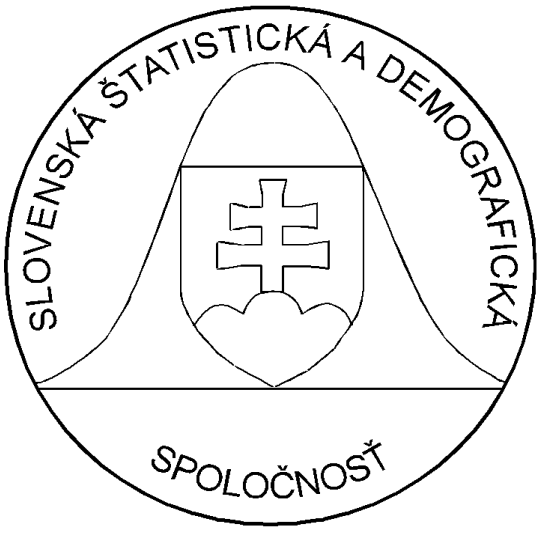
7/2008

FORUM STATISTICUM SLOVACUM



ISSN 1336-7420
87
9 771336 742001

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

Naše najbližšie akcie:

(pozri tiež www.ssds.sk, blok Poriadané akcie)

NITRIANSKE ŠTATISTICKÉ DNI
5. – 6. 2. 2009, Nitra

Pohľady na ekonomiku Slovenska 2009
7. 4. 2009, Bratislava

EKOMSTAT 2009, 23. škola štatistiky
tematické zameranie: *Štatistické metódy vo vedecko-výskumnej, odbornej a hospodárskej praxi.*
31. 5. 2009 – 5. 6. 2009, Trenčianske Teplice

Aplikácie metód na podporu rozhodovania vo vedeckej, technickej a spoločenskej praxi
jún 2009, STU Bratislava

Slovensko-česká konferencia PRASTAN
jún 2009 Kočovce

12. SLOVENSKÁ DEMOGRAFICKÁ KONFERENCIA
tematické zameranie: Demografická budúcnosť Slovenska
23. – 25. 9. 2009, Trenčiansky kraj

FernStat 2008
VI. medzinárodná konferencia aplikovanej štatistiky
(**F**inancie, **E**konomika, **R**iadenie, **N**ázory)
tematické zameranie: *Aplikovaná, demografická, matematická štatistika, štatistické riadenie kvality.*
október 2009, hotel Lesák, Tajov pri Banskej Bystrici

18. Medzinárodný seminár VÝPOČTOVÁ ŠTATISTIKA,
3. – 4. 12. 2009, Bratislava, Infostat

Prehliadka prác mladých štatistikov a demografov
3. 12. 2009, Bratislava, Infostat

Regionálne akcie
priebežne

Pokyny pre autorov

Jednotlivé čísla vedeckého časopisu FORUM STATISTICUM SLOVACUM 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. Názov word-súboru uvádzajte a posielajte v tvare: ***priezvisko_nazovakcie.doc***

Forma: Príspevky písané výlučne len v textovom editore MS WORD, verzia 6 a vyššia do verzie 2007, 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 členom redakčnej rady alebo externým oponentom.

Štruktúra príspevku: ***(Pri písaní príspevku využite elektronickú šablónu: http://www.ssds.sk/ v časti Vedecký časopis, Pokyny pre autorov.)***

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, centrovať)

Vynechať riadok

Abstract: Text abstraktu v anglickom jazyku (resp. v slovenskom jazyku ak je článok napísaný v anglickom jazyku), max. 10 riadkov (štýl normálny: Time New Roman 12).

Vynechať riadok

Key words: Kľúčové slová v anglickom jazyku, max. 2 riadky (štýl normálny: Time New Roman 12).

Vynechať riadok

Kľúčové slová: Kľúčové slová v slovenskom jazyku, resp. v jazyku akom je napísaný príspevok, max. 2 riadky (štýl normálny: Time New Roman 12).

Vynechať riadok

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 2. . .**
4. **Záver** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

Vlastný text jednotlivý 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 nevynechávajte.

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) *(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
Ulica1
970 00 Mesto1
meno1.priezvisko1@mail.sk

Meno2 Priezvisko2, tituly2
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
02/63812565

e-mail
chajdiak@statis.biz
Jan.Luha@statistics.sk

Registráciu vykonalo
Ministerstvo kultúry Slovenskej republiky

Registračné číslo
3416/2005

Tematická skupina
B1

Dátum registrácie
22. 7. 2005

Objednávky
Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava 24
Slovenská republika
IČO: 178764
Číslo účtu: 0011469672/0900

ISSN 1336-7420

Ročník
IV.

Číslo
7/2008

Cena výtlačku 500 SKK / 20 EUR
Ročné predplatné 1500 SKK / 60 EUR

OBSAH

Prehliadka prác mladých štatistikov a demografov

CONTENT

Review of papers of young statisticians and demographers

ÚVOD

Vážené kolegyne, vážení kolegovia,

siedme číslo štvrtého ročníka vedeckého recenzovaného časopisu Slovenskej štatistickej a demografickej spoločnosti (SŠDS) je zostavené z príspevkov, ktoré sú obsahovo orientované v súlade s tematikou 17. medzinárodného seminára Výpočtová štatistika a Prehliadkou prác mladých štatistikov a demografov. Tieto akcie sa uskutočnili v dňoch 4. a 5. decembra 2008 na Infostate v Bratislave. Organizátormi oboch podujatí v tomto roku boli SŠDS v spolupráci s Fakultou managementu Univerzity Komenského v Bratislave, Infostatom, Štatistickým úradom SR a spoločnosťou SAS Slovensko.

Akcie, z poverenia Výboru SŠDS, zorganizoval organizačný a programový výbor v zložení: doc. Ing. Jozef Chajdiak, CSc. – predseda, Ing. Iveta Stankovičová, PhD. - podpredseda, RNDr. Ján Luha, CSc. – tajomník, Prof. RNDr. Beáta Stehlíková, CSc., Doc. RNDr. Bohdan Linda, CSc., Doc. Dr. Jana Kubanová, CSc., Ing. Mária Kanderová, PhD., Ing. Vladimír Úradníček PhD., RNDr. Samuel Koróny.

Na príprave a zostavení tohto čísla participovali: doc. Ing. Jozef Chajdiak, CSc., RNDr. Ján Luha, CSc., Ing. Iveta Stankovičová, PhD.

Recenziu príspevkov zabezpečili: doc. Ing. Jozef Chajdiak, CSc., RNDr. Ján Luha, CSc., Ing. Iveta Stankovičová, PhD., RNDr. Samuel Koróny, RNDr. Mária Bohdalová, 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 zvyšuje aj odbornú úroveň mladých štatistikov a demografov.

Výbor SŠDS

Z histórie seminárov Výpočtová štatistika
From the History of Seminars Computational Statistics

Pri príležitosti 17. ročníka semináru Výpočtová štatistika uvádzame stručnú chronológiu predošlých 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 a väčšina seminárov sa organizovala v spolupráci so Štatistickým úradom SR (resp. SŠU v Bratislave) a Infostat-om Bratislava (resp. VUSEIaR Bratislava).

Druhý seminár prebehol 8. 12. 1987, tretí seminár 11. - 12. 12.1990. Pád socializmu a spoločenské zmeny spôsobili určitú prestávku v organizácii seminárov Výpočtovej štatistiky. 4. seminár sa uskutočnil až 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 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 a

17. medzinárodný seminár Výpočtová štatistika sa uskutočnil 4. - 5. 12. 2008.

Príspevky 2. seminára boli opublikované v Informáciách SDŠS č. 1/1999 a od 3. seminára sa publikujú v samostatnom Zborníku príspevkov príslušného seminára. Od 14. seminára sú príspevky publikované vo vedeckom recenzovanom časopise SŠDS s názvom FORUM STATISTICUM SLOVACUM.

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 boli 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 aktuálnom ročníku 17. seminára Výpočtová štatistika bolo prihlásených 15 prác mladých autorov.

Prípadní záujemcovia z radov mladých štatistikov a demografov (za mladých považujeme štatistikov a demografov pred ukončením vysokej školy) môžu získať informácie na www.ssds.sk , blok akcie a na e-mailových adresách:

chajdiak@statis.biz
Jan.Luha@statistics.sk
iveta.stankovicova@fm.uniba.sk

Informácie o najbližšom seminári získate na webovskej stránke SŠDS *http://www.ssds.sk/* resp. *http://www.statistics.sk* v bloku Slovenská štatistická a demografická spoločnosť.

Doc. Ing. Jozef Chajdiak, CSc.
vedecký tajomník SŠDS

RNDr. Ján Luha, CSc.
člen sekretariátu Výboru SŠDS

Value at risk of an option contract based upon a Monte Carlo simulation
Odhad value at risk opčného kontraktu založený na simulácii Monte Carlo

Martin Boďa, Tomáš Osička

Abstrakt: Článok predkladá návrh simulácie Monte Carlo pre odhad value at risk opcie na akcie nevyplácajúcu dividendu. Hlavná myšlienka spočíva v použití Taylorovho vzorca na Blackovu-Scholesovou opčnú oceňovaciu formulu a v predpoklady viacrozmerného normálneho rozdelania výnosov rizikových faktorov.

Kľúčové slová: value at risk, európska opcia na akciu nevyplácajúcu dividendu, delta-gama aproximácia, Choleského rozklad, simulácia Monte Carlo

Key words: value at risk, European option on a no-dividend paying stock, delta-gamma approximation, the Cholesky decomposition, the Monte Carlo simulation.

1. The introduction

It has been known since 1973 that the price of a European option on a no-dividend paying stock can be explicitly written as a function of five variables: the spot price of the underlying stock (S), the volatility of the stock price (σ), the riskless rate of interest (r), the exercise time (T), and the strike price (K). Here it must be said that it is assumed that the stock price follows a geometric Brownian motion and that the parameter σ is the respective noise factor of this process. In the modeling of value at risk of an option contract, these variables are viewed as a source of risk and are called risk factors. These variables, but one, change over time and their changes affect the market price of the option. The investor owning such an option either loses or gains in consequence of these changes. The task is then to determine, analytically or by way of simulation, the potential loss of the option contract that is not to be exceeded at a specified confidence level. In the article, the joint normality of the risk factor returns is assumed and a Monte Carlo simulation, based on the Cholesky factorization of the covariance matrix of the considered risk factor returns, is employed so as to arrive at the desired estimate of value at risk.

2. The delta-gamma approximation of the changes in the option value

The best theoretical approximation of the prices of a European-style option on a stock paying no dividend is the Black-Scholes fair value and, in accord with the notes above, the value of an option is then $V \equiv V(S, \sigma, r, T, K)$. For the strike price K is constant over the entire time to maturity, in no measure does it contribute to the riskiness of the option and there is no need to be allowed for. Taking a second-order Taylor series approximation of the change in the option value, one obtains

$$\Delta V \approx \Delta S \frac{\partial V}{\partial S} + \Delta\sigma \frac{\partial V}{\partial \sigma} + \Delta r \frac{\partial V}{\partial r} + \Delta T \frac{\partial V}{\partial T} + \frac{1}{2}(\Delta S)^2 \frac{\partial^2 V}{\partial S^2}, \tag{1}$$

in which the other terms of the expansion are neglected as comparatively very small. Making a consideration for the fact that, all though T varies over time, its changes can be determined at any time for the future (a one-day change of T is always $\Delta T = 1/365$); the approximation (1)

can be split into a deterministic component $\Delta V_{\text{deterministic}}$ and into a stochastic, risky, component $\Delta V_{\text{stochastic}}$. The consideration leads to

$$\Delta V \approx \Delta V_{\text{deterministic}} + \Delta V_{\text{stochastic}}, \tag{2a}$$

with

$$\Delta V_{\text{deterministic}} = \Delta T \theta, \tag{2b}$$

$$\Delta V_{\text{stochastic}} = \Delta S\, \delta + \Delta\sigma \upsilon + \Delta r\, \rho + \frac{1}{2}(\Delta S)^2 \gamma, \tag{2c}$$

where the partial derivatives are replaced by the Greek letters. The details upon the fair value as determined by the Black-Scholes option pricing equation and the Greeks are given in Scheme 1.

	CALL OPTION	**PUT OPTION**
FAIR VALUE	$V = S\,\Phi(d_1) - K e^{-rT}\,\Phi(d_2)$	$V = Ke^{-rT}\,\Phi(-d_2) - S\,\Phi(-d_1)$
DELTA $\delta = \frac{\partial V}{\partial S}$	$\Phi(d_1)$	$-\Phi(-d_1)$
RHO $\rho = \frac{\partial V}{\partial r}$	$Ke^{-rT}T\,\Phi(d_2)$	$-Ke^{-rT}T\,\Phi(-d_2)$
VEGA $\upsilon = \frac{\partial V}{\partial \sigma}$	$S\phi(d_1)\sqrt{T}$	
THETA $\theta = \frac{\partial V}{\partial T}$	$-\frac{S\sigma\phi(d_1)}{2\sqrt{T}} - \frac{r}{T}\rho_C$	$-\frac{S\sigma\phi(d_1)}{2\sqrt{T}} - \frac{r}{T}\rho_P$
GAMMA $\gamma = \frac{\partial^2 V}{\partial S^2}$	$\frac{\phi(d_1)}{S\sigma\sqrt{T}}$	

Whilst $d_1 = \dfrac{\ln(S/K) + (r + 0.5\sigma^2)T}{\sigma\sqrt{T}}$, $d_2 = \dfrac{\ln(S/K) + (r - 0.5\sigma^2)T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T}$,

$\phi(x) = (2\pi)^{-1} e^{-x^2/2}$ is the probability distribution function of $N(0,1)$ evaluated at x, and

$\Phi(x) = \int_{-\infty}^{x} \phi(y)\,dy$ is the cumulative probability distribution function of $N(0,1)$ evaluated at x.

Scheme 1 *The Black-Scholes option pricing formulæ and the Greeks for European-style call and put options*

It is evident that, to derive value at risk of an option, it suffices to work only with (2c) which advances into the form

$$\Delta V_{\text{stochastic}} = r_S\left[\delta S\right] + r_\sigma\left[\upsilon\sigma\right] + r_r\left[\rho r\right] + \frac{1}{2}(r_S)^2\left[\gamma S^2\right], \tag{3}$$

in which $r_S = \Delta S/S$, $r_\sigma = \Delta\sigma/\sigma$ and $r_r = \Delta r/r$ are the discrete returns of the risk factors S, σ, and r. It is suggestive itself that r_S, r_σ and r_r might be jointly normal, id est

$$\begin{pmatrix} r_S \\ r_\sigma \\ r_r \end{pmatrix} \sim N_3(\mathbf{0}; \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} - p.d. \tag{4}$$

It follows forthwith that

$$(r_S)^2 \sim \{\Sigma\}_{SS}\, \chi^2(1) \tag{5}$$

and that r_S and $(r_S)^2$ are dependent. One might attempt at deriving the distribution for $\Delta V_{\text{stochastic}}$, however, it may happen that these attempts are of less avail. Thus, it is advisable to employ a Monte Carlo procedure instead.

3. The simulation framework

It is naïve to build upon the assumption that the only unknown parameter $\mathbf{\Sigma}$ is readily available, and it is clear that it must needs be estimated out of a sufficient number of historical observations of r_S, r_σ and r_r . Therefore the procedure takes these steps:

1. Compute the historical risk factor returns r_S, r_σ and r_r .
2. Estimate the covariance matrix $\mathbf{\Sigma}$ in (4). The method of estimation is a matter of choice: it can be a conventional non-weighted case, an approach with exponential weights or an estimate from a multivariate GARCH model, for example. The criterion is that the method should produce a positively definite matrix.
3. Effect the Cholesky decomposition of est$\mathbf{\Sigma}$, so that est$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^{\mathrm{T}}$.
4. Generate a large number n (say n = 10 000 or n = 50 000) of $N_3(\mathbf{0}, \mathbf{I}_3)$ drawings $\mathbf{U}_{3,\,n}$.
5. Compute the product $\mathbf{BU}$. The rows of $(\mathbf{BU})_{3,\,n}$ correspond to random numbers of r_S, r_σ and r_r . Square the row of r_S to obtain random numbers for $(r_S)^2$.
6. For each of n quadruplets of r_S, r_σ, r_r and $(r_S)^2$ determine the value of $\Delta V_{\text{stochastic}}$. Compute $\Delta V_{\text{deterministic}}$, which is the same for each of the simulated quadruplets. The expressions δS, ρr, $\upsilon\sigma$, γS^2 and $\Delta T\theta$ are computed from the newest observations of S, r, σ, and from the value of T and K.
7. Sum the value of $\Delta V_{\text{stochastic}}$ and the value of $\Delta V_{\text{deterministic}}$ for each simulation case.
8. Find the α-percentile of the sums in step 7.
   To find the α-quantile in the final step means to determine the value at risk at level α.

4. The summary

To estimate value at risk of an option contract it appears convenient to utilize a simulation approach rather than an analytic one. This preference stems from the fact that option contracts are far from linear forms and it is difficult and even beyond any benefit to derive the distribution of changes in the option value. On the foundation of normality of the relevant risk factor returns of the option contract, in the article a Monte Carlo simulation approach is presented. Under this approach value at risk of the option is obtained (at a given level α) as a sum of the deterministic changes in the option value (due to the changing exercise time) and of the stochastic changes (due to the stock price, its volatility and the riskless rate).

5. The references

[1.] ANDĚL, Jiří 2005. Základy matematické statistiky. Praha : Matfyzpress, 2005. 358 pp. ISBN 80-86732-40-1.

[2.] MELICHERČÍK, Igor, OLŠAROVÁ, Ladislava, ÚRADNÍČEK, Vladimír 2005. Kapitoly z finančnej matematiky. Bratislava : Epos, 2005. 242 pp. ISBN 80-8057-651-3.

The authors‘address

Martin Boďa, Ing. et Bc.
Univerzita Mateja Bela v Banskej Bystrici
Ekonomická fakulta, KKMaIS
Tajovského 10, 975 90 Banská Bystrica
martin.boda@umb.sk

Tomáš Osička
Univerzita Mateja Bela v Banskej Bystrici
Ekonomická fakulta
Tajovského 10, 975 90 Banská Bystrica
tomas.osicka@gmail.com

Properties of the insensitivity region for a linear function of the fixed effects parameters in the mixed linear model
Vlastnosti oblasti necitlivosti pro lineární funkci parametrů střední hodnoty ve smíšeném lineárním modelu

Hana Boháčová, Jana Heckenbergerová

Abstract: The aim of this paper is to find an explicit form of an insensitivity region for a linear function of the fixed effects parameters in a regular mixed linear regression model without constraints, to explore a possible graphical representation and check the influence of the used linearization.

Key words: Insensitivity region, mixed linear regression model, fixed effects parameters, locally best linear unbiased estimator of the fixed effects parameters, variance components.

Abstrakt: Cílem příspěvku je najít explicitní vyjádření oblasti necitlivosti pro lineární funkci parametrů střední hodnoty ve smíšeném lineárním modelu bez podmínek, vysvětlit, jak je možné tuto oblast graficky interpretovat a ověřit, jaký vliv má linearizace použitá při odvozování.

Klíčová slova: Oblast necitlivosti, smíšený lineární regresní model, parametry střední hodnoty, lokálně nejlepší lineární nestranný odhad, varianční komponenty.

1. Introduction

The real values of the variance components are usually not known in a mixed linear model. It means we need to estimate them from the measured data and these estimates are to be used to estimate the fixed effects parameters instead of their actual, proper values.

2. Basic symbols

$\boldsymbol{I}$	identity matrix
$\boldsymbol{A}^+$	Moore-Penrose generalized inverse of the matrix $\boldsymbol{A}$
$\mathcal{M}(\boldsymbol{A})$	column space of the matrix $\boldsymbol{A}$
$\boldsymbol{M}_{\boldsymbol{A}}$	projection matrix (in the Euclidean norm) on a space orthogonal to $\mathcal{M}(\boldsymbol{A})$, $(\boldsymbol{M}_{\boldsymbol{A}} = \boldsymbol{I}-\boldsymbol{A}\boldsymbol{A}^+)$
$r(\boldsymbol{A})$	rank of the matrix $\boldsymbol{A}$
$tr(\boldsymbol{A})$	trace of the matrix $\boldsymbol{A}$, it is defined for square matrices as a sum of its diagonal elements
$\boldsymbol{Y} \sim (\boldsymbol{X\beta}, \boldsymbol{\Sigma})$	the mean value of the n-dimensional random vector $\boldsymbol{Y}$ is $\boldsymbol{X\beta}$ and its covariance matrix is $\boldsymbol{\Sigma}$
$Var_{\theta 0}\,[\,\hat{\beta}\,(\boldsymbol{\theta})]$	the covariance matrix of the estimator $\hat{\beta}$ when the parameter $\boldsymbol{\theta_0}$ is under consideration
θ_i	the i-th component of the vector $\boldsymbol{\theta}$

3. Estimator of the fixed effects parameters in the mixed linear model

Let us consider a mixed linear model

$$\boldsymbol{Y} \sim (\boldsymbol{X\beta}, \boldsymbol{\Sigma_\theta}). \tag{1}$$

$\boldsymbol{X}$ is a known matrix of a $n \times k$ dimension which is of a full column rank k. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is a vector of unknown fixed effects parameters. The covariance matrix is of a form

$$\text{Var } \boldsymbol{Y} = \boldsymbol{\Sigma_\theta} = \sum_{i=1}^{r} \theta_i \boldsymbol{V}_i , \tag{2}$$

where $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_r$ are known linear independent symmetrical positive definite matrices and $\theta_1, \ldots, \theta_r > 0$ are unknown variance components. As we consider a regular model we take into account only $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)' \in \mathcal{S}$, where $\mathcal{S}$ is a set of such $\boldsymbol{\theta}$ for which $\boldsymbol{\Sigma_\theta}$ is positive definite.

Let us use a given $\boldsymbol{\theta_0} \in \mathcal{S}$, obviously we receive $\boldsymbol{\theta_0}$ as some estimate of $\boldsymbol{\theta}$, MINQUE is an often used method of determining the estimates, see [5] for more details on this method. We know the $\boldsymbol{\theta_0}$–locally best linear unbiased estimator of fixed effects parameters is (see [1])

$$\hat{\beta}(\boldsymbol{\theta_0}) = [\boldsymbol{X}'(\boldsymbol{\Sigma_{\theta 0}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma_{\theta 0}})^{-1}\boldsymbol{Y}. \tag{3}$$

Here $\boldsymbol{\Sigma_{\theta 0}}$ denotes a matrix of type (2) with $\boldsymbol{\theta_0}$ instead of $\boldsymbol{\theta}$. The covariance matrix of this estimator (when we consider $\boldsymbol{\theta_0}$ to be the real value of $\boldsymbol{\theta}$) is

$$\text{Var}_{\boldsymbol{\theta 0}}[\hat{\beta}(\boldsymbol{\theta_0})] = [\boldsymbol{X}'(\boldsymbol{\Sigma_{\theta 0}})^{-1}\boldsymbol{X}]^{-1}. \tag{4}$$

Let us further consider a linear function $\boldsymbol{h'\beta}$ of the fixed effects parameters, $\mathrm{h} \in \mathrm{R}^{\mathrm{k}}$. Its $\boldsymbol{\theta_0}$–locally best linear unbiased estimator is (according to (3))

$$\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta_0}) = \boldsymbol{h}'[\boldsymbol{X}'(\boldsymbol{\Sigma_{\theta 0}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma_{\theta 0}})^{-1}\boldsymbol{Y}. \tag{5}$$

The covariance matrix of the estimator (5) is according to (4)

$$\text{Var}_{\boldsymbol{\theta 0}}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta_0})] = \boldsymbol{h}'[\boldsymbol{X}'(\boldsymbol{\Sigma_{\theta 0}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{h}. \tag{6}$$

As we can see estimator (5) is a function of $\boldsymbol{\theta}_0$ (via matrix $\boldsymbol{\Sigma_{\theta 0}}$). We can ask what happens when we change the input value $\boldsymbol{\theta_0}$ of the variance components with some small $\delta\boldsymbol{\theta} = (\delta\theta_1, \ldots, \delta\theta_r)$. We will look for a set of such admissible input values $\boldsymbol{\theta}_0 + \delta\boldsymbol{\theta}$ that cause an ε-multiple increase of the standard deviation of the estimator of $\boldsymbol{h'\beta}$ at most when compared to the standard deviation of the estimator of $\boldsymbol{h'\beta}$ based on $\boldsymbol{\theta}_0$.

4. Insensitivity region of the linear function $\boldsymbol{h'\beta}$

According to the above mentioned idea we will look for a set $\mathcal{N}_{\boldsymbol{h'\beta},\boldsymbol{\theta 0}}$ of those input values $\boldsymbol{\theta}_0 + \delta\boldsymbol{\theta}$ which satisfy

$$\boldsymbol{\theta}_0 + \delta\boldsymbol{\theta} \in \mathcal{N}_{\boldsymbol{h'\beta},\boldsymbol{\theta 0}} \Rightarrow$$
$$\Rightarrow \{Var_{\boldsymbol{\theta 0}}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta_0}+\delta\boldsymbol{\theta})]\}^{1/2} \leq (1+\varepsilon)\{Var_{\boldsymbol{\theta 0}}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta_0})]\}^{1/2} . \tag{7}$$

This set is called an insensitivity region for the linear function $\boldsymbol{h'\beta}$ of the fixed effects parameters.

$Var_{\boldsymbol{\theta 0}}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta_0})]$ is specified in (6), $Var_{\boldsymbol{\theta 0}}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta_0}+\delta\boldsymbol{\theta})]$ that is used on the right hand side of the inequality (7) needs to be found. We can approximate $\hat{\beta}(\boldsymbol{\theta_0}+\delta\boldsymbol{\theta})$ with help of the Taylor polynomial of the first order as

$$\hat{\beta}(\boldsymbol{\theta_0}+\delta\boldsymbol{\theta}) \approx \hat{\beta}(\boldsymbol{\theta_0}) + \frac{\partial\hat{\beta}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\delta\boldsymbol{\theta}. \tag{8}$$

As we have

$$\frac{\partial \hat{\beta}(\theta_0)}{\partial \theta_i} = - [\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\, \boldsymbol{V}_i(\boldsymbol{\Sigma}_{\theta 0})^{-1}[\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}(\boldsymbol{\theta}_0)] \tag{7}$$

we can write in accordance with (8)

$$\beta(\boldsymbol{\theta}_0+\delta\boldsymbol{\theta}) \approx \beta(\boldsymbol{\theta}_0) - \sum_{i=1}^{r} [\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\, \boldsymbol{V}_i(\boldsymbol{\Sigma}_{\theta 0})^{-1}[\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}(\boldsymbol{\theta}_0)]\delta\theta_i. \tag{8}$$

Following approximate expression can be based on (8)

$$Var_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0+\delta\boldsymbol{\theta})] \approx Var_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0)] +$$
$$+ \boldsymbol{h}'[\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{\Sigma}_{\delta\theta}(\boldsymbol{M}_X\boldsymbol{\Sigma}_{\theta 0}\boldsymbol{M}_X)^{+}\boldsymbol{\Sigma}_{\delta\theta}(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}[\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{h}. \tag{9}$$

$\boldsymbol{\Sigma}_{\delta\theta}$ denotes the matrix $\sum_{i=1}^{r}\delta\theta_i\,\boldsymbol{V}_i$. The resulting insensitivity region is (cf. [4], page 167, [2] or [3]):

$$\mathcal{N}_{h'\beta,\theta 0} = \{\boldsymbol{\theta}_0+\delta\boldsymbol{\theta}: (\delta\boldsymbol{\theta})'\boldsymbol{W}_h\delta\boldsymbol{\theta} \le (2\varepsilon+\varepsilon^2)\boldsymbol{h}'[\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{h}\}. \tag{10}$$

$\boldsymbol{W}_h$ is a matrix whose elements are

$$\{\boldsymbol{W}_h\}_{i,j} = \boldsymbol{h}'[\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{V}_i(\boldsymbol{M}_X\boldsymbol{\Sigma}_{\theta 0}\boldsymbol{M}_X)^{+}\boldsymbol{V}_j(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}[\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{h} \tag{11}$$

for i,j = 1, …, r.

Remark 1: *ε is chosen to be a small positive number, usually we have $\varepsilon < 0{,}5$, it means we can replace $(2\varepsilon+\varepsilon^2)$ with 2ε in the set described in (10) without doing any serious intervention into $\mathcal{N}_{h'\beta,\theta 0}$.*

In [4] following assertion can be found:

Lemma 1: *The matrix $\boldsymbol{W}_h$ which is assigned to the linear function $\boldsymbol{h}'\boldsymbol{\beta}$ satisfies*

$$\mathcal{M}(\boldsymbol{W}_h) \perp \boldsymbol{\theta}_0. \tag{12}$$

The lemma says the column space of the matrix $\boldsymbol{W}_h$ is orthogonal to the given vector $\boldsymbol{\theta}_0$. It means $\boldsymbol{W}_h$ (which is of the dimension r×r) needs to be singular. When we consider the special case of $r = 2$ we can interpret the lemma above in the following way: The quadratic form inside of the set (10) determines a singular conic – with respect to (12) it is a zone midway between a couple of the parallel lines whose direction vector is $\boldsymbol{\theta}_0$.

Remark 2: *We have to comprehend that a linearization was used when deriving the insensitivity region. It means that the inequality (7) may be valid only in the immediate neighbourhood of the point $\boldsymbol{\theta}_0$.*

Let us have a look at the fulfilment of this inequality within the whole zone in a concrete case.

5. Insensitivity region for a linear function of the fixed effects parameters in case of Michaelis-Menten model

Let the Michaelis-Menten regression function

$$f(x) = \frac{\gamma_1 x}{\gamma_2 + x} \tag{13}$$

with two unknown parameters γ_1 and γ_2 be considered. This kind of function is used in the kinetics of enzymes. Following design of the experiment is proposed: we will measure the function values for x_1=0,5, x_2=1,5 with the unknown dispersion of σ_1^2 and for x_3=7, x_4=9

with the dispersion of σ_2^2 which is also unknown. We have two measurements in each of the points, so our observation vector is $\boldsymbol{Y}=(Y_1, ..., Y_8)'$. Our task is to estimate γ_1 and γ_2 using the observation vector. The estimates will be based on the following mixed linear model that arises from the linearization of (13) using prior values γ_1^0, γ_2^0

$$\boldsymbol{Y}-\boldsymbol{f_0} \sim (\boldsymbol{X\beta}, \boldsymbol{\Sigma_\theta}), \tag{14}$$

where

$$\boldsymbol{f_0}=\left(\frac{\gamma_1^0 x_1}{\gamma_2^0+x_1}, \frac{\gamma_1^0 x_1}{\gamma_2^0+x_1}, \frac{\gamma_1^0 x_2}{\gamma_2^0+x_2}, \frac{\gamma_1^0 x_2}{\gamma_2^0+x_2}, \frac{\gamma_1^0 x_3}{\gamma_2^0+x_3}, \frac{\gamma_1^0 x_3}{\gamma_2^0+x_3}, \frac{\gamma_1^0 x_4}{\gamma_2^0+x_4}, \frac{\gamma_1^0 x_4}{\gamma_2^0+x_4}\right)', \tag{15}$$

$$\boldsymbol{X}=\begin{pmatrix} \frac{x_1}{\gamma_2^0+x_1} & \frac{x_1}{\gamma_2^0+x_1} & \frac{x_2}{\gamma_2^0+x_2} & \frac{x_2}{\gamma_2^0+x_2} & \frac{x_3}{\gamma_2^0+x_3} & \frac{x_3}{\gamma_2^0+x_3} & \frac{x_4}{\gamma_2^0+x_4} & \frac{x_4}{\gamma_2^0+x_4} \\ -\frac{\gamma_1^0 x_1}{(\gamma_2^0+x_1)^2} & -\frac{\gamma_1^0 x_1}{(\gamma_2^0+x_1)^2} & -\frac{\gamma_1^0 x_2}{(\gamma_2^0+x_2)^2} & -\frac{\gamma_1^0 x_2}{(\gamma_2^0+x_2)^2} & -\frac{\gamma_1^0 x_3}{(\gamma_2^0+x_3)^2} & -\frac{\gamma_1^0 x_3}{(\gamma_2^0+x_3)^2} & -\frac{\gamma_1^0 x_4}{(\gamma_2^0+x_4)^2} & -\frac{\gamma_1^0 x_4}{(\gamma_2^0+x_4)^2} \end{pmatrix}', \tag{16}$$

$$\boldsymbol{\beta}=\begin{pmatrix}\delta\gamma_1\\ \delta\gamma_2\end{pmatrix}=\begin{pmatrix}\gamma_1-\gamma_1^0\\ \gamma_2-\gamma_2^0\end{pmatrix},\ \boldsymbol{\theta}=\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}=\begin{pmatrix}\sigma_1^2\\ \sigma_2^2\end{pmatrix}, \tag{17}$$

$$\boldsymbol{\Sigma_\theta}=\theta_1\begin{pmatrix}1&0&0&0&0&0&0&0\\0&1&0&0&0&0&0&0\\0&0&1&0&0&0&0&0\\0&0&0&1&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\end{pmatrix}+\theta_2\begin{pmatrix}0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&1&0&0&0\\0&0&0&0&0&1&0&0\\0&0&0&0&0&0&1&0\\0&0&0&0&0&0&0&1\end{pmatrix}. \tag{18}$$

We will not focus on the estimates of $\boldsymbol{\beta}$ in what follows, we can get them in accordance with (3). The insensitivity region for a linear function of $\boldsymbol{\beta}$ is in the centre of our attention. In our case r=2 (as we have two variance components σ_1^2 and σ_2^2). Let us use $\boldsymbol{h}=(1,0)'$. In (10) and (11) we can see we need vector $\boldsymbol{h}$, matrices $\boldsymbol{X}$ (which is given in (16) and depends on the prior values γ_1^0, γ_2^0), $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$ (known from 18) and $\boldsymbol{\Sigma_{\theta 0}}$ (this matrix will be of the same form as matrix $\boldsymbol{\Sigma_\theta}$ in (18) but using the prior value $\boldsymbol{\theta_0}$ instead of $\boldsymbol{\theta}$). It means we need to get some initial values γ_1^0, γ_2^0 and $\boldsymbol{\theta_0}$. They will be based on the concrete observation vector

$$\boldsymbol{Y}=(1{,}245;\ 0{,}779;\ 2{,}264;\ 2{,}258;\ 6{,}612;\ 5{,}909;\ 6{,}827;\ 6{,}301)'. \tag{19}$$

We will use γ_1^0=16,068; γ_2^0=8,250; $\boldsymbol{\theta_0}$=(0,0543; 0,1927)'. The appropriate insensitivity region for ε=0,1 is a set

$$\mathcal{N}_{h'\beta,\theta 0}=\left\{\boldsymbol{\theta_0}+\delta\boldsymbol{\theta}:\ (\delta\boldsymbol{\theta})'\begin{pmatrix}24{,}171 & -6{,}811\\ -6{,}811 & 1{,}919\end{pmatrix}\delta\boldsymbol{\theta}\le 0{,}619\right\}. \tag{20}$$

The insensitivity region is shown in figure 1 together with the relative position of the insensitivity region (20) and vector $\boldsymbol{\theta_0}$ resulting from lemma 1. Let us target the problem mentioned in remark 2. We will use three points from the insensitivity region $\boldsymbol{\theta_1}$=(0,203; 0,151)', $\boldsymbol{\theta_2}$=(0,301; 0,500)' and $\boldsymbol{\theta_3}$=(0,100; 0,200)' and investigate the dispersion of the estimator of $\boldsymbol{h'\beta}$ based on these points but calculated for the real value of $\boldsymbol{\theta}$ considered to be equal to $\boldsymbol{\theta_0}$ according to

$$\mathrm{Var}_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0+\delta\boldsymbol{\theta})] = \boldsymbol{h}'[\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0+\delta\theta})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0+\delta\theta})^{-1}\boldsymbol{\Sigma}_{\theta 0}(\boldsymbol{\Sigma}_{\theta 0+\delta\theta})^{-1}\boldsymbol{X}[\boldsymbol{X}'(\boldsymbol{\Sigma}_{\theta 0+\delta\theta})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{h}. \quad (21)$$

This dispersion was replaced by its linearized form (9) when the insensitivity region was derived. This linearization may cause that the real dispersion $\mathrm{Var}_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0+\delta\boldsymbol{\theta})]$ exceeds $\mathrm{Var}_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0)]$ more than by the admissible (1+ε)-multiple and we have to check this for being sure.

The changes $\delta\boldsymbol{\theta}$ corresponding to $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ are

$$\begin{aligned}(\delta\boldsymbol{\theta})_1 &= \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0 = (0{,}149;\ -0{,}0417)',\\ (\delta\boldsymbol{\theta})_2 &= \boldsymbol{\theta}_2 - \boldsymbol{\theta}_0 = (0{,}247;\ -0{,}307)',\\ (\delta\boldsymbol{\theta})_3 &= \boldsymbol{\theta}_3 - \boldsymbol{\theta}_0 = (0{,}0457;\ -0{,}0073)'.\end{aligned} \quad (22)$$

The location of the points within the insensitivity region is also displayed in figure 1.

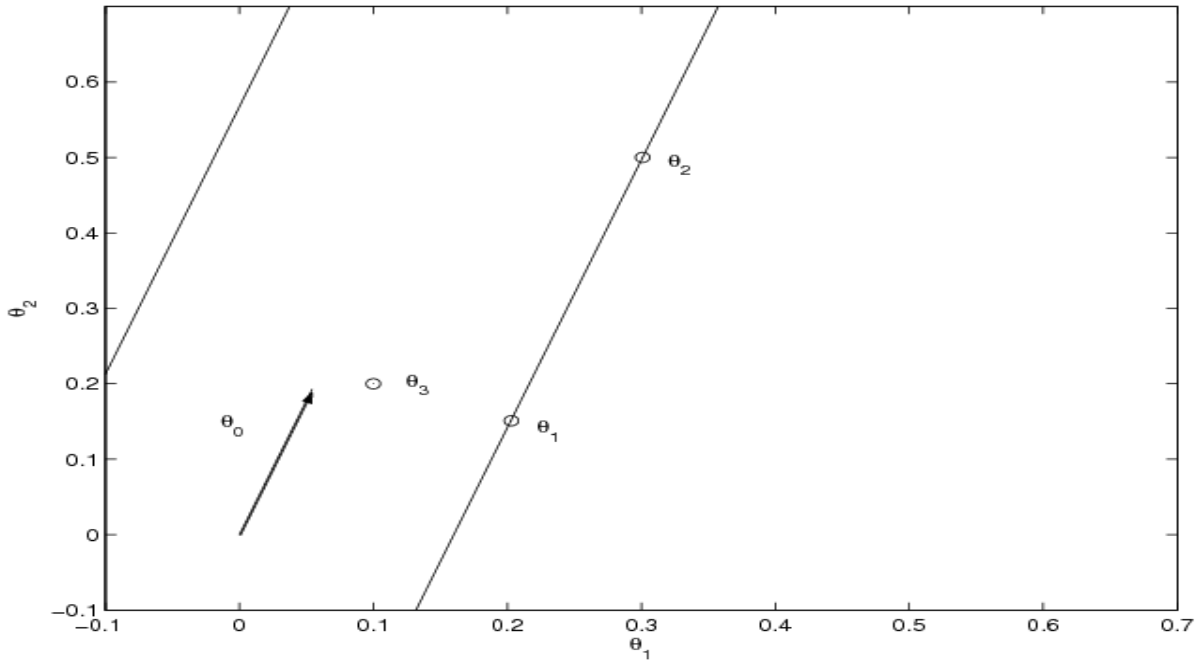


Figure 1: *Insensitivity region for a linear function* $\boldsymbol{h}'\boldsymbol{\beta}$.

Using (21) we get

$$\begin{aligned}\mathrm{Var}_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0+(\delta\boldsymbol{\theta})_1)] &= 3{,}854,\\ \mathrm{Var}_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0+(\delta\boldsymbol{\theta})_2)] &= 3{,}177,\\ \mathrm{Var}_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0+(\delta\boldsymbol{\theta})_3)] &= 3{,}133.\end{aligned} \quad (23)$$

According to (7) it should be compared to $Var_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0)]$ (counted from (6))

$$Var_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0)] = 3{,}094. \quad (24)$$

As stated in (7) we are interested in the ratio of the standard deviations $\{Var_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0+\delta\boldsymbol{\theta})]\}^{1/2}$ and $\{Var_{\theta 0}[\boldsymbol{h}'\hat{\beta}(\boldsymbol{\theta}_0)]\}^{1/2}$. It should not exceed (1+ε)=1,1. From (23) and (24) we have

$$\frac{\sqrt{Var_{\theta 0}\left[h'\beta(\theta_1)\right]}}{\sqrt{Var_{\theta 0}\left[h'\beta(\theta_0)\right]}} = 1{,}116 \quad (25)$$

As we can see in the figure 1 the point $\boldsymbol{\theta}_1$ is on the bound of the insensitivity region. The increase of the standard deviation is slightly higher than 1,1.

Next we have

$$\frac{\sqrt{Var_{\theta 0}\left[h'\beta(\theta_2)\right]}}{\sqrt{Var_{\theta 0}\left[h'\beta(\theta_0)\right]}} = 1{,}013 \tag{26}$$

and

$$\frac{\sqrt{Var_{\theta 0}\left[h'\beta(\theta_1)\right]}}{\sqrt{Var_{\theta 0}\left[h'\beta(\theta_0)\right]}} = 1{,}007, \tag{27}$$

which is in conformity with the used ε=0,1.

6. Conclusion

As shown at the close of the previous paragraph the real increase of the standard deviation of the estimator can be a little bit higher than the one given by the selected ε. When we tried to compute the standard deviation for some more points inside (and also outside) the insensitivity region we realized that the largest increase is realized in the direction orthogonal to the vector $\boldsymbol{\theta_0}$. This is the case of the point $\boldsymbol{\theta_1}$. This is a bit surprising. We expected the largest increase of the standard deviation in case of points more remote from $\boldsymbol{\theta_0}$ - as for example point $\boldsymbol{\theta_2}$, which is on the bound of the insensitivity region and its distance from the point $\boldsymbol{\theta_0}$ is larger than the one of $\boldsymbol{\theta_1}$. (26) shows that the ratio of the appropriate standard deviations does not even reach the (1+ε) boundary. The utilization of the insensitivity regions is described e.g. in [3].

7. References

[1] ANDĚL, J. Statistical Methods. (in Czech), Praha: MATFYZPRESS, 2003. ISBN 80-86732-08-8.

[2] BOHÁČOVÁ, H., HECKENBERGEROVÁ, J. Insensitivity regions for the fixed effects parameters in the mixed linear model with type I constraints and related computational problems. (in Czech) In: Forum Statisticum Slovacum, No. 6, 2007, pp. 31-35. ISSN 1336-7420.

[3] BOHÁČOVÁ, H. 2008. Insensitivity region for variance components in general linear model. In: Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica, Vol. 49, 2008, pp.7-22. ISSN 0231-9721.

[4] KUBÁČEK, L., KUBÁČKOVÁ, L. Statistics and Metrology, Univerzita Palackého v Olomouci– vydavatelství, 2000. ISBN 80-244-0093-6.

[5] RAO, C.R., KLEFFE, J. Estimation of Variance Components and Applications, Amsterdam – New York – Oxford – Tokyo: North-Holland, 1988. ISBN 0-444-70023-4.

Authors addresses:

Mgr. Hana Boháčová
Institute of Mathematics
Faculty of Economics and Administration
Studentská 84
532 10 Pardubice
hana.bohacova@upce.cz

Mgr. Jana Heckenbergerová
Institute of Information Technologies
Faculty of Electrical Engineering and Informatics
Studentská 95, 532 10 Pardubice
jana.heckenbergerova@upce.cz

Inflačné vplyvy cien energií na Slovensku v posledných rokoch
Inflation influences of the energy prices in Slovakia in last years

Mikuláš Cár

Abstract: Energy prices for producers have risen in Slovakia recently on the year-on-year basis much more than energy prices for households.

Key words: price of oil, energy prices, energy prices for producers, energy prices for households, inflation

Kľúčové slová: cena ropy**,** ceny energií, ceny energií pre výrobcov, ceny energií pre domácností, inflácia.

1. Úvod

Priemerná cena ropy Brent prekonala v júli 2007 (mesačný priemer 76,3 USD/barel) jej maximum z augusta 2006 (74,5 USD/barel) a po prakticky permanentnom raste dosiahla v júli 2008 nový priemerný mesačný vrchol (137,2 USD/barel).

Zvyšovanie energetickej závislosti väčšiny štátov[1] a turbulencie cien ropy na svetových trhoch si vyžiadali zvýšenú pozornosť nielen energetikov, výrobcov, domácností ako malospotrebiteľov rôznych druhov energií, politikov, ale aj analytikov a centrálnych bánk v súvislosti so sledovaním cenovej stability.

Vláda SR prerokovala materiál, ktorého cieľom je zabezpečiť realizáciu takých krokov, ktoré umožnia dosiahnuť konkurencieschopnú energetiku, zabezpečujúcu bezpečnú, spoľahlivú a efektívnu dodávku všetkých foriem energie za prijateľné ceny s prihliadnutím na ochranu odberateľa, ochranu životného prostredia, trvalo udržateľný rozvoj, bezpečnosť zásobovania a technickú bezpečnosť.[2]

2. Vývoj cien energií pre výrobcov

Skôr, ako začneme analyzovať ceny energií pre výrobcov, je potrebne uviesť, že predstavujú jednu zo základných súčasti úhrnných cien priemyselných výrobcov. Miesto cien energií pre priemyselných výrobcov ako aj ich aktuálny medziročný vývoj sú zrejme z nasledujúcej tabuľky.

---

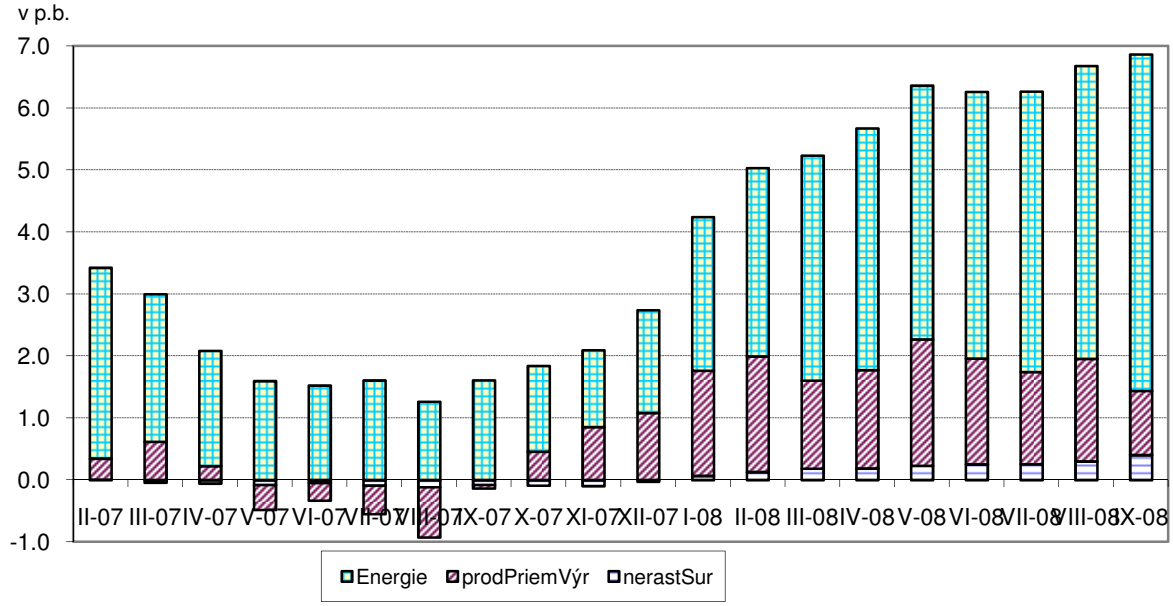
[1] Napr. v roku 2006 dosiahla miera energetickej závislosti Slovenska 64%, čo je viac ako priemer za krajiny EÚ-27 na úrovni 53,8%. Viac pozri: EU27 energy dependence rate at 54% in 2006. In.: Newsrelease 98/2008, Eurostat.

[2] Pozri materiál: *Návrh stratégie energetickej bezpečnosti SR*, prerokovaný na zasadnutí Vlády SR dňa 15.10.2008.

Tabuľka 1: Váhy a medziročný vývoj zložiek cien priemyselných výrobcov

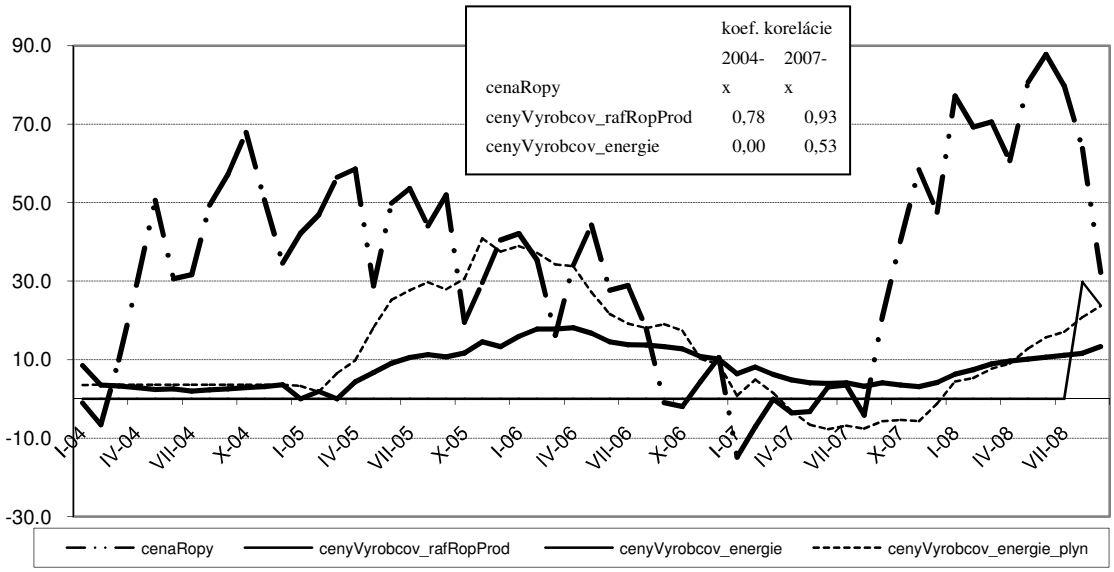
Zložky ceny priemyselných výrobcov		**Váhy**	**IX. 08**
PRIEMYSEL SPOLU		1000,00	106,8
Produkty banskej a povrchovej ťažby		18,21	122,6
Produkty priemyselnej výroby		596,48	101,8
Koks, rafinérske ropné produkty a jadrové palivá		*33,72*	*123,8*
Elektrická energia, plyn, para a teplá voda		**385,31**	113,3
E 40	*Elektrická energia, plyn, para a teplá voda*		*113,4*
E 401	*Výroba a distribúcia elektriny*		*109,5*
E 402	*Výroba plynu, distribúcia plynných palív potrubím*		*123,7*
E 403	*Dodávka pary a teplej vody*		*109,3*
E 41	*Úprava a distribúcia vody*		*108,5*

Opodstatnenosť zvýšenej pozornosti cenám energií pre priemyselných výrobcov je daná nielen tým, že slovenský priemysel je považovaný za značne energeticky náročný, ale aj pomerne výrazne sa zvyšujúcim príspevkom cien energií pri vývoji úhrnných cien priemyselných výrobcov v poslednom období.



Graf 1: Vývoj príspevkov hlavných zložiek k medziročným zmenám PPI_tuzemsko

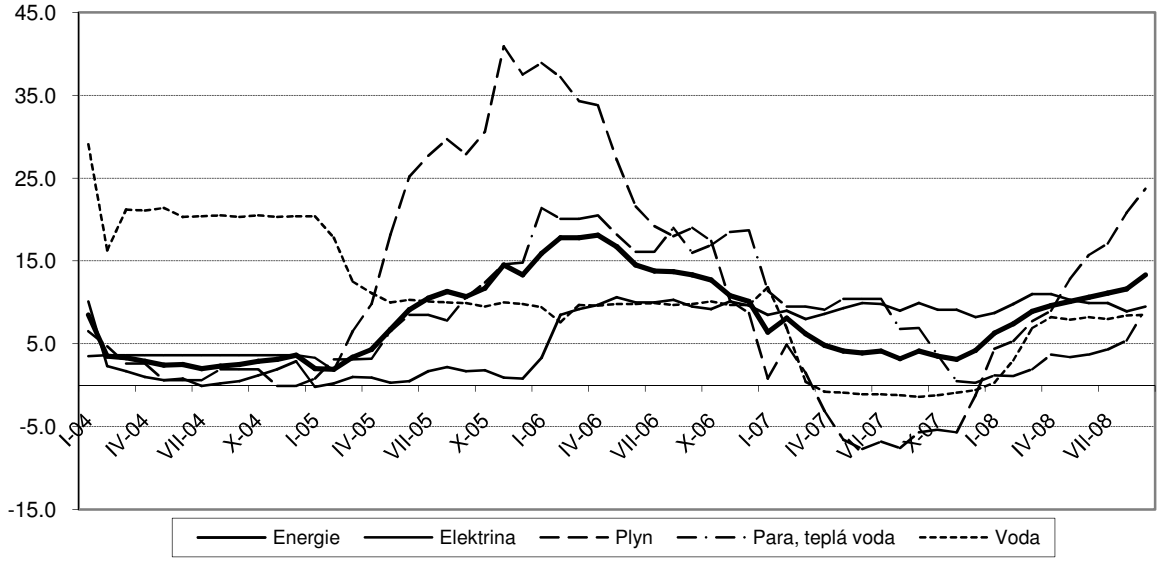
Ceny energií sa už aj v priebehu roku 2007 najvýraznejšie podieľali na medziročných zmenách cien priemyselných výrobcov pre tuzemsko. Od začiatku roka 2008 je však evidentný rastúci vplyv cien energií na vývoj úhrnných cien priemyselných výrobcov. Kým v januári 2008 predstavoval ich príspevok na zmene úhrnnej ceny priemyselných výrobcov takmer 60%, tak napr. v septembri 2008 predstavoval tento podiel takmer 80%. Tento rastúci podiel bol sprevádzaný poklesom podielu príspevku cien produktov priemyselnej výroby (zo 40% na 15%). Ceny nerastných surovín aj vzhľadom na ich zanedbateľnú váhu prispievajú len nepatrne ku zmene úhrnnej ceny priemyselných výrobcov, ale od začiatku roku 2008 majú jednoznačne rastúci trend.



Graf 2: Vývoj cien ropy, rafinérskych ropných produktov a energií (medziročná zmena v %)

Ceny energií sú v podstatnej miere závislé na vývoji cien ropy na svetových trhoch. Od začiatku roku 2007 najvýraznejšie s vývojom cien ropy koreloval v podmienkach Slovenska vývoj cien rafinérskych ropných produktov v rámci produktov priemyselnej výroby (koeficient korelácie r= 0,93). Menej výrazný bol vzťah medzi vývojom ceny ropy a ceny plynu v rámci skupiny energie (r = 0,60). Podobné, ale menej tesné väzby medzi vývojom ceny ropy a cien rafinérskych ropných produktov a plynu pre priemyselných výrobcov boli v dlhodobejšom horizonte od roku 2004 (r = 0,78 a r = 0,18).

Ceny energií ako celku sa od začiatku roku 2007 menili na medziročnej báze pomerne v malej miere a s priemernou cenou ropy korelovali menej výrazne (r = 0,53) ako ceny plynu. Ešte nižšiu koreláciu zaznamenali voči cenám ropy ceny elektriny (r = 0,44) a ceny úpravy a distribúcie vody (r = 0,34). Ceny dodávok pary a teplej vody boli k cenám ropy v hodnotenom období vo výrazne zápornej korelácii (r = -0,84)



Graf 3: Vývoj cien energií a ich zložiek v rámci PPI (medziročná zmena v %)

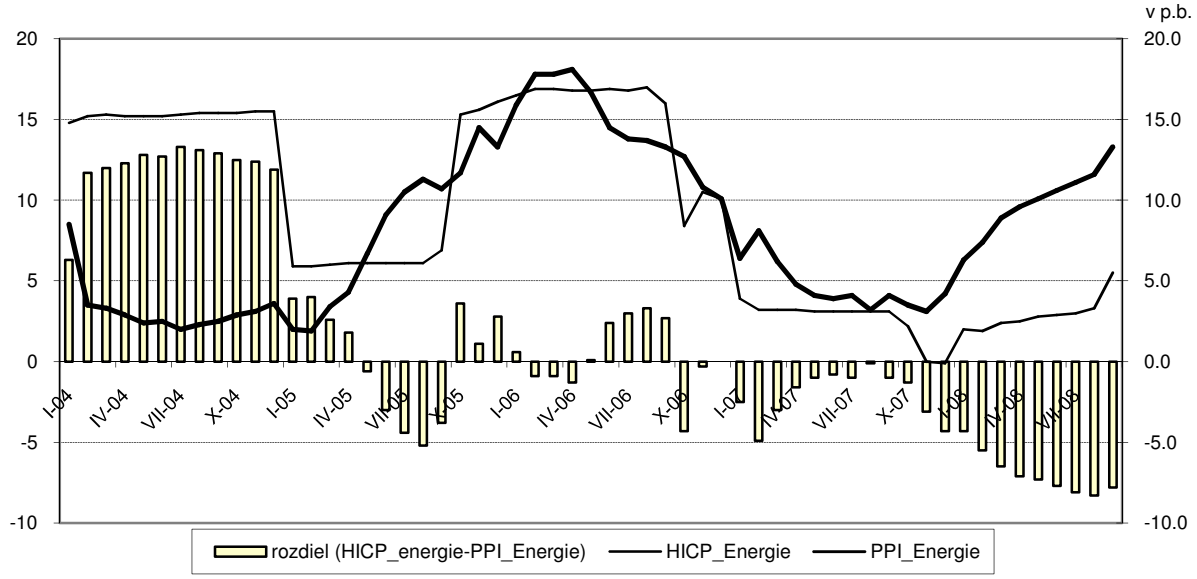
Z dlhodobejšieho pohľadu sú zmeny úhrnných cien energií v priemysle ovplyvňované hlavne vývojom cien plynu. Vývoj cien elektriny bol v posledných rokoch voči vývoju úhrnných cien energií v relatívne neutrálnom vzťahu. Ceny dodávok pary a teplej vody a ceny úpravy a distribúcie vody sa vyvíjali pomerne diferencovane. Ceny úpravy a distribúcie vody sa až do konca roku 2006 vyvíjali značne rozdielne ako úhrnné ceny energií, ale od začiatku roku 2007 pomerne výrazne korelujú. Celkom opačný vývoj je charakteristický pre vzťah medzi vývojom cien úpravy a distribúcie vody a úhrnnými cenami energií.

Od cien priemyselných výrobcov sa do značnej miery odvíjajú aj spotrebiteľské ceny. V prípade cien energií možno sledovať niekoľko rozdielnosti medzi medziročnými zmenami cien energií v rámci úhrnných cien priemyselných výrobcov a v rámci spotrebiteľských cien.

3. Vývoj cien energií v rámci spotrebiteľských cien

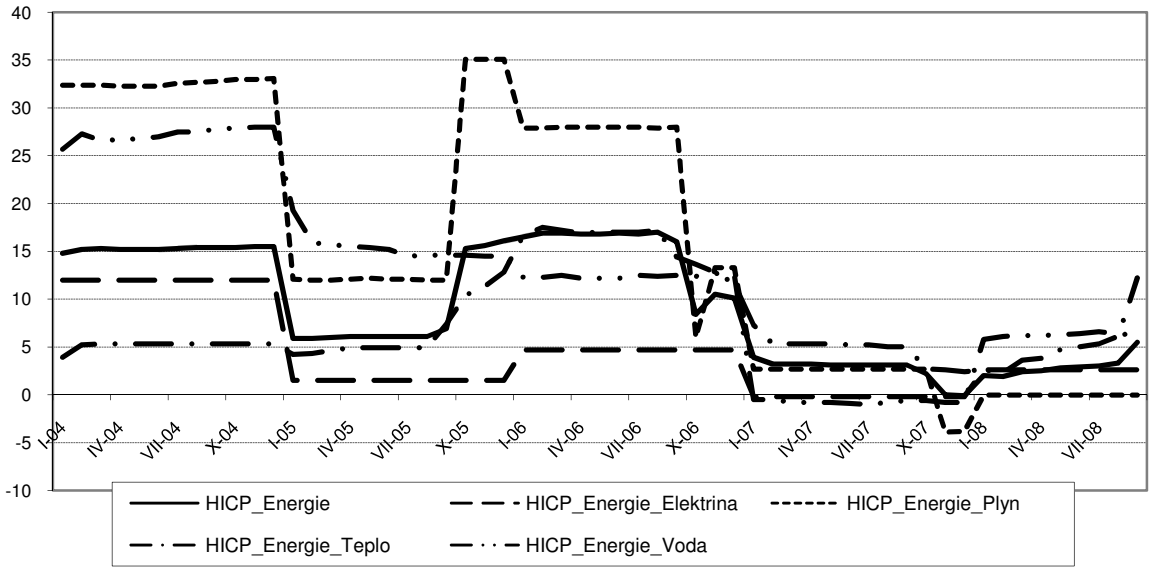
V rámci spotrebného koša postupne rastie váha výdavkov domácností spojených s bývaním a energiami. Energie majú v spotrebnom koši pre domácností relatívne nižšiu váhu (13,24%), ako energie v rámci úhrnných cien priemyselných výrobcov (38,53%).

V jednotlivých mesiacoch od začiatku roku 2004 až do polovice roku 2005 zaznamenali ceny energií pre priemyselných výrobcov značne odlišné medziročné zmeny ako ceny energií pre domácností, keď spotrebiteľské ceny energií rástli omnoho výraznejšie ako ceny energií v priemysle. Súviselo to aj s rozdielnymi metodikami stanovovania cien pre výrobcov a malospotrebiteľov. Od konca roku 2005 až do 4. štvrťroku 2007 bol vývoj cien energii pre výrobcov aj obyvateľstvo veľmi podobný. Od začiatku roku 2008 dochádzalo postupne k roztváraniu nožníc, keď bola dynamika rastu cien pre obyvateľstvo výrazne nižšia ako pre výrobcov.



Graf 4: Vývoj cien energií v rámci PPI a HICP (medziročná zmena v %)

Medziročné zmeny cien energií pre domácností a ich jednotlivých zložiek mali v posledných rokoch zaujímavý priebeh. Z dlhodobého hľadiska bol vývoj cien energií pre domácností dosť výrazne ovplyvňovaný vývojom cien všetkých základných zložiek, ale jednoznačne hlavne vývojom cien plynu. Zásadný zlom v medziročných zmenách cien energií pre domácností je evidentný od začiatku roku 2007, keď s vývojom súhrnných cien energií najvýraznejšie koreluje vývoj cien tepla (r = 0,86). Stále výrazný vplyv, ale miernejší než v dlhodobejšom horizonte má od začiatku roku 2007 vývoj cien plynu (r = 0,65), kým vplyv vývoja cien elektriny a vodného a stočného výrazne poklesol (r = 0,14 a 0,15).



Graf 4: Vývoj cien energií v rámci HICP (medziročná zmena v %)

Výraznejšie rozdiely vo vývoji cien energií pre domácností a pre výrobcov v priebehu roku 2004 boli spôsobené do značnej miery aj rozdielnou metodikou stanovovania týchto cien. Rastúce rozdiely medzi vývojom oboch skupín súhrnných cien energií najmä od začiatku roku 2008 sú spôsobované aj ochranou politikou najmä domácností zo strany štátu.

Pri detailnejšom porovnaní medziročných zmien cien jednotlivých energetických médií pre domácností a výrobcov zisťujeme, že od začiatku roku 2006 je najvýraznejší rozdiel v neprospech výrobcov pri medziročných zmenách cien elektriny a od začiatku roku 2008 rastú aj rozdiely medziročných zmien cien plynu v neprospech výrobcov. Rozdiely medzi medziročným rastom cien dodávok a úpravy vody a tepla nezaznamenali v posledných rokoch medzi domácnosťami a výrobcami výraznejšie hodnoty.

4. Záver

Vývoj ceny ropy na medziročnej báze koreloval s vývojom cien energií pre priemyselných výrobcov od začiatku roku 2007 na Slovensku podstatne výraznejšie (r = 0,53) a dokonca protichodne ako s vývojom cien energií v rámci spotrebiteľských cien (r = -0,37). Znamená to, že ceny ropy výraznejšie ovplyvňujú vývoj cien priemyselných výrobcov ako vývoj spotrebiteľských cien.

Určujúcou zložkou pre vývoj cien energií je cena plynu, pričom výraznejšie ovplyvňuje vývoj cien energií pre priemyselných výrobcov ako pre domácností. Ceny tepla vo väčšej miere ovplyvňujú vývoj úhrnnej ceny energií pre domácností ako vývoj úhrnnej ceny energií pre priemyselných výrobcov. Ceny elektriny sú vo vzťahu k úhrnným cenám energií pre výrobcov relatívne neutrálne a v prípade domácností sa ich vplyv na úhrnné ceny energií hlavne od roku 2007 postupne znižuje.

Úhrnné ceny energií pre priemyselných výrobcov sa len v určitých obdobiach vyvíjali na medziročnej báze podobne ako úhrnné ceny energií pre domácností. Od záveru roku 2007 dochádza postupne k evidentnému zväčšovaniu sa ich rozdielov hlavne v dôsledku výrazne nižšej dynamiky rastu cien energií pre domácností ako pre priemyselných výrobcov.

Používanie rozdielnych kritérií pri určovaní cien energií pre výrobcov a domácností môže mať aj deformačné účinky na tvorbu cien energetických komodít. Je potrebné nájsť

transparentný a jednoduchý mechanizmus, ktorý zohľadní tak oprávnené záujmy výrobcov a dodávateľov energií, ako aj primeranú ochranu spotrebiteľov energií.

Dosiahnuť takú liberalizáciu sektora energetiky, ktorá by bola schopná zabezpečiť spoľahlivé a efektívne dodávanie všetkých foriem energií za prijateľné ceny konečným spotrebiteľom aj s prihliadnutím na objektívne záujmy výrobcov a dodávateľov energií, je stále jednou z prioritných úloh nielen centrálnych inštitúcií EÚ, ale aj jej jednotlivých členských krajín.

Literatúra:

[1] Cár, M. 2007: Energetika a otázky vývoja cien energií na Slovensku v posledných rokoch. In.: Biatec č. 6, 2007, s. 2 - 7.

[2] EU27 energy dependence rate at 54% in 2006. In.: Newsrelease 98, 2008, Eurostat.

[3] EUROSTAT Website/Home page/Environment and energy/Data.

Adresa autora:

Mikuláš Cár, Ing., PhD.
Národná banka Slovenska
Imricha Karvaša 1
813 25 Bratislava
Mikulas.Car@nbs.sk

Introduction to the Discrete GNSS Position Integrity Monitoring Algorithms
Úvod do diskrétních GNSS-PIM algoritmů

Jana Heckenbergerová, Hana Boháčová

Abstrakt: V současnosti existuje řada metod pro určení polohy vlaku na trati. Výzkumy v oblasti Globálních Navigačních Satelitních Systémů (GNSS) ukazují, že je výhodné využívat satelitní navigaci i pro tento účel. Popisované algoritmy monitorují integritu GNSS polohy. Cílem algoritmů je ověřit, zda GNSS poloha koresponduje s danou předpokládanou tratí vlaku, tedy zjistit, zda získaná GNSS poloha splňuje požadavky na bezpečnost – Safety Integrity Level (SIL) pro danou pravděpodobnost nedetekované chyby. Trať vlaku je definována diskrétně pomocí časově ekvidistantní množiny bodů. Příspěvek je věnován seznámení s diskrétními GNSS-PIM algoritmy, které jsou založeny na mnohorozměrné statistické analýze.

Klíčová slova: GNSS, monitoring integrity, mnohorozměrná statistická analýza, Wishartova matice, Hottelingova t-kvadrát testovací statistika, prahová oblast, protection level (PL)

Key words: GNSS, Integrity Monitoring, Multivariate Statistical Analysis, Wishart matrix, Hotteling's test statistics, Threshold Domain, Protection Level (PL)

Introduction

Global Navigation Satellite Systems

The name Global Navigation Satellite System (GNSS) covers all existing satellite systems (e.g. GPS-USA, GLONASS-Russia, GALILEO-EU). GNSS receivers of satellite signal provide information about position and its precision to their users. Some GNSS receivers are able to collect signals from different satellite systems. More information about operation GNSS can be found in (Mervart, 1993), where algorithms for (x,y)-position of GNSS receiver determination are also described. GNSS have been successfully used in the automotive field (GPS navigation), geodesy, and modern telecommunication technology. In air transport, elaborate methods for safety position determination by the help of GNSS have been developed and used.

GNSS in railway transport

Czech Railways perform tests and experiments with this new technology. GNSS receivers are installed on several selected locomotives and obtained GNSS data is compared with the map of train tracks. Trains can move only on tracks, so they can move only on the route, which has been strictly defined. This is the greatest advantage of GNSS usage for railway transport. For safety determination of locomotive position, it is enough to recognize in which track the reference point of the locomotive is situated. Then the determination of train position on this track is quite easy. Reasons for usage of this modern technology are evident – to reduce running costs and to increase effectiveness, safety and capacity of train operation.

Nowadays GNSS database system for monitoring and prediction of locomotive position is also developed. Locomotive equipped with GNSS receiver send own positions every 5sec to central computer. All GNSS positions (with other information like train identification

number, conductor number and track number) are saved to GNSS database. By the help of database it could by possible to describe (analytically or discretely) all used tracks, view driven trace of some train, safety verify current position of given locomotive. Algorithms which worked above GNSS database are named as GNSS Simplification (GNSS-SIM) algorithms.

GNSS-PIM algorithms

The GNSS Position Integrity Monitoring (GNSS-PIM) algorithms monitor the position integrity of the GNSS receiver. Let GNSS position of locomotive reference point be given by (x,y)-coordinates. The aim of these algorithms is to verify that GNSS position corresponds to the given train track and to find out that obtained GNSS position complies with safety requirements – Safety Integrity Level (SIL) – for given probability of missed detection (P_{MD}). For every GNSS position the Protection Level (PL) is computed. In statistical point of view the PL is the threshold domain, which is limited by probability of the second kind error. GNSS position integrity monitoring is provided by comparing of obtained PL with Horizontal Alert Limit (HAL), which is for given SIL already defined by user. If PL value is same or lower than HAL, then GNSS position has enough integrity and it could be regarded as safe. If PL is greater than HAL, the GNSS position doesn't satisfy safety requirements and mustn't be used for following safety related applications.

For correct operation of GNSS-PIM algorithms it is necessary to ensure sufficient integrity of the GNSS signal. The GNSS signal integrity can be obtained by Ground Integrity Channel (GIC) report and by usage of Receiver Autonomous Integrity Monitoring (RAIM). Another necessary demand for correct operation of monitoring algorithms is the usage of safety track map. For this reason the track map must satisfy safety requirements and it is necessary to ensure sufficient integrity of this map. If the map contains position errors and hasn't demanded precision, it mustn't be used for safety related applications.

The GNSS-PIM algorithms could be divided in analytical, parametrical and discrete according to the train track definition. Another division of GNSS-PIM algorithms could be done by a character of the track. Tracks specification for railway traffic is linear, arch and cubic. Algorithms development begins from linear character of the track, arch and cubic algorithms are explored after. They are solved by linearization of gain statistical models.

Analytical algorithms are used when the train track is defined by analytical function $f(x, y) = 0$. Incepted problems could be described by statistical model with constraints. Analytical GNSS-PIM algorithms for linear and arch track are solved in M.S. theses (Dvorakova, 2004).

Parametrical algorithms are used when the train track is defined by parametrical equations $\begin{cases} x = \varphi(t) \\ y = \psi(t) \end{cases}$ with parameter t. Statistical description of these algorithms appears in (Heckenbergerova, 2007) and (Heckenbergerova, 2008).

If the train track is defined by discrete timely equidistant set of position coordinates $Y = \left\{ \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} \right\}_{i=1}^{n}$, then GNSS-PIM algorithms are called discrete. **Discrete algorithms** have great advantage against others, because knowledge of track character (line, arch, cubic) isn't demanded. Discrete problems could be described by multivariate statistical model. This is the main disadvantage. Features of multivariate models like confidence area and insensitivity region hasn't been described. Incepted problems in discrete GNSS-PIM algorithms are main aim of current research. Some basic discrete algorithms are described in conference proceedings (Dvorakova et al., 2005).

Discrete GNSS-PIM algorithms

Theoretical basement for discrete problem

Definition 1:

Let $X_1 \sim N_p(\mu_1,\Sigma),...,X_n \sim N_p(\mu_n,\Sigma)$ be independent stochastic vectors, μ_i be p-dimensional columned vector for $\forall i = 1,...,n$ and Σ be matrix of type $(p \times p)$.

Let $\mathbf{X}_{n\times p} = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}$ and $\mathbf{M}_{n\times p} = \begin{pmatrix} \mu_1^T \\ \vdots \\ \mu_n^T \end{pmatrix}$.

Conjugate distribution of the elements of matrix $\mathbf{Y} = \mathbf{X}^T\mathbf{X}$ is called **p-dimensional Wishart distribution with *n* degrees of freedom** and with parameters Σ, $\mathbf{M}$:

$\mathbf{Y} \sim W_p(n,\Sigma)$ – central Wishart distribution ($\mathbf{M} = \mathbf{0}$),

$\mathbf{Y} \sim W_p(n,\Sigma,\mathbf{M})$ – noncentral Wishart distribution ($\mathbf{M} \neq \mathbf{0}$).

Definition 2:

Let $\mathbf{y} \sim N_p(\mu, {}^1\!/_c\,\Sigma)$, $\mathbf{Y} \sim W_p(k,\Sigma)$ be independent and Σ be positive definite matrix.

Test statistics $T^2 = c \cdot k \cdot \mathbf{y}^T\mathbf{Y}^{-1}\mathbf{y}$

is called **Hotelling t-square test statistics**.

Theorem 1:

If $k > p\text{-}1$ then $F = \dfrac{k-p+1}{p} \cdot \dfrac{T^2}{k} \sim F_{p,k-p+1}(\delta)$,

where $\delta = c \cdot \mu^T\Sigma^{-1}\mu$ is noncentrality parameter of the *F*-distribution.

Formulation of discrete problem

The GNSS Position Integrity Monitoring (GNSS-PIM) algorithms monitor the position integrity of the GNSS receiver. Let $\mathbf{Z}$ be vector of GNSS position of locomotive reference point be given by (*x*,*y*)-coordinates. The aim of Discrete GNSS-PIM algorithms is to verify that GNSS position corresponds to the given train track defined by polygon $\mathbf{Y}$. And then find out that obtained GNSS position complies with safety requirements – Safety Integrity Level (SIL) – for given probability of missed detection (P_{MD}). For every GNSS position the Protection Level (PL) is computed. GNSS position integrity monitoring is provided by comparing of obtained PL with Horizontal Alert Limit (HAL). If PL value is same or lower than HAL, then GNSS position has enough integrity and it could be regarded as safe. If PL is greater than HAL, the GNSS position doesn't satisfy safety requirements and mustn't be used for following safety related applications.

Let $\mathbf{Y} = \left\{ \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} \right\}_{i=1}^{n}$ be vector of the (x, y)-coordinates vector track axis,

$\mathbf{Z} = \left\{ \begin{pmatrix} Z_{j,1} \\ Z_{j,2} \end{pmatrix} \right\}_{j=1}^{m}$ be vector of the GNSS train positions,

$\tilde{\mathbf{Z}} = \left\{ \begin{pmatrix} \tilde{Z}_{j,1} \\ \tilde{Z}_{j,2} \end{pmatrix} \right\}_{j=1}^{m}$ be vector of the orthogonal projections $\mathbf{Z}$ to $\mathbf{Y}$ based polygon and

$\Delta = \left\{ \begin{pmatrix} \Delta_{j,1} \\ \Delta_{j,2} \end{pmatrix} \right\}_{j=1}^{m}$ be vector of the projection residual vectors $\Delta_j = \begin{pmatrix} \Delta_{j,1} \\ \Delta_{j,2} \end{pmatrix} = \begin{pmatrix} Z_{j,1} - \tilde{Z}_{j,1} \\ Z_{j,2} - \tilde{Z}_{j,2} \end{pmatrix}$.

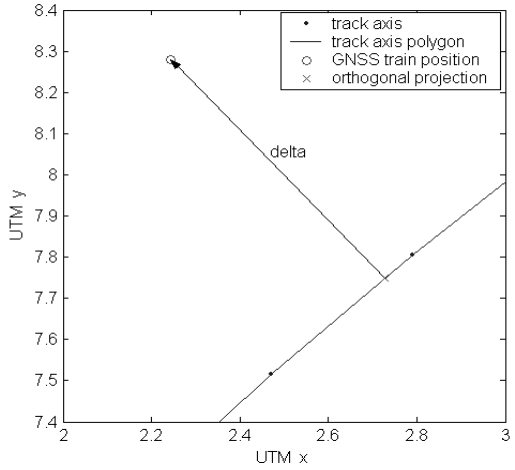


Fig 1. *Graphical representation of described vectors*

First aim of the lower described algorithm is testing hypotheses, whether true positions of given locomotive lie on the supposed track given by discrete timely equidistant set of position coordinates **Y**.
So null hypotheses:

$$H_0 : E(\Delta_j) = 0$$

is tested against alternative hypotheses:

$$H_a : E(\Delta_j) \neq 0$$

for every *j=1,…m* so for every measured GNSS train position.

Test statistics properties are introduced in following theorem. This theorem can be proved by Theorem 1 and Definitions 1, 2 using.

Theorem 2: Test of fit

Let Δ_j for *j=1,… ,m* be a sequence of the projection residuals vectors and $W_{2,m} = \sum_{j=1}^{m} \Delta_j \Delta_j^T$ be Wishart matrix type (2x2). Let Δ_{m+1} be the projection residual vector correspondent with (*m+1*) GNSS position, then test statistics

$$T = \Delta_{m+1}^T W_{2,m}^{-1} \Delta_{m+1} \sim \frac{\chi_2^2}{\chi_{m-2+1}^2} = \frac{2}{m-1} F_{2,m-1} .$$

It is obvious that minimal three GNSS positions are essential for algorithm first using. Let first three projection residual vectors Δ_1, Δ_2, Δ_3 be evaluate and test significance level α (probability of the first order error, false alarm probability) be select. Now, perform test of fit for third GNSS train position $\begin{pmatrix} Z_{3,1} \\ Z_{3,2} \end{pmatrix}$, where the test statistics $T = \Delta_3^T W_{2,2}^{-1} \Delta_3 \sim \frac{1}{2} F_{2,1}$.

If $T \leq 2 \cdot F_{2,1}(1-\alpha)$, hypothesis H_0 can not be refused on the given test significance level α, if $T > 2 \cdot F_{2,1}(1-\alpha)$, hypothesis H_0 is refused in behalf of alternative hypothesis H_a on the given test significance level α. In case of hypothesis H_0 refusal, Δ_3 is significantly differ from other projection residual vectors, so it's necessary to generated error report and delete given GNSS

position from safety related applications.. Then algorithm restart must be carry out. GNSS position seems to be right, when hypothesis H_0 isn't refused. Then PL for this measurement can be determinate and algorithm can pass on the next iteration. If obtained PL value is same or less than HAL value, which is previously defined by user, then the GNSS position has enough integrity and it can be regard as safety. If PL is greater than HAL, the GNSS position doesn't satisfy safety requirements.

Protection Level is the threshold domain, which is limited by the probability of the second kind error (probability of undetected failure, P_{MD}). The threshold domain determinate how big difference between projection residual vector and mean-value of these vectors can be detected with given probability of undetected failure, P_{MD} value is defined by safety requirements (SIL).

Let Ω be the set of all plane vectors and let $\delta_{\max}$ be solution of equation

$$P\{F_{2,m-1}(\delta) \geq F_{2,m-1}(0, 1-\alpha)\} = 1 - P_{MD},$$

then

$$\mathrm{PL} = \mathrm{T}_{\kappa_\alpha} = \left\{u : u \in \Omega,\ \tfrac{m-1}{2} \cdot u^T W_{2,m}^{-1} u \leq \delta_{\max}\right\} \quad .$$

Discrete GNSS-PIM algorithm

Now the algorithm m-iteration description can be introduced:

- Evaluation of Wishart matrix $W_{2,m}$

  $$W_{2,m} = \sum_{j=1}^{m} \Delta_j \Delta_j^T$$

- Determination of the projection residual vector Δ_{m+1}
- Evaluation of the test criteria value T

  $$T = \Delta_{m+1}^T W_{2,m}^{-1} \Delta_{m+1}$$

- Comparing of values $\frac{m-1}{2} \cdot T$ and $F_{2,m-1}(1-\alpha)$
  - $\frac{m-1}{2} \cdot T \leq F_{2,m-1}(1-\alpha)$ – PL determination and passing on the next iteration $m+1$
  - $\frac{m-1}{2} \cdot T > F_{2,m-1}(1-\alpha)$ – generation of error report and algorithm restart.

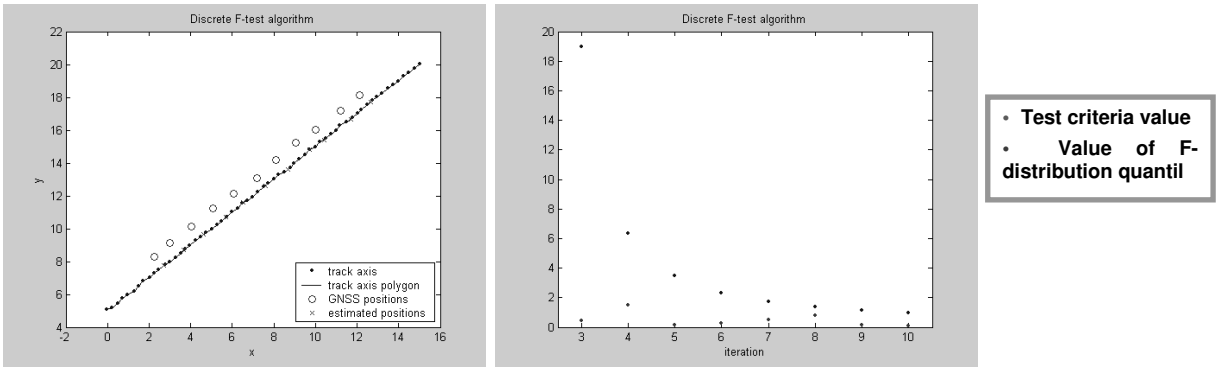


***Fig 2**. Numerical results of Discrete GNSS-PIM algorithm*

Discrete GNSS-PIM algorithm can be negatively influenced by dependence of GNSS positions and their orthogonal projection, Wishart matrix become singular and can not be inverted. Because of singularity, it is necessary to decrease degree of freedom of F-

distribution and generalized inversion must be used for the test criteria. This problem increases time costingness of the whole algorithm.

Conclusion

During the realization of the mentioned research many difficult mathematical problems can be expected. For example up to now normal distribution of GNSS position errors is assumed and due to this assumption estimation procedures and testing of statistical hypotheses have been used in standard way. Usually variances of the actual errors in determination of train position are unknown and must be estimated. This fact influences heavily the procedure of confidence domain determination and power functions of statistical tests. Utilization of other measure techniques e.g. odometers, gyroscopes and Doppler radar, make the set of results heterogeneous, what need development of special numerical algorithms. Only satisfactory solution of mentioned problems has chance to utilize in the practice successfully and this way to increase effectiveness, safety and capacity of train operation.

Aim of future research is integration PIM algorithms into existing software for train position determination, laboratory and online tests of whole positioning system. Confidence area, threshold domains and insensitivity regions in discrete PIM algorithms are essential part of future statistical problem. PIM algorithms could be used for car position verification too. Evolution of automobile safety position monitoring is also direction of future development.

References

KUBÁČEK, L., KUBÁČKOVÁ, L. *Statistika a metrologie*. Univerzita Palackého v Olomouci 2000, ISBN 80-244-0093-6.

RAO, R.C. : *Lineární metody statistické indukce a jejich aplikace.* Academia Praha (1978)

MERVART, L. *Základy GPS*. Vydavatelství ČVUT, Praha, 1993.

DVOŘÁKOVÁ, J. *Statistické metody pro identifikaci polohy vlaku.* Olomouc 2004.

DVOŘÁKOVÁ, J., MOCEK, H., MAIXNER, V., *Statistical Approach to the Train Position Integrity Monitoring*. Sborník „Reliability, safety and diagnostics of transport structures and means 2005“, Pardubice, ISBN 80-7194-769-5.

HECKENBERGEROVÁ, J. *Parametrické algoritmy pro ověření GNSS polohy vlaku.* Sborník Infotrans 2007, Pardubice, ISBN 978-80-7194-989-3.

HECKENBERGEROVA, J.: *GNSS train position verification by the help of parametrical PIM algorithms.* 7th International Conference „Aplimat 2008“, Bratislava 2008, ISBN: 978-80-89313-03-7.

Current address:

Heckenbergerová Jana, Mgr.
University of Pardubice,
Faculty of Electrical Engineering and Informatics,
Department of Information Technologies,
Studentská 95, 532 10 Pardubice
phone: +420 466 036 726
e-mail: jana.heckenbergerova@upce.cz

Boháčová Hana, Mgr.
University of Pardubice,
Faculty of Economics and Administration
Institute of Mathematics,
Studentská 84, 532 10 Pardubice
phone: +420 466 036 170
e-mail: hana.bohacova@upce.cz

Analýza BMI a hmotnosti dieťaťa a matky v súbore abruptio placentae praecox
Analysis values of BMI and weight of children and mother in file patient with indication abruptio placentae praecox

Jozef Chajdiak, Maroš Suchánek

Abstract: Paper includes statistical analysis values of BMI and weight of children and mother in file patient with indication abruptio placentae praecox.

Key words: abruptio placentae praecox, correlation matrix.

Kľúčové slová: abruptio placentae praecox, korelačná matica.

1. Úvod

Analyzuje sa súbor pacientok s indikáciou abruptio placentae praecox vo verzii 2-11, ktorá obsahuje súbor údajov pri vitalite plodu rovnej nula (dieťa zomrelo) a vo verzii 12-106, ktorá obsahuje súbor údajov pri vitalite plodu rovnej jedna (dieťa prežilo). Vo verzii 2-110 sú údaje z referenčného súboru rodičiek bez indikácie abruptio placentae praecox V analýze sú použité údaje o výške (DLZKA) a hmotnosti (HMd) dieťaťa, výške (VYSKA) matky a hmotnosti matky pri (HMpor) a po (HMnor) pôrode, veku matky (VEK) a týždni pôrodu (TYZDEN). Z údajov o výškach a hmotnostiach sú vypočítané hodnoty BMI dieťaťa (BMId), BMI matky pri (BMIpor) a po (BMInor) pôrode. Realizoval sa základný štatistický rozbor (rozsah, min, max, priemer, mesián a std) jednotlivých premenných a analýza korelačnej matice jednotlivých premenných.

2. Základný štatistický rozbor

Základný štatistický rozbor jednotlivých premenných (rozsah, min, max, priemer, medián, smerodajná odchýlka) sme realizovali pomocou funkcií COUNT, MIN, MAX, AVERAGE, MEDIAN a STDEV v systéme Excel. Výsledky sú nasledujúce:

PLACE, vitalita=0

2-11	vek	týždeň	VYSKA	HMnor	HMpor	kg+	BMInor	BMIpor	HMd	DLZKA	BMId
n	10	10	10	10	10	10	10	10	10	10	10
min	25	26	147	47	62	9	20,5	23,9	650	29	7,73
max	36	39	171	70	83	16	24,8	29,4	3040	51	12,49
priemer	31,5	33	159,1	56,3	68,1	11,8	22,2	26,9	1865	41,2	10,30
median	33	33	157,5	54,5	65,5	10	21,8	26,6	1955	41,5	10,43
std	3,78	3,92	8,41	6,41	7,23	2,66	1,22	1,64	806,63	7,16	1,76

PLACE,vitalita=1

12-106	*vek*	*týždeň*	*VYSKA*	*HMnor*	*HMpor*	*kg+*	*BMInor*	*BMIpor*	*HMd*	*DLZKA*	*BMId*
n	96	96	96	96	96	96	96	96	96	96	96
min	19	27	153	46	53	0	16,96	18,78	730	30	7,81
max	56	42	181	95	97	21	31,20	34,78	4950	56	16,36
priemer	30,73	36,63	165,90	59,81	71,61	11,9	21,71	26,00	2695	46,96	11,87
median	30	37	165,5	58	71	11,5	21,37	26,17	2680	48	11,74
std	5,84	3,47	6,18	8,61	9,30	4,34	2,68	2,84	798,0	4,75	1,79

REFER, vitalita=1

2-110	vek	týždeň	VYSKA	HMnor	HMpor	kg+	BMInor	BMIpor	HMd	DLZKA	BMId
n	110	110	110	110	110	110	110	110	110	110	110
min	19	37	154	48	56	0	16,67	19,38	2550	37	11,00
max	57	42	180	110	124	35	37,18	41,91	4390	56	20,38
priemer	33,87	39,71	167,76	62,52	77,03	14,47	22,22	27,39	3428	50,48	13,44
median	33	40	168	60	75	14	21,36	27,07	3430	51	13,43
std	8,02	1,03	5,40	10,52	11,80	5,54	3,59	4,06	424,89	2,03	1,38

Z vypočítaných štatistík môžeme konštatovať, že dĺžka tehotenstva je u matiek z referenčného súboru dlhšia ako u pacientok s dieťaťom, ktoré prežilo aj ktoré neprežilo pôrod. Podobná situácia je u výšky ženy, normálnej hmotnosti ženy, hmotnosti pri pôrode, prírastku hmotnosti počas tehotenstva. BMI žien z referenčného súboru je vyššie ako u žien pacientok ale pacientky s vitalitou plodu 0 (úmrtie dieťaťa) majú vyššie BMI pri a aj po pôrode ako pacientky s vitalitou plodu1. Miery u plodu sú najvyššie v referenčnom súbore matiek, potom u pacientok s vitalitou plodu 1 a najhoršia je situácia u pacientok s vitalitou plodu 0.

3. Korelačná matica

Súbor PLACE obsahuje údaje o 106 a súbor REFER obsahuje údaje o 110 rodičkách. K výpočtu korelačných matíc medzi hodnotami jednotlivých premenných sa použil nástroj Correlation z Excelu. Výsledné korelačné matice sú nasledujúce:

PLACE, vitalita=0

2-11	vek	týždeň	VYSKA	HMnor	HMpor	kg+	BMInor	BMIpor	HMd	DLZKA	BMId
vek	1										
týždeň	0,068	1									
VYSKA	-0,34	-0,779	1								
HMOTnor	-0,37	-0,509	0,8583	1							
HMOTpor	-0,16	-0,628	0,8124	0,9311	1						
kg+	0,454	-0,48	0,1401	0,1212	0,475	1					
BMInor	-0,09	0,4405	-0,144	0,3825	0,3149	-0,07	1				
BMIpor	0,334	0,2585	-0,34	0,0745	0,2706	0,557	0,7247	1			
HMOTd	0,173	0,9431	-0,73	-0,509	-0,574	-0,33	0,3531	0,2613	1		
DLZKA	0,127	0,9786	-0,806	-0,517	-0,58	-0,33	0,4636	0,3763	0,96	1	
BMId	0,238	0,715	-0,555	-0,442	-0,457	-0,18	0,164	0,1613	0,89	0,778	1

Medzi vekom pacientky a jej výškou aj hmotnosťami je nepriama závislosť (záporné koeficienty korelácie). Medzi týždňom pôrodu a výškou a hmotnosťami pacientky je silnejšia nepriama závislosť a naopak mierami plodu (hmotnosť, dĺžka, BMI) silnejšia priama závislosť. Silné závislosti medzi BMI a výšky a hmotnosti, z ktorých je vypočítaný je daný vo veľkej miere vzorcom výpočtu BMI. Vyššie hodnoty koeficientov korelácie (abstrahujúc od znamienka) sú dané nižším rozsahom súboru (n=10).

Medzi vekom pacientky a BMI dieťaťa je síce malá, ale nepriama závislosť (staršie rodičky – nižšie BMI dieťaťa). Týždeň pôrodu a parametre dieťaťa sú vo vysokej priamej závislosti. Medzi BMI matky a BMI dieťaťa je nepriama slabá závislosť.

Koeficienty korelácie pre súbor REFER naznačujú vyššiu mieru závislosti ako pre súbor PLACE s vitalitou 1.

PLACE, vitalita=1

12-106	*vek*	*týždeň*	*VYSKA*	*HMnor*	*HMpor*	*kg+*	*BMInor*	*BMIpor*	*HMd*	*DLZKA*	*BMId*
vek	1,000										
týždeň	0,008	1,000									
VYSKA	0,018	0,116	1,000								
HMOTnor	0,164	-0,016	0,476	1,000							
HMOTpor	0,154	0,142	0,510	0,881	1,000						
kg+	0,041	0,360	0,168	-0,066	0,401	1,000					
BMInor	0,177	-0,090	-0,044	0,855	0,699	-0,177	1,000				
BMIpor	0,167	0,089	-0,071	0,703	0,820	0,355	0,841	1,000			
HMOTd	0,032	0,885	0,044	0,023	0,175	0,339	-0,004	0,174	1,000		
DLZKA	0,045	0,902	0,076	0,044	0,173	0,297	0,003	0,154	0,908	1,000	
BMId	-0,010	0,766	0,020	-0,030	0,125	0,334	-0,054	0,129	0,914	0,697	1,000

REFER, vitalita=1

2-110	*vek*	*týždeň*	*VYSKA*	*HMnor*	*HMpor*	*kg+*	*BMInor*	*BMIpor*	*HMd*	*DLZKA*	*BMId*
vek	1,000										
týždeň	0,068	1,000									
VYSKA	-0,002	-0,004	1,000								
HMnor	0,051	0,001	0,267	1,000							
HMpor	0,005	-0,043	0,240	0,881	1,000						
kg+	-0,115	-0,092	0,006	-0,029	0,445	1,000					
BMInor	0,053	0,004	-0,109	0,927	0,806	-0,051	1,000				
BMIpor	0,009	-0,044	-0,183	0,774	0,909	0,453	0,860	1,000			
HMd	0,081	0,217	0,110	0,121	0,239	0,287	0,074	0,194	1,000		
DLZKA	0,085	0,297	0,088	0,277	0,363	0,253	0,250	0,332	0,630	1,000	
BMId	0,031	-0,011	0,057	-0,081	-0,013	0,133	-0,115	-0,041	0,686	-0,124	1,000

4. Záver

Vypočítané štatistiky potvrdzujú fakt, že pôrodné miery dieťaťa v rozhodujúcej miere ovplyvňuje týždeň pôrodu.

5. Literatúra

[1]CHAJDIAK, J. 2003. Štatistika jednoducho. Bratislava: Statis 2003. 194 s. ISBN 80-85659-28-X.

[2]CHAJDIAK, J. 2003. Štatistické úlohy a ich riešenie v Exceli. Bratislava: Statis 2005. 262 s. ISBN 80-85659-39-5.

Adresa autorov:

Jozef Chajdiak, Doc., Ing., CSc.
Ústav manažmentu STU
810 00 Bratislava
chajdiak@statis.biz

Maroš Suchánek, MUDr.
FNsP Ružinov
920 00 Bratislava
marossuchanek@gmail.com

Odhadnutie parametrov regresného modelu reznej rýchlosti kvapalinového abrazívneho lúča v systéme SAS
Parameter estimation of regression model of waterjet cutting speed in system SAS

Ivan Janiga, Iveta Stankovičová, Marek Kall

Abstract: The contribution deals with the relationship between cutting speed abrasive Waterjet and relevant quantities. On the basis of measured data we want to develop the best model of the relationship. Problem solving method is regression analysis, and selection of variables and model building. The stepwise regression method given in software SAS is used in estimation.

Key words: model building, regression analysis, stepwise regression.

Kľúčové slová: modelovanie, regresná analýza, kroková regresia.

1. Úvod

Na základe nameraných dát získaných z experimentu sme hľadali závislosť reznej rýchlosti abrazívneho kvapalinového lúča R od nezávislých premenných, ktorými sú hrúbka materiálu H, priemer trysky P, tlak T a prietok abrazíva A. Vychádzali sme z viacnásobného lineárneho regresného modelu

$$R_i = \beta_0 + \beta_1 H_i + \beta_2 P_i + \beta_3 T_i + \beta_4 A_i + \beta_5 H_i^2 + \beta_6 P_i^2 + \beta_7 T_i^2 + \beta_8 A_i^2 + \\ + \beta_9 H_i P_i + \beta_{10} H_i T_i + \beta_{11} H_i A_i + \beta_{12} P_i T_i + \beta_{13} P_i A_i + \beta_{14} T_i A_i + \varepsilon_i , \qquad (1)$$

kde ε_i je chyba merania a $i = 1, 2, ..., 80$.

Pri vyhodnotení modelu (1) na celej oblasti nameraných dát sme spätnou krokovou regresiou dostali najlepší model so závislými premennými *TA, H, P,* T^2, H^2, *HP, HT,* pri ktorom upravený koeficient viacnásobnej determinácie $R^2_{\text{adj}} = 0{,}738$ a reziduálny rozptyl regresie $MS_E = 75{,}89$. Znamená to, že iba približne 75% variability vysvetľuje uvedený model, čo možno považovať za slabý výsledok. Na základe tohto výsledku by nebolo možné predikovať vstupné veličiny pri rezaní abrazívnym kvapalinovým lúčom (H, P, T, A).

2. Spracovanie nameraných dát v programovom systéme SAS

V systéme SAS sme zrealizovali analýzu údajov na celej oblasti hrúbky materiálu. V prvom kroku sme spracovali viacrozmerný regresný model pozostávajúci zo závislej premennej R a regresných premenných *H, P, T, A,* H^2, P^2, T^2, A^2, *HP, HT, HA, PT, PA, TA,* pri fixácií nezávislých premenných najprv troch a potom dvoch z premenných – priemer trysky P, tlak T a hmotnosť prietoku abrazíva A. Pri každej fixácií parametrov sme urobili samostatnú analýzu pomocou spätnej krokovej regresie, ktorá vylúčila nevýznamné regresné premenné. Získali sme tak v prvom prípade jeden regresný model a v druhom prípade tri regresné modely.

V tabuľke 1 sú uvedené regresné premenné najlepších modelov vybraných pri fixácií dvoch premenných a troch premenných. Z tabuľky vidieť, že najlepší je model s regresnými premennými *T, H,* H^2, *HT*, pri ktorom fixujeme dve regresné premenné – priemer trysky P a prietok abrazíva A. Hoci miera celkovej strednej kvadratickej chyby regresného modelu $C_p = 3{,}1412$ nie je najmenšia v porovnaní s ostatnými modelmi, upravený koeficient

viacnásobnej determinácie $R^2_{\text{adj}} = 0{,}8592$ je zo všetkých štyroch výrazne najväčší. Znamená to, že len približne 86% variability vysvetľuje uvedený model. Takýto model je však tiež nedostačujúci pre nastavenie reznej rýchlosti v riadiacom systéme stroja na základe definovaných premenných – hrúbka materiálu *H*, prietok trysky *P*, tlak *T* a hmotnosť prietoku abrazíva *A*.

Tabuľka 1 Zoznam vybraných modelov 1

Fixácia	**Regresné premenné v modeli**	R^2_{adj}	C_p
P, T, A	$H,\ H^2$	0,6818	3,0000
P, T	$H, A,\ H^2$	0,7412	3,1127
P, A	$T, H,\ H^2,\ HT$	**0,8592**	**3,1412**
T, A	$H, P,\ H^2$	0,7259	2,7795

Definovali sme preto nové regresné premenné. Odvodili sme ich z hrúbky materiálu berúc do úvahy nelineárnu závislosť reznej rýchlosti na hrúbke materiálu pri fixovaní priemeru trysky *P*, tlaku *T* a hmotnosti prietoku abrazíva *A*. Nové regresné premenné sú H^{-1}, $(\ln H)^{-1}$, e^{-H}.

Celkove máme k dispozícii regresné premenné *H, P, T, A,* H^2, P^2, T^2, A^2, *HP, HT, HA, PT, PA, TA,* H^{-1}, $(\ln H)^{-1}$, e^{-H} a závislú premennú *R*. Pri fixácii nezávislých premenných najprv troch a potom dvoch z premenných – priemer trysky *P*, tlak *T* a hmotnosť prietoku abrazíva *A* – sme spätnou krokovou regresiou dostali významné regresné premenné pre jednotlivé modely.

Regresné premenné najlepších modelov vybraných pri fixácií dvoch premenných a troch premenných sú uvedené v tabuľke 2.

Tabuľka 2 Zoznam vybraných modelov 2

Fixácia	**Regresné premenné v modeli**	R^2_{adj}	C_p
P, T, A	$H^{-1},\ (\ln H)^{-1},\ \mathrm{e}^{-H}$	**0,9997**	**2,3627**
P, T	$A,\ A^2,\ HA,\ H^{-1},\ (\ln H)^{-1},\ \mathrm{e}^{-H}$	0,9828	5,2963
P, A	$H,\ T^2,\ HT,\ H^{-1},\ (\ln H)^{-1},\ \mathrm{e}^{-H}$	0,9940	4,0617
T, A	$H, P, HP,\ H^{-1}$	0,9918	3,5348

Z tabuľky 2 vidieť, že najlepší model dostaneme pri fixovaní troch regresných premenných – priemer trysky *P*, tlak *T* a prietok abrazíva *A*. Tento model sa skladá s regresných premenných H^{-1}, $(\ln H)^{-1}$ a e^{-H}. Hodnota upraveného koeficientu viacnásobnej determinácie modelu je $R^2_{\text{adj}} = 0{,}9997$, čo znamená, že približne 99,9% variability vysvetľuje uvedený model. Hodnota miery celkovej strednej kvadratickej chyby regresného modelu $C_p = 2{,}3627$ je zo všetkých najmenšia, čo potvrdzuje, že model je z uvedených štyroch modelov najlepší.

Model môžeme považovať za veľmi dobrý, avšak pri fixácii troch regresných premenných ide o jednoduchý model závislosti reznej rýchlosti R od hrúbky materiálu H. Teda tento model nerieši náš problém.

V ďalšej časti sme sa preto rozhodli spracovať regresný model pomocou krokovej regresie so všetkými regresnými premennými, teda bez fixácie. Zobrali sme do úvahy regresné premenné $H, P, T, A, H^2, P^2, T^2, A^2, HP, HT, HA, PT, PA, TA, H^{-1}, (\ln H)^{-1}, \mathrm{e}^{-H}$ a všetky namerané dáta.

Vychádzali sme z viacnásobného lineárneho regresného modelu

$$\begin{aligned} R_i = {} & \beta_0 + \beta_1 H_i + \beta_2 P_i + \beta_3 T_i + \beta_4 A_i + \beta_5 H_i^2 + \beta_6 P_i^2 + \beta_7 T_i^2 + \beta_8 A_i^2 + \\ & + \beta_9 H_i P_i + \beta_{10} H_i T_i + \beta_{11} H_i A_i + \beta_{12} P_i T_i + \beta_{13} P_i A_i + \beta_{14} T_i A_i + \varepsilon_i , \\ & + \beta_{15} H^{-1} + \beta_{16} (\ln H)^{-1} + \beta_{17} \mathrm{e}^{-H} + \varepsilon_i \end{aligned} \qquad (2)$$

kde ε_i je chyba merania a $i = 1, 2, ..., 80$.

Potom sme spätnou krokovou regresiou vypočítali odhady parametrov regresných modelov pre príslušné regresné premenné, ktoré sú uvedené v tabuľke 3.

Tabuľka 3 Zoznam vybraných modelov 3

Model	**Regresné premenné v modeli**	R^2_{adj}	C_p
1	$H, P, T, A, H^2, P^2, T^2, A^2, HP, HT, HA, PT, PA, H^{-1}, (\ln H)^{-1}, \mathrm{e}^{-H}$	0,9821	16,0000
2	$H, T, A, H^2, P^2, T^2, A^2, HP, HT, HA, PT, PA, H^{-1}, (\ln H)^{-1}, \mathrm{e}^{-H}$	0,9821	14,0250
3	$H, T, A, H^2, P^2, T^2, A^2, HP, HT, HA, PA, H^{-1}, (\ln H)^{-1}, \mathrm{e}^{-H}$	0,9824	14,0250
4	$H, A, H^2, P^2, T^2, A^2, HP, HT, HA, PA, H^{-1}, (\ln H)^{-1}, \mathrm{e}^{-H}$	0,9824	12,0505
5	$A, H^2, P^2, T^2, A^2, HP, HT, HA, PA, H^{-1}, (\ln H)^{-1}, \mathrm{e}^{-H}$	0,9827	10,1064
6	$A, H^2, T^2, A^2, HP, HT, HA, PA, H^{-1}, (\ln H)^{-1}, \mathrm{e}^{-H}$	**0,9829**	**9,2878**

Z tabuľky 3 vidieť, že najlepší je model 6 s regresnými premennými $A, H^2, T^2, A^2, HP, HT, HA, PA, H^{-1}, (\ln H)^{-1}, \mathrm{e}^{-H}$. Tento model má najväčšiu hodnotu upraveného koeficienta viacnásobnej determinácie $R^2_{\mathrm{adj}} = 0{,}9829$ a najmenšiu hodnotu miery celkovej strednej kvadratickej chyby regresného modelu $C_p = 9{,}2878$, čo potvrdzuje, že uvedený model je najlepší. Odhad regresnej funkcie je

$$\begin{aligned} \hat{R} = {} & 1247{,}96174 + 0{,}92360A + 0{,}05927H^2 + 0{,}00224T^2 - 0{,}00031151A^2 + \\ & + 5{,}07725HP - 0{,}03912HT - 0{,}00501HA - 0{,}40752PA + \\ & + 11881H^{-1} - 5991{,}73271(\ln H)^{-1} + 8047{,}79008\,\mathrm{e}^{-H} \end{aligned} \qquad (3)$$

Test významnosti regresie v ANOVA tabuľke (tabuľka 4) nám hovorí, že zamietame nulovú hypotézu, že všetky parametre modelu sú rovné nule. Podobne testovanie významnosti parametrov a absolútneho člena modelu *t*-testom uvedené v tabuľke 5 ukazuje, že všetky parametre vrátane absolútneho člena sú vysoko významné. Teda približne 98,3% variability vysvetľuje uvedený model. Závislosť nameranej reznej rýchlosti *R* oproti predikovanej reznej rýchlosti *R* (pozri obrázok 1) je lineárna, čo potvrdzuje, že model je vhodný pre uvedené ciele.

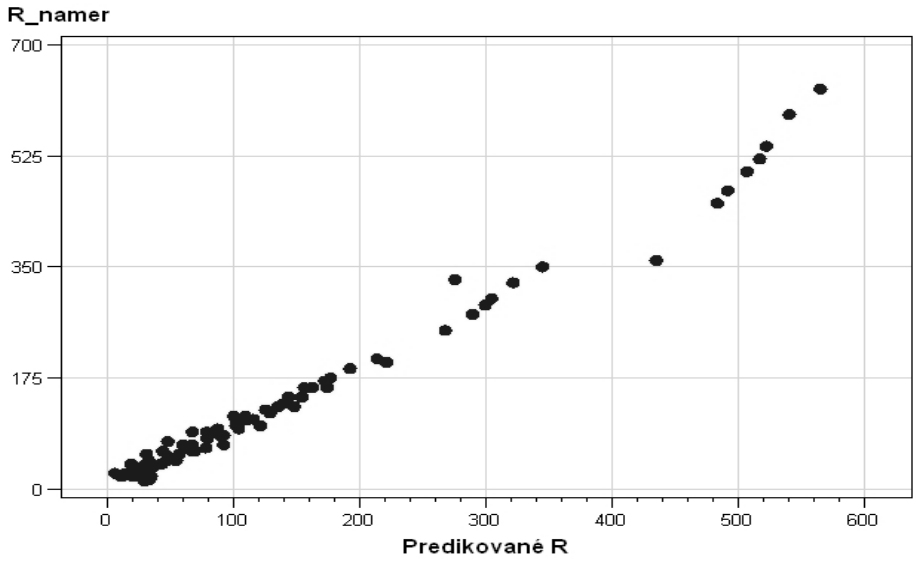
Uvedený model možno teda použiť v riadiacom systéme stroja pri nastavení reznej rýchlosti rezania abrazívnym kvapalinovým lúčom na základe zadaných veličín hrúbka materiálu *H*, priemer trysky *P*, tlak *T* a hmotnosť prietoku abrazíva *A* pri rezaní nehrdzavejúcej ocele a dosiahnutí strednej kvality rezu.

Tabuľka 4 ANOVA tabuľka pre výsledný model 6

Zdroj variability	**Súčet štvorcov**	**Stupne voľnosti**	**Priemerný štvorec**	**Testovacia štatistika**	***P*-hodnota**
Regresia	1714124	11	155829	412,86	<0,0001
Reziduálny	25666	68	377,43669		
Celkový	1739790	79			

Tabuľka 5 Testovanie významnosti parametrov výsledného modelu 6 *t*-testom

Parametre	**Hodnota odhadu parametra**	**Smerodajná chyba**	***t*-hodnota**	***P*-hodnota**
β_0	1247,96174	394,37695	3,16	0,0023
β_4	0,92360	0,13760	6,71	<0,0001
β_5	0,05927	0,02146	2,76	0,0074
β_7	0,00224	0,00027928	8,03	<0,0001
β_8	-0,00031151	0,00014723	-2,12	0,0380
β_9	5,07725	1,17157	4,33	<0,0001
β_{10}	-0,03912	0,00712	-5,50	<0,0001
β_{11}	-0,00501	0,00141	-3,54	0,0007
β_{13}	-0,40752	0,06762	-6,03	<0,0001
β_{15}	11881	2680,96904	4,43	<0,0001
β_{16}	-5991,73271	1520,69437	-3,94	0,0002
β_{17}	8047,79008	1851,95100	4,35	<0,0001



Obrázok 1 Grafické posúdenie kvality výsledného modelu 6

3. Záver

Pri hľadaní modelu na celej oblasti hrúbky materiálu sme získali najlepší model pri zaradení nových regresných premenných H^{-1}, $(\ln H)^{-1}$, e^{-H} do regresného modelu. Začínali sme teda s regresnými premennými *H, P, T, A,* H^2, P^2, T^2, A^2, *HP, HT, HA, PT, PA, TA,* H^{-1}, $(\ln H)^{-1}$, e^{-H}. Pri použití krokovej regresie (spätná eliminácia) sme vylúčili nevýznamné regresné premenné a dostali sme odhad regresnej závislosti (2), čo zodpovedá modelu 6 z tabuľky 3. Model má vysokú hodnotu upraveného koeficienta viacnásobnej determinácie ($R^2_{\text{adj}} = 0{,}9829$), je to teda dobrý model. Potvrdzuje nám to aj test významnosti regresie (pozri tabuľku 4), kde *P*-hodnota je nižšia ako 0,0001. Tiež *t*-testy (pozri tabuľku 5) poukazujú na vysokú významnosť všetkých parametrov modelu vrátane absolútneho člena (všetky *P*-hodnoty v tabuľke 5 sú veľmi nízke).

Vypočítaná rezná rýchlosť $\hat{R}$ pomocou získaného regresného modelu (3) pre zadané veličiny *H, P, T*, *A* je uvedená v tabuľke 6.

Na základe definovaných veličín hrúbka materiálu *H*, priemer trysky *P*, tlak *T* a hmotnosť prietoku abrazíva *A* vieme určiť reznú rýchlosť, ktorú je potrebné nastaviť na riadiacom systéme stroja pri rezaní nehrdzavejúcej ocele a dosiahnutí strednej kvality rezu.

4. Literatúra

[1]JANIGA, I., KÁLL, M., GELETA, V., 2006, Model Building of Relationship between of Waterjet Cutting Speed and Relevant Quantities Influencing Cutting Quality. In FORUM STATISTICUM SLOVACUM. ISSN 1336-7420. 3/2006, s. 97 – 104.

[2]JANIGA, I., KÁLL, M., GELETA, V., 2006, Relationship between Waterjet Cutting Speed and Relevant Quantities Influencing Cutting Quality. In FORUM STATISTICUM SLOVACUM. ISSN 1336-7420. 5/2006, s. 79 – 82.

[3]STANKOVIČOVÁ, I., VOJTKOVÁ, M., 2007: Viacrozmerné štatistické metódy s aplikáciami. 1. vyd. Bratislava: Iura Edition, s. 261. ISBN 978-80-8078-152-1

Tabuľka 6 Porovnanie nameranej a vypočítanej reznej rýchlosti na základe modelu 6

H [mm]	***P*** [mm]	***T*** [MPa]	***A*** [g / min]	***R*_nam** [mm / min]	***R*_vyp** [mm / min]
3,00	0,90	300,00	450,00	520,00	516,42
5,00	1,10	330,00	450,00	300,00	303,85
5,00	0,76	300,00	450,00	325,00	321,10
8,00	1,10	300,00	450,00	145,00	143,34
8,00	1,10	330,00	450,00	175,00	176,29
8,00	1,10	300,00	650,00	160,00	161,86
8,00	0,76	300,00	450,00	190,00	191,88
8,00	0,90	300,00	450,00	170,00	171,89
10,00	1,10	300,00	650,00	125,00	125,05
10,00	1,10	300,00	450,00	110,00	108,54
12,00	1,10	300,00	650,00	100,00	101,44
12,00	0,90	300,00	450,00	110,00	111,42
15,00	1,10	300,00	450,00	70,00	67,14
15,00	1,10	300,00	650,00	80,00	78,64
15,00	0,90	300,00	450,00	85,00	88,58
20,00	1,10	300,00	450,00	50,00	49,94
20,00	1,10	300,00	650,00	55,00	56,43
30,00	1,10	300,00	450,00	35,00	35,94
30,00	1,10	330,00	450,00	40,00	43,06
30,00	1,10	300,00	650,00	35,00	32,41
30,00	0,76	300,00	450,00	45,00	46,50
30,00	0,90	300,00	450,00	40,00	42,15
40,00	1,10	330,00	450,00	30,00	27,14
40,00	1,10	300,00	250,00	20,00	20,37
40,00	0,90	300,00	450,00	25,00	27,80
50,00	0,90	300,00	450,00	20,00	20,18

Adresa autorov:

Ivan Janiga, doc, RNDr, PhD.
Strojnícka fakulta STU
Nám, slobody17
812 31 Bratislava
Ivan.janiga@stuba.sk

Iveta Stankovičová, Ing., PhD.
Fakulta managementu UK
Katedra informačných systémov
Odbojárov 10, 820 05 Bratislava
iveta.stankovicova@fm.uniba.sk

Marek Kall, Ing.
Vajnorská 98/C
831 04 Bratislava
mkal@gratex.com

Poďakovanie:

Tento príspevok vznikol s podporou grantových projektov VEGA č. 1/3182/06 Zlepšovanie kvality produkcie strojárskych výrobkov pomocou moderných štatistických metód a VEGA č. 1/1247/08 Kvantitatívne metódy v stratégii šesť sigma.

Viacrozmerný regulačný diagram v procese testovania presnosti polohovania rezacieho zariadenia
Multivariable control chartin testing process of cutting device positioning accuracy

Ivan Janiga, Jana Škorvagová

Abstract: There are situations in practice when two or more characteristics has to be controlled on the same product or testing process. In checking the process it can be used multivariate control chart. The contribution deals with using of the multivariate control chart in testing process of cutting device accuracy.

Key words: multivariate control chart, cutting device positioning accuracy.

Kľúčové slová: viacrozmerný regulačný diagram, presnosť polohovania meracieho zariadenia.

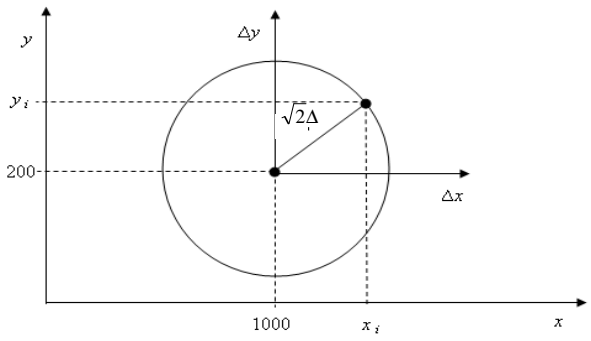
1. Úvod

Plazmové a kyslíkovo-acetylénové rezacie centrá PLASMACUTTER sú určené na presné tvarové delenie materiálov. Sú programovateľné a riadené CNC riadiacim systémom. Riadiaci systém stroja môže riadiť spojite 6 pohybových osí stroja, čo umožňuje spojitú zmenu uhla naklonenia horáka vzhľadom k zvislej súradnicovej osi z v procese rezania.

Na stanovenie presnosti polohovania sa vykonal experiment. Presnosť polohovania sa merala pomocou softvérového kľúča pri trajektórii pohybu do stanoveného bodu $[x_0, y_0]$ = [1000, 200]. V i-tom pokuse sa merala odchýlka polohy ramena od stanoveného bodu $[\Delta x_i; \Delta y_i] = [x_i, -1000; y_i - 200]$, pričom $i = 1, 2, \ldots, 100$. Namerané dvojrozmerné dáta obsahovali 100 hodnôt odchýlok.

Pôvodné vyhodnotenie presnosti sa vykonávalo tak, že najväčšia odchýlka z odchýlok Δx_i, Δy_i, $i = 1, 2, \ldots, 100$ sa považovala za presnosť polohovania. Znamenalo to, že za rovnakú presnosť sa považovala poloha na obvode štvorca so stranami $2\Delta x = 2\Delta y = 2\Delta$, pričom stred štvorca je v bode [1000, 200]. Podľa tejto metodiky je presnosť rovná Δ. Ak však zoberieme napríklad štyri koncové body štvorca ich vzdialenosť od stanoveného bodu [1000, 200] je $\sqrt{2}\Delta$, čo je väčšie ako Δ. Uvedená metodika na stanovenie presnosti polohovania je teda nesprávna, preto sme pristúpili k inému riešeniu.

Metodika, ktorú navrhujeme, považuje za rovnakú presnosť polohu na obvode kružnice s polomerom Δ a stredom v stanovenom bode [1000, 200], ako vidieť na obrázku 1.



Obrázok 1: Odchýlky od stanoveného bodu

2. Overenie podmienok na regulácia procesu presnosti polohovania

Na regulovanie procesu polohovania možno použiť regulačný diagram. Pretože ide o dvojrozmerné dáta, možno použiť viacrozmerný regulačný diagram, musí byť však splnená podmienka normality.

Označme ako X náhodnú premennú, ktorej realizácie sú merania na x-ovej osy a Y náhodnú premennú s realizáciami na y-ovej osy. Pretože merané hodnoty sú párované, máme k dispozícii dvojrozmernú spojitú náhodnú premennú (X, Y), ktorej združenú funkciu hustoty označíme ako $f_{XY}(x, y)$. Keď poznáme združenú hustotu, poznáme aj marginálne funkcie hustoty $f_X(x)$ a $f_Y(y)$.

Podmienka použitia viacrozmerného regulačného diagramu je, že dvojrozmerné dáta sú hodnoty náhodného výberu z dvojrozmerného normálneho rozdelenia. Na overenie tejto podmienky postačuje testovať nulové hypotézy pre marginálne rozdelenia.

$H_0^{(X)}$: X má rozdelenie pravdepodobnosti s funkciou hustotou $f_X(x)$,

$H_0^{(Y)}$: Y má rozdelenie pravdepodobnosti s funkciou hustotou $f_Y(y)$.

Hodnoty nameraných odchýlok na x-ovej resp. y-ovej osy považujeme za realizáciu náhodného výberu z rozdelenia s funkciou hustoty $f_X(x)$ resp. $f_Y(y)$. a označíme ako $V_X = \{\Delta x_1, \Delta x_2, \ldots, \Delta x_{100}\}$ resp. $V_Y = \{\Delta y_1, \Delta y_2, \ldots, \Delta y_{100}\}$. Na základe výberov V_X, V_Y môžeme testovať hypotézy $H_0^{(X)}$, $H_0^{(Y)}$. Na testovanie normality sme použili päť testov (Kolmogorovov-Smirnovov D, Kuiperov V, Cramerov-Von Misesov W^2, Watsonov U^2 a Andersonov-Darlingov A^2). V tabuľke 1 sú uvedené *P*–hodnoty pre jednotlivé testy.

Tabuľka 1: Testy normality nameraných dát (P–hodnoty)

Výber	Test D	Test V	Test W^2	Test U^2	Test A^2	Normálne rozdelenie	1 - α
V_X	<0,10	<0,05	0,1615	0,1390	0,1948	nemá	95%
V_Y	<0,01	<0,01	0,0001	0,0002	0,0000	nemá	99%

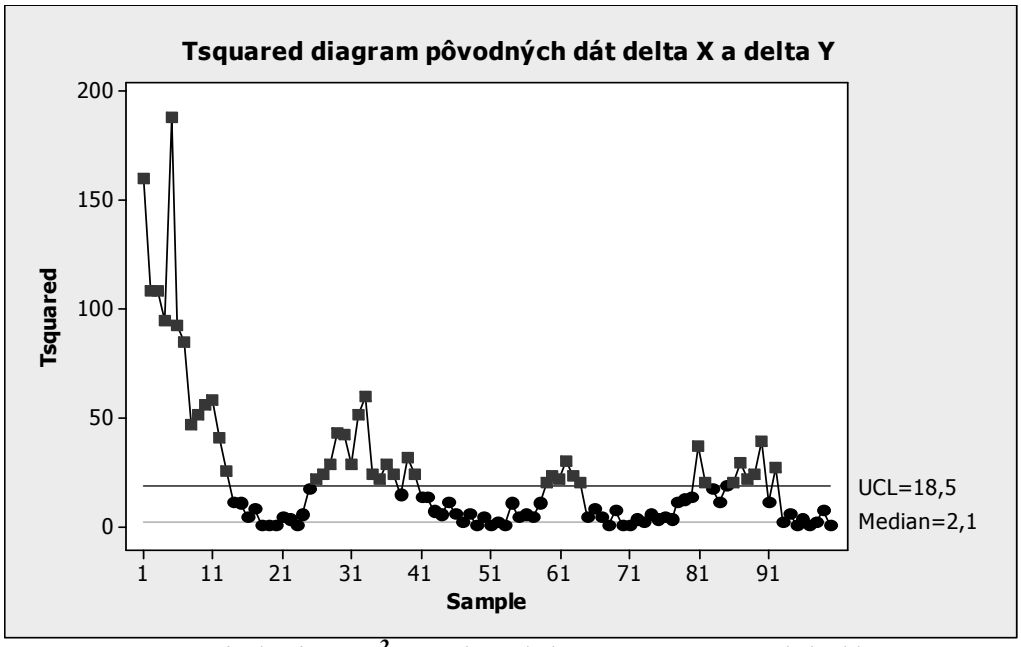
Z tabuľky 1 vidieť, že obidve nulové hypotézy zamietame a tvrdíme, že ani jedna z náhodných premenných nemá normálne rozdelenie. Teda podmienka použitia viacrozmerného regulačného diagramu nie je splnená.

Napriek nesplneniu predpokladov o normalite sme vyhodnotili dáta pomocou viacrozmerného regulačného diagramu. Na obrázku 2 je T^2 regulačný diagram, z ktorého vidieť, že dáta vykazujú silnú nestabilitu. Približne polovica hodnôt na diagrame je nad hornou regulačnou medzou. Podobne je to aj na obrázku 3, kde je uvedený diagram všeobecného rozptylu. Tento diagram ukazuje tiež nestabilitu, pretože tri body ležia nad hornou regulačnou medzou.

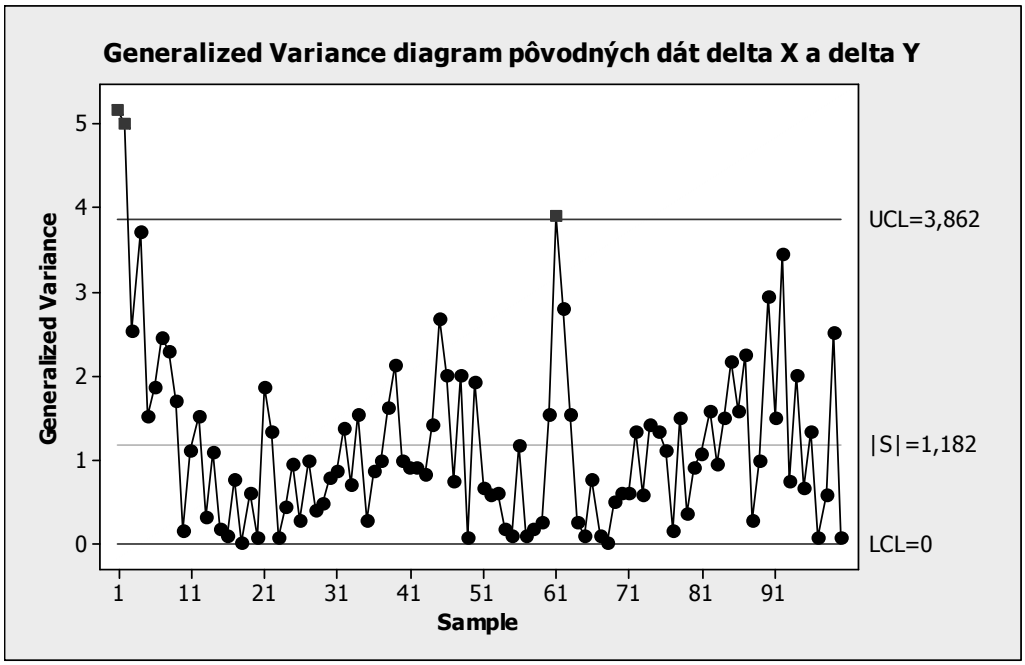
Na záver zhrnieme to, čo sme doteraz zistili. Namerané dvojrozmerné dáta nepochádzajú z dvojrozmerného normálneho rozdelenia, čo znamená, že proces testovania presnosti rezacieho zariadenia s nameranými dátami sa nedá regulovať pomocou viacrozmerného regulačného diagramu. Z uvedeného dôvodu sme pristúpili k transformácii nameraných dát. Hľadali sme vhodnú transformáciu dát tak, aby sme získali dvojrozmerné normálne rozdelenie.

3. Regulácia procesu presnosti polohovania pomocou transformovaných dát

Z nameraných dát sme v programe Mathematica 5.1 vypočítali výberovú kovariančnú maticu, ktorú sme následne použili v programe Minitab 15 na transformáciu nameraných dát



Obrázok 2: T^2 regulačný diagram nameraných dát



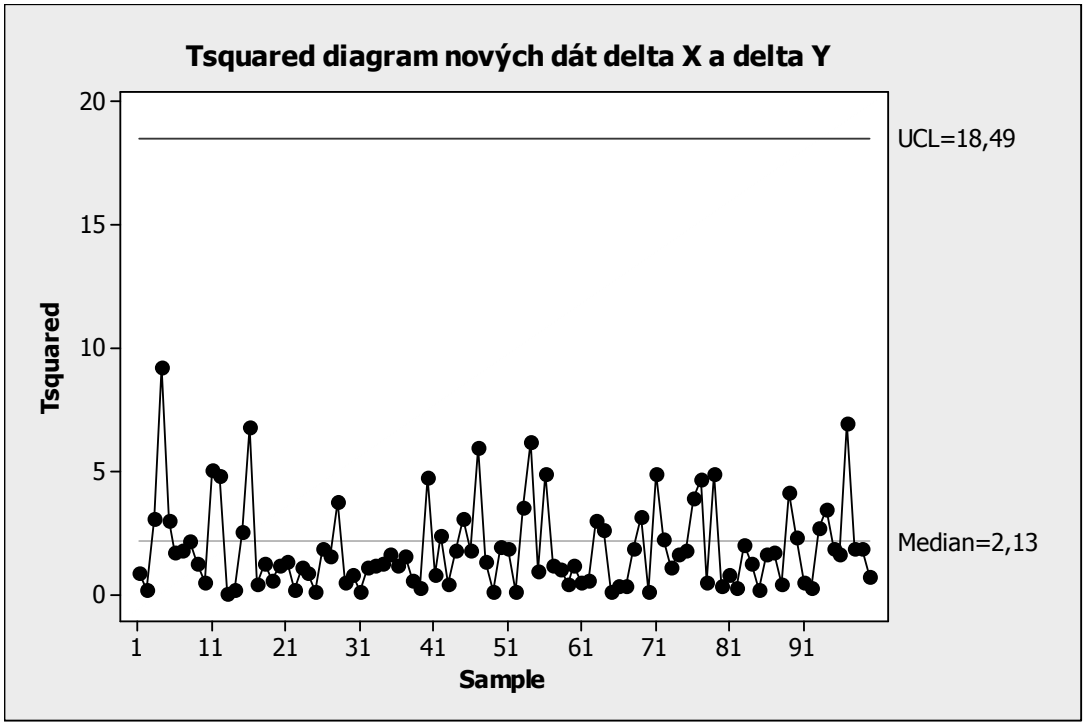
Obrázok 3: Diagram všeobecného rozptylu nameraných dát

na dáta s normálnym rozdelením. Normalitu transformovaných dát sme overovali pomocou testov, ktorými sme overovali normalitu nameraných dát. Výsledky sú uvedené v tabuľke 2.

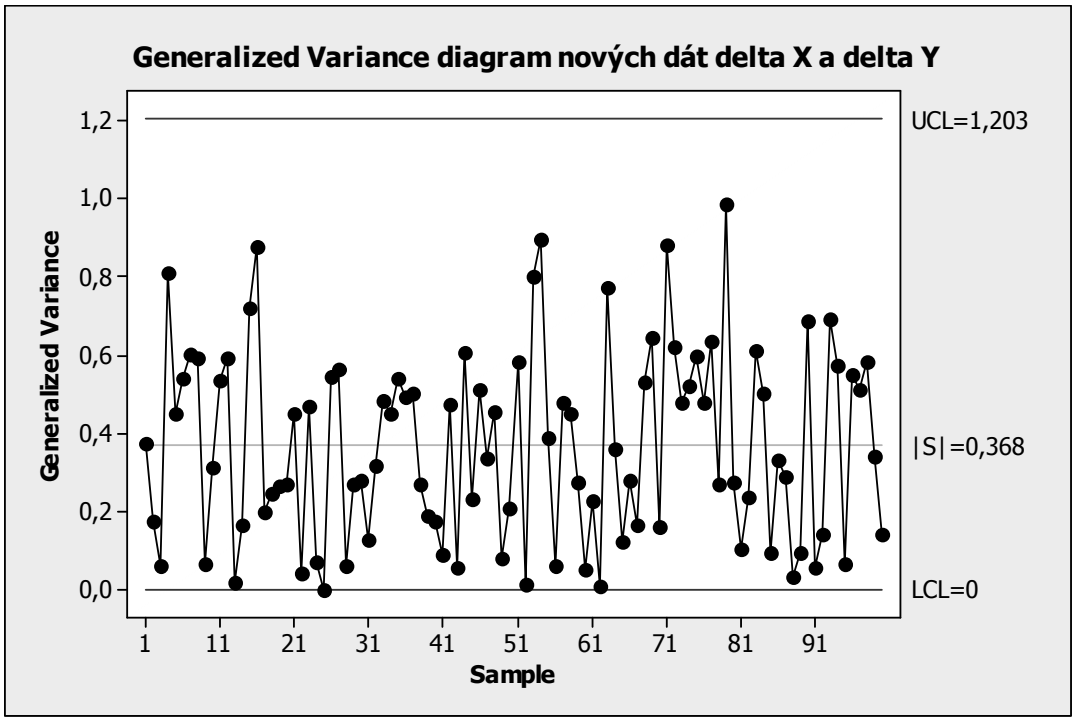
Tabuľka 2: Testy normality trasformovaných dát (P–hodnoty)

Výber	Test D	Test V	Test W^2	Test U^2	Test A^2	Normálne rozdelenie	1 - α
V_X	≥0,10	≥0,10	0,6762	0,6282	0,7622	má	≥90 %
V_Y	≥0,10	≥0,10	0,4827	0,5044	0,5455	má	≥90 %

Z tabuľky 2 vidieť, že obidve nulové hypotézy prijímame a tvrdíme, že obidve náhodné premenné majú normálne rozdelenie. Teda podmienka normality na korektné použitie viacrozmerného regulačného diagramu je splnená. Na transformované dáta sme použili viacrozmerný regulačný diagram.



Obrázok 4: T^2 regulačný diagram transformovaných dát



Obrázok 5: Diagram všeobecného rozptylu transformovaných dát

Z obrázkov 4 a 5 vidieť, že v obidvoch diagramoch hodnoty kolíšu náhodne a žiadna hodnota nie je mimo regulačných medzí, čo znamená, že proces je štatisticky stabilný. Proces presnosti polohovanie teda môžeme regulovať pomocou viacrozmerného regulačného diagramu transformovaných odchýlok od stanoveného bodu.

4. Záver

V príspevku sme sa zaoberali možnosťou regulácie v procese testovania presnosti polohovania rezacieho zariadenia. Zistili sme, že namerané dvojrozmerné dáta nemajú normálne rozdelenie. Nebolo možné implementovať viacrozmerný regulačný diagram v procese testovania. Našli sme vhodnú transformáciu dát, ktoré už mali normálne rozdelenie. Na transformované dáta sme použili viacrozmerný regulačný diagram, ktorý ukazoval stabilitu procesu testovania. Teda proces testovania presnosti polohovania rezacieho zariadenia sa dá regulovať, avšak iba pomocou transformovaných dát. Štatistickými metódami v riadení kvality sa zaoberajú publikácie [1], [2], [3], [4], [5].

5. Literatúra

[1]MASON, R.L., YOUNG, J.C. 2002. *Multivariate Statistical Process Control with Industrial Applications*. SIAM, ASA, s. 261, ISBN 0-89871-496-6.

[2]ŠKORVAGOVÁ, J. 2008. Praktické využitie viacrozmerných regulačných diagramov. Slovenská technická univerzita v Bratislave, Strojnícka fakulta, Bakalárska práca.

[3]JANIGA, I., CISKO, P. 2008. Modified control charts in proces control. In: Quality, Environment, Health Protection and Safety Management Development Trends: Proceedings of International scientific conference. NEUM (Bosnia and Herzegovina). 2008, ISBN 978-80-7399-479-2, s. 120 – 125.

[4]TEREK, M., HRNČIAROVÁ, Ľ. *Štatistické riadenie kvality*. Vydavateľstvo IURA EDITION, 2004, 234 s. ISBN 80-89047-97-1.

[5]GARAJ, I. Požiadavky na rozsah náhodného výberu jednostranných preberacích plánov meraním. In *FORUM STATISTICUM SLOVACUM*. ISSN 1336-7420, 1/2006, s.38-43.

Poďakovanie:

Tento príspevok vznikol s podporou grantových projektov VEGA č, 1/3182/06 Zlepšovanie kvality produkcie strojárskych výrobkov pomocou moderných štatistických metód a VEGA č, 1/1247/08 Kvantitatívne metódy v stratégii šesť sigma,

Adresa autorov:

Ivan Janiga, doc. RNDr. PhD.
Strojnícka fakulta STU
Nám, slobody17
812 31 Bratislava
ivan.janiga@stuba.sk

Jana Škorvagová, Bc. - diplomantka
Strojnícka fakulta STU
Nám, slobody17
812 31 Bratislava
janka200@zoznam.sk

Rodinná struktura na Chrudimsku v roce 1651
Family structure in Chrudim area in 1651

Eva Kačerová

Abstract: One of the most valuable sources for Czech historical demography is the List of Serfs according to Faith of 1651. It captured the situation immediately following the end of the Thirty Years War. The most complicated pitfall for the researcher studying the age structure of the population on the basis of the List is the sub-registration of children before the age at which they can make confessions on certain estates (or in certain communities). The children are recorded systematically only from the age of 12. This article is focused on family structure at Chrudim area. This article came into being within the framework research project IGA VŠE 15/08.

Key words: historical demography, 17th century, Chrudim area, family structure

Klíčová slova: historická demografie, 17. století, Chrudimsko, struktura dle rodinného stavu

1. Úvod

Soupis podaných podle víry z roku 1651 dává jedinečným způsobem nahlédnout do složení tehdejší rodiny. Rozbor struktury obyvatelstva podle rodinného stavu umožňuje přiblížit se odpovědi k jakému typu evropské rodiny patřila rodina žijící v polovině 17. století v Čechách. Podle J. Hajnala byla západoevropská rodina tvořená jen manželským párem a jeho dětmi. Sňatkový věk byl poměrně vysoký, protože do manželství vstupovali jen ti, kteří měli zajištěnou existenci. Východoevropská rodina byla rodinou komplexní, kde možnost uzavření manželství nebyla vázána na samostatné bydlení, a proto byl sňatkový věk poměrně nízký. Střední Evropa je oblastí, o jejímž zařazení k západoevropskému nebo východoevropskému typu se dosud vedou mezi historiky diskuse [1].

2. Struktura obyvatel Chrudimska dle rodinného stavu

Studium struktury obyvatelstva podle rodinného stavu ze Soupisu poddaných podle víry z roku 1651 je však zatíženo jistými problémy. Kolonka „rodinný stav“ mezi rubrikami Soupisu nebyla, a proto ne vždy je snadné jej určit. Situace je celkem jasná u osob žijících v manželství. Soupisem bylo na Chrudimsku zachyceno 46 626 osob, avšak na panství Bystrá a Čankovice u 1 958 nebyl evidován věk, a tak je dále pojednáváno pouze o 44 668 osobách. Děti byly systematicky evidovány zhruba od 12 let věku. Manželé jsou v Soupise téměř vždy psaní pod sebou (s výjimkou tří případů nejdříve muž a pod ním žena) tak, že v kolonce povolání je u muže zapsáno řemeslo, funkce v obecní samosprávě nebo označení podle rozlohy vlastněné půdy (chalupník, sedlák, zahradník) a u ženy „manželka“ nebo „žena jeho“.

Poněkud složitější situace je u ostatních osob. V mnoha případech najdeme označení vdovec nebo vdova, avšak patrně tak nebyly označeny všechny ovdovělé osoby. Není výjimkou, že Soupis eviduje domácnost, v níž žijí děti a muž bez manželky. Protože je vysoce nepravděpodobné, že by děti, které navíc jsou označeny jako syn či dcera, byly nemanželskými dětmi tohoto muže a dokonce jím byly vychovávány, byli muži z těchto rodin považováni za ovdovělé. Pokud podobná situace nastala u ženy, tj. v domácnosti s ní žily její děti avšak nikoli manžel, vzniká pochybnost, zda-li se nejednalo o svobodnou matku. Na tuto otázku však nelze dát jednoznačnou odpověď. Pokud byla taková žena v čele domácnosti nebo pokud měla dvě a více dětí, byla považována za vdovu, v ostatních případech nebylo možné rodinný stav určit.

Poměrně často není o rodinném stavu jednotlivých osob zapsáno nic. To se téměř bez výjimky týká veškeré čeledi a osob s označením syn a dcera. Pokud byl jejich věk nižší než 35 let, byly automaticky zařazeny mezi osoby svobodné. U čeledi tomu tak s nejvyšší pravděpodobností i bylo.U synů a dcer tato jistota chybí. Nejsou totiž ojedinělé případy, kdy u rodičů žil „syn“ a „manželka jeho“, případně „dcera“ a „zeť“. Jak by asi byl Soupisem označen bezdětný syn (nebo s dětmi tak malými, že je Soupis nezaznamenal) žijící u rodičů poté, co ovdověl? Odpověď, že jednoduše „syn“, není sice nijak ověřená, ale není vyloučená.

U podruhů a především podruhyň žijících bez manželského partnera najdeme jen označení jejich sociálního postavení, nikoli však zmínku o tom, zda ve svém životě někdy byli či nebyli ženati nebo vdány, a tak mladší 35 let byli považováni za svobodné, u 35letých a starších rodinný stav nebyl zjištěn.

Věk při prvním sňatku je jedním z kritérií zařazení do typologie historického utváření rodiny. Z pramenů, z nichž nelze zjistit přesný věk při prvním sňatku a k nimž patří i Soupis poddaných podle víry z roku 1651, lze zjistit podíly vdaných žen v jednotlivých věkových skupinách.

Reprodukční věk ženy bývá vymezován věkem 15–50 let, avšak v tradiční společnosti se dbalo na to, aby nedocházelo k plození dětí mimo manželství, a tak věk při vstupu do manželství významně ovlivňoval délku reprodukční aktivity ženy. Reprodukční aktivita častěji než překročením výše zmíněné věkové hranice bývala ukončena buď smrtí ženy, nebo jejího manžela, což ji natrvalo či jen dočasně vyřazovalo z reprodukčního procesu. Nejmladší vdanou ženou na Chrudimsku byla teprve 11letá Salomena, žena 36letého sedláka Matouše Kryštofa z Dlouhé Třebové na panství Lanškroun. Takto nízký sňatkový věk je dokonce pod hranicí, kterou tehdy pro ženy připouštělo kanonické právo – ta byla stanovena na 12 let pro ženy a 14 let pro muže [2]. Ženy se začínaly vdávat v 15 letech, avšak vdané 15, 16 a 17leté byly spíše výjimkou, je-li možné na Chrudimsku hovořit o výjimce, pokud v této věkové skupině byla více než pětina žen vdaných (tab. 1). Maximální podíl vdaných žen byl na Chrudimsku ve věkové skupině 25–29 let. Dívky se tedy vdávaly spíše v mladším věku a se vrůstajícím věkem jejich šance na sňatek klesaly, a tak ve věkové skupině 30–34 bylo svobodných žen 18 % z dané věkové skupiny zatímco u mužů to bylo pouze 12 %.

Samozřejmě je zde na místě připomenout problematiku zaokrouhlování věku. U mnoha mladých vdov byl zapsán věk 30 let. Lze říci, že prostě k tomuto rodinnému stavu v takto mladém věku příslušelo stáří právě 30 let. Větší naději na opětovný sňatek měly ženy mladé a majetné nebo alespoň majetné. Ve městech se relativně mnoho vdov po cechovních mistrech provdávalo za tovaryše, pro než to byla cesta do vyšších pater cechovní organizace. To je zřejmě případ 40leté Kateřiny z České Třebové, která byla provdána za 30letého řezníka Jana Weydu. Spolu s nimi bydleli v témže domě také její dva synové ve věku 21 a 15 let. Lze se tedy domnívat, že ovdověla ještě v době, kdy synové byli příliš mladí na to, aby živnost převzali. Co se sociální kategorie týče, pak vdovy byly nejčastěji podruhyně žijící buď osaměle, nebo u svých dětí. Celkový podíl ovdovělých žen byl o 6 % vyšší než podíl vdovců mezi muži. Nejmladší vdovy nalezneme již ve věkové skupině 15–19letých. Některé z takto mladých vdov žily ve vlastní domácnosti – 18letá Kateřina Hurtovská ze vsi Bohoňovice na panství Litomyšl, jiné se vrátily k rodičům, např. 18letá Anna, dcera Šimona Hudečka ze vsi Plchovice na Choceňsku. Je nutno připustit, že od nich ani po sňatku neodešla a bydleli s manželem ve společné domácnosti s jejími rodiči.. Některé se dostaly do podružského postavení, když domácnost převzal syn zemřelého manžela. To je případ 18leté Doroty ze vsi Makov na panství Poličském, která žila ve společné domácnosti s vyvdaným ženatým 18letým synem. Ve věkové skupině 35–39 již podíl vdov přesahoval 10 % a mezi ženami staršími 60 let již vdovy tvořily většinu. Vdov bylo v tehdejší populaci 4krát více než vdovců.

Muži se zřídkakdy ženili před 20. narozeninami (samozřejmě je ze potřeba opět vzít v úvahu zaokrouhlování věku). Nejmladším ženatým mužem byl teprve 14letý Jan Doucha ze vsi Topol na panství Chrudim. Jako manželku má uvedenou 22letou Dorotu.

Tabulka 1: Struktura obyvatel na Chrudimsku dle věku a rodinného stavu

Věk	Muži					Ženy					Celkem
	svobodní	ženatí	vdovci	nezj.	celkem	svobodné	vdané	vdovy	nezj.	celkem	
0–4	34	0	0	0	34	24	0	0	0	24	58
5–9	96	0	0	0	96	107	0	0	0	107	203
10–14	2 454	1	0	0	2 455	2 671	7	0	0	2 678	5 133
15–19	3 054	80	0	0	3 134	3 672	783	6	3	4 464	7 598
20–24	1 560	946	4	3	2 513	1 362	2 324	45	5	3 736	6 249
25–29	485	1 688	15	1	2 189	520	2 391	92	2	3 005	5 194
30–34	316	2 364	28	4	2 712	539	2 277	166	1	2 983	5 695
35–39	13	1 579	28	122	1 742	17	1 199	180	216	1 612	3 354
40–44	7	1 785	55	156	2 003	17	1 235	395	347	1 994	3 997
45–49	3	997	45	61	1 106	3	590	208	153	954	2 060
50–54	4	1 090	94	97	1 285	10	539	333	207	1 089	2 374
55–59	1	344	28	32	405	1	152	117	50	320	725
60–64	3	485	67	66	621	2	145	221	101	469	1 090
65–69	1	136	18	20	175	0	42	36	17	95	270
70–74	0	138	26	16	180	2	29	55	20	106	286
75–79	2	32	10	4	48	0	6	8	6	20	68
80–84	0	40	5	6	51	0	6	39	14	59	110
85–89	0	13	4	4	21	0	1	4	2	7	28
90–94	0	8	4	2	14	0	0	1	1	2	16
95+	0	7	1	1	9	0	1	0	2	3	12
nezj.	21	25	1	24	71	15	31	2	28	76	147
Celkem	8 054	11 758	433	619	20 864	8 962	11 758	1 908	1 176	23 804	44 668

Podíly svobodných mužů se na Chrudimsku se vzrůstajícím věkem snižovaly. Svobodných starších 35 let bylo zanedbatelné procento. Nejmladším vdovcem, můžeme-li Soupisu věřit, byl 22letý zahradník-krčmář Jan Vach, který žil ve společné domácnosti s 18letou dcerou! Pakliže tomu tak skutečně bylo, musel si vzít pravděpodobně ženu o dost starší než byl sám, která relativně brzy po svatbě zemřela. Podobná situace nastala v domácnosti 23letého sedláka-hajného Jiříka Mudrocha ze vsi Lažany na panství Rychmburk, s nímž v domácnosti žil evidentně vyženěný 28letý syn Havel, Havlova 20letá manželka a kromě dvou žen ve služebném postavení ještě Havlův o rok mladší bratr. Stejně tak v domácnosti ovdovělého 23letého sedláka Jiříka Experata z Vlčí Habřiny na Pardubicku žila vyženěná 16letá dcera (protože Soupis nezaznamenal děti mladší 12 let, je možné, že měla mladší sourozence). V podružském postavení je uveden jeho 20letý bratr Petr a 19letý bratr Jakub s 18letou manželkou Kateřinou. V městském prostředí byl nejmladší vdovec ve věkové skupině 25–29 let – Tomáš N. z Chrudimi. 10% podíl překročili vdovci ve věkové skupině 60–64 let. Se vzrůstajícím věkem se také zvyšovalo zastoupení mužů, u nichž nebylo možno na základě zápisu uvedeného v Soupisu rozhodnout o jejich rodinném stavu. Na celém Chrudimsku bylo takovýchto mužů 619, tj. 3 %.

Tabulka 1: Struktura obyvatel na Chrudimsku dle věku a rodinného stavu

Věk	Muži					Ženy				
	svobodní	ženatí	vdovci	nezj.	celkem	svobodné	vdané	vdovy	nezj.	celkem
0–4	100,0	0,0	0,0	0,0	100,0	100,0	0,0	0,0	0,0	100,0
5–9	100,0	0,0	0,0	0,0	100,0	100,0	0,0	0,0	0,0	100,0
10–14	100,0	0,0	0,0	0,0	100,0	99,7	0,3	0,0	0,0	100,0
15–19	97,4	2,6	0,0	0,0	100,0	82,3	17,5	0,1	0,1	100,0
20–24	62,1	37,6	0,2	0,1	100,0	36,5	62,2	1,2	0,1	100,0
25–29	22,2	77,1	0,7	0,0	100,0	17,3	79,6	3,1	0,1	100,0
30–34	11,7	87,2	1,0	0,1	100,0	18,1	76,3	5,6	0,0	100,0
35–39	0,7	90,6	1,6	7,0	100,0	1,1	74,4	11,2	13,4	100,0
40–44	0,3	89,1	2,7	7,8	100,0	0,9	61,9	19,8	17,4	100,0
45–49	0,3	90,1	4,1	5,5	100,0	0,3	61,8	21,8	16,0	100,0
50–54	0,3	84,8	7,3	7,5	100,0	0,9	49,5	30,6	19,0	100,0
55–59	0,2	84,9	6,9	7,9	100,0	0,3	47,5	36,6	15,6	100,0
60–64	0,5	78,1	10,8	10,6	100,0	0,4	30,9	47,1	21,5	100,0
65–69	0,6	77,7	10,3	11,4	100,0	0,0	44,2	37,9	17,9	100,0
70–74	0,0	76,7	14,4	8,9	100,0	1,9	27,4	51,9	18,9	100,0
75–79	4,2	66,7	20,8	8,3	100,0	0,0	30,0	40,0	30,0	100,0
80–84	0,0	78,4	9,8	11,8	100,0	0,0	10,2	66,1	23,7	100,0
85–89	0,0	61,9	19,0	19,0	100,0	0,0	14,3	57,1	28,6	100,0
90–94	0,0	57,1	28,6	14,3	100,0	0,0	0,0	50,0	50,0	100,0
95+	0,0	77,8	11,1	11,1	100,0	0,0	33,3	0,0	66,7	100,0
nezj.	29,6	35,2	1,4	33,8	100,0	19,7	40,8	2,6	36,8	100,0
Celkem	38,6	56,4	2,1	3,0	100,0	37,6	49,4	8,0	4,9	100,0

U žen častěji než u mužů se nepodařilo rodinný stav určit. Mezi těmi, u nichž zůstal rodinný stav nezjištěn je mnoho těch, které možná byly svobodnými matkami nebo to byly ženy ovdovělé s jedním dítětem. Bez výslovného záznamu v evidenci to však nebylo možné zjistit. Na celém Chrudimsku jich bylo 1 176, tj. téměř 5 %.

3. Závěr

Věková struktura svobodných byla poměrně jednoduchá, jejich podíl se od 20 let rychle snižoval až do věku 40 let. Starší osoby, u nichž bylo zřetelně uvedeno, že jsou svobodné, byly velkou výjimkou. Ženy vstupovaly do manželství dříve než muži, mezi 15–19letými jich bylo ženato pouze 2,6 %, mezi ženami už 17,5 %. Muži častěji než ženy uzavírali po ovdovění další sňatky, a tak je podíl ženatých vysoký i v pozdějším věku. U žen ve vyšších věkových skupinách převažoval vdovský stav.

4. Literatura

[1] HORSKÝ, J.– SELIGOVÁ, M. 1997. Rodina našich předků, s. 60–65.

[2] TRETERA, J. R. – Církevní právo. 1992.

[3] Soupis poddaných podle víry – Chrudimsko. 2002.

Adresa autora

Eva Kačerová, RNDr.
VŠE Praha
nám. W. Churchilla 4, 130 67 Praha
kacerova@vse.cz

Analýza korelace hospodářského cyklu České republiky a Eurozóny - růstové pojetí
Correlation analysis of the business cycle in the Czech republic and Eurozone – growth cycle approach

Svatopluk Kapounek, Jitka Poměnková

Abstract: One of the main determinants of the success of the Eurozone is whether the commond monetary policy set by the European central bank is suitable fo all member countries. The theory of optimal currency areas argued that if countries are prone to asymmetric shocks, the costs related with the autonomous monetary policy loss will be minimized.

Key words: asymmetric shock, business cycle synchronisation, Hodrick-Presott filter

Klíčová slova: asymetrický šok, sladěnost hospodářského cyklu, Hodrick-Prescott filter

1. Úvod

Jeden z hlavních předpokladů úspěšného fungování Evropské měnové unie je efektivita společné monetární politiky ve smyslu její stabilizační funkce. V situaci, kdy je společná monetární politika Evropské centrální banky (ECB) aplikována na všechny členské státy Eurozóny současně, lze o její úspěšnosti ve smyslu stabilizace cenové hladiny hovořit pouze za předpokladu nízké pravděpodobnosti výskytu asymetrických šoků. Pravděpodobnost výskytu asymetrických šoků je přitom nepřímo úměrná vzájemné sladěnosti hospodářských cyklů jednotlivých zemí. Podmínkou sladěnosti hospodářských cyklů ve spojení s úspěšností fungování měnových unií se zabývá teorie optimálních měnových oblastí (Optimum Currency Area Theory – OCA Theory), vyplývající z původních myšlenek prof. Mundella v 60. letech 20. století. (De Grauwe, 2005)

Pro účely tohoto článku je hospodářským cyklem myšleno „více či méně pravidelně se opakující fáze expanze a kontrakce v ekonomické aktivitě kolem dlouhodobého růstového trendu." (Dornbusch, 1984, pp. 9, přeloženo autorem) Jednotlivé fáze hospodářského cyklu a střídající se vrcholy a dna, kdy je ekonomická aktivita vyšší či nižší než je dlouhodobý trend, jsou úzce provázány s krátkodobými výkyvy agregátní nabídky a poptávky s následným vlivem na míru inflace a nezaměstnanosti.

V historických přehledech lze nalézt mnoho logických zdůvodnění vzniku kolísání ekonomické aktivity. Od přírodních vlivů způsobujících neúrodu, přes nesoulad výroby organických a neorganických produktů či politické vlivy hledají současní ekonomové zdroje hospodářských cyklů v ekonomické oblasti. První náznaky hospodářských cyklů popisuje Sherman a Kolk (1996) v 18. století v Anglii. Jejich existenci tak lze spojit s určitou úrovní ekonomické vyspělosti a je možné je chápat jako součást ekonomického vývoje. Základním předpokladem pro existenci cyklického kolísání jsou přitom inovace. Tyto mají charakter strukturálních zlomů vyvolávající investiční aktivitu. V krátkém období jdou přitom investiční aktivity nad reálné možnosti dané ekonomiky, v období dlouhém se pak vrací do rovnovážného stavu. Opakující se výskyt inovací zapříčiňuje hospodářský cyklus, který se však v reálné ekonomice vyznačuje nepravidelností jak v čase, tedy frekvenci střídání jednotlivých fází, tak míře kolísání. (Schumpeter, 1939)

Podle Keynese je hlavním zdrojem cyklického kolísání agregátní poptávka a indukované investice. Tyto jsou funkcí míry reálné ekonomické aktivity, která poroste v případě kladných nenulových indukovaných investic. Akcelerované tempo růstu ekonomické aktivity se však zabrzdí ve chvíli, kdy ekonomika dosáhne hranice svých produkčních možností. Poté dochází

k poklesu investic, který je následně zastaven nutností nahrazení opotřebeného kapitálu (fyzického). Monetaristé naopak vyzdvihují negativní efekt poptávkově orientované hospodářské politiky státu v souvislosti s možností prohloubení hospodářských výkyvů. Hlavním zdrojem cyklického kolísání ekonomické aktivity jsou dle monetaristů externí šoky, jejichž důsledky je možné regulovat prostřednictvím monetární politiky. Nová klasická ekonomie doplňuje uvedené přístupy o prvek racionálního očekávání, kdy je reálný produkt a tedy ekonomická výkonnost ovlivněná neočekávanými změnami v hospodářské politice. Vliv exogenních šoků na ekonomickou aktivitu vyzdvihuje dále teorie reálných hospodářských cyklů. Lze ji označit za stochastickou verzi Walrasiánských principů dokonalé rovnováhy trhů. Předpokladem je dokonalá informovanost ekonomických subjektů a identifikace pouze reálného kolísání ekonomické aktivity, nikoliv nominálního. Nedokonalostmi na trzích, jako zdrojem kolísání ekonomické aktivity, především rigiditou cen a mezd se zabývá nová keynesovská ekonomie. (Holman, 2001)

V souladu s předpoklady Solowova modelu dlouhodobého ekonomického růstu osciluje ekonomická aktivita kolem stabilního (stálého) stavu daného rovnovážným ekonomickým růstem, kdy tempo růstu kapitálu a tempo růstu potenciálního produktu je rovno tempu růstu počtu obyvatel. Nechť je dlouhodobý trend identifikovaný v ekonomickém růstu stabilním (stálým) stavem, pak hospodářský cyklus tvoří krátkodobé výkyvy, které jsou předmětem stabilizační funkce společné měnové politiky v eurozóně.

Autoři tohoto příspěvku používají pro identifikaci trendu vývoje ekonomického růstu pro odvození hospodářského cyklu Hodrick-Prescottův filtr (HP filtr). Tato metodika je oblíbená především pro svou jednoduchost a snadnou aplikovatelnost. Je propojením statistické metodiky a ekonomického přístupu k modelování hospodářského cyklu. Hodrick a Prescott (1980) při konstrukci HP filtru vycházeli z empirických údajů ekonomické aktivity U.S. v letech 1947-1973.

Cílem příspěvku je identifikovat vzájemnou sladěnost cylklického kolísání v ekonomické aktivitě mezi Eurozónou a Českou republikou a nepřímo se tak vyjádřit k efektivitě fungování Evropské měnové unie jako optimální měnové oblasti v případě přijetí společné měny euro ze strany České republiky.

2. Metodika

Nechť Y_t označuje makroekonomickou časovou řadu tvořící absolutní hodnoty HDP. Vzhledem k vysokým hodnotám ekonomického růstu v posttransformačních ekonomikách bude následující empirická analýza zaměřena na identifikaci růstového cyklu, definovaného prostřednictvím prvních diferencí Y_t. Základní myšlenka modelování ekonomického růstu prosřednictvím HP filtru spočívá v rozkladu Y_t na trendovou složku g_t (růstovou komponentu) a stacionární reziduální složku c_t (cyklickou komponentu)

$$Y_t = g_t + c_t,\ t=1,...,T, \qquad (1)$$

kde g_t a c_t jsou nepozorované. HP filtr je přitom navržen tak, aby při extrakci trendu bral v úvahu dvě kritéria, a to velikost reziduí a míru vyhlazení trendu. Omezení růstové komponenty plyne z řešení následujícího problému, a to

$$\min_{\{g_t\}_{t=1}^{T}} \sum_{t=1}^{T}(Y_t - g_t)^2 + \lambda \sum_{t=1}^{T}\left[(g_{t+1} - g_t) - (g_t - g_{t-1})\right]^2, \qquad (2)$$

kde cyklická komponenta $c_t = Y_t - g_t$ představuje odchylky od dlouhodobého trendu a hladkost cyklické komponenty je měřena prostřednictvím kvadrátu druhých diferencí. Parametr λ určuje průběh odhadnuté trendové komponenty. Je-li $\lambda=0$, pak řešení výše uvedené minimalizační úlohy vede k tomu, že trendová složka je shodná s původní řadou. Pokud naopak $\lambda \to \infty$, je kladena větší váha na druhý člen a g_t se přibližuje k lineárnímu trendu. Lze

ukázat, že řešení minimalizační úlohy, a tedy hledání odhadu růstové komponenty, můžeme zapsat maticově ve formě

$$Ag=Y,$$

kde $Y=[Y_1,...Y_T]'$, $g=[g_1,...g_T]'$ a pro matici A platí $A=\lambda F+I$, kde

$$F=\begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & & & & & & \cdots & 0 \\ -2 & 5 & -4 & 1 & 0 & \cdots & & & & & \cdots & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 & \cdots & & & & \cdots & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & & & & & \vdots \\ \vdots & \ddots & & & & & & \ddots & & & & \\ & & & & & & & \ddots & & & \ddots & \vdots \\ & & & & & & & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & & & & & & & \cdots & 0 & 1 & -4 & 6 & -4 & 1 \\ \vdots & & & & & & & & \cdots & 0 & 1 & -4 & 5 & -2 \\ 0 & \cdots & & & & & & & & \cdots & 0 & 1 & -2 & 1 \end{bmatrix}$$

Pak trendová a cyklická komponenta může být identifikovaná jako

$$\hat{g} = A^{-1}Y,\ c = Y - \hat{g}\,. \tag{3}$$

Jak ukázal Harvey a Jager (1993) v nekonečné verzi lze HP filtr interpretovat jako optimální lineární filtr trendové komponenty. King a Rebelo (1993) uvádějí, že komponenta hospodářského cyklu a druhá diference růstové komponenty musí mít stejnou reprezentaci klouzavých průměrů pro HP filtr, aby byl lineárním filtrem, který minimalizuje střední kvadratickou chybu. Tedy, aby tento filtr minimalizoval chybu

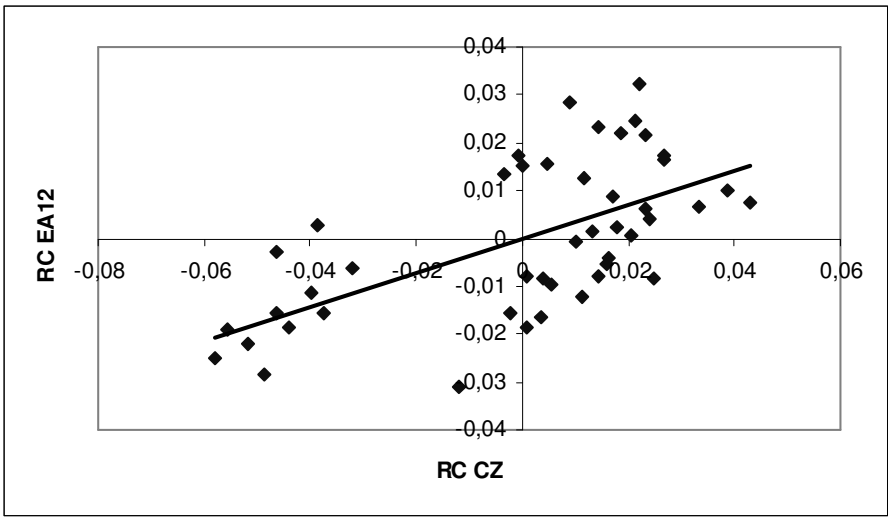
$$MSE = (1/T)\sum_{t=1}^{T}\left(ets(c_t) - c_t\right)^2\,, \tag{4}$$

kde c_t je skutečná cyklická komponenta a (c_t) je její odhad. Hodrick and Prescot zjistili, že jestliže jsou cyklická komponenta (c_t) a druhá diference růstové komponenty $(\Delta^2 g_t)$ identicky nezávisle normálně rozložené typu $c_t \sim N(0,\sigma_c^2)$, $\Delta^2 g_t \sim N(0,\sigma_g^2)$, pak nejlepší výběr vyhlazovacího parametr λ, ve smyslu minimalizace MSE, je $\lambda = \sigma_c^2 / \sigma_g^2$. V mnoha studiích je doporučovanou hodnotou λ pro čtvrtletní periodicitu dat *λ=1600*, například Ahumada, Garegnani (1990), Guy (1997), Hodrick-Prescott (1981) a další.

Pro posouzení vzájemné sladěnosti cyklů lze využít několik metod. Mezi široce užívané patří například korelační analýza (Wooldridge, 2003), OCA index, index cyklické shody nebo VAR a impulse-response analýza, SVAR. Aplikace těchto metod byla prezentována např. Darvas, Szapáry (2004), Artis at all (2004). Vzhledem k faktu, že česká ekonomika posttransformační, jeví se některé metody nevhodné, tj. např. kladou požadavek na znalost bodů zlomů, na rozsah datového souboru, na počet cyklů v ekonomice apod. Pak, z důvodů "menších" požadavků, se jeví korelační analýza jako vhodná úvodní analýza při měření sladěnosti hospodářských cyklů.

3. Empirická analýza

Pro emipirickou analýzu byly zvoleny čtvrtletní absolutní hodnoty HDP ve stálých cenách roku 2000, sezónně očištěné v období 1998/Q1 – 2008/Q2 pro Českou republiku a Eurozónu (EA12). S cílem linearizace časové řady byla data transformována přirozeným logaritmem. Pro získání hodnot růstového pojetí hospodářského cyklu (dále označováno RC) bylo provedeno detrendování logaritmovaných hodnot vstupní časové řady pomocí diferencí prvního řádu. Obrázek 1 prezentuje vztah meziročního růstu HDP mezi Českou republikou a Eurozónou.



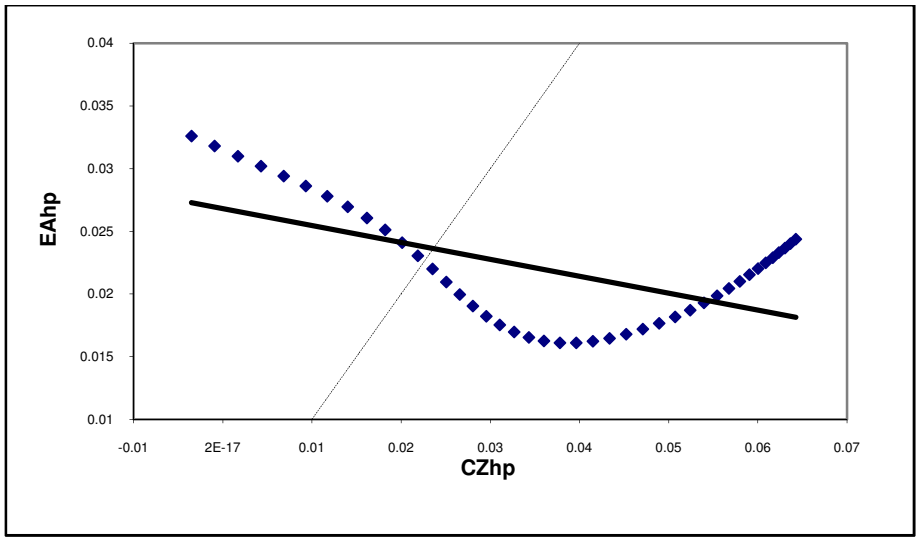
Obrázek 1: Meziroční růst HDP mezi EA a ČR
(body – hodnoty RC, plná čára – regresní přímka)

Pro posouzení sladěnosti cyklického kolísání je nutné identifikovat trend vývoje růstového cyklu prostřednictvím HP filtru. Sladěnost hospodářských cyklů byla měřena pomocí korelačního koeficientu. Poznamenejme, že v grafech budeme pod označením CZhp, resp. RC CZ, rozumět trend RC hodnot ČR, resp. růstové hodnoty ČR. Analogicky EAhp, resp. RC EA, rozumět trend RC hodnot EA12, resp. růstové hodnoty EA12. Zjištěné hodnoty uvádí tabulka 1.

Tabulka 1: Korelační koeficienty pro růstové pojetí hospodářského cyklu EA a ČR

EA versus	*CZ – HP trend*	*CZ – RC hodnoty*
korelace	*-0,5793*	*0,6230*
p-value	*0,0001*	0,0000

Statisticky vysoce významná korelace na 1% hladině významnosti byla prokázána jak v případě vlastního růstu HDP, tedy růstového cyklu (RC) tak i jeho trendu. Zatímco hodnoty RC vykazují pozitivní vztah mezi Českou republikou a Eurozónou (Obrázek 1 a Tabulka 1), vztah trendu vývoje je negativní.



Obrázek 2: Závislost trendu růstového pojetí hospodářského cyklu EA a ČR (body – hodnoty trendů, tečkovaná čára – osa prvního kvadrantu)

Bližší analýzu vzájemného vztahu mezi trendy RC v České republice a Eurozóně poskytuje Obrázek 2. Symetrické cyklické kolísání ekonomické aktivity, kdy lze společnou monetární politiku ECB označit za efektivní i pro Českou republiku a tedy náklady spojené se ztrátou autonomie monetární politiky ČNB za nevýznamné, v grafu prezentuje osa prvního kvadrantu. Asymetrické kolísání trendu RC mezi Českou republikou a Eurozónou je přitom z obrázku 2 patrné pouze u méně výrazných výkyvů, v případě České republiky do hodnoty 0,04. Výraznější cyklické výkyvy od dlouhodobého trendu ekonomické aktivity mají tendenci potvrdit symetrický charakter.

4. Závěr

Cílem předkládaného příspěvku byla identifikace sladěnosti cyklického kolísání ekonomické aktivity mezi Českou republikou a Eurozónou ve vztahu k možné symetričnosti či asymetričnosti krátkodobých výkyvů od dlouhodobě stabilního, stálého stavu.

Jak bylo nastíněno v úvodu, autoři předpokládají, že dlouhodobý trend ekonomické aktivity souvisí se zásobou kapitálu v ekonomice a jeho produktivitou. Cyklické výkyvy v ekonomické aktivitě jsou pak krátkodobými odchylkami od stálého, stabilního stavu ekonomiky. Cílem společné monetární politiky v Eurozóně je pak stabilizovat ekonomickou aktivitu na úrovni již zmíněného dlouhodobě rovnovážného stavu, který pro jednoduchost označme potenciálním produktem. Efektivnost její stabilizační funkce je pak významně ovlivněna výskytem asymetrických makroekonomických šoků. V případě očekávaného a plánovaného přijetí společné měny euro ze strany České republiky se jako nejvýznamnější asymetrický šok jeví nárůst inflace (Lacina, L. a kol., 2008).

Empirická analýza sice prokazuje určitou míru asymetričnosti výkyvů ekonomické aktivity kolem svého potenciálu, ale v případech výrazného vychýlení je patrná tendence k symetričnosti. Autoři se proto domnívají, že nebezpečí dopadu asymetrických šoků na českou ekonomiku po zavedení eura a s tím spojený nárůst nákladů souvisejících se ztrátou autonomie monetární politiky ČNB se může projevovat pouze v mírném cyklickém kolísání způsobeném ve výkyvech v agregátní poptávce a nabídce. V případě, že by česká ekonomika byla zasažena výraznějším asymetrickým šokem s následnou tendencí k vyšší inflaci, lze stejný vývoj očekávat i v Eurozóně. Těžiště odůvodnění uvedeného závěru leží ve výrazné otevřenosti české ekonomiky a jejího provázání s Eurozónou prostřednictvím zahraničně obchodních vztahů.

Předkládaný příspěvek vznikl za podpory výzkumného záměru „Česká ekonomika v procesech integrace a globalizace a vývoj agrárního sektoru a sektoru služeb v nových podmínkách evropského integrovaného trhu“.

5. Literatura

[1] AHMUDA, H., GAREGNANI, M. L., 1999: Hodrick-Prescott Filter in practice, UNLP;

[2] ARAUJO, F., AREOSA, M. B. M., NETO, J. A. R. 2003 r-filters: a Hodrick-Prescott Filter Generalization Working paper Series 69, Banco Central do Brasil, 37 pp., ISSN 1518-3548

[3] ARTIS, M.J., MARCELLINO, M.G., PROIETTI, T., 2004: Dating the Euro Area Business Cycle: A Methodological Contribution with an Application to the Euro Area, *Oxford Bulletin of Economics and Statistics*, 66, pp. 537 – 565.

[4] DARVAS, Z., SZAPÁRY, G. (2007): Business Cycle Synchronization in the Enlarged EU. *Open Economies Review*, Springer Netherlands, pp. 1-19, ISSN 0923-7992

[5]DE GRAUWE P. (2005): Economics of Monetary Union. Sixth Edition. Oxford: OUP. 2005. ISBN: 0-19-927700-1;

[6]DORNBUSCH, F. (1984): Macroeconomics. Third Edition. New York, McGraw-Hill Book company, 1984, pp. 5-18, ISBN: 0-07-017770-8;

[7] GUAY,A. ST-AMANT, P. Do the Hodrick-Prescott and Baxter-king Filters Provide a Good Approximation of Business Cycles? Université a Québec á MOntreál, Working paper No. 53, 1997

[8] HARVEY, A.C., JAEGER, A. (1993): Detrending, Stylized FActs and the Business Cycle. *Journal of Applied Econometrics 8*: pp. 231-47

[9] HODRICK, R.J., PRESCOTT, E.C. Post-war U.S. Business Cycles: An Empirical Investigation, mimeo, Carnegie-Mellou University, Pitsburgh, PA. 1980, 24pp.

[10] HOLMAN, R. a kol. Dějiny ekonomického myšlení. 2. vydání. Praha, C.H.Beck, 2001, pp.544, ISBN: 80-7179-631-X;

[11] KING, R.G., REBELO, S.T., (1993): Low frequency filtering and real busines cycles. *Journal of economic dynamics and Control 17*. pp. 207-232.

[12] LACINA, L. a kol. (2008): Studie vlivu zavedení eura na ekonomiku ČR. 1. vyd. Bučovice: Ministerstvo financí ČR a nakladatelství Martin Stříž, 2008. 295 s. ISBN 978-80-87106-08-2;

[13] SCHUMPETER, J.A. Business Cycles: A Theoretical Historical and Statistical Analysis of The Capitalist Process. New York, McGraw-Hill Book Company, 1939, pp.461, ISBN 1578985560;

[14] SHERMAN, H.J., Kolk,D.X. Business Cyccles and Forecasting. New York, Addison Wesley Publishing Company, 1995, ISBN: 0-06-501139-5;

[15] WOOLDRIDGE, J. M. 2006. *Introductory Econometrics: A modern approach*. South-Western Cengage Learning, 865 p., ISBN-10: 0-324-66054-5.

Adresa autora:

Jitka Poměnková, Ph.D., RNDr.
Ústav financí PEF MZLU v Brně
Zemědělská l
Česká republika
613 00 Brno
pomenka@mendelu.cz

Svatopluk Kapounek, Ph.D., ing.
Ústav financí PEF MZLU v Brně
Zemědělská l
Česká republika
613 00 Brno
skapounek@mendelu.cz

Rozsah výroby Cobbovej-Douglasovej produkčnej funkcie stavebných podnikov pre rôzne chyby v premenných
Production scale of Cobb-Douglas production function of construction companies for various errors in variables

Samuel Korónyi

Abstract: Griliches approach to consistent bounds of Cobb-Douglas production function parameters can offer possible values of production scale values with errors in variables assumption. Then larger production scale values correspond to larger errors in variables. Examined Cobb-Douglas production function was estimated from absolute economic production indicators of Slovak construction joint stock and ltd. companies from the year 2001 (value added, sum of assets and average number of employees).

Keywords: Regression analysis, Errors in variables, Production functions, Construction sector

Kľúčové slová: Regresná analýza, Chyby v premenných, Produkčné funkcie, Stavebníctvo

1. Úvod

Príspevok voľne nadväzuje na Grilichesove ohraničenia regresných parametrov (Griliches 1971) aplikované na Cobbovu-Douglasovu produkčnú funkciu (ďalej CDPF) slovenských stavebných podnikov (Korónyi 2008). Grilichesov postup pre zistenie konzistentných ohraničení regresných parametrov vychádza z predpokladu znalosti podielu rozptylu chybovej zložky voči príslušnej premennej (Maddala 1992).

2. Dáta

Pre zistenie rozsahu výroby CDPF boli použité vybrané ukazovatele - pridaná hodnota (v mil. Sk) ako výstup, priemerný evidenčný počet zamestnancov vo fyzických osobách a celkový kapitál (v mil. Sk) ako vstupy z reálnych údajov súkromných slovenských stavebných podnikov právnej formy a. s. a s. r. o. z výkazu ŠÚ SR Prod 3-04 za rok 2001. Regresná analýza dát bola urobená v štatistickom systéme SPSS verzia 13.

3. Grilichesove konzistentné hodnoty rozsahu výroby - teória

V krátkosti zopakujeme základné kroky Grilichesovho postupu pri uvažovaní o chybách v premenných. Majme lineárny regresný model s dvomi nezávislými premennými v normalizovanom tvare

$$y = \beta_1 x_1 + \beta_2 x_2 + e,$$

v ktorom všetky premenné sú merané s chybou. Nech pozorované premenné sú

$$Y = y + v, X_1 = x_1 + u_1, X_2 = x_2 + u_2.$$

Za navzájom nekorelované považujeme chybové zložky u_1, u_2, v a tiež nekorelované s premennými x_1, x_2, y. Chybové zložky e a v zahrnieme do jednej zloženej chyby

v premennej Y, preto $\sigma^2 = \text{var}(e+v)$. Nech $\text{cov}(X_1, X_2) = \rho$. Rozptyly relatívnych chýb nezávislých premenných označíme

$$\lambda_1 = \frac{\text{var}(u_1)}{\text{var}(X_1)} = \text{var}(u_1),\ \lambda_2 = \frac{\text{var}(u_2)}{\text{var}(X_2)} = \text{var}(u_2).$$

Po úpravách dostaneme pravdepodobnostné limity $\hat{\beta}_1, \hat{\beta}_2$ pre skutočné regresné parametre β_1, β_2 získané metódou najmenších štvorcov

$$\text{plim}\left(\hat{\beta}_1 - \beta_1\right) = -\frac{\beta_1\lambda_1 - \rho\beta_2\lambda_2}{1-\rho^2},$$

$$\text{plim}\left(\hat{\beta}_2 - \beta_2\right) = -\frac{\beta_2\lambda_2 - \rho\beta_1\lambda_1}{1-\rho^2}. \qquad (1)$$

Zo vzťahu (1) vyplýva, že na základe predpokladaných hodnôt (intervalov) pre parametre relatívnej chyby λ_1, λ_2 je možné dostať im odpovedajúce hodnoty (intervaly) konzistentných odhadov β_1, β_2. Vychýlenie odhadu je navyše zväčšené faktorom $1/\left(1-\rho^2\right)$, pričom pre väčšie ρ, je väčšie aj vychýlenie. Multikolinearita zhoršuje uvedené odhady.

Sčítaním obidvoch častí vzťahu (1) dostaneme vyjadrenie odchýlky súčtu regresných parametrov lineárneho regresného modelu ako funkciu relatívnych chýb oboch premenných a ich vzájomnej korelácie (Griliches 1986)

$$\text{plim}\left[\left(\left(\hat{\beta}_1 + \hat{\beta}_2\right) - \left(\beta_1 + \beta_2\right)\right)\right] = -\frac{\beta_1\lambda_1 + \beta_2\lambda_2}{1+\rho}. \qquad (2)$$

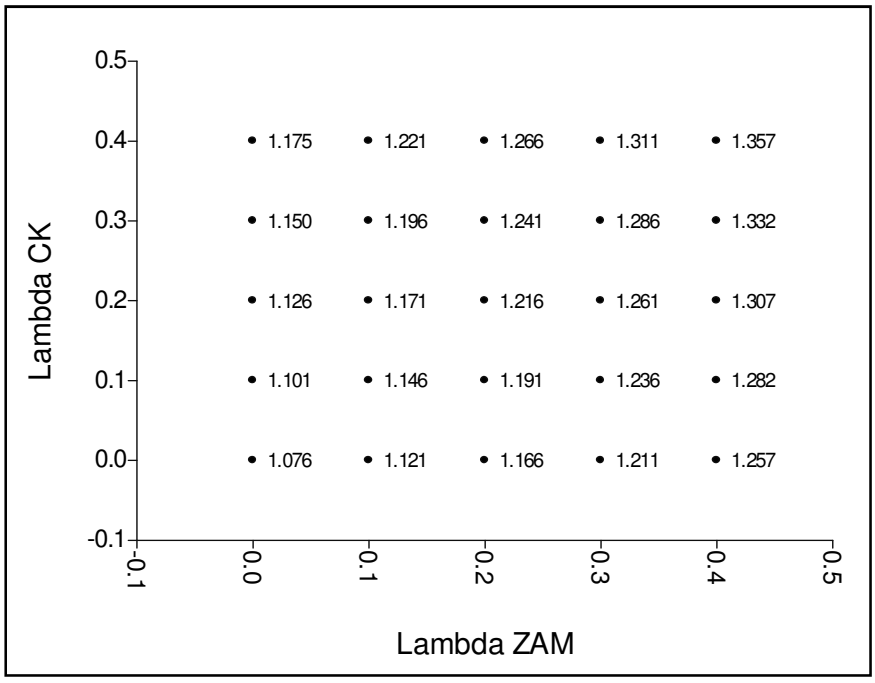
Súčet regresných parametrov β_1, β_2 je z ekonomického hľadiska rozsah produkcie CDPF. Na základe vzťahu (2) je zrejmé, že skutočný súčet odhadovaných parametrov je vždy vychýlený smerom k nule pri nenulových chybách nezávislých premenných. V tomto prípade je vychýlenie klesajúcou funkciou ρ. Čím je korelácia medzi nezávislými premennými väčšia, tým je menší vplyv nezávislých chýb v nich.

4. Výsledky pre Grilichesove konzistentné hodnoty rozsahu výroby CDPF

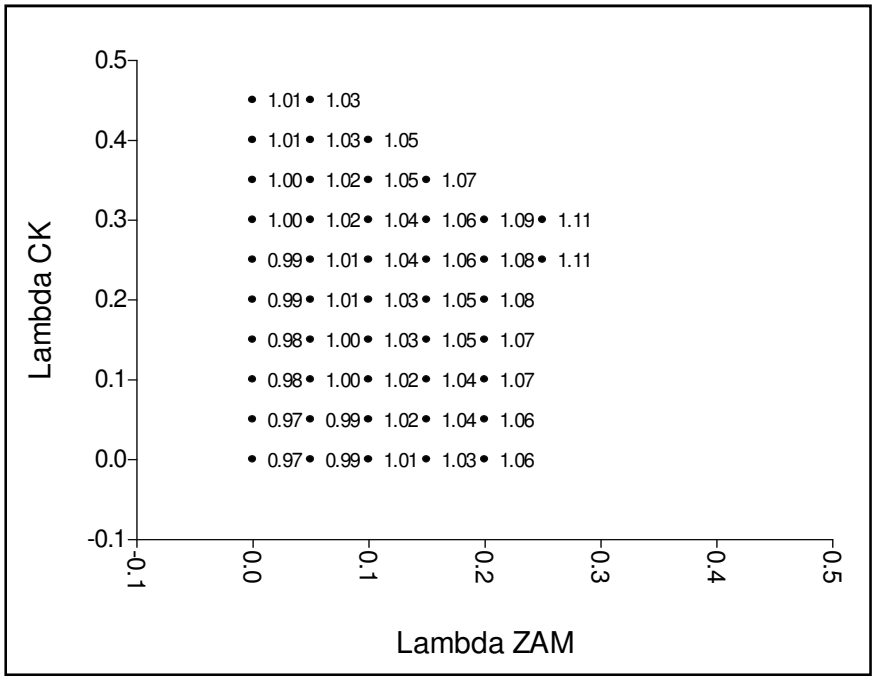
Na základe Grilichesovho postupu je teda možné zistiť hodnotu súčtu regresných parametrov CDPF stavebných podnikov (efekt rozsahu výroby) pre rôzne chyby v premenných. Na grafoch 1 a 2 je zobrazená množina bodov s hodnotami, ktoré sa rovnajú súčtu konzistentných hodnôt účinností pracovnej sily a celkového kapitálu pre vybrané relatívne chyby produkčných faktorov priemerného evidenčného počtu zamestnancov a celkového úhrnu aktív CDPF. Zreteľne vidieť, že efekt rozsahu výroby je stúpajúcou funkciou relatívnych chýb vstupov obidvoch vstupných produkčných faktorov. Chyby prítomné vo vstupoch potom okrem vychýlenia regresných parametrov produkčnej funkcie zvyšujú aj efekt jej rozsahu výroby.

Hodnoty rozsahu produkcie pre stavebné podniky právnej formy s. r. o. sú na grafe 1 zobrazené s krokom relatívnej chyby v nezávislých premenných 0,1. Na grafe 2 je krok 0,05 (stavebné podniky právnej formy a. s.). Súčet parametrov účinnosti CDPF je 1,076 pri nulovej chybe oboch vstupných premenných pre stavebné podniky s. r. o. (graf 1). Hodnota 1,121 odpovedá chybe 0,1 pre vstup priemerného evidenčného počtu zamestnancov vo fyzických osobách a nulovej chybe v hodnote celkového kapitálu. Pri veľkosti relatívnych chýb 0,4 je súčet parametrov - rozsah produkcie CDPF rovný 1,357.

Pre podniky formy a. s. je situácia komplikovanejšia, lebo väčšina hodnôt parametrov vychádza záporná (porušenie základných ekonomických predpokladov) a navyše sú obe premenné viac korelované (koeficient korelácie 0,722). Preto má aj oblasť konzistencie nepravidelný tvar.



Graf 1 Hodnoty rozsahu výroby pre rôzne chyby vstupov CDPF podnikov s. r. o.



Graf 2 Hodnoty rozsahu výroby pre rôzne chyby vstupov CDPF podnikov a. s.

5. Záver

Grilichesov postup pre zistenie konzistentných hodnôt rozsahu produkcie CDPF je vhodný vtedy, ak máme aspoň určitú predstavu o veľkosti chýb v nezávislých premenných. Pritom pre väčšie chyby v premenných je väčší aj skutočný rozsah produkcie. Jeho ďalšou výhodou je názorná geometrická interpretácia, ktorá sa samozrejme stráca pri použití viac ako dvoch nezávislých premenných.

6. Literatúra

[1] GRILICHES, Z. – RINGSTAD, V. Economies of Scale and the Form of the Production Function. Amstredam : North-Holland, 1971. ISBN 9780720431728

[2] KORÓNY, S. Grilichesove konzistentné hranice pre parametre Cobbovej-Douglasovej produkčnej funkcie stavebných podnikov. In: Forum Statisticum Slovacum. Roč.IV., č.6, 2008, s. 55-66. ISSN 1336-7420

[3] MADDALA, G. S. Introduction to Econometrics. London: Prentice-Hall, 1992. ISBN 0-13-880352-8

[4] GRILICHES, Z. Economic Data Issues. In: Handbook of Econometrics, Volume III, Ed. Griliches Z. and Intriligator M.D. Elsevier, 1986. ISBN 0-13-880352-8

Adresa autora:

RNDr. Samuel Koróny
Ústav vedy a výskumu UMB
Cesta na amfiteáter 1
974 01 Banská Bystrica
Email: samuel.korony@umb.sk

On-line slovník
On-line version of dictionary

Löster Tomáš

Abstract: In this text, I would like to present our on-line statistical, economic and demographic dictionary, which is being developed in terms of project IGA University of Economics, Prague, No. 25/2008. On-line version of the dictionary is available at http://slovnik.vse.cz. The dictionary contains Czech, English, German, French and Polish language. We would like to add more languages in the future and more of sciences, for example mathematics.

Key words: electronical dictionary, statistical terms, economical terms, demographical terms.

Klíčová slova: elektronický slovník, statistické pojmy, ekonomické pojmy, demografické pojmy.

1. Úvod

S masivním rozvojem elektronické verze knih, článků a dalších materiálů, ve spojitosti s rozvojem internetu, je stále vyšší poptávka po elektronických odborných slovnících. Tlak na vznik nových odborných slovníků v elektronické podobě se projevuje také do požadavků na jejich funkčnost a schopnost. Uživatel chce nejen samotný překladový slovník, který je schopen přeložit pojmy mezi libovolnou dvojicí zvolených jazyků. Uživatel dále požaduje, aby byl slovník schopen poskytnout stručný výklad jednotlivých pojmů v různých jazykových kombinacích. Kromě využití schopností existujících slovníků by se uživatel mnohdy rád se svými znalostmi a zkušenostmi podílel na zlepšení existujících slovníků. Cílem tohoto příspěvku je stručně seznámit s on-line verzí statisticko-ekonomicko-demografického slovníku, který je jako víceletý grant IGA VŠE v Praze č. 25/2008 projektován a rozšiřován nejen o nové pojmy, vědní obory, ale také funkčnosti. Podrobný popis současných funkcí a možností slovníku je uveden např. v [7], a proto jejich popisu bude věnována pouze krátká poznámka v následující kapitole.

2. Popis funkcí a obsahu slovníku

Slovník je koncipován jako otevřený a neustále se rozvíjející bez stanoveného limitu a konce. V současné době je připravena ke spuštění databáze vědního oboru demografie, která bude během počátku roku 2009 spuštěna v první verzi. V rámci řešení současného projektu IGA VŠE v Praze č. 25/2008 je připravována i první verze výkladové části některých pojmů statistického slovníku. Případná spolupráce s externími spolupracovníky nejen z České republiky, ale i ze Slovenska může pomoci k vytvoření velmi přínosného projektu nejen pro akademickou půdu. Slovník může být využíván také širokou veřejností, což je díky umístění slovníku na internetu jeho zásadní přínos. Současný projekt navazuje na dva úspěšně ukončené projekty a to 15/06 a 24/07, které byly řešeny v roce 2006 a 2007. Výchozí myšlenkou projektu 15/06 bylo pomocí programovacího jazyku Visual Basic .NET vytvořit nový, prozatím nikde neexistující software, který bude dostupný nejen studentům a zaměstnancům VŠE, ale všem, kdo mají přístup k internetu. Do nového softwaru, který byl vytvořen, byla postupně digitalizována část pojmů tištěných slovníků, které jsou uvedeny v seznamu referencí elektronické verze. Tyto pojmy byly postupně digitalizovány, kontrolovány a doplňovány. Během prvních let byla databáze statistických pojmů nejprve

naplňována digitalizovanými pojmy, v současné době obě funkční databáze obsahují cca 30 % nových pojmů a 70 % digitalizovaných pojmů ze starých tištěných slovníků. Celkový počet pojmů v říjnu 2008 je uveden v tabulkách č. 1 a 2.

Tabulka 1: Počty pojmů ve statistickém slovníku

Jazyk	Počet pojmů
Český	1 895
Anglický	1 895
Německý	1 894
Francouzský	1 579
Polský	1 253

Z tabulky 1 vyplývá, že v současné době je celkem 1253 pojmů, které je možné přeložit v libovolné kombinaci jazyků Český-Anglický-Německý-Francouzský-Polský. U ostatních výrazů byl v některých jazycích nejednoznačný překlad, a proto byly pojmy prozatím vynechány.

Tabulka 2: Počty pojmů v ekonomickém slovníku

Jazyk	Počet pojmů
Český	300
Anglický	300
Německý	300

Z tabulky 2 vyplývá, že všech tři sta pojmů je přeložitelné v libovolné kombinaci jazyků Český-Anglický-Německý.

Další pojmy a světové jazyky budou nejen ve statistické, ale také v ekonomické databázi postupně doplňovány.

Jak již bylo uvedeno výše, on-line slovník pojmů z oblasti statistiky, ekonomie a demografie (databáze bude spuštěna během roku 2009) je umístěn na webu http://slovnik.vse.cz. Slovník je volně přístupný všem uživatelům internetu. Jeho jedinečnost je tvořena následujícími znaky. Slovník:

- Umožňuje on-line překlad statistických, ekonomických a již brzy i demografických termínů a frází mezi všemi zahrnutými jazyky.
- Je vyvinut jako otevřený, tj. umožňuje přidávání nových a korekci existujících termínů ze strany jeho uživatelů.
- Umožňuje dva způsoby vyhledávání pojmů: full-textové vyhledávání, ale také vyhledávání pojmů pomocí rejstříků.
- Umožňuje zobrazení synonym (termínů se stejným výrazem) i homonym (dalších významů jednoho slova).
- U termínů, jejichž překlad je nejednoznačný, u archaismů a termínů s více možnými překlady je zobrazena značka.
- Libovolný uživatel může přidat návrh na opravu existujícího termínu nebo doplnění nového záznamu ve strukturované formě (tj. nikoliv jen jako nestrukturovaný komentář).

3. Závěr

On-line slovník různých vědních oborů, který je v současné době vytvářen jako několikaletý projekt Interní grantové agentury Vysoké školy ekonomické v Praze, reaguje na dlouhodobou absenci odborných slovníků na internetu, tj. v elektronické podobě. Jeho počátek byl reprezentován digitalizací tištěných forem různě zaměřených statistických slovníků. Následovalo rozšíření o další pojmy a novou ekonomickou databázi. Během roku 2009 bude spuštěna první verze demografické databáze a první verze výkladové části slovníku. Mezi současné střednědobé cíle řešitelů projektu patří, kromě průběžného doplňování databází novými pojmy, také další obor – matematika či ekonometrie. V současných databázích však chybí mj. Slovenský jazyk. Zařazením Slovenského jazyka by se samozřejmě využitelnost on-line slovníku zvýšila. Pokud by se kdokoliv rád podílel na jeho rozvoji, zlepšení či návrzích dalších odvětví a oborů, může kontaktovat řešitelský tým na níže uvedené adrese. Mezinárodní spolupráce je vítána a přispěje k zaplnění mezery v odborných slovnících na internetu a absenci Slovenského jazyka v tomto slovníku.

4. Literatura

[1] HEBÁK, P., HUSTOPECKÝ, J.: *Šestijazyčný slovník termínů ze statistické dynamiky,* VÚSEI, Praha 1980.

[2] HEBÁK, P., HUSTOPECKÝ, J.: *Šestijazyčný slovník termínů z regresní analýzy*, VÚSEI, Praha 1978.

[3] HEBÁK, P., HUSTOPECKÝ, J.: *Šestijazyčný slovník termínů z matematické statistiky*, VÚSEI, Praha 1979.

[4] HEBÁK, P., HUSTOPECKÝ, J.: *Šestijazyčný slovník termínů z teorie výběrových šetření*, VÚSEI, Praha 1981.

[5] *SQL Server 2000 Books Online, www.microsoft.com/downloads, Microsoft, 2004.*

[6] *Microsoft Developer Network (MSDN), msdn.microsoft.com, Microsoft.*

[7] LÖSTER, Tomáš, VOHLÍDAL, Jiří. *On-line verze statisticko-ekonomického slovníku*. Forum Statisticum Slovacum. Bratislava: ISSN 1336-7420

Adresa autora:

Tomáš Löster, Ing.
Vysoká škola ekonomická v Praze
Katedra statistiky a pravděpodobnosti
nám. W. Churchilla 4
130 67 Praha 3
losterto@vse.cz

Metodologické aspekty zberu a záznamu dát otázok s možnosťou viac odpovedí
Methodological aspects collection and entering data for multiresponse questions

Ján Luha

Abstract: This article deals with methodological aspects collection and entering data for multiresponse questions.

Key words: Multiresponse questions, data collection, data entering, methodological aspects.

Kľúčové slová: Otázky s možnosťou viac odpovedí, zber dát, záznam dát, metodologické aspekty.

1. Úvod

Príspevok je venovaný špecifickým otázkam zberu a záznamu otázok s možnosťami viacerých odpovedí (multiresponse questions), v texte ich budeme ďalej označovať ako multiresponse otázky. Takéto otázky sa často vyskytujú v empirických prieskumoch. Keďže nie sú jednoznačnou transformáciou, nespĺňajú definíciu štatistického znaku a vyžadujú preto špecifický prístup pri zbere, zázname a pri ich štatistickom spracovaní.

Multiresponse otázky môžu mať vopred stanovenú škálu odpovedí alebo môžu byť otvorené, teda respondent formuluje odpovede sám. V druhom prípade je potrebné pred záznamom dát vytvoriť kódovací kľúč, kde zhlukujeme určité „podobné“ typy odpovedí a priradíme im kód a názov, ktorý tieto odpovede reprezentuje. Dôležité je pred záznamom dát mať k dispozícii „finálnu“ škálu odpovedí (kódovací kľúč). V príspevku sa zaoberáme metodologickými otázkami a preto sa nezaoberáme technikami záznamu dát do PC.

2. Otázka s možnosťou viacerých odpovedí (multiresponse question)

Uvažujme otázku Z s množinou možných odpovedí {1, 2, ... , r} a možnosťou dať maximálne m odpovedí, kde: m <= r.

Z	**1**	**2**	**...**	**r**	**∑**
abs. počet. odpovedí	K_1	K_2	...	K_r	**K**
rel. počet. vzhľadom na počet respondentov	P_1	P_2	...	P_r	**P**
rel. počet. vzhľadom na počet odpovedí	p_1	p_2	...	p_r	**1**

Kde K_i je počet odpovedí s hodnotou i, n je počet odpovedajúcich respondentov, K je počet všetkých odpovedí a kde $P_i = K_i/n$, $p_i = K_i/K$.

Platí: $\sum_i K_i = K$, $\sum_i P_i = P$ a $\sum_i p_i = 1$, **pričom P >= 1** a **K <= m.n**.

Príklad: Otázka: Ktorý z politikov pôsobiacich na Slovensku má v súčasnosti vašu dôveru? Možno uviesť najviac tri odpovede. (Príklad je zo spojeného súboru štyroch výskumov ÚVVM, ktorý je využitý aj v práci [7].)

Case Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
$polit(a)	4347	97,3%	122	2,7%	4469	100,0%

a Group

$polit Frequencies

		Responses		Percent of Cases
		N	Percent	
$polit(a)	1 žiadnemu	1094	16,2%	25,2%
	2 M.Dzurinda	212	3,1%	4,9%
	3 I.Gašparovič	727	10,8%	16,7%
	4 R.Fico	1467	21,7%	33,7%
	5 V.Mečiar	341	5,0%	7,8%
	6 M.Flašíková-Beňová	58	,9%	1,3%
	7 J.Slota	469	6,9%	10,8%
	8 B.Bugár	235	3,5%	5,4%
	9 V.Tomanová	90	1,3%	2,1%
	10 D.Čaplovič	69	1,0%	1,6%
	11 J.Počiatek	97	1,4%	2,2%
	12 V.Palko	39	,6%	,9%
	13 F.Mikloško	56	,8%	1,3%
	14 P.Hrušovský	139	2,1%	3,2%
	15 D.Lipšic	94	1,4%	2,2%
	16 E.Kukan	27	,4%	,6%
	17 P.Paška	60	,9%	1,4%
	18 I.Mikloš	175	2,6%	4,0%
	19 A.Belousovová	72	1,1%	1,7%
	20 I.Radičová	349	5,2%	8,0%
	21 V.Veteška	18	,3%	,4%
	22 R.Kaliňák	420	6,2%	9,7%
	23 Z.Martináková	24	,4%	,6%
	24 P.Csáky	86	1,3%	2,0%
	25 iní	335	5,0%	7,7%
Total		6753	100,0%	155,3%

a Group

3. Záznam otázok s viacerými možnosťami odpovede

Kódovanie a záznam multiresponse môžeme realizovať dvomi spôsobmi:

Prvý spôsob: Okódujeme všetky možnosti odpovedí od 1 po r a na záznam vytvoríme toľko stĺpcov, koľko je maximálny počet dovolených odpovedí, napr. m. Kódy v rozmedzí od 1 po r zaznamenávame do m stĺpcov.

Príklad: Maximálny počet dovolených odpovedí:

Uvažujme ilustratívnu otázku kde počet odpovedí r=25 a maximálny počet možností odpovedať m=3. Prvý spôsob záznamu redukuje počet stĺpcov. Záznam, možno realizovať pomocou tabuľky:

porcis	otazka_a	otazka_b	otazka_c
1			
2			
3			
.			
.			

Chýbajúci záznam pre multiresponse otázku identifikujeme, ak v príslušnom riadku nie sú odpovede vo všetkých stĺpcoch.

Tento spôsob záznamu ***umožňuje*** zaznamenať v jednom riadku jeden kód viackrát, čo pri rekódovaní pomocou transformácií na systém jedna/nula spôsobuje vypadnutie „redundantných" hodnôt a tým aj odlišné výsledky v štatistických analýzach. V ilustratívnom príklade, ktorý sme uviedli v 2. kapitole, je kód odpovede 25=iní politici zaznamenaný potenciálne viackrát. Naozaj sa vyskytol v kombinácii prvého a druhého stĺpca 1 krát a 4 krát v kombinácii druhého a tretieho stĺpca. Tento problém sa priamo vo výsledkoch, ktoré produkuje SPSS neprejavuje, iba ak potrebujeme túto premennú rekódovať, prípadne transformovať na 25 indikátorových premenných, ktoré skúmajú elementárne javy vyjadrené v jednotlivých odpovediach v škále otázky.

Druhý spôsob: Na záznam multiresponse otázky pripravíme r, stĺpcov kde r je maximálne možný počet odpovedí a zaznamenáme kód 1 v každom stĺpci pre záznam príslušnej odpovede, ak si túto respondent vybral, a kód 0, keď si túto možnosť nevybral (takto sú definované indikátorové premenné). Aj v tomto prípade ***chýbajúci údaj*** znamená, že sme od daného respondenta nezískali na túto otázku žiadnu odpoveď. Môžeme použiť aj ***modifikáciu:*** v i-tom stĺpci pre i-tu odpoveď zaznamenáme kód i, ak respondent volil túto odpoveď a kód 0, keď nie. Pre štatistické vyhodnotenie je to ale menej výhodné ako pomocou indikátorových premenných.

Príklad: Maximálny počet odpovedí:

Druhý spôsob záznamu multiresponse otázok môže mať určitú výhodu, pretože zaznamenáva všetky skúmané elementárne javy, ale má menšiu nevýhodu vo viac stĺpcoch potrebných na záznam dát. Tento spôsob pôsobí určité problémy v prípade otvorených multiresponse otázok, ako napríklad o dôveryhodnosti politikov, pretože nepoznáme vopred maximálny počet odpovedí a kvôli potenciálnej potrebe použiť kód pre „zvyškové" odpovede typu „iní" a pod. Odpovede 1 alebo 0 „zapíšeme" do príslušného stĺpca tabuľky:

porcis	p_01	p_02	p_03	p_04	p_05	p_06	p_07	p_08	p_09		...		p_23	p_24	p_25
1															
2															
3															
.															
.															

Kvôli úspore miesta neuvádzame konkrétny výstup pre druhý spôsob záznamu multiresponse otázky. Výsledok je samozrejme rovnaký.

Poznámka: Ak z nejakých dôvodov, napríklad kvôli lepšej prehľadnosti výstupov, potrebujeme zmeniť kódovací kľúč otvorenej multiresponse otázky, musíme skúmať či neporušíme konzistenciu výstupov pri rôznych kódovacích kľúčoch. To môže nastať ak iným spôsobom preskupíme odpovede a teda inak definujeme „podobné" zhluky odpovedí. V takomto prípade nemôžeme vytvoriť nové premenné napríklad rekódovaním už zaznamenaných údajov, ale musíme ich znova zaznamenať, čo je pri väčšom rozsahu výberového súboru značne prácne.

4. Indikátorové premenné otázok s viacerými možnosťami odpovede

Pri transformáciách multiresponse otázok na indikátorové premenné ***(nadobúdajú hodnotu 1, ak respondent uviedol v odpovedi daného politika, inak 0)***, napríklad pri otázke o dôveryhodnosti politikov, musíme najprv identifikovať chýbajúce odpovede. Kódovanie chýbajúcich hodnôt musí byť realizované takým kódom, ktorý sa nemôže vyskytnúť ako „hodnota" odpovede na danú otázku. Uvažujme prvý spôsob záznamu multiresponse otázok

prezentovaný v 3. kapitole. Musíme rešpektovať kódovanie chýbajúcich hodnôt v tomto zázname, aby sme správne definovali chýbajúce hodnoty indikátorových premenných.

V ilustratívnom príklade sú v pôvodnom zázname ***chýbajúce hodnoty označené kódom 0***.

Príklad:

```
*Najprv zabezpečíme aby hodnoty 0 deklarované ako chýbajúce hodnoty sa pre účely transformácie zmenili na „reálne“ nuly.
Potom vytvoríme pre transformované premenné patričné chýbajúce hodnoty. Keďže pracujeme s dvojcifernými kódmi, označili sme chýbajúce hodnoty indikátorových premenných kódom 99.
Na záver definujeme nuly (označujú „neprítomnosť“ politika v odpovedi respondenta) a znovu označíme chýbajúce hodnoty ako systémové „missingy“.
*korektné definovanie chýbajúcich hodnôt pri multiresponse. Kvôli úspore miesta ilustrujeme syntax iba na niekoľkých politikoch.
MISSING VALUES polit_a TO polit_c ().
IF (polit_a = 0 & polit_b = 0 & polit_c = 0) ziadny = 99 .
IF (polit_a = 0 & polit_b = 0 & polit_c = 0) dzurinda = 99 .
IF (polit_a = 0 & polit_b = 0 & polit_c = 0) gasparovic = 99 .
IF (polit_a = 0 & polit_b = 0 & polit_c = 0) fico = 99 .

IF (polit_a = 0 & polit_b = 0 & polit_c = 0) ini = 99 .
*definovanie prítomnosti daného politika v odpovedi respondentov kódom 1.
IF (polit_a = 1 or polit_b = 1 or polit_c = 1) ziadny = 1 .
IF (polit_a = 2 or polit_b = 2 or polit_c = 2) dzurinda = 1 .
IF (polit_a = 3 or polit_b = 3 or polit_c = 3) gasparovic = 1 .
IF (polit_a = 4 or polit_b = 4 or polit_c = 4) fico = 1 .

IF (polit_a = 4 or polit_b = 4 or polit_c = 4) ini = 1 .
EXECUTE.
*definovanie neprítomnosti politika v odpovediach respondentov kódom 0 a označenie chýbajúcich hodnôt.
RECODE
  ziadny to ini  (SYSMIS=0)  (99=SYSMIS)  .
```

Výsledkom sú indikátorové premenné reprezentujúce „prítomnosť“ daného politika. Ako sme už uviedli prv, pre iných politikov je výnimkou kód 25 a teda premenná **ini**, kde sa po transformácii už nemôžu vyskytnúť prípadné multiplicity tohto kódu. Ilustrujeme to vo výstupe indikátorovej premennej, kde je percento jej výskytu 7,4%, hoci reálne bolo 7,7%.

ini		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	4025	86,7	92,6	92,6
	1	322	6,9	7,4	100,0
	Total	4347	93,6	100,0	
Missing	System	295	6,4		
Total		4642	100,0		

Tento „nedostatok“ sa dá odstrániť rozšírením kódovacieho kľúča. Keďže takáto zmena neovplyvňuje výsledky „skúmaných“ politikov a rozšírenie o ďalších politikov s malým „výskytom“ nie je až také zaujímavé, tak sa zvyčajne od rozširovania rozsahu kódov upúšťa.

5. Úprava dátovej matice pre spracovanie multiresponse otázky.

Ďalší spôsob riešenia spracovania multiresponse otázok. ***Táto úprava umožňuje analýzu „multiresponse" otázky z bázy podľa počtu odpovedí a nie podľa počtu respondentov, preto pri prípadných ďalších analýzach je nutné túto úpravu vziať do úvahy, pretože nie všetky štatistické analýzy spojené aj s inými premennými takto upraveného súboru možno korektne realizovať'***.

Úpravu ilustrujeme na príklade otázky o dôveryhodnosti politikov. V tomto prípade máme maximálny počet odpovedí r=3. Dátovú maticu skopírujem trikrát „pod seba" a tri stĺpce záznamu premennej polit_a, polit_b a polit_c „nastavíme postupne „pod seba". Schematické vyjadrenie vysvetlí celú manipuláciu s dátami:

Pôvodný dátový súbor:

pagina	o1	**polit_a**	**polit_b**	**polit_c**	o2	...	om
1							
2							
. . .							
n							

skopírujeme trikrát (r=3) „pod seba", pričom pri premennej polit kopírujeme postupne „pod seba" hodnoty stĺpcov polit_a, poli_b a polit_c:

pagina	o1	**polit**	o2	...	om
1					
2					
. . .					
n					
1					
2					
. . .					
n					
1					
2					
. . .					
n					

6. Štatistické spracovanie otázok s viacerými možnosťami odpovede

Ako už bolo uvedené, otázka s viacerými možnosťami odpovede nie je štatistický znak a preto vyžaduje jej spracovanie špecifický postup. SPSS má najlepšie vyriešené štatistické spracovanie multiresponse otázok. Príklad výstupu frekvenčnej tabuľky je v kapitole 2. Ďalšou možnosťou je transformácia na indikátorové premenné a ich následné spracovanie. V prípade duplicitného výskytu rovnakého kódu v odpovediach respondentov nastáva problém, pretože redundantné kódy sa transformáciou na indikátorové premenné „zrušia".

SPSS dáva možnosť tvorby kontingenčných tabuliek pre kombinácie multiresponse otázky so štatistickým znakom a tiež s multiresponse otázkou, pričom príslušné percentuálne vyjadrenia počíta z dvoch báz – z počtu respondentov a z počtu odpovedí.

7. Literatúra

[1] Chajdiak J.: Štatistické úlohy a ich riešenie v Exceli. STATIS, Bratislava 2005, ISBN 80-85659-39-5.

[2] Kanderová, M. – Úradníček, V.: Štatistika a pravdepodobnosť pre ekonómov. 1. časť. OZ Financ, Banská Bystrica 2007, ISBN 978-80-969535-5-4.

[3] Kanderová, M. – Úradníček, V.: Štatistika a pravdepodobnosť pre ekonómov. 2. časť. OZ Financ, Banská Bystrica 2007, ISBN 987-80-696535-6-1.

[4] Luha J.: Skúmanie súboru kvalitatívnych dát. EKOMSTAT 2003. SŠDS Bratislava 2003.

[5] Luha J.: Matematickoštatistické aspekty spracovania dotazníkových výskumov. Štatistické metódy vo vedecko-výskumnej práci 2003, SŠDS, Bratislava 2003. ISBN 80-88946-32-8.

[6] Luha J.: Viacrozmerné štatistické metódy analýzy kvalitatívnych znakov. *EKOMSTAT 2005*, Štatistické metódy v praxi. Trenčianske Teplice 22.–27. 5. 2005. SŠDS Bratislava 2005.

[7] Luha J.: Korelácie medzi politikmi a stranami. FORUM STATISTICUM SLOVACUM 7/2008. SŠDS Bratislava 2008. ISSN 1336-7420.

[8] Pecáková I.: Statistika v terénních průzkumech. Proffessional Publishing, Praha 2008. ISBN 978-80-86946-74-0.

[9] Řehák J., Řeháková B.: Analýza kategorizovaných dat v sociologii. Academia Praha 1986.

[10] Řezanková A.: Analýza dat z dotazníkových šetření. Proffessional Publishing, Praha 2007. ISBN 978-80-86946-49-8.

[11] Stankovičová I., Vojtková M.: Viacrozmerné štatistické metódy s aplikáciami. IURA EDITION, Bratislava 2007, ISBN 978-80-8078-152-1.

Adresa autora:

RNDr. Ján Luha, CSc.
Jan.Luha@statistics.sk

Korelácie medzi politikmi a stranami
Correlations between politicians and parties

Ján Luha

Abstract: This article present relations on the political scene represented correlation between indicator variables of politicians, parties and politicians and parties. Results are illustrate on researches Institute for public opinion research at the Statistical Office of the Slovak Republic.

Key words: relations, correlation coefficients, indicator variables, politicians, parties, correlation profile.

Kľúčové slová: vzťahy, korelačné koeficienty, indikátorové premenné, politici, strany, korelačný profil.

1. Úvod

V príspevku sa venujeme špecifickým vzťahom vyjadrených koreláciami politikov, strán a taktiež politikov a strán na konkrétnych ilustráciách výsledkov Ústavu pre výskum verejnej mienky pri ŠÚ SR (ďalej ÚVVM) za mesiace máj až august 2008. Aby sme vylúčili potenciálny vplyv „terénnej“ fázy budeme pracovať so spojeným súborom za uvedené štyri vybrané mesiace.

Skúmame korelácie (Pearsonov korelačný koeficient) ***medzi politikmi, stranami a medzi politikmi a stranami*** počítané pomocou indikátorových premenných jednotlivých politikov a strán.

Poznámka: Korelácie počítané cez indikátorové premenné sú determinované veľkosťou podielu výskytu javov charakterizovaných indikátorovými premennými. Explicitný vzťah korelácie medzi politikmi je odvodený v práci [13] Luha J. (2008).

Otázka skúmajúca dôveryhodnosť politikov má možnosť (obvykle) troch odpovedí a teda nepredstavuje štatistický znak v zmysle jeho definície. Pri jej analýze využívame v prostredí SPSS spracovanie otázok s viacerými možnosťami odpovede „multiresponse“.

Znenie otázky vo výskumoch ÚVVM:

Ktorý z politikov pôsobiacich na Slovensku má v súčasnosti vašu dôveru ?
Možno uviesť najviac tri osobnosti.

Uvádzame výsledok za spojený súbor zo štyroch výskumov:

Poznámka: Pri kóde iní politici (=25), po transformácii na indikátorovú premennú, zaniknú prípadné „multiplicity“ tohto kódu, pretože kód iný politik sa môže u respondenta vyskytnúť viackrát (maximálne 3 krát – vzhľadom na konštrukciu otázky).

Multiple Response

Case Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
$polit	4347	97,3%	122	2,7%	4469	100,0%

$polit Frequencies

	N	Percent	Percent of Cases	
4 R.Fico	1467	21,7	**33,7**	
1 žiadnemu	1094	16,2	**25,2**	
3 I.Gašparovič	727	10,8	**16,7**	
7 J.Slota	469	6,9	**10,8**	
22 R.Kaliňák	420	6,2	**9,7**	
20 I.Radičová	349	5,2	**8,0**	
5 V.Mečiar	341	5,0	**7,8**	
25 iní	335	5,0	**7,7**	
8 B.Bugár	235	3,5	**5,4**	
2 M.Dzurinda	212	3,1	**4,9**	
18 I.Mikloš	175	2,6	**4,0**	
14 P.Hrušovský	139	2,1	**3,2**	
11 J.Počiatek	97	1,4	**2,2**	
15 D.Lipšic	94	1,4	**2,2**	
9 V.Tomanová	90	1,3	**2,1**	
24 P.Csáky	86	1,3	**2,0**	
19 A.Belousovová	72	1,1	**1,7**	
10 D.Čaplovič	69	1,0	**1,6**	
17 P.Paška	60	0,9	**1,4**	
6 M.Flašíková-Beňová	58	0,9	**1,3**	
13 F.Mikloško	56	0,8	**1,3**	
12 V.Palko	39	0,6	**0,9**	
16 E.Kukan	27	0,4	**0,6**	
23 Z.Martináková	24	0,4	**0,6**	
21 V.Veteška	18	0,3	**0,4**	
Total	**6753**	**100**	**155,3**	

Otázka na zistenie volebných preferencií strán vo výskumoch ÚVVM má znenie:

Predstavte si, že by sa voľby do Národnej rady SR konali už teraz. Ktorú stranu, hnutie, koalíciu by ste volili?

Výsledok za spojený súbor zo štyroch výskumov, ***bez prepočítania*** na bázu bez tých čo by sa nezúčastnili volieb a tých čo nevedia koho by volili:

strany Ktorú stranu by ste volili do NR SR?

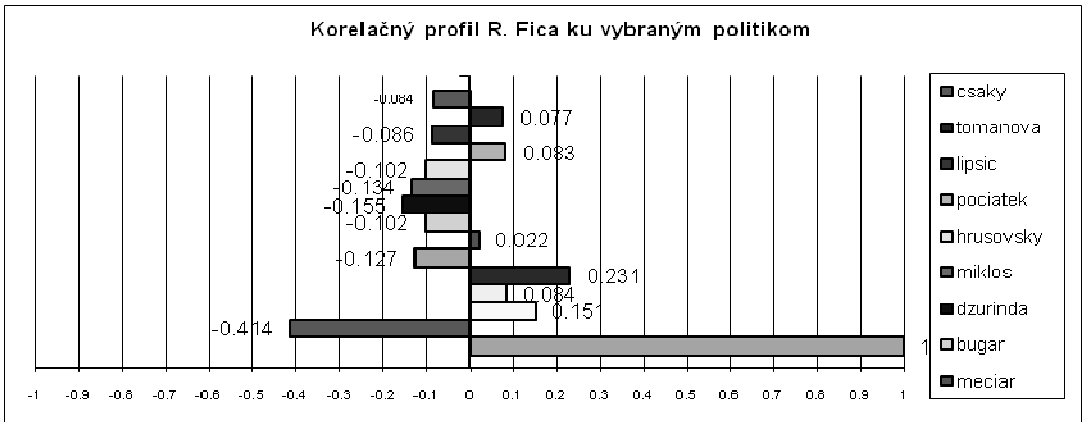
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	9 SMER	1330	29,8	**29,8**	29,8
	12 nezúčastnil by sa	830	18,6	**18,6**	48,5
	13 neviem	717	16,0	**16,1**	64,6
	7 SDKÚ	363	8,1	**8,1**	72,7
	8 SNS	327	7,3	**7,3**	80,0
	10 SMK	281	6,3	**6,3**	86,4
	5 ĽS-HZDS	257	5,8	**5,8**	92,1
	4 KDH	236	5,3	**5,3**	97,4
	3 KSS	41	,9	**,9**	98,3
	2 HZD	35	,8	**,8**	99,1
	6 SF	23	,5	**,5**	99,6
	11 inú	11	,2	**,2**	99,9
	1 ANO	5	,1	**,1**	100,0
	Total	**4456**	**99,7**	**100,0**	
Missing	0	13	,3		
Total		4469	100,0		

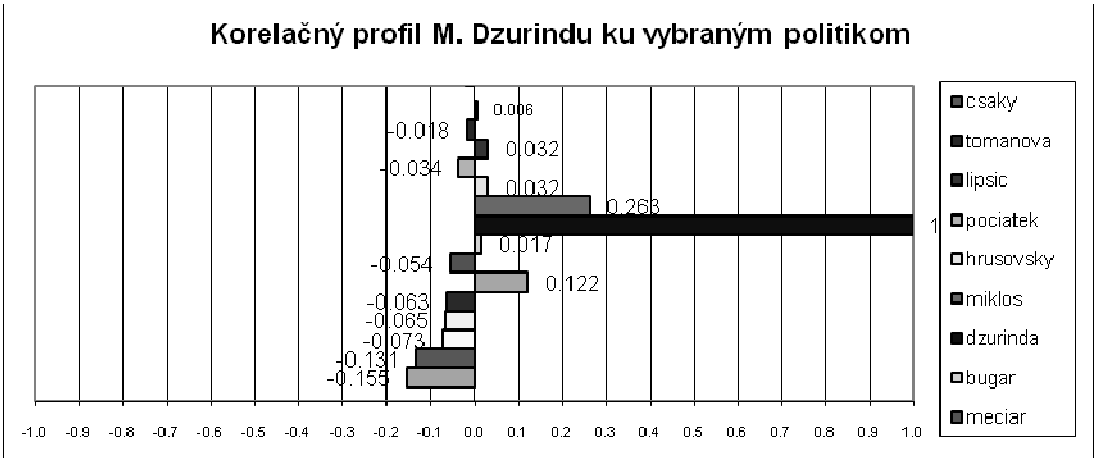
2. Závislosť medzi politikmi

Uvedieme korelácie medzi politikmi s dôveryhodnosťou aspoň 2%: P. Csáky, V. Tomanová, D. Lipšic, J. Počiatek, P. Hrušovský, I. Mikloš, M. Dzurinda, B. Bugár, V. Mečiar, I. Radičová, R. Kaliňák, J. Slota, I. Gašparovič, „žiadnemu“, R. Fico. Korelačná matica vybraných politikov:

	fico	ziadny	gasparovic	slota	kalinak	radicova	meciar	bugar	dzurinda	miklos	hrusovsky	pociatek	lipsic	tomanova	csaky
fico	1	-0,414	0,151	0,084	0,231	-0,127	0,022	-0,102	-0,155	-0,134	-0,102	0,083	-0,086	0,077	-0,084
ziadny	-0,414	1	-0,260	-0,202	-0,190	-0,171	-0,169	-0,139	-0,131	-0,119	-0,105	-0,088	-0,086	-0,084	-0,082
gasparovic	0,151	-0,260	1	0,019	0,004	-0,048	-0,046	-0,058	-0,073	-0,079	-0,036	-0,030	-0,058	0,026	-0,042
slota	0,084	-0,202	0,019	1	-0,003	-0,086	0,114	-0,063	-0,065	-0,067	-0,051	-0,007	-0,042	-0,019	-0,049
kalinak	0,231	-0,190	0,004	-0,003	1	-0,057	-0,061	-0,051	-0,063	-0,043	-0,055	0,156	-0,043	0,007	-0,041
radicova	-0,127	-0,171	-0,048	-0,086	-0,057	1	-0,058	0,030	0,122	0,181	0,033	0,001	0,026	-0,025	-0,005
meciar	0,022	-0,169	-0,046	0,114	-0,061	-0,058	1	-0,062	-0,054	-0,055	-0,038	-0,038	-0,043	-0,012	-0,041
bugar	-0,102	-0,139	-0,058	-0,063	-0,051	0,030	-0,062	1	0,017	0,013	-0,026	-0,029	-0,001	-0,006	0,178
dzurinda	-0,155	-0,131	-0,073	-0,065	-0,063	0,122	-0,054	0,017	1	0,263	0,032	-0,034	0,032	-0,018	0,006
miklos	-0,134	-0,119	-0,079	-0,067	-0,043	0,181	-0,055	0,013	0,263	1	-0,004	-0,015	0,026	-0,022	-0,021
hrusovsky	-0,102	-0,105	-0,036	-0,051	-0,055	0,033	-0,038	-0,026	0,032	-0,004	1	-0,027	0,243	-0,026	-0,007
pociatek	0,083	-0,088	-0,030	-0,007	0,156	0,001	-0,038	-0,029	-0,034	-0,015	-0,027	1	-0,022	0,000	-0,021
lipsic	-0,086	-0,086	-0,058	-0,042	-0,043	0,026	-0,043	-0,001	0,032	0,026	0,243	-0,022	1	0,001	0,002
tomanova	0,077	-0,084	0,026	-0,019	0,007	-0,025	-0,012	-0,006	-0,018	-0,022	-0,026	0,000	0,001	1	-0,009
csaky	-0,084	-0,082	-0,042	-0,049	-0,041	-0,005	-0,041	0,178	0,006	-0,021	-0,007	-0,021	0,002	-0,009	1

Zaujímavé je grafické vyjadrenie korelačného profilu politika voči vybraným politikom vrátane samotného politika. Kvôli úspore miesta uvedieme iba korelačné profily R. Fica a M. Dzurindu. Poradie politikov v grafe je zvolené náhodne, ale rovnaké vo všetkých ilustratívnych grafoch. Škála osi znázorňujúcej hodnoty korelácií je od –1 po +1.





Vzťahy politika ku vybraným politikom vyjadrené koreláciami možno lepšie skúmať z hore uvedeného grafického vyjadrenia. Všímame si znamienko vzťahov a tiež veľkosť korelačných koeficientov. Napríklad v grafe korelačného profilu R. Fica sme zistili kladnú koreláciu vo vzťahoch fico x tomanova, pociatek, meciar, kalinak, slota a gasparovic, teda ku skúmaným koaličným politikom a ku prezidentovi SR. Zápornú koreláciu má R. Fico ku všetkým skúmaným opozičným politikom a ku „politikovi" ziadny. Táto korelácia je u R. Fica najväčšia.

V korelačnom profile M. Dzurindu: Kladnú koreláciu sme zistili pri politikoch: csaky, lipsic, hrusovsky, miklos, bugar a radicova a zápornú ku terajším koaličným politikom a taktiež ku politikovi ziadny, ale značne menšej hodnoty ako u R. Fica.

Detailnejšiu analýzu možno realizovať pomocou kontingenčných tabuliek. Uvádzame príklady vzťahu fico x ziadny a fico x dzurinda:

Crosstab

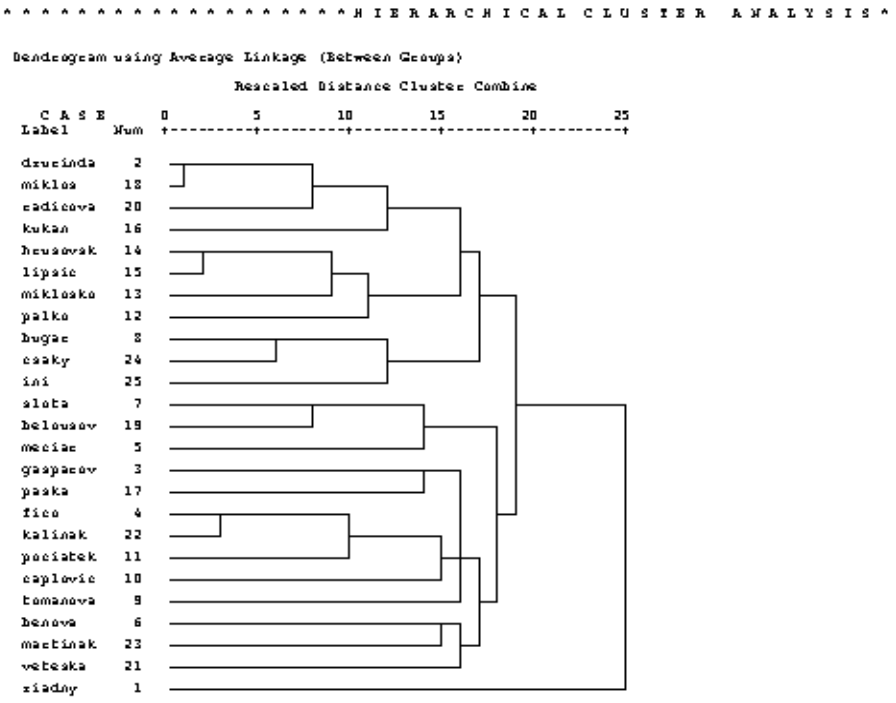
			ziadny		Total
			0	1	0
fico	0	Count	1786	1094	2880
		% within fico	62,0%	38,0%	100,0%
		% within ziadny	54,9%	100,0%	66,3%
		Adjusted Residual	-27,3	27,3	
	1	Count	1467	0	1467
		% within fico	100,0%	,0%	100,0%
		% within ziadny	45,1%	,0%	33,7%
		Adjusted Residual	27,3	-27,3	
Total		Count	3253	1094	4347
		% within fico	74,8%	25,2%	100,0%
		% within ziadny	100,0%	100,0%	100,0%

Crosstab

			dzurinda		Total
			0	1	0
fico	0	Count	2671	209	2880
		% within fico	92,7%	7,3%	100,0%
		% within dzurinda	64,6%	98,6%	66,3%
		Adjusted Residual	-10,2	10,2	
	1	Count	1464	3	1467
		% within fico	99,8%	,2%	100,0%
		% within dzurinda	35,4%	1,4%	33,7%
		Adjusted Residual	10,2	-10,2	
Total		Count	4135	212	4347
		% within fico	95,1%	4,9%	100,0%
		% within dzurinda	100,0%	100,0%	100,0%

Vo vzťahu ***fico x ziadny*** vidno ešte potenciál pre R. Fica. z 2880, ktorí mu nedôverujú „dôveruje" žiadnemu politikovi 1094 a „zostáva" 1786, ktorí nedôverujú nikomu a tiež R. Ficovi. Vo vzťahu s M. Dzurindom máme troch respondentov, ktorí dôverujú obom politikom. Z 2880 tých čo nedôverujú R. Ficovi je 209 takých čo dôverujú M. Dzurindovi ale 2671 mu nedôveruje.

Zaujímavý „obrázok" vzťahu politikov dáva zhluková (cluster) analýza.



3. Závislosť medzi stranami

Špecifikom korelácií indikátorových premenných kvalitatívneho znaku je ich ***zápornosť***, viď práca Luha J.(2008): Korelácie disjunktných a komplementárnych javov. FORUM STATISTICUM SLOVACUM 6/2008. SŠDS Bratislava 2008. ISSN 1336-7420.

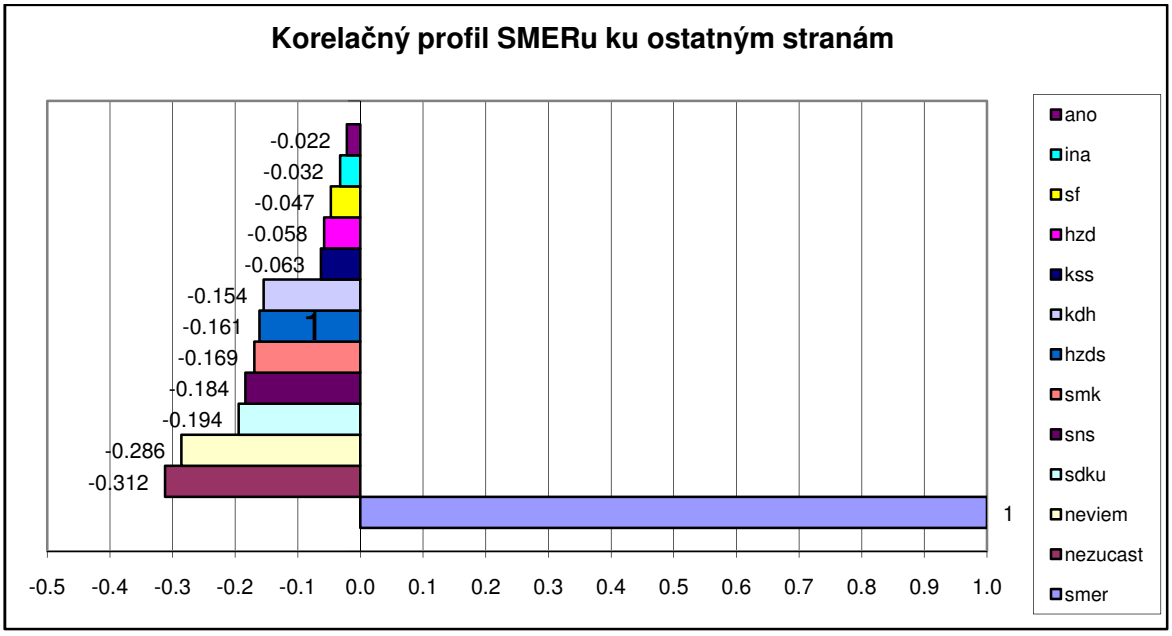
V citovanej práci je dokázaný vzťah, podľa ktorého, je veľkosť korelačného koeficienta určená veľkosťou preferencií strán. Na ilustráciu uvedieme malú tabuľku:

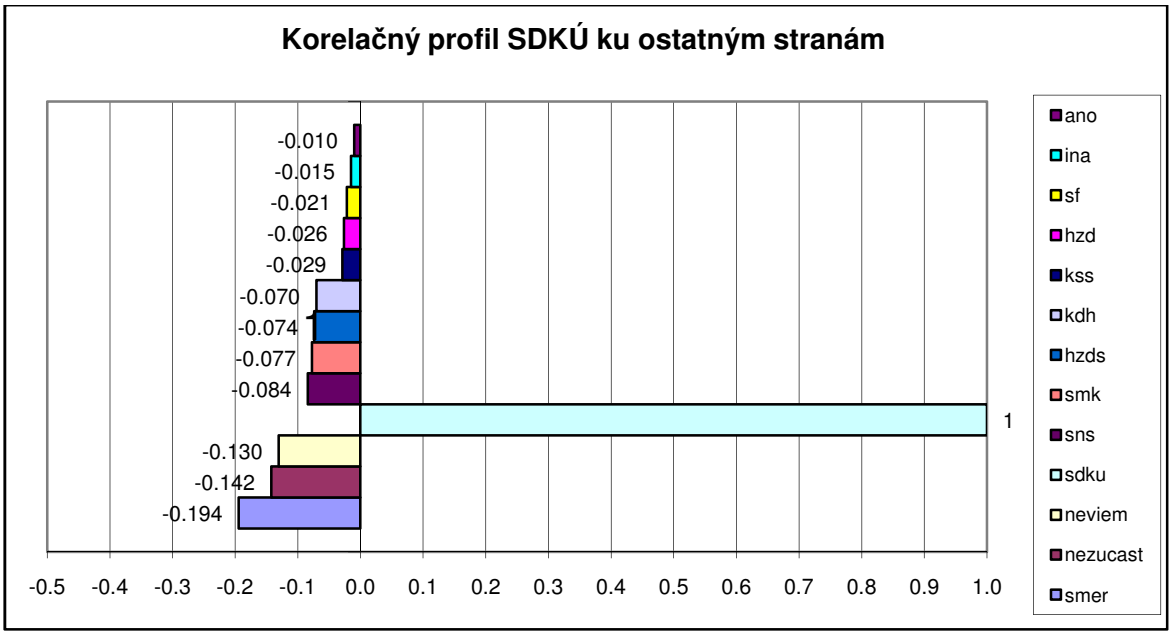
p_1	0,1	0,1	0,1	0,2	0,2	0,3	0,3	0,3	0,4	0,4
p_2	0,1	0,2	0,3	0,4	0,6	0,3	0,5	0,6	0,5	06
r	-0,012	-0,026	-0,048	-0,167	-0,583	-0,184	-0,429	-0,643	-,667	-1

V prípade $p_1+p_2=1$ ide o komplementárne javy a ich korelačný koeficient je rovný -1. Vzťahy všetkých strán, vrátane nezúčastnených a neviem, uvádzame v 13x13 korelačnej matici. Vieme, že vzájomné korelácie sú záporné a tak môžeme hľadať v korelačných profiloch strán najväčšie záporné korelácie. Prípadná veľká korelácia so „stranami“ nezucast a neviem naznačuje potenciál strán na možné zisky.

	smer	nezucast	neviem	sdku	sns	smk	hzds	kdh	kss	hzd	sf	ina	ano
smer	1	-0,312	-0,286	-0,194	-0,184	-0,169	-0,161	-0,154	-0,063	-0,058	-0,047	-0,032	-0,022
nezucast	-0,312	1	-0,210	-0,142	-0,135	-0,124	-0,118	-0,113	-0,046	-0,043	-0,034	-0,024	-0,016
neviem	-0,286	-0,210	1	-0,130	-0,123	-0,114	-0,108	-0,104	-0,042	-0,039	-0,032	-0,022	-0,015
sdku	-0,194	-0,142	-0,130	1	-0,084	-0,077	-0,074	-0,070	-0,029	-0,026	-0,021	-0,015	-0,010
sns	-0,184	-0,135	-0,123	-0,084	1	-0,073	-0,070	-0,067	-0,027	-0,025	-0,020	-0,014	-0,009
smk	-0,169	-0,124	-0,114	-0,077	-0,073	1	-0,064	-0,061	-0,025	-0,023	-0,019	-0,013	-0,009
hzds	-0,161	-0,118	-0,108	-0,074	-0,070	-0,064	1	-0,059	-0,024	-0,022	-0,018	-0,012	-0,008
kdh	-0,154	-0,113	-0,104	-0,070	-0,067	-0,061	-0,059	1	-0,023	-0,021	-0,017	-0,012	-0,008
kss	-0,063	-0,046	-0,042	-0,029	-0,027	-0,025	-0,024	-0,023	1	-0,009	-0,007	-0,005	-0,003
hzd	-0,058	-0,043	-0,039	-0,026	-0,025	-0,023	-0,022	-0,021	-0,009	1	-0,006	-0,004	-0,003
sf	-0,047	-0,034	-0,032	-0,021	-0,020	-0,019	-0,018	-0,017	-0,007	-0,006	1	-0,004	-0,002
ina	-0,032	-0,024	-0,022	-0,015	-0,014	-0,013	-0,012	-0,012	-0,005	-0,004	-0,004	1	-0,002
ano	-0,022	-0,016	-0,015	-0,010	-0,009	-0,009	-0,008	-0,008	-0,003	-0,003	-0,002	-0,002	1

Napríklad smer má najväčšiu koreláciu so stranami sdku, kdh, smk, ale aj hzds, sns a čo je vzhľadom na konkurenčnosť prirodzené. Korelácie so stranami s malými preferenciami sú malé. Treba si však všimnúť značné korelácie so „stranami“ nezucast a neviem. Na ilustráciu uvedieme korelačné profily SMERu a SDKÚ. Škálu x-ovej osi sme volili jednotne od –0,5 po 1. Poradie strán v ilustratívnych grafoch je zvolené podľa veľkosti preferencií.



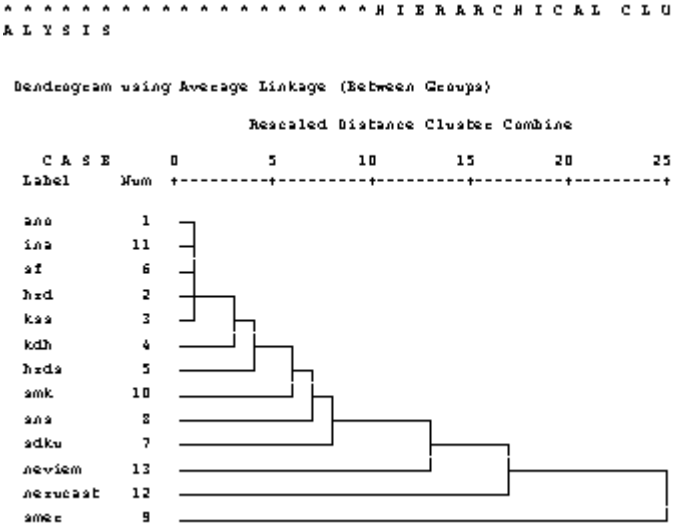


Korelačný koeficient vyjadruje určitú mieru vzťahu. Na podrobnejšiu analýzu využívame kontingenčné tabuľky vzájomných vzťahov strán. Na ilustráciu uvedieme dve kontingenčné tabuľky.

			sdku		Total
			0	1	0
smer	0	Count	2763	363	3126
		% within smer	88,4%	11,6%	100,0%
		% within sdku	67,5%	100,0%	70,2%
		Adjusted Residual	-13,0	13,0	
	1	Count	1330	0	1330
		% within smer	100,0%	,0%	100,0%
		% within sdku	32,5%	,0%	29,8%
		Adjusted Residual	13,0	-13,0	
Total		Count	4093	363	4456
		% within smer	91,9%	8,1%	100,0%
		% within sdku	100,0%	100,0%	100,0%

			hzds		Total
			0	1	0
smer	0	Count	2869	257	3126
		% within smer	91,8%	8,2%	100,0%
		% within hzds	68,3%	100,0%	70,2%
		Adjusted Residual	-10,8	10,8	
	1	Count	1330	0	1330
		% within smer	100,0%	,0%	100,0%
		% within hzds	31,7%	,0%	29,8%
		Adjusted Residual	10,8	-10,8	
Total		Count	4199	257	4456
		% within smer	94,2%	5,8%	100,0%
		% within hzds	100,0%	100,0%	100,0%

Zaujímavý pohľad na vzájomný vzťah strán dáva dendogram.



4. Závislosť medzi politikmi a stranami

Skúmanie závislosti medzi dôveryhodnosťou politikov a volebnými preferenciami strán dáva zaujímavý pohľad na situáciu na politickej scéne. Azda najzaujímavejším sa ukazuje vzťah lídra strany ku svojej materskej strane a taktiež korelácie ďalších politikov najmä ku „svojej“ strane. Korelačné koeficienty politikov s dôveryhodnosťou 2 a viac percent so stranami sú v nasledovnej tabuľke. n=4338.

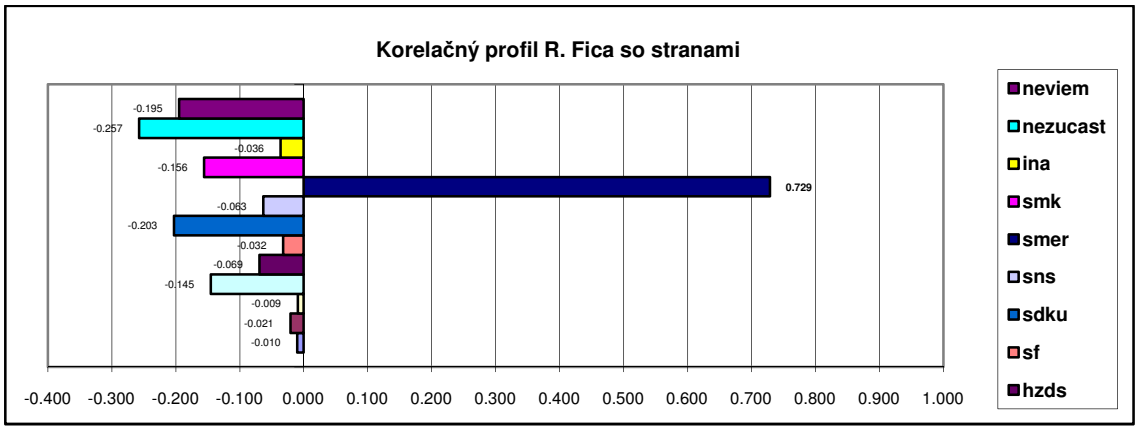
	ano	hzd	kss	kdh	hzds	sf	sdku	sns	smer	smk	ina	nezucast	neviem
fico	-0,010	-0,021	-0,009	-0,145	-0,069	-0,032	-0,203	-0,063	**0,729**	-0,156	-0,036	-0,257	-0,195
ziadny	0,012	-0,052	0,020	-0,092	-0,127	-0,006	-0,123	-0,141	-0,336	-0,051	-0,019	**0,513**	**0,266**
dzurinda	-0,008	0,003	-0,022	0,021	-0,048	-0,017	**0,555**	-0,057	-0,146	-0,024	-0,011	-0,075	-0,073
gasparovic	0,003	**0,111**	0,046	-0,031	-0,021	-0,016	-0,086	0,017	**0,182**	-0,086	-0,010	-0,109	-0,009
meciar	-0,010	0,080	0,007	-0,054	**0,694**	-0,021	-0,088	-0,008	-0,081	-0,072	-0,015	-0,104	-0,098
benova	-0,004	-0,010	-0,011	-0,019	-0,004	0,019	-0,013	-0,026	**0,071**	-0,022	0,034	-0,044	0,011
slota	-0,012	-0,015	-0,011	-0,077	0,014	-0,025	-0,099	**0,629**	-0,043	-0,090	-0,018	-0,109	-0,097
bugar	-0,008	-0,022	-0,013	-0,035	-0,056	0,011	-0,006	-0,049	-0,112	**0,560**	0,008	-0,053	-0,074
tomanova	-0,005	0,005	0,019	-0,013	-0,009	-0,011	-0,038	-0,005	**0,079**	0,015	-0,007	-0,042	-0,022
caplovic	-0,004	-0,011	-0,012	-0,022	-0,016	-0,009	-0,018	-0,008	**0,108**	-0,018	-0,006	-0,050	-0,019
pociatek	-0,005	0,021	-0,015	-0,036	-0,038	-0,011	-0,034	0,004	**0,130**	-0,039	-0,008	-0,066	0,004
palko	-0,003	-0,009	-0,009	**0,193**	-0,024	-0,007	0,033	-0,027	-0,053	-0,025	0,044	-0,025	0,000
miklosko	-0,004	-0,010	-0,011	**0,270**	-0,029	-0,008	0,054	-0,017	-0,071	-0,021	0,075	-0,043	-0,032
hrusovsky	-0,006	-0,016	-0,018	**0,633**	-0,046	-0,013	-0,022	-0,042	-0,118	-0,047	-0,009	-0,064	-0,056
lipsic	-0,005	-0,013	-0,015	**0,398**	-0,037	0,011	0,030	-0,042	-0,092	-0,026	0,024	-0,057	-0,020
kukan	-0,003	-0,007	-0,008	-0,006	-0,020	-0,006	**0,199**	-0,023	-0,046	0,003	-0,004	-0,029	-0,026
paska	-0,004	0,011	-0,012	0,007	-0,005	-0,009	-0,029	0,049	**0,063**	-0,031	0,033	-0,035	-0,040
miklos	-0,007	-0,018	-0,020	-0,013	-0,051	-0,015	**0,494**	-0,050	-0,121	-0,020	-0,010	-0,080	-0,033
belousovova	-0,004	-0,012	-0,013	-0,023	-0,017	-0,009	-0,033	**0,285**	-0,031	-0,034	-0,007	-0,046	-0,036
radicova	-0,010	0,002	-0,029	0,023	-0,052	0,037	**0,353**	-0,071	-0,120	-0,021	0,019	-0,078	0,050
veteska	-0,002	-0,006	-0,006	0,000	**0,090**	-0,005	-0,019	-0,005	-0,004	-0,002	-0,003	-0,030	0,002
kalinak	-0,011	-0,003	-0,016	-0,071	-0,049	-0,002	-0,087	-0,013	**0,309**	-0,076	-0,017	-0,106	-0,069
martinakova	-0,003	-0,007	0,025	-0,004	-0,005	**0,380**	-0,011	-0,009	-0,029	-0,019	0,058	-0,027	0,011
csaky	-0,005	-0,013	-0,014	-0,034	-0,036	-0,010	-0,031	-0,041	-0,094	**0,506**	-0,007	-0,058	-0,052

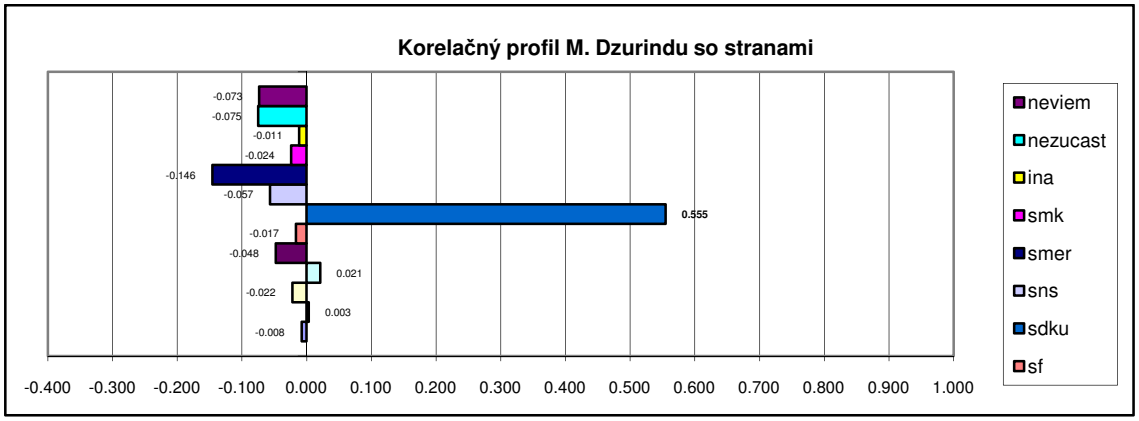
Na základe konkrétnych výskumov ÚVVM za máj až august 2008 možno, v spojenom súbore za uvedené štyri mesiace, zistiť dominantné korelácie lídrov strán a „zaujímavé“ korelácie niektorých ďalších predstaviteľov niektorých strán:

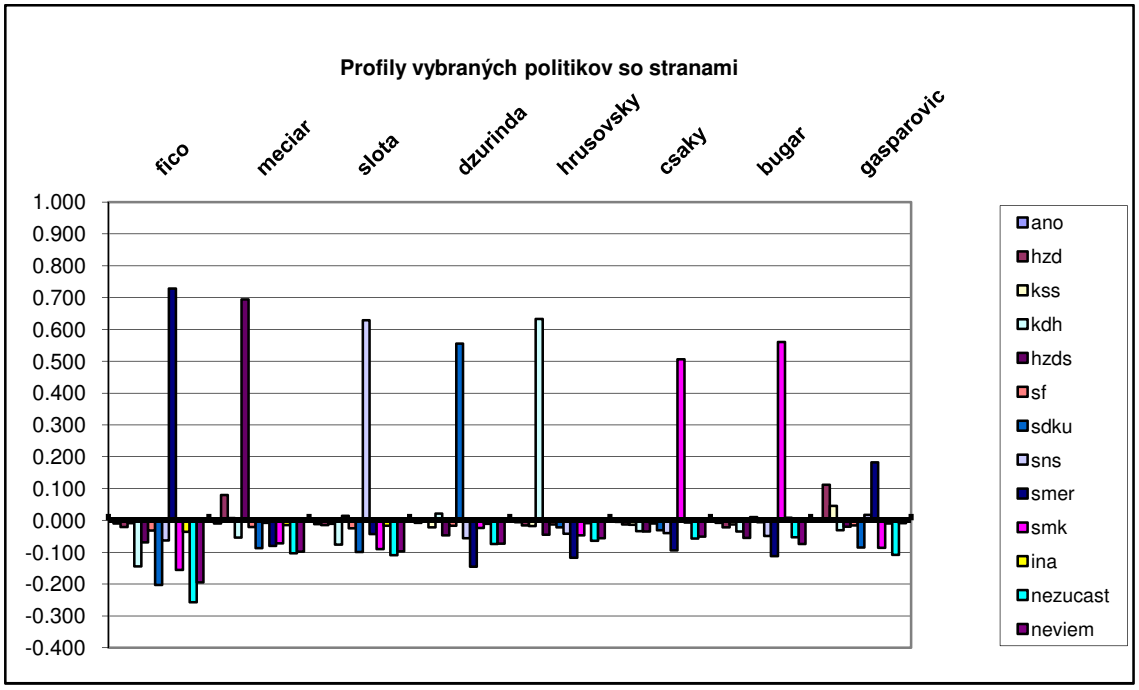
- Najsilnejšia je, podľa očakávania, ***korelácia fico x smer 0,729***. Z politikov SMERU sú ešte zaujímavé korelácie ***kalinak 0,309, caplovic 0,108***, ***pociatek 0,130, tomanova 0,079, benova 0,071 a paska 0,063***.
- V HZDS je to korelácia ***meciar x hzds 0,694.*** Z ostatných skúmaných politikov má kladnú koreláciu s HZDS ***veteska x hzds 0,090***.
- V SNS dominuje ***slota x sns 0,629 a zaujímavá je korelácia belousovova 0,285***.
- V opozičných stranách:
  - SDKÚ má najsilnejší vzťah ***dzurinda x sdku 0,555, ďalej miklos 0,494, radicova 0,353 a kukan 0,199.***
  - V KDH ***hrusovsky x kdh 0,633 a lipsic 0,398*** a veľmi zaujímavé sú korelácie „odídencov“ z kdh ku tejto strane: ***miklosko 0,270*** a ***palko 0,193***.
  - SMK „má“ vlastne dvoch lídrov ***csaky x smk 0,506*** a ***bugar x smk 0,560***, pričom bývalý líder má väčšiu dôveryhodnosť.
  - Korelácie prezidenta SR I. Gašparoviča sú kladné so 4 stranami: ***smer 0,182***, ***hzd 0,111***, ***kss 0,045, a sns 0,017.***
  - SF je strana s malými preferenciami a tak korelácia ***martinakova x sf 0,380*** nehrá veľkú úlohu.

Poznámka: Okrem hodnoty korelačných koeficientov je potrebné pri analýzach brať do úvahy aj „silu“ politika vyjadrenú veľkosťou percenta dôveryhodnosti a pri strane veľkosťou percenta preferencií. V súčasnosti je „sila“ R. Fica násobne väčšia ako iných politikov a podobne aj preferencie SMERu a tak ich vzťah je aj tým silnejší a vzájomná väzba je určujúca.

Grafické vyjadrenie korelačného profilu politika oproti stranám ilustrujeme na korelačných profiloch R. Fica a M. Dzurindu a na súhrnnom grafe skúmaných politikov.







Ďalšiu možnosť podrobnejšieho skúmania vzťahu politikov so stranami poskytujú kontingenčné tabuľky. Pre úsporu miesta uvedieme iba kontingenčnú tabuľku fico x smer:

fico * smer Crosstabulation

			smer		Total
			0	1	0
fico	0	Count	2684	190	2874
		% within fico	93,4%	6,6%	100,0%
		% within smer	89,1%	14,3%	66,3%
		Adjusted Residual	48,0	-48,0	
	1	Count	328	1136	1464
		% within fico	22,4%	77,6%	100,0%
		% within smer	10,9%	85,7%	33,7%
		Adjusted Residual	-48,0	48,0	
Total		Count	3012	1326	4338
		% within fico	69,4%	30,6%	100,0%
		% within smer	100,0%	100,0%	100,0%

Keď skúmame iba ***vzťah lídra a jeho materskej strany*** môžeme zistiť prípadný jeho potenciál z množstva respondentov, ktorí mu neprejavujú dôveru. Napríklad dôveru ku R. Ficovi deklarovalo 33,7% a teda má možnosť osloviť ešte 66,3% respondentov.

Jednoznačne možno konštatovať že ***najsilnejšie pôsobí na preferencie materskej strany R. Fico,*** nasledujú V. Mečiar a HZDS, J. Slota a SNS a P. Hrušovský a KDH. Pri SMK sa prejavuje rozštiepenosť a bývalý líder B. Bugár má stále väčšiu koreláciu s materskou stranou ako súčasný líder P. Csáky. Líderská pozícia bývalého premiéra M. Dzurindu má síce dominantnú ale menšiu koreláciu ku SDKÚ, ako vykazujú lídri súčasnej vládnej koalície.

Vzťah politika s materskou stranou, ***v prípade vyšších preferencií a väčšej dôveryhodnosti***, možno skúmať aj z priebehu jeho výsledkov dôveryhodnosti a preferencií „jeho“ strany ***pomocou regresnej analýzy***.

Keďže uvedené podmienky v súčasnosti najviac spĺňa vzťah SMERu a R. Fica, budeme ilustrovať iba tento regresný vzťah na základe údajov za obdobie prvých 10 mesiacov roka 2008. Uvedieme iba regresiu zavislosti smer od fico, kde uvažujeme neprepočítané preferencie SMERu.

vyskum	fico	smer
2008_01_jan	34,96	30,50
2008_03_feb	37,81	31,65
2008_05_mar	33,96	29,09
2008_06_apr	32,71	27,78
2008_08_maj	31,16	27,83
2008_09_jun	33,33	29,55
2008_11_jul	38,43	33,96
2008_14_aug	32,40	28,39
2008_17_sep	32,30	30,45
2008_20_okt	31,42	29,58

Výsledky regresného modelu závislosti smer od fico:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,849(a)	,721	,686	1,060146

a Predictors: (Constant), fico

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23,206	1	23,206	20,647	,002(a)
	Residual	8,991	8	1,124		
	Total	32,197	9			

a Predictors: (Constant), fico

b Dependent Variable: smer

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	8,306	4,759		1,745	,119
	fico	,637	,140	,849	4,544	,002

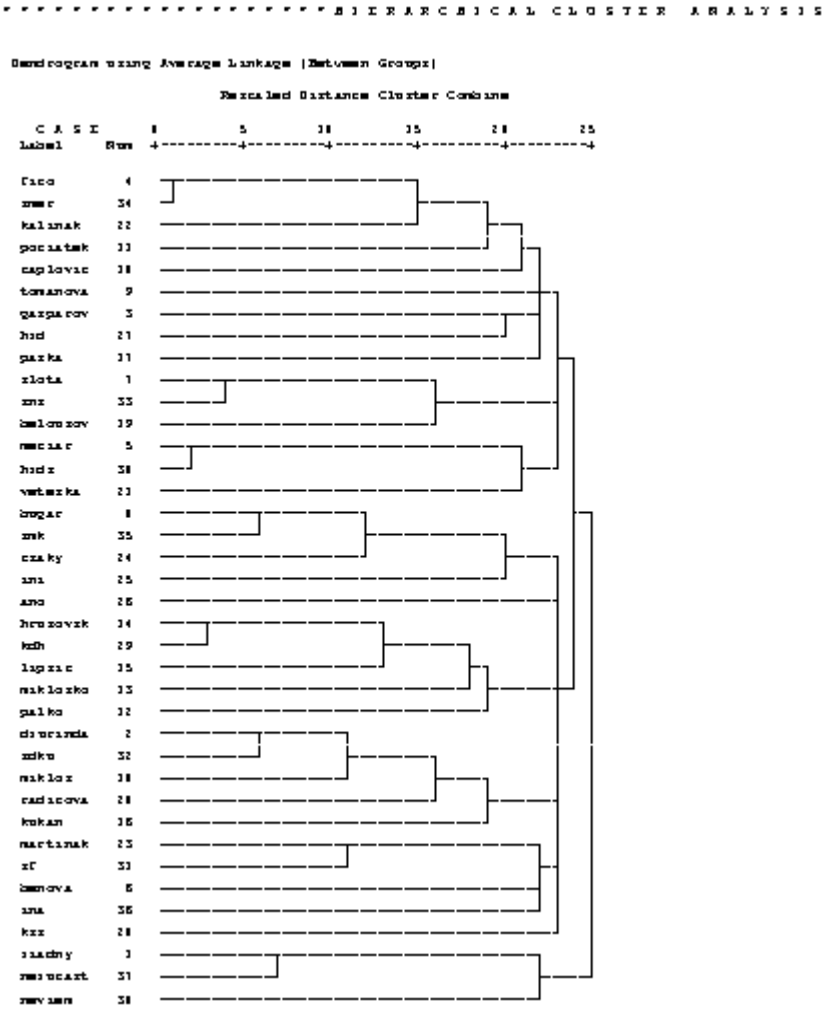
a Dependent Variable: smer

Poznámka: Výsledky vyjadrené v tvare percent preferencií a dôveryhodnosti sú značne agregované a preto sú tu skúmané korelácie prirodzene väčšie ako pri koreláciách počítaných pomocou indikátorových premenných.

Regresná závislosť smer_prep od fico nezahŕňa tú časť respondentov, ktorí by sa nezúčastnili volieb a tiež tých, ktorí by nešli voliť. ***Výsledok tejto závislosti nedáva signifikantný model*** a koeficient determinácie ukazuje, že regresný model vysvetľuje menej ako 25% variability. Na skúmanie závislosti ***smer_prep od fico*** musíme uvažovať aj „strany“ ***nezucast a neviem***. Je to prirodzené, pretože sme ich pri prepočítaní vlastne zo skúmaných vzťahov vylúčili.

Kvôli úspore miesta tieto výsledky neuvedieme.

Dendogram dáva veľmi zaujímavý pohľad na vzájomné vzťahy strán aj politikov.



V predošlej kapitole sme skúmali aj vzťah koalície - opozície pomocou spojenia odpovedajúcich strán. Môžeme skúmať aj vzťah politikov ku takto spojeným stranám, ale ako sme už poznamenali prv, takéto spojenie môže „zastrieť“ jemnejšie vzťahy s pôvodnými stranami. Na ilustráciu uvedieme ***vzťah lídrov najväčších strán so spojenými stranami***:

koalicia= smer + hzds + sns,

opozicia= sdku + kdh + smk,

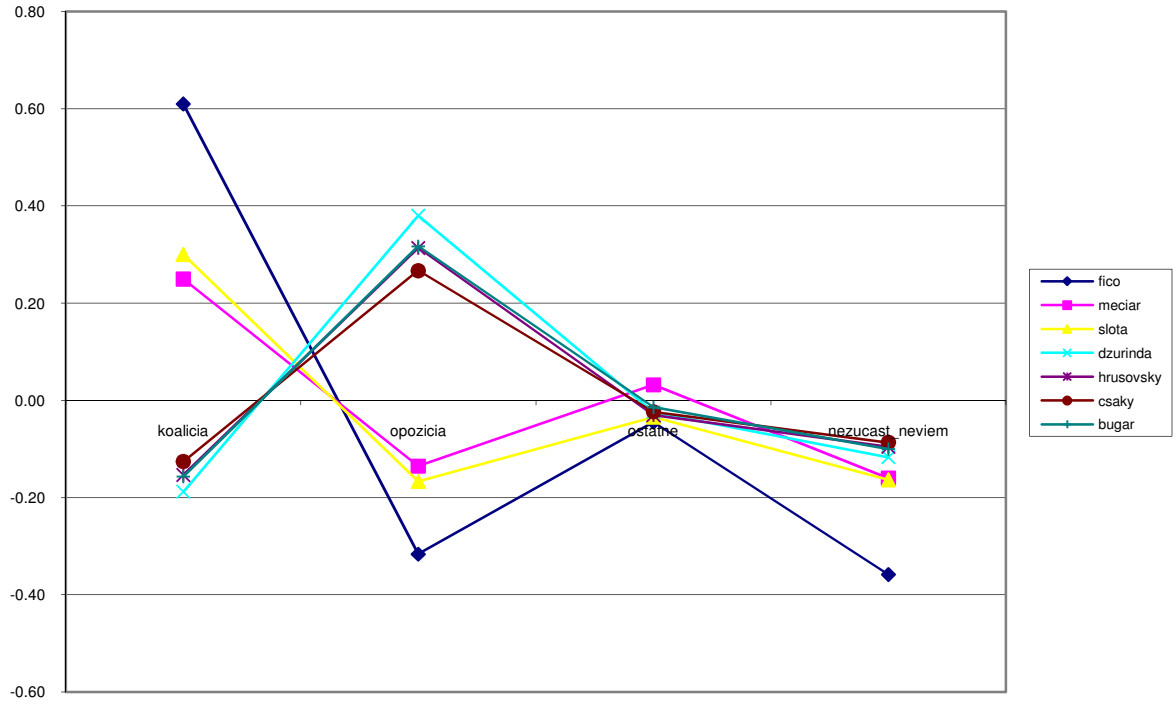
ostatne= ano + ina + hzd + kss + sf,

neucast_neviem = nezucast + neviem.

Vybrané korelačné vzťahy sú v tabuľke, n=4338:

politik	koalicia	opozicia	ostatne	nezucast_neviem
fico	0,61	-0,32	-0,04	-0,36
meciar	0,25	-0,13	0,03	-0,16
slota	0,30	-0,17	-0,03	-0,16
dzurinda	-0,19	0,38	-0,02	-0,12
hrusovsky	-0,15	0,31	-0,03	-0,10
csaky	-0,13	0,27	-0,02	-0,09
bugar	-0,16	0,32	-0,01	-0,10

Graficky ilustrujeme uvedené vzťahy nasledovne:



5. Záver

Skúmanie vzťahov medzi politikmi, stranami a najmä medzi politikmi a stranami, môžeme robiť pomocou metodolodógie prezentovanej v príspevku. Personov korelačný koeficient dáva dobrý nástroj pre výpočet korelácií pre indikátorové premenné. Na ďalší hlbší výskum môžeme následne využiť kontingenčné tabuľky už identifikovaných závislostí, tak ako je v článku naznačené.

6. Literatúra

[1] Herzmann J., Novák I., Pecáková I.: Výzkumy veřejného mínění. Skriptum VŠE Praha, Fakulta informatiky a statistiky, 1995.

[2]Chajdiak, J.: Štatistika jednoducho. Statis Bratislava 2003, ISBN 808565928-X.

[3]Chajdiak J.: Štatistické úlohy a ich riešenie v Exceli. STATIS, Bratislava 2005, ISBN 80-85659-39-5.

[4]Kanderová, M. – Úradníček, V.: Štatistika a pravdepodobnosť pre ekonómov. 1. časť. OZ Financ, Banská Bystrica 2007, ISBN 978-80-969535-5-4.

[5]Kanderová, M. – Úradníček, V.: Štatistika a pravdepodobnosť pre ekonómov. 2. časť. OZ Financ, Banská Bystrica 2007, ISBN 987-80-696535-6-1.

[6]Luha J.: Skúmanie súboru kvalitatívnych dát. EKOMSTAT 2003. SŠDS Bratislava 2003.

[7]Luha J.: Matematickoštatistické aspekty spracovania dotazníkových výskumov. Štatistické metódy vo vedecko-výskumnej práci 2003, SŠDS, Bratislava 2003. ISBN 80- 88946-32-8.

[8]Luha J.: Viacrozmerné štatistické metódy analýzy kvalitatívnych znakov. EKOMSTAT 2005, Štatistické metódy v praxi. Trenčianske Teplice 22.–27. 5. 2005. SŠDS Bratislava 2005.

[9] Luha J.: Reprezentatívnosť vo výskumoch verejnej mienky. FORUM STATISTICUM SLOVACUM 2/2005. ISSN 1336-7420. SŠDS Bratislava 2005.

[10] Luha J.: Overovanie reprezentatívnosti výberového súboru. FORUM STATISTICUM SLOVACUM 5/2006. ISSN 1336-7420. SŠDS Bratislava 2006.

[11]Luha J.: Kvótový výber. FORUM STATISTICUM SLOVACUM 1/2007. SŠDS Bratislava 2007. ISSN 1336-7420.

[12]Luha J.: Korelácie vnorených javov. FORUM STATISTICUM SLOVACUM 5/2008. SŠDS Bratislava 2008. ISSN 1336-7420.

[13]Luha J.: Korelácie disjunktných a komplementárnych javov. FORUM STATISTICUM SLOVACUM 6/2008. SŠDS Bratislava 2008. ISSN 1336-7420.

[14]Pecáková I.: Statistické aspekty terénních průzkumů I. Skriptum VŠE Praha 1995.

[15] Príručky SPSS. Ver. 14.

[16]Pecáková I.: Statistika v terénních průzkumech. Proffessional Publishing, Praha 2008. ISBN 978-80-86946-74-0.

[17]Řehák J., Řeháková B.: Analýza kategorizovaných dat v sociologii. Academia Praha 1986.

[18]Řezanková A.: Analýza dat z dotazníkových šetření. Proffessional Publishing, Praha 2007. ISBN 978-80-86946-49-8.

[19]Stankovičová I., Vojtková M.: Viacrozmerné štatistické metódy s aplikáciami. IURA EDITION, Bratislava 2007, ISBN 978-80-8078-152-1.

Adresa autora:

RNDr. Ján Luha, CSc.

Jan.Luha@statistics.sk

Získavanie údajov pre štatistiky www stránok ako súčasť e-commerce
Retrieving data for web statistics as part of e-commerce

Branislav Mišota, Ľuboš Horka, Marián Stenchlák

Abstract: This article is focused on the retrieving data for statistics of web sites as part of e-commerce. Next, it describe possible approaches of retrieving data from http protocol, especially by using log file analysis and page tagging. The work clarifies theoretical bases of the mentioned area.

Key words: statistics of web sites, e-business, log file analysis, page tagging

Kľúčové slová: štatistiky www stránok, e-podnikanie, analýza log súborov, tagovanie stránok

1. Úvod

Získavanie údajov pre štatistiky webu je dôležitý proces predchádzajúci monitoringu správania sa návštevníkov na stránkach. K stavu v akom sa nachádza v súčasnosti prešiel tento proces komplikovaným vývojom. V neustále sa meniacom prostredí internetu sa pomerne jednoduché technologické postupy výrazne zmenili a priniesli aj novú kvalitu pre použitie dát, získaných z prevádzky webu.

Potreba merania návštevnosti web stránok vyplýva z nutnosti riešenia šiestich kľúčových podnikateľských problémov, pri podnikaní v elektronickom svete tzv. e-business. Tieto sú podrobne popísané v [Tab.1].

Tabuľka 2: Riešenie šiestich kľúčových podnikateľských problémov

NÁVŠTEVNÍCI	porozumenie správania anonymných návštevníkov
MARKETING	maximalizácie návratnosti výdavkov (ROI) na marketing
OBCHOD	vyhodnocovanie efektívnosti obchodu
UDRŽANIE ZÁKAZNÍKOV	maximálne zvýšenie vernosti zákazníkov a ich hodnoty (lifetime value)
OBSAH	analýza efektívnosti obsahu
PREDAJNÉ KANÁLY	posudzovanie partnerov

2. Metódy zberu údajov

Na zbieranie štatistických dát potrebných pre meranie návštevnosti webových stránok existuje veľké množstvo rôznych technologických prístupov. Merania sú prevažne vykonávané on-line metódami na rozdiel od off line metód kde sa napríklad zisťuje znalosť respondentov jednotlivých spravodajských portálov. On-line merania možno uskutočňovať nasledovnými spôsobmi:

- merania vykonávané na strane klienta (traffic volume measurement from panels),
- merania vykonávané analýzou TCP / IP paketov (TCP / IP packet sniffing),
- merania vykonané pomocou plug-inu na serveri (direct server measurement)
- merania vykonávané analýzou prevádzky ISP (internet traffic measurement)
- merania vykonávané prostredníctvom log súboru (logfile analysis),
- merania pomocou aktívneho obsahu - tagovanie stránok (page tagging),

V praxi sa hlavne používajú posledné dva spomenuté druhy meraní.

Analýza log súborov (log file analysis) získava dáta zo záznamov (logov), do ktorých web server zaznamenáva všetky svoje transakcie. Webové servery zapisujú všetky transakcie do logu. Už na začiatku 90. rokov vznikali prvé programy na analýzu týchto súborov za účelom získania informácií o popularite jednotlivých serverov. Na začiatku bol jedinou štatistikou počet požiadaviek (hits) na server. Keď neskôr pribudli na stránky obrázky (stiahnutie obrázku je rovnako hit na server) a stránky sa rozdelili na viacero súborov, tento ukazovateľ stratil výpovednú hodnotu.

Zaviedli sa dva nové ukazovatele:

- page views (počet požiadaviek na zobrazenie stránky)
- visits / sessions (počet sekvencií požiadaviek unikátneho klienta v časovom úseku, napr. 30min)

Ďalšie problémy so získavaním dát z logov spôsobilo používanie web proxy, vyrovnávacích cache, a dynamické prideľovanie IP adries. V logoch sú tiež zaznamenané návštevy vyhľadávacích robotov, ktoré treba oddeliť od návštev reálnych používateľov. Tieto záznamy môžu byť na druhej strane vhodne použité pri optimalizácii web stránok pre vyhľadávače.

Tagovanie stránok (page tagging) notifikuje pri každom prístupe na stránku externý monitorovací server pomocou JavaScriptu vloženého v kóde stránky. V súčasnosti vieme získať najpresnejšie informácie o návštevnosti webových stránok využitím služieb niektorého z dostupných monitorovacích serverov :

- Slovenské
  - TopList: www.toplist.sk
  - Naj: www.naj.sk
  - Monitoring návštevnosti: http://www.monitoring-navstevnosti.sk
- Najpoužívanejší celosvetový:
  - Google Analytics: http://www.google.com/analytics/

Tieto služby sa vyvinuli z JavaScript počítadiel návštev. Informácie o návštevníkoch stránky získavajú monitorovacie servery prostredníctvom skriptu umiestneného v HTML kóde monitorovanej stránky. Pri každom zobrazení takto monitorovanej stránky sa monitorovacej agentúre odošlú informácie o prístupe. Pri prvom prístupe na monitorovanú stránku sa môže v prehliadači uložiť tzv. Tracking Cookie, pomocou ktorého je možné sledovať postupnosť podstránok, ktoré daný používateľ na serveri navštívi. Pomocou Tracking Cookie vieme identifikovať používateľa aj medzi návštevami, kedy mu už session expirovala.

3. Záver

Na základe získaných údajov o návštevnosti stránok nasleduje proces vyhodnotenia návštevnosti. Tento proces nazývaný aj ako analýza click-stream dát alebo tzv. e-metriky umožňujú zaznamenávať a vyhodnocovať viac ako len konečnú transakciu kúpy, ale vyhodnocujú každé kliknutie návštevníka.

4. Literatúra

[1]REINER, D.: E-Metrics Solutions For The New Economy: The NetGenesis Enterprise. Architecture. URL: *ftp://ftp.spss.com/pub/web/wp/ArchWhitepaper.pdf*

[2]STUCHLÍK, P., DVOŘÁČEK. Marketing na Internetu. Praha, Grada Publishing 2000.

[3]STUCHLÍK, P., DVOŘÁČEK, M. Reklama na Internetu. Praha, Grada Publishing 2002.

[4]MROSKO, M., MIKLOVIČOVÁ, E., MURGAŠ, J.: Predictive Control Using Genetic Algorithms and Multicriterial Optimization. 2nd International Conference on Advanced Control Circuits & Systems (ACCS'08). Cairo, Egypt: 30.3.-2.4.2008, s. CD Rom

[5]www.toplist.sk

[6]www.naj.sk

[7]http://www.google.com/analytics/

Adresa autorov:

Branislav Mišota, Ing.	Ľuboš Horka, Ing.	Marián Stenchlák, Ing.
Ústav manažmentu STU	Ústav manažmentu STU	Ústav manažmentu STU
Branislav.misota@stuba.sk	Lubos.horka@stuba.sk	Marian.stenchlak@stuba.sk

E-metriky a hodnotenie štatistiky návštevnosti webových stránok. E-metrics and evaluation traffic statistics of web site

Branislav Mišota, Ľuboš Horka, Marián Stenchlák

Abstract: In this article we will attempt to define measurement tools of e-commerce. In e-businesses e-metrics are important assessment of traffic statistics web site. E-metrics make it possible to record and evaluate more than just the final transaction of purchase, but evaluated every click a visitor.

Key words: e-metrics**,** e-commerce, web traffic statistics, visitor of web site

Kľúčové slová: e-metrika, e-commerce, štatistiky návštevnosti webových stránok, návštevníci web stránok

1. Úvod

Tradičné metriky vychádzajúca prevažne z tokov ziskov a strát a ani zďaleka nezobrazujú úplný obraz správania zákazníka aké by sme potrebovali pre popísanie zákazníckeho správania na webe. Tieto metriky v podstate vychádzajúca iba zo zhromažďovania transakcií v čase predaja. V kontraste s tým, metriky pre e-commerce alebo tiež nazývané ako e-metriky ponúkajú príležitosť monitorovať viac ako len konečnú transakciu kúpy. Komplexnosť týchto dát je v tom, že zachytia každé kliknutie návštevníkov, a každú stránku, ktorú vidia. V komerčnom prostredí sa analýza webu zameriava najmä na použitie dát, získaných z prevádzky webovej stránky na určenie toho, ktoré aspekty webovej stránky podporujú obchodné ciele prevádzkovateľa stránky, napr. ktoré podstránky webovej prezentácie prinášajú najviac nákupov.

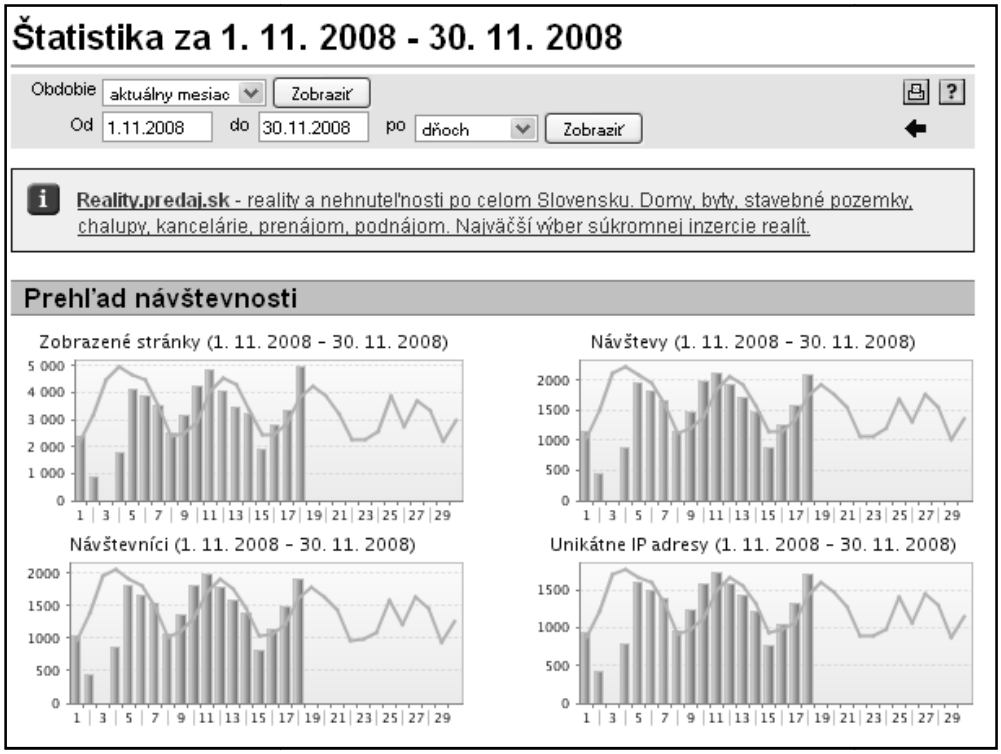
2. Monitorovanie a vyhodnotenie návštevnosti

Cieľom monitorovania návštevnosti webových stránok je hlavne získanie informácií ako:

- Prehľad počtov návštev webovej stránky za obdobie (deň, mesiac, ...)
  - Počet unikátnych návštevníkov za obdobie
  - Počet návštev
  - Počet zobrazených stránok (vrátane podstránok)
  - Počet IP adries
  - Prehľad návštevnosti podľa času v rámci dňa
- Odkiaľ návštevníci prichádzajú
  - Z ktorých vyhľadávačov prichádzajú návštevníci
  - Prehľad odkazov na iných stránkach, z ktorých návštevníci prichádzajú
  - Prehľad vyhľadávacích fráz (kľúčových slov), ktoré návštevníkov privádzajú na našu stránku
- Štatistiky podstránok a sekcií
  - Najnavštevovanejšie stránky na monitorovanom serveri
  - Vstupné stránky (landing pages)
  - Výstupné stránky (exit pages)
  - Najdlhšie zobrazované stránky

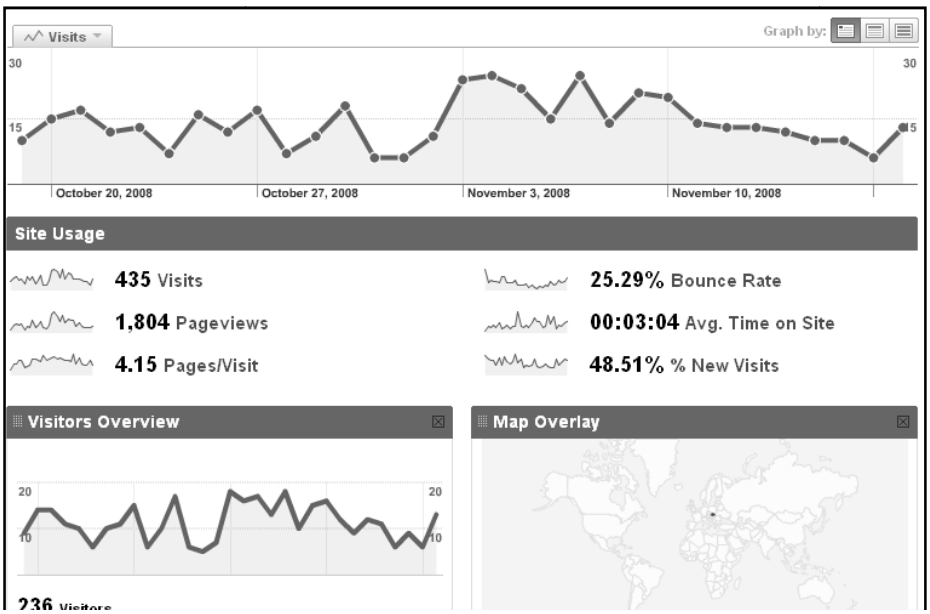
  - Návštevnosť podstránok podľa sekcií
  - Najčastejšie cesty medzi podstránkami
- Prehľad správania sa návštevníkov na stránke
  - Počty návštev na návštevníka (podľa skupín, napr. 1, 2-4, 5-10, ...)
  - Frekvencia návštev
  - HĺbkaHĺbka návštev (počet zobrazených podstránok)
- Odkiaľ návštevníci pochádzajú
  - Prehľad podľa krajiny pôvodu
  - Prehľad podľa domény najvyššej úrovne
  - Prehľad podľa poskytovateľa pripojenia

Na základe získaných údajov o návštevnosti webových stránok je hlavné spracovanie informácií. V tomto procese nám veľakrát pomáhajú vyhodnocovať výsledky nástroje, ktoré sú súčasťou niektorého z dostupných monitorovacích serverov. Napríklad predikcie návštevnosti, ktorú znázorňuje obrázok 1.



Obrázok 1: Predikcie návštevnosti.

Ďalšou užitočnou štatistikou je Bounce Rate, ktorá podáva obraz o správnosti nastavenia marketingovej kampane slúžiacej na propagáciu hodnoteného webu. Hodnota Bounce Rate ako pridaná funkcia serveru Google Analytics sa udáva v percentách. Vyjadruje percento návštevníkov stránky, ktorí sa na stránku dostali náhodou, nesprávnym odkliknutím atď. Prípadne ich stránka, na ktorej sa ocitli prostredníctvom reklamy nezaujala a ihneď stránku opustia. Vysoká hodnota bounce rate znamená, že stránka nezodpovedá dobre reklame alebo odkazu, ktorý na stránku privádza návštevníkov.



Obrázok 2: Sumárny prehľad o návštevnosti v Google Analytics.

3. Záver

Štatistické údaje o návštevnosti stránok, získané z výsledkov monitorovania , je potrebné správne interpretovať a použiť aby sme prostredníctvom nich mohli ďalej zvyšovať návštevnosť našich webových stránok prípadne priamo zvyšovať obrat nášho e-commerce riešenia. E-metriky ponúkajú bohatý prehľad o všetkých udalostiach, ktoré viedli zákazníka k nákupu.

4. Literatúra

[1] REINER, D.: E-Metrics Solutions For The New Economy: The NetGenesis Enterprise. Architecture. URL: ftp://ftp.spss.com/pub/web/wp/ArchWhitepaper.pdf

[2] STUCHLÍK, P., DVOŘÁČEK. Marketing na Internetu. Praha, Grada Publishing 2000.

[3] STUCHLÍK, P., DVOŘÁČEK, M. Reklama na Internetu. Praha, Grada Publishing 2002.

[4] RUBENS, P.: Analyzing Web Analytics.

URL: http://www.aspnews.com/trends/article/0,2350,9921_1492401,00.html

[5] VESELÝ, V.: Metriky a metodologie pro Internet reklamu.

URL: http://www.park.cz/vypis.asp?kod_cl=61

[6] www.naj.sk

[7] http://www.google.com/analytics/

Adresa autorov:

Branislav Mišota, Ing.
Ústav manažmentu STU
Branislav.misota@stuba.sk

Ľuboš Horka, Ing.
Ústav manažmentu STU
Lubos.horka@stuba.sk

Marián Stenchlák, Ing.
Ústav manažmentu STU
Marian.stenchlak@stuba.sk

Optimalizácia webu na základe sekvenčných pravidiel
Web optimalization base on sequence rules

Michal Munk

Abstract: In the article we deal with the possibility of web optimalization base on sequence rules. In the paper we derive sequence rules from common association rules. Consequently, we focus on a distinctive task from web usage mining, particularly for searching sequence rules in data.

Key words: web usage mining, sequence rules, web log files, optimalization

Kľúčové slová: objavovanie znalostí na základe používania webu, sekvenčné pravidlá, logovacie súbory, optimalizácia

1. Úvod

Cieľom objavovania znalostí na základe používania webu (web usage mining) je analýza správania sa používateľa pri prechádzaní webu [1], [2]. Napríklad, zdrojom dát môžu byť automaticky ukladané dáta v logovacích súboroch. V takýchto dátach sledujeme postupnosti – sekvencie v návšteve jednotlivých stránok používateľom, ktorého identifikuje IP adresa. V sekvenciách môžeme hľadať vzorce správania sa používateľov. Za týmto účelom je najlepšie použiť sekvenčnú analýzu, ktorej cieľom je extrakcia sekvenčných pravidiel. Prostredníctvom týchto pravidiel predikujeme sekvencie návštev rôznych webových častí používateľom. Táto metóda bola odvodená z asociačných pravidiel a je príkladom metódy zohľadňujúcej zvláštnosti Internetu.

Cieľom príspevku je preskúmať možnosť optimalizácie webu na základe sekvenčných pravidiel. Aplikáciou nájdených znalostí môžeme prispieť k efektívnejšiemu používaniu portálu. V závere rozoberáme výsledky prvotnej analýzy a naznačujeme ďalší postup, aby nájdené znalosti mohli byť použité pri optimalizácii portálu.

2. Sekvenčné pravidlá

Pri riešení niektorých problémov potrebujeme nájsť asociácie medzi udalosťami, ktoré sa odohrávajú v rôznom čase. Typickým problémom je analýza správania sa používateľa pri prechádzaní webu, ktorá sa aplikuje na automaticky ukladané dáta v logovacích súboroch. Niekedy presný časový údaj o udalosti nie je potrebný, stačí nám iba informácia o tom, že niektorá udalosť predchádzala inej udalosti. V prípade analýzy správania sa používateľa pri prechádzaní webu udalosťou by bola návšteva konkrétnej webovej časti a sekvenciou zoznam navštívených webových častí. Používateľa by sme identifikovali IP adresou a navštívenú webovú časť URL adresou, inak povedané IP adresa by identifikovala sekvenciu a URL adresa by predstavovala udalosť v rámci tejto sekvencie. Časový údaj by v tomto prípade tiež nebol potrebný stačila by nám iba informácia o tom, že niektorá udalosť predchádzala inej udalosti.

Ako sme už spomínali sekvenčné pravidlá sú odvodené od asociačných, rozdiely preto nie sú príliš veľké [3]. *K* - sekvencia je sekvencia dĺžky *k*, t. j. obsahuje *k* stránok. Frekventovaná *k* - sekvencia je obdobou frekventovanej *k* - položkovej množiny, resp. kombinácie. Zvyčajne, frekventovaná 1 - položková množina je rovnaká ako frekventovaná 1 - sekvencia.

Na nasledujúcom príklade naznačíme rozdiely medzi algoritmom Apriori a AprioriAll, kde algoritmus Apriori slúži na hľadanie asociačných pravidiel a AprioriAll na hľadanie sekvenčných pravidiel: $D = \{S1 = \{U1, <A, B, C>\}, S2 = \{U2, <A, C>\}, S3 = \{U1, <B, C,$

E>}, *S*4 = {*U*3, <*A, C, D, C, E*>}}, kde *D* je databáza transakcií s časovou nálepkou, *A, B, C, D, E* sú webové časti a *U*1, *U*2, *U*3 sú používatelia. Každá transakcia je identifikovaná používateľom. Predpokladajme, že minimálna podpora (*min support*) je 30%.

V tomto príklade má používateľ *U*1 aktuálne dve transakcie. Pri hľadaní sekvenčných pravidiel uvažujeme jeho sekvenciu ako aktuálne spojenie webových častí v transakciách *S*1 a *S*3, t. j. sekvencia môže pozostávať z viacerých transakcií, pričom sa nevyžadujú súvislé prístupy na stránky. Rovnako podpora sekvencie je určená nie percentom transakcií, ale percentom používateľov, ktorý majú danú sekvenciu. Sekvencia je veľká/frekventovaná ak sa prinajmenšom nachádza v jednej sekvencii identifikovanej používateľom a spĺňa podmienku minimálnej podpory. Prvým krokom je zoradenie transakcií podľa používateľa s časovou nálepkou každej ním navštívenej stránky, zostávajúce kroky sú podobné ako v algoritme Apriori. Po zoradení sme získali aktuálne sekvencie identifikované používateľom, ktoré predstavujú kompletné odkazy od jedného používateľa: *D* = {*S*1 = {*U*1, <*A, B, C*>}, *S*3 = {*U*1, <*B, C, E*>}, *S*2 = {*U*2, <*A, C*>}, *S*4 = {*U*3, <*A, C, D, C, E*>}}. Podobne ako v algoritme Apriori začíname generovaním množiny kandidátov dĺžky 1: *C*1 = {<*A*>, <*B*>, <*C*>, <*D*>, <*E*>}, z nej určíme množinu *L*1 = {<*A*>, <*B*>, <*C*>, <*D*>, <*E*>} jednoprvkových sekvencií takých, kde každá stránka je odkazovaná prinajmenšom jedným používateľom. Následne množiny kandidátov *C*2 vygenerujeme z *L*1 tzv. úplným spájaním, t. j. zohľadňujeme, že používateľ webu prehľadáva stránky dopredu alebo dozadu. Z tohto dôvodu algoritmus Apriori nie je vhodný na objavovanie znalostí z webových prístupov (web log mining), na rozdiel od algoritmu AprioriAll, ktorý spomínanú skutočnosť zohľadňuje. Z množiny kandidátov dĺžky 2: *C*2 = *{<A, B>, <A, C>, <A, D>, <A, E>, <B, A>, <B, C>, <B, D>, <B, E>, <C, A>, <C, B>, <C, D>, <C, E>, <D, A>, <D, B>, <D, C>, <D, E>, <E, A>, <E, B>, <E, C>, <E, D>*}, určíme množinu *L*2 = *{<A, B>, <A, C>, <A, D>, <A, E>, <B, C>, <B, E>, <C, B>, <C, D>, <C, E>, <D, C>, <D, E>*} dvojprvkových sekvencií takých, kde sa každá sekvencia prinajmenšom nachádza v jednej sekvencii identifikovanej používateľom. Analogicky postupujeme ďalej.

3. Hľadanie sekvenčných pravidiel

Cieľom analýzy je popísať typické správanie sa používateľov anonymného portálu Univerzity Konštantína Filozofa (UKF) v prvý deň akademického roka a zistiť, či nájdené pravidlá odhalia možné nedostatky vyplývajúce z používania portálu UKF.

Dáta sme získali z logovacieho súboru, v ktorom sú zapísané prístupy na portál UKF využívajúci CMS Joomla. Sledované boli prístupy používateľov na jednotlivé webové časti univerzitného portálu v priebehu prvého týždňa akademického roka.

Vytvorená dátová matica obsahuje IP adresu, deň a čas prístupu a ID webovej časti. Každá webová časť má svoje ID, ktoré vzniklo odfiltrovaním z URL adresy. IP adresou identifikujeme používateľov, pričom z dátovej matici je vidieť, v akom poradí navštevovali jednotlivé webové časti portálu. Pracovať budeme s dvoma dátovými maticami (tab. 1), kde prvú sme získali z logovacieho súboru a druhú z mapy webu, kde prostredníctvom ID webovej časti identifikujeme jej obsah.

Tabuľka 1: Dátové matice

IP adresa	Deň	Čas	Web ID
147.175.131.20	25	16:19:20	a211
147.175.131.20	25	16:19:57	a212
147.175.131.20	25	16:20:02	a492
...	...	...	...

Web ID	Obsah
a31	Náš čas
a32	Kontaktné informácie
a33	Mapa objektov UKF
...	...

Na spracovanie spomínaných dátových matíc nepriamo použijeme nástroj určený pre taxonómiu hierarchickej asociačnej analýzy. Tento nástroj nám umožňuje stanoviť tzv.

taxonomické premenné, pričom nemusíme vytvárať dotaz nad spomínanými tabuľkami, medzi ktorými je definované spojenie 1:n.

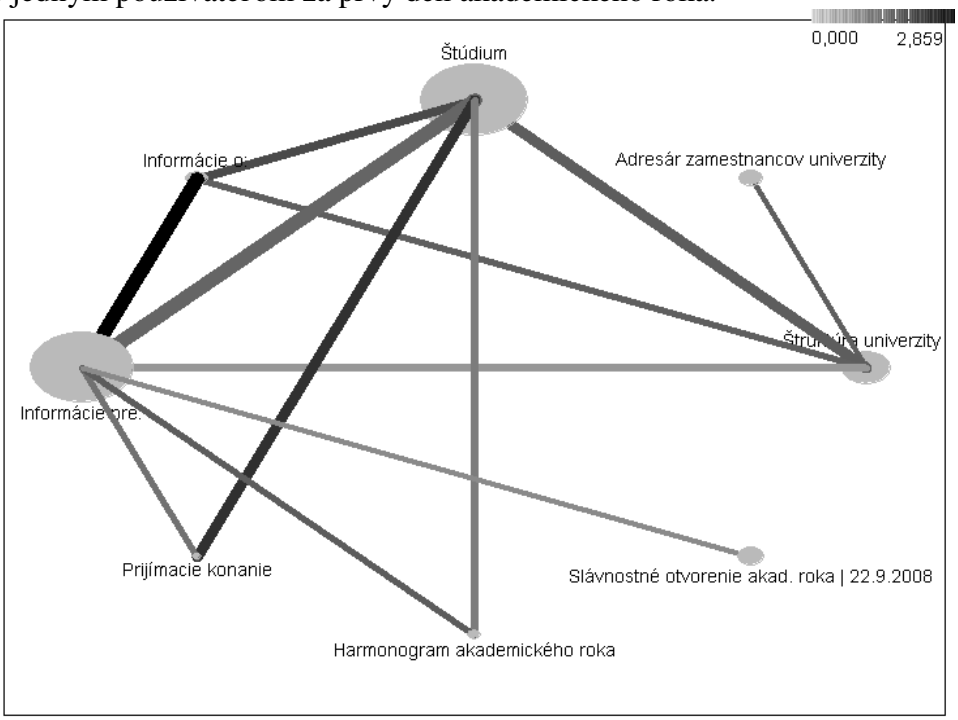
Ešte pred samotnou analýzou je potrebné identifikovať a následne vylúčiť z analýzy prehľadávacie stroje (google, yahoo a pod.) na základe IP adresy, najlepšie z tabuľky početností (tab. 2), kde už samotné vysoké prístupy na nich poukazujú. Ak by sme ich nevylúčili získali by sme skreslené, resp. nezmyselné výsledky v podobe pravidiel. Ďalším problémom sú IP adresy smerovačov, ktoré je tiež nutné identifikovať.

Tabuľka 2: Tabuľka početností pre premennú IP adresa – identifikácia IP adries

IP adresa	Abs. poč.	Rel. poč.	Identifikácia IP adresy
66.249.65.201	665	1,34	crawl-66-249-65-201.googlebot.com
74.6.17.172	490	0,99	llf520198.crawl.yahoo.net
195.80.177.66	167	0,34	router.tekov.sk
...	...	...	...
Celkom	6236	12,60	

Z celkového počtu 49502 prístupov na portál UKF počas prvého týždňa akademického roka sme odstránili 6236 prístupov, čo predstavuje iba 12,6%. Odstránené prístupy boli identifikované ako prehľadávacie stroje, respektíve ako smerovače. Z tabuľky vidíme, že práve prehľadávacie stroje boli identifikované ako najpočetnejšie, čo do počtu prístupov.

V nasledujúcom prípade si najskôr ukážeme nie najvhodnejší nesekvenční prístup. Nebudeme analyzovať sekvencie, ale transakcie, t. j. nezahrnieme do analýzy časovú premennú. Podobne ako v analýze nákupného košíka transakcia predstavuje jeden nákup, v našom prípade bude predstavovať množinu navštívených stránok jedným používateľom. Vzhľadom na naše dáta, budeme za jeden „nákupný kôš“ - transakciu považovať stránky navštívené jedným používateľom za prvý deň akademického roka.

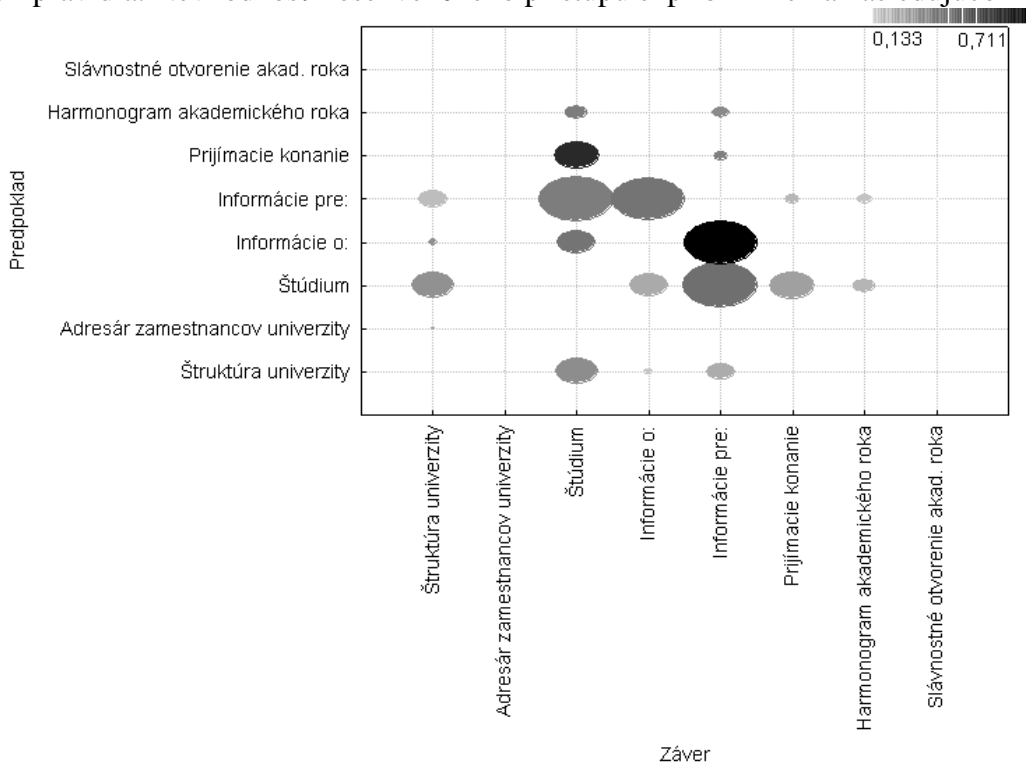


Obrázok 1: Webový graf – vizualizácia nájdených pravidiel

Webový graf (obr. 1) vizualizuje vybrané asociačné pravidlá, konkrétne veľkosť uzla znázorňuje podporu (*support*) prvku, hrúbka čiary podporu pravidla a jas čiary zdvih (*lift*) pravidla. Z predchádzajúceho grafu, ktorý prehľadne popisuje vybrané asociácie, vidíme, že najnavštevovanejšou webovou časťou je *Štúdium* a *Informácie pre:*, rovnako ako aj

kombinácia týchto dvoch častí (*support* = 9,18) alebo napríklad, že sa webové časti *Informácie pre:* a *Informácie o:* vyskytujú častejšie spolu v množinách navštívených stránok jednotlivými používateľmi ako zvlášť (*lift* = 2,86). To isté platí pre webové časti *Štúdium* a *Prijímacie konanie* (*lift* = 2,50). V týchto prípadoch sa dosiahla najväčšia miera zaujímavosti – zdvih, ktorá určuje koľko krát častejšie sa navštívené webové časti vyskytujú spolu, než by to bolo, keby boli štatisticky nezávislé. Vo všetkých prípadoch, okrem dvojici webových častí *Informácie pre:* a *Slávnostné otvorenie akad. roka*, je zdvih väčší ako jedna, t. j. všetky vybrané dvojice sa vyskytujú častejšie spolu ako zvlášť v množine navštívených webových častí jednotlivými používateľmi.

Treba si však uvedomiť, že pri charakteristike zaujímavosti – zdvih nezáleží na orientácii pravidla. Nevhodnosť nesekvenčného prístupu si priblížime na nasledujúcom grafe.



Obrázok 2: Graf pravidiel – vizualizácia nájdených pravidiel

Graf pravidiel (obr. 2) vizualizuje vybrané asociačné pravidlá, kde veľkosť uzla znázorňuje podporu (support) pravidla a jas uzlu spoľahlivosť (confidence) pravidla.
Pri spoľahlivosti už záleží na orientácii pravidla, vezmime do úvahy pravidlá, kde bola dosiahnutá najväčšia miera zaujímavosti.

Z grafu vidíme, že pravidlo *Informácie o:* ==> *Informácie pre:* (*confidence* = 71,07), má väčšiu spoľahlivosť ako jeho opačná orientácia - pravidlo *Informácie pre:* ==> *Informácie o:* (*confidence* = 36,29). Pravdepodobnosť výskytu webovej časti *Informácie pre:* za podmienky, že sa webová časť *Informácie o:* v množine navštívených stránok vyskytuje je podstatne väčšia ako v prípade podmienenej pravdepodobnosti opačne orientovaného pravidla. Z toho sa dá usudzovať, že je tu pomerne významná časť používateľov, ktorí nenájdu potrebné informácie pod odkazom *Informácie o:*, respektíve chýba tu jednoznačné rozlíšenie týchto dvoch webových častí. Takto môžeme usudzovať v prípade webových častí na rovnakej úrovni. Nezmyselné výsledky prinášajú nájdené pravidlá pozostávajúce z webových častí na rôznej úrovni, napríklad pravidlo *Prijímacie konanie* ==> *Štúdium* (*confidence* = 63,11) bude mať podstatne väčšiu spoľahlivosť ako opačne orientované pravidlo *Štúdium* ==> *Prijímacie konanie*, (*confidence* = 26,97), vzhľadom na to, že väčšina

používateľov sa dostane na webovú časť *Prijímacie konanie* cez webovú časť na vyššej úrovni *Štúdium*. Z čoho vyplýva logický záver, že ak sa v množine navštívených stránok vyskytuje webová časť *Prijímacie konanie*, potom sa bude s vysokou pravdepodobnosťou vyskytovať aj webová časť *Štúdium*, ktorá je na hierarchicky vyššej úrovni a väčšina používateľov cez ňu prechádza na webovú časť *Prijímacie konanie*.

V nasledujúcej časti popíšeme sekvenčnú analýzu, ktorú môžeme považovať za validnú, vzhľadom na charakter dát. Zahrnieme do analýzy časovú premennú, t. j. zohľadníme poradie v akom jednotliví používatelia navštevovali jednotlivé webové časti portálu univerzity v rámci sledovaného obdobia. Výsledkom analýzy môžu byť sekvenčné pravidlá (tab. 3), ktoré získame z frekventovaných sekvencií spĺňajúcich minimálnu podporu (v našom prípade *min support* = 0,03). Frekventované sekvencie sme získali z 1907 identifikovaných sekvencií, t. j. návštev jednotlivých používateľov portálu UKF za prvý deň akademického roka.

Tabuľka 3: Tabuľka vybraných sekvenčných pravidiel

Predpoklad	==>	Záver	Podpora(%)	Spoľahlivosť(%)
(Študentské domovy)	==>	(Rozhodnutie ubytovacej komisie)	3,20	47,66
(Informácie o:)	==>	(Informácie pre:)	3,15	38,46
(Informácie pre:)	==>	(Informácie pre:)	7,39	37,11
(Informácie pre:)	==>	(Informácie pre študentov UKF)	5,72	28,68
(Adresár zamestnancov univerzity)	==>	(Adresár zamestnancov univerzity)	3,09	27,70
(Štúdium)	==>	(Štúdium)	4,46	21,09
(Štúdium)	==>	(Možnosti štúdia na UKF v 09/10)	4,14	19,60
(Štúdium)	==>	(Prijímacie konanie)	4,14	19,60
(Štúdium)	==>	(Akreditované študijné programy)	3,51	16,63

Všimnime si najskôr pravidlo z najvyššou spoľahlivosťou (*confidence* = 47,66) - používateľ, ktorý navštívil webovú časť *Študentské domovy*, navštívi následne taktiež webovú časť *Rozhodnutie ubytovacej komisie*, treba však podotknúť, že takéto pravidlo má skôr sezónny charakter a predstavuje vzorec správania sa používateľov v prvých dňoch akademického roka. Druhú najvyššiu spoľahlivosť malo pravidlo *Informácie o:* ==> *Informácie pre:*, t. j. ak používateľ navštívil webovú časť *Informácie o:* s viac ako 38% pravdepodobnosťou navštívi webovú časť *Informácie pre:*.

Zaujímavými znalosťami sú nájdené pravidlá, ktoré odhaľujú nepresnosti a môžu byť podkladom pre optimalizáciu webu. Napríklad pravidlo *Informácie pre:* ==> *Informácie pre:*, ktoré ma najväčšiu podporu a tretiu najvyššiu spoľahlivosť. Neznamená to v tomto prípade, že ak používateľ klikne na webovú časť *Informácie pre:*, tak na ňu klikne s 37% pravdepodobnosťou znovu, ale že sa na nižšej úrovni nachádza webová časť s rovnakým názvom po výbere, ktorej nám neponúka ďalšie informácie. Zo sekvenčnej analýzy naviac vyplýva, že ak používateľ navštívi webovú časť *Adresár zamestnancov univerzity*, takmer s 28% pravdepodobnosťou klikne na názov tejto webovej časti, ktorý sa javí ako odkaz, ale samozrejme neponúka ďalšie informácie. Tieto nedostatky vyplývajú z používaného CMS Joomla. Predpokladáme, že pod názvom webovej časti *Adresár zamestnancov univerzity* hľadajú mnohí používatelia zoznam zamestnancov s kontaktnými údajmi, vzhľadom na to, že spomínaná webová časť obsahuje iba možnosť vyhľadávania zamestnancov.

Nájdené pravidlá predstavujú vzorce správania sa používateľov univerzitného portálu počas prvého dňa akademického roka, prostredníctvom ktorých môžeme predikovať návštevy rôznych webových častí, respektíve optimalizovať portál, napr. objavením zavádzajúcich odkazov a pod.

4. Záver

Prostredníctvom sekvenčnej analýzy môžeme odhaliť zaujímavé vzorce správania sa používateľov rôznych portálov/systémov. Dokážeme popísať, ktoré webové časti jednotliví používatelia navštívili za sledované obdobie, a v akom poradí sa návštevy uskutočnili.

Časť nájdených znalostí nám odhalila nedostatky na portáli UKF, ktoré vyplývajú aj z používaného CMS Joomla. Odstránením identifikovaných nedostatkov môžeme prispieť k efektívnejšiemu používaniu portálu.

Druhá časť pravidiel nám odhaľuje vzorce správania sa používateľov pri prechádzaní portálu. Dôležité si je uvedomiť, že niektoré nájdené pravidlá môžu mať sezónny charakter, vzhľadom na špecifické obdobie. Príkladom takéhoto pravidla je pravidlo, že ak používateľ navštívi webovú časť *Študentské domovy*, potom s takmer 50% pravdepodobnosťou navštívi webovú časť *Rozhodnutie ubytovacej komisie*, treba však podotknúť, že takéto pravidlo predstavuje vzorec správania sa používateľov v prvých dňoch akademického roka. Používatelia pri prechádzaní portálu sa budú pravdepodobne správať inak počas semestra, a inak počas skúšobné obdobie, resp. na začiatku akademického roka, cez zápočtový týždeň, v období podávania prihlášok na vysokú školu a pod.

Navrhujeme nasledovný postup:

1. získavanie dát - definovať logovací modul z pohľadu získania potrebných dát (IP adresa, dátum a čas prístupu, URL adresa, protokol, webový prehliadač a pod.),
2. príprava dát – vytvorenie dátových matíc z logovacieho súboru (informácie o prístupoch) a mapy webu (informácie o obsahu webu),
3. analýza dát - pri analýze zohľadňovať faktor sezónnosti a analýzu realizovať v jednotlivých časových obdobiach,
4. optimalizácia webu – odstránenie identifikovaných nedostatkov vyplývajúcich z používania portálu a predikovanie správania sa používateľov.

5. Literatúra

[1]BERKA, P. 2003. Dobývání znalostí z databází. Praha: Academia, 2003. 392 s. ISBN 80-200-1062-9.

[2]SRIVASTAVA, J. - COOLEY, R. - DESHPANDE, M. - TAN, P. 2000. Web usage mining: discovery and applications of web usage patterns from web data. In: SIGKDD Explorations, č. 1, 2000, s. 12 - 23.

[3]WANG TONG - HE PI-LIAN. 2005. Web Log Mining by an Improved AprioriAll Algorithm. In: Engineering and Technology, č. 4, 2005, s. 97-100.

Adresa autora:

Michal Munk, PhD., RNDr.
Tr. A. Hlinku 1
949 74 Nitra
mmunk@ukf.sk

Time-series and causality
Časové rady a kauzalita

Oľga Nánásiová, Martin Kalina, Mária Bohdalová

Abstrakt: V príspevku sa venujeme metóde na analýzu časových radov, ktorá je priamou aplikáciou teórie kvantových logík. V metóde sme použili nesymetrickú kovariančnú maticu, ktorá je odvodená z teórie o s-zobrazeniach (funkcia pre súčasnú merateľnosť) a umožňuje uvažovať kauzalitu v časových radoch. Na reálnych a generovaných údajoch sme ukázali použitie tejto metódy.

Kľúčové slová: kvantová logika, pozorovateľná, lineárny model, časový rad, s-zobrazenie

Key words: Quantum logic, observable, linear model, time series, s-map

1. Introduction

Covariance is a non-normalized measure of dependence, and as such it is assumed to be symmetric. I.e., if ξ and η are two random variables, then cov (ξ, η) = cov (η, ξ). However, there are some generalized models of sample spaces, called orthomodular lattices (see. e.g. [11]), where we might get

$$\text{cov}\,(\xi, \eta) \neq \text{cov}\,(\eta, \xi). \tag{1}$$

In such a case we say that ξ and η are incompatible. Now, we are not going to present the teoretical background for this theory. Rather we would like to show how it works on real data. Readers interested in this theory we refer to papers, e.g. [1, 3, 4, 6-12]. In the next part we just present the modification of the MinQE estimator for the case when (1) occurs.

2. Time series and quantum logic

When considering time-series, it is quite natural to assume that some event A, occurring at a time-instant t_1, may influence another event B occurring at a time-instant $t_2 > t_1$, but not vice-versa. In such a case we say that A is a cause and B its consequence. If we consider a regression model with some symmetric covariance matrix $\mathbf{\Sigma}$ of the vector of residuals ε, $\mathbf{Y}$ the response vector and $\mathbf{X}$ the design matrix, then the MinQE estimator, $\hat{\beta}$, of the vector of parameters is given by

$$\hat{\beta} = (\mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{Y} \tag{2}$$

In a causal system where we cannot assume that the covariance matrix $\mathbf{\Sigma}$ is symmetric, we must modify the above estimator into the following form:

$$\tilde{\beta} = (\mathbf{X}^T (\mathbf{\Sigma}^{-1/2})^T \mathbf{\Sigma}^{-1/2} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{\Sigma}^{-1/2})^T \mathbf{\Sigma}^{-1/2} \mathbf{Y} \tag{3}$$

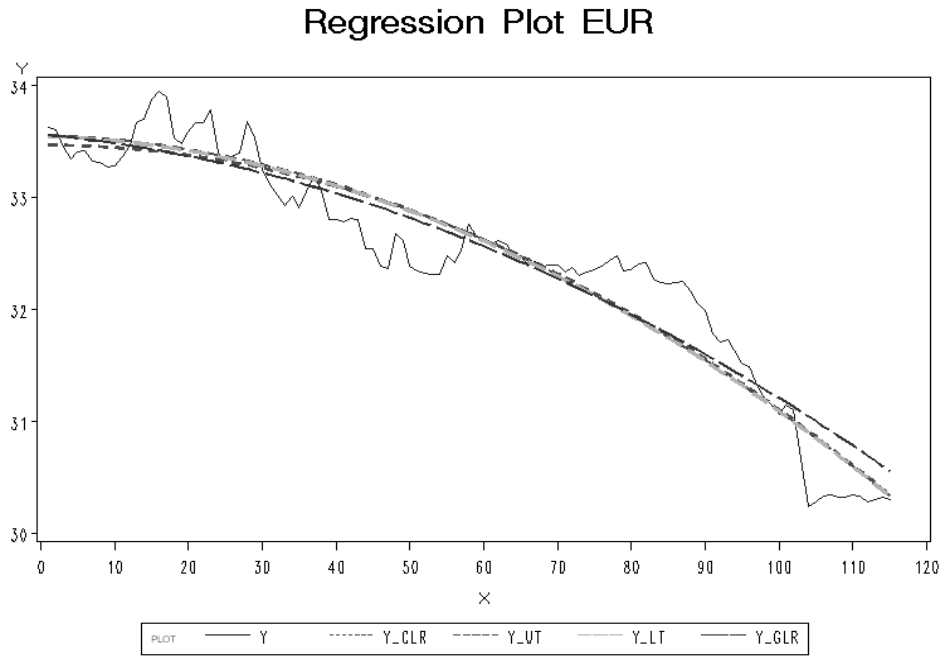
where the matrix $(\mathbf{\Sigma}^{-1/2})^T \mathbf{\Sigma}^{-1/2}$ is symmetric, but $(\mathbf{\Sigma}^{-1/2})^T \mathbf{\Sigma}^{-1/2} \neq \mathbf{\Sigma}^{-1}$.

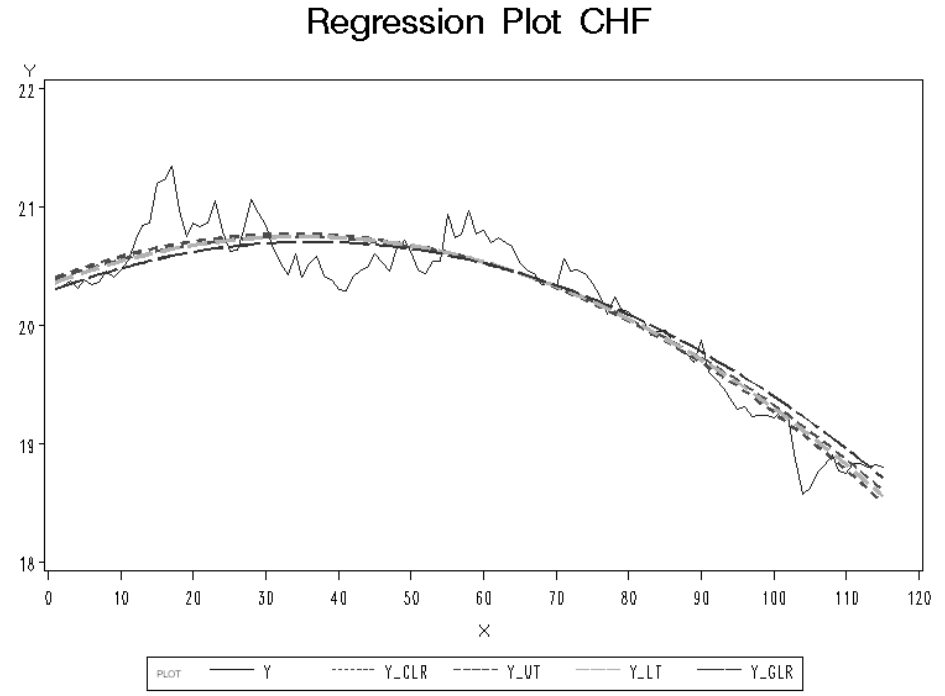
In what follows we have used the above estimator (3) in two ways:

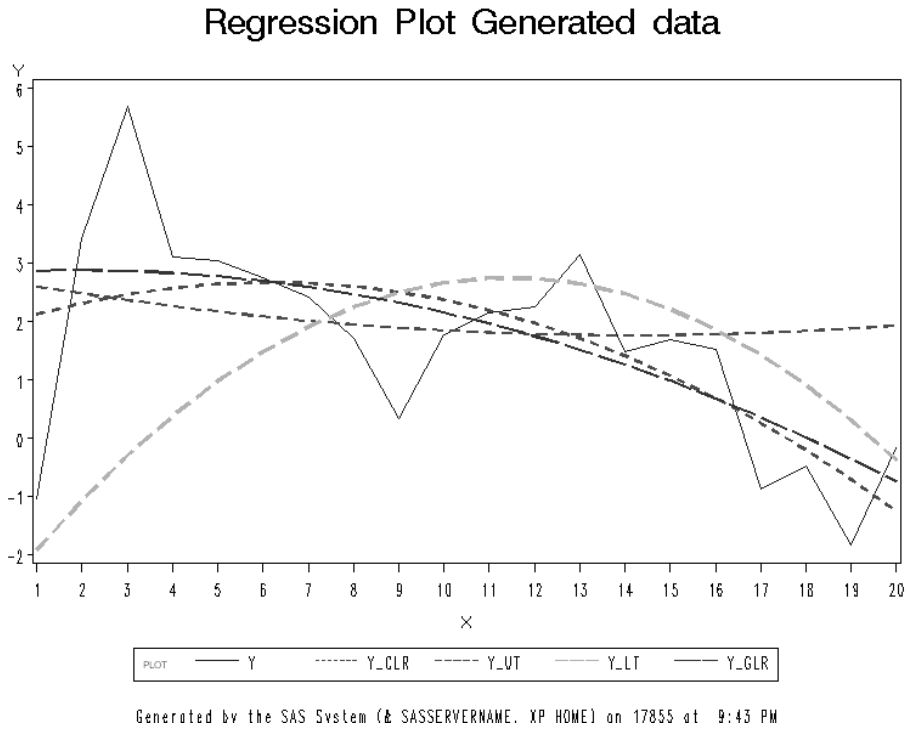
- we assumed $\mathbf{\Sigma}$ to be the upper triangle-matrix (*UT*),
- we assumed $\mathbf{\Sigma}$ to be the lower triangle-matrix (*LT*).

This method was programmed in the system SAS. As test-data we took the exchange rates of Euro (EUR) and Switzerland frank (CHF) to Slovak Crown, see www.nbs.sk. Finally we used also generated data of an AR(2) process with a parabolic trend. In each case we compared the above two models having triangle covariance matrix with the classical one (*CLR*), having its

covariance matrix $\Sigma = \sigma^2 I$ and with one having symmetric covariance matrix Σ (*GLR*). We assumed a parabolic trend in our considerations. The results we can see on the next pictures.







3. Conclusion

From the last picture we can conclude that the model with the upper triangle covariance matrix tends better to copy the trend of the end of the time-interval in question, the model with the lower triangle covariance matrix tends better to copy the trend of the beginning of our time-interval, while the model with the full covariance matrix tends more to copy the trend in the middle part of the time-interval. However, it needs more investigation to decide wether this hypothesis is true.

4. Acknowledgements

The work on this paper has been supported by Science and Technology Assistance Agency under the contract No. APVV-0375-06, and by the VEGA grant agency, grant numbers 1/3014/06, 1/4024/07 and 1/0373/08.

5. References

[1] Beltrametti, E., Cassinelli, G. 1981. The logic of quantum mechanics. Addison-Wesley; Reading, Mass.

[2] Cipra, T. 1981. Analýza časových řad s aplikacemi v ekonómii. SNTL, Praha.

[3] Dohnal, G. 2008. Markov Property in Quantum Logic. A reflection, Inf.Sci. in press

[4] Dvurečenskij, A., Pulmannová S. 2000. New Trends in Quantum Structures. Kluwer Acad. Publ.

[5] Anděl, J. 1978. Matematická statistika. SNTL, Praha

[6] Gudder, S., P. 1969. Quantum probability spaces. AMS, 21, 296-302.

[7] Khrennikov, A., Yu. 2003. Contextual viewpoint to quantum stochastic, J. Math. Phys., 44, 2471-2478.

[8] Nánásiová, O. 1987 On conditional probabilities on quantum logic. Int. Jour. of Theor. Phys., 25, 155-162.

[9] Nánásiová, O. 2004. Principle conditioning, Int. Jour. of Theor. Phys., 43, 1383-1395.

[10] Nánásiová, O. 2003. Map for Simultaneous Measurements for a Quantum Logic, Int. Journ. of Theor. Phys., 42, 1889-1903.

[11] Nánásiová O., Khrennikov A. 2006. Representation theorem for observables on quantum system. Int. J. Theoret. Phys. 45 , 481-494.

[12] Varadarajan V. 1968. Geometry of quantum theory. Princeton, NewJersey, D. Van Nostrand.

Address authors :

Oľga Nánásiová, doc.,
SvF STU
Radlinského 11
813 68 Bratislava
nanasiova@math.sk

Martin Kalina, doc.,
SvF STU
Radlinského 11
813 68 Bratislava
kalina@math.sk

Mária Bohdalová, PhD.
FM UK
Odbojárov 10
820 05 Bratislava
maria.bohdalova@fm.uniba.sk

Konkurenční výhoda ekonomik zemí EU
Competitive advantage of economics of EU countries

Jakub Odehnal, Marek Sedlačík, Jaroslav Michálek

Abstract: The goal of the contributions is the application of multivariate classification techniques to assessment of competitive advantage of chosen countries of the European Union. Used data set was obtained from the Knowledge Assesment Matrix published by the World Bank. There are variables: annual GDP growth, human development index, tariff barriers, regulatory quallity, rule of law, royalty payments and receipts, technical journal articles, patents granted by USPTO, adult literacy rate, gross secondary enrollment, gross tertiary enrollment, total telephones, computers and internet users. Result of classification are two groups of countries of European Union: traditional contries and new member countries.

Key words: competitive advantage, knowledge-based economy, cluster analysis, factor analysis

Klíčová slova: konkurenční výhoda, znalostní ekonomika, shluková analýza, faktorová analýza

1. Úvod

Konkurenční výhoda založená na tvorbě inovací a kreativitě umožňuje dosahování dlouhodobého růstu konkurenceschopnosti ekonomik sledovaných zemí a regionů. Přirozené rozdíly mezi zeměmi a regiony odpovídající odlišnému faktorovému vybavení tak umožňují identifikovat konkurenční výhodu zemí a regionů a sledovat tak přechod ekonomik od konkurenční výhody pramenící z levného faktorového vybavení zemí a regionů ke znalostní ekonomice. Identifikace znalostně založené konkurenční výhody je založena na výsledcích šetření Světové banky Knowledge Assessment Matrix, které hodnotí konkurenční výhody vybraných ekonomik prostřednictvím složek indexu znalostní ekonomiky [2]. Cílem příspěvku je klasifikovat země EU na základě proměnných indexu znalostní ekonomiky a analyzovat tak rozdíly mezi proměnnými identifikované vzniklou klasifikací.

2. Zdroje dat a jejich charakteristika

Index znalostní ekonomiky prezentovaný Světovou bankou [4] sleduje a hodnotí zdroje konkurenční výhody založené na znalostech a inovacích pro vybraných 140 národních ekonomik ve 4 základních oblastech (ekonomický systém, inovační systém, systém vzdělání a kvalita lidských zdrojů, přístup k informačním a komunikačním technologiím). Ze zastoupených 4 oblastí bylo vybráno 14 základních proměnných reprezentujících složky indexu znalostní ekonomiky prezentovaného v [4]. Data byla v rámci původního šetření Světovou bankou upravena tak, že pro další analýzu nabývají standardizovaných hodnot na škále 0 až 10. Použité proměnné jsou: a) průměrný roční růst HDP, b) index lidského rozvoje, c) existence obchodních bariér, d) kvalita regulace, e) kvalita právního řádu, f) licenční politika, g) počty odborných článků ve vědeckých publikacích na mil. obyvatel, h) počty přijatých patentů a zlepšení na mil. obyvatel, i) vzdělanostní úroveň obyvatelstva, j) podíl zapsaných studentů v soustavě sekundárního vzdělávání, k) podíl zapsaných studentů v soustavě terciárního vzdělávání, l) celkový počet telefonních přístrojů na tis. obyvatel, m) celkový počet osobních počítačů na tis. obyvatel, n) počet uživatelů internetového připojení na tis. obyvatel.

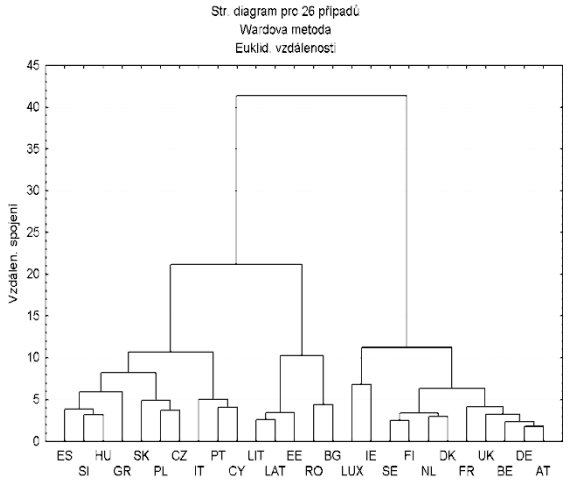
Země vybrané ke klasifikaci tvoří 26 členských zemí Evropské unie s výjimkou Malty, jejíž data o znalostní konkurenční výhodě nebyla v rámci šetření publikována. Pro zpracování dat byly použity mnohorozměrné statistické metody: shluková analýza k vytvoření klasifikace zemí a faktorová analýza pro redukci počtu proměnných.

3. Použité metody

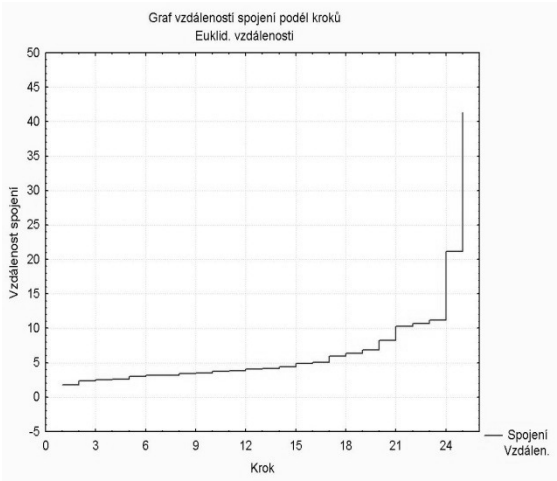
Cílem zvolené shlukové analýzy je nalézt skupiny podobných objektů, konkrétně v případě uvažované klasifikace zemí Evropské unie dle charakteristik indexu znalostní ekonomiky klasifikujeme 26 objektů prostřednictvím 14 proměnných. Ve shlukovacím algoritmu používáme Wardovu metodu. Tato metoda vybírá takové objekty ke sloučení, kde je minimální součet čtverců odchylek všech hodnot od příslušných shluků [3]. V další části práce provedeme redukci dat pomocí faktorové analýzy [1]. Tímto způsobem vytvořená klasifikace je následně porovnána s výsledky shlukové analýzy před a po redukci počtu proměnných. Srovnáním tak získáme informaci o vzniklých rozdílech v klasifikaci zemí EU prostřednictvím 14 proměnných indexu znalostní ekonomiky a pomocí vytvořených 3 faktorů znalostní ekonomiky.

4. Shluková analýza zemí EU užitím hodnot složek indexu znalostní ekonomiky

Výsledek mnohorozměrné klasifikace použitím hierarchického shlukování [3] (Wardova metoda, euklidovská vzdálenost) zobrazuje dendrogram na obrázku 1.



Obrázek 2: Výsledek shlukování



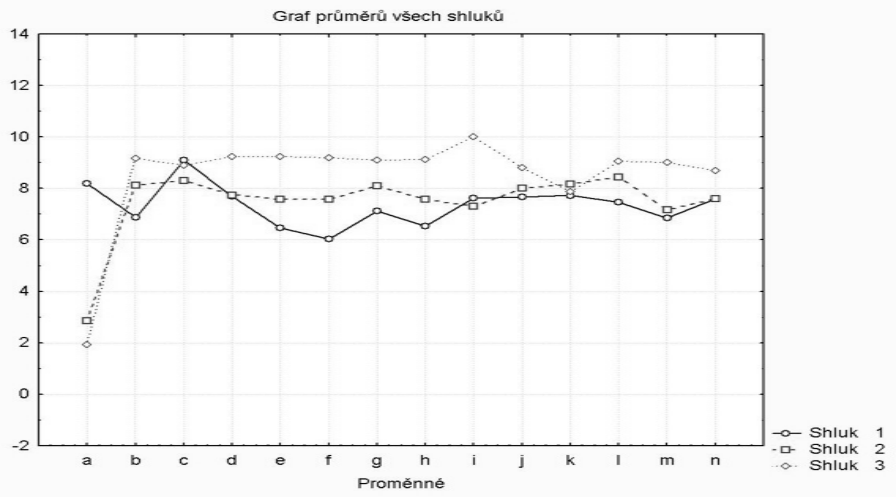
Obrázek 2: Graf vzdálenosti spojení

Z charakteru větvení a s přihlédnutím k výsledkům rozvrhu shlukování (obrázek 2) byly identifikovány tři shluky obsahující hodnocené země EU. První skupina zemí zahrnuje 11 zemí (Lucembursko, Irsko, Švédsko, Finsko, Nizozemí, Dánsko, Francie, Velká Británie, Belgie, Německo, Rakousko), tedy země uchovávající si oproti zbývajícím shlukům patrný náskok v dílčích proměnných indexu znalostní ekonomiky, který se projevuje zejména ve vyšších hodnotách proměnných inovačního charakteru značícího existenci konkurenční výhody založené na inovacích jako zdroje růstu konkurenceschopnosti těchto národních ekonomik.

Druhá skupina obsahuje 5 nových členských zemí EU (Lotyšsko, Litva, Estonsko, Rumunsko, Bulharsko), jejichž hodnoty u sledovaných proměnných charakterizujících možný přechod na znalostní ekonomiku oproti tradičním zemím EU spíše zaostávají.

Třetí skupina, kterou tvoří 10 zemí (Španělsko, Slovinsko, Maďarsko, Řecko, Slovensko, Polsko, Česká republika, Itálie, Portugalsko, Kypr), je složená jak z tradičních členů EU, jejichž hodnoty proměnných však nedosahují nadprůměrných hodnot jako v případě první skupiny, tak z ostatních nových členských zemí, které se však oproti druhé

skupině identifikované jako nejvíce inovačně zaostalé výrazně v jednotlivých proměnných odlišují. Detailní rozdíly mezi vytvořenými shluky demonstruje graf průměrů tří shluků (obrázek 3) charakterizující průměrné hodnoty proměnných v každém shluku.



Obrázek 3: Graf průměrů shluků

Z grafu je dobře patrný rozdíl mezi hodnotami proměnných u tří identifikovaných shluků, přičemž pořadí shluků odpovídá jejich postavení v rámci sledovaných hodnot indexu znalostní ekonomiky. Výsledky tak potvrzují existenci rozdílů zejména u proměnných[3] inovačního a znalostního charakteru (e, f, g, h) u tradičních zemí Evropské unie (shluk 3) a u „nových členských zemí“ (shluk 1 a 2). Odlišná situace je patrná v případě tempa růstu HDP (proměnná a), kde „nové“ členské země dosahují vyššího tempa růstu potvrzujícího vzájemnou konvergenci národních ekonomik zemí EU.

V další části textu provedeme alternativní klasifikaci zemí EU prostřednictvím faktorů znalostní ekonomiky a srovnáním výsledků původní klasifikace pomocí 14 proměnných a alternativní klasifikace identifikujeme vzniklé rozdíly ve výsledcích dílčích klasifikací.

5. Alternativní klasifikace zemí EU pomocí faktorů znalostní ekonomiky

K provedení alternativní klasifikace vybraných zemí EU byla použita faktorová analýza, umožňující redukci počtu proměnných z původních 14 na výsledné 3 faktory znalostní ekonomiky. Bližší charakteristiku výsledků faktorové analýzy popisuje tabulka 1 charakterizující faktorové zátěže jednotlivých proměnných. Tučně jsou znázorněny zátěže s hodnotou větší než 0,5. Počet vytvořených faktorů byl zvolen s přihlédnutím k výsledkům sutinového grafu a k procentu vyčerpané variability.

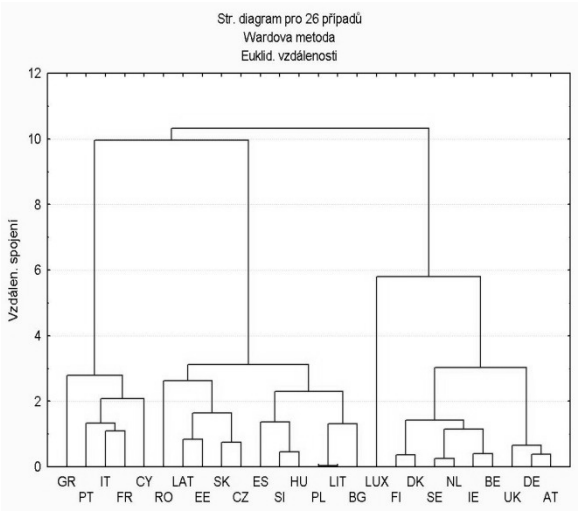
3

a) průměrný roční růst HDP, b) index lidského rozvoje, c) existence obchodních bariér, d) kvalita regulace, e) kvalita právního řádu, f) licenční politika, g) počty odborných článků ve vědeckých publikací na mil. obyvatel, h) počty přijatých patentů a zlepšeních na mil. obyvatel, i) vzdělanostní úroveň obyvatelstva, j) podíl zapsaných studentů v soustavě sekundárního vzdělávání, k) podíl zapsaných studentů v soustavě terciárního vzdělávání, l) celkový počet telefonních přístrojů na 1000 obyvatel, m) celkový počet osobních počítačů na tis. obyvatel, n) počet uživatelů internetového připojení na tis. obyvatel

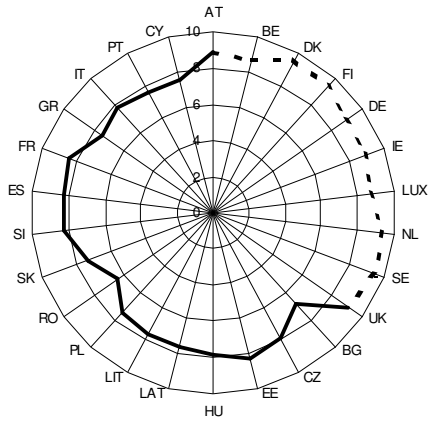
Tabulka 3: Matice faktorových zátěží

	Bez rotace			varimax normalizovaný		
	Faktor 1	***Faktor 2***	***Faktor 3***	**Faktor 1**	**Faktor 2**	**Faktor 3**
a) průměrný roční růst HDP,	***0,662594***	*-0,427970*	*0,276929*	-0,567132	**0,589202**	-0,173436
b) index lidského rozvoje,	***-0,913215***	*0,251814*	*-0,010738*	**0,804515**	-0,329829	0,376104
c) existence obchodních bariér,	*-0,042366*	***-0,759516***	*0,448263*	0,158643	**0,868571**	0,003905
d) kvalita regulace,	***-0,844767***	*-0,281863*	*-0,021452*	**0,879311**	0,115627	0,083553
e) kvalita právního řádu,	***-0,929865***	*-0,070209*	*-0,003151*	**0,902270**	-0,060698	0,227625
f) licenční politika,	***-0,871290***	*-0,025712*	*-0,154817*	**0,861488**	-0,172755	0,108487
g) počty odborných článků ve vědeckých publikací na mil. obyvatel,	***-0,893022***	*0,190829*	*0,291849*	**0,748462**	-0,112040	0,588499
h) počty přijatých patentů a zlepšeních na mil. obyvatel,	***-0,958525***	*0,029717*	*-0,040652*	**0,910193**	-0,167601	0,254491
i) vzdělanostní úroveň obyvatelstva,	***-0,669336***	*-0,468419*	*0,213423*	**0,719989**	0,420293	0,133961
j) podíl zapsaných studentů v soustavě sekundárního vzdělávání,	***-0,539856***	*0,389264*	*0,470907*	0,330432	-0,134674	**0,733077**
k) podíl zapsaných studentů v soustavě terciárního vzdělávání,	*-0,130025*	*0,400259*	*0,791634*	-0,117625	0,082401	**0,884970**
l) celkový počet telefonních přístrojů na 1000 obyvatel,	***-0,702379***	*0,208157*	*-0,313701*	**0,668161**	-0,431904	0,045752
m) celkový počet osobních počítačů na tis. obyvatel,	***-0,861905***	*-0,305766*	*-0,100705*	**0,915561**	0,090165	0,011555
n) počet uživatelů internetového připojení na tis. obyvatel	***-0,707707***	*-0,262509*	*-0,119093*	**0,761055**	0,063717	-0,026059
% vysvětlené variability	*0,558707*	*0,120189*	*0,100409*	0,515531	0,121300	0,142474

První část tabulky zobrazuje výsledek extrakce metody hlavních komponent bez rotace, která se jeví jako méně jasná z hlediska korelační struktury mezi faktory a proměnnými. Druhá část tabulky zobrazuje výsledek rotace pomocí metody varimax normalizovaný, vedoucí k přehlednější korelační struktuře. První faktor získaný rotací má vysoké hodnoty zátěže (v tabulce znázorněny tučně) v případě proměnných (b, d, e, f, g, h, i, l, m, n,) , druhý faktor v případě proměnných (a, c) a u třetího faktoru pozorujeme vysoké hodnoty zátěže u proměnných (j, k). Dle vlastností sledovaných proměnných tak jednotlivé faktory znalostní ekonomiky můžeme slovně popsat: faktor 1 – faktor institucionálních a inovačních charakteristik, faktor 2 – faktor ekonomických charakteristik a faktor 3 – faktor vzdělanostních charakteristik.



Obrázek 4: Výsledek shlukování po redukci počtu proměnných



Obrázek 5: Hodnoty indexu znalostní ekonomiky

Výsledek klasifikace zemí EU na základě faktorového skóre zjištěného z výsledků faktorové analýzy zobrazuje dendrogram na obrázku 4. Porovnáním s obrázkem 1 tak pozorujeme potvrzení identifikovaných rozdílů v inovační úrovni sledovaných zemí.

Z dominantního shluku tradičních zemí EU identifikovaného na obrázku 1 došlo ke změně v pozici Francie, která se tak zařadila mezi zbylé země EU. O rozdílných úrovních inovačních charakteristik jednotlivých zemí Evropské unie se můžeme přesvědčit i z obrázku 5 znázorňujícího průměrné hodnoty indexu znalostní ekonomiky zemí EU. Čárkovaně jsou znázorněny země klasifikované pomocí shlukové analýzy do shluku 3, tedy na obrázku 5

země s nejvyššími průměrnými hodnotami sledovaného indexu. Na první pohled překvapivému postavení některých tradičních členských zemí EU (Řecka, Portugalska, Itálie, Francie a Španělska) ve shluku nových zemí EU odpovídá i nižší průměrná hodnota souhrnného indexu oproti ostatním tradičním zemím EU. Obdobný závěr je patrný i z postavení v žebříčku 140 sledovaných ekonomik publikovaného Světovou bankou [4], ve kterém země shluku 3 předčí země ostatních identifikovaných shluků. Dodatečným srovnáním hodnot indexu znalostní ekonomiky Řecka, Portugalska, Itálie, Francie a Španělska z roku 2005 s rokem 1995 získáváme informaci o snížení hodnot indexu u sledovaných zemí a také o poklesu umístění v rámci sestaveného žebříčku. V případě Řecka o 5 míst, v případě Portugalska o 7 míst, v případě Itálie o 2 místa, v případě Francie o 4 místa a Kypru o 1 místo. Zařazení těchto zemí do společného shluku s novými členskými zeměmi EU tak odpovídá jejich aktuální situaci u sledovaných faktorů znalostní ekonomiky.

6. Závěr

Příspěvek předkládá výsledky mnohorozměrné klasifikace zemí Evropské unie na základě proměnných indexů znalostní ekonomiky. Inovace, znalosti a kreativita jako zdroj růstu konkurenceschopnosti národních ekonomik vedou k ekonomickému růstu ústícímu k zvyšování životní úrovně a bohatství obyvatel. Užitím shlukové a faktorové analýzy byly prokázány rozdíly mezi zeměmi Evropské unie zejména u proměnných inovačního charakteru, kdy nové členské země v těchto ukazatelích zaostávají oproti tradičním zemím Evropské unie. Řecko, Portugalsko, Itálie, Francie a Španělsko, tedy tradiční země EU, jsou však klasifikovány mimo skupinu tradičních zemí EU, což ukazuje rozdílný vývoj těchto „jižních“ zemí a řadí je tak v rámci sledovaných charakteristik spíše mezi nové členské země EU.

7. Literatura

[1] HEBÁK, P.: Vícerozměrné statistické metody. Vyd. 1. Praha: Informatorium, 2004.239 s.

[2] Kadeřábková, A. a kol.: Proces konvergence v nových členských zemích EU**.** Praha, CES VŠEM 2007, Working Paper No. 6.

[3] LUKASOVÁ, A., ŠARMANOVÁ, J. Metody shlukové analýzy. Vyd. 1. Praha: SNTL, 1985. 210 s.

[4] WORLD BANK: Knowledge Assessment Methodology.
http://info.worldbank.org/etools/kam2/KAM_page3.asp?default=1

Adresa autora (-ov):

Ing. Jakub Odehnal
Katedra ekonomie
Kounicova 65
Univerzita obrany
612 00 Brno
jakub.odehnal@unob.cz

Mgr. Marek Sedlačík, Ph.D.
Katedra ekonometrie
Kounicova 65
Univerzita obrany
612 00 Brno
marek.sedlacik@unob.cz

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

Příspěvek vznikl za podpory výzkumného záměru MSM0021622418 a VZ04-FEM-K101

Analýza predikovateľnosti akciových trhov počas finančnej krízy
The analysis of predictability of financial markets during a financial crisis

Tomáš Osička, Martin Boďa

Abstract: The aim of this paper is to determine changes in market efficiency due to the financial crisis. Several degrees of market efficiency are introduced and each degree is issued with test to judge whether historical returns follow the assumptions of a given degree of market efficiency. These tests connect their results to the ongoing financial crisis.

Key words: random walk 1, random walk 2, random walk 3, sequences and reversals, runs, filter rules, Dow Jones Industrial Average

Kľúčové slová: náhodná prechádzka 1, náhodná prechádzka 2, náhodná prechádzka 3, série a zvraty, trendy, filtračné pravidlá, Dow Jonesov priemyselný index

1. Úvod

Finančné trhy sú už desaťročia predmetom záujmu mnohých ekonómov, matematikov či štatistikov. Výsledkom ich skúmania je množstvo modelov. Niektoré dávajú návod na tvorbu portfólií, pomocou iných je možné ohraničiť stratu z uzavretej pozície. Spoločným menovateľom viacerých z nich je predpoklad o efektívnom trhu, na ktorom investori pôsobia. Efektívnym trhom sa pritom podľa vymedzenia Fischera Blacka rozumie trh, na ktorom investori bez špeciálnych informácií o spoločnosti nemôžu dosiahnuť výnosy, a aj pre investorov, ktorí takéto informácie majú, je dosahovanie výnosov náročné, pretože cena sa prispôsobuje príliš rýchlo po tom, čo sa informácia stane dostupnou. Preto sa dá očakávať, že sa v cenách akcií bude prejavovať skôr náhodnosť ako kontinuita ich vývoja. Náhodnosťou rozumieme, že série malých nárastov (alebo poklesov) ceny sú veľmi nepravdepodobné a v prípade nárastu cena akcie vzrastie na vyššiu cenovú úroveň v jednom kroku, a nie vo veľkom počte malých nárastov. V prípadoch vážneho narušenia efektívnosti je ohrozená správnosť výstupov z modelov založených na týchto predpokladoch.

Je nesporné, že finančné krízy ovplyvňujú správanie sa investorov a celého trhu (resp., že správanie sa investorov a celého trhu spôsobuje finančnú krízu). Zaujímavé je v tomto kontexte zaoberať sa, ako je prepojená finančná kríza a trhová efektívnosť. Je zaujímavé zistiť, či existuje kauzálny vplyv finančných kríz na trhovú efektívnosť. Jedným z možných prístupov je použitie časti časového radu cien akcií na zostrojenie kĺzavého indikátora efektívnosti a vyhodnotenie jeho vývoja.

V príspevku sa obmedzíme na aktuálne súvislosti prebiehajúcej svetovej finančnej krízy. Strata náhodnosti v cenových pohyboch na burzách podľa nášho názoru môže sťažiť pozíciu regulátorov, ktorí sa pri posudzovaní rizikovej expozície dohliadaných subjektov často spoliehajú práve na výstupy modelov merajúcich riziko.

2. Predpoklady o charaktere finančných trhov

Náhodnosť vo vývoji cien aktív na finančných trhoch býva popisovaná rôznymi prístupmi, ktoré sa líšia v prísnosti podmienok, ktoré na proces kladú. Procesy s veľmi prísnymi podmienkami majú užitočné matematické vlastnosti, umožňujúce jednoduchšie modelovanie, naproti tomu sú však vzdialenejšie trhovej realite. Základnými formami procesov pre finančné trhy sú martingál a rôzne verzie náhodnej prechádzky (*random walk*).

2.1 Martingál

Ide o historicky najstarší prístup, popisujúci vývoj cien ťažko splniteľnými podmienkami. Ide však o základný koncept, preto ho podrobne popíšeme.

Martingál vzhľadom na filtráciu $(\mathcal{F}_t)_{t\geq 0}$ je stochastický proces $P = (P_t)_{t\geq 0}$ na $(\Omega, \mathcal{F}, \mathbf{P})$ taký, že (1.) P je adaptovaný na filtráciu $(\mathcal{F}_t)_t$, (2.) $\mathrm{E}[|P_t|] < \infty$ pre $\forall\ t \geq 0$, (3.) $\mathrm{E}[P_t \mid \mathcal{F}_s] = \mathrm{P}_s$ pre $\forall\ 0 \leq s \leq t$.

Voľnejšie môžeme povedať, že martingál je stochastický proces (čiže funkciu času t a náhodného javu ω označujúceho nárast či pokles ceny), ktorého funkčná hodnota v čase t závisí výlučne na vývoji do času t (adaptovanosť). Ako proces má konečnú strednú hodnotu v ľubovoľnom čase a očakávaná hodnota procesu v budúcnosti vzhľadom na množinu pohybov do času s, teda s použitím informácií dostupných v čase s, bude P_s. Inak povedané, minulý vývoj nedáva žiadny základ predpokladom rastu ani poklesu hodnoty procesu v budúcnosti.

Vyššie popísaný proces výborne spĺňa požiadavky kladené na efektívny trh, nakoľko nie je možné dosahovať výnosy na základe informácií o minulom vývoji. Avšak viacero súčasných prístupov nachádza vzťah medzi rizikom a zodpovedajúcim výnosom, čo martingál nerešpektuje.

2.2 Náhodná prechádza 1

Náhodná prechádzka v podobe najsilnejších predpokladov je proces, ktorý je možné popísať nasledovným vzťahom:

$$P_t = \mu + P_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{IID}(0,\sigma^2),$$

kde μ je očakávaný cenový nárast, označovaný aj ako *drift* a stochastické zložky ε_t sú nezávislé a rovnako rozdelené (IID: *independent and identically distributed*) so strednou hodnotou 0 a disperziou σ^2. Na základe tejto definície vieme určiť podmienenú strednú hodnotu a disperziu procesu v čase t.

$$\mathrm{E}[P_t \mid P_0] = P_0 + \mu_t,$$
$$\mathrm{Var}[P_t \mid P_0] = \sigma^2 t.$$

2.3 Náhodná prechádzka 2

Rovnako rozdelené stochastické zložky sú veľmi silným predpokladom, ktorý je nesplniteľný najmä ak sledujeme vývoj za dlhšie časové obdobie. Finančné trhy sa v čase vyvíjajú a zmeny v ich štruktúre, alebo v reálnej ekonomike sa sotva neprejavia na charaktere vývoja cien finančných titulov. Z tohto dôvodu sa dajú upraviť predpoklady predchádzajúceho modelu a pre proces náhodnej prechádzky 2 postačí aby boli prírastky nezávislé, ale už nemusia byť rovnako rozdelené (INID: *independent but not identically distributed*):

$$P_t = \mu + P_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{INID}(0,\sigma^2).$$

2.4 Náhodná prechádzka 3

Posledná verzia modelu náhodnej prechádzky upúšťa oproti modelu náhodnej prechádzky 2 od nezávislosti náhodných prírastkov. Pre model postačuje ak sú prírastky nekorelované v lineárnom zmysle slova. Preto aby časový rad cien finančného titulu spĺňal predpoklady modelu náhodnej prechádzky 3, nemusia byť prírastky rovnako rozdelené, ale musia byť nekorelované (lineárne nezávislé) (UCNID: *uncorrelated and not identically distributed*):

$$P_t = \mu + P_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{UCNID}(0,\sigma^2).$$

Je zrejmé, že náhodná prechádzka 3 je špeciálny prípad modelu náhodnej prechádzky 2, za predpokladu, že prírastky okrem nekorelovanosti (lineárnej nezávislosti) budú aj nezávislé. A v prípade, že by boli navyše aj rovnako rozdelené, pôjde o náhodnú prechádzku 1.

3. Testy na meranie trhovej efektívnosti

V článku budeme overovať, ako dáta rešpektujú predpoklady jednotlivých verzií náhodnej prechádzky. Pre lepšiu názornosť výstupov stručne popíšeme použité metódy.

3.1 Série a zvraty

Test skúmajúci početnosti sérií a zvratov vo vývoji cien časového radu je jedným z prvých testov, skúmajúcich, či časový rad spĺňa predpoklady modelu náhodnej prechádzky 1. Zostrojili ho Cowles a Jones v roku 1937, ktorí vychádzali z logaritmickej verzie náhodnej prechádzky 1 ktorú možno zapísať nasledovne:

$$p_t = p_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{IID}(0,\sigma^2),$$

kde $p_t = ln(P_t)$. Vytvorí sa premenná I_t, ktorá nadobúda hodnotu 1 v prípade, že cena v čase t vzrástla a hodnotu 0 pri poklese, alebo nezmenenej cene. Na základe I_t definujeme N_s ako počet sérií (*sequences*) dvoch po sebe nasledujúcich rovnakých pohybov v cene (2-krát nárast, resp. 2-krát pokles) a N_r ako počet zvratov (*reversals*).

$$N_s = \sum_{t=1}^{n} Y_t, Y_t = I_t I_{t+1} + (1-I_t)(1-I_{t+1}),\ N_r = n - N_s .$$

Testovacou štatistikou je Cowlesov-Jonesov pomer, definovaný ako $CJ = N_s/N_r$. Dá sa ukázať, že ak $\varepsilon_t \sim \mathrm{N}(0,\sigma^2)$ a pravdepodobnosť nárastu ceny akcie označíme π a pravdepodobnosť série označíme π_s, Cowlesov-Jonesov pomer bude mať asymptoticky normálne rozdelenie s nasledovnými parametrami:

$$CJ \sim_{as.} N\left(\frac{\pi_s}{1-\pi_s}, \frac{\pi_s(1-\pi_s)+2(\pi^3+(1-\pi)^3-\pi_s^2}{n(1-\pi_s)^4}\right).$$

Tento vzťah je pre nás zároveň základom na výpočet p-hodnoty v testoch, ktorým sme dáta podrobili.

3.2 Trendy

Ďalší test predpokladov náhodnej prechádzky počíta rastové alebo klesajúce trendy. Použili sme verziu modelu počítajúcu s dvoma stavmi sledovanej premennej, 1 v prípade rastu a 0 v prípade poklesu či nezmenenej ceny. Možno použiť premennú I_t z Cowlesovho-Jonesovho testu, ktorá nadobúda práve také hodnoty. Oproti modelu sledujúcemu série a zvraty budeme počítať, koľko trendov sa v dátach nachádza, keď I_t opakovane dosahovala tie isté hodnoty. Ak by napríklad počas piatich období I_t dosahoval hodnoty 00101, napočítali by sme 1 sériu (00) a tri zvraty (1, 0, 1). Ale trendy by v postupnosti boli štyri, najprv jeden dvojitý (00) a potom tri trendy trvajúce jedno obdobie (1, 0, 1).

Označme počet trendov v dátach N_{runs} a pravdepodobnosť, že I_t nadobudne hodnotu 1 ako π (túto pravdepodobnosť určíme empiricky ako podiel nárastov na celkovom sledovanom období). Je možné ukázať, že náhodná premenná N_{runs} bude mať za platnosti H_0 o tom, že prírastky hodnôt v časovom rade sa riadia predpokladmi náhodnej prechádzky 1, normálne rozdelenie s parametrami

$$N_{runs} \sim N(2n\pi(1-\pi) - \frac{1}{2}, 4n\pi(1-\pi)\left[1-3\pi(1-\pi)\right]).$$

3.3 Filtračné pravidlá

Odporúča sa testovať model náhodnej prechádzky 2 pomocou fiktívnych obchodných operácií s aktívom, riadiacich sa jednoduchým súborom pravidiel. V prípade *x*- percentného poklesu ceny treba aktíva predať a pri *x*- percentnom náraste, naopak, kúpiť. Na jednej strane umožní tento postup odhaliť trendy skryté za malými výkyvmi trhovej ceny, pretože si budeme všímať iba výrazné výkyvy. Na druhej strane, ak by bol vývoj cien finančného aktíva nepredikovateľný, rozdiel medzi výnosom, dosiahnutým použitím ľubovoľného pravidla, by dlhodobo nemal byť odlišný oproti výnosu z kúpy aktíva a jeho držby počas celého sledovaného obdobia. My sme v príspevku sledovali práve bonusový výnos dosiahnutý aplikáciou pravidla oproti výnosu z držby aktív. Podstatným detailom testu je voľba hladiny, pri ktorej aktíva predávame, resp. kupujeme. Alexander odporúča voliť hladinu v rozmedzí 0,5% až 1,5%.

3.4 Autokorelácia časového radu

Nekorelovanosť náhodných prírastkov sme testovali pre model náhodnej prechádzky 3 na základe známej Ljungovej-Boxovej štatistiky.

3.5 Normalita výnosov

Normalita výnosov finančných titulov nebola posudzovaná formou štatistického testu. Dôraz sme kládli na identifikáciu zdrojov odchýlok od normálneho rozdelenia, aby sme mohli lepšie charakterizovať vývoj dát v čase. Posudzovanie bolo založené na kĺzavých hodnotách koeficientu šikmosti, koeficientu špicatosti a Jarqueovej-Berovej štatistiky.

4. Výsledky testov

Efektívnosť trhov sme testovali na dátach Dow Jonesovho priemyselného indexu (*Dow Jones Industrial Average*) za časové obdobie od 24. 11. 2004 do 07. 11. 2008. Zvolili sme štvorhodinové dáta, aby sme mali dostatok pozorovaní pre testy a mohli sme počítať hodnoty pre kratšie časové obdobia. Dôsledkom tejto voľby je jednak vyššia sila testov, jednak aj väčšia dynamiku výstupov. Dáta sme stiahli cez demo účet maklérskej spoločnosti *X-Trade Brokers*. Všetky testy sme konštruovali na podobnom princípe ako pohyblivé priemery. Zvolili sme si 101 pozorovaní, čo vzhľadom na to, že za deň máme dve štvorhodinové hodnoty, zodpovedá obdobiu zhruba dvoch obchodných mesiacov. Schéma na ďalšej strane graficky ilustruje niektoré výstupy nášho testovania.

5. Interpretácia a závery

Zo schémy vidíme, že hodnota Cowlesovho-Jonesovho pomeru za posledné roky kolísala okolo 1. Podľa p-hodnoty boli odchýlky viackrát príliš výrazné, aby mohol platiť predpoklad náhodnej prechádzky 1, avšak jeho porušenie bolo len krátkodobé. Čo bolo pre nás pomerne prekvapivé, je blízkosť Cowlesovho-Jonesovho pomeru jednej práve v období najvýraznejších prepadov na akciových trhoch reflektovaných prepadmi Dow Jonesovho priemyselného indexu. Cowlesov-Jonesov pomer sa práve v období najvýraznejších prejavov krízy ustálil na jednej. Podobne aj p-hodnota odvodená na základe trendového modelu koncom leta 2008 prudko narástla na úroveň blízkej jednej. Ani jeden z modelov nevedie k záveru, že finančné trhy sa počas krízy nesprávajú podľa modelu náhodnej prechádzky 1. Musíme však zohľadniť konštrukciu modelov. Oba modely nerešpektujú veľkosť prepadu či nárastu. Sledujú iba smer, ktorým sa cena uberala. Priemerné kĺzavé poklesy a nárasty potvrdili, že prepad nebol spôsobený frekventovanejšími poklesmi oproti nárastom alebo prítomnosťou klesajúcich trendov. Priemerná výška poklesov bola vyššia oproti priemernej výške nárastu. Z toho vyplýva, že počas prepadov v súčasnej finančnej kríze dochádzalo k prudkým poklesom, po ktorých nasledovali len mierne korekcie.

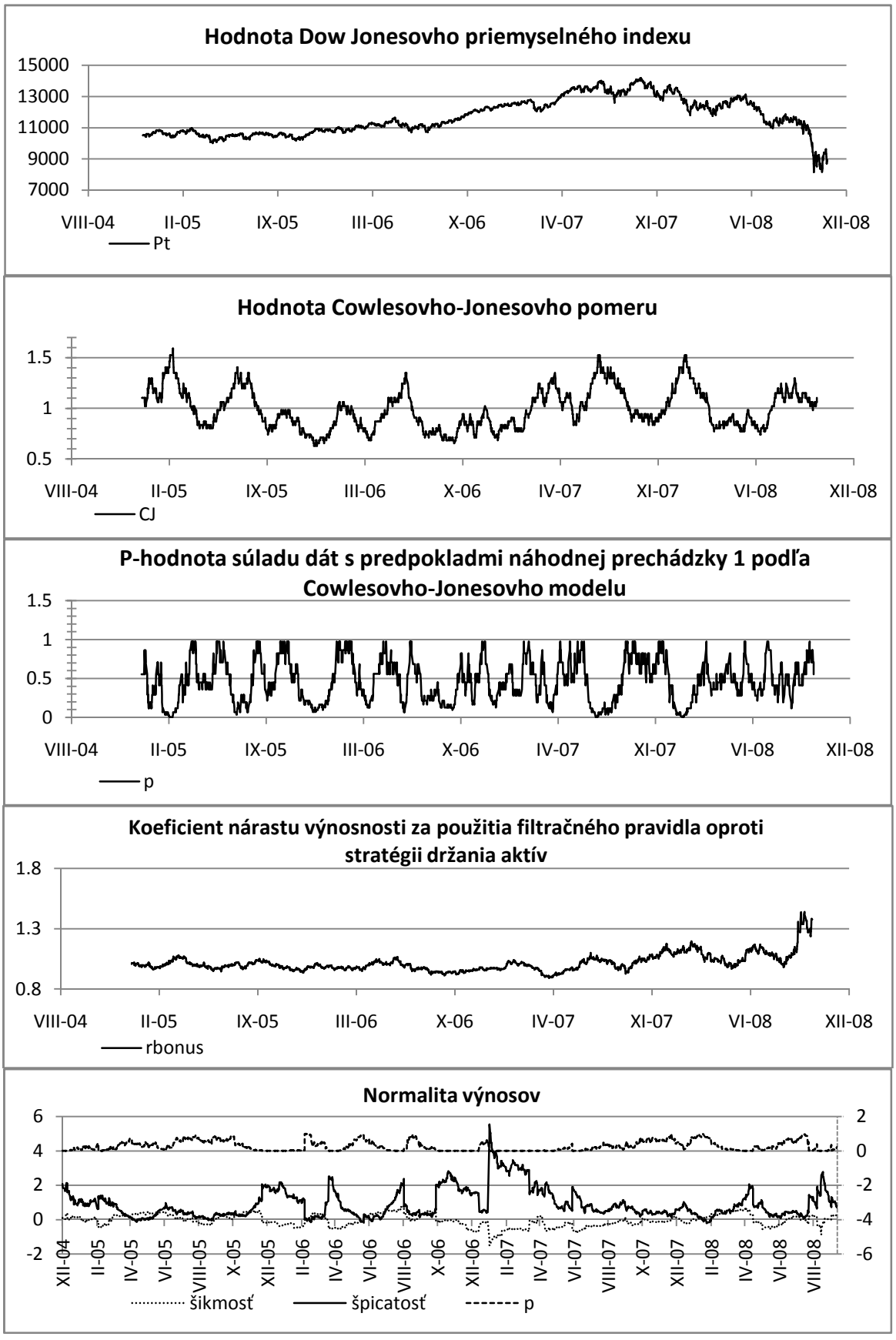


Schéma (začiatok) *Výstupy z overovania trhovej efektívnosti*

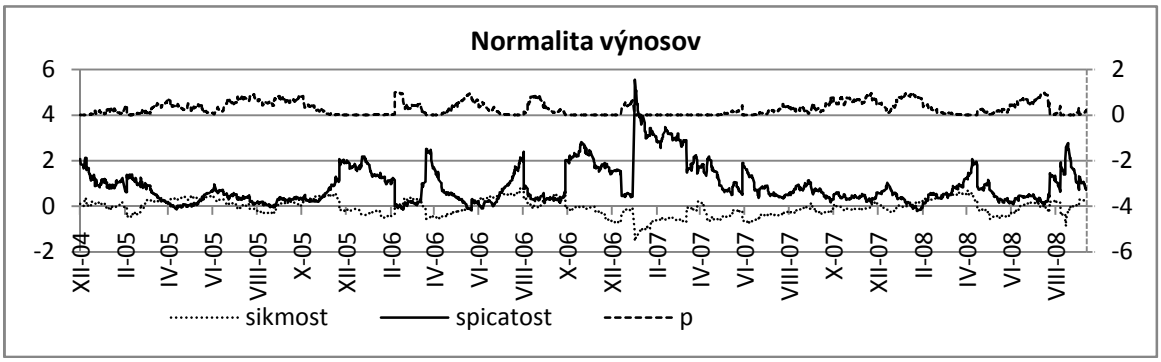


Schéma (koniec) *Výstupy z overovania trhovej efektívnosti*

Filtračné pravidlo veľmi jasne odhaľuje trend vznikajúci počas prepadu cien na finančných trhoch. Od roku 2005 až do konca roku 2007 je bonusový výnos z aplikácie filtračnej obchodnej stratégie prakticky nulový, koeficient striedavo kolíše okolo jednej. Začiatkom roka 2008 sa už mierne prejavuje nárast výnosovej prémie oproti trhu, dosiahnutý najmä zhoršujúcimi sa výsledkami na trhu. V septembri rozdiel prekročil 20% a zamieril až na vyše 40%. Výnosnosť založená na filtračnej obchodnej stratégii bola nízka, avšak prepady vyslali predajný signál, čo zabránilo ďalším stratám a zvýraznilo relatívny rozdiel oproti vývoju indexu. Realizácia výrazne vyššieho zisku (resp. nižšej straty v našom prípade) potvrdzuje prítomnosť trendu počas finančnej krízy a istú mieru predikovateľnosti.

Ljungove-Boxove testovacie štatistiky sme zostrojili do ôsmeho rádu autokorelácie, z rozsahových dôvodov však nie sú prezentované. Realizácie Ljungovej-Boxovej štatistiky poukázali na štatisticky významnú (aj na hladine významnosti 0.01) autokoreláciu okolo štvrtého až siedmeho oneskorenia. Ostatné štatistiky boli štatisticky významné aspoň na hladine 0.10. Ak teda zatiaľ opomenieme, či cena klesá alebo stúpa, a zvážime iba mieru poklesu, zamietame predpoklad o správaní sa Dow Jonesovho priemyselného indexu podľa modelu náhodnej prechádzky 3 (nota bene, čo implikuje aj zamietnutie predpokladu o modeli náhodnej prechádzky 1). Počas rokov 2005 – 2007 s výnimkou krátkodobých, skôr výnimočných porušení, hypotézu o platnosti náhodnej prechádzky nezamietame.

Normalita výnosov sa ukázala byť veľmi naivnou predstavou, keďže ani počas relatívne stabilného vývoja výnosy Dow Jonesovho priemyselného indexu neboli normálne rozdelené.

Zhrnutím je zistenie, že Dow Jonesov priemyselný index počas prepadu, ktorý bol spôsobený finančnou krízou, skutočne stratil časť svojej efektivity, čo dáva predpoklady pre úspešné modelovanie. Zistili sme však, že trendy sa neprejavovali nezvyčajne častými poklesmi (poklesy a nárasty boli počas krízy náhodné), ale nezvyčajnou silou týchto prepadov, ktoré neboli kompenzované dostatočne výraznými nárastmi.

6. Literatúra

[1.] ARLT, J.; ARLTOVÁ, M. 2003. *Finanční časové řady.* Praha : Grada, 2003. 220 s. ISBN 80-247-0330-0.

[2.] CAMPBELL, J. Y.; LO, A. W.; MACKINLAY, A. C. 1997. *The econometrics of financial markets*. Princeton : Princeton Univerity Press, 1997. 611 s. ISBN 0-691-04301-9.

Adresa autorov

Tomáš Osička
Univerzita Mateja Bela v Banskej Bystrici
Ekonomická fakulta
Tajovského 10, 975 90 Banská Bystrica
tomas.osicka@gmail.com

Martin Boďa, Ing. et Bc.
Univerzita Mateja Bela v Banskej Bystrici
Ekonomická fakulta, KKMaIS
Tajovského 10, 975 90 Banská Bystrica
martin.boda@umb.sk

Srovnání modelů s alternativní vysvětlovanou proměnnou
Comparison of models for binary response data

Iva Pecáková

Abstract: Generalized linear models for binary response data are recently used increasingly in a wide variety of applications. Although the logit is the most popular link function for probabilities, statistical systems offer models with other links too. The paper compares logit, probit, log-log and clog-log models.

Key words: logit, probit, log-log, clog-log models for binary response data

Klíčová slova: logit, probit, log-log, clog-log modely pro binární vysvětlovanou proměnnou

1. Metodologie

Uvažujme alternativní proměnnou Y, jež nabývá s pravděpodobností π hodnoty 1 (nastoupení určitého jevu) a s pravděpodobností $1 - \pi$ hodnoty 0 (nenastoupení tohoto jevu). Představuje-li vektor

$$\mathbf{x}_i^{/} = \left[x_{i1}, x_{i2}, ..., x_{ik} \right]$$

i-tou jedinečnou kombinaci hodnot k nenáhodných vysvětlujících proměnných, i = 1, 2, …, n, i-té podmíněné rozdělení veličiny Y je alternativní s parametrem (a střední hodnotou veličiny Y) πi a pravděpodobnostní funkcí

$$P(y_i / \pi_i) = \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}, \; i = 1, 2, \ldots, n. \tag{1}$$

Vyskytují-li se kombinace hodnot (kategoriálních) vysvětlujících proměnných opakovaně, jsou data obvykle nejprve roztříděna do kontingenční tabulky. Představuje-li nyní y_i počet případů, kdy pro i-tou kombinaci hodnot vysvětlujících proměnných (z C možných) veličina Y nabývá hodnoty 1, a n_i celkový počet případů pro jednotlivé kombinace hodnot vysvětlujících proměnných, $\sum_C n_i = n$, pak tyto četnosti mají binomické rozdělení s pravděpodobnostní funkcí

$$P(y_i / n_i, \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}, \; i = 1, 2, \ldots, C. \tag{2}$$

Zůstává-li pro různé vektory **x** podmíněné rozdělení pravděpodobnosti veličiny Y stejné, pak vysvětlující proměnné X_1, X_2, …, X_k zřejmě nastoupení sledovaného jevu neovlivňují, Y na nich nezávisí. Pokud však různé kombinace hodnot vysvětlujících proměnných znamenají různé pravděpodobnosti π, lze zřejmě uvažovat o nějakém typu závislosti Y na **x** a pokusit se o jeho modelování.

V případě opakovaného výskytu kombinací hodnot vysvětlujících proměnných je relativní četnost nastoupení sledovaného jevu p ovlivněna jednak pozorovanými nenáhodnými veličinami **x,** jednak nepozorovanými veličinami, jež považujeme v souhrnu za náhodnou složku modelu ε;

$$p = \mathbf{x}'\boldsymbol{\beta} + \varepsilon . \qquad (3)$$

Na pravděpodobnostní rozdělení náhodné složky jsou kladeny požadavky, jejichž splněním je podmíněna kvalita odhadů parametrů, a tedy užitečnost odhadnutého modelu.

Pro jedinečné kombinace vysvětlujících proměnných je situace podobná. Součet (3) je však v takovém případě výhodné považovat za veličinu, představující „sklon“ či „ochotu“ jednotky k realizaci sledovaného jevu (užitečnost výrobku či služby v ekonometrii, tolerance k insekticidu v biologii či k léku v medicíně, preference ve společenských vědách apod.).

V obecném lineárním modelu je podmíněná střední hodnota vysvětlované proměnné determinována lineární kombinací vysvětlujících proměnných,

$$\eta = \pi = \beta_0 + \sum_j^k \beta_j x_j = \mathbf{x}'\boldsymbol{\beta} , \qquad (4)$$

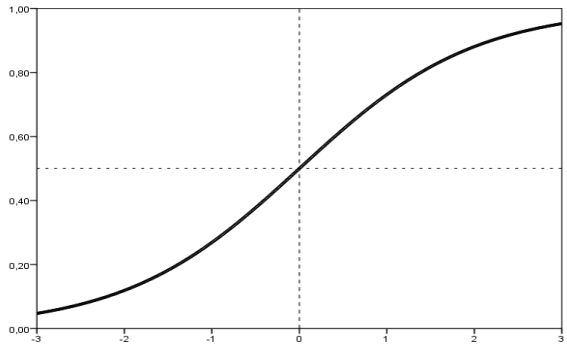
a rozdělení náhodné složky je považováno za normální s nulovou střední hodnotou. Charakter vztahu mezi pravděpodobností π a vysvětlujícími proměnnými je však významně ovlivněn omezením jejích hodnot pouze na interval od 0 do 1. To použití takového modelu obvykle vylučuje.

Deterministická složka *zobecněného* lineárního modelu je nelineární funkcí této podmíněné střední hodnoty,

$$\eta = g(\pi) = \mathbf{x}'\boldsymbol{\beta} . \qquad (5)$$

Použití nelineární funkce $g(\pi)$ má zde zajistit transformaci intervalu možných hodnot pravděpodobnosti $\langle 0,1 \rangle$ na interval reálných čísel. Budeme-li pak předpokládat, že sledovaný jev nastane (Y = 1), pokud $\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0$, potom pravděpodobnost nastoupení tohoto jevu lze považovat za distribuční funkci rozdělení náhodné složky v bodě $\mathbf{x}'\boldsymbol{\beta}$.

I z věcného hlediska lze pro modelování pravděpodobnosti považovat za vhodnější použití monotónní funkce, jejíž nárůst (či pokles) se v blízkosti 0 a 1 zpomaluje a jež tak nabývá v grafu tvaru s-křivky s asymptotami v bodech 0 a 1 rovnoběžnými s osou x (obrázek 1).



***Obrázek* 1**. *Příklad symetrické s-křivky (logit)*

Typ této nelineární funkce tedy ovlivňuje uvažované rozdělení náhodné složky.

1)

Je-li považováno rozdělení náhodné složky za normované logistické s nulovou střední hodnotou a rozptylem $\pi^2/3 = 3{,}287$ (π je zde Ludolfovo číslo), jehož distribuční funkci lze zapsat jako

$$F(\varepsilon) = 1/(1 + e^{-\varepsilon}), \tag{6}$$

pak lze pravděpodobnost π vyjádřit jako

$$\pi = \left[1+\exp(-\mathbf{x}'\boldsymbol{\beta})\right]^{-1} = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}. \tag{7}$$

Z (7) plyne, že

$$\frac{\pi}{1-\pi} = \exp(\mathbf{x}'\boldsymbol{\beta}), \text{ a tedy } \ln\frac{\pi}{1-\pi} = g(\pi) = \mathbf{x}'\boldsymbol{\beta}. \tag{8}$$

Pro uvedený typ transformace se vžilo označení <u>logit</u> a pro model s logitovou transformací logistický regresní model. Funkce (7) má v grafu tvar symetrické s-křivky s jedním inflexním bodem (viz obrázek 1; jako model se v logistické regresi používá i analogická křivka klesající). Protože (7) je distribuční funkce logistického rozdělení, logit je jeho 100π-procentním kvantilem.

Obliba logitové transformace pramení z jednoduché interpretace parametrů lineární kombinace vysvětlujících proměnných, které po odlogaritmování představují změnu šance $\pi(1-\pi)$ nastoupení sledovaného jevu (tedy $Y = 1$) vyvolanou (obecně řečeno) jednotkovou změnou odpovídající vysvětlující proměnné, pokud se ostatní nezmění.

2)

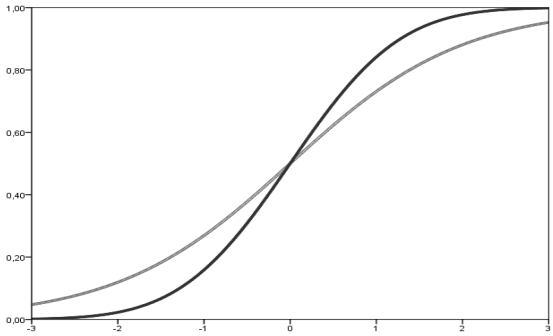
Pokud je rozdělení náhodné složky považováno za normované normální s nulovou střední hodnotou a jednotkovým rozptylem, pak lze π vyjádřit s užitím jeho distribuční funkce jako

$$\pi = \Phi(\mathbf{x}'\boldsymbol{\beta}), \tag{9}$$

a tedy

$$\mathbf{x}'\boldsymbol{\beta} = \Phi^{-1}(\pi). \tag{10}$$

Lineární kombinace vysvětlujících proměnných $\mathbf{x}'\boldsymbol{\beta}$ je v tomto případě 100π-procentním kvantilem normovaného normálního rozdělení. Pro uvedený typ transformace se vžil termín <u>probit</u>. Křivka grafu distribuční funkce normovaného normálního rozdělení má stejný inflexní bod jako křivka grafu distribuční funkce normovaného logistického rozdělení, má však v tomto bodě větší směrnici (cca 1,6krát), tj. je strmější. Vzhledem k rozptylu $\pi^2/3$ (π je zde Ludolfovo číslo) má logistické rozdělení také cca 1,8krát větší směrodatnou odchylku. Důsledkem těchto skutečností je, že odhady parametrů logitového modelu jsou u týchž dat cca 1,6 −1,8 násobkem parametrů modelu probitového.



***Obrázek 2**. Srovnání logitové (světlejší) a probitové křivky*

3)

Symetrie obou předchozích transformací, kdy

$$\text{logit}(\pi) = \ln\frac{\pi}{1-\pi} = -\ln\frac{1-\pi}{\pi} = -\,\text{logit}(1-\pi),$$

a také

$$\Phi^{-1}(\pi) = -\,\Phi^{-1}(1-\pi),$$

může být na závadu, pokud pravděpodobnost π klesá k nule či stoupá k jedné odlišným způsobem. Z předpokladu normovaného Gumbelova rozdělení náhodné složky plyne

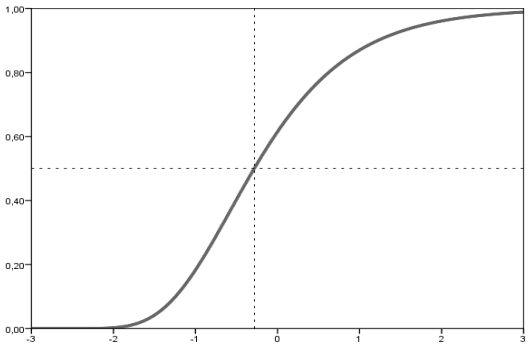
$$\pi = 1-\exp\left[-\exp(\mathbf{x}'\boldsymbol{\beta})\right], \tag{9}$$

$$\mathbf{x}'\boldsymbol{\beta} = \ln\left[-\ln(1-\pi)\right] \tag{10}$$

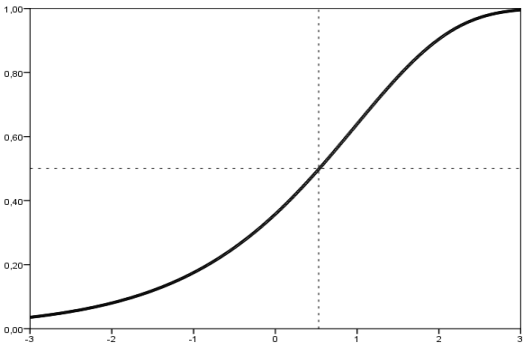
(tzv. clog-log, *complementary log-log*, transformace). V tomto případě je s-křivka nesymetrická. Pro malé pravděpodobnosti se clog-log model podobá logitu, k jedné se však blíží pomaleji. Jeho název je důsledkem toho, že z (9) dostáváme

$$\pi' = 1-\pi = \exp\left[-\exp(\mathbf{x}'\boldsymbol{\beta})\right], \text{ a tedy } \mathbf{x}'\boldsymbol{\beta} = \ln\left[-\ln(\pi')\right]$$

(tzv. log-log transformace). Střední hodnota normovaného Gumbelova rozdělení je dána Eulerovou-Mascheroniho konstantou (0.577), rozptyl $\pi^2/6$ = 1,644 (π je zde Ludolfovo číslo) je tedy oproti normovanému logistickému rozdělení poloviční.



***Obrázek 3**. Model clog-log (příklad)*



***Obrázek 4**. Model log-log (příklad)*

2. Analýza

Součástí omnibusového výzkumu agentury[4] bylo kupní chování respondentů. Zjišťována byla jejich motivace při rozhodování o koupi českého či zahraničního výrobku, výrobku značkového či neznačkového, co je případně vede ke změně značky. Rozsah výběrového souboru je 1006 osob. Datová matice obsahuje dále některé identifikační údaje o vybraných respondentech, jako je jejich pohlaví (muž = 1, žena = 2), věk (v dokončených letech), vzdělání (základní = 1, středoškolské = 2, vysokoškolské = 3), příjem domácnosti na hlavu (v tisících Kč), atd. Analyzujeme zjištění, zda respondent dává při koupi oblečení přednost českým či zahraničním výrobkům, v závislosti na pohlaví, věku, vzdělání a příjmu domácnosti na hlavu. Výpočty byly provedeny v systému SPSS 16.0.

Podobnost modelů je zřejmá z následujících tabulek. Parametry logitového modelu jsou cca 1,6 násobkem parametrů modelu probitového, parametry clog-log modelu jsou cca jejich 1,2 násobkem. Zaznamenejme ještě prakticky stejné výsledky testů u všech modelů a také skutečnost, že ve všech případech došlo ke zhruba stejnému významnému snížení deviance (na hladině 0,001) a hodnoty Nagelkerkeovy statistiky se pohybují na solidní úrovni mezi 0,26 – 0,28.

Parameter Estimates

		Estimate	S. E.	Wald	df	Sig.
Threshold	[zn1e = 1]	-2,274	,636	12,802	1	,000
Location	vek	-,059	,007	68,477	1	,000
	prnahlav	6,290E-5	,000	4,567	1	,033
	[sex=1]	,027	,229	,013	1	,908
	[sex=2]	0[a]	.	.	0	.
	[vzd=1]	-,870	,472	3,405	1	,065
	[vzd=2]	-,983	,486	4,095	1	,043
	[vzd=3]	0[a]	.	.	0	.

Link function: Logit.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.
Threshold	[zn1e = 1]	-1,355	,376	12,978	1	,000
Location	vek	-,035	,004	75,142	1	,000
	prnahlav	3,623E-5	,000	4,283	1	,038
	[sex=1]	,029	,135	,045	1	,832
	[sex=2]	0[a]	.	.	0	.
	[vzd=1]	-,511	,278	3,389	1	,066
	[vzd=2]	-,595	,287	4,292	1	,038
	[vzd=3]	0[a]	.	.	0	.

Link function: Probit.

[4] Data poskytla výzkumná agentura Factum Invenio (2004)

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.
Threshold	[zn1e = 1]	-1,696	,408	17,303	1	,000
Location	vek	-,033	,004	68,378	1	,000
	prnahlav	3,233E-5	,000	3,005	1	,083
	[sex=1]	,049	,134	,135	1	,713
	[sex=2]	0[a]	.	.	0	.
	[vzd=1]	-,497	,301	2,732	1	,098
	[vzd=2]	-,603	,311	3,758	1	,053
	[vzd=3]	0[a]	.	.	0	.

Link function: Complementary Log-log.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.
Threshold	[zn1e = 1]	-1,363	,439	9,640	1	,002
Location	vek	-,047	,005	76,989	1	,000
	prnahlav	4,910E-5	,000	5,362	1	,021
	[sex=1]	,020	,171	,013	1	,908
	[sex=2]	0[a]	.	.	0	.
	[vzd=1]	-,627	,318	3,893	1	,048
	[vzd=2]	-,709	,330	4,630	1	,031
	[vzd=3]	0[a]	.	.	0	.

Link function: Negative Log-log.

3. Závěr

V praxi si získaly oblibu modely s logitovou transformací (tzv. logistická regrese) na úkor ostatních. Důvod je zřejmý – přirozená interpretace parametrů použitého modelu a fakt, že jiná transformace *obvykle* nepředstavuje pro zkoumanou souvislost významnější přínos.

4. Literatura

[1] Agresti, A.: Categorical Data Analysis, John Wiley & Sons, 2002
[2] Pecáková, I.: Statistika v terénních průzkumech, Praha, Professional Publishing 2008
[3] Powers, D.A. – Xie, Yu: Statistical Methods for Categorical Data Analysis, Academic Press 2000
[4] Train, K.: Discrete Choice Methods with Simulation, Cambridge University Press, 2003
[5] http://en.wikipedia.org/wiki/Gumbel_distribution
[6] http://data.princeton.edu/wws509/notes/c3s7.html

Adresa autorky:

Doc. Ing. Iva Pecáková, CSc.
VŠE Praha
nám. W. Churchilla 4, 130 67 Praha 3
e-mail: pecakova@vse.cz

Ověření platnosti keynesiánské spotřební funkce pro Českou republiku
Validation of Keynesians short-run consumption function in the atmosphere of Czech Republic

Jitka Poměnková, Zuzana Toufarová

Abstract: The paper is validation of Keynesians short-run consumption function in the Czech Republic environment for the period 2001/Q1–2008/Q2. This paper also converses factor of inflation into the consumption function. On the basis of the statistical data it also analyses justification this factor into the consumption function. The authors want, in the context of the results of the empiric analyses and in the context of the new suggested factor of the inflation, to create modified consumption function for the atmosphere of the Czech republic. New founded results will be interpreted in the conclusion of this paper.

Key words: Consumer, Consumption expenditures, Revenue of Consumer, Savings, Inflation, Consuption funkcion, Correlation, Partial Correlation

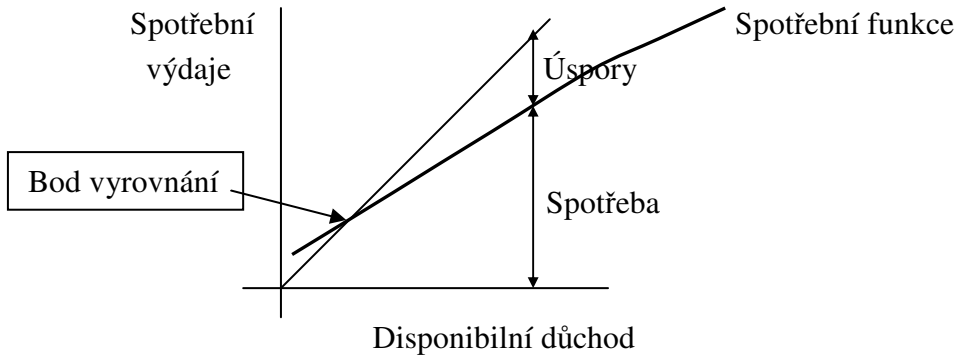
Kľúčové slová: spotřebitel, spotřební výdaje, příjmy spotřebitele, úspory, inflace, spotřební funkce, korelace, parciální korelace

1. Úvod

V současném silném konkurenčním boji mezi firmami je až nezbytné soustředit pozornost na zkoumání chování spotřebitele. Firmy bez odbytu svých produktů nebo služeb, tj. bez jejich prodeje konečným zákazníkům (spotřebitelům), nebudou moci „přežít“ na trhu. Proto je třeba nejen zákazníky bedlivě sledovat, přizpůsobovat nabídku jejich potřebám a přáním, zjišťovat jejich preference, ale také sledovat vývoj jejich příjmů. Obyvatelé dostávají důchod ze své práce a jako vlastníci kapitálu, platí daně a poté se rozhodují, kolik důdou o zdanění, tj. disponibilního důchodu, spotřebují a kolik ušetří.

Sledování vývoje spotřebních výdajů i příjmů obvyatel umožňuje nejen zjištění vývojových tendencí v jednotlivých skupinách spotřebních výdajů, ale například i možnost konstrukce odhadu budoucích hodnot, účinnou pomoc při strategickém plánování a rozhodování firem nebo plánování výdajů na reklamu a propagaci výrobků. Díky těmto krokům mohou firmy lépe přizpůsobit nabídku svých produktů potřebám zákazníků a zajistit tím přínos vlastní firmě. Pokud sledujeme chování spotřebitel je důležité sledovat nejen jejich preference, přání, výši příjmů, výdajů či úspor, ale také změny v cenách produktů, legislativu, hygienické omezení, dovozní či vývozní kvóty, konkurenci a v neposlední řadě také výši a vývoj inflace.

Vztah mezi úrovní spotřebních výdajů a úrovní disponibilního důchodu domácností zobrazuje spotřební funkce, viz obrazek 1. Bod vyrování je bod, kdy se přesně vyrovnávají spotřební výdaje a důchody domácností. Domácnosti v tomto bodě ani nespoří, ani nemá záporné úspory. Jestliže spotřební funkce leží nad osou kvadrantu, domácnosti vykazují záporné úspory, a jestliže leží pod osou kvadrantu, mají domácnosti kladné úspory. Výše kladných nebo záporných úspor se vždy měří jako svislá vzdálenost mezi spotřební funkcí a osou kvadrantu.



***Obrázek 3**: Spotřební funkce*

Další skupinou, která má na spotřebitele vliv, je velmi často reklama, doporučení, životní úroveň a vzdělanost národa apod. Některé z těchto faktorů jsou však obtížně sledovatelné a měřitelné. Nelze proto obsáhnout a měřit podrobně všechny faktory, které mohou spotřebitelské chování ovlivnit.

2. Cíl příspěvku

Cílem příspěvku je ověření platnosti keynesiánské spotřební funkce pro Českou republiku v období 2001/Q1–2008/Q2. Dále provést rozšíření klasické spotřební funkce o faktor inflace a na základě statistických testů vyhodnotit oprávněnost zapojení tohoto faktoru do spotřební funkce. V souvislosti s výsledky empirické analýzy nově navrhovaného faktoru inflace následně vytvořit upravenou spotřební funci pro Českou republiku a zjištěné výsledky interpretovat.

3. Metodika

Pro dosažení daného cíle příspěvku a současně pro analýzu spotřební funkce autorky shromáždily databázi sekundárních dat. Ukazatelé jako jsou příjmy, úspory, harmonizovaný index spotřebitelských cen získaly z EUROSTATu. A ukazatele výdaje (spotřební výdaje) z Českého statistického úřadu, kde jsou sledovány v rubrice statistika rodinných účtů. Spotřební výdaje jsou v České republice tříděny podle mezinárodně srovnávané klasifikace CZ-COICOP (Toufarová, 2008).

Analýza agregátní spotřeby na základě spotřební funkce umožňuje zkoumat ty faktory, jež určují způsob, jakým spotřebitel dělí svůj celkový disponibilní důchod na spotřební výdaje a úspory. Na základě předpokladů krátkodobé keynesiánské spotřební funkce rozdělují spotřebitelé svůj důchod na část spotřeby a úspor (S)

$$C_t = \beta_0 + \beta_1 Y_t - \beta_2 S_t, \quad t = 1,...,n. \tag{1}$$

Vztah (9) ovšem neobsahuje plně nevystihuje veškeré faktory mající vliv na agregátní spotřebu v České republice. Pozorování pro empirickou analýzu vzhledem k posttransformačnímu charakteru české ekonomiky obsahuje několik strukturálních zlomů, např. měnová krize v roce 1997, vstup do EU v roce 2004 apod. (Širůček, 2007 s. 237 - 248). Tyto strukturální zlomy označované za šoky měly na vývoj ekonomiky nemalý vliv. Dalším významným faktorem je omezený rozsah pozorování. To spolu s vyskytujícími se strukturálními zlomy způsobuje například nedostatečnou statistickou významnost standardně zahrnovaných proměnných než je tomu při aplikaci na země s rozvinutou tržní ekonomikou. Navrhovanou úpravou klasického modelu dle Huška (2003) se pak jeví přidání proměnné inflace

$$C_t = \beta_0 + \beta_1 Y_t - \beta_2 S_t + \beta_3 P_t, \quad t = 1,...,n \tag{2}$$

u které se očekává statistická nevýznamnost odhadnutého parametru, resp. nulová hodnota parametru, neboť růst cen při konstantních reálných příjmech a úsporách vyvolává ekviproporcionální změnu příjmů i úspor v peněžním vyjádření, tedy úroveň reálné spotřeby se nemění v závislosti na změně cenové hladiny. Pokud je odhadnutá hodnota koeficientu β_3 statisticky významná, pak dochází k peněžní iluzi, kdy spotřebitelé vnímají růst svých reálných mezd pouze jako růst nominální. Tedy předpokládají, že v ekonomice dochází k poklesu tzv. reálných peněžních zůstatků (Piguův efekt). V případě růstu cenové hladiny tak spotřebitelé snižují reálné množství spotřebovaných statků a služeb.

Pro empirickou analýzu výše popsané problematiky bude využito vícenásobného lineárního regresního modelu, který bude testován z hlediska významnosti regresních parametrů, významnosti modelu jako celku. Pozornost bude rovněž věnována testování autokorelace reziduí (Wooldridge, 2003). Pro stanovení vhodných proměnných pro formulaci vícerozměrného regresního modelu byla využita korelační analýza, a to konkrétně výběrový korelační koeficient a výběrový parciální korelační koeficient (Hebák, Hustopecký, Malá, 2005).

4. Vlastní práce

Pro emipirickou analýzu spotřebních výdajů v České republice byly zvoleny čtvrtletní hodnoty spotřebních výdajů (C), příjmů obyvatel (měřeno formou HDP, Y), úspor (S) a inflace (HICP) v období 2001/Q1–2008/Q2. Uvedené hodnoty byly použity jako meziroční čtvrtletní změny.

V první fázi byla provedena korelační analýza s cílem zjištění délky zpoždění proměnných zahrnutých do regresního modelu (10). Uvažovaná délka zpoždění byla do čtyř čtvrtletí. Ani v jednom případě však nebyl zjištěn výrazný nárůst korelace ve vztahu k hodnotě bez zpoždění, proto byly do modelu zahrnuty všechny proměnné v čase *t*. Následně byl proveden odhad regresního modelu, „model 1“, spotřebitelské funkce

$est(C_t)$	=	$0{,}8081Y_t$	$- 0{,}0764S_t$	$- 0{,}3643P_t$
		$SE=(0{,}1274)$	$(0{,}0347)$	$(0{,}1697)$
		$p\text{-}value=0{,}0000$	$0{,}0364$	$0{,}0410$
$n=30$		$R^2_{adj}=0{,}6632$	$F=19{,}71$ $(p\text{-}value = 0{,}0000)$	$d=1{,}2407.$

Z výsledků odhadu modelu 1 vyplývá, že se neprokázala statistická nevýznamnost parametru inflace (P). Na základě negativního znaménka odhadnutého koeficientu se lze domnívat, že spotřebitelé v České republice podléhají peněžní iluzi. Spotřebitelé předpokládají, že roste-li inflace, klesá jejich reálný příjem i úspory, přestože tomu tak není, a proto mají tendenci ke snižování svých spotřebitelských výdajů. Svým způsobem tudíž zaznamenají zvýšení cen, avšak nikoliv ekviproporcionální růst důchodů i úspor v peněžním vyjádření. Z uvedeného odhadu modelu je dále patrné, že rezidua jsou autokorelovaná (d=1,2407). V tomto případě se jedná o autokorelaci, která je způsobená pravděpodobně chybějícími proměnnými nebo opomenutými vlivy. Problematika spotřebitelského chování ukazuje na existenci vlivů (a tedy potenciálních nezávislých proměnných), jejichž použitelnost je pro účely této práce problematická, neboť neexistuje odpovídající běžně dostupná metodika pro měření jejich vlivu. Charakter těchto faktorů je většinou subjektivní. Příkladem může být sezónní reklama na dovolenou, pod jejímž vlivem se spotřebitelé rozhodují o nákupu. Problém autokorelace reziduí budeme řešit prostřednictvím rozšíření modelu o prvek náhodnosti.

Jako doplňující informaci uveďme vypočtené hodnoty korelačního koeficientu proměnných zařazených do modelu 1 pro posouzení vzájemné závislosti jednotlivých proměnných a to jak z pohledu klasické, tak parciální korelace. Hodnoty uvedené v tabulce 1

v závorkách (p-value) označuje hladinu významnosti α. Pro statistickou významnost korelačního koeficientu požadujeme, aby p-value < α (α = 1, 5, 10 %).

Tabulka 4: *Výběrové korelační a parciální korelační koeficienty*

	I.: Výběrový korelační koeficient				*II.: Výběrový parciální korelační koeficient*			
	C	*Y*	*S*	*P*	*C*	*Y*	*S*	*P*
C	*1*	*(0,1296)*	*(0,8682)*	*(0,0385)*	*1*			
Y	*0,2831*	*1*	*(0,0003)*	*(0,6991)*	*0,4005*	*1*		
S	*-0,0316*	*0,6095*	*1*	*(0,6722)*	*-0,3083*	*0,6527*	*1*	
P	*-0,3797*	*-0,0736*	*-0,0806*	*1*	*-0,4025*	*0,1352*	*-0,1634*	*1*

Z vypočtených hodnot výběrových korelací je patrná statistická významnost, a tedy korelovanost mezi příjmy a úsporami (r = 0,6095) a mezi spotřebou a inflací (r = -0,3797). Korelace mezi spotřebou a úsporami (r = -0,0316) se jeví jako statisticky nevýznamná, avšak splňuje očekávání ve smyslu znaménka, tedy negativní korelace, kdy s rostoucí spotřebou klesají úspory. Korelace mezi spotřebou a příjmy (r = 0,2813) se jeví statisticky významnou až na 13% hladině významnosti. Vzhledem k uvedeným výsledkům byl proveden výpočet parciálních korelačních koeficientů, kdy posuzované proměnné byly očištěny o vliv zbývajících proměnných, přičemž výchozí skupinu proměnných tvořily spotřeba, příjmy, úspory a inflace. Porovnáme-li zjištěné výsledky parciálních korelačních koeficientů s korelačními koeficienty uvedenými v tabulce 1, části II. vidíme, že významněji se nezměnila hodnota korelace mezi spotřebou a inflací (r = -0,4025) a dále mezi příjmy a úsporami (r = 0,6527). V případě ostatních posuzovaných a očištěných korelací došlo ke změnám. V případě korelace mezi spotřebou a příjmy (r = 0,4005) došlo k významnému nárůstu hodnoty korelace (výběrová korelace činila 0,2831). Analogicky v případě spotřeby a úspor (r = -0,3083) došlo k významnému nárůstu hodnoty korelace (výběrová korelace činila -0,0316). Oba poslední uvedené parciální korelační koeficienty můžeme považovat za indikující střední závislost mezi posuzovanými očištěnými znaky.

Z uvedených výpočtů korelační analýzy lze vyvodit dva následující závěr. Budeme-li posuzovat očištěnou korelaci mezi spotřebou a ostatními faktory, pak se korelace mezi spotřebou a příjmy jeví stejná jako mezi spotřebou a inflací. Tudíž, zapojení proměnné inflace do regresního modelu spotřební funkce je oprávněné odpovídající charakteru posttransformační ekonomiky České republiky.

Předpokládejme, že spotřebitel využívá optimálně dostupné informace v období, kdy se rozhoduje o nákupu, stejně jako v následujícím období. Podle Romera (2001; s. 337-338) můžeme v obecnosti říci, že v každém období je očekávaná spotřeba v příštím období rovna současné spotřebě. Toto implikuje, že změny ve spotřebě jsou nepredikovatelné a chovají se náhodně. V souvislosti se spotřebitelským chováním může být zmíněná náhodnost způsobena neočekávanými událostmi souvisejícími s nutností krytí krátkodobých nepředvídatelných výdajů či neočekávaných potřeb ekonomických subjektů. Podle definice očekávání lze pak toto chování popsat procesem náhodné procházky. Odhad modelu náhodné procházky pro spotřebitelské chování v České republice popisuje následující „model 2“:

$$\begin{array}{lll} est(C_t) & = & 0{,}8263C_{t-1} \\ & & SE=(0{,}1072) \\ & & p\text{-}value=0{,}0000 \\ \hline n=29 & & R^2_{adj}=0{,}6795 \\ \hline d=2{,}1632. & & F=59{,}39 \ (p\text{-}value = 0{,}0000) \end{array}$$

Na základě výsledků odhadu modelu 2 se skutečně můžeme domnívat, že spotřebitelé chování v České republice v krátkodobém horizontu jednají náhodně. Zapojíme-li mezi nezávisle proměnné modelu 1 proměnnou spotřeba zpožděnou o jedno časové období vzad (C_{t-1}), získáme následující odhad regresního modelu s označením „model 3“:

est(C)	=	*0,4816Y_t*	*+ 0,4673C_{t-1}*	*- 0,0504S_t*	*- 0,2756P_t*
		SE=(0,1612)	*(0,1608)*	*(0,0325)*	*(0,1595)*
		p-value=0,0062	*0,0076*	*0,1340*	*0,0965*
n=29			*R^2_{adj}=0,7365*	*F=20,32 (p-value = 0,0000)*	*d=2,0443.*

Výsledný odhad modelu 3 je statisticky významný jako celek s nekorelovanými rezidui. V případě proměnných příjmů (Y_t) a spotřeby (C_{t-1}) jsou parametry významné na 5% hladině významnosti. V případě proměnné úspory (S_t) je parametr statisticky významný na 14% hladině významnosti s negativním znaménkem. Přestože je významnost tohoto parametru nižší než očekávaná, v modelu jej ponecháme, neboť jeho vyloučení by odporovalo předpokladu krátkodobé keynesiánské spotřební funkce, a to rozdělení důchodu spotřebitelů na část spotřeby a úspor. Autorky předpokládají, že při větším rozsahu souboru by se parametr spotřeby projevil jako statisticky významnější v odhadu modelu 3. Poslední zapojená proměnná, inflace (P_t), se jeví v odhadu parametru jako statisticky významná na 10% hladině významnosti s negativním odhadnutým znaménkem. I zde tedy můžeme konstatovat peněžní iluzi, které podléhají spotřebitelé v České republice, jak již bylo komentováno v případě modelu 1.

5. Závěr

Cílem příspěvku je ověření platnosti keynesiánské spotřební funkce pro Českou republiku v období 2001/Q1–2008/Q2. Dále provést rozšíření klasické spotřební funkce o faktor inflace a na základě statistických testů vyhodnotit oprávněnost zapojení tohoto faktoru do spotřební funkce. V souvislosti s výsledky empirické analýzy nově navrhovaného faktoru inflace následně vytvořit upravenou spotřební funci pro Českou republiku a zjištěné výsledky interpretovat.

V první fázi analýzy byl proveden odhad krátkodobé keynesiánské spotřební funkce. Vzhledem ke statistickým vlastnostem výsledného odhadu došlo k doplnění modelu. Nejprve byl model doplněn o proměnnou inflace. Jako doplňující zde byla provedena korelační analýza. Z výsledků korelační analýzy vyplývá, že budeme-li posuzovat očištěnou korelaci mezi spotřebou a ostatními faktory, pak se korelace mezi spotřebou a příjmy jeví stejně významná jako mezi spotřebou a inflací. Tudíž, zapojení proměnné inflace do regresního modelu spotřební funkce je oprávněné odpovídající charakteru posttransformační ekonomiky České republiky. A dále, korelace mezi spotřebou a příjmy, úsporami a příjmy je pozitivní, resp. mezi spotřebou a úsporami negativní. Toto lze chápat jako potvrzující keynesiánské pojetí krátkodobé spotřební funkce a rozdělením důchodu spotřebitelů na část spotřeby a úspor. Druhou úpravou regresního modelu popisující spotřební funkci bylo zapojení faktoru náhodnosti. Autorky zde vyšly z předpokladu, že spotřebitel využívá optimálně dostupné informace v období, kdy se rozhoduje o nákupu, stejně jako v následujícím období. Tedy, že v každém období je očekávaná spotřeba v příštím období rovna současné spotřebě. Toto implikuje, že změny ve spotřebě jsou nepredikovatelné a chovají se náhodně. V souvislosti se spotřebitelským chováním může být zmíněná náhodnost způsobena neočekávanými událostmi souvisejícími s nutností krytí krátkodobých nepředvídatelných výdajů či neočekávaných potřeb ekonomických subjektů. Podle definice očekávání lze pak toto chování popsat procesem náhodné procházky. Skupina nezávislých proměnných byla proto rozšířena o proměnnou spotřeby zpožděnou o jedno časové období vzad.

Získaný regresní model popisující spotřební funkci v České republice, tedy spotřebu obyvatel jako závislá proměnná, byl koncipován na základě nezávislých proměnných příjmů obyvatel, spotřeby zpožděné o jedno časové období vzad, úspor obyvatel a inflace. Statistické vlastnosti modelu byly vyhovující s výjimkou významnosti proměnné úspory (14%). Vzhledem k tomu, že proměnná úspory vychází ze samotného keynesiánského pojetí krátkodobé spotřební funkce nedomnívají se autorky, že by v tomto případě šlo

o nadbytečnou proměnnou. Důvod snížené statistické významnosti shledávají v rozsahu souboru dat, kdy při jeho rozšíření lze očekávat vylepšení statistické významnosti. V případě zapojení proměnné inflace se projevilo, že vnímání spotřebitele v České republice je ovlivněno inflací. Pokud by tomu tak nebylo, proměnná inflace by se v odhadnutém modelu jevila nevýznamnou. V případě ČR však z negativního znaménka odhadu můžeme usuzovat, že dochází k tzv. peněžní iluzi, kdy spotřebitelé předpokládají, že roste-li inflace, klesá jejich reálný příjem i úspory, přestože tomu tak není, a proto mají tendenci ke snižování svých spotřebitelských výdajů. Svým způsobem tudíž zaznamenají zvýšení cen, avšak nikoliv ekviproporcionální růst důchodů i úspor v peněžním vyjádření.

6. Literatura

[1]Eurostat [online] [cit. 2008-10-02]. Dostupné z: <http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1090,30070682,1090_30298591&_dad=portal&_schema=PORTAL>

[2]HEBÁK, J., HUSTOPECKÝ, J., MALÁ, J. 2005. Vícerozměrné statistické metody (2). Informatorium, Praha. 2005. ISBN 80-7333-036-9

[3]HUŠEK, R., PELIKÁN, J. 2003. Aplikovaná ekonometrie, teorie a praxe, Professional Publishing, Praha, 2003. 263 s., ISBN 80-86419-29-0

[4]ROMER, D. 2001. Advanced macroeconomic. Published by McGraw-Hill/Irwin company, 2001. 651 s., ISBN 0-07-231885-4

[5]Statistika rodinných účtů [online] Poslední revize 23. 3. 2005 [cit. 2008-09-16]. Dostupné z: <http://www.czso.cz/csu/2004edicniplan.nsf/publ/3005-04-za_4__ctvrtleti_2004>

[6]ŠIRŮČEK, P. a kol. 2007 Hospodářské dějiny a ekonomické teorie (Vývoj – současnost – výhledy), Melandrium, Slaný, 2007. 511 s. ISBN 978-80-86175-03-4

[7]TOUFAROVÁ, Z. 2008 Problémy v metodice ukazatele spotřební výdaje. In: MendelNet PEF 2008. Brno: Mendelova Zemědělská a Lesnická Univerzita v Brně, 2008. s 245. ISBN 978-80-87222-03-4

[8]WOOLDRIGGE, J. M. 2003. Introductory Econometrics: A modern approach. Ohio, 863 s., ISBN 0-324-11364-1.

Adresa autorů:

Jitka Poměnková, RNDr., Ph.D.
Ústav financí, PEF MZLU v Brně
Zemědělská 1
613 00 Brno
pomenka@mendelu.cz

Zuzana Toufarová, Ing.
Ústav marketingu, PEF MZLU v Brně
Zemědělská 1
613 00 Brno
xtoufaro@mendelu.cz

Data mining – moderní způsob získávání informací
Data mining – the modern way of information acquiring

Martin Řezáč

Abstract: Data mining is the process of sorting through large amounts of data and picking out relevant information. It has been described as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Regarding increasing amount of data that are collected and kept in databases, it is still increasing the number of fields where it is possible, and often essential, to use data mining methodology. The contribution deals with basic concepts of data mining from explanation why it is so popular and needed in current days, through methodology (CRISP-DM), up to overview of application fields. A short overview of data mining software is also included.

Key words: Data mining, methodology, CRISP-DM, data mining software, applications.

Klíčová slova: Data mining, metodologie, CRISP-DM, data miningový software, aplikace.

1. Úvod

Data mining (DM), nebo také dolování z dat či vytěžování dat, je analytická metodologie získávání netriviálních skrytých a **potenciálně užitečných informací**. V průběhu času a v závislosti na uživatelích této metodologie se postupně objevovali názvy:

- 1960: Data Fishing, Data Dredging… užíváno statistiky
- 1989: Knowledge Discovery in Databases (KD, KDD)…užíváno komunitou zabývající se umělou inteligencí a strojovým učením
- 1990: Data Mining (DM)…užíváno v komerční sféře a databázové komunitě.

Další názvy, jako Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction a další, byly postupem času zapomenuty a současné době se používá téměř výhradně název Data Mining. Více viz [2].

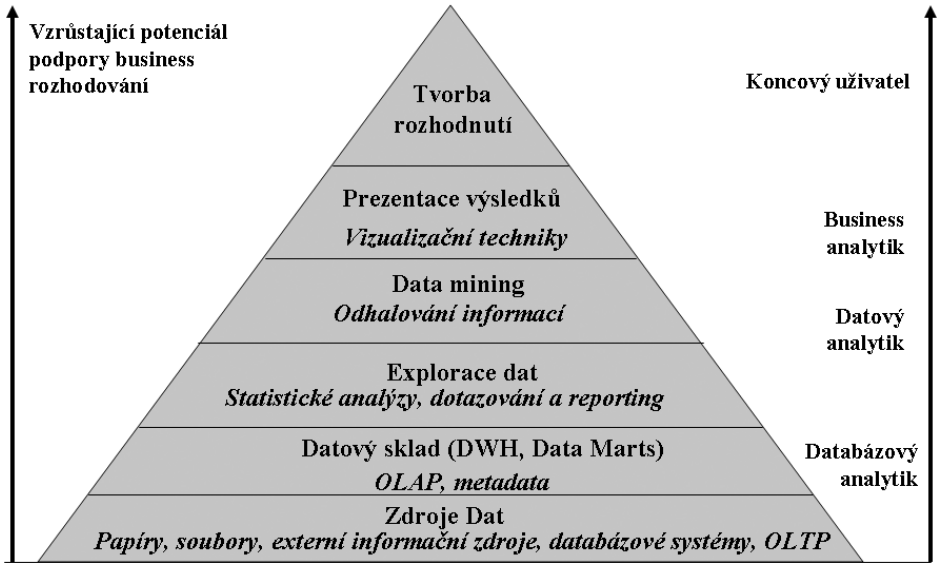
Jedním ze základních principů data miningu je odhalování nových neočekávaných vztahů. Obecně lze říci, že:

- zajímavé jsou ty vztahy, které se liší od obecných očekávání.
- data mining se vyplácí díky objevování doposud neznámých a překvapivých vztahů.

2. Proč data mining

Chápeme-li data mining jakožto podporu business rozhodování, lze využít schématu na obrázku 1, které zobrazuje míru této podpory v závislosti na míře zpracování primárních dat. Na základní úrovni jsou primární data a o úroveň výše jsou data shromážděná v datovém skladu (Data Warehouse, Data Marts). O zpracování dat na těchto úrovních se většinou starají databázoví odborníci. Následující dvě úrovně typicky představují vytváření různých pravidelných reportů, ad hoc statistických analýz a data miningových analýz, které zpracovávají analytici různých oddělení. Ve spojení s business analytiky jsou pak vytvářeny prezentace výsledků, na základě kterých je již management firmy schopen činit svá rozhodnutí. Více viz [1].

Pokud jde o úroveň datového skladu, je naprosto nezbytné, aby na jeho návrhu spolupracovali všechny zmíněné profese, tj. od databázových odborníků, přes datové a business analytiky až po zástupce managementu.

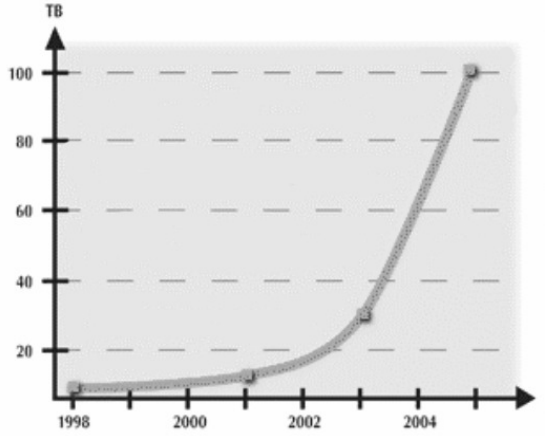


Obrázek.1 :Data mining jakožto podpora business rozhodování. Zdroj:[2]

Dalším aspektem, proč se data mining prosazuje čím dál častěji, je rozsah dat, která jsou k dispozici a ze kterých je tedy možné dolovat informace. V roce 2005 patřili mezi největší databáze světa:

- Max Planck Inst. for Meteorology ~ 222 TB
- Yahoo ~ 100 TB
- AT&T ~ 94 TB

V současné době se velikost největších databází pohybuje již v řádech yottabajtů, tj. 10^{24} bajtů. Protože terabajt představuje 10^{12} bajtů, jsou dnešní nejrozsáhlejší databáze 10^{12}-krát, tj. bilion-krát, větší než před třemi lety. Pro představu jak ohromný je to nárůst je na obrázku 2 uveden graf představující velikost největších světových databází od roku 1998 do roku 2005. V posledních 3 letech, tj. 2005-2008, je nárůst ještě strmější.



Obrázek 2: Největší světové databáze v r. 2005. Zdroj [3].

Co je zdrojem tak ohromných dat? Z pohledu velikosti databází je možné uvést následující příklady: data obchodních řetězců a bank (terabajty, tj. 10^12 B), geografická data (petabajty, tj. 10^15 B), národní databáze zdravotních záznamů v USA (exabajty, tj. 10^18 B), meteorologická data (zettabajty, tj. 10^21 B), video-databáze a data z kosmických teleskopů

(yottabajty, tj. 10^24 B). Celkem lze mezi hlavní důvody stále častějšího využití data miningu v dnešní době zařadit:

- Data jsou produkována.
- Data jsou skladována.
- Výpočetní síla je dostupná.
- Výpočetní síla je cenově dostupná.
- Konkurenční tlak je velice silný.
- Komerční produkty (software) jsou k dispozici.

3. Metodologie data miningu

Data mining se často definuje jakožto proces (polo-) automatické analýzy (rozsáhlých) databází k identifikaci vztahů, které jsou:

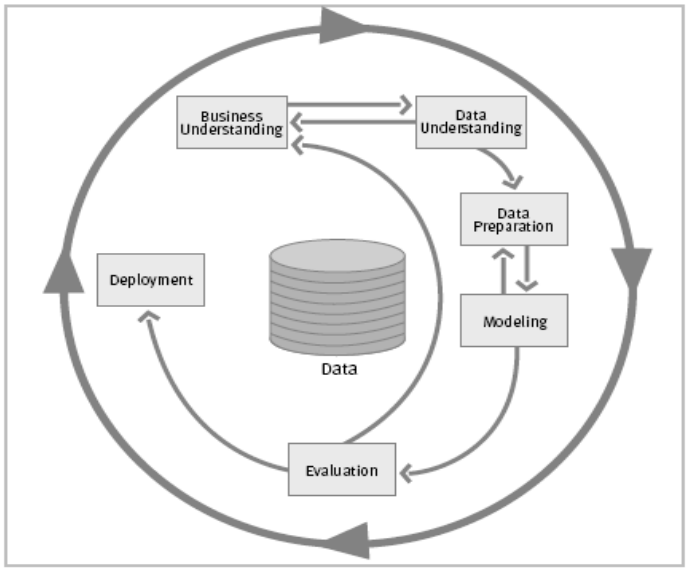
- Validní: platí na nových datech s určitou jistotou obecné platnosti.
- Nové: doposud neznámé.
- Užitečné: dají se v praxi nějak použít.
- Srozumitelné: (vždy) se nalezený vztah dá nějak vysvětlit.

První tři atributy jsou zcela logické a všichni „uživatelé“ data miningu se na nich shodnou. Problém někdy nastává u čtvrtého atributu. Ačkoli zní požadavek na vysvětlitelnost identifikovaných vztahů logicky, stává se, že vysvětlení vztahu není snadné nalézt nebo se jej nalézt vůbec nepodaří. V takovém případě je třeba zvážit zda jsme ochotni v praxi použít vztah, kterému sice nerozumíme, ale jeho použití významně zvyšuje efektivitu naší činnosti, tj. např. zisk firmy.

Dolování dat je dnes již velmi vyspělá disciplína, takže pro její nasazení do praxe bylo vyvinuto několik standardů. Mezi dva nejpoužívanější patří CRISP-DM a SEMMA. Metodologie CRISP-DM *(CRoss Industry Standard Process for Data Mining)* je vyvíjena nezávisle na dolovacím nástroji, nicméně u zrodu tohoto projektu stála firma SPSS. Naproti tomu konkurenční standard SEMMA *(Sample, Explore, Modify, Model, Assess)* je výhradním dílem firmy SAS. Metodologie CRISP-DM popisuje dolování dat jako proces sestávající z následujících kroků:

- Pochopení obchodních, či jiných důležitých, souvislostí.
- Pochopení dat.
- Příprava dat.
- Modelování.
- Vyhodnocení modelu.
- Nasazení modelu do praxe.

Jednotlivé kroky přitom na sebe navazují pouze formálně, protože v procesu modelování často zjistíme, že data je nutno připravit trochu jiným způsobem a musíme se proto vrátit o krok zpět. V závěru z vyhodnocení získaných modelů získáme další poznatky, např. o širších obchodních souvislostech, což opět často vede k návratu do fáze přípravy dat. Schematicky je celý proces zachycen na obrázku 3.



Obrázek 3: Metodologie CRISP-DM.Zdroj: [4].

Mohlo by se zdát, že samotné modelování je klíčovým bodem celého procesu. Do určité míry je to sice pravda, ovšem tento krok v praxi zabírá pouze nepatrný zlomek zdrojů, neboť je již do značné míry zautomatizován vyspělými dolovacími nástroji. Těžiště úspěchu data miningového projektu stále zůstává v prvních třech výše uvedených bodech: správné stanovení cíle, pochopení dat a příprava dat totiž často zabírá cca 90% celkového času. Samotná příprava dat totiž v sobě skrývá řadu činností, jako jsou:

- Sběr dat:
  - Mnohdy jsou data uložena v několika databázích založených na různých platformách (Oracle, MS SQL, Excel, …).
  - Zdrojem dat mohou být veřejné databáze na internetu nebo jiné externí databáze.
- Konsolidace a čištění:
  - Využití/vytvoření vazebních tabulek, agregace hodnot, diskretizace spojitých ukazatelů,…
  - Odstranění chybějících hodnot (nahrazení průměrem/mediánem/modem).
  - Odstranění odlehlých pozorování – nahrazení maximální/minimální hodnotou/průměrem/mediánem/modem, …
- Selekce:
  - Ignorování neužitečných dat.
  - Výběr dat – v případě rozsáhlých dat je vhodné a některých případech dokonce nutné pracovat pouze s náhodných výběrem z dostupných dat.
- Transformace – vytváření nových odvozených proměnných.

Jak už bylo zmíněno výše, proces modelování je do značné míry zautomatizován. Nicméně již z povahy dat, širších souvislostí, případně cílů, které si stanovíme na samém počátku data miningového projektu, často vyplývá jaký typ modelu bude použit. Základní typy modelů jsou modely prediktivní (regrese, rozhodovací stromy, neuronové sítě,…) a deskriptivní (klastrová analýza, asociační pravidla,…).

K vyhodnocení modelu jsou většinou použity různé statistické charakteristiky jako jsou Kolmogorovova-Smirnovova statistika, Giniho index, Lift chart a další. Další možností je pak vyhodnocení pomocí celkové přínosu daného modelu, tj. např. odhad zisku, který bude

generován při použití modelu. Často jsou také modely podrobovány různým testům stability, testům citlivosti na změnu vstupních údajů atp.

Nasazení modelu do praxe typicky představuje algoritmizaci nalezeného modelu do produkčního systému. S tím je spojena tvorba různých případových studií a testování shody výstupů modelu připraveném v modelovacím nástroji a výstupů modelu nově integrovaného v produkčním systému. V některých případech jsou ovšem získané výstupy data miningového procesu natolik přímočaré, že jejich nasazení do praxe je zcela triviální.

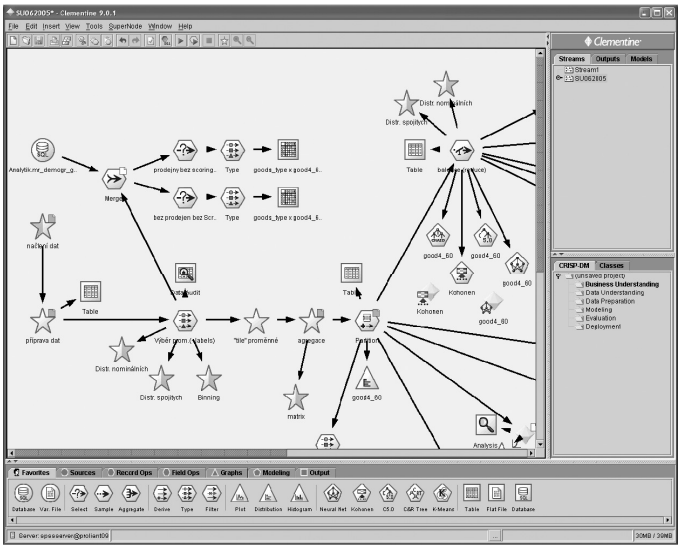
Princip SEMMA je v porovnání s metodologií CRISP-DM o něco specifičtější a zaměřuje se na jádro dolovacího procesu, v němž je zapotřebí provést tyto kroky:

- *Sample* – identifikovat vhodná učící data, určit odpovídající rozsah dat, a to jak z pohledu časového okna tak i z pohledu počtu případů.
- *Explore* – připravit popisné statistiky, které poskytnou základní představu o obsahu a kvalitě podkladových dat.
- *Modify* – na základě předchozího kroku konsolidovat data a odvodit nové proměnné. Následně transformovat data do tvaru vhodného pro modelování.
- *Model* – vytvořit příslušný model, přičemž beze zbytku platí poznámka uvedená výše.
- *Assess* – vyhodnotit úspěšnost modelu a případně implementovat model do praxe.

4. Software

V současné době existuje cca 20 až 30 dodavatelů data miningového softwaru. Mezi nejvýznamnější lze zařadit:

- Clementine,
- IBM's Intelligent Miner,
- SGI's MineSet,
- SAS's Enterprise Miner,
- STATISTICA Data Miner.



Obrázek 4: Vývoj skóringové funkce v systému Clementine. Zdroj: vlastní konstrukce.

Charakteristickým znakem těchto systémů je jejich uživatelská přívětivost, kdy je s daty manipulováno pomocí grafického uživatelského rozhraní (GUI). Na obrázku 4 je příklad pracovní plochy v systému Clementine.

5. Aplikace

Data mining je v současné době aplikován v opravdu široké řadě oblastí. Mezi nejvýznamnější patří bankovnictví (schvalování úvěrů/kreditních karet pomocí predikce dobrých zákazníků), CRM (churn modely, cross-sell, up-sell, cílený marketing), fraud management (online detekce podvodného chování), medicína (optimalizace léčebné péče), farmacie (identifikace nových léků), vědecká analýza dat (identifikace nových galaxií), design webových stránek (změna podoby stránek na míru aktuálního návštěvníka stránek), supermarkety (identifikace současně nakupovaného zboží), průmysl (automatické přenastavení ovládacích prvků při změně parametrů procesu). Své uplatnění si však data mining našel i takových oblastech jako je rozpoznávaní textu, řeči a obrázků nebo dokonce sport – v NBA je používán k optimalizaci herní strategie.

6. Závěr

Klasická statistická analýza často staví na modelech kladoucích řadu předpokladů na zkoumaná data. Navíc nejsou klasické modely a výpočetní postupy ani zdaleka optimalizovány pro rozsáhlé datové soubory. Naproti tomu data miningové postupy na data nekladou žádné nebo jen minimální předpoklady a lze pomocí nich zpracovávat extrémně rozsáhlá data. Mnohem větší důraz je kladen na porozumění datům z business pohledu než na znalost pokročilých statistických metod. Díky produkci ohromného množství dat, schopnosti tato data uchovávat, dostupnosti dostatečné výpočetní síly a existenci sofistikovaného data miningového softwaru je v současnosti data mining užit v široké řadě aplikačních oblastí.

7. Literatura

[1] HAN, J. – KAMBER, M. 2000. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers. 550 s. ISBN 1-55860-489-8.

[2] HAN, J. – KAMBER, M. 2000. Data Mining: Concepts and Techniques [online]. [cit.2008-11-11] URL: <http://www.cs.sfu.ca/~han/bk/1intro.ppt >.

[3] DITTRICH, J.P., 2007. Data Warehousing FS 2007 [online]. [cit. 2008-11-11]. URL: <http://www.dbis.ethz.ch/education/ ss2007/07_dbs_datawh/Data_Mining.pdf>.

[4] CHAPMAN, P., at al. CRISP-DM 1.0 [online]. [cit. 2008-11-11]. URL: <http://www.spss.com/media/collateral/CRISP-DM_1.0__Step-by-Step_Data_Mining_Guide.pdf>

Adresa autora:

Martin Řezáč, Mgr. Ph.D.
Ústav statistiky a operačního výzkumu PEF MZLU
Zemědělská 1
613 00 Brno
rezac@mendelu.cz

Vizualizace při analýze dat
Visualization in data analyses

Martin Řezáč

Abstract: The amount of huge data sources is still increasisng. Therewith it is still increasing the potency of usage of informations, which are possible to extract from that data. The amount of information, which is possible to get from visualized data, is much higher than in case of primary data files or in case of several numerical characteristics derived from that data files. The visualization can be used in begining stages of data analyses, when it is possible to get insight into the data much more effectively. The role of visualization is totally unsubstitutable in stage of presentation of obtained results.

Key words: Visualization, histogram, joint density function, joint distribution function.

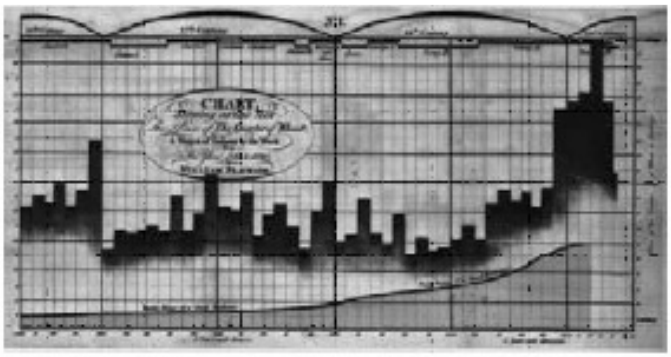
Klíčová slova: Vizualizace, histogram, sdružená hustota, sdružená distribuční funkce.

1. Úvod

V posledních cca dvaceti letech dochází k masivnímu rozvoji zpracování dostupných datových souborů. Zdrojů generujících ohromná množství dat neustále přibývá a s tím rostou i možnosti využití informací, které lze z těchto dat extrahovat. I přes rozvoj počítačové techniky zůstává role lidského faktoru nezastupitelná. Nelze si dost dobře představit, že by od základního zpracování dat přes jejich transformace, tvorbu různých modelů, zpracování výstupů těchto modelů až po interpretaci výsledků a vyvození patřičných důsledků zvládl pouze a jen počítač. Na druhou stranu lze těžko očekávat, že je člověk schopen efektivně zpravovat stovky či tisíce popisných statistik, indexů a dalších charakteristik odvozených z právě analyzovaných dat. Významnou pomocí je pak vizualizace dat. Lze ji využít již v prvních fázích datové analýzy, kdy je možné mnohem efektivněji získat přehled o charakteru zkoumaných dat. Naprosto nezastupitelnou roli pak má vizualizace ve fázi prezentace dosažených výsledků.

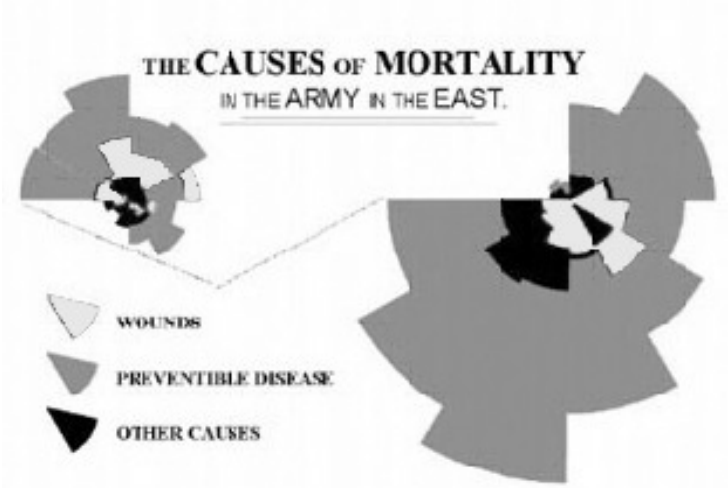
2. Stručná historie vizualizace

I když by se mohlo zdát, že analýza dat a jejich vizualizace je doménou několika posledních desetiletí, první publikovaná prezentační grafika se objevila již v osmnáctém století. Autorem byl William Playfair a ukázka jeho práce z roku 1786 je zobrazena na obrázku 1.



Obrázek 1:První publikovaná prezentační grafika. Zdroj: [1].

Za průkopníka a velkého popularizátora vizualizace je považován Florence Nightingale. V roce 1858 zpracoval a publikoval výsledky analýzy důvodů úmrtí vojáků v průběhu krymské války. Výsledek byl více než překvapivý. Většina vojáků nezemřela v důsledku zranění nýbrž v důsledku nedostatečného dodržovaní základních hygienických pravidel. Vizualizované výsledky analýzy jsou zobrazeny na obrázku 2.

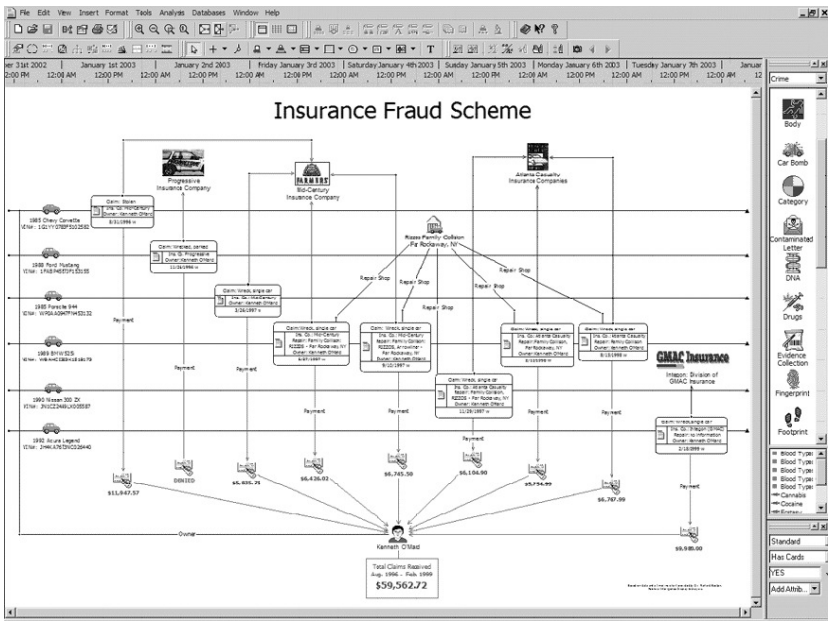


Obrázek 2:Důvody úmrtí v průběhu Krymské války (1853-1856). Zdroj: [1].

3. Současnost vizualizace

Jak už bylo naznačeno v úvodu, je v současné době vizualizace dat využívána mnohem masivněji než historii. Jmenovat lze oblasti jako jsou biologie, geologie, meteorologie, astronomie, bankovnictví, nebo sport. Prezentační grafika je běžnou současí denního tisku a například volební výsledky, výsledky průzkumu veřejného mínění, výsledky různých testů spotřebního zboží a mnoho dalšího si lze již jen těžko představit bez využití právě prezentační grafiky.

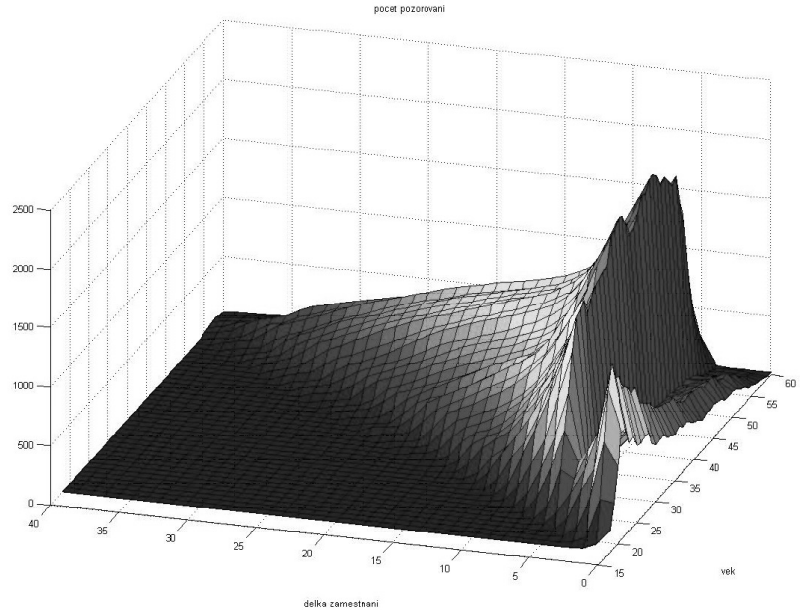
S úspěchem je vizualizace využívána ve finančních institucích, a to jak při vlastní analýze dat, tak i při následné prezentaci získaných výsledků. Na obrázku 3 je uveden příklad využití při odhalování pojistných podvodů.



Obrázek 3: Odhalování pojistných podvodů. Zdroj: [2].

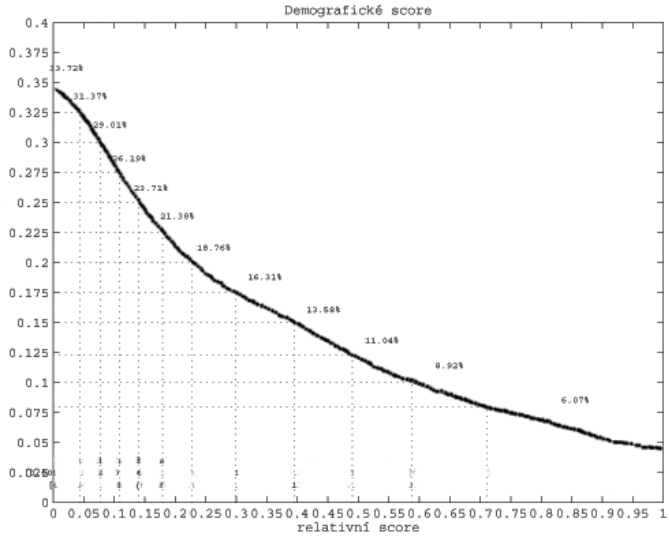
Často je třeba získat představu o datech už na samém počátku jejich analýzy. I obyčejný histogram má mnohem vyšší vypovídací hodnotu než je sada popisných statistik jako je průměr, směrodatná odchylka či v lepším případě hodnoty kvantilů nebo hodnoty šikmosti a špičatosti. V případě sofistikovanějších metod vizualizace je možné získat velmi dobrou představu o celkové povaze zkoumaných dat. Využijeme-li navíc zobrazení ve 3D, jsme schopni analyzovat vztahy mezi dvojicemi datových entit. V tomto případě lze použít prostorový histogram nebo obecně nějaký odhad sdružené pravděpodobnostní hustoty pro danou dvojici datových entit.

Na obrázku 4 je vynesena absolutní četnost klientů jedné finanční společnosti podle věku a délky zaměstnání. Použit byl prostorový histogram zobrazený pomocí spojité plochy spojující hodnoty sdružených četností v jednotlivých třídách.



Obrázek 4: Prostorový histogram. Zdroj: Vlastní konstrukce.

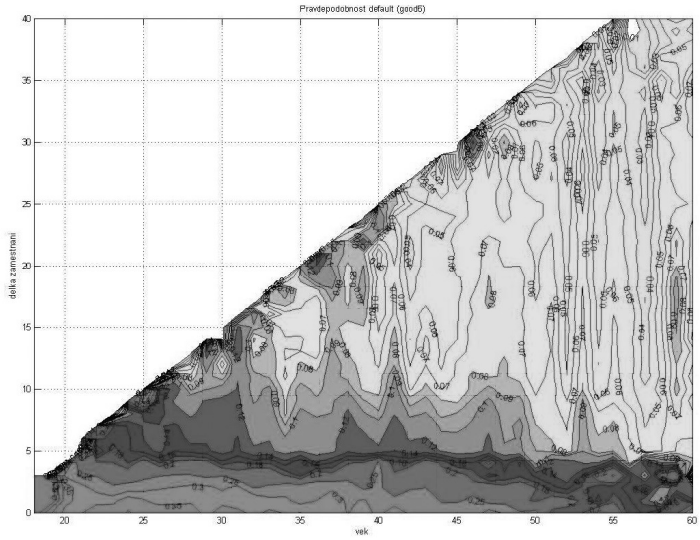
Dalším často využívaným typem vizualizačního postupu je zobrazení zkoumané entity v závislosti na jiné entitě.



Obrázek 5: Závislost default rate na relativním skóre . Zdroj: Vlastní konstrukce.

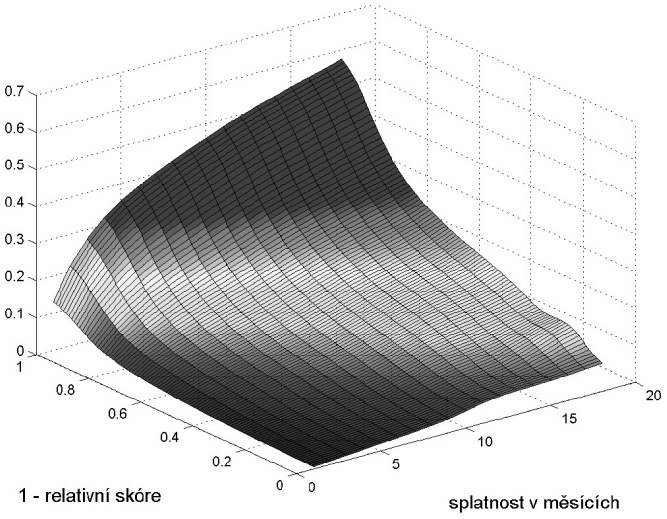
Na obrázku 5 je zobrazena závislost default rate, tj. poměru nesplácejících klientů, na relativním skóre. Pomocí tohoto grafu, mimo jiné, lze ověřit správnost modelu predikujícího default rate pomocí skóre.

Chceme-li zobrazit závislost nějaké entity, tj. například default rate, na dvou jiných entitách, lze využít následujícího typu grafu. Jde o odhad sdružené pravděpodobnostní hustoty default rate vzhledem k věku a délce zaměstnání. Na obrázku 6 jsou zobrazeny vrstevnice odhadnuté hustoty, přičemž je využita barevná škála pro zvýraznění přislušných hodnot. Jak je z tohoto obrázku patrné, nejnižší default rate je pozorován u klientů s věkem nad 40 let a délkou zaměstnání nad 10 let.



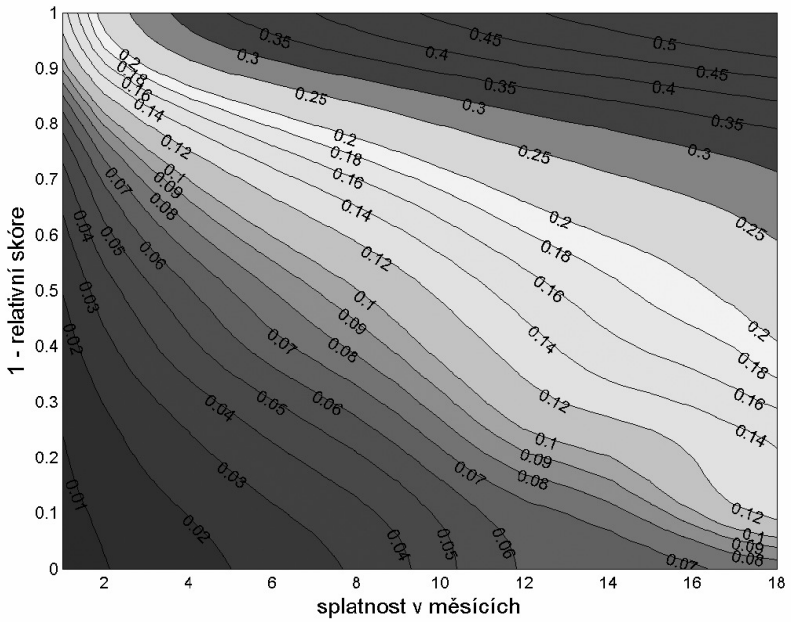
Obrázek 6: Vrstevnice sdružené hustoty default rate. Zdroj: Vlastní konstrukce.

Další typem často užívaného grafu je graf distribuční funkce dané entity. I v tomto případě lze s úspěchem využít 3D zobrazení sdružené distribuční funkce. Na obrázku 7 je uveden příklad odhadu sdružené distribuční funkce doby do defaultu v závislosti na skóre (respektive 1- relativní skóre) a splatnosti poskytovaného úvěru.



Obrázek 7: Sdružená distribuční funkce doby do defaultu . Zdroj: Vlastní konstrukce.

Vyjdeme-li z předchozího grafu a zaměříme-li se na otázku: „je odhad distribuční funkce funkcí neklesající ve všech vertikálních i horizontálních řezech, tj. jsou příslušné marginální distribuční funkce neklesající?“, lze snadno využít následující typ grafu. Jde o vrstevnicový graf srdužené distribuční funkce, ze kterého lze vizuálně snadno ověřit zda je splněn výše zmíněný požadavek, tj. zda ve všech vertikálních i horizontálních řezech je pozorován neklesající průběh. Na obrázku 8 je uveden příklad vrstevnicového grafu sdružené distribuční funkce doby do defaultu v závislosti na skóre (respektive 1- relativní skóre) a splatnosti poskytovaného úvěru.



Obrázek 8: Závislost default rate na relativním skóre . Zdroj: Vlastní konstrukce.

4. Závěr

Vizualizace je v současné době využívána ve stále se zvyšující míře. Množství informace, které lze velmi rychle získat z vizualizovaných dat, je mnohonásobně vyšší než v případě primarních dat nebo několika číselných charakteristik z primárních dat odvozených. I člověk naprosto neznalý jakýchkoli kvantitativních metod má tak možnost snadno získávat nové informace. V případě odborníků pracujících s daty je pak nesporný přínos vizualizačních technik ke zvyšující se efektivitě práce.

5. Literatúra

[1] ZELLWEGER, P. 2006. What Is Information Visualization [online]. [cit. 2008-11-11]. URL: <http://vadl.cc.gatech.edu/documents/62_Stone_L2_What_is_Visualization.pdf>.

[2] I2- Solutions [online]. [cit 2008-11-11]. URL: <http://www.i2inc.com/solutions/fraud/>.

Adresa autora:

Martin Řezáč, Mgr. Ph.D.
Ústav statistiky a operačního výzkumu PEF MZLU
Zemědělská 1
613 00 Brno
rezac@mendelu.cz

Komparace síly vybraných testů normality proti těžkochvostým, lehkochvostým a kontaminovaným alternativám
Comparison of selected tests of normality against heavy tailed, light tailed and contaminated alternatives

Luboš Střelec

Abstract: This paper deals with comparison of selected tests of normality – the Shapiro-Wilk test, the classical Jarque-Bera test, the robust Jarque-Bera test, the Lilliefors (Kolmogorov-Smirnov) test and the Medcouple test. Consequently, paper demonstrates results of simulation studies of power of such tests for the various alternatives as heavy tailed alternatives (Cauchy, Laplace, t_3, t_5, t_7 and logistic distributions), light tailed and very short tailed alternatives (exponential, lognormal, gamma, beta and uniform distributions) and contaminated normal distributions simulated outlier values.

Key words: Tests of normality, Monte Carlo simulation, Power comparison, Heavy tailed alternatives, Light tailed alternatives, Contaminated alternatives.

Klíčová slova: testy normality, Monte Carlo simulace, komparace síly testu, těžkochvosté alternativy, lehkochvosté alternativy, kontaminované alternativy.

1. Úvod

V současné době existuje velké množství odborné literatury zabývající se testy normality a jejich statistickými vlastnostmi (např. Thode (2002)). Mimo klasických formálních testovacích procedur určených k testování normality existují rovněž metody založené na grafickém posouzení či testy testující odlehlé hodnoty či jiné odchylky od normality. Přesto však nejvíce používanými metodami jsou formální testovací procedury, které doplňují prvotní posouzení normality vycházející z grafických metod. Všeobecně nejběžněji využívaným omnibus testem je Shapiro-Wilkův test (Shapiro, Wilk (1965)). V ekonomii či financích je naopak nejvíce využívaný Jarque-Bera test (Jarque, Bera (1980)). Nejznámějším testem založeným na kumulativní distribuční funkci je Lillieforsův (Kolmogorov-Smirnovův) test (Lilliefors (1967)). V současné době se mimo klasických testů objevují také nové testy normality, které jsou založeny na robustních charakteristikách. Jako příklad lze uvést robustní Jarque-Bera test zkonstruovaný Gelovou a Gastwirthem (2007) či testy normality založené na robustní charakteristice šikmosti (medcouple) zkonstruované Brysem et al. (2004).

2. Materiál a metodika

Pro potřeby komparační analýzy bude sledována síla vybraných klasických testů normality – konkrétně Shapiro-Wilkova testu (SW), Lilliefors (Kolmogorov-Smirnovova) testu (LT), Jarque-Bera testu (JB) – a vybraných robustních testů normality – konkrétně robustního Jarque-Bera testu (RJB) a medcouple testu (MC), který využívá medcouple charakteristiku pro celý soubor v kombinaci s medcouple charakteristikou pro hodnoty menší než medián a medcouple charakteristikou pro hodnoty větší než medián (Brys et al. (2004)). Pro všechny sledované testy bude využito exaktních kritických hodnot získaných na základě Monte Carlo simulací, které pro malé a středně velké rozsahy souborů v případě Jarque-Bera testu, robustního Jarque-Bera testu a medcouple testu v porovnání s aproximací chí-kvadrát rozdělení věrohodněji zachycují rozdělení testových statistik těchto testů normality (viz níže).

Jako alternativy jsou zvoleny vybraná rozdělení s těžkými chvosty (Cauchyho rozdělení, Laplaceovo rozdělení, Studentovo *t*-rozdělení se třemi, pěti a sedmi stupni volnosti a logistické rozdělení), rozdělení s lehkými či velmi krátkými chvosty (lognormální rozdělení, exponenciální rozdělení, gamma rozdělení, beta rozdělení a uniformní (rovnoměrné) rozdělení) a kontaminovaná normální rozdělení (CN) daná následující funkcí

$$CN = (1-p)N(\mu_1, \sigma_1^2) + pN(\mu_2, \sigma_2^2), \tag{1}$$

kde $0 \le p \le 1$. V naší komparační analýze budeme konkrétně předpokládat následující hodnoty jednotlivých parametrů: $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 \in \{0,1,2,3,4,5\}$, $\sigma_2^2 \in \{1,9,25\}$ a $p = 0{,}1$ a 0,5. Výběr těchto parametrů umožňuje analyzovat unimodální i binominální rozdělení, symetrická rozdělení s krátkými i těžkými chvosty či asymetrická rozdělení s různými hodnotami šikmosti. Kontaminovaná normální distribuční funkce CN uvedená v (1) pak může být interpretována následovně: s pravděpodobností $(1 - p)$ pocházejí hodnoty z normálního rozdělení se střední hodnotou $\mu_1 = 0$ a rozptylem $\sigma_1^2 = 1$ a s pravděpodobností p pocházejí hodnoty z normálního rozdělení se střední hodnotou μ_2 (pro $\mu_2 \in \{0,1,2,3,4,5\}$) a rozptylem σ_2^2 (pro $\sigma_2^2 \in \{1,9,25\} \rightarrow \sigma_2 \in \{1,3,5\}$).

3. Komparace síly sledovaných testů normality

Tabulka 1 uvádí empirické kritické hodnoty sledovaných testů normality pro hladinu významnosti $\alpha = 0{,}05$ a pro rozsahy souborů n = 25, 100, 200, 500, 1000 a 5000. Pro JB, RJB a MC testy jsou rovněž uvedeny asymptotické kritické hodnoty příslušeného chí-kvadrát rozdělení – pro JB a RJB testy se jedná o kvantil $\chi^2_{1-\alpha}(2)$ a pro MC test o kvantil $\chi^2_{1-\alpha}(3)$. Z uvedené tabulky je patrné, že JB test založený na asymptotických kritických hodnotách je pro uvedené rozsahy souborů konzervativní, tj. hypotéza o normalitě zůstává nezamítnuta častěji než odpovídá empirickému rozdělení JB testové statistiky. Aproximace chí-kvadrát rozdělením tak není vhodná jak pro malé, tak dokonce i velké rozsahy souborů, jelikož konvergence empirických hodnot k asymptotickým hodnotám je velmi pomalá. Uvedený poznatek lze nalézt i v jiných studií zabývajících se komparacemi empirických a asymptotických kritických hodnot JB testové statistiky – např. Würtz a Katzgraber (2005) či Thadewald a Büning (2007).

Podobně i MC test založený na asymptotických hodnotách je pro sledované rozsahy souborů s výjimkou pro $n = 5000$ konzervativní. Naopak RJB test založený na asymptotických hodnotách je pro malé a středně velké rozsahy souborů ($n \le 200$) příliš striktní, zatímco pro velké rozsahy souborů ($n > 200$) konzervativní. Ve všech uvedených případech se tak jako vhodnější jeví využití empirických kritických hodnot.

Tabulka 5: Empirické kritické hodnoty pro hladinu významnosti $\alpha = 0{,}05$ a různé rozsahy souborů

	25	100	200	500	1000	5000	$\rightarrow\infty$
SW	0,919	0,975	0,986	0,994	0,997	0,999	.
LT	0,173	0,089	0,063	0,040	0,029	0,013	.
JB	4,156	5,423	5,668	5,847	5,930	5,976	5,991
RJB	7,289	6,496	6,100	5,842	5,782	5,706	5,991
MC	6,807	7,182	7,462	7,689	7,735	7,885	7,815

Pozn.: Počet Monte Carlo simulací je 100000.

Tabulka 2 uvádí sílu sledovaných testů normality proti těžkochvostým alternativám zahrnující Cauchyho rozdělení, Laplaceovo rozdělení, Studentovo *t*-rozdělení se třemi, pěti a sedmi stupni volnosti a logistické rozdělení jednak pro malý rozsah souboru $n = 25$, tak

i pro středně velký rozsah souboru $n = 100$ a to vše pro hladinu významnosti $\alpha = 0{,}05$. Z uvedených výsledků je patrné, že největší sílu proti těžkochvostým alternativám má RJB test, následován JB testem a SW testem. Naopak nejnižší sílu má MC test, jenž např. v případě Cauchyho rozdělení pro středně velký rozsah souboru zamítá hypotézu o normalitě pouze v 65,3 % případů či v případě Laplaceova rozdělení dokonce pouze ve 13,3 % případů.

Tabulka 2: Komparace síly testů – těžkochvosté alternativy pro hladinu významnosti $\alpha = 0{,}05$

	n=25					n=100				
	SW	**LT**	**JB**	**RJB**	**MC**	**SW**	**LT**	**JB**	**RJB**	**MC**
Cauchy	0,924	0,906	0,922	***0,949***	0,172	***1,000***	***1,000***	***1,000***	***1,000***	0,653
Laplace	0,307	0,255	0,353	***0,417***	0,054	0,799	0,710	0,798	***0,889***	0,133
t_3	0,402	0,301	0,459	***0,487***	0,052	0,876	0,736	0,905	***0,926***	0,092
t_5	0,218	0,146	0,269	***0,283***	0,049	0,562	0,333	0,645	***0,674***	0,060
t_7	0,149	0,100	0,190	***0,197***	0,049	0,371	0,182	0,461	***0,481***	0,053
logistic	0,130	0,088	0,169	***0,177***	0,050	0,305	0,157	0,394	***0,422***	0,054

Pozn.1: Počet Monte Carlo simulací je 100000.
Pozn.2: Zvýrazněny jsou testy s nejvyšší sílou pro danou alternativu a rozsah souboru.

Tabulka 3 pak uvádí sílu sledovaných testů normality proti lehkochvostým alternativám zahrnující lognormální rozdělení, exponenciální rozdělení, gamma (2,1) rozdělení a alternativám s velmi krátkými chvosty zahrnující beta rozdělení a uniformní (rovnoměrné) rozdělení opět pro malý rozsah souboru $n = 25$ i pro středně velký rozsah souboru $n = 100$ a to vše pro hladinu významnosti $\alpha = 0{,}05$. Z výsledků uvedených v Tabulce 3 je patrné, že největší sílu proti těmto alternativám má SW test, následován pro lehkochvosté alternativy JB testem a pro alternativy s velmi krátkými chvosty MC a LT testy pro malé rozsahy souborů a JB, LT a MC testy pro středně velké rozsahy souborů. Naopak RJB test nedokáže alternativy s velmi krátkými chvosty odhalit ani ve středně velkých souborech.

Tabulka 3: Komparace síly testů – lehkochvosté alternativy a alternativy s velmi krátkými chvosty pro hladinu významnosti $\alpha = 0{,}05$

	n=25					n=100				
	SW	**LT**	**JB**	**RJB**	**MC**	**SW**	**LT**	**JB**	**RJB**	**MC**
lnorm (0,1)	***0,975***	0,876	0,903	0,866	0,510	***1,000***	***1,000***	***1,000***	***1,000***	0,989
exp (1)	***0,924***	0,694	0,737	0,662	0,404	***1,000***	***1,000***	***1,000***	***1,000***	0,957
gamma (2,1)	***0,650***	0,395	0,500	0,435	0,174	***1,000***	0,954	0,995	0,979	0,632
beta (2,2)	0,063	0,054	0,003	0,003	***0,067***	***0,458***	0,156	0,047	0,000	0,097
uniform	***0,281***	0,118	0,001	0,001	0,128	***0,997***	0,589	0,742	0,007	0,324

Pozn.1: Počet Monte Carlo simulací je 100000.
Pozn.2: Zvýrazněny jsou testy s nejvyšší sílou pro danou alternativu a rozsah souboru.

Tabulky 4 a 5 uvádějí sílu sledovaných testů normality proti kontaminovaným normálním rozdělením s různou vahou kontaminace a různými parametry μ_2 a σ_2^2. V tabulce 4 je uvedena síla sledovaných testů normality pro váhu kontaminace $p = 0{,}10$ pro rozsahy souborů $n = 25$ a $n = 100$. Váha kontaminace $p = 0{,}10$ umožňuje zkoumat jak symetrické alternativy s lehkými či těžkými konci (pro hodnoty parametrů $\mu_2 = 0$), tak i asymetrické alternativy (pro hodnoty parametrů $\mu_2 \neq 0$) a především umožňuje zkoumat robustnost sledovaných testů díky přítomnosti několika odlehlých hodnot v souboru.

V případě symetrických alternativ $(\mu_2 = 0)$ s lehkými či těžkými konci $(\sigma_2 \in \{1,3,5\})$ pro váhu kontaminace $p = 0{,}10$ je z Tabulky 4 patrné, že největší sílu testu mají JB a RJB testy následovány SW testem, LT testem a MC testem. Síla JB, RJB, SW a LT testů roste

s růstem rozptylu σ_2^2 a tím i s růstem špičatosti daného rozdělení (viz Tabulka 6). Naopak síla MC testu roste s růstem rozptylu σ_2^2 pouze nepatrně, přičemž vykazuje sílu testu rovnou přibližně hodnotě zvolené hladiny významnosti. V případě asymetrických alternativ $(\mu_2 \neq 0)$ pro $\sigma_2 \in \{1,3,5\}$ jsou nejsilnější testy JB a RJB test, které mají přibližně shodnou sílu, následovány SW testem a LT testem. MC test opět vykazuje nízkou sílu testu a to jak pro malý, tak i středně velký rozsah souboru. Nízká síla MC testu je však způsobena jeho robustností, kdy hypotézu o normalitě nezamítá tak často jako ostatní sledované testy, které nejsou natolik robustní. MC testu tak lze úspěšně využít tam, kde lze v souboru očekávat několik málo odlehlých hodnot, které nelze jednoduše identifikovat a díky čemuž ostatní testy hypotézu o normalitě zamítají.

Tabulka 4: Komparace síly testů – kontaminované normální rozdělení pro p = 0,10 a hladinu významnosti $\alpha = 0,05$

σ_2	μ_2	n=25					n=100				
		SW	**LT**	**JB**	**RJB**	**MC**	**SW**	**LT**	**JB**	**RJB**	**MC**
1	**0**	0,046	0,048	0,048	0,048	0,051	0,050	0,049	0,053	0,051	0,050
	1	0,056	0,049	***0,058***	***0,058***	0,049	0,060	0,055	***0,063***	***0,063***	0,051
	2	0,107	0,080	***0,126***	0,123	0,051	0,280	0,170	0,283	***0,287***	0,071
	3	0,343	0,219	***0,376***	***0,376***	0,062	0,876	0,678	0,861	***0,874***	0,199
	4	0,723	0,503	0,738	***0,741***	0,095	**0,999**	0,987	***0,999***	***0,999***	0,408
	5	0,947	0,784	0,947	***0,948***	0,114	**1,000**	***1,000***	***1,000***	***1,000***	0,564
3	**0**	0,379	0,229	***0,453***	0,452	0,052	0,824	0,504	***0,877***	0,876	0,057
	1	0,414	0,259	***0,485***	0,480	0,052	0,868	0,592	***0,908***	0,907	0,060
	2	0,523	0,350	***0,587***	0,584	0,057	0,941	0,782	0,960	***0,961***	0,086
	3	0,659	0,484	***0,707***	0,703	0,072	0,989	0,928	***0,992***	***0,992***	0,134
	4	0,796	0,631	***0,835***	0,829	0,086	***0,999***	0,987	***0,999***	***0,999***	0,196
	5	0,883	0,764	***0,908***	0,904	0,095	***1,000***	0,998	***1,000***	***1,000***	0,288
5	**0**	0,705	0,544	***0,750***	0,749	0,060	0,992	0,949	***0,995***	***0,995***	0,070
	1	0,716	0,564	0,763	***0,767***	0,062	0,992	0,951	***0,996***	0,995	0,074
	2	0,744	0,599	0,785	***0,786***	0,062	0,997	0,975	***0,998***	***0,998***	0,088
	3	0,782	0,655	***0,819***	0,818	0,067	0,998	0,984	0,999	***1,000***	0,110
	4	0,843	0,729	***0,873***	0,872	0,073	***1,000***	0,994	***1,000***	***1,000***	0,153
	5	0,880	0,788	***0,904***	0,903	0,092	***1,000***	0,998	***1,000***	***1,000***	0,196

Pozn.1: Počet Monte Carlo simulací je 100000.
Pozn.2: Zvýrazněny jsou testy s nejvyšší sílou pro danou alternativu a rozsah souboru.

V tabulce 5 je uvedena síla sledovaných testů normality pro váhu kontaminace $p = 0,50$ pro rozsahy souborů $n = 25$ a $n = 100$. Váha kontaminace $p = 0,50$ umožňuje zkoumat jednak alternativy symetrické unimodální s lehkými a těžkými konci (pro $\mu_2 = 0$ a $\sigma_2 \in \{1,3,5\}$) či alternativy asymetrické (pro $\mu_2 \neq 0$ a $\sigma_2 \neq 1$), tak i alternativy symetrické bimodální (pro $\mu_2 \geq 3$ a $\sigma_2 = 1$) vyznačující se mírně podnormální špičatostí.

V případě symetrických unimodálních alternativ s lehkými i těžkými chvosty ($\mu_2 = 0$ a $\sigma_2 \in \{1,3,5\}$) vykazuje podobně jako u těžkochvostých rozdělení největší sílu RJB test a naopak nejnižší sílu MC test. Nízká síla MC testu lze opět připsat na vrub jeho velké robustnosti, avšak zde je již tato vlastnost spíše na škodu, neboť váhu kontaminace $p = 0,50$ nelze považovat za váhu, která by simulovala malé množství odlehlých hodnot.

V případě symetrických bimodálních alternativ ($\mu_2 \geq 3$ a $\sigma_2 = 1$) vykazuje nejvyšší sílu SW test. Naopak nulovou sílu vůči těmto alternativám pro malé rozsahy souborů mají JB test a RJB test. Pro středně velký rozsah souboru ($n = 100$) je síla JB testu pro $\mu_2 \geq 4$ již

srovnatelná s SW testem, zatímco síla RJB testu je srovnatelná až pro $\mu_2 \geq 5$. Důvodem je symetrie rozdělení a mírně podnormální špičatost, což zapříčiní, že pro malé rozsahy souborů JB a RJB testy nedokáží tato bimodální rozdělení identifikovat. Síla MC testu proti bimodální alternativám je nízká jak pro malé rozsahy souborů, tak i pro středně velké rozsahy souborů.

V případě asymetrických alternativ s různě těžkými chvosty ($\mu_2 \neq 0$ a $\sigma_2 \neq 1$) se jako nejvhodnější jeví SW test a LT test. JB, RJB a MC testy mají sílu mírně nižší, přičemž je nutné zdůraznit, že MC test je v některých případech dokonce silnější než JB a RJB testy. Důvodem je asymetrie rozdělení v kombinaci se špičatostí, jejíž hodnoty se nepříliš liší od hodnoty 3, což je hodnota špičatosti pro normální rozdělení. Např. pro $n = 25$, $\mu_2 = 5$ a $\sigma_2 = 3$, kdy šikmost je rovna 0,80 a špičatost 2,76, je síla MC testu 0,450, JB testu 0,169 a RJB testu 0,112 nebo pro $n = 100$, $\mu_2 = 5$, $\sigma_2 = 5$, kdy šikmost je rovna 1,07 a špičatost 3,95, je síla MC testu 0,998, JB testu 0,985 a RJB testu 0,991. Hodnoty šikmosti a špičatosti analyzovaných kontaminovaných rozdělení jsou pak uvedeny v Tabulce 6.

Tabulka 5: Komparace síly testů – kontaminované normální rozdělení pro p = 0,50 a α = 0,05

σ_2	μ_2	n=25					n=100				
		SW	**LT**	**JB**	**RJB**	**MC**	**SW**	**LT**	**JB**	**RJB**	**MC**
1	**0**	0,052	0,050	0,052	0,053	0,050	0,048	0,051	0,050	0,050	0,049
	1	***0,050***	0,048	0,043	0,041	0,048	0,041	***0,047***	0,034	0,033	0,048
	2	0,041	0,048	0,010	0,008	***0,063***	***0,108***	0,095	0,006	0,002	0,080
	3	***0,136***	0,118	0,000	0,000	0,099	***0,735***	0,627	0,218	0,001	0,203
	4	***0,466***	0,384	0,000	0,000	0,169	***0,999***	0,994	0,911	0,278	0,301
	5	***0,852***	0,734	0,000	0,000	0,318	***1,000***	***1,000***	***1,000***	0,962	0,317
3	**0**	0,303	0,263	0,328	***0,416***	0,060	0,835	0,760	0,775	***0,909***	0,196
	1	0,338	0,308	0,337	***0,411***	0,084	0,890	0,848	0,803	***0,910***	0,322
	2	***0,470***	0,430	0,359	0,389	0,152	***0,973***	0,964	0,888	0,929	0,613
	3	***0,593***	0,535	0,340	0,321	0,250	***0,997***	0,995	0,950	0,941	0,863
	4	***0,695***	0,627	0,271	0,223	0,362	***1,000***	***1,000***	0,982	0,934	0,962
	5	***0,781***	0,693	0,169	0,112	0,450	***1,000***	***1,000***	0,995	0,900	0,990
5	**0**	0,556	0,608	0,453	***0,675***	0,115	***0,998***	***0,998***	0,928	***0,998***	0,736
	1	0,575	0,635	0,461	***0,664***	0,138	0,998	***0,999***	0,939	0,997	0,757
	2	0,636	***0,693***	0,478	0,644	0,211	***1,000***	***1,000***	0,950	0,997	0,826
	3	0,723	***0,771***	0,500	0,612	0,322	***1,000***	***1,000***	0,964	0,996	0,925
	4	0,806	***0,835***	0,513	0,556	0,459	***1,000***	***1,000***	0,977	0,993	0,985
	5	0,870	***0,881***	0,489	0,483	0,585	***1,000***	***1,000***	0,985	0,991	0,998

Pozn.1: Počet Monte Carlo simulací je 100000.
Pozn.2: Zvýrazněny jsou testy s nejvyšší sílou pro danou alternativu a rozsah souboru.

Tabulka 6: Charakteristiky šikmosti a špičatosti pro vybraná kontaminovaná rozdělení pro p = 0,10 a 0,50

p	0,1						0,5					
σ_2	1		3		5		1		3		5	
μ_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0	0,00	3,00	0,00	8,35	0,00	16,45	0,00	3,00	0,00	4,92	0,00	5,56
1	0,06	3,04	0,86	8,82	1,01	16,62	0,00	2,92	0,50	4,73	0,37	5,46
2	0,36	3,36	1,54	9,82	1,86	16,99	0,00	2,50	0,82	4,28	0,69	5,20
3	0,80	4,02	2,00	10,58	2,47	17,23	0,00	2,04	0,92	3,72	0,91	4,81
4	1,21	4,78	2,27	10,92	2,87	17,17	0,00	1,72	0,89	3,19	1,02	4,38
5	1,54	5,45	2,43	10,91	3,08	16,81	0,00	1,51	0,80	2,76	1,07	3,95

4. Závěr

Z provedených simulačních analýz je patrné, že síla sledovaných testů se liší dle jednotlivých alternativ. Proti těžkochvostým alternativám vykazuje nejvyšší sílu RJB test, nejnižší sílu pak vykazuje MC test. Pro lehkochvosté alternativy je nejsilnějším testem SW test, přičemž nejnižší sílu vykazuje opět MC test. Důvodem je velká robustnost MC testu, který vykazuje nízkou sílu rovněž v případě unimodálních symetrických i asymetrických alternativ a to pro nízkou váhu kontaminace ($p = 0{,}10$). Jedná se však o alternativy obsahující pouze několik málo odlehlých hodnot a v takovém případě lze robustnost MC testu považovat za výhodu, neboť hypotézu o normalitě nezamítá tak často jako ostatní sledované testy.

Váha kontaminace $p = 0{,}50$ již však generuje takové množství odlehlých hodnot, že robustnost MC testu je škodlivá. Pro symetrické unimodální alternativy je nejsilnějším testem RJB test, pro asymetrické alternativy pak SW a LT testy a v případě symetrických bimodálních alternativ pro malé i středně velké rozsahy souborů je nejsilnější SW test, zatímco JB a RJB testy pro malé rozsahy souborů svoji sílu zcela ztrácejí – důvodem ztráty síly je symetrie a mírně podnormální špičatost těchto bimodálních rozdělení, což u nich pro malé rozsahy souborů nevede k zamítnutí hypotézy o normalitě rozdělení.

5. Literatura

[1]BRYS, G., HUBERT, M., STRUYF, A. 2004. A Robustification of the Jarque-Bera Test of Normality. In: COMPSTAT 2004, Proceedings in Computational Statistics, edited by J. Antoch, Springer, Physica Verlag, 2004, s. 753 – 760.

[2]GEL, Y. R., GASTWIRTH, J. L. 2007. A robust modification of the Jarque-Bera test of normality. In: Economics Letters. 2007. doi:10.1016/j.econlet.2007.05.022.

[3]JARQUE, C. M., BERA, A. K. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. In: Economics Letters, č. 6, 1980, s. 255 – 259.

[4]LILLIEFORS, H. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. In: Journal of the American Statistical Association. Vol. 62. s. 399 – 402.

[5]SHAPIRO, S. S., WILK, M. B. 1965. An analysis of variance test for normality (complete samples). In: Biometrika. Vol. 52. č. 3 and 4. s. 591 – 611.

[6]THADEWALD, T., BÜNING, H. 2007. Jarque-Bera test and its competitors for testing normality – a power comparison. In: Journal of Applied Statistics, Taylor and Francis Journals, Vol. 34(1), s. 87 – 105.

[7]THODE, H. C. 2002. Testing for normality. New York: Marcel Dekker, 2002. 368 s. ISBN 978-0824796136.

[8]WÜRZT, D., KATZGRABER, H. G. 2005. Precise finite-sample quantiles of the Jarque-Bera adjusted Lagrange multiplier. In: ETH Zürich Preprint. Comp. Stat. submitted, (math.ST/0509423).

Adresa autora:

Luboš Střelec, Ing.
Zemědělská 1
691 25 Brno
xstrelec@node.mendelu.cz

Příspěvek vznikl za podpory programu AKTION Česká republika – Rakousko, spolupráce ve vědě a vzdělání, projektu č. 51p7 s názvem „Robust testing for the normality and its applications for the economy, sports and Basel II“.

Problém stability rozdelenia pri modelovaní celkovej výšky škôd
On the problem of stability in the modelling of total claim amount

Alena Tartaľová

Abstract: In this paper some properties of stable distribution and the relation between them are studied. The main reason why we are interested in stable distribution is, that for extreme insurance claims the classical central limit theorem condition (finite mean and variance) don't hold. We launched the generalization of central limit theorem and explain using it for calculating risk premium.

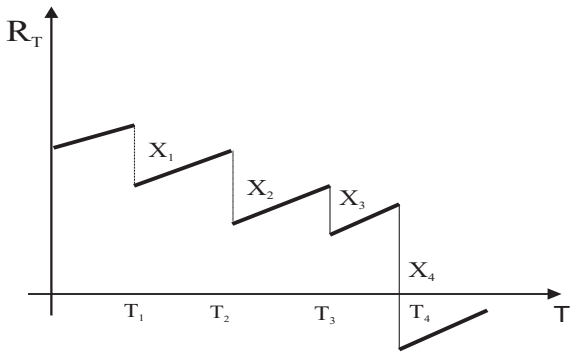
Key words: stable distribution, central limit theorem, extreme insurance losses, risk premium

Kľúčové slová: stabilné rozdelenie, centrálna limitná veta, extrémne poistné plnenia, riziková prirážka

1. Stochastický model rizika

Základným pojmom v poisťovníctve je pojem rizika, ktoré predstavuje možnosť vzniku poistnej udalosti. O procese rizika ako o náhodnom procese prvýkrát uvažoval švédsky matematik Filip Lundberg už v roku 1903. Ukázal, že Poissonov proces je vhodný na modelovanie poistných plnení. Túto myšlienku ďalej rozvíjal Harold Cramér. Základný Cramér-Lundbergov model je daný nasledujúcimi predpokladmi:

1. *Výšky poistných plnení* X_i, sú nezávislé, identicky rozdelené náhodné veličiny s distribučnou funkciou $F(x)$, strednou hodnotou $\mu = E(X_i) < \infty$ a disperziou $\sigma^2 = D(X_i) \leq \infty$.
2. *Okamžiky poistných plnení* $0 < T_1 < T_2 < \ldots$ sú náhodné veličiny.
3. *Počet poistných udalostí* v závislosti na čase t je náhodný proces $\{N_t, t \geq 0\}$, kde
$$N_t = \max\{n \geq 1; T_n \leq t\}.$$
4. Doby medzi susednými udalosťami $Y_1 = T_1$, $Y_k = T_k - T_{k-1}$, $k = 2,3,\ldots$ sú nezávislé, exponenciálne rozdelené náhodné veličiny so strednou hodnotou $E(Y_i) = 1/\lambda$.
5. Postupnosti $\{X_k\}_k$ a $\{Y_k\}_k$ sú nezávislé.
6. *Riziková rezerva poisťovne* v závislosti na čase je náhodný proces $\{R_t, t \geq 0\}$:
$$R_t = R_0 + ct - S_t. \qquad (1)$$



Obrázok 4: Proces rizika

Vo vzťahu (1) je R_0 – počiatočná rezerva poisťovne, c – intenzita toku poistného (v našom prípade je to lineárna funkcia, teda ide o konštantný prírastok c za jednotku času), $S_t = \sum_{i=1}^{Nt} X_i$ je výška celkového poistného plnenia za čas t (tzv. kolektívne riziko poisťovne), X_i je výška i-tého poistného plnenia, pre $i = 1,2,\ldots, N_t$, N_t je celkový počet poistných plnení za čas t.

Za daných predpokladov je $\{N_t, t \geq 0\}$ homogénny Poissonov proces s konštantnou intenzitou $\lambda > 0$, teda

$$P(N_t = k) = \frac{(\lambda t)^k}{k!} e^{-kt}, \; k = 0,1,2,\ldots$$

Jedným z najdôležitejších problémov poistnej matematiky je nájsť pravdepodobnostný model (rozdelenie) celkovej výšky škôd S_t.

Distribučnú funkciu celkového poistného plnenia $G(x)$ vyjadríme ako:

$$G(x) = P(S \leq x) = P\left(\bigcup_{n=0}^{\infty} (S \leq x \wedge N = n)\right) = \sum_{n=0}^{\infty} P(S \leq x \cap N = n) = \sum_{n=0}^{\infty} P(S \leq x / N = n).P(N = n).$$

Keďže máme fixný počet poistných plnení $\{X_i\}_{i=1}^{n}$, tak

$$P(S \leq x / N = n) = P(X_1 + X_2 + \cdots + X_n \leq n) = F^{*n}(x),$$

kde $F^{*n}(x)$ je n - násobná konvolúcia distribučných funkcií. Úpravou dostávame:

$$G(x) = \sum_{n=0}^{\infty} P(N = n).F^{*n}(x).$$

V prípade, že rozdelenie individuálnych poistných plnení je stabilné, tak zachováva typ rozdelenia.

2. Stabilné rozdelenia

Stabilné rozdelenia sú bohatou triedou distribučných funkcií, ktoré majú veľa zvláštnych matematických vlastností. Túto triedu prvýkrát charakterizoval Paul Lévy v roku 1920 v práci o súčte nezávislých, identicky rozdelených (skrátene iid – independent identically distributed) náhodných veličinách. Niekoľko objavov urobil aj taliansky sociológ a ekonóm Vilfredo Pareto. Približne v tom istom čase ako Paul Lévy, pracoval na podobnom probléme aj ruský matematik Alexander Chinčin.

Označme $S_0 = 0, S_n = X_1 + X_2 + \cdots + X_n, n \geq 1$. Pretože $S_n = \sum_{i=1}^{n} X_i$ je súčtom iid náhodných veličín, podľa centrálnej limitnej vety (skrátene CLV) sa ponúka aproximácia normálnym rozdelením. Podľa klasickej CLV, normovaný súčet iid náhodných veličín , ktoré majú konečnú strednú hodnotu a disperziu, konvergujú v distribúcii k normálnemu rozdeleniu, teda:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \approx N(0;1), \text{ pre } n \to \infty. \qquad (2)$$

Tento vzťah možno prepísať na tvar:

$$a_n S_n - b_n \xrightarrow{d} Z \approx N(0;1), \text{ pre } n \to \infty. \tag{3}$$

Kde $a_n = \frac{1}{\sigma\sqrt{n}}, b_n = \frac{\sqrt{n}\mu}{\sigma}$.

Ak stredná hodnota alebo disperzia nie sú konečné alebo neexistujú, tak „klasické normovanie“ nie je možné a otázkou je, či je možné zovšeobecniť princíp normovania. Teda sa pýtame:

Aké limitné rozdelenia prichádzajú do úvahy pre vhodne normované súčty nezávislých identicky rozdelených náhodných veličín?

Teda, pre ktoré rozdelenia platí:

$$c_1 X_1 + c_2 X_2 = bX + a. \tag{4}$$

Inak povedané pýtame sa, ktoré distribučné funkcie sú uzavreté vzhľadom na konvolúciu a násobenie reálnymi číslami. Možnými limitnými rozdeleniami pre súčty iid náhodných veličín sú iba distribučné funkcie, ktoré spĺňajú vzťah (4).

Dôležitou vlastnosťou Gaussovho (normálneho) rozdelenia je, že súčet iid náhodných veličín s normálnym rozdelením má normálne rozdelenie, t.j.: ak $X_1 \approx N(\mu_1;\sigma_1^2)$ a $X_2 \approx N(\mu_2;\sigma_2^2)$, potom $X_1 + X_2 \approx N(\mu_1+\mu_2;\sigma_1^2+\sigma_2^2)$. V práci (Tartaľová, 2006) je ukázané, že túto vlastnosť spĺňajú aj Poissono a Gama rozdelenie.

Definícia. Hovoríme, že náhodná veličina má stabilné rozdelenie, ak pre všetky n>1 existujú konštanty $b_n > 0$ a $a_n \in R$ také, že:

$$X_1 + X_2 + \cdots + X_n \overset{d}{=} b_n X + a_n, \tag{5}$$

kde $X_1, X_2, \ldots, X_n$ sú iid náhodné veličiny s rovnakým rozdelením ako *X*. Konštanty a_n, b_n nazývame normovacie konštanty.

Poznámka. Symbol „$\overset{d}{=}$“ znamená rovnosť v distribúcii, t.j. náhodná veličina na ľavej aj na pravej strane majú rovnaké pravdepodobnostné rozdelenie.

Ak súčet náhodných veličín $X_1, X_2, \ldots, X_n$ označíme ako S_n, potom dostávame:

$$S_n \overset{d}{=} b_n X + a_n, \text{ pre } n \geq 1.$$

Tento vzťah môžeme prepísať na tvar:

$$b_n^{-1}(S_n - a_n) \overset{d}{=} X, \text{ pre } n \geq 1.$$

Pre stabilné rozdelenia sa používa názov „stabilné“ preto, lebo typ funkcie je stabilný, nemení sa vplyvom operácie sčítania.

Niekedy je však jednoduchšie zadať stabilné rozdelenie pomocou charakteristickej funkcie (podľa Embechts et. al, 1997):

Veta. (Lévyho reprezentácia stabilných rozdelení)

Rozdelenie dané charakteristickou funkciou $\varphi(t)$, je stabilné práve vtedy, ak platí:

$$\varphi(t) = \exp\left\{i\delta t - \gamma|t|^{\alpha}\left[1 - \beta(\operatorname{sgn}(t))z(t,\alpha)\right]\right\}, \tag{6}$$

kde $\alpha \in (0;2]$, $\beta \in [-1;1]$, $\gamma \geq 0$, $\delta \in R$ a

$$z(t,\alpha)=\begin{cases} tg\left(\frac{\pi\alpha}{2}\right); & \alpha\neq 1 \\ -\frac{2}{\pi}\ln|t|; & \alpha=1. \end{cases}$$

Funkcia *sgn(t)* vo vzťahu (6) je definovaná nasledovne:

$$\operatorname{sgn}(t)=\begin{cases} -1, & t<0 \\ 0, & t=0 \\ 1, & t>0. \end{cases}$$

Poznámka. Parameter α v charakteristickej funkcii stabilného rozdelenia (6) sa nazýva charakteristický exponent alebo index stability. Odpovedajúce rozdelenie nazývame α-stabilné. Napr. pre $\alpha=2$ dostávame normálne rozdelenie.
Dá sa ukázať, že aj Lévyho, Cauchyho a Paretovo rozdelenie sú stabilné (pozri Tartaľová, 2006).

Individuálne poistné plnenia pri viacerých typoch najmä neživotného poistenia majú niektoré spoločné vlastnosti. Väčšina z nich má hodnotu nižšiu ako priemerné poistné plnenie, pritom sú dosť pravdepodobné extrémne poistné plnenia. To spôsobuje veľký rozptyl, a teda pri aproximácii normálnym rozdelením (pri CLV) sa môže stať, že rozptyl nie je konečný, t.j. $D(X_i)=\infty$. Vážnym nedostatkom normálnej aproximácie je aj to, že hustota normálneho rozdelenia je symetrická a jej pravý koniec veľmi rýchlo konverguje k nule. Pri extrémnych poistných plneniach má rozdelenie tzv. tučné chvosty s pomalým klesaním k osi x. Pre takéto hodnoty má normálna aproximácia tendenciu podhodnocovať hodnoty na pravom chvoste rozdelenia. Z pohľadu poisťovateľa je to veľmi nežiaduca skutočnosť, pretože vysoké poistné plnenia majú pre neho závažný finančný dopad. Preto takýto prípad uvedieme zovšeobecnenie CLV (podľa Embrechts et.al, 1997) .

Veta. (Zovšeobecnenie CLV)

(a) Ak $E(X_i^{\,2})<\infty$, potom:

$$P\left[\frac{S_n-n\mu}{\sigma n^{\frac{1}{2}}}\le x\right]\to\Phi(x), \text{ pre } n\to\infty,$$

kde $\Phi(x)$ je distribučná funkcia štandardného normálneho rozdelenia a $\mu=E(X_i)$, $\sigma^2=D(X_i)$.

(b) Ak $E(X_i^{\,2})=\infty$, potom:

$$P\left[\frac{S_n-\tilde{a}n}{n^{\frac{1}{\alpha}}L(n)}\le x\right]\to G_\alpha(x), \text{ pre } n\to\infty,$$

kde $G_\alpha(x)$ je α-stabilná distribučná funkcia, *L(n)* je vhodná pomaly sa meniaca funkcia a $\tilde{a}$ je definované ako:

$$\tilde{a}=\begin{cases} \mu, & \alpha\in(1;2> \\ 0, & \alpha\le 1 \end{cases}.$$

Poznámka. Typickými príkladmi pomaly sa meniacich funkcií sú napríklad: kladné konštanty, funkcie konvergujúce kladnej konštante, logaritmus.

3. Určenie rizikovej prirážky k čistému poistnému

Aproximácia kolektívneho modelu rizika S v praxi často využívame na určenie tzv. rizikovej prirážky θ k čistému poistnému, ktorým rozumieme strednú hodnotu $E(S)$ kolektívneho rizika. Je to konštanta, ktorá zaručí, že príjem z poistného sa bude rovnať hodnote zvoleného kvantilu rozdelenia celkového poistného plnenia S (napr. $S_{0,95}$). Ak priemerné poistné plnenie $E(S)$ je tzv. čisté (netto) poistné, potom požadujeme, aby sa celkové brutto poistné $(1+\theta)E(S)$ rovnalo γ% - ného kvantilu rozdelenia S, t.j.

$P(S \le S_\gamma) = P(S \le (1+\theta)E(S)) = \gamma$, t.j.

$$P\left(\frac{S - E(S)}{\sqrt{D(S)}} \le \frac{\theta E(S)}{\sqrt{D(S)}}\right) = \gamma .$$

Podľa klasickej CLV platí:

$\frac{S - E(S)}{\sqrt{D(S)}}$ má asymptoticky štandardné normálne rozdelenie $N(0;1)$,

tak dostávame:

$$\frac{\theta E(S)}{\sqrt{D(S)}} = z_\gamma$$

a teda ak označíme $\sigma(S) = \sqrt{D(S)}$ úpravou dostávame:

$$\theta = \frac{z_\gamma \sigma(S)}{E(S)} .$$

V prípade, že rozptyl nie je konečný podľa zovšeobecnenej CLV, normovaný súčet iid náhodných veličín má stabilné rozdelenie, t.j.

$$P\left(\frac{S - E(S)}{n^{-\frac{1}{\alpha}} L(n)} \le \frac{\theta E(S)}{n^{-\frac{1}{\alpha}} L(n)}\right) = \gamma .$$

Podobnou úpravou ako v predchádzajúcom prípade, dostávame pre rizikovú prirážku:

$$\theta = \frac{x_\gamma n^{-\frac{1}{\alpha}} L(n)}{E(S)} .$$

Rozdelenie kolektívneho modelu rizika vieme aproximovať aj pomocou posunutého gamma rozdelenia(Pacáková, 2004), čím čiastočne znížime podhodnotenie pravdepodobnosti vysokých poistných plnení v porovnaní s normálnou aproximáciou. Ale pri tejto aproximácii potrebujeme odhadnúť prvé tri začiatočné momenty rozdelenia S, čo v prípade extrémnych poistných plnení môže byť nemožné.

Tento príspevok je podporovaný grantom VEGA č. 1/0724/08, „Riadenie rizík neživotného poistenia podľa direktívy Európskej komisie SOLVENCY II“.

4. Literatúra

[1]BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J. and TEUGELS, J., 2004. Statistics of Extremes: Theory and Applications, Wiley, New York, 2004. ISBN 0-471-97647-4

[2]EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T., 1997. Modelling extremal events for insurance and finance, Springer, Berlin, 1997. ISBN 3-540-60931-8

[3]MIKOSCH, T. 2003. Non-Life insurance mathemathics (A Primer), Laboratory of Actuarial Mathematic, University of Copenhagen.(dostupné na www.math.ku.dk)

[4]PACÁKOVÁ, V.2004. Aplikovaná poistná štatistika, Edícia Ekonómia, 2004, 261 s. ISBN 80-8087-004-8

[5]SKŘIVÁNKOVÁ, V- TARTAĽOVÁ, A.,2008.Catastrofic risk management in non-life insurance. E+M Ekonomie a Management

[6]TARTAĽOVÁ, A.2006. Modelovanie extrémnych poistných udalostí, 2006. Diplomová práca na UPJŠ

Adresa autora:

Alena Tartaľová, Mgr.
Technická univerzita v Košiciach
Ekonomická fakulta
Katedra aplikovanej matematiky a hospodárskej informatiky
Nemcovej 32
040 01 Košice
alena.tartalova@tuke.sk

On-line splajny a aproximačné modely
On-line splines and approximation models

Csaba Török

Abstract: The paper proposes for approximation of complex dependences a two-part approximation model, spline, components of which are computed successively. In derivation of the central part of the model the key role is played by a special representation of polynomials.

Key words: splines, polynomial models, LS, piecewise approximation, reference points, reparametrization

Kľúčové slová: splajny, polynomiálne modely, MNŠ, úseková aproximácia, referenčné body, preparametrizácia

1. Úvod

V práci na aproximáciu komplexných závislostí predkladáme aproximačný model z dvoch častí, splajn, komponenty ktorého sa vypočítavajú postupne. Základnú úlohu pri odvodení centrálnej časti modelu zohráva špeciálna reprezentácia polynómov.

Klasická teória splajnov je o off-line alebo globálnych splajnoch. Musíme mať všetky informácie, body, aby sme mohli na základe k-diagonálnej sústavy zostrojiť splajn k-ho stupňa, teda úsekové polynómy k-ho stupňa, kde v spoločných uzloch sa rovnajú nielen susedné polynómy, ale aj ich k-1derivácie.

Pod on-line splajnami budeme rozumieť lokálne splajny, kde na základe postupujúcich informácií zostrojíme najprv prvú komponentu, potom druhú a postupne aj ďalšie komponenty splajnu. Pričom väčšinou postupujeme zľava doprava: po zostrojení jedného polynómu a po získaní ďalších informácií, bodov, zostrojíme nasledujúci polynóm k-ho stupňa dodržujúc základnú požiadavku, aby aj prvé k-1 derivácie práve počítaného nového a už hotového posledného polynómu, ľavého suseda, sa rovnali v spoločnom uzle. Takto postupujúc najprv získame lokálny splajn z dvoch komponentov, potom z troch a postupne aj z ďalších komponentov.

Spoločné v off-line a on-line splajnoch je to, že v uzloch susedné polynómy a ich derivácie sa rovnajú. Líšia sa v tom, že komponenty off-line splajnu sa vypočítajú naraz pomocou jednej sústavy, a komponenty on-line splajnu postupne na základe susedných komponentov a následných meraní. V dôsledku toho off-line splajny sú optimálne globálne a off-line splajny lokálne. Ďalej, kým off-line splajny štandardne predpokladajú presné hodnoty druhých koordinát bodov (až napr. na neparametrické smoothing splajny [2]), on-line splajny môžeme používať aj pri meraní zaťaženými chybami.

V druhej časti práce sformulujeme reprezentačnú vetu. V tretej predstavíme aproximačný model z dvoch častí, ktorý ilustrujeme príkladom v štvrtej časti.

2. Reprezentácia polynómov

Nedávno sme odvodili pomocou r-bodovej transformácie všeobecnú reprezentáciu polynómov pomocou neúplných interpolačných polynómov [6]

Veta. Nech $p \geq r \geq 2$. Potom ľubovoľný polynóm $P_p(x) = a_0 + a_1 x + ... + a_p x^p$ môže byť vyjadrený pomocou jeho *r* rozličných bodov $R = \{[x_i, y_i],\ y_i = P_p(x_i),\ i = \overline{0, r-1}\}$ ako

$$P_p(x) = I_{r-1}(x) + Z_r(x)\, A_{p-r,r}(x), \tag{1}$$

kde

$$I_{r-1}(x) = \sum_{i=0}^{r-1} \Pi_i(x)\, y_i$$

je neúplný interpolačný polynóm,

$$\Pi_i(x) = \prod_{v \in V_i} \frac{x - v}{x_i - v},\quad V_i = V \setminus \{v_i\},\quad V = \{x_0,\ x_1,\ ...\ , x_{r-1}\},\quad i = \overline{0, r-1},$$

$$Z_r(x) = \prod_{i=0}^{r-1} (x - x_i),\quad A_{p-r,\,r}(x) = S^T \cdot \alpha,$$

$$\alpha = (a_r,\ a_{r+1}, ..., a_p)^T,\quad S = (S_{0,r},\ S_{1,r}, ..., S_{p-r,r})^T,\quad S_{1,r} = 1,\ r \geq 0,\quad S_{j,0} = x_0^j,\ j \geq 0,$$

$$S_{j,r} = S_{j,r-1} + x_r\, S_{j-1,r},\quad j \geq 1,\ r \geq 1.$$

Napríklad, kubický polynóm $P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ pomocou troch referenčných bodov $R = \{[v_0, P_0], [v_1, P_1], [v_2, P_2]\}$ ležiacich na *P*(*x*) môžeme vyjadriť v súlade s reprezentačnou vetou nasledovne

$$P(x) = \frac{(x - v_1)(x - v_2)}{(v_0 - v_1)(v_0 - v_2)} P_0 + \frac{(x - v_0)(x - v_2)}{(v_1 - v_0)(v_1 - v_2)} P_1 + \frac{(x - v_0)(x - v_1)}{(v_2 - v_0)(v_2 - v_1)} P_1 + (x - v_0)(x - v_1)(x - v_2) a_3.$$

Vzťah (1) predstavuje prechod od meraní, bodov polynómu k čistej funkčnej závislosti a preparametrizuje pôvodný polynóm: namiesto prvých *r* koeficientov $a_0, a_1, ..., a_{r-1}$ zavedie cez referenčné body parametre $x_0, x_1, ..., x_{r-1}$ a $y_0, y_1, ..., y_{r-1}$. Problém pri jeho aplikácii nastáva v prípade, keď $y_0, y_1, ..., y_{r-1}$ sú zaťažené chybami. Na riešenie toho sme zaviedli model z dvoch častí.

Reprezentačnú vetu sme úspešne aplikovali pri rôznych aproximačných modeloch [1,4]. Vďaka C^2 kvázi-spojitosti, modely z dvoch častí [3,5] majú také relevantné asymptotické vlastnosti ako konzistentnosť alebo asymptotická normalita. Na rozdiel od týchto prác, v nasledujúcej časti ukážeme, ako je možné odvodiť explicitný vzťah, ktorý zabezpečuje nielen C^2 kvázi-spojitosť susedných dvoch polynómov v spoločnom bode, ale aj C^2 spojitosť.

3. Model z dvoch častí

Uvažujme nad intervalmi $[u_0, u_1]$, $[u_1, u_2]$ polynómy

$$P(x) = a_0 + a_1 x + ... + a_p x^p,$$

$$Q(x) = b_0 + b_1 x + ... + b_q x^q$$

medzi ktorými predpokladáme C^2 spojitosť

$$P(u_1) = Q(u_1),$$

$$P'(u_1) = Q'(u_1),$$

$$P''(u_1) = Q''(u_1),$$

kvôli hladkému prechodu medzi $P(x), Q(x)$ v spoločnom uzle u_1 a zašumené dáta

$$\{[x_{i,M}, \tilde{y}^*_{i,M}],\ \tilde{y}^*_{i,M} = \tilde{P}(x_{i,M}) = P(x_{i,M}) + \varepsilon^*_{i,M},\quad i = \overline{1,M}\},$$
$$\{[x_{i,N}, \tilde{y}_{i,N}],\ \tilde{y}_{i,N} = \tilde{Q}(x_{i,N}) = Q(x_{i,N}) + \varepsilon_{i,N},\quad i = \overline{1,N}\}.$$

Nech $v_0 = u_1$, $v_1 = u_1 - \tau$, $v_2 = u_1 - 2\tau$. Je možné ukázať, že ak na základe reprezentačnej vety vyjadríme Q(x) ako (pre jednoduchosť naďalej nech je p=q=3)

$$Q(x) = \frac{(x-v_1)(x-v_2)}{(v_0-v_1)(v_0-v_2)}P(v_0) + \frac{(x-v_0)(x-v_2)}{(v_1-v_0)(v_1-v_2)}P(v_1) + \frac{(x-v_0)(x-v_1)}{(v_2-v_0)(v_2-v_1)}P(v_2) + (x-v_0)(x-v_1)(x-v_2)b_3 .$$

Potom pri $\tau \to 0$ dostaneme

$$Q(x) = P(x) + (x-u_1)^3(b_3 - a_3). \tag{2}$$

Táto voľba $Q(x)$ skutočne zabezpečí splnenie podmienok rovností polynómov $P(x)$, $Q(x)$ a ich prvých dvoch derivácií v u_1. Preto budeme uvažovať aproximačný model z dvoch častí

$$\begin{cases} \tilde{y}^*_{i,M} = P(x_{i,M}) + \varepsilon^*_{i,M}, \\ \tilde{y}_{i,N} = P(x_{i,N}) + (x_{i,N} - u_1)^3(b_3 - a_3) + \varepsilon_{i,M}, \end{cases} \tag{3}$$

v ktorom neznámy vektor parametrov $a = (a_0, a_1, a_2, a_3)^T$ prvej časti odhadujeme pomocou klasickej MNŠ a jediný neznámy parameter b_3 druhej časti, ktorá zodpovedá (2), na základe vzťahu

$$\tilde{y}_{i,N} - \hat{P}(x_{i,N}) + (x_{i,N} - u_1)^3 \hat{a}_3 = (x_{i,N} - u_1)^3 b_3,\quad i = \overline{1,N}, \tag{4}$$

v ľavej časti ktorého vystupujú už vypočítané odhady $\hat{a} = (\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3)^T$, tiež pomocou MNŠ s bázovou funkciou $(x-u_1)^3$.

4. Príklad

Uvažujme nad intervalmi [0; 1], [1; 1.9] polynómy (u_1=1)

$P(x) = 0.03 + 0.432x - 1.2x^2 + 0.8x^3$,

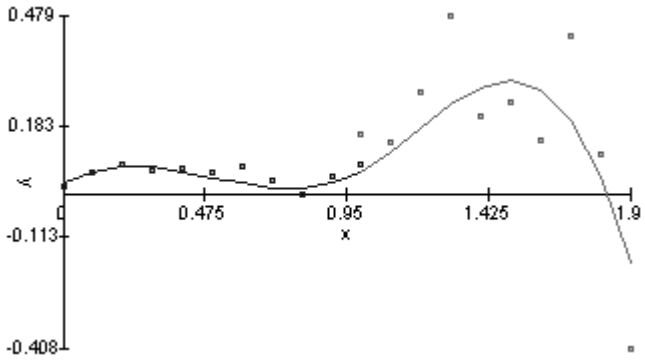
$Q(x) = 0.03 + 0.432x - 1.2x^2 + 0.8x^3 - 3(x - 1)^3 = 3.03 - 8.568x + 7.80x^2 - 2.2x^3$

a ich pomocou generované dáta (SNR=1):

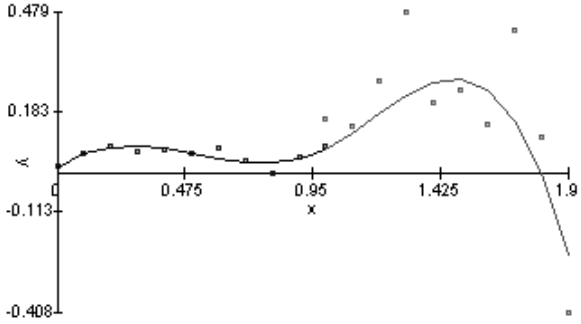
x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\tilde{P}(x)$	0.02	0.057	0.079	0.066	0.071	0.058	0.078	0.037	0.001	0.046	0.079

x	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
$\tilde{Q}(x)$	0.16	0.141	0.273	0.479	0.207	0.246	0.144	0.424	0.108	-0.41

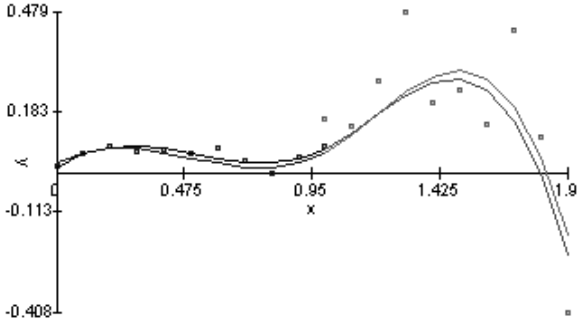
Obrázok 1 znázorňuje $P(x)$, $Q(x)$ a generované dáta $[x_i, \tilde{P}(x_i)]$, $i = \overline{0,11}$ a $[x_i, \tilde{Q}(x_i), i = \overline{0,10}]$.



Obrázok 5:Polynómy a generované dátové body



Obrázok 2:Odhady polynómov



Obrázok 3: Polynómy a ich odhady

V prvom kroku na základe prvej sady dát a polynomiálneho regresného modelu z prvej časti (3) obdržíme

$$\hat{P}(x) = 0.01776 + 0.51123x - 1.25114x^2 + 0.79389x^3,$$

a v druhom kroku na báze druhej sady dát a (4) získame odhad $\hat{b}_3 = -2.16342$, teda

$$\hat{Q}(x) = 0.01776 + 0.51123x - 1.25114x^2 + 0.79389(x^3 - (x - 1)^3) - 2.16342(x - 1)^3 =$$

$$= 2.97507 - 8.36070x + 7.62079x^2 - 2.16342x^3.$$

Obrázok 2 znázorňuje body a odhady polynómov $\hat{P}(x)$, $\hat{Q}(x)$ a obrázok 3 body, polynómy a ich odhady. Vidíme, že v $x = u_1 = 1$ prechod medzi $\hat{P}(x)$, $\hat{Q}(x)$ je hladký.

Prehod od základného modelu z dvoch častí k všeobecnému úsekovému aproximačnému modelu s uzlami $u_0, u_1, \ldots, u_k$ vďaka (2) s $u_1, u_2, \ldots, u_{k-1}$ je prirodzený.

5. Záver

Na rozdiel od klasických off-line splajnov navrhovaný aproximačný model vypočíta komponenty splajnu postupne. Model z dvoch častí sme demonštrovali na príklade. Jeho zovšeobecnenie na parametrickú úsekovú aproximáciu a vyšetrenie štatistických vlastností modelu je cieľom blízkej budúcnosti.

6. Literatúra

[1]DIKOUSSAR N.D., TÖRÖK CS., 2006, Automatic knot finding fo piecewise-cubic aproximation, Mat.model., T-17, N.3

[2]Eubank R.L., Nonparametric regression and spline smoothing, Marcel Dekker, Inc., 1999

[3]MATEJČÍKOVÁ A., TÖRÖK CS., 2007, Rekurzívny odhad parametra v kvadratickom modeli,VIII konferencia SvF TU Košice

[4]RÉVAYOVÁ M., TÖRÖK CS.: Piecewise approximation and neural networks, Kybernetika, Kybernetika – Volume 43 (2007), N4, pp. 547-559

[5]RÉVAYOVÁ M., TÖRÖK CS.,2007, Reference points based approximation, VIII konferencia SvF TU Košice

[6]Török Cs., The Fusion of Interpolation and Approximation, MD CEF TU Košice, to appear

Adresa autora:

Csaba Török, doc. RNDr. CSc.
Ústav informatiky, PF UPJŠ
Jesenná 5, 04001 Kosice
Csaba.torok@upjs.sk

***Poďakovanie**: Výskum bol podporený z grantového projektu VEGA 1/4003/07.*

Matematické modelovanie riadenia technologických procesov s využitím metód viacrozmernej štatistiky
Mathematical Modeling of Technological Processes Utilizing Methods of Multidimensional Statistics

Alena Vagaská

Abstract: Technological processes are complex dynamic processes in which parameters determining a process running occur. The fact that a considerable number of parameters enter the process makes mathematical models designing by theoretical-analytical methods difficult. Based on the mentioned, within the meaning of a quality statistical control, the article will deal with the design of mathematical models of selected technological processes using methods of multidimensional statistics and program systems.

Key words: control**,** technological processes, mathematical modeling, design of experiments – DoE, statistical optimalization, process factors.

Kľúčové slová: riadenie, regulácia, technologické procesy, matematické modelovanie, plánované experimenty (DoE), štatistická optimalizácia, procesné faktory

1. Úvod

Zavádzanie matematických metód modelovania do technologických procesov vo výrobnej sfére má na zreteli jasný cieľ – pomocou modelov a simulácie riadiť a udržiavať technologické procesy na optimálnej úrovni s požadovanou kvalitou výstupu. Dosiahnuť požadovanú kvalitu výstupu, napr. požadované kvalitatívne charakteristiky obrobenej plochy, nám umožňuje ako využitie moderných, vysoko presných obrábacích strojov s CNC riadením, tak aj poznanie funkčných závislostí medzi parametrami kvality výrobku a procesnými faktormi výrobného systému. Z tohto uhla pohľadu sa v článku zameriame na riadenie a reguláciu (NC strojov) a na matematické modelovanie technologických procesov využitím metód viacrozmernej štatistiky. Prínosom je stanovenie pôvodných funkčných závislostí výstupných parametrov kvality na parametroch vstupujúcich do technologického procesu a vytvorenie pôvodných matematických modelov na základe spracovania štatistických súborov získaných experimentálnych meraním.

2. Numerické riadenie a číslicová riadiaca technika (CNC technika)

NC-riadenie (Numerical control) riadi pohyby strojov a zariadení na základe numerických údajov. Hlavnou oblasťou využitia je výroba geometricky definovaných výrobkov, t.j. výrobkov popísaných technickými výkresmi. Číslicové riadenie spočíva v číslicovom nastavovaní a regulácii pohyblivých častí a to suportu stroja alebo tiež jednotlivých osí NC-stroja.

Pri obrábaní na NC-strojoch nás zaujíma predovšetkým riadenie a regulácia polohy a otáčok. U obrábacieho stroja môže byť poloha suportu alebo pohyblivého stola snímaná pomocou analógového snímača a regulovaná pomocou operačného zosilovača. Pomocou pohonov sú suporty nastavované do zadaných pozícii tak, že sú v rýchlom slede zadávané požadované polohy a regulačný systém nastavuje suporty zo skutočných do požadovaných polôh. Regulácia polohy sa realizuje prevažne ako kaskádová regulácia. Pri kaskádovej regulácii sa pomocou regulátora polohy získava požadovaná hodnota rýchlosti $v_{pož}$. Ak vznikne regulačná diferencia polohy $\Delta x = x_{pož} - x_{sk}$ (spôsobená napr. mechanickým odporom

obrábania), je následne zosilená a ovplyvní otáčky motora v zmysle zmenšenia regulačnej diferencie Δx. Rýchlosť posuvu v je úmerná polohovej diferencii Δx. U číslicovo riadených strojov vykonáva regulátor polohy výpočet regulačnej diferencie Δx vždy číslicovo. U obrábacích strojov by mali skutočné otáčky n_{sk} hlavného pohonu i pohonu vretien (posuvov) čo najpresnejšie sledovať požadované hodnoty $n_{pož}$, ktoré sú zadávané numerickým riadiacim (NC) systémom.

Vzhľadom na zrealizovaný experiment v rámci optimalizácie technologických procesov, ktorý je v článku prezentovaný, je vhodné spomenúť v súvislosti s obrábaním na NC-strojoch ešte ďalšie skutočnosti. Z ekonomických dôvodov je často požadované kompletné opracovanie výrobku pri jednom upnutí. Z toho dôvodu sú NC-stroje často vybavené poháňanými nástrojovými hlavami s pevne upnutými nástrojmi nastavovanými do pracovnej pozície natočením hlavy. S odpovedajúcimi nástrojmi môžu byť potom vyvŕtané alebo vystruhované otvory, rezané závity alebo frézované drážky. Súčasťou systému automatickej výmeny nástrojov u číslicovo obrábaných strojov sú zásobníky nástrojov. Ich výhodou oproti revolverovej hlave je väčší počet nástrojov, nevýhodou je naopak dlhšia doba výmeny nástroja. Takúto výmenu nástrojov prostredníctvom zásobníka nástrojov pri jednom upnutí obrobku sme využili v rámci nášho experimentu.

Pred samotnou realizáciou obrábacieho procesu na NC strojoch je potrebné dopredu nastaviť nástroj a materiál (obrobok) do určitej vzájomnej polohy. Dnes sa už používa riadenie NC-osi. K najdôležitejším funkciám NC-riadenia osí patrí napr. pracovný posuv (suportu), rýchloposuv, absolútne krokové vystavenie do cieľovej polohy a vystavenie do referenčného bodu začiatku určitého programu. Pri riadení pohybu a pracovnej dráhy nástroja je potrebné vedieť popísať pracovnú dráhu (trajektóriu) referenčného bodu nástroja (TCP), poznať okamžitú polohu (súradnice TCP) a charakter dráhy (krivku) medzi týmito bodmi. Potom je nutné prepočítavať súradnice medziľahlých bodov, t.j. vykonať interpoláciu. Podľa charakteru medziľahlej dráhy (krivky) rozlišujeme napr. lineárnu, kruhovú, či parabolickú interpoláciu a interpoláciu splinom pri dráhe skladanej z rôznych kriviek. Riadiaci systém stačí počítať súradnice medziľahlých bodov v časových intervaloch niekoľko milisekúnd, čo odpovedá bežným požiadavkám a možnostiam pohonov. Dráha referenčného bodu nástroja TCP je odvodená z plánovaného tvaru obrábaného predmetu. Pri vytváraní programu opracovania obrobku na CNC stroji, t.j. pri programovaní dielčích operácií a plánovaného pohybu obrábacieho nástroja musíme mať vždy k dispozícii údaje vzťažné na nejaký súradnicový systém. Komunikácia medzi riadiacim systémom číslicovo riadeného stroja a obsluhou sa uskutočňuje prostredníctvom obslužného panelu, ktorý umožňuje vkladanie príkazov riadiacich programov aj špeciálne parametre stroja.

V rámci štatistickej optimalizácie technologických procesov sme pri realizácii experimentu potrebovali použiť vhodný obrábací stroj. Pri výrobe sa obrábací stroj vyberá na základe technologických požiadaviek a ekonomickej efektívnosti. Čím menšia je výrobná dávka, tým dôležitejšia je schopnosť výrobného zariadenia adaptabilne sa prispôsobovať zmenám. Pri našich experimentálnych skúškach, zameraných na vyjadrenie vplyvu rezných podmienok (pri frézovaní) na morfológiu obrobeného povrchu bolo vhodné použiť univerzálny obrábací stroj. Na Fakulte výrobných technológií TU v Košiciach sa v Laboratóriu výrobných technológií nachádza CNC vertikálne obrábacie centrum PK-VMC 650 S, ktoré bolo použité v rámci nášho experimentu orientovaného na optimalizáciu rezných podmienok pri čelnom frézovaní z hľadiska morfológie obrobeného povrchu. Toto obrábacie centrum je určené na presné a rýchle obrábanie obecných tvarových povrchov, vŕtanie, vyvrtávanie, vystruhovanie, rezanie závitov a frézovanie väčších a tvarovo zložitých dielcov. Automatická výmena nástrojov umožňuje prácu v cykloch.

V experimente sa z hľadiska kvality obrobeného povrchu skúmal vplyv štyroch faktorov vstupujúcich do procesu frézovania: priemer frézy $d\,[\mathrm{mm}]$, otáčky $n[\mathrm{min}^{-1}]$, posuv na zub $f_z\,[\mathrm{mm}]$, hĺbka rezu $a_p\,[\mathrm{mm}]$. Spomínané faktory nadobúdali dve úrovne (hladiny) pri obrábaní. Preto prišlo vhod, že obrábacie centrum PK-VMC 650S umožňuje nielen automatickú výmenu nástrojov (výmena nástroja kvôli zmene priemeru čelnej frézy), ale aj možnosť zmeny otáčok, hĺbky rezu a posuvu na zub pri jednom upnutí obrobku. Aktuálne rezné podmienky experimentu sú uvedené v tab. 1. Tieto podmienky ako aj teoretické poznatky zhrnuté v predchádzajúcom texte boli potom zohľadnené pri naprogramovaní CNC stroja pre realizáciu čelného frézovania drážok do vopred zvoleného materiálu. Výber rezných podmienok bol podriadený druhu použitého polovýrobku (rozmery a trieda materiálu), materiálu rezného nástroja, možnostiam obrábacieho stroja a v neposlednom rade aj čo najširšiemu záberu rezných parametrov pre dosiahnutie komplexných výsledkov.

3. Štatistická optimalizácia a plánované experimenty

Zavádzanie matematických modelov do praxe je v súlade s cieľmi komplexného manažérstva kvality, kde štatistické metódy v matematickom modelovaní sú hlavným nástrojom zlepšovania kvality. Uvedieme si príklad efektívneho využitia štatistických metód, konkrétne plánovaných experimentov (DoE) pri riadení technologických procesov (frézovanie) za účelom štatistickej optimalizácie drsnosti povrchu, t.j. optimalizácie kvality obrobeného povrchu.

Obrobený povrch sa vytvára kopírovaním tvaru reznej hrany na obrobenú plochu pri jej relatívnom pohybe voči obrobku. Výsledný efekt preto závisí na geometrickom tvare reznej hrany a kinematike jej pohybu oproti obrobku. Výrazný vplyv majú mechanické vlastnosti obrábaného materiálu a tribologické vzťahy medzi nástrojom a obrobkom. Poznanie týchto faktorov a faktorov pôsobiacich v danej technológii obrábania a ich optimalizácia vedie k splneniu stúpajúcich požiadaviek na kvalitu obrobenej plochy.

Optimalizácia v najširšom slova zmysle predstavuje získanie najlepších výsledkov za určitých podmienok, t.j. na daný variant vstupov získame optimálny výstup. Matematická formulácia optimalizácie sa vo všeobecnosti zakladá na hľadaní maxima (a pri zmene znamienka tiež minima) funkcie:

$$q = f(\,x_0, x_1, x_2, \ldots, x_N, y_0, y_1, y_2, \ldots, y_N\,) \tag{1}$$

kde $x_0, x_1, \ldots, x_N, y_1, y_2, \ldots, y_N$ je konečný počet reálnych premenných. Funkcia q sa nazýva kritériom optimalizácie (kriteriálna, účelová funkcia, optimalizačný funkcionál) a charakterizuje kvantitatívne získané výsledky [2]. Riadené, tzv. predikčné resp. kontrolované premenné $x_0, x_1, \ldots, x_N$ charakterizujú podmienky, na ktoré je možné vplývať ľubovoľným spôsobom s cieľom získania najlepších podmienok podľa vybraného kritéria. Neriadené premenné $y_1, y_2, \ldots, y_N$ charakterizujú podmienky, na ktoré nie je možné vplývať a sú určované nezávislými činiteľmi. Väzbu oboch druhov premenných s kritériom optimalizácie q určuje predpis funkcie (1). Teória regresnej analýzy je založená na výpočte extrémov funkcií a v základnej úlohe predpokladá, že veličina y (meraná veličina) stochastickým spôsobom závisí na premenných x_j, $j = 1,2,3,\ldots,k$ (k – je počet nezávisle premenných). Ak sú známe hodnoty y_j, $j = 1,2,3,\ldots,N$ (N – je počet nezávislých meraní) spolu s odpovedajúcimi x_{ij}, potom úlohou regresie je odhad parametrov $b_0, b_1, \ldots, b_k$ vo funkcii (2):

$$\hat{y} = f(\,x_0, x_1, x_2, \ldots, x_N, b_0, b_1, b_2, \ldots, b_k\,) \tag{2}$$

Parametre $b_0, b_1, \ldots, b_k$ sa určujú tak, aby minimalizovali funkciu odchýlky (3):

$$u_i = \hat{y}_j - f(x_0, x_1, x_2, ..., x_N, b_0, b_1, b_2, ..., b_k) \quad (3)$$

kde $j = 1, 2, 3, \ldots, N$ a $i = 1, 2, 3, \ldots, k$.

Experiment možno definovať ako ľubovoľný zásah do systému s cieľom pozorovať alebo merať účinky tohto zásahu. V experimente vystupuje jedna (alebo viac) premenná, ktorá sa nazýva ozva (ozvová, výstupná premenná) a reprezentuje výstup z experimentu; a niekoľko predikčných premenných. Predikčná premenná, ktorá sa mení s cieľom posúdiť jej vplyv na ozvu, sa nazýva faktor, ktorý môže mať viac úrovní (hodnôt, či vymedzení). Účinkom (efektom) faktora je odchýlka ozvy, spôsobená odchýlkou úrovne faktora od zvolenej strednej úrovne. Plánovanie experimentu, jeho realizácia, analýza a reakcia na výsledky experimentu sú aktivity, ktoré musia mať istú postupnosť – vyjadruje to tzv. Demingov cyklus [3].

Najjednoduchším typom experimentu sú jednofaktorové experimenty. Ak nás však zaujíma vplyv súčasne viacerých faktorov na ozvu, je potrebné realizovať viacfaktorový experiment. V praxi totiž faktory, ktoré vstupujú do procesu, nepôsobia vždy len aditívnym spôsobom, ale spoločne vo vzájomnej interakcii. Faktoriálny experiment umožňuje analyzovať daný model súčasným variovaním všetkých faktorov. Faktory môžu mať nekonečné množstvo úrovní, avšak pri plánovaní experimentu s cieľom dosiahnuť výsledky v rozumných tvaroch stačí skúmať faktory na dvoch, troch alebo až piatich úrovniach. Experiment, v ktorom sú realizované všetky možné kombinácie hladín faktorov bez opakovaní, nazývame úplným faktorovým experimentom. Ak je počet faktorov k a počet hladín každého z nich je λ, potom počet kombinácii N pri úplnom faktorovom experimente zodpovedá počtu pokusov a platí vzťah $N = \lambda^k$. Pri dvoch úrovniach jednotlivých faktorov, ktorých jednotlivé úrovne sa kódujú ako -1 a $+1$ ide o faktorový experiment typu 2^k. Experimentálne body so súradnicami $x_i = -1$ a $x_i = +1$ $(i = 1,2,...,k)$ sú umiestnené vo vrcholoch štvorca (pre $k = 2$), kocky ($k = 3$), alebo hyperkocky ($k > 3$). Takýto dvojhladinový faktorový experiment je vhodné použiť na štatistickú optimalizáciu procesných faktorov, ktoré štatisticky významne ovplyvňujú variabilitu hodnôt profilov drsnosti Ra, Rz a Rq. Tieto veličiny charakterizujú drsnosť povrchu. Najčastejším hodnotiacim kritériom drsnosti povrchu je Ra - stredná aritmetická hodnota absolútnych odchýliek profilu v rozsahu základnej dĺžky, Rq je stredná kvadratická odchýlka profilu a Rz je najväčšia výška nerovnosti profilu. Pri skúmaní vplyvu rezných podmienok na skutočný profil drsnosti obrobeného povrchu sme si vytvorili dva modely plánovaného experimentu.

4. Vplyv procesných faktorov na morfológiu obrobeného povrchu

V prípade modelu plánovaného experimentu nás zaujímal vplyv rezných podmienok na morfológiu obrobeného povrchu pri čelnom frézovaní ušľachtilej ocele ISO 1052-82. Sledovali sme štyri faktory vstupujúce do procesu čelného frézovania drážok (tzv. nezávislé premenné): priemer frézy $d\,[\mathrm{mm}]$, otáčky $n\left[min^{-1}\right]$, posuv na zub $f_z\,[\mathrm{mm}]$ a hĺbku rezu $a_p\,[\mathrm{mm}]$. Každý z faktorov nadobúdal dve úrovne. V takom prípade ide o tetradimenzionálny lineárny model, pre tieto faktory boli realizované pokusy v 2^4 rôznych kombináciách úrovní. Z hľadiska optimalizácie rezných podmienok vzhľadom na hodnotu najväčšej výšky nerovnosti profilu drsnosti Rz a strednej aritmetickej odchýlky profilu drsnosti Ra sú v tab.1 uvedené Kódované podmienky pokusov. Z tejto tabuľky je zrejmé, na akých dvoch úrovniach boli konkrétne faktory skúmané. Po zrealizovaní čelného frézovania drážok na vyššie spomínanom CNC stroji sme na meranie parametrov drsnosti povrchu použili Mitutoyho Surftest SJ – 301, pri vyhodnocovaní experimentu softvér Matlab a Statistica.

Tabuľka 1: Kódované podmienky pokusov

Poradie pokusov	Kódované podmienky pokusov				Aktuálne podmienky pokusov			
	x_1	x_2	x_3	x_4	d[mm]	n[min^{-1}]	f_z[mm]	a_p[mm]
1	1	1	1	1	15	260	0,154	2
2	1	1	1	-1	15	260	0,154	1
3	1	1	-1	1	15	260	0,132	2
4	1	1	-1	-1	15	260	0,132	1
5	1	-1	1	1	15	300	0,154	2
6	1	-1	1	-1	15	300	0,154	1
7	1	-1	-1	1	15	300	0,132	2
8	1	-1	-1	-1	15	300	0,132	1
9	-1	1	1	1	13	260	0,154	2
10	-1	1	1	-1	13	260	0,154	1
11	-1	1	-1	1	13	260	0,132	2
12	-1	1	-1	-1	13	260	0,132	1
13	-1	-1	1	1	13	300	0,154	2
14	-1	-1	1	-1	13	300	0,154	1
15	-1	-1	-1	1	13	300	0,132	2
16	-1	-1	-1	-1	13	300	0,132	1

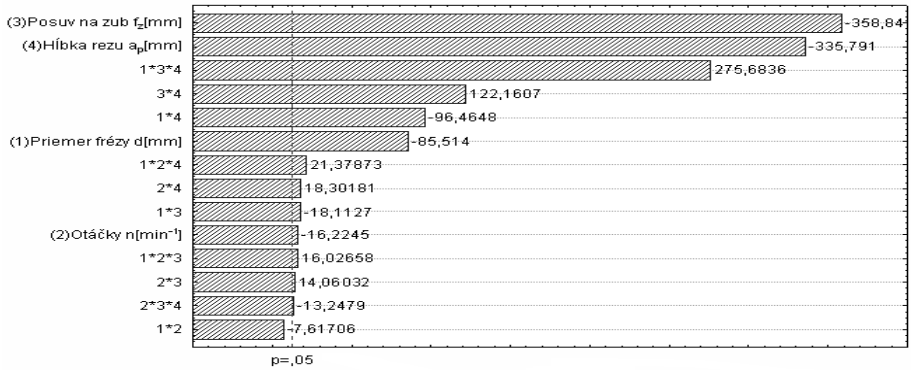
Pri štatistickej optimalizácii pomocou faktorového experimentu 2^4 sa sledovala významnosť vplyvu štyroch procesných faktorov frézovania ušľachtilej ocele (priemer frézy, otáčky, posuv na zub a hĺbka rezu) na kvalitu obrobeného povrchu. Matematický model hodnotenia kvality obrobeného povrchu, ktorý popisuje závislosť ozvy y na faktoroch $x_1,...,x_4$ pri čelnom frézovaní drážok, sa vyjadruje polynómom prvého stupňa v 4-rozmernom priestore [1].Rovnica modelu pre hodnotenie skutočného profilu drsnosti povrchu typu 2^4 pre závislú premennú $\tilde{y}_b$ je:

$$\tilde{y}_b = b_0x_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{14}x_1x_4 + b_{23}x_2x_3 + b_{24}x_2x_4 + \\ + b_{34}x_3x_4 + b_{123}x_1x_2x_3 + b_{134}x_1x_3x_4 + b_{234}x_2x_3x_4 + b_{1234}x_1x_2x_3x_4$$

Na základe hypotézy o významnosti jednotlivých koeficientov rovnice z lineárnej regresie bola získaná výsledná rovnica (4), ktorá charakterizuje maximálnu výšku drsnosti Rz

$$\tilde{y}_{Rz} = -0{,}58x_1 - 0{,}111x_2 - 2{,}287x_3 - 0{,}475x_4 + 0{,}137 \qquad (4)$$

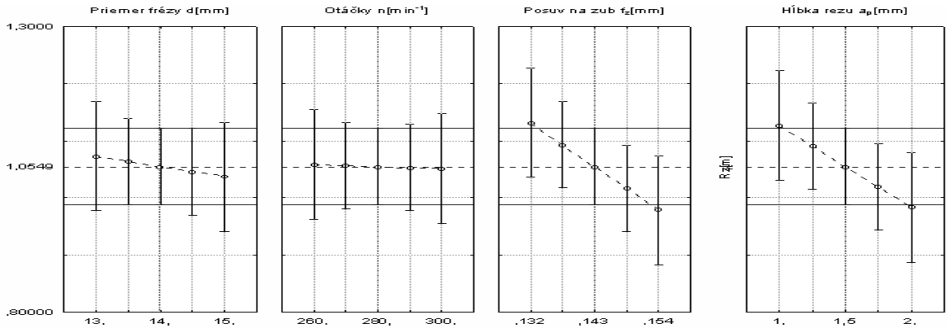
Z výsledkov experimentu vyplynulo, že najväčší vplyv na celkovú hodnotu Rz má posuv na zub (64% vplyv). Ďalšími signifikantnými faktormi sú priemer frézy (16% podiel) a hĺbka rezu (13% vplyv na maximálnu výšku profilu drsnosti Rz). Ostatné faktory ako napr. geometria rezných klinov frézy, materiálové vlastnosti, tuhosť stroja a pod. vyjadrené absolútnym členom predstavujú 4% - ný vplyv. Štatistický nevýznamné otáčky predstavujú 3%. Ako samostatne pôsobiaci faktor nemá výrazne ovplyvňujúci dopad na kvalitu obrobenej plochy, ale vo vzájomnej interakcii s ostatnými do technologického procesu vstupujúcimi faktormi a spoločnej synergii jeho pôsobenie výraznejšie ovplyvňuje maximálnu výšku profilu drsnosti Rz.



Obrázok 1: Vplyv faktorov pri vzájomných interakciách na Rz

Ďalej z výsledkov experimentu vyplynulo, že parameter 0,132 mm posuvu na zub f_z pri priemere frézy $d = 13\,\text{mm}$ má menší vplyv na klesanie drsnosti Rz v hĺbke rezu a_p 1 až 2 mm ako parameter posuvu na zub 0,154 mm v rovnakom intervale hĺbky rezu. Pri priemere

frézy $d = 15\,\text{mm}$ pre posuv na zub 0,132 mm sa ukázal markantný vplyv na pokles Rz v tom istom intervale hĺbky rezu 1 až 2 mm voči parametru 0,154 mm posuvu na zub.



Obrázok 2: Vplyv jednotlivých faktorov na Rz

Z grafu na obr. 2 je zrejmé, že posuv na zub je štatistický významný faktor a so svojou zvýšujúcou sa hodnotou spôsobuje pokles skúmaného paramertra Rz. Podobný vplyv má aj hĺbka rezu a priemer frézy. Vplyv otáčok na celkovú hodnotu maximálnej výšky profilu drsnosti Rz sa ukázal ako zanedbateľný. Na základe analýzy sa preukázal až 96% vplyv rezných podmienok na maximálnu výšku profilu drsnosti Rz, zvyšné 4% predstavujú zanedbateľné faktory. Experimentálne sa potvrdila rôzna významnosť sledovaných procesných faktorov vplývajúcich na drsnosť a asymetriu povrchu pri čelnom frézovaní. Výsledky štatistickej optimalizácie ako aj prezentovaný štatistický model sú platné pre obrábanie ušľachtilej ocele Fe490 (ISO 1052 – 82) v intervale použitých rezných podmienok.

5. Záver

Prínosom príspevku je stanovenie pôvodných funkčných závislostí výstupných parametrov kvality na parametroch vstupujúcich do technologického procesu a vytvorenie pôvodných matematických modelov na základe spracovania štatistických súborov získaných experimentálnych meraním.

6. Literatúra

[1]MIŠÍK, L. – GREGOVÁ, L. – MATIJA, R. 2007. eStatistical Significance of Factors on a Real Profil of Surface Rouhgness after Side Milling for High-grade Steel Fe490.In. TSO´2007. Prešov: FVT TU Košice. ISBN 978-80-8073-900-3.

[2]KAŽIMÍR, I. – BEŇO, J. 1989. Teória obrábania. Bratislava: ALFA, 1989. 280 s. ISBN 063-720-89.

[3]TEREK, M. – HRNČIAROVÁ, Ľ. 2004. Štatistické riadenie kvality. Bratislava: Ekonómia, 2004. 240 s. ISBN 80-89047-97-1.

Adresa autora:

Alena Vagaská, PaedDr., PhD.
Bayerova 1
080 01 Prešov
alena.vagaska@tuke.sk

Hodnotenie zamestnanosti vo vybraných regiónoch členských krajín Európskej únie
Regional evaluation of employment in selected countries of European Union

Mária Vojtková

Abstract: This article deals with multivariate evaluation of 55 selected regions of European Union in 2005. Because the input variables are dependent, we have to transform the dependent variables to a independent principal component and then carry out the evaluation. The result is evaluation of EU regions according to different point of view employment.

Key words: strategy of employment, regional analysis, principal component method

Kľúčové slová: stratégia zamestnanosti , regionálna analýza, metóda hlavných komponentov

1. Úvod

V roku 1997 odštartoval Luxemburský summit o pracovných miestach myšlienku **Európskej stratégie zamestnanosti**. Stratégia má *tri ciele*: dosiahnutie plnej zamestnanosti, zvýšenie produktivity a kvality práce a podpora kohézie.

V odpovedi na dvojitú výzvu globalizácie a demografických zmien (starnutie populácie) si dala Európska rada (v Lisabone a neskôr Solúne) ambiciózne ciele v oblasti zamestnanosti. *V roku 2010* sa má dosiahnuť:

- 70% celková úroveň zamestnanosti ,
- 60% úroveň zamestnanosti žien,
- 50% úroveň zamestnanosti starších pracovníkov.

Jednou z výziev, ktorým sa politiky zamestnanosti a sociálne systémy v EÚ museli efektívne postaviť, bolo rozšírenie únie o **desať nových členských krajín** v roku 2004 a **dve nové členské krajiny** v roku 2007.

Cieľom tohto článku je hodnotenie zamestnanosti v 55 regiónoch nových členských krajín EÚ (okrem pôvodných 15 členských krajín) za rok 2005 na úrovni NUTS 2 (nomenklatúra územných štatistických jednotiek).

2. Regionálna analýza zamestnanosti v krajinách EÚ

Na hodnotenie zamestnanosti vybraných členských krajín Európskej únie na regionálnej úrovni v roku 2005 bol stanovený súbor vybraných štruktúrnych ukazovateľov:

1. **Miera zamestnanosti celková (v %) (Zam_15-64)** – Zamestnané osoby vo veku od 15-64 rokov ako podiel na celkovej populácií v rovnakej vekovej skupine – harmonizované s dátami obyvateľstva a národnými účtami. Dáta sú vykazované 12 týždňov po skončení príslušného roka ako ročný priemer.
2. **Miera zamestnanosti žien (v %) (Zam_zeny)** – Zamestnané ženy vo veku od 15-64 rokov ako podiel na celkovej populácií v rovnakej vekovej skupine – harmonizované s dátami obyvateľstva a národnými účtami. Dáta sú vykazované 12 týždňov po skončení príslušného roka ako ročný priemer.
3. **Miera zamestnanosti starších pracovníkov (v %) (Zam_starsi)** – Zamestnané osoby vo veku od 55-64 rokov ako podiel na celkovej populácií v rovnakej vekovej skupine –

harmonizované s dátami obyvateľstva a národnými účtami. Dáta sú vykazované 12 týždňov po skončení príslušného roka ako ročný priemer.

Základným predpokladom usporiadania vybraných regiónov EÚ je predpoklad vzájomnej nezávislosti (nekorelovanosti) vybraných štruktúrnych ukazovateľov, proti nemu však stojí všeobecná súvislosť javov. Dôsledkom nesplnenia tohto predpokladu je duplicitnosť analyzovaných informácií obsiahnutých vo vstupných ukazovateľoch, ktorá môže viesť ku značnému skresleniu výsledkov. Riešenie tohto problému je možné dosiahnuť použitím analýzy hlavných komponentov pred samotným hodnotením zamestnanosti regiónov EÚ.

Metóda hlavných komponentov spočíva v transformácii k-rozmerného vektora premenných X_j na q-rozmerný vektor hlavných komponentov F_h ($q \leq k$) tak, aby jednotlivé hlavné komponenty boli navzájom ortogonálne a vyčerpávali maximum celkového rozptylu:

$F_h = \alpha_{1h} X_1 + \alpha_{2h} X_2 + ... + \alpha_{kh} X_k$, kde $h = 1, 2, ..., q$.

Nové (skyté, latentné) premenné musia spĺňať nasledovné vlastnosti:

– výberové hlavné komponenty (HK) sú lineárnou kombináciou pôvodných štandardizovaných premenných X_j,

– maximálne možno vytvoriť rovnaký počet HK ako pôvodných premenných,

– nové hlavné komponenty sú vzájomne nekorelované (nezávislé, ortogonálne).

Tabuľka 1: Vlastné čísla korelačnej matice

Vlastné čísla korelačnej matice: Celkovo = 3 Priemer = 1				
	Vlastné číslo	**Rozdiel**	**Podiel**	**Kumulatívny podiel**
1	2.69758310	2.48280094	0.8992	0.8992
2	0.21478216	0.12714743	0.0716	0.9708
3	0.08763473		0.0292	1.0000

V našom prípade vstupujú do analýzy tri štruktúrne ukazovatele zamestnanosti, čiže z nich môžeme skonštruovať iba tri hlavné komponenty. Obvyklým cieľom metódy hlavných komponentov je vysvetliť maximum rozptylu medzi vstupnými premennými pomocou minimálneho počtu komponentov, čiže redukcia dát. Výsledkom je dosiahnutie nezávislých premenných, ktoré môžeme použiť na hodnotenie zamestnanosti.

Vzhľadom k vysokej závislosti všetkých vstupných štruktúrnych ukazovateľov a tiež vzhľadom k ich malému počtu je zrejmé, že na vysvetlenie podstatnej časti variability pôvodných vstupných ukazovateľov bude stačiť jeden hlavný komponent. Vlastné číslo charakterizuje rozptyl každého komponenta (tabuľka 1), pričom celkový rozptyl vzhľadom k tomu, že pracujeme so štandardizovanými premennými je rovný 3. V poslednom stĺpci tabuľky 1 je uvedený kumulatívny podiel variability vysvetlený daným počtom komponentov.

Náš predpoklad, že za dostačujúci počet hlavných komponentov budeme považovať prvý hlavný komponent sa potvrdil, pretože vysvetľuje až 89,92% celkovej variability vstupných ukazovateľov. Hodnotenie zamestnanosti vybraných regiónov EÚ sa teda stáva jednorozmerným hodnotením.

Výsledkom aplikácie metódy hlavných komponentov je komponentná matica obsahujúca komponentné saturácie pre jednotlivé štruktúrne ukazovatele a komponenty

(Tabuľka 2). V tomto prípade ide o párové koeficienty korelácie príslušnej premennej a prvého hlavného komponenta. Na základe ich veľkosti, ktorá sa vo všetkých prípadoch blíži k 1, môžeme skonštatovať, že medzi prvým hlavným komponentom a jednotlivými štruktúrnymi ukazovateľmi zamestnanosti je štatisticky významná lineárna závislosť od všetkých.

Tabuľka 2: Matica komponentnej štruktúry

Komponentná matica		
		HK1
Zam15-64	Zam15-64	0.96332
Zam_zeny	Zam_zeny	0.95696
Zam_starsi	Zam_starsi	0.92403

Podľa matice komponentných saturácii je možné pre jednotlivé regióny vybraných členských krajín EÚ vypočítať hodnoty komponentných skóre a následne ich usporiadať na základe veľkosti. Výsledkom je ligová tabuľka 3, ktorá obsahuje hodnotenie zamestnanosti vybraných regiónov EÚ.

Tabuľka 3: Usporiadanie vybraných regiónov krajín EÚ podľa zamestnanosti

Poradie	Regióny	Krajina	Zam_15-64	Zam_zeny	Zam_starsi	HK1
1	Praha	Czech republic	71,3	64,5	58,5	2,42
2	Bratislavský	Slovakia	69,6	63,6	52,2	2,05
3	Estonia	Estonia	64,5	62,1	56,0	1,80
4	Cyprus	Cyprus	68,5	60,3	50,6	1,75
5	Jihozápad	Czech republic	67,8	58,9	45,9	1,46
6	Střední Čechy	Czech republic	67,0	57,9	47,7	1,42
7	Nord-Est	Romania	61,4	59,0	54,9	1,41
8	Latvia	Latvia	63,3	59,3	49,5	1,35
9	Lithuania	Lithuania	62,6	59,4	49,2	1,30
10	Severovýchod	Czech republic	65,7	56,3	43,4	1,11
11	Slovenija	Slovenia	66,0	61,3	30,7	0,96
12	Sud-Vest Oltenia	Romania	60,0	54,3	51,9	0,96
13	Jihovýchod	Czech republic	64,1	55,4	41,6	0,90
14	Yugozapaden	Bulgaria	61,5	57,8	39,0	0,79
15	Severozápad	Czech republic	61,5	53,2	43,7	0,70
16	Střední Morava	Czech republic	62,1	52,8	39,6	0,57
17	Nyugat-Dunántúl	Hungary	62,1	55,9	34,6	0,57
18	Sud-Muntenia	Romania	57,9	50,2	42,5	0,28
19	Közép-Dunántúl	Hungary	60,2	52,9	34,0	0,27
20	Moravskoslezsko	Czech republic	59,3	51,7	36,5	0,23
21	Západné Slovensko	Slovakia	60,6	53,7	28,8	0,15
22	Dunántúl	Hungary	58,7	52,3	32,2	0,08
23	Mazowieckie	Poland	57,6	51,8	33,3	0,03
24	Bucureşti-Ilfov	Romania	59,3	53,4	26,6	-0,02
25	Nord-Vest	Romania	55,9	51,1	35,5	-0,03
26	Lubelskie	Poland	56,0	51,0	34,9	-0,06
27	Podlaskie	Poland	56,9	50,4	32,0	-0,14

Poradie	Regióny	Krajina	Zam_15-64	Zam_zeny	Zam_starsi	HK1
28	Yuzhen tsentralen	Bulgaria	54,7	50,5	33,5	-0,21
29	Vest	Romania	56,5	49,5	31,9	-0,21
30	Małopolskie	Poland	55,0	49,8	33,2	-0,24
31	Yugoiztochen	Bulgaria	54,2	48,1	36,9	-0,25
32	Severen tsentralen	Bulgaria	53,4	50,8	31,9	-0,32
33	Severoiztochen	Bulgaria	53,9	48,2	34,4	-0,35
34	Sud-Est	Romania	54,6	46,2	36,1	-0,36
35	Stredné Slovensko	Slovakia	55,2	48,1	27,6	-0,52
36	Podkarpackie	Poland	52,3	48,0	32,2	-0,54
37	Dél-Alföld	Hungary	53,8	47,4	28,4	-0,61
38	Dél-Dunántúl	Hungary	53,4	47,8	27,6	-0,64
39	Centru	Romania	54,0	46,6	28,4	-0,65
40	Łódzkie	Poland	54,1	49,2	23,3	-0,67
41	Świętokrzyskie	Poland	51,6	47,1	30,8	-0,68
42	Wielkopolskie	Poland	54,0	45,8	26,8	-0,75
43	Opolskie	Poland	52,5	44,4	26,1	-0,94
44	Kujawsko-Pomorskie	Poland	51,5	44,9	25,2	-1,00
45	Východné Slovensko	Slovakia	51,5	44,6	24,4	-1,04
46	Pomorskie	Poland	51,0	43,3	27,2	-1,05
47	Észak-Alföld	Hungary	50,2	43,9	26,3	-1,09
48	Lubuskie	Poland	51,1	44,4	22,3	-1,15
49	Severozapaden	Bulgaria	47,0	44,9	27,7	-1,18
50	Észak-Magyarország	Hungary	49,5	44,7	23,5	-1,19
51	Dolnośląskie	Poland	49,3	44,0	23,2	-1,25
52	Malta	Malta	53,9	33,6	30,8	-1,29
53	Zachodniopomorskie	Poland	48,3	41,8	25,5	-1,35
54	Warmińsko-Mazurskie	Poland	48,7	42,4	23,2	-1,37
55	Śląskie	Poland	49,5	43,8	18,6	-1,41

Na prvom mieste výrazne dominuje región Praha, ktorý dosahuje najvyššie hodnoty vo všetkých troch ukazovateľoch zamestnanosti a už v roku 2005 spĺňal ciele stanovené Európskou radou na rok 2010. Na druhom mieste je región Bratislava, ktorý dosahuje druhé najvyššie hodnoty celkovej zamestnanosti a zamestnanosti žien, avšak pri hodnotení samotného ukazovateľa zamestnanosti starších by bol až na 4. mieste. Celkovo región Bratislava len máličko zaostáva za cieľom Európskej rady avšak na jeho splnenie má ešte dostatočnú časovú rezervu. Na treťom mieste sa nachádza región Cyprus, ktorý vzhľadom k veľkosti zahŕňa celú členskú krajinu. Toto umiestnenie dosiahol predovšetkým vďaka vysokej hodnote ukazovateľa celkovej zamestnanosti, pretože v oblasti zamestnanosti žien sa nachádza až na 5. mieste a v oblasti zamestnanosti starších na 6. mieste.

Na druhej strane ligovej tabuľky sa nachádzajú tri regióny z Poľska Śląskie, Warmińsko-Mazurskie a Zachodniopomorskie. Región Śląskie, ktorý sa umiestnil z pohľadu zamestnanosti na tom najhoršie dosahuje predovšetkým veľmi nízku úroveň zamestnanosti starších pracovníkov, pričom vo zvyšných dvoch ukazovateľoch zamestnanosti sa jednotlivo nachádza na 51. mieste. Región Warmińsko-Mazurskie sa pri hodnotení jednotlivých ukazovateľov zamestnanosti nachádza zhodne na 53. mieste avšak celkovo je hodnotený ako druhý najhorší. Tretí najhorší región dosahuje 54. miesto za celkovú zamestnanosť i zamestnanosť žien avšak za zamestnanosť starších pracovníkov je na 47. mieste.

Čo sa týka samotného hodnotenia regiónov Slovenska je možné pozorovať značné rozdiely. Zatiaľ, čo región Bratislava je na druhom mieste, ostatné regióny sú v rebríčku umiestnené nižšie. Región západné Slovensko je na 21. mieste, pričom sa nachádza tesne pred regiónom, v ktorom je hlavné mesto Maďarska (22) a Poľska (23). Stredné Slovensko je na 35. mieste a región východné Slovensko až na 45. mieste. Celkovo na Slovensku badať z hľadiska splnenia stanovených cieľov Európskou radou najväčšie rezervy hlavne v oblasti zamestnanosti starších pracovníkov. Podľa veľkosti komponentného skóre je možné pozorovať aj určité zoskupenia regiónov EÚ na základe podobnosti štruktúrnych ukazovateľov zamestnanosti, ktoré by mohlo byť predmetom ďalšej analýzy.

4. Záver

Obnovená lisabonská agenda identifikovala tri zásadné oblasti v opätovnom spustení svojej stratégie: posilňovanie vedomostí a inovácií, ako motorov trvalo udržateľného rozvoja, zabezpečenie, aby sa EÚ stala príťažlivou oblasťou pre investície a prácu a uznanie, že rast a zamestnanosť sú najlepšími prostriedkami na podporu vzájomnej súdržnosti jednotlivých regiónov EÚ. V tomto zohrávajú vlády rozhodujúcu úlohu, pretože štrukturálne reformy sú mimoriadne dôležité, ak sa tieto ciele majú splniť. Vo svojich národných reformných programoch z roku 2005 sa členské štáty zameriavajú na otázky, ktoré úzko súvisia s Integrovaným usmernením pre rast a zamestnanosť.

Národné reformné programy naznačujú posun v politike smerom k výskumu a inováciám, účinnosti zdrojov a energie, podnikaniu a vzdelávaniu, investíciám do ľudského kapitálu a modernizácii trhov práce spolu so zabezpečním vysokej úrovne sociálnej ochrany pre budúcnosť. Nové formy sociálnej ochrany sa nemajú zameriavať na udržanie zamestnancov na jednom mieste po celý život, ale na „schopnosť ľudí ostať súčasťou trhu práce a postupovať na ňom“ (dostatočný plat, prístup k celoživotnému vzdelávaniu, pracovné podmienky, ochrana pred nespravodlivým prepustením, podpora v prípade straty zamestnania, právo na transfer sociálnych práv v prípade zmeny zamestnania).

Celkovo môžeme zhodnotiť, že dvanásť nových členských krajín omnoho inovatívnejšie využíva pracovné možnosti (dočasné kontrakty, využitie sprostredkovateľských agentúr, samozamestnávanie formou živnosti, či práca na čiastočný úväzok), než je tomu v 15 pôvodných členských krajinách a teda intenzívne smeruje k dosiahnutiu cieľa stanoveného Európskou radou na rok 2010.

5. Literatúra

[1] KHATTREE, R. – NAIK, N. D.*: Multivariate Data Reduction and Discrimination with SAS® Software.* First edition, Cary, NC: SAS Institute Inc., 2000.
[2] SHARMA, S.: *Applied multivariate techniques*. New York: John Wiley & Sons, 1996.
[3] STANKOVIČOVÁ, I.: Viackriteriálne hodnotenie zamestnanosti členských krajín EÚ na základe vybraných ukazovateľov Lisabonskej stratégie. In: Statistika–ekonomicko-statistický časopis, Roč. 45, č. 4, (2008), s. 304-312
[4] STANKOVIČOVÁ, I. - VOJTKOVÁ, M.: *Viacrozmerné štatistické metódy s aplikáciami*. Bratislava: Iura Edition, 2007.
[5] ŠOLTÉS, E.: *Regresná a korelačná analýza s aplikáciami*. Bratislava: Iura Edition, 2008.
[6] Štvrtá správa o hospodárskej a sociálnej kohézii.
http://ec.europa.eu/regional_policy/sources/docoffic/official/reports/cohesion4/pdf/4cr_sk.pdf

Adresa autora:

Mária Vojtková, Ing. PhD.
Katedra štatistiky, FHI, Ekonomická univerzita
Dolnozemská 1/b, 852 35 Bratislava
E-mail: vojtkova@euba.sk

Príspevok bol vypracovaný v rámci riešenia APVV 0649-07 ANARED.

Odhad subjektívnej chudoby na Slovensku založený na distribučnej funkcii rozdelenia príjmov

Estimation of Subjective Poverty in Slovakia Based on Cumulative Distribution Function of Income

Tomáš Želinský, Oto Hudec

Abstract: The paper deals with the perception of subjective poverty in Slovakia. The aim of the presented paper is to estimate subjective poverty line using approach of „discrete information". We propose a model of estimation subjective poverty line based on comparison of empirical cumulative distribution functions of income. Using the EU SILC microdata, three alternatives are compared (in terms of who can be considered poor). If the lowest (out of three) estimated poverty line in terms of equivalent disposable income is used, the proportion of poor people is three-times higher than the proportion of poor calculated using the official EU poverty line. As for the level of households, there are approximately 25% of subjectively pool households.

Key words: poverty, subjective poverty, poverty line, EU SILC

Kľúčové slová: chudoba, subjektívna chudoba, hranica chudoby, EU SILC.

1. Úvod

Je vždy náročné zachytiť a presne popísať, prípadne kvantifikovať sociálne javy subjektívnej povahy. Vnímanie subjektívnej chudoby je nepochybne jedným z najtypickejších prípadov takých javov.

Existuje viacero teoretických prístupov zameraných na odhad hraníc subjektívnej chudoby. V tomto príspevku navrhujeme model na odhad hranice subjektívnej chudoby založený na porovnávaní empirických distribučných funkcií rozdelenia príjmov dvoch skupín obyvateľov – tých, ktorých môžeme považovať za chudobných a tých ostatných. Model je navrhnutý v rámci prístupu založeného na tzv. „diskrétnom" položení otázky ohľadom vnímania subjektívnej príjmovej situácie respondentov.

2. Vymedzenie pojmu subjektívna chudoba

Väčšina bežne používaných ekonomických definícií chudoby má dva spoločné prvky. Prvým krokom je určenie *indikátora blahobytu.* Následne je potrebné stanoviť „*deliaci bod*" – hranicu chudoby určujúcu úroveň daného indikátora pre jednotlivca, pod ktorou je považovaný za chudobného (napr. [1], [2], [3], [4]).

Existuje viacero prístupov k určeniu hranice chudoby, ktoré je možné členiť z viacerých typologických hľadísk – napríklad subjektívny/objektívny prístup, absolútny/relatívny prístup.

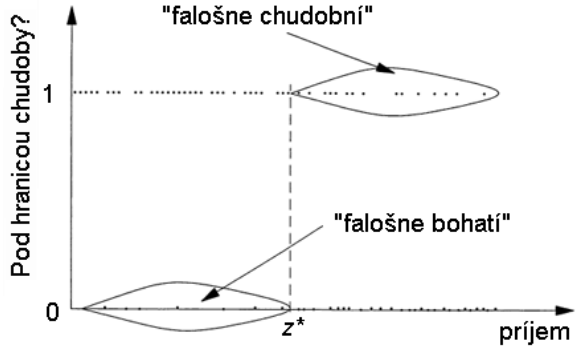
Ako už vyplýva z názvu príspevku, svoju pozornosť sústredíme na subjektívny koncept chudoby.

Subjektívny prístup vyjadruje, že hranice chudoby sú inherentnými subjektívnymi úsudkami ľudí o tom, čo považujú za sociálne akceptovateľný minimálny životný štandard v určitej spoločnosti [5]. Prístup vychádza z predpokladu, že podmienky, v ktorých sa jednotlivec nachádza v porovnaní k okolnostiam ostatných členov referenčnej skupiny tak, ako ich vníma, ovplyvňujú vnímanie jeho osobného blahobytu relatívne k ostatným členom referenčnej skupiny. [6].

2.1 Určenie hraníc subjektívnej chudoby

Určenie hraníc subjektívnej chudoby môže vychádzať *zo stanovenia takej sumy, s ktorou by domácnosť „vystačila“* [4], resp. by uspokojili svoje (subjektívne vnímané) nevyhnutné potreby. Odpoveď na túto otázku by hypoteticky mala byť rastúcou funkciou aktuálneho príjmu jednotlivca [5]. Pri použití tejto metódy je potrebné zohľadniť skutočnosť, že ľudia neradi odpovedajú na senzitívne otázky, akou otázka o príjme určite je. Ďalšou nevýhodou metódy môže byť veľká variabilita odpovedí a z toho vyplývajúce náročné (až nemožné) určenie funkcie minimálneho subjektívneho príjmu v závislosti od aktuálneho príjmu.

Ďalšia metóda, ako určiť subjektívnu hranicu chudoby vychádza z „diskrétne“ položenej otázky [7]. Diskrétnosť otázky spočíva v tom, že respondenti nemajú stanoviť absolútnu hodnotu príjmu, ktorý považujú za minimálne požadovaný. Majú v odpovedi uviesť, či vzhľadom na svoj aktuálny príjem sa cítia byť chudobní (hodnota „1“), alebo nie (hodnota „0“). Následne je žiaduce odhadnúť hodnotu subjektívnej hranice z^*. Hodnota z^* je určená ako hodnota, ktorej zodpovedá najvyššia pravdepodobnosť, že odpovede respondentov o ich statuse chudoby korešpondujú so statusom, ktorý by sme zistili porovnaním hodnoty z^* s ich príjmom.



Obr. 2.1: Odhad subjektívnej hranice chudoby
Zdroj: [7, s. 125]

Ak situáciu zachytíme graficky, odhadom z^* sa minimalizuje pravdepodobnosť výskytu pozorovaní nachádzajúcich sa vo vyznačených elipsách na obr. 2.1 - tzn. pozorovaní, v ktorých osoby s príjmom nižším ako z^* sa necítia byť chudobné (falošne bohaté) a naproti tomu osoby s príjmom vyšším ako z^* sa chudobnými cítia (falošne chudobné).

Ani v tomto prípade sa prirodzene nevyhneme problému citlivosti údajov o aktuálnom príjme respondentov.

Vychádzajúc z tohto predpokladu môžeme na odhad hranice chudoby využiť napríklad logistickú regresiu (napr. [4]), pozornosť je však možné sústrediť aj na porovnanie empirických distribučných funkcií (ďalej len ECDF – z angl. *empirical cummulative distribution function*) „chudobných“ a „nechudobných“.

3. Metódy

3.1 Odhad hranice chudoby

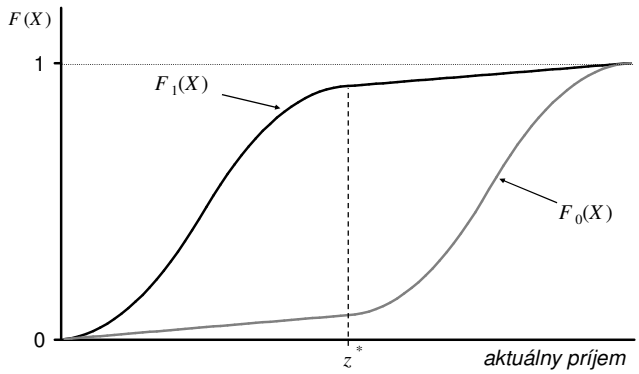
Majme dvojrozmerné pozorovania $[X_i; Y_i]$, kde X_i je aktuálny príjem i-tej osoby a Y_i je binárna premenná popisujúca subjektívne vnímanie pocitu chudoby i-tej osoby, a teda môže nadobúdať hodnotu 0 (nie je subjektívny pocit chudoby) alebo 1 (je subjektívny pocit chudoby).

Nech rozdelenie príjmov osôb, ktoré uviedli, že sa cítia byť chudobné ($Y_i = 1$), je dané spojitou funkciou hustoty $f_1(x)$ so strednou hodnotou $E_1(X)$ a rozdelenie príjmov osôb, ktoré uviedli, že sa necítia byť chudobné ($Y_i = 0$), je dané spojitou funkciou hustoty $f_0(x)$ so strednou hodnotou $E_0(X)$. Nech obe funkcie sú definované na intervale $[0; \infty)$ a zodpovedajú im distribučné funkcie $F_1(X)$ a $F_0(X)$. Ďalej nech platí $E_1(X) < E_0(X)$.

Za hranicu subjektívnej chudoby z^* budeme považovať takú hodnotu príjmu X, pri ktorej maximalizujeme rozdiel empirických distribučných funkcií, t. j.

$$z^* = x_0 : \max_{x \in \langle 0;\infty)} \left[F_1(x) - F_0(x) \right]. \tag{1}$$

Situáciu je možné znázorniť graficky:



Obr. 2.2: Odhad subjektívnej hranice chudoby na základe ECDF
Zdroj: vlastný obrázok

Predpokladajme, že $F_1(X)$ môžeme na určitom ohraničenom intervale aproximovať polynomickou funkciou n-tého stupňa $g_1(x)$ a $F_0(X)$ funkciou $g_0(x)$. Označme ich rozdiel ako funkciu $g(x) = g_1(x) - g_0(x)$. Potom za predpokladu $\frac{\partial^2 g(x)}{\partial x^2} < 0$ je odhadom hranice chudoby z^* taký kladný koreň x_0 rovnice

$$\frac{\partial g(x)}{\partial x} = 0, \tag{2}$$

pre ktorý platí:

$$x_0 : \max_x \{ g(x) \}. \tag{3}$$

3.2 Popis vstupných údajov

Vstupnými údajmi pre analýzu sú mikroúdaje zisťovania EU SILC 2006 [8]. Pre detaily týkajúce sa popisu vzorky a zberu údajov, pozri [9].

Za príjem budeme v modeli pokladať *(1.) ekvivalentný disponibilný príjem domácnosti*, t. j. podiel celkového ročného disponibilného príjmu domácnosti a ekvivalentnej veľkosti domácnosti [9] a (2.) celkový disponibilný príjem domácnosti (tzn. prepočet uskutočníme s použitím oboch druhov príjmov).

Domácnosťou rozumieme hospodáriacu domácnosť, t. j. súkromnú domácnosť tvorenú 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 za bývanie, elektrina, plyn a pod.) [9]

Vnímanie subjektívnej chudoby domácnosťou hodnotíme na základe otázky v dotazníku: *„Domácnosť môže mať rôzne zdroje príjmu a viac ako jeden člen domácnosti môže prispievať do celkového príjmu. Ak sa zamyslíte nad celkovým mesačnými príjmom Vašej domácnosti, s akými ťažkosťami je Vaša domácnosť schopná splatiť zvyčajné výdavky?“*

1. s veľkými ťažkosťami,
2. s ťažkosťami,
3. s určitými ťažkosťami,
4. pomerne ľahko,
5. ľahko,
6. veľmi ľahko.

Vnímanie subjektívnej chudoby domácnosťou je binárnou premennou, preto je potrebné škálu odpovedí tejto otázky transformovať. Označme ako y_i odpoveď i-tej domácnosti, pričom $y_i \in \{1, 2, ..., 6\}$ a ako y_i^* transformovanú odpoveď tej istej domácnosti, kde $y_i^* \in \{0, 1\}$.

V snahe vyhnúť sa subjektivite pri tejto transformácii, uskutočníme tri porovnania vychádzajúce z troch transformácií premennej *Y*:

1. $\forall y_i = 1 : y_i^* = 1 \,\&\, \forall y_i > 1 : y_i^* = 0$, (i)
2. $\forall y_i \leq 2 : y_i^* = 1 \,\&\, \forall y_i > 2 : y_i^* = 0$, (ii)
3. $\forall y_i \leq 3 : y_i^* = 1 \,\&\, \forall y_i > 3 : y_i^* = 0$. (iii)

(V prvom prípade považujeme za subjektívne chudobných len tých, ktorí označili možnosť „1“, v druhom prípade tých, ktorí označili možnosti „1“ alebo „2“ a v treťom prípade tých, ktorí označili „1“, „2“ alebo „3“.)

Vzhľadom na to, že v súbore sa nachádzalo niekoľko málo hodnôt príjmu, ktoré boli výrazne vysoké v porovnaní s ostatnými (ekvivalentný disponibilný príjem vyšší ako 400 000 Sk), dochádzalo k skresľovaniu výsledkov, a preto sme sa rozhodli tieto pozorovania zanedbať. Dokopy šlo o cca 60 pozorovaní.

4. Výsledky

Pozrime sa najskôr na rozdelenie príjmov respondentov v závislosti od vnímania subjektívnej chudoby (tab. 4.1).

Tab. 4.1: Rozdelenie príjmov v závislosti od vnímania subjektívnej chudoby

Interval	1		2		3		4		5		6		Spolu
(4453,30; 43786,18]	**49**	*7,36*	**37**	*3,08*	**32**	*1,40*	**14**	*1,87*	**0**	*0,00*	**2**	*10,53*	**134**
	36,57		*27,61*		*23,88*		*10,45*		*0,00*		*1,49*		*100*
(43786,18; 83119,02]	**214**	*32,13*	**206**	*17,14*	**232**	*10,17*	**47**	*6,28*	**8**	*7,08*	**1**	*5,26*	**708**
	30,23		*29,10*		*32,77*		*6,64*		*1,13*		*0,14*		*100*
(83119,02; 122451,87]	**257**	*38,59*	**531**	*44,18*	**798**	*34,97*	**164**	*21,93*	**12**	*10,62*	**3**	*15,79*	**1765**
	14,56		*30,08*		*45,21*		*9,29*		*0,68*		*0,17*		*100*
(122451,87; 161784,71]	**111**	*16,67*	**290**	*24,13*	**686**	*30,06*	**193**	*25,80*	**21**	*18,58*	**2**	*10,53*	**1303**
	8,52		*22,26*		*52,65*		*14,81*		*1,61*		*0,15*		*100*
(161784,71; 201117,55]	**21**	*3,15*	**88**	*7,32*	**322**	*14,11*	**145**	*19,39*	**26**	*23,01*	**4**	*21,05*	**606**
	3,47		*14,52*		*53,14*		*23,93*		*4,29*		*0,66*		*100*
(201117,55; 240450,40]	**8**	*1,20*	**34**	*2,83*	**121**	*5,30*	**86**	*11,50*	**20**	*17,70*	**2**	*10,53*	**271**
	2,95		*12,55*		*44,65*		*31,73*		*7,38*		*0,74*		*100*
(240450,40; 279783,24]	**3**	*0,45*	**8**	*0,67*	**52**	*2,28*	**53**	*7,09*	**10**	*8,85*	**1**	*5,26*	**127**
	2,36		*6,30*		*40,94*		*41,73*		*7,87*		*0,79*		*100*
(279783,24; 319116,09]	**0**	*0,00*	**5**	*0,42*	**20**	*0,88*	**24**	*3,21*	**6**	*5,31*	**2**	*10,53*	**57**
	0,00		*8,77*		*35,09*		*42,11*		*10,53*		*3,51*		*100*
(319116,09; 358448,93]	**3**	*0,45*	**2**	*0,17*	**15**	*0,66*	**14**	*1,87*	**6**	*5,31*	**1**	*5,26*	**41**
	7,32		*4,88*		*36,59*		*34,15*		*14,63*		*2,44*		*100*
(358448,93; 397781,80]	**0**	*0,00*	**1**	*0,08*	**4**	*0,18*	**8**	*1,07*	**4**	*3,54*	**1**	*5,26*	**18**
	0,00		*5,56*		*22,22*		*44,44*		*22,22*		*5,56*		*100*
Spolu	**666**	*100*	**1202**	*100*	**2282**	*100*	**748**	*100*	**113**	*100*	**19**	*100*	**5030**
	(13%)		***(24%)***		***(45%)***		***(15%)***		***(2%)***		***(0%)***		

Zdroj: vytvorené autormi podľa mikroúdajov EU SILC 2006 [8]

Zaujímavé je všimnúť si, že najpočetnejšou odpoveďou v prieskume bola odpoveď „3“. Ak by sme respondentov rozdelili na „chudobných“ a „nechudobných“, resp. transformovali škálu na binárnu premennú tak, ako je uvedené v bode (iii) v kapitole 3.2 (tzn. za „subjektívne chudobných“ by sme považovali všetkých tých, ktorí zvolili jednu z možností „1“ – „3“), podiel „subjektívne chudobných“ by bol až 82,5%.

4.1 Porovnanie výsledkov – ekvivalentný disponibilný príjem

Transformáciou premennej *Y* na binárnu premennú *Y** a aproximáciou údajov o ekvivalentnom disponibilnom príjme polynómom *n*-tého stupňa na určitom ohraničenom intervale dostávame dvojice odhadnutých funkcií korelácie (tab. 4.2).

Podľa (1) a (2) dostávame tri odhady hranice subjektívnej chudoby (tab. 4.3) zodpovedajúce funkciám korelácie z tab. 4.2.

Tab. 4.2: Odhad distribučnej funkcie rozdelenia

	Aproximovanie distribučnej funkcie polynómom	**R^2**
(i)	$g_1(x) = 2.10^{-32}x^6 - 3{,}2.10^{-26}x^5 + 1{,}8.10^{-20}x^4 - 5{,}1.10^{-15}x^3 + 6{,}6.10^{-10}x^2 - 2{,}8.10^{-5}x + 0{,}385$ $g_2(x) = -3{,}6.10^{-27}x^5 + 4{,}4.10^{-21}x^4 - 2{,}0.10^{-15}x^3 + 3{,}7.10^{-10}x^2 - 2{,}3.10^{-5}x + 0{,}416$	0,997 0,996
(ii)	$g_1(x) = 7.10^{-33}x^6 - 1{,}4.10^{-26}x^5 + 1{,}1.10^{-20}x^4 - 3{,}6.10^{-15}x^3 + 5{,}5.10^{-10}x^2 - 2{,}8.10^{-5}x + 0{,}434$ $g_2(x) = -2.10^{-32}x^6 - 2{,}4.10^{-26}x^5 + 8{,}7.10^{-21}x^4 - 1{,}01.10^{-15}x^3 + 4{,}0.10^{-11}x^2 - 6{,}9.10^{-6}x + 0{,}162$	0,995 0,997
(iii)	$g_1(x) = -5{,}0.10^{-27}x^5 + 5{,}8.10^{-21}x^4 - 2{,}4.10^{-15}x^3 + 4{,}3.10^{-10}x^2 - 2{,}5.10^{-5}x + 0{,}423$ $g_2(x) = 3{,}1.10^{-28}x^5 + 2{,}2.10^{-22}x^4 - 3{,}7.10^{-16}x^3 + 1{,}2.10^{-10}x^2 - 8{,}9.10^{-6}x + 0{,}192$	0,996 0,997

Zdroj: vlastné výpočty podľa mikroúdajov EU SILC 2006 [8]

Tab. 4.3: Odhad subjektívnej hranice chudoby v SR

Prístup	$\max_{x\in\langle 0;\infty)}\left[F_1(X) - F_0(X)\right]$	**Hranica chudoby z*** *(roč. hodnota ekviv. disp. príjmu)*	**Podiel chudobného obyvateľstva**
(i)	0,3054	110 940	36 %
(ii)	0,2581	121 038	46 %
(iii)	0,3070	151 021	69 %

Zdroj: vlastné výpočty podľa mikroúdajov EU SILC 2006 [8]

Za hranicu subjektívnej chudoby podľa prístupu popísaného v príspevku možno v podmienkach SR považovať osoby, u ktorých hodnota mesačného ekvivalentného disponibilného príjmu je nižšia ako 9 245 (resp. 10 087 alebo 12 585) Sk.

Je evidentné, že aj pri uplatnení „najprísnejšieho" kritéria (t. j. za chudobných považujeme len tých, ktorí „vystačia s peniazmi s veľkými ťažkosťami"), dostávame až trojnásobne vyšší podiel chudobných, ako je to pri uplatnení hranice chudoby definovanej EÚ (12%).

4.2 Porovnanie výsledkov – celkový disponibilný príjem domácnosti

Rovnakým postupom, ako je uvedené vyššie, sme analyzovali situáciu z pohľadu domácnosti (bez prepočtu na ekvivalentný disponibilný príjem na jedného člena).

Tab. 4.4: Odhad subjektívnej chudoby v domácnostiach SR

Prístup	$\max_{x\in\langle 0;\infty)}\left[F_1(X) - F_0(X)\right]$	**Hranica chudoby z*** *(roč. hodnota ekviv. disp. príjmu)*	**Podiel chudobného obyvateľstva**
(i)	0,2548	191 345	26 %
(ii)	0,2254	209 026	30 %
(iii)	0,2424	281 649	50 %

Zdroj: vlastné výpočty podľa mikroúdajov EU SILC 2006 [8]

Ak sledujeme subjektívnu chudobu v domácnostiach bez prepočtu na ekvivalentného člena domácnosti (tab. 4.4), dostávame významne nižšie miery subjektívnej chudoby ako pri jeho uplatnení.

5. Záver

Hlavným výstupom článku je model odhadu hraníc subjektívnej chudoby založený na porovnaní empirických distribučných funkcií príjmov „subjektívne chudobných" a ostatných – subjektívne nie chudobných. Výhodou tohto prístupu je skutočnosť, že pri samotnom

odhade nie je potrebná informácia o príjme, ktorý domácnosti považujú za hranicu subjektívnej chudoby, čo je v zásade citlivý údaj.

Aby sme sa vyhli subjektivite pri rozdelení respondentov na chudobných a nechudobných, uskutočnili sme tri porovnania, ktorým zodpovedajú tri rôzne hranice chudoby. Uskutočnili sme odhad subjektívnej chudoby na základe ekvivalentného disponibilného príjmu a na základe celkového disponibilného príjmu.

Uplatnením napríklad „najprísnejšej" hranice chudoby dostávame v prvom prípade, že viac ako tretina obyvateľstva vníma svoju subjektívnu situáciu ako stav chudoby, tzn. že nevystačia s peniazmi (resp. vystačia s veľkými ťažkosťami). V druhom prípade, t. j. z pohľadu celkového počtu domácností, pociťuje stav subjektívnej chudoby asi štvrtina domácností. To nasvedčuje záveru, že porovnávanie vlastnej situácie s inými, resp. životný pocit sa vyznačuje vo väčšine prípadov na Slovensku nespokojnosťou so stavom svojich príjmov a predstavami o chudobe výrazne sa odlišujúcimi od „objektívnej" skutočnosti v podobe svetovo uznávaných indikátorov chudoby.

6. Literatúra

[1] TOWNSEND, P.**:** Measuring Poverty. In: *The British Journal of Sociology*. Vol. 5, No. 2 (Jun., 1954), pp. 130 – 137.

[2] DESAI, M., SHAH, A.: An Econometric Approach to the Measurement of Poverty. In: *Oxford Economic Papers. New Series*. Vol. 40, No. 3 (Sep., 1988), pp. 505 – 522.

[3] RAVALLION, M.: Issues in Measuring and Modelling Poverty. In: *The Economic Journal.* Vol. 106, No. 438 (Sep., 1996), pp. 1328 – 1343.

[4] KAPTEYN, A., KOOREMAN, P., WILLEMSE, R.: Some Methodological Issues in the Implementation of Subjective Poverty Definitions. In: *The Journal of Human Resources*. Vol. 23, No. 2 (Spring, 1988), pp. 222 – 242.

[5] RAVALLION, M.: *Poverty comparisons: A guide to concepts and methods*. Washington, D. C.: Svetová banka, 1992. ISBN 0-8213-2036-X.

[6] RAVALLION, M.: *Poverty Lines in Theory and Practice*. LSMS Working Paper 133. Washington, D. C.: Svetová banka, 1998. ISBN. 0-8213-4226-6.

[7] DUCLOS, J. Y., ARAAR, A.: *Poverty and Equity: Measurement, Policy and Estimation with DAD*. New York: Springer, 2006. ISBN 978-0387-33318-2.

[8] ŠÚ SR: *EU SILC 2006, UDB verzia 15/05/2007*.

[9] ŠÚ SR: *EU SILC 2006: Zisťovanie o príjmoch a životných podmienkach domácností v SR*. Bratislava: Štatistický úrad SR, 2007.

Adresy autorov:

Tomáš Želinský, Ing.
TU Košice, Ekonomická fakulta
Letná 9, 040 01 Košice
tomas.zelinsky@tuke.sk

Oto Hudec, doc. RNDr. CSc.
TU Košice, Ekonomická fakulta
Letná 9, 040 01 Košice
oto.hudec@tuke.sk

Príspevok bol napísaný s podporou Vedeckej grantovej agentúry MŠ SR a SAV v rámci riešenia vedecko-výskumného projektu *VEGA 1/0370/08 Regionálny prístup k meraniu sociálneho kapitálu a chudoby.*

Relation between GDP and Total Public Expenditure on Education
Vztah mezi HDP a celkovými veřejnými výdeji na vzdělání

Bohdan Linda, Jana Kubanová

Abstrakt: Hlavním cílem tohoto článku je naleznout vztah mezi ekonomickou úrovní krajiny a výdaji, které společnost je ochotná investovat do vzdělávacího systému.

Klíčové slova: HDP, celkové veřejné výdaje na vzdělání, koeficient korelace, resampling, bootstrap replication, bootstrapový odhad

Abstract: The main target of the paper is searching a relation between the economic level of the country and expenditure that is the society willing to invest into educational system.

Key words: gross domestic product, total public expenditure on education, correlation coefficient, resampling, bootstrap replication, bootstrap estimate

1. Introduction

Very often target of statistical investigation is to determine any relation of two or more random variables. This relation is expressed by the population correlation coefficient ρ. We don´t usually know the distribution of random variables, so that we have to try to estimate this value of ρ. The characteristics, that can help us to estimate parameter ρ and determine the degree of linear dependence, is the sample correlation coefficient r. Application of this estimate is usually not very simple, especially when the basic assumption are not granted, when probability distribution of population is unknown and when it is impossible to estimate it. The certain possibility, how to consider about the accuracy of the estimate, about confidence intervals for correlation coefficient or hypotheses testing provide bootstrap method.

2. Interpretation of solved problem

To demonstrate the very problem, we searched relation between basic economical indicator - gross domestic product[5] and total public expenditure on education[6]. The data are from Eurostat [4], from the year 2005, 23 European countries were analysed (table 1). The calculated value of the estimate of correlation coefficient is: $r = 0.4184$.

The question that should be answered is how well is ρ estimated by the value r. This question can be answered without complications in the case of the normal distribution. We don´t have guaranteed a normal distribution of population so that the bootstrap method is used to estimate the bias and standard error of transformed correlation coefficient.

---

[5] GDP (gross domestic product) is an indicator for a nation´s economic situation. It reflects the total value of all goods and services produced less the value of goods and services used for intermediate consumption in their production. Expressing GDP in PPS (purchasing power standards) eliminates differences in price levels between countries, and calculations on a per head basis allows for the comparison of economies significantly different in absolute size.

[6] Total public expenditure on education: Generally, the public sector funds education either by bearing directly the current and capital expenses of educational institutions (direct expenditure for educational institutions) or by supporting students and their families with scholarships and public loans as well as by transferring public subsidies for educational activities to private firms or non-profit organisations (transfers to private households and firms); both types of transactions together are reported as total public expenditure on education.

3. Bootstrap process

We have a sample S of n = 23 pairs of independent bivariate observations (x_i,y_i) from a bivariate population. The correlation coefficient estimate r was computed. The bootstrap process will start with this sample S of 23 observations. B bootstrap replications of the original sample S are created.

Table 6: Data for analysis-Ed2005 are total public expenditure on education, Gdp2005 is gross domestic product

country	*Ed2005*	*Gdp2005*	*country*	*Ed2005*	*Gdp2005*
Bulgaria	2756.9	7900	Lithuania	2013.9	12000
Czech Republic	7464.4	17100	Hungary	7913	14400
Denmark	12736.9	28300	Malta	477.1	17500
Germany	96326.7	25800	Netherlands	24883.7	29600
Estonia	924.6	13900	Austria	12915	28700
Ireland	6365.7	32299	Poland	23948.9	11500
Greece	9514.4	21400	Portugal	9620.8	16900
Spain	42354.7	23100	Slovakia	2813.2	13600
France	89365.5	25200	Finland	8535.5	25800
Italy	61199.5	23600	Sweden	17453.6	27700
Cyprus	1088.6	20800	United Kingd.	85005.5	27100
Latvia	1300.2	11200			

By repetition of the bootstrap process we obtain many new samples and compute r^* and t^*. The distribution of T can be approximated by the distribution of the bootstrapped variable T^*, where the values of the random variable T^* are calculated according to the formula $t^* = \frac{1}{2}\ln\frac{1+r^*}{1-r^*}$. To investigate the properties of the estimator, we did 3000 bootstrap samples, calculated 3000 values r^* and t^*. This process was repeated ten times, so we have 10 series, each has 3000 replications.

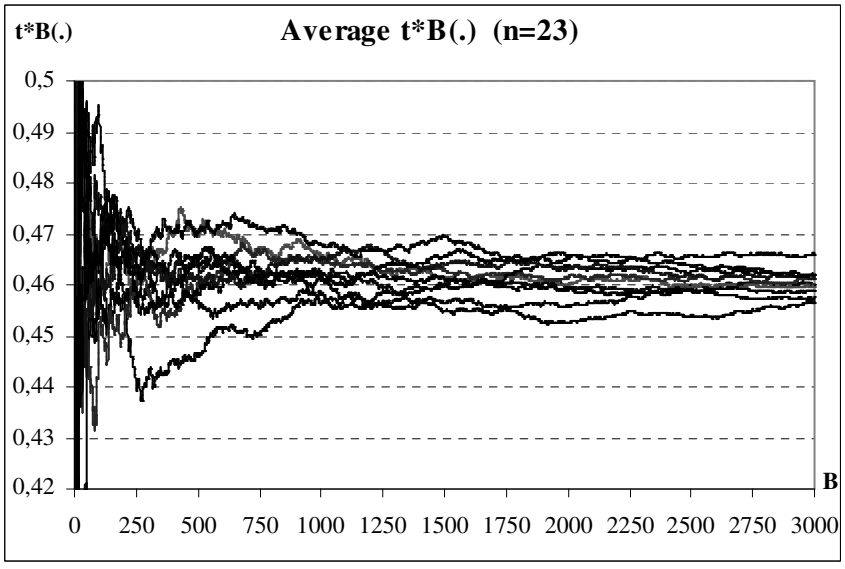


Figure 1: Development of the average of t* values for 1 to 3000 bootstrap replications

We can see the development of the average values of the parameter t^* after 3000 bootstrap replications in the figure 1. We can see that the values don´t differ any more after 1000 bootstrap replications. Convergence of the values t^* to the value 0,46 is evident as well. The values of average of t^* after 3000 bootstrap replications are stated in the second row of the table 2. This process was repeated 10 times.

Values of bias of t^* and standard error of t^* are stated in the third and four row in the table 2. We can see that values of bias minimally differ in individual series of bootstrap, the average value for bias $t^*_{(3000)}$ is 0,0148. The similar conclusion can be stated about the values of standard error, the average value for SE $t^*_{(3000)}$ is 0,1472.

Table 2: Results of ten series of bootstrap

Serie boot	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
average t^*	0,4600	0,4598	0,4588	0,4576	0,4613	0,4619	0,4567	0,4613	0,4620	0,4661
Bias t^*	0,0143	0,0140	0,0131	0,0139	0,0156	0,0162	0,0150	0,0155	0,0163	0,0144
SE t^*	0,1498	0,1475	0,1441	0,1469	0,1449	0,1465	0,1465	0,1492	0,1471	0,1496

Development of bias and standard error depending on the number of bootstrap replications is illustrated in the figure 2. We can see that the values of bias don´t change for B=1800, the values of standard error don´t change for B=1125 (B is the number of bootstrap replications).

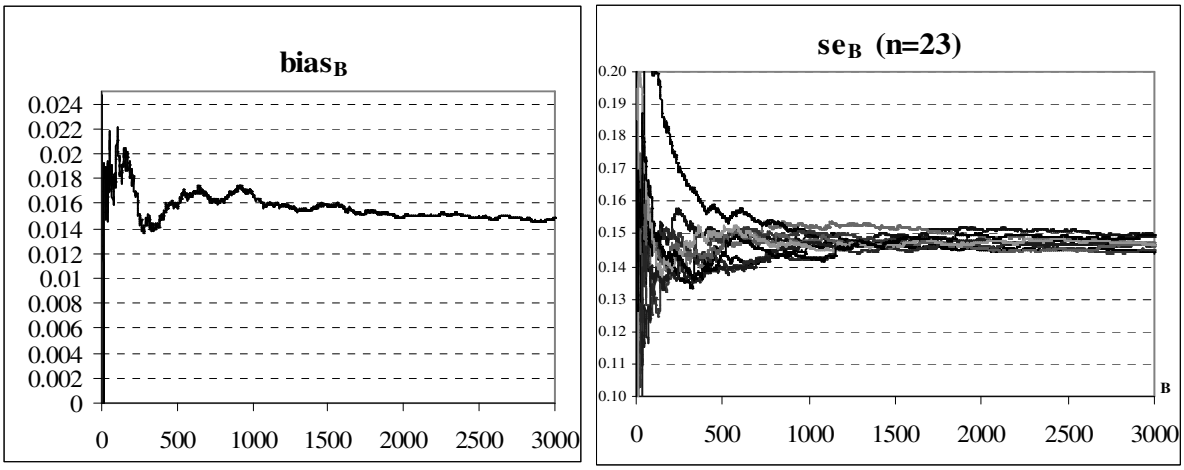


Figure 2: Bias and standard deviation of the bootstrap estimate of parameter t

Very important question that should be solved is the shape of probability distribution of the estimator T, respectively R.

Two histograms were constructed from the bootstrapped values of t^* and r^*. Two matters of interest are visible for the first sight.

a) Both histograms impress that both random variable T and random variable R have an approximate normal distribution. This statement is supported by the result of the χ^2 goodness of fit test. We did 3000 replications, and according to the central limit theorem the random variable T has an approximate normal distribution.

b) Both histograms are very similar. This similarity appears only for small values of R, because values of T and R are very similar for values from 0 till 0,6. On the base of stated results we could construct the confidence intervals in a known way, as if the assumption of normality is true.

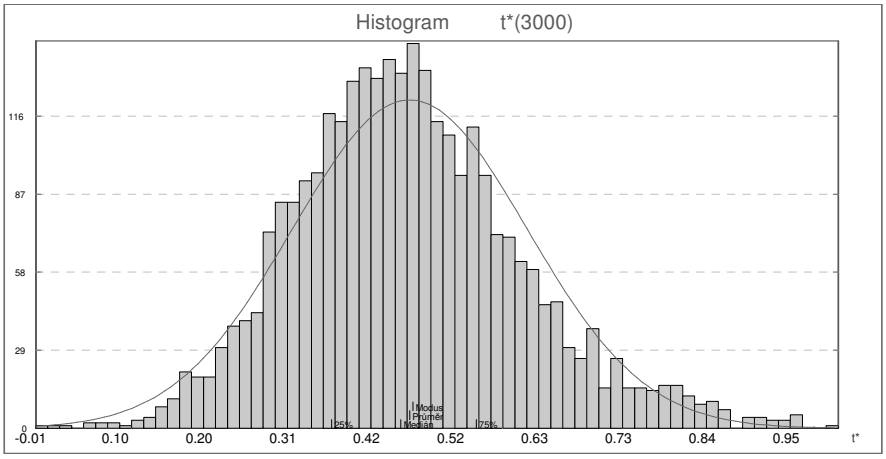


Figure 3: Histogram of values t*

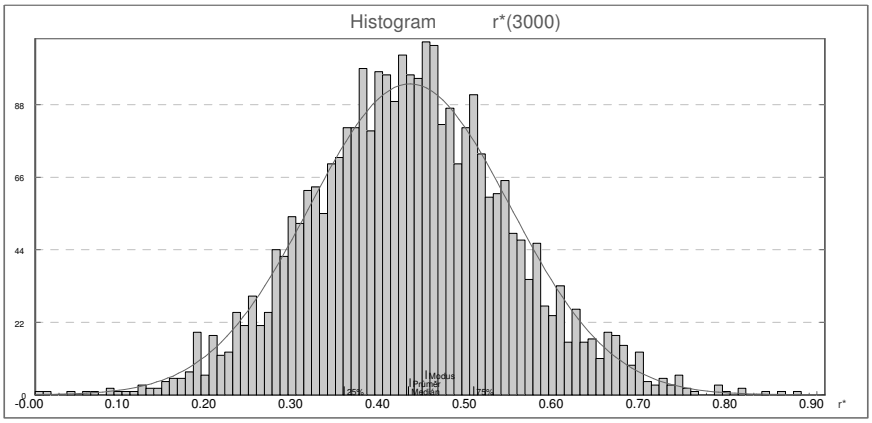


Figure 4: Histogram of values r*

The last task was to construct the confidence interval. If we try to calculate the 95% confidence interval for the parameter ρ on the assumption normality, we obtain interval (0,007; 0,708). On the other hand, the 95% bootstrap quantile confidence interval for the parameter ρ is (0,203; 0,655).

References

[1.] Efron, B., Tibshirany, R.: An Introduction to the Bootstrap. Chapman &Hall, New York, 1993

[2.] Linda, B., Kubanová, J.: Comparison of the Estimates of Correlation Coefficient by Bootstrap and Classical Methods. 2008 International Conference on Applied Probability and Statistics (CAPS 2008), Hanoi 2008

[3.] Stankovičová, I., Vojtková, M.: Viacrozmerné štatistické metódy s aplikáciami. Bratislava. IURA EDITION 2007, p. 261. ISBN 978-80-8078-152-1

[4.] http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1996,45323734&_dad=portal&_schema=PORTAL&screen=welcomeref&open=/&product=REF_TB_national_accounts&depth=3

Addresses of authors:

Jana Kubanová, doc., PaedDr, CSc.
Studentská 95, 532 10 Pardubice
jana.kubanova@upce.cz

Bohdan Linda, doc., RNDr., CSc.
Studentská 95, 532 10 Pardubice
bohdan.linda@upce.cz

Příspěvek k charakterizaci rozdělení příjmů ve vybraných regionech ČR
Article on Characteristics of Income Distribution in Some Regions in the Czech Republic

Jitka Bartošová

Abstract: The submitted article deals with characteristics of present income distribution of Czech households in some regions. The description of the level, differentiation and shape of the distribution is conducted based on simple, and usual as well as more complex, and less usual types of sample characteristics. The article also provides a description of the annual survey Household Income and Living Condition (program EU – SILC) and a brief description of the present regional distribution in the Czech Republic.

Keywords: EU – SILC, regional distribution, income distribution, sample characteristics

Klíčová slova: EU – SILC, regionální uspořádání, rozdělení příjmů, výběrová charakteristika

1. Úvod

Příjmy domácností řadíme mezi základní statistické ukazatele, které nám poskytují informace o životní úrovni obyvatelstva. Síla ekonomiky se často porovnává pomocí velikosti HDP, deficitu/přebytku obchodní bilance apod. O skutečném bohatství občanů, žijících společně v jednom státu (regionu, společenství), však více vypovídá výše peněžních příjmů domácností. Tento fakt velice dobře pociťují právě občané Česka, kde příjmová konvergence k úrovni zemí západní Evropy zaostává za přibližováním ekonomického výkonu (HDP) i cenové hladiny. Neméně důležitá je také struktura rozdělení příjmů mezi jednotlivé sociální skupiny, kraje, rodiny s nezaopatřenými dětmi apod. Zajímavé jsou také informace o vlivu vzdělání, věku, pohlaví atd. na příjem domácnosti.

K vystižení důležitých rysů rozdělení příjmů obyvatelstva slouží souhrnné charakteristiky, které nám podávají informaci o úrovni, diferenciaci i o nerovnoměrnosti rozdělení bohatství mezi občany státu. Vzhledem k tomu, že rozdělení příjmů bývá mnohdy značně komplikované, snažíme se je také aproximovat vhodnými teoretickými modely. Nalezením takových modelů získáváme nástroj, který nás nejenom rychle a přehledně informuje o celkové distribuci bohatství mezi občany, ale který slouží i k dalším hloubkovým analýzám současného stavu a k predikcím do budoucna. Pro státní správu mohou být takovéto modely podkladem pro nastavení parametrů daňového zatížení občanů nebo určování výše sociálních dávek a podpor. Pro ziskové i neziskové organizace soukromého sektoru zase znamenají podporu při zavádění výrobků speciálně zacílených na vybrané skupiny obyvatel či expanzi do nových regionů.

2. Datová základna

Po vstupu do Evropské unie v roce 2004 harmonizovala ČR legislativu v oblasti statistiky s příslušnými zákony EU a nahradila dosavadní nepravidelné šetření (Mikrocensus) každoročním zjišťováním příjmů a životních podmínek EU – SILC (European Union – Statistics on Income and Living Conditions), které bylo poprvé provedeno v roce 2005. Hlavním rozdílem oproti dříve prováděným šetřením metodikou Mikrocensů je vedle menšího vzorku domácností především větší detailnost zjišťovaných informací a zpracování výstupů za jednotlivce. Výhodou je jednak velmi podobná metodika výběru domácností, která umožňuje následnou korekci dat. Veškeré přepočty se provádějí dle metod používaných u

Mikrocensů, a to včetně odhadů podcenění příjmů a eliminace chybějících dat. O tom, jak důležité je provedení korekce sebraných údajů, se můžeme přesvědčit vyhodnocením nekorigovaných dat.

SILC 2005 byl proveden na vzorku 7000 bytů, tj. asi 0,16% z celkového počtu všech obydlených bytů v ČR. Z tohoto počtu se ukázalo 354 jednotek jako neobydlených, příp. adresa nebyla nalezena nebo nebyla dostupná. Z výsledků výběrového šetření SILC 2005 vyplývá, že struktura neúspěšných odpovědí je v podstatě stejná jako při posledním mikrocensním šetření, které bylo provedeno v roce 2002. Stejně jako u tohoto šetření se opět lišila také úspěšnost tazatelů v jednotlivých krajích. Nejmenší úspěšnosti bylo dosaženo v Praze (51,1%), největší v Moravskoslezském kraji (73,9%) (viz tabulka 1). To ukazuje bohužel na ještě nižší celkovou úspěšnost něž v roce 2002. Především výsledky z Prahy jsou už povážlivě nízké, protože se snižujícím se počtem vyšetřených domácností pochopitelně kvalita výběrových dat a tím i následných analýz, na těchto datech provedených, klesá.

Tabulka 1: Úspěšnost sběru dat v jednotlivých krajích, SILC 2005 (Zdroj: www.czso.cz)

Kraj	HD v šetření	z toho vyšetřeno		Kraj	HD v šetření	z toho vyšetřeno	
		počet	%			počet	%
Hl. m. Praha	917	469	51,1	Královéhradecký	364	229	62,9
Středočeský	721	459	63,7	Pardubický	304	207	68,1
Jihočeský	396	249	62,9	Vysočina	317	233	73,5
Plzeňský	375	275	73,3	Jihomoravský	708	425	60,0
Karlovarský	193	118	61,1	Olomoucký	414	308	74,4
Ústecký	560	362	64,6	Zlínský	358	241	67,3
Liberecký	272	174	64,0	Moravskoslezský	815	602	73,9

3. Členění výběrového souboru do skupin podle krajů

Česká republika se od roku 2000 územně člení na čtrnáct vyšších samosprávných celků, tzv. krajů. Toto členění je odlišné od předchozího krajského uspořádání, které zahrnovalo osm oblastí:

- Středočeský kraj se sídlem v Praze,
- Jihočeský kraj se sídlem v Českých Budějovicích,
- Západočeský kraj se sídlem v Plzni,
- Severočeský kraj se sídlem v Ústí nad Labem,
- Východočeský kraj se sídlem v Hradci Králové,
- Jihomoravský kraj se sídlem v Brně,
- Severomoravský kraj se sídlem v Ostravě,
- samostatnou jednotkou v mnoha ohledech postavenou na stejnou úroveň bylo území hlavního města Prahy.

Hlavním důvodem k tomuto kroku bylo „uspokojení“ požadavků některých větších měst, ve kterých zatím krajské sídlo nebylo (např. Liberec, Pardubice, Jihlava). Tímto zákonem tak vzniklo čtrnáct krajů s velmi různou velikostí a u některých z nich také s velice problematicky vymezeným územím, jako je tomu např. v případě Královéhradeckého, Pardubického, či Olomouckého kraje. Zde krajské sídlo rozhodně netvoří přirozené epicentrum, takže obyvatelé ze vzdálenějších obcí jsou nuceni při vyřizování svých úředních povinností často daleko cestovat. Dalším dosud nedořešeným problémem jsou v některých krajích chybějící krajské státní zastupitelství a krajské soudy.

I přes tyto objektivní problémy se krajské uspořádání v České republice za sedm let od jeho vzniku zavedlo poměrně hladce a nově vzniklé vyšší samosprávné celky jsou dnes samozřejmou součástí samosprávné struktury. Jednotlivé kraje mají velmi různou ekonomickou úroveň, a tím i rozdílnou výši příjmů obyvatel. Svou roli v tom hraje geografická poloha kraje, jeho napojení na dálniční tahy do západoevropských zemí nebo také podíl zemědělství, průmyslu a služeb na tvorbě HDP.

Vzhledem k velkému počtu krajů byly ke statistickému zpracování vybrány pouze tři kraje. Kritériem výběru byla velikost kraje. Jedná se kraj Olomoucký, který je zástupcem krajů se střední velikostí, dále pak kraj Moravskoslezský, který se řadí do skupiny krajů velkých (s více než milionem obyvatel) a konečně o kraj Praha, který představuje co do příjmů obyvatelstva speciální případ.

4. Charakteristiky příjmů

Analýzami a modelováním stavu rozdělení příjmů a platů v Čechách (a na Slovensku) po roce 1989 se zabývá řada autorů (viz např. Bartošová (2007, 2006a,b), Pacáková – Sipková – Sodomová (2005), Sipková (2005)), Marek – Vrabec (2008)). Vzhledem k tomu, že příjmy i platy závisí na mnoha faktorech, můžeme pro jejich vyhodnocení použít také metody vícerozměrné analýzy (viz např. Stankovičová – Vojtková (2007)).

Základ kvantitativního popisu empirického rozdělení příjmů tvoří metody deskriptivní statistiky. Souhrnné charakteristiky, které používáme k popisu rozdělení, mohou vycházet

- z úplné množiny hodnot náhodné proměnné (momentové charakteristiky),
- z některých význačných hodnot (kvantilové charakteristiky).

Pro vystižení úrovně příjmů byly v této práci použity především jednoduché charakteristiky, jako je prostý aritmetický průměr a dále pak medián a dolní a horní kvartily, tedy hodnoty v 50%, 25% a 75% souboru. Ze složitějších ukazatelů polohy byl vybrán odhad BES. Tento souhrnný ukazatel využívá poslední tři uvedené kvantilové charakteristiky a je definován vztahem

$$BES = 0{,}25\tilde{x}_{0{,}25} + 0{,}5\tilde{x} + 0{,}25\tilde{x}_{0{,}75}, \qquad (1)$$

kde $\tilde{x}_{0{,}75}, \tilde{x}_{0{,}25}, \tilde{x}$ jsou hodnoty horního kvartilu, dolního kvartilu a mediánu (viz např. Bartošová (2006b)).

Další velmi důležitou charakteristikou příjmů je jejich variabilita, která vystihuje různost hodnot této zkoumané proměnné. Při nízké variabilitě dat se od sebe hodnoty liší jen málo, čímž stoupají na významu charakteristiky polohy. Pokud je však variabilita souboru vysoká, jejich vypovídací hodnota ztrácí na síle. Hlavními momentovými charakteristikami variability pro vzorek hodnot ze základního souboru je výběrový rozptyl s^2 a z rozptylu odvozená směrodatná odchylka s (viz např. Cyhelský – Kahounová – Hindls (1999)) a dále pak variační koeficient, který vystihuje variabilitu dat relativně – vůči průměrné hodnotě. Pro posouzení variability výběrového souboru, která není zkreslená odlehlými hodnotami (zde především hodnotami v pravé části rozdělení příjmů), byly použity některé kvantilové charakteristiky variability, a to kvartilové rozpětí a poměrná kvartilová odchylka (viz např. Cyhelský – Kahounová – Hindls (1999), Čermák (1993)).

Tvar rozdělení příjmů vystihuje jeho šikmost a špičatost. Šikmost vypovídá o tom, jak vypadá rozdělení „dolní“ poloviny příjmů vůči polovině příjmů vyšších. Pokud zabírá první polovina hodnot větší část variačního rozpětí (a je tedy méně nahuštěna než polovina druhá), vykazuje soubor zápornou šikmost a aritmetický průměr je v takovém případě menší než medián. U příjmových rozděleních je situace opačná. Medián má nižší hodnotu než aritmetický průměr, takže je na první pohled zřejmé, že vyšší příjmy jsou méně nahuštěné a rozdělení vykazuje kladnou šikmost. K vyjádření míry šikmosti se běžně používá známá momentová charakteristika – koeficient šikmosti, který je dán hodnotou třetího normovaného

momentu rozdělení četností. Z kvantilových ukazatelů byl zde použit kvantilový koeficient šikmosti, který je dán vztahem

$$\tau = \frac{\tilde{x}_{0,75} - 2\tilde{x} + \tilde{x}_{0,25}}{\tilde{x}_{0,75} - \tilde{x}_{0,25}}, \tag{2}$$

kde $\tilde{x}_{0,75}, \tilde{x}_{0,25}, \tilde{x}$ jsou hodnoty horního kvartilu, dolního kvartilu a mediánu (viz např. Čermák (1993)).

Co se týče charakterizace špičatosti, rozdělení je tím špičatější, čím jsou hodnoty prostřední velikosti více nahuštěny kolem centra ve srovnání s hodnotami okrajovými. Pro vyjádření špičatosti se velmi často používá koeficient špičatosti, tj. čtvrtý normovaný moment zmenšený o 3. Z kvantilových charakteristik pak byl k vyhodnocení použit známý Moorsův koeficient

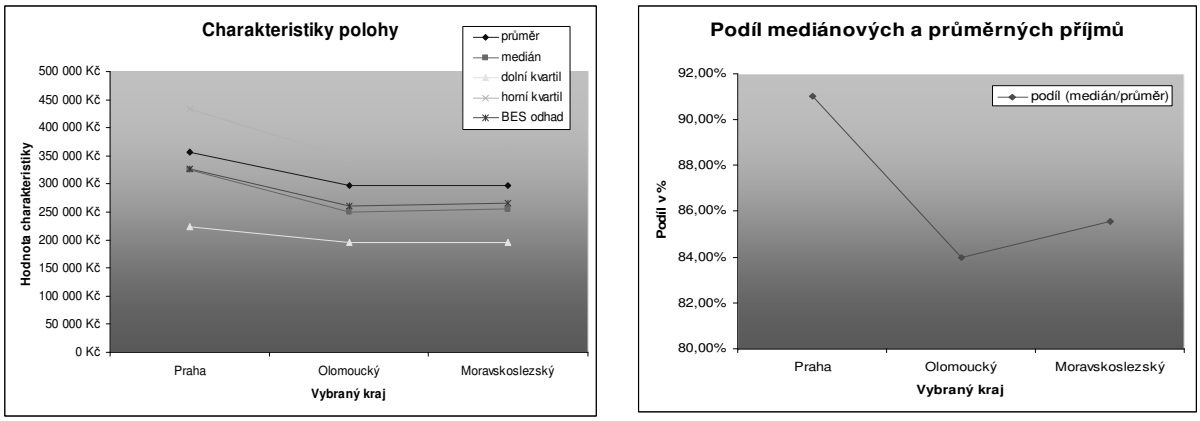
$$b_Q = \frac{\tilde{x}_{0,875} - \tilde{x}_{0,625} + \tilde{x}_{0,375} - \tilde{x}_{0,125}}{\tilde{x}_{0,75} - \tilde{x}_{0,25}}, \tag{3}$$

kde $\tilde{x}_{0,75}, \tilde{x}_{0,25}, \tilde{x}$ jsou hodnoty horního kvartilu, dolního kvartilu a mediánu (viz např. Čermák (1993)).

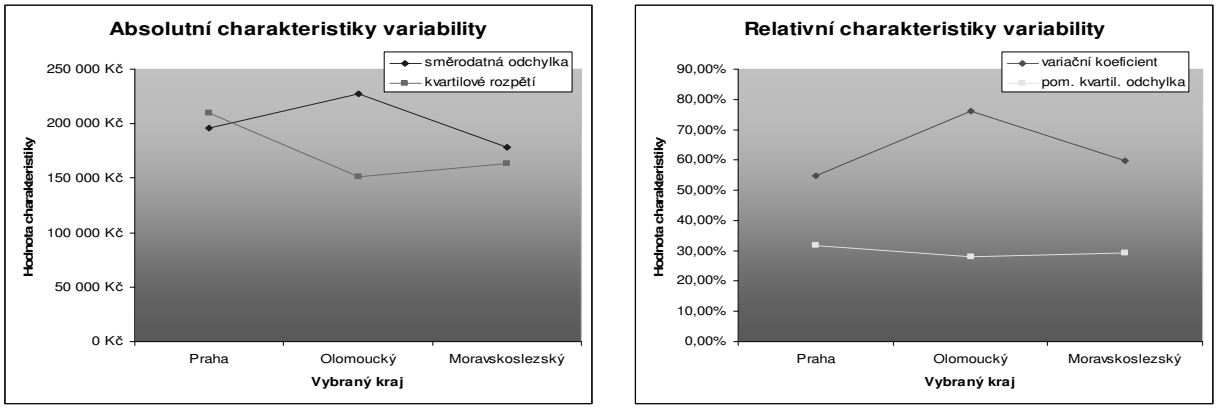
Hodnoty charakteristik rozdělení příjmů ve vybraných krajích ČR v roce 2005 jsou uvedeny v tabulce 2 a získané výsledky jsou pro zvýšení názornosti graficky zobrazeny na obr. 1 – 3.

Tabulka 2: Charakteristiky rozdělení příjmů ve vybraných krajích ČR (Zdroj dat: SILC 2005)

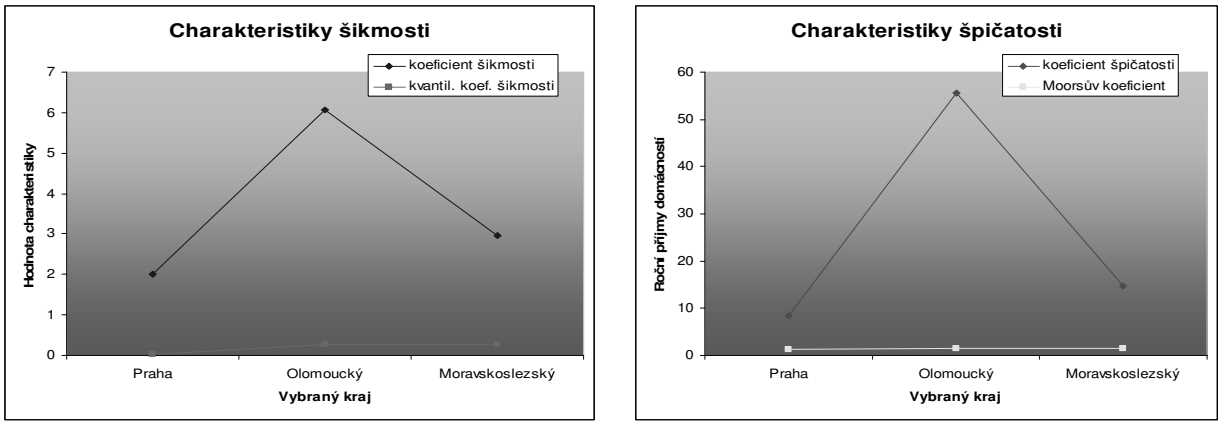
Charakteristika	Praha	Olomoucký	Moravskoslezský
počet domácností	327	207	366
podíl ve vzorku	36,29%	22,97%	40,62%
průměr	357 491	297 816 Kč	297 760 Kč
medián	325 444	250 140 Kč	254 704 Kč
podíl (medián/průměr)	**91,04%**	**83,99%**	**85,54%**
dolní kvartil	223 780	195 589 Kč	195 640 Kč
horní kvartil	433 730	347 000 Kč	358 782 Kč
BES odhad	327 100	260 717 Kč	265 958 Kč
směrodatná odchylka	195 780	226 996 Kč	178 297 Kč
kvartilové rozpětí	209 950	151 411 Kč	163 142 Kč
variační koeficient	**54,8%**	**76,2%**	**59,9%**
pom. kvartil. odchylka	31,9%	27,9%	29,4%
koeficient šikmosti	2,000	6,081	2,958
kvantil. koef. šikmosti	0,032	0,279	0,276
koeficient špičatosti	8,348	55,654	14,760
Moorsův koeficient	1,324	1,371	1,539



Obrázek 1: Porovnání různých typů charakteristik polohy rozdělení příjmů ve vybraných krajích ČR.



Obrázek 2: Porovnání různých typů charakteristik variability rozdělení příjmů ve vybraných krajích ČR.



Obrázek 3: Porovnání různých typů charakteristik tvaru rozdělení příjmů ve vybraných krajích ČR.

4. Závěr

V porovnání s posledním Mikrocensem, realizovaným v roce 2002, se u šetření EU – SILC 2005 snížil zhruba o 10% podíl „vyšetřenosti“ vybraných domácností v Praze (z 61,9% na 51,1%) a tím klesl i podíl tohoto kraje na celkovém vzorku.

Problém vysokého podílu „vyšetřených“ domácností se výrazně odrazil mj. také na hodnotách charakteristik rozdělení. Sice jsme podle očekávání dostali v Praze nejvyšší hodnoty charakteristik polohy, ale řada dalších výsledků odporuje předpokladům o vyšší variabilitě i šikmosti rozdělení příjmů. Zajímavé jsou např. poměry mediánu a průměru, které vyšly v roce 2005 přesně v opačném pořadí než by se dalo očekávat, takže to vypadá, že průměrné a mediánové příjmy jsou nejvíce vyrovnané v Praze (medián zde tvoří 91% průměru). Teprve potom by následovaly kraje Moravskoslezský (s 85,5%) a Olomoucký (s 84%). Tento fakt je zřejmě způsoben tím, že u údajů sebraných v Praze chybí právě údaje za domácnosti s velmi vysokými příjmy, které by měly charakter odlehlých, popřípadě extrémních hodnot. Jejich absence způsobuje snížení hodnot všech momentových charakteristik a vede tak k mylným představám především o variabilitě příjmů v Praze a o šikmosti a špičatosti jejich rozdělení. Těžko lze zdůvodnit např. to, že variační koeficient i koeficient šikmosti a špičatosti jsou v roce 2005 v Praze nižší než v obou sledovaných krajích.

Z grafických výstupů je zřejmé, že nejvyšší variabilita, šikmost a špičatost rozdělení příjmů je v Olomouckém kraji, kde je také největší disproporce mezi průměrnými a mediánovými příjmy. Variační koeficient zde dosahuje dokonce 76%, což je zřejmě zapříčiněno mj. přítomností odlehlých hodnot ve vzorku.

Literatura

[1] BARTOŠOVÁ, J. 2007. Pravděpodobnostní model rozdělení příjmů v České republice. In: ACTA OECONOMICA PRAGENSIA 15, č. 1, Praha, 2007, s. 7 – 12.

[2] BARTOŠOVÁ, J. 2006a. Logarithmic-Normal Model of Income Distribution in the Czech Republic. In: Austrian Journal of Statistics 35, Nr.2&3, Vienna, 2006, s. 215 – 222.

[3] BARTOŠOVÁ, J. 2006b. Volba a aplikace metod analýzy stavu rozdělení příjmů domácností v České republice po roce 1990. Doktorská disertační práce. Praha: Fakulta informatiky a statistiky VŠE, 2006.

[4] CYHELSKÝ, L., KAHOUNOVÁ, J., HINDLS, R. 1999. Elementární statistická analýza. Management Press, Praha, 1999.

[5] ČERMÁK, V. 1993. Diskrétní a spojitá rozdělení – vzorce, grafy, tabulky. VŠE, Praha, 1993.

[6] PACÁKOVÁ, V., SIPKOVÁ, Ľ., SODOMOVÁ, E. 2005. Štatistické modelovanie príjmov domácností v Slovenskej republike. In: Ekonomický časopis 4, č. 53, Bratislava, 2005, s. 427 – 439.

[7] MAREK, L., VRABEC, M. 2008. Regionální rozdíly ve vývoji mezd v ČR. Brno 13.03.2008 – 14.03.2008. In: Firma a konkurenční prostředí 2008 [CD-ROM]. Brno : MZLU, 2008, s. 297–301. ISBN 978-80-7392-020-3.

[8] SIPKOVÁ, Ľ. 2005. Štatistická analýza rozdelenia príjmov domácností v Slovenskej republike. Doktórská dizertačná práca. Bratislava: Fakulta hosdárskej informatiky a štatistiky EU, 2005.

[9] STANKOVIČOVÁ, I., VOJTKOVÁ, M. 2007. Viacrozmerné štatistické metódy s aplikáciami. Bratislava: IURA Edition, 2007. 261 s. ISBN 978-80-8078-152-1

Adresa autora:

Jitka Bartošová, RNDr., Ph.D.
Vysoká škola ekonomická v Praze, Fakulta managementu
Jarošovská 1117/II, 377 01 Jindřichův Hradec, e-mail: bartosov@fm.vse.cz

Prehliadka prác
mladých štatistikov a demografov

Analýza cien bytov na pobreží Čierneho mora
Analysis of prices for flats on the Black Sea-shore

Peter Bialoň

Abstract: The core of my labour is an effort to sketch a picture of price conditions for flats and their relations to different variables in area called Vantage Heights, in Bulgaria. Vantage Heights is project in progress, which is planned to be completed during spring 2009. This project comprises complex of buildings which main purpose will be to provide residential and social premises both for the locals and mainly for foreign investors. Convenience location, complex is situated only 20km for international airport Varna, promises this project broad demand among persons concerned. Therefore, my intention is to provide potential investors with handy information referring to floorage, prices for square meter, district, count of rooms, floor level and how they all together influence the price for accommodation unit. This heading will be properly described, where for base or origin will be taken the up-mentioned price.

Key words: descriptive statistics, distribution analysis, analysis of variance (ANOVA), correlation analysis, linear regression model

Kľúčové slová: popisné štatistiky, distribučná analýza, analýza rozptylu (ANOVA), korelačná analýza, lineárny regresný model

1. Úvod

V príspevku budeme analyzovať bytový projekt, ktorý má byť predbežne dokončený na jar roku 2009. Jedná sa o výstavbu ubytovacieho komplexu s polohou na pobreží Čierneho mora v Bulharsku. Úlohou bude skúmať prevažne jeden štatistický znak, a to cenu bytov u 190 bytových jednotiek v oblasti Vantage Heights. Celý objekt sa rozprestiera na pomerne rozľahlom území, a preto jednotlivé byty alebo jednotky majú rozličné atribúty. Stavba je tvorená niekoľkými budovami, z pomedzi ktorých hlavnou je budova s označením "M". Ďalej sa tu nachádza desať samostatných budov, ktoré majú podobu oddelených víl, pričom každá obsahuje dvanásť až osemnásť apartmánov alebo bytov. Vychádzajúc z poskytnutých dát, môžeme hovoriť o diverzifikácii bytov v týchto znakoch: lokalita, počet miestností, rozloha bytu, cena za meter štvorcový, poschodie, celková cena, prípadne dostupnosť danej jednotky. Skúmať budeme jednotlivé štatistické znaky celého súboru bytov. Začneme jednoduchou popisnou analýzou konkrétnych znakov s ich interpretáciou. Nasledovať bude analýza rozptylu tzv. ANOVA. Východiskovým štatistickým znakom však bude cena, o ktorej sa budem snažiť dokázať, že koreluje teda závisí od iných faktorov. Od akých ostatných znakov a v akej miera cena bytu závisí, sa budeme usiloval zistiť prostredníctvom modelov lineárnej regresie. Výpočty sme urobili v štatistickom softvére SAS Entreprise Guide.

2. Analýza údajov

Lokalita: Rozlišujeme jedenásť základný lokalít s označením "A" až "J" a špecifickú lokalitu "M", ktorá sa nachádza v srdci komplexu. Existuje tu podoba, pretože lokality A, B, C, E, G, H, I ponímajú rovnaký počet bytov. Na druhú stranu lokalita M ponúka 32 bytových priestorov a doplnkovou množinou sú lokality D, F, J, ktoré tvoria 20 bytové zoskupenia.

Čo sa priemerných cien bytov v jednotlivých lokalitách týka, zistili sme, že priemerná cena bytu je totožná v lokalitách A, B, C, E, G, H, I a jej výška je 80.512,30 €, pričom

priemerná rozloha týchto bytov v spomínaných lokalitách bola opätovne rovnaká (66,695 m^2). V zostávajúcich lokalitách, a to D, F, J boli priemerné ceny bytov tiež rovnaké vo výške 56.374,91 € s priemernou celkovou rozlohou 46,6855 m^2. Výrazne sa však odlišovala priemerná cena bytu v strede komplexu vo vile M, kde jej hodnota bola 100.441,08 € pri priemernej rozlohe 82,3603 m^2.

Počet miestností: Odlišnosť bytových jednotiek sa týka aj počtu miestností, ktoré byt tvoria. Vo všeobecnosti môžeme tvrdiť, že existuje typológia bytov a to: 1-izbové, 2-jizbové a 3-izbové. Z frekvenčnej analýzy môžeme vyvodiť, že jednoizbových bytov je 37, čo predstavuje 30% komplexu. 2-izbových bytov v celom areáli je najviac, a to 92. Tento údaj predstavuje niečo viac ako 48% percent bytov. Nakoniec sú to 3-izbové byty, ktorých je naopak najmenej, a to 41 s podielom okolo 22%.

Rozloha bytu: Snáď najviac diverzifikovanou štatistikou je rozloha, ktorá sa odlišuje u väčšiny bytových jednotiek. Zavádzame pojem celková rozloha resp. výmer, ktorý je tvorený dvoma zložkami, obytnou plochou bytu a spoločenskými priestormi, využívanými jednak vlastníkom bytu, ale aj ostatnými obyvateľmi. Priemerná rozloha bytu vo Vantage Heights sa pohybuje okolo 63 m^2 a extrémami v tejto oblasti sú byty s rozlohami 31,26 m^2 a 164,58 m^2.

Cena za m^2: Táto cena predstavuje finančné ohodnotenie jednotky výmeru bytu, pričom pri jej skúmaní sme prišli k záveru, že rozlišujeme tri typy cien za meter štvorcový. Tieto ceny sú vo výške 1160 €, 1200 € a 1300 € za jednotku rozlohy. Z frekvenčnej analýzy sme zistili, že práve 52 bytov je s cenou 1160€/m^2, 112 bytov je za 1200€/m^2 a nakoniec 26, čiže najmenej, je bytov, ktoré sa predávajú za 1300€/m^2. Pretože tieto ceny sú v každej časti komplexu, či vo vilách, alebo tzv. M lokalite rôzne, táto variabilita ceny jednoznačne nezávisí len od umiestnenia bytu v komplexe, ale treba zohľadniť aj ostatné faktory.

Umiestnenie bytu na poschodí: Architekti daného komplexu považovali za najvyhľadávanejšie práve prvé a druhé poschodia obytných budov, preto je najväčší objem bytov práve na týchto poschodiach (56 pre každé poschodie). Naopak najmenšej obľube sa tešia suterénne príbytky, ktorých bude vyhotovený iba limitovaný počet – len 3 byty z celého 190-bytového komplexu. Podkrovné byty, ktoré sú taktiež navrhované iba v malom množstve (6 bytových jednotiek). Nemožno ich však so suterénnymi bytmi porovnávať, pretože ich môžeme označiť za lukratívny artikel predaja, preto je aj ich cena vyššia a sú k dispozícii iba v centrálnej lokalite "M".

Celková cena: Finálna cena bytu je jedným z variabilných údajov už aj preto, že vychádza z celkového výmeru bytu, ktorý sa u jednotlivých jednotiek rôzni. Vychádzajúc z popisných štatistík môžem skonštatovať, že priemerná cena bytovej jednotky v celom komplexe Vantage Heights sa pohybuje okolo 76.246,20 €. Ďalej, variabilita ceny okolo priemernej ceny sa pohybuje v rozmedzí ±31.161,60 €, pričom najlacnejší byt v tomto prímorskom letovisku sa dá zaobstarať za 36.231,60 €, a naopak najdrahší byt sa dá kúpiť za 213.954,00 €. Tieto štatistiky len poukazujú pestrosť výberu, a teda na svoje si prídu menej i viac solventní zákazníci. Uvedené tvrdenie sa dá podložiť i ďalšími ukazovateľmi. Napr. mediánom, ktorý nám rozdeľuje súbor bytov na dve rovnaké skupiny, a to byty drahšie ako 74.412,00 € a byty lacnejšie. Nasledovne z kvartilov sa dá vyvodiť, že 25% bytov je lacnejších ako 47.508,00 € a 25% bytov je drahších ako 97.708,00 €. V nižšie uvedenej tabuľke sú zobrazené dôležité cenové štatistiky bytov v jednotlivých lokalitách komplexu (tabuľka 1).

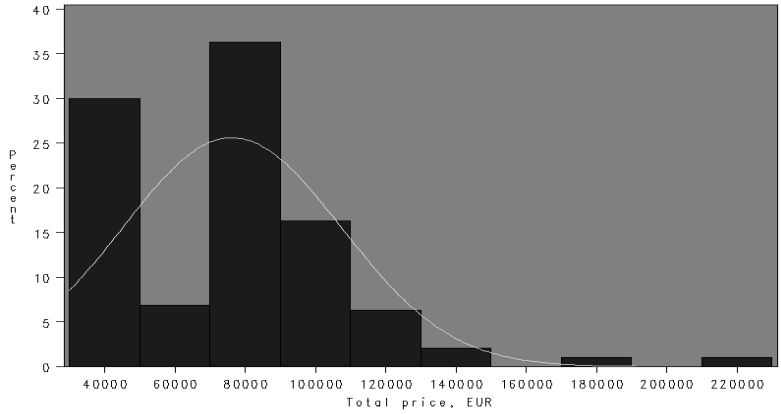
Na základe testovania hypotézy, či cena bytov má normálne rozdelenie, môžeme tvrdiť, že tento štatistický znak, nemá normálne rozdelenie. Tento záver potvrdil aj grafický výstup (obrázok 1). Všetky p-hodnoty použitých testov (Shapiro-Wilkov, Kolmogoro-Smirnov, Cramerov, Anderson-Darlingov) boli nižšie ako hladina významnosti 5%.

Pomocou analýzy rozptylu sme sa pokúsili rozhodnúť, či priemerné ceny bytov v lokalitách sú približne rovnaké. Pre použitie parametrickej ANOVA metódy musia byť

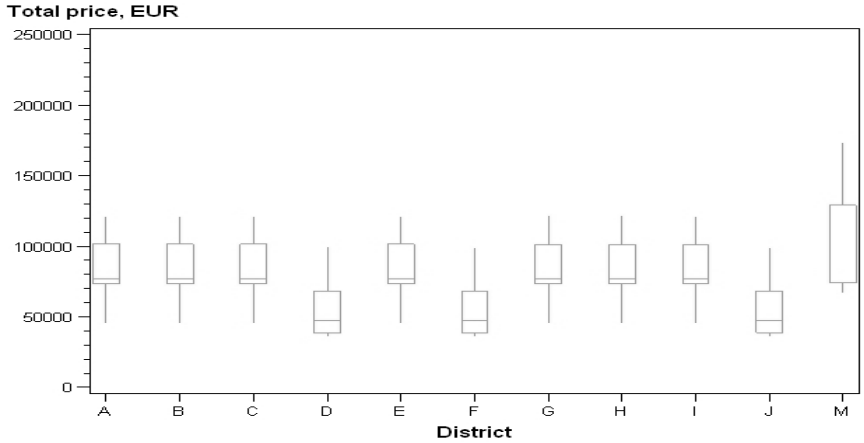
splnené tri podmienky a to *normalita ceny, nezávislosť súborov a homogenita rozptylov.* Nakoľko sme už zistili, že ceny nemajú normálne rozdelenie, je nutné použiť neparametrickú ANOVA metódu. Použitím testov (Kruskal-Wallisov test, mediánové skóre) sme zistili, že priemerná cena bytov v jednotlivých lokalitách komplexu nie je rovnaká. Ani v jednom prípade nenadobudla p-hodnota vyššiu hodnotu ako 0,01.

Tabuľka 1:Vybrané popisné štatistiky cien bytov v jednotlivých lokalitách komplexu

	Priemerná cena	**Štand. odchýlka**	**Min. cena**	**Max. cena**	**Dolný kvartil**	**Medián**	**Horný kvartil**
A	80512.30€	23668.81€	45402.40€	121069€	73230.80€	76626€	101581.0€
B	80512.30€	23668.81€	45402.40€	121069€	73230.80€	76626€	101581.0€
C	80512.30€	23668.81€	45402.40€	121069€	73230.80€	76626€	101581.0€
D	56374.75€	23769.52€	36273.20€	123188€	38769.60€	47263€	68343.8€
E	80512.30€	23668.81€	45402.40€	121069€	73230.80€	76626€	101581.0€
F	56374.75€	23769.52€	36273.20€	123188€	38769.60€	47263€	68343.8€
G	80512.30€	23668.81€	45402.40€	121069€	73230.80€	76626€	101581.0€
H	80512.30€	23668.81€	45402.40€	121069€	73230.80€	76626 €	101581.0€
I	80512.30€	23668.81€	45402.40€	121069€	73230.80€	76626€	101581.0€
J	56374.75€	23769.52€	36273.20€	123188€	38769.60€	47263€	68343.8€
M	100441.0€	42620.71€	66690.00€	213954€	73908.00 €	74496€	128988.0€



Obrázok 1: Histogram cien bytových jednotiek



Obrázok 2: Box ploty pre priemerné ceny bytov v jednotlivých lokalitách komplexu

3. Lineárny regresný model

V tejto časti sme sa pokúsili odhadnúť parametre lineárneho modelu, ktorý má potvrdiť resp. vyvrátiť existenciu závislosti ceny od ostatných faktorov. Budeme skúmať či celková cena závisí od iných atribútov bytu ako sú: rozloha, počet izieb, poloha, poschodie.

V prvom kroku som zhodnotil kvantitatívne premenné. Nasledovne sme chceli zistiť, či medzi premennými existuje závislosť, teda či cena bytu koreluje s ostatnými premennými ako: počet izieb bytu, celkový výmer bytu, obytná plocha atď. Vypočítali sme teda párové koeficienty korelácie (tabuľka 2).

Tabuľka 2:Koeficienty korelácie ceny bytov od ostatných premenných

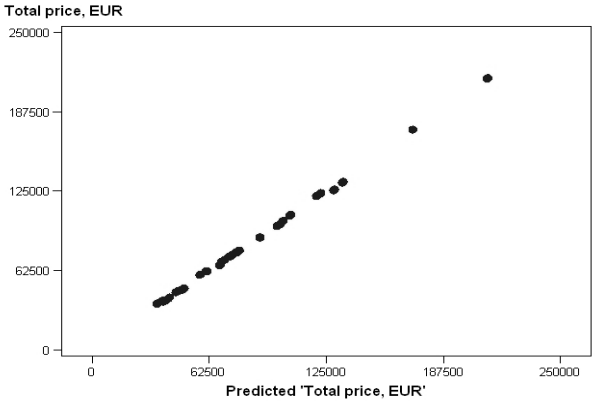
Premenná	**Celková cena bytovej jednotky**		
	Pearsonov koef.	**Spearmanov koef.**	**Kendallov koef.**
Počet izieb (X2)	0.88597	0.92185	0.80057
Celkový výmer (X3)	0.99468	0.99393	0.96443
Obytná plocha	0.99472	0.97472	0.91300
Spoločné priestory	0.95155	0.94167	0.82267
Poschodie (X1)	0.43712	0.38978	0.31451

Z výsledkov v tabuľke 2 jednoznačne vidíme, že medzi cenou a ostatnými premennými existuje v každom prípade korelácia, pričom táto závislosť je značná, až na výnimku poschodia. Najväčší korelačný koeficient je badateľný u celkového výmeru, kde sa pohybuje až nad hranicou 0,96 z čoho pre nás vyplýva, že cena bytu najviac závisí od celkového výmeru bytu. Potrebné je azda ešte dodať, že p- hodnoty všetkých spomínaných koeficientov boli menšie ako 0,0001.

V nasledovnej časti vytvoríme konkrétny model a budeme sa snažiť o jeho čo najväčšiu výpovednú hodnotu. Pri tvorbe budeme testovať tieto premenné: počet izieb, celkový výmer, spoločné priestory, obytné priestory a cenu za m^2 resp. ich dopad na celkovú cenu za byt. Skúšali sme rôzne kombinácie regresných premenných a na základe vysokej hodnoty koeficienta determinácie R^2 a vhodnej Mallowsovej C(p) štatistiky sme vybrali nasledovný model:

$$Y_i = -96714 - 270{,}22X_1 - 2446{,}32X_2 + 1285{,}63X_3 + 80{,}64X_4 + e_i \qquad (1)$$

, kde X_1 - poschodie na ktorom sa obytná jednotka nachádza, X_2 - počet izieb bytu, X_3 - celkový výmer bytu, X_4 - cena bytu za m^2. Model je ako celok významný. Kvalita modelu je dobrá: R^2 = 0,9994 pri štatistike C(p) = 4,9491. Všetky spomínané premenné uvedeného modelu sú významné.



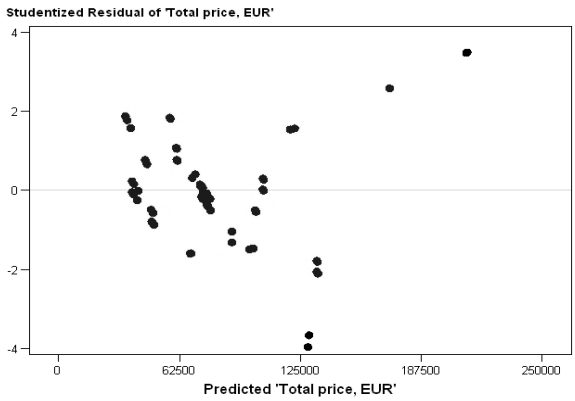
Obrázok 3: Lineárny regresný model závislosti celkovej ceny bytu od ostatných premenných

Na prvý pohľad z modelu vidíme, že priesečník má len lokovaciu povahu, teda určuje polohu priamky. Pokračujúc v interpretácii výsledkov: v prípade, ak by sme uvažovali o kúpe ľubovoľného bytu, tak za akýkoľvek iný byt o poschodie vyššie by sme zaplatili v priemere o 270€ menej. Táto skutočnosť však platí iba v prípade, že všetky ostatné podmienky modelu zostanú nezmenené. Ďalej, v prípade, že sa počet izieb v byte zvýši o jednu, tak nasledovne priemerná cena bytu klesne o 2.446,32 €. Už táto úvaha a interpretácia modelu sa javí nelogická a nepravdepodobná, preto je neskôr nutné model podrobiť ďalším testom. V prípade ak by sa celkový výmer bytu zvýšil o jeden m^2 a ostatné premenné by sa nezmenili, tak by sa cena za bytovú jednotku zvýšila o 1.285,63 €. A nakoniec, ak sa cena za m^2 zvýši o 1 € a všetky ostatné premenné zostanú konštantné (resp. bude dodržaná podmienka ceteris paribus), priemerná cena bytu sa zvýši o 80,64 €. Pre tento model môžeme jasne povedať, že až okolo 98% variability ceny spôsobuje celkový výmer bytu, nasledovne približne 1% variability môžeme pripísať cene za m^2 a ostatné faktory sú v podstate zanedbateľné.

Kvalitu tohto modelu som skúmal z dvoch hľadísk: R-square štatistika a rezíduá. R-square štatistika modelu je vysoká (0,999), a však je nutné zohľadniť rezíduá. Rezíduá musia mať normálne rozdelenie a pomocou filtrovania studentizovaných rezíduí sa pokúsim vylúčiť tzv. outliers . Nový model už neobsahuje štyri mimo ležiace body a má podobu:

$$Y_i = -99671 - 427{,}25X_1 - 1612{,}20X_2 + 1257{,}47X_3 + 83{,}40X_4 + e_i \qquad (2)$$

, pričom jeho interpretácia je podobná ako pri prvom modeli a tento model je možné označiť ako kvalitnejší, lebo nezahŕňa štyri mimo ležiace body, ktoré sú vyznačené na obrázku nižšie.

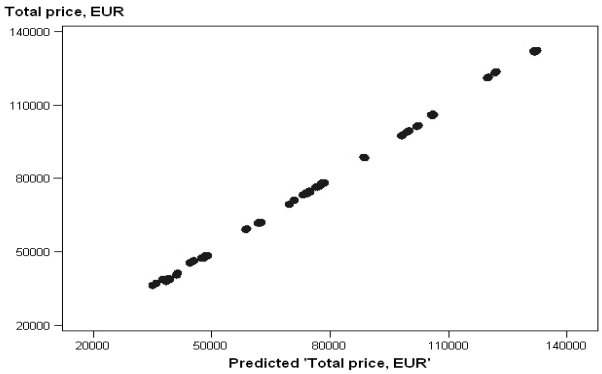


Obrázok 4:Graf rezíduí s vyznačenými mimo ležiacimi bodmi (outliers)

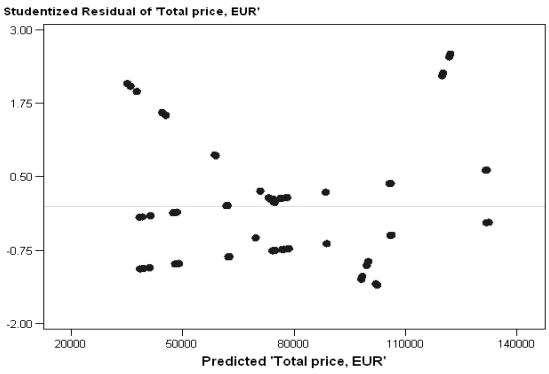
Napriek vyššie uvedenej korektúre vykazuje model stále nedostatky. Tie sa prejavili v prítomnosti kolinearity, teda závislosti medzi premennými v modeli. Variačný inflačný faktor u premennej výmer spoločenských plôch bytu presahuje hodnoty 10,0 , čo sa javí ako problém. Z tohto dôvodu sme sa rozhodoli danú premennú z modelu vylúčiť, a tak sme dostali konečný model, ktorý nadobúda nasledovnú podobu:

$$Y_i=-104.625-497{,}886X_1+1.194{,}25195X_2+88.26036X_3+e_i \qquad (3)$$

,kde X_1-poschodie na ktorom sa obytná jednotka nachádza, X_2-celkový výmer bytu, X_3-cena bytu za m^2. Model je ako celok významný, pričom hladina významnosti bola opäť 5%. Kvalita modelu je v porovnaní s prvý a druhým modelom o niečo vyššia a to: $R^2 = 0{,}9995$ pri štatistike C(p) = 3,4934. Všetky spomínané premenné uvedeného modelu sú opäť významné ako aj priesečník. Kolinearita a rezíduá v tomto modeli sú taktiež v poriadku resp. významné, čo zaisťuje, že výsledky sú neskreslené. Model je zobrazený na obrázku.



Obrázok 5: Konečný lineárny model



Obrázok 6: Rezíduá konečného lineárneho modelu

4. Záver

Príspevok bol zameraný na analýzu cien ponúkaných bytov reálneho projektu v prímorí Bulharska, kde sa plánuje výstavba bytového komplexu, ktorý bude mať charakter malomesta s názvom Vantage Heights. V prvej časti sme poskytli deskriptívne charakteristiky a frekvenčnú analýzu rôznych bytov projektu. Táto časť bola určená na vytvorenie si všeobecného obrazu a mala osloviť najmä záujemcu, ktorý plánuje v danej lokalite investovať.

V druhej časti projektu sme sa zamerali na analýzu cien bytov v danom komplexe, ktorá by parciálne mohla slúžiť pre potenciálnych investorov so širším finančným zázemím. Analyzované ceny bytov boli skúmané z rôznych hľadísk, pričom nechýbala ani ich korelačná analýza. Poukázali sme na faktory, ktoré prevažne vplývajú na cenové ohodnotenie bytov. Z výsledkov, ktoré sme získali lineárnym modelovaním, môžeme konštatovať, že najviac cenu za bytovú jednotku ovplyvňujú faktory ako celkový výmer bytovej jednotky, poschodie, na ktorom sa bytová jednotka nachádza (v zápornom zmysle - čím je byt na vyššom poschodí tým jej jeho cena nižšia) a nakoniec je to cena bytu za m^2.

5. Literatúra

[1]STANKOVIČOVÁ, IVETA Ako robiť štatistiku v systéme SAS In: Výpočtová štatistika. - Bratislava: SŠDS, 2000. - S. 74-78. - ISBN 80-88946-09-03

[2]PACÁKOVÁ, VIERA A KOL.: Štatistika pre ekonómov. Bratislava. IURA Edition. 2003. ISBN 80-89047-74-2

Adresa autora :

Peter Bialoň, študent, 3. ročník,
Fakulta managementu
Univerzita Komenského v Bratislave
Odbojárov 10, 820 05 Bratislava
e-mail: *peter.bialon@post.sk*

Formovanie a rozpad rodiny v priestore miest a vidieka Slovenska
Formation and desintegration of family in the urban and rural space of Slovakia

Katarína Čupeľová

Abstract: This work analyses demographic processes of formation and desintegration of families (nuptality, divorce) in the urban and rural population. These influence the main demographic process - fertility. Analysis is elaborated in level of two subpopulation – urban and rural and in level of municipalities, that are classed into categories according to number of inhabitants. Synthesis of acquired knowledges evaluates demographic behavior of urban and rural population and its dependence on number of inhabitants in municipality. Especially this work evaluates differences in demographic behavior of population in town and country or in small municipalities and big cities.

Key words: family, demographic process, nuptality, divorce, urban and rural population, categories of municipalities according to number of inhabitants, demographic behavior

1. Úvod

Súčasný vývoj v Európe, tiež nazývaný aj druhá demografická revolúcia, priniesol zmeny v správaní ale aj hodnotovom systéme populácie, ktoré vedú k oslabovaniu funkcie manželstva a rodiny a redukujú úroveň plodnosti na úroveň pod záchovnou hodnotou – 2,1 dieťaťa na 1 ženu počas jej reprodukčného obdobia. Aj keď na Slovensku tradične prebiehal proces formovania aj rozpadu rodiny vo vidieckom a mestskom prostredí odlišne, zmeny v rodinnom ale aj reprodukčnom správaní sa nevyhli obom prostrediam.

Cieľom tohto príspevku je poukázať na špecifické črty sobášnosti a rozvodovosti, ktoré v konečnom dôsledku vplývajú na plodnosť populácií vidieka a miest a aj jednotlivých veľkostných kategórií obcí.

Za mestá sme považovali tie obce, ktoré majú priznaný štatút mesta (138 obcí), ostatné predstavujú obce vidieckeho charakteru. Druhým delením boli veľkostné kategórie obcí, ako ich uvádza slovenská štatistika v pravidelných ročných publikáciách o populačných procesoch. Uvedomujeme si, že tieto delenia majú len zmluvný charakter a demografické správanie ich obyvateľstva je tiež potrebné chápať v prechodných formách. Tento príspevok je súčasťou širšieho výskumu správania mestského a vidieckeho obyvateľstva, skúmané obdobie sú roky 2001 – 2003, keďže boli spracúvané aj štruktúrne charakteristiky obyvateľstva (sčítanie v roku 2001), ktoré na charakteristiky populačných procesov priamo vplývajú.

2. Sobášnosť mestského a vidieckeho obyvateľstva

Sobášnosť je osobitný populačný proces, ktorým sa začína formovať rodina a to uzatvorením manželstva. Rodina ako trvalé spoločenstvo partnerov rozličného pohlavia a ich detí má nezastupiteľnú funkciu v procese reprodukcie obyvateľstva najmä v silnej väzbe s jej rozhodujúcou zložkou – pôrodnosťou. Sobáš je akt právneho uzatvorenia manželstva, ktorý nemusí podstúpiť každý člen populácie a charakteristickou črtou posledných rokov je zvyšovanie podielu neformálnych partnerských zväzkov.

Počet uzatvorených manželstiev ovplyvňujú faktory demografické (zloženie populácie podľa pohlavia a veku) a spoločensko – ekonomické (úroveň ekonomického vývoja, jeho

krízové javy, celospoločenská populačná klíma, individuálne predstavy o založení rodiny, o neformálnych partnerských zväzkoch atď.) [3].

K hlavným rysom vývoja sobášnosti na Slovensku patrili po dlhé desaťročia kombinácia vysokej miery sobášnosti a nízkeho sobášneho veku. Obidve charakteristiky sa odvíjali od kresťanských hodnôt a spôsobu života tradičnej vidieckej rodiny. Manželská rodina sa považovala za univerzálny spôsob životnej dráhy, pričom počas socializmu sobášnosť podporovali aj viaceré výhody napr. prideľovanie bytov prevažne rodinám, poskytovanie mladomanželských pôžičiek štátom s možnosťou odpisu za každé narodené dieťa a pod. Mladí sa museli vyrovnať s obmedzenou príležitosťou študovať na stredných a vysokých školách, nemotivujúcim ohodnotením práce, uzatvorenými hranicami, ktoré im bránili vycestovať a nízkou úrovňou antikoncepcie, ktorá často viedla k nechcenému otehotneniu partnerky. Skoré uzatvorenie manželstva bolo prvým a často jediným krokom, ako sa osamostatniť od rodičov.

Nízky vek pri sobáši sa dá vysvetliť aj tzv. teóriou vzniku partnerských zväzkov ekonóma Garyho Beckera. Snaha maximalizovať zisk z manželstva vedie bežne k predĺženiu doby chodenia. Šanca nájsť ekonomicky zdatného partnera bola v dobe reálneho socializmu mizivá, a preto sa mladí ľudia zbytočne nezdržiavali pátraním po výhodnejšej partii a zakladali manželstvo s prvou známosťou [1].

Tabuľka č. 1: Vybrané ukazovatele sobášnosti mestského a vidieckeho obyvateľstva Slovenska (2001 – 2003)

Veľkosť a typy obcí	Hrubá miera sobášnosti (‰)	Nepriamo štandard. hms (‰)	Priemerný vek nevesty	Priemerný vek ženícha	Podiel 1. sobášov
mesto	4,84	-	27,00	29,94	86,39
vidiek	4,38	-	24,74	27,74	92,35
0-199	3,92	4,49	25,03	28,37	92,08
200-499	4,20	4,46	24,73	27,93	92,95
500-999	4,31	4,54	24,71	27,70	92,50
1000-1999	4,45	4,53	24,74	27,77	92,36
2000-4999	4,49	4,54	24,82	27,72	91,99
5000-9999	4,56	4,49	25,70	28,67	89,58
10000-19999	4,68	4,51	26,27	29,14	87,66
20000-49999	4,87	4,71	26,55	29,41	86,82
50000-99999	4,90	4,80	27,02	29,91	86,61
100000+	5,00	4,94	28,67	31,78	83,08
SR	4,64	4,64	26,05	29,02	88,91

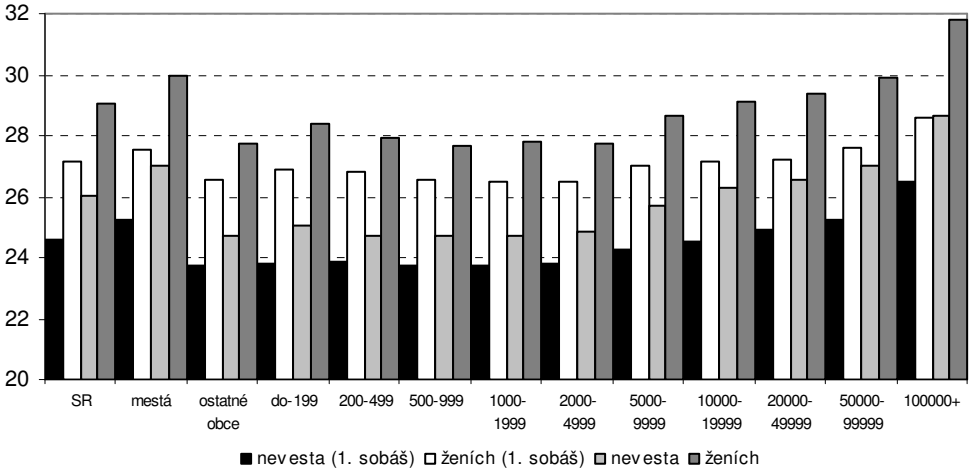
Zdroj: Stav a pohyb obyvateľstva v Slovenskej republike 2001, 2002, 2003, ŠÚ SR; vlastné výpočty

Hrubá miera sobášnosti ukazuje, že intenzita sobášnosti sa s rastúcou veľkosťou obce zvyšuje. No keďže tento ukazovateľ je ovplyvnený vekovou štruktúrou obyvateľstva, je vhodnejšie použiť nepriamo štandardizovanú mieru, ktorej hodnoty nám ukazujú presne opačnú situáciu. Najvyššiu intenzitu sobášnosti má obyvateľstvo väčších miest, zatiaľ čo u vidieckych sídiel je táto intenzita nižšia. To je spôsobené najmä sociálnou a ekonomickou situáciou na vidieku, najmä čo sa týka trhu s bytmi a rodinnými domami, resp. ich finančnej nedostupnosti a neistote nájsť si primerané zamestnanie a uplatniť sa v ňom [4].

Priemerný vek ženíchov a neviest pri prvom sobáši ako pri sobáši vôbec dosahuje vyššie hodnoty v mestách ako na vidieku. Priemerný vek neviest pri prvom sobáši je v obciach pod 5 tis. obyvateľov nižší ako 24 rokov, vek ženíchov sa pohybuje okolo 25 rokov. S rastúcou veľkosťou obce rastie aj vek pri prvom sobáši, najrýchlejšie v súboroch obcí nad 10 tis. obyvateľov. Model skorej sobášnosti je na vidieku ovplyvnený tradíciami, mierou religiozity a celkovým vnímaním týchto spoločností v otázke rodiny a manželstva. Rozdiely

medzi jednotlivými súbormi obcí sa ešte zväčšujú, ak uvažujeme priemerný vek pri sobáši. V najväčších mestách sa tento vek u ženíchov dostal nad 30 rokov. V týchto mestách majú vysoké podiely na uzavretých sobášoch sobáše vyšších poradí, ktoré uzatvárajú obyvatelia spravidla vo vyššom veku. I. Možný vysvetľuje zvyšujúci sa vek pri sobáši vyšším vzdelaním a odbornou kvalifikáciou žien spojenou s profesionálnym úspechom a ekonomickou nezávislosťou [1].

Graf č. 1: Priemerný vek nevesty a ženícha pri prvom sobáši a sobáši u mestského a vidieckeho obyvateľstva Slovenska (2001 – 2003)



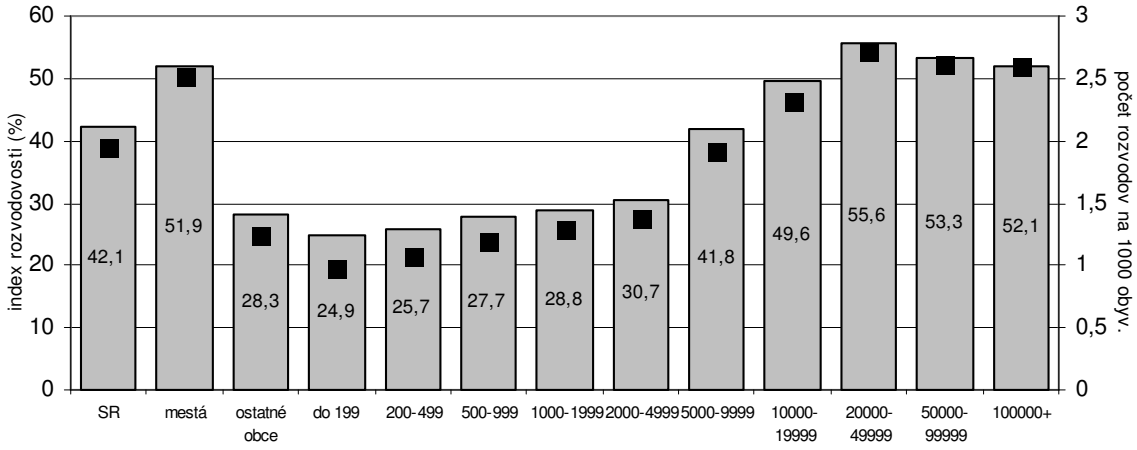
Zdroj: Stav a pohyb obyvateľstva v Slovenskej republike 2001, 2002, 2003, ŠÚ SR; vlastné výpočty

3. Rozvodovosť mestského a vidieckeho obyvateľstva

Rozvod je jedna z foriem rozpadu manželstva, založená na zákonnom spôsobe zrušenia manželstva. Rozvodovosť charakterizuje úroveň populácie z hľadiska stability rodín. Rozvod možno chápať ako významnú udalosť v živote ľudí s množstvom sociálnych dôsledkov, medzi ktoré patrí narušenie funkcie rodiny, ale často i výrazná zmena ekonomického aj sociálneho statusu niektorých členov pôvodnej rodiny, zmena ich spôsobu života, zamestnania a pod. [4,5]. Intenzita rozvodovosti je spojená a celospoločenským vývojom, so zmenami ekonomického postavenia obyvateľov, s novou pozíciou žien v spoločnosti, ale odráža aj niektoré menej riešené otázky ako je bytový problém, nízky sobášny vek a pod. [3].

Všeobecne sa v demogeografii uznáva silná väzba úrovne rozvodovosti s procesmi urbanizácie. Mestské prostredie sa vyznačuje prítomnosťou viacerých faktorov, ktoré pôsobia v smere znižovania stability manželstva, zoslabujú funkciu rodiny. V mestských rodinách sa rodí menej detí, a preto môže byť do značnej miery oslabená integračná funkcia detí v rodine, anonymita prostredia je vyššia, podobne aj možnosti nadväzovania nových partnerských kontaktov. Vysoká rozvodovosť je evidentne jedným z prejavov najvyššej sociálnej mobility, ktorá sa vyznačuje vysokou pracovnou fluktuáciou, častou zmenou bydliska. Pri ľahšej možnosti získania bytu v mestách je potlačený jeden z brzdiacich faktorov rozvodovosti, ktorý sa uplatňuje hlavne vo vidieckych sídlach [5]. Hodnota indexu rozvodovosti nám nehovorí len o rastúcej intenzite rozvodovosti, ale aj o klesajúcej intenzite sobášnosti, keďže dáva do pomeru jednotlivé hodnoty v danom roku a počet rozvodov nevzťahuje k východiskovému počtu sobášov. Hlavne v obciach nad 5000 obyvateľov hodnota tohto ukazovateľa prudko rastie, keďže v týchto obciach sa v dôsledku zvyšujúcej sa preferencie neformálnych partnerských zväzkov, typickej pre mestské prostredie, znižuje počet sobášov.

Graf č. 2: Index rozvodovosti a hrubá miera rozvodovosti mestského a vidieckeho obyvateľstva Slovenska (2001 – 2003)



Zdroj: Stav a pohyb obyvateľstva v Slovenskej republike 2001, 2002, 2003, ŠÚ SR; vlastné výpočty

Počet detí je významným faktorom determinujúcim úroveň rozvodovosti. Viacdetné manželstvá (3 a viac závislých detí) sa vyznačujú väčšou stabilitou (významnú úlohu zohráva aj náboženské vyznanie, ale aj ekonomické problémy rodiny) [2]. Vo všetkých veľkostných súboroch obcí tvorila takáto skupina rozvodov najnižší podiel. Štruktúra rozvodov podľa počtu detí odráža najmä štruktúru rodín v jednotlivých obciach. V menších obciach s viacdetným modelom rodiny je aj vyšší podiel rozvádzajúcich sa manželstiev s 2 a viac deťmi, v mestách nad 20 000 obyvateľov pribúda podiel rozvodov bezdetných manželstiev. Mnohé z týchto manželstiev odložili splodenie dieťaťa na neskôr, čo oslabilo funkciu v manželstva v už aj tak kritických prvých rokoch.

4. Záver

Štatistické vyhodnotenie dát potvrdilo hypotézu, že mestské a vidiecke obyvateľstvo sa vyznačuje špecifickými črtami zakladania, stability aj rozpadu rodiny.

Pre mestské prostredie je typické znižovanie intenzity plodnosti, posun plodnosti do vyššieho veku v dôsledku uprednostnenia individuálnej životnej dráhy pred založením rodiny. V mestách je preferovaný model neskorej sobášnosti a model jednodetnej max. dvojdetnej rodiny, zatiaľ čo vidiecke obyvateľstvo svoju plodnosť realizuje už v nižšom veku a model rodiny je prevažne dvoj- a viacdetný. Mestské prostredie je vhodné na realizovanie neformálnych partnerských zväzkov, túto skutočnosť dokazuje aj podiel mimomanželsky narodených detí. Anonymné prostredie mesta tiež podporuje rozvodovosť ako takú a jej akceptácia je vyššia ako na vidieku.

Na druhej strane prebiehajú v posledných rokoch silné suburbanizačné procesy, najmä v zázemí našich najväčších miest. Týmito procesmi sa do vidieckeho prostredia dostáva populačné správanie mestského obyvateľstva. Celkove možno predpokladať, že rozdiely v demografickom správaní obyvateľstva sledovaných súborov sa budú postupne zmenšovať.

5. Použitá literatúra

[1] BENEŠOVÁ, V. (2001): Současné demografické změny podle výsledků sociologických výzkumů. In: Demografie, roč. 43, č. 2, s. 111 – 124

[2] MICHÁLEK, A. (2000): Rozvodovosť, jej príčiny, dôsledky, vývoj a regionálne diferencie. In: Slovenská štatistika a demografia, roč. 10, č. 3, s. 24 - 34

[3] MLÁDEK, J. a kol. (1998): Demografia Slovenska. Vývoj obyvateľstva, jeho dynamika, vidiecke obyvateľstvo. Univerzita Komenského, Bratislava, s. 194

[4] MLÁDEK, J., MARENČÁKOVÁ, J., ŠIROČKOVÁ, J. (2006): Demografické správanie vysokoškolákov Slovenska. 2. časť – rodinné správanie. In: Slovenská štatistika a demografia, roč. 16, č. 2, s. 71 – 88

[5] MLÁDEK, J., ŠIROČKOVÁ, J. (2002): Vplyv urbanizácie na rozvodovosť obyvateľstva Slovenska. In: Geografické informácie 7, Nitra, s. 177 – 186

[6] Stav a pohyb obyvateľstva v Slovenskej republike 2001, 2002, 2003. Štatistický úrad Slovenskej republiky, Bratislava

Adresa autora

Katarína Čupeľová, 2. ročník magisterského štúdia,
Univerzita Komenského v Bratislave
Prírodovedecká fakulta,
Katedra humánnej geografie a demogeografie,
katka_cupelova@centrum.sk

Obchodný systém AUD/USD – zostavenie a analýza efektívnosti systému
AUD/USD trading system – designing the system and analyzing its efficiency

Jakub Martinovič Husár

Abstract: The aim of this paper is the presentation of self trading online system on a foreign exchange rate of AUD/USD, test of the system using the statistical methods and identification of the weak points of the system. We will create the decision tree that will system use for decision making while opening short or long position. We suppose the system trading using, Awesome Oscillator, Relative Strength Index (RSI), Average True Range (ATR) which are technical indicators. We do not include influence of the fundamental indicators to our analysis. We will test the decision tree system several times with the different values of parameters and at the various time series in order to find optimal setting of system parameters. Final result analyzing the model will be then reported and results interpreted. The aim of this analysis is to find out optimal trading self-decision making system that will be profitable.

Keywords: AUD/USD, trading system, Awesome Oscilator, Relative Strength Index, Average True Range, decision tree

Kľúčové slová: AUD/USD, obchodný systém, Awesome Oscilator, Relative Strength Index, Average True Range, rozhodovací strom

1. Úvod

Technická analýza vývoja obchodovaného aktíva sa opiera o dve základné skutočnosti. Prvou je existencia reakcie trhu na zmenu fundamentálnych ukazovateľov, ktorá sa prejaví na grafe obchodovaného aktíva a druhou skutočnosťou je neistota, že skutočne trh bude na tieto informácie reagovať podľa teórie. Praktické skúsenosti ukázali, že v reálnom obchodovaní je možné docieliť 60-%-nú úspešnosť technickej analýzy, pričom 40 % signálov ktoré ponúka, je nepoužiteľných. V tomto bode nastáva priestor pre analýzu efektívnosti zostavenej techniky obchodovania, ktorá je testovaná na časových radoch obchodovaného aktíva v dlhom časovom období. Pri testovaní zostavenej techniky obchodovania, či už pomocou rozhodovacieho stromu, integrálneho ukazovateľa, alebo ich kombináciou, musíme správne nastaviť kritériá splnenia podmienok a vstupné parametre, ktoré v našom modely vystupujú ako konštanty. Následne posledným krokom je zostavenie hodnotenia modelu, kde si všímame jeho štatistickú významnosť, teda kvalitu modelu, rozptyl hodnoty účtu, maximálny pokles hodnoty obchodného účtu a celkovú ziskovosť, prípadne taktiež stabilitu obchodovania. Našim cieľom je nájsť ziskový obchodný systém, ktorý by bol stabilný a neznamenal pre obchodníka vysoké riziko. Celá simulácia obchodovania je vykonávaná systémom MetaQuotes (terminál MetaTrader) na účte spoločnosti ActivTrades UK Ltd. Rozhodovací strom je naprogramovaný v jazyku MQL a výstupy sú následne analyzované a doplnené spracovaním v programe Microsoft Excel. Obchodovaným aktívom je devízový pár AUD/USD.

2. Postup a použité metódy

Predpokladáme, že na finančnom trhu vznikajú krátkodobé a dlhodobé výkyvy od rovnovážnej hodnoty ceny obchodovaného aktíva. Ďalej predpokladáme existenciu nelogického správania investorov, ktoré vplýva na cenu aktíva v nesúlade s teóriou.

Postupným testovaním si zvolíme ukazovatele. Awesome Oscillator (ďalej aj AO) determinuje momentum trhu na posledný piatich baroch (bar je časový úsek zobrazený na grafe ako jeden bod v prípade sviečkového grafu ako jedna sviečka, rozhodujúcou cenou je uzatváracia)(5 períód), porovnávajúc ich s momentom na 34 poslených baroch (perióda 34). Awesome Oscillator je rozdielom medzi 34-periódovým a 5-periódovým jedoduchých kĺzavých priemerov stredov cien jednotlivých barov (High+Low)/2. Awesome oscillator je tradične zobrazený vo forme histogramu. Pre výpočet použijeme nasledovný vzťah:

$$x_i = \frac{High(i) - Low(i)}{Low(i)}, \tag{1a}$$

$$AO_i = SMA(x_i,5) - SMA(x_i,35), \tag{1b}$$

kde High je najvyššia cena zobchodovaná v danom časovom období, Low je najnižšia cena zobchodovaná v danom časovom období a SMA (Simple moving average) je jednoduchý kĺzavý priemer, pričom vstupmi sú použitá cena a perióda výpočtu

Signálom pre nákup, alebo predaj pri tomto indikátore je zvýšenie, alebo zníženie hodnoty, pričom nákup sa vykoná, ak hodnota indikátora súčasná je vyššia, ako hodnota indikátora predošlého baru (predošlá časová jednotka – v našom prípade 5 minút).

Ďalším použitým indikátorom je Relative Strength Index (ďalej RSI)[1]. Bol vyvinutý J. Wellsom Wilderom, ktorý zverejnil v knihe New Concepts in Technical Trading Systems v roku 1978. Indikátor Relative Strength Index (RSI) sa pokladá za veľmi užitočný indikátor. Patrí do skupiny momentových oscilačných indikátorov RSI, porovnáva magnitúdu zisku aktíva s magnitúdou súčasných strát aktíva a vytvára číselný rozsah v rozmedzí od 0 do 100. Zahŕňa jednoduchý parameter a tým je číslo períód pre výpočet indikátora. Wilder odporúča vo svojej knihe indikátor nastavený na 14 períód.

$$RSI = 100 - \frac{100}{1 + RS}, \tag{2a}$$

$$\text{RS(relative strength)} = \text{Priemerné zisky/Priemerné straty}, \tag{2b}$$

$$\text{Priemerné zisky} = [(\text{Priemerný zisk(t-1)}) \times 13 + \text{Súčasný zisk}] / 14, \tag{2c}$$

$$\text{Prvý priemerný zisk} = \text{Suma ziskov za 14 períód} / 14, \tag{2d}$$

$$\text{Priemerné straty} = [(\text{Priemerné straty(t-1)}) \times 13 + \text{Súčasná strata}] / 14, \tag{2e}$$

$$\text{Prvá priemerná strata} = \text{Úplné straty za 14 períód} / 14. \tag{2f}$$

Pri tomto indikátore by sme nákup a predaj realizovali podľa polohy indikátora v intervale 0 až 100. Štandardne by mal obchodník nakupovať, ak index klesne pod hodnotu 30 a predávať, ak klesne nad hodnotu 70. Je možné aj analyzovať pozitívne, alebo negatívne divergencie, na základe ktorých môžeme odhadovať moment divergencie vývoja cien a jeho potenciálnu silu. Indikátor budeme využívať pre potvrdenie divergencie, ktorú bude indikovať indikátor Awesome oscillator. V prípade, že jeho hodnota poklesne pod 35, bude to pre systém indikovať potvrdenie dlhej pozície nákupu. V prípade, že jeho hodnota vzrastie nad 65, systému bude indikovať RSI potvrdenie otvorenia krátkej pozície.

Pre meranie volatility trhov bol vyvinutý Average True Range index (ďalej ako ATR). Index vyvinul J. Welles Wilder, ktorý zverejnil v knihe New Concepts in Technical Trading Systems. Wilder pôvodne používal index ATR pre komodity, ale index je možné používať

[1] Zdroj: Spracované podľa J. J. Murphyho, strana 239

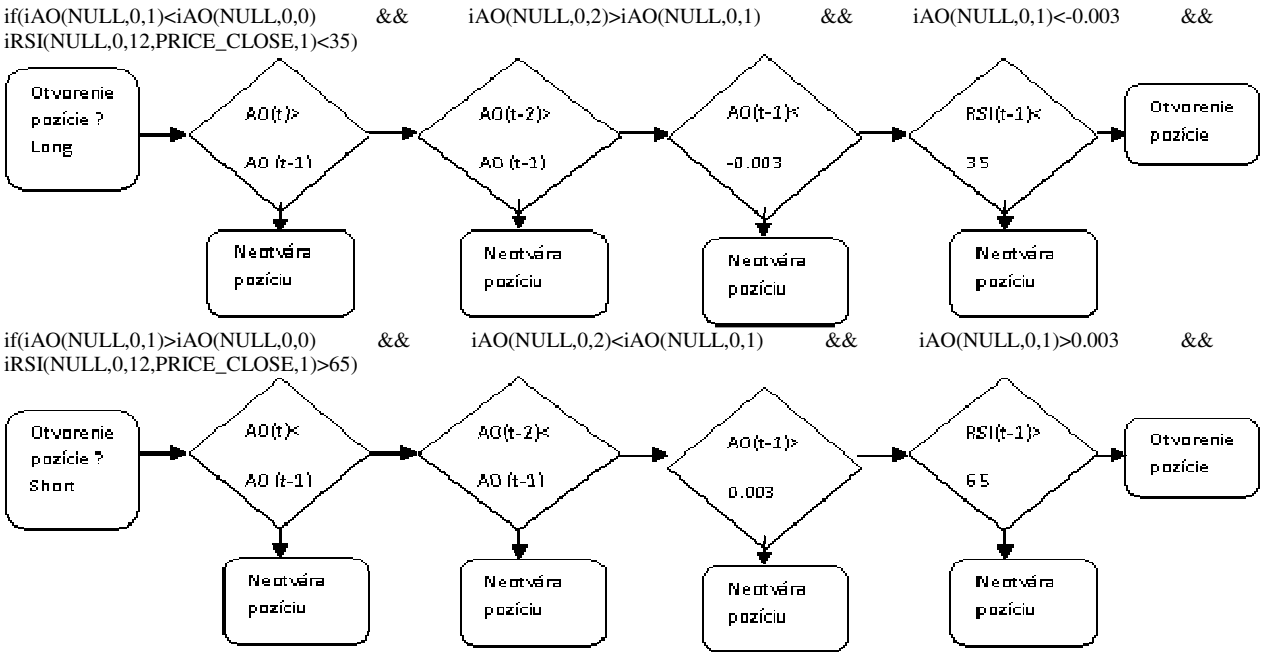
taktiež pre indexy akcie a menové páry. Meranie ATR je zamerané na meranie bezpečnej volatility. Vysoké hodnoty ATR indikujú vysokú volatilitu, zatiaľ čo nízke indikujú nízku volatilitu. True Range je definovaný ako najväčšia divergencia:

– súčasných high mínus súčasných low,
– absolútna hodnota súčasných high mínus predošlá uzatváracia cena,
– absolútna hodnota súčasných low mínus predošlá uzatváracia cena.

Average True Range (ATR) kĺzavý priemer vypočítaný pre niekoľkodňové časové periódy.

Dôkladnou analýzou vytvoríme rozhodovací strom, ktorý bude uzatvárať, alebo otvárať pozície na základe stavu troch spomenutých ukazovateľov.

Diagram č. 1: Diagramy rozhodovacích stromov otvárania pozícií



Zdroj: Vlastné spracovanie

3. Parametre systému

Najdôležitejšou časťou pre efektívne fungovanie systému je nastavenie parametrov. Parametre sme testovali niekoľko krát s rôznymi hodnotami a rôznymi kombináciami. Nakoniec sme si zvolili konštanty.

Tabuľka č. 1: Parametre systému

Typ parametra	Názov parametra	Hodnota parametra
extern int	TakeProfit	100
extern int	StopLoss	35
extern int	TrailingStop	36
extern double	Lots	2

Zdroj: Vlastné spracovanie

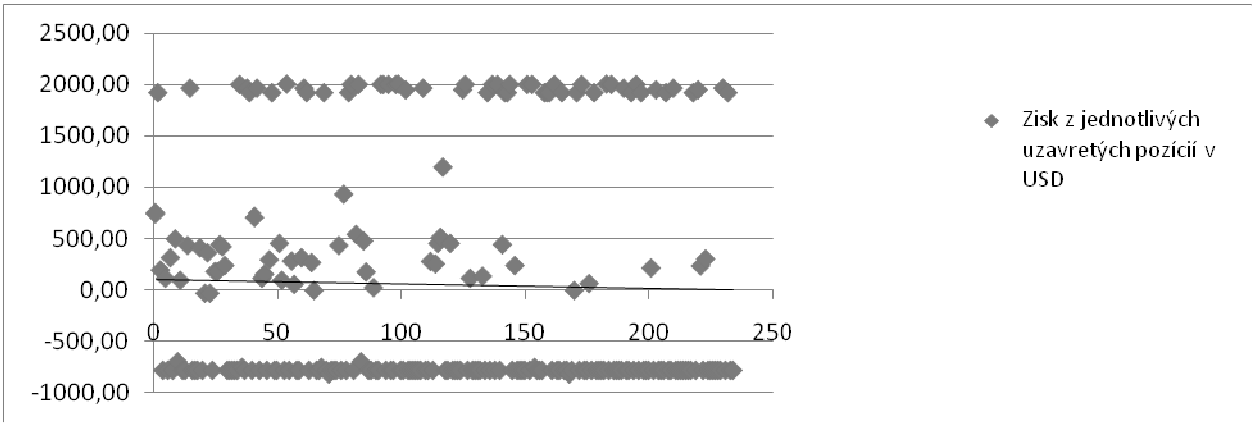
4. Dáta

Za účelom otestovania naprogramovaného systému sme ponechali obchodovať simulátor MetaQuotes na historických dátach 2008.04.02 08:35 po 2008.10.30 23:55 na 5 minútovom grafe[2]. Následne sme ešte vykonali niekoľko testov na iných periódach, ale bližšie

[2] Zdroj: dáta boli poskytnuté spoločnosťou Activtrades UK Ltd.

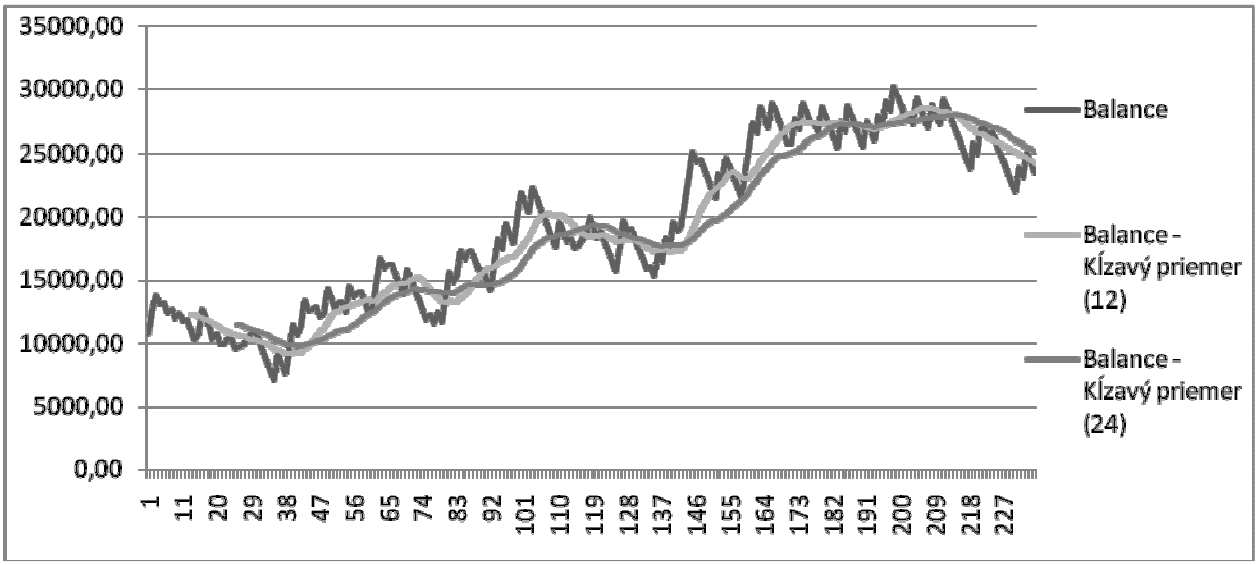
opíšeme predovšetkým výsledky testu na spomenutej perióde. Musíme ale zdôrazniť, že model, ktorý bol zostavený a otestovaný na anteriórnych dátach podlieha nepriaznivým vplyvom ako slippage a problémy vyplývajúce z nedostatočného množstva informácii o objeme dopytu a ponuky. Systém po dotestovaní na dátach dosiahol nasledovné výsledky.

Graf č. 1: Rozdelenie zisku z uzavretých pozícií



Zdroj: Vlastné spracovanie podľa výstupu obchodovaného modelu zo simulácie MetaQuotes v programe MS Excel

Graf č. 2: Graf vývoja vývoja obchodného účtu, s kĺzavými priemermi 12tej a 24tej periódy



Zdroj: Vlastné spracovanie v programe MS Excel podľa výstupu simulácie programu MetaQuotes

Systém otvoril a uzavrel 235 obchodov na menovom páre. Systém uzavrel 97 obchodov formou krátkych pozícií , z čoho 42,27% bolo úspešných a 138 dlhých pozícií, z ktorých úspešných bolo 42,03% . Stratových obchodov bolo 138 (42.03%) a ziskových obchodov bolo 99 (42.13%). Najvýnosnejší obchod dosiahol zisk 2004 amerických dolárov (USD) a najvyššia strata bola 815,8 amerických dolárov (USD). Priemerná hodnota ziskových obchodov je 1190,97 (USD) a priemernou hodnotou stratových obchodov je -767,61. Hrubý zisk testovaného obchodného systému je 117906.00 amerických dolárov (USD) hrubá strata je -104394.40 amerických dolárov (USD). Maximálny drawdown systému bol 9527.60 amerických dolárov (USD), čo predstavuje riziko v hodnote 30.89%. Depozit na účte mal počiatočnú hodnotu 10000 amerických dolárov (USD), pričom systém v sledovanom období vyprodukoval zisk 13511,60 dolárov (USD). Systém detegoval výskyt 15tich chýb v priebehu

generovania, pričom kvalita modelu bola 58.23%, tá ale nie je voľne interpretovateľná. Využíva sa predovšetkým pri porovnávaní niekoľkých modelov

5. Záver

Na základe použitých metód a aplikovaných anteriórnych testov sme zostavili model obchodovania s menovým párom AUD/USD. Za predpokladu nezmeneného charakteru trhu a dlhodobého obchodovania, by mal obchodný systém vykazovať zisky. Pomerne nízke riziko, ktoré chápeme ako potenciálny pokles hodnoty účtu (account drawdown), ktorý dosiahol hodnotu 30.89% je v prípade pozitívnym znakom kvality a stability systému. Musíme ale zdôrazniť, že testovaný model na anteriórnych dátach mal kvalitu iba 58.23% čo je značne nízke percento na to, aby sme model pokladali za použiteľný bez testovania dát v reálnom obchodnom prostredí. Pre ohodnotenie modelu by následne mohli byť použité ešte iné ukazovatele, ktoré by podali presnejší obraz o kvalite a efektivite obchodného systému, pre ktoré by sme ale potrebovali väčšie množstvo dát.

6. Literatúra

[1.] BOJSA, M. 2007. Obchodovanie na amerických akciových trhoch, Banská Bystrica: Active Trader.

[2.] KAHN, M. N. 2006. . Technical Analysis Plain and Simple, Second Edition , Charting the Markets In Your Language, New Jersey: FT Prentice Hall, 2006 by Pearson Education,Inc.

[3.] MURPHY, J. J. 1999. Technical Analysis of The Financial Markets, New Yourk: Institute of finance, 1999.

Adresa autora

Jakub Martinovič Husár
Univerzita Mateja Bela v Banskej Bystrici
Ekonomická fakulta
Tajovského 10
975 90 Banská Bystrica
jakub.martinovic.husar@gmail.com

Využitie štatistických metód pri finančno-ekonomickej analýze odvetvia
Using the Statistical Methods in the Financial-Economical Analysis of the Industry

Miroslav Kello

Abstract: In this paper we tried to classify the companies into different groups, so that the companies in each subset shared some common trait. As methods we used factor and cluster analysis together with multivariate statistical methods. The results were obtained using the tools in SAS Learning Edition 2.0 and in MS® Excel environment.

Key words: cluster analysis, factor analysis, multivariate statistical methods, Kaiser's rule

Kľúčové slová: zhluková a faktorová analýza, viacrozmerné štatistické metódy, Kaiserovo pravidlo

1. Úvod

Metódy analýzy finančno-ekonomickej pozície podniku v priestore možno využiť pri porovnávaní podobnej skupiny podnikov, teda napr. skupiny s rovnakou ekonomickou činnosťou. Pre korektný priebeh analýz je dôležitý ako výber vhodných podnikov, tak aj výber nezávislých kritérií, podľa ktorých sa budú podniky rozlišovať. Uvedeným skutočnostiam je potrebné venovať zvýšenú pozornosť, keďže v konečnom dôsledku ovplyvňujú výsledok medzipodnikového porovnávania. Cieľom tejto práce je využitie vybraných štatistických metód pri analýze podnikov hutníckeho priemyslu v Slovenskej republike, a to za účelom ich rozdelenia do zhlukov, obsahujúcich podobné – porovnateľné podniky. Analýzy boli prevedené v programoch SAS Learning Edition 2.0 a MS® Excel.

2. Dáta, metodický postup a empirické výsledky

Ako predmet analýzy sme si vybrali 30 priemyselných podnikov, patriacich podľa ich hlavnej činnosti pod OKEČ číslo 27 – Výroba kovov. Skúmanými charakteristikami je celkovo 19 pomerových ukazovateľov rozdelených do jednotlivých skupín podľa ich charakteru : ukazovatele likvidity, zadlženosti, aktivity, rentability. Zo všetkých ukazovateľov sme sa na základe nášho úsudku rozhodli vybrať nasledujúcich 9 ukazovateľov:

L2 - bežná likvidita
DSP - doba splatnosti pohľadávok
DSZ - doba splácania záväzkov
CZA - celková zadlženosť aktív
FP - finančná páka
ROE - rentabilita vlastného imania
ROS - rentabilita tržieb
PPHT - podiel PH v tržbách
PNHvT - podiel novovytvorenej hodnoty v tržbách

Na začiatku analýzy je vhodné preskúmať vzťahy medzi premennými, a to napr. prostredníctvom korelačnej matice. V našom prípade sme zistili existujúcu závislosť medzi

viacerými dvojicami ukazovateľov, z dôvodu prehľadnosti však nie je možné uviesť celú maticu. Využitie zhlukovej analýzy predpokladá (pri euklidovskej vzdialenosti) nezávislosť premenných. Tento problém je možné riešiť vynechaním štatisticky významných dvojíc ukazovateľov, avšak v našom prípade ide o relatívne vysoký počet premenných. Ako vhodná metóda sa v tomto prípade javí tzv. faktorová analýza.

Faktorová analýza je metóda zameraná na vytváranie nových hypotetických premenných a na zníženie rozsahu údajov za podmienky čo najmenšej straty informácie, ktorú tieto údaje nesú. Jedným z jej základných cieľov je preskúmanie vzťahov medzi premennými a zistiť tak, či je možné ich rozdeliť do skupín, ktoré by mali vysokú vnútornú koreláciu avšak medzi týmito skupinami by bola korelácia nízka.

Matematicky možno vyjadriť faktorovú analýzu nasledujúcim spôsobom:

$$X_j = \mu_j + a_{j1}F_1 + a_{j2}F_2 + ... + a_{jq}F_q + \varepsilon_j \qquad (1)$$

kde

X_j - j-ta pozorovateľná premenná

$F_1, F_2, ..., F_j$ - spoločné faktory

ε_j - náhodné (chybové) zložky

a_{jk} - faktorové váhy, ktoré vyjadrujú vplyv k-teho spoločného faktora na premennú X_j

Na odhad parametrov modelu sme použili tzv. metódu hlavných komponentov (PCA – principal component analysis) s využitím ortogonálnej rotácie. Ako kritérium vhodnosti použitia dát sme využili KMO štatistiku. Táto štatistika skúma rozdiely medzi zodpovedajúcimi si párovými a parciálnymi koeficientmi korelácie a možno ju vyjadriť nasledujúcim spôsobom:

$$KMO = \frac{\sum_{i \neq j}^{p} \sum_{i \neq j}^{p} r_{ij}^2}{\sum_{i \neq j}^{p} \sum_{i \neq j}^{p} r_{ij}^2 + \sum_{i \neq j}^{p} \sum_{i \neq j}^{p} r_{parc.,ij}^2} \qquad (2)$$

kde

r_{ij} - párový koeficient korelácie medzi X_j a X_i

$r_{parc.,ij}$ - parciálny koeficient korelácie

Za vhodné ukazovatele sme považovali tie, pri ktorých ukazovateľ KMO dosiahol hodnotu minimálne 0,5.

Tabuľka č.1 Hodnoty KMO pre pôvodné ukazovatele

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.56205392								
L2	**DSP**	**DSZ**	**CZA**	**FP**	**ROE**	**ROS**	**PNHvT**	**PPHT**
0.5860	0.4187	0.5887	0.6269	0.4995	0.4648	0.6216	0.5138	0.7838

Zdroj: vlastné spracovanie

Ako môžeme vidieť, ukazovatele DSP, FP a ROE nespĺňajú danú podmienku. Ako prvý sme sa rozhodli vylúčiť z analýzy ukazovateľ DSP. Následne sme celý proces opakovali, až kým sme nedospeli k výsledku, kde je hodnota KMO pre všetky ukazovatele vyššia ako 0,5.

Tabuľka č.2 Hodnoty KMO pre zvolené ukazovatele

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.66709678						
DSZ	**CZA**	**L2**	**ROE**	**ROS**	**PNHvT**	**PPHT**
0.75809460	0.64623452	0.55443148	0.50993070	0.75221889	0.69670308	0.77555540

Zdroj: vlastné spracovanie

Z pôvodného súboru premenných sme v konečnom dôsledku vynechali ukazovatele DSP a FP, celková miera KMO sa zvýšila z 0,562 na 0,667.

Takto zredukované ukazovatele sme následne transformovali na nezávislé hlavné komponenty. V nasledujúcej tabuľke je možné vidieť vlastné čísla a podiel variability vysvetlený jednotlivými komponentmi, podľa ktorých je možné odhadnúť ich počet.

Tabuľka č.3 Vlastné čísla a podiel variability vysvetlený jednotlivými komponentmi

Eigenvalues of the Correlation Matrix: Total = 7 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	2.91678636	0.70288273	0.4167	0.4167
2	2.21390363	1.29852029	0.3163	0.7330
3	0.91538333	0.47729366	0.1308	0.8637
4	0.43808968	0.19862808	0.0626	0.9263
5	0.23946159	0.08251089	0.0342	0.9605
6	0.15695070	0.03752600	0.0224	0.9829
7	0.11942471		0.0171	1.0000

Zdroj: vlastné spracovanie

Na základe Kaiserovho pravidla za hlavné komponenty považujeme tie, ktoré majú vlastné číslo väčšie ako 1. Tieto komponenty však vysvetľujú iba 73,30 % variability pôvodných premenných. Preto sme sa rozhodli pridať aj tretí hlavný komponent, ktorý dohromady s ostatnými vysvetľuje 86,37 %, čo ešte stále nie je optimálne, ale pre účely tejto analýzy postačujúce. Príslušnosť ukazovateľov k jednotlivým hlavným komponentom je vyjadrená v nasledujúcej tabuľke. V našom prípade ide o párové koeficienty korelácie príslušnej premennej a daného hlavného komponentu.

Tabuľka č.4 Príslušnosť ukazovateľov k jednotlivým faktorom

Rotated Factor Pattern				
		Factor1	**Factor2**	**Factor3**
PNHvT	PNHvT	0.93309	-0.06099	-0.15049
PPHT	PPHT	0.90499	0.12959	0.03594
ROS	ROS	0.88615	-0.04937	-0.30484
ROE	ROE	0.10270	0.94961	-0.06107
L2	L2	-0.14191	0.88362	0.32281
CZA	CZA	-0.31115	-0.04277	0.87101
DSZ	DSZ	0.13058	0.29899	0.69771

Zdroj: vlastné spracovanie

Takto zostavené faktory sú nezávislé, a teda sú vhodné ako vstupné premenné v zhlukovej analýze. Ako metódu zhlukovania sme si zvolili tzv. Wardovu metódu za použitia

euklidovskej vzdialenosti. Táto metóda je v praxi najpoužívanejšia, zhluky sa podľa nej formujú prostredníctvom maximalizácie vnútrozhlukovej homogenity. Jej mierou je vnútrozhluková suma štvorcov odchýlok od priemeru zhluku:

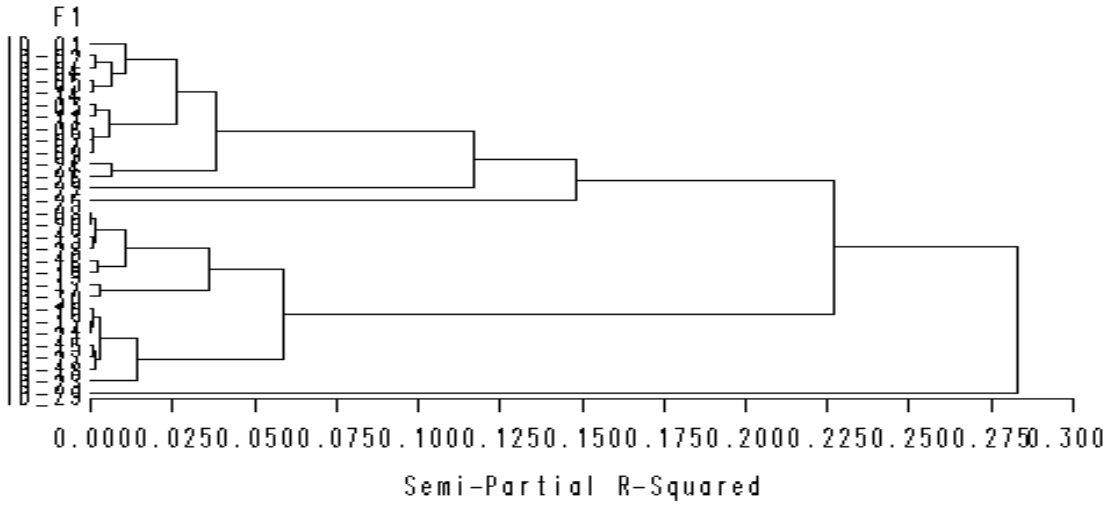
$$ESS = \sum_{i=1}^{n_h} \sum_{h=1}^{q} (x_{hi} - \bar{x}_{c_h})^2 \qquad (3)$$

kde

n_h - je počet objektov v zhluku C_h

$\bar{x}_{c_h}$ - vektor priemerov hodnôt znaku v zhluku C_h

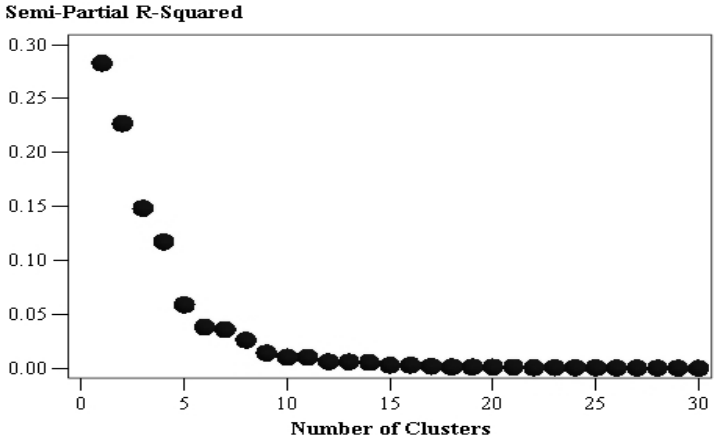
x_{hi} - vektor hodnôt znaku i-teho objektu v zhluku C_h



Graf č.1 Dendogram zhlukovania podnikov

Zdroj: vlastné spracovanie

Na zistenie počtu významných zhlukov využijeme graf znázorňujúci vývoj semiparciálneho koeficientu determinácie v závislosti od počtu zhlukov, pričom si všímame na akom stupni zhlukovania nadobúda tento koeficient minimálnu hodnotu pri čo najnižšom prírastku pre ďalšie úrovne zhlukovania.



Graf č.2 Vývoj semiparciálneho koeficienta determinácie

Zdroj: vlastné spracovanie

Ako vidíme, na piatom, resp. šiestom stupni zhlukovania dochádza už iba k minimálnemu prírastku koeficienta. Príslušnosť podnikov do jednotlivých zhlukov zobrazuje nasledujúca tabuľka:

Tabuľka č.5 Príslušnosť podnikov k jednotlivým zhlukom

CLUSTER	meno	CLUSTER	meno
1	Schüle Slovakia, s.r.o.	3	ZLIEVÁREN TRNAVA, s.r.o.
	METALURG STEEL, spol. s r.o		LLEMI SLOVAKIA s.r.o.
	MEDEKO CAST s.r.o.		Zlieváreň SEZ Krompachy a.s.
	KOVOMONT KRÁSNO, s.r.o.		Nemak Slovakia s.r.o.
	ZSNP, a.s.		MOPS PRESS, s.r.o.
	ZAL - Zlievareň Al,spol. s r.o.		ZLIEVAREN SVIT, a.s.
	STRIP, a.s. Košice		EUROCAST Košice, s.r.o.
	VESUVIUS SLOVAKIA, s.r.o.		ZGH, s.r.o
2	VALSABBIA SLOVAKIA, s.r.o.		TZB Global, s.r.o.
	TUBAPACK, a.s.		Zlieváren Zábrež, a.s.
	Med Slovakia a.s.		FERPLAST SLOVAKIA, s.r.o.
	Sapa Profily a.s.		CMK, s.r.o.
	BOHUŠ s.r.o.	4	TAVAL, spol s r.o.
	KOVOHUTY, a.s.	5	Valcovňa profilov a.s.
	Fagor Ederlan Slovensko, a.s.	6	PRAKOVSKÁ OCELIARSKA SPOL.s.r.o.

Zdroj: vlastné spracovanie

Hodnotenie jednotlivých zhlukov sme uskutočnili na základe metód viacrozmerného hodnotenia, konkrétne metódy normovanej premennej. Podstatou tejto metódy je prevod rôznych hodnôt ukazovateľov na bezrozmerné číslo – normovanú premennú.

Tabuľka č.6 Poradie jednotlivých zhlukov

CLUSTER	1	2	3	4	5	6
počet podnikov	8	7	12	1	1	1
Integr. ukazovateľ	0,20661038	-0,32024	-0,25368	-0,69029	1,433009	-0,36447
poradie	2	4	3	6	1	5

Zdroj: vlastné spracovanie

3. Záver

Prostredníctvom faktorovej a zhlukovej analýzy sme rozdelili pôvodný súbor podnikov na 6 skupín, pričom posledné tri skupiny obsahovali každá po jednom podniku. V tomto prípade môže ísť o podniky, ktoré zahŕňajú extrémne hodnoty ukazovateľov. Pre nás sú tak dôležité najmä zhluky č. 1, 2 a 3. Prvý zhluk obsahuje podniky s relatívne nízkym zadlžením, krátkou dobou splatnosti záväzkov a relatívne vysokou rentabilitou vlastného imania a tržieb. Skupina podnikov č. 2 a 3 dosahuje v tejto oblasti horšie výsledky, ukazovateľ podielu pridanej hodnoty v tržbách je však najlepší práve pri podnikoch patriacich do tretej skupiny. Cieľom tejto analýzy bolo využiť vybrané štatistické metódy pri analýze odvetvia, v rámci hlbšej analýzy by bolo vhodné vybrať z každej skupiny podnikov jeden reprezentatívny a na základe ich účtovných závierok a iných informácií skúmať príčiny rozdielov medzi týmito podnikmi.

4. Použitá literatúra

[1] HEBÁK, P. a kol.: *Vícerozmerné statistické metody*, Praha: Informatorium, 2005

[2] STANKOVIČOVÁ, I., VOJTKOVÁ, M.: *Viacrozmerné štatistické metódy s aplikáciami*, Bratislava: Iura Edition, 2007

[3] ZALAI, K. a kol.: *Finančno-ekonomická analýza podniku*, Bratislava: SPRINT, 2007

Adresa autora:

Miroslav Kello, 5. ročník
Ekonomická fakulta
Univerzita Mateja Bela
Tajovského 10
975 90 Banská Bystrica
miroslavkello@yahoo.com

Porovnanie úrovne terciárneho vzdelávania v krajinách Európskej únie
The comparison of the level of tertiary education in countries of European union

Jana Lacková

Abstract: The work deals with a comparison of the European Union (EU) countries from the aspect of level and opportunities of education at universities and other institutions that belong to the 5. and 6. level of International Standard Classification of Education. Different level of educational systems is the result of different political, social and business progress in particular states. The work comes out from one of the main goals of Lisboan strategy that is the creation of knowledge society that would be attractive for all groups of scientists. As it is a complex problem it was needed to select some criteria (7) and analyse the level by usage of multidimensional statistic methods. In more details it deals with multivariate statistic methods analysis of independency and classificatory methods. In the conclusion of this document are completed results of using stated methods.

Key words: Lisboan strategy, factor analysis, cluster analysis

Kľúčové slová: Lisabonská stratégia, faktorová analýza, zhluková analýza

1. Úvod

Uvedená práca prináša a systematizuje dostupné informácie o stave terciárneho vzdelávania v jednotlivých krajinách Európskej únie (EÚ). Vychádza pritom z jedného z hlavných cieľov Lisabonskej stratégie, a to zo snahy o vytvorenie znalostnej spoločnosti atraktívnej pre všetky skupiny vedcov a výskumníkov.

Vzhľadom na cieľ práce, ktorým je porovnanie krajín EÚ na základe úrovne ich vysokoškolských systémov, pozornosť je sústredená najmä na analýzu spomínanej úrovne. Keďže ide o problém, ktorý nie je možné vyjadriť jedným ukazovateľom, bolo potrebné vybrať niekoľko relevantných ukazovateľov (7 ukazovateľov) a analyzovať ich pomocou viacrozmerných štatistických metód. Výber ukazovateľov bol do veľkej miery ovplyvnený aj dostupnosťou údajov. Získané údaje pochádzajú z internetového portálu Eurostat a Inštitútu pre štatistiky UNESCO. Na analýzu boli použité údaje z roku 2005, keďže mnoho krajín EÚ novšie údaje z tejto oblasti nevykazuje.

Keďže sa jedná o skorelované ukazovatele, na analýzu bolo potrebné použiť faktorovú analýzu a následne aplikovať zhlukovú analýzu. Rozsah štatistického súboru zahŕňal 26 štátov EÚ. Cyprus bolo nutné z analýzy vylúčiť kvôli neprimerane odlišným hodnotám určitých ukazovateľov a následnému skresľovaniu výsledkov analýzy.

2. Hodnotiace kritériá

V marci roku 2000 na mimoriadnom zasadnutí Rada Európskej únie prijala tzv. Lisabonskú stratégiu. Jednalo sa o plán zvyšovania hospodárskej úrovne pri zachovaní princípov sociálneho modelu EÚ. Tento program radikálnej obnovy bol neskôr rozšírený aj na environmentálnu a sociálnu oblasť, čo malo za následok zneprehľadenie cieľov ako aj možnosti ich vyhodnocovania. Nedokonalosť plánu ešte umocňovali nejasne definované pravidlá zodpovednosti za plnenie prijatých cieľov. Tento stav viedol v roku 2004 predstaviteľov EÚ k prehodnoteniu vytýčených cieľov a k ich úprave, ako aj k zvýšeniu zodpovednosti členských krajín za ich plnenie. Táto práca monitoruje schopnosť jednotlivých

krajín zabezpečiť občanom EÚ dostupné a kvalitné vzdelávanie na úrovni terciárneho vzdelávania, zahŕňajúce vysokoškolské a postgraduálne štúdium.

Obmedzujúce podmienky dostupnosti údajov o úrovni vzdelávania v kombinácii so snahou čo najlepšie odraziť realitu viedli k výberu nasledujúcich ukazovateľov:

1. relatívny počet študentov (X1) - ukazovateľ vyjadruje percentuálny podiel študentov vzhľadom na celkovú populáciu,
2. podiel študentov študujúcich v zahraničí (X2) - ukazovateľ vyjadruje percentuálny podiel študentov odchádzajúcich študovať do zahraničia vzhľadom na celkový počet študentov danej krajiny,
3. podiel zahraničných študentov na celkovom počte študentov (X3) - je vyjadrením percentuálneho podielu zahraničných študentov študujúcich v jednotlivých krajinách EU vzhľadom na počet domácich študentov,
4. priemerné ročné výdavky na študenta (X4) - tento ukazovateľ je odrazom priemerných výdavkov na jedného študenta, pričom sú tu zahrnutí študenti súkromných aj verejných inštitúcií tretieho stupňa. Taktiež sú v tomto ukazovateli zahrnuté výdavky na personál, ale aj ďalšie bežné a fixné výdavky,
5. priemerné ročné výdavky na študenta v porovnaní s HDP na obyvateľa (X5) - tieto výdavky predstavujú fixné aj variabilné výdavky, ktoré sú vynakladané na vzdelávacie procesy terciárneho stupňa. Vyjadrenie pomocou HDP na obyvateľa umožňuje porovnanie úrovne ekonomickej aktivity ekonomík rôznych veľkostí bez ohľadu na cenovú úroveň danej krajiny,
6. miera nezamestnanosti vysokoškolsky vzdelaných ľudí (X6) - predstavuje mieru nezamestnanosti populácie vo veku 25- 64 rokov vzhľadom na dosiahnutú úroveň vzdelania tretieho stupňa,
7. miera zamestnanosti vysokoškolsky vzdelaných ľudí (X7) - je percentuálnym vyjadrením podielu zamestnaných ľudí vo veku 25- 64 rokov, ktorí dosiahli vysokoškolský stupeň vzdelania, a celkovej populácie danej vekovej kategórie.

Už z názvu ukazovateľov je zrejmé, že medzi hodnotami jednotlivých ukazovateľov budú existovať určité súvislosti. Tieto súvislosti je možné poodhaliť pomocou matice korelačných koeficientov, ktorá je zobrazená v nasledujúcej tabuľke:

Tabuľka 2: Korelačná matica s hodnotami Pearsonových koeficientov korelácie

	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**
X1	1,000	-0,115	**-0,511**	-0,225	-0,103	0,104	0,222
X2	-0,115	1,000	-0,118	-0,309	-0,335	0,093	-0,110
X3	**-0,511**	-0,118	1,000	**0,668**	**0,467**	-0,301	0,119
X4	-0,225	-0,309	**0,668**	1,000	**0,580**	-0,221	0,254
X5	-0,103	**-**0,335	**0,467**	**0,580**	1,000	-0,324	0,239
X6	0,104	0,093	-0,301	-0,221	-0,324	1,000	**-0,590**
X7	0,222	-0,110	0,119	0,254	0,239	**-0,590**	1,000

Zvýraznené hodnoty predstavujú významnú koreláciu medzi premennými.

3. Analýza údajov

Na základe tejto korelácie premenných je možné nahradiť uvedené premenné umelými faktormi prostredníctvom faktorovej analýzy. Ak pôvodné premenné boli X_j, kde j = 1, ..., p, tak ich lineárnu závislosť možno vyjadriť pomocou nepozorovateľných premenných (spoločných faktorov) f_i, kde i = 1, ..., m; m < p a chybových zložiek (špecifických faktorov)

ε_j, j = 1, ..., p. Spoločné faktory spôsobujú medzi premennými závislosť jednotlivých premenných a špecifické faktory, naopak, vyvolávajú rozptyl.

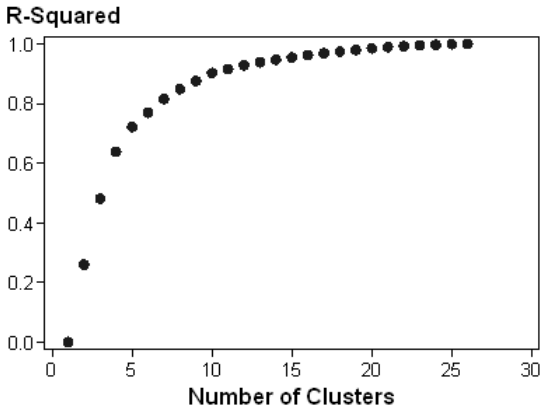
Aplikáciou faktorovej analýzy na túto skorelovanú skupinu ukazovateľov sa dopracujeme k vlastným číslam korelačnej matice:

Tabuľka 3: Hodnoty vlastných čísel korelačnej matice

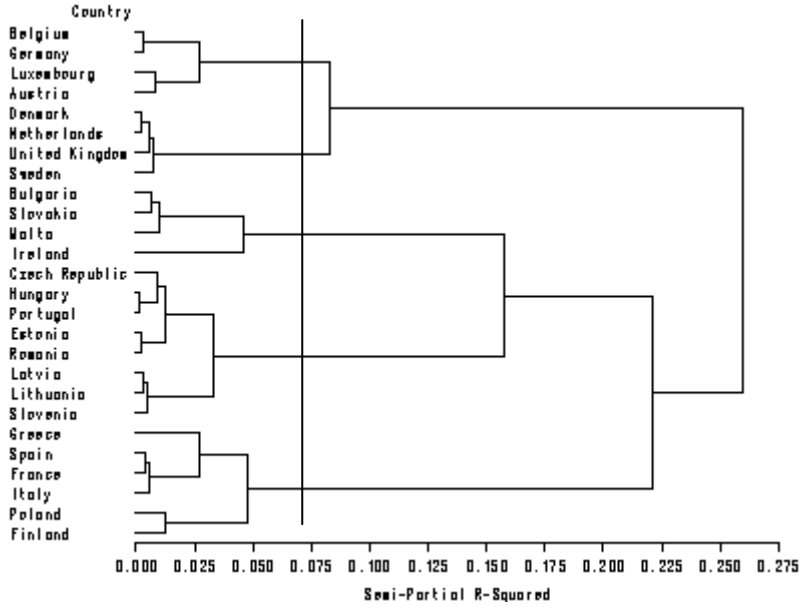
	Eigenvalue	Difference	Proportion	Cumulative
1	**2.719**	**1.204**	**0.388**	**0.388**
2	**1.515**	**0.384**	**0.217**	**0.605**
3	**1.131**	**0.511**	**0.162**	**0.767**
4	0.62	0.136	0.089	0.855
5	0.484	0.185	0.069	0.924
6	0.299	0.067	0.043	0.967
7	0.231		0.033	1

Podľa Kaiserovho pravidla by sme mali vybrať tri spoločné faktory. Tieto vysvetľujú takmer 77% celkovej variability, čo je pre naše účely dostatočné. Novovzniknuté ukazovatele - faktory už nie sú vzájomne skorelované. Nekorelované ukazovatele možno využiť na rozdelenie krajín EÚ do určitých, relatívne homogénnych, skupín použitím zhlukovej analýzy. Podstatou tejto metódy je rozdelenie množiny objektov do niekoľkých pomerne homogénnych zhlukov (skupín). Cieľom je vytvoriť zhluky C_1, ..., C_m ($2 \le m \le n$), v rámci ktorých sú si objekty X_1, ..., X_n čo najviac podobné, pričom objekty z rôznych zhlukov by sa mali čo najviac odlišovať.

Počet zhlukov možno určiť aplikáciou Wardovej metódy z RSQ - charakteristiky kvality zhlukovania (obrázok 1). Kvalita zhlukovania pri piatich zhlukoch je vysoká a radikálne sa už nemení, preto tento počet zhlukov určíme za optimálny. Proces zhlukovania na jednotlivých úrovniach zhlukovania možno zobraziť pomocou dendogramu (obrázok 2).



Obrázok 3: Graf závislosti kvality zhlukovania od počtu zhlukov



Obrázok 4: Grafické znázornenie procesu zhlukovania

Z dendogramu je rovnako možné vyčítať rozdelenie krajín EÚ do jednotlivých zhlukov:

- **zhluk č.1:** Estónsko, Lotyšsko, Litva, Portugalsko, Slovinsko, Rumunsko, Maďarsko, Česká republika
- **zhluk č.2:** Belgicko, Luxembursko, Nemecko, Rakúsko
- **zhluk č.3:** Holandsko, Dánsko, Veľká Británia, Švédsko
- **zhluk č.4:** Grécko, Taliansko, Španielsko, Francúzsko, Poľsko, Fínsko
- **zhluk č.5:** Bulharsko, Slovensko, Malta, Írsko

Základné popisné charakteristiky vstupných premenných podľa jednotlivých zhlukov sú zobrazené v tabuľke 3.

4. Záver

Táto práca sa zaoberala analýzou úrovne terciárneho vzdelávania v krajinách EÚ. Keďže sa jedná o viacrozmerný problém, boli na analýzu využité viacrozmerné štatistické metódy. Na základe zistených skutočností možno konštatovať, že krajiny EÚ možno rozdeliť do piatich pomerne homogénnych zhlukov.

Prvý zhluk tvoria krajiny vyznačujúce sa vysokým relatívnym počtom študentov. Záujem zahraničných študentov o štúdium v krajinách tohto zhluku je nízky. Naopak druhý zhluk spája krajiny s vysokými priemernými výdavkami a nízkym počtom študentov. Tieto krajiny však spája aj obrovský záujem zo strany zahraničných študentov o štúdium v nich. Krajiny tretieho zhluku vynakladajú na vzdelávanie značné investície a dosahujú vysokú mieru zamestnanosti vysokoškolsky vzdelaných ľudí. Štvrtý zhluk spája krajiny s vysokou mierou nezamestnanosti vysokoškolsky vzdelaných ľudí. Krajiny, ktoré doň patria vynakladajú nízky podiel zo svojho HDP na študentov, pričom relatívny počet študentov je pomerne vysoký. Nízku úroveň vzdelávania v krajinách piateho zhluku indikujú najmä nízke výdavky na študentov a nízka miera zamestnanosti vysokoškolsky vzdelaných ľudí. Do tohto zhluku patrí aj Slovensko. Vzhľadom na podobnosť základných charakteristík možno Cyprus dodatočne priradiť do prvého zhluku.

Napriek relatívne vysokej úrovni terciárneho vzdelávania v niektorých krajinách žiadna z krajín EÚ neexcelovala vo všetkých sledovaných kritériách. Kontinuálne zvyšovanie úrovne

vzdelávania by malo byť prioritou všetkých krajín EU. Je dôležité aby si každá krajina uvedomila svoje slabé stránky a tie sa snažila v čo najviac eliminovať.

Tabuľka 4: Prehľad základných charakteristík jednotlivých zhlukov

CLUSTER	Statistic	Variable						
		X1	X2	X3	X4	X5	X6	X7
1	Mean	0.05	0.02	0.01	4746.85	34.9	3.43	84.76
	Std Dev	0.01	0.01	0.01	1555.71	3.11	1.08	1.36
2	Mean	0.04	0.02	0.04	12498.48	44.58	3.28	86.25
	Std Dev	0.01	0.01	0.01	759.41	3.56	1.05	0.84
3	Mean	0.03	0.03	0.07	11177.28	40.1	3.73	83.23
	Std Dev	0.01	0.01	0.02	1210.54	3.14	1.2	0.75
4	Mean	0.05	0.02	0.01	7508.83	34.8	5.73	80.55
	Std Dev	0.01	0.02	0.01	2315.56	6.86	0.85	2.14
5	Mean	0.03	0.09	0.02	5321.98	33.03	4.23	82.9
	Std Dev	0.01	0.01	0.01	2416.64	10.31	1.8	2.14

5. Literatúra

[1] Stankovičová, Iveta: Viackriteriálne hodnotenie zamestnanosti členských krajín EÚ na základe vybraných ukazovateľov Lisabonskej stratégie. In: STAKAN 2007- Sborník příspěvku. ISBN 978-80-87106-07-5, s.162- 177., dostupné z <http://www.statspol.cz/stakan/stakan-v-2007/stakan2007.pdf>

[2] Stankovičová, Iveta – Vojtková, Mária: Viacrozmerné štatistické metódy s aplikáciami. Bratislava: IURA Edition, 2007. 261 s. ISBN 978-80-8078-152-1

6. Zdroje údajov

[1] Eurostat Home page: <http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1090,30070682,1090_33076576&_dad=portal&_schema=PORTAL>

[2] UNESCO Institute for Statistics: <http://www.uis.unesco.org/ev_en.php?ID=6513_201&ID2=DO_TOPIC>

Adresa autora :

Jana Lacková, Bc.
FMFI a FM UK v Bratislave, 4. ročník
Mládežnícka 7
935 21 Tlmače
Jana.lackova1@gmail.com

Behaviorálne kreditné hodnotenie
Behavioral Credit Scoring

Marián Magna

Abstract: The object of our project was to realize credit scoring for corporate clients of one Slovak commercial bank. On the basis of input data and their character we decided to accomplish the behavioral credit scoring. Our effort was to build a model, which will be logically interpretable as well as statistically strong.

Key words: credit scoring, loans, SAS Enterprise Miner, logistic regression, scorecard, Gini coefficient, ROC curve

Kľúčové slová: kreditné hodnotenie, úvery, SAS Enterprise Miner, logistická regresia, skórovacia karta, Gini koeficient, ROC krivka

1. Úvod

V dnešných časoch, keď nielen banky čelia následkom svojich nesprávnych rozhodnutí z minulosti sa hlási k slovu otázka dôkladnejšieho preverovania budúcich dlžníkov pri žiadaní o úver. I keď súčasnú ekonomickú krízu priniesol práve laxný prístup spoločností pôsobiacich na trhu s hypotekárnymi úvermi pre širokú verejnosť, nemenej pozornosti si zaslúži aj kreditné hodnotenie firemných zákazníkov.

Kreditné hodnotenie, v angličtine nazývané credit scoring, je proces, prostredníctvom ktorého sa určuje miera rizika pre potenciálneho (alebo existujúceho) klienta. Kreditné hodnotenie sa využíva najviac v bankovom sektore, avšak metodológia kreditného hodnotenia je rovnako dobre použiteľná aj v poisťovníctve (pri odhaľovaní potenciálnych podvodníkov) v úverových spoločnostiach nebankového charakteru, u lízingových spoločností, vládnych organizácií ako aj u spoločností poskytujúcich telekomunikačné služby.

Kreditné hodnotenie samotné je založené na štatistických výpočtoch, vychádzajúcich z údajov, ktoré má poskytovateľ služby k dispozícii. Informácie o žiadateľoch môžu poskytovatelia služieb získať z viacerých zdrojov. Najjednoduchšie dostupné sú interné údaje o zákazníkoch, ktorí využívajú niektoré produkty poskytovateľa (v prípade banky je to napríklad bežný podnikateľský účet, kreditná karta, ...). Tieto údaje sú nie vždy postačujúce, preto sa nezriedka poskytovatelia dopytujú žiadateľov o dodatočné údaje (napríklad veľkosť firmy, počet zamestnancov, finančné ukazovatele, vek firmy, odvetvie, osobné údaje o manažmente, právna forma spoločnosti,...). Okrem toho, poskytovatelia služieb sa zaujímajú aj o existujúce záväzky budúcich dlžníkov (splátky iných úverov, lízingové splátky, splátky kreditných kariet, záväzky voči dodávateľom,...). Na základe týchto údajov je možné vykonať kreditné hodnotenie žiadateľov o úver.

2. Ciele práce

Cieľom práce je na základe operatívnych dát poskytnutých bankou vytvoriť taký model, podľa ktorého by bolo možné čo najpresnejšie odhadnúť riziko, za akých podmienok sa zákazník dostane do situácie, že nedokáže splácať svoje záväzky. Inak povedané, aký typ zákazníka na základe stanovených premenných nebude schopný splácať svoje záväzky.

Celkový cieľ práce je možné rozdeliť na čiastkové ciele:

- vykonať podrobnú analýzu všetkých premenných (nepotrebné premenné vylúčiť),
- vytvoriť spôsob, akým sa môžu nahradiť alebo odstrániť chýbajúce hodnoty, prípadne zvážiť možnosť pracovať s nimi,

- vytvoriť skórovaciu kartu použitím najvhodnejšieho modelu a so zohľadnením špecifík vstupných dát,
- vyhodnotiť vytvorenú skórovaciu kartu na základe logickej interpretovateľnosti ako aj na základe štatistických ukazovateľov.

3. Dátový súbor

Zdrojové dáta sa týkali korporátnej klientely istej slovenskej banky. Jednalo sa o interné dáta existujúcich klientov, kde každý záznam predstavoval jeden úverový obchod (hypotekárny úver alebo spotrebný úver). Dáta boli verejne dostupné v roku 2007 ako podklad k súťaži organizovanej touto bankou. Obsahovali niekoľko druhov premenných, z ktorých sme vybrali len behaviorálne premenné zhromažďované v informačnom systéme banky počas uplynulých dvoch rokov.

Dátový súbor obsahoval 15 007 štatistických jednotiek (klientov) a nasledovné skupiny premenných:

- premenné týkajúce sa počtu platieb splátok úveru – 9 premenných,
- premenné týkajúce sa dlžných súm v omeškaní v uplynulom období – 20 premenných,
- premenné týkajúce sa počtu omeškaní splátok v uplynulom období – 25 premenných,
- jedna cieľová premenná, ktorá je binárna.

4. Postup analýzy

V prvom kroku sme si analyzovali všetky premenné. Zaujímalo nás najmä ich distribučné rozdelenie, formáty, minimá, maximá a priemery (prípadne mediány a modusy). Z počiatočnej analýzy sme zistili, že dátový súbor je „zašumený“ a v niekoľkých premenných sú chýbajúce hodnoty. Chýbajúce hodnoty sme nahradili priemerom alebo mediánom podľa povahy premennej.

V nasledujúcom kroku sme načítali vstupný súbor vo formáte MS Excel do programu SAS 9.1. Každej premennej sme priradili adekvátny formát. Samotnú analýzu sme sa rozhodli vykonať v prostredí SW SAS v module Enterprise Miner 5.3, ktorý je vhodným nástrojom pre riešenie tohto problému. Ako prvé sme si vytvorili nový projekt, do ktorého sme si vložili upravený dátový súbor.

V programe SAS Enterprise Miner sme si vytvorili pracovný diagram s nasledovnými uzlami:

- uzol „Input Data“ slúži na zadefinovanie dátového súboru,
- uzol „Impute“ slúži na nahradenie chýbajúcich hodnôt v štatistickom súbore,
- uzol „Data Partition“ slúži na rozdelenie štatistického súboru na dve časti – tréningovú 55%) a validačnú (45%),
- uzol „Interactive Grouping“ slúži na vytvorenie hraníc v každej premennej,
- uzol „Scorecard“ slúži na vytvorenie modelu a zostavenie výslednej skórovacej karty.

Pracovný diagram (flow diagram) je znázornený na obrázku (Obrázok 5) a vyjadruje postup analýzy.

Obrázok 5: Pracovný diagram pre model s premennými pre behaviorálnu skórovaciu kartu

V uzle „Scorecard“ sme vytvorili hodnotiaci model pomocou logistickej regresie. Logistická regresia slúži na predikovanie výsledkov dvoch úrovní závislej premennej. Ako

závislá premenná v modeli vystupuje kategoriálna premenná (v našom prípade binárna cieľová premenná; nadobúda hodnoty 0 v prípade, že sa klient nedostal do platobnej neschopnosti a hodnotu 1, ak sa dostal do platobnej neschopnosti) a ako vysvetľujúce premenné sú intervalové premenné (12 premenných z uzla „Interactive Grouping Beh“).

Vo výslednej skórovacej karte (Tabuľka 5) zostalo nasledujúcich 5 premenných:

- Celkový počet dní v omeškaní za posledných 24 mesiacov
- Najvyšší počet dní v omeškaní za posledných 24 mesiacov
- Počet omeškaní 11 až 30 dní za posledných 6 mesiacov
- Počet omeškaní 31 až 60 dní za posledných 6 mesiacov
- Celkový počet omeškaní za posledných 6 mesiacov

Tabuľka 5: Výsledná skórovacia karta

Premenná	**Interval hodnôt**	**Bodové hodnotenie (Scorecard Points)**
Celkový počet dní v omeškaní za posledných 24 mesiacov	0	40
	1-7	35
	8-30	33
	31-60	20
	61-120	15
	>120	2
Najvyšší počet dní v omeškaní za posledných 24 mesiacov	0	35
	1-7	31
	8-14	28
	15-30	22
	>30	10
Počet omeškaní 11 až 30 dní za posledných 6 mesiacov	0	29
	1-3	22
	>3	17
Počet omeškaní 31 až 60 dní za posledných 6 mesiacov	0	34
	>0	4
Celkový počet omeškaní za posledných 6 mesiacov	0	48
	1-3	22
	4-6	1
	>6	-24

Ako na prvý pohľad vidieť, vo výslednej skórovacej karte sa objavili premenné, ktoré súvisia s časovým rozpätím 6 a 24 mesiacov. V Tabuľka 1 je zobrazená výsledná skórovacia karta. Skórovacia karta v takejto forme je priamo použiteľná pre praktické hodnotenie existujúcich zákazníkov na základe ich správania.

5. Vyhodnotenie modelu

Pre overenie kvality a stability modelu sa používa niekoľko štatistík. Jednou z nich sú hodnoty Giniho koeficientov pre každú z významných premenných, ktoré sú vo výslednej

skórovacej karte (Tabuľka 6). Koeficienty Giniho indexu by mali byť čo najbližšie k hodnote 100 (ak sa jedná o percentá), prípadne k hodnote 1.

Okrem Giniho koeficientov pre každú jednotlivú premennú sú dôležité aj hodnoty Giniho koeficientov aj pre celú skórovaciu kartu. Navyše, pre celú skórovaciu kartu nás zaujímajú aj iné štatistiky, ako napríklad Average Square Error, prípadne Mean Square Error, ktoré sú pre tréningový a validačný súbor uvedené v Tabuľka 7.

Tabuľka 6: Gini koeficienty

Premenná	**Gini koeficient**
Celkový počet dní v omeškaní za posledných 24 mesiacov	88,857
Najvyšší počet dní v omeškaní za posledných 24 mesiacov	88,269
Počet omeškaní 11 až 30 dní za posledných 6 mesiacov	86,081
Počet omeškaní 31 až 60 dní za posledných 6 mesiacov	78,405
Celkový počet omeškaní za posledných 6 mesiacov	87,352

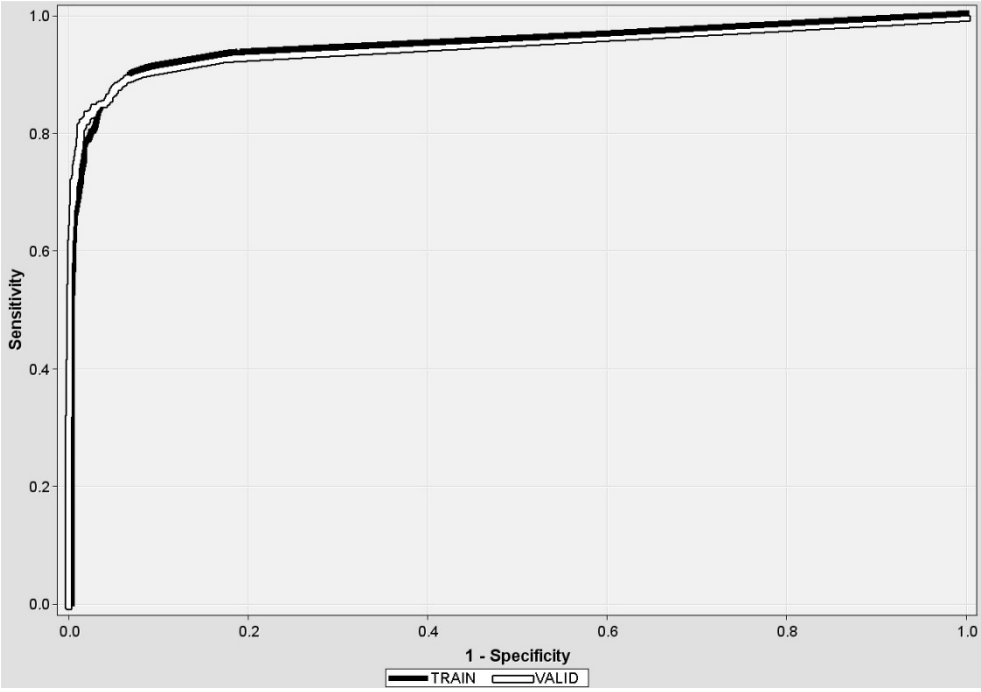
Hodnoty Misclassification Rate, Average Square Error, prípadne Mean Square Error vyjadrujú chybovosť modelu, preto je potrebné aby boli čo najnižšie, tzn. aby sa čo najviac blížili k nule. Hodnoty Giniho koeficientu by v tomto prípade mali byť čo najbližšie k hodnote 1. Okrem toho platí, že pre tréningovú a validačnú sadu by mali byť jednotlivé hodnoty štatistík pre tréningovú aj validačnú sadu čo najbližšie pri sebe, v ideálnom prípade by mali byť rovnaké.

Tabuľka 7: Hodnotiace štatistiky

Názov štatistiky	**Tréningový súbor**	**Validačný súbor**
Misclassification Rate	0,0236	0,0199
Average Squared Error	0,0185	0,0161
Mean Square Error	0,0185	0,0161
Gini Coefficient	0,9165	0,9191

Ďalším spôsobom, ako hodnotiť kvalitu a stabilitu modelu je pohľad na ROC krivku. ROC krivka reprezentuje vzťah medzi hodnotami špecificity (x-ová os: 1-Specificity) a senzitivity (y-ová os: Sensitivity). Krivka smeruje vždy z ľavého dolného rohu do pravého horného rohu, pričom čím je väčší obsah plochy pod krivkou, tým je model stabilnejší.

Ako vidíme na obrázku (***Obrázok 6***), obsah plochy pod ROC krivkou je veľmi blízko k číslu 1 (1x1), a teda možno konštatovať, že model dosahuje veľmi dobrú stabilitu.



Obrázok 6: ROC krivka

6. Záver

Analýzou a vytvorením skóringového modelu nám vznikla jednoduchá metodika pre hodnotenie potenciálnej kvality existujúcich bankových zákazníkov. Z pôvodných 55 premenných vstupujúcich do analýzy nám zostalo päť významných premenných, na ktorých sme postavili model. Všetkých päť premenných pochádza zo skupiny 25 premenných, ktoré sa týkali počtu omeškaní splátok v uplynulom období. Tieto premenné, i keď sú z jednej skupiny, dokážu sami o sebe poskytnúť relatívne dobrú vypovedaciu schopnosť o charaktere dlžníka.

Na základe vyhodnocovacích štatistík je zrejmé, že model dosahuje dobrú stabilitu. Rovnako, aj hodnoty jednotlivých skóre pre každú premennú sú primerané, čo tiež dokazuje reálnosť modelu.

7. Literatúra

[1]SIDDIQI NAEEM: Credit Risk Scorecard, Developing and Implementing Intelligent Credit Scoring, Wiley Publishing, 2006, ISBN: 978-0-471-75451-0

[2]Scorto Credit Scoring Solutions [online] dostupné na internete:
http://www.scorto.com/credit_scoring.htm

[3]Applied Analytics Using SAS Enterprise Miner Software 5 Course Notes, ISBN: 978-1-59994-515-6

Adresa autora:

Marián Magna, Bc. et Bc., 5. ročník
Fakulta managementu, Univerzita Komenského,
Odbojárov 10, Bratislava
mariaan.magna@gmail.com

Projekcia indexu dôchodkovej závislosti v SR
Projection of a dependency pension ratio in SR

Jana Maniačková

Abstract: Using Leslie model as discrete dynamical system to describe the population dynamics of the female portion of closed species, we have the projection of population of SR to the year 2106. From the people in active age (from 22 years to the pension age) work just 73,3 %, but all people older than pension age have claim for retiring pay. While average retiring pay is 44 % and pension contribution is 18 % of the average gross earnings, ideal dependency pension ratio defined as a ration of those who pay pension contribution and those whom retiring pay is paid to is 2,44. Dependency pension ratio curve for different pension ages (from 60 to 68) quickly (from 4 to 40 years) fall under ideal dependency pension ratio and assuming the constant TFR it is being fixed between 0,93 and 1,83 in the time.

Key words: Leslie model, population projection, dependency pension ratio, pension system.

Kľúčové slová: Leslieho model, predikcia populácie, index dôchodkovej závislosti, dôchodkový systém.

1. Úvod

Nepriaznivý demografický vývoj na Slovensku, ale aj v celej Európe, vyvoláva nutnosť zavádzania zmien v penzijnom zabezpečení [3], [4]. Cieľom článku je pomocou diskrétneho deterministického Leslieho modelu získať projekciu populácie v jednotlivých ročných skupinách. Skúmame vývoj pomeru medzi prispievateľmi a poberateľmi dôchodkových dávok. Ukážeme, že zmena dôchodkového veku musí byť nevyhnutne sprevádzaná aj zmenou plodnosti, úmrtnosti, výšky dôchodkového odvodu, či zamestnanosti pre udržanie tohto pomeru na prijateľnej úrovni.

2. Projekcia počtu obyvateľov SR

Na popis časovej zmeny vo veľkosti populácie SR použijeme diskrétny deterministický model, ktorý uvažuje vekové zloženie populácie s dĺžkou intervalu medzi jednotlivými pozorovaniami populácie 1 rok. Uvažujeme uzavretú populáciu. Metódou posúvania vekových skupín v čase $t = 0, 1, ..., 100$ (roky 2006 až 20106) prichádza k presunu z mladších do starších vekových skupín a počiatočná veková skupina sa vytvára na základe prirodzenej populačnej reprodukcie. O vlastnosti deterministického modelu hovorí Lotkov zákon latencie (Veta o silnej ergodicite): Pri nezmenených mierach pôrodnosti a úmrtnosti sa veková štruktúra asymptoticky blíži do ustáleného stavu, ktorý nezávisí na počiatočnej vekovej štruktúre [2]. Použitím modelu populačného rastu so ženskou dominanciou, teda počet narodených detí závisí na počte žien. Mužská časť populácie sa pomocou prislúchajúceho pomeru odvodzuje z chovania ženskej populácie. Predpokladáme, že pôrodnosť a úmrtnosť sa pre jednotlivé veky v uvažovanej populácii v čase nemení a maximálny vek dožitia je 99 rokov ($x = 0, 1, ..., 99$), kde vek x rokov je totožný s vekových intervalom $\langle x; x+1)$. Ide však o predpoveď konštruovanú pomocou matematických metód odvodených od matematických predpokladov a preto idealizuje skutočnosť, nakoľko uvažujeme uzavretú populáciu, v ktorej nedochádza k migrácii, neuvažujeme sociálno-ekonomické aspekty vývoja populácie a iné.

Pri projekcii sme použili úmrtnostné tabuľky VDC SR vychádzajúce z údajov pre rok 2006 o počte žijúcich, pravdepodobnosti úmrtia v ročných vekových skupinách (q_x) osobitne pre ženskú aj mužskú časť populácie a o špecifickej miere plodnosti (F_x) [8] v SR.

Pre špecifickú mieru reprodukcie platí:

$$f_x^f = \frac{L_0^f}{2 \cdot l_0^f} \cdot \frac{1}{1 + SRB} \cdot (\mathrm{F_x} + \mathrm{s_x} \cdot \mathrm{F_{x+1}}), \tag{1}$$

kde SRB je sekundárny index maskulinity a predpokladajme, že jeho výška je 1,062. f_x^f je špecifická miera reprodukcie vo veku x (počet dcér narodených za rok ženám vo veku $\langle x; x+1)$). s_x je podiel x - ročných žien v čase t, ktoré sa dožijú veku $x+1$ v čase $t+1$. Reprodukcia mužskej populácie mužov je odvodená od reprodukcie ženskej populácie:

$$f_x^m = SRB \cdot f_x^f . \tag{2}$$

3. Leslieho model v maticovom zápise

Leslieho model používa maticový zápis pre zjednodušenie formálneho zápisu kohorentno komponetnej metódy pri konštrukcii populačnej projekcie.

Označíme K_{xt} počet žien, ktoré v čase t prislúchajú vekovému intervalu $\langle x; x+1)$. Hodnoty K_{xt} budeme združovať do 100 - rozmerných stĺpcových vektorov. Zatiaľ známy je vektor v čase $t = 0 : K_x = (\, K_{0t}, K_{1t}, ..., K_{99\,t}\,)^T$, ktorý tvoria údaje o počte žien v jednotlivých vekových skupinách k 1.1.2006. Predpokladáme, že projekčné koeficienty s_x nezávisia na čase t, teda sa v čase nemenia a tak platia lineárne rekurentné vzťahy, ktoré popisujú dynamiku prechodu uvažovanej populácie medzi časovými okamžikmi t a $t+1$:

$$K_{x+1,t+1} = s_x \cdot K_{xt}, \tag{3}$$

$$K_{0,t+1} = \sum_{x=0}^{99} f_x K_{xt} , \tag{4}$$

čo formálne zapíšeme pomocou matice L (tzv. Leslieho matica) [1]:

$$\begin{pmatrix} K_0^{t+1} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ K_{99}^{t+1} \end{pmatrix} = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{98} & f_{99} \\ s_0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & s_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & s_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & & s_{98} & 0 \end{pmatrix} \bullet \begin{pmatrix} K_0^t \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ K_{99}^t \end{pmatrix}. \tag{5}$$

Lineárne rekurentné vzťahy sa dajú pomocou matíc zapísať:

$$K_{t+1} = L \cdot K_t . \tag{6}$$

Pomocou indukcie teda platí vzťah označujúci Leslieho model:

$$K_{t+n} = L^n \cdot K_t \quad ,$$

(7) kde K_0 je počiatočné vekové zloženie v čase $t = 0$, teda pri dostatočne veľkom n bude veková štruktúra K_t konštantná a aj populačný rast v každom projekčnom intervale bude konštantný. Uvedená úvaha vedie k stabilnej populácii, pričom maticová algebra ponúka elegantný spôsob ako odvodiť simuláciu stabilnej populácie:

$$K_{t+1}^s = L \cdot K_t^s = \lambda \cdot K_t^s , \tag{8}$$

kde nárast za 1 rok λ je dominantná vlastná hodnota Leslieho matice [6]. Vlastné hodnoty Leslieho matice sú počítané v programe R. Tempo populačného rastu je teda *0,982,* konkrétne ak $0 < \lambda < 1$ tak sa jedná o pokles populácie [1]. V stabilnej populácii bude teda veľkosť populácie o sto rokov λ^{100} = 15,9 % z veľkosti populácie v roku 2006. V projekcii sa populácia znížila na 26,8 % z pôvodnej veľkosti, nakoľko sme nevychádzali zo stabilnej populácie, ale z reálnych dát z roku 2006.

Maticovo pomocou vlastného vektora Leslieho matice možno zapísať:

$$K_{t+1}^{s} = (L - \lambda \cdot I) \cdot K_{t}^{s} = 0, \tag{9}$$

kde I je jednotková matica a 0 je stĺpcový nulový vektor [6]. Pri aplikáciách sa pracuje s prirodzeným logaritmom vlastnej hodnoty λ, ktorý vyjadruje vnútornú mieru prirodzeného prírastku, tzv. Lotkovu mieru $r = ln(\lambda) = -0{,}018$ [1].

Periodicita chovania sa populácie sú cyklicky sa opakujúce vlny, ktorých intenzita postupne slabne a každá populačná nepravidelnosť sa prejaví slabnúcim účinkom v nasledujúcich generáciách. Pre periodicitu populačných vĺn $2\pi/q$ závislú na druhej, resp. treťej vlastnej hodnote $\lambda^{T}_{2,3} = 0{,}9345732 \pm 0{,}2043354$i určenej ako $\alpha \pm \beta i$ platí:

$$\frac{2\pi}{q} = \frac{2\pi}{\arcsin\left(\frac{\beta}{\sqrt{\alpha^2 + \beta^2}}\right)} = 29{,}190, \tag{10}$$

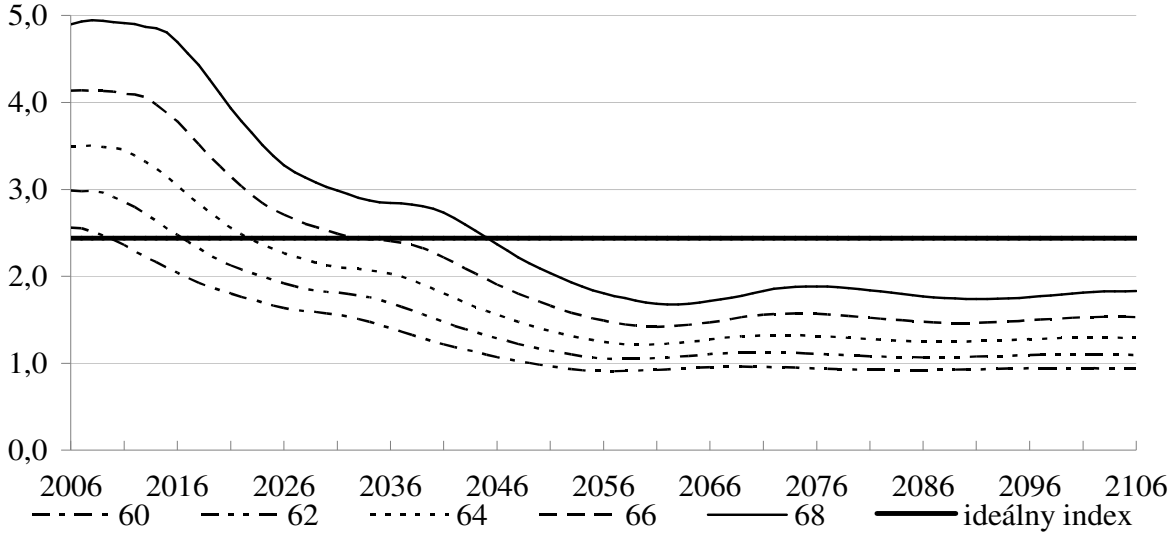
čiže dĺžka jednej generácie je 29 rokov [1].

4. Index dôchodkovej závislosti

Index dôchodkovej závislosti definujeme ako pomer pracujúcich (t. j. prispievateľov do dôchodkového systému) ku poberateľom dôchodku. Uvažujem, dovŕšenie veku 22 rokov ako vstup do produktívnej populácie a výstup z nej pri dovŕšení veku 62 rokov jednotne pre mužov a ženy, teda osoba sa nachádza produktívnom veku 40 rokov. Počas tejto doby prispieva svojimi odvodmi len časť osôb v produktívnom veku, avšak dôchodkové dávky poberá 100% ľudí, ktorí sa nachádzajú v dôchodkovom veku. Počet pracujúcich v hospodárstve SR spolu za rok 2006 je 2 301,4 tisíc osôb (vrátane osôb na riadnej materskej dovolenke, profesionálnych príslušníkov ozbrojených zložiek a pracujúcich v zahraničí) [7]. Hoci profesionálni príslušníci ozbrojených zložiek majú osobitný dôchodkový systém, nebudeme ich uvažovať osobitne. Počet pracujúcich osôb závisí od rôznych sociálno-ekonomických faktorov, preto sa pomer v budúcnosti môže meniť, avšak pre daný model ho uvažujeme ako konštantný. Podiel pracujúcich ku osobám v produktívnom veku pre rok 2006 je nasledujúci: $\frac{2301400}{3138898} = 0{,}733$. 73% ľudí v produktívnom veku prispieva svojimi odvodmi a dôchodkové dávky bude poberať 100% tých, ktorí sa dožijú dôchodkového veku.

V hospodárstve SR (rok 2006) bola priemerná hrubá mesačná mzda 18761 Sk a dôchodok 8226 Sk (t. j. 44% z výšky hrubej mzdy) [7]. Na výplatu jednej dôchodkovej dávky je teda treba 2,44 pracujúcich, ktorí na dôchodkové dávky prispievajú 18% z hrubej mzdy, čo hovorí, že pre daný model je ideálny index dôchodkovej závislosti vo výške 2,44.

Zvýšenie dôchodkového veku (o 2 roky) na 62 rokov spôsobilo, že do dôchodkového systému bude prispievať o 73170 pracujúcich viac (nárast o 3,28 %) a počet dôchodcov sa zníži o 99797 (pokles o 11,47 %) v porovnaní s dôchodkovým vekom 60 rokov. Index dôchodkovej závislosti sa vďaka tomu zvýšil z 2,56 na 2,99, avšak v čase sa tento index značne znižuje, teda je prirodzené predpokladať, že dôchodkový vek sa bude postupne zvyšovať, nakoľko práve v najbližších rokoch prichádza k presunu silných generačných ročníkov do poproduktívnej generácie. Keďže v modeli sa nemenilo aktuálne nastavenie parametroch v čase, tak je nereálne dosiahnuť ustálenie indexu dôchodkovej závislosti v ideálnej výške len zmenou dôchodkového veku. Výsledky výpočtov sú znázornené na grafe č. 1.



Graf 7: Index dôchodkovej závislosti v SR pri zmene dôchodkového veku (Na osi y je znázornený celkový počet pracujúcich na jedného dôchodcu, na osi x je čas, t. j. roky)

5. Záver

Stabilizácia indexu dôchodkovej závislosti na udržateľnej hladine 2,44 sa podľa štúdie pri stálosti parametrov použitých pri výpočtoch ukazuje ako nerealizovateľná. Na krátkodobé vykrytie strát spôsobených prechodom silných populačných ročníkov do poproduktívneho veku je postačujúce zvýšenie dôchodkového veku na 62 rokov, avšak z dlhodobého hľadiska je posúvanie dôchodkového veku nepostačujúce a najmä nereálne vzhľadom na ideálny index dôchodkovej závislosti 2,44. Aktuálny demografický vývoj značne ovplyvňuje rovnováhu medzi generáciami, preto sú v penzijnom systéme potrebné viaceré zmeny a to nie len zvýšenie dôchodkového veku (ale aj zmena plodnosti, úmrtnosti, výšky dôchodkového odvodu, zamestnanosti...).

6. Literatúra

[1]CIPRA, T. 1990. Matematické metódy v demografii a pojištění. Praha: SNTL, 1990.

[2]KEYFITZ, N. 1977. Introduction to the Mathematics of Population with Revisions. Reading, Mass.: Addison - Wesley, 1977.

[3]KOMISIA EURÓPSKYCH SPOLOČENSTIEV, KOM (2005) 94 v konečnom znení: Zelená kniha „Ako čeliť demografickým zmenám, nová solidarita medzi generáciami", Brusel, 2005.

[4]KOMISIA EURÓPSKYCH SPOLOČENSTIEV, KOM (2006) 244 v konečnom znení: Demografická budúcnosť Európy – pretvorme výzvu na príležitosť, Brusel, 2006.

[5]MANIAČKOVÁ J., 2008. Bakalárska práca: Prognózy hospodárenia dôchodkového systému SR. FMFI UK, Bratislava, 2008.

[6]PRESTON, S.H., HEUVELINE, P., GUILLOT, M. 2001. Measuring and Modeling Population Processes. Oxford: Blackwell, 2001.

[7]ŠTATISTICKÝ ÚRAD SR. http://portal.statistics.sk/showdoc.do?docid=1642

[8]VÝSKUMNÉ DEMOGRAFICKÉ CENTRUM. http://www.infostat.sk/slovakpopin/

Adresa autora:

Bc. Jana Maniačková
Dlhá 44, 949 01 Nitra
maniackova@gmail.com

Porovnanie krajín Európskej únie na základe ukazovateľov výskytu HIV a AIDS
European union´s member states comparison by the incidence of HIV and AIDS indicators

Martina Mihalíková

Abstract: European union´s memberstate´s (15 states) comparison by the incidence of HIV and AIDS indicators (growth index of AIDS, number of HIV incidence, HIV positive drug use, number of prisoner´s, incidence of tuberculosis) per 100 000 population at the age 15-64 years during 2000-2005. Memberstate´s comparison is realized by the method of principal components and cluster analysis.

Key words: AIDS, HIV, European union, method of principal components, cluster analysis

Kľúčové slová: AIDS, HIV, Európska únia, metóda hlavných komponentov, zhluková analýza

1. Úvod

AIDS je ochorenie zapríčinené vírusom HIV, ktorý napáda ľudský imunitný systém tela. Prvé infekcie HIV boli zaznamenané v 30. rokoch v Afrike, v súčasnosti je infikovaných až 40 miliónov ľudí na celom svete. Skupiny vystavené najväčšiemu riziku infikovania vírusom HIV zahŕňajú užívateľov injekčných drog, mužov, ktorí majú pohlavný styk s mužmi, sexuálnych pracovníkov, prisťahovalcov, väzňov a mladých ľudí do 25 rokov. Vývoj HIV na AIDS urýchľuje i tuberkulóza.
Cieľom práce je porovnať krajiny Európskej únie na základe ukazovateľov výskytu HIV a AIDS s využitím vybraných viacrozmerných štatistických metód (metódy hlavných komponentov a zhlukovej analýzy).
Na analýzu použijeme program MS Excel a SAS® Enterprise Guide 4.0

2. Popis a výber vhodných ukazovateľov

Analýza je uskutočnená na 25 krajinách Európskej únie – Rakúsko, Belgicko, Cyprus, Česká republika, Dánsko, Estónsko, Fínsko, Francúzsko, Nemecko, Grécko, Maďarsko, Írsko, Taliansko, Lotyšsko, Litva, Luxembursko, Malta, Holandsko, Poľsko, Portugalsko, Slovensko, Slovinsko, Španielsko, Švédsko a Veľká Británia (Spojené kráľovstvo).

Väčšina údajov je z roku 2005, v niektorých prípadoch z intervalu rokov 2000 až 2005. Veková skupina, na ktorej boli ukazovatele prepočítavané je 15 – 64 rokov.

Vzhľadom na dostupnosť údajov je vytvorených nasledujúcich 9 ukazovateľov pôsobiacich na výskyt HIV a AIDS, ktoré môžeme zoradiť do troch skupín:

Prvú skupinu predstavujú ukazovatele AIDS, druhú skupinu ukazovatele HIV a tretiu skupinu ukazovatele počtu väzňov a výskytu TBC.

Pred výberom štatistických metód viacrozmerného porovnávania, analyzujeme vzťahy medzi premennými. Vzhľadom na relatívne silné vzťahy medzi dvojicami premenných, použijeme metódu hlavných komponentov. Z analýzy vylúčime premenné KrHIV, MrtviAIDS, PocetAIDS.

Tabuľka 1*: *Premenné použité v analýze

KrAIDS	Koeficient rastu výskytu AIDS na 100 000 obyvateľov v období r. 2000-2005
PocetAIDS	Počet prípadov AIDS na 100 000 obyvateľov vo veku 15-64 r.
MrtviAIDS	Počet zomrelých na AIDS na 100000 obyvateľov vo veku 15-64 r.
PocetHIV	Počet prípadov HIV-pozitívnych na 100000 obyvateľov vo veku 15-64 r.
NovoinfHIV	Novoinfikovaní užívatelia injekčne aplikovateľných drog na 100 000 obyvateľov vo veku 15-64r.
KrHIV	Koeficient rastu výskytu HIV na 100 000 obyvateľov v období r. 2000-2005
PocetVAZNI	Počet väzňov na 100 000 obyvateľov vo veku 15-64 r.
VyskytTBC	Výskyt tuberkulózy na 100 000 obyvateľov vo veku 15-64 r.

Tabuľka 2: Párové koeficienty korelácie ukazovateľov použitých pri analýze

	KrAIDS	**PocetVAZNI**	**NovoinfHIV**	**VyskytTBC**	**PocetHIV**
KrAIDS	1.00000	0.63045 0.0007	0.82787 <.0001	0.45182 0.0234	0.77793 <.0001
PocetVAZNI	0.63045 0.0007	1.00000	0.65401 0.0004	0.81268 <.0001	0.50755 0.0096
NovoinfHIV	0.82787 <.0001	0.65401 0.0004	1.00000	0.59137 0.0018	0.91781 <.0001
VyskytTBC	0.45182 0.0234	0.81268 <.0001	0.59137 0.0018	1.00000	0.37543 0.0644
PocetHIV	0.77793 <.0001	0.50755 0.0096	0.91781 <.0001	0.37543 0.0644	1.00000

Na hladine významnosti α=0,01 sú všetky premenné štatisticky závislé a vhodné na použitie metódy hlavných komponentov.

3. Použitie metódy hlavných komponentov

Pri voľbe počtu hlavných komponentov zohľadňujeme vlastné čísla korelačnej matice a podiel vysvetlenej variability. Na základe týchto dvoch kritérií sme vybrali prvé dva hlavné komponenty, ktoré vysvetľujú až 90,82 % celkovej variability údajov.

Tabuľka 3 Vlastné čísla

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	3.63746725	2.73385685	0.7275	0.7275
2	0.90361040	0.63964778	0.1807	0.9082
3	0.26396262	0.11353832	0.0528	0.9610
4	0.15042430	0.10588888	0.0301	0.9911
5	0.04453542		0.0089	1.0000

Pomocou tabuľky vektora vlastných čísel (výstupy neuvedené z dôvodu obsahových obmedzení) usudzujeme, že druhý hlavný komponent je ovplyvnený premennými *PocetVAZNI* a *VyskytTBC*. Na hodnoty prvého hlavného komponentu významne vplývajú všetky ostatné premenné.

Vhodnosť výberu premenných potvrdzuje aj celková hodnota Kaiser-Meyer-Olkinovej štatistiky (0,69110266), ktorá prekračuje doporučenú hranicu 0,5.

Na základe hodnôt prvého hlavného komponentu sme krajiny usporiadali.

Z Tabuľky 3 môžeme usúdiť, že najlepšie zo všetkých krajín obstál Cyprus. Táto krajina vykazuje najnižšie hodnoty koeficienta rastu AIDS, novoinfikovaných HIV z injekčne aplikovateľných drog a počtu HIV pozitívnych.

Naopak najhoršie výsledky (najvyššie hodnoty) spomínaných ukazovateľov sú v Estónsku. Slovenská republika sa umiestnila na trinástom mieste.

Tabuľka 4: Usporiadanie krajín na základe prvého hlavného komponentu

Štáty	PRIN1	Poradie	Štáty	PRIN1	Poradie
Cyprus	-0.705264234	1	Francúzsko	-0.317305359	14
Malta	-0.623566192	2	Belgicko	-0.305530735	15
Švédsko	-0.596296201	3	Španielsko	-0.216943649	16
Slovinsko	-0.573426643	4	Maďarsko	-0.202137515	17
Nemecko	-0.56250738	5	Írsko	-0.027674277	18
Dánsko	-0.55805118	6	Veľká Británia	-0.023912535	19
Taliansko	-0.558033686	7	Luxembursko	0.0381647157	20
Fínsko	-0.557021053	8	Poľsko	0.0709353178	21
Grécko	-0.426366218	9	Litva	0.8436114394	22
Holandsko	-0.358176641	10	Portugalsko	0.9953438626	23
Rakúsko	-0.352732861	11	Lotyšsko	1.7279805138	24
Česká republika	-0.332678213	12	Estónsko	3.9437097378	25
Slovensko	-0.322121019	13			

4. Použitie zhlukovej analýzy

Zhlukovanie krajín uskutočníme na základe hodnôt prvých dvoch hlavných komponentov. Pri zhlukovaní použijeme Wardovu hierarchickú metódu zhlukovania. Počet zhlukov sme určili na základe zlomu v grafe hodnôt semi-parciálneho koeficienta determinácie (SPRSQ). Rozhodli sme sa pre šesť zhlukov.

Výsledky zhlukovania využitím Wardovej hierarchickej metódy zhlukovania sú v nasledujúcej tabuľke 5.

Prvý zhluk tvoria zväčša krajiny, ktoré stáli pri vzniku Európskej únie. Vyznačujú sa najnižšími hodnotami *počtu väzňov* a *výskytom tuberkulózy* a relatívne nízkou hodnotou *počtu nakazených vírusom HIV*.

Druhý zhluk tvorí Veľká Británia, Írsko a Luxembursko. Vyznačujú sa nízkym *počtom väzňov*, nízkym *výskytom tuberkulózy* a *novoinfikovaných vírusom HIV* injekčne aplikovateľnými drogami.

Tretí zhluk tvoria krajiny strednej Európy, medzi ktorými je aj Slovenská republika. Tento zhluk krajín sa vyznačuje najnižšími hodnotami *počtu infikovaných a novoinfikovaných vírusom HIV* a nízkym *koeficientom rastu AIDS*. Naopak vysoké hodnoty dosahuje premenná *počet väzňov* a *výskyt tuberkulózy*.

Štvrtý zhluk tvorí Lotyšsko a Litva, ktorý sa vyznačuje najvyššími hodnotami *výskytu tuberkulózy* a vysokými hodnotami *počtu väzňov* a *koeficienta rastu AIDS*.

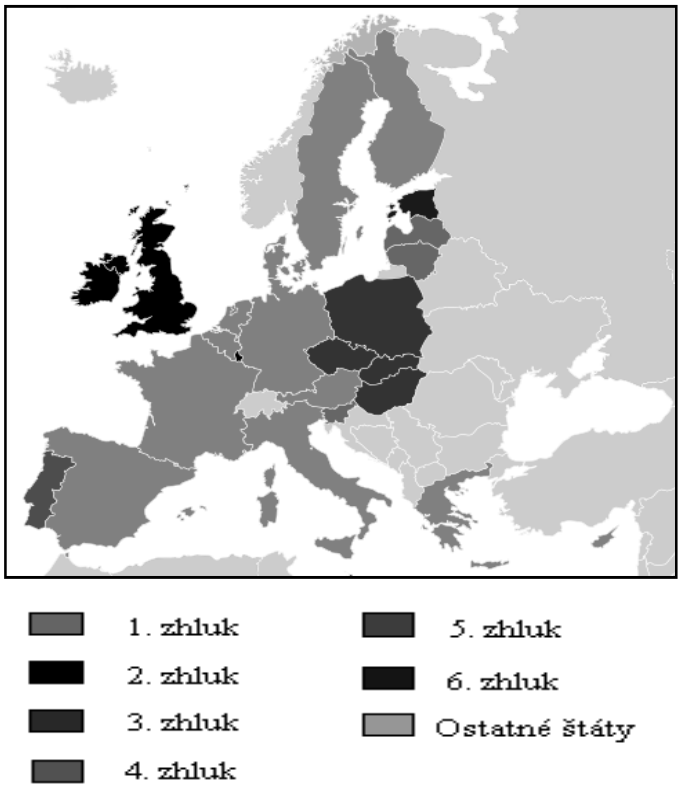
Piaty zhluk predstavuje Portugalsko. Samostatnosť tejto krajiny súvisí s najnižšou hodnotou *koeficienta rastu AIDS* a zároveň najvyšším počtom novoinfikovaných injekčne aplikovateľnými drogami a už nakazených vírusom HIV.

Šiesty zhluk tvorí Estónsko, ktoré si vyslúžilo výnimočnosť tým, že má najvyššie hodnoty *koeficienta rastu AIDS, počtu väzňov, novoinfikovaných HIV* injekčne aplikovateľnými drogami a *počtu HIV pozitívnych* obyvateľov.

Tabuľka 5: Rozdelenie krajín do homogénnych zhlukov

Poradie zhluku	Krajiny patriace do zhluku
Zhluk 1	Rakúsko, Belgicko, Cyprus, Dánsko, Fínsko, Francúzsko, Nemecko, Grécko, Taliansko, Malta, Holandsko, Slovinsko, Švédsko
Zhluk 2	Írsko, Luxembursko, Veľká Británia
Zhluk 3	Česká republika, Maďarsko, Poľsko, Slovensko, Španielsko
Zhluk 4	Lotyšsko, Litva
Zhluk 5	Portugalsko
Zhluk 6	Estónsko

Prehľad vytvorených zhlukov sledovaných krajín je na Obrázku 1.



Obrázok 1: Prehľadná mapa vytvorených zhlukov

5. Záver

V predloženom príspevku sme sa pokúsili o priestorovú diferenciáciu krajín EÚ na základe premenných, ktoré opisujú výskyt HIV a AIDS ochorenia v týchto krajinách. Pri priestorovom porovnávaní sme použili metódu hlavných komponentov a zhlukovú analýzu.

6. Literatúra a ostatné zdroje

[1] Stankovičová, I., Vojtková, M. - Viacrozmerné štatistické metódy s aplikáciami. Bratislava: Iura Edition, 2007

[2] Pažitná, M., Labudová, V. - Metódy štatistického porovnávania. Bratislava : Vydavateľstvo EKONÓM, 2007.

[3] www.who.int

[4] http://epp.eurostat.ec.europa.eu

[5] http://hdr.undp.org/en/statistics/data

Adresa autora:
Martina Mihalíková
FHI, 5. ročník
Ekonomická univerzita v Bratislave

Aproximácia výšky poistných plnení v neživotnom poistení
Approximation of the individual claim amount in non-life insurance

Jana Michaliková

Abstract: The amount of the individual claims is one of the basic information which has to be known by the insurance companies so that they can forecast the trend of the future claims and determine the height of the premium. Therefore it is very important to correctly approximate this data. The paper introduces main distributions with heavy tails which are most suited distribution group for this type of random variables. In addition following work contains an analysis of the concrete individual claims and a brief description of the process how to find the best distribution of the claim amount.

Key words: heavy-tailed distributions, Pareto distribution, lognormal distribution, non-life insurance.

Kľúčové slová: rozdelenia s ťažkými chvostami, Paretovo rozdelenie, lognormálne rozdelenie, neživotné poistenie.

1. Úvod

Oblasť neživotného poistenia je charakteristická tým, že sa so značnou pravdepodobnosťou vyskytujú aj individuálne poistné plnenia s extrémne vysokou hodnotou, čo spôsobuje, že tieto náhodné premenné sú veľmi rozptýlené, aj keď väčšina z nich sa nachádza pod úrovňou strednej hodnoty výšky poistných plnení daného portfólia. Ako najvhodnejšie typy rozdelení na aproximáciu sa preto ukazujú tie, ktoré sú pravostranne zošikmené a konvergencia k nule nie je príliš rýchla. Takéto rozdelenia sa nazývajú rozdelenia s ťažkými chvostami a medzi najznámejšie a najpoužívanejšie patria Paretovo rozdelenie (európska a americká verzia), lognormálne rozdelenie a Weibullovo rozdelenie za určitých ohraničení pre parametre. Zaveďme si teraz pojem chvosta a ťažkého chvosta exaktne a takisto definujme prvé dve zo spomínaných rozdelení, ktoré budeme využívať v našej analýze.

Ak F(x) je distribučná funkcia náhodnej premennej definovaná vzťahom: $F(x) = P\,(X < x)$, potom pravým chvostom označujeme $\overline{F}\,(x) \triangleq P(X > x) = 1 - F(x)$. Ďalej hovoríme, že rozdelenie má ťažký chvost, ak je konvergencia rozdelenia k nule pomalšia ako pri exponenciálnom rozdelení, formálne zapísané, ak pre každé $\lambda > 0$ platí: $\lim_{x\to\infty} e^{\lambda x}\overline{F}\,(x) = \infty$.

Hovoríme, že náhodná premenná X má **európske Paretovo rozdelenie** $\boldsymbol{Pa(\alpha, c)}$, ak hustota pravdepodobnosti má tvar: $f(x) = \frac{\alpha c^{\alpha}}{x^{\alpha+1}}$, kde $\alpha > 0, c > 0, x > c$. Distribučná funkcia tohto rozdelenia je: $F(x) = 1 - (\frac{c}{x})^{\alpha}$, a teda chvost je: $\overline{F}\,(x) = (\frac{c}{x})^{\alpha}$.

Hovoríme, že náhodná premenná X má **americké Paretovo rozdelenie** $\boldsymbol{Pa(\alpha, b)}$, ak hustota pravdepodobnosti má tvar: $f(x) = \frac{\alpha b^{\alpha}}{(b+x)^{\alpha+1}}$, kde $\alpha > 0, b > 0, x > 0$. Distribučná funkcia tohto rozdelenia je: $F(x) = 1 - (\frac{b}{b+x})^{\alpha}$, a teda chvost je: $\overline{F}\,(x) = (\frac{b}{b+x})^{\alpha}$.

Je ľahké ukázať, že ak náhodná premenná X má európske Paretovo rozdelenie $\boldsymbol{Pa(\alpha, c)}$, potom náhodná premenná $X - c$ má Americké Paretovo rozdelenie $\boldsymbol{Pa(\alpha, b)}$, pričom $b = c$.

Hovoríme, že náhodná premenná X má **lognormálne rozdelenie** $\boldsymbol{LN(\mu, \sigma^2)}$, ak náhodná premenná $Y = lnX$ má normálne rozdelenie $N(\mu, \sigma^2)$. Hustota pravdepodobnosti má

tvar: $f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}$, kde $x > 0$. Distribučná funkcia tohto rozdelenia je: $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{lnx} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, dy$, a teda chvost je: $\overline{F}(x) = 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_0^{lnx} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, dy$.

Všetky tieto rozdelenia sú definované len pre kladné hodnoty x, pre $x < 0$ je hustota všetkých spomínaných rozdelení rovná nule. Tento predpoklad je ale v praxi splnený, keďže poistné plnenia nemôžu nadobúdať záporné hodnoty.

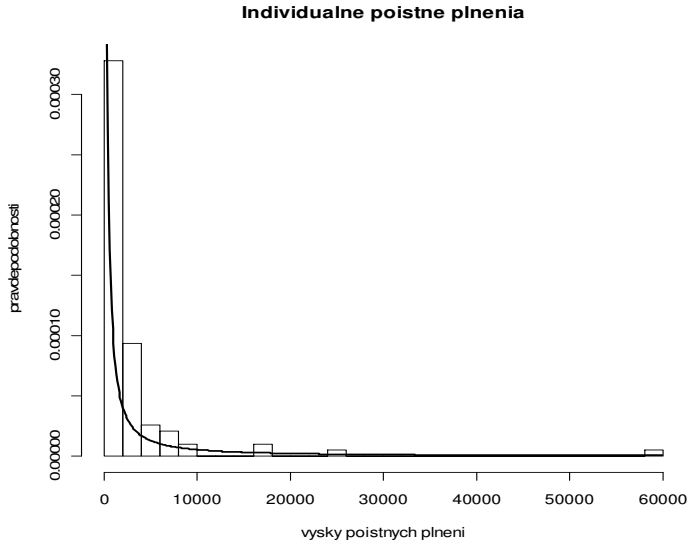
2. Aproximácia výšky poistných plnení Paretovým rozdelením

Aby bol postup aproximácie poistných plnení názornejší, uskutočníme analýzu na konkrétnych dátach, ktoré sú uvedené v tabuľke 1 a predstavujú výšku 96 poistných plnení za vybrané portfólio poistiek v priebehu jedného roka.

Tabuľka 8: Poistné plnenia (v 100 Sk)

24	26	73	84	102	115	132	159	207	240	241	254
268	272	282	300	302	329	346	359	367	375	378	384
452	475	495	503	631	543	563	594	609	671	687	691
716	757	821	829	885	893	968	1053	1081	1083	1150	1205
1262	1270	1351	1385	1498	1546	1564	1635	1671	1706	1820	1829
1855	1873	1914	2030	2066	2240	2413	2421	2521	2586	2727	2797
2850	2989	3110	3166	3383	3443	3515	3521	3531	4068	4527	5006
5065	5481	6045	7003	7245	7477	8738	9197	16370	17605	25318	58524

Voľba najvhodnejšieho rozdelenia na aproximáciu dát sa vo všeobecnosti skladá z troch krokov: výber určitého typu rozdelenia (napríklad na základe grafického zobrazenia alebo popisných charakteristík dát), odhadnutie parametrov tohto rozdelenia jednou zo známych metód (my použijeme metódu maximálnej vierohodnosti) a na záver overenie vhodnosti zvoleného rozdelenia (napr. Kolmogorovým-Smirnovým testom).



Obrázok 8: Histogram rozdelenia výšky poistných plnení a krivka hustoty Paretovho rozdelenia

Na vytvorenie určitej predstavy o možnom rozdelení našich dát použijeme histogram, ktorý je znázornený na obrázku 1. Na grafické porovnanie zhody s Paretovým rozdelením je

do obrázku 1 nakreslená aj krivka hustoty tohto rozdelenia v americkej verzii (európska verzia sa výrazne nelíši).

Ako je z histogramu zrejmé, na aproximáciu týchto dát je vhodné použiť niektoré z rozdelení s ťažkými chvostami, keďže sa tu vyskytuje niekoľko hodnôt výrazne vyšších, ako sú ostatné. Najprv otestujeme zhodu s európskym a americkým Paretovým rozdelením. Toto rozdelenie má dva parametre a na ich odhad použijeme metódu maximálnej vierohodnosti (ďalšie známe metódy sú napríklad metóda momentov alebo metóda kvantilov). Použitím týchto odhadov a Kolmogorovho-Smirnovho testu overíme hypotézu, či naše dáta pochádzajú z Paretovho rozdelenia. Výsledky sú zhrnuté v tabuľke 2.

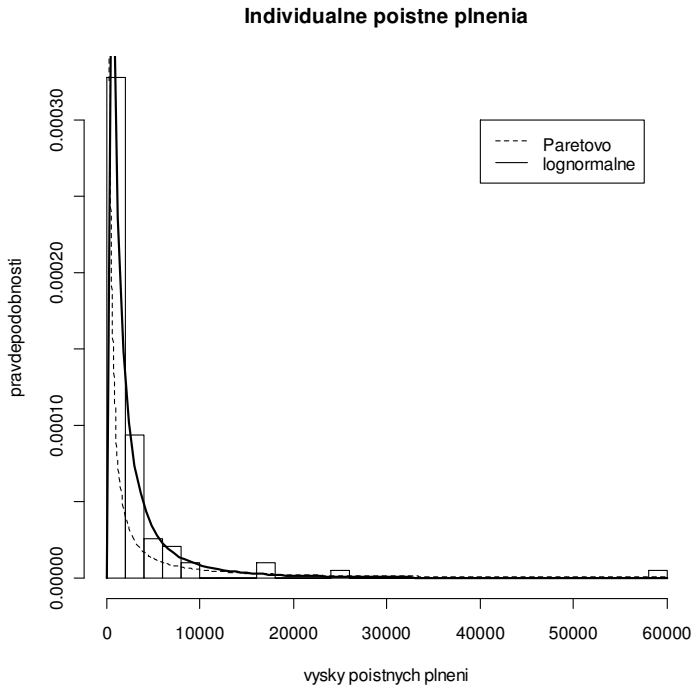
Tabuľka 2: Odhady parametrov Paretovho rozdelenia a p-hodnota KS-testu

	α	c, resp. b	p-hodnota
európska verzia	0.2574727	24	5.869e-11
americká verzia	0.2539768	24	8.44e-12

Výsledok Kolmogorovho-Smirnovho testu spolu s grafickým porovnaním dát s krivkou hustoty nás privádzajú k záveru, že ani jedno z týchto rozdelení nie je dobrým teoretickým modelom výšky našich poistných plnení, preto v ďalšej časti zvolíme iné rozdelenie zo skupiny rozdelení s ťažkými chvostami.

3. Aproximácia výšky poistných plnení lognormálnym rozdelením

Aby sme videli, či je lognormálne rozdelenie vhodnejšie ako Paretovo, na obrázku 2 je znázornený histogram dát spolu s krivkami hustôt Paretovho aj lognormálneho rozdelenia.



Obrázok 2: Histogram rozdelenia výšky poistných plnení a krivky hustôt Paretovho a lognormálneho rozdelenia

Na základe tohto grafického znázornenia môžeme predpokladať, že lognormálne rozdelenie je vhodnejšie a mohlo by dobre popisovať naše dáta. Na overenie tohto predpokladu použijeme viacero testov, ktoré sú výhodné na testovanie hypotéz o normálnom a lognormálnom rozdelení. Parametre tohto rozdelenia opäť odhadneme metódou maximálnej

vierohodnosti, čím dostaneme nasledovné výsledky: $\hat{\mu} = 2989.917$ a $\widehat{\sigma^2} = 6856.112$. Tieto odhady sme použili v 3 testoch a výsledky sú uvedené v tabuľke 3.

Tabuľka 3: p-hodnoty testov o lognormálnom rozdelení

test	p-hodnota
Kolmogorov – Smirnov test	0.9539
χ^2- test dobrej zhody	0.7140
Shapirov – Wilkov test	0.8854

P-hodnoty vo všetkých troch testoch sú výrazne vyššie ako hladina významnosti $\alpha = 0{,}05$, takže hypotézu o lognormálnom rozdelení nezamietame. Toto rozdelenie teda môžeme považovať za vhodný teoretický model rozdelenia výšky poistných plnení nami analyzovaných dát.

4. Záver

Rozdelenia s ťažkými chvostami sú veľmi dôležitým okruhom a ich aplikácia v poisťovniach je veľmi rozšírená, pretože výšky poistných plnení, hlavne v neživotnom poistení, majú vlastnosti týchto rozdelení. Preto sa stále hľadajú metódy, ako čo najpresnejšie a zároveň najjednoduchšie odhadovať parametre tejto skupiny rozdelení. Takisto sa využívajú v teórii ruinovania na modelovanie procesu celkových nárokov a aj v iných oblastiach pravdepodobnosti i štatistiky.

5. Literatúra

[1]PACÁKOVÁ, V. 2000. Aplikovaná poistná štatistika. Bratislava: Vydavateľstvo ELITA, 2000. 160 s. ISBN 80-8044-073-5.

[2]PLEVOVÁ, J. 1999. Risk in Insurance Master Thesis. Bratislava: Comenius University, 1999.

[3]STEHLÍK, M. – POTOCKÝ, R. – WALDL, H. – FABIÁN, Z. 2008. IFAS Research Paper Series 2008-32. Linz: Johannes Kepler University, Department for Applied Statistics, 2008, s. 18 – 21.

[4] KATINA, S. 2006. Vybrané kapitoly z počítačovej štatistiky I. Bratislava: Univerzita Komenského, Katedra aplikovanej matematiky a štatistiky. 2000. 82 s.

Adresa autora:

Jana Michaliková, Bc.
FMFI UK Bratislava, 5. ročník
Smreková 3
080 01 Prešov
jana.michalikova@gmail.com

Analýza cien bytov v mestách Viedeň, Budapešť, Praha a Bratislava
Analysis of prices for apartments in Vienna, Budapest, Prague and Bratislava

Andrej Sýkora, Ján Juriga, Ladislav Stankovits

Abstract: The object of this labour is the analysis and comparison of prices for flats in capitals of the so called V4: Austria, Hungary, Czech Republic and Slovak Republic, as well as factors influencing these prices. We suppose the significance of this study is ample, since we have used the data about more than 400 flats (around 100 per city) during observations and calculations. Implicitly we have sorted out outliers (units with extreme values of prices, area, etc.), because these would misrepresent our statistics. The data were gathered during the first months of this year (2008), such that we can see the situation just before the mortgage-crisis, when the prices were probably on their top.

Key words: descriptive statistics, linear regression model, analysis of variance (ANOVA)

Kľúčové slová: popisné štatistiky, lineárny regresný model, analýza rozptylu (ANOVA)

1. Úvod

Obsahom tejto práce je analýza a porovnanie cien bytov v metropolách Slovenska a troch našich najbližších susedov: Rakúska, Maďarska a Českej republiky, ako aj faktorov, ktoré na tieto ceny vplývajú. Pri pozorovaniach a následných výpočtoch boli použité údaje o viac než 400 bytoch, čerpaných prevažne z internetových realitných portálov. Dáta boli zozbierané začiatkom tohto roka, takže nie sú ovplyvnené hypotekárnou krízou, ktorá vypukla neskôr. Najskôr sme vytriedili priveľké, pridrahé, alebo inak nevhodné byty, ktoré by príliš skresľovali štatistiky, pričom zostalo 364 bytov, s ktorými sme ďalej pracovali v štatistickom softvéri SAS Enterprise Guide.

Cieľom práce je poukázať hlavne na rozdiely v cenách, resp. spoločné faktory vplývajúce na ceny bytov v týchto mestách.

Zaznamenali, resp. analyzovali sme nasledovné premenné:

- Cena – štandardne uvádzaná alebo prepočítaná na eurá [EUR] pre porovnávanie
- Poloha – každé mesto sme rozdelili na 3 zóny: „centrum“, „širšie centrum“ a „mimo centra“
- Rozloha – v štvorcových metroch [m^2]
- Počet izieb – jedno- až trojizbové byty
- Stav – kategorizovaný do hodnôt: „nový“, „rekonštruovaný“ a „starý“.

2. Popisné štatistiky

Na úvod sme pre každý byt vypočítali pomerovú premennú cena za štvorcový meter, ktorú budeme ďalej používať. Taktiež sme si kvalitatívne premenné stav a poloha pretransformovali na kvantitatívne, čo bolo potrebné pre ďalšie výpočty. Pomocou štatistického softvéru SAS sme zistili nasledovné ukazovatele bytov v jednotlivých mestách:

Priemerná cena za m^2 bytov vo **Viedni** (zaokrúhlená na dve desatinné miesta) je 2133,22 EUR a štandardná odchýlka je 793,73 EUR. Variačné rozpätie, teda rozdiel medzi najnižšou (1065 EUR) a najvyššou (6000 EUR) cenou za m^2 je 4935 EUR. Medián je 1977,78 EUR a rozdeľuje štatistický súbor na 2 rovnako veľké časti podľa ceny za štvorcový meter. S pravdepodobnosťou 95% je priemerná cena za m^2 v intervale <1960,97 ; 2305,47>.

Priemerná cena za m^2 bytov v **Budapešti** je 1295,01 EUR a štandardná odchýlka je 473,81 EUR. Najnižšia hodnota ceny za m^2 je 682,93 EUR, pričom najvyššia hodnota je

2926,67 EUR a z toho sa dá následne vypočítať variačné rozpätie, a to je 2243,74 EUR. To znamená, že rozdiel medzi bytom, ktorého cena za m^2 je najnižšia a medzi bytom, ktorého cena za m^2 je najvyššia je v tomto meste 2243,74 EUR. Medián je 1147,99 EUR, čo znamená, že 50% bytov má cenu za m^2 nižšiu ako 1147,99 EUR a zvyšných 50% bytov má cenu za m^2 vyššiu ako 1147,99 EUR. S pravdepodobnosťou 95% je priemerná cena za m^2 z intervalu <1194,62 ; 1395,40>.

Priemerná cena za m^2 bytov v **Prahe** je 2644,87 EUR a štandardná odchýlka je 907,46 EUR. Najnižšia hodnota ceny za m^2 je 1369,05 EUR, pričom najvyššia hodnota je 6165,52 EUR. Variačné rozpätie je teda 4796,47 EUR. Medián je 2436,41 EUR. S pravdepodobnosťou 95% je priemerná cena za m^2 z intervalu <2474,96 ; 2814,79>.

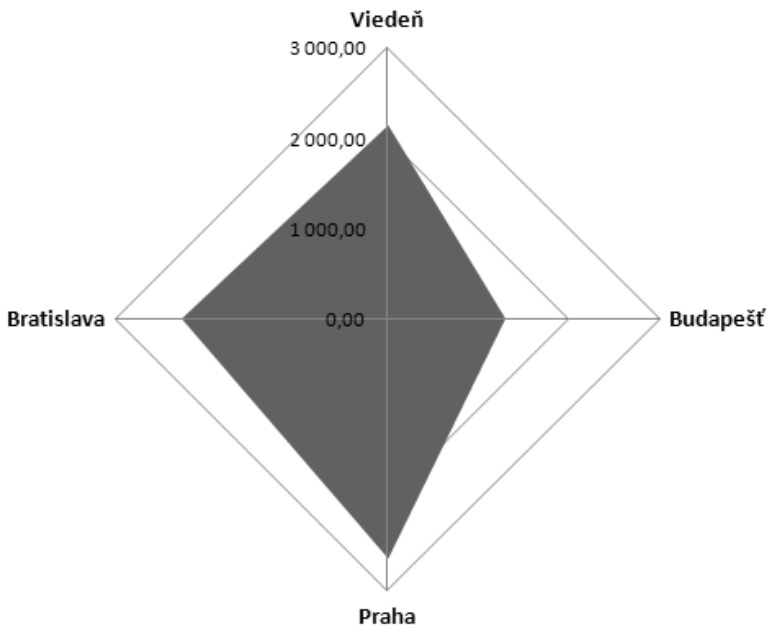
Priemerná cena za m^2 bytov v **Bratislave** zaokrúhlená na dve desatinné miesta je 2280,20 EUR a štandardná odchýlka je 604,86 EUR. Variačné rozpätie, teda rozdiel medzi najnižšou (1351 EUR) a najvyššou (4 253 EUR) cenou za m^2 je 2902 EUR. Medián je 2138,64 EUR a rozdeľuje štatistický súbor zoradený podľa ceny za štvorcový meter na 2 rovnako veľké časti.

V nasledovnej tabuľke vidíme priemery, štandardné odchýlky a mediány ďalších skúmaných premenných.

Tabuľka 9: Popisné štatistiky bytov v jednotlivých mestách

		Viedeň	**Budapešť**	**Praha**	**Bratislava**
Priemer	*cena za m^2*	2 133,22	1 295,01	2 644,87	2 280,20
	cena [EUR]	167 146,42	72 500,00	172 483,23	141 776,37
	počet izieb	2,30	1,85	2,23	2,13
	rozloha	73,62	55,84	65,69	63,17
Št. odchýlka	*cena za m^2*	793,73	473,81	907,46	604,86
	cena [EUR]	113 876,41	37 164,72	84 115,57	60 664,19
	počet izieb	0,69	0,65	0,67	0,82
	rozloha	27,38	18,83	21,76	20,83
Medián	*cena za m^2*	1 977,78	1 147,99	2 436,41	2 138,64
	cena [EUR]	129 500,00	62 000,00	142 462,51	124 975,10
	počet izieb	2,00	2,00	2,00	2,00
	rozloha	70,00	54,00	62,65	62,50

Priemerné jednotkové, ale aj celkové ceny (v EUR) sa teda líšia. Najvyššie sú v Prahe, a suverénne najnižšie v Budapešti. Taktiež môžeme z tabuľky vyčítať, že Budapešť má najmenšie byty s najmenším počtom izieb, pričom Viedeň ich má najväčšie. Značné rozdiely v priemerných cenách za m^2 vidieť najlepšie v nasledujúcom grafe (obrázok 1).



Obrázok 9: Graf porovnávajúci priemerné jednotkové ceny za m^2

3. Lineárny regresný model

V tejto časti sme sa pokúsili vytvoriť lineárny regresný model závislosti premennej *cena za m^2* od polohy, stavu a počtu izieb, a to najskôr osobitne v jednotlivých mestách (okrem Bratislavy, ktorú sme z technických príčin do modelu nezahrnuli). Premenné *Stav* a *Poloha* sme pre túto potrebu pretransformovali na kvantitatívne premenné s možnými hodnotami: *Stav*: 1=nový, 2=rekonštruovaný, 3=starý. *Poloha*: 1=centrum, 2=širšie centrum, 3=mimo centra.

V rámci linear regression sme vybrali „stepwise selection“, aby sme z modelu vylúčili nevýznamné premenné. To sa stalo v meste **Viedeň** s premennou počet izieb, pretože na hladine významnosti 15%, ktorú sme stanovili, nebola dostatočne významná. Výsledný lineárny model teda vyzeral nasledovne:

$$\hat{y}_i = 4313{,}29 - 333{,}70 * Poloha - 706{,}05 * Stav \qquad (1)$$

Interpretácia hodnôt:

$\beta_0 = 4313{,}29$ (intercept) – keby sa *poloha* aj *stav* rovnali nule (iba teoreticky), tak *cena za m^2* bytu vo Viedni by bola 4313,29 EUR.

$\beta_1 = -333{,}7$ – ak sa *poloha* bytu zmení o jednotku (miesto mimo centra na blízko centra, alebo z blízko centra na centrum), pričom ostatné parametre ostanú nezmenené, tak *cena za m^2* klesne o 333,7 EUR.

$\beta_2 = -706{,}05$ – ak sa *stav* bytu zmení o jednotku (z nový na rekonštruovaný alebo z rekonštruovaný na starý) a ostatné parametre ostanú nezmenené, *cena za m^2* klesne o 706,05 EUR.

Podľa semiparciálnych štatistík vplýva na cenu za štvorcový meter premenná stav (38,65%) výraznejšie než poloha (2,44%)

Koeficient determinácie, teda r^2 vyšiel 0,4015, čo znamená, že tento model vysvetľuje iba 40,15% prípadov, teda ostatné prípady týmto modelom nie sú vysvetlené. Kvalitu modelu teda nepovažujeme za postačujúcu. To možno pripísať tomu, že na cenu bytu za m^2 vplývajú aj ďalšie faktory, ktoré sme do modelu nezahrnuli.

Premenná *počet izie*b nebola použitá ani v modeli pre **Budapešť**. Lineárny regresný model sme preto vytvorili s použitím premenných *poloha* a *stav*:

$$\hat{y}_i = 2135{,}96 + 94{,}62 * Poloha - 544{,}81 * Stav \quad (2)$$

Koeficient determinácie vyšiel 0,3506, teda tento model vysvetľuje len 35,06% prípadov, a ostatných 64,94% ostáva týmto modelom nevysvetlených. Zo semiparciálnych štatistík vyčítame, že *stav* bytu prispieva k modelu najväčšou mierou, vysvetľuje 33,08% prípadov, a *poloha* bytu vysvetľuje len 1,98% prípadov.

Pre **Prahu** boli pri použití stepwise selection v modeli obsiahnuté všetky premenné, ktoré sme zahrnuli, teda *počet izieb*, *poloha* aj s*tav:*

$$\hat{y}_i = 5475{,}3 - 186{,}9 * Pocet\_izieb - 1189{,}9 * Poloha - 207{,}4 * Stav \quad (3)$$

Podľa Mallowovho C_p môžeme považovať model za kvalitný. C_p =3,27 čo je menšie ako 4 (počet parametrov v modeli). r^2=0,6 znamená, že model popisuje 60% všetkých pozorovaní. Táto hodnota nie je úplne uspokojivá. Pripisujeme to aj tomu, že v modeli nie sú zahrnuté ďalšie dôležité faktory určujúce ceny bytov. Medzi ne patrí napr. poschodie, na ktorom sa byt nachádza, hlučnosť okolia, infraštruktúra atď.

Vytvorili sme aj lineárny regresný model pre **všetky byty vo všetkých mestách** (ako sme už spomenuli, okrem Bratislavy), pričom sme taktiež použili metódu „stepwise“, teda že premenné, ktoré nie sú z hľadiska modelu štatisticky významné, sú v priebehu tvorby modelu vynechané. Vyšiel nám nasledujúci model:

$$\hat{y}_i = 4120{,}19 - 675{,}31 * Poloha - 340{,}47 * Stav \quad (4)$$

Interpretácia koeficientov je rovnaká ako pri lineárnom regresnom modeli pre mestá jednotlivo.

Koeficient determinácie tohto modelu je 0,28, teda tento model vysvetľuje len 28% prípadov, pričom ostatných 72% ostáva týmto modelom nevysvetlených. Väčšou mierou k modelu prispieva premenná *poloha*, teda 21,31%, a *stav* prispieva len 6,69%.

4. ANOVA

Pomocou analýzy rozptylu sme sa snažili zistiť, či je priemerná cena za m^2 bytu v týchto troch mestách (Viedeň, Budapešť a Praha) približne rovnaká. Keďže táto premenná nemá normálne rozdelenie, neboli splnené podmienky pre použitie parametrickej ANOVY a museli sme použiť ANOVU neparametrickú.

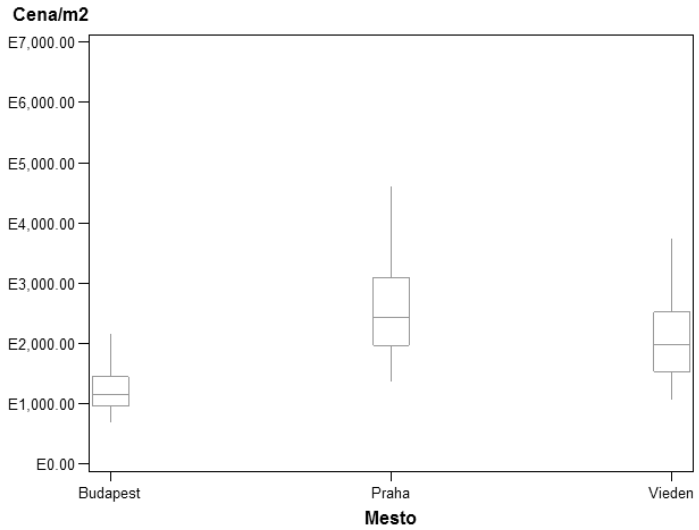
Testovali sme nasledujúce hypotézy:

H_0: Priemerné ceny za m^2 v jednotlivých mestách sú približne rovnaké.

H_1: Aspoň v jednom meste je priemerná cena bytu za m^2 významne odlišná.

Hladina významnosti α=0,05

Hypotézu sme testovali pomocou testu mediánu, Van der Waerdenovej analýzy a Savageovej analýzy, ale ani v jednom prípade sme nemohli prijať hypotézu H_0, pretože p-hodnoty sa pohybovali výrazne pod úrovňou 0,05. Pre ilustráciu pripájame aj krabicový graf (box plot), ktorý prehľadne znázorňuje priemerné ceny za m^2 aj veľkosť rozptylov. Ako môžete z neho vyčítať, *cena za* m^2 je v priemere najvyššia v Prahe.



Obrázok 10: Box plot znázorňujúci rozdiely v priemerných cenách za m^2

5. Záver

V prvej časti našej práce sme sa zaoberali popisnými štatistikami bytov v jednotlivých mestách (Viedeň, Budapešť, Praha a Bratislava), a ich vzájomným porovnaním. Vyšli nám zaujímavé závery, napríklad že suverénne najlacnejšie byty sa predávajú v Budapešti, zatiaľ čo najdrahšie v Prahe.

Ďalej sme sa pokúsili vytvoriť lineárne regresné modely cien bytov za štvorcový meter a zistiť, ktoré premenné akou mierou do týchto modelov prispievajú. Najkvalitnejší model vznikol pre byty v Prahe, pretože vysvetľoval vyše 60 percent prípadov. Jednoznačne najviac vplývala na cenu p*oloha* vo všetkých modeloch. Do lineárneho modelu sme z technických príčin nezahrnuli Bratislavu.

V poslednej časti sme si len potvrdili neparametrickou analýzou rozptylu (ANOVA), že priemerné ceny bytov v jednotlivých mestách sa významne líšia. Do tohto testu sme taktiež nezahrnuli byty v Bratislave.

6. Literatúra

[1]STANKOVIČOVÁ, IVETA Ako robiť štatistiku v systéme SAS In: Výpočtová štatistika. - Bratislava: SŠDS, 2000. - S. 74-78. - ISBN 80-88946-09-03

[2]PACÁKOVÁ, VIERA A KOL.: Štatistika pre ekonómov. Bratislava. IURA Edition. 2003. ISBN 80-89047-74-2

Adresa autorov:

Andrej Sýkora, Ján Juriga, Ladislav Stankovits, študenti 3. roč.
Fakulta managementu, Univerzita Komenského v Bratislave
ul. Odbojárov 10, 820 05 Bratislava
e-mail: *Andrej.Sykora@gmail.com*
Jan.Juriga@st.fm.uniba.sk
Ladislav.Stankovits@st.fm.uniba.sk

Rozdelenia s ľahkými chvostami a ich využitie pri aproximovaní výšky poistných plnení
Light-tailed distributions and their application by aproximating claims amount

Eva Uhríková

Abstract: To approximate claims amount seems to be one of the most important things for insurance company to do. This paper deals with so-called light-tailed distributions and their using by approximation, because in some cases, there is more suitable to use light-tailed distributions for approximation than heavy-tailed. We will also offer general steps for finding the best approximating distribution. At the end, we ask a question, if it is better to use the mixtures of distributions or not.

Key words: Light-tailed distributions, claim, Kolmogorov-Smirnov test, Pearson's Chi-squared test, mixtures of exponentials.

Kľúčové slová: Rozdelenia s ľahkými chvostami, poistné plnenie, χ^2-test dobrej zhody, Kolmogorov-Smirnovov test, exponenciálne mixy

1. Úvod

Pre aproximovanie počtu poistných plnení sa v poisťovacej praxi využívajú diskrétne rozdelenia pravdepodobnosti. Výšku týchto plnení však vhodnejšie aproximujú spojité rozdelenia.

Zameriame sa na rozdelenia, ktoré rýchlo konvergujú k nule, presnejšie na rozdelenia s ľahkými chvostami. Pre pochopenie tohto pojmu je dôležité si ľahké chvosty definovať.

Ak máme rozdelenie s distribučnou funkciou F(x), ktorá je definovaná ako $F(x) = P(X < x)$, potom pravým chvostom tohto rozdelenia bude funkcia $\overline{F}(x) = P(X > x)$. Ďalej hovoríme, že rozdelenie má ľahký chvost, ak je konvergencia rozdelenia k nule nie je pomalšia ako pri exponenciálnom rozdelení.

Medzi najznámejšie rozdelenia s ľahkými chvostami patria okrem exponenciálneho aj gamma-rozdelenie.

Každé z nich je špecifické a nemôžme jednoznačne povedať, ktoré je pre aproximáciu poistných plnení vhodnejšie. Vždy to závisí od konkrétnych poistných plnení, ktoré chceme popisovať.

Náhodná premenná X má **Exponenciálne rozdelenie** s parametrom λ, ak pre jej hustotu platí vzťah : $f(x) = \lambda e^{-\lambda x}$, kde $x > 0$, $\lambda > 0$. Distribučná funkcia je potom definovaná : $F(X) = 1 - e^{-\lambda x}$ pre $x \geq 0$, $F(X) = 0$ pre $x < 0$.

Hovoríme, že náhodná premenná X má **Gamma-rozdelenie** s parametrami α, β, ak jej hustota má funkčné vyjadrenie $f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

2. Modelovanie rozdelenia výšky poistných plnení

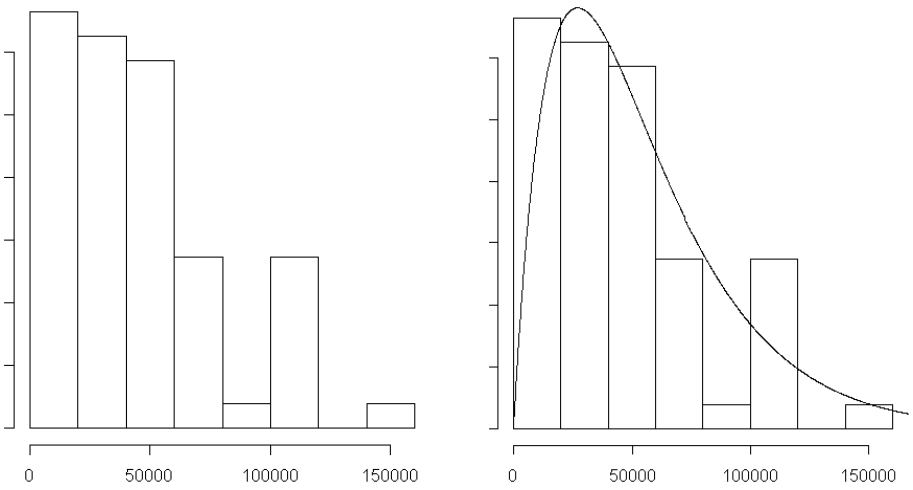
Exponenciálne rozdelenie je vzhľadom na pomerne jednoduché vyjadrenie hustoty a distribučnej funkcie veľmi obľúbené pri modelovaní výšky individuálnych poistných plnení.

Na druhej strane, keďže nedostatočne modeluje interval najnižších a najvyšších hodnôt poistných plnení, nie je vždy použiteľné.

Malá flexibilita exponenciálneho rozdelenia pri modelovaní poistných plnení je spôsobená aj tým, že tvar hustoty pravdepodobnosti závisí od jediného parametra. Preto je omnoho flexibilnejšie rozdelenie gamma, ktorého hustota sa mení v závislosti od dvoch parametrov. Napriek tomu, že je rozdelenie gama v porovnaní s exponenciálnym rozdelením flexibilnejšie, neodhaduje presne pravdepodobnosť príliš vysokých poistných plnení. Je vhodné pre modelovanie výšky poistných plnení, kde variabilita poistných náhrad nie je veľká.

Aby mali výsledky nášho testovania nejaký zmysel, použijeme na aproximáciu reálne dáta. Použitím týchto dát budeme, pomocou štatistického programu R, popisovať výšky poistných plnení pri troch typoch poistných zmlúv. Konkrétne to bude poistenie proti krádeži, protiživelné poistenie a poistenie zodpovednosti za škodu.

Všeobecný postup výberu najvhodnejšieho rozdelenia pre aproximáciu výšky poistných plnení môžeme (podľa [1].) zhrnúť do troch bodov. V prvom kroku navrhneme predpokladaný typ rozdelenia (na základe skúsenosti, prípadne grafického zobrazenia poistných plnení, atď.), následne odhadneme parametre predpokladaného zobrazenia (my použijeme kvantilovú metódu) a nakoniec overíme hypotézu, že nami predpokladané rozdelenie je naozaj to vhodné. Na overenie platnosti hypotézy o zvolenom rozdelení sa najčastejšie používajú Pearsonov χ^2 - test dobrej zhody a Kolmogorov-Smirnovov test.



Obrázok 11: Histogram rozdelenia početností výšky poistných plnení a jeho porovnanie s gamma –rozdelenie výšky poistných plnení pri poistení proti krádeži

Základné informácie o rozdeleniach výšok poistných plnení pri jednotlivých poisteniach sme získali na základe histogramu (Obrázok 1) a tiež vypočítaním si niektorých výberových charakteristík. Z nich vyplýva, že pri poistení proti krádeži a protiživelnom poistení má väčšina poistných plnení nižšie hodnoty a napriek tomu, že sa objavia aj extrémne plnenia, variačný koeficient nie je vysoký. Naopak je tomu pri poistení zodpovednosti za škodu. Popri veľkom počte nízkych nárokov sa objavujú extrémne hodnoty poistných plnení naozaj sporadicky, čo variačný koeficient zvyšuje.

Na základe týchto poznatkov vyslovíme nulovú hypotézu o tom, že poistné plnenia pri všetkých troch typoch poistných zmlúv budeme modelovať pomocou gamma-rozdelenia.

Nulovú hypotézu si najprv overíme **χ^2-testom dobrej zhody.**

Tento test slúži na overenie zhody empirického rozdelenia, t.j. rozdelenia početností výberových údajov s predpokladaným rozdelením. Postup pri testovaní je nasledovný : každý

zo súborov poistných plnení pri jednotlivých poisteniach si rozdelíme do niekoľkých skupín na základe nami zvoleného rozhodovacieho pravidla. Ako príklad môžeme uviesť, že v prvej skupine budú plnenia do 500 Sk, v druhej väčšie ako 500 Sk, ale menšie ako 1000 Sk atď. Potom si vypočítame skutočné početnosti v jednotlivých skupinách a tiež zodpovedajúce teoretické, očakávané početnosti na základe rozdelenia, ktorého vhodnosť testujeme. V poslednom kroku vypočítame hodnotu testovacej štatistiky :

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

ktorá má χ^2-rozdelenie s *k-1-p* stupňami voľnosti, pričom *k* je počet skupín a *p* je počet odhadnutých parametrov. E_i sú už spomínané očakávané a O_i skutočné, zistené početnosti. Nakoniec nájdeme v tabuľkách kvantil χ^2-rozdelenia pri požadovanej hladine významnosti a porovnáme ju s hodnotu testovacej štatistiky. Ak je hodnota testovacej štatistiky menšia, nulovú hypotézu prijímame.

Aplikovaním tohto postupu na naše konkrétne dáta sme zistili, že nulovú hypotézu o Gamma-rozdelení výšky poistných plnení prijímame len pri prvých dvoch typoch poistných zmlúv (poistenie proti krádeži a protiživelné poistenie).

V prípade, že rozsah súboru dát nie je postačujúci na využitie χ^2-testu dobrej zhody, môžeme použiť **Kolmogorov-Smirnovov test**. Ten na rozdiel od χ^2-testu vychádza z netriedených dát.

V Tabuľke 1. uvádzame k jednotlivým poisteniam nulovú hypotézu o predpokladanom rozdelení výšky poistných plnení a tiež výsledok p-hodnotu, vypočítanú Kolmogorovovým-Smirnovovovým testom. Pomocou nej buď prijmeme alebo zamietneme nulovú hypotézu.

Tabuľka 10: Výsledky testovania

	Nulová hypotéza	Kolmogorov-Smirnovov test (p-hodnota)
Poistenie proti krádeži	Dáta majú gamma-rozdelenie	0,27750
Protiživelné poistenie	Dáta majú gamma-rozdelenie	0,06557
Poistenie zodpovednosti za škodu	Dáta majú gamma-rozdelenie	0,02062

Z tabuľky je zrejmé, že keďže pri prvých dvoch typoch poistných zmlúv je p-hodnota väčšia ako hladina významnosti (zvoľme si hladinu významnosti $\alpha = 0.05$), nulovú hypotézu nezamietneme. Naopak, pri poistení zodpovednosti za škodu, ju z opačného dôvodu zamietame.

Testy, ktoré sme v článku spomenuli majú veľkú výhodu v tom, že sú široko využiteľné a dá sa nimi testovať vhodnosť použitia viacerých pravdepodobnostných rozdelení.

Je však vhodné spomenúť, že existujú aj iné testy, nie asymptotické, ale presné, ktoré sa využívajú pri testovaní gamma modelov. Ich výhodou je, že dokážu presne spočítať p-hodnotu a tiež určiť silu testu. Viac v [3].

3. Záver

Z nášho testovania vyplýva, že poistné plnenia pri poistení proti krádeži, tak ako aj pri protiživelnom poistení sa dajú veľmi vhodne aproximovať pomocou gamma-rozdelenia. Naopak pri poistení zodpovednosti za škodu oba testy nulovú hypotézu zamietli. To môže vyplývať z toho, že výber z rozdelenia s ľahkými chvostami obsahuje menší podiel extrémne nízkych alebo aj vysokých hodnôt ako je to u rozdelení s ťažkými chvostami. Poistné plnenia

pri poistení zodpovednosti za škodu sú z veľkej časti nízke a preto sa nedajú aproximovať rozdelením s ľahkými chvostami.

Ak by nejaký súbor poistných plnení nebolo možné aproximovať pomocou niektorého zo spomenutých rozdelení, v praxi sa často využívajú ich mixy. Ich využitie je vhodné najmä vo chvíli, keď sa v súbore poistných plnení vyskytuje viacero extrémnych hodnôt.

Spomenieme napríklad **mix exponenciálnych rozdelení**, nazývaný tiež hyper-exponenciálne rozdelenie. Je to spojité rozdelenie s nasledovnou hustotou :

$$f_X(x) = \sum_{i=1}^{n} f_{Y_i}(y)\, p_i \,, \tag{2}$$

kde Y_i je náhodná premenná s exponenciálnym rozdelením s parametrom λ_i a p_i je pravdepodobnosť, že náhodná premenná X bude vhodne pasovať do exponenciálneho rozdelenia s parametrom λ_i.

Keďže hustota exponenciálneho mixu je vlastne konvexná kombinácia hustôt jednotlivých komponentov mixu, tak pre pre strednú hodnotu a disperziu platia nasledovné vzťahy :

Očakávaná hodnota : $E(X) = \int_{-\infty}^{\infty} xf(x)dx = p_1 \int_0^{\infty} x\lambda_1 e^{-\lambda_1 x} dx + \ldots + p_n \int_0^{\infty} x\lambda_n e^{-\lambda_n x} dx = \sum_{i=1}^{n} \frac{p_i}{\lambda_i}$ (3)

Druhý centrálny moment : $E(X^2) = \sum_{i=1}^{n} \frac{2}{\lambda_i^2} p_i$ (4)

Z toho jednoduchým výpočtom dostaneme disperziu : $D(X) = \sum_{i=1}^{n} \frac{2}{\lambda_i^2} p_i - (E(X))^2$ (5)

Vlastnosti exponenciálnych mixov. Napriek tomu, že mixy rozdelení vo všeobecnosti, vytvárajú množstvo rôznych tvarov hustôt, mixy exponenciálnych rozdelení sú menej variabilné a odchýlky v tvaroch hustôt exponenciálnych mixov nie sú významné.

Dôležitou otázkou na záver teda zostáva, či je pre pre štatistickú prax dostačujúce využitie jednoduchých exponenciálnych rozdelení alebo je lepšie využiť ich mixy.

4. Literatúra

[1]PACÁKOVÁ, V. 2004. Aplikovaná poistná štatistika. Bratislava : IURA EDITION, 2004. 261 s. ISBN 80-8078-004-8

[2] WAGNER, H. 2008. Bayesian analysis of mixtures of exponentials. In: Journal of Applied Matematics and Informatics, č. 5, 2008.

[3]POTOCKÝ, R. – STEHLÍK, M. 2005. Stochastic Models in Insurance, Risk and Pension Funds. In : Journal of Applied Matematics and Informatics, č. 1, 2005.

Adresa autora:

Eva Uhríková, Bc., FMFI UK, 5.ročník
Povoznícka 16, 841 03 Bratislava, E-mail: eva.uhrikova@gmail.com

Optimalizácia trhového portfólia
Market portfolio optimalization

Juraj Varga

Abstract: The aim of this paper is modelling an efficient portfolio of risky assets, including risk-free asset. In the first chapters we are describing theory of portfolio optimalization and in the last part creating an optimal portfolio of assets. During optimalization we can encounter some problems with data which we are discussing in summary.

Keywords: efficient portfolio, SML, CAPM, envelope portfolio, risky and risk-free assets

Kľúčové slová: efektívne portfólio, priamka trhu cenných papierov, riziková a bezriziková aktíva, CAPM, obalové portfólio

1. Úvod

Moderná teória portfólia sa spája s menami ako Hary Markowitz, John Lintner, Jan Mossin a William Sharpe. Teória portfólia je mikroekonomická disciplína, ktorá skúma, aké kombinácie aktív je vhodné držať spoločne, aby takto vytvorené portfólio malo isté, vopred dané vlastnosti. Investorov na kapitálových trhoch môžeme rozdeliť na dve skupiny a to na špekulantov a investorov, ktorý sledujú dlhodobú udržateľnosť hodnoty nimi voleného portfólia, inak, chcú sa zaistiť voči nadmernej volatilite. V ostatných dňoch sú odborné i menej odborné state o alokačnej efektivite trhu, bezrizikových portfóliách, burzách o akomkoľvek nástroji finančného trhu, skoro taký prečin ako v stredoveku kacírstvo, našťastie sa dnes už neupaľuje na hranici. Napriek tomu sa bude tento príspevok venovať efektívnemu trhovému. Pri konštrukcií efektívneho trhového portfólia využijeme poznatky diskutované v úvode príspevku. Budeme sa venovať základným definíciám tykajúcich sa modelu, pričom cieľom tejto práce je popísať teoretické východiská, na ktorých sa zakladajú teórie efektívneho portfólia a následne zostrojiť portfólio aktív, ktorého efektivita sa testuje v reálnych podmienkach kapitálového trhu. Autor tejto práce si je vedomí rizík, ktoré sú spojené s overením efektívnosti kapitálového trhu ale aj meniacich sa podmienok v rámci globálneho pohľadu na finančné trhy ako také. Zmyslom príspevku je jednoducho popísať a názorne ukázať proces optimalizácie trhových portfólií. Pri konštrukcii modelu budeme používať aplikácie programu Excel 2007.

2. Postup a použité metódy

V úvode časti zadefinujeme pojmy a premenné, s ktorými model pracuje:

Trhové portfólio je portfólio, ktoré ja tvorené investíciami do všetkých cenných papierov v takom pomere, že proporcia investovaná do jedného cenného papiera zodpovedá jej relatívnej trhovej hodnote, kde portfólio obsahuje N rizikových aktív, pričom $E[r_i]$ je očakávaný výnos *i-tého* aktíva. Premenná $\boldsymbol{R}$ je stĺpcový vektor očakávaných výnosov aktív portfólia a premenná $\boldsymbol{S}$ je $N \times N$ je kovariančná matica očakávaných výnosov. Portfólio rizikových aktív je stĺpcový vektor $\boldsymbol{x}$, pričom platí, že:

$$\sum_{i=!}^{N} x_i = 1, \tag{1}$$

Každá súradnica x_i reprezentuje pomer daného rizikového aktíva ***i*** v portfóliu.

Očakávaný výnos portfólia : $E[r_x] = x^T \times R = \sum_{i=1}^{N} x_i E[r_i]$ (2)

Disperzia výnosov portfólia: $\sigma^2 = \sigma_{xx} = x^T S x = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \sigma_{ij}$ (3)

Kovariancia výnosov dvoch portfólií: $Cov(r_x, r_y) = \sigma_{xy} = x^T S y = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i y_j \sigma_{ij}$ (4)

Základom teórie efektívneho portfólia je (kovariančný) vzťah medzi výnosom aktíva. Podľa Markowitzovej teórie je celková miera rizika celého portfólia v skutočnosti menšia ako vážený priemer štandardných odchýlok jednotlivých aktív v portfóliu. Na základe tohto prístupu rozvinul pojem "*efficient portfolio*" (*efektívne portfólio*). Efficient portfolio je také portfólio, kde pri danej miere rizika je možné dosiahnuť vyšší výnos, resp. daný výnos je možné dosiahnuť pri nižšej miere rizika. Riziko celého portfólia však nie je len vážený priemer štandardných odchýlok jednotlivých aktív, ale dôležité je brať do úvahy ich vzájomnú kovarianciu. Teda portfólio ***x*** je efektívne, ak neexistuje také portfólio ***y***, pre ktoré platí, že $E[R_y] > E[R_x]$ a zároveň $\sigma_y \le \sigma_x$. Skupinu efektívnych portfólií nazývame účinným ohraničením – *efficient frontier*.

2.1. Predpoklady o efektívnom portfóliu a modely CAPM

*Predpoklad 1.:*Predpokladajme že *c* je konštanta, potom môžeme vektor **R** upraviť ako $R - c$. Ak preriešime systém lineárnych rovníc cez vektor z, teda $R - c = Sz$, môžeme urobiť nasledujúce závery:

$$z = S^{-1}\{R - c\} \tag{5}$$

$$x = \{x_1, ..., x_N\}, \text{ kde } x_i = \frac{z_i}{\sum_{j=1}^{N} x_j} \tag{6}$$

Tento predpoklad nám pomôže pochopiť geometrické riešenie pri hľadaní efektívneho portfólia, kde existuje tangencia medzi konštantou c a priestorom v ktorom existujú efektívne portfóliá (graf č.1).

Predpoklad 2.: ak *y* je obalové portfólio[3], tak pre iné portfólio (obalové alebo nie) platí vzťah (konštrukcia priamky trhu cenných papierov):

$$E[r_x] = c + \beta_x [E(r_y) - c], \text{ kde } \beta_x = \frac{Cov(x, y)}{\sigma_y^2} \tag{7}$$

Zároveň platí, že *c* je očakávaný výnos portfólia *z,* pričom kovariancia jeho portfólia s portfóliom *y* je nulová: $c = E[r_z]$, $Cov(y, z) = 0$ (Poznámka[4]). V prípade ak je trhové portfólio *M* efektívne , tak môžeme výraz (7) zovšeobecniť pre každé portfólio, pre ktoré platia vyššie uvedené podmienky. Konštruujeme SML, kde *c* je substituované $E[r_z]$:

---

[3] Obalové portfólio (*envelope portfolio)*, je krivka, na ktorej ležia portfólia s minimálnym rozptylom.

[4] Ak je *y* na obale (ohraničenie efektívnosti - *efficient frontier*) existuje medzi portfóliami z a y lineárna závislosť. Táto konštrukcia CAPM je Security Market Line – Sharpe-Lintner-Mossin nahradená SML s bezrizikovým aktívom nula beta faktorom rešpektuje predpoklady envelope porfolio *y*.

$$E[r_x] = E[r_z] + \beta_x \left[E(r_M) - E(r_z) \right] \tag{8}$$

Predpoklad 3.: Ak na trhu existuje bezrizikové aktívum s výnosom r_f, tak existuje envelope portfolio M, ktorého výnos močme zapísať v tvare rovnice (8) kde nahrádzame výnos portfólia r_z za bezrizikovú úrokovú mieru na danom trhu - r_f (poznámka)[5].

2.2. Odvodenie rizikového beta faktora

CAPM model je odvodený z teórie portfólia a z teórie kapitálového trhu. Model je použitý najme pri:

- Skúmaní správania sa aktív v závislosti od správania sa trhového portfólia
- Vysvetľovaní procesu tvorby cien rizikových aktív na kapitálovom trhu
- Objasňovaní vzťahu medzi očakávaným výnosom aktíva a jeho rizikom za podmienok rovnováhy na trhu, keď všetci investori volia optimálne portfólio, pričom také portfólio je kombináciou rizikových aktív a bezrizikového aktíva.

Pri odvodení použijeme výraz (7), kde jednoduchými úpravami vyjadríme β_x, pričom predpokladáme že aproximáciu $c = E[r_z]$ a $Cov(z, y) = 0$. Ďalej nech y je envelope portfólio a x iné portfólio, potom: $\beta_x = \frac{Cov(x, y)}{\sigma_y^2} = \frac{x^T S y}{y^T S y}$. Pokiaľ je y na efektívnej hranici, už vieme že existuje taký vektor W a konštanta c, ktoré riešia systém $Sw = R - c$ a vieme, že $y = w / \sum_i w_i = w / a$, môžeme vykonať substitúciu:

$$\beta_x = \frac{Cov(x, y)}{\sigma_y^2} = \frac{x^T S y}{y^T S y} = \frac{x^T (R - c)/a}{y^T (R - c)/a} = \frac{x^T (R - c)}{y^T (R - c)} \tag{9}$$

Za predpokladu, že suma váh aktív v portfóliu nám dáva hodnotu jedna, vieme upraviť vzťah (9) nasledovne:

$$\beta_x = \frac{E(r_x) - c}{E(r_y) - c} \tag{10}$$

Pri zavedenom označení premenných môžeme odhadnúť β_x aj za predpokladu existencie N rizikových aktív a bezrizikového aktíva. Použime označenie podľa vzťahu (8), resp. predpokladu 3.

---

[5] Ak všetci investori vytvárajú portfóliá iba na základe ich štatistických charakteristík a to stredného výnosu a disperzie výnosov, potom *M* je portfólio vytvorené zo všetkých rizikových aktív v ekonomike v hodnotách daných aktív. Predpokladáme, že máme *N* rizikových aktív a hodnota každého aktíva *i* je V_i. Potom majú aktíva portfólia nasledujúce váhy: váha aktíva *i* portfólia $M = \frac{V_i}{\sum_{h=1}^{N} V_h}$

3. Dáta a konštrukcia efektívneho portfólia rizikových aktív a bezrizikového aktíva

Nami konštruované portfólio obsahuje zatváracie ceny piatich titulov[6] obchodovaných na NYSE[7] od 18.11.2007 do 17.11.2008, čiže údaja za jeden rok. Tituly boli zvolené v záujme pokryť viaceré segmenty ekonomiky. Konštrukcia vychádza z predpokladu 3, teda zahŕňa bezrizikové aktívum, ktorým je ročný US Streasury Bond[8].

Úloha výberu optimálneho portfólia je stanoviť váhy tak aby sme splnili požiadavky na optimálne portfólio. Na základe definície efektívneho portfólia chceme minimalizovať jeho riziko $D[p]=\sigma^2$. Riešením, za podmienky hodnoty sumy (1) váh titulov portfólia, pričom rizikové aktívum má nulovú váhu, je portfólio s výnosom $E[p]$ a rizikom $D[p]=\sigma^2$ meraným smerodajnou odchýlkou $\sqrt{D[p]}$, počítaných podľa vzorca (2) resp. (3) (tabuľka č.1,2) Ním predchádza konštrukcia kovariančnej matice výnosov titulov portfólia – (4).

Tabuľka č.1,2

Tituly p.	váhy i-tého aktíva
Arcelor Mittal	0
Disney	0.399249521
Exxon	0.13213139
Google	0.206825438
Vodafone	0.26179365

očakávaný výnos p.	-0.2408%
riziko p.	***0.061%***
suma váh aktiv p.	1
smernica SML - β	-0.190821

Zdroj: vlastné spracovanie

Predpoklad o efektívnom portfóliu overíme, tak že budeme maximalizovať očakávaný výnos pri dodržaní rovnakého rizika. Váha bezrizikového aktíva je nenulová, platia všetky predpoklady platné pri minimalizácií rizika v prvom prípade. Ak maximalizáciou očakávaného výnosu dosiahneme vyšší výnos (nižšiu stratu), dostávame trhovo efektívne portfólio rizikových aktív a bezrizikového aktíva (tabuľka č. 3,4) (proces optimalizácie portfólia). Oba výpočty uskutočníme aplikáciou Riešiteľ (Solver) MS Excel 2007.

Tabuľka č.3,4

Tituly p.	váhy i-tého aktíva
Arcelor Mittal	0
Disney	0
Exxon	0.811707821
Google	0
Vodafone	0
US Treasury Bond	0.188292179

očakávaný výnos p.	**0.0067%**
riziko p.	*0.061%*
suma váh aktiv p.	1
smernica SML - β	-0.09057

Zdroj: Vlastné spracovanie

4. Diskusia a záver

Cieľom práce bolo popísať a skonštruovať trhovo efektívne portfólio. V časti 3. sme popísali dáta, podmienky modelu a postup konštrukcie trhovo efektívneho portfólia, na základe teoretických poznatkov z predchádzajúcich častí príspevku. Ako môžeme pozorovať

[6] Portfólio je zložené zo zatváracích cien akcii spoločností Arcelor Mittal, Disney, Exxon Mobile, Google Inc. a Vodafone.

[7] Zdroj:finance.yahoo.com

[8] Zdroj:bonds.yahoo.com

z dátového zobrazenia optimalizácie nami zvoleného portfólia (tabuľka č. 3,4), vyriešili sme portfólio s minimálnym rozptylom s výnosom $E[p]$ (tabuľka č.1,2), ktorý sme následne maximalizovali a tak získali váhy aktív portfólia. Pri riešení sme uvažovali o povolení krátkych predajov, avšak pri nami zvolených tituloch by sme museli konštruovať portfólio výhradne s krátkymi predajmi, čo je v rozpore s definíciou portfólia ako takého, investor chce držať aktíva a nie sa ich zbavovať. Zvolili sme obmedzenie nezáporných váh a zopakovali optimalizáciu. Riešením sú váhy optimalizovaného portfólia (tabuľka č.3,4). Súčasný stav na svetových finančných trhoch (najmä v USA) je vysvetlením práve tohto výsledku optimalizácie (v prvom prípade záporných váh aktív), kde očakávaný výnos je výnos bezrizikového aktíva znížený o záporný výnos akcii Exoon Mobile. Nulové váhy ostatných aktív sú dôsledkom vývoja cien akcii týchto aktív, ktoré za sledované obdobie generovali stratu na akciách. Pri tvorení portfólia bez obmedzení krátkych predajov sú tak záporné váhy aktív logickým dôsledkom maximalizácie očakávaného výnosu portfólia (minimalizácie straty). Pri ich obmedzení, zase ich nulové hodnoty. Ukázalo sa, že optimalizovať (maximalizácia výnosu, pri danom riziku) portfólio v stave recesie finančných trhov nemá požadovaný efekt z pohľadu investora, ba dokonca je disfunkčné. Treba však poznamenať, že investor, ktorý investuje do akciových titulov a tak vytvára portfólio rizikových aktív, pozerá aj za horizont jedného roku. Optimalizáciou sa nedá zabrániť prepadu cien aktív, ani ich predikovať, preto by nebolo na mieste zavrhnúť štatistické metódy vytváraní a optimalizácií portfólií aktív, hoci dnešný stav dáva mnohým priestor tak činiť.

5. Literatúra

[1] BENNINGA, S.: Financial Modeling, MIT Press, 2000
[2] MLYNAROVIČ, V.: Finančné investovanie. Bratislava: Iura Edition, 2001
[3] SIVÁK, R., GERTLER, Ľ.: Teória a prax vybraných druhov finančných rizík (kreditné, trhové, operačné). Bratislava

Adresa autora:

Juraj Varga, 5. ročník
Ekonomická fakulta UMB
Tajovského 10
975 90 Banská Bystrica
e-mail: juraj.varga@yahoo.com

Asymptotické vlastnosti kvadratického modelu z dvoch častí
Asymptotic normality of a two-part quadratic model

Jana Horňanová

Abstract: The paper studies a two-part approximation model of constrains. It shows that thanks to a new type of constraint and a special polynomial representation one can get estimates with such properties as asymptotic normality.

Key words: approximation, polynomial model, reference points, recursive estimate

Kľúčové slová: aproximácia, polynomický model, referenčné body, rekurzívny odhad

1. Úvod

V článku sa budeme zaoberať vlastnosťami odhadu polynómu druhého stupňa v rámci aproximačného modelu z dvoch častí. Tento model je opísaný v druhej kapitole. Dôležitú úlohu v druhej časti modelu zohrávajú referenčné body. V tretej kapitole sú obsiahnuté základné výsledky, kde ukážeme, že pre rekurzívny odhad platí nie len konzistentnosť ale aj asymptotická normalita.

Aproximačný model z dvoch častí je založený na vete o reprezentácii polynómov [1, 3]:

Veta: Nech $p > 2$, p je prirodzené číslo. Potom pre ľubovoľné nerovnaké $v_0, v_1 \in R$, platí:

$$P_p(x) = I(x) + Z(x)A(x), \tag{1}$$

Kde:

$$I(x) = \frac{x - v_1}{v_0 - v_1} P_p(v_0) + \frac{x - v_0}{v_1 - v_0} P_p(v_1), \tag{2}$$

$$Z(x) = (x - v_0)(x - v_1),$$

$$A(x) = \sum_{i=2}^{p} a_i T_{i,x}, \quad T_{i,x} = \sum_{j=0}^{i-2} x^j S_{i-2-j}, \quad S_j = \sum_{k=0}^{j} v_1^k v_0^{j-k}.$$

Množinu $B = \{[v_0, P_p(v_0)], [v_1, P_p(v_1)]\}$ budeme nazývať množinou referenčných bodov.

2. Model z dvoch častí

Uvažujme na intervale [-1,0] a [0,1] kvadratické polynómy $P_a(x)$, $P_b(x)$, kde $a^T = (a_0, a_1, a_2)$, $b^T = (b_0, b_1, b_2)$ a zodpovedajúce množiny meraní

$$\left\{[x_i, \tilde{y}_i] : x_i = -\frac{i}{M}, \ \tilde{y}_i \equiv P_a(x_i) + \varepsilon_i^*, \ i = 1, M\right\},$$

$$\left\{[x_i, \tilde{y}_i] : x_i = -\frac{i}{N}, \ \tilde{y}_i \equiv P_b(x_i) + \varepsilon_i, \ i = 1, N\right\},$$

kde $M, N >> 1$, $M = \kappa N$, $\kappa \in R^+$, $\kappa < \infty$, $\{\varepsilon_i\}_{i=1}^{M}$ a $\{\varepsilon_i^*\}_{i=1}^{N}$ sú nekorelované postupnosti chýb so strednou hodnotou nula, $E(\varepsilon_i \varepsilon_j) = \sigma^2 \delta_{i,j} < \infty$, $E(\varepsilon_k^* \varepsilon_l^*) = \sigma^2 \delta_{k,l} < \infty$, δ_{**} je Kroneckerovo delta. Písmenami *E, D* označujeme strednú hodnotu a disperziu náhodnej veličiny.

Model pre aproximáciu sa skladá z dvoch častí [2]:

$$\begin{cases} \tilde{y} = P_a(x) + \varepsilon^*, & x \in [-1,0], \\ \tilde{y} = I^B(x) + Z(x)b + \varepsilon, & x \in [0,1]. \end{cases}$$

Koeficienty polynómu P_a odhadneme pomocou metódy najmenších štvorcov a odhad polynómu označíme $P_{\hat{a}}$. Na zabezpečenie hladkého prechodu medzi polynómami v bode nula budeme predpokladať, že platí:

$$P_a(0) = P_b(0),\ P_b'(0) = P_a'(0) + o(\tau), \tag{3}$$

kde τ je malé kladné číslo. Je možné ľahko ukázať, že definovaním referenčných bodov spôsobom: $B = B_a = \{[-\tau, P_a(-\tau)],[0, P_a(0)]\}$ je podmienka (3) splnená. Teda pre referenčné body bude platiť: $v_0 = 0, v_1 = -\tau$ a $P_a(0) = P_b(0)$, $P_a(-\tau) = P_b(-\tau)$.

3. Rekurzívny odhad parametra

Všimnime si, že druhá časť modelu vychádza z reprezentácie (1). Vďaka zvolenému modelu a výberu ref. bodov polynóm $P_b(x)$ obsahuje iba jeden neznámy parameter $b \equiv b_2$.

Pre odhad parametra *b* použijeme iteračnú schému [1]:

$$b_i = b_{i-1} + \frac{Z_i}{\varsigma_i}(\tilde{y}_i - \hat{I}_i - Z_i b_{i-1}),\ b_0 = 0,\ i = 1,..,N$$

kde $\hat{I}_i \equiv I^{\hat{B}}(x_i)$, $\hat{B} = B_{\hat{a}} = \{[-\tau, P_{\hat{a}}(-\tau)],[0, P_{\hat{a}}(0)]\}$, $Z_i = Z(x_i)$, $\varsigma_i = \sum_{k=1}^{i} Z_k^2$. Zdôrazňujeme, že druhé súradnice referenčných bodov druhej časti modelu sú odhadnuté pomocou odhadu polynómu $P_{\hat{a}}$ z prvej časti modelu. Pre teoretické výpočty používame vyjadrenie:

$$b_i = b + \frac{1}{\varsigma_i}\left(\sum_{k=1}^{i} Z_k(I_k - \hat{I}_k + \varepsilon_k)\right),\ i = 1,..,N. \tag{4}$$

Pomocou (1) a (2) vyjadríme rozdiel interpolačných polynómov:

$$I_k - \hat{I}_k = \sum_{i=0}^{1} p_i(x_k)\left(P_a(i\tau) - P_{\hat{a}}(i\tau)\right) = (a_0 - \hat{a}_0) + (a_1 - \hat{a}_1)\frac{k}{N} + (a_2 - \hat{a}_2)(-\tau)\frac{k}{N}.$$

Označme: $\Delta a = a - \hat{a} = -(X^T X)^{-1} X^T \varepsilon^*$,

$$\begin{aligned} v^T &= \left(\sum_{k=1}^{N} Z_k, \sum_{k=1}^{N} \frac{k}{N} Z_k, \sum_{k=1}^{N} -\tau\frac{k}{N} Z_k\right) = \\ &= \left(N\left(\frac{1}{3} + \frac{1}{2}\tau\right) + o(1), N\left(\frac{1}{4} + \frac{1}{3}\tau\right) + o(1), N\left(-\frac{1}{4}\tau - \frac{1}{3}\tau^2\right) + o(1)\right). \end{aligned}$$

Vidíme, že platí: $\sum_{k=1}^{N} Z_k(I_k - \hat{I}_k) = v^T \Delta a$.

Pomocou (4) dostávame:

$$b_N = b + \frac{1}{\varsigma_N}\left(v^T \Delta a + \sum_{k=1}^{N} Z_k \varepsilon_k\right), \tag{5}$$

kde

$$\varsigma_N = \sum_{k=1}^{N} Z_k^2 = \frac{1}{N^4}\sum_{k=1}^{N}((k + N\tau)k)^2 = \left(\frac{1}{5} + \frac{\tau^2}{3} + \frac{\tau}{2}\right)N + o(1). \tag{6}$$

V článku [2] je dokázané, že $E(b_k) = b$, $D(b_n) = \frac{1}{O(N)}$ a že odhad b_N je konzistentný.

Najprv spresníme, čomu sa rovná limita $D\left(b_N\sqrt{N}\right)$ a potom dokážeme asymptotickú normalitu.

Lema: *Nech* b_N *je odhad voľného parametra b pre druhú časť aproximačného modelu.*

Potom :
$$\lim_{N\to\infty} D(b_N\sqrt{N}) = \frac{\sigma^2}{T^2}\left(\frac{1}{\kappa}c^T U^{-1}c + T\right),$$

kde $c = \lim_{N\to\infty}\frac{\nu_N}{N} = \left(\frac{1}{2}\tau+\frac{1}{3},\frac{1}{4}+\frac{1}{3}\tau,\frac{1}{4}\tau+\frac{1}{3}\tau^2\right)^T$, $T = \frac{1}{5}+\frac{\tau^2}{3}+\frac{\tau}{2}$,

$$U = \lim_{N\to\infty}\frac{X^T X}{M} = \begin{pmatrix} 1 & -\frac{1}{2} & \frac{1}{3} \\ -\frac{1}{2} & \frac{1}{3} & -\frac{1}{4} \\ \frac{1}{3} & -\frac{1}{4} & \frac{1}{5} \end{pmatrix}.$$

Dôkaz: Použitím vzťahu (5) dostaneme:

$$D(b_N\sqrt{N}) = \frac{N}{\varsigma_N^2}\left(D(\nu_N^T\Delta a) + D(\sum_{k=1}^{N}\varepsilon_k Z_k)\right) = \frac{N^2}{\varsigma_N^2}\left(\frac{\nu_N^T}{N}ND(\Delta a)\frac{\nu_N}{N} + \sigma^2\frac{\varsigma_N}{N}\right).$$

Pozrime sa teraz na limity jednotlivých členov posledného výrazu.

Na základe (6) dostávame: $\frac{N^2}{\varsigma_N^2}\to\frac{1}{T^2}$,

ďalej: $ND(\Delta a) = N\sigma^2(X^T X)^{-1} = \frac{1}{\kappa}\sigma^2\left(\frac{X^T X}{M}\right)^{-1}\to\frac{U^{-1}\sigma^2}{\kappa}$.

Z toho vyplýva:

$$\frac{N^2}{\varsigma_N^2}\left(\frac{\nu_N^T}{N}ND(\Delta a)\frac{\nu_N}{N}+\sigma^2\frac{\varsigma_N}{N}\right)\to\frac{1}{T^2}\left(c^T\frac{U^{-1}\sigma^2}{\kappa}c+\sigma^2 T\right)$$

q.e.d.

Dôsledok : $D\left(b_N\frac{\sqrt{\varsigma_N}}{\sigma}\right) = \frac{1}{\kappa T}c^T U^{-1}c + 1$**.**

Veta: *Nech sú predpoklady lemy splnené, potom platí*:

$$(b_N - b)\frac{\sqrt{\varsigma_N}}{\sigma}\xrightarrow{d}\mathrm{N}(0,\frac{1}{\kappa\mathrm{T}}\mathrm{c}^{\mathrm{T}}\,\mathrm{U}^{-1}\mathrm{c}+1).$$

Dôkaz: Postupujeme podobne ako v [5]. Najprv ukážeme, že:

$$(b_N - b)\frac{\varsigma_N}{\sigma\sqrt{N\delta_N}}\xrightarrow{d}N(0,1),\ \text{kde}\ \delta_N = \frac{1}{\kappa}c^T U^{-1}c + \frac{\varsigma_N}{N}.$$

Označme výraz:

$$(b_N - b)\varsigma_N = \nu^T\Delta a + \sum_{k=1}^{N}Z_k\varepsilon_k = \sum_{k=1}^{M}d_k\varepsilon_k^* + \sum_{k=1}^{N}Z_k\varepsilon_k = \sum_{k=1}^{L}\xi_k = S_L,$$

kde $d_k = \nu^T(X^T X)^{-1}X^T$ a $L = M(1+\kappa)$. S_L označuje súčet nekonečne veľa náhodných veličín s konečnou disperziou.

$$D(S_L) = \nu^T(X^T X)^{-1}\sigma^2\nu + \sigma^2\varsigma_N = \frac{\sigma^2}{M}\nu^T U^{-1}\nu + \sigma^2\varsigma_N = \frac{\sigma^2 N}{k}c^T U^{-1}c + \sigma^2\varsigma_N =$$

$$= \sigma^2 N(\frac{1}{\kappa}c^T U^{-1}c + \frac{\varsigma_N}{N}) = \sigma^2 N\delta_N$$

Teda: $(b_N - b)\dfrac{\varsigma_N}{\sigma\sqrt{N\delta_N}} = \dfrac{S_L}{\sqrt{D(S_L)}}$.

Na dôkaz štandartného normálneho rozdelenia pre $L \to \infty$ stačí ukázať, že pre $\forall \lambda > 0$ platí Lindebergova podmienka.

$$\lim_{L\to\infty} \frac{1}{D(S_L)} \sum_{k=1}^{L} \int_{\{x:|x-E(\xi_k)|\geq \lambda\sqrt{D(S_L)}\}} (x - E(\xi_k))^2 dF_k(x) = 0 \qquad (7)$$

Ľahko vidieť, že $D(S_L) = O(L)$, teda máme:

$$\frac{1}{D(S_L)} \sum_{k=1}^{L} \int_{\{x:|x|\geq \lambda\sqrt{D(S_L)}\}} x^2 dF_k(x) \leq \frac{L}{O(L)} \max_{1\leq k\leq L} \int_{\{x:|x|\geq \lambda O(\sqrt{L})\}} x^2 dF_k(x).$$

Keďže $D(\xi_k) = \int_{-\infty}^{\infty} x^2 dF_k(x) < \infty$ a $\{x : |x| \geq \lambda O(\sqrt{L})\} \downarrow \emptyset$ pre $L \to \infty$, podmienka (7) bude splnená, lebo: $\lim_{L\to\infty} \int_{\{x:|x|\geq\lambda O(\sqrt{L})\}} x^2 dF_k(x) = 0$.

Keďže $\lim_{N\to\infty} \dfrac{N\delta_N}{\varsigma_N} = \dfrac{1}{\kappa T} c^T U^{-1} c + 1$, tak pomocou štandardných techník ľahko dostaneme:

$$(b_N - b)\frac{\sqrt{\varsigma_N}}{\sigma} \xrightarrow{d} N(0, \frac{1}{\kappa T} c^T U^{-1} c + 1). \qquad \textit{q.e.d.}$$

4. Záver

V práci sme ukázali asymptotickú normalitu rekurzívneho odhadu v kvadratickom modeli z dvoch častí, ktorý je možné využiť ako základ pre úsekovú aproximáciu dát s komplexnou závislosťou [4].

5. Literatúra

[4] DIKOUSSAR N.D., TÖRÖK CS., Automatic knot finding fo piecewise-cubic aproximation, 2006, Mat.model., T-17, N.3

[5] MATEJČÍKOVÁ A., TÖRÖK CS., Two-point Transformation and polynomials, 2007, Kybernetika, odoslaný na publikáciu

[6] RÉVAYOVÁ M., TÖRÖK CS., Piecewise approximation and neural networks, Kybernetika, Kybernetika – Volume 43 (2007), N4, pp. 547-559

[7] RÉVAYOVÁ M., TÖRÖK CS., Úseková parametrická aproximácia, 2007, VIII konferencia SvF TU Košice

[8] RÉVAYOVÁ M., TÖRÖK CS., Reference points based approximation, 2007, Theory of probab. & its applic., odoslaný na publikáciu

Adresa autora:

Jana Horňanová, Bc.
Univerzita Pavla Jozefa Šafárika
Prírodovedecká fakulta, 5. ročník
Trieda SNP 10
040 11 Košice
e-mail: hornanova@centrum.sk

Zmeny v rodinnom a reprodukčnom správaní obyvateľov Bratislavy po roku 1989

The Changes in Family and Reproductive Behaviour of Bratislava Population

Michal Katuša

Abstract: This article is focused on the changes in family and reproductive behaviour of Bratislava population after the year 1989, when the socio-economical changes has been made in Slovakia. These changes are typical for second demographic transition, and are very important for the future development of Bratislava population This work point to the difference between the changes and intensity of changes in Slovakia and in Bratislava.

Key word: Bratislava, second demographic transition, socio-economical transformation

Kľúčové slová: Bratislava, druhý demografický prechod, socio-ekonomická tranformácia

1. Úvod

V druhej polovici 20. storočia zaznamenala väčšina štátov západnej a severnej Európy zmeny v demografickom správaní obyvateľstva, ktoré sú považované za jedny z najvýznamnejších zmien v histórii a preto ich môžeme nazvať revolučnými a tvoriacimi druhý demografický prechod (van de Kaa, D. J., 1987).

Populácia hlavného mesta Slovenska Bratislavy, tak ako aj celého Slovenska, začala podliehať koncom 20. storočia taktiež zmenám v reprodukčnom a rodinnom správaní. Avšak podmienky, za ktorých tieto zmeny prebiehali a prebiehajú v štátoch východnej Európe sú odlišné ako tie, ktoré ovplyvňovali tieto zmeny v západnej Európe. Vo východnej Európe môžeme nazvať tieto zmeny viac krízou ako zámerom (Rychtaříková, J., 1999). Mládek, Marenčáková a Káčerová poukazujú na to, že zmeny vo východnej Európe sú spojené s negatívnym ekonomickým vývojom (Mládek, J., 2006), ktorý nastal po roku 1989, kedy došlo k celkovej socio-ekonomickej tranformácii spoločnosti. Tento negatívny ekonomický vývoj ešte výrazne zintenzívnil priebeh zmien.

V hodnotovom rebríčku obyvateľov došlo k významným zmenám prevláda rast individualizmu a emancipácie žien (Marenčáková, J., 2006, str.198). Dôraz sa kladie na sebarealizáciu, kariéru a osobný úspech. Tradičné hodnoty rodiny sú tak ohrozené, čo ovplyvňuje hlavne sobášnosť a následne aj pôrodnosť ako kľúčový proces demografického vývoja obyvateľstva ,čo sa prejavuje novými vzorcami rodinného a reprodukčného správania obyvateľstva.

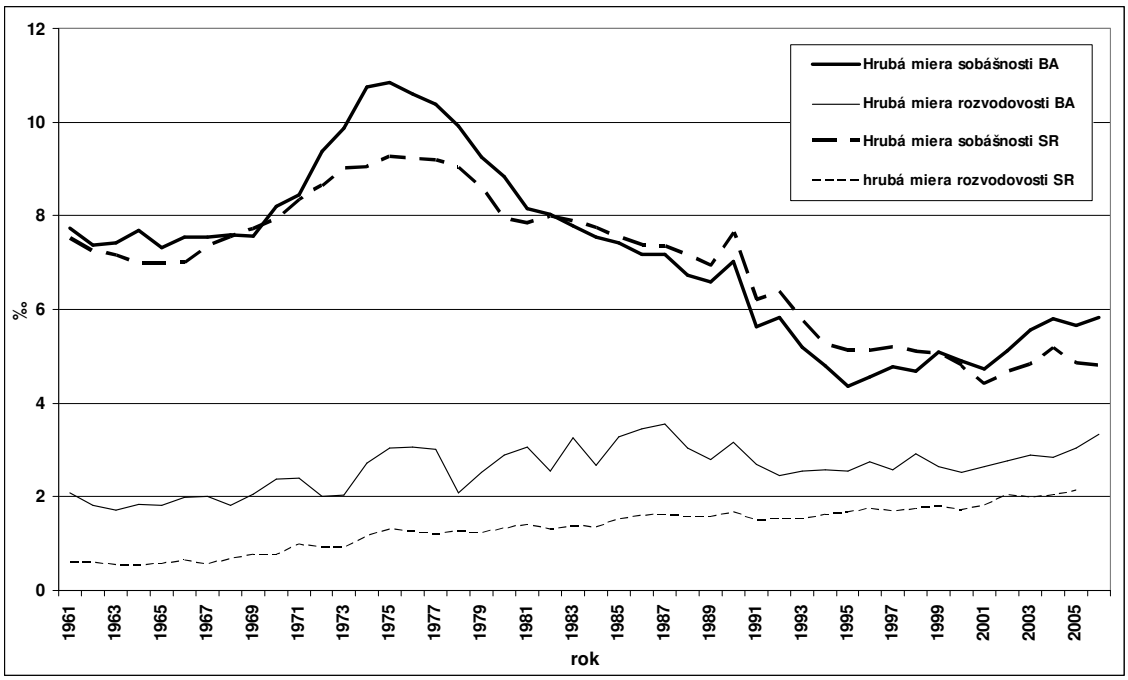
Mestské obyvateľstvo sa pritom vyznačuje rozdielnym demografickým správaním a má rozdielne špecifiká. Vyznačuje sa nižšou plodnosťou, vyššou strednou dĺžkou života, lepším zdravím. Môžeme teda povedať, že urbanizácia ešte zrýchľuje priebeh druhého demografického prechodu (Population reference bureau, 2004, str.12).

2. Sobášnosť a rozvodovosť

Od roku 1975 kedy sobášnosť (vyjadrená hrubou mierou) obyvateľov Bratislavy dosahovala maximálnu hodnotu 10,85 ‰, začala postupne klesať. Po roku 1991 zaznamenáva

sobášnosť prudký pokles a v polovici 90. rokov dosiahla sobášnosť minimálnu hodnotu 4,36 ‰. V absolútnych číslach bol počet sobášov v roku 1995 iba 1 968 čo je takmer o polovicu menej ako v roku 1975 kedy ich bolo 3 657. Podobný pokles v sobášnosti po roku 1991 zaznamenáva aj celé Slovensko. Tento pokles bol zapríčinený odkladaním sobášov kvôli zlým socio-ekonomickým podmienkam v tom čase (Mládek, J., 1999, str.65). Vaňo poukazuje na nové možnosti osobnej realizácie, ktoré začali konkurovať zakladaniu rodín, tak isto negatívny vplyv spôsobil rast životných nákladov, inflácia, nezamestnanosť, obmedzenie bytovej výstavby a obmedzenie podpory rodín štátom (Vaňo, B., 2005, str.103). Po roku 1995 však zaznamenáva Bratislava nárast sobášov, čo môže byť spôsobené príchodom početných vekových kategórií, narodených v 70 rokoch, do veku kedy sa obyvatelia Bratislavy sobášia (graf č.2). Podiel na náraste sobášnosti obyvateľstva môže mať aj pozitívny ekonomický vývoj Bratislavy a teda realizácia odložených sobášov z počiatočného obdobia socio-ekonomickej transformácie.

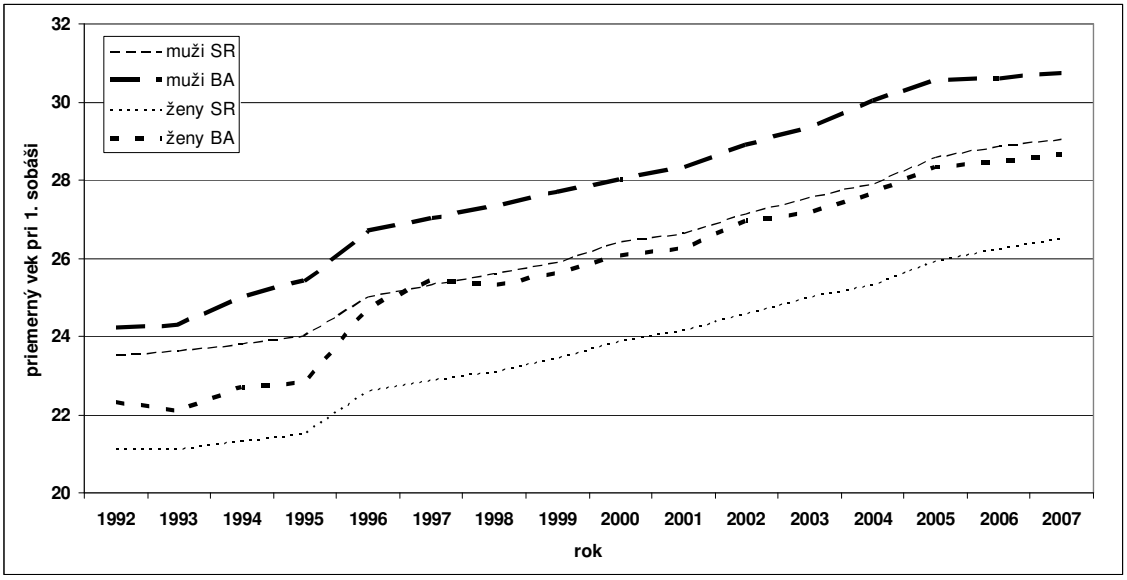
Rozvodovosť (vyjadrená hrubou mierou) obyvateľov Bratislavy zaznamenáva v celom sledovanom období iba pomalý nárast a počas celého obdobia dosahuje výrazne vyššie hodnoty ako celoslovenský priemer. Vyššie hodnoty rozvodovosti sú typické pre mestské obyvateľstvo, ktoré má výrazne odlišné demografické správanie ako obyvateľstvo vidiecke. Obyvatelia miest upúšťajú od tradícií rodiny a náboženských pravidiel a sú otvorenejší aj k otázke rozvodu. Rozvodovosť obyvateľov Bratislavy dosahovala úroveň 2 ‰ už v 60. rokoch, pokiaľ celoslovenský priemer dosiahol túto hodnotu až v roku 2002.



Graf č.1: Vývoj sobášnosti a rozvodovosti obyvateľov Bratislavy a SR (vyjadrené hrubými mierami)

Priemerný vek ženícha a nevesty pri 1. sobáši zaznamenal od začiatku 90. rokov extrémny nárast. Nárast priemerného veku pri sobáši, resp. pri 1. sobáši, je prejavom druhého demografického prechodu, kedy mladí ľudia odkladajú sobáše, z dôvodu štúdia, kariéry, alebo preferujú niekoľkoročné spolužitie ako skúšku pred svadbou. Zvyšovanie priemerného veku pri 1. sobáši sa začalo prejavovať v 70. rokoch v západnej Európe. V priebehu 25 rokov sa zvýšil priemerný vek pri 1. sobáši neviest o 4 - 6 rokov (Bartoňová, D., 2001, str.50).

Avšak taký extrémny nárast ako zaznamenalo Slovensko a Bratislava spôsobila socio-ekonomická transformácia po roku 1989, kedy sa pre mladých ľudí strácajú sociálne a ekonomické istoty, ktoré im poskytoval socialistický štát, čo zapríčinilo odkladanie sobášov. Priemerný vek ženícha pri 1. sobáši vzrástol iba v priebehu 14 rokov z 24 až na viac ako 31 rokov pre ženíchov a z 22 na viac ako 28 rokov pre nevesty. Hodnoty celoslovenského priemeru sú naproti tomu výrazne nižšie, čo svedčí o výrazne odlišnom správaní mestského obyvateľstva. Výrazne vyšší vek pri 1. sobáši najmä u neviest vedie k nepriaznivému vývoju pôrodnosti a reprodukcii populácie



Graf č.2: Priemerný vek pri 1. sobáši obyvateľov Bratislavy a SR

3. Plodnosť

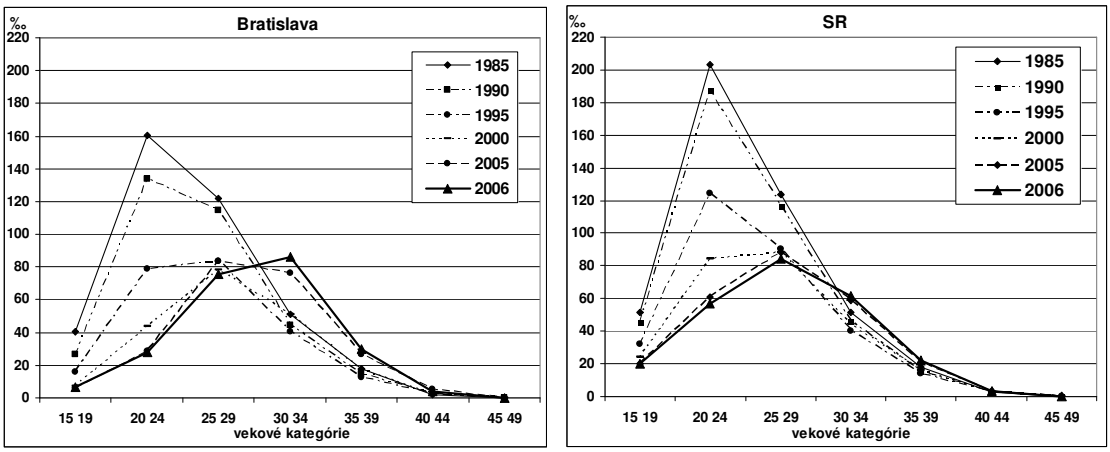
Vývoj vekovo špecifických mier plodnosti, tak ako Bratislavy, aj Slovenska bol významne zasiahnutý zmenami v demografickom správaní typickými pre druhý demografický prechod. U žien aj u mužov dochádza k zmenám v hodnotovom rebríčku obyvateľov, kedy na prvom mieste nie je dieťa, ale kariéra. Individualizmus, ktorý je typickým prejavom post-industriálnej spoločnosti, nahrádza altruizmus (Mládek, J., 1999, str.60). K tomu sa po roku 1989 pridali aj zmeny v socio-ekonomickej situácii na Slovensku a to rast nákladov spojených so starostlivosťou o dieťa, dočasný pokles reálnych príjmov domácností, finančná nedostupnosť bytov, hrozba nezamestnanosti a nová sociálna politika voči rodinám (zníženie novomanželských pôžičiek, regulácia prídavkov na deti) (Marenčáková, J., 2006, str.198). V neposlednom rade sa na znížení plodnosti žien podieľalo aj výrazne vyššie používanie antikoncepcie.

Tieto všetky faktory majú vplyv v prvom rade na pôrodnosť. Pokým v roku 1978 bol počet narodených detí 7 474 v roku 2001 ich bolo iba 3 139 čo je viac ako o polovicu menej. Druhým dôsledkom spomínaných zmien je posun maximálnej hodnoty vekovo špecifických mier plodnosti do vyšších vekov a znižovanie plodnosti v nižších vekových kategóriách – odkladanie narodenia dieťaťa.

Pokým v roku 1985 dosahovali vekovo špecifické miery plodnosti Bratislavy aj Slovenska jedno maximum vo vekovej kategórií 20-24 ročných žien (Bratislava 160,6 ‰, Slovensko 203,4 ‰). V roku 2006 dosahuje Slovensko jedno maximum (84,5 ‰) vo vekovej kategórii 25-29 ročných a zvýšili sa aj špecifické miery plodnosti vo vyšších vekových kategóriách 30-34 a 35-39 ročných. V prípade Bratislavy bol tento vývoj ešte rýchlejší a jedno maximum vo vekovej kategórií 25-29 ročných zaznamenala Bratislava už v roku 2000, pričom v priebehu ďalších šiestich rokov sa situácia zmenila a Bratislava zaznamenáva

jedno maximum (86,2 ‰) už vo vekovej kategórií 30-34 ročných. Tento posun zodpovedá zvyšovaniu priemerného veku pri 1. sobáši, čo následne posúva aj narodenie dieťaťa. Pozitívne môžeme hodnotiť pokles plodnosti v oboch prípadoch v najnižšej vekovej kategórií 15-19 ročných žien.

Úhrnná plodnosť žien Bratislavy dosahovala už v roku 1985 hodnotu 1,9 čo je pod hranicou rozšírenej reprodukcie (2,1). Slovensko dosahovalo v roku 1985 ešte stále hodnotu nad hranicou rozšírenej reprodukcie a to 2,25. Úhrnná plodnosť žien Bratislavy rapídne klesala najmä medzi rokmi 1990 a 1995 kedy klesala z 1,7 na 1,17, čo už sa hodnotí ako extrémne nízka plodnosť. Najnižšiu hodnotu dosiahla úhrnná plodnosť v roku 2000 a to alarmujúcich 0,99. Po roku 2000 však Bratislava zaznamenáva mierny nárast v sobášnosti a pôrodnosti čo zdvihlo aj ukazovateľ úhrnnej plodnosti na 1,15 v roku 2006 čo sa približuje úhrnnej plodnosti Slovenska v roku 2006 na hodnotu 1,24. Tento mierny nárast v úhrnnej plodnosti môže byť spôsobený priaznivým ekonomickým vývojom v regióne Bratislavy a čiastočnou rekuperáciou odložených pôrodov.



Graf č.3: Vekovo špecifické miery plodnosti žien Bratislavy a SR

4. Záver

Bratislava, tak ako aj celé Slovensko, zaznamenala v posledných 20 rokoch, významné zmeny v demografickom správaní obyvateľstva, ktoré môžeme nazývať druhým demografickým prechodom. Tieto zmeny sa prejavili najmä v reprodukčnom správaní a rodinnom správaní. Po roku 1989 bol priebeh zmien zintenzívnený celkovou socio-ekonomickou transformáciou spoločnosti.

V prípade Bratislavy však nachádzame mnohé výraznejšie rozdiely v priebehu a intenzite zmien oproti celému Slovensku. Bratislavu môžeme v tomto smere nazvať jadrovým regiónom mnohých zmien. Bratislava už v roku 1985 dosiahla hodnotu úhrnnej plodnosti pod hranicu rozšírenej reprodukcie (2,1) a to 1,97, kedy ešte celé Slovensko dosahovalo hodnotu 2,25. Celá Bratislava zaznamenala prirodzený úbytok už v roku 1995 oproti Slovensku v roku 2001. V roku 2000 dosahuje hodnota úhrnnej plodnosti žien Bratislavy dokonca iba 0,99, čo je hlboko pod hodnotou extrémne nízkej plodnosti. Môžeme teda povedať, že zmeny v pôrodnosti a plodnosti žien Bratislavy nastávajú omnoho skôr a dosahujú vyššiu intenzitu.

Odlišnosti nachádzame aj v ukazovateľoch rodinného správania, ktoré sa výrazne mení so zmenou v životnom štýle a priorít mladých ľudí. Muži aj ženy v Bratislave vstupujú omnoho neskôr do manželstva čo dokumentuje omnoho vyšší priemerný prvosobášny vek Bratislavčanov v porovnaní s celým Slovenskom, tak pre ženícha ako aj nevestu. To sa premieta aj do neskoršej pôrodnosti. Bratislava v roku 2006 zaznamenáva maximálne hodnoty

plodnosti vo vekovej kategórii 30 – 34 ročných žien, pokým Slovensko v kategórii 25 - 29 ročných. Priemerný vek pri 1. pôrode je tak isto výrazne vyšší v Bratislave v porovnaní s celoslovenským priemerom. Ženy aj muži v Bratislave dosahujú viac ako o 2 roky vyšší priemerný vek pri 1. pôrode ako celoslovenský priemer. Výrazne vyššie hodnoty, v porovnaní so Slovenskom, zaznamenáva dlhodobo aj rozvodovosť, ktorá negatívne vplýva na výchovu detí a reprodukciu.

5. Použitá literatúra

[1]BARTOŇOVÁ, D., 2001, Demografické chování populace České republiky v regionálním a Evropském kontextu

[2]MARENČÁKOVÁ, J., 2003, Pôrodnosť obyvateľstva Slovenska a jej vzťah s vybranými demografickými a spoločenskými javmi. In: Acta Facultatis Rerum Naturalium Universitatis Comenianae Geographica Nr. 44, 3-38.

[3]MARENČÁKOVÁ, J., 2006, Reprodukčné a rodinné správanie obyvateľstva Slovenska po roku 1989 z časového a priestorového aspektu. In: Geografický časopis, 58, 3, 197-223

[4]MLÁDEK, J., 1999, Population developement in Slovakia in the European context. In: Acta Facultatis Rerum Naturalium Universitatis Comenianae Geographica Supplementum No 2/I, 59-71.

[5]MLÁDEK, J., KÁČEROVÁ, M., MARENČÁKOVÁ, J., et al., 2006, New Trends of Popilation Development in Slovakia at the end of the 20th Century and Begining of the 21st Century. In: Acta Geographica Universitatis Comenianae No. 48, 189-210.

[6]MLÁDEK, J., KUSENDOVÁ, D., MARENČÁKOVÁ, J., et al. 2006, Demografická analýza Slovenska. Univerzita Komenského v Bratislave.

[7]POPULATION REFERENCE BUREAU, 2004, Transitions in World Population. Population Bulletin, 59, 1, 1-40.

[8]RYCHTAŘÍKOVÁ, J., 1999, Is Eastern Europe expriencing a second demographic transition?. In: Acta Universitatis Carolinae Geographica, No 1, 19-44.

[9]VAN DE KAA, D. J., 1987, Europe's Second Demographic Transition. In: Population Bulletine, 42, 1, 1-57.

[10] VAŇO, B., 2005, Populačný vývoj na Slovensku po roku 1990. In: Demografie, 47, 2, 103-112.

6. Zdroje

[1]KRAJSKÁ SPRÁVA ŠTATISTICKÉHO ÚRADU SR (1993-2007). Štatistická ročenka hlavného mesta SR Bratislavy (1993-2007). Bratislava (ŠU SR).

[2]MESTSKÁ SPRÁVA SLEVENSKÉHO ŠTATISTICKÉHO ÚRADU V HLAVNOM MESTE SSR BRATISLAVE, (1985-1986), Štatistická ročenka hlavného mesta SSR Bratislavy (1985-1986), Bratislava.

[3]ŠTATISTICKÝ ÚRAD SR, (1992-2006), Pohyb obyvateľstva v Slovenskej republike v roku (1992-2006) (Pramenné diela). Dostupné na internete: http://portal.statistics.sk/showdoc.do?docid=6674

[4]VÝSKUMNÉ DEMOGRAFICKÉ CENTRUM, 2007, Population information (POPIN). dostupné na internete: http://www.infostat.sk/slovakpopin/data/maindata.xls

Michal Katuša Bc.

Čipkárska 3, 821 09 Bratislava

michal.katusa@gmail.com, aso@chello.sk,

Zamestnanosť, vzdelanosť a príjmy v high-tech odvetviach krajín EÚ
Employment, education and earnings in the EU high-tech sectors

Andrej Bajdich

Abstract: This article is focusing on the analysis of the employment, education level and the earnings in the high-tech manufacturing sector and in the high-tech KIS sector in the EU-27 countries and in selected countries. It also details the difference in the wages of men and women depending on theyre level of education and explains which sub-sectors are included in the high-tech manufacturing and in the KIS sector. Since the high-tech manufacturing sector and in the high-tech KIS sector are often seen as major engines of growth in modern economics, it's quite interesting, how the individual countries are ranked, when they're compared with each other.

Key words: Employment, education, earnings, EU, high-tech sectors, high-tech manufacturing, share of women and men in high-tech sectors

Kľúčové slová: zamestnanosť, príjmy, EÚ, high-tech odvetvia, high-tech spracovateľský priemysel, podiel žien a mužov v high-tech odvetviach EÚ

1. Úvod

Príspevok sa zameriava na porovnanie zamestnanosti a príjmov v „high-tech" odvetviach 27 krajín EÚ a niektorých vybraných krajín. Snahou je aj poukázať na rozdielnosť úrovne jednotlivých skúmaných ukazovateľov podľa pohlavia.

Pod „high-tech" odvetviami sa rozumie hlavne sektor *high-tech spracovateľského odvetvia* a *sektor služieb s vysokou technológiou*, ktoré sú uvedené v medzinárodnej klasifikácii ekonomických činností NACE[9] pod kódmi 24, 30-34, 352, 354, 355 resp. 64, 72 a 73. Pre jednoduchšiu formuláciu budeme používať Pre služby s vysokou technológiou v nasledujúcom texte požívanú anglickú skratku KIS (Knowledge-intensive services), ktorou sú označené aj v medzinárodnej klasifikácii ekonomických činností – NACE.

2. Zamestnanosť a vzdelanosť v high-tech odvetviach krajín EÚ a vybraných krajín

V roku 2006 v krajinách EÚ-27 pracovalo v službách 142 miliónov ľudí, v porovnaní so spracovateľským priemyslom, v ktorom bolo zamestnaných 39 miliónov ľudí. Z uvedených 39 miliónov osôb zamestnaných v priemysle, ktorý zahŕňa výrobu farmaceutických výrobkov, kancelárskych strojov a počítačov, rádiových, televíznych a komunikačných zariadení a zdravotníckych prístrojov , bolo v high-tech sektore zamestnaných zhruba 2,3 milióna osôb, čo v relatívnom vyjadrení predstavovalo zhruba 1,1 % z celkového počtu zamestnaných osôb v 27 krajinách EÚ (Tabuľka č.1). Z celkového počtu 142 miliónov pracujúcich v službách v EÚ-27 pracovala v KIS odvetviach menej ako polovica. Z toho však v high-tech odvetviach KIS, ktoré zahŕňajú služby pôšt a telekomunikácií, počítačové a súvisiace činnosti ako aj služby vo výskume a vývoji, pracovalo len cca 7 miliónov ľudí. V high-tech odvetviach spracovateľského sektoru a high-tech KIS bolo v rámci EÚ-27 v roku 2006 spolu zamestnaných do 9,3 milióna osôb.

[9] Rozdelenie je uvedené podľa NACE revízie č. 1.1. V súčasnosti bola publikovaná 2.revízia NACE. Pre podrobnejšie informácie viď [1]

Najvýznamnejšie absolútnym počtom zamestnaných osôb prispelo Nemecko, ktorého zamestnanosť v týchto sektoroch predstavovala viac ako 1,9 mil. ľudí. Tento výsledok možno nie je až taký prekvapivý, keďže nemecká vláda každoročne vyčleňuje na podporu výskumu a vývoja niekoľko miliárd EUR. V roku 2009 plánuje Nemecko pre tento sektor vyčleniť až 15 miliárd EUR.[10] Absolútnym počtom zamestnaných v high-tech odvetviach bola druhá Veľká Británia a tretie Severné Írsko. Obe krajiny zamestnávali po viac ako 1 mil. ľudí v high-tech KIS sektore. Porovnaním relatívnych hodnôt sa najlepšie umiestnilo Fínsko a Írsko, ktoré dosiahli najvyššie hodnoty podielu zamestnaných osôb v oboch sektoroch z počtu zamestnaných – Fínsko 6,7 % a Írsko 6,6 %. Z grafu č.1 je zrejmé, že zamestnanosť v high-tech sektoroch vo vzťahu k celkovej zamestnanosti je najvyššia v severných regiónoch EÚ[11].

Tabuľka č.1: Zamestnanosť a vzdelanosť v high-tech odvetviach spracovateľského priemyslu a v sektore služieb za rok 2006 v krajinách EÚ-27 roztriedená podľa pohlavia a zobrazená za 1000 osôb a ako percento z celkového počtu zamestnaných osôb.

	Spracovateľský priemysel						KIS					
	High-tech						High-tech					
	Celkom		Pohlavie (%)		Vzdelanie (%)		Celkom		Pohlavie (%)		Vzdelanie (%)	
	v 1000	% z celk. zam.	ženy	muži	základné a stredoškolské	vysoko-školské	v 1000	% z celk. zam.	ženy	muži	základné a stredoškolské	vysoko-školské
Austria	53,44	1,4	30,5	69,5	1,2	1,6	107,84	2,8	28,5	71,5	1,9	3,7
Belgium	28,35	0,7	33,2	66,8	0,5	0,9	166,95	3,9	29,7	70,3	2,8	5,5
Bulgaria	16,43	0,5	52,8	47,2	-	-	80,43	2,6	47,4	52,6	-	4,9
Cyprus	0,52	0,1	-	-	-	-	7,01	2,0	31,1	68,9	-	3,5
Czech Republic	80,77	1,7	49,1	50,9	1,8	1,9	141,61	2,9	43,1	56,9	3,1	6,2
Denmark	22,16	0,8	42,4	57,6	0,8	1,0	123,14	4,4	33,9	66,1	2,9	5,7
Estonia	6,82	1,1	-	-	-	-	16,42	2,5	-	59,9	-	3,8
Finland	51,18	2,1	29,1	70,9	1,2	3,5	112,92	4,6	36,2	63,8	2,8	5,7
France	276,87	1,1	34,0	65,5	0,8	1,8	967,82	3,9	37,3	62,7	3,5	6,9
Germany	635,01	1,7	31,8	68,2	1,5	2,2	1294,17	3,5	32,5	67,5	2,3	4,6
Greece	10,59	0,2	24,9	75,1	-	0,4	88,25	2,0	31,0	69,0	1,6	3,2
Hungary	97,57	2,5	51,2	48,8	3,1	1,6	134,31	3,4	40,5	59,5	3,2	6,4
Ireland	53,17	2,7	40,9	59,1	2,1	3,5	78,02	3,9	28,0	72,0	3,1	6,2
Italy	293,69	1,3	31,6	68,4	1,3	1,2	702,49	3,1	34,4	65,6	2,5	5,1
Latvia			-	-	-	-	27,44	2,5	48,9	51,1	-	3,9
Lithuania	9,33	0,6	-	-	-	-	30,72	2,1	54,0	46,0	-	3,2
Luxembourg			-	-	-	-	6,40	3,3	27,0	73,0	2,2	4,4
Malta	4,77	3,1	44,0	56,0	-	-	4,73	3,1	-	73,8	-	-
Netherlands	51,16	0,6	21,1	78,9	0,5	0,8	312,44	3,8	26,1	73,9	2,6	5,2
Poland	84,48	0,6	43,7	56,3	-	0,6	345,92	2,4	39,5	60,5	2,4	4,8
Portugal	21,70	0,4	43,5	56,4	-	-	94,03	1,9	32,7	67,3	2,1	4,2
Romania	28,60	0,3	37,7	62,3	-	0,8	149,85	1,6	46,3	53,7	2,4	4,8
Slovakia	41,01	1,8	59,9	40,1	2,4	1,2	58,87	2,6	43,7	56,3	-	5,3
Slovenia	10,49	1,1	47,3	52,7	1,2	0,9	26,15	2,7	28,6	71,4	-	4,6
Spain	87,57	0,4	32,5	67,5	0,4	0,7	588,85	3,0	31,6	68,4	2,7	5,3
Sweden	39,81	0,9	32,1	68,0	0,8	1,0	223,97	5,1	31,7	68,3	3,9	7,8
UK	287,53	1,0	29,8	70,2	0,8	1,5	1186,06	4,2	24,2	75,8	2,8	5,5
EU-27	2295,02	1,1	34,8	65,2	0,9	1,4	7076,80	3,3	32,9	67,1	2,7	5,4

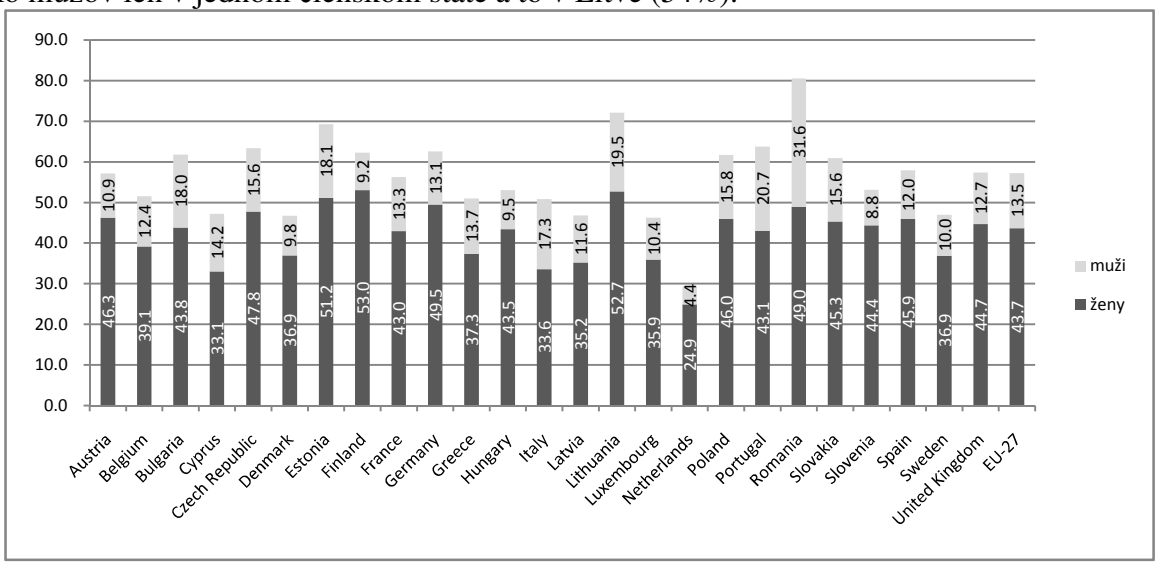
% z celk.zam : percento z celkového počtu zamestnaných osôb
- : chýbajú hodnoty
Zdroj: Eurostat, High-tech statistics

Umiestnenie krajín V4 podľa relatívneho vyjadrenia zamestnanosti v high-tech odvetviach spracovateľského sektoru a KIS – Maďarsko, Česko a Slovensko nad resp. na úroveň 4,4 % z celkového počtu zamestnaných je na priemernej úrovni všetkých 27 krajín EÚ a len Poľsko mierne zaostalo, pričom dosiahlo hodnotu 3,0 % z celkového počtu zamestnaných osôb. V roku 2006 predstavoval pomer žien zamestnaných v high-tech sektoroch v krajinách EÚ zhruba 1 tretinu zo všetkých zamestnancov v týchto sektoroch. V high-tech sektore spracovateľského priemyslu predstavoval tento pomer 34,8%, čo bolo viac ako pomer žien pracujúcich v high-tech sektore KIS, ktorý predstavoval 32,9%. Pritom hranicu 50% zamestnaných žien sa v sektore high-tech spracovateľského priemyslu podarilo

[10] BUNDESMINIST ERIUM FÜR BILDUNG UND FORSCHUNG . The High-Tech Strategy for Germany. In: Bundesministerium für Bildung und Forschung, 2006, s. 2.

[11] [3] FELIX, B . Statistics in focus - Employment and earnings in high-tech sectors. In: European Communities, 2007, s. 2.

prekročiť len v pomerne mladších, resp. nových členským štátoch: Slovensku (59,9%), Maďarsku(51,2%) a Bulharsku(52,8%). V high-tech sektore KIS bolo viac zamestnaných žien ako mužov len v jednom členskom štáte a to v Litve (54%).[12]



V grafe nie sú, kvoli chýbajúcim údajom zahrnuté nasledovné krajiny: Írsko a Malta
Zdroj: Eurostat, High-tech statistics

Graf č.1: Pomer terciálne vzdelaných žien a mužov v high-tech odvetviach krajinách EÚ-27 za rok 2006 z hľadiska pohlavia a zobrazená ako percento z ekonomicky aktívneho obyvateľstva.

Graf č.1 zobrazuje pomer vysokoškolsky vzdelaných mužov a žien pracujúcich v high-tech odvetviach EÚ za rok 2006, ktorí ukončili a pracujú v jednom z odborov zameraných na vedu, matematiku, informatiku, stavebníctvo, techniku alebo výrobu. Z grafu je zrejmé, že v spomínaných high-tech odvetviach pracuje viac vysokoškolsky vzdelaných žien ako mužov. Najväčší počet vysokoškolsky vzdelaných ľudí so zameraním na spomínané odbory pracovalo v Rumunsku(80,6%), Litve(72,1%) a v Estónsku (69,3%), pričom vo Fínsku, Litve a v Estónsku pracovalo v týchto odboroch viac ako polovica žien, ktoré mali ukončené vysokoškolské vzdelaní, kde pomer terciálne vzdelaných žien predstavoval 53%, resp. 52,7%, resp. 51,2 %. Hoci ženy dosahujú vysoké profesionálne vzdelanie ich podiel na celkovej zamestnanosti v rámci high-tech sektorov v EÚ prestavuje zhruba 1 tretinu.

2. Príjmy v high-tech odvetviach krajín EÚ a vybraných krajín

Nevyváženosť medzi pohlaviami je možné pozorovať nielen z hľadiska zamestnanosti a vedomostí, ale aj z hľadiska príjmov. V tabuľke č.2 sú zobrazené ročné príjmy prepočítané v EUR na jedného zamestnanca v odvetví spracovateľského priemyslu a v sektore služieb v roku 2002 v krajinách EÚ-27. Sú v členení podľa pohlavia. Vyplýva z nej, že ženy zarábali menej ako muži, bez ohľadu na sektor alebo krajinu. Vo všeobecnosti možno sledovať väčšiu mzdovú diferenciáciu pohlaví v high-tech odvetvi spracovateľského priemyslu oproti high-tech odvetviu KIS, pričom najväčší rozdiel v ročných platoch mužov a žien sa vyskytol v Belgicku, Nemecku a v Spojenom kráľovstve. Na druhej strane, okrem 5 krajín - Rakúska, Belgicka, Fínska, Francúzka a Talianska, kde mali muži väčšie platy v high-tech spracovateľskom sektore, boli mzdy oboch pohlaví vyššie v high-tech sektore KIS. V tomto sektore zarábali v roku 2002 najviac ženy a muži žijúci v Dánsku, Luxembursku a vo Spojenom Kráľovstve najmenej v Litve, Rumunsku a Bulharsku. Dôležitý je aj fakt, že

[12] FELIX, B . Statistics in focus - Employment and earnings in high-tech sectors. In: European Communities, 2007, s. 2.

priemerné ženské príjmy v high-tech odvetviach, na rozdiel od tých mužských, nedosahujú v niektorých krajinách(napr. Rakúsko, Slovinsko), ani len úroveň priemerných miezd zamestnancov s ukončeným základným resp. stredoškolským vzdelaním.

Tabuľka č.2: Ročné príjmy prepočítané v EUR na 1 zamestnanca v odvetví spracovateľského priemyslu a v sektore služieb za rok 2002 v krajinách EÚ-27 a vybraných krajinách roztriedená podľa pohlavia a úrovne ukončeného vzdelania.

	Spracovateľský priemysel				KIS			
	High-tech				**High-tech**			
	Pohlavie		**Vzdelanie**		**Pohlavie**		**Vzdelanie**	
	ženy	muži	základné a stredoškolské	vysoko-školské	ženy	muži	základné a stredoškolské	vysoko-školské
Austria	30333	45060	36071	57661	30442	39815	35339	58294
Belgium	26614	44518	29211	51996	31454	38062	25977	47332
Bulgaria	1645	1853	1808	2378	2231	2739	1978	3241
Cyprus	-	-	4167	-	20039	27519	22562	29380
Czech Republic	5183	8078	5095	12629	6341	10428	5370	11166
Denmark	30657	43904	33154	49586	42619	55443	45296	58253
Estonia	3838	6082	4243	6328	5010	9335	4784	8943
Finland	29149	38993	27659	41296	29987	35334	27637	39248
France	26462	38604	17349	47092	30429	38424	27665	42645
Germany	30745	46473	38204	65410	33668	46663	37053	58342
Greece	13795	24566	15664	28354	17661	26486	20090	28782
Hungary	4658	6735	3100	12705	6954	10069	3473	16112
Ireland	29630	38081	25442	42646	31464	37708	30211	41031
Italy	21933	29659	25881	40150	25047	27211	22555	38211
Latvia	2793	3438	2805	3792	4389	7570	3739	8255
Lithuania	3486	5699	3725	6297	4117	6655	2732	7192
Luxembourg	22458	32808	38191	45720	41471	52609	43627	57911
Netherlands	28970	40314	32224	47262	36983	38566	35020	47463
Poland	6389	8615	6802	13256	8754	10958	7754	14736
Portugal	11519	20653	12696	36742	21679	25319	20079	37245
Romania	2135	2681	1774	4441	3466	3985	2938	5760
Slovakia	4023	6381	3656	9230	4448	7075	4018	9771
Slovenia	8014	13707	8539	21544	16108	17545	12234	23130
Spain	20006	27990	13694	30801	22897	31585	20936	32371
Sweden	-	-	-	-	34409	46192	40438	47379
United Kingdom	28805	44035	31811	52119	40491	50916	39490	59441

- : chýbajú hodnoty
Zdroj: Eurostat, High-tech statistics

3. Záver

Sektory high-tech spracovateľského priemyslu a služby v týchto odvetviach (KIS) sa často pokladajú za hnací motor moderných ekonomík. V roku 2006 bolo v uvedených sektoroch zamestnaných okolo 9,3 milióna ľudí z z celkového počtu zamestnaných osôb v 27 krajinách EÚ. Z hľadiska ročného príjmu a pohlavia je možné tvrdiť, že ženy zarábajú vo všeobecnosti menej ako muži bez ohľadu na sektor alebo krajinu. Hoci ženy dosahujú vysoké profesionálne vzdelanie ich podiel na celkovej zamestnanosti v rámci týchto sektorov v EÚ prestavuje približne jednu tretinu. Hranica 50 % pomeru zamestnaných žien high-tech odvetviach bola dosiahnutá a prekročená len u troch členských krajinách. Európska únia by sa mala zamerať na zvyšovanie počtu žien a dosiahnutie väčšej porovnateľnosti platov žien v rámci high-tech odvetví a smerovať k napĺňaniu cieľa EÚ, ktorý bol stanovený v Lisabonskej zmluve. Je ním do roku 2010 dosiahnuť konkurencieschopnosť ekonomiky založenej na vedomostiach.

4. Literatúra

[1] EUROSTAT. 2008. NACE Rev. 2 - Statistical classification of economic activities in the European Community. In: Eurostat, 2008.

http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_NACE_REV_2/PGE_DS_NACE_REV_2/NACEREV2STRUCTURE.PDF

[2] BUNDESMINIST ERIUM FÜR BILDUNG UND FORSCHUNG . 2006. The High-Tech Strategy for Germany. In: Bundesministerium für Bildung und Forschung, 2006, s. 2.

http://www.bmbf.de/pub/bmbf_hts_lang_eng.pdf

[3] FELIX, B . 2007. Statistics in focus - Employment and earnings in high-tech sectors. In: Eurostat, 2007.

http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-SF-07-032/EN/KS-SF-07-032-EN.PDF

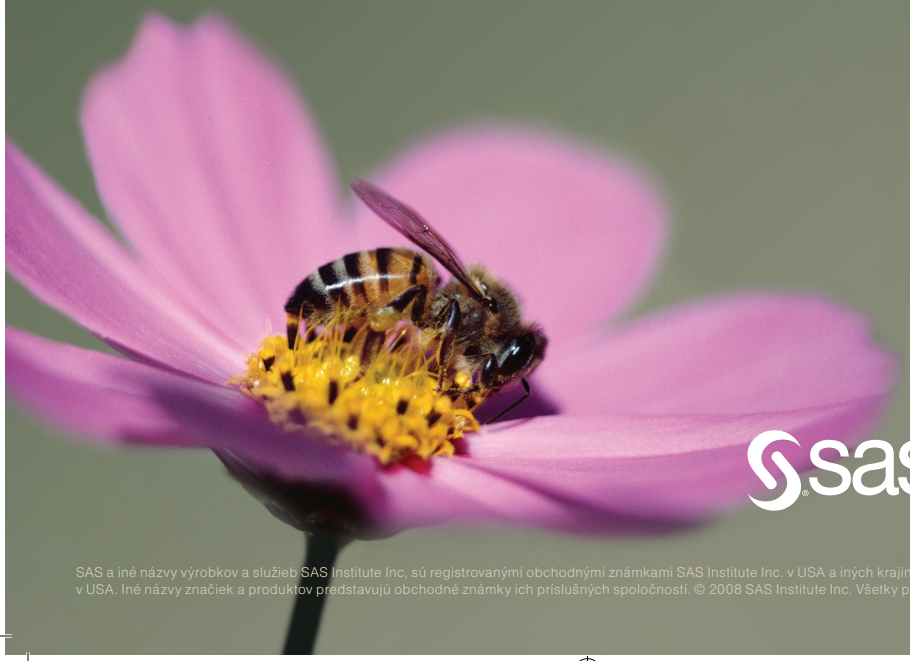
Adresa autora:

Andrej Bajdich, Bc., 5. ročník,
Ekonomická univerzita v Bratislave
E-mail: cquence1@gmail.com

Jedna včela vyrobí za svoj život jednu dvanástinu lyžičky medu

Nevidí ako sa jej tvrdá práca pretaví na niečo veľké a aký má skutočne význam.
Vy to však vidieť môžete. S osvedčeným softvérom pre vzdelávanie od spoločnosti SAS.

www.sas.com/slovakia/uni
www.sas.com/bees



sas | THE POWER TO KNOW.