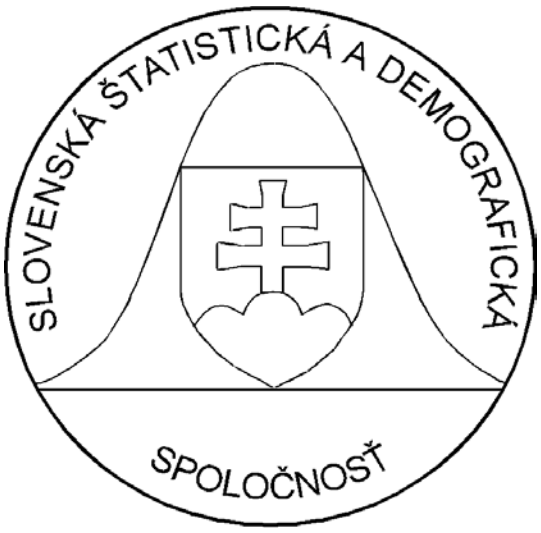
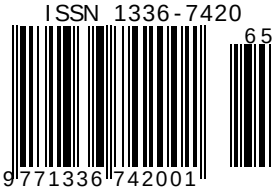
5/2006

FORUM STATISTICUM SLOVACUM



ISSN 1336-7420



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

Naše najbližšie akcie:

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

KONFERENCIA POHĽADY NA EKONOMIKU SLOVENSKA 2007,
tematické zameranie: *Makroekonomický vývoj; zamestnanosť*
27. 3. 2007, Bratislava

EKOMSTAT 2007, 21. škola štatistiky
Tematické zameranie: *Štatistické metódy v praxi*
3. – 8. 6. 2007, Trenčianske Teplice

11. SLOVENSKÁ DEMOGRAFICKÁ KONFERENCIA,
tematické zameranie: *Migrácia*
17. – 19. 9. 2007, hotel Čingov, Slovenský raj

FernStat 2007
IV. medzinárodná konferencia aplikovanej štatistiky
(**F**inancie, **E**konomika, **R**iadenie, **N**ázory)
tematické zameranie: *Aplikovaná, demografická, matematická štatistika, štatistické riadenie kvality*.
4. – 5. 10. 2007, hotel Lesák, Tajov pri Banskej Bystrici

16. Medzinárodný seminár VÝPOČTOVÁ ŠTATISTIKA,
6. – 7. 12. 2007, Bratislava

Prehliadka prác mladých štatistikov a demografov
6. 12. 2007, Bratislava

SLÁVNOSTNÁ KONFERENCIA 40 ROKOV SŠDS,
marec 2008, Bratislava

ÚVOD

Vážené kolegyne, vážení kolegovia,
piate číslo druhého ročníka časopisu, ktorý vydáva Slovenská štatistická a demografická spoločnosť (SŠDS) je zostavené z príspevkov, ktoré autori pripravili pre 15. Medzinárodný seminár výpočtová štatistika a na Prehliadku prác mladých štatistikov a demografov. Táto akcia sa uskutočila v dňoch 7. a 8. decembra na Infostate v Bratislave.

Akciu, z poverenia Výboru SŠDS, zorganizoval Organizačný a programový výbor: doc. Ing. Jozef Chajdiak, CSc. – predseda, RNDr. Ján Luha, CSc. – tajomník, RNDr. Peter Mach, Doc. RNDr. Beáta Stehlíková, CSc., Doc. RNDr. Bohdan Linda, CSc., Doc. Dr. Jana Kubanová, CSc., RNDr. Jitka Bartošová, Ing. Vladimír Úradníček PhD., RNDr. Samuel Koróny.

Na príprave a zostavení tohoto čísla participovali: doc. Ing. Jozef Chajdiak, CSc., RNDr. Ján Luha, CSc.

Recenziu príspevkov zabezpečili: doc. Ing. Jozef Chajdiak, CSc., RNDr. Ján Luha, CSc., RNDr. Peter Mach, RNDr. Samuel Koróny.

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.

Organizátori seminára si považujú za milú povinnosť poďakovať za podporu predsedovi Štatistického úradu SR RNDr. Petrovi Machovi a Infostatu.

Výbor SŠDS

FORUM STATISTICUM SLOVACUM

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
II.

Číslo
5/2006

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

Z HISTÓRIE SEMINÁROV VÝPOČTOVÁ ŠTATISTIKA

Pri príležitosti 15. 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 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 a
15. medzinárodný seminár Výpočtová štatistika prebieha v dňoch 7. - 8. 12. 2006.

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 v novozaloženom vedecko-odbornom časopise SŠDS FORUM STATISTICUM SLOVACUM.

Zameraním seminára je problematika na rozhraní počítačových vied a štatistiky.

Tematické okruhy posledných seminárov sa nemenia:

- praktické použitie systémov štatistických a príbuzných predmetov,
- práca s rozsiahlymi súbormi údajov,
- vyučovanie výpočtovej štatistiky a príbuzných predmetov,
- praktické aplikácie výpočtovej štatistiky,
- iné.

V čase konania seminára Výpočtová štatistika sa uskutočňuje aj ***prehliadka prác mladých štatistikov a demografov***. Táto akcia prebieha od 7. seminára. Na 8. medzinárodnom seminári prezentovalo svoje práce 5 mladých štatistikov a demografov, na 9. medzinárodnom seminári už bolo 20 prác mladých štatistikov a demografov, na 10. bolo prihlásených 26 prác a na 11. bolo prihlásených 18 prác, ale vzhľadom na niekoľko prác vypracovaných skupinou autorov bol počet účastníkov vyšší než predošlý rok. Na 12. seminári bolo prihlásených 19 prác, pričom niektoré sú prácou viacerých autorov. Na ďalšom 13. seminári bolo prihlásených 9 prác od 12 autorov. V rámci 14. seminára bolo prihlásených 15 sólových prác mladých autorov. V aktuálnom ročníku, v rámci 15. seminára bolo prihlásených 20 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 resp. Jan.Luha@statistics.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

Probability Models – Theory and Applications

Jitka Bartošová

Abstract. Probability models of income distribution provide for evaluation of the living standard of population of a country at whole as well as for comparison of the living standard of different social classes or regions in the country. For creation of a probability model it is important both, to find a theoretical distribution function that characterizes empirical frequency distribution and to choose suitable methods to calculate parameters of the model. To construct a suitable parametric model of income distribution one can apply the maximization of coincidence of theoretical and empirical frequencies or quintiles. This paper concentrates on methods of probability modeling and their application to creation of a model of household income distribution.

Key words: Income distribution; probability model; validity of the model.

1 Introduction

Examination of income distribution and its comparison from the prospective of various social-economical and time and space views forms basis for evaluation of the living standard of population, the level of social security and social justice in distribution of material values created by the society. Therefore, statistical analyses of population income distribution represent grounds for decision making regarding budgets and social policy.

Knowledge of income distribution enables to assess the living standard of all inhabitants of a country at whole as well as to compare living standards of members of different social groups or different regions. Income also designates the relative living standard of a country population in comparison with other countries. The living standard that is, in broad terms, determined by complex of all material and social living conditions, cannot be quantified in this general notion. Therefore, for the purpose of statistical analyses of the living standard, we focus merely on the measurable elements of the living standard. For correct quantification of the element of the living standard, which is directly dependent on income, it is necessary to describe the level and structure of population income in a complex manner and find suitable probability models of income distribution for individual social groups as well as the population as whole without regard to social groups.

The analysis of income distribution that leads and quantitative description of the living standard is, in general, focused on characterization of the income inequality. In past, such inequality was usual determined by Lorenzo's curve and basic characteristics of income differentiation computed from it and derived by Pareto, Bresciani, Gini, Pietro, etc. In present literature, there are three basic approaches to quantification of the income inequality:

- First, traditional approach is directed towards measurement of the relative level of the income inequality within the frame of a population of one economic complex. This method was published by Pareto in 1896 a built up by Gini in 1955.
- Another approach to measurement of the income inequality was developed by Gastwirth in 1975 (see [7]) and further worked on by, for example, Dagum (see [5]). These measurements show differences among several populations that differ either from social-economical or geographical prospective.
- Last decomposition approach focuses on expressing the contribution of individual subpopulations to the income inequality of the whole heterogenic population. Each subpopulation is specified by a value of a certain social-economical feature and the level of its impact may be determined with respect to the total inequality or the sum of

the whole population income. This approach, which was used first by Thiel in 1967, was further developed by Dagum in his models.

2 Probabilistic modeling

To measure level of standard and level of social security of the population within the frame of one state as well as to compare levels of standard in different countries in terms of income inequalities we often use simple probabilistic models. Complicated empiric income distribution of the population is then replaced by simple approximations – models. In their construction we may consider two ways different on principle – using distribution function (or its derivation) or quintile function (or its derivation). This article focuses on a construction of a probabilistic model of households income distribution using distribution function. Constructions using quintile function can be found in e.g. Pacáková, Sipková a Sodomová [9], Sipková [10].

2.1 Logarithm-normal model

Economic quantities, such as income, wages, turn-out, profits, expenses etc., are bounded below by nonnegative values. In past the three-parameter logarithm-normal distribution with parameters μ, σ^2 and γ, where γ is the theoretical minimum, represented a good approximation of income distribution. Therefore, the probability distribution function of the chosen model is determined by the following relation

$$f(x;\mu,\sigma^2,\gamma) = \begin{cases} \left(2\pi\sigma^2\right)^{-\frac{1}{2}}(x-\gamma)^{-1}e^{-\frac{(\ln(x-\gamma)-\mu)^2}{2\sigma^2}} & \text{if } x\in(\gamma,\infty), \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\mu \in R$, $\sigma^2 \in R^+$ and $\gamma \in R$.

Initially, development of the logarithm-normal distribution was studied by Aitchison and Brown in 1957 [1]. They were interested mainly in its application in astronomy, biology, sociology, economics or simulation of physical processes. These authors also formulated special reasons for using the logarithm-normal distribution in income distribution models Applications of the logarithm-normal distribution in economics was studied mainly by Gibrat [6]. The biggest "competitor" of the logarithm-normal distribution in income modeling is the Pareto's distribution. The logarithm-normal curve corresponds to empirical income distribution in a large central area, while in extremes it significantly diverges. On the contrary, Pareto's curve is a suitable model of income distribution in extreme values (see [8]). It means that it is suitable to use the logarithm-normal model for depicting income distribution of majority of households of the central part of range and the Pareto's distribution for depicting extremes.

2.2 Selection of the model of income distribution after 1990

After bad success with Pareto's distribution and its modifications (e.g. Champernown's model) economists' opinion settled down saying that the logarithm-normal model is satisfactory accurate approximation of income distribution. Logarithm-normal distribution was found to be suitable for models of households' income distribution both in particular social classes and as a whole without regarding the classes. However if we consider more detailed distribution of income sets, this model does not suit any more.

Validity of this probabilistic model after 1990 was tested on net income distribution of the households in 1996. Logarithm-normal distribution was one of the most used models of income distribution in the Czechoslovakia before 1990 and therefore we could expect that would hold on. But in regard of changes in incomes of the population in the Czech Republic as a result of transformation from planned economy to market economy, this conjecture must be verified.

3 Modeling of income distribution in the Czech Republic in 1996

3.1 Data Set

In 1996 sample survey Mikrocensus (including households' income data) was made in about 1% of households, which was about 28 000 flats that time. While making statistical analysis of income distribution in 1996, I came out from complete data set that I got from Czech Statistical Office. Complete non-aggregated sample set enabled me to gain quality estimates of parameters of the models of the distribution. For purposes of research following data were chosen:¨

- Social class of the head of household
- Number of members in household
- Net income of household (CZK per year)

Social structure of sample set of households' incomes from 1996 is in Table 1.

Table 1.
Structure of Set of Households' Incomes in 1996.

Social class	Size	%
Worker	8856	31,5
Self-employed	1748	6,2
Employees	6915	24,6
Self-employed farmers	131	0,5
Farmers–member of cooperative	195	0,7
Retired with EA members	1156	4,1
Retired without EA members	8651	30,7
Unemployed	260	0,9
Others	236	0,8

3.2 Construction of the model

To construct the probabilistic model of households' income distribution in 1996 we use the method of maximum likelihood to estimate of parameters μ a σ^2 of logarithm-normal models combined with other types of estimates of theoretical minimum (of parameter γ) (see [3], [4]).

The level of conformity of the empirical household income distribution with the logarithm-normal model was quantified by the likelihood ratio, i.e. the following statistic

$$LR(\mu,\sigma^2,\gamma|n) = 2[\ell(\vec{p}|n) - \ell(\vec{\pi}(\mu,\sigma^2,\gamma)|n)], \tag{2}$$

where $\vec{p}$ is the vector of income empirical probability, $\vec{\pi}(\mu,\sigma^2,\gamma)$ is the vector of probabilities of occupation of particular classes and $\ell(\vec{p}|n)$, $\ell(\vec{\pi}(\mu,\sigma^2,\gamma)|n)$ are corresponding logarithm-normal likelihood functions (see [2]). Statistic LR was chosen because the method of maximum likelihood and its modifications were used to estimate parameters of the model.

The results are also influenced by the number of classes, in which data are classified. The problem of optimal number of classes m has been a subject matter of numerous papers. In this case we choose

$$m = 15 \cdot \sqrt[5]{\left(\frac{n}{100}\right)^2}\,, \tag{3}$$

which is suitable for a large sample, i.e. for $n > 80$ (see [11]).

3.3 Quality of the model

We concentrate on construction of a theoretical distribution model for two statistical features – income per household and income per head. These two features may behave in different manner, since households with larger number of members usually have lower income per head. For creation of the logarithm-normal model, we use net year incomes of households in current prices. Likelihood ratio comparison (statistic LR) of the models, which are based on combination of maximum likelihood estimate of μ and σ^2 and estimate of γ by means of a different estimation method, are contained in Tables 2 and 3. Besides likelihood ratio, there are also values of corresponding 95% quintiles $\chi^2(m-4)$, where m is defined by (3).

Table 2.
Comparison of Conformity of Empirical Distribution with Logarithm-Normal Models.
(Income per Households)

Social class	**Likelihood ratio** LR				
	$\hat{\gamma} = 0$	$\hat{\gamma} = x_{\min}$	$\hat{\gamma}\_Cohen$	$\hat{\gamma}\_LR$	$\chi^2_{0,95}(m-4)$
Worker	475,9	907,3	300,8	232,8	108,6
Self-employed	77,4	108,6	81,3	77,1	59,3
Employees	145,2	264,6	114,1	109,4	99,6
Self-employed farmers	31,3	34,3	32,0	31,0	22,4
Farmers–member of cooperative	14,8	28,8	15,0	14,8	26,3
Retired with EA members	32,6	38,8	29,6	21,4	51,0
Retired without EA members	**3874,5**	**3690,6**	**4027,7**	**3698,5**	**108,0**
Unemployed	29,3	51,7	26,7	23,0	28,9
Others	26,9	21,8	33,3	25,3	27,6
All	**3239,8**	**3233,1**	**3311,7**	**3224,4**	**167,5**

As we can see in Tables 2 and 3, in almost all social classes $LR \approx \chi^2_{0,95}(m-4)$ applies, and thus our logarithm-normal curves represent a good approximation of empirical distribution of household income in 1996. Only in case of retired without economically active members and all households (irrespective of social classes), the logarithm-normal model is totally unsuitable. In both these cases, the above statistic LR was more then ten times higher than corresponding quintile. Moreover, graphs of kernel estimate of density of probability of distribution, from which samples of households` incomes originate, are two-peaked in both cases and it clearly indicates commixture of two one-peaked curves. This means that it is not possible to find a one-peaked parametric model that would well approximate this empirical distribution. On the other hand, both corresponding income distributions per head in the classes of retired without economically active members and all households (irrespective of

social classes) are one-peaked. Consequently, we should search for other more convenient parametric models.

Table 3.
Comparison of Conformity of Empirical Distribution with Logarithm-Normal Models.
(Income per Head)

Social class	**Likelihood ratio** LR				$\chi^2_{0,95}(m-4)$
	$\hat{\gamma}=0$	$\hat{\gamma}=x_{\min}$	$\hat{\gamma}\_Cohen$	$\hat{\gamma}\_LR$	
Worker	170,6	279,7	174,7	163,1	108,6
Self-employed	94,6	55,3	97,2	55,8	59,3
Employees	167,4	189,1	104,5	104,2	99,6
Self-employed farmers	29,0	23,0	23,8	22,7	22,4
Farmers–member of cooperative	11,5	41,8	12,4	11,0	26,3
Retired with EA members	86,0	120,3	101,7	84,8	51,0
Retired without EA members	**1520,6**	**1672,2**	**2675,8**	**1519,3**	**108,0**
Unemployed	17,8	25,1	18,0	17,8	28,9
Others	76,9	38,3	58,4	44,4	27,6
All	**2702,5**	**2533,3**	**3253,9**	**2534,2**	**167,5**

Notation

$\hat{\gamma}=0$	LR for estimate of $\hat{\gamma}$ with null (two-parametric model)
$\hat{\gamma}=x_{\min}$	LR for estimate of $\hat{\gamma}$ with sample minimum $x_{\min}$
$\hat{\gamma}\_Cohen$	LR for estimate of $\hat{\gamma}$ with Cohen's method, where the sample minimum is used as $\frac{100}{n+1}\%$ quintile of the theoretical distribution
$\hat{\gamma}\_LR$	LR for estimate of $\hat{\gamma}$ with method of likelihood ratio minimization
bald face	sharp disagreement of empiric distribution with the model $LR >> \chi^2_{0,95}(m-4)$

4 Conclusions

The most important result of the analysis is that the formerly used probabilistic model of income distribution (logarithm-normal model), which is suitable mostly to model empiric distribution can be considered as suitable in 1996 as well.

The results of the analysis show that the logarithm-normal distribution represents a suitable model for the most of social classes. However, this model is totally unsuitable for income distribution in the class of households of retired without economically active members and the class of all households. The comparison of quality of logarithm-normal models, in which different methods were used to estimate theoretical minimum, shows that the achieved results are similar.

Surprising results brings comparison of agreement of logarithm-normal models where maximum likelihood estimates of parameters μ and σ^2 are combined with different types of estimates of theoretical minimum γ (see Tables 2 and 3).). It turns out that in all cases we achieved similar agreements of empiric distribution with the model, although the estimates of theoretical minimum are roughly distinct. The source of this phenomenon is existence of negative correlation between parameters μ, σ^2 and γ, that compensates the influence of estimate of theoretical minimum.

References:

[1] Aitchison, J., Brown, J.A.C.. *The Lognormal Distribution with Special Reference to Its Uses in Economics*. Cabridge University Press, Cambridge, 1957.

[2] Anděl, J.. *Základy matematické statistiky*. Univerzita Karlova v Praze, Praha, 2002.

[3] Bartošová, J. Odhady parametrů lognormálního modelu rozdělení příjmů domácností. Acta *Universitatis Bohemiae Meridionales, **8(1)***, Jihočeská univerzita, České Budějovice, 2005, pp. 39-44.

[4] Bartošová, J. Maximal Likelihood Estimates of Parameters of Model of Household Income Distribution in the Czech Republic. *Forum Statisticum Slovacum*, **3/2005**, SŠDS, Bratislava, 2005, pp. 150-155.

[5] Barry, C.A., Balakrishnan, N., Nagaraja, H. N.: *A First Course in Order Statistics*, John Wiley & Sons, New York, 1992.

[6] Gibrat, R.: *Les inegalités économiques*, Libraire du Recueil Sirey, Paris, 1931.

[7] Champernowne, D.G.: A Model of Income Distribution, *Economic Journal **63***, June, 1953, pp. 318-351.

[8] Johnson, J.N.L., Kotz, S., Balakrishnan, N.: *Continuous Univariate Distributions*, vol. **1**, 2nd edition, J. Wiley, New York, 1994.

[9] Pacáková, V., Sipková, Ľ., Sodomová, E. Štatistické modelovanie príjmov domácností v Slovenskej republike. *Ekonomický časopis, **4(53)***, 2005, pp. 427 – 439.

[10] Sipková, Ľ. Modelovanie príjmov domácností zovšeobecneným lambda rozdelením. *Ekonomika a informatika*, ***1(3)***, 2005, pp. 90 – 164.

[11] Williams, D.: *Wieghing the Odds, A Course in Probability and Statistics*. Cambridge Univ. Press, Cambridge, 2001.

Jitka Bartošová
University of Economics Prague
Faculty of Management,
Department of Information Management
Jarošovská 1117/II
377 01 Jindřichův Hradec
Czech Republic
bartosov@fm.vse.cz

ODHADY BUDOUCÍ DÉLKY ŽIVOTA V ČR V KOMBINACI S VÝSLEDKY SOCIOŠETŘENÍ

RNDr. Jaromír Běláček, CSc.[1], Mgr. Kryštof Zeman[2], RNDr. Petr Hrala[3]
[1]Ústav biofyziky a informatiky 1.lékařské fakulty UK Praha, [2]Český statistický úřad,
[3]Opinion Window Research International Praha

Abstract: This contribution tries to communicate the possibilities, which give or can give in demography not frequently used data sources useful to estimation of the future life expectancies by age, sex and education. The mortality tables surveyed in this text have been calculated from Population census and from numbers od dead persons in Czech Republic during 2001 at regional detail of 14 Czech provinces by sex, 5-year age groups and by four stages of graduated ones. These estimators are subsequently compared with the results of quantitative sociological survey with about 1250 adults, whese the respondents has been asked about their own imagination of his/her own complete life expectancy (this survey was realised by Opinion Window Research International agency in 2006).

The results used the demographic data from 2001 confirm that the more education of people corresponds with the higher chance forward longer life expectancy of that ones. But the results of omnibus survey in 2006 show that the own imagination of adults about their future lifetime are different against the real or the more probable values in less graduated social groups. In similar way, the subpopulation of men assume their own life expectancy generally higher to life expectancy lever comparable with that one of women. But this statement is in a disagreement with real lower life expectancies for men towards women (this conclusion holds as the known fact for the whole European countries from the long term point of view).

Key words: *Mortality tables, life expectancy by age, stages of finished education, omnibus survey, demographical prognoses, public health.*

1. Cíle a záměry příspěvku

Střední délka života anebo také naděje dožití (při narození i ve věku x) je v současnosti snad jednou z nejčastěji uváděných populačních charakteristik. V posledním čtvrtstoletí byla metodicky rozpracována do celé řady speciálních demografických a geodemografických aplikací (tabulky sňatečnosti, rozvodovosti, vícerozměrné demografické modely), v posledních 10-15ti letech se hojně rozvíjejí modifikace, které zohledňují např. sníženou fyzickou schopnost (disabilitu) a skutečný zdravotní stav u lidských populací (viz např. koncept „zatížení obyvatelstva nemocí“ tzv. „burdon of disease“ - viz Murray-Lopez/1996/). Posledně zmíněné koncepty, které jsou systematicky cíleny na odhady charakteristik jako je „střední délka života prožitá ve zdraví“, „potenciální roky ztraceného života“ atd., jsou ve svých složitěji propracovaných i jednodušších verzích obvykle nemyslitelné bez nezávislých epidemiologických či sociologických průzkumů u populace. Na základě těchto externích průzkumů se pak kvantifikují parametry pro modifikace běžných odhadů střední délky života.

Tento víceméně referenční text je jedním z příspěvků připravených v rámci dílčího výzkumného tématu, které jsme si stanovili pro rok 2006 v prvním roce řešení tříletého grantového projektu u GA ČR (viz Běláček a kol./2006/). Cílem zmíněného tématu (pod názvem „Vývoj úmrtnosti v ČR z pohledu regionálních a socio-ekonomických struktur obyvatelstva“) je kvalifikované stanovení budoucích hladin střední délky života v České republice resp. v krajích ČR, které by byly použitelné do demografické projekce s horizontem roku 2025. Za tímto účelem byl proveden výpočet zkrácených úmrtnostních tabulek ve všech čtrnácti krajích ČR a podle vzdělání (viz také Zeman/2006/). Tabulky byly vypočteny

z průřezových údajů Sčítání lidu, domů a bytů v ČR a počtů zemřelých osob v ČR v r. 2001 vytříděných podle pohlaví, pětiletých věkových skupin a čtyř kategorií dokončeného vzdělání.

Paralelně s řešeným tématem o úmrtnosti jsme připravili a provedli sociovýzkum zaměřený (v rámci dalšího dílčího výzkumného tématu) na ohodnocení úrovně nemocnosti a disability (nesoběstačnosti) u obyvatelstva ČR. Při této příležitosti byla reprezentativnímu vzorku 1250 dospělých osob (vybraných reprezentativně v krajích ČR podle pohlaví, věku a velikosti místa bydliště – průzkum provedla agentura Opinion Window Research International v r.2006 – viz také na www.morbidity.wz.cz) položeny následující dvě otázky:

H11a. „Víte, jaká je aktuální naděje dožití (často se říká také střední délka života) pro právě narozeného chlapce nebo dívku v České republice?“ (Odpovědi ANO – NE);

H11b. „Zajímalo by mne ale, jakého přesného nebo alespoň přibližného věku předpokládáte, že se dožijete (nebo byste se chtěl/a dožít) Vy sám/sama? (UVEĎTE OČEKÁVANÝ VĚK DOŽITÍ V LETECH).“

Průměrné hodnoty *respondenty očekávaného věku dožití* byly pro účely tohoto příspěvku vytříděny podle pohlaví, desetiletých věkových skupin a skupin vzdělání a porovnány s výsledky získanými na úrovni úmrtnostních tabulek podle vzdělání (viz výše).

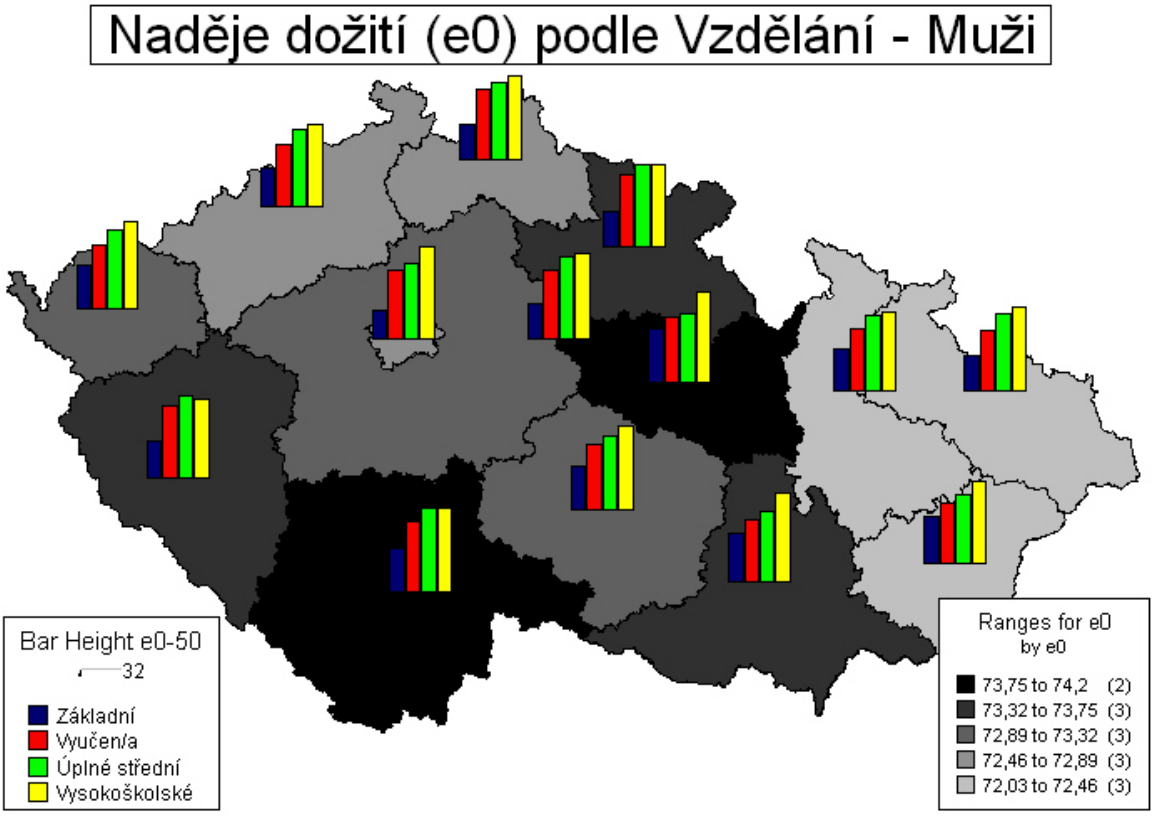
2. Prezentace a porovnání výsledků

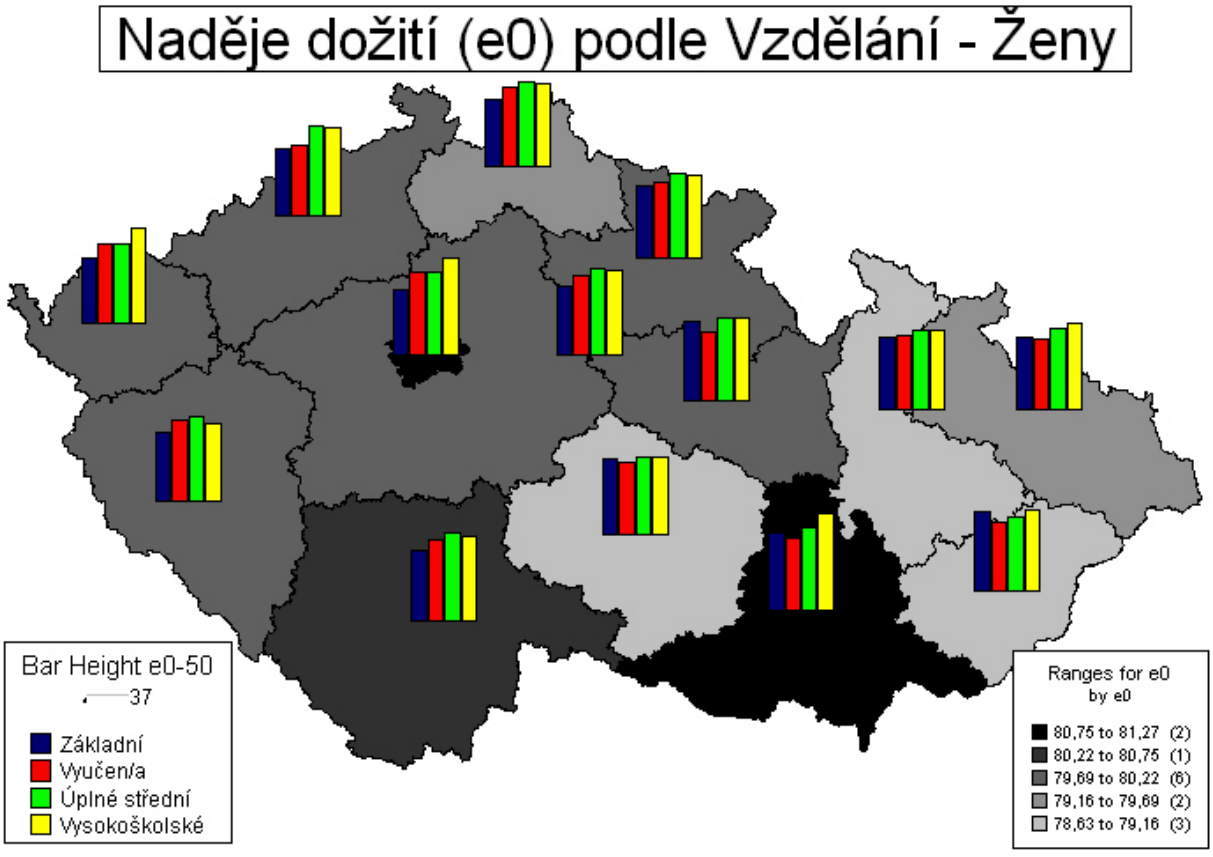
Zkrácené úmrtnostní tabulky byly vypočteny v dokončeném věku „0“, ve věkové skupině „1-4 let“, a dále v pětiletkách s poslední skupinou „85+“, ve čtyřech kategoriích vzdělání a podle pohlaví. Pro naději dožití při narození jsou výsledky shrnuty v Tabulce č.1:

Tabulka 1. Naděje dožití při narození (e0) v ČR a v krajích ČR podle skupin vzdělání

Muži:	Základní	Vyučen/a	Úplné střední	VŠ	Celkem e0	ČR 2001 (dle ČSÚ)
ČR	63,79	72,54	76,09	79,64	71,87	72,14
Praha	59,77	73,30	75,86	81,31	73,52	73,46
Středočeský	62,11	73,73	77,91	79,37	71,74	71,58
Jihočeský	65,17	74,03	78,65	78,92	72,35	72,31
Plzeňský	62,95	74,86	78,51	77,18	72,29	72,10
Karlovarský	64,81	71,81	76,73	79,50	71,03	70,60
Ústecký	63,55	71,60	76,68	78,29	69,94	70,09
Liberecký	62,47	74,40	76,18	78,44	71,33	71,22
Králové-hradecký	62,48	75,10	78,25	78,62	72,79	72,54
Pardubický	68,58	72,93	73,72	80,99	72,79	72,64
Vysočina	65,48	72,98	75,51	79,01	72,70	72,63
Jihomoravský	67,19	71,71	74,16	80,49	72,36	72,31
Olomoucký	64,69	71,37	75,85	77,40	71,52	71,41
Zlínský	66,35	70,94	73,56	77,87	71,68	71,37
Moravsko- slezský	62,39	70,70	76,15	78,92	70,42	70,27
Ženy:						
ČR	77,04	78,60	81,35	83,19	78,51	78,45
Praha	74,53	81,11	81,43	86,14	79,22	78,95
Středočeský	76,20	79,94	82,30	81,49	78,20	77,95
Jihočeský	76,95	80,54	83,40	81,86	78,65	78,36
Plzeňský	76,03	80,64	82,43	79,70	78,49	78,03
Karlovarský	74,85	79,84	79,50	86,03	77,23	76,90
Ústecký	75,18	77,08	84,14	83,51	76,67	76,50
Liberecký	75,38	80,17	82,09	81,03	77,96	77,86
Králové-hradecký	77,69	79,12	82,16	81,60	79,21	78,96
Pardubický	80,23	76,52	81,67	81,62	78,91	78,59
Vysočina	78,83	77,60	79,77	79,59	79,08	78,64
Jihomoravský	79,30	77,79	81,40	86,55	79,15	78,96
Olomoucký	77,64	78,32	80,09	80,00	78,73	78,59
Zlínský	79,86	75,97	78,13	80,58	79,18	78,72
Moravsko- slezský	76,98	76,68	80,77	82,40	78,00	77,82

Regionální diference jsou přehledně vyjádřeny prostřednictvím Kartodiagramů č.1-2:





Pro očekávaný věk dožití, který vyjádřili dospělí respondenti svými odpověďmi na otázku č. *H11b*, ovšem nemůžeme získat odhad pro e_0 t.j. pro naději dožití *při narození*, nýbrž pouze pro e_x (t.j. pro naději dožijí *v jejich přesném věku x*) anebo pro *odhad jejich celkové délky života* (ozn. c_x), což je vzhledem k formulaci *H11b* evidentně metodicky korektnější (platí rovnost: $c_x = e_x + x$). Pro porovnatelnost výsledků s úmrtnostními tabulkami, kde jsou hodnoty e_x resp. c_x kalkulovány *přesně* pro levé krajní body tabulkových intervalů (vyberme třeba za krajní body $x_0, x_{10}, x_{20}, \ldots, x_{60}, x_{70+}$), musíme individuální odhady c_x ze sociošetření vyčíslit jako průměrné hodnoty získané v rámci vhodně zvolených desetiletých intervalů (zvolíme-li za tyto věkové intervaly „15-24“, ... , „65-74“, „75+“, mohly by být výsledky metodicky srovnatelné).

Pro lepší srozumitelnost si ukažme výsledky pouze na regionální úrovni ČR (tabulka níže).

Tabulka 2. Porovnání odhadů celkové délky života podle věku v ČR (ÚT vs. OPW RI)

x	Celková délka života (c(x))v r. 2001					"x"	Odhad očekávané délky života v r. 2006				
Muži:	Základní	Vyučen	Úplné střední	VŠ	Celkem c_x		Základní	Vyučen	Úplné střední	VŠ	Celkem c_x
0	63,79	72,54	76,09	79,64	71,87						
10	64,22	73,03	76,59	80,17	72,35						
20	64,42	73,27	76,85	80,44	72,59	15-24	81,18	81,95	79,08	80,00	80,68
30	65,38	73,84	77,16	80,84	73,07	25-34	73,52	78,20	82,08	82,99	80,30
40	66,50	74,50	77,55	81,09	73,63	35-44	72,49	76,97	77,60	83,85	77,69
50	68,75	75,79	78,42	81,63	74,84	45-54	80,76	78,08	78,15	77,41	78,08
60	72,98	78,27	80,20	82,78	77,19	55-64	79,62	79,55	82,34	85,14	81,15
70	78,13	82,16	83,22	84,90	80,91	65-74	82,26	81,98	83,84	82,77	82,78
80+	84,26	87,82	87,70	88,71	86,11	75+	94,06	88,39	91,56	85,36	89,08
Ženy:											
0	77,04	78,60	81,35	83,19	78,51						
10	77,44	79,00	81,77	83,62	78,91						
20	77,58	79,14	81,92	83,78	79,05	15-24	81,90	79,15	81,33	77,65	81,10
30	77,98	79,39	82,04	83,91	79,23	25-34	85,58	74,45	77,62	79,13	77,19
40	78,54	79,70	82,25	84,07	79,50	35-44	73,48	80,48	79,34	77,93	79,31
50	79,38	80,44	82,77	84,40	80,13	45-54	81,55	77,94	79,28	78,13	79,01
60	80,76	81,96	83,83	85,13	81,36	55-64	83,37	79,84	83,27	77,87	82,09
70	82,81	84,46	85,83	86,45	83,44	65-74	82,89	82,67	81,48	83,85	82,47
80	86,70	88,43	89,93	89,01	87,26	75+	87,10	86,86	85,34	84,00	86,69

Na první pohled je zřejmé, že s rostoucím věkem (x) se zvyšuje i celková délka života (c_x). Tato vlastnost vyplývá přímo z konstrukce úmrtnostních tabulek, takže platí rovněž pro všechny kategorie vzdělání. Diference hodnot mezi nejvyšší a nejnižší vzdělanostní skupinou markantní zvláště u mužů se počínaje věkem 60 a více let poměrně rychle snižuje.

Vzhledem ku známému dlouhodobému nárůstu středních délek života po celé Evropě bychom očekávali, že odhady budoucí délky života učiněné v r.2005 budou oproti odhadům z úmrtnostních tabulek znatelně vyšší. Toto se však ukázalo jako pravdivé pouze u populace mužů. U žen (s výjimkou nejmladší věkové skupiny „15-24“, tomu však bylo (alespoň co se týče námi již provedeného sociovýzkumu), naopak. Je zvláštní, že dospělé osoby dotazované na jejich vlastní očekávanou délku života v podstatě „v průměru“ nereflektují ani dobře známou skutečnost, že ženy v ČR se dožívají stále cca o 5 let vyššího věk nežli muži. Přestože se tyto tzv. „nůžky mezi oběma pohlavími“ z dlouhodobého hlediska zavírají, nedá se toto považovat za vysvětlení v podstatě rovnostářských odpovědí mužů a žen na uvedenou otázku (H11b). Do interpretace těchto záležitostí nevnesla světlo ani respondentem deklarovaná znalost aktuální naděje dožití při narození (otázku H11a zodpovědělo kladně více než 30% respondentů). U mužů se naopak zdá, jakoby kladná odpověď na H11a vedla k ještě

většímu rovnostářství z hlediska rozlišování odhadu budoucí délky života vzhledem k jejich věku.

3. Závěr

Při zvoleném rozsahu výběru respondentů a provedeném způsobu vyšetřování otázek vztažených k budoucí střední délce života v České republice a v krajích ČR nelze vyvozovat o reálné budoucí délce života v populaci žádné statisticky významné závěry. Spíše lze říci, že populace v ČR trpí možná až příliš vysokou neinformovaností v oblasti relevantních demografických informací. Problematika úmrtnosti by musela být v dalším sociovýzkumu ošetřena nepochybně mnohem pečlivějším způsobem.

Naopak údaje odhadnuté na základě speciálních demografických třídění podle pohlaví, věku, regionu a vzdělání se jeví jako velmi perspektivní nástroj pro projektování budoucí dynamiky procesů vztažených k úmrtnosti (jmenovitě máme teď na mysli projekce nemocnosti nebo disability ve vybraných regionech nebo větších sociálních skupinách obyvatelstva). Každopádně je potřebné hledat a snažit se využívat i jiné alternativy odhadů budoucí střední délky života, ať jde o již používané expertní hodnocení či zkušenosti z vývoje v sousedních, zejména ekonomicky vyspělejších zemích, o odhady postavené na bázi generační úmrtnosti (viz příspěvek T.Fialy na jiném místě tohoto Sborníku) anebo alternativní možnosti formálního matematického modelování (např. na bázi tzv. Brassova relačního modelu).

Tento text byl vytvořen jako příspěvek řešení grantovému projektu č.403/06/1836 „Operacionalizace projekčního modelu pro regionální projekce zdravotního stavu obyvatelstva v České republice“ podporovaného z prostředků Grantové agentury ČR – viz www.morbidity.wz.cz .

4. Literatura

Běláček, J.,: *Je třeba více pozornosti důchodcům anebo teenagerům?* In: Vomáčková H.(ed): „Ekonomické aspekty vzdělanosti v regionálním kontextu III“. Sborník příspěvků z vědeckého semináře s mezinárodní účastí 21.října 2005. Ústí nad Labem: Fakulta sociálně-ekonomická Univerzity J.E.Purkyně (*FSE ÚJEP*), ISBN 80-7044-727-3, 2005c), str. 30-47;

Běláček, J.-Fiala, T.-Geryk, E.-Hrala, P.-Kokavec, P.: Analýza zdravotního stavu obyvatelstva v ČR – východiska pro řešení grantového projektu v r.2006. Sborník příspěvků „Medsoft 2006“. Praha: Zeithamlová Milena – Agentura Action M, ISBN 80-86742-12-1, 2006, str.23-30;

Murray, Ch.J.L-Lopez, Q.D.(eds.): *The Global Burden of Disease*. Geneva: World Health Organization (WHO), Harward School of Public health, 1996;

Zeman, K.: Úmrtnostní tabulky podle nejvyššího ukončeného vzdělání – Česká republika. Demografie, 2006, v tisku;

Adresa autora textu:
Ústav biofyziky a informatiky 1.LF UK Praha
Jaromír Běláček, RNDr., CSc.
Salmovská 1
120 00 Praha 2
jaromir.belacek@lf1.cuni.cz

Analýza ekonomickej efektívnosti v súbore leasingových spoločností v roku 2005 pomocou frekvenčných tabuliek

Zuzana Berčačinová[1]

Abstract: The paper consists the analysis of economical efficiency of leasing firms and constructions their table of frequency.

Keywords: profitability, productivity of labor, profitableness of equity, table of frequency, Excel, leasing firms.

K dispozícii sme mali súbor údajov o výnosoch spolu, výsledku hospodárenia, pridanej hodnote, osobných nákladoch, vlastnom imaní za 68 leasingových firiem v roku 2005. Úlohou bolo vytvoriť frekvenčné tabuľky z ukazovateľov ekonomickej efektívnosti, konkrétne

- ziskovosti
- finančnej produktivity práce
- rentability vlastného imania.

Ziskovosť predstavuje podiel zisku k výnosom, zvyčajne sa udáva v halieroch zisku na jednu korunu výnosov. Predstavuje prakticky najčastejšie používaný ekonomický ukazovateľ efektívnosti spoločnosti z pohľadu súkromného vlastníka.

Finančná produktivita práce sa počíta ako pomer pridanej hodnoty k osobným nákladom. Vyjadruje efektívnosť spotrebovanej ľudskej práce vo firme. Vyjadruje sa v korunách pridanej hodnoty na korunu osobných nákladov.

Rentabilita vlastného imania sa vypočíta ako podiel zisku a vlastného imania. Vyjadruje sa v halieroch zisku na korunu vlastného imania. V teoretickom aspekte je to najdôležitejší ukazovateľ efektívnosti z pohľadu súkromného vlastníka. Praktický život okolo vykazovania zisku a niektoré problémy ohľadne špecifikovania konkrétnej číselnej hodnoty vlastného imania, však tento teoretický význam výrazne znižujú.

1. Výpočet frekvenčnej tabuľky

K vlastným výpočtom frekvenčnej tabuľky sme použili systém Excel. Prvým krokom bol výpočet z východiskových hodnôt ukazovateľov hodnoty troch relatívnych ukazovateľov efektívnosti, podľa jednoduchých vzorcov.

Druhým krokom je teoretické určenie hraníc jednotlivých triednych intervalov. V nasledujúcich tabuľkách 1,2,3 sú uvedené horné hranice pre príslušné triedy.
K určeniu triednych početností v systéme Excel sa môže použiť funkcia FREQUENCY. Treba špecifikovať hodnoty, ktoré budeme triediť, špecifikácia sa realizuje v políčku DATA_ ARRAY a ďalej treba špecifikovať horné hranice jednotlivých tried rozdelení početností príslušného rozdelenia v políčku BINS_ARRAY. V tabuľke máme vysvietených o jedno políčko viac ako je ich špecifikovaných v BINS_ARRAY a súčastným stlačením klávesov CTRL + SHIFT + ENTER systém zrealizuje vytriedenie údajov.

[1] Ing. Zuzana Berčačinová, Patria, a.s.

Vlastná špecifikácia excelovskej tabuľky obsahujúcej rad rozdelenia početností (frekvenčnej tabuľky) obsahuje okrem stĺpca z hornými hranicami tried a stĺpca absolútnymi triednymi početnosťami ešte tri stĺpce, konkrétne stĺpec relatívnych početností, stĺpec kumulatívnych absolútnych početností a stĺpec relatívnych kumulatívnych početností.
V poslednom riadu tabuľky sú údaje o pre súbor spolu. Ich výpočtu sa používajú jednoduché vzorce resp. funkcia SUM.

2. Frekvenčné tabuľky ukazovateľov ekonomickej efektívnosti

Vypočítané frekvenčné tabuľky sú uvedené v tejto časti príspevku. V tab. č.1 je rad rozdelenia početností pre ziskovosť, v tab. č.2 je uvedený rad rozdelenia pre finančnú produktivitu práce a v tab. č. 3 je uvedený rad rozdelenia pre rentabilitu vlastného imania.

Tabuľka č. 1 – Frekvenčná tabuľka pre ziskovosť

triedy	n	f	kn	kf
-20	14	20,59%	14	20,59%
-15	1	1,47%	15	22,06%
-10	1	1,47%	16	23,53%
-5	2	2,94%	18	26,47%
-1E-12	3	4,41%	21	30,88%
0	9	13,24%	30	44,12%
5	15	22,06%	45	66,18%
10	8	11,76%	53	77,94%
15	5	7,35%	58	85,30%
viac	10	14,71%	68	100,00%
Spolu	68	100,00%		

Tabuľka č. 2 – Frekvenčná tabuľke pre finančnú produktivitu práce

triedy	n	f	kn	kf
-1	5	7,35%	5	7,35%
-0,0000001	4	5,88%	9	13,24%
0	31	45,59%	40	58,82%
1	0	0,00%	40	58,82%
2	1	1,47%	41	60,29%
3	4	5,88%	45	66,18%
5	3	4,41%	48	70,59%
8	4	5,88%	52	76,47%
10	2	2,94%	54	79,41%
viac	14	20,59%	68	100,00%
Spolu	68	100,00%		

Tabuľka č. 3 – Frekvenčná tabuľka pre rentabilitu vlastného imania

triedy	n	f	kn	kf
VI<0	15	22,06%	15	22,06%
VI=0	0	0,00%	15	22,06%
-10	7	10,29%	22	32,35%
-5	3	4,41%	25	36,76%
-0,0000001	6	8,82%	31	45,59%
0	3	4,41%	34	50,00%
2	2	2,94%	36	52,94%
5	3	4,41%	39	57,35%
10	3	4,41%	42	61,76%
viac	26	38,24%	68	100,00%
Spolu	68	100%		

3. Analýza ekonomickej efektívnosti

Pri ziskovosti môžeme konštatovať:

- 21 spoločností resp. 31% spoločností má zápornú ziskovosť, z toho jedna pätina menej ako –0,20 halierov výsledku hospodárenia na korunu výnosov,
- 9 spoločností malo vykázaný nulový výsledok hospodárenia, čiže aj nulovú ziskovosť,
- 15 spoločností má ziskovosť menej ako 5 halierov zisku na korunu výnosov.

Pri finančnej produktivite práce môžeme konštatovať:

- 31 spoločností resp. 46% spoločností má vykázanú nulu pri pridanej hodnote,
- 9 spoločností má vykázanú zápornú hodnotu, čiže zápornú produktivitu práce,
- 28 spoločností pôsobilo s rozumnou finančnou produktivitou práce (pridaná hodnota bola vyššia ako osobné náklady), pričom 14 organizácii vykázalo extrémne vysokú produktivitu práce vyššiu ako 10,– Sk pridanej hodnoty na korunu osobných nákladov.

Pri rentabilite vlastného imania môžeme konštatovať:

- 15 spoločností resp. 22% vykázalo záporné vlastné imanie,
- 16 spoločností vykázalo zápornú rentabilitu vlastného imania,
- 29 spoločností má rentabilitu vlastného imania vyššiu ako 5 halierov zisku na jednu korunu vlastného imania, to je viac ako úroková miera na ročné vklady alebo inflácia za rok 2005.

Literatúra:

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

Chajdiak, J.: Ekonomická analýza stavu a vývoja firmy, Bratislava, STATIS, 2004, ISBN: 80-85659-32-8

Berčačinová, Z.: Kvantitatívna analýza ekonomických výsledkov lízingových spoločností. In: Štatistické metódy vo vedecko-výskumnej práci 2003, Bratislava, SŠDS, 2003, ISBN: 80-88946-32-8, s.67-69

Adresa autora:
zuzanapatria@mail.t-com.sk

Spracovanie rozsiahlych súborov údajov

Martin Blahušiak

Summary:

Saving and manipulating data was very important in the past and is even more important in our modern society. The best way to save and manipulate huge amount of data is a database. Therefore database systems, which support the creation and administration of databases, exist. The paper tells about the basics of databases, database systems, structured query language SQL and functionality of few popular database systems.

Úvod

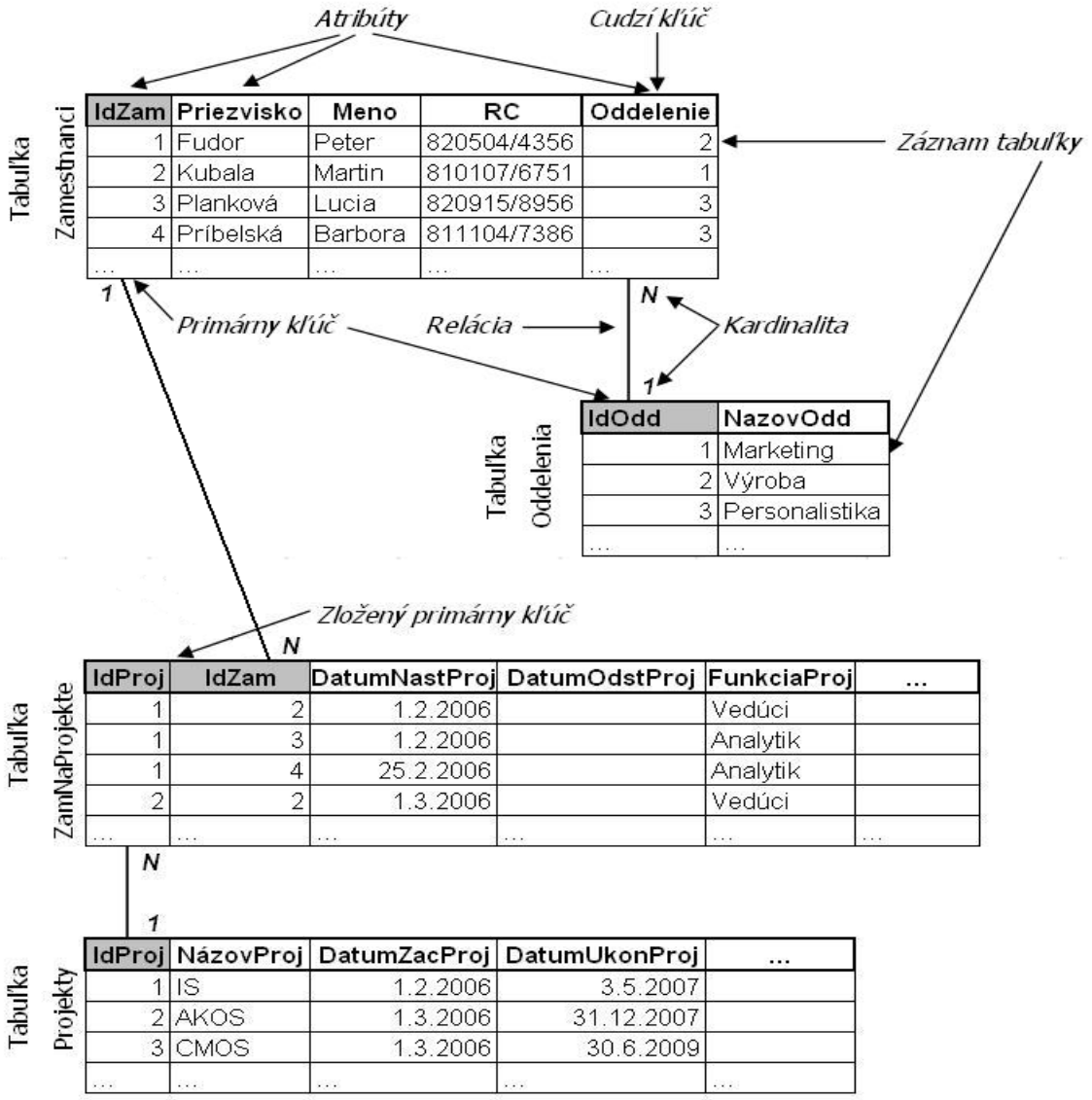
Spracovanie údajov bolo aj je jednou s najdôležitejších vecí v oblasti výpočtovej techniky a samozrejme aj štatistiky. Väčšina sa s ním stretáva skoro na každom kroku a veľakrát si to ani neuvedomuje. Veľa ľudí si robí zoznam kníh, kompaktných diskov, alebo videokaziet, firmy spracúvajú údaje o svojich výrobkoch, odberateľoch, dodávateľoch a podobne. Ak ide o spracovanie rozsiahlych súborov údajov, tak je v súčasnosti najlepšie využiť relačné databázy a databázové systémy.

1. Čo je to databáza

Databáza (DB) predstavuje množinu samostatných tabuliek vytvorených na základe logickej príslušnosti k určitej entite. Entita je objekt reálneho sveta, ktorý je schopný nezávislej existencie a je jednoznačne odlišný od ostatných objektov (napr. zamestnanec, študent). Tabuľka je množina entít, ktoré majú rovnaké atribúty. Napríklad entitou je Zamestnanec a tabuľkou budú Zamestnanci, pretože všetci zamestnanci majú rovnaké atribúty, ako napr. meno, priezvisko, rodné číslo atď. Tabuľky sú navzájom poprepájané reláciami s násobnosťou väzby 1:N. V tabuľke sa konkréte súbory údajov uchovávajú do riadkov a stĺpcov. Stĺpce predstavujú atribúty popisujúce daný objekt a riadky predstavujú jednotlivé záznamy. V každej tabuľke existuje minimálne jeden atribút, ktorého hodnota je jedinečná pre každý záznam tabuľky a nazýva sa primárny kľúč. Na základe primárnych kľúčov sú potom tvorené relácie v databáze. Ak je medzi entitami relácia s násobnosťou väzby 1:N, tak aj tabuľky budú mať medzi sebou reláciu s násobnosťou 1:N a primárny kľúč z prvej tabuľky bude mať priradený v druhej tabuľke atribút nazývaný cudzí kľúč, v ktorom sa budú uchovávať hodnoty primárneho kľúča prvej tabuľky. V prípade násobnosti väzby medzi entitami 1:1 sa atribúty druhej entity stávajú atribútmi prvej entity, takže nakoniec vzniká jedna tabuľka. Ak je medzi entitami násobnosť väzby M:N, potom vzniká medzi pôvodnými tabuľkami vytvorenými z entít nová tabuľka nazývaná relačná. Táto relačná tabuľka prevezme primárne kľúče z obidvoch pôvodných tabuliek a v nej sa tieto kľúče stávajú zloženým primárnym kľúčom. K zloženému primárnemu kľúču treba dodefinovať aspoň jeden ďalší atribút (tzv. nekľúčový), ktorý vyšpecifikuje zložený kľúč a novo vzniknutú tabuľku. Pôvodná relácia s násobnosťou väzby M:N je rozdelená na dve relácie s násobnosťou 1:N, pričom jednotky sú vždy na strane tabuliek vzniknutých z pôvodných entít a N-ká sa stretávajú v novovzniknutej relačnej tabuľke.

Príklad databázy a popisu jej prvkov je znázornený na Obr. 1. Databáza obsahuje štyri tabuľky: Zamestnanci, Oddelenia, Projekty a ZamNaProjekte. Prvé tri tabuľky vznikli

z pôvodných entít Zamestnanec, Oddelenie a Projekt. Obsahujú jednoduchý primárny kľúč, popisujúce atribúty a tabuľka Zamestnanci aj cudzí kľúč. Posledná tabuľka ZamNaProjekte predstavuje relačnú tabuľku, ktorá vznikla z relácie s násobnosťou väzby M:N medzi entitami Zamestnanec a Projekt. Obsahuje zložený primárny kľúč IdProj a IdZam a nekľúčové atribúty DatumNastProj (dátum nástupu zamestnanca na projekt), DatumOdstProj (dátum odstúpenia zamestnanca od projektu), FunkciaProj (funkcia zamestnanca na projekte) a podobne.



Obr. 1 Príklad databázy

2. Databázové systémy a jazyk SQL

Na tvorbu, definovanie a spravovanie databáz, uchovávanie a spracovanie dát v databázach slúžia databázové systémy (DBS). Databázové systémy predstavujú systémy skladajúce sa z databázy a programového vybavenia určeného na správu tejto databázy a v nej uložených dát. Okrem základného programového vybavenia môžu databázové systémy obsahovať aj ďalšie programové vybavenie, ktoré slúži na tvorbu podporných

a doplnkových objektov celého systému. V súčasnosti existuje široká škála databázových systémov. Medzi najznámejšie patria Microsoft SQL Sever, Oracle Database, IBM DB2, Sybase Adaptive Server, MySQL, PostgreSQL a aj Microsoft Access. Jednotlivé databázové systémy existujú vo viacerých verziách a odlišujú sa prostredím, poskytovanou funkcionalitou, náročnosťou práce a podobne, ale všetky vychádzajú a pracujú s jazykom SQL (Structured Query Language).

Structured Query Language, v preklade štruktúrovaný dotazovací jazyk, je všeobecný nástroj na definíciu databázy a manipuláciu, správu a organizovanie dát v databázach. S pomocou príkazov SQL možno definovať štruktúru jednotlivých tabuliek, vytvárať relácie, napĺňať, aktualizovať, vyberať a mazať dáta v stĺpcoch tabuľky a definovať organizáciu a vzťahy medzi položkami dát. Okrem bežných aktualizačných činností a priamych výstupov výsledkov dotazov možno príkazmi SQL riadiť prístup k dátam, teda udeľovať a odoberať prístupové oprávnenia na rôznych úrovniach, a chrániť tak dáta pred náhodným alebo úmyselným zničením, neautorizovaným čítaním alebo manipuláciou. Príkazy SQL zaisťujú aj integritu databáz tým, že sa nepovolí nekonzistentná a nesprávna aktualizácia oprávneným používateľom. Ak pristupuje k dátam viac používateľov súčasne umožňuje SQL zdieľané využívanie dát a zaisťuje hladký priebeh činností. SQL nie je plnohodnotným samostatným programovacím jazykom, pretože neobsahuje riadiace programové konštrukcie a požadované prvky, ktoré by mal obsahovať každý obecný programovací jazyk (priraďovací a podmienený príkaz, príkaz cyklu atď.). Z uvedeného dôvodu sa jeho príkazy vyvolávajú buď z iného tzv. hosťovského programovacieho jazyka (napríklad C, Visual Basic, C++) alebo sa využíva neštandardizovaná nadstavba jazyka SQL, tzv. jazyk štvrtej generácie (4GL).

Spomenuté databázové systémy možno rozdeliť z hľadiska dostupnosti na dve základné skupiny: komerčné produkty (Microsoft SQL Sever, Oracle Database, IBM DB2, Sybase Adaptive Server, MS Access) a voľne šíriteľné produkty (MySQL, PostgreSQL). Ich použitie a nasadenie v jednotlivých firmách závisí od viacerých kritérií a podmienok, napríklad objem finančných prostriedkov, účel použitia systému, objem spracovávaných dát, veľkosť databázových súborov, spoľahlivosť, bezpečnosť a podobne.

2.1 MySQL

MySQL vyvinula švédska spoločnosť TcX v roku 1996 a v súčasnosti predstavuje jednotku medzi otvorenými databázovými systémami. Patrí medzi jednoduchšie databázové servery, pretože implementuje len určitú podmnožinu funkcií veľkých DB serverov ako Oracle a MS SQL Server. Podporuje širokú škálu platforiem a operačných systémov (Linux, Solaris, OS/2, Unix). Medzi hlavné prednosti a výhody patria jednoduchá inštalácia a obsluha, vysoký výkon, nulová cena, rýchlosť, prenositeľnosť, dobrá podpora prepojení s okolitým svetom a mnoho SQL funkcií. MySQL je tiež kompatibilná so štandardom ANSI SQL92 a poskytuje aj príkazy na podmienené mazanie tabuliek. MySQL dokáže uchovať 4 GB dát v jednej tabuľke. Pre zaručenie bezpečnosti má v sebe zabudovaný systém užívateľov, hesiel a prístupových práv. Nevýhodou je, že nepodporuje transakčné spracovanie, uložené procedúry, triggery a vnorené dotazy.

V súčasnosti možno hovoriť o troch verziách MySQL: 3, 4 a 5. Verzia 3 patrí už k zastaraným a väčšinou sa nepoužíva. Najrozšírenejšou je verzia 4. Verzia 5 sa len

vyvíja a je budúcnosťou MySQL, ktorá odstráni nedostatky v oblasti uložených procedúr, kurzorov, triggerov a pohľadov.

2.2 PostgreSQL

PostgreSQL je multiplatformový klient – server databázový systém podporujúci všetky moderné operačné systémy (OS/2, Novell, Linux). Bol vyvinutý na Berkeleyskej katedre počítačových vied Kalifornskej univerzity v USA. Je voľne šíriteľný pod BSD licenciou umožňujúcou vlastné úpravy a šírenie binárneho kódu. Predstavuje objektovo – relačný systém podporujúci vnorené dotazy, definovanie vlastných funkcií a triggerov, vytváranie vlastných dátových typov a možnosť dedenia definícií tabuliek. Okrem klasického prístupu chráneného heslom umožňuje overenie pomocou tzv. identifikačného protokolu.

Stabilnou verziou používanou aj v súčasnosti je verzia 8.0.4, ktorá má oproti predošlej verzii zlepšený výkon a dostupnosť, podporu polí, ľahšiu konfiguráciu, polymorfné funkcie atď. Od verzie 8.1 sa očakáva podpora dvojfázového potvrdzovania, partitioning, vyšší výkon, podpora rekurzívnych dotazov, balík funkcií na správu databázy a optimalizáciu dotazov.

2.3 MS Access

MS Access patrí medzi produkty firmy Microsoft a zaraďuje sa do oblasti jednoduchých a desktopových databázových systémov. MS Access je určený pre verzie operačného systému Windows a buď je šírený ako súčasť balíka MS Office alebo ako samostatná aplikácia. Nie je voľne šíriteľný. Je to výkonný databázový program s obrovským množstvom nástrojov a funkcií. Aj keď je veľmi jednoduché pochopiť koncept rozhrania programu, systém pomocníka a vstavaných sprievodcov, ktorí do istej miery uľahčujú realizáciu základných úloh, je tu stále pomerne rozsiahla oblasť na preskúmanie. V aplikácii možno ukladať dáta vnútorne alebo sa pripojiť k dátam z vonkajších zdrojov. Je možné si vytvoriť jednoduchú databázu, alebo s určitou znalosťou programovania vytvoriť aplikáciu pre používateľov v malých firmách (napríklad s prepojením na Microsoft Word alebo Excel), alebo vytvoriť prepracovanú databázu klient - server pre viacero užívateľov, ktorá bude fungovať na sieti intranet či Internet (s prepojením na dáta uložené v databáze Microsoft SQL Server). Aplikácia Access poskytuje skvelé nástroje na vytváranie formulárov určených na zadávanie a upravovanie dát a nástroje na vytváranie zostáv, ktoré možno použiť na zhromažďovanie a prezentovanie dát. Navyše umožňuje vytvárať dátové stránky na zobrazovanie informácií a dát na sieti WWW alebo firemnej sieti intranet a množstvo iných funkcií.

MS Access je poskytovaný tiež vo viacerých verziách. V súčasnosti je najpoužívanejšia verzia 2003, ale postupne sa na trh dostáva verzia 2007.

2.4 Microsoft SQL Server

Firma Microsoft má významné postavenie na databázovom trhu vďaka produktu SQL Server. Najnovšou verziou je SQL server 2005, pričom sú ešte stále využívaný jeho predchodcovia SQL Server 2003 a SQL Server 2000. Tento databázový systém má prínos hlavne pre veľké podniky. Je veľmi výkonný a jednoducho použiteľný. Jediný veľký nedostatok je, že pracuje len pod jednotlivými verziami operačného systému Windows.

Je integrovateľný s používateľskými kontami a bezpečnostnými prvkami operačných systémov Windows. Okrem komerčnej verzie existuje aj voľne šíriteľná verzia s označením MSDE (Microsoft Desktop Engine), ktorá obsahuje plne funkčné jadro systému MS Server. Primárne je určená na vývoj a testovanie databázových aplikácií a pri prevádzke aplikácií na lokálnych počítačoch. MSDE má obmedzenia v počte súčasných prístupov procesov (maximálne 5) a vo veľkosti databázových súborov (maximálne 1 GB), ale výhodou je, že kedykoľvek možno MSDE vymeniť za plnú verziu MS SQL Server bez potreby zmeny aplikácie. Za nástupcu MSDE možno považovať verziu MS SQL Server 2005 Express Edition, ktorá predstavuje odľahčenú verziu plnej verzie databázového systému MS SQL 2005 Server. Ale aj tak predstavuje spoľahlivý a bezpečný systém a systém s jednoduchou administráciou, podporou XML a dynamického ladenia parametrov databázy. Medzi obmedzenia tejto verzie sa zaraďujú: podpora len jedného procesora, podpora maximálne 1GB RAM a maximálnej veľkosti databázy 4GB.

Okrem primárnych úloh databázového servera poskytuje MS SQL Server 2005 aj nasledovné služby: analytické, reportovacie, notifikačné, služby pre transformáciu údajov, replikačné, integračné, administrátorské, vývojárske a podobne. Do podstatných vlastností systému patria: odolnosť voči chybám, on-line vykonávanie obnovovacích operácií, rýchla obnova údajov, šifrovanie obsahu databázy, presnejšie špecifikovanie a vymedzenie prístupových privilégií, integrovaný nástroj na správu databázy a prácu s databázovými objektmi, trvalý snapshot na disku a možnosť rozdelenia tabuliek na partície. Okrem spomenutých služieb obsahuje podporu pre dátový sklad (Data Warehouse) a dolovanie dát (Data Mining).

2.5 Oracle Database

Firma Oracle je najväčším dodávateľom podnikového softvéru. V súčasnosti má na trhu databázový systém Oracle Database 10g, ktorý je pokračovateľom stále používanej verzie Oracle9i Database. Oracle Database 10g predstavuje kompletnú platformu na ukladanie, správu a analýzu dát. Popri unikátnej bezpečnosti, výkonnosti a spoľahlivosti prináša rozšírenie jadra o podporu Enterprise Grid-u, výrazné zjednodušenie a automatizáciu správy a finančne výhodnú verziu pre malé firmy (Oracle Database 10g Standard Edition One). Oracle Database 10g je možné efektívne nasadiť na akejkoľvek hardvérovej platforme a operačnom systéme. Spolu s Oracle Application Server 10g a Oracle Enterprise Manager 10g tvoria Oracle Grid Computing, ktorý predstavuje novú softvérovú architektúru, ktorá spája pamäte, servery, aplikácie a disky, tak aby dokázalo vyhovieť neustále sa meniacim požiadavkám a aby spracovanie dát bolo čo najlacnejšie a najspoľahlivejšie. Oracle Enterprise Manager zabezpečuje automatickú správu a riadenie databázy pomocou grafického diagnostického okna Database Control. Správcovia môžu aktívne monitorovať databázu, prijímať výstrahy a rady na zabezpečenie optimálneho výkonu a spoľahlivosti, automaticky vyladiť databázu, eliminovať zložité a opakujúce sa úlohy. Dôležitou vlastnosťou Oracle 10g Database je integrovaný clusterware, ktorý predstavuje súbor spoločných klastrovacích služieb na jednoduchšie vytváranie a prevádzku databázových klastrov. Ďalším podstatným softvérom je Automatic Storage Management (ASM), ktorý zabezpečuje výrazné zjednodušenie konfigurácie a správy fyzického úložiska dát pre databázu.

Podobne ako Microsoft aj Oracle má bezplatnú edíciu Oracle databázy, ktorá má názov Oracle Database 10g Express Edition. Obmedzenia tejto edície sú podobné ako pri

MS SQL Server 2005 Express Edition: podpora len jedného procesora, podpora maximálne 1GB RAM, obmedzenie jednej relácie v systéme a maximálnej veľkosti databázy 4GB.

Zjednodušenou, ale nie bezplatnou je edícia Oracle Database 10g Enterprise Edition. Je ideálna pre podniky, ktoré potrebujú zabezpečiť veľkoobjemové on-line spracovanie transakcií a dopytovo intenzívne aplikácie dátových skladov. Podporuje všetky štandardné relačné dátové typy ako aj úložiská dát pre XML, text, dokumenty, obrázky, audio, video a podobne. Navyše pomáha administrátorom ľahko diagnostikovať a odstrániť dopady chýb, ochranu pred kompletným zlyhaním systému, chráni dôveryhodnosť dát v sieti, umožňuje centrálne spravovanie používateľských kont a poskytuje komplexný súbor automatickej diagnostiky výkonnosti a monitorovacích funkcií databázy. Samozrejme obsahuje aj podporu pre dátový sklad (Data Warehouse) a dolovanie dát (Data Mining).

Záver

Spracovanie veľkých súborov dát je najvýhodnejšie pomocou niektorého zo spomenutých a popísaných databázových systémov. Je len na konkrétnej firme alebo fyzickej osobe aké zdroje má k dispozícii a na základe akých kritérií sa rozhodne. Databázové systémy MySQL, PostgreSQL a MS Access sú použiteľné v malých a stredných firmách a MS SQL Server a Oracle Database v niektorých stredných a prevažne vo veľkých firmách.

Článok je spracovaný v rámci riešenia grantovej úlohy VEGA 1/2631/05 „Analýza možností aplikácie viacrozmerných štatistických metód na skúmanie ekonomických výsledkov na príklade priemyslu SR prípadne iných oblastí ekonomiky,“

Literatúra:

1. Blahušiak, M.: Uchovávanie údajov a tvorba databázy v MS Access. Bratislava: Statis, 2006.
2. Momjian B.: PostgreSQL Praktický prúvodce. Computer press, 2004.
3. Stránka Oracle: http://www.oracle.com
4. Stránky Microsoft: http://www.microsoft.cz

Kontakt:

Ing. Martin Blahušiak,
KAI, FHI, Ekonomická univerzita v Bratislave
Tel: 02/672 95 861
E-mail: martin.blahusiak@fhi.sk

Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/2631/05 „Analýza možnosti aplikácie viacrozmerných štatistických metód na skúmanie ekonomických výsledkov na príklade priemyslu SR prípadne iných oblastí ekonomiky“

Value at risk II. Základné prístupy k modelovaniu

Martin Boďa[†]

Abstract: The aim of the paper is to follow up with the article published in Forum Statisticum Slovacum No 4/2006 wherein, amongst others, the framework of Value at risk (VaR) was introduced and its utilization as a risk measure was discussed. In this continuation of the series, a more concrete insight into the workings of VaR is offered and its methodologies are briefly summarized. An especial accent is laid upon the importance of risk mapping in VaR forecasting as a prerequisite step in the process of the application of a respective VaR method. Of divers approaches to constructing VaR forecasts, the focus of the paper is directed to those most applied, which are (i) variance-covariance method, (ii) synthetic parametric models, (iii) Monte Carlo simulation method, (iv) delta-gamma method, and eventually, (v) historical simulation method.

Keywords: value at risk (VaR), risk mapping, variance-covariance method, synthetic parametric models, Monte Carlo simulation, delta-gamma method, historical simulation.

1. Úvod

Predložený príspevok je prirodzeným pokračovaním článku *Value at risk I. Value at risk ako miera rizika, alternatívy, nedostatky a regulačný aspekt* publikovanom vo *Forum Statisticum Slovacum 4/2006*, v ktorom bola predstavená koncepcia value at risk (VaR) a naznačený potenciál jeho využitia ako miery finančného (trhového) rizika. V tomto príspevku je prvotné koncepčné poňatie value at risk dotvorené už konkrétnejším aplikačným rozmerom, a síce uvedením základných metodológií využívaných pri samotnom výpočte v praxi.

Koncept value at risk ako miery rizika vychádza zo stotožnenia rizika s maximálnou očakávateľnou stratou hodnoty portfólia pri špecifikovanej hladine spoľahlivosti $1-\alpha$. Z aplikačného hľadiska zodpovedá výpočet value at risk konštrukcii ľavostranného intervalu spoľahlivosti výnosov portfólia R (definovaných ako náhodná premenná na pravdepodobnostnom priestore (Ω, F, P)). Value at risk je teda kvantilovým zobrazením rizika, pre ktoré platí $VaR_\alpha : [V \subseteq \Omega] \longrightarrow [-\inf\{x:\ F_R(r) \geq \alpha\} \in \Re^*]$, kde $F_R(r)$ je distribučná funkcia výnosov portfólia.

Žiada sa zdôrazniť, že **kým koncept value at risk je jednoduchý a ľahko pochopiteľný, jeho praktická aplikácia už nie je tak ideálne naplniteľná** a realizuje sa vo forme výstavby relatívne komplikovaných modelov pri predpokladoch, ktoré sa v niektorých prípadoch vzďaľujú od reality. Divergencia niektorých predpokladov bola diskutovaná podrobne v Boďa (2006).

2. Základné východiská

Ústrednými pojmami v manažmente finančného rizika sú *(obchodné) portfólio*[1] a *výnos*. Portfólio bude chápané ako určitá množina finančných inštrumentov $I_\bullet$ v určitej štruktúre: $\Pi = \{i \in \Omega:\ [I_i; w_i]\}$, pričom sa pripúšťajú dlhé i krátke pozície a vyznačujú sa znamienkom pri váhach $w_\bullet$. Portfólio bude v danom čase konštantné a v priebehu uvažovaného obdobia sa jeho zloženie nebude meniť, preto sa pri váhach a symbole portfólia abstrahuje od časového indexu.

[†] Bc. Martin Boďa. Univerzita Mateja Bela v Banskej Bystrici, Ekonomická fakulta, Tajovského 10, 975 90 Banská Bystrica. E-mail: ma_bo@azet.sk.

[1] Finančné riziko sa neprisudzuje cenným papierom obstarávaným najmä za účelom kontroly, a preto sa v tejto nadväznosti hovorí o obchodnom portfóliu, ktoré tvoria cenné papiere a ostatné inštrumenty určené na obchodovanie alebo zabezpečenie. Analyticky sa portfólio skladá z finančných aktív a ich derivátov (ktoré sú spoločne označované pojmom finančné inštrumenty).

Jednotlivé inštrumenty okrem kvantitatívneho zastúpenia v portfóliu charakterizuje ich trhová cena (v istom čase t) $P_{\bullet t}$. Trhovú hodnotu portfólia v čase t potom možno zapísať $\mathbf{P}_t(\Pi) = \sum_{i=1}^{i=n} w_i P_{it}$. Inštrumentom za obdobie držby τ prislúcha výnos $R_{\bullet}^{\tau}$ definovaný ako logaritmická zmena trhovej hodnoty $R_{\bullet}^{\tau} := \ln(P_{\bullet t+\tau}/P_{\bullet t})$. Znamienko výnosu potom určuje, či sa hovorí o zisku alebo strate. Analogicky sa konštruuje výnos portfólia R^{τ} za obdobie držby τ a teda $R^{\tau} := \ln(\mathbf{P}_{t+\tau}(\Pi)/\mathbf{P}_t(\ ))$.[2]

3. Zobrazenie rizika

Primárnym krokom pri kalkulácii value at risk je zobrazenie rizika (*risk mapping*)[3] do peňažných tokov asociovaných s príslušným portfóliom. Táto operácia spočíva v tom, že peňažným tokom plynúcim z držby daného inštrumentu sa priradia rizikové faktory a inštrumenty sú rizikovo kvalifikované z hľadiska rizikovej expozície ich peňažných tokov. Hodnota inštrumentov je vyjadrená trhovou hodnotou ich peňažných tokov ovplyvňovanou vytypovanými rizikovými faktormi. Potom je portfólio dezagregované do identifikovaných N peňažných tokov $CF_{\bullet k}$ a im priradeným rizikovým faktorom $f_{\bullet k} \sim R_{\bullet k}$.

Zobrazenie rizika nie je samoúčelné, ale je motivované snahou

(i) identifikovať peňažné toky a ich expozíciu rizikovým faktorom: $I_{\bullet} \longrightarrow \{[CF_{\bullet k};\ f_{\bullet k}]\}_{k\in\Pi}$ a

(ii) zredukovať dimenziu rizikových faktorov: $\dim(\mathbf{f}^{\underline{\Pi}}) = \underline{N} << \dim(\mathbf{f}^{\Pi}) = N$.

Redukcia dimenzie rizikovej expozície aproximáciou niektorých pozícií zastupiteľnými (pre model výhodnejšími) pozíciami zjednodušuje výpočet value at risk (eventuálne ho vôbec umožňuje). Obchodné portfólio finančnej inštitúcie tvoria rádovo desiatky až stovky finančných inštrumentov, ktorých verné zobrazenie produkuje realisticky až $N > 5000$ peňažných tokov. Napríklad použitie variančno-kovariančnej metódy v tomto prípade by si vyžadovalo odhad kovariančnej matice typu $N \times N$ a teda odhad $C_2(5000)$ parametrov. Každé zníženie dimenzie podstatne uľahčuje výpočet. Databáza RiskMetrics™ napríklad používa 380 rizikových faktorov.

Redukciu dimenzie si možno predstaviť ako zobrazenie $\mathsf{R}\colon \Pi \longrightarrow \underline{\Pi}$ za podmienok:

1. $\mathbf{P}_t(\Pi) = \mathbf{P}_t(\underline{\Pi})$ alebo $\mathbf{P}_t(\Pi) \approx \mathbf{P}_t(\underline{\Pi})$, tzn. že trhová hodnota portfólia sa (veľmi) nezmení a
2. $VaR_{\alpha}(\Pi \mid \Psi_t) = VaR_{\alpha}(\underline{\Pi} \mid \Psi_t)$ alebo $VaR_{\alpha}(\Pi \mid \Psi_t) \approx VaR_{\alpha}(\underline{\Pi} \mid \Psi_t)$, tzn. že odhad value at risk na čas $t+\tau$ sa pri danej informácii Ψ_t v čase t (veľmi) nezmení.

Výsledkom zobrazenia rizika môžu byť (vzhľadom na diferencovateľnosť podľa času) lineárne pozície alebo nelineárne pozície, ktoré potom podmieňujú použitie konkrétnej metódy. Klasifikáciu finančných inštrumentov z hľadiska tohto kritéria približuje schéma č. 1. Príkladom nelineárnej pozície môže byť napríklad európska kúpna opcia na cudziu menu a príkladom

---

[2] Použitie logaritmických výnosov rezultuje z prístupu, ktorý uvažuje o cenách akcií a odvodene aj o cenách niektorých ďalších finančných aktív generovaných spojitým stochastickým procesom v tvare $P_{\bullet t+\tau} = P_{\bullet t} \exp(\mu t + \sigma W_t)$, kde konštantyμ a $\sigma > 0$ sú postupne drift a volatilita a W_t je Wienerov proces na danom priestore $(\Omega, \mathsf{F}, \mathsf{P})$. Výnosy sú potom výsledkom procesu ξ: $\mathrm{d}P_{\bullet t+\tau}/P_{\bullet t} = \mu \mathrm{d}t + \sigma \mathrm{d}W_t$, ktorý po uvážení vlastností Wienerovho procesu a zavedení prírastku času $\mathrm{d}t = \tau$ nadobúda tvar ξ: $\mathrm{d}P_{\bullet t+\tau}/P_{\bullet t} = \mu\tau + \sigma\sqrt{\tau}\,\varepsilon$, kde ε je premenná s rozdelením $N(0, 1)$. Tento zápis je ale ekvivalentný so zápisom ξ: $\ln(P_{\bullet t+\tau}/P_{\bullet t}) \equiv \mathrm{d}P_{\bullet t+\tau}/P_{\bullet t} = \mu_\tau + \sigma_\tau\varepsilon$, kde μ_τ je konštantný drift za dané obdobie držby (resp. stredná hodnota procesu ξ) a σ_τ volatilita za dané obdobie držby (resp.σ_τ^2 disperzia procesu ξ), a implikuje tvrdenie, že ceny aktív majú logaritmicko-normálne rozdelenie $LN(\mu_\tau, \sigma_\tau^2)$ (resp. výnosy normálne rozdelenie $N(\mu_\tau, \sigma_\tau^2)$).

[3] V článku sa z dôvodu lepšej vypovedateľnosti preferuje matematický význam anglického pojmu *mapping*, t. j. zobrazenie. Česko-slovenská literatúra venovaná problematike value at risk pojem zjednodušuje a hovorí v tomto prípade o mapovaní (rizika).

lineárnej pozície dlhopis s plávajúcimi kupónmi viazanými na referenčný úrokový vzor.[4] Isté technické ťažkosti vznikajú v situáciách, keď peňažné platby vznikajú v období, pre ktoré nie sú dostupné úrokové kotácie (napr. kupón dlhopisu je splatný o 11 mesiacov a sú dostupné iba 9-mesačné a 12-mesačné úrokové sadzby). Obyčajne sa použije lineárna (alebo iná) interpolácia sadzieb alebo peňažných tokov: tak, aby bola zachovaná trhová hodnota aj expozícia trhovému riziku. Bližšie sa problematike zobrazenia rizika venujú podrobne Cipra (2002), RiskMetrics™ (1996) a Mina a Xiao (2001).

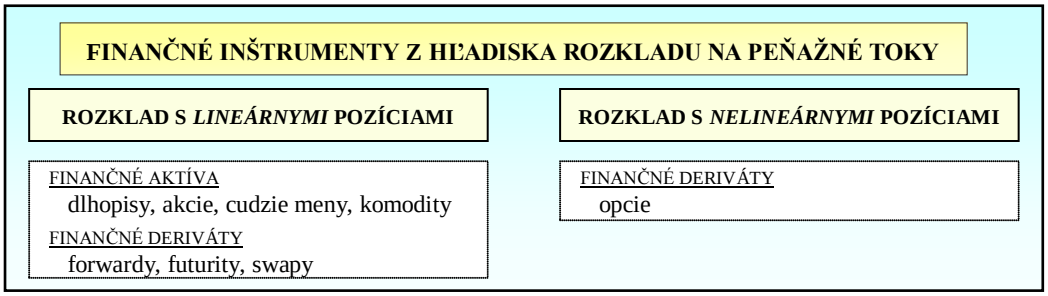


Schéma č. 1 *Finančné inštrumenty z hľadiska charakteru rozkladu* (Zdroj: vlastné spracovanie)

4. Modely value at risk

V princípe možno pri odhade value at risk vyjsť z modelov, ktoré sú fundované na predpokladoch o tvare rozdelenia rizikových faktorov, alebo z modelov, ktoré takéto predpoklady neprijímajú a vychádzajú výlučne z historických dát. Toto členenie znázorňuje schéma č. 2.

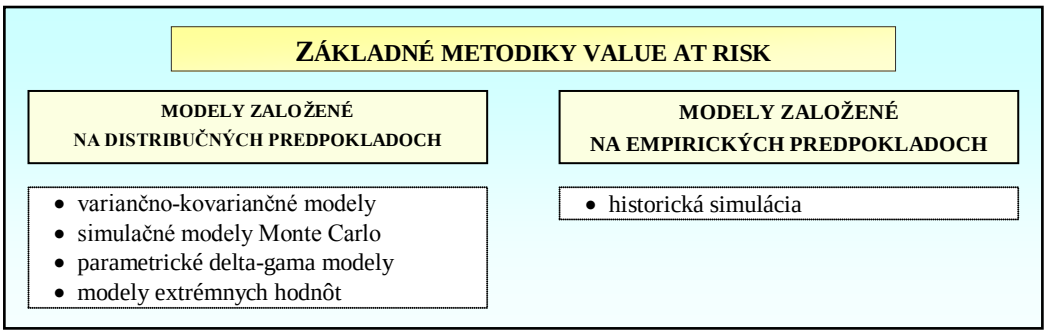


Schéma č. 2 *Najpoužívanejšie metodiky odhadu value at risk* (Zdroj: vlastné spracovanie)

Voľbu konkrétneho modelu value at risk okrem vybavenosti používateľa výpočtovou technikou a schopnosti vysporiadať sa s výpočtovými nárokmi ovplyvňuje najmä:

- *Charakter pozícií v portfóliu: linearita* vs. *nelinearita.* V optimálnom prípade (keď opčné inštrumenty v pôvodnom portfóliu neboli zastúpené alebo sa dali efektívne zobraziť do

[4] V prvom – nelineárnom – príklade je trhová cena menovej opcie v čase *t* určená Blackovým-Scholesovým vzťahom

$$P_t\left(O_{FX}^{C}\right) = S_t e^{-r_f \Delta t}\Phi(d_1) - X e^{-r\Delta t}\Phi(d_2), \quad \text{kde} \quad d_1 = \left[\ln\left(S_0 / X\right) + \left(r - r_f + 0.5\sigma^2\right)/\Delta t\right] \Big/ \left(\sigma\sqrt{\Delta t}\right) \quad \text{a} \quad d_2 = d_1 - \sigma\sqrt{\Delta t},$$

pričom S_t je devízový kurz v čase *t*, *r* je tuzemská bezriziková úroková miera a r_f zahraničná bezriziková úroková miera, symbol Δt označuje čas zostávajúci do expirácie opcie, *X* realizačnú cenu opcie a symbol σ zapisuje volatilitu devízového kurzu. V druhom – lineárnom – prípade je hodnota dlhopisu zastupiteľná zápisom

$$P_t(B) = r_{t+\delta t}^{\#} N e^{-\delta t \cdot r_{t|\delta t}} + r_{t+2\delta t}^{\#} N e^{-2\delta t \cdot r_{t|2\delta t}} + \ldots + r_{t+2\delta t}^{\#} N e^{-(n-1)\delta t \cdot r_{t|(n-1)\delta t}} + \left(1 + r_{t+n\delta t}^{\#}\right) N e^{-n\delta t \cdot r_{t|n\delta t}},$$

kde *N* je nominálna hodnota dlhopisu, $r_T^{\#}$ referenčná sadzba určujúca výšku kupónovej platby v čase *T*, δt denotuje (v tomto prípade rovnaké) časové rozpätie medzi dvoma kupónmi, $r_{t|T}$ úrokovú sadzbu bezkupónových dlhopisov v čase *t* so splatnosťou v čase $(t + T)$.

lineárnych pozícií) je výsledkom zobrazenia rizika lineárne zložené portfólio $\underline{\Pi}$, ktorého trhová hodnota je vyjadriteľná vo vektorovom tvare $\mathbf{P}_t(\underline{\Pi}) = {}^{T}_{\underline{\Pi}}$[5].

- *Predpokladaná forma závislosti a konštruovaný vzor volatility rizikových faktorov v portfóliu.* Pri modelovaní sa *vždy* prijíma domnienka, že rizikové faktory sú tendenčné vo vzájomnom vývoji, a spravidla sa predpokladá, že ich vzájomný vývoj je lineárny. Tento výhodný (uľahčujúci) predpoklad ústi do zapisovania vzájomných vzťahov do kovariančnej matice. Ďalšia voľba sa potom týka konkrétneho modelu volatility a motivuje spôsob, akým sa kovariančná matica vôbec konštruuje. Toto rozhodnutie sa obyčajne odvíja od charakteru finančného makroprostreda alebo konkrétnych podmienok modelovania.
- *Historické dáta a ich informačná hodnota.* Historické dáta indikujú, aké predpoklady o rizikových faktoroch sú (ešte) realistické a tiež podmieňujú rozhodnutie o modeli volatility.

V prípade lineárne zobrazeného portfólia $\underline{\Pi}$ je – nakoľko uvažujeme v logike statického portfólia – zmena jeho hodnoty určená zmenami (výnosmi) rizikových faktorov r_f. Vzťah $\mathbf{P}_{t+\tau}(\underline{\Pi}) - \mathbf{P}_t(\underline{\Pi}) = {}^{T}\left(\underline{\Pi}_{t+\tau} - \underline{\Pi}_{t}\right) = \sum_{i=1}^{i=N} \omega_i f_i (e^{r_i - r_{fi}} - 1)$ podmieňuje konštrukciu kovariančnej matice medzi výnosmi rizikotvorných faktorov, prostredníctvom ktorej sa meria volatilita hodnoty portfólia. Klasický vzor volatility by zostavoval kovariančnú maticu podľa schémy č. 3a a každému pozorovaniu zúčastnenému na odhade prvkov matice by priraďoval rovnakú váhu. Volatilné podmienky na finančných trhoch si vynútili inováciu kovariančnej matice, aby lepšie indikovala zmenu pôsobiacich faktorov tým, že novším dátam sa prisudzuje vyššia váha v porovnaní so staršími údajmi – napr. model EWMA (*exponentially weighted moving averages*) v schéme č. 3b. Alebo sa začali používať dynamické modely podmienenej heteroskedasiticity, najčastejšie Bollerslevov model GARCH(1, 1) (*generalized autoregressive conditional heteroskedasticity*) uvedený v schéme č. 3c. Najmä modely GARCH(1, 1) a EWMA (ako špeciálny prípad modelu GARCH(1, 1)) sa zakladajú na spektre špecifických predpokladov: o problémoch modelovania volatility možno nájsť bližšie informácie napr. v Cipra (2002), Arlt a Arltová (2002), Andersen et al. (2005). Zápis modelov v podobe uvedenej v schéme č. 3 je založený na predpoklade $Er_{f\Box} = 0$.[6]

(a) KLASICKÝ MODEL	**(b) MODEL EWMA**[7]	**(c) MODEL GARCH(1, 1)**
$\boldsymbol{\Sigma}_{t\|t+1} = \begin{pmatrix} \sigma^2_{1[t]} & \gamma_{12[t]} & \cdots & \gamma_{1\underline{N}[t]} \\ \gamma_{21[t]} & \sigma^2_{2[t]} & \cdots & \gamma_{2\underline{N}[t]} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{\underline{N}1[t]} & \gamma_{\underline{N}2[t]} & \cdots & \sigma^2_{\underline{N}[t]} \end{pmatrix}$	${}_E\boldsymbol{\Sigma}_{t\|t+1} = \begin{pmatrix} {}_E\sigma^2_{1[t]} & {}_E\gamma_{12[t]} & \cdots & {}_E\gamma_{1\underline{N}[t]} \\ {}_E\gamma_{21[t]} & {}_E\sigma^2_{2[t]} & \cdots & {}_E\gamma_{2\underline{N}[t]} \\ \vdots & \vdots & \ddots & \vdots \\ {}_E\gamma_{\underline{N}1[t]} & {}_E\gamma_{\underline{N}2[t]} & \cdots & {}_E\sigma^2_{\underline{N}[t]} \end{pmatrix}$	${}_G\boldsymbol{\Sigma}_{t\|t+1} = \begin{pmatrix} {}_G\sigma^2_{1[t]} & {}_G\gamma_{12[t]} & \cdots & {}_G\gamma_{1\underline{N}[t]} \\ {}_G\gamma_{21[t]} & {}_G\sigma^2_{2[t]} & \cdots & {}_G\gamma_{2\underline{N}[t]} \\ \vdots & \vdots & \ddots & \vdots \\ {}_G\gamma_{\underline{N}1[t]} & {}_G\gamma_{\underline{N}2[t]} & \cdots & {}_G\sigma^2_{\underline{N}[t]} \end{pmatrix}$
Špecifikácia modelu volatility		
$\hat{\sigma}^2_{i[t]} = \frac{1}{m-1}\sum_{k=t-1}^{k=t-m} r_{ik}{}^2$	${}_E\hat{\sigma}^2_{i[t]} \approx (1-\lambda)\cdot\sum_{k=t-1}^{k=t-m}\lambda^{t-k-1} r_{ik}{}^2$ *rekurzívna forma* ${}_E\hat{\sigma}^2_{i[t]} \approx \lambda\cdot {}_E\hat{\sigma}^2_{i[t-1]} + (1-\lambda)\cdot r^2_{i\ t-1}$	*ekonometrický model* ${}_G\sigma^2_{i[t]} = \upsilon_i + \alpha_i \cdot r^2_{i\ t-1} + \beta_i \cdot {}_G\sigma^2_{i[t-1]}$ (kde $\upsilon_i > 0, \quad \alpha_i, \beta_i \geq 0, \quad \alpha_i + \beta_i < 1$)
$\hat{\gamma}_{ij[t]} = \frac{1}{m-1}\sum_{k=t-1}^{k=t-m} r_{ik} r_{jk}$	${}_E\hat{\gamma}_{ij[t]} \approx (1-\lambda)\cdot\sum_{k=t-1}^{k=t-m}\lambda^{t-k-1} r_{ik} r_{jk}$ *rekurzívna forma* ${}_E\hat{\gamma}_{ij[t]} \approx \lambda\cdot {}_E\hat{\gamma}_{ij[t-1]} + (1-\lambda)\cdot r_{i\ t-1} r_{j\ t-1}$	*ekonometrický model* ${}_G\gamma_{ij[t]} = \upsilon_{ij} + \alpha_{ij}\cdot r_{i\ t-1} r_{j\ t-1} + \beta_{ij}\cdot {}_G\gamma_{ij[t-1]}$ (kde $\upsilon_{ij} > 0, \quad \alpha_{ij}, \beta_{ij} \geq 0, \quad \alpha_{ij} + \beta_{ij} < 1$)

Schéma č. 3 *Základné modely volatility: konštrukcia kovariančnej matice* (Zdroj: vlastné spracovanie)

[5] Možno ľahko odvodiť $\mathbf{P}_t(\underline{\Pi}) = \sum_{i=1}^{i=n} w_i P_{it} = \left\| I_i = \{[CF_{ik};\ f_{ik}]\}_{k=1}^{k=m(I_i)} \right\| = \sum_{i=1}^{i=n} w_i \sum_{k=1}^{k=m(I_i)} CF_{ikt} f_{ikt} = \sum_{i=1}^{i=\pi} \sum_{k=1}^{k=m(I_i)} w_{ik} CF_{ikt} f_{ikt} = \sum_{i=1}^{i=N} \omega_i f_{it} = {}^{T}\underline{\Pi} = {}_t(\underline{\Pi}) \cdot$

[6] I keď sa vo finančnej matematike venuje *správnej* funkčnej špecifikácii kovariančnej matice veľká pozornosti, Lopez a Walter (2000) komparatívne ukázali, že konštrukcia kovariančnej matice pre praktické modelovanie value at risk nemá podstatný význam. Odhady value at risk sú ovplyvňované predovšetkým distribučnými (alt. empirickými) predpokladmi modelu.

[7] Pre model EWMA *exponenciálneho zabúdania staršej informácie* sa volí $\lambda = 0.94$ (pre denné údaje), resp. $\lambda = 0.97$ (pre mesačné dáta). Hodnoty parametra boli optimalizované empiricky v RiskMetrics™ (1996) a aj iné štúdie potvrdili vhodnosť voľby hodnôt.

Pri výstavbe modelov sa obyčajne oddeľujú predpoklady týkajúce sa modelu volatility (a konštrukcie kovariančnej matice) od predpokladov týkajúcich sa distribučnej špecifikácie modelu. Navyše sa tradične predpokladá – v súlade s modelmi volatility v schéme č. 3 – a rešpektuje sa to i v tomto texte, že výnosy rizikových faktorov sú *v priemere* nulové: $Er_{f\bullet} = 0$.

4.1 Variančno-kovariančný parametrický model (*VCV model: variance-covariance model*)

Najjednoduchší distribučný model[8] vychádza z výhodného predpokladu, že *rozdelenie výnosov rizikových faktorov je normálne* (tzn. $r_{f\square} \sim N(0;\sigma^2_{\bullet})$), a je zrejmé, že je oprávnene aplikovateľný vtedy, keď vzťahy v zobrazenom portfóliu $\underline{\Pi}$ sú lineárne. Individuálna normalita výnosov rizikových faktorov vysvetľuje $\underline{N}$ -rozmerné normálne rozdelenie pre vektor rizikových faktorov $\mathbf{r}_f$ so všetkými dôsledkami na rozdelenie výnosov celého portfólia $r_{\underline{\Pi}}$. Ak je totiž zaručená normalita $\mathbf{r}_f \,\Sigma\, N_{\underline{N}}(\mathbf{0};\ )$, je rozdelenie výnosov portfólia opäť normálne a platí $r_{\underline{\Pi}} \sim N\left(\boldsymbol{\omega}^{\mathrm{T}}\mathbf{0};\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\omega}\right)$. Celý problém nájdenia odhadu value at risk sa redukuje na výpočet kvantilu rozdelenia $r_{\underline{\Pi}}$, ktorý je určený vzťahom $VaR_{\alpha}\left(r_{\underline{\Pi}}\right) = -u_{1-\alpha}\sqrt{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\omega}}$, kde $u_{1-\alpha}$ je α.100%-ný kvantil normovaného normálneho rozdelenia[9].

4.2 Syntetické parametrické modely a modely teórie extrémnych hodnôt

Už článok Boďa (2006) poukázal na spornosť validity predpokladu o normalite rozdelenia výnosových faktorov. Vzhľadom na tento poznatok pri modelovaní (ak to nie je náležité)

- sa upúšťa od predpokladu normality výnosov portfólia pre vlastnú distribučnú špecifikáciu modelu, ale súčasne <u>často</u>
- sa zachováva predpoklad viacrozmernej normality výnosov pre špecifikáciu modelu volatility.

Výsledkom je pseudosyntetický model, v ktorom majú výnosy portfólia *vhodné* rozdelenie so strednou hodnotou $Er_{\underline{\Pi}} = 0$ a disperziou $Dr_{\underline{\Pi}} = \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\omega}$. Vlastná predpoveď value at risk sa zostavuje vzťahom $VaR_{\alpha}\left(r_{\underline{\Pi}}\right) = -q_{1-\alpha}\sqrt{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\omega}}$, kde $q_{1-\alpha}$ je α.100%-ný kvantil rozdelenia *vhodného* pre štandardizované výnosy portfólia $z = r_{\underline{\Pi}}/\hat{\sigma}_{\underline{\Pi}}$ s charakteristikami $Ez = 0$ a $Dz = 1$.[10]

Ako rozdelenie výnosov portfólia sa volia najčastejšie

- Studentovo t-rozdelenie (pre svoju symetrickosť a leptokurtickosť) a zovšeobecnené Studentovo t-rozdelenie (pre svoju prispôsobivosť empirickým dátam) alebo
- zmes normálnych rozdelení (ktoré vnímajú výnosy portfólia ako výber obyčajne z dvoch *normálne* rozdelených populácií; kde jedna populácia predstavuje tiché obchodné dni a druhá populácia hektické dni s abnormálnymi ziskami alebo stratami – teda hrubé konce).[11]

Osobitným prístupom sa kvalifikuje **teória extrémnych hodnôt** (*EVT: extreme value theory*), ktorá umožňuje flexibilne pristupovať k modelovaniu extrémnych hodnôt, čo si svojou povahou za cieľ stavia i koncept value at risk. Rozdelenia extrémnych hodnôt umožňujú vzdať sa obmedzujúceho predpokladu o špecifickom type rozdelenia a modelovať správanie sa veličiny,

---

[8] Variančno-kovariančná metóda sa zvykne nazývať o i. aj delta-normálna metóda alebo portfólio-normálna metóda.

[9] Proti tomuto rýdzoštatistickému zápisu sa v niektorej literatúre preferuje analytický zápis, ktorý umožňuje identifikovať value at risk individuálnych pozícií: $VaR_{\alpha}\left(r_{\underline{\Pi}}\right) = \sqrt{\mathbf{var}^{\mathrm{T}}\mathbf{C}\mathbf{var}}$, kde $\mathbf{var} = (u_{\alpha}\omega_{\bullet}\sigma_{\bullet})^{T}$ a $\mathbf{C}$ je korelačná matica výnosov rizikových faktorov.

[10] Je nutné poznamenať, že pri takomto prístupe nie je možná analógia analytickej formy naznačenej v poznámke [9]. Separácia distribučného predpokladu a predpokladu vzoru volatility je vyvolaná snahou vyhnúť sa komplikovaným distribučným formám, ale i tak je možné pre jednotlivé faktory zvoliť vhodné marginálne rozdelenia a value at risk odhadovať ich agregáciou.

[11] Podrobnejší výklad možno nájsť v Mina a Xiao (2001), Hull a White (1998) a Cipra (2002) a nadväzujúcich odkazoch.

až keď presiahne stanovenú hranicu. V prípade výnosov portfólia to znamená, že sú uvažované iba (veľké) straty portfólia (ktoré presiahnu určitú referenčnú hranicu), pričom z pohľadu tvaru funkcie hustoty pravdepodobnosti to znamená, že sú uvažované iba straty portfólia v (hrubých) koncoch. Táto skutočnosť akoby oprávňovala použitie tohto prístupu pre finančné dáta, ktoré sú tendenčné svojou leptokurtickosťou (a teda hrubými koncami). Takto pristupuje k modelovaniu prístup *POT* (*peaks-over-threshold;* maximá za prahom), obyčajne založený na dvojparametrickom zovšeobecnenom Paretovom rozdelení (*GPD: generalized Pareto distribution*) s distribučnou funkciou $G_{\xi,\beta}(x) = 1-(1+\xi x/\beta)^{-1/\xi}$, ak $\xi \neq 0$, resp. $G_{\xi,\beta}(x) = 1-e^{-x/\xi}$, ak $\xi = 0$ (pre x za vhodne zvoleným prahom t). Toto rozdelenie je vhodné pre modelovanie rizika pre $\xi > 0$, keď má hrubé konce. Na základe špecifikácie distribučnej funkcie je možné zostrojiť odhadnú funkciu pre value at risk (pozri napr. Harmantzis et al. (2005) alebo McNeil (1999)).[12] Teóriu extrémnych hodnôt je možné aplikovať synteticky pre výnosy portfólia analogicky ako vyššie naznačené prístupy alebo analyticky vo viacrozmernej podobe (*MEVT: multivariate extreme value theory*). Z citovanej literatúry ju vysvetľuje McNeil (1999).

4.3 Modely stochastickej simulácie Monte Carlo

Metóda Monte Carlo je koncepčne zrozumiteľná, avšak výpočtovo a technicky značne náročná. Spočíva v simulovaní stochastického vývoja hodnoty portfólia a v následnom spočítaní odhadu value at risk zo simulovaných výnosov portfólia.

Východiskom pre aplikáciu metódy je vzor zobrazenia rizika: portfólio $\underline{\Pi}$, o ktorom už bolo poznamenané, že ideálne preň sú lineárne vzťahy medzi identifikovanými rizikovými faktormi $\mathbf{f}_{\underline{\Pi}}$ a výškou ich rizikovej expozície $\boldsymbol{\omega}$. Na základe historického vývoja sa pre rizikové faktory určí vhodný parametrický model ich spoločného stochastického vývoja reprezentovaný *vhodným* $\underline{N}$-rozmerným (združeným) rozdelením. Z posledných m pozorovaní (napr. $m = 251$) sa odhadnú parametre združeného rozdelenia a generujú sa $\underline{N}$-tice pseudonáhodných čísel zodpovedajúcich zvolenému združenému rozdeleniu tak, aby bola uchovaná empirická korelačná štruktúra. Počet simulácií je považovaný za dostatočný, ak obsahuje obyčajne viac ako 1000 (až 50000) $\underline{N}$-tíc náhodných čísiel. Simulované hodnoty sú bázou pre spočítanie hodnoty portfólia pre jednotlivé simulácie a výnosov portfólia určených týmito simuláciami. Odhad value at risk $VaR_\alpha(r_{\underline{\Pi}})$ je daný α.100%-ným kvantilom simulovaných výnosov.[13]

Hlavná výhoda tejto metódy vyplýva z možnosti simulovať vývoj aj nelineárnych vzťahov v portfóliu a teda uschopňuje celkom presne odhadnúť value at risk pre nelineárne portfólia. Jej použitie v praxi je podmienené predovšetkým týmto faktorom, pretože v prípade lineárnych portfólií sa adekvátnejšie javí použiť napr. variančno-kovariančnú metódu (v prípade viacrozmerných normálnych vzťahov) alebo jej modifikácie. Korektnosť výpočtu value at risk

---

[12] Pre úplnosť treba dodať, že existuje i prístup BMM (*block maxima method:* metóda bloku maxím), ktorý si na rozdiel od prístupu POT vyžaduje veľké množstvo pozorovaní. Podľa Fisherovej-Tippettovej vety normalizované maximum sérií (bloku) pozorovaní (vysokých hodnôt) $(max(X_n) - d_n)/c_n$ konverguje k limitnému rozdeleniu extrémnych hodnôt (*GEV: generalized extreme value*) $H_\xi(x)$, ktoré je definované $H_\xi(x) = \exp(-(1 + \xi x)^{-1/\xi})$ pre $\xi \neq 0$, resp. $H_\xi(x) = \exp(-\exp(-x))$ pre $\xi = 0$. Potom sa abnormálne straty vyjadrené value at risk modelujú napr. touto funkciou. Cf. McNeil (1999).

[13] Pretože je voľba správneho združeného rozdelenia problémovou záležitosťou, sprostredkuje sa často simulácia vhodnou kopulovou funkciou, pre ktorú sa vyžadujú iba marginálne distribučné funkcie rizikových faktorov. Podľa Sklarovej vety totiž ku každej združenej distribučnej funkcii so spojitými marginálnymi distribúciami existuje jednoznačne určená funkcia (kopula) $\boldsymbol{C}$: $\langle 0, 1\rangle^d \rightarrow \langle 0, 1\rangle$, pre ktorú platí $\boldsymbol{F}(x_1, x_2, ..., x_d) = \boldsymbol{C}_\theta[F_1(x_1), F_2(x_2), ..., F_d(x_d)]$. Potom stačí nájsť marginálne distribučné funkcie $F_\bullet$, definovať kopulu $\boldsymbol{C}_\theta$ s parametrami θ (a odhadnúť ich), generovať Monte Carlo realizácie d-rozmerného rovnomerného rozdelenia z $\langle 0, 1\rangle^d$ a cez kopulu $\boldsymbol{C}_\theta$ zostrojiť realizácie vektora $(x_1, x_2, ..., x_d)^{\mathrm{T}}$, tzn. teraz realizácie rizikových faktorov.

stochastickou simuláciou je úzko asociovaná s voľbou korektného modelu vývoja rizikových faktorov.

4.4 Delta-gama parametrické modely

Delta-gama metóda je vhodná pre nelineárne portfóliá a je v tomto prípade azda preferovanejšia než simulácie Monte Carlo. Vychádza z vhodnej aproximácie zmien v hodnote portfólia; obyčajne druhým (kvadratickým) rádom Taylorovho rozvoja. Pre zmenu hodnoty portfólia $\underline{\Pi}$ môžeme písať $\Delta\mathbf{P}_{t+\tau|}(\underline{\Pi}) = \sum_{i=1}^{i=N} \omega_i f_i (e^{r_i \cdot r_{fi}} - 1) \approx \Delta^{\delta\gamma} = \sum_{i=1}^{i=N} \delta_i r_{f_i} + 0.5 \sum_{i=1}^{i=N} \sum_{j=1}^{j=N} \gamma_{ij} r_{fi} r_{f_j}$, kde $\delta_i = f_i \cdot [\partial \mathbf{P} / \partial f_i]$ a $\gamma_{ij} = f_i f_j \cdot \left[\partial^2 \mathbf{P} / \partial f_i \partial f_j\right]$. Značenie δ (pre lineárne členy) a γ (pre kvadratické členy) determinovalo pomenovanie metódy. Nie je nutné ani žiaduce pre všetky inštrumenty v portfóliu použiť až kvadratickú aproximáciu: vo väčšine inštrumentov si vystačíme s lineárnym (delta) rozvojom. Toto platí i pre finančné inštrumenty denominované v cudzej mene, keď na hodnotu portfólia pôsobia *zložené* rizikové faktory $\tilde{f}_{\cdot} = f_{\circ} f_{*}$, z ktorých jeden je zahraničný rizikový faktor (napr. cena komodity denominovaná v cudzej mene) a druhý faktor je konverzný devízový kurz (na tuzemskú menu). Kvadratický (gama) rozvoj je obzvlášť vhodný pre opčné inštrumenty, kde súbežne pôsobiacimi faktormi sú tradične cena podkladového aktíva, bezriziková úroková miera, volatilita, doba do splatnosti a realizačná cena (pozri príklad v poznámke [4)]).[14]

Je evidentné, že na použitie metódy sa vyžaduje znalosť analytického vyjadrenia hodnoty portfólia, a je potrebné podotknúť, že použitie metódy na odhad (správneho) value at risk je *aproximatívne* a môže byť zdrojom nepresností, pretože aj pre druhý rád aproximácie je presnosť iba lokálna (pozri Mina a Ulmer (1999)).[15]

4.5 Historická simulácia

Historická simulácia (alebo *bootstrapping*: svojpomoc) na rozdiel od ostatných spomínaných metód nenanucuje rozdeleniu výnosov ani volatilite žiaden funkčný tvar (tzn. abstrahuje od distribučných predpokladov a predpokladov o modeli volatility) a vychádza z empirických pozorovaní. Avšak explicitne predpokladá, že rizikové podmienky finančných trhov sa nezmenia a historické dáta sú (veľmi) vhodné na formovanie očakávaní budúcnosti (*implicite* to vedie k tomu, že spočítané výnosy sú považované za nekorelované a identicky rozdelené). Portfólio v aktuálnom zložení sa precení podľa minulých cien a rizikových faktorov, spočítajú sa historické výnosy portfólia a value at risk sa určí ako α.100%-ný kvantil historicky simulovaných výnosov portfólia. Zvyčajne sa *generuje* podstatne viac ako 251 historických simulácií. Metóda si vyžaduje veľkú databázu minulých pozorovaní a môže byť (podobne ako simulácie Monte Carlo) komputačne náročná. Sinha a Chamú (2000) poukazujú na antagonizmus počtu simulácií: nižší počet simulácií (tzn. menej ako 251) nie je spôsobilý dostatočne identifikovať extrémne hodnoty a odhady value at risk sú nepresnejšie (optimistickejšie), kým vyšší počet minulých simulácií narušuje implicitný predpoklad o identickom rozdelení výnosov, pretože finančné dáta sú

---

[14] V prípade, že sa použije Taylorov rozvoj iba prvého rádu a prijme sa predpoklad o normalite rozdelenia výnosov, hovorí sa o delta-normálnej metóde. Je potom náležité využitie variančno-kovariančnej metódy a tieto dve metodológie splývajú.

[15] Nakoľko nie je možné nájsť analytické vyjadrenie rozdelenia Taylorovho rozkladu $\Delta^{\delta\gamma}$ (ktorý reprezentuje zmeny v hodnote portfólia), používa sa napr. Johnsonova transformácia na získanie hustoty rozdelenia $\Delta^{\delta\gamma}$. Johnson navrhol tri transformácie (napr. $f(X) = \lambda.\sinh[(X - \gamma)/\delta] + \xi$, kde $X \sim N(0, 1)$), ktoré závisia na štyroch parametroch, ktoré sú odhadované algoritmicky na základe prvých štyroch centrálnych momentov $\Delta^{\delta\gamma}$. Ďalšími zaujímavými prístupmi sú Cornishova-Fisherova aproximácia a Furierova inverzia, o ktorých možno získať informácie napr. v Mina a Ulmer (1999) alebo v matematicky intenzívnejšej forme v Härdle, Kleinow a Stahl (2002).

tendenčné meniacou sa volatilitou. Manganelli a Engle (2001) v tomto kontexte hovoria o *drastickom zjednodušovaní* a poukazujú zo spomenutých dôvodov na logickú inkonzistenciu metódy.[16] Skutočnosťou však zostáva, že metóda historickej simulácie je používaná a obľúbená, pretože je koncepčne jednoduchá a *transparentná*.

5. Zhrnutie a záver

Zo spôsobu, akým boli jednotlivé metódy prezentované, je zrejmé, že každá z nich má svoje slabiny a prednosti, ktoré limitujú alebo určujú rozsah jej použitia. Určujúcimi faktormi pre voľbu konkrétnej výpočtovej metodiky value at risk sú najmä

(1.) charakter a zloženie portfólia (osobitne z pohľadu linearity) a miera jeho expozície fluktuácii finančných trhov,

(2.) konkrétne podmienky na finančných trhoch, ktoré vplývajú na správanie sa výnosov a ich schopnosť zachovať si volatilitu konštantnú v čase, a

(3.) implementovateľnosť metódy value at risk v informačnom systéme inštitúcie, integrácia výpočtu value at risk do infraštruktúry systému a flexibilita systému vo výpočte value at risk.

Boli vypracované viaceré komparatívne štúdie zamerané na použitie týchto metód v praxi, ale ich výsledky sú *de facto* diferencované, čo možno uzavrieť tým, že správnosť voľby metodiky je v rozhodujúcej miere ovplyvnená individuálnymi okolnosťami používajúcej inštitúcie a prostredím, v ktorom sa nachádza. Pre každú metódu však platí, že jej spoľahlivosť je determinovaná intenzitou predpokladov, na ktorých je explicitne a implicitne vystavaná, a na miere, ako sú tieto predpoklady rešpektované pri technickej realizácii.

Spomedzi uvádzaných *základných* praxou preferovaných metód akýsi tradičný triumvirát tvoria *variančno-kovariančná metóda* (vďaka proliferácii investičnou spoločnosťou J. P. Morgan Chase a jej systémom a databázou RiskMetrics™), *metóda Monte Carlo* a *historická simulácia* (pre svoju jednoduchosť). Jednotlivé prístupy vo svetle základných hodnotiacich kritérií sú prezentované v nasledujúcej schéme č. 4.

ATRIBÚT	VCV metóda (delta-normálna metóda)	Syntetické EVT modely	Simulácia Monte Carlo	Delta-gama metóda	Historická simulácia
Distribučné predpoklady	Predpoklad o normalite rozdelenia výnosov a predpoklad iid	Predpoklad o rovnakom rozdelení a a predpoklad iid	Predpoklad o spoločnom vývoji rizikových faktorov (a ich nejakom rozdelení)	Predpoklad o rozdelení rizikových faktorov a a predpoklad iid	Distribučný predpoklad nahradený empirickou spoľahlivosťou minulých dát
Kovariančná matica (vzor volatility)	Nevyhnutná (musí byť pozitívne definitná)	Nevyhnutná (musí byť pozitívne definitná)	Nevyhnutná (musí byť pozitívne definitná)	Nevyhnutná (musí byť pozitívne definitná)	Nepotrebná
Množstvo pozorovaní	Na odhad kovariancií stačí ca. 150 údajov (EWMA alebo klasický prípad)	Je potrebný väčší počet pozorovaní (500 a viac) na zabezpečenie realizácie extrémnych hodnôt	Na odhad kovariancií stačí ca. 150 údajov (EWMA alebo klasický prípad)	Na odhad kovariancií stačí ca. 150 údajov (EWMA alebo klasický prípad)	Spravidla minimálne ca. 250 pozorovaní (ročné denné údaje)
Aktualizácia dát	Postačuje s určitým (aj mesačným ap.) časovým oneskorením	Postačuje s určitým (aj mesačným ap.) časovým oneskorením	Postačuje s určitým (aj mesačným ap.) časovým oneskorením	Postačuje s určitým (aj mesačným ap.) časovým oneskorením	Denná (absolútne závisí na najnovších dátach)
Riešenie nelinearity (opcie v portfóliu)	Nevhodná (aproximuje nepresne riziko iba lineárnym členom rozvoja – deltou)	Nevhodná (aproximuje riziko nepresne ako VCV metóda)	Vhodná (na nelinearite opčných inštrumentov nezávisí)	Presnejšie ako VCV metóda (aproximuje riziko po kvadratický člen rozvoja – deltou aj gamou)	Vhodná (na nelinearite opčných inštrumentov nezávisí)
Horizont predpovede	Predpovede sú statické (predpokladá ohodnotenie portfólia k dátumu predpovede). Možnosť konverzie.	Predpovede sú denné (alt. na krátke obdobie projekcie) s absenciou konverzie	Zvolený trhový horizont	Predpovede sú statické (denné). Možnosť konverzie závisí od predpokladov a ich splnení	Možno upravovať v závislosti na dostupnosti minulých pozorovaní

./.

[16] Problém meniacej sa volatility Hull a White (1998b) navrhli riešiť preškálovaním relatívnych výnosov (resp. výnosov rizikových faktorov) pomocou vzoru $\tilde{r}_{i[t]} = \hat{\sigma}_{i[T]} r_{i[t]} / \hat{\sigma}_{i[t-1]}$. Historický výnos $r_{i[t]}$ v čase *t* sa upraví o historickú predpoveď volatility pre tento deň $\hat{\sigma}_{i[t-1]}$ a okamžité trhové podmienky sa zachytia prostredníctvom odhadnutej hodnoty súčasnej volatility $\hat{\sigma}_{i[T]}$ v čase *T*, keď sa aplikuje metóda historickej simulácie. O rozdelení preškálovaných výnosov $\tilde{r}_{i[t]}$ sa ďalej predpokladá nemennosť v čase.

ATRIBÚT	VCV metóda (delta-normálna metóda)	Syntetické EVT modely	Simulácia Monte Carlo	Delta-gama metóda	Historická simulácia
Presnosť predpovedí	Závisí od intenzity a platnosti predpokladov	Závisí od splnenia predpokladov a voľby prahovej konštanty	Závisí od predpokladov o volatilite a rozdelení rizikových faktorov a na počte simulácií	Vo všeobecnejšie presnejšie ako pri VCV metóde, ale závisia od splnenia predpokladov	Závisí od stability podmienok na finančných trhoch
Komputačná náročnosť	Vyžaduje iba násobenie matíc aktuálnym zložením portfólia. Relatívne rýchla pre väčšinu portfólií	Nízka. Vyžaduje zostavenie predpovede volatility násobením matíc súčasným zložením portfólia	Výpočtovo náročná (všetky inštrumenty musia byť prehodnotené pre každý cenový scenár)	Po zadefinovaní vzájomných vzťahov nízka. Vyžaduje predpovede volatility násobením matíc	Ľahko implementovateľné (ak sú nastavené oceňovacie funkcie)

Schéma č. 4 *Charakteristika základných prístupov* (Zdroj: upravené podľa Blanco (1998) a vlastné spracovanie)

V neposlednom rade voľbu modelu ovplyvňujú požiadavky kladené na model value at risk regulátorom, ktorému inštitúcia podlieha. Tieto podmienky boli diskutované v predošlom príspevku Boďa (2006).

Je evidentné, že výber vhodného modelu pre kalkuláciu value at risk (či už pre vlastné alebo pre regulačné účely) nie je jednoduchou záležitosťou: pokiaľ je, samozrejme, cieľom správne mapovať okamžitú situáciu na finančných trhoch a zostavovať nevychýlené predpovede limitnej straty na danej hladine spoľahlivosti a pre daný horizont držby portfólia. Azda treba ešte zdôrazniť pre voľbu modelu dve okolnosti.

Everything should be made as simple as possible, but not simpler. (Albert Einstein)
All models are wrong but some are useful. (George E. P. Box)

Ďalší príspevok sa bude týkať niektorých ďalších prístupov k modelovaniu value at risk a spätného overovania kvality používaného modelu.

Bibliografia

ANDERSEN, Torben G., BOLLERSLEV, Tim, CHRISTOFFERSEN, Peter F., DIEBOLD, Francis X. 2005. *Volatility Forecasting.* In: *Handbook of Economic Forecasting. Year 2006.* Ed. G. Elliott, C. W. J. Granger, A. Timmermann. Amsterdam: Elsevier. ISBN 0-444-51395-7. [Pripravované.]

ARLT, Josef, ARLTOVÁ, Markéta 2003. *Finanční časové rady. Vlastnosti, metody modelování, příklady a aplikace.* Praha: Grada Publishing 2003. 220 s. ISBN 80-247-0330-0.

BAXTER, Martin, RENNIE, Andrew 1996. *Financial calculus. An introduction to derivative pricing.* Cambridge: Cambridge University Press 1999. 234 s. ISBN 0 521 55289 3.

BLANCO, Carlos 1998. *Value at risk for energy: Is VaR useful to manage energy price risk?* In: *Commodities Now*. 1998, č. 12 (december). 11 s.

BOĎA, Martin 2006. *Value at risk I. Value at risk ako miera rizika, alternatívy, nedostatky a regulačný aspekt.* In: *Forum Statisticum Slovacum.* 2006, č. 4, roč. 2. S. 15-24.

BRITTEN-JONES, Mark, SCHAEFER, Stephen M. 1999. *Non-Linear Value-at-Risk.* In: *European Finance Review*. 1999, č. 2, roč. 2 S. 161-187.

CASSIDY, Colleen, GIZYCKI, Marianne 1997. *Measuring Traded Market Risk: Value-at-risk and Backtesting Techniques* . In: *Research Discussion Papers of Reserve Bank of Australia.* 1997, č. 9708. 37 s.

CIPRA, Tomáš 2002. *Kapitálová přiměřenost ve financích a solventnost v pojišťovnictví.* Praha: Ekopress 2002. 272 s. ISBN 80-86119-54-8.

DUFFIE, Darrell, PAN, Jun 1997. *An Overview of Value at Risk*. In: *Journal of Derivatives.* 1997, č. 1, roč. 4. S. 7-49.

GIACOMINI, Enzo, HÄRDLE, Wolfgang 2005. *Value-at-Risk Calculations with Time Varying Copulae.* In: *SFB 649 Economic Risk.* Discussion Paper 2005-004. Berlin: SFB 649 2005.

HARDY, Mary 2003. *Investment guarantees: modeling and risk management for equity-linked life insurance.* New Jersey: Wiley 2003. 286 s. ISBN 0-471-39290-1.

HARMANTZIS, Fotios, MIAO, Linyan, CHIEN Yifan 2005. *Empirical Study of Value-at-Risk and Expected Shortfall Models with Heavy Tails.* [Acrobat ® pdf online]. Hoboken [USA]: Stevens Institute of Technology 2005. [Cit. 29. 09. 2006]. Dostupné na World Wide Web: <http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID788624_code389663.pdf>.

HÄRDLE, Wolfgang, KLEINOW, Torsten, STAHL, Gerhard 2002. *Applied quantitative finance: theory and computational tools.* Berlin: Springer Verlag 2002. 401 s. ISBN 3-540-43460-7.

HÄRDLE, Wolfgang, HLÁVKA, Zdeněk, STAHL, Gerhard 2006. *On the appropiateness of inappropriate VaR models.* In: *SFB 649 Economic Risk.* Discussion Paper 2006-003. Berlin: Physica 2006. ISBN 0002-6018.

HULL, John C. 2000. *Options, Futures & Other Derivatives.* Fourth Edition. Upper Saddle River [USA]: Prentice Hall 2000. 698 s. ISBN 0-13-015822-4.

HULL, John, WHITE, Alan 1998a. *Value at Risk When Daily Changes in Market Variables Are Not Normally Distributed.* In: *Journal of Derivatives.* 1998, č. 1 (jar), roč. 5. S. 9-19.

HULL, John, WHITE, Alan 1998b. *IncorporatingVolatility Updating into the Historical Simulation Method for Value at Risk.* In: *Journal of Risk.* 1998, č. 1, roč. 1. S. 5-19.

LINSMEIER, Thomas, PEARSON, Neil 2000. *Value at Risk*. In: *Financial Analyst Journal*. 2000, č. 2 (marec/ apríl), roč. 56. S. 47-67.

LOPEZ, Jose A., WALTER, Christian A 2000. *Evaluating Covariance Matrix Forecasts in a Value-at-Risk Framework.* In: *Working Papers in Applied Economic Theory.* Federal Reserve Bank of San Francisco. 2000, č. 2000-21. 50 s.

MANGANELLI, Simone, ENGLE, Robert F. 2001. *Value at risk models in finance.* In: *European Central Bank Working Paper Series.* 2001, č. 75 (august 2001). 40 s. ISSN 15-61-0810.

NCNEIL, Alexander J. 1999. *Extreme Value Theory for Risk Managers*. In: *Internal Modelling and CAD II.* Risk Books. 1999. S. 93-113.

MELICHERČÍK, Igor, OLŠAROVÁ, Ladislava 2005. *Kapitoly z finančnej matematiky 2.* Zvolen: Sabovci 2005. 122 s. ISBN 80-89029-93-0.

MINA, Jorge, XIAO, Jerry Yi 2001. *Return to RiskMetrics: The Evolution of a Standard*. New York: RiskMetrics Group 2001. 110 s.

MINA, Jorge, ULMER, Andrew 1999. *Delta-Gamma Four Ways.* New York: RiskMetrics Group 1999. 12 s.

RAO, C. Radharkishna 1978. *Lineární metody statistické indukce a jejich aplikace.* Prel. Josef Machek. Z am. or. *Linear Statistical Inference and Its Applications* (Wiley 1973). Praha: Academia 1978. 668 s. Bez ISBN.

RANK, Jörn 2000. *Copulas in Financial Risk Management.* [Diploma thesis in Mathematical Finance.] [Acrobat ® pdf online] Oxford: University of Oxford, Department of Continuing Education. 2000. 40 s. [Cit. 14. 07. 2006]. Dostupné na World Wide Web: <http://www.gloriamundi.org/picsresources/jr.pdf >.

RiskMetrics™ - Technical Document. Fourth Edition (December 1996). New York: Morgan Guaranty Trust Company of New York / Reuters 1996. 284 s.

SINHA, Tapen, CHAMÚ, Francisco 2000. *Comparing Different Methods of Calculating Value at Risk.* [Acrobat ® pdf online]. Nottingham [UK]: Nottingham University Business School 2000. [Cit. 29. 09. 2006]. Dostupné na World Wide Web: <http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID706582_code58705.pdf>.

VRIES, Andreas de 2000. *The Value at Risk.* [Acrobat ® pdf online]. Düsseldorf: Fachhochshule Sudwestfalen 2000. [Cit. 14. 07. 2006]. Dostupné na World Wide Web: <http://www.gloriamundi.org/download.asp?ResourceID=453057827>.

ZMEŠKAL, Zdeněk et al. 2004. *Finanční modely.* Praha: Ekopress 2004. 236 s. ISBN 80-86119-87-4.

Analýza časových řad ve výuce studentů oboru Cestovní ruch

Jana Borůvková

Abstract: This article is concerned with a teaching of the statistical methods applied on the field of a tourist trade. A data from the travel movement are employed not only to expose the students to the statistical methods used in the time series processing but also to demonstrate the interpretation of results and their practical exploitation. MS Excel Software is employed in this educational project.

Key words: Education; time series; trend; prognosis; seasonal influence; MS Excel

1. Výuka statistiky na oboru Cestovní ruch

Při výuce statistiky na oboru Cestovní ruch využíváme z metodických důvodů data z oblasti cestovního ruchu. Studenti při studiu statistiky nám i sobě často kladou otázku: „Na co mi tohle bude?“. Tím, že používáme data ze studovaného oboru, částečně této otázce předcházíme. Navíc studenti jsou díky této volbě dat vedeni k tomu, aby si na základě zpracovaných dat vytvořili konkrétní představu o významu vypočítané hodnoty. Cílem výuky tedy není pouze data statisticky zpracovat, ale též naučit se zpracovaná data interpretovat.

Vzhledem k tomu, že VŠPJ je nově vzniklá vysoká škola, nemají studenti a učitelé zatím k dispozici žádný statistický software, jehož pořízení je finančně velmi nákladné. Ke zpracování dat je využíván pouze Excel. Je zřejmé, že toto řešení má své nevýhody, ale má také jednu výhodu oproti použití statistického softwaru. Studenti se během studia statistiky naučí velmi efektivně pracovat s programem MS Excel, neboť program v tu chvíli pro ně není cílem studia, ale pracovním nástrojem. Velmi rychle se ukáže, že efektivní využití tohoto softwaru usnadňuje statistické výpočty. Dokonce umožňuje získat některé správné výsledky i v případě, že student nezná příslušné vzorce.

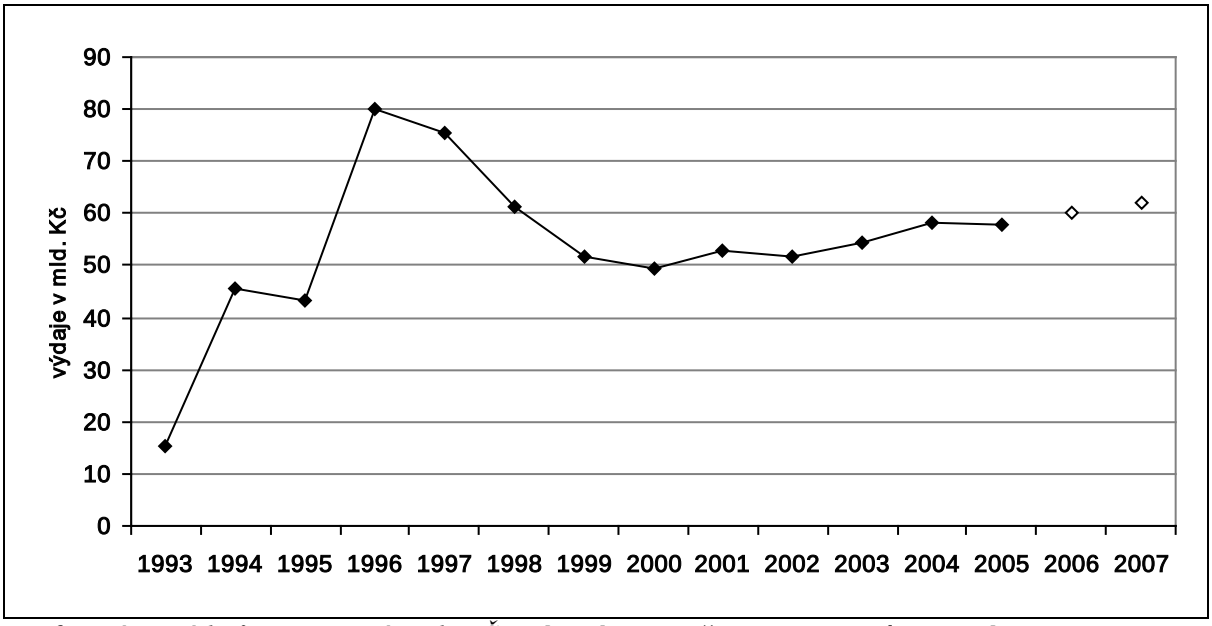
Necelá polovina jednoho semestru je věnována studiu a analýze časových řad. Tuto problematiku je samozřejmě možné vysvětlit na jakékoli časové řadě, jak již však bylo řečeno, z metodických důvodů volíme data z oblasti cestovního ruchu, a to například údaje popisující vývoj příjmů z cestovního ruchu a výdajů na cestovní ruch, které jsou poskytované průběžně Českou národní bankou[1], případně počty návštěvníků, kteří navštívili ČR v letech 1993–2004[2], uváděné ve statistických ročenkách.

2. Popis trendové složky časové řady

Popis vývoje analyzované řady je jeden z nejdůležitějších úkolů analýzy časových řad. Naším cílem je popsat časovou řadu (a zejména její poslední fázi) pomocí jednoduchých matematických vztahů, tzv. trendových funkcí. Vzhledem k tomu, že časová řada může popisovat i velmi dlouhé období, je zřejmé, že na vývoj časové řady v různých obdobích mohou mít vliv různé faktory. Proto není výhodné hledat funkci popisující celé období, neboť ta by byla velmi komplikovaná a také nepřesná. V konkrétních případech je lepší rozdělit časovou řadu na kratší etapy, které je možné popsat skutečně jednoduchými matematickými funkcemi (např. lineární, kvadratickou, polynomickou atd.).

[1] http://www.cnb.cz/www.cnb.cz/cz/statistika/platebni_bilance_stat/platebni_bilance/BOP_CS.XLS
[2] Statistické ročenky ČR pro roky 1997–2004

Tyto etapy určíme na základě grafického znázornění časové řady. Z možností, které poskytuje Excel při tvorbě grafů, je v tomto případě nejlepší volbou spojnicový graf, který kromě příslušných hodnot ukazatele (body) zachycuje též jeho vývoj mezi jednotlivými obdobími (spojnice). Časová řada vývoj výdajů na cestovní ruch v letech 1993 až 2005 a též prognóza vývoje na roky 2006 a 2007 je zobrazena v grafu 1. Na základě rozboru grafu 1 rozdělíme vývoj popsaný touto časovou řadu do tří etap. V první etapě od roku 1993 do roku 1996 hodnoty prudce narůstají. Ve druhé etapě v letech 1997–2000 dochází nejen k zastavení růstu, ale je zde patrný i výrazný pokles hodnot. Ve třetí etapě od roku 2000 se jedná o mírný nárůst výdajů, který v podstatě kopíruje vývoj inflace v ČR.



Graf 1. Vývoj výdajů na cestovní ruch v ČR v letech 1993 až 2005 a prognóza na roky 2006 a 2007

Nyní se snažíme vytvořit modely pro jednotlivé etapy vývoje. Jinými slovy hledáme jednoduché matematické vztahy, které by popisovaly vývoj v jednotlivých etapách. První a zejména třetí etapu je možné popsat nejjednodušším možným vztahem – lineární funkcí, kterou je možné obecně zapsat $y = ax + b$.

Nalezení koeficientů *a* a *b* je možné v Excelu několika způsoby.

1. Přidáním spojnice trendu do grafu příslušné etapy. Na kartě „Typ“ zvolíme lineární trend a na kartě „Možnosti“ zatrhneme volby Zobrazit rovnici regrese a Zobrazit hodnotu spolehlivosti R.
2. Použitím analytického nástroje Regrese. Do vstupní oblasti Y zadáme zkoumaná data, do vstupní oblasti X čísla 1 až *n* (počet zkoumaných dat). V zobrazené tabulce najdeme (mimo jiné) Koeficienty i Hodnotu spolehlivosti R.
3. Koeficient *a* je možné získat statistickou funkcí LINREGRESE a koeficient *b* funkcí LINTREND.

Po nalezení koeficientů *a* a *b* je možné první etapu popsat modelem $y = 19{,}213x - 1{,}914$, který vystihuje 87 % variability trendové složky a třetí etapu modelem $y = 1{,}763x + 47{,}883$, který vystihuje 86,8 % variability trendové složky.

Použití lineárního modelu pro druhou etapu vývoje výdajů není příliš vhodné. V tomto případě je vhodnější použít polynomický model a druhou etapu popsat polynomem 2. stupně $y = ax^2 + bx + c$.

Nalezení koeficientů *a*, *b* a *c* je v Excelu možné též více způsoby.

1. Přidáním spojnice trendu do grafu příslušné etapy. Na kartě „Typ“ zvolíme polynomický trend, stupeň 2 a na kartě „Možnosti“ zatrhneme volby Zobrazit rovnici regrese a Zobrazit hodnotu spolehlivosti R.

2. Použitím analytického nástroje Regrese. Do vstupní oblasti Y zadáme zkoumaná data, do vstupní oblasti X čísla 1 až *n* (počet zkoumaných dat). V zobrazené tabulce najdeme (mimo jiné) Koeficienty i Hodnotu spolehlivosti R.

Druhou etapu vystihuje model $y = 2{,}997x^2 - 23{,}761x + 96{,}362$, který vystihuje 99,9 % variability trendové složky.

Dále studenty necháme samostatně zpracovat data, která popisují vývoj příjmů z cestovního ruchu v letech 1993 až 2005. Výsledky jsou uvedeny v [1].

3. Prognóza vývoje na následující období

Hlavním důvodem vytvoření matematického modelu poslední etapy časové řady je jeho využití pro prognózu dalšího vývoje. V našem případě se jedná o prognózu vývoje příjmů i výdajů spojených s cestovním ruchem na roky 2006 a 2007.

Model poslední etapy časové řady *příjmy* byl vytvořen z hodnot v letech 2002 až 2005 a předpokládal lineární trend časové řady. Získaná přímka má rovnici $y = 5089{,}8x + 90970$. Dosadíme-li do rovnice přímky za x postupně čísla 0 až 3, získáme vypočítané hodnoty časové řady příjmy z cestovního ruchu za sledované období. Tyto vypočítané příjmy se mohou od skutečných hodnot lišit.

Dosadíme-li do rovnice za x číslo 4 (resp. 5), dostaneme prognózu na další dvě období. V roce 2006 lze tedy očekávat příjem z cestovního ruchu 111,3 mld. Kč a v roce 2007 116,4 mld. Kč.

Odhad *výdajů* na cestovní ruch na roky 2006 a 2007 necháme studenty provést samostatně. Vzhledem k tomu, že model posledního úseku časové řady výdajů byl vytvořen z hodnot v letech 2000 až 2005, je nutné za x do rovnice popisující vývoj v této etapě dosadit číslo 6 (resp. 7). Výsledky jsou rovněž uvedeny v [1].

Slabým místem těchto prognóz je předpoklad neměnnosti dosavadních vývojových tendencí prognózovaného jevu. Model jednak nepopisuje 100 % variability trendové složky a jednak nepopisuje nové vlivy, které se ve vývoji mohou objevit. Proto je nutné studentům zdůraznit, že i prognóza na jedno období je velmi nejistá a nemá tedy smysl činit prognózy na více období dopředu.

4. Vliv sezóny na časovou řadu

Vzhledem k tomu, že ve většině časových řad s periodicitou zjišťování kratší než jeden rok (velmi často s periodicitou čtvrtletní nebo měsíční) existují sezónní vlivy, lze ke studiu vlivu sezóny na časovou řadu použít časové řady příjmy a výdaje v oblasti cestovního ruchu, neboť ČNB poskytuje informace o obou časových řadách čtvrtletně. Pro analýzu sezónního průběhu obou časových řad ve výuce byla použita data od 1. čtvrtletí 1999 po 4. čtvrtletí 2005.

Prvním úkolem je dokázat, že sezónní výkyvy jsou skutečně statisticky významné. V případě časových řad příjmy a výdaje v oblasti cestovního ruchu se jedná o velmi jednoduché situace, ve kterých je možné existenci sezónnosti odhalit intuitivně.

Dalším úkolem je kvantifikace sezónních výkyvů (např. pomocí průměrných sezónních indexů), jejichž znalost je možné využít jak pro zpřesnění prognózy na následující období, tak i k tzv. sezónnímu očišťování, jehož úkolem je vyloučit sezónní složku z analyzované řady.

Pro výpočet sezónních indexů máme k dispozici hodnoty za 7 roků (7 ucelených sezónních period, 4 hodnoty pro každý rok). Pro každý rok vypočítáme ze 4 hodnot zjištěných v jednotlivých čtvrtletích příslušného roku průměrnou hodnotu. Touto průměrnou hodnotou vydělíme všechny čtyři hodnoty zjištěné v jednotlivých čtvrtletích a získáme odpovídající sezónní indexy. Z jednotlivých sezónních indexů vypočítáme průměrný sezónní index pro jednotlivá čtvrtletí. K výpočtu tohoto průměru je vhodnější použít geometrický průměr.

Na tomto místě je nutné zdůraznit, že tento výpočet je svým způsobem problematický, neboť předpokládá nepřítomnost trendů v jednotlivých periodách. Jedná se však o způsob nejjednodušší, a proto jej při výuce používáme.

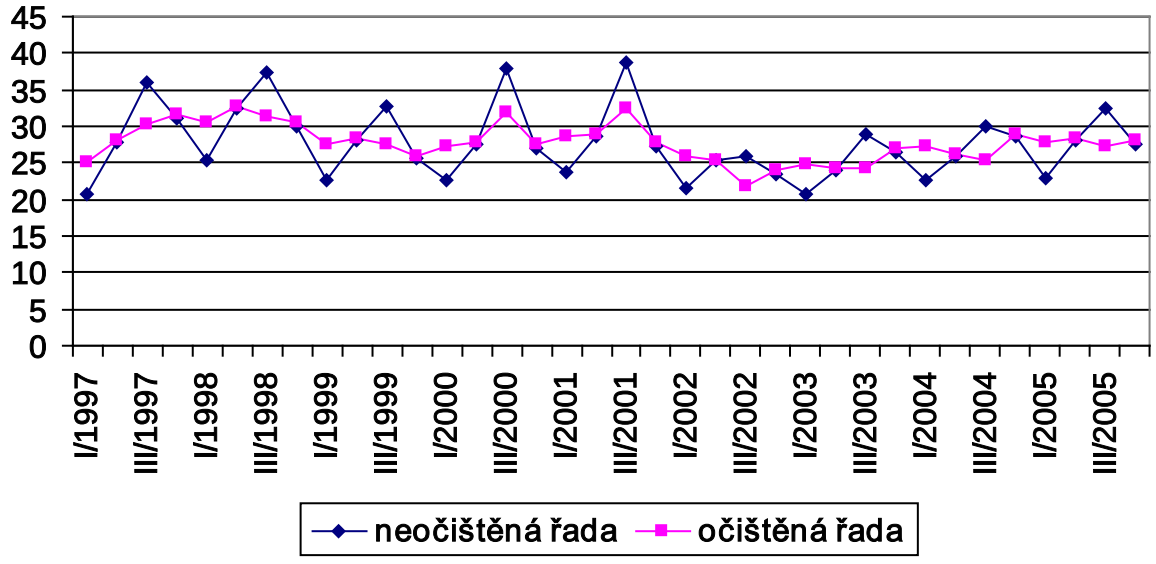
Výpočet průměrných sezónních indexů pro jednotlivá čtvrtletí je uveden v tabulce 1.

Tabulka 1. Výpočet průměrných sezónních indexů

	I.	II.	III.	IV.	průměr
1999	22 766,0	28 143,0	32 723,0	25 510,0	27 285,5
2000	22 525,0	27 472,0	38 011,0	27 063,0	28 767,8
2001	23 660,0	28 523,0	38 648,0	27 302,0	29 533,3
2002	21 584,0	25 268,0	25 893,4	23 543,8	24 072,3
2003	20 706,7	24 048,0	28 989,1	26 566,3	25 077,5
2004	22 655,8	25 939,6	30 087,4	28 549,0	26 808,0
2005	23 000,8	28 063,7	32 372,5	27 511,0	27 737,0

	individuální sezónní indexy			
1999	0,834	1,031	1,199	0,935
2000	0,783	0,955	1,321	0,941
2001	0,801	0,966	1,309	0,924
2002	0,897	1,050	1,076	0,978
2003	0,826	0,959	1,156	1,059
2004	0,845	0,968	1,122	1,065
2005	0,829	1,012	1,167	0,992
průměr	0,830	0,991	1,190	0,983

Pro možnost porovnávat po sobě jdoucí údaje v časové řadě uvnitř roku i tehdy, jsou-li aktuálně ovlivněny sezónností, je nutné údaje časové řady sezónně očistit. Jde o modelové rozdělení časové řady na složku sezónní, trendovou a náhodnou a sezónní složku z časové řady odstranit. Nejjednodušší způsob, jak toho docílit je vydělit hodnoty původní časové řady příslušným průměrným sezónním indexem. Výsledkem je sezónně očištěná časová řada. Porovnání sezónně očištěné i neočištěné řady výdaje na cestovní ruch je vidět v grafu 2.



Graf 2. Porovnání sezónně očištěné a neočištěné řady

5. Závěr

Zkoumáním konkrétních časových řad z oblasti cestovního ruchu v ČR je možné nejen procvičit statistické výpočty umožňující analýzu časových řad pomocí Excelu, ale též seznámit studenty s interpretací výpočtů a jejich případným využitím v praxi.

Studenti se naučí vytvořit matematický model časové řady a odhadnout pomocí něj vývoj v nejbližších obdobích. Zde je ovšem nutné zdůraznit, že lze modelovat pouze dvě složky časových řad – trendovou a sezónní. Třetí náhodnou složku není možné statistickými metodami zachytit. Bohužel však v cestovním ruchu tato složka může mít v některých letech velký vliv a učiněnou předpověď může naprosto znehodnotit.

Studenti na základě svých zkušeností a znalostí ví, že na časové řady výdaje na cestovní ruch a příjmy z cestovního ruchu má vliv sezóna, tzn. tyto řady obsahují sezónní složku. Tyto intuitivně vnímané skutečnosti se studenti naučí popsat pomocí sezónních indexů, případně i časovou řadu sezónně očistit.

Pro výše popsané výpočty ve většině případů stačí běžná uživatelská znalost Excelu (tvorba grafů, výpočty pomocí vzorců, absolutní a relativní adresy, běžné matematické funkce – součet, průměr, maximum, minimum). Výjimku tvoří pouze pasáž věnovaná vytvoření regresního modelu. Zde je nutné znát také některé statistické funkce (Linregrese, Lintrend), práci se spojnicí trendu, případně práci s analytickým nástrojem Regrese.

Při výuce metod používaných k analýze časových řad na oboru Cestovní ruch se ukazuje, že přínos kompletního zpracování několika časových řad, které souvisí s problematikou cestovního ruchu, je nejen v oblasti statistického zpracování dat, ale též v oblasti praktického využití programu MS Excel.

6. Literatura

[1] Borůvková, J. 2006. *Příjmy a výdaje spojené s cestovním ruchem v České republice v letech 1993–2005.* FORUM STATISTICUM SLOVACUM. Ročník II, č. 4/2006. Str. 31–36. ISSN 1336-7420

[2] Hindls, R., Hronová, S., Seger, J. 2006. *Statistika pro ekonomy.* Professional Publishing Praha. 2006. 415 s. ISBN 80-86419-99-1

[3] Chajdiak, J. 2003. *Štatistika jednoducho*. STATIS Bratislava. 2003. 194 s. ISBN 80-85659-28

[4] Chajdiak, J. 2005. *Štatistické úlohy a ich riešenie v Exceli.* STATIS Bratislava. 2005. 262 s. ISBN 80-85659-39-5

[5] *Platební bilance ČR.* [Internet]. [cit. 25.8.2006]. Dostupné z http://www.cnb.cz/www.cnb.cz/cz/statistika/platebni_bilance_stat/platebni_bilance/BOP_CS.XLS

Adresa autora:
RNDr. Jana Borůvková, Ph.D.
Vysoká škola polytechnická Jihlava
Tolstého 16
58601 Jihlava
boruvkova@vspji.cz

Softvér na výpočet minimálnej detegovateľnej hodnoty v prípade konštantnej smerodajnej odchýlky

Peter Cisko1[1], Ivan Janiga[2], Ivan Garaj[3]

Abstract: The minimum detectable value obtained from a particular calibration is the smallest value of the net state variable which can be detected with a probability of $1-\beta$ as different from zero. The presentation of the software for the computation of the critical values and the minimum detectable value from the linear calibration function is given. The linear calibration model with constant standard deviation is concidered in the paper.

Key words: minimum detectable value, critical values, linear calibration model.

1. Úvod

Minimálna detegovateľná hodnota získaná s príslušnej kalibrácie je najmenšia hodnota redukovanej stavovej premennej, ktorá môže byť detegovaná s pravdepodobnosťou $1-\beta$ ako hodnota rôzna od nuly (pozri obrázok). V [1] sa predpokladala lineárna kalibračná funkcia s konštantnou smerodajnou odchýlkou $Y_{ij} = a + bx_i + \varepsilon_{ij}$, $i = 1, 2, ..., I$, $j = 1, 2, ..., J$ kde x_i je redukovaná stavová premenná v stave i. Náhodné premenné ε_{ij}, ktoré reprezentujú náhodnú zložku chyby vzorkovania, preparátu a merania, sú nezávislé, normálne rozdelené s nulovou strednou hodnotou a smerodajnou odchýlkou σ, t. j. $\varepsilon_{ij} \sim N(0, \sigma^2)$. Merania ozvovej premennej pre všetky referenčné stavy (I) a preparáty (J) Y_{ij} sú nezávislé náhodné premenné, ktoré majú normálne rozdelenie so strednou hodnotou $E(Y_{ij}) = a + bx_i$ a rozptylom $D(Y_{ij}) = \sigma^2$, ktorý nezávisí od hodnôt x_i, t. j. $Y_{ij} \sim N(a + bx_i, \sigma^2)$, $i = 1, 2, ..., I$, $j = 1, 2, ..., J$.

2. Odhad minimálnej detegovateľnej hodnoty

Odhad minimálnej detegovateľnej hodnoty sme dostali z odhadu kalibračnej priamky. Pre odhady *a*, *b* a σ^2 platí:

$$\hat{b} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(x_i - \bar{x})(\bar{y}_{ij} - \bar{y})}{s_{xx}}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{\sigma}^2 = \frac{1}{I \cdot J - 2}\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij} - \hat{a} - \hat{b}x_i)^2$$

kde $s_{xx} = J\sum_{i=1}^{I}(x_i - \bar{x})^2$.

Kritická hodnota ozvovej premennej sa vypočíta ako hornú hranicu jednostranného predikčného intervalu pre priemer K budúcich meraní:

[1] Ing. Peter Cisko, SjF STU v Bratislave, doktorand
[2] Doc. RNDr. Ivan Janiga, PhD., Katedra matematiky SjF STU, Nám. slobody 17, 812 31 Bratislava; Katedra aplikovanej matematiky, FPV UCM, Nám. J. Herdu 2, 917 01 Trnava, e-mail: ivan.janiga@stuba.sk
[3] RNDr. Ivan Garaj, PhD., Ústav informatizácie, automatizácie a matematiky, FCHPT STU, Radlinského 9, 812 37 Bratislava, tel.: +421-2-59325 297, e-mail: ivan.garaj@stuba.sk

$$y_c = \hat{a} + t_{1-\alpha}(v)\hat{\sigma}\sqrt{\frac{1}{K} + \frac{1}{I \cdot J} + \frac{\bar{x}^2}{s_{xx}}}$$

Pre zodpovedajúcu kritickú hodnotu redukovanej stavovej premennej platí vzťah

$$x_c = \frac{y_c - \hat{a}}{\hat{b}} = t_{1-\alpha}(v)\frac{\hat{\sigma}}{\hat{b}}\sqrt{\frac{1}{K} + \frac{1}{I \cdot J} + \frac{\bar{x}^2}{s_{xx}}}$$

a minimálna detegovateľná hodnota x_d je hodnota pre ktorú s pravdepodobnosťou $1-\beta$ platí

$$x_d = \delta\frac{\hat{\sigma}}{\hat{b}}\sqrt{\frac{1}{K} + \frac{1}{I \cdot J} + \frac{\bar{x}^2}{s_{xx}}}$$

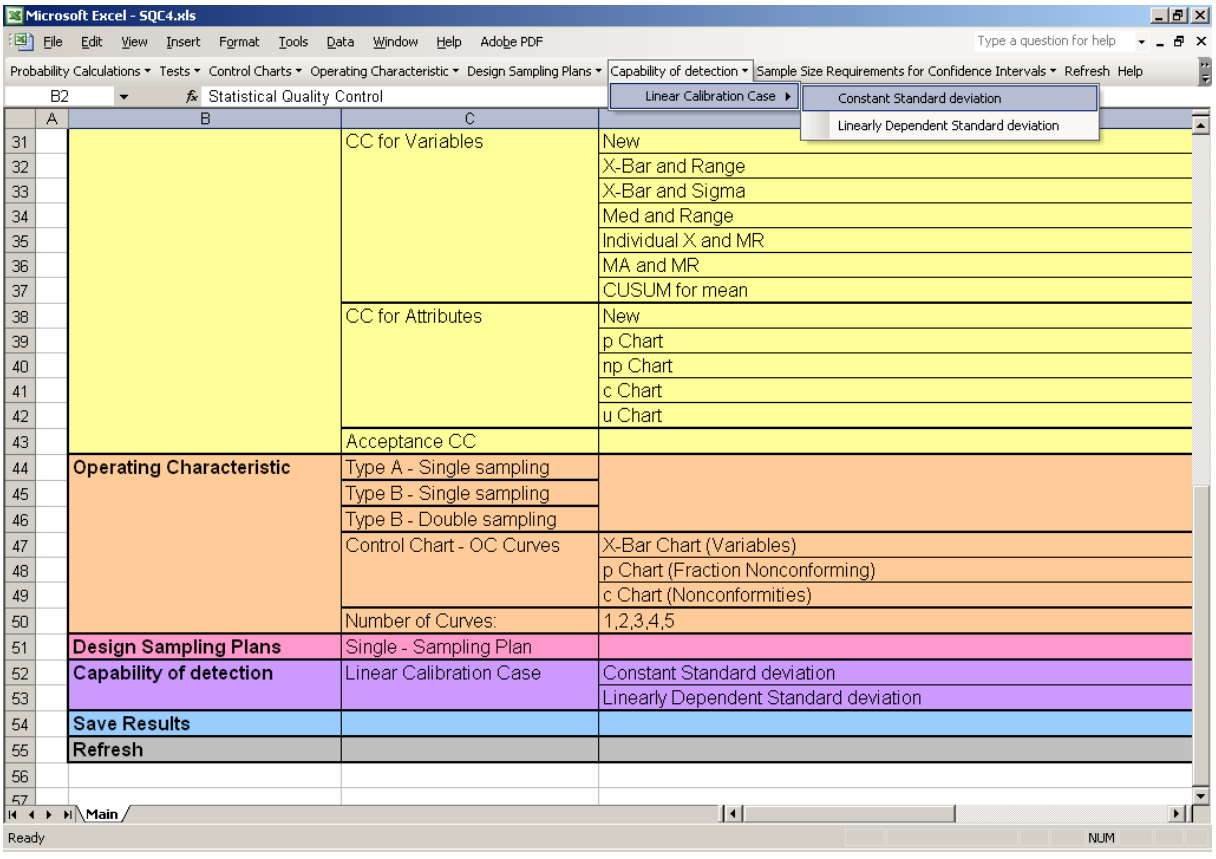
kde $\delta = \delta(v, \alpha, \beta)$ je hodnota parametra necentrality náhodnej premennej $T(v,\delta)$ s necetrálnym *t*-rozdelením so stupňami voľnosti $v = I \cdot J - 2$ a parametrom necentrality δ.

3. Softvér na výpočet minimálnej detegovateľnej hodnoty

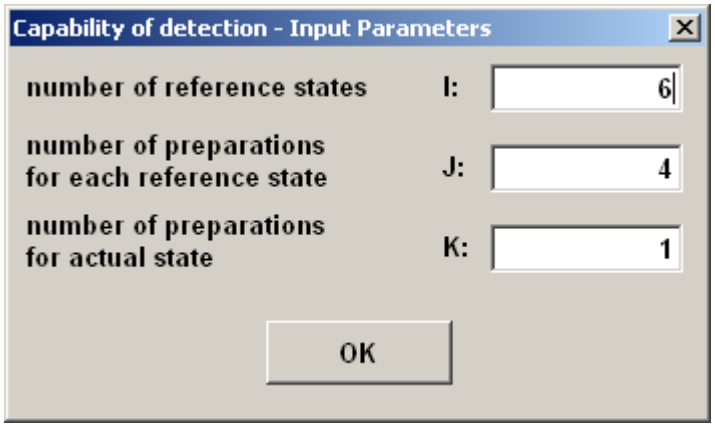
Výpočet minimálnej detegovateľnej hodnoty sa vykonáva v rámci softvéru SQC (Statistical Quality Control). Základné používateľské prostredie softvéru SQC je program MS Excel, v ktorom sú jednotlivé metódy naprogramované v jazyku Visual Basic. Celý softvér je obsiahnutý v jednom excelovskom súbore s príponou xls, ktorý obsahuje zdrojový kód. Softvér svojou štruktúrou umožňuje pridávanie ďalších programových modulov, načítavanie dát, ukladanie výsledkov, výber jednotlivých metód a zostrojovanie grafov.

Program sa spúšťa jednoduchým otvorením súboru v aplikácii MS Excel, kde je potrebné mať povolené makrá. Spustením programu sa do menu Excelu pridá nové menu, ktoré obsahuje jednotlivé štatistické metódy.

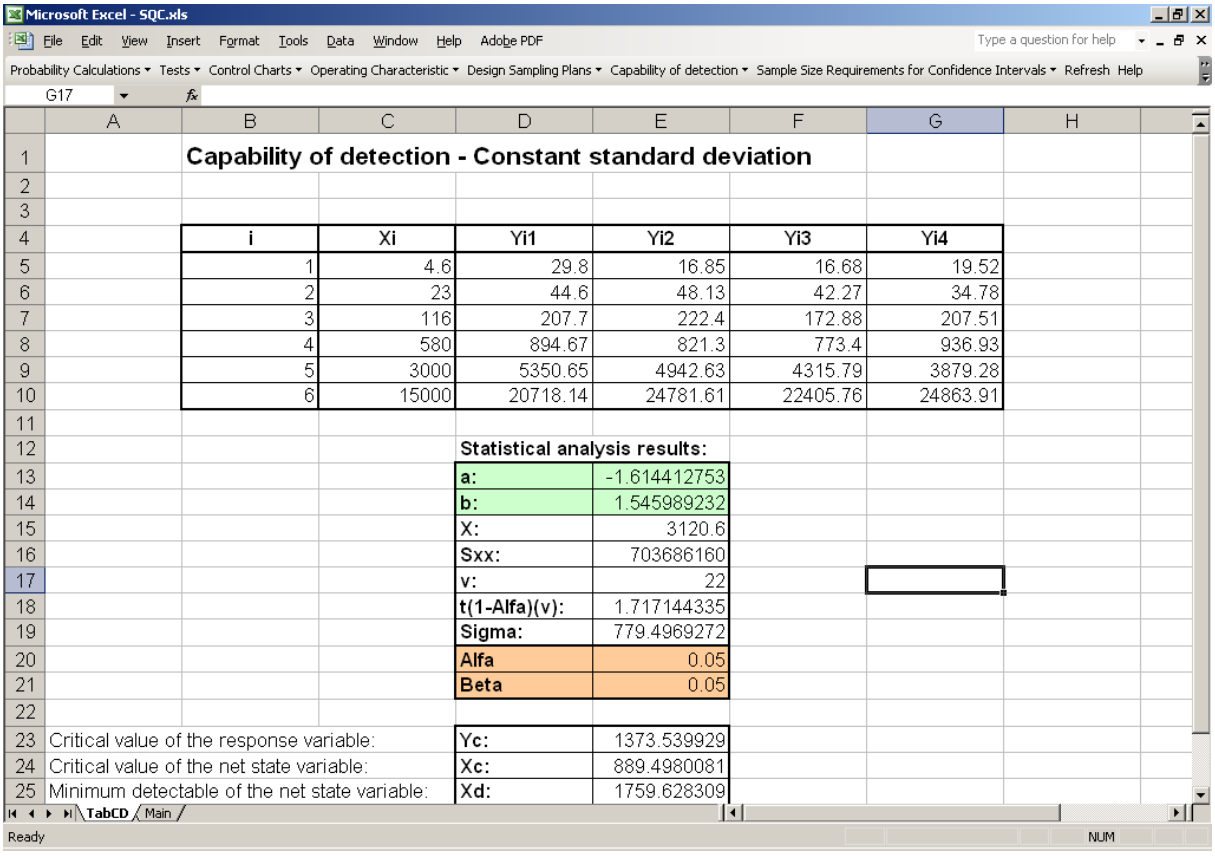
V základnom menu je možné zvoliť rôzne výpočtové metódy, kde jedná z metód je detekčná schopnosť v prípade konštantnej smerodajnej odchýlky.



Vstupnými hodnotami sú požiadavky používateľa na počet referenčných stavov, počet preparátov pre každý stav a počet preparátov pre aktuálny stav (hodnoty I, J, K).



Na základe týchto hodnôt je následne vytvorená vstupná tabuľka do ktorej používateľ zadá vstupné hodnoty. Po zadaní všetkých vstupných hodnôt nasleduje odhad kalibračnej funkcie, výpočet kritických hodnôt a minimálnej detegovateľnej hodnoty.



5. Literatúra

[1] JANIGA, I., GARAJ, I., CISKO, P. Estimation of Minimum Detectable Value by Using Linear Calibration with Constant Standard Deviation. In *FORUM METRICUM SLOVACUM*. ISBN 80-88946-23-9. 2004, Tom VIII., pp. 68-73.

[2] JANIGA, I., GARAJ, I. Necentrálne *t*-rozdelenie v odhadovaní minimálnej detegovateľnej hodnoty pomocou lineárnej kalibračnej funkcie s konštantnou smerodajnou odchýlkou. In *FORUM STATISTICUM SLOVACUM*. ISSN 1336-7420. 2006, roč. II, č. 1, s. 44-50.

[3] GARAJ, I., JANIGA, I. *Dvojstranné tolerančné medze normálnych rozdelení s neznámymi strednými hodnotami a s neznámym spoločným rozptylom.* Bratislava: STU, 2004. 218 s. ISBN 80-227-2019-4.

[4] GARAJ, I. – JANIGA, I. *Dvojstranné tolerančné medze pre neznámu strednú hodnotu a rozptyl normálneho rozdelenia.* Bratislava: STU, 2002. 147 s. ISBN 80-227-1779-7.

[5] GARAJ, I., JANIGA, I. *Jednostranné tolerančné medze normálneho rozdelenia s neznámou strednou hodnotou a rozptylom. One sided tolerance limits of normal distribution with unknown mean and variability.* Bratislava: STU, 2005. 214 s. ISBN 80-227-2218-9.

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

[7] TEREK, M., HRNČIAROVÁ, Ľ. *Analýza spôsobilosti procesu*. Vydavateľstvo EKONÓM, Ekonomická univerzita v Bratislave, 2001. 205 s. ISBN 80-225-1443-8.

[8] HRNČIAROVÁ, Ľ., TEREK, M. Analýza zoskupení bodov v regulačných diagramoch. In *FORUM STATISTICUM SLOVACUM*. ISSN 1336-7420, 2005, roč. I., č. 1, s. 56-61.

[9] TEREK, M.- HRNČIAROVÁ, Ľ- LIŠKOVÁ, I :Navrhovanie regulačných diagramov v štatistickej regulácii procesu In *Ekonomika a informatika*. ISSN 1336-3514, 2005, roč. III, č. 1, s. 126-137.

[10] TEREK, M.- HRNČIAROVÁ, Ľ.: Zisk zo stratifikácie. In *FORUM STATISTICUM SLOVACUM*. ISSN 1336-7420, 2006, roč. II., č. 1, s. 51-55.

[11] PALENČÁR, R., RUIZ, J.M., JANIGA, I., HORNÍKOVÁ, A. *Štatistické metódy v skúšobných a kalibračných laboratóriách*. Bratislava: Grafické štúdio Ing. Peter Juriga, 2001. 380 s. ISBN 80-968449-3-8.

[12] GROSS, P., KUREKOVA, E. Advanced Experiments Design for the Three-Torch Plasma Cutter Testing. In *Proceedings of the 4th International Conference MEASUREMENT 2005*. Bratislava: VEDA, 2005. ISBN 80-967402-8-8, p. 530-533.

[13] PALENČÁR, R., KUREKOVÁ, E., VDOLEČEK, F., HALAJ, M. *Systém riadenia merania*. Bratislava: Grafické štúdio Ing. Peter Juriga, 2001. 208 s. ISBN 80-968449-7-0

[14] HALAJ, M. A contribution to calibration of piezorezistive tactile matrix sensors. In *Strojnícky časopis*. ISSN 0039-2472, 2003, roč. 54, č. 4, s. 228-238.

Tento článok vznikol s podporou grantových projektov VEGA č. 1/3182/06 Zlepšovanie kvality produkcie strojárskych výrobkov pomocou moderných štatistických metód, VEGA č. 1/3584/06 Príprava, charakterizácia a osobitné vlastnosti heterocyklických a aromatických zlúčenín a VEGA č. 1/1247/04 Progresívne štatistické techniky a rozhodovanie v procese zlepšovania kvality

Vplyv internacionalizácie maloobchodu na Slovensku na nákupné správanie sa zákazníkov vybraných veľkoplošných predajní

Jana Fertaľová

Abstract: This article points out the current development of retail trade, intensive integration and concentration of business activities on the Slovak market. It is indicated by permeate of big supranational companies and this brings changes in the shopping behaviour of citizens.

The consumers prefer large shopping centres on the periphery of the town with their typical pleasant atmosphere, the opportunity to park free and their multifunction in front a specialized shop in its centre. This creates space for the maximalization of stay of the consumer in the shopping centre and increase in average expenses spent for purchase of goods and services connected with it.

A questionnaire was carried out in the **shopping centres Optima** and **Cassovia in Košice** and in the **hypermarkets Tesco** and **Hypernova in Prešov** during the months October to December in the year 2005, when 974 customers were questioned, 600 in Košice and 374 in Prešov.

Key words: shopping behaviour, retail trade, large-scale store units, life quality

1. Úvod

V druhej fáze (tj. fáza koncentrácie, prebiehajúca od konca 90-tych rokov 20. storočia) prevláda spontánna transformácia, spojená s internacionalizáciou maloobchodu a vstupom nadnárodných maloobchodných reťazcov na náš trh. Táto etapa koncentrácie a internacionalizácie priniesla podstatné zmeny na náš trh, objavujú sa postupne siete supermarketov, hypermarketov, diskontných predajní, hobbymarketov a iných špecializovaných veľkopredajní. Veľkoplošné predajne bývajú súčasťou veľkoformátovej maloobchodnej štruktúry nákupných centier, lokalizovaných nie len v periférnych lokalitách veľkých miest, ale začínajú sa objavovať aj v ich centrách.

Ak **atomizáciu** maloobchodnej siete (ako jej prvú vývojovú etapu) hodnotíme z časového hľadiska na území Českej a Slovenskej republiky, môžeme badať nasledujúce rozdiely. Na území Českej republiky dosiahla svoj vrchol v polovici 90. rokov 20. storočia a hneď prešla do druhej vývojovej etapy **koncentrácie** a **internacionalizácie** obchodných aktivít. Na Slovensku vplyvom odlišnej ekonomickej a politickej situácie, ktorá zbrzdila príchod zahraničných obchodných reťazcov, atomizácia maloobchodnej siete pokračovala ešte v druhej polovici 90. rokov 20. storočia. Veľmi dobre túto situáciu ilustruje napr. okamih uvedenia do prevádzky prvého veľkého hypermarketu, ktorým bol v Českej republike v roku 1996 *Globus* (Brno), na Slovensku *Tesco* o tri roky neskôr (Nitra). Tiež v prípade ďalších obchodných formátov a reťazcov bol pozorovateľný časový nesúlad. Napríklad rozvoj siete supermarketov *Billa* alebo *Delvita* na Slovensku bol v porovnaní s vývojom v Českej republike zhruba o tri až štyri roky oneskorený. Iným dokladom oneskoreného vývoja sú aktivity firmy *Hornbach*, ktorá v Čechách svoj prvý hobbymarket otvorila už koncom roku 1998, na Slovensku až v auguste v roku 2004. Naopak, malý časový rozdiel je medzi Slovenskom a Českom pri otváraní diskontov nemeckého reťazca Lidl. V Čechách to bolo v júni, roku 2003 a na Slovensku v septembri 2004 (Szczyrba, 1998a, Fertaľová – Szczyrba, 2006, Fertaľová, 2005a).

Ako bolo vyššie spomenuté, v roku 1996 vstupuje na slovenský trh odvtedy najsilnejšia nadnárodná spoločnosť Tesco Stores SR, a.s., ktorá si od toho roku permanentne udržiava vedúcu pozíciu v TOP 50 obchodných spoločností na Slovensku (viď tabuľka 1). Nasledujúca tabuľka dokumentuje taktiež charakter prehlbujúcej sa internacionalizácie a koncentrácie

slovenského maloobchodného prostredia. Zatiaľ čo pred niekoľkými rokmi prevládali domáce firmy, dnes je situácia odlišná.

Tabuľka 1. Vývoj štruktúry TOP 10 obchodných spoločností na Slovensku podľa obratu v mld. Sk za roky 1998 a 2005

TOP 1998				TOP 2005		
por.	spoločnosť	obrat v mld.Sk	počet prevádzok	spoločnosť	obrat v mld.Sk	počet prevádzok
1.	Tesco Stores SR, a.s. *	4,50	7	Tesco Stores SR, a.s. *	22,00	37
2.	Interkontakt Group *	3,93	42	Metro Cash&Carry Slovakia, s.r.o.*	16,20	5
3.	Prima Zdroj Holding a.s.	3,62	59	Billa s.r.o. *	10,50	79
4.	Smoker	2,30	4	Kaufland SK v.o.s. *	8,50	26
5.	Opal-Fytos Group	2,05	20	Ahold Retail Slovakia k.s. *	7,00	24
6.	Euroholding Verex	1,67	18	Carrefour Slovensko a.s. *	5,50	4
7.	Kon-rad s.r.o.	1,60	2	CBA SK, s.r.o.	4,50	196
8.	Essex a.s.	1,60	31	bauMax SR, s.r.o. *	3,80	11
9.	M-Market a.s.	1,51	78	BARCZI s.r.o.	3,60	75
10.	Jednota SD Krupina	1,45	90	Lidl SR. *	3,40	56
1. – 10.		24,23 mld. Sk		1. – 10.	85,0 mld. Sk	

Zdroj: Moderní obchod, 5/1997 – 5/2006
Vysvetlivky: * zahraničná majetková účasť

2. Nákupné správanie sa

Medzinárodné obchodné spoločnosti, ktoré tvoria väčšinu z prvej skupiny najúspešnejších TOP spoločností na Slovensku, sa etablovali okrem hlavného mesta prakticky vo všetkých krajských a okresných mestách, nevynímajúc Košice a Prešov. Výskum nákupného správania sa, sa vo východoslovenských mestách Košice a Prešov realizoval za účelom identifikovania zmien v nákupnom správaní sa spotrebiteľa po penetrácii nadnárodných obchodných spoločností na slovenský trh koncom 90-tych rokov 20. storočia a následnej výstavby veľkoplošných predajní na periférіách týchto miest. V **nákupných centrách Optima** a **Cassovia v Košiciach** a v **hypermarketoch Tesco** a **Hypernova v Prešove** prebiehal dotazníkový prieskum v mesiacoch október až december v roku 2005, kedy bolo oslovených 974 zákazníkov, 600 v Košiciach a 374 v Prešove.

Dotazníkovým výskumom bola okrem iného zisťovaná frekvencia a dôvod návštevy zákazníkov týchto veľkoplošných formátov a časová dostupnosť spotrebiteľov. Osobitne bolo niekoľko otázok venovaných aj preferencii veľkoplošných predajní lokalizovaných na periférii miest pred nákupmi a trávením voľného času v predajniach situovaných priamo v centrách miest Prešova a Košíc.

Ako vidíme aj v nasledujúcej tabuľke, na otázku:

- **Ako často navštevujete toto nákupné centrum, resp. hypermarket?**

odpovedali nasledovne. V prešovských hypermarketoch nakupuje minimálne raz týždenne viac ako 63% respondentov, v košických nákupných centrách je to takmer 50%.

Tabuľka 2. Frekvencia nakupovania zákazníkov vo vybraných veľkoplošných predajniach Prešova a Košíc

mesto		počet zákazníkov	% podiel	∑ (v%)
Prešov	pravidelne viackrát v týždni	137	36,6	36,7
	najviac jedenkrát v týždni	100	26,7	63,5
	najviac jedenkrát za dva týždne	35	9,4	72,9
	najviac jedenkrát v mesiaci	27	7,2	80,2
	nepravidelne/občas	64	17,1	97,3
	som tu prvý krát	7	1,9	99,2
	chýbajúci údaj	4	1,1	100,0
	SPOLU	374	100,0	
Košice	pravidelne viackrát v týždni	153	25,5	25,6
	najviac jedenkrát v týždni	137	22,8	48,5
	najviac jedenkrát za dva týždne	90	15,0	63,5
	najviac jedenkrát v mesiaci	61	10,2	73,7
	nepravidelne/občas	123	20,5	94,3
	som tu prvý krát	26	4,3	98,7
	chýbajúci údaj	10	1,6	100,0
	SPOLU	600	100,0	

Zdroj: vlastný výskum realizovaný v mesiacoch október – december 2005

Pri porovnaní frekvencie nakupovania zákazníkov v Prešove a v Košiciach sme zistili nasledovné:

- viac ako 94% košických a viac ako 97% prešovských zákazníkov navštívilo predmetnú veľkopredajňu v ktorej boli oslovení minimálne raz, teda má osobnú skúsenosť s nakupovaním v takomto formáte
- v prešovských hypermarketoch nakupuje pravidelne viackrát v týždni takmer 37% a vo vybraných predajniach v Košiciach 25,6%.

Tento rozdiel je spôsobený tým, že:

- v Košiciach je dvojnásobný počet veľkoplošných predajní a tak dochádza k väčšiemu rozptylu zákazníkov medzi ne
- prešovské hypermarkety sú svojou lokalizáciou lepšie dostupné väčšiemu počtu peších zákazníkov a preto sa tu realizuje viac denných nákupov
- prešovský hypermarket Tesco je otvorený 24 hodín denne na rozdiel od košických nákupných centier

Dôležitým ukazovateľom hodnotenia nákupného správania sa a jeho zmien po internacionalizácii slovenského trhu je čas strávený zákazníkmi v nákupných centrách, hypermarketoch a iných veľkoplošných predajniach. Na otázku:

- **Koľko času (okrem príchodu a odchodu) strávite zvyčajne v nákupnom centre, resp. hypermarkete/v centre mesta?**

odpovedali zákazníci nasledovne:

Tabuľka 3. Čas strávený v nákupnom centre/hypermarkete

mesto		počet zákazníkov	podiel v %
Prešov	menej ako 30 minút	91	24,3
	30 až 60 minút	204	54,5
	60 až 120 minút	66	17,6
	viac ako 120 minút	11	2,9
	chýbajúci údaj	2	0,6
SPOLU		374	100,0

Košice		menej ako 30 minút	68	11,3
		30 až 60 minút	236	39,3
		60 až 120 minút	211	35,2
		viac ako 120 minút	73	12,2
		chýbajúci údaj	12	2,0
	SPOLU		600	100,0

Zdroj: vlastný výskum realizovaný v mesiacoch október – december 2005

U zákazníkov bol predmetom výskumu aj ich zvyčajný dôvod návštevy konkrétnej predajne, v ktorej boli oslovení. Respondenti si mohli vybrať viacero možností, a tak okrem percentuálneho podielu zo všetkých odpovedí sú v tabuľke uvedené aj kumulatívne početnosti zákazníkov, ktorí v dotazníku túto možnosť uviedli.

Na otázku:

- **Za akým účelom zvyčajne navštevujete toto nákupné centrum/hypermarket? (môžete označiť viacero možností)**

takmer 95% zákazníkov v Prešove a 91% v Košiciach uviedlo ako dôvod svojej návštevy veľkoplošnej predajne **nakupovanie**. To, že **je tu všetko pod jednou strechou** a zákazník si môže porovnať rôzny tovar, je druhým najčastejšie sa opakujúcim dôvodom návštevy. Zatiaľ čo v Košiciach využije možnosť občerstviť sa, či najesť v niektorej z reštaurácií, či kaviarní štvrtina zákazníkov, v Prešove je to iba každý deviaty zákazník. Dôvodom je nepomer v počte gastronomických zariadení v Prešovských hypermarketoch a Košických nákupných centrách (v Optime sa ich nachádza 10, v Cassovii 5 reštaurácií a kaviarní a v prešovských hypermarketoch iba 1 reštaurácia).

Účel návštevy nákupného centra/hypermarketu je podrobne analyzovaný v nasledujúcej tabuľke 4.

Tabuľka 4. Účel návštevy nákupného centra/hypermarketu? (môžete označiť viacero možností)

Za akým účelom zvyčajne navštevujete toto NC/HM?	Prešov		Košice	
	% podiel odpovedí	∑ (v %) *	% podiel odpovedí	∑ (v %) *
nakupovať	35,1%	94,6%	27,1%	90,6%
pretože je tu všetko pod jednou strechou a môžem si porovnať rôzny tovar	13,7%	37,0%	11,8%	39,6%
pretože je tu oveľa väčší výber	9,5%	25,7%	11,3%	37,9%
pretože ma oslovila reklama	9,4%	25,4%	7,1%	23,8%
iba tak popozerať sa, informovať sa	9,1%	24,6%	10,4%	34,9%
pretože je tu vždy výhodnejšia ponuka	7,7%	20,8%	8,5%	28,4%
najesť sa / vypiť si niečo	4,2%	11,4%	7,5%	25,2%
stretnúť sa s priateľmi a so známymi	3,6%	9,7%	6,2%	20,8%
oddýchnuť si	2,8%	7,6%	4,8%	16,1%
pretože sa tu cítim bezpečne (napr. nebol som tu nikdy okradnutý...)	2,7%	7,3%	1,9%	6,4%
pretože tu nachádzam doplnkové služby (čistiareň, kaderníctvo, kino, ...)	1,8%	4,9%	2,6%	8,7%
pretože tu môžem nakupovať na svojom vozíčku pre invalidov bez obmedzenia	0,3%	0,8%	0,6%	2,0%
SPOLU	100,0%	269,7%	100,0%	334,4%

Zdroj: vlastný výskum realizovaný v mesiacoch október – december 2005

Vysvetlivky: ∑ (v %) * - početnosť zákazníkov (ktorí v dotazníku túto možnosť uviedli) vyjadrená v %

3. Záver

Záverom boli vyhodnotené niektoré spoločné a rozdielne znaky v nákupnom správaní sa spotrebiteľa v prešovských hypermarketoch a košických nákupných centrách. S rozvojom siete veľkoplošných predajní na území miest Prešova a Košíc došlo k zmene nákupného správania sa obyvateľov týchto miest a ich zázemia a tí to vnímajú ako zvyšovanie kvality života. Väčšina z oslovených zákazníkov vyjadrila svoj pozitívny postoj k výstavbe a prenikaniu týchto veľkoplošných formátov na náš trh.

- Takmer 50% zákazníkov v Košiciach a viac ako 60% v Prešove navštevuje nákupné centrum/hypermarket minimálne 1x v týždni
- Časová dostupnosť zákazníkov košických nákupných centier je väčšia ako zákazníkov prešovských hypermarketov
- Zákazníci košických nákupných centier strávia v priemere viac času v predmetnej predajni ako zákazníci prešovských hypermarketov
- Najčastejším dôvodom návštevy veľkoplošných predajní je nakupovanie (v Košiciach je tiež v porovnaní s Prešovom častejšie uvádzaný ako dôvod oddych a relaxácia)

4. Literatúra

- FERTAĽOVÁ, J. (2005a): Regionálnogeografické aspekty hodnotenia vývoja maloobchodu na Slovensku po roku 1989. In: Acta Facultatis Studiorum Humanitatis et Naturae Universitatis Prešoviensis, Prírodné vedy, XLIII., Folia geographica 8. Prešov: FHPV PU, 2005, p. 5-12.
- FERTAĽOVÁ, J. (2005b): Some methodological issues in classification of retail stores (with examples from European countries). In: Acta Facultatis Studiorum Humanitatis et Naturae Universitatis Prešoviensis, Prírodné vedy, XLIII., Folia geographica 8. Prešov: FHPV PU, 2005, p. 13-19, ISSN 1336-6157.
- FERTAĽOVÁ, J. – SZCZYRBA, Z. (2006): Globalisation in the Czech and Slovak retail: Common and Specific Features. In.: Siwek, T., Baar, V. (eds.): Globalisation and its Impact to Society, Regions and States (sborník příspěvků z mezinárodní konference). Ostravská univerzita, Ostrava, s. 164-173, ISBN 80-7368-256-7.
- CHAJDIAK, J.(2005): Štatistické úlohy a ich riešenie v exceli. Statis, Bratislava.
- KANDEROVÁ, M., ÚRADNÍČEK V. (2005): Štatistika a pravdepodobnosť pre ekonómov, 1. časť. OZ FINANC, Banská Bystrica.
- LUHA, J. (1985): Testovanie štatistických hypotéz pri analýze súborov charakterizovaných kvalitatívnymi znakmi. Vydal Odbor Výskumu programov ČST a divákov v SR. Bratislava.
- LUHA, J.(2006): Štatistické metódy analýzy kvalitatívnych znakov. FORUM STATISTICUM SLOVACUM 2/2006. SŠDS Bratislava.
- SZCZYRBA, Z. (1998a) Dimenze maloobchodní sítě v České republice [Dimension of the retail network in the Czech republic]. In: Dubcová, A. (ed.): Úlohy regionálnej geografie Slovenskej a Českej republiky v podmienkach transformujúcich sa ekonomík, Pedagogická fakulta UKF Nitra 5, s. 136-143.
- SZCZYRBA, Z. (2000a): Transformace struktur maloobchodní sítě České republiky (Regionálně geografická analýza s důrazem na Olomoucko). [Disertační práce]. Katedra geografie Přírodovědecké Fakulty MU, Brno, p. 145.
- SZCZYRBA, Z. (2003): Širší souvislosti transformace struktury TOP 10 obchodních společností – Česká republika. Malé a střední podniky před a po vstupu do Evropské unie (sbrník referátů), Obchodně podnikatelská fakulta Slezské univerzity, Karviná, p. 328 – 334.

- SZCZYRBA, Z. (2005b): Maloobchod v ČR po roce 1989. Vývoj a trendy se zaměřením na geografickou organizaci. PřF Univerzity Palackého Olomouc. p.126, ISBN 80-244-1274-8.
- SZCZYRBA, Z. (2005c): Územní rozvoj obchodních sítí v České republice a Slovenské republice. AFRNUC, Geographica, No 3, Bratislava, p. 574 – 581.

Adresa autora:
RNDr. Jana Fertaľová
Katedra geografie a regionálneho rozvoja
FHPV Prešovskej univerzity v Prešove
ul. 17. novembra 1
080 01 Prešov
jankageo@unipo.sk

Príspevok je súčasťou grantov:
VEGA: *Kvalita života - konceptuálny rámec komlexnej geografickej interpretácie priestorovej štruktúry mesta* (ved. R. Matlovič)

KEGA: *Koncepcia výuky a vypracovanie učebnice "Regionálny rozvoj a regionálna politika pre geografov" pre kľúčovú jednotku nového študijného programu geografia v regionálnom rozvoji* (ved. E. Michaeli)

Projekce vývoje nákladů na zdravotní péči

Tomáš Fiala[1], Jitka Langhamrová[2]

Abstract: The ageing of Czech population will bring increase of proportion of older persons**.** The expenditure of health care will grow. All variants of population projection show increasing gap between returns and expenditure of health insurance corporations.

Key words: population ageing, population projection, health insurance

1. Úvod

Ve svém příspěvku jsme se pokusili nastínit několik variant projekce budoucího vývoje nákladů na zdravotní péči a úhrnu vybraného pojistného na zdravotní pojištění v souvislosti s očekávaným stárnutím obyvatelstva ČR. Projekce byla vypočítána za předpokladu, že se nebudou měnit následující charakteristiky: průměrné náklady na zdravotní péči na jednoho pojištěnce podle pohlaví v jednotlivých věkových skupinách, výše průměrné mzdy podle pohlaví v jednotlivých věkových skupinách, míry zaměstnanosti podle pohlaví v jednotlivých věkových skupinách ani předpisy týkající se platby zdravotního pojištění. Projekce vývoje nákladů na zdravotní péči a pojistného na zdravotní pojištění vycházela z vlastní projekce vývoje obyvatelstva ČR. Cílem této projekce nebylo odhadnout co nejpřesněji budoucí vývoj obyvatelstva ČR jako spíš nastínit možné varianty vývoje při poměrně velkých rozdílech mezi minimální a maximální variantou úmrtnosti, plodnosti i migrace.

2. Projekce obyvatelstva

Jako výchozí věková struktura byla použita poslední dostupná data, tj. demografická struktura obyvatelstva ČR k 1. 1. 2006, horizontem projekce byl 1. leden 2056. Po celé období projekce se předpokládal lineární nárůst střední délky života (u mužů o něco vyšší než u žen). Pro plodnost se předpokládal do roku 2020 lineární nárůst se současnou změnou struktury plodnosti aby v roce 2020 byla již stejná jako současná struktura plodnosti žen Nizozemska. V období 2020–2055 se předpokládal pomalejší lineární nárůst bez dalších změn struktury. Roční migrační přírůstek se předpokládal konstantní po celé období projekce. Vývoj střední délky života, úhrnné plodnosti i migrace byl uvažován ve třech alternativách.

Tabulka 1. Základní alternativy vývoje plodnosti, úmrtnosti a migrace

Alternativa vývoje	*Roční nárůst střední délky života*		*Úhrnná plodnost*		*Roční migrační přírůstek*
	muži	*ženy*	*2020*	*2055*	
nízká	0,10	0,08	1,4	1,5	10 000
střední	0,20	0,16	1,5	1,7	30 000
vysoká	0,30	0,24	1,6	1,9	50 000

K největšímu stárnutí však dochází při vysokém růstu střední délky života v kombinaci s nízkou plodností. Proto byl výpočet projekce proveden celkem v devíti variantách, byly uvažovány různé kombinace alternativ vývoje plodnosti, úmrtnosti i migrace. Jejich přehled a značení udává následující tabulka.

[1]Tomáš Fiala, katedra demografie fakulty informatiky a statistiky Vysoké školy ekonomické v Praze
[2]Jitka Langhamrová, katedra demografie fakulty informatiky a statistiky Vysoké školy ekonomické v Praze

Tabulka 2. Přehled variant projekce

Varianta projekce	*Nárůst střední délky života*	*Nárůst plodnosti*	*Migrační přírůstek*
NNN	nízký	nízký	nízký
NNV	nízký	nízký	vysoký
NVN	nízký	vysoký	nízký
NVV	nízký	vysoký	vysoký
SSS	střední	střední	střední
VNN	vysoký	nízký	nízký
VNV	vysoký	nízký	vysoký
VVN	vysoký	vysoký	nízký
VVV	vysoký	vysoký	vysoký

2. Projekce poměru pojistného a nákladů na zdravotní péči

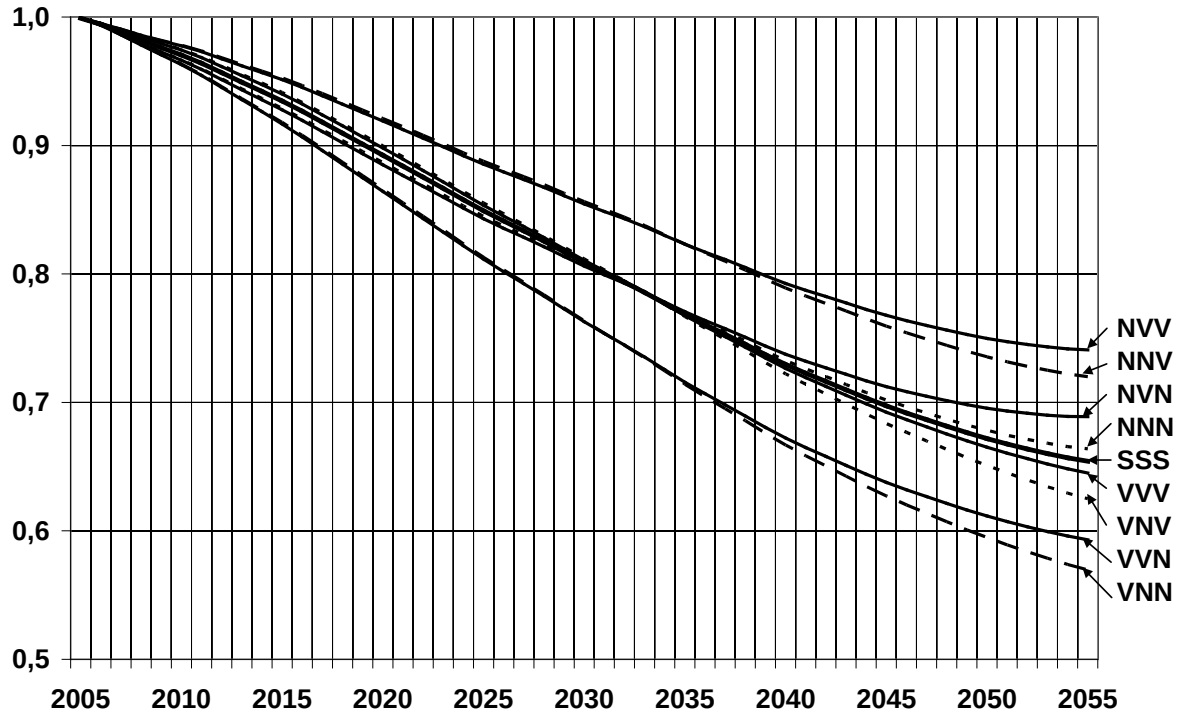
Na základě projekce vývoje obyvatelstva byly provedeny projekce výdajů na zdravotní péči i projekce vybraného pojistného na zdravotní pojištění. Projekce výdajů na zdravotní péči vycházela z předpokladu, že náklady na jednoho pojištěnce podle pohlaví a pětiletých věkových skupin zůstanou po celé období projekce na úrovni roku 2004. Odhad vývoje vybraného pojistného byl obtížnější, neboť nebyly k dispozici analogické údaje o průměrné výši vybraného pojistného podle pohlaví a věku. Vycházelo se proto ze zjednodušeného předpokladu, že osoba zaměstnaná platí pojistné z částky odpovídající 80 % hrubé průměrné mzdy[3] pro dané pohlaví a daný věk v roce 2004 zatímco za osobu nezaměstnanou platí pojistné stát v předepsané výši (opět pro rok 2004). Podle definice se za zaměstnané považují nejen osoby v pracovním či obdobném poměru, ale i podnikatelé. Lze předpokládat, že osoby, jejichž jediným nebo hlavním příjmem je příjem ze samostatné výdělečné činnosti, odvádějí na zdravotní pojištění v průměru (možná podstatně) nižší pojistné, než osoby téhož věku a pohlaví v pracovním nebo obdobném poměru. Jako míry zaměstnanosti podle pohlaví a věku byly použity údaje z výběrového šetření pracovních sil ČSÚ .

Vývoj finanční zátěže systému zdravotního pojištění můžeme charakterizovat podílem úhrnu předpokládaného vybraného pojistného ku úhrnu předpokládaných nákladů na zdravotní péči hrazených pojišťovnou. Vývoj této charakteristiky pro jednotlivé varianty projekce zobrazuje Graf 1.

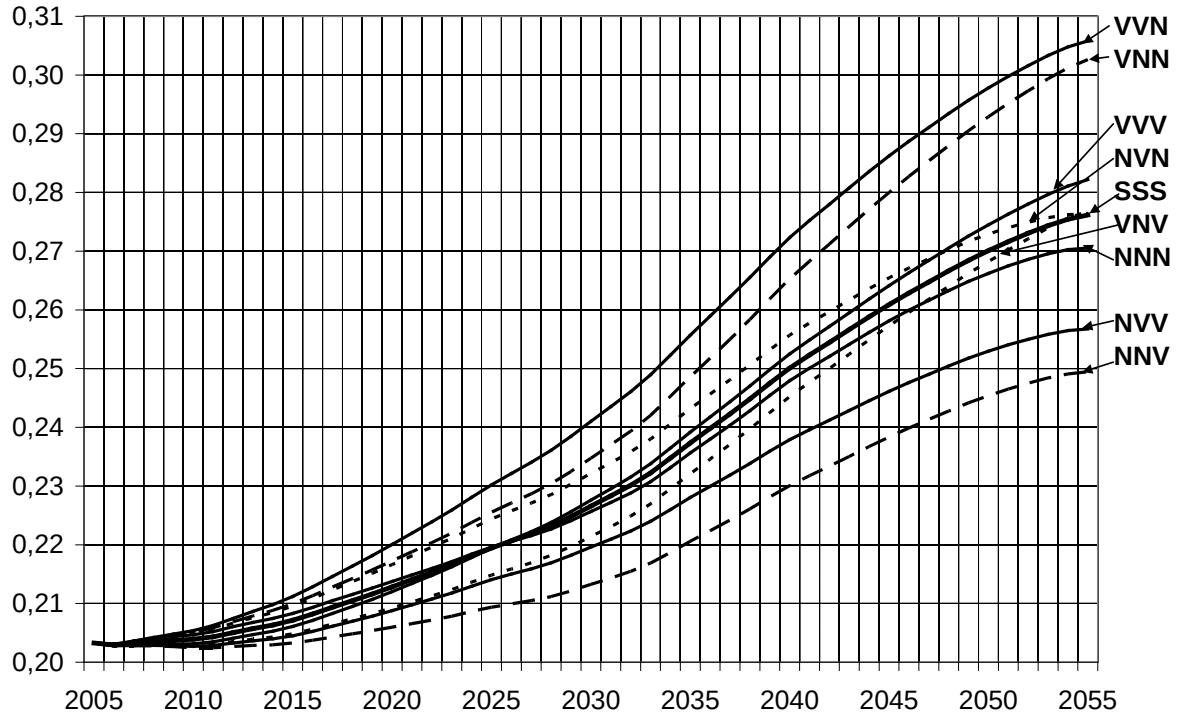
Vidíme, že při všech alternativách demografického vývoje by docházelo k poklesu podílu pojistného ku nákladům hluboko pod jednotkovou hodnotu. To znamená, že pojistné by pojišťovnám nestačilo na pokrytí nákladů zdravotní péče. Nejmenší pokles by nastal podle očekávání v případě malého nárůstu střední délky života a vysoké imigrace, naopak největší pokles by přinesl vysoký nárůst střední délky života při současné nízké imigraci. Ostatní alternativy vývoje se zejména v prvních letech příliš neliší od střední varianty projekce. Vliv tempa růstu plodnosti by se začínal výrazněji projevovat až za několik desítek let, vyšší nárůst plodnosti by znamenal nižší pokles podílu pojistného ku nákladům.

Za většinu osob, které nejsou zaměstnané, platí pojistné na zdravotní pojištění zpravidla stát (děti, žáci, studenti, nepracující důchodci, nezaměstnaní aktivně hledající práci atd.). Na Grafu 2 je vidět, jak by se měnil podíl pojistného placeného státem v souvislosti se změnami demografické struktury české populace.

[3] Pro osoby v pracovním nebo obdobném poměru je tzv. vyměřovacím základem, z něhož se platí pojistné, výše hrubé mzdy, pro osoby samostatně výdělečně činné pak 50 % rozdílu příjmů a výdajů. V roce 2004 činila úhrnná výše pojistného vybraného od zaměstnaných osob zhruba 80 % hodnoty vypočtené za předpokladu, že by každá zaměstnaná osoba platila pojistné z průměrné mzdy pro své pohlaví a svůj věk.



Graf 1: Podíl vybraného pojistného ku nákladům na zdravotní péči hrazeným pojišťovnami podle jednotlivých alternativ vývoje

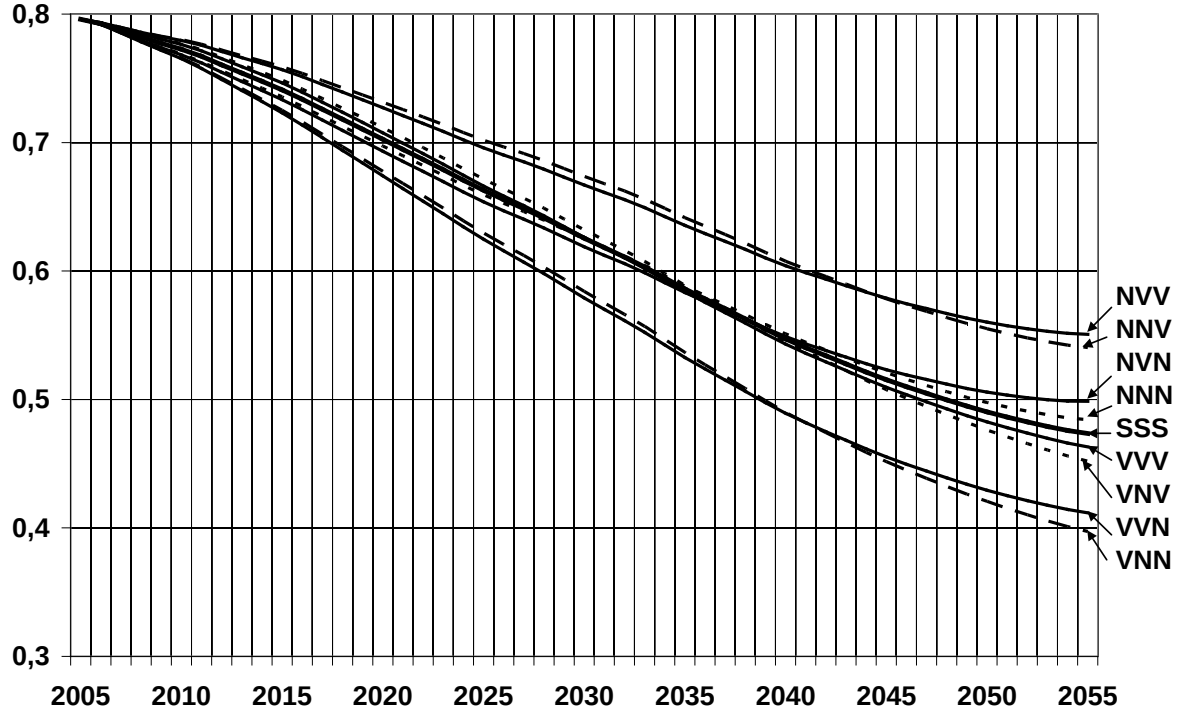


Graf 2: Podíl úhrnu pojistného placeného státem ku celkovému pojistnému na zdravotní pojištění podle jednotlivých alternativ vývoje

Při všech alternativách vývoje by podíl pojistného placeného státem rostl. Nejvíce v případě vysokého růstu střední délky života, vysoké plodnosti a nízké imigrace, nejméně

v případě opačném. Vysoká plodnost by znamenala opět v prvních letech zvýšení zátěže, později by vedla k jejímu snížení.

Celkovou zátěž státu související s úhradou nákladů na zdravotní péči můžeme proto charakterizovat podílem úhrnné výše pojistného vybraného pouze od zaměstnaných osob ku úhrnné výši nákladů na zdravotní péči. Vývoj této charakteristiky vidíme na Grafu č. 3.



Graf 3: Podíl vybraného pojistného od zaměstnaných osob ku nákladům na zdravotní péči hrazeným pojišťovnami podle jednotlivých alternativ vývoje.

4. Závěry:

Při všech variantách demografického vývoje by tedy docházelo k podstatnému zvyšování rozdílu mezi předpokládanými příjmy a předpokládanými výdaji zdravotních pojišťoven. Jaké jsou možnosti řešení tohoto nepříznivého trendu?

První možností je samozřejmě zvyšování odvodů na zdravotní pojištění. Pokud bychom například chtěli, aby byl podíl vybraného pojistného ku nákladům na zdravotní péči hrazeným pojišťovnami roven přibližně jedné i v dalších letech, musela by se (za předpokladu nezměněných dalších charakteristik) sazba pojistného do roku 2055 postupně zvýšit z dnešních 13,5 % na 18–24 % vyměřovacího základu (viz následující tabulka).

Tabulka 3. Potřebné sazby pojistného na zdravotní pojištění v roce 2055 (v %)

Varianta	*SSS*	*MMM*	*MMV*	*MVM*	*MVV*	*VMM*	*VMV*	*VVM*	*VVV*
Sazba	20,6	20,3	18,7	19,6	18,2	23,7	21,6	22,8	20,9

Jinou možností, jak zvýšit úhrn vybraného pojistného a snížit zátěž státu, je zvyšování míry zaměstnanosti, především v souvislosti s pokračujícím zvyšováním důchodového věku. Pokud by se například do roku 2055 zvýšila míra zaměstnanosti 55–59letých mužů na 80 % a 60–64letých mužů na 75 % a míra zaměstnanosti žen v těchto věkových skupinách by byla jen o 5 procentních bodů nižší než u mužů, byl by v roce 2055 v případě střední varianty demografického vývoje podíl vybraného pojistného ku nákladům na zdravotní péči roven zhruba 70 % (zatímco při zachování současné míry zaměstnanosti pouze 65 %). Zvyšování

zaměstnanosti tedy nárůst rozdílů mezi příjmy a výdaji pouze sníží, ale neeliminuje jej zcela. Důležité by bylo rovněž zlepšit výběr pojistného od osob, jejichž hlavním zdrojem příjmu je samostatná výdělečná činnost. Jiným ekonomickým nástrojem snižujícím narůstající rozdíl mezi příjmy a výdaji zdravotních pojišťoven by bylo zvýšení spoluúčasti pojištěnců.

Budoucí vývoj je velmi těžké odhadnout. Je však pravděpodobné, že pokračujícího zlepšování zdravotního stavu obyvatelstva bude možno (v souvislosti s vývojem nových zdravotnických technologií) v budoucnu dosahovat s nižšími relativními náklady než v současné době.

5. Literatura

DURDISOVÁ, J. – LANGHAMROVÁ, J. Úvod do teorie zdravotní politiky. 1. vyd. Praha.VŠE, 2001.126 s.ISBN 80-245-0217-8

FIALA, T. Dva přístupy modelování vývoje úmrtnosti v populační projekci a jejich aplikace na populaci ČR, In: FORUM STATISTICUM SLOVACUM. 4/2006, Slovenská štatistická a demografická spoločnosť, Bratislava 2006, ISSN 1336-7420

VÝROČNÍ ZPRÁVY VZP

DOKUMENTY Z INTERNETOVÝCH STRÁNEK:

Analýza zdravotnických účtů ČR časová řada 2000-2002
http://www.czso.cz/csu/2004edicniplan.nsf/publ/1524-04-casova_rada_2000_2002
Analýza zdravotnických účtů ČR (2000 - 2004)
http://www.czso.cz/csu/2006edicniplan.nsf/publ/3306-06-(2000___2004)
Analýza zdravotních účtů 2005
http://www.czso.cz/csu/2005edicniplan.nsf/publ/1524-05-
Návrh koncepce zdravotnictví na léta 2005 – 2009)
http://www.lekarnici.cz/module.php?module=36&article=1643
Světové šetření o zdraví v České republice 2003
http://www.uzis.cz/download.php?ctg=10&search_name=světové&region=100&kind=2&mnu_id=5300
Vliv lékařského výzkumu na zdraví a ekonomiku
(Vybráno z: Pardes H, Manton KG, Lander ES, Tolley HD, Ullian AD, Palmer H. Effects of Medical Research on Health Care and the Economy. Science 1999;283:36-37)
http://www.tigis.cz/PSYCHIAT/PSYCH299/11zpravy.htm#1
Zdravotnické ročenky České republiky v roce 2000-2004
http://www.uzis.cz/download.php?ctg=10&search_name=ročenka&region=100&mnu_id=5300
Zdravotnictví ČR ve statistických údajích 2000-2005
http://www.uzis.cz/download.php?ctg=10&search_name=Zdravotnictví%20ČR&region=100&kind=2&mnu_id=5300

Adresy autorů:
Tomáš Fiala, RNDr., CSc., Jitka Langhamrová, Ing., CSc.
Katedra demografie fakulty informatiky a statistiky VŠE
130 67 Praha 3
fiala@vse.cz, langhamj@vse.cz

Průřezová a kohortní analýza úmrtnosti v ČR pomocí životních potenciálů

Tomáš Fiala[1]

Abstract: The cohort analysis of mortality in the Czech Republic was made. Because of incomplete age-specific mortality rates series the life potential (generalized life expectancy) was used for the analysis. The mortality of women was decreasing all the period investigated. On the other hand the mortality of man at the age over 40 was increasing for some generations.

Key words: life tables, life expectancy, life potential, cohort analysis

1. Úvod

Demografická analýza úmrtnosti se provádí zpravidla na základě specifických měr úmrtnosti, resp. z nich vypočtených pravděpodobností přežití či úmrtí a dalších biometrických měr. Průřezová analýza úmrtnosti charakterizuje úmrtnost celé populace v krátkém časovém období, zpravidla v jednom roce. Vypočtené charakteristiky délky života se proto netýkají skutečné populace, ale populace hypotetické, pro niž by po celou dobu jejího života byla úmrtnost stejná jako ve sledovaném roce.

Doplňkem průřezové analýzy úmrtnosti je proto analýza generační vycházející ze specifických měr úmrtnosti osob určitého ročníku narození během jejich celého života. Vypočtené charakteristiky délky života se pak týkají osob daného ročníku narození. Protože podrobná statistika úmrtnosti v ČR je dispozici až od roku 1920, nemáme zatím pro žádnou generaci úplnou řadu specifických měr úmrtnosti. Není proto možné vypočítat střední délku života novorozence pro žádnou generaci, lze však použít určitou analogii střední délky života pro užší věková rozmezí.

2. Používané charakteristiky úmrtnosti

Označme l_x počet dožívajících, L_x pak počet prožitých let (počet žijících) z úplných úmrtnostních tabulek. Střední délka e_x^0 života osoby v přesném věku x je definována následovně

$$e_x^0 = \frac{\sum_{u=x}^{\omega-1} L_u}{l_x}, \tag{1}$$

udává průměrnou délku zbývajícího života (při dané úmrtnosti) osoby v přesném věku x, charakterizuje tedy úmrtnost pouze osob x-letých a starších. Často se uvádí její hodnota pouze pro věk x=0, tzv. střední délka života novorozence.

Pokud předchozím vzorec dále zobecníme v tom smyslu, že neuvažujeme úmrtnost až do konce života, ale pouze po dobu h let, dostáváme tzv. střední životní potenciál v intervalu x–(x+h-1) dokončených let

$${}_h e_x^0 = \frac{\sum_{u=x}^{h-1} L_u}{l_x}, \tag{2}$$

Jedná se tedy o charakteristiku úmrtnosti osob v dokončeném věku x–(x+h-1), tedy v intervalu od x do x+h let přesného věku. Životní potenciál tedy udává, kolik let (z maximálního počtu h let) prožije v průměru osoba v přesném věku x během následujících h let.

[1]Tomáš Fiala, katedra demografie fakulty informatiky a statistiky Vysoké školy ekonomické v Praze

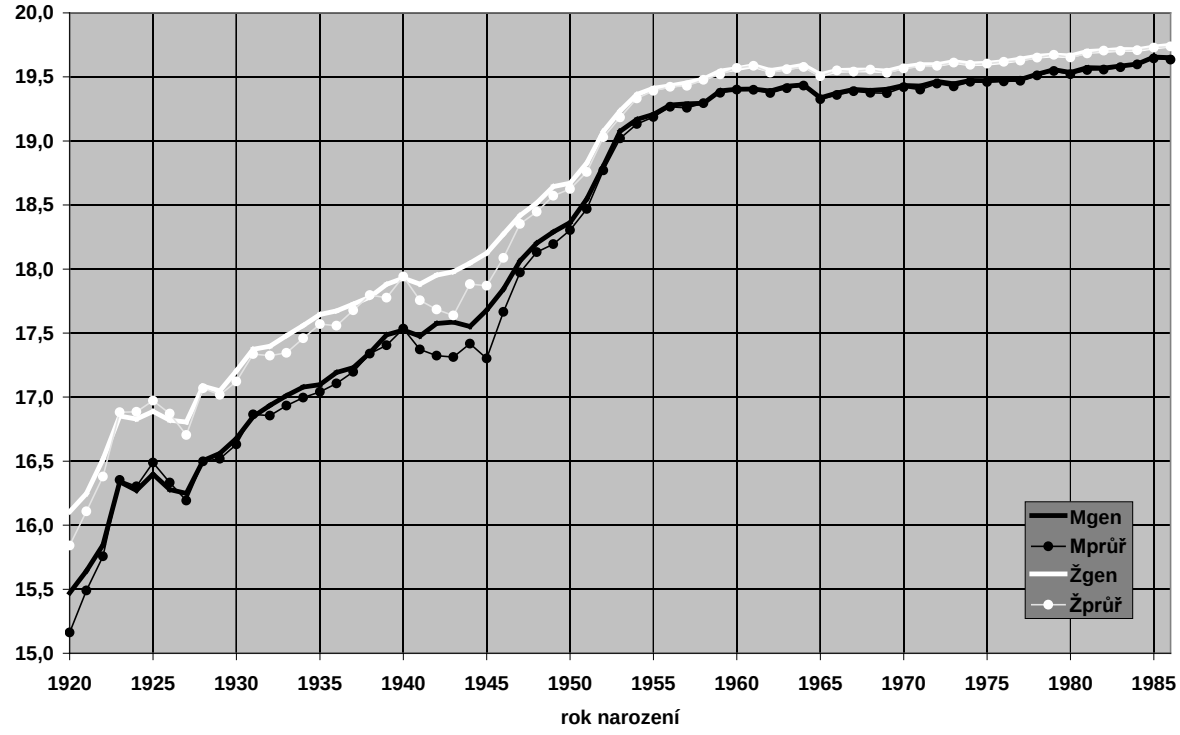
Pro výpočet životního potenciálu tedy stačí mít řadu specifických měr úmrtnosti odpovídající délce příslušného intervalu.

3. Hlavní výsledky generační analýzy úmrtnosti v ČR od roku 1920

Analýza vycházela z dat ČSÚ od roku 1920 do roku 2005. Za jednotlivé roky byly vypočteny úmrtnostní tabulky, za léta 1938–44 byly použity přímo tabulky ČSÚ. Výpočet generačních pravděpodobností přežití se prováděl na základě průřezových pravděpodobností pro vzájemně se překrývající dvouleté generace. (Pravděpodobnost přežití osoby ve věku *x* v roce *t* byla považována za pravděpodobnost přežití ve osoby věku *x* z generace narozených ve dvouletí *t*-*x* a *t*-*x*-1.) Pro zjednodušení značení však byla každá dvouletá generace označena pouze vyšším ročníkem narození.

Na základě průřezových i generačních úmrtnostních tabulek byly provedeny výpočty středních délek života a středních životních potenciálů pro jednotlivé roky i pro jednotlivé generace. Pro generace bylo pochopitelně možné vypočítat jen životní potenciály týkající se období, které členové generace prožili od roku 1920 do roku 2005.

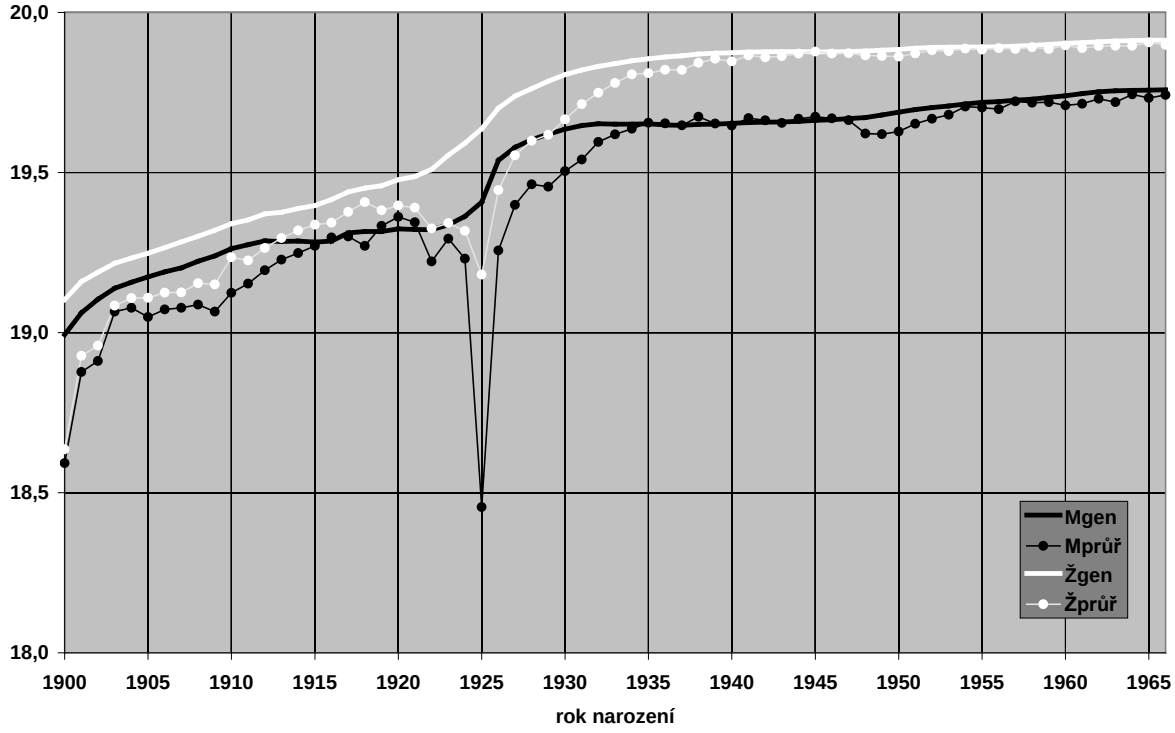
Základní výsledky analýzy vidíme na následujících grafech. Hodnoty průřezových životních potenciálů zvlášť pro muže (M) a ženy (Ž) zobrazené v grafech se týkají vždy roku, kdy příslušná generace dosáhla dolní meze intervalu životního potenciálu. (Např. v grafu středních životních potenciálů ve věku 20–39 let je u generace narozených např. v r. 1940 zobrazen pro srovnání průřezový životní potenciál v roce 1960, tedy v roce, kdy generace dosáhla věku 20 let.) Průřezový životní potenciál udává hodnotu generačního potenciálu za předpokladu, že by se po následujících *h* let (*h* je šířka intervalu potenciálu) úmrtnost neměnila. Většinou je průřezový potenciál (průř) nižší než generační (gen) , což svědčí o tom, že v dalších letech se úmrtnost v daném věkovém intervalu snižovala. Pokud je průřezový potenciál vyšší než generační, znamená to, že se úmrtnost zvýšila).



Graf 1. Střední životní potenciál ve věku 0–19 dokončených let

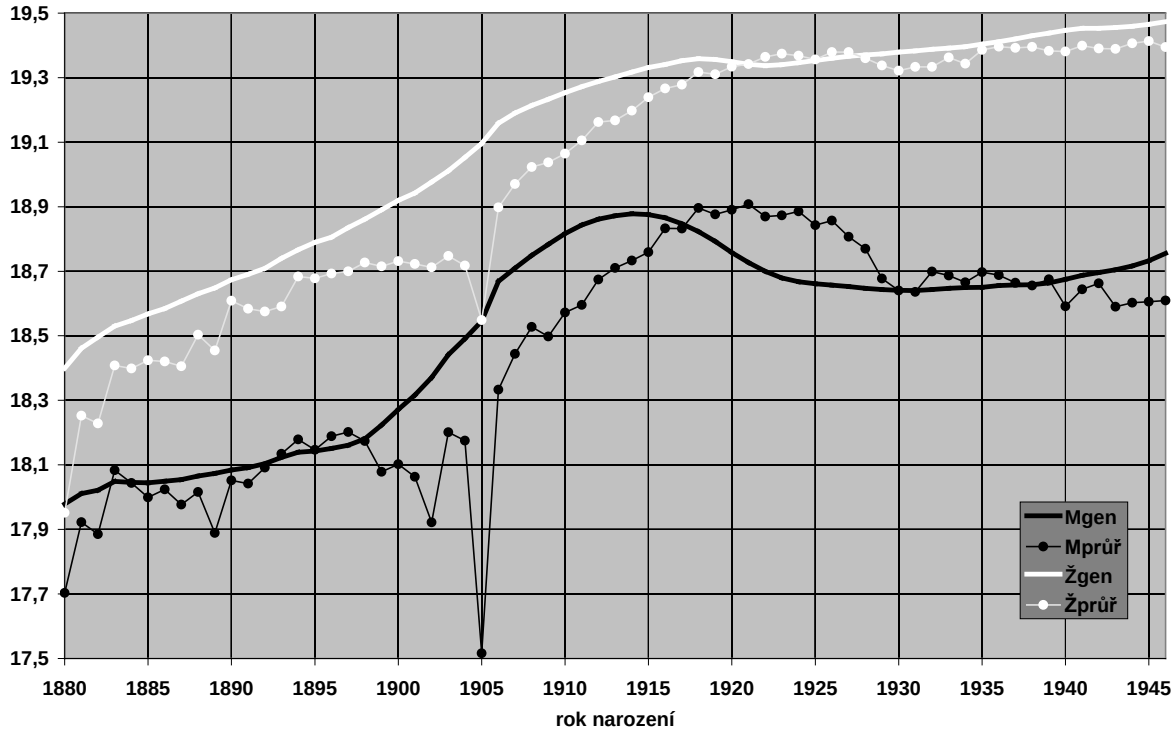
Na prvním grafu vidíme úmrtnost v prvních dvaceti letech života. Růst životního potenciálu u generací narozených v první polovině minulého století svědčí o trvalém poklesu

úmrtnosti v daném věku, způsobeném především výrazným poklesem úmrtnosti kojenecké. U osob narozených po roce 1955 je již úmrtnost nízká a její další pokles poměrně pomalý.



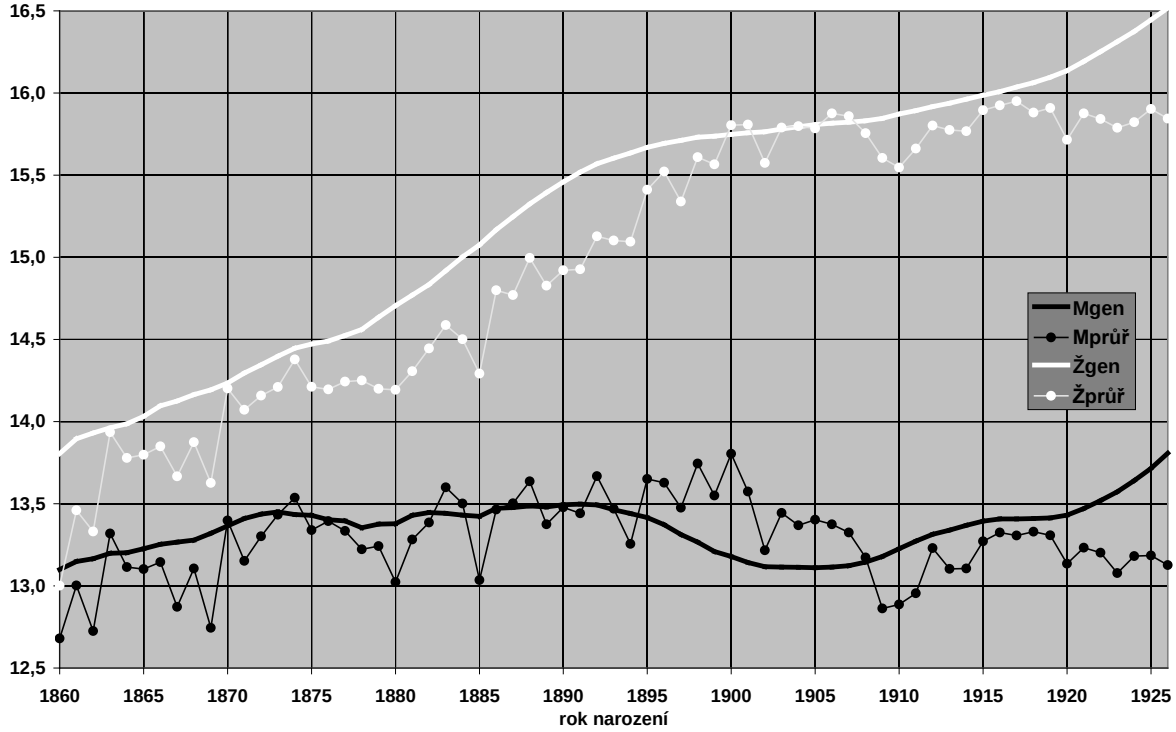
Graf 2. Střední životní potenciál ve věku 20–39 dokončených let

Podobný trend vykazuje vývoj úmrtnosti ve věku 20–39 dokončených let (viz Graf 2). Růst se však zpomaluje již počínaje generacemi narozenými po roce 1930, tedy osobami, které dosáhly věku 20 let po roce 1950. Vidíme tedy, že ve druhé polovině minulého století se úmrtnost osob do 40 let již dále výrazně nesnižovala.

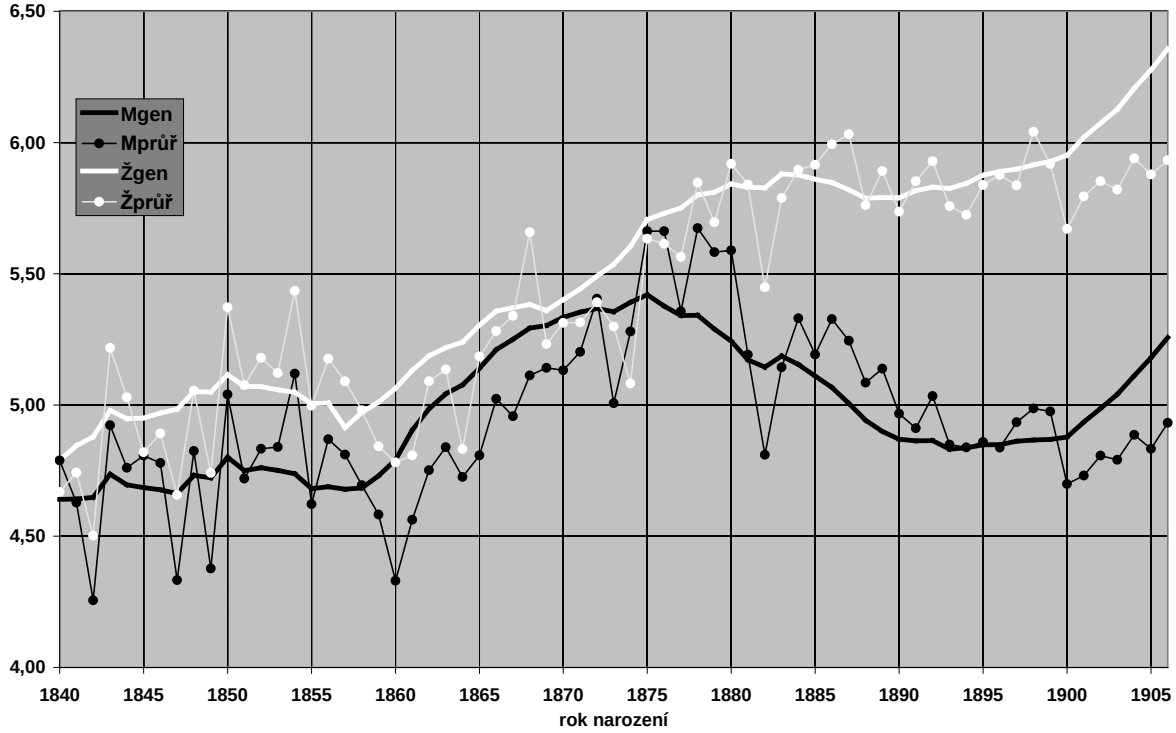


Graf 3. Střední životní potenciál ve věku 40–59 dokončených let

Trochu jiný trend má vývoj úmrtnosti ve věku 40–59 let (viz Graf 3). Po počátečním nárůstu se u mužů narozených po roce 1915 (tedy mužů, kteří dosáhli 40 let po roce 1955) začíná úmrtnost v daném intervalu zvyšovat a ani u mužů narozených těsně po druhé světové válce nedosahuje životní potenciál úrovně generací narozených kolem roku 1915. U žen dochází pouze k výraznému zpomalení růstu životního potenciálu.

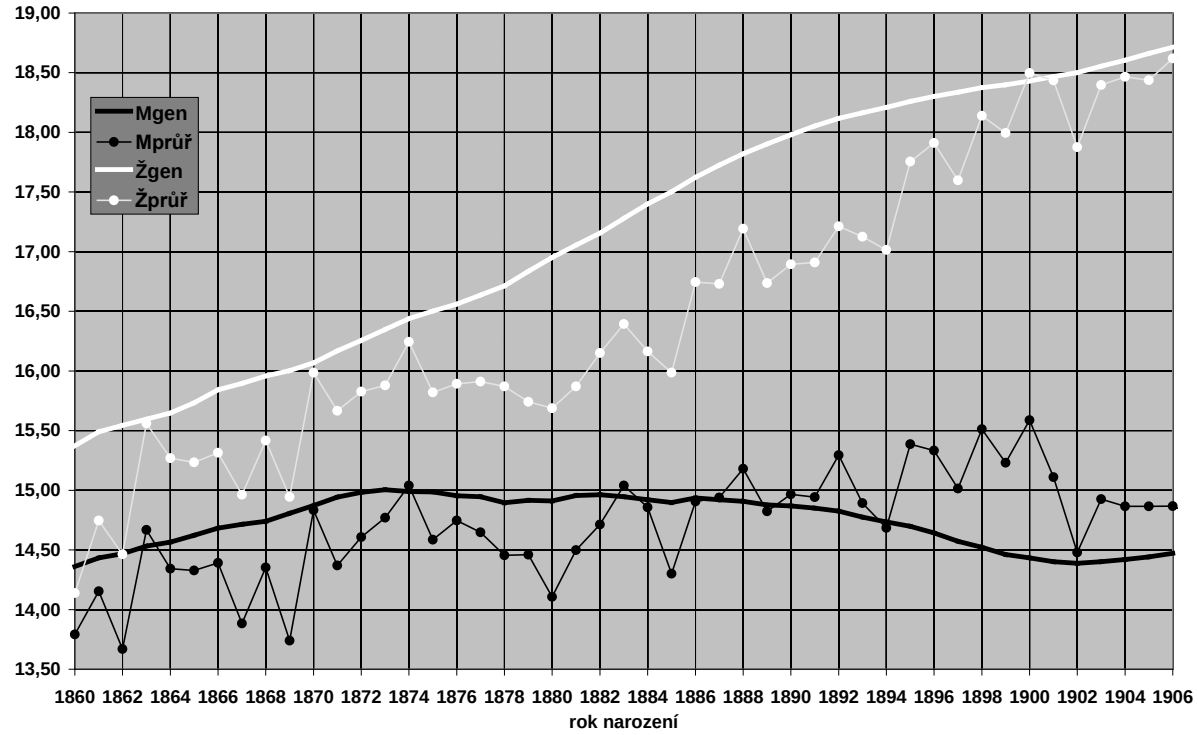


Graf 4. Střední životní potenciál ve věku 60–79 dokončených let

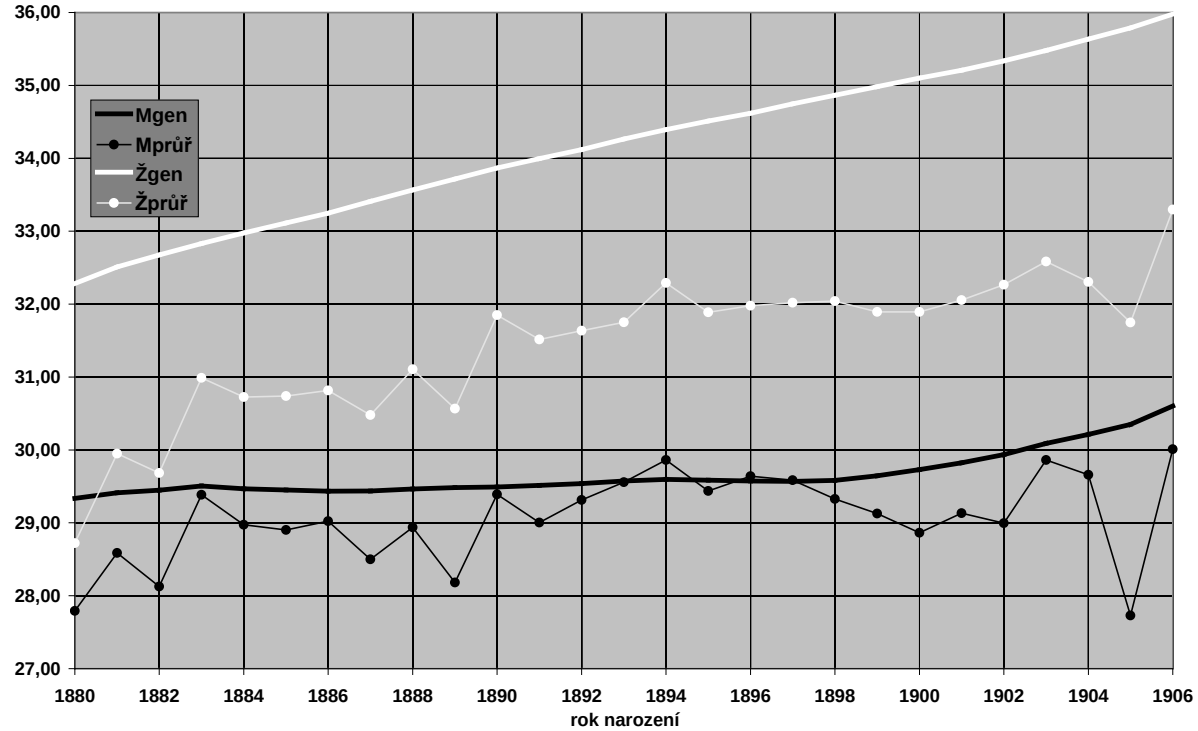


Graf 5. Střední délka života ve věku 80 let

Rovněž ve věkovém intervalu 60–79 let (viz Graf 4), pozorujeme pokles životního potenciálu u mužů narozených koncem 19. století a počátkem století 20, tedy těch, kteří dosáhli 60 let věku po roce 1955. Teprve muži narození po roce 1920 měli v uvedeném věku opět nižší úmrtnost než muži narození v letech 1870–90.



Graf 6. Střední délka života ve věku 60 let



Graf 7. Střední délka života ve věku 40 let

Úmrtnost na samém konci lidského života charakterizuje střední délka života ve věku 80 let (Graf 5). tedy opět těch, kteří dosáhli 60 let věku po roce 1955. Obdobně jako v předchozích věkových skupinách i v tomto případě pozorujeme u mužů zvyšování úmrtnosti, v tomto případě u mužů narozených v poslední čtvrtině devatenáctého století.

Vidíme tedy, že u mužů docházelo ve vyšším věku ke zvyšování generační úmrtnosti. Toto zvýšení však bylo později u některých generací částečně vykompenzováno. Například u generací narozených po roce 1920 vidíme ve srovnání s předchozími generacemi vyšší úmrtnost ve věku 40–59 let, ale o něco nižší úmrtnost ve věku 60–79 let

Poslední dva grafy zachycují generační střední délku života ve věku 60, resp. 40 let. (Grafy 6, resp. 7). Zatímco střední délky života mužů ve věku 40 let stagnuje nebo mírně roste, střední délka života ve věku 60 let u mužů narozených koncem 19. století mírně klesá.

4. Závěr

Vývoj úmrtnosti jednotlivých generací nevykazuje tak velké náhodné odchylky v jednotlivých letech jako vývoj průřezový. Zatímco u jednotlivých generací žen pozorujeme trvalý pokles nebo stagnaci úmrtnosti, u některých generací mužů došlo ke vzrůstu úmrtnosti ve věku nad 40 let, který byl kompenzován jen částečně. Uvedený vývoj souvisí s dlouhodobou stagnací či zvyšováním úmrtnosti mužů ve věku nad 40 let v 60. a 70. letech minulého století.

5. Literatura

BĚLÁČEK, J. – FIALA, T. 2003. Analýza zdravotního stavu obyvatelstva ČR a kraje Vysočina. Studie vypracována v souladu s Výzkumným záměrem MZ0 2002 01 IZPE. [Výzkumná zpráva]. Praha: STADEA, 2003. 21 s.

FIALA, T. 1999. Analýza úmrtnosti ve středním věku v České republice v letech 1950-1998. Trenčianské Teplice 13.09.1999 – 15.09.1999. In: Demografické, zdravotné a sociálno-ekonomické aspekty úmrtnosti. Bratislava : Slovenská štatistická a demografická spoločnosť, 1999, s. 19–30. ISBN 80-88946-00-X.

KOSCHIN, F. – FIALA, T. – LANGHAMROVÁ, J. – ROUBÍČEK, V. 1998. Úmrtnost v českých zemích v devadesátých letech. Praha: VŠE, 1998. 68 s. ISBN 80-7079-574-3.

Tento příspěvek vznikl za podpory prostředků poskytnutých Grantovou agenturou České republiky projektu č. 403/06/2006 Operacionalizace projekčního modelu pro regionální prognózy zdravotního stavu obyvatelstva ČR" – viz www.morbidity.wz.cz .

Adresa autora:
Tomáš Fiala, RNDr., CSc.
Katedra demografie fakulty informatiky a statistiky VŠE
130 67 Praha 3
fiala@vse.cz

ČÍSLO π A ROZVOJE FUNKCIÍ b_f A $\ln(b_f)$

Ivan GARAJ

Abstract. In the paper are given the asymptotic expansions of functions b_f and $\ln(b_f)$. Both expansions are used for approximation formula for number π. Three programs in Mathematica system are presented.

Key words: Number π, Asymptotic Series, Gamma Function, Stirling formula

1. Úvod

Nech $(X_1, X_2, \ldots, X_n)$ je náhodný výber z náhodnej premennej X, ktorá má normálne rozdelenie $N(\mu, \sigma^2)$. Parametre μ a σ^2 sú neznáme. Nevychýlené odhady týchto parametrov sú $\overline{X}$ a S^2, kde $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ je výberový aritmetický priemer a $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ výberový rozptyl, lebo stredná hodnota $E(\overline{X}) = \mu$ a $E(S^2) = \sigma^2$. Rozptyl $D(S) = E(S^2) - E^2(S) = \sigma^2 - E^2(S) \geq 0$, z čoho vyplýva, že $E(S) \leq \sigma$ a teda výberová smerodajná odchýlka S je asymptoticky nevychýleným odhadom smerodajnej odchýlky σ, pričom pre $f = n-1$ platí

$$E(S) = \sqrt{\frac{2}{f}}\,\frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)}\,\sigma = b_f\,\sigma \tag{1}$$

kde $\Gamma(f) = \int_0^\infty x^{f-1} e^{-x}\,dx$ je gama funkcia $(f > 0)$. Zo vzťahu (1) vyplýva, že $E\left(\frac{S}{b_f}\right) = \sigma$ a $D(S) = (1 - b_f^2)\sigma^2$. Štatistika $\frac{S}{b_f}$ je preto [1] nevychýleným odhadom parametra σ, no rozptyl $D\left(\frac{S}{b_f}\right) = \sigma^2\left(\frac{1-b_f^2}{b_f^2}\right) > D(S) = (1-b_f^2)\sigma^2$. Hodnoty funkcie

$$b_f = \sqrt{\frac{2}{f}}\,\frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)} \tag{2}$$

závisia iba od stupňov voľnosti $f = n-1$.

2. Rozvoj funkcie b_f do nekonečného radu.

Využitím Stirlingovho rozvoja gama funkcie [2] do nekonečného radu

$$\Gamma(f+1) = f^f e^{-f} \sqrt{2\pi f} \left(1 + \frac{1}{12\,f} + \frac{1}{288\,f^2} - \frac{139}{51\,840\,f^3} - \frac{571}{2\,488\,320\,f^4} + O\left(\frac{1}{f^5}\right)\right) \tag{3}$$

možno pomerne komplikovane [3] získať prvé štyri členy rozvoja funkcie b_f

$$b_f = \sqrt{\frac{2}{f}} \frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)} = 1 - \frac{1}{4\,f} + \frac{1}{32\,f^2} + \frac{5}{128\,f^3} + O\left(\frac{1}{f^4}\right) \tag{4}$$

Kompletný rozvoj [4] je daný vzťahom

$$b_f = 1 + \sum_{i=1}^{\infty} (-1)^i \frac{c_i}{2^{3i-1}\,f^i} \tag{5}$$

kde $c_1 = 1$ a c_i $(i = 2, 3, ...)$ sa vypočítajú podľa rekurentného vzťahu

$$c_i = 2^{i-2} \binom{2i}{i} - \frac{1}{2} \sum_{j=1}^{i-1} \binom{i-1}{j-1} 2^{3(i-j)}\, c_j - \sum_{k=1}^{i-1} \sum_{j=1}^{i-k} \binom{i-k-1}{j-1} 2^{3(i-j-k)}\, c_j\, c_k \tag{6}$$

V systéme Mathematica [5] výpočet konštánt c_i $(i = 1, 2, ..., 16)$ podľa vzťahu (6) bol realizovaný programom č. 1. Numerický výstup možno nájsť v tabuľke č. 1. Prvých desať členov rozvoja (5) bude v tvare

$$b_f = 1 - \frac{1}{4\,f} + \frac{1}{32\,f^2} + \frac{5}{128\,f^3} - \frac{21}{2\,048\,f^4} - \frac{399}{8\,192\,f^5} + \frac{869}{65\,536\,f^6} + \frac{39\,325}{262\,144\,f^7} -$$

$$- \frac{334\,477}{8\,388\,608\,f^8} - \frac{28\,717\,403}{33\,554\,432\,f^9} + O\left(\frac{1}{f^{10}}\right) \tag{7}$$

Tabuľka č. 1. Výpočet konštánt c_i

i	c_i
1	1
2	1
3	- 10
4	- 21
5	798
6	1 738
7	- 157 300
8	- 334 477
9	57 434 806
10	119 394 366
11	- 33 601 489 740
12	- 68 858 583 810
13	28 797 022 447 980
14	58 526 378 304 180
15	- 34 009 655 736 503 400
16	- 68 787 420 596 367 128

Rozvoj funkcie $\ln(b_f)$ do nekonečného radu konverguje ešte rýchlejšie, ako rozvoj b_f. V [3] možno nájsť prvé tri členy tohto rozvoja pomocou Stirlingovho rozvoja [2] funkcie

$$\ln\Gamma(f+1) = \frac{1}{2}\ln(2\pi) - f + \left(f+\frac{1}{2}\right)\ln f + \sum_{i=1}^{\infty} \frac{B_i}{i(i+1)\,f^{2i-1}} \tag{8}$$

kde B_i sú Bernulliho čísla, pričom prvých osem [2] nadobúda nasledujúce hodnoty:

$$B_1 = \frac{1}{6},\ B_2 = \frac{1}{30},\ B_3 = \frac{1}{42},\ B_4 = \frac{1}{30},\ B_5 = \frac{5}{66},\ B_6 = \frac{691}{2\,730},\ B_7 = \frac{7}{6},\ B_8 = \frac{3617}{510} \tag{9}$$

Pomocou rozvoja (7) a tabuľky č. 1 možno v systéme Mathematica [5] programom č. 2 získať prvých osem členov rozvoja $\ln(b_f)$:

$$\ln(b_f) = -\frac{1}{4f} + \frac{1}{24f^3} - \frac{1}{20f^5} + \frac{17}{112f^7} - \frac{31}{36f^9} + \frac{691}{88f^{11}} - \frac{5\,461}{52f^{13}} + \frac{929\,569}{480f^{15}} - O\left(\frac{1}{f^{17}}\right) \tag{10}$$

3. Približný výpočet čísla π pomocou funkcie b_f

Existuje veľa veľmi rýchlo konvergujúcich radov [6], vhodných na výpočet čísla π. Nasledujúca možnosť bola použitá predovšetkým preto, aby sa dala spoľahlivo otestovať konvergentnosť rozvojov (7) a (10). Z vlastnosti gama funkcie [2] vyplýva, že pre párne f možno číslo π približne vypočítať pomocou funkcie b_f nasledovne:

$$\pi \approx \frac{2}{f}\ b_f^2 \left[\frac{2^{\frac{f}{2}} \left(\frac{f}{2}\right)!! \left(\frac{f}{2}-1\right)!!}{(f-1)!} \right]^2 \tag{11}$$

pričom $\left(\frac{f}{2}\right)!!$ a $\left(\frac{f}{2}-1\right)!!$ sú dvojné faktoriály a zo vzťahu (10) vyplýva, že

$$b_f \approx Exp\left(-\frac{1}{4f} + \frac{1}{24f^3} - \frac{1}{20f^5} + \frac{17}{112f^7} - \frac{31}{36f^9} + \frac{691}{88f^{11}} - \frac{5\,461}{52f^{13}} + \frac{929\,569}{480f^{15}} \right) \tag{12}$$

Číslo π sa dá približne vypočítať v systéme Mathematica [5] podľa vzťahu (11) najviac pre $f = 86\,181\,396$ pomocou programu č. 3. Dosiahnutá presnosť je 140 platných cifier. Podobnú problematiku možno nájsť v prácach [7], [8], [9] a [10.]

Tento článok vznikol s podporou grantových projektov VEGA č. 1/3182/06 Zlepšovanie kvality produkcie strojárskych výrobkov pomocou moderných štatistických metód a VEGA č. 1/1247/04 Progresívne štatistické techniky a rozhodovanie v procese zlepšovania kvality.

Literatúra

[1] LIKEŠ, J., MACHEK, J. *Matematická statistika*. Praha: SNTL, 1983. 180 s. (in Czech).

[2] DWIGHT, H. B. *Tables of Integral and other Mathematical Data. Fourth Edition, New York 1961.*

[3] KENDALL. M. G., STUART, A. *The Advanced Theory of Statistics. Vol 1. Distributions Theory. 2nd Ed. London, Griffin 1962.*

[4] KNUTH, D. E., VARDI, I. RICHBERG, R. The asymptotic expansion of the middle binomial coefficient. Amer. Math. Monthly 97, 1990, p. 626-630.

[5] WOLFRAM, S. *The Mathematica Book.* 3rd ed. Wolfram Media/Cambridge University Press, 1996. 1403 p. ISBN 0-521-58889-8.

[6] BECKMANN. P. A. *History of* π. 5. vydanie, THE GOLEM PRESS 1982.

[7] JANIGA, I., MIKLÓŠ, R. Statistical Tolerance Intervals for a Normal Distribution. In *Measurement Science Review*. ISSN 13, 2001, vol. 1, no. 1, p. 29-32.

[8] GARAJ, I., JANIGA I. *Dvojstranné tolerančné medze pre neznámu strednú hodnotu a rozptyl normálneho rozdelenia*. Bratislava: Vydavateľstvo STU, 2002. 147 s. ISBN 80-227-1779-7.

[9] GARAJ, I., JANIGA I. *Dvojstranné tolerančné medze normálnych rozdelení s neznámou strednou hodnotou a spoločným rozptylom. Two Sided Tolerance Limits of Normal Distributions with Unknown Means and Unknown Common Variability.* Bratislava, Vyd. STU, 2004, 218 s. ISBN 80-227-2019-4.

[10] GARAJ, I., JANIGA I. Jednostr*anné tolerančné medze normálneho rozdelenia s neznámou strednou hodnotou a rozptylom. One Sided Tolerance Limits of Normal Distributions with Unknown Mean and Variability.* Bratislava:. STU, 2005, 214 s. ISBN 80-227-2218-9.

Dodatok. Použité programy v systéme Mathematica

Program č. 1. Výpočet konštánt c_i $(i = 1, 2,...,16)$ pomocou vzťahu (6)

```
c[1] = 1;
a1 = 2^(i - 2) * Binomial[2 * i, i];
a2 = -0.5 * Sum[c[j] * 2^(3 * i - 3 * j) * Binomial[i - 1, j - 1], {j, 1, i - 1}];
a3 = Sum[-c[k] * Sum[c[j] * 2^(3 * i - 3 * j - 3 * k) * Binomial[i - k - 1, j - 1],
      {j, 1, i - k}], {k, 1, i - 1}];
c[i] = a1 + a2 + a3;
v4 = SetPrecision[Table[Flatten[{i, c[i] = a1 + a2 + a3}], {i, 1, 16}], 16];
bb = TableForm[v4, TableSpacing → {0, 1}, TableAlignments → Center]
```

Program č. 2. Rozvoj funkcie $\ln(b_f)$

```
t = -x / 4 + (x^2) / 32 + (x^3) * (5/128) - (x^4) * (21 / 2048) -
   (x^5) * (399 / 8192) + (x^6) * (869 / 65536) + (x^7) * (39325 / 262144) -
   (x^8) * (334477 / 8388608) - (x^9) * (57434806) / (2)^26 +
   (x^10) * (119394366) / (2)^29 + (x^11) * (33601489740) / (2)^32 -
   (x^12) * (68858583810) / (2)^35 - (x^13) * (28797022447980) / (2)^38 +
   (x^14) * (58526378304180) / (2)^41 + (x^15) * (34009655736503400) / (2)^44 -
   (x^16) * (68787420596367128) / (2)^47
Series[Log[1 + t], {x, 0, 16}]
```

Program č. 3. Výpočet čísla π pomocou funkcie b_f

```
r2 = (((2)^(f / 2)) * (((f / 2)!!) * (f / 2 - 1)!!) / ((f - 1)!!))^2;
r1 = (-1 / (4 * f) + 1 / (24 * f^3) - 1 / (20 * f^5) + 17 / (112 * f^7) -
     31 / (36 * f^9) + 691 / (88 * f^11) - 5461 / (52 * f^13) +
     929569 / (480 * f^15));
pi = SetPrecision[(r2 * Exp[2 * r1]) * (2 / f), 150];
v4 = Table[Flatten[{f, pi}], {f, 86181398, 86181398}];
bb = TableForm[v4, TableSpacing → {0, 1}, TableAlignfents → Center]
```

Kontaktná adresa autora:

RNDr. Ivan Garaj, PhD.,
Ústav informatizácie, automatizácie a matematiky,
FCHPT STU, Radlinského 9, 812 37 Bratislava,
Tel.: +421-2-59325 297, E-mail: ivan.garaj@stuba.sk

Bankruptcy prediction in Slovak companies using linear probability models[1]

Rudolf Gavliak

Abstract: In this contribution we show the possible application of linear probability models (LPM) for bankruptcy prediction in Slovak condition. We suppose the most significant factor, which is determining the company surviving, to be the company financial situation. According to this assumption, we've used the chosen financial analysis ratios as the independent variables, to explain the zero-one dependent variable, representing the bankruptcy occurrence. We have found the best fitting model for logit, probit and gompit models class and compared the ability of the models to discriminate among the bankrupting and non-bankrupting companies in selected data set.
Key words: Bankruptcy prediction, LPM models, Logit, Probit, Gompit, Likelihood estimation.

1. Introduction

We make use of binary linear probability models to predict probability of bankruptcy in a group of Slovak small and medium sized companies. This contribution is connected to already published paper concerning applied logit models to solve the firm insolvency problem. In this contribution we use binary dependent variable models with not only logit, but also with probit and gompit (extreme value) link function, to handle the bankruptcy probability. We suppose the extreme value link function to better fit the zero – one insolvency. Now let's shortly introduce the theoretical background of estimated models.

2. Estimation techniques and procedures

In models we used in this contribution, the dependent variable, may have only two values (in our case one for bankruptcy and zero otherwise) There is a strong assumption of the bankruptcy to be influenced by different financial analysis ratios (independent variables). The goal is to quantify the relationship between the individual financial characteristics and the probability of become insolvent.

A simple linear regression of is not appropriate, since among other things, the implied model of the conditional mean places inappropriate restrictions on the residuals of the model (e_t). Furthermore, the fitted value ($\hat{y}$) from a simple linear regression is not restricted to lie between zero and one. So instead of the dummy variable as dependent variable, we replace it with the probability of default:

$$P\left(y_t = 1 \middle| x_t, \beta\right) = 1 - F\left(-x_t^T \beta\right) \tag{1}$$

The function *F* is a continuous, strictly increasing function, which takes a real value and returns a value ranging from zero to one. The choice of the function determines the type of binary model. It results, that the probability of survival of a company is following:

$$P\left(y_t = 0 \middle| x_t, \beta\right) = F\left(-x_t^T \beta\right) \tag{2}$$

Model with such a specification can't be estimated using standard least squares procedures. But we can estimate the parameters of this model using the method of maximum likelihood. In this case the likelihood function is given by following specification:

$$\log\left(\beta\right) = \sum_{t=1}^{n} y_t \log\left(1 - F\left(-x_t^T \beta\right)\right) + \left(1 - y_t\right) \log\left(F\left(-x_t^T \beta\right)\right) \tag{3}$$

[1] This paper was presents the results of Faculty of Economics UMB research grant FG 67

The first order conditions for this likelihood are nonlinear so that obtaining parameter estimates requires an iterative solution. We used a second derivative method for iteration and computation of the covariance matrix of the parameter estimates.
There are two alternative interpretations of this specification. The model divides the results space into two subspaces. Suppose that there is an unobserved latent variable that is linearly related to explanatory variables:

$$y_t^* = x_t^T \beta + e_t \tag{4}$$

According to the values of unobserved latent variable we can decide, whether the object can be classified to one or another subspace:

$$y_t \begin{cases} 1, & y_t^* = x_t^T \beta + e_t > 0 \\ 0, & y_t^* = x_t^T \beta + e_t \le 0 \end{cases}, \quad t = 1, 2 \ldots n \tag{5}$$

In this case, the classification threshold is set to zero, but the choice of a threshold value is irrelevant, because if constant term is included in regressor vector x_t^T, the notation in (5) becomes a general form. Then the probability of default could be rewritten in form:

$$P\left(y_t = 1 \middle| x_t, \beta\right) = P\left(y_t^* > 0\right) = P\left(x_t^T \beta + e_t > 0\right) = F_e\left(-x_t^T \beta\right) \tag{6}$$

The notation F_e denotes the cumulative distribution function of the residuals (e_t). In our case study we will use the probit (standard normal), logit (logistic), and gompit (extreme value) specifications for the link function *F*.
Coding the dependent variable as zero – one variable implies that the expected value of *y* is simply the probability that *y = 1*:

$$E\left(y_t = 1 \middle| x_t, \beta\right) = 1 \cdot P\left(y_t = 1 \middle| x_t, \beta\right) + 0 \cdot P\left(y_t = 0 \middle| x_t, \beta\right) \tag{7}$$

It follows that we can write the binary model as a regression model:

$$y_t = \left(1 - F\left(-x_t^T \beta\right)\right) + \varepsilon_t . \tag{8}$$

In this equation the series ε_t denotes the deviations of y_t from its conditional mean. The basic properties of the deviations are:

$$\begin{aligned} &E\left(\varepsilon_t \middle| x_t, \beta\right) = 0, \\ &D\left(\varepsilon_t \middle| x_t, \beta\right) = F\left(-x_t^T \beta\right) \cdot \left(1 - F\left(-x_t^T \beta\right)\right). \end{aligned} \tag{9}$$

The difference among the logit, probit and gompit models is different distribution of the error term in equation (6). The probability of firm insolvency in the probit models class is computed in following manner:

$$P\left(y_t = 1 \middle| x_t, \beta\right) = 1 - \Phi\left(-x_t^T \beta\right) = \Phi\left(x_t^T \beta\right) = \Phi\left(\hat{y}\right) \tag{10}$$

The sign of common cumulative distribution function changes to cumulative distribution function of standard normal distribution (Φ). So if the parameters vector β is known, the default probability could be easily estimated using standard normal distribution tables. The probability of insolvency in the logit models class differs in this way:

$$P\left(y_t = 1 \middle| x_t, \beta\right) = 1 - \left(\frac{e^{-x_t^T \beta}}{1 + e^{-x_t^T \beta}}\right) = \frac{e^{x_t^T \beta}}{1 + e^{x_t^T \beta}} = \frac{e^{\hat{y}}}{1 + e^{\hat{y}}}. \tag{11}$$

The probability of default computation is based upon the cumulative distribution function for the logistic distribution.
The last possibility to cumulative distribution function definition is the extreme value specification:

$$P\left(y_t = 1 \middle| x_t, \beta\right) = 1 - \left(1 - \exp\left(-e^{-x_t^T \beta}\right)\right) = \exp\left(-e^{-x_t^T \beta}\right) = e^{-e^{-\hat{y}}}. \tag{12}$$

This means, that the cumulative distribution function has the Gompertz curve shape, it is important to note, that such distribution is skewed. In next section we present our input data, the data frequency and received estimation results.

3. Data and estimation results

The basic input data set contains of financial data obtained from 856 Slovak companies active in different industries across the economy structure. For each of the company we know the values of 36 financial ratios for the years 2002 until 2005. Hiadlovský and Kráľ (2006) have shown that only 18 of these indicators have statistically significant discriminant capability. The labels and the formulas used for construction of these financial indicators are introduced in table 1.

Table 1. Specification of input variables (ratios) for LPM construction

Label	Characteristic/Variable	Label	Characteristic/Variable
Y	dummy (zero-one) variable representing financial insolvency appearance	KCZOM	Current assets / Total debt
OKEC	classification according to prevailing business activity (ordered variable)	KKZOM	Current assets / Short term liabilities
BU_CZ	Banking loans / Total debt ratio	KZ_MC	Current assets / Total assets
CZA	Total debt / Total assets	L_CF	Liquidity calculated with cash flow
DOZ	Inventory / Daily sales	L1	Cash Ratio
DZA	Long term debt / Total assets	L2	Quick ratio
EBIT_CK	EBIT / Total assets	L3	Current ratio
FP1	Financial leverage measure	OA	Assets turnover ratio
HVB_KZ	EBT / Short term liabilities	STZ	Total assets / Total debt
KCZBCF	Cash flow / Total debt	UZA	Loans / Total assets

Source: INFIN, s.r.o., Bratislava

The numbers of bankrupting companies in years 2002-2005 is 15 and the companies are active in different industries. To assure the comparability among the bankrupting and successful companies we selected from the surviving companies only 67 companies from the same industries, as the less successful companies come from. The final data set also comprises financial indicators described in table 1 for 82 companies, where the insolvent firms take 18,29 percent and surviving 81,71 percent share[2].

4. Empirical results

In this section we will present three models, the best data fitting model for each respective class of used models (logit, probit, gompit). We decided the fitting criterion to be represented by Schwarz criterion, the Hannan – Quinn info criterion and McFadden R – squared. There is of course the request for statistical validity of the model as whole. The statistical significance of binary linear probability models is tested by following likelihood ratio statistic:

$$LR = -2\left(l\left(\tilde{\beta}\right) - l\left(\beta\right)\right) \sim \chi_k^2 \qquad (13)$$

The test statistic compares the restricted log likelihood function $l\left(\tilde{\beta}\right)$ and the log likelihood function of estimated model. Restricted log likelihood is the maximized log likelihood value, when all slope coefficients in estimated model are restricted to zero. Since the constant term is included, this specification is equivalent to estimating the unconditional mean probability of "insolvency". If the difference among the log likelihood functions is rather big we will more likely to reject the zero hypothesis of no jointly significance of the estimated model. The test

[2] We would like to thank to INFIN, s.r.o. for providing data and further helpful recommendation.

statistic is chi-squared distributed with degrees of freedom equivalent to number of independent variables included to model (excepting the constant term). We followed also the request of statistical significance of single parameters. In the three presented models, all estimated parameters are significant.
The best model in the class of logit models according to mentioned criteria is:

$$P\left(y_t = 1 \middle| x_t, \beta\right) = \frac{e^{x_t^T \beta}}{1 + e^{x_t^T \beta}}, \quad McFadden\ R^2 = 55,08\%$$

$$x_t^T \beta = 7,49 - 5,23L3 - 3,26\log(OKEC) - 7,39KCZBCF + 0,23EBIT\_CK \qquad (14)$$

$$z_{b_j} \quad (3,23)\ (-3,98)\ (-2,72) \qquad (-2,14) \qquad (3,44).$$

The best fitting model, if the link function was changed to standardized normal cumulative distribution function has the following form:

$$P\left(y_t = 1 \middle| x_t, \beta\right) = \Phi\left(x_t^T \beta\right), \quad McFadden\ R^2 = 59,18\%$$

$$x_t^T \beta = 3,31 + 3,75L\_CF - 2,75L3 - 8,28HVB\_KZ - 1,51\log(OKEC) + 0,11EBIT\_CK \qquad (15)$$

$$z_{b_j} \quad (3,23)\ (3,58) \qquad (-3,77) \qquad (-3,54) \qquad (-2,68) \qquad (3,74).$$

The McFadden R^2 could be interpreted in the same manner, than the R^2 in a simple regression. So we can state, that the explained variability of the dependent variable increased to more than 59 percent.
We have assumed the extreme value binary models to be the most successful in discriminating the bankrupting and successful companies. The best data fitting gompit model can be formally written like:

$$P\left(y_t = 1 \middle| x_t, \beta\right) = \exp\left(-e^{-x_t^T \beta}\right), \quad McFadden\ R^2 = 65,80\ \%$$

$$\hat{y} = 2,99 - 1,17KKZOM - 8,94L1/L3 - 20,41HVB\_KZ + 16,15KCZBCF - 2,84KCZOM \qquad (16)$$

$$z_{b_j} \quad (3,38)\ (-3,62) \qquad (-2,66) \qquad (-2,64) \qquad (2,71) \qquad (-3,38).$$

The gompit model explains the biggest part of the variability of the dependent variable. The interpretation of the results is not as straightforward as in linear regression models.
Interpretation of the coefficient values is complicated by the fact that estimated coefficients from a binary model cannot be interpreted as the marginal effect (coefficients of elasticity) on the dependent variable. The marginal effect of x_j on the conditional probability is given by:

$$\frac{\partial E\left(y_t \middle| x_t, \beta\right)}{\partial x_{tj}} = \frac{dF\left(-x_t^T \beta\right)}{dx} \cdot \beta_j = f\left(-x_t^T \beta\right)\beta_j. \qquad (17)$$

The first derivative of the link function *F* is the density function *f*. The value of estimated parameter at *j*-th financial characteristic is weighted by the value of density function, which depends on the values of all of the regressors in input vector. The direction of the effect of a change in x_j depends only on the sign of the β_j coefficient. Positive values of β_j imply that increasing x_j will increase the probability of firm insolvency. Negative values of estimated parameter imply that increase of the selected repressor causes decrease in default probability. An alternative interpretation of the coefficients results from the fact, that the ratios of coefficients provide a measure of the relative changes in the probabilities.

5. Conclusion

For better comparison of estimated models let's present the ability of the models to correctly classify the companies in the mentioned two groups. The fitting ability is presented in expectation-prediction table. To construct the proposed contingency table, we have to enter

the cutoff probability. If the conditional probability of bankruptcy is higher than the specified cutoff value, the company is classified as bankrupting. The classification results are shown in table 2.

Table 2. Classification results for best fitting logit, probit and gompit model

Binary Logit				**Binary Probit**				**Binary Extreme Value**			
Prediction Evaluation (success cutoff C = 0,35)				Prediction Evaluation (success cutoff C = 0,35)				Prediction Evaluation (success cutoff C = 0,3)			
	Estimated Equation				Estimated Equation				Estimated Equation		
	Dep=0	Dep=1	Total		Dep=0	Dep=1	Total		Dep=0	Dep=1	Total
P(Dep=1)<=C	62	3	65	P(Dep=1)<=C	63	2	65	P(Dep=1)<=C	62	1	63
P(Dep=1)>C	5	12	17	P(Dep=1)>C	4	13	17	P(Dep=1)>C	5	14	19
Total	67	15	82	Total	67	15	82	Total	67	15	82
Correct	62	12	74	Correct	63	13	76	Correct	62	14	76
% Correct	92,54	80,00	90,24	% Correct	94,03	86,67	92,68	% Correct	92,54	93,33	92,7
% Incorrect	7,46	20,00	9,76	% Incorrect	5,97	13,33	7,32	% Incorrect	7,46	6,67	7,32

Source: software output and own calculations

As seen the best classification results were obtained, if the probability cutoff point was set to 0,35 for logit and probit model and 0,3 for the extreme value (gompit) model. The classification performance of the logit model is significantly lower, than the performance of the other two models. One fifth of badly classified bankrupting companies are quite dangerous results, because an insolvent firm, which considered prospering is a source of potential losses for financial institutions. The other two models have the same total classification success rate. The difference is, that extreme value model failed to correctly diagnose just one bankrupting company (for cost of badly classifying one prospering company).

6. References

[B.a.] 2005. EViews 5.1 User's Guide. Irvine : Quantitative Micro Software, 2005. dostupné na internete: http://www.eviews.com/eviews5/eviews5/EViews51Manuals.zip (19.11.2006)

FICZOVÁ, I. – SEDLÁČEK, J. – ÚRADNÍČEK, V. 2002. Finančno-ekonomická analýza podniku. Praktikum. Časť I. Banská Bystrica : OZ Financ, 2002. ISBN: 80-968702-1-1.

GREENE, W.H. 1997. Econometric Analysis. New Jersey : Prentice Hall, Upper Saddle River, 1997.

GUJARATI, N. D. 2003. Basic Econometrics. 4th Edition. New York : McGraw-Hill, 2003.

HIADLOVSKÝ, V. – KRÁĽ, P. 2005. Využitie diskriminačnej analýzy na predikovanie finančnej situácie podnikov v SR. In: Forum Statisticum Slovacum, No. 1, 2005, p. 44-50. ISSN 1336-7420.

HIADLOVSKÝ, V. – KRÁĽ, P. 2006. Možnosti predikovania finančnej situácie podnikov v SR s využitím SPSS. In: Forum Statisticum Slovacum, No. 4, 2006, p. 90-95. ISSN 1336-7420.

HUŠEK, R. 1999. Ekonometrická analýza. Praha : Ekopress, 1999. ISBN 80-86119-19-X.

ZALAI, K. A KOL. 2002. Finančno-ekonomická analýza podniku. 4. doplnené vyd. Bratislava : Sprint, 2002. ISBN: 80-88848-96-6.

Authors address:

Ing. Rudolf Gavliak, Department of Quantitative Methods and Informatics, Faculty of Economics UMB,Tajovského 10, 975 90 Banská Bystrica, rudolf.gavliak@umb.sk

Diskrétní pravděpodobnostní rozdělení v MS Excel

Michal Vrabec, Luboš Marek

Abstract
The main aim of this paper is the describing of probability discrete distributions in MS Excel software. The each probability distribution is described at first in theoretical level (including formulas for mean and variance) and then the method of calculation for probability function and distribution function is following (including the exact syntax).

Key words
MS Excel, probability distribution, probability function, distribution function, percentile, critical value.

Mezi statistickou obcí se často diskutuje, který statistický program je nejlepší, přičemž se tyto programy posuzují z různých hledisek. Každý, kdo někdy pracoval s několika různými programy, ví, že na tuto jednoduchou otázku není jednoznačná odpověď. Každý z programů má totiž některé slabší a některé silnější stránky, a tak asi nikoho nepřekvapí, když uvedeme, že statistický program si často vybíráme až na základě úlohy, kterou chceme zpracovat. Podmínkou pochopitelně je, aby bylo z čeho vybírat.
Mezi programy, které by asi málokdo zařadil mezi statistické, patří i statistiky často odmítaný MS Excel. Přitom, pokud pomineme nekvalitní místy až zoufalý překlad z angličtiny v oblasti statistiky, se jedná o program, který umožňuje kvalitní aplikaci různých procedur, mnohdy na úrovni srovnatelné s procedurami ve specializovaných (a také daleko dražších) statistických programech. Umožňuje např. velmi snadnou aplikaci statistických funkcí z nejrůznějších oblastí. Nesmíme totiž zapomínat, že právě jednoduché ovládání a dostupnost (tento tabulkový procesor je dnes instalován téměř na každém počítači) je velkou devizou tohoto programu. V tomto článku bychom chtěli zhodnotit MS Excel z hlediska pravděpodobnostních rozdělení. V první řadě zhodnotíme nabídku pravděpodobnostních rozdělení v tomto programu. U každého rozdělení popíšeme možnosti výpočtů hodnot distribuční funkce a kvantilů a syntaxi příslušných funkcí. U každého spojitého rozdělení uvedeme obrázek s ukázkou průběhu hustoty pravděpodobnosti pro konkrétní parametry (je ovšem třeba ihned poznamenat, že při jiné volbě parametrů bychom obdrželi jiný tvar hustoty). V závěru rovněž zhodnotíme možnosti generování náhodných hodnot z pravděpodobnostních rozdělení.

V oblasti diskrétních (nespojitých) rozdělení obsahuje MS Excel následující rozdělení, u kterých zároveň uvádíme název příslušné funkce:

rozdělení	distribuční funkce	pravděpodobnostní funkce	kvantily
binomické	BINOMDIST	BINOMDIST	CRITBINOM
negativně binomické	NE	NEGBINOMDIST	NE
Poissonovo	POISSON	POISSON	NE
hypergeometrické	NE	HYPGEOMDIST	NE

Jedná se tedy o naprosto o základní typy rozdělení, navíc ne vždy je možné spočítat distribuční funkci a kvantily. To ale není žádné neštěstí, neboť oboje jsme schopni poměrně snadno spočítat z hodnot pravděpodobnostní funkce.
Podívejme se nyní na jednotlivá rozdělení podrobněji.

Binomické rozdělení

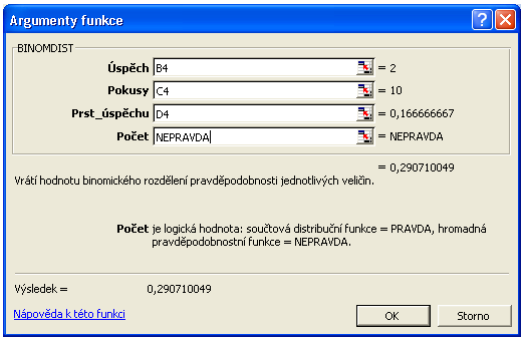
Náhodná veličina X má binomické rozdělení s parametry n a π, jestliže její pravděpodobnostní funkce má tvar

$$P(x)=\binom{n}{x}\pi^x(1-\pi)^{n-x},\qquad x=0,1,\ldots,n,\quad 0<\pi<1,\quad n\in\mathrm{N}.$$

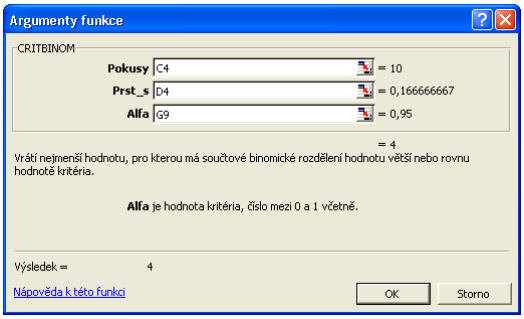
Střední hodnota a rozptyl mají tvar:

$$E(X)=n\pi \qquad\qquad D(X)=n\pi(1-\pi).$$

V Excelu se jak pro distribuční funkci i pro pravděpodobnostní funkci používá funkce BINOMDIST. Její argumenty mají následující význam:



Úspěch - x (počet úspěchů). Hodnota, ve které počítáme F(x) či P(x).
Pokusy - n (počet pokusů).
Prst_úspěchu - π. Pravděpodobnost úspěchu.
Počet - NEPRAVDA pro hodnotu pravděpodobnostní funkce $P(x)$, PRAVDA pro hodnotu distribuční funkce $F(x)$.



Jako pro jediné z nespojitých rozdělení je v Excelu uvedena i funkce pro výpočet kvantilů: CRITBINOM. Její argumenty mají obdobný význam, jako u funkce BINOMDIST.
Pokusy - n (počet pokusů).
Prst_s - π. Pravděpodobnost úspěchu.
Alfa - pravděpodobnost P pro hodnotu kvantilu x_p.

Negativně binomické rozdělení

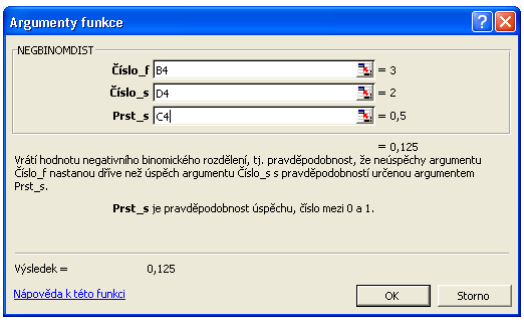
Náhodná veličina X má negativně binomické rozdělení s parametry n a π, jestliže její pravděpodobnostní funkce má pro n celočíselné tvar

$$P(x)=\binom{n+x-1}{x}\pi^n(1-\pi)^x,\qquad x=0,1,\ldots\quad 0<\pi<1,\quad n\in\mathrm{N}.$$

Připomeňme, že pro přirozená n můžeme náhodnou veličinu X chápat jako počet neúspěchů před n-tým úspěchem. Úmyslně uvádíme pravděpodobnostní funkci ve zjednodušeném tvaru, neboť takto je chápána v Excelu.

Střední hodnota a rozptyl mají tvar:

$$E(X)=\frac{n(1-\pi)}{\pi} \qquad\qquad D(X)=\frac{n(1-\pi)}{\pi^2}.$$



V Excelu se pro pravděpodobnostní funkci používá funkce NEGBINOMDIST. Její argumenty mají následující význam:
Číslo_f - x (počet neúspěchů před n-tým úspěchem). Hodnota, ve které počítáme $F(x)$ či $P(x)$.

Číslo_s - n (počet pokusů).
Prst_s - π. Pravděpodobnost úspěchu.

Hodnoty distribuční funkce je nutné napočítat z hodnoty pravděpodobnostní funkce, stejně tak i hodnoty kvantilů.
Speciálním případem negativně binomického rozdělení je rozdělení geometrické. Toto rozdělení obdržíme velmi snadno, pokud v negativně binomickém rozdělení položíme n=1. Potom se předchozí pravděpodobnostní funkce zjednoduší do tvaru

$$P(x) = \pi(1-\pi)^x, \qquad x = 0,1, \ldots, \quad 0 < \pi < 1.$$

Náhodnou veličinu X lze potom chápat jako počet neúspěchů před prvním úspěchem.
Pro střední hodnotu a rozptyl obdržíme

$$E(X) = \frac{1-\pi}{\pi} \qquad D(X) = \frac{1-\pi}{\pi^2}.$$

V Excelu použijeme funkci NEGBINOMDIST, ve které položíme Číslo_s = 1.

Poissonovo rozdělení

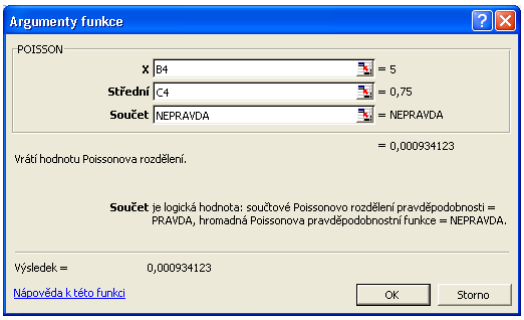
Náhodná veličina X má Poissonovo rozdělení s parametrem λ, jestliže její pravděpodobnostní funkce má tvar

$$P(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \qquad x = 0,1, \ldots, \quad \lambda > 0.$$

Střední hodnota a rozptyl mají tvar:

$$E(X) = \lambda \qquad D(X) = \lambda.$$

V Excelu se jak pro distribuční funkci i pro pravděpodobnostní funkci používá funkce POISSON. Její argumenty mají následující význam:
X - x. Hodnota, ve které počítáme $F(x)$ či $P(x)$.
Střední - λ. Parametr a zároveň střední hodnota rozdělení.
Součet - NEPRAVDA pro hodnotu pravděpodobnostní funkce $P(x)$, PRAVDA pro hodnotu distribuční funkce $F(x)$.



Hypergeometrické rozdělení

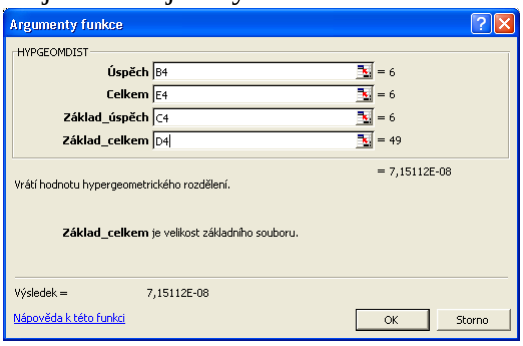
Náhodná veličina X má hypergeometrické rozdělení s parametry N, M a n, jestliže její pravděpodobnostní funkce má tvar

$$P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad x = \max(0, M-N+n), \ldots, \min(M, n).$$

Přitom N, M a n jsou přirozená čísla, $1 \le n < N$, $1 \le M < N$.
Střední hodnota a rozptyl mají tvar

$$E(X) = n\frac{M}{N} \qquad D(X) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}.$$

V Excelu se pro pravděpodobnostní funkci používá funkce HYPGEOMDIST. Její argumenty mají následující význam:



Úspěch - x. Hodnota, ve které počítáme $P(x)$.

Celkem - n (rozsah výběru).

Základ_úspěch - M. Počet prvků s vlastností M.

Základ_celkem - N (rozsah základního souboru).

Literatura:

Nápověda k programu MS Excel
Nápověda k programu Statgraphics Centurion

Doc. RNDr. Luboš Marek, CSc.
Katedra statistiky a pravděpodobnosti
Fakulta informatiky a statistiky
Vysoká škola ekonomická Praha
marek@vse.cz

Mgr. Michal Vrabec
Katedra statistiky a pravděpodobnosti
Fakulta informatiky a statistiky
Vysoká škola ekonomická Praha
vrabec@vse.cz

Vzdialenosti medzi hodnotiteľmi a zhluky hodnotiteľov

Jozef Chajdiak,[1]

Abstract: The paper deals with analysis of evaluation Slovak macro economy from experts. The experts are clustering into the clusters by their evaluations.

Keywords: Macro economy, evaluation, distance, cluster, Excel, SAS, Ward method

Jednou zo súčastí konferencie Pohľady na ekonomiku Slovenska 2006, ktorá sa uskutočnila 4. 4. 2006 v Bratislave bolo expertné hodnotenie stavu a vývoja makroekonomiky Slovenska zúčastnenými prednášajúcimi. Hodnotenie sa zakladalo na posúdení hodnôt piatich ukazovateľov:
X1 – hrubý domáci produkt na obyvateľa,
X2 – inflácia,
X3 – miera nezamestnanosti,
X4 – saldo štátneho rozpočtu k HDP,
X5 – saldo zahraničného obchodu k HDP.

Na hodnotenie stavu a vývoja ukazovateľov sa použila bodová stupnica hodnotení od -2 (veľmi zle) cez 0 (neutrálny stav alebo vývoj) po +2 (veľmi dobre) s krokom po pol bode. Hodnotil sa stav v roku 2005 a vývoj v roku 2005 oproti roku 2004. Intenzita počtu bodov sa prideľovala v porovnaní so stavom a vývojom v EÚ. Výsledné hodnotenia boli nasledujúce:

stav05	x1	x2	x3	x4	x5
Gábriš M.	-2	-1	-1	-1	-2
Haluška J.	-2	-1	-2	-1	-2
Chajdiak J.	-2	-1	-2	-1	-1
Olexa M.	-2	-1	-2	-1	-2
Pálenik V.	-2	-1	-2	0	1
Ševcovic P.	-2	-1	-2	0	-1
Tóth J.	-2	1	-2	-1	-1

vyvoj0504	x1	x2	x3	x4	x5
Gábriš M.	2	-1	1	1	-1
Haluška J.	2	2	1	2	-2
Chajdiak J.	2	-1	2	1	-1
Olexa M.	2	2	2	1	-2
Pálenik V.	2	1	1	2	0
Ševcovic P.	2	1	1	1	-1
Tóth J.	2	1	2	2	-1

[1] Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/2631/05 „Analýza možnosti aplikácie viacrozmerných štatistických metód na skúmanie ekonomických výsledkov na príklade priemyslu SR prípadne iných oblastí ekonomiky“

Na posúdenie homogenity súboru hodnotiteľov použijeme vzdialenosti medzi hodnoteniami jednotlivých dvojíc hodnotiteľov a hodnotiteľov zoskupíme do zhlukov. Na určenie vzdialenosti medzi dvoma hodnotiteľmi použijeme Mahalanobisovú vzdialenosť:

$$(\mathbf{x_i} - \mathbf{x_j})\, S^{-1}(\mathbf{x_i} - \mathbf{x_j})'$$

kde i a j predstavujú označenia hodnotiteľov, **x** je vektor bodových hodnotení ukazovateľov hodnotiteľom a S je výberová kovariančná matica vypočítaná z jednotlivých bodových hodnotení.

Na výpočet matice vzdialeností hodnotení medzi jednotlivými expertmi môžeme použiť Excel a v jeho rámci nástroj Covariation (vypočíta kovariačnú maticu) a z nej prenásobením jednotlivých prvkov matice konštantou n/(n-1) dostaneme výberovú kovariačnú maticu S. Na výpočet inverznej matice S^{-1} použijeme matematickú funkciu MINVERSE. K výpočtu rozdielov dvoch vektorov $(\mathbf{x_i} - \mathbf{x_j})$ a ich transponovanej hodnoty môžeme použiť priamo výpočtové vzorce v Exceli (operáciu mínus a operáciu rovná sa). Z vektorov rozdielov a inverznej matice použitím matematickej funkcie MMULT (dva krát – prvý krát rozdiel krát matica a druhý krát medzivýsledok krát transponovaný rozdiel) vypočítame Mahalanobisovú vzdialenosť medzi príslušnými dvoma vektormi hodnotení.

Kovariačná matica z hodnotení stavu je nasledujúca (Excel počíta kovariančnú maticu pre základný súbor s delením lomeno n):

	x1	*x2*	*x3*	*x4*	*x5*
x1	0	0	0	0	0
x2	0	0,49	-0	-0,1	0,04
x3	0	-0	0,12	-0	-0,1
x4	0	-0,1	-0	0,2	0,33
x5	0	0,04	-0,1	0,33	0,98

Z nej vypočítaná výberová kovariačná matica S je nasledujúca:

x1	*x2*	*x3*	*x4*	*x5*
0	0	0	0	0
0	0,57	-0,05	-0,1	0,05
0	-0,05	0,14	-0,05	-0,14
0	-0,1	-0,05	0,24	0,38
0	0,05	-0,14	0,38	1,14

Pri analýze bodových hodnotení hodnotiteľov vidíme úplnú zhodu všetkých hodnotiteľov pri stave ukazovateľa hrubý domáci produkt na obyvateľa (hodnotenia -2, t.j. veľmi zlý stav), čo prakticky znamená, že daný ukazovateľ nemá vplyv na vzdialenosti hodnotení medzi jednotlivými ukazovateľmi. Praktickým dôsledkom je, že inverznú maticu musíme počítať z nenulového zvyšku – z matice S vynecháme riadok a stĺpec X1 zodpovedajúci hodnoteniu hrubého domáceho produktu na obyvateľa. Vypočítaná inverzná matica je nasledujúca:

x2	*x3*	*x4*	*x5*
2,25	0,75	2,25	-0,75
0,75	8,25	0,75	0,75
2,25	0,75	11,3	-3,75
-0,75	0,75	-3,75	2,25

Nasleduje prácna časť výpočtu jednotlivých vzdialeností. Musíme vypočítať n*(n-1)/2 vzdialeností (t.j. 21 krát vypočítať vektor rozdielov, transponovať ho, príslušne ich prenásobiť s inverznou kovariančnou maticou) a nakoniec výsledky usporiadať do matice vzdialeností. Matica vzdialeností hodnotení medzi jednotlivými hodnotiteľmi je nasledujúca (hodnotitelia sú uvedení prvými písmenami svojho priezviska a mena):

stav05	gm	hj	ch	om	pv	sp	tj
gm	0	8,25	9	8,25	11,25	11,25	12
hj	8,25	0	2,25	0	9	6	8,25
ch	9	2,25	0	2,25	5,25	11,25	9
om	8,25	0	2,25	0	9	6	8,25
pv	11,25	9	5,25	9	0	9	11,25
sp	11,25	6	11,25	6	9	0	11,25
tj	12	8,25	9	8,25	11,25	11,25	0

Analogicky by sme mohli vypočítať maticu vzdialeností medzi hodnoteniami vývoja makroekonomiky Slovenska jednotlivými hodnotiteľmi:

vyv0504	gm	hj	ch	om	pv	sp	tj
gm	0	7,48	3,87	10,9	8,99	5,22	9,49
hj	7,48	0	11,45	9,59	10,74	9,29	7,58
ch	3,87	11,45	0	7,33	9,59	7,98	4,32
om	10,9	9,59	7,33	0	11,45	4,77	7,18
pv	8,99	10,74	9,59	11,45	0	5,17	5,62
sp	5,22	9,29	7,98	4,77	5,17	0	9,99
tj	9,49	7,58	4,32	7,18	5,62	9,99	0

Pri analýze vzdialeností jednotlivých hodnotiteľov pri stave makroekonomiky Slovenska vidíme, že na diagonále sú nuly (vzdialenosť hodnotiteľa od seba samého je nulová). Ďalej vidíme zhodu hodnotení Halušku J. a Olexu M. (vzdialenosť medzi nimi je nulová). Najbližšie k J. Haluškovi a M. Olexovi je Chajdiak J. (vzdialenosť 2,25). Najväčšia vzdialenosť hodnotení (12) je medzi Gábrišom M. a Tóthom J.

Z analýzy vzdialeností jednotlivých hodnotiteľov pri vývoji makroekonomiky Slovenska v roku 2005 oproti roku 2004 vidno, že najbližšie k sebe pri hodnotení vývoja majú Gábriš M. a Chajdiak J. (3,87) a najďalej dvojica Haluška J. a Chajdiak J. a dvojica Olexa M. a Páleník V. (11,45).

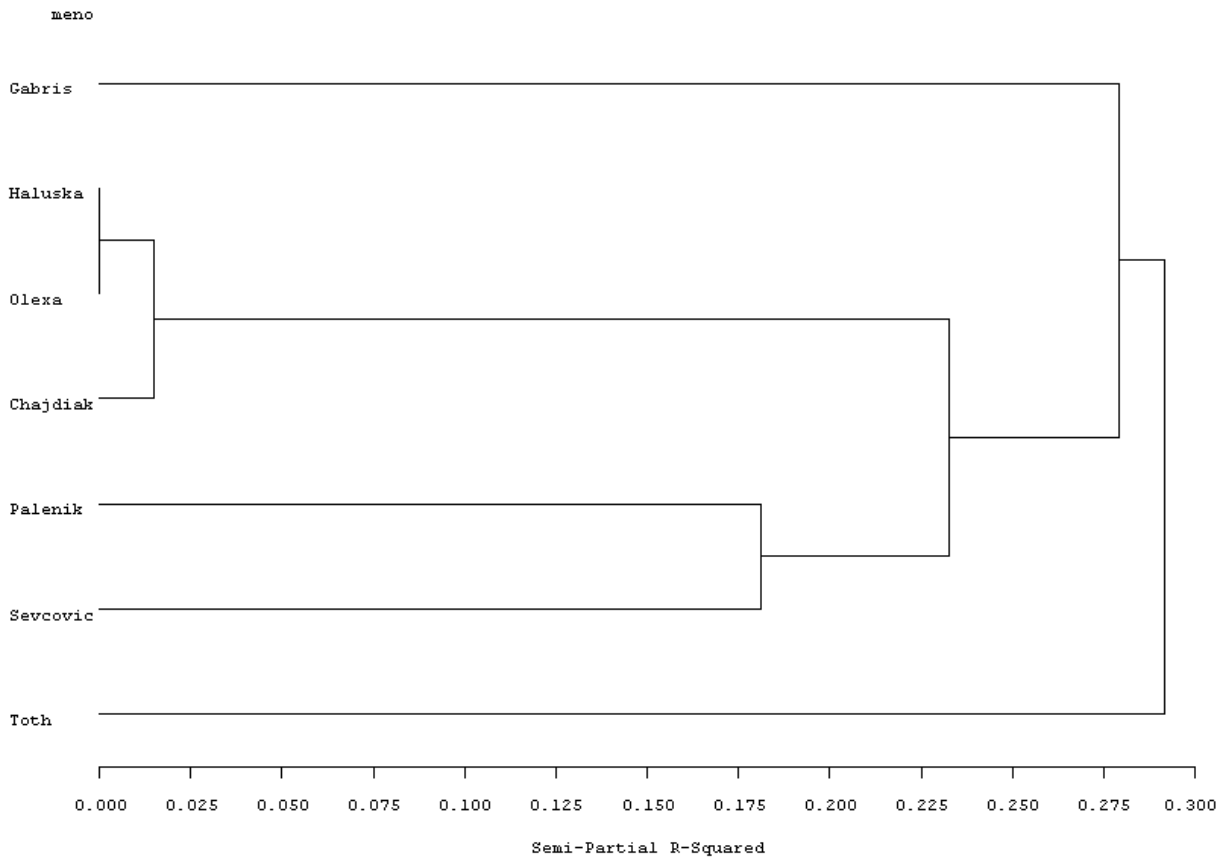
Matica vzdialeností je vhodným základom pre zoskupenie hodnotiteľov do príbuzných skupín – zhlukov. Systém Excel neobsahuje podsystém pre zhlukovú analýzu. Preto použijeme systém SAS a k vlastnému zhlukovaniu Wardovú metódu. Zápis programu v SAS-e je nasledujúci:

```
data pes05 (type=distance);
input (gm    hj     ch     om     pv     sp     tj) 6.2  @50 meno $12.;
cards;
0                                                        Gabriš M.
8.25   0                                                 Haluška J.
9.     2.25   0                                         Chajdiak J.
8.25   0.     2.25    0                                 Olexa M.
11.25  9.     5.25    9.      0                         Páleník V.
11.25  6.    11.25    6.      9.      0                 Ševčovic P.
12.    8.25   9.     8.25 11.25 11.25         0         Tóth J.
```

```
;
proc cluster data=pes05 method=ward pseudo;
id meno;
run;
proc tree horizontal space=2;
id meno;
run;
```

Výstupom je, okrem iného, dendrogram, na ktorom môžeme vidieť príbuznosť hodnotení jednotlivých expertov a ich zoskupenie do zhlukov.

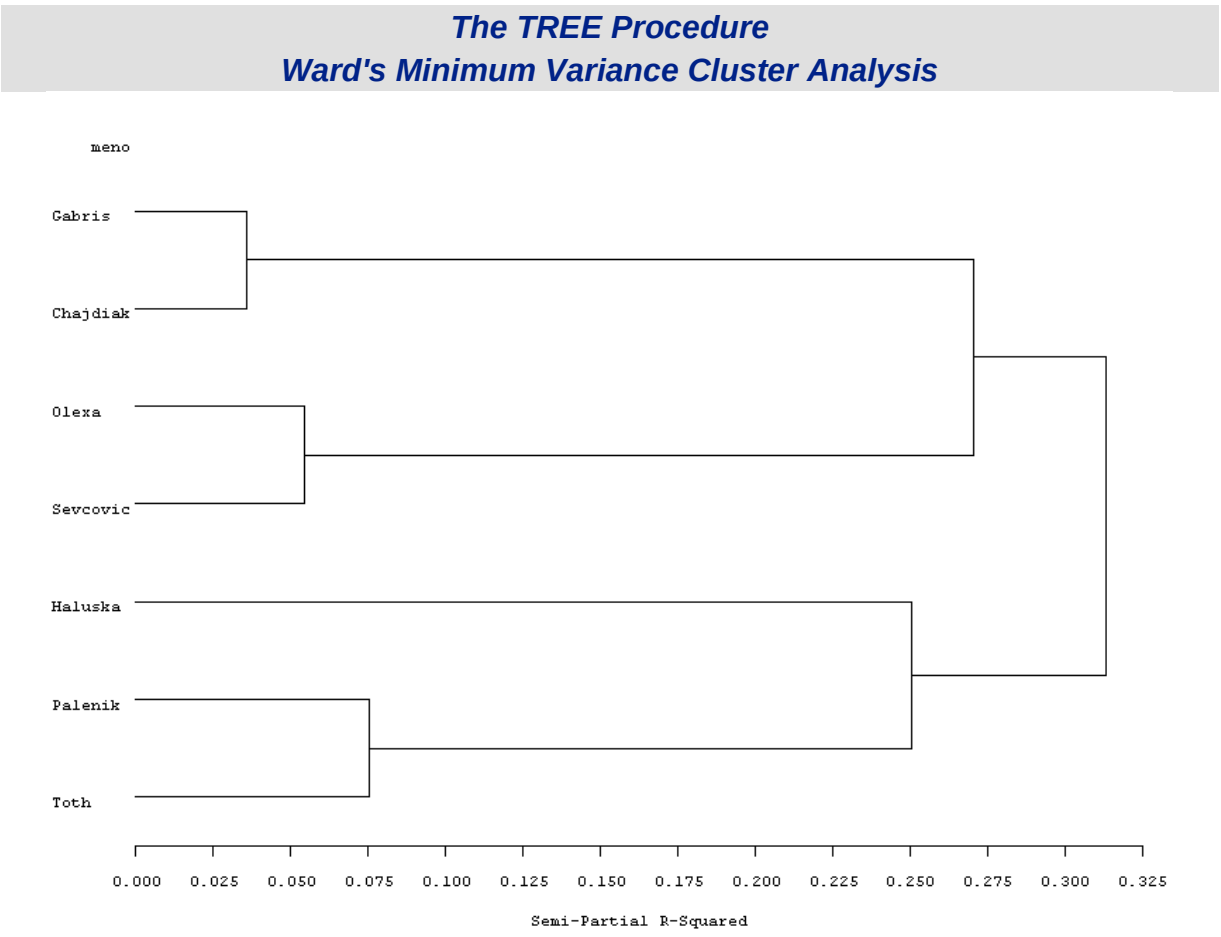
The TREE Procedure
Ward's Minimum Variance Cluster Analysis



Obr. 1 Dendrogram pre stav

Podľa výsledkov v dendrograme prvý zhluk tvoria hodnotitelia Haluška, Olexa a Chajdiak; do druhého zhluku možno zaradiť hodnotiteľov Páleník a Ševčovic. Tretí zhluk predstavuje hodnotiteľ Gábriš a štvrtý zhluk hodnotiteľ Tóth.

Zápis programu v SAS-e pre zoskupenie hodnotiteľov vývoja makroekonomiky je analogický. Časť výstupu obsahujúca dendrogram je na obr.2. Podľa vývoja máme tri zhluky hodnotiteľov. Prvý predstavujú Gábriš a Chajdiak, druhý Olexa a Ševčovic, tretí Haluška, Páleník a Tóth.



Obr.2 Dendrogram pre vývoj

Literatúra :

Expertné hodnotenie stavu a vývoja ekonomiky Slovenska za rok 2005 s oproti roku 2004. In: Pohľady na ekonomiku Slovenska 2006. Bratislava, SŠDS 2006, str. 4

Chajdiak, J.: Štatistické úlohy a ich riešenie v Exceli. Bratislava, Statis 2005

Chajdiak, J.: Štatistický analytický systém. Bratislava, Statis 1994

Luha J., Chajdiak J.: Pohľad na predošlé konferencie PODĽADY NA EKONOMIKU SLOVENSKA. Pohľady na ekonomiku Slovenska 2006. Bratislava, SŠDS 2006

Vojtková, H.: Priestorová klasifikácia priemyselných podnikov podľa efektívnosti. In. Forum Statisticum Slovacum 1/2006, Bratislava, SŠDS 2006

Berčačinová, Z.: Kvantitatívna analýza ekonomických výsledkov lízingových spoločností. In: Štatistické metódy vo vedecko-výskumnej práci 2003, Bratislava, SŠDS, 2003

Berčačinova Z.: Zhlukovanie lízingových spoločností podľa veľkosti. In: FernStat 2004, Tajov, SŠDS 2004

Adresa autora:

Doc. Ing. Jozef Chajdiak, CSc., Statis, Bratislava
chajdiak@statis.biz

Relationship between Waterjet Cutting Speed and Relevant Quantities Influencing Cutting Quality

Ivan Janiga, Marek Káll, Vojtech Geleta

Abstract. The contribution deals with the relationship between cutting speed abrasive Waterjet and relevant quantities. Problem solving method is regression analysis, and selection of variables and model building.

1 Introduction

Waterjet cutting has one basic positive attribute: it is a „cold“ production operation with minimum of invasive (power, temperature) effects on machined material with minimal risk of formatting auto-created vibrations, exact cut and minimum of waste. At the beginning, Waterjet cutting was used only for cleaning. Working pressure of water was 100 Mpa and the area of use was chemical and food industry. For cutting materials (metals, plastics, gum) high-speed Waterjets were used, with high pressure 450 MPa in combination with abrasive pieces, which are added to a mixing post located before the exit point.

Further, the paper focuses on the cutting speed specification, which should be set in the expert system of the machine on the basis of relevant values as material width H [mm], jet diameter P [mm] and flow of abrasive A [g/min] at tool steel cutting and medium quality of the cut. Problem solving method is regression analysis and selection of variables and model building.

2 Problem solving

For finding the relationship between abrasive Waterjet cutting speed and relevant values a multiple regression model can be used. The full model consists of the dependent variable or response R and regressor variables *H, P, A, HH, PP, AA, HP, HA, PA.* When applying the full regression model to all data, the least squares estimation of the regression yields the adjusted R^2 only $R^2_{\text{adj}} = 0.874$, that is a poor result. Therefore we decided to investigate the relationship between the response variable *R* and a one of the quantities *H, P, A* provided that the three others were fixed at constant values. It can be seen from the results (see Fig. 1, Fig. 2, Fig. 3) that only the material width *H* influences the cutting speed non-linearly. This finding was very important for making a decision.

Figure 1

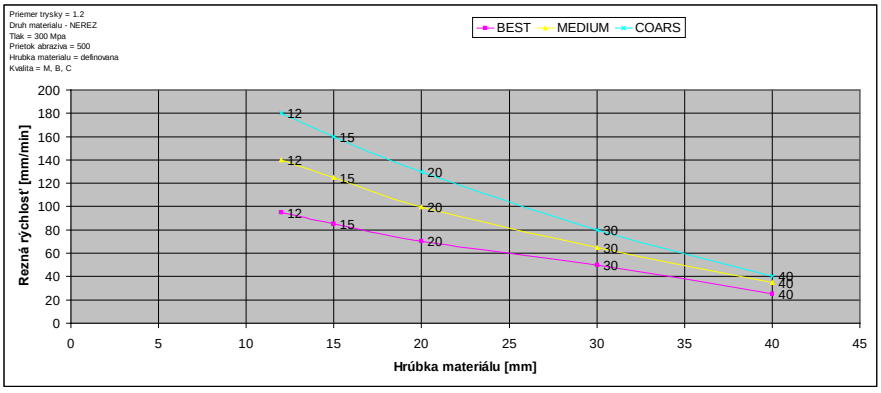


Figure 2

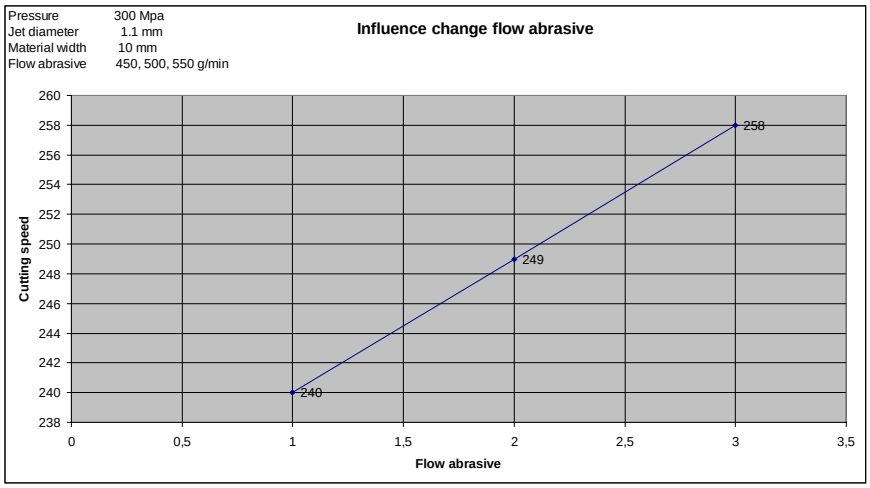
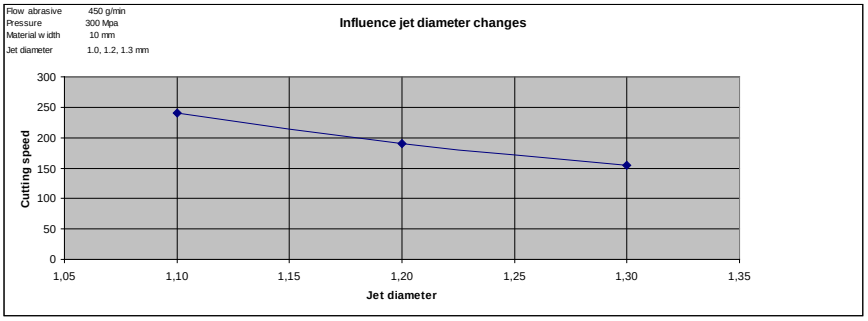


Figure 3



Therefore we decided to divide the range of the material width $\{1,2,3,5,8,10,12,15,20,30,40,50,70,80,100\}$ into two partitions as follows:

A) Two partitions

A	I. partition [mm]	II. partition [mm]
1.	1, 2, 3, 5, 8, 10, 12	12, 15, 20, 30, 40,50, 70, 80, 100
2.	1, 2, 3, 5, 8, 10, 12, 15	15, 20, 30, 40,50, 70, 80, 100
3.	1, 2, 3, 5, 8, 10, 12, 15, 20	20, 30, 40,50, 70, 80, 100

We used three methods for model building: *Backward Elimination*, *Forward Selection* and *Stepwise Regression*. Since the Forward Selection and Stepwise Regression yield the same results we present results only of the first method. Therefore the summary of the most convenient models for each partition and the selected methods of model building is the following:

Partitions A1 (material width [mm]: 1, 2, 3, 5, 8, 10, 12, 15)

1. *Backward Elimination* gives us the best model:

$$\hat{R} = f(PP, HH, HA)$$
$$R^2_{\text{adj}} = 0.999 \qquad MS_E = 11{,}268$$

2. *Forward Selection* gives us the best model:

$$\hat{R} = f(A)$$
$$R^2_{\text{adj}} = 0.873 \qquad MS_E = 105{,}640$$

Partition A1 (material width: 15, 20, 30, 40,50, 70, 80, 100**)**

1. *Backward Elimination* gives us the best model:

$$\hat{R} = f(PA, HH, H)$$
$$R^2_{\text{adj}} = 0.981 \qquad MS_E = 6.582$$

2. *Forward Selection* gives us the best model:

$$\hat{R} = f(H, HH, PA)$$
$$R^2_{\text{adj}} = 0.981 \qquad MS_E = 6.582$$

The particular results are summarized in the next table

***H* [mm]**	**Model's Repressors**	R^2_{adj}	**Std. Error**
1, 2, 3, 5, 8, 10, 12	P, H, HH, HA	0,998	11,421
12, 15, 20, 30, 40,50, 70, 80, 100	PA, HH, H	0,986	6,315
1, 2, 3, 5, 8, 10, 12, 15	PP, HH, HA	0,999	11,268
15, 20, 30, 40,50, 70, 80, 100	PA, HH, H	0,973	6.582
1, 2, 3, 5, 8, 10, 12, 15, 20	PA, HH, P, H	0,991	27,79
20, 30, 40,50, 70, 80, 100	PA, HH, H	0,996	2,074

3 Conclusion

The most convenient of the presented models are the three obtained from dividing the whole set of values of the material width $\{1,2,3,5,8,10,12,15,20,30,40,50,70,80,100\}$ into the two partitions $\{1,2,3,5,8,10,12,15\}$ and $\{15,20,30,40,50,70,80,100\}$. In future we will try to join the given two partitions by two regression functions in such a way that they form one continuous function. The response function together with confidence interval on the mean response at any point could by used in cutting process control. Bibliography includes also publications which deal with process control and statistical quality control.

4 Bibliography

[1] KÁLL, M., GELETA, V., BERNÁT, F. Spôsoby ovplyvňovania vysokotlakého kvapalinového lúča. In. *9. medzinárodná konferencia Strojné inžinierstvo 2005*. Bratislava 2005. ISBN 80-227-2314-2, s. 779-786.

[2] GELETA, V., MELICHER, I., KÁLL, M. Dynamic Jet Control in WJ scope. In. *9. medzinárodná konferencia Strojné inžinierstvo 2005*. Bratislava 2005. ISBN 80-227-2314-2, s. 749-752.

[3] MONTGOMERY, D. C., RUNGER, G. C. *Applied statistics and probability for engineers.* John Wiley & Sons, Inc., 2003. 706 pp. ISBN 0-471-20454-4.

[4] GARAJ, I., JANIGA, I. *Dvojstranné tolerančné medze pre neznámu strednú hodnotu a rozptyl normálneho rozdelenia* . Bratislava: STU, 2002. 147 s. ISBN 80-227-1779-7.

[5] GARAJ, I., JANIGA, I. *Dvojstranné tolerančné medze normálnych rozdelení s neznámymi strednými hodnotami a s neznámym spoločným rozptylom. Two sided tolerance limits of normal distributions with unknown means and unknown common variability.* Bratislava: STU, 2004. 218 s. ISBN 80-227-2019-4.

[6] GARAJ, I., JANIGA, I. *Jednostranné tolerančné medze normálneho rozdelenia s neznámou strednou hodnotou a rozptylom. One sided tolerance limits of normal distribution with unknown mean and variability.* Bratislava: STU, 2005. 214 s. ISBN 80-227-2218-9.

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

[8] TEREK, M., HRNČIAROVÁ, Ľ. *Analýza spôsobilosti procesu*. Vydavateľstvo EKONÓM, Ekonomická univerzita v Bratislave, 2001. 205 s. ISBN 80-225-1443-8.

[9] HRNČIAROVÁ, Ľ., TEREK, M. Analýza zoskupení bodov v regulačných diagramoch. In *FORUM STATISTICUM SLOVACUM*. ISSN 1336-7420, 2005, roč. I., č. 1, s. 56-61.

[10] TEREK, M.- HRNČIAROVÁ, Ľ- LIŠKOVÁ, I :Navrhovanie regulačných diagramov v štatistickej regulácii procesu In *Ekonomika a informatika*. ISSN 1336-3514, 2005, roč. III, č. 1, s. 126-137.

[11] TEREK, M.- HRNČIAROVÁ, Ľ.: Zisk zo stratifikácie. In *FORUM STATISTICUM SLOVACUM*. ISSN 1336-7420, 2006, roč. II., č. 1, s. 51-55.

[12] GROSS, P., KUREKOVA, E. Advanced Experiments Design for the Three-Torch Plasma Cutter Testing. In *Proceedings of the 4th International Conference MEASUREMENT 2005*. Bratislava: VEDA, 2005. ISBN 80-967402-8-8, p. 530-533.

[13] PALENČÁR, R., RUIZ, J.M., JANIGA, I., HORNÍKOVÁ, A. *Štatistické metódy v skúšobných a kalibračných laboratóriách*. Bratislava: Grafické štúdio Ing. Peter Juriga, 2001. 380 s. ISBN 80-968449-3-8.

[14] GARAJ, I. Sequential sampling plan of Poisson distribution. In *Mechanical Enginneering: International Conference: Proceedings.* Bratislava: SjF STU, 2001. ISBN 80–227-1616-2, p. 670-675.

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

Acknowledgement

This contribution was supported by the Grant VEGA No. 1/3182/06 as well as the Grant VEGA No. 1/1247/04.

Addresses of authors:

Doc. RNDr. Ivan Janiga, PhD., Katedra matematiky, SjF STU, Nám. slobody 17,
812 31 Bratislava, tel.: +421-2-57296-160, e-mail: ivan.janiga@stuba.sk

Ing. Marek Káll, Katedra výrobných systémov, SjF STU, Nám. slobody 17,
812 31 Bratislava, tel.: +421-2-44442456, e-mail: mkal@gratex.com

Doc. Ing. Vojtech Geleta, PhD., Katedra výrobných systémov, SjF STU, Nám. slobody 17,
812 31 Bratislava, tel.: +421-2-57296-160, e-mail: vojtech.geleta@stuba.sk

Cobbove-Douglasove produktivitné funkcie stavebných podnikov pri uvažovaní chyby v premenných

Samuel Korónyi

Abstract: Almost all economic variables are measured with error but surprisingly most of econometrics textbooks mention only very briefly the problem. Paper deals with application of errors in variables models of Slovak construction companies' productivity functions.
Key words: Errors in variables regression, Construction sector

1. Úvod

Klasické predpoklady lineárneho regresného modelu obsahujú aj podmienku o deterministickom charaktere nezávislých premenných. Koncom 19. storočia sa objavili metódy zohľadňujúce chyby v premenných napr. Passingova-Blalockova regresia. Až začiatkom 70. – tych rokov 20. storočia nastalo znovu oživenie záujmu o modely s chybami v premenných – tzv. EIV (errors in variables) modely (Maddala 1988). Je to prekvapujúce vzhľadom na zrejmý fakt nepresného merania väčšiny ekonomických ukazovateľov.

Pri uvažovaní chyby v premenných sa používajú funkčné a štruktúrne vzťahy, ktoré nepatria k lineárnym regresným modelom, aj keď ich v mnohom pripomínajú (Zvára 1989).

Ďalej uvedieme odhad regresných parametrov pomocou klasickej metódy najmenších štvorcov („naivné riešenie problému" ako sa uvádza v renomovaných učebniciach ekonometrie). Pre viac nezávislých premenných je explicitné vyjadrenie regresného parametra značne komplikovanejšie a z praktického hľadiska nevhodné. Už v prípade dvoch nezávislých premenných sú nutné ďalšie doplňujúce predpoklady (na základe väčšinou neznámych alebo otáznych informácií) pre určenie smeru a veľkosti vychýlenia regresných parametrov.

2. Riešenie klasickou metódou najmenších štvorcov pre jednoduchý regresný model

Nech je skutočný model v tvare

$$y = \alpha + \beta x + e\,, \tag{1}$$

kde α, β sú neznáme parametre. Namiesto premenných y a x však máme k dispozícii premenné

$$Y = y + v \text{ a } X = x + u, \tag{2}$$

kde u a v sú chyby premenných, x a y sú systematické zložky. Predpokladáme, že chyby majú nulové stredné hodnoty a rozptyly σ_u^2 a σ_v^2. Ďalej predpokladáme, že chyby sú nekorelované medzi sebou a tiež so systematickými zložkami. Formálne

$$E(u) = E(v) = 0, \text{var}(u) = \sigma_u^2, \text{var}(v) = \sigma_v^2,$$
$$\text{cov}(u, x) = \text{cov}(u, y) = \text{cov}(v, x) = \text{cov}(v, y) = 0. \quad (3)$$

Rovnica v pozorovaných premenných má tvar

$$Y - v = \alpha + \beta(X - u) + e \quad (4)$$

alebo

$$Y = \alpha + \beta X + w, \quad (5)$$

kde $w = e + v - \beta u$. Na tento model nemôžeme použiť metódu najmenších štvorcov, pretože $\text{cov}(w, X) \neq 0$. Po dosadení totiž máme

$$\text{cov}(w, X) = \text{cov}(e + v - \beta u, x + u) = \text{cov}(-\beta u, x + u) = -\beta\sigma_u^2. \quad (6)$$

Jeden z hlavných predpokladov metódy najmenších štvorcov nie je splnený. Regresný parameter nie je len vychýlený, ale navyše aj nekonzistentný, pretože ostáva vychýlený aj v prípade neohraničeného rastu veľkosti súboru. Ak len premenná *y* má chybu a premenná *x* nie, potom $\text{cov}(w, X) = 0$ a problém nevzniká. Problémom sú chyby v premennej *x*. Pri výpočte regresného parametra β klasickou metódou najmenších štvorcov dostávame

$$b_{YX} = \frac{\sum XY}{\sum X^2} = \frac{\sum (x+u)(y+v)}{\sum (x+u)^2}. \quad (7)$$

Pre veľké súbory v limite

$$p\lim b_{YX} = \frac{\text{cov}(xy)}{\text{var}(x) + \text{var}(u)} = \frac{\sigma_{xy}}{\sigma_x^2 + \sigma_u^2}, \quad (8)$$

pretože zmiešané súčiny vypadnú. Pre $\beta = \sigma_{xy} / \sigma_x^2$ máme

$$p\lim b_{YX} = \frac{\beta}{1 + \frac{\sigma_u^2}{\sigma_x^2}}. \quad (9)$$

Preto b_{YX} znižuje skutočnú hodnotu β. Miera zníženia závisí od podielu rozptylov chybovej a systematickej zložky v nezávislej premennej σ_u^2 / σ_x^2. Dostali sme jednoduchý klasický odhad metódou najmenších štvorcov pre hodnotu lineárneho regresného parametra v závislosti od relatívnej chyby určenia nezávislej premennej.

3. Hodnoty parametrov produktivitnej funkcie pri uvažovaní chyby v premenných

Cobbova-Douglasovou produktivitná funkcia sa odvodzuje z klasickej Cobbovej-Douglasovej produkčnej funkcie (Korónyi 2002) vydelením vstupným faktorom práce a po úprave dostaneme

$$\frac{y}{L} = \alpha\left(\frac{K}{L}\right)^{\beta}. \qquad (10)$$

kde *y/L* je podiel výstupu a vstupu práce (produktivita práce) a *K/L* je podiel vstupu kapitálu a vstupu práce (vybavenosť zamestnancov kapitálom). Ide o vzťah produktivity práce ako lineárnej alebo mocninovej funkcie vybavenosti zamestnancov kapitálom.

Na základe vzťahu (9) je možné simulovať regresný parameter pre vybrané podiely rozptylov chybovej a systematickej zložky nezávislej premennej. V ekonometrickej literatúre sa tento pomer označuje λ.

Pre stavebné podniky právnej formy s. r. o. je regresná rovnica linearizovanej produktivitnej funkcie $y = -0,984 + 0,428.x$. Hodnoty regresných parametrov pre podniky právnej formy s. r. o. v závislosti od pomeru rozptylov chybovej a systematickej zložky sú uvedené v tabuľke 1.

Tabuľka 1. Hodnoty regresných parametrov produktivitnej funkcie pre stavebné podniky právnej formy s. r. o. v závislosti od λ

λ	β	α
0.00	0.428	-0.984
0.05	0.449	-0.962
0.10	0.471	-0.940
0.15	0.492	-0.918
0.20	0.514	-0.895
0.25	0.535	-0.873
0.30	0.556	-0.851
0.35	0.578	-0.829
0.40	0.599	-0.807
0.45	0.621	-0.785
0.50	0.642	-0.763

λ = pomer rozptylov chybovej a systematickej zložky nezávislej premennej, β= lineárny regresný parameter produktivitnej funkcie

Z tabuľky 1 je vidieť, že pre rastúci podiel relatívnej chyby určenia nezávislej premennej (vybavenosti pracovnej sily kapitálom) narastá aj lineárny regresný parameter β (zväčšuje sa sklon regresnej priamky) od 0,428 po 0,642, rovnako rastie aj absolútny člen α od - 0,984 po - 0,763.

Tento jav je univerzálny - odhad metódou najmenších štvorcov (λ = 0), vychyľuje regresný parameter β smerom k nule, parameter α naopak smerom od nuly.

Pre stavebné podniky právnej formy a. s. je regresná rovnica produktivitnej funkcie $y = -1,208 + 0,270.x$. Hodnoty regresných parametrov pre podniky formy a. s. v závislosti od pomeru rozptylov chybovej a systematickej zložky sú uvedené v tabuľke .

Tabuľka 2. Hodnoty regresných parametrov produktivitnej funkcie pre stavebné podniky právnej formy a. s. v závislosti od λ

λ	β	α
0.00	0.270	-1.208
0.05	0.284	-1.202
0.10	0.297	-1.197
0.15	0.311	-1.192
0.20	0.324	-1.186
0.25	0.338	-1.181
0.30	0.351	-1.175
0.35	0.365	-1.170
0.40	0.378	-1.164
0.45	0.392	-1.159
0.50	0.405	-1.153

λ = pomer rozptylov chybovej a systematickej zložky nezávislej premennej, β= lineárny regresný parameter produktivitnej funkcie

Z tabuľky 2 je opäť vidieť nárast lineárneho regresného parametra β od 0,270 po 0,405 a parametra α od -1,208 po -1,153 pre rastúci podiel relatívnej chyby určenia nezávislej premennej.

4. Záver

Príspevok uvádza do problematiky odhadov regresných parametrov pri uvažovaní chyby v premenných. Pre riešenie uvedeného problému sú vypracované určité postupy, všetky však vyžadujú vstup ďalších parametrov, informácií ap. a to je často najväčší problém. V našom prípade práve odhad chyby v nezávislej premennej je sám najväčší problém.

5. Literatúra

GREEN, W. H. 1997. Econometric Analyses. Londýn: Prentice – Hall, 1997. ISBN 0-13-7246659-5.

KORÓNY, S. 2002. Lineárna produktivitná funkcia slovenských stavebných podnikov. In: Zborník 11. medzinárodného seminára Výpočtová štatistika. Bratislava : SŠDS, 2002, s. 39 – 41. ISBN 80-88946-20-4

MADDALA, G. S. 1988. Introduction to Econometrics. London: Macmillan, 1988. ISBN 0-02-374530-4

ZVÁRA, K. 1989. Regresní analýza. Praha: Academie, 1989. ISBN 80-200-0125-5

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

Intervaly spoľahlivosti v bayesovskej štatistike.

Eva Kotlebová, Daniela Sivašová

Abstract

The contribution is concerned with comparing classic and bayesian confidens intervals. The concept of highest density region is defined.

The difference between classic an bayesian interval is ilustrated on the beta distribution with parameters 20 and 5. Two softwares were used to find the highest density region: *Statgraphics Plus* and *EXCEL*. It was shown, that the calculations are more simple when using *Excel*, but *Statgraphics Plus* provides more precise result.

Úvod

Bayesovská štatistika predstavuje oproti klasickej alternatívny prístup k riešeniu induktívnych úloh – bodových a intervalových odhadov parametrov základného súboru, testovaniu hypotéz, a osobitným spôsobom sa podieľa aj pri riešení rozhodovacích úloh.

Zásadný rozdiel bayesovskej a klasickej štatistiky spočíva v tom, že kým v klasickej štatistike je výlučným zdrojom informácií výberový súbor, bayesovská štatistika berie do úvahy aj iné informácie o sledovanom parametri, ktoré spolu s výberovými údajmi vhodne skombinuje, a takto vytvorené induktívne závery môžu byť presnejšie a spoľahlivejšie.

Pri určovaní intervalov spoľahlivosti ako aj pri testovaní hypotéz sa v bayesovskej štatistike vychádza z tzv. *aposteriórneho* rozdelenia náhodnej premennej, ktoré vzniká matematicky odôvodneným skombinovaním výberového rozdelenia a tzv. *apriórneho* rozdelenia.

Cieľom príspevku nie je venovať sa súvislostiam medzi uvedenými rozdeleniami. Budeme predpokladať, že rozdelenie, prostredníctvom ktorého treba vytvárať induktívne závery, je pevne dané. Naším cieľom je poukázať na rozdiely, ktoré existujú pri konštrukcii intervalov spoľahlivosti v klasickej a bayesovskej štatistike, ktoré sú badateľné predovšetkým u nesymetrických rozdelení.

Regióny s najvyššou hustotou

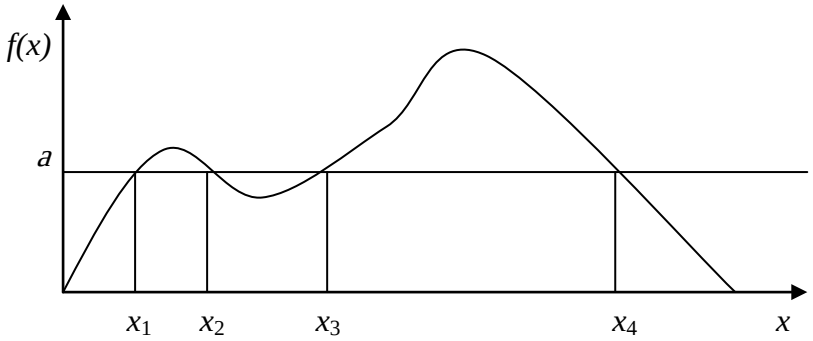
Klasická štatistika pristupuje k určovaniu dvojstranných intervalov spoľahlivosti štandardným spôsobom: pre danú spoľahlivosť $1-\alpha$ sú hranicami dvojstranného intervalu spoľahlivosti pevne stanovené kvantily uvažovaného rozdelenia: $x_{\frac{\alpha}{2}}$ a $x_{1-\frac{\alpha}{2}}$. Takéto intervaly sa v literatúre označujú ako *equal-tailed* (rovnochvostové – obidva kvantily oddeľujú na koncoch rozdelenia rovnako veľké plochy pod grafom funkcie hustoty $f(x)$ príslušného rozdelenia), čo možno symbolicky vyjadriť takto:

$$\int_{-\infty}^{\frac{\alpha}{2}} f(x)dx = \int_{1-\frac{\alpha}{2}}^{\infty} f(x)dx = \frac{\alpha}{2}.$$

V bayesovskej štatistike sa ku konštrukcii intervalov spoľahlivosti pristupuje iným spôsobom: každému zjednoteniu spojitých intervalov je možné priradiť pravdepodobnosť toho, že odhadovaný parameter nadobúda hodnotu práve z tejto množiny (vypočíta sa určitý integrál z funkcie hustoty na uvažovanej množine). Ak má táto pravdepodobnosť hodnotu $1-\alpha$, nazveme príslušnú množinu $(1-\alpha)\cdot 100\%$-nou *oblasťou spoľahlivosti*. Z takéhoto vymedzenia oblasti spoľahlivosti je zrejmé, že oblastí s rovnakou spoľahlivosťou je nekonečne veľa. Je teda vhodné určiť kritérium, ktoré by z tejto množiny oblastí vybralo tú najvhodnejšiu – tak, aby určený interval, (resp. zjednotenie intervalov) bol čo najužší. Toto kritérium možno vyjadriť takto:

Ak U je oblasťou spoľahlivosti (s požadovanou vlastnosťou), tak pre ľubovoľné $x_1 \in U$, $x_2 \notin U$ platí: $f(x_1) \geq f(x_2)$, t.j. funkcia hustoty nadobúda na príslušnom intervale (intervaloch) väčšiu, príp. rovnakú hodnotu, ako mimo tohto intervalu. Takto vymedzenú oblasť voláme región s najvyššou hustotou – *highest density region* (HDR). Situáciu ilustruje nasledujúci obrázok:

Obr. 1: Znázornenie regiónu s najvyššou hustotou.



$$U = \langle x_1; x_2 \rangle \cup \langle x_3; x_4 \rangle \qquad \int_{x_1}^{x_2} f(x)dx + \int_{x_3}^{x_4} f(x)dx = 1 - \alpha .$$

Je zrejmé, že hranice regiónu s najvyššou hustotou sú také hodnoty náhodnej premennej, v ktorých hustota rozdelenia nadobúda rovnakú hodnotu (na obr. 1 je to hodnota *a*).

Klasické intervaly spoľahlivosti a regióny s najvyššou hustotou nemusia byť nevyhnutne rôzne – napríklad v prípade normálneho rozdelenia sú vďaka jeho vlastnostiam úplne totožné. Ak je však uvažované rozdelenie zošikmené, intervaly budú odlišné. Dôjde nielen k posunu, ale aj k zúženiu intervalu, čo znamená vyššiu presnosť.

Budeme sa zaoberať beta rozdelením, ktoré má v bayesovskej štatistike dosť široké uplatnenie. Pretože pre rôzne dvojice hodnôt parametrov je rôzne zošikmené, prejaví sa rozdiel medzi klasickým intervalom spoľahlivosti a bayesovskou oblasťou spoľahlivosti.

Číslené porovnanie klasického a bayesovského intervalu spoľahlivosti .

Na ilustráciu sme použili rozdelenie $Be(20;5)$, ktoré má tieto číselné charakteristiky: $E(X) = \frac{20}{20+5} = 0{,}8$, $D(X) = \frac{20 \cdot 5}{(20+5)^2 \cdot (20+5+1)} = 0{,}00615$. Je ľavostranne zošikmené.

Počítali sme hranice dvojstranného 95%-ného intervalu spoľahlivosti. Použili sme pritom dva štandardné softvéry: *Statgraphics Plus* a *EXCEL*:

Riešenie v systéme *Statgraphics Plus*:

Z procedúry *Plot* sme vybrali *Probability Distributions* a zvolili sme rozdelenie *Beta*. V *Tabular options* sme vybrali funkciu *Inverse CDF*, ktorá zadanej pravdepodobnosti priradí zodpovedajúcu hodnotu kvantilu príslušného rozdelenia (ide teda o inverznú funkciu k distribučnej funkcii). Pomocou *Analysis Options* sme zadali parametre rozdelenia a v *Pane Options* sme zvolili hodnoty 0,025 a 0.975 (kvantily $x_{0,025}$ a $x_{0,975}$ sú hranicami klasického dvojstranného 95%-ného intervalu spoľahlivosti). Z výstupu vidno, že hranice tohto intervalu sú 0,626158 a 0,692723.

Obr. 2: Výstup z procedúry *Inverse CDF*.

```
Inverse CDF
-----------

Distribution: Beta

CDF           Dist. 1
0,025         0,626158
0,0975        0,692723
```

Aby sme zistili, či tento interval možno považovať aj za región s najvyššou hustotou pre spoľahlivosť 0,95, vybrali sme v *Tabular Options* funkciu *Cumulative Distribution*, ktorá okrem iného uvádza na výstupe hodnotu hustoty pre zadaný argument. Do *Pane Options* sme teda zadali hranice klasického intervalu spoľahlivosti:

Obr. 3: Výstup z procedúry *Cumulative Distribution*.

```
Cumulative Distribution
-----------------------
Distribution: Beta

              Probability Density
Variable      Dist. 1
0,626158      0,569045
0,692723      1,77073
```

Pretože hustota v krajných bodoch intervalu je rôzna, tento interval nie je regiónom s najvyššou hustotou. *Statgraphics Plus* neobsahuje procedúru, pomocou ktorej by bolo možné priamo určiť HDR, bolo treba postupovať v podstate mechanicky pomocou takéhoto algoritmu:

Hľadali sa hodnoty premennej x_1 a x_2 tak, aby spĺňali 2 podmienky:

1) hustota v oboch bodoch je rovnaká, t.j. $f(x_1) = f(x_2)$

2) $\int_{x_1}^{x_2} f(x)dx = 0{,}95$, čo znamená, že $F(x_2) = F(x_1) + 0{,}95$.

Vzhľadom na ľavostranné zošikmenie rozdelenia je zrejmé, že poloha hľadaných bodov x_1 a x_2 bude viac vpravo oproti počiatočným hodnotám, zodpovedajúcim štandardným kvantilom, ktoré sú hranicami klasického intervalu spoľahlivosti. V nasledujúcej tabuľke uvádzame prehľad dvojíc hodnôt x_1, x_2, ktoré sa postupne stále viac približovali hľadaným číslam:

Tabuľka 1: Kvantily rozdelenia Be(20;5) a im zodpovedajúce hodnoty funkcie hustoty:

$F(x_1)$	**$F(x_2)$**	**x_1**	**x_2**	**$f(x_1)$**	**$f(x_2)$**	**$f(x_2)$-$f(x_1)$**
0,025	0,975	0,626158	0,928681	0,569045	1,3514	0,782355
0,026	0,976	0,627886	0,929433	0,588625	1,31205	0,723425
0,028	0,978	0,631176	0,931002	0,627387	1,23825	0,610863
0,03	0.98	0,63427	0,932669	0,665658	1,16167	0,496021
0,04	0,99	0,647512	0,943379	0,850522	0,721685	-0.128837
0,037	0,987	0,643868	0,939595	0,79613	0,866093	0,069963
0,038	0,988	0,64511	0,940781	0,81436	0,819457	0,005097
0,0381	0,9881	0,645232	0,940904	0,816168	0,814691	-0,001477
0,03808	0,98808	0,645208	0,940879	0,815812	0,815659	-0,000153

V poslednom riadku je rozdiel hustôt už zanedbateľný, takže hodnoty x_1 a x_2 z posledného riadku možno považovať za hranice 95%-ného bayesovského intervalu spoľahlivosti, t.j. regiónu s najvyššou hustotou pre zvolenú spoľahlivosť 0,95.

Interval (0,645208;0,940879) je oproti klasickému (0,626158;0,928681) posunutý doprava a užší: šírka klasického je 0,302523 a šírka bayesovského 0,295671.

Riešenie v *EXCELi*:

Ani *EXCEL* neumožňuje priamo nájsť hranice HDR-regiónu s konkrétne určenou spoľahlivosťou. Jeho nevýhodou oproti systému *Statgraphics Plus* je aj to, že neobsahuje funkciu hustoty uvažovaného rozdelenia, takže sme boli nútení jej hodnoty naprogramovať z distribučnej funkcie, čo sa do určitej miery prejavilo na presnosti výpočtu.

Opäť sme hľadali také dve čísla, ktoré spĺňajú podmienky (1) a (2). Na určenie hodnoty zadaných kvantilov sme použili funkciu *BETAINV* s parametrami 20 a 5. Hodnoty hustoty v *i*-tom riadku sme určili „ručne“ použitím vzťahu

$$f(x_i) = \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}}.$$

Zodpovedajúce výpočty sú uvedené v nasledujúcej tabuľke:

Tabuľka 2: Kvantily rozdelenia Be(20;5) a im zodpovedajúce hodnoty funkcie hustoty:

i	$F(x_1)$	$F(x_2)$	x_1	x_2	$f(x_1)$	$f(x_2)$	$f(x_2)-f(x_1)$
1	0,0375	0,9875	0,644492	0,940180			
2	0,03755	0,98755	0,644554	0,940239	0,806597	0,845626	0,039029
3	0,0376	0,9876	0,644616	0,940299	0,806597	0,832203	0,025606
4	0,03765	0,98765	0,644678	0,940358	0,806597	0,845626	0,039029
5	0,0377	0,9877	0,644740	0,940418	0,806597	0,832203	0,025606
6	0,03775	0,98775	0,644802	0,940478	0,812850	0,832203	0,019354
7	0,0378	0,9878	0,644863	0,940538	0,812850	0,832203	0,019354
8	0,03785	0,98785	0,644925	0,940599	0,806597	0,819200	0,012603
9	0,0379	0,9879	0,644987	0,940660	0,812850	0,832203	0,019354
10	0,03795	0,98795	0,645048	0,940721	0,812850	0,819200	0,006350
11	0,038	0,988	0,645110	0,940782	0,812850	0,819200	0,006350
12	0,03805	0,98805	0,645171	0,940843	0,812850	0,819200	0,006350
13	***0,0381***	***0,9881***	***0,645232***	***0,940904***	***0,819200***	***0,819200***	***0,000000***
14	0,03815	0,98815	0,645293	0,940965	0,819200	0,812850	-0,006350
15	0,0382	0,9882	0,645354	0,941027	0,819200	0,812850	-0,006350
16	0,03825	0,98825	0,645416	0,941089	0,812850	0,806597	-0,006253
17	0,0383	0,9883	0,645477	0,941151	0,819200	0,806597	-0,012603
18	0,03835	0,98835	0,645537	0,941213	0,825650	0,806597	-0,019053
19	0,0384	0,9884	0,645598	0,941276	0,819200	0,794376	-0,024824
20	0,03845	0,98845	0,645659	0,941338	0,819200	0,806597	-0,012603
21	0,0385	0,9885	0,645720	0,941401	0,819200	0,794376	-0,024824
22	0,03855	0,98855	0,645781	0,941463	0,832203	0,794376	-0,037827
23	0,0386	0,9886	0,645842	0,941526	0,819200	0,794376	-0,024824
24	0,03865	0,98865	0,645902	0,941590	0,832203	0,782519	-0,049684
25	0,0387	0,9887	0,645963	0,941653	0,819200	0,794376	-0,024824
26	0,03875	0,98875	0,646023	0,941717	0,832203	0,782519	-0,049684
27	0,0388	0,9888	0,646083	0,941781	0,825650	0,782519	-0,043131

Ako v tabuľke vidno, minimálny rozdiel medzi hodnotami funkcie hustoty je v riadku 13, ktorý zodpovedá intervalu (0,645232;0,940904). Jeho šírka je 0,295708, čo je o 0,001409 viac

ako šírka intervalu, ktorý sme dostali pomocou systému *Statgraphics Plus*, ale o 0,006815 menej ako je šírka klasického intervalu spoľahlivosti.

Záver:

Pri určovaní bayesovského intervalu spoľahlivosti je možné využiť oba softvéry, majú však isté nedostatky, ktoré neumožnia hľadaný interval určiť priamo.

Systém *Statgraphics Plus* vyžaduje ručné zadávanie postupne získavaných hraníc intervalov, čo je pomerne zdĺhavé. Naviac pri niektorých rozdeleniach nedokáže pracovať s väčšími hodnotami parametrov, napr. pre rozdelenie Be(50;20) je už nepoužiteľný.

Výhodou *EXCEL*-u je to, že číselné nastavenie hodnôt v prvých dvoch riadkoch a naprogramovanie jednotlivých stĺpcov umožní rýchlejší výpočet. Absencia funkcie hustoty však znižuje presnosť výsledku.

Aj keď obe procedúry vykazujú vzhľadom na stanovený cieľ isté nedostatky, predsa umožňujú vypočítať hranice hľadaného intervalu. Sú teda vhodným nástrojom na dôkaz toho, že bayesovský interval spoľahlivosti je užší ako klasický.

Zoznam použitej literatúry

[1] BAKYTOVÁ H. – HÁTLE J. – NOVÁK.I.- UGRON M.: *Statistická indukce pro ekonomy*, Praha, SNTL, ALFA, 1986

[2] KOTLEBOVÁ E.: *Redukované bayesovské odhady v lineárnej regresii*, Nové Zámky, EDAMBA – zborník príspevkov, 2005

[3] KOTLEBOVÁ E.: *Redukované bayesovské odhady v analýze rozptylu*, Bratislava, 13. medzinárodný seminár VÝPOČTOVÁ ŠTATISTIKA, 2004,

[4] LEE PETER M: *Bayesian statistics*, New York: Oxford University Press, 1989

[5] PACÁKOVÁ V. a kolektív: *Štatistika pre ekonómov*, Bratislava: IURA EDITION, 2003

[6] PACÁKOVÁ V.: *Aplikovaná poistná štatistika*, Bratislava: IURA EDITION, 2004

[7] PACÁKOVÁ V. – KOTLEBOVÁ E.: *Bayesovská štatistika v poisťovníctve*, Bratislava, Ekonomika a informatika 2/2004, IURA EDITION

[8] PACÁKOVÁ V. – SIVAŠOVÁ D. – TÓTHOVÁ M: *Teória pravdepodobnosti*, Bratislava, ES VŠE, 1990

[9] SODOMOVÁ E. a kol.: *Štatistika. Modul A*, Bratislava, Ekonóm, 2001

[10] TEREK M.: *Úvod do analýzy rozhodovania a bayesovskej indukcie*, Bratislava: EKONÓM, 2003

[11] WONNACOT T.H. – WONNACOT R.J.: *Statistika pro obchod a hospodářství*, Praha: VICTORIA PUBLISHING, 1993

[12] http://en.wikipedia.org/wiki/Bayes'_theorem

Adresa autoriek

RNDr. Eva Kotlebová, RNDr. Daniela Sivašová, PhD.
Katedra štatistiky FHI, Ekonomická univerzita, Dolnozemská cesta 1, 852 35 Bratislava
e-mail: eva.kotlebova@inmail.sk, sivasova@euba.sk

Použitie entropie pri posudzovaní koncentrácie a asymetrie rozdelenia

Viera Labudová, Katarína Sušienková

Abstract
The entropy of a system is related to the amount of information it contains. A highly ordered system can be described using fewer bits of information than a disordered one. Shannon's formula gives the entropy $H = -\sum_i p_i \log_2 p_i$. The article presents the entropy as a measure of inequality of a distribution, entropy is used as a parameter that describes asymmetry of a distribution too.

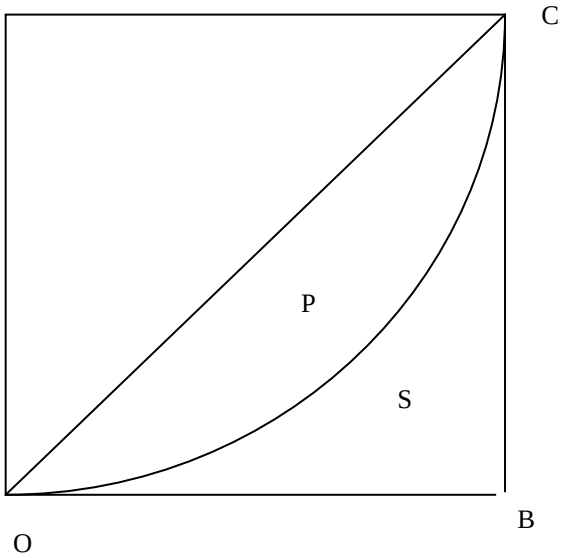
Entropia ako miera koncentrácie

Nerovnomernosť rozdelenia celkovej sumy (úhrnu) hodnôt znaku medzi jednotlivé štatistické jednotky štatistického súboru nazývame v štatistike koncentráciou.

Na posúdenie sily koncentrácie možno použiť grafické zobrazenie rozdelenia hodnôt znaku v pravouhlom súradnicovom systéme, tzv. Lorenzovu krivku (koncentračnú krivku). Lorenzova krivka vznikne tak, že na os x nanášame kumulované podiely štatistických jednotiek F_i a na os y kumulované podiely hodnôt znaku Z_i, ktoré prislúchajú danému podielu štatistických jednotiek.

Vychádzajúc z Lorenzovej krivky, možno odvodiť vzťah na výpočet koeficienta koncentrácie.

Graf 1 *Lorenzova krivka*



Ak P je plocha ohraničená diagonálou a Lorenzovou krivkou, T je plocha trojuholníka 0BC, S plocha pod Lorenzovou krivkou a plocha štvorca so stranami o dĺžke jedna sa rovná jednej, potom plocha trojuholníka T=0,5.
Koeficient koncentrácie možno určiť[1]:

[1]Uvedný vzťah možno ďalej upresniť, viď. Terek (2002), Pacáková ().

$$K_k = \frac{P}{T} = \frac{T-S}{T} = \frac{0,5-S}{0,5} = 1-2S \quad (1)$$

Koeficient koncentrácie nadobúda hodnoty od 0 do 1. Čím viac sa jeho hodnota blíži k jednej, tým je koncentrácia vyššia, t. j. hodnoty znaku sú rozdelené nerovnomernejšie a naopak hodnoty blížiace sa k 0 svedčia o nízkom stupni koncentrácie, teda o rovnomernom rozdelení sledovaných hodnôt znaku.

Na meranie nerovnomernosti rozdelenia hodnôt premennej možno použiť entropiu.

V teórii informácií sa entropia používa ako miera apriórnej neurčitosti systému. Ak skúmame systém X, ktorý je schopný prijímať konečné množstvo stavov x_1, x_2, ...,x_n s pravdepodobnosťami p_1, p_2,...,p_n, entropiou systému sa nazýva súčet súčinov pravdepodobností rôznych stavov systému a logaritmov týchto pravdepodobností. s opačným znamienkom.

$$H = -\sum_i p_i \log_2 p_i \,. \quad (2)$$

Systém X môžeme nahradiť diskrétnou náhodnou premennou X, ktorá nadobúda hodnoty x_1, x_2, ...,x_n (môžu to byť napríklad poradové čísla stavu systému) s pravdepodobnosťami p_1, p_2,...,p_n[2] a vzťah (2) použiť na vyčíslenie entropie empirického súboru s premennou X.

Ak sa entropia rovná nule, všetky štatistické jednotky nadobúdajú rovnakú hodnotu x_j. Entropia nadobúda maximálnu hodnotu vtedy, keď sú rôzne hodnoty x_1, x_2, ...,x_n rovnako pravdepodobné, resp. relatívne početnosti tried x_1, x_2, ...,x_n sú rovnaké. Maximálna hodnota entropie sa rovná počtu rôznych obmien premennej X.

$$H_{max} = \log_2 n \,. \quad (3)$$

Vzhľadom na to, že entropia závisí od počtu obmien premennej, pri porovnávaní entropie rôznych súborov sa používa koeficient entropie , ktorý je vyjadrený pomocou jej maximálnej hodnoty:

$$WH = 1 - \frac{H}{H_{max}} \,. \quad (4)$$

Uvažujme nasledovné príklady rozdelenia pravdepodobnosti (relatívnej početnosti) náhodnej premennej X.

Tab. 1...*Empirické súbory pre výpočet entropie*

A		**B**		**C**		**D**	
x_i	p_i	x_i	p_i	x_i	p_i	x_i	p_i
1	0,2	1	0,05	1	0,02	1	0,40
2	0,2	2	0,10	2	0,03	2	0,50
3	0,2	3	0,70	3	0,05	3	0,05
4	0,2	4	0,10	4	0,50	4	0,03
5	0,2	5	0,05	5	0,40	5	0,02

[2] Pre opísanie miery neurčitosti systému nie sú dôležité hodnoty x_1, x_2, ...,x_n, podstatný je len počet týchto hodnôt (stavov) a ich pravdepodobnosti.

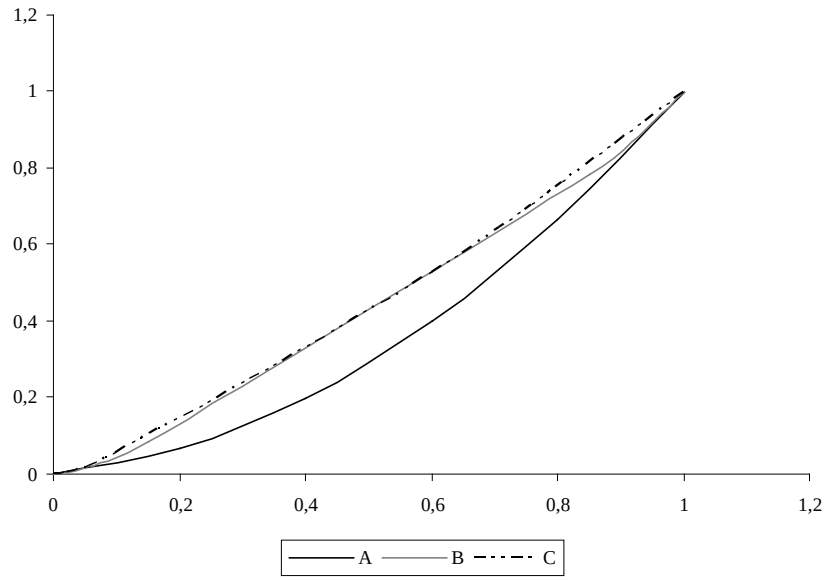
Pre každý prípad A, B, C, D sú vyčíslené hodnoty entropie H a koeficienta entropie WH. Koncentrácia hodnôt znaku je najväčšia v prípade A, kedy dosahuje entropia maximum a koeficient entropie hodnotu 0. Koncentrácia hodnôt znaku v prípadoch B, C, D je nižšia (v prípade C a D rovnaká).

Tab. 2 *Hodnoty rôznych mier koncentrácie*

	A	**B**	**C**	**D**
H	2,3219	1,4568	1,5095	1,5095
WH	0	0,3723	0,3499	0,3499

Rôzny stupeň koncentrácie hodnôt znaku pre súbory A, B, C je zobrazený pomocou Lorenzovej krivky.

Graf 2 *Lorenzova krivka pre súbory A, B, C*



Použitie entropie pri posudzovaní asymetrie rozdelenia

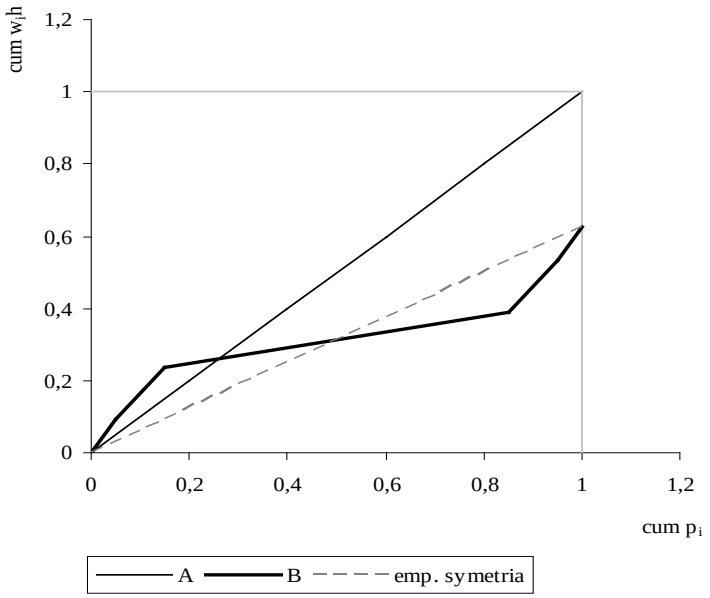
Grafické zobrazenie hodnôt entropie možno použiť na posúdenie asymetrie rozdelenia náhodnej premennej X.

Graf pre posúdenie asymetrie rozdelenia je konštruovaný z kumulatívnych pravdepodobností (relatívnych početností) jednotlivých hodnôt znaku *cum* p_i, ktoré sú nanesené na x-ovú os a kumulatívnych „čiastkových“ koeficientov entropie *cum* $w_i h$, ktoré sú nanesené na y-ovú os. Čiastkové koeficienty entropie sa počítajú podľa vzťahu:

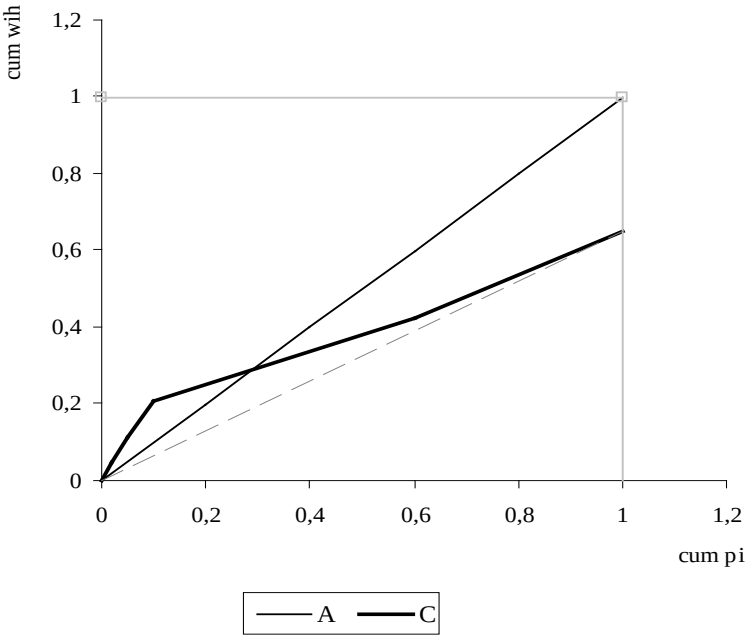
$$w_i = \frac{\left| p_i \log_2 p_i \right|}{H_{\max}}. \tag{5}$$

Pri určovaní asymetrie (zošikmenia) rozdelenia hodnôt znaku v štatistickom súbore je dôležitá tzv. čiara „empirickej symetrie“. V prípade symetrického rozdelenia, graf entropie pretína túto čiaru (B), ak je zošikmenie rozdelenia pravostranné, graf entropie prechádza ponad túto čiaru (D), pri ľavostrannej asymetrii, leží čiara empirickej symetrie nad grafom entropie (C).

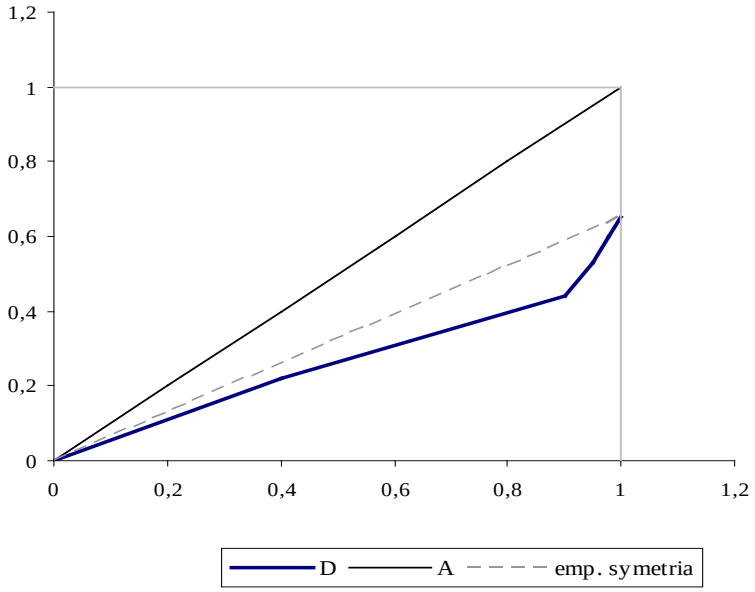
Graf 3 *Enropia súboru B*



Graf 4 *Entropia súboru C*



Graf 5 *Entropia súboru D*



Literatúra

1. PACÁKOVÁ V.: Štatistika pre ekonómov. Edícia Ekonómia, Bratislava, 2003
2. ROESKE - SLOMKA, I.: Entropia jako miara koncentracji i asymetrii rozkladu.. In.: Przeglad statystyczny, Nr.1, 1993, s. 61- 69.
3. TEREK M.: Miery koncentrácie, Slovenská štatistika a demografia, č. 3, 2002, ISSN 1210-1095
4. VENTCELOVÁ, J. S.: Teória pravdepodobnosti. ALFA, 1973.
5. http://dictionary.reference.com/browse/entropy

Kontakt

RNDr. Viera Labudová, PhD.,
Katedra štatistiky FHI; Ekonomická univerzita v Bratislave, Dolnozemská cesta 1, 852 35
labudova@dec.euba.sk
Ing. Katarína Sušienková,
Katedra štatistiky FHI; Ekonomická univerzita v Bratislave, Dolnozemská cesta 1, 852 35
susienko@dec.euba.sk

Uvedený príspevok vznikol ako súčasť projektu VEGA 1/2631/05

Analysis of time series of the selected interventions of the fire-fighting rescue brigades in the Czech Republic

Bohdan Linda, Jana Kubanová

Abstract: The paper deals with time series analysis of the number of fires and claim amounts put there with these fires in selected department of national economy.

Key words: Fire and Rescue Service of the Czech Republic, expenses from the state budget, claims caused by fires, preserved value

Though the Fire and Rescue Service of the Czech Republic is a coordinator of the integrated rescue system, nevertheless its widest sphere of activity is both fire liquidation and its fall-out liquidation. The second wide sphere of its operation is to provide subservience at traffic accidents. The number of fires and traffic accidents, at which intervened fire and rescue service troops**,** is on a long-term basis approximately identical and moves about 20 000 interventions per a year. The third significant sphere, requiring intervention of the fire and rescue service troops, are dangerous chemicals escapes, above all escapes of petrol products.
In terms of numbers about one quarter of interventions is in evidence in this area in comparison with fires and traffic accidents.

In terms of financial losses the biggest claims rise in consequence upon fires. That is why we deal with these losses in the article, classified according to individual branches. However several global statistics will be presented at first. The following table 1 shows the time series of the indicators, comparison of them, predicates about efficiency of the financial resources, issuing by the state for the Fire and Rescue Service of the Czech Republic working in the years 2001 - 2005.
It deals with expenses from the state budget for the fire and rescue service, with direct claims caused by fire and with preserved values at the fires. The data are presented in milliards and in current prices. With regard to comparison of trends, this presumption can be accepted.

Table 1: Financial resources, issuing by the state for the Fire and Rescue Service of the Czech Republic

year	**2001**	**2002**	**2003**	**2004**	**2005**
expenses	5	5,702	5,895	6,707	7,127
claims	2,055	3,732	1,837	1,669	1,634
preserved values	6,23	6,252	7,647	6,977	7,11

When we illustrate these time series in a grafical way (fig. 1) we can see, that although the expenses from state budget for the fire and rescue service have increasing trend (tangent of the regression line is 0,52 and the significance test rejects the hypotheses about its zero value), preserved values don't follow this trend (the regression line has tangent 0,2 and the significance test didn't reject the hypotheses about her zero value), which would be expected at the first sight. This phenomenon can be explained by a thought that the long-term average of numbers of fires stagnates, so that preserved values stagnate as well.

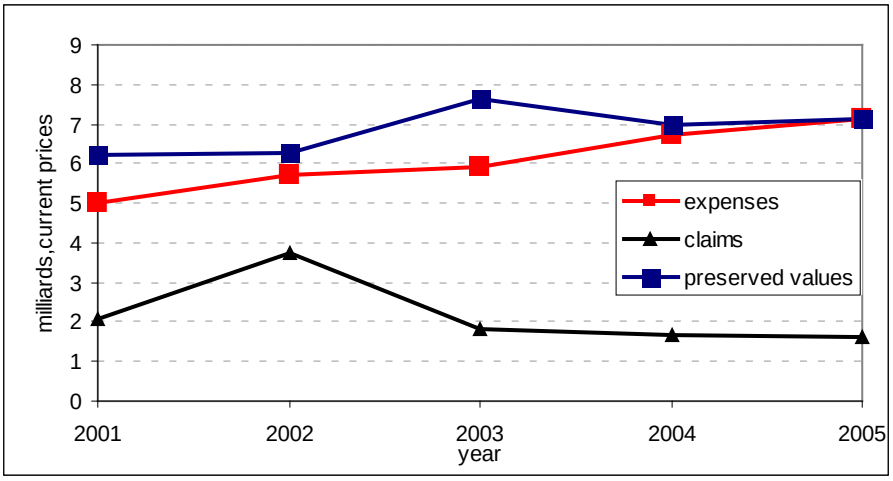


Fig.1 Expenses, claims and preserved values

The increasing expenses can be placed to debit of the reality, that the number of actions of the fire and rescue service troops increases at other kinds of crisis events, however here is no reliable methodology for their valuation as it is for fires.

It can be interesting to go over from aggregate data to data in individual branches. The greatest claims, in capacity of finance, are caused by industry fires. The following graph 2 presents their development both in the number and financial losses.

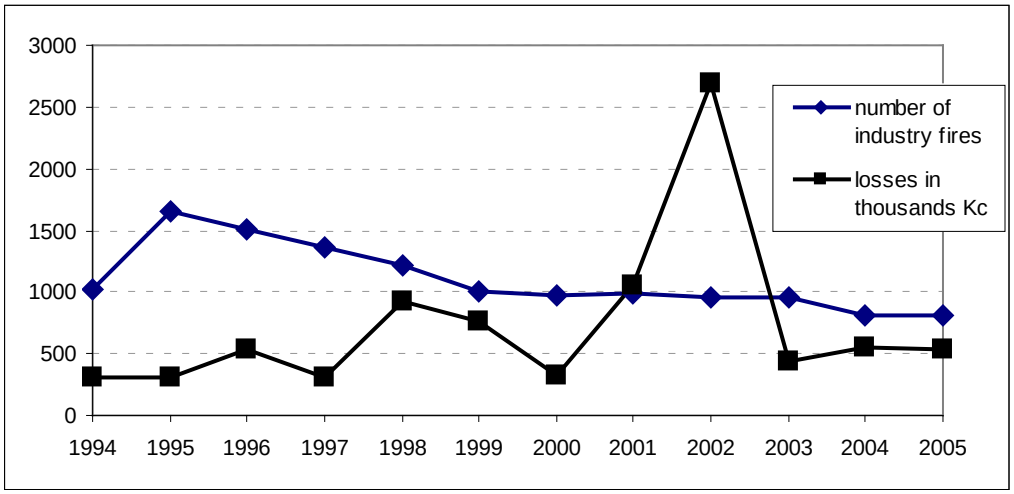


Fig. 2. Industry fires and caused losses (in thousands of Czech Crowns)

When the number of actions was analysed, the statistical test of the estimate of the tangent of the line confirms the decreasing trend. On the other hand, in terms of losses, the test of the tangent of the regression line doesn’t reject the hypothesis about its stagnation.

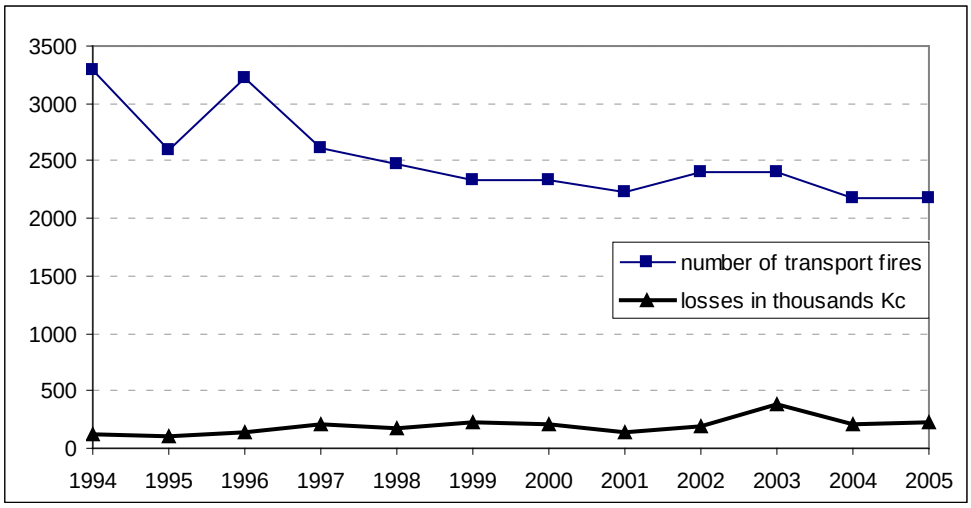


Fig. 3. Transport fires and caused losses (in thousands of Czech Crowns)

Transport is the following branch in succession in terms of amount of losses. The presented graph 3 illustrates the development of the identical indicators in transport. When tangents of both regression lines were tested we discovered that *p*-value is 0,012 for the number of fires, it means that at the significance level 0,05 we can declare that the trend is decreasing but the same statement is not true for the significance level 0,01. We have not reason to reject, the stagnation of the trend at the regression line presenting the amount of loses, *p*-value is 0,0495.

The household fires are the most interesting for the common residents. This kind of fires can be, in size of losses, classified together with fires in trading organisations by the third step. The graph 4 illustrates again the number of fires and amount of caused losses.

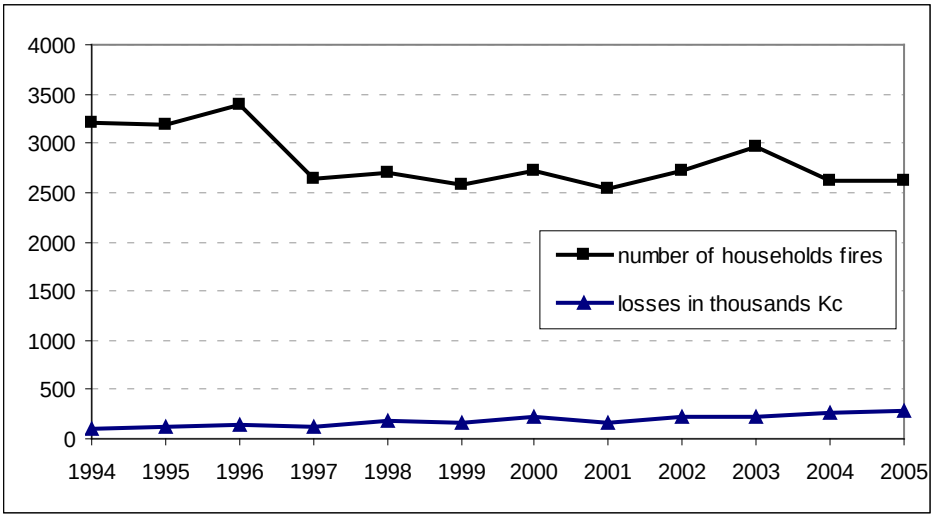


Fig. 4. Household fires and caused losses (in thousands Czech Crowns)

When tangent of the trend of the fires was analysed, we can not claim on the base of *p*-value, that the trend is decreasing. On the other hand the same test claimed the increasing trend in the case of losses. With regard to not changing number of fires we can infer from this that the lifestyle of population can be increasing.

It is possible to say in the end that all figures present sure stagnation of the trends of time series of the number of fires and claims, both according to the branches and in total. It could predicate, with regard to increasing budget grants, needs to take a deep think over organisation, managenemt and decision making at solution of the crisis situations.

Literature:

[1] Linda, B. – Kubanová, J.: Fire statistics in the Czech Republic in the years 1996-2005. Forum Statisticum Slovacum 4/2006. SŠDS Bratislava 2006. ISSN 1336-7420
[2] http://www.czso.cz
[3] http://www.mvcr.cz
[4] Rublíková, E., Hill, M. M.: Modelling Time Series with ConditionalHeteroscedasticity. The Simple ARCH Model. In: Scientific Papers Fakultaekonomicko-správní Univerzity Pardubice, č. 10/2006. Ser. D. ISSN 1211-555X,ISBN 80-7194-851-9, s. 139-147.
[5] Rublíková Eva, Magalhaes-Hill Manuela: A Non-linear Approach to ModellingAsymmetry in a Time Series. In: Temas em Métodos Quantitativos. No 4s.51-59. - Lisboa : Edicoes Sílabo 2004.- 273 s. ISBN 972-618-329-4
[6] MOLNÁR Štefan, RUBLÍKOVÁ Eva: Moderné metódy modelovania finančných časových radov / Modern Methods of Financial Time Series Modelling. In: Burza, 1/2004. s. 39-42. ISSN 1335-1435.

Acknowledgment:
The article was worked-out with support of the grant GAČR 402/06/0084 "Modelling and Optimisation of the Decision-making Processes within Municipal and Regional Administration".

Authors:
Doc.RNDr. Bohdan Linda, CSc.
Doc.PaedDr. Jana Kubanova, CSc.
University of Pardubice
Studentska 95
53009 Pardubice
e-mail: bohdan.linda@upce.cz
jana.kubanova@upce.cz

Overovanie reprezentatívnosti výberového súboru

Ján Luha

Abstract: Article deals with definition representativeness of sample survey and their formalisation. Then we discussed basic method for testing of representativeness.

1. Úvod

Pri výberových zisťovaniach ako sú napríklad výskumy verejnej mienky je jedným z najdôležitejších atribútov reprezentatívnosť získaných datových súborov. Najpoužívanejšie metódy na získanie reprezentatívnych výberových súborov sú metóda náhodného výberu (uvažujeme oblastný náhodný výber) a metóda kvótového výberu.
V príspevku uvedieme formalizujeme definíciu reprezentatívnosti a uvedieme základné testy reprezentatívnosti výberových súborov. Zároveň spresňujeme pojem reprezentatívnosti uvedený v práci Luha J. (2005).

2. Definícia reprezentatívnosti

Na začiatok uvedieme pojem reprezentatívneho výberu z ESOMAR Marketing Research Glossary: **Representative Sample** is a sample that contains units in the same proportion as the population of interest.
V závislosti od skúmaných znakov (charakteristík) a od spôsobu vytvorenia výberového súboru možno pojem reprezentatívnosti ďalej bližšie špecifikovať. Budeme sa zaoberať kvalitatívnymi znakmi a dvoma spôsobmi konštrukcie výberových súborov, ktoré zabezpečujú ich reprezentatívnosť:oblastný náhodný výber a kvótový výber.
Keď vychádzame zo všeobecného ponímania reprezentatívnosti je v prípade dobrej realizácie náhodný výber reprezentatívny a obyčajne aj príslušné odhadové štatistiky majú vlastnosť nevychýlenosti výberových odhadov.
Pri výberových skúmaniach však neskúmame obvykle iba jeden štatistický znak a tým je problematika reprezentatívnosti zložitejšia. Obyčajne vyberieme rozhodujúce relevantné znaky a overujeme zhodu ich distribúcií vo výberovom a základnom súbore.
Pri oblastnom náhodnom výbere za rozhodujúce relevantné znaky vyberáme znaky-kritériá podľa ktorých konštruujeme oblasti a pri kvótovom výbere ako napr. výskumy verejnej mienky, sa obvykle vyberajú demografické znaky ako pohlavie, vek, vzdelanie, národnosť, veľkostná skupina obce, kraj, prípadne aj iné znaky v závislosti od tematického zamerania prieskumu.

2.1 Formalizácia definície reprezentatívnosti:

Uvažujme kvalitatívny znak Z s r hodnotami, ktoré okódujeme číslami 1, 2, ..., r.
Označme rozdelenie pravdepodobnosti tohoto znaku v skúmanom základnom súbore:

Л= ($л_1$, $л_2$, ..., $л_r$), kde $\sum_{i=1}^{r} \pi i = 1$

a rozdelenie pravdepodobnosti daného znaku vo výberovom súbore (empirické relatívne početnosti a absolútne početnosti):

P= (p_1, p_2, ..., p_r), kde $\sum_{i=1}^{r} pi = 1$ a N= (n_1, n_2, ..., n_r), pričom $\sum_{i=1}^{r} ni = n$.

V prípade reprezentatívneho výberu platí hypotéza H_0 o zhode uvedených rozdelení pravdepodobnosti. Hypotézu môžeme formulovať nasledovne:

H_0: Rozdelenie N je multinomické M(л,n).
Obvykle stanovíme alternatívnu hypotézu H_1: neplatí H_0.

Reprezentatívnosť v prípade viac znakov môžeme uvažovať za každý znak zvlášť. Výberový súbor potom považujeme za reprezentatívny, ak je reprezentatívny za všetky kontrolované znaky.

3. Testy reprezentatívnosti pre kvalitatívne znaky

3.1 CHí-kvadrát test dobrej zhody

Na overenie hypotézy H_0 používame testy dobrej zhody. Testovacia štatistika CHí-kvadrát testu dobrej zhody je:

$$\mathbf{X^2 = \sum_{i=1}^{r} (n_i - e_i)^2 / e_i, \text{ kde } e_i = n.л_i, \; i = 1, 2, ..., r.}$$

Táto štatistika má asymptoticky X^2 rozdelenie s r – 1 stupňami voľnosti.
Podmienkou použitia testu je splnenie podmienok aproximovateľnosti. Tieto sú známe z teoretických štúdiií a praxe. Ak platí nulová hypotéze a rozsah výberu rastie, tak rozdelenie štatistiky X^2 konverguje ku chi-kvadrát rozdeleniu s r – 1 stupňami voľnosti, pritom však musí aspoň 20% teoretických početností e_i rovných, alebo väčších ako 5 a žiadna z nich nemá byť menšia ako 1. Pre menšie rozsahy výberu sa odporúča aplikovať exaktný test.

3.2 Exaktný test (SPSS)

SPSS generuje exaktné multinomické rozdelenie M(л,n) za platnosti nulovej hypotézy a potom počíta exatnú p-hodnotu testu. Tento test je teda vhodnejší. Pre veľmi rozsiahle výberové súbory postačuje Monte Carlo odhad p-hodnoty testu.

Poznámka:
Pre numerické premenné možno na overenie reprezentatívnosti využiť Kolmogorov-Smirnovov test dobrej zhody alebo hore uvedené testy po vhodnej kategorizácii numerickej premennej.

4. Testy reprezentatívnosti pre kvalitatívne znaky s dvomi hodnotami:

Špecifickým prípadom sú znaky s dvomi možnými hodnotami. Ako príklad môžeme uviesť pohlavie, alebo znak zisťujúci prítomnosť určitej vlastnosti. Kvôli jednoduchosti budeme označovať hodnoty skúmaného znaku číslami 0 a 1.

V tomto prípade sa hypotéza zhody distribúcií vyjadruje vzťahom:
H_0: p=л, kde p je podiel vo výberovom súbore a л je podiel v základnom súbore.

Ako alternatívu možno stanoviť obojstrannú alternatívu H_1: p≠л, alebo jednostranné alternatívy H_1: p<л, resp. H_1: p>л.

Na overenie danej hypotézy, pri výbere o rozsahu n možno použiť:

4.1 Binomický test

Za predpokladu platnosti H_0 počítame prislušnú exaktnú p-hodnotu.
Za predpokladu platnosti nulovej hypotézy má náhodná veličina p binomické rozdelenie B(n, л).

Nech p=i/n, potom pre pravdepodobnosť možných výsledkov platí:

$P(i)= (n!/((n-i)!.i!)\ л^{i}\ (1- л)^{n-i}$, pre i=0,1, 2, ..., r.

a exaktné p-hodnoty počítame podľa vzorcov:

$$\sum_{j=0}^{i} P(j)\text{, ak i <= n. л} \qquad \text{(alternatíva: p < л)}$$

p1 ={ , pre jednostranné alternatívy

$$\sum_{j=i}^{n} P(j)\text{, ak i >= n. л} \qquad \text{(alternatíva: p > л)}$$

$p2 = \sum_{j \in S(i)} P(j)$, kde S(i) = (j : P(j) i <= P(i) , pre obojstrannú alternatívu.

Vyhodnotenie testu: Zvolíne si úroveň významnosti α,
potom ak príslušná p-hodnota < α, hypotézu o zhode zamietame a v prípade keď,
ak p-hodnota > α, tak hypotézu o zhode môžeme na danej úrovni významnosti prijať a výsledok považovať za reprezentatívny.

4.2 z-test

No overenie zhody podielu znaku vo výberovom a základnom súbore možno použiť aj z-test (ak sú splnené podmienky aproximácie, ako sú uvedené v časti 6 a Tabuľke 1).

$$z= (p - л)/\sqrt{(л(1 – л)/n)}.$$

Táto štatistika má asymtoticky N(0,1) zozdelenie.
V tomto prípade môžeme využiť kritické hodnoty štandardizovaného normálneho rozdelenia pre uvedené alternatívne hypotézy.
Pre H_1: p<л je kritická oblasť daná nerovnosťou $z<z_{\alpha}$.
Pre H_1: p<л je kritická oblasť daná nerovnosťou $z>z_{1-\alpha}$.
Pre H_1: p≠л je kritická oblasť daná nerovnosťou $|z|>z_{1-\alpha/2}$. Kde |z| značí absolútnu hodnotu z.
Príklad: Pre α=0,05, platí $z_{1-\alpha}$=1,64, z_{α}= -1,64 a $z_{1-\alpha/2}$=1,96.
Vyhodnotenie testu: Uvedieme na príklade pre α=0,05. Ak hodnota testovej štatistiky z nie je v kritickej oblasti, tak nulovú hypotézu môžeme prijať a môžeme tvrdiť že výberový súbor je reprezentatívny podľa daného znaku.
Teda výberový súbor považujeme za reprazentatívny:
pre H_1: p<л, ak z > -1,64,
pre H_1: p<л, ak z < 1,64 a
pre H_1: p≠л ak |z| < 1,96.

5. Príklad

Uvedieme síce konkrétny, ale anonymizovaný príklad:
Všeobecná zdravotná poisťovňa, a.s. (ďalej VšZP), v spolupáci s ÚVVM, (podľa metodiky, ktorú navrhol autor tejto práce) v decembri 2005 a v januári 2006 uskutočnila reprezentatívny prieskum náhodne vybraných poistencov, ktorí v priebehu 2. až 4. štvrťroka 2005 absolvovali liečbu v niektorej nemocnici alebo v kúpeľnom zariadení na Slovensku (viď: http://www.vszp.sk/showdoc.do?docid=231). Vybraných poistencov oslovila písomne a požiadala ich, aby vyplnili dotazník, ktorý mapoval úroveň ich spokojnosti s konkrétnou nemocnicou alebo kúpeľným zariadením.

Špecifikom prieskumu bol fakt, že každé skúmané zariadenie predstavovalo základný súbor, takže bolo nutné overovať veľa výberových súborov. Na ilustráciu postupu vyberieme tri nemocnice.

Výberové súbory boli konštruované pomocou optimálneho proporcionálneho oblastného náhodného výberu. ***Oblasti*** boli konštruované pomocou vybraných relevantných znakov: pohlavie, vek, pričom znak vek pozostával z dvoch kategórií: 1 = do 50 rokov a 2 = nad 50 rokov. Skombinovaním oboch znakov sme dostali štyri oblasti: 1 "muž do 50 rokov", 2 "muž nad 50 rokov", 3 "žena do 50 rokov", 4 "žena nad 50 rokov". Takto môžeme overiť reprezentatívnosť za vybrané znaky simultánne.

Na overenie reprezentatívnosti vypočítame testy 3.1 - CHí-kvadrát test dobrej zhody a 3.2 – Exaktný test, pričom na interpretáciu budeme využívať exaktný test.

5.1 SPSS procedúry

```
*Najprv vypočítame kombináciu pohl x vek .
IF (o18 = 1 and o19=1) pxv = 1 .
IF (o18 = 1 and o19=2) pxv = 2 .
IF (o18 = 2 and o19=1) pxv = 3 .
IF (o18 = 2 and o19=2) pxv = 4 .
EXECUTE .
 VARIABLE LABELS pxv "pohlavie x vek".
 VALUE LABELS pxv
  1 "muž do 50 r."
  2 "muž nad 50 r."
  3 "žena do 50 r."
  4 "žena nad 50 r."
 .
FORMATS pxv (F5.0) .
VARIABLE WIDTH pxv (5) .


*Testovanie reprezentatívnosti výberových súborov Nemocnice – na ilustráciu
sme vybrali nemocnice č. 7, 24.
************************************************************    7   .
USE ALL.
COMPUTE filter_$=(o16 = 7).
VARIABLE LABEL filter_$ 'o16 = 1 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
* pohlavie x vek .
TITLE " Nemocnica 7" .
NPAR TEST
  /CHISQUARE=pxv
  /EXPECTED=18.00 23.26 26.09 32.64
  /MISSING ANALYSIS
  /METHOD=EXACT TIMER(5).
FILTER OFF.
USE ALL.
EXECUTE .
************************************************************   24   .
USE ALL.
COMPUTE filter_$=(o16 = 24).
VARIABLE LABEL filter_$ 'o16 = 1 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
```

```
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
* pohlavie x vek .
TITLE " Nemocnica 24" .
NPAR TEST
  /CHISQUARE=pxv
  /EXPECTED=19.70 23.12 23.26 33.93
  /MISSING ANALYSIS
  /METHOD=EXACT TIMER(5).
FILTER OFF.
USE ALL.
EXECUTE .
****************.
```

5.2 Výsledky SPSS procedúr: Chi-Square Test a Exaktný test

Nemocnica 7: pxv pohlavie x vek

	Observed N	Expected N	Residual
1 muž do 50 r.	10	11,7	-1,7
2 muž nad 50 r.	17	15,1	1,9
3 žena do 50 r.	13	17,0	-4,0
4 žena nad 50 r.	25	21,2	3,8
Total	65		

Test Statistics

	pxv pohlavie x vek
Chi-Square(a)	2,080
df	3
Asymp. Sig.	,556
Exact Sig.	,560
Point Probability	,001

a 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 11,7.

Nemocnica 24: pxv pohlavie x vek

	Observed N	Expected N	Residual
1 muž do 50 r.	10	13,4	-3,4
2 muž nad 50 r.	15	15,7	-,7
3 žena do 50 r.	17	15,8	1,2
4 žena nad 50 r.	26	23,1	2,9
Total	68		

Test Statistics

	pxv pohlavie x vek
Chi-Square(a)	1,354
df	3
Asymp. Sig.	,716
Exact Sig.	,720
Point Probability	,001

a 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 13,4.

6. Literatúra

1. Agresti A., Coull Ba.: Approximate is better than "Exact" for interval estimation of binomial proportions. The American Statistician. 52:119-126, 1998.
2. Anděl M., Černý R., Charamza P., Neustadt J.: Přehled metod odhadu statistické chyby ve výběrových šetřeních. Informační bulletin České statistické společnosti. Číslo2-3/2004.
3. Cochran G.W.: Sampling Techniques. Third Edition. Wiles, New York 1977.
4. Čermák V.: Výběrové statistické zjišťování. SNTL/ALFA Praha 1980.
5. Elliott, C. and Ellingworth, D. (1997) 'Assessing the Representativeness of the 1992 British Crime Survey: The Impact of Sampling Error and Response Biases' *Sociological Research Online*, vol. 2, no. 4, http://www.socresonline.org.uk/socresonline/2/4/3.html
6. Gourieroux Christian: Théorie des Sondages. INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES, Economica. 1987.
7. Hastings R.: ESOMAR Marketing Research Glossary. http://www.esomar.org/web/show/id=45195

8. Herzmann J., Novák I., Pecáková I.: Výzkumy veřejného mínění. Skriptum VŠE Praha, Fakulta informatiky a statistiky, 1995.
9. Chajdiak J.: Štatistika jednoducho. Statis Bratislava 2003.
10. Kanderová, M. – Úradníček, V.: Aplikovaná štatistika vo finančno-ekonomickej praxi. OZ Financ, Banská Bystrica 2005.
11. Kanderová, M. – Úradníček, V.: Štatistika a pravdepodobnosť pre ekonómov. 1. časť. OZ Financ, Banská Bystrica 2005.
12. Likeš J., Laga J.: Základní statistické tabulky. SNTL, Praha 1978.
13. Luha J.: Testovanie štatistických hypotéz pri analýze súborov charakterizovaných kvalitatívnymi znakmi. STV Bratislava, 1985.
14. Luha J.: Meranie spoľahlivosti výsledkov výskumu verejnej mienky. Zborník príspevkov Využitie štatistických metód v sociálno - ekonomickej praxi, EKOMSTAT´94 29.5.-3.6. 1994 Trenčianske Teplice, SŠDS.
15. Luha J.: Exaktné intervaly spoľahlivosti pre podiely. Slovenská štatistika a demografia 1/96.
16. Luha J.: Matematickoštatistické aspekty spracovania dotazníkových výskumov. Štatistické metódy vo vedecko-výskumnej práci 2003, SŠDS, Bratislava 2003.
17. Luha J.: Reprezentatívnosť vo výskumoch verejnej mienky. FORUM STATISTICUM SLOVACUM 2/2005. SŠDS Bratislava 2005.
18. Mace A. E.: Samle-size determination. Reihold. New York 1964.
19. Motulsky H.: GraphPad PRISM, Version 4.0 Statistical Guide. Statisstical analyses for laboratory and clinical research. 2005. GraphPad Software, Inc. http://www.graphpad.com/
20. Pecáková I.: Statistické aspekty terénních pruskumu I. Skriptum VŠE Praha 1995.
21. Príručky SPSS. Ver. 14.
22. http://www.vszp.sk/showdoc.do?docid=231

Adresa autora

RNDr. Ján Luha, CSc.
Ústav pre výskum verejnej mienky pri ŠÚ SR
(Adresa sídla: Hanulova 5/c, 841 01 Bratislava 42)
Poštová adresa: Miletičova 3, 824 67 Bratislava 26
Jan.Luha@statistics.sk

Ekonomická aplikace statistických klasifikačních metod

Jakub Odehnal[1], Lucie Doudová[2], Jaroslav Michálek[3]

Abstract: The contribution is focused on application of statistical software to classification candidates countries of European Union and for comparison with members countries of European Union. Classification was made by softwares Matlab and Statistica with using cluster and diskriminant analysis. ROC curves popular in medical research were used to comparison variables (voice and accountability, political stability, government effectiveness, regulatory quality, rule of law and control of corruption).

Key words: MATLAB, STATISTICA, Cluster Analysis, Diskriminant Analysis, ROC curve, Governance Matters

1. Úvod

V příspěvku je ukázáno použití standardního statistického software pro klasifikaci možných kandidátských zemí EU a pro jejich srovnání se zeměmi, které již členy EU jsou. Klasifikace byla provedena pomocí shlukové analýzy a diskriminační analýzy s použitím balíků STATISTICA a MATLAB. Kromě toho pro detailní srovnání jednotlivých klasifikačních proměnných byla použita metoda založená na ROC křivkách (Receiver Operator Characteristic Curves), které se v poslední době hojně užívají ke klasifikaci rizikových stavů v lékařském výzkumu i v klinické praxi [3]. Protože při klasifikací zemí EU jsou k dispozici statistická data malého rozsahu (statistický soubor je tvořen potenciálními kandidátskými zeměmi EU a členskými zeměmi EU a omezeným výběrem rozvojových zemí), bylo zapotřebí vyjít z odhadů ROC křivek získaných z náhodných výběrů malého rozsahu. Standardní statistický software, který při odhadu ROC křivky využívá pouze výběrových distribučních funkcí a dává nespojité odhady není pro klasifikaci ekonomických subjektů nejvhodnější. Proto bylo ke srovnání použito hladkých odhadů ROC křivek založených na Kolmogorovském odhadu distribuční funkce [2]. Příslušné programy byly implementovány v prostředí MATLAB.

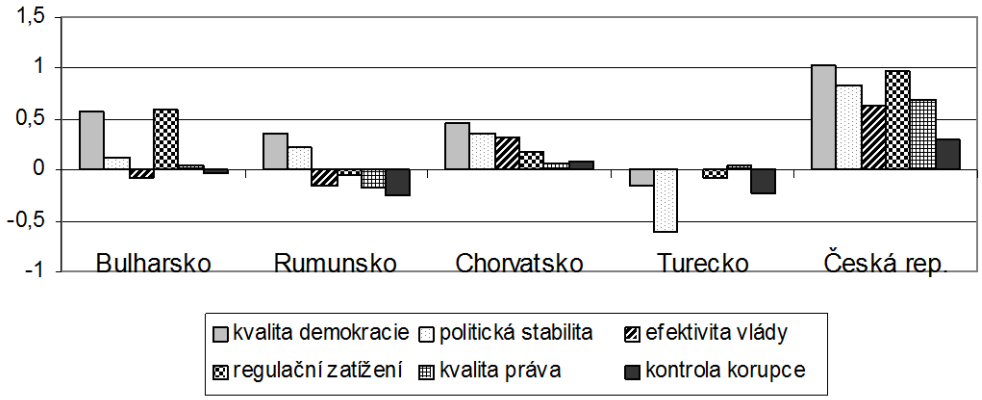
1. Charakteristika datového souboru

Výchozí datový soubor je tvořen současnými členskými státy Evropské unie (25 zemí), které byly pro potřeby dalšího zkoumání rozklasifikovány prostřednictvím shlukové analýzy [4] do třech výchozích shluků (skupin). Jednotlivé sledované proměnné (GM1 kvalita demokracie, GM2 politická stabilita, GM3 efektivita vlády, GM4 regulační zatížení, GM5 kvalita práva, GM6 stav a kontrola korupce) byly získány z mezinárodního projektu Světové Banky Governance Matters a mohou nabývat standardizovaných hodnot na škále -2,5 až 2,5 (vyšší hodnota znamená lepší výsledek). Základní datový soubor byl následně rozšířen o potenciální kandidátské státy (Bulharsko, Rumunsko, Chorvatsko, Turecko), s cílem

[1] Ing. Jakub Odehnal, Katedra aplikované matematiky a informatiky, Ekonomicko správní fakulta, Masarykova univerzita, Lipová 41a, 600 00 Brno e-mail: odehnal@mail.muni.cz
[2] Mgr. Lucie Doudová, Katedra aplikované matematiky, Přírodovědecká fakulta, Masarykova Univerzita, Janáčkovo nám. 2a, 602 00 Brno e-mail: ldoudova@centrum.cz
[3] Doc. RNDr. Jaroslav Michálek, CSc., Katedra aplikované matematiky a informatiky, Ekonomicko správní fakulta, Masarykova univerzita, Lipová 41a, 600 00 Brno e-mail: michalek@econ.muni.cz
Příspěvek vznikl za podpory grantu GAČR 402/04/1308 a grantu pro specifický výzkum ESF MU 561404.

objektivního zařazení příslušných kandidátských zemí charakterizovaných pomocí 6 proměnných do předem vytvořených shluků (skupin). Kvalita rozhodovacích proměnných byla následně hodnocena pomocí grafického zobrazení ROC křivek, znázorňujících vhodnost použití proměnných k výsledné klasifikaci. Pro odlišení členských států Evropské unie od ostatních zemí byla dodatečně vytvořena 4. skupina složená zejména z rozvojových zemí s nízkou ekonomickou úrovní (Kongo, Etiopie, Ghana, Nigerie, Sudán, Yemen, Zambie, Zimbabwe).

Podrobnější charakteristika vybraných států EU byla již provedena v [4], zaměřme se proto pouze na popis jednotlivých charakteristik Bulharska, Rumunska, Chorvatska a Turecka tedy vybraných států, které mají do Evropské unie již významně nakročeno. Pro výchozí porovnání sledovaných států využijme následující obr. 1 zobrazující jednotlivé proměnné. Pro lepší orientaci jsou do grafu vyneseny i hodnoty pro Českou republiku jakožto reprezentanta 3. shluku z předchozí shlukové analýzy. Již na první pohled je zřejmý výrazný rozdíl mezi hodnocenými zeměmi a zemí Evropské unie zastoupené Českou republikou. Přítomnost záporných hodnot u některých veličin Bulharska, Rumunska a Turecka významně napovídá o kvalitě správy ve sledovaných zemích. Kritická situace je především v ukazateli stavu a kontroly korupce, což potvrzuje i umístění sledovaných zemí v žebříčku mezinárodní organizace Transparency International zabývající se právě mezinárodním hodnocením stavu korupce. Výsledné korupční prostředí tak drtivým způsobem doléhá na podnikatelské prostředí sledovaného regionu a snižuje tak jeho konkurenceschopnost a atraktivitu v očích případných investorů. Naopak nejvyšší hodnoty jsou dosahovány (s výjimkou) Turecka u proměnných hodnotící kvalitu demokracie a politickou stabilitu. Průměrný ukazatel kvality správy je nejvyšší (v případě, že ve srovnání neuvažujeme Českou republiku) u Chorvatska (0,25) dále pak u Bulharska (0,21), Rumunska (-0,01) a Turecka (-0,17). Zřejmé tedy je, že na základě sledovaní 6 ukazatelů kvality správy můžeme tedy pozorovat určitou výhodu v relativní připravenosti na vstup do evropského uskupení Chorvatska a Bulharska před Rumunskem a především Tureckem.



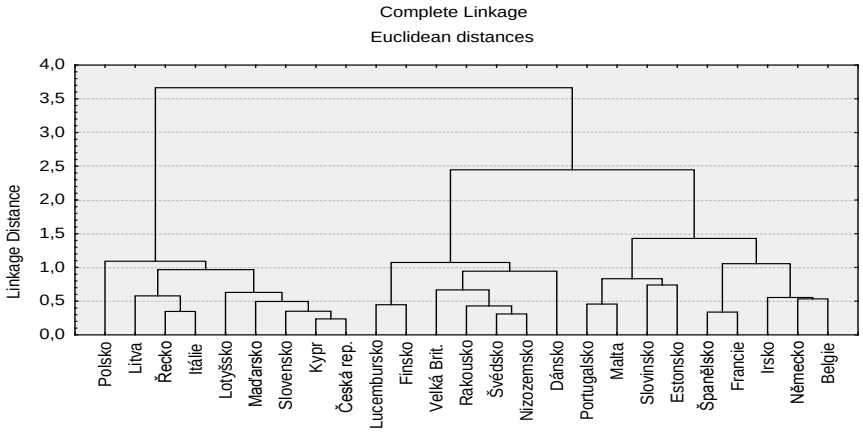
obr. 1. Srovnání kandidátských zemí dle sledovaných proměnných

3. Shluková analýza

Za pomocí programového vybavení programu STATISTICA, (modul Shluková analýza, použita Euklidovská míra vzdáleností jednotlivých objektů a metoda nejvzdálenějšího souseda pro spojování shluků) byl vytvořen následující dendrogram (obr. 2) z něhož je dobře patrná klasifikace sledovaných států Evropské unie do 3 shluků (skupin).

První skupina (na obr. 2 zprava) je tvořena 9 státy a obsahuje převážně „tradiční státy“ Evropské unie. Při následném snížení shlukovací hladiny pod hodnotu 1,5 je patrný rozpad shluku na 2 menší podshluky (Belgie, Německo, Irsko, Francie, Španělsko) a (Estonsko, Slovinsko, Malta, Portugalsko).

Druhý shluk je tvořený Dánskem, Nizozemskem, Švédskem, Rakouskem, Velkou Británií, Finskem a Lucemburskem. Z charakteru dat je dobře patrné, že se jedná o státy s nejvyššími hodnotami sledovaných veličin, tedy o státy s nejvyšší kvalitou správy v rámci Evropské unie.

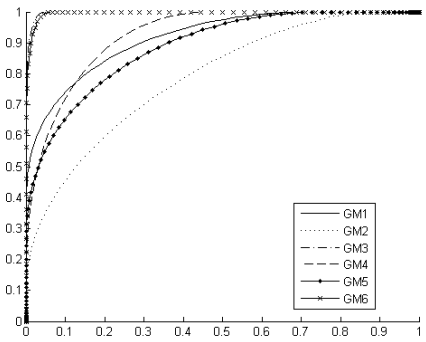


Obr. 2. Dendrogram vytvořený shlukovou analýzou

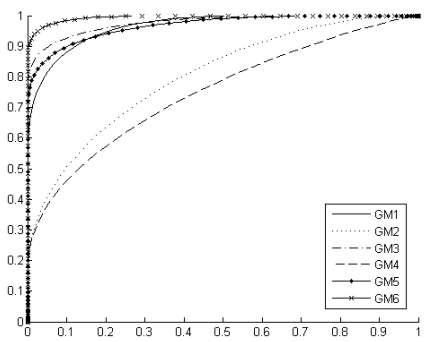
Třetí shluk tvořený Českou republikou, Kyprem, Slovenskem, Maďarskem, Lotyšskem, Itálií, Řeckem, Litvou a Polskem je složen převážně se států (s výjimkou Itálie), které řadíme mezi nováčky v rámci evropského uskupení, jedná se o státy, které prošly náročným procesem transformace a jejich kvalita správy měřená sledovanými proměnnými stále zatím patří mezi nejnižší v rámci EU.

4. Klasifikace zemí pomocí ROC křivek

Kvalita klasifikačního kritéria byla ověřena pomocí zavedení ROC křivek a jejich vzájemného grafického porovnání. ROC křivky byly zavedeny pro všechny sledované klasifikační proměnné (GM1 – GM6), určené k postupnému srovnání jejich kvalit mezi první skupinou států (tradiční země EU) a druhou (nejvyšší kvalita správy), první skupinou (tradiční země EU) a třetí (nově přistoupené země) a třetí skupinou (nově přistoupené země) se čtvrtou (rozvojové země). Kvalitu jednotlivých klasifikačních proměnných zjistíme ze tvaru zavedených ROC křivek (viz obr. 3, 4). Vysoké kvalitě rozhodovacích proměnných a tedy schopnosti správné klasifikace odpovídá takový tvar ROC křivky, kdy křivka nejprve strmě roste a dále je téměř konstantní. Opačné situaci, tedy nízké kvalitě rozhodovací proměnné odpovídá tvar ROC křivky přibližující se diagonále.



obr. 3. země 1 x země 2



obr. 4. země 1 x země 3

Z charakteru tvaru ROC křivek (viz obr. 3) je tedy zřejmé, že při klasifikaci zemí mezi první (tradiční země EU) a druhou skupinou (nejvyšší kvalita správy) můžeme usuzovat především na vysokou důležitost u rozhodovacích proměnných GM3 (efektivita vlády) a GM6 (stav a kontrola korupce). Patrná je tedy zřejmá odlišnost především v těchto dvou ukazatelích kvality správy mezi první skupinou států a druhou. Tento závěr je dobře patrný i ze samotného charakteru proměnných, kdy u druhé skupiny států všechny státy dosahují hodnoty vyšší než 2.0 v ukazateli stavu a kontroly korupce (nejvyšší hodnoty v rámci EU), zatímco u skupiny první žádný ze sledovaných států hodnotu 2.0 vůbec nepřesáhne (Portugalsko 1,9). U ukazatele efektivity vlády je situace téměř obdobná.

Podobný průběh křivek je vidět i na obr. 4 (rozhodování mezi první a třetí skupinou). Pozorujeme tedy opět vysokou kvalitu rozhodovacího kritéria GM6 (stav a kontrola korupce) a GM4 (efektivita vlády), doprovázené ukazateli GM5 (kvalita práva), GM3 (efektivita vlády) a GM1 (kvalita demokracie). O nižší kvalitě rozhodovacího kritéria vypovídá tvar křivek GM2 (politická stabilita) a GM4 (regulační zatížení). Rozdílnost mezi první skupinou států a třetí je tedy především v ukazatelích GM1, GM3, GM5, GM6. Výraznější rozdíly jsou patrné zejména v ukazateli stavu a kontroly korupce, který je v případě nových členských států výrazně níží (ČR 0,30 – 3. nejnižší hodnota v rámci EU), což svědčí o výrazně nepříznivém korupčním prostředí ve sledovaných zemích.

Propastný rozdíl mezi proměnnými kvality správy je však nejvíce patrný při srovnání třetí skupiny zemí (nově přistoupené země) se skupinou čtvrtou (rozvojové země). ROC křivka v tomto případě u všech proměnných kopíruje levý horní roh co svědčí o zmíněném rozdílu. Zřejmé a nikterak překvapivé tedy je, že kvalita správy je výrazně vyšší u členských států EU, zaměřme se tedy dále k jakému výsledku povede samotná diskriminační analýza, jejíž cílem je objektivně klasifikovat nové kandidátské země (Bulharsko, Rumunsko, Chorvatsko, Turecko) do výchozích shluků – skupin států.

4. Diskriminační analýza

Výsledky diskriminační analýzy (zpracované podle [1]) získáváme v podobě následující tabulky:

	Lineární diskriminační analýza				Kvadratická diskriminační analýza			
Stát	d1	d2	d3	d4	D1	D2	D3	D4
Bulharsko	-0,8338	-3,5398	**-0,2216**	-1,3010	2,2098	-0,4962	**2,8220**	1,7425
Rumunsko	-1,9437	-4,2101	-2,0497	**-1,6232**	3,1339	0,8675	3,0279	**3,4544**
Chorvatsko	-2,0392	-3,1469	**-1,8252**	-1,9202	2,9583	1,8506	**3,1723**	3,0773
Turecko	-1,4790	-3,3481	**-0,9663**	-1,3654	1,9705	0,1014	**2,4832**	2,0841

Tab. 1. Lineární a kvadratická diskriminační analýza

Z tabulky vybereme do příslušné skupiny zemí (d1 – d4, D1 – D4) právě tu zemi, které odpovídá nejvyšší hodnota jednotlivých výsledků. Vidíme tedy, že na základě lineární diskriminační analýzy (levá část tab. 1) se 3 země (Bulharsko, Chorvatsko, Turecko) řadí do třetí skupiny zemí (d3) a pouze 1 země (Rumunsko) do skupiny čtvrté (d4). Obdobně při kvadratické diskriminační analýze dosahujeme shodných zařazení sledovaných kandidátských zemích (viz pravá část tab. 1). Důležitost jednotlivých klasifikačních proměnných byla zjištěna za pomoci vypočtení vah těchto proměnných (viz. [1]). Zřejmé tedy je, že na klasifikaci Rumunska do čtvrté skupiny zemí má největší podíl ukazatel GM1 (kvalita demokracie, hodnota vah 3,07), dále pak GM6 (stav a kontrola korupce, 1,02), GM3

(efektivita vlády, -0,24), GM4 (regulační zatížení, -0,76), GM5 (kvalita práva, -1,21) a GM2 (politická stabilita, -2,56).

Pro klasifikace zemí do 1. skupiny se zdá být důležitá především proměnná GM1 (kvalita demokracie), dále pak GM5 (kvalita práva), GM6 (stav a kontrola korupce), GM3 (efektivita vlády), GM4 (regulační zatížení), a GM2 (politická stabilita). Důležitosti proměnných pro klasifikaci zemí do 2 skupiny odpovídá pořadí GM3 (efektivita vlády), GM6 (stav a kontrola korupce), GM1 (kvalita demokracie), GM4 (regulační zatížení), GM2 (politická stabilita) a GM5 (kvalita práva). Pro 3. skupinu zemí je pořadí následující GM1, GM3, GM4, GM5, GM6, GM2.

5. Závěr

Porovnáním výsledků klasifikace pomocí diskriminační analýzy s výsledky použitelnosti ROC křivek je zřejmá takřka shoda v důležitosti proměnných GM3 (efektivita vlády) a GM6 (stav a kontrola korupce), které mají na klasifikaci zemí mezi prvními dvěmi skupinami největší podíl. Prostřednictvím diskriminační analýzy byly klasifikovány potenciální kandidátské země Bulharsko, Chorvatsko a Turecko mezi skupiny zemí EU a pouze Rumunsko shodně u lineární i kvadratické diskriminační analýzy mezi země ostatní.

6. Literatura

[1] Anděl, J.: Matematická statistika. Praha: SNTL/ALFA, 1978.

[2] Michálek, J. - Veselý, V. Odhady ROC a ODC křivky v binormálním modelu pomocí nejlepších nestranných odhadů distribuční funkce. In XXIII. mezinárodní kolokvium o řízení osvojovacího procesu, Sborník abstraktů a elektronických verzí příspěvků na CD-ROMu. Brno : UO 2005, s. 34-39

[3] Zhou X.H., McClish D.K. and Obuchowski N.A.: Statistical Methods in Diagnostic Medicine, Wiley, New York, 2002

[4] Odehnal, J - Michálek, J. Alternativní přístup k hodnocení kvality správy zemí EU. Forum Statisticum Slovacum, Bratislava, SSDS. ISSN 1336-7420, 2006, vol. 2, no. 4, pp. 146-151.

[5] Kaufmann, D., Kraay, A., Mastruzzi, M.: Governance Matters IV: Governance Indicators for 1996-2004. Washington, D.C., World Bank 2005 (Working Paper No. 3630)

Adresa autorů:
Ing. Jakub Odehnal, Doc. RNDr. Jaroslav Michálek, CSc., Mgr. Lucie Doudová
Katedra aplikované matematiky a informatiky
Ekonomicko správní fakulta Masarykovy univerzity
Lipová 41a, 600 00 Brno
odehnal@mail.muni.cz, michalek@econ.muni.cz, ldoudova@centrum.cz

Vzťahy medzi podrobnou a skrátenou úmrtnostnou tabuľkou

Karol Pastor, Jana Simonidesová

Relations between complete and abridged life table. Abstract. The paper describes and compares four methods for the construction of the tables: 1) Abridged life table (ALT) from the complete life tables (CLT), 2) ALT from the data for 5-year age groups, 3) CLT from ALT using the Sprague multipliers and recurrent calculation for number of survivors in exact age x, and 4) CLT from ALT using desaggregation of both table-number of living and dying.

1. ÚVOD

Úmrtnostné tabuľky predstavujú základný matematický model v demografii. Podrobná (úplná) úmrtnostná tabuľka (PUT alebo UUT, angl. *complete life table*) je počítaná pre základné jednoročné vekové intervaly, skrátená úmrtnostná tabuľka (SUT, angl. *abridged LT*) pre viacročné, spravidla 5-ročné vekové intervaly.

Základnou výhodou PUT je, že z nej ľahko možno získať ďalšie informácie. Jej nevýhodou je nielen pomerne veľký rozmer a objemnejšie výpočty, ale i väčšia zaťaženosť náhodnými chybami (tzv. problém malých čísel), nie je teda vhodná pre malé súbory. Ďalej, na výpočet PUT treba mať podrobne členené vstupné dáta, čo často nie je splnené. Ako ukážeme nižšie, túto nevýhodu možno prekonať.

Keďže PUT aj SUT sú dva modely tej istej populácie, hodnoty veličín počítaných z rovnakých dát a majúcich rovnakú interpretáciu by sa mali rovnať. Opak znižuje ich dôveryhodnosť a použiteľnosť. Lenže na výpočet PUT a SUT sa používajú odlišné metódy, ktoré vedú k rozdielnym výsledkom. Jedeným z mnohých príkladov sú tabuľky ŠÚ SR, napr. [10]. V tomto príspevku ukážeme niekoľko metód, ktoré zabezpečujú, aby zhoda medzi údajmi PUT a SUT bola čo najlepšia.

2. Označenie

Použijeme označenie obvyklé v demografickej a aktuárskej literatúre (napr. [1] - [7], [9]). Označme okamžikavé (stavové) veličiny

x vek (nezávisle premenná, *presný* vek), $0 \le x \le \infty$

$l_x = l(x)$ počet dožívajúcich sa (presného) veku x

(obvykle, ale nie nutne $l_0 = 100000$ a $l_x = 0$ pre $x > \omega$);

$T_x = T(x)$ celkový počet rokov, ktoré ešte prežijú v súbore osoby presne x- ročné,

$$T_x = \int_x^\infty l(t)\,dt \ ,$$

$e_x = T_x / l_x$ stredná dĺžka (ďalšieho) života osoby presne x -ročnej.

Pre vekový interval $\langle x; x+n)$ potom označme intervalové (tokové) veličiny

${}_n d_x = l_x - l_{x+n}$ počet zomrelých vo vekovom intervale $\langle x; x+n)$

$_n p_x = l_{x+n} / l_x$ pravdepodobnosť, že osoba presne x-ročná sa dožije veku $x + n$ rokov; (tento stĺpec sa spravidla nepíše)

$_n q_x = 1 - {}_n p_x$ pravdepodobnosť, že osoba presne x-ročná zomrie pred dovŕšením veku $x + n$;

$_n L_x$ celkový počet rokov, ktoré prežijú osoby x-ročné vo vekovom intervale $\langle x; x+n)$, pričom

$$ {}_n L_x = \int_x^{x+n} l(t)\,dt = l_{x+n} + {}_n a_x \; {}_n d_x $$

$_n a_x$ číslo, pre ktoré platí uvedená rovnosť, t.j. priemerný počet rokov, ktoré prežijú vo vekovom intervale $\langle x; x+n)$ osoby, ktoré v tom intervale zomreli (ani tento stĺpec sa spravidla nepíše).

Zrejme platí

$$ {}_n L_x = \sum_{i=0}^{n-1} {}_1 L_{x+i} \qquad \text{a} \qquad T_x = {}_n L_x + T_{x+n} . $$

Pre PUT je n=1 a ľavý index v intervalových veličinách sa vynecháva, čo neskúseného čitateľa môže pomýliť. V SUT je zaužívané písať intervaly $\langle x; x+n)$ v tvare x až $x+n$-1, napr. „5-9“ miesto „⟨5;10)“, čo je tiež zavádzajúce, ale asi sa to nezmení. Poznamenajme, že ŠÚ SR (a predtým FSU) v SUT uvádzal v stĺpci L_x priemer $l_{x+\,n}$ a l_x , od r. 1998 uvádza $_nL_x$ / n. Stĺpce q_x a l_x majú celkom inú interpretáciu.

Predpoklad $_n a_x = n$ /2 zodpovedá predpokladu o rovnomernom rozdelení úmrtí v intervale $\langle x; x+n)$. Pre PUT (n=1) je to postačujúca presnosť, takže a_x = 0,5 s výnimkou vekovej skupiny 0-ročných, kde sa kladie napr. a_0 = 0,1. Pre SUT je rozdiel dosť veľký a môže viesť k značným skresleniam. V práci [7] sú vypočítané a publikovaná koeficienty $_na_x$ pre mužov aj ženy SR za rok 1997. Napr. pre 50-ročných mužov je $_5a_{50}$ = 2,66 ≠ 2,5. Tieto koeficienty sú pomerne stabilné a použiteľné pre aj pre iné súbory toho istého kultúrneho okruhu, resp. v iných rokoch. Je to nástroj, ktorý umožňuje vysporiadať sa so zakrivením $l(x)$ na pomerne dlhom n-ročnom úseku.

3. Výpočet SUT z PUT

Postup pri konštrukcii je nasledovný:

1. Z PUT sa opíšu hodnoty l_x a T_x v riadkoch 0, 1, 5, 10, ... , 85.
2. Dopočítajú sa ostatné hodnoty podľa vzorcov pre x < 85 vo všetkých riadkov okrem posledného,

   $_n d_x = l_x - l_{x+\,n}$ $\qquad$ $_n q_x = {}_n d_x / l_x$ $\qquad$ $_n L_x = T_x - T_{x+n}$ $\qquad$ $e_x = T_x / l_x$
3. V poslednom riadku (85+) ide o vekovú skupinu $\langle x; \infty)$. Keďže tabuľky sú modelom stacionárnej populácie (koľko osôb za rok do skupiny 85+ vstúpi, toľko i zomrie), definuje sa

   $$ L_{85+} = {}_\infty L_x = T_{85} \qquad \text{a} \qquad d_{85+} = {}_\infty d_x = l_{85} , $$

4. Napokon sa doplní (číslo l_{90} prevzaté z PUT) $e_{85} = T_{85} / l_{85}$ a

$$ {}_5q_{85} = 1 - l_{90} / l_{85} \qquad \text{alebo} \qquad q_{85+} = {}_\infty q_x = d_{85+} / l_{85} = 1 . $$

4. Výpočet SUT priamo z dát

Často sa stáva, že vyššie uvedený postup nie je možné použiť, lebo napr. k dispozícii sú iba dáta za 5-ročné vekové skupiny. Popíšeme výpočet využívajúci základné vzorce z § 2.

1. Z pozorovaných dát za 5-ročné skupiny sa vypočítajú špecifické miery úmrtnosti ${}_n m_x$ a odhady pravdepodobnosti úmrtia pre $x < 85$ podľa vzorca

$$ {}_nq_x = \frac{n \, {}_n m_x}{1 + (n - {}_n a_x) \, {}_n m_x} . $$

   Ak nie sú známe jednoročné dáta, nemôžu byť známe ani koeficienty ${}_n a_x$. Preto sa použijú koeficienty už prv vypočítané pre nejakú štandardnú populáciu alebo napr. $n/2$.

2. Vypočítajú sa hodnoty ${}_n d_x$ a l_x podľa známych vzorcov.

3. Vypočíta sa ${}_n L_x = l_{x+n} + {}_n a_x \, {}_n d_x$ a podľa známych vzorcov T_x a e_x.

4. V poslednom riadku (85+) sa položí

$$ l_{85} = (1 - {}_5q_{80}) \, l_{80} \qquad \text{a} \qquad d_{85+} = {}_\infty d_x = l_{85} , $$

5. Pretože pomer d_{85+} / L_{85+} (tabuľková špecifická miera úmrtnosti) sa odhaduje pomerom D_{85+} / P_{85+}, kde D_{85+} resp. P_{85+} je pozorovaný počet zomrelých resp. žijúcich v najstaršej vekovej skupine, položí sa

$$ L_{85+} = d_{85+} \cdot P_{85+} / D_{85+} $$

   a ďalej

$$ T_{85} = L_{85+} \qquad \text{a} \qquad e_{85} = T_{85} / l_{85} . $$

6. Napokon (nie je to nutné) sa doplní $q_{85+} = 1$ alebo častejšie ${}_5q_{85}$ ako (lineárna) extrapolácia predchádzajúcich hodnôt ${}_5q_x$.

Tento výpočet je pomerne citlivý na nepravidelnosti štruktúry reálnej populácie vo veku nad 85 rokov. Na druhej strane, vo veku nad 85 rokov Kingova – Hardyho metóda konštrukcie PUT predpokladá exponenciálny rast intenzity úmrtnosti a ignoruje reálnu populáciu. Z tohoto dôvodu rozdiely medzi PUT a SUT zostrojenou z dát môžu byť vo vyšších vekových skupinách väčšie.

Pre porovnanie v Tab. 1 uvádzame l_x v SUT vypočítanej z PUT, ďalej zo SUT s koeficientami ${}_n a_x$ za rok 1997 prevzatými z [7], $a_0 = 0{,}1$, a potom s koeficientami ${}_n a_x = n/2$. Pre porovnanie je pripojený aj stĺpec z publikácie ŠÚ SR [10]. Pochopiteľne, najväčšia zhoda by sa dosiahla pre koeficienty ${}_n a_x$ zostrojené priamo z dát za rok 2001.

Tab. 1. Porovnanie hodnôt l_x v SUT SR 2001 vypočítaných rôznymi metódami.

	Muži				Ženy			
Vek	exaktne z UUT	${}_5a_x$´97 z dát	${}_5a_x = n/2$ z dát	ŠÚ SR	exaktne z UUT	${}_5a_x$´97 z dát	${}_5a_x = n/2$ z dát	ŠÚ SR
x	l_x	l_x	l_x	l_x	l_x	l_x	l_x	l_x
0	100000	100000.0	100000.0	100000	100000	100000.0	100000.0	100000
1	99316	99317.7	99317.7	99316	99476	99477.2	99477.2	99476
5	99100	99102.8	99102.8	99265	99348	99351.5	99351.5	99444
10	98960	98952.7	98952.6	99172	99278	99283.1	99283.1	99398
15	98817	98804.0	98804.0	99072	99211	99215.3	99215.3	99340
20	98530	98535.4	98535.4	98924	99060	99060.6	99060.6	99259
25	98079	98082.6	98082.7	98686	98932	98940.5	98940.5	99161
30	97478	97485.9	97486.0	98345	98740	98743.9	98743.9	99040
35	96681	96665.7	96666.0	97891	98474	98486.5	98486.6	98878
40	95505	95488.0	95488.8	97257	98074	98083.0	98083.1	98661
45	93495	93480.5	93482.9	96270	97323	97323.3	97323.7	98319
50	90225	90204.9	90211.8	94656	96027	96033.6	96034.7	97732
55	85172	85189.9	85205.4	92059	94085	94115.6	94117.7	96737
60	77920	78033.5	78061.0	88116	91081	91090.9	91096.6	95201
65	67659	67724.1	67779.7	82236	86440	86497.8	86509.9	92808
70	55137	55242.9	55325.3	73930	79287	79341.1	79373.9	89174
75	40879	40829.4	40912.5	63558	68014	68041.3	68124.1	83421
80	26738	26735.8	26749.2	51170	51725	51505.7	51632.8	74329
85	13867	13938.3	13796.6	37486	31330	32262.0	32372.0	60633

5. Výpočet PUT priamo zo SUT - dezagregáciou ${}_nL_x$

Aby bolo možné počítať PUT, treba intervalové 5-ročné údaje dezagregovať na jednoročné. To možno urobiť viacerými spôsobmi, najlepšie však pomocou Spragueových multiplikátorov (napr. [5], [8] a i., kde sú aj popísané). Je to určitý typ kĺzavých priemerov, ktorý zodpovedá interpolácii kumulovaných súčtov ${}_nL_x$ pomocou kubických splajnov. Postup konštrukcie PUT je nasledovný:

1. Zo stĺpca ${}_nL_x$ v SUT sa pripravia päťročné skupiny tak, že prvé dve sa zlúčia, ${}_5L_0 = {}_1L_0 + {}_4L_1$, a posledná ($L_{85+}$) sa rozdelí na štyri v pomere :

   $$0{,}725 : 0{,}230 : 0{,}040 : 0{,}005 .$$

   Tento pomer približne zodpovedá pomeru (získaného z údajov ŠÚ SR), v akom sa delia počty osoborokov v príslušných skupinách v PUT v SR za posledné desaťročia a možno ho teda zovšeobecniť. Je pre mužov aj ženy v podstate rovnaký. Súvisí to s Kingovou - Hardyho metódou graduácie konca PUT. Posledná z týchto skupín je L_{100+}.

2. Skupiny ${}_5L_x$, $x = 0, 5, 10, ..., 95$ sa dezagregujú na jednoročné skupiny L_x, $x = 0, 1, ..., 99$ pomocou Spraguových multiplikátorov.

3. Položí sa $l_0 = 100000$ a vypočítajú sa počty dožívajúcich sa veku x, $x = 1, 2, ..., 100$.

   $$l_{x+1} = 2\,L_x - l_x .$$

4. Dopočítajú sa stĺpce d_x , q_x , T_x a e_x podľa známych vzorcov.

Táto metóda má jednu dôležitú nevýhodu, a totiž že nezaručuje hladký priebeh dopočítaných tabuľkových funkcií, ba dokonca ani ich nezápornosť. Rekurentný spôsob výpočtu zapríčiňuje, že numerické chyby narastajú. Táto nevýhoda sa dá čiastočne eliminovať napr. tým, že hodnoty l_x získané v 3. kroku sa pred ďalším výpočtom vhodne graduujú. Celkový výsledok je potom už uspokojivý, resp., ak odhliadneme od veľmi nízkych a veľmi vysokých vekov, veľmi dobrý.

6. Výpočet PUT z SUT - klasicky cez odhad miery úmrtnosti m_x

Táto metóda využíva dezagregáciu na jednoročné skupiny nielen počtu odžitých rokov ${}_nL_x$, ale aj počtu zomrelých ${}_nd_x$. Predtým treba aj skupinu d_{85+} rozdeliť na štyri, tentoraz ako vhodné sa ukazujú pomery iné pre mužov ako pre ženy.

0,650 : 0,290 : 0,055 : 0,005 (muži)

0,590 : 0,330 : 0,075 : 0,005 (ženy).

V ďalšom kroku sa vypočíta špecifická miera úmrtnosti

$$m_x = d_x / L_x$$

a PUT sa dopočíta klasickou metódou, napr. podľa metodiky ŠÚ SR. Tento postup dáva najvernejšiu rekonštrukciu PUT. Ak sú k dispozícii, možno dezagregovať priamo pozorované 5-ročné počty žijúcich a zomrelých.

7. Záver

Uvedené metódy možno ďalej zdokonaľovať. Treba zistiť, ako by sa dali vybrať nejaké ešte presnejšie a pritom univerzálne koeficienty ${}_na_x$, a tiež pomery, v ktorých sa delia počty žijúcich a zomrelých v poslednej vekovej skupine. Otvorenou otázkou zostáva, ako a kedy možno zaručiť monotónny priebeh po graduácii Spragouvými multiplikátormi. No už teraz vyššie popísané metódy umožňujú kvalitný obojstranný prepočet medzi PUT a SUT.

Literatúra

[1] Browers Jr., N.L., Gerber, H.V., Hickman, J.C., Jones, D.A., Nesbitt, C.J.: *Actuarial Mathematics*. The Soc. of Actuaries, Itasca, 1986.

[2] Chiang, C.L.: *Life Tables and Mortality Analysis*. WHO 1978.

[3] Cipra, T.: *Matematické modely demografie a pojištení*. Praha, SNTL 1990.

[4] Koschin, F.: *Aktuárská demografie (úmrtnost a životní pojištění).* VŠE, Praha 1997.

[5] Keyfitz, N.: *Introduction to the Mathematics of Population with Revisions.* Addison - Wesley, Reading, Mass. 1977.

[6] Newell, C.: *Methods and Models in Demography*. Guilford, New York 1988.

[7] Pastor, K.: *Spresnený výpočet skrátených úmrtnostných tabuliek*. 9. medzinárodný seminár Výpočtová štatistika, Bratislava 7.-8.12.2000. Zborník príspevkov. SŠDS Bratislava 2000, s. 65-69.

[8] Pastor, K.: *Dezagregácia sumovaných údajov pomocou interpolačných polynómov.* In: 13. medzinárodný seminár Výpočtová štatistika, Bratislava 2.-3.12.2004. Zborník príspevkov. SŠDS Bratislava 2004, s. 101-105.

[9] Preston, S.H., Heuveline, P., Guillot, M.: *Demography. Measuring and Modeling Population Processes*. Blackwell, Oxford 2001.

[10] *Úmrtnostné tabuľky za Slovenskú republiku 2001.* ŠÚ SR Bratislava 2002.

Riešené v rámci úlohy VEGA 1/3016/06.

Adresy autorov

Doc. RNDr. K.Pastor, CSc. Katedra aplikovanej matematiky a štatistiky FMFI UK, Mlynská dolina, 842 48 Bratislava. e-mail: pastor@fmph.uniba.sk

Mgr. Jana Simonidesová, Wüstenrot poisťovňa, a.s., Karadžičova 17, 825 22 Bratislava. e-mail: Simonidesova@wuestenrot.sk.

Analýza kategoriálních dat ze dvou závislých výběrů

Iva Pecáková

Abstract: Statistically dependent samples are the samples result from repeated observations on the response variable for a set of subjects (two samples that have the same subjects). Such samples also occur in so called matched-pair designs, when two samples have a natural pairing (married couples, two people rate the same fact, etc.). Both variables have the same categories in this situation and if responses are summarized by a two-way contingency table, this table is square. The paper presents analyses methods for dependent samples classified in such square contingency tables.

Key words: categorical variables, dependent samples, square contingency tables

1. Metodologie

Provádíme-li zjišťování hodnot nějaké kategoriální proměnné u týchž jednotek v časovém odstupu, tedy opakovaně (například v panelových šetřeních), či tvoří-li jednotky ve dvou souborech přirozené dvojice (například manželský pár, rodič s dítětem apod.), jsou výběrové soubory závislé (v prvním případě jde o tzv. longitudinální, ve druhém případě o párově závislá data). Výsledkem dvourozměrného třídění takových údajů je čtvercová kontingenční tabulka r x r, kdy počet řádků v tabulce (r) odpovídá počtu sloupců, neboť kategorie v řádcích a ve sloupcích tabulky jsou stejné. V tabulkách tohoto typu se mnohdy největší četnosti nacházejí na hlavní diagonále. Vzhledem k charakteru dat je obvykle cílem analýzy zjištění, zda je dvourozměrné populační rozdělení, z něhož byl proveden výběr, podle této hlavní diagonály symetrické, a také nakolik se shodují, popřípadě liší obě rozdělení marginální.

Je-li sledovaná proměnná alternativní, tedy $r = 2$, tabulka je čtyřpolní (sdružené absolutní četnosti budeme značit n_{ij}, $i = 1, 2$; $j = 1, 2$, sdružené relativní četnosti $p_{ij} = n_{ij}/n$). Pokud pro populační dvourozměrné rozdělení platí

$$\pi_{12} = \pi_{21} \tag{1}$$

(kdy π_{ij} jsou odpovídající sdružené pravděpodobnosti), je symetrické, a protože v tom případě také

$$\pi_{i+} = \pi_{+i}, \; i = 1, 2, \tag{2}$$

obě populační marginální rozdělení se shodují.Četnost n_{12} ve čtyřpolní kontingenční tabulce má binomické rozdělení s parametry $n_{12} + n_{21}$ a 0,5. P-hodnota dvoustranného testu hypotézy vyjádřené výrazem (1), případně (2), tedy testu symetrie a zároveň marginální homogenity dvourozměrného populačního rozdělení, je rovna dvojnásobku pravděpodobnosti $P[n_{12} \leq \min(n_{12}^*, n_{21}^*)]$, n_{12}^*, n_{21}^* jsou konkrétní ve výběru zjištěné četnosti. Statistika

$$U = \frac{n_{12} - 0,5(n_{12} + n_{21})}{0,5\sqrt{n_{12} + n_{21}}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \tag{3}$$

má potom, jak známo, asymptoticky normované normální rozdělení a její kvadrát

$$U^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \tag{4}$$

tedy asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti (známé McNemarovo testové kritérium).

Je-li sledovaná proměnná vícekategoriální, tedy $r > 2$, hypotézu symetrie vyjádříme jako

$$\pi_{ij} = \pi_{ji}\,,\ i = 1, 2 \ldots, r;\ j = 1, 2 \ldots, r, \tag{5}$$

hypotézu marginální homogenity jako

$$\pi_{i+} = \pi_{+i}\,,\ i = 1, 2 \ldots, r. \tag{6}$$

U symetrického rozdělení jsou marginální rozdělení identická, naopak to platí pouze ve čtyřpolní tabulce, pro $r > 2$ však nikoliv.

K analýze četností v kontingenční tabulce lze použít loglineární model. Četnosti v dvourozměrné kontingenční tabulce vyjadřuje loglineární model

$$\ln m_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij}\,, \tag{7}$$

$$\text{kde}\quad \lambda = \frac{1}{r^2}\sum_i\sum_j \ln m_{ij}\,,\ \lambda_i = \frac{1}{r}\sum_j \ln m_{ij} - \lambda\,,$$

$$\lambda_j = \frac{1}{r}\sum_i \ln m_{ij} - \lambda\,,\ \lambda_{ij} = \ln m_{ij} - \lambda_i - \lambda_j - \lambda\,.$$

Parametry tohoto modelu – celkový efekt λ, efekt řádku λ_i, efekt sloupce λ_j a efekt asociace veličin λ_{ij} – jsou lineární kombinace logaritmů očekávaných četností m_{ij}, $i = 1, 2 \ldots, r$; $j = 1, 2, \ldots, r$. Aby model nebyl přeurčený, musí parametry splňovat nějakou podmínku, například $\Sigma\lambda_i = 0$, $\Sigma\lambda_j = 0$, $\Sigma\lambda_{ij} = 0$, čímž je jejich počet omezen na $1 + r - 1 + r - 1 + (r - 1)^2 = r^2$ a model je tedy saturovaný. Očekávané četnosti m_{ij} se odhadují na základě sdružených četností n_{ij} získaných dvourozměrným tříděním dat do tabulky.

V modelu **nezávislosti** jsou parametry λ_{ij} v (7) rovny nule a očekávané četnosti se odhadují jako $np_{i+}p_{+j}$. Shodu zjištěných a modelem odhadnutých četností lze hodnotit užitím známé Pearsonovy statistiky X^2. Alternativní test je založen na věrohodnostním poměru

$$G^2 = 2\sum_i^r\sum_j^r n_{ij} \ln\frac{n_{ij}}{\hat{m}_{ij}} \tag{8}$$

(deviance). Obě statistiky mají asymptoticky identické chí-kvadrát rozdělení s $(r - 1)^2$ stupni volnosti, jeho uplatnění však je podmíněno (přísněji u G^2) dostatečně velkými očekávanými četnostmi. Pouze druhá z těchto dvou statistik, věrohodnostní poměr G^2, má však vlastnost užitečnou pro konfrontaci různých modelů, a sice tu, že ji lze podle potřeby rozložit na součet odpovídajících složek.

Vzhledem ke koncentraci největších četností na hlavní diagonále čtvercové kontingenční tabulky uvedeného typu model nezávislosti bývá celkem hladce zamítnut. Pro analýzu je zde ovšem zajímavější struktura četností mimo hlavní diagonálu. Pokud eliminujeme při posuzování nezávislosti hlavní diagonálu, hovoříme o **quasi nezávislosti**. Loglineární model zapíšeme jako

$$\ln m_{ij} = \lambda + \lambda_i + \lambda_j + \delta_i I_{ij}, \tag{9}$$

$I_{ij} = 1$ pro $i = j$ a $I_{ij} = 0$ pro $i \neq j$ jsou identifikátory polí na hlavní diagonále. Zjištěné diagonální četnosti tak přímo odhadují očekávané, pro mimodiagonální četnosti však je nutné k pořízení maximálně věrohodných odhadů řešit soustavu nelineárních věrohodnostních rovnic (nejčastěji se používá Newtonův-Raphsonův iterační algoritmus). Vzhledem k počtu parametrů modelu quasi nezávislosti, $p = [1 + 2(r - 1) + r]$, je počet stupňů volnosti chí-kvadrát rozdělení testových kritérií $df = r^2 - p = (r - 1)^2 - r$.

Doplníme-li model (7) o podmínku $\lambda_{ij} = \lambda_{ji}$, získáváme model symetrie. Diagonální očekávané četnosti jsou opět přímo odhadnuty četnostmi získanými tříděním, a jelikož také parametry λ_i, $i = 1, 2 ..., r$, jsou identické pro odpovídající řádkovou a sloupcovou kategorii, lze mimodiagonální očekávané četnosti odhadnout jako $(n_{ij} + n_{ji})/2$. Počet parametrů modelu symetrie,

$$p = 1 + (r - 1) + r(r - 1)/2,$$

pak vede k počtu stupňů volnosti rozdělení testových kritérií $df = r(r - 1)/2$. Pearsonovu statistiku X^2 lze v tomto případě upravit na tvar

$$X^2 = \sum_{i<j} \sum \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}, \tag{10}$$

což je pro $r = 2$ statistika (4).

Především kvůli odlišným marginálním rozdělením model symetrie jen málokdy vyhovuje reálným datům. V modelu **quasi symetrie** je proto podmínka shodných parametrů $\lambda_{ij} = \lambda_{ji}$ vyslovena pouze pro $i \neq j$. Model tak není marginálně homogenní, parametry λ_i, $i = 1, 2 ..., r$ se liší. Symetrické jsou hodnoty poměrů šancí (*odds ratio*)

$$\theta_{ij} = \frac{m_{ij} m_{rr}}{m_{ir} m_{rj}} = \frac{m_{ji} m_{rr}}{m_{jr} m_{ri}} = \theta_{ji} \text{ pro všechna } i,\ j. \tag{11}$$

Diagonální očekávané četnosti jsou opět odhadnuty přímo, maximálně věrohodné odhady ostatních jsou obvykle získávány iteračním řešením soustavy nelineárních věrohodnostních rovnic. Počet stupňů volnosti rozdělení testových kritérií je v tomto případě

$$df = r^2 - [1 + 2(r - 1) + r(r - 1)/2] = (r - 1)(r - 2)/2.$$

Je-li dvourozměrné rozdělení symetrické, je rovněž quasi symetrické a marginální rozdělení jsou homogenní. Je-li dvourozměrné rozdělení quasi symetrické, pak marginální homogenita znamená také jeho symetrii, liší-li se marginální rozdělení, pak je nesymetrické. Hypotézu **marginální homogenity** lze tak testovat konfrontováním hodnoty statistiky G^2 (deviance) modelu symetrie (S) a quasi symetrie (QS):

$$G^2(S / QS) = G^2(S) - G^2(QS). \tag{12}$$

Tato diference má chí-kvadrát rozdělení s (r – 1) stupni volnosti. Starší Stuartův-Maxwellův test používaný pro stejný účel (resp. jeho modifikace – Bhapkarův test) poskytuje obvykle velmi podobné výsledky. Testové kritérium tohoto testu

$$X^2 = \mathbf{d}' \mathbf{S}^{-1} \mathbf{d}, \tag{13}$$

kde $\mathbf{d}' = [d_1, d_2 \ldots, d_{r-1}]$, $d_i = n_{i+} - n_{+i}$, $i = 1, 2..., r - 1$ a $\mathbf{S}$ je kovarianční matice prvků vektoru $\mathbf{d}$, má asymptoticky chí-kvadrát rozdělení s (r – 1) stupni volnosti.

2. Analýza

Data pro následující analýzu pocházejí z průzkumů volebních preferencí realizovaných firmou Factum Invenio, s. r. o., v České republice v červnu 2003 (rok po parlamentních volbách), v dubnu 2004 (na konci vlády premiéra Špidly), v červnu 2005 (po období vlády premiéra Grosse) a v dubnu 2006 (krátce před posledními parlamentními volbami). Ve všech průzkumech odpovídali respondenti na otázky „Kterou stranu jste volil v parlamentních volbách 2002?“ (proměnná „vote 2002“) a „Kterou stranu byste volil, kdyby se volby konaly v tuto chvíli?“ (proměnná „preference“). U každého respondenta tak bylo zjišťováno, zda v uvedeném období své volební preference změnil či nikoliv.

Kategorie sledovaných veličin bylo nutno nejprve sjednotit, a sice proto, že v parlamentních volbách v roce 2002 tvořily strany ČSL a US koalici, po volbách již vystupovaly samostatně (jejich preference bylo tedy nutné nadále shrnovat). Kromě této dvojice jsou sledovanými stranami ČSSD, ODS a KSČM, jiné strany tvoří kategorii „ostatní“. Do šesté kategorie pak jsou zahrnuti nevoliči. Ukázku dat z dubna 2006 obsahuje tabulka 1. Všechny dále uvedené výpočty byly provedeny s užitím výpočetního systému SPSS.

Tabulka 1. Data z června 2006

vote 2002 * preference 2006 Crosstabulation

Count

		preference 2006						
		ČSSD	ODS	KSČM	ČSL, US	other	no vote	Total
vote 2002	ČSSD	157	9	6		12	7	191
	ODS	8	174		1	11	4	198
	KSČM	3	2	87			8	100
	ČSL, US	3	2	1	51	2	1	60
	other	7	2	2		32	4	47
	no vote	14	46	5	7	32	132	236
Total		192	235	101	59	89	156	832

Ve čtvercových kontingenčních tabulkách, které vznikly tříděním výběrových údajů z jednotlivých šetření, jsou podle očekávání největší četnosti soustředěny na hlavní diagonále. Hypotéza nezávislosti tak byla ve všech případech jednoznačně zamítnuta (hodnoty testového kritéria byly 1876 a vyšší při 25 stupních volnosti). Také model symetrie celkem podle očekávání nevyhovuje, marginální rozdělení se zdají být vesměs celkem jasně nehomogenní i bez testu. Hodnoty testových kritérií a p-hodnoty jednotlivých testů jsou obsaženy v tabulce 2.

Pomineme-li hlavní diagonálu v tabulce, ověřujeme hypotézu quasi nezávislosti, resp. quasi symetrie. Výsledky testů jsou obsaženy v tabulce 3. V roce 2003 a 2005 na 5% hladině významnosti k zamítnutí testovaných hypotéz, i když ve druhém případě je výsledek testu nezávislosti celkem těsný. Ke zcela jasnému zamítnutí hypotézy quasi nezávislosti i quasi symetrie dochází u dat z roku 2006. U dat z roku 2004 vedou p-hodnoty těsně k zamítnutí hypotézy quasi nezávislosti, hypotéza quasi symetrie vyvrácena nebyla.

Tabulka 2. Výsledky analýzy

	X^2	p-value	df	G^2	p-value
Year	Symmetry				
2003	51,3	0,00	15	59,2	0,00
2004	83,4	0,00	15	95,7	0,00
2005	55,6	0,00	15	64,4	0,00
2006	83,5	0,00	15	97,8	0,00
	Marginal homogeneity				
2003			5	54,8	0,00
2004			5	82,9	0,00
2005			5	51,3	0,00
2006			5	67,2	0,00

Tabulka 3. Výsledky analýzy

	X^2	p-value	df	G^2	p-value
Year	Quasi independence				
2003	21,4	0,32	19	25,9	0,13
2004	29,7	0,05	19	30,7	0,04
2005	27,9	0,09	19	28,1	0,08
2006	44,6	0,00	19	47,9	0,00
	Quasi symmetry				
2003	4,6	0,92	10	4,4	0,93
2004	11,8	0,30	10	12,8	0,24
2005	11,9	0,29	10	13,1	0,22
2006	29,1	0,00	10	30,6	0,00

Tabulka 4. Znaménková schémata

		ČSSD	ODS	KSČM	ČSL,US	other	no
ČSSD	QI/04	.	++	.	.	.	.
	QI/06	.	.	.	.	++	– – –
	QS/06	.	.	–	.	.	.
ODS	QI/04	– –	.	–	+	.	+
	QI/06	.	.	.	.	– –	+++
	QS/06	.	.	.	.	– –	+
KSČM	QI/04	+++	.	.	.	.	.
	QI/06	+++	.	.	.	.	.
	QS/06	+	.	.	.	+++	– – –
ČSL,US	QI/04	.	.	.	.	.	.
	QI/06	.	.	.	.	.	+
	QS/06	.	.	.	.	.	.
other	QI/04	.	.	.	.	.	.
	QI/06	.	.	– –	.	.	.
	QS/06	.	++	– –	.	.	.
no vote	QI/04	.	.	+++	.	.	.
	QI/06	.	.	+++	.	.	.
	QS/06	.	–	+++	.	.	.

Pro podrobnější rozbor výsledků testů sestavme ještě znaménková schémata (tabulka 4). Pokud byl některý výše uvedený test významný (model quasi nezávislosti 2004 a 2006, model

quasi symetrie 2006), byla spočtena adjustovaná rezidua. Znaménky v jednotlivých polích tabulky jsou vyznačena rezidua překračující kritické hodnoty na hladině významnosti (pro individuální, nikoliv simultánní testy) 0,1 (jedno znaménko), 0,05 (dvě znaménka), resp. 0,01 (tři znaménka).

Z tabulky 4. lze vyčíst některé změny ve voličských preferencích.

Rok 2004 (konec vlády premiéra Špidly):

- určitý přesun sympatií voličů ČSSD k ODS;
- u voličů KSČM přechod k ČSSD a naopak významně více nevoličů udávajících sympatie KSČM, než by odpovídalo modelu quasi nezávislosti.

Rok 2006 (před parlamentními volbami):

- přechod některých voličů KSČM k ČSSD. Touto skutečností je obecně zdůvodňován nižší volební výsledek KSČM a vyšší volební výsledek ČSSD v červnových parlamentních volbách – voliči KSČM se rozhodli svou volbou patrně podpořit ČSSD, a tedy vyhlídky na případné volební vítězství levice. V tabulce je však rovněž patrný přechod voličů KSČM k ostatním, menším stranám (Zelení?), a dále určitá volební „disciplinovanost“ voličů KSČM, neboť mezi momentálně rozhodnutými nevolit jich je významně méně, než by odpovídalo modelu quasi symetrie
- Mezi osobami, jež v posledních volbách nevolily, je opět významně více nových příznivců KSČM.
- Příznivci ODS odcházeli k jiným stranám významně méně, než by odpovídalo modelům, spíše se rozhodovali nevolit.
- K ODS přešlo více voličů některých menších stran, a od ODS k nim odešlo méně voličů, než by odpovídalo modelu quasi symetrie.

Je samozřejmě třeba zaznamenat, že analyzované čtvercové kontingenční tabulky jsou řídké a mnohé odhadované četnosti jsou malé. Korektnost používaných rozdělení testových kritérií tak vyvolává pochybnosti. Podle [1] lze ukázat, že při použití testových kritérií založených na konfrontaci modelů, jako je např. kritérium (12), je konvergence k chí-kvadrát rozdělení daleko rychlejší. Nicméně pochybnost existuje a výsledky zde provedené analýzy považujme proto jen za orientační. Prověření použitelnosti uvedené metodologie v řídkých kontingenčních tabulkách se hodláme věnovat v nějakém dalším příspěvku.

3. Literatura:

[1] AGRESTI, A.: Categorical Data Analysis, John Wiley & Sons, 1995
[2] ANDĚL, J.: Matematická statistika, SNTL, Praha 1978
[3] JOBSON, J. D.: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods, 1991
[4] SPSS Manuals, SPSS Inc., 1994 –1999
[5] SIMONOFF, J. S.: Analyzing Categorical Data, Springer-Verlag Inc., New York 2003
[6] STOKES, M. E.- DAVIS, C. S.- KOCH, G. G.: Categorical data Analysis Using the SAS System, SAS Institute Inc., 1995

Doc. Ing. Iva Pecáková, CSc.
nám. W. Churchilla 4
130 67 Praha 3
e-mail: pecakova@vse.cz

Aproximačná webová služba

Martina Révayová, Csaba Török

Abstract

The paper deals with a continuous global piecewise cubic approximation of noisy data based on a local model. The method is implemented as a Web application and a Web service.

Úvod

V [1] bol predstavený kubický model na riešenie lokálnych aproximačných úloh. V [2] sme sa zaoberali zovšeobecnením daného aproximačného modelu. V terajšej práci ukážeme využitie lokálneho modelu na riešenie globálnej úsekovej aproximácie pomocou distribuovanej Webovej aplikácie, implementovanej vo Visual C# v prostredí MS .NET Framework [3] .NET dovoľuje prepojiť systémy, ľudí a informácie takými technológiami, ako napr. Remoting [4] alebo Webové služby (Web services). Tu sa budeme opierať o Webové služby ako celky kódu, knižnice na vzdialených počítačoch, serveroch.

Prvá časť obsahuje stručný popis nášho lokálneho modelu a prístupu ku globálnej spojitej aproximácii. Ďalšia časť predstaví priemyselné štandardy pre Webové služby: XML, SOAP, WSDL a UDDI. Posledná časť uvádza najdôležitejšie kódové riadky Webovej aplikácie pomocou Webovej služby.

1 Úseková kubická aproximácia

V tejto časti uvedieme aproximačný model polynómu n-tého stupňa

$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n .$$

Reprezentácia polynómu vychádza z priamej a inverznej Diskrétnej Projektívnej Transformácie [1,2].

Veta: Nech stupeň polynómu $n \geq 3$ a majme body $[x_0, P_n(x_0)]$, $[x_1, P_n(x_1)]$, $[x_2, P_n(x_2)]$. Potom polynóm môže byť vyjadrený v tvare

$$P_n(x) = I(x) + Z(x)A(x), \qquad (1)$$

kde

$$I(x) = p_0(x)P_n(x_0) + p_1(x)P_n(x_1) + p_2(x)P_n(x_2),$$

$$p_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)},\ p_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)},\ p_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)},$$

$$Z(x) = (x - x_0)(x - x_1)(x - x_2),$$

$$A(x) = a_3 + a_4 T_1(x) + \ldots + a_n T_{n-3}(x),$$

$$T_i(x) = x^i + R_1 x^{i-1} + \ldots + R_{i-1}x + R_i,\ R_i = \sum_{k_0=0}^{i} \sum_{k_1=0}^{i-k_0} x_0^{k_0} x_1^{k_1} x_2^{i-k_0-k_1} .$$

Model (1) má niekoľko základných vlastností:

Článok bol podporený grantovým projektom VEGA 1/1006/04 9150/MŠ

- $I(x_0) = P_n(x_0), I(x_1) = P_n(x_1), I(x_2) = P_n(x_2),$
- $Z(x_0) = 0, Z(x_1) = 0$ a $Z(x_2) = 0,$
- $n-3$ parametrov.
- y_0, y_1, y_2 sú riadiace parametre, aproximácia prechádza cez body $[x_0, P_n(x_0)]$, $[x_1, P_n(x_1)]$, $[x_2, P_n(x_2)]$

Dôsledok: Polynóm $P_3(x)$ môžeme vyjadriť v tvare

$$P_3(x) = I(x) + Z(x)A. \qquad (2)$$

Môžeme si všimnúť, že v tejto reprezentácii vystupuje jeden parameter $A = a_3$.

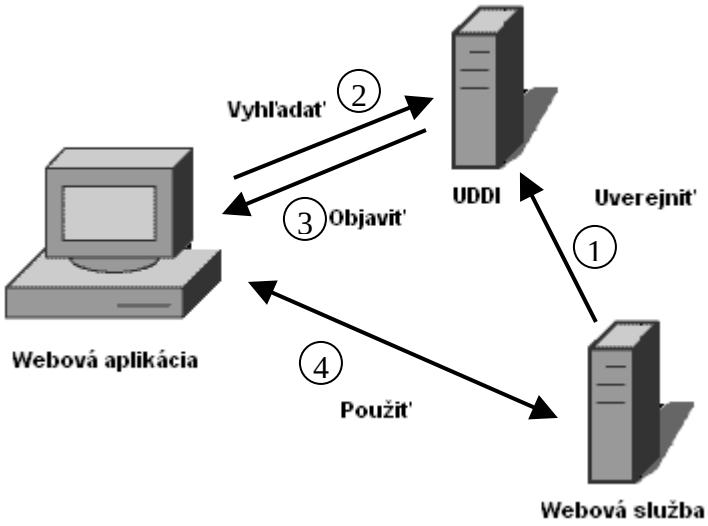
Uvažujme zašumené dáta $\{[x_i, \tilde{y}_i], i = \overline{0, N}\}$, ktoré rozdelíme na niekoľko úsekov, na ktorých vieme nájsť lokálne kubické aproximácie. Na nepárnych úsekoch použijeme klasický kubický regresný polynóm. Párne úseky aproximujeme pomocou modelu (2), v ktorom potrebujeme tri body. Dva body si môžeme zvoliť z okrajových bodov susedných nepárnych úsekov x_k, x_K, ktoré sme odhadli klasickou regresiou. Za tretí bod si zvolíme $x_k + h = x_1$, pre ktorý $\hat{y}_1 = y_k' h + y_k$. Teda poznáme odhady bodov $[x_k, \hat{y}_k]$, $[x_1, \hat{y}_1]$, $[x_K, \hat{y}_K]$. Na základe kubického aproximačného modelu (2) odhadneme parameter $\hat{A}$ vzťahom

$$\hat{A} = \frac{\sum (\tilde{y}_i - \hat{I}_i) z_i}{\sum z_i^2}.$$

Vďaka štvrtej vlastnosti takto zostrojená globálna kubická aproximácia je spojitá bez použitia splajnov.

2 Webové služby

Webové služby sú aplikácie, ktoré na výmenu informácii používajú štandardizovaný prenos, kódovanie a protokoly. Webové služby sú revolučné tým, že poskytujú univerzálny dátový formát pre komunikáciu s inými aplikáciami. Umožňujú počítačovým systémom na ľubovoľnej platforme komunikovať cez intranety, extranety a Internet.



Obr.1 Infraštruktúra práce s Webovou službou

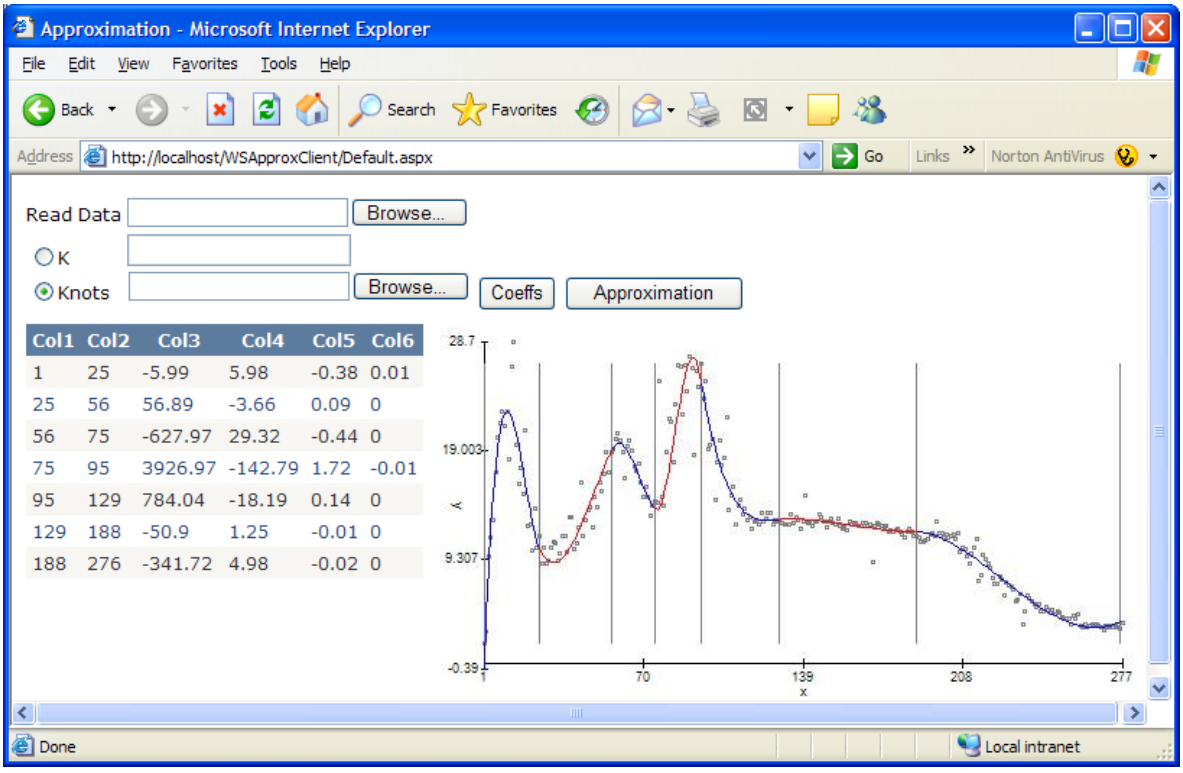
Infraštruktúra webových služieb je tvorená na základe nasledujúcich štandardov, ktoré popisujú syntax a sémantiku programovej komunikácie:

- XML - Extensible Markup Language
- SOAP - Simple Object Access Protocol
- WSDL - Web Services Description Language
- UDDI - Universal Description, Discovery, and Integration

XML poskytuje syntax na reprezentáciu dát. Na základe XML Webové služby môžu komunikovať cez platformy a operačné systémy, bez ohľadu na programovací jazyk, v ktorom je aplikácia napísaná.
SOAP poskytuje sémantiku pre výmenu dát. Je to protokol určený na výmenu štrukturovaných informácii v distribuovanom prostredí založenom na XML. SOAP je založený na XML správach. Bol štandardizovaný konzorciom W3C, ktoré špecifikuje všetky potrebné pravidlá pre umiestnenie Webovej služby, jej začlenenie do aplikácie a komunikácie medzi nimi.
WSDL poskytuje mechanizmus na opísanie schopností Webových služieb.
UDDI je verejný register kde je možné uverejňovať vlastné Webové služby a získavať informácie o iných Webových službách.

3 Webová aplikácia

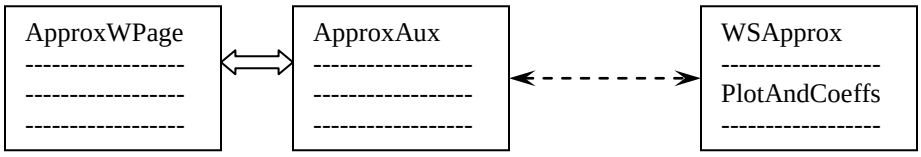
Táto časť demonštruje použiteľnosť uvažovanej úsekovej aproximácie na reálnych údajoch a uvádza najpodstatnejšie kódové riadky interaktívnej Webovej aplikácie a Webovej služby.



Obr.2 Web aplikácia a apr. prierezov π–p kolízie

Obr. 2 zobrazuje aplikáciu, v ktorej treba načítať dáta a zadať buď počet segmentov, alebo ich koncové body. Obrázok ukazuje globálnu úsekovú aproximáciu prierezov π–p kolízie [P] a koeficienty príslušných lokálnych kubických aproximácií.

Obr.3, ako časť obrázku 1, schematicky ukazuje dve triedy *ApproxWPage*, *ApproxAux* Webovej aplikácie a dve triedy *WSApprox*, *PlotAndCoeffs* Webovej služby. *ApproxWPage* je hlavnou stránkou aplikácie a *WSApprox* je Webová služba. Pomocná stránka je nutná k



Obr.3 Hlavná a pomocná stránka Webovej aplikácie a dve triedy Webovej služby

zobrazeniu Grafických výstupov, ako je to popísané v [T]. Pomocná štvorriadková trieda

```
public class PlotAndCoeffs  {
   public byte[ ] BArrayImage;    // graf sa posiela ako byte array
   public string[ ] Coeffs;       // koeficienty kubických aproximácií
}
```

zjednodušuje získanie z Webovej služby obrázku ako byte array a koeficientov kubických polynómov jednotlivých segmentov.

Po klepnutí na knoflík *Approximation* sa vykoná metóda *butApprox_Click*, ktorá nastaví hodnoty štyroch session premenných [T] pre komunikáciu s pomocnou stránkou *WSApproxAux.aspx* a načíta ju do riadiacieho prvku *Image1:*

```
//class ApproxWPage : System.Web.UI.Page
protected void butApprox_Click(object sender, EventArgs e)  {
   string savePath;
   butCoeffs.Enabled = true;
   Session["RadioBut"] = RadioButtonList1.SelectedItem.Text;
   Session["KValue"] = TextBox1.Text;
   if (this.FileUpload1.HasFile)
   {
      savePath = HostingEnvironment.ApplicationPhysicalPath + FileUpload1.FileName;
      FileUpload1.SaveAs(savePath);              // názov súboru s x, y
      Session["xyFileName"] = savePath;
   }
   if (this.FileUpload2.HasFile)
   {
      savePath = HostingEnvironment.ApplicationPhysicalPath + FileUpload2.FileName;
      FileUpload2.SaveAs(savePath);              // názov súboru s uzlami
      Session["nodsFileName"] = savePath;
   }
   this.Image1.ImageUrl = "WSApproxAux.aspx";// zobrazí graf z w.služby cez pomocnú stránku
}
```

Komunikáciu s aproximačnou Webovou službou vykoná metóda Page_Load pomocnej Web stránky:

```
//class ApproxAux : System.Web.UI.Page
protected void Page_Load(object sender, EventArgs e)  {
   readXY(Session["xyFileName"].ToString());
```

```
    localhost.WSApprox ws = new localhost.WSApprox(); // inštancia webovej služby
    localhost.PlotAndCoeffs pc;                       // inštancia PloAndtCoeffs
    if (Session["RadioBut"].ToString() == "K")
    {
        k = Int32.Parse(Session["KValue"].ToString());
        pc = ws.Approximation1PointsNb(k, _x, _y, w, h);
    }
    else
    {
        knots = VectorD.FileRead(Session["nodsFileName"].ToString());
        pc = ws.Approximation2Knots(knots, _x, _y, w, h);
    }
    coeffsToFile(pc.Coeffs);                  // načíta to GriedView z ApproxWPage
    bitmapToOutputStream(pc.BArrayImage);     // zobrazí to Image1 z ApproxWPage
}
```

Posledné kódové riadky ilustrujú z implementačného hľadiska najpodstatnejšie miesta Webovej služby:

```
[WebService(Namespace = "http://tempuri.org/")]
[WebServiceBinding(ConformsTo = WsiProfiles.BasicProfile1_1)]
public class WSApprox : System.Web.Services.WebService  {
...
[WebMethod]
public PlotCoeffs Approximation2Knots(double[] knots, double[] dataX, double[] dataY,
                                      int w, int h)    {
        iniData(dataX, dataY, w, h);
        _ppI = _x.IndicesOfNear(knots);
        iniSubIntervals();
        return plotAndCoeffs();                  // hlavná metóda aproximácie a vizualizácie
    }
...
}
```

Záver

V práci sme sa venovali úsekovej kubickej aproximácii, pre ktorú sme vytvorili Webovú aplikáciu a Webovú službu. Zatiaľ sme Webovú službu pre aproximáciu odskúšali lokálne http://localhost/WSApprox/WSApprox.aspx ale v budúcnosti ju plánujeme zverejniť.

Literatúra

[1] N.D. Dikoussar, Cs. Török, Automatic Knot Finding for Piecewise-Cubic Approximation. Mathem. Model., 18, 3, (2006), 23-40
[2] M. Révayová , Cs. Török, Piecewise Approximation and Neural Networks, Kybernetika, sent for publication, 2006
[3] Gunnerson E.: Začíname programovat v C#, Computer Press, Praha, 2001
[4] Révayová M., Török Cs.: Distribuované počítanie v MS .NET Framework, Forum Statisticum Slovacum, 3/2005, Bratislava, ISSN 1336-7420, pp. 215-222
[5] Freeman A., Iannuzzi S., Jones A.: Beginning ASP.NET 2.0 Web Services in C#
[6] N.D. Dikoussar, Cs. Török, Data Smoothing by Splines with Free Knots, 2006 Tatry
[T] Török Cs., Graphics and Data Visualization in ASP.NET, 12-th International Workshop of Computational Statistics, Bratislava 2003, pp.122-126
[P] http://pdg.lbl.gov/2006/hadronic-xsections/hadron.html

Kontakt: Martina.Revayova@tuke.sk, Csaba.Torok@tuke.sk

Metody odhadu ROC křivky [1,2]

Marek Sedlačík, Jaroslav Michálek

Abstract: The receiver operating characteristic (ROC) curve plays an important role in many branches when it is necessary to classify between two populations. In the contribution the attention is concentrated on methods of ROC curve estimation. Three approaches are discussed: classical estimator based on the empirical CDF, the weighted regression estimator, the estimator based on the best unbiased CDF estimator. Then the procedures are applied to the medical real data processing.

Key words: ROC curve, binormal model, latent random variable, estimation

1 Úvod

V tomto příspěvku se zaměříme na různé techniky odhadu ROC křivky (Receiver Operator Characteristic Curve). Budeme uvazovat data spojitého typu i data kategorizovaná. Výpočty a grafické výstupy byly implementovány v prostředí MATLAB 7.0.

2 ROC a ODC křivka

Uvažujme diagnostický test, jehož cílem je na základě měření na spojité škále zařadit objekt do jedné ze dvou disjunktních skupin. Například do skupiny „zdravý“ nebo do skupiny „nemocný“. Předpokládáme, že měřené testové skóre je ve skupině zdravých pacientů náhodná veličina X, která má rozdělení pravděpodobností dané distribuční funkcí F. Ve skupině nemocných pacientů je měřené testové skóre náhodná veličina Y s rozdělením pravděpodobností o distribuční funkci G. Předpokládáme, že obě náhodné veličiny X a Y jsou nezávislé.

ROC křivku, která popisuje kvalitu testového kritéria, zavádíme analyticky vztahem $ROC(t) = 1 - G(F^{-1}(1-t))$ pro $t \in (0,1)$, pokud uvedená inverzní funkce $F^{-1}(t)$ existuje. Obdobně zavádíme ODC křivku (Ordinal Dominance Curve) jako $ODC(t) = F(G^{-1}(t))$, $t \in (0,1)$. ODC křivka charakterizuje kvalitu testového kritéria analogicky jako ROC křivka, ovšem po formální stránce je její funkční předpis jednodušší. Ekvivalentně může být ROC a ODC křivka vyjádřena parametricky pomocí senzitivity a specificity testového kritéria (viz např. [4]).

Diagnostický test, který má dobrou rozlišovací schopnost mezi oběma populacemi (viz [5]), je charakteristický tím, že jeho ROC křivka zpočátku rapidně roste a potom je téměř konstantní. Naopak u diagnostického testu s malou rozlišovací schopností se jeho ROC křivka přibližuje diagonále. Základní vlastnosti ROC křivek lze nalézt např. v [2].

3 Empirický odhad ROC křivky

Nechť $X_1,\ldots,X_m$ je náhodný výběr z rozdělení o distribuční funkci F. Označme $F_m(x)$ výběrovou distribuční funkci příslušnou náhodnému výběru $X_1,\ldots,X_m$. Podobně pro náhodný výběr $Y_1,\ldots,Y_n$ z rozdělení o distribuční funkci G označíme příslušnou výběrovou distribuční funkci $G_n(x)$.

[1]Příspěvek byl vypracován jako součást řešení grantu GA ČR č. 402/04/1308.
[2]Příspěvek byl vypracován jako součást řešení grantu VZ04-FEM-K01-13-SJA.

Když v definičním vzorci ROC křivky nahradíme distribuční funkce F a G příslušnými výběrovými protějšky $F_m(x)$ a $G_n(x)$ a místo inverzní funkce použijeme kvantilovou funkci dostaneme empirický odhad ROC křivky

$$\overline{ROC}(t) = 1 - G_n\left(F_m^{-1}(1-t)\right), \quad 0 < t < 1.$$

Tento odhad budeme v dalším textu označovat jako "Sample ROC". Jeho vlastnosti lze nalézt např. v [2].

4 Binormální model

Další odhady, jimiž se budeme zabývat, budou vycházet z předpokladu, že obě náhodné veličiny X i Y mají normální rozdělení. Mluvíme potom o binormálním modelu. V některých situacích lze k binormálnímu modelu dospět tak, že předpokládáme existenci monotónní transformace H, která simultánně transformuje F a G na distribuční funkce normálního rozdělení. Tedy po transformaci H lze bez újmy na obecnosti předpokládat, že platí

$$H(X) \sim N(0,1) \text{ a } H(Y) \sim N(\mu, \sigma^2).$$

V binormálním modelu potom pracujeme s náhodnými veličinami *H(X)* a *H(Y)* místo s X a Y. V tomto modelu lze ROC křivku analyticky vyjádřit a když odhadneme distribuční funkce F a G jejich výběrovými protějšky, lze odhady parametrů μ a σ získat pomocí váženého lineárního regresního modelu.

4.1 Odhad ODC křivky pomocí zobecněné metody nejmenších čtverců

Pro odhad parametrů μ a σ lze použít iterativní váženou metodu nejmenších čtverců (metoda GLS - generalized least squares method). Podrobnosti např. v [2] .

4.2 Adaptabilní procedura pro odhad ROC křivky

Předchozí metodu lze modifikovat tak, že se ve výpočtu zohlední rychlost růstu ROC křivky (metoda AGLS - adaptive generalized least squares method) – viz [2].

4.3 Odhad ROC křivky založený na nejlepším nestranném odhadu distribuční funkce

Opět vyjdeme ze vztahu $ROC(t) = 1 - G(F^{-1}(1-t))$ pro $t \in (0,1)$, kde distribuční funkce F a G nahradíme jejich nejlepším nestranným (kolmogorovským) odhadem. Simulační výsledky ukazují (viz [3]), že odhady ROC a ODC křivek založené na nejlepším nestranném odhadu distribuční funkce EBBUCE (z anglického Estimate Based on the Best Unbiased Cdf Estimate) jsou poměrně kvalitní a vyniknou zejména při malých rozsazích výběrů. Podrobnosti viz [3].

4.4 Odhad ROC křivky pro kategorizovaná data

Při odhadu ROC křivek pro diskrétní nebo setříděná data (viz [4]) se běžně používá procedura, kterou navrhli Dorfman a Alf (viz [1]). Zde se zaměříme na proceduru navrženou v [2], která předpokládá existenci diagnostické latentní náhodné veličiny *W*. Odhad ROC křivky se opět provádí za předpokladů binormálního modelu.

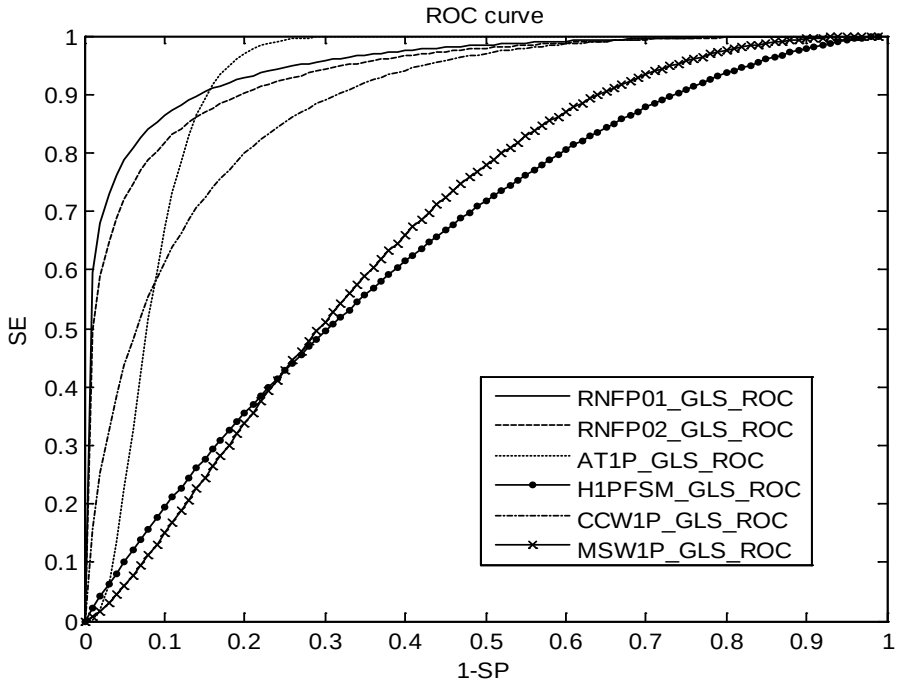
5 Programové implementace

Pro odhad ROC křivek některou z výše popsaných metod byl v systému MATLAB vytvořen program, který při vhodné volbě parametrů počítá požadované typy odhadů ROC křivek, případně také odhady parametrů μ a σ postupem, jež odpovídá zvolené metodě.

6 Příklad

Na závěr srovnáme popsané techniky odhadu ROC křivky na reálných datech, které pochází z prostředí očního lékařství Fakultní nemocnice u svaté Anny v Brně. Úkolem je ověřit klasifikační schopnosti dostupných prediktorů při diagnostice zeleného zákalu (glaukomu). Důvodem je skutečnost, že jen včasná diagnóza této nemoci a okamžité zahájení léčby vedou k zpomalení progrese choroby, která je nevyléčitelná. I z ekonomického hlediska je důležité, co nejlépe rozlišit zdravé a nemocné, protože jakmile jednou zařadíme osobu mezi nemocné, zahájí se doživotní ekonomicky nákladná léčba.

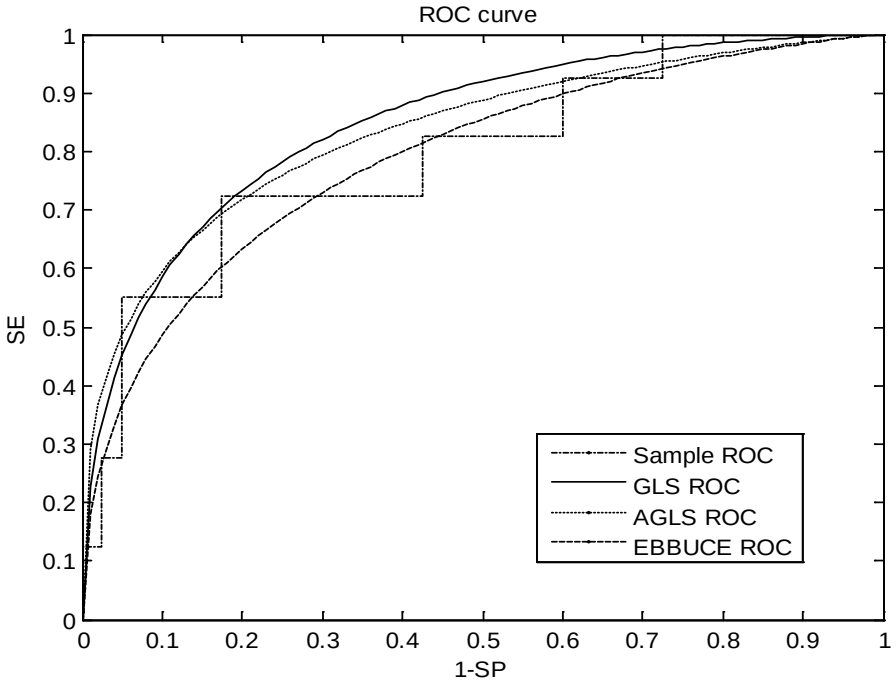
Datový soubor obsahuje celkem 80 pacientů, z nichž 40 bylo zdravých a 40 nemocných, pro ilustraci odhadů ROC křivek se omezíme pouze na diagnostiku pravého oka (více viz [6]). U každého pacienta je k dispozici 28 spojitých prediktorů. Klasifikační sílu jednotlivých prediktorů lze demonstrovat pomocí ROC křivek. Cílem bude porovnat různé metody odhadu na reálných datech a pro daný pevně zvolený odhad provést srovnání vybraných prediktorů. Klasifikační schopnosti prediktorů RNFP01 (hodnocení vrstvy nervových vláken sítnice pozorovatelem č.1), RNFP02 (hodnocení vrstvy nervových vláken sítnice pozorovatelem č.2), AT1P (hodnota nitroočního tlaku), H1PFSM (hodnota Mikelbergovy diskriminační analýzy), CCW1P (centrální citlivost zorného pole bílé perimetrie) a MSW1P (průměrná citlivost zorného pole bílé perimetrie) provedeme pomocí odhadnuté ROC křivky metodou GLS (více viz [6]). Výsledky jsou na obr. 1.



Obrázek 1: ROC křivky pro jednotlivé proměnné spočtené metodou GLS (viz odstavec 3.3)

Z tvaru ROC křivek lze usoudit, že pro detekci glaukomu lze využít klasifikačních schopností především proměnných RNFP01, RNFP02, AT1P a CCW1P. Naopak klasifikace pacientů pomocí proměnných H1PFSM a MSW1P vykazuje velký podíl chybně klasifikovaných pacientů.

Abychom demonstrovali vlastnosti popsaných odhadů v kapitole 3, provedeme odhad ROC křivky u proměnné CCW1P. Výsledky jsou na obr. 2 a celkem výstižně charakterizují vlastnosti jednotlivých typů odhadů (viz. [3])).



Obrázek 2: Odhad ROC křivky pro proměnnou CCW1P různými technikami (viz kapitola 3)

Reference

1. DORFMAN, D. D., ALF, E. Maximum likelihood estimation for parameters of signal. *Math.~Psychol*. 1969, N. 6, s. 487-496.
2. HSIEH, F., TURNBULL, B. W. Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. *The Annals of Statistics.* 1996, vol. 24, no. 1, s. 25-40.
3. MICHÁLEK, J., SEDLAČÍK, M., DOUDOVÁ, L. *A Comparison of Two Parametric ROC Curves Estimators in Binormal Model*, 23 rd International Conference Mathematical Methods in Economics 2005, Hradec Králové, Czech Republic, 14. - 16.9. 2005.
4. SEDLAČÍK, M. ROC curve for Binormal Latent Variable. In *XXIV. mezinárodní kolokvium o řízení osvojovacího procesu : sborník abstraktů a elektronických verzí příspěvků na CD-ROMu* [CD-ROM]. Brno : UO, 2006.
5. SEDLAČÍK, M., MICHÁLEK, J. ROC křivky a jejich využití při konstrukci klasifikačních a regresních stromů. In *Zborník referátov, XVI. letná škola biometriky*. Pribylina (Slovakia): 2004. ISBN 801-89162-06-1.
6. SKORKOVSKA, S., MASKOVA, Z.,Koci J., MICHALEK, J., SEDLACIK, M.: Search for Optimal Combination of Structural and Functional Methods for the diagnosis and follow-up of glaucoma. *EVER - European Association for Vision and Eye Research.* Villamoura, Portugalsko: 2006.

Mgr. Marek Sedlačík, Ph.D.
Katedra ekonometrie, UO
Kounicova 65, Brno
E-mail: marek.sedlacik@unob.cz
Telefon: +420 973 443 591

Doc.RNDr. Jaroslav Michálek, CSc.
Katedra aplikované matematiky a informatiky ESF MU
Lipová 41a, Brno
E-mail: michalek@econ.muni.cz
Telefon: +420 549 496 075

Ako postupovať pri faktorovej analýze v SAS Enterprise Guide

Iveta Stankovičová

Abstract: The purpose of *common factor analysis* is to explain the correlations or covariances among a set of variables in terms of a limited number of unobservable, latent variables. The latent variables are not generally computable as linear combinations of the original variables. In common factor analysis, it is assumed that the variables are linearly related if not for uncorrelated random error or *unique variation* in each variable; both the linear relations and the amount of unique variation can be estimated.

Key words: common factor analysis, multivariate procedures, rotation, SAS Enterprise Guide

1. Úvod

Faktorová analýza (FA = Factor Analysis) je metóda vhodná na zjednodušenie štatistických analýz. Základným cieľom FA je posúdiť štruktúru vzťahov medzi sledovanými premennými a zistiť, či je ich možné rozdeliť do skupín, v ktorých by ich vzájomné korelácie boli významné a medzi týmito skupinami by zase neboli významné. Cieľom je vytvoriť nové premenné (tzv. *faktory*), ktoré sú v praxi priamo nemerateľné a umožňujú pochopiť analyzované dáta a prípadne je ich možné použiť v ďalších analýzach.

Faktorová analýza vznikla v psychológii (Ch. Spearman, 1904), kde sa aj dlho výhradne používala a aj dnes sa často používa. O jej ďalší rozvoj a zdokonalenie sa zaslúžili vedci z oblasti spoločenskovedných disciplín (L. L. Thurstone, R. B. Cattel, C. Burt, G. Thomson a iní) a štatistici (D. N. Lawley, M. S. Bartlett, C. R. Rao a ďalší). V posledných 40-tich rokoch prenikla aj do iných vedných odborov.

Faktorová analýza je však štatistikmi často kritizovaná pre jej nejednoznačné riešenie, pre subjektivitu v niektorých jej cieľoch a krokoch, hmlistú interpretáciu a približnosť výsledkov. Povaha FA je skôr heuristická a prieskumná (exploratívna) ako overovacia (konfirmatívna). Jej použitie vyžaduje rešpektovanie predpokladov a podmienok tejto metódy, skúsenosti s jej použitím a hlavne tiež znalosti predmetu aplikačnej oblasti. Mnohí jej prívrženci súčasne varujú pred rutinným použitím FA vo výskume a pred „klamaním“ pomocou FA.

Termín *spoločný faktor (common factor)* sa vo FA používa na označenie nepozorovateľnej, hypotetickej premennej, ktorá prispieva k vysvetleniu najmenej dvoch pôvodných premenných. *Špecifický faktor (unique factor)* je zase nepozorovateľná, hypotetická premenná, ktorá prispieva k vysvetleniu len jednej pôvodnej premennej. Model FA predpokladá len jeden špecifický faktor pre každú pôvodnú premennú.

Poznáme dva druhy faktorovej analýzy:

1. prieskumnú (exploratory) a
2. potvrdzujúcu (confirmatory) faktorovú analýzu.

Ak výskumník nemá alebo má len málo znalostí o faktorovej štruktúre premenných, tak použije na jej odhalenie prieskumnú FA. Na druhej strane, ak výskumník predpokladá, že faktorová štruktúra je známa, respektíve má nejakú hypotézu *a priori*, tak na jej overenie použije potvrdzujúcu FA. V tomto príspevku sa budeme ďalej zaoberať len prieskumnou FA.

2. Všeobecná schéma aplikovania faktorovej analýzy

Pri aplikovaní faktorovej analýzy sa odporúča dodržať nasledovné kroky:

1. krok: Výber premenných

V úvode je veľmi dôležitý výber premenných (indikátorov) do FA, ktorý sa robí na základe vecnej analýzy problému. Po výbere premenných (indikátorov) je treba zvážiť aj veľkosť vzorky. Pri počte *p* zvolených premenných do FA je potrebné mať dostatočný počet štatistických jednotiek *n* (pozorovaní). Neexistuje presné pravidlo pre určenie rozsahu vzorky. Môžeme použiť len niektoré z empirických pravidiel, napr. že rozsah vzorky *n* by mal spĺňať kritérium $n>10p$.

2. krok: Odhad korelačnej matice a posúdenie vhodnosti dát pre faktorovú analýzu

Po výbere premenných je potrebné vypočítať ich výberovú kovariančnú, resp. korelačnú maticu (čo je v praxi bežnejší prípad, lebo zvyčajne sú premenné vyjadrené v rôznych merných jednotkách).

Z hľadiska FA by mali byť vybrané premenné závislé, lebo len vtedy je možné predpokladať existenciu spoločných faktorov – latentných premenných a následne redukovať dimenziu údajov. K tomuto účelu nám poslúži analýza korelačnej matice.

Bartlettov test sféričnosti[1] testuje nulovú hypotézu, že korelačná matica sa rovná jednotkovej matici, čiže že všetky koeficienty korelácie mimo hlavnej diagonály sú rovné 0. Tento test vyžaduje, aby náhodný vektor indikátorov **X** mal viacrozmerné normálne rozdelenie. Dáta považujeme vhodné na FA, ak zamietneme nulovú hypotézu. Bartlettov test má viacero podôb. Tento test SAS EG neposkytuje.

Sú aj iné možnosti, ako posúdiť vhodnosť výberových dát na FA. Obľúbená je štatistika KMO (Kaiser – Meyer – Olkin), ktorá je založená na porovnaní koeficientov korelácie s parciálnymi koeficientmi korelácie:

$$KMO = \frac{\sum_{i \neq j}^{p} \sum r_{ij}^2}{\sum_{i \neq j}^{p} \sum r_{ij}^2 + \sum_{i \neq j}^{p} \sum r_{parc.,ij}^2} \quad , \qquad (1)$$

kde r_{ij} je párový koeficient korelácie medzi X_i a X_j a $r_{parc.,ij}$ koeficient korelácie. Veľké rozdiely medzi odpovedajúcimi si párovými a parciálnymi koeficientmi korelácie signalizujú silné závislosti v skupine premenných.

KMO štatistika sa počíta ako celková miera adekvátnosti (vhodnosti) výberových dát pre FA (*MSA - Kaiser's measure of sampling adequacy overall*) a aj ako čiastková miera adekvátnosti pre jednotlivé indikátory. Je to miera homogenity premenných. Hodnoty KMO miery sa netestujú, ale používa sa nasledovná tabuľka (*Tabuľka 1*) odporučení podľa Kaisera a Ricea (1974)[2]:

Ako môžeme vidieť z tabuľky 1, odporúčajú sa vyššie hodnoty KMO miery pre vzorku dát ako 0,5. Dobré je, keď KMO miera je vyššia ako 0,8. Miera okolo 0,6 je ešte v tolerancii. Ak je KMO miera nižšia ako 0,5, tak môžeme zvýšiť jej hodnotu, ak nepoužijeme vo FA takú premennú, ktorá má nízku individuálnu hodnotu KMO miery. Hodnoty KMO miery SAS EG vo výstupe procedúry FACTOR poskytuje.

---

[1] Testovacie kritérium pre Bartlettov test je uvedené napr. v knihe Hebák P. a kolektív: Vícerozmerné statistické metody (2). Informatorium. Praha 2005. 9. kapitola.

[2] Tabuľka odporučení je prebratá z knihy Subhash Sharma: Applied Multivariate Techniques. New York, John Wiley & Sons, Inc., 1996.

Tabuľka 1: Hodnoty KMO miery pre adekvátnosť výberových dát na FA

Hodnota KMO štatistiky	Odporúčanie pre adekvátnosť výberových dát na FA
≥ 0,9	vynikajúce
<0,8; 0,9)	chvályhodné
<0,7; 0,8)	stredne užitočné
<0,6; 0,7)	priemerné
<0,5; 0,6)	slabé
<0,5	nedostatočné

3. krok: Odhad parametrov faktorového modelu

Po tom čo sme posúdili vhodnosť dát pre FA, treba sa rozhodnúť, ktorú z metód odhadovania parametrov (faktorových váh) FA modelu zvolíme. Na odhad parametrov FA modelu bolo vyvinutých mnoho postupov, resp. metód *extrakcie* faktorov. Veľká časť vyvinutých metód má len historický význam a dnes sa už nepoužíva. V systéme SAS EG procedúra FACTOR ponúka 7 metód na odhad parametrov faktorového modelu:

1. metóda hlavných komponentov (*PCA – Principal component analysis*),
2. iteračná metóda hlavných faktorov, ktorá predstavuje modifikáciu metódy PCA a označuje sa skratkou PAF (*Iterated principal factor analysis*),
3. metóda maximálnej vierohodnosti (*ML – Maximum-likelihood factor analysis*),
4. Harrisova komponentná analýza (*Harris component analysis*),
5. *Image covariance matrix*,
6. *Alpha factor analysis*,
7. *Unweighted least squares factor analysis*.

Nedá sa jednoznačne povedať, ktorá metóda je najlepšia. Medzi najobľúbenejšie patrí metóda hlavných faktorov PAF. Metóda maximálnej vierohodnosti má oproti ostatným metódam výhody v tom, že je nezávislá na použitých merných jednotkách a poskytuje určité kritéria pre odhad vhodného počtu spoločných faktorov.

Existuje celý rad objektívnych a subjektívnych rád a možností ako aspoň približne určiť počet spoločných faktorov *q*:

1. Začať FA metódou PCA a určiť počiatočný počet faktorov, ktorý je len orientačný, ale niekde treba začať.
2. Odhadom počtu spoločných faktorov môže byť počet vlastných čísiel redukovanej korelačnej matice väčších ako jedna.
3. Niekedy apriórne vieme z iných analýz alebo z teoretickej analýzy problematiky, koľko spoločných faktorov je potrebných na charakterizovanie vzťahov medzi indikátormi.
4. Spoločné faktory by mali vysvetliť čo najviac celkového rozptylu. V exaktných vedách by to malo byť 90-95% a v spoločenských vedách viac ako 60-70%.
5. Spoločné faktory by mali reprezentovať viac ako 90% celkovej komunality, ktorá je daná súčtom komunalít všetkých *p* vstupných premenných (indikátorov).
6. Môžeme použiť graf „*scree plot*“, ktorý zobrazuje počet faktorov na osi x a na osi y percento vysvetlenej variability, t.j. hodnoty vlastných čísiel redukovanej kovariančnej, resp. korelačnej matice. Za optimálny počet faktorov je treba považovať hodnotu na x-ovej osi pred bodom zlomu na krivke vlastných čísiel.
7. Do konečného riešenia sa nemajú zahrňovať tzv. *triviálne* faktory. Triviálne faktory sú také, ktoré významne korelujú len s jedným indikátorom. Lepšie je takýto indikátor z FA vylúčiť a začať znovu. To neznamená, že daná premenná je nepodstatná, ale nehodí sa do faktorovej analýzy a môže byť uvažovaná samostatne.

Metóda maximálnej vierohodnosti poskytuje *objektívnejšie* kritériá pre hodnotenie počtu faktorov. Jednou z možností je test počtu faktorov vierohodnostným pomerom, ktorý

vyžaduje predpoklad viacrozmerného normálneho rozdelenia indikátorov a dostatočne veľký rozsah vzorky *n*.

4. krok: Rotácia a interpretácia

V treťom kroku FA získame len počiatočný odhad matice faktorových váh. Odporúča sa použiť viacero techník *rotácie* a výsledky porovnať a posúdiť z hľadiska interpretovateľnosti faktorov. Treba zistiť, či výsledok rotácie je logický, či podporuje teóriu alebo empirické zistenia z danej vecnej problematiky.

Ortogonálna transformácia faktorov (tzv. rotácia faktorov) je výpočtová operácia, ktorou sa z matice faktorových váh získava nová matica. Pojem rotácie sa do FA preniesol z geometrického zobrazenia transformácie faktorových váh. Ortogonálna transformácia je geometricky pevná (rigidná) rotácia *q* súradnicových osí v *p*-rozmernom priestore.

Teória FA a štatistické programy poskytujú celý rad metód transformácie (rotácie) faktorov. Je potrebné sa rozhodnúť, či použijeme *ortogonálnu* (pravouhlú, kolmú) rotáciu alebo *kosouhlú* (šikmú) rotáciu. Ortogonálna rotácia vedie k riešeniu s nekorelovanými (nezávislými) faktormi. Prvky matice faktorových váh je možné interpretovať ako regresné koeficienty závislosti indikátorov na faktoroch a tiež ako korelačné koeficienty medzi nimi. Kosouhlá rotácia vedie ku získaniu závislých faktorov, čiže naviac ešte poskytuje korelačnú maticu medzi faktormi. Kosouhlú rotáciu niektorí autori odmietajú, kým iní autori ju vítajú. Tvrdia, že pre prax sú reálnejšie korelované (závislé) spoločné faktory.

5. krok: Odhad faktorových skóre a aplikácia výsledkov faktorovej analýzy

Ak je to potrebné, tak odhadneme faktorové skóre, tzn. hodnoty spoločných faktorov pre jednotlivé štatistické jednotky. Hodnoty FA skóre sa dajú ďalej použiť v iných štatistických analýzach, napr. v regresných modeloch ako menší počet nezávislých premenných oproti pôvodnému počtu, v zhlukovej analýze ako vstupné premenné, ktoré sú nekorelované (len po použití ortogonálnej rotácie), v diskriminačnej analýze a pod.

3. Príklad

Postupnosť krokov pri faktorovej analýze budeme demonštrovať na demografických údajoch za 79 okresov Slovenska. K dispozícii máme nasledované priemerné demografické ukazovatele za obdobie rokov 2001 až 2005:

Tabuľka 2: Zoznam vstupných premenných

Premenná	**Popis**
cislo okresu	číslo okresu
okres	názov okresu
us_m	úhrnná sobášnosť muži
sr_m	štandardizovaná rozvodovosť muži
us	úhrnná sobášnosť ženy
sr	štandardizovaná rozvodovosť ženy
up	úhrnná plodnosť
prvs_m	priemerný vek pri prvom sobáši muži
prvr_m	priemerný vek rozvode muži
prvs	priemerný vek pri prvom sobáši
prvr	priemerný vek rozvode
prvp	priemerný vek pri prvom pôrode

Na základe analýzy korelačnej matice pre 10 indikátorov a ich KMO miery (Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.76341472) sme usúdili, že dáta sú vhodné na faktorovú analýzu. Na odhad parametrov faktorového modelu sme najskôr použili metódu hlavných komponentov a získali sme 2 faktory, ktoré môžeme pokladať za štatisticky

významné (použili sme Kaiserovo pravidlo o vlastných číslach väčších ako 1, viď *Tabuľka 3*). Tieto dva odhadnuté latentné faktory vysvetľujú až 83,12% celkovej variability dát. Na zlepšenie ich interpretácie sme použili ortogonálnu transformáciu (rotáciu) varimax. Po rotácii sme získali až 4 významné faktory, ktoré vysvetľujú až 92,26% variability (*Tabuľka 4*). Môžeme konštatovať, že prvý faktor je pozitívne skorelovaný so sobášnosťou mužov a žien, plodnosťou a záporne s rozvodovosťou mužov a žien. Druhý faktor je vysoko pozitívne skorelovaný hlavne s priemerným vekom rozvodov mužov a žien. Tretí faktor je korelovaný s ostatnými priemernými vekmi a štvrtý faktor len s rozvodovosťou mužov a žien (*Tabuľka 5*).

Tabuľka 3: Vlastné čísla (časť, PCA metóda)

Eigenvalues of the Correlation Matrix: Total = 10 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	**6.330**	**4.349**	**0.633**	**0.633**
2	**1.982**	**1.232**	**0.198**	**0.831**
3	0.749	0.280	0.075	0.906
4	0.470	0.189	0.047	0.953

Tabuľka 4: Vysvetlená variabilita faktormi po rotácii

Variance Explained by Each Factor			
Factor1	**Factor2**	**Factor3**	**Factor4**
3.0887	2.8278	2.1083	1.2006

Tabuľka 5: Faktorové váhy (po rotácii varimax)

Rotated Factor Pattern				
	Factor1	**Factor2**	**Factor3**	**Factor4**
us_m	**0.990**	-0.079	-0.066	-0.089
us	**0.874**	-0.254	-0.134	-0.178
sr_m	**-0.586**	0.329	0.155	**0.714**
sr	**-0.610**	0.251	0.093	**0.735**
up	**0.658**	-0.215	-0.265	-0.192
prvs_m	-0.275	**0.472**	**0.751**	0.166
prvs	-0.177	**0.547**	**0.786**	0.143
prvr_m	-0.168	**0.949**	0.195	0.125
prvr	-0.141	**0.953**	0.221	0.092
prvp	-0.203	**0.456**	**0.846**	0.051

4. Záver

FA je podobná metóde analýzy hlavných komponentov (PCA), pretože tiež je určená na vytváranie nových premenných a na zníženie dimenzie dát s čo najmenšou stratou informácie. Do určitej miery je možné FA považovať za rozšírenie PCA. Na rozdiel od PCA však vychádza zo snahy vysvetliť závislosti medzi pôvodnými premennými. Medzi nedostatky PCA patrí, že je závislá na merných jednotkách premenných, neposkytuje jednoznačné kritérium pre rozhodnutie či zvolený počet HK vysvetľuje dostatočné percento celkovej variability a nezaoberá sa chybovým rozptylom premenných. Prístup FA čiastočne odstraňuje tieto nedostatky PCA, ale má však iné slabé miesta. FA má veľa subjektívnych aspektov a nejednoznačnosť odhadov faktorových parametrov. Prednosťou FA je jej väčšia obecnosť a úspornosť, i keď niektoré odhady vyžadujú splnenie aspoň približného viacrozmerného normálneho rozdelenia. Dôležité je aj určenie počtu spoločných faktorov pred vykonaním FA, ktoré musí vychádzať z hypotéz výskumníka v predmetnej aplikačnej oblasti.

5. Literatúra

HEBÁK, PETR A KOLEKTÍV: Vícerozmerné statistické metody (2). Informatorium. Praha 2005.

STANKOVIČOVÁ, IVETA: Viacrozmerná analýza rentability poisťovní SR pomocou Enterprise Guide, In: 10. medzinárodný seminár Výpočtová štatistika. Bratislava: SŠDS 2001. ISBN 80-88946-14-X

SHARMA, SUBHASH: Applied Multivariate Techniques. New York, John Wiley & Sons, Inc., 1996.

VOJTKOVÁ, MÁRIA: Viackriteriálne hodnotenie podnikov priemyslu Slovenskej republiky, Ekonomické rozhľady. Roč. 2003, č. 3, s. 320-331, ISSN 0323-262X

Zdroj dát: Výskumné demografické centrum, INFOSTAT Bratislava

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

Fuzzy *c* zhluková analýza pomocou softvéru Mathematica

Fuzzy *c* cluster analysis using softvare Mathematica

Beáta Stehlíková

Abstrakt. *Cieľom príspevku je prezentovať použitie softvéru Mathematica pre výpočet fuzzy zhlukovej analýzy.*
Kľúčové slová: fuzzy *c* zhluková analýza, Mathematica

Abstract. *The aim of the paper is to presentat calculation of the fuzzy cluster analysis using software Mathematica.*
Key words: fuzzy cluster analysis, Mathematica

Úvod

Pojem podobnosti je jedným z najdôležitejších pojmov teórie fuzzy množín. Dva objekty sú rovnaké, ak sú vzájomne zameniteľné. Vzájomná zameniteľnosť znamená, že každý z týchto objektov obsahuje úplnú informáciu o druhom objekte, ktorá je v danej situácii dôležitá. Vzájomná zameniteľnosť objektov je zhoda znakov, kritérií dôležitých v danej situácii. Podobnosť objektov znamená ich čiastočnú vzájomnú zameniteľnosť, t.j. možnosť vzájomnej zámeny objektov s určitou stratou informácie, ktorej veľkosť je v danej situácii prípustná. Každý prvok z množiny podobných objektov nesie v sebe určitú informáciu o jemu podobných objektoch. Nie je to však úplná informácia, ako v prípade rovnakých prvkov. Na druhej strane však nie je tu iba jediná možnosť: úplná informácia alebo žiadna informácia. Sú tu rôzne úrovne informácie, ktorú jeden objekt obsahuje o inom podobnom objekte. Ak je pre množinu objektov zadaná len podobnosť, potom túto množinu nevieme rozložiť do presne určených tried tak, aby objekty vnútri jednej triedy boli podobné, ale aby objekty rôznych tried podobné neboli. Vieme dosiahnuť len to, aby objekty vnútri jednej triedy boli viac podobné, ako je ich podobnosť s objektmi iných tried. Pojem vzájomnej podobnosti objektov a kvantitatívne vyjadrenie tejto podobnosti je kľúčovým problémom zhlukovej analýzy (Lukasová, Šarmanová, 1985). Popri metódach klasickej zhlukovej analýzy existuje fuzzy zhluková analýza (Klir, Yuan, 1995).

Materiál a metódy

Nech $Z = \{z_1, z_2, ..., z_n\}$ je n objektov, z ktorých každý je charakterizovaný p znakmi, t.j. $z_i = (z_{i1}, z_{i2}, ..., z_{ip})$, $i = 1, 2, ..., n$. Úlohou fuzzy zhlukovania je nájsť fuzzy c rozklad $\wp$ a stredy rozkladov, ktoré by čo najlepšie charakterizovali štruktúru údajov.

Fuzzy c rozklad objektov je systém fuzzy podmnožín $\wp = \{P_1, P_2, ..., P_c\}$, pre ktorý platí

$$\sum_{i=1}^{c} P_i(z_k) = 1 \text{ pre všetky } x_k,\ k = 1, 2, ..., n,$$

kde $P_i(z_k)$ je hodnota funkcie príslušnosti P_i (i =1, 2, ..., c) objektu z_k (k = 1, 2, ..., n). S fuzzy c rozkladom $\wp$ korešponduje c stredov $t_1, t_2, ..., t_c$ zhlukov

$$t_i = \frac{\sum_{k=1}^{n} [P_i(z_k)]^m x_k}{\sum_{k=1}^{n} [P_i(z_k)]^m},\ i = 1, 2, ..., c,$$

kde $m > 1$ je reálne číslo. Stred t_i ($i = 1, 2, ..., c$) je vážený priemer údajov v P_i. Váha objektu charakterizovaného z_k je jeho stupeň príslušnosti do podmnožiny P_i. Pre m blízke 1 fuzzy c priemery konvergujú ku klasickej zhlukovej analýze c priemerov. Rozklad je tým viac fuzzy, čím je hodnota m väčšia. Pre $m \to \infty$ stredy zhlukov t sa stáčajú v smere centroidu údajov.

Softvér Mathematica balík Fuzzy Logic obsahuje procedúru na výpočet fuzzy *c* zhlukovej analýzy. Pomocou príkazu

```
FCMCluster[Udaje, InitializeU[Udaje, počiatočný počet zhlukov],
          koeficient fuzzyčnosti, epsilon]
```

kde *Udaje* je názov súboru s údajmi, *počiatočný počet zhlukov* je celočíselná hodnota udávajúca predpokladaný počet zhlukov, ktorý je z intervalu $\langle 2, n \rangle$, kde *n* je rozsah súboru, *koeficient fuzzyčnosti* je hodnota obvykle z intervalu $\langle 1{,}5; 2{,}5 \rangle$ udávajúca hodnotu, nakoľko je zhlukovanie fuzzy. V literatúre sa doporučuje hodnota 2. Hodnota *epsilon* určuje iteračnú presnosť. Príkaz

InitializeU[Udaje, počiatočný počet zhlukov]

udáva počiatočné delenie údajov do zhlukov. Príkaz

ShowCenters

umožňuje znázorniť stredy zhlukov v prípade dvoch zhlukov. Pomocou príkazu

ShowCenterProgression

znázorníme iteračný proces výpočtu stredov zhlukov v prípade dvoch zhlukov. V prípade troch zhlukov môžeme využiť 3D grafiku softvéru Mathematica. V prípade viac ako troch zhlukov sa musíme uspokojiť s vyjadrením výsledkov v numerickej forme.

Výsledky a diskusia

Použitie softvéru si ukážeme na príklade. Na základe dvoch znakov Znak1 a Znak2 máme rozdeliť objekty, na ktorých boli znaky merané pomocou fuzzy c zhlukovej analýzy do dvoch zhlukov s iteračnou presnosťou 0,01. Namerané hodnoty sú nasledovné.

```
Udaje2 = {{1, 2}, {1, 2.5}, {2, 2}, {2, 2.5}, {7, 8}, {7.2, 8.1}, {7.9, 8.3}}
```

Zhlukovanie vykonáme pomocou príkazu

```
Vysledok = FCMCluster[Udaje2, InitializeU[Udaje2, 2], 2, 0.01]
```

Výstup je nasledovný.

```
0.474587
0.303759
0.123156
0.00159406
```

Hodnoty epsilon v jednotlivých iteráciách

```
{{{1.50079, 2.25017}, {7.36488, 8.13267}},
```

Stredy zhlukov

```
 {{0.996005, 0.995683, 0.995326, 0.994876, 0.00237557, 0.000423374, 0.00403722},
  {0.00399541, 0.00431697, 0.00467439, 0.00512364, 0.997624, 0.999577, 0.995963}},
 {{{3.24436, 3.91813}, {4.0092, 4.77012}}, {{2.6923, 3.42443}, {5.78026, 6.56078}},
  {{1.57715, 2.31056}, {7.2907, 8.0702}}, {{1.50079, 2.25017}, {7.36488, 8.13267}}}}
```

Hodnoty funkcie prislušnosti objektov do zhlukov **Súradnice stredov v jednotlivých iteráciách**

Je trochu neprehľadný, bez označenia obsahu jednotlivých častí. Znalosť členenia výstupu však pomôže v kombinácii s ďalšími príkazmi softvéru Mathematica dosiahnúť zrozumiteľnejšiu formu výstupu. Pomocou príkazu

```
{{7.365075576466693`, 8.132739375646299`},
 {1.5004547857579358`, 2.250090872468123`}} // MatrixForm
```

dostaneme výstup v tvare matice

$$\begin{pmatrix} 7.36508 & 8.13274 \\ 1.50045 & 2.25009 \end{pmatrix}$$

t.j. stredy zhlukov sú

$$\left(7.36508, 8.13274\right) \text{ a } \left(1.50045, 2.25009\right)$$

Pomocou príkazov

```
<< LinearAlgebra`MatrixManipulation`
Transpose[
  {{0.00399043467316385`, 0.00431273121820389 8`, 0.00467855918916089 9`,
    0.00512957182064868 1`, 0.997622026250113 2`, 0.999575605104600 9`,
    0.995966078503707 9`},
   {0.996009565326836 2`, 0.995687268781796 1`, 0.995321440810839`,
    0.994870428179351 5`, 0.00237797374988678 9`,
    0.00042439489539904346`, 0.00403392149629200 86`}}] //
 MatrixForm
```

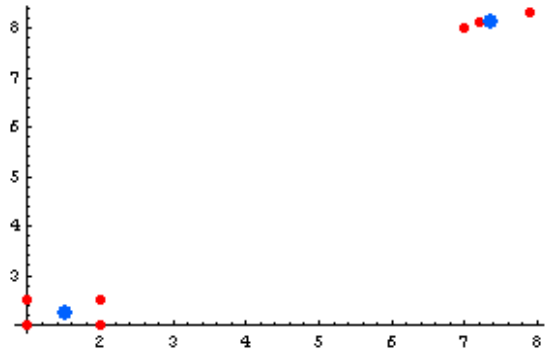
dostaneme výstup, ktorý obsahuje prehľadne hodnoty funkcií príslušnosti objektov do jednotlivých zhlukov

$$\begin{pmatrix} 0.00399043 & 0.99601 \\ 0.00431273 & 0.995687 \\ 0.00467856 & 0.995321 \\ 0.00512957 & 0.99487 \\ 0.997622 & 0.00237797 \\ 0.999576 & 0.000424395 \\ 0.995966 & 0.00403392 \end{pmatrix}$$

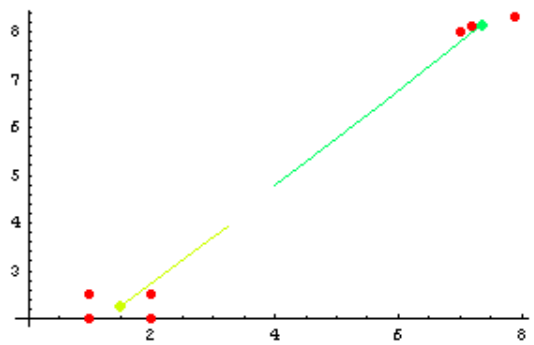
Hodnota funkcie príslušnosti prvého objektu do prvého zhluku je 0,00399043 a do druhého zhluku 0,99601. Znamená to, že prvý objekt patrí do druhého zhluku. Rovnako do neho patrí druhý, tretí a štvrtý objekt. Do prvého zhluku patria zvyšné tri objekty. V prípade, že funkcia príslušnosti do žiadneho zhluku nie je vyššia ako 0,80, hovoríme, že objekt je nevyhranený, ale inklinuje do zhluku, pre ktorý je funkcia príslušnosti maximálna. Rovnakým spôsobom môžeme vypísať súradnice stredov zhlukov v jednotlivých iteráciách.

Aplikácia nasledovných príkazov nám v prípade dvoch zhlukov umožní grafické znázornenie stredov zhlukov ako aj iteračný proces ich výpočtu.

```
body = ListPlot[Udaje2, Axes → True, PlotStyle → {PointSize[0.02], Hue[0]}]
Graf2 = ShowCenters[body, Vysledok]
ShowCenterProgression[ body , Vysledok]
```



Obrázok 1 Objekty a stredy zhlukov



Obrázok 2 Iterácia určenia stredov zhlukov

V mnohých prípadoch sú výsledky fuzzy zhlukovej analýzy vzhľadom na existenciu nevyhranených objektov bližšie k realite. Podmienkou širšieho využívania fuzzy zhlukovej analýzy nematematikmi je dostupnosť vhodného softvéru, akým Mathematica je. Prezentovaný príspevok pomôže správne čítať získané výsledky a využiť ich v plnej miere.

Literatúra

KWANG H. LEE: *First Course On Fuzzy Theory and Applications.* Berlin: Springer-Verlag, 2005, ISBN: 3540229884

KLIR, G.J., YUAN, B.: *Fuzzy Sets and Fuzzy Logic; Theory and Applications.* Prentice Hall, Upper Saddle River, New York, 1995

LUKASOVÁ, A., ŠARMANOVÁ, J.: *Metody shlukové analýzy*. Praha: SNTL, 1985

www.wolfram.com

Príspevok je súčasťou riešenia projektu KEGA 3/2060/04.

Kontaktná adresa

doc. RNDr. Beáta Stehlíková, CSc.
Katedra štatistiky aoperačného výskumu, Fakulta ekonomiky a manažmentu, Slovenská poľnohospodárska univerzita v Nitre, Tr. Andreja Hlinku 2, 949 76 Nitra
e-mail: Beata.Stehlikova@uniag.sk

Simulácia náhodného vývoja ceny akcie pomocou programov SPSS a MS® Excel[1]

Vladimír Úradníček

Abstract:
Any variable whose value changes over time in an uncertain way is said to follow a stochastic process. Particular type of stochastic process is a Markov process. Stock prices are usually assumed to follow a Markov process. In the contribution, we have developed a plausible Markov stochastic process for the behavior of a stock price over time. The process is widely used in the valuation of derivatives. It is know as geometric Brownian motion. Under this process, the proportional rate of return to the holder of the stock in any small interval of time is normally distributed and the returns in any two different small intervals of time are independent. This contribution is comparing different scenarios of MSFT share price forecast for next one year. Every scenario was formed as a Monte Carlo simulation – one triple of the simulations was created employing SPSS software and the second one employing MS®Excel.

Key words:
Stochastic process, Markov process, Brownian motion, Wiener process, Itô process, stock prices, geometric Brownian motion, Monte Carlo simulation

1. Úvod

Pri skúmaní vývoja hodnoty opčných derivátov je dôležité skúmať a modelovať aj očakávaný vývoj ceny podkladového finančného aktíva. Pre finančné podkladové aktíva je charakteristický náhodný vývoj v čase, pričom tento priebeh býva označovaný ako stochastický proces. Cieľom predkladaného príspevku je ilustrovať na príklade akcie firmy Microsoft Corp.(MSFT) simuláciu náhodného vývoja ceny akcie na obdobie jedného roku. Simulácia bola uskutočnená metódou Monte Carlo pre tri rôzne scenáre variantne so softvérovými produktami SPSS a MS®Excel.

2. Teoretické východiská

Existuje niekoľko formulácií hypotézy o náhodnom pohybe ceny podkladových aktív, avšak všetky majú spoločné nasledovné dva predpoklady:

- súčasná cena akcie plne a výlučne odzrkadľuje všetky minulé udalosti, t.j. nezahŕňa už žiadne ďalšie informácie;
- trhy odpovedajú okamžite na každú ďalšiu informáciu o podkladovom aktíve.

Z uvedeného vyplýva, že modelovanie cien podkladových aktív je v skutočnosti modelovaním vplyvov nových informácií, ovplyvňujúcich ich cenu. Za platnosti dvoch horeuvedených predpokladov tvoria nepredpovedané zmeny v cene podkladového aktíva tzv. Markovov proces. *Markovov proces* je taký stochastický proces[2] , pre ktorý platí, že ak je daná hodnota *X(s)*, tak budúce hodnoty *X(t)* pre $t > s$ môžu závisieť iba na *X(s)*, nie však na predošlých hodnotách *X(u)* pre $u < s$. Požiadavka Markovovského charakteru stochastického

[1] Tento príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/2594/05 „Analýza vybraných otázok finančného a bankového trhu po vstupe SR do EÚ“.

[2] *Stochastický proces* je *t*-parametrický systém náhodných premenných $\{X(t), t \in I\}$, kde I je interval alebo diskrétna množina indexov.

vývoja cien akcií je v súlade *so slabou formou trhovej efektívnosti*, nakoľko jedine súčasné hodnoty cien akcií by mali slúžiť na vytváranie budúcich hodnôt.

Takýto proces môžeme popísať diskrétne s aplikáciami pri simuláciách alebo spojite s využitím najmä pri analytickom riešení.

V roku 1827 objavil škótsky botanik Brown[3] pri skúmaní peľových zrniek pod svetelným mikroskopom, že malinké častice peľových zrniek sa pohybujú v rastlinnej šťave zdanlivo plne nekontrolovateľným pohybom vo všetkých smeroch. Neskôr bol tento pohyb označený ako Brownov pohyb. Jeho fyzikálnu podstatu vysvetlili v rokoch 1905 – 1906 Albert Einstein a Marian von Smoluchowski. Exaktnú matematickú formuláciu poskytol až v roku 1923 Wiener.[4] Nezávisle od uvedených fyzikálnych interpretácií rozvinul už v roku 1900 Bachelier v svojej práci *Théorie de la spéculation* viaceré aspekty Wienerovho procesu vo vzťahu k finančným trhom.

Na matematický popis pohybu peľových zrniek možno využiť stochastický proces $W_t(\omega)$, ktorý možno chápať ako polohu peľového zrnka ω v čase t. Pre jednoduchosť sa obmedzíme na jednorozmerný Brownov pohyb, tj. $W_t(\omega)$ bude nadobúdať hodnoty len v **R**. Celý postup však možeme ľahko zovšeobecniť na n-rozmerný Brownov pohyb.
Brownov pohyb je stochastický proces s nasledujúcimi vlastnosťami:

(i) *s pravdepodobnosťou 1 sú trajektórie $W_t(\omega)$ spojité a platí $W_0 = 0$;*
(ii) *náhodná premenná W_t má normálne rozdelenie N(0,t)*;
(iii) *$W_{t+s} - W_s$ má N(0,t) rozdelenie. Ďalej platí, že W_t má nezávislé prírastky, t.j. W_{t1}, $W_{t2} - W_{t1}$, ..., $W_{tk} - W_{tk-1}$ sú nezávislé pre všetky $0 \le t_1 < t_2 < \ldots < t_k$.*

Brownov pohyb (proces) s parametrami: strednou hodnotou $\mu = 0$ a disperziou $\sigma^2 = 1$ nazývame Wienerov proces.

Z predošlého všeobecného definovania Brownovho procesu vyplýva, že ak $\{w_t(\omega), t \ge 0\} = \{w(t), t \ge 0\}$ je Wienerov proces, tak pre jeho štatistické parametre strednej hodnoty a disperzie platí:

$$E[w(t)] = 0 \quad \text{a} \quad D[w(t)] = t\,. \tag{1}$$

Okrem toho sa dá ukázať, že pre distribučnú funkciu rozdelenia pravdepodobnosti Wienerovho procesu platí:

$$P(w(t) < x) = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{x} e^{-\xi^2/2t} d\xi\,. \tag{2}$$

Na základe definície Wienerovho procesu sú jeho prírastky $dw(t)$ navzájom nekorelované a pre ich základné charakteristiky platí:

$$E[dw(t)] = 0 \quad \text{a} \quad D[dw(t)] = dt\,. \tag{3}$$

Prírastky dw môžeme teda zapísať v tvare

$$dw = \tilde{Z}\sqrt{dt}\,, \tag{4}$$

kde $\tilde{Z}$ je náhodná premenná s *N(0,1) rozdelením*.

---

[3] Robert Brown sa narodil 21.12.1773 v Montrose (Škótsko) a zomrel 10. 6. 1858 v Londýne.
[4] Norbert Wiener sa narodil v roku 1894 ako syn excentrického harvardského docenta pre slovanské jazyky Leo Wienera v Columbii (Missouri). Norbert Wiener je označovaný za jedného zo zakladateľov nového vedného odboru -kybernetiky. Zomrel v roku 1964.

Ak uvažujeme vývoj ceny v čase za niekoľko intervalov, potom

$$\sum_{t=1}^{T} \tilde{Z}_t \cdot \sqrt{dt} = \tilde{Z}_T - Z_0 . \qquad (5)$$

Pre základné charakteristiky potom platí:

$$E\left[\tilde{z}_t\right] = 0 \quad \text{a} \quad D\left[\tilde{z}_t\right] = T . \qquad (6)$$

Jedným zo všeobecných typov stochastických procesov, ktorý zahŕňa ako zvláštne prípady Wienerove a Brownove procesy, je Itôov proces, ktorý je pre premennú x definovaný nasledovne:

$$dx = a(x;t) \cdot dt + b(x;t) \cdot dw \quad , \qquad (7)$$

kde *a(.)* je prírastok a *b(.)* je smerodajná odchýlka zmeny premennej.

Veľké uplatnenie vo finančnom modelovaní má tzv. *Brownov geometrický proces*, u ktorého sa cena vyvíja exponenciálnym trendom. Je určený ako:

$$dx = \mu \cdot x \cdot dt + \sigma \cdot x \cdot dw \quad . \qquad (8)$$

Vzťah (8) môžeme zapísať aj ako

$$\frac{dx}{x} = \mu \cdot dt + \sigma \cdot dw \quad . \qquad (9)$$

μ udáva priemerný výnos (spravidla za obdobie jedného roku) a σ = smerodajná odchýlka za rok.

Pri analytickom oceňovaní opcií sa využíva geometrický Brownov proces s logaritmickými cenami. Potom

$$x_t = x_{t-1} \cdot \exp(\mu dt + \sigma dw) \quad . \qquad (10)$$

$$E\left[x_T\right] = x \cdot \exp(\mu T), \qquad D\left[x_T\right] = x^2 \cdot \exp(2\mu T)\left(\exp\left[\sigma^2 T\right] - 1\right) .$$

Hodnota kvantilu logaritmicko-normálneho rozdelenia na hladine pravdepodobnosti α sa môže vypočítať zo vzťahu:

$$x_T^{\alpha} = x \cdot \exp(T + \Phi^{-1}(\alpha) \cdot \sigma \cdot \sqrt{T}) \quad , \qquad (11)$$

kde Φ^{-1} je kvantilová funkcia zodpovedajúca distribučnej funkcii Φ, čo je distribučná funkcia *N(0,1) rozdelenia* a hodnoty $\Phi^{-1}(\alpha)$ sú kvantily *N(0,1) rozdelenia*.

Náhodný vývoj ceny akcie S_t , $t = 1, 2, \ldots, T$ potom môžeme modelovať ako

$$S_t = S_{t-1} \cdot \exp(\mu + \Delta\sigma \cdot dw) = S_{t-1} \cdot \exp(\mu + \Delta\sigma \cdot z \cdot \tilde{\Delta}t) \quad . \qquad (12)$$

Stredná hodnota ceny akcie

$$E\left[S_T\right] = S_0 \cdot \exp(\mu T) \cdot \qquad (13)$$

a

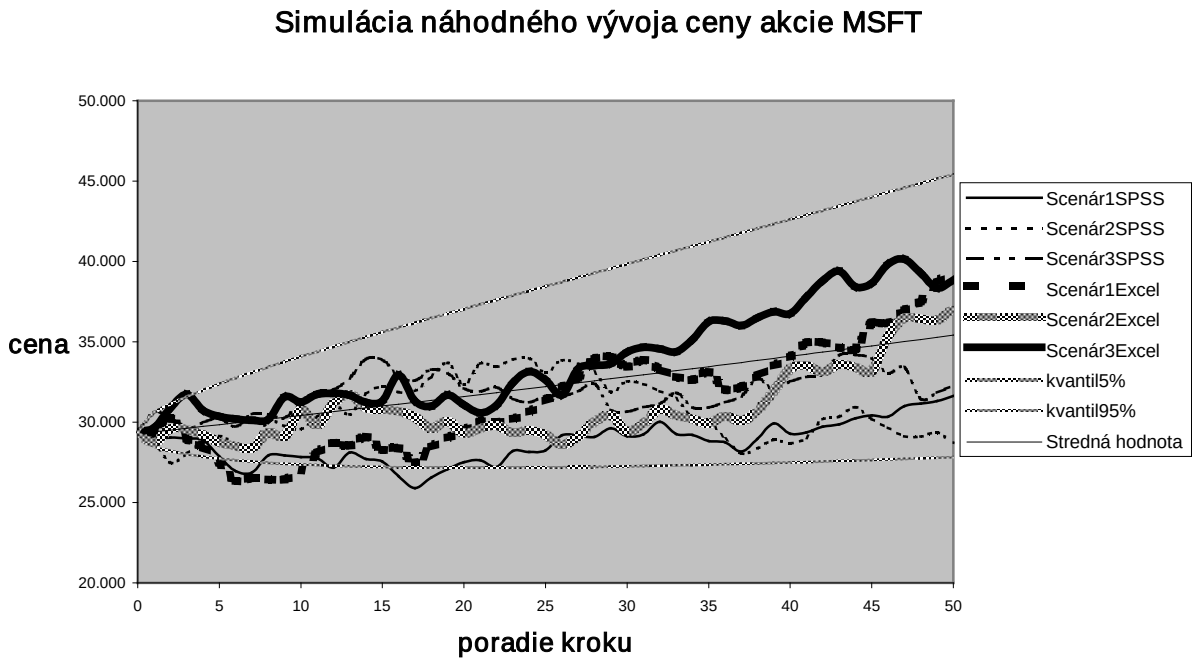
α . *100 %-ný* kvantil logaritmicko-normálneho rozdelenia pravdepodobnosti ceny akcie

$$S_T^{\alpha} = S_0 \cdot \exp(\mu \cdot n + \Phi^{-1}(\alpha) \cdot \sigma \cdot \sqrt{n}) \cdot \quad . \qquad (14)$$

3. Analýza

Pomocou softvérových produktov SPSS a MS® Excel sme variante modelovali vždy tri rôzne scenáre vývoja ceny akcie Microsoft Corp. (MSFT). Simulácia pre každý scenár bola na báze geometrického Brownovho procesu s logaritmickými cenami pre 50 krokov, pričom dĺžka jedného kroku bola $\frac{1}{50} = 0{,}02$, keďže cieľom bola simulácia pre obdobie $T = 1$ rok. Vstupné dáta boli získané z webovských stránok http://finance.yahoo.com, resp. http://finance.google.com a http://bonds.yahoo.com . Aktuálna spotová cena akcie MSFT 17. novembra 2006 bola vo výške 29,40 USD. Rizikovosť akcie bola odhadnutá na základe implikovanej volatility – 14,9 %. Stredná hodnota spojitého výnosu ceny akcie za jeden rok bola odhadovaná na 19 %.

Najprv boli pomocou generátorov náhodných čísel – *Random Numbers Rv.Normal* v prípade SPSS (v časti *Transform → Compute*), resp. *Generátora pseudonáhodných čísel v Exceli* (v časti *Nástroje→ Analýza dát*) vygenerované po tri náhodné scenáre. Následne boli pomocou vzťahou (12), (13), resp. (14) vypočítané hodnoty rôznych scenárov vývoja cien akcie MSFT, jej strednej hodnoty a hranice 5 % a 95 % -ného kvantilu lognormálneho rozdelenia pravdepodobnosti ceny akcie pre nasledujúce jednoročné obdobie. Výsledky sú znázornené na *Obrázku 1*.



Obrázok 1: Simulácia náhodného vývoja ceny akcie MSFT
Prameň: Vlastné spracovanie

4. Záver

Z *obrázku 1* vyplýva, že na konci nasledujúceho jednoročného obdobia môžeme očakávať pohyby cien akcie MSFT z intervalu od 28,709 (Scenár2SPSS) do 39,464 (Scenár1Excel) USD. Vypočítané odhady nám samozrejme umožňujú stanoviť aj očakávané hodnoty v ľubovoľnom časovom okamihu sledovaného obdobia. Zaujímavosťou môže byť, že scenáre získané na základe hodnôt náhodnej premennej $\tilde{Z}$ vygenerovaných v programe MS®Excel

vykazujú v naprostej väčšine prípadov vyššie hodnoty očakávanej ceny akcie MSFT ako je tomu v prípade scenárov vypočítaných na báze vygenerovaných hodnôt N(0,1) rozdelenia pomocou programu SPSS. Pre korektnosť je vhodné poznamenať, že generátor pseudonáhodných čísel programu MS®Excel nespĺňa podľa názorov niektorých autorov (napr. [9]) úplne požiadavky na profesionálnu kvalitu, napriek tomu sa dajú aj výsledky, získané pomocou tohto generátora považovať za vcelku veľmi dobré a vierohodné.
V *obrázku 1* sú znázornené aj odhadnuté kvantily, vymedzujúce hranice, v ktorých by sa mali ceny akcie MSFT v analyzovanom období náhodne pohybovať, pričom môžeme očakávať, že zhruba 10 % pokusov sa môže vyskytnúť aj mimo stanovené hranice.

5. Literatúra

[1] BARTOŠOVÁ, J. 2006. *Základy statistiky pro manažery*. Praha : Oeconomica, 2006. ISBN 80-245-1019-7.

[2] BOHDALOVÁ, M. – STANKOVIČOVÁ, I. 2006. *Using the PCA in the Analyse of the risk Factors of the investment Portfolio*. In: Forum Statisticum Slovacum 3/2006. Bratislava : SŠDS, 2006, s. 41 – 51. ISSN 1336-7420.

[3] GAVLIAK, R. 2005. *Možnosti predikcie variability finančných časových radov v podmienkach SR.* In: Forum Statisticum Slovacum 3/2005. Bratislava : SŠDS, 2005, s. 112 - 119. ISSN 1336-7420.

[4] HULL, J.C. 2000. *Options, Futures and Others Derivates*. London : Prentice-Hall, Inc., 2000. 698 s. ISBN 0-13-015822-4.

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

[6] KANDEROVÁ, M. 2006. *Testovanie hypotézy efektívnosti slovenského kapitálového trhu*. In: Forum Statisticum Slovacum 4/2006. Bratislava : SŠDS, 2006, s. 73 – 81. ISSN 1336-7420.

[7] LUHA, J. 2000. *Pravdepodobnostné kalkulátory*. In: Zborník príspevkov z 9. medzinárodného seminára Výpočtová štatistika, Bratislava : SŠDS, 7. – 8. 12. 2000.

[8] MELICHERČÍK, I. – OLŠAROVÁ, L. – ÚRADNÍČEK, V. 2005. *Kapitoly z finančnej matematiky*. Bratislava : Epos, 2005. 242 s. ISBN 80-8057-651-3.

[9] ZMEŠKAL, Z. a kol. 2004. *Finanční modely*. Praha : Ekopress, 2004. 236 s. ISBN 80-86119-87-4.

Adresa autora
Ing. Vladimír Úradníček, PhD.
Ekonomická fakulta UMB
Katedra kvantitatívnych metód a informatiky
Oddelenie štatistiky a ekonomickej analytiky
Tajovského 10
975 90 Banská Bystrica
vladimir.uradnicek@umb.sk

ŠTATISTICKÁ ANALÝZA NAJČASTEJŠÍCH DIAGNÓZ VO VYBRANOM ODBORE A ICH LIEČBY V SLEDOVANOM OBDOBÍ NA SLOVENSKU

Marianna Vavrová

Abstract

What were the 10 most frequent diagnoses? For how many diagnoses was prescribed a product? What were 10 drugs prescribed most often? … There are too much questions, but one source of knowledge for answers - a medicine prescription.

Úvod

Lekárska preskripcia je okrem základných informácií o pacientovi, predpísanom medikamente a jeho dávkovaní, tiež základným prvotným zdrojom informácií pre získanie vedomostí o zdravotnom stave populácie, výskyte najčastejších chorôb, ich liečbe a iných dôležitých poznatkov v tejto oblasti.

Cieľom tohto článku je poukázať na nevyhnutnosť využívania týchto vedomostí, či už pre zvýšenie informovanosti tých povolaných, ktorým sa zverujeme so svojimi zdravotnými problémami, alebo nás samotných.

Základné informácie

Údajová základňa pre túto analýzu bola zisťovaná v priebehu 2. štvrťroka 2006 na území Slovenska na reprezentatívnej vzorke ambulancií odborných lekárov - špecialistov v odbore interná medicína.

Následná analýza bola vykonaná na základe spracovania databázy údajov získaných z lekárskych preskripcií vybraného odboru - interná medicína.

Výsledky analýzy

Najčastejšie diagnózy

Tab. 1 Najčastejšie sa vyskytujúce diagnózy u pacientov na Slovensku v sledovanom období stanovených jedným náhodne vybratým lekárom internej medicíny

NÁZOV DIAGNÓZY	%
Primárna hypertenzia	49,1
Ischemická choroba srdca	21,3
Familiárna hypercholesterolémia	11,1
Prekonaný infarkt myokardu	3,7
Zmiešaná hyperlipoproteinémia	2,8
Hyperurikémia a dnavá artritída	1,9
Ischemická choroba srdca - nebolestivá forma	1,9
Predsieňová fibrilácia, Flutter v predsieni	0,9
Bifascikulárna blokáda	0,9
Iné	6,9

Zdroj: IMS

Ako vidno z tabuľky 1, alarmujúcich, takmer 50 % pacientov trpí primárnou hypertenziou (vysoký krvný tlak), viac ako jedna pätina pacientov trpí ischemickou chorobou srdca a viac ako každý desiaty pacient má vrodený vysoký cholesterol. Tieto výsledky boli zistené u konkrétneho (náhodne vybratého) lekára špecialistu v odbore interná medicína v sledovanom období.

Ostatné ochorenia nebudeme podrobnejšie interpretovať, ich výskyt viditeľne percentuálne klesá a čitateľ si ich vie zinterpretovať sám.

Tieto výsledky nie sú v rozpore s dlhoročným štatistickým výskumom v oblasti civilizačných chorôb, ktorý upozorňuje, že približne až 20 % zo všetkých prostriedkov na liečbu je vynakladaných práve na choroby srdcovo-cievneho a centrálneho nervového systému.

Päť najčastejších diagnóz a päť najčastejšie predpisovaných liekov na ich liečbu

Z dôvodu názornosti uvádzame výsledky tejto analýzy v tabuľke 2.

Tab. 2 Najčastejšie diagnózy a ich najčastejšia liečba jedným konkrétnym lekárom špecialistom v odbore interná medicína a lekárom špecialistom internej medicíny v priemere počas sledovaného obdobia

DIAGNÓZA	LIEKY PREDPÍSANÉ JEDNÝM LEKÁROM INTERNEJ MEDICÍNY	LIEKY PREDPÍSANÉ LEKÁROM TEJTO ŠPECIALIZÁCIE V PRIEMERE
Primárna hypertenzia	AMLOPIN	AGEN
	MONOZIDE	PRESTARIUM
	PIRAMIL	ACCUZIDE
	CARVEDILOL RAT	DIROTON
	MOXOGAMMA	TRITACE
Ischemická choroba srdca	ANOPYRIN	CORVATON
	SORBIMON	ANOPYRIN
	BISOGAMMA	PREDUCTAL MR
	DIGOXIN	TRIMETAZIDIN RAT
	CORVATON	SORBIMON
Familiárna hypercholesterolémia	SIMVOR	SIMVOR
	LIPANTHYL SUPRA	TORVACARD
	LIPANTHYL	SIMGAL
	-	SIMVASTATIN RAT
	-	SIMVACARD
Prekonaný infarkt myokardu	SIMVOR	SIMVOR
	PIRAMIL	PIRAMIL
	MONOPRIL	SORBIMON
	TRIMETAZIDIN RAT	TRIMETAZIDIN RAT
	CORVATON	CORVATON
Zmiešaná hyperlipoproteinémia	SIMVOR	TORVACARD
	ANOPYRIN	CRESTOR
	LIPANTHYL SUPRA	LESCOL XL
	-	LIPANTHYL
	-	ATORIS

Zdroj: IMS

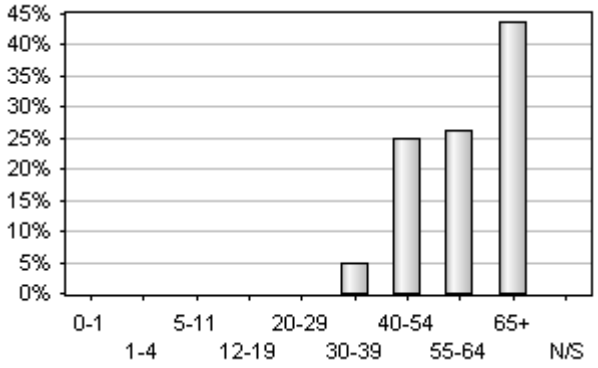
Údaje v predchádzajúcej tabuľke (Tab. 2) podrobnejšie neinterpretujeme, keďže predpokladáme, že čitateľ sa dokáže v tabuľke veľmi jednoducho orientovať.

Triedenie pacientov podľa pohlavia a veku

Z dôvodu zachovania anonymity pacientov možno výsledky triedenia podľa pohlavia a veku interpretovať iba veľmi opatrne.

I tak je zrejmé, že lekárov špecialistov v uvažovanom odbore navštevuje približne o 20 % viac žien ako mužov. Možná príčina tohto javu tkvie v tom, že pravdepodobne ženy viac dbajú o svoj zdravotný stav; avšak je to skôr konštatovanie ako tvrdenie.

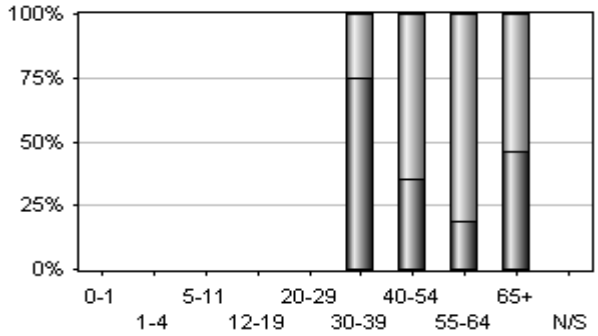
Graf 1 Triedenie pacientov podľa veku



Zdroj: IMS

Ako vidno z grafu 1 najviac pacientov so spomínanými ochoreniami je starších ako 65 rokov (takmer 45 % z daného počtu), avšak rizikovou vekovou skupinou, ktorá predstavuje 5 % z daného počtu pacientov je už v rozmedzí 30 - 39 rokov. Spomínanými ochoreniami trpí približne 25 % pacientov vo veku 40 - 54 rokov a približne rovnaké percento pacientov tvorí i skupina 55 - 64 ročných.

Graf 2 Triedenie pacientov podľa veku a pohlavia



Zdroj: IMS

Na grafe 2 je znázornené podvojné triedenie pacientov podľa veku a pohlavia súčasne. Bledosivú časť stĺpca predstavujú ženy a tmavosivou farbou sú znázornení muži. Zaujímavou

sa javí najmladšia veková skupina pacientov (30 - 39 roční), kde až 75 % zo všetkých pacientov tvoria práve muži. Naopak, uvedenými ochoreniami trpia ženy najviac (približne 80 %) vo veku 55 - 64 rokov, avšak i tak, okrem prvej uvažovanej vekovej skupiny, je všeobecne viac pacientiek ako pacientov.

Záver

Cieľom tohto príspevku bolo poukázať na využiteľnosť poznatkov z databáz v oblasti zdravotníctva so zreteľom zvyšovania informovanosti populácie. Závažným faktom je skutočnosť, že choroby srdcovo-cievneho systému sa objavujú už u 30 - 39 ročných mladých ľudí a z toho 75 % tvoria muži. Ďalším závažným faktom je, že choroby srdcovo-cievneho a centrálneho nervového systému odčerpávajú na Slovensku až 20 % všetkých prostriedkov určených na liečebné účely.

Literatúra

IMS Health: Doctor Feedback Report, Slovak Republic, 2 Q 2006, Docnum: 1085

Poďakovanie

Ďakujeme spoločnosti IMS Health za pomoc pri tvorbe tohto článku, ale i za konzultácie a poskytnutie odbornej rady.

Kontakt

Ing. Marianna Vavrová
INTERMEDI CENTRUM, s. r. o.
A. Hlinku 27
920 01 Hlohovec

e-mail: vm0011@gmail.com

ŠTATISTICKÁ ANALÝZA NÁKLADOVOSTI MEDIKAMENTÓZNEJ LIEČBY „NAJDRAHŠÍCH DIAGNÓZ“ VO VYBRANOM ODBORE V SLEDOVANOM OBDOBÍ NA SLOVENSKU

Marianna Vavrová, Ján Vavro

Abstract

What was the average value of a script written in the reporting time, and total cost? (A "script value" is equivalent to a "value-per-visit" and is calculated by totaling the value of all products written on the same script.) What are the "most expensive diagnoses" in the reported time? Which age group of patients is associated with the highest treatment cost? This paper contains some answers on these questions from area consultations by doctor specialist.

Úvod

V podstate nie je známa jediná vyspelá krajina, ktorá by nemala zabezpečenú starostlivosť o podporu prevencie, ochranu a rozvoj zdravia svojej populácie. Odborníci, zaoberajúci sa skúmaním vynakladaných prostriedkov na oblasť zdravotníctva sa vo všeobecnosti zhodujú, že náklady v tejto oblasti majú stúpajúcu tendenciu.

Cieľom tohto článku je poukázať na konkrétne náklady na liečbu vyskytujúce sa v ordináciách lekárov špecialistov vo vybranom odbore - interná medicína. Článok je zameraný hlavne na monitoring jedného konkrétneho lekára špecialistu v odbore interná medicína. Analýza bola vykonaná na základe spracovania databázy údajov získaných z lekárskych preskripcií vybraného odboru (interná medicína) za obdobie druhého štvrťroka 2006 na vybranej reprezentatívnej vzorke slovenských lekárov.

1. Nákladová analýza medikamentóznej liečby

Tab. 1 Nákladová analýza medikamentóznej liečby charakterizovaná vybranými základnými štatistikami z hľadiska jedného konkrétneho lekára špecialistu internej medicíny za jeden týždeň (vybraný bol náhodne 19. týždeň) a lekára špecialistu v danom odbore za celé sledované obdobie v priemere na týždeň

Názov základnej štatistiky	Lekár špecialista (interná medicína)	Lekári špecialisti (interná medicína)
Minimum	77 Sk	6 Sk
Aritmetický priemer	872 Sk	1 670 Sk
Maximum	3 288 Sk	127 023 Sk
Celkové náklady na preskripciu	69 788 Sk	120 725 Sk
Celkový počet preskripcií	80	72,3

Zdroj: IMS; MZ SR

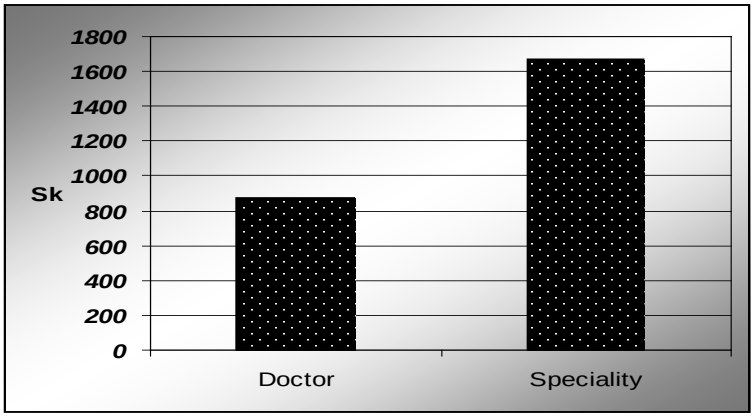
Výsledky z tabuľky 1 hovoria, že priemerné týždenné náklady na preskripciu v ambulancii lekára špecialistu v odbore interná medicína sú 872 Sk, najdrahší predpísaný liek jedným lekárom v sledovanom týždni bol za 3 288 Sk a najlacnejší v hodnote 77 Sk. Celkové týždenné náklady predstavujú približne sumu 69 788 Sk pri celkovom počte 80 vystavených lekárskych preskripcií jedným lekárom (uvažujeme, že jedna preskripcia obsahuje v priemere 5 druhov medikamentov).

Situácia v ambulanciách lekárov špecialistov daného odboru počas celého sledovaného obdobia v konkrétnom vybranom týždni je: najlacnejší predpísaný liek bol v priemernej hodnote 6 Sk a najdrahší v priemernej sume 127 023 Sk na lekára. Priemerné štvrťročné náklady na preskripciu u týchto lekárov boli v hodnote 1 670 Sk na lekára. Celková hodnota preskripcie predstavovala priemernú sumu 120 725 Sk pri priemernom počte približne 72 vystavených lekárskych preskripcií na jedného lekára špecialistu v odbore interná medicína (uvažujeme, že jedna preskripcia môže obsahovať 1 - 10 druhov medikamentov).

Pre názornosť uvádzame graf 1 týkajúci sa priemerných týždenných nákladov (19. týždeň) na preskripciu na jedného konkrétneho lekára špecialistu v odbore interná medicína (stĺpec

„Doctor“); a priemerných nákladov na týždennú preskripciu na jedného lekára špecialistu v odbore interná medicína (stĺpec „Speciality“) v celom sledovanom období.

Graf 1 Priemerné týždenné náklady na medikamentóznu liečbu v prepočte na jedného lekára



Zdroj: IMS; MZ SR

2. „Najdrahšie diagnózy“ z hľadiska medikamentóznej ambulantnej liečby

V tejto analýze sme sa zamerali iba na jedného konkrétneho lekára špecialistu v odbore interná medicína a skúmali sme iba 19. týždeň z celkového uvažovaného obdobia.

Výsledkom analýzy je 5 najnákladnejších diagnóz, ktorých kódy a názvy, celkový počet lekárom preskribovaných medikamentov za sledovaný týždeň, priemerná hodnota na jedno balenie a celková hodnota všetkých lekárom predpísaných balení medikamentov v sledovanom týždni (pri priemerných výpočtoch boli hodnoty zaokrúhľované na celé čísla, čo mohlo spôsobiť nepatrné rozdiely pri výpočtoch v poslednom stĺpci tabuľky).
Výsledky sú uvedené v tabuľke 2.

Tab. 2 Najnákladnejšie diagnózy z hľadiska medikamentóznej ambulantnej liečby lekára špecialistu v odbore interná medicína v 19. sledovanom týždni

DIAGNÓZA		**Lekár špecialista v odbore interná medicína**		
Kód	**Názov**	**Počet predpísaných medikamentov / týždeň**	**Priemerné týždenné náklady v Sk**	**Celkové týždenné náklady v Sk**
I100	Primárna hypertenzia	89	427	38 003
I250	Ischemická choroba srdca	47	305	14 335
E780	Familiárna hypercholesterolémia	13	587	7 631
I252	Prekonaný infarkt myokardu	10	339	3 390
E782	Zmiešaná hyperlipoproteinémia	3	437	1 311

Zdroj: IMS; MZ SR

Z výsledkov v tabuľke 2 vyplýva, že najnákladnejšiu liečbu si vyžaduje primárna hypertenzia (vysoký krvný tlak), na liečbu ktorej sa vynakladá najviac peňažných prostriedkov a lekár predpisoval na túto diagnózu najväčší počet liekov. Na druhom mieste najnákladnejších diagnóz sa nachádza ischemická choroba srdca, na treťom familiárna hypercholesterolémia (vrodený zvýšený cholesterol), za ňou na štvrtom mieste je prekonaný infarkt myokardu a posledné piate miesto najnákladnejších diagnóz patrí zmiešanej hyperlipoproteinémii. Výsledky v tabuľke sú jednoducho interpretovateľné a z tohto dôvodu sa ich ďalšou interpretáciou v tomto článku podrobnejšie nebudeme zaoberať.

3. Náklady na pacientov podľa jednotlivých vekových skupín

V tabuľke 3 sú uvedené náklady na pacientov podľa jednotlivých vekových skupín, a to v priemere na liečbu jedného pacienta podľa hodnoty predpisovaných medikamentov, ako i v celkovej výške nákladov podľa ceny predpisovaných medikamentov na všetkých pacientov danej vekovej skupiny. Analýza sa týkala iba konkrétneho lekára špecialistu v odbore interná medicína v náhodne vybranom 19. sledovanom týždni.

Tab. 3 Náklady na liečbu pacientov podľa jednotlivých vekových skupín

Vek	Priemerné náklady na pacienta v Sk	Celková suma nákladov v Sk
0 - 1	0	0
12 - 19	0	0
20 - 29	0	0
30 - 39	611	2 443
40 - 54	847	16 932
55 - 64	916	19 238
65 -	891	31 176

Zdroj: IMS; MZ SR

Je zrejmé, že najvyššie náklady na liečbu spomínaných druhov ochorení sú tvorené práve skupinou najstarších pacientov (65 ročných a starších), ktorá je z uvažovaných vekových skupín najpočetnejšia. Smerom k mladším skupinám pacientov náklady v celkovej hodnote klesajú. Treba poznamenať, že vo vekovej skupine 30 - 39 ročných pacientov sú priemerné náklady na pacienta (vzhľadom na vek) pomerne vysoké, hoci celková suma nákladov svedčí o tom, že táto veková skupina nie je veľmi početná.

Interpretáciou ostatných výsledkov v tabuľke 3 sa nebudeme z dôvodu dostatočnej názornosti podrobnejšie zaoberať.

Záver

Cieľom tohto článku bola nákladová analýza medikamentóznej liečby v ordináciách lekárov špecialistov vo vybranom odbore interná medicína, s prevažným zameraním na konkrétneho lekára špecialistu v odbore interná medicína v náhodne vybranom 19. týždni.

Z pohľadu autorov tohto článku bol cieľ splnený, keďže 1. časť bola zameraná na nákladovú analýzu medikamentóznej liečby prostredníctvom vybraných základných štatistík, v druhej časti boli vyhodnotené „najdrahšie diagnózy“ z hľadiska medikamentóznej ambulantnej liečby a posledná tretia časť bola zameraná na náklady vynakladané na liečbu pacientov podľa vekových skupín.

Treba poznamenať, že práca v tejto oblasti je iba v začiatočnom štádiu a autori by radi pokračovali v jej ďalšom napredovaní.

Literatúra

[1] IMS Health: Doctor Feedback Report, Slovak Republic, 2 Q 2006, Docnum: 1085

[2] Opatrenenie MZ SR č. 07045/2003 OAP z 31. 12. 2003, ktorým sa ustanovuje rozsah regulácie cien v oblasti zdravotníctva v znení neskorších predpisov

[3] Rovný - I., Mihinová - D.: Európska únia a verejné zdravotníctvo, Lekársky obzor 5/2006, roč. LV., ISSN 0457-4214, str. 220 - 222

Poďakovanie

Ďakujeme spoločnosti IMS Health za pomoc pri tvorbe tohto článku, ale i za konzultácie a poskytnutie odbornej rady.

Kontakty

Ing. Marianna Vavrová
INTERMEDI CENTRUM, s. r. o.
A. Hlinku 27
920 01 Hlohovec
e-mail: vm0011@gmail.com

MUDr. Ján Vavro
INTERMEDI CENTRUM, s. r. o.
A. Hlinku 27
920 01 Hlohovec
e-mail: intvav@mail.t-com.sk

Metóda hlavných komponentov ako prostriedok prieskumnej analýzy efektívnosti priemyselných podnikov v okresoch SR

Mária Vojtková

Abstract: Principal component analysis (PCA) is one of the most widely used multivariate techniques of exploratory data analysis. The principal components are such that most of the information, measured in terms of total variance, is preserved in only a few of them. In this artictle PCA has been applied on example of the economic efficiency of enterprises.

Key words: Multivariate Analysis, Principal Component Analysis, Measure of Sampling Adequacy, Number of Components, Ortogonal Rotations, Estimation of Factor Scores.

1. Úvod

Na poznanie a pochopenie dát celá rada rôznych autorov doporučuje takmer každú viacrozmernú úlohu začať výpočtom prípadne zobrazením hlavných komponentov. Metóda hlavných komponentov je považovaná za jednu z metód prieskumu dát, pričom dáva odpoveď na mnohé otázky. S jej pomocou je možné identifikovať odľahlé resp. extrémne pozorovania alebo riešiť problém multikolinearity vysvetľujúcich premenných. Považuje sa tiež za metódu analýzy skrytých vzťahov alebo zníženia dimenzie. V tomto príspevku sme sa rozhodli uskutočniť prieskumnú analýzu efektívnosti hospodárenia priemyselných podnikov v jednotlivých okresoch SR pomocou metódy hlavných komponentov s využitím rotácie, ktorá bola historicky vyvinutá až o niečo neskôr v súvislosti s faktorovou analýzou. Samotnú aplikáciu sme uskutočnili v SAS Enterprise Guide verzia 4.1.

Predmetom analýzy je skupina kvantitatívnych premenných, ktoré sú vyjadrené v rôznych merných jednotkách. Z toho dôvodu, aby sme ich previedli na spoločný základ pracujeme s ich normovanými hodnotami. 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:

$$F_h = \alpha_{1h} X_1 + \alpha_{2h} X_2 + ... + \alpha_{kh} X_k \qquad (1)$$

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).

2. Popis vstupných premenných

Predmetom analýzy sú anonymizované priemyselné podniky charakterizované ukazovateľmi výstupu z hospodárskeho procesu, poskytnuté pre vedecké účely zo Štatistického úradu SR za roky 2003. Vzhľadom k tomu, že ide o anonymné podniky, samotná analýza je zameraná na okresy, z ktorých jednotlivé podniky pochádzajú. Každý okres je charakterizovaný týmito syntetickými ukazovateľmi efektívnosti:

Z/V - ziskovosť,
PH/ZAM - produktivita práce z pridanej hodnoty,
ODP/V - nákladovosť odpisov,

ZAV/POH - koeficient prvotnej platobnej neschopnosti,
Z/CK - rentabilita celkového kapitálu,
Z/PH - rentabilita pridanej hodnoty,
V/ZAM - produktivita práce z výnosov,
TRZBY/ZI - obrátkovosť základného imania.

Kde Z je výsledok hospodárenia pred zdanením, V sú výnosy, TRZBY sú tržby z predaja, CK je celkový kapitál (pasíva spolu), PH je pridaná hodnota, ZI je základné imanie,ODP sú odpisy, ZAV sú záväzky, POHL sú pohľadávky a ZAM je priemerný počet zamestnancov v prepočítaných osobách.

3. Aplikácia metódy hlavných komponentov

Predpokladom použitia metód zníženia dimenzie je závislosť medzi sledovanými ukazovateľmi. Koreláciu medzi ukazovateľmi je možné vyšetriť viacerými spôsobmi:

1. Ak predpokladáme, že dáta pochádzajú z viacrozmerného normálneho rozdelenia, môžeme testovať hypotézu, že korelačná matica je jednotkovou maticou pomocou Bartlettovho testu.
2. Parciálne korelačné koeficienty sú indikátormi sily vzťahu medzi znakmi. Ak sa za závislosťou ukazovateľov skrývajú spoločné skyté premenné, parciálne korelačné koeficienty medzi pôvodnými premennými sú blízke nule.
3. Kaiser-Meyerova-Olkinova (KMO) miera je index porovnávajúci veľkosť korelačných koeficientov voči veľkosti parciálnych korelačných koeficientov. Jej doporučované hodnoty sú nad 0,5, pričom vyššia hodnota signalizuje vhodnejšie použitie príslušného ukazovateľa (maximálna hodnota je 1). Zjednodušenou analogickou mierou je MSA (Measure of Sampling Adequacy), ktorú môžeme interpretovať rovnako ako KMO.

Vzhľadom k veľkému rozsahu výstupov si uvedieme iba tretí spôsob posúdenia vhodnosti dát (tabuľka 1) pomocou mier KMO pre všetky vstupné premenné. Ukazovateľ nákladovosť odpisov a obrátkovosť základného imania nadobúda veľmi nízke hodnoty KMO, takže bude z analýzy vylúčený.

Tabuľka 1. KMO miery adekvátnosti všetkých vstupných ukazovateľov

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.56991155							
Z/V	**PH/ZAM**	**ZAV/POH**	**ODP/V**	**Z/CK**	**Z/PH**	**V/ZAM**	**TRZBY/ZI**
0.628919	0.467999	0.679752	0.069138	0.625686	0.764405	0.451208	0.233366

Po vylúčení nevhodných vstupných ukazovateľov všetky miery KMO dosahujú dostatočne vysokú hodnotu (viď tabuľka 2), čiže koreláciu medzi nimi bude možné vysvetliť pomocou skrytých premennných.

Tabuľka 2. KMO miery adekvátnosti vybraných vstupných ukazovateľov

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.72310040					
Z/V	**PH/ZAM**	**ZAV/POH**	**Z/CK**	**Z/PH**	**V/ZAM**
0.72170248	0.56782621	0.91026944	0.83450975	0.77232456	0.56011616

Celkovo je možné vytvoriť rovnaký počet spoločných hlavných komponentov ako je počet analyzovaných premenných, čiže šesť. Vlastné číslo charakterizuje rozptyl každého komponenta, pričom celkový rozptyl vzhľadom k tomu, že pracujeme so štandardizovanými

premennými je rovný 6. Najviac nás bude zaujímať posledný stĺpec v tabuľke 3, ktorý vyjadruje kumulatívny podiel vysvetlený daným počtom komponentov. V záujme zachovania pôvodných informácií by jeho hodnota mala byť, čo najvyššia (pri exaktných vedách sa doporučuje okolo 90%). Ďalším kritériom pre výber štatisticky významných komponentov je porovnanie vlastných čísel s priemerným vlastným číslom, čiže 1. Celkovo teda môžeme uvažovať s 2-3 hlavnými komponentami. Keďže tretí hlavný komponent vysvetľuje ešte 12,18% variability, čo je celkom zaujímavý prínos, budeme uvažovať i s ním. Správnosť zaradenia tohto komponenta si dokážeme ďaľšou analýzou.

Tabuľka 3. Rozptyly jednotlivých spoločných faktorov

Vlastné čísla korelačnej matice: Úhrn = 6 Priemer = 1				
	Vlastné čísla	**Rozdiel**	**Podiel**	**Kumulatívny podiel**
1	3.49812141	1.97358259	0.5830	0.5830
2	1.52453882	0.79376136	0.2541	0.8371
3	0.73077746	0.56154562	0.1218	0.9589
4	0.16923184	0.11369241	0.0282	0.9871
5	0.05553943	0.03374839	0.0093	0.9964
6	0.02179104		0.0036	1.0000

Rotovaná komponentná matica (tabuľka 4) obsahuje komponentné saturácie pre jednotlivé znaky a hlavné komponenty. V prípade ortogonálnej rotácie ide o korelačné koeficienty medzi každým vybraným ukazovateľom a komponentom. Rotácia je prostriedkom zjednodušenia komponentnej štruktúry pri interpretácii a v našom prípade bola použitá ortogonálna rotácia Equamax. Za významné komponentné saturácie sa považujú váhy väčšie ako 0,5. V našom prípade sme na základe ich hodnôt vytvorili tri hlavné komponenty (HK):

1. hlavný komponent – rentabilita,
2. hlavný komponent – produktivita,
3. hlavný komponent – likvidita.

Tabuľka 4. Rotovaná komponentná matica

Rotated Factor Pattern			
	HK1	**HK2**	**HK3**
Z/V	0.95667	0.15092	-0.21200
Z/CK	0.95185	0.07758	-0.21540
Z/PH	0.93180	0.20672	-0.25029
V/ZAM	0.06795	0.95066	-0.10541
PH/ZAM	0.17220	0.93912	-0.08676
ZAV/POH	-0.21488	-0.10629	0.97075

Na základe tabuľky 5 môžeme zhodnotiť kvalitu získaného 3-komponentného modelu. V absolútnom vyjadrení z celkového rozptylu spomínané 3 komponenty dokážu vysvetliť 5,753438 t.j. 95,89%. Pri jednotlivých premenných je podiel variability vysvetlený

uvedenými 3 hlavnými komponentami veľmi vysoký. Napríklad pri ukazovateli ziskovosť môžeme jeho variabilitu na 98,3 % vysvetliť pomocou uvedeného 3-komponentného modelu atď..

Tabuľka 5. Odhad rozptylu vysvetlený 3-komponentným modelom pre jednotlivé ukazovatele

Final Communality Estimates: Total = 5.753438					
Z/V	**PH/ZAM**	**ZAV/POH**	**Z/CK**	**Z/PH**	**V/ZAM**
0.98294336	0.91912783	0.99982824	0.95842972	0.97362120	0.91948733

V prípade, že by sme uvažovali iba s 2-komponentným modelom podiel variabiltiy pri ukazovateli koefiecient prvotnej platobnej schopnosti sa značne zníži (tvorí iba 34,3%), čo iba potvrdzuje správnosť troch zvolených hlavných komponentov. Celkovo môžeme teda uvedený model považovať za prijateľný a interpretovateľný o čom svedčí i reziduálna korelačná matica so špecifickými rozptylmi na diagonále (tabuľka 6). Špecifické rozptyly charakterizujú nevysvetlenú variabilitu spôsobenú náhodou. Ich hodnoty na diagonále tvoria doplnok k vysvetlenej variabilite vyjadrenej v tabuľke 5 a v prípade kvalitného modelu by mali byť, čo najnižšie. Ďalšou charakteristikou umožňujúcou zhodnotiť kvalitu získaného riešenia je charakteristika RMS (tabuľka 7). Je to priemerná suma štvorcov reziduí, pričom jej hodnota pre dobrý model nesmie presiahnuť 0,5.

Tabuľka 6. Reziduálna korelačná matica

Residual Correlations With Uniqueness on the Diagonal						
	X1	**X2**	**X4**	**X6**	**X7**	**X8**
Z/V	0.01706	0.00198	-0.00035	-0.01585	-0.00155	-0.00307
PH/ZAM	0.00198	0.08087	-0.00343	-0.01721	0.00568	-0.08042
ZAV/POH	-0.00035	-0.00343	0.00017	0.00023	0.00055	0.00332
Z/CK	-0.01585	-0.01721	0.00023	0.04157	-0.02453	0.02149
Z/PH	-0.00155	0.00568	0.00055	-0.02453	0.02638	-0.00904
V/ZAM	-0.00307	-0.08042	0.00332	0.02149	-0.00904	0.08051

Tabuľka 7. Priemerná suma štvrocov reziduí pri 2-komponentnom modeli

Root Mean Square Off-Diagonal Residuals: Overall = 0.02342464					
Z/V	**PH/ZAM**	**ZAV/POH**	**Z/CK**	**Z/PH**	**V/ZAM**
0.00730728	0.03690875	0.00215633	0.01794860	0.01198514	0.03749912

Výsledné 3 hlavné komponenty sú vzájomne nezávislé, čo umožňuje ich ďalšie použitie v metódach vyžadujúcich vstup nekorelovaných premenných. Miesto pôvodných znakov je teda možné pracovať s hodnotami komponentných skóre. Ide o kombinovanú mieru každého komponenta vyčíslenú pre jednotlivé okresy. Na základe jej veľkosti je možné sledovať poradie skrytých premenných vyjadrujúcich určitú vlastnosť skúmaného súboru ukazovateľov.

Napríklad pre prvý hlavný komponent môžeme podľa veľkosti komponentných skóre vyhodnotiť okres s priemyselnými podnikmi dosahujúcimi najvyššiu rentabilitu - Zlaté Moravce a okres s podnikmi charakteristickými najnižšou rentabilitou – Detva.

Druhý hlavný komponent dosahuje najvyššie hodnoty ukazovateľov produktivity práce v okresoch Bratislava IV, II, I, čo je typické pre zastúpenie dvoch priemyselných gigantov v prvých dvoch okresoch. Najnižšie hodnoty produktivity práce sú charakteristické pre okresy na východe Slovenska Bardejov, Veľký Krtíš a Snina.

Z hľadiska skrytého ukazovateľa likvidity sa ako najlepšie javia prekvapujúco okres Svidník, Levoča, Pezinok a ako najhoršie okresy Banská Štiavnica, Košice – okolie a Hlohovec. Ukazovateľ - koeficient prvotnej platobnej neschopnosti je možné hodnotiť i samostatne, avšak skrytý hlavný komponent, ktorý sme pomenovali na základe vysokej saturácie pri tomto ukazovateli v sebe zahŕňa aj informáciu o veľkosti ostatných ukazovateľov efektívnosti, avšak v menšej miere.

4. Záver

Silná závislosť sledovaných ukazovateľov, optimálna voľba relatívne malého počtu hlavných komponentov, silná korelácia medzi vstupnými ukazovateľmi a ortogonálnymi komponentami, sú dôležité podmienky použitia hodnôt hlavných komponentov pri sledovaných okresoch. Okrem posúdenia kvality dát vytvárajú nové skryté hlavné komponenty príležitosť k ich použitiu v rôznych oblastiach života.

5. Literatúra

CHAJDIAK, J.: *Ekonomická analýza stavu a vývoja firmy*. Bratislava: Statis, 2004.

KHATTREE, R. – NAIK, N. D.*: Multivariate Data Reduction and Discrimination with SAS® Software.* First edition, Cary, NC: SAS Institute Inc., 2000.

LUHA, J.: Viacrozmerné štatistické metódy analýzy kvalitatívnych znakov. *EKOMSTAT 2005,* Štatistické metódy v praxi. SŠDS Trenčianske Teplice 22. – 27. 5. 2005.

MELOUN, M. – MITICKÝ, J.: *Statistická analýza experimentálních dat*. Praha: ACADEMIA, 2004.

SHARMA, S.: *Applied multivariate techniques*. New York: John Wiley & Sons, 1996.

ŠOLTÉS, E. – ŠOLTÉSOVÁ, T: Analýza vplyvu vybraných faktorov na výšku poistných plnení v povinnom zmluvnom poistení SR. In.: *Forum Staticum Slovacum. 1/2006*, s. 151-157, Bratislava: SŠDS, 2006. ISSN 1336-7420

Adresa autora:
Mária Vojtková, Ing. PhD.
Katedra štatistiky, FHI, Ekonomická univerzita
Dolnozemská 1/b
852 35 Bratislava
E-mail: vojtkova@euba.sk

Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/2631/05 „Analýza možností aplikácie viacrozmerných štatistických metód na skúmanie ekonomických výsledkov na príklade priemyslu SR prípadne iných oblastí ekonomiky“.

Rizikové faktory aterosklerózy u rómskej a slovenskej populácie

Ladislava Wsólová, Zuzana Bašistová, Daniela Siváková, Mária Zacharová

Abstract

The influence of apolipoprotein E genotypes (APOE) on plasma lipid levels and the interaction with other risk factors of atherosclerosis was determined in two population samples in Slovakia: 150 Romany and 348 Slovak individuals. We detected decreased LDL cholesterol concentrations in males with APOE*2 genotype. The ethnic samples differ significantly in total cholesterol, LDL-cholesterol and HDL-cholesterol.

Key words: ApoE, lipids, risk factors, atherosclerosis, Slovakia

1. Úvod

Ateroskleróza, chronické degeneratívne ochorenie cievnej steny, je hlavnou príčinou srdcovo cievnych ochorení a sprevádza ľudstvo niekoľko tisícročí. Aterosklerotické lézie boli opísané už pri pitvách egyptských faraónov (Češka 1999). Dodnes však neexistuje exaktná definícia aterosklerózy. Jedným z dôvodov je neznámy presný mechanizmus jej vzniku.

Klinický obraz aterosklerózy je veľmi variabilný. Rozoznávame tri štádiá aterosklerózy (Gavorník 1999):

1.latentné štádium – chorý nepociťuje žiadne ťažkosti napriek dokázateľným aterosklerotickým zmenám,

2.manifestné štádium – je charakteristické subjektívnymi ťažkosťami pacienta, ktoré vyvoláva ischémiou postihnutý orgán,

3.štádium s komplikáciami – v popredí sú príznaky vyplývajúce z ischémie alebo nekrózy postihnutého orgánu.

Nakoľko etiopatogenéza aterosklerózy je mimoriadne zložitá, prevencia a liečba tohto ochorenia musí byť komplexná. Na vzniku aterosklerózy sa podieľa viacero exogénnych a genetických faktorov, ktoré sa najčastejšie rozdeľujú na ovplyvniteľné a neovplyvniteľné.

Ovplyvniteľných rizikových faktorov aterosklerózy je podľa mnohých štúdií viac ako dvesto, preukázateľne významných je zatiaľ iba niekoľko. Medzi najvýznamnejšie **ovplyvniteľné rizikové faktory** patrí: fajčenie, stres, nedostatočná telesná aktivita, nadmerná konzumácia alkoholu, diabetes mellitus, hyperhomocysteinémia, artériová hypertenzia, obezita, nesprávne stravovacie návyky, dyslipoproteinémia a iné.

Medzi **neovplyvniteľné rizikové faktory** patrí: vek, pohlavie, genetická predispozícia a pozitívna rodinná anamnéza. Keďže ateroskleróza je dlhodobý proces, riziko jej vzniku vzrastá s vekom. Fakt, že muži majú výrazne vyššie riziko vzniku aterosklerózy ako ženy do menopauzy je jednoznačne preukázaný. Za pozitívnu rodinnú anamnézu z hľadiska aterosklerózy sa považuje výskyt infarktu myokardu alebo náhlej smrti otca či prvostupňového mužského príbuzného vo veku nižšom ako 55 rokov, u matky a prvostupňových príbuzných ženského pohlavia vo veku nižšom ako 65 rokov (Češka 1999). Z génov, resp. ich polymorfizmov, u ktorých sa predpokladá vzťah k ateroskleróze, sme sa zamerali na gén pre apolipoproteín E (APOE). ApoE je kľúčovým regulátorom hladín lipidov v plazme.

Genetický polymorfizmus prvýkrát opísal Utermann et al. (1975, 1977). Gén APOE je kódovaný tromi alelami označovanými ako APOE*2, APOE*3 a APOE*4. V kaukazoidnej populácii sa alela APOE*4 vyskytuje vo frekvencii 0,15, APOE*3 0,77 a APOE*2 0,08. Populačné štúdie ukázali, že nositelia APOE*2 majú v porovnaní s nositeľmi alely APOE*3 nižšie hodnoty celkového cholesterolu, LDL-cholesterolu a o niečo vyššie hladiny triglyceridov. Nositelia alely APOE*4 majú opačné charakteristiky, preto sú z hľadiska vzniku a rozvoja aterosklerózy rizikovejší (Weisgraber et al. 1982, Mahley et Inerarity 1983, Rahl et Mahley 1992, Ordovas et Schaefer 1999, Zacharová 2003, Siváková et al. 2006). Niektorí autori zaraďujú medzi neovplyvniteľné rizikové faktory aj etnické faktory, ktoré je však ťažko odlíšiť od vplyvu prostredia. Preto bolo aj cieľom tejto práce zistiť vplyv etnika na hladiny jednotlivých lipidov.

2. Materiál a metódy

Rómsku populáciu tvorili obyvatelia obce Zlaté Klasy, ktorá leží neďaleko Bratislavy. V čase zberu dát mala obec 3322 obyvateľov, asi polovica z nich bola rómskeho pôvodu. Do štúdie vstúpilo 150 dospelých dobrovoľníkov (68 mužov a 82 žien).
Slovenskú populáciu zastupovalo 348 dospelých dobrovoľníkov (172 mužov a 176 žien) z okresu Banská Bystrica, ktorí sa zúčastnili projektu CINDI (Countrywide integrated non-communicable diseases intervention program).

U všetkých probandov boli zisťované nasledovné antropometrické údaje: telesná výška, telesná hmotnosť, obvod pása a obvod bokov. Z nich bol vypočítaný BMI (body mass index) ako pomer hmotnosti v kg a druhej mocniny výšky v m a WHR index (waist to hip ratio) ako pomer obvodu pása a bokov. Zo vzoriek krvi sa stanovovali hladiny celkového cholesterolu (TC), triglyceridov (TG), HDL-cholesterolu (HDL-C). Hodnoty LDL-cholesterolu sa vypočítali podľa Friedewaldovej rovnice. DNA bola izolovaná fenolovou metódou z leukocytov. Genotypy APOE boli stanovené PCR (polymerase chain reaction) metódou. Dosiahnuté vzdelanie bolo rozdelené na tri kategórie: základné (1), učňovské (2) a vyššie (3). Na štatistickú analýzu bol použitý štatistický softvér SPSS 13.0 pre Windows.

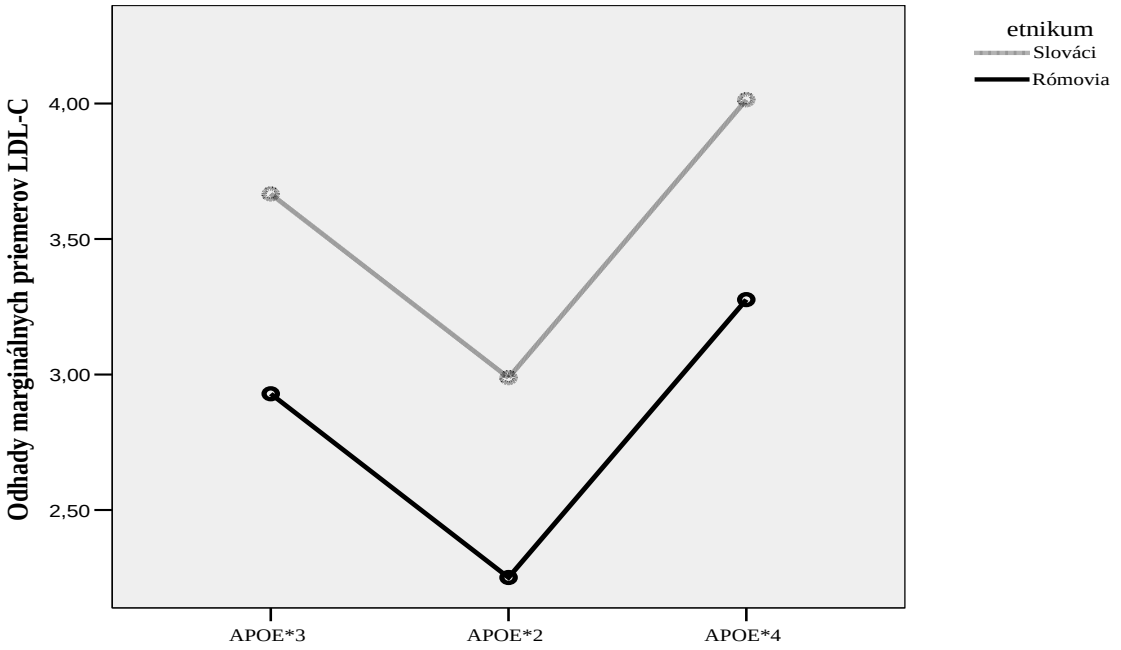
3. Výsledky

Hodnoty celkového cholesterolu, LDL-cholesterolu, HDL-cholesterolu a triglyceridov predstavovali závislé premenné, ktoré môžu do značnej miery charakterizovať stav tepien. Vek, pohlavie, etnikum, genotypy APOE, vzdelanie a antropometrické údaje tvorili skupinu nezávislých premenných. Naším cieľom bolo nájsť vhodné štatistické modely pre vzťah medzi jednotlivými závislými premennými a skupinou nezávislých premenných. Tieto modely sme robili pre každé pohlavie zvlášť, nakoľko sú medzi mužmi a ženami preukázateľné rozdiely v antropometrických parametroch, ale aj v hodnotách HDL-cholesterolu a triglyceridov. Uvedené výsledky predstavujú z nášho pohľadu modely s najlepšou výpovednou hodnotou.

Muži:

Hodnoty celkového cholesterolu boli u mužov štatisticky významne ovplyvňované predovšetkým vekom a etnikom. So zvyšujúcim sa vekom, hodnoty celkového cholesterolu stúpali, pričom nižšie hodnoty mala rómska populácia.
Hodnoty HDL-cholesterolu boli štatisticky významne ovplyvňované etnikom a obvodom pása. Rómska populácia mala nižšie hodnoty a s rastúcim obvodom pása klesali hodnoty HDL-cholesterolu.
Hodnoty LDL-cholesterolu boli štatisticky významne ovplyvňované etnikom, genotypmi APOE a BMI. Najnižšie hodnoty LDL cholesterolu mali nositelia alely APOE*2, najvyššie

hodnoty nositelia alely APOE*4, pričom rómska populácia dosahovala nižšie hodnoty vo všetkých genotypoch. So stúpajúcim BMI sa hodnoty LDL-cholesterolu zvyšovali.
Hodnoty triglyceridov boli štatisticky významne ovplyvňované iba indexom WHR, pričom s rastúcimi hodnotami WHR sa zvyšovala aj hladina TG.



Graf 1. Vzťah medzi APOE, etnikom a LDL-cholesterolom u mužov.

Ženy:

Hodnoty celkového cholesterolu boli štatisticky významne ovplyvňované predovšetkým etnikom a vekom. So zvyšujúcim sa vekom hodnoty celkového cholesterolu stúpali, pričom nižšie hodnoty mala rómska populácia.
Hodnoty HDL-cholesterolu boli štatisticky významne ovplyvňované etnikom a obvodom pása. S rastúcim obvodom pása klesali hodnoty HDL-cholesterolu, pričom rómska populácia mala nižšie hodnoty.
Hodnoty LDL-cholesterolu boli štatisticky významne ovplyvňované etnikom, vekom a telesnou výškou. Rómska populácia mala nižšie hodnoty, so zvyšujúcim sa vekom hodnoty LDL-cholesterolu stúpali a s rastúcou telesnou výškou klesali.
Hodnoty triglyceridov boli štatisticky významne ovplyvňované obvodom pása, vekom a telesnou výškou. S rastúcim obvodom pása a so zvyšujúcim sa vekom hodnoty T
G stúpali a s rastúcou telesnou výškou klesali.

4. Záver

Ako z uvedených výsledkov vyplýva, etnikum výrazne ovplyvňuje celkový cholesterol, HDL aj LDL-cholesterol. Hodnoty triglyceridov sa nelíšili medzi etnikami ani u jedného pohlavia. Pre celkový cholesterol a HDL-cholesterol sme získali u oboch pohlaví podobný model. Avšak pri hodnotách LDL-cholesterolu a triglyceridoch je medzi modelmi pre rôzne

pohlavia značný rozdiel. Štatisticky významný vplyv genotypov APOE bol zistený iba u LDL-cholesterolu u mužov.

5. Literatúra

1. Češka, R., 1999: Cholesterol a ateroskleróza. Léčba hyperlipidémií. 2. vyd., MAXDORF, Praha, 226 s.
2. Gavorník, P., 1999: Ateroskleróza a iné choroby tepien. Univerzita Komenského, Bratislava, 216 s.
3. Mahley, R.W., Innerarity, T.L., 1983: Lipoprotein receptors and cholesterol homeostasis. Biochem. Biophys. Acta, 737:197-222.
4. Ordovas, J.M., Schaefer, E.J., 1999: Genes, variation of cholesterol and fat intake and serum lipids. Curr Opin Lipidol, 10:15-22.
5. Rahl, S.C. Jr., Mahley, R.W., 1992: The role of apolipoprotein E genetic variants in lipoprotein disorders. J Intern Med, 231.633-659.
6. Siváková, D., Zacharová, M., Gašparovič, J., Rašlová, K., Wsólová, L., Bašistová, Z., Blažíček, P.: Apolipoprotein E Polymorhism in Relation to Plasma Lipid Levels and Other Risk Factors of Atheosclerosis in Two Ethnic Groups from Slovakia. Coll.Antropol. 30 (2006) 2: 387-394.
7. Utermann, G., Jaeschke, M., Menzel, J., 1975: Familiar hyperlipoproteinemia type III: deficiency of a specific apolipoprotein (apoE III) in the very-low-density lipoprotein. FEBS Lett., 56:352-355.
8. Utermann, G., Hees, M., Steinmetz, A., 1977: Polymorphism of apolipoprotein E and occurrence of dysbetalipoproteinemia in man. Nature, 269:604-607.
9. Weisgraber, K.H., Innerarity, T.L., Mahley, R.W., 1982: Abnormal lipoprotein receptor biology activity of human E apoprotein due to cystein – argine interchange at a single site. J. Biol. Chem., 257:2518-2521.
10. Zacharová, M., 2003: Globálne riziká aterosklerózy v rómskej populácii Slovenska. Dizertačná práca. Univerzita Komenského, Prírodovedecká fakulta. Bratislava.

Adresa autorov
RNDr. Ladislava Wsólová
Slovenská zdravotnícka univerzita, Limbová 12
833 03 Bratislava
ladislava.wsolova@szu.sk
Mgr. Zuzana Bašistová,
Prof. RNDr. Daniela Siváková, CSc.,
RNDr. Mária Zacharová, PhD.
Prírodovedecká fakulta Univerzity Komenského, Bratislava, SR

Shlukování ve velkých souborech dat

Marta Žambochová

Abstract:

The paper discusses about clustering methods for large datasets. First of all it focuses on methods that use assorted tree structures. Common clustering methods fail in very large dataset processing. Be needed to find some new alternative of clustering algorithms. The advantage of using trees in algorithms is the fact that more primary data are collected in the node of the tree and we can process they together. By it we down the number of operations. We usually create the tree structure only once and we optimize this structure after it. Thereby it happens to some reduction of a processing term and reduction of HW demands.

Key words:

Data Clustering, Large Datasets, R*-trees, Mrkd-trees, CF-trees

Úvod

Shluková analýza se stala jednou z hlavních metod používaných v data mining, v oblasti dobývání znalostí z databází. Jedná se o jeden z druhů klasifikace. Obecně je klasifikace metodou pro rozdělování dat do skupin dle jistých kriterií. Pokud jsou tato kriteria předem známa, alespoň pro vzorek dat, můžeme vytvořit klasifikační model. Častější je však případ klasifikace, kdy výsledná kriteria nejsou předem známa a úlohou klasifikace je jejich nalezení. Používanou technikou v takovýchto případech je shluková analýza

Shlukování je důležitá technika pro odhalování struktury dat. Shluková analýza se zabývá podobností datových objektů. Řeší dělení množiny datových bodů do několika skupin (shluků, clusters) tak, aby si objekty uvnitř jednotlivých shluků byly co nejvíce podobny a objekty z různých shluků si byly podobny co nejméně. Při tom každý datový bod je popsán skupinou znaků (proměnných). Výsledky analýzy závisí na volbě proměnných, zvolené míře podobnosti mezi datovými body a shluky a na zvoleném algoritmu výpočtu.

Pomocí nalezených shluků pak můžeme lépe vyšetřit strukturu dat a s jednotlivými shluky můžeme následně pracovat hromadně jako s jedním objektem. Nalezené shluky ale vystihují strukturu dat pouze z pohledu vybraných znaků. Nevhodný výběr znaků může vést k zavádějícím závěrům.

Je popsáno mnoho metod shlukové analýzy, ale většina z nich má jednu společnou negativní vlastnost. Není rozumně použitelná pro velké objemy dat. Jako velké množství dat se označuje již počet nad 250 objektů. Mnohé zpracovávané soubory jsou však mnohokrát větší. Proto se v poslední době objevují nové algoritmy shlukové analýzy, které jsou určeny právě pro zpracování datových souborů s velkým počtem objektů. Tyto algoritmy jsou často založeny na principu různých typů kořenových stromů.

Stromem se rozumí souvislý acyklický graf. Kořenovým stromem pak strom, jehož jeden uzel je vybrán jako hlavní. Tento vybraný uzel se nazývá kořen. U stromů využívaných v metodách shlukové analýzy reprezentují uzly stromu určitou množinu datových bodů. Kořen reprezentuje celý zpracovávaný soubor. Větve stromu vždy představují určité dělení množiny dat obsažených v rodičovském uzlu do několika podmnožin dat reprezentovaných uzly – potomky.

Výhodou takovéto reprezentace dat je fakt, že mnohé početní i rozhodovací operace v příslušném algoritmu se mohou nahradit jediným hromadným úkonem a mohou se provádět místo s jednotlivými datovými objekty s celou množinou objektů spojenou ve společném uzlu. Navíc většina těchto algoritmů buduje strom pouze jedenkrát a pak jej

v jednotlivých iteračních krocích upravuje a optimalizuje. Tím se velmi sníží počet přístupů k datům i aritmetických a rozhodovacích operací.

Varianta algoritmu CLARANS pro velké databáze

Jedním z algoritmů pro shlukování ve velkých souborech dat je algoritmus zvaný CLARANS (Clustering Large Application based on RANdomized Search). Základem tohoto algoritmu je metoda k-medoidů.

Algoritmus CLARANS řeší problematiku shlukování množiny objektů pro předem daný počet shluků *k* a pro definovanou míru vzdálenosti mezi dvěmi objekty (pro prostorová data je přirozenou měrou vzdálenosti eukleidovská vzdálenost). Algoritmus hledá množinu *k* vybraných objektů (tzv. medoidů) takových, že průměrná vzdálenost všech objektů ze zkoumaného souboru k jejich nejbližšímu medoidu je minimální. Medoidem tedy nazýváme objekt z podmnožiny, který je umístěn nejblíže středu této podmnožiny objektů. Výsledkem algoritmu bude *k* shluků, kde každý z těchto *k* shluků přísluší k jednomu z *k* medoidů a každý shluk příslušný k danému medoidu obsahuje ty objekty souboru, pro které je vzdálenost k tomuto medoidu ze všech vzdáleností ke *k* medoidům minimální.

Algoritmus CLARANS v prvním kroku náhodně vybere *k* medoidů. V druhém kroku vybere z těchto medoidů náhodně jednoho zástupce a taktéž náhodně vybere jeden z objektů zkoumaného souboru, který není medoidem.

Dále algoritmus zjistí, zda by záměnou těchto dvou objektů došlo ke zlepšení. To znamená, že zjistí, jestli se sníží průměrná vzdálenost objektů od „jejich“ medoidů, po přeorganizování příslušnosti objektů k medoidům, pokud by se z vybraného objektu, který není medoidem stal medoid a z vybraného medoidu se stal „běžný“ objekt. Pokud by ke zlepšení došlo, provede záměnu.

Mezi vstupní parametry algoritmu patří počet, kolikrát algoritmus tento krok provádí. Po provedení daného počtu testů a případných záměn algoritmus spočítá a uloží aktuální průměrnou vzdálenost.

Postup se opakuje od prvního kroku, od náhodného výběru *k* medoidů. Po průchodu tímto vnějším cyklem algoritmu se opět porovná, jestli došlo k vylepšení. To znamená, jestli druhému náhodnému výběru *k* medoidů nepřísluší nižší průměrná vzdálenost. Počet průchodů vnějším cyklem je omezen, a to druhým vstupním parametrem algoritmu.

Algoritmus končí po provedení předem daného počtu iteračních kroků. Tento počet je určen vstupními parametry. Doporučuje se první vstupní parametr (počet náhodných výběrů *k* medoidů – počet průchodů vnějším cyklem) nastavit na 2 a druhý vstupní parametr (počet pokusů o záměnu medoidů – počet průchodů vnitřním cyklem) nastavit na hodnotu $max\{1{,}25\% * k * (n-k);\ 250\}$. Výstupem algoritmu je množina medoidů a jim příslušných shluků.

Nejproblematičtější částí algoritmu z hlediska efektivity pro velké soubory dat je vyhodnocení zlepšení průměrné vzdálenosti. V algoritmu pro velké prostorové databáze je tento problém vyřešen tak, že je vybrán jen relativně malý počet reprezentantů ze souboru a postup algoritmu CLARANS je aplikován pouze na tyto reprezentanty. Toto je způsobem vzorkování používaného v databázových systémech, a je zde využito tzv. R*-stromů. Ze všech objektů zkoumaného souboru vytvoříme R*-strom a pomocí tohoto stromu vybereme vhodný počet reprezentantů (v závislosti na velikosti souboru a požadovaném počtu shluků). Poté pomocí algoritmu CLARANS zjistíme *k* medoidů. Nakonec všechny datové body přiřadíme k jednotlivým medoidům a tím vytvoříme požadované shluky.

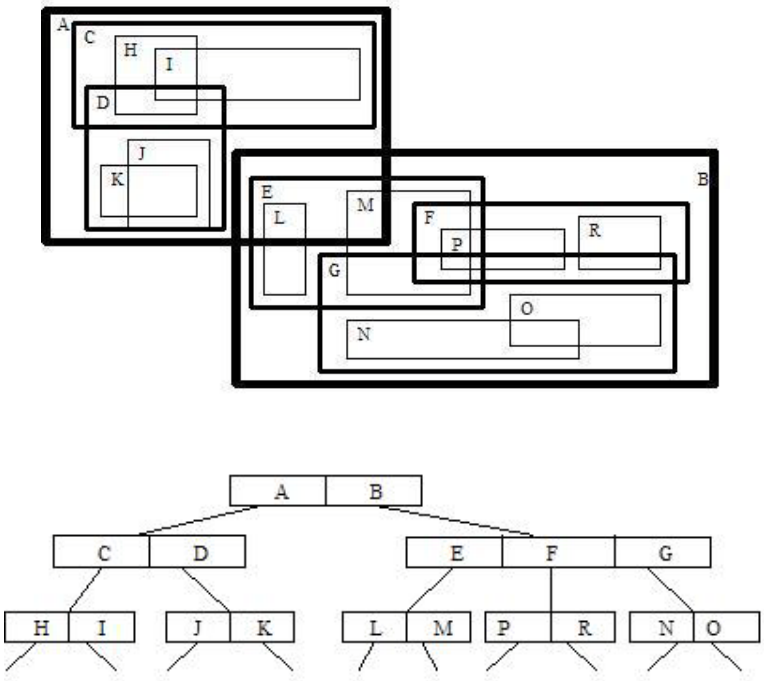
R*-stromy jsou datové struktury využívané k indexování prostorových informací. Jsou obdobou R-stromů, hlavním rozdílem je řešení situace, kdy při vkládání nového objektu přetéká nějaký uzel. V R-stromech je tento problém řešen štěpením uzlu, v R*-stromech

vyjmutím a opětovným vložením některých objektů tohoto uzlu. Satisfakcí za menší efektivitu je stabilnější struktura, která je méně závislá na pořadí v jakém jsou objekty vkládány.

R-strom [4] představuje jednoduchou modifikaci B-stromů, kde záznamy v listových uzlech stromu obsahují ukazatele k datovým objektům reprezentujícím prostorové objekty. Na rozdíl od B-stromu není striktně dodrženo pravidlo o polovičním naplnění uzlů v nejhorším případě. R-strom řádu (m_1, m) je m-ární strom, v němž minimální počet následníků libovolného vnitřního uzlu je m_1. R-strom řádu (m_1, m) je tedy m-ární strom, který má následující vlastnosti. Kořen, není-li listem, má nejméně dva bezprostřední následníky. Každý vnitřní uzel má n bezprostředních následníků, $n \in < m_1, m >$. Každý listový uzel obsahuje n indexových záznamů $n \in < m_1, m >$. Všechny cesty v R-stromu jsou stejně dlouhé.

Každý uzel R-stromu je multidimenzionálním pravoúhelníkem, jenž obsahuje všechny multidimenzionální pravoúhelníky obsažené v uzlech – potomcích. Jednotlivé pravoúhelníky se mohou překrývat.

Všechny zmíněné typy stromů jsou blíže popsány například v [8].



Obr. 1 ***Příklad R-stromu***

Mezi výhody tohoto algoritmu patří mimo jiné fakt, že jej lze použít nejen pro shlukování objektů, pro které je možno definovat průměr, ale i pro objekty, pro které je definována míra podobnosti mezi dvěmi objekty. Výhodou je také robustnost vůči odlehlým bodům, což je vlastnost všech k-medoid algoritmů. Třetí pozitivní vlastností algoritmu je jeho značná efektivnost při zpracování velkých souborů dat. Algoritmus je podrobněji popsán v [2].

Algoritmus BIRCH

Dalším z algoritmů umožňujícím shlukování ve velkých souborech dat je shlukovací metoda BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). Metoda BIRCH patří mezi k-medoids metody shlukování. Tento algoritmus shlukuje data postupně,

dynamicky tak, jak vstupují do procesu. V první fázi vytváří CF-strom, do kterého zařazuje postupně přicházející data. V druhé fázi kondenzuje vytvořený CF- strom a optimalizuje jeho velikost pomocí úpravy prahové hodnoty (jeden z parametrů CF-stromu) a pomocí patřičného přestavění stromu, zároveň umožní odstranění outlierů. Ve třetí fázi se minimalizuje dopad citlivosti na pořadí vstupních dat. Algoritmus zde shlukuje listové uzly pomocí aglomerativního hierarchického algoritmu shlukování. Ve čtvrté, nepovinné, fázi algoritmus přerozděluje referenční body k jejich nejbližším centrům a tím získává nové složení shluků. Tato fáze umožňuje odstranění jednoho z negativních důsledků výstavby CF-stromu, že fakticky jeden referenční bod vstupující formálně algoritmem ve dvou různých okamžicích se může přiřadit do dvou různých shluků. Přerozdělením se tyto dvě instance dostanou do společného shluku. Bohužel je to druhá fáze, kde je nutno opětovně procházet celý soubor, objekt po objektu. Algoritmus je blíže popsán v [9] a [10].

Struktura CF-stromů je založena na principech B-stromů a R-stromů (varianta B-stromu), tj. stromových strukturách pro indexování. B-stromem stupně *n* rozumíme strom, splňující následující podmínky. Kořen má nejméně dva potomky, pokud není listem, Každý uzel kromě kořene a listu má nejméně *n/2-1* a nejvýše *n* potomků. Navíc všechny cesty od kořene k listům jsou stejně dlouhé a data (klíče) v uzlu jsou organizována vzestupně. Každý klíč je asociován s potomkem, který je kořenem podstromu, který obsahuje klíče, které jsou menší nebo rovny, než tento klíč, ale větší než klíč předchozí.

CF-stromy využívají tzv. CF-(Clustering Feature) charakteristiku (CF-vektor) shluku. Tato charakteristika je uspořádanou trojicí $CF = (N, LS, SS)$, kde N je počet datových bodů ve shluku, LS je součtem všech datových bodů ve shluku a SS je součtem druhých mocnin těchto datových bodů.

$$(LS = \sum_{i=1}^{N} X_i , \; SS = \sum_{i=1}^{N} X_i^2)$$

Důležitou vlastnost této charakteristiky popisuje věta o CF-additivitě:

Mějme $CF_1 = (N_1, LS_1, SS_1)$ a $CF_2 = (N_2, LS_2, SS_2)$, CF-charakteristiky dvou disjunktních shluků. CF-charakteristika shluku, který vznikne spojením původních dvou shluků je rovna:

$CF = CF_1 + CF_2 = (N_1 + N_2, LS_1 + LS_2, SS_1 + SS_2)$

Údaje obsažené v CF-charakteristice jsou dostačující k výpočtu centroidů, míry vzdálenosti shluků a míry kompaktnosti shluků.

CF-stromy jsou vysoce vybalancované stromy se dvěma parametry. Prvním parametrem je faktor větvení (F, L) a druhým práh P. Pro každý vnitřní uzel CF-stromu platí, že obsahuje maximálně F vstupů ve tvaru $[CF_i, potomek_i]$, kde $i = 1, \ldots, F$, „$potomek_i$“ je ukazatel na i-tý podřízený uzel a CF_i je CF-charakteristika podshluku, který je reprezentován tímto potomkem.

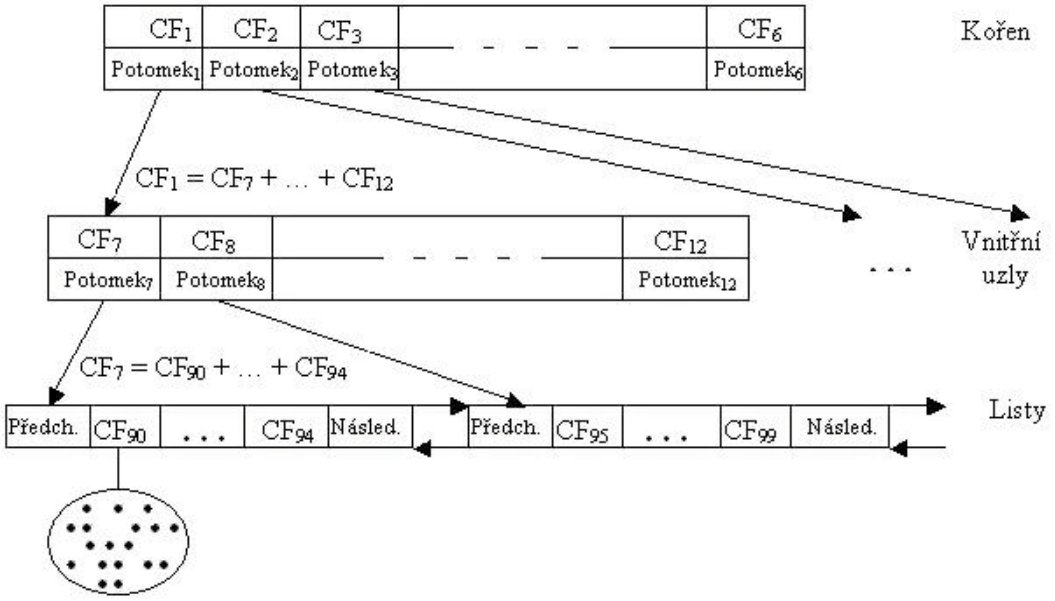
Úkolem vnitřních uzlů je umožnění nalezení správného listu pro zařazení nového datového bodu.

Každý list obsahuje nejvýše L vstupů ve formě $[CF_i]$, kde $i = 1, \ldots, L$, navíc obsahuje ukazatel na předchozí list a ukazatel na následný list, takže všechny listy jsou vzájemně propojeny pomocí ukazatelů „*předchůdce*“ a „*následník*“. Každý listový uzel reprezentuje shluk vytvořený všemi podshluky reprezentovanými jednotlivými vstupy daného listu. Pro každý vstup listu ale musí platit prahové pravidlo, že rozpětí (poloměr) vstupu je menší než práh P.

Je zřejmé, že velikost stromu je funkcí prahu P. Platí, že čím větší je hodnota prahu P, tím je menší velikost stromu a naopak.

CF-strom je budován dynamicky, postupně jsou vkládány nové datové body. Proces vkládání začíná v kořenu stromu. Nalezne se nejbližší vstup a přes něj se přejde do patřičného

uzlu-potomka. Takto se pokračuje, až se dojde do listu. V listu se opět nalezne nejbližší vstup. Pro podshluk reprezentovaný tímto vstupem ověříme platnost prahového pravidla. Pokud toto pravidlo platí, zařadíme nový datový bod do příslušného vstupu a upravíme příslušnou CF-charakteristiku. Pokud prahové pravidlo po zařazení nového objektu neplatí, musíme pro tento datový bod vytvořit nový shluk a jemu příslušný vstup. Následně se musí přepočítat všechny inkriminované CF-charakterististiky. Pokud se vytvořením nového vstupu dostaneme do konfliktu s faktorem větvení *F* (počet vstupů žádného uzlu nesmí být větší než *F*, resp. *L*), musíme rozštěpit daný uzel, opět s přepočtem příslušných CF-charakteristik. Obdobně je přebudován celý strom.



Obr. 2*Ukázka CF-stromu pro F = 7 a L = 5 (dle [11])*

Výhodou tohoto postupu je, že prochází datovým souborem pouze jedenkrát pro prvotní načtení všech objektů do CF-stromu. Reflektuje přirozenou tendenci datových bodů k seskupování a rozlišuje oblasti s velkou hustotou datových bodů a výskyt osamocených bodů, tím dobře odhaluje odlehlé body a umožňuje tyto outliery optimálně vymazat.

Nevýhodou je citlivost na pořadí vstupujících datových bodů.

Tento algoritmus je součástí algoritmu dvoukrokové shlukové analýzy ve statistickém programovém balíku SPSS (verze 11.5). Tato metoda odstraňuje výše zmíněnou nevýhodu shlukování pomocí CF-stromů.

Filtrovací algoritmus

Tzv. Filtrovací algoritmus je jednou z implementací Lloydova shlukovacího algoritmu využívající mrkd-stromy. Tento algoritmus je podrobněji popsán v [5], principy, na kterých je algoritmus postaven v [6] a [7].

Mrkd (Multiresolution kd) stromy [1] jsou speciálním, binárním, případem kd-stromů. Kd-strom (kvadrantový strom) je datová stromová struktura, která reprezentuje rekurzivní dělení konečné množiny bodů z d-dimenzionálního prostoru na k částí (d-dimenzionálních hyperkvádrů), pomocí d-1 dimenzionálních ortogonálních nadrovin. Existuje mnoho způsobů dělení, jeden jednoduchý je rozdělení ortogonálně k nejdelší straně hyperkvádru na úrovni mediánu ze všech bodů hyperkvádru. Mrkd-stromy jsou tedy binární stromy (tzn. každý z vnitřních uzlů se štěpí na dva podřízené uzly), přičemž každý z vrcholů obsahuje informaci

o všech bodech z příslušného hyperkvádru. Kořen stromu reprezentuje hyperkvádr obsahující všechny sledované body. List (koncový uzel) obsahuje jeden bod (nebo obecněji počet bodů menší než daná malá konstanta).

Lloydův algoritmus [3] je jedním z k-means shlukovacích algoritmů. Je založen na tzv. centroidech shluků (tj. vektorech, jejichž každá složka je vypočítána jako aritmetický průměr příslušných složek vektorů náležících datovým bodům patřícím do daného shluku).V inicializačním kroku Lloydova algoritmu se zvolí náhodně k center. Následně se pro každý datový bod *x* nalezne centrum, které je tomuto bodu nejbližší a datový bod se přiřadí tomuto centru (tj. přiřadí se do shluku příslušného tomuto centru). V dalším kroku algoritmu se pro každé z *k* center vypočítá centroid ze všech datových bodů přiřazených k tomuto centru a všechna centra se přesunou do příslušných centroidů. Pokud nedošlo k přesunu žádného datového bodu z jednoho shluku do jiného, nebo sice k takovémuto přesunu došlo, ale nedošlo tím k podstatnému vylepšení, algoritmus končí. V opačném případu se postup opakuje od přiřazování jednotlivých datových bodů k příslušným centrům.

Postup tzv. Filtrovacího algoritmu začíná vytvořením mrkd-stromu pro daná data. Dále se pro každý vnitřní uzel mrkd-stromu, resp. pro data asociovaná s tímto vrcholem, vypočítá centroid. Stejně, jako v Lloydově algoritmu, se určí inicializační množina k center (např. náhodným zvolením). Navíc se vytvoří množiny tzv. kandidátských center pro každý z vrcholů, a to následujícím způsobem. Množina kandidátských center pro kořen obsahuje všech *k* center. Jednotlivá kandidátská centra se budou „prosívat“ stromem dolů následovně. Pro každý uzel *u* se označí *H* - hyperkvádr náležící tomuto uzlu a *K* - množina kandidátských center náležících vrcholu *u*. Nechť $c^* \in K$, c^* je ze všech $c \in K$ nejblíže středu hyperkvádru *H*. Z množiny $K \setminus \{c^*\}$ se odeberou, „odfiltrují“, ta kandidátská centra *c*, pro která platí, že žádná část hyperkvádru *H* není blíže k *c* než k c^*, protože z předchozího vyplývá, že toto *c* není nejbližším centrem pro žádný z datových bodů patřících k tomuto uzlu. Pokud, po „odfiltrování“ všech nežádoucích kandidátských center z *K*, obsahuje *K* právě jediný prvek (který byl dříve označen c^*), pak je zřejmé, že c^* je nejbližším centrem pro všechny datové body asociované s daným uzlem a všechny tyto datové body se mohou tomuto centru přiřadit. Pokud množina *K* po „odfiltrování“ nežádoucích kandidátských center obsahuje více než jeden prvek a pokud uzel *u* není listem stromu, přejde se rekurzivně k dceřinému uzlu. Pokud je u listem, spočítá se vzdálenost všech datových bodů asociovaných s tímto vrcholem od všech zbylých kandidátských center z množiny K a přiřadí se vždy k nejbližšímu centru. Tímto jsou přiřazeny všechny datové body jednotlivým centrům a může se dokončit iterační krok dle Lloydova algoritmu – mohou se spočítat patřičné centroidy a přemístit centra do těchto centroidů. Průběh algoritmu se ukončí, stejně jako v případě Lloydova algoritmu, pokud se buď v posledním iteračním kroku „nepřesunul“ žádný datový bod od jednoho centra k druhému, nebo pokud i přes případný přesun došlo jen k nevýraznému (menšímu než předem daná konstanta) zlepšení.

Klasický iterativní k-means algoritmus není založen na matici vzdáleností a proto je využitelný i pro shlukování ve větších souborech dat. Nevýhodou je ovšem časová náročnost zpracování. Algoritmus využívající mrkd-stromy je velkým zefektivněním klasického přístupu. Je zřejmé, že mrkd-strom je zkonstruován pouze jedenkrát pro dané datové body a celá struktura nemusí být přepočítávána v každém iteračním kroku algoritmu.

Autoři algoritmu provedli a v [5] popsali mnoho srovnávacích testů a na základě podrobné analýzy došli k závěru, že algoritmus běží rychleji, čím více jsou shluky separované, ale je účinný i pokud nejsou shluky dobře oddělené. Dále ověřili srovnatelnost účinnosti s algoritmem BIRCH.

Závěr:

Výhodou využití stromů v těchto metodách je fakt, že v uzlech stromů shromáždí větší množství původních datových bodů a poté s nimi pracují vcelku, čímž se sníží objem zpracovávaných údajů. Stromovou strukturu z původních dat vytváříme většinou jen jedenkrát a pak ji pouze optimalizujeme. Tím dochází k velké redukci potřebného času na zpracování, ale i minimalizaci I/O ceny a tím redukci požadavků na HW.

Literatura:

[1] BENTLEY, J. L.: Multidimensional Binary Search Trees Used for Associative Searching. Comm. ACM, vol. 18, pp. 509-517, 1975

[2] ESTER, M., KRIEGEL, H-P., XU, X..: A database Interface for Clustering in Large Spatial Databases, Proc. of 1st Int‘1 Conf. on Knowledge Discovery and Data Mining, 1995

[3] FABER, V.: Clustering and the Continuous k-Means Algorithm. Los Alamos Science, vol. 22, pp. 138-144, 1994

[4] GUTTMAN A., STONEBRAKER M.: R-Trees: A Dynamic Index Structure for Spatial Searching, EECS Department University of California, Berkeley, Technical Report No. UCB/ERL M83/64, 1983

[5] KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., PIATKO, CH. D., SILVERMAN, R., WU, A. Y.: An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on pattern analysis and machina inteligence, vol. 24, no. 7, july 2002

[6] MOORE, A.: Very Fast EM-based Mixture Model Clustering usány Multiresolution kd-trees. Proc. Conf. Neural Information Processing Systéme, 1998

[7] PELLEG, D., MOORE, A.: Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 277-281, Aug. 1999

[8] POKORNÝ, J., Prostorové datové struktury a jejich použití k indexaci prostorových objektů, GIS Ostrava 2000 Editor J. Růžička, Pg. 146-160, Inst. ekonomiky a systémů řízení Ostrava

[9] ZHANG, T., RAMAKRISHNAN, R., LIVNY, M.: BIRCH: An Efficient Data Clustering Method for Very Large Databases. ACM SIGMOD Record, 25, 2. 1996 (s. 103-114)

[10] ZHANG, T., RAMAKRISHNAN, R., LIVNY, M.: BIRCH: A New Data Clustering Algorithms and Its Applications. Journal of Data Mining and Knowledge Discovery, vol. 1, no. 2, 1997 (s. 141-182)

[11] http://www.cs.uic.edu/~dyu/birch.ppt

Adresa

RNDr. Marta Žambochová
Fakulta sociálně ekonomická, Univerzita J.E. Purkyně, Ústí nad Labem
Katedra matematiky a statistiky
zambochova@fse.ujep.cz

PREHLIADKA PRÁC MLADÝCH ŠTATISTIKOV A DEMOGRAFOV

Value at risk: on the naïvete of the correct specification of a volatility model

Martin Boďa[†]

Abstract: In value at risk forecasting, an especial attention is given to the volatility specification. As there were, some time ago, a shadow of doubt cast upon its importance, this article briefly examines its role and juxtaposes the choice of the volatility model and the selection of the distributional assumptions.

Keywords: value at risk (VaR), exponentially weighted moving averages (EWMA), generalized auto-regressive conditional heteroskedasticity (GARCH), synthetic parametric models, historical simulation.

To T. Š.

1. Introductory notes and portfolio selection

Contemporary financial management is dominated by several convenient suppositions that are broadly accepted, which renders them facts; and one of them is the conviction that financial risk can be controlled, measured and forecasted. This establishes attempts of integrating financial data into some useful quantitative characteristic that would reflect the riskiness of positions assumed by traders and brokers at financial markets. All though many a possibility in theory and usage exists, it is value at risk (VaR) that enjoys the overall popularity in comparison to other risk measures applicable in assessing a portfolio's risk. Value at risk is an estimate of the potential loss of the market value of a *trading* portfolio that is not to be exceeded at the chosen confidence level over a pre-specified holding horizon. By definition, it is a quantile of the profit & loss distribution of a portfolio and as a measure is characterized by the parameters *holding horizon* (commonly 1 day or 10 days) and "*confidence level"* (normally 0.95 or 0.99).

Simple as the concept of value at risk may be, the aspects of its practical calculation render value at risk unduly vulnerable to the real changing environment of financial markets and erode its capability of faithful representation of risk. It is again assumptions and misspecifications that are a source of failures in modelling risk and there are doubts as to whether value at risk itself can be considered a reliable measure. On these issues the reader may consult e. g. Boďa (2006a). If we adopt the assumption that value at risk is eligible to capture risk and that from historical observations future expectations can be derived, we can take the procedure of calculation of value at risk – without prejudice to its complexity – as a four-stage process: (1.) the mapping of risk, (2.) the specification of volatility pattern, (3.) the choice and application of method, and (4.) the performance evaluation. Some basic notions on these steps are given in Boďa (2006b) and can be tracked on through the associated references.

This article selects some of possible approaches as indicated in Boďa (2006b) so as to perform a meaningful simulated computation of this measure. Even though it must be stressed that all three steps are vital for the real-life estimation of value at risk, the focus rests upon the last two steps. This choice results from the knowledge that risk mapping is more of a technical matter distinctive to a portfolio's composition than the employment of statistical techniques. As a result, the two aspects of value at risk modelling are confronted: 1. the *correct* specification of volatility model, and 2. the choice of distributional assumptions. The motivation for such an outline of the task may be seen in the study of Lopez and Walter (2000) who spotlighted the first topic and concluded that the underlying distributional assumptions are of greater importance than the volatility specification.

With this intent in mind, a portfolio of assets from the perspective of a U. S. holder was chosen to be composed of four equities and two currencies.[1] The basket of shares consisted of those of *The Coca-Cola Company* (CC), *Microsoft Corporation* (MS), *Hilton Hotels Corporation* (HH), *Bank of America*

[†] Bc. Martin Boďa. Univerzita Mateja Bela v Banskej Bystrici, Ekonomická fakulta, Tajovského 10, 975 90 Banská Bystrica. E-mail: ma_bo@azet.sk.

[1] It is quite customary to perform such studies on a market index which itself is a trustworthy representative of a real portfolio and makes the entire process of value at risk estimation substantially simpler. The substitution of a sole index for a portfolio would not allow the demonstration of the different performances of different volatility patterns.

Corporation (BA) and the foreign currency part of the portfolio was represented by *the euro* (EUR) and *the Czech koruna* (CZK).[2] The assets were allocated to the portfolio under the following weights:

CC: $w_1 = 0.042$, **MS**: $w_2 = 0.058$, **HH**: $w_3 = 0.311$, **BA**: $w_4 = 0.293$, **CZK**: $w_5 = 0.234$, **EUR**: $w_6 = 0.062$.[3]

2. Modelling issues

Both the multivariate form of volatility and the univariate alternative were examined. In the former case, the volatility was interpreted through the covariance matrix representing single asset returns movement and co-movement. In the univariate framework, the asset returns were integrated on the assumption that a single variance is sufficient to describe the dynamics of the entire portfolio. Of a variety of estimation possibilities under consideration were the static model with volatility constant over time, the model with volatility updated daily (simple & exponentially weighted averages), and GARCH models. The specifications are presented in Scheme 1. The returns of assets r_i and of the portfolio r_{II} are logarithmic and their expectations are assumed respective $Er_i = 0$ and $Er_{II} = 0$.

Model specification	UNIVARIATE MODELS	MULTIVARIATE MODELS
VOLATILITY CONSTANT	$\hat{\sigma}^2_{\Pi} = \frac{1}{M-1}\sum_{k\in in-sample} r_{\Pi k}{}^2$	$\hat{\sigma}^2_i = \frac{1}{M-1}\sum_{k\in in-sample} r_{ik}{}^2$ $\overline{\mathrm{cov}}_{ij} = \frac{1}{M-1}\sum_{k\in in-sample} r_{ik} r_{jk}$
VOLATILITY UPDATED (unweighted moving averages)	$\hat{\sigma}^2_{\Pi[t]} = \frac{1}{m-1}\sum_{k=t-1}^{k=t-m} r_{\Pi k}{}^2$	$\hat{\sigma}^2_{i[t]} = \frac{1}{m-1}\sum_{k=t-1}^{k=t-m} r_{ik}{}^2$ $\overline{\mathrm{cov}}_{ij[t]} = \frac{1}{m-1}\sum_{k=t-1}^{k=t-m} r_{ik} r_{jk}$
VOLATILITY UPDATED (exponentially weighted moving averages)[4]	$\hat{\sigma}^2_{\Pi[t]} = (1-\lambda)\cdot\sum_{k=t-1}^{k=t-m}\lambda^{t-k-1} r_{\Pi k}{}^2$ ***recursive form*** $\hat{\sigma}^2_{\Pi[t]} = \lambda\cdot\hat{\sigma}^2_{\Pi[t-1]} + (1-\lambda)\cdot r^2_{\Pi\ t-1}$	$\hat{\sigma}^2_{i[t]} \approx (1-\lambda)\cdot\sum_{k=t-1}^{k=t-m}\lambda^{t-k-1} r_{ik}{}^2$ $\overline{\mathrm{cov}}_{ij[t]} = (1-\lambda)\cdot\sum_{k=t-1}^{k=t-m}\lambda^{t-k-1} r_{ik} r_{jk}$ ***recursive forms*** $\hat{\sigma}^2_{i[t]} = \lambda\cdot\hat{\sigma}^2_{i[t-1]} + (1-\lambda)\cdot r^2_{i\ t-1}$ $\overline{\mathrm{cov}}_{ij[t]} = \lambda\cdot\overline{\mathrm{cov}}_{ij[t-1]} + (1-\lambda)\cdot r_{i\ t-1} r_{j\ t-1}$
GARCH(1, 1)	$\sigma^2_{\Pi[t]} = \upsilon + \alpha\cdot r^2_{\Pi\ t-1} + \beta\cdot\sigma^2_{\Pi[t-1]}$ (where $\upsilon > 0,\quad \alpha,\beta \geq 0,\quad \alpha+\beta < 1$)	***DVEC representation of MGARCH*** $\sigma^2_{i[t]} = \upsilon_i + \alpha_i r^2_{i\ t-1} + \beta_i \sigma^2_{i[t-1]}$ (where $\upsilon_i > 0,\quad \alpha_i,\beta_i \geq 0,\quad \alpha_i+\beta_i < 1$) $\mathrm{cov}_{ij[t]} = \upsilon_{ij} + \alpha_{ij} r_{i\ t-1} r_{j\ t-1} + \beta_{ij}\,\mathrm{cov}_{ij[t-1]}$ (where $\upsilon_{ij} > 0,\quad \alpha_{ij},\beta_{ij} \geq 0,\quad \alpha_{ij}+\beta_{ij} < 1$)
EGARCH(1, 1)	$\lg\left(\sigma^2_{\Pi[t]}\right) = \upsilon + \beta\sigma^2_{\Pi[t-1]} + \alpha\left\lvert\frac{r_{\Pi\ t-1}}{\sigma_{\Pi[t-1]}}\right\rvert + \gamma\frac{r_{\Pi\ t-1}}{\sigma_{\Pi[t-1]}}$	

Scheme 1 *Volatility models entering the analysis*

In the determination of distributional assumptions, for the sake of convenience, a simple parametric synthetic model of stochastic volatility was employed, which resulted in describing the behaviour of portfolio returns by the generic formula $r_{\Pi[t]} = \sigma_{\Pi[t]} z_{\Pi[t]}$ where $\sigma_{\Pi[t]}$ stands for the conditional or non-conditional volatility holding for a day t and $z_{\Pi[t]}$ represents the corresponding innovative component of the process generating returns. This definition implies that it only suffices to select a suitable distributional form for the innovations $z_{\Pi[t]}$ (standardized returns effectively) as the volatility component $\sigma_{\Pi[t]}$ was

[2] Of the companies, the shares of Microsoft Corporation are listed on NASDAQ and those of the rest on NYSE. It is rational to assume that the area of business of the companies in focus is clear and there is no need to engage in the description.

[3] The procedure to yield the structure of portfolio ran under this algorithm: (1.) For each asset its order i was (non-randomly) determined. The order was respected in the enumeration above. (2.) 6 realizations u_i from $U(0, 1)$ were generated and in the ascending order indexed to the assets. (3.) For the asset indexed $i = 1$ (CC) the weight was set equal to its u-number, i. e. $w_1 = u_1$. The assets indexed $i = 2, 3, 4, 5$ were assigned their weights by the formula $w_i = u_i(1 - \Sigma w_{i-1})$ and the weight of the last position (EUR) was fixed as $w_6 = (1 - \Sigma w_5)$.

[4] The λ-factor was determined in accord with the custom value of this parameter for daily returns, to wit $\lambda = 0.94$.

identified in the earlier step. Again, several options were studied and $z_{\Pi\,[t]}$ was accounted for the normal distribution, the t-distribution, the non-central t-distribution, and the generalized Pareto distribution (GPD).[5]

The forecast of value at risk made in a day t for the following trading day $t + 1$ was generated according to the formula $VaR_{t,\,\alpha}(r_\Pi) = -q_{1-\alpha}\,\sigma_{\Pi\,[t+1]}$, in which $q_{1-\alpha}$ is the appropriate α-quantile of a distribution selected for z_Π and $\sigma_{\Pi\,[t+1]}$ denotes (the forecast of) the volatility related to the day $t + 1$. It is needful to add that the holding period was set to one trading day, i. e. $\tau = 1$ (which permitted the omission of the rough and cumbersome scaling the value at risk forecasts for a longer period of time), and that the "confidence level" was chosen 0.95 in line with the established practise, i. e. $\alpha = 0.05$.

As to the forecast evaluation, Kupiec's failure rate test was made use of. Analogously, for the inter-model comparison the absolute percentage exceedance (APE) and the average number of failures (ANF) over the previous 250 trading days were applied. [6]

3. Estimation issues and results

The daily stock prices and foreign exchange rates were obtained from http://finance.yahoo.com. The time series counted 1978 observations, of which the in-sample period encompassed 1221 observations (from 4 Jan 1999 to 10 Nov 2003) and the remaining 757 data were out-of-sample (from 11 Nov 2003 to 10 Nov 2006). The interesting statistic properties of returns are displayed in Scheme 2. Notwithstanding that the data manifest themselves to meet the zero expectation, it is evident that they are far from being normal (which is frequently assumed). Both the QQ graphs and Fisher's g2 suggest that the data suffer from heavy tails; likewise, the tests of normality are rejective.

Returns	Mean	SD	Max	Min	Fisher's g1	Fisher's g2
CC	-0.0002 (0.0001)	0.0194 (0.0086)	-0.1061 (-0.0807)	0.0937 (0.0402)	-0.0152 (-1.0517)	2.9680 (12.1333)
MS	-0.0002 (0.0002)	0.0275 (0.0119)	-0.1698 (-0.1205)	0.1788 (0.0595)	-0.1291 (-1.3535)	4.3556 (17.4248)
HH	0.0000 (-0.0009)	0.0381 (0.0322)	-0.2977 (0.1371)	0.1371 (0.1550)	-1.2207 (0.2629)	9.6342 (2.8238)
BA	0.0003 (0.0006)	0.0226 (0.0090)	-0.1067 (-0.1611)	0.0824 (0.0308)	-0.0233 (-2.1090)	1.4269 (25.2779)
EUR	0.0000 (0.0002)	0.0066 (0.0058)	-0.0232 (-0.1067)	0.0226 (0.0187)	0.0489 (-0.0296)	0.5866 (0.8838)
CZK	0.0001 (0.0004)	0.0076 (0.0067)	-0.0445 (-0.0291)	0.0420 (0.0253)	0.2143 (1.3709)	2.6960 (1.3709)

Scheme 2 *Descriptive statistics of gross 1220 in-sample returns and 757 out-of-sample returns (in brackets)*

In the estimation of volatilities from the in-sample returns EViews 4.0 and an ADModel-Builder-based program of Otter Research Ltd for DVEC-GARCH modelling were helpful; however, in the majority of computations the utilization of MS ® Excel 2000 dominated. Besides the traditional option of the normal distribution, in specifying the distribution of the innovative term in GARCH and EGARCH equations the generalized error distribution (GED) and the Student's t-distribution were also experimented. This said, the innovations in MGARCH's specifications were opted for to comply with the conservative choice of the normal distribution. The constructed GARCH models possessed desirable econometric properties, with the exception of a few cases in which location $\upsilon_\bullet$-term was found on the verge of significance. As suggestive of the graphs presented in Figure 1, the models captured the dynamics of volatility after their own fashion, and yet some general patterns (at least pairwise) are traceable.

Subsequently, it was possible to carry out the procedure outlined above, i. e. to construct the standardized returns $z_{\Pi\,[\bullet]}$, to estimate the parameters of the distributions under consideration and compute the respective α-quantile, and eventually to piece the components q_α and $\sigma_{\Pi\,[\bullet]}$ together so as to form

[5] For a clarification of possible misunderstanding as to the specification, the representation of the non-central t-distribution is to be found e. g. in Goorbergh (1999), and that of the generalized Pareto distribution in Embrechts et al. (1997) or McNeil (1999).

[6] Kupiec's test bases on the idea that the frequency of exceedances of an adequate model answers to the significance level α:. In this case, statistics $LR = -2\lg[(1-\alpha)^{n-x}\,\alpha^x] + 2\lg[(1-x/n)^{n-x}\,(x/n)^x]$, where x is the number of exceedances over n forecasts, follows an asymptotical $\chi^2(1)$ distribution. The absolute percentage exceedance measures the magnitude of exceedances over a given period of n predictions and is specified, for the purpose of this article, by the formula $APE = \Sigma|\varepsilon_i|$, where ε_i denotes the amplitude of exceedances, if any, or 0, if there were none.

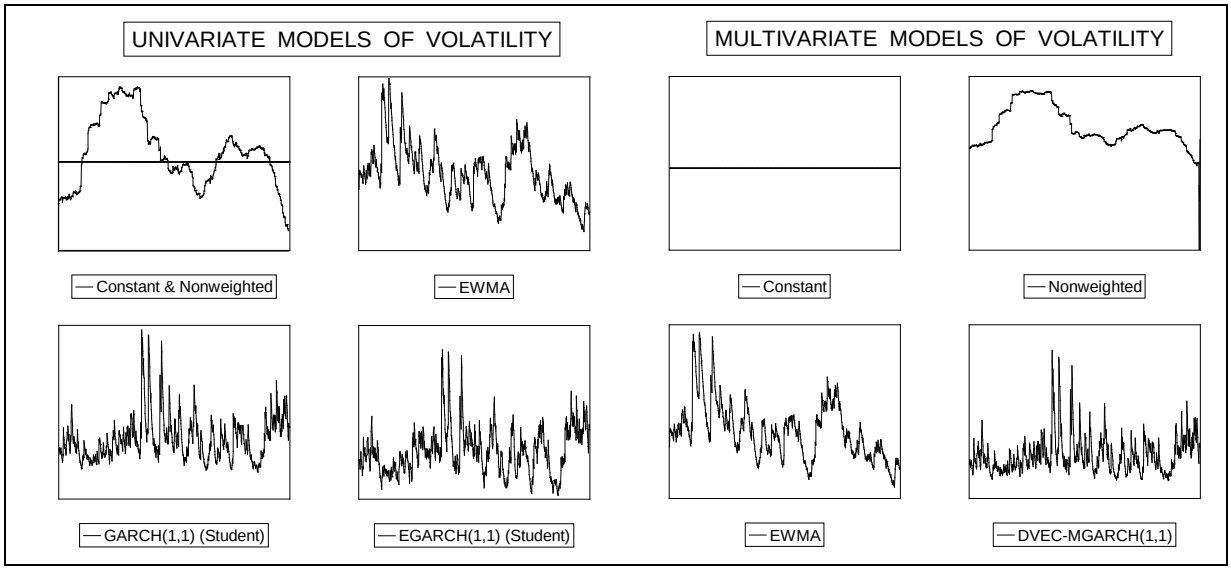


Figure 1 *In-sample volatilities estimates for (most of) the individual models*

out-of-sample estimates of value at risk. The fitting of distributions to z_{Π} was facilitated chiefly through Xtremes 3.01.[7, 8]

All in all, a collection of the 13 volatility models (of which 9 univariate and 4 multivariate) and 4 distributional forms was considered. The results for each model comprising the corresponding quantile $q_{1-\alpha}$ for the distribution of z_{Π}, the average number of failures of the value at risk forecasts over 250 trading days and the value of APE indicator are presented in Scheme 3.

Even though the diversity of the results is striking, they appear to be differentiated both by the volatility specification and by the distributional form. When inspecting the performance of the individual volatility specifications, one sees that their propensity to fail varies roughly and their predictive capability is influenced by the magnitude of the quantile. The possibly lower predictive capability of a volatility specification is generally compensated by the higher quantile number. Of the volatility models in focus, the best performance is recorded with the multivariate GARCH specification, and, furthermore and surprisingly, with the either model of constant volatility. The other models, which in fact are founded on the daily updating volatility scheme, seem to be comparatively proner to fail as to their capability to predict the loss occurred. Another ascertainment may be induced when comparing the performance of the single distributions. Save the generalized Pareto distribution, the t-distribution manifest the satisfactory eligibility to cope with heavy tails and its estimates are relatively satisfactory. As for the generalized Pareto distribution, its utility comes when losses exceed a pre-determined threshold and their occurrence is observed rare.[9] At large, the use of the non-central t-distribution was not very successful.

---

[7] The parameters (μ, σ^2) of the normal distribution were estimated in an unbiased form and the degrees of freedom ν of the Student t-distribution were picked up so that the theoretical dispersion $\nu/(\nu - 2)$ (of course, for $\nu > 2$) might correspond to the actual dispersion of z_{Π} . In some cases the t-distribution proved itself unfit to the data (when the dispersion was less than 0) or ν was very large (in that event the t-distribution was virtually identical with the normal distribution). For the other two distributions the method of maximum likelihood was employed. Specifically, the parameters (ξ, ν, β) of the generalized Pareto distribution were received from the re-signed positive standardized returns $z_{\Pi[\bullet]}$, and the computed $(1 - \alpha)$-quantile was converted back to the negative part of the number axis by adding a negative sign to it, becoming the α-quantile. The selection of thresholds was opted automatic and left to the program.

[8] The application of the maximum likelihood method requires the data be independent and identically distributed. The iid property of the series was tested by the BDS test for independence as included in EViews 4.0. The test indicated that most series are prone to evince dependence. The outcome correspondent with independence was paradoxically received in the case of constant and nonweighted volatility models (both for the univariate and multivariate specification), and with DVEC-MGARCH(1, 1) specification. Then, the standardized returns of GARCH(1, 1) with normal innovations bordered on independence. The estimated parameters in the unfavourable cases may thus be liable to produce distorted results.

[9] On no account is this meant to challenge the significance of the generalized Pareto distribution in modelling; however, one should beware that its foundations are in extreme value theory and their usage and interpretation should be adjusted accordingly.

MODEL OF VOLATILITY	$z_\Pi \sim N(\mu, \sigma^2)$			$z_\Pi \sim T(\nu)$			$z_\Pi \sim nctT(\mu, \xi, \nu)$			$z_\Pi \sim GPD(\xi, \nu, \beta)$		
	$q_{0.95}$	mean # of fails	APE	$q_{0.95}$	mean # of fails	APE	$q_{0.95}$	mean # of fails	APE	$q_{0.95}$	mean # of fails	APE
UNIVARIATE MODELS												
CONSTANT	-1.626	3.97 ± 1.87	1.559	$\nu \to +\infty$ (T → N(0, 1))			-1.633	3.97 ± 1.87	1.528	-2.701	0.35 ± 0.48	0.165
NONWEIGHTED	-1.660	13.1 ± 1.86	4.793	-1.987	7.8 ± 1.43	2.412	-1.633	13.3 ± 1.95	5.061	-2.774	1.88 ± 0.56	0.534
EWMA	-1.725	12.8 ± 2.26	4.552	-2.086	8.41 ±1.54	1.820	-1.656	14.4 ± 2.66	5.240	-2.730	1.76 ± 0.78	0.592
GARCH(1, 1) (ged)	-1.634	9.15 ± 1.24	2.459	-1.968	2.79 ± 0.79	0.931	-1.606	9.18 ± 1.25	2.691	-2.813	0.35 ± 0.48	0.168
GARCH(1, 1) (n)	-1.628	8.77 ± 1.17	2.514	$\nu \to +\infty$ (T → N(0, 1))			-1.601	9.14 ± 1.24	2.722	-2.814	0.35 ± 0.48	0.171
GARCH(1, 1) (St)	-1.629	9.19 ± 1.25	2.461	$\nu \to +\infty$ (T → N(0, 1))			-1.597	9.39 ± 1.30	2.730	-2.781	0.35 ± 0.48	0.169
EGARCH(1, 1) (ged)	-1.639	6.72 ± 1.23	2.719	-1.972	3.35 ±0.98	1.257	-1.596	9.19 ± 1.89	3.013	-2.710	0.81 ± 0.71	0.271
EGARCH(1, 1) (n)	-1.633	7.71 ± 1.35	2.721	$\nu \to +\infty$ (T → N(0, 1))			-1.589	9.71 ± 1.35	3.051	-3.304	0.35 ± 0.48	0.090
EGARCH(1, 1) (St)	-1.689	9.82 ± 1.61	3.255	-2.045	4.29 ± 1.79	1.285	-1.644	10.7 ± 1.58	3.619	-3.001	0.81 ± 0.71	0.117
MULTIVARIATE MODELS												
CONSTANT	-1.612	3.97 ± 1.87	1.559	inadequate ($D(z_\Pi) < 1$)			-1.626	3.97 ± 1.87	1.492	-2.986	0.35 ± 0.48	0.068
NONWEIGHTED	-1.761	13.7 ± 1.95	5.258	-1.772	13.2 ± 2.02	5.152	-1.684	15.5 ± 1.97	6.138	-3.252	1.22 ± 0.42	0.361
EWMA	-1.717	11.2 ± 1.41	5.791	-1.717	11.2 ± 1.41	5.791	-1.647	12.4 ± 1.46	6.702	-3.907	0.46 ± 0.50	0.012
DVEC-MGARCH(1, 1)	-2.223	0.81 ± 0.71	0.250	-2.132	1.05 ± 0.94	0.343	-2.187	0.81 ± 0.71	0.279	-4.910	0.00 ± 0.00	0.000

Scheme 3 *The results structured by the volatility model and the distributional form*

In conclusion, it must be said that there comes to be no relevant evidence to assume further that the volatility specification is a minor factor in the value at risk estimation. Contrariwise, the correctness of the volatility specification determines the quality of value at risk estimates as the distributional form cannot satisfactorily compensate for the deficiencies caused by the incorrect volatility specification. The importance of the correct volatility specification, thus, is not naïve. None the less, it must needs be accompanied by the choice of suitable distributional form.

References

ANDREEV, Andriy, KANTO, Antti 2004. *A note on calculation of CVAR for Student's distribution.* In: *Helsinki School of Economics Working Papers.* 2004, č. W369 (máj). 8 s.

BENSALAH, Younes: *Steps in Applying Extreme Value Theory to Finance: A Review.* In: *Bank of Canada Working Papers.* 2000, č. 20 (november). 22 s.

BOĎA, Martin 2006a. *Value at risk I. Value at risk ako miera rizika, alternatívy, nedostatky a regulačný aspekt.* In: *Forum Statisticum Slovacum.* 2006, č. 4, roč. 2. S. 15-24.

BOĎA, Martin 2006b. *Value at risk II. Základné prístupy k modelovaniu.* In: *Forum Statisticum Slovacum.* 2006, č. 5, roč. 2. 10 s.

CASSIDY, Colleen, GIZYCKI, Marianne 1997. *Measuring Traded Market Risk: Value-at-risk and Backtesting Techniques .* In: *Research Discussion Papers of Reserve Bank of Australia.* 1997, č. 9708. 37 s.

CIPRA, Tomáš 2002. *Kapitálová přiměřenost ve financích a solventnost v pojišťovnictví.* Praha: Ekopress 2002. 272 s. ISBN 80-86119-54-8.

EMBRECHTS, Paul et al. 1997. *Modelling extremal events for insurance and finance.* Berlin: Springer-Verlag. 648 s. ISBN 0172-4568.

ENGLE, Robert F., KRONER, Kenneth F. 1995. *Multivariate Simultaneous Generalized ARCH.* In: *Econometric Theory.* 1995, č. 11, roč. 1. S. 122-150.

GOORBERGH, Rob van der 1999. *Value-at-Risk analysis and least squares tail index estimation.* In: *Research Memorandum WO&E.* De Nederlandsche Bank. 1999, č. 578 (marec). 18 s.

GOORBERGH, Rob van der, VLAAR, Peter 1999. *Value-at-Risk analysis of stock returns: Historical simulation, variance techniques or tail index estimation?* In: *Research Memorandum WO&E.* De Nederlandsche Bank. 1999, č. 579 (marec). 37 s.

LOPEZ, Jose A., WALTER, Christian A 2000. *Evaluating Covariance Matrix Forecasts in a Value-at-Risk Framework.* In: *Working Papers in Applied Economic Theory.* Federal Reserve Bank of San Francisco. 2000, č. 2000-21. 50 s.

NCNEIL, Alexander J. 1999. *Extreme Value Theory for Risk Managers.* In: *Internal Modelling and CAD II.* Risk Books. 1999. S. 93-113.

SINHA, Tapen, CHAMÚ, Francisco 2000. *Comparing Different Methods of Calculating Value at Risk.* [Acrobat ® pdf online]. Nottingham [UK]: Nottingham University Business School 2000. [Cit. 29. 09. 2006]. Dostupné na World Wide Web: <http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID706582_code58705.pdf>.

POUŽITIE METÓDY WEBBOVHO GRAFU V ANALÝZE DYNAMIKY OBYVATEĽSTVA TRENČIANSKEHO KRAJA V ROKOCH 1999 - 2003

Zuzana Boriová

Abstract: In the following article we are discussing about the total increase of population in the Trencin´s county in the period of years 1999 – 2003. As interpretation method for typology of communities we have used the „Webb´s graph“.

Kľúčové slová: Celkový pohyb obyvateľstva, Webbov graf

Úvod a metodika

Príspevok približuje parciálnu časť širšie riešenej regionálno-demografickej analýzy Trenčianskeho kraja (Boriová 2005) s cieľom priblížiť metódu Webbovho grafu, ktorá sa používa pri analýzach demografického procesu pohybu obyvateľstva. Použitie tejto metódy ukazujeme na príklade hodnotenia okresov a obcí Trenčianskeho kraja s použitím zpriemerovaných údajov z rokov 1999 až 2003.

Veľká pozornosť sa v demografii venuje celkovej dynamike (pohybu) obyvateľstva, ktorá je výsledkom prirodzeného a migračného pohybu (resp. je výsledkom súčtu hrubej miery prirodzeného pohybu a hrubej miery migračného pohybu).

Dobrým ukazovateľom hodnotiacim celkový pohyb obyvateľstva je hrubá miera celkového prírastku, ktorá vyjadruje súčet hrubej miery prirodzeného prírastku a hrubej miery migračného salda v sledovanom časovom období.

$$Hmcp = (N – M + I – E) / S * 1000$$

kde: *hmcp – hrubá miera celkového prírastku*
N – počet narodených
M – počet zomretých
I – počet imigrantov
E – počet emigrantov
S – stredný stav obyvateľstva

Pokiaľ sú obe miery (prirodzeného a migračného pohybu) kladné, celkový pohyb obyvateľstva je tiež kladný, ak sú obe záporné, celkový pohyb obyvateľstva je záporný. V prípade, že jedna z mier je kladná a druhá záporná, tak o celkovom pohybe rozhoduje tá, ktorej absolútna hodnota je väčšia.

Vhodnou metódou na typológiu regionálnych jednotiek na základe dynamiky obyvateľstva poskytuje grafický spôsob podľa J. Webba, tzv. Webbov kríž alebo graf (Mládek 1992, Pavlík 1986). Vychádza z toho, že pohyb obyvateľstva charakterizujú dve zo štyroch teoreticky

možných zložiek: prirodzený prírastok (PP), prirodzený úbytok (PÚ), migračný prírastok (MP) a migračný úbytok (MU). Zo vzájomne možných kombinácií týchto zložiek vychádzajú štyri typy populácií s celkovým prírastkom obyvateľstva (označené ako oktanty A, B, C a D) a štyri typy populácií s celkovým úbytkom obyvateľstva (oktanty E, F, G a H). Jednoduchý výpočet a konštrukcia grafu, názornosť celej metódy prispeli k tomu, že je táto metóda často používaná. Zvolený súbor územných jednotiek sa zobrazuje na grafe pomocou karteziánskeho systému súradníc. Vertikálna os tohto systému zobrazuje bilanciu prirodzeného pohybu (v obore kladných hodnôt je to prirodzený prírastok a v obore záporných hodnôt prirodzený úbytok obyvateľstva) a horizontálna os bilanciu migračného pohybu (v obore kladných hodnôt je to migračný prírastok a v obore záporných hodnôt migračný úbytok obyvateľstva). Každá územná jednotka je v grafe zobrazená bodom, ktorého súradnice tvoria bilancie prirodzeného a migračného pohybu. Dôležité sú aj uhlopriečky grafu, ktoré spolu so súradnicovými osami rozdeľujú pole grafu na osem častí (už spomínané oktanty), označených znakmi A až H a zároveň v každej z nich je umiestnený jeden typ územných jednotiek. V sektoroch A, B, C, D sú umiestnené typy, ktoré charakterizuje celkový prírastok obyvateľstva. Odlišujú sa vzájomným vzťahom prirodzenej a migračnej zložky pohybu.

Typológia obcí kraja na základe prirodzeného a migračného pohybu

Hodnota celkového prírastku obyvateľstva v Trenčianskom kraji v sledovanom období rokov 1999 – 2003 bola záporná, - 1,57 ‰. Kraj bol v sledovanom období celkovo úbytkový, pričom Slovenská republika bola v tomto istom období prírastková, a to na úrovni 0,39 ‰. Ani jeden okresov Trenčianskeho kraja nezaznamenal celkový prírastok obyvateľstva. Najvýraznejší úbytok obyvateľstva dosiahol Myjavský okres s -6,11 ‰, najmenší úbytok okres Trenčín s -0,38 ‰.

Pri analýze Webbovho grafu zostrojeného pre obce Trenčianskeho kraja sú v *sektore A* obce, ktoré majú migračný úbytok menší ako prirodzený prírastok. V Trenčianskom kraji v skúmanom období rokov 1999 – 2003 tu pripadlo 8 obcí, čo je len 2,3 % všetkých obcí. V sektore A bolo len jedno mesto, a to Bánovce nad Bebravou. Ďalšie dve kategórie B a C sú charakteristické tým, že majú len prírastky (migračné i prirodzené). Do druhého *sektoru B* patria všetky obce, kde je prirodzený prírastok väčší ako migračný prírastok. Spadlo sem 8 obcí kraja (2,3 %) a len jedno mesto – Nemšová. V treťom *sektore C* prevažuje, opačne ako to bolo v sektore B, migračný prírastok nad prirodzeným prírastkom. V sledovanom období sa sem zaradilo 25 obcí kraja, ktoré tvorili 9,06 %. Išlo len o vidiecke obce, nenachádza sa tu žiadne mesto. Vo štvrtom *sektore D* je migračný prírastok väčší ako prirodzený úbytok. Do

tohto sektoru spadlo najviac obcí kraja a to 86, čo je takmer tretina všetkých obcí (31,2 %). Nováky sú jediným mestom patriacim do tejto skupiny. Nasledujúce sektory E, F, G a H sú charakteristické celkovým úbytkom obyvateľstva. Do *sektoru E*, v ktorom prirodzený úbytok je väčší ako migračný prírastok, spadlo 52 obcí (18,8 % podiel). Je to početnosťou obcí druhá najväčšia skupina a nachádza sa tu len jedno mesto, Bojnice. Sektory F a G majú obe zložky (migračný i prirodzený pohyb) úbytkové. V *sektore F* je prirodzený úbytok väčší ako migračný úbytok a patrí sem 43 obcí kraja, čo je 15,6 %. Nachádzajú sa tu dve mestá: Myjava a Ilava. V *sektore G* prevažuje migračný úbytok nad prirodzeným úbytkom. Spadá sem 13 % všetkých obcí Trenčianskeho kraja (36). Nájdeme tu päť miest: Handlová, Nová Dubnica, Stará Turá, Brezová pod Bradlom a Trenčianske Teplice. V poslednom *sektore H* sa nachádza 18 obcí kraja čo je 6,5 %. V sektore je väčší migračný úbytok ako prirodzený prírastok. Nachádza sa tu väčšina okresných miest (6 z 9): Trenčín, Považská Bystrica, Púchov, Prievidza, Nové Mesto nad Váhom, Partizánske, ale i mesto Dubnica nad Váhom. Mestá zaznamenali celkový úbytok obyvateľstva (-3,16 ‰). Do sektorov E až H, ktoré sú charakteristické celkovým úbytkom obyvateľstva spadlo 15 z 18 miest Trenčianskeho kraja, 8 z 9 okresných miest. Rozdelenie početnosti obcí v jednotlivých kvadrantoch obsahuje Tab. 1.

Záver

Sledovanie a analýza prirodzeného a migračného pohybu obyvateľstva je jednou z primárnych problematík riešených v demografických a demogeografických analýzach regiónov. Demografia poskytuje mnoho grafických možností na interpretáciu výsledkov celkového pohybu obyvateľstva a práve Webbov graf je častou voľbou vďaka svojej jednoduchosti zostrojenia a výbornej schopnosti interpretácie získaných výsledkov. Okrem grafického výstupu v podobe Webbovho kríža (Obr. 1) sa v praxi často vyskytuje mapový výstup, ktorý vizuálne znázorňuje priestorové rozmiestnenie javu v hodnotenom regióne.

Tab. 1 Rozdelenie obcí podľa Webbovho grafu

Oktant	Charakteristika	Počet obcí	Podiel v %
A	PP > MÚ	8	2,90
B	PP > MP	8	2,90
C	MP > PP	25	9,06
D	MP > PÚ	86	31,16
E	PÚ > MP	52	18,84
F	PÚ > MÚ	43	15,58
G	MÚ > PÚ	36	13,04
H	MÚ > PP	18	6,52
Spolu		276	100,00

Zdroj: Bilancia pohybu obyvateľstva 1999, 2000, 2001, 2002, 2003. ŠÚSR

Obr. 1 Webbov krížový graf

Použitá literatúra a štatistické zdroje:

Boriová, Z. (2005): Vývoj a rozmiestnenie obyvateľstva v Trenčianskom kraji. Bakalárska práca. Prírodovedecká fakulta UK. Bratislava

Mládek, J. (1992): Základy geografie obyvateľstva. SPN, Bratislava.

Pavlík, Z., Rychtaříková, J., Šubrtová, A. (1986): Základy demografie. Academia, Praha.

Bilancia pohybu obyvateľstva 1999, ŠÚSR

Bilancia pohybu obyvateľstva 2000, ŠÚSR

Bilancia pohybu obyvateľstva 2001, ŠÚSR

Bilancia pohybu obyvateľstva 2002, ŠÚSR

Bilancia pohybu obyvateľstva 2003, ŠÚSR

Adresa:
Zuzana Boriová
Sibírska 18/32
911 01 Trenčín
e-mail: zuzana_boriova@yahoo.com.au

Modelovanie výšky výdavkov domácností

Peter Hrubina

Abstract: The aim of this paper is to determine main factors which can moderately influence the amount of household expenditures. Thus reveal to reader a problematic of household production ability which is eventually a part of gross domestic product. The aim was not to describe household consumption structure, but to find main factors which can possibly determine its total amount. Firstly, five regressors were chosen to describe expenditure function. After statistical verification two of them were omitted. Thus the final model contains three regressors which particularly are – income, loans and natality.

Key words: Household expenditures, Income, Loans, Natality, Statistical, Economical and Econometrical verification, Model prediction ability

1. Úvod

Mojím cieľom bude určenie ekonometrického regresného modelu, ktorý bude kvantifikovať výšku výdavkov domácností(Expenses – E). Teda cieľom nebude zisťovať štruktúru výdavkov, ale zistenie niekoľkých faktorov, ktoré budú môcť popísať funkciu výšky spotreby domácností. Za východiskové regresory som zvolil 5 premenných:

1. Príjmy domácností – Income (I)
2. Sociálne dávky – Transfers (T)
3. Úvery – Loans (L)
4. Natalita – Natality (N)
5. Inflácia – Inflation (If)

Výška výdavkov domácností, inak povedané, konečná spotreba domácností je súčasťou hrubého domáceho produktu. Teda ich výška priamo ovplyvňuje produkčnú schopnosť ekonomiky a iné nadväzujúce veličiny a javy, akými sú bezpochybne aj inflácia. V súčasnosti, pol druha roka pred plánovaným zavedením spoločnej európskej meny EURO, má Slovenská republika problém práve s inflačným kritériom. Tento jav sa centrálna banka snaží regulovať zvyšovaním úrokových sadzieb, pričom predpokladá, že tieto obmedzujú prístup k úverom. Znemožnovanie prístupu k úverom spôsobí menšie výdavky, teda pomalší rast hospodárstva. Príjmy domácností sa budú znižovať, svoju budúcu spotrebu budú odkladať. T.j. očakáva sa nižší prírastok natality. Pomalší hospodársky rast si vyžiada vyššie sociálne dávky z dôvodu obmedzovania zamestnanosti. Daný ekonometrický model má nasledovnú podobu:

$$E\sim(I, T, L, N, If)$$

2. Charakteristika dát a metodika

Daný ekonometrický model budeme konštruovať pre podmienky Slovenskej republiky v kvartálnom časovom rade o dĺžke 27 období, od prvého kvartálu 2000 po tretí kvartál roku 2006. Na odhad parametrov a jeho verifikáciu som použil dostupné metódy a testy obsiahnuté v študentskej verzii ekonometrického softvéru Eviews 4.1. V nasledujúcej tabuľke uvádzam názvy, merné jednotky a zdroje potrebných dát.

Tabuľka 1: Popis premenných

Položka	**Názov**	**Merná Jednotka**	**Zdroj**
Výdavky domácností	Expenses	mil. Sk bežných cien	Konečná spotreba domácností podľa Klasifikácie individuálnej spotreby (COICOP) *Štatistický úrad SR*
Príjmy	Income	mil. Sk bežných cien	Tvorba a použitie dôchodkov v sektore domácností Bežné príjmy spolu *Štatistický úrad SR*
Sociálne dávky	Transfers	mil. Sk bežných cien	Tvorba a použitie dôchodkov v sektore domácností v mil. Sk bežných cien - Sociálne dávky okrem naturál.sociálnych transferov *Štatistický úrad SR*
Úvery	Loans	mld. Sk	Menový prehľad vo fixnom kurze
Natalita	Natality	počet obyvateľov	Prehľad pohybu obyvateľstva *Štatistický úrad SR*
Inflácia	Inflation	index	Indexy spotrebiteľských cien podľa Klasifikácie individuálnej spotreby podľa účelu (COICOP) *Štatistický úrad SR*

Prameň: Vlastné spracovanie

Predpokladané závislosti uvedených vysvetľujúcich premenných na vysvetľovanú premennú, vyjadrené prostredníctvom znamienka príslušného beta koeficientu, sú nasledovné. Pri všetkých regresoroch, okrem inflácie, predpokladáme priamo úmernú závislosť, teda kladné znamienka. T.j. ich zvýšenie spôsobí aj zvýšenie výdavkov domácností. Pri inflácií predpokladáme opačnú závislosť.

3. Konštrukcia modelu

Na základe kvantifikácie vplyvu regresorov na regresant, uvádzam prvý ekonometrický model vo forme zápisu.

$$EXPENSES_t = 121\,470{,}1 + 0{,}35*INCOME_t - 855{,}71*INFLATION_t + 302{,}23*LOANS_t +$$
$$\text{Std. Error} \quad (60\,754{,}08) \quad (0{,}06) \quad (383{,}28) \quad (100{,}19)$$
$$t_j \quad (1{,}99) \quad (5{,}54) \quad (-2{,}23) \quad (3{,}02)$$

$$+ 1{,}92*NATALITY_t + 0{,}11*TRANSFERS_t + e_t \qquad (1)$$
$$(1{,}62) \quad (0{,}96)$$
$$(1{,}18) \quad (0{,}12)$$

Je zrejmé, že daný model obsahuje pomerne veľké množstvo problémov, ktoré musia byť vyriešené, aby bol model aplikovateľný. Postupnou úpravou premenných logickými krokmi, akými sú napríklad logaritmy či prírastky sa mi nepodarilo zvýšiť štatistickú významnosť vysvetľujúcich premenných inflácie a sociálnych dávok. T.j. nemohol som zamietnuť nulovú hypotézu o nulovej hodnote príslušných koeficientov. V ďalšom ponímaní som musel od týchto premenných abstrahovať. Regresor - pôrodnosť som upravil na prírastky dvoch po sebe nasledujúcich období, teda som definoval novú premennú – Natalityprirastky.

Rozhodol som sa kvantifikovať model iným spôsobom. Do modelu som zaradil premennú úvery oneskorenú o jedno obdbobie. Vzhľadom na to, že úver musí človek najskôr obrdžať

a až potom ho môže použiť. Terajší úver je nositeľom budúcej spotreby. Premennú Income som ponechal a taktiež aj premennú prírastky natality. Výsledný model, na ktorom môžeť vykonávať verifikačné testy je nasledovný:

$$EXPENSES_t = 32\,511{,}34 + 0{,}49*INCOME_t + 4{,}90*NATALITYPRIRASTKY_t + 265{,}48*LOANS_{t-1} + e_t \quad (2)$$

Std. Error	(8 935,59)	(0,05)	(1,00)	(43,48)
t_j	(3,64)	(9,86)	(4,90)	(6,11)

4. Verifikácia modelu

Každý ekonometrický modelu musí byť zverifikovaný z troch oblastí, ktorými sú – ekonomická, štatistická a ekonometrická verifikácia.

Ekonomická verifikácia

V prípade, že sa všetky vysvetľujúce nezmenia, budú výdavky na úrovni 32,5 miliardy korún, keďže všetci musíme jesť, piť, niekde bývať atď. V prípade nárastu príjmu o jeden milión Sk, vzrastú výdavky o 0,5 milióna korún. V prípade nárastu úverov v predchádzajúcom období o jednu miliardu, vzrastú výdavky v bežnom období o 265 miliónov korún. A v prípade prírastku obyvateľstva, vzrastú výdavky o 4,9 milióna korún. Všetko za predpokladu nezmenených ostatných parametrov.

Štatistická verifikácia

Testovacie štatistiky a p-hodnoty jednotlivých parametrov sú štatisticky významné. Štatistickú významnosť modelu ako celku demoštrujem na upravenom koeficiente determinácie R^2, ktorého hodnota dosahuje 98,7%, čo je najviac zo všetkých testovaných modelov. Akaikovo informačné kritérium(19.26) a Schwarcovo kritérium(19.46) je najnižšie zo všetkých modelov. Ich relatívne veľká hodnota môže súvisieť s nie príliš veľkým počtom pozorovaní.

Ekonometrická verifikácia

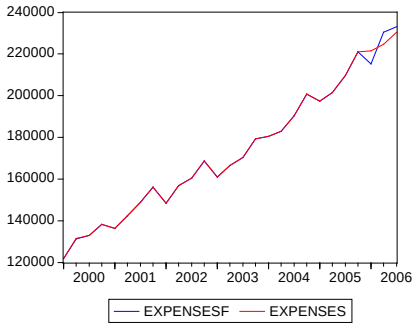
Testovali sme tri typy porušení klastických predpokladov – multikolinearita, autokorelácia a heteroskedasticita. Na základe korelačnej matice je predpoklad, že v danom modeli sa vyskytuje multikolinearita, keďže medzi výškou príjmu a úvermi je veľmi silná priama lineárna závislosť. Logicky, úver sa stáva príjem, takže danú koreláciu som očakával. Avšak vykonal som test na základe Kleinovej metodiky, ktorá porovnáva koeficient viacnásobnej determinácie s párovými koeficientami determinácie z korelačnej matice. Pričom nulová hypotéza predpokladá neexistenciu multikolinearity. Daný test potvrdil neexistenciu multikolinearity. Daný stav som sa rozhodol, vzhľadom na silnú koreláciu spomínaných premenných, overiť ešte testom Farara a Glaubera. Tento, v prvom rade porovnáva determinant korelačnej matice, ktorý ak je veľmi malý(t.j. blíži sa k nule) demonštruje možnosť multikolinearity. Následne sa vypočíta testovacia štatistika. Táto v porovnaní s tabelovanými hodnotami kvantilov pre hladinu významnosti α=0,01 a α=0,05 jednoznačne zamieta nulovú hypotézu. T.j. v danom modeli sa vyskytuje multikolinearita. Čo sa však dalo očakávať. Pretože úver sa stáva príjmom. Ale pre naše ponímanie modelu nejde o špecifikáciu momentálneho stavu, cieľom je vytvoriť model, ktorý bude mať predikčnú schopnosť, a ktorý bude teda možné použiť na predikovanie výšky spotreby domácností závislé od výšky príjmu, úverov a natality.

Heteroskedasticitu, teda závisloť reziduí na niektorom regresore, som overoval grafickými aj analytickými metódami. Nulová hypotéza hovorí o konštatných disperziách reziduí. Danú hypotézu na všetkých bežných hladinách významnosti nemôžem zamietnuť, na základe grafického znázornenia závislosti reziduí od odhadnutých výdavkov domácností a Whitovho testu Heteroskedasticity.

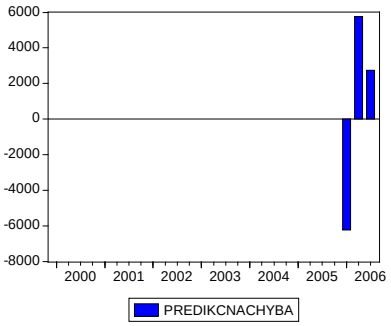
Autokorelácia, čiže porušenie predpokladu sériovej nekorelovanosti reziduí, teda $Cov(\varepsilon_i,\varepsilon_j)=0$, pre všetky $i \neq j$. Na detekciu či sa daný problém vyskytuje v mojom modeli, som použil údaj z výstupu, Durbin-Watsonova štatistika. Táto má, pri konečnom modeli, hodnotu 2,31(pričom ideálna hodnota sériovej nekorelovanosti je 2,00, pretože dané rozdelenie je symetrické okolo 2). T.j. sa jedná o sériovú nekorelovanosť vektora náhodných chýb daného ekonometrického modelu. Avšak pojednáva iba o sériovej nekorelovanosti prvého stupňa. Obšírnejší test je Breusch-Godfrey Serial Correlation LM Test, ktorého nulová hypotéza znie: Sériová nekorelovanosť reziduí, teda neexistencia autokorelácie. Na základe výstupu nemôžem zamietnuť nulovú hypotézu na všetkých bežných hladinách významnosti. T.j. jav autokorelácie sa v danom ekonometrickom modeli nevyskytuje.

5. Výsledky a diskusia

Na základe dostupných metód som dospel ku konečnému modelu determinácie výšky spotreby domácností. Tento model má vypovedaciu schopnosť. Hlavná úloha daného modelu je jeho predikčná schopnosť, ktorú uvádzam v nasledujúcich dvoch grafoch.



Graf 1: Predikčná schopnosť modelu
Prameň: Vlastné spracovanie



Graf 2: Predikčná chyba
Prameň: Vlastné spracovanie

Môžeme vidieť, že predikčná schopnosť modelu je vyhovujúca. Predikčné chyby sú v prvom kvartíli 2006 -6,2 mld. Sk(čo v porovnaní s výdavkami predstavuje odchýlku vo výške 2,81%), v druhom kvartíli 5,7 mld. Sk(2,55%) a v treťom kvartáli 2,7 mld. Sk(1,19%). Dané hodnoty môžeme považovať za uspokojivé. Avšak pre zlepšenie predikčnej schopnosti by sme mohli použiť metódu SARIMA.

6. Literatúra

GRANGER, C.W. – NEWBOLD, P. 1974. Spurious Regression in Econometrics. In: Journal of Econometrics, č. 2, 1974, s. 111 – 120.
GREEN, W. H. 1997. Econometric Analyses. Londýn: Prentice – Hall, 1997. 1076 s. ISBN 0-13-7246659-5.
KANDEROVÁ, M. 2004. Metódy prognózovania sezónnosti v ekonomických časových radoch. In: ACTA FACULTATIS AERARI PUBLICI. 2004, č. 1. Banská Bystrica : Fakulta financií UMB, 2004, s. 53 – 60.
Verejná štatisticko-demografická databáza SLOVSTAT, dostupná na www.statistics.sk, 2005.

Adresa autora:
Peter Hrubina
Ekonomická fakulta UMB
Tajovského 10
975 90 Banská Bystrica
peterhrubina@centrum.sk

Vplyv AIDS na prirodzený pohyb obyvateľstva sveta

Andrej Chromeček

Abstrakt

"In June of 1981 we saw a young gay man with the most devastating immune deficiency we had ever seen. We said, 'We don't know what this is, but we hope we don't ever see another case like it again'." (WHO, 1994)

This year marks a quarter century since the first cases of AIDS were reported. In that time, AIDS has fundamentally changed our world - killing more than 25 million men and women, orphaning millions of children, rising poverty and hunger, and, in some countries, even reversing human development altogether. Nearly 40 million people are living with HIV today.What was first reported as a few cases of a mystery illness is now a pandemic that poses among the greatest threats to global progress in the 21st century.

Úvod

Počet ľudí infikovaných HIV, vírusom, ktorý spôsobuje ochorenie AIDS stále vzrastá. Nakazený môže byť každý bez ohľadu na rasu, pohlavie, vierovyznanie, alebo sexuálnu orientáciu. Každých 8 sekúnd sa nakazí jeden ďalší človek. Aj preto je AIDS problém ktorý sa týka nás všetkých.

Definícia ochorenia

Aids je skratka pre syndróm fatálneho zlyhania imunitného systému, spôsobeného vírusom HIV. Vývoj infekcie prechádza niekoľkými fázami. V priebehu prvých niekoľko týždňov infekcie sa u človeka rozvinie akútne ochorenie podobné mononukleóze, ktoré po pár týždňoch odoznie . Vtedy sa môžu vyskytnúť rôzne príznaky ako celková únava, bolestivosť svalov a kĺbov, horúčka, vyrážky, bolesti hrdla, hnačka, stuhnutosť, zväčšenie lymfatických uzlín, demencia a nevysvetliteľný úbytok váhy. Po odoznení týchto počiatočných príznakov človek chorobu niekoľko nasledujúcich rokov vôbec neregistruje. Medzi infikovaním vírusom HIV a rozvinutím záverečného štádia AIDS, u neliečeného jedinca, existuje priemerný časový interval 10 rokov. Počas tohto obdobia môže infikovaná osoba nakaziť vírusom ďalších ľudí. V konečnom štádiu choroby AIDS človeku úplne zlyhá imunitný systém, a teda telo už nie je schopné brániť sa žiadnemu ochoreniu.

Spôsob prenosu

Vírus HIV sa prenáša prevažne krvou, spermiami a ostatnými pohlavnými sekrétmi a materským mliekom. Vírus môže byť prenesený aj z matky na dieťa počas tehotenstva cez placentu. Riziko prenosu z HIV pozitívnej matky na dieťa počas tehotenstva je pri 20-30%. Vírus sa však neprenáša pri kašli, kýchaní alebo bežnom telesnom kontakte.

región	*Najčastejší spôsob prenosu*
Svet	Heterosexuálny styk
Subsaharská Afrika	Heterosexuálny styk
Južná a JV Ázia	Heterosexuálny styk, IUD
Latinská Amerika	Homosexuálny styk, IUD, Heterosexuálny styk
Východná Európa / Stredná Ázia	IUD
Vých. Ázia a Oceánia	IUD, Homosexuálny styk, Heterosexuálny styk
Severná Amerika	Homosexuálny styk, IUD, Heterosexuálny styk
Západná Európa	Homosexuálny styk, IUD
Sev.Afrika / BlízkyVýchod	Heterosexuálny styk, IUD
Karibská oblasť	Heterosexuálny styk, Homosexuálny styk

IUD – injekčné užívanie drog

Zdroj: Population Reference Bureau staff (2004): Transitions in World Population in: Population Bulletin, vol59, no.1, p33

Typy vírusov

Sú známe dva hlavné typy vírusov HIV-1 a HIV-2. HIV-1 je viac celosvetovo rozšírený. HIV-2 sa vyskytuje prevažne v oblasti západnej Afriky. Oba vírusy pôsobia podobne, ale HIV-2 spôsobuje konečné štádium AIDS oveľa pomalšie.

Globálny pohľad

Prvýkrát bolo ochorenie AIDS zaznamenané presne pred štvrť storočím, v roku 1981, v Los Angeles v Kalifornii. Pôvodca ochorenia, vírus HIV, bol prvý krát objavený v roku 1983, dvoma na sebe nezávislými vedeckými týmami vo Francúzsku a v USA. Predpokladá sa však že vírus HIV má svoj pôvod v oblasti západnej Afriky, kde sa vyvinul z vírusu vyskytujúceho sa u divo žijúcich šimpanzov.

K explozívnemu rozšíreniu ochorenia prispelo niekoľko faktorov:

- masívny rozmach cestovného ruchu, ktorý uľahčil, šírenie mikroorganizmov do celého sveta
- uvolnenie sexuálneho života
- všeobecné sprístupnenie transfúzií a krvných derivátov do celého sveta
- zdieľanie použitých ihiel u injekčných užívateľov drog
- v Afrike a mnohých iných rozvojových krajinách spôsobil príliv populácie do miest prehĺbenie chudoby a rozklad tradičnej spoločenskej štruktúry, čo u niektorých skupín viedlo k zvýšeniu promiskuity.

Tab č.1.

región	Počet osôb žijúcich s vírusom HIV	Novoinfi- kované osoby v roku 2005	Miera prevalencie dospelých 15-49r. (%)	Počet zomretých 15r.+
Subsaharská Afrika	**24,5 mil**	**2,7 mil**	**6,1**	**2,0 mil**
Sev.Afrika a Blízky Východ	**440 000**	**64 000**	**0,2**	**37 000**
Ázia	**8,3 mil**	**930 000**	**0,4**	**600 000**
Oceánia	**78 000**	**7 200**	**0,3**	**3 400**
Latinská Amerika	**1,6 mil**	**140 000**	**0,5**	**59 000**
Karibik	**330 000**	**37 000**	**1,6**	**27 000**
Východná Európa a Stredná Ázia	**1,5 mil**	**220 000**	**0,8**	**53 000**
Severná Amerika Západná a Stredná Európa	**2,0 mil**	**65 000**	**0,5**	**30 000**
Rozvinuté krajiny	**3 mil**	**-**	**0,5**	**70 000**
Rozvojové krajiny	**35,6 mil**	**-**	**1,3**	**2,8 mil**
Najmenej rozvinuté krajiny*	**12,3 mil**	**-**	**4,3**	**1 mil**
Svet spolu:	**38,6 mil**	**4,1 mil**	**1**	**2,8 mil**

***** -50 najchudobnejších krajín v rámci rozvojových krajín

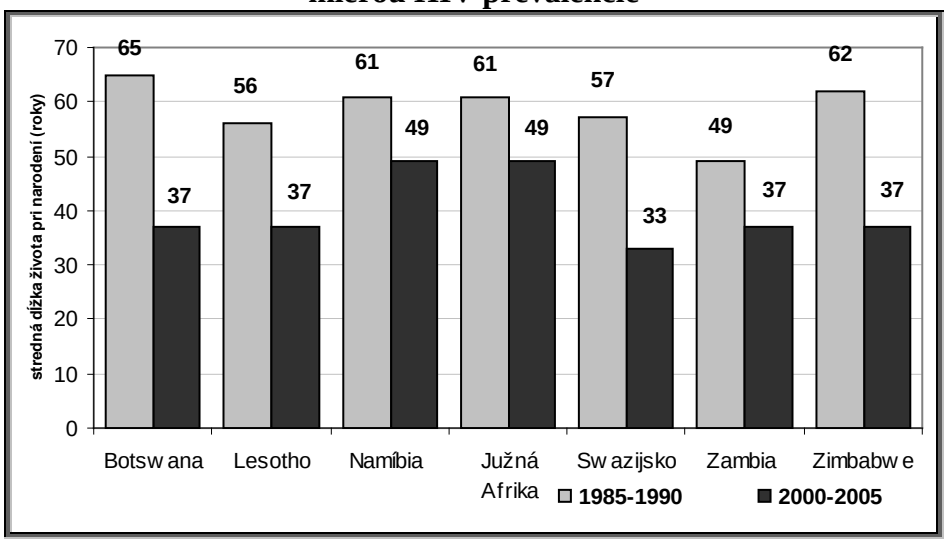
Zdroj: 2006 Report on the global AIDS epidemic, A UNAIDS 10th anniversary special edition

Podľa poslednej správy organizácie spojených národov bolo k roku 2005 na svete 38,6 miliónov ľudí infikovaných vírusom HIV. Z tohto počtu tvoria 4,1milióna len novoinfikované osoby nakazené v tomto roku. Priestorové rozšírenie epidémie AIDS je však krajne nerovnomerné. Najviac sú touto chorobou zasiahnuté rozvojové krajiny. Zatiaľ čo v rozvinutých krajinách je infikovaný len každý 200 obyvateľ, v rozvojových krajinách je to jeden človek zo sto. V päťdesiatke najmenej rozvinutých krajín je to dokonca už každý dvadsiaty obyvateľ. (Bližšie viď. tabuľka č.1)

V subsaharskej Afrike ktorá je domovom iba niečo viac ako jednej desatiny obyvateľstva sveta, sa koncentruje 64% ľudí infikovaných vírusom HIV. Ešte horšiu štatistiku má Afrika v počte nakazených detí do 15 rokov. Žijú ich tu asi 2 milióny, čo je 9/10 zo všetkých detí infikovaných vírusom HIV na svete. Práve čierny kontinent, a hlavne jeho južné oblasti, majú dlhodobo najväčší počet ochorení ako aj úmrtí na AIDS. Nie je to prekvapujúce keď si uvedomíme, že z dvadsiatky krajín s najvyšším podielom obyvateľstva infikovaného vírusom HIV, sa až devätnásť nachádza práve v Afrike (viď graf č.3)

Niekoľko najviac postihnutých afrických krajín má tak vysokú úmrtnosť obyvateľstva spôsobenú pandémiou AIDS, že v nich v súčasnosti dochádza aj napriek vysokej miere úhrnnej plodnosti k prirodzenému úbytku obyvateľstva. Sú to: Botswana s prirodzeným úbytkom v roku 2005 (-0,3%) a Lesotho (-0,1%). Treba si pritom uvedomiť že sa jedná o krajiny s prevahou mladého obyvateľstva, v ktorých je prirodzená úmrtnosť vstupujúca do bilancie prirodzeného pohybu nízka. V prakticky všetkých juhoafrických krajinách došlo za posledné desaťročie k veľmi výraznému skráteniu strednej dĺžky života pri narodení. A tak krajiny s najvyšším podielom obyvateľstva infikovaného vírusom HIV ako Botswana, Lesotho, Swazijsko, Zambia či Zimbabwe dosahujú v súčasnosti celosvetovo najnižšiu strednú dĺžku života iba od 35 do 40 rokov (viď graf č.1). Porovnateľne nízke hodnoty pritom tieto krajiny dosahovali naposledy zhruba pred pol storočím.

Graf č.1. Pokles strednej dĺžky života pri narodení v siedmych krajinách s najvyššou mierou HIV prevalencie

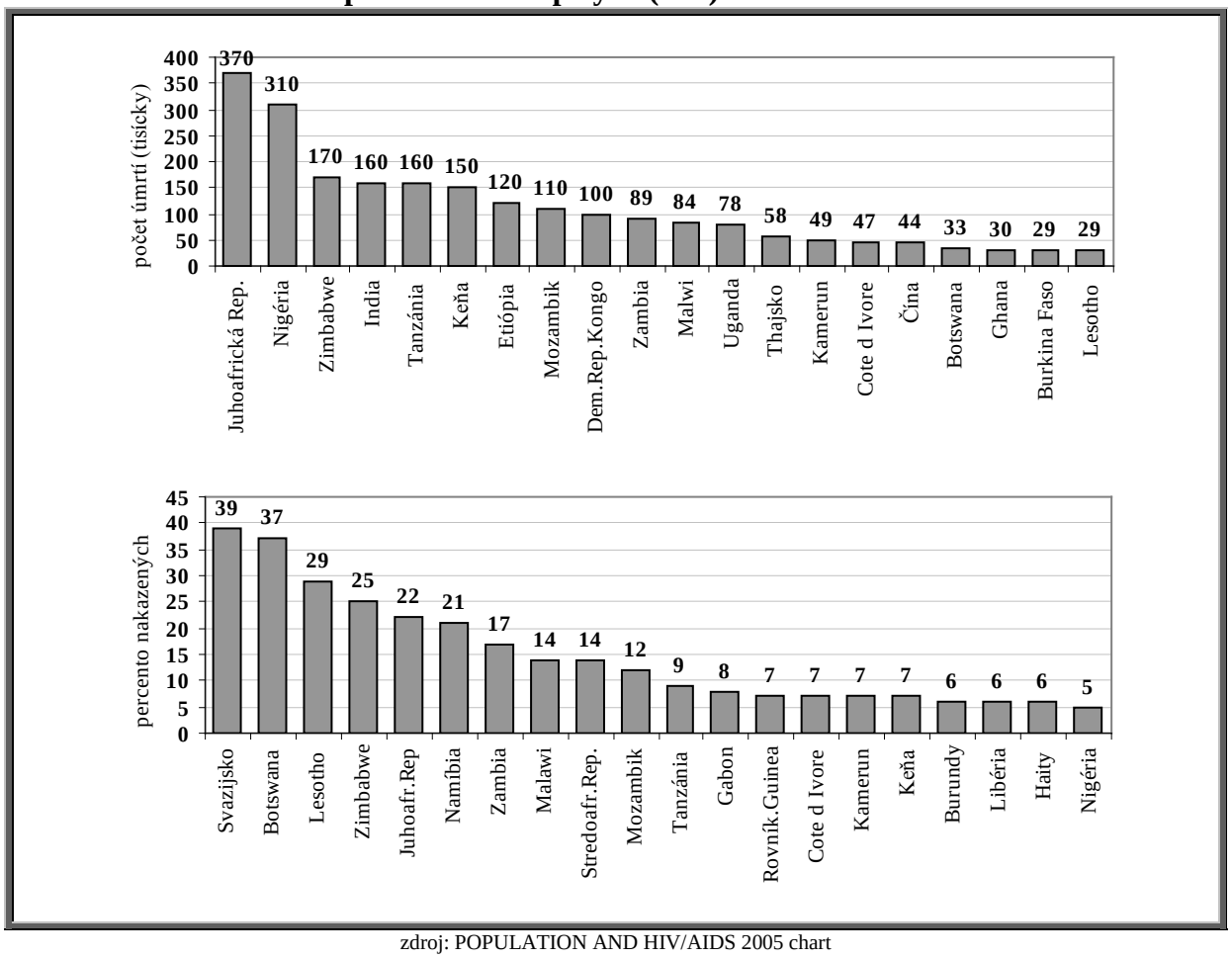


zdroj: POPULATION AND HIV/AIDS 2005 chart

Aj keď nikde vo svete nie je problém tak vypuklý a jasne viditeľný ako práve na Africkom kontinente aj v iných častiach sveta nájdeme mnoho príkladov krajín, ktoré viac alebo menej úspešne vedú boj s touto pandémiou.

V Ázii žije v súčasnosti 8,3 milióna obyvateľov infikovaných vírusom HIV. Pre viac ako 2/3 z nich je domovom jediná krajina – India. Práve v Indii len v roku 2005 zomrelo viac ako 150 000 ľudí na AIDS. Veľmi rýchlo rastie počet infikovaných aj v najľudnatejšej krajine sveta v Číne. A aj keď podielom na populácii to zatiaľ nie je znepokojivé číslo v absolútnych hodnotách je to okolo 650 000 infikovaných. Veľmi vysoký podiel nakazených majú krajiny juhovýchodnej Ázie ako Thajsko, Kambodža alebo Myanmarsko, ktoré slúžia ako obľúbené destinácie „sexuálnej turistiky“. Práve krajiny južnej a juhovýchodnej Ázie zaznamenávajú obrovský nárast nových prípadov AIDS.

Grafy č 2. a 3. **Krajiny s najvyšším počtom úmrtí a najvyššou mierou HIV prevalencie dospelých (15+) v roku 2003**



zdroj: POPULATION AND HIV/AIDS 2005 chart

Vo vyspelých krajinách Európy a Severnej Ameriky sa výskyt choroby AIDS obmedzuje prevažne na homosexuálne komunity, a injekčných užívateľov drog. Dá sa preto povedať že vírus tu pôsobí selektívne a postihuje vo zvýšenej miere tieto 2 skupiny obyvateľstva, zatiaľ čo v zbytku populácie nie je príliš rozšírený. Pozoruhodný je najmä rozmach tohto ochorenia v regióne východnej Európy. Najväčší nárast infikovaných zaznamenali v Ruskej federácii a na Ukrajine, kde sa v roku 2006 nakazilo vírusom HIV až 270 000 ľudí. Z nich takmer 1/3 je vo veku 15 -24 rokov. Šírenie HIV v týchto oblastiach je spôsobené najmä užívaním drog a používaním nesterilných injekčných striekačiek (až 63%). Epidémia HIV/AIDS na Ukrajine podľa UNAIDS narastá, pretože iba 13 percent ľudí z celkového počtu 190 000 nakazených má prístup k liekom. Štatistika uvádza, že na Ukrajine bolo do konca roka 2005 infikovaných vírusom HIV 97 000 ľudí, avšak podľa organizácie UNAIDS je toto číslo v skutočnosti oveľa vyššie a nakazených je až 377 000 ľudí. V Rusku má AIDS okolo 940 000 ľudí, pričom štyria z piatich sú vo veku 15 až 30 rokov.

Záver

Odhaduje sa že doteraz si vírus epidémia AIDS vybrala ako svoju daň za 25 miliónov obetí počas svojho 25 ročného pôsobenia. Napriek všemožnej snahe svetového spoločenstva k nim pribúdajú každý rok milióny a milióny ďalších. Najhoršie sú na tom rozvojové krajiny, ktorých vlády nemajú dostatok prostriedkov na boj s touto epidémiou. Najdôležitejšia je pritom osveta a prevencia. Krajiny ktoré sa vydali touto cestou pred pár rokmi spravidla zaznamenali značné medziročné úbytky v počte novoinfikovaných osôb.

Použité zdroje literatúry

Population Reference Bureau (2005): World Population Data Sheet 2005

United Nations , Department of Economic and Social Affairs , Population Division (2005): POPULATION AND HIV/AIDS 2005 CHART

Joint Uniteed Nations Programme on HIV – AIDS: 2006 Report on the global AIDS epidemic, A UNAIDS 10th anniversary special edition

Population Reference Bureau staff (2004): Transitions in World Population in: Population Bulletin, vol59, no.1

http://www.aids-pomoc.cz/pdf/meli_byste_byt_informovani_o_hiv_aids.pdf (7.11.2006)

autor
Bc. Andrej Chromeček
Katedra humannej geografie a demogeografie
Prirodovedecka fakulta UK
Mlynska dolina
842 15 Bratislava
email: KimiMouse@azet.sk

Vývoj pôrodnosti a plodnosti na Slovensku a v Maďarsku v období 1950-2004 – komparatívna analýza

Justína Jakúbeková

Abstract: My thesis provides overview about evolution of birthrate and fertility of population in former eastern block countries (Slovakia, Hungary) between years 1950 and 2004. You will find in my thesis description how the political-economic situation had influenced the demographic behaviour of population.

Zo všetkých demografických procesov sa najväčšia pozornosť sústreďuje na pôrodnosť, pretože táto sa najsilnejšie viaže s budúcim rozvojom spoločnosti. Je určujúcim procesom 2. demografickej revolúcie a súčasne zaznamenávame výrazne zmeny tohto procesu v poslednom období v Európe.

Vo vývoji základných ukazovateľov pôrodnosti po 2. sv. vojne môžeme vidieť medzi oboma krajinami rozdiely. Týkajú sa na jednej strane úrovne (výšky), na druhej strane intenzity zmien v priebehu posledných desaťročí. To čo spája obe krajiny, je neustály pokles hrubej miery pôrodnosti a následne aj úrovne všetkých ostatných špecifických ukazovateľov plodnosti.

Po skončení 2. sv. vojny v celej Európe doznievali kompenzačné populačné trendy, prejavujúce sa predovšetkým nárastom intenzity pôrodnosti a plodnosti (Graf 1). Najvyššiu hodnotu hrubej miery živorodenosti malo v roku 1950 Slovensko 28 ‰. Maďarsko svoje maximum hrubej miery živorodenosti dosiahlo 24% až o 4 roky neskôr v roku 1954. Od tohto obdobia pôrodnosť začala klesať a však, nie nadlho. Pokles pôrodnosti na Slovensku bol zosilnený prijatím interrupčného zákona v roku 1958.Výraznejší nárast živorodenosti mali obe krajiny v 70. rokoch. Na Slovensku to bol hlavne dôsledok uskutočnenia pronatalitných opatrení a prijatia programu na pomoc rodinám s deťmi (materské príspevky, predĺženie materskej dovolenky, poskytovanie mladomanželských pôžičiek). V Maďarsku to boli v roku 1973 navrhnuté, prijaté a zrealizované opatrenia populačnej politiky k tomu, aby povzbudili plodnosť. Na Slovensku bola hrubá miera pôrodnosti v celom období vyššia než v Maďarsku. Výrazný vplyv na úroveň pôrodnosti mala aj religiozita obyvateľstva. Na Slovensku hrubá miera pôrodnosti vystúpila z 18 ‰ v roku 1967 na 21 ‰ v roku 1974. V Maďarsku to bolo 13 ‰ roku 1962 na 18 ‰ roku 1974. Obdobie baby boomu a relatívne vysokej plodnosti 70. rokov vystriedalo od 80. rokov obdobie nepretržitého poklesu úrovne ukazovateľov pôrodnosti, ktoré bolo najintenzívnejšie najmä v prvej polovici 90. rokov.

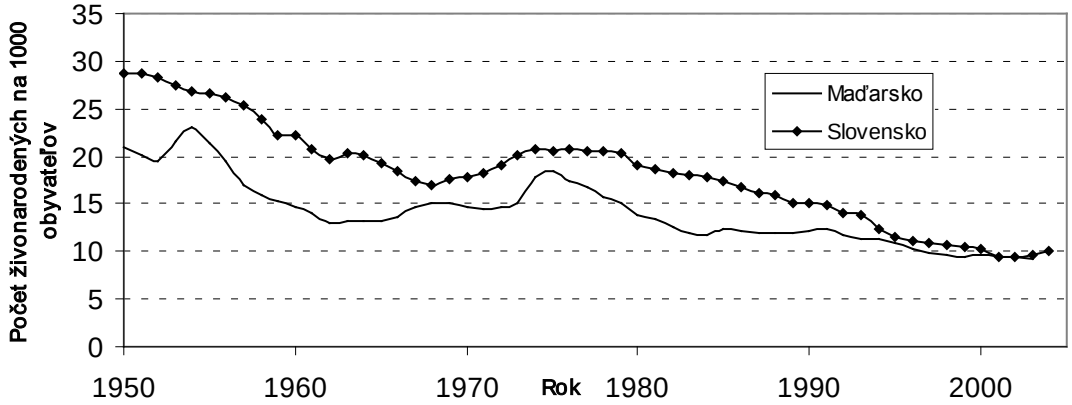
Slovensko patrilo ešte začiatkom 80. rokov ku krajinám s vysokou pôrodnosťou a plodnosťou. V tomto období sa rodilo vyše 90-tisíc detí ročne. V tom období už v Maďarsku prebiehal pokles výrazne. Najmä vplyvom politicko-ekonomických zmien, ako i celkovou zmenou populačnej klímy v 90. rokoch, sa počty živo narodených detí začal znižovať. Od roku 1996 sa v SR znížil počet živo narodených detí z hodnoty 60,1 tisíc na hodnotu 51,7 tisíc v roku 2003, čo predstavuje pokles o 14 %, a to i napriek tomu, že vo fertilnom veku sa nachádzajú ženy zo silnej pronatalitnej vlny 70. rokov. Zatiaľ čo v roku 1996 sa jednej žene počas jej reprodukčného veku priemerne narodilo 1,5 dieťaťa, do roku 2003 sa narodilo už len 1,2. Slovensko sa v súčasnosti zaraďuje k štátom s veľmi nízkou plodnosťou (druhá najnižšia hodnota úhrnnej plodnosti v Európe za Českou republikou). Nízka úroveň úhrnnej plodnosti v SR je determinovaná jednak celkovým poklesom realizovaných tehotenstiev, ako aj poklesom počtu vydatých žien v populácii. Po novembri 1989 sa politické zmeny nemohli prejaviť

zvýšením plodnosti, lebo boli súčasne sprevádzané oslabením sociálnych istôt rodín s deťmi, nezamestnanosťou a otvorením nových osobných perspektív. Rozhodnutiu mať deti konkurovala možnosť vlastnej sebarealizácie a snaha o zvyšovanie životnej úrovne.

Maďarsko bolo jednou z prvých krajín, v ktorej začala plodnosť klesať po 2 sv. vojne a kde tiež zmeny v štruktúre plodnosti začali prebiehať v 80. rokoch. Prejavili sa poklesom miery plodnosti žien mladších než 25 rokov, rastom miery plodnosti žien starších ako 25 rokov a zvýšením priemerného veku matiek pri pôrode. V prvej polovici 90. rokov nie je vidieť až taký výrazný pokles ako na Slovensku. Jedno z vysvetlení maďarských demografov je, že po roku 1990, po zmene ekonomického a politického režimu, si nová vláda zachovala väčšinu opatrení na podporu rodiny, prídavky na deti zvýšila a zaviedla niektoré ďalšie opatrenia, napr. príspevok na výchovu dieťaťa (tzv. materskú na plný úväzok). Na tento príspevok získali nárok rodičia, ktorí vychovávajú 3 a viac detí. Príspevok sa vypláca medzi 3. a 8. rokom života najmladšieho dieťaťa. Až nová vláda v rokoch 1994-1998 pristúpila k reštriktívnym opatreniam a zaviedla adresnosť v poskytovaní štátneho príspevku rodinám. Rodičovský príspevok odvodený od predchádzajúceho zárobku bol nahradený príjmovo testovanou dávkou a prídavky na deti sa tak stali príjmovo testované. Po zavedení týchto úsporných opatrení sa zrýchlil pokles úrovne plodnosti (Kocourková, 2002). V 1992 došlo v Maďarsku aj k sprísneniu interrupčného zákona, ale nemalo to žiadny bezprostredný vplyv na vývoj mier plodnosti (Poppová 2004).

Priaznivú zmenu vo vývoji pôrodnosti a plodnosti zaznamenávame na Slovensku v roku 2003, kedy sa narodilo o 872 detí viac ako v roku 2002. Nárast počtu živo narodených detí sprevádzal i mierny nárast úhrnnej plodnosti, z hodnoty 1,10 v roku 2002 na hodnotu 1,20 v roku 2003. S veľkou pravdepodobnosťou možno konštatovať, že ide najmä o realizáciu odložených pôrodov silnej generácie zo 70.rokov. Je pravdepodobné, že realizácia odložených pôrodov bude ešte určité obdobie pokračovať. V Maďarsku bol zaznamenaný v roku 2003 pokles narodených oproti roku 2002 o 2157 detí. Poklesla aj úhrnná plodnosť z 1,30 na 1,27.

Graf 1 Vývoj hrubej miery živorodenosti na Slovensku a v Maďarsku (1950-2004)

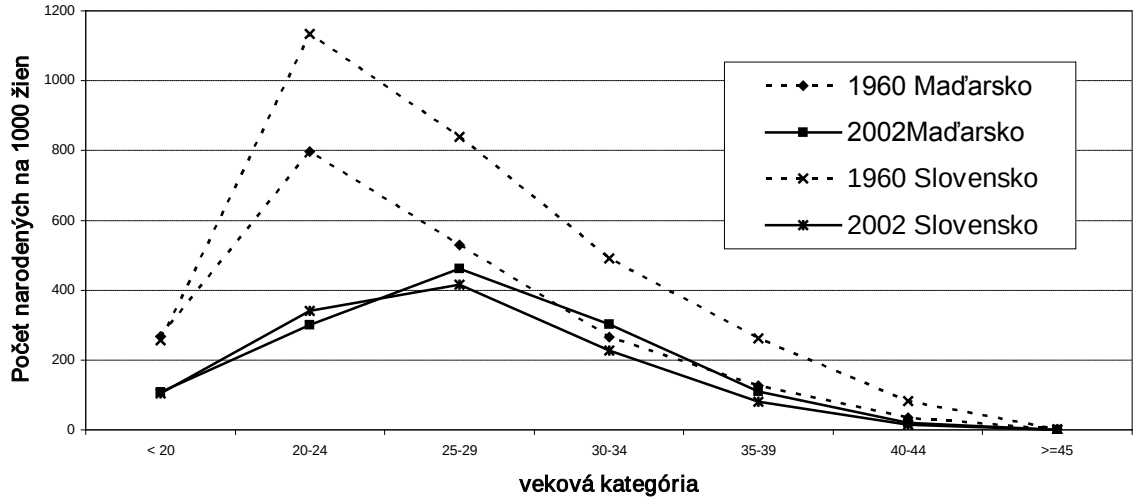


Ukazovatele ako priemerný vek ženy pri pôrode aj pri prvom pôrode sa dlhodobo veľmi nemenili ani v jednej krajine. Posun úrovne plodnosti do vyšších vekových kategórií potvrdzuje nárast priemerného veku žien pri pôrode. Na začiatku skúmaného obdobia v roku 1950 bol priemerný vek žien pri pôrode dieťaťa na Slovensku 27,6 a v Maďarsku 25,0 rokov. V 1996 bol priemerný vek žien pri pôrode v SR 25,5 roka, do konca sledovaného obdobia sa zvýšil na hodnotu 27,0 rokov. V Maďarsku dosiahol tento ukazovateľ v 1996 hodnotu 26,5 a 27,6 roka v roku 2001. Narástol i priemerný vek prvorodičiek, v priemere o dva roky. Zatiaľ

čo v SR roku 1950 rodili ženy svoje prvé dieťa priemerne vo veku 23 rokov, v súčasnosti sa táto hranica posunula k 25,0 roku života ženy. V Maďarsku sa tento vek zvýšil zhruba o 4 roky. V 1950 ženy porodili prvé dieťa vo veku 21,0 rokov, v 2001 sa zvýšil na 25,3 rokov.

Mení sa i rozloženie aj úroveň plodnosti podľa veku (Graf 2). V minulosti prevládal na Slovensku i v Maďarsku model skorej plodnosti, väčšina pôrodov sa realizovala hneď na začiatku reprodukčného obdobia ženy. V súčasnosti sa najväčšia časť realizovanej plodnosti posúva nad hranicu 25 rokov. Zatiaľ čo do roku 2000 bola na Slovensku najplodnejšou veková skupina 20 - 24 ročných žien, v roku 2001 sa vrchol maximálnej plodnosti posunul do vekovej skupiny 25 - 29 ročných žien. V Maďarsku nastala táto zmena v roku 1996. V 60. rokoch bol značný rozdiel v úrovni špecifickej plodnosti medzi oboma krajinami. Na Slovensku sú v celom sledovanom období hodnoty špecifickej plodnosti vyššie, ale aj rýchlejšie klesajú. Postupne sa však rozdiel medzi oboma krajinami zmenšoval, až sú skoro na jednej úrovni. O odklade pôrodov svedčí i nárast plodnosti vekovej kategórie 30 - 34 ročných žien, ktoré si počas celého pozorovaného obdobia zachovávajú vyššiu intenzitu plodnosti ako najmladšia veková skupina žien 15 - 19 ročných. V 90. rokoch plodnosť poklesla vo všetkých vekových skupinách.

Graf 2 Špecifická plodnosť na Slovensku a v Maďarsku (1960;2002)



Meniaci sa vzťah k materstvu a rodičovstvu sprevádza i zmena vo veľkostnej štruktúre rodín, čo sa prejavuje najmä uprednostňovaním menšieho počtu detí. Dlhodobo zaužívanú dvoj- až trojdetnú rodinu ale v poslednej dobe narastá početnosť jednodetné rodiny. Podiel detí narodených v prvom poradí mierne narastá. Druhé a tretie poradie vykazujú klesajúce tendencie. Možno predpokladať, že celkový pokles pôrodnosti a plodnosti je zapríčinený najmä poklesom narodených detí v druhom a treťom poradí. Príčiny takého stavu treba hľadať vo viacerých aspektoch. Zmenila sa hodnota detí, tie v súčasnosti predstavujú najmä emocionálny prínos pre rodičov a nie ekonomický, ako tomu bolo v minulosti (Marenčáková 2006). Väčšina žien využíva modernú antikoncepciu ako na odloženie, tak aj na ukončenie tehotenstva.

Ženy posúvajú prvé pôrody do vyššieho veku, čím sa skracuje ich reálne reprodukčné obdobie. Významnou mierou sa na poklese počtu detí v rodinách podieľajú nevyhovujúce ekonomické a bytové podmienky mladých ľudí. Reprodukčné správanie v Maďarsku je ovplyvnené spoločenskými zmenami, obzvlášť zlepšením postavenia žien.

Vývoj mimomanželskej pôrodnosti je podobný v oboch krajinách. Napriek poklesu celkovej pôrodnosti, mimomanželská pôrodnosť rastie. Pri znižujúcom sa počte narodených detí vzrástol počet narodených detí mimo manželstva. V 60. rokoch pôrodnosť mimo manželstva v oboch krajinách do roku 1975 stagnovala na 5 %. Od tohto roku začína výrazný nárast na Slovensku ako aj v Maďarsku. Na Slovensku do roku 2001 narástla táto pôrodnosť na 20 % a v Maďarsku na 30 %. Pričom v Maďarsku je od roku 1975 pôrodnosť mimo manželstva vyššia než na Slovensku a tento rozdiel sa od tohto obdobia zvyšuje.

O reprodukčných procesoch do veľkej miery vypovedá čistá miera reprodukcie, nakoľko ukazuje počet potencionálnych matiek v jednotlivých populáciách. V súčasnosti je čistá miera reprodukcie na Slovensku 0,59, čo znamená, že počas jednej generácie pri zachovaní súčasnej úrovne plodnosti a úmrtnosti žien by prirodzeným pohybom ubudlo 41% potenciálnych matiek. V 90. rokoch poklesla na nedostatočnú úroveň. V Maďarsku išlo o rozšírenú reprodukciu do roku 1957 a v období rokov 1974 - 1977. Aj keď v minulosti malo Maďarsko miery reprodukcie nižšie, od roku 1994 sú na rovnakej úrovni ako Slovensko.

Použitá literatúra

Filadelfiová, J., Guráň, P. (1997): Demografické trendy a rodina v postkomunistických krajinách Európy. BICFS, Bratislava
Lukáčová, M. (2005): Plodnosť žien vo veľmi nízkom veku. Slovenská štatistika 3-4/2005.23-31
Marenčáková, J. (2006): Reprodukčné a rodinné správanie obyvateľstva Slovenska po roku 1989 z časového a priestorového aspektu. Geografický časopis 3/2006.197-223
Poppová, Z. (2004): Regionální rozdíly ve vývoji úrovňe plodnosti v období 1988-1998 v České republice, Maďarsku a Polsku. Demografie 4, ročník 46. Český statistický úřad 2004 Praha, 264-275
Recent demographic developments in Europe 2002, Council of Europe, Strasbourg
Stav a pohyb obyvateľstva v Slovenskej republike 2004, Štatistický úrad SR 2005
United Nations 1961-2002. Demographic Yearbook 1960-2001. New York
Vaňo, B. (2005): Populačný vývoj na Slovensku po roku 1990. Demografie 2/2005.103-112
Vaňo, B. (2001): Obyvateľstvo Slovenska 1950-2000. Infostat. Bratislava 2001.
Kocourková, J.(2002): Má populační politika v České republice perspektivu?- http://cepin.cz/cze/prednaska.php?ID=495 (3.11.2006)
Changing fertility and relationship dynamics in Hungary- http://www.demogr.mpg.de/generalstructure/division2/ab-ceffd/142.htm (25.2.2006)
United Nations Statistics Division . Demogarphic Yearbook 2003- http://unstats.un.org/unsd/demographic/products/dyb/dyb2.htm (25.2.2006)
Výskumné demografické centrum. Základné demografické údaje- http://www.infostat.sk/vdc/data/databoris.htm (25.2. 2006)

Bc. Justína Jakúbeková
SNP 71/24-25
Nová Dubnica 01851

jastin@post.sk

Modelovanie bankových úverov poskytnutých obyvateľstvu v rokoch 2000-2004

Zuzana Kisková

Abstract

The aim of this contribution is to construct an econometric model of bank loans provided to inhabitants in Slovakia in the years 2000 – 2004. The research investigated the influence of average interest rate on bank loans, average disposable income of households and population employed on bank loans provided to inhabitants. Quarterly data were used. An evidence of strong dependence among changes of listed variables has been found. This resulted in construction of an econometric model explaining a significant part of the dependent variable variance.

1. Špecifikácia dát

Závislou premennou sú bankové úvery poskytnuté obyvateľstvu v rokoch 2000 až 2004. Skúmali sme ich závislosť od priemernej úrokovej miery bankových úverov (UM), priemerného disponibilného príjmu domácností vypočítaného ako rozdiel príjmov a výdavkov domácností (DP) a od počtu ekonomicky aktívnych obyvateľov, ktorí sú zamestnaní (ZAM). Použité boli štvrťročné dáta získané z verejne dostupných databáz NBS a ŠÚSR.

Keďže úroková miera je jedným z hlavných faktorov dopytu po úveroch, predpokladali sme, že bude mať na závisle premennú najväčší vplyv. Pri raste úrokovej miery sú úvery drahšie a ich objem klesá. Pokles úrokovej miery vyvoláva opačný efekt. Medzi úrokovou mierou a objemom poskytnutých úverov teda predpokladáme nepriamu závislosť.

Disponibilný príjem domácností môže nadobúdať kladné alebo záporné hodnoty v závislosti od toho, či sú vyššie príjmy alebo výdavky domácností a môže mať na objem úverov dvojaký vplyv. Na jednej strane má pokles disponibilného príjmu ako zdroja splácania potenciálnych úverov tendenciu vyvolať pokles objemu úverov (lebo banky neodsúhlasia úvery, ktoré nie sú riadne zabezpečené). Na strane druhej však vedie k rastu dopytu po úveroch (ako náhrady strateného príjmu) a má tendenciu vyvolať ich nárast. Celkový vplyv tejto premennej bude teda závisieť od toho, ktorá tendencia je prevládajúca.

Pokles počtu zamestnaných znamená zníženie počtu klientov, ktorí by mohli splácať úvery zo získanej mzdy, čo vedie k poklesu objemu úverov. Budeme teda predpokladať priamu závislosť medzi počtom zamestnaných a objemom úverov.

Prvotným problémom modelu bola multikolinearita nezávislých premenných. Preto sme z hodnôt počtu ekonomicky aktívnych obyvateľov, ktorí sú zamestnaní (ZAM) vypočítali kĺzavé ročné priemery a absolútne hodnoty takto vypočítanej veličiny sme nahradili absolútnymi zmenami (D_VZAM). Týmto sme odstránili sezónne vplyvy v časovom rade zamestnanosti. Následne vypočítané hodnoty párových koeficientov korelácie medzi jednotlivými vysvetľujúcimi premennými nadobúdajú hodnoty -0,0284 medzi DP a D_VZAM, -0,1236 medzi DP a UM a 0,3660 medzi UM a D_VZAM, a preto vylučujeme pravdepodobnosť výskytu multikolinearity.

Nezávislú premennú sme zmenili na absolútne zmeny bankových úverov (D_UV).

2. Kvantifikácia ekonometrického modelu

V prvotnom modeli sme odhadli parametre viacnásobného lineárneho regresného modelu, ktorý má nasledujúci tvar:

$$D\_UV = 16{,}3751 - 1{,}3158*UM - 0{,}0050*DP - 0{,}0383*D\_VZAM$$
$$t_{bj} \qquad (6{,}5566) \quad (-5{,}0032) \quad (-2{,}4159) \quad (-0{,}6636) \qquad (1)$$

Kde:
D_UV - je zmena bankových úverov poskytnutých obyvateľstvu za daný štvrťrok v mld. v Sk
UM - je priemerná úroková miera z úverov obchodných bánk v danom štvrťroku v %
DP - je štvrťročný priemer disponibilného príjmu domácností na mesiac a na osobu v Sk
D_VZAM - je zmena kĺzavého ročného priemeru počtu zamestnaných oproti predchádzajúcemu štvrťroku

Tento model má však viacero nedostatkov. Prvým je, že záporný koeficient nezávisle premennej D_VZAM je v rozpore s ekonomickou interpretáciou a ďalšia analýza tohto modelu preto nedáva zmysel. Napriek tomu uvediem aj ďalšie nedostatky, ktorými sú:

- Príliš vysoké hodnoty probability premennej D_VZAM, ktoré neumožňujú na bežných hladinách významnosti zamietnuť nulovú hypotézu o nevýznamnosti, a teda nulovej hodnote koeficienta tejto vysvetľujúcej premennej.
- Hodnota Durbin-Watsonovej štatistiky sa nachádza v zóne, kde nie je možné jednoznačne prijať či zamietnuť nulovú hypotézu o nezávislosti reziduí. Nemôžme teda vylúčiť autokoreláciu reziduí.

Na odstránenie nedostatkov prvotného modelu sme skúšali použiť logaritmickú transformáciu aj oneskorenie nezávislých premenných. Na základe posudzovania významnosti jednotlivých parametrov pomocou t-testov, významnosti modelu ako celku pomocou F-testu a koeficienta determinácie, sme dospeli k záveru, že najlepší model má nesledujúci tvar:

$$D\_UV = 16{,}7816 - 1{,}4278*UM - 0{,}0043*DP + 0{,}1069*D\_VZAM(-2)$$
$$t_{bj} \quad (7{,}3053) \quad (-5{,}6957) \quad (-3{,}2235) \quad (2{,}6237) \qquad (2)$$

kde:
D_UV – je zmena bankových úverov poskytnutých obyvateľstvu za daný štvrťrok v mld. Sk
UM – je priemerná úroková miera z úverov obchodných bánk v danom štvrťroku v %
DP – je štvrťročný priemer disponibilného príjmu domácností na mesiac a na osobu v Sk
D_VZAM(-2) – je zmena kĺzavého ročného priemeru počtu zamestnaných v štvrťroku t-2 oproti štvrťroku t-3, t.j. ($VZAM_{t-2} - VZAM_{t-3}$)

Na základe F-testu bola potvrdená významnosť modelu ako celku. Hodnota upraveného koeficienta determinácie hovorí, že 88,09 % celkového rozptylu tvorí rozptyl spôsobený regresiou. U všetkých použitých parametrov sa nám na bežných hladinách významnosti podarilo zamietnuť nulovú hypotézu t-testu o nevýznamnosti parametra. Bol odstránený aj problém ekonomickej verifikácie. Naďalej však zostáva hodnota DW-štatistiky v tzv. šedej zóne, a preto overíme predpoklad sériovej nekorelovanosti rezíduí aj Breusch-Godfreyovym testom.

3. Testovanie modelu

Kvalitu výsledného modelu sme overili viacerými testami. Zamerali sme sa na overenie klasických predpokladov lineárneho regresného modelu.

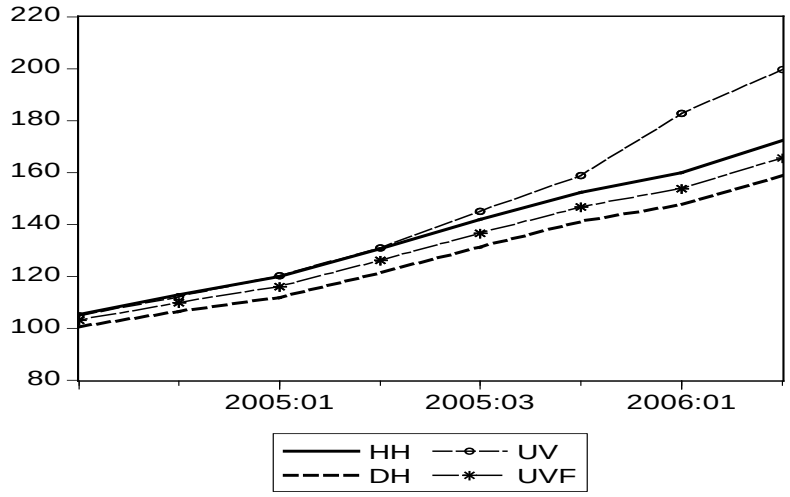
Test normality rezíduí overuje nulovú hypotézu, ktorá hovorí, že rezíduá modelu majú normálne rozdelenie. Na základe p-hodnoty 0,705 nezamietame nulovú hypotézu na bežných hladinách významnosti. Rozdelenie rezíduí je podľa hodnoty koeficienta špicatosti (1,9479) špicatejšie ako normálne rozdelenie.

Breusch-Godfreyov LM test sériovej korelácie sme použili na overenie sériovej nekorelovanosti rezíduí, keďže výsledok Durbin-Watsonovho testu jednoznačne nepotvrdil ani nevyvrátil jej existenciu. Na základe p-hodnoty 0,3553 nemôžeme na bežných hladinách významnosti zamietnuť nulovú hypotézu, ktorá hovorí o sériovej nekorelovanosti rezíduí. Nepotvrdila sa ani autokorelácia vyššieho rádu.

Whiteov test heteroskedasticity testuje nulovú hypotézu, ktorá predpokladá homoskedasticitu rezíduí. Na základe p-hodnoty (0,5258) nezamietame nulovú hypotézu na bežných hladinách významnosti.
Test Farrara a Glaubera sme použili na overenie neexistencie multikolinearity vysvetľujúcich premenných. Na bežných hladinách významnosti nezamietame nulovú hypotézu, ktorá tvrdí, že determinant korelačnej matice vysvetľujúcich premenných sa rovná jednej, a teda že premenné nie sú lineárne závislé (p-hodnota = 0,8911).
Ramseyov RESET Test je komplexnejší ako predchádzajúce testy. Používa sa na overenie nulovej hypotézy, ktorá tvrdí, že náhodné chyby modelu majú n-rozmerné normálne rozdelenie, ktorého stredná hodnota sa rovná nulovému vektoru a kovariančná matica sa rovná jednotkovej matici, ktorá má na hlavnej diagonále prvky σ^2. Podľa výsledku testu (p-hodnota = 0,9928) nemôžeme zamietnuť nulovú hypotézu.

4. Predikcia

Pre porovnanie predpovedaných (odhadovaných) a skutočných dát sme zostrojili nasledujúci graf, v ktorom sme znázornili aj interval, v ktorom by sa skutočné hodnoty mali nachádzať s pravdepodobnosťou 95%. Horná a dolná hranica tohto intervalu je vymedzená ako predpovedaná hodnota +/- dva krát smerodajná odchýlka odhadovaných dát.



Graf 1: Predikcia
Prameň: Vlastné spracovanie

Kde:
HH – je horná hranica intervalu predpovede
DH – je dolná hranica intervalu predpovede
UV – je skutočný objem bankových úverov
UVF – je predpovedaný objem bankových úverov

Z grafu vyplýva, že na základe modelu sa dá spočiatku veľmi dobre predpovedať vývoj absolútnych hodnôt úverov v budúcnosti. Je však zrejmé, že asi v druhom kvartáli roku 2005 došlo k zmene trendu vývoja a skutočné dáta rastú rýchlejším tempom ako predikované (vzďaľujú sa od hornej hranice). Potvrdil to aj Theilov koeficient nesúladu, ktorý pre prvú polovicu dát predpovede nadobúda hodnotu 3,10%, a teda predikované dáta nedosahujú významnejšie odchýlky od skutočných dát. Pre druhú polovicu dát, však dosahuje hodnotu 13,7%, čo znamená, že model stráca predikčnú schopnosť.

Na základe zistenia zlej predikčnej schopnosti modelu v dlhšom období sme na overenie jeho stability ešte použili **Chow Forecast Test** a **Chow breakpoint test.** Obidva testy slúžia na nájdenie významného zlomu, resp. štrukturálnej zmeny v závislosti

analyzovaných premenných, ale ich výsledky môžu byť rozdielne. Výsledok prvého testu identifikoval významnú zmenu od obdobia 2003:3 na základe p - hodnoty 0,0894, zatiaľ čo druhý test od obdobia 2003:1 na základe p - hodnoty 0,0571. Napriek nejednoznačnému záveru, budeme predpokladať, že v jednom z týchto dvoch období došlo k významnej zmene v závislosti sledovaných premenných. Mohlo to byť spôsobené predovšetkým zrýchlením rastu hypotekárnych úverov. Nevieme presne aký podnet to vyvolal, pretože k významnej zmene úrokovej sadzby v tomto období nedošlo. Je však možné že dlhodobý pokles úrokovej sadzby prekročil určitú „psychologickú" hranicu, kedy začali byť pre obyvateľstvo hypotekárne úvery vysoko atraktívne. Tento nárast úverov viedol postupne k zrušeniu bonifikácie hypoték, pretože ich poskytovanie by sa stalo pre štátny rozpočet neúnosné. Prvá významná zmena nastala v roku 2003, kedy sa bonifikácia znížila zo 4,5 % na 2,5 % s účinnosťou od 1. júla 2003. V prvom polroku preto došlo k náhlemu nárastu objemu hypotekárnych úverov, keďže klienti sa snažili využiť ešte platnú 4,5-percentnú bonifikáciu. Banky sa v konkurenčnom boji snažili prilákať čo najviac klientov zvýhodnením podmienok ich poskytovania, čo prispelo k ďalšiemu nárastu hypotekárnych úverov. Z uvedených dôvodov by bolo vhodné odhadnúť dva samostatné modely pre každé z týchto období – prvý pre obdobie 2000/1 až 2003/2 a druhý pre obdobie začínajúce kvartálom 2003/3, čo však pre nedostatok priestoru v tejto práci už nebudeme realizovať.

5. Záver

Výsledný model potvrdil apriórne hypotézy, že existuje priama lineárna závislosť medzi zmenou objemu úverov a zmenou počtu zamestnaných a ďalej, že existuje nepriama lineárna závislosť medzi zmenou objemu úverov a úrokovou mierou i disponibilným príjmom domácností.

Z ekonometrického modelu vyplýva, že v sledovanom období za predpokladu konštantných hodnôt všetkých vysvetľujúcich premenných, môžeme očakávať v priemere štvrťročný nárast prírastku bankových úverov o 16,782 mld. SKK. Pri konštantnej úrovni ostatných nezávislých premenných v sledovanom období vyvolá rast úrokovej miery o jeden percentuálny bod pokles prírastku bankových úverov o 1,428 mld. SKK, rast disponibilného príjmu domácností o 1 SKK vyvolá pokles prírastku bankových úverov o 4 mil. SKK a rast prírastku priemerného počtu zamestnaných o 1 tis. osôb vyvolá nárast prírastku bankových úverov o 107 mil. SKK s polročným oneskorením.

6. Literatúra

HUŠEK, R., PELIKÁN, J. 2003. Aplikovaná ekonometrie. Teorie a praxe. Praha: Proffesional Publishing, 2003. ISBN 80-86419-29-0.
HUŠEK, R.1999. Ekonometrická analýza. Praha: Ekopress, 1999. ISBN 80-86119-19-X.
[b.a.] 2003. Bonifikácia hypoték sa má znížiť na 1 percento In: Trend, 8/2003.
[b.a.] 2003. Zmena systému štátnej podpory zvýšila apetít po hypoúveroch In: Trend, 8/2003.
[b.a.] 2003. Od júla sa zmenila štátna podpora hypoúverov In: Trend, 7/2003.

Adresa autora:
Zuzana Kisková
Ekonomická Fakulta UMB
Tajovského 10
975 90 Banská Bystrica
zuzanakiskova@gmail.com

Predikcia vývoja počtu hydiny v SR pomocou neurónovej siete

Simona Kišková

Abstrakt
Základným cieľom príspevku je analýza časového radu počtu hydiny v SR pomocou neurónovej siete.

Kľúčové slová: hydina, neurónová sieť

Úvod

Hydinárske odvetvie v SR patrí medzi stabilizujúce prvky poľnohospodárstva. Jeho zastúpenie v celkovej spotrebe mäsa na obyvateľa v SR má vzrastajúci trend. V súčasnej dobe sa spotreba pohybuje na úrovni 25 %. Na vzraste spotreby hydinového mäsa sa podieľa nielen prístupnosť v cenových reláciách, kde hydina zastáva funkciu sociálneho mäsa, ale aj jeho mnohostranné využitie v príprave zdravých a dietetických jedál. Hydinové mäso svojím zastúpením vo svete je jednoznačne svetovou komoditou, nakoľko v najrozvinutejších štátoch sveta stúpa požiadavka na biele mäso a prudko klesá dopyt po červenom mäse.

Do odvetvia hydinárstva v roku 2005 zasiahol výskyt vtáčej chrípky, ktorá sa rozšírila z Ázie a Ruska do Západnej Európy a Afriky. Zmeny v dopyte po hydine sa z tohto dôvodu v jednotlivých krajinách prejavili rôzne. Možno však konštatovať, že napriek početným vypuknutiam vtáčej chrípky produkcia, spotreba a obchodovanie zostali u väčšiny významných producentov na svojich predchádzajúcich pozíciách a pokračujú v raste.

V monitorovaní chorôb hydiny bola pozornosť zameraná na najzávažnejšie ochorenia hydiny, ktoré priamo ohrozujú zdravie nielen samotnej hydiny, ale aj ľudskej populácie. Dôraz bol kladený na dve závažné ochorenia spôsobujúce zdravotné problémy a nemalé finančné straty u samotnej hydiny a u ľudí a to monitorovanie salmonelóz a aviárnej influenzy.

Po vstupe Slovenskej republiky do Európskej únie sa zmenili podmienky obchodu s poľnohospodárskymi a potravinárskymi výrobkami. Zahraničným obchodom sa stal len obchod mimo spoločného územia EÚ a výrazne sa zmenila štruktúra slovenského obchodu s poľnohospodárskymi a potravinárskymi výrobkami.

Z hľadiska výrobkovej skladby nášho vývozu do krajín Európskej únie nastala radikálna zmena. Vstup SR do EÚ spôsobil, že naši producenti a vývozcovia sa na jednotnom trhu presadzujú nie len tradičnými poľnohospodárskymi výrobkami, ale vo väčšej miere aj so spracovanými, potravinárskymi výrobkami s vyššou pridanou hodnotou.

Od vstupu do EÚ má Slovenská republika možnosť využívať celý rad koncesií, poskytovaných tretími krajinami na základe asociačných a iných typov dohôd o zónach voľného obchodu. Väčšina preferenčných dohôd sa neustále vyvíja. Vo väčšine prípadov boli upravené a zohľadňujú sa aj predchádzajúce obchodné toky a predchádzajúce preferencie nových členských štátov.

Na zmeny vo výrobkovej skladbe malo vplyv odbúranie všetkých bariér v obchodovaní na spoločnom trhu EÚ, čo umožnilo slovenským vývozcom vo väčšej miere presadzovať sa nielen na tradičných trhoch strednej Európy, ale aj na náročnejších trhoch rozvinutých „starých“ členských krajín Európskej únie, a naopak širší prienik výrobkov pôvodom z krajín EÚ – 25 na domáci slovenský trh.

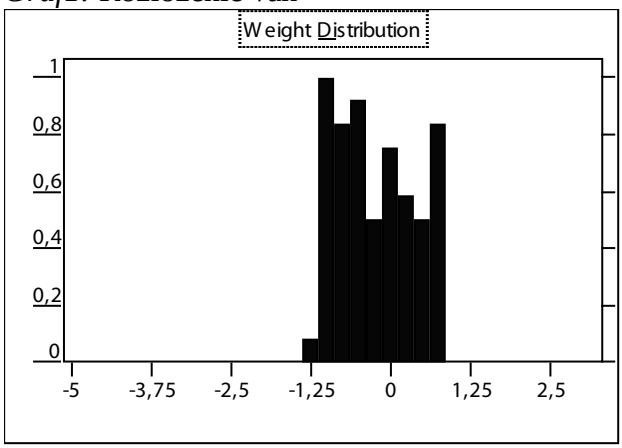
Na zmenách vo vývoji výrobkovej skladby v obchode s tretími krajinami sa pri vývozoch podieľala predovšetkým zmena v štruktúre poskytovaných vývozných náhrad.

Po vstupe do EÚ Slovenská republika prevzala do svojho právneho poriadku všetky opatrenia, resp. nariadenia Rady alebo Komisie tak na úseku spoločnej organizácie trhu a vnútorného trhu spoločenstva, ako aj na úseku obchodných mechanizmov uplatňovaných v rámci spoločnej obchodnej politiky vo vzťahu k tretím krajinám.

Výsledky a diskusia

Pre analýzu údajov pomocou neurónových sietí sme použili program – System STATISTICA Neural Networks. Ako najvhodnejšia sa ukázala neurónová sieť MLP s jednou skrytou vrstvou a ôsmimi neurónami. Má veľmi dobrú výkonnosť. Jej podrobné charakteristiky sú v nasledujúcich tabuľkách.

Graf1: Rozloženie váh



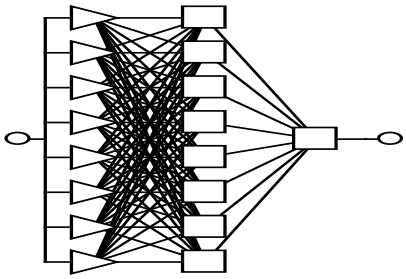
Zdroj údajov: Vlastné výpočty

Tabuľka1: Hodnoty chyby najmenších štvorcov pre jednotlivé skupiny údajov

Charakteristika	trénovacie údaje	verifikačné údaje	testovacie údaje
RMS Error	1,29E+06	7,30E+05	9,17E+05

Zdroj údajov: Vlastné výpočty

Graf2: Neurónová sieť MLP s ôsmimi skrytými neurónmi v skrytej vrstve.



Zdroj údajov: Vlastné výpočty

Tabuľka 2: Charakteristiky regresie

Charakteristika	trénovacie údaje	verifikačné údaje	testovacie údaje
Data Mean	1,48E+07	1,46E+07	1,55E+07
Data S.D.	1888448	1209516	1349929
Error Mean	480936,5	159763	-420277,8
Error S.D.	1263592	755431,6	871484,8
Abs E. Mean	1085663	518491,6	897237,7
S.D. Ratio	0,6691169	0,6245734	0,6455784
Correlation	0,9234123	0,7813892	0,7649731

Zdroj údajov: Vlastné výpočty

Tabuľka 3: Skutočné hodnoty

Rok	**Skutočnosť**	**Odhad**	**Chyba**
1979	1,58E+07	1,64E+07	0,1300905
1980	1,58E+07	1,58E+07	0,001978
1981	1,59E+07	1,60E+07	0,01391
1982	1,63E+07	1,71E+07	0,1836736
1983	1,62E+07	1,68E+07	0,1368806
1984	1,60E+07	1,66E+07	0,1399484
1985	1,59E+07	1,64E+07	0,1055473
1986	1,55E+07	1,66E+07	0,2366638
1987	1,57E+07	1,63E+07	0,1435466
1988	1,60E+07	1,64E+07	0,08548
1989	1,57E+07	1,66E+07	0,2063329
1990	1,54E+07	1,65E+07	0,2445035
1991	1,55E+07	1,39E+07	0,3541032
1992	1,52E+07	1,33E+07	0,4216851
1993	1,46E+07	1,22E+07	0,522797
1994	1,41E+07	1,43E+07	0,02902
1995	1,42E+07	1,34E+07	0,1743197
1996	1,38E+07	1,42E+07	0,0672469
1997	1,33E+07	1,42E+07	0,2117958

1998	1,35E+07	1,31E+07	0,08643
1999	1,43E+07	1,23E+07	0,4528995
2000	1,47E+07	1,36E+07	0,2514901
2001	1,45E+07	1,56E+07	0,242278
2002	1,47E+07	1,40E+07	0,1566003
2003	1,50E+07	1,42E+07	0,1676889
2004	1,45E+07	1,37E+07	0,1670369
2005	1,43E+07	1,41E+07	0,0537642

Zdroj údajov: Infostat a vlastné výpočty

Neurónová sieť umožňuje predikciu hodnôt. V rokoch 1979 – 2005 bola najpresnejšie aproximovaná hodnota pre rok 1980 s chybou 0,0019 a najmenej presne pre rok 1993 s chybou 0,5227.

Literatúra

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford: University Press.
Carling, A. (1992). *Introducing Neural Networks*. Wilmslow, UK: Sigma Press.
Fausett, L. (1994). *Fundamentals of Neural Networks*. New York: Prentice Hall.

Kontaktná adresa

Simona Kišková, študentka, FEM SPU v Nitre, Tr. A. Hlinku 2, 949 76 Nitra

Analýza vývoja pôrodnosti, sobášnosti a rozvodovosti v SR od roku 1946

Dušan Leitner[1]

Abstract: In this project I have studied how the natality, marriage rate and the divorce rate evolve since the Second World War. We have then also asked the question, weather governmental decisions in any way have influenced the evolution of the born ratio. We have found out that this was true.

Key words: natality, divorce rate, marriage rate, Linear (Holt) Exponential Smoothing, Damped Trend Exponential Smoothing, Pearson correlation coefficient, nonparametric ANOVA, System SAS V9.1, Time Series Forecasting System

1. Úvod

Cieľom projektu bolo analyzovať vývoj demografických ukazovateľov pôrodnosti, sobášnosti a rozvodovosti v Slovenskej republike a ich prípadné vzťahy. Ukazovatele boli vyjadrené ako počet danej demografickej udalosti na 1000 obyvateľov priemerného stavu obyvateľstva SR v danom roku. Analyzované časové rady, z ktorých sme vychádzali pri modelovaní vývoja, zachytávali obdobie od roku 1946 až do roku 2002.

Formulácia otázok:

- Aká je predpoveď ukazovateľov do budúcnosti?
- Vplývajú historické udalosti na vývoj pôrodnosti v SR?
- Existuje vzťah medzi pôrodnosťou a sobášnosťou na Slovensku? Je vývoj ukazovateľov na Slovensku a v Čechách porovnateľný?

Tabuľka 1: Výsledné modely časových radov

Premenná	Model	Počet meraní	Krajina
SOBÁŠE	Linear (Holt) Exponential Smoothing	57	SR
ROZVODY	Linear (Holt) Exponential Smoothing	57	SR
NARODENÍ	Damped Trend Exponential Smoothing	57	SR

Na odhadnutie parametrov modelov vývoja daných demografických ukazovateľov sme použili modul *Time Series Forecasting System* zo Systému SAS 9.1. Využili sme umelú inteligenciu systému a použili sme funkciu *fit model automatically*.

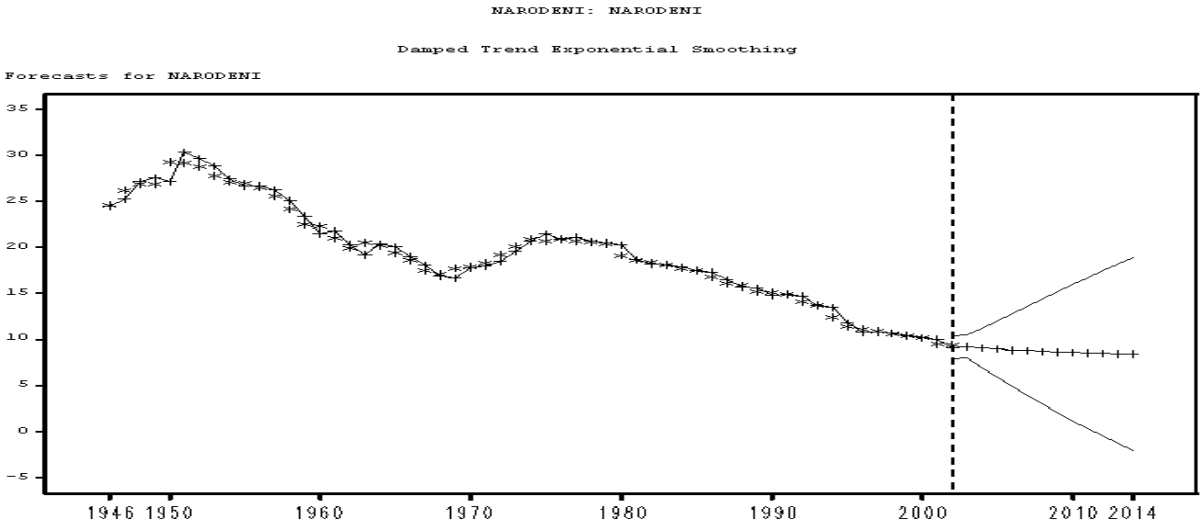
Všetky tri výsledné modely patria do skupiny modelov exponenciálneho vyrovnávania, čiže medzi adaptabilné modely.

3. Aká je predpoveď ukazovateľov do budúcnosti?

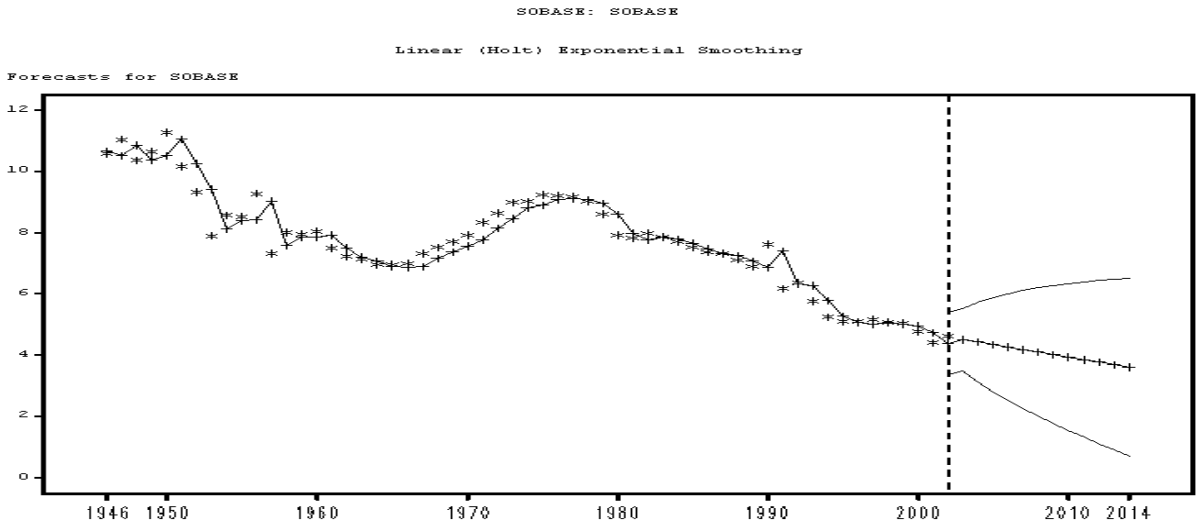
Hodnoty predpovedí premenných sa dajú zistiť aj z predchádzajúcich grafov, ale pre väčšiu zrozumiteľnosť uvádzame číselné hodnoty do roku 2006. Dodatočne sme získali i skutočné hodnoty ukazovateľov z Výskumného demografického centra v Bratislave[2] za roky 2003 a 2004. Z porovnania modelov ukazovateľov a ich skutočných hodnôt nám vyplynulo,

[1] Dušan Leitner[1], študent 3. ročníka Fakulty managementu Univerzity Komenského v Bratislave
[2] http://www.infostat.sk/vdc/

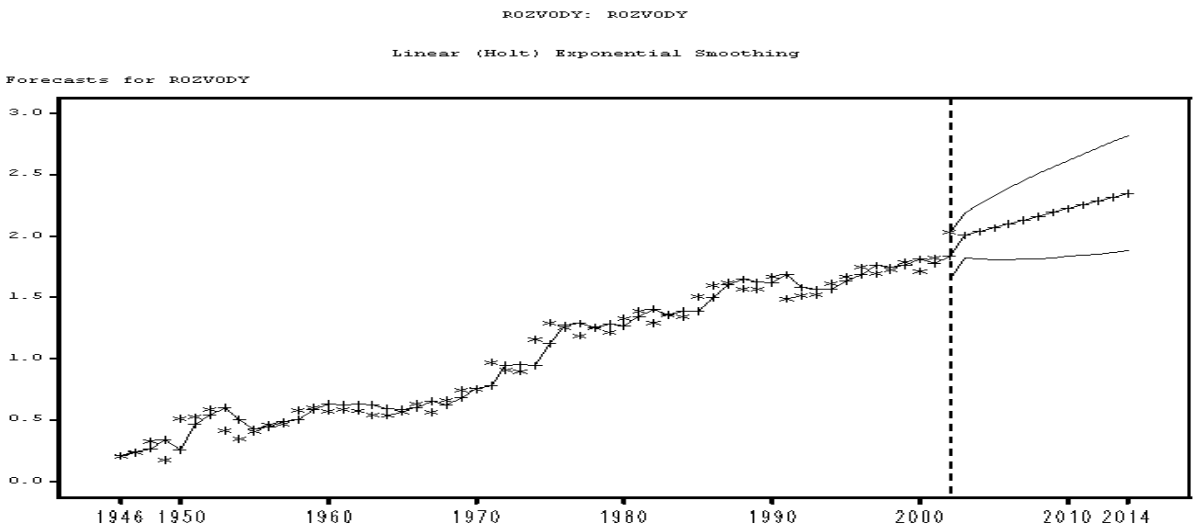
že namodelované trendy dobre kopírujú vývoj a od skutočnosti sa iba málo líšia v krátkom časovom období.



Graf 1: Vývoj počtu narodených na 1000 obyvateľov



Graf 2: Vývoj počtu sobášov na 1000 obyvateľov



Graf 3: Vývoj počtu rozvodov na 1000 obyvateľov

Tabuľka 2: Predpoveď premenných (2003 – 2006)

Year	Predicted Value	Upper 95% Confidence Limit	Lower 95% Confidence Limit	Country	Variable
2003	9.303	10.569	8.036	SR	Narodení
2004	9.148	11.305	6.992	SR	Narodení
2005	9.019	12.067	5.970	SR	Narodení
2006	**8.910**	**12.852**	**4.967**	**SR**	**Narodení**
2003	2.010	2.195	1.825	SR	Rozvody
2004	2.041	2.267	1.815	SR	Rozvody
2005	2.072	2.333	1.811	SR	Rozvody
2006	**2.103**	**2.394**	**1.812**	**SR**	**Rozvody**
2003	4.528	5.540	3.516	SR	Sobáše
2004	4.446	5.747	3.144	SR	Sobáše
2005	4.363	5.902	2.825	SR	Sobáše
2006	**4.281**	**6.024**	**2.537**	**SR**	**Sobáše**

Tabuľka 3: Skutočné hodnoty premenných za roky 2003 a 2004

Year	Real value	Predicted value	Difference	Variable
2003	9.665	9.303	0.363	Narodení
2004	10.028	9.148	0.880	Narodení
2003	2.007	2.010	-0.002	Rozvody
2004	2.043	2.041	0.002	Rozvody
2003	4.833	4.528	0.305	Sobáše
2004	5.200	4.446	0.754	Sobáše

4. Vplývajú historické udalosti na vývoj pôrodnosti v SR?

Ďalším predmetom našej analýzy bol vplyv historických udalostí (etáp) na pôrodnosť v SR. Pri rozdelení súboru na 4 etapy sme vychádzali z materiálu, v ktorom sa píše:[3] "Koniec

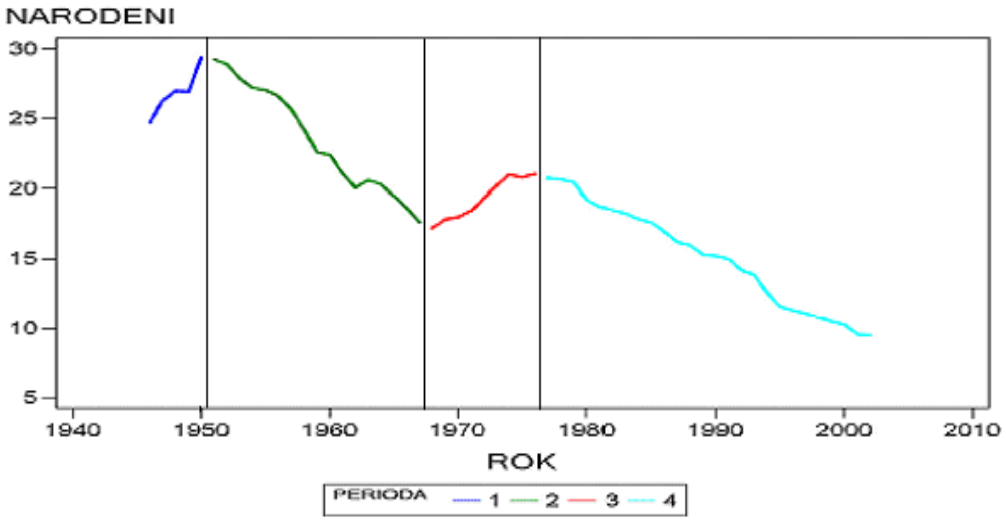
[3] Populačný vývoj v Slovenskej republike 1999. Edícia: Akty, Bratislava, september 2000, http://www.infostat.sk/vdc/sk/index.html

40. rokov a prvá polovica 50. rokov boli v celej Európe poznamenané kompenzačným nárastom pôrodnosti a plodnosti po druhej svetovej vojne. 60. roky, ktoré sú dôležitým medzníkom pre súčasný populačný vývoj vo vyspelých krajinách, priniesli významné zmeny aj do vývoja plodnosti. Začalo sa obdobie poklesu plodnosti, ktoré so zníženou intenzitou trvá prakticky až do súčasnosti. Je to tiež obdobie, kedy sa pod vplyvom politických pomerov začal demografický vývoj, a teda aj vývoj plodnosti v Európe, diferencovať. V krajinách strednej a východnej Európy bol pokles plodnosti prerušovaný hlavne v 60. a 70. rokoch obdobiami stagnácie resp. rastu. Ani na Slovensku nemal pokles plodnosti taký priebeh ako vo vyspelých krajinách západnej Európy. Na konci 50. rokov ho urýchlilo prijatie interrupčného zákona. Ďalej nasledovali dve prerušenia – začiatkom 60. rokov a v prvej polovici 70. rokov. Bol to dôsledok prísľubu a neskôr aj uskutočnenia pronatalitných opatrení (predĺženie materskej dovolenky, vyplácanie materského príspevku po skončení materskej dovolenky, zvýšenie prídavkov na deti, zavedenie mladomanželských pôžičiek, zvýšenie počtu miest v predškolských zariadeniach).

Od polovice 70. rokov nastáva aj na Slovensku neprerušený pokles plodnosti, od konca 70. rokov aj pokles počtu narodených. Až do konca 80. rokov však Slovensko patrilo ku krajinám s najvyššou plodnosťou v Európe. Príčiny tohto stavu treba hľadať okrem iného aj vo vtedajšom spoločenskom a politickom vývoji.”

Vo vyššie uvedenom texte sa hovorí o viacerých poklesoch a vzostupoch premennej natalita (NARODENI), ale my sme ju rozdelili iba do 4. období (periód), aby sme mali dostatok pozorovaní. Použili sme 4 etapy:

1946 – 1950: perióda 1
1951 – 1966: perióda 2
1967 – 1975: perióda 3
1976 – 2002: perióda 4



Graf 4: Časové etapy pre premennú natalita v SR

Rozdelenia vývoja na 4 etapy sme využili pre ďalšiu analýzu. Použili sme metódu *neparametrická ANOVA* (procedúra *NPARWAY* v SAS Enterprise Guide). Nemohli sme použiť procedúru *One-Way-ANOVA*, pretože naše dáta nemali normálne rozdelenie a ani početnosť periód nebola dostatočná. Testovali sme hypotézu:

H0: rozdelenie natality v etapách je identické
H1: rozdelenie natality v etapách nie je identické

Na hladine významnosti α=0,05 zamietame hypotézu H0, lebo p -hodnota pre Kruskal-Walisov test bola blízka 0. Môžme povedať, že periódy sme zvolili vhodne a vývoj natality v čase sa mení aj v závislosti od historických, politických alebo kultúrnych udalostí. Z priemerných hodnôt skóre v tabuľke 5 je zrejmé, že natalita v časovom slede klesá, i keď v tretej etape sme zaznamenali dočasný rast tohto ukazovateľa.

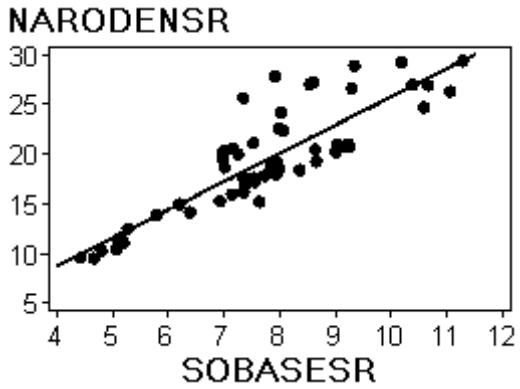
Tabuľka 5: Neparametrická ANOVA, faktor = časová etapa

Period	N	Sum of scores	Expected under H0	Std Dev under H0	Mean Score
1	5	275	145	35.449	55.0
2	17	523	493	57.329	30.7
3	8	307	232	43.527	38.3
4	27	548	783	62.570	20.3

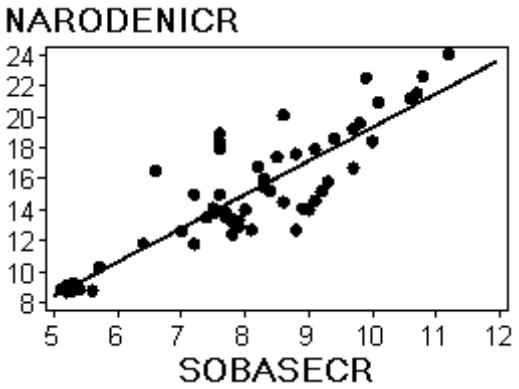
5. Existuje vzťah medzi pôrodnosťou a sobášnosťou na Slovensku? Je vývoj ukazovateľov na Slovensku a v Čechách porovnateľný?

Na internete sú k dispozícii demografické údaje o nami sledovaných demografických ukazovateľoch aj za Českú republiku. Pokúsili sme sa zistiť, či existuje nejaká súvislosť medzi vývojom časových radov v ČR a SR. Vypočítali sme koeficienty korelácie medzi príslušnými časovými radmi v rámci krajín a aj medzi nimi. Rozhodli sme sa zamerať na analýzu korelácie medzi sobášnosťou a natalitou v SR a ČR.

Ako vidieť z grafov, sobášnosť a natalita sú silno korelované. Pearsonove korelačné koeficienty (SR: *0,8687,* ČR: *0,8685)* sa takmer rovnajú a závislosť má identický priebeh, preto môžme povedať, že vývoj v týchto krajinách je podobný.



Graf 5. Korelácia časových radov v SR



Graf 6. Korelácia časových radov v ČR

6. Záver

Modelovali sme časové rady premenných narodení, rozvody a sobáše na 1000 obyvateľov a podľa týchto modelov sme urobili predikciu na roky 2003 až 2006. Na základe vládnych rozhodnutí sme rozdelili premennú natalita na 4 periódy, ktoré sme potom porovnávali. Poslednou vecou, ktorú sme robili bolo zisťovanie, či je nejaký vzťah medzi premennou sobáše a narodení na 1000 obyvateľov. Zistili sme, že tieto časové rady majú podobný priebeh.

7. Literatúra

Populačný vývoj v Slovenskej republike 1999, Edícia: Akty, Bratislava, september 2000

Stankovičová I.: Štatistická analýza krajín sveta podľa vybraných demografických ukazovateľov v roku 1998. Zborník príspevkov zo 7. demografickej konferencie s medzinárodnou účasťou Demografické, zdravotné a sociálno-ekonomické aspekty úmrtnosti, str. 126 - 136, Trenčianske Teplice, 13. - 15. 9. 1999

Adresa autora:
Dušan Leitner
Jarmočná 309/6, 992 01 Modrý Kameň
dusan.leitner@gmail.com

Banka roka 2006 v SR

Marián Magna[1]

Abstract: The object of our project was to compare list of banks in Slovakia scheduled by TREND Analyses with our list scheduled according to weighted principal components. Moreover, we tried to explain our results and found the reasons of other rank.

Comment [F1]: Predmetom nášho projektu bolo porovnať zoznam bánk na SLovensku zoradených podľa TREND Analyses s našim zoznamom zoradeným podľa vážených hlavných komponentov. Navyše sme sa snažili vysvetliť naše výsledky a nájsť dôvody pre rozličné poradie.

Key words: TREND, Trend Analyses, SAS Enterprise Guide, rank of Slovakian Banks, Principal Component Analysis

1. Úvod

V dnešných časoch, keď banky preklenuli obdobie zlého hospodárenia a nesprávnych rozhodnutí došlo k stabilizácii bankového sektora. Zaslúžila sa o to privatizácia všetkých bankových domov na Slovensku, ktorá bola zavŕšená 1. novembra 2005 sprivatizovaním najmenšej slovenskej banky (Banky Slovakia) a jej následným premenovaním na Privatbanku. Práve teraz, keď všetky komerčné banky na trhu sú ovládané a riadené zahraničným kapitálom dochádza k intenzívnejšiemu boju medzi všetkými hráčmi na bankovom trhu. Banky prešli od zavádzania zmien a novej firemnej kultúry k agresívnejšiemu boju o klienta ako to bolo pred niekoľkými rokmi. Dopyt obyvateľstva po úveroch, ktoré rozbehli v minulých rokoch najmä nákupy na splátky, prilákal aj pozornosť bánk, ktoré sa začali biť o klienta aj na trhu úverov.

Podľa analýz spoločnosti TREND Analyses sa za uplynulé roky vyšplhala na prvé miesto Slovenská sporiteľňa, za čo si zaslúžila Výročnú cenu týždenníka TREND Banka roka 2006. Na základe údajov spoločnosti TREND Analyses (tie isté údaje, z ktorých súčasne vychádzame) sme sa rozhodli rozpracovať projekt, v ktorom sme vytvorili rebríček bánk zoradený podľa váženého priemeru hlavných komponentov. Na výpočet sme použili štatistický softvér SAS Enterprise Guide.

2. Dátový súbor

Na Slovensku pôsobí 23 bánk. Z tohto súboru sme vyradili dva štátne špecializované peňažné ústavy (Eximbanku a Slovenskú záručnú a kreditnú banku) a dvoch novoetablovaných hráčov na slovenskom trhu, ktorí nedosiahli v roku 2004 hospodársky výsledok (HSBC bank a Banco Mais). Zostalo nám 19 bánk (štatistických jednotiek), ktoré analyzujeme v našom projekte.

Použité štatistické premenné:

- Zmena trhových podielov poskytnutých úverov 2005/2004 (%): **zm_uvery**
- Zmena trhových podielov prijatých vkladov 2005/2004 (%): **zm_vynosy**
- Prevádzkové náklady / prevádzkové výnosy 2005 (%): **nakladovost05**
- Zmena prevádzkových nákladov / prevádzkových výnosov 2005-04 (%): **nakladovost0504**
- Zisk pred zdanením 2005 (tis. Sk): **zisk05**
- Zmena zisku pred zdanením 2005/2004 (tis. Sk): **zisk0504**
- Rentabilita kapitálu 2005 (%): **ROE05**
- Rentabilita aktív 2005 (%): **ROA05**

[1] Marián Magna, študent 3. ročníka Fakulty managemetu Univerzity Komenského v Bratislave

- Prevádzkové náklady/aktíva 2005 (%): **nakl_ku_aktivam05**
- Zmena prevádzkových nákladov/aktíva 2005/2004 (%): **nakl_ku_aktivam0504**

3. Analýza hlavných komponentov a jej výsledky

Ako prvé sme použili analýzu hlavných komponentov na zredukovanie skorelovaných vstupných premenných. Výsledné umelé premenné – hlavné komponenty – sú lineárnou kombináciou pôvodných premenných a už prvé štyri vysvetľujú 87,27% variability vstupných dát. Na základe toho budeme ďalej uvažovať len prvé štyri hlavné komponenty.

Tabuľka 1: Vlastné vektory

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	**4.26**	2.51	**0.43**	**0.43**
2	**1.75**	0.17	**0.18**	**0.60**
3	**1.59**	0.46	**0.16**	**0.76**
4	**1.13**	0.59	**0.11**	**0.87**
5	0.54	0.08	0.05	0.93
6	0.46	0.36	0.05	0.97
7	0.10	0.02	0.01	0.98
8	0.08	0.01	0.01	0.99
9	0.07	0.05	0.01	1.00
10	0.02		0.00	1.00

Druhým krokom bolo zistenie sily závislosti hlavných komponentov od jednotlivých premenných. V tabuľke možno vidieť, že prvý hlavný komponent (PRIN1) je závislý prevažne od koeficientov rentability (aktív aj vlastného kapitálu), zo zisku v roku 2005 a zo zmeny zisku oproti roku 2004. Druhý hlavný komponent (PRIN2) sa skladá zo zmeny trhového podielu na trhu vkladov oproti roku 2004 a z dvoch pomerov: pomer nákladov k výnosom a nákladov k aktívam. Tretí hlavný komponent (PRIN3) je zložený zo zmeny trhového podielu na úverovom trhu a zo zmeny pomeru nákladov k aktívam oproti roku 2004. Posledný hlavný komponent (PRIN4) je zložený zo zmeny pomeru nákladov k výnosom oproti roku 2004 a z pomeru nákladov k aktívam v roku 2005.

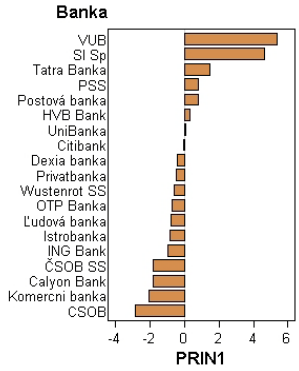
Comment [Fm2]: Je vhodný výraz „zložený“? Nebude lepšie napísať „závislý“

Tabuľka 2: Závislosť hlavných komponentov a vstupných premenných

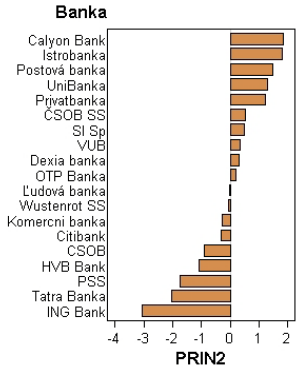
Pearson Correlation Coefficients, N = 19				
	PRIN1	**PRIN2**	**PRIN3**	**PRIN4**
zm_uvery	0,50	0,39	**-0,70**	0,08
zm_vklady	-0,56	**-0,58**	0,17	0,04
nakladovost05	-0,49	**0,75**	-0,20	-0,07
nakladovost0504	-0,18	0,45	0,25	**0,81**
zisk05	**0,90**	-0,18	-0,21	0,12
zisk0504	**0,92**	0,09	-0,24	-0,08
ROE05	**0,89**	-0,12	0,05	0,16
ROA05	**0,86**	-0,21	0,42	0,05
nakl_ku_aktivam05	0,28	**0,51**	0,45	**-0,63**
nakl_ku_aktivam0504	0,40	0,37	**0,70**	0,13

V treťom kroku som vytvoril grafy, v ktorých je zobrazené poradie bánk podľa veľkosti hodnôt jednotlivých hlavných komponentov.

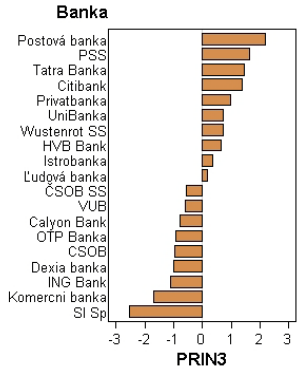
Z prvého grafu vyplýva, že najvyššie absolútne zisky a súčasne najlepšie ukazovatele rentability má Všeobecná úverová banka nasledovaná Slovenskou sporiteľňou a Tatra bankou, najhoršie je na tom Československá obchodná banka. Na druhom grafe si môžeme všimnúť, že najlepšie pomery nákladovosti dosahuje Calyon bank a najhoršie ING banka. Treba si uvedomiť, že náklady nie sú pozitívny faktor pri hodnotení bánk, preto sa na tento graf treba pozerať opačne: najnižšia hodnota hlavného komponentu znamená najlepšiu banku. Tretí graf vyjadruje, že najlepšiu zmenu pomeru nákladov ku aktívam oproti roku 2004 a súčasne zmenu na úverovom trhu dosiahla Poštová banka a najhorší Slovenská sporiteľňa. Takisto ako pri predošlom hlavnom komponente si treba uvedomiť, že najnižší koeficient znamená najlepší koeficient, preto je Slovenská sporiteľňa prvá. Štvrtý graf hovorí o zostavení bánk podľa zmeny nákladov k výnosom oproti roku 2004 a súčasne o pomere nákladov k aktívam v roku 2005. Prvá miesto patrí Calyon banke a posledné Poštovej banke.



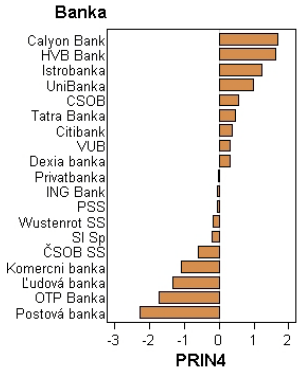
Graf 1: Poradie bánk podľa prvého hlavného komponentu



Graf 2: Poradie bánk podľa druhého hlavného komponentu



Graf 3: Poradie bánk podľa tretieho hlavného komponentu



Graf 4: Poradie bánk podľa štvrtého hlavného komponentu

Ďalším krokom našej analýzy bolo zostavenie tabuľky, v ktorej sú zoradené jednotlivé banky na základe vážených komponentných skóre. Koeficienty váh pre jednotlivé hlavné komponenty sme získali z tabuľky vlastných hodnôt v stĺpci „Proportion“: pre PRIN1 = 0,43;

pre PRIN2 = 0,18; pre PRIN3 = 0,16; pre PRIN4 = 0,11. Zoradili sme si banky podľa hlavných komponentov 1 a 4 zostupne a podľa hlavných komponentov 2 a 3 vzostupne. Na výpočet váženého priemeru poradia sme použili vzorec (1):

$$\frac{\sum_{i=1}^{4} x_i.n_i}{\sum_{i=1}^{4} n_i} \quad , \qquad (1)$$

kde x_i je poradie banky podľa daného hlavného komponentu a n_i je váha pre daný hlavný komponent. Po vykonaní tejto operácie sme dostali výslednú tabuľku poradia (stĺpec „Vážený priemer"). Okrem toho sme do našej tabuľky doplnili stĺpce „Skóre" a „Poradie podľa TRENDu" zostavené analytikmi spoločnosti TREND Analyses.

Tabuľka 3: Výsledné poradie bánk (naše a podľa TREND Analyses)

Naše poradie	Banka	Vážený priemer	Skóre podľa TRENDu	Poradie podľa TRENDu
1.	Všeobecná úverová banka, a.s., Bratislava	5,39	**14,05**	3.
2.	Slovenská sporiteľňa, a.s., Bratislava	5,58	**14,45**	1.
3.	Tatra banka, a.s., Bratislava	5,73	**12,20**	5.
4.	HVB Bank Slovakia, a.s., Bratislava	6,17	**11,15**	6.
5.	Prvá stavebná sporiteľňa, a.s., Bratislava	7,38	**12,40**	4.
6.	Dexia banka Slovensko, a.s., Žilina	8,49	**11,10**	7.
7.	Citibank (Slovakia), a.s., Bratislava	8,92	**8,50**	11.
8.	ING Bank N.V., pobočka zahraničnej banky	9,49	**14,35**	2.
9.	UniBanka, a.s., Bratislava	9,69	**9,90**	8.
10.	Wüstenrot stavebná sporiteľňa, a.s., Bratislava	11,02	**8,15**	14.
11.	OTP Banka Slovensko, a.s., Bratislava	11,28	**9,55**	9.
12.	Poštová banka, a.s., Bratislava	11,76	**9,35**	10.
13.	Československá obchodná banka, a.s.	11,83	**8,45**	12.
14.	Privatbanka, a.s., Bratislava	11,91	**8,05**	16.
15.	Ľudová banka, a.s., Bratislava	12,17	**8,05**	15.
16.	Komerční banka Bratislava, a.s., Bratislava	12,62	**8,25**	13.
17.	Istrobanka, a.s., Bratislava	12,84	**8,00**	17.
18.	Calyon Bank Slovakia, a.s., Bratislava	13,52	**6,65**	19.
19.	ČSOB stavebná sporiteľňa, a.s., Bratislava	14,19	**7,40**	18.

4. Záver

Ako prvé je potrebné uviesť, že čím má banka vyššie skóre podľa TRENDu, tým je lepšia, zatiaľ čo v našom poradí je najlepšia banka s najnižším váženým priemerom.

Z našej analýzy vyplynulo, že najlepšou bankou podľa ukazovateľov za rok 2005 je Všeobecná úverová banka, a.s., pričom v pôvodnom zoradení jej patrilo tretie miesto. Druhá priečka rebríčka patrí Slovenskej sporiteľni a.s., ktorá získala ocenenie Banka roka 2006. Na tretiu pozíciu v našej tabuľke sa dostala Tatra banka, a.s.

Rozdielnosť obidvoch rebríčkov bola spôsobená najpravdepodobnejšie váhou koeficientov, ktorými sa štatistické premenné násobili. Spoločnosť TREND Analyses dáva najväčšiu váhu zmene trhových podielov na úverovom trhu (15%), 5% váhu prikladá zmene prevádzkových nákladov ku aktívam oproti roku 2004 a ostatné premenné majú rovnakú, 10% váhu. Naproti tomu, v našom modeli mal najsilnejšiu váhu (43%) prvý hlavný komponent zložený z koeficientov rentability (kapitálu a aktív), zisku pred zdanením a zmeny zisku oproti roku 2004.

5. Literatúra

MARHOLD, K., SUDA, J.: Analýza multivariačních dat v taxonomii. (Fenetické metody). Přírodovědcká fakulta UK v Praze. Katedra botaniky. Praha 2001.

SHARMA, S.: Applied Multivariate Techniques. New York, John Wiley & Sons, Inc., 1996.

STANKOVIČOVÁ, I.: Viacrozmerná analýza rentability poisťovní SR pomocou Enterprise Guide, In: 10. medzinárodný seminár Výpočtová štatistika. Bratislava: SŠDS 2001. ISBN 80-88946-14-X

Zdroj dát: časopis TREND

Adresa autora:
Marián Magna
Tatranská 25
974 11 Banská Bystrica
marian.magna@st.fm.uniba.sk

Analýza vplyvu environmentálnych premenných na efektívnosť poľnohospodárskej výroby

Martina Majorová

ABSTRACT

Efficiency of a company is nowadays considered to be the key factor of the success of a company. However, efficiency of a company is not an abstract quantity; it can be measured and evaluated through various indicators. In the paper we deal with the analysis of the influence of environmental variables on the agricultural production efficiency. We used the methodology of *DEA* (***D**ata **E**nvelopment **A**nalysis*) to examine the above-mentioned questions. This method belongs to non-parametric approaches to the efficiency measurement of companies and it is based on utilizing the linear programming calculus.

KEY WORDS

efficiency, agricultural production, DEA, data envelopment analysis, environmental variables, linear programming

ÚVOD

Pojem efektívnosť v ostatnom čase rezonuje v sektorovo rozličných oblastiach ľudskej činnosti (napr. poľnohospodárstvo, zdravotníctvo, verejná správa, bankový a finančný sektor a pod.) a mnohokrát sa zamieňa s pojmom produktivita. Snáď najčastejšie sa s ním stretávame v spojitosti s termínom ekonomická efektívnosť. Keďže existencia podnikov v akomkoľvek odvetví súčasnej modernej spoločnosti 21.storočia je podmienená množstvom faktorov, či už vonkajších (napr. aktuálna ekonomická a politická klíma v krajine, kúpyschopnosť obyvateľstva, demografická štruktúra obyvateľstva a pod.) alebo vnútorných (napr. organizačná štruktúra podniku, vedenie podniku a pod.), je zrejmé, že racionálny vlastník podniku (resp. predstavitelia jeho top manažmentu alebo samotný podnikateľ) sa bude snažiť o maximálny výnos (vo väčšine prípadov to bývajú tržby) pri minimálnych hodnotách vynakladaných vstupov (nákladov v akejkoľvek podobe). Riešenie tejto optimalizačnej úlohy je možné dosiahnuť uplatením aparátu lineárneho programovania.

V praxi jestvujú rôzne metódy hodnotenia efektívnosti podnikov. Vo všeobecnosti ich možno rozdeliť na dve skupiny (dva prístupy): parametrický a neparametrický.

Parametrický prístup, často označovaný aj ako ekonometrický, vychádza z predpokladu, že poznáme explicitné vyjadrenie produkčnej funkcie, ale nepoznáme parametre produkčnej funkcie.

Neparametrický prístup je zastúpený analýzou dátových obalov – **DEA** (Data Envelopment Analysis), metodológiu založenou na aplikácii matematického programovania.

Vzhľadom k tomu, že príspevok sa zaoberá skúmaním vplyvu environmentálnych premenných na efektívnosť poľnohospodárskej výroby, považujeme za potrebné definovať si aj tento pojem. Environmentálne premenné vyjadrujú *vplyv prostredia na podnik* a charakterizujú *prostredie* (environment) a jeho vlastnosti, *v ktorom rozhodovacia jednotka*

operuje. Je známych niekoľko spôsobov začleňovania premenných tohto druhu do DEA modelov – buď priamo alebo s využitím doplnkových premenných ([URL 1]).

Cieľom príspevku bolo overiť hypotézu, či je efektívnosť podniku ovplyvená prítomnosťou environmentálnej premennej v modeli, t.z. či je technická efektívnosť podniku závislá od kvality a vlastností pôdy, na ktorej hospodári.

MATERIÁL A METÓDY

Analýza efektívnosti podnikov bola realizovaná na databáze údajov o pšenici za rok 2000, ktorú sme získali z Výskumného ústavu ekonomiky poľnohospodárstva a potravinárstva (VÚEPP Bratislava).

Pri výpočte mier efektívnosti bolo použitých päť štandardných vstupných premenných (*osivá, hnojivá, ostatný materiál, mzdové náklady, ostatné náklady*), jedna environmentálna premenná (*skupina ceny pôdy – SCP*) a jedna štandardná výstupná premenná (*produkcia*). Premenná SCP nadobúdala tri kategórie: podniky hospodáriace v horších výrobných podmienkach (SCP 1-10), podniky hospodáriace v priemerných výrobných podmienkach (SCP 11-15) a podniky hospodáriace v lepších výrobných podmienkach (SCP 16-20).

Výberový súbor pozostával zo 122 podnikov, ktoré bolo potrebné redukovať na 98, aby bola splnená podmienka pre aplikáciu DEA prístupu, t.j. všetky vstupné, resp. výstupné premenné museli vykazovať kladné hodnoty. Popisné charakteristiky týchto premenných zobrazuje Tabuľka 1.

Medzi základné DEA modely patria modely za podmienok konštantných výnosov z rozsahu (CCR[1] modely) a modely za podmienok variabilných výnosov z rozsahu (BCC[2] modely). V príspevku boli všetky miery technickej efektívnosti[3] počítané za podmienok variabilných výnosov z rozsahu z toho dôvodu, že slovenská ekonomika ešte takmer ani po 15 rokoch sústavných reforiem nemôže byť považovaná za trhovú ekonomiku, skôr transformujúcu sa. Práve BCC DEA modely zohľadňujú túto skutočnosť pri výpočtoch mier technickej efektívnosti (TE). Matematický zápis inputovo-orientovaného BCC DEA modelu (1) môžeme definovať takto:

$$
\begin{aligned}
\max \varpi &= \overline{1}^T \bar{s}^+ + \overline{1}^T \bar{s}^- \\
Y\overline{\lambda}^T - \bar{s}^+ &= \bar{y}_0 \\
-\theta \bar{x}_0 + X\overline{\lambda}^T + \bar{s}^- &= \overline{0} \\
1^T \overline{\lambda}^T &= 1 \\
\overline{\lambda}^T, \bar{s}^+, \bar{s}^- &\geq \overline{0} \qquad (1)
\end{aligned}
$$

θ – miera technickej efektívnosti
$Y\lambda$ – virtuálny output
$X\lambda$ – virtuálny input
$\bar{x}_0$ – vstupy hodnoteného podniku (DMU_0)
$\bar{y}_0$ – výstupy hodnoteného podniku (DMU_0)
$\bar{s}^+$ – deficit výstupov
$\bar{s}^-$ – exces vstupov

kde θ je skalár a λ je N×1 vektor konštánt. Výstupom modelu (1) je miera technickej efektívnoti θ, ktorá udáva, ako by mal hodnotený podnik proporcionálne redukovať objem svojich vstupov, aby bol efektívny a odchýlkové premenné $\bar{s}^+$ a $\bar{s}^-$ (tzv. *neradiálne zdroje*

[1] CCR je akronym z mien autorov modelu: Charnes, Cooper, Rhodes.
[2] BCC je akronym z ien autorov modelu: Banker, Charnes, Cooper.
[3] Miera technickej efektívnosti je v príspevku počítaná tak, ako ju vo svojej práci definoval Farrell (1957).

neefektívnosti), ktoré určujú, ako môže podnik ešte neradiálne redukovať objem svojich vstupov (premenné $\bar{s}^-$) alebo expandovať objem svojich výstupov (premenné $\bar{s}^+$) tak, aby sa stal efektívnym. Podmienka konvexnosti $\bar{1}^T \bar{\lambda}^T = 1$ zaručuje, že hodnotený podnik bude porovnávaný iba s podnikmi podobného rozsahu. Podľa modelu (1) je hodnotený podnik považovaný za efektívny vtedy a len vtedy, ak sú súčasne splnené tieto podmienky:

1. $\theta = 1$
2. všetky doplnkové premenné $\bar{s}^+$ a $\bar{s}^-$ sa rovnajú nule (**Pareto-Koopmansova efektívnosť**[4] **alebo celková efektívnosť**).

Tabuľka 1 Základné štatistické charakteristiky premenných vstupov a výstupu kategorizované podľa environmentálnej premennej

Metódy hodnotenia efektívnosti, v ktorých pri hodnotení efektívnosti možno použiť

Premenná	SCP	Minimum	Maximum	Priemer	Štandardná odchýlka
Osivá (tis. Sk)	1-10	306,00	2 761,00	1 057,50	598,22
	11-15	409,00	4 667,00	1 676,16	1 134,02
	16-20	651,00	6 572,00	2 366,97	1 499,16
Hnojivá (tis. Sk)	1-10	370,00	2 914,00	1 160,75	631,47
	11-15	450,00	4 940,00	1 792,12	1 189,79
	16-20	689,00	7 054,00	2 504,62	1 608,82
Ostatný materiál (tis. Sk)	1-10	32,00	1 723,00	419,66	397,33
	11-15	117,00	2 992,00	733,72	733,72
	16-20	92,00	5 572,00	1 193,55	1 285,44
Mzdy (tis. Sk)	1-10	793,00	6 808,00	2 637,91	1 572,95
	11-15	1 217,00	12 599,00	4 202,00	2 915,20
	16-20	1 592,00	19 198,00	6 065,14	4 223,15
Ostatné náklady (tis. Sk)	1-10	148,00	10 004,00	2 425,52	2 367,19
	11-15	405,00	17 380,00	4 426,00	3 830,24
	16-20	2 446,00	22 096,00	7 431,59	5 738,84
Produkcia (tis. Sk)	1-10	57,00	3 361,00	696,17	751,18
	11-15	323,20	6 767,00	1 668,12	1 439,33
	16-20	910,00	10 858,00	2 988,35	2 378,90

environmentálne premenné a ktoré sú založené na aplikácii matematického programovania možno rozdeliť do štyroch základných skupín ([6]):

Prvá metóda predpokladá, že hodnoty environmentálnej premennej možno zoradiť z hľadiska vplyvu na efektívnosť od najlepšej po najhoršiu. Potom možno použiť prístup podľa Bankera a Moreya ([1]). Podľa tejto metódy sa efektívnosť i-teho podniku porovnáva iba s tými podnikmi, ktorých hodnota environmentálnej premennej je horšia nanajvýš rovná hodnote i-teho podniku.

[4] Koopmans (1951) definuje technickú efektívnosť striktnejšie: podnik je technicky efektívny iba vtedy, ak sa nachádza na hranici produkčných možností a všetky odchýlkové premenné (prebytok vsupov, resp. deficit výstupov) sú rovné nule.

Druhá metóda, navrhnutá Charnesom, Cooperom a Rhodesom ([2]), sa odporúča použiť, ak nie je možné prirodzene zoradiť podniky podľa environmentálnej premennej. Táto metóda pozostáva z troch etáp:

1. rozdelenie súboru hodnotených podnikov do podskupín a výpočet mier efektívnosti pomocou matematického programovania v rámci každej skupiny,

2. projekcia údajov hodnotených podnikov na hranice produkčných možností jednotlivých skupín podnikov,

3. riešenie jedného modelu matematického programovania s použitím projektovaných hodnôt a analýza štatistickej významnosti rozdielu stredných hodnôt efektívnosti uvažovaných skupín podnikov.

Nevýhodou prvej a druhej metódy je, že rozdelením podnikov na podskupiny sa môže významne redukovať počet porovnávaných podnikov. To môže viesť k tomu, že veľa podnikov bude hodnotených ako efektívnych, čím sa významne redukuje diskriminačná sila analýzy. Druhou nevýhodou je, že sa môže použiť iba jedna environmentálna premenná.

Tretia metóda vychádza z predpokladu, že environmentálna(e) premenná(é) sú súčasťou úlohy matematického programovania. Environmentálna premenná môže byť do úlohy zaradená buď ako vstup, výstup, alebo ako neutrálna premenná a môže to byť tak premenná pod kontrolou manažéra, ako aj mimo kontroly manažéra. Uvedené možnosti environmentálnych premenných vedú minimálne k trom typom úloh matematického programovania, ktorých charakteristiku možno nájsť napr. v práci Coelliho et al. ([3]).

Štvrtá metóda je dvojetapová procedúra. Prvá etapa pozostáva z riešenia úloh lineárneho programovania, pričom sú použité iba tradičné (neenvironmentálne) faktory. V druhej etape sa pomocou regresnej analýzy skúma závislosť mier efektívnosti získaných v prvej etape od environmentálnych faktorov. Znamienko parametrov regresnej funkcie indikuje smer závislosti a štandardné testovanie hypotéz možno použiť k analýze sily závislosti. Regresnú funkciu možno taktiež použiť ku "korekcii" mier efektívnosti vzhľadom na environmentálne faktory. Táto metóda umožňuje použiť tak spojité, ako aj kategorické environmentálne premenné. Vzhľadom na to, že miery efektívnosti sú vždy z intervalu <0,1>, k odhadu parametrov regresnej funkcie sa namiesto metódy najmenších štvorcov odporúča použiť tzv. Tobit regresiu (Kmenta [7]). Predpokladom použitia tejto metódy je, že premenné použité v prvej etape nekorelujú s premennými druhej etapy.

VÝSLEDKY A DISKUSIA

Vplyv prírodných podmienok na technickú efektívnosť sme hodnotili prostredníctvom faktora „Skupina ceny pôdy (SCP)“. Keďže ide o kategorickú premennú, nadobúdala tieto tri hodnoty:

- podniky hospodáriace v horších výrobných podmienkach (HVP), SCP 1-10,
- podniky hospodáriace v priemerných výrobných podmienkach (PVP), SCP 11-15,
- podniky hospodáriace v lepších výrobných podmienkach (LVP), SCP 16-20.

Analýzu efektívnosti podnikov sme realizovali tromi metódami, ktoré sú určené pre riešenie DEA modelov s environmentálnymi premennými. Metóda 3 nebola použitá pri výpočtoch mier efektívnosti, pretože charakter environmentálnej premennej môže viesť pre každú

rozhodovaciu jednotku minimálne k trom typom úloh lineárneho programovania, čo je značne náročné z hľadiska výpočtu. Štandardná metóda, t.j. miery technickej efektívnosti bez zohľadnenia environmentálnej premennej, bola aplikovaná výlučne pre transparentnejšiu komparáciu dosiahnutých výpočtov (viď. Tabuľka 2).

Tabuľka 2 Metódy merania technickej efektívnosti (BCC model) a komparácia ich hodnôt

Výrobné podmienky	Štand. metóda	Metóda 1	Metóda 2	Metóda 4
HVP	0,6837	0,8308	0,8331	0,7119
PVP	0,7391	0,8161	0,8228	0,7749
LVP	0,8391	0,8439	0,9141	0,8537
Minimum	0,2938	0,3373	0,4140	0,3360
Priemer	0,7428	0,8309	0,8546	0,7699
Štand. odchýlka	0,1810	0,1569	0,1769	0,1818

Z tabuľky 2 vyplýva, že podľa všetkých metód dosahujú podniky hospodáriace v lepších výrobných podmienkach vyššie hodnoty technickej efektívnosti. Najnižšie hodnoty technickej efektívnosti v priemere poskytuje štandardná metóda, čo je z veľkej miery podmienené absenciou environmentálnej premennej v DEA modeli. Na druhej strane, najvyššie miery technickej efektívnosti môžeme pozorovať pri metóde 2, čo je to pravdepodobne dané metodikou výpočtu.

Analýzou rozptylu sme zistili, že medzi vypočítanými mierami technickej efektívnosti existujú štatisticky vysoko preukazné rozdiely pri aplikácii štandardnej metódy a metódy 4 (hodnoty p-value 0,001332355, resp. 0,003959801, viď. Tabuľka 3). Naopak, pri použití metód 1 a 2 sa tieto rozdiely nepreukázali (hodnoty p-value 0,812954923, resp. 0,093253, viď. Tabuľka 3). Kvôli podrobnejším informáciám, sme analýzu rozptylu ďalej použili aj pri zisťovaní rozdielov v mierach technickej efektívnosti medzi dvojicou kategórií (HVP-PVP, PVP-LVP a HVP-LVP). Zaujímavosťou je, že pokým rozdiely medzi PVP-LVP a HVP-LVP v štandardnej metóde sú štatisticky (vysoko) preukazné (p-value 0,01720955, resp. 0,000440592, viď. Tabuľka 3), medzi mierami technickej efektívnosti podnikov HVP a PVP pri aplikácii tej istej metódy je tento rozdiel štatisticky nevýznamný (p-value 0,2226412, viď. Tabuľka 3). Dôvodom môže byť nie veľká variabilita v sledovaných údajoch medzi spomínanými kategóriami.

Tabuľka 3 Výsledky analýzy rozptylu pri jednotlivých metódach merania efektívnosti

Výsledky ANOVY	Štand. metóda	Metóda 1	Metóda 2	Metóda 4
HVP-PVP	0,2226412	0,725213848	0,826635	0,180280069
PVP-LVP	0,01720955	0,472148859	0,041772	0,068750099
HVP-LVP	0,000440592	0,736610037	0,056619	0,001051907
HVP+PVP+LVP	0,001332355	0,812954923	0,093253	0,003959801

Pre overenie si týchto výsledkov sme na údaje aplikovali aj analýzu kontrastov medzi uvedenými kategóriami SCP, konkrétne *Scheffeho test* v štatistickom softvéri SAS, ktorý sme použili kvôli rozdielnemu rozsahu súborov týchto podnikov. Z obrázkov 1 a 2 vyplýva, že štatisticky vysoko preukazné rozdiely existujú medzi kategóriami HVP-LVP, čím sme dospeli k rovnakému vyhodnoteniu ako v prípade štandardného parametrického testu ANOVA.

Obr.1, 2 Výsledok Scheffeho testu pre štandardnú metódu a metódu 4

Štandardná metóda

Scheffe's Test for TE

Comparisons significant at the 0.05 level are indicated by ***.

SCP Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
LVP - PVP	0.10001	-0.01554 0.21555	
LVP - HVP	0.15536	0.05280 0.25792	***
PVP - LVP	-0.10001	-0.21555 0.01554	
PVP - HVP	0.05535	-0.04959 0.16030	
HVP - LVP	-0.15536	-0.25792 -0.05280	***
HVP - PVP	-0.05535	-0.16030 0.04959	

Metóda 4

Scheffe's Test for TE

Comparisons significant at the 0.05 level are indicated by ***.

SCP Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
LVP - PVP	0.07015	-0.04551 0.18580	
LVP - HVP	0.14851	0.04562 0.25141	***
PVP - LVP	-0.07015	-0.18580 0.04551	
PVP - HVP	0.07837	-0.02802 0.18475	
HVP - LVP	-0.14851	-0.25141 -0.04562	***
HVP - PVP	-0.07837	-0.18475 0.02802	

V metóde 4 sme realizovali výpočet mier technickej efektívnosti pomocou tzv. cenzorovaného regresného modelu Tobit. Ide o špeciálny typ regresie, v ktorej závislou premennou sú vypočítané miery technickej efektívnosti bez environmentálnej premennej a nezávislou premennou sú kategórie skupiny ceny pôdy formulované cez techniku umelých premenných. Tobit regresia sa aplikuje z toho dôvodu, že technická efektívnosť vždy nadobúda hodnoty z intervalu <0,1>[5]. Je teda zrejmé, že nemôže byť aplikovaná klasická regresia, pretože nie je splnená požiadavka normálneho rozdelenia a teda koeficienty vypočítané pomocou tejto regresie by mohli poskytovať skreslené hodnoty parametrov.

Z výsledkov Tobit regresie (viď. Tabuľka 4) vyplýva, že priemerná miera technickej efektívnosti rozhodovacích jednotiek hospodáriacich v lepších výrobných podmienkach je 79,63%. Podniky hospodáriace v horších výrobných podmienkach dosahujú v priemere technickú efektívnosť nižšiu o 4,25% a podniky hospodáriace v priemerných výrobných podmienkach vykazujú technickú efektívnosť nižšiu v priemere o 3,57% v porovnaní s podnikmi hospodáriacimi v lepších výrobných podmienkach. Koeficienty pri oboch parametroch sú však štatisticky nevýznamné (p-value 0,340363, resp. 0,423313).

Tabuľka 4 Koeficienty Tobit regresie a im prislúchajúce hodnoty p-value

Parametre	Odhad	Pravdepodobnosť
b1(LVP)	0,796313	2,89E-15
b2 (HVP)	-0,0425549	0,340363
b3 (PVP)	-0,0357367	0,423313

Na základe týchto skutočností môžeme konštatovať, že pracovná hypotéza o vplyve environmentálnej premennej charakterizovanej skupinou ceny pôdy sa nepotvrdila, t.j. *skupina ceny pôdy nemá významný vplyv na efektívnosť podnikov pestujúcich pšenicu*.

SÚHRN

V súčasnosti sa za kľúčový faktor úspešnosti podniku považuje predovšetkým jeho efektívnosť. Efektívnosť podniku nie je abstraktná veličina, je možné ju merať a hodnotiť prostredníctvom viacerých ukazovateľov. V príspevku sa zaoberáme analýzou vplyvu environmentálnych premenných na efektívnosť poľnohospodárskej výroby. Pri skúmaní

[5] V literatúre sa takáto premenná nazýva truncated variable (useknutá premenná).

uvedenej problematiky sme vychádzali z metodológie analýzy dátových obalov (*DEA–**D**ata **E**nvelopment **A**nalysis*), ktorá patrí medzi neparametrické metódy hodnotenia efektívnosti podnikov a je založená na využití aparátu lineárneho programovania.

KĽÚČOVÉ SLOVÁ

efektívnosť, poľnohospodárska výroba, DEA, analýza dátových obalov, environmentálne premenné, lineárne programovanie

LITERATÚRA

[1] BANKER, R. D. – MOREY, R. C.: Efficiency Analysis for Exogenously Fixed Inputs and Outputs. In: Operations Research, zv.34, 1986, č.4, s.513–521

[2] CHARNES, A. – COOPER, W. W. – RHODES, E.: Data Envelopment Analysis as an Approach for Evaluating Program and Managerial Efficiency – with an Illustrative Application to the Program Follow Through Experiment in U.S. Public School Education. In: Management Science, zv.27, 1981, s.668–697

[3] COELLI, T. – RAO, D. S. P. – BATTESE, G. E.: An Introduction to Efficiency and Productivity Analysis. Boston: Kluwer Academic Publishers, 1998

[4] COOPER, W. W – SEIFORD, L. M – TONE, K.: Data envelopment analysis. A Comprehensive Text with Models, Applications, References and Dea-Solver Software. Boston: Kluwer Academic Publishers, 2000

[5] FANDEL, P.: Globálne miery efektívnosti. In: Acta oeconomica et informatica, roč.5, č.2, 2002, s. 47-50

[6] FANDEL, P.: Environmentálne faktory v hodnotení efektívnosti poľnohospodárskej výroby. In: Zborník z medzinárodnej vedeckej konferencie "Medzinárodné vedecké dni 2001 - sekcia Kvantitatívny manažment a informatika". Nitra: SPU Nitra, 2001, s. 609-614, ISBN 80-7137-868-2

[7] KMENTA, J.: Elements of Econometrics. New York: Macmilan Publishing Company, 1986

[8] MAJOROVÁ, M.: Environmentálne premenné a ich vplyv na efektívnosť poľnohospodárskej výroby. Práca ŠVČ. Nitra: SPU, Fakulta ekonomiky a manažmentu, 2006, 38 s. Dostupné na Internete: <http://www.fem.uniag.sk/Martina.Majorova/SVC_2006/SVC_2006.pdf>

[URL 1] Banxia Software - Decision Support Software for Professionals. [online]. [cit. 2006-11-23]. Dostupné na Internete: <http://www.banxia.com/frontier/glossary.html>

KONTAKTNÁ ADRESA

Martina Majorová, Centrum informačných technológií
Fakulta ekonomiky a manažmentu, Slovenská poľnohospodárska univerzita v Nitre
Trieda A. Hlinku 2, 949 76 Nitra, tel. 037/6414 813
e-mail: martina.majorova@fem.uniag.sk

Oponent: doc.Ing. Peter Fandel, CSc.

The people's Republic of Bangladesh

Megbah Ahmed[1]

Introduction

Bangladesh officially the people's Republic of Bangladesh, is a country in South Asia. It is surrounded by India on all sides except a small border with Myanmar to the far southeast and the Bay of Bengal to the south.

East Bengal the region that was to become East Pakistan and now Bangladesh was a prosperous region of South Asia until modern times. The border of Bangladesh were set by the partition of India 1947 when it became the eastern wing of Pakistan separated from western wing by 1600 km. Despite their common religion of Islam, the ethnic and liguistic gulf between the two wings, compared by an apathetic government based in West Pakistan resulted in the independence of Bangladesh under the leadership of Sheikh Mujibor Rahman in 1971 after the blody liberation war.

Bangladesh has the advantages of a mild, almost tropical climate, fertile soil. ample water and abundance of fish wildlife and fruits. The standard of living compared favorably with other parts of South Asia.

Population

Bangladesh is a country having an area of about 147000 square kilometres. It is burned with about 147 million people. About 10000 people live per square kilometer. So It is one of the most densely populated country in the world. Over population adversely affects the economic developments and progress of a country. In Bangladesh population explosion is so high that it creats so many grievous problems which are not very easy to overcome .98 percents of the people of bangladesh are Bangalees. It is the third largest muslim majority nation.The major religion is Muslim with 85 percents of total population. The second major religion is Hinduism which constitutes 13 percents other religions includes Buddhism and Christiany. Minorities include Bharis and tribes Among the tribes Chakma is the biggest.
The following table lists estimates of the population.

[1] Megbah Ahmed, Fakulta sociálnych a ekonomických vied UK Bratislava

Demograpic data

Age structure:

0-14 years :32.9%

15-64 years:63.6%

65 years over:3.5%

Population growth rate :2.09%

Birth rate :29.8 births/1000 population

Death rate :8.27 deaths/1000 population

Net migration rate :-0.68 migrants/1000 population.

Sex ratio

At birth:1.06 Male/ Female

Under 15 years:1.06 male/Female

15-64 years:1.09 Male /Female

65 years and over:1.16 Male/Female

Total population:1.05 male/Female

Life exectancy at birth

Total population:62.46 years

Male:62.47 years

Female:62.47 years

Nationality: Bangladeshi

Ethic groups: Bengali 98%,tribal groups, non Bengali muslim

Religion: Muslim 85%. Hindus 13% others 2%

Language: Begali,English

Literacy: Defination:age 15 and over can read and write

Total population:43.1%(Literate)

Male:53.9%(Literate)

Female:31.8%(Literate)

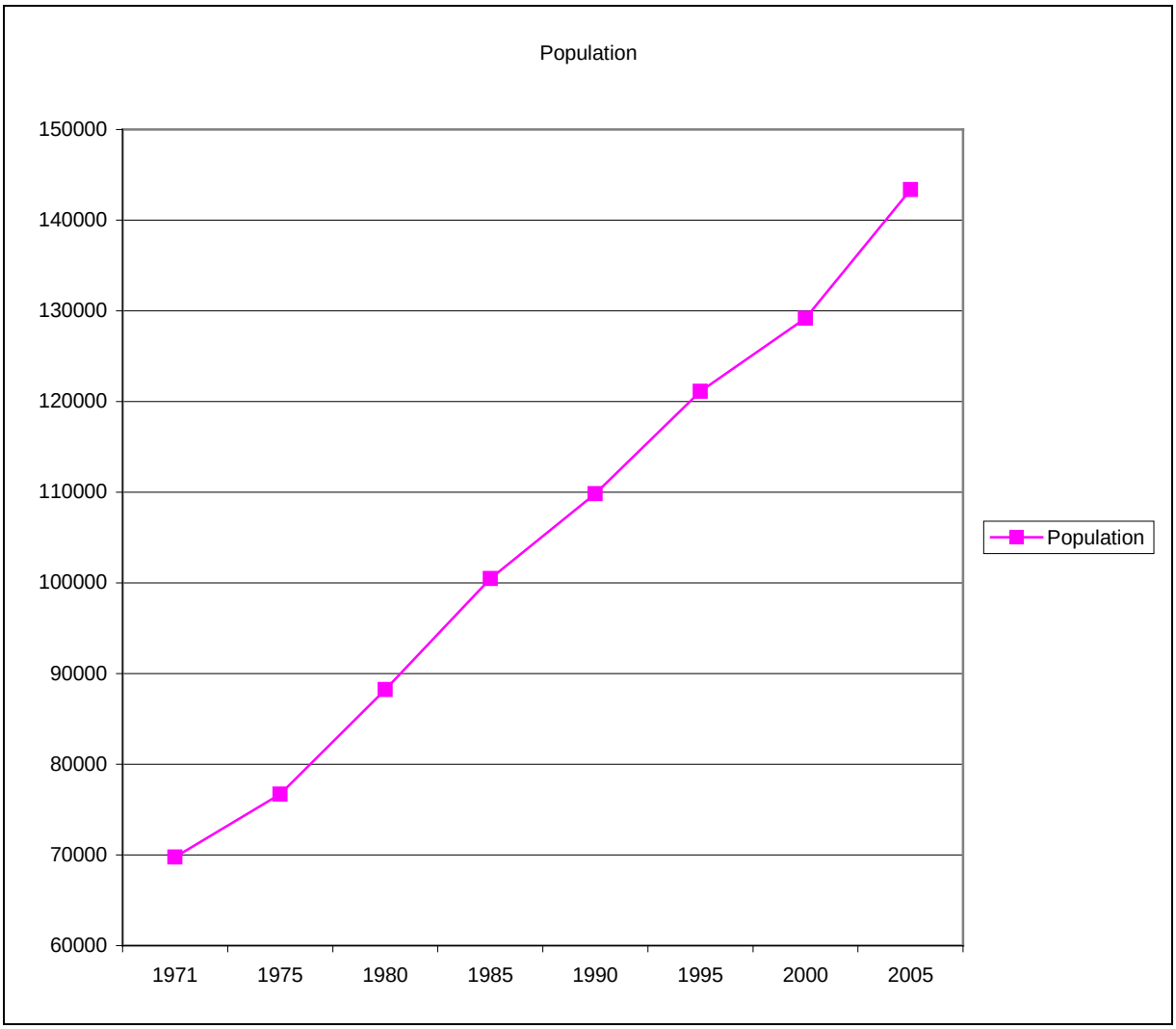

Population
150000
140000
130000
120000
110000
100000
90000
80000
70000
60000
1971
1975
1980
1985
1990
1995
2000
2005
Population

Figure 1

Economy

The backbone of the bangladesh economy agriculture contributes about 45% of gross domestic product. Approximately two thirds of the land area is cultivated. Rice, jute, tea, sugarcane, tobacco and wheat are the chief crops. Fishing is also an important economic activity and beef dairy products and poultry are also produced.

Dhaka and Chattagong are the principle industrial centers, clothing and cotton textiles, jute products, processed and chemical fertilizers are manufactured. In addition to clothing textiles, jute and jute products export include tea, leather. Fish and shrimp. Remittances from several millon bangladeshis working abroadare the second largest foreign income. Western Europe, The USA, India and China are the main trading partner.

This is a chart of trend of gross domestic product of bangladesh at market prices estimated by the international Monetary Fund with figures in millons of Bangladeshi Takas.

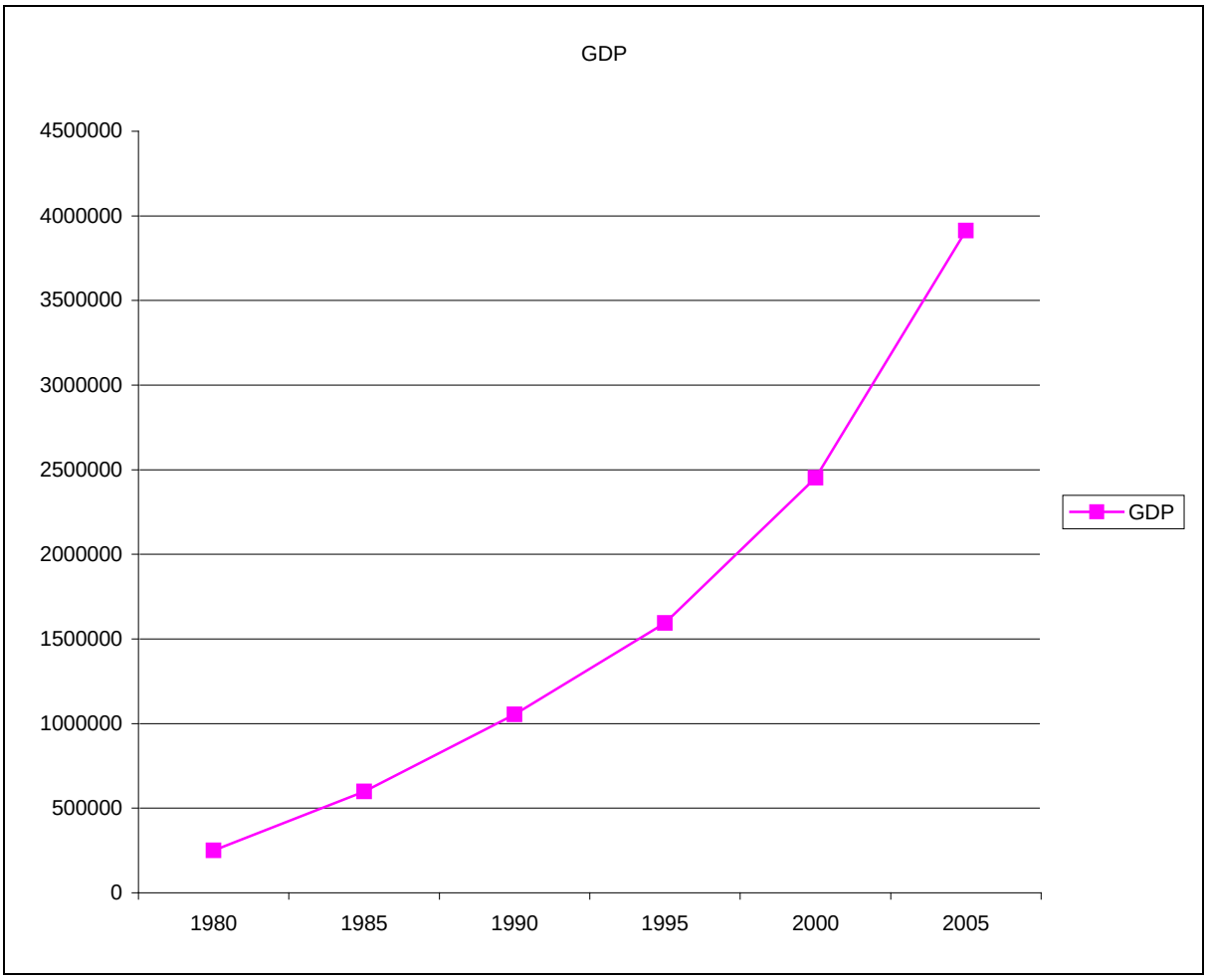


Figure 2

Annual GDP growth rate(2006 est):6.7%

GDP:$60 billion

Exchange rate(2006):1$=Taka 66.70

Per capital GDP(2006):$470

Author:

Bc. Megbah Ahmed

študent 1. ročníka

Fakulta sociálnych a ekonomických vied UK Bratislava

zinslov@yahoo.com

Analýza tržieb predajní vybranej firmy v SR

Matej Mikuška, Martin Hvizdoš[1]

Abstract: We analyzed revenues of a Slovak company in this project, which has a chain of branches in 81 cities in Slovak republic. We analyzed general information of this subject and then we studied the dependency according to size of the cities and number of citizens. We have founded very important dependency between these factors and the fact that they are statistically significant.

Key words: statistical analysis revenues of a Slovak company, test normality, ANOVA, regression model

1. Úvod

V príspevku budeme analyzovať údaje, ktoré sme získali od nemenovanej firmy pôsobiacej na území Slovenska za rok 2005. Firma poskytuje rôzne služby v oblasti telekomunikácií a svoje obchodné prevádzky má rozmiestnené vo väčšine miest Slovenska.

Základné údaje o dátovom súbore:

- rozsah súboru je 130 štatistických jednotiek (predajní),
- 2 kvalitatívne premenné: názov mesta, región (Z – západné Slovensko, S – stredné Slovensko, V – východné Slovensko).
- 4 kvantitatívne premenné: počet obyvateľov v meste, počet predajní v meste, objem ročných tržieb v jednotlivých predajniach v Sk, počet obslúžených zákazníkov za sledovaný rok.
- 1 vytvorená kategoriálna premenná veľkosť obce na základe počtu obyvateľov v obci nasledovne (5 kategórií):

Tabuľka 1. Rozdelenie obcí podľa počtu obyvateľov

počet obyvateľov	veľkosť obce
do 10 000	1
do 20 000	2
do 50 000	3
do 100 000	4
nad 100 000	5

Keďže vo väčších mestách je viac ako jedna predajňa a chceli sme vypočítať priemerné tržby na obyvateľa mesta, tak sme sa rozhodli vytvoriť novú premennú vyjadrujúcu priemernú predajňu v každom meste. Tým sa nám rozsah súboru zmenšil z pôvodných 130 (predajné miesta) na 81 štatistických jednotiek (mestá), tzv. reprezentatívnych predajní. Na základe tejto úpravy sme si mohli vytvoriť ďalšiu dôležitú premennú (priemerná tržba na predajňu v meste), ktorú sme si vypočítali ako Tržby v jednotlivom meste/ Počet predajni v obci.

2. Charakteristika premenných

Tabuľka 2. Popisná štatistika

Premenná	Maximum	Priemer	Minimum	N	Štand. odchýlka	Suma
Počet predajni v obci	11	1.6049383	1	81	1.3479523	130
Tržby v jednotlivých mestách	46066207	4982007.19	1793423	81	5749090.34	403542582
Priemerné tržby na predajni	4187837	2830455.58	1793423	81	554975.10	
Počet obslúžených zákazníkov	17672.55	13120.40	9846	81	1683.49	1062752.3

[1] Fakulta managementu, Univerzita Komenského v Bratislave, študenti 3. ročníka

Z tabuľky č. 2 sme zistili, že najnižšie tržby boli v obci Sobrance 1 793 423 Sk a maximálne v Bratislave 46 066 207 Sk. Priemerné tržby sú 4 982 007 Sk so štandardnou odchýlkou 5 749 090 Sk. Celkové tržby na území SR boli 403 542 582 Sk.

Pri analyzovaní reprezentatívnych (priemerné tržby) predajní sme zistili, že minimálne tržby na predajňu sú v meste Sobrance 1 793 423 Sk a maximum bolo v Bratislave 4 187 837 Sk Priemerné tržby na predajňu sú 2 830 455 Sk so štandardnou odchýlkou 554 975 Sk. Ďalej sme zistili, že minimálny počet obslúžených zákazníkov bol 9846 v meste Sobrance a maximálny počet obslúžených zákazníkov bol 17 673 v Bratislave. Priemerne bolo obslúžených 13 121 ľudí so štandardnou odchýlkou 1684. Firma spolu obslúžila za dané obdobie vo svojich predajniach 1 062 752 zákazníkov.

Na 95 % intervale spoľahlivosti sa priemerné tržby na predajňu pohybujú v rozmedzí od 2 707 741 Sk do 2 953 171 Sk so štandardnou odchýlkou od 480 708 Sk do Sk 656 596 Sk. Tento interval nám odstránil prílišné odchýlky od priemeru čiže hodnoty lepšie zodpovedajú celkovej vzorke.

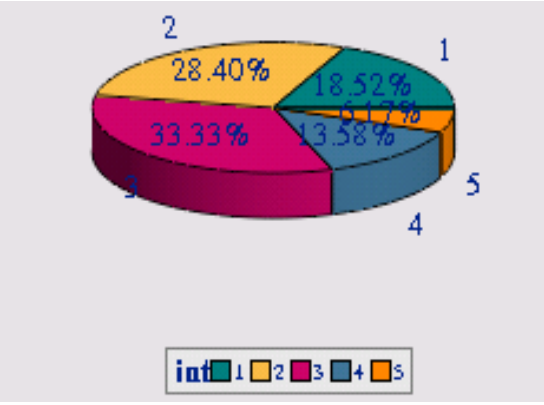
2. Analýza priemerných tržieb na predajňu za dané obdobie

Podľa dosiahnutých tržieb sme rozdelili predajne do 5 príjmových intervalov a to nasledovne:

- 1 - slabá: do 2 280 000
- 2 - podpriemerná od 2 280 001 do 2 760 000
- 3 - priemerná od 2 760 001 do 3 240 000
- 4 - nadpriemerná od 3 240 001 do 3 720 000
- 5 - výborná od 3 720 001

Tabuľka 4. Frekvenčná tabuľka priemerných tržieb na predajňu

int	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	15	18.52	15	18.52
2	23	28.40	38	46.91
3	27	33.33	65	80.25
4	11	13.58	76	93.83
5	5	6.17	81	100.00



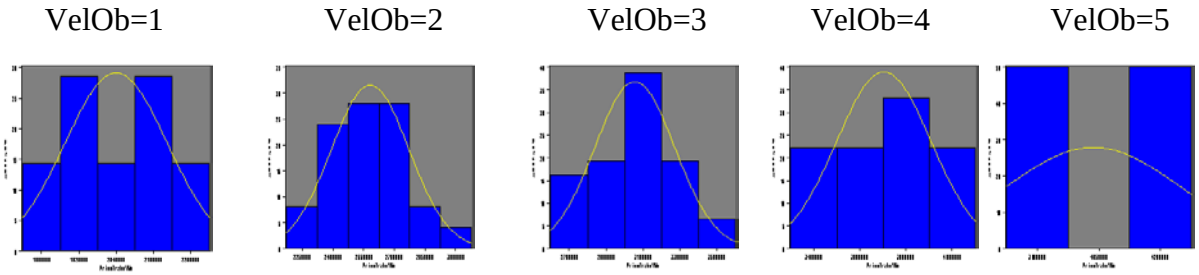
Graf 1. Ziskovosť reprezentatívnych predajní

Z nameraných výsledkov z tabuľky č. 4 vidíme, že až 38 z 81 reprezentatívnych predajní nedosahuje ani priemerné tržby, čo znamená, že zaostávajú v poskytovaní služieb. V kategórii najvyššej ziskovosti sa nachádza 5 reprezentatívnych predajní (Banská Bystrica, Bratislava, Košice, Nitra, Trnava).

3. Vplyv veľkosti obce na tržby

Tabuľka 5. Test normálového rozdelenia priemerných tržieb v meste podľa veľkosti obce

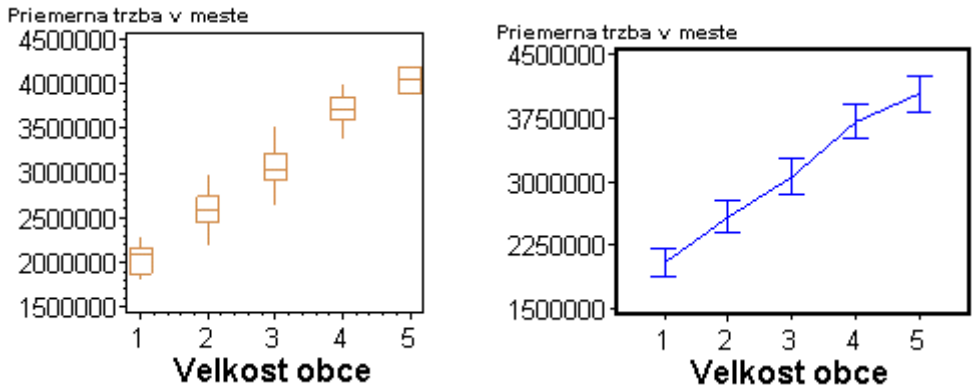
Veľkosť obce	Kolmogorov-Smirnov; p Value	Normálne rozdelenie
1	0.148	áno
2	0.15	áno
3	0.15	áno
4	0.15	áno
5	0.15	áno



Graf 2. Normálne rozdelenie podľa veľkosti obcí

Z výsledkov (tabuľka č. 5 a graf č. 2) analýzy a grafov nám vyplýva, že premenná priemerné tržby v meste má normálne rozdelenie vo všetkých kategóriách rozdelených podľa veľkosti obce tento poznatok sme využili pri výbere metódy zisťovania závislosti – One-way Anova.

Pomocou One-way Anova sme skúmali mieru závislosti medzi premennými Priemerná tržba v meste a Veľkosť obce.



Graf 3. Rozptyly premennej Veľkosť obce

Na základe výsledkov Bartlettovho testu sme zistili, že rozptyly sa zhodujú (ChiSq = 0,8644), čo vidno aj z grafu č. 3, a tým sme sa presvedčili, že môžeme pokračovať ďalej v interpretácii tohto modelu.

Tabuľka 6. Model One-way Anova

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	2.1631815E13	5.4079537E12	136.64	<.0001
Error	76	3.0079738E12	39578602546		
Corrected Total	80	2.4639789E13			

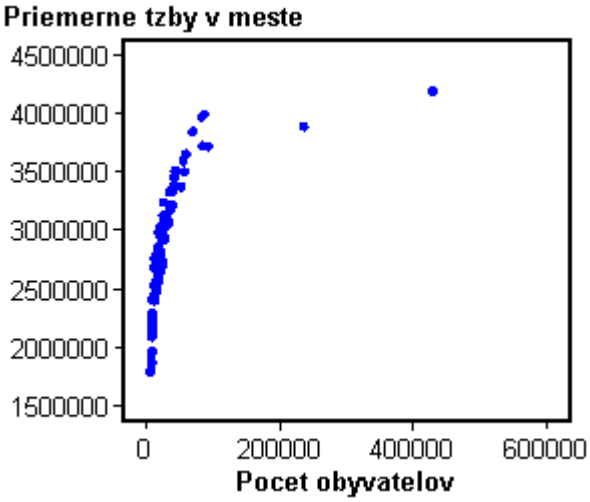
Keďže sme ďalej zistili, že P hodnota (0.0001) je menšia ako alfa (0.05), overili sme si, že dané premenné sú závislé (tabuľka č. 6). Miera závislosti je vyjadrená v tabuľke č. 7 hodnotou R- Square, ktorá sa rovná 0.8779, tzn. 87.8 % zmien výšky tržieb nám vysvetľuje daný model. Ostatné faktory ovplyvňujúce zmenu výšky tržieb tento model nevyjadruje.

Tabuľka 7. Významnost modelu One-way Anova

R-Square	Coeff Var	Root MSE	PriemTrzbaVMe Mean
0.877922	7.028682	198943.7	2830456

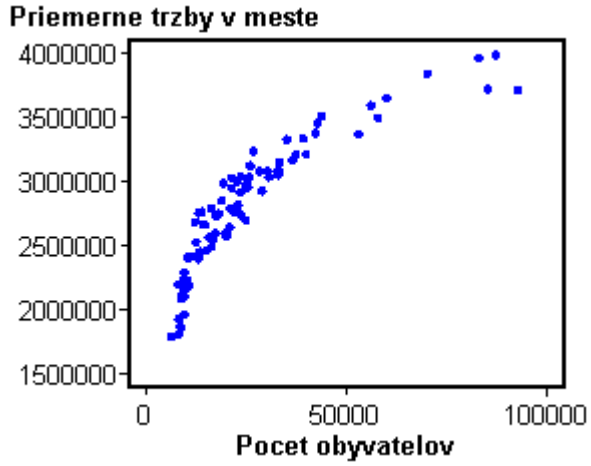
4. Miera vplyvu počtu obyvateľov na priemerné tržby v meste.

Z korelácie premenných Počet obyvateľov a Priemerné tržby v meste nám vyšiel korelačný koeficient 0.61963. Značí nám to skutočnosť, že keď sa zmení počet obyvateľov o 1 jednotku, priemerné tržby sa v závislosti na charaktere zmeny znížia alebo zvýšia o 0.61963 jednotky.



Graf 4. Graf korelácie

Následne sme upravili vzorku vyňatím dvoch hodnôt, ktoré nezapadajú do modelu (viď graf 4), a to mestá Bratislava a Košice. Čiže sa nám vzorka zúžila na 79 údajov. Túto upravenú vzorku sme opätovne korelovali:



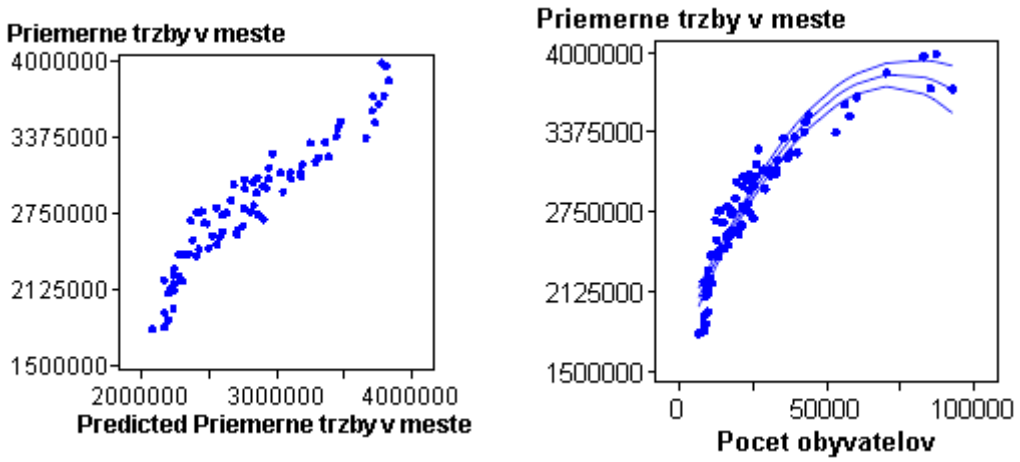
Graf 5. Graf korelácie

Korelačný koeficient sa zvýšil na 0,88241 oproti predchádzajúcej korelácii, náš model nám teraz oveľa lepšie vystihuje vzorku. Keď sa zmení počet obyvateľov o 1 jednotku, priemerné tržby sa v závislosti na charaktere zmeny znížia alebo zvýšia o 0,88241 jednotky.

Zobrazenie z grafu č. 5 je rozhodujúce v určení ďalšieho postupu, a to vo výbere regresie, keďže namerané hodnoty nemajú ani približný tvar priamky (nedajú sa preložiť priamkou) použijeme pri výpočte regresnej priamky nelineárnu regresiu.

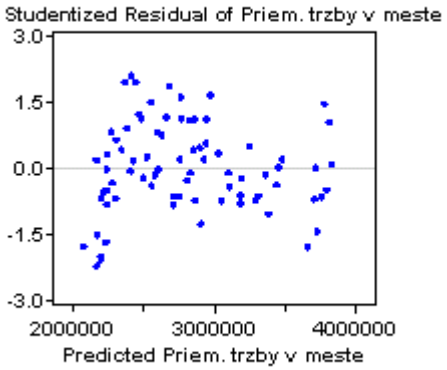
Usporiadanie hodnôt v grafe č. 5 nám najlepšie vystihuje vzorec:

*Priemerna trzba v meste = a + b * Pocet obyvatelov + c * Pocet obyvatel **2*



Graf 6. Grafy nelineárnej regresie

Z grafu č. 6 je možné vidieť, že použitý model (vzorec) vystihuje analyzované hodnoty.



Graf 7. Reziduá nelineárneho modelu

Graf č. 7 nám takisto potvrdzuje vhodný výber modelu, lebo všetky hodnoty sa nachádzajú v intervale <-3,3> od regresného modelu.

Tabuľka 8. Hodnoty regresnej priamky

Parameter	Estimate	Approx. Std Error	Approximate 95% Confidence Limits	
a	1742692	53780.6	1635578	1849805
b	55.9294	3,3484	49.2605	62.5983
c	-0.00037	0.000037	-0.00045	-0.00030

*Priemerna trzba v meste = 1742692 + 55.9294 * Pocet ob. + (- 0.00037)* Pocet ob. **2*

3. Záver

Z predchádzajúcich analýz môžeme jednoznačne usúdiť, že faktory Veľkosť mesta a Počet obyvateľov sú štatistický významné a vieme nimi do značnej miery interpretovať ich vplyvy na vývoj tržieb na území Slovenskej republiky.

Adresy autorov:
Matej Mikuška
Komenského 1627/5
020 01 Púchov
mato.mik@orangemail.sk

Martin Hvizdoš
Dukelské náměstí 30/6
693 01 Hustopeče, Česká republika
martin.hvizdos@centrum.cz

Infraštruktúrna vybavenosť rómskych osídlení na Slovensku a v Bratislavskom kraji

Jana Pukačová

ABSTRACT

The article is focused on sociographic mapping of romes settlements in region of Bratislava. Research was carried out in 2004 and was realized on whole area of Slovak Republic. Article is focused mainly on presentation results in Bratislava region and to their comparation with general state. Another part is analysis and comparation of aquipment and ingineering networks in Bratislava region with average values in all Slovak Republic.

ÚVOD

Posudzovať situáciu v rómskych komunitách je veľmi problematické, nakoľko na Slovensku je nedostatok údajov, na základe ktorých by to bolo možné. Podrobná a presne definovaná štatistika rómskych osád a osídlení na Slovensku v podstate neexistovala až do roku 2004, kedy boli sprístupnené výsledky sociografického mapovania rómskych osídlení. Výskum sa zameral na zisťovanie úrovne rómskych osídlení s ohľadom na možnosti využívania základnej infraštruktúry, dostupnosti rôznych služieb a celkovej úrovne bývania Toto sociografické mapovanie sa uskutočnilo na celom území Slovenska. Neboli však doň zahrnuté tie obce, kde bol počet obyvateľov rómskeho osídlenia nižší ako 15. Pokiaľ však v obci bola situácia obyvateľov rómskej komunity problematická (z hľadiska životnej úrovne, bývania, vzťahu s majoritou) bol výskum realizovaný aj v rámci týchto obcí.

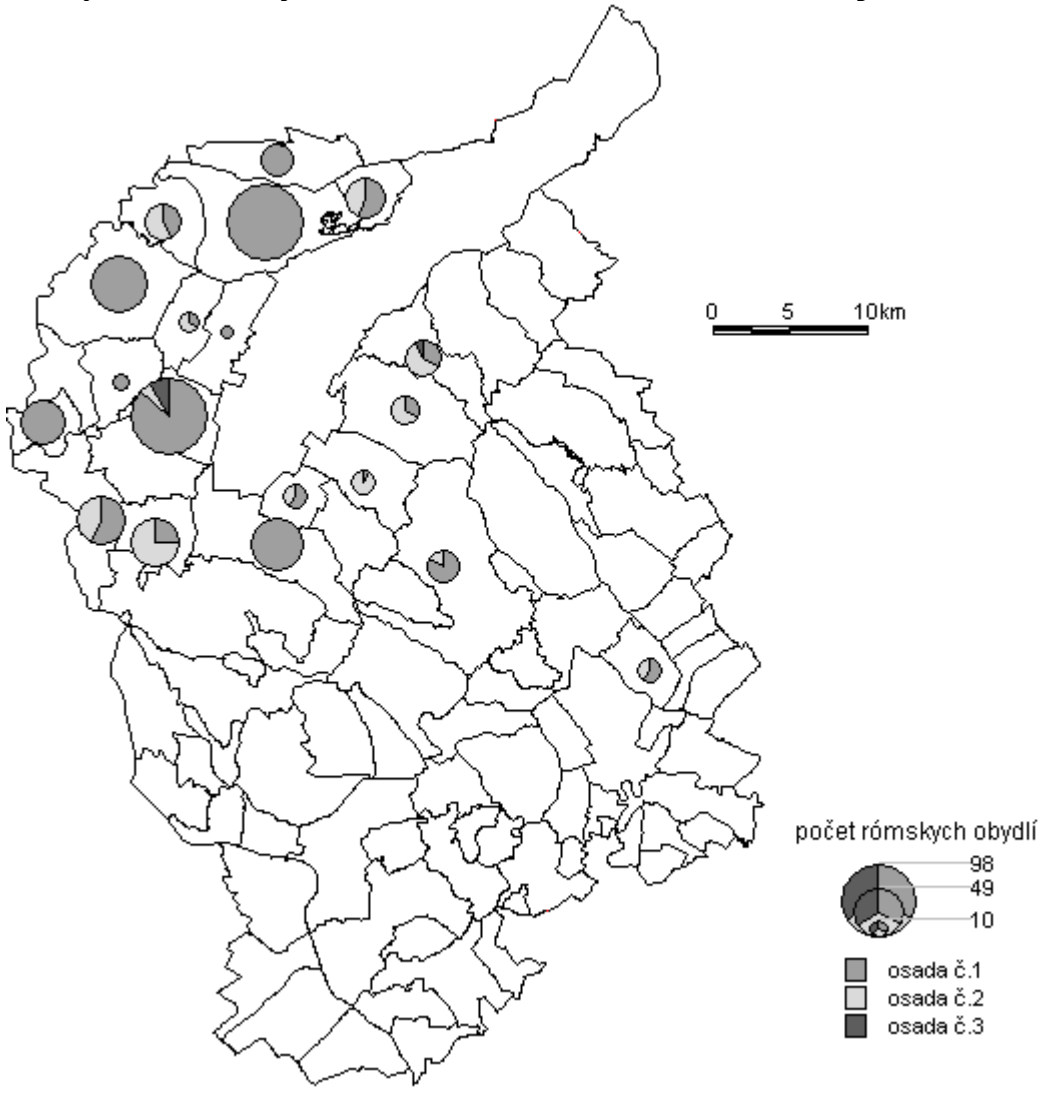
Podľa výsledkov výskumu, Rómovia žijú v 1087 obciach. V týchto obciach bolo definovaných 1575 osídlení rôzneho typu, v ktorých žije 282 315 Rómov. Osídlenie bolo pre potreby výskumu v roku 2004 definované ako zoskupenie minimálne troch domov obývaných komunitou, ktorú majoritné obyvateľstvo označilo ako rómsku (Radičová, 2004). Približne v polovici zmapovaných rómskych osídlení žijú ich obyvatelia rozptýlene medzi majoritou. Zvyšok týchto osídlení má charakter obecných resp. mestských koncentrácií, alebo sa nachádzajú na okraji obce/mesta, eventuálne sú vzdialené.

Pri sociografickom mapovaní sa zisťoval aj počet obydlí v jednotlivých rómskych osídleniach. Medzi obydlia boli zaradené byty v bytových domoch, murované domy, drevené domy, nebytové budovy, chatrče, tzv. unimobunky a iné objekty (napr. stany, autobusy, fóliovníky a pod.). Na Slovenku bolo zistených celkovo 37 211 rôznych obydlí.

V regióne Bratislavského kraja sa vyskytuje v porovnaní s ostatnými regiónmi Slovenska najmenej rómskych osídlení (32 osídlení), v ktorých žije spolu 3 535 Rómov (1,3% všetkých Rómov). Tieto rómske osídlenia sú situované v troch okresoch kraja, najviac ich je v okrese Malacky (28 osídlení), ďalej potom v okrese Pezinok ako aj Senec Rómovia žijú len v dvoch osídleniach (Mapa č.1.)

Mapa č.1.

Počet obydlí v rómskych osídleniach v Bratislavskom kraji v roku 2004



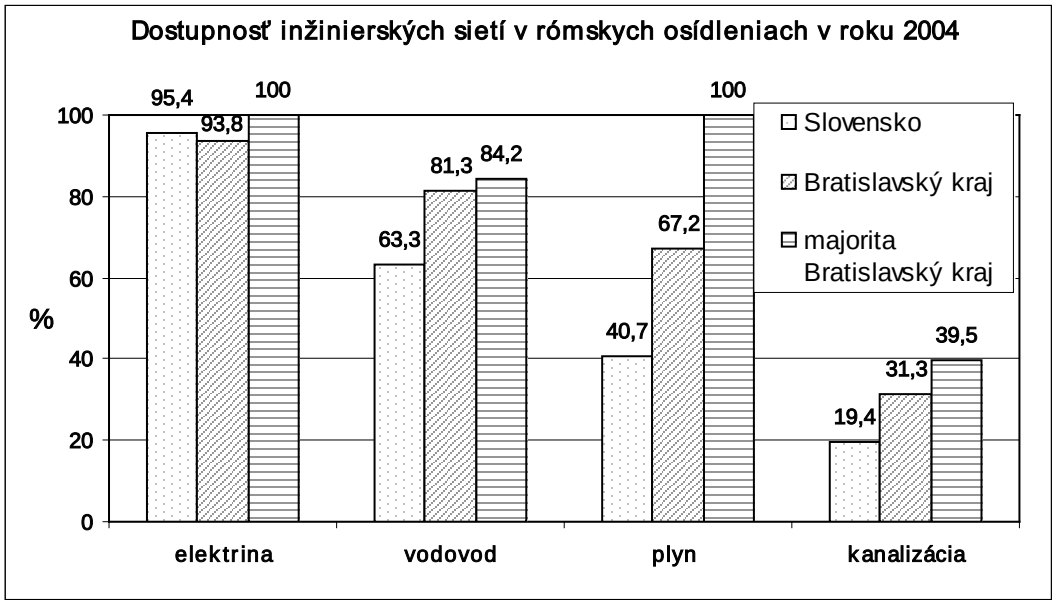
Zdroj: Atlas rómskych komunít na Slovensku, 2004

Vybavenosť rómskych osídlení technickou infraštruktúrou je na území Slovenska do veľkej miery diferencovaná. Táto rôznorodosť a od nej závislá kvalita života obyvateľov týchto osídlení je podmienená viacerými faktormi, ktoré je potrebné podrobnejšie popísať. Najlepšie sú vybavené osídlenia, ktoré sa nachádzajú priamo v obci/meste. Čím je však rómske osídlenie viac umiestnené na periférii tj. vzdialenejšie od obce/mesta, tým je v ňom úroveň života nižšia. Pri vybavenosti rómskych osídlení infraštruktúrou sa brali do úvahy primárne dve charakteristiky.

V prvom prípade sa jednalo o dostupnosť inžinierskych sietí (elektrina, vodovod, plyn, kanalizácia) v obci a v rómskych osídleniach, teda ich prístupnosť a využiteľnosť zo strany obyvateľstva. Pri porovnávaní dostupnosti inžinierskych sietí (Graf č.1.) je infraštruktúrna vybavenosť rómskych osídlení v Bratislavskom kraji jednoznačne lepšia ako priemer pre celú

Slovenskú republiku (s výnimkou dostupnosti elektriny, ktorá je naopak na Slovensku vyššia avšak len o 1,6%, tento rozdiel je tak zanedbateľný). Ostatné inžinierske siete sú v prípade Bratislavského kraja vo väčšine rómskych osídlení dostupnejšie než v ostatných regiónoch na Slovensku. Dostupnosť niektorých inžinierskych sietí v rómskych osídleniach sa dokonca približuje k dostupnosti týchto sietí pre majoritné obyvateľstvo žijúce v predmetných obciach. V reálnych číslach sa to prejavuje nasledovne. Dostupnosť vodovodu v rámci rómskych osídlení v Bratislavskom kraji je 81,3 %, dostupnosť v obci pre tam žijúcu majoritu je 84,2% (rozdiel 2,9% je tak prakticky bezvýznamný). Podobne je to napríklad aj s dostupnosťou elektrickej energie (rozdiel tu činí celkovo 6,2%) alebo kanalizácie (v tomto prípade je rozdiel 8,2%). Najvýraznejšie sa rozdiely prejavujú v dostupnosti, a to nielen na Slovensku, ale aj v prípade Bratislavského kraja medzi rómskymi osídleniami a majoritou, v oblasti prístupu k prípoju plynu resp. v celkovej úrovni plynofikácie týchto osídlení. V tejto problematike sa však objavuje aj výrazná odlišnosť dostupnosti v rámci vzájomnej komparácie rómskych osídlení na území Slovenskej republiky. Kým v rámci Slovenska je dostupnosť prívodu plynu pre domácnosti len pre 40,7 % rómskych osídlení, v kraji Bratislava je to až pre 67,2 % v rovnakom type osídlenia. Napriek tomu však v porovnaní miery plynofikácie rómskych osídlení v regióne Bratislavy s majoritným obyvateľstvom, je tento rozdiel pomerne výrazný. Kým pre majoritu je plyn dostupný vo všetkých obciach, v rómskych osídleniach je to len 67,2 % dostupnosť. V niektorých prípadoch však nemožno vnímať nedostupnosť jednotlivých inžinierskych sietí len z pohľadu rómskeho obyvateľstva, pretože v niektorých obciach Bratislavského kraja nie je napr. vodovod a kanalizácia zavedená vôbec (jedná sa o obce Jablonové a Pernek, kde obe tieto siete chýbajú).

Graf č.1.

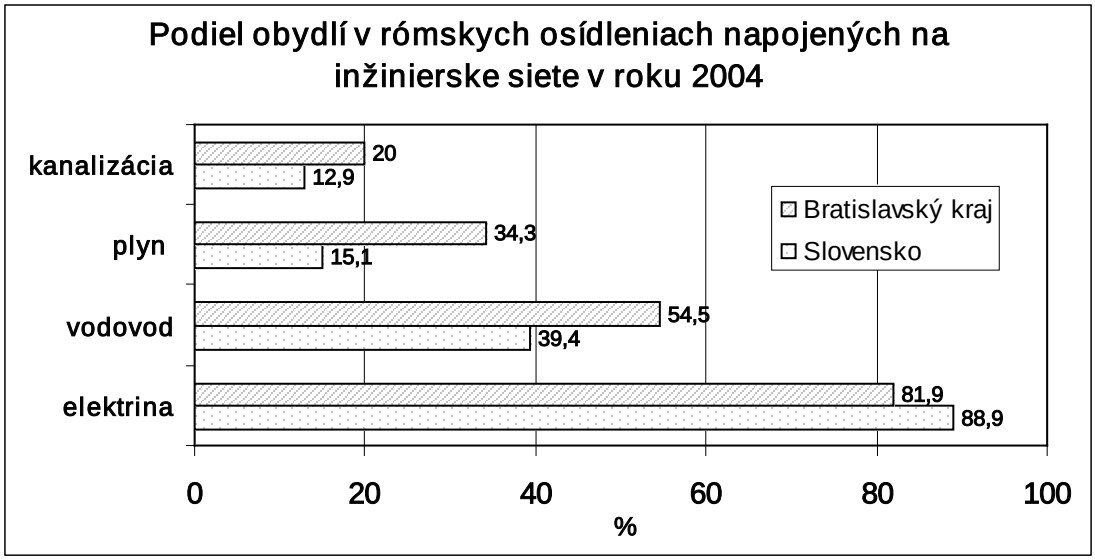


Zdroj: Atlas rómskych komunít na Slovensku, 2004

Druhou charakteristikou pri vybavení rómskych osídlení infraštruktúrou je počet obydlí napojených na tieto siete (Graf č.2.). V rámci Slovenska z inžinierskych sietí Rómovia najčastejšie využívajú elektrickú energiu. Z hľadiska dostupnosti, tej ktorej inžinierskej siete, sú potom Rómovia najviac napojený na kanalizáciu (rozdiel medzi dostupnosťou a napojenosťou je 6,5 %), pri napojenosti na vodovod a plyn je tento rozdiel podstatne vyšší (cca 25 %). Podiel

obydlí napojených na jednotlivé inžinierske siete je v Bratislavskom kraji vyšší ako v ostatných krajoch Slovenska (s výnimkou napojenia na elektrinu). Ale taktiež ako na Slovensku aj v Bratislavskom kraji je podstatný rozdiel medzi dostupnosťou a napojenosťou na jednotlivé inžinierske siete. Podobne ako na Slovensku aj v kraji Bratislava je najviac rómskych obydlí napojených na kanalizáciu (rozdiel medzi dostupnosťou a napojenosťou je 11,3 %). Zo všetkých rómskych obydlí, ktoré mali možnosť napojiť sa na vodovod to využilo len 54,5 % obydlí (zvyšných 26,8 % sa nenapojilo). Na prípojku plynu sa dokonca nenapojilo až 33 % rómskych obydlí. Medzi hlavné dôvody nenapojenia sa na plyn je predovšetkým vysoká cena za prípojky, ktoré si väčšina rómskych obyvateľov v súčasnej sociálnej situácii nemôže dovoliť.

Graf č.2.



Zdroj: Atlas rómskych komunít na Slovensku, 2004

ZÁVER

Sociografické mapovanie ukázalo, že rómske osídlenia na Slovensku sú vysoko rozdielne z hľadiska množstva faktorov. Zistená vybavenosť infraštruktúrou v rómskych osídleniach je v Bratislavskom kraji na vyššej úrovni ako na Slovensku. Jednou z hlavných príčin môže byť, že v regióne Bratislava sa nenachádzajú osídlenia s veľkým počtom obyvateľov a hlavne tieto osídlenia sú situované v obci respektíve na jej okraji. Mimo obce sa nachádza len päť osídlení. Práve vzdialenosť od obce diametrálne znižuje možnosť napojiť sa na jednotlivé inžinierske siete dostupné v danej obci/meste. Vzhľadom na oveľa vyššiu koncentráciu rómskych osídlení ako aj obydlí na východe Slovenska (V Košickom a v Prešovskom kraji je situovaných 50% všetkých rómskych osídlení a 54% obydlí, v ktorých žije až 60% všetkých Rómov) sa dá predpokladať, že práve na tomto území žije najviac týchto komunít v segregácii resp. separovane na okraji obce/mesta a tu je možnosť pripojenia sa na inžinierske siete najmenšia. Možno sa domnievať, že práve vplyvom vyššej vybavenosti rómskych osídlení na území Bratislavského kraja, je v nich aj vyššia celková životná úroveň obyvateľov.

POUŽITÁ LITERATÚRA:

Jurásková, M., Kriglerová, E., Rybová, J.: *Rómovia*. In: Kollár, M., Mesežnikov, G. (ed.): Slovensko 2004. Súhrnná správa o stave spoločnosti. Bratislava, Inštitút pre verejné otázky 2004, pp. 229 – 256

Kadlečíková, J., Kriglerová, E.: *Rómovia*. In: Kollár, M., Mesežnikov, G., Bútora, M. (ed.): Slovensko 2005. Súhrnná správa o stave spoločnosti. Bratislava, Inštitút pre verejné otázky 2005, pp. 169 - 187

Radičová, I. (ed.): *Atlas rómskych komunít na Slovensku*. Bratislava, Inštitút pre verejné otázky 2004

Adresa:

Bc. Jana Pukačová, Prírodovedecká fakulta UK BA, 1.roč. humánna geografia a demografia
jankap83@seznam.cz

Štúdium morbidity obyvateľstva ako indikátora jeho zdravotného stavu

Magdaléna Pullmannová

Abstract
This contribution is focused on morbidity as one of the factors which describes health of the population. The attention was paid especially to diseases of circulation system, respiratory system, digestive system, tumour diseases and external injuries and poisons.

1. Úvod

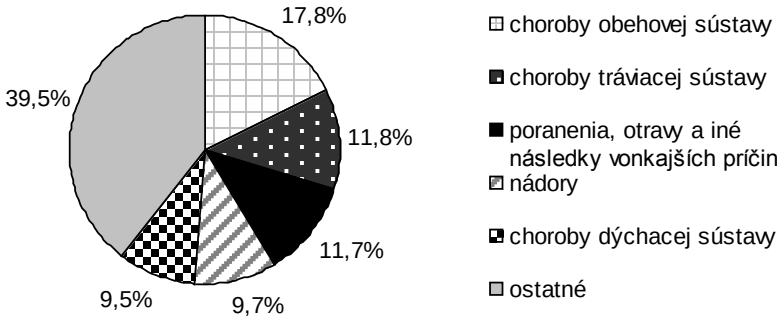
Zravie je podľa definície WHO „celkový stav úplnej telesnej, duševnej i socálnej pohody". Ak sa popisuje zdravotný stav obyvateľstva, používajú sa častejšie nepriame ukazovatele, ako je úmrtnosť alebo chorobnosť, alebo výsledky cielených vyšetrení určitej špeciálnej stránky zdravia na reprezentatívnom súbore ľudí. Je potrebné si uvedomiť, že žiadny z ukazovateľov zdravotného stavu obyvateľstva, ktoré sa dnes používajú, nevypovedá o zdraví komplexne, ale len o jeho niektorej stránke. Úmrtnosť je ukazovateľom len u chorôb, ktoré sú príčinou smrti. Chorobnosť je ukazovateľom u chorôb, s ktorými ľudia prídu k lekárovi a ak sú to choroby, ktoré podliehajú povinnému hláseniu. Takých chorôb nie je veľa (infekcie, nádory, choroby s pracovnou neschopnosťou). (Kríž a kol. 1997)
V tomto príspevku sa zameriavam na morbiditu – chorobnosť. Sledovať, vyhodnocovať a v konečnom dôsledku zlepšovať zdravotný stav populácie je a aj bude aktuálne a relevantné.

2. Štruktúra chorôb - príčin hospitalizácie

2.1. Príčiny hospitalizácie mužov v roku 2004

Najčastejšou príčinou hospitalizácie u mužov sú choroby obehovej sústavy. Ich podiel na celkovom počte hospitalizovaných mužov bol 17,8 %. Na druhom mieste sú choroby tráviacej sústavy, tvoriac 11,8 %. Na treťom mieste sa poranenia a otravy a niektoré iné následky vonkajších príčin podieľali 11,7 %. Štvrté miesto zaujali nádory, a to 9,7 %. Piate miesto zaujali choroby dýchacej sústavy s podielom 9,5 %. Týchto päť skupín chorôb, ktoré sú najčastejšími príčinami hospitalizácie mužov, sa v roku 2004 podieľali 60,5 % z celkového počtu hospitalizácií.

Graf 1 Päť najčastejších príčin hospitalizácie mužov v roku 2004

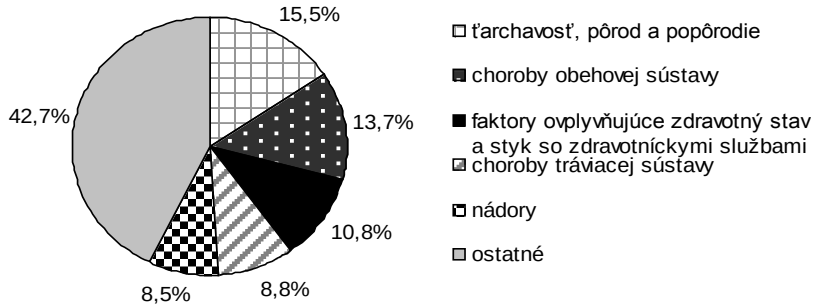


Zdroj : NCZI 2005

2.2. Príčiny hospitalizácie žien v roku 2004

Najčastejšou príčinou hospitalizácie u žien je ťarchavosť, pôrod a popôrodie, ktoré sa na celkovom počte hospitalizovaných žien podieľalo 15,5 %. Na druhom mieste sú choroby obehovej sústavy s podielom 13,7 %. Na treťom mieste sú faktory ovplyvňujúce zdravotný stav a styk so zdravotníckymi službami, tvoriac 10,8 %. Štvrté miesto zaujali choroby tráviacej sústavy s podielom 8,8 %. Piate miesto zaujali nádory s podielom 8,5 %. Týchto päť skupín chorôb, ktoré sú najčastejšími príčinami hospitalizácie žien, sa v roku 2004 podieľali 57,3 % z celkového počtu hospitalizácií.

Graf 2 Päť najčastejších príčin hospitalizácie žien v roku 2004



Zdroj : NCZI 2005

3. Vývoj počtu hospitalizovaných podľa vybraných príčin hospitalizácie a pohlavia v rokoch 1994 až 2004

Vybrala som päť skupín chorôb, ktoré sú najčastejšími príčinami hospitalizácie mužov v celom časovom horizonte rokov 1994 až 2004 a navzájom som porovnala vývoj hospitalizácie mužov a žien.

3.1. Choroby obehovej sústavy

Časový vývoj hospitalizácie na choroby obehovej sústavy je u mužov a žien identický. Rozdiel je v tom, že muži dosahujú vyššie hodnoty. V r. 1994 bolo hospitalizovaných 2 335 mužov a 2 229 žien na 100 000 obyv. (to bola minimálna hodnota u oboch pohlaví). Do r. 1999 tento počet u oboch pohlaví kontinuálne narastal, v r. 1999 mierne klesol a od r. 2000 sa zaznamenal opätovný nárast do r. 2002, kedy opäť poklesol a rast nastal až od roku 2004. V r. 2004 bolo hospitalizovaných 2 952 mužov a 2 902 žien na 100 000 obyv. Maximálna hodnota bola zaznamenaná v r. 2001, pre mužov 3 005/100 000 a pre ženy 2 945/100 000.

3.2. Choroby tráviacej sústavy

Časový vývoj hospitalizácie na choroby tráviacej sústavy je u mužov a žien veľmi podobný. Muži dosahujú vyššie hodnoty.

Muži : V r. 1994 bolo hospitalizovaných 2 136 mužov na 100 000 obyv., následne počet nepatrne poklesol, ale od r. 1996 zaznamenávame nárast (maximálna hodnota v r. 1997, a to 2 244/100 000) až do r. 1998, odkedy znova poklesáva, v r. 2000 nastal mierny nárast, avšak od r. 2002 sledujeme pokles až do r. 2004, kedy dosiahol v sledovanom období minimálnu hodnotu 1 956/100 000.

Ženy : V r. 1994 bolo hospitalizovaných 1 980 žien na 100 000 obyv. U žien počet hospitalizácií na choroby tráviacej sústavy kontinuálne vzrastal (maximálna hodnota v r. 1998, a to 2 148/100 000) do roku 1999, kedy nastal pokles, už v r. 2000 opäť nárast a od roku 2002 sledujeme pokles až do r. 2004, kedy dosiahol v sledovanom období minimálnu hodnotu 1 854/100 000.

3.3. Poranenia a otravy a niektoré iné následky vonkajších príčin

V tejto XIX. kapitole skupín chorôb je tradične veľký rozdiel medzi pohlaviami, čo sa týka početnosti v celom sledovanom období. Vývoj v čase je veľmi podobný.

Muži : V r. 1994 bolo hospitalizovaných 1 668 mužov na 100 000 obyv., čo bola minimálna hodnota, následne počet vzrastal až do r. 2000, kedy sa zaznamenal pokles, v r. 2001 počet vzrástol na maximum 1 979/100 000. V r. 2002 počet poklesol a od roku 2003 narastal. V r. 2004 dosiahli hospitalizácie mužov 1 943/100 000.

Ženy : V r. 1994 bolo hospitalizovaných 821 žien na 100 000 obyv., čo bola minimálna hodnota, následne počet vzrastal až do r. 2000, kedy sa zaznamenal pokles, od r. 2001 počet narastal a v r. 2004 dosiahol maximálnu hodnotu 1 089/100 000.

3.4. Nádory

Z analyzovaných skupín chorôb iba v tejto dosahujú ženy vyššie hodnoty než muži. Vývoj v čase je u oboch pohlaví identický.

Muži : V r. 1994 bolo hospitalizovaných 1 226 mužov na 100 000 obyv., čo bola minimálna hodnota, následne počet vzrastal až do r. 1999, vtedy sa zaznamenal pokles, od r. 2000 sledujeme nárast (v r. 2001 dosiahli maximálnu hodnotu 1 711/100 000) až do r. 2002, keď počet opäť poklesol a nárast sa zaznamenal až v r. 2004 (1 608/100 000).

Ženy : V r. 1994 bolo hospitalizovaných 1 364 žien na 100 000 obyv., čo bola minimálna hodnota, následne počet vzrastal (maximálna hodnota v r. 1998, a to 1 967/100 000) až do r. 1999, vtedy sa zaznamenal pokles, od r. 2000 sledujeme nárast až do r. 2002, keď počet opäť poklesol a nárast sa zaznamenal až v r. 2004 (1 791/100 000).

3.5. Choroby dýchacej sústavy

Muži dosahujú vyššie hodnoty ako ženy.

Muži : V r. 1994 bolo hospitalizovaných 1 928 mužov na 100 000 obyv., následne počet narastal (maximum v r. 1997, a to 1 998/100 000) do r. 1998, od toho roku bol zaznamenaný pokles, nárast prechodne v r. 2001, od r. 2002 pokles, pričom minimálna hodnota bola zistená v r. 2003 (1 568/100 000). V r. 2004 počet hospitalizovaných vzrástol na 1573/100 000.

Ženy : V r. 1994 bolo hospitalizovaných 1 492 žien na 100 000 obyv., potom počet narastal (maximum v r. 1997, a to 1 529/100 000) do roku 1998, od toho roku počet klesal (minimum 1 197/100 000 v r. 2003) a až v r. 2004 nepatrne vzrástol na 1 223/100 000.

4. Veková a pohlavná štruktúra hospitalizovaných v nemocniciach v roku 1994 a v roku 2004

4.1. Veková štruktúra hospitalizovaných mužov v r. 2004

Do 1 roku života bolo v r. 2004 hospitalizovaných 9,6 % mužov zo všetkých hospitalizovaných mužov. Vo vekovej kategórii 1 – 4 podiel poklesol na 4,5 % a v nasledujúcich vek. kategóriách postupne narastal. Vo vek. kategórii 45 – 54 dosiahli hospitalizácie podiel 14,6 % a v nasledujúcej vek. kategórii 55 – 64 maximum 15,2 %. V ďalších vek. kategóriách sú podiely stále významné, ale klesajú až do veku 85 +, kedy je podiel hospitalizovaných mužov najmenší (1,4 %).

Porovnanie rokov 1994 a 2004

V r. 2004 sa znížil podiel hospitalizovaných mužov v prvých troch vek. kategóriách (0 – 14 rokov) a tiež vo vek. kategórii 35 – 44 rokov. Napr. vo vek. kategórii do 1 roku podiel poklesol z 12,7 % (1994) na 9,6 % (2004). Naopak vo vek. kategóriách 45 – 84 rokov nastal oproti roku 1994 nárast, napr. v r. 1994 bolo hospitalizovaných 12,7 % mužov vo vek. kategórii 55 – 64 rokov, o 10 rokov neskôr to bolo už 15,2 %. Aj to je jeden z prejavov starnutia populácie.

4.2. Veková štruktúra hospitalizovaných žien v r. 2004

Do 1 roku života bolo v r. 2004 hospitalizovaných 6,2 % žien zo všetkých hospitalizovaných žien. Vo vekovej kategórii 1 – 4 podiel poklesol na 2,6 % a v nasledujúcich vek. kategóriách

narastal na maximum 17,4 % vo vek. kategórii 25 – 34 rokov. V ďalších vek. kategóriách sa podiel hospitalizovaných striedavo zmenšuje a zväčšuje až do veku 85 +, kedy je podiel hospitalizovaných žien najmenší (2,1 %).

Porovnanie rokov 1994 a 2004

Oproti roku 1994 nastal v r. 2004 pokles vo vek. kategóriách 0 – 24 ročných, pričom najvýraznejšie u 15 – 24 ročných, kde podiel hospitalizovaných žien v r. 1994 bol 18,6 % a v r. 2004 iba 11,8 %. Vo vek. kategórii 25 – 34 podiel vzrástol zo 16,6 % (1994) na 17,4 % (2004). Zatiaľ čo v r. 1994 bolo maximum hospitalizácie žien u 15 – 24 ročných, v r. 2004 toto prvenstvo prevzala vek. kategória 25 – 34 ročných. Je to dôsledok posunu maxima plodnosti do vek. kategórie 25 – 29 ročných (v r. 2004 bol priemerný vek pri pôrode 27,2 rokov). Pokles oproti r. 1994 nastal už iba u 35 – 44 ročných. Vo vek. kategórii 45 – 85+ sa zväčšil podiel hospitalizovaných žien, pričom najviac u 75 – 84 ročných, kde podiel vzrástol zo 5,4 % (1994) na 11,7 % (2004).

4.3. Štruktúra hospitalizovaných podľa pohlavia

Keď porovnám vekovú štruktúru hospitalizovaných mužov a žien v roku 1994 a 2004 navzájom, tak sa ukáže iba jedna zmena v čase, a to vo vek. kategórii 75 – 84 ročných v r. 2004 podiel hospitalizovaných žien prevýšil podiel mužov. V ostatných vek. kategóriách ostali trendy rovnaké : väčší podiel hospitalizovaných mužov než žien je vo veku 0 – 14 a 45 – 74. Vo veku 15 – 44 a 85 + je väčší podiel hospitalizovaných žien ako mužov. To teda dokazuje väčšiu zraniteľnosť chlapcov ako dievčat a mužskú nadúmrtnosť.

5. Záver

Medzi prvých päť najpočetnejších hospitalizácií u oboch pohlaví patria ch. obehovej sústavy, ch. tráviacej sústavy a nádory. Ostatné sa odlišujú, dokumentujúc odlišnú biologickú funkciu muža a ženy, u mužov to boli poranenia, otravy a niektoré iné následky vonkajších príčin a ch. dýchacej sústavy. U žien - ťarchavosť, pôrod a popôrodie a faktory ovplyvňujúce zdravotný stav a styk so zdravotníckymi službami. Vývoj počtu hospitalizovaných v rokoch 1994 – 2004 je u mužov a žien veľmi podobný. Iba pri nádorových ochoreniach dosahujú ženy vyššie hodnoty. Najväčší rozdiel medzi mužmi (1,8 krát vyššie hodnoty) a ženami je pri poraneniach a otravách … . Počet hospitalizovaných vzrástol pri ch. obehovej sústavy (najvýraznejšie), nádoroch, poraneniach a otravách … . Naopak počet hospitalizovaných klesol pri ch. tráviacej sústavy a ch. dýchacej sústavy.

Pri porovnávaní rokov 1994 a 2004 možno sledovať u oboch pohlaví znižovanie podielu hospitalizovaných v najnižších vek. kategóriách (0 – 14 rokov) a ich nárast vo vyšších vek. kategóriách (45 – 85 + rokov). Väčšiu „zraniteľnosť“ chlapcov dokumentuje vyšší podiel hospitalizovaných mužov ako žien vo veku 0 – 14 rokov. Vo veku 15 – 44 rokov výrazne dominujú ženy, čo súvisí s ťarchavosťou. Vo veku 45 – 74 rokov prevládajú muži, v tomto období sa postupne v populácii zväčšuje podiel žien, a tak vo vek. kategóriách 75 – 85 + rokov je hospitalizovaný vyšší podiel žien.

Zdroj:

Národné centrum zdravotníckych informácií, 2005: Hospitalizovaní, ošetrovacie dni a priemerný ošetrovací čas v nemocniciach podľa skupín diagnóz základného ochorenia (MKCH-10), pohlavia a vek. skupín. Interné pramene.

Kríž, J. a kol.(1997): Jak jsme na tom se zdravím. Státní zdravotní ústav, Praha, pp. 90

Krajčír, A. (1970). Vývoj a súčasný stav medicínskej geografie. Geografický časopis, 22, 51 – 65.

Krajčír, A. (1971). Teoretická problematika medicínskej geografie. Geografický časopis, 23, 339 – 353.

Autorka

Bc. Magdaléna Pullmannová

Katedra humánnej geografie, Prírodovedecká fakulta Univerzity Komenského, Bratislava

Partizánska 29, Bánovce nad Bebravou, 95701, pulim@zoznam.sk

Modelování rozdělení výšky pojistných škod v systému STATGRAPHICS Plus

Jana Rybářová[1]

1. Úvod

Pro pojišťovnu je modelování rozdělení výšky škod v různých typech pojištění velmi důležité. Znalost správného rozdělení pravděpodobností základního souboru ji může být užitečná např. při výpočtu pojistného, rozhodování o výšce pojistných rezerv, či v otázkách zajištění. Následujícím příspěvkem chceme popsat, jakým způsobem je možné rozdělení výšky škod modelovat a předvést to na konkrétním souboru výšek škod u havarijního pojištění motorových vozidel použitím statistického programového balíku Statgraphics Plus.

2. Testy dobré shody

Pro výběr nejvhodnějšího rozdělení výšky pojistných plnění nám systém Statgraphics Plus poskytuje testy dobré shody: χ^2- test (Chí-kvadrát test) a Kolmogorův-Smirnovův test. U těchto testů se hypotézy týkají rozložení pravděpodobností základního souboru. Testujeme shodu skutečného rozdělení pravděpodobností základního souboru s rozdělením předpokládaným (teoretickým).

Testujeme hypotézu ve tvaru: Statistický soubor má rozdělení pravděpodobností s hustotou $f(x;\Theta)$. Pro to, abychom si utvořili představu o typu rozdělení pravděpodobností, bývá účelné si před stanovením hypotézy sestavit histogram nebo polygon četností náhodného výběru. Při formulaci hypotézy pak nemusíme brát v úvahu rozdělení pravděpodobností s podstatně odlišným průběhem.

- *Chí-kvadrát test*

Testujeme hypotézu H_0: náhodná proměnná X má rozdělení s hustotou $f(x;\Theta)$. Testovací kritérium má tvar:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (1)$$

s *k - 1- p* stupni volnosti, kde p je počet odhadnutých parametrů předpokládaného teoretického rozdělení, O_i jsou empirické, skutečně zjištěné četnosti hodnot x_i diskrétní proměnné nebo intervalů $(x_{i-1}, x_1\rangle$ hodnot x spojité proměnné X a E_i jsou příslušné teoretické, očekávané četnosti, vyjádřené vztahem

$$E_i = np_i. \qquad (2)$$

Přičemž n je rozsah výběrového souboru a p_i je pravděpodobnost hodnoty x_i diskrétní proměnné s předpokládaným rozdělením, tedy $p_i = P(x_i) = P(X = x_i)$, resp. pravděpodobnost intervalu hodnot $x \in (x_{i-1}, x_1\rangle$ spojité proměnné,tj.

$$p_i = P(x_{i-1} < X \le x_i) = F(x_i) - F(x_{i-1}) \qquad (3)$$

kde F(x) je distribuční funkce předpokládaného teoretického rozdělení.

Čím větší je neshoda mezi předpokládaným a skutečným rozdělením pravděpodobností základního souboru, tím budou větší rozdíly mezi skutečnými a teoretickými četnostmi. Proto budeme nulovou hypotézu zamítat při velkých hodnotách těchto rozdílů a tudíž i velké hodnotě testovacího kritéria χ^2. Nejčastěji se při testu používá hladina významnosti $\alpha = 0{,}05$,

[1] Fakulta ekonomicko-správní, Univerzita Pardubice

tedy $\chi^2_{1-\alpha} = \chi^2_{0,95}$ je 95. percentil χ^2 rozdělení s k-1-p stupni volnosti. Kritickou oblast tedy volíme ve tvaru $\chi^2_{1-\alpha} < \chi^2$.

- *Kolmogorův-Smirnovův test*

Tento test je založen na porovnání rozdílů mezi distribuční funkcí ověřovaného a výběrového rozložení pravděpodobností. Pokud náhodný výběr x_1, x_2, …, x_n pochází ze spojitého rozdělení specifikovaného distribuční funkcí F(x), a rozsah n výběrového souboru není dostatečný na použití χ^2-testu, potom pro test hypotézy H_0: náhodná proměnná X má rozdělení s distribuční funkcí F(x) můžeme použít Kolmogorův-Smirnovův test dobré shody. Tento test na rozdíl od χ^2-testu vychází z netříděných, vzestupně uspořádaných výběrových údajů: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Uspořádanému náhodnému výběru ($x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$) přísluší výběrová (empirická) distribuční funkce, definovaná vztahem:

$$F_n(x) = \begin{cases} 0 & x \leq x_1 \\ \dfrac{j}{n} & x_{(j)} < x \leq x_{(j+1)} \quad j = 1,2,...,n-1 \\ 1 & x > x_{(n)} \end{cases} \tag{4}$$

Kolmogorova-Smirnovova testovací charakteristika má tvar:

$$d_n = \sup_x |F_n(x) - F(x)|, \tag{5}$$

což je maximální absolutní odchylka výběrové (empirické) distribuční funkce $F_n(x)$ od spojité distribuční funkce F(x), kterou předpokládá nulová hypotéza. Při určování hodnoty testovacího kritéria d_n zkoumáme v bodech $x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$ absolutní odchylky $\left|F_n(x_{(j)}) - F(x_{(j)})\right|$ a $\left|F_n(x_{(j+1)}) - F(x_{(j)})\right|$ pro j = 1,2,…,n, přičemž $F_n(x_{(n+1)}) = 1$. Největší z těchto odchylek považujeme za hodnotu testovacího kritéria d_n.

Hypotézu, že náhodný výběr pochází z rozdělení s distribuční funkcí F(x) přijmeme tehdy, když $d_n < d_{n;1-\alpha}$, přičemž α = 0,05, resp. α = 0,01 je hladina významnosti.

3. Modelování výšky škod u havarijního pojištění motorových vozidel

Následující část textu se zabírá modelováním rozdělení výšky pojistných škod u havarijního pojištění motorových vozidel. Základní údaje, poskytla nejmenovaná pojišťovna.

Tabulka 1. Výška pojistných škod havarijního pojištění motorových vozidel v Kč.

412	2704	4153	7255	11071	18849	35147	78738
494	2704	4319	7600	11315	19158	39515	84653
534	2766	4428	7635	11622	19952	39616	88487
655	2923	4778	7647	12320	20466	40817	98705
895	2950	4850	8218	12353	20945	40988	99766
1462	3014	4924	8340	12451	21905	45911	117842
1697	3104	4940	8482	12602	22410	46371	128324
1838	3168	5148	8703	13972	22563	49191	143822
1840	3392	5277	8913	14556	27548	49607	145226
1865	3680	5343	8965	14988	27575	53404	196769
2084	3692	5343	8976	16110	28067	53610	
2190	3722	5344	9042	16464	31663	62099	
2335	3774	6887	9476	16494	31863	64473	
2550	3791	6958	9778	17317	32775	68168	
2685	4060	7162	10210	17368	33445	75690	

Vhodné prostředí pro modelování rozdělení výšky pojistných škod poskytuje program Statgraphics Plus, ovšem je možno využít i modelování v Excelu. Statgraphics nabízí výběr z mnoha druhů rozdělení pravděpodobností. V případě havarijního pojištění motorových vozidel jsou pro nás zajímavá pravostranně zešikmená rozdělení, která naznačují vysoký výskyt nižších škod. Mezi ně patří exponenciální, Paterovo, Weibullovo gama a lognormální rozdělení.

Podle metod dobré shody bylo zjištěno, že nejvhodnějším teoretickým rozdělením výšky pojistných škod je lognormální rozdělení se střední hodnotou 26 517,1 a směrodatnou odchylkou 60 044,4.

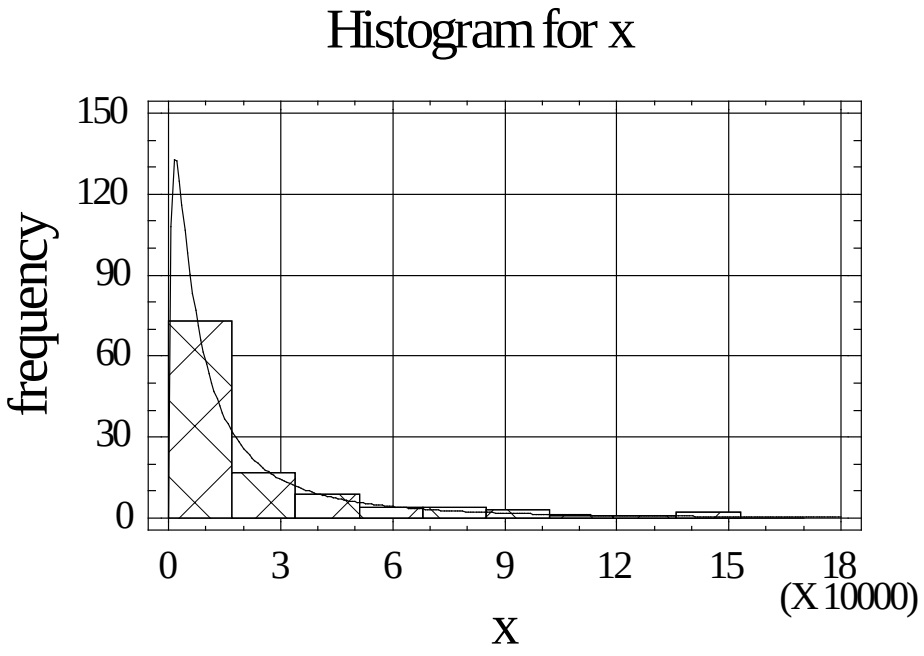
Tabulka 2: Test dobré shody pro lognormální rozdělení pravděpodobností (Statgraphics)

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
	at or below	17000,0	73	72,93	0,00
	17000,0	34000,0	17	19,58	0,34
	34000,0	51000,0	9	8,31	0,06
	51000,0	85000,0	8	7,04	0,13
above	85000,0		8	7,13	0,11

Chi-Square = 0,634891 with 2 d.f. P-Value = 0,728006

Estimated Kolmogorov statistic DPLUS = 0,0624723
Estimated Kolmogorov statistic DMINUS = 0,042539
Estimated overall statistic DN = 0,0624723
Approximate P-Value = 0,760514



Graf 1. Lognormální rozdělení pravděpodobností výšky škod základního souboru

Program Statgraphics poskytuje další funkce, které umožňují detailní popis rozdělení. Jsou jimi Tail areas a Critical values. Tail areas umožňuje výpočet hodnot distribuční funkce definované vztahem $F(x) = P(X \leq x)$, pro každou reálnou hodnotu x. Critical values poskytuje výpočet kvantilů. Je to inverzní procedura k Tail areas. Pro zvolenou hodnotu distribuční funkce F(x) najde právě takové x, pro které platí $F(x) = P(X \leq x)$.

Tabulka 3. Výstup z procedury Tail Areas a Critical Values

Tail Areas for x
area below 5000,0 = 0,285717
area below 30000,0 = 0,777822
area below 60000,0 = 0,89967
area below 85000,0 = 0,938021
area below 110000,0 = 0,958173

Critical Values
area below 98103,7 = 0,95
area below 113128,0 = 0,96
area below 134787,0 = 0,97
area below 170129,0 = 0,98
area below 245570,0 = 0,99

V tabulce lze např. vyčíst, že pravděpodobnost, že výška škod nepřesáhne 60 000 je 0,89967 nebo že 95% škod nabývá hodnoty do 98103,7 Kč. Do hodnot v tabulce Critical Values byly vybrány vysoké percentily, které popisují pravděpodobnost vzniku nejvyšších škod.

4. Závěr

Pro modelování výšky škod byly využity testy dobré shody, konkrétně chí-kvadrát test a Kolmogorův-Smirnovův test. V programu Statgraphics bylo zjištěno, že pro popis rozdělení výšky pojistných škod u havarijního pojištění motorových vozidel se jako nejvhodnější jeví lognormální rozdělení. V závěru byl detailněji popsán tvar pravého konce rozdělení, který ukazuje na to, že vysoké škody jsou sice málo pravděpodobné, ale nejsou nemožné.

5. Literatura

CIPRA, T. Teorie rizika v neživotním pojištění. Praha: MFF UK a Česká pojišťovna, 1991.

CURRIE, I.D.: Loss Distributions. London and Edinburgh: Institute of Actuaries and Faculty of Actuaries, 1993.

KUBANOVÁ, J. Statistické metody pro ekonomickou a technickou praxi. Bratislava: STATIS, 2003. 247 s. ISBN 80-85659-31-X.

PACÁKOVÁ, V. Aplikovaná poistná štatistika. Bratislava: IURA EDITION, 2004. 261 s. ISBN 80-8078-004-4.

Adresa autora:
Jana Rybářová, Bc.
R. Jesenské 282
500 09 Hradec Králové
jana.ryb@seznam.cz

Stanovení pojistného v neživotním pojištění

Denisa Sehnoutková[1]

1. Úvod

Cílem této práce je stanovit pojistné v havarijním pojištění. Budou použity dva způsoby výpočtu. Nejprve vypočítáme brutto pojistné pomocí kolektivního rizika a poté pomocí $E(S) + 3\sigma$.

2. Lognormální rozdělení – LN (μ; σ^2)

Toto kladně zešikmené rozdělení, při kterém velmi malé hodnoty sledované náhodné proměnné mají malou pravděpodobnost, středně velké jsou nejpravděpodobnější a vysoké se stávají méně pravděpodobné, přičemž i velmi vysoké hodnoty mají kladnou pravděpodobnost, má široké využití v ekonomické oblasti. Je vhodné i na modelování výšky pojistných plnění např. v havarijním pojištění, v požárním pojištění zděných budov a pojištění proti vichřicím.

Hovoříme, že X má lognormální rozdělení LN (μ; σ^2), jestliže ln(X) má normální rozdělení N (μ; σ^2).

Z této definice vyplývá funkční vyjádření hustoty lognormálního rozdělení ve tvaru

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0. \tag{2.1}$$

Označme Y = ln X. Protože $Y \sim N(\mu; \sigma^2)$, hustotu g (y) vyjadřuje vztah

$$g(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < +\infty . \tag{2.2}$$

Nechť F(x) je distribuční funkce proměnné X a G (y) je distribuční funkce proměnné Y. Platí

$$F(x) = P(X < x) = P(Y < \ln x) = G(\ln x). \tag{2.3}$$

Odtud

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{x} g(\ln x). \tag{2.4}$$

Distribuční funkce rozdělení LN (μ; σ^2) má funkční vyjádření

$$F(x) = \begin{cases} 0 & \text{když } x \le 0 \\ \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_0^{\ln x} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy & \text{když } x > 0 \end{cases} \tag{2.5}$$

Základní charakteristiky rozdělení LN (μ; σ^2):

$$E(X) = e^{\mu+\frac{\sigma^2}{2}} \tag{2.6}$$

$$D(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \tag{2.7}$$

$$\gamma_1 = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1} \tag{2.8}$$

$$E(X^k) = e^{k\mu+\frac{k^2\sigma^2}{2}} \tag{2.9}$$

$$M_x(z) = e^{\mu z+\frac{\sigma^2 z^2}{2}} \tag{2.10}$$

Tyto charakteristiky vyplynou z možnosti jednoduchého vyjádření začátečních momentů rozdělení LN (μ; σ^2) pomocí momentů N (μ; σ^2).

[1] Fakulta ekonomicko-správní, Univerzita Pardubice

Když $Y = \ln X \sim N(\mu; \sigma^2)$, potom momentová vytvářející funkce $M_Y(z)$ má tvar

$$M_Y(z) = E(e^{zY}) = e^{\mu z + \frac{\sigma^2 z^2}{2}} \tag{2.11}$$

Nyní použijeme jednoduchou, velmi výhodnou substituci. Protože $X = e^Y$, platí $E(X) = E(e^Y)$, což je vlastně $M_Y(z = 1)$. Proto platí

$$E(X) = e^{\mu + \frac{\sigma^2}{2}} \tag{2.12}$$

Analogicky

$$E(X^2) = E(e^{2Y}) = M_Y(z = 2) = e^{2\mu + 2\sigma^2} \tag{2.13}$$

Všeobecně

$$E(X^k) = E(e^{kY}) = M_Y(z = k) = e^{k\mu + \frac{\sigma^2}{2}k^2} \tag{2.14}$$

Z možnosti vyjádření centrálních momentů pomocí začátečních momentů vyplynou vztahy pro všechny uvedené charakteristiky rozdělení $LN(\mu; \sigma^2)$.

3. Složené Poissonovo rozdělení

Nechť S je celkové pojistné plnění, přičemž N má Poissonovo rozdělení s parametrem λ, symbolicky $N \sim Po(\lambda)$, a F(x) je společná distribuční funkce identicky rozdělených individuálních pojistných plnění X_i.

Budeme hovořit, že S má složené Poissonovo rozdělení s parametrami λ a F(x) a označovat $CoPo(\lambda, F(x))$. Potom platí

$$E(N) = D(N) = \lambda \quad \text{a} \quad M_N(z) = e^{\lambda(e^z - 1)} \tag{3.1}$$

Základní charakteristiky složeného rozdělení $CoPo(\lambda, F(x))$:

$$E(S) = \lambda m_1 \tag{3.2}$$

$$D(S) = \lambda(m_2 - m_1^2) + \lambda m_1^2 = \lambda m_2 \tag{3.3}$$

$$M_S(z) = e^{\lambda[M_x(z)-1]} \tag{3.4}$$

Vztah (3.4) využijeme na vyjádření třetího centrálního momentu proměnné S:

$$\begin{aligned} \mu_3(S) &= E[(S - E(S))^3] = E[(S - \lambda m_1)^3] = \frac{\partial^3}{\partial z^3} \ln M_S(z)\Big|_{z=0} = \\ &= \frac{\partial^3}{\partial z^3}(\lambda M_x(z) - 1) = \lambda \frac{\partial^3}{\partial z^3} M_x(z)\Big|_{z=0} = \lambda m_3 \end{aligned} \tag{3.5}$$

Vyjádříme ještě koeficient šikmosti γ_1 rozdělení $CoPo(\lambda, F(x))$:

$$\gamma_1 = \frac{\mu_3(S)}{[D(S)]^{\frac{3}{2}}} = \frac{\lambda m_3}{(\lambda m_2)^{\frac{3}{2}}} > 0. \tag{3.6}$$

Protože $\lambda > 0$ a $X_i \geq 0$, platí též $\gamma_1 > 0$. Složené Poissonovo rozdělení je tedy vždy pozitivně zešikmené, dokonce i tehdy, když jsou individuální pojistná plnění negativně zešikmené. Pro velké hodnoty parametru λ se rozdělení $CoPo(\lambda, F(x))$ stává symetričtějším, protože podle (3.6) platí, že $\gamma_1 \to 0$, když $\lambda \to \infty$.

4. Aproximace normálním rozdělením

Předpokládáme, že všechno, co víme anebo můžeme odhadnou o kolektivním riziku S, jsou základní charakteristiky – střední hodnota $E(S) = \mu$ a rozptyl $D(S) = \sigma^2$. Protože S = je součtem nezávislých a identicky rozdělených náhodných proměnných, podle centrální limitní věty se nabízí normální aproximace rozdělení S. Tedy pro libovolné s platí

$$G(s) = P(S \leq s) = P\left(\frac{S - \mu}{\sigma} \leq \frac{s - \mu}{\sigma}\right) \sim \Phi\left(\frac{s - \mu}{\sigma}\right) \tag{4.1}$$

Čím větší je počet N pojistných plnění, tím je aproximace G(s) pomocí distribuční funkce normovaného normálního rozdělení lepší.

Při odvozování koeficientu šikmosti složeného Poissonova rozdělení jsme dospěli k závěru, že čím větší je hodnota λ, tím je rozdělení S symetričtější, tím lepší je aproximace normálním rozdělením.

5. Výpočet brutto pojistného

E(X)	26517,1
σ(X)	60044,4
D(X)	3605329971
m1	26517,1
m2	3,099364661
m3	3,583377882

$e^{\mu+(\sigma 2)/2} = 26517,1$

$3605329971 = e^{2\mu+\sigma^2}(e^{\sigma 2} - 1)$

$D(X) = \nu_2 - E(X)^2 = e^{2\mu+2\sigma^2} = 26517,1^2$

$2(\mu + \sigma^2) = 22,18385253$

$\mu + \sigma^2 = 11,09192627$

$\mu + (\sigma^2)/2 = 10,185545 \quad * 2$

$2\mu + \sigma^2 = 20,37109$

$\mu = 9,279$

$\sigma^2 = 1,812926$

$\sigma = 1,34645$

λ^*	1400
E(S)	37123940
D(S)	4339,110525
μ_3(S)	5016,729035
γ_1	0,017551716

a) $S_{0,95} = 37124048,36$

b) $E(S) + 3\sigma = 37124137,62$

6. Závěr

Pojišťovna musí na pojistném vybrat alespoň Sk 37124137,62, aby byla schopna dostát svým závazkům z titulu pojistného plnění.

7. Literatura

CIPRA, T.: Pojistná matematika v praxi, Edice HZ, Praha 1994.

PACÁKOVÁ, V.: Aplikovaná poistná štatistika, EKONÓMIA, Bratislava 2004, s. 65 - 111.

Adresa autora:
Denisa Sehnoutková
Starý Mateřov 125
530 02 Pardubice
dseha@email.cz

Paterovo rozdělení výšky škod v neživotním pojištění

Lucie Šedová[1]

1. Úvod

Cílem práce je najít vhodné rozdělení výšky pojistných plnění při havarijním pojištění motorových vozidel na základě znalostí výšky pojistných plnění z portfolia pojistek. Práce je zaměřena na ověření, zda *Paretovo rozdělení* je vhodným modelem výšky pojistných plnění. K ověření, zda výšky individuálních pojistných plnění mají tento typ rozdělení použijeme testy dobré shody v tabuľkovém procesoru Excel.

2. Paretovo rozdělení Pa (α,λ)

Vzhledem k nedostatkům, které má exponenciální rozdělení při modelu rozdělení pojistných plnění je třeba najít takové rozdělení, které při pravděpodobnosti nejvyšších hodnot konverguje k nule pomaleji než při rozdělení exponenciálním. Takovým rozdělením je rozdělení Paretovo.

Distribuční funkci Paretova rozdělení vyjadřuje vztah:

$$F(x) = 1 - \left(\frac{\lambda}{\lambda + x} \right)^{\alpha} \tag{2.1}$$

Náhodná proměnná X má Paretovo rozdělení Pa (α,λ) právě tehdy, když funkční vyjádření její hustoty pravděpodobnosti má tvar

$$f(x) = \begin{cases} \dfrac{\alpha \lambda^{\alpha}}{(\lambda + x)^{\alpha+1}} & \text{pro } x > 0 \quad \alpha > 0 \quad \lambda > 0 \\ 0 & \text{pro } x \le 0 \end{cases} \tag{2.2}$$

Paretovo rozdělení se používá jako rozdělení pojistných plnění nejvíce při existenci extrémních hodnot v nemocenském pojištění a v pojištění proti požáru.

3. Ověření Paretova rozdělení s parametry, odhadnutými metodou momentů

Abychom zjistili, zda daný soubor má tento typ rozdělení, musíme nejprve odhadnout parametry α, λ *Paretova rozdělení*. Použijeme metodu momentů a dosadíme do vzorců

[1] Fakulta ekonomicko-správní, Univerzita Pardubice

$$\tilde{\alpha} = \frac{2s^2}{s^2 - \bar{x}^2} \qquad \tilde{\lambda} = (\tilde{\alpha} - 1)\bar{x} \tag{2.3}$$

Dostaneme odhadnuté hodnoty parametrů:

alfa	3,9608
lambda	72583,36

Pro zjištění, zda je pro naše data daný typ rozdělení vhodný, slouží testy dobré shody. Vypočítáme nejprve Pearsonovu chí-kvadrát testovací charakteristiku pomocí vzorce

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{2.4}$$

přičemž O_i jsou skutečně získané početnosti intervalů hodnot x_i a E_i jsou příslušné teoretické, očekávané početnosti, vyjádřené vztahem

$$E_i = np_i. \tag{2.5}$$

kde n je rozsah výběrového souboru a p_i je pravděpodobnost hodnoty x_i.

Tabulka 1: Chí-kvadrát test pro Paretovo rozdělení (odhad metodou momentů)

x	n	F(x)	p	E	chi^2
8000	49	0,3390832	0,3390832	38,994566	2,5672
20000	29	0,6186183	0,2795351	32,146541	0,3080
50000	21	0,8745288	0,2559105	29,429709	2,4146
100000	11	0,9676333	0,0931045	10,707015	0,0080
nad100000	5		0,0323667	3,722169	0,4387
					5,7365

Určíme kritickou hodnotu $\chi 2 = 5{,}991$ pro 2 stupně volnosti a $\alpha = 0{,}05$. a porovnáme ji s výsledkem chí-kvadrát testu. Hodnota testovacího kritéria 5,7365 je menší ako kritická hodnota, z čehož vyplývá, že P*aretovo rozdělení* s parametry, odhadnutými metodou momentů, je vhodným modelem výšky pojistných plnění.

Stejný závěr potvrzuje i výsledek Kolmogorova-Smirnovova testu, kde hodnota testovacího kritéria, tj. d = 0,11238 padla do oblasti přípustných hodnot.

3. Ověření Paretova rozdělení s parametry, odhadnutými metodou maximální věrohodnosti

Parametre, odhadnuté metodou momentů, jsme použili jako výchozí hodnoty pro odhad metodou maximální věrohodnosti. Použitím funkce Solver (řešitel) v Excelu jsme měnili parametr λ tak, že výraz (2.65) v publikaci Pácáková, V.: Aplikovaná poistná

štatistika, str.72, nadobude nulovou hodnotu. Získáme tak maximálně věrohodné odhady parametrů $\alpha = 2{,}0435$ a $\lambda = 28126{,}74$. Stejným postupem jako v tabulce 1 provedeme $\chi 2$ test dobré shody (tabulka 2).

Tabulka 2: Chí-kvadrát test pro Paretovo rozdělení (odhad metodou maximální věrohodnosti)

x	n	F(x)	p	E	chi^2
8000	49	0,4004127	0,3390832	38,994566	2,5672
20000	29	0,6663287	0,2659160	30,580340	0,0817
50000	21	0,8760235	0,2096948	24,114899	0,4023
100000	11	0,9548858	0,0788622	9,069158	0,4111
nad100000	5		0,0451142	5,188137	0,0068
					3,4692

Ve srovnání s tabulkou 1 je hodnota testovací charakteristiky 3,4692 nižší, teda Paretovo rozdělení s parametry, odhadnutými metodou maximální věrohodnosti, je lepším modelem pojistných škod při havarijním pojištění jako v případě odhadu metodou momentů. Stejný závěr vyplyne i z Kolmogorovova-Smirnovova testu, kde testovací charakteristika má hodnotu d = 0,066 ve srovnání s hodnotou 0,11238 v předchozím případě.

4. Závěr

Na základě výsledků testů dobré shody jsme zjistili, že Paretovo rozdělení s parametry, odhadnutými metodou momentů i metodou maximální věrohodnosti dobře modeluje pojistná plnění při havarijním pojištění. Lepším modelem je Paratův model s parametry, odhadnutými metodou maximální věrohodnosti.

Literatura

LINDA, B.: Využití rozdělení pravděpodobností v pojišťovnictví, Pardubice 1998.

PACÁKOVÁ, V.: Aplikovaná poistná štatistika, Iura Edition, Bratislava 2004.

STRAUB, B.: Non-Life Insurance Mathematics, Springer-Verlag, Zürich 1988.

Adresa autora:

Lucie Šedová

Koldín 97

565 01 Choceň

lucka.sedova@centrum.cz

Medzikultúrne správanie v manažmente (kultúrne dimenzie podľa Geerta Hofsteda)

Krystyna Šípošová[1]

Abstract: In the present world of expanding globalization and international trade it is crucial to be aware of cross-cultural differences among countries in order to be successful when doing business. Dutch professor Geert Hoffstede has identified five independent cultural dimensions based on detailed research in numerous countries and companies. Following his research, the aim of the project would be to find out whether geographical position of the countries has some impact on the way business is done and to illustrate the different business behavior using appropriate statistical methods.

Key words: five independent cultural dimensions, cross-cultural differences among countries, business behavior, SAS Enterprise Guide

1. Úvod

Holandský profesor Geert Hofstede realizoval zatiaľ najrozsiahlejšiu štúdiu o interkultúrnych rozdieloch a ich dopadoch na pracovný proces. Svoj výskum uskutočnil v rokoch 1967–1973. Autor administroval dotazníky hodnôt zamestnancom IBM v 64 krajinách. Na pracovníkov IBM sa zameral kvôli tomu, aby mal relatívne homogénnu vzorku účastníkov. Celkový počet dotazníkov dosiahol pôsobivé číslo 116 000. Nasledovali ďalšie doplňujúce štúdie na študentoch v 23 krajinách, vybranej vzorke inteligencie („elites“) v 19 krajinách, pilotoch v 23 krajinách, spotrebiteľoch v 15 krajinách a pracovníkov štátnej správy v 14 krajinách. Tieto štúdie identifikovali a potvrdili 5 nezávislých **dimenzií národných kultúrnych rozdielov,** ktoré budú v našom projekte figurovať ako kvantitatívne premenné.
Charakteristika vstupných premenných:

1. Dimenzia **odstup od moci** (**PDI -** power distance) ukazuje, do akej miery ľudia akceptujú nerovnosť v práci, v rodine a v spoločnosti ako celku.
2. Dimenzia **individualizmus – kolektivizmus** (**IDV -** individualism vs. collectivism) poukazuje na pevnosť väzby medzi jednotlivcami a spoločenskými skupinami. Vyjadruje, v akom rozsahu spoločnosť podporuje individualistické správanie a samostatné dosahovanie výsledkov.
3. Dimenzia **maskulinita – femininita** (**MAS -** masculinity) určuje, do akej miery spoločnosť podporuje tradičný maskulínny pracovný model. Poukazuje na maskulínne črty ako úspech, asertívnosť a výkon; alebo femininne črty ako solidarita, osobné vzťahy, služby, kvalita života.
4. **Vyhýbanie sa neistote** (**UAI -** uncertainty avoidance) poukazuje na preferenciu štruktúrovaných alebo neštruktúrovaných situácií. Prejavy vysokej miery vyhýbania sa neistote sú dôraz na formálne pravidlá a jasné, predpovedateľné a čitateľné štruktúry, inštitúcie a vzťahy. Naopak, nízka miera vyhýbania sa neistote znamená menej formálne vzťahy, ľudia radi riskujú.
5. **Dlhodobá vs. krátkodobá orientácia** hovorí, do akej miery sú ciele spoločnosti v súlade s tradíciami.

[1] Krystyna Šípošová, študentka 3. ročníka Fakulty managementu Univerzity Komenského v Bratislave

2. Cieľ projektu a popis dát

V súčasnom svete globalizácie, kedy medzinárodný obchod pomaly smeruje k vytvoreniu jednotného svetového trhu, je priam nevyhnutné poznať, tolerovať ale často aj prispôsobiť sa kultúrnym rozdielom, ktoré majú veľký vplyv na spôsob vedenia manažmentu, pracovný proces a zvyky pri vyjednávaní a obchodovaní. Mnohé dôležité medzinárodné spolupráce, obchodné zmluvy či investície stroskotali na nedorozumeniach v dôsledku nevedomosti a netolerancii kultúrnych odlišností. Nadväzujúc na štúdiu Geerta Hofsteda o interkultúrnych rozdieloch by mal náš projekt nasledovnou analýzou prispieť k podrobnejšiemu vysvetleniu kultúrnych odlišností jednotlivých krajín a to prostredníctvom základných popisných štatistík a následne využitím zhlukovej analýzy.

Zdroj dát: **http://www.geert-hofstede.com/hofstede_dimensions.php**
Štatistická jednotka: **štát**
Rozsah súboru: **96** štatistických jednotiek – štátov sveta, 5 premenných (4 numerické a 1 charakterová)
Premenné:

- charakterová: kontinent,
- numerické: **PDI -** power distance, **IDV -** individualism vs. collectivism, **MAS -** masculinity, **UAI -** uncertainty avoidance

Piaty index (dlhodobá vs. krátkodobá orientácia) nebol analyzovaný, keďže nebol zistený pre všetky štáty. Premenné sú indexované a ich mernou jednotkou sú body od 0 do 105, kde 0 znamená nízku mieru daného indexu a 105 znamená vysokú mieru.

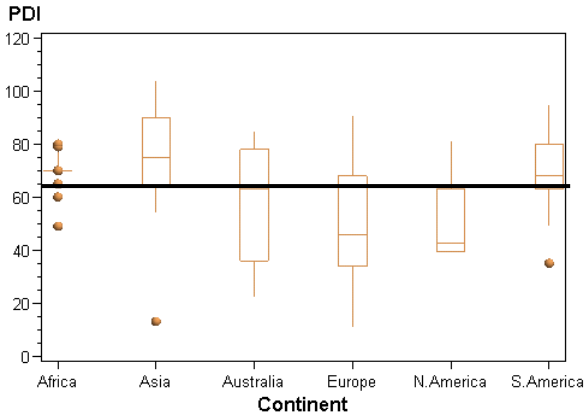
3. Výsledky štatistickej analýzy

Na úvodnú charakteristiku situácie medzikultúrneho správania boli využité základné popisné štatistiky. Na nasledovných box-plotoch (grafy 1 až 4) je znázornené porovnanie jednotlivých kultúrnych dimenzií medzi kontinentmi ako aj voči celosvetovému priemeru. (Poznámka: Vodorovná čiara v grafoch box-plotov znázorňuje celosvetový priemer pre sledovaný ukazovateľ.)

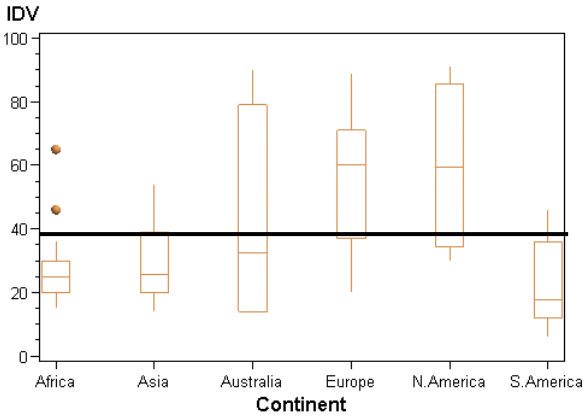
Porovnanie kultúrnych odlišností medzi jednotlivými kontinentmi je veľmi všeobecné a nie vždy smerodajné vzhľadom k tomu, že aj v rámci kontinentov ako takých sa prejavujú tieto rozdiely pod vplyvom mnohých socioekonomických, kultúrnych, náboženských a iných faktorov, ktoré vplývajú aj na správanie sa v manažmente týchto krajín.

Keďže diferenciácia krajín podľa kontinentov je veľmi široká a všeobecná, cieľom tohto projektu bolo jednotlivé krajiny, ktorých správanie sa v manažmente je čo najpodobnejšie, zlúčiť do zhlukov. Na základe vyšpecifikovaných zhlukov budeme vedieť lepšie porovnať a poukázať na kultúrne *podobnosti v rámci zhluku a na odlišnosti medzi jednotlivými zhlukmi.*

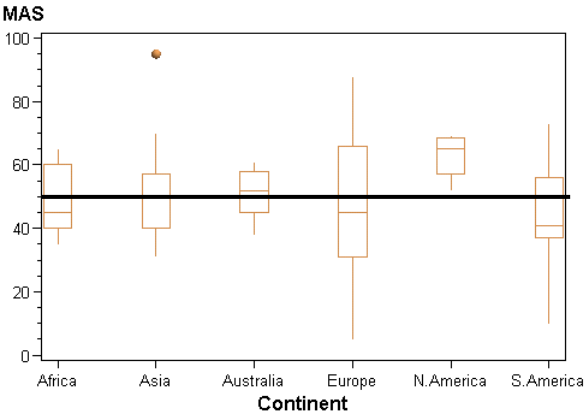
Analyzované ukazovatele sú kvantitatívne a spojité a nie sú korelované, spĺňajú podmienky pre uskutočnenie zhlukovej analýzy. Použili sme hierachický aglomeratívny zhlukovací postup prostredníctvom Systému SAS Enterprise Guide V4. Výsledné zhluky sme získali Wardovou metódou. Na základe vecnej analýzy problematiky a aj na základe analýzy hodnôt koeficientu determinácie (*R-Squared*) na grafe 5, sme sa rozhodli pre 12 zhlukov.



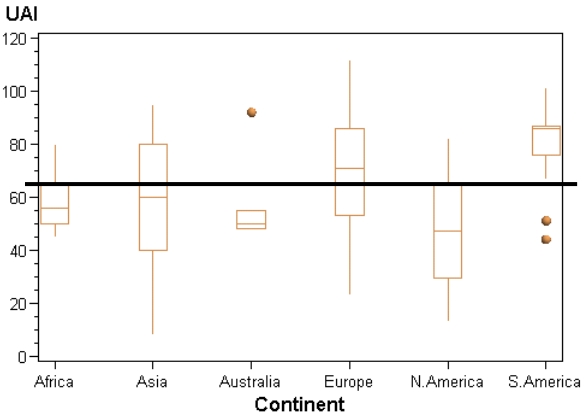
Graf 1: PDI pre kontinenty



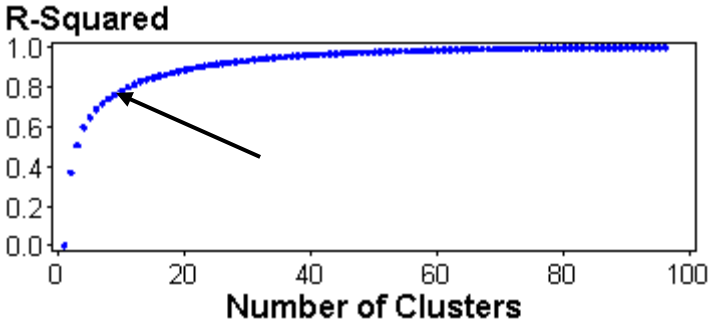
Graf 2: IDV pre kontinenty



Graf 3: MAS pre kontinenty



Graf 4: UAI pre kontinenty



Graf 5: Hodnoty koeficientu determinácie pre počty zhlukov

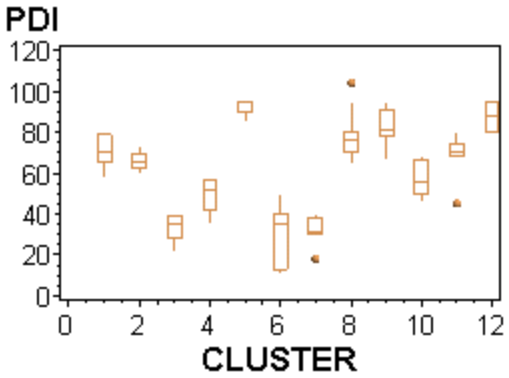
Porovnanie jednotlivých kultúrnych dimenzií, ktoré charakterizujú správanie v manažmente krajín v rámci vytvorených zhlukov, je znázornené na nasledujúcich box-plotoch (grafy 6 až 9). Prehľadne vidieť rozdiel medzi danými zhlukmi, čo vysvetľuje ich kultúrnu rôznorodosť. Naopak štáty v rámci jedného zhluku majú podobné postoje k nerovnosti v rodine, práci, ako aj v spoločnosti; do podobnej miery inklinujú k individualistickému alebo kolektívnemu správaniu; do podobnej miery prevažujú v krajinách klasické rodové rozdiely a ich roly v pracovnom procese a podobne podstupujú bezrizikové a vopred štruktúrované situácie.

Pravdaže je potrebné poznamenať, že rozdelenie do zhlukov nie je u všetkých štátov opodstatnené vzhľadom k tomu, že je prakticky nemožné, aby si boli krajiny podobné v rámci všetkých kultúrnych dimenzií súčasne. Niektoré krajiny zaradené do zhluku na základe

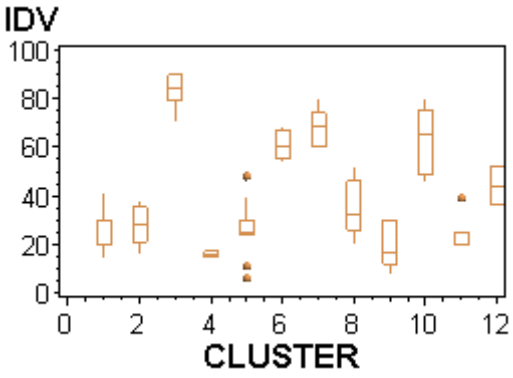
viacerých podobností môžu ovplyvniť správanie sa celého zhluku nesprávnym spôsobom (napríklad Japonsko v zhluku č. 10). Preto niektoré výsledky je potrebné interpretovať s nadhľadom a berúc do úvahy spomínanú skutočnosť.

Na grafe č. 6 je znázornený stupeň mocenskej vzdialenosti v jednotlivých zhlukoch. Najnižší mocenský odstup vidieť v 3. zhluku (anglicko-hovoriace krajiny) v 6. zhluku a 7. zhluku (škandinávske krajiny). Tieto štáty sú charakteristické menej formálnymi vzťahmi, snahou eliminácie spoločenskej nerovnosti a striktne hierarchických štruktúr v organizáciách. Na druhej strane napríklad zhluk č. 11 s vysokým stupňom mocenskej vzdialenosti (ázijské krajiny) poukazuje na formálne vzťahy v manažmente, hierarchické usporiadanie v organizáciách a dôraz na diferenciáciu spoločenských vrstiev. Rovnako aj 9. zhluk s najvyšším indexom mocenskej vzdialenosti zlučuje krajiny ako Rusko, Saudská Arábia alebo Srí Lanka, pre ktoré je typická spoločenská nerovnosť a formalizované vzťahy.

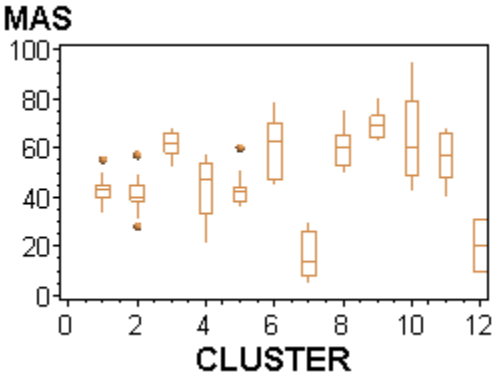
Zaujímavé je si všimnúť na grafe č. 8 siedmy zhluk škandinávskych krajín s najnižším indexom maskulinity. Tieto krajiny majú najväčší počet žien v manažmente, snažia sa eliminovať rodové rozdiely, presadzujú práva žien a bojujú za ich emancipáciu. Rovnakým spôsobom môžeme interpretovať aj ďalšie kultúrne dimenzie na základe grafov č. 7, 8 a 9.



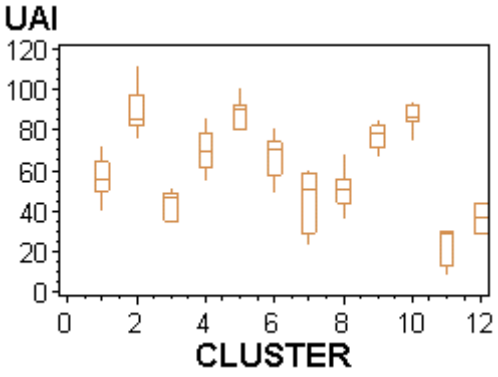
Graf 6: PDI pre 12 zhlukov



Graf 7: IDV pre12 zhlukov



Graf 8: MAS pre12 zhlukov



Graf 9: UAI pre12 zhlukov

4. Záver

Treba brať do úvahy, že výskum bol robený v minulom storočí a mnohé manažérske praktiky sa neustále zdokonaľujú a preberajú od iných krajín, kde sa ich používanie pozitívne uplatnilo. Aj samotná integrácia svetového trhu speje k tomu, že vplyv kultúrnych odlišností na manažérske správanie sa postupne stráca. Napriek tomu výsledky projektu môžu slúžiť ako orientačné informácie, na základe ktorých je možné urobiť si prehľad o jednotlivých krajinách.

5. Literatúra

HEBÁK, PETR A KOLEKTÍV: Vícerozmerné statistické metody (2). Informatorium. Praha 2005.

ROBBINS, S. P.: Management. Prentice Hall, New Jersey. 1996

SHARMA, SUBHASH: Applied Multivariate Techniques. New York, John Wiley & Sons, Inc., 1996.

STANKOVIČOVÁ, IVETA: Stredná dĺžka života v krajinách sveta v roku 1998. Slovenská štatistika a demografia č. 3/2000, str. 4 - 19.

Tabuľka : Zhluky štátov sveta podľa 4 ukazovateľov kultúrneho správania

C1 (n = 19)	**C2 (n = 12)**	**C3 (n = 6)**	**C4 (n = 4)**	**C5 (n = 9)**	**C6 (n = 7)**
Fiji	Peru	Australia	Pakistan	Romania	Germany
Indonesia	South Korea	United States	Taiwan	Serbia & Montenegro	Switzerland
Namibia	Bulgaria	Canada	Trinidad	Arab Emirates	Czech Republic*
Nepal	Croatia	United Kingdom	Costa Rica	Saudi Arabia	Luxemburg
Malawi	Egypt	Ireland		Russia	South Africa
Tanzania	Brazil	New Zealand		Sri Lanka	Israel
Burkina Faso	Turkey			Kuwait	Austria
Caucasus	El Salvador			Guatemala	
Cape Verde	Portugal			Panama	
Ghana	Uruguay				
Senegal	Chile				
Honduras	Greece				
Iran					
Zambia					
Sierra Leone					
Angola					
Bangladesh					
Jordan					
Thailand					
C7 (n = 7)	**C8 (n = 11)**	**C9 (n = 6)**	**C10 (n = 8)**	**C11 (n = 5)**	**C12 (n = 2)**
Estonia	Ethiopia	Iraq	Belgium	Hong Kong	Bhutan
Finland	Kenya	Mexico	France	China*	Surinam
Iceland	Nigeria	Ecuador	Argentina	Singapore	
Norway	Syria	Venezuela	Spain	Vietnam	
Denmark	Dominican Rep.	Colombia	Hungary*	Jamaica	
Sweden	India	Albania	Italy		

Adresa autora:

Krystyna Šípošová
Myrtina 81
951 12 Ivanka pri Nitre
krystyna.sipos@post.sk

Príloha:

Tabuľka 2. Popisné štatistiky ukazovateľov kultúrneho správania pre výsledných 12 zhlukov

CLUSTER	N obs.	Variable	Mean	Std Dev	Minimum	Maximum	N	Median	Coeff of Variation
1	19	PDI	71.42	7.07	58.00	80.00	19	70.00	9.90
		IDV	23.63	7.51	14.00	41.00	19	20.00	31.80
		MAS	43.00	5.14	34.00	55.00	19	43.00	11.96
		UAI	55.47	8.41	40.00	72.00	19	56.00	15.16
2	12	PDI	65.42	4.32	60.00	73.00	12	65.00	6.60
		IDV	28.08	7.88	16.00	38.00	12	28.50	28.06
		MAS	41.17	7.61	28.00	57.00	12	40.00	18.50
		UAI	89.50	10.83	76.00	112.00	12	85.50	12.10
3	6	PDI	33.33	6.98	22.00	40.00	6	35.50	20.93
		IDV	83.17	8.28	70.00	91.00	6	84.50	9.96
		MAS	61.17	5.74	52.00	68.00	6	61.50	9.39
		UAI	44.00	7.16	35.00	51.00	6	47.00	16.26
4	4	PDI	49.00	10.23	35.00	58.00	4	51.50	20.88
		IDV	15.75	1.50	14.00	17.00	4	16.00	9.52
		MAS	43.50	15.93	21.00	58.00	4	47.50	36.61
		UAI	70.00	12.68	55.00	86.00	4	69.50	18.11
5	9	PDI	90.78	3.93	85.00	95.00	9	90.00	4.33
		IDV	25.78	12.76	6.00	48.00	9	25.00	49.52
		MAS	43.33	7.70	36.00	60.00	9	42.00	17.76
		UAI	88.44	7.49	80.00	101.00	9	90.00	8.46
6	7	PDI	31.00	13.96	11.00	49.00	7	35.00	45.05
		IDV	61.00	5.72	54.00	68.00	7	60.00	9.37
		MAS	60.00	12.91	45.00	79.00	7	63.00	21.52
		UAI	66.71	10.58	49.00	81.00	7	70.00	15.86
7	7	PDI	31.57	7.09	18.00	40.00	7	31.00	22.46
		IDV	68.14	7.56	60.00	80.00	7	69.00	11.09
		MAS	15.57	9.31	5.00	30.00	7	14.00	59.77
		UAI	46.43	14.56	23.00	60.00	7	51.00	31.36
8	11	PDI	77.82	11.70	65.00	104.00	11	76.00	15.04
		IDV	35.00	10.34	20.00	52.00	11	32.00	29.55
		MAS	60.36	7.31	50.00	75.00	11	60.00	12.11
		UAI	50.64	9.22	36.00	68.00	11	51.00	18.21
9	6	PDI	82.17	9.93	67.00	95.00	6	81.00	12.08
		IDV	18.83	9.47	8.00	30.00	6	16.50	50.31
		MAS	69.83	6.24	63.00	80.00	6	69.50	8.94
		UAI	76.83	6.85	67.00	85.00	6	78.00	8.92
10	8	PDI	57.13	8.85	46.00	68.00	8	55.50	15.50
		IDV	63.13	14.13	46.00	80.00	8	65.50	22.38
		MAS	64.00	19.50	42.00	95.00	8	60.00	30.47
		UAI	86.75	6.34	75.00	94.00	8	86.00	7.31
11	5	PDI	67.40	13.33	45.00	80.00	5	70.00	19.78
		IDV	24.80	8.23	20.00	39.00	5	20.00	33.18
		MAS	55.80	11.88	40.00	68.00	5	57.00	21.30
		UAI	22.00	10.65	8.00	30.00	5	29.00	48.43
12	2	PDI	87.50	10.61	80.00	95.00	2	87.50	12.12
		IDV	44.00	11.31	36.00	52.00	2	44.00	25.71
		MAS	20.50	14.85	10.00	31.00	2	20.50	72.44
		UAI	36.50	10.61	29.00	44.00	2	36.50	29.06

Porovnanie odhadov kovariančnej matice ako vstupu do optimalizačného procesu tvorby portfólia

František Štulajter

Abstrakt

Markowitz first introduced portfolio selection using a quantitative optimization procedure that balances the trade-off between risk and return. Necessary iputs needed for classical mean – variance optimization are expected returns of assets and their covariance matrix. The purpose of this case study is comparison of different covariance matrix estimates.

1. Úvod

Cieľom práce je porovnať štyri odhady kovariančnej matice v kontexte riešenia úlohy výberu optimálneho portfólia. Motiváciou testovania rôznych odhadov vstupných parametrov optimalizačnej úlohy je niekoľko. Ako uvádza Fabozzi a kol. (1), mnoho investičných spoločností sa v súčasnosti neopiera o optimalizačné úlohy z dôvodu ich vysokej citlivosti na hodnoty vstupných premenných (hlavne odhadov očakávaných výnosov aktív) a problémov s odhadom kovariančnej matice pri veľkom množstve finančných aktív. Jednou z možností ako znížiť citlivosť optimalizačného procesu na odhad očakávaného výnosu je použitie robustných odhadov, ako napríklad Black – Littermanov model. Príspevok popisuje a porovnáva odhady kovariančnej matice. Jedná sa o výberovú kovariančnú maticu (sample covariance matrix), kovariančnú maticu generovanú jednoduchým indexovým modelom (single index model), kovariančnú maticu s konštantnou koreláciou a kovariančnú maticu odhadnutú metódou shrinkage Oliviera Ledoita a Michaela Wolfa (4).
Príspevok je metodicky rozdelený do dvoch častí. Teoretická časť popisuje prezentované odhady a praktická na základe odhadnutých parametrov stanovuje váhy modelových portfólií.

Výberová kovariančná matica (Sample covariance matrix)

Výberová kovariančná matica môže byť odhadnutá ako $\hat{\Sigma} = \frac{1}{N-1} X'X$, kde X je $T*N$ matica T pozorovaní N náhodných čísiel, ktoré reprezentujú výnosy finančných aktív. $\hat{\Sigma}$ je maximálne vierohodný odhad skutočnej kovariančnej matice za predpokladu normality. Avšak ak sa počet cenných papierov N približuje k počtu pozorovaní T, celkový počet odhadovaných parametrov je príliš veľký, čo vedie k vysokej chybe odhadu[1].

Kovariančná matica generovaná SIM (Single-Index covariance matrix estimator ***)***

Single Index Model je jedno faktorový model rovnováhy na finančných trhoch. Bol vytvorený Williamom F. Sharpom v roku 1963. Podľa Eltona a Grubera (2) je v pozadí SIM empirické pozorovanie cien akcií, ktorých ceny sa vo všeobecnosti zvyšujú, ak rastie celkový trh a vice versa. Toto naznačuje jednu príčinu korelácie výnosov akcií, a síce ich citlivosť na pohyby agregovaného trhu. Z tohto dôvodu je možné vyjadriť výnos akcie i v závislosti od výnosu celkového trhu, najčastejšie reprezentovaného trhovým indexom ako $R_i = \alpha_i + \beta_i R_M + e_i$. $\beta_i R_M$ je časť výnosu spôsobená trhovým výnosom, α_i je očakávaná hodnota výnosu, nezávislá na trhovom výnose a e_i je náhodná zložka s nulovou strednou hodnotou. β_i je konštanta prislúchajúca k CP i a vyjadrujúca očakávanú zmenu R_i pri jednotkovej zmene R_M, označovaná ako miera systematického rizika. Prvým predpokladom SIM je ne korelovanosť

[1] Podľa Ledoit, O – Wolf, M (4)

náhodnej zložky e_i s výnosom trhu R_M. Dôležitejším predpokladom definujúcim celý model je nezávislosť náhodných zložiek e_i a e_j pre všetky hodnoty i a j, teda $E(e_i e_j) = 0$. To znamená, že jediným zdrojom spoločného pohybu jednotlivých CP je ich spoločný pohyb s trhom. Okrem trhu neexistuje žiadna iná spojitosť, ktorá môže vysvetliť vzájomnú závislosť výnosov CP. Tento zjednodušujúci predpoklad je aproximáciou reality. Odhad kovariančnej matice implikovanej SIM má tvar: $F = s_{00}^2 \hat{\beta}\hat{\beta}' + D$ [2], kde s_{00}^2 je výberová volatilita trhových výnosov, $\hat{\beta}$ je vektor odhadov príslušných beta faktorov[3] a D je diagonálna matica obsahujúca odhady variancií rezíduí. Ako pripomína Ledoit a Wolf (4), implicitným predpokladom je pozitívna variancia trhového portfólia.

Kovariančná matica s konštantnou koreláciou

Odhad kovariančnej matice môžeme rozložiť na $\hat{\Sigma} = \Lambda C \Lambda'$, kde Λ je diagonálna matica volatilít výnosov finančných aktív a C je výberový odhad korelačnej matice (sample correlation matrix). Maticu C nahradíme korelačnou maticou s konštantným korelačným koeficientom, ktorej hlavná diagonála obsahuje jednotky a ostatné prvky matice sú rovné $\hat{\rho}$ [4]. Tento rozklad kovariančnej matice umožňuje oddelene modelovať korelačné koeficienty a volatility skúmaných premenných.

Kovariančná matica odhadnutá metódou shrinkage

Základy štatistickej metódy odhadov parametrov shrinkage (zrážanie) položil svojou prácou Charles Stein v roku 1956. Metóda shrinkage je forma priemerovania rôznych odhadov parametrov, pričom shrinkage estimátor zväčša pozostáva z troch komponentov: 1.) odhad parametra s malou resp. žiadnou štruktúrou (napr.: výberová volatilita alebo aritmetický priemer ako odhad strednej hodnoty); 2.) odhad parametra so značnou štruktúrou (označovaný ako shrinkage target); a 3.) váha priemeru (shrinkage intensity). Podľa Fabozziho a kol. (1) shrinkage intensity môže byť zvolená na základe teoretických implikácií alebo na základe číselnej simulácie. Metóda shrinkage sa používa aj na odhad kovariančnej matice pre potreby optimalizácie investičných portfólií. Ako odhad s malou štruktúrou sa používa výberová kovariančná matica, ktorá je maximálne vierohodným odhadom, avšak môže prinášať veľkú chybu odhadu. Shrinkage target na druhej strane prináša len malú chybu odhadu, ale jedná sa o vychýlený odhad. V práci porovnávame odhad kovariančnej matice prezentovaný v Ledoit, O. – Wolf, M. (4). Autori za shrinkage target uvažujú kovariančnú matica s konštantnou koreláciou alebo kovariančnú maticu generovanú jedno faktorovým modelom. Obe možnosti podľa Fabozziho a kol. (1) ponúkajú podobné výsledky, avšak prvá menovaná je kvôli jednoduchšej implementácii v praxi používanejšia. V našej práci má porovnávaný odhad tvar $\hat{\Sigma}_{LW} = w\hat{\Sigma}_{CC} + (1-w)\hat{\Sigma}$, kde $\hat{\Sigma}$ je výberový odhad kovariančnej matice a $\hat{\Sigma}_{CC}$ je kovariančná matica s konštantnou koreláciou. Optimálna váha w (shrinkage intensity) je vyjadrená formulou: $w = \max\left\{0, \min\left\{\frac{\hat{\kappa}}{T}, 1\right\}\right\}$ pričom $\hat{\kappa} = \frac{\hat{\pi} - \hat{c}}{\hat{\gamma}}$ a T je počet pozorovaní. Pre presné matematické odvodenie prezentovaných odhadov odkazujeme čitateľa na príspevky (3) a (4), ako aj na internetovú stránku www.ledoit.net, kde sa nachádzajú naprogramované funkcie, ktoré implementujú model Ledoita a Wolfa v prostredí Matlabu. Odhad kovariančnej matice ako ju prezentujú Ledoit a Wolf, predpokladá že výnosy cenných papierov sú nezávislé a identicky rozdelené (IID) a prvé štyri momenty rozdelenia ich pravdepodobnosti sú konečné.

---

[2] Podľa Ledoit, O. – Wolf, M. (4)

[3] Podľa (2) strana 140 alebo regresnou analýzou

[4] $\hat{\rho} = \frac{1}{(N-1)N}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\rho}_{ij} - N\right)$

2. Postup a dáta

Naša práca má za cieľ porovnať možnosti odhadu kovariančnej matice ako vstupného parametra pre M-V optimalizáciu, ktorej úlohou je zostrojenie optimálnych portfólií na základe stanovených kritérií. Jednou možnosťou porovnania rôznych odhadov kovariančnej matice je riešenie úlohy stanovenia váh portfólia s globálne najmenšou varianciou (global minimum variance portfolio). Úloha je riešená kvadratickou optimalizáciou, pričom optimalizačný problém má tvar: $\min_w w'\hat{\Sigma} w$, za podmienok $w'\iota = 1$, $\iota' = [1,1,\ldots,1]$. Optimalizačná úloha je myopická, zameriava sa na stanovenie váh optimálneho portfólia v horizonte jednej periódy, preto sme použili koncept preskupovania portfólií. Z dôvodu podľa nášho názoru jednoduchšej správy portfólia sme stanovili obmedzenie váh portfólií, v tvare „zakázania“ krátkych predajov ($w \geq 0$). V procese optimalizácie dochádza z dôvodu zmenených vstupných parametrov k zmenám zastúpení jednotlivých aktív. Tieto zmeny môžu z praktického hľadiska spôsobovať vysoké transakčné náklady. Tento problém je možné znížiť stanovením obmedzenia $|x_i| \leq U_i$, kde $x = w - w_0$, pričom w_0 sú súčasné váhy portfólia a w sú cieľové váhy danej parciálnej časti optimalizácie.

V práci bolo skúmaných sedem finančných aktív reprezentovaných indexmi spoločnosti MSCI Barra. Jednalo sa o regionálne indexy: Europe, North America, Japan, Singapore a zástupcov rozvojových krajín, Latin Amrerica, Far East a Eastern Europe. Index MSCI World bol použitý ako trhové portfólio pri odhade parametrov beta. Pracovali sme s dennými hodnotami indexov denominovaných v EUR v období od 16.8.2001 do 5.10.2006[5]. Tento súbor dát bol rozdelený do 6 období, vždy s posunom o 60 obchodných dní. Každá perióda má 720 pozorovaní a najstarších 60 pozorovaní sa vždy vylučuje. Na konci prvej periódy 19.5.2005 bol uskutočnený odhad kovariančnej matice a stanové optimálne portfólio, ktoré bolo držané do 11.8.2005. Následne v perióde O2 bol odhadnutý nový model a skonštruované portfólio bolo držané do 3.11.2005 atď.. Popis období zobrazuje tabuľka č.1.

Tab.č. 1 Prehľad skúmaných periód

O1 Odhadná perióda / vždy 720 dní, cca 3 roky	16.08.2001 – 19.05.2005
O2 1. posun / 60 dní, cca 3 mesiace	20.05.2005 – 11.08.2005
O3 2. posun / 60 dní, cca 3 mesiace	12.08.2005 – 03.11.2005
O4 3. posun / 60 dní, cca 3 mesiace	04.11.2005 – 26.01.2006
O5 4. posun / 60 dní, cca 3 mesiace	27.01.2006 – 20.04.2006
O6 5. posun / 60 dní, cca 3 mesiace	21.04.2006 – 13.07.2006
60 dní, cca 3 mesiace	14.07.2006 – 05.10.2006

Výpočty boli uskutočnené softvérovým produktom Matlab 5.3. a MS Excel. Riešením úlohy vznikli 4 portfóliá, ktorých hodnoty začínajú koncom prvej periódy (O1). Cieľom bolo nájsť portfólio s najnižšou varianciou. Tabuľka č.2 zobrazuje hodnoty ročných štandardných odchýlok realizovaných výnosov jednotlivých modelových portfólií [6]. Ako je možné vidieť, odhad kovariančnej matice s konštantnou koreláciou umožnil vytvorenie portfólia s najnižšou realizovanou varianciou v oboch periódach. V šiestom stĺpci tabuľky č. 2 je znázornená štandardná odchýlka portfólia s globálne najnižšou varianciou, ktoré bolo vytvorené na základe poznania realizovaných výnosov v daných periódach. Jedná sa o portfólio, ktoré nie je možné zostrojiť, pretože pri tvorbe tohto portfólia bola použitá kovariančná matica výnosov, ktoré skutočne nastali. Je to minimálna hodnota štandardnej odchýlky portfólia.

[5] Zdroj: www.MSCIBarra.com

[6] Prevedenie realizovanej štandardnej odchýlky denných dát na ročné hodnoty, predpokladá nezávislosť a identické rozdelenie výnosov. Pri porušení tohto predpokladu je potrebné upraviť odhad ročnej štandardnej odchýlky o sériovú koreláciu výnosov (LITTERMAN, R. – RESOURCES GROUP QUANTITATIVE 2003. Modern Investment Management. Hoboken NJ: John Wiley & Sons Ltd, 2003, str. 232 – 240).

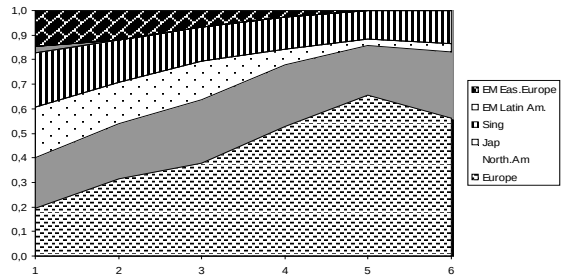
Tab.č.2 Porovnanie realizovaných štandardných odchýlok optimálneho portfólia

perióda \ σ_p	výberová C matica	odhadnutá SIM	C matica s konšt. ρ	Odhadnutá m. Shrinkage	In – Sample Variance
20.5.05-26.1.06	9,5510 %	10,2187 %	9,1070 %	9,4765 %	8,640 %
27.1.06-5.10.06	11,2445 %	12,6301 %	11,2198 %	11,2362 %	10,212 %
celkom	10,4412 %	11.5262 %	10,2200 %	10,4019 %	9,618 %

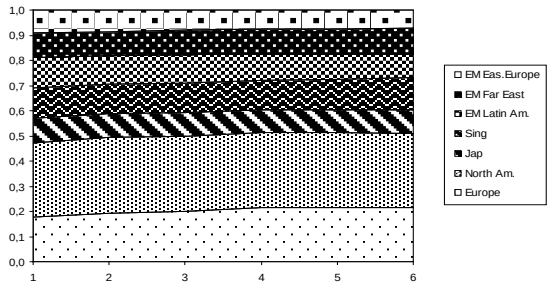
Zdroj: výpočty autora

3. Záver

Na začiatku práce sme očakávali, že najlepší odhad prinesie metóda shrinkage. Sledovali sme ako sa menilo zastúpenie jednotlivých aktív v modelovom portfóliu. U všetkých odhadov okrem odhadu generovaného SIM bolo zastúpenie veľmi podobné a je znázornené grafom č.1.



Graf č.1: Vývoj váh portfólia s konšt. koreláciou



Graf č.2: Vývoj váh portfólia odhadnutého SIM

Ako je vidieť z grafu č.2, zmeny váh aktív v modelovom portfóliu odhadnutého pomocou jednoduchého indexového modelu nie sú také rôznorodé. Z hľadiska správy portfólia predstavuje toto portfólio oveľa lepšiu diverzifikáciu rizika ako ostatné portfóliá. Príčinou tejto rôznosti modelových portfólií je vývoj na finančných trhoch v sledovanom období. Volatilita na európskom a severo – americkom trhu výrazne klesala, čo prostredníctvom odhadov variančno – kovariančných matíc spôsobilo preferovanie týchto aktív v optimalizačnom procese. Odhad kovariančnej matice pomocou jednoduchého indexového modelu za jedinú príčinu spoločného pohybu aktív považuje pohyb trhového portfólia, v našom prípade indexu MSCI World. Z grafu č.2 môžeme vidieť, že váhové zastúpenia v modelom portfóliu sú približne proporcionálne k váhovému zastúpeniu daného aktíva v index MSCI World, pričom indexy rozvíjajúcich sa krajín v ňom nie sú obsiahnuté. Riešením môže byť použitie rovnako váženého trhového indexu. Z dôvodu zhlukovania volatility možno navrhnúť modelovať kovariančnú maticu pre účely optimalizácie portfólia riešením dvoch parciálnych úloh - modelovaním variability napr. pomocou GARCH modelov a modelovaním korelácie výnosov jednotlivých finančných aktív.

4. Literatúra

(1) FABOZZI, F – FOCARDI, S – KOLM, P. 2006. Financial Modeling of the Equity Market: From CAPM to Cointegration. Hoboken, NJ: John Wiley & Sons Ltd, 2006.
(2) ELTON, E – GRUBER, M. 2003. Modern Portfolio Theory and Investment Analysis 6th. Ed.. Hoboken, NJ: John Wiley & Sons Ltd, 2003.
(3) LEDOIT, O – WOLF, M. 2003. Impoved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection, In: Journal of Empirical Finance 10, pp.603 – 621, 2003.
(4) LEDOIT, O – WOLF, M. 2004. Honey, I shrunk the Sample Covariance Matrix, In: The Journal of Portfolio Management 30 (2004), pp. 110 – 621 , 2004.

František Štulajter
stulajterf@yahoo.com
Ekonomická fakulta UMB
Tajovského 10
975 90 Banská Bystrica

ANALÝZA A PREDIKCIA VÝVOJA AGRÁRNEJ ZAMESTNANOSTI

Andrea Vladárová

Abstrakt
Základným cieľom príspevku je analýza časového radu agrárnej zamestnanosti v SR v rokoch 1991 až 2005 a prognóza. Očakáva sa pokles vývoja agrárnej zamestnanosti. Účelom analýzy je nájdenie a pomenovanie jednoznačných trendov v oblasti vývoja agrárnej zamestnanosti. Pre predikciu boli použité dva prístupy – prognóza vývoja agrárnej zamestnanosti – Holtov model, prognóza vývoja agrárnej zamestnanosti pomocou neurónovej siete.
Kľúčové slová: nezamestnanosť, ARIMA, neurónová sieť

Úvod

Poľnohospodárska výroba patrí medzi najstaršie odvetvia hospodárstva každej krajiny. V období globalizácie však poľnohospodárstvo postupne stráca svoje niekdajšie významné postavenie. V transformačnom období bolo poľnohospodárstvo ako prvé spomedzi ekonomických odvetví národného hospodárstva vystavené tlaku globálnych trhov. Vývoj zamestnanosti v agrárnom sektore na Slovensku bol ovplyvnený mnohými faktormi. Pokles zamestnanosti ovplyvnil na jednej strane rast produktivity práce a rozšírenie ekonomickej základne smerom k nepoľnohospodárskym odvetviam.

Metodika

Metodický aspekt projektu je reprezentovaný nasledovnými štatistickým metódami použitými tak pre analýzu agrárnej zamestnanosti v SR v rokoch 1991 až 2005 ako aj pre potreby predikcie: analýza časového radu, prognóza vývoja agrárnej zamestnanosti – Holtov model, prognóza vývoja agrárnej zamestnanosti pomocou neurónovej siete

Výsledky a diskusia

Keďže počet ekonomicky aktívnych ľudí rastie, preto aj zamestnanosť v absolútnych hodnotách rastie. K analýze vývoja nám preto budú slúžiť podiely – podiely agrárnej zamestnanosti na celkovej zamestnanosti. K dispozícii máme časový rad údajov v rozsahu 15 rokov, od roku 1991 do roku 2005. Pomocou špeciálnych charakteristík časového radu sa oboznámime s vývojovými tendenciami agrárnej zamestnanosti na Slovensku. V roku 1991 tvoril podiel zamestnaných v poľnohospodárstve 12,63% z celkového počtu zamestnaných v národnom hospodárstve. Z grafu vyplýva, že agrárna zamestnanosť za sledované obdobia najviac klesla v roku 1994 oproti roku 1993 na 85,44 %, čo predstavuje úbytok o 15,93% . Nepatrný nárast bol zaznamenaný iba v dvoch rokoch, a to v roku 1993 oproti roku 1992 na 101,37% teda o 1,37 % a roku 2002 o 0,45%. Počet zamestnancov v transformovaných poľnohospodárskych družstvách a štátnych majetkoch klesol, pretože mnohí ľudia si našli iné zamestnanie a iných prepustili. Značné množstvo pracovníkov v poľnohospodárstve odišlo do dôchodku, niektorí si našli prácu v iných odvetviach hospodárstva a pre časť z nich prekážkou je aj nízka úroveň vzdelania. V tomto období sa poľnohospodárstvo na celkovej zamestnanosti podieľa už len 4,7 %. Pokles agrárnej zamestnanosti sa prejavil tiež v poklese relatívnej zamestnanosti v poľnohospodárstve o 0,4 osôb na 100 ha poľnohospodárskej pôdy.
Uvedenými ukazovateľmi sme sa dostali na úroveň EÚ. Predpokladá sa, že tieto trendy znižovania agrárnej zamestnanosti budú pokračovať, a to hlavne v dôsledku týchto skutočností:

- Pomer starších pracovníkov v poľnohospodárstve rastie.

Na úbytku pracovných síl sa budú vo veľkej miere podieľať starší ľudia, ktorí sú

v súčasnosti ešte zamestnaní v poľnohospodárstve a v budúcom desaťročí odídu do dôchodku.

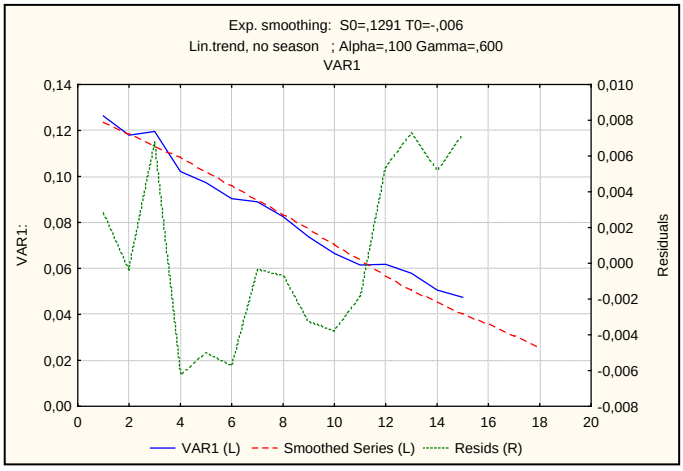
- Reštrukturalizácia hospodárskej sféry v neprospech poľnohospodárstva
  Pod vplyvom ďalšieho ekonomického rastu odíde značné množstvo pracovníkov z poľnohospodárstva za prácou do iných odvetví .
- Nízka vzdelanostná úroveň
  Väčšie dôvody na migráciu by mohli mať ľudia, ktorí boli prepustení v dôsledku reštrukturalizácie fariem a nemôžu si nájsť iné zamestnanie, alebo tí, ktorí sú na farmách nadbytoční. Pracovníci, ktorí si nemôžu nájsť zamestnanie v prosperujúcich odvetviach sú zväčša starší a menej vzdelaní, a neprimerane vysoký podiel pracovníkov oboch skupín je zamestnaný práve v poľnohospodárstve.

Prognóza vývoja agrárnej zamestnanosti – Holtov model

Na prognózu sme použili štatistický softvér Statistics 6.0. K dispozícii máme údaje za 15 rokov, od roku 1991 po rok 2005. Pri výbere vhodnej metódy musíme zohľadniť, že máme relatívne málo pozorovaní, z čoho usudzujeme, že výber ARIMA modelu by nebol vhodný. Ako prvý model sme použili Holtov model zohľadňujúci trend. Takýto model je charakterizovaný 2 konštantami (alfa - vyrovnávacia váha pre priemer, gama - vyrovnávacia váha pre trend). Miery kvality modelu merajú kvalitu modelu cez nevysvetlenú variabilitu, ktorá má podobu reziduálov.

	Error
Mean error –ME	0,00048255260672
Mean absolute error - MAE	0,00412302558840
Sums of squares - SS	0,00033898332230
Mean square - MS	0,00002259888815
Mean percentage error - MPE	1,47646297975475
Mean abs. perc. error - MAPE	5,78045611538929

Obrázok 1 Holtov model zohľadňujúci trend



Prognóza vývoja agrárnej zamestnanosti pomocou neurónovej siete

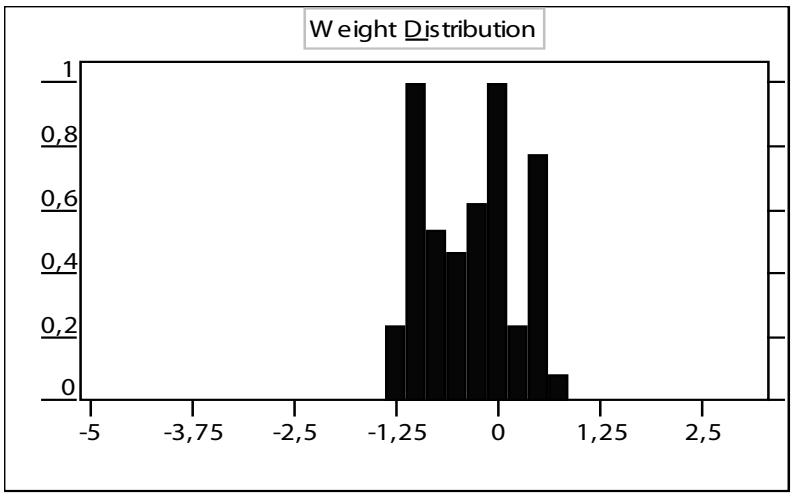
Na predikciu vývoja podielu agrárnej zamestnanosti sme použili aj neurónovú sieť. Najvhodnejšia sa ukázala neurónová sieť MLP s jednou skrytou vrstvou a ôsmimi neurónami. Jej podrobné charakteristiky sú v nasledujúcich tabuľkách.

Tabuľka 1 Hodnoty chyby najmenších štvorcov pre jednotlivé skupiny údajov

Charakteristika	trénovacie údaje	verifikačné údaje	testovacie údaje
RMS error	1,303e-05	0,002464	0,003541

Zdroj: Vlastné výpočty

Obrázok 2 Rozloženie váh



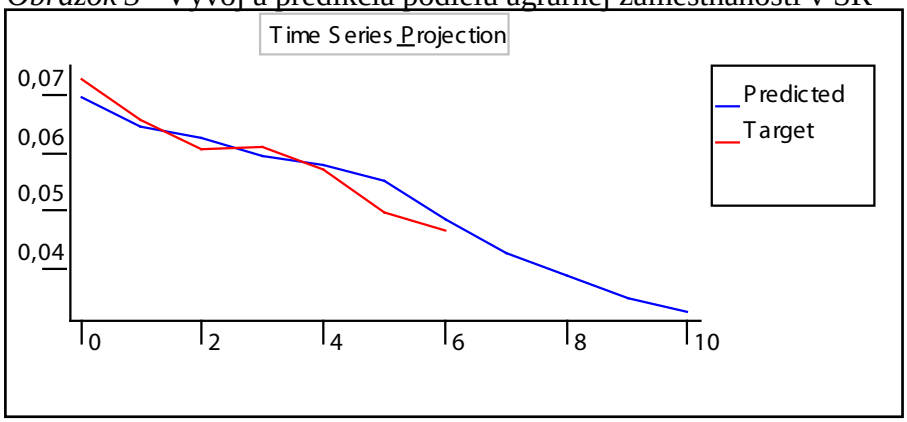
Zdroj: Vlastné výpočty

Tabuľka 2 Charakteristiky regresie

Charakteristika	trénovacie údaje	verifikačné údaje	testovacie údaje
Data Mean	0,05697	0,06437	0,05618
Data S.D.	0,0135	0,008304	0,007911
Error Mean	-1,30E-05	8,61E-05	0,0009825
Error S.D.	2,97E-07	0,00302	0,004812
Abs E. Mean	1,30E-05	0,002348	0,003402
S.D. Ratio	2,20E-05	0,363671	0,6082367
Correlation	0,9987621	0,9878849	0,9978655

Zdroj: Vlastné výpočty

Obrázok 3 Vývoj a predikcia podielu agrárnej zamestnanosti v SR



Zdroj: Vlastné výpočty

Tabuľka 3 Predikované a skutočné hodnoty

Rok	Skutočnosť	Odhad	Chyba
1999	0,07034	0,07373	0,1777208
2000	0,0665	0,06652	0,0006931
2001	0,06352	0,0615	0,1060419
2002	0,05936	0,06178	0,1267308
2003	0,05951	0,05789	0,08521
2004	0,05497	0,05059	0,229641
2005	0,0474107	0,04742	0,0006711

Zdroj: Vlastné výpočty

V rokoch 1999-2005 bola najpresnejšie aproximovaná hodnota pre rok 2000 s chybou 0,0006931 a najmenej presne pre rok 2004 s chybou 0,229641.

Literatúra

[1] DUFEK, J.: Výzkum demografie venkova České republiky v procesu ekonomické transformace. In: Zborník príspevkov z medzinárodného seminára Štatistika sociálno-demografickej situácie vidieka. Nitra: SPU, 1998, s. 1 - 5

[2] HRUBÝ, J.: Problematika prirodzeného prírastku obyvateľstva vo vzťahu k vidieku. In: Slovenská štatistika a demografia, ŠU SR, roč. 7, 1997, č. 1, s. 40-47, ISSN 1210-1095

[3] STEHLÍKOVÁ., B. - MECHÍROVÁ., A.: Porovnanie nezamestnanosti vybraných štátov na základe indikátorov Svetovej banky. In Zborník príspevkov z AP IX, Praha: Czech University of Agriculture, 2000, ISBN 80-213-0657-2

Kontaktná adresa

Andrea Vladárová, študentka, FEM SPU v Nitre, Tr. A. Hlinku 2, 949 76 Nitra

Recenzia knihy

Jozef Chajdiak: „Štatistické úlohy a ich riešenie v Exceli“

Ing. Zuzana Berčačinová[1]

Abstract: The paper consists describe book of Jozef Chajdiak „ The statistical Tasks and their Solve in Excel“.

Nedávno som bola obdarená kvalitne napísanou knihou „Štatistické úlohy a ich riešenie v Exceli“ od skúseného pedagóga na Katedre štatistiky FHI EU v Bratislave Doc. Ing. Jozefa Chajdiaka, CSc. Kniha je akýmsi voľným pokračovaním a nadstavbou už staršej publikácie od tohto autora „Štatistika v Exceli“. Kniha bola vydaná v Bratislave vydavateľstvom STATIS v roku 2005. Obsahuje 266 strán textu s bohatou ilustráciou, na ktorých je prehľadne a zaujímavým spôsobom odprezentovaná štatistika a úlohy zo štatistiky riešené v programe Excel.

Kniha sa vyznačuje presnosťou v štylizácii textu k teórii, podrobným rozpracovaním jednotlivých výstupov z počítača pri každom spracovávanom príklade. V tejto knihe, ktorá je rozčlenená do viacerých kapitol, autor podáva stručný prierez všetkými najpoužívanejšími oblasťami štatistiky. Okrajovo sa venuje aj zložitejšej problematike, ktorá je určená už pre profesionálnych odborníkov v štatistike.

Celý text knihy nesie rukopis pedagóga. Svojou knihou rozširuje a upevňuje vedomosti z oblasti štatistiky, podporuje rozvoj myslenia a tvorivú činnosť čitateľov. Jednotlivé kapitoly majú svoju presnú štruktúru. V úvode kapitoly autor stručne čitateľa prenesie teoretickou časťou danej problematiky, rozpíše členenie a metodiku riešenia. Po teoretickej časti prichádza aplikačná časť jednotlivých kapitol, ktorá sa skladá zo zadania príkladu, podrobného popisu postupu jednotlivých úkonov a príkazov k dosiahnutiu želaného výsledku v aplikácii programu Excel. Jednotlivé počítačové výstupy graficky dopĺňajú riešenie príkladov. Nakoniec autor presne a podrobne opisuje zobrazený konečný výstup. Opisuje a zdôvodňuje číselné hodnoty na zadanie príkladu.

Kniha je rozdelená do 16. základných kapitol. Prvá kapitola obsahuje všeobecné poznámky. Veľmi dôležitým je opis postupu ako získať na internete súbory údajov, s ktorými sa pracuje v ďalších kapitolách, časť z nich je aj priamo uvedená v tejto prvej kapitole.

V druhej kapitole je spracovaná problematika analýzy súboru nameraných hodnôt jednej premennej. Začína základným štatistickým rozborom (výpočet jednoduchých štatistík, vážených štatistík a ďalších štatistík), pokračuje radom rozdelenia početností v tvare frekvenčnej tabuľky a v tvare histogramu. Dôležitou súčasťou je použitie metódy usporiadania súboru. Ďalej nasleduje výpočet kvantilov, mier koncentrácie a štandardizácia hodnôt. V závere kapitoly je ukážka metód analýzy poradia.

Tretia kapitola sa venuje grafickej analýze, uvedené sú tri základné grafy – stĺpcový, bodový a spojnicový.

Štvrtá kapitola je venovaná problematike analýzy štatistických závislosti hodnôt premenných (korelačná a regresná analýza).

Piata kapitola sa venuje problematike časových radov. Prezentácia časového radu, elementárne charakteristiky časového radu, kĺzavé priemery a kĺzavé úhrny, modelovanie trendu, analýza sezónnosti, prognózovanie budúceho obdobia. V závere kapitoly sa autor venuje úmernosti vývoja osobných nákladov na produktivitu práce.

Šiesta kapitola je venovaná problematike indexov a rozkladov. Obsahuje tiež jednoduchý model rozkladu zisku a pyramídové modely rozkladu produktivity práce.

[1] Ing. Zuzana Berčačinová, Patria, a.s. Bratislava

Prognóza vývoja počtu obyvateľov je rozpracovaná v siedmej kapitole. Ďalších šesť kapitol sa venuje metódam a postupom matematickej štatistiky. V ôsmej kapitole je opísaný proces randomizácie výberu, v deviatej intervalové odhady v desiatej testovanie hypotéz, v jedenástej vyhodnocovanie experimentov a v dvanástej kapitole viacrozmerné testy.

Trinásta kapitola svojim obsahom je určená čitateľom zaoberajúcim sa riadeným výrobných procesov. Je v nej rozpracovaná problematika štatistického riadenia kvality, konkrétne Paretovej analýzy, regulačných diagramov a analýzy spôsobilosti procesu.

Štrnásta kapitola je určená pre záujemcov o analýzy rozsiahlych súborov. Opísaný je podsystém analýzy kontingenčnej tabuľky (Pivot Table). Konkrétne je v nej realizované jedno, dvoj a trojstupňové triedenie. Opísaný je spôsob určenia možných hodnôt v jednotlivých bunkách tabuľky. V danej kapitole je tiež uvedená metóda vytvorenia kontingenčného grafu.

V pätnástej kapitole je uvedená množina modelov rozdelení pravdepodobnosti s členením na diskrétne a spojité rozdelenia. Posledná kapitola sa venuje problematike generovania náhodných čísel.

Určite by som túto knihu odporučila všetkým študentom na stredných a vysokých školách, ktorí majú v rozvrhoch predmet štatistiku. Rada by som ju odporučila aj ekonomickým pracovníkom, účtovníkom, finančným analytikom ale aj samoukom, a všetkým tým, ktorý dennodenne pracujú s číslami a finančnými analýzami. Všetkým tým, ktorí chcú lepšie ovládať program Excel pri svojej práci a ktorí majú potrebu si prehlbovať svoje vedomosti v tejto oblasti.

A nakoniec ešte pár slov. Som rada, že som bola obdarená touto knihou. Stala sa pre mňa akousi encyklopédiou pri dennodennom spracovaní ekonomických výsledkov a ukazovateľov o vývoji spoločnosti aj je napredovaní.

Adresa autora:
zuzanapatria@mail.t-com.sk

Pravděpodobnostní rozdělení v MS Excel

Luboš Marek, Michal Vrabec

Abstract

The main aim of this paper is the describing of probability distributions in MS Excel software. The each probability distribution is described at first in theoretical level (including formulas for mean and variance) and then the method of calculation for density function and distribution function is following (including the exact syntax). The way how to calculate the density function and the values of percentiles is shown, too.

Key words

MS Excel, probability distribution, density function, distribution function, percentile, critical value.

Tento článek navazuje na stať věnovanou popisu diskrétních rozdělení v programu MS Excel autorů M. Vrabec a L. Marek. Zde se soustředíme na rozdělení spojitá.

Spojitá rozdělení

V oblasti spojitých rozdělení obsahuje MS Excel řadu rozdělení, u kterých opět uvádíme název příslušné funkce pro výpočet distribuční funkce, hustoty a kvantilu:

rozdělení	distribuční funkce	Hustota	kvantily
normální	NORMDIST	NORMDIST	NORMINV
normované normální	NORMSDIST	NE	NORMSINV
logaritmicko normální	LOGNORMDIST	NE	LOGINV
exponenciální	EXPONDIST	EXPONDIST	NE
Weibullovo	WEIBULL	WEIBULL	NE
Studentovo (t)	TDIST	NE	TINV
Fischer-Schnedecorovo (F)	FDIST	NE	FINV
chí-kvadrát	CHIDIST	NE	CHIINV
Beta	BETADIST	NE	BETAINV
Gama	GAMMADIST	GAMMADIST	GAMMAINV

Je tedy zřejmé, že nabídka spojitých rozdělení je podstatně širší, než je tomu u rozdělení nespojitých. Pouze čtyři rozdělení však mají uveden vzorec pro výpočet hustoty, což v ostatních případech pochopitelně bude komplikovat její výpočet a případné grafické zobrazení. V takových případech bude nutné zadat vzorec hustoty pravděpodobnosti ručně dle tvaru příslušné funkce.

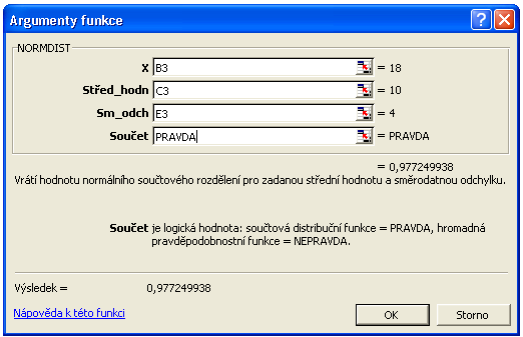
Normální rozdělení

Náhodná veličina X má normální rozdělení s parametry μ a σ^2, jestliže její hustota pravděpodobnosti má tvar

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < +\infty, \ -\infty < \mu < +\infty, \ \sigma > 0 \ .$$

Střední hodnota a rozptyl mají tvar

$$E(X) = \mu \qquad D(X) = \sigma^2 .$$



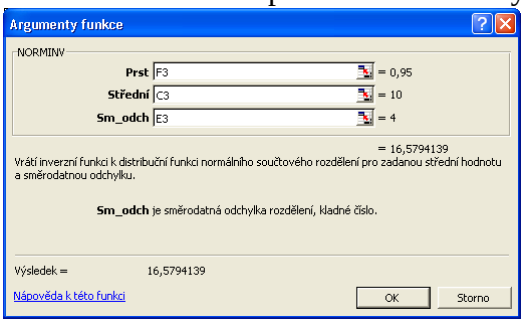
V Excelu se pro distribuční funkci a hustotu používá funkce NORMDIST. Její argumenty mají následující význam:

X - x. Hodnota, ve které počítáme $F(x)$, resp. $f(x)$.

Střed_hodn - μ. Parametr rozdělení a zároveň střední hodnota.

Sm_odch - σ. Parametr rozdělení a zároveň odmocnina z rozptylu (tedy směrodatná odchylka).

Součet - NEPRAVDA pro hodnotu hustoty $f(x)$, PRAVDA pro hodnotu distribuční funkce $F(x)$.



Funkce pro výpočet kvantilů normálního rozdělení má v Excelu název NORMINV. Jedná se skutečně o kvantilovou funkci $F(x_P) = P$, která má následující argumenty:

Prst - pravděpodobnost P pro hodnotu kvantilu x_P.

Střední - μ. Parametr rozdělení a zároveň střední hodnota.

Sm_odch - σ. Parametr rozdělení a zároveň odmocnina z rozptylu (tedy směrodatná odchylka).

Kromě normálního rozdělení s obecnými parametry μ a σ^2 nabízí Excel i normované normální rozdělení, tedy rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$.

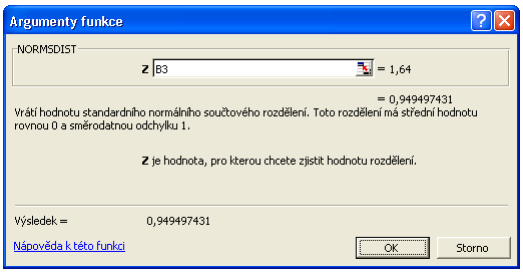
Normované normální rozdělení

Náhodná veličina U má normální rozdělení s parametry 0 a 1, jestliže její hustota pravděpodobnosti má tvar

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} \qquad -\infty < u < +\infty \ .$$

Střední hodnota a rozptyl mají tvar

$$E(U) = 0 \qquad D(U) = 1.$$



V Excelu se pro distribuční funkci používá funkce NORMSDIST. Tato funkce má jediný argument:

Z - u. Hodnota, ve které počítáme $F(u)$.

Hodnoty hustoty lze pomocí této funkce počítat nelze, je třeba je napočítat z obecného normálního rozdělení při vhodné volbě parametrů. To ostatně platí i pro hodnoty distribuční funkce a kvantily.

Logaritmicko normální rozdělení

Náhodná veličina X má logaritmicko normální rozdělení s parametry μ a σ^2, jestliže její hustota pravděpodobnosti má tvar

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}} \qquad x > 0,\ -\infty < \mu < +\infty,\ \sigma > 0$$
$$= 0 \qquad x \le 0$$

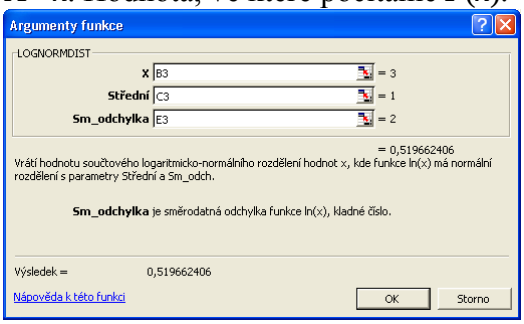
.

Střední hodnota a rozptyl mají tvar

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \qquad D(X) = \exp(2\mu + \sigma^2)\,[\exp(\sigma^2) - 1].$$

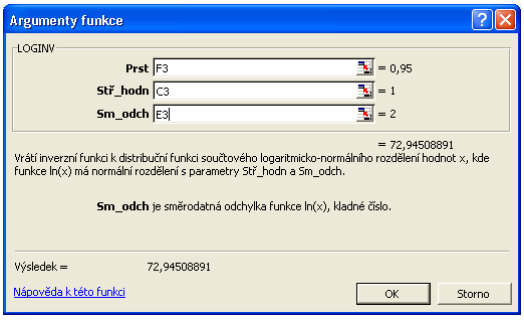
Připomeňme, že náhodná veličina $Y = \ln(X)$, má potom normální rozdělení s parametry μ a σ^2 - tedy přirozený logaritmus náhodné veličiny s logaritmicko normálním rozdělením má normální rozdělení se stejnými parametry μ a σ^2.

V Excelu se pro výpočet hodnot distribuční funkce používá funkce LOGNORMDIST. Její argumenty mají následující význam:
X - x. Hodnota, ve které počítáme $F(x)$.



Střední - μ. Parametr rozdělení. Pozor, nejedná se o střední hodnotu X, nýbrž o střední hodnotu ln(X).
Sm_odchylka - σ. Parametr rozdělení. Opět se nejedná o směrodatnou odchylku X, nýbrž o směrodatnou odchylku hodnoty ln(X).

Funkce pro výpočet kvantilů logaritmicko normálního rozdělení má v Excelu název LOGINV. Jedná se o kvantilovou funkci $F(x_P) = P$, která má následující argumenty:



Prst - pravděpodobnost P pro hodnotu kvantilu x_P.
Stř_hodn - μ. Parametr rozdělení a zároveň střední hodnota veličiny ln(X).
Sm_odch - σ. Parametr rozdělení, odmocnina ze σ^2. Zároveň směrodatná odchylka veličiny ln(X).

Exponenciální rozdělení

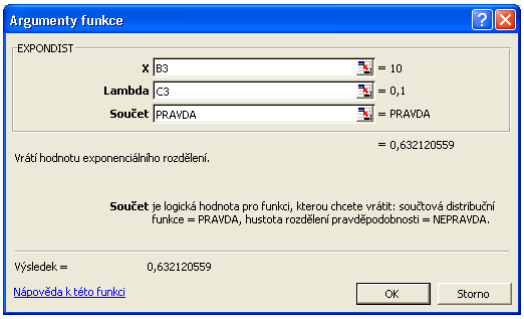
Náhodná veličina X má exponenciální normální rozdělení s parametrem λ, jestliže její hustota pravděpodobnosti má tvar

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} \quad x > 0,\ \lambda > 0 \\ &= 0 \qquad\quad x \le 0 \end{aligned} .$$

Pokud položíme $\lambda = \frac{1}{\delta}$, dostali bychom tvar rozdělení, v jakém je obvykle uváděn v literatuře. Tento tvar je prezentován pro hodnoty $x > 0$, neuvažujeme tedy možné posunutí A.
Střední hodnota a rozptyl mají tvar

$$E(X) = \frac{1}{\lambda} \qquad D(X) = \frac{1}{\lambda^2}.$$

V Excelu se pro výpočet hodnot distribuční funkce a hustoty používá funkce EXPONDIST. Její argumenty mají následující význam:



X - x. Hodnota, ve které počítáme $F(x)$, resp. f(x).
Lambda - λ. Parametr rozdělení.
Součet - NEPRAVDA pro hodnotu hustoty $f(x)$, PRAVDA pro hodnotu distribuční funkce $F(x)$.

Funkce pro výpočet kvantilů tohoto rozdělení není k dispozici. S jejím výpočtem si však snadno poradíme, neboť pro $100P\%$ kvantil exponenciálního rozdělení platí vztah

$$x_P = \frac{-\ln(1-P)}{\lambda}, \quad 0 < P < 1$$

Weibullovo rozdělení

Náhodná veličina X má Weibullovo rozdělení s parametry δ a c, jestliže její hustota pravděpodobnosti má tvar

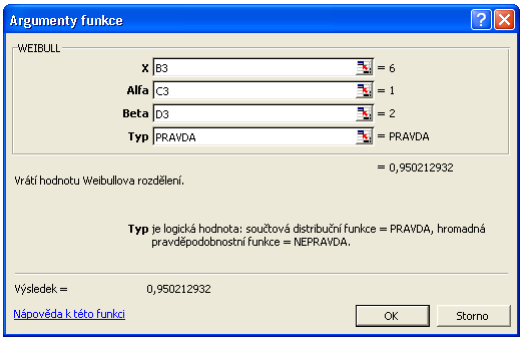
$$f(x) = \frac{cx^{c-1}}{\delta^c}\exp\left[-\left(\frac{x}{\delta}\right)^c\right], \quad x>0,\ \delta>0, c>0$$
$$= 0 \qquad\qquad x \le 0 .$$

Střední hodnota a rozptyl mají tvar vyjádřený pomocí gama funkce

$$E(X) = \Gamma\left(\frac{1}{c}+1\right)\delta \qquad D(X) = \left[\Gamma\left(\frac{2}{c}+1\right) - \Gamma^2\left(\frac{1}{c}+1\right)\right]\delta^2 .$$

Speciálním případem Weibullova rozdělení je pro $c = 1$ exponenciální rozdělení.

V Excelu se pro výpočet hodnot distribuční funkce a hustoty používá funkce WEIBULL. Její argumenty mají následující význam:



X - x. Hodnota, ve které počítáme $F(x)$, resp. $f(x)$.
Alfa - c. Parametr rozdělení.
Beta - δ. Parametr rozdělení.
Typ - NEPRAVDA pro hodnotu hustoty $f(x)$, PRAVDA pro hodnotu distribuční funkce $F(x)$.

Funkce pro výpočet kvantilů tohoto rozdělení není v Excelu k dispozici. Pro výpočet $100P\%$ kvantilu Weibullova rozdělení můžeme použít vztah

$$x_P = \delta\left[-\ln(1-P)\right]^{1/c}, \quad 0<P<1.$$

Studentovo rozdělení (t-rozdělení)

Náhodná veličina X má Studentovo rozdělení s parametrem n (počet stupňů volnosti), jestliže její hustota pravděpodobnosti má tvar

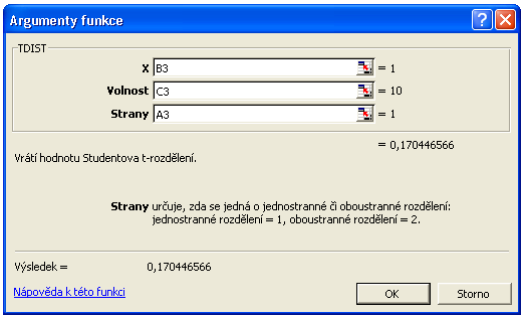
$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad -\infty < x < +\infty,\ n \in N$$

Střední hodnota existuje, pokud $n>1$ a je rovna $E(X)=0$.

Rozptyl existuje, pokud $n>2$ a je roven $D(X) = \frac{n}{n-2}$.

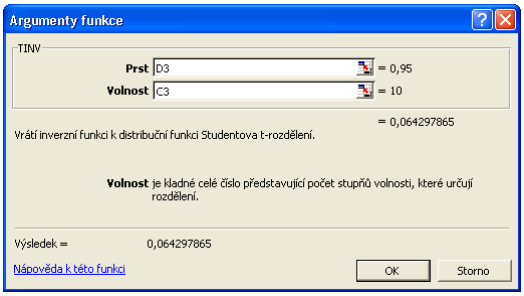
V Excelu se pro výpočet hodnot distribuční funkce používá funkce TDIST. Pozor - Excel nepočítá přímo hodnotu distribuční funkce, ale počítá $P(X>x)$, tedy výraz $1-F(x)$! Navíc není možné za x dosadit záporné číslo - pro záporná x je tedy nutné využít symetrie Studentova rozdělení kolem nuly ($F(-x) = 1-F(x)$).

Argumenty funkce TDIST mají následující význam:



X - *x*. Hodnota, ve které počítáme výraz 1-*F*(*x*).
Volnost - *n*. Parametr rozdělení, počet stupňů volnosti.
Strany - lze dosadit hodnoty 1 a 2. Pro 1 se počítá výraz
1-*F*(*x*), pro 2 se počítá pravděpodobnost $2*(1-F(x))$ tj. $1-P(-x<X<x)$.

Funkce pro výpočet kvantilů Studentova rozdělení má v Excelu název TINV. Výpočet kvantilů se přitom vymyká postupům u předchozích rozdělení. Nepočítá se totiž kvantilová funkce, počítá se funkce kritických hodnot $F'(x_P)=P(|X|>x_P)=P$. Pro výpočet kvantilu tedy platí, že pro zadanou pravděpodobnost *P* počítá funkce TINV kvantil $x_{1-P/2}$ - a pozor, nerespektuje se znaménko u kvantilu! Znaménko tedy musí uživatel doplnit sám tak, že od 0 do 50% kvantilu přiřadí znaménko záporné, od 50% do 100% znaménko kladné. Funkce TINV je tedy vlastně inverzní funkcí k TDIST pro hodnotu argumentu Strany = 2. Funkce TINV má následující argumenty:



Prst - pravděpodobnost *P* pro hodnotu kvantilu $x_{1-P/2}$ (až na znaménko).
Volnost - *n*. Parametr rozdělení, počet stupňů volnosti.

Fischer-Schnedecorovo rozdělení (F rozdělení)

Náhodná veličina *X* má Fischer-Schnedecorovo rozdělení s parametry *n* a *m* (počty stupňů volnosti), jestliže její hustota pravděpodobnosti má tvar

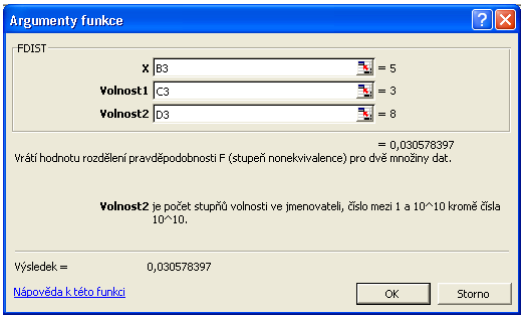
$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{n}{m}\right)^{n/2} x^{n/2-1}\left(1+\frac{n}{m}x\right)^{-\frac{n+m}{2}} \qquad x>0,\ n\in N,\ m\in N$$

$$= 0 \qquad x\le 0$$

Střední hodnota existuje, pokud $m>2$ a je rovna $E(X)=\frac{m}{m-2}$.

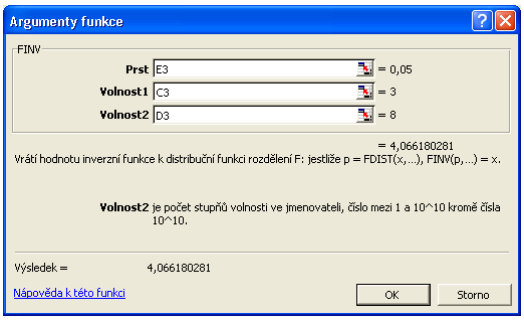
Rozptyl existuje, pokud $m>4$ a je roven $D(X)=\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$.

V Excelu se pro výpočet hodnot distribuční funkce používá funkce FDIST. Pozor - Excel nepočítá přímo hodnotu distribuční funkce, ale počítá $P(X>x)$, tedy výraz $1-F(x)$! Argumenty funkce FDIST mají následující význam:



X - *x*. Hodnota, ve které počítáme výraz 1-*F*(*x*).
Volnost1 - *n*. Parametr rozdělení, počet stupňů volnosti.
Volnost2 - *m*. Parametr rozdělení, počet stupňů volnosti.

Funkce pro výpočet kvantilů Fischer-Schnedecorova rozdělení má v Excelu název FINV. Název je opět zavádějící, protože se nejedná o kvantilovou funkci, nýbrž o funkci kritických hodnot $F'(x_P) = P(X > x_P) = P$. Pro zadanou pravděpodobnost P se tedy počítá kvantil x_{1-P}! Funkce FINV má následující argumenty:



Prst - pravděpodobnost P pro hodnotu kvantilu x_{1-P}.
Volnost1 - n. Parametr rozdělení, počet stupňů volnosti.
Volnost2 - m. Parametr rozdělení, počet stupňů volnosti.

Při výpočtu kvantilů Fischer-Schnedecorova rozdělení můžeme využít vztah

$$x_P(n,m) = \frac{1}{x_{1-P}(m,n)}$$

Chí kvadrát rozdělení (χ^2 rozdělení)

Náhodná veličina X má chí-kvadrát rozdělení s parametrem n (počet stupňů volnosti), jestliže její hustota pravděpodobnosti má tvar

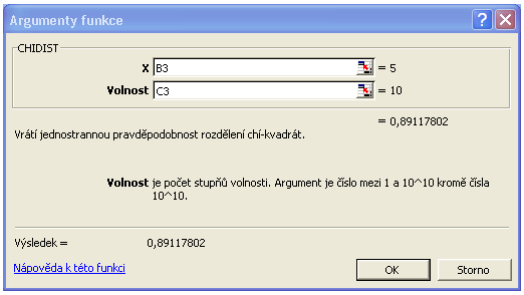
$$f(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{n/2-1} e^{-x/2} \quad x > 0, \quad n \in N$$

$$= 0 \quad x \le 0$$

Střední hodnota a rozptyl mají tvar

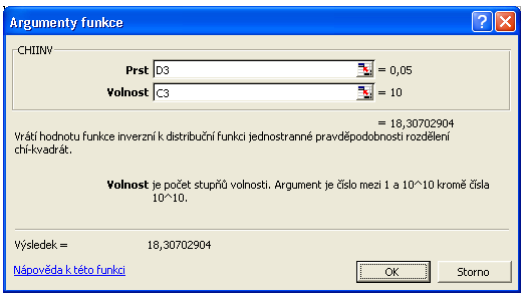
$E(X) = n$ $D(X) = 2n$.

V Excelu se pro výpočet hodnot distribuční funkce používá funkce CHIDIST. Pozor - Excel nepočítá přímo hodnotu distribuční funkce, ale počítá $P(X > x)$, tedy výraz $1 - F(x)$! Argumenty funkce CHIDIST mají následující význam:



X - x. Hodnota, ve které počítáme výraz 1-$F(x)$.
Volnost - n. Parametr rozdělení, počet stupňů volnosti.

Funkce pro výpočet kvantilů chí-kvadrát rozdělení má v Excelu název CHINV. Opět platí, že název je zavádějící, protože se nejedná o kvantilovou funkci, nýbrž o funkci kritických hodnot $F'(x_P) = P(X > x_P) = P$. Pro zadanou pravděpodobnost P se tedy počítá kvantil x_{1-P}! Funkce CHINV má následující argumenty:



Prst - pravděpodobnost P pro hodnotu kvantilu x_{1-P}.
Volnost - n. Parametr rozdělení, počet stupňů volnosti.

Beta rozdělení (4 parametrické)

Náhodná veličina X má Beta rozdělení s parametry a, b, α, β, jestliže její hustota pravděpodobnosti má tvar

$$f(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta-1}} \quad a<x<b,\ \ \alpha>0,\ \ \beta>0,\ \ b>a$$
$$= 0 \qquad \text{jinak}$$

Připomeňme, že $B(\alpha,\beta)$ je Beta funkce, definovaná jako

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx, \quad \alpha>0, \quad \beta>0$$

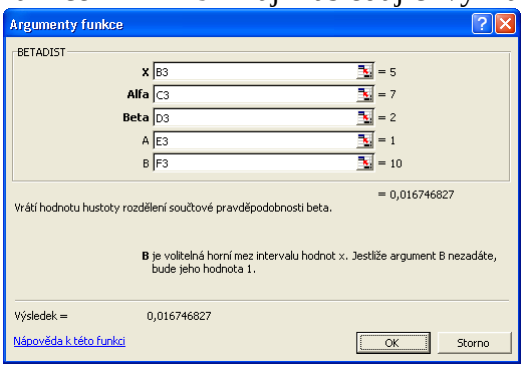
Střední hodnota a rozptyl mají tvar

$$E(X)=a+\frac{b\alpha}{\alpha+\beta} \qquad D(X)=\frac{\alpha\beta(b-a)^2}{(\alpha+\beta)^2(\alpha+\beta+1)^2}\ .$$

Pokud bychom položili $a=0$ a $b=1$, obdržíme „klasické“ dvouparametrické Beta rozdělení ve tvaru

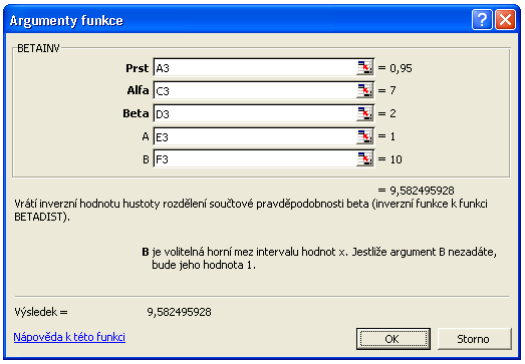
$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad 0<x<1,\ \ \alpha>0,\ \ \beta>0$$
$$= 0 \qquad \text{jinak}$$

V Excelu se pro výpočet hodnot distribuční funkce používá funkce BETADIST. Argumenty funkce BETADIST mají následující význam:



X - x. Hodnota, ve které počítáme hodnotu distribuční funkce $F(x)$.
Alfa - α, parametr rozdělení.
Beta - β, parametr rozdělení.
A - a, parametr rozdělení, dolní mez pro hodnoty x. Jedná se o nepovinný argument.
B - b, parametr rozdělení, horní mez pro hodnoty x. Jedná se o nepovinný argument.
Pokud nejsou argumenty A a B zadány, automaticky platí A = 0 a B = 1.

Funkce pro výpočet kvantilů Beta rozdělení má v Excelu název BETAINV. Jedná se o kvantilovou funkci



$$F(x_P)=P(X<x_P)=P\,,$$

která má následující argumenty:
Prst - pravděpodobnost P pro hodnotu kvantilu x_P.
Alfa - α, parametr rozdělení.
Beta - β, parametr rozdělení.
A - a, parametr rozdělení, dolní mez pro hodnoty x. Jedná se o nepovinný argument.
B - b, parametr rozdělení, horní mez pro hodnoty x. Jedná se o nepovinný argument.
Pokud nejsou argumenty A a B zadány, automaticky platí A = 0 a B = 1.

Gama rozdělení

Náhodná veličina X má Gama rozdělení s parametry α, β, jestliže její hustota pravděpodobnosti má tvar

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\,\Gamma(\alpha)} \quad x>0,\ \alpha>0,\ \beta>0$$
$$= 0 \qquad \text{jinak}$$

Připomeňme, že $\Gamma(\alpha)$ je Gama funkce, definovaná jako

$$\Gamma(\alpha)=\int_0^{\infty} x^{\alpha-1}e^{-x}dx, \quad \alpha>0$$

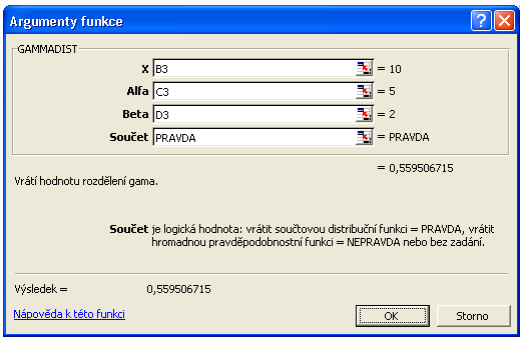
Střední hodnota a rozptyl mají tvar

$E(X)=\alpha\beta$ $\qquad$ $D(X)=\alpha\beta^2$.

Excel nemá přímo funkci, která by počítala hodnoty funkce Gama. Obsahuje však funkci GAMMALN, která vrací hodnotu přirozeného logaritmu funkce Gama. Hodnotu Gama funkce v bodě α pak snadno získáme složením funkce EXP a GAMMALN ve tvaru
EXP(GAMMALN(α)).
V Excelu rovněž není funkce Beta (viz předchozí rozdělení). Pro její výpočet je možné využít vztah

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Pro výpočet hodnot distribuční funkce rozdělení gama se v Excelu používá funkce GAMMADIST. Argumenty funkce GAMMADIST mají následující význam:



X - x. Hodnota, ve které počítáme hodnotu distribuční funkce $F(x)$.

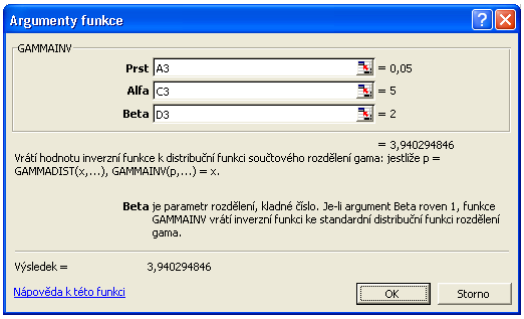
Alfa - α, parametr rozdělení.

Beta - β, parametr rozdělení.

Součet - NEPRAVDA pro hodnotu hustoty $f(x)$,
PRAVDA pro hodnotu distribuční funkce $F(x)$.

Pokud položíme parametr $\alpha=1$, obdržíme exponenciální rozdělení ($\lambda=1/\beta$).

Funkce pro výpočet kvantilů Gama rozdělení má v Excelu název GAMMAINV. Jedná se o kvantilovou funkci $F(x_P)=P(X<x_P)=P$, která má následující argumenty:



Prst - pravděpodobnost P pro hodnotu kvantilu x_P.
Alfa - α, parametr rozdělení.
Beta - β, parametr rozdělení.

Literatura

Nápověda k programu MS Excel
Nápověda k programu Statgraphics Centurion

Doc. RNDr. Luboš Marek, CSc.
Katedra statistiky a pravděpodobnosti
Fakulta informatiky a statistiky
Vysoká škola ekonomická Praha
marek@vse.cz

Mgr. Michal Vrabec
Katedra statistiky a pravděpodobnosti
Fakulta informatiky a statistiky
Vysoká škola ekonomická Praha
vrabec@vse.cz