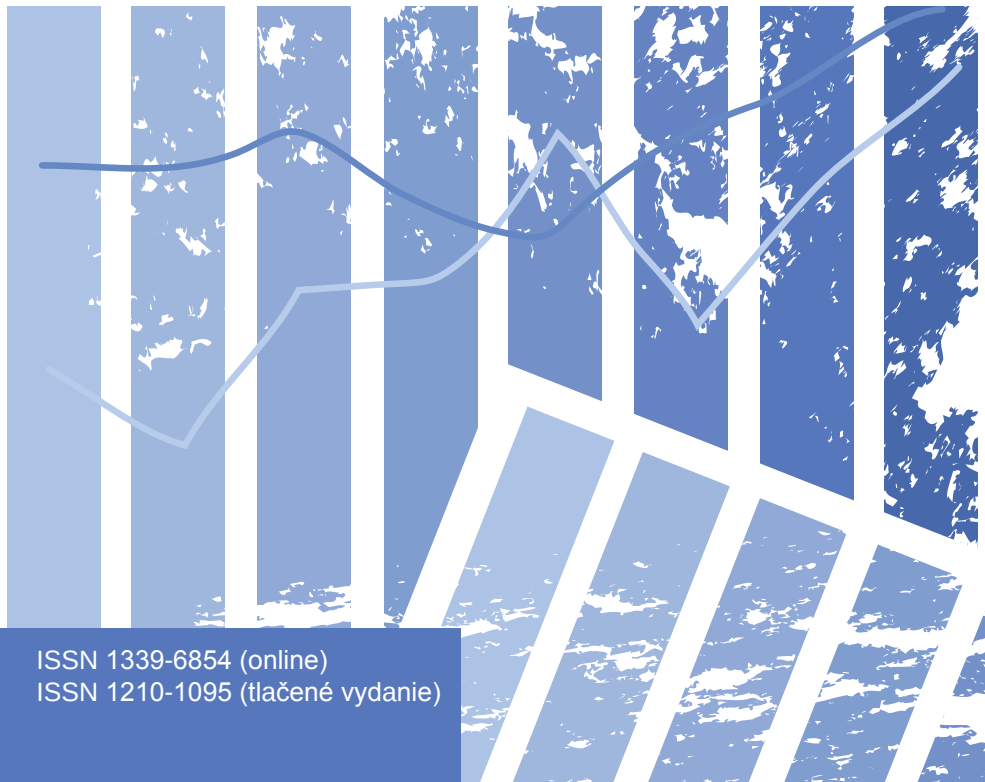
Štatistický úrad Slovenskej republiky
The Statistical Office of the Slovak Republic

# SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

## SLOVAK STATISTICS and DEMOGRAPHY

vedecký časopis/scientific journal

3/2017
ročník 27



ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY

ISSN 1339-6854 (online)
ISSN 1210-1095 (tlačené vydanie)

## SLOVENSKÁ ŠTATISTIKA A DEMOGRAFIA

Recenzovaný vedecký časopis založený v roku 1991. Od roku 2014 sú jednotlivé čísla dostupné čitateľskej verejnosti s trojmesačným odstupom aj v elektronickej forme na www.statistics.sk. Názory autorov článkov sa nemusia zhodovať s názormi vydavateľa.





## SLOVAK STATISTICS AND DEMOGRAPHY

The scientific peer-reviewed journal founded in 1991. From 2014 individual copies of the journal will be available at intervals of three-months also in electronic form at the website www.statistics.sk. The opinions of the authors do not necessarily correlate with the opinions of the publisher.





**Adresa redakcie/Address of Editorial Office**
Slovenská štatistika a demografia
Štatistický úrad SR
Miletičova 3, 824 67 Bratislava
Slovenská republika

**E-mailová adresa/E-mail adress**
SSaD@statistics.sk

**www.statistics.sk**

## OBSAH/CONTENTS

## EDITORIÁL

**Doc. Mgr. Erik Šoltés, PhD.**

Vážení čitatelia,

je už tradíciou, že tretie číslo vedeckého časopisu *Slovenská štatistika a demografia* má monotematické zameranie. Na želanie našich čitateľov sme sa tento rok rozhodli prezentovať v monotematickom čísle metodológie rôznych štatistických metód a postupov s cieľom poukázať na širokú paletu matematicko-štatistických metód využívaných v štatistike ako vednom odbore.

Štatistika má čoraz významnejšiu úlohu vo vede, v priemysle, zdravotníctve, ako aj podnikaní. Adekvátne využitie štatistických metód a správna interpretácia výsledkov získaných z analýz údajov môže relevantne prispieť k takým rozhodnutiam, ktoré zásadne šetria čas a financie. O narastajúcom význame štatistiky pre spoločnosť svedčí aj fakt, že v roku 2010 Organizácia Spojených národov vyhlásila 20. október za Svetový deň štatistiky.

Bežný smrteľník si ani neuvedomuje, že štatistika ho dennodenne ovplyvňuje. Nejde pritom len o štatistické informácie, ktoré prijíma z médií, alebo informácie obsiahnuté v predpovediach počasia. Štatistické skúmanie má niekoľko etáp. S veľkou pravdepodobnosťou je každý z nás pravidelne účastníkom prvej etapy štatistického skúmania, ktorou je štatistické zisťovanie. Cieľom štatistického zisťovania je získavať údaje o hromadných javoch a procesoch. Napríklad už pri bežnom nakupovaní v kamenných obchodoch, používaním zákazníckych kariet, vyhľadávaním informácií a nákupmi cez internet alebo návštevami sociálnych sietí prispievame k zberu údajov bez toho, aby sme si to uvedomovali.

Štatistické metódy využíva každá marketingová kampaň. Dokonca ani umiestňovanie tovarov v predajniach obchodných reťazcov nie je náhodné, ale je výsledkom sledovania správania zákazníkov a jeho analyzovania prostredníctvom štatistiky. Uvediem aj iný príklad využitia štatistiky. Pri výrobe liekov, výživových doplnkov a kozmetických prípravkov sa pomocou štatistických metód nastavuje koncentrácia účinnej látky tak, aby dávka bola bezpečná, ale účinná. A nakoniec aj výrobný proces podlieha kontrole kvality, ktorá je založená na štatistických metódach.

Na jednej strane každý z nás poskytuje údaje pre štatistické skúmania, na druhej strane každého z nás ovplyvňujú závery a rozhodnutia založené na štatistických analýzach. V súčasnosti, keď nás všade obklopujú informačné technológie a všetko je merané, máme množstvo údajov (prierezových a longitudinálnych), ktoré sa stávajú užitočnými len vďaka správnemu aplikovaniu štatistických metód.

Štatistika ako matematická vedná disciplína sa člení na mnoho oblastí, a ako sme už uviedli, jej využitie je širokospektrálne. Aktuálne číslo časopisu *Slovenská štatistika a demografia*, ktoré držíte v rukách, sa zameriava len na vybrané oblasti. Keďže každá štatistická analýza má byť založená na kvalitnej databáze, prvý článok je z oblasti výberového skúmania. V ďalších článkoch sú prezentované matematicko-štatistické metódy z oblasti štatistickej indukcie (analýza rozptylu, bayesovská

štatistika), viacrozmerných štatistických metód (zhluková analýza) a prediktívneho modelovania (rozhodovacie stromy).

Autormi článkov sú erudovaní vedecko-pedagogickí pracovníci renomovaných slovenských a českých univerzít. Vo svojej výskumnej činnosti a pedagogickom procese využívajú rôzne štatistické a analytické softvéry, napr. SAS, SPSS, STATISTICA, SYSTAT, STATGRAPHICS, alebo programovacie jazyky, ako napr. jazyk a prostredie R, bez ktorých by bola nepredstaviteľná aplikácia sofistikovaných štatistických metód na údaje obsiahnuté vo veľkých databázach. V príspevkoch, ktoré vám prinášame v tomto čísle, nájdete aplikácie v softvéroch SAS, SYSTAT a STATGRAPHICS, ako aj v programovacom jazyku R.

Veríme, že vedecké články a ďalšie príspevky publikované v čísle 3/2017 *Slovenskej štatistiky a demografie* budú pre našich čitateľov obohacujúce a podnetné. Tým, ktorí sa zaujímajú o iné oblasti štatistiky, napr. o regresnú a korelačnú analýzu, analýzu časových radov a analýzu kategoriálnych údajov, odporúčam do pozornosti vedecké články z niektorých minulých, ale perspektívne aj budúcich čísel nášho časopisu.

**Doc. Mgr. ERIK ŠOLTÉS, PhD.**

*Autor je prodekanom Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave pre vedu a doktorandské štúdium. Ako člen redakčnej rady časopisu Slovenská štatistika a demografia bol spolu s doc. Ing. Ivetou Stankovičovou, PhD., gestorom prípravy monotematického čísla 3/2017.*

EDITORIAL

Dear readers,

It is a tradition that the third issue of the scientific journal *Slovak Statistics and Demography* has a monothematic orientation. According to our readers´ wishes, this year we decided to present in this monothematic issue, the methodologies of various statistical methods and procedures in order to emphasize the wide range of mathematical and statistical methods used in statistics as a scientific discipline.

Statistics has an increasingly significant role in science, industry, health care and also in business. An adequate use of statistical methods and the proper interpretation of the results obtained from data analyses may relevantly contribute to time-saving and economic decisions. The increasing importance of statistics for the society is evidenced by the fact that in 2010, the October 20$^{th}$ was declared as the World Statistics Day by the United Nations.

Ordinary mortals hardly realize the impact of statistics on everyday life. This is more than just statistical information received by media or information contained in weather forecasts. Statistical examination has several stages. Every one of us is likely to regularly participate in the first stage of the examination which is the statistical survey. The aim of a statistical survey is to obtain data on collective phenomena and processes. For example while ordinary shopping in brick-and-mortar stores, using store cards, by searching information and online shopping or visiting social networking sites we contribute to data collection without being aware of it.

Every marketing campaign uses statistical methods. Even the products in retail stores of supermarket chains are not randomly placed but according to consumer behaviour and its statistical analysis. I will give you another example of the uses of statistics. In the manufacture of medicinal products, food supplements and cosmetic products, the concentration of the active substance is adjusted so that the dose will be safe but efficient. Lastly, the production process is subject to quality inspection using statistical methods.

On the one hand, every one of us is providing data for the statistical examinations, on the other hand, we are all affected by conclusions and decisions based on statistical analyses. Currently, when we are everywhere surrounded by information technology and everything is quantified, we have a great deal of data (cross-sectional and longitudinal) which are becoming useful only thanks to the proper application of statistical methods.

Statistics as a mathematical science is divided into several fields and, as already indicated, it has a wide-ranging use. The current issue of the journal *Slovak Statistics and Demography*, you are obtaining, is devoted only to specific fields. Whereas every statistical analysis shall be based on a high quality database, the first article is a sample examination. Other articles present mathematical and statistical methods from the field of statistical induction (analysis of variance, Bayesian statistics), multi-dimensional statistical methods (cluster analysis) and predictive modelling (decision trees).

The authors of the articles are qualified education and science workers of renowned Slovak and Czech universities. They use in their research activity and in the pedagogical process various statistical and analytical software, e.g. SAS, SPSS, STATISTICA, SYSTAT, STATGRAPHICS or programming languages such as the R language and environment, without which the application of sophisticated statistical methods on data contained in large databases would be inconceivable. In the articles we bring you in this issue, you will find applications in SAS, SYSTAT and STATGRAPHICS software, as well as in the R programming language.

We do believe that our readers may find the scientific articles and other contributions published in the issue No 3 (2017) of the *Slovak Statistics and Demography* enriching and inspiring. For those interested in other statistical fields, for example in the regression and correlation analysis, time series and categorical data analysis, I highly recommend the scientific articles of some of the previous or the prospective future issues of our Journal.

**Assoc. Prof. ERIK ŠOLTÉS, PhD.**

*The author is a Vice-Dean of the Faculty of Economic Informatics of the University of Economics in Bratislava for Science and Doctoral Studies. As a member of the Editorial Board of the Journal Slovak Statistics and Demography together with the Assoc. Prof. Iveta Stankovičová, PhD. he was responsible for the preparation of the monothematic issue No 3 (2017).*

**Milan TEREK**
**Katedra štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave**

# NAVRHOVANIE KOMPLEXNÝCH ŠTATISTICKÝCH PRIESKUMOV A NIEKTORÉ MOŽNOSTI ANALÝZY Z NICH ZÍSKANÝCH DÁT

# DESIGNING COMPLEX STATISTICAL SURVEYS AND SOME POSSIBILITIES OF DATA ANALYSIS OBTAINED THEREFROM

## ABSTRAKT
Článok sa zaoberá možnosťami navrhovania komplexných štatistických prieskumov a vybranými analýzami dát získaných z týchto prieskumov. Uvádza sa v ňom všeobecný postup návrhu výberovej schémy komplexného štatistického prieskumu. Analýzu dát z komplexného prieskumu možno realizovať postupnosťou krokov odhadovania charakteristík z najnižšej po najvyššiu úroveň výberovej schémy. Využitie výberových váh môže proces analýzy dát z komplexného prieskumu často značne zjednodušiť. Článok sa venuje opisu formulácie empirickej pravdepodobnostnej funkcie, empirickej distribučnej funkcie a odhadovania mediánu s využitím výberových váh. Na základe dát z komplexného prieskumu EU SILC 2014 v Slovenskej republike sa odhadujú a porovnávajú mediány celkových hrubých príjmov domácností v ôsmich slovenských regiónoch.

## ABSTRACT
The paper deals with the possibilities of designing complex statistical surveys and selected analysis of data obtained from these surveys. A description is provided for the general approach to the sampling scheme of the complex statistical survey. Data analysis from the complex survey can be realized by sequence of actions estimating the characteristics from lowest to the highest level of the sampling scheme. The use of sampling weights can often greatly simplify the data analysis process from complex surveys. The paper describes the construction of empirical probability mass function, empirical cumulative distribution function and estimation of median with the use of sampling weights. The medians of the total gross household incomes in eight Slovak regions were estimated and compared on the basis of the data from the Slovak Republic: EU-SILC, 2014.

## KĽÚČOVÉ SLOVÁ
komplexný štatistický prieskum, výberové váhy, empirická pravdepodobnostná funkcia, empirická distribučná funkcia, odhadovanie mediánu

## KEYWORDS
complex statistical survey, sampling weights, empirical probability mass function, empirical cumulative distribution function, median estimation

## 1. ÚVOD

Štatistický prieskum[1], ktorý obsahuje viacero takých komponentov, ako je napríklad náhodné vyberanie[2], stratifikácia, skupinové vyberanie, vyberanie s nerovnakými pravdepodobnosťami, pomerové odhadovanie a podobne, sa zvyčajne nazýva komplexný štatistický prieskum. Uvedieme moduly na konštrukciu komplexných štatistických prieskumov a základné možnosti analýzy dát, ktoré sme z nich získali.

Všetky prezentované procedúry vyžadujú využitie pomocných informácií. Všeobecne, pomocné informácie sú ľubovoľné informácie, ktoré nepochádzajú z výberu a ktoré môžu zlepšiť presnosť hodnôt odhadov[3]. Pomocné informácie možno využiť v etape tvorby plánu výberového skúmania aj v etape odhadovania pri konštrukcii bodových odhadov[4]. Slúžia na vytvorenie vhodnej výberovej schémy (*sampling design*) a/alebo na výpočet hodnôt odhadov. V oboch prípadoch sa nazývajú pomocné premenné[5]. Hodnoty pomocných premenných možno často získať z rozličných registrov, napríklad z obchodného registra alebo z registra obyvateľov. V návrhoch komplexných štatistických prieskumov je veľmi dôležitá voľba vhodných pomocných premenných. Metódy na identifikáciu najvhodnejších pomocných premenných v súvislosti s vychýlením bodových odhadov sú podrobne opísané v [10].

V článku sa podrobnejšie zmienime o možnostiach využitia výberových váh pri konštrukcii empirickej pravdepodobnostnej funkcie, empirickej distribučnej funkcie a pri odhadovaní niektorých charakteristík základného súboru. Postupy budeme ilustrovať na analýze dát z komplexného štatistického prieskumu EU SILC (European Union Statistics on Income and Living Conditions), ktorý sa na Slovensku realizoval v roku 2014. Budeme analyzovať regionálnu štruktúru príjmov na základe hodnôt odhadov mediánu celkových hrubých príjmov domácností v ôsmich slovenských regiónoch.

## 2. MATERIÁL A METÓDY

Článok poskytne charakteristiku niektorých základných komponentov komplexných štatistických prieskumov, spôsob ich navrhovania a základné spôsoby analýzy dát z komplexných štatistických prieskumov. Uvedieme možnosti využívania výberových váh pri odhadovaní rozdelenia pravdepodobnosti pre základný súbor a pri odhadovaní niektorých charakteristík základného súboru.

---

[1] *Štatistický prieskum (statistical survey, survey) je proces zhromažďovania dát prostredníctvom zisťovania odpovedí jednotiek (osôb, domácností, firiem a pod.) na otázky. V rovnakom význame sa používajú aj termíny anketa alebo zisťovanie.*

[2] *Výberový súbor alebo výber (sample) je vybratá časť základného súboru. Postup získavania výberu nazveme vyberanie (sampling). Náhodné alebo pravdepodobnostné vyberanie (random sampling, probability sampling) je také vyberanie, že pravdepodobnosť každého výberu z daného základného súboru je známa. Množina hodnôt pozorovaní, ktorá sa získala náhodným vyberaním, sa nazýva náhodný výber (random sample). V induktívnej štatistike sa hodnoty v náhodnom výbere považujú za realizácie náhodných premenných – pozorovaní. Množina týchto pozorovaní sa tiež nazýva náhodný výber.*

[3] *Keď sa použije hodnota v nejakej výberovej charakteristike V na odhadnutie charakteristiky θ základného súboru, ide o bodové odhadovanie (point estimation). Samotná výberová charakteristika V sa nazýva bodovým odhadom (point estimator) charakteristiky θ a jej hodnota v sa nazýva hodnotou bodového odhadu (point estimate) V charakteristiky θ. Pri bodovom odhadovaní sa charakteristika základného súboru odhaduje jediným číslom alebo jediným bodom na osi reálnych čísel.*

[4] *Podrobnejšie pozri v [4].*

[5] *Podrobnejšie pozri napríklad v [9], [10].*

## 2.1. Navrhovanie komplexných štatistických prieskumov

Budeme charakterizovať viaceré komponenty komplexného štatistického prieskumu: jednoduché náhodné vyberanie, stratifikácia, skupinové vyberanie atď. Ďalej uvedieme, ako z nich možno vytvoriť jedinú výberovú schému.

### 2.1.1. Komponenty komplexného štatistického prieskumu

Jednoduché náhodné vyberanie je najjednoduchšia forma náhodného vyberania. Ide o náhodné vyberanie jednotiek[6] bez opakovania alebo s opakovaním, ktoré sa realizuje priamo z celého základného súboru. Pri jednoduchom náhodnom vyberaní má každá možná podmnožina *n* jednotiek v základnom súbore rovnakú pravdepodobnosť tvoriť náhodný výber rozsahu *n*. Výsledkom jednoduchého náhodného vyberania je jednoduchý náhodný výber.

Jednoduché náhodné vyberanie je najjednoduchšia výberová schéma. Všimnime si teraz stratifikované náhodné vyberanie. Ide o náhodné vyberanie, v ktorom sa základný súbor delí na vzájomne sa vylučujúce a základný súbor celkom pokrývajúce podsúbory, ktoré sa nazývajú stratá. Stratá sa vzhľadom na skúmanú premennú považujú za viac homogénne ako celý základný súbor. Z každého strata sa získa jednoduchý náhodný výber, pričom jednotlivé výbery sa zo strát vyberajú nezávisle. Súbor vytvorený zo všetkých získaných výberov tvorí stratifikovaný náhodný výber zo základného súboru. Stratá sa najčastejšie definujú na základe záujmových podskupín základného súboru (domén), napríklad môže ísť o regióny krajiny pri prieskumoch, ktoré sa týkajú obyvateľstva krajiny, alebo veľkostné kategórie firiem pri prieskumoch, ktoré sa týkajú firiem. Jednotky v tom istom strate majú obyčajne tendenciu viac sa podobať ako jednotky náhodne vybraté z celého základného súboru, preto stratifikácia často zvyšuje presnosť hodnôt odhadov. Stratifikácia sa často používa na redukciu variability bodových odhadov a na získavanie separovaných hodnôt odhadov pre domény. Stratifikované viacstupňové vyberanie je založené na využití hierarchickej štruktúry jednotiek v každom strate.

Skupinové vyberanie je náhodné vyberanie, v ktorom každá jednotka patrí do určitej skupiny a skupiny sa vyberajú podľa konkrétnej výberovej schémy. Pri jednostupňovom skupinovom vyberaní je každá jednotka z náhodne vybratých skupín zaradená do výberu. Pri dvojstupňovom skupinovom vyberaní sa náhodne vyberú len niektoré jednotky z náhodne vybratých skupín. Všeobecne je možné navrhnúť ľubovoľný konečný počet stupňov. Skupinové vyberanie, v ktorom všetky skupiny v základnom súbore nemajú rovnakú pravdepodobnosť byť vybraté do výberu, sa nazýva skupinové vyberanie s nerovnakými pravdepodobnosťami. Často sa navrhuje skupinové vyberanie s pravdepodobnosťami výberu úmernými veľkosti skupiny. Skupinové vyberanie, niekedy s nerovnakými pravdepodobnosťami, sa zvyčajne navrhuje z dôvodu redukcie nákladov spojených s prieskumom. Často sa v komplexných štatistických prieskumoch využíva pomerové a regresné odhadovanie.

### 2.1.2. Postup pri navrhovaní komplexných štatistických prieskumov

Pri návrhu komplexného štatistického prieskumu sa odporúča postupovať tak, že sa známe koncepty formulujú v modulárnej forme[7]. Moduly sa potom v rozličných možných kombináciách integrujú do jedinej výberovej schémy. Prvý modul je

[6] *Nie skupín jednotiek.*
[7] *Podrobnejšie v [7], s. 281 – 282.*

založený na stratifikácii, druhý na skupinovom vyberaní s opakovaním[8] a tretí na skupinovom vyberaní bez opakovania[9].

Stratifikácia je obyčajne jadrom výberovej schémy. Stratá môžu byť napríklad regióny v krajine, typy sídiel a podobne. Skupiny (niekedy viacero stupňov skupín) sa vyberajú z každého strata v návrhu a vnútri skupín sa môže objaviť dodatočná stratifikácia. Veľa prieskumov používa stratifikované viacstupňové vyberanie, v ktorom sa využíva stratifikovaný výber primárnych jednotiek a podvýbery sekundárnych jednotiek sa vyberajú z každej vybratej primárnej jednotky. Keď sa komplexný prieskum skladá z viacstupňového skupinového vyberania a stratifikácie, je užitočné vytvoriť diagram alebo tabuľku výberovej schémy. Niekedy sa oplatí využiť pri odhadovaní aj pomerové odhadovanie. To sa bežne využíva na ľubovoľnej úrovni výberovej schémy.

V [7] sa na s. 282 – 283 uvádza zaujímavý príklad. V roku 1991 sa v Gambii vo vidieckych sídlach realizoval prieskum zameraný na mieru používania ochranných sietí na posteľ impregnovaných insekticídom proti komárom prenášajúcim maláriu. Výberová báza (opora výberu) pozostávala z vidieckych sídiel v Gambii s 3 000 a menej obyvateľmi. Vidiecke sídla boli stratifikované podľa dvoch stratifikačných premenných − regiónu (východný, centrálny a západný) a existencie verejnej kliniky (áno, nie). Stratifikácia sa realizovala v troch stupňoch. V každom regióne sa vybralo päť okresov (*districts*) s pravdepodobnosťami úmernými počtu obyvateľov okresu. V druhom stupni sa v každom vybratom okrese vybrali štyri vidiecke sídla opäť s pravdepodobnosťami úmernými počtu obyvateľov – dve s verejnými klinikami a dve bez nich. Nakoniec sa v každom vidieckom sídle náhodne vybralo šesť objektov (*compounds*) a v nich sa zistil počet postelí a ochranných sietí spolu s inými informáciami. Výberová schéma je charakterizovaná v tabuľke 1.

***Tabuľka č. 1: Výberová schéma***

| Stupeň | Výberová jednotka | Stratifikácia |
|---|---|---|
| 1 | okres | región |
| 2 | vidiecke sídlo | existencia verejnej kliniky |
| 3 | objekt | |

***Zdroj údajov: vlastné spracovanie podľa [7], s. 283***

Pri výpočte hodnôt odhadov alebo smerodajných chýb sa začína na 3. stupni a postupuje sa smerom nahor. Odhadovanie úhrnu (celkového počtu) ochranných sietí vo vidieckych sídlach v Gambii pozostávalo zo 6 krokov. Zaznamenal sa celkový počet ochranných sietí pre každé vidiecke sídlo, vypočítala sa hodnota odhadu úhrnu sietí pre každé vidiecke sídlo, hodnota odhadu úhrnu pre vidiecke sídla s verejnou klinikou a rovnako bez verejnej kliniky v každom okrese, hodnota odhadu počtu sietí v každom okrese, hodnota odhadu úhrnu sietí pre každý región a nakoniec hodnota odhadu úhrnu sietí v Gambii. Podobne sa odhadli aj rozptyly. Celý postup je pomerne komplikovaný. Vo všeobecnosti nie je nevyhnutné pri odhadovaní v komplexných štatistických prieskumoch používať pomerne zložité postupy ako v spomínanom príklade. V mnohých prípadoch odhadovanie uľahčuje využitie výberových váh.

[8] *Náhodným vyberaním s opakovaním sa vyberie n skupín – primárnych jednotiek.*
[9] *Náhodným vyberaním bez opakovania sa vyberie n primárnych jednotiek.*

## 2.2. Výberové váhy a ich využitie pri odhadovaní

Vo výberových skúmaniach sa najčastejšie odhaduje stredná hodnota, úhrn, podiel alebo pomer. Niekedy je potrebné odhadovať medián a iné kvantily[10]. Vhodným prostriedkom nato sú výberové váhy. Výberové váhy umožňujú aj konštrukciu empirického rozdelenia pre základný súbor[11]. Výberové váhy možno vypočítať na základe pomocných informácií.

Predpokladajme že poznáme rozsah *N* konečného základného súboru *U*. Symbolom *x* označíme študovanú premennú aj jej hodnoty, $U = \{1, 2, \ldots N\}$ je množina indexov jednotiek v základnom súbore. Symbol *S* označuje výber zo základného súboru – podmnožinu, ktorá obsahuje *n* jednotiek z *U*. Nech $\pi_i$ je pravdepodobnosť, že jednotka $i \in U$ bude v náhodnom výbere. Výberové váhy $w_i$ pre ľubovoľnú výberovú schému sú definované takto:

$$w_i = \frac{1}{\pi_i} \ . \tag{1}$$

Výberovú váhu jednotky *i* možno interpretovať ako počet jednotiek v základnom súbore, ktoré sú reprezentované jednotkou *i* vo výbere.

V jednoduchom náhodnom výbere má každá jednotka v základnom súbore pravdepodobnosť $\pi_i = n/N$ byť v náhodnom výbere. V tejto výberovej schéme je výberová váha každej jednotky v základnom súbore $w_i = 1/\pi_i = N/n$. Každá jednotka v jednoduchom náhodnom výbere reprezentuje samu seba a $N/(n - 1)$ ďalších nevybratých jednotiek v základnom súbore. V jednoduchom náhodnom výbere

$$\sum_{i \in S} w_i = \sum_{i \in S} \frac{N}{n} = N \ . \tag{2}$$

Je známe, že stredná hodnota $\mu_K$ premennej *x* konečného základného súboru je definovaná takto:

$$\mu_K = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{3}$$

a úhrn $\tau$ konečného základného súboru je

$$\tau = \sum_{i=1}^{N} x_i = N \mu_K \ . \tag{4}$$

V jednoduchom náhodnom výbere je výberový priemer $\bar{X}$ nevychýleným bodovým odhadom strednej hodnoty $\mu_K$. Jeho hodnota $\bar{x}$ sa vypočíta podľa vzťahu

$$\bar{x} = \frac{1}{n} \sum_{i \in S} x_i \ . \tag{5}$$

[10] *Podrobnejšie o kvantiloch pozri napríklad v [13].*
[11] *V skutočnosti ide o empirické rozdelenie pozorovania zo skúmaného konečného základného súboru.*

Na odhadovanie úhrnu $\tau$ možno použiť nevychýlený bodový odhad $N\overline{X}$ .

Pomocou výberových váh možno úhrn a strednú hodnotu konečného základného súboru odhadnúť nasledujúcimi hodnotami:

$$N\overline{x} = \sum_{i \in S} \frac{N}{n} x_i = \sum_{i \in S} w_i x_i \tag{6}$$

$$\overline{x} = \frac{N\overline{x}}{N} = \frac{\sum_{i \in S} w_i x_i}{\sum_{i \in S} w_i} \ . \tag{7}$$

Všimnime si teraz stratifikovaný základný súbor. Nech je základný súbor rozsahu *N* rozdelený na *H* strát. Stredná hodnota *h*-teho strata je definovaná takto:

$$\mu_h = \frac{1}{N_h} \sum_{i=1}^{N_h} x_{hi} \ , \tag{8}$$

kde $x_{hi}$ je hodnota premennej *x*, *i*-tej jednotky v *h*-tom strate, $N_h$ – rozsah *h*-teho strata v základnom súbore.

Stredná hodnota $\mu_K$ premennej *x* je definovaná takto:

$$\mu_K = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h \tag{9}$$

Úhrn $\tau$ v stratifikovanom základnom súbore je definovaný takto:

$$\tau = \sum_{h=1}^{H} N_h \mu_h \tag{10}$$

Hodnota $\overline{x}_h$ výberového priemeru $\overline{X}_h$ v stratách sa vypočíta podľa vzťahu

$$\overline{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi} \ , \tag{11}$$

kde $n_h$ je rozsah výberu z *h*-teho strata.

Z každého strata sa náhodným vyberaním bez opakovania získa výberový súbor. Výberový priemer

$$\overline{X}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{X}_h = \sum_{h=1}^{H} \frac{N_h}{N} \cdot \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi} = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} X_{hi} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{N_h}{n_h} X_{hi} \tag{12}$$

je nevychýleným bodovým odhadom strednej hodnoty $\mu_K$ stratifikovaného základného súboru. Výberový úhrn

$$N\overline{X}_{str} = \sum_{h=1}^{m} N_h \overline{X}_h = \sum_{h=1}^{m} \frac{N_h}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{h=1}^{m} \frac{N_h}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{N_h}{n_h} X_{hi} \tag{13}$$

je nevychýleným bodovým odhadom úhrnu $\tau$ stratifikovaného základného súboru. Výberová váha *i*-tej jednotky z *h*-teho strata je

$$w_{hi} = \frac{N_h}{n_h} . \tag{14}$$

Výberovú váhu $w_{hi}$ možno interpretovať ako počet jednotiek v *h*-tom strate základného súboru reprezentovaných *i*-tou jednotkou z *h*-teho strata vo výbere. Zrejme

$$\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} = N. \tag{15}$$

Pomocou výberových váh možno strednú hodnotu a úhrn stratifikovaného základného súboru odhadnúť hodnotami

$$\overline{x}_{str} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} x_{hi} \tag{16}$$

a

$$N\overline{x}_{str} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} x_{hi} . \tag{17}$$

Majme základný súbor, v ktorom je napríklad 2 000 mužov a 500 žien. Nech je pohlavie stratifikačnou premennou. Uvažujeme teda o dvoch stratách – muži a ženy. Jednoduchým náhodným vyberaním sme z každého strata vybrali 100 jednotiek. V stratifikovanom náhodnom výbere je potom 100 mužov a 100 žien. Pravdepodobnosť výberu muža je 100/2 000 = 1/20 a pravdepodobnosť výberu ženy je 100/500 = 1/5. Výberová váha každého muža vo výbere je 20, výberová váha každej ženy vo výbere je 5. Každý muž vo výbere reprezentuje 20 mužov v základnom súbore, každá žena vo výbere reprezentuje 5 žien v základnom súbore. Zrejme suma váh sa rovná rozsahu základného súboru

$$\sum_{h=1}^{2} \sum_{i=1}^{100} w_{hi} = 100 \cdot 20 + 100 \cdot 5 = 2\,500.$$

Bodové odhady uvedených a iných charakteristík konečného základného súboru v skupinovom vyberaní a v iných výberových schémach vrátane ich kombinácií takých, ako napríklad viacstupňové stratifikované skupinové vyberanie, možno vyjadriť pomocou výberových váh.

Výberové váhy sa dajú modifikovať vzhľadom na neodpovedanie a na chyby pokrytia. Doteraz sme uvažovali len o tzv. základných váhach (*base weights*), ktoré sú odvodené z plánu výberového skúmania. Označme ich $w_{Bi}$. Váhy $w_{NRi}$ sa považujú

za faktory úpravy vzhľadom na neodpovedanie. Váhy $w_{NCi}$ sa považujú za faktory kompenzácie nepokrytia[12]. Posledné dva uvedené typy váh sa považujú za faktory úpravy základných váh[13].

Konečná výberová váha $w_i$ pre *i*-tú jednotku vo výbere je

$$w_i = w_{Bi} \cdot w_{NRi} \cdot w_{NCi} . \tag{18}$$

Výberové váhy pre všetky pozorovania sú rovnaké v samovážiacich výberoch. Takéto výbery možno považovať[14] za reprezentatívne v tom zmysle, že každá pozorovaná jednotka reprezentuje rovnaký počet nepozorovaných jednotiek v základnom súbore[15]. Výberové váhy pre všetky pozorovania nie sú rovnaké v nesamovážiacich výberoch.

### 2.2.1. Odhadovanie empirickej pravdepodobnostnej funkcie, empirickej distribučnej funkcie a niektorých charakteristík základného súboru pomocou výberových váh

Predpokladajme, že sú známe hodnoty premennej *x* pre všetkých *N* jednotiek v základnom súbore. Hodnota pravdepodobnostnej funkcie v bode *x* je

$$p(x) = \frac{N_{(x)}}{N} , \tag{19}$$

kde $N_{(x)}$ je počet jednotiek, ktoré majú hodnotu premennej *x*. Hodnota distribučnej funkcie v bode *x* je

$$F(x) = \sum_{y \le x} p(y) . \tag{20}$$

Poznamenajme, že ide o pravdepodobnostnú funkciu a distribučnú funkciu pozorovania zo základného súboru[16].

Výberové váhy umožňujú formulovať empirickú pravdepodobnostnú funkciu a empirickú distribučnú funkciu. Empirická pravdepodobnostná funkcia $\hat{p}(x)$ je definovaná ako suma váh všetkých pozorovaní, ktoré majú hodnotu *x,* delená sumou všetkých váh

$$\hat{p}(x) = \frac{\sum_{i \in S ; x_i = x} w_i}{\sum_{i \in S} w_i} . \tag{21}$$

---

*[12] Jednotky, ktoré sú v cieľovom základnom súbore, ale nie sú vo výberovej báze, vytvárajú nepokrytie alebo neúplné pokrytie.*

*[13] Podrobnejšie o výberových váhach pozri napríklad v [6] alebo v [11].*

*[14] V prípade absencie nevýberových chýb.*

*[15] Termín „reprezentatívny výber“ môže mať rozličné významy. V [2], s. 23 – 24 je uvedených deväť rozličných významov tohto termínu, ktoré sa bežne používajú. Odporúča sa používať tento termín len s vysvetlením jeho chápania v konkrétnom texte.*

*[16] Uvažujeme o prístupe k výberovému skúmaniu známemu ako prístup „bez modelu“ alebo „bez rozdelenia“ (podrobnejšie pozri v [3], s. 8 – 9).*

Empirická distribučná funkcia $\hat{F}(x)$ je

$$\hat{F}(x) = \sum_{y \le x} \hat{p}(y). \tag{22}$$

### 2.2.2. Odhadovanie niektorých charakteristík základného súboru pomocou výberových váh

Pomocou empirickej pravdepodobnostnej funkcie $\hat{p}(x)$ a empirickej distribučnej funkcie $\hat{F}(x)$ možno odhadovať charakteristiky základného súboru. Napríklad strednú hodnotu základného súboru možno odhadnúť hodnotou

$$\hat{\mu}_K = \sum_x x\hat{p}(x) = \frac{\sum_{i \in S} w_i x_i}{\sum_{i \in S} w_i}. \tag{23}$$

Keby bola distribučná funkcia *F* spojitá, medián základného súboru by bol definovaný ako hodnota $\tilde{\mu}$, pre ktorú $F(\tilde{\mu}) = \frac{1}{2}$. V diskrétnom prípade je medián konečného základného súboru definovaný ako hodnota $\tilde{\mu}$, pre ktorú $F(\tilde{\mu}) = \frac{1}{2}$, ak taká hodnota existuje. Inak je medián konečného základného súboru definovaný ako ľubovoľná hodnota z intervalu $[\tilde{\mu}_1, \tilde{\mu}_2]$, kde $\tilde{\mu}_1$ je najväčšia hodnota *x* v základnom súbore, pre ktorú $F(x) < \frac{1}{2}$ a $\tilde{\mu}_2$ je najmenšia hodnota *x,* pre ktorú $F(x) > \frac{1}{2}$. Všeobecne $Q_p$ je $p \cdot 100$ % kvantil (percentil), keď $F(Q_p) = p$, ak taká hodnota existuje, inak $Q_p \in [a, b]$, kde *a* je najväčšia hodnota *x* v základnom súbore, pre ktorú $F(x) < p$ a *b* je najmenšia hodnota *x,* pre ktorú $F(x) > p$. Keď $p < \frac{1}{N}$, $Q_p$ je najmenšia hodnota *x,* a keď $p > 1 - \frac{1}{N}$, $Q_p$ je najväčšia hodnota *x*.

V ďalšej časti článku ukážeme, ako sa odhadujú kvantily v základnom súbore. Pretože empirická distribučná funkcia $\hat{F}$ je kroková funkcia, na nájdenie jedinej hodnoty kvantilu je obyčajne potrebná interpolácia. Nech $y_1$ je najväčšia hodnota vo výbere, pre ktorú $\hat{F}(y_1) \le p$ a, nech $y_2$ je najmenšia hodnota vo výbere, pre ktorú $\hat{F}(y_2) \ge p$, potom

$$\hat{Q}_p = y_1 + \frac{p - \hat{F}(y_1)}{\hat{F}(y_2) - \hat{F}(y_1)}(y_2 - y_1). \tag{24}$$

Poznamenajme, že bodové odhady založené na použití výberových váh nie sú nevyhnutne nevychýlené alebo numericky stabilné. Napriek tomu hodnoty štatistík, ktoré sa počítajú pomocou výberových váh, sú obyčajne oveľa bližšie skutočným hodnotám charakteristík základného súboru ako v prípade použitia bodových odhadov, ktoré neberú do úvahy štruktúru dát ([7], s. 293).

## 3. ANALÝZA REGIONÁLNEJ ŠTRUKTÚRY PRÍJMOV NA BÁZE DÁT Z EU SILC 2014

Rozdelenia príjmov sú obyčajne výrazne zošikmené a obsahujú odľahlé hodnoty[17]. Výpovedná schopnosť strednej hodnoty v takýchto rozdeleniach je veľmi malá[18] a stredná hodnota sa nepovažuje za vhodnú charakteristiku centra rozdelenia. Stredná hodnota príjmu nie je vhodnou charakteristikou „typického“ príjmu. V takýchto rozdeleniach sa všeobecne považuje za dobrú charakteristiku centra rozdelenia medián. Ide o stabilnú charakteristiku, robustnú voči odľahlým hodnotám. Alternatívne možno odporúčať ako vhodné charakteristiky centra rozdelenia aj niektoré netradičné charakteristiky, napríklad zastrihnutú strednú hodnotu (*trimmed mean*)[19], winsorizovanú strednú hodnotu alebo M-estimátory[20]. Podobné výsledky poskytujú aj tradičné charakteristiky aplikované na redukovaný súbor dát[21].

Prieskumy EU SILC sa vykonávajú každoročne vo všetkých krajinách Európskej únie vrátane Slovenskej republiky. Týkajú sa domácností a osôb. Na úrovni domácností sa zhromažďujú dáta o viacerých kategóriách príjmov. Podobné prieskumy sa vykonávajú aj v mnohých krajinách mimo Európskej únie. V Slovenskej republike sa prieskum EU SILC realizuje ako stratifikovaný, dvojstupňový s dvomi stratifikačnými premennými – región a veľkosť sídla. Zisťovanie EU SILC 2014 sa uskutočnilo vo vybraných 6 010 domácnostiach. Databázu tvoria dáta o 5 490 domácnostiach a 13 433 osobách starších ako 16 rokov. Vypočítané výberové váhy boli modifikované vzhľadom na neodpovedanie. Tieto váhy možno využívať na tvorbu induktívnych úsudkov o domácnostiach v Slovenskej republike. Iné váhy boli vypočítané pre každú z osôb. Výberové dáta z EU SILC sú vo všeobecnosti dátami z komplexného štatistického prieskumu a výberový súbor je nesamovážiaci.

Dáta z EU SILC 2014 sa nachádzajú vo viacerých súboroch. Každej domácnosti je priradené identifikačné číslo[22]. Bola vykonaná analýza celkových hrubých príjmov domácností v ôsmich doménach – slovenských regiónoch. Tieto domény korešpondujú s hodnotami jednej zo stratifikačných premenných. Najskôr sa vykonalo spárovanie dát – výberových váh[23] a celkových hrubých príjmov domácností[24] podľa čísel domácností. Potom sa spárovali dáta rozdelené podľa regiónov. Takýmto spôsobom vzniklo osem súborov dát, jeden pre každý región. Každý región sa analyzoval separovane. Potom sa pre každý región podľa (21) vypočítali hodnoty empirickej pravdepodobnostnej funkcie. Na základe týchto hodnôt sa podľa (22) vypočítali pre každý región hodnoty empirickej distribučnej funkcie. Nakoniec sa pre každý región podľa vzťahu (24) vypočítala hodnota odhadu mediánu celkových hrubých príjmov domácností. Všetky výpočty sa realizovali v programe Excel 2013. Získané výsledky sú v tabuľke č. 2. Hodnotami odhadu mediánu v treťom stĺpci tabuľky odhadujeme mediány celkových hrubých príjmov domácností v slovenských regiónoch v roku 2014.

---

[17] *Odľahlú hodnotu v množine dát možno definovať ako hodnotu (alebo podmnožinu hodnôt), ktorá sa zdá nekonzistentná s ostatnými hodnotami v množine dát ([1], s. 7).*
[18] *Podrobnejšie pozri v [5].*
[19] *Pozri napríklad v [8], s. 55.*
[20] *Podrobnejšie pozri napríklad v [12] alebo v [14].*
[21] *Podrobnejšie pozri v [12].*
[22] *Hodnoty premenných DB030, HB030.*
[23] *Hodnoty premennej DB090.*
[24] *Hodnoty premennej HY010.*

***Tabuľka č. 2: Medián celkových hrubých príjmov domácností v slovenských regiónoch v roku 2014***

| Číslo regiónu | Región | Hodnota odhadu mediánu celkových hrubých príjmov domácností v roku 2014 (v eurách) | Poradie regiónu podľa mediánu celkových hrubých príjmov domácností |
|---|---|---|---|
| 1 | Bratislava | 14 491,37 | 1. |
| 2 | Trnava | 13 969,12 | 4. |
| 3 | Trenčín | 14 368,47 | 2. |
| 4 | Nitra | 12 379,67 | 7. |
| 5 | Žilina | 14 054,85 | 3. |
| 6 | Banská Bystrica | 11 746,41 | 8. |
| 7 | Prešov | 13 595,22 | 5. |
| 8 | Košice | 13 118,16 | 6. |

***Zdroj údajov: vlastné výpočty***

Medián celkového hrubého príjmu domácností v celej Slovenskej republike v roku 2014 odhadujeme hodnotou 13 305,83 eura.

## 4. ZÁVER

Pri navrhovaní komplexných štatistických prieskumov možno využiť tri základné moduly − stratifikáciu, skupinové vyberanie s opakovaním a skupinové vyberanie bez opakovania. Ich vhodnou kombináciou spolu s prípadným zaradením viacstupňového vyberania, vyberania s nerovnakými pravdepodobnosťami a pomerového alebo regresného odhadovania možno vytvoriť výberovú schému na komplexný štatistický prieskum.

Niekedy sa dajú pri induktívnych úsudkoch o základnom súbore na báze dát z výberového súboru získaného komplexným štatistickým prieskumom použiť štandardné štatistické metódy a bežný softvér, niekedy nie.

Keď bol výber z konečného základného súboru získaný náhodným vyberaním s opakovaním, pozorovania sú štatisticky nezávislé a rovnako rozdelené. Keď sú vo výberovom súbore všetky pozorovania štatisticky nezávislé a rovnako rozdelené, možno na induktívne úsudky o základnom súbore použiť bežné štatistické metódy a bežný štatistický softvér.

Pri náhodných výberoch získaných z komplexných štatistických prieskumov pomocou zložitejších výberových schém nie sú dva spomenuté predpoklady splnené. Keď je výber samovážiaci[25], možno bežné induktívne štatistické metódy a bežný softvér použiť na získanie hodnôt bodových odhadov. Smerodajné chyby, intervaly spoľahlivosti a testy hypotéz, ktoré poskytne bežný softvér, budú už nesprávne.

Keď je výber nesamovážiaci, nemožno bežné induktívne štatistické metódy a bežný softvér použiť ani na bodové odhadovanie. V uvedenej aplikácii bol k dispozícii nesamovážiaci výber, a preto sa nedal odhadovať medián základného

[25] *Napríklad, keď má stratifikovaný náhodný výber rovnaký výberový pomer (z každého strata sa vyberie rovnaké percento jednotiek), majú všetky jednotky vo výbere rovnakú výberovú váhu – výber je samovážiaci.*

súboru pomocou výberového mediánu; bolo nevyhnutné zohľadniť štruktúru dát. Použili sa výberové váhy.

V [7] na s. 287 – 288 sa uvádza: „Keď čítate článok alebo knihu, v ktorej autori analyzujú dáta z komplexného štatistického prieskumu, všimnite si, či zobrali do úvahy štruktúru analyzovaných dát alebo či len realizovali výpočty pomocou bežného softvéru, ktorý nie je určený na analýzy dát z komplexných štatistických prieskumov. Ak nezobrali do úvahy štruktúru dát, mali by ste sa na výsledky, ktoré prezentujú, pozerať s veľkým podozrením."

Analýza regionálnej štruktúry príjmov domácností v roku 2014 podľa mediánu ich celkových hrubých príjmov poskytla zaujímavé výsledky. Obyčajne sa predpokladá, že bratislavský región v príjmoch domácností výrazne prevyšuje ostatné slovenské regióny. Analýza ukázala, že rozdiel medzi prvou Bratislavou a druhým Trenčínom nie je veľký, predstavuje len 122,9 eura. Mediány príjmov tretej Žiliny, štvrtej Trnavy a piateho Prešova sú pomerne blízke a tiež sa priveľmi neodlišujú od bratislavského regiónu. Rozdiel medzi prvou Bratislavou a piatym Prešovom je „len" 896,15 eura. Väčšie rozdiely sú medzi poslednými tromi regiónmi – Košicami, Nitrou a Banskou Bystricou. Medián celkových hrubých príjmov v regióne Banská Bystrica je prekvapivo nízky v porovnaní s Bratislavským krajom – je až o 2 744,96 eura nižší.

Na porovnanie sa vykonal aj odhad mediánu celkových hrubých príjmov domácností v slovenských regiónoch pomocou výberového mediánu, teda bez zohľadnenia štruktúry dát prostredníctvom výberových váh[26]. Rozdiel medzi hodnotami odhadov získanými s váhami a bez váh kolíše od (−1,82 %) do 7,50 %, čo nie sú zanedbateľné rozdiely. Pri odhadovaní bez váh vyšlo aj odlišné poradie regiónov: 1. Bratislava, 2. Trenčín, 3. Žilina, 4. Prešov, 5. Košice, 6. Trnava, 7. Nitra, 8. Banská Bystrica.

***Tento článok vznikol s prispením grantovej agentúry VEGA v rámci projektu číslo 1/0092/15 Moderné prístupy k navrhovaniu komplexných štatistických prieskumov.***

## LITERATÚRA

[1] BARNETT, V. − LEWIS, T.: Outliers in Statistical Data. Hoboken: Wiley and Sons, 1994. ISBN 0-471-93094-5.
[2] BETHLEHEM, J.: Applied Survey Methods. A Statistical Perspective. Hoboken: Wiley and Sons, 2009. ISBN 978-0-470-37308-8.
[3] COCHRAN, W. G.: Sampling Techniques. New York: Wiley and Sons, 1977. ISBN 0-471-16240-X.
[4] FULLER, W. A.: Sampling Statistics. USA: Wiley and Sons, 2009. ISBN 978-0-470-45460-2.
[5] HALLEY, R. M.: Measures of Central Tendency, Location, and Dispersion in Wage Survey Research. In: Compensation and Benefits, 2004, No. 36, p. 39-52.
[6] LEVY, P. S. − LEMESHOW, S.: Sampling of Populationas. Methods and Applications. Fourth Edition. Hoboken: Wiley and Sons, 2008. ISBN 978-0-470-04007-2.

[26] *Výpočet bol realizovaný pomocou funkcie Medián v Exceli.*

[7] LOHR, S. L.: Sampling: Design and Analysis. 2nd edition. Boston: Brooks/Cole, 2010. ISBN-10: 0-495-11084-1.
[8] PIEGORSCH, W. W.: Statistical Data Analysis. Foundations for Data Mining, Informatics, and Knowledge Discovery. Chichester: Wiley and Sons, 2015. ISBN 978-1-118-61965-0.
[9] SÄRNDAL, C. E. − SWENSSON, B. − WRETMAN, J.: Model Assisted Survey Sampling. New York: Springer, 2003. ISBN 0-387-40620-4.
[10] SÄRNDAL, C. E. − LUNDSTRÖM, S.: Estimation in Surveys with Nonresponse. Chichester: Wiley and Sons, 2005. ISBN 0-470-01133-5.
[11] TEREK, M.: Možnosti riešenia problému neodpovedania v štatistických prieskumoch. In: Ekonomické rozhľady, 2014, č. 2,, s. 150 – 165.
[12] TEREK, M.: Odľahlé dáta a charakteristiky polohy v analýzach miezd a príjmov. In: Revue sociálno-ekonomického rozvoja, 2016, No. 1.
[13] TEREK, M.: Interpretácia štatistiky a dát. Štvrté doplnené vydanie. Košice: Eqilibria, 2016. ISBN 978-80-8143-177-7.
[14] WILCOX, R. R.: Applying Contemporary Statistical Techniques. Burlington: Academic Press, 2003. ISBN 0-12-751541-0.

**RESUME**

The statistical survey consisting of more components such as random sampling, stratification, cluster sampling, sampling with unequal probabilities, ratio estimating etc. is commonly referred to as a comprehensive statistical survey. The paper presents the modules for the construction of comprehensive statistical surveys and the main data analysis options from the comprehensive statistical surveys. The above-mentioned procedures require the use of auxiliary information. These information can be used in the planning stage of the sample survey, as well as in the stage of point estimation construction. The article discusses in a detailed manner the use of the construction of sampling weights in empirical probability mass function, empirical cumulative distribution function and in the estimation of some characteristics of the basic file. The procedures are illustrated on the data analysis from the complex EU-SILC survey realized in the Slovak Republic in 2014. The regional structure of incomes was analyzed on the basis of the total gross household incomes median estimates in eight Slovak regions.

***PROFESIJNÝ ŽIVOTOPIS***

***Prof. Ing. Milan Terek, PhD.**, od roku 1977 do roku 1987 pôsobil na Katedre operačného výskumu a ekonometrie, od roku 1987 pôsobí na Katedre štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave. V súčasnosti vyučuje predmety štatistika (v slovenčine aj francúzštine) a aplikácie štatistických metód na prvom stupni štúdia, štatistické riadenie kvality, hĺbková analýza dát a analýza rozhodovania na druhom stupni štúdia a predmety výberové skúmanie a hĺbková analýza dát na treťom stupni štúdia. Na Vysokej škole manažmentu Trenčín/City University of Seattle vedie na doktorandskom štúdiu slovenský aj anglický kurz Kvantitatívne metódy vo výskume v oblasti podnikového manažmentu. Vo výskume sa venuje hlavne aplikáciám výberového skúmania, štatistického riadenia kvality a analýzy rozhodovania v ekonómii a manažmente.*

**KONTAKT**

milan.terek@euba.sk

**Gábor SZŰCS**
**Fakulta matematiky, fyziky a informatiky Univerzity Komenského v Bratislave**

# VIACROZMERNÁ ANALÝZA ROZPTYLU A JEJ APLIKÁCIE

# MULTIVARIATE ANALYSIS OF VARIANCE AND ITS APPLICATIONS

## ABSTRAKT
Analýza rozptylu bola navrhnutá predovšetkým na posúdenie podobností a odlišností medzi viacerými súbormi, pričom jej prostredníctvom možno testovať efekt jedného alebo viacerých faktorov. Ak sa na pozorovaných objektoch súčasne sledujú viaceré štatistické znaky, odporúča sa použiť viacrozmernú analýzu rozptylu, ktorá ponúka komplexnejšie možnosti testovania odlišností medzi súbormi alebo efektmi. Tento článok poskytuje krátky úvod do teórie jednorozmernej analýzy rozptylu a zovšeobecnenie modelu pre prípad viacerých rozmerov. Uvádza teoretický aparát na testovanie hypotéz pomocou viacrozmernej analýzy rozptylu a ukážky jej použitia na reálnych dátach v softvéri R.

## ABSTRACT
Analysis of variance was primarily designed to assess similarities and differences between more groups and by means of which the effect of one or more factors can be tested. If various statistical features are being pursued simultaneously on the observed objects, it is recommended to use the multivariate analysis of variance offering more complex possibilities of the difference testing between two or more groups or effects. This paper provides a brief introduction to the theory of one-dimensional analysis of variance and the generalization of the model to the multivariate case. It presents the theoretical apparatus for hypothesis testing using the multivariate analysis of variance and its illustrations on real data sets in R software.

## KĽÚČOVÉ SLOVÁ
analýza rozptylu, viacrozmerná analýza rozptylu, testovanie rovnosti vektorov stredných hodnôt, testovanie významnosti vplyvu vysvetľujúcich premenných

## KEY WORDS
analysis of variance, multivariate analysis of variance, testing the equality of mean vectors, significance test of the impact of explanatory variables

## 1. ÚVOD
Analýza rozptylu (známa aj pod skratkou ANOVA, Analysis of Variance) patrí medzi najpoužívanejšie parametrické metódy matematickej štatistiky. Je zovšeobecnením dvojvýberového Studentovho t-testu, pretože pomocou nej môžeme porovnať nielen dva, ale aj viaceré súbory a zistiť podobnosti a odlišnosti medzi nimi. Model analýzy rozptylu je založený na lineárnom regresnom modeli, v ktorom kvantitatívnu vysvetľovanú premennú popisujú neznáme regresné koeficienty, matica plánu a náhodný šum. ANOVA má široké spektrum praktického využitia, napr. vo farmácii, v biológii, genetike, poľnohospodárstve či priemysle.

Model ANOVA vychádza zo základného predpokladu, že rozdelenie pravdepodobnosti vysvetľovanej premennej je normálne (Gaussovo rozdelenie). Tento predpoklad je jednou z najväčších slabín analýzy rozptylu z hľadiska praktickej aplikácie metódy, pretože sledované štatistické znaky sa v každom prípade neriadia normálnym rozdelením. Pri silnom porušení predpokladu normality sa odporúča, aby sa štatistický test a model analýzy rozptylu nahradil jeho neparametrickým ekvivalentom, takzvaným Kruskalovým-Wallisovým testom, ktorý nevyžaduje, aby sa vysvetľované premenné správali podľa určitého rozdelenia pravdepodobnosti. Pomocou Kruskalovho-Wallisovho neparametrického testu môžeme testovať aj hypotézu o rovnosti stredných hodnôt v jednotlivých súboroch, teda v konečnom dôsledku môžeme dospieť k podobnému štatistickému vyhodnoteniu ako pri analýze rozptylu.

V praxi sa stretávame aj s takými situáciami, v ktorých je potrebné súčasne vysvetliť viaceré štatistické znaky. Vtedy už základná jednorozmerná analýza rozptylu nemusí stačiť na adekvátny opis premenných prostredníctvom vysvetľujúcich efektov. Tým, že pri analýze by sa naraz bral do úvahy iba jeden štatistický znak, mohli by sa stratiť cenné informácie zo simultánneho sledovania všetkých vysvetľovaných premenných. Viacrozmerná analýza rozptylu (v medzinárodnej literatúre nazývaná ako Multivariate Analysis of Variance, MANOVA) práve pre takéto situácie ponúka čiastočné riešenie a komplexnejšie testovanie odlišností medzi viacerými súbormi, na ktorých sa súčasne sledujú minimálne dva štatistické znaky.

Teoretické základy modelu MANOVA a možnosti jeho využitia boli vysvetlené vo viacerých výborných publikáciách, napríklad v [4], [5], [8], [9], [13]. S konkrétnymi aplikáciami MANOVA sa môžeme stretnúť v rôznych odvetviach vedeckého výskumu, napríklad v genetike [17], hydrológii [18], neurológii alebo v oblasti potravinárstva. Postupy viacrozmernej analýzy rozptylu by sa dali využiť napríklad aj pri práci s dátami z oblasti antropometrie [6], psychológie [15] a pri tvorbe ekonomických štúdií [10]. Náš ilustračný príklad, ktorý prezentujeme v 4. časti tohto článku, je z oblasti papierenského priemyslu [1], [7].

## 2. MODEL JEDNOROZMERNEJ ANALÝZY ROZPTYLU

Ako sme už v úvode spomínali, model jednorozmernej analýzy rozptylu je vlastne lineárnym regresným modelom, ktorý môžeme zapísať v tvare $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$, kde $\boldsymbol{Y}$ je náhodný vektor vysvetľovaných premenných dĺžky $n$, $\boldsymbol{X}$ je tzv. matica plánu, $\boldsymbol{B}$ je vektor efektov (prípadne vektor stredných hodnôt) a $\boldsymbol{\varepsilon}$ je náhodný vektor dĺžky $n$, s nezávislými, rovnako rozdelenými zložkami, ktoré majú normálne rozdelenie $N(0, \sigma^2)$, pričom $\sigma^2$ je rozptyl (alebo disperzia) zložiek náhodného vektora $\boldsymbol{Y}$, aj zložiek náhodného vektora $\boldsymbol{\varepsilon}$.

*Prípad dvoch súborov a jedného faktora*

V prípade, keď potrebujeme porovnať dva súbory, vektor $\boldsymbol{Y}$ je stĺpcovým náhodným vektorom $\boldsymbol{Y} = \left(Y_{11}, Y_{12}, \ldots, Y_{1n_1}, Y_{21}, Y_{22}, \ldots, Y_{2n_2}\right)$, kde zložky $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ sú hodnoty sledovaného štatistického znaku na prvom objekte, $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ sú hodnoty sledovaného štatistického znaku na druhom objekte a $n_1 + n_2 = n$. V súlade s vyššie uvedenou konštrukciou predpokladáme, že $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ je náhodný výber z rozdelenia $N(\mu_1, \sigma^2)$ s neznámou strednou hodnotou $\mu_1 \in \mathbb{R}$, a analogicky

$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ je náhodný výber z rozdelenia $N(\mu_2, \sigma^2)$ s neznámou strednou hodnotou $\mu_2 \in \mathbb{R}$. Tiež predpokladáme, že všetky zložky vektora $\boldsymbol{Y}$ sú navzájom nezávislé. Maticu plánu $\boldsymbol{X}$ v tomto prípade môžeme zapísať v tvare:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}.$$

Vektor efektov $\boldsymbol{B}$ je v tomto prípade dvojzložkovým stĺpcovým vektorom neznámych stredných hodnôt, môžeme teda písať v tvare: $\boldsymbol{B} = (\mu_1, \mu_2)$. Poznamenáme, že počet jednotiek v hornej časti matice $\boldsymbol{X}$ je $n_1$, kým v dolnej časti matice ich počet je $n_2$. Matica plánu $\boldsymbol{X}$ vlastne „zapína“ a „vypína“ efekty stredných hodnôt na 1., resp. 2. súbor.

V praxi obvykle hľadáme odpoveď na otázku, či sú medzi dvoma sledovanými súbormi štatisticky významné rozdiely alebo nie. Presnejšie, či sa stredná hodnota 1. súboru štatisticky významne líši od strednej hodnoty 2. súboru alebo nie. Nulová hypotéza príslušného štatistického testu sa zvyčajne formuluje v tvare $H_0: \mu_1 = \mu_2$ a testuje sa oproti alternatívnej hypotéze $H_1: \mu_1 \neq \mu_2$. Opísaný model je najjednoduchším prípadom analýzy rozptylu, ktorý sa nazýva (jednorozmernou) jednofaktorovou analýzou rozptylu pre prípad dvoch súborov a je totožný s modelom dvojvýberového Studentovho t-testu. Ďalšie detaily modelu a odvodenie testovej štatistiky sú dostupné napr. v [8], [9].

Ako jednoduchý príklad by sme mohli uviesť aplikáciu z poľnohospodárstva, v ktorej potrebujeme porovnať hektárové výnosy dvoch odrôd pšenice. V tomto prípade by premenná $Y_{11}$ predstavovala hektárový výnos prvej odrody pšenice na prvom poli (pri prvom meraní), premenná $Y_{12}$ by znamenala hektárový výnos prvej odrody pšenice na druhom poli (pri druhom meraní) atď., až veličina $Y_{1n_1}$ by špecifikovala hektárový výnos prvej odrody pšenice pri $n_1$-om meraní. Analogicky premenné $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ by postupne určovali hektárový výnos druhej odrody pšenice pri prvom, druhom až $n_2$-om meraní.

Predpokladajme, že hektárové výnosy v prípade oboch odrôd pšenice sa správajú podľa normálneho rozdelenia s rovnakou smerodajnou odchýlkou $\sigma > 0$ a merania sú navzájom nezávislé. Pomocou testu jednorozmernej jednofaktorovej analýzy rozptylu by sme mohli testovať, či sú štatisticky významné rozdiely medzi dvoma odrodami pšenice, čo sa týka ich očakávaného (priemerného) hektárového výnosu. Táto analýza je *jednorozmerná*, pretože pri každom meraní sledujeme jediný štatistický znak – hektárový výnos. Keby sme potrebovali súčasne vysvetliť aj ďalšie štatistické znaky, napríklad dĺžku stebla pšenice, hmotnosť slamy alebo dĺžku klasu, tak by sme používali viacrozmernú analýzu rozptylu (pozri v 3. časti tohto príspevku). Ďalej, náš pôvodný príklad je *jednofaktorový*, pretože priemerný hektárový výnos vysvetľujeme len pomocou jedného faktora – odrody pšenice. Ak by sme pri opise hektárového výnosu chceli brať do úvahy efekt viacerých vysvetľujúcich premenných, napríklad efekt typu pôdy alebo aj efekt hnojenia, tak by sme aplikovali dvoj- či trojfaktorovú analýzu rozptylu (pozri záverečnú časť tejto kapitoly).

*Prípad viacerých súborov a jedného faktora*

Predchádzajúci jednoduchý model bol navrhnutý na porovnanie dvoch súborov (v našom ilustračnom príklade dvoch odrôd pšenice), ľahko ho však môžeme zovšeobecniť pre prípad $\ell$ súborov (napríklad pre $\ell$ rôznych odrôd pšenice). Potom náhodný vektor $\boldsymbol{Y}$ možno napísať v tvare stĺpcového vektora dĺžky $n$ ako $(Y_{11}, Y_{12}, \dots, Y_{1n_1}, Y_{21}, Y_{22}, \dots, Y_{2n_2}, \dots, Y_{\ell 1}, Y_{\ell 2}, \dots, Y_{\ell n_\ell})$, kde $n_1, n_2, \dots, n_\ell$ sú počty meraní v jednotlivých súboroch a $n_1 + n_2 + \cdots + n_\ell = n$, pričom náhodná premenná $Y_{ik}$ vyjadruje hodnotu štatistického znaku $k$-teho objektu ($k$-teho pozorovania) v $i$-tom súbore pre $k = 1,2, \dots, n_i$ a $i = 1,2, \dots, \ell$. Opäť predpokladajme, že $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ je náhodný výber z rozdelenia $N(\mu_i, \sigma^2)$ s neznámou strednou hodnotou $\mu_i \in \mathbb{R}$ pre $i = 1,2, \dots, \ell$ a zložky vektora $\boldsymbol{Y}$ sú nezávislé. Vektor efektov $\boldsymbol{B}$ v lineárnom regresnom modeli $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$ je v tomto prípade $\ell$-zložkovým stĺpcovým vektorom neznámych stredných hodnôt, t. j. $\boldsymbol{B} = (\mu_1, \mu_2, \dots, \mu_\ell)$, kým maticu plánu $\boldsymbol{X}$ môžeme písať v tvare

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

kde počty jednotiek v jednotlivých blokoch matice $\boldsymbol{X}$ sú postupne $n_1, n_2, \dots, n_\ell$.

Základnou úlohou je overiť platnosť nulovej hypotézy o rovnosti stredných hodnôt sledovaného štatistického znaku v rámci všetkých $\ell$ súborov, t. j. $H_0: \mu_1 = \mu_2 = \cdots = \mu_\ell$. Táto nulová hypotéza sa testuje oproti alternatívnej hypotéze $H_1: \exists i, h \in \{1,2, \dots, \ell\}, i \neq h$, pre ktoré $\mu_i \neq \mu_h$. V nulovej hypotéze tvrdíme, že medzi súbormi nie sú štatisticky významné rozdiely v zmysle strednej hodnoty sledovaného štatistického znaku. Ak túto spoločnú strednú hodnotu z nulovej hypotézy označíme symbolom $\mu$, tak model jednofaktorovej analýzy rozptylu môžeme prepísať do tvaru

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik} = \mu_i + \varepsilon_{ik},$$

kde $\alpha_i$ je čistý efekt $i$-teho súboru na hodnotu sledovaného štatistického znaku ($i \in \{1,2, \dots, \ell\}$) a platí vzťah $\mu_i = \mu + \alpha_i$. Formálne sme teda pôvodný model iba reparametrizovali: strednú hodnotu $i$-teho súboru ($\mu_i$) sme nahradili súčtom spoločnej strednej hodnoty (spoločného efektu $\mu$) a čistého efektu $i$-teho súboru ($\alpha_i$). Potom nulovú hypotézu o rovnosti stredných hodnôt môžeme zapísať aj v ekvivalentnom tvare $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_\ell$.

Vzniká prirodzená otázka: prečo sa táto parametrická štatistická metóda nazýva analýzou rozptylu, ak primárne slúži na testovanie rovnosti stredných hodnôt? Odpoveď treba hľadať v odvodení a vo výslednom tvare testovej štatistiky spomínaného testu (pozri napríklad v [8], [9]), ktorá dáva do pomeru sumu štvorcov

odchýlok vysvetleného modelu (akýsi rozptyl alebo variabilitu medzi súbormi) a celkovú sumu štvorcov odchýlok (celkový rozptyl vysvetľovanej premennej). Testovú štatistiku v reči matematickej štatistiky zapisujeme v tvare

$$F = \frac{\frac{\mathrm{ESS}}{\ell - 1}}{\frac{\mathrm{TSS}}{n - \ell}} = \frac{\frac{\mathrm{TSS} - \mathrm{RSS}}{\ell - 1}}{\frac{\mathrm{TSS}}{n - \ell}},$$

kde $\mathrm{ESS} = \sum_{i=1}^{\ell} n_i (\bar{Y}_{i\cdot} - \bar{Y})^2$ je suma štvorcov odchýlok vysvetleného modelu (*explained sum of squares*, ESS), $\bar{Y}_{i\cdot}$ je aritmetický priemer vo vnútri $i$-teho súboru, $\bar{Y}$ je celkový aritmetický priemer vektora $\boldsymbol{Y}$, $\mathrm{RSS} = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_{i\cdot})^2$ je suma štvorcov rezíduí (*residual sum of squares*, RSS) a $\mathrm{TSS} = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y})^2$ je celková suma štvorcov odchýlok (*total sum of squares*, TSS). Medzi zavedenými veličinami platí vzťah $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$. Za predpokladu normality, nezávislosti premenných $Y_{ik}$ a platnosti nulovej hypotézy platí, že testová štatistika $F$ má Fisherovo F-rozdelenie so stupňami voľnosti $\ell - 1$ a $n - \ell$. Spomínanú $F$-štatistiku môžeme interpretovať nasledovne: ak medzi strednými hodnotami sledovaných súborov sú len malé rozdiely, tak rozptyl medzi súbormi (ESS) bude malý a celá hodnota $F$-štatistiky bude nízka. Ak hodnota štatistiky je menšia ako kritická hodnota F-rozdelenia, tak hypotézu o rovnosti stredných hodnôt nezamietame. V opačnom prípade, ak už len jediný súbor je iný ako ostatné, tak hodnota čitateľa $F$-štatistiky bude vysoká (pre vysokú variabilitu medzi súbormi), teda $F$-hodnota pravdepodobne presiahne kritickú hodnotu F-rozdelenia a hypotézu $H_0$ zamietneme.

*Prípad viacerých súborov a viacerých faktorov*

Najkomplexnejším prípadom jednorozmernej analýzy rozptylu je ten, v ktorom porovnávame viaceré súbory a vysvetľované premenné opisujeme pomocou viacerých efektov a ich interakcií. Idea a predpoklady ANOVA aj v tomto prípade zostávajú rovnaké, ako sme ich opísali predtým, avšak lineárny regresný model a testové štatistiky budú z matematického hľadiska už o niečo zložitejšie. Ako príklad dvojfaktorovej analýzy rozptylu by sme mohli uviesť istú modifikáciu ilustračnej štúdie z oblasti poľnohospodárstva (uvedenej na začiatku tejto kapitoly). Predpokladajme teda, že potrebujeme porovnať hektárové výnosy siedmich odrôd pšenice, ktoré sa pestovali v troch rôznych typoch pôdy. Nech $Y_{ijk}$ je hektárový výnos $i$-tej odrody pšenice siatej do $j$-teho typu pôdy v prípade $k$-teho pozorovania (pri danej kombinácii odrody pšenice a typu pôdy), pričom v našom príklade $i = 7$ odrôd pšenice, $j = 3$ a $k = 1,2, \ldots, n_{ij}$, kde $n_{ij}$ je počet meraní v prípade $i$-tej odrody a $j$-teho typu pôdy. Podrobný všeobecný popis dvojfaktorovej analýzy rozptylu a odvodenie testových štatistík na testovanie hypotézy o rovnosti stredných hodnôt je možné nájsť napríklad v knihe [9].

Možnosti používania modelu ANOVA a testovanie hypotéz v tomto modeli ponúka viacero matematických či štatistických softvérov, ako napríklad SPSS, Microsoft Excel, SAS alebo R. V prípade posledného menovaného softvéru možno aplikovať napríklad funkciu `aov`, ktorá je súčasťou základného štatistického balíka programu [12].

## 3. MODEL VIACROZMERNEJ ANALÝZY ROZPTYLU

Model viacrozmernej analýzy rozptylu vznikol zovšeobecnením jednorozmerného modelu pre situácie, keď sa na skúmaných objektoch súčasne sleduje $p$ štatistických znakov, kde $p$ je prirodzené číslo väčšie ako $1$. Podkladový lineárny regresný model viacrozmernej analýzy rozptylu (tzv. všeobecný model MANOVA) môžeme zapísať v tvare

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E},$$

kde $\boldsymbol{Y}$ je náhodná matica typu $n \times p$, pričom $n$ je počet všetkých sledovaných objektov (počet riadkov matice $\boldsymbol{Y}$), $p$ je počet sledovaných štatistických znakov (počet stĺpcov matice $\boldsymbol{Y}$), $\boldsymbol{X}$ je známa nenáhodná matica plánu rozmerov $n \times \ell$, pričom $\ell$ je počet regresných parametrov (efektov alebo ošetrení) používaných pre každý sledovaný štatistický znak, $\boldsymbol{B}$ je nenáhodná neznáma matica efektov typu $\ell \times p$ a $\boldsymbol{E}$ je náhodná matica rozmerov $n \times p$, ktorej každý riadok je náhodným výberom z $p$-rozmerného normálneho rozdelenia $N_p(\boldsymbol{0},\boldsymbol{\Sigma})$, kde $\boldsymbol{\Sigma}$ je kovariančnou maticou $p$-tice sledovaných štatistických znakov.

*Prípad jedného faktora*

Ak sledovanú závislú premennú vysvetľujeme pomocou jedného faktora, tak vyššie uvedený všeobecný model viacrozmernej analýzy rozptylu môžeme napísať v tvare

$$\boldsymbol{Y}_{ki} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ki},$$

kde $\boldsymbol{Y}_{ki}$ je $p$-rozmerný vektor hodnôt sledovaných štatistických znakov v prípade $k$-teho pozorovania v $i$-tom súbore (alebo v prípade $i$-teho ošetrenia) pre $k = 1,2,\dots,n_i$ a $i = 1,2,\dots,\ell$, pričom $n_i$ je počet pozorovaní (meraní) vykonaných v $i$-tom súbore a $n = \sum_{i=1}^{\ell} n_i$ je celkový počet pozorovaní. Ďalej, symbolom $\boldsymbol{\mu}$ označujeme vektor stredných hodnôt sledovaných štatistických znakov nezávislý od súborov ($\boldsymbol{\mu}$ je tzv. spoločný alebo celkový efekt). Podobne ako pri jednorozmernom modeli ANOVA, aj v tomto prípade $\boldsymbol{\alpha}_i$ označuje čistý efekt $i$-teho súboru na hodnoty sledovaných štatistických znakov. Poznamenajme, že $\boldsymbol{\alpha}_i$ je tiež $p$-rozmerným vektorom, pre ktorý platí vzťah $\boldsymbol{\mu} + \boldsymbol{\alpha}_i = \boldsymbol{\mu}_i$, kde $\boldsymbol{\mu}_i$ je vektorom stredných hodnôt $i$-teho súboru. Posledným členom modelu jednofaktorovej viacrozmernej analýzy rozptylu je $p$-rozmerný náhodný šum $\boldsymbol{\varepsilon}_{ki}$, o ktorom predpokladáme, že pochádza z $p$-rozmerného normálneho rozdelenia $N_p(\boldsymbol{0},\boldsymbol{\Sigma})$.

Hlavnou úlohou MANOVA je testovanie vplyvu súborov na hodnoty sledovaných štatistických znakov, t. j. či jednotlivé súbory majú vplyv na výsledné hodnoty alebo nie. Nulovú hypotézu o rovnosti vektorov stredných hodnôt sledovaného štatistického znaku v rámci všetkých $\ell$ súborov môžeme písať v tvare $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_\ell$. Ak platí nulová hypotéza, tak to znamená, že medzi súbormi nie sú štatisticky významné rozdiely v zmysle strednej hodnoty a vysvetľované premenné možno popísať len pomocou spoločného efektu $\boldsymbol{\mu}$. Uvedená nulová hypotéza sa obvykle testuje oproti alternatívnej hypotéze $H_1: \exists i,h \in \{1,2,\dots,\ell\}, i \neq h$, pre ktoré $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_h$. Tvrdí, že medzi súbormi je aspoň jeden, ktorý sa štatisticky významne líši od ostatných súborov v zmysle vektora stredných hodnôt. Na definovanie testových štatistík pre uvedenú

nulovú hypotézu je potrebné zaviesť niekoľko ďalších označení a definovať rozdelenia pravdepodobnosti pre viacrozmerné náhodné výbery.

**Definícia 1.** Nech $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n$ je náhodný výber z $p$-rozmerného normálneho rozdelenia $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\Sigma}$ je pozitívne definitná matica typu $p \times p$. Nech $\boldsymbol{Z}^\top = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n)$ je náhodná matica typu $p \times n$ a nech $\boldsymbol{M} = \boldsymbol{Z}^\top \boldsymbol{Z}$ je náhodná matica typu $p \times p$. Rozdelenie pravdepodobnosti matice $\boldsymbol{M}$ nazývame Wishartovým rozdelením s parametrom $\boldsymbol{\Sigma}$ a stupňami voľnosti $n$; označujeme to zápisom $\boldsymbol{M} \sim W_p(\boldsymbol{\Sigma}, n)$.

**Definícia 2.** Nech $\boldsymbol{M}$ a $\boldsymbol{N}$ sú nezávislé náhodné matice typu $p \times p$ a nech $\boldsymbol{M} \sim W_p(\boldsymbol{I}_p, m)$, $\boldsymbol{N} \sim W_p(\boldsymbol{I}_p, n)$, pričom $m \geq p$ a symbolom $\boldsymbol{I}_p$ označujeme identickú maticu typu $p \times p$ (maticu, ktorá má na svojej hlavnej diagonále samé jednotky a mimo hlavnej diagonály samé nuly). Definujme náhodnú premennú $\Lambda$ predpisom

$$\Lambda = \frac{\det(\boldsymbol{M})}{\det(\boldsymbol{M} + \boldsymbol{N})} = \frac{1}{\det(\boldsymbol{I}_p + \boldsymbol{M}^{-1}\boldsymbol{N})},$$

kde $\det(\boldsymbol{M})$ označuje determinant matice $\boldsymbol{M}$ a $\boldsymbol{M}^{-1}$ je inverznou maticou matice $\boldsymbol{M}$. Rozdelenie pravdepodobnosti náhodnej premennej $\Lambda$ nazývame Wilksovým Lambda-rozdelením s parametrami $p, m, n$; označujeme to zápisom $\Lambda \sim \Lambda(p, m, n)$.

Definujme teraz viacrozmerné ekvivalenty veličín $\mathrm{ESS}$, $\mathrm{RSS}$ a $\mathrm{TSS}$ zavedených v 2. časti tohto príspevku pri modeli jednorozmernej jednofaktorovej analýzy rozptylu.

**Definícia 3.** Nech pre náhodnú maticu $\boldsymbol{Y}$ platia všetky predpoklady modelu MANOVA a nech

$$\boldsymbol{H} = \sum_{i=1}^{\ell} n_i (\overline{\boldsymbol{Y}}_{\cdot i} - \overline{\boldsymbol{Y}})(\overline{\boldsymbol{Y}}_{\cdot i} - \overline{\boldsymbol{Y}})^\top$$

je matica súčtov štvorcov a súčinov odchýlok medzi súbormi, kde $\overline{\boldsymbol{Y}}_{\cdot i}$ je $p$-rozmerný vektor aritmetických priemerov hodnôt sledovaných štatistických znakov vo vnútri $i$-teho súboru a $\overline{\boldsymbol{Y}}$ je $p$-rozmerný vektor celkových aritmetických priemerov počítaných po stĺpcoch matice $\boldsymbol{Y}$. Ďalej nech

$$\boldsymbol{E} = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} (\boldsymbol{Y}_{ki} - \overline{\boldsymbol{Y}}_{\cdot i})(\boldsymbol{Y}_{ki} - \overline{\boldsymbol{Y}}_{\cdot i})^\top$$

je matica súčtov štvorcov a súčinov odchýlok vo vnútri súborov a

$$\boldsymbol{T} = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} (\boldsymbol{Y}_{ki} - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_{ki} - \overline{\boldsymbol{Y}})^\top$$

je matica celkových súčtov štvorcov a súčinov odchýlok.

Poznamenáme, že medzi vyššie definovanými maticami platí vzťah $\boldsymbol{T} = \boldsymbol{E} + \boldsymbol{H}$, ktorý je viacrozmernou analógiou vzťahu $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$. Tiež dodávame, že za

predpokladu, že riadky matice $\boldsymbol{Y}$ pochádzajú z náhodného výberu z rozdelenia $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, náhodné matice $\boldsymbol{E}$ a $\boldsymbol{H}$ majú Wishartovo rozdelenie.

**Veta 1.** Nech platia všetky vyššie zavedené označenia a predpoklady. Potom testová štatistika pomerom vierohodností pre hypotézu o rovnosti vektorov stredných hodnôt sledovaného štatistického znaku v rámci $\ell$ súborov v modeli MANOVA ($H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_\ell$) má tvar

$$\Lambda = \frac{\det(\boldsymbol{E})}{\det(\boldsymbol{T})} = \frac{\det(\boldsymbol{E})}{\det(\boldsymbol{E}+\boldsymbol{H})} = \frac{1}{\det(\boldsymbol{I}_p + \boldsymbol{E}^{-1}\boldsymbol{H})},$$

kde testová štatistika $\Lambda$ má Wilksovo Lambda-rozdelenie s parametrami $p, \nu_{\boldsymbol{E}}, \nu_{\boldsymbol{H}}$, kde $\nu_{\boldsymbol{E}} = n - \ell$ je hodnosť náhodnej matice $\boldsymbol{E}$ a $\nu_{\boldsymbol{H}} = \ell - 1$ je hodnosť náhodnej matice $\boldsymbol{H}$.

Dôkaz Vety 1 možno nájsť napríklad v [9].

Test zavedený vo Vete 1 nazývame Wilksovým testom pomerom vierohodností a interpretujeme ho podobne ako F-test definovaný pre jednorozmernú analýzu rozptylu. Ak determinant matice $\boldsymbol{E}^{-1}\boldsymbol{H}$ je malý, tak to znamená, že medzi súbormi nie sú veľké odchýlky. Potom aj determinant matice $\boldsymbol{I}_p + \boldsymbol{E}^{-1}\boldsymbol{H}$ bude malý, teda to znamená, že nulovú hypotézu nezamietame pri vysokých hodnotách $\Lambda$-testovej štatistiky. Hodnoty testovej štatistiky sa v praxi obvykle porovnávajú s kritickými hodnotami aproximatívneho Fisherovho F-rozdelenia alebo s kritickými hodnotami aproximatívneho $\chi^2$-rozdelenia (ďalšie detaily sú uvedené v publikáciách [8], [9], [14]).

Pri testovaní hypotézy o rovnosti vektorov stredných hodnôt v modeli MANOVA sa okrem Wilksovho testu pomerom vierohodností používajú aj ďalšie testy, napríklad Pillaiov test, Lawleyov-Hotellingov test alebo Royov test.

Nech platia všetky vyššie zavedené označenia a predpoklady. Definujme testovú štatistiku $V$ ako

$$V = \mathrm{st}((\boldsymbol{E}+\boldsymbol{H})^{-1}\boldsymbol{H}),$$

kde $\mathrm{st}$ označuje stopu matice, teda súčet diagonálnych prvkov matice. Testová štatistika $V$ sa nazýva Pillaiovou stopou [11], [14]. Ďalej, definujme testovú štatistiku $U$ predpisom

$$U = \mathrm{st}(\boldsymbol{E}^{-1}\boldsymbol{H}),$$

ktorá sa nazýva Lawleyovou-Hotellingovou stopou [8], [14]. Označme symbolom $\lambda_1$ najväčšie vlastné číslo matice $\boldsymbol{E}^{-1}\boldsymbol{H}$, ktoré sa nazýva Royovo najväčšie vlastné číslo. Potom testová štatistika definovaná ako

$$\theta = \frac{\lambda_1}{1+\lambda_1}$$

sa nazýva Royovou testovou štatistikou pomocou najväčšieho vlastného čísla [8], [14]. V teoretických prácach autorov týchto testových štatistík sa ukázalo, že všetky

tri sú aplikovateľné pri testovaní hypotézy $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_\ell$ v modeli MANOVA. Hoci všetky tri testové štatistiky (náhodné premenné $V$, $U$ a $\theta$) majú odvodené svoje rozdelenie pravdepodobnosti, v praxi sa obvykle aplikujú aproximácie založené na Fisherovom F-rozdelení. Ďalšie podrobnosti o aproximáciách a ich použiteľnosti obsahuje napríklad práca [14].

## 4. PRAKTICKÁ APLIKÁCIA VIACROZMERNEJ ANALÝZY ROZPTYLU

Praktickú ukážku testovania hypotéz v modeli MANOVA sme zvolili z oblasti papierenského priemyslu. Pôvodné dáta boli publikované v článku [1] a ďalej sa podrobnejšie analyzovali v knihe [7]. V oboch prípadoch však išlo o analýzy zamerané na použiteľnosť tzv. zovšeobecnených lineárnych modelov (*generalized linear models*, GLM), takže naše skúmania o aplikovateľnosti MANOVA sa od nich odlišujú.

Experiment sa uskutočnil v roku 1985 v papierni *Norske Skog* v nórskom meste Skogn a bol zameraný na meranie kvality papiera pri rôznych nastaveniach výrobného procesu. Trval 30 hodín, pričom výskumníci v hodinových intervaloch merali a zaznamenávali 13 rôznych ukazovateľov kvality papiera, ktoré teraz pracovne označíme ako $Y_1, Y_2, \ldots, Y_{12}, Y_{13}$. Počas experimentu sa podarilo zaznamenať 29 úspešných meraní, jedno meranie sa nepodarilo. Cieľom pôvodného experimentu bolo zistiť, ako vplývajú 3 ovplyvniteľné nastavenia procesu výroby papiera (označíme ich ako $X_1, X_2, X_3$) na kvalitu papiera. Ukázalo sa, že výstupné premenné $Y_1, Y_2, \ldots, Y_{12}, Y_{13}$ majú zložitú multivariačnú štruktúru, a preto už autor pôvodného článku [1] na vyhodnotenie experimentu používal metodiky viacrozmerných štatistických analýz.

V našich analýzach sa zameriame na testovanie hypotézy o rovnosti vektorov stredných hodnôt v modeli MANOVA, t. j. otestujeme, či rôzne nastavenia troch vstupov výroby majú alebo nemajú vplyv na 13 súčasne sledovaných štatistických znakov (na ukazovatele kvality papiera). Dodatočne vykonáme aj testovanie významnosti vplyvu vysvetľujúcich premenných (nastavení výrobného procesu) na sledované vysvetľované premenné (ukazovatele kvality papiera). Všetky naše analýzy sa uskutočnili v prostredí softvéru R [12], značnú časť postupu preto uvádzame heslovito a v programátorskom formáte.

```
# načítanie dát (zdroj dát: [1])
np <- read.table(
"http://www.iam.fmph.uniba.sk/ospm/Szucs/data/norwaypaper.csv",
header=TRUE); attach(np);

# názvy používaných premenných:
# Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9 Y10 Y11 Y12 Y13 X1 X2 X3
# vypísanie matice plánu:
cbind(rep(1,length(X1)), X1, X2, X3)
```

Keďže model MANOVA je založený na predpoklade normality vysvetľovaných premenných, v prvom kroku vykonáme zovšeobecnený Shapirov-Wilkov test na viacrozmerné normálne rozdelenie publikovaný v článku [16]. Pri tomto teste v softvéri R používame balík `mvShapiroTest` [3].

```
install.packages("mvShapiroTest"); library(mvShapiroTest);
mvShapiro.Test(cbind(Y1,Y2,Y3,Y4,Y5,Y6,Y7,Y8,Y9,Y10,Y11,Y12,Y13))
# Výstup:
# Generalized Shapiro-Wilk test for Multivariate Normality by
# Villasenor-Alva and Gonzalez-Estrada
# MVW = 0.96663, p-value = 0.698
```

Skratkou `MVW` sa v predošlom výstupe označuje hodnota testovej štatistiky zovšeobecneného Shapirovho-Wilkovho testu na viacrozmerné normálne rozdelenie (matematické vyjadrenie testovej štatistiky možno nájsť napríklad v článku [16]). Nulovou hypotézou testu je, že dáta pochádzajú z normálneho rozdelenia. P-hodnota testu (`p-value`) vyšla pomerne vysoká (0,698), takže nulovú hypotézu by sme nezamietli na bežne používaných hladinách významnosti, teda hodnoty našich ukazovateľov kvality papiera by sa mohli riadiť podľa 13-rozmerného normálneho rozdelenia. Výsledok tohto testu však musíme brať s určitou rezervou, pretože v dátovom súbore sa nachádza len 29 pozorovaní, čo je pomerne malý počet na adekvátne posúdenie normality.

Keďže viacrozmernú normalitu vysvetľovaných premenných sme nezamietli, môžeme pristúpiť k hlavným testom v modeli MANOVA. V nulovej hypotéze testu o rovnosti vektorov stredných hodnôt predpokladáme, že rôzne nastavenia troch vstupov výrobného procesu nemajú štatisticky významný vplyv na výslednú kvalitu papiera, t. j. $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ alebo $H_0: \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3$. Pri testovaní používame všetky štyri testové štatistiky uvedené na konci predchádzajúcej časti a výpočty vykonáme pomocou balíka `car` [2].

```
# vytvorenie lineárneho regresného modelu,
# ktorý obsahuje aj celkový efekt
model1 <- lm(cbind(Y1,Y2,Y3,Y4,Y5,Y6,Y7,Y8,Y9,Y10,Y11,Y12,Y13) ~
             1+X1+X2+X3, data=np)
#
install.packages("car"); library(car);
# CAR = Companion to Applied Regression [2]
#
# používame príkaz lht = Linear Hypothesis Test
lht(model1, c("X1", "X2", "X3"))
#
# Výstup:
# Multivariate Tests:
#              Df test stat approx F num Df  den Df     Pr(>F)
# Pillai        3  2.269414  3.584183     39 45.00000 2.6920e-05 ***
# Wilks         3  0.006956  4.380605     39 39.24367 5.2064e-06 ***
# Hot.-Lawley   3 17.176653  5.138315     39 35.00000 1.7985e-06 ***
# Roy           3 11.897814 13.728247     13 15.00000 4.8994e-06 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Stĺpce výstupu sú nasledovné:
# názov testovej štatistiky (Pillai, Wilks, Hot.-Lawley, Roy),
# Df = počet stupňov voľnosti (počet testovaných premenných),
# test stat = hodnota testovej štatistiky,
# approx F = hodnota aproximatívnej F-štatistiky,
# num Df, den Df = počty stupňov voľnosti (parametre) F-štatistiky,
```

```
# Pr(>F) = p-hodnota testu počítaná z aproximatívnej F-štatistiky.
#
# V poslednom riadku výstupu sú uvedené vizuálne kódy
# signifikantnosti testov medzi rôznymi hladinami významnosti.
```

Z posledného stĺpca výstupu, označeného ako `Pr(>F)`, je zrejmé, že p-hodnota všetkých štyroch testov vyšla veľmi nízka. To znamená, že pri testoch nastalo významné porušenie nulovej hypotézy, ktorú preto s veľkou istotou zamietame. To znamená, že aspoň jedno z troch nastavení vstupov výrobného procesu má štatisticky významný vplyv na výslednú kvalitu papiera, teda efekt aspoň jedného ošetrenia sa líši od ostatných dvoch efektov.

V predchádzajúcom teste sa dáta testovali ako jeden celok a zistilo sa, že niektoré nastavenia vstupných premenných by mohli mať vplyv na kvalitu papiera. Môžeme preto položiť prirodzené otázky: Ktorá vstupná premenná z $X_1, X_2, X_3$ má najväčší vplyv na 13 pozorovaných ukazovateľov? Všetky vysvetľujúce premenné majú štatisticky významný vplyv alebo len niektoré z nich? Odpoveď na tieto otázky nám dá tzv. test významnosti vplyvu vysvetľujúcich premenných (pozri [8], [9]), ktorý tiež vykonáme v softvéri R pomocou balíka `car` [2] a príkazu `Anova`. Nulovou hypotézou testu je, že testovaná vstupná premenná nemá štatisticky významný vplyv na hodnoty výstupnej premennej (t. j. má nulový efekt na hodnoty výstupnej premennej).

```
Anova(model1, test.statistic="Wilks")
Anova(model1, test.statistic="Pillai")
Anova(model1, test.statistic="Hotelling-Lawley")
Anova(model1, test.statistic="Roy")
#
# výstupy týchto príkazov sú podobné, preto uvádzame len prvý z nich
# Type II MANOVA Tests: Wilks test statistic
#    Df test stat approx F num Df den Df   Pr(>F)
# X1  1  0.080818  11.3734     13     13 4.73e-05 ***
# X2  1  0.247041   3.0479     13     13  0.02717 *
# X3  1  0.283753   2.5242     13     13  0.05369 .
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Stĺpce výstupu sú skoro rovnaké, ako v prípade výstupu
# pri príkaze lht (viď vyššie).
```

Výsledok testu založený na Wilksovej testovej štatistike preukázal, že najvýznamnejšou vstupnou premennou je premenná $X_1$, pretože p-hodnota testu významnosti (uvedená v poslednom stĺpci výstupu) vyšla veľmi nízka ($0{,}0000473$), teda nulovú hypotézu o nulovom efekte premennej $X_1$ by sme zamietli na všetkých bežne používaných hladinách významnosti. Ďalej na hladine významnosti $0{,}05$ by aj premenná $X_2$ mohla mať štatisticky významný efekt na kvalitu papiera, pretože p-hodnota testu vyšla $0{,}02717 < 0{,}05$, teda nulovú hypotézu o nevýznamnom vplyve premennej $X_2$ by sme zamietli. Z výstupu tiež vidíme, že najslabší efekt na kvalitu papiera má premenná $X_3$, pri ktorej by sme nulovú hypotézu na hladine významnosti $0{,}05$ nezamietli (p-hodnota $= 0{,}05369 > 0{,}05$). Pre úplnosť dodávame, že tieto testy významnosti vplyvu vysvetľujúcich premenných by sme mohli vykonať aj pomocou

funkcie `manova`, ktorá je súčasťou základného štatistického balíka softvéru R. Použitie a výstup funkcie uvádzame ďalej.

```
m <- manova(model1)
summary(m)
#           Df  Pillai approx F num Df den Df   Pr(>F)
# X1         1 0.91918  11.3734     13     13 4.73e-05 ***
# X2         1 0.76630   3.2790     13     13  0.02048 *
# X3         1 0.71625   2.5242     13     13  0.05369 .
# Residuals 25
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Stĺpce výstupu sú skoro rovnaké, ako v prípade výstupu
# pri príkaze lht (viď vyššie).
```

Výsledok tohto testu pri použití Pillaiovej testovej štatistiky vyšiel veľmi podobne ako v predchádzajúcom výstupe. Môžeme teda zhrnúť, že najvýznamnejšou premennou je prvá vstupná premenná výrobného procesu, kým zvyšné dve vstupné nastavenia majú pravdepodobne len menší vplyv na výslednú kvalitu vyrábaného papiera.

## 5. ZÁVER

Viacrozmerná analýza rozptylu má široké spektrum využitia od lekárskej vedy cez poľnohospodárstvo až po priemyselné aplikácie. Jej najväčšou výhodou je, že komplexne, v rámci jediného modelu, zohľadňuje všetky vstupné a výstupné premenné, a to aj v prípade, keď sú viacrozmerné. V tomto článku sme zhrnuli najdôležitejšie teoretické poznatky o modeli viacrozmernej analýzy rozptylu a testovaní hypotéz v tomto modeli. Poukázali sme aj na možné slabé miesta testovania pomocou modelu MANOVA, najmä čo sa týka prípadného porušenia predpokladu normality kvantitatívnych vysvetľovaných premenných. V článku sme uviedli štyri štatistické testy vhodné na testovanie rovnosti vektorov stredných hodnôt a ich použitie sme ilustrovali reálnym príkladom z oblasti výroby papiera. Do riešenia ilustračnej úlohy sme doplnili aj test významnosti vplyvu vysvetľujúcich premenných v modeli MANOVA. Jedným z hlavných výstupov tohto príspevku je ukážka postupu testovania hypotéz v modeli MANOVA v prostredí softvéru R. Tento postup je do veľkej miery všeobecný, a preto sa dá použiť aj pri iných výskumoch a dátových súboroch.

***Tento článok vznikol s podporou grantov VEGA 2/0047/15, VEGA 1/0251/16 a APVV-0465-12.***

## LITERATÚRA

[1] ALDRIN, M.: Moderate projection pursuit regression for multivariate response data. In: Computational Statistics & Data Analysis, Elsevier, 1996, No. 5, p. 501-531.

[2] FOX, J. – WEISBERG, S.: An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage, 2011. Dostupné na: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion [prístup k 24. 4. 2017].

[3] GONZALEZ-ESTRADA, E. – VILLASENOR-ALVA, J. A.: mvShapiroTest: Generalized Shapiro-Wilk test for multivariate normality. R package version 1.0,

2013. Dostupné na: https://CRAN.R-project.org/package=mvShapiroTest [prístup k 23. 11. 2016].
[4] HÄRDLE, W. K. – HLÁVKA, Z.: Multivariate Statistics: Exercises and Solutions. New York: Springer, 2007. 368 p. ISBN 978-0-387-73508-5.
[5] HÄRDLE, W. K. – SIMAR, L.: Applied Multivariate Statistical Analysis. Heidelberg: Springer, 2012. 516 p. ISBN 978-3-642-17229-8.
[6] HEINZ, G. – PETERSON, L. J. – JOHNSON, R. W. – KERK, C. J.: Exploring Relationships in Body Dimensions. In: Journal of Statistics Education, 2003, No. 2.
[7] IZENMAN, A. J.: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. 1st Edition. New York: Springer, 2008. ISBN 978-0-387-78189-1.
[8] JOHNSON, R. A. – WICHERN, D. W.: Applied Multivariate Statistical Analysis. 6th Edition. Harlow: Pearson Education Limited, 2014. 770 p. ISBN 13: 978-1-292-02494-3.
[9] LAMOŠ, F. – POTOCKÝ, R.: Pravdepodobnosť a matematická štatistika: Štatistické analýzy. Bratislava: Univerzita Komenského, 1998. 343 s. ISBN 80-223-1262-2.
[10] MURA, L. – BULECA, J. – HAJDUOVA, Z. – ANDREJKOVIC, M.: Quantitative financial analysis of small and medium food enterprises in a developing country. In: Transformations in business & economics. 2015, No 1, p. 161-173.
[11] PILLAI, K. C. S.: On the Distribution of the Largest Root of a Matrix in Multivariate Analysis. In: Ann. Math. Statist., 1967, No. 2, p. 616-617.
[12] R CORE TEAM: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016. Dostupné na: https://www.R-project.org/ [prístup k 10. 4. 2017].
[13] RENCHER, A. C.: Multivariate Statistical Inference and Applications. New York: Wiley, 1998. 592 p. ISBN: 978-0-471-57151-3.
[14] SVETLÍKOVÁ, B.: Rôzne spôsoby testovania v MANOVA. Bratislava: Fakulta matematiky, fyziky a informatiky Univerzity Komenského v Bratislave, 2015.
[15] TIMM, N. H.: Multivariate Analysis with Applications in Education and Psychology. Wadsworth, 1975.
[16] VILLASENOR-ALVA, J. A. – GONZALEZ-ESTRADA, E.: A generalization of Shapiro-Wilk's test for multivariate normality. In: Communications in Statistics: Theory and Methods, 2009, No. 11, p. 1870-1883.
[17] YANG, Q. – WANG, Y.: Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. In: Journal of Probability and Statistics, 2012.
[18] WANG, X. – MELESSE, A. M. – YANG, W.: Development of a Multivariate Regression Model for Soil Nitrate Nitrogen Content Prediction. In: Journal of Spatial Hydrology, 2006, No. 2.

**RESUME**

Multivariate analysis of variance (MANOVA) has a wide range of applications, from medical science through agriculture to industrial applications. The main advantage of MANOVA is the simultaneous consideration of all input variables, even if the variables follow a complex multivariate distribution. This paper presents the theoretical background of the univariate and multivariate analysis of variance and hypothesis testing in these models. We have also pointed out possible limitations of MANOVA, especially the violation of normality assumption of explanatory variables of the linear regression model. To test the equality of mean vectors the following four

test statistics were used: Wilk's lambda, Pillai's trace, Hotelling-Lawley's trace and Roy's largest eigenvalue. To illustrate the usage of these test statistics, a research study from the paper industry was mentioned. The significance test of the impact of explanatory variables in the MANOVA model was carried out as well. The main contribution of this paper is to demonstrate the MANOVA model construction and testing of hypotheses in R software. Our approach is general enough to be applied to other studies and datasets as well.

***PROFESIJNÝ ŽIVOTOPIS***

***Mgr. Gábor Szűcs, PhD.**, vyštudoval pravdepodobnosť a matematickú štatistiku na Fakulte matematiky, fyziky a informatiky Univerzity Komenského v Bratislave a podnikové hospodárstvo a manažment na Univerzite J. Selyeho v Komárne. Doktorandské štúdium absolvoval na Fakulte matematiky, fyziky a informatiky Univerzity Komenského v odbore aplikovaná matematika a od roku 2015 pôsobí na tejto fakulte ako odborný asistent. Vo svojej výskumnej činnosti sa venuje oceňovaniu poistných dôchodkov, skúmaniu rozdelení pravdepodobnosti v neživotnom poistení a ďalším oblastiam aplikovanej štatistiky.*

**KONTAKT**

Gabor.Szucs@fmph.uniba.sk

**Eva KOTLEBOVÁ**
**Katedra štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave**

# VYUŽITIE BAYESOVSKÝCH METÓD PRI ANALÝZE DOSTUPNOSTI ZDRAVOTNEJ STAROSTLIVOSTI NA SLOVENSKU

# THE USE OF BAYESIAN METHODS FOR ANALYSING THE ACCESSIBILITY TO HEALTH CARE IN SLOVAKIA

## ABSTRAKT
Problémy s dostupnosťou zdravotnej starostlivosti možno považovať za indikátor chudoby, pretože jednou z jej hlavných príčin je nedostatok finančných zdrojov. Príspevok sa zaoberá odhadom podielu obyvateľov Slovenskej republiky, ktorí si nemôžu dovoliť zdravotnú starostlivosť najmä z finančných dôvodov, ale aj z iných príčin, ktoré však tiež úzko súvisia s finančnou situáciou obyvateľov. Podkladom na analýzu sú údaje z výberového zisťovania EHIS 2014, ktoré mapujú situáciu z pohľadu poberateľov zdravotnej starostlivosti. Okrem odhadu podielov obyvateľov Slovenskej republiky, pre ktorých je nedostupná niektorá z foriem zdravotnej starostlivosti, sa v článku analyzujú aj niektoré faktory ovplyvňujúce mieru nedostupnosti zdravotnej starostlivosti. Na spresnenie odhadov sa využívajú aj údaje z výberového zisťovania EU SILC, ktoré možno považovať za zdroj apriórnej informácie, takže nám umožnili aplikovať bayesovské metódy.

## ABSTRACT
Problems with access to the health care can be regarded as an indicator of poverty, because one of its leading causes is the lack of financial resources. The paper deals with estimating the proportion of the population of the Slovak Republic that cannot afford medical treatment mainly due to financial but also, for other reasons which are closely related to the population´s financial situation. The analysis is based on data from the EHIS 2014 sample survey, monitoring the situation from the recipients´ perspective. In addition to estimating the proportion of the Slovak population without access to one of the forms of health care, some factors determining the degree of inaccessibility to health care are analysed in the paper. In order to make the estimates more accurate, the data from the EU SILC sample survey was also used which can be regarded as a source of a priori information thus allowing us to apply the Bayesian methods.

## KĽÚČOVÉ SLOVÁ
zdravotná starostlivosť, bayesovský bodový odhad podielu, konjugovaný systém binomické/beta, EHIS – európske zisťovanie o zdraví, EU SILC – štatistika EÚ o príjmoch a životných podmienkach

## KEY WORDS
Health care, Bayesian point estimation of proportion, conjugate family beta-binomial, EHIS – European Health Interview Survey, EU SILC – European Union Statistics on Income and Living Conditions

## 1. ÚVOD

Jedným z dôležitých problémov, ktoré Európska únia v súčasnosti rieši, je znižovanie počtu ľudí ohrozených chudobou a sociálnym vylúčením. Stratégia Európa 2020 stanovila pre sociálnu oblasť konkrétny cieľ – vymaniť z ohrozenej časti populácie 20 miliónov obyvateľov. Na monitorovanie aktuálneho stavu bol vyvinutý agregovaný ukazovateľ AROPE (at risk of poverty or social exclusion), ktorý obsahuje niekoľko konkrétnych čiastkových indikátorov (podrobnejšie v [7]). Popri nich však existujú aj ďalšie ukazovatele odvíjajúce sa od sociálnej situácie obyvateľov, ktoré tiež naznačujú príslušnosť k ohrozenej časti populácie. Jedným z nich je problém s dostupnosťou zdravotnej starostlivosti.

Úroveň zdravotnej starostlivosti v Európskej únii pravidelne monitoruje Svetová zdravotnícka organizácia (WHO – World Health Organization). Analytická a informačná organizácia Health Consumer Powerhouse každoročne stanovuje hodnotu komplexného ukazovateľa EHCI (European Health Consumer Index). Podľa tohto kritéria má úroveň zdravotnej starostlivosti na Slovensku od roku 2013 klesajúci trend. V roku 2016 bola hodnota ukazovateľa 678, čo zaraďuje Slovensko na 23. miesto spomedzi 35 európskych krajín (podrobnejšie v [8] a [9]).

V príspevku sme sa zaoberali najmä takými indikátormi kvality zdravotnej starostlivosti, ktoré priamo vyplývajú zo sociálnej situácie jej poberateľov – nedostupnosťou zdravotnej starostlivosti z finančných dôvodov, z dôvodu príliš veľkej vzdialenosti medzi zariadením poskytujúcim zdravotnú starostlivosť a jej poberateľom a z dôvodu príliš dlhej čakacej lehoty na zdravotnú starostlivosť. Na analýzu sme využili databázy výberových zisťovaní EHIS a EU SILC, ktoré v pravidelných intervaloch realizuje Štatistický úrad Slovenskej republiky a v ktorých sa uvedené aspekty monitorujú z pohľadu respondentov – poberateľov zdravotnej starostlivosti.

V 2. kapitole sme analyzovali údaje z databázy EHIS: okrem podielov obyvateľov, pre ktorých je zdravotná starostlivosť nedostupná, sme identifikovali niektoré faktory, ktoré ovplyvňujú mieru jej nedostupnosti. V 3. kapitole sme na odhad niektorých podielov využili bayesovské metódy, preto sme do analýzy zaradili aj údaje z výberového zisťovania EU SILC, ktoré boli podkladom na modelovanie apriórnej informácie.

## 2. ANALÝZA NEDOSTUPNOSTI ZDRAVOTNEJ STAROSTLIVOSTI NA ZÁKLADE UKAZOVATEĽOV Z VÝBEROVÉHO ZISŤOVANIA EHIS

Európske zisťovanie o zdraví EHIS 2014 je už druhou vlnou tohto zisťovania (prvá prebehla v roku 2009). Realizovalo sa na základe nariadenia Európskej komisie č. 141/2013 z 19. februára 2013, ktorým sa vykonáva nariadenie Európskeho parlamentu a Rady (ES) č. 1338/2008 o štatistikách Spoločenstva v oblasti verejného zdravia a bezpečnosti a ochrany zdravia pri práci (podrobnejšie informácie o zisťovaní sú v [3]).

Dotazník predložený respondentom obsahoval štyri moduly. My sme sa zaoberali predovšetkým časťou „Európsky modul o zdravotnej starostlivosti“, v ktorej sú (aj) otázky týkajúce sa nenaplnenej potreby zdravotnej starostlivosti. Z ostatných modulov sme analyzovali len tie premenné, ktoré podľa nášho názoru môžu nejakým spôsobom ovplyvniť dostupnosť zdravotnej starostlivosti.

Z už spomenutého modulu sme sa venovali skupine otázok UN (Unmet needs for healthcare). Vzhľadom na to, že v 3. kapitole budeme analyzovať niektoré s nimi súvisiace otázky, ktoré však nemajú rovnaké znenie, je potrebné uviesť presnú štylizáciu jednotlivých otázok:

UN1a: Neuspokojená potreba zdravotnej starostlivosti za posledných 12 mesiacov z dôvodu dlhej čakacej doby[1].
UN1b: Neuspokojená potreba zdravotnej starostlivosti za posledných 12 mesiacov z dôvodu veľkej vzdialenosti alebo problémov s dopravou.
UN2a: Respondent si lekársku starostlivosť za posledných 12 mesiacov nemohol dovoliť.
UN2b: Respondent si zubnú starostlivosť za posledných 12 mesiacov nemohol dovoliť.
UN2c: Respondent si za posledných 12 mesiacov nemohol dovoliť lieky.
UN2d: Respondent si za posledných 12 mesiacov nemohol dovoliť starostlivosť týkajúcu sa duševného zdravia (napr. psychologickú alebo psychiatrickú starostlivosť).

Pri všetkých otázkach si respondent mohol vybrať jednu z troch odpovedí: „áno“, „nie“ a „nepotreboval(a) som zdravotnú starostlivosť“. Takto formulované odpovede poskytujú možnosť stanovenia dvoch druhov podielov: podiel tých, ktorých potreba starostlivosti nebola uspokojená, zo všetkých respondentov, ale aj podiel tých, ktorých potreba nebola uspokojená, z tých, ktorí ju naozaj potrebovali. Je zrejmé, že druhý z uvedených podielov je vyšší, ale poskytuje hodnotnejšiu informáciu o nedostupnosti niektorej z foriem starostlivosti. Porovnanie podielov je v tabuľke č. 1.

***Tabuľka č. 1: Podiely respondentov, ktorých potreba zdravotnej starostlivosti nebola uspokojená (a – podiel zo všetkých respondentov, b – podiel z tých, ktorí potrebovali starostlivosť)***

| Otázka | UN1a | UN1b | UN2a | UN2b | UN2c | UN2d |
|---|---|---|---|---|---|---|
| Podiel respondentov (a) | 4,83 % | 1,06 % | 1,68 % | 4,95 % | 3,39 % | 0,29 % |
| Podiel respondentov (b) | 6,26 % | 1,42 % | 2,21 % | 6,95 % | 4,89 % | 1,80 % |

***Zdroj údajov: vlastné výpočty z databázy EHIS 2014***

Z údajov v tabuľke vyplýva, že podiely neuspokojených respondentov vypočítané z tých, ktorí potrebovali niektorú z foriem zdravotnej starostlivosti, sú približne o 30 až 40 percent vyššie ako podiely získané z celkového počtu respondentov. Výnimku tvorí porovnanie podielov pri otázke UN2d, kde je rozdiel v podieloch s odstupom najvyšší. Táto disproporcia je pravdepodobne spôsobená obsahom otázky – časť populácie môže mať isté zábrany poskytovať informácie o potrebe (rovnako uspokojenej aj neuspokojenej) starostlivosti o duševné zdravie. Táto otázka by asi potrebovala osobitnú analýzu, preto sa jej v ďalšej časti budeme venovať len okrajovo.

[1] *Analýze dlhých čakacích lehôt pri poskytovaní lekárskej starostlivosti je venovaný príspevok [5].*

Z vypočítaných percentuálnych podielov sa ukazuje ako veľký problém nedostupnosť zubnej starostlivosti (takmer 7 percent z tých, ktorí ju potrebujú, si ju z finančných dôvodov musia odoprieť). Z hľadiska optimalizácie siete poskytovateľov sa ako oveľa závažnejší problém javia príliš dlhé čakacie lehoty (viac ako 6 percent respondentov nemalo uspokojenú potrebu starostlivosti z tohto dôvodu) ako veľká vzdialenosť od poskytovateľa (menej ako 1,5 percenta ju uviedlo ako dôvod neuspokojenej potreby).

Popri premenných zo skupiny UN (v nasledujúcom texte budeme úplné znenie otázok väčšinou nahrádzať kódmi otázok UN1a, UN1b, UN2a, UN2b, UN2c a UN2d) sme venovali pozornosť aj takým premenným, ktoré môžu nejakým spôsobom súvisieť s dostupnosťou niektorej z foriem starostlivosti alebo ju priamo ovplyvniť. Ide o tieto premenné zo skupiny hlavných sociálnych premenných:

SEX: *pohlavie,*
AGE: *vek,*
REGION: *región trvalého bydliska* (2-miestny kód na základe klasifikácie NUTS),
DEG_URB: *stupeň urbanizácie* (najmä v súvislosti s otázkou UN1b),
HATLEVEL: *najvyššia úroveň dosiahnutého vzdelania* (na základe klasifikácie ISCED 2011),
MAINSTAT: *respondentom uvádzaný stav z hľadiska zamestnania*,
JOBSTAT: *postavenie v zamestnaní*,
HHINCOME: *čistý mesačný príjem ekvivalentný domácnosti* (rozdelený na kategórie podľa kvintilov).

Zo skupiny premenných týkajúcich sa zdravia sme vybrali premennú HS1: *všeobecný zdravotný stav vnímaný respondentom*.

Pre všetky uvedené premenné sme vykonali testy nezávislosti s každou z premenných zo skupiny UN, pričom každý test sme realizovali dvakrát: v prvom prípade (možnosť označíme ako (a)) sme brali do úvahy všetky (tri možné) odpovede na otázky zo skupiny UN („áno“, „nie“, „nepotreboval(a) som zdravotnú starostlivosť“), v druhom prípade (b) sme brali do úvahy len odpovede „áno“ a „nie“; respondentov s odpoveďou „nepotreboval(a) som“ sme z analýzy vylúčili. Všetky testy sme realizovali pomocou štatistického softvéru Statgraphics Centurion.

V tabuľke č. 2 sú uvedené p-hodnoty testov nezávislosti[2] medzi jednotlivými premennými skupiny UN a premennými, ktoré sme vybrali ako potenciálne faktory. Premennú *vek* sme pretransformovali do skupín tvorených intervalmi ⟨15; 20⟩, (20; 30⟩, ..., (60; 70⟩, pričom do poslednej skupiny sme zaradili vyše 70-ročných respondentov.

Porovnaním zodpovedajúcich p-hodnôt v tabuľkách vidíme, že v prípade troch uvažovaných odpovedí (časť (a)) na otázky zo skupiny UN (s výnimkou faktora s*tupeň urbanizácie*) je závislosť medzi premennými silnejšia ako v prípade dvoch uvažovaných odpovedí (časť (b)). My sa však budeme venovať časti (b), pretože

[2] *p-hodnoty sú zaokrúhlené na 4 desatinné miesta; v takejto podobe ich uvádza výstup z použitého softvéru Statgraphics Centurion, číslo 0,0000 v tabuľke tak vyjadruje p-hodnotu, ktorá je menšia ako 0,00005.*

práve táto možnosť podľa nás lepšie vystihuje mieru nedostupnosti niektorej z foriem zdravotnej starostlivosti.

Z vypočítaných p-hodnôt je zrejmé, že ak by sme hypotézy o nezávislosti testovali na hladine významnosti 0,05, z 54 testovaných dvojíc znakov by sa až v 31 prípadoch preukázala závislosť. Vzhľadom na limitovaný rozsah príspevku nie je reálne interpretovať a analyzovať všetky uvažované dvojice, preto počet závislých dvojíc znakov čiastočne zredukujeme znížením hladiny významnosti na 0,01. Dostaneme tak 26 dvojíc závislých znakov, pri ktorých uvedieme len najpodstatnejšie výsledky porovnania jednotlivých úrovní faktorov.

***Tabuľka č. 2 (a) p-hodnoty testov nezávislosti medzi premennými zo skupiny UN s tromi obmenami a uvažovanými faktormi***

| | UN1a | UN1b | UN2a | UN2b | UN2c | UN2d |
|---|---|---|---|---|---|---|
| SEX | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0006 |
| AGE | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| REGION | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| DEG_URB | 0,2438 | 0,2703 | 0,0137 | 0,0000 | 0,1679 | 0,0059 |
| HATLEVEL | 0,0000 | 0,0004 | 0,0000 | 0,0000 | 0,0000 | 0,0003 |
| MAINSTAT | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| JOBSTAT | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0002 |
| HHINCOME | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0003 |
| HS1 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |

***Zdroj údajov: vlastné výpočty s použitím softvéru Statgraphics Centurion z databázy EHIS 2014***

***Tabuľka č. 2 (b) p-hodnoty testov nezávislosti medzi premennými zo skupiny UN s dvomi obmenami a uvažovanými faktormi***

| | UN1a | UN1b | UN2a | UN2b | UN2c | UN2d |
|---|---|---|---|---|---|---|
| SEX | 0,3291 | 0,3067 | 0,4869 | 0,0018 | 0,4782 | 0,6850 |
| AGE | 0,0449 | 0,0673 | 0,0023 | 0,0014 | 0,0000 | 0,6639 |
| REGION | 0,0001 | 0,6864 | 0,0000 | 0,0000 | 0,0000 | 0,0572 |
| DEG_URB | 0,0745 | 0,2985 | 0,3589 | 0,0119 | 0,3268 | 0,1076 |
| HATLEVEL | 0,1008 | 0,2724 | 0,0008 | 0,0000 | 0,0000 | 0,0136 |
| MAINSTAT | 0,0181 | 0,0024 | 0,0000 | 0,0000 | 0,0000 | 0,1317 |
| JOBSTAT | 0,2506 | 0,0484 | 0,0004 | 0,0000 | 0,0000 | 0,6041 |
| HHINCOME | 0,1486 | 0,1776 | 0,0000 | 0,0000 | 0,0000 | 0,2739 |
| HS1 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0006 | 0,2667 |

***Zdroj údajov: vlastné výpočty s použitím softvéru Statgraphics Centurion z databázy EHIS 2014***

Faktor *pohlavie* štatisticky významne ovplyvňuje len otázku UN2b – medzi mužmi a ženami je štatisticky významný rozdiel pri nedostupnosti zubnej starostlivosti; nemohlo si ju dovoliť až 8,02 % žien, ale len 5,45 % mužov[3].

Faktor *vek* štatisticky významne ovplyvňuje odpovede na otázky UN2a, UN2b a UN2c. Ukázalo sa, že finančná dostupnosť jednotlivých foriem starostlivosti je

[3] *Percentuálne podiely sú súčasťou výstupov zo softvéru; vzhľadom na limitovaný rozsah príspevku nie je reálne prezentovať všetky frekvenčné tabuľky.*

najproblematickejšia pre vekovú skupinu 50- až 60-ročných respondentov (najvyšší podiel sa týka zubnej starostlivosti – 9,47 %), najlepšia dostupnosť je v kategórii do 20 rokov.

Faktor *kraj* (REGION) štatisticky významne ovplyvňuje odpovede na všetky otázky okrem UN1b a UN2d. Z hľadiska dlhej čakacej lehoty je na tom najhoršie Východoslovenský kraj (až pre 8,85 % respondentov je starostlivosť z tohto dôvodu nedostupná); najnižší podiel je v Stredoslovenskom kraji (4,21 %). Finančná nedostupnosť niektorej z foriem starostlivosti je najvyššia v Bratislavskom kraji (UN2a – 5,49 %, UN2b – 19,03 % a UN2c – 12,58 %). Najnižšie podiely sú v Stredoslovenskom kraji.

Faktor *stupeň urbanizácie* sa ukázal ako najmenej významný spomedzi všetkých uvažovaných faktorov. Tento výsledok je pre nás prekvapujúci najmä v súvislosti s otázkou UN1b, pretože sme očakávali, že v husto obývanej oblasti majú respondenti lepšiu dostupnosť (z hľadiska vzdialenosti a dopravy) ako v riedko obývaných oblastiach. Podiely sú síce rozdielne, ale vzhľadom na vysokú p-hodnotu nie sú štatisticky významné. Podrobnejšia analýza tohto faktora (Šoltés, Gajdošík, 2016) ukazuje aj spolupôsobenie iných faktorov, ktorých úrovne sa v jednotlivých regiónoch líšia.

Faktor HATLEVEL (*najvyššia úroveň dosiahnutého vzdelania*) štatisticky významne ovplyvnil premenné UN2a, UN2b a UN2c. Ukázalo sa, že finančná dostupnosť troch foriem starostlivosti je priamo úmerná dosiahnutému vzdelaniu respondentov. Podiel respondentov, ktorí si nemohli dovoliť zdravotnú starostlivosť, je najvyšší v skupine respondentov s nižším sekundárnym vzdelaním (4,06 %), najnižší v skupine s bakalárskym stupňom vzdelania. Podiel respondentov, ktorí si nemohli dovoliť zubnú starostlivosť, je najvyšší v skupine respondentov s primárnym vzdelaním (16,3 %), najnižší v skupine respondentov s vysokoškolským vzdelaním druhého stupňa (3,5 %). Najväčší problém s finančnou dostupnosťou liekov mali respondenti s primárnym vzdelaním (14,71 %), najmenší respondenti s bakalárskym stupňom vzdelania (1,09 %).

Faktor MAINSTAT (*respondentom uvádzaný stav z hľadiska zamestnania*) významne ovplyvnil premenné UN1b, UN2a, UN2b a UN2c. Z hľadiska nedostupnosti zdravotnej starostlivosti z dôvodu vzdialenosti alebo problémov s dopravou sú na tom najhoršie osoby s trvalým zdravotným postihnutím (3,66 %), na druhej strane v skupine študentov a osôb vykonávajúcich domáce práce je podiel 0 %. Zdravotnú starostlivosť si z finančných dôvodov nemohlo dovoliť najviac respondentov s trvalým zdravotným postihnutím (7,29 %), najnižší podiel je v skupine zamestnaných osôb (1,06 %). Najvyšší podiel respondentov, ktorí si nemohli dovoliť zubnú starostlivosť, je v skupine nezamestnaných osôb (17,89 %), najnižší podiel je medzi študentmi (3,23 %). Osoby, ktoré si nemohli dovoliť kúpiť lieky, tvoria najvyšší podiel v skupine nezamestnaných osôb (13,99 %), najnižší podiel je medzi študentmi (1,38 %).

Faktor JOBSTAT (*postavenie v zamestnaní* – týka sa len zamestnaných osôb) má vplyv na premenné UN2a, Un2b a UN2c. Najlepší výsledok je v kategórii samostatne zárobkovo činných osôb, kde sú podiely pre všetky premenné najnižšie (UN2a – 0,5 %, UN2b – 3,13 %, UN2c – 1,16 %). Najvyšší podiel tých, ktorí si nemohli dovoliť

zdravotnú starostlivosť, je v skupine osôb so stálym zamestnaním (1,15 %). U osôb s dočasným zamestnaním boli najvyššie podiely pri nedostupnosti zubnej starostlivosti (7,41 %) a nedostatku finančných prostriedkov na kúpu liekov (3,23 %).

Ako sme predpokladali, faktor HHINCOME (*čistý mesačný ekvivalentný príjem domácnosti*) štatisticky významne ovplyvnil premenné týkajúce sa nedostatku finančných prostriedkov: UN2a, UN2b a UN2c. Pre všetky tieto premenné platí, že najvyššie podiely osôb, ktoré si nemohli dovoliť jednu z foriem starostlivosti, boli v skupine ľudí s najnižšími príjmami (pod 1. kvintilom) (UN2a – 4,65 %, UN2b – 16,93 %, UN2c – 12,17 %); najnižšie podiely boli v skupine s príjmom medzi 4. a 5. kvintilom (UN2a – 0,74 %, UN2b – 3,48 %, UN2c – 1,81 %), pričom pri každej premennej sa podiely s prechodom do vyššej príjmovej skupiny znižovali.

Faktor HS1 (*všeobecný zdravotný stav vnímaný respondentom*) sa ukázal ako najvýznamnejší spomedzi všetkých uvažovaných – s výnimkou premennej UN2d totiž štatisticky významne ovplyvnil všetky premenné zo skupiny UN. Náš predpoklad, že najmenší problém s dostupnosťou zdravotnej starostlivosti majú osoby s dobrým zdravotným stavom, sa splnil. Pri všetkých premenných bol najnižší podiel v skupine osôb, ktoré svoj zdravotný stav hodnotia ako veľmi dobrý (UN1a – 3,41 %, UN1b – 0 %, UN2b – 3,84 %, UN2c – 1,88 %), resp. ako dobrý (UN2a – 0,91 %). Najvyššie podiely boli zaznamenané u osôb, ktoré hodnotia svoj zdravotný stav ako zlý (UN1a – 10,95 %, UN2b – 14,29 %, UN2c – 12,02 %), resp. veľmi zlý (UN1b – 5,88 %, UN2a – 5,63).

## 3. MOŽNOSTI APLIKÁCIE BAYESOVSKÝCH METÓD PRI ODHADE PODIELU OBYVATEĽOV SR, KTORÍ MAJÚ PROBLÉM S DOSTUPNOSŤOU ZDRAVOTNEJ STAROSTLIVOSTI

Bayesovská štatistika ([1],[2]) je vhodným nástrojom na spresnenie odhadov neznámych parametrov v prípade, ak okrem údajov z výberového zisťovania existuje ešte iný dôveryhodný zdroj informácie o odhadovanom parametri. Základný matematický rozdiel medzi klasickou induktívnou a bayesovskou štatistikou je v ponímaní odhadovaného parametra. V klasickej štatistike ho považujeme za konštantu, bayesovská štatistika k nemu pristupuje ako k náhodnej premennej (označujeme symbolom $\Theta$), ktorej rozdelenie sa vplyvom nových informácií postupne aktualizuje. Vzťah medzi hustotou rozdelenia odhadovaného parametra a hustotou rozdelenia, z ktorého pochádza náhodný výber, definuje spojitá verzia Bayesovej vety (uvádzame jej skrátený tvar podľa [6]):

$$f_{\Theta}(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta) \cdot f_{\Theta}(\theta) \ , \qquad (1)$$

v ktorej jednotlivé symboly majú nasledujúci význam:

$f_{\Theta}(\theta)$ je hustota *apriórneho rozdelenia* parametra $\Theta$, ktoré stvárňuje informáciu o odhadovanom parametri pochádzajúcu z iného zdroja, ako je aktuálny náhodný výber,

$f(\mathbf{x} \mid \theta)$ je funkcia vierohodnosti ($f(x)$ je hustota rozdelenia, z ktorého pochádza náhodný výber),

$f_{\Theta}(\theta \mid \mathbf{x})$ je hustota aposteriórneho rozdelenia, ktoré je výsledkom aktualizácie apriórneho rozdelenia na základe údajov z náhodného výberu.

Vzťah (1) teda vyjadruje fakt, že hustota aposteriórneho rozdelenia je proporcionálna súčinu hustoty apriórneho rozdelenia a funkcie vierohodnosti. Aposteriórne rozdelenie je podkladom na induktívne závery o odhadovanom parametri.

Ak sú apriórne a aposteriórne rozdelenia rovnakého typu, voláme ich *konjugované* rozdelenia k rozdeleniu, z ktorého pochádza náhodný výber. Spolu s výberovým rozdelením tvoria tzv. *konjugovaný systém* (conjugate family). Existencia konjugovaných rozdelení značne zjednodušuje výpočty, pretože pre veľkú časť z nich boli odvodené vzorce, pomocou ktorých možno vypočítať parametre aposteriórneho rozdelenia pomocou parametrov apriórneho rozdelenia a niektorých výberových charakteristík. Najviac využívanými konjugovanými systémami sú binomické/beta, Poissonovo/gama a normálne/normálne. My sa budeme venovať prvému z nich, preto uvedieme základné fakty a súvislosti medzi nimi.

V konjugovanom systéme binomické/beta predpokladáme, že výber pochádza z binomického rozdelenia, v ktorom odhadujeme parameter $\pi$. Ak apriórnym rozdelením tohto parametra je beta rozdelenie s parametrami $\alpha, \beta,$ tak aposteriórne rozdelenie tohto parametra je tiež rozdelenie typu beta, pričom pre jeho parametre $\alpha', \beta'$ platia vzťahy (uvedené v [6])

$$\alpha' = \alpha + x \tag{2}$$

$$\beta' = \beta + n - x \ , \tag{3}$$

v ktorých $n$ označuje počet pokusov (rozsah výberového súboru) a $x$ je počet pokusov, pri ktorých nastala sledovaná udalosť (počet štatistických jednotiek s určitou vlastnosťou). Bayesovským bodovým odhadom podielu (parametra $\pi$) je stredná hodnota aposteriórneho rozdelenia

$$E(\pi / \mathbf{x}) = \frac{\alpha + x}{\alpha + \beta + n} \ . \tag{4}$$

Intervalový odhad možno stanoviť viacerými spôsobmi; najjednoduchšie je za hranice intervalu určiť kvantily aposteriórneho rozdelenia $\pi / \mathbf{x}_{\frac{\alpha}{2}}$ a $\pi / \mathbf{x}_{1-\frac{\alpha}{2}}$ zodpovedajúce zvolenej spoľahlivosti $1 - \alpha$.

Naším cieľom bolo odhadnúť podiely osôb, ktoré si nemôžu dovoliť zdravotnú starostlivosť, presnejšie, ako sme to urobili v časti 2. Preto sme do analýzy popri údajoch z výberového zisťovania EHIS zaradili aj údaje z výberového zisťovania EU SILC, ktoré sme využili ako apriórnu informáciu[4]. V časti P (Personal data) sú otázky, ktorými sa tiež zisťuje miera dostupnosti zdravotnej starostlivosti a hlavné príčiny, pre ktoré nebola niektorá z foriem starostlivosti poskytnutá. Formulácia

[4] *V [4] autori využili databázu EU SILC na analýzu nedostupnosti zdravotnej starostlivosti v súvislosti so systémom zdravotnej starostlivosti v krajinách Európskej únie.*

a logická postupnosť otázok je však odlišná ako v dotazníku EHIS, preto sme mohli využiť len niektoré z nich. Konkrétne sme sa zaoberali týmito otázkami:

PH040: Nenaplnená potreba medicínskeho vyšetrenia alebo liečby počas posledných 12 mesiacov.
PH050: Hlavný dôvod pre nenaplnenie potreby medicínskeho vyšetrenia alebo liečby.
PH060: Nenaplnená potreba zubného vyšetrenia alebo liečby počas posledných 12 mesiacov.
PH070: Hlavný dôvod pre nenaplnenie potreby zubného vyšetrenia alebo liečby.

Pri otázkach PH040 a PH060 si respondenti mohli vybrať z dvojice odpovedí „áno“ (bola aspoň jedna taká situácia) alebo „nie“ (nebola žiadna taká situácia). Odpovede na otázky PH050 a PH060 si respondenti vyberali z nasledujúcich možností:

1 nemohli ste si to dovoliť (príliš drahé)
2 poradovník
3 nemohli ste sa uvoľniť kvôli práci, starostlivosti o deti alebo iné
4 museli by ste cestovať príliš ďaleko, nemali by ste sa tam ako dopraviť
5 strach z lekára (zubára)/nemocníc/vyšetrení/liečby
6 chceli ste počkať a zistiť, či sa problém nezlepší (nevyrieši sám)
7 nepoznali ste dobrého lekára/špecialistu
8 iné príčiny

Je zrejmé, že uvedené poradie a formulácia otázok delí respondentov na iné skupiny, ako je to v prípade zisťovania EHIS, takže vypočítané podiely by boli neporovnateľné a neprispeli by tak k spresneniu odhadov vypočítaných v časti 2.

Po logickej analýze otázok v obidvoch zisťovaniach sme našli indikátor, ktorý má rovnaký obsah v obidvoch zisťovaniach: zaoberali sme sa podielom respondentov, ktorých potreba zdravotnej starostlivosti nebola naplnená z niektorej (práve z jednej) z nasledujúcich príčin: nemohli si to dovoliť (1), poradovník (2) alebo by museli cestovať príliš ďaleko a nemali sa tam ako dopraviť (4). Uvažované podiely sa týkajú všetkých oslovených respondentov, pretože ponuka odpovedí v EU SILC neobsahuje možnosť „nepotreboval som“.

Stanovený indikátor z údajov zisťovania EHIS vypočítame ako podiel tých respondentov (zo všetkých), ktorí (práve) v jednej z otázok UN1a, UN1b a UN2a odpovedali „áno“. Ukázalo sa, že takýchto respondentov je spolu 292 ($x = 292$) z celkového počtu 5 490 ($n = 5490$) respondentov; výberový podiel má teda hodnotu 0,0541 (5,41 %).

Z údajov zisťovania EU SILC mu zodpovedá podiel tých respondentov, ktorí na otázku PH040 odpovedali „áno“ a v otázke PH050 zvolili jednu z možností (1), (2) alebo (4). Na modelovanie apriórneho rozdelenia sme využili údaje zo zisťovaní z rokov 2005 – 2015, ktoré sú v tabuľke č. 3.

Vypočítané podiely sme použili na modelovanie apriórneho beta rozdelenia. Metódou maximálnej vierohodnosti sme odhadli jeho parametre: $\alpha = 15,466$, $\beta = 746,788$ a pomocou Kolmogorovovho-Smirnovovho testu sme overili vhodnosť

navrhnutého rozdelenia. Výsledok testu je na obrázku 1 (výstup zo Statgraphics Centurion).

***Tabuľka č. 3: Počty respondentov, ktorí si pri výberovom zisťovaní EU SILC na otázku PH050 zvolili jednu z možností (1), (2) alebo (4) v rokoch 2005 – 2015***

| ROK | (1) | (2) | (4) | (1) + (2) + (4) | Počet respondentov | Výberový podiel |
|---|---|---|---|---|---|---|
| 2005 | 324 | 44 | 24 | 392 | 12 868 | 0,030463 |
| 2006 | 278 | 42 | 33 | 353 | 12 630 | 0,027949 |
| 2007 | 109 | 46 | 18 | 173 | 12 570 | 0,013763 |
| 2008 | 67 | 71 | 41 | 179 | 13 645 | 0,013118 |
| 2009 | 84 | 81 | 60 | 225 | 13 580 | 0,016568 |
| 2010 | 80 | 101 | 47 | 228 | 13 907 | 0,016395 |
| 2011 | 89 | 149 | 43 | 281 | 13 261 | 0,021190 |
| 2012 | 112 | 135 | 35 | 282 | 13 502 | 0,020886 |
| 2013 | 92 | 135 | 23 | 250 | 13 044 | 0,019166 |
| 2014 | 114 | 130 | 32 | 276 | 13 187 | 0,020930 |
| 2015 | 102 | 154 | 52 | 308 | 13 535 | 0,022756 |

***Zdroj údajov: vlastné výpočty z databázy EU SILC 2005 – 2015***

***Obrázok č. 1: Výstup z procedúry Distribution Fitting softvéru Statgraphics Centurion***

**Goodness-of-Fit Tests for podiely**

Kolmogorov-Smirnov Test

| | *Beta* |
|---|---|
| DPLUS | 0,126352 |
| DMINUS | 0,123817 |
| DN | 0,126352 |
| P-Value | 0,994681 |

***Zdroj: Statgraphics Centurion, vlastné výpočty***

Vysoká p-hodnota jednoznačne potvrdzuje vhodnosť navrhnutého rozdelenia, preto ho možno použiť ako apriórne rozdelenie na bayesovský odhad podielu. Pomocou vzťahov (2) a (3) sme vypočítali parametre aposteriórneho beta rozdelenia:

$$\alpha' = 15,466 + 297 = 312,466$$

$$\beta' = 746,788 + 5490 - 297 = 5939,788$$

Bayesovským bodovým odhadom podielu je stredná hodnota tohto aposteriórneho rozdelenia, ktorú sme vypočítali pomocou vzťahu (4):

$$\hat{\pi}_B = E(\pi / \mathbf{x}) = \frac{15,466 + 297}{15,466 + 746,788 + 5490} = 0,04998 \quad = 5\ \%$$

Vypočítaný podiel 5 % je kompromisom medzi dvomi vstupnými informáciami: výberovým podielom 5,41 % a podielom, ktorý možno stanoviť len na základe apriórneho rozdelenia – jeho stredná hodnota je 0,0203 = 2,03 %. Údaje z výberového zisťovania EU SILC tak skorigovali pôvodný odhad smerom dolu.

Napriek tomu však tento výsledný podiel považujeme za vysoký – znamená, že takmer každý dvadsiaty obyvateľ Slovenskej republiky nemá uspokojenú potrebu

zdravotnej starostlivosti z niektorého z troch predtým uvedených dôvodov. Aj keď len v otázke UN2a je dôvodom neuspokojenia potreby zdravotnej starostlivosti priamo nedostatok finančných prostriedkov, aj otázky UN1a (príliš dlhá čakacia lehota) a UN2a (priveľká vzdialenosť od poskytovateľa, problémy s dopravou) majú istý sociálny podtón. Obidve uvedené prekážky by sa do určitej miery dali prekonať, ak by bola k dispozícii dostatočná finančná suma. Navyše prispôsobením analýzy možnostiam odpovedí vo výberovom zisťovaní EU SILC sa odhadnuté podiely týkajú všetkých obyvateľov, teda nielen tých, ktorí zdravotnú starostlivosť naozaj potrebujú. Vzhľadom na zastúpenie jednotlivých odpovedí vo výberovom zisťovaní EHIS možno predpokladať, že takýto podiel by bol aspoň o 20 % vyšší.

Pomocou aposteriórneho rozdelenia môžeme analyzovaný podiel odhadnúť aj pomocou intervalu spoľahlivosti. Pre 95-percentnú spoľahlivosť sú hranicami intervalu zodpovedajúce kvantily aposteriórneho rozdelenia: dolná hranica je $\pi / \mathbf{x}_{0,025} = 0{,}0447$ a horná $\pi / \mathbf{x}_{0,975} = 0{,}0555$. (Na porovnanie: interval spoľahlivosti vypočítaný len z apriórneho rozdelenia má hranice 0,0115 a 0,0314 a klasický interval spoľahlivosti vypočítaný z výberových údajov (EHIS) má hranice 0,0481 a 0,0601. Obidva tieto intervaly sú širšie v porovnaní s bayesovským intervalom: interval z apriórneho rozdelenia má rozpätie 0,0199, klasický interval 0,012 a bayesovský len 0,0108. Bayesovský interval tak poskytuje najpresnejší odhad.)

## 4. ZÁVER

Nedostupnosť zdravotnej starostlivosti (najmä z finančných, ale aj z iných dôvodov) možno oprávnene považovať za indikátor chudoby a sociálneho vylúčenia. Pomocou údajov z výberového zisťovania EHIS 2014 sme odhadli, že až 6,26 % tých obyvateľov Slovenskej republiky, ktorí potrebujú zdravotnú starostlivosť, ju nedostanú z dôvodu príliš dlhej čakacej lehoty, 1,42 % z dôvodu príliš veľkej vzdialenosti od poskytovateľa zdravotnej starostlivosti a 2,21 % preto, že si ju nemôžu dovoliť. Nedostatok finančných prostriedkov je aj hlavnou príčinou, pre ktorú 6,95 % obyvateľov nedostáva potrebnú zubnú starostlivosť, 4,89 % obyvateľov sa nedostane k predpísaným liekom a 1,8 % má nenaplnenú potrebu starostlivosti o duševné zdravie napriek tomu, že uvedené formy starostlivosti podľa vlastného vyjadrenia naozaj potrebujú.

Analýza faktorov ovplyvňujúcich nedostupnosť jednotlivých foriem starostlivosti preukázala, že tento problém sa týka najmä najzraniteľnejších skupín obyvateľstva – starších ľudí, ľudí so zlým zdravotným stavom, s trvalým zdravotným postihnutím, nízkym stupňom vzdelania a nízkymi príjmami. V mnohých prípadoch je podiel ľudí s neuspokojenou potrebou zdravotnej starostlivosti v týchto skupinách vyšší ako 10 %. Táto situácia si rozhodne vyžaduje pozornosť kompetentných orgánov. Je žiaduce prijať také opatrenia, ktoré by práve takýmto znevýhodneným skupinám zlepšili dostupnosť zdravotnej starostlivosti.

Na spresnenie odhadov z výberového zisťovania EHIS sme využili aj databázu z výberových zisťovaní EU SILC za roky 2005 – 2015, z ktorých sme modelovali apriórnu situáciu ako podklad na bayesovský odhad analyzovaných podielov. Podrobnou analýzou znenia otázok v obidvoch výberových zisťovaniach sme však nenašli širšie spektrum porovnateľných množín respondentov (a z nich vypočítaných podielov), preto sme sa sústredili len na odhad podielu respondentov, ktorí nedostali zdravotnú starostlivosť z jedného z trojice dôvodov: príliš dlhá čakacia lehota, príliš

veľká vzdialenosť od poskytovateľa, nedostatok financií. Využitie apriórnej informácie spôsobilo korekciu pôvodného podielu (5,41 %) smerom nadol na hodnotu 5,00 %. Výsledný podiel však aj tak považujeme za príliš veľký. Ak z aktuálneho počtu obyvateľov Slovenskej republiky (cca 5 400 000) približne 84,7 % má vek aspoň 15 rokov (tak ako respondenti zisťovania EHIS), ide o vyše 220-tisíc ľudí, pre ktorých je niektorá z foriem zdravotnej starostlivosti nedostupná z dôvodov, ktoré možno zlepšením ich sociálnej situácie eliminovať.

Niektoré výsledky, ktoré sa v tomto príspevku prezentujú, sa môžu javiť ako prekvapujúce alebo ako málo pravdepodobné. Treba si však uvedomiť, že všetky odhady boli stanovené na základe unikátneho zisťovania (EHIS 2014), v ktorom sa pohľad na dostupnosť zdravotnej starostlivosti prezentuje výlučne z pohľadu prijímateľov zdravotnej starostlivosti. Porovnateľná databáza údajov (z hľadiska kvality, metodiky zisťovania a rozsahu) neexistuje, a preto aj závery našej analýzy nemožno konfrontovať s publikáciami venujúcimi sa danej problematike.

Pri aplikácii bayesovských metód na odhad podielov obyvateľov s neuspokojenou potrebou zdravotnej starostlivosti sa ukázalo, že presná formulácia a poradie otázok zohráva kľúčovú úlohu pri výsledkoch zisťovania. V dvoch použitých zdrojoch (EHIS 2014 a EU SILC) boli skupiny otázok venujúcich sa problematike odlišne sformulované, čo mohlo do určitej miery ovplyvniť veľkosť zodpovedajúcich podielov. Rozdiely v logickej štruktúre otázok pri týchto zisťovaniach nám neumožnili využitie bayesovského prístupu v takom rozsahu, ako sme pôvodne zamýšľali.

***Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/0548/16 Pokrok SR pri napĺňaní stratégie EURÓPA 2020 v oblasti znižovania chudoby a sociálneho vylúčenia.***

## LITERATÚRA

[1] BERNARDO, JOSÉ M. – SMITH, ADRIAN F. M.: Bayesian theory. 2. edition, Chichester: John Wiley & Sons Ltd., 2000. 640 p. ISBN 978-0-471-49464-5.
[2] BOLSTAD, W. M.: Introduction to Bayesian statistics. 2. edition. New Yersey, USA: John Wiley & Sons, Inc., 2007. 437 p. ISBN 978-0-470-14115-1.
[3] Eurostat. European Health Interview Survey (EHIS wave 2)-Methodological manual. 2013. (Working paper No. KS-RA-13-018-EN-N).
[4] CHAUPMAN-GUILLOT, S. – GUILLOT, O.: Health system characteristics and unmet care needs in Europe: an analysis based on EU-SILC data. In: European Journal of Health Economics, 2015, No 7, p. 781-796.
[5] MUZIK, R. – SZALAYOVA, A.: Measuring Waiting Times in Slovakia. In: Proceedings of the 17th International Conference on Current Trends in Public Sector Research. Slapanice, Czech Republic: Masarykova univ., 2013, p. 66-74.
[6] PACÁKOVÁ, V. a kol.: Štatistická indukcia pre ekonómov. 1. vyd. Vydavateľstvo EKONÓM, 2012. 362 s. ISBN 978-80-225-3382-9.
[7] ŠOLTÉS, E. – GAJDOŠÍK, M.: Dopad revízie deprivačných položiek na hodnotenie materiálnej deprivácie slovenských domácností na základe databázy EU SILC 2014. In: Ekonomika a informatika, 2016, č. 2, s. 178 – 196.
[8] http://www.healthpowerhouse.com/media/prerelease/EHCI_2015_media.pdf. (posledný prístup k 30. 1. 2017).

[9] http://www.statistics.sk/pls/elisw/objekt.sendName?name=m_silk (posledný prístup k 10. 3. 2017).

**RESUME**

The article deals with estimating a proportion of the Slovak population with unmet needs for health care for reasons closely associated with their social situation.

The first section of the paper analyses the EHIS survey data, in which the availability of different types of health care is monitored from the recipients´ perspective. In addition to proportion estimates some relevant factors determining the degree of health care availability were identified: sex, age, region, educational attainment, health status, employment status and income level. For each factor, the variations with the highest and with the lowest proportion of people with unmet needs for health care are indicated.

In the second section of the paper, in order to make more precise estimates, the Bayesian approach is applied, using more than one source of information. Therefore, the data from the EU SILC survey were included in the analysis by means of which the a priori information (for Bayesian estimation, the conjugate family beta-binomial) was modelled. Considering the different sequence and the logical interconnection of questions in both surveys, it was not possible to use a Bayesian approach for the entire spectrum of proportions from the first section. Therefore, we used this technique to estimate only the proportion of those people for whom the medical care was not available for one of these three reasons: lack of financial resources, too long waiting time and too long distance from the health care provider. By using the Bayesian approach, the proportion was partially reduced and became more precise.

***PROFESIJNÝ ŽIVOTOPIS***

***RNDr. Eva Kotlebová, PhD.***, *je absolventkou Matematicko-fyzikálnej fakulty Univerzity Komenského v Bratislave (vedecký smer matematika – teória systémov). Po ukončení vysokoškolského štúdia bola tri roky na študijnom pobyte na Katedre štatistiky Fakulty riadenia Vysokej školy ekonomickej v Bratislave. Potom pôsobila niekoľko rokov ako stredoškolská učiteľka matematiky na gymnáziu v Bratislave. Od roku 2003 pracuje na Katedre štatistiky Fakulty hospodárskej informatiky v Bratislave. V roku 2008 ukončila doktorandské štúdium. Venuje sa štatistickej indukcii, bayesovskej štatistike a aplikácii štatistických metód v poisťovníctve.*

**KONTAKT**

eva.kotlebova@gmail.com

**Tomáš LÖSTER**
**Vysoká škola ekonomická v Praze**

# RŮZNÉ ZPŮSOBY STANOVENÍ POČTU SHLUKŮ VE SHLUKOVÉ ANALÝZE

## VARIOUS METHODS OF DETERMINING THE NUMBER OF CLUSTERS IN CLUSTER ANALYSIS

**ABSTRAKT**
V současné odborné literatuře existuje celá řada způsobů, jak stanovit optimální počet shluků. Běžný způsob, který je často v literatuře uváděn a v praxi využíván, spočívá v nalezení počtu shluků na základě grafu – dendrogramu. Jedná se však o značně subjektivní záležitost, a tak bývá doporučeno využít některý z koeficientů pro stanovení počtu shluků. Těchto koeficientů je celá řada a neexistuje jednoznačné pravidlo, které by definovalo použitelnost daných koeficientů. Cílem tohoto článku je ukázat vybrané možnosti stanovení počtu shluků v různých podmínkách při pevném hierarchickém shlukování.

**ABSTRACT**
In the current specialised literature, there are many methods to determine the optimal number of clusters. A common method which is often mentioned in literature and used in practice lies in determining the number of clusters on the basis of the graph - the dendrogram. However, it is a quite subjective matter, thus it is recommended to use one of the coefficients for determining the number of clusters. There is a large number of such coefficients and there is no clear rule for the use of these coefficients. The aim of this article is to show the selected possibilities of determining the number of clusters under various conditions in a stable hierarchical clustering.

**KLÍČOVÁ SLOVA**
shlukování, hodnocení shlukování, koeficienty pro stanovení optimálního počtu shluků

**KEY WORDS**
Clustering, Evaluation of clustering, Coefficients for determining the optimal number of clusters.

## 1. ÚVOD

Základní úlohou, kterou řeší mnoho vědních disciplín, je vytváření skupin objektů. Ty mohou být representovány zákazníky, pacienty, automobily, dokumenty, atd. K vytváření skupin mohou být využity různé matematicko-statistické metody a postupy. Pokud je objekt (pozorování) zařazován do existující skupiny, využívá se k tomu diskriminační analýza, pokud je objekt zařazován do tříd, které nemusí být předem známé, využívá se k tomu shluková analýza. Ta představuje vícerozměrnou statistickou metodu, jejímž cílem je vytváření skupin objektů, které se nazývají shluky. Uplatnění shlukové analýzy lze najít v mnoha odvětvích. Často se využívá při řešení ekonomických úloh. Vytváří se skupiny klientů podle různé strategie či rizika, shlukují se firmy, země, atd.

Základním cílem metod shlukové analýzy je vytvářet skupiny objektů (shluky), které jsou charakterizovány pomocí různých proměnných. Při vytváření shluků je důležité, aby si objekty, které jsou zařazeny uvnitř jednoho shluku, byly co nejvíce podobné a objekty, které jsou zařazeny do dvou různých shluků, si byly co nejméně podobné.

V současné odborné literatuře existuje mnoho metod shlukové analýzy. Ty mohou být členěny pomocí různých kritérií. Vývoj metod shlukové analýzy je spojen jednak se vznikem nových metod, jednak s modifikacemi stávajících metod. V současné době tak díky celé řadě softwarových produktů existuje velké množství metod a postupů, které může konečný uživatel aplikovat. Neexistuje však pravidlo, které by určilo, jak zvolit vhodnou kombinaci metod a algoritmů k tomu, aby výsledné rozdělení objektů do shluků bylo nejlepší. Počet výsledných skupin objektů často není předem známý, a proto součástí shlukové analýzy bývá stanovení optimálního počtu shluků, do kterého mají být objekty klasifikovány. Cílem tohoto článku je ukázat možnosti stanovení počtu shluků na základě vybraných koeficientů, které jsou dostupné v softwaru a tím pádem využívané z řad analytiků.

## 2. TEORETICKÝ RÁMEC

Shluková analýza je oblíbená vícerozměrná metoda, která se využívá v řadě ekonomických oblastí, mezi které může být zařazena například klasifikace domácností podle typu a vztahu k materiální deprivaci, viz například [17]. Samotný proces shlukování a jeho výsledky jsou velmi intenzivně závislé na volbě proměnných, pomocí kterých jsou jednotlivé objekty charakterizovány. Jak je uvedeno v [16], výsledkem shlukové analýzy není stanovení významných či nevýznamných proměnných, nýbrž vytvoření shluků, na základě vhodně vybraných vlastností objektů. Základní členění tradičních metod shlukování spočívá v rozdělení na *hierarchické* (slouží k vytváření stromovité struktury) a *nehierarchické* metody shlukování, které představují metody rozkladu, viz například [12], [16]. Speciálním případem je tzv. *fuzzy* shlukování, kde zařazení objektu do shluku je dáno tzv. mírou příslušnosti. Ta představuje hodnotu z intervalu od 0 do 1, která vyjadřuje příslušnost, že daný objekt je zařazen do daného shluku. Vyplývá z toho, že objekt může být zařazen do více shluků současně a míra příslušnosti představuje „pravděpodobnost“, že objekt bude klasifikován do daného shluku.

Pro shlukovou analýzu představují klíčovou informaci *míry podobnosti*. Při měření podobností záleží na typu proměnných, které charakterizují jednotlivé objekty. Mezi nejznámější charakteristiky lze zařadit Euklidovu vzdálenost či Mahalanobisovu vzdálenost, které se zejména používají v případě, že jsou objekty charakterizovány pomocí kvantitativních proměnných. Euklidova vzdálenost není vhodná pro případ, kdy jsou jednotlivé proměnné, které charakterizují jednotlivé objekty, velmi silně korelované. Jak je uvedeno například v [4], Mahalanobisova vzdálenost, na rozdíl od Euklidovy míry vzdálenosti, odstraňuje problém, který vzniká při použití nestandardizovaných dat, které mohou způsobit rozdíly mezi shluky, v důsledku odlišností měrných jednotek. Tato míra vzdálenosti je navíc použitelná i tehdy, jestliže jsou jednotlivé proměnné vzájemně závislé.

Jak je uvedeno například v [16] nebo [12], mezi nejznámější a nejpoužívanější metody shlukování lze zařadit například metodu nejbližšího souseda, metodu nejvzdálenějšího souseda, metodu průměrné vzdálenosti, Wardovu metodu, atd.

Tyto metody se liší nejen dobou vzniku, ale také přístupem ke shlukování. Jejich podrobný popis lze nalézt například v [4], [16].

## 3. KOEFICIENTY PRO STANOVENÍ OPTIMÁLNÍHO POČTU SHLUKŮ PŘI HIERARCHICKÉ SHLUKOVÉ ANALÝZE

Součástí shlukové analýzy velmi často bývá stanovení počtu shluků, do kterých mají být dané objekty rozděleny. Ke stanovení optimálního počtu shluků existuje celá řada kritérií a postupů. Použití různých metod shlukování může přinášet rozdílná rozdělení objektů do shluků. Problematice stanovení optimálního počtu shluků pro případ, že jsou shluky charakterizovány kvantitativními proměnnými, je věnována celá řada odborných článků, mezi které patří například [1], [2], [5], [6], [7], [8], [14], atd. Problematice stanovení optimálního počtu shluků v případě jiných proměnných se věnují například v [9] či [10]. V současné odborné literatuře neexistuje jednoznačné pravidlo, které by určilo použití konkrétních koeficientů v různých podmínkách. Různí autoři vytváří či modifikují koeficienty, někdy částečně své a vybrané koeficienty porovnávají s již existujícími. Mezi vybraná kritéria pro stanovení optimálního počtu shluků je možné zařadit Daviesův-Bouldinův (DB) index, Dunnův index, RMSSTD index, CHF index, PTS index.

V této části článku jsou dále popsány výše uvedené vybrané koeficienty pro disjunktní shlukování. Předpokládá se rozdělení množiny *n* objektů do *k* disjunktních shluků, přičemž každý objekt je zařazen právě do jednoho shluku. U koeficientů pro stanovení počtu shluků většinou nezáleží, zda jsou shluky výsledkem metod rozkladu anebo výsledkem hierarchického shlukování, viz [12]. Některé koeficienty jsou však určeny výlučně pro shluky, které vznikly na základě hierarchického shlukování.

Použití **Daviesova-Bouldinova indexu**, jak je uvedeno v [11], nezávisí na vybrané metodě shlukování. Aby bylo možné určit hodnoty Daviesova-Bouldinova indexu, je nejprve nutné definovat tzv. *disperzi h*-tého shluku $S_h$, která se stanoví jako

$$S_h = \sqrt{\frac{\sum_{x_i \in C_h} D^2(\mathbf{x}_i, \bar{\mathbf{x}}_h)}{n_h}},$$

kde význam jednotlivých symbolů je následující: $n_h$ je počet objektů v *h*-tém shluku; $\mathbf{x}_i \in \mathrm{C}_h$ představuje označení, že *i*-tý objekt se nachází v shluku $\mathrm{C}_h$, $\bar{\mathbf{x}}_h$ je centroid *h*-tého shluku, a pro kterou platí následující podmínky

$S_h \geq 0$,

$S_h = 0$, v případě, že jsou objekty ve shluku charakterizovány identickými vlastnostmi.

Pro měření vzdáleností shluků *D* platí

$$D_{hh'} = D(\bar{\mathbf{x}}_h, \bar{\mathbf{x}}_{h'}),$$

$$D_{hg} = D(\bar{\mathbf{x}}_h, \bar{\mathbf{x}}_g),$$

kde $\bar{\mathbf{x}}_h$ je centroid *h*-tého shluku, $\bar{\mathbf{x}}_{h'}$ je centroid *h´*-tého shluku a $\bar{\mathbf{x}}_g$ je centroid *g*-tého shluku.

Nechť **míra podobnosti** mezi *h*-tým a *h´*-tým shlukem se značí $A_{hh´}$ a je založena na disperzích těchto shluků a podle [12] musí splňovat následující podmínky:

1. $A_{hh´} \geq 0$ ,
2. $A_{hh´} = A_{h´h}$ ,
3. $A_{hh´} = 0$ , pokud $S_h = S_{h´}$ ,
4. $A_{hh´} > A_{hg}$ , pokud $S_{h´} = S_g$ a $D_{hh´} < D_{hg}$ ,
5. $A_{hh´} > A_{hg}$ , pokud $S_{h´} > S_g$ a $D_{hh´} = D_{hg}$ ,

kde $S_h$, $S_{h´}$, $S_g$ jsou disperze *h*-tého, *h´*-tého a *g*-tého shluku, $D_{hh´}$, $D_{hg}$ jsou vzdálenosti mezi jednotlivými shluky.

V následujícím kroku se určí míra podobnosti mezi *h*-tým a *h´*-tým shlukem podle vzorce

$$A_{hh´} = \frac{S_h + S_{h´}}{D_{hh´}} .$$

Maximální míra podobnosti mezi shluky *h* a *h´* se dále označí jako $A_h$, tj.

$$A_h = \max_{h, h \neq h´} A_{hh´} .$$

Daviesův-Bouldinův index se nakonec určí jako aritmetický průměr maximálních měr podobností $A_h$, tedy podle vzorce

$$I_{\mathrm{DB}}(k) = \frac{\sum_{h=1}^{k} A_h}{k} ,$$

kde *k* je počet shluků.

Vyhodnocení optimálního počtu shluků se pomocí Daviesova-Bouldinova indexu provádí nalezením minimální hodnoty tohoto indexu, která indikuje kompaktní a dobře separované shluky. Základem je stanovit hodnoty tohoto indexu na předem stanoveném maximálním počtu shluků, tj.

$$I_{\mathrm{DB}}(k^*) = \min_{2 \leq k \leq n-1} I_{\mathrm{DB}}(k) ,$$

kde *k** je optimální počet shluků.

Další možností, jak stanovit počet shluků, je využít **Dunnův index**. Pomocí tohoto indexu je opět možné najít kompaktní a dobře separované shluky, viz [3].

Vzdálenost mezi *h*-tým a *h´*-tým shlukem je definována jako minimální vzdálenost dvou objektů z těchto různých shluků

$$D_{hh´} = \min_{\mathbf{x}_i \in C_h, \mathbf{x}_j \in C_{h´}} D(\mathbf{x}_i, \mathbf{x}_j) .$$

Nechť $M_h$ je definována jako maximální vzdálenost dvou objektů ze stejného ($h$-tého) shluku, tedy

$$M_h = \max_{\mathbf{x}_i, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j) .$$

Dunnův index je následně definován jako

$$I_D(k) = \min_{1 \le h \le k} \left\{ \min_{1 \le h' \le k} \frac{D_{hh'}}{\max_{1 \le h \le k} M_h} \right\} .$$

Při stanovení optimálního počtu shluků se, na rozdíl od Daviesova-Bouldinova indexu, hledá maximální hodnota tohoto indexu v rámci předem stanoveného počtu shluků, který je opět menší než počet objektů, tj.

$$I_D(k^*) = \max_{2 \le k \le n-1} I_D(k) .$$

Vysoké hodnoty tohoto indexu indikují kompaktní a dobře separované shluky.

V situaci, kdy jsou jednotlivé objekty charakterizovány pouze pomocí $t$ kvantitativních proměnných, pro měření variability je možné využít rozptyl. Další koeficienty pro hodnocení výsledků shlukování jsou založeny na rozkladu celkové variability na vnitroshlukovou a mezishlukovou složku variability. Jde o analogii analýzy rozptylu. Do této skupiny patří například **RS index** (též R-kvadrát, RSQ index), který je možné využít pro srovnávání různých postupů shlukování, nejen při hierarchickém shlukování. Tento index je možné využít také k vyjádření kvality shluků, viz [12]. Jeho myšlenka je založena na rozkladu celkového součtu čtverců na vnitroshlukovou a mezishlukovou složku variability.

Nechť jsou označeny následující součty čtverců, viz [12]:
$SS_B$ = součet čtverců mezi shluky (charakteristika mezishlukové variability),
$SS_W$ = součet čtverců uvnitř shluků (charakteristika vnitroshlukové variability),
$SS_T$ = celkový součet čtverců (charakteristika celkové variability).

Jednotlivé součty čtverců se stanoví podle následujících vzorců

$$SS_W = \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} \sum_{l=1}^{t} (x_{it} - \bar{x}_{hl})^2 ,$$

$$SS_T = \sum_{i=1}^{n} \sum_{t=1}^{m} (x_{it} - \bar{x}_t)^2 ,$$

$$SS_B = SS_T - SS_W .$$

RS index je následně definován jako podíl mezishlukového a celkového součtu čtverců, tj. podle vzorce

$$I_{RS} = \frac{SS_B}{SS_T} = \frac{SS_T - SS_W}{SS_T} .$$

Tento koeficient není vhodné používat pro stanovení optimálního počtu shluků, protože s rostoucím počtem shluků (tj. se snižujícím se počtem objektů v nich) se dílčí shluky stávají více homogenní. V důsledku toho dochází k nárůstu mezishlukového součtu čtverců, a tedy k nárůstu hodnoty tohoto indexu. Z tohoto

důvodu je vhodné používat tento index pro srovnání úspěšnosti různých shlukovacích metod. Index nabývá hodnot z intervalu od 0 do 1, přičemž hodnota 0 vyjadřuje, že nejsou žádné rozdíly mezi shluky, hodnota 1 vyjadřuje významný rozdíl mezi shluky, které jsou homogenní.

**RMSSTD index** (*root-mean-square standard deviation index*) je index, který opět využívá rozklad celkového součtu čtverců na dílčí složky. Měří homogenitu nových shluků a jeho výpočet je založen pouze na vnitroshlukové variabilitě, viz [5]. Jeho výpočet je definován podle vzorce

$$I_{\mathrm{RMSSTD}}\ (k) = \sqrt{\frac{SS_{\mathrm{W}}}{t \cdot (n - k)}}\ .$$

Hodnoty tohoto koeficientu je možné využít také ke stanovení optimálního počtu shluků. Nízké hodnoty RMSSTD indexu opět indikují lepší rozdělení objektů do výsledných shluků. V případě, že tento index nabývá vysokých hodnot, jedná se o nehomogenní shluky. Při grafickém vyhodnocení hodnot tohoto indexu pro jednotlivé počty shluků se optimální počet shluků stanoví podle „bodu zlomu“ křivky.

Jak uvádí samotní autoři koeficientů v [6], při vyhodnocení výsledků shlukování pomocí těchto koeficientů je vhodné stanovit hodnoty všech těchto koeficientů současně, aby výsledné hodnocení byly co „nejobjektivnější“.

Další koeficient, který vychází z rozkladu celkového součtu čtverců, je **CHF index** (též pseudo *F* index), který byl navržen autory Calinski a Habarasz, viz [12]. Dále pak byl zkoumán autory Maulik a Bandyopadhyay, viz [14]. CHF index je definován jako podíl průměrné mezishlukové a průměrné vnitroshlukové variability, tj. podle vzorce

$$I_{\mathrm{CHF}}\ (k) = \frac{\frac{SS_{\mathrm{B}}}{k - 1}}{\frac{SS_{\mathrm{W}}}{n - k}} = \frac{(n - k) \cdot SS_{\mathrm{B}}}{(k - 1) \cdot SS_{\mathrm{W}}}\ .$$

Tento koeficient představuje analogii *F*-testu, který se používá v analýze rozptylu. Je možné jej využít pro stanovení optimálního počtu shluků. Vysoké hodnoty tohoto koeficientu indikují dobře separované shluky, tj. při stanovení optimálního počtu shluků se hledá maximální hodnota tohoto indexu v rámci předem stanoveného počtu shluků

$$I_{\mathrm{CHF}}\ (k^*) = \max_{2 \le k \le n-1} I_{\mathrm{CHF}}\ (k)\ .$$

Tento koeficient byl také modifikován, viz [12], pro hodnocení výsledků shlukování v případě kvalitativních proměnných a proměnných různých typů. V těchto případech uvedený koeficient poskytoval nejlepší výsledky v porovnání se známým počtem shluků.

**PTS index** (též pseudo T-kvadrát index) opět využívá myšlenky rozkladu celkového součtu čtverců na jednotlivé složky. Je možné jej využít pro stanovení optimálního počtu shluků, viz [15]. Vychází z vyhodnocení spojení *h*-tého a *h´*-tého shluku. Stanoví se podle vzorce

$$I_{\mathrm{PTS}}(k) = \frac{SS_{\mathrm{B}_{hh'}}}{\frac{SS_{\mathrm{W}_h} + SS_{\mathrm{W}_{h'}}}{n_h + n_{h'} - 2}},$$

kde $SS_{\mathrm{B}_{hh'}}$ je mezishlukový součet čtverců a $SS_{\mathrm{W}_h}$ a $SS_{\mathrm{W}_{h'}}$ jsou vnitroshlukové součty čtverců.

Vyhodnocení se provádí tak, že v případě, že je pro *k* shluků hodnota tohoto indexu větší než pro (*k* – 1) a (*k* + 1) shluků zároveň, optimální počet shluků je (*k* + 1).

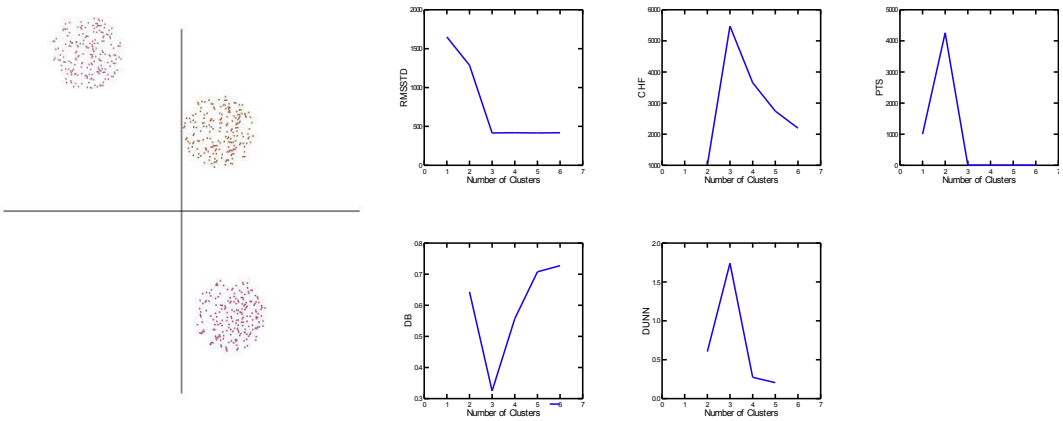
Ke shlukování lze využít celou řadu softwarových produktů. Mezi nejznámější a velmi často používané lze zařadit například systémy IBM SPSS, SAS, STATISTICA, S-PLUS, SYSTAT, STATGRAPHICS, atd. V nich jsou implementovány zejména tradiční metody shlukování, včetně případných metod rozkladu. Některé z nich, jako je třeba systém SAS, neumožňují uživateli volit příslušné kombinace shlukovacích metod a měr vzdálenosti. Systém nabízí kombinace jako dané.

## 4. STANOVENÍ POČTU SHLUKŮ V PRAKTICKÝCH ÚLOHÁCH

V této části článku je představen postup, jak vybrat počet shluků na základě výše popsaných koeficientů bez ohledu na zvolenou metodu shlukování a míru vzdálenosti. V každém z grafů bude v levé části zobrazeno reálné rozdělení objektů do shluků, které se liší barvou. V pravé části grafu je zobrazen grafický výstup pěti výše popsaných koeficientů ze systému SYSTAT pro dané rozdělení objektů do shluků. Není zde hodnocena úspěšnost vybraných metod shlukování v kombinaci s různými měrami vzdáleností, ale pouze schopnost stanovit počet shluků pro danou situaci.

První situace, která je zobrazena na obrázku č. 1, vyjadřuje tři dobře separované shluky. Z obrázku 1 vyplývá, že uvedené koeficienty jsou pro tři dobře separované shluky schopny správně odhalit jejich počet.

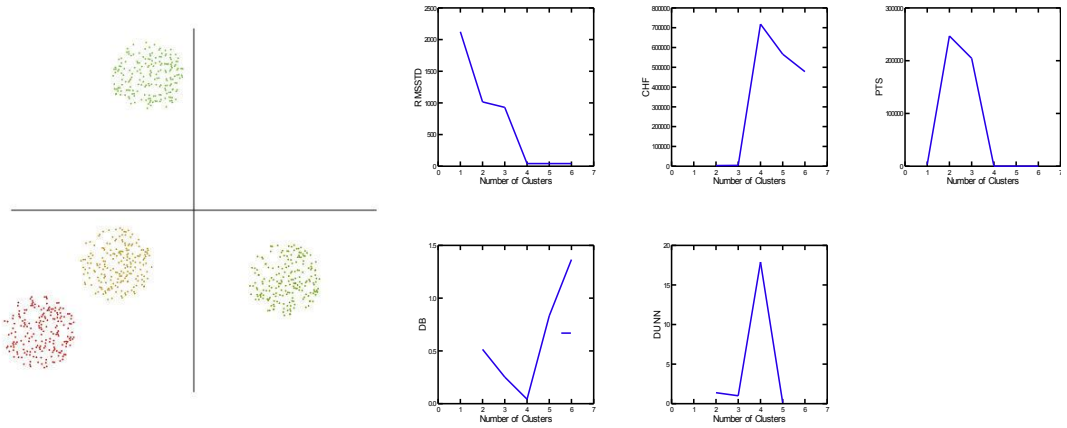
***Obrázek č. 1: Tři dobře separované shluky***




***Zdroj: vlastní zpracování***

Druhá situace, která je zobrazena na obrázku č. 2, zachycuje čtyři dobře separované shluky.

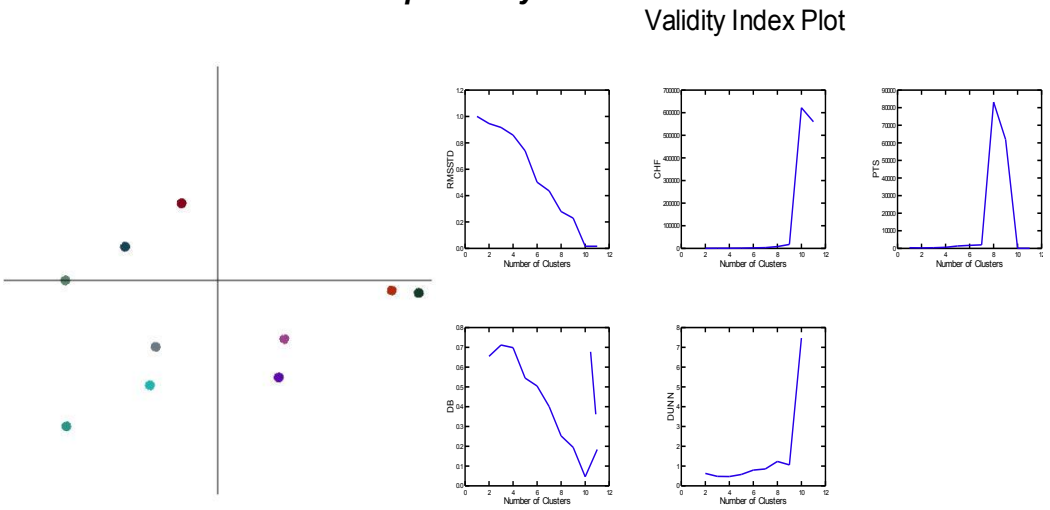
***Obrázek č. 2: Čtyři dobře separované shluky***




***Zdroj: vlastní zpracování***

I v této situaci je zřejmé, že uvedené koeficienty jsou schopny správně stanovit počet shluků. Další situace, která je zobrazena na obrázku č. 3, zachycuje deset dobře separovaných shluků. V případě deseti dobře separovaných shluků je opět zřejmé, že uvedené koeficienty jsou schopny správně odhalit počet shluků.

***Obrázek č. 3: Deset dobře separovaných shluků***

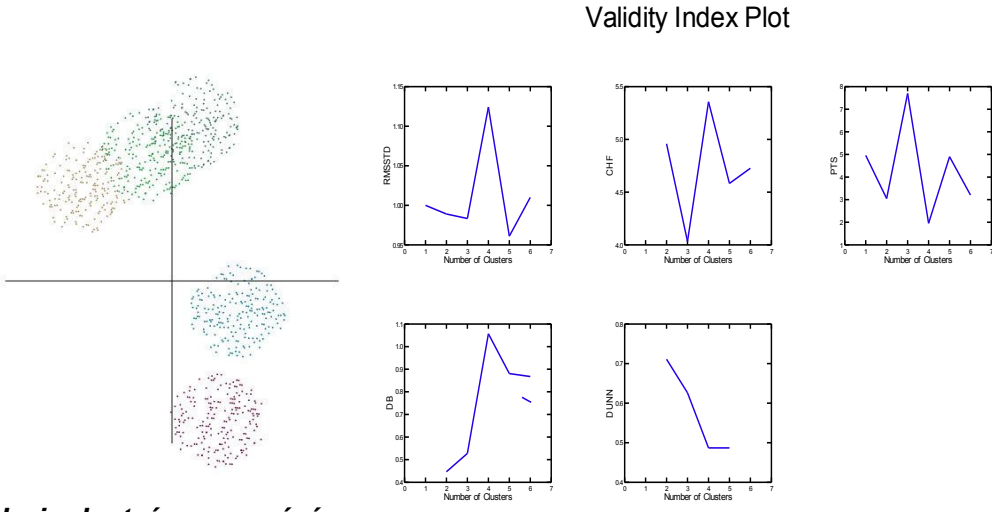



***Zdroj: vlastní zpracování***

Jak vyplývá z těchto příkladů, je možné konstatovat, že v případě dobře separovaných shluků se koeficienty chovají stabilně a počet shluků bývá dobře stanovitelný. Je zřejmé, že počet shluků je správně nalezen nejen v případě malého počtu výsledných shluků, ale i v případě většího počtu shluků.

Poslední popsaná situace, která je zobrazena na obrázku č. 4, zachycuje pět shluků, které se vzájemně překrývají. Prakticky to znamená, že se objekty různých barev (tedy z různých shluků) nachází v jedné oblasti (průnik dvou a více shluků).

***Obrázek č. 4: Pět překrývajících se shluků***




***Zdroj: vlastní zpracování***

V případě pěti překrývajících se shluků je již zřejmé, že uvedené koeficienty se jeví jako nestabilní, poskytují různé hodnoty a je tedy nutné hledat průnik co nejvyššího počtu koeficientů. V praktických úlohách se však může stát, že neexistuje žádný průnik a tak volba počtu shluků závisí na analytikovi.

Jak je z výše uvedených příkladů zřejmé, v případě dobře separovaných shluků koeficienty jejich počet odhalují správně. Úspěšnost koeficientů klesá v případě, že se výsledné shluky překrývají. Z tohoto důvodu je vhodné tuto situaci analyzovat podrobněji. Pro tuto analýzu byly vybrány soubory z databáze *The UCI Machine Learning Repository* (*http://archive.ics.uci.edu/ml/datasets.html*). Analýze byly podrobeny soubory, které jsou určené ke klasifikaci objektů s možností analyzovat všechny proměnné současně, tj. do výběru nevstupovaly soubory, které obsahovaly kvalitativní proměnné. Následnému hodnocení byly podrobeny ty soubory, u nichž je uveden skutečný počet výsledných shluků (zařazení objektů do shluků), aby bylo následně možné uskutečnit srovnání. V případě, že se v některých datových souborech vyskytly chybějící hodnoty u některých proměnných, uvedené objekty byly z dalších analýz vyřazeny. V případě nestejných měrných jednotek byla provedena standardizace pomocí tzv. Z-skórů. Pro účely vyhodnocení postupů pro stanovení optimálního počtu shluků byly vybrány a hodnoceny následující datové soubory, u nichž je uváděno zařazení objektů do shluků. Jedná se o soubory (seřazeny abecedně): *Abalone, Banknote Authentication, Blood Transfusion Service Center, Cardiotocography, Connectionist Bench (Vowel Recognition – Deterding Data), Energy Efficiency, Indian Liver Patient, Ionosphere, Iris, Musk (Version 1), QSAR Biodegradation, Statlog (Vehicle Silhouettes) a+b, Susy, Vertebral Column 2c, Vertebral Column 3c, Wall-Following Robot Navigation Data.* Uvedené soubory se týkají různých oblastí. Jedná se například o bankovky, pacienty, usně, atd. Při analýze každého souboru byly aplikovány následující metody: metoda nejbližšího souseda, nejvzdálenějšího souseda, metoda průměrné vazby, centroidní metoda a Wardova metoda. Každá z uvedených metod byla aplikována společně s Euklidovou vzdáleností a Mahalanobisovou vzdáleností, o které se uvádí, že odstraňuje potenciální problém s vzájemnou závislostí mezi proměnnými, které charakterizují jednotlivé objekty. V daných souborech se běžně vyskytuje překrytí shluků, tj. jeden objekt se nachází současně v prostoru dvou nebo více shluků.

V následujících tabulkách jsou uvedeny hodnoty úspěšností jednotlivých koeficientů, které byly získány srovnáním známého (skutečného) počtu shluků a nalezeného počtu shluků s danou kombinací metody a vzdálenosti pro všechny výše popsané soubory dohromady.

***Tabulka č. 1: Úspěšnost koeficientů při použití Euklidovy vzdálenosti (v %)***

| Metoda/koeficient | RMSSTD | CHF | PTS | D-B | Dunn |
|---|---|---|---|---|---|
| Nejbližšího souseda | 10,53 | 47,37 | 42,11 | 36,84 | 57,89 |
| Nejvzdálenějšího souseda | 26,32 | 15,79 | 21,05 | 63,16 | 42,11 |
| Centroidní metoda | 31,58 | 36,84 | 26,32 | 42,11 | 47,37 |
| Průměrná vzdálenost | 26,32 | 31,58 | 26,32 | 42,11 | 42,11 |
| Wardova metoda | 21,05 | 36,84 | 36,84 | 15,79 | 42,11 |

***Zdroj: vlastní zpracování***

Jak vyplývá z tabulky 1, při užití Euklidovy míry vzdálenosti v případě, že se shluky překrývají, je úspěšnost vybraných koeficientů nižší, než v případě dobře separovaných shluků. Nejlepších výsledků (63, 16 %) bylo dosaženo při použití Daviesova-Bouldinova indexu za současné aplikace s metodou nejvzdálenějšího souseda.

***Tabulka č. 2: Úspěšnost koeficientů při použití Mahalanobisovy vzdálenosti (v %)***

| Metoda/koeficient | RMSSTD | CHF | PTS | D-B | Dunn |
|---|---|---|---|---|---|
| Nejbližšího souseda | 5,26 | 47,37 | 52,63 | 52,63 | 47,37 |
| Nejvzdálenějšího souseda | 21,05 | 26,32 | 31,58 | 36,84 | 36,84 |
| Centroidní metoda | 0,00 | 52,63 | 42,11 | 63,16 | 36,84 |
| Průměrná vzdálenost | 5,26 | 47,37 | 52,63 | 63,16 | 52,63 |
| Wardova metoda | 26,32 | 42,11 | 21,05 | 5,26 | 57,89 |

***Zdroj: vlastní zpracování***

Jak vyplývá z tabulky 2, při užití Mahalanobisovy míry vzdálenosti je úspěšnost vybraných koeficientů opět nižší. Nejlepších výsledků bylo dosaženo opět při použití Daviesova-Bouldinova indexu za současné aplikace s metodou průměrné vzdálenosti či centroidní metody (63,16 %).

V tabulce č. 3 jsou uvedeny rozdíly v úspěšnosti jednotlivých koeficientů při jejich aplikaci s různými metodami shlukování a oběma měrami vzdáleností. Z tabulky vyplývá, že lepších výsledků je ve většině případů dosaženo při aplikaci Mahalanobisovy míry vzdálenosti, zejména při užití centroidní metody a metody průměrné vzdálenosti. Rozdíl v tomto případě činí 21,05 % ve prospěch Mahalanobisovy míry vzdálenosti.

***Tabulka č. 3: Rozdíly v úspěšnostech koeficientů (Mahalanobisova – Euklidova vzdálenost) v %***

| Metoda/koeficient | RMSSTD | CHF | PTS | D-B | Dunn |
|---|---|---|---|---|---|
| Nejbližšího souseda | -5,26 | 0,00 | 10,53 | 15,79 | -10,53 |
| Nejvzdálenějšího souseda | -5,26 | 10,53 | 10,53 | -26,32 | -5,26 |
| Centroidní metoda | -31,58 | 15,79 | 15,79 | 21,05 | -10,53 |
| Průměrná vzdálenost | -21,05 | 15,79 | 26,32 | 21,05 | 10,53 |
| Wardova metoda | 5,26 | 5,26 | -15,79 | -10,53 | 15,79 |

***Zdroj: vlastní zpracování***

## 5. ZÁVĚR

Shluková analýza je vícerozměrná statistická metoda, jejímž cílem je klasifikace objektů do skupin. Ke stanovení optimálního počtu shluků existuje mnoho způsobů (koeficientů), které je možné kombinovat s různými metodami a různými měrami vzdáleností.

Cílem tohoto článku bylo ukázat příklad stanovení počtu shluků u vybraných koeficientů, které jsou aplikovány do oblíbených softwarových produktů. K vyhodnocení schopnosti uvedených koeficientů správně stanovit počet shluků bylo provedeno mnoho analýz, jako například také v [5], [6]. V tomto článku byla podrobně analyzována skutečnost překrývajících se shluků na skutečných datových souborech z databáze *The UCI Machine Learning Repository*. V analýzách, které jsou dále uvedeny v [11] a [13], je provedeno srovnání na mnoho generovaných souborech, aby bylo možné stanovit srovnatelné podmínky a určit schopnost použití těchto koeficientů v různých podmínkách.

Na základě uvedených analýz je možné konstatovat, že u dobře separovaných shluků nemá počet výsledných shluků ani počet proměnných vliv na úspěšnost jednotlivých koeficientů. Lepších výsledků je obecně dosaženo při použití Euklidovy vzdálenosti, viz [13]. Čím je však separace shluků nižší, tím jsou koeficienty pro stanovení počtu shluků méně úspěšné. Jak bylo uvedeno výše, nejvyšší úspěšnost při analýze vybraných reálných datových souborů měl Daviesův-Bouldinův index, u kterého bylo při použití Mahalanobisovy míry vzdálenosti dosaženo vyšší úspěšnosti o 21,05 %. Použitelnost koeficientů pro stanovení optimálního počtu shluků v případě, že se shluky značně překrývají, byla velmi nízká. V takovémto případě je tedy lepších výsledků dosaženo při použití Mahalanobisovy míry vzdálenosti.

***Poděkování***

***Tento článek byl vytvořen za podpory prostředků dlouhodobé institucionální podpory číslo IP400040 Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze.***

## LITERATURA

[1] CANNON, R. L. – DAVE, J. V. – BEZDEK, J. C.: Efficient Implementation of the Fuzzy c-means Clustering Algorithms. IEEE Transactions On Pattern Analysis and Machine Inteligence, 1989, No. 7, p. 773-781.

[2] DAVIES, D. L. – BOULDIN, D. W.: A Cluster Separation Measure. IEEE Transactions On Pattern Analysis and Machine Inteligence, 1979, No. 4, p. 224-227.

[3] DUNN, J.: Well Separated Clusters and Optimal Fuzzy Partitions. In: Journal of Cybernetics, 1974, No. 4, p. 95-104.
[4] GAN, G. – MA, CH. – WU, J.: Data Clustering Theory, Algorithms, and Applications. Philadelphia: ASA, 2007. ISBN: 978-0-898716-23-8.
[5] HALKIDI, M. – Vazirgiannis, M.: Clustering Validity Assessment: Finding the optimal partitioning of a data set. The Proceedings of ICDM. California, 2001, p. 1-9.
[6] HALKIDI, M. – BATISTAKIS, Y. – VAZIRGIANNIS, M.: On Clustering Validation Techniques. Journal of Intelligent Information System, 2001, No. 2-3, p. 107-145.
[7] HALKIDI, M. – Vazirgiannis, M. – BATISTAKIS, I.: Quality scheme assessment in the clustering proces. Proceedings of PKDD, 2000, p. 265-276.
[8] KOVÁCS, F. – LEGÁNY, C. – BABOS, A.: Cluster Validity Measurement Techniques. World Scientific and Engineering Academy and Society (WSEAS), 2006, p. 388-393.
[9] LÖSTER, T. – ŘEZANKOVÁ, H.: Evaluation of Clustering with Categorical and Mixed Type Variables and Cluster Number Determination. ISI 2011. Dublin, p. 1-6.
[10] LÖSTER, T.: Modification of CHF and BIC Coefficients for Evaluation of Clustering with Mixed Type Variables. In: Research Journal of Economics, 2013, No. 2, p. 1-4.
[11] LÖSTER, T.: The Evaluation of CHF coefficient in determining the number of clusters using Euclidean distance measure. The 8th International Days of Statistics and Economics. Praha, 2014, p. 858-869. ISBN 978-80-87990-02-5.
[12] LÖSTER, T.: Metody shlukové analýzy a jejich hodnocení. 1. vyd. Slaný: Melandrium, 2014. 132 s. ISBN 978-80-86175-88-1.
[13] LÖSTER, T.: The Evaluation of CHF coefficient in determining the number of clusters using Mahalanobis distance measure. 14th Conference on Applied Mathematics – Aplimat 2015 Bratislava, 2015, p. 546-554. ISBN 978-80-227-4314-3.
[14] MAULIK, U. – BANDYOPADHYAY, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2002, No. 12, p. 1650-1654.
[15] ŘEZANKOVÁ, H.: Hodnocení kvality shluků, Analýza dat 2008/II, TriloByte Statistical Software, Pardubice, 2009, s. 19–40. ISBN 978-80-904053-1-8.
[16] ŘEZANKOVÁ, H. – HÚSEK, D. – SNÁŠEL, V.: Shluková analýza dat. 2. rozšíř. vyd. Praha: PROFESSIONAL PUBLISHING, 2009. 218 s. ISBN 978-80-86946-81-8.
[17] ŘEZANKOVÁ, H. – ŽELINSKÝ, T.: Faktory míry materiální deprivace v České republice a jejich vztahy k typu domácnosti. In: Ekonomický časopis, 2014, č. 4, s. 394–410. ISSN 0013-3035.
[18] http://archive.ics.uci.edu/ml/datasets.html

**RESUME**

In case of the well-separated clusters, neither the number of the resulting clusters, nor the number of variables affect the efficiency of individual clusters. Better results are generally achieved by using the Euclidean distance. However, the lower the separation of clusters, the less efficient are the coefficients determining the number of clusters. The applicability of coefficients for determining the optimal number of clusters is very low if the clusters are significantly overlapped. In this case, better results are achieved by using Mahalanobis distance.

***PROFESNÍ ŽIVOTOPIS***

*Ing. Tomáš Löster, Ph.D., působí na Fakultě informatiky a statistiky Vysoké školy ekonomické v Praze. Ve vědecko-výzkumné práci se zaměřuje na vícerozměrné statistické metody a statistické výpočetní prostředí. Největší pozornost ze statistických vícerozměrných metod je soustředěna na shlukovou analýzu. V této oblasti uchazeč publikoval řadu článků. Vyučuje předměty z oblasti statistiky a statistických metod. Je autorem či spoluautorem několika monografií, učebnic, skript a mnoha vědeckých článků. Působí v České statistické společnosti jako hospodář.*

**KONTAKT**

tomas.loster@vse.cz

**Viera LABUDOVÁ**
**Katedra štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave**

# ROZHODOVACIE STROMY AKO PREDIKTÍVNA MODELOVACIA TECHNIKA

# DECISION TREES AS A PREDICTIVE MODELING METHOD

**ABSTRAKT**
Rozhodovacie stromy sú silným nástrojom, ktorý sa používa na predikciu a klasifikáciu. Príťažlivosť metód založených na rozhodovacích stromoch je daná skutočnosťou, že rozhodovacie stromy predstavujú pravidlá. Ak sa rozhodovací strom používa na klasifikáciu (cieľová premenná je kategoriálna), nazýva sa klasifikačný strom. Ak sa použije na regresné úlohy (cieľová premenná je spojitá), nazýva sa regresný strom. Článok sa venuje opisu štruktúry rozhodovacích stromov a základným algoritmom na konštrukciu rozhodovacích stromov.

**ABSTRACT**
Decision trees are powerful tools for classification and prediction. The attractiveness of tree-based methods is largely due to the fact that decision trees represent rules. When a decision tree is used for classification tasks (the target variable is categorical), it is referred to as a classification tree. When it is used for regression tasks (the target variable is continuous), it is called a regression tree. This article describes the structure of decision trees and the basic algorithm for their construction.

**KĽÚČOVÉ SLOVÁ**
rozhodovacie stromy, entropia, Giniho index, algoritmy rozhodovacích stromov

**KEY WORDS**
decision trees, entropy, Gini index, decision tree algorithms

## 1. ÚVOD

Modely rozhodovacích stromov založené na niekoľkých vstupných a jednej výstupnej premennej patria do skupiny viacrozmerných štatistických metód. Na základe hodnôt vstupných (vysvetľujúcich) premenných sa odhaduje hodnota spojitej závislej (výstupnej) premennej alebo sa jednotlivé objekty zaraďujú do príslušných skupín zodpovedajúcich kategóriám závislej premennej. Rozhodovacie stromy možno preto považovať za alternatívny prístup k lineárnej regresnej analýze, ak je závislá premenná číselná spojitá, alebo k logistickej regresii a diskriminačnej analýze, ak je závislá premenná kategoriálna [14]. Rozhodovacie stromy možno zaradiť tiež do skupiny hierarchických zhlukovacích metód, keďže ich výstupom sú disjunktné podmnožiny pôvodného súboru objektov [5].

Rozhodovacie stromy sa používajú v situáciách, keď potrebujeme predikovať hodnoty spojitej závislej premennej, alebo v prípadoch, keď predikujeme príslušnosť objektov do vopred zvolených tried.

V literatúre sa môžeme stretnúť s rôznymi definíciami rozhodovacieho stromu.

Rozhodovací strom je štruktúra, ktorá sa využíva na rozdelenie veľkého súboru prípadov v databáze na menšie súbory prípadov pri postupnej aplikácii jednoduchých rozhodovacích pravidiel. Rozhodovací strom pozostáva zo súboru pravidiel (predpisov)[1] na rozdelenie veľkej heterogénnej populácie do menších, homogénnejších skupín s rešpektovaním príslušnej výstupnej premennej [2].

Rozhodovací strom predstavuje reprezentáciu rozhodovacej procedúry na klasifikáciu prípadov do príslušných tried. Je to grafová štruktúra vo forme stromu obsahujúca koreňový uzol, nelistové a listové uzly. Uzly reprezentujú triedu alebo testovací znak. Hrany reprezentujú hodnoty testovacieho znaku [8].

## 2. GENEROVANIE ROZHODOVACIEHO STROMU

Rozhodovacie stromy sa generujú postupom, ktorý sa nazýva TDIDT (*top down induction of decision trees*) – indukcia rozhodovacích stromov zhora nadol. Algoritmus na generovanie rozhodovacieho stromu sa začína na trénovacej množine, ktorá sa nazýva aj priestor prípadov, množina prípadov alebo základný priestor. Ten je tvorený hodnotami vstupných premenných (znakov) $X_1$, $X_2$... $X_k$, ktoré môžu byť číselné (diskrétne, spojité) aj slovné, a hodnotami výstupnej premennej $Y$, zistených na množine prípadov (objektov). Hodnoty výstupnej premennej vytvárajú triedy.

Základný priestor sa v procese generovania rozhodovacieho stromu delí na podpriestory, ktoré sú charakterizované hodnotami testovacích znakov. Delenie sa uskutočňuje rekurzívne, kým nie je splnená tzv. ukončovacia podmienka. Pri tomto postupe sa množina prípadov postupne delí na menšie a menšie podmnožiny (podpriestory), v ktorých prevládajú prípady jednej triedy alebo prípady s podobnou hodnotou znaku.

Tento algoritmus, ktorý sa považuje za všeobecný postup generovania rozhodovacieho stromu zhora nadol, môžeme zapísať takto [8]:

1. Ak je pre každý podpriestor splnené ukončovacie kritérium, generovanie sa ukončí.
2. Inak:
3. Zvolí sa podpriestor obsahujúci prípady klasifikované do viacerých tried.
4. Pre zvolený podpriestor sa vyberie jeden testovací znak, ešte nepoužitý pre daný podpriestor prípadov.
5. Zvolený podpriestor prípadov sa rozdelí na ďalšie podpriestory podľa hodnôt zvoleného testovacieho znaku.

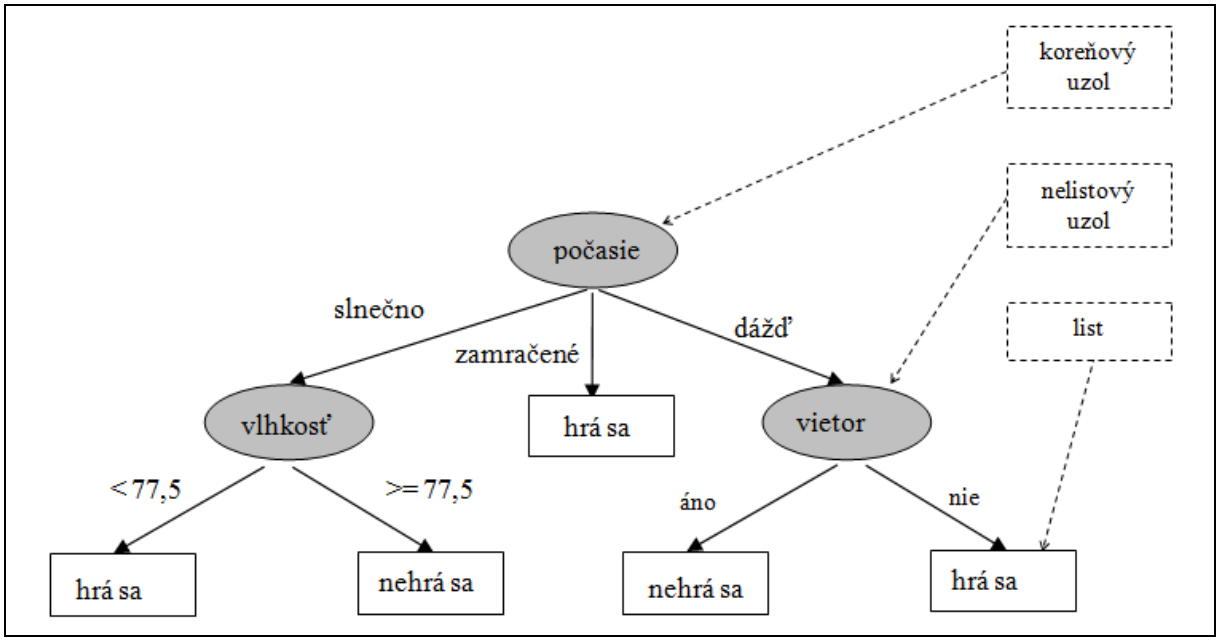
Krok 1 zodpovedá prípadu, keď generovanie rozhodovacieho stromu je už ukončené. Krok 2 pokračuje v generovaní, keď expanduje jeden listový uzol o jednu úroveň.

Uvedený algoritmus okrem pravidla budovania stromu zhora nadol zahŕňa aj pravidlo „rozdeľuj a panuj“, ktoré sa uplatňuje pri indukovaní podstromov, pri ktorom sa úlohy delia na podúlohy. Pri tomto postupe sa množina prípadov postupne delí na menšie a ešte menšie podmnožiny (podpriestory), v ktorých prevládajú prípady jednej triedy.

[1] *Pravidlami sú v tomto prípade vzťahy medzi premennými.*

Rozhodovacie stromy sa zobrazujú najčastejšie v grafickej podobe, pričom existuje veľa možností ich znázornenia. Obrázok 1 ilustruje jednu z možností zobrazenia rozhodovacieho (klasifikačného) stromu.

***Obrázok č. 1: Grafické zobrazenie rozhodovacieho stromu***




***Zdroj: [10], vlastné spracovanie***

Klasifikačný strom na obrázku 1 vznikol na trénovacej množine uvedenej v tabuľke 1. Prípadmi sú dni, počas ktorých sa sledovali atribúty počasia (vysvetľujúce premenné), závislou, modelovanou premennou je premenná hra, ktorá nadobúda dve obmeny: hrá sa, nehrá sa.

***Tabuľka č. 1: Trénovacia množina prípadov***

| Vstupné premenné | | | | Cieľová premenná |
|---|---|---|---|---|
| **počasie** | **teplota** | **vlhkosť** | **vietor** | **hra** |
| Slnečno | 29 | 85 | nie | nehrá sa |
| Slnečno | 27 | 90 | áno | nehrá sa |
| Zamračené | 28 | 78 | nie | hrá sa |
| Dážď | 21 | 96 | nie | hrá sa |
| Dážď | 20 | 80 | nie | hrá sa |
| Dážď | 18 | 70 | áno | nehrá sa |
| Zamračené | 18 | 65 | áno | hrá sa |
| Slnečno | 22 | 95 | nie | nehrá sa |
| Slnečno | 21 | 70 | nie | hrá sa |
| Dážď | 24 | 80 | nie | hrá sa |
| Slnečno | 24 | 70 | áno | hrá sa |
| Zamračené | 22 | 90 | áno | hrá sa |
| Zamračené | 27 | 75 | nie | hrá sa |
| Dážď | 22 | 80 | áno | nehrá sa |

***Zdroj: [10]***

Vytvorený klasifikačný strom je trojúrovňovou[2] hierarchickou štruktúrou, ktorá má osem uzlov. V koreňovom uzle sa nachádza celá množina prípadov (objekty opísané hodnotami vstupných premenných a hodnotami výstupnej premennej). Nelistové (neterminálové, medziľahlé uzly) sú v tomto grafe zobrazené oválom. Tie reprezentujú testovací znak (vetviaci znak). Zhora nadol orientované hrany (vetvy stromu), ktoré vychádzajú z týchto uzlov, zodpovedajú kategóriám testovacích znakov. Podpriestory, ktoré sa už ďalej nedelia, sa nazývajú listy (listové, terminálové uzly), tie sú zobrazené obdĺžnikom. Listy najčastejšie obsahujú informáciu o zaradení objektu do klasifikačnej triedy (v prípade klasifikačných stromov) alebo informáciu o odhadnutej strednej hodnote modelovanej premennej (v prípade regresných stromov). Okrem toho obsahujú aj informáciu o početnostiach príslušných tried závislej premennej.

Keďže sa každý trénovací príklad asociuje s jedným listovým uzlom, stromovú štruktúru možno prepísať do súboru rozhodovacích (klasifikačných, produkčných) pravidiel. Každé klasifikačné pravidlo obsahuje opis jednej cesty od koreňového uzla po niektorý listový uzol. Pravá strana pravidla obsahuje názov triedy zodpovedajúcej listovému uzlu, v ktorom sa cesta končí. Do tejto triedy je zaradený každý trénovací príklad spĺňajúci podmienky ľavej strany pravidla. Ako príklad rozhodovacích pravidiel uvedieme pravidlá, do ktorých je prepísaná stromová štruktúra uvedeného rozhodovacieho stromu (obr. 2).

***Obrázok č. 2: Rozhodovacie pravidlá***

```
Node = 4
if pocasie IS ONE OF: ZAMRAČENÉ  then Tree Node Identifier = 4
Number of Observations = 4  Predicted: hra = nehrá sa  = 0.00  Predicted: hra = hrá sa  = 1.00

Node = 5
if vlhkost < 77.5 AND pocasie IS ONE OF: SLNEČNO or MISSING  then Tree Node Identifier = 5
Number of Observations = 2  Predicted: hra = nehrá sa  = 0.00  Predicted: hra = hrá sa  = 1.00

Node = 6
if vlhkost >= 77.5 or MISSING AND pocasie IS ONE OF: SLNEČNO or MISSING then Tree Node
Identifier  = 6
Number of Observations = 3  Predicted: hra = nehrá sa  = 1.00  Predicted: hra = hrá sa  = 0.0

Node = 7
if vietor IS ONE OF: ÁNO AND pocasie IS ONE OF: DÁŽĎ then  Tree Node Identifier = 7
Number of Observations = 2  Predicted: hra = nehrá sa  = 1.00  Predicted: hra = hrá sa  = 0.00

Node = 8
if vietor IS ONE OF: NIE or MISSING AND pocasie IS ONE OF: DÁŽĎ then  Tree Node
Identifier = 8
Number of Observations = 3  Predicted: hra = nehrá sa  = 0.00  Predicted: hra = hrá sa  = 1.00
```

***Zdroj: vlastné spracovanie, SAS EM***

Ak je výstupná premenná kategoriálna, každý listový uzol predstavuje niektorú z kategórií, tried výstupnej premennej. Vtedy hovoríme o klasifikačných stromoch. Ak je výstupná premenná spojitá, každý list reprezentuje odhadnutú hodnotu výstupnej premennej. V takomto prípade hovoríme o regresných stromoch. V uvedenom

[2] *Úroveň obsahujúca koreňový uzol sa považuje za nultú.*

príklade má modelovaná premenná dve triedy, preto listy reprezentujú jednu z tried znaku hra: hrá sa, nehrá sa.

## 2.1. Kritériá výberu testovacích znakov

### Kritériá výberu testovacích znakov pre klasifikačné stromy

Kritérium na výber premennej, ktorá sa použije na príslušnej úrovni vetvenia, závisí od charakteru výstupnej premennej. Základná idea rastu stromu súvisí s teóriou čistoty údajov. Kritériom výberu vetvenia je zvyšovanie čistoty dcérskych uzlov[3]. Výber testovacieho znaku sa môže uskutočniť rôznymi postupmi. Ak je výstupná premenná kategoriálna, používa sa pri výbere testovacieho znaku Giniho index, entropia, informačný zisk alebo chí-kvadrát test nezávislosti. Ak je výstupná premenná spojitá, jednou z možností je kategorizácia jej hodnôt a použitie niektorej z už spomenutých mier. Pri zachovaní jej pôvodného charakteru sa uplatňuje redukcia rozptylu alebo F-test.

Pri posudzovaní kvality delenia (vetvenia) sa využívajú miery čistoty vzniknutých dcérskych uzlov, ktoré vychádzajú z entropie: informačný zisk (*Information Gain*) a pomerný informačný zisk.

#### *Entropia*

Uvažujme o trénovacej množine *n* prípadov. Každý prípad je opísaný hodnotou vstupného znaku *A*[4] a hodnotou výstupného znaku *Y*. Nech nadobúda vstupný znak hodnoty $a_i$ ($i = 1,2,\ldots k$) a nech má výstupný znak *m* rôznych hodnôt – tried $y_j$, ($j = 1,2,\ldots m$). Pravdepodobnosť výskytu triedy $y_j$, ($j = 1,2,\ldots m$) výstupného znaku *Y* označme $p_j$ ($j = 1,2,\ldots m$).

Entropiu výstupného znaku *Y* vyjadríme takto:

$$H(Y) = -\sum_{j=1}^{m} \left( p_j \log_2 p_j \right), \tag{1}$$

kde $p_j$ je pravdepodobnosť výskytu *j*-tej triedy výstupného znaku *Y*.

Pravdepodobnosť $p_j$ môžeme odhadnúť pomocou relatívnej početnosti $\frac{n_j}{n}$, kde $n_j$ je absolútna početnosť triedy $y_j$, $j = 1,2,\ldots m$ v množine trénovacích prípadov. Vzťah (1) potom upravíme na tvar

$$H(Y) = -\sum_{j=1}^{m} \left( \frac{n_j}{n} \log_2 \frac{n_j}{n} \right). \tag{2}$$

[3] *Za čistý uzol sa považuje taký, ktorý obsahuje len prípady jednej triedy výstupného znaku.*

[4] *A budeme považovať za kategoriálnu premennú. Vstupnými premennými môžu byť aj spojité premenné. Pri tvorbe rozhodovacieho stromu nie je možné vytvárať vetvy pre každú hodnotu premennej, preto dochádza v procese rastu stromu ku kategorizácii hodnôt spojitých premenných.*

Ak má výstupný znak *Y* len dve kategórie, entropia nadobúda minimálnu hodnotu 0 vtedy, ak všetky prípady patria do tej istej triedy. Ak je početnosť obidvoch tried výstupného znaku rovnaká, entropia dosahuje maximálnu hodnotu 1.

Použitím vstupného znaku *A* rozdelíme množinu prípadov do *k* tried (*i*-tá trieda obsahuje všetky prípady s hodnotou $a_i$, $i = 1, 2, \ldots k$ ). Očakávaná entropia znaku *A* je vyjadrená vzťahom

$$H(A) = \sum_{i=1}^{k} p_i H(a_i) = \sum_{i=1}^{k} \frac{n(a_i)}{n} H(a_i), \tag{3}$$

kde $n(a_i)$ je počet prípadov trénovacej množiny, ktoré nadobúdajú hodnotu $a_i$ znaku *A*, $H(a_i)$ je entropia na množine prípadov, ktoré majú hodnotu $a_i$ znaku *A,* a *n* je počet všetkých prípadov tejto množiny. $H(a_i)$ vypočítame takto:

$$H(a_i) = -\sum_{j=1}^{m} \frac{n_j(a_i)}{n(a_i)} log_2 \frac{n_j(a_i)}{n(a_i)}, \tag{4}$$

kde $n_j(a_i)$ je počet prípadov množiny *j*-tej triedy výstupného znaku *Y*, ktoré majú hodnotu $a_i$ znaku *A*. Na vetvenie množiny sa vyberá vstupný znak, pre ktorý je hodnota očakávanej entropie najmenšia.

***Informačný zisk***

Pre entropiu, ktorá je mierou nečistoty, je stanovený informačný zisk $Z(A)$ znaku *A* takto:

$$Z(A) = H(Y) - H(A). \tag{5}$$

Informačný zisk znaku *A* je očakávané zmenšenie entropie zapríčinené rozdelením prípadov na základe kategórií znaku *A*. Informačný zisk na množinách vytvorených vetvením na základe kategórií premennej *A* je definovaný ako rozdiel entropie vyčíslenej na celej množine údajov $H(Y)$ a entropie $H(A)$ na podmnožinách, ktoré dostaneme vetvením uzla (množiny prípadov, ktoré uzol obsahuje) na základe kategórií premennej A. Na vetvenie sa vyberie znak s najvyššou hodnotou informačného zisku.

***Pomerný informačný zisk***

Pomerný informačný zisk na rozdiel od entropie a informačného zisku zohľadňuje počet hodnôt znaku *A*, ktorý sa použil pri vetvení. Pomerný informačný zisk $PZ(A)$ je definovaný ako podiel informačného zisku $Z(A)$ a tzv. vetvenia $V(A)$

$$PZ(A) = \frac{Z(A)}{V(A)}, \tag{6}$$

kde je vetvenie $V(A)$ definované takto:

$$V(A) = -\sum_{i=1}^{k} \frac{n(a_i)}{n} log_2 \frac{n(a_i)}{n} \quad . \tag{7}$$

***Giniho index***

Giniho index má pri výbere premenných a postupnosti ich zaraďovania v procese generovania stromu podobnú funkciu ako entropia. Giniho index je definovaný

$$G = 1 - \sum_{j=1}^{m} p_j^2 \ , \tag{8}$$

kde $p_j$ je pravdepodobnosť výskytu *j*-tej triedy výstupného znaku *Y*.

Podobne ako pri entropii odhadneme pravdepodobnosti pomocou relatívnych početností a vzťah na výpočet Giniho indexu na celej množine prípadov potom upravíme na tvar

$$G(Y) = 1 - \sum_{j=1}^{m} \left( \frac{n_j}{n} \right)^2 \quad . \tag{9}$$

Očakávanú hodnotu Giniho indexu pre znak *A* určíme analogicky ako pri entropii

$$G(A) = \sum_{i=1}^{k} \frac{n(a_i)}{n} G(a_i) \ , \tag{10}$$

kde $G(a_i)$ je Giniho index na množine prípadov, ktoré majú hodnotu $a_i$ znaku *A*

$$G(a_i) = 1 - \sum_{j=1}^{m} \left( \frac{n_j(a_i)}{n(a_i)} \right)^2 \quad . \tag{11}$$

Na vetvenie sa použije znak s najmenšou očakávanou hodnotou Giniho indexu.

Tak ako pri entropii sme definovali informačný zisk, aj pri Giniho indexe môžeme zaviesť podobnú mieru, ktorou je redukcia nečistoty

$$Z_G(A) = G(Y) - G(A) \, . \tag{12}$$

Pri vetvení sa vyberie znak, ktorého použitie pri vetvení vedie k najväčšej redukcii nečistoty v uzle.

Alternatívne možno na vetvenie stromu použiť aj chí-kvadrát test nezávislosti. Na vetvenie na príslušnej úrovni vetvenia sa použije znak, ktorý má najväčšiu asociáciu

s výstupným znakom. Sila asociácie sa porovnáva pomocou *p*-hodnoty testu nezávislosti. Na vetvenie sa vyberie znak, pre ktorý je *p*-hodnota najmenšia.

**Kritériá výberu testovacích znakov pre regresné stromy**

Iné spôsoby výberu znakov na vetvenie sa používajú pri regresných stromoch, ktoré modelujú spojitú výstupnú premennú. Regresné stromy sa používajú na odhad očakávanej hodnoty výstupného znaku. Listové uzly v regresných stromoch obsahujú priemernú hodnotu výstupného znaku pre prípady v danom uzle. V regresných stromoch sa na voľbu znaku na vetvenie množiny prípadov používa redukcia smerodajnej odchýlky výstupného znaku alebo F-test.

***Redukcia smerodajnej odchýlky***

V regresných stromoch možno považovať redukciu smerodajnej odchýlky výstupného znaku za alternatívu k informačnému zisku. Smerodajná odchýlka hodnôt výstupného znaku je na množine *n* prípadov vyjadrená vzťahom

$$s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (i = 1,2,\ldots n)\,. \tag{13}$$

Použitím vstupného znaku *A* rozdelíme množinu *n* prípadov do *k* tried (*i*-tá trieda obsahuje všetky prípady s hodnotou $a_i$, ($i = 1,2,\ldots k$). Očakávaná smerodajná odchýlka na podmnožinách, ktoré vzniknú vetvením množiny prípadov na základe hodnôt $a_i$ znaku *A*, je určená vzťahom

$$s_y(A) = \sum_{i=1}^{k} \frac{n(a_i)}{n} s_y(a_i)\,, \tag{14}$$

kde $n(a_i)$ je počet prípadov množiny, ktoré nadobúdajú hodnotu $a_i$ znaku *A*, $s_y(a_i)$ je smerodajná odchýlka výstupného znaku *Y* na množine prípadov, ktoré majú hodnotu $a_i$ znaku *A*, a *n* je počet prípadov danej množiny.

Namiesto informačného zisku sa na výber znaku používa redukcia smerodajnej odchýlky

$$Zs_y(A) = s_y - s_y(A)\,. \tag{15}$$

Na vetvenie sa vyberie znak, pre ktorý je redukcia smerodajnej odchýlky (15) najväčšia.

***F- test***

Pri tomto teste sa porovnávajú stredné hodnoty $\mu_i$ ($i = 1,2,\ldots k$) výstupného znaku *Y* na podmnožinách, ktoré dostaneme rozdelením množiny prípadov podľa hodnôt vetviaceho znaku *A*. Pri tomto teste sa overuje platnosť nulovej hypotézy

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ pre $k > 2$

oproti alternatívnej hypotéze

$H_1$ : aspoň dve stredné hodnoty sa nerovnajú.

Hodnotu testovacej štatistiky možno za predpokladu platnosti nulovej hypotézy vypočítať podľa vzťahu

$$F = \frac{\frac{\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n(a_i)}{k-1}}{\frac{\sum_{i=1}^{k} \sum_{j} (y_{ij} - \bar{y}_i)}{n-k}}, \qquad (16)$$

kde $\bar{y}_i$ je priemerná hodnota znaku *Y* na množine prípadov, ktoré majú hodnotu $a_i$ znaku *A*, $\bar{y}$ je priemerná hodnota znaku *Y* na množine všetkých prípadov, $n(a_i)$ je počet prípadov, ktoré nadobúdajú hodnotu $a_i$ znaku A, $y_{ij}$ sú hodnoty znaku *Y* na množine prípadov, ktoré majú hodnotu $a_i$ znaku *A*, *n* je početnosť množiny prípadov (uzla), ktorú vetvíme, a *k* je počet hodnôt znaku A.

Príslušná testovacia štatistika má Fisherovo rozdelenie s počtom stupňov voľnosti $v_1 = k - 1$ a $v_2 = n - k$.

Nulovú hypotézu na úrovni významnosti *α* zamietneme vtedy, keď $F \geq F_{1-\alpha}(v_1, v_2)$, kde $F_{1-\alpha}(v_1, v_2)$ je (1–α)100-percentný kvantil Fisherovho rozdelenia so stupňami voľnosti $v_1 = k - 1$ a $v_2 = n - k$.

Na vetvenie vyberieme znak s najväčšou hodnotou *F* štatistiky, respektíve s najmenšou *p*-hodnotou *F*-testu. Keď na príslušnej úrovni vetvenia pre žiadny znak nezamietneme $H_0$, vetvenie skončíme.

Podobne možno na vetvenie použiť aj chí-kvadrát test nezávislosti.

## 3. ALGORITMY GENERUJÚCE ROZHODOVACIE STROMY

V praxi sa využívajú rôzne softvérové balíky, ako napr. Salford Systems CART, IBM SPSS Modeler, Rapid Miner, SAS Enterprise Miner, Matlab, R, Weka atď., v ktorých sú implementované algoritmy na generovanie rozhodovacích stromov.

Algoritmy generujúce rozhodovacie stromy sa líšia aplikovaným kritériom, ktoré sa používa pri výbere vetviacej premennej, druhom vetvenia (binárne, viacnásobné vetvenie), spôsobom narábania s chýbajúcimi hodnotami, charakterom cieľovej premennej a nastaveniami regulujúcimi rast stromu.

Prvým počítačovo implementovaným algoritmom bol algoritmus AID (*Automatic Iteraction Detection*). Vyvinuli ho v roku 1963 John Sonquist a James Morgan [15].

Na modelovanie hodnôt výstupnej premennej využíva binárne vetvenie. Vstupnými premennými môžu byť nominálne aj ordinálne premenné, výstupná premenná je spojitá. Pri generovaní stromu rozdeľuje algoritmus najskôr uzly, v ktorých je súčet štvorcov odchýlok hodnôt výstupnej premennej od priemeru najväčší. Vetvenie sa

ukončí, ak je pokles súčtu štvorcov odchýlok nižší ako hraničná hodnota, ktorá sa rovná súčinu zvolenej konštanty a celkovej sumy štvorcov odchýlok.

Nasledovníkom tohto algoritmu bol klasifikačný algoritmus THAID (*Theta-Automatic Iteraction Detection*), vyvinutý Jamesom N. Morganom a Robertom C. Messengerom v roku 1973 [3].

Jedným z najstarších a súčasne jedným z najrozšírenejších algoritmov v komerčnej oblasti je CHAID (*CHi-squared Automatic Interaction Detection*). Jeho autor V. Gordon Kass [6] zdokonalil predchádzajúce algoritmy AID a THAID.

CHAID sa používa na modelovanie nominálnej výstupnej premennej, pričom využíva nominálne aj ordinálne vysvetľujúce premenné[5]. Na rozdiel od systému AID generuje nebinárny strom, pričom delenie údajov v jednotlivých uzloch je rekurzívne, t. j. každý uzol sa delí podľa rovnaného predpisu. Na delenie pozorovaní a výber vetviacich znakov využíva chí-kvadrát test.

V koreňovom uzle sa pre každý vstupný znak $A_i$ vytvorí kontingenčná tabuľka rozmerov $k$ x $l_i$, kde $k$ je počet hodnôt (kategórií) výstupnej premennej $Y$ a $l_i$ je počet hodnôt (kategórií) vstupnej premennej $A_i$. Na podtabuľkách rozmerov $k$ x 2, ktoré sa vytvárajú pre každú kombináciu dvojíc hodnôt vstupnej premennej $A_i$, sa pomocou chí-kvadrát testu nezávislosti testuje podobnosť týchto dvoch hodnôt, kategórií. Postupne nastáva zhlukovanie tých dvojíc kategórií vstupnej premennej, pre ktoré je výsledok chí-kvadrát testu štatisticky nevýznamný, a to v poradí rastúcej hodnoty chí--kvadrát štatistiky. Po každom zlúčení kategórií sa prepočítava hodnota chí-kvadrát štatistiky vytvorenej tabuľky. Po ukončení zhlukovania sa hľadá najlepšie vetvenie pre kategórie, ktoré vznikli zlúčením aspoň troch pôvodných kategórií vstupnej premennej. Ak je výsledok chí-kvadrát testu štatisticky významný, uskutoční sa dané vetvenie, ak nie je výsledok štatisticky významný, zachová sa táto zlúčená kategória a prejde sa na ďalšiu premennú. Po dokončení optimálneho zlučovania kategórií pre každú vysvetľujúcu premennú sa vyberie najvhodnejšia premenná na vetvenie, a to na základe výsledku chí-kvadrát testu ($p$-hodnoty testu po Bonferroniho korekcii). Začiatkom 90. rokov minulého storočia vytvoril Barry de Ville algoritmus Exhaustive CHAID, ktorý uskutočňuje podrobnejšie prehľadávanie. Výsledkom je strom s väčším počtom vetiev [3].

Algoritmus ID3 (*Iterative Dichotomizer 3*) [11] je klasickým príkladom algoritmu, ktorý buduje rozhodovací strom metódou zhora nadol TDIDT. Strom vytvorený s využitím algoritmu ID3 pracuje ako klasifikátor, pri ktorom je výstupná premenná kategoriálna. Kategoriálnymi sú aj vstupné premenné. Ako kritérium výberu deliacich znakov využíva entropiu. Na každej úrovni vetvenia sa zo všetkých potenciálnych vstupných premenných vyberie tá, ktorej použitím nastane rozštiepenie množiny (materského uzla) na také podmnožiny (dcérske uzly), na ktorých je celková entropia najmenšia. Vetvenie sa končí, ak každý list obsahuje pozorovania patriace do jednej triedy, t. j. pozorovania nadobúdajú rovnakú hodnotu výstupnej premennej (hodnota entropie v každom liste je nulová). Aby sa dosiahol čo najjednoduchší strom, entropia by mala čo najrýchlejšie klesnúť na nulovú hodnotu. Strom generovaný algoritmom ID3 vedie k maximálnemu poklesu entropie lokálne v každom kroku. Algoritmom

[5] *Vstupné spojité premenné sa kategorizujú, pričom sa vytvárajú približne rovnako početné intervaly.*

dokáže vytvoriť stromy s vysokým stupňom generalizácie. Je použiteľný len pri riešení neinkrementálnych úloh[6].

V roku 1986 vytvorili matematici Schlimmer a Fisher algoritmus ID4, ktorý bol inkrementálnou modifikáciou algoritmu ID3. Oveľa známejším je algoritmus ID5R, ktorý bol odvodený priamo z algoritmu ID4.

Algoritmus ID5R (*Inductive Dichotomizer 5 Recursive*) je inkrementálnou modifikáciou algoritmu ID3, pričom sa pri ňom nevyskytujú problémy algoritmu ID4. Využíva sa v situáciách, keď nie sú známe všetky trénovacie prípady naraz, do trénovacej množiny sa pridávajú postupne. Ak by sa v takejto situácii použil niektorý neinkrementálny algoritmus, viedlo by to po príchode každého nového prípadu k zrušeniu už existujúceho stromu a indukcia stromu by musela prebehnúť od začiatku. Pri novej indukcii by sa nevyužili informácie získané v predchádzajúcich krokoch. Pri inkrementálnej indukcii každé pridanie nových prípadov vedie k modifikácii už vytvoreného rozhodovacieho stromu. Rozhodovací strom sa rekurzívne aktualizuje pod aktuálnym uzlom pozdĺž vetvy zodpovedajúcej tej hodnote znaku, ktorá sa vyskytla v novom trénovacom prípade. Algoritmus ID5R využíva pri výbere deliacich znakov entropiu, resp. informačný zisk.

Algoritmus C4.5 pracuje na podobnom princípe ako ID3 [13]. Ako vstupné premenné používa nominálne aj číselné spojité premenné, dokáže pracovať s pozorovaniami, pri ktorých chýbajú hodnoty niektorých premenných. Ide o neinkrementálny algoritmus, ktorý buduje strom zhora nadol. Aj pri tomto algoritme musí byť na vytvorenie perfektného rozhodovacieho stromu splnená podmienka neprotirečivosti trénovacích prípadov. Na výber testovacej podmienky využíva pomerný informačný zisk. Algoritmus C4.5 pracuje iba v textovom režime. V oblasti strojového učenia sa považuje za štandard tvorby rozhodovacích stromov. Jeho najnovšia verzia je implementovaná ako algoritmus C5.0 a jeho unixový duplikát See 5.

## 4. UKONČENIE RASTU STROMU, PREREZÁVANIE ROZHODOVACÍCH STROMOV

Rast, vetvenie stromu sa ukončí, ak je splnená niektorá z nasledujúcich podmienok:

- uzol je čistý, t. j. obsahuje rovnaké hodnoty výstupnej premennej,
- všetky pozorovania v uzle majú rovnaké hodnoty vstupných premenných,
- strom dosiahol používateľom definovanú hĺbku vetvenia,
- počet pozorovaní v rodičovskom uzle je menší ako používateľom definovaný minimálny počet pozorovaní,
- počet pozorovaní v dcérskych uzloch je menší ako používateľom definovaná minimálna hranica,
- redukcia nečistoty uzla, ktorý by sa mal optimálne rozštiepiť, je nižšia, ako ju používateľ definoval.

[6] *Pri inkrementálnych úlohách sa postupne spracúva jeden trénovací prípad za druhým. Po každom prípade použitý algoritmus poskytuje riešenie. Pri neinkrementálnych úlohách sa spracuje naraz celá množina prípadov.*

Pri generovaní rozhodovacích stromov vedie snaha o podrobný opis údajov trénovacej množiny k vytvoreniu stromu, ktorý bezchybne klasifikuje na množine trénovacích prípadov. Takýto strom býva často preučený. Preučenie (*overfitting*) stromu v praxi znamená, že model síce kvalitne vysvetľuje vzťahy na trénovacej množine, tie však nie sú všeobecne platné, preto pri jeho aplikácii na inej množine údajov nastáva vysoká chybovosť. Typickými znakmi preučeného stromu sú jeho prílišná košatosť, tenké vetvy obsahujúce často iba jeden tréningový prípad a málopočetné listové uzly na spodných úrovniach stromu.

Vygenerované stromy sú preto modifikované tzv. orezávaním (*tree-pruning*). Používajú sa dva spôsoby orezávania:

- orezávanie pri konštrukcii (*prepruning*),
- orezávanie po konštrukcii (*postpruning*).

Pri prvom spôsobe orezávania počas rastu stromu sa predčasne pomocou modifikovaného algoritmu ukončí rast niektorých vetiev. Dôvodom ukončenia môže byť napríklad dostatočne vysoká pravdepodobnosť, že údaje príslušnej vetvy patria do tej istej klasifikačnej triedy. V praxi môže byť problémom určenie tejto hranice pravdepodobnosti.

Ďalšou možnosťou, ako zlepšiť predikčnú schopnosť a chybu klasifikácie stromu, je orezanie už vygenerovaného stromu. Pri tomto spôsobe sa vygeneruje úplný strom. Pri prerezávaní, keď sa nelistové uzly nahrádzajú listovými, sa posudzuje, ako sa zhorší jeho klasifikačná schopnosť. Tento spôsob sa v praxi považuje za jednoduchší, hodnovernejší, hoci je časovo náročnejší.

Techniky orezávania sa líšia od seba aj tým, aká množina údajov sa pri raste stromu a jeho spätnom orezaní používa. Techniky orezávania podľa toho rozdeľujeme do dvoch skupín:

- orezávanie používajúce len trénovaciu množinu,
- orezávanie používajúce trénovaciu aj testovaciu, resp. validačnú množinu.

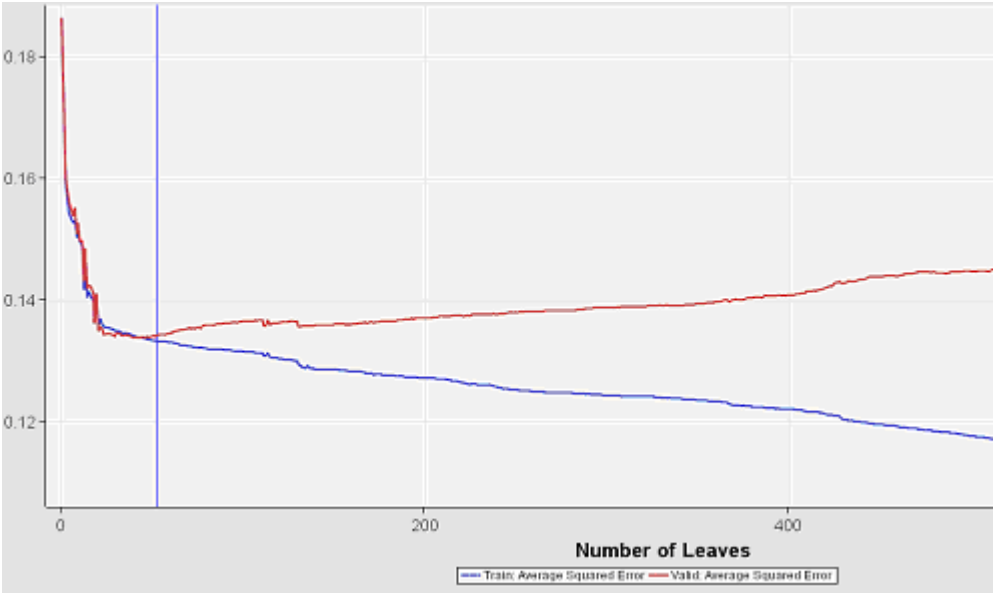
Pri prvej technike sa na tej istej dátovej množine, na ktorej sa nechá strom narásť, robí rozhodnutie o tom, ako ho orezať.

Pri druhej technike sa na jednej (trénovacej) množine strom vygeneruje, druhá množina slúži na výber podstromu z množiny všetkých kandidátskych podstromov, ktorým je pôvodný strom nahradený. Pri výbere sa zohľadňuje miera nesprávnej klasifikácie.[7]

Na obrázku 3 je znázornená závislosť priemernej štvorcovej chyby od počtu listov pri raste stromu. S rastúcim počtom listov priemerná štvorcová chyba na trénovacej množine systematicky klesá, na validačnej množine začne od istého počtu listov rásť. Problém sa rieši orezaním stromu na taký počet listov, aby chyba na trénovacej aj validačnej množine dosiahla minimum (na obrázku je orezanie naznačené zvislou čiarou).

[7] *Postup orezávania podrobnejšie pozri v [2] na s. 184 – 192.*

***Obrázok č. 3: Orezanie rozhodovacieho stromu***




***Zdroj: vlastné spracovanie***

## 5. PRAKTICKÁ UKÁŽKA ROZHODOVACIEHO STROMU

Ukážkou rozhodovacieho stromu, ktorý bol vytvorený na reálnej databáze údajov, je klasifikačný strom vytvorený na základe údajov pochádzajúcich zo štatistického zisťovania EU SILC 2015 (zdroj: ŠÚ SR, EU SILC 2015, UDB 26/09/2016). Použili sme R_súbor (Register osôb), ktorý obsahuje záznam za každú osobu, ktorá v čase zisťovania žila v domácnosti zahrnutej do databázy alebo bola dočasne neprítomná. Tento súbor obsahoval 16 181 prípadov (osôb). Databáza bola rozdelená na trénovaciu množinu (60 % prípadov) a validačnú množinu (40 % prípadov).

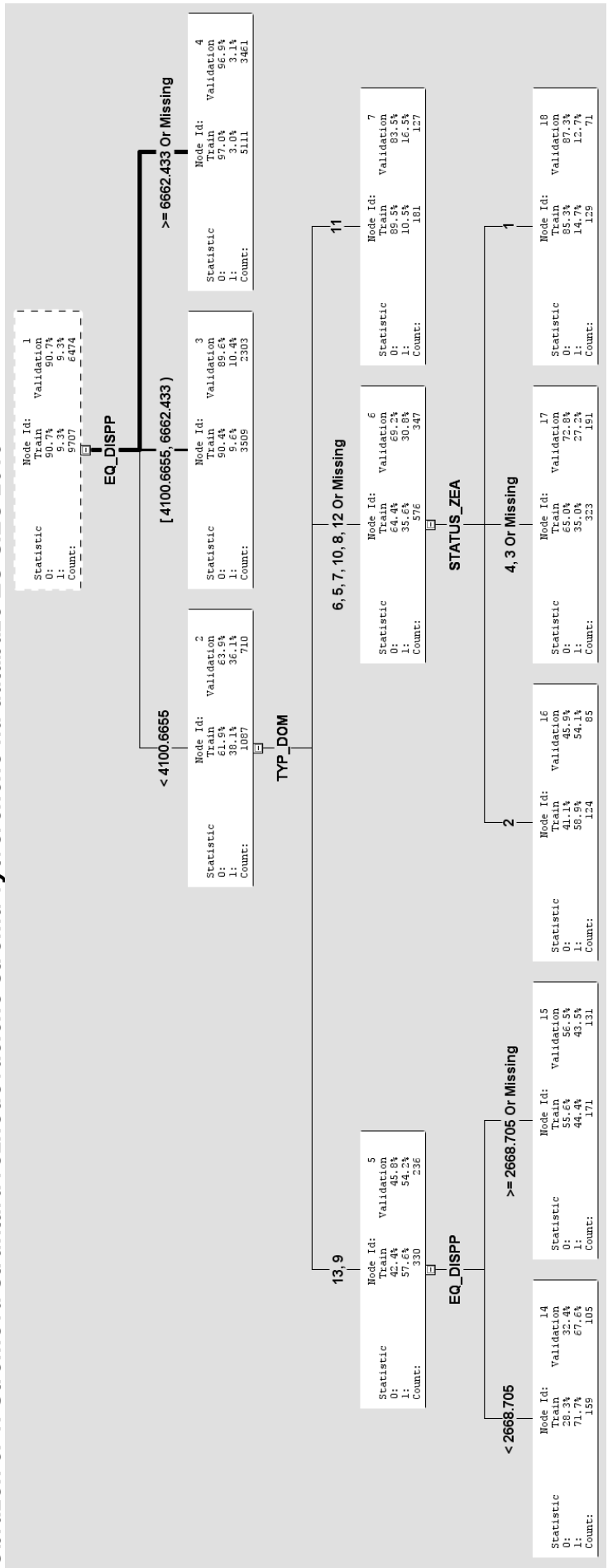
Na modelovanie sme definovali skupinu vstupných (vysvetľujúcich, nezávislých) premenných a modelovanú (vysvetľovanú, závislú) premennú. Nezávislými premennými boli premenné: **RB090**: *pohlavie,* (1 – muž, 2 – žena), **RB210**: *status základnej ekonomickej aktivity* (1 – pracujúci, 2 – nezamestnaný, 3 – starobný dôchodca, osoba v predčasnom dôchodku, 4 – iná neaktívna osoba), **HT** *typ domácnosti* (5 – jednočlenná domácnosť, 6 – domácnosť 2 dospelých bez závislých detí – obaja vo veku pod 65 rokov, 7 – domácnosť 2 dospelých bez závislých detí – aspoň jeden dospelý vo veku 65 rokov a viac, 8 – ostatné domácnosti bez závislých detí, 9 – domácnosť s 1 rodičom a s 1 alebo viac závislými deťmi, 10 – domácnosť 2 dospelých s 1 závislým dieťaťom, 11 – domácnosť 2 dospelých s 2 závislými deťmi, 12 – domácnosť 2 dospelých s 3 alebo viac závislými deťmi, 13 – ostatné domácnosti so závislými deťmi), **EQ_INC20**: ekvivalentný disponibilný príjem domácnosti (ročná suma). Závislou premennou bola premenná **SEV_DEP**: *závažná materiálna deprivácia*[8] (1 – áno, 0 – nie), ktorá vyjadrovala, či je osoba alebo domácnosť ohrozená rizikom chudoby.

[8] *Za závažne deprivované osoby sa považujú osoby, ktoré uvádzajú neprítomnosť alebo vynútený nedostatok aspoň v štyroch z týchto deviatich položiek: čeliť neočakávaným výdavkom, v priebehu jedného roka jeden týždeň dovolenky mimo domova, platiť za nedoplatky (hypotéky alebo nájomné, účty alebo kúpy na splátky), jedlo s mäsom, hydinou alebo rybou každý druhý deň, udržiavať primerane vykurovaný domov. Ďalšie položky vyjadrujú, že domácnosť si nemôže dovoliť (hoci chce): vlastniť práčku; vlastniť farebný televízor; vlastniť telefón; vlastniť osobný automobil. Išlo o predmety dlhodobej spotreby alebo činnosti, ktorých neprítomnosť, resp. nedostatok boli vynútené (ľudia by to chceli vlastniť, ale zdroje im to nedovoľujú).*

Klasifikačný strom sme vytvorili v programe SAS Enterprise Miner™, ktorý využíva vlastnú metodológiu SEMMA. Ako mieru čistoty údajov sme zvolili entropiu, umožnili sme najviac trojnásobné vetvenie, rast stromu sme regulovali tým, že sme povolili maximálne štyri úrovne vetvenia a určili sme minimálny počet prípadov v listoch. Vytvorený klasifikačný strom je na obrázku č. 4.

Najväčší vplyv na závažnú materiálnu depriváciu má výška ekvivalentného disponibilného príjmu, vplyv pohlavia je zanedbateľný. Najvyšší podiel materiálne deprivovaných je v skupine osôb, u ktorých výška ekvivalentného disponibilného príjmu neprekračuje 4 101 eur, pričom v tejto skupine sa podiel materiálne deprivovaných líši v závislosti od typu domácnosti, v ktorej osoba žije. Napríklad v kategórii domácností 2 dospelých s 2 závislými deťmi je 16,5 % deprivovaných (údaje validačnej množiny), v kategórii domácností s 1 rodičom a s 1 alebo viac závislými deťmi a ostatných domácností so závislými deťmi je až 54,2 % materiálne deprivovaných. V prípade, že výška ich ročného ekvivalentného disponibilného príjmu klesne pod 2 668,70 eura, zvýši sa podiel materiálne deprivovaných na 67,6 % (údaje na validačnej množine). Ak by sme rozdelili osoby s výškou ročného ekvivalentného disponibilného príjmu pod hranicou 4 101 eur žijúce v ostatných domácnostiach (*typ domácnosti*: 5, 6, 7, 8, 10 a 12) podľa statusu základnej ekonomickej aktivity, najvyšší podiel materiálne deprivovaných by bol v skupine nezamestnaných (54,1 % na validačnej množine).

***Obrázok č. 4: Stromová štruktúra rozhodovacieho stromu vytvoreného na databáze EU SILC 2015***




Poznámka: EQ_DISPP – ekvivalentný disponibilný príjem domácnosti, TYP_DOM – typ domácnosti, STATUS_ZEA – status základnej ekonomickej aktivity.
***Zdroj údajov: ŠÚ SR, EU SILC 2015, UDB 26/09/2016, vlastné spracovanie (SAS Enterprise Miner™)***

## 6. ZÁVER

V štatistickej praxi sa často stretávame so situáciami, keď je našou úlohou analyzovať dátové súbory vyznačujúce sa vysokou dimenzionalitou, súvisiacou so snahou komplexne opísať zložený jav, pričom premenné, ktoré tento jav opisujú, môžu mať rôzny charakter (nominálne, ordinálne, kardinálne). Rozhodovacie stromy (klasifikačné, regresné stromy) patria k viacrozmerným metódam, ktoré sú schopné analyzovať takéto dátové súbory. Rozhodovacie stromy možno považovať za alternatívny nástroj k regresnej analýze (v situáciách, keď je závislá premenná číselná spojitá); používajú sa ako alternatíva k logistickej regresii a diskriminačnej analýze (ak je modelovaná premenná kategoriálna). Niekedy sa táto metóda zaraďuje do skupiny hierarchických zhlukovacích metód, pretože jej produktom je vytvorenie disjunktných podskupín z pôvodného súboru objektov [5].

V článku sme sa venovali opisu rozhodovacích stromov, metódam výberu premenných (testovacích znakov) na vetvenie a stručnému opisu najčastejšie používaných algoritmov tvorby rozhodovacích stromov.

***Článok vznikol s podporou grantovej agentúry VEGA v rámci projektu VEGA 1/0770/17 Dostupnosť bývania na Slovensku.***

## LITERATÚRA

[1] BERKA, P.: Dobývání znalostí z databází. Praha: Academia, 2003. ISBN 80-200-1062-9.
[2] BERRY, M. J. A. – LINOFF, G. S.: Data mining Techniques. For Marketing, Sales, and Customer Relationship management. Indianapolis: Wiley Publishing, Inc., 2004.
[3] BIGGS, D. – DE VILLE, B. – SUEN, E.: A method of choosing multiway partitions for classification and decision trees. In: Journal of Applied Statistics [online], 1991, No. 1, p. 49-62. Dostupné na internete: <http://dx.doi.org/10.1080/02664769100000005> [prístup k 11. 2. 2017].
[4] BREIMAN, L. – FRIEDMAN, J. H. – OLSHEN, R. A. – STONE, C. J.: Classication and Regression Trees. Wadsworth, 1984. ISBN 9780412048418.
[5] HENDL, J.: Přehled statistických metod: Analýza a metaanalýza dat. Praha: Portál, 2009. ISBN 978-80-7367-482-3.
[6] KASS, G. V.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. In: Applied Statistics, 1980, No. 2, p. 119-127.
[7] KLASCHKA, J. – KOTRČ, E.: Klasifikační a regresní lesy. In: Sborník konference ROBUST [online]. Jednota českých matematiků a fyziků, 2004, s. 177 – 184 Dostupné na internete: <http://statspol.cz/robust/robust2004/klaschka.pdf> [prístup k 15. 2. 2017].
[8] MACHOVÁ, K.: Strojové učenie: Princípy a algoritmy. Košice, 2002.
[9] MORGAN, J. N. – MESSENGER, R. C.: THAID: a sequential program for the analysis of nominal scale depend variables [online]. University of Michigan, 1973. Dostupné na internete: <http://hdl.handle.net/2027/mdp.39015071883859> [prístup k 17. 2. 2017].
[10] PARALIČ, J.: Umelá inteligencia 1: Objavovanie znalostí [online]. Dostupné na internete: <http://people.tuke.sk/jan.paralic/prezentacie/UI/objavovanie_znalosti.pdf> [prístup k 17. 2. 2017].
[11] QUINLAN, J. R.: Induction of Decision Trees. In: Machine Learning [online], 1986, No. 1, p. 81-106. Dostupné na internete: <https://doi.org/10.1007/BF00116251> [prístup k 17. 2. 2017].

[12] QUINLAN, J. R.: Simplifying decision trees. In: International Journal of Man-Machine Studies [online], 1987, No. 3, p. 221-234. Dostupné na internete: <https://doi.org/10.1016/S0020-7373(87)80053-6> [prístup k 17. 2. 2017].
[13] QUINLAN, J. R.: C4.5 Programs for Machine Learning. San Francisco: Morgan Kaufmann, 1993. ISBN 1-55860-238-0.
[14] ŘEZANKOVÁ, H.: Analýza dat z dotazníkových šetření. Praha: Professional Publishing, 2007. ISBN 978-80-86946-49-8.
[15] SONQUIST, J. N. – MORGAN, J. A.: Problems in the Analysis of Survey Data, and a Propsal. In: Journal of the American Statistical Association, 1963, No. 302, p. 415-435.
[16] STANKOVIČOVÁ, I.: Rozhodovacie stromy v marketingových analýzach. In: Nová ekonomika, 2006, č. 1, s. 105 – 111.
[17] TEREK, M. – HORNÍKOVÁ, A. – LABUDOVÁ, V.: Hĺbková analýza údajov. Bratislava: Iura Edition, 2010. ISBN 978-80-8078-336-5.
[18] WILKINSON, L.: Tree Structured Data Analysis: AID, CHAID and CART [online]. Dostupné na internete: http://www.spss.com/research/wilkinson/Publications/c&r trees.pdf [prístup k 17. 2. 2017].

**RESUME**

The tree based learning algorithms are considered to be one of the best and most common supervised learning methods. They are suitable for exploratory data analysis in obtaining information on the impact of the large number of candidate input variables on the target variable. Decision tree models can be effectively used to determine the most important attributes in a dataset. Decision tree is a data mining technique used for the classification and the prediction of values. In data mining, it represents classifications and regression models. When a decision tree is used for classification tasks (the target variable is categorical), it is referred to as a classification tree. When it is used for regression tasks (the target variable is continuous), it is called a regression tree. Decision trees have many appealing properties: understandability, flexibility in handling a variety of input data: nominal, numeric and textual, adaptability in processing datasets with error or missing values, achieving high predictive performance for a relatively small computational effort, being part of various data mining software. Because decision trees combine both data exploration and modeling techniques, they are a powerful first step in the modeling process even in the construction of the final model using some other technique. This article describes the structure of decision trees and the basic algorithm for their construction.

***PROFESIJNÝ ŽIVOTOPIS***

***Doc. RNDr. Viera Labudová, PhD.,*** *je absolventkou Matematicko-fyzikálnej fakulty Univerzity Komenského v Bratislave. Na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave pôsobila od roku 2000 ako odborná asistentka, od roku 2014 vo funkcii docentky v študijnom odbore kvantitatívne metódy v ekonómii. Vo svojej vedeckovýskumnej a pedagogickej činnosti sa venuje aplikácii štatistických metód pri analýzach sociálno-ekonomických javov, analýzam sociálno-patologických javov s osobitným zreteľom na výskyt chudoby, aplikácii metód hĺbkovej analýzy údajov, analýze kategoriálnych údajov a regionálnej štatistike.*

**KONTAKT**

viera.labudova@euba.sk

Informácia/Information

# ŠTATISTICI PRIJALI ZÁKLADNÝ SÚBOR UKAZOVATEĽOV NA MERANIE TRVALO UDRŽATEĽNÉHO ROZVOJA VO SVETE
## Postrehy zo 48. zasadania Štatistickej komisie Organizácie Spojených národov

# STATISTICIANS APPROVED AN INDICATOR FILE FOR MEASURING SUSTAINABLE DEVELOPMENT IN THE WORLD
## Observations from the 48th session of the United Nations Statistical Commission

Štatistická komisia Organizácie Spojených národov (ďalej len „OSN“) vznikla v roku 1947 a zasadá pravidelne v marci. Tohto roku sa konalo v New Yorku už jej 48. zasadanie. Účastníci, medzi ktorými boli aj zástupcovia Štatistického úradu SR, si pripomenuli 70. výročie spolupráce štatistikov z celého sveta. Bohatý oficiálny program tohtoročného zasadania doplnili početné sprievodné podujatia na rôzne odborné témy. Pripravili ich národné štatistické inštitúcie a medzinárodné organizácie s ambíciou osloviť jednotlivé štatistické komunity.

Na 48. zasadaní boli Štatistickej komisii OSN predložené na informáciu a diskusiu správy podľa jednotlivých oblastí štatistiky. Krajiny Európskeho štatistického systému, Európskej únie a Európskeho združenia voľného obchodu už tradične poskytli k vybraným bodom programu spoločné stanovisko. Tento rok sa vyjadrili k údajom a ukazovateľom **Agendy trvalo udržateľného rozvoja 2030** (ďalej „SDGI“), k **Big Datam pre oficiálne štatistiky a k integrácii štatistických a geopriestorových informácií**. Spoločný postup si predbežne odsúhlasili ešte v novembri 2016 na rokovaní Výboru pre európsky štatistický systém. Definitívne sa na ňom dohodli počas koordinačného stretnutia pod vedením Eurostatu tesne pred začatím oficiálneho programu 48. zasadania Štatistickej komisie OSN. Spoločné stanovisko predniesol predstaviteľ Malty ako predsedajúcej krajiny Rady Európskej únie. Zohľadňujú ho aj závery, ktoré štatistická komisia schválila.

V oblasti agendy 2030 sa dohodlo, že jej ukazovatele budú predmetom kontinuálnej kontroly a úprav, ktoré z nej vyplynú. Pripúšťa sa tiež možnosť zaradenia nových ukazovateľov po ich dôkladnej revízii v rokoch 2020 a 2025, pričom sa bude prihliadať na to, aby krajinám nepriniesli záťaž navyše. Rozvíjať sa budú indikátory 3. stupňa, t. j. tie, pre ktoré nie sú v súčasnosti vyvinuté normy a metodika.

K problematike SDGI prijali účastníci 48. zasadania Štatistickej komisie OSN **návrh rezolúcie**, ktorá bude predložená na schválenie Valnému zhromaždeniu OSN. V dokumente sa píše:

Štatistická komisia OSN:
– prijíma rámec SDGI ako dobrovoľný a krajinami vytvorený nástroj. Rámec zahŕňa základný súbor ukazovateľov, ktoré sa budú každoročne spresňovať a komplexne preskúmavať v rokoch 2020 a 2025. Základný súbor ukazovateľov sa doplní o ukazovatele na národnej a regionálnej úrovni,

– požaduje vývoj medzinárodných noriem, metód a usmernení zameraných na úplnú implementáciu SDGI,

– požaduje ďalšie spresňovanie ukazovateľov, vytvorenie metaúdajov k nim a ich periodické revidovanie v súlade s cieľmi trvalo udržateľného rozvoja,

– požaduje, aby generálny sekretariát udržiaval databázu ukazovateľov, informoval raz ročne o dosiahnutom pokroku v tejto oblasti a zabezpečil transparentnosť údajov,

– požaduje, aby naďalej pokračovala spolupráca medzi národnými štatistickými systémami a relevantnými medzinárodnými organizáciami zameraná na harmonizáciu a konzistentnosť údajov,

– zdôrazňuje, že oficiálna štatistika je základom pre SDGI, pričom odporúča využitie nových zdrojov údajov. Národné štatistické inštitúcie plnia dôležitú úlohu koordinátora,

– žiada medzinárodné organizácie, aby využívali údaje národných štatistických inštitúcií; v prípade nedostupnosti týchto údajov budú medzinárodné organizácie svoje odhady pred ich publikovaním konzultovať s príslušnou krajinou,

– víta Globálny akčný plán trvalo udržateľného rozvoja v oblasti štatistiky prijatý na prvom svetovom fóre o dátach v Kapskom meste v januári 2017,

– zdôrazňuje potrebu diskusie o zabezpečení dostatočných kapacít v súvislosti s SDGI na vysokých politických fórach,

– žiada krajiny, medzinárodné a nadnárodné organizácie, špecializované agentúry a ďalšie inštitúcie, aby zintenzívnili podporu zberu údajov a posilnili štatistické kapacity.

**Big Data pre oficiálne štatistiky** boli ďalšou z aktuálnych diskutovaných tém. Krajiny Európskej únie požadovali opatrný a postupný prístup k ustanoveniu svetovej platformy. Účastníci Štatistickej komisie OSN sa dohodli, že svetová pracovná skupina pre túto oblasť spracuje štúdiu o globálnej platforme o dátach, službách a aplikáciách. Štúdia bude vychádzať z partnerstva so spoločnosťami zaoberajúcimi sa technológiami, s poskytovateľmi údajov a s akademickou obcou. Toto partnerstvo sa zameria na formuláciu politického rámca na riadenie údajov a informačný manažment vrátane záležitostí týkajúcich sa dôvery, rešpektovania súkromia, dôvernosti a bezpečnosti údajov. O výsledkoch svojej práce bude pracovná skupina informovať na najbližšom zasadaní štatistickej komisie.

Vo svete existuje viacero skupín, ktoré sa zaoberajú **integráciou štatistických a geopriestorových informácií** s osobitným zameraním na monitorovanie ukazovateľov trvalo udržateľného rozvoja. Aktivitu týchto skupín bude koordinovať expertná skupina OSN (Global Geospation Information Management, GGIM).

Správa zo zasadania oceňuje doterajšiu prácu v jednotlivých oblastiach štatistiky, ktorú vykonali expertné skupiny na rôznych úrovniach. Stanovili sa priority na

nasledujúce obdobie, schválili sa nové a predĺžil sa mandát niektorým doterajším pracovným skupinám.

Jednou zo sprievodných akcií 48. zasadania Štatistickej komisie OSN bol aj seminár o otvorených dátach (Open Data). Usporiadala ho nezisková mimovládna organizácia Open Data Watch, ktorá monitoruje pokrok v tejto oblasti a poskytuje informácie a pomoc pri implementácii otvorených dát.

Otvorené dáta sú v praktickom zmysle slova definované ako údaje strojovo čitateľné, bez vlastníckeho práva, selektovateľné používateľom, doplnené presnými metaúdajmi, ako dáta, ktoré sú vhodné na akékoľvek použitie. ODIN, Open Data Inventory (hodnotí pokrytie a otvorenosť oficiálnej štatistiky s cieľom identifikovať nedostatky, propagovať politiku otvorených dát, zlepšovať prístup k nim a podporovať dialóg medzi používateľmi údajov a národnými štatistickými úradmi), vyhodnotilo stav v 173 krajinách, Slovensko so skóre 52/100 sa umiestnilo na 42. mieste. Diskusia sa zamerala na úlohu otvorených dát pri zostavovaní ukazovateľov Agendy 2030.

Ďalším zaujímavým sprievodným podujatím bolo stretnutie o cieľoch agendy trvalo udržateľného rozvoja, ktoré zorganizovala Organizácia pre ekonomickú spoluprácu a rozvoj (ďalej „OECD“) spolu so Stálym zastúpením Českej republiky pri OSN. Diskutovalo sa na ňom o národných prístupoch k tvorbe a sledovaniu ukazovateľov. Vnútroštátna koordinácia je v jednotlivých krajinách odlišná, napríklad v Česku je celonárodným koordinátorom zodpovedným za ukazovatele Český štatistický úrad, v iných krajinách sú to napr. ministerstvá, prípadne úrady vlády.

Hlavná štatistička OECD Martine Durandová na tematickom stretnutí informovala o pilotnom projekte OECD zameranom na analýzu situácie vo vzťahu k dosiahnutiu cieľov stanovených Agendou 2030. Výsledkom pilotného projektu, na ktorom participovalo Dánsko, Fínsko, Holandsko, Nórsko, Slovinsko a Švédsko, bolo vyhodnotenie štartovacej pozície na ceste k SDGI. OECD bude nápomocná pri spracovaní stratégie na dosiahnutie cieľov v ďalších krajinách vrátane Slovenskej republiky.

49. zasadanie Štatistickej komisie OSN sa bude konať 6. – 9. marca 2018 v sídle OSN v New Yorku.

**Ing. ELENA BENKOVÁ**

*Autorka je riaditeľkou odboru európskych záležitostí a medzinárodnej spolupráce Štatistického úradu SR.*

Názory/Opinions

# MÁLO INŠPIRATÍVNE INŠPIRÁCIE

# LACK OF INSPIRATIONAL INSPIRATION

V čísle 2/2017 Slovenskej štatistiky a demografie som si so záujmom prečítal informáciu s názvom *I. svetové fórum o dátach prinieslo veľa inšpirácie aj pre štátnu štatistiku*. Ako človeka, ktorého celý doterajší profesijný život je spojený s údajmi, ma táto téma výrazne oslovila. Po prečítaní informácie som mal však zmiešané pocity. Predovšetkým preto, že článok na mňa zapôsobil ako spleť množstva informácií bez jednoznačnejšieho inovatívneho odkazu odbornej štatistickej i laickej verejnosti. Vzhľadom na to, že išlo o historicky prvé svetové fórum o údajoch, autor článku podľa môjho názoru dostatočne nevyužil príležitosť na propagáciu významu štatistiky a štatistických údajov v každodennom živote. Najviac som bol sklamaný z toho, že v publikovanej informácii chýbala zmienka o globálnom akčnom pláne pre údaje, ktorý je najdôležitejším odkazom a výzvou svetového fóra pre národné štatistické úrady.

Sprostredkovane cez internet sa možno dozvedieť, že vecná problematika svetového fóra bola rozdelená do týchto šiestich tematických okruhov:

1. Nové prístupy k vytváraniu kapacít na získavanie lepších údajov (*New approaches to capacity development for better data*),
2. Inovácie a synergie medzi údajmi rozdielnych ekosystémov (*Innovations and synergies across different data ecosystems*),
3. Nenechať nikoho mimo štatistiky (*Leaving no one behind*)
4. Chápanie sveta prostredníctvom údajov (*Understanding the world through data*),
5. Otázky princípov zberu, spracovania, zverejňovania a celkového spravovania údajov (*Data principles and governance*),
6. Najbližšie úlohy: globálny akčný plán pre údaje (*The way forward: A Global Action Plan for data*).

Podstata zvýšenej celosvetovej pozornosti štatistickým údajom vychádza zo snahy vytvoriť predpoklady na efektívne naplnenie Agendy 2030 pre udržateľný rozvoj[1]. Táto úloha je objektívne spojená s požiadavkami na zber, spracovanie a zverejňovanie obrovského množstva údajov na miestnej, národnej, regionálnej a globálnej úrovni a týka sa viacerých zainteresovaných strán. Agenda 2030 výslovne požaduje posilnenie budovania kapacít na podporu národných plánov realizácie cieľov trvalo udržateľného rozvoja. Národné štatistické systémy budú čeliť naliehavej potrebe uspokojiť zvyšujúce sa nároky používateľov údajov vrátane úplného zabezpečenia Agendy 2030. To bude vyžadovať aktívnu a cielenú angažovanosť aj pri presviedčaní predstaviteľov výkonnej moci o nevyhnutnosti financovať modernizáciu národných štatistických systémov.

Cieľom spomínaného globálneho akčného plánu je vytvoriť rámec na diskusiu o prioritách plánovania a budovania štatistických kapacít potrebných na dosiahnutie zámerov Agendy 2030. Hlavný dôraz pri jeho rozpracovaní a realizácii sa však kladie

[1] *Tento materiál bol oficiálne schválený Štatistickou komisiou OSN na jej 48. zasadnutí 7. – 10. 3. 2017.*

na štatistické úrady, ktoré by pri tvorbe indikátorov a budovaní kapacít mali zohľadniť aj národné a regionálne špecifiká.

Pred národnými štatistickými systémami stojí výzva urobiť rozhodné opatrenia pri transformácii doterajšieho zberu štatistických údajov a spôsobu ich zverejňovania, ktoré by mali zabezpečiť efektívne využívanie štatistických údajov pri tvorbe rôznych politických rozhodnutí. V tomto procese sa javí ako životne dôležitá podpora národných vlád, ale aj užšia spolupráca medzi zainteresovanými stranami z akademickej obce, občianskej spoločnosti, súkromného sektora a širokej verejnosti.

Globálny akčný plán pre udržateľný rozvoj navrhuje šesť strategických oblastí, z ktorých každá je spojená s určitými cieľmi (v rámci nich sú naznačené aj príslušné konkrétne aktivity).[2] Sú v ňom zakomponované nasledujúce oblasti a ciele:

*1. Koordinácia činností a stratégia využívania údajov pre trvalo udržateľný rozvoj*
- *1.1. Posilnenie národných štatistických systémov a koordinačnej úlohy národných štatistických úradov.*
- *1.2. Posilnenie spolupráce medzi národnými štatistickými systémami, regionálnymi a medzinárodnými organizáciami pôsobiacimi v oblastiach práce s údajmi a vytvárania štatistík pre udržateľný rozvoj.*

*2. Inovácia a modernizácia národných štatistických systémov*
- *2.1. Modernizovanie správy vecí verejných a inštitucionálneho rámca, aby národné štatistické systémy zodpovedali požiadavkám a možnostiam neustáleho vývoja dátových ekosystémov.*
- *2.2. Modernizovanie štatistických noriem, najmä tých, ktorých cieľom je uľahčiť integráciu údajov a automatizáciu ich zdieľania v rôznych fázach procesu tvorby štatistík.*
- *2.3. Uľahčenie používania nových technológií a nových dátových zdrojov v hlavných štatistických aktivitách.*

*3. Posilnenie základných štatistických aktivít a programov s osobitným dôrazom na riešenie potrieb monitoringu Agendy 2030*
- *3.1. Posilnenie a rozšírenie programov zisťovaní v domácnostiach, integrovaných systémov zisťovaní, priemyselných a iných ekonomických zisťovaní, cenzov, ďalších veľmi potrebných štatistík a medzinárodných porovnaní s prihliadnutím na potreby Agendy 2030.*
- *3.2. Zlepšenie kvality národných štatistických registrov a rozšírenie využívania administratívnych zdrojov údajov integrovaných s údajmi z rôznych štatistických zisťovaní a ďalších nových zdrojov údajov na zostavovanie integrovaných sociálnych, ekonomických a environmentálnych štatistík vo väzbe na Agendu 2030.*
- *3.3. Posilnenie a rozšírenie systému národných účtov a systému environmentálnych ekonomických účtov.*
- *3.4. Integrácia geopriestorových údajov do štatistických programov na všetkých úrovniach.*
- *3.5. Posilnenie a rozšírenie zberu, spracovania a zverejňovania údajov o všetkých skupinách obyvateľstva z dôvodu ich komplexného informačného pokrytia.*
- *3.6. Posilnenie a rozšírenie údajov v doménach, ktoré nie sú v súčasnosti dobre vyvinuté v rámci oficiálnych štatistík.*

[2] *Podrobnejšie informácie o akčnom pláne pre údaje pozri na: http://undataforum.org/WorldDataForum/wp-content/uploads/2017/01/Cape-Town-Action-Plan-For-Data-Jan2017.pdf (dostupné k 18. 4. 2017).*

*4. Zverejňovanie a využívanie údajov o udržateľnom rozvoji*

*4.1. Rozvoj a podpora inovačných stratégií s cieľom zabezpečiť riadne šírenie a využívanie údajov na udržateľný rozvoj.*

*5. Spolupráca rôznych záujmových skupín pri poskytovaní údajov na udržateľný rozvoj*

*5.1. Rozvoj a posilnenie spolupráce národných a medzinárodných štatistických systémov s vládami, univerzitami, občianskou spoločnosťou, súkromným sektorom a inými zainteresovanými stranami zapojenými do tvorby a využívania údajov na udržateľný rozvoj.*

*6. Mobilizácia zdrojov a koordinovanie úsilia pri budovaní štatistických kapacít*

*6.1. Uistenie sa, že zdroje údajov na uskutočnenie potrebných programov a akcií sú dostupné v takom rozsahu, ako je uvedené v globálnom akčnom pláne.*

Najväčšími výzvami pre národné štatistické úrady z odporúčaných aktivít by mali byť predovšetkým tie, ktoré sa im doteraz nedarilo úspešne a efektívne napĺňať. Týkajú sa napríklad podpory, resp. revízie platnej legislatívy smerujúcej k posilneniu ich nezávislosti a koordinačnej úlohy pri vytváraní celoštátnych informačných systémov. Potrebné je ďalej zvýšiť koordinačné aktivity národných štatistických úradov pri interoperabilite rôznych rezortných informačných systémov, aby boli schopné uľahčiť a zlepšiť integráciu údajov zautomatizovaním možností ich využívania pre rôzne inštitúcie. V neposlednom rade národné štatistické úrady musia začať zohrávať výraznejšiu úlohu v rámci iniciatívy za otvorené údaje. Očakáva sa napríklad vytváranie príležitostí na zapojenie neštátnych subjektov do financovania štatistických činností prostredníctvom inovatívnych finančných mechanizmov v súlade so základnými princípmi Organizácie Spojených národov platnými pre oficiálne štatistiky.

Cieľom globálneho akčného plánu je urýchliť riešenie určitých nedostatkov v národných štatistikách a aj v koordinácii tvorby národných informačných systémov, ktoré sa zistili v reakcii na pripravovanú Agendu 2030. Posilnenie národných štatistických systémov v tomto kontexte znamená nielen zabezpečiť ich schopnosť plniť potreby Agendy 2030, ale plnením cieľov a aktivít Agendy 2030 celkovo zlepšiť a zefektívniť činnosť národných štatistických systémov. V prípade Štatistického úradu SR je v tomto smere nevyhnutná užšia aktívna spolupráca s Úradom podpredsedu vlády SR pre investície a informatizáciu.

**Ing. MIKULÁŠ CÁR, PhD.**

*Autor je členom Slovenskej štatistickej a demografickej spoločnosti. Aktuálne sa zaoberá makroekonomickými súvislosťami trhu s bývaním.*

Informácia/Information

# POHĽADY NA EKONOMIKU SLOVENSKA 2017

Stručné poznámky zo 17. ročníka konferencie s rovnomenným názvom

# 2017 SLOVAKIA´S ECONOMY AT A GLANCE

Brief notes on the 17th conference with the same title

Už po 17. raz sa koncom apríla konala v Bratislave konferencia *Pohľady na ekonomiku Slovenska 2017*. Usporiadala ju Slovenská štatistická a demografická spoločnosť (ďalej „SŠDS"). Tradíciu konferencií tohto druhu založil ešte v roku 2001 docent Jozef Chajdiak, dlhoročný vedecký tajomník a predseda SŠDS, ktorý 16 rokov stál aj na čele programového a organizačného výboru. Programovému výboru v tomto roku prvýkrát predsedal podpredseda Štatistického úradu SR František Bernadič, ktorý bol aj autorom témy konferencie. Záštitu nad podujatím s názvom *Meranie výkonnosti ekonomiky SR v kontexte nových spoločensko-ekonomických fenoménov (brexit, globalizácia, princíp ekonomického vlastníctva)* prevzal predseda Štatistického úradu SR Alexander Ballek.

Rokovanie otvorila predsedníčka SŠDS *Iveta Stankovičová*. Zaujímavé fakty z histórie pripomenul podpredseda SŠDS *Peter Mach*. Vyzdvihol o. i. fakt, že od roku 2009 má konferencia trvalé miesto na Ekonomickej univerzite v Bratislave. V mene jej rektora pozdravila účastníkov podujatia prorektorka pre vzdelávanie *Zuzana Juhászová.* Potom sa už ujal slova podpredseda Štatistického úradu SR *František Bernadič*, aby zhodnotil minuloročné slovenské predsedníctvo v Rade Európskej únie, ktoré bolo témou predchádzajúceho ročníka konferencie. Výsledky v oblasti štatistiky možno pokladať za pozitívne a veľmi významné.

Program konferencie prebiehal formou prezentácií pozvaných účastníkov v dvoch častiach. Na úvod prvej časti podujatia vystúpila riaditeľka odboru štatistiky zahraničného obchodu sekcie makroekonomických štatistík Štatistického úradu SR *Alžbeta Ridzoňová* s prezentáciou na tému Prečo zahraničné firmy vykazujú slovenský zahraničný obchod? Aj ďalší prezentujúci *Jaroslav Dolinič* zastupoval na podujatí Štatistický úrad SR. Témou jeho vystúpenia bola problematika globalizácie a jej vplyv na meranie výkonnosti ekonomiky v jednotlivých krajinách Európskej únie. Potvrdilo sa, že fenomén globalizácie, ktorého významným dôsledkom je napríklad narastajúci medzinárodný obchod a stále užšia previazanosť národných ekonomík, zamestnáva nielen ekonómov a sociológov, ale i štatistikov. Veľký rozsah a rôznorodá forma medzinárodného obchodu predstavuje výzvu pre jednotlivé štatistické autority, ako správne zaznamenať veľkosť a výkonnosť národných ekonomík. Udalosti posledných mesiacov dôrazne upozornili (nielen) štatistikov, že globalizácia môže významne ovplyvňovať základné makroekonomické ukazovatele a od nich odvodené indikátory.

Ďalší dvaja prezentujúci sa venovali vplyvu brexitu na politiku súdržnosti (*Martin Hulényi* z Úradu vlády SR) a rozpočet Európskej únie (*Viliam Páleník* z Ekonomického ústavu Slovenskej akadémie vied, ďalej „EÚ SAV"). Viliam Páleník sa zameral na možné dôsledky a alternatívne riešenia dosahu brexitu na finančné

zdroje Únie. Predstavil hlavné závery zo správy Montiho skupiny, ktorá sa zoberala prípravou reformy príjmových zdrojov Európskej únie (HLGOR).[1]

Druhá časť konferencie bola zameraná viac na budúcnosť. Zástupcovia troch popredných slovenských prognostických inštitúcií, INFOSTAT-u, Národnej banky Slovenska a EÚ SAV, predstavili makroekonomické prognózy vývoja ekonomiky SR. Závery, ku ktorým dospel kolektív z INFOSTAT-u (*Andrej Hamara*, *Branislav Pristáč*), prezentoval tretí člen autorského tímu *Ján Haluška*. Predstavil výsledky analýzy vývoja makroekonomickej výkonnosti slovenskej ekonomiky v roku 2016 (meranej tvorbou reálneho hrubého domáceho produktu, ďalej „HDP"), ktoré sú východiskom na odhad jej výkonnosti v roku 2017. Konštatoval, že nominálny HDP stúpol o 2,9 %, čo znamená, že úhrnná hladina cien v ekonomike (meraná deflátorom HDP) sa na medziročnej báze znížila o 0,4 %, čo predstavuje hlbší pokles (o 0,2 p. b.) v porovnaní s rokmi 2014 a 2015. Minulý rok agregátny dopyt v slovenskej ekonomike stúpol len o 2,9 %, zatiaľ čo v roku 2015 vzrástol až o 5,9 %. Na spomalení dynamiky jeho rastu sa podieľal domáci aj vonkajší dopyt (objem celkového vývozu), pretože ich medziročné prírastky boli vlani podstatne nižšie (o 0,9 %, resp. o 4,8 %) ako v roku 2015 (o 4,8 %, resp. o 7,0 %). Ján Haluška na záver uviedol, že na základe aktuálneho vývoja predstihových indikátorov u nás i v relevantnom zahraničí možno usudzovať, že v súčasnosti sú negatívne a pozitívne riziká viac-menej vybilancované.

*Ján Beka* z Národnej banky Slovenska konštatoval, že svižné tempo rastu slovenskej ekonomiky sa udržalo aj koncom roka 2016. Hlavným zdrojom rastu bol export a súkromná spotreba. Zrýchlenie rastu ekonomiky v strednodobom horizonte by malo vyplývať predovšetkým z nábehu novej produkcie automobiliek, k čomu sa intenzívnejšie pridá domáci dopyt. Slovenská ekonomika by tak mala vzrásť o 3,2 % v tomto roku s následnou akceleráciou na 4,2 % v roku 2018 a 4,6 % v roku 2019. Priaznivá situácia na trhu práce by mala pokračovať, ekonomika bude generovať nové pracovné miesta a miera nezamestnanosti klesne na konci horizontu predikcie na historicky najnižšiu úroveň 6,9 %. Napätie na trhu práce a rast produktivity práce sa premietnu do rastu nominálnych miezd. Tie by mali podporiť dopytové tlaky a spolu s rastom cien komodít by sa mala inflácia dostať nad 2 % v horizonte predikcie.

*Ivan Lichner* z EÚ SAV predstavil strednodobú makroekonomická prognózu vývoja ekonomiky SR, ktorú vypracoval v spolupráci s *Marekom Radvanským*. Predložená prognóza na najbližšie štyri roky (2017 – 2020) je založená na ekonometrickom modeli IER_ECM_16q4. Prezentovaná verzia modelu predstavuje najnovšiu aktualizáciu makroekonomického modelu postupne vyvíjaného v EÚ SAV, ktorý obsahuje aj člen korigujúci chyby. Na základe tejto prognózy sa v roku 2017 očakáva oživenie rastu HDP oproti minulému roku na úrovni 3,7 % a v roku 2018 rast na úrovni 3,8 %. Tento rast by mohol prispieť k tvorbe nových pracovných miest a na trhu práce by malo i naďalej dochádzať k pozitívnemu vývoju, ktorý však bude sprevádzaný problémami s nedostatkom kvalifikovanej pracovnej sily a tlakom na rast miezd.

[1] *Skupina na vysokej úrovni pre vlastné zdroje (HLGOR – The high-level group on own resources) vznikla vo februári 2014 s cieľom preskúmať, ako sa dá príjmová strana rozpočtu EÚ zjednodušiť, urobiť transparentnejšou, spravodlivejšou a demokraticky zodpovednou. Vedúcou osobnosťou expertnej skupiny je bývalý predseda talianskej vlády a eurokomisár Mário Monti.*

Výhľadové prognózy sú v súčasnom období relatívne výrazne závislé od vývoja a stability vonkajšieho dopytu, ako aj dodržiavania cieľov hospodárskej politiky. V porovnaní s predchádzajúcou prognózou vznikli nové riziká, ako brexit a zmena hospodárskej politiky USA, ktoré môžu zásadne ovplyvniť ekonomický vývoj u nás v nasledujúcich rokoch.

*Prednášatelia konferencie počas záverečnej diskusie (zľava J. Dolinič zo Štatistického úradu SR, I. Lichner z Ekonomického ústavu SAV, F. Bernadič, podpredseda Štatistického úradu SR, J. Haluška z INFOSTAT-u, J. Beka z Národnej banky Slovenska, V. Páleník z Ekonomického ústavu SAV a M. Hulényi z Úradu vlády SR).*

Na záver podujatia sa uskutočnila diskusia pod odborným vedením podpredsedu Štatistického úradu SR Františka Bernadiča. Rezonovali v nej otázky brexitu, globalizácie a ich vplyv na meranie výkonnosti našej ekonomiky. Poslucháčov zaujal príklad Írska. Zavedením novej metodiky na zostavovanie národných účtov ESA 2010 (platnej od roku 2014) namiesto metodiky ESA 95 nastali významné zmeny vo vykazovaných objemoch HDP v krajinách Európskej únie. Bolo to spôsobené hlavne zmenami vo vykazovaní zahraničného obchodu, v princípe vlastníctva a rezidencie ekonomických jednotiek. Zavedením tejto zmeny v metodike bolo potrebné prepočítať vykazované objemy HDP v jednotlivých krajinách a to malo výrazný prorastový efekt. Tento efekt sa výrazne prejavil hlavne na ukazovateľoch HDP Írska, kde zmena údajov znamenala medziročný nárast HDP o 26,3 % a nárast hrubého národného dôchodku o 18,7 % (obidva ukazovatele v stálych cenách). Táto nezvyčajne veľká revízia bola predmetom skúmania Eurostatu, ktorý ju predbežne považuje za opodstatnenú v súlade s metodikou národných účtov.

Konferencia *Pohľady na ekonomiku Slovenska 2017* poskytla priestor na konštruktívnu odbornú diskusiu o aktuálnych makroekonomických prognózach vývoja ekonomiky SR a problémoch s meraním jej výkonnosti vo svetle súčasných

fenoménov. Podujatie bolo príležitosťou na vecnú výmenu názorov a osvedčilo sa aj ako platforma na nadviazanie kontaktov, ktoré sú prísľubom novej spolupráce.

**Doc. Ing. IVETA STANKOVIČOVÁ, PhD.**

*Autorka je predsedníčkou Slovenskej štatistickej a demografickej spoločnosti.*

*Konferencia Pohľady na ekonomiku Slovenska 2017 sa od roku 2009 pravidelne koná v priestoroch Ekonomickej univerzity v Bratislave. Priaznivé prostredie na výmenu poznatkov a nadviazanie nových kontaktov tu našli aj účastníci 17. ročníka podujatia.*

Recenzia publikácie/Review of Publication

Šprocha, B. – Vaňo, B. – Jurčová, D. – Pilinská, V. – Mészáros, J. – Bleha. B.:
**DEMOGRAFICKÝ OBRAZ NAJVÄČŠÍCH MIEST SLOVENSKA**
**THE DEMOGRAPHIC PICTURE OF THE LARGEST CITIES IN SLOVAKIA**
Bratislava: INFOSTAT, 2016. 100 s.
ISBN 978-80-89398-33-1


Publikáciu s názvom *Demografický obraz najväčších miest Slovenska* vydalo Výskumné demografické centrum INFOSTAT-u, Inštitútu informatiky a štatistiky. Ide o prvé vydanie tejto publikácie, ktoré v náklade 150 kusov vyšlo s prispením Agentúry na podporu výskumu a vývoja Slovenskej republiky (v rámci riešenia projektu APVV-0018-12 a ako čiastkový výstup z projektu VEGA č. 1/0745/16).

Autorský kolektív zložený z pracovníkov Výskumného demografického centra a zástupcu Prírodovedeckej fakulty Univerzity Komenského v Bratislave pod vedením editora RNDr. Branislava Šprochu, PhD., prezentuje čitateľom výsledky analýzy demografického vývoja v jedenástich populačne najväčších mestách Slovenskej republiky – v Bratislave, Košiciach, Prešove, Banskej Bystrici, Nitre, Žiline, Trnave, Trenčíne, Martine, Poprade a Prievidzi. Publikácia sa skladá z 9 hlavných kapitol a 7 podkapitol. Jednotlivé kapitoly obsahujú podrobné hodnotenie populačného vývoja v každom zo spomínaných miest v období rokov 1992 až 2015 a jeho porovnanie s celorepublikovým vývojom.

Autori vypracovali veľmi podrobnú analýzu všetkých hlavných (pôrodnosť, úmrtnosť, migrácia), ako aj vedľajších (sobášnosť a rozvodovosť, potratovosť) demografických procesov, ktorú doplnili o hodnotenie základných demografických štruktúr, veku a pohlavia. Treba zdôrazniť, že zistené skutočnosti sú metodicky založené na postupoch, ktoré očisťujú priebeh jednotlivých demografických procesov od rozdielnej vekovej štruktúry. Tento transverzálny pohľad autori vždy dopĺňajú triedením jednotlivých účastníkov demografických udalostí do rôznych kategórií, z ktorých možno upozorniť na najdôležitejšie: vek pri sobáši a pôrode, rozvodovosť podľa dĺžky trvania manželstva, úroveň bezdetnosti u žiadateliek o interrupciu či iné. Pridanú hodnotu práce predstavuje aj aplikovanie viacerých menej používaných postupov, ako napríklad analýzy transformácie plodnosti v kontexte modelu odkladania a rekuperácie. Za originálne možno označiť hodnotenie jednotlivých príčin úmrtí, ktoré v takomto časovom rozsahu a geografickej mierke nie je celkom bežné. Publikácia okrem analytickej časti venovanej konkrétnym demografickým procesom obsahuje aj syntézu tých častí, ktoré sa zaoberajú celkovým prírastkom, jeho zložkami a výsledným počtom obyvateľov. Záverečné časti publikácie autori orientovali na analýzu vekovej štruktúry, kde sú popri intenzite starnutia naznačené aj

budúce scenáre demografického vývoja a možné dôsledky prebiehajúcich reprodukčných zmien v príslušných mestách.

Pre akademické prostredie je publikácia zaujímavá najmä z hľadiska použitých progresívnych metodických postupov a hĺbkou prezentovanej analýzy. Dátový obsah publikácie a naznačenie vývoja v 11 sledovaných mestách predstavuje cenný materiál pre ich decíznu sféru. V neposlednom rade je publikácia svojou zrozumiteľnosťou vhodná aj pre študenta či bežného čitateľa so záujmom o demografiu.

**Mgr. PAVOL ĎURČEK, PhD.**

*Autor je vedeckým pracovníkom na Katedre humánnej geografie a demografie Prírodovedeckej fakulty Univerzity Komenského v Bratislave. Venuje sa výskumu geografickej a štatistickej diferenciácie demografických procesov a demografických štruktúr.*

PRIPRAVUJEME/COMING SOON

Boris VAŇO
**ŽENY PODĽA POČTU ŽIVONARODENÝCH DETÍ V OKRESOCH SR**
WOMEN BY NUMBER OF LIVE BIRTHS IN DISTRICTS OF SLOVAKIA

Branislav ŠPROCHA
**ZMENY V KOHORTNEJ PLODNOSTI ŽIEN SLOVENSKA V SPOJITOSTI S NAJVYŠŠÍM DOSIAHNUTÝM VZDELANÍM**
CHANGES IN COHORT FERTILITY OF WOMEN IN SLOVAKIA IN CONNECTION TO THE EDUCATIONAL ATTAINMENT

* * *

***ONLINE VERZIA KOMPLETNÉHO ČÍSLA 3/2017 SLOVENSKEJ ŠTATISTIKY A DEMOGRAFIE BUDE VEREJNE DOSTUPNÁ*** *na internetovej stránke Štatistického úradu SR www.statistics.sk* ***15. OKTÓBRA 2017****.*

***THE FULL ONLINE VERSION OF THE JOURNAL SLOVAK STATISTICS AND DEMOGRAPHY No 3 (2017) WILL PUBLICLY BE AVAILABLE*** *at the website of the Statistical Office of the SR www.statistics.sk* ***ON OCTOBER 15, 2017****.*

## INFORMÁCIE PRE PRISPIEVATEĽOV

Príspevky prijímame v slovenskom, v českom a v anglickom jazyku. Musia rešpektovať odborné zameranie časopisu a jeho vedecký charakter. Zaslaný príspevok nesmie byť v recenznom konaní v inom časopise, ani uverejnený v odbornej a inej tlači.

Príspevky zasielajte v elektronickej forme vo formáte MS Word alebo Open Office, typ písma Arial, veľkosť 12, riadkovanie 1. Nad titulkom treba uviesť meno autora a jeho pracovisko.

Súčasťou príspevku je abstrakt (základný popis cieľa a spôsobu spracovania faktov v rozsahu do 100 slov), kľúčové slová (maximálne 5), resumé (stručné zhrnutie obsahu článku s dôrazom na jeho prínos a najvýznamnejšie závery v rozsahu do 500 slov), profesijný životopis (v rozsahu do 120 slov) a kontakt (e-mailová adresa autora). Názov článku, abstrakt, kľúčové slová a resumé poskytne autor aj v anglickom jazyku. Zoznam použitej literatúry v abecednom poradí s úplnými bibliografickými údajmi sa uvádza na konci článku. Odkazy na literatúru sa uvádzajú v texte číslami v hranatých zátvorkách. Poznámky s poradovým číslom sú umiestnené pod čiarou na príslušnej strane textu, ku ktorému sa vzťahujú. Podrobnejšie pokyny nájdu autori na www.statistics.sk.

Maximálny rozsah vedeckých článkov je 15 normostrán, informatívnych článkov 6 normostrán, recenzie, rozhovory a informácie publikujeme v rozsahu maximálne 3 normostrany. Tabuľky, mapy, grafy a obrázky musia mať názov a uvedený zdroj údajov; odporúčame, aby kopírovali šírku textu. Skratky sa používajú len minimálne, pri prvom použití je potrebné skratku v zátvorke rozpísať. Redakcia zabezpečuje jazykovú úpravu textu.

Príspevky sú recenzované. Oponentské konanie je obojstranne anonymné. Konečné rozhodnutie o publikovaní článku vydáva redakčná rada.

Redakcia si vyhradzuje právo zverejniť články schválené redakčnou radou v tlačenej podobe a s odstupom troch mesiacov aj v elektronickej forme na internetovej stránke Štatistického úradu SR.

## INFORMATION FOR AUTHORS

Articles are accepted in Slovak, Czech and English languages and must comply with the journal´s professional specialisation and scientific nature as well. The submitted articles should not be peer-reviewed by another journal and should not have already been published in any specialised or other press.

Please submit your articles in electronic form, in MS Word or Open Office format, Arial font, size 12 and typed in single spacing. The author´s name and workplace should be indicated above the heading.

Articles should contain an abstract (general description of the objective and the processing methods used up to 100 words), key words (max. 5), resume (brief summary of the article´s content emphasizing its contribution and the most important conclusions up to 500 words), curriculum vitae of the author (no more than 120 words) and the author´s contact (e-mail address). The author should submit the article´s title, abstract, key words and resume in English language. List of the literature used with full bibliographic data should be given in alphabetical order at the end of an article. Bibliographic citations should be given in square brackets. References are indicated by numbers in a text in square brackets. Footnotes should be numbered in the order of the corresponding page of a text. Authors can find more details at the website www.statistics.sk.

Maximum scope of a scientific article is up to 15 standard pages, informative articles should be up to 6 standard pages in length, reviews, discussions and information not more than 3 standard pages. Tables, maps, graphs and pictures should have a title and the data source indicated, it is also advised to copy the width of a text. Abbreviations should be used only rarely and should be appropriately ex-plained in parentheses when first used. Language text revisions are provided by the editorial office.

Articles are reviewed. The opponent procedure is mutually anonymous. The final decision on the article´s publication is made by the editorial board.

The editorial office reserves the right to publish articles approved by the editorial board in printed form at intervals of at least three months also in electronic form at the website of the Statistical Office of the SR.

## SLOVENSKÁ ŠTATISTIKA A DEMOGRAFIA

je jediný recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov. Propagujeme miesto a význam slovenskej štatistiky v Európskom štatistickom systéme, spoluprácu Eurostatu a národných štatistických úradov pri harmonizácii zisťovaní a multidimenzionálny rozmer štatistiky. Podporujeme rozvoj štatistickej teórie a jej prepojenie s praxou. Naším cieľom je prispievať k využiteľnosti štatistických výstupov v rôznych oblastiach a k zvyšovaniu ich kvality a efektivity.

Publikujeme analytické články, prognózy, názory, diskusné príspevky, recenzie, rozhovory, informácie a oznamy z rôznych oblastí štatistiky (národné účty, produkčné štatistiky, sociálne štatistiky, štatistika životného prostredia a pod.) a demografie (demografická štatistika, teoreticko-metodologické východiská demografie, historická demografia a pod.), vrátane sčítania obyvateľov, domov a bytov ako neodmysliteľnej súčasti demografickej štatistiky.

**Vydáva:**
Štatistický úrad SR

**Identifikačné číslo vydavateľa:**
IČO 00166197

**Vychádza:**
Štyrikrát ročne

**Dátum vydania:**
15. júl 2017

**Tlač:**
Reprografické stredisko
Štatistického úradu SR

**Predplatné:**
20 eur (na rok)
5 eur (za jeden výtlačok)

**Objednávky prijíma:**
Informačný servis
Štatistického úradu SR
Tel.: +4212/502 36 339
+4212/502 36 335
E-mail: info@statistics.sk

Evidenčné číslo/Evidence number 272/08

## SLOVAK STATISTICS AND DEMOGRAPHY

is the only scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures. Our aim is to promote the position and importance of Slovak statistics in the European statistical system, cooperation between the Eurostat and the national statistical offices in the field of survey harmonisation and the multidimensional character of statistics as well. We support the development of statistical theory and its connection with practice. We aim to contribute to the utility of statistical outputs in various fields and to the improvement of quality and efficiency.

We publish analytic articles, prognoses, views, discussion contributions, reviews, discussions, information and announcements from various statistical fields (national accounts, production statistics, social statistics, environmental statistics etc.) and demography (demographic statistics, theoretical and methodological bases of demography, historical demography etc.) including the population and housing census as an essential part of demographic statistics.

**Issued by:**
Statistical Office of the SR

**Company registration number:**
00166197

**Published:**
Four times a year

**Date of issue:**
15th July 2017

**Press:**
Reprographic centre of the
Statistical Office of the SR

**Subscription:**
20 Eur (per year)
5 Eur (for one copy)

**Orders are to be addressed to:**
Information Service of the
Statistical Office of the SR
Tel.: +4212/502 36 339
+4212/502 36 335
E-mail: info@statistics.sk