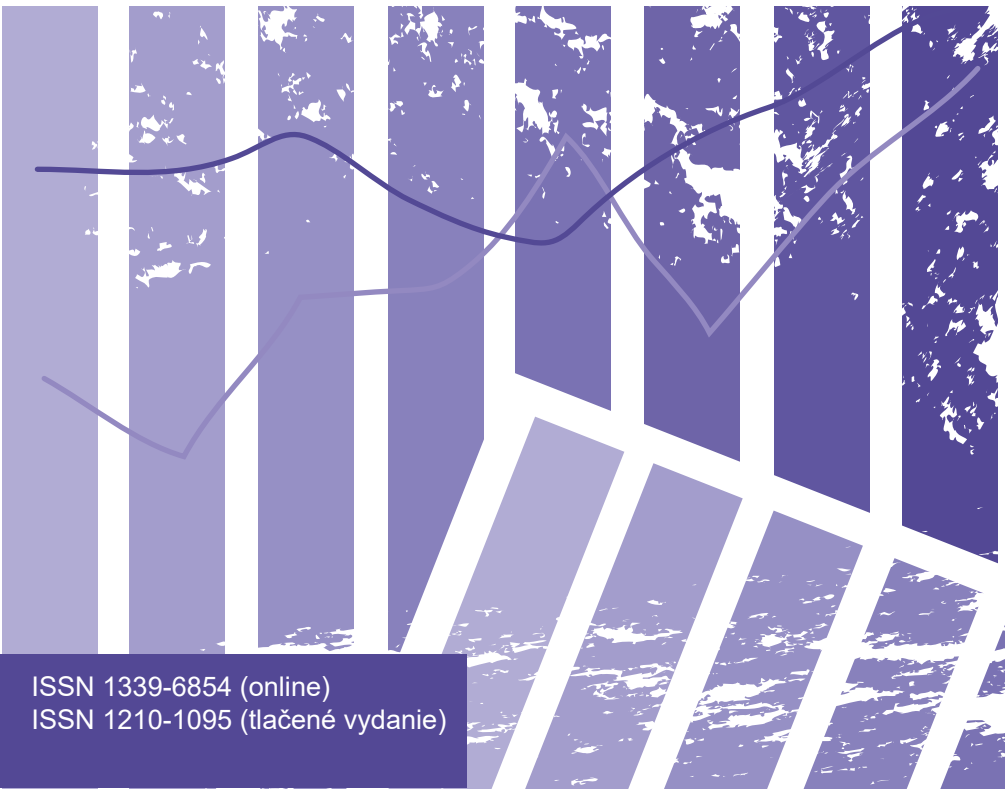
Štatistický úrad Slovenskej republiky
The Statistical Office of the Slovak Republic

SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

vedecký časopis/scientific journal

3/2020
ročník 30



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 jednotlivé čísla časopisu zverejňujeme aj v elektronickej podobe na ssad.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 are available to readers in electronic form at the website ssad.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

ssad.statistics.sk
www.statistics.sk

OBSAH/CONTENTS

EDITORIÁL

Vážení čitatelia,

Doc. Ing. Iveta Stankovičová, PhD.

tretie číslo vedeckého časopisu *Slovenská štatistika a demografia* je tradične monotematické, snažíme sa v ňom prezentovať vybranú skupinu štatistických metód a postupov. V tomto čísle našim čitateľom predstavujeme vybrané viacrozmerné metódy a ich aplikácie. Viacrozmerné štatistické metódy sú vhodné na analýzu zložitých javov. Práve zložité javy a procesy sa v objektívnej realite vyskytujú oveľa častejšie ako javy a procesy jednoduché. Zložité javy a procesy dokážeme merať a charakterizovať iba pomocou viacerých merateľných ukazovateľov. Súbory rôznorodých ukazovateľov potom možno skúmať len pomocou viacrozmernej analýzy, pri ktorej sa využívajú práve tieto metódy.

Pod pojmom *metódy viacrozmernej analýzy* rozumieme široký okruh postupov, ktoré sa venujú viacrozmerným problémom z rôznych hľadísk. K najrozšírenejším metódam patrí: viacrozmerná regresná analýza, kanonická korelačná analýza, analýza hlavných komponentov, faktorová analýza, zhluková analýza, diskriminačná analýza, viacrozmerné škálovanie, viackriteriálne metódy hodnotenia ale aj iné. Jednotlivé metódy netvoria uzatvorené okruhy, ale sú neustále obohacované o nové prístupy viacrozmerného riešenia a možnosti ich aplikácií v praxi.

Metódy viacrozmernej štatistickej analýzy zaznamenali prudký rozvoj najmä v posledných desaťročiach, hoci ich korene siahajú už do 2. polovice 19. storočia. Stretávame sa s nimi prakticky vo všetkých vedných odboroch, počínajúc prírodnými a technickými vedami a končiac spoločenskými vedami. Najčastejšie aplikácie týchto metód sa uskutočnili v oblasti psychológie a sociológie, medicíny, biológie, v rôznych populačných a regionálnych analýzach. Až neskôr sa viacrozmerné prístupy začali používať aj v hospodárskej praxi.

Súčasná doba je charakteristická explóziou dát, ktoré sa ukladajú v databázach. Tým rastie aj význam ich spracovania pomocou metód, ktoré dokážu z dát vyťažiť informácie. Práve informácie sú potrebné na efektívne rozhodovanie a riadenie podniku či spoločnosti. Význam viacrozmerných štatistických metód preto rastie nielen vo vede, ale aj pri riadení procesov v podnikoch a v spoločnosti. Vhodné využitie štatistických metód a správna interpretácia výsledkov získaných z analýz údajov prispievajú k takým rozhodnutiam, ktoré významne šetria čas a finančné prostriedky.

Štatistika ako veda sa rýchlo rozvíja a neustále ponúka nové metódy. Viacrozmerné štatistické metódy majú široké uplatnenie v rôznych oblastiach, avšak vyznačujú sa výpočtovou náročnosťou, čiže ich aplikácia je nemysliteľná bez využitia výpočtovej techniky. To sa na prvý pohľad javí ako prekážka ich širšieho používania. V konečnom dôsledku je to však jediný možný spôsob, ako zabezpečiť široké a systematické využitie

týchto metód v praxi. V súčasnosti je k dispozícii celý rad rôznych softvérových nástrojov, ktoré obsahujú širokú ponuku viacrozmerných metód a neustále vznikajú nové.

Štatistika ako vedná disciplína sa člení na mnoho oblastí a ako sme už uviedli, jej využitie v praxi je širokospektrálne. Aktuálne číslo časopisu *Slovenská štatistika a demografia* sa zameriava len na vybrané viacrozmerné štatistické metódy, ktoré boli aplikované pri analýzach vybraných ekonomických javov a problémov.

Prvý článok je rozsiahly a zaoberá sa v praxi často používanými viacrozmernými metódami: *analýza hlavných komponentov, faktorová analýza a zhluková analýza*. Autorky opisujú teoretické princípy a podmienky správneho použitia týchto metód. Teoretický koncept sa potom názorne aplikuje na hodnotenie a klasifikáciu 34 európskych krajín v oblasti ľudského rozvoja. Pomocou faktorovej analýzy bol vytvorený integrálny ukazovateľ ľudského rozvoja a na základe toho bolo zostavené poradie krajín. Všetky výpočty sa uskutočnili v programovom systéme SAS Enterprise Guide (verzia 7.1).

Ďalší príspevok sa venuje problematike výberu vysvetľujúcich premenných v *klasifikačných stromoch*. Týmto názvom je označovaná skupina metód, ktoré boli navrhnuté na riešenie klasifikačných úloh s vysvetľovanou premennou kategoriálneho typu na základe hodnôt vysvetľujúcich pramenných. Obsah článku je ladený teoreticky, pretože obsahuje množstvo známych aj menej známych vzorcov na výpočet výberových kritérií. V závere autorka opísanú teóriu aplikovala aj na ilustračný príklad. Výpočty boli vykonané v programovom systéme IBM SPSS Decision Trees (verzia 26).

Tretí článok sa zaoberá jednoduchými metódami viacrozmerného hodnotenia (*metóda poradí, bodovacia metóda a metóda vzdialenosti od fiktívneho objektu*) a ich praktickou aplikáciou pri analýze zadlženosti domácností vo vybraných krajinách EÚ. V prvej časti autorka opisuje princíp metód a v druhej časti prezentuje aplikáciu týchto metód v priestorovej analýze dlhu domácností v krajinách, ktoré sa zúčastnili v druhej vlne prieskumu o financovaní a spotrebe domácností (HFCS). Výpočty sa takisto uskutočnili v programovom systéme SAS Enterprise Guide.

Posledný vedecký článok je venovaný *regresnej analýze* a jej aplikáciám v oblasti modelovania úmrtnosti v životnom poistení. Modely úmrtnosti sú štatistické (stochastické) modely, pre ktoré je potrebné odhadnúť parametre regresnej funkcie. Parametre vybraných regresných modelov úmrtnosti boli odhadované iteračnými metódami nelineárnej regresnej analýzy v softvéri SAS Enterprise Guide.

Všetky vedecké články pre toto číslo časopisu napísali ženy, ktoré patria medzi erudované vedecko-pedagogické pracovníčky renomovaných slovenských a českých univerzít. Vo svojej výskumnej činnosti a pedagogickom procese využívajú rôzne štatistické a analytické softvéry alebo aj programovacie jazyky, bez ktorých by bola aplikácia viacrozmerných štatistických metód na veľkých databázach údajov nepredstaviteľná a neuskutočniteľná. V predkladaných článkoch metodicky opísali a ukázali, ako treba v praxi pristupovať k hodnoteniu zložitých javov. Často sa na to využívajú viacrozmerné metódy a postupy, ktoré pri dodržaní podmienok ich použitia prinášajú nové, užitočné, logické a interpretovateľné výsledky. V prípade ich nedodržania,

výsledky môžu byť skreslené, nelogické až nesprávne. Preto je potrebné a dôležité v analytickej praxi ako prvý krok uskutočniť prieskum vstupných údajov a overiť podmienky použitia metód, ktoré plánujeme využiť v našej analýze.

Veríme, že vedecké články a ďalšie príspevky publikované v čísle *3/2020 Slovenskej štatistiky a demografie* budú pre našich čitateľov obohacujúce a podnetné. Tým, ktorí sa zaujímajú o iné oblasti štatistiky, odporúčam do pozornosti vedecké články z niektorých minulých, ale perspektívne aj budúcich čísel nášho časopisu.

Doc. Ing. Iveta STANKOVIČOVÁ, PhD.

Autorka pôsobí na Fakulte managementu Univerzity Komenského v Bratislave. Je členka redakčnej rady časopisu Slovenská štatistika a demografia a bola gestorkou prípravy monotematického čísla 3/2020.

EDITORIAL

Dear readers,

the third issue of the scientific Journal *Slovak Statistics and Demography* is traditionally monothematic, where we always try to present a selected group of statistical methods and procedures. In this issue, we present our readers with selected multidimensional methods and their applications. Multivariate statistical methods are suitable for the analysis of complex phenomena. It is the complex phenomena and processes that occur in objective reality much more frequently than simple phenomena and processes. We can measure and characterize complex phenomena and processes only with the help of several measurable indicators. Sets of diverse indicators can then only be examined using multivariate analysis, with the use of these methods.

By the *methods of multivariate analysis* we shall mean a wide range of procedures addressing multidimensional problems from different perspectives. The most common methods include: multidimensional regression analysis, canonical correlation analysis, principal components analysis, factor analysis, cluster analysis, discriminant analysis, multidimensional scaling, multicriteria evaluation methods, and others as well. The individual methods do not form closed circuits, but are constantly enriched with new approaches of multidimensional solutions and the possibility of their application in practice.

The methods of multivariate statistical analysis have developed rapidly, especially in recent decades, although their roots go back to the second half of the 19th century. We come across with them practically in all scientific disciplines, from natural and technical sciences to social sciences. These methods were the most frequently used in the field of psychology and sociology, medicine, biology, in various population and regional analyses. Only later the multidimensional approaches started to be used in the economic practice.

The current period is characterized by the explosion of data which are stored in databases. This also increases the importance of their processing, using methods enabling to extract information from the data. It is the information that is necessary for the effective decision-making and management of a company or a society. Therefore the importance of multidimensional statistical methods is growing not only in science, but also in process management in companies and in a society. The appropriate use of statistical methods and the correct interpretation of results obtained from data analyses contribute to a decisions that significantly save time and financial resources.

Statistics as a science is evolving rapidly and constantly offering new methods. Multivariate statistical methods have a wide range of application in various fields, but they are characterized by computational complexity, thus their application is unthinkable without the use of computational technology. At first glance, this seems to be an obstacle in their wider use. However, ultimately, this is the only possible way to ensure the wide and systematic use of these methods in practice. There are currently a number of different software tools available that include a wide range of multivariate methods and are constantly evolving.

Statistics as a scientific discipline is divided into many areas and, as we have already mentioned, it has a large-scale use in practice. The current issue of the Journal *Slovak Statistics and Demography* focuses only on selected multidimensional statistical methods that have been applied in the analysis of selected economic phenomena and problems.

The first article is extensive and focuses on the multidimensional methods frequently used in practice: *principal component analysis, factor analysis and cluster analysis*. The authors describe the theoretical principles and conditions of the correct use of these methods. The theoretical concept is then clearly applied to the evaluation and classification of 34 European countries in the field of human development. Using factor analysis, an integral indicator of human development was created, on the basis of which, the ranking of countries was compiled. All calculations are performed in the SAS Enterprise Guide software (version 7.1).

Another article deals with the issue of selecting explanatory variables in *classification trees.* This name refers to a group of methods that have been designed to solve classification problems with an explanatory variable of the categorical type, based on the values of the explanatory variables. The content of the article is theoretical, because it contains a number of known and less known formulae for calculating selection criteria. In the end, the author applied the described theory to an illustrative example. The calculations were performed on the IBM SPSS Decision Trees software (version 26).

The third article deals with simple methods of multidimensional evaluation (*order method, scoring method and method of distance from a fictitious object*) and their practical application in the analysis of household indebtedness in selected EU countries. In the first part, the author describes the principle of the methods and in the second part presents the application of these methods in the spatial analysis of household debt in countries participating in the second wave of the Survey on household finance and consumption (HFCS). The calculations are also performed in the SAS Enterprise Guide software.

The last scientific article is devoted to the *regression analysis* and its applications in the field of modelling mortality in life insurance. Mortality models are statistical (stochastic) models for which it is necessary to estimate the parameters of the regression function. The parameters of selected regression models of mortality were estimated by iterative methods of nonlinear regression analysis in the SAS Enterprise Guide software.

All scientific articles for this issue of the journal were written by women who are members of the erudite scientific and pedagogical staff of the prestigious Slovak and Czech universities. In their research activities and pedagogical process, they use various statistical and analytical software, or even programming languages, without which the application of multidimensional statistical methods on large data databases would be unimaginable and unfeasible. In the presented articles, they methodically described and showed how to evaluate the complex phenomena in practice. To this end, multidimensional methods and procedures are often used, which under the conditions of their use bring new, useful, logical and interpretable results. In the event of non-compliance, these results can be skewed, illogical or incorrect results. Therefore, it is

necessary and important in the analytical practice as a first step to conduct a survey of input data and verify the conditions of using the methods planned to use in our analysis.

We believe that the scientific articles and other contributions published in the issue *3/2020 of Slovak Statistics and Demography* will be enriching and stimulating for our readers. I commend to those who are interested in other areas of statistics, the scientific articles from some of the past, but also future issues of our journal.

Doc. Ing. Iveta STANKOVIČOVÁ, PhD.

The author works at the Faculty of Management of Comenius University in Bratislava. She is a member of the Editorial Board of the Journal Slovak Statistics and Demography and was responsible for the preparation of the monothematic issue 3/2020.

Iveta STANKOVIČOVÁ
Fakulta managementu, Univerzita Komenského v Bratislave
Alena MOJSEJOVÁ
Ekonomická fakulta, Technická univerzita v Košiciach

INDEX ĽUDSKÉHO ROZVOJA: HODNOTENIE A KLASIFIKÁCIA EURÓPSKYCH KRAJÍN POMOCOU VIACROZMERNÝCH METÓD

HUMAN DEVELOPMENT INDEX: RATING AND CLASSIFICATION OF EUROPEAN COUNTRIES BY MULTIVARIATE METHODS

ABSTRAKT
Kvalita života a ľudský rozvoj v krajinách sveta sú v poslednej dobe často diskutované témy. Ide o zložené javy, ktoré vieme merať len pomocou viacerých ukazovateľov. Experti z Rozvojového programu Organizácie Spojených národov (UNDP) navrhli metodiku výpočtu integrálneho ukazovateľa na hodnotenie krajín sveta v oblasti ľudského rozvoja. Ukazovateľ sa volá index ľudského rozvoja (HDI) a v súčasnosti sa skladá zo štyroch čiastkových ukazovateľov, ktoré vieme zmerať. Na základe dosiahnutých hodnôt HDI, navrhli aj klasifikáciu krajín (resp. regiónov) do skupín. Cieľom príspevku je porovnať úroveň HDI v 34 krajinách Európy a vytvoriť hodnotiaci rebríček. Na tento účel použijeme hodnoty HDI vypočítané podľa metodiky UNDP, ale aj HDI vypočítaný viacrozmernými štatistickými metódami, ako sú analýza hlavných komponentov a faktorová analýza. Na klasifikáciu krajín do skupín použijeme zhlukovú analýzu. Výsledky porovnáme a zhodnotíme výhody a nevýhody týchto rôznych prístupov pri hodnotení ľudského rozvoja.

ABSTRACT
Quality of life and human development in the countries of the world have recently been frequently discussed. This compound phenomena can be measured only by using several indicators. Experts from the United Nations Development Program (UNDP) have proposed a methodology for calculating an integral indicator for assessing the countries of the world in the area of human development. The indicator is called the Human Development Index (HDI) and currently consists of four measurable sub-indicators. Based on the achieved HDI values, they also proposed the classification of countries (or regions) into groups. The aim of the paper is to compare the HDI values in 34 European countries and to create a ranking. For this purpose, we will use HDI values calculated by the UNDP methodology, but also the HDI calculated using multivariate statistical methods such as principal component analysis and factor analysis. We will use cluster analysis for the classification of countries into groups. We will compare the results and evaluate the advantages and disadvantages of these different approaches for evaluating human development.

KĽÚČOVÉ SLOVÁ
index ľudského rozvoja, analýza hlavných komponentov, faktorová analýza, zhluková analýza, SAS Enterprise Guide

KEY WORDS
human development index, principal component analysis, factor analysis, cluster analysis, SAS Enterprise Guide

1. ÚVOD

Súčasná doba je charakteristická komparáciami rôznych javov. Často sa navzájom porovnávajú ekonomiky krajín, resp. zoskupení. Ako však môžeme jednoducho porovnávať národné ekonomiky? V minulosti sa používali hlavne makroekonomické ukazovatele, ako napríklad hrubý domáci produkt, inflácia, miera nezamestnanosti alebo zamestnanosti, platobná bilancia a dlh (pozri [4, 7, 10, 13]). Problémom týchto ukazovateľov je ich zameranie na stav a vývoj hospodárstva v krajine, ako na stav a vývoj krajiny ako celku. Makroekonomické premenné nehovoria o kvalite života a ľudskom rozvoji občanov v sledovanej krajine (pozri [5]).

Ľudský rozvoj je predovšetkým o snahe umožniť ľuďom viesť životy, ktoré si cenia a ktoré im umožňujú uvedomiť si ich potenciál ako ľudských bytostí. Trvalo udržateľné rozvojové ciele sú ciele v medzinárodne dohodnutom súbore čiastkových cieľov (známy pod názvom Agenda 2030, [15]) na zníženie extrémnej chudoby a rozšírenie rodovej rovnosti, rovnosti príležitostí na zdravie a vzdelávanie. Redukovanie nerovností v rôznych oblastiach sa stalo kľúčovým cieľom v Agende 2030.

Porovnávanie krajín na základe rôznych ukazovateľov predstavuje dôležitý pohľad na postavenie jednotlivých štátov a analýzu nerovnosti medzi nimi. Nerovnosť medzi krajinami ale aj v ich vnútri krajín je naďalej vážnym problémom napriek pokroku a snahám o zmenšovanie týchto rozdielov.

Nerovnosti v oblasti ľudského rozvoja sú najviac viditeľné. Podľa Správy o ľudskom rozvoji [18], nerovnosti v tejto oblasti poškodzujú spoločnosť, oslabujú sociálnu súdržnosť, dôveru ľudí vo vládu a v inštitúcie, ale aj vzájomnú dôveru ľudí. Navyše, tieto nerovnosti sú prekážkou pri dosahovaní cieľov Agendy 2030 pre trvalo udržateľný rozvoj.

Cieľom príspevku je analyzovať stav a vývoj ľudského rozvoja v európskych krajinách na základe známej a používanej miery, konkrétne indexu ľudského rozvoja (HDI). Téma analýzy ľudského rozvoja je veľmi aktuálna, o čom svedčí aj množstvo publikácií, ktoré sa danej tematike venujú, napr. [1, 2, 3]. V prácach [3] a [18] autori pomocou zhlukovej analýzy klasifikujú krajiny sveta do skupín podľa dosiahnutého stupňa ľudského rozvoja meraného pomocou HDI. Zhlukovej analýze európskych krajín podľa HDI na regionálnej úrovni sa venuje práca [9].

2. MERANIE ĽUDSKÉHO ROZVOJA

Ľudský rozvoj a kvalita života sú veľmi blízke pojmy a veľmi často sa s nimi stretávame v spoločných analýzach. Kvalita života je výsledkom vzájomného pôsobenia sociálnych, zdravotných a environmentálnych podmienok, týkajúcich sa ľudského a spoločenského rozvoja. Pojem kvalita života sa prvýkrát objavil v práci amerického ekonóma J. K. Galbraitha, ktorému sa pripisuje jeho autorstvo.

Dosiahnuté výsledky v oblasti ľudského rozvoja vo vyše 150 krajinách sveta, od roku 1990 monitoruje Rozvojový program Organizácie Spojených národov (UNDP – United Nations Development Programme). Ľudský rozvoj je podľa Organizácie spojených národov (OSN) definovaný ako proces rozširovania ľudských možností a obsiahnutý stupeň blahobytu. Pri ľudskom rozvoji rozlišujeme dva pohľady na získané schopnosti. Na jednej strane je to formovanie nadania, ktoré sa môže merať kvalitným vzdelávaním alebo dostupnou zdravotnou starostlivosťou. Na druhej strane uvažujeme využitie získaných schopností v pracovnom i voľnom čase (definícia podľa UNDP). Tieto možnosti sa môžu neustále meniť a pribúdať alebo sa môžu stať neaktuálnymi. V tomto prípade sa jedná o tri nevyhnutné činitele (podľa [17 a 19]):

- dĺžka života obyvateľstva danej krajiny,
- dosiahnutá vzdelanostná úroveň obyvateľstva,
- kvalita života ľudí vyjadrená reálnym HDP pripadajúcim na jedného obyvateľa.

2.1 MERANIE ĽUDSKÉHO ROZVOJA PODĽA METODIKY UNDP

Index ľudského rozvoja (z angl. Human Development Index – HDI) môžeme zaradiť medzi komplexné indikátory trvalo udržateľného rozvoja sociálno-ekonomickej povahy. Prvýkrát bol publikovaný v roku 1990 v správe Human Development Report. Hodnotí dosiahnuté priemerné výsledky krajiny v troch dimenziách ľudského rozvoja: dlhý a zdravý život (index zdravia – IH), prístup k vedomostiam (index vzdelania – IE) a primeraný životný štandard (príjmový index – II). Zhrnutie jednotlivých indexov do jedného ukazovateľa sa vytvára na báze kompozitného indikátora, pričom UNDP používa rovnaké váhy pre všetky tri ukazovatele (indexy). Výsledný HDI sa vypočítava ako geometrický priemer troch indexov, tzv. dimenzií:

$$HDI = \sqrt[3]{Index\ zdravia * Index\ vzdelania * Príjmový\ index}\ . \qquad (1)$$

Po viacerých diskusiách odborníkov bola zvolená *očakávaná dĺžka života pri narodení* (angl. life expectancy at birth) ako najvhodnejší indikátor prvej dimenzie predstavujúcej "dlhý a zdravý život". Má dôležitú výpovednú hodnotu, keďže vek obyvateľstva krajiny môže vyjadrovať v určitých prípadoch napríklad aj zdravotný status ľudí tu žijúcich. Druhá dimenzia predstavujúca vzdelanie je vyjadrená mierou gramotnosti. Aktuálnymi indikátormi sú dva ukazovatele: *predpokladané roky štúdia* (angl. expected years of schooling) a *priemerný počet rokov vzdelávania* (angl. mean years of schooling). Poslednú dimenziu predstavuje kúpyschopnosť obyvateľstva, čiže uspokojovanie základných potrieb. Ukazovateľom, ktorý sa využíva v terajšom vzťahu je hrubý národný produkt (HNP) na obyvateľa vyjadrený v PPP $[1] (angl. GNI per capita in PPP $). Tento ukazovateľ nahradil predtým používaný logaritmus hrubého domáceho produktu (HDP).

[1] *PKS – parita kúpnej sily (angl. PPP – Purchasing Power Parity) je ekonomický ukazovateľ, na základe ktorého je možné porovnávať ekonomickú silu jednej krajiny (regiónu) oproti inej. Je to špecifický prevodový cenový index, ktorý dáva do súvisu cenové rozdiely tovarov a služieb v určitom okamihu v rôznych krajinách (regiónoch) pri eliminovaní rozdielov v cenovej úrovni medzi krajinami. V praxi sa často využíva ukazovateľ HDP na obyvateľa v parite kúpnej sily (HDP/PKS alebo angl. GDP/PPP), vyjadrený v US dolároch v parite kúpnej sily.*

Podľa viacerých autorov (pozri napr. [1, 6, 8, 13]), HDI poskytuje presný spôsob výpočtu ľudského rozvoja, čo je rozhodujúce pri meraní celkovej výkonnosti ktorejkoľvek krajiny. HDI navyše umožňuje ľahké a spravodlivé hodnotenie krajín na základe získanej hodnoty.

Výpočet výsledného indexu ľudského rozvoja HDI možno podľa [19] zhrnúť do dvoch krokov. V tabuľke č. 1 je zoznam ukazovateľov, ktoré sa používali v minulosti a ktoré sa používajú v súčasnosti (od roku 2009). Z týchto štyroch ukazovateľov sa počíta výsledná hodnota HDI.

Tabuľka č. 1: Porovnanie komponentov indexu ľudského rozvoja (HDI)

Pôvodný index z roku 1990		**Index aktuálne využívaný (od roku 2009)**		**Index**
Komponent	**Indikátor**	**Dimenzia**	**Indikátor**	
Dlhý život	Očakávaná dĺžka života pri narodení	Dlhý a zdravý život	Očakávaná dĺžka života pri narodení	**Index zdravia**
Vedomosti	Miera gramotnosti	Vedomosti	Očakávaný počet rokov štúdia	**Index vzdelania**
			Priemerný počet rokov štúdia	
Kúpna sila	Logaritmus hrubého domáceho produktu na obyvateľa	Primerané životné štandardy	Hrubý národný produkt na obyvateľa v PPP $ (2011)	**Príjmový index**

Zdroj: vlastné spracovanie podľa [4] a [6]

Prvým krokom je vytvorenie tzv. indexu dimenzie. Pre každý zo štyroch indikátorov odborníci určili hranice, minimálnu a maximálnu hodnotu (tabuľka č. 2).

Tabuľka č. 2: Minimálne a maximálne hodnoty komponentov HDI

Dimenzia	**Indikátor (ukazovateľ)**	**Min.**	**Max.**
Zdravie	Očakávaná dĺžka života pri narodení (v rokoch)	20	85
Vzdelanie	Očakávaný počet rokov štúdia	0	18
	Priemerný počet rokov štúdia	0	15
Primerané životné náklady	Hrubý národný produkt na obyvateľa (2011 PPP $)	100	75 000

Zdroj: spracované podľa [19]

Podľa [19] tieto hranice sú určené kvôli preškálovaniu indikátorov vyjadrených v rôznych jednotkách na škálu od 0 po 1. Minimálna hodnota očakávanej dĺžky života je stanovená na 20 rokov, keďže žiadna krajina v 20. storočí nedosahovala nižšiu hodnotu. Maximálna dĺžka života je stanovená na 85 rokov, čo predstavuje cieľ pre mnoho krajín za posledných 30 rokov. Spoločnosti môžu existovať bez formálneho vzdelania, čo odôvodňuje nastavenie minimálnej hodnoty očakávaného počtu rokov štúdia na hodnotu 0. Maximálne očakávaný počet rokov štúdia je 18, čo je rovnocenné získaniu druhého stupňa vysokoškolského štúdia vo väčšine krajín. Maximálna hodnota priemerného počtu rokov štúdia je 15, čo je predpokladané maximum tohto ukazovateľa na rok 2025. Nízka minimálna hodnota hrubého národného dôchodku na obyvateľa je 100 dolárov. Podľa UNDP predstavuje nezmerateľnú časť HND, ktorá nie je zachytená v oficiálnych štatistikách. Maximálna výška je stanovená na 75 000 dolárov na obyvateľa. V súčasnosti (r. 2018) iba štyri krajiny (Brunej, Lichtenštajnsko, Katar a Singapur)

presahujú HND vo výške 75 000 dolárov na obyvateľa. Je však dokázané, že príjem nad 75 000 dolárov neprináša vysoký zisk pre ľudský rozvoj v krajine.[2]

Po definovaní minimálnej a maximálnej hodnoty, môžeme každý indikátor preškálovať na hodnoty z intervalu <0, 1> podľa nasledujúceho vzorca:

$$Index\ dimenzie = \frac{aktuálna\ hodnota - minimálna\ hodnota}{maximálna\ hodnota - minimálna\ hodnota}. \quad (2)$$

V prípade indexu vzdelania, ktorý sa skladá až z dvoch čiastkových ukazovateľov (indikátorov), sa preškálovaný index vypočíta pre každý indikátor zvlášť a výsledný index je priemerom týchto dvoch hodnôt.

Výpočty čiastkových indexov a HDI ilustrujeme na príklade Slovenska (výpočty na rok 2018):

$$Index\ zdravia = \frac{77{,}4 - 20}{85 - 20} = 0{,}883 \quad (3)$$

$$Index\ očakávaného\ počtu\ rokov\ štúdia = \frac{14{,}5 - 0}{18 - 0} = 0{,}806 \quad (4)$$

$$Index\ priemerného\ počtu\ rokov\ štúdia = \frac{12{,}6 - 0}{15 - 0} = 0{,}840 \quad (5)$$

$$Index\ vzdelania = \frac{0{,}806 + 0{,}840}{2} = 0{,}823 \quad (6)$$

$$Príjmový\ index = \frac{\ln(30672) - \ln(100)}{\ln(75000) - \ln(100)} = 0{,}865 \quad (7)$$

Tabuľka č. 3: Vstupné hodnoty ukazovateľov na výpočet komponentov HDI pre Slovensko (rok 2018)

Indikátor	Hodnota
Očakávaná dĺžka života pri narodení (v rokoch)	77,4
Očakávaný počet rokov štúdia	14,5
Priemerný počet rokov štúdia	12,6
Hrubý národný produkt na obyvateľa (2011 PPP $)	30672,0

Zdroj: vlastné spracovanie podľa [16]

Výsledný index ľudského rozvoja (HDI) pre SR sa vypočíta ako geometrický priemer z hodnôt indexu zdravia, indexu vzdelania a príjmového indexu:

[2] *V roku 2007 dosiahlo Luxembursko veľmi vysokú hodnotu HND (91 519 dolárov na obyvateľa v PPP $ 2011) , t. j. nad hranicou 75 000 dolárov na obyvateľa.*

$$HDI = \sqrt[3]{Index\ zdravia * Index\ vzdelania * Príjmový\ index} = \\ = (0{,}883 * 0{,}823 * 0{,}865)^{1/3} = 0{,}857 \quad (8)$$

Hodnoty HDI a jeho jednotlivých komponentov pre Slovensko majú od začiatku merania v roku 1990 až po súčasnosť rastúci trend (tabuľka č. 4). Najväčší nárast môžeme pozorovať pri indexe vzdelania. V roku 2019 sa Slovensko v hodnotení HDI nachádzalo na 36. mieste medzi 189 hodnotenými krajinami sveta na základe údajov z roku 2018 ([18]).

Tabuľka č. 4: Vývoj HDI a jeho komponentov na Slovensku (1990 – 2018)

Rok	Index vzdelania	Index zdravia	Príjmový index	HDI
1990	0,679	0,788	0,753	0,739
2000	0,712	0,820	0,762	0,763
2007	0,779	0,842	0,823	0,814
2010	0,802	0,854	0,832	0,829
2012	0,824	0,863	0,838	0,842
2016	0,822	0,877	0,855	0,851
2017	0,824	0,880	0,859	0,854
2018	0,824	0,883	0,865	0,857

Zdroj: spracované podľa [16]

Všetky krajiny sú každoročne klasifikované v oblasti stupňa ľudského rozvoja a zatriedené do jednej zo štyroch skupín podľa skóre HDI (tabuľka č. 5). Klasifikačná schéma sa vzťahuje aj na každý z troch čiastkových indexov.

Tabuľka č. 5: Klasifikácia ľudského rozvoja podľa UNDP

Hodnota HDI	Skupina rozvoja
0,800 a viac	Veľmi vysoký stupeň ľudského rozvoja
0,700 – 0,799	Vysoký stupeň ľudského rozvoja
0,550 – 0,699	Stredný stupeň ľudského rozvoja
menej ako 0,550	Nízky stupeň ľudského rozvoja

Zdroj: spracované podľa [19]

V článku hodnotíme ľudský rozvoj na základe HDI a jeho komponentov vo vybraných európskych krajinách (spolu 34 krajín). Z popisnej štatistiky indexov (tabuľka č. 6) a z údajov v prílohách článku (príloha č. 2) je zrejmé, že skoro všetky európske krajiny sa v roku 2007 zaradili do skupiny s veľmi vysokým stupňom ľudského rozvoja (tabuľka č 5, HDI ≥ 0,8). Do nižšej skupiny, skupiny krajín s vysokým stupňom rozvoja, patrilo len 6 štátov, konkrétne Rumunsko, Čierna Hora, Bulharsko, Srbsko, Albánsko a Turecko. Väčšinou išlo o štáty, ktoré ešte nepatrili do EÚ ale boli kandidátskymi krajinami na vstup do Európskej únie. Najvyššiu hodnotu HDI dosiahlo Nórsko (0,938) a najnižšiu Turecko (0,709).

Tabuľka č. 6: Popisné štatistiky indexov HDI pre 34 vybraných európskych krajín (r. 2007 a 2018)

Štatistika	IE_2007	IH_2007	II_2007	IE_2018	IH_2018	II_2018	HDI 2007	HDI 2018	Rank diff
Min	0.557	0.794	0.674	0.712	0.845	0.727	0.709	0.791	-3
Max	0.915	0.948	1.000	0.946	0.979	0.985	0.938	0.954	6
Medián	0.804	0.908	0.864	0.850	0.941	0.874	0.849	0.884	-1
Priemer	0.796	0.885	0.858	0.847	0.924	0.878	0.845	0.882	0

Vysvetlivky: IE – index vzdelania, IH – index zdravia, II – príjmový index, HDI – index ľudského rozvoja, Rank – poradie krajiny v hodnotení podľa HDI, Rank diff – rozdiel v poradí 2007 vs. 2018.
Zdroj: vlastné spracovanie v SAS EG na základe údajov z prílohy č. 1.

V roku 2018 sa hodnoty HDI v sledovaných krajinách zvýšili a do skupiny s veľmi vysokým stupňom ľudského rozvoja (tabuľka č. 5, HDI ≥ 0,8) sa dostali skoro všetky analyzované krajiny. Do nižšej skupiny, skupiny krajín s vysokým stupňom rozvoja, patria už len 2 štáty, konkrétne Srbsko (0,799) a Albánsko (0,791). Turecko dosiahlo hodnotu HDI nad hranicou 0,8 (0,803). Najvyššiu hodnotu HDI dosiahlo opäť Nórsko (0,954).

Záverom tejto časti možno konštatovať, že na základe klasifikácie UNDP, rozdeľujeme vybrané európske krajiny len do dvoch skupín. Zaujímavé bude zistiť, aké skupiny, resp. zhluky krajín sa vytvoria použitím viacrozmerných štatistických metód.

2.2 MERANIE ĽUDSKÉHO ROZVOJA VIACROZMERNÝMI ŠTATISTICKÝMI METÓDAMI

Ľudský rozvoj v krajine je zložený jav. Index ľudského rozvoja ako ukazovateľ na meranie tohto zloženého javu môžeme považovať za integrálny ukazovateľ. Hodnota HDI je výsledkom matematických výpočtov na základe hodnôt merateľných čiastkových ukazovateľov (indikátorov). Na základe metodiky OSN (UNDP) sa ako výsledný vzorec na výpočet HDI používa geometrický priemer z 3 indexov, konštruovaných podľa vzorcov uvedených v predchádzajúcej časti.

V tejto časti článku ukážeme, či by na výpočet HDI neboli vhodné viacrozmerné štatistické metódy ako analýza hlavných komponentov, resp. faktorová analýza a na klasifikáciu krajín do skupín zase zhluková analýza.

Viacrozmerné štatistické metódy sa používajú na zjednodušenie štatistických analýz. Často sa v praxi stretávame s tým, že počiatočný počet analyzovaných znakov (premenných) je veľmi vysoký, čo sťažuje analýzu a interpretáciu výsledkov. Pre zjednodušenie je vhodné nahradiť veľký počet prvotných premenných menším počtom podstatných znakov bez toho, že by sme stratili väčšiu časť informácie. Na riešenie tohto problému boli vytvorené dve príbuzné metódy, a to metóda hlavných komponentov (PCA - principal components analysis) a faktorová analýza (FA – factor analysis).

Obe metódy patria do skupiny metód slúžiacich na analýzu skrytých vzťahov medzi premennými. Cieľom je pochopiť a identifikovať, ako sú premenné prepojené, to znamená, ako sú skorelované. Ak sú premenné skorelované, tak je možné rovnaký objem informácie, ktorý vyjadrujú, vyjadriť menším počtom premenných, čiže môžeme tak znížiť dimenziu. Obe tieto metódy preto vychádzajú z analýzy kovariančnej, resp. korelačnej

matice pôvodných východiskových premenných a pokúšajú sa nájsť skryté, nemerateľné (tzv. latentné) premenné. Tieto nové premenné sa nedajú priamo zmerať, ale majú určitú schopnosť vecnej interpretácie.

2.2.1 Analýza hlavných komponentov

Analýza hlavných komponentov je metódou na tvorbu nových premenných, ktoré sú lineárnymi kombináciami pôvodných *p* premenných. Maximálny počet nových, už nekorelovaných premenných, ktoré môžeme vytvoriť pomocou PCA, sa rovná počtu pôvodných premenných. Nové umelé premenné nazývame hlavné komponenty (PRIN). Hlavné komponenty sú navzájom už nekorelované. Prvý hlavný komponent je najdôležitejší, lebo vysvetľuje najväčšiu časť celkovej variability údajov a ďalšie hlavné komponenty vysvetľujú postupne zvyšnú časť variability, a to tak, že podiel vysvetlenej variability postupne klesá a na posledný hlavný komponent ostáva len nepatrná časť.

Metóda hlavných komponentov sa vo všeobecnosti používa:

- na identifikáciu odľahlých, resp. vplyvných pozorovaní v údajoch (angl. outliers), ktoré niekedy veľmi silno ovplyvňujú výsledky štatistických analýz;
- na zníženie dimenzie viacrozmernej analýzy;
- na odstránenie závislosti medzi premennými a následné použitie nezávislých hlavných komponentov v iných metódach, napr. v zhlukovej analýze alebo pri tvorbe lineárnych regresných modelov na odstránenie multikolinearity.

Predpokladajme, že súbor *p* pozorovaných premenných X_1, X_2, X_3, ..., X_p transformujeme do súboru nových premenných Y_1, Y_2, Y_3, ..., Y_p (hlavných komponentov) tak, že sú lineárnou kombináciou pôvodných premenných a môžeme ich zapísať nasledujúcim spôsobom:

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\vdots \\ Y_p &= a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{aligned} \qquad (9)$$

Koeficienty (váhy, saturácie) a_{ij} každej z týchto lineárnych kombinácií sú odhadované tak, aby spĺňali 5 základných vlastností (pozri [14], kapitola 2).

2.2.2 Faktorová analýza

Faktorová analýza vznikla v psychológii (Ch. Spearman, 1904), kde sa dlho výhradne používala a aj dnes sa často používa. Postupne však prenikla aj do iných vedných odborov. Túto metódu však niektorí štatistici kritizujú pre jej nejednoznačné riešenie, pre subjektivitu v niektorých jej cieľoch a krokoch, hmlistú interpretáciu a približnosť výsledkov.

Faktorová analýza je podobná metóde analýzy hlavných komponentov, pretože takisto je určená na vytváranie nových premenných a na zníženie dimenzie dát s čo najmenšou stratou informácie. Do určitej miery je možné FA považovať za rozšírenie PCA. Na rozdiel

od PCA však vychádza zo snahy vysvetliť závislosti medzi pôvodnými premennými. Medzi nedostatky PCA patrí, že je závislá od merných jednotiek premenných, neposkytuje jednoznačné kritérium na rozhodnutie, či zvolený počet hlavných komponentov vysvetľuje dostatočné percento celkovej variability, a nezaoberá sa chybovým rozptylom premenných. Prístup FA čiastočne odstraňuje tieto nedostatky PCA, má však iné slabé miesta. FA má veľa subjektívnych aspektov a nejednoznačnosť odhadov faktorových parametrov. Prednosťou FA je jej väčšia všeobecnosť a úspornosť, i keď niektoré odhady vyžadujú splnenie aspoň približného viacrozmerného normálneho rozdelenia. Dôležité je aj určenie počtu spoločných faktorov pred vykonaním FA, ktoré musí vychádzať z hypotéz výskumníka v predmetnej aplikačnej oblasti.

Predpokladajme, tak ako pri PCA, že máme súbor *p* pozorovateľných premenných (náhodných veličín) $X_1, X_2, ..., X_p$, ktoré majú viacrozmerné rozdelenie s *p*-členným vektorom stredných hodnôt $\boldsymbol{\mu}_X$ a s kovariančnou maticou $\sum_X$ s hodnosťou *p*. Všeobecný model FA predpokladá, že existuje *q* v pozadí stojacich *spoločných faktorov* $F_1, F_2, ...$, F_q, ktorých je menej ako *p* (najlepšie výrazne menej). Tieto faktory umožňujú *j*-tu pozorovateľnú náhodnú premennú X_j, $j = 1, 2, ..., p$ vyjadriť týmto spôsobom:

$$X_j = \mu_j + a_{j1}F_1 + a_{j2}F_2 + ... + a_{jq}F_q + \varepsilon_j \tag{10}$$

kde ε_j , $j = 1, 2, ..., p$ sú náhodné (chybové) zložky, označované ako *špecifické faktory* a a_{jk} sú *faktorové váhy (náklady, saturácie, záťaže)*, ktoré vyjadrujú vplyv *k*-tého spoločného faktora na premennú X_j.

Ak použijeme terminológiu z regresnej analýzy, tak faktorové váhy (saturácie) α_{jk} predstavujú regresné koeficienty medzi pozorovateľnými premennými a nepozorovateľnými faktormi. Pri splnení určitých podmienok sú to vlastne kovariancie medzi nimi. Za predpokladu, že pozorované premenné sú merané v rovnakých merných jednotkách, môžeme faktorové saturácie interpretovať ako príspevok *k*-teho faktora k *j*-tej vysvetľovanej premennej. V maticovom tvare môžeme model FA zapísať takto:

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\mu}_\mathbf{X} + \mathbf{AF} + \boldsymbol{\varepsilon} \text{ resp.} \\ \mathbf{X} - \boldsymbol{\mu}_\mathbf{X} &= \mathbf{AF} + \boldsymbol{\varepsilon} \end{aligned} \tag{11}$$

kde **A** je matica faktorových váh typu *p* x *q*, **F** je *q*-členný vektor spoločných faktorov a $\boldsymbol{\varepsilon}$ je *p*-členný vektor špecifických faktorov. Vektor **X** je vektorom pôvodných merateľných premenných, ktoré sa nazývajú tiež *indikátory.*

Bez straty všeobecnosti, môžeme vychádzať aj z normovaných premenných Z_j, čo je pri FA bežnejší prípad. Maticový model môžeme potom zapísať v tvare:

$$\begin{aligned} \frac{\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}}{\boldsymbol{\sigma}_\mathbf{X}} &= \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon}^* \\ \mathbf{Z} &= \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon}^* \end{aligned} \tag{12}$$

V tomto modeli je rozdiel hlavne v matici faktorových záťaží $\mathbf{A}^*$, kde prvky α^*_{jk} predstavujú korelačné koeficienty medzi premennými X_j a faktormi F_k . Pre model FA v ktoromkoľvek tvare predpokladáme:

P1. Spoločné faktory F_k , k = 1, 2, ..., *q* sú nezávislé a rovnako rozdelené náhodné veličiny s nulovými strednými hodnotami E(F_k) = 0 a jednotkovými rozptylmi D(F_k) = 1.

P2. Špecifické faktory ε_j , j = 1, 2, ..., *p* sú náhodné premenné s nulovými strednými hodnotami E(ε_j) = 0 a s rozptylmi D(ε_j) = e_j, ktoré sú po dvojiciach nezávislé, čiže cov(ε_i, ε_j) = 0.

P3. Faktory F_k a ε_j sú nezávislé náhodné premenné pre každú kombináciu k = 1, 2, ..., *q* a j=1, 2, ..., *p*., čiže cov(F_k, ε_j) = 0.

Faktorový model $\mathbf{\Sigma_X = AA^T + E}$, v ktorom symbol Σ_x označuje kovariančnú maticu, neurčuje maticu faktorových váh **A** jednoznačne (okrem prípadu keď *q = 1*). Ak nejaká matica **A** vyhovuje FA modelu, tak pre maticu **B = AT**, kde **T** je ortogonálna matica (jej stĺpce sú ortogonálne s normou 1), platí: $\mathbf{\Sigma_X = BB^T + E}$. Matica **B** je tiež riešením FA modelu. Hovoríme, že sme ju získali ortogonálnou rotáciou matice **A**.

Matica faktorových váh nám pomáha identifikovať vzťah medzi spoločnými faktormi a identifikátormi. Rotáciou[3] faktorov sa snažíme nájsť takú maticu váh, ktorá je prijateľnejšie interpretovateľná. Pretože existuje nekonečne veľa faktorových riešení z jednej kovariančnej, resp. korelačnej matice, vzniká otázka, či existuje *optimálna* množina spoločných faktorov.

Americký vedec Louis L. Thurston (1887 – 1955) vo svojich prácach zaviedol pojem jednoduchej štruktúry, ktorý môže slúžiť ako kritérium na hľadanie optimálneho riešenia, a naformuloval 5 pravidiel pre jednoduchú štruktúru FA modelu:

1. Každý riadok matice faktorových váh by mal obsahovať aspoň jednu nulu.
2. Každý stĺpec matice faktorových váh by mal obsahovať aspoň q núl.
3. Pre každú dvojicu stĺpcov matice faktorových váh by malo byť viacero premenných, ktoré s jedným faktorom majú nulovú váhu, ale s ostatnými faktormi ich majú vysoké.
4. Pre 5 a viac faktorov by už v každej dvojici stĺpcov matice faktorových váh malo byť čo najviac premenných, ktoré majú nulové váhy v obidvoch stĺpcoch.
5. Pre každú dvojicu stĺpcov matice faktorových váh by malo byť málo premenných, ktoré majú vysoké váhy v obidvoch stĺpcoch.

Cieľom tohto návodu je (po vylúčení triviálnych faktorov), aby každá korelácia dvoch premenných bola vysvetlená čo najmenším počtom faktorov. Na bežné používanie tohto

[3] Vhodnejší a výstižnejší je pojem transformácia ako rotácia. Ortogonálna transformácia faktorov (tzv. rotácia faktorov) je výpočtová operácia, ktorou sa z matice faktorových váh získava nová matica. Pojem rotácie sa do FA preniesol z geometrického zobrazenia transformácie faktorových váh. Ortogonálna transformácia je geometricky pevná (rigidná) rotácia q súradnicových osí v p-rozmernom priestore.

návodu je potrebné pojem „nulová“ faktorová váha nahradiť pojmom „malá“ faktorová váha.

Prvé faktorové riešenie získané napr. metódou PCA sa často ani nedá rozumne interpretovať. Je ďalej potrebné tieto faktory transformovať (rotovať) a tak získať zmysluplnejšie (interpretovateľnejšie) faktory. Väčšina metód rotácie sa snaží vo výsledku získať čo najviac faktorových váh blízkych nule a zároveň čo najviac ostatných (zvyšných) váh blízkych jednej.

Teória FA a štatistické programy poskytujú celý rad metód transformácie (rotácie) faktorov. Je potrebné sa rozhodnúť, či použijeme ortogonálnu (pravouhlú, kolmú) rotáciu alebo kosouhlú (šikmú) rotáciu. Ortogonálna rotácia vedie k riešeniu s nekorelovanými (nezávislými) faktormi. Prvky matice faktorových váh **A** možno interpretovať ako regresné koeficienty závislosti indikátorov od faktorov a tiež ako korelačné koeficienty medzi nimi. Kosouhlá rotácia vedie k získaniu závislých faktorov, čiže navyše poskytuje korelačnú maticu medzi faktormi. Kosouhlú rotáciu niektorí autori odmietajú, kým iní ju vítajú. Tvrdia, že pre prax sú reálnejšie korelované (závislé) spoločné faktory.

Faktorový model vychádzajúci z korelačnej matice indikátorov môžeme zapísať v maticovom tvare nasledovným spôsobom: $\mathbf{R} = \mathbf{A}^* (\mathbf{A}^*)^T + \mathbf{E}^*$.

Predpokladom faktorového modelu je lineárny vzťah medzi indikátormi a faktormi pre jeho jednoduchosť a výhody pri interpretácii faktorových koeficientov ako koeficientov korelácie ako aj predpoklad práve o *q* počte spoločných faktorov. Pri odhade parametrov FA modelu je veľa štatistických problémov (bližšie pozri [14], kapitola 3). Medzi tie najdôležitejšie patrí nejednoznačnosť FA modelu, odhad parametrov FA modelu, rotácia faktorov a taktiež problém ako určiť čo najmenší počet spoločných faktorov, ktoré by čo najlepšie vysvetľovali koreláciu medzi vstupnými premennými.

Na určenie počtu spoločných faktorov existuje celý rad objektívnych a subjektívnych rád a možností:

1. Začať FA metódou PCA a určiť počiatočný počet faktorov, ktorý je len orientačný, ale niekde treba začať.
2. Odhadom počtu spoločných faktorov môže byť počet vlastných čísiel redukovanej korelačnej matice väčších ako jedna.
3. Niekedy apriórne vieme z iných analýz alebo z teoretickej analýzy problematiky, koľko spoločných faktorov je potrebných na charakterizovanie vzťahov medzi indikátormi.
4. Spoločné faktory by mali vysvetliť čo najviac celkového rozptylu. V exaktných vedách by to malo byť 90 – 95 % a v spoločenských vedách viac ako 60 – 70 %.
5. Spoločné faktory by mali reprezentovať viac ako 90 % celkovej komunality, ktorá je daná súčtom komunalít všetkých p indikátorov.
6. Môžeme použiť, tak ako v PCA, graf „scree plot“, ktorý zobrazuje počet faktorov na osi x a na osi y percento vysvetlenej variability, t. j. hodnoty vlastných čísiel redukovanej kovariančnej, resp. korelačnej matice. Za optimálny počet faktorov treba považovať hodnotu na x-ovej osi pred bodom zlomu na krivke vlastných čísiel.

7. Do konečného riešenia sa nemajú zahŕňať tzv. triviálne faktory. Triviálne faktory sú také, ktoré významne korelujú len s jedným indikátorom. Lepšie je takýto indikátor z FA vylúčiť a začať znovu. To neznamená, že daná premenná je nepodstatná, ale nehodí sa do faktorovej analýzy a môže sa brať do úvahy samostatne.

Pri odhade parametrov FA modelu rôznymi metódami je podmienka, aby matica **E** (matica špecifických faktorov) bola pozitívne definitná. Niektoré riešenia FA modelu vedú však k nevhodným (nesprávnym) riešeniam. Takéto prípady dostali názov Heywoodove prípady. Heywoodov prípad nastáva vtedy, keď počas iteračného procesu najmenej jeden špecifický faktor je odhadovaný ako nepozitívny (rovný nule alebo záporný). Matica **E** teda nemusí byť nulovou maticou, ale obsahuje najmenej jeden záporný prvok, a to vedie ku komunalitám väčším ako 1, čo nie je vo FA prípustné.

Nie je však pravda, že všetky nevhodné riešenia vyvoláva Heywoodov prípad. Veľa empirických štúdií dokázalo, že príčinami nesprávnych riešení sú:

- malý rozsah vzorky,
- nízky počet vstupných premenných,
- príliš vysoký počet extrahovaných spoločných faktorov,
- príliš nízky počet extrahovaných spoločných faktorov,
- zlá voľba počiatočného odhadu komunalít,
- použitie nevhodného faktorového modelu pre údaje,
- prítomnosť „outlierov“ v dátach.

Faktorovú analýzu môžeme do určitej miery považovať za rozšírenie metódy PCA. Obe metódy majú spoločné, ale aj odlišné vlastnosti, majú svoje výhody, ale aj nedostatky, ktoré môžeme zhrnúť do nasledovných bodov:

- Obe metódy nemá význam použiť, keď pôvodné premenné nie sú korelované, lebo FA nemá čo vysvetliť a PCA vedie k výsledným hlavným komponentom, ktoré sú zhodné s pôvodnými premennými.
- FA sa pokúša vysvetliť kovariancie a korelácie pôvodných premenných pomocou niekoľkých málo spoločných faktorov (tzv. latentných premenných). PCA vysvetľuje len rozptyl premenných.
- Výpočty v PCA sú jednoduchšie a priamočiarejšie. Model FA má veľa predpokladov a výpočty sú náročnejšie a zložitejšie. Bolo vyvinutých veľa metód odhadu parametrov FA modelu a veľa spôsobov rotácie, ktoré vedú k rôznym riešeniam.
- Medzi nedostatky PCA patrí, že je závislá od zmien merných jednotiek premenných. Riešenie FA modelu metódou maximálnej vierohodnosti je invariantné k takýmto zmenám, čo je výhodou FA.

Pri posúdení vhodnosti vstupných údajov pre PCA a FA sa vychádza z korelačnej matice vstupných premenných. Použitie týchto metód vyžaduje významnú vzájomnú koreláciu vstupných premenných. Okrem korelačnej matice je vhodné použiť aj KMO kritérium (Kaiser – Meyer – Olkin), ktoré je založené na porovnaní jednoduchých a parciálnych koeficientov korelácie. KMO sa vypočíta podľa vzorca ([14]):

$$KMO = \frac{\sum^{p}\sum_{i \neq j} r_{ij}^2}{\sum^{p}\sum_{i \neq j} r_{ij}^2 + \sum^{p}\sum_{i \neq j} r_{parc.,ij}^2} \quad (13)$$

KMO štatistika sa počíta ako celková miera adekvátnosti (vhodnosti) výberových dát pre FA (v SAS EG výstupoch ako MSA – Kaiser's measure of sampling adequacy overall) aj ako čiastková miera adekvátnosti jednotlivých indikátorov. Je to miera homogenity premenných. Hodnoty KMO sa netestujú, ale v praxi sa používa tabuľka odporúčaní podľa Kaisera a Ricea (1974) (tabuľka č. 7).

Tabuľka č. 7: FA - Odporúčania pre hodnoty KMO miery

Hodnota KMO štatistiky	Odporúčanie pre adekvátnosť výberových dát
≥ 0,9	Vynikajúce
<0,8; 0,9)	Chvályhodné
<0,7; 0,8)	stredne užitočné
<0,6; 0,7)	Priemerné
<0,5; 0,6)	Slabé
<0,5	Nedostatočné

Zdroj: [14] podľa Kaisera a Ricea (1974)

2.2.3 Zhluková analýza

Zhluková analýza (angl. cluster analysis – CA) zahŕňa širokú škálu metód a postupov, ktoré sa používajú pri riešení problémov typológie objektov a ich klasifikácie. Cieľom zhlukovej analýzy je rozklad súboru objektov opísaných viacerými ukazovateľmi na niekoľko relatívne rovnorodých podmnožín (zhlukov, tried, segmentov) tak, aby objekty patriace do toho istého zhluku si boli čo najviac podobné a objekty patriace do rôznych zhlukov si boli podobné čo najmenej.

Prvotnú úlohu zhlukovej analýzy môžeme matematicky sformulovať takto: ide o zoskupenie objektov X_i (i = 1, 2, …, n) do zhlukov C_1, C_2, … , C_q (2 ≤ q ≤ n) tak, aby objekty patriace do toho istého zhluku si boli blízke, podobné a objekty patriace do rôznych zhlukov si boli vzdialené, odlišné. V porovnaní s inými metódami sú predmetom zhlukovej analýzy objekty (štatistické jednotky, pozorovania), a nie premenné (štatistické znaky, ukazovatele). Neskôr sa metódy zhlukovania použili aj na premenné, čo znamenalo aplikáciu špecifického prístupu faktorovej analýzy v tejto oblasti. Zhlukovanie premenných sa dnes využíva v data miningu, kde sa často stretávame s veľký počtom skorelovaných vstupných premenných.

Pri aplikovaní zhlukovej analýzy v praxi musí analytik riešiť tieto problémové okruhy:

- výber miery podobnosti alebo vzdialenosti štatistických jednotiek (objektov),
- výber druhu zhlukovacieho postupu,
- výber zhlukovacej metódy,
- určenie počtu významných zhlukov,
- interpretácia zhlukov.

V teórii sú definované 4 skupiny mier podobnosti medzi objektmi (bližšie pozri [14], kapitola 5). V počítačových štatistických programoch je najčastejšie implementovaná euklidovská vzdialenosť.

Podľa spôsobu zoskupovania objektov do zhlukov rozoznávame hierarchické (aglomeračné a divízne) a nehierarchické postupy. V procese hierarchického zhlukovania nie je potrebné dopredu poznať optimálny počet zhlukov. Ten sa určuje dodatočne na základe grafu dendrogramu alebo rôznych mier homogenity, resp. heterogenity zhlukovania. Medzi miery homogenity patria semiparciálny koeficient determinácie (semipartial R-Squared – SPRQ) a vzdialenosť zhlukov (cluster distance – CD). Koeficient determinácie (R-Squared – RSQ) je zase miera heterogenity zhlukov. Čím je táto miera bližšie k číslu 1, tým vyššia je medziskupinová variabilita.

V prípade nehierarchických postupov musí byť vopred známy (daný) počet zhlukov (zhlukovanie s konštantným počtom zhlukov), pričom tento počet sa môže počas zhlukovania meniť (zhlukovanie s optimalizovaným počtom zhlukov, napr. metóda K- means). V princípe pre malé súbory objektov sú vhodné hierarchické postupy a pre veľké objemy zase nehierarchické postupy zhlukovania. Tieto postupy sa navzájom doplňujú.

Pre oba zhlukovacie postupy bolo vyvinutých niekoľko zhlukovacích metód, ktoré súvisia so zvoleným spôsobom vyjadrenia vzdialenosti. Pri hierarchických postupoch zhlukovania môžeme použiť napr. tieto metódy: metódu najbližšieho suseda (angl. single linkage), metódu najvzdialenejšieho suseda (angl. complete linkage), metódu priemernej vzdialenosti (angl. average linkage), centroidnú metódu (angl. centroid method), mediánovú metódu (angl. median method) a Wardova metóda (angl. Ward's minimum variance method). Wardova metóda je obľúbená u štatistikov, lebo neoptimalizuje vzdialenosť medzi zhlukmi, ale minimalizuje heterogenitu (rozptyl) vo vnútri zhlukov. Vyžaduje vyjadrenie vzdialenosti objektov pomocou štvorca euklidovskej vzdialenosti. Pri použití rôznych zhlukovacích postupov a metód dostávame rôzne výsledky. Nedá sa všeobecne povedať, ktorá z uvedených metód je najlepšia. Všetky sú však citlivé na prítomnosť extrémnych hodnôt (angl. outliers) v dátach.

Zhluková analýza vyžaduje splnenie určitých predpokladov. Práve na základe vlastností dátovej množiny sa rozhodujeme, ktorá zhlukovacia metóda je pre konkrétny prípad najvhodnejšia. Dáta by nemali obsahovať odľahlé pozorovania (angl. outliers) a chýbajúce hodnoty. Premenné musia byt' často štandardizované (normované), aby sa odstránil vplyv rôznych merných jednotiek. Ak by sme premenné nenormovali, prejavili by sa v našej zhlukovej analýze s rôznou dôležitosťou, ktorá by sa odvíjala od ich merných jednotiek, nie charakteru problému. Po odstránení uvedených problémov vyberáme vhodnú metódu na základe rozsahu dátového súboru z hľadiska počtu pozorovaní a podielu počtu premenných k počtu pozorovaní, charakteru premenných (intervalové alebo poradové premenné) a prípadnej apriórnej informácie o počte zhlukov.

Výsledky zhlukovej analýzy negatívne ovplyvňuje aj závislosť medzi vstupnými premennými. Tento jav je možno eliminovať použitím PCA alebo FA, ktoré vytvárajú nové nezávislé premenné (hlavné komponenty alebo spoločné faktory). Takisto je možné

riešenie, že ako vzdialenosť objektov vyberieme Mahalanobisovu vzdialenosť, ktorá dokáže odstrániť vplyv korelácie medzi premennými. V softvéri SAS EG je však možno použiť len euklidovskú vzdialenosť, ktorá to nedokáže.

2.2.4 Použitie analýzy hlavných komponentov a faktorovej analýzy na výpočet HDI

Pri výpočte HDI pomocou PCA je potrebné vychádzať z pôvodných čiastkových ukazovateľov a tie musia byť významne korelované. Ako uvádzame vyššie (tabuľka č. 1), odborníci na skúmanie problematiky v oblasti ľudského rozvoja ako čiastkové indikátory na výpočet HDI vybrali nasledujúce 4 merateľné ukazovatele (indikátory):

- očakávaná dĺžka života pri narodení (LE),
- očakávaný počet rokov štúdia (Exp_Edu),
- priemerný počet rokov štúdia (Mean_Edu) a
- hrubý národný produkt na obyvateľa v PPP $ v cenách roku 2011 (GNI).

Na účely článku sme použili údaje o týchto štyroch merateľných ukazovateľoch za 34 európskych krajín, ktoré sú dostupné na webovej stránke UNDP [16]. Analyzovali sme dva roky, rok 2007 a 2018. Na štatistické výpočty bol použitý softvér SAS Enterprise Guide.

Popisná štatistika 4 vstupných ukazovateľov (indikátorov) v rokoch 2007 a 2018 za 34 krajín EÚ je v tabuľke č. 8. Z výsledkov je zrejmé, že hodnoty sledovaných ukazovateľov v roku 2018 oproti roku 2007 v analyzovaných 34 európskych krajinách vzrástli. Priemerná očakávaná dĺžka života sa zvýšila zo 77,5 na 80 rokov, t. j. o 2,5 roka. Vzrástli aj oba ukazovatele počtu rokov štúdia. Vzrast HDP na obyvateľa (GNI) za toto obdobie bol v priemere až o 2 894 PPP $ v cenách roku 2011. Je potrebné si však všimnúť, že maximálna hodnota HDP v roku 2007 bola až 91 519 a dosiahlo ju Luxembursko. Tieto výsledky signalizujú, že kvalita života v krajinách za sledované obdobie vzrástla.

Tabuľka č. 8: Hodnoty popisných štatistík 4 indikátorov pre 34 vybraných európskych krajín (roky 2007 a 2018)

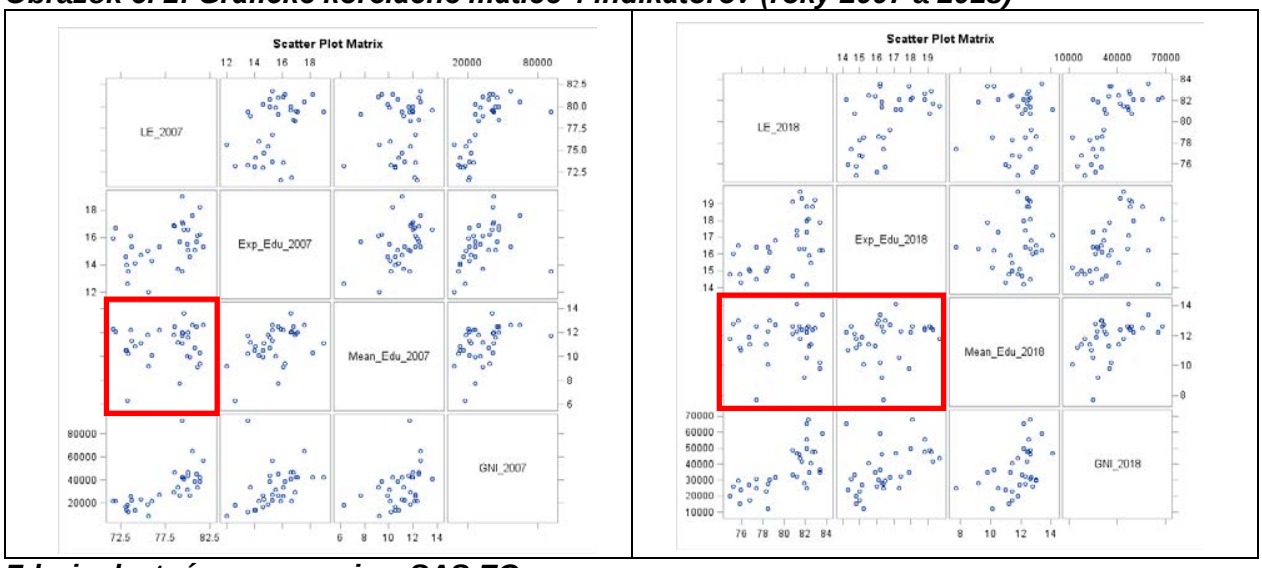
Rok 2007	N	Mean	Std Dev	Median	Min	Max
LE	34	77.5	3.2	79.0	71.6	81.7
Exp_Edu	34	15.5	1.5	15.4	12.0	19.0
Mean_Edu	34	11.0	1.5	11.3	6.3	13.6
GNI	34	33177	16850	30363	8671	91519

Rok 2018	N	Mean	Std Dev	Median	Min	Max
LE	34	80.0	2.7	81.2	74.9	83.6
Exp_Edu	34	16.6	1.6	16.3	14.2	19.7
Mean_Edu	34	11.8	1.3	12.2	7.7	14.1
GNI	34	36071	13953	32622	12300	68059

Zdroj: vlastné spracovanie v SAS EG

Na základe výsledkov korelačných matíc 4 indikátorov (obrázok č. 1) môžeme konštatovať, že v roku 2007 nie je významná korelácia medzi dvoma ukazovateľmi: očakávanou dĺžkou života pri narodení (LE) a priemerným počtom rokov štúdia (Mean_Edu) (r = 0,11 a p-hodnota = 0,537). V roku 2018 až 2 korelačné koeficienty nie sú významné, okrem nevýznamnej korelácie medzi očakávanou dĺžkou života pri narodení (LE) a priemerným počtom rokov štúdia (Mean_Edu) (r = 0,036 a p-hodnota = 0,842), je aj nevýznamná korelácia medzi priemerným počtom rokov štúdia (Mean_Edu) a očakávaným počtom rokov štúdia (Exp_Edu), (r = 0,181 a p-hodnota = 0,307). Toto sa prejaví vo veľkosti vysvetlenej variability hlavnými komponentmi, resp. faktormi.

Obrázok č. 1: Grafické korelačné matice 4 indikátorov (roky 2007 a 2018)



Zdroj: vlastné spracovanie v SAS EG

Aj na základe veľkosti hodnôt celkových KMO mier vidíme, že ukazovatele patria do kategórie slabé na použitie PCA, resp. FA (tabuľka č. 7). Celková KMO miera dosahuje hodnoty len tesne nad 0,5 (tabuľka č. 9: KMO = 0,504 za r. 2007 a KMO = 0,523 za r. 2018. Spôsobili to hlavne nízke hodnoty individuálnych KMO mier pre 2 ukazovatele (sú pod hodnotou 0,5): očakávaná dĺžka života (LE) a priemerný počet rokov štúdia (Mean_Edu).

Tabuľka č. 9: Hodnoty KMO mier pre 4 indikátory (roky 2007 a 2018)

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.504				Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.523			
LE_2007	Exp_Edu_2007	Mean_Edu_2007	GNI_2007	LE_2018	Exp_Edu_2018	Mean_Edu_2018	GNI_2018
0.469	0.614	0.414	0.530	0.498	0.820	0.301	0.530

Zdroj: vlastné spracovanie v SAS EG

Pri PCA sme vychádzali z korelačnej matice, lebo vstupné indikátory sú vyjadrené v rôznych merných jednotkách. Z dosiahnutých výsledkov v SAS EG vidíme, že 1. hlavný komponent (PRIN1), ktorého vlastné číslo je vyššie ako 1 (t. j. priemer všetkých vlastných čísiel (angl. Eigenvalues)), vysvetľuje len necelých 60 % celkovej variability 4 indikátorov (obrázok č. 2 a tabuľka č. 10: 56,1 % za r. 2007 a 57,3 % za r. 2018. V roku 2018 aj vlastné číslo pre 2. hlavný komponent (PRIN2) je ešte vyššie ako 1 (1,006) a vysvetľuje ešte vysoký podiel variability premenných (až 25,2 %).

Aj na základe odhadnutých komponentových váh (tabuľka č. 11) je zrejmé, že PCA metóda nie je najlepším riešením na konštrukciu integrálneho ukazovateľa pre ľudský rozvoj, čiže na výpočet HDI. Komponentové váhy PRIN1 pre všetky vstupné ukazovatele sa pohybujú len okolo hodnoty 0,5 a v roku 2018 sa vyskytla veľmi nízka váha pre indikátor Mean_Edu (len 0,3). V ďalších stĺpcoch sa vyskytujú zase vysoké komponentové váhy pre PRIN2 až PRIN4 a aj tieto hlavné komponenty vysvetľujú ešte vysoký podiel variability údajov (nad 5 %, pozri tabuľku č. 10).

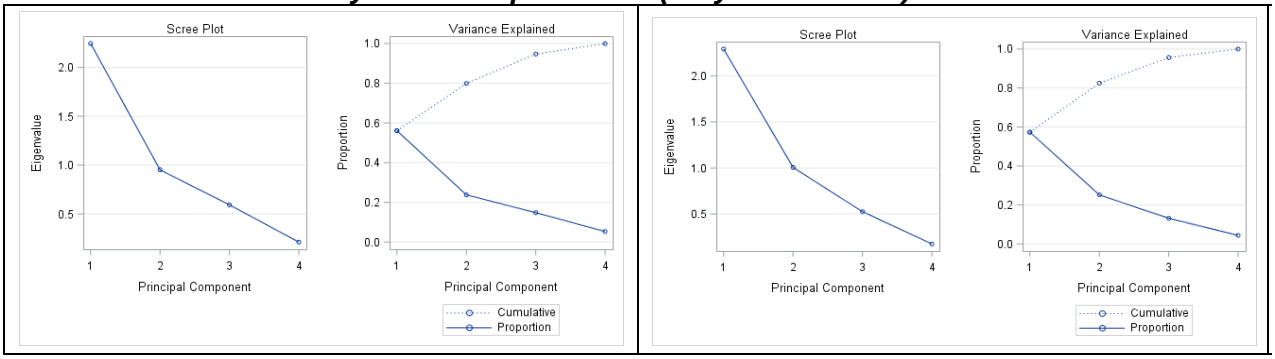
Tabuľka č. 10: PCA – hodnoty vlastných čísiel (roky 2007 a 2018)

Eigenvalues of the Correlation Matrix (2007)				
	Eigenvalue	Difference	Proportion	Cumulative
1	**2.244**	**1.293**	**0.561**	**0.561**
2	**0.951**	**0.359**	**0.238**	**0.799**
3	0.593	0.380	0.148	0.947
4	0.212		0.053	1.000

Eigenvalues of the Correlation Matrix (2018)				
	Eigenvalue	Difference	Proportion	Cumulative
1	**2.292**	**1.286**	**0.573**	**0.573**
2	**1.006**	**0.480**	**0.252**	**0.825**
3	0.526	0.351	0.132	0.956
4	0.175		0.044	1.000

Zdroj: vlastné spracovanie v SAS EG

Obrázok č. 2: PCA – Grafy Sree Plot pre PRIN1 (roky 2007 a 2018)



Zdroj: vlastné spracovanie v SAS EG

Tabuľka č. 11: PCA – hodnoty komponentových váh (roky 2007 a 2018)

Eigenvectors				
	PRIN1	PRIN2	PRIN3	PRIN4
LE_2007	**0.517**	**-0.572**	-0.091	**0.630**
Exp_Edu_2007	**0.491**	0.297	**-0.781**	-0.246
Mean_Edu_2007	**0.409**	**0.721**	0.412	0.379
GNI_2007	**0.570**	-0.254	0.461	**-0.631**

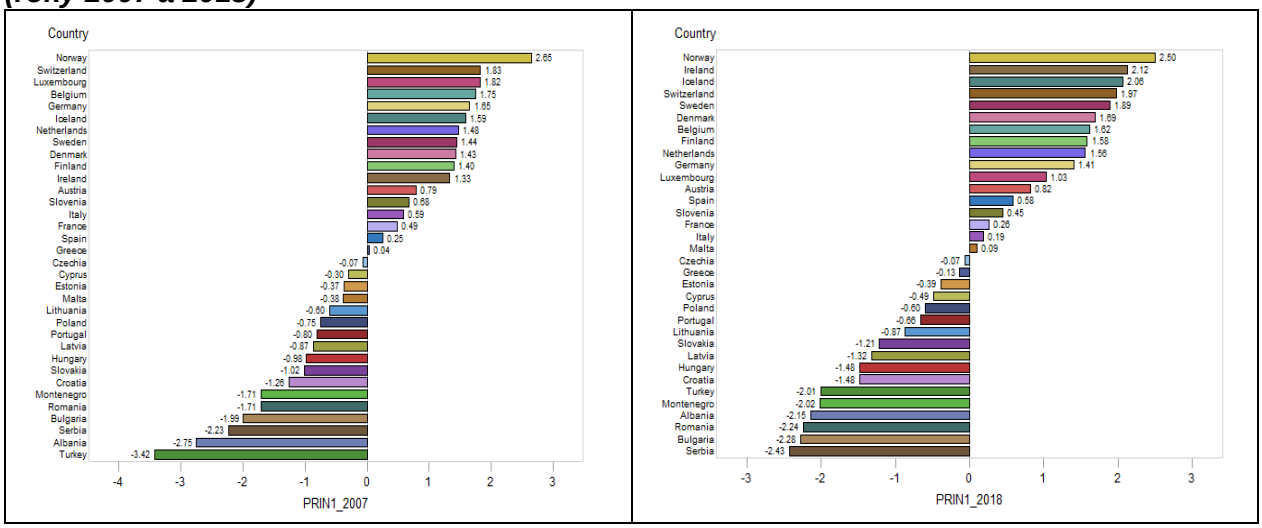
Eigenvectors				
	PRIN1	PRIN2	PRIN3	PRIN4
LE_2018	**0.546**	-0.411	-0.390	**0.617**
Exp_Edu_2018	**0.505**	-0.232	**0.828**	-0.078
Mean_Edu_2018	**0.300**	**0.872**	0.096	0.376
GNI_2018	**0.598**	0.134	-0.392	**-0.687**

Zdroj: vlastné spracovanie v SAS EG

Ak aj napriek tomu použijeme hodnoty komponentových skóre PRIN1 pre analyzované krajiny na zostavenie ich poradia podľa úrovne ľudského rozvoja, výsledok môžeme vidieť na obrázku č. 3. V roku 2007 dosiahlo najvyššiu hodnotu Nórsko (kladná hodnota PRIN1 = 2,65) a najnižšiu hodnotu Turecko (záporná hodnota PRIN1 = -3,42). V roku 2018 sa poradie zmenilo hlavne na konci rebríčka. Nórsko síce opäť dosiahlo najvyššiu hodnotu (kladná hodnota PRIN1 = 2,50), ale najnižšiu hodnotu malo Srbsko (záporná hodnota PRIN1 = -2,43). Turecko si polepšilo svoju pozíciu v oblasti ľudského rozvoja (záporná hodnota PRIN1 = -2,01). Záverom môžeme konštatovať, že výsledné poradie 34 krajín na základe hodnôt PRIN1 (metóda PCA) ktoré vychádza zo 4 merateľných indikátorov je podobné aké bolo na základe hodnôt HDI poskytnutých metodikou UDNP (pozri príloha č. 2). Len poradie Luxemburska v roku 2007 je výrazne iné (3. miesto), lebo výsledné hodnoty PRIN1 ovplyvnila hlavne extrémne vysoká hodnota ukazovateľa hrubý domáci produkt (GNI).

Pri výpočte HDI pomocou faktorovej analýzy z pôvodných 4 čiastkových ukazovateľov by sme dostali podobné výsledky ako pri PCA. Prvý spoločný faktor však vysvetľuje len necelých 60 % závislosti premenných. Výsledky v článku neuvádzame.

Obrázok č. 3: PCA – poradie 34 krajín podľa hodnôt komponentových skóre pre PRIN1 (roky 2007 a 2018)



Zdroj: vlastné spracovanie v SAS EG

Ak ako vstupné premenné pre PCA, resp. FA použijeme 3 indexy (dimenzie), ktoré boli vypočítané podľa metodiky UDNP, tak postup aj výsledok bude iný. Ako je zrejmé z korelačných matíc (tabuľka č. 12, obrázok č. 4), indexy sú skorelované, ale aj medzi nimi sa objavila slabá, resp. nevýznamná korelácia: konkrétne medzi indexom vzdelania (IE) a indexom zdravia (IH) za obidva sledované roky (2007: r = 0,288 a 2018: r=0,372). Môže to znamenať, že vstupné premenné nie sú vhodné na použitie PCA, resp. FA. Hodnoty celkových aj individuálnych KMO mier sú však vysoké (tabuľka č. 13), preto pomocou FA vypočítame z nich integrálny ukazovateľ HDI ako prvý spoločný faktor (Factor1). Uprednostníme použitie FA namiesto PCA, lebo výsledky FA sa dajú v SAS EG rotovať a to nám umožní lepšiu interpretáciu HDI.

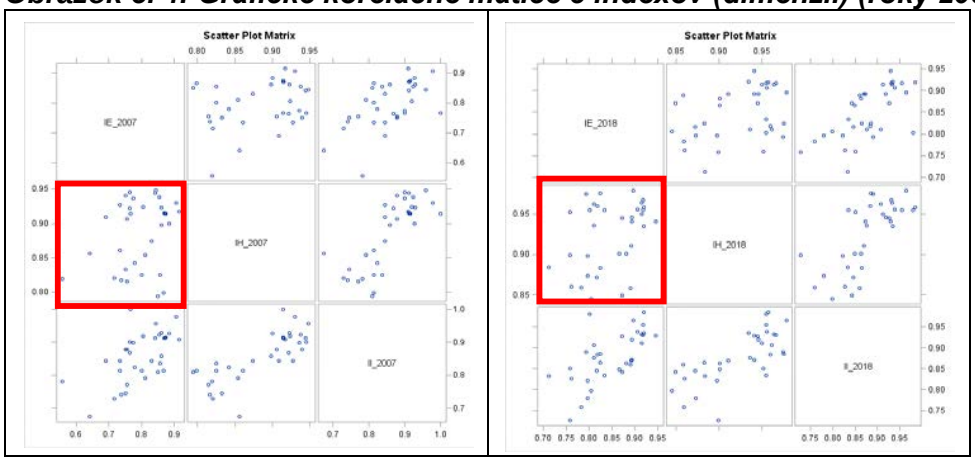
Tabuľka č. 12: Hodnoty Pearsonových koeficientov korelácie medzi 3 indexmi (dimenziami) na výpočet HDI pre 34 vybraných európskych krajín (roky 2007 a 2018)

Pearson Correlation Coefficients, N = 34			
Prob > |r| under H0: Rho=0			
	IE_2007	IH_2007	II_2007
IE_2007	1	0.2880	0.5949
		0.0986	0.0002
IH_2007	0.2880	1	0.7803
	0.0986		<.0001
II_2007	0.5949	0.7803	1
	0.0002	<.0001	

Pearson Correlation Coefficients, N = 34			
Prob > |r| under H0: Rho=0			
	IE_2018	IH_2018	II_2018
IE_2018	1	0.3716	0.6576
		0.0305	<.0001
IH_2018	0.3716	1	0.7187
	0.0305		<.0001
II_2018	0.6576	0.7187	1
	<.0001	<.0001	

Zdroj: vlastné spracovanie v SAS EG

Obrázok č. 4: Grafické korelačné matice 3 indexov (dimenzií) (roky 2007 a 2018)



Vysvetlivky: IE – index vzdelania, IH – index zdravia, II – príjmový index
Zdroj: vlastné spracovanie v SAS EG

Tabuľka č. 13: Hodnoty KMO mier pre 3 indexy (roky 2007 a 2018)

Kaiser's Measure of Sampling Adequacy:		
Overall MSA = 0.760		
IE_2007	IH_2007	II_2007
0.8801	0.7465	0.6801

Kaiser's Measure of Sampling Adequacy:		
Overall MSA = 0.779		
IE_2018	IH_2018	II_2018
0.8544	0.7851	0.7103

Zdroj: vlastné spracovanie v SAS EG

Pretože vstupné indexy sú bezrozmerné ukazovatele, tak pri výpočtoch PCA, resp. FA je potrebné vychádzať z kovariančnej matice. Hodnoty vlastných čísiel z kovariančnej matice sú uvedené v ďalšej tabuľke. Je zrejmé, že 1. spoločný faktor (Factor1) vysvetľuje takmer 100 % variability údajov (tabuľka č. 14: 2007: 99,79 % a 2018: 99,89 % a obrázok č. 5: Graf Scree Plot). Faktorové váhy všetkých 3 indexov sú veľmi vysoké (blízke číslu 1) v tomto spoločnom faktore (tabuľka č. 15).

Takto získaný Factor1 bude preto vhodný ako integrálny ukazovateľ HDI. Poradie 34 krajín na základe hodnôt faktorových skóre pre Factor1 sme zobrazili na obrázku č. 6) a hodnoty uvádzame aj v prílohe č. 3. Z výsledkov je zrejmé, že poradie krajín na základe hodnôt HDI vypočítaných metodikou UNDP a hodnôt faktorových skóre pre týmto spôsobom získaný Factor1 sú si veľmi podobné (príloha č. 3).

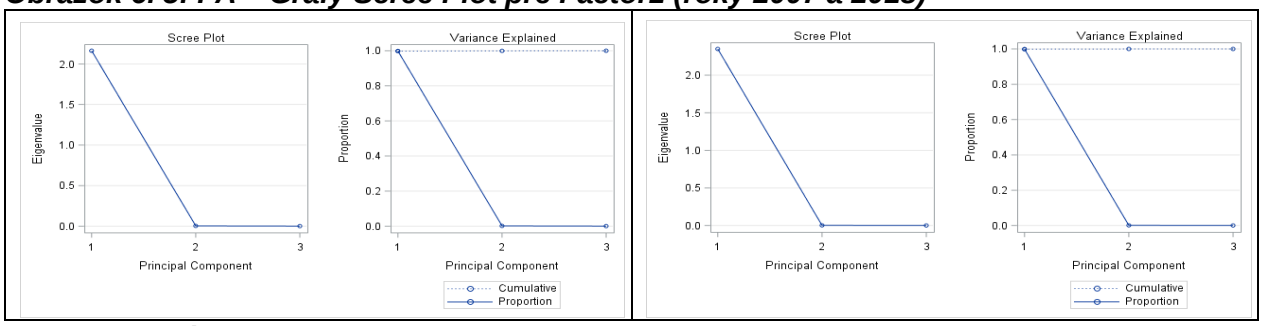
Tabuľka č. 14: Hodnoty vlastných čísiel z kovariančnej matice pre 3 indexy (roky 2007 a 2018)

Eigenvalues of the Uncorrected Covariance Matrix (2007)				
	Eigenvalue	Difference	Proportion	Cumulative
1	**2.1627**	**2.1593**	**0.9979**	**0.9979**
2	0.0034	0.0024	0.0016	0.9995
3	0.0010		0.0005	1.0000

Eigenvalues of the Uncorrected Covariance Matrix (2018)				
	Eigenvalue	Difference	Proportion	Cumulative
1	**2.3480**	**2.3463**	**0.9989**	**0.9989**
2	0.0018	0.0010	0.0008	0.9996
3	0.0008		0.0004	1.0000

Zdroj: vlastné spracovanie v SAS EG

Obrázok č. 5: FA – Grafy Scree Plot pre Factor1 (roky 2007 a 2018)



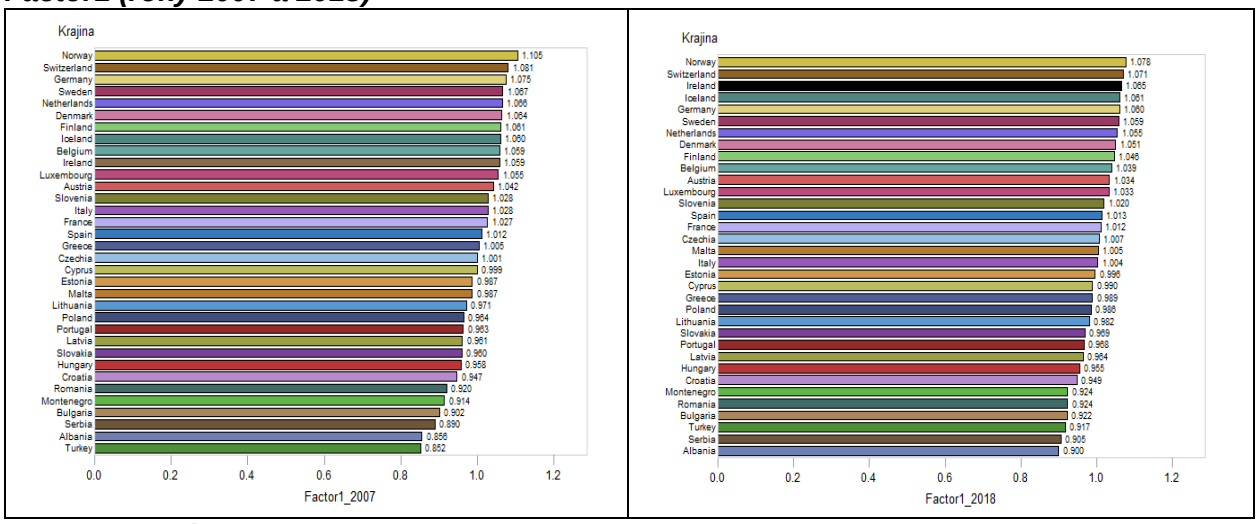
Zdroj: vlastné spracovanie v SAS EG

Tabuľka č. 15: Hodnoty faktorových váh pre Factor1 (roky 2007 a 2018)

Factor Pattern		Factor Pattern	
	Factor1		**Factor1**
IE_2007	0.9982	**IE_2018**	0.9992
IH_2007	0.9991	**IH_2018**	0.9995
II_2007	0.9995	**II_2018**	0.9996

Zdroj: vlastné spracovanie v SAS EG

Obrázok č. 6: FA – poradie 34 vybraných európskych krajín podľa faktorových skóre pre Factor1 (roky 2007 a 2018)

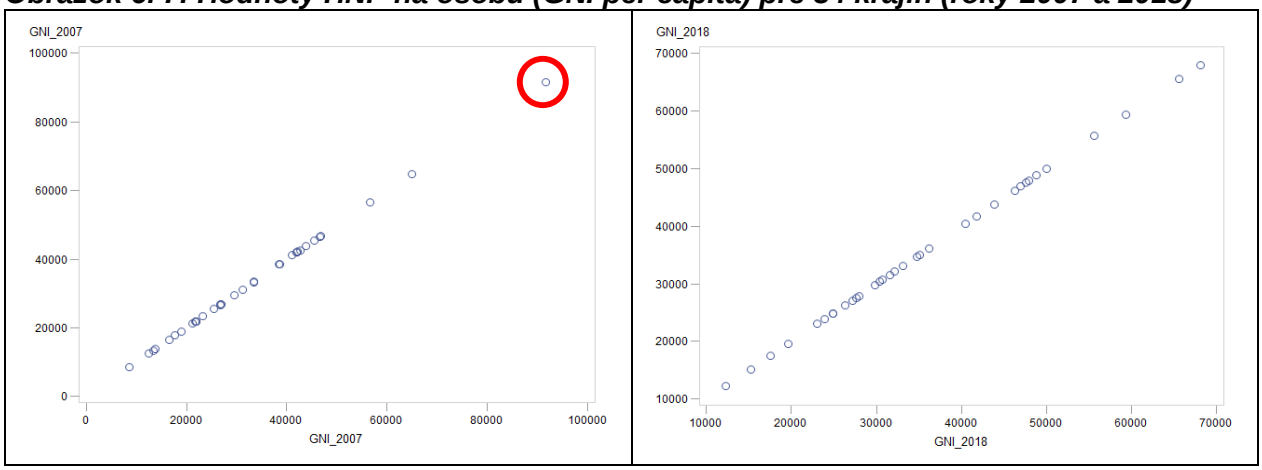


Zdroj: vlastné spracovanie v SAS EG

2.2.5 Použitie zhlukovej analýzy na zaradenie krajín do skupín podľa úrovne HDI

V tejto časti sa pomocou zhlukovej analýzy pokúsime zoskupiť 34 krajín EÚ do vhodného počtu skupín (zhlukov) na základe 4 merateľných ukazovateľov HDI. Zhluková analýza kladie na údaje viacero požiadaviek. Jednou z nich je, aby vstupné premenné neboli navzájom korelované. Druhou požiadavkou je, aby sa v údajoch nevyskytovali extrémne hodnoty (angl. outliers). Veľmi dôležité je aj to, aby vstupné premenné boli vyjadrené v rovnorodých meracích jednotkách. Na základe predchádzajúcich zistení môžeme konštatovať, že tieto požiadavky naše údaje nespĺňajú, lebo vstupné premenné sú korelované, v dátach sa vyskytujú aj extrémne hodnoty, napríklad hodnota HNP (GNI) v Luxembursku za rok 2007 (obrázok č. 7,) a pôvodné 4 ukazovatele sú rôznorodé.

Obrázok č. 7: Hodnoty HNP na osobu (GNI per capita) pre 34 krajín (roky 2007 a 2018)



Zdroj: vlastné spracovanie v SAS EG na základe údajov: http://hdr.undp.org/en/indicators/141706#

Prvý a tretí problém vieme vyriešiť pomocou PCA alebo FA. Použijeme FA, lebo umožňuje aj rotáciu, čiže nájsť vhodné riešenie z množstva dobrých riešení. Výpočty urobíme z korelačnej matice metódou PCA s ortogonálnou Varimax rotáciou faktorov. Získame tak 4 triviálne nekorelované a normované spoločné faktory v každom sledovanom roku zvlášť (tabuľka č. 16: Factor1 až Factor4).

Tabuľka č. 16: FA – faktorové váhy po ortogonálnej Varimax rotácii (roky 2007 a 2008)

Rotated Factor Pattern				
	Factor1	**Factor2**	**Factor3**	**Factor4**
LE_2007	-0.0032	**0.9145**	0.2086	0.3467
Exp_Edu_2007	0.2249	0.1926	**0.9465**	0.1285
Mean_Edu_2007	**0.9601**	0.0065	0.2152	0.1784
GNI_2007	0.2280	0.3879	0.1448	**0.8812**

Rotated Factor Pattern				
	Factor1	**Factor2**	**Factor3**	**Factor4**
LE_2018	-0.0334	**0.9029**	0.2843	0.3208
Exp_Edu_2018	0.0834	0.2502	**0.9482**	0.1770
Mean_Edu_2018	**0.9797**	-0.0135	0.0730	0.1864
GNI_2018	0.3040	0.4182	0.2357	**0.8229**

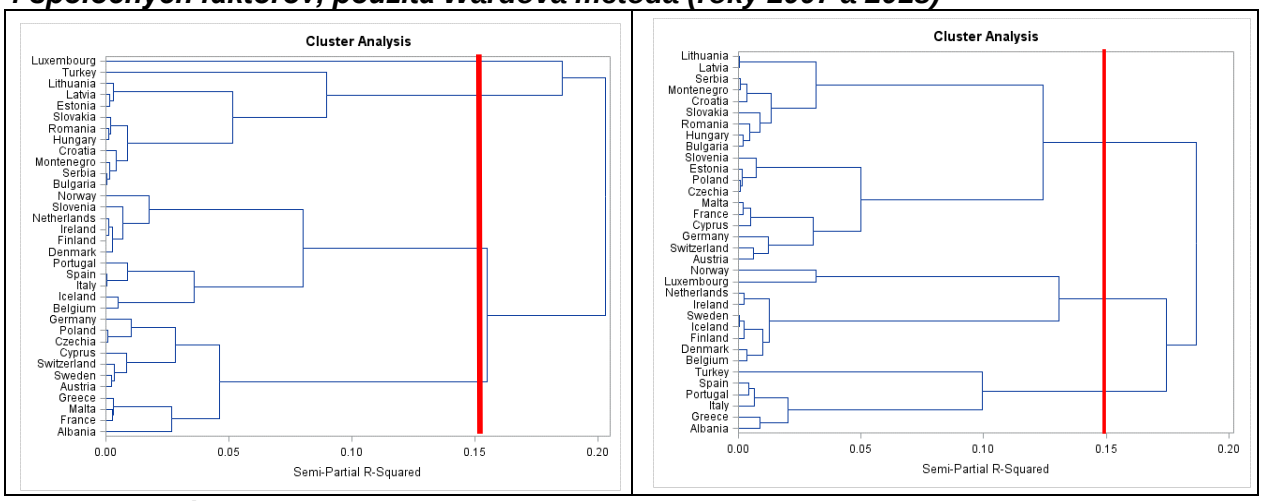
Zdroj: vlastné spracovanie v SAS EG

Tieto nekorelované triviálne faktory použijeme v zhlukovej analýze namiesto 4 merateľných, ale vzájomne skorelovaných ukazovateľov, aby sme získali homogénne skupiny (zhluky) krajín, ktoré sú si navzájom blízke v úrovni ľudského rozvoja.

V SAS EG sme použili z ponúkaných metód hierarchický postup Wardovou metódou. Na výsledných grafoch zhlukovania – dendrogramoch (obrázok č. 8), je zrejmé, že v analyzovaných rokoch 2007 a 2018 sa zhluky krajín líšia. Ak použijeme na určenie počtu zhlukov rovnakú hodnotu štatistiky semi-parciálny koeficient determinácie (SPRQ) na úrovni 0,15, tak v roku 2007 dostaneme 4 zhluky a v roku 2018 len 3 zhluky. V roku 2007 je Luxembursko zaradené do samostatného zhluku, lebo jeho HNP na hlavu (GNI per capita) je vysoké (nad 75 000 $), čiže ide o extrémnu hodnotu, ktorá negatívne ovplyvnila výsledky zhlukovania. V roku 2018 je už Luxembursko zaradené do zhluku spolu s Nórskom, lebo jeho hodnota HNP sa znížila (pod 75 000 $).

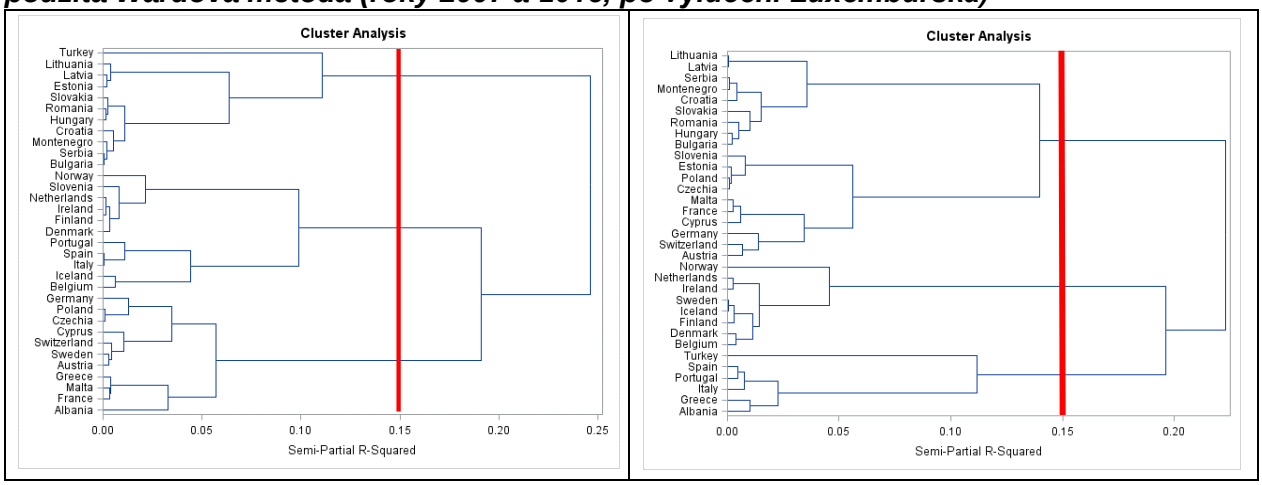
Po vylúčení Luxemburska z analýzy dostávame zmenené dendrogramy (obrázok č. 9). Pri voľbe tej istej hodnoty (0,15) pre štatistiku semi-parciálny koeficient determinácie (SPRQ) dostávame zhodne 3 zhluky za obidva sledované roky.

Obrázok č. 8: CA – dendrogramy zhlukovania 34 vybraných európskych krajín na základe 4 spoločných faktorov, použitá Wardova metóda (roky 2007 a 2018)



Zdroj: vlastné spracovanie v SAS EG

Obrázok č. 9: CA – dendrogramy zhlukovania 33 krajín na základe 4 spoločných faktorov, použitá Wardova metóda (roky 2007 a 2018, po vylúčení Luxemburska)



Zdroj: vlastné spracovanie v SAS EG

S výsledkami zhlukovej analýzy nemôžeme byť spokojní, lebo zhluky sú nelogické. Je to spôsobené výberom len 4 ukazovateľov na meranie úrovne HDI. Druhý dôvod je veľkosť a štruktúra použitých údajov za vybrané krajiny. Ide o pomerne malý súbor štatistických jednotiek (n = 34) a premenné majú zošikmené rozdelenia, lebo v každej zo štyroch premenných sa nachádzajú nejaké vybočené, resp. extrémne hodnoty.

Ďalší dôvod je, že sme vylúčili Luxembursko až v zhlukovej analýze. Správny postup je, že extrémne hodnoty treba vylúčiť už pri faktorovej analýze, lebo narušíme normovanie vstupných premenných, čiže faktorov (pozri tabuľka č. 17), ktoré potom používame v zhlukovej analýze. Po úprave výpočtov vo faktorovej analýze (vylúčili sme Luxembursko, n = 33) a po opätovnom uskutočnení zhlukovej analýzy s takto upravenými triviálnymi faktormi sme dostali nové dendrogramy (obrázok č. 10).

Tabuľka č. 17: FA – popisná štatistika faktorových skóre pre 34 a 33 krajín (roky 2007 a 2008)

rok 2007	Mean	Std Dev	Min	Max	N	Median
Factor1	0.00	1.00	-3.15	1.79	34	0.22
Factor2	0.00	1.00	-2.11	1.34	34	0.20
Factor3	0.00	1.00	-2.15	2.52	34	-0.10
Factor4	0.00	1.00	-1.30	4.51	34	-0.19

rok 2018	Mean	Std Dev	Min	Max	N	Median
Factor1	0.00	1.00	-3.52	1.84	34	0.13
Factor2	0.00	1.00	-1.87	1.55	34	0.04
Factor3	0.00	1.00	-2.23	2.09	34	-0.05
Factor4	0.00	1.00	-1.58	2.99	34	-0.14

Po vylúčení Luxemburska len v CA

rok 2007	Mean	Std Dev	Min	Max	N	Median
Factor1	0.00	1.02	-3.15	1.79	33	0.23
Factor2	0.02	1.01	-2.11	1.34	33	0.33
Factor3	0.06	0.96	-2.15	2.52	33	0.03
Factor4	-0.14	0.61	-1.30	1.79	33	-0.20

rok 2018	Mean	Std Dev	Min	Max	N	Median
Factor1	0.00	1.02	-3.52	1.84	33	0.13
Factor2	-0.01	1.01	-1.87	1.55	33	0.01
Factor3	0.07	0.93	-1.42	2.09	33	0.06
Factor4	-0.09	0.86	-1.58	2.71	33	-0.20

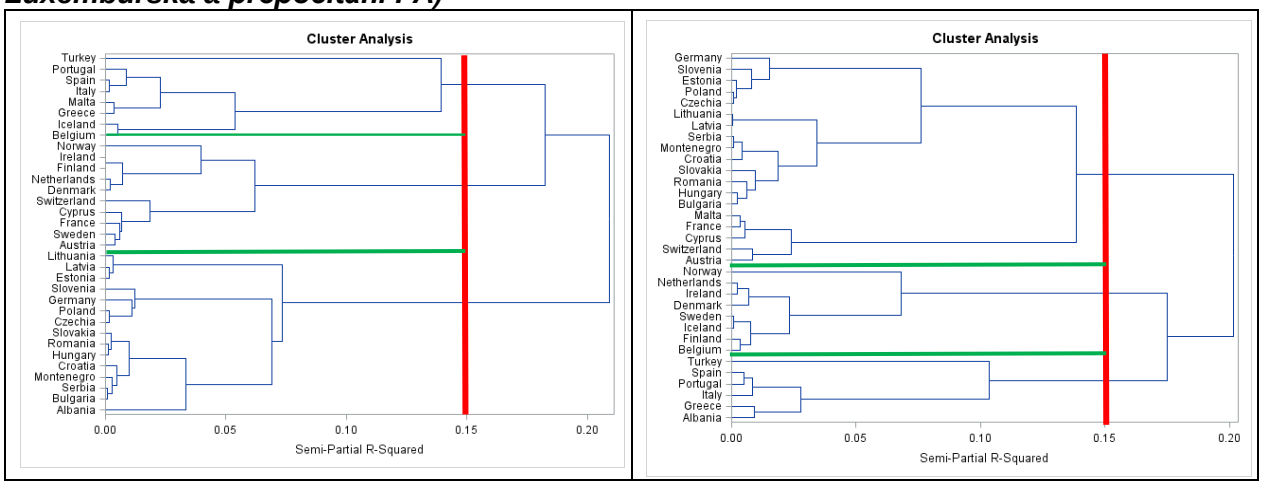
Po vylúčení Luxemburska vo FA, pre novú CA

rok 2007	Mean	Std Dev	Min	Max	N	Median
Factor1	0.00	1.00	-2.01	1.43	33	0.18
Factor2	0.00	1.00	-3.07	1.75	33	0.24
Factor3	0.00	1.00	-2.18	2.60	33	0.01
Factor4	0.00	1.00	-1.46	3.05	33	-0.08

rok 2018	Mean	Std Dev	Min	Max	N	Median
Factor1	0.00	1.00	-1.84	1.54	33	0.02
Factor2	0.00	1.00	-3.47	1.81	33	0.12
Factor3	0.00	1.00	-1.64	2.18	33	-0.06
Factor4	0.00	1.00	-1.67	3.25	33	-0.10

Zdroj: vlastné spracovanie v SAS EG

Obrázok č. 10: CA – dendrogramy zhlukovania 33 európskych krajín na základe 4 spoločných faktorov, použitá Wardova metóda (roky 2007 a 2018, po vylúčení Luxemburska a prepočítaní FA)



Zdroj: vlastné spracovanie v SAS EG

Pri zvolenom počte zhlukov 3 dostávame iné zoskupenia na roky 2007 a 2018 (obrázok č. 10) v porovnaní s výsledkom na obrázku č. 9. Tieto zhluky sú už pomerne rovnorodé a logické z hľadiska úrovne ľudského rozvoja. Zhluky je vhodné interpretovať pomocou popisnej štatistiky pôvodných vstupných ukazovateľov (pozri tabuľka č.18), alebo graficky na základe krabicových grafov (príloha č. 4). Označenie zhlukov číslom 1 až 3 v sledovaných rokoch vytvoril softvér. Zhluky nemožno porovnávať podľa označenia číslom, ale podľa výsledných štatistík. V roku 2007 sú zaradené krajiny s najvyššou úrovňou HDI do zhluku 1 (10 krajín) a v roku 2018 do zhluku 2 (8 krajín). Rozdelenie ukazovateľov v niektorých zhlukoch je pomerne zošikmené a vyskytujú sa aj extrémne hodnoty, napr. pri ukazovateľoch očakávaná stredná dĺžka života (LE) a hrubý národný produkt na hlavu (GNI) (pozri príloha č. 4).

Tabuľka č. 18: Popisná štatistika výsledných zhlukov pre 33 európskych krajín (roky 2007 a 2018)

CL_2007	N	Variable	Mean	Std Dev	Min	Max	CL_2018	N	Variable	Mean	Std Dev	Min	Max
1	10	LE_2007	80.03	0.99	78.50	81.70	1	19	LE_2018	78.74	2.77	74.9	83.6
		Exp_Edu_2007	16.00	1.21	13.70	17.60			Exp_Edu_2018	15.71	0.93	14.3	17.4
		Mean_Edu_2007	11.96	0.64	10.70	12.60			Mean_Edu_2018	12.23	0.84	11	14.1
		GNI_2007	46066	8939	33342	64859			GNI_2018	31365	10861	15218	59375
2	15	LE_2007	74.73	2.28	71.60	79.60	2	8	LE_2018	82.01	0.67	80.8	82.9
		Exp_Edu_2007	15.00	1.34	12.00	16.90			Exp_Edu_2018	18.88	0.58	18	19.7
		Mean_Edu_2007	11.31	1.12	9.20	13.60			Mean_Edu_2018	12.38	0.27	11.8	12.6
		GNI_2007	21136	8005	8671	41080			GNI_2018	50461	8226	41779	68059
3	8	LE_2007	79.46	2.68	73.20	81.40	3	6	LE_2018	81.12	2.56	77.4	83.4
		Exp_Edu_2007	15.99	1.99	12.60	19.00			Exp_Edu_2018	16.55	0.94	15.2	17.9
		Mean_Edu_2007	9.23	1.54	6.30	11.10			Mean_Edu_2018	9.58	1.02	7.7	10.5
		GNI_2007	32351	8523	17734	42179			GNI_2018	26872	8647	12300	36141

Zdroj: vlastné spracovanie v SAS EG

3. ZÁVER

V príspevku sme opísali princíp výpočtu indexu ľudského rozvoja (HDI) podľa metodiky UNDP. Ľudský rozvoj je zložený jav, ktorý sa priamo nedá zmerať jedným ukazovateľom. Experti vybrali 4 merateľné ukazovatele a vytvorili z nich 3 dimenzie (indexy). Výsledný HDI sa potom vypočíta ako geometrický priemer z týchto 3 indexov.

V článku sme sa pokúsili vypočítať HDI pomocou viacrozmerných štatistických metód, konkrétne pomocou analýzy hlavných komponentov a faktorovej analýzy z pôvodných 4 merateľných ukazovateľov ako aj z jeho 3 dimenzií (indexov). Použili sme údaje z 34 európskych krajín z dvoch rokov, 2007 a 2018. Aj keď použité premenné boli korelované, čo vyžadujú tieto viacrozmerné metódy, údaje obsahovali extrémne hodnoty a to naše riešenie a výsledky negatívne ovplyvnilo. Záverom musíme konštatovať, že výsledky HDI podľa metodiky UNDP sú v princípe lepšie ako sme dostali pomocou viacrozmerných metód, lebo ukazovatele sú škálované (normované) na základe logicky vybraných hodnôt a nielen na základe matematických výpočtov. Podobné výsledky sme dostali len použitím faktorovej analýzy z troch čiastkových indexov (pozri prílohu č. 3). Medzi poradím (Rank) na základe HDI a spoločného faktora (Factor1) sú len minimálne rozdiely v oboch sledovaných rokoch. Toto riešenie pokladáme za vhodnú alternatívu metodiky UNDP na hodnotenie ľudského rozvoja, čiže na výpočet integrálneho ukazovateľa HDI.

Rozvojová komisia OSN (UNDP) každý rok zostavuje poradie 189 krajín sveta v závislosti od veľkosti vypočítaného HDI, čiže sleduje stav a vývoj v oblasti ľudského rozvoja vo svete. UNDP zaraďuje krajiny sveta do 4 skupín podľa úrovne HDI. My sme sa pokúsili pomocou zhlukovej analýzy zoskupiť európske krajiny do zhlukov s podobnou úrovňou ukazovateľov ľudského rozvoja. Aj táto naša snaha bola negatívne ovplyvnená extrémnymi hodnotami použitých ukazovateľov. Výsledné zhluky nie sú rovnorodé a zoskupenia krajín sú nelogické.

Záverom konštatujeme, že metodika UNDP prináša pomerne uspokojivé výsledky pri hodnotení ľudského rozvoja. Výpočet integrálneho ukazovateľa HDI pomocou viacrozmerných štatistických metód narazil na viaceré problémy spojené so štruktúrou údajov a s výskytom extrémnych hodnôt v čiastkových ukazovateľoch.

Cieľom článku bolo metodicky opísať a ukázať, ako pristupovať k hodnoteniu zložitých javov v praxi. Na konštrukciu integrálnych ukazovateľov je v teórii odvodených niekoľko viacrozmerných metód a postupov, ktoré pri dodržaní podmienok ich použitia prinášajú dobré výsledky. V prípade ich nedodržania, výsledky môžu byť skreslené, nelogické až nesprávne.

POĎAKOVANIE

Túto prácu podporili:

- Agentúra na podporu výskumu a vývoja na základe zmluvy č. APVV-16-0091,
- Vedecká grantová agentúra MŠVVaŠ SR a SAV (VEGA) na základe zmluvy č. VEGA 2/0002/19.

LITERATÚRA

[1] ALTAS, D. – ARIKAN, G. : The Analysis of Human Development Index with Cluster Analysis Techniques. In: Social Sciences Research Journal, 2007, č. 3, s. 126 –138.

[2] AMALUDDIN, A. et al.: A Modified Human Development Index and Poverty in the Villages of West Seram Regency, Maluku Province, Indonesia. In: International Journal of Economics and Financial Issues, 2018, č. 2, s. 325 – 330.

[3] BERENGER, V.: Multidimensional measures of well-being: Standard of living and quality of life across countries. In: World Development, 2007, č. 7. s. 1259 – 1276.

[4] DESAI, M.: Human development: Concept and measurement. In: European Economic Review, 1991, č. 2 – 3, s. 350 – 357.

[5] DIENER, E. – SUH, E.: Measuring quality of life: economic, social, and subjective indicators. In: Social Indicators Research, 1997, 40, s. 189 – 216.

[6] KLUGMAN, J. – RODRIGUEZ, F. – CHOI, H. J.: The HDI 2010: new controversies, old critiques. In: The Journal of Economic Inequality, 2011, č. 2, s. 249 – 288.

[7] KRÁĽ, P. et al.: Viacrozmerné štatistické metódy so zameraním na riešenie problémov ekonomickej praxe. Banská Bystrica: Univerzita Mateja Bela, Ekonomická fakulta, 2009. 175 s. ISBN 978-80-8083-840-9.

[8] MAJEROVÁ, I.: Comparison of Old and New Methodology in Human Development and Poverty Indexes: A case of the Least Developed Countries. In: Journal of Economics Studies and Research, 2012, s. 1 – 15.

[9] MAJEROVÁ, I. – NEVIMA, J.: The measurement of Human Development using the Ward Method of Cluster Analysis. In: Journal of International Studies, 2017, č. 2, 239 – 257.

[10] MURTAGH, F. – LEGENDRE, P.: Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. In: Journal of Classification, 2011, 31, s. 274 – 295.

[11] OECD: Measuring well-being and progress. Well-being Research. Paris: OECD Statistics and Data, 2019. [online]. [cit. 25.3.2020] Dostupné na: https://www.oecd.org/statistics/measuring-well-being-and-progress.htm

[12] OECD: How's Life? 2020: Measuring Well-being. Paris: OECD Publishing. 2020. [online]. [cit. 25.3.2020] Dostupné na: https://doi.org/10.1787/9870c393-en

[13] SOMARRIBA, N. – PENA, B.: Synthetic indicators of Quality of Life in Europe. In: Social Indicators Research, 2009, č. 1, s.115 – 133.

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

[15] Sustainable Development Goals. [online]. [cit. 25. 3. 2020] Dostupné na: https://sustainabledevelopment.un.org/sdgs

[16] UNDP: Human Development Data (1990 – 2018) [online]. [cit. 25.3.2020] Dostupné na: http://hdr.undp.org/en/data

[17] UNDP Human Development Indices and Indicators: 2018 Statistical Update . 2018. [online]. [cit. 25.3.2020] Dostupné na:

http://hdr.undp.org/sites/default/files/2018_human_development_statistical_update.pdf

[18] UNDP: Human Development Report 2019. Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century. 2019. USA: PBM Graphics, 2014. ISBN 978-92-1-126368-8. [online]. [cit. 25.3.2020] Dostupné na: http://hdr.undp.org/sites/default/files/hdr2019.pdf

[19] UNDP: Technical notes. 2019, [online]. [cit. 25.3.2020] Dostupné na: http://hdr.undp.org/sites/default/files/hdr2019_technical_notes.pdf

[20] YANG, Y. – HU, A.: Investigating Regional Disparities of China´s Human Development with Cluster Analysis: A historical Perspective. In: Social Indicators Research, 2017, č. 3, s. 417 – 432.

RESUMÉ

Článok sa zaoberá metódami výpočtu indexu ľudského rozvoja (HDI) ako integrálneho ukazovateľa zloženého z viacerých čiastkových indikátorov. V prvej časti predstavujeme oficiálnu metodiku výpočtu HDI stanovenú a používanú v Rozvojovom programe Organizácie Spojených národov (UNDP). Takto vypočítaný HDI použijeme na zostavenie poradia vybraných európskych krajín (spolu 34 štátov) za roky 2007 a 2018. V druhej časti článku použijeme na zostavenie poradia týchto krajín v oblasti ľudského rozvoja viacrozmerné štatistické metódy, ako PCA, faktorovú analýzu a zhlukovú analýzu. Výsledky navzájom porovnáme a posúdime výhody a nevýhody oboch prístupov k výpočtu HDI.

RESUME

The paper deals with the methods of calculating the Human Development Index (HDI) as an integral indicator composed of several sub-indicators. In the first part we present the official HDI calculation methodology established and used by the United Nations Development Programme (UNDP). The HDI calculated in this way will be used to compile the ranking of selected European countries (34 countries in total) for years 2007 and 2018. In the second part of the article, we will use multivariate statistical methods such as PCA, factor analysis and cluster analysis to create countries' ranking in human development. We will compare the results and assess the advantages and disadvantages of both approaches for the HDI calculation.

PROFESIJNÝ ŽIVOTOPIS

Mgr. Alena Mojsejová, PhD., *ukončila magisterské štúdium na Fakulte prírodných vied UPJŠ v Košiciach. Titul PhD., v odbore ekonometria a operačný výskum, získala na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave. Pôsobí na Ekonomickej fakulte Technickej univerzity v Košiciach. V pedagogickej oblasti sa venuje výučbe pravdepodobnosti a štatistických metód. Vo vedeckej oblasti sa venuje aplikácii štatistických metód v rôznych oblastiach, najmä však v ekonomickej a sociálnej praxi.*

Doc. Ing. Iveta Stankovičová, PhD., *pôsobí na Fakulte managementu Univerzity Komenského v Bratislave. Dlhodobo sa venuje problematike využitia kvantitatívnych metód v ekonómii a manažmente, čo sa odráža v jej vedeckej profilácii a prednáškovej praxi pre študentov. Vyučuje predmety z oblasti štatistických metód a hĺbkovej analýzy údajov (data mining). Je autorkou vedeckej monografie, spoluautorkou niekoľkých učebníc, skrípt a mnohých vedeckých článkov publikovaných doma i v zahraničí. Od novembra 2014 je predsedníčkou Slovenskej štatistickej a demografickej spoločnosti.*

KONTAKT
alena.mojsejova@tuke.sk
iveta.stankovicova@fm.uniba.sk

Príloha č. 1: Hodnoty 4 merateľných ukazovateľov pre 34 vybraných európskych krajín (roky 2007 a 2018)

ID	Krajina	LE 2007	LE 2018	Exp_Edu 2007	Exp_Edu 2018	Mean_Edu 2007	Mean_Edu 2018	GNI 2007	GNI 2018
1	Albania	75.6	78.5	12.0	15.2	9.2	10.1	8671	12300
2	Austria	80.0	81.4	15.1	16.3	11.6	12.6	43896	46231
3	Belgium	79.4	81.5	19.0	19.7	11.1	11.8	42179	43821
4	Bulgaria	73.1	74.9	14.0	14.8	10.5	11.8	13447	19646
5	Croatia	76.0	78.3	14.3	15.0	10.1	11.4	21755	23061
6	Cyprus	78.9	80.8	13.7	14.7	11.2	12.1	33342	33100
7	Czechia	76.8	79.2	15.3	16.8	12.2	12.7	26958	31597
8	Denmark	78.5	80.8	16.8	19.1	12.5	12.6	46609	48836
9	Estonia	73.6	78.6	16.1	16.1	12.2	13.0	25555	30379
10	Finland	79.4	81.7	17.1	19.3	12.0	12.4	42604	41779
11	France	80.8	82.5	15.1	15.5	10.7	11.4	38425	40511
12	Germany	79.6	81.2	16.6	17.1	13.6	14.1	41080	46946
13	Greece	79.9	82.1	15.5	17.3	10.0	10.5	31243	24909
14	Hungary	73.7	76.7	15.3	15.1	11.3	11.9	21944	27144
15	Iceland	81.4	82.9	18.2	19.2	10.3	12.5	42014	47566
16	Ireland	79.5	82.1	17.0	18.8	11.9	12.5	42012	55660
17	Italy	81.4	83.4	16.2	16.2	9.4	10.2	38642	36141
18	Latvia	71.6	75.2	15.9	16.0	12.3	12.8	21284	26301
19	Lithuania	71.9	75.7	16.7	16.5	12.1	13.0	21881	29775
20	Luxembourg	79.4	82.1	13.5	14.2	11.7	12.2	**91519**	65543
21	Malta	80.2	82.4	14.6	15.9	9.9	11.3	26809	34795
22	Montenegro	74.1	76.8	14.1	15.0	10.8	11.4	13767	17511
23	Netherlands	80.0	82.1	16.6	18.0	12.0	12.2	46861	50013
24	Norway	80.5	82.3	17.6	18.1	12.6	12.6	64859	68059
25	Poland	75.5	78.5	15.0	16.4	11.8	12.3	18903	27626
26	Portugal	79.1	81.9	15.7	16.3	7.7	9.2	26694	27935
27	Romania	73.0	75.9	14.6	14.3	10.5	11.0	16494	23906
28	Serbia	73.3	75.8	13.5	14.8	10.2	11.2	12506	15218
29	Slovakia	74.7	77.4	14.7	14.5	11.1	12.6	23310	30672
30	Slovenia	78.4	81.2	16.9	17.4	11.8	12.3	29483	32143
31	Spain	81.1	83.4	16.1	17.9	9.1	9.8	33494	35041
32	Sweden	81.0	82.7	15.7	18.8	12.5	12.4	45438	47955
33	Switzerland	81.7	83.6	15.3	16.2	12.6	13.4	56611	59375
34	Turkey	73.2	77.4	12.6	16.4	6.3	7.7	17734	24905

Vysvetlivky: LE – očakávaná dĺžka života pri narodení v rokoch, Exp_Edu – očakávaný počet rokov štúdia, Mean_Edu – priemerný počet rokov štúdia, GNI – hrubý národný produkt na obyvateľa v PPP $ (v cenách roku 2011)

Zdroje údajov: http://hdr.undp.org/en/content/human-development-index-hdip a http://hdr.undp.org/en/indicators/141706#

Príloha č. 2: Hodnoty čiastkových indexov na výpočet HDI a poradie krajín podľa HDI (roky 2007 a 2018)

ID	Krajina	IE 2007	IH 2007	II 2007	IE 2018	IH 2018	II 2018	HDI 2007	HDI 2018	Rank_HDI 2007	Rank_HDI 2018	Rank Diff
1	**Albania**	**0.642**	**0.856**	**0.674**	**0.758**	**0.899**	**0.727**	**0.718**	**0.791**	**33**	**34**	**-1**
2	Austria	0.806	0.924	0.919	0.871	0.945	0.927	0.881	0.914	12	11	1
3	Belgium	0.871	0.914	0.913	0.893	0.946	0.919	0.899	0.919	8	10	-2
4	Bulgaria	0.738	0.817	0.740	0.805	0.845	0.798	0.764	0.816	31	29	2
5	Croatia	0.734	0.861	0.813	0.796	0.898	0.822	0.801	0.838	28	28	0
6	Cyprus	0.755	0.906	0.878	0.811	0.936	0.876	0.844	0.873	19	20	-1
7	**Czechia**	**0.830**	**0.874**	**0.845**	**0.892**	**0.911**	**0.869**	**0.849**	**0.891**	**17**	**16**	**1**
8	Denmark	0.884	0.900	0.928	0.920	0.935	0.935	0.904	0.930	6	8	-2
9	Estonia	0.856	0.825	0.837	0.881	0.901	0.863	0.839	0.882	20	19	1
10	Finland	0.874	0.914	0.915	0.915	0.950	0.912	0.901	0.926	7	9	-2
11	France	0.775	0.936	0.899	0.811	0.962	0.907	0.867	0.891	14	15	-1
12	Germany	0.915	0.917	0.909	0.946	0.941	0.929	0.914	0.939	3	4	-1
13	Greece	0.764	0.922	0.868	0.833	0.955	0.834	0.849	0.872	18	21	-3
14	Hungary	0.801	0.825	0.814	0.816	0.872	0.846	0.813	0.844	26	27	-1
15	Iceland	0.842	0.944	0.912	0.918	0.967	0.931	0.898	0.938	10	5	5
16	Ireland	0.870	0.915	0.912	0.918	0.955	0.955	0.899	0.943	9	3	6
17	Italy	0.765	0.945	0.900	0.793	0.975	0.890	0.867	0.883	15	18	-3
18	Latvia	0.850	0.794	0.810	0.871	0.849	0.842	0.818	0.854	24	25	-1
19	Lithuania	0.867	0.799	0.814	0.890	0.858	0.860	0.826	0.869	22	23	-1
20	**Luxembourg**	**0.766**	**0.914**	**1.000**	**0.802**	**0.955**	**0.980**	**0.888**	**0.909**	**11**	**12**	**-1**
21	Malta	0.734	0.927	0.845	0.818	0.960	0.884	0.832	0.885	21	17	4
22	Montenegro	0.751	0.833	0.744	0.797	0.873	0.780	0.775	0.816	30	30	0
23	Netherlands	0.862	0.923	0.929	0.906	0.956	0.939	0.904	0.933	5	7	-2
24	**Norway**	**0.907**	**0.930**	**0.978**	**0.919**	**0.958**	**0.985**	**0.938**	**0.954**	**1**	**1**	**0**
25	Poland	0.811	0.854	0.792	0.866	0.901	0.849	0.819	0.872	23	22	1
26	Portugal	0.691	0.909	0.844	0.759	0.952	0.851	0.809	0.850	27	26	1
27	Romania	0.756	0.815	0.771	0.762	0.860	0.827	0.780	0.815	29	31	-2
28	**Serbia**	**0.716**	**0.820**	**0.729**	**0.783**	**0.859**	**0.759**	**0.754**	**0.799**	**32**	**33**	**-1**
29	**Slovakia**	**0.779**	**0.842**	**0.823**	**0.824**	**0.883**	**0.865**	**0.814**	**0.857**	**25**	**24**	**1**
30	Slovenia	0.862	0.898	0.859	0.893	0.941	0.872	0.873	0.902	13	13	0
31	Spain	0.751	0.940	0.878	0.824	0.976	0.885	0.853	0.893	16	14	2
32	Sweden	0.854	0.938	0.924	0.914	0.964	0.932	0.905	0.936	4	6	-2
33	**Switzerland**	**0.844**	**0.948**	**0.958**	**0.896**	**0.979**	**0.965**	**0.915**	**0.946**	**2**	**2**	**0**
34	**Turkey**	**0.557**	**0.819**	**0.782**	**0.712**	**0.884**	**0.833**	**0.709**	**0.806**	**34**	**32**	**2**

Vysvetlivky: IE – index vzdelania, IH – index zdravia, II – príjmový index, HDI – index ľudského rozvoja, Rank – poradie, Rank diff = Rank_2007 – Rank_2018 – zmena v poradí (kladné hodnoty znamenajú zlepšenie postavenia krajiny v rebríčku podľa HDI, resp. záporné hodnoty znamenajú zhoršenie postavenia krajiny).

Zdroj údajov: http://hdr.undp.org/en/data a vlastné spracovanie v SAS EG

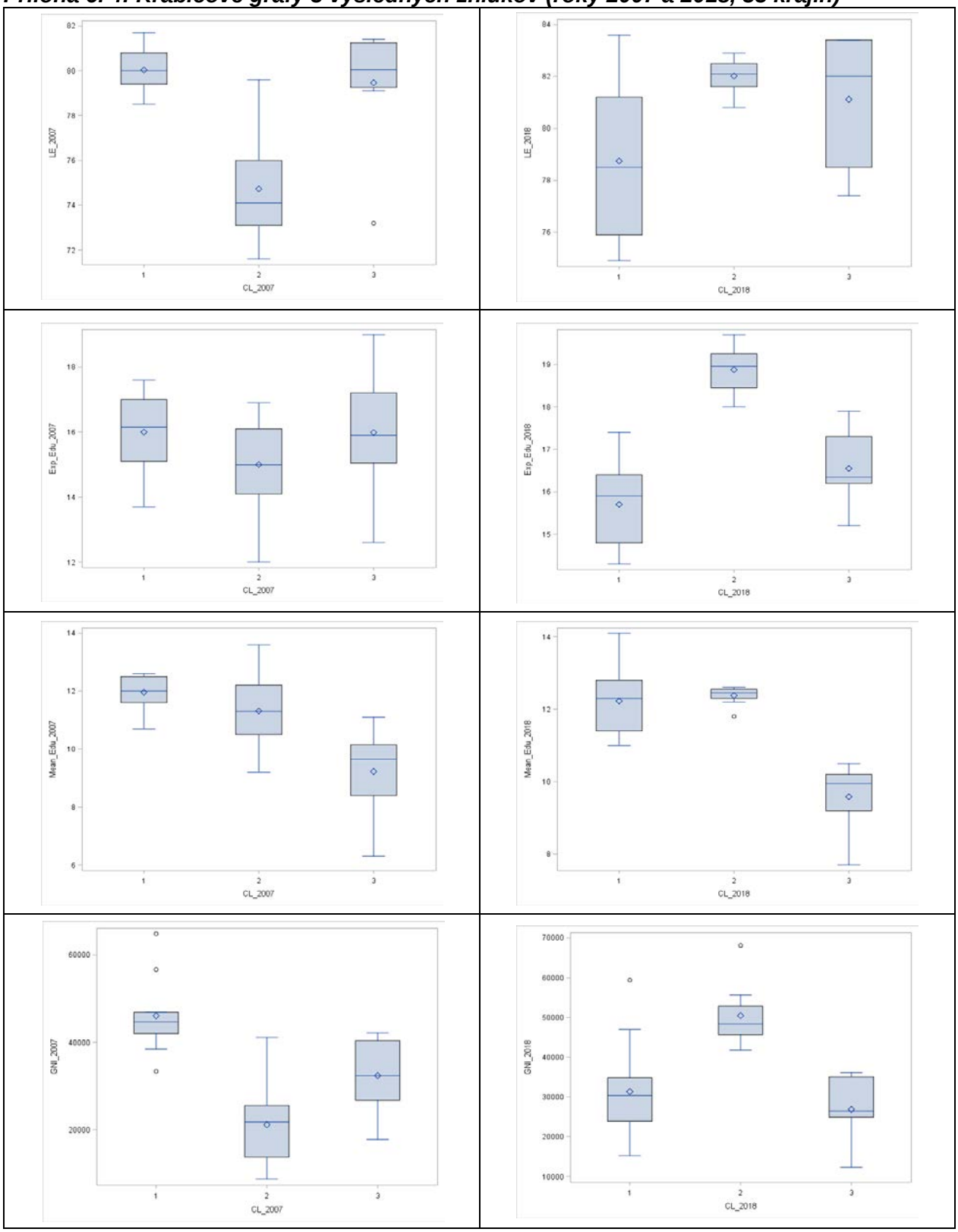
Príloha č. 3: Hodnoty a poradia pre ukazovatele HDI, 1. hlavný komponent (PRIN1) a 1. spoločný faktor (Factor1) pre 34 vybraných európskych krajín (roky 2007 a 2018)

ID	Country	HDI 2007	Rank	HDI 2018	Rank	PRIN1 2007	Rank	PRIN1 2018	Rank	Factor1 2007	Rank	Factor1 2018	Rank
1	**Albania**	**0.718**	**33**	**0.791**	**34**	**-2.746**	**33**	**-2.146**	**31**	**0.856**	**33**	**0.900**	**34**
2	Austria	0.881	12	0.914	11	0.790	12	0.816	12	1.042	12	1.034	11
3	Belgium	0.899	8	0.919	10	1.745	4	1.618	7	1.059	9	1.039	10
4	Bulgaria	0.764	31	0.816	29	-1.991	31	-2.277	33	0.902	31	0.922	31
5	Croatia	0.801	28	0.838	28	-1.261	28	-1.484	28	0.947	28	0.949	28
6	Cyprus	0.844	19	0.873	20	-0.298	19	-0.487	21	0.999	19	0.990	20
7	**Czechia**	**0.849**	**17**	**0.891**	**16**	**-0.066**	**18**	**-0.068**	**18**	**1.001**	**18**	**1.007**	**16**
8	Denmark	0.904	6	0.930	8	1.430	9	1.691	6	1.064	6	1.051	8
9	Estonia	0.839	20	0.882	19	-0.369	20	-0.390	20	0.987	20	0.996	19
10	Finland	0.901	7	0.926	9	1.398	10	1.584	8	1.061	7	1.046	9
11	France	0.867	14	0.891	15	0.488	15	0.257	15	1.027	15	1.012	15
12	Germany	0.914	3	0.939	4	1.654	5	1.410	10	1.075	3	1.060	5
13	Greece	0.849	18	0.872	21	0.039	17	-0.134	19	1.005	17	0.989	21
14	Hungary	0.813	26	0.844	27	-0.976	26	-1.480	27	0.958	27	0.955	27
15	Iceland	0.898	10	0.938	5	1.586	6	2.063	3	1.060	8	1.061	4
16	Ireland	0.899	9	0.943	3	1.335	11	2.124	2	1.059	10	1.065	3
17	Italy	0.867	15	0.883	18	0.589	14	0.189	16	1.028	14	1.004	18
18	Latvia	0.818	24	0.854	25	-0.870	25	-1.320	26	0.961	25	0.964	26
19	Lithuania	0.826	22	0.869	23	-0.601	22	-0.867	24	0.971	22	0.982	23
20	**Luxembourg**	**0.888**	**11**	**0.909**	**12**	**1.821**	**3**	**1.027**	**11**	**1.055**	**11**	**1.033**	**12**
21	Malta	0.832	21	0.885	17	-0.378	21	0.095	17	0.987	21	1.005	17
22	Montenegro	0.775	30	0.816	30	-1.707	29	-2.021	30	0.914	30	0.924	29
23	Netherlands	0.904	5	0.933	7	1.478	7	1.560	9	1.066	5	1.055	7
24	**Norway**	**0.938**	**1**	**0.954**	**1**	**2.649**	**1**	**2.497**	**1**	**1.105**	**1**	**1.078**	**1**
25	Poland	0.819	23	0.872	22	-0.751	23	-0.597	22	0.964	23	0.986	22
26	Portugal	0.809	27	0.850	26	-0.805	24	-0.663	23	0.963	24	0.968	25
27	Romania	0.780	29	0.815	31	-1.713	30	-2.240	32	0.920	29	0.924	30
28	**Serbia**	**0.754**	**32**	**0.799**	**33**	**-2.233**	**32**	**-2.428**	**34**	**0.890**	**32**	**0.905**	**33**
29	**Slovakia**	**0.814**	**25**	**0.857**	**24**	**-1.016**	**27**	**-1.215**	**25**	**0.960**	**26**	**0.969**	**24**
30	Slovenia	0.873	13	0.902	13	0.677	13	0.449	14	1.028	13	1.020	13
31	Spain	0.853	16	0.893	14	0.253	16	0.584	13	1.012	16	1.013	14
32	Sweden	0.905	4	0.936	6	1.439	8	1.890	5	1.067	4	1.059	6
33	**Switzerland**	**0.915**	**2**	**0.946**	**2**	**1.828**	**2**	**1.972**	**4**	**1.081**	**2**	**1.071**	**2**
34	**Turkey**	**0.709**	**34**	**0.806**	**32**	**-3.420**	**34**	**-2.008**	**29**	**0.852**	**34**	**0.917**	**32**

Vysvetlivky: HDI – hodnoty indexu ľudského rozvoja (podľa metodiky UNDP), PRIN1 – komponentné skóre pre 1. hlavný komponent (vypočítaný z korelačnej matice 4 vstupných indikátorov), Factor1 – faktorové skóre pre 1. spoločný faktor (vypočítaný z kovariančnej matice 3 dimenzií – indexov), Rank – poradie krajiny v rebríčku podľa hodnoty ukazovateľa v sledovanom roku.

Zdroj údajov: http://hdr.undp.org/en/data a vlastné spracovanie v SAS EG

Príloha č. 4: Krabicové grafy 3 výsledných zhlukov (roky 2007 a 2018, 33 krajín)



Zdroj: vlastné spracovanie v SAS EG

Hana ŘEZANKOVÁ
Fakulta informatiky a statistiky Vysoké školy ekonomické v Praze

ZPŮSOBY VÝBĚRU VYSVĚTLUJÍCÍCH PROMĚNNÝCH V KLASIFIKAČNÍCH STROMECH

METHODS OF SELECTING EXPLANATORY VARIABLES IN CLASSIFICATION TREES

ABSTRAKT
Článek se zaměřuje na různá hodnocení vztahů mezi kategoriálními proměnnými a jejich aplikaci na problematiku výběru vysvětlujících proměnných v klasifikačních stromech. Jsou jednak diskutovány postupy dostupné v komerčních programových systémech (chí-kvadrát testy a porovnávání variability vysvětlované proměnné v různých skupinách objektů pomocí Giniho míry a entropie), jednak naznačeny další možnosti vývoje v této oblasti. Stávající postupy jsou ilustrovány na analýze dat v programovém systému IBM SPSS Decision Trees. Výzkum se v poslední době zaměřuje na hodnocení jednostranné závislosti ordinální vysvětlované proměnné na proměnné nominální a implementaci v rámci klasifikačních stromů. Takové přístupy již byly realizovány v balíčcích v prostředí R.

ABSTRACT
The paper focuses on different evaluations of relationships between categorical variables and their application to the explanatory variables selection in classification trees. On the one hand approaches available in commercial software systems (chi-square tests and comparison of variability explained in different groups of objects using the Gini measure and the entropy) are discussed and on the other hand, further development possibilities of development are outlined. The well-known possibilities are illustrated on the data analysis in the IBM SPSS Decision Trees system. Recently research focuses on the evaluation of directional association of the target variable on the nominal variable and the implementation in classification trees. Such approaches have been realized in the packages in the R environment.

KLÍČOVÁ SLOVA
klasifikační stromy, kategoriální proměnná, nominální proměnná, ordinální proměnná

KEY WORDS
classification trees, categorical variable, nominal variable, ordinal variable

1. ÚVOD
Názvem *klasifikační stromy* je označována skupina metod, které byly navrženy pro řešení klasifikačních úloh s vysvětlovanou proměnnou. V klasifikačních úlohách tohoto typu jsou na základě známých hodnot kategoriální vysvětlované proměnné s využitím vysvětlujících proměnných vytvářeny modely či pravidla tak, aby mohly být odhadovány hodnoty vysvětlované proměnné v případě, kdy nejsou známy. Cílem je získání návodu k zařazování (klasifikaci) objektů charakterizovaných vektory hodnot vysvětlujících proměnných do skupin (tříd) daných množinou kategorií vysvětlované proměnné.

Klasifikační stromy tedy slouží ke stejným účelům jako diskriminační analýza nebo logistická regrese. Jejich základní odlišností od dvou dalších zmíněných metod je to, že vysvětlující proměnné jsou uvažovány jako kategoriální. Buď do analýzy vstupují jako kategoriální, nebo jsou na kategoriální převedeny (v případě vstupních kvantitativních spojitých proměnných). Navíc v procesu analýzy může docházet k překódování vysvětlujících proměnných, aby výsledné vztahy byly co nejvýstižnější.

Základním principem klasifikačních stromů je postupný výběr vysvětlujících proměnných z množiny vstupních proměnných (v případě potřeby překódovaných do vhodného počtu kategorií). Je vytvářena stromová hierarchická struktura, při které je původní soubor objektů postupně rozdělován na podsoubory. Není ovšem vhodné označovat klasifikační stromy jako metodu hierarchické shlukové analýzy (jak se někdy mylně v literatuře uvádí), neboť shluková analýza klasifikuje objekty na zcela jiném principu, a to bez využití reálné vysvětlované proměnné, navíc ani není znám počet tříd, viz [9].

Přestože jsou metody pro tvorbu klasifikačních stromů v dnešní době již poměrně dobře známé, v literatuře se někdy vyskytují nepřesnosti. Cílem tohoto článku je diskutovat některé způsoby výběru vysvětlujících proměnných. Nebude zde detailně pojednáno o konstrukci stromové struktury, ani o možných podmínkách ukončení větvení stromu. Základy těchto postupů jsou uvedeny např. v článku [7], kde jsou též charakterizovány nejznámější metody, včetně jejich historie. Budou však naznačeny další možnosti vývoje v této oblasti, k nimž patří např. aplikace speciálních postupů pro ordinální vysvětlovanou proměnnou.

2. KRITÉRIA PRO VÝBĚR VYSVĚTLUJÍCÍCH PROMĚNNÝCH

Při výběru vysvětlujících proměnných je postupně pro všechny vstupní proměnné (s různými vhodnými počty kategorií) buď testována nezávislost mezi vysvětlovanou a vysvětlující proměnnou, nebo je posuzována vnitroskupinová, resp. meziskupinová variabilita vysvětlované proměnné při rozdělení objektů do skupin podle kategorií zkoumané vysvětlující proměnné. Pokud jsou vysvětlující proměnné ordinální, pak se pořadí kategorií zohledňuje pouze při překódování do menšího počtu kategorií. Při samotném výběru vysvětlujících proměnných jsou pak všechny vstupní proměnné považovány za nominální. Většina používaných postupů je vhodná pro nominální vysvětlovanou proměnnou.

2.1. VYUŽITÍ CHÍ-KVADRÁT TESTŮ

Chí-kvadrát testy jsou určeny pro zkoumání závislosti dvou nominálních proměnných; v klasifikačních stromech se zpravidla používají bez ohledu na typ kategoriálních proměnných. Při jejich aplikaci se na všech vytvářených úrovních, tzn. pro různé skupiny objektů, provádějí testy o nezávislosti vysvětlované proměnné postupně s jednotlivými vysvětlujícími proměnných, přičemž pro vícekategoriální proměnné (tj. proměnné s více než dvěma kategoriemi) jsou prováděna překódování do nových proměnných s různými počty kategorií (původní kategorie jsou různými způsoby sdružovány). Pro větvení stromu se vybírá vysvětlující proměnná (původní nebo překódovaná), pro kterou je p-hodnota menší nebo rovna stanovené hladině významnosti. Pokud je takových proměnných více, vybírá se proměnná s nejnižší p-hodnotou. Protože jde o opakované statistické testování,

p-hodnota je obvykle modifikována Bonferroniho metodou. Pokud je p-hodnota větší než stanovená hladina významnosti, strom se dále nevětví.
Chí-kvadrát test může být proveden buď pomocí Pearsonovy chí-kvadrát statistiky, nebo s využitím věrohodnostního poměru. Označme vysvětlovanou proměnnou jako Y a její kategorie y_j, kde $j = 1, 2, \ldots, s$, a zkoumanou vysvětlující proměnnou jako X s kategoriemi x_i, kde $i = 1, 2, \ldots, r$. Dále označme počet objektů v rozdělované skupině symbolem n, sdružené absolutní četnosti v dvourozměrné kontingenční tabulce pro danou skupinu jako n_{ij}, řádkové marginální četnosti jako n_{i+} a sloupcové marginální četnosti jako n_{+j}. Pearsonova chí-kvadrát statistika se počítá jako:

$$\chi_P^2 = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}. \qquad (1)$$

Při platnosti nulové hypotézy o nezávislosti má tato statistika přibližně chí-kvadrát rozdělení s počty stupňů volnosti $(r-1)(s-1)$. Vzorec pro věrohodnostní poměr je:

$$\chi_{LR}^2 = -2\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}\ln\left(\frac{\frac{n_{i+}n_{+j}}{n}}{n_{ij}}\right). \qquad (2)$$

Při platnosti nulové hypotézy o nezávislosti má tato statistika rovněž přibližně stejné rozdělení jako Pearsonova statistika.

2.2. Využití principu analýzy rozptylu pro nominální vysvětlovanou proměnnou

Jinou z používaných technik je nalezení vždy takové vysvětlující proměnné, pomocí jejichž kategorií lze vytvořit podmnožiny objektů tak, aby vnitroskupinová variabilita vysvětlované proměnné byla co nejmenší. Jde tedy o aplikaci poznatků z analýzy rozptylu. Protože je ale vysvětlovaná proměnná kategoriální, místo součtů čtvercových odchylek (jako základu pro výpočet rozptylu) se používají speciální míry pro nominální proměnnou, a to buď nominální rozptyl (Giniho míra mutability, viz [5]), nebo entropie. Ve statistických programových systémech IBM SPSS Decision Trees a Statistica je nabízena pouze Giniho míra. Využijme symboliku z části 2.1 s tím rozdílem, že n bude vždy vyjadřovat celkový počet objektů. Za předpokladu, že vysvětlovaná proměnná je v kontingenční tabulce sloupcová, pak celkovou variabilitu proměnné Y lze pomocí Giniho míry vyjádřit jako:

$$G(Y) = \sum_{j=1}^{s} \frac{n_{+j}}{n}\left(1 - \frac{n_{+j}}{n}\right), \qquad (3)$$

přičemž podíly $\frac{n_{+j}}{n}$ jsou marginální relativní četnosti, které lze označit též jako p_{+j}. Tato míra zohledňuje podíl počtu párů objektů s různými hodnotami. Nabývá hodnot z intervalu

od 0 do $(s-1)/s$; 0 odpovídá konstantě, nejvyšší hodnota pak stejným četnostem pro všechny kategorie, viz [12] a [13].

Při rozdělení původní množiny objektů na základě kategorií vybrané vysvětlující proměnné X pak lze vyjádřit variabilitu vysvětlované proměnné pro každou podmnožinu objektů (tj. pro každý řádek v kontingenční tabulce). Vnitroskupinová variabilita je váženým průměrem z hodnot Giniho míry získaných pro všechny podmnožiny objektů, tj.:

$$G(Y|X) = \sum_{i=1}^{r} \frac{n_{i+}}{n} \sum_{j=1}^{s} \frac{n_{ij}}{n_{i+}} \left(1 - \frac{n_{ij}}{n_{i+}}\right). \tag{4}$$

Při prvním větvení stromu se porovnává celková variabilita s vnitroskupinovou variabilitou rozdílem, to znamená, že je spočtena meziskupinová variabilita jako

$$G(Y) - G(Y|X). \tag{5}$$

Pro větvení se vybere taková vysvětlující proměnná, pro kterou byla zjištěna největší meziskupinová variabilita (tudíž nejmenší vnitroskupinová variabilita).

Při dalším větvení je postupováno analogicky. Porovnává se vážená variabilita proměnné Y v určité skupině objektů a vnitroskupinová variabilita při rozdělení dané skupiny objektů do dílčích skupin. Váhy se počítají vždy jako podíl počtu objektů v dané skupině k celkovému počtu objektů n. Obecně lze vzorec pro variabilitu proměnné Y v množině objektů odpovídající u-tému uzlu zapsat jako:

$$G_u(Y) = \frac{n_u}{n} \sum_{j=1}^{s} \frac{n_{+ju}}{n_u} \left(1 - \frac{n_{+ju}}{n_u}\right), \tag{6}$$

kde symbol u označuje uzel, v kterém je prováděno větvení. Např. n_u označuje počet objektů v u-tém uzlu a platí, že $n_0 = n$.

Vnitroskupinovou variabilitu pro u-tý uzel vyjádříme opět jako vážený průměr z hodnot Giniho míry získaných pro všechny podmnožiny objektů, tj.:

$$G_u(Y|X) = \sum_{i=1}^{r} \frac{n_{i+u}}{n} \sum_{j=1}^{s} \frac{n_{iju}}{n_{i+u}} \left(1 - \frac{n_{iju}}{n_{i+u}}\right). \tag{7}$$

Meziskupinová variabilita pro u-tý uzel je dána vztahem:

$$G_u(Y) - G_u(Y|X). \tag{8}$$

Obdobně by mohla být vyjádřena variabilita vysvětlované proměnné Y pomocí entropie. V metodách statistické analýzy se entropie vyjadřuje pomocí přirozeného logaritmu, který je využit i v jiných postupech (viz např. výše uvedený věrohodnostní poměr). V metodách

„data mining“ se využívá dvojkový logaritmus. Tento postup je v klasifikačních stromech využit např. v systému SAS Enterprise Miner, viz příklad v [7].

Celkovou variabilitu proměnné Y lze pomocí entropie vyjádřit jako:

$$H(Y) = \sum_{j=1}^{s} \frac{n_{+j}}{n} \log_2 \left(\frac{n_{+j}}{n}\right). \tag{9}$$

Při rozdělení původní množiny objektů na základě kategorií vybrané vysvětlující proměnné X pak lze vyjádřit variabilitu vysvětlované proměnné Y pro každou podmnožinu objektů. Vnitroskupinová variabilita je váženým průměrem z hodnot získaných pro všechny podmnožiny objektů, tj.:

$$H(Y|X) = \sum_{i=1}^{r} \frac{n_{i+}}{n} \sum_{j=1}^{s} \frac{n_{ij}}{n_{i+}} \log_2 \left(\frac{n_{ij}}{n_{i+}}\right). \tag{10}$$

Při prvním větvení stromu se porovnání celková variabilita s vnitroskupinovou variabilitou rozdílem, to znamená, že je spočtena meziskupinová variabilita jako:

$$H(Y) - H(Y|X). \tag{11}$$

Výsledná hodnota je označována jako „informační zisk“ (viz [7]). Při dalším větvení je postupováno analogicky s postupem vysvětleným při aplikaci Giniho indexu. Celková variabilita určité skupiny je vždy vážena relativní četností objektů vzhledem k celkovému původnímu počtu objektů.

Variabilita nominální vysvětlované proměnné a její rozklad jsou v praxi aplikovány také ke konstrukci koeficientů jednostranné závislosti (na jiné nominální proměnné), pro které byly rovněž navrženy testy na nulovost těchto koeficientů, tj. nezávislost. Jednostranná závislost je posuzována na základě podílu meziskupinové variability na celkové variabilitě. Podle způsobu vyjádření variability nominální proměnné jsou v praxi využívány koeficienty Goodmanovo-Kruskalovo lambda, Goodmanovo-Kruskalovo tau (využívá Giniho míru) a koeficient neurčitosti či nejistoty neboli informační koeficient (využívá entropii vyjádřenou pomocí přirozeného logaritmu), viz např. [13]. Tyto postupy však nejsou v klasifikačních stromech implementovány.

Využití měr variability nominální proměnné v analýze dat je samozřejmě mnohem širší, jako příklad lze uvést konstrukce nových měr podobnosti, které mohou být aplikovány ve shlukové analýze kategoriálních dat, viz [14].

2.3. POSTUPY PRO ORDINÁLNÍ VYSVĚTLOVANOU PROMĚNNOU

Pro ordinální vysvětlovanou proměnnou je obvykle postupováno analogicky jako pro nominální proměnnou, ovšem s využitím vhodného testu (v systému IBM SPSS Decision Trees je nabízen pouze chí-kvadrát test s využitím věrohodnostního poměru) či vhodné míry variability pro ordinální proměnnou (v systému SAS Enterprise Miner jsou speciálně

upraveny výpočty pro Giniho míru a entropii). Existují ale i další přístupy vyjádření míry závislosti ordinální proměnné na nominální, viz např. [1].

Variabilita ordinální proměnné je obvykle vyjadřována pomocí míry známé pod označením *dorvar* (diskrétní ordinální variance), viz např. [12]. Za předpokladu využití v kontingenční tabulce ji můžeme zapsat jako:

$$dorvar(Y) = 2\sum_{j=1}^{s} F_{+j}\left(1 - F_{+j}\right), \tag{12}$$

kde F_{+j} je marginální kumulativní relativní četnost pro j-tou kategorii proměnné Y. Tato míra nabývá hodnot od 0 do $(s-1)/2$. Variabilita se zvyšuje se vzrůstajícími četnostmi v krajních kategoriích (v první a poslední kategorii), viz [13]. Vztah této míry k Giniho míře mutability je vysvětlen v [10], kde je rovněž navržena míra závislosti charakterizující závislost ordinální (vysvětlované) proměnné na proměnné nominální (vysvětlující). Podrobněji se měřením variability ordinální proměnné zabývá článek [3], v němž je kromě řady jiných měr zmíněna normalizovaná varianta míry *dorvar* (vyjádřená na intervalu od 0 do 1). Vyjádřením variability ordinální proměnné a jeho využitím při ohodnocení závislosti se zabývá též článek [8].

Možné využití vyjádření variability ordinální proměnné pomocí kumulativních relativních četností k výběru vysvětlujících proměnných při konstrukci klasifikačních stromů je navrženo v článku [11]. Na návrhy publikované v tomto článku navazuje Archer, který v prostředí R vytvořil balíček rpartOrdinal, viz [2]. Další autoři navrhli modifikaci Archerova balíčku a vytvořili balíček rpartScore, viz [4]. Problematika ordinální vysvětlované proměnné při konstrukci náhodných lesů (rozšíření problematiky klasifikačních stromů pro případy datových souborů o velké dimenzionalitě) je zohledněna např. v [6].

Giniho míra mutability a míra *dorvar* mohou být vyjádřeny zobecněným vzorcem, jehož speciálním případem je rovněž míra pro diskrétní kvantitativní vysvětlovanou proměnnou. Je to Giniho průměrná diference (mean difference) daná vztahem:

$$D(Y) = 2\sum_{j=1}^{s-1}\left(y_{j+1} - y_j\right)F_{+j}\left(1 - F_{+j}\right), \tag{13}$$

odvození viz [10]. Pro kvantitativní proměnnou je však možné použít také rozptyl.

3. ILUSTRACE APLIKOVÁNÍ VYBRANÝCH KRITÉRIÍ

Pro účely ilustrace výše uvedených postupů je vybrán jednoduchý příklad s 11 objekty a třemi vysvětlujícími proměnnými. Objekty (tj. statistické jednotky pro klasifikační stromy) jsou vybrané metody shlukové analýzy ze tří skupin, kterými jsou *metody pevného rozkladu*, *metody fuzzy rozkladu* a *metody hierarchického shlukování* (vysvětlovaná proměnná *skupina*), viz tabulka č. 1. První vysvětlující proměnnou je *shlukování*, která indikuje, zda jde o shlukování *pevné* (každý objekt je přiřazen právě do jednoho shluku),

nebo *fuzzy* (každé kombinaci objekt a shluk je přiřazen stupeň příslušnosti na škále od 0 do 1). Podle této proměnné je možné jednoznačně identifikovat metody fuzzy rozkladu. Druhá proměnná *centroid* nabývá tří kategorií podle toho, zda jsou v průběhu analýzy pro jednotlivé shluky vytvářeny centroidy (vektory charakteristik vstupních proměnných) jako *vektory průměrných hodnot* pro daný shluk, nebo jako *vektory mediánů*, nebo jestli *se centroidy nevytvářejí*. Třetí vysvětlující proměnná *vzdálenosti* charakterizuje, jaké typy vzdáleností jsou v průběhu analýzy počítány. Možnosti jsou *vzdálenosti objektů od centroidu*, *vzdálenosti objektů od medoidu* (konkrétní objekt ze souboru, který reprezentuje danou skupinu), *vzdálenosti mezi objekty z různých shluků* a *vzdálenosti mezi centroidy* (pro jednotlivé páry shluků).

K aplikaci klasifikačních stromů byl využit programový systém IBM SPSS Decision Trees (verze 26). Chí-kvadrát testy byly aplikovány v algoritmu CHAID, princip analýzy rozptylu pomocí Giniho míry v algoritmu CRT. U obou algoritmů byl nastaven malý minimální počet objektů v koncových uzlech (vzhledem k celkově malému počtu objektů v souboru); lze nastavit např. hodnoty 2 nebo 3. Jde o ilustraci používaných postupů na malém datovém souboru; v případě aplikace chí-kvadrát testů nejsou splněny podmínky pro jejich použití. Správně by měl být použit exaktní Fisherův test, příklad je proto pro srovnání doplněn p-hodnotami pro tento test.

Tabulka č. 1: Vstupní datová matice pro ilustraci výběru vysvětlujících proměnných

Metoda	*Shlukování*	*Centroid*	*Vzdálenosti*	*Skupina*
k-průměrů (HCM)	pevné	vektor průměrů	vzdálenosti objektů od centroidu	pevného rozkladu
k-mediánů	pevné	vektor mediánů	vzdálenosti objektů od centroidu	pevného rozkladu
k-medoidů (PAM)	pevné	nestanovuje se	vzdálenosti objektů od medoidu	pevného rozkladu
CLARA	pevné	nestanovuje se	vzdálenosti objektů od medoidu	pevného rozkladu
fuzzy *k*-průměrů (FCM)	fuzzy	vektor průměrů	vzdálenosti objektů od centroidu	fuzzy rozkladu
PCM	fuzzy	vektor průměrů	vzdálenosti objektů od centroidu	fuzzy rozkladu
fuzzy *k*-medoidů	fuzzy	nestanovuje se	vzdálenosti objektů od medoidu	fuzzy rozkladu
průměrného spojení	pevné	nestanovuje se	vzdálenosti mezi objekty z různých shluků	hierarchické metody
jednoduchého spojení	pevné	nestanovuje se	vzdálenosti mezi objekty z různých shluků	hierarchické metody
úplného spojení	pevné	nestanovuje se	vzdálenosti mezi objekty z různých shluků	hierarchické metody
centroidní	pevné	vektor průměrů	vzdálenosti mezi centroidy	hierarchické metody

Zdroj: vlastní zpracování

Z tabulky č. 1 je zřejmé, že lze jednoznačně identifikovat buď skupinu metod fuzzy rozkladu (na základě kategorie *fuzzy* proměnné *shlukování*), nebo skupinu hierarchických metod, která neobsahuje vzdálenosti objektů od centroidu, ani od medoidu, charakteristických pro metody rozkladu (proměnná *vzdálenosti*). Pomocí použitých klasifikačních stromů byly v různém pořadí vybírány právě proměnné s*hlukování* a *vzdálenosti*, jejichž kombinace vede k jednoznačnému přiřazení metod shlukové analýzy do stanovených skupin. V další části této kapitoly bude pro zjednodušení pozornost věnována těmto dvěma vysvětlujícím proměnným.

3.1. Aplikace chí-kvadrát testů

Při aplikaci chí-kvadrát testů se v prvním kroku provádějí testy o nezávislosti vysvětlované proměnné s jednotlivými vysvětlujícími proměnných, přičemž pro vícekategoriální proměnné jsou prováděna překódování do nových proměnných s různými počty kategorií. V tabulce č. 2 jsou uvedeny hodnoty získané pro Pearsonův chí-kvadrát test a Fisherův exaktní test pro vysvětlující proměnnou *Shlukování*, původní proměnnou *Vzdálenosti* a proměnné odvozené z této proměnné překódováním do dvou kategorií (výsledky pro proměnné vzniklé překódováním do tří kategorií nejsou pro zjednodušení uvedeny).

Tabulka č. 2: Hodnoty získané pro Pearsonův chí-kvadrát test a Fisherův exaktní test

	Pearsonova statistika	P-hodnota pro chí-kvadrát test	P-hodnota pro Fisherův test
Závislost na shlukování	11,000	0,004	0,006
Závislost na vzdálenostech (4 kategorie)	11,306	0,079	0,050
Závislost na vzdálenostech (2 kategorie – 1. varianta)	0,557	0,757	1,000
Závislost na vzdálenostech (2 kategorie – 2. varianta)	1,253	0,535	0,766
Závislost na vzdálenostech (2 kategorie – 3. varianta)	1,925	0,382	1,000
Závislost na vzdálenostech (2 kategorie – 4. varianta)	2,597	0,273	0,418
Závislost na vzdálenostech (2 kategorie – 5. varianta)	3,798	0,150	0,309
Závislost na vzdálenostech (2 kategorie – 6. varianta)	7,219	0,027	0,055
Závislost na vzdálenostech (2 kategorie – 7. varianta)	11,000	0,004	0,006

Zdroj: vlastní zpracování

Nejmenší p-hodnota je jak v případě Pearsonova testu (0,004), tak Fisherova testu (0,006) menší než 0,05 a je shodná pro dvě proměnné, kterými jsou *shlukování* a *vzdálenosti* překódované do dvou kategorií (7. varianta). Kontingenční tabulky pro dvě shodně nejlépe ohodnocené závislosti jsou uvedeny jako tabulky č. 3 a 4.

Tabulka č. 3: Kontingenční tabulka pro vztah skupiny metod a typu shlukování

		Skupina metod			Celkem
		metody pevného rozkladu	metody fuzzy rozkladu	hierarchické metody	
Typ shlukování	pevné	4	0	4	8
	fuzzy	0	3	0	3
Celkem		4	3	4	11

Zdroj: vlastní zpracování

Tabulka č. 4: Kontingenční tabulka pro vztah skupiny metod a výpočtu vzdálenosti

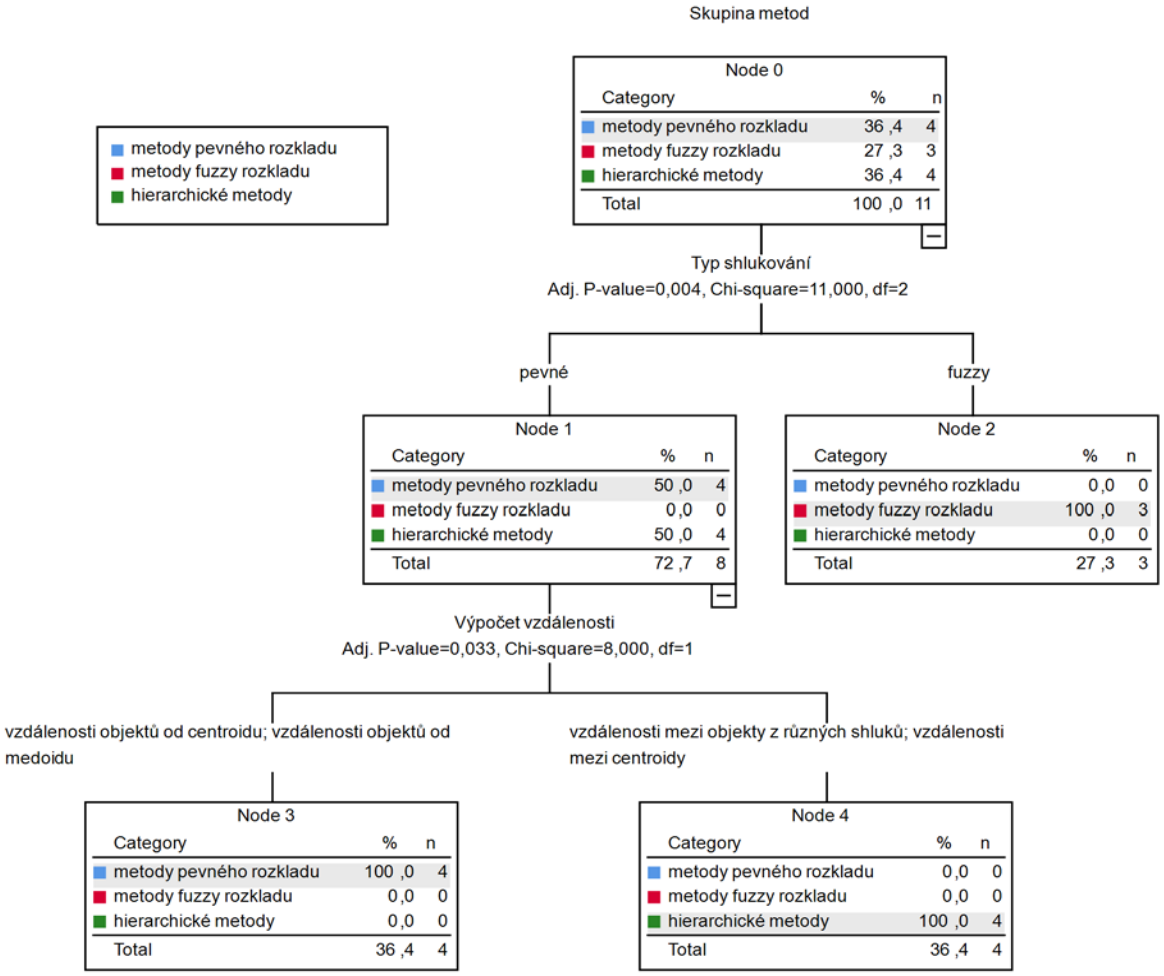
		Skupina metod			Celkem
		metody pevného rozkladu	metody fuzzy rozkladu	hierarchické metody	
Výpočet vzdálenosti	vzdálenosti objektů od centroidu nebo medoidu	4	3	0	7
	vzdálenosti mezi shluky	0	0	4	4
Celkem		4	3	4	11

Zdroj: vlastní zpracování

Algoritmem CHAID byla pro větvení použita první vysvětlující proměnná podle pořadí, tj. proměnná *shlukování*. Podle jejích dvou kategorií byly vytvořeny dvě podmnožiny objektů, viz obrázek č. 1. V každém uzlu grafu (anglicky „node“) je zobrazena tabulka četností pro hodnoty vysvětlované proměnné, které odpovídají dané skupině objektů. Do první skupiny byly zařazeny metody pevného shlukování, do druhé metody fuzzy shlukování. Druhá podmnožina tedy obsahuje pouze metody odpovídající kategorii „metody fuzzy rozkladu“ vysvětlované proměnné a další větvení se neprovádí. První podmnožina obsahuje metody ze dvou skupin, tudíž se zkoumá, zda by ji bylo možné dále rozdělit.

V úvahu přicházejí dvě zbývající vysvětlující proměnné a proměnné vytvořené jejich překódováním. Nejnižší p-hodnota byla získána při testu o nezávislosti vysvětlované proměnné s proměnnou *vzdálenosti* překódované do dvou kategorií, přičemž test byl proveden pro objekty z podmnožiny metod pevného shlukování. Odpovídající kontingenční tabulka pro dané dvě proměnné je označena jako tabulka č. 5. Hodnota Pearsonovy statistiky je 8, p-hodnota pak 0,005 a p-hodnota upravená Bonferroniho metodou 0,033. Jde o hodnotu menší než 0,05, proto se provádí další větvení stromu (při použití Fisherova testu by byla získána p-hodnota 0,014). Tím byly získány další dvě podmnožiny objektů, které jednoznačně odpovídají zbývajícím skupinám metod – metodám pevného rozkladu a hierarchickým metodám.

Obrázek č. 1: Klasifikační strom vytvořený metodou CHAID (Pearsonova statistika)



Zdroj: vlastní zpracování

Tabulka č. 5: Kontingenční tabulka pro vztah skupiny metod a výpočtu vzdálenosti pro metody pevného shlukování

		Skupina metod		Celkem
		metody pevného rozkladu	hierarchické metody	
Výpočet vzdálenosti	vzdálenosti objektů od centroidu nebo medoidu	4	0	4
	vzdálenosti mezi shluky	0	4	4
Celkem		4	4	8

Zdroj: vlastní zpracování

Obdobně se postupuje při aplikaci testu o nezávislosti s využitím věrohodnostního poměru, pouze jsou v klasifikačním stromu uváděny hodnoty této statistiky a odpovídající p-hodnoty, resp. p-hodnoty upravené Bonferroniho metodou.

V různých programových systémech se i při použití stejných algoritmů se stejným nastavením mohou výsledky lišit. Odlišné mohou být např. způsoby výběru vysvětlující proměnné v případě, kdy jsou získány dvě (příp. více) minimální p-hodnoty, jak je tomu v tabulce č. 2. V programovém systému IBM SPSS Decision Trees byla vybrána proměnná *shlukování*. Pokud bychom stejnou analýzu provedli v systému Statistica, byla by vybrána proměnná *vzdálenost*.

3.2 APLIKACE PRINCIPU ANALÝZY ROZPTYLU

Jak bylo zmíněno v části 2.2, při porovnání celkové a vnitroskupinové variability se v případě kategoriální vysvětlované proměnné využívá např. Giniho míra mutability. Na základě analýzy dat z tabulky č. 1 byl pomocí algoritmu CRT vytvořen klasifikační strom, který je uveden na obrázku č. 2. Ze znázorněného postupu je zřejmé, že pro klasifikaci byly využity dvě z původních tří proměnných. K rozlišení hierarchických metod a metod rozkladu byl využit způsob výpočtu vzdáleností. K rozlišení dvou skupin metod rozkladu byl použit typ shlukování.

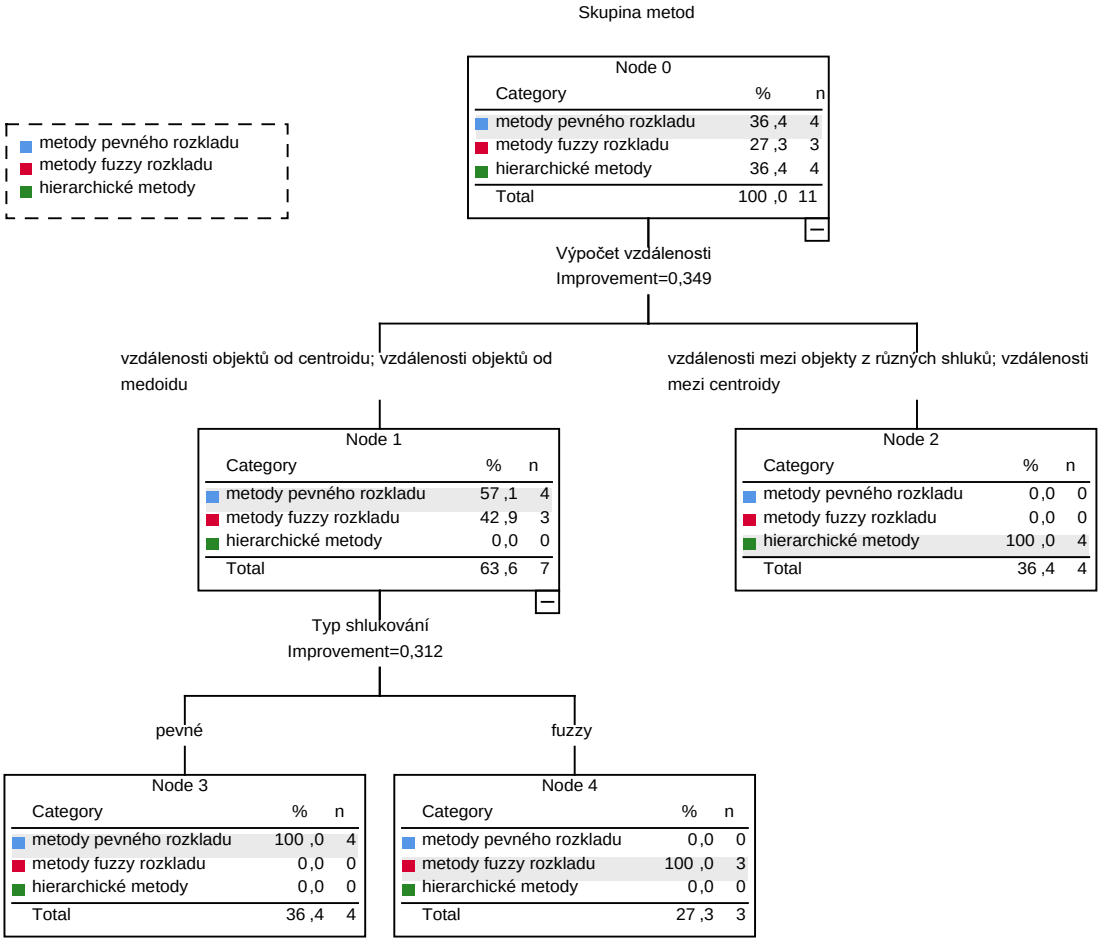
Hodnoty „improvement“ vyjadřují meziskupinovou variabilitu při rozdělení (větvení) určité skupiny objektů do menších skupin. Je vybráno takové rozdělení, kde je meziskupinová variabilita největší. V uzlu 0 jsou zahrnuty všechny objekty. Variabilita vysvětlované proměnné pomocí Giniho míry variability je 0,661 (viz tabulka č. 5). Tento uzel je rozdělen do uzlu 1 s 63,64 % objektů a s variabilitou 0,4898 a do uzlu 2 s 36,36 % objektů a s nulovou variabilitou. Průměrná variabilita ve skupinách je 0,6364·0,4898 + 0,3636·0 = 0,312. Rozdíl mezi celkovou variabilitou v původní množině objektů a vnitroskupinovou variabilitou při rozdělení do uzlů 1 a 2 (tj. meziskupinová variabilita) je tedy 0,661 – 0,312 = 0,349 (viz hodnota „improvement“ při rozdělení uzlu 0 do uzlů 1 a 2 na obrázku č. 2). Uzel 1 je rozdělen do dvou uzlů s nulovou variabilitou. Jako poslední hodnota je tedy uvedena variabilita v uzlu 1 (vážená), tj. 0,312 (viz hodnota „improvement“ při rozdělení uzlu 1 do uzlů 3 a 4 na obrázku č. 2). Pomocné výpočty variabilit pro uzly 0 a 1 jsou v tabulce č. 6 (podíl $\frac{n_{+ju}}{n_u}$ je označen jako p_{+ju} a platí, že $\frac{n_{+j1}}{n_1} = \frac{n_{1j0}}{n_{1+0}}$).

Tabulka č. 6: Pomocné výpočty pro klasifikační strom na obrázku č. 2

	Uzel 0		Uzel 1	
j	p_{+j0}	$p_{+j0}\ (1 - p_{+j0})$	p_{+j1}	$p_{+j1}\ (1 - p_{+j1})$
1	0,3636	0,2314	0,5714	0,2449
2	0,2727	0,1983	0,4286	0,2449
3	0,3636	0,2314	0,0000	0,0000
Součet	1,0000	0,6611	1,0000	0,4898

Zdroj: vlastní zpracování

Obrázek č. 2: Klasifikační strom vytvořený metodou CRT (Giniho míra)



Zdroj: vlastní zpracování

4. ZÁVĚR

Při výběru vysvětlujících proměnných v klasifikačních stromech by mělo být zohledněno, zda je vysvětlovaná proměnná nominální, nebo ordinální. V některých programových systémech je však zohlednění pouze částečné. Chí-kvadrát test se pro ordinální proměnnou buď nepoužívá vůbec (SAS Enterprise Miner), nebo se aplikuje pouze věrohodnostní poměr (IBM SPSS Decision Trees). V systému SAS Enterprise Miner jsou pro ordinální proměnnou výpočty Giniho míry a entropie speciálně upraveny. Problematika ordinální vysvětlované proměnné je postupně dále zkoumána a navržené postupy již byly implementovány v balíčcích v prostředí R.

V případě nominální vysvětlované proměnné nejsou dosud zohledněny všechny možnosti zkoumání závislostí. Používány jsou např. chí-kvadrát testy o nezávislosti, které hodnotí vzájemnou závislost proměnných. I když je jednostranná závislost její součástí, vhodnější by bylo aplikování speciálních měr pro jednostrannou závislost, resp. příslušné testy o nezávislosti. Na druhou stranu je třeba konstatovat, že tyto míry jsou založeny na zkoumání variability vysvětlované proměnné a jejího rozkladu, což je stejný princip jako při využití Giniho míry a entropie.

Dosud není speciálně řešena problematika diskrétní kvantitativní vysvětlované proměnné, pro kterou lze aplikovat neparametrické testy zaměřené na jednostrannou závislost, např. Kruskalův-Wallisův test. Při rozkladu variability je možné kromě rozptylu použít i jiné míry, např. Giniho průměrnou diferenci.

Poděkování

Tento článek byl připraven za podpory projektu IGA F4/44/2018 Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze.

LITERATURA

[1] AGRESTI, A.: Measures of nominal-ordinal association. In: Journal of the American Statistical Association, 1981, č. 375, s. 524 – 529.
[2] ARCHER, K. J.: rpartOrdinal: An R package for deriving a classification tree for predicting an ordinal response. In: Journal of Statistical Software, 2010, č. 7, s. 1 – 17. [online]. [cit. 30. 3. 2020]. Dostupné na: http://www.jstatsoft.org/v34/i07/.
[3] BLAIR, J. – LACY, M. G.: Statistics of ordinal variation. In: Sociological Methods & Research, 2000, č. 3, s. 251 – 280.
[4] GALIMBERTI, G. – SOFFRITTI, G. – MASO, M. D.: Classification trees for ordinal responses in R: The rpartScore package. In: Journal of Statistical Software, 2012, č. 10, s. 1 – 25. [online]. [cit. 30. 3. 2020]. Dostupné na: http://www.jstatsoft.org/v47/i10/.
[5] GINI, C. W.: Variability and Mutability. Contribution to the Study of Statistical Distributions and Relations. Studi Economico-Giuridici della R. Universita de Cagliari, 1912.
[6] JANITZA, S. – TUTZ, G. – BOULESTEIX, A.-L.: Random forest for ordinal responses: Prediction and variable selection. In: Computational Statistics & Data Analysis, April, 2016, s. 57 – 73.
[7] LABUDOVÁ, V.: Rozhodovacie stromy ako prediktívna modelovacia technika. In: Slovenská štatistika a demografia, 2017, č. 3, s. 60 – 76.
[8] LACY, M. G.: An explained variation measure for ordinal response models with comparisons to other ordinal R^2 measures. In: Sociological Methods & Research, 2006, č. 4, s. 469 – 520.
[9] LÖSTER, T.: Různé způsoby stanovení počtu shluků ve shlukové analýze. In: Slovenská štatistika a demografia, 2017, č. 3, s. 47 – 59.
[10] PICCARRETA, R.: A new measure of nominal-ordinal association. In: Journal of Applied Statistics, 2001, č. 1, s. 107 – 120.
[11] PICCARRETA, R.: Classification trees for ordinal variables. In: Computational Statistics, 2008, č. 3, s. 407 – 427.
[12] ŘEHÁK, J. – ŘEHÁKOVÁ, B.: Analýza kategorizovaných dat v sociologii. Praha: Academia, 1986. 397 s.
[13] ŘEZANKOVÁ, H.: Analýza dat z dotazníkových šetření. 4. vydání. Praha: Professional Publishing, 2017. 225 s. ISBN 978-80-906594-8-3.
[14] ŠULC, Z. – ŘEZANKOVÁ, H.: Comparison of similarity measures for categorical data in hierarchical clustering. In: Journal of Classification, 2019, č. 1, s. 58 – 72.

RESUMÉ

Článek pojednává o způsobech výběru vysvětlujících proměnných v klasifikačních stromech. Zaměřuje se jednak na dnes již poměrně dobře známé přístupy, kdy se předpokládá nominální vysvětlovaná proměnná. Vysvětlující proměnné jsou vybírány buď na základě chí-kvadrát testu (v programových systémech jsou obvykle nabízeny testy s využitím Pearsonovy statistiky a věrohodnostního poměru), nebo pomocí rozkladu variability, obdobně jako je zkoumán rozklad rozptylu v případě kvantitativní proměnné. Protože je však vysvětlovaná proměnná kategoriální, využívá se buď Giniho míra mutability, nebo entropie. Tyto známé přístupy jsou ilustrovány na analýze jednoduchého datového souboru v programovém systému IBM SPSS Decision Trees pomocí algoritmů CHAID a CRT. Kromě toho článek poukazuje také na současné trendy ve výzkumu v oblasti klasifikačních stromů a náhodných lesů, kterými jsou aplikace speciálních přístupů pro ordinální vysvětlovanou proměnnou.

RESUME

The paper deals with ways of explanatory variable selection in classification trees. It focuses on the well-known approaches when the nominal explanatory variable is expected. Explanatory variables are selected either on the basis of a chi-square test (in software systems the Pearson statistics and the likelihood ratio are usually available) or by means of variability decomposition, similarly as the variance decomposition is investigated in the case of the quantitative variable. However, for the reason that the target variable is categorical, either the Gini measure of mutability or the entropy are applied. These well-known approaches are illustrated on the analysis of the simple dataset using the CHAID and CRT algorithms in the IBM SPSS Decision Trees software system. Moreover, the actual research trends in the field of classification trees and the random forests are the applications of special techniques for the ordinal target variable.

PROFESNÍ ŽIVOTOPIS

Prof. Ing. Hana Řezanková, CSc., *absolvovala obor ekonomicko-matematické výpočty na Vysoké škole ekonomické v Praze, kde pôsobí v súčasnej dobe na katedře statistiky a pravděpodobnosti Fakulty informatiky a statistiky. Je členkou vedeckej rady a akademického senátu na tejto fakulte a predsedníčkou odborovej rady pre doktorský študijný program statistika. V rokoch 2013 – 2017 bola predsedníčkou České statistické společnosti a v rokoch 2015 – 2019 členkou České statistické rady. Vo svojej vedecko-výskumnej činnosti sa zameriava na analýzu kategoriálnych údajov a na metódy zhlukovej analýzy. Je autorkou či spoluautorkou niekoľkých knižných publikácií.*

KONTAKT

hana.rezankova@vse.cz

Viera LABUDOVÁ
Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave

POUŽITIE JEDNODUCHÝCH METÓD VIACROZMERNÉHO POROVNÁVANIA: ANALÝZA ZADLŽENOSTI DOMÁCNOSTÍ

THE USE OF SIMPLE METHODS OF MULTI-DIMENSIONAL COMPARISON: THE ANALYSIS OF HOUSEHOLD DEBT

ABSTRAKT
Článok sa zaoberá jednoduchými metódami viacrozmerného porovnávania (metóda poradí, bodovacia metóda a metóda vzdialenosti od fiktívneho objektu) a ich praktickou aplikáciou pri analýze zadlženosti domácností vybraných krajín EÚ. V prvej časti článku opisujeme princíp týchto metód. V druhej časti sa zaoberáme priestorovou analýzou dlhu domácností v krajinách, ktoré sa zúčastnili druhej vlny prieskumu o financovaní a spotrebe domácností HFCS – Household Finance and Consumption Survey. Pri porovnávaní krajín sme použili údaje z druhej vlny tohoto zisťovania.

ABSTRACT
This article deals with simple methods of the multi-dimensional comparison (the ranking method, the scoring method and the distance method from a fictitious object) and their practical application in the analysis of households' indebtedness of selected EU countries. In the first part of the article we describe the principle of these methods. In the second part we deal with the spatial analysis of household debt in the countries that participated in the second wave of the Household Finance and Consumption Survey (HFCS). We used data from the second wave of this survey to compare the countries.

KĽÚČOVÉ SLOVÁ
zadlženosť domácností, HFCS, metódy viacrozmerného porovnávania

KEY WORDS
household indebtedness, HFCS, methods of multi-dimensional comparison

1. ÚVOD
Pri porovnávacích analýzach sa veľmi často stretávame s úlohou usporiadania a klasifikácie objektov (krajín, regiónov rôzneho stupňa regionálneho členenia...) na základe sledovaného zloženého javu. Najčastejšie ide o porovnávanie objektov vzhľadom na dosiahnutú úroveň sociálneho, ekonomického, sociálno-ekonomického rozvoja, dosiahnutej životnej úrovne, kvality života atď.

Zloženým javom nazývame pri takýchto analýzach jav, na ktorého opis potrebujeme konečnú množinu premenných (ukazovateľov) $X_1, X_2,... X_k$. Ich počet a štruktúra závisí od mnohých skutočností, napr. od hĺbky výskumu, hierarchickej úrovne sledovaného objektu, dostupnosti a spracovateľnosti údajov [32].

Pri analýzach, ktoré sú založené na sledovaní zloženého javu, možno použiť rôzne metodologické prístupy.

Prvý prístup spočíva v budovaní systému ukazovateľov, podrobne charakterizujúcich sledovaný jav. Uvedený systém umožňuje analyzovať jednotlivé prvky zloženého javu a to ustavičným dopĺňaním ukazovateľov, t. j. nie redukciou, ale rozširovaním rozmerov obsiahnutých v danom jave[1] [23].

Druhý prístup je založený na konštrukcii syntetickej premennej (taxonomickej miery rozvoja[2]), ako funkcie premenných meritórne zviazaných na nižších úrovniach agregácie, pričom konštrukcia tejto miery je spojená s transformáciou viacrozmerného priestoru vybraných znakov do jednorozmerného priestoru agregovanej premennej, ktorej realizácia závisí od všetkých pôvodných premenných [33].

Cieľom jednoduchých metód viacrozmerného porovnávania[3] je nahradiť niekoľko vybraných ukazovateľov, pomocou ktorých chceme porovnávať vybrané objekty, jedným kvantitatívne vyjadreným integrálnym ukazovateľom (syntetickou premennou). Vzhľadom na to, že vybrané ukazovatele bývajú spravidla heterogénne (vyjadrené v rôznych meracích jednotkách, ich veľkosti sú rádovo rôzne), nemôžeme ich agregovať priamym sčitovaním. Nerovnorodé ukazovatele sa preto menia (transformujú) na rovnorodé ukazovatele, z ktorých sa vytvára tzv. syntetická premenná. Hodnoty tejto premennej možno využiť na lineárne usporiadanie objektov (určenie poradia, ktoré vyjadruje dosiahnutú úroveň sledovaného javu).

V českej literatúre sa týmto metódam[4] venujú predovšetkým práce Křováka [13, 14, 15, 16, 17]. Autor v nich opisuje metódu váženého súčtu poradí, bodovaciu metódu, metódu normovanej premennej a metódu vzdialenosti od fiktívneho objektu, pričom súčasne rieši problém ich aplikácie pre prípad chýbajúcich hodnôt. Príspevky poskytujú jednoduchý návod na použitie uvedených metód spolu s aplikačnou ukážkou na reálnych súboroch údajov opisujúcich priemyselné podniky.

Jednoduché metódy hodnotenia prezentoval aj Kejkula v článku, zaoberajúcom sa porovnávaním krajín na základe dosiahnutej životnej úrovne obyvateľstva [11]. Vybrané krajiny porovnával pomocou desiatich ukazovateľov, ktoré opisujú jednotlivé zložky životnej úrovne.

Každá z týchto metód hodnotenia sa v uvedených prácach opisuje samostatne bez zdôrazňovania ich spoločného základu. Ten spočíva v uplatňovaní niektorej z metód normovania hodnôt vstupných premenných a ich následnej agregácii. Celý proces vytvárania tzv. syntetických premenných možno zovšeobecniť, pričom všetky metódy možno potom chápať ako špecifické varianty tohto všeobecného návodu.

Metodike tvorby syntetickej premennej sa venuje nadštandardný priestor v poľskej literatúre. Priekopníkom pri využívaní syntetických mier na hodnotenie regiónov bol

---

[1] Zložený jav sa považuje za viacprvkový, posudzovaný z hľadiska viacerých aspektov [23].

[2] Uvedený pojem sa používa hlavne v poľskej literatúre , v ktorej sú tieto metódy spracované najkomplexnejšie.

[3] Niektorí autori používajú pojem viackriteriálne hodnotenie, napr. [30], alebo viacrozmerné hodnotenie.

[4] Tu sa použil názov jednoduché metódy hodnotenia.

Drewnowski, ktorý vstúpil do povedomia štatistickej verejnosti prácou On Measuring and Planning the Quality of Life. V Poľsku na jeho prácu nadviazal A. Luszniewicz , ktorý svoje výsledky zverejnil v publikácii Poziom zycia ludności Polski w latach 1980-1986, raport nr.1 [citované podľa 26].

Osobitnú pozornosť si zaslúži vroclavská škola, ktorá vytvorila skupinu metód viacrozmernej porovnávacej analýzy. Východiskom pre túto skupinu bola práca Zastosowania metody taksonomicznej do typologicznego podzialu krajów ze wzgledu na poziom ich rozwoju i strukture wykwalifikowanych kadr, ktorá sa stala všeobecne známou cez taxonomickú metódu Hellwiga. Ide o tzv. „vzorcovú metódu rozvoja“, ktorá sa stala koncepčným východiskom pre iných autorov, aplikujúcich túto metódu doslovne alebo v rôznych modifikáciách [24].

Za metodologicky najúplnejšie možno považovať práce Grabinského [8], Nowaka [27] a Zeliasia [21]. Okrem spôsobov normovania a agregovania hodnôt premenných sa v nich venuje nezanedbateľná pozornosť aj výberu premenných.

Ani v jednej z týchto publikácií sa priamo nestretávame s metódou súčtu poradí, ktorá je jedným z variantov tvorenia syntetickej premennej. Z tohto hľadiska ojedinelou je práca Nykowského, ktorá predstavuje metódy poradí od tých najjednoduchších po metódy zložitejšie, ktorých aplikácia je časovo pomerne náročná [28].

Tento príspevok plní dve úlohy. V prvej časti článku poskytujeme základné informácie o teoretickom princípe jednoduchých metód viacrozmerného porovnávania, konkrétne metódy poradí, bodovacej metódy, metódy normovanej premennej a metódy vzdialenosti od fiktívneho objektu. Druhá časť je analytická, v nej sú niektoré z týchto postupov aplikované na konkrétnych dátach, ktoré opisujú vybrané európske krajiny z hľadiska zadlženosti ich domácností.

2. KONŠTRUKCIA SYNTETICKEJ PREMENNEJ

Pri tvorbe syntetickej premennej sa musí riešiť niekoľko problémov súvisiacich s výberom premenných, určovaním typu premenných z hľadiska ich vplyvu na sledovaný jav, s výberom váh jednotlivých premenných, normovaním premenných a agregáciou hodnôt premenných do konečného tvaru syntetickej premennej.

2.1 VÝBER PREMENNÝCH

Premenné sa vyberajú pri súčasnom rešpektovaní zásad univerzálnosti (musia mať všeobecný význam), merateľnosti (musia byť priamo alebo nepriamo merateľné), dostupnosti hodnôt (ide o možnosť získania dostupných číselných informácií o každej premennej použitej v analýze), kvality údajov (toto kritérium súvisí s presnosťou práce s premennými), ekonomickosti (finančná náročnosť zberu údajov) a interpretovateľnosti (pri výbere treba uprednostniť premenné, ktoré sú jednoznačne interpretovateľné) [21].

Po výbere premenných spĺňajúcich uvedené kritériá sa počet takto vybraných premenných redukuje, pričom sa zohľadňuje ich priestorová variabilita a informačná hodnota.

2.2 REDUKCIA POČTU PREMENNÝCH

Premenné, ktoré opisujú zložený jav, sú často vzájomne závislé, čo v praxi znamená, že môžu byť nositeľmi podobných informácií o sledovanom jave. V súvislosti s tým je potrebné pôvodnú množinu premenných redukovať. Pri redukcii sa zohľadňuje objem informácií o sledovanom jave, ktoré premenné obsahujú a vzájomné väzby medzi nimi. Po redukcii vstupnej množiny premenných dostávame skupinu tzv. diagnostických premenných.

Pri výbere diagnostických premenných, sa uplatňujú dva prístupy. Pri prvom sa vyberajú reprezentanti skupín, do ktorých sú premenné rozdelené na základe ich podobnosti. Pri druhom prístupe sa vyberajú reprezentanti množiny premenných, bez toho, že by boli predtým premenné rozdelené do skupín.

Redukcia pôvodnej množiny premenných na základe ich informačnej hodnoty využíva rôzne miery, kvantifikujúce väzby medzi jednotlivými premennými. Medzi najčastejšie využívané miery patrí lineárny koeficient korelácie.

Predpokladajme, že sledujeme výskyt hodnôt premenných $X_1, X_2,... X_k$ na objektoch $O_1, O_2,... O_m$. Výsledkom je matica pozorovaní X:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mk} \end{pmatrix} \tag{1}$$

Každý riadkový vektor ($x_{i1}, x_{i2}, ..., x_{ik}$) matice **X**, pre $i = 1, 2, ..., m$, je jednou konkrétnou realizáciou k– tice ukazovateľov $X_1, X_2, ... X_k$, vyjadruje teda úroveň, stav sledovanej kategórie, javu, alebo procesu v i-tom objekte.

V prvej etape redukcie je východiskom matica párových koeficientov korelácie premenných vstupujúcich do analýzy:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & 1 \end{pmatrix} \tag{2}$$

Za podobné premenné X_i, X_j sa považujú tie, pre ktoré platí:

$$\left| r_{ij} \right| > r^* \tag{3}$$

pričom hodnota r^* je definovaná vzťahom [2]:

$$r^* = \min_i \max_j \left| r_{ij} \right| \tag{4}$$

kde r_{ij} *(i* = 1, 2,... *m; j* = 1, 2,... *k)* je hodnota párového koeficienta korelácie premenných X_i, X_j. Za *r** možno zvoliť aj ľubovoľné číslo z intervalu ⟨0, 1⟩ [4].

Postup pri redukcii jednotlivých premenných, ktorý je založený na zohľadňovaní ich „podobnosti“ býva najčastejšie výsledkom subjektívneho rozhodnutia. Existujú však metódy, ktoré vylučujú toto subjektívne hľadisko. Sem patrí tzv. parametrická metóda Hellwiga [21], [27]. Pri použití tejto metódy sa postupuje takto:

V matici *R* sa určí súčet prvkov každého stĺpca:

$$R_j = \sum_{i=1}^{k} \left| r_{ij} \right|, \ (i, j = 1, 2, \ldots k; i \neq j) \tag{5}$$

Vyznačí sa *s*-tý stĺpec, ktorý spĺňa podmienku:

$$R_s = \max_j \left\{ R_j \right\}, \ (j = 1, 2, \ldots k) \tag{6}$$

V *s*-tom stĺpci sa vyznačia prvky vyhovujúce nerovnosti

$$\left| r_{is} \right| \geq r^* \tag{7}$$

Premenná X_s určujúca *s*-tý stĺpec sa považuje za tzv. centrálnu premennú, premenné odpovedajúce riadkom, pri ktorých je splnený vzťah (7), sú tzv. satelitné premenné. Matica ***R*** sa redukuje o riadky a stĺpce, ktoré prislúchajú centrálnej premennej a satelitným premenným. Postup sa opakuje až do okamihu vyčerpania množiny premenných. Premenná, ktorá spĺňa vzťah (6), ale nevyhovuje podmienke (7), sa považuje za izolovanú premennú. Centrálne a izolované premenné tvoria množinu diagnostických premenných.

Od diagnostických premenných sa požaduje splnenie týchto podmienok: nízky stupeň „skorelovania“ premenných medzi sebou a vysoký stupeň „skorelovania“ s premennými, ktoré boli z analýz vylúčené. Vylúčené premenné sa „nepodieľajú“ na vytváraní syntetickej premennej priamo, ale prostredníctvom diagnostických premenných, s ktorými sú „skorelované“.

V druhej etape redukcie počtu premenných sa uplatňuje štatistické hľadisko založené na priestorovej variabilite premenných. Z analýzy sú vylúčené premenné, ktorých variabilita vyjadrená variačným koeficientom je veľmi nízka, čo je vyjadrené nerovnosťou [21]:

$$V_j \leq \varepsilon \tag{8}$$

kde ε je arbitrálne daná hodnota ($\varepsilon > 0$), najčastejšie ε = 0,1 a V_j je variačný koeficient premennej X_j.

2.3 URČOVANIE TYPU PREMENNÝCH Z HĽADISKA ICH VPLYVU NA SLEDOVANÝ JAV

Premenné môžeme na základe toho, či sú pri analýze sledovaného javu žiaduce ich čo najvyššie, resp. najnižšie hodnoty, rozdeliť na stimulujúce premenné (stimulanty), destimulujúce premenné (destimulanty) a nominanty.

Stimulanty sú také premenné, ktorých vyššie hodnoty deklarujú vyššiu úroveň rozvoja sledovaného javu (pozitívny je rast hodnôt premennej). Destimulantami nazývae premenné, ktoré dokazujú vyššiu úroveň rozvoja sledovaného javu svojimi nízkymi hodnotami (pozitívnym javom je pokles hodnôt premennej) [27], [13]. Nominanty sú premenné, ktorých rastúce hodnoty vplývajú pozitívne na sledovaný jav, ale len po určitú hodnotu. Po prekročení tejto hodnoty je ich vplyv na sledovaný jav negatívny [3].

Určovanie charakteru premenných je založené na meritórnych úvahách. Správnosť určenia typu premennej možno overiť ex post. Vychádzame pritom z toho, že stimulanty sú skorelované pozitívne, podobne aj všetky destimulanty. Koeficienty korelácie medzi stimulujúcimi a destimulujúcimi premennými sú vždy záporné. Vzhľadom na charakter tretej skupiny premenných (nominanty) nie je možné takéto vzťahy overiť ex post.

Niektoré konštrukcie syntetických mier vyžadujú jeden charakter premenných, najčastejšie stimulujúci. Destimulujúce premenné sa preto upravujú na stimulujúce premenné. Pri ich transformácii možno použiť niektorý z týchto vzťahov [27]:

$$x_{ij}^S = 1 - x_{ij}^D \tag{9}$$

$$x_{ij}^S = \frac{1}{x_{ij}^D} \tag{10}$$

$$x_{ij}^S = c_j - x_{ij}^D \tag{11}$$

kde x_{ij}^S sú hodnoty stimulujúcej premennej, x_{ij}^D sú hodnoty destimulujúcej premennej ($i = 1, 2,..., m; j = 1, 2,..., k$).

V poslednom vzťahu je c_j konštantou, ktorá má spĺňať podmienku:

$$c_j \geq \max_i \left\{ x_{ij} \right\} \tag{12}$$

Iný spôsob zmeny charakteru premennej uvádza Malina [23]:

$$x_{ij}^{S} := 2\overline{x_j} - x_{ij}^{D} \tag{13}$$

Majewski navrhuje použiť vzťah [22]:

$$x_{ij}^{S} = -x_{ij}^{D} \tag{14}$$

Spôsob zmeny charakteru premenných nemôže byť náhodný. Treba dbať na to, aby sa premenná, ktorej charakter sa mení, vyznačovala podobnou variabilitou ako pôvodná premenná. Zmena charakteru premennej nesmie mať vplyv na výsledok usporiadania objektov [27].

2.4 URČOVANIE VÁH PREMENNÝCH

Ďalším problémom praktickej tvorby syntetickej premennej je určenie váh premenných. Aldenderfer a Blashfield tvrdia, že váženie premenných je vlastne manipulovaním s ich hodnotami [12]. Dufek a Minařík vo svojom článku uvádzajú, že problém voľby váh netreba, najmä pri veľkom počte ukazovateľov, preceňovať. Na druhej stane pripúšťajú, že voľba váh môže do istej miery predstavovať jeden z činiteľov, ktorý ovplyvní výsledky hodnotenia [6]. Abrahamovicz a Zajac vyjadrili potrebu váženia premenných: „Skúsenosť či intuícia ukazujú, že relatívna váha jednotlivých premenných nemôže byť rovnaká“ [1].

Ak máme analyzovať objekty na základe *k* ukazovateľov, potom by mali ich váhy spĺňať podmienku nezápornosti a ich súčet sa má rovnať 1 alebo *k.*

Metodika tvorby váh premenných rozlišuje vo všeobecnosti dva prístupy. Váhy môžu byť založené na expertných hodnoteniach, alebo sú východiskom pri ich tvorbe štatistické vlastnosti premenných [19].

Pri tvorbe váh na základe expertných hodnotení sa vychádza z hodnôt, ktoré priraďujú jednotlivým premenným odborníci v danej oblasti. Nevýhodou tohto systému je príliš vysoká diferenciácia pri hodnotení tej istej premennej [1]. Sem možno zaradiť aj tvorbu váh kvalitatívne porovnateľných premenných. Podrobne je spôsob tvorby týchto váh opísaný v [6], [10], [18]. Váhy tzv. kvalitatívne porovnateľných premenných sa využívajú hlavne v ekonomickej oblasti. Ide o situácie, keď nie sme schopní priamo stanoviť číselné hodnoty váh, ale môžeme určiť poradie, prípadne intenzitu preferencie jednotlivých premenných. Na tomto princípe je založená bodovacia metóda, metóda párového porovnania, Saatyho metóda, metóda najmenších štvorcov, metóda postupného rozvrhu váh, metóda stromu kritérií atď.

Pri formálno-štatistickom prístupe, ktorý vychádza z predpokladu, že „dôležitosť danej diagnostickej premennej je proporcionálna jej informačnému obsahu“, možno váhy tvoriť prakticky dvoma spôsobmi [1].

Prvý spôsob využíva variačný koeficient premenných. Tento spôsob tvorenia váh preferuje premenné s relatívne vysokou variabilitou. Váhy sa v tomto prípade tvoria podľa vzťahu [12]:

$$w_j = \frac{V_j}{\sum_{j=1}^{k} V_j}, \quad (j = 1, 2, .. k) \tag{15}$$

kde V_j je variačný koeficient premennej X_j[5].

Druhý spôsob je založený na tzv. mierach jednoznačnosti charakteru premenných, t. j. na „istote", na základe ktorej je premenná zaradená do množiny stimulujúcich, resp. destimulujúcich premenných. Jeho základom je korelačná matica premenných. Váhy sa tvoria na základe vzťahu [9]:

$$w_j = \frac{\left|\sum_{i=1}^{m} r_{ij}\right|}{\left|\sum_{j=1}^{k}\sum_{i=1}^{m} r_{ij}\right|} \tag{16}$$

kde r_{ij} je korelačný koeficient premenných X_i a X_j.

Korelačný systém váh preferuje premenné, ktoré sú silne skorelované s ostatnými premennými. Vysoké hodnoty váh nadobúdajú „centrálne" premenné, nízke hodnoty „izolované" alebo „satelitné" premenné.

Okrem systému rôznych váh sa v analýzach využíva systém rovnakých váh, pri ktorom všetky premenné vstupujú do analýzy ako rovnako dôležité:

$$w_j = \frac{1}{k} \quad (j = 1, 2, .. k) \tag{17}$$

Názory na používanie váh premenných sú rôzne. Kontroverzný je spôsob ich tvorby, nie je jasné, kedy používať systém rovnakých, kedy rozdielnych váh, prípadne, či vôbec premenné „prevažovať". Pri tvorbe syntetických premenných sa stretávame skôr s metódami, ktoré váhy nevyužívajú.

2.5 NORMOVANIE PREMENNÝCH

Jednou z úloh, ktoré treba riešiť vo viacrozmernej porovnávacej analýze je normovanie premenných, ktoré vstupujú do analýzy. Nevyhnutnosť ich normovania súvisí s tým, že jednotlivé premenné sú vyjadrené v rôznych merných jednotkách, prípadne ide o premenné, ktorých hodnoty sú rádovo rôzne.

Vo všeobecnosti možno normovanie vyjadriť vzťahom:

[5] *Spôsob výpočtu koeficienta V_j v závislosti od toho, na akej stupnici sú merané hodnoty premennej X_j uvádza [12].*

$$Z = \left(\frac{X - a}{b} \right)^p \tag{18}$$

kde X je pôvodná premenná, Z je normovaná premenná, a, b ($b \neq 0$), p sú parametre normovania. Najčastejšie sa pri normovaní používa spôsob, pri ktorom $p = 1$).

Výber parametrov a, b súvisí s požiadavkami, ktoré sa kladú na normovanú premennú, resp. na jej základné popisné charakteristiky. Vzhľadom na to, že pri výbere diagnostických premenných je rozhodujúcou variabilita ich hodnôt, pri transformovaní (normovaní) sa snažíme často o to, aby sa variabilita, vyjadrená variačným koeficientom, výrazne nezmenila.

Pri transformovaní hodnôt premennej X na hodnoty premennej Z použitím vzťahu (18), pri ktorom $p = 1$, sa jej základné popisné charakteristiky (priemerná hodnota, štandardná odchýlka a variačný koeficient) zmenia takto [21]:

$$\bar{z} = \frac{\bar{x} - a}{b} \tag{19}$$

$$s_z = \frac{s}{|b|} \tag{20}$$

$$V_z = \frac{s}{|\bar{x} - a|} \tag{21}$$

Podrobnosti o spôsoboch voľby parametrov normovania a a b tak, aby bola zachovaná variabilita pôvodných premenných ($V_x = V_z$) uvádza napríklad [21].

Najčastejšie sa normovaním pôvodných premenných sleduje splnenie nasledujúcich postulátov:

- aditívnosti (rôznorodé premenné – vyjadrené v rôznych jednotkách sú neporovnateľné, nemožno ich jednoducho agregovať, preto ich normovaním zbavíme prirodzených jednotiek),
- rovnakého charakteru (charakter premenných sa najčastejšie zjednotí prevodom destimulujúcich premenných na stimulujúce premenné)[6],
- nezápornosti (transformované premenné nadobúdajú nezáporné hodnoty),
- konštantného rozpätia (normované premenné nadobúdajú hodnoty z určitého dopredu stanoveného intervalu hodnôt, najčastejšie z intervalu $< 0;1 >$.

[6] *Niektoré metódy tvorby syntetickej premennej nevyžadujú rovnaký charakter premenných.*

Normovanie premenných vyjadrené vzťahom (18) možno previesť v závislosti od voľby parametrov normovania niekoľkými spôsobmi: štandardizáciou, unitarizáciou, transformáciou podielom a prevodom hodnôt na poradia.

Unitarizácia sa používa pri úprave hodnôt premenných, ktoré sú vyjadrené v rovnakých jednotkách, rádovo sú však rôzne. Hodnoty pôvodných premenných sú prepočítavané na hodnotu variačného rozpätia. Pri štandardizácii sa prepočítavajú na hodnotu štandardnej odchýlky. Tento spôsob normovania využíva metóda normovanej premennej a metóda vzdialenosti od fiktívneho objektu. Normovanie podielom má, podobne ako predchádzajúce metódy, niekoľko rôznych variantov. Jeden z nich využíva bodovacia metóda.

Samotná konštrukcia syntetickej premennej závisí od použitého spôsobu normovania pôvodných premenných a ich následnej agregácie. Tento príspevok opisuje najčastejšie používané metódy, ktoré nesú názov podľa toho, aký spôsob normovania pôvodných premenných bol použitý.

2.6 METÓDY TVORBY SYNTETICKEJ PREMENNEJ

Najčastejšie používanými metódami konštrukcie syntetickej premennej sú metóda poradí, bodovacia metóda, metóda normovanej premennej a metóda vzdialenosti od fiktívneho objektu. V článku sú uvádzané pre situácie, kedy je objekt opísaný aj stimulujúcimi aj destimulujúcimi premennými[7].

2.6.1 Metóda poradí

Metóda poradí je najjednoduchšou metódou vytvárania syntetickej premennej. Na základe hodnôt, ktoré nadobúda ukazovateľ X_j (j = 1, 2 ... k) na objektoch O_i (i =1, 2... m), objekty usporiadame. Ak je premenná X_j stimulujúca (pozitívny je rast jej hodnôt), poradie m priradíme tomu objektu, v ktorom táto premenná nadobúda maximálnu hodnotu. Poradie 1 je priradené objektu s najnižšou hodnotou tohto ukazovateľa. V prípade destimulujúcej premennej (pozitívny je pokles jej hodnôt) najvyššie *poradie m* priradíme objektu s najnižšou hodnotou a poradie 1 objektu, v ktorom daný ukazovateľ nadobudol najvyššiu hodnotu. Uvedený postup zopakujeme pre každý zo sledovaných ukazovateľov.

Pôvodné hodnoty premenných x_{ij} sa pretransformujú takto [21]:

$$z_{ij} = \begin{matrix} 1 \quad pre \quad \min\limits_i \{x_{ij}\} \, (j = 1, 2, \ldots k) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ m \quad pre \quad \max\limits_i \{x_{ij}\} \end{matrix} \quad \text{,ak je } X_j \text{ stimulujúca premenná} \quad (22)$$

[7] *Pri aplikácii týchto metód možno postupovať aj tak, že sa zjednotí charakter premenných a pracuje sa napríklad len so stimulujúcimi premennými. Z praktického hľadiska je tento postup vhodnejší.*

$$z_{ij} = \begin{cases} 1 \quad pre \quad \max_i \{x_{ij}\} (j = 1, 2, \ldots k) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ m \quad pre \quad \min_i \{x_{ij}\} \end{cases}$$,ak je X_j destimulujúca premenná (23)

Nedostatkom uvedenej metódy je prechod od „silnejšej“ stupnice hodnôt (číselnej) k stupnici „slabšej“ (poradovej). Ďalším problémom je to, že rozdiel dvoch po sebe idúcich normovaných hodnôt je maximálne 1 bez ohľadu na to, aký je skutočný rozdiel hodnôt pôvodnej premennej. Uvedené nedostatky sa pokúšajú odstrániť ďalšie u nás prakticky nepoužívané metódy, ako je napríklad metóda Capelanda [28].

Z takto určených poradí všetkých ukazovateľov vypočítame hodnotu syntetickej premennej:

$$d_i^{(1)} = \sum_{j=1}^{k} z_{ij} \text{ , } (i = 1, 2 \ldots m) \tag{24a}$$

Častejšie sa používa jednoduchý aritmetický priemer hodnôt z_{ij}, v tomto prípade tzv. priemerné poradie:

$$d_i^{(2)} = \frac{1}{k} \sum_{j=1}^{k} z_{ij} \text{ , } (i = 1, 2 \ldots m) \tag{24b}$$

Na základe hodnoty syntetickej premennej d_i určíme poradie jednotlivých objektov. V poradí prvý bude objekt s najvyššou hodnotou d_i, posledný bude objekt, ktorého hodnota syntetickej premennej d_i je najnižšia.

Opísaný postup tvorby syntetickej premennej vychádza z predpokladu rovnakej dôležitosti pozorovaných premenných (ukazovateľov). Pokiaľ by sa ukázalo potrebným priradiť jednotlivým ukazovateľom rôzne váhy, možno ich veľmi jednoducho zabudovať do vzťahov (24a) alebo (24b):

$$d_i^{(3)} = \sum_{j=1}^{k} z_{ij} w_j \text{ , } (i = 1, 2, \ldots m) \tag{25a}$$

$$d_i^{(4)} = \frac{1}{k} \sum_{j=1}^{k} z_{ij} w_j \text{ , } (i = 1, 2, \ldots m) \tag{25b}$$

kde w_j je váha j-teho ukazovateľa a z_{ij} sú hodnoty syntetickej premennej.

2.6.2 Bodovacia metóda

Pri aplikácii tejto metódy nahradíme hodnoty jednotlivých premenných X_j (j = 1, 2,... k) pozorované na objektoch Q_i (i = 1, 2,... m) príslušným počtom bodov. Pre každý ukazovateľ X_j nájdeme objekt, v ktorom tento ukazovateľ dosahuje maximálnu hodnotu, ak je pozitívnym javom rast hodnôt ukazovateľa, alebo minimálnu hodnotu, ak je pozitívnym javom pokles hodnôt ukazovateľa. Uvedenému objektu obvykle priradíme za uvedený ukazovateľ 100 bodov. Ostatné objekty získajú od 0 do 100 bodov, podľa toho, koľko percent predstavuje hodnota x_{ij} ukazovateľa pozorovaná na danom objekte z maximálnej hodnoty, resp. minimálnej hodnoty tohto ukazovateľa.

V prípade, že je pozitívnym javom rast hodnôt ukazovateľa priradíme objektom počet bodov podľa vzťahu [31]:

$$z_{ij} = \frac{x_{ij}}{x_{\max.j}} \cdot 100 \tag{26}$$

kde z_{ij} je počet bodov pre j-ty ukazovateľ v i-tom objekte, x_{ij} je hodnota j-teho ukazovateľa prislúchajúca i-tému objektu a $x_{max.j}$ je maximálna hodnota j–teho ukazovateľa.

V prípade, že je pozitívnym javom pokles hodnôt ukazovateľa, ako základ výpočtu použijeme jeho minimálnu hodnotu. Objektom priradíme počet bodov podľa vzťahu [31]:

$$z_{ij} = \frac{x_{\min.j}}{x_{ij}} \cdot 100 \tag{27}$$

kde z_{ij} je počet bodov pre j-ty ukazovateľ v i–tom objekte, x_{ij} je hodnota j-teho ukazovateľa prislúchajúca i–tému objektu a $x_{min.j}$ je minimálna hodnota j–teho ukazovateľa. Vyhodnotením úrovne ukazovateľov na jednotlivých objektoch a sčítaním ich bodového hodnotenia dostaneme celkové bodové hodnotenie, na základe ktorého možno objekty usporiadať (vzťah 24a alebo 25a).

V prípade opakovaného použitia bodovacej metódy s rozličným počtom ukazovateľov treba pri porovnávaní výsledkov použiť priemerný počet bodov (vzťah 12b alebo 13b). Priemerný počet bodov sa používa pri porovnávaní najčastejšie.

Existuje niekoľko modifikácií uvedenej metódy. Uvádzajú sa pod všeobecným názvom metódy využívajúce normovanie podielom. Líšia sa od seba tým, ktorú hodnotu považujeme za základnú. Za základnú hodnotu môžeme zvoliť aj priemernú hodnotu ukazovateľa. Body jednotlivým objektom priraďujeme potom podľa toho, aký je podiel ich hodnoty ukazovateľa na tejto priemernej hodnote. Ďalej možno vybrať jeden objekt a hodnoty ukazovateľov tohto objektu považovať za základné (každej hodnote priradíme 100 bodov) [21], [22].

2.6.3 Metóda normovanej premennej

Ukazovatele, ktoré sú vyjadrené v rôznych meracích jednotkách, prípadne ukazovatele, ktorých hodnoty sú rádovo rôzne, najčastejšie upravujeme na porovnateľný tvar normovaním.

V praxi je veľmi často používaným spôsobom normovania štandardizácia**,** ktorú možno, ak ide o stimulujúcu premennú, vyjadriť vzťahom [21], [22]:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \ (i = 1, 2,... \ m; \ j = 1, 2,... \ k) \tag{28}$$

V prípade destimulujúcej premennej sa potom používa vzťah:

$$z_{ij} = \frac{\bar{x}_j - x_{ij}}{s_j}, \ (i = 1, 2,... \ m; \ j = 1, 2,... \ k) \tag{29}$$

Normované premenné majú strednú hodnotu rovnajúcu sa nule a ich štandardná odchýlka sa rovná 1.

Hodnotu syntetickej premennej dostaneme, podobne ako pri ostatných metódach, ako súčet normovaných hodnôt ukazovateľov.

Výhodou metódy normovanej premennej oproti bodovacej metóde je skutočnosť, že okrem absolútnych rozdielov medzi jednotlivými objektmi zohľadňuje rôznu variabilitu jednotlivých ukazovateľov. Objekt, ktorý dosiahol „dobré“ umiestnenie, musí mať „dobré“ výsledky vo všetkých skúmaných ukazovateľoch, nestačí dosiahnuť výborný výsledok pri jednom alebo malom počte ukazovateľov.

2.6.4 Metóda vzdialenosti od fiktívneho objektu

Podstatou tejto metódy je vytvorenie fiktívneho (optimálneho) objektu O_0, v ktorom všetky destimulujúce ukazovatele nadobúdajú minimálnu hodnotu $\min_i \{x_{ij}\}$ a všetky ukazovatele so stimulujúcim charakterom maximálnu hodnotu $\max_i \{x_{ij}\}$ z hodnôt vyskytujúcich sa v súbore porovnávaných objektov $\{O_i\}$.

Hodnoty všetkých ukazovateľov sa najskôr vyjadria v normovanom tvare. Najčastejšie sa pri normovaní používa štandardizácia vyjadrená vzťahom (28) alebo (29).

Súradnice fiktívneho objektu O_0 sú potom takéto:

$$z_{0j} = \max_i \{z_{ij}\}, \ (i = 1, 2\ldots \ m), \text{ ak je premenná stimulujúca} \tag{30}$$

$$z_{0j} = \min_i \{z_{ij}\}, \ (i = 1, 2\ldots \ m), \text{ ak je premenná destimulujúca} \tag{31}$$

kde z_{ij} (i=1, 2... m; j = 1, 2... k) sú normované hodnoty ukazovateľov.

Pre každý objekt O_i (i = 1, 2… n) sa vypočíta priemerná vzdialenosť od tohto fiktívneho objektu O_0. Najčastejšie sa používa euklidovská vzdialenosť:

$$d_i = \left[\frac{1}{k} \sum_{j=1}^{k} \left(z_{ij} - z_{0j} \right)^2 \right]^{\frac{1}{2}} \tag{32}$$

Čím podobnejší je vybraný objekt fiktívnemu objektu, tým menšia je jeho vzdialenosť od neho. Najnižšiu dosiahnuteľnú hodnotu $d_{i0} = 0$ by dosiahol objekt, ktorý by vo všetkých ukazovateľoch nadobudol najlepšie hodnoty (z jeho hodnôt ukazovateľov by bol vymodelovaný fiktívny objekt) [14].

Konečné poradie objektov určíme tak, že najlepší objekt s poradím 1 bude ten, ktorý má najmenšiu vzdialenosť d_{i0} od optimálneho (fiktívneho) objektu, najhorší, s poradím m, bude ten, ktorý má najväčšiu vzdialenosť d_{i0} od fiktívneho objektu.

Vzhľadom na to, že táto metóda pracuje so štvorcami odchýlok, je citlivejšia na zmeny hodnôt ukazovateľov. V prípade, že sú rozdiely medzi analyzovanými objektmi výrazné, možno očakávať, že použitie tejto metódy bude viesť k rovnakému usporiadaniu objektov ako pri použití predchádzajúcich troch metód.

2.7 UKÁŽKA APLIKÁCIE METÓD PRI ANALÝZE ZADLŽENOSTI DOMÁCNOSTÍ

Uvedené metódy viacrozmerného porovnávania objektov, využívajúce rôzne spôsoby vytvárania syntetickej premennej, sme aplikovali v analýze zadlženosti domácností vo vybraných krajinách Európskej únie. Zadlžovanie domácností je vo vyspelých krajinách s modernými finančnými systémami v súčasnosti prirodzeným a bežným fenoménom. V posledných desaťročiach sa postoj k úverom zmenil a stal sa súčasťou modernej spotrebiteľskej spoločnosti [20]. Zámerom tejto časti príspevku je ilustrovať použitie opísaných metód. Použili sme len bodovaciu metódu, metódu poradí a metódu vzdialenosti od fiktívneho objektu.

Na sledovanie zadlženosti domácností v krajinách EÚ možno použiť údaje, ktoré pochádzajú zo Zisťovania o príjmoch a životných podmienkach domácností (EU SILC – European Union Statistics on Income and Living Conditions), údaje z Európskeho inštitútu pre výskum úverov (ECRI – European Credit Research Institute) alebo údaje zo Zisťovania o financiách a spotrebe domácnosti (HFCS – Household Finance and Consumption Survey).

V tomto článku boli pri viacrozmernej analýze zadlženosti krajín EÚ použité údaje pochádzajúce z druhej vlny zisťovania HFCS.

2.7.1 Zdroj údajov: Zisťovanie o financiách a spotrebe domácnosti (HFCS)

Zisťovanie o financiách a spotrebe domácnosti je spoločným projektom národných centrálnych bánk Eurosystému a národných štatistických úradov Francúzska, Fínska a Portugalska. Doteraz sa toto výberové zisťovanie uskutočnilo v troch vlnách. Prvá vlna

pokryla všetky krajiny v eurozóne okrem Írska a Estónska. Celkovú veľkosť vzorky tvorilo vyše 62 000 domácností, pričom v každej krajine sa jej veľkosť pohybovala medzi 340 a 15 000 domácností. Všetky výberové štatistiky boli prepočítané použitím váh na celú populáciu. Zisťovanie bolo realizované v období od konca roku 2008 do polovice roku 2011, prevažujúcim referenčným obdobím bol rok 2010. Druhá vlna HFCS poskytla harmonizované údaje o jednotlivých domácnostiach v 18 krajinách eurozóny (t. j. vo všetkých krajinách Eurozóny okrem Litvy), ako aj v Maďarsku a Poľsku. Zisťovanie bolo realizované na vzorke viac ako 84 000 domácností. Aj keď sa prieskum netýkal rovnakého časového obdobia vo všetkých krajinách, najbežnejším referenčným obdobím pre údaje bol rok 2014. Tretia vlna pokryla všetky krajiny eurozóny, plus Poľsko, Maďarsko, Chorvátsko, Rumunsko, Česko. Referenčným bol rok 2017. V čase, kedy boli robené analýzy prezentované v článku, neboli ešte dostupné údaje z tretej vlny zisťovania [7].

HFCS poskytuje podrobné údaje na úrovni domácností o rôznych aspektoch hospodárenia domácností a súvisiacich hospodárskych a demografických premenných vrátane príjmu, súkromných dôchodkov, zamestnanosti a mier spotreby. Cieľovou referenčnou skupinou prieskumu boli všetky súkromné domácnosti; neboli sem zahrnutí ľudia žijúci v kolektívnych domácnostiach a v inštitúciách, ako sú starší ľudia žijúci v inštitucionalizovaných domácnostiach [5], [29].

Hlavým ťažiskom záujmu HFCS je čisté bohatstvo domácností, ktoré odráža výšku celkových aktív a pasív domácností. Predmetom našej analýzy boli nesplatené záväzky domácností. Výšku nesplatených záväzkov domácností vyjadrujú celkové nesplatené záväzky domácností, ktoré sa skladajú z nesplatenej časti hypotekárneho (zabezpečeného) úveru, ktorý dlhujú domácnosti za všetky vlastnené nehnuteľnosti a nehypotekárneho úveru. Hypotekárny úver sa skladá z hypoték s hlavnou nehnuteľnosťou ako zábezpekou a hypoték zabezpečených ostatnými nehnuteľnosťami domácnosti. Nesplatený zostatok iných, nehypotekárnych úverov (celkový nezabezpečený dlh) zahŕňa nesplatené zostatky na kontokorentných účtoch, kreditných kartách, prečerpania limitov na kreditných kartách, za ktoré musí majiteľ platiť úroky a nesplatené zostatky na všetkých ostatných pôžičkách (lízingy na autá, spotrebné úvery, privátne pôžičky od príbuzných, známych, zamestnávateľov a pod.) [25].

2.7.2 Premenné použité pri analýze

Pri viacrozmernej analýze krajín, ktorej výsledkom bolo ich lineárne usporiadanie, boli použité vybrané ukazovatele zadlženosti domácností z druhej vlny zisťovania HFCS: DL1110i – domácnosť má hypotekárne úvery na hlavnú nehnuteľnosť (podiel domácností v %), DL1120i – domácnosť má hypotekárne úvery na inú nehnuteľnosť (podiel domácností v %), DL1200i – domácnosť má nehypotekárne úvery (podiel domácností v %) a ďalej sme použili premenné, ktoré vyjadrujú rozdelenie podielu zadlžených domácností podľa príjmových skupín (pod 20 %, 20 – 40 %, 40 – 60 %, 60 – 80 %, 80 – 90 %, 90 – 100 % príjmu) a podľa kategórií veku referenčnej osoby v domácnosti (16 – 34, 35 – 44, 45 – 54, 55 – 64, 65 – 74, 75+ rokov) [25]. Použili sme aj premenné DODARATIO (podiel dlhu k aktívam), DODIRATIO (podiel dlhu kpríjmom) a DOLTVRATIO (podiel dlhu k hodnote hlavnej nehnuteľnosti). (Údaje boli získané zo stránky Európskej centrálnej banky). Zoznam použitých premenných je v tabuľke č. 1.

Tabuľka č. 1: Označenie a definície premenných použitých vo viacrozmernej analýze

Označenie	Definícia
DL1000i	Podiel domácností, ktoré majú nesplatené záväzky
DL1100i	Podiel domácností, ktoré majú hypotekárny úver
DL1110i	Podiel domácností, ktoré majú hypotekárne úvery na hlavnú nehnuteľnosť
DL1120i	Podiel domácností, ktoré majú hypotekárne úvery na ostatné nehnuteľnosti
DL1200i	Podiel domácností, ktoré majú nehypotekárne úvery
Príjem pod 20 %	Podiel zadlžených domácností s príjmom v kategórii pod 20 %
Príjem 20 – 40 %	Podiel zadlžených domácností s príjmom v kategórii 20 – 40 %
Príjem 40 – 60 %	Podiel zadlžených domácností s príjmom v kategórii 40 – 60 %
Príjem 60 – 80 %	Podiel zadlžených domácností s príjmom v kategórii 60 – 80 %
Príjem 80 – 90 %	Podiel zadlžených domácností s príjmom v kategórii 80 – 90 %
Príjem 90 – 100 %	Podiel zadlžených domácností s príjmom v kategórii 90 – 100 %
Vek RO 16 – 34	Podiel zadlžených domácností, v ktorých je vek referenčnej osoby (RO) 16 – 34 rokov
Vek RO 35 – 44	Podiel zadlžených domácností, v ktorých je vek referenčnej osoby (RO) 35 – 44 rokov
Vek RO 45 – 54	Podiel zadlžených domácností, v ktorých je vek referenčnej osoby (RO) 45 – 54 rokov
Vek RO 55 – 64	Podiel zadlžených domácností, v ktorých je vek referenčnej osoby (RO) 55 – 64 rokov
Vek RO 65 – 74	Podiel zadlžených domácností, v ktorých je vek referenčnej osoby (RO) 65 – 74 rokov
Vek RO 75+	Podiel zadlžených domácností, v ktorých je vek referenčnej osoby (RO) 75+ rokov
DODARATIO	podiel dlhu k aktívam
DODIRATIO	podiel dlhu k príjmom
DOLTVRATIO	podiel dlhu k hodnote hlavnej nehnuteľnosti

Zdroj: [7], [25]

Pri určovaní charakteru premenných (stimulujúce, destimulujúce) sme vychádzali z toho, že ich smer pôsobenia je rovnaký. Ďalší metodický postup súvisiaci s tým, či s nimi pracujeme ako so stimulujúcimi alebo destimulujúcimi premennými, závisí od zorného uhla pohľadu na zadlžovanie domácností. Hoci rast hodnôt premenných indikuje narastanie zadlženosti a vzhľadom na túto skutočnosť ide o premenné stimulujúce, naše hľadisko považuje premenné za indikátory, ktorých rast zhoršuje z hľadiska prehlbujúcej sa zadlženosti, sociálnoekonomickú situáciu v krajine. Pri analýzach sme preto pracovali s uvedenými premennými ako s destimulantmi. V lineárnom usporiadaní krajín po aplikácii konkrétnej metódy usporiadania sa umiestnili na prvých miestach krajiny, v ktorých je zadlžovanie najnižšie. Vzťahy medzi jednotlivými premennými boli analyzované pomocou matice Pearsonových koeficientov korelácie.

Redukcia premenných bola uskutočnená v súlade s postupom metódy Hellwiga. Výsledkom tejto redukcie sú centrálne premenné DL1000i – podiel domácností, ktoré majú nesplatené záväzky (satelitné premenné: DL1100i, DL1110i, DL1200i, príjem pod 20 %, príjem 20 – 40 %, príjem 40 – 60 %, príjem 60 – 80 %, príjem 80 – 90 %, príjem

90 – 100 %, vek RO 16 – 34, vek RO 35 – 44, vek RO 45 – 54 a vek RO 55 – 64), DODARATIO – podiel dlhu k aktívam (satelitná premenná DOLTVRATIO), vek RO 75+ (satelitná premenná vek RO 65 – 74) a izolovaná premenná DODIRATIO – podiel dlhu k príjmom, ktoré boli použité pri usporiadaní krajín.

2.7.3 Lineárne usporiadanie krajín

Usporiadanie krajín podľa hodnôt syntetických premenných vytvorených metódou poradí, bodovacou metódou a metódou vzdialenosti od fiktívneho objektu je v tabuľke č. 2, po použití váh, ktoré vychádzajú zo vzťahov závislosti medzi jednotlivými premennými (vzťah 16) sú poradia usporiadaných krajín v tabuľke č. 3.

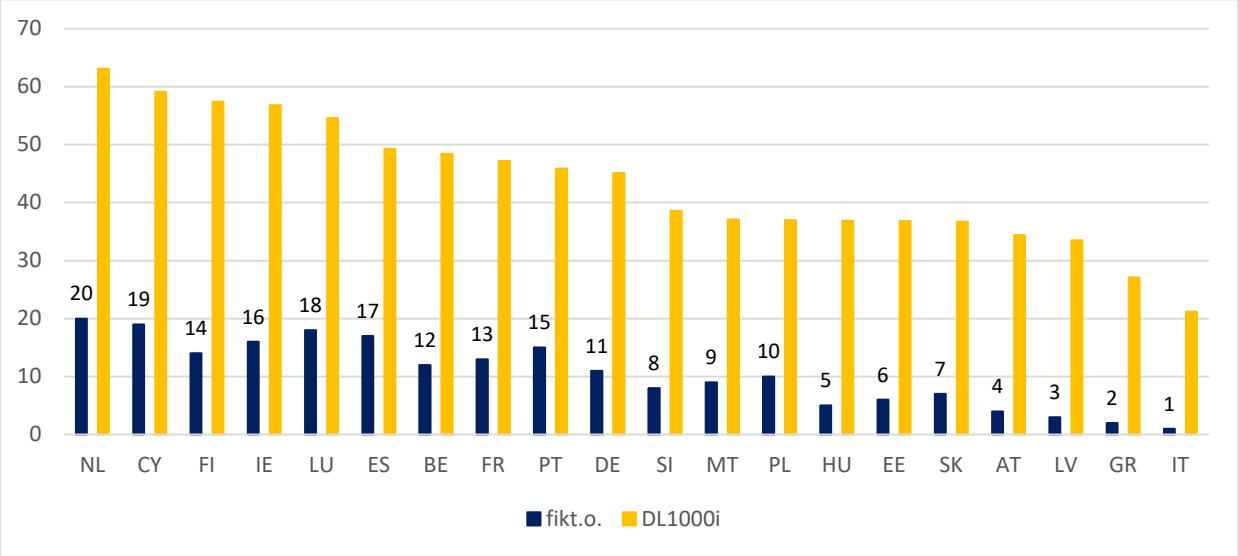
Tabuľka č. 2: Poradie krajín určené metódou poradí, bodovacou metódou a metódou vzdialenosti od fiktívneho objektu

Krajina	Metóda poradí		Bodovacia metóda		Metóda vzdialenosti od fiktívneho objektu	
	Hodnota	Poradie	Hodnota	Poradie	Hodnota	Poradie
AT	14	7	41.09	9	0.7863	4
BE	9.8	11	30.92	11	1.4021	12
CY	5	18	22.57	16	2.8421	19
DE	8	12-13	26.80	12	1.3787	11
EE	15.8	3	52.42	5	0.8999	6
ES	6	15	23.01	15	1.9510	17
FI	5.6	16-17	20.90	18	1.7651	14
FR	8	12-13	26.22	13	1.6373	13
GR	14.6	5	42.63	8	0.6388	2
HU	12.4	10	36.71	10	0.8396	5
IE	4.2	19	19.00	19	1.8940	16
IT	16	2	63.48	1	0.5730	1
LU	5.6	16-17	22.19	17	2.0328	18
LV	13.6	8	48.27	6	0.7025	3
MT	13.4	9	45.51	7	1.1866	9
NL	3.6	20	16.67	20	3.1018	20
PL	14.4	6	62.80	2	1.2537	10
PT	7.6	14	25.09	14	1.7813	15
SI	15.6	4	55.67	4	1.1208	8
SK	16.8	1	57.48	3	0.9360	7

Zdroj: vlastné spracovanie autorky (excel)

Vzhľadom na to, že prvou centrálnou premennou bola premenná DL1000i – podiel domácností, ktoré majú nesplatené záväzky, ktorá pri tvorbe syntetickej premennej „reprezentovala“ najväčší počet premenných, porovnali sme usporiadanie krajín vytvorené pomocou tejto premennej a poradie na základe výsledkov metódy vzdialenosti od fiktívneho objektu (graf č. 1). Poradia sa zhodujú na konci (krajiny s najhoršími výsledkami) a na začiatku usporiadania (krajiny s najlepšími výsledkami).

Graf č. 1: Porovnanie usporiadania krajín na základe premennej DL1000i a poradia určeného metódou vzdialenosti od fiktívneho objektu



Zdroj: vlastné spracovanie autorky (excel)

Tabuľka č. 3: Poradie krajín určené metódou poradí, bodovacou metódou a metódou vzdialenosti od fiktívneho objektu pri použití váh

Krajina	Metóda poradí vážená		Bodovacia metóda vážená		Metóda vzdialenosti od fiktívneho objektu vážená	
	Hodnota	Poradie	Hodnota	Poradie	Hodnota	Poradie
AT	2.8443	7	8.4343	9	0.3609	4
BE	1.9427	11	6.3389	11	0.6534	12
CY	0.9571	18	4.6578	17	1.2729	19
DE	1.6683	12	5.7028	12	0.6156	11
EE	3.1576	3	10.3844	5	0.4250	6
ES	1.2115	15	4.8646	15	0.8680	16
FI	1.0837	17	4.3680	18	0.8227	14
FR	1.6467	13	5.5656	13	0.7265	13
GR	2.9694	6	8.8397	8	0.2960	2
HU	2.4684	10	7.4744	10	0.3953	5
IE	0.8248	19	3.9977	19	0.8741	17
IT	3.1631	2	12.4476	2	0.2681	1
LU	1.1534	16	4.7024	16	0.9108	18
LV	2.6934	9	9.4894	7	0.3156	3
MT	2.7119	8	9.4948	6	0.5516	9
NL	0.6380	20	3.4421	20	1.3818	20
PL	2.9747	5	13.2805	1	0.5826	10
PT	1.4305	14	5.0675	14	0.8260	15
SI	3.1254	4	11.4503	3	0.5335	8
SK	3.3352	1	11.2879	4	0.4478	7

Zdroj: vlastné spracovanie autorky (excel)

Na prvých miestach sa umiestnili krajiny, v ktorých je zadlžovanie domácností, merané pomocou použitých premenných, najnižšie (IT, PL), na posledných miestach krajiny s najvyššou zadlženosťou domácností (NL, IE, CY). Najviac sa zhodujú poradia priradené jednotlivým krajinám pri aplikovaní metódy poradí a bodovacej metódy. Použitím váh pri metóde poradí sa zvýšila presnosť usporiadania. Najmenší vplyv na zmenu výsledkov malo prevažovanie normovaných hodnôt pri metóde vzdialenosti od fiktívneho objektu. Samozrejme, že spôsob usporiadania krajín závisí od relatívneho pohľadu na analyzovaný jav a je možné použiť presne opačný prístup k určovaniu poradia. Zaujímavé by bolo porovnanie výsledkov – poradí určených jednoduchými metódami viacrozmerného porovnávania s usporiadaním krajín pomocou hodnôt syntetických premenných, konštruovaných napríklad na princípe hlavných komponentov.

3. ZÁVER

V predloženom príspevku sme opísali princíp použitia jednoduchých metód viacrozmerného porovnávania. Ich podstata spočíva v tom, že hodnoty pôvodných premenných opisujúce vlastnosti nejakých objektov, sú pretransformované a následne agregované, čím dostávame hodnoty tzv. syntetickej premennej. Krajiny boli usporiadané na jej základe. Rozdiely medzi jednotlivými metódami súvisia s rôznym spôsobom transformácie hodnôt (pôvodné hodnoty sú nahradené poradím, bodmi, normovanými hodnotami alebo vzdialenosťami). Tri z týchto metód (metóda poradí, bodovacia metóda a metóda vzdialenosti od fiktívneho objektu) boli použité pri porovnávaní vybraných krajín, ktoré sa zapojili do druhej vlny zisťovania HFCS. Pri výbere premenných bola použitá metóda, ktorá je založená na analýze vzťahov medzi premennými pomocou koeficienta korelácie. Použitím váh, ktoré sú založené na hodnotách koeficientov korelácie sa poradie krajín výrazne nezmenilo.

LITERATÚRA

[1] ABRAHAMOWICZ, M. – ZAJAC, K.: Metoda wazenia zmiennych w taksonomii numerycznej i procedurach porzadkowania liniowego. In: Prace naukowe AE we Wroclawiu: Metody taksonomiczne. Wroclaw, 1986, č. 328.

[2] BARTOSZEWICZ, S.: Propozycja metody tworzenia zmiennych syntetycznych. In: Prace naukowe AE we Wroclawiu. Wroclaw, 1976, č. 84.

[3] BORYS, T.: Wezlowe problemy statystyki transgranicznej. Wroclaw: AE we Wroclawiu, 2000.

[4] CIEŚLAK, M.: Taksonomiczna procedura programowania rozwoju gospodarczego i określania potrzeb na kadry kwalifikowane. In: Przeglad Statystyczny, 1974, č. 1.

[5] CUPÁK, A. – STRACHOTOVÁ, A.: Výsledky druhej vlny zisťovania finančnej situácie a spotreby domácností. 2015. [online]. [cit. 2020-04-05]. Dostupné na: https://www.nbs.sk/_img/Documents/_komentare/AnalytickeKomentare/2016/AK39_HFCS2.pdf.

[6] DUFEK, J. – MINAŘÍK, B.: Poznámka ke stanovení vah ukazatelů. In: Statistika, 1984, č. 11, s. 486 – 489.

[7] ECB. The Household Finance and Consumption Survey: results from the second wave. 2016. [online]. [cit. 2020-04-05]. Dostupné na https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp18.en.pdf.

[8] GRABINSKI, T.: Wielowymiarowa analiza porównawcza w badaniach dynamiky zjawisk ekonomicznych. Zeszyty naukowe: Seria specjalna: Monografie. Kraków: AE, 1984, č. 61. 265 s.
[9] GRABINSKI, T. – WYDYMUS, S. – ZELIAS, A.: Metody prognozowania rozwoju spoleczno-gospodarczego. Kraków: AE w Krakowie, 1993.
[10] JABLONSKÝ, J. – FIALA, P. – MAŇAS, M.: Vícekriteriální optimalizace. Praha: SPN, 1985. 245 s.
[11] KEJKULA, J.: Možné přístupy porovnání zemí podle stupně životní úrovně. In: Statistika, 1981, č. 3, s.122 – 136.
[12] KROK, E.: Ocena mozliwości stosowania formalno-statystycznych systemów wag dla zmiennych w procese doboru kadr. In: Prace naukowe AE we Wroclawiu: Taksonomia 10: Klasyfikacja i analiza danych – teoria i zastosowania, 2003, č. 988.
[13] KŘOVÁK, J. Možnosti víceaspektního hodnocení podniků. In: Statistika, 1981, č. 6, s. 264 – 282.
[14] KŘOVÁK, J.: Možnosti víceaspektního hodnocení průmyslových organizací: studijní materiál č.136. Praha, 1982. 67 s.
[15] KŘOVÁK, J.: K problematice víceaspektního hodnocení průmyslových organizací. In: Statistika, 1983, č. 2, s.61 – 69.
[16] KŘOVÁK, J.: Vybrané metody víceaspektního hodnocení průmyslových organizací: výzkumná práce č. 96. Praha: 1983. 59 s.
[17] KŘOVÁK, J.: Jednoduché metody víceaspektního hodnocení v případě chybějících hodnot. In: Statistika, 1984, č. 9 – 10, s. 420 – 427.
[18] KŘOVÁK, J. – ŠTUDLAR, J.: Metody stanovení vah ukazatelů. In: Statistika, 1983, č. 12, s. 543 – 550.
[19] KURKIEWICZ, J. – POCIECHA, J. – ZAJAC, K.: Metody wielowymiarowej analizy porównawczej w badaniach rozwoju demograficznego: Monografie i Opracowania. Warszawa, 1991, č. 336.
[20] LEA, S. E. – WEBLEY, P. – WALKER, C. M.: Psychological factors in consumer debt: Money management, economic socialization, and credit use. In: Journal of Economic Psychology, 1995, č. 4, s. 681 – 701.
[21] LIPIETA, A. et al.: Taksonomiczna analiza przestrzennego zróznicowania poziomu zycia w Polsce w ujeciu dynamicznym. Krakow: Wydawnictwo AE w Krakowie, 2000. 295 s.
[22] MAJEWSKI, S.: Szeregowanie krajów przy pomocy Diagramu Czekanowskiego i Taksonomicznego Mierniku Rozwoju. In: Wiadomośći statystyczne, 1999, č. 8, s. 76 – 84.
[23] MALINA, A. – ZELIAŚ, A.: Taksonomiczna analiza przestrzennego zróznicowania. In: Prace Naukowe, Chrzanow: Wyźsa Szkola Przedsiebiorczósci i Marketingu w Chrzanowie, 1998, č. 2, s. 23 – 43.
[24] MLODAK, A.: Taksonomiczne mierniki przestrzennego zróznicowania rynku pracy. In: Wiadomośći statystyczne, 2002, č. 4, s. 16 – 25.
[25] NBS. Zoznam odvodených premenných v 2. vlne zisťovania HFCS (Household Finance and Consumption Survey). 2016.
[26] NGUYEN, T. H.: Poziom zycia ludności w Wietnamie. In: Wiadomośći statystyczne, 1999, č. 3, s. 83 – 93.
[27] NOWAK, E.: Metody taksonomiczne w klasifikacji obiektów spoleczno-gospodarczych. Warszawa: PWE, 1990. 201 s.

[28] NYKOWSKI, I.: O rankingach skończonego zbioru obiektów ocenianych wielokryterialnie. Kraków: WEA, 2001.17 s. ISSN 1230 – 1477.
[29] SENAJ, M. – ZAVADIL, T.: Výsledky prieskumu finančnej situácie slovenských domácností. 2012. [online]. [cit. 2020-04-05]. Dostupné na: http://www.nbs.sk/_img/Documents/PUBLIK/OP_1-2012_Senaj_Zavadil_hfcs.pdf.
[30] STANKOVIČOVÁ, I. – VOJTKOVÁ, M.: Viacrozmerné štatistické metódy s aplikáciami. Bratislava: Iura Edition, 2007. 261 s. ISBN 978-80-8078-152-1.
[31] STRAHL, D.: Propzycja konstrukcji miary syntetycznej. Przeglad Statystyczny, 1978, č. 2.
[32] STRAHL, D.: Modelowanie zjawisk zlozonych modele infrastruktury spolecznej: Praca habilitacyjna. In: Prace naukowe AE we Wroclawiu, 1980, zeszyt 158. 356 s.
[33] SYLABY, T.: Systemy vskaźnikow spolecznych w polskich warunkach transformacji rynkowej: Monografie I opracowania. Warszawa: SGH, 1994, č. 392.

RESUMÉ

Článok sa zaoberá jednoduchými metódami viacrozmerného hodnotenia a ich praktickou aplikáciou pri analýze zadlženosti domácností vybraných krajín EÚ. V prvej časti článku opisujeme princíp týchto metód. V druhej časti článku sa zaoberáme priestorovou analýzou dlhu domácností v krajinách, ktoré sa zúčastnili v druhej vlne prieskumu o financovaní a spotrebe domácností HFCS – Household Finance and Consumption Survey. Výsledky metódy poradí, bodovacej metódy a metódy vzdialenosti od fiktívneho objektu boli v závere porovnané.

RESUME

This article deals with simple methods of the multi-dimensional comparison and their practical application in the analysis of households‘ indebtedness of selected EU countries. In the first part of the article we describe the principle of these methods. In the second part, we deal with spatial analysis of household debt in the countries that participated in the second wave of the Household Finance and Consumption Survey (HFCS). We used data from the second wave of this survey to compare the countries. We compared the results of the ranking method, the scoring and the distance method from the fictitious object.

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 vedecko-výskumnej a pedagogickej činnosti sa venuje aplikácii štatistických metód pri analýzach sociálno-ekonomických javov, analýzam sociálno-patologických javov s osobitným zreteľom na výskyt chudoby, meraniu príjmovej nerovnosti, aplikácii metód hĺbkovej analýzy údajov, analýze kategoriálnych údajov a regionálnej štatistike.

KONTAKT

viera.labudova@euba.sk

Tatiana ŠOLTÉSOVÁ
Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky Ekonomickej univerzity v Bratislave
Jana KÚTIKOVÁ
Katedra štatistiky, Fakulta hospodárskej informatiky Ekonomickej univerzity v Bratislave

VYUŽITIE REGRESNEJ ANALÝZY PRI MODELOVANÍ ÚMRTNOSTI V ŽIVOTNOM POISTENÍ

THE USE OF REGRESSION ANALYSIS IN MODELING OF MORTALITY IN LIFE INSURANCE

ABSTRAKT
Cieľom príspevku je predstaviť a aplikovať vybrané parametrické modely úmrtnosti na modelovanie úmrtnosti žien a mužov na Slovensku v roku 2018. Vybrané modely úmrtnosti sú analyzované ako regresné modely, ktoré zahŕňajú regresnú funkciu a náhodnú zložku. Na modely úmrtnosti teda nazeráme ako na štatistické (stochastické) modely, pre ktoré na základe empirických údajov o pravdepodobnosti úmrtia v rôznom veku osôb (osobitne mužov a žien) je potrebné odhadnúť parametre regresnej funkcie. Parametre vybraných modelov úmrtnosti budeme odhadovať iteračnými metódami nelineárnej regresnej analýzy v softvéri SAS, resp. prostredníctvom jeho aplikácie SAS Enterprise Guide.

ABSTRACT
The aim of the paper is to introduce and apply the selected parametric mortality models to modelling of male and female mortality in Slovakia in 2018. The selected mortality models are analysed as regression models including regression function and a random component. We are looking at mortality models as statistical (stochastic) models for which it is necessary to estimate the parameters of regression function based on empirical data of age specific mortality rates (separately for men and women). The parameters of selected mortality models will be estimated by iterative methods of nonlinear regression analysis in the SAS software, through its application SAS Enterprise Guide.

KĽÚČOVÉ SLOVÁ
úmrtnosť, úmrtnostné tabuľky, parametrický model úmrtnosti, iteračná metóda, regresná analýza

KEY WORDS
mortality, life tables, parametric model of mortality, iterative method, regression analysis

1. ÚVOD
Všetky zmeny prebiehajúce v spoločnosti, teda aj demografické, sa priamo odrážajú v poisťovacích činnostiach. Z demografických ukazovateľov je pre životné poistenie dôležité sledovať v rámci populácie hlavne úmrtnosť. V súčasnosti, keď v celej Európe už niekoľko desiatok rokov prebieha starnutie populácie, je problematika sledovania vplyvu očakávanej dĺžky života populácie na dôchodkový systém často diskutovanou témou.

Nepriaznivý demografický vývoj, ktorý zasahuje aj Slovensko, je charakterizovaný nízkou pôrodnosťou a predlžovaním strednej dĺžky života. Tieto dva demografické javy (pôrodnosť a úmrtnosť), ktoré významne menia štruktúru populácie, majú v súčasnosti za následok starnutie populácie. Hlavnými príčinami starnutia populácie je teda znižovanie úmrtnosti (zlepšujúce sa životné podmienky a zdravotná starostlivosť), pokles pôrodnosti a v niektorých regiónoch je to aj emigrácia obyvateľstva v produktívnom, resp. reprodukčnom veku.

Priemerná dĺžka života za posledné desaťročia výrazne vzrástla a stále sa mení, čoho dôsledkom je dlhšia doba vyplácania predovšetkým doživotných poistných dôchodkov (napríklad z nasporenej sumy v II. pilieri pri odchode do dôchodku) ako aj ich zvýšená očakávaná súčasná hodnota. Oceňovanie produktov životného poistenia prebieha na základe súčasných očakávaní o budúcom vývoji úmrtnosti. Zmenou očakávanej úmrtnosti sa poisťovne vystavujú riziku, že ich záväzky z poistných produktov môžu výrazne prekročiť očakávanú hodnotu. Toto riziko vzniká pri poklese úmrtnosti v prípade dôchodkových produktov alebo pri zvýšení úmrtnosti v prípade produktov kryjúcich riziko smrti, a preto problematika vekového modelovania a analýzy úmrtnosti je a vždy bude dôležitou súčasťou práce v aktuárskej praxi.

2. PARAMETRICKÉ MODELY ÚMRTNOSTI

Hlavným parametrom modelov slúžiacich na stanovenie primeraného kapitálu na krytie rizík v životnom poistení je *ročná miera úmrtnosti*, ktorá predstavuje istú štatisticky stanovenú pravdepodobnosť úmrtia osoby vo veku x, resp. pravdepodobnosť, že osoba vo veku x sa nedožije nasledujúceho roka [16]. Ide o stochastický parameter, ktorého odchýlky a výkyvy môžu mať výrazný dosah na záväzky poisťovne, a tým aj na jej solventnosť. Tieto odchýlky a celkový vývoj miery úmrtnosti si vyžadujú neustálu pozornosť, pretože je od nich závislá finančná výkonnosť a stabilita poisťovní.

Údaje o vekovej úmrtnosti populácie potrebné na výpočty v životnom poistení sú uvedené v úmrtnostných tabuľkách, ktoré si môže poisťovňa vytvárať sama na základe svojich skúseností o úmrtnosti vo svojom poistnom portfóliu alebo môže použiť informácie o úmrtnostnom správaní populácie v danej krajine. Tieto informácie sa nachádzajú na webovej stránke Štatistického úradu Slovenskej republiky (ďalej „ŠÚ SR“) v časti s názvom Tabuľky života [18] a sú členené nielen podľa pohlavia, ale aj podľa regionálneho členenia (kraje, okresy, mestá a iné obce).

Zákony úmrtnosti, resp. parametrické modely úmrtnosti patria k deterministickým modelom. Predpokladáme pri nich, že hodnoty (napr. vekové miery úmrtnosti, intenzity úmrtnosti) pozorované v danom roku môžu byť preložené nejakým trendom, teda pomocou nejakej matematickej funkcie a môžeme tak predpokladať, že tento trend bude pokračovať aj v najbližších rokoch. Parametrické modely úmrtnosti sú použiteľné v populačných projekciách vďaka analýze historických trendov vývoja uvažovaných parametrov modelov. Napríklad v publikácii z roku 1979 sa v analýze historických zmien vývoja úmrtnosti švédskych mužov ukázalo, že v Gompertzovom-Makehamovom modeli úmrtnosti vykazovala zložka, ktorá je závislá od veku, historickú stabilitu, a to aj napriek rýchlemu poklesu zložky, ktorá je od veku nezávislá. Ďalšie analýzy v publikácii [4] potvrdili platnosť tohto javu a aj výskum historických časových radov zostavených

z údajov o úmrtnosti v 17 krajinách potvrdil tieto závery. Podľa Gavrilovovej a Gavrilova (2011) predstavujú parametrické modely úmrtnosti užitočný nástroj v demografických i aktuárskych prognózach úmrtnosti.

V príspevku odhadneme parametre nelineárnych funkcií parametrických zákonov úmrtnosti [1], ktorých predpis je určený:

- Heligmanovým-Pollardovým modelom,
- Kanistovým modelom,
- Coaleovým-Kiskerovým modelom.

Vybrané modely sme zvolili preto, lebo patria k najrozšírenejším a najviac používaným zákonom úmrtnosti, sú vhodné pre vyššie veky [11, 12], majú rôzny počet odhadovaných parametrov (od 2 do 8 parametrov) a sú to rôzne typy matematických funkcií.

Heligmanov-Pollardov model

Heligman a Pollard navrhli v roku 1980 niekoľko parametrických modelov úmrtnosti, ktoré sú schopné modelovať trend úmrtnostného správania ľudskej populácie pre všetky vekové kategórie a majú biologickú interpretáciu. Práve z tohto dôvodu patria k najrozšírenejším a najviac používaným zákonom úmrtnosti v demografii.

V našej aplikácii využijeme tzv. druhý tvar Heligmanovho-Pollardovho zákona úmrtnosti, ktorý opisuje mieru úmrtnosti, tzn. pravdepodobnosť, že *x* ročná osoba zomrie v priebehu roka. Druhý tvar tohto zákona vyjadruje mieru úmrtnosti takto:

$$q_x = A^{(x+B)^C} + De^{-E(lnx-lnF)^2} + \frac{GH^x}{1+GH^x} \qquad (1)$$

Heligman a Pollard zistili, že pre vyššie veky (v našom prípade 50 až 105 rokov) je vhodné použiť len tretí (logistický) člen modelu (1), ktorý vychádza z Gompertzovho zákona úmrtnosti. Teda pre mieru úmrtnosti q_x platí vzťah:

$$q_x = \frac{GH^x}{1+GH^x} \qquad (2)$$

ktorý prepíšeme pomocou parametrov *a*, *b* do tvaru:

$$q_x = \frac{ae^{bx}}{1+ae^{bx}} \qquad (3)$$

Uvedený vzťah sme použili aj v aplikácii, čiže odhadovali sme dva parametre *a*, *b* Helimanovho-Pollardovho modelu.

Kannistov model

Kannistov model (1994) patrí do skupiny logistických modelov [7], ktoré sú v poslednom čase čoraz viac populárne pri projekcii úmrtnosti. Tento model na rozdiel od predchádzajúceho modelu opisuje intenzitu úmrtnosti *x*-ročnej osoby μ_x. Ide o dôležitý

pojem v aktuárskej matematike súvisiaci s modelovaním budúcej dĺžky života osoby. Platí, že čím je hodnota μ_x väčšia, tým väčšia je pravdepodobnosť, že osoba vo veku x zomrie v krátkom časovom intervale. Intenzitu úmrtnosti môžeme opísať aj ako „rýchlosť vymierania“. V monografii [16] sú uvedené rôzne vzťahy jej vyjadrenia. Kannistov model vyjadruje intenzitu úmrtnosti v závislosti od dvoch parametrov a, b takto:

$$\mu_x = \frac{ae^{bx}}{1+ae^{bx}} \tag{4}$$

Aby sme mohli modelovať úmrtnosť z údajov ŠÚ SR, kde sú uvedené miery úmrtnosti, vyjadríme z aproximácie $\mu_x \sim -\ln(1-q_x)$, ktorá platí pre vyššie vekové kategórie [11] mieru úmrtnosti q_x a dostaneme vzťah s parametrami a, b:

$$q_x = 1 - e^{-\frac{ae^{bx}}{1+ae^{bx}}} \tag{5}$$

Coaleov-Kiskerov model

Coaleho-Kiskerov model (1990) bol alternatívou Gompertzovho-Makehamovho zákona úmrtnosti, ktorý zdôrazňoval zmenu miery úmrtnosti v dvoch po sebe idúcich rokoch a uvažoval so „spomalením“ úmrtnosti vo vysokom veku. Aplikáciu uskutočníme použitím nasledujúceho tvaru Coaleovho-Kiskerovho úmrtnostného modelu s troma parametrami a, b, c, pre ktorý platí:

$$q_x = e^{ax^2+bx+c} \tag{6}$$

Jeho prednosťou je predovšetkým vyššia flexibilita, ako aj numerická ustálenosť pri odhade parametrov.

3. ITERAČNÉ METÓDY ODHADU PARAMETROV NELINEÁRNYCH REGRESNÝCH MODELOV

V 2 kapitole sme opísali parametrické modely úmrtnosti, tzv. zákony úmrtnosti. Keďže zákony úmrtnosti sú vyjadrené nelineárnymi funkciami, ktoré obyčajne nie sú linearizovateľné (nedajú sa jednoduchými matematickými transformáciami previesť do lineárneho tvaru), na odhad parametrov týchto funkcií využívame nelineárnu regresnú analýzu. Nelineárna regresná analýza na rozdiel od lineárnej regresnej analýzy nevyužíva metódu najmenších štvorcov, ale iteračné metódy.

Opíšeme tri iteračné metódy používané na odhad parametrov nelineárnych regresných modelov. Patria k nim:

- Gaussova-Newtonova metóda,
- gradientná metóda,
- Levenbergova-Marquardtova metóda.

Metódy aplikujeme v softvéri SAS Enterprise Guide. V SAS Enterprise Guide sú v ponuke len niektoré nelineárne funkcie, preto sme museli urobiť zásah do

programovacieho kódu, kde sme požadované nelineárne funkcie zapísali v rámci procedúry nelineárnej regresie (*PROC NLIN*) v príkaze *MODEL*. Zároveň bolo potrebné zadefinovať parametre funkcie a zadať vstupné odhady parametrov, čo sme realizovali vďaka príkazu *PARMS*. Naším cieľom bolo porovnať aj výsledky jednotlivých iteračných metód, a preto sme nevyužili len štandardne nastavenú Gaussovu-Newtonovu metódu, ale prostredníctvom príkazu *METHOD* sme aplikovali aj gradientnú metódu a Levenbergovu-Marquardtovu metódu.

Regresná funkcia je matematická funkcia, ktorá je daná príslušným predpisom v zákone úmrtnosti, a jej parametre sú neznáme. Vo všetkých troch metódach sa uvažuje o nelineárnej regresnej funkcii s vektorom parametrov $\boldsymbol{\theta} = (\theta_1 \ \theta_2 \ ... \ \theta_p)$. Cieľom je získať také odhady parametrov regresnej funkcie, pre ktoré je súčet štvorcov odchýlok regresnej funkcie od hodnôt vysvetľovanej premennej čo najmenší, čo zapíšeme takto:

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[y_i - f\left(x_i, \theta\right) \right]^2 \rightarrow min. \tag{7}$$

Všetky metódy podľa [14] vychádzajú z počiatočných (vstupných) odhadov parametrov $\boldsymbol{\theta}^0 = (\theta_1^0 \ \theta_2^0 \ ... \ \theta_p^0)$, ktoré sa v jednotlivých iteráciách vylepšujú. *V*-te vylepšené odhady získané vo *v*-tej iterácii, pričom $v = 0, 1, 2...$, sa označujú takto $\boldsymbol{\theta}^v = (\theta_1^v \ \theta_2^v \ ... \ \theta_p^v)$. V každej iterácii sa počíta hodnota funkcie $S(\boldsymbol{\theta}^v)$, pričom by malo platiť $S(\boldsymbol{\theta}^{v+1}) \leq S(\boldsymbol{\theta}^v)$, čo znamená, že regresná funkcia s parametrami odhadnutými v iterácii (v+1) lepšie opisuje cieľovú premennú (vysvetľovanú premennú) ako regresná funkcia s parametrami odhadnutými vo *v*-tej iterácii.

Gaussova-Newtonova metóda (*metóda linearizácie*) využíva vo *v*-tej iterácii linearizáciu prostredníctvom Taylorovho rozvoja 1. rádu v bode $\boldsymbol{\theta}^v$:

$$f\left(\mathbf{x}_i, \boldsymbol{\theta}\right) \doteq f\left(\mathbf{x}_i, \boldsymbol{\theta}^v\right) + \sum_{j=1}^{p} \left\{ \left[\frac{\partial f\left(\mathbf{x}_i, \boldsymbol{\theta}\right)}{\partial \theta_j} \right]_{\boldsymbol{\theta} = \boldsymbol{\theta}^v} \cdot \left(\theta_j - \theta_j^v\right) \right\} \tag{8}$$

V každej iterácii ($v = 0, 1, 2, \ldots$) sa urobia substitúcie:

$$f_i^v = f\left(\mathbf{x}_i, \boldsymbol{\theta}^v\right) \qquad \beta_j^v = \theta_j - \theta_j^v \qquad z_{ij}^v = \left[\frac{\partial f\left(\mathbf{x}_i, \boldsymbol{\theta}\right)}{\partial \theta_j} \right]_{\boldsymbol{\theta} = \boldsymbol{\theta}^v} \tag{9}$$

čím sa získa lineárny regresný model:

$$\underbrace{y_i - f_i^v}_{y_i^v} = \sum_{j=1}^{p} \beta_j^v z_{ij}^v + \varepsilon_i \quad \text{resp.} \quad \mathbf{y}^v = \mathbf{Z}^v \cdot \boldsymbol{\beta}^v + \boldsymbol{\varepsilon}^v \tag{10}$$

Vektor parametrov $\boldsymbol{ß}^{v}$ tohto regresného modelu sa odhadne metódou najmenších štvorcov a z jeho odhadu $\mathbf{b}^{v}$ sa vypočíta vektor vylepšených odhadov pre iteráciu $(v + 1)$ takto:

$$\boldsymbol{\theta}^{v+1} = \boldsymbol{\theta}^{v} + \mathbf{b}^{v} \tag{11}$$

Celý proces pokračuje až do iterácie, v ktorej je splnené konvergenčné kritérium (bližšie pozri [12]). Nevýhodou tejto metódy je, že niekedy konverguje veľmi pomaly, čo vedie k veľkému počtu iterácií. V určitých prípadoch môže dokonca oscilovať. Vtedy sa rast a pokles súčtu štvorcov $S(\boldsymbol{\theta})$ opakuje až pokiaľ sa hodnota nestabilizuje.

Základná stratégia *gradientnej metódy* (*metódy najstrmšieho zostupu*) spočíva v iteračnom spôsobe hľadania globálneho minima funkcie $S(\boldsymbol{\theta})$. Začína sa v začiatočných odhadoch $\boldsymbol{\theta}^0 = (\theta_1^0\ \theta_2^0\ \ldots\ \theta_p^0)$, pre ktoré sa vypočíta súčet štvorcov $S(\boldsymbol{\theta})$. Hodnota parametra sa zvýši o malú hodnotu. Ak súčet štvorcov $S(\boldsymbol{\theta})$ poklesne, pokračuje sa vo zvyšovaní hodnoty parametra. Ak však hodnota $S(\boldsymbol{\theta})$ vzrastie, hodnota parametra sa vráti na pôvodnú úroveň a následne sa zníži. Tento postup sa opakuje veľakrát, pričom každý krok by mal viesť k zníženiu hodnoty súčtu štvorcov $S(\boldsymbol{\theta})$. Ak hodnota $S(\boldsymbol{\theta})$ namiesto toho vzrastie, tak krok bol veľký a preskočilo sa minimum funkcie $S(\boldsymbol{\theta})$. V takomto prípade sa veľkosť kroku zmenší a pokračuje sa uvedeným algoritmom. Proces sa končí, ak sa nájde bod, ktorý je podľa konvergenčného kritéria dostatočne blízky globálnemu minimu funkcie $S(\boldsymbol{\theta})$.

V porovnaní s Gaussovou-Newtonovou metódou gradientná metóda pracuje:

- lepšie v prvých iteráciách (zo zlých počiatočných odhadov dokáže efektívnejšie nájsť vhodný smer k minimu funkcie $S(\boldsymbol{\theta})$),
- horšie v posledných iteráciách (v oblasti okolo minima funkcie $S(\boldsymbol{\theta})$ často konverguje pomaly alebo osciluje).

Levenbergova-Marquardtova metóda sa niekedy označuje ako *Marquardtov kompromis*, čo vystihuje jej snahu spojiť prednosti oboch predchádzajúcich metód. V začiatočných iteráciách využíva gradientnú metódu a s približovaním sa k oblasti minima funkcie $S(\boldsymbol{\theta})$ postupne prepína na Gaussovu-Newtonovu metódu.

Podrobnosti o matematickom aparáte iteračných metód určených na odhad nelineárnych regresných modelov záujemcovia nájdu napríklad v [3, 6, 9, 10]. V týchto prácach sú uvedené aj možnosti aplikácie nelineárnej regresie v rôznych komerčných alebo open-source softvéroch. V príspevku budú na odhad nelineárnych regresných modelov aplikované všetky tri uvedené iteračné metódy, a to prostredníctvom procedúry nelineárnej regresie (*PROC NLIN*) v aplikácii Enterprise Guide štatisticko-analytického softvéru SAS.

4. ODHADY PARAMETROV MODELOV ÚMRTNOSTI V SR V ROKU 2018

Modelovanie úmrtnosti nelineárnou regresiou nám umožňuje určiť bodový ako aj intervalový odhad parametrov jednotlivých modelov zákonov úmrtnosti. V článku sú tiež prezentované bodové aj intervalové odhady pravdepodobností úmrtnosti žien, resp.

mužov získané pre jednotlivé vekové skupiny života na základe uvedených štyroch parametrických zákonov úmrtnosti.

Analýzu úmrtnosti uskutočníme pre populáciu žien a populáciu mužov na Slovensku v roku 2018 vo veku od 50 do 105 rokov, pretože vybrané modely sú vhodné pre vyššie veky. Použili sme miery úmrtnosti uvedené na stránke ŠÚ SR v Tabuľke života [18]. V tabuľke č. 1 a tabuľke č. 2 uvádzame počiatočné a výsledné odhady parametrov regresných funkcií modelov úmrtnosti (H-P – Heligmanov-Pollardov model, K – Kannistov model, C-K – Coaleov-Kiskerov model), osobitne pre ženy a pre mužov vo veku 50 až 105 rokov. Tieto súhrnné tabuľky navyše uvádzajú hodnoty funkcie $S(\boldsymbol{\theta})$ pri počiatočných a pri výsledných odhadoch parametrov. Na základe hodnôt $S(\boldsymbol{\theta}^v)$ výsledných odhadov parametrov vieme porovnať kvalitu modelov. Z výsledkov vyplýva, že vo vekovom intervale 50 až 105 rokov bol v roku 2018 z uvedených modelov najlepší Kannistov model, a to pre populáciu žien (tabuľka č. 1), ako aj pre populáciu mužov (tabuľka č. 2). Tabuľka č. 1 a tabuľka č. 2 porovnávajú aj jednotlivé iteračné metódy. Gradientná metóda nesplnila konvergenčné kritérium pri odhade žiadneho modelu. Poznamenajme, že pri zmenených vstupných odhadoch parametrov a ak by sme zvýšili maximálny počet iterácii, ktorý bol nastavený na hodnote 100, metóda by možno splnila konvergenčné kritérium, ale tomuto problému sme sa nevenovali, pretože Gaussova-Newtonova a Levenbergova-Marquardtova metóda dosiahli uspokojivé výsledky. Obidve iteračné metódy viedli k rovnakým výsledným odhadom parametrov modelov.

Tabuľka č. 1: Súhrnná tabuľka odhadov parametrov modelov úmrtnosti, porovnanie kvality modelov prostredníctvom funkcie $S(\boldsymbol{\theta})$ a porovnanie efektívnosti použitých iteračných metód pre populáciu žien vo veku 50 až 105 rokov

Charakteristika		**Model**		
		H-P	**K**	**C-K**
Počiatočné odhady	a^0	0,001	0,001	0,001
	b^0	0,001	0,001	0,001
	c^0	–	–	0,001
	$S(\boldsymbol{\theta}^0)$	4,9535	4,9326	1,38E10
Výsledné odhady	a^v	8,088E-9	1,137E-7	-0,00211
	b^v	0,1915	0,1537	0,5067
	c^v	–	–	-30,1198
	$S(\boldsymbol{\theta}^v)$	0,0135	**0,00122**	0,00293
v – počet iterácií	**Iteračná metóda**			
	Gauss-Newton	33	30	19
	Gradient	nekonverguje	nekonverguje	nekonverguje
	Marquardt	28	27	19

Vysvetlivky: H-K – Heligmanov-Pollardov model, K – Kannistov model,
C-K – Coaleov-Kiskerov model.

Zdroj: [18], spracované v SAS Enterprise Guide

Tabuľka č. 2: Súhrnná tabuľka odhadov parametrov modelov úmrtnosti, porovnanie kvality modelov prostredníctvom funkcie $S(\boldsymbol{\theta})$ a porovnanie efektívnosti použitých iteračných metód pre populáciu mužov vo veku 50 až 105 rokov

Charakteristika		Model		
		H-P	**K**	**C-K**
Počiatočné odhady	a^0	0,001	0,001	0,001
	b^0	0,001	0,001	0,001
	c^0	–	–	0,001
	$S(\boldsymbol{\theta}^0)$	3,8957	3,8753	1,38E10
Výsledné odhady	a^v	1,216E-6	3,819E-6	-0,00066
	b^v	0,1376	0,1159	0,2105
	c^v	–	–	-15,0651
	$S(\boldsymbol{\theta}^v)$	0,0107	**0,00239**	0,00246
***v* – počet iterácii**	**Iteračná metóda**			
	Gauss-Newton	32	21	20
	Gradient	nekonverguje	nekonverguje	nekonverguje
	Marquardt	26	24	20

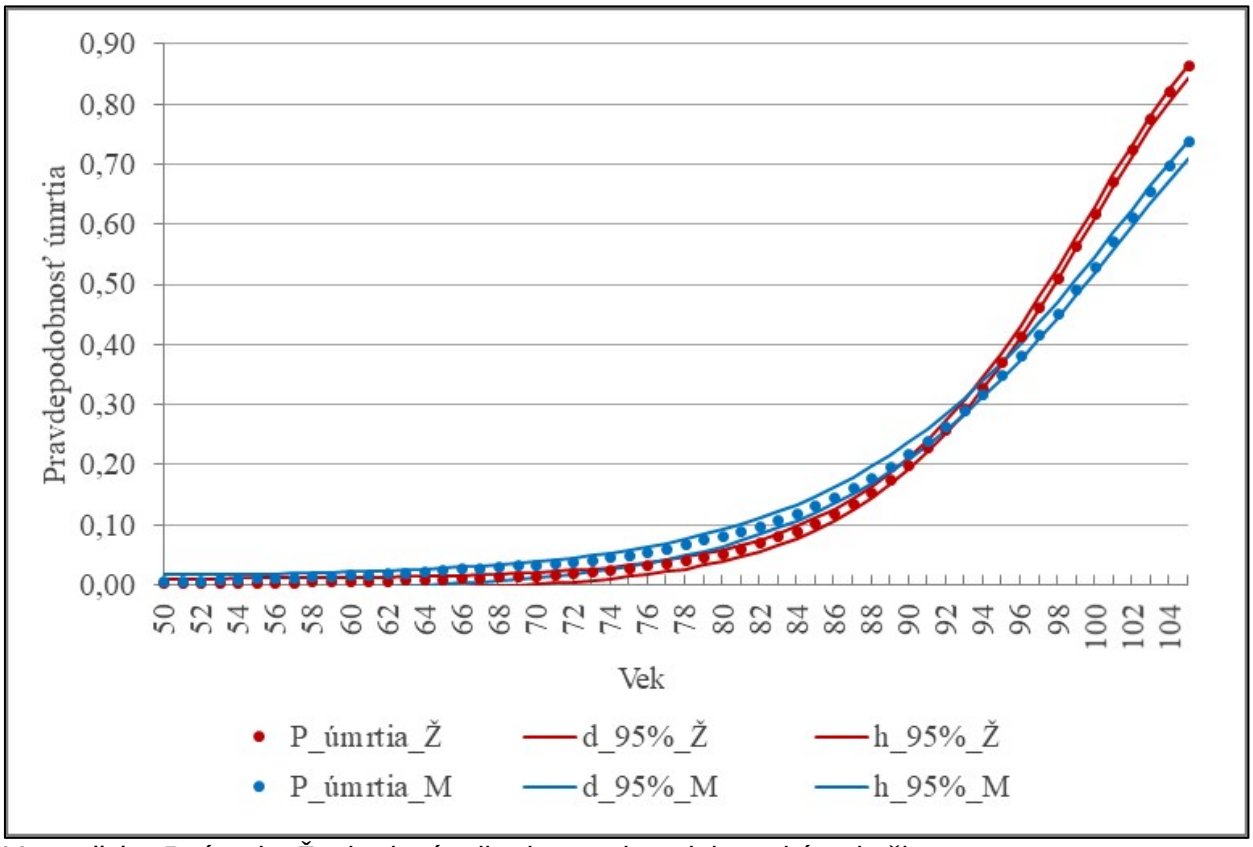
Vysvetlivky: H-K – Heligmanov-Pollardov model, K – Kannistov model,
C-K – Coaleov-Kiskerov model a CoDe – CoDe model.

Zdroj: [18], spracované v SAS Enterprise Guide

Ako sme uviedli, vo vekovom intervale 50 až 105 rokov sme najlepšie výsledky pre populáciu žien aj pre populáciu mužov v roku 2018 dosiahli Kanistovým modelom úmrtnosti. Preto sme na základe odhadnutých Kanistových modelov vypočítali aj 95 % intervaly spoľahlivosti pre pravdepodobnosti úmrtia vzťahujúce sa na analyzované roky veku života žien a mužov. Porovnanie týchto intervalov spoľahlivosti pre populáciu žien a populáciu mužov poskytuje graf č. 1.

V prípade populácie žien sme získali kvalitnejší model, pretože hodnota funkcie $S(\boldsymbol{\theta})$ bola v poslednej iterácii v populácii žien [$S(\boldsymbol{\theta}^v) = 0{,}00122$] nižšia ako v populácii mužov [$S(\boldsymbol{\theta}^v) = 0{,}00239$]. Túto skutočnosť odzrkadľujú aj intervalové odhady pravdepodobnosti úmrtností na grafe č. 1, ktoré sú v prípade populácie žien užšie ako v prípade mužov, a to vo všetkých sledovaných rokoch veku života.

Graf č. 1: 95 % intervaly spoľahlivosti pre pravdepodobnosti úmrtia žien a mužov vo veku od 50 do 105 rokov odhadnuté na základe Kannistovho modelu



Vysvetlivky: P_úmrtia_Ž – bodové odhady pravdepodobnosti úmrtia žien,
d_95 %_Ž; h_95 %_Ž – dolné (d) a horné (h) hranice 95 % intervalových odhadov pravdepodobnosti úmrtia žien,
P_úmrtia_M – bodové odhady pravdepodobnosti úmrtia mužov,
d_95 %_M; h_95 %_M – dolné (d) a horné (h) hranice 95 % intervalových odhadov pravdepodobnosti úmrtia mužov.

Zdroj: [18], spracované v SAS Enterprise Guide

Parametrické modely (zákony úmrtnosti) môžu byť užitočné aj v súvislosti s prognózovaním úmrtnosti populácie na základe analýzy historických trendov vývoja uvažovaných parametrov modelov. Istým obmedzením týchto modelov je však ich závislosť od konkrétneho matematického vzťahu, čo znemožňuje, aby model reagoval na možné zmeny vo vývoji úmrtnosti v budúcnosti.

5. ZÁVER

Hlavným parametrom modelov slúžiacich na stanovenie primeraného kapitálu na krytie rizík v životnom poistení je ročná miera úmrtnosti, ktorá predstavuje istú štatisticky stanovenú pravdepodobnosť úmrtia osoby vo veku *x*, resp. pravdepodobnosť, že osoba vo veku *x* sa nedožije nasledujúceho roka. Ide o stochastický parameter, ktorého odchýlky a výkyvy môžu mať výrazný dosah na záväzky poisťovne, a tým aj na jej solventnosť. Tieto odchýlky a celkový vývoj miery úmrtnosti si vyžadujú neustálu pozornosť, pretože je od nich závislá finančná výkonnosť a stabilita poisťovní.

Výhodou využitia nelineárnej regresnej analýzy pri odhade parametrov zákonov úmrtnosti je aj skutočnosť, že okrem bodových odhadov parametrov modelu získame štandardné chyby odhadov, ktoré umožňujú testovať štatistickú významnosť odhadnutých parametrov a konštruovať intervalové odhady pre parametre. Nelineárna regresná analýza navyše poskytuje možnosť realizovať intervalové odhady predikovaných hodnôt cieľovej premennej, ktoré prislúchajú konkrétnym hodnotám vysvetľujúcich premenných. Kannistove modely, ktoré boli najvhodnejšie na modelovanie mier úmrtnosti žien a mužov vo veku nad 50 rokov dobre vystihovali empirické miery úmrtnosti, čo potvrdili nielen nízke hodnoty súčtov štvorcov rezíduí, ale aj pomerne úzke 95 % intervalové odhady pravdepodobnosti úmrtia prislúchajúce analyzovaným rokom života.

LITERATÚRA

[1] FORFAR, D. O.: Mortality laws. Wiley StatsRef: Statistics Reference Online, 2014, s. 1 – 11.
[2] GAMPE, J.: Human mortality beyond age 110. In: Supercentenarians. Springer, Berlin, Heidelberg, 2010, s. 219 – 230.
[3] GAVIN, H. P.: The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems: Department of Civil and Environmental Engineering. Duke University, 2019.
[4] GAVRILOV, L. A. – GAVRILOVA, N. S.: The biology of life span: a quantitative approach. New York: Harwood Academic Publisher, 1991.
[5] GAVRILOVA, N. S. – GAVRILOV, L. A.: Stárnutí a dlouhověkost: Zákony a prognózy úmrtnosti pro stárnoucí populace. In: Demografie, 2011, č. 2, s. 109 – 128.
[6] GOULD, N. I. M. – REES, T. – SCOTT, J.: A higher order method for solving nonlinear least-squares problems. RAL Preprint RAL-P-2017-010, STFC Rutherford Appleton Laboratory, 2017.
[7] KANNISTO, V.: Development of oldest-old mortality, 1950-1990: Evidence from 28 developed countries (No. 1). University Press of Southern Denmark, 1994.
[8] MARQUARDT, D.: An Algorithm for Least Squares Estimation of Nonlinear Parameters. In: Journal of the Society for Industrial and Applied Mathematics, 1963, č. 2, s. 431 – 441.
[9] MONDRAGON, P. F. – BORCHERS, B.: A comparison of nonlinear regression codes. In: Journal of Modern Applied Statistical Methods, 2005, č. 1, s. 343 – 351.
[10] PANIK, M.: Regression modeling: Methods, theory, and computation with SAS. Chapman and Hall/CRC, 2009. ISBN 978-1420091977.
[11] PITACCO, E.: High age mortality and frailty. Some remarks and hints for actuarial modeling. CEPAR Working Paper, 2017, č. 1.
Available at: http://www. cepar. edu. au/working-papers/working-papers-2016. aspx.
[12] SAS Institute Inc.: The NLIN Procedure. In: SAS/STAT 13.1® User´s Guide. Cary, NC: SAS Institute Inc., 2013.
[13] SLUD, E. V.: Actuarial mathematics and life-table statistics. Chapman & Hall/CRC, 2012. ISBN 978-1439861974.
[14] ŠOLTÉS, E.: Regresná a korelačná analýza: s aplikáciami v softvéri SAS. Bratislava: Letra Edu, 2019. 238 s. ISBN 978-80-89962-38-9.
[15] ŠOLTÉSOVÁ, T.: Analýza úmrtnosti na Slovensku vo vzťahu k parametrickým modelom úmrtnosti. In: Softvérová podpora v predmetoch študijného programu Aktuárstvo: vedecká konferencia. Bratislava: EKONÓM, 2016, s. 76 – 81.

[16] ŠOLTÉSOVÁ, T.: Aktuárske modelovanie v životnom poistení. Bratislava: Vydavateľstvo Letra Edu, 2019. 148 s. ISBN 978-80-89962-36-5.
[17] ŠPROCHA, B. – MAJO, J.: Storočie populačného vývoja Slovenska I: demografické procesy. Bratislava: INFOSTAT, 2016.
[18] DATAcube, Štatistický úrad SR [online] [cit. 11. 3. 2020]. Dostupné na: https://slovak.statistics.sk/wps/portal/ext/themes/demography/population/indicators/!ut/p/z1/jdBBDoIwEAXQs3iCTkFpWRYMpWkDtFDAbgwr00TRhfH8GnRrYXaTvD-ZfOTQiNw8vfxlevr7PF0_-8kl5141NMswg9rgAjSArCTthM4JGhagifgCmvURCFIZLLXm0hLktuRzzso9UQBU8QMIVlqT6jgGFm_Lw59hsC0fAC58fkBuIaEGFlC0TcTSPc-Ppi5AdHlEW5VEwJMVAPgHQiWtvfm4WWtH8MKz3Ru7ePyp/dz/d5/L2dJQSEvUUt3QS80TmxFL1o2X1E3SThCQjFBMDhCVjIwSTdOUjFLUVFHSTky/.

RESUMÉ

V príspevku sme odhadli parametre troch vybraných parametrických modelov úmrtnosti (Heligmanov-Pollardov, Coaleov-Kiskerov a Kannistov model), tzv. zákonov úmrtnosti. Použili sme najaktuálnejšie dostupné údaje o miere úmrtnosti mužov a žien na Slovensku v roku 2018 zverejňované Štatistickým úradom Slovenskej republiky. Z dôvodu presnejšej analýzy sme uvedené modely použili zvlášť pre mužov a zvlášť pre ženy. Z údajov v intervale od 0 do 105 rokov sme modelovali a porovnávali úmrtnosť v období tzv. neskorej dospelosti a staroby, čiže vo veku od 50 rokov. Z použitých modelov mal pre obe pohlavia najlepšie výsledky Kanistov zákon úmrtnosti, a preto sme na základe odhadnutých parametrov Kanistových modelov vypočítali aj 95 % intervaly spoľahlivosti pre miery úmrtnosti prislúchajúce analyzovaným rokom života mužov a žien.
Odhad parametrov zákonov úmrtnosti sme uskutočnili pomocou nelineárnej regresnej analýzy v softvéri SAS Enterprise Guide, kde sme okrem bodových odhadov parametrov modelu získali aj štandardné chyby odhadov a určili sme intervalové odhady parametrov. Tieto odhady parametrov modelov môžu byť užitočné aj v súvislosti s prognózovaním úmrtnosti populácie na základe analýzy historických trendov ich vývoja.

RESUME

In this paper we estimated the parameters of three selected parametric models of mortality (Heligman-Pollard, Coale-Kisker and Kannist model), the so-called mortality laws. We used the most recently available data of the mortality rates of men and women in Slovakia in 2018 published by the Statistical Office of the Slovak Republic. For more accurate analysis, we used the above-mentioned models separately for men and for women. Based on data from the age interval 0 to 105 years, we modelled and compared the mortality in the period of the so-called late adulthood and old age, i. e. for ages 50 years and older. Out of these models used, he the Kannist law of mortality had the best results for both sexes and therefore based on the estimated parameters of the Kannist models, we also calculated 95% confidence intervals for mortality rates corresponding to the analyzed ages of men and women.
We estimated the parameters of mortality laws using nonlinear regression analysis in the SAS Enterprise Guide software, where in addition to the point estimates of model parameters, we also obtained standard errors and determined interval estimates for parameters. These estimates of model parameters may also be useful in connection with

forecasting population mortality based on an analysis of historical trends of their development.

PROFESIJNÝ ŽIVOTOPIS

***Doc. Mgr. Tatiana Šoltésová, PhD.**, od roku 1998 pôsobí na katedre matematiky (v roku 2011 premenovaná na katedru matematiky a aktuárstva) Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave. Titul PhD. získala v roku 2007 vo vednom odbore štatistika na FHI EU v Bratislave. V januári 2020 absolvovala habilitačné konanie v odbore kvantitatívne metódy v ekonómii, prepojenom na študijný odbor ekonómia a manažment. V rámci pedagogickej činnosti sa venuje výučbe matematickej analýzy a jej využitiu v ekonomických príkladoch, výučbe lineárnej algebry, finančnej a aktuárskej matematiky. Jej vedecká činnosť sa zameriava na využitie stochastického prístupu pri modelovaní v životnom poistení.*

***Ing. Jana Kútiková, PhD.,** pôsobí ako doktorandka v študijnom programe kvantitatívne metódy v ekonómii na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave. V roku 2019 ukončila štúdium v študijnom programe aktuárstvo na Fakulte hospodárskej informatiky a získala titul Ing. V súčasnosti vyučuje predmety na katedre štatistiky a venuje sa využitiu štatistických metód v životnom poistení.*

KONTAKT

tatiana.soltesova@euba.sk
jana.kutikova@euba.sk

Recenzia publikácie/Review of Publication

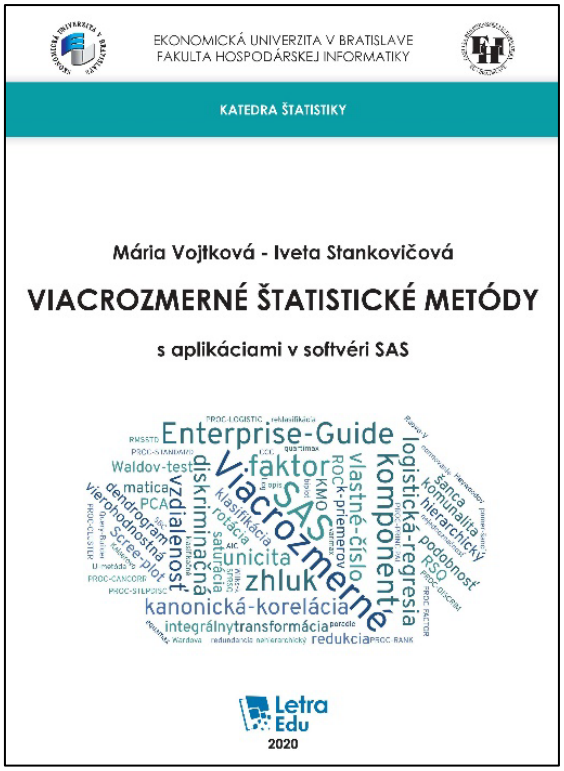
Mária Vojtková – Iveta Stankovičová
VIACROZMERNÉ ŠTATISTICKÉ METÓDY S APLIKÁCIAMI V SOFTVÉRI SAS

Mária Vojtková – Iveta Stankovičová
MULTIVARIATE STATISTICAL METHODS WITH APPLICATIONS IN THE SAS SOFTWARE

Druhé, doplnené a prepracované vydanie.
Bratislava: LetraEdu, 2020. 320 s.

ISBN: 978-80-89962-58-7 (print)
ISBN: 978-80-89962-59-4 (online)

Viacrozmerné štatistické metódy sa používajú na analýzu zložitých javov, ktoré vieme merať len pomocou viacerých ukazovateľov. Existujú rôzne definície vymedzenia viacrozmerných štatistických metód a ich rôzne klasifikácie, ale už podnázov knihy – *s aplikáciami v softvéri SAS* naznačuje, že táto vysokoškolská učebnica sa zameriava na viacrozmerné štatistické metódy, ktoré sú primárne označené prívlastkom viacrozmerné v štatisticko-analytickom softvéri SAS. Kniha sa teda zameriava na metódy ako sú metóda hlavných komponentov (2. kapitola), faktorová analýza (3. kapitola), kanonická korelačná analýza (4. kapitola), zhluková analýza (5. kapitola) a diskriminačná analýza (6. kapitola), ktoré sú začlenené do ponuky „Multivariate“ v module Enterprise Guide softvéru SAS, ale venuje sa aj metódam viackriteriálneho hodnotenia (1. kapitola) a logistickej regresii (7. kapitola). Zaradenie uvedených metód do tejto vysokoškolskej učebnice je determinované sylabami predmetu *viacrozmerné štatistické metódy*, ktorý je ovplyvnený podmieňujúcimi predmetmi, napr. *regresná a korelačná analýza* a nadväzujúcimi predmetmi, napr. *hĺbková analýza údajov* tak, aby tento predmet primerane prispieval k profilu absolventa študijného programu *štatistické metódy v ekonómii*, ktorý poskytuje Fakulta hospodárskej informatiky Ekonomickej univerzity v Bratislave na 2. stupni štúdia.



Autorkami druhého, doplneného a prepracovaného vydania vysokoškolskej učebnice *Viacrozmerné štatistické metódy s aplikáciami v softvéri SAS* sú doc. Ing. Mária Vojtková, PhD., a doc. Ing. Iveta Stankovičová, PhD., ktoré sa výučbe predmetných metód na svojich domovských fakultách (Fakulta hospodárskej informatiky EU v Bratislave a Fakulta managementu UK v Bratislave) a ich aplikácii vo svojej vedeckovýskumnej

činnosti venujú už viac ako 20 rokov. Ide o erudované odborníčky, o čom svedčí aj množstvo citácií ich publikačnej činnosti v oblasti viacrozmerných štatistických metód, pričom na 1. vydanie učebnice *Viacrozmerné štatistické metódy s aplikáciami*, ktorá bola vydaná v roku 2007 vo vydavateľstve Iura Edition, je v súčasnosti evidovaných viac ako 80 citácií, z ktorých približne polovica je v renomovaných citačných databázach Web of Science alebo SCOPUS.

Ako autorky uvádzajú, táto vysokoškolská učebnica je praktická pomôcka na zvládnutie vybraných metód viacrozmernej štatistickej analýzy. Každá z prvých 7 kapitol obsahuje:

- úvod do problematiky, opis príslušných metód a postupov a návod na interpretáciu výsledkov,
- ilustratívne príklady v SAS Enterprise Guide, na ktorých sa demonštrujú aplikácie príslušných metód a postupov,
- rekapituláciu riešenia príkladov v prostredí SAS Enterprise Guide, v ktorej je na ilustratívne príklady opísaný vytvorený projekt (vo forme súboru (.egp)) a programovací kód v programovacom jazyku SAS,
- opis možností príslušnej procedúry v systéme SAS.

Učebnica teda poskytuje metodický aparát príslušných metód a postupov, opis postupov ich adekvátnej aplikácie v SAS Enterprise Guide (vrátane interpretácie výsledkov), syntax a opis príslušných procedúr a ich možností v programovacom jazyku SAS, ktoré možno využiť vo všetkých moduloch softvéru SAS (nielen v SAS Enterprise Guide). Kniha je nielen učebnicou, ale aj manuálom na aplikáciu nižšie uvedených metód s využitím príslušných procedúr v jazyku SAS:

- metódy viackriteriálneho hodnotenia s využitím procedúr PROC RANK a PROC STANDARD,
- metóda hlavných komponentov (PROC PRINCOMP),
- faktorová analýza (PROC FACTOR),
- kanonická korelačná analýza (PROC CANCORR),
- zhluková analýza (hlavne PROC CLUSTER, ale aj PROC FASTCLUS, PROC MODECLUS, PROC VARCLUS, PROC TREE a PROC ACECLUS)
- diskriminačná analýza (hlavne PROC DISCRIM, ale aj PROC CANDISC a PROC STEPDISC),
- logistická regresia (PROC LOGISTIC).

Okrem uvedených procedúr je v učebnici využitá aj procedúra korelačnej analýzy (PROC CORR) a ďalšie procedúry ako sú PROC SORT, PROC STDIZE, PROC FREQ, PROC MEANS, PROC UNIVARIATE, PROC TTEST, PROC GPLOT, PROC GCHART, PROC PRINT.

Napriek tomu, že učebnica obsahuje syntaxe kódov v programovacom jazyku SAS, poskytuje návody aj na používanie aplikácie (resp. GUI – grafického používateľského rozhrania) SAS Enterprise Guide, ktorá je medzi používateľmi veľmi obľúbená, a to aj z toho dôvodu, že nevyžaduje programovanie. Pre tých, ktorí nemajú skúsenosti s prácou v SAS Enterprise Guide je určená 8. kapitola, v ktorej čitateľ nájde základné informácie

o prostredí tejto aplikácie, o práci s dátovými súbormi a nastaveniach na tvorbu výstupov v rôznych formátoch a štýloch.

Keďže mnohé sociálne a ekonomické javy sú multidimenzionálne a charakterizujú ich viaceré ukazovatele, indikátory alebo štatistické premenné, je prirodzené, že analytici a vedeckí pracovníci v oblasti sociálnych a ekonomických vied čoraz častejšie využívajú viacrozmerné štatistické metódy, vďaka ktorým je možné získať oveľa komplexnejší obraz o skúmaných javoch ako v prípade použitia jednorozmerných analýz. Vysokoškolská učebnica *Viacrozmerné štatistické metódy s aplikáciami v softvéri SAS*, aj vďaka vyššie uvedeným prednostiam, je určená nielen vysokoškolským študentom študijných programov, ktoré sa orientujú na kvantitatívne metódy, ale aj pre všetkých analytikov a vedeckých pracovníkov, ktorí skúmajú vzájomné vzťahy medzi kvantitatívnymi premennými v ekonómii a v iných vedných disciplínach.

Knihu považujem za veľmi dobrú vysokoškolskú učebnicu, ktorá využíva vhodné didaktické postupy na vysvetlenie metód a postupov vybraných viacrozmerných štatistických metód. Ide o komplexnú učebnú pomôcku, ktorá poskytuje návody na riešenie praktických úloh v profesionálnom štatistickom softvéri SAS, ktorý je v slovenskej praxi pomerne rozšírený.

***doc. Mgr. Erik ŠOLTÉS, PhD**.*

Autor je člen Katedry štatistiky a prodekan pre vedu a doktorandské štúdium na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave.

Názor/Opinion

MEDIÁN? PRIEMER? ALEBO KOREKTNÁ ANALÝZA PROBLÉMU? MEDIAN? AVERAGE? OR CORRECT ANALYSIS OF THE PROBLEM?

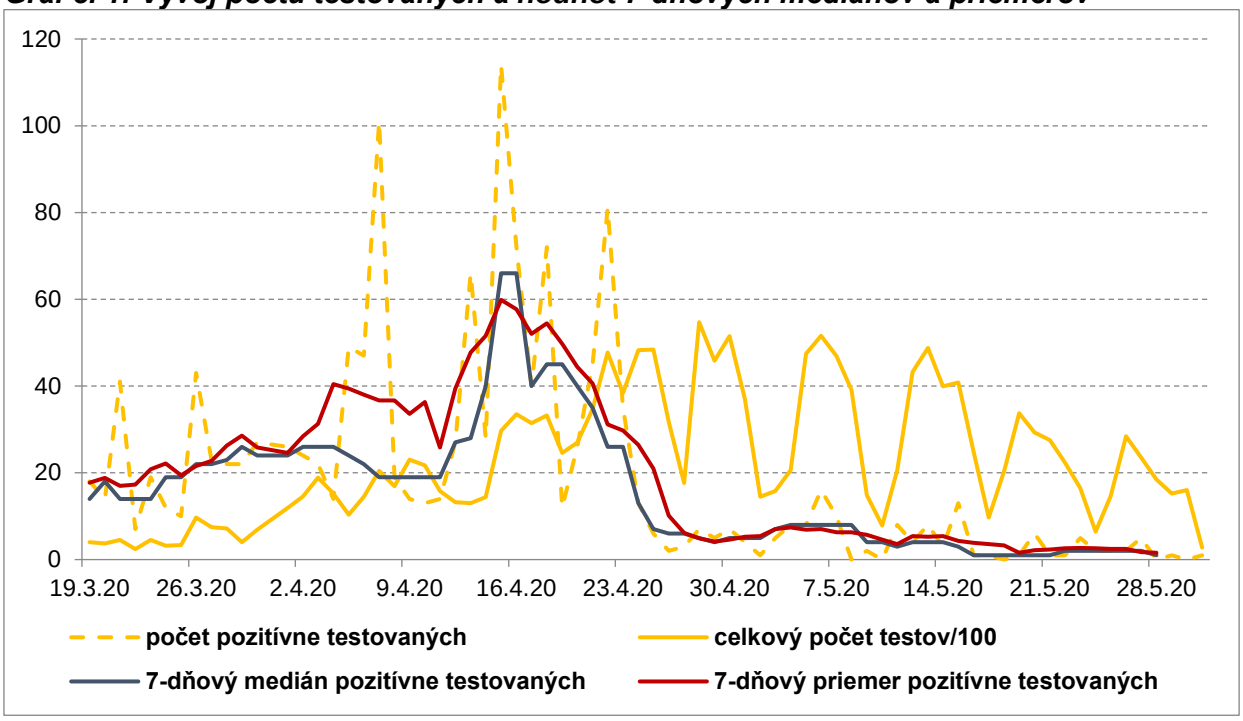
Koronavírus presiakol v posledných mesiacoch do všetkých oblasti nášho každodenného života. Ľudia venujú zvýšenú pozornosť neustále opakovaným bezpečnostným opatreniam Ústredného krízového štábu a vlády SR v boji s touto zákernou pandémiou. V spleti rôznych problémov, ktoré je potrebné priebežne riešiť, sa bežný občan len výnimočne pozastaví nad detailnejším obsahom jednotlivých nástrojov, ktoré sú súčasťou súboru prijímaných opatrení. Jedným z nástrojov, ktorý by mal podmieňovať uvoľňovanie opatrení prijatých v boji proti pandémii koronavírusu na Slovensku je tzv. kĺzavý medián[1] z denného počtu nových nakazených za uplynulý týždeň.

S podrobnejším zdôvodnením použitia kĺzavého mediánu na hodnotenie pandemickej situácie na Slovensku sme sa v žiadnom materiáli nestretli. Rozhodnutie preň sa pravdepodobne odvinulo od všeobecného konštatovania, že výhodou mediánu, ako popisnej štatistickej charakteristiky je skutočnosť, že nie je ovplyvnený extrémnymi hodnotami sledovaného znaku. Využitie mediánu pri analýze vývoja hodnôt konkrétneho znaku by však malo mať aj jasné vecné opodstatnenie.

V prípade hodnotenia vývoja počtu pozitívne testovaných na COVID-19 v podmienkach Slovenska je potrebné si uvedomiť, že denné počty sa získavajú z neustále sa dosť výrazne meniaceho počtu celkovo testovaných obyvateľov. **Sledovanie mediánových hodnôt počtu pozitívne testovaných preto korektne nereflektuje reálnu rizikovosť vývoja pandémie, a preto je uplatnenie takého prístupu vecne nesprávne.** Logiku by to malo len v prípade, keby sa denne realizoval zhruba rovnaký celkový počet testov na koronavírus. Podľa grafu č. 1 sa podľa tohto prístupu (na základe 7 dňových mediánov aj priemerov počtu pozitívne testovaných) Slovensko nachádzalo v najkritickejšej situácii v polovici apríla, čo nezodpovedá realite. Výraznejšie sa zvyšujúce počty pozitívne testovaných na koronavírus v prvých dvoch dekádach apríla boli značne relativizované rastúcimi počtami celkovo testovaných osôb. Grafické zobrazenie vývoja 7 dňového mediánu a priemeru počtu pozitívne testovaných v polovici apríla 2020 dokonca svedčí proti očakávaniu, že medián je menej ovplyvňovaný extrémnymi hodnotami počtu pozitívne testovaných ako aritmetický priemer.

[1] Podľa Google trends sa o tento pojem zaujímalo najviac ľudí 21. 4. 2020, keď bol tento ukazovateľ vyhlásený za kľúčový pre postupné otváranie slovenskej ekonomiky. Odvtedy už záujem oň klesá. Podľa Inštitútu pre stratégie a analýzy Úradu vlády SR sedemdňový kĺzavý medián denného prírastku infikovaných osôb, dokonca očistený o importované prípady či prípady v karanténnych oblastiach, nepoužíva okrem Slovenska nikto (pozri https://dennikn.sk/blog/1867449/klzavy-median-najsledovanejsie-cislo-buducich-tyzdnov/).

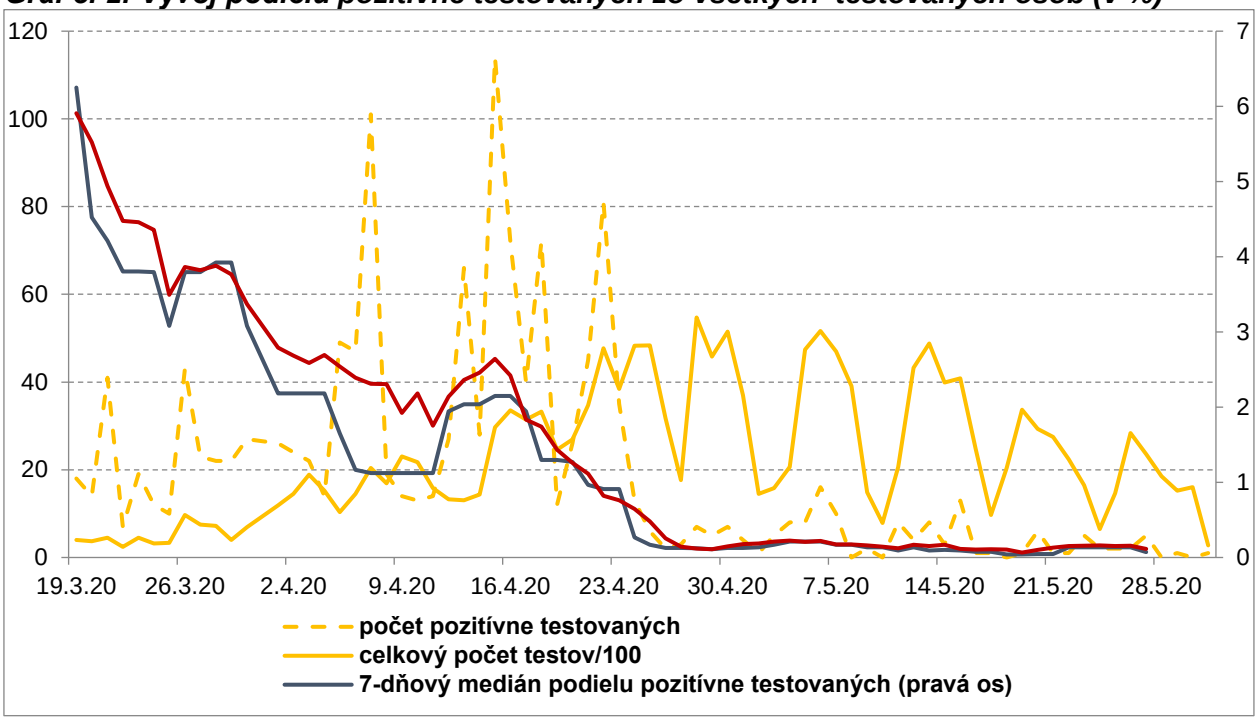
Graf č. 1: Vývoj počtu testovaných a hodnôt 7-dňových mediánov a priemerov



Zdroj údajov: MZ SR, vlastné prepočty

V pravidlách prechodu medzi jednotlivými fázami uvoľňovania slovenskej ekonomiky sú zadefinované tri možné varianty (medián do 100 nakazených, od 100 do 150 nakazených a nad 150 nakazených). Vzhľadom na existujúce poznatky o podmienkach a vývoji testovania na koronavírus i vzhľadom na prvé hodnoty kĺzavého mediánu, boli uvedené hranice uvoľňovania nastavené v druhej polovici apríla 2020 skôr intuitívne ako vedecky. **Použitá metodika kĺzavého mediánu nie je v danej situácii z nášho pohľadu vôbec odborne vyargumentovaná, a preto je diskutabilná a v konečnom dôsledku aj neakceptovateľná.**

Graf č. 2: Vývoj podielu pozitívne testovaných zo všetkých testovaných osôb (v %)



Zdroj údajov: MZ SR, vlastné prepočty

Reálnejší pohľad na vývoj rizika pandémie koronavírusu na Slovensku možno získať **prostredníctvom vývoja podielu pozitívne testovaných na celkovom počte testovaných osôb**. Podľa tohto prístupu bolo na Slovensku najväčšie riziko pandémie zaznamenané pred záverom druhej dekády v marci 2020, keď podiel pozitívne testovaných na celkovom počte testovaných osôb atakoval 10 %. Odvtedy zaznamenávame trend postupného znižovania rizika pandémie koronavírusu na Slovensku, čo je zrejmé aj z grafu č. 2.

Krivky hodnôt 7-dňového mediánu a priemeru podielu pozitívne testovaných z celkového počtu testovaných osôb v grafe č. 2 ukazujú, že obe majú podobný trend a na základe toho je veľmi ťažko favorizovať jednu z nich. Vedie to k poznatku, že prvoradé pri hodnotení vývoja spoločenských javov a procesov je správne vecné uchopenie problematiky a až následne treba venovať pozornosť výberu analytických nástrojov.

Trend znižovania rizika pandémie koronavírusu je na Slovensku tak podľa mediánu aj priemeru podielu pozitívne testovaných zrejmý a od konca apríla 2020 sa trvalejšie pohybuje hlboko pod úrovňou 0,5 %. Stále je však namieste zvýšená obozretnosť a osobná zodpovednosť každého obyvateľa pri dodržiavaní celého komplexu odporúčaní.

Zámerom vyjadrenia tohto názoru nie je vôbec kritika prijímaných centrálnych opatrení v boji s pandémiou koronavírusu, ale **upozornenie na nekritické a nedostatočne zdôvodnené používanie analytických postupov a nástrojov pri hodnotení reálnej situácie**. Platí to aj pri korektnom používaní mediánu ako analytického nástroja v ľubovoľnej spoločenskej oblasti. Akýkoľvek poznatok, záver či opatrenie vyznejú

dôveryhodnejšie, ak sú získané na základe nie deklarovaného, ale skutočného vedeckého prístupu.

Na záver si dovolíme len zopakovať, že **použitá metodika kĺzavého mediánu je v prípade hodnotenia rizikovosti pandémie koronavírusu na Slovensku** z nášho pohľadu **veľmi diskutabilná, nie je vôbec odborne vyargumentovaná, a preto je vecne neakceptovateľná.** Tento prístup nie je navyše použiteľný ani na žiadne medzinárodné porovnanie.

Ing. Mikuláš CÁR, PhD.

Autor je bývalý dlhoročný člen výboru Slovenskej štatistickej a demografickej spoločnosti.

Informácia/Information

ŠTÁTNA ŠTATISTIKA V OBDOBÍ PANDÉMIE COVID-19

STATE STATISTICS DURING THE COVID-19 PANDEMIC

Epidémia COVID-19 bezprecedentným spôsobom zasiahla celú spoločnosť a významne ovplyvnila prácu Štatistického úradu SR (ďalej len „ŠÚ SR“) s dosahom na štatistickú produkciu. Zdravie a bezpečnosť zamestnancov a ich rodín sa stali prioritou. Rovnako dôležité bolo zohľadnenie zdravotného rizika a administratívnej záťaže spravodajských jednotiek.

Na druhej strane požiadavka na zabezpečenie údajov a informácií na posúdenie ekonomických a sociálnych dôsledkov krízy postavila štatistické úrady pred nové výzvy, pričom bolo potrebné zabezpečiť aj štandardnú štatistickú produkciu.

Na ilustráciu ekonomického a sociálneho dopadu krízy COVID-19 na Slovensku a v Európe z pohľadu dimenzií kvality štatistických výstupov prioritu dostala včasnosť, spoľahlivosť a komplexnosť na úkor presnosti. V praxi to znamená publikovanie predbežných štatistických výstupov, čo si vyžiada ich následnú revíziu.

Na novovzniknutú situáciu okamžite reagoval aj ŠÚ SR. Predseda ŠÚ SR zriadil svojím príkazom komisiu na riešenie otázok pri výkone štátnej štatistiky v súvislosti s mimoriadnou situáciou vyhlásenou uznesením vlády Slovenskej republiky.

Na základe jej rozhodnutia prijal ŠÚ SR viaceré opatrenia:
– niektoré zisťovania vyžadujúce osobné návštevy boli pozastavené,
– pri ďalších zisťovaniach bolo osobné opytovanie nahradené telefonickým,
– niektoré zisťovania, pri ktorých je to možné, majú odsunuté termíny zberu údajov na letné, resp. jesenné mesiace a pri niektorých sa zvažuje posun zberu na koniec roka,
– pri zisťovaní spotrebiteľských cien boli prijaté opatrenia na zabezpečenie náhradných foriem získania údajov (z internetu, webscrappingom, z transakčných údajov obchodných reťazcov a iných dostupných náhradných externých zdrojov).

ŠÚ SR pokračuje v príprave Sčítania obyvateľov, domov a bytov v roku 2021 (SODB 2021). Vzhľadom na to, že zber údajov o domoch a bytoch realizovaný obcami sa začal v júni 2020, ŠÚ SR zohľadnil mimoriadny stav dištančným vzdelávaním, kontaktom s obcami formou e-mailov a telefonátov, diseminá ciou informácií formou videokonferencií a inštruktážnych videí.

V troch prierezových štvrťročných zisťovaniach (zisťovania produkčných odvetví vo veľkých a malých podnikoch a zisťovanie o práci), ktoré sa dotýkajú najväčšieho počtu podnikov a organizácií (cca. 30 000 subjektov), ŠÚ SR upravil rozsah spravodajských povinností a obrátil sa na respondentov so žiadosťou o vyplnenie aspoň kľúčových ukazovateľov, ktoré predstavujú 23 % až 30 % z celkového počtu požadovaných premenných.

Všetky prijaté opatrenia mali dopad na prípadné zabezpečenie náhradných zdrojov údajov a na spôsob spracovania štatistických údajov, pričom bolo potrebné urýchlene implementovať nové metodické postupy a matematicko-štatistické metódy.

ŠÚ SR je v intenzívnej komunikácii a kontakte s Eurostatom pri analýze špecifík národnej situácie a určovaní a zabezpečovaní priorít štatistickej produkcie v rámci plnenia záväzkov Slovenskej republiky voči Európskej únii.

V oblasti legislatívy bol na rokovanie vlády Slovenskej republiky predložený a následne v Národnej rade SR schválený Zákon č. 107/2020 Z. z., ktorým sa dopĺňa zákon č. 540/2001 Z. z. o štátnej štatistike v znení neskorších predpisov s účinnosťou od 5. 5. 2020.

Naliehavosť prijatia novely zákona vyplynula z aktuálnej situácie z dôvodu šírenia ochorenia COVID-19 na území Slovenskej republiky, ktorá si vyžadovala dočasné opatrenia v oblasti štátnej štatistiky umožňujúce spravodajským jednotkám sústrediť sa na riešenie najvážnejších dopadov pandémie na ich vlastnú činnosť. Cieľom novely zákona bola pomoc podnikateľom a ostatným spravodajským jednotkám spočívajúca v možnosti orgánov vykonávajúcich štátnu štatistiku vydávať opatrenia, ktoré v súčasnej vládou vyhlásenej mimoriadnej situácii a bezprostredne po nej modifikujú ich spravodajskú povinnosť.

Je však potrebné upozorniť, že dôležitosť plnenia spravodajskej povinnosti sa vyhlásením výnimočného stavu, núdzového stavu alebo mimoriadnej situácie nezmenšuje, pretože získané štatistické údaje sú podkladom na analýzu sociálno-ekonomického vývoja spoločnosti a pre svoju objektívnu výpovednú hodnotu umožňujú prijímať verejnosťou očakávané, adekvátne opatrenia v hospodárskej politike a iných verejných politikách. Uvedené znamená, že opatrenia orgánov vykonávajúcich časť štatistických zisťovaní, ktoré spadajú pod aktuálny Európsky štatistický program, musia byť v súlade s príslušnými právne záväznými aktami Európskej únie v oblasti európskej štatistiky.

Nad rámec programu štátnych štatistických zisťovaní ŠÚ SR vykonal 14. – 20. apríla 2020 prieskum zameraný na zmapovanie aktuálneho stavu podnikateľského prostredia po vyhlásení mimoriadnej situácie na území Slovenskej republiky v súvislosti so šírením ochorenia COVID-19. Cieľom tejto iniciatívy bolo prispieť k lepšej, objektívnej informovanosti vlády Slovenskej republiky a verejnosti o dopadoch, ktoré aktuálne ovplyvnili, resp. ovplyvnia vývoj slovenskej ekonomiky v najbližšej budúcnosti.

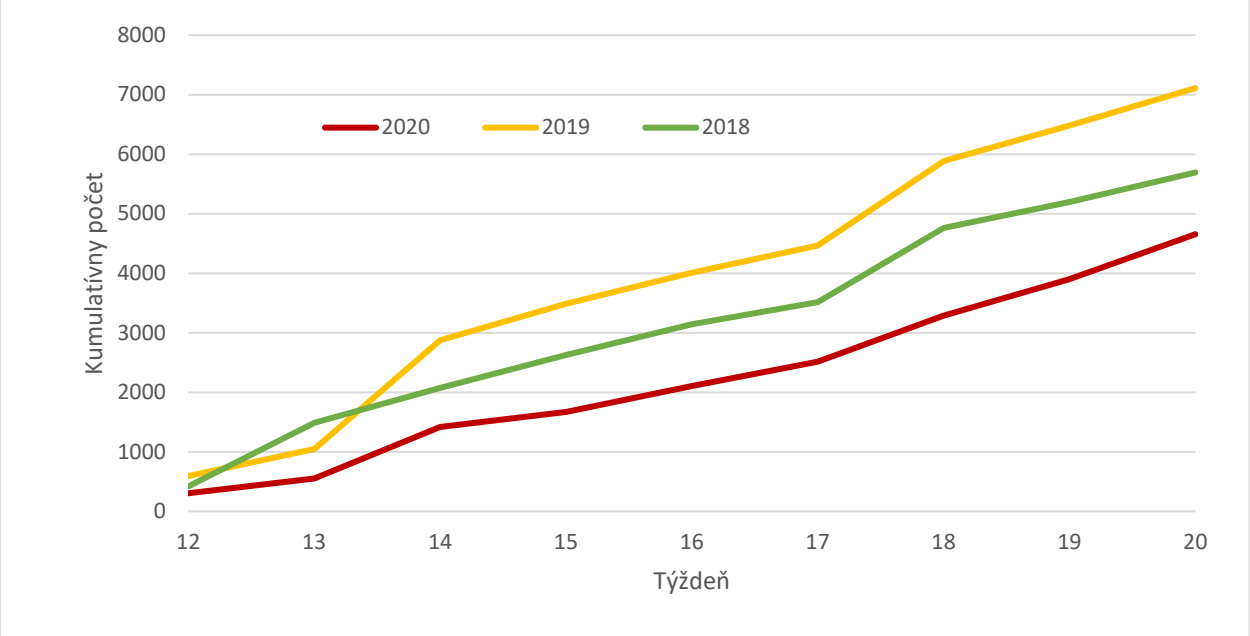
V rámci prieskumu, ktorý bol organizovaný na báze dobrovoľnosti, bolo oslovených spolu 1690 podnikateľských subjektov z odvetvia obchodu, priemyslu, stavebníctva a vybraných trhových služieb. Otázky položené v prieskume sa týkali vývoja tržieb od vyhlásenia mimoriadnej situácie a predpokladaného vývoja v oblasti zamestnanosti a tiež budúcich zámerov podnikateľských subjektov v oblasti vlastnej podnikateľskej činnosti. Napriek pretrvávajúcej zložitej situácii, ŠÚ SR zaznamenal odpovede od 434 respondentov.

V oblasti tržieb najhoršia situácia bola zistená v odvetví obchodu, kde najviac podnikov deklarovalo pokles tržieb o viac ako 50 % v kategórii s počtom zamestnancov 20 – 249. Z hľadiska vývoja zamestnanosti podľa odvetví a počtu zamestnancov podnikateľské subjekty zo všetkých odvetví zhodne deklarovali pokles zamestnanosti do 25 % v kategórii 20 až 249 zamestnancov. O dočasnej zmene zamerania podnikateľskej činnosti uvažovalo len 3,9 % odpovedajúcich subjektov, hlavne podniky v odvetví obchodu (6,5 %). Najviac podnikov (57 %) uvažovalo o skrátenom režime v odvetví priemyslu, naopak, najmenej (28,8 %) v stavebníctve. Pomerne vysoké percento (až 38,2 %) podnikov neuvažovalo o skrátenom režime vo svojom podniku. Najväčší záujem o „Kurzarbeit“ deklarovali subjekty v odvetví priemyslu (42,4 %). Z celkového počtu odpovedajúcich subjektov 45,6 % ešte nevedelo, či ho využije.

Na základe údajov obsiahnutých v Registri právnických osôb, podnikateľov a orgánov verejnej moci (RPO) boli spracované informácie o vývoji oprávnení na podnikanie fyzických osôb so stavom k 15. 5. 2020. Na účely porovnania vývoja boli použité údaje za obdobie 1. január až 15. máj za roky 2018, 2019 a 2020 o všetkých fyzických osobách – podnikateľoch (SZČO).

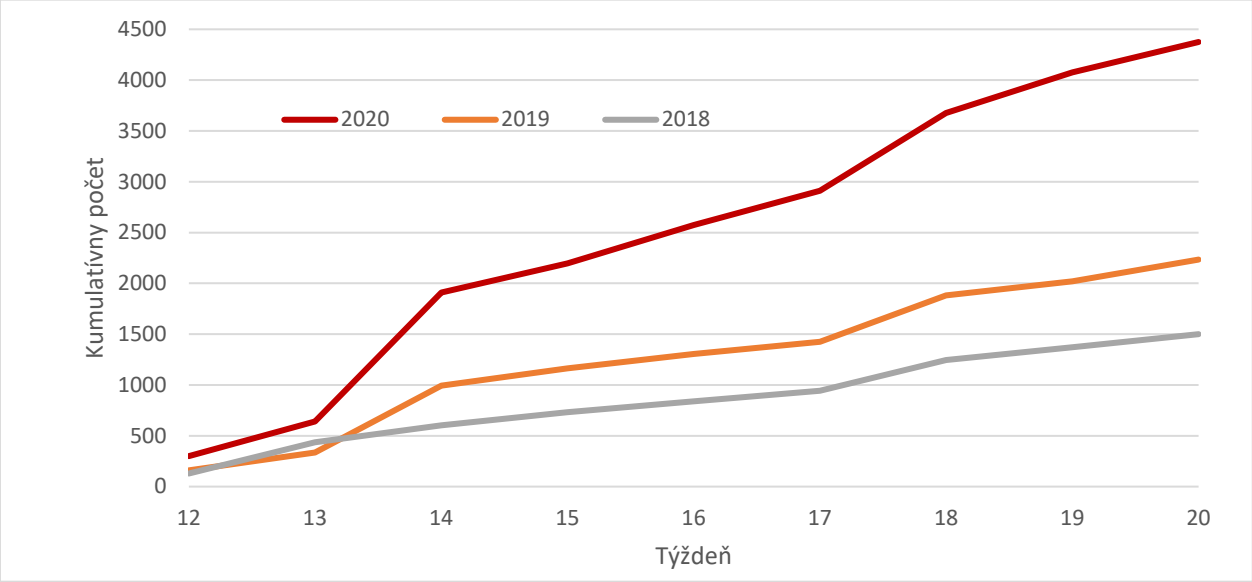
Informácie sa týkali životných situácií typu: vznik, pozastavenie a zánik oprávnenia na podnikanie. Bolo zistené, že vývoj uvedených životných situácií je silne ovplyvnený vývojom v živnostenskom registri, keďže živnostníci tvoria zo skupiny všetkých podnikateľov viac ako 90 %.

Graf č. 1: Vznik nových živností



Zdroj údajov: ŠÚ SR – RPO

Graf č. 2: Pozastavenie živností



Zdroj údajov: ŠÚ SR – RPO

Čo sa týka vývoja jednotlivých životných situácií, v období rokov 2018, 2019 a 2020 je zrejmý porovnateľný trend v prípade zániku živností. Výnimkou je vznik nových a pozastavenie existujúcich živností (graf č.1 a č. 2), čo ovplyvnilo aj vývoj oprávnení na podnikanie fyzických osôb – podnikateľov ako celku. Zmena oproti predchádzajúcim rokom bola zistená v 12. týždni 2020 (marec), odkedy sa tento vývoj v dôsledku vyhlásenia mimoriadnej situácie mení, najmä v prípade pozastavenia živností nadobudol výraznejší stúpajúci trend.

V súčasnej mimoriadnej situácii, ktorá má na viaceré špecifické oblasti podnikateľského prostredia stále nepredvídateľný dopad, existuje mnoho ďalších, dodnes nezodpovedaných otázok, ktoré má ŠÚ SR v zmysle svojich zákonných kompetencií oprávnenie zisťovať a verejnosť informovať o ich výsledkoch.

Kríza spôsobená pandémiou na jednej strane vyvolala obrovské množstvo problémov a prekážok pri plnení úloh štátnej štatistiky, ale na druhej strane urýchlila prijatie niektorých opatrení v oblasti metodológie, nových foriem práce a využívania externých zdrojov údajov.

ŠÚ SR bude i naďalej pokračovať vo svojom úsilí a inovatívnom prístupe k hľadaniu odpovedí na výzvy vyvolané krízou šírenia COVID-19.

Ing. Helena GLASER-OPITZOVÁ

Autorka je generálnou riaditeľkou sekcie všeobecnej metodiky a registrov Štatistického úradu SR.

PRIPRAVUJEME/COMING SOON

Boris VAŇO
DOPADY ZMIEN REPRODUKČNÉHO SPRÁVANIA NA POČET, PRÍRASTOK A VEKOVÉ ZLOŽENIE OBYVATEĽSTVA
THE IMPACT OF CHANGES IN THE REPRODUCTIVE BEHAVIOUR ON THE NUMBER, INCREASE AND AGE STRUCTURE OF THE POPULATION

Boris VAŇO, Ľudmila IVANČÍKOVÁ
HODNOTENIE KVALITY ADMINISTRATÍVNYCH ZDROJOV ÚDAJOV – TEORETICKO-METODOLOGICKÉ ASPEKTY.
QUALITY ASSESSMENT OF ADMINISTRATIVE DATA SOURCES - THEORETICAL AND METHODOLOGICAL ASPECTS

Milan TEREK
MOŽNOSTI RIEŠENIA PROBLÉMU NEODPOVEDANIA V ANALÝZACH DÁT Z DOTAZNÍKOVÝCH PRIESKUMOV PRI VYČERPÁVAJÚCOM SKÚMANÍ
POSSIBILITIES OF SOLVING THE PROBLEM OF NONRESPONSE IN ANALYSES OF DATA FROM QUESTIONNAIRE SURVEYS IN CENSUSES

Michal MAJTÁN
RODOVÁ ROVNOSŤ
GENDER EQUALITY

* * *

ONLINE VERZIA ČÍSLA 3/2020 SLOVENSKEJ ŠTATISTIKY A DEMOGRAFIE JE VEREJNE DOSTUPNÁ *na internetovej stránke ssad.statistics.sk od* ***15. JÚLA 2020****.*

THE ONLINE VERSION OF THE JOURNAL SLOVAK STATISTICS AND DEMOGRAPHY No 1 (2020) IS PUBLICLY BE AVAILABLE *at the website ssad.statistics.sk from* ***JULY 15, 2020.***

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 ssad.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 a elektronickej podobe na ssad.statistics.sk.

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 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 title.

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 ssad.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 explained 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 and electronic form at the website ssad.statistics.sk.

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 2020

Tlač:
Reprografické stredisko
Štatistického úradu SR

Predplatné:
20 € (na rok)
5 € (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 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 2020

Press:
Reprographic centre of the
Statistical Office of the SR

Subscription:
€20 (per year)
€5 (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