Slovenská štatistická a demografická spoločnosť



Výpočtová štatistika 2017

zborník abstraktov z 26. medzinárodného vedeckého seminára
Výpočtová štatistika 2017
a zborník príspevkov z konferencie
Prehliadka prác mladých štatistikov a demografov 2017
konaných v Bratislave v dňoch 7. – 8. decembra 2017

2017



Slovenská štatistická a demografická spoločnosť

Výpočtová štatistika 2017

zborník abstraktov z 26. medzinárodného vedeckého seminára

Výpočtová štatistika 2017

a zborník príspevkov z konferencie

Prehliadka prác mladých štatistikov a demografov 2017

konaných v Bratislave v dňoch 7. – 8. decembra 2017

Bratislava 2017

Výpočtová štatistika 2017

zborník abstraktov z 26. medzinárodného vedeckého seminára **Výpočtová štatistika 2017**
a zborník príspevkov z konferencie **Prehliadka prác mladých štatistikov a demografov 2017**
konaných v Bratislave v dňoch 7. – 8. decembra 2017

Editori: Iveta Stankovičová, Janka Medová

Vydala Slovenská štatistická a demografická spoločnosť, Bratislava, 2017.
Text neprešiel jazykovou úpravou. Za obsah príspevkov zodpovedajú jednotliví autori.

ISBN 978-80-88946-79-3

Obsah

Časť 1

Zborník abstraktov z medzinárodného seminára Výpočtová štatistika 2017

Mzdy upravené podle parity kupní síly ve vybraných evropských a mimoevropských zemích
Purchasing Power Parity – Adjusted Wages in Selected European and Non-European Countries
Diana Bílková 6

Zvyšuje rebalansovanie výkonnosť trackovacieho portfólia?
Does rebalancing boost performance of a tracking portfolio?
Martin Boďa, Mária Kanderová, Ľubomír Pinter 7

Změny složení obyvatelstva ČR podle rodinného stavu a vzdělání po roce 1989
Changes in the Population Structure of the Czech Republic by Marital Status and Education Attained after 1989
Tomáš Fiala, Jitka Langhamrová, Jana Vrabcová 8

Zhluková analýza na panelových dátach
Clustering analysis on panel data
Samuel Hudec 9

Komparácia krajín CEE a EÚ-15 v kontexte cyklickosti fiškálnej politiky – panelový VAR model
Comparison of the CEE and the EU-15 Countries in the Context of the Fiscal Policy's Cyclicality – Panel VAR Model
Martin Kameník, Ľubica Štiblárová 10

Vliv věkové struktury obyvatelstva na výskyt a léčbu chronických onemocnění v rámci Evropy
Influence of the age structure of the population on occurrence and treatment of chronic diseases within Europe
Lucie Kopecká 11

Demografický grid, jeho tvorba a použitie v geografických informačných systémoch
Demographic grid, its creation and use in geographic information systems
Dagmar Kusendová 12

Různé způsoby stanovení počtu shluků při aplikaci shlukové analýze v softwaru R
Different ways how to find optimal number of clusters using the cluster analysis with software R
Tomáš Löster, Jakub Danko 13

Porovnanie sociálnych zariadení na Slovensku
Comparison of social service facilities in the Slovak republic
Ľubica Sipková 14

Využitie modelov so zmiešanými efektami a metódy RE-EM stromu pri predikcii finančnej tiesne
Utilization of Mixed Effects Models and RE-EM Tree Method in Financial Distress Prediction
Lukáš Sobíšek, Karel Helman, Mária Stachová 15

Plnenie cieľov Stratégie Európa 2020 v sociálnej oblasti v SR a ČR – ukazovateľ AROPE
Achieving the Europe 2020 Strategy in the Social Field in the Slovakia and the Czech Republic - AROPE indicator
Iveta Stankovičová, Jitka Bartošová 16

Viacrozmerná parametrická a neparametrická analýza rozptylu a možnosti jej využitia v softvéri R
Multivariate parametric and non-parametric analysis of variance and its application in software R
Gábor Szűcs 17

Exekuce na důchodech seniorů
Execution of senior's retirement pensions
Jana Vrabcová 18

Časť 2
Zborník článkov z konferencie Prehliadka prác mladých štatistikov a demografov 2017

Analýza položkových dát
Itemset Mining
Michal Árvai 20

Aktuárske modelovanie vybraných poistných produktov dôchodkového sporenia
Actuarial modelling of selected insurance products of pension savings
Tatiana Duchoňová 30

Loggama rozdelenie v demografickej štatistike
Log-gamma distribution in demographic statistics
Daniel Hladký 39

Identifikace statistických distribucí v Preisach-Mayergoyzově prostoru hysteronů jako úloha optimalizace
Identifications of statistical distributions in Preisach-Mayergoyz hysteron space as optimization task
Colette Kožená 49

Modelovanie závislosti finančno-ekonomických ukazovateľov pomocou kopula funkcií
Dependence modelling of financial-economic indicators using copula functions
Nikoleta Lengyelová 58

Vzdelanie ako diferenčný faktor plodnosti na Slovensku
Educational as a differential factor of fertility in Slovakia
Michaela Lukáčová 67

Predikcia poradia v ankete Športovec roka 2016 lineárnou regresiou
The prediction of results in Sportsman of 2016 poll using linear regression
Michaela Mihoková 76

Časť 1

Výpočtová štatistika 2017

Zborník abstraktov

Mzdy upravené podle parity kupní síly ve vybraných evropských a mimoevropských zemích
Purchasing Power Parity – Adjusted Wages in Selected European and Non-European Countries

Diana Bílková
Vysoká škola ekonomická v Praze; Fakulta informatiky a statistiky
University of Economics, Prague; Faculty of Informatics and Statistics
diana.bilkova@vse.cz

Abstrakt: *Tento příspěvek se zaměřuje na situaci týkající se mezd upravených podle parity kupní síly ve vybraných evropských a mimoevropských zemích. Speciální pozornost je věnována genderovému mzdovému rozdílu a jeho vývoji v posledních letech. Při pohledu na mapu východní Evropy je zřejmé, že oblast poměrně nízkých mezd se stále rozkládá směrem od severu na jih. Začíná v Estonsku, pokračuje přes Lotyšsko, Litvu, Polsko, Česko, Slovensko, Maďarsko, Chorvatsko, Srbsko, Rumunsko, Bulharsko, Makedonii a končí v Turecku. Náklady na práci jsou vyšší ve všech zemích západně od Česka.*

Abstract: *The present paper focuses on the situation of purchasing power parity-adjusted wages in selected European and non-European countries. Special attention is paid to the gender wage gap and its development during last years. Looking at the map of Eastern Europe, the area of relatively low wages stretches from north to south. It begins in Estonia, continues through Latvia, Lithuania, Poland, Czechia, Slovakia, Hungary, Croatia, Serbia, Romania, Bulgaria, Macedonia and ends in Turkey. Labour costs are higher in all countries west of Czechia.*

Klíčová slova: *Mzdy upravené podle parity kupní síly, roční tempo růstu mezd, genderový mzdový rozdíl.*

Key words: *Purchasing power parity adjusted wages, annual wage growth rate, gender wage gap.*

Zvyšuje rebalansovanie výkonnosť trackovacieho portfólia? Does rebalancing boost performance of a tracking portfolio?

Martin Boďa, Mária Kanderová, Ľubomír Pinter
Ekonomická fakulta Univerzity Mateja Bela v Banskej Bystrici
Faculty of Economics of Matej Bel University in Banská Bystrica
martin.boda@umb.sk, maria.kanderova@umb.sk, lubomir.pinter@umb.sk

Abstrakt: *Príspevok sa venuje praktickým aspektom rebalansovania pri riadení výkonnosti trackovacieho portfólia. Príspevok sumarizuje existujúce prístupy k rebalansovaniu a diskutuje ich silné a slabé stránky, pričom rozširuje túto diskusiu na malú prípadovú štúdiu s cieľom zistiť, ktorý z týchto prístupov možno považovať za superiórny. Prípadová štúdia sa orientuje na trackovanie indexu S&P 500 a situáciu malého investora ochotného vytvoriť portfólio najviac 50 akcií v podmienkach existencie transakčných nákladov. Výsledky sú vyhodnocované za použitia štatistických metód. Tento výskum vznikol v rámci čiastočné plnenia projektu VEGA č. 1/0554/16.*

Abstract: *The presentation is devoted to practical aspects of rebalancing in managing the performance of a tracking portfolio. The presentation summarizes the existing approaches to rebalancing and discusses its strengths and weaknesses and extends this discussion to a small case study in order to find out which of this approaches may be thought of as superior. The case study is oriented upon tracking the S&P 500 Index and a situation of a small investor who is willing to create a portfolio of 50 equities at most and faces transaction costs. The results are evaluated by means of statistical method. This research arose in partial fulfilment of the project VEGA # 1/0554/16.*

Kľúčové slová: *rebalansovacia stratégia, trackovanie, index S&P 500, výkonnosť*
Key words: *rebalancing strategy, tracking, S&P 500 Index, performance*

Změny složení obyvatelstva ČR podle rodinného stavu a vzdělání po roce 1989
Changes in the Population Structure of the Czech Republic by Marital Status and Education Attained after 1989

Tomáš Fiala, Jitka Langhamrová, Jana Vrabcová
Vysoká škola ekonomická v Praze, Fakulta Informatiky a statistiky, Česká republika
University of Economics, Prague; Faculty of Informatics and Statistics, Czech Republic
fiala@vse.cz; langhamj@vse.cz; vrabcova.jana@post.cz

Abstrakt: *Změna politického režimu v ČR v roce 1989 se výrazně projevila nejen v oblasti ekonomické a sociální, ale došlo i ke změnám demografického chování obyvatelstva. Zvýšil se růst střední délky života, poměrně rychle klesala plodnost, narůstala zahraniční migrace. Současně s odkládáním plodnosti žen do vyššího věku a poklesem počtu dětí v rodině docházelo k obdobným změnám i ve sňatečnosti. Zvyšoval se průměrný věk snoubenců, přibývalo nesezdaných soužití. Na druhé straně rostla úroveň vzdělanosti, zejména mladých lidí. Zvýšil se počet středních i vysokých škol, rostl počet přijímaných studentů.*
Příspěvek zachycuje změny složení obyvatelstva ČR podle pohlaví a věku v kombinaci s rodinným stavem a úrovní nejvyššího dokončeného vzdělání. Vychází z porovnání výsledků posledních tří sčítání lidu, domů a bytů v ČR v letech 1991, 2001 a 2011.
Abstract: *The change of political regime in the Czech Republic in 1989 had a very considerable impact not only in the economic and social sphere but also important changes in demographic behaviour occurred. The rise in life expectancy accelerated, fertility fell down relatively rapidly, international migration increased. Females postponed births to higher age, number of children in the family diminished and similar changes were apparent in marriage behaviour. Average age of groom as well as of bride raised and proportion of unmarried couples increased. On the other hand the education level grew especially among young people. The number of secondary and tertiary schools increased the number of students raised.*
The paper reflects changes in the population structure of the Czech Republic by sex and age in combination with marital status and education level attained. It brings comparison of the results of the last three censuses in the Czech Republic in 1991, 2001 and 2011.
Klíčová slova: *složení obyvatelstva, rodinný stav, dosažené vzdělání, sčítání lidu.*
Key words: *Population structure, marital status, education attained, census.*

Poděkování: Příspěvek vznikl za podpory GA ČR 15-13283S *Projekce populace České republiky podle úrovně vzdělání a rodinného stavu.*

Zhluková analýza na panelových dátach
Clustering analysis on panel data

Samuel Hudec
Fakulta prírodných vied, Univerzity Mateja Bela
Faculty of Natural Sciences, Matej Bel University
samuel.hudec@umb.sk

Abstrakt: *Príspevok uvádza jeden zo spôsobov použitia známej zhlukovej analýzy na panelových dátach za použitia základných vedomostí o analýze mnohorozmerných dát. Náš postup je prezentovaný na makroekonomických dátach. Detailnejšie sa venuje zobrazeniu, obhliadke a príprave dát pred samotným zhlukovaním, následne rieši otázku voľby metódy zhlukovania k danému problému. Kľúčovým krokom príspevku je zapasovanie informácie o čase do postupu zhlukovania na panelových dátach. Spracovanie, analýza a grafické znázornenie dát bolo uskutočnené v štatistickom prostredí R.*

Abstract: *This paper shows one of the way of using well-known clustering analysis on panel data with respect to basic knowledge of multidimensional data analysis. Our approach "Cook book" is presented on a macroeconomic dataset. In details, we present visualization, viewing and transformation of the data before the clustering analysis, then we present our solution of selecting the clustering method. Inclusion of time information into the analysis is a key step of our paper. The computation for this paper was undertaken in statistical environment R.*

Kľúčové slová: *Panelové dáta, analýza hlavných komponentov, zhluková analýza, robustná metóda, krajiny Európskej únie.*

Key words: *Panel data, principal component analysis, clustering analysis, robust method, countries of the European Union.*

Komparácia krajín CEE a EÚ-15 v kontexte cyklickosti fiškálnej politiky – panelový VAR model
Comparison of the CEE and the EU-15 Countries in the Context of the Fiscal Policy's Cyclicality – Panel VAR Model

Martin Kameník, Ľubica Štiblárová
Technická univerzita v Košiciach, Ekonomická fakulta, Němcovej 32, 040 01 Košice, Slovensko
Technical University of Košice, Faculty of Economics, Němcovej 32, 040 01 Košice, Slovakia
martin.kamenik@tuke.sk, lubica.stiblarova@tuke.sk

Abstrakt: *Cieľom tohto príspevku je preskúmať cyklické správanie fiškálnej politiky v rámci krajín EÚ-27 a identifikovať možné rozdiely v cyklickosti fiškálnej politiky medzi krajinami CEE a EÚ-15 s využitím panelového VAR modelu. Výsledky východiskového modelu naznačujú pro-cyklický charakter fiškálnej politiky; čiastkové modely krajín CEE a EÚ-15 potvrdzujú predchádzajúce výsledky, avšak výrazné rozdiely môžu byť identifikované v rámci výdavkových multiplikátorov. Okamžitý výdavkový multiplikátor pre krajiny EÚ-15 je viac než dvojnásobne vyšší ako pre CEE krajiny a tieto rozdiely pretrvávajú dokonca aj rok po šoku.*
Abstract: *The aim of this paper is to examine cyclical behavior of the fiscal policy within the EU-27 countries and to identify possible differences in the fiscal policy's cyclicality among the CEE and the EU-15 countries using panel VAR model. Results of the default model suggest pro-cyclical character of the fiscal policy; partial models of the CEE and EU-15 countries confirm previous results, although remarkable differences can be identified concerning spending multipliers. Impact multiplier for the EU-15 countries is more than twice as high as for the CEE countries and these differences remain even a year after the shock.*
Kľúčové slová: *Fiškálna politika, cyklickosť, EÚ-15, krajiny CEE, panelový VAR model.*
Key words: *Fiscal policy, cyclicality, EU-15, CEE countries, panel VAR model.*

Vliv věkové struktury obyvatelstva na výskyt a léčbu chronických onemocnění v rámci Evropy
Influence of the age structure of the population on occurrence and treatment of chronic diseases within Europe

Lucie Kopecká
Univerzita Pardubice, Fakulta ekonomicko-správní, Ústav matematiky a kvantitativních metod
University of Pardubice, Faculty of Economics and Administration, Institute of Mathematics and Quantitative Methods
Lucie.Kopecka1@student.upce.cz

Abstrakt: *Chronická onemocnění postihují evropskou populaci skrze všechny generace stále častěji. Tato onemocnění se vyznačují dlouhodobou léčbou, nedají se zcela vyléčit a zanechávají následky na zbytek života. Dlouhodobá léčba je nákladná a tím ovlivňuje výdaje do zdravotnictví. Cílem článku je odhalit vliv věkové struktury obyvatelstva na výskyt a léčbu chronických onemocnění ve vybraných zemích Evropy pomocí vícerozměrných statistických metod za využití programů EXCEL, STATISTICA a R. Data byla získána z databází Eurostat, OECD a UNITED NATIONS.*

Abstract: *Chronic diseases affect European population through all generations more and more. These diseases are characterized by long-term treatment, they cannot be completely cured and they leave consequences for the rest of life. Long-term treatment is expensive and influences health expenditure. The main aim of this article is to reveal influence of the age structure of the population on incidences and treatment of chronic diseases in the selected European countries by multivariate statistical methods and EXCEL, STATISTICA and R programs. Data were obtained from databases of Eurostat, OECD and UNITED NATIONS.*

Klíčová slova: *Věková struktura obyvatelstva, chronická onemocnění, výdaje na zdravotnictví, vícerozměrné statistické metody.*

Key words: *Age structure of population, chronic diseases, health expenditure, multivariate statistical methods.*

Demografický grid, jeho tvorba a použitie v geografických informačných systémoch
Demographic grid, its creation and use in geographic information systems

Dagmar Kusendová
Katedra humánnej geografie a demografie, Prírodovedecká fakulta, Univerzita Komenského v Bratislave
Department of Human Geography and Demography, Faculty of Natural Sciences, Comenius University in Bratislava
dagmar.kusendova@uniba.sk

Abstrakt: *Príspevok objasňuje základné pojmy a postupy spojené s vizualizáciou rozmiestnenia a charakteristík obyvateľstva na báze pravidelných priestorových sieťových štruktúr (gridov) v programovom prostredí geografických informačných systémov. Grid predstavuje univerzálnu a relatívne novú formu kartografickej prezentácie výsledkov sčítaní, ale aj ďalších demografických štatistík. Ide v princípe o rastrovú prezentáciu geograficky lokalizovaných dát, ktorá efektívne rieši problematiku heterogénnosti a časových zmien hraníc územných štatistických jednotiek v demografickej štatistike. V závere prezentujeme stručnú históriu vzniku a organizáciu tvorby „demografických gridov“ v medzinárodnom a národnom kontexte a niektoré možnosti ich budúceho vývoja a využitia v praxi a výskume.*

Abstract: *The paper addresses the basic concepts and procedures concerned with the visualization of distribution and further characteristics of population based on regular spatial network structures (grids) in the environment of GIS software. Grid is a universal and relatively new form of cartographic presentation of the censuses results and further demographic statistics. Basically, it's a raster presentation of geographically localized data that effectively solves issues of heterogeneity and temporal changes in the boundaries of territorial statistic units in demographic statistics. In the end we briefly view the history of the creation and organization of the "demographic" grids production in the international and national context and some perspectives of their future development and use in practice and research.*

Kľúčové slová: *grid, populačný raster, geografické informačné systémy, vizualizácia, demografické štatistiky*

Key words: *grid, population raster, geographic information systems, visualization, demographic statistics*

Různé způsoby stanovení počtu shluků při aplikaci shlukové analýze v softwaru R
Different ways how to find optimal number of clusters using the cluster analysis with software R

Tomáš Löster[a], Jakub Danko[b]
[a] Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky, Česká republika
[a] University of Economics, Prague; Faculty of Informatics and Statistics, Czech Republic
[b] Technická univerzita Košice, Ekonomická fakulta, Slovenská republika
[b] Technical University of Košice, Faculty of Economics, Slovak Republic
tomas.loster@vse.cz, jakub.danko@tuke.sk

Abstrakt: *Současná odborná literatura nabízí velké množství koeficientů a postupů, jak stanovit optimální počet shluků, do kterých mají být jednotlivé objekty při aplikaci shlukové analýzy klasifikovány. Na současném trhu existuje celá řada softwarových produktů, které lze pro shlukovou analýzu využít. Pouze omezená část z nich však umožňuje využít koeficienty pro stanovení počtu shluků. V některých softwarech jsou koeficienty implementovány pouze v podobě grafických výstupů, v jiných v podobě konkrétních hodnot. Společné bývá, že jsou implementovány pouze vybrané koeficienty. Jako nejvíce propracovaný z hlediska stanovení počtu shluků se jeví R-software, v němž je v rámci balíčků aplikováno velké množství těchto koeficientů. Cílem tohoto příspěvku je ukázat vybrané možnosti, jak využít R-software ke stanovení počtu shluků ve shlukové analýze.*

Abstract: *Current literature offers a large number of coefficients and procedures how to determine the optimal number of clusters into which individual objects are classified when we use cluster analysis. There are a number of software products in the current software market that can be used for clustering. However, only a limited number of them allow using coefficients for determining the number of clusters. In some software, the coefficients are implemented only in the form of graphical outputs, in others in the form of specific values. It is common for all software products that only selected coefficients are implemented. The most sophisticated regarding finding of a number of clusters is R-software, in which is implemented a large number of these coefficients in many packages. This paper aims to show selected options how to determine the number of clusters in cluster analysis using R-software.*

Kľúčové slová: *Shluková analýza, koeficienty, optimální počet shluků, software, R-software*
Keywords: *Cluster analysis, coefficients, optimal number of clusters, software, R-software*

Porovnanie sociálnych zariadení na Slovensku
Comparison of social service facilities in the Slovak republic

Ľubica Sipková
Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Bratislava, Slovensko
The University of Economics in Bratislava, Faculty of Economic Informatics, Bratislava, Slovakia
lubica.sipkova@euba.sk

Abstrakt: *V príspevku regionálne a štrukturálne porovnávame počty zariadení sociálnych služieb a zariadení na vykonávanie opatrení sociálnoprávnej ochrany a sociálnej kurately v Slovenskej republike. Rôznorodá deskriptívna grafická a numerická analýza zastúpenia zariadení podľa ôsmich krajov Slovenska vzhľadom na typ vlastníctva zariadenia, počet osôb v starostlivosti zariadenia a ich veku vychádza z individuálnych údajov oficiálneho vyčerpávajúceho pravidelného ročného štatistického zisťovania, s označením SOC 1-01, ktoré poskytol Štatistický úrad SR na vedecké účely. Prezentované sú výsledky a opísané aj niektoré charakteristické črty a významné rozdiely vo finančnej situácii v štruktúre týchto zariadení, hlavne čo sa týka príjmových a výdavkových rozdelení za rok 2015.*

Abstract: *By the regional and structural comparison, in the article we describe the structure of the facilities providing social services and facilities for the implementation of social protection in the Slovak republic. Specification of deployment of facilities by the results of various descriptive graphical and numerical analysis according to the eight regions of Slovakia, with regard to the type of ownership of the facility, the number of persons in the care of the facility and their age is based on the individual data of the official regular annual statistical survey SOC 1-01. The survey data was provided by the Statistical Office of the Slovak republic for research. Some features and noteworthy differences in the financial situation are described, particularly in terms of income and expenditure distributions in 2015.*

Kľúčové slová: *Sociálne služby, finančná situácia, príjmy, výdavky, sociálne zariadenia, NUTS regióny SR*

Key words: *Social services, financial situation, incomes, expenditures, social facilities, NUTS regions SR*

Poďakovanie: Článok bol vypracovaný za podpory grantovej agentúry VEGA č. 1/0092/15*: Moderné prístupy k navrhovaniu komplexných štatistických prieskumov* a VEGA č. 1/0770/17*: Dostupnosť bývania na Slovensku.* / This article was developed with support of the grant agency VEGA 1/0092/15 and VEGA 1/0770/17.

Využitie modelov so zmiešanými efektami a metódy RE-EM stromu pri predikcii finančnej tiesne
Utilization of Mixed Effects Models and RE-EM Tree Method in Financial Distress Prediction

Lukáš Sobíšek[a], Karel Helman[a], Mária Stachová[b]
[a] Vysoká škola ekonomická v Prahe, nám. W. Churchilla 4, 130 67 Praha 3, ČR
[a] University of Economics in Prague, W. Churchill sqr. 4, 130 67 Prague 3, Czech Republic
[b] Univerzita Mateja Bela, Tajovského 10, 974 90 Banská Bystrica, SR
[b] Matej Bel University, Tajovského 10, 974 90 Banská Bystrica, Slovakia
lukas.sobisek@vse.cz, helmank@vse.cz, maria.stachova@umb.sk

Abstrakt: *Existuje viacero prístupov k finančným ex-ante analýzam. Niektoré sú založené na statických klasifikačných modeloch, napríklad logistickej regresii, klasifikačných stromoch alebo diskriminačnej analýze. Tieto analýzy odhaľujú súvislosti medzi medzi indikátormi, resp. ukazovateľmi finančného zdravia podnikov. Finančné ukazovatele však väčšinou bývajú zaznamenávané počas niekoľkých, po sebe nasledujúcich rokov a tým vytvárajú dátovú množinu, ktorá má charakter panelových dát. V našom príspevku sa zameriavame na regresnú analýzu aplikovanú na panelové dáta, ktorá nám pomôže predikovať finančné zdravie podnikov a zároveň zahrnúť časovú dynamiku dát. Konkrétne sa zameriavame na regresný model so zmiešanými efektami, pomocou ktorého sa snažíme zvýšiť predikčnú presnosť RE-EM klasifikátora.*

Abstract: *There are several approaches to financial analysis ex-ante. Some of them are based on static classification models, e.g. logistic regression, classification trees or discriminant analysis. These analyses are carried out to study a relationship between financial health indicators of companies. These financial indicators are often collected over several consecutive years, thus the data have the form of panel data. In our paper, we focus on regression analysis of panel data to predict financial health of companies. We fit generalized mixed effects model in order to improve the prediction ability the RE-EM classifier.*

Kľúčové slová: *Zovšeobecnený regresný model so zmiešanými efektami, finančné zdravie podnikov, RE-EM model*

Key words: *Generalized regression mixed effect model, Financial health of companies, RE-EM tree model*

Plnenie cieľov Stratégie Európa 2020 v sociálnej oblasti v SR a ČR – ukazovateľ AROPE

Achieving the Europe 2020 Strategy in the Social Field in the Slovakia and the Czech Republic - AROPE indicator

Iveta Stankovičová[a], Jitka Bartošová[b]
[a] Fakulta managementu, Univerzita Komenského, Bratislava, SR
[a] Faculty of Management, Comenius University, Bratislava, Slovakia
[b] Vysoká škola ekonomická v Prahe, Fakulta managementu, Jindřichův Hradec, ČR
[b] University of Economics Prague, Faculty of Management, Jindřichův Hradec, Czech Republic
iveta.stankovicova@fm.uniba.sk; bartosov@fm.vse.cz

Abstrakt: *Viac ako 120 miliónov ľudí je ohrozených chudobou alebo sociálnym vylúčením v krajinách EÚ. Vedúci predstavitelia EÚ sa preto zaviazali, že do roku 2020 vymania z chudoby a sociálneho vylúčenia najmenej 20 miliónov ľudí. Boj proti chudobe a sociálnemu vylúčeniu je jadrom stratégie Európa 2020 pre inteligentný, udržateľný a inkluzívny rast. Každý členský štát prijal jeden alebo viacero národných cieľov. Článok skúma agregovaný ukazovateľ chudoby a sociálneho vylúčenia AROPE v SR a ČR. Ukazovateľ AROPE je súčet osôb, ktoré sú vystavené riziku chudoby alebo sú výrazne znevýhodnené alebo žijú v domácnostiach s veľmi nízkou intenzitou práce ako podiel na celkovom počte obyvateľov. Zdrojom výpočtu ukazovateľa je harmonizované štatistické zisťovanie EÚ SILC. Zameriavame sa na rozdelenie chudoby a sociálneho vylúčenia v regiónoch SR a ČR. Opisujeme súčasné trendy pre agregovaný ukazovateľ na Slovensku a v Českej republike a porovnávame naše hodnoty a trendy s ostatnými krajinami EÚ. Práca využíva vybrané štatistické a matematické modely a postupy.*

Abstract: *More than 120 million people are at risk of poverty or social exclusion in the EU. EU leaders have pledged to bring at least 20 million people out of poverty and social exclusion by 2020. The fight against poverty and social exclusion is at the heart of the Europe 2020 strategy for smart, sustainable and inclusive growth. Each individual member state will have to adopt one or several national targets. Presented article examines the aggregate indicator of poverty and social exclusion AROPE in the Slovak and Czech Republic. Indicator AROPE is the sum of persons who are at-risk-of-poverty or severely materially deprived or living in households with very low work intensity as a share of the total population. Source for calculating of this indicator is harmonized EU SILC statistical survey. We focus on distribution of poverty and social exclusion in the regions of Slovakia and Czech Republic. We describe current trends for aggregate indicator in Slovakia and Czech Republic and compare our values and trends with others EU countries. The paper uses selected statistical and mathematical models and procedures.*

Kľúčové slová: *Stratégia Európa 2020, riziko chudoby, materiálna deprivácia, nízka pracovná intenzita, región, databáza EU SILC*

Key words: *Europe 2020 Strategy, At-Risk-of-Poverty, Material Deprivation, Low Work Intensity, Region, EU SILC database.*

Viacrozmerná parametrická a neparametrická analýza rozptylu a možnosti jej využitia v softvéri R
Multivariate parametric and non-parametric analysis of variance and its application in software R

Gábor Szűcs
Fakulta matematiky, fyziky a informatiky Univerzity Komenského v Bratislave
Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava
Gabor.Szucs@fmph.uniba.sk

Abstrakt: *Analýza rozptylu (skratka ANOVA, ANalysis Of VAriance) je parametrickou štatistickou metódou, ktorá bola navrhnutá predovšetkým na posúdenie podobností a odlišností viacerých súborov, pričom prostredníctvom nej možno testovať efekt jedného alebo viacerých faktorov. Ak sa na pozorovaných objektoch súčasne sledujú viaceré štatistické znaky, odporúča sa použiť viacrozmernú analýzu rozptylu (Multivariate ANalysis Of VAriance, MANOVA), ktorá ponúka komplexnejšie možnosti testovania odlišností medzi súbormi alebo efektmi. Najväčšou nevýhodou modelov ANOVA a MANOVA je ich závislosť na predpoklade normálneho rozdelenia vysvetľovanej premennej. V praktických aplikáciách sa tento predpoklad často poruší, preto je potrebné hľadať iné alternatívy testovania. Riešenie ponúkajú neparametrické ekvivalenty analýzy rozptylu, napríklad jednorozmerný či viacrozmerný Kruskalov-Wallisov test. Cieľom tohto príspevku je porovnať medzi sebou spomínané parametrické a neparametrické modely v prípade viacrozmerných pozorovaní a ukázať možnosti ich aplikácie v prostredí softvéru R.*

Abstract: *Analysis of variance (also known as ANOVA, ANalysis Of VAriance) is a parametric statistical method, which was primarily designed to assess similarities and differences between more groups and to test the effect of one or more factors. In case of multiple dependent statistical variables, it is recommended to use the multivariate analysis of variance (MANOVA) which offers more sophisticated testing options to find differences between two or more groups. The main limitation of the ANOVA and MANOVA models is their dependence on the assumption of normality of the dependent variable. In practical applications, this assumption is often violated, so it is necessary to seek other alternatives for testing similarities and differences between treatments. One possibility is to apply non-parametric equivalent of analysis of variance, the one-dimensional or multivariate Kruskal-Wallis test. The aim of this paper is to compare aforementioned parametric and non-parametric models in case of multivariate observations and to present their applications in software R.*

Kľúčové slová: *viacrozmerná analýza rozptylu, MANOVA, viacrozmerné neparametrické testy, Kruskalov-Wallisov test.*

Key words: *multivariate analysis of variance, MANOVA, multivariate non-parametric tests, Kruskal-Wallis test.*

Poďakovanie: Tento výskum bol podporený grantmi VEGA 2/0047/15, VEGA 1/0251/16 a APVV-0465-12.

Exekuce na důchodech seniorů
Execution of senior's retirement pensions

Jana Vrabcová
Výzkumný ústav práce a sociálních věcí, Praha, Česká republika
Research Institute for Labour and Social Affairs, Prague, Czech Republic
janalanghamrova@seznam.cz

Abstrakt: *V posledních letech roste zadlužení celé populace České republiky, které se nevyhýbá ani seniorům. Senioři jsou specifickou skupinou, která je také ohrožena exekucemi. Mnoho z nich má se svými dluhy problém a dostává se do fáze předlužení, kdy již nejsou schopni hradit své závazky. V tom případě nastupuje instituce exekuce, která je uvalena na starobní důchod. Podle dat České správy sociálního zabezpečení se počet poživatelů důchodu v České republice od roku 2004 do roku 2014 více než ztrojnásobil z téměř 24 tisíc poživatelů důchodu s nařízenou exekucí v roce 2004 na více než 75 tisíc poživatelů důchodu s nařízenou exekucí v roce 2014. Na Slovensku eviduje a také provádí exekuce na důchodech Sociální pojišťovna na základě příkazu na vykonání srážky od soudního exekutora, soudu nebo daňového úřadu. Důvody pro exekuci starobních důchodů jsou obdobné jako v případě České republiky. Může se jednat o přenesení dluhu z aktivního pracovního života, problémy s placením výživného, daní atp.*

Abstract: *In recent years, the indebtedness of the entire population of the Czech Republic has risen, which is not avoided even by seniors. Seniors are a specific group that is also threatened with execution. Many of them have a problem with their debts and get into the over-indebtedness stage, when they are no longer able to pay their debts. In that case, the begins execution of the retirement pension. According to the data of the Czech Social Security Administration, the number of pensioners in the Czech Republic has more than tripled from 2004 to 2014 from nearly 24 thousand pensioners in 2004 to more than 75 thousand in 2014. In Slovakia the Social Insurance Company records and also executes retirement executions. The reasons for the execution of retirement pensions are similar to those of the Czech Republic. It can be a transfer of debt from active working life, problems with maintenance payments, taxes, etc.*

Klíčové slova: *exekuce, starobní důchod, senioři*
Key words: *execution, retirement pension, seniors*

Časť 2

Výpočtová štatistika 2017
Prehliadka prác mladých štatistikov a demografov

Zborník článkov

Analýza položkových dát
Itemset Mining

Michal Árvai
Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, magisterské štúdium, 1. ročník
Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics
michalarvai@gmail.com

Abstrakt: *Analýza položkových dát je jedným z mnohých nástrojov hĺbkovej analýzy dát, ktorá slúži na objavenie asociačných pravidiel v dátových súboroch. Prototypickým problémom analýzy položkových dát je tzv. problém nákupných košov, ktorý sa zaoberá hľadaním kombinácií položiek, ktoré boli často zakúpené spoločne, t.j. boli často súčasťou transakcií. V tomto článku uvažujeme známy pravdepodobnostný model na simulovanie transakcií pomocou nezávislých Bernoulliho pokusov a zároveň predstavíme vlastnú, alternatívnu metódu simulovania pomocou Markovových reťazcov. Funkcionalita vybranných algoritmov je ilustrovaná na simulovaných dátach, ktoré sú analyzované v štatistickom softvéri R.*

Abstract: *The presented paper discusses the problem of itemset mining which is one of the main data mining tasks. The framework of the so-called market-basket problem is introduced which is the prototypical application of itemset mining. It aims to discover interesting association rules in transaction data to provide insights into customer behaviour. A known probabilistic framework is presented which enables transaction data simulation based on independent Bernoulli trials. Building on this framework, a new, Markov chains based, simulation is proposed. The paper concludes with an analysis of simulated transactions in the statistical software R.*

Kľúčové slová: *analýza položkových dát, hľbková analýza dát, asociačné pravidlá, nezávislé Bernoulliho pokusy, Markovove reťazce.*

Key words: *itemset mining, data mining, association rules, independent Bernoulli trials, Markov chains.*

1 Úvod

Prototypická a jedna z najznámejších aplikácií analýzy položkových dát je analýza nákupných košov (angl. *market basket analysis*) (Agrawal, Imielinski, Swami, 1993). Cieľom tejto analýzy je identifikovať položky, ktoré sú často zakúpené spoločne, resp. nájsť pravidelnosti v nakupovaní zákazníkov. Pomocou algoritmov sa pokúšame nájsť také pravidlá, ktoré nám predpovedajú výskyt položiek podľa toho, ktoré položky sa už nachádzajú v transakcii.

1.1 Základné pojmy a označenia

Nech $I = \{i_1, i_2, \dots, i_m\}$ je množina položiek; táto množina vyjadruje napr. množinu všetkých možných položiek, ktoré môžeme zakúpiť v danom obchode. Množina $X \subseteq I$ sa nazýva položková množina.

Nech $\boldsymbol{T} = \{t_1, t_2, \dots, t_m\}$ je množina prvkov, ktoré nazývame identifikátory transakcií; táto množina vyjadruje napr. množinu všetkých zákazníkov daného obchodu. Množina $T \subseteq \boldsymbol{T}$ sa nazýva množina identifikátorov transakcií.

Transakciou nazývame dvojicu v tvare $\langle t, X \rangle$, kde $t \in \boldsymbol{T}$ je individuálny identifikátor transakcie, a $X \subseteq I$ je položková množina. Ľubovoľnú množinu dvojíc $\langle t, X \rangle$ nazývame databázou transakcií, ozn. D.

Asociačné pravidlo je taká usporiadaná dvojica (X, Y), ktorú je zvykom značiť pomocou symbolu implikácie $X \Rightarrow Y$, kde $X, Y \subseteq I$, a položkové množiny X, Y sú disjunktné, t.j. $X \cap Y = \emptyset$. Pomocou asociačných pravidiel hľadáme pre ktoré položkové množiny X, Y platí, že ak zákazník zakúpil X, tak zakúpil aj Y.

1.2 Miery zaujímavosti

Keďže chceme analyzovať len zaujímavé asociačné pravidlá musíme uvažovať rôzne miery zaujímavosti. Najzaužívanejšie sú elementárne miery početnosť (angl. *support*) a spoľahlivosť (angl. *confidence*). Početnosť vyjadruje počet transakcií ktoré obsahujú obe strany pravidla, teda je vyjadrená v tvare

$$\sup(X \Rightarrow Y) = \sup(X \cup Y) = \sup(XY) = |\boldsymbol{t}(XY)|, \qquad (1)$$

kde $\boldsymbol{t}(X) = \{t \in \boldsymbol{T} \mid \text{transakcia } t \text{ obsahuje } X\}$. Ak je početnosť vyjadrená v tvare zlomku, tak sa jedná o relatívnu početnosť, ktorá korešponduje s odhadom pravdepodobnosti javu, že transakcia bude obsahovať danú položkovú množinu, ozn. $\hat{P}(\boldsymbol{T}_X)$. Spoľahlivosť asociačného pravidla je definovaná pomocou podmienenej pravdepodobnosti javu, že transakcia obsahuje Y za podmienky, že obsahuje X. Spoľahlivosť je definovaná v tvare

$$conf(X \Rightarrow Y) = \hat{P}(\boldsymbol{T}_Y | \boldsymbol{T}_X) = \frac{\hat{P}(\boldsymbol{T}_X \cap \boldsymbol{T}_Y)}{\hat{P}(\boldsymbol{T}_X)}. \qquad (2)$$

Ďalšia miera zaujímavosti je `lift`[1], ktorá meria závislosť medzi jednotlivými položkovými množinami, definovaná v tvare

$$lift(X \Rightarrow Y) = \frac{\hat{P}(\boldsymbol{T}_Y | \boldsymbol{T}_X)}{\hat{P}(\boldsymbol{T}_Y)} = \frac{\hat{P}(\boldsymbol{T}_X \cap \boldsymbol{T}_Y)}{\hat{P}(\boldsymbol{T}_X)\hat{P}(\boldsymbol{T}_Y)}. \qquad (3)$$

Ak sa $lift = 1$, tak sú dané položkové množiny nezávislé, pričom $lift > 1$ poukazuje na mieru závislosti daných položkových množín.

1.3 Postup analýzy položkových dát a pravidiel

Hlavná myšlienka analýzy položkových dát je zúžiť dátový súbor len na množiny položiek, ktoré sa často objavili spoločne v transakciách, čiže tie, ktoré spĺňajú kritérium minimálnej početnosti dané parametrom `minsup`. V druhej časti analýzy sa z množiny častých položkových množín pokúšame nájsť silné asociačné pravidlá, čiže tie, ktoré spĺňajú kritérium parametra minimálnej spoľahlivosti, ozn. `minconf`.

[1] Miera `lift` bola zavedená s názvom záujem (angl. *interest*) (Brin, Motwani, Ullmann, Tsur, 1997).

2 Generovanie syntetických položkových dát

Syntetické dáta nám slúžia na ohodnotenie efektivity a na porovnanie implementovaných algoritmov na analýzu položkových dát, resp. na zavedenie a následné testovanie nových mier zaujímavosti. Cieľom simulácie je navrhnúť vhodný pravdepodobnostný model na získanie databázy transakcií D, ktorá sa dá v našej interpretácií chápať ako súbor transakcií, ktorý vhodne (spĺňajúc naše úvodné požiadavky) opisuje správanie sa zákazníkov.

2.1 Simulovanie transakcií pomocou nezávislých Bernoulliho pokusov

Model simulovania transakcií pomocou nezávislých Bernoulliho pokusov bol predstavený v článku (Hahsler, Hornik, Reutterer, 2005). Pomocou tejto metódy simulujeme transakcie bez prítomnosti asociácií s nezávisle vyskytujúcimi sa položkami. Transakcie sledujeme na pevnom časovom intervale $[0, h]$ s náhodnou realizáciou, pričom postupnosť okamihov transakcií nastáva v súlade s Poissonovým procesom s parametrom θ, t.j. $|\boldsymbol{T}| = M \sim Po(\theta h)$. Predpokladáme, že položky sú zakúpené nezávisle od seba a pre každú položku $i_k \in I$, pre $k = 1, \dots, n$, existuje fixná pravdepodobnosť p_k výskytu v transakcii. Každá transakcia je teda výsledok z n nezávislých Bernoulliho pokusov, zvlášť pre každú položku s pravdepodobnosťami danými vektorom $p = (p_1, \dots, p_n)$.

	i_1	i_2	$\dots$	i_n
p	p_1	p_2	$\dots$	p_n
t_1	1	0	$\dots$	1
t_2	0	1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_m	1	0	$\dots$	0
c	c_1	c_2	$\dots$	c_n

Obr. 1 Binárna reprezentácia dát s pridanými riadkami pre vektor pravdepodobnosti $\boldsymbol{p}$ a pre vektor súčtov jednotlivých stĺpcov $\boldsymbol{c}$, t.j. počty jednotlivých položiek v databáze tranaskcií. *(Zdroj:* (Hahsler, Hornik, Reutterer, 2005), *vlastné spracovanie)*

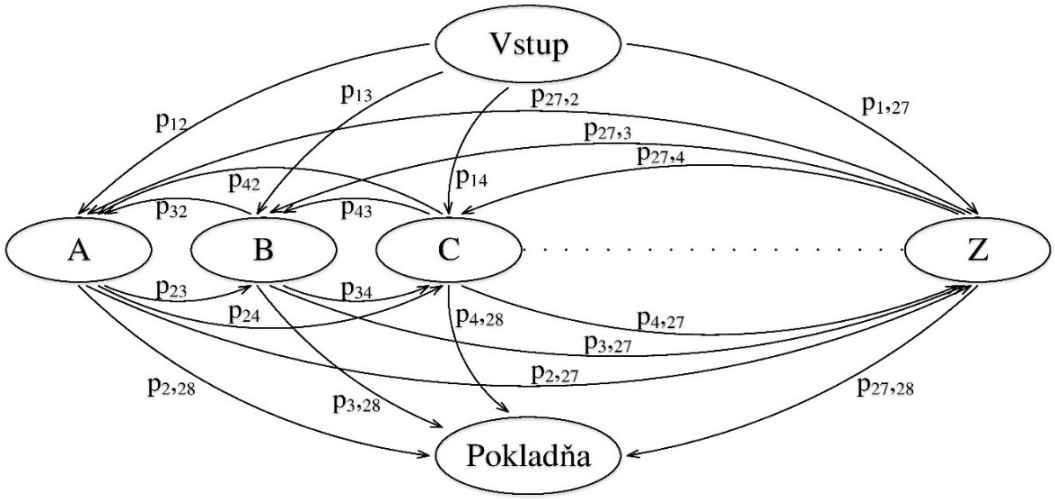
Hodnoty jednotlivých počtov položiek c_i z obr. 1 interpretujeme ako realizáciu náhodnej premennej C_i. Jednotlivé pravdepodobnosti výskytu, využijúc vlastnosti Poissonovho procesu, vieme vyjadiť v tvare

$$P(C_i = c_i) = \frac{e^{(-p_i\theta h)}(p_i\theta h)^{c_i}}{c_i!} \sim Po(p_i\theta h) = Po(\lambda_i). \tag{4}$$

Hlavná myšlienka modelu je odhadnúť neznáme parametre Poissonovho rozdelenia a následne simulovať transakcie s využitím odhadnutých pravdepodobností. Podľa vzoru z článku (Hahsler, Hornik, Reutterer, 2005) na odhad parametra $\lambda = (\lambda_1, \dots, \lambda_n)$ použijeme Gamma rozdelenie.

2.2 Simulovanie transakcií pomocou Markovových reťazcov

V tomto modeli predstavujeme alternatívnu metódu simulovania transakcií pomocou homogénnych Markovových reťazcov s diskrétnym časom. Pokúšame sa simulovať správanie sa zákazníkov navrhnutím pravdepodobností prechodov na danom priestore stavov. Pod pojmom nachádzania sa v stave (okrem začiatočného a koncového stavu) rozumieme pridanie položky k sledovanej transakcii. Kvôli jednoduchosti uvažujme obchod s $|I| = 26$ položkami, pričom ku každej položke priradíme jedno písmeno anglickej abecedy, $I = \{\mathrm{A}, \mathrm{B}, \dots, \mathrm{Z}\}$.



Obr. 2 Markovov reťazec v podobe orientovaného grafu, v ktorom každý vrchol predstavuje stav reťazca a orientované hrany znázorňujú možné prechody medzi stavmi s prislúchajúcimi pravdepodobnosťami prechodov $0 \leq p_{ij} \leq 1$, pre $i, j \in S = \{Vstup, A, \dots, Z, Pokladňa\}$ (*Zdroj: vlastné spracovanie*)

Na obr. 2 je ilustrovaný orientovaný graf Markovovho reťazca, ktorý sa dá v našej interpretácií chápať ako štruktúra rozmiestnenia položiek v obchode. Množina stavov je daná v tvare $S = \{\mathrm{Vstup}, \mathrm{A}, \dots, \mathrm{Z}, \mathrm{Pokladňa}\} = \{1, \dots, 28\}$. Stav $\{1\}$ označuje začiatočný stav všetkých zákazníkov, teda stav z ktorého zákazníci vstupujú do modelu, ale už sa do neho nevrátia, t.j. $p_{i1} = 0$, pre všetky $i \in S$. Stav $\{28\}$ je koncový (absorbčný) stav, v ktorom nahráme transakciu do databázy a následne končíme pozorovanie zákazníka, t.j. $p_{28,28} = 1$.

Markovov reťazec je jednoznačne určený maticou pravdepodobnosti prechodu, ktorá je zadefinovaná v tvare $P = (p_{ij})_{i,j \in S}$. Z tohto vyplývajúc

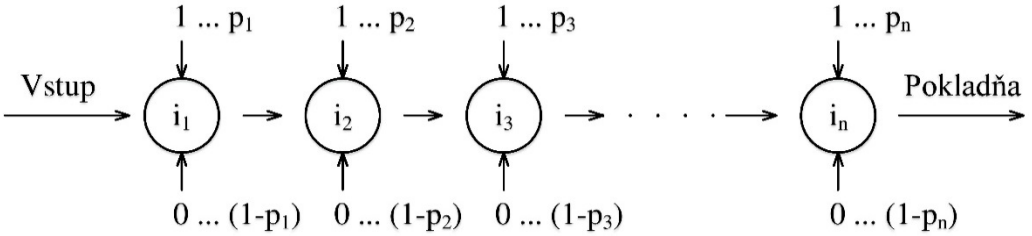
medzi orientovaným grafom ilustrovaným na obr. 2 a maticou prechodu P je vzájomne jednoznačné zobrazenie, t.j. na popísanie Markovovho reťazca nám stačí jeho matica prechodu.

Matica P pre Markovov reťazec z obr. 3 je daná v tvare

$$P = \begin{pmatrix} & Vstup & A & B & \cdots & Z & Pokladna \\ Vstup & 0 & p_{12} & p_{13} & \cdots & p_{1,27} & 0 \\ A & 0 & 0 & p_{23} & \cdots & p_{2,27} & p_{2,28} \\ B & 0 & p_{32} & 0 & \cdots & p_{3,27} & p_{3,28} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ Z & 0 & p_{27,2} & p_{27,3} & \cdots & 0 & p_{27,28} \\ Pokladna & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{5}$$

V matici P je viditeľné, že stav $\{Vstup\}$ môže byť jedine začiatočný a taktiež, že stav $\{Pokladňa\}$ je absorbčný. Na hlavnej diagonále sú, okrem posledného stavu, samé nuly, keďže zaznamenávame len prítomnosť položky v transakcii, nezávisle od jeho množstva.

Naším cieľom je vybrať model prechádzania cez množinu stavov a následne vhodne navrhnúť maticu prechodu, t.j. určíme akým spôsobom nakupujú zákazníci v danom obchode a podľa toho simulujeme databázu transakcií.



Obr. 3 Postup transakcie v simulovaní pomocou nezávislých Bernoulliho pokusov (*Zdroj: vlastné spracovanie*)

2.3 Alternatívna interpretácia metódy simulovania transakcií pomocou nezávislých Bernoulliho pokusov

V tomto modeli sa pokúšame navrhnúť takú maticu prechodu podľa ktorej vieme skonštruovať analógiu známej metódy z kapitoly 2.1. Máme k dispozícii rovnaké pravdepodobnosti zakúpenia položky $p = (p_1, \dots, p_n)$, pri rovnakom počte transakcií $|\boldsymbol{T}|$. Prechod zákazníka na priestore stavov je jednoducho ilustrovaný na obr. 3, kde zákazník vstupuje do obchodu a postupne prechádza cez množinu stavov, pričom pri každej položke sa rozhoduje s príslušnými pravdepobodnosťami, až kým sa nedostane do stavu $\{Pokladňa\}$. Z predpokladu nezávislosti položiek vieme vyjadriť pravdepodobnosť priameho prechodu zo vstupu do stavu $\{i_k\}$ v tvare $\prod_{m=1}^{k-1}(1-p_m)p_k$, pričom sa zákazník zo stavu $\{i_k\}$ nevracia naspäť. Ostatné pravdepodobnosti vyjadríme analogicky.

Na potrvdenie tejto úvahy stačí overiť, či je splnená podmienka $\sum_{j\in S} p_{ij} = 1$, kde $p_{ij} = (P)_{ij\in S}$.

Náš fiktívny obchod disponuje s $|I| = 26$ položkami, t.j. počítame s vektorom $p = (p_1, \dots, p_{26})$ v matici prechodu danej v tvare

$$P = \begin{pmatrix} 0 & p_1 & (1-p_1)p_2 & \cdots & \prod_{m=1}^{25}(1-p_m)p_{26} & \prod_{m=1}^{26}(1-p_m) \\ 0 & 0 & p_2 & \cdots & \prod_{m=2}^{25}(1-p_m)p_{26} & \prod_{m=2}^{26}(1-p_m) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{26} & (1-p_{26}) \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \qquad (6)$$

podľa ktorej implementujeme náš algoritmus do štatistického softvéru R. Na odhad neznámeho parametra λ a následného generovania vektora pravdepodobnosti $p = (p_1, \dots, p_n)$ sme použili podľa vzoru z článku (Hahsler, Hornik, Reutterer, 2005) Gamma rozdelenie s parametrami $shape = 0{,}75$ a $scale = 250$. Sledovaný časový interval je $h = 30$ dní, pričom intenzita nastátia transakcií za jednotku času je $\theta = 60$, keďže uvažujeme obchod s malým počtom položiek, $|I| = 26$.

Tab. 2 Početnosti jednopoložkových množín z množiny položiek $\boldsymbol{I}$, ktoré sme dostali zo simulácií pomocou matice prechodu ($\boldsymbol{S_M}$) a známou metódou pomocou nezávislých Bernoulliho pokusov ($\boldsymbol{S_B}$). *(Zdroj: vlastné spracovanie)*

I	A	B	C	D	E	F	G	H	I	J	K	L	M
$\boldsymbol{S_M}$	625	448	472	457	313	316	286	212	188	187	115	130	101
$\boldsymbol{S_B}$	642	430	419	454	303	297	293	263	204	163	121	95	111
I	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
$\boldsymbol{S_M}$	104	80	51	46	34	48	38	35	29	24	11	7	5
$\boldsymbol{S_B}$	105	58	57	50	39	32	46	33	31	18	11	8	4

Z tabuľky 2 je viditeľné, že sme našou metódou dostali podobné počty c_i, pre jednotlivé položky $I = \{A, \dots, Z\}$. V simuláciách sme použili rovnaký počet transakcií $|\boldsymbol{T}|$ s rovnakým vektorom pravdepodobnosti úspechu p, ktorý sme usporiadali zostupne kvôli lepšej interpretácií.

2.4 Jednoduché modely na generovanie databáz transakcií

Existuje široké spektrum príkladov, ktoré vieme modelovať nami navrhnutou metódou využitím Markovových reťazcov. Tieto modely nám slúžia na generovanie takých štruktúr databáz, ktoré obsahujú presne to, čo hľadáme pomocou algoritmov na analýzu položkových dát, teda tieto modely sú vhodné

z dôvodu testovania správnosti implementovaných algoritmov. Ako príklad uvedieme triviálny model usporiadaného obchodu.

Model usporiadaného obchodu

Uvažujme obchod, ktorý už v minulosti analyzoval transakcie svojich zákazníkov a z toho dôvodu navrhol nové usporiadanie položiek vo svojich priestoroch. Položky usporiadal tak, aby tie, ktoré mali tendeciu byť komplementárne, resp. tie, ktoré sa často nachádzali spolu v transakciách boli blízko seba. Z tohto dôvodu je rozumné simulovať transakcie pre nové usporiadanie položiek a sledovať ako sa zmenia návyky zákazníkov.

3 Analýza dát v softvéri R

Simulované dáta analyzujeme v štatistickom sofrvéri R, využívajúc vopred implementovaný algoritmus `apriori` (Agrawal, Srikant, 1994) v balíku `arules` (Hahsler et al., 2016). Implementácie algoritmov pre simulácie s využitím Markovových reťazcov sú uvedené v bakalárskej práci (Árvai, 2016), avšak boli použité ich mierne modifikácie, pričom sú uvedené správne interpretácie výsledkov.

3.1 Analýza dát simulovaných podľa alternatívnej interpretácie metódy nezávislých Bernouilloho pokusov

Po načítaní potrebných funkcií na vytvorenie matice prechodu P, a po následnom simulovaní a eliminovaní prázdnych transakcií dostávame nasledovnú štruktúru položkových dát.

```
R> summary(dataMR)
    transactions as itemMatrix in sparse format with
    1661 rows (elements/itemsets/transactions) and
    26 columns (items) and a density of 0.101005
    most frequent items:
       A       C       D       B       F  (Other)
     625     472     457     448     316    2044
    element (itemset/transaction) length distribution:
    sizes
     1   2   3   4   5   6   7  8  9
    350 525 401 239  90  42  12  1  1
```

Zo sumarizácie vlastností dátového súboru je viditeľné, že sme dospeli najmä k 1 až 4-položkovým množinám, pričom sa budeme zaoberať hlavne s 2 a 3-položkovými. Po úvodných výpočtoch nastavíme parametre početnosti (`minsup`), resp. spoľahlivosti (`minconf`) pre naše pravidlá na 1%. Využitím implementácie algoritmu `apriori` dostávame nasledujúcu sumarizáciu štruktúry asociačných pravidiel.

```
R> summary(pravidlaMR)
set of 107 rules
rule length distribution (lhs + rhs):sizes
 2  3
83 24
summary of quality measures:
    support         confidence        lift
 Min.   :0.01023   Min.   :0.1062   Min.   :0.6573
 1st Qu.:0.01355   1st Qu.:0.2000   1st Qu.:0.9143
 Median:0.01927   Median:0.2524   Median :1.0040
 Mean   :0.02639   Mean   :0.2621   Mean   :1.0309
 3rd Qu.:0.03070   3rd Qu.:0.3207   3rd Qu.:1.1554
 Max.   :0.10054   Max.   :0.4375   Max.   :1.5292
```

Výsledky ukazujú, že sme sa dopracovali k 107 asociačným pravidlám pri 1%-nej požadovanej hodnote parametrov `minsup` a `minconf`. Hodnota parametra `lift` je blízko k 1, čiže v nami navrhnutom modeli majú položky tendenciu byť nezávislými. Taktiež je viditeľné, že sledované položkové množiny majú nízke hodnoty početnosti, avšak hodnoty spoľahlivosti sú vyššie ako by sme očakávali pri predpoklade nezávislosti. Pre zistenie príčiny by sme potrebovali hlbšie analyzovať náš dátový súbor, avšak môže to byť spôsobené tým, že uvažujeme obchod s malým počtom položiek, resp. metódou, ktorou generujeme pravdepodobnosti zakúpenia jednotlivých položiek.

```
R> summary(pravidlaNBP)
set of 94 rules
rule length distribution (lhs + rhs):sizes
2  3
76 18
summary of quality measures:
support         confidence        lift
Min.   :0.01021   Min.   :0.1043   Min.   :0.5731
1st Qu.:0.01276   1st Qu.:0.1966   1st Qu.:0.8816
Median:0.01892   Median :0.2528   Median :0.9895
Mean   :0.02650   Mean   :0.2529   Mean   :1.0040
3rd Qu.:0.03408   3rd Qu.:0.2986   3rd Qu.:1.0904
Max.   :0.10030   Max.   :0.4373   Max.   :1.4605
```

Vo vyššie uvedenom výstupe je sumarizácia pravidiel pre simuláciu transakcií pomocou nezávislých Bernoulliho pokusov, pri rovnako nastavených parametroch `minsup` a `minconf`. Dospeli sme k veľmi podobným hodnotám

sledovaných mier zaujímavosti, podľa čoho vieme skonštatovať, že sa nám podarilo vytvoriť analógiu známeho modelu.

3.2 Analýza dát pre model usporiadaného obchodu

Po implementácií a načítaní potrebných funkcií a pri rovnakom nastavení parametrov `minsup`, `minconf` ako v predchádzajúcom modeli dostávame nasledujúcu štruktúru asociačných pravidiel.

```
summary of quality measures:
    support         confidence          lift
 Min.  :0.01765   Min.   :0.3158   Min. :1.054
 1st Qu. :0.02497   1st Qu. :0.4099   1st Qu. :1.776
 Median:0.03354   Median:0.4709   Median :2.202
 Mean  :0.04328   Mean  :0.4702   Mean  :2.214
 3rd Qu.:0.05095   3rd Qu. :0.5260   3rd Qu. :2.618
 Max.  :0.16116   Max.  :0.6223   Max.  :3.611
mining info:
 data  ntransactions  support  confidence
 TUSP          3965         0.01        0.01
```

Parameter `lift` poukazuje na vyššiu mieru závislosti medzi položkovými množinami ako v predchádzajúcom prípade. Rovnako parametre početnosti a spoľahlivosti sú vyššie, keďže sme maticu prechodu definovali spôsobom v ktorom sa zákazník vie voľne pohybovať, resp. vrátiť sa k predchádzajúcej položke. Zvolili sme maximálny počet položiek v transakciách (10 položiek), aby sme nedospeli k zbytočne veľkým transakciám.

```
R> inspect(sort(pravidlaUSP, by="lift")[1:3])
      lhs     rhs     support     confidence     lift
203 {F,I}  => {H}  0.03253468   0.6201923  3.610958
92  {H,J} => {I}   0.02774275   0.5527638  3.604784
95  {G,J} => {I}   0.02547289   0.5489130  3.579671
```

Na vybraných pravidlách je viditeľné, že sme dospeli k očakávaným výsledkom, keďže poukazujú na súvislosti práve medzi blízkymi položkami, t.j. správne opisujú model usporiadaného obchodu.

4 Záver

V tomto článku sme sa venovali problematike simulovania položkových dát, ktoré vieme analyzovať s cieľom nájsť asociácie medzi jednotlivými položkami. Navrhli sme alternatívnu metódu simulovania transakcií pomocou Markovových reťazcov, pričom sme stručne predstavili známy model, ktorý je

založený na nezávislých Bernoulliho pokusoch. Ukázali sme, že známy model predstavený v článku (Hahsler, Hornik, Reutterer, 2005) je jedným špeciálnym prípadom nami navrhnutého modelu. Autori v článku v ktorom bola predstavená známa metóda poukázali aj na potenciálny prínos takého modelu, ktorý nám umožní simulovať transakcie s prítomnosťou asociácií medzi položkami. Pomocou nami predstavenej metódy vieme naplniť aj túto požiadavku, keďže vieme navrhnúť viacero scenárov, ako sme to aj predstavili na jednoduchom príklade v modeli usporiadaného obchodu.

Uvedený model by sa dal samozrejme zlepšiť, napríklad efektívnejším navrhnutím matice prechodu, alebo generovaním celej stochastickej simulačnej matice. Taktiež by sa mohla navrhnúť implementácia simulácie pre obchod s väčším, ľubovoľným počtom položiek.

Predstavená metóda simulovania vytvára nové príležitosti výskumu nielen v oblasti simulačných metód, ale aj priamo v rámci analýzy položkových dát, keďže otvára nové otázky z pohľadu testovania, resp. zavedenia nových mier zaujímavosti.

5 Literatúra

Agrawal, R., Imielinski, T., Swami, A. (1993*). Mining association rules between sets of items in large databases.* In Proc. of the AMC SIGMOD International Conference on Management of Data, Washington D.C., 207-216.

Agrawal, R., Srikant, R. (1994*). Fast algorithms for mining association rules.* Proc. 20th Int. Conf. Very Large Data Bases, Santiago, Chile, 487-499.

Árvai, M. (2016). *Analýza položkových dát.* Bakalárska práca, FMFI UK, Bratislava.

Brin, S., Motwami, R., Ullman, J.D., Tsur, S. (1997). *Dynamic itemset counting and implication rules for market basket data.* In Proceedings of the 1997 AMC SIGMOD international conference on Management of data, New York, NY, USA, 255-264.

Hahsler, M., Buchta, Ch., Gruen, B., Hornik, K. (2016). arules: *Mining Association Rules and Frequent Itemsets.* R package version 1.4-1. [Dostupné na: https://cran.r-project.org/web/packages/arules/vignettes/arules.pdf].

Hahsler, M., Hornik, K., Reutterer, T. (2005). *Implications of probabilistic data modeling for rule mining.* Research Report Series. Department of Statistics and Mathematics, 14. Institut für Statistik und Mathematik, WU Vienna University of Economics and Business, Vienna. [Dostupné na: http://epub.wu.ac.at/764/].

Aktuárske modelovanie vybraných poistných produktov dôchodkového sporenia
Actuarial modelling of selected insurance products of pension savings

Tatiana Duchoňová
Univerzita Mateja Bela, Ekonomická fakulta, inžinierske štúdium, I. ročník
Matej Bel University, Faculty of Economics
tana.duchonova@gmail.com

Abstrakt: *Článok sa zaoberá faktormi, ktoré ovplyvňujú výšku poistného v životnom poistení. Ponúka poistno-matematické modelovanie vybraných poistných produktov dôchodkového sporenia. Základ modelovania tvoria úmrtnostné tabuľky platné v Slovenskej republike v roku 2015 a v súčasnosti platná maximálna výška technickej úrokovej miery 0,7 % p. a..*

Abstract: *The paper is focused on factors which influence the amount of life insurance premiums. It offers actuarial modelling of selected insurance products as a part of pension savings. The basis for modelling is the life table valid in the Slovak Republic of 2015 and the currently valid maximum amount of technical interest rate 0.7% p. a..*

Kľúčové slová: *anuita, dôchodok, životné poistenie, úmrtnostné tabuľky, technická úroková miera*

Key words: *annuity, pension, life insurance, life tables, technical interest rate*

1 Úvod

Jedným z najväčších súčasných problémov mnohých krajín je riešenie otázky dôchodkového zabezpečenia obyvateľstva. Výnimkou nie je ani Slovenská republika. Počet dôchodcov stále narastá a naopak, ekonomicky aktívne obyvateľstvo ubúda. Väčšina ľudí sa spolieha na to, že štát bude schopný zabezpečiť všetky ich finančné potreby v budúcnosti a nevytvárajú si iné možnosti zaistenia na dôchodok. Ak chcú mať súčasní aktívne pracujúci obyvatelia vyšší dôchodok, mali by využiť aj iné, ako povinné formy zabezpečenia sa do budúcnosti. Napríklad, momentálne existuje široká škála rôznych poistných produktov, ako možností dôchodkového sporenia.

Cieľom článku je na základe poistno-matematického modelovania poukázať na faktory ovplyvňujúce vybrané produkty dôchodkového sporenia.

2 Vybrané faktory ovplyvňujúce poistenie osôb

Za najdôležitejšie faktory, ovplyvňujúce kalkulácie v životnom poistení, považujeme:

- pohlavie poisteného,
- vek poisteného,
- dĺžku poistnej doby,

- technickú úrokovú mieru,
- úmrtnostné tabuľky,
- náklady poisťovne.

Medzi tieto faktory sme zaradili aj pohlavie, avšak smernica Rady 2004/113/ES z 13. decembra 2004[1] o vykonávaní zásady rovnakého zaobchádzania medzi mužmi a ženami v prístupe k tovaru a službám a k ich poskytovaniu zakazuje používať tento faktor ako poistno-matematické kritérium. Aby sa zabezpečilo rovnaké zaobchádzanie medzi mužmi a ženami, použitie pohlavia ako poistno-matematického kritéria by nemalo mať za následok rozdiely v poistnom, poistných sumách, prémiách a zľavách. Toto pravidlo sa vzťahuje len na nové zmluvy, ktoré boli uzavreté po dátume predloženia tejto smernice, teda na zmluvy uzatvorené po decembri 2012.

Autorky Špirková a Urbaníková (2012) považujú za poistnú udalosť v rámci poistenia osôb v životnom poistení okrem iného aj úmrtie alebo dožitie sa určitého veku. Preto dôležitú úlohu pri poistno-matematických výpočtoch majú úmrtnostné tabuľky, technická úroková miera ale aj náklady poisťovne. Vzhľadom na rozsah článku abstrahujeme od nákladov poistenia a počítame len tzv. čisté „netto“ dôchodky, ktoré budú ovplyvnené len vekom poisteného, teda v našom prípade, vekom nástupu na dôchodok a strednou dĺžkou dožitia sa.

2.1 Technická úroková miera

Technická úroková miera (*i % p.a.*) je úroková miera, ktorú poisťovňa používa pre všetky výpočty týkajúce sa produktov poistenia a na výpočet poistného. Predstavuje zhodnotenie technickej rezervy životného poistenia. Inak povedané, predstavuje minimálny výnos, o ktorý sa poisťovňa zaväzuje zhodnotiť finančné prostriedky poisteného, ktoré vložil do poistenia (Urbaníková et al., 2014).

Na základe Opatrenia Národnej banky Slovenska č. 25/2015 z 1. decembra 2015, o maximálnej výške technickej úrokovej miery, je v súčasnosti platná maximálna výška technickej úrokovej miery pre poisťovne 0,7 % p. a..

Smernica Solventnosť II[2] v článku 77, odseku 2 popisuje zhodnotenie technickej rezervy životného poistenia tak, že najlepší odhad by mal korešpondovať pravdepodobnosťou váženému priemeru budúcich peňažných tokov s prihliadaním na časovú hodnotu peňazí (očakávaná súčasná hodnota budúcich peňažných tokov) a s použitím príslušnej časovej štruktúry bezrizikovej úrokovej miery. Vyjadrené jazykom praktických výpočtov, táto

[1] Známa pod názvom Gender smernica

[2] Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II)

smernica vyžaduje okrem použitia pravdepodobností úmrtia alebo dožitia sa určitého veku aj štruktúru bezrizikovej úrokovej miery pomocou tzv. výnosových kriviek bezrizikových cenných papierov. V našich prvotných výpočtoch avšak používame konštantnú technickú úrokovú mieru. Modelovaniu pomocou výnosových kriviek sa budeme venovať v našej následnej práci.

Európska centrálna banka ECB (2017) popisuje výnosovú krivku ako časovú štruktúru bezrizikových úrokových mier z bezrizikových štátnych dlhopisov. ECB zverejňuje parametre výnosových kriviek na dennej báze.

2.2 Úmrtnostné tabuľky

Úmrtnostné tabuľky (ÚT) sú nástrojom na zisťovanie úmrtnostných pomerov určitej populácie. Poskytujú pohľad podľa vekovej štruktúry na pravdepodobnosť úmrtia a dožitia (Meszároš, 2000).

Pre účely výpočtov sme zostavili ÚT na základe údajov Štatistického úradu Slovenskej republiky za rok 2015 pri konštantnej technickej úrokovej miere 0,7% p. a..

V poistnej matematike využívame základný princíp poistno-matematických výpočtov, tzv. *princíp ekvivalencie*, čiže pri uzatváraní súboru poistných zmlúv musia byť všetky príjmy poisťovne rovné výdavkom, keď sa diskontujú k tomu istému časovému okamihu (Špirková, Urbaníková, 2012).

V našom prípade bude tento časový okamih vek jedinca v čase uzavretia poistenia, pričom diskontný fakor

$$v = \frac{1}{1 + \frac{i}{100}} \tag{1}$$

Ako uvádza Dickson et al. (2009), v prípade použitia bezrizikových úrokových sadzieb z výnosových kriviek je odúročiteľ v^* počítaný pomocou spojitej úrokovej miery podľa vzťahu

$$v^* = e^{-\frac{\delta}{100}} \tag{2}$$

kde δ predstavuje spotovú bezrizikovú spojitú úrokovú mieru.

V našom článku modelujeme produkty s použitím konštantného odúročiteľa, vypočítaného podľa vzťahu (1).

V jednotlivých stĺpcoch ÚT sme počítali nasledujúce údaje označené podľa Dicksona et al. (2009):

- x – jedinec žijúci vo veku x, pričom $x \geq 0$, čiže $x = 1, 2, 3, ..., \omega$, kde ω je maximálny vek, ktorého sa môže jedinec dožiť,

- q_x - pravdepodobnosť, že x-ročný jedinec zomrie vo veku *x*,
- p_x - pravdepodobnosť, že x-ročný jedinec prežije aspoň 1 rok, kde:

$$p_x = 1 - q_x \tag{3}$$

Keďže dôchodky z II. a III. piliera dôchodkového sporenia sa vyplácajú mesačne, tak na výpočet tzv. področnej (v našom prípade mesačnej) pravdepodobnosti úmrtia jedinca vo veku x do veku $x+s$, respektíve področnej pravdepodobnosti prežitia jedinca vo veku x aspoň veku $x+s$, môžeme použiť dva nasledujúce spôsoby:

1. Predpoklad zlomkového veku, ktorý vo svojej knihe opisuje Dickson et al. (2009). Na výpočet používa predpoklad rovnomerného rozloženia úmrtí[3]. Pre x-ročného človeka a v prípade, že $0 \le s \le 1$, predpokladáme že:

$$_s q_x = s \cdot q_x \tag{4}$$

a následne

$$_s p_x = 1 - _s q_x \tag{5}$$

2. Balducciho predpoklad, ktorý vo svojej knihe opisuje Potocký (2012). Predpokladá, že pre $x \in N$ a $0 < s < 1$ ročná miera úmrtnosti spĺňa vzťah

$$_{1-s} q_{x+s} = (1-s) \cdot q_x \tag{6}$$

a zároveň platí

$$p_x = {}_s p_x \cdot {}_{1-s} p_{x+s}, \tag{7}$$

odkiaľ

$$_s p_x = \frac{1}{1 + s \cdot \frac{q_x}{p_x}} \tag{8}$$

V našej úmrtnostnej tabuľke sme počítali podľa prvého spôsobu, čiže podľa vzťahov (4) a (5). Časť úmrtnostnej tabuľky môžeme vidieť v Tabuľke 1.

Tab. 1 Úmrtnostná tabuľka *(Zdroj:vlastné spracovanie podľa údajov Štatistického úradu SR)*

x	q_x	p_x	s	$_s q_x$	$_s p_x$
65	0,016722	0,983278	0,500000	0,000062	0,999938
66	0,018433	0,981567	0,583333	0,000072	0,999928
67	0,020132	0,979868	0,666667	0,000082	0,999918
...	...	...	...	...	...
98	0,521100	0,478900	0,250000	0,000020	0,999980
99	0,569990	0,430010	0,333333	0,000027	0,999973
100	0,620000	0,380000	0,416667	0,000033	0,999967

[3]Uniform distribution of deaths assumption

3 Modelovanie vybraných poistných produktov

3.1 Doživotný starobný dôchodok z II. piliera

Osoba vo veku x si v životnej poisťovni kúpi doživotný dôchodok z II. piliera, vyplácaný 12-krát ročne vo výške *MD*, čiže uzavrie poistnú zmluvu, podľa ktorej jej bude poisťovňa z nasporenej sumy *S* vyplácať dôchodok v stanovenej výške, pokiaľ bude poistená osoba nažive. Podľa Zákona č. 43/2004 Z. z. o starobnom dôchodkovom sporení a o zmene a doplnení niektorých zákonov sa dôchodok z II. piliera vypláca polehotne, čo v našom prípade znamená, že osobe vo veku x bude vyplácaný na konci mesiaca. Ak by poberateľ doživotného dôchodku z II. piliera zomrel skôr, ako mu bolo vyplatených 84 mesačných anuít doživotného dôchodku, sumu zodpovedajúcu rozdielu sumy určenej na výplatu týchto 84 mesačných anuít doživotného dôchodku a súčtu vyplatených mesačných anuít doživotného dôchodku prevedie oprávnenej osobe. To znamená, že týchto 84 mesačných súm doživotného dôchodku, predstavuje určitý garantovaný dôchodok a označujeme ho $(Ga)^{(12)}_{x:\overline{7}|}$.

Na výpočet garantovaného dôchodku sme odvodili vzťah:

$$(Ga)^{(12)}_{x:\overline{7}|} = \sum_{r=1}^{83} \frac{84-r}{12} \cdot {}_{\frac{r}{12}|\frac{1}{12}}q_x \cdot v^{\frac{r+1}{12}} \tag{9}$$

kde ${}_{\frac{r}{12}|\frac{1}{12}}q_x$ predstavuje pravdepodobnosť, že jedinec vo veku x prežije $\frac{r}{12}$ a potom počas nasledujúcej $\frac{1}{12}$ roka zomrie, to znamená medzi vekom $x+\frac{r}{12}$ a $x+\frac{r}{12}+\frac{1}{12}$. Túto pravdepodobnosť sme počítali pomocou vzorca, ktorý uvádza Dickson et al. (2009) v tvare

$${}_{u|t}q_x = {}_up_x - {}_{u+t}p_x \tag{10}$$

V prípade, keď chceme zistiť, akú výšku netto mesačného dôchodku *MD* bude x-ročný jedinec doživotne poberať z nasporenej sumy *S,* použijeme vzťah

$$MD = \frac{S}{12 \cdot (a^{(12)}_x + (Ga)^{(12)}_{x:\overline{7}|})}, \tag{11}$$

kde $a^{(12)}_x$ predstavuje očakávanú súčasnú hodnotu doživotného anuitného dôchodku, splatného polehotne v mesačných intervaloch, vo výške 1/12 p. j.

$$a^{(12)}_x = \sum_{r=1}^{12\cdot(\omega-x)} \frac{1}{12} \cdot {}_{\frac{r}{12}}p_x \cdot v^{\frac{r}{12}} \tag{12}$$

Modelová situácia

Predstavme si situáciu, že máme troch ľudí vo veku x, konkrétne 65, 70 a 75-ročného človeka. V poisťovni si kúpili doživotný starobný dôchodok z II. piliera, ktorý im bude vyplácaný na konci každého mesiaca. Každý z nich má nasporených 10 000 EUR. Chceme vypočítať, akú výšku čistého mesačného dôchodku *MD* bude dostávať 65, 70 a 75-ročný človek. Za rizikový faktor považujeme len vek a nasporenú sumu, abstrahujeme od ostatných faktorov.

Najskôr vypočítame garantovaný mesačný dôchodok $(Ga)^{(12)}_{x:\overline{7}|}$, podľa vzťahu (9). Ďalej potrebujeme vypočítať očakávanú súčasnú hodnotu polehotného mesačného doživotného dôchodku $a_x^{(12)}$ vo výške 1/12 p. j. podľa vzťahu (12).

Po dosadení našich hodnôt do vzťahov (9), (11) a (12) dostávame výsledky zosumarizované v Tabuľke 2.

Tab. 2 Netto mesačný doživotný dôchodok z II. piliera *(Zdroj: vlastné spracovanie)*

(*x*) vek jedinca v rokoch	**65**	**70**	**75**
$a_x^{(12)}$ (p.j.)	0,453847	0,653387	1,043700
$(Ga)^{(12)}_{x:\overline{7}|}$ (p.j.)	15,640039	12,550312	9,564245
S (EUR)	10 000	10 000	10 000
MD (EUR)	**51,78**	**63,11**	**78,56**

Na jednotlivých sumách netto mesačného dôchodku *MD* v Tabuľke 2, môžeme vidieť, ako naň vplýva vek a nasporená suma, pričom ostatné podmienky sú rovnaké. Je zrejmé, že čím vo vyššom veku začne jedinec poberať dôchodok, tým ho bude mať z rovnakej nasporenej sumy vyšší. V tom prípade starší poistený bude dostávať vyšší dôchodok z danej nasporenej sumy, ako mladší.

3.2 Doplnkový doživotný starobný dôchodok z III. piliera

Osoba vo veku x, ktorá uzavrela účastnícku zmluvu pred 31. decembrom 2013 a nepodpísala dodatok k účastníckej zmluve podľa Zákona č. 650/2004 Z. z. o doplnkovom dôchodkovom sporení a o zmene a doplnení niektorých zákonov, požiada o vyplácanie doplnkového doživotného starobného dôchodku z III. piliera, vyplácaného predlehotne 12-krát ročne MD^* zo svojej aktuálnej hodnoty dôchodkového osobného účtu S.

V prípade, keď chceme zistiť, akú výšku čistého mesačného dôchodku MD^* bude x-ročný jedinec doživotne poberať z nasporenej sumy *S,* použijeme vzťah

$$MD^* = \frac{S}{12 \cdot \ddot{a}_x^{(12)}} \tag{13}$$

kde $\ddot{a}_x^{(12)}$ predstavuje očakávanú súčasnú hodnotu doživotného dôchodku, splatného predlehotne v mesačných intervaloch vo výške 1/12 p.j. a vypočítame ho podľa vzťahu

$$\ddot{a}_x^{(12)} = \sum_{r=0}^{12\cdot(\omega-x)} \frac{1}{12} \cdot {}_{\frac{r}{12}}p_x \cdot v^{\frac{r}{12}} \qquad (14)$$

Modelová situácia

Predpokladajme, že máme troch ľudí vo veku x, a to 65, 70 a 75 rokov. Uzavreli účastnícku zmluvu s doplnkovou dôchodkovou spoločnosťou, v ktorej majú obsiahnutý dávkový plán. Požiadajú o vyplácanie doplnkového starobného dôchodku z III. piliera, ktorý im bude vyplácaný predlehotne každý mesiac, pokiaľ budú nažive. Aktuálna hodnota ich dôchodkových účtov je 10 000 EUR. Chceme vypočítať, akú výšku netto mesačného dôchodku MD^* bude dostávať 65, 70 a 75-ročný človek. Za rizikový faktor považujeme len vek a dĺžku poistnej doby, abstrahujeme od ostatných činiteľov.

V prvom rade potrebujeme vypočítať očakávanú súčasnú hodnotu mesačného predlehotného doživotného dôchodku $\ddot{a}_x^{(12)}$ vo výške 1/12 p. j., ktorú dostaneme dosadením hodnôt do vzťahu (14).

Po dosadení všetkých požadovaných hodnôt do vzťahov (13) a (14) dostávame výsledky, zosumarizované v Tabuľke 3.

Tab. 3 Netto mesačný doživotný predlehotný dôchodok z III. piliera dôchodkového sporenia*(Zdroj: vlastné spracovanie)*

(*x*) vek jedinca v rokoch	**65**	**70**	**75**
$\ddot{a}_x^{(12)}$ (p.j.)	15,723372	12,633646	9,647578
S (EUR)	10 000	10 000	10 000
MD^* (EUR)	**53,00**	**65,96**	**86,38**

Tak ako v prípade doživotného dôchodku z II. piliera, môžeme aj tento krát sledovať (Tabuľka 3), ako vplýva vek a poistná doba na výšku netto mesačného dôchodku MD^*, pričom ostatné podmienky sú rovnaké. V tom prípade bude starší človek dostávať vyšší dôchodok z danej nasporenej sumy ako mladší. Treba si všimnúť, že aj keď máme v prípade II. a III. piliera rovnaký vek x a rovnakú dĺžku poistnej doby, výška netto mesačných dôchodkov *MD* a MD^* je odlišná. Je to spôsobené rozdielnym časom výplaty(predlehotne, polehotne) a samozrejme, v prípade dôchodku z II. piliera aj vyplatením doposiaľ nevyplatených dôchodkov pozostalým, v prípade úmrtia dôchodcu.

4 Záver

Môžeme konštatovať, že v oblasti dôchodkového zabezpečenia je momentálne najväčším problémom negatívny demografický vývoj. V SR štát zabezpečuje dôchodcov v rámci I. piliera, ktorý je ale závislý na príspevkoch aktívnej časti obyvateľstva. Avšak z dlhodobého hľadiska výskumy ukazujú rýchlejšie starnutie obyvateľstva, čo môže spôsobiť veľké problémy vo financovaní povinného dôchodkového poistenia. Preto si myslíme, že by sa občania nemali spoliehať len na to, že štát ich bude schopný v budúcnosti zabezpečiť, a mali by využívať aj iné, ako povinné formy zabezpečenia sa do dôchodku. Jednou z možností je prispievanie do II. a III. piliera, ktoré sme sledovali na modelových situáciách.

Na základe výpočtov dôchodku z II. a III. piliera sme sledovali, ako vplýva vek a poistná doba na výšku netto mesačného dôchodku, pričom ostatné faktory boli rovnaké. V tom prípade bude starší človek dostávať vyšší dôchodok z danej nasporenej sumy ako mladší človek. Aj keď sme v prípade II. a III. piliera mali rovnaký vek a rovnakú dĺžku poistnej doby, výška netto mesačného dôchodku bola odlišná. Doživotný starobný mesačný dôchodok z II. piliera je vyplácaný na konci mesiaca a doplnkový doživotný starobný mesačný dôchodok z III. piliera je vyplácaný na začiatku každého mesiaca. Taktiež je to spôsobené tým, že pri II. pilieri figuruje riziko garantovaného dôchodku počas prvých sedem rokov, čo znižuje dôchodkovú dávku. Na záver by sme chceli upozorniť, že naše výpočty poukazujú na spôsob výpočtu jednotlivých dôchodkov, avšak tieto výšky dôchodkov nepredstavujú reálne dôchodky, pretože sme abstrahovali od väčšieho počtu faktorov vplývajúcich na výšku dôchodkov. Vplyvu viacerých faktorov na výšku dôchodkov sa plánujeme venovať v našej ďalšej práci.

5 Literatúra

Dickson, D., et al. 2012. Actuarial Mathematics for Life Contingent Risks. Cambridge: Cambridge University Press, 2012, 493 p.

Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II) [Dostupné na: http://www.tsb.org.tr/images/Documents/SolvencyIIDirektifi. pdf]

European central bank.(2017).*Euro area yield curves.* Frankfurt am Main Germany: 2017. [Dostupné na: https://www.ecb.europa.eu/stats/financial_markets_and_interest_rates/euro_area_yield_curves/html/index.en.html]

Európska únia. (2004). *Smernica Rady 2004/113/ES z 13. decembra 2004 o vykonávaní zásady rovnakého zaobchádzania medzi mužmi a ženami v prístupe k tovaru a službám a k ich poskytovaniu*. [Dostupné na: http://eurlex.europa.eu/legal-content/SK/TXT/?uri=celex%3A32004L0113]

Mézsáros, J. (2000). *Metodický materiál - Výpočet úmrtnostných tabuliek, Výpočet stratených rokov života úmrtím*. Bratislava: INFOSTAT - Inštitút informatiky a štatistiky, 2000, 14s. [Dostupné na: http://www.infostat.sk/vdc/pdf/metodika_ut.pdf]

Národná banka Slovenska. (2015).*Opatrenie Národnej banky Slovenska č. 25/2015 z 1. decembra 2015, o maximálnej výške technickej úrokovej miery* [Dostupné na: http://www.nbs.sk/_img/Documents/_Legislativa/_Vestnik/OPAT25-2015.pdf]

Potocký, R. (2012). *Modely v životnom a neživotnom poistení*. Bratislava: Statis, 2012. 216 s.

Špirková, J., Urbaníková, M. (2012). *Aktuárska matematika: životné poistenie*. Bratislava: IURA Edition, 2012, 241 s.

Štatistický úrad SR. (2017) [Dostupné na: https://slovak.statistics.sk/]

Urbaníková, M., et al. (2014). *Aplikovaná matematika: vybrané kapitoly z finančnej a poistnej matematiky.* Bratislava : Slovenská technická univerzita, 2014, 201 s.

Zákon č. 650/2004 Z. z. o doplnkovom dôchodkovom sporení a o zmene a doplnení niektorých zákonov.

Zákon č. 43/2004 Z. z. o starobnom dôchodkovom sporení a o zmene a doplnení niektorých zákonov.

Loggama rozdelenie v demografickej štatistike
Log-gamma distribution in demographic statistics

Daniel Hladký
Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, magisterské štúdium, 1.ročník
Commenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, 1st year of magister level of study
danohladky@gmail.com

Abstrakt: *Táto práca skúma rôzne typy log-gama rozdelení a vzťahy medzi nimi. Za špeciálny prípad zovšeobecneného log-gama rozdelenia možno považovať aj Coale-McNeilovo rozdelenie, ktoré je základom známeho Coale-McNeilovho modelu prvosobášnosti. V práci je na dátach pre Slovenskú republiku ukázané, že tento model možno s úspechom použiť aj na fitovanie špecifických mier sobášnosti, plodnosti a prvoplodnosti. Špecifické miery týchto demografických procesov sú porovnávané s gama a log-normálnym modelom pomocou súčtu štvorcov odchýlok.*

Abstract: *This thesis examines different types of log-gamma distributions and relations between them. As a special case of generalized log-gamma distribution, it can be considered Coale-McNeil distribution, which is basis of the well-known Coale-McNeil nuptiality model. In the data for Slovak republic, there is shown, that this model can be also successfully used on fitting specific rates of nuptiality, fertility, and first birth. Specific rates of these demographic processes are compared with gamma and log-normal model by sum of squared errors.*

Kľúčové slová: *Sobášnosť, plodnosť, špecifické miery, log-gama rozdelenie, Coale-McNeilov model*

Key words: *Nuptiality, Fertility, Specific rates, Log-gamma distribution, Coale-McNeil model*

1 Úvod

Táto práca skúma rôzne typy log-gama rozdelení, vzťahy medzi nimi a ich použitie pri modelovaní špecifických mier vybraných demografických procesov. Zamerali sme sa na 6 procesov: prvosobášnosť mužov a žien, sobášnosť mužov a žien, prvoplodnosť a plodnosť.

Plodnosť charakterizuje špecifická miera plodnosti podľa veku (*age-specific fertility rate* - f_x), ktorá vyjadruje počet živonarodených detí N_x ženám v určitom veku x k strednému stavu žien P_x v tom istom veku za jeden rok. Je definovaná rovnosťou:

$$f_x = \frac{N_x}{P_x}. \tag{1}$$

Hodnota f_x bude v texte vyjadrovať vždy inú špecifickú mieru v závislosti od demografického procesu a takisto aj hodnoty P_x a N_x. Súčet špecifických mier plodnosti vyjadrujúcich intenzitu plodnosti danej populácie v danom časovom období (väčšinou jeden kalendárny rok) sa nazýva úhrnná plodnosť (*total fertility rate* - TFR). Udáva počet detí, ktoré by sa narodili jednej žene počas

reprodukčného obdobia za predpokladov nulovej úmrtnosti a toho, že by sa hodnoty špecifických mier plodnosti nemenili počas celého reprodukčného obdobia. Platí teda:

$$P = \sum_x f_x,$$

kde P v tomto prípade predstavuje TFR.

Analogicky to platí aj u ostatných demografických procesov. Pre úhrnné miery budeme v práci používať nasledovné označenia: úhrnná prvoplodnosť ($TFR1$), úhrnná miera sobášnosti (pre mužov $TMRm$ a pre ženy $TMRf$), úhrnná miera prvosobášnosti (pre mužov $TMR1m$ a pre ženy $TMR1f$).

2 Typy log-gama rozdelení a vzťahy medzi nimi

2.1 Transformácia náhodných premenných

Transformáciu náhodných premenných používame v úlohách, v ktorých chceme nájsť rozdelenie náhodnej premennej Y, ktorá je zviazaná s náhodnou premenou X funkčných vzťahom $Y = g(X)$ a pričom poznáme rozdelenie náhodnej premennej X. Pre prípad spojitej náhodnej premennej si uvedieme $Vetu\ 1$. prevzatú z (Lamoš, Potocký, 1998).

$Veta\ 1$. Nech spojitá náhodná premenná X má hustotu f. Nech g je rýdzo monotónna funkcia, ktorá má spojitú prvú deriváciu rôznu od nuly. Potom pre hustotu pravdepodobnosti $h(y)$ náhodnej premennej $Y = g(X)$ platí:

$$h(y) = f\left(g^{-1}(y)\right)\left|\frac{d}{dy}\left(g^{-1}(y)\right)\right|,$$

kde g^{-1} je inverzná funkcia k funkcii g.

2.2 Gama rozdelenie

Rozdelenie gama je spojité rozdelenie pravdepodobnosti s dvomi parametrami $\alpha > 0$ a $\beta > 0$. Funkcia hustoty gama rozdelenia pre $x > 0$ má tvar:

$$f(x) = \frac{x^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\, exp\left\{-\frac{x}{\beta}\right\},$$

inak $f(x) = 0$. Označenie $\Gamma(\alpha)$ predstavuje gama funkciu, definovanú Eulerovým integrálom 2.druhu ako:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} exp\{-x\}\, dx, \qquad \alpha > 0.$$

2.3 Log-gama rozdelenie (1)

Log-gama rozdelenie (1) (LG_1) je naše označenie pre prvý typ log-gama rozdelenia (Hogg, Klugman, 2008) s funkciou hustoty pravdepodobnosti:

$$f(x) = \frac{\lambda^{\alpha}(\ln x)^{\alpha-1}}{x^{\lambda+1}\Gamma(\alpha)}, \tag{2}$$

definovanou pre $x > 1$ s parametrami $\alpha > 0, \lambda > 0$. Vzťah medzi gama rozdelením a log-gama rozdelením(1) je nasledovný. Ak náhodná premenná X sa riadi log-gama rozdelením(1), tak potom náhodná premenná $Y = ln(X)$ sa riadi gama rozdelením. Platnosť tohto vzťahu si ukážeme pomocou $Vety\ 1$. Nech $Y = ln(X) = g(x)$ a nech $X = e^{Y} = g^{-1}(y)$. Podľa $Vety\ 1$ dostaneme:

$$h(y) = f(e^{Y})\left|\frac{d}{dy}(e^{Y})\right| = \frac{\lambda^{\alpha}(\ln(e^{y}))^{\alpha-1}}{(e^{y})^{\lambda+1}\Gamma(\alpha)}e^{y} = \frac{\lambda^{\alpha}y^{\alpha-1}}{(e^{y})^{\lambda+1}\Gamma(\alpha)}e^{y} = \frac{\lambda^{\alpha}y^{\alpha-1}}{e^{y\lambda}\Gamma(\alpha)}.$$

Ak parameter $\lambda = 1/\beta$, tak dostaneme nasledovnú rovnosť:

$$\frac{\left(\frac{1}{\beta}\right)^{\alpha}y^{\alpha-1}}{\Gamma(\alpha)}\exp\left(-\frac{y}{\beta}\right) = \frac{y^{\alpha-1}\exp\left(-\frac{y}{\beta}\right)}{\beta^{\alpha}\Gamma(\alpha)},$$

čo predstavuje funkciu hustoty gama rozdelenia. Čiže ak $X{\sim}LG_1(\alpha,\lambda) => Y = ln(X){\sim}G(\alpha,\lambda)$.

2.4 Log-gama rozdelenie (2)

Ako log-gama rozdelenie (2) (LG_2) sme označili rozdelenie prevzatého z (WM, online) s dvomi kladnými parametrami α a β, ktorého funkcia hustoty pravdepodobnosti náhodnej premennej X je daná ako:

$$f(x) = \frac{e^{\alpha x}e^{-e^{x}/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, \tag{3}$$

kde $-\infty < x < \infty$.

Pre náhodnú premennú X riadiacu sa log-gama rozdelením(2), a náhodnú premennú Y, ktorá má gama rozdelenie platí nasledujúci vzťah:

$$X{\sim}LG_2(\alpha,\beta) => Y = e^{x}{\sim}G(\alpha,\beta).$$

Na dokázanie tohto vzťahu sme použili $Vetu\ 1$. V našom prípade chceme ukázať, že funkcia $h(y)$ vyjadruje hustotu pravdepodobnosti gama rozdelenia náhodnej premennej Y. Funkcia $f(x)$ vyjadruje hustotu pravdepodobnosti log-gama rozdelenia(2) náhodnej premennej X. Pre náhodné premenné Y a X platia vzťahy $Y = g(x) = e^{X}$ a $X = g^{-1}(y) = \ln Y$. Potom podľa $Vety\ 1$ sme vyjadrili funkciu hustoty pravdepodobnosti $h(y)$:

$$h(y) = f(\ln y)\left|\frac{d}{dy}(\ln y)\right| = \frac{e^{\alpha\ln y}e^{-e^{\ln y}/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\frac{1}{y} = \frac{\left(e^{\ln y}\right)^{\alpha}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\frac{1}{y} = \frac{y^{\alpha}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\frac{1}{y} = \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)},$$

a tým sme dokázali, že ak náhodná premenná X má log-gama rozdelenie(2), tak potom náhodná premenná $Y = e^{x}$ sa riadi gama rozdelením. Modus log-gama rozdelenia(2) je $M = \ln(\alpha\beta)$.

2.5 Zovšeobecnené log-gama rozdelenie

Zovšeobecnené log-gama (*generalized log-gamma*-GLG) rozdelenie dostaneme nasledovnou transformáciou náhodnej premennej X:

$$X \sim LG_2(\alpha, \beta) => Y = (mX + u) \sim GLG(\alpha, \beta, u, m).$$

Podľa $Vety\ 1$ platí $Y = mX + u = g(x), X = \frac{(Y-u)}{m} = g^{-1}(y)$ a $\frac{d}{dy}\left(g^{-1}(y)\right) = 1/m$. Potom platí nasledovný vzťah:

$$h(y) = f\left(g^{-1}(y)\right)\left|\frac{d}{dy}\left(g^{-1}(y)\right)\right| = f\left((Y-u)/m\right)\left|\frac{1}{m}\right| =$$

$$= \frac{exp\left\{\alpha\left(\frac{y-u}{m}\right) - e^{\left(\frac{y-u}{m}\right)}/\beta\right\}}{\beta^{\alpha}\Gamma(\alpha)}\left|\frac{1}{m}\right|,$$

kde výsledná funkcia hustoty pravdepodobnosti predstavuje GLG rozdelenie. Ak $m = 1$ a $u = 0$, tak dostaneme funkciu hustoty log-gama rozdelenia(2). Modus GLG rozdelenia s parametrami α, β, m a u je $M = u + m\ ln(\alpha\beta)$.

Ak vo funkcii hustoty pravdepodobnosti GLG rozdelenia preznačíme parametre nasledovným spôsobom, že $m = \frac{b}{\lambda}$, $\alpha = \lambda^{-2}$ a $\beta = \lambda^2$, tak dostaneme:

$$h(y) = \frac{exp\left\{\lambda^{-2}\left(\frac{\lambda(y-u)}{b}\right) - \frac{e^{\left(\frac{\lambda(y-u)}{b}\right)}}{\lambda^2}\right\}}{\lambda^{2^{(\lambda^{-2})}}\Gamma(\lambda^{-2})}\left|\frac{\lambda}{b}\right| = \tag{4}$$

$$= \frac{|\lambda|(\lambda^{-2})^{\lambda^{-2}} exp\left\{\lambda^{-1}\left(\frac{y-u}{b}\right) - \lambda^{-2} exp\left[\lambda\left(\frac{y-u}{b}\right)\right]\right\}}{b\Gamma(\lambda^{-2})},$$

čo zodpovedá funkcii hustoty pravdepodobnosti špeciálneho prípadu GLG rozdelenia z práce (Liang, 2000).

2.6 Coale-McNeilovo rozdelenie

V (Coale, McNeil, 1972) je uvedená všeobecná forma funkcie hustoty pravdepodobnosti pre Coale-McNeilovo (CM) rozdelenie a je definovaná nasledovne:

$$g(x) = \frac{\beta}{\Gamma(\alpha/\beta)} exp[-\alpha(x-\mu) - exp\{-\beta(x-\mu)\}], \tag{5}$$

kde $\alpha > 0, \beta > 0,\ \mu(-\infty < \mu < \infty)$ sú tri parametre tohto rozdelenia.

Podľa práce (Kaneko, 2003) si teraz uvedieme vzťah medzi CM rozdelením a špeciálnym prípadom GLG rozdelenia (4). Nasledujúca alternatívna parametrizácia CM rozdelenia umožňuje zastúpenie plného rozsahu parametrického priestoru špeciálneho prípadu GLG rozdelenia:

$$g(x) = \frac{|\beta|}{\Gamma(k)} exp[-k\beta(x-\mu) - exp\{-\beta(x-\mu)\}],$$

kde $k > 0$ je nový parameter a korešponduje s $\alpha/|\beta|$ z rovnosti (5), čiže parameter β teraz môže nadobúdať aj záporné hodnoty. Ak pôvodné CM rozdelenie zodpovedá špeciálnemu prípadu GLG rozdelenia so zápornou hodnotou λ a ak zvažujeme iba situácie, v ktorých λ nadobúda zápornú hodnotu, môžeme považovať špeciálny prípad GLG rozdelenia ako ekvivalentný ku CM rozdeleniu. Súlad medzi parametrami CM a GLG rozdelenia môže byť vyjadrený nasledovne:

$$\alpha = -\frac{1}{b\lambda}, \beta = -\frac{\lambda}{b}, \mu = u - \frac{b}{\lambda} ln\lambda^2.$$

Alebo:

$$\lambda = -\left(\frac{\alpha}{\beta}\right)^{-\frac{1}{2}}, b = (\alpha\beta)^{-\frac{1}{2}}, u = \mu - \frac{1}{\beta} ln\left(\frac{\alpha}{\beta}\right).$$

3 Coale-McNeilov model

Nasledovné je upravená verzia všeobecného štandardného modelu, kde rozdelenie veku prvého sobáša má strednú hodnotu nula a štandardnú odchýlku jedna, prevzatého z literatúry (Rodriguez, Trussel, 1980) a získaného nastavením parametrov $\alpha = 1{,}145, \beta = 1{,}896$ a $\mu = -0{,}805$ do rovnice (5):

$$g(x) = 1{,}2813 exp[-1{,}145(x + 0.805) - exp\{-1{,}896(x + 0{,}805)\}]. \quad (6)$$

Nech $g(x; u, b)$ popisuje rozdelenie veku prvého sobáša akéhokoľvek ženskej kohorty s jeho pozorovanou strednou hodnotou u a štandardnou odchýlkou b. S použitím CM rozdelenia z rovnosti (5) dostaneme normovanú funkciu prvosobášnosti:

$$g(x; u, b) \cong \frac{1}{b} g_s\left(\frac{x-u}{b}\right).$$

Sobášnu schému $f^{CM}(x)$, ktorá stelesňuje hustoty pravdepodobnosti rozdelenia veku prvého sobáša počas veku x pre všetkých členov kohorty, dostaneme vynásobením hustoty pravdepodobnosti tohto rozdelenia parametrom P, ktorý predstavuje v tomto prípade úhrnnú mieru $TMR1m$ alebo $TMR1f$. Teda je definovaná rovnosťou:

$$f^{CM}(x) = Pg(x; u, b) \cong P\frac{1}{b} g_s\left(\frac{x-u}{b}\right).$$

Výsledný CM model prvosobášnosti, ktorým sme fitovali špecifické miery vybraných demografických procesov teda vyzerá nasledovne:

$$f^{CM}(x) = P\frac{1}{b}\frac{\beta}{\Gamma\left(\frac{\alpha}{\beta}\right)}exp\left[-\alpha\left(\left(\frac{x-\mathrm{u}}{b}\right)-\mu\right)\right.$$
$$\left.-exp\left\{-\beta\left(\left(\frac{x-\mathrm{u}}{b}\right)-\mu\right)\right\}\right]. \quad (7)$$

4 Odhad parametrov CM modelu

Pre výpočet momentov empirických dát sme používali nasledovné charakteristiky uvedené v (Pastor, 2013). Prvý počiatočný moment vypočítame nasledovne:

$$m_1 = \frac{\sum_x xf(x)}{P}.$$

Podobne druhý počiatočný moment:

$$m_2 = \frac{\sum_x x^2 f(x)}{P}.$$

Bodovým odhadom pre strednú hodnotu náhodnej premennej X je aritmetický priemer $\bar{x}$:

$$\bar{x} = m_1 = \frac{\sum_x xf(x)}{P},$$

ktorý predstavuje priemerný vek muža alebo ženy pri danom demografickom procese a pre disperziu je bodovým odhadom výberový rozptyl s^2:

$$s^2 = m_2 - m_1^2 = \frac{\sum_x x^2 f(x)}{P} - \left(\frac{\sum_x xf(x)}{P}\right)^2.$$

Štandardná odchýlka s je odmocnina z výberového rozptylu.

Pri CM modeli sme normovali rozdelenie veku jednotlivých demografických procesov $g(x; u, b)$, kde u predstavuje hodnotu $\bar{x}$ a b predstavuje hodnotu s podľa (7), kde $(x - \bar{x})/s$ môže nadobúdať aj záporné hodnoty. Počiatočné hodnoty parametrov modelu podľa (Rodriguez, Trussell, 1980) boli nastavené nasledovne: $\alpha = 1{,}145, \beta = 1{,}896$ a $\mu = -0{,}805$.

Aplikovaním týchto charakteristík dostaneme prvotné fitovania špecifických mier. Na zistenie kvality modelu sme zvolili metódu súčet štvorcov odchýlok (označujeme SSE), ktorá bude slúžiť taktiež ako porovnávacie kritérium a je definovaná nasledovne:

$$SSE = \sum_x \left(f(x) - f^M(x)\right)^2,$$

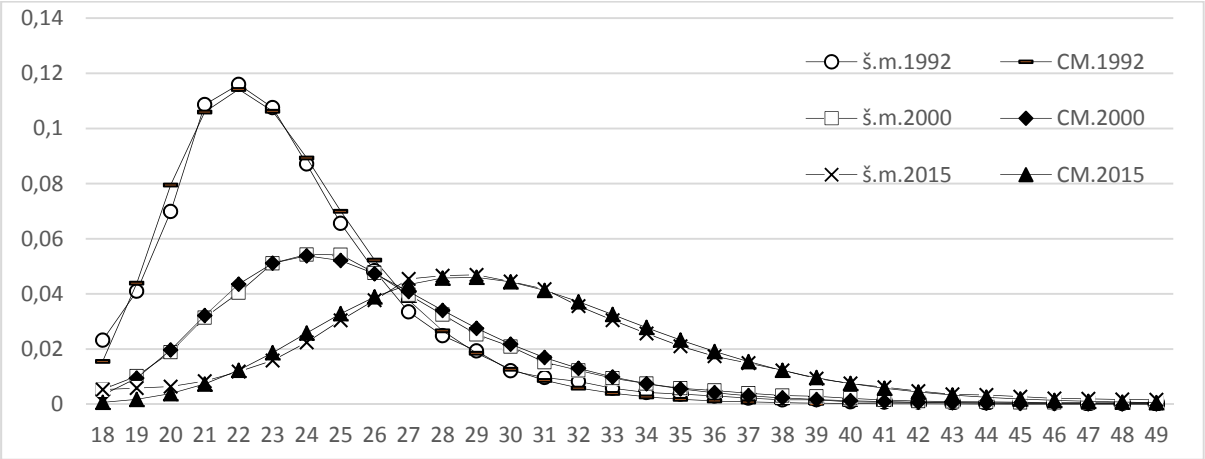
kde $f(x)$ sú empirické špecifické miery a $f^M(x)$ sú modelované špecifické miery. Našim cieľom je nájsť také hodnoty α, β a μ, ktoré minimalizujú SSE.

Najmenšiu hodnotu SSE a najlepšie fitovanie špecifických mier sme dosiahli za pomoci excelovského nástroju $SOLVER$.

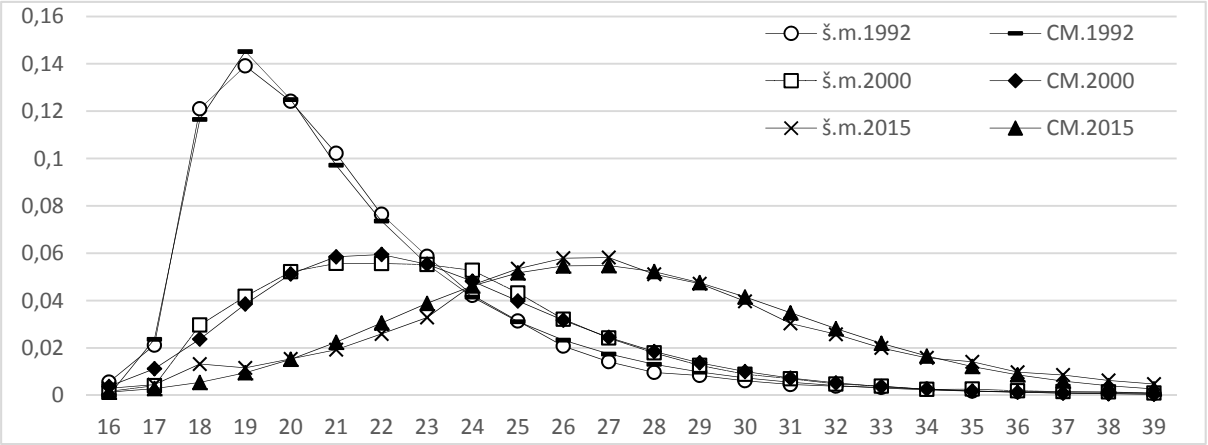
5 Použitie CM modelu na dáta SR a porovnanie s inými modelmi

V práci sme využívali údaje pre Slovenskú republiku, ktoré spracúva a publikuje Štatistický úrad Slovenskej republiky (označujeme ŠÚ SR). Empirické tabuľky pre nami skúmané demografické procesy sú zverejnené aj Inštitútom informatiky a štatistiky (Infostat, online), ktorý nadväzuje na spracovania ŠÚ SR.

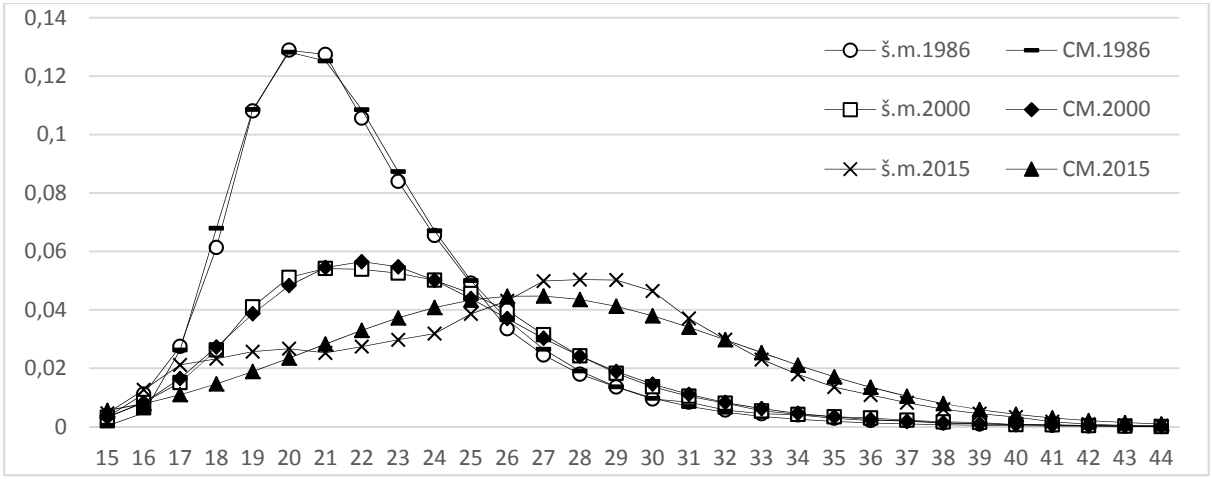
Výsledky fitovania pre vybrané roky sú ilustrované na grafoch, kde vidieť, že CM model s úspechom možno použiť nielen na modelovanie špecifických mier prvosobášnosti, ale aj ostatných demografických procesov. Podrobnosti a viac výsledkov možno nájsť v práci (Hladký, 2017).



Graf 1 Vývoj špecifických mier prvosobášnosti mužov v jednotlivých rokoch a vylepšené odhady CM modelu prvosobášnosti mužov (*Zdroj: vlastné spracovanie*)

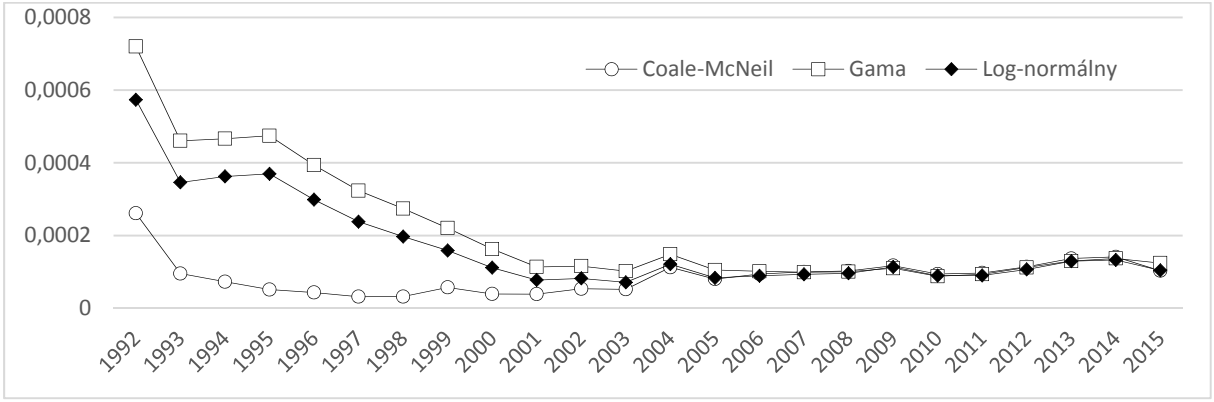


Graf 2 Vývoj špecifických mier prvosobášnosti žien v jednotlivých rokoch a vylepšené odhady CM modelu prvosobášnosti žien *(Zdroj: vlastné spracovanie)*

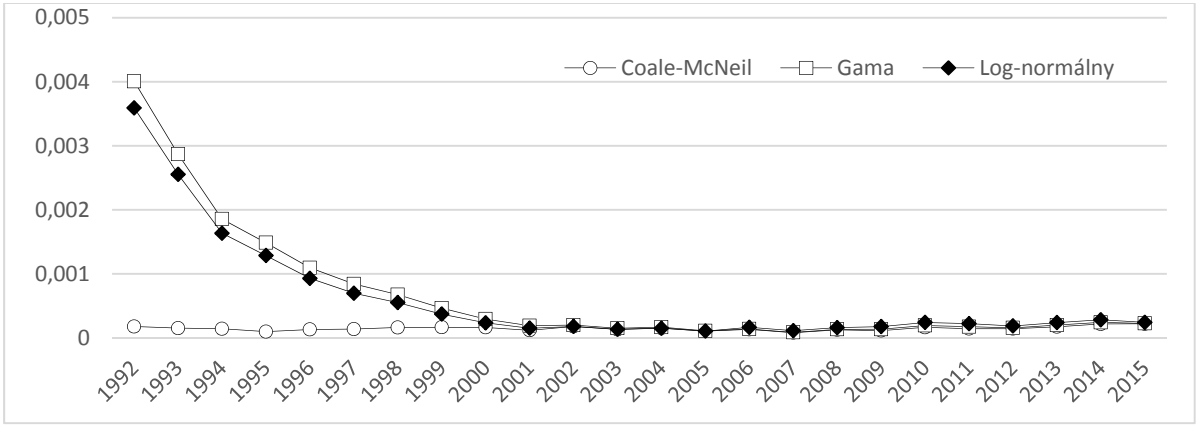


Graf 3 Vývoj špecifických mier prvoplodnosti v jednotlivých rokoch a vylepšené odhady CM modelu prvoplodnosti *(Zdroj: vlastné spracovanie)*

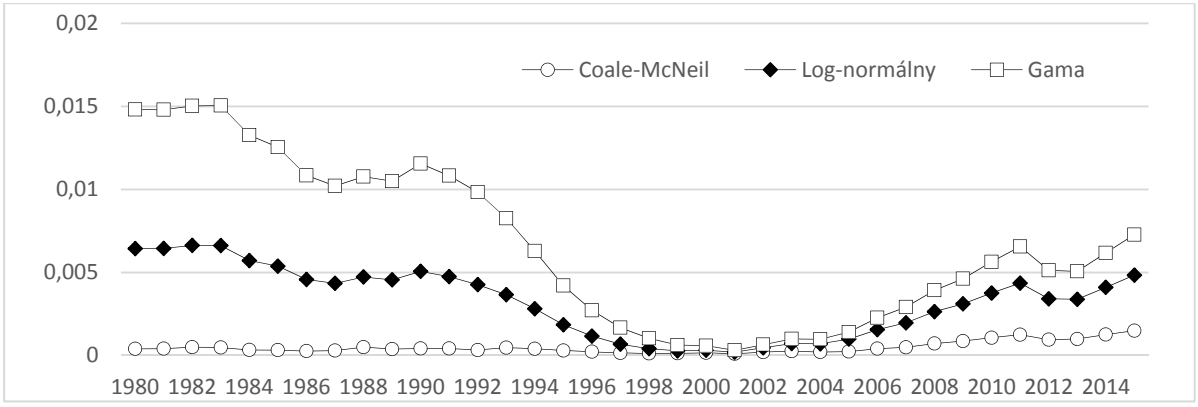
CM model sledovaných demografických procesov sme porovnávali s gama a log-normálnym modelom. V nasledovných grafoch sú zobrazené výsledky vývoja SSE gama, log-normálneho a CM modelu pre jednotlivé demografické procesy, kde CM model sa javí ako najlepší.



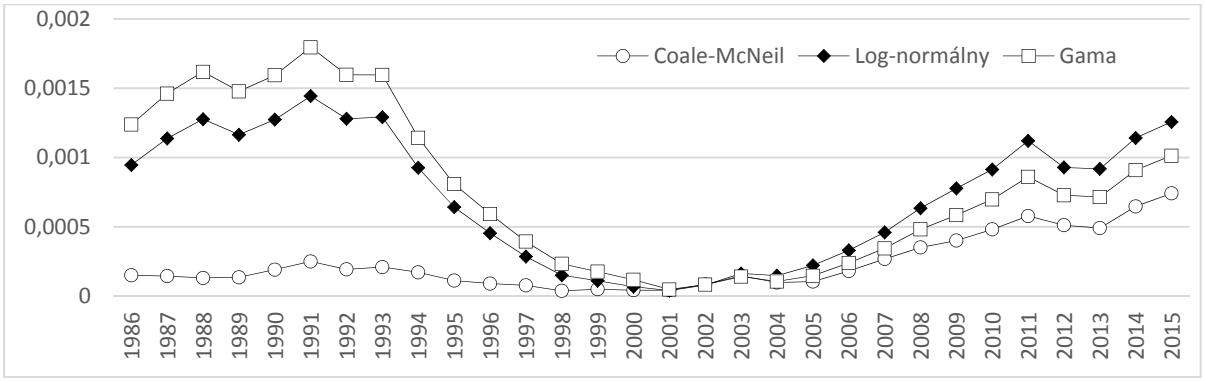
Graf 4 Vývoj SSE jednotlivých modelov prvosobášnosti mužov v jednotlivých rokoch (*Zdroj: vlastné spracovanie*)



Graf 5 Vývoj SSE jednotlivých modelov prvosobášnosti žien v jednotlivých rokoch (*Zdroj: vlastné spracovanie*)



Graf 6 Vývoj SSE jednotlivých modelov plodnosti v jednotlivých rokoch
(*Zdroj: vlastné spracovanie*)



Graf 7 Vývoj SSE jednotlivých modelov prvoplodnosti v jednotlivých rokoch
(*Zdroj: vlastné spracovanie*)

6 Záver

V práci sú uvedené rôzne typy log-gama rozdelení a vzťahy medzi nimi. Najviac nás zaujímal vzťah s CM rozdelením, ktoré bolo použité pri modelovaní špecifických mier sobášnosti, prvosobášnosti, plodnosti a prvoplodnosti. Ako vidno pomocou CM modelu prvosobášnosti založeného na CM rozdelení sa dajú modelovať nielen špecifické miery prvosobášnosti, ale aj špecifické miery ostatných demografických procesov. Podľa grafického znázornenia vývoja súčtu štvorcov odchýlok je tento model oproti gama a log-normálnemu modelu lepší a drží si konštantnú hodnotu počas celého sledovaného obdobia.

7 Literatúra

Coale, A.J., McNeil, D.R. (1972). *The Distribution by Age of the Frequency of First Marriage in Female Cohort*. Journal of American Statistical Association 67: 743-749.

Hladký, D. (2017). *Loggama rozdelenie v demografickej štatistike.* Bakalárska práca, FMFI UK Bratislava.

Hogg, R.V., Klugman, S. A. (2008). *Loss Distributions*. Wiley Series in Probability and Mathematical Statistics, Department of Statistics and Actuarial Science, The University of Iowa.

Infostat, (online). *Rozšírené tabuľky základných údajov SR. [Dostupné na:* http://www.infostat.sk/vdc/sk/index.php?option=com_content&view=category&layout=blog&id=13&Itemid=58]

Kaneko, R. (2003*). Elaboration of the Coale-McNeil Nuptiality Model as the Generalized Log Gamma Distribution: A New Identity and Empirical Enhancements*. Demographic research, Volume 9, Article 10, 223-262. [Dostupné na: www.demographic-research.org/Volumes/Vol9/10.]

Lamoš, F., Potocký, R. (1998). *Pravdepodobnosť a matematická štatistika*. 2.vydanie, UK Bratislava.

Liang, Z. (2000). *The Coale-McNeil model*. Theory, generalisation and application. Amsterdam, The Netherlands: Thela Thesis.

Pastor, K. (2013). *Matematické modely špecifických mier plodnosti*. Forum statisticum Slovacum IX, č.1, 71-75.

Rodriguez, G., Trussel, J. (1980). *Maximum Likelihood Estimation of the Parameters of Coale's Model Nuptiality Schedule from Survey Data*. World Fertility Survey, Technical Bulletins, No.7.

WM, (online)**.** *Univariate Distribution Relationships. [Dostupné na:*

http://www.math.wm.edu/~leemis/chart/UDR/UDR.html]

Identifikace statistických distribucí v Preisach-Mayergoyzově prostoru hysteronů jako úloha optimalizace
Identifications of statistical distributions in Preisach-Mayergoyz hysteron space as optimization task

Colette Kožená
Fakulta jaderná a fyzikálně inženýrská, České vysoké učení technické v Praze
Faculty of Nuclear Sciences and Physical Engineering, Czech technical university in Prague
kozencol@fjfi.cvut.cz

Abstrakt: *Za účelem detekce struktury a mechanických vlastností hysterezních materiálů je aplikován Preisach-Mayegoyzův model (PM model). Je představen identifikační proces nalezení distribuční směsi rozdělení hysteronů v PM prostoru s využitím optimalizačních algoritmů, implementovaných do programu s pracovním názvem PM identifikátor. Zaměřili jsme se na stochastický algoritmus Simulovaného žíhání (SA), jeho rychlejší variantu (FSA) a také rozšíření těchto metod o algoritmus slepého náhodného prohledávání za účelem nalezení počátečního řešení. Byla též navržena a testována modifikace metody SA. Implementované algoritmy, založené na Φ -divergencích či L2 normě, byly testovány vzhledem k časové náročnosti a přesnosti nalezeného řešení.*

Abstract: *Preisach-Mayergoyz model (PM model) of hysteresis can be used to structure detection and evaluate the mechanical properties of hysteresis materials. Identification of the PM space probability density within a class of distribution mixtures by means of numerical optimization techniques with L2 distance and Φ -divergence measures is presented. We focus on Simulated Annealing (SA), its alternative Fast Simulated Annealing (FSA) and these methods in combination with an algorithm of blind random search to find an initial solution. An extension of SA was proposed and all of these implemented methods were tested with respect to time and accuracy of the solution found.*

Kľúčové slová: *Preisach-Mayergoyzův model hystereze, Φ-divergence, optimalizace, simulované žíhání, distribuční směsi*

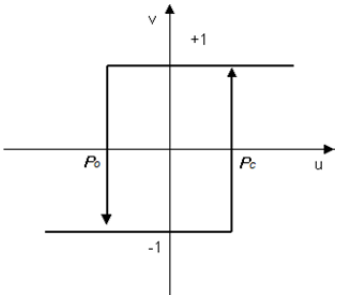
Key words: *Preisach-Mayergoyz model of hysteresis, Φ-divergences, optimization, Simulated Annealing, distribution mixtures*

1 Úvod

V tomto článku je představen Preisach-Mayergoyzův (PM) model jako vhodný nástroj k popisu mechanických vlastností materiálů, který vykazuje hysterezní chování. My se zabýváme identifikací tzv. PM prostoru daného materiálu a následně se snažíme navrhnout vyhodnocení indexu elasticity či indexu poškození materiálu. Jde o nedestruktivní a výpočetně méně náročnou metodu, což představuje značnou výhodu.

1 Preisach-Mayergoyzův model

Preisach-Mayergoyzův (Mayergoyz, 2003) model je založen na myšlence, že každý materiál je tvořen velkým množstvím malých jednotek – tzv. hysteronů. Hysteron je charakterizován otevírací hodnotou tlaku P_o a zavírací hodnotou tlaku P_c (viz graf na obrázku 1).



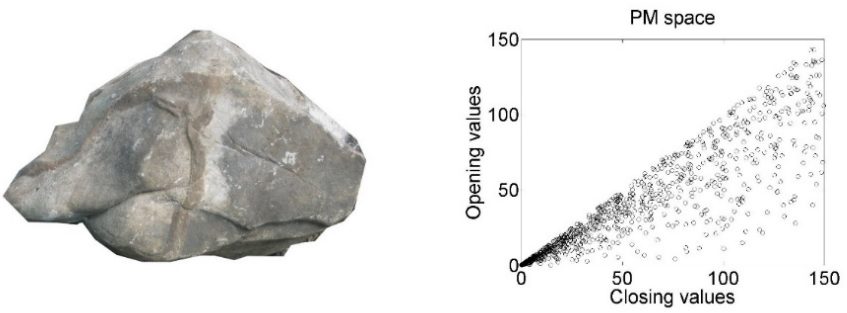
Obr. 1 Hysteron může být ve stavu zavřeném (značíme hodnotou +1) či ve stavu otevřeném (značíme hodnotou -1). (*Zdroj: vlastní zpracování*)

Hysterony se zakreslují jako body (P_c, P_o) do pravoúhlého trojúhelníku, kde $P_o \leq P_c$. Preisachův trojúhelník představuje strukturu daného materiálu (viz obr. 2). Čím vyšší je hodnota zavíracího tlaku oproti tlaku otevíracímu, tím méně je hysteron elastický a tím více vykazuje hysterezní chování. Elastické jednotky se tedy nacházejí na přeponě Preisachova trojúhelníku.

Rozložení hysteronů v PM prostoru lze popsat hustotou rozdělení. Cílem je tedy nalezení vhodné metody k identifikaci této hustoty, jež může být hledána i ve tvaru distribuční směsi o M komponentách a váhách $\lambda_i > 0$

$$p(x) = \sum_{i=1}^{M} \lambda_i \, p_i(x), \qquad \sum_{i=1}^{M} \lambda_i = 1, \tag{1}$$

kde $p_i(x)$ jsou hustoty pravděpodobnosti. Distribuční směs (McLachlan, Peel, 2006) by mohla mnohem lépe vystihovat reálnou strukturu materiálu nebo lépe popsat složitější ekonomické či biologické jevy s hysterezním chováním. Tento předpoklad byl již potvrzen zpracováním experimentálně naměřených dat v dřívějším výzkumu.



Obr. 2 Příklad hysterezního materiálu a korespondujícího PM prostoru hysteronů (*Zdroj: vlastní zpracování*)

Přistoupíme k matematickému popisu hysteronů, nebo-li Preisachova operátoru $\hat{\gamma}_{P_c,P_o}$ hysteronu

$$\hat{\gamma}_{P_c,P_o}\big(u(t)\big) = \begin{cases} -1, & u(t) \le P_o, \\ +1, & u(t) \ge P_c, \\ k, & u(t) \in (P_o, P_c), \end{cases} \tag{2}$$

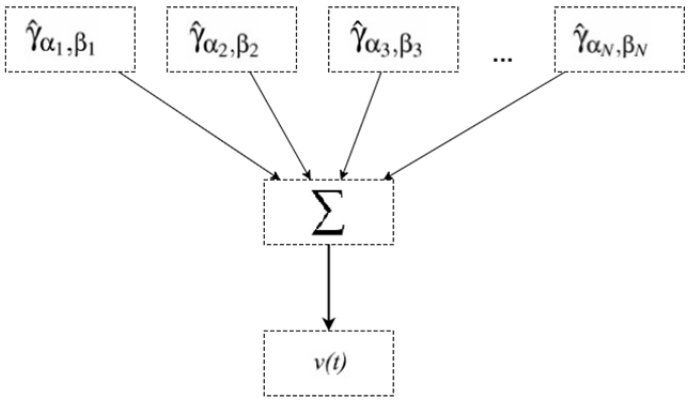
kde $u(t)$ je vstupní signál a

$$k = \begin{cases} +1, & \exists t^*: u(t^*) > P_c \text{ and } \forall \tau \in (t^*, t)\ u(\tau) \in (P_o, P_c), \\ -1, & \exists t^*: u(t^*) < P_c \text{ and } \forall \tau \in (t^*, t)\ u(\tau) \in (P_o, P_c). \end{cases} \tag{3}$$

Nyní již můžeme vyjádřit výstupní signál $v(t)$ po průchodu vstupního signálu $u(t)$ daným materiálem

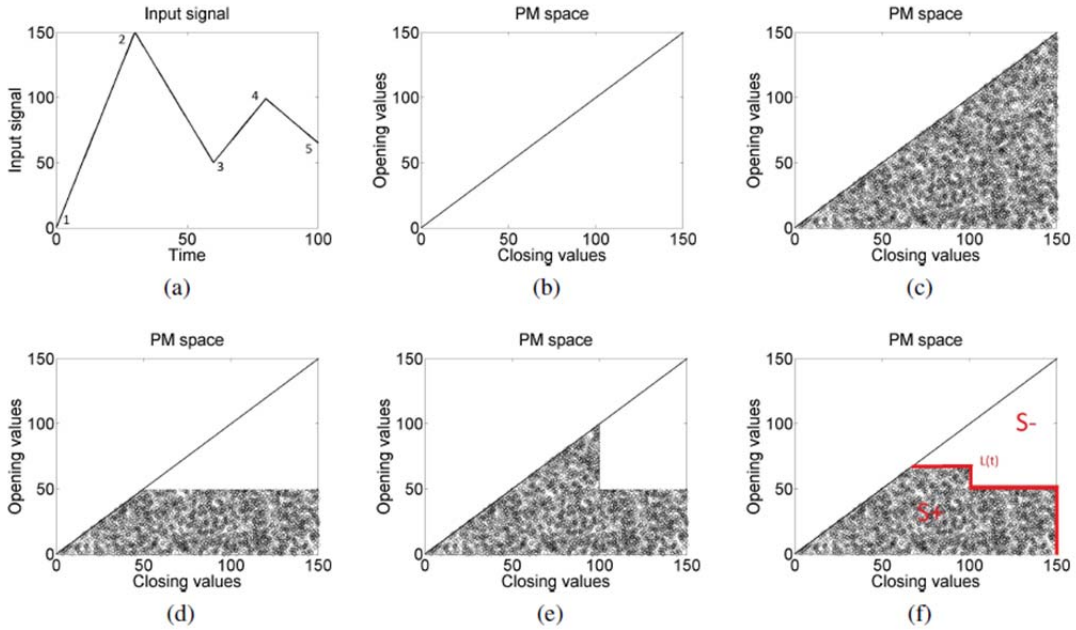
$$y(t) = \iint_{P_o \le P_c} \mu(P_c, P_o)\ \hat{\gamma}_{P_c,P_o}\big(u(t)\big)\ dP_c\, dP_o, \tag{4}$$

kde $\mu(P_c, P_o)$ je hustota pravděpodobnosti. V diskrétním případě bude pak Preisachův model definován jako součet všech hysteronů v závislosti na jejich hustotě výskytu $\mu(P_c, P_o)$ – viz grafická interpretace diskrétního Preisachova modelu na obrázku 3.



Obr. 3 Grafická interpretace diskrétního Preisachova modelu (*Zdroj: vlastní zpracování*)

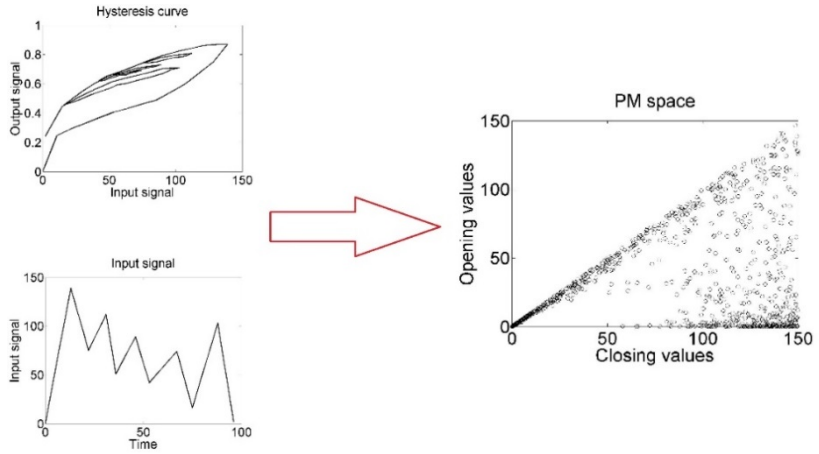
Model lze přiblížit jeho geometrickou interpretací na obr. 4. Graf (4a) značí průběh vstupního zatížení (tlaku). Obrázek (4b) představuje PM prostor před průběhem vstupního zatížení, tedy v bodě 1 na grafu (4a). Všechny hysterony jsou nyní ve stavu -1, tedy ve stavu otevřeném (indikováno bílou barvou bodů (P_c, P_o) v PM prostoru). Následně roste tlak až na hodnotu označenou číslem 2 na podgrafu (4a). Hysterony, jejichž zavírací hodnota P_c je menší než aktuální hodnota tlaku, mění svůj stav na zavřený (hysteronu jsou označeny černou barvou) - viz obr. (4c). Dále vstupní signál klesá (do hodnoty 3 na obr (4a)) a všechny hysterony, jež jsou ve stavu zavřeném a jejichž otevírací hodnota P_o je větší než aktuální hodnota vstupu, se vrátí zpět do stavu otevřeného (obr. (4d)). Obdobná situace bude při dalším nárůstu a poklesu tlaku. Na obr. (4f) je pak finální PM prostor. Rozložení hysteronů pro tuto ilustraci bylo zvoleno uniformní.



Obr. 4 Geometrická interpretace Preisach-Mayergoyzova modelu (*Zdroj: vlastní zpracování*)

2 Úloha optimalizace

Hlavním cílem je tedy nalezení vhodné parametrické distribuce, resp. distribuční směsi rozdělení (1), tedy příslušných parametrů rozdělení pouze na základě znalosti vstupního zatížení a hysterezní křivky daného materiálu (obr. 5).

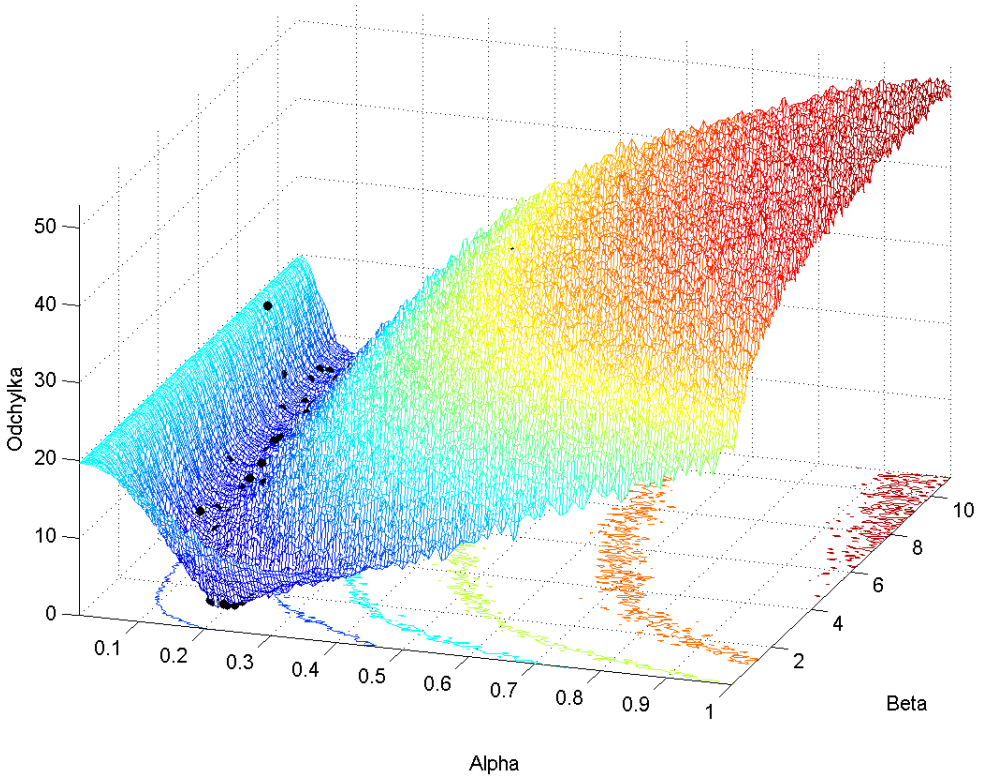


Obr. 5 Identifikace PM prostoru na základě znalosti vstupního zatížení a hysterezní křivky (*Zdroj: vlastní zpracování*)

Za účelem identifikace rozložení hysteronů v PM prostoru byl v prostředí Matlab vytvořen program s názvem PM identifikátor.

Na identifikaci lze pohlížet jako na optimalizační úlohu, kde funkcionál $\mathcal{F}$, který se snažíme minimalizovat, představuje v našem případě odchylku naměřené hysterezní křivky a hysterezní křivky vygenerované v každém kroku iteračního algoritmu. Příklad vygenerované účelové funkce lze vidět na grafu 6,

kde je také vyobrazen postup jednoho z implementovaných algoritmů pro nalezení přibližného globálního minima – algoritmu Simulovaného žíhání (Kirkpatrick et al., 2008).



Obr. 6 Testovací příklad postupu algoritmu SA pro volbu rozdělení Guyer s parametry α=0.2 a β=1. Rozdělení pochází z článků R.A. Guyera a K.R. McCalla, kteří se zabývali převážně hysterezí skal. (*Zdroj: vlastní zpracování*)

Pro vyjádření odchylky hysterezních křivek je použita klasická L2 norma, Kolmogorova vzdálenost nebo robustnější Φ-divergence, zejména Hellingerova a LeCamova divergence. Více informací týkajících se teorie Φ-divergencí lze nalézt v (Kůs et al., 2008).

3 Stochastické optimalizační algoritmy

Do programu byly implementovány stochastické algoritmy, zejména klasické simulované žíhání (SA), rychlejší varianta FSA (Metropolis et al., 1953), později také genetický algoritmus (GA) a byla také navržena modifikace SA, kterou označíme jako SA*. Dále byla do programu přidána možnost přiblížení počátečního řešení samotných algoritmů pomocí slepého náhodného prohledávání, avšak tato možnost se při testování jevila spíše negativně na rychlost algoritmů i přesnost nalezeného řešení.

Algorithm 1 Simple Simulated Annealing

1: initial temperature $T = T_{\max}$
2: **while** $i < K$ **do** ▷ K is maximal number of steps
3: initial solution x_0, $i \leftarrow 0$
4: evaluation of the cost function $t_0 = \mathcal{F}(x_0)$
5: new solution $x_1 = \Pi(x_0)$ ▷ Π is called change operator
6: evaluation $t_1 = \mathcal{F}(x_1)$
▷ Metropolis criterion for accepting the new state
7: **if** $\mathcal{F}(x_0) \geq \mathcal{F}(x_1)$ **then**
8: accept the new state x_1, $x_0 \leftarrow x_1$, $t_0 \leftarrow t_1$, $i \leftarrow i + 1$
9: repeat the procedure from step 5
10: **else**
11: **if** rand $(1, 1) \leq p = \exp\left(\frac{-(\mathcal{F}(x_1) - \mathcal{F}(x_0))}{T}\right)$ **then**
12: accept the new state x_1, $x_0 \leftarrow x_1$, $t_0 \leftarrow t_1$, $i \leftarrow i + 1$
13: repeat the procedure from step 5
14: **else**
15: accept the original state x_0, $i \leftarrow i + 1$
16: repeat the procedure from step 5
17: **end if**
18: **end if**
19:
20: **end while**
21: temperature change $T_{\text{new}} = \text{koef} \cdot T$
22: repeat the procedure from step 2

Obr. 7 Pseudokód: implementovaný algoritmus SA (*Zdroj: vlastní zpracování*)

Modifikace klasického SA spočívá v přidání několika dalších parametrů z důvodu statistické povahy naší úlohy. Jde o parametr L, který určuje opakování algoritmu SA, ze kterých pak vybíráme řešení s nejnižší hodnotou hysterezních křivek. Vzhledem k tomu, že vygenerování prostoru hysteronů dle určitého rozdělení s danými parametry je stochastickou záležitostí, je přidán parametr M, tedy počet přegenerování PM prostoru a určení průměrné odchylky hysterezních křivek. Parametr označený jako TT značí počet iterací samotného SA. Pokud bychom zvolili hodnoty parametrů $M = 1$ a $L = 1$, dostaneme klasický algoritmus SA.

Jednotlivé parametry modifikovaného algoritmu SA* jsme testovali a byla zjištěna statistická významnost parametrů L, M, TT a jejich interakcí. Boxploty pro jednotlivé parametry lze vidět na obrázku 9. Pro testování byly zvoleny 3 sady vstupních dat a k výpočtu odchylky hysterezních křivek byla zvolena Hellingerova divergence. Součástí analýzy bylo také testování závislosti algoritmu na vstupních datech (označeno jako parametr R na obr. 9). Nebyla prokázána signifikantnost tohoto parametru na hladině významnosti 5 % a algoritmus se tedy zdá být vhodný k našemu účelu identifikace.

U parametrů T_{max} a $koef$ také nebyla prokázána signifikantnost na 5% hladině významnosti. Výrazný vliv na velikost odchylky hysterezních křivek

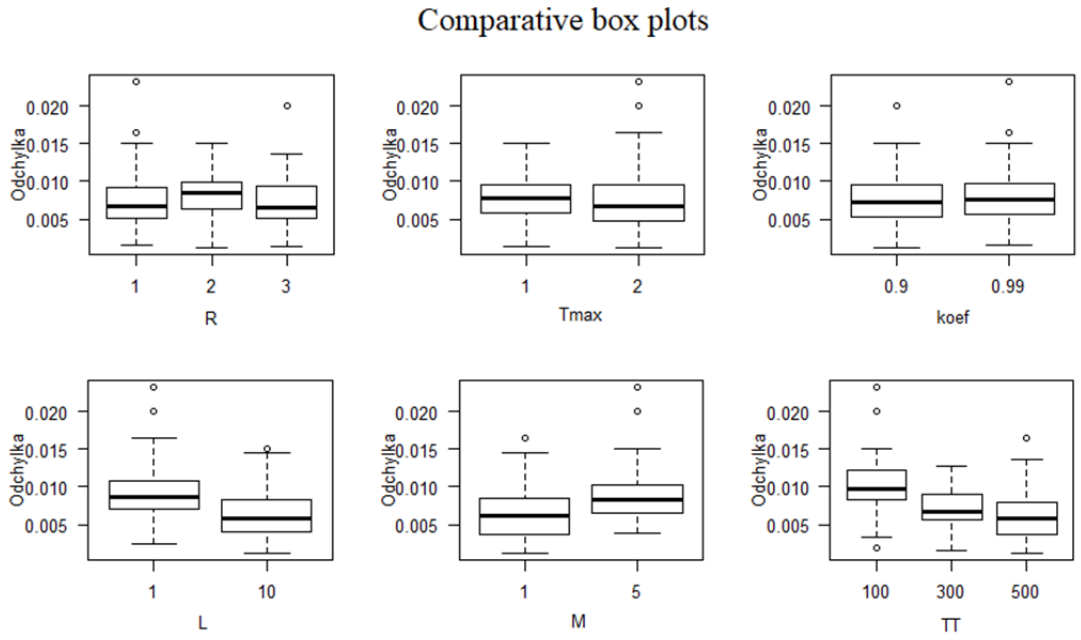
můžeme pozorovat u parametru M, avšak v negativním smyslu. Zvýšíme-li počet přegenerování PM prostoru, dochází k výraznému nárůstu odchylky. Naopak zvýšení parametru L má výrazný vliv na snížení odchylky, stejně tak jako u parametru TT. Se zvyšováním těchto parametrů však roste také čas výpočtu. Po podrobnější analýze příslušných contourplotů jsme došli k závěru, že nejvhodnějšími volbami jsou $M = 1$, $L \in \langle 10, 15 \rangle$ a $TT \in \langle 500, 1\,000 \rangle$.

Pseudokód modifikovaného algoritmu SA* je následující:

Algorithm 2 Modification of Simulated Annealing SA*

1: **for** $l = 1 : L$ **do** ▷ Number of SA repetitions
2: initial solution and evaluation of the cost function - functional $\mathcal{F}$:
3: **for** $m = 1 : M$ **do**
4: generation of PM space ▷ We get M deviations L_m
5: **end for**
6: compute the average of M hysteresis deviations $\overline{L_1} = \sum_{m=1}^{M} L_m$
7: **for** $t = 1 : TT$ **do** ▷ TT iterations in SA algorithm
8: new state (newly generated parameters of distribution)
9: cost function evaluation:
10: **for** $m = 1 : M$ **do**
11: generation of PM space ▷ We get M deviations L_m
12: **end for**
13: the average of M deviations calculation
14: **end for**
▷ We get TT possible states with the corresponding deviations
15: select the lowest deviation from $(\overline{L_1}, \overline{L_2}, \ldots, \overline{L_{TT}})$ as $L_{\min}^{(l)} = \min_{k \in \{1,\ldots,TT\}} \overline{L_k}$
16: **end for**
17: select the lowest deviation from $(L_{\min}^{(1)}, L_{\min}^{(2)}, \ldots, L_{\min}^{(L)})$ and the corresponding state

Obr. 8 Pseudokód: modifikovaný algoritmus SA* (*Zdroj: vlastní zpracování*)



Obr. 9 Boxploty pro všechny testované parametry SA* (*Zdroj: vlastní zpracování*)

Modifikace SA* se jeví jako vhodný nástroj k identifikaci hysteronů v PM prostoru, avšak je výrazně časově náročnější než klasické SA, FSA či SA v kombinaci se slepým náhodným prohledáváním – viz tabulka 1. Rozložení hysteronů v PM prostoru pocházela pro toto testování z rozložení typu Guyer (McCall, Guyer, 1996), (McCall, Guyer, 1995) s předem zvolenými parametry α=2 a β=0.5. Rozdělení typu Guyer je definováno následovně

$$P_c = \max r_c^{\alpha}, \quad P_o = P_c\, r_o^{\beta}, \quad \alpha, \beta \in \mathbb{R}_0^+, \tag{5}$$

kde ‚max' je maximální hodnota vstupního zatížení, r_c a r_o náhodně rovnoměrně generovaná čísla z intervalu $(0, 1)$ a α, β představují parametry tohoto rozdělení.

Pro SA* byly zvoleny následující hodnoty parametrů $M = 1$, $L = 10$ a $TT = 500$. Dále hodnota koeficientu ochlazování byla nastavena na hodnotu $koef = 0.99$ a počáteční teplota $T_{max} = 1$ pro všechny typy zkoumaných algoritmů.

V tabulce 1 lze vidět průměrné hodnoty a směrodatné odchylky obou hledaných parametrů, Hellingerovy divergence a celkového času pro 150 opakování. Jak je kromě výrazné časové náročnosti algoritmu SA* vidět, tento algoritmus dosahoval znatelně nižších odchylek hysterezních křivek v porovnání s ostatními sledovanými algoritmy.

SA	α	β	Hellinger	celkový čas [s]
Průměr	1.97	0.50	0.030771	5.84
σ	0.2765	0.2333	0.021193	6.3776

FSA	α	β	Hellinger	celkový čas [s]
Průměr	1.97	0.52	0.037882	3.68
σ	0.2659	0.4152	0.086383	3.7522

BLIND + SA	α	β	Hellinger	celkový čas [s]
Průměr	1.98	0.45	0.007092	12.81
σ	0.1263	0.0666	0.002109	23.1072

SA*	α	β	Hellinger	celkový čas [s]
Průměr	1.97	0.47	0.002591	172.17
σ	0.0883	0.0299	0.000543	6.2778

Tab. 2 Srovnání identifikačních metod: SA, FSA, SA v kombinaci se slepým náhodným prohledáváním a SA* z hlediska přesnosti a časové náročnosti (*Zdroj: vlastní zpracování*)

4 Závěr

Znalost PM prostoru materiálu, např. kůže by mohla mít zásadní význam v oblasti dermatologie či v kosmetickém a farmaceutickém průmyslu. Využití je však mnohem rozsáhlejší. Na základě znalosti příslušného PM prostoru materiálu lze odhadovat aktuální poškození či elasticitu materiálu. Abychom však odhadli co nejpřesnější rozložení hysteronů v PM prostoru, využili jsme optimalizačních stochastických algoritmů. Implementovány byly jak standardní algoritmy (SA, FSA), tak i různé modifikace. Pro modifikaci navrženou speciálně pro tuto identifikační úlohu, označenou jako SA*, jsme testovali signifikantnost jednotlivých parametrů a nalezli jsme jejich optimální hodnoty. Dále jsme srovnávali tuto metodu se standardními algoritmy z hlediska přesnosti a časové náročnosti a dospěli jsme k závěru, že se tato metoda jeví jako vhodná pro nalezení přesnějšího řešení než klasické metody SA či FSA (případně v kombinaci s algoritmem slepého náhodného prohledávání), avšak jde o metodu časově náročnější.

5 Literatura

Mayergoyz, I. D. (2003). *Mathematical models of hysteresis and their applications*. Amsterdam: Elsevier.

McCall, K. R., & Guyer, R. A. (1996). A new theoretical paradigm to describe hysteresis, discrete

memory and nonlinear elastic wave propagation in rock. *Nonlinear Processes in Geophysics,3*(2), 89-101. doi:10.5194/npg-3-89-1996

McCall, K. R., & Guyer, R. A. (1996). A new theoretical paradigm to describe hysteresis, discrete memory and nonlinear elastic wave propagation in rock. *Nonlinear Processes in Geophysics,3*(2), 89-101. doi:10.5194/npg-3-89-1996

Kůs, V., Morales, D., & Vajda, I. (2008). Extension of the Parametric Families of Divergences Used in Statistical Inference. *Kybernetika*, *44*, 95-112.

McLachlan, G. J., & Peel, D. (2006). *Finite mixture models*. New York: John Wiley & Sons.

McCall, K. R., & Guyer, R. A. (1995). Equation of state and wave propagation in hysteretic nonlinear elastic materials. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts,32*(6). doi:10.1016/0148-9062(95)99156-r

Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1987). Optimization by Simulated Annealing. *Readings in Computer Vision,*606-615. doi:10.1016/b978-0-08-051581-6.50059-3

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal Chemical Physics,21*(6), 1087-1092. doi:10.2172/4390578

Modelovanie závislosti finančno-ekonomických ukazovateľov pomocou kopula funkcií
Dependence modelling of financial-economic indicators using copula functions

Nikoleta Lengyelová
Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, magisterské štúdium, 4.ročník
Comenius University in Bratislava, Faculty od Mathematics, Physics and Informatics
niki.lengyelová@gmail.com

Abstrakt: *Kopula funkcie sú vo veľkej miere využívané, ako nová štatistická metóda, na riešenie mnohých finančných problémov. Táto práca má za cieľ teóriu riešenej problematiky premietnuť do praktickej oblasti a na konkrétnych dátach hľadať vzájomnú závislosť finančno-ekonomických ukazovateľov. Práca je rozdelená do 4 kapitol. V prvej časti sú priblížené základy teórie kopula funkcií, najdôležitejšie vlastnosti, a to najmä Sklarova veta. Uvedené sú aj hlavné triedy kopula funkcií, ktoré sú využívané vo výskume. Nakoľko sa všetky výpočty uskutočňujú v štatistickom softvéri R, časť práce je venovaná aplikácii kopula funkcií v tomto softvéri, čo zahŕňa základné balíky a príkazy. Posledná kapitola práce je venovaná výskumu najvhodnejšej kopula funkcie pre reálne, finančno-ekonomické dáta.*

Abstract: *Copula functions are widely used, as a new statistical method, to solve many financial problems. The aim of this thesis is to put the theoretical background of the discussed issues into practice and with the concrete data to look for the interdependence between financial-economic indicators. The thesis is divided into 4 chapters. In the first part, the basic theory of copula functions, the most important properties, especially Sklar's theorem, are explained. It also contains main classes of copula functions that are used in research. Because all calculations are performed in R software, part of the work is devoted to the application of copula functions in this software, which includes basic packages and commands. The last chapter is devoted to the research of the most appropriate copula function on real, financial-economic data.*

Kľúčové slová: *kopula funkcie, triedy kopula funkcií, časové rady, závislosť časových radov.*
Key words: *copula functions, classes of copula functions, time series, dependence of time series.*

1 Úvod

Modelovanie korelácie medzi premennými je už dlhé stáročia predmetom diskusií nejedného matematika, štatistika či ekonóma. Popri historických metódach sa nie tak dávno dostali do popredia aj kopula funkcie, ktoré slúžia ako mocný nástroj na určenie závislosti vhodne zvolených ukazovateľov. Pojem kopula bol po prvýkrát využitý v Sklarovom výskume (1959) na pomenovanie funkcií, ktoré spájajú združené distribučné funkcie s častokrát jednoduchšími marginálnymi. Postupom času sa stávali kopuly viac a viac skúmanou oblasťou matematiky.

Táto práca sa zameriava predovšetkým na skúmanie závislosti bežných finančno-ekonomických ukazovateľov na základe historických dát. Cieľom tejto práce je zmapovať typ a silu závislosti vývoja skúmaných ukazovateľov použitím už spomínaných kopula funkcií. Základnou metódou pre dosiahnutie cieľa, je prehĺbenie si vedomostí teórie kopula funkcií a naštudovanie si podstaty odhadovania parametrov najznámejších a najpoužívanejších kopúl. Časť práce slúži na oboznámenie sa so štatistickým softvérom R, konkrétne so základným balíkom výskumu, *copula*. Posledná kapitola je venovaná aplikovaniu poznatkov na reálnych dátach. Túto kapitolu tvorí vlastný výskum na dvojici ukazovateľov získaných z reálnych, finančno-ekonomických dát a záver ponúka stručné zhrnutie výsledkov. Cieľom tohto výskumu je poskytnúť univerzálny návod na modelovanie závislosti časových radov a poukázať na dôležitosť testovania prítomnosti autokorelácie.

2 Teória kopula funkcií

Na odhadovanie vhodnej kopula funkcie pre reálne dáta je potrebné si v prvom rade priblížiť pojem kopula funkcie spolu s najdôležitejšími vlastnosťami a základnými triedami (Valková, 2013), pričom hlavný dôraz kladieme na bivariačné, teda dvojrozmerné kopuly. Pre podrobnejší výklad pozri (Nelsen, 2006).

2.1 Základné definície a vlastnosti

Definícia 2.1. Funkciu $C: I^n \to \mathbb{R}$ nazývame n-rozmernou kopulou (pričom $I = \langle 0,1 \rangle$), ak sú splnené nasledujúce vlastnosti:

- C je rastúca vo všetkých zložkách,
- C má jednorozmerné marginálne funkcie C_k, pre ktoré platí: $C_k(u) = u, \forall u \in I$,
- $\forall u \in I_n$ platí: ak $C(u) = 0$, tak aspoň jedna zložka vektora u je nulová,
- $\forall u \in I_n$ platí: ak $C(u) = u_i$, tak všetky zložky u, okrem u_i, sú jednotky,
- pre $\forall \mathrm{a} = (\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_\mathrm{n}), b = (b_1, b_2, \dots, b_n) \in I^n, \mathrm{a} \leq \mathrm{b}$, a n-rozmernú kocku $B = \langle \mathrm{a}, \mathrm{b} \rangle = \langle \mathrm{a}_1, \mathrm{b}_1 \rangle$ x $\langle \mathrm{a}_2, \mathrm{b}_2 \rangle$ x ... x $\langle \mathrm{a}_\mathrm{n}, \mathrm{b}_\mathrm{n} \rangle$, ktorej vrcholy ležia v definičnom obore C platí, že objem takejto kocky B je $V_c(B) = V_C(\langle \mathrm{a}, \mathrm{b} \rangle) \geq 0$.

Jednou zo základných vlastností kopula funkcií je schopnosť spájať marginálne distribučné funkcie náhodných veličín so združenou distribučnou funkciou náhodného vektora. Tento atribút je presnejšie popísaný v Sklarovej vete, avšak pre potreby tohto výskumu si ho zadefinujeme iba pre bivariačné rozdelenie.

Veta 2.1. Nech H je združená distribučná funkcia s marginálnymi distribučnými funkciami F a G. Potom existuje taká kopula C, že pre $\forall x, y \in \mathbb{R}$ platí:

$$H(x, y) = C(F(x)G(y))$$

Ak F a G sú spojité, potom C je jediná, inak C je jednoznačne definovaná na priestore ohraničenom obormi hodnôt F a G. Naopak, ak C je kopula a F a G sú distribučné funkcie, potom H je združená distribučná funkcia s marginálnymi distribučnými funkciami F a G.

2.2 Triedy kopula funkcií

Medzi základné triedy kopula funkcií patria:

- Archimedovské kopuly: Claytonova kopula, Frankova kopula, Gumbelova kopula, Joeova kopula,
- eliptické kopuly: Gaussova kopula, Studentova t-kopula,
- kopuly s extrémnymi hodnotami.

3 Aplikácia kopula funkcií v softvéri R

Nakoľko odhadovanie vhodnej kopula funkcie na reálnych finančno-ekonomických dátach prebieha v softvéri R (R Core Team, 2016), táto časť práce podrobnejšie popisuje využitie rôznych balíkov a príkazov, ktoré slúžia na modelovanie kopula funkcií a následné vyhodnocovanie údajov. Základný a zároveň najviac potrebný balík nášho výskumu je balík s názvom *copula*. Pre jeho spustenie je nevyhnutné ho v prvom rade nainštalovať pomocou príkazu *install.packages("copula")* a následne načítať použitím *library(copula).* V rámci balíka *copula* sa nám automaticky načítajú aj rôzne ďalšie knižnice ako napríklad *graphics*, potrebný pre grafické vykresľovanie výsledkov, *mvtnorm*, *sn* a množstvo iných. V tomto kontexte si uvedieme zopár funkcií a príkazov, s ktorými najčastejšie pracujeme, pričom komplexnejší prehľad poskytuje oficiálna príručka balíka *copula* (Hofert, Kojadinovic, Maechler, Yan, 2016), z ktorej v tejto kapitole vychádzame.

Medzi základné príkazy využívané vo výskume patria:

- funkcia pre distribučnú funkciu *pCopula*,
- funkcia pre hustotu *dCopula*,
- archimedovské kopuly *archmCopula*: *claytonCopula, gumbelCopula, frankCopula, joeCopula,*
- eliptické kopuly *ellipCopula*: *normalCopula*, *tCopula*,
- kopuly s extrémnymi hodnotami *evCopula*: *galambosCopula*, *huslerReissCopula*,
- test dobrej zhody, goodness-of-fit test: *gofCopula*,

- metóda maximálnej vierohodnosti, používaná ako globálny ukazovateľ pri výbere vhodnej kopula funkcie: *loglikCopula*.

4 Markovovská vlastnosť a jej testovanie

Definícia 4.1. Markovov reťazec popisuje diskrétny stochastický proces, pre ktorý platí, že pravdepodobnosť prechodu do nasledujúceho budúceho stavu závisí len na súčasnom stave.

Na testovanie Markovovskej vlastnosti nám slúži *Ljungov-Boxov test*, ktorý overuje, či niektorá zo skupín autokorelácií časového radu sa líši od nuly. *Ljungov-Boxov test* môžeme použiť na testovanie nasledujúcich hypotéz:

- H_0: Zložky časového radu sú nezávislé, čo znamená, že jednotlivé autokorelácie časového radu sú nulové.
- H_1: Zložky časového radu nie sú nezávislé a vykazujú sériovú závislosť.

Testová štatistika je definovaná ako:

$$\boldsymbol{Q = n(n+2)\sum_{k=1}^{h}\frac{\widehat{\rho}_k^2}{n-k}},$$

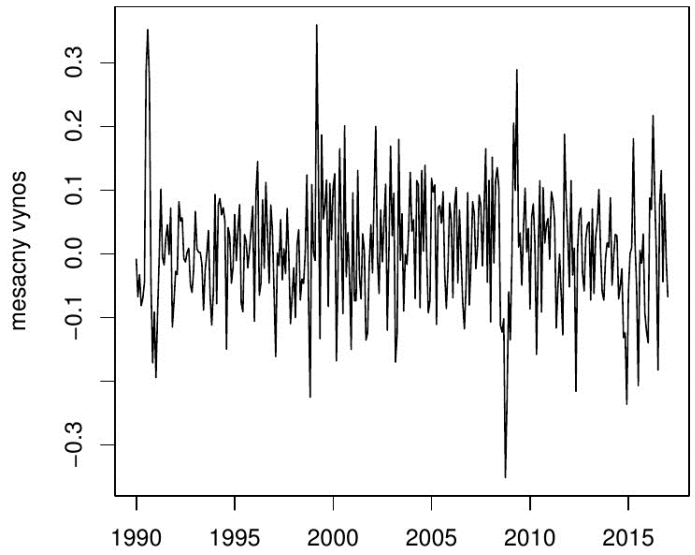
kde n je počet dát, $\hat{\rho}_k$ je autokorelácia k-teho stupňa a h je počet testovaných oneskorení. Hypotézu H_0 zamietame, ak testová štatistika $Q > \chi^2_{1-\alpha,\mathrm{h}}$ na hladine významnosti α (obvykle $\alpha = 5\%$), kde $\chi^2_{1-\alpha,h}$ je $1-\alpha$ kvantil χ^2 rozdelenia s počtom stupňov voľnosti h.

5 Výskum závislosti časových radov na vlastných dátach

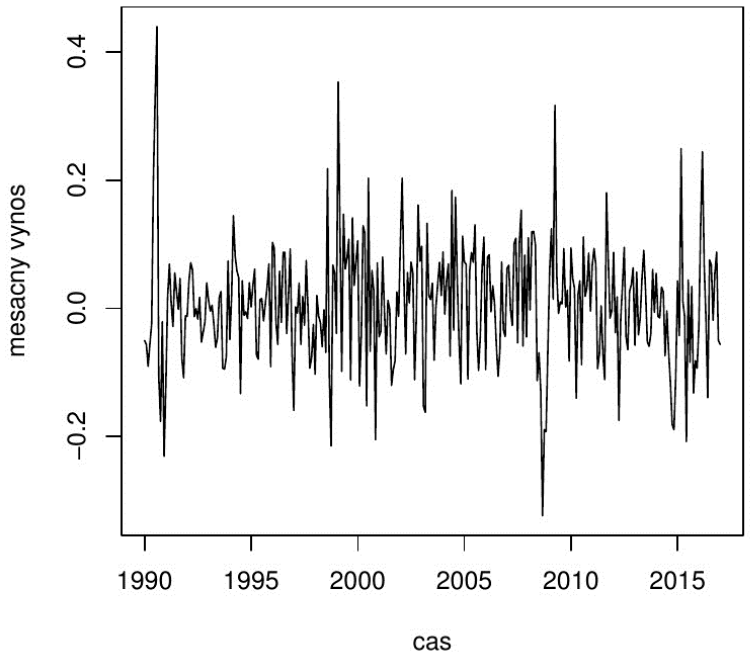
Cieľom tejto práce je mapovať typ a silu závislosti vývoja skúmaných ukazovateľov využitím teórie kopula funkcií. Dôležitým aspektom, ktorý musia tieto ukazovatele spĺňať, je Markovovská vlastnosť. Záverečná časť práce sa preto zaoberá výskumom nami zvolených časových radov, u ktorých bola Markovovská vlastnosť viditeľná jednak graficky, z autokorelačných grafov Obr. 3 a Obr.4, na druhej strane sa prítomnosť Markovovskej vlastnosti potvrdila aj po spustení *Ljungovho-Boxovho* testu. Dáta, ktoré sme si pre tento výskum zvolili, sú ceny futurít ropy *Oklahoma Crude Oil Future* (ďalej budeme označovať ako NYMEX) a *West Texas Intermediate Crude Oil Prices* (ďalej budeme označovať ako WTI) získané zo zdrojov (Fred Economic Data, 2017) a (U.S. Energy Information Administration, 2017) v mesačnej frekvencii z časového obdobia 1990-2017. Jednotky, v ktorých sú dáta udávané, sú doláre za barel ropy. Z mesačných dát sme následne definovali relatívne mesačné výnosy pomocou nasledujúcej schémy:

$$\frac{P_t - P_{t-1}}{P_{t-1}},$$

kde P_t je cena barelu ropy v čase t.Tieto výnosy sú zobrazené na nasledujúcich grafoch Obr.1 a Obr.2.

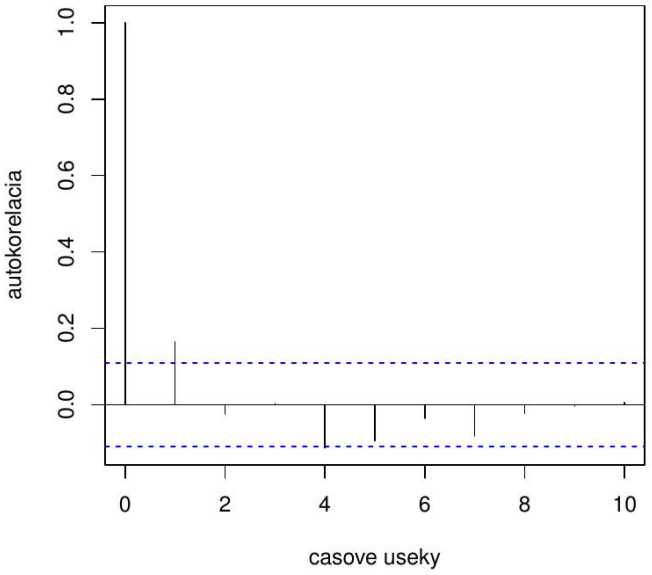


Obr. 1: Graf mesačných výnosov pre časový rad NYMEX *(Zdroj: vlastné spracovanie)*

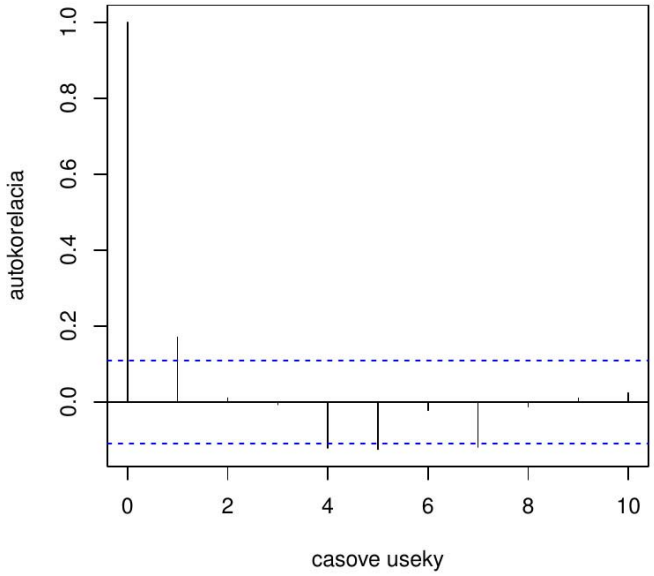


Obr. 2: Graf mesačných výnosov pre časový rad WTI (Zdroj: vlastné spracovanie)

Grafy Obr.3 a Obr.4 predstavujú autokorelačné grafy časových radov WTI a NYMEX.



Obr. 3: Autokorelačný graf pre časový rad WTI
(Zdroj: vlastné spracovanie)



Obr. 4: Autokorelačný graf pre časový rad NYMEX
(Zdroj: vlastné spracovanie)

Môžeme si všimnúť, že v oboch prípadoch je korelácia medzi dvomi susednými zložkami časového radu najsilnejšia, čo nám napovedá možnú prítomnosť Markovovskej vlastnosti. Treba podotknúť, že autokorelačný graf slúži len ako pomocný nástroj na posúdenie prítomnosti alebo neprítomnosti Markovovskej vlastnosti v sledovanom časovom rade. Na presnejšie určenie prítomnosti Markovovskej vlastnosti slúži už spomínaný *Ljungov-Boxov test*.

Vo výskume sme uvažovali Archimedovské a eliptické kopula funkcie, konkrétne Gaussovu, Studentovu, Claytonovu, Frankovu, Gumbelovu a Joeovu. Pre všetky tieto triedy kopula funkcií sme odhadli parameter. Nasledujúca Tabuľka 1 uvádza odhadnuté hodnoty parametrov kopula funkcií v prípade, že uvažujeme prítomnosť Markovovskej vlastnosti. Pre Studentovu kopulu uvádzame aj počet stupňov voľnosti.

Tabuľka 1: Hodnoty parametrov kopula funkcií *(Zdroj: vlastné spracovanie)*

Triedy kopúl	**Gaussova**	**Studentova**	**Claytonova**
Hodnota parametra	0,210	0,157	0,352
Stupne voľnosti		2,457	

Triedy kopúl	**Frankova**	**Gumbelova**	**Joeova**
Hodnota parametra	1,277	1,168	1,156

Na zistenie, ktorá kopula nám najlepšie fituje dáta sme použili globálny ako ukazovateľ hodnotu funkcie *loglikCopula*. Na základe Tabuľky 2 môžeme skonštatovať, že najvyššiu hodnotu vykazuje Claytonova kopula. Túto kopulu by

sme teda odporučili na modelovanie štruktúry závislosti medzi výnosmi NYMEX Futures a WTI.

Tabuľka 2: Hodnoty funkcie *loglikCopula* pre jednotlivé triedy kopula funkcií
(Zdroj: vlastné spracovanie)

Triedy kopúl	**Gaussova**	**Studentova**	**Claytonova**
loglikCopula	6,91	6,21	**8,13**

Triedy kopúl	**Frankova**	**Gumbelova**	**Joeova**
loglikCopula	4,99	6,50	4,74

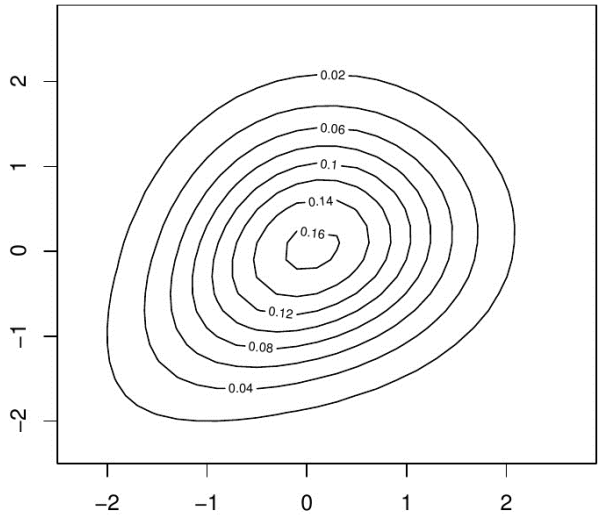
Na overenie správnosti našich výsledkoch bol použitý aj *goodness-of-fit test*, ktorý slúži na meranie dobrej zhody.

Tabuľka 3: P-hodnoty funkcie *gofCopula* pre jednotlivé triedy kopula funkcií
(Zdroj: vlastné spracovanie)

Triedy kopúl	**Gaussova**	**Studentova**	**Claytonova**
gofCopula	0,33	0,55	**0,71**

Triedy kopúl	**Frankova**	**Gumbelova**	**Joeova**
gofCopula	0,21	0,16	0,02

Obr. 5 ilustruje Claytonovu kopula funkciu s parametrom 0,352, ktorú nám všetky testy prehlásili za najlepšiu. Na osiach x a y sú znázornené argumenty kopula funkcie.



Obr. 5: Claytonova kopula s parametrom 0,352
(Zdroj: vlastné spracovanie)

5.1 Dopad zanedbania autokorelácie časového radu na výsledok

Pre porovnanie výsledkov sme sa rozhodli zrealizovať aj výskum v prípade zanedbania sériovej závislosti. Uvažovali sme rovnaké dáta, s akými sme doteraz pracovali v kapitole 5, teda NYMEX a WTI, ktoré vykazovali signifikantnú Markovovskú vlastnosť. Tabuľka 4 predstavuje výsledky funkcie *loglikCopula* v prípade, ak pri výskume zanedbáme dôležitý faktor, autokoreláciu.

Tabuľka 4: Hodnoty funkcie *loglikCopula* pre jednotlivé triedy kopula funkcií, v prípade zanedbania Markovovskej vlastnosti (Zdroj: vlastné spracovanie)

Triedy kopúl	**Gaussova**	**Studentova**	**Claytonova**
loglikCopula	6,91	**9,82**	8,50

Triedy kopúl	**Frankova**	**Gumbelova**	**Joeova**
loglikCopula	5,09	6,84	4,78

Ako si môžeme všimnúť, najvyššie hodnoty nadobudla Studentova t-kopula, kým Claytonova skončila až na druhom mieste. V tomto prípade, by sme prehlásili, že Studentova kopula je najvhodnejšia na modelovanie našich dát. My však teraz už vieme, že tento záver by nebol najsprávnejší. Z výskumu teda vyplýva, že keď zanedbáme signifikantnú autokoreláciu v rámci časových radov, môžeme dostať úplne iný kopula model, ako v prípade pokročilejšieho prístupu, ktorý berie do úvahy aj prítomnosť Markovovskej vlastnosti v rámci časových radov.

6 Záver

Kopula funkcie sú v dnešnej dobe často využívaným nástrojom na určenie štruktúry závislosti medzi finančno-ekonomickými ukazovateľmi. Našou úlohou v tejto práci bolo priblížiť problematiku kopula funkcií od základnej teórie, až po praktické využitie a aplikáciu na reálnych dátach. Práca je prínosom nielen pre matematikov, či štatistikov, ktorí sa s pojmom kopula funkcie stretávajú na dennej báze, ale je určená aj pre tých, ktorí by sa chceli dozvedieť niečo nové z oblasti skúmania závislosti finančno-ekonomických ukazovateľov. Hlavnú časť práce tvorí výskum, z ktorého vyplynulo, že ak zanedbáme signifikantnú autokoreláciu v rámci časových radov, môžeme dostať úplne iný kopula model, než v prípade, ak berieme do úvahy aj prítomnosť Markovovskej vlastnosti v rámci časových radov. Z tohto dôvodu za hlavný prínos práce považujeme poukázanie na dôležitosť testovania prítomnosti Markovovskej vlastnosti v dátach.

7 Literatúra

Fred Economic Data, [cit. 29.04.2017, dostupné na: fred.stlouisfed.org/series/DCOILWTICO]

Gudendorf, G., Segers, J. (2009). *Extreme-Value Copulas*, Institut de statistique, Louvain-la Neuve: Universite catholique de Louvain

Hofert, M., Kojadinovic, I., Maechler, M., Yan, J. (2016). *Copula: Multivariate Dependence with Copulas. R package version* 0.999-16, [dostupné na:CRAN.Rproject.org/package=copula]

Nelsen, R. B. (2006). *An Introduction to Copulas*, New York : Springer

R Core Team. (2016). R: *A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, [dostupné na: https://www.R-project.org/]

Rémillard, B., Papageorgiou, N., Soustra, F. (2012). Copula-based semiparametric models for multivariate time series, *Journal of Multivariate Analysis*, [cit. 22.04.2017, dostupné na: http://www.brunoremillard.com/Papers/dyncop.pdf]

U.S. Energy Information Administration, [cit. 29.04.2017, dostupné na: www.eia.gov]

Valková, M. (2013). *Bivariačné kopuly v poistnej matematike*, bakalárska práca, Bratislava : Comenius University

Lengyelová, N. (2017). *Modelovanie závislosti finančno-ekonomických ukazovateľov pomocou kopula funkcií*, bakalárska práca, Bratislava : Comenius University

Vzdelanie ako diferenčný faktor plodnosti na Slovensku
Educational as a differential factor of fertility in Slovakia

Michaela Lukáčová
Prírodovedecká fakulta Univerzity Komenského v Bratislave, typ štúdia: magisterské, ročník: druhý
Faculty of Natural Sciences of the Comenius University in Bratislava
michaela.lukac@gmail.com

Abstrakt: *Tento príspevok sa snaží identifikovať vplyv dosiahnutého vzdelania ako aj dopady jeho dynamických zmien v spojitosti s plodnosťou žien na Slovensku a jej vývojom po roku 1989. Dôraz sa kladie najmä na časovanie plodnosti v poslednom štvrťstoročí s aplikáciou transverzálneho pohľadu, ktorý je založený na dátach zo sčítaní obyvateľov (1991-2001-2011).*

Abstract: *The aim of this paper is to identify the impact of achieved education as well as effects of its dynamic changes on female fertility in the Slovak Republic after 1989. The key element is to focus on timing of fertility in the last quarter-century by application of a transversal perspective based on data obtained from last three realized censuses (1991-2001-2011).*

Kľúčové slová: *vzdelanie, plodnosť, transformácia, Slovensko*
Key words: *educational, fertility, transformation, Slovakia*

1 Úvod

Úroveň dosiahnutého vzdelania predstavuje v populácií Slovenska jeden z najdynamickejšie sa meniacich znakov, ktorý po roku 1989 prešiel, respektíve stále prechádza historicky nebývalými kvalitatívno-kvantitatívnymi zmenami. Jeden z najtypickejších znakov je rastúci dopyt po vyššej forme vzdelania, v poslednom desaťročí sa jedná najmä o terciárne vzdelanie. Spolu s ním dochádza k zmene nielen pri samotnom procese vzdelávania, ale i pri vytváraní inštitucionálnych podmienok. Terciárne vzdelanie a jeho nárast od začiatku nového tisícročia je spojený s neustále sa rozširujúcimi formami tohto typu vzdelávania, no zároveň je ovplyvnený dopytom spoločnosti a pracovného trhu po tejto forme vzdelania. Tento vzájomný vzťah, ktorý dominuje v transformačnom procese spoločnosti výrazne ovplyvňuje intenzitu a charakter plodnosti žien na Slovensku. Najvyššie dosiahnuté vzdelanie, pred ukončením demografickej revolúcie aj gramotnosť, bývajú považované za jedny z najvýznamnejších faktorov reprodukčného správania.

2 Metodika práce a získavanie zdrojov

Pri písaní tohto príspevku sme čerpali zo zahraničnej i domácej literatúry, ktorá je uvedená v poslednej kapitole „Použitá literatúra“. Literatúra slúžila na hlbší a odbornejší prehľad v danej problematike a pri samotných výpočtoch bola jedným z dôležitých podkladov na vysvetlenie jednotlivých grafov.

Zdrojom pre jednotlivé číselné údaje bol slovenský štatistický úrad a jeho internetová databáza (www.slovak.statistics.sk), údaje zo sčítania pochádzajú z internetovej databázy Infostat-u (www.sodb.infostat.sk).

Úhrnná plodnosť bol prvý ukazovateľ, ktorý sme počítali. Podľa Jurčovej (2005) úhrnná plodnosť znamená, priemerný počet živonarodených detí pripadajúcich na jednu ženu počas jej celého reprodukčného obdobia (15-49 rokov), pri zachovaní úrovni plodnosti a potratovosti sledovaného roka a za predpokladu nulovej úmrtnosti. Pri výpočte sme postupovali podľa výpočtu Rochovskej a kol (2014).

up= $\sum_{x=15}^{49} f_x$ ***,*** *kde sa sčítavajú úrovne špecifických plodností v jednotlivých vekových kategóriách. Fx je špecifická plodnosť vo veku* ***x.***

Ďalší ukazovateľ pomocou, ktorého sme analyzovali plodnosť je ***priemerný vek matky pri pôrode*** (podľa poradia dieťaťa).

$$xf = \frac{\sum_{x=15}^{49} x \times fx^{poradie}}{\sum_{x=15}^{49} fx^{poradie}}$$

Pri počítaní priemernej dĺžky jedného života, ktorú žena strávi ako bezdetná (v rokoch) sme použili metódu tzv. ***SMAM***. Túto metódu prvýkrát použil Hajnal v roku 1953. Aplikoval ju na manželstvo, preto sa oficiálne volá ***S****ingulate* ***M****ean* ***A****t* ***M****arriage (SMAM).*

SMAM=(15+5P-50*N)/(1-N), kde **„15“** je veková hranica, **„5P“** sú 5 ročné vekové kategórie (dá sa pracovať i s jednoročnými), **„N“** znamená presnú hodnotu vo veku 50 (Booth, H., 2001).

Hodnota „15“ sa nám menila podľa potreby, teda pre ktorú vzdelanostnú skupinu sme priemernú dĺžku jedného života ženy ako bezdetnej počítali. Pre vzdelanostnú skupinu „základné a bez vzdelania“ to bola už vyššie spomínaná hodnota 15 rokov. Pre „stredoškolské vzdelanie bez maturity“ sme určili hodnotu 16 rokov, pre „stredoškolské s maturitou“ sme stanovili hranicu 18 rokov a pre „vysokoškolské vzdelanie“ sa určila hranica 22 rokov. V závere príspevku je uvedené najdôležitejšie zhrnutie získaných poznatkov.

3 Vzájomný vzťah medzi plodnosťou a najvyšším dosiahnutým vzdelaním

Vo všeobecnosti je čas strávený nadobúdaním vyššej formy vzdelania nekompatibilný s plodnosťou a formovaním rodiny (Baizán, P., a kol., 2003). Často krát je snaha získať vyššie vzdelanie braná ako odkladanie dospelosti, či už sa to týka ekonomickej nezávislosti, vstupu na trh práce, či zakladania rodiny (Kohler, H., Billary, F., 2002). Všetky tieto priame vplyvy sú zároveň doplnené o nepriame pôsobenie ďalších faktorov – napríklad socioekonomická situácia,

preferencie a hodnoty, možnosti uplatnenia na trhu práce. Vo finálnej fáze, na základe stupňa dosiahnutého vzdelania, sa ich výsledky vplyvu diferencujú.

Vzdelanejšie ženy sa snažia oveľa častejšie nadobudnúť stabilnú pracovnú pozíciu, dosiahnuť adekvátne finančné ohodnotenie a primerané bývanie. Ak ženy s vyšším vzdelaním dosahujú uspokojivú vlastnú ekonomickú situáciu, vedie to k ekonomickej nezávislosti a z toho dôvodu, nestabilný príjem, alebo neistota na trhu práce môže pôsobiť negatívne na plodnosť (Sobotka, T., a kol., 2008). U žien s nižším vzdelaním je tento vzťah zložitejší. V niektorých prípadoch dochádza dokonca k opačnému efektu. Prostriedok na redukciu neistoty, tak je vnímané materstvo a rodičovstvo u žien s nižším vzdelaním, často krát tesne po dokončení určitej formy vzdelania (Friedaman, D., a kol., 1994). Rovnako je založenie rodiny, respektíve nadobudnutie statusu matky, brané ako jeden z istejších zdrojov príjmu. Keď sa zameriame na vzťah medzi vzdelanostnou štruktúrou a postavením na trhu práce, dostávame sa k tzv. nákladom stratených príležitostí (stagnácia na trhu práce, platobných podmienok, strata kontaktu s pracovným prostredím). Tie predstavujú pre ženy s vyšším vzdelaním často krát väčšie prekážky ako samotná finančná záťaž, ktorá vzniká pri starostlivosti o dieťa.

Pri snahe objasniť vzťah medzi vzdelaním a poklesom plodnosti existuje niekoľko dôvodov. Cieľom vzdelanejších rodičov je primerane sa postarať o svoje deti (dieťa), tým pádom vyššie výdavky prispievajú k prípadnému obmedzeniu počtu detí (Van de Walle, F., 1980). Obmedzeniu detskej práce prispela i snaha vyššej gramotnosti detí a ich následné začlenenie do vzdelávacieho procesu. Z toho dôvodu klesal prínos detí pre domácnosť, z hľadiska financií. V mestskom prostredí príliš veľa detí nevyhovovalo bytovým podmienkam. Vzdelanejšie ženy sa museli po každom narodenom dieťati vyrovnať s vysokými nákladmi stratených príležitostí (Hašková, H., 2010). Medzi hrubou mierou pôrodnosti a viacerými ekonomickými premennými, napr. priemerná mzda, podiel ekonomicky aktívnych žien, vybavenosť domácností, existuje štatisticky významný nepriamy vzťah. Čím väčšia je hodnota takéhoto faktora, tým je miera pôrodnosti menšia, a tým výraznejší je jej pokles. V neposlednom rade treba podotknúť, že ženy s vyšším vzdelaním mali lepšie podmienky a príležitosti na prípadnú redukciu neželaných tehotenstiev, zároveň boli ochotnejšie tieto prostriedky využiť (Coale, A., 1973).

Na vzťah vzdelania a plodnosti sa odborníci pozerajú i pomocou niekoľkých teórií. Teória preferencií (Hakim, C., 2003) reprezentuje 3 typy vzťahu ženy k rodine.

1 *Orientované profesijne* - ženy s vyšším stupňom vzdelania budú vo väčšej miere orientované na profesiu, viac kompromisov ohľadne načasovania materstva a veľkosti rodiny.

2 *Orientované na rodinu* - ženy s nižším stupňom vzdelania výraznejšia orientácia na rodinu.
3 *Tzv. Adaptívne ženy* - väčšina žien sa bude snažiť jednotlivé dráhy harmonizovať.

Klasické ekonomické teórie vysvetľujú nižší počet detí a odkladanie rodičovstva u žien s vyšším dosiahnutým vzdelaním pomocou nákladov stratených príležitostí. Vzdelanejšie ženy sú ekonomicky nezávislé a zároveň majú lepšie poznatky o antikoncepčných prostriedkoch a ich správnom používaní.

Mikroekonomická teória „relatívneho príjmu“ (Easterlin, 1897) tvrdí, že úroveň plodnosti nie je ovplyvnená aktuálnym stavom ekonomiky, v ktorej sa jedinec nachádza, ale závisí skôr od toho, ako tento stav jedinci vnímajú.

Sobotka (2004) uviedol niekoľko faktorov vplyvu pri odkladaní plodnosti:

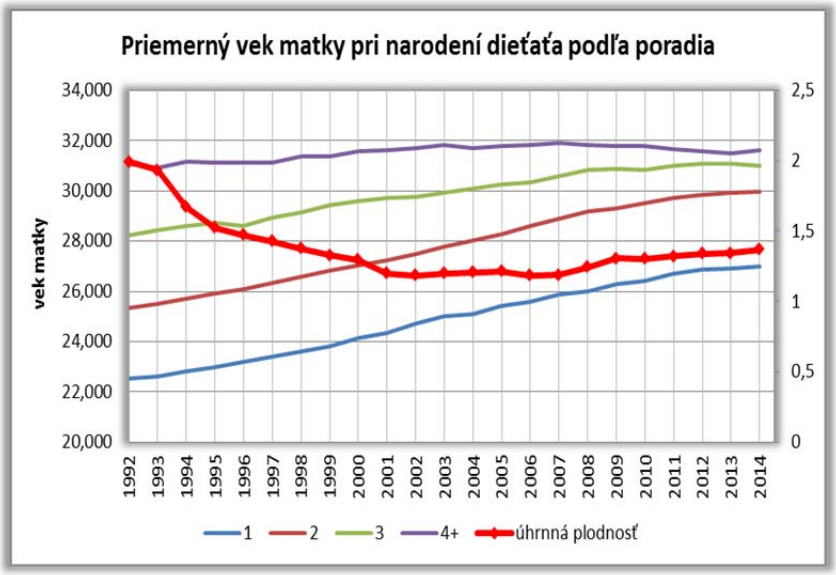
- vplyv úrovne vzdelanosti
- konflikt medzi zamestnaním a rodičovstvom
- nezamestnanosť, neistota, ekonomické podmienky a načasovanie prvého pôrodu
- transformácia rodiny a partnerských vzťahov
- “antikoncepčná” revolúcia odklad pôrodov prvého poradia

4 Úhrnná plodnosť a priemerný vek matky pri narodení dieťaťa podľa poradia

Graf č. 1 nám odzrkadľuje situáciu od roku 1992 až po rok 2014. Môžeme vidieť, že všetky typy poradia (1,2,3,4+) zaznamenali posun k vyšším vekom matiek. Do určitého obdobia (2007-2008) si držali pomerne rovnaký rozstup medzi sebou, avšak po tomto bode sa poradie 4+ ustálilo a stagnuje pod hranicou veku 31. To znamená, že žena, ktorá sa rozhodne stať matkou po štvrtý krát, prípadne piaty a viac, bude mať 31 rokov. Matky s troma deťmi sa k tejto vekovej hranici taktiež priblížili. Matky s dvoma deťmi dosahujú vek 30 rokov. Najvýraznejšiu zmenu sme zaznamenali pri prvorodičkách. Zatiaľ čo v roku 1992 mali okolo 22,5 roka, v roku 2014 sa ženy prvýkrát stanú matkami až, keď dovŕšia 27 rokov. Za toto približne 20-ročné obdobie je to najvýraznejší posun.

Tieto výsledky nám potvrdzuje i úhrnná plodnosť, ktorá naopak na grafe klesá. Po politickej transformácií na začiatku 90-tych rokov došlo k výraznej zmene trendu vývoja plodnosti. Aj keď Slovensko oproti Českej republike bolo charakteristické vyššími hodnotami úhrnnej plodnosti, od roku 2005 tento fakt neplatí (Rabušic a kol., 2012). Môžeme vidieť, že z hodnôt cca 2 deti na jednu ženu z roku 1992 sa behom tohto obdobia klesla pod hodnotu 1,5 dieťaťa na

ženu. Najvýraznejší pokles bol do roku 2001, kedy dokonca dosahovala ešte nižšie hodnoty ako v roku 2014. Od obdobia 2006-2008 dochádza k stabilizácií hodnôt úhrnnej plodnosti.



Graf 1 Priemerný vek matky pri narodení dieťaťa podľa poradia
Zdroj: ŠÚ SR (2016), výpočty autora

5 Vývoj jednotlivých vzdelanostných skupín žien na Slovensku

Vzťah plodnosti a úrovne dosiahnutého vzdelania je dôležitý najmä u žien a to predovšetkým v reprodukčnom veku. Keď sa zameriame na vývoj jednotlivých vzdelanostných skupín (základné, stredoškolské bez maturity, stredoškolské s maturitou a vysokoškolské) podľa autorov Šprocha a kol. (2010), ako je uvedené na grafoch č. 2-5, základné vzdelanie u žien reprodukčnom veku dominovalo (90%) od roku 1950 vo všetkých vekových skupinách. V 70-tych rokoch dosahovalo hodnoty od 50% (veková kategória 25-29) až po 80% (veková hranica 40-49). V nasledujúcom sčítaní v roku 1991 dosahovalo hodnoty okolo v rozmedzí 10% až 30%. Posledné sčítanie za rok 2011 vykazuje nízke hodnoty vo všetkých vekových kategóriách do 10%. Je to jediná vzdelanostná skupina, ktorá sa pohybuje v takmer celej percentuálnej škále počas skúmaného obdobia.

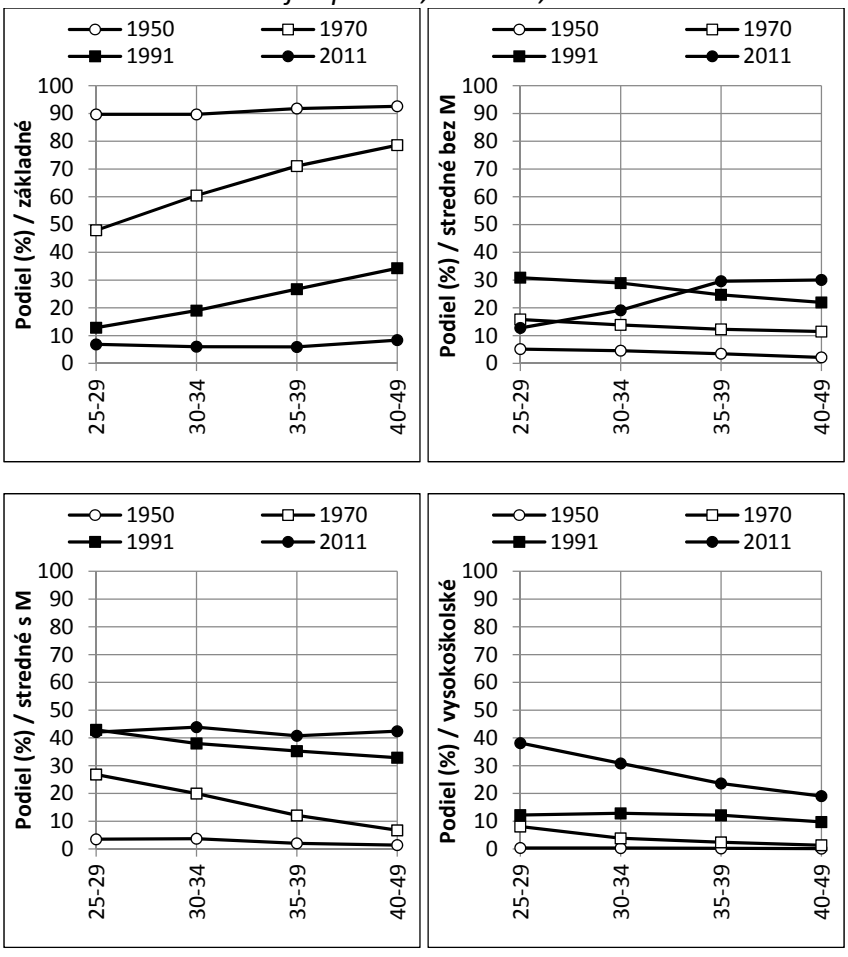
Pri stredoškolskom vzdelaní, respektíve pri všetkých ostatných vzdelanostných skupinách dochádza k opačnému efektu, nárastu. Stredoškolské bez maturity dosahuje svoje maximum pri 30-tich percentách a to v období 1991 pri vekových kategóriách 25-34 a potom v období 2011 pri vekových kategóriách 35-49 rokov. Stredoškolské s maturitou dosahuje svoje maximum pri hodnote okolo 40% v poslednom sčítaní. Túto hodnotu dosahovalo ešte v predchádzajúcom sčítaní v roku 1991, ale len u vekovej kategórie 25-29, následne došlo k poklesu.

Vysokoškolsky vzdelané ženy dosahujú najmenšie hodnoty za všetky sčítania a logicky v poslednom sčítaní sú tieto hodnoty najvyššie – okolo 40% (veková kategória 25-29).

Posledné sčítanie ukazuje, že stredoškolské s maturitou a vysokoškolské je najviac zastúpené u žien vo vekovej kategórií 25-29 (okolo 40%). U žien 30-34 rokov dominuje stredoškolské s maturitou (nad 40%), vysokoškolské (30%), stredoškolské bez maturity (20%) a základné vzdelanie (pod 10%). Vo vyššej vekovej kategórií 35-39 je to veľmi podobné. Najviac zastúpené je tu stredoškolské s maturitou (40%), rozdiel je, že za ním nasleduje stredoškolské bez maturity (30%), vysokoškolské (nad 20%) a základné vzdelanie (pod 10%). Najvyššia veková kategória z hľadiska reprodukčného správania 45-49 vykazuje rovnaké zastúpenie ako mladšia veková kategória 35-39. Dominuje stredoškolské s maturitou (nad 40%), stredoškolské bez maturity (30%), nasleduje vysokoškolské vzdelanie (20%) a vzdelanostné skupiny uzatvára s hodnotu 10% základné vzdelanie.

Grafy 2-5 Vývoj podielu žien so základným, stredoškolským bez maturity, s maturitou a vysokoškolským vzdelaním vo veku 25-49 rokov, sčítanie 1950, 1970, 1991 a 2011

Zdroj: Šprocha, B. a kol., 2010



Najväčší nárast zastúpenia žien na vysokých školách prebehol behom posledných desiatich rokov. Výsledkom tohto pomerne dynamického záujmu je fakt, že od roku 2002 ich podiel na celkovom počte študentov vysokých škôl presahuje hranicu 50% a od akademického roku 2007/2008 tvoria ženy už 60% vysokoškolákov (Šprocha, B., Potančoková, M., 2010). Významnú úlohu zohráva i rozmach externého štúdia, ktoré ešte v minulom režime slúžilo len na zvýšenie už získaného vzdelania u pracujúcich osôb. Možnosť byť súčasťou terciárneho štúdia pomohol i novo prijatý zákon o vysokých školách (Zákon č. 131/2002 Zb. z.), ktorý upravuje aj financovanie vysokých škôl. Výsledkom bolo zväčšenie kapacity novoprijatých študentov, otváranie nových odborov, ale taktiež nových verejných ale najmä súkromných škôl so externou možnosťou výučby.

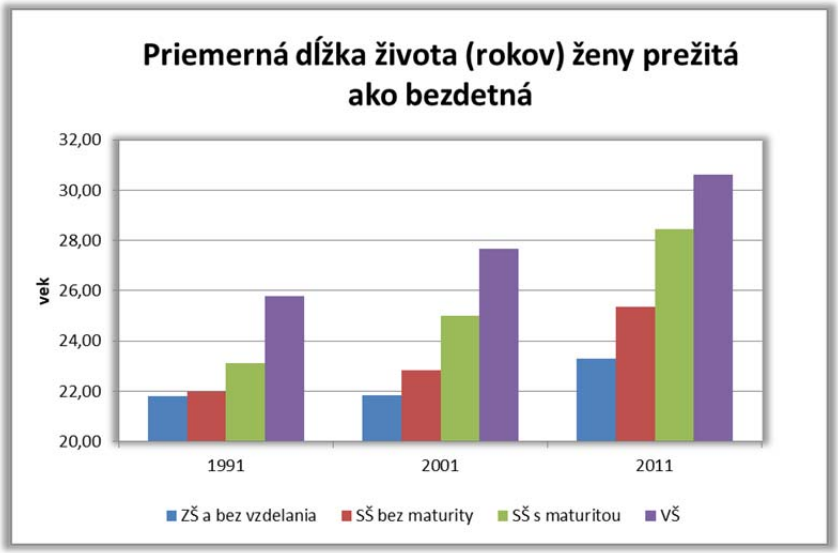
6 Priemerná dĺžka života ženy prežitá ako bezdetná podľa vzdelanostných skupín

Na nasledujúcom grafe č. 6 je znázornená priemerná dĺžka života (rokov) ženy, ktorú prežila ako bezdetná. Ako sme očakávali vo všetkých vzdelanostných skupinách má rastúci trend. Ženy so základným vzdelaním a bez vzdelania sa stávajú matkami ako prvé. I v tejto kategórií je vidno posun smerom k vyššiemu veku – z roku 1991 (21,8 rokov) na 23,29 rokov v roku 2011. Stredoškolsky vzdelané ženy bez maturity mali veľmi podobné hodnoty v roku 1991 (21,98). Avšak v roku 2011 rozdiel dosahoval takmer 2 roky. Ženy so strednou školou bez maturity prežijú 25,36 roka ako bezdetné, podľa údajov z roku 2011. Stredoškolské vzdelanie s maturitou si udržuje približne dvojročný odstup od žien so strednou školou bez maturity za všetky tri sčítania. Vysokoškolsky vzdelané ženy dosahujú najvyšší vek. V roku 2011 prežili ako bezdetné 30,52 rokov. Najväčšou zmenou prešli však ženy so strednou školou s maturitou. Rozdiel medzi rokom 1991 a 2011 je až 5,29 rokov. Väčší nárast dosahujú i medzi rokmi 1991 a 2001, rozdiel je 1,88 rokov. Medzi rokmi 2001 a 2011, to je 3,4. Ženy s terciárnym vzdelaním vo všetkých troch porovnaniach dosahujú menší nárast.

Tabuľka 1 Priemerná dĺžka života ženy prežitá ako bezdetná (v rokoch)

	ZŠ a bez vzdelania	SŠ bez maturity	SŠ s maturitou	Vysokoškolské
1991	21,80	21,98	23,10	25,77
2001	21,84	22,83	24,99	27,41
2011	23,29	25,36	28,39	30,52

Graf 6 Priemerná dĺžka života (rokov) ženy prežitá ako bezdetná
Zdroj: ŠÚ SR(2016), Infostat, výpočty autora



7 Záver

Najvyššie dosiahnuté vzdelanie sa stalo z jedným z najvýznamnejších diferenčných faktorov reprodukčného správania a to najmä v spoločnostiach, ktoré prešli transformáciou spoločnosti. To, že medzi jednotlivými vzdelanostnými skupinami a ich reprodukčným správaním existuje niekoľko dôležitých rozdielov sa potvrdilo i v prípade Slovenska. Terciárne vzdelanie a jeho nárast od začiatku nového tisícročia je spojený s neustále sa rozširujúcimi formami tohto typu vzdelávania, no zároveň je ovplyvnený dopytom spoločnosti a pracovného trhu po tejto forme vzdelania. Tento vzájomný vzťah, ktorý dominuje v transformačnom procese spoločnosti výrazne ovplyvňuje intenzitu a charakter plodnosti žien na Slovensku.

Okrem poklesu úhrnnej plodnosti k hodnote 1,5 dieťaťa na ženu od začiatku 90-tych rokov, dochádza i k zvyšovaniu priemerného veku matky podľa poradia dieťaťa. Najvyššie hodnoty pri priemernej dĺžke života ženy ako bezdetná dosahujú vysokoškolsky vzdelané ženy, avšak najväčšou zmenou si prešli ženy so strednou školu s maturitou. Práve v ich prípade jednotlivé rozdiely medzi obdobiami 1991 až 2011 boli najväčšie.

Vzájomné pôsobenie plodnosti a vzdelania je teda nepopierateľné. Snaha dosahovať vyššie vzdelanie prináša so sebou odkladanie rodičovstva. Naopak, skoré materstvo v konečnom dôsledku znižuje dosiahnuté materstvo.

8 Použitá literatúra

BAIZÁN, Pau - AASSVE, Arnstein - BILLARI, C. Francesco. (2003) Cohabitation, marriage, and first birth: The interrelationship of family formation events in Spain. In: *European Journal of Population* 19, č. 2, s. 147-169

BOOTH, H. (2001) Trends in mean age at first birth and first birth intervals in the Pacific Islands. In: *Genus*. January, s. 165-190.

EASTERLIN, R.(1974) Does economic growth improve the human lot? Some empirical evidence. In. *Nations and households in economic growth*, s. 89-125.

FRIEDAMAN, D. (1994) Theory of the Value of Children. In. *Demography*, s. 375-401.

HAKIM, C. (2003) A new aproach to explaining fertility patterns: Preference theory. In: *Population and Development Review*, s. 349-374.

HAŠKOVA, H. (2010) *Fenomén bezdětnosti*. Praha: Slon, s. 42-44.

JURČOVÁ, D. (2005) *Slovník demografických pojmov*. Bratislava: INFOSTAT – Inštitút informatiky a štatistiky, Akty. 72s. ISBN 0-85659-40-9. Dostupné z: http://www.infostat.sk/vdc/pdf/slovnik_2verdd.pdf

KOHLER, H., BILLARI, C. (2002) *The Emergence of Lowest-Low Fertility in Europe During the 1990s*., s. 641-680.

SOBOTKA, T., ŠŤASTNÁ, A. (2008) A rapid tranformation of fertility and family behaviour after the collapse of state socialism. In: *Demography Research*, s.403-454.

ROCHOVSKÁ, A., KAČEROVÁ, M. (2014) *Výskumné metódy v humánnej geografii a ich aplikácie*. Bratislava: Univerzita Komenského, 219 s. ISBN 978-80-223-3787-8.

VAN DE WALLE, F. (1980) Education and the Demographic Transition in Switzerland. In *Population and Development Review*, s. 463-472.

ŠPROCHA a kol. (2010) *Vzdelanie ako diferenčný faktor reprodukčného správania*. Bratislava: INFOSTAT.

Štatistické zdroje

ŠTATISTICKÝ ÚRAD SR <www.slovak.statistics.sk>

INFOSTAT SR < http://sodb.infostat.sk>

Predikcia poradia v ankete Športovec roka 2016 lineárnou regresiou
The prediction of results in Sportsman of 2016 poll using linear regression

Michaela Mihoková
Univerzita Mateja Bela, Fakulta prírodných vied, magisterské štúdium, 2. ročník
Matej Bel University, Faculty of Natural Sciences
micha.miho@gmail.com

Abstrakt: *Cieľom tohto príspevku je predikovať umiestnenie najlepších slovenských športovcov v ankete* Športovec roka 2016 *rôznymi regresnými modelmi. Na predikciu výsledkov je použitá jednoduchá lineárna regresia a zovšeobecnené lineárne modely. Predikované poradia na základe jednotlivých modelov sú porovnávané medzi sebou a tiež so skutočnými výsledkami. Všetky výpočty a grafy boli urobené výpočtovým softvérom R.*

Abstract: *The main purpose of this paper is to predict the final results of Sportsman of 2016 poll using various regression models. To predict the results a simple linear regression is used together with generalized linear models. The predictions based on these models are compared to one another and also to the real results. All calculations and graphs were made with statistical computing software R.*

Kľúčové slová: *Lineárna regresia, lineárny model, poissonovská distribúcia, predikcia, šport.*
Key words: *Linear regression, linear model, Poisson distribution, prediction, sport.*

1 Úvod

Rok 2016 bol pre slovenských športovcov mimoriadne vydarený. Začal sa skvelými výkonmi lyžiarky Veroniky Velez-Zuzulovej, ktorej úspešne sekundovala Petra Vlhová, a 2. miestom motocyklistu Štefana Svitka na Rely Dakar. Pokračoval splnením sna slovenskej futbalovej reprezentácie, ktorá si prvýkrát v ére samostatnosti zahrala na Majstrovstvách Európy vo futbale, kde sa mohla spoľahnúť najmä na fenomenálne kľučky a prihrávky Mareka Hamšíka. V lete sa slovenskí fanúšikovia tešili zo zisku prvej zlatej olympijskej medaily v atletike, ktorú vybojoval v chôdzi na 50 km Matej Tóth. Medailovú bilanciu našej krajiny na Olympijských hrách v Riu vylepšili aj vodní slalomári, keď Matej Beňuš bol strieborný v kategórii C1 a bratranci Ladislav a Peter Škantárovci dokonca zlatí v C2. Na jeseň Slováci zažiarili na majstrovstvách sveta - Peter Sagan obhájil dúhový dres v cyklistike a Dominika Cibulková ovládla šampionát v tenise. Hokejoví fanúšikovia si zas celú sezónu vychutnávali rýchlosť, obratnosť a góly Mariana Hossu.

Tým, ktorý z týchto fantastických športovcov bol v minulom roku najlepší, sa zaoberá tradičná slovenská anketa Športovec roka. Minulý rok v nej hlasovalo 122 športových odborníkov. Každý z nich si zvolil svoju desiatku najlepších,

pričom prvému vo svojom poradí (teda najlepšiemu športovcovi podľa hlasujúceho) pridelil 10 bodov, druhému 9, ..., desiaty športovec v poradí hlasujúceho získal 1 bod. Nakoniec sa pre každého športovca, ktorý sa objavil aspoň na jednom hlasovacom lístku, spočítali jemu pridelené body. Z takéhoto hlasovania vzišla spomenutá desiatka najlepších. Titul Športovec roka tak získal Matej Tóth, za ktorým o 75 hlasov zaostal Peter Sagan.

Je naozaj olympijská medaila viac než titul majstra sveta a výborné výsledky počas celej sezóny? Ak áno, prečo sa potom Škantárovci umiestnili v ankete až na štvrtom mieste, keď ich predbehla aj ďalšia majsterka sveta, Cibulková?

To, ako objektívne pristupovali k hlasovaniu odborníci, si musia zodpovedať oni sami. Čo však môžeme urobiť my, je pozrieť sa na túto situáciu z pohľadu matematiky. Cieľom tohto príspevku je predikcia poradia minuloročných najlepších slovenských športovcov v uvedenej ankete pomocou regresných modelov. Použijeme na to výsledky ankety z prechádzajúcich desiatich rokov a prediktory zahŕňajúce nielen športové úspechy, ale aj popularitu či vek športovca.

Najprv budeme predpovedať výsledky pomocou jednoduchej lineárnej regresie. Neskôr sa pozrieme na dáta z iného pohľadu, a síce použitím zovšeobecnených lineárnych modelov. Nakoniec jednotlivé predikcie porovnáme s reálnymi výsledkami, aby sme zistili, ktorý model predpovedal výsledky najlepšie. Poznamenajme, že všetky použité modely boli zostavené pred vyhlásením výsledkov ankety, teda neboli vykonávané žiadne umelé zásahy.

2 Odozva a prediktory

Umiestnenie v ankete Športovec roka 2016 budeme predpovedať na základe výsledkov ankety v rokoch 2006 až 2015. V každom z týchto rokov sa pozrieme na bodové zisky desiatich najlepších športovcov, čo bude odozva v modeloch.

Prediktory modelov by mali zohľadniť najmä športové úspechy. Športovec roka by mal byť predsa ten, ktorému sa v danom roku darilo najviac. Najväčšími športovými sviatkami sú olympijské hry a majstrovstvá sveta. Na vyjadrenie toho, ako sa darilo športovcom práve na týchto podujatiach, použijeme premenné *OH* a *MS*. Prvá premenná nadobúda hodnoty 0 až 5, kde 5 je za prvé miesto na olympijských hrách v danom roku, 4 za druhú priečku, a tak ďalej až po hodnotu 1 za piate miesto a hodnotu 0 za horšie umiestnenie. Nula bude pridelená aj v prípade, že sa v príslušnom roku olympijské hry nekonali. Druhá premenná nadobúda hodnoty 0 až 8, pričom ich interpretácia je podobná ako v prechádzajúcom prípade. Tu sa hodí bodovanie až do 8, pretože v mnohých športoch (plávanie, atletika,...) sa do finále dostáva práve osmička najlepších. Ostatné športové úspechy v príslušnom roku sú zohľadnené

v premennej *ine*, ktorá bude mať hodnotu 0 za veľmi slabé ďalšie výsledky až hodnotu 5 za výborné výsledky v danej sezóne.

Športy sa dajú rozdeliť na letné a zimné alebo na individuálne a kolektívne, preto by bolo vhodné mať v modeloch prediktory zahŕňajúce práve tieto informácie. Budú to teda dummy premenné a nazveme ich *leto* a *kolektiv*. V prípade, že šport, v ktorom je športovec úspešný, je letný, resp. kolektívny, má premenná *leto*, resp. *kolektiv*, hodnotu 1. V opačných prípadoch nadobúda 0.

Môže sa stať, že v nejakom roku dosiahnu viacerí športovci podobné úspechy a zároveň robia rovnaký typ športu. Bolo by preto dobré použiť ďalšie premenné na ich odlíšenie. Napríklad *vek*, čo poslúži aj na zistenie toho, ktorí športovci dosahujú lepšie výsledky - či mladší, či starší. Hodnoty tohto prediktora budú rovnaké ako vek športovca v danom roku. Ďalšou premennou slúžiacou na odlíšenie bude *pohlavie* nadobúdajúce hodnotu 0 pri mužoch a 1 pri ženách. Táto premenná zároveň slúži na zodpovedanie otázky, či hlasujúci odborníci preferujú vo významnej miere jedno pohlavie pred druhým. V sledovanom období sa totiž športovcom roka stal sedemkrát muž a iba trikrát žena. Podobne, aj celkové zastúpenie mužov v uvedených desiatich rokoch je takmer 70% ku iba 30% žien.

Poslednou premennou v našich modeloch je *popularita*. Budeme tak vidieť, ako veľmi vplýva obľúbenosť u ľudí a frekvencia výskytu v médiách na bodové zisky v ankete. V prípade veľmi populárneho športovca má táto premenná hodnotu 2, v prípade takmer žiadnej popularity má hodnotu 0. Za priemernú obľúbenosť je pridelená 1.

3 Lineárne modely

Na predikciu výsledkov ankety najprv použijeme obyčajnú lineárnu regresiu. Z nej vychádzajúci model nemá zložitú interpretáciu, preto sa z neho ľahko môžeme dozvedieť veľa užitočných informácií.

Východiskovým modelom je jednoduchý lineárny model so všetkými premennými uvedenými v prechádzajúcej časti. Predpis takéhoto modelu je uvedený rovnosťou (1), kde dolný index *i* vyjadruje poradie pozorovanej osoby v dátach, teda nadobúda celé čísla od 1 po 100 (pracujeme s najlepšou desiatkou športovcov v každom z desiatich rokov). Napríklad hodnota premennej vek_{33} je rovná veku 33. športovca v dátach. Konštanty β_j, kde *j* je celé číslo od 0 po 8, vyjadrujú optimálne hodnoty v zmysle minimalizovania štvorcov odchýliek ε_i.

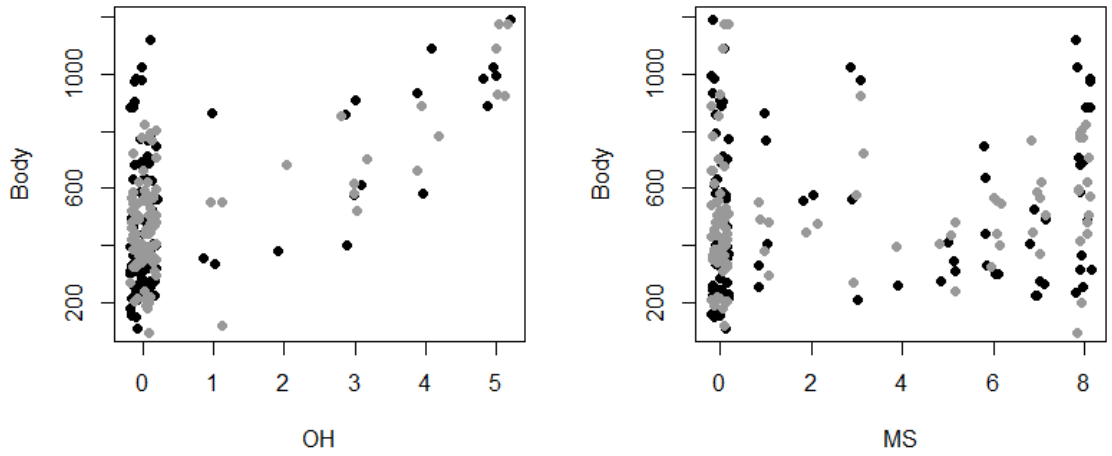
$$body_i = \beta_0 + \beta_1 vek_i + \beta_2 pohlavie_i + \beta_3 OH_i + \beta_4 MS_i + \beta_5 ine_i + \\ + \beta_6 leto_i + \beta_7 kolektiv_i + \beta_8 popularita_i + \varepsilon_i \qquad (1)$$

V takto zostavenom modeli sa prediktor *leto* ukázal ako výrazne nesignifikantný, preto ho vynecháme a zostavíme nový model, ktorého predpis je podobný ako v predchádzajúcom prípade, no člen s indexom 6 je vynechaný. Nesignifikantná je aj konštanta β_0, avšak tú v modeli z dôvodu zachovania konvencií ponecháme[1]. Ukazuje sa, že po vynechaní pozorovaní podozrivých na základe diagnostických grafov sa hodnoty parametrov veľmi nezmenia. Zostaneme teda pri modeli so siedmimi parametrami, ktorých hodnoty sú uvedené v prvom stĺpci Tabuľky 1.

Tab. 3 Hodnoty parametrov *(Zdroj: vlastné spracovanie)*

Parameter	Lineárny model	Poissonove rozdelenie	Negatívne binomické rozdelenie
$\boldsymbol{\beta_0}$	100.30	5.36	4.95
$\boldsymbol{\beta_1}$	-6.59	-0.01	-----
$\boldsymbol{\beta_2}$	-102.99	-0.20	-0.16
$\boldsymbol{\beta_3}$	111.59	0.17	0.19
$\boldsymbol{\beta_4}$	22.55	0.04	0.04
$\boldsymbol{\beta_5}$	96.16	0.20	0.20
$\boldsymbol{\beta_6}$	-----	-0.08	-----
$\boldsymbol{\beta_7}$	-114.23	-0.24	-0.19
$\boldsymbol{\beta_8}$	113.59	0.21	0.23

Z Tabuľky 1 možno vyčítať napríklad to, že v prípade dvoch športovcov, ktorí sa v sledovaných hodnotách líšia len vekom, by podľa lineárneho modelu získal viac bodov mladší z nich. Zaujímavé je aj to, že tento model predikuje o takmer 103 bodov menej pre ženu, ktorá je tak úspešná i populárna ako muž v jej veku robiaci rovnaký šport z hľadiska individuálnosti.



Graf 2 Porovnanie skutočných bodov (čierna farba) a odhadnutých (sivá farba) v závislosti od premennej OH, resp. MS (*Zdroj: výpočtový softvér R*)

[1] Konštanta β_0 bude z rovnakého dôvodu ponechaná aj vo všetkých ďalších modeloch bez ohľadu na jej signifikanciu.

Ďalej sa môžeme pozrieť na to, ako blízko sú skutočné a odhadnuté hodnoty premenných *OH* a *MS* (Graf 1). Z grafov vidno, že pre tých, ktorí sa stali majstrami sveta, model predpovedá menší bodový zisk než v skutočnosti dosiahli, kým pre olympijských víťazov sú skutočné a odhadnuté hodnoty bližšie. Znamená to, že takýto lineárny model znevýhodňuje víťazov MS pred víťazmi OH. Taktiež vidno, že lineárna závislosť je pre premennú OH naozaj dosť rozumná (až na to množstvo núl).

Lineárnym modelom predikované bodové zisky najlepšej desiatky v ankete Športovec roka 2016 možno usporiadať, na základe čoho sa zostaví poradie, ktoré je uvedené v Tabuľke 2. Vidno z nej, že skutočné a odhadnuté výsledky sú celkom blízko, keďže poradie štyroch športovcov je odhadnuté úplne presne a ďalšie štyri miesta sú predikované len o jednu priečku nižšie či vyššie. Tento pomerne jednoduchý model teda odhaduje poradie dosť dobre.

Tab. 2 Predikované poradia a body *(Zdroj: vlastné spracovanie)*

Športovec	Skutočnosť		Lineárny model		Poissonove rozdelenie		Negatívne binomické rozdelenie	
	Poradie	Body	Poradie	Body	Poradie	Body	Poradie	Body
M. Beňuš	5.	619	5.	644	5.	503	5.	540
D. Cibulková	3.	934	4.	708	3.	699	3.	708
M. Hamšík	7.	320	9.	311	9.	310	9.-10.	327
M. Hossa	9.	237	10.	258	10.	306	9.-10.	327
P. Sagan	2.	1076	2.	817	1.	862	2.	830
Š. Svitko	8.	296	7.	375	8.	366	7.	384
L. a P. Škantárovci	4.	888	3.	723	4.	561	4.	653
M. Tóth	1.	1151	1.	860	2.	710	1.	844
V. Velez-Zuzulová	6.	585	6.	381	6.	409	6.	399
P. Vlhová	10.	172	8.	357	7.	380	8.	328

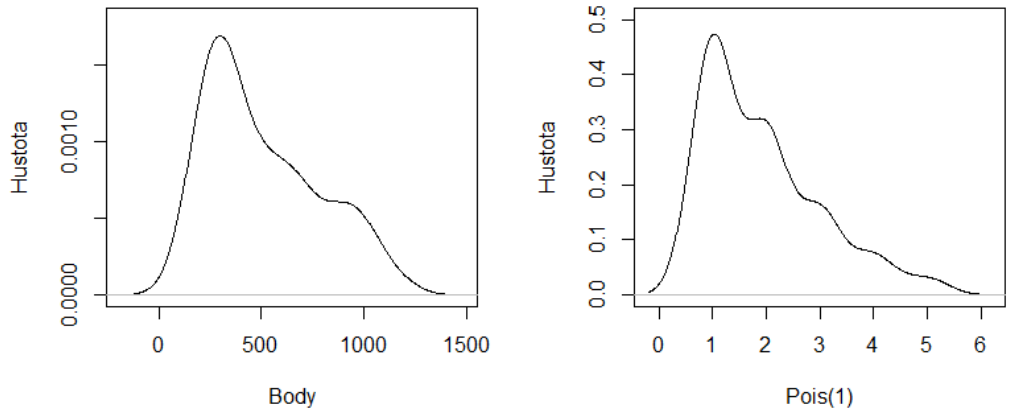
4 Zovšeobecnené lineárne modely

I keď je obyčajný lineárny model pomerne jednoduchý na interpretáciu, jeho nevýhodou je, že môže predpovedať aj záporné hodnoty. Odozvou v ankete sú však body, teda nezáporné celé čísla. Navyše, zovšeobecnené lineárne modely popisujú pravdepodobnostné rozdelenie odozvy presnejšie. V tejto časti sa preto pozrieme na dáta z iného pohľadu.

4.1 Poissonove rozdelenie

Na Grafe 2 je zobrazená hustota bodov z dát o športovcoch a hustota 100 náhodne vygenerovaných hodnôt z Poissonovho rozdelenia s parametrom 1. Tieto grafy sú si podobné, preto by pri predikovaní poradia v ankete Športovec

roka mohlo byť užitočné urobiť to pomocou zovšeobecneného lineárneho modelu s poissonovskou rodinou.

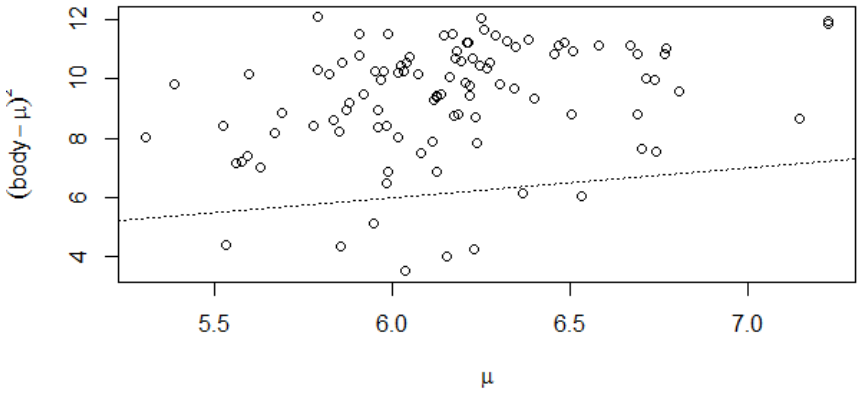


Graf 2 Vľavo hustota bodov z dát, vpravo hustota náhodného výberu 100 hodnôt z Poissonovho rozdelenia s parametrom 1 (*Zdroj: výpočtový softvér R*)

Predpis takéhoto modelu sa dá zapísať rovnosťou (2), pričom dolné indexy *i* aj parametre β_j, majú rovnakú interpretáciu ako v rovnosti (1). Takýto model uvažuje, že body prislúchajúce *i*-tej osobe v dátach majú Poissonove rozdelenie s parametrom μ_i.

$$\log(\mu_i) = \beta_0 + \beta_1 vek_i + \beta_2 pohlavie_i + \beta_3 OH_i + \beta_4 MS_i + \\ + \beta_5 ine_i + \beta_6 leto_i + \beta_7 kolektiv_i + \beta_8 popularita_i \qquad (2)$$

Po zostavení vyššie uvedeného modelu sa ukazuje, že všetky prediktory sú veľmi signifikantné, čo je podozrivé. Náhodná premenná s Poissonovým rozdelením má tú vlastnosť, že jej stredná hodnota a variancia sa rovnajú parametru rozdelenia. Graf 3 však ukazuje, že v takomto modeli to tak nie je, pretože pozorované hodnoty sú v oveľa väčšom počte nad diagonálou prvého kvadrantu, než pod ňou. Vyskúšajme preto ešte nejaké iné rozdelenie, napríklad negatívne binomické.



Graf 3 Vzťah medzi strednou hodnotou a varianciou (*Zdroj: výpočtový softvér R*)

Pre úplnosť informácií je v Tabuľke 2 uvedené aj poradie a body predikované zovšeobecneným lineárnym modelom s poissonovskou rodinou a v Tabuľke 1 ním odhadnuté koeficienty. Treba si však uvedomiť, že tu je interpretácia parametrov iná, než pri jednoduchom lineárnom modeli. V tomto prípade zmena prediktora vek_i o jeden rok vynásobí priemerný bodový zisk μ_i faktorom e^{β_1} za predpokladu fixovaných ostatných premenných (analogicky pre ďalšie prediktory). Všimnime si, že znamienka všetkých koeficientov sú pri oboch doterajších modeloch rovnaké, teda v oboch prípadoch má ľubovoľná zmena hociktorého prediktora rovnaký vplyv na odozvu z pohľadu jej zvýšenia, respektíve zníženia.

4.2 Negatívne binomické rozdelenie

Negatívne binomické rozdelenie je bohatšie rozdelenie než Poissonove, ktoré je jeho limitným prípadom. Takéto rozdelenie nemá strednú hodnotu a varianciu rovné tomu istému parametru, ale k variancii je pripočítaný nejaký násobok druhej mocniny strednej hodnoty. Tento násobok sa nazýva disperzný parameter.

Predpis zovšeobecneného lineárneho modelu s negatívne binomickou rodinou je uvedený rovnosťou (3), kde *k* je prevrátená hodnota disperzného parametra a μ_i je stredná hodnota predikovaných bodov pre *i*-te pozorovanie.

$$\log\frac{\mu_i}{\mu_i + k} = \beta_0 + \beta_1 vek_i + \beta_2 pohlavie_i + \beta_3 OH_i + \beta_4 MS_i + \\ + \beta_5 ine_i + \beta_6 leto_i + \beta_7 kolektiv_i + \beta_8 popularita_i \tag{3}$$

Použitím tohto modelu na naše dáta sa výrazne nesignifikantnými javia prediktory *vek* a *leto*. Koeficienty modelu zostrojenom ich vynechaním sú uvedené v poslednom stĺpci Tabuľky 1. Môžeme vidieť, že znamienka koeficientov sú opäť pri všetkých premenných rovnaké ako v prechádzajúcich modeloch. V posledných dvoch stĺpcoch Tabuľky 2 nájdeme predikované poradie a body. Tento model predpovedá až sedem miest úplne presne, čo je najviac spomedzi všetkých použitých modelov. Posledné dve miesta sú však delené, pretože športovci na nich umiestnení majú hodnoty všetkých použitých premenných rovnaké. Ak by sme na ich odlíšenie pridali späť do modelu prediktor *vek*, poradie by bolo rovnaké ako pri modeli s poissonovskou rodinou.

5 Záver

Cieľom tohto príspevku bolo predikovať poradie desiatich najlepších športovcov v ankete Športovec roka 2016 použitím regresných modelov. Najprv sme uviedli použité regresory. Potom sme pomocou nich zostavili jednoduchý lineárny model a neskôr zovšeobecnené lineárne modely s Poissonovým a negatívnym binomickým rozdelením.

Všetky zostrojené modely predikovali umiestnenia blízke realite. Najbližšie skutočným výsledkom bol model s negatívnym binomickým rozdelením, ktorý až sedem miest predpovedal úplne presne. Spomedzi zostavených modelov je však najzložitejší na interpretáciu. Naopak, najjednoduchším použitým modelom je obyčajný lineárny model. Ten presne predikoval štyri miesta, no ani zvyšných šesť priečok nebolo úplne ďaleko od skutočnosti. V príspevku je okrem týchto dvoch modelov použitý ešte zovšeobecnený lineárny model s Poissonovou distribúciou, ktorý pridelil piatim športovcom tie isté miesta, na akých naozaj skončili. Problémom tohto modelu je však nesplnenie predpokladu o vzťahu medzi strednou hodnotou a varianciou odhadovaných bodov.

Okrem použitých modelov by mohlo byť užitočné modelovanie pomocou tzv. zero-inflated modelu, nakoľko sa v dátach vyskytuje veľa núl. Taktiež by sme mohli dáta prerobiť na panelové, no vzhľadom na veľkosť vzorky by to zrejme nepomohlo.

Zaujímavé je, že koeficient pri každej premennej má rovnaké znamienko pre všetky použité modely, čo vypovedá o kvalitatívne rovnakom vplyve prediktora na zisk bodov v ankete. Z tohto hľadiska stojí za povšimnutie najmä skutočnosť, že premenná hovoriaca o pohlaví športovca bola v každom modeli signifikantná, teda to, či je športovec muž alebo žena, má zrejme vplyv na bodový zisk v ankete. Ak by v nej boli žena a muž s rovnakými hodnotami všetkých ostatných prediktorov, viac bodov by podľa každého modelu získal muž. Treba si však uvedomiť, že muži majú v našich dátach oveľa vyššie zastúpenie ako ženy a taktiež to, že môže existovať množstvo premenných korelovaných s pohlavím, ktoré nekontrolujeme, no majú vplyv na odozvu.

Spomedzi sledovaných premenných sa najmenej významnou ukázala byť tá, ktorá hovorí o tom, či sa športovec venuje letnému alebo zimnému športu. Veľký vplyv na zisk bodov zas mali premenné týkajúce sa športových úspechov, čo je tiež ukazovateľom správnosti modelov. Nezanedbateľnú rolu pri bodovaní má však aj popularita športovcov. Aj tento fakt vysvetľuje, prečo športovci, ktorí v roku 2016 získali zlaté olympijské medaily, neskončili v ankete tesne za sebou, ale sa medzi nich dostali majstri sveta vo svojich disciplínach.

Výsledky ankety boli pre mnohých redaktorov i bežných fanúšikov športu prekvapivé. My však vidíme, že z pohľadu regresie až o také veľké prekvapenia nešlo.

6 Literatúra

Faraway, J.J. (2006). *Extending the Linear Model with R.* Boca Raton: Taylor & Francis Group, 345p.

Faraway, J.J. (2015). *Linear Models with R.* Boca Raton: Taylor & Francis Group, Second Edition, 2015, 284p.

www.sportovecroka.eu/sk/Sportovec-roka.html

Výpočtová štatistika 2017

zborník abstraktov z 26. medzinárodného vedeckého seminára **Výpočtová štatistika 2017**
a zborník príspevkov z konferencie **Prehliadka prác mladých štatistikov a demografov 2017**
konaných v Bratislave v dňoch 7. – 8. decembra 2017

Editori: Iveta Stankovičová, Janka Medová

Vydala Slovenská štatistická a demografická spoločnosť, Bratislava v roku 2017.
Prvé vydanie

ISBN 978-80-88946-79-3



SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies. © 2015 SAS Institute Inc. All rights reserved. S138136US.0315

ISBN 978-80-88946-79-3