Slovenská štatistická a demografická spoločnosť



Výpočtová štatistika 2018

zborník abstraktov z 27. medzinárodného vedeckého seminára
Výpočtová štatistika 2018
a zborník príspevkov z konferencie
Prehliadka prác mladých štatistikov a demografov 2018
konaných v Bratislave v dňoch 6. – 7. decembra 2018

2018



Slovenská štatistická a demografická spoločnosť

Výpočtová štatistika 2018

zborník abstraktov z 27. medzinárodného vedeckého seminára

Výpočtová štatistika 2018

a zborník abstraktov a článkov z konferencie

Prehliadka prác mladých štatistikov a demografov 2018

konaných v Bratislave v dňoch 6. – 7. decembra 2018

Bratislava 2018

Výpočtová štatistika 2018

zborník abstraktov z 27. medzinárodného vedeckého seminára **Výpočtová štatistika 2018** a zborník abstraktov a článkov z konferencie **Prehliadka prác mladých štatistikov a demografov 2018** konaných v Bratislave v dňoch 6. – 7. decembra 2018

Editori: Iveta Stankovičová, Janka Medová

Vydala Slovenská štatistická a demografická spoločnosť, Bratislava, 2018.
Text neprešiel jazykovou úpravou. Za obsah príspevkov zodpovedajú jednotliví autori.

ISBN 978-80-88946-83-0

Obsah

Časť 1

Zborník abstraktov z medzinárodného seminára
Výpočtová štatistika 2018

Pokroky v modelování ekonomických dat pomocí spojitých pravděpodobnostních rozdělení
Advances in Modeling Economic Data Using Continuous Probability Distributions
Diana Bílková 6

Rebalansovanie založené na investičnom štýle: poznatky pre malého investora
Investment-Style Based Rebalancing: Lessons Learned for a Small Investor
Martin Boďa, Mária Kanderová 7

Predikcia finančného zdravia vybratých podnikov v Slovenskej republike a Poľsku
Financial Distress Prediction of Selected Companies in Slovak Republic and Poland
Mária Danišovská, Mária Stachová, Lukáš Sobíšek 8

Důchodový věk na Slovensku a v Česku stanovený na základě relativního prospektivního věku
Retirement Age in Slovakia and Czechia Based on Relative Prospective Age Concept
Tomáš Fiala, Jitka Langhamrová 9

Analýza rozvodovosti v SR viacrozmernými štatistickými metódami
Analysis of Divorce rate in Slovakia by Multivariate Statistical Methods
Alžbeta Garajová 10

Príjmová nerovnosť a riziko chudoby v krajinách V4
Income Inequality and Poverty Risk in V4 Countries
Lucia Janovičová, Ľubica Bartová 11

Porovnání výskytů závažných onemocnění v rámci krajů v České republice
Comparison of Incidences of Serious Diseases within Regions in the Czech Republic
Lucie Kopecká, Viera Pacáková 12

Štatistické mapy a atlasy
Statistical Maps and Atlases
Dagmar Kusendová 13

Analýza chování koeficientů pro stanovení počtu shluků ve shlukové analýze
Analysis of the Behavior of Coefficients for Determining the Number of Clusters in Cluster Analysis
Tomáš Löster, Jakub Danko 14

Determinanty chudoby na Slovensku
Determinants of Poverty in Slovakia
Alena Mojsejová 15

Použití výstupního systému SAS pro vytváření pracovních listů aplikace Microsoft Excel
Using the SAS Output Delivery System to Create Microsoft Excel Worksheets
Roman Pavelka 16

Identifikácia a použitie metód hĺbkovej analýzy údajov na maximalizáciu ziskovosti retenčnej kampane
Identification and Application of Data Mining Methods in Order to Maximize the Profitability of a Retention Campaign
Iveta Stankovičová, Daniel Gábriš 17

Zhlukovanie časových radov dlhodobých úrokových mier
Clustering Time Series of Long-Term Interest Rates
Beáta Stehlíková 18

Analýza doživotných dôchodkov vyplácaných zo starobného dôchodkového sporenia v Slovenskej republike
Analysis of Whole Life Annuities Paid out from the Old-Age Pension Scheme in the Slovak Republic
Gábor Szűcs 19

Alternatívne prístupy k identifikácii subjektívne chudobných
Alternative Approaches to the Identification of the Subjectively Poor
Tomáš Želinský, Martina Mysíková 20

Časť 2

Zborník abstraktov a článkov z konferencie Prehliadka prác mladých štatistikov a demografov 2018

Príjmová nerovnosť vo vybraných členských krajinách EÚ
Income Inequality in Selected EU Countries
Lucia Janovičová 22

Predikcia výskytu rakoviny pľúc
Prediction of Lung Cancer Diagnosis
Lucia Plichtová 23

Trendy migračných smerov mesta Skalica
Trends of Migration Flows in Skalica Town
Nikoleta Slavíková 37

Modelovanie slovenských parlamentných volieb
Modelling of Slovak Parliamentary Elections
Dominika Zorvanová 47

Časť 1

Výpočtová štatistika 2018

Zborník abstraktov

Pokroky v modelování ekonomických dat pomocí spojitých pravděpodobnostních rozdělení
Advances in Modeling Economic Data Using Continuous Probability Distributions

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

Abstrakt: *Tento článek se zabývá pokročilými nástroji pro modelování ekonomických dat s využitím robustních metod bodového odhadu parametrů. Představeny jsou výhody využití metody L-momentů oproti běžné metodě využívající klasických centrálních momentů. Článek se dále zabývá rovněž výhodami použití metody TL-momentů a metody LQ-momentů oproti použití původní metody L-momentů. TL-momenty a LQ-momenty představují alternativní robustní verze L-momentů. Vydatnost TL-statistik a LQ-statistik však často předčí vydatnost klasických L-statistik.*

Abstract: *This paper deals with advanced tools for modeling economic data using robust methods of point estimation of parameters. The advantages of using the L-moment method versus the conventional method using classical central moments are presented. The paper also deals with the advantages of using the TL-moment method and LQ-moment method versus the use of the original L-moment method. TL-moments and LQ-moments represent an alternative robust version of L-moments. However, the efficiency of TL-statistics and LQ-statistics often exceeds the efficiency of standard L-statistics.*

Kľúčové slová: *Vydatnost statistik, L-momenty pravděpodobnostních rozdělení, TL-momenty pravděpodobnostních rozdělení, LQ-momenty pravděpodobnostních rozdělení.*

Key words: *Efficiency of statistics, L-moments of probability distribution, TL-moments of probability distribution, LQ-moments of probability distribution.*

Rebalansovanie založené na investičnom štýle: poznatky pre malého investora
Investment-Style Based Rebalancing: Lessons Learned for a Small Investor

Martin Boďa, Mária Kanderová
Univerzita Mateja Bela v Banskej Bystrici, Ekonomická fakulta, Tajovského 10, 975 90 Banská Bystrica, Slovenská republika
Matej Bel University in Banská Bystrica, Faculty of Economics, Tajovského 10, 975 90 Banská Bystrica, Slovak Republic
martin.boda@umb.sk, maria.kanderova@umb.sk

Abstrakt: *Prezentácia poskytuje prehľad výsledkov získaných počas riešenia projektu VEGA č. 1/0554/16. S dôrazom na i.) elimináciu obmedzení konvenčných a alternatívnych prístupov k trackovaniu portfólia a ii.) formuláciu vhodnej trackovacej stratégie rešpektujúcej nevyhnutnosť častého rebalansovania portfólia pri transakčných nákladoch, prezentácia sumarizuje základné výsledky projektu a načrtá odporúčania pre teóriu a prax finančného investovania. Navrhnuté odporúčania sa týkajú situácie malého investora, ktorý presadzuje pasívny prístup k výberu portfólia a musí urobiť viacero rozhodnutí od predvýberu aktív až po monitorovanie portfólia. Pokiaľ ide o počiatočný krok, sú skúmané rozličné skrínovacie prístupy k predvýberu kandidátskych aktív, a v rámci posledného kroku sa rozhodujú efekty periodických a intervenčných rebalančných stratégií.*

Abstract: *The presentation gives an overview of the results attained in the process of executing project VEGA # 1/0554/16. Putting an emphasis upon i) elimination of shortcomings of conventional and alternative approaches to portfolio tracking, and ii) formulation of a suitable tracking strategy respecting a necessity of frequent portfolio rebalancing under transaction costs, the presentation summarizes the main results of the project and draws recommendations for the theory and practice of financial investing. The suggested recommendations relate to a situation of a small investor who follows a passive approach to portfolio selection and must make a variety of choices starting with asset pre-selection and ending with portfolio monitoring. In the initial step, diverse screening approaches to pre-selecting candidate assets are explored based on market capitalization and price multipliers, whereas the final step distinguishes the effects of periodic and threshold rebalancing strategies.*

Kľúčové slová: *Rebalansovanie, investičný štýl, transakčné náklady, trackovacie portfólio, výkonnosť.*

Key words: *Rebalancing, investment style, transaction costs, tracking portfolio, performance.*

Predikcia finančného zdravia vybratých podnikov v Slovenskej republike a Poľsku
Financial Distress Prediction of Selected Companies in Slovak Republic and Poland

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

Abstrakt: *Keďže finančné problémy podnikov môžu prerásť do obrovských rozmerov a spôsobiť tak veľké straty či ovplyvniť fungovanie viacerých subjektov z blízkeho ale aj vzdialeného okolia podniku, je dôležité s dostatočnou presnosťou a predstihom finančnú tieseň rozpoznať, predpovedať a následne sa pokúsiť jej zabrániť. Preto hlavným cieľom tohto príspevku je vytvorenia mechanizmu, ktorý by dokázal podnik upozorniť na blížiace sa finančné ohrozenie. Príspevok sa zaoberá možnosťami predikcie finančného zdravia vybratých podnikov Slovenskej republiky a Poľska, pričom sa porovnávajú vytvorené matematicko-štatistické modely (založené na stromových štruktúrach) aplikované na viacročné reálne dáta tvorené finančnými indikátormi podnikov z pohľadu presnosti ich predikcie.*

Abstract: *Financial problems of companies can lead into huge distress and caused big loss or can have negative impact on many subjects from companies' environments. Thus, it is important to recognize, predict and identify the financial distress with high accuracy and try to avoid this situation. The main goal of this contribution is to create a mechanism that is able to notice the financial menace. The study shows possibilities for financial distress prediction of selected companies of Slovakia and Poland and also it compares estimated mathematical statistical models (tree-based) that are built on real data sets consist of financial indicators.*

Kľúčové slová: *stromové štruktúry, predikčné modely, finančné zdravie podnikov, finančné indikátory*

Key words: *tree-based models, prediction models, financial health of companies, financial indicators*

Důchodový věk na Slovensku a v Česku stanovený na základě relativního prospektivního věku
Retirement Age in Slovakia and Czechia Based on Relative Prospective Age Concept

Tomáš Fiala, Jitka Langhamrová
Katedra demografie, fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze, Česká republika
Department of Demography, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic
fiala@vse.cz, langhamj@vse.cz

Abstrakt: *V Evropě probíhá již několik desítek let stárnutí populace. To má mimo jiné za následek prodlužování průměrné doby pobírání starobních důchodů, což může vést k ohrožení finanční udržitelnosti důchodových systémů. V řadě zemí Evropy proto proběhly nebo probíhají reformy penzijního systému, jejichž součástí je často zvyšování důchodového věku, někdy i nad obvyklou hranici 65 let.*

Objevuje se pochopitelně požadavek, aby hranice důchodového věku byla nějakým způsobem svázána s očekávanou délkou života. Jedním z možných přístupů je určení hranice důchodového věku na základě myšlenky tzv. relativního prospektivního věku, tj. na základě průměrného očekávaného podílu zbývající délky života. Článek obsahuje definici a metodologii výpočtu prospektivního věku a relativního prospektivního věku a dále pak výpočty a odhady vývoje hodnot relativního prospektivního věku na Slovensku a v Česku v období 1960–2080. Je zde rovněž návrh zvyšování důchodového věku v uvedených zemích po roce 2030vycházející z uvedeného principu.

Abstract: *In many European countries there is going on population ageing and there are intensions to rise the retirement age. The reason for this is the rise in life span and increase the period of old-age pension receipt, which can cause financial unsustainability of pension systems. A possible adjustment of the retirement age to the level of mortality can be based on the concept of so called relative prospective age: the ratio of the remaining life expectancy to the total expected life span. This paper brings methodological background of definition of relative prospective age and computations of its value for Slovakia and Czechia since 1960 until 2080. Recommended increase of the retirement age in countries mentioned after the year 2030 based on this principle is shown.*

Klíčová slova: *prospektivní věk, relativní prospektivní věk, důchodový věk, Slovensko, Česko*

Key words: *prospective age, relative prospective age, retirement age, Slovakia, Czechia*

Analýza rozvodovosti v SR viacrozmernými štatistickými metódami
Analysis of Divorce Rate in Slovakia by Multivariate Statistical Methods

Alžbeta Garajová
Univerzita Komenského v Bratislave, Prírodovedecká fakulta, Katedra humánnej geografie a demografie, Slovenská republika
Comenius University in Bratislava, Faculty of Natural Sciences, Department of Human Geography and Demography, Slovak Republic
garajova29@uniba.sk

Abstrakt: *Rozvodovosť ako demografický proces je jeden z najmenej samostatne analyzovaných procesov v slovenskej demografii. Rovnako ojedinelá je v demografii aj problematika využitia viacrozmerných štatistických metód. Práve z tohto dôvodu si v príspevku dávame za cieľ zhodnotiť trend rozvodovosti za posledných 20 rokov v Slovenskej republike. Taktiež by sme sa v danej analýze radi zamerali aj na zhodnotenie súčasného stavu rozvodovosti na Slovensku prostredníctvom vybraných viacrozmerných štatistických metód. Na analýzu súčasných trendov rozvodovosti sme použili faktorovú a zhlukovú analýzu. Metódy sme aplikovali na dáta okresnej rozvodovosti SR z roku 2017. Na základe faktorovej analýzy je možné určiť 3 až 4 najvýznamnejšie faktory, ktoré sme ďalej uplatnili aj v zhlukovej analýze. Všetky výpočty sme uskutočnili pomocou systému SAS Enterprise Guide.*

Abstract: *Divorce as a demographic process is one of the least independently analyzed processes in Slovak demography. Equally unique in demography is the use of multivariate statistical methods. Consequently, we are trying to evaluate the trend of divorce in the last 20 years in the Slovak Republic. We would also like to focus on analyzing the current state of divorce in Slovakia through selected multivariate statistical methods. To analyze current trends in divorce we used factor and cluster analysis. The methods have been applied to divorces at district level of Slovak Republic in year 2017. Based on the analysis of the factors, it is possible to determine 3 or 4 the most important factors that we used in cluster analysis. All calculations were made by using Enterprise SAS.*

Kľúčové slová *demografia, rozvodovosť, sobášnosť, Slovenská republika, viacrozmerné štatistické metódy*

Key words *demography, divorces, marriages, Slovak Republic, multivariate statistical methods*

Príjmová nerovnosť a riziko chudoby v krajinách V4
Income Inequality and Poverty Risk in V4 Countries

Lucia Janovičová, Ľubica Bartová
Katedra štatistiky a operačného výskumu, Fakulta ekonomiky a manažmentu, Slovenská poľnohospodárska univerzita v Nitre, Slovenská republika
Department of Statistics and Operations Research, Faculty of Economics and Management, Slovak University of Agriculture in Nitra, Slovak Republic
lubica.bartova@uniag.sk, xjanovicova@uniag.sk

Abstrakt: *Cieľom príspevku bolo hodnotenie vývoja príjmovej nerovnosti, miery rizika chudoby v jednotlivých krajinách V4 v rokoch 2005 - 2017. Použili sme panel ročných údajov 13 ukazovateľov (Eurostat, 2018). Vplyv vybraných faktorov na príjmovú nerovnosť a mieru rizika chudoby sme odhadli ekonometrickými modelmi. Príjmová nerovnosť a miera rizika chudoby v PL boli v sledovanom období najvyššie ale klesajúce. Ich vývoj v SK a v CZ sa výrazne nelíšil a ich úroveň bola relatívne nižšia ako v PL a HU. Príjmovú nerovnosť v krajinách V4 (so aj bez zohľadnenia sociálnych transferov) a mieru rizika chudoby znižoval rast HDP na obyvateľa. Sociálne transfery prispeli k zníženiu príjmovej nerovnosti v HU. Podiel obyvateľov vo veku 65 rokov a viac mal štatisticky významný pozitívny vplyv na rast miery rizika chudoby celej a ženskej populácie. V PL a HU bolo riziko chudoby štatisticky preukazne vyššie v porovnaní s CZ a SK.*

Abstract: *The aim of the contribution was to assess the development of income inequality, the poverty risk rate in V4 countries over the years 2005-2017. We used the panel of annual data of 13 indicators (Eurostat, 2018). The impact of selected factors on income inequality and at the risk of poverty rate was estimated by econometric models. The income inequality and at the risk of poverty rate in PL were the highest but declining over observed period. Their development in SK and CZ did not differ significantly and their level was relatively lower than in PL and HU. Reduction of income inequality in the V4 countries (both, with and without taking into account social transfers) and at the risk of poverty rate by GDP growth per capita was statistically significant. Social transfers have contributed to income inequality decline in HU. The share of the population aged 65 and over, had a statistically significant positive effect on the at the risk of poverty rate growth of the entire and female population. In PL and HU, the at the risk of poverty rate was significantly higher than in CZ and SK.*

Kľúčové slová*: príjmová nerovnosť, miera rizika chudoby, panelové údaje, krajiny V4*

Key words: *income inequality, risk of poverty rate, panel data, V4 countries*

Porovnání výskytů závažných onemocnění v rámci krajů v České republice
Comparison of Incidences of Serious Diseases within Regions in the Czech Republic

Lucie Kopecká, Viera Pacáková
Univerzita Pardubice, Fakulta ekonomicko-správní, Katedra matematiky a kvantitativních metod, Studentská 95, 532 10 Pardubice, Česká republika
University of Pardubice, Faculty of Economics and Administration, Department of Mathematics and Quantitative Methods, Studentská 95, 532 10 Pardubice, Czech Republic
Lucie.Kopecka1@student.upce.cz; Viera.Pacakova@upce.cz

Abstrakt: *Výskyty závažných onemocnění mají rostoucí trend, oproti úmrtím na tato onemocnění. Tak jako existují rozdíly ve výskytech závažných onemocnění mezi jednotlivými státy, existují tyto rozdíly rovněž mezi regiony v rámci jednotlivých států. Rozdíly ve výskytech mohou být vyvolány mnoha faktory, a to přírodními, socio-ekonomickými, zdravotnickými, demografickými, etnickými aj. Je proto nutné tyto rozdíly sledovat v rámci jednotlivých států a přijímat odpovídající opatření. Cílem článku je porovnat výskyty závažných onemocnění v rámci krajů v České republice pomocí hybridního přístupu, který kombinuje metodu vícerozměrného škálování s lineárním uspořádáním, a to na datech získaných z ÚZIS ČR. Hybridní přístup je vhodný pro vizualizaci objektů a určení vzdálenosti od ideálního objektu na základě použitých dat.*

Abstract: *Incidences of serious diseases have growing trend in contrast with mortalities cause by these diseases. Just like differences caused by serious diseases exist among individual countries, these differences exist within individual countries as well. Differences in case of incidences could be caused by many factors such as natural, socio-economic, healthcare, demographic, ethnic and others. It is necessary to observe these differences within individual countries and to take appropriate measures. The main aim of article is to compare incidences of serious diseases within the regions in the Czech Republic by using hybrid approach which combine multidimensional scaling with linear ordering on data from ÚZIS ČR. Hybrid approach is suitable for visualisation objects and determination of distances from ideal object based on used data.*

Kľúčové slová: *Výskyty, závažná onemocnění, hybridní přístup.*

Key words: *Incidences, serious diseases, hybrid approach.*

Štatistické mapy a atlasy
Statistical Maps and Atlases

Dagmar Kusendová
Prírodovedecká fakulta Univerzity Komenského v Bratislave, Mlynská dolina, 84215 Bratislava 4, Slovenská republika
Comenius University Faculty of Natural Sciences, Mlynská dolina, 84215 Bratislava 4, Slovak Republic
kusendova@fns.uniba.sk

Abstrakt: *Štatistické atlasy sú významným nástrojom kartografickej vizualizácie výsledkov sčítania obyvateľstva. V minulosti, ale aj dnes predstavujú komplexný systémovo usporiadaný súbor zväčša analytických máp, ktoré prezentujú štatisticky spracované dáta zo sčítaní v rôznej územnej mierke. V ére digitálnej kartografie, geografických databáz, internetu a distribuovaných počítačových sietí sa však menia spôsoby tvorby týchto atlasov. Príspevok sa zaoberá zmenami a aktuálnym trendom atlasovej tvorby s dôrazom na vybrané aspekty v národnom a medzinárodnom rámci. Relatívna bohatosť česko-slovenskej atlasovej tvorby nás inšpirovala aj k analýze demograficky zameraných častí atlasov so zameraním sa na štyri národné atlasy vydané v rokoch 1935, 1966, 1980 a 2002, ktoré sme porovnali na základe vybraných kvantitatívnych a kvalitatívnych charakteristík zachycujúcu ich štruktúru, obsah, komponiku, mierkovú rozlíšiteľnosť máp a použité kartografické vyjadrovacie metódy. Cieľom je ukázať, v čom sú atlasy rovnaké, resp. rozdielne v demograficky zamerných kapitolách.*

Abstract: *Atlases, which are referred to statistical ones, are the important tools for cartographic visualization of the results gained from population censuses. In the past, as well as today, they represent a comprehensive and systematically structured set of mostly analytical maps that contain statistically processed data from censuses in different spatial scales. In the era of digital cartography, geographical databases, the Internet and variety of distributed computer networks the methods of such atlases production are changing. The paper analyses changes and current trends in atlases production with the emphasis on selected aspects in the national and international framework. The relative richness of the Czech-Slovak atlases production inspired us to choose the study focused on four national atlases published in 1935, 1966, 1980 and 2002, which were compared on the basis of selected quantitative and qualitative characteristics capturing their structure, content, composition, scales of maps resolution and applied cartographic methods. The aim is to show the scope of the atlases specific differences, resp. similarity in demographically targeted parts.*

Kľúčové slová: *štatistická mapa a atlas, sčítanie obyvateľov, kartografická vizualizácia.*

Key words: *statistical map and atlas, census, cartography visualization.*

Analýza chování koeficientů pro stanovení počtu shluků ve shlukové analýze
Analysis of the Behavior of Coefficients for Determining the Number of Clusters in Cluster Analysis

Tomáš Löster, Jakub Danko
Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky, Nám. W. Churchilla 4, 130 67 Praha 3, Česká republika
tomas.loster@vse.cz, danj01@vse.cz

Abstrakt: *Shluková analýza je vícerozměrná klasifikační metoda, jejímž cílem je vytvářet skupiny objektů. Výstupem shlukové analýzy bývá také stanovení počtu shluků. K tomuto cíli existuje mnoho koeficientů a postupů. Tyto koeficienty se liší dobou vzniku a principem, jak počet shluků stanovují. V současné době existuje mnoho softwarových produktů, které lze ke stanovení počtu shluků využít. Mezi tyto produkty patří i R-software, ve kterém je v rámci různých balíčků aplikováno zřejmě největší množství těchto koeficientů. Cílem tohoto příspěvku je ukázat chování vybraných koeficientů pro stanovení počtu shluků v různých podmínkách. Jedná se o simulaci chování na generovaných datech. Analyzovány jsou různé skupiny souborů, v rámci kterých je vždy dostatečně velký počet shluků. Za všeobecně nejstabilnější koeficienty pro stanovení počtu shluků lze považovat PTS a CHF koeficient, které se liší úspěšností podle stupně překrytí jednotlivých shluků. Z používaných měr vzdáleností bývá obvykle nejlepších výsledků dosaženo při Euklidovské míře vzdálenosti.*

Abstract: *Cluster analysis is a multivariate classification method that aim is to create groups of objects. The output of cluster analysis is also the determination of the number of clusters. There are many coefficients and procedures for this goal. These coefficients differ by the time of occurrence and the principle of process how the number of clusters are determined. There are currently many software products that can be used to determine the number of clusters. These coefficients are also include in R-software, in which the largest number of these coefficients are apparently applied in the various packages. The aim of this paper is to show the behavior of selected coefficients for determining the number of clusters under different conditions. This is a simulation of behavior on the generated data. Different groups of files are analyzed, within which there is always a large number of files. For the most stable coefficients for determining the number of clusters can be considered PTS and CHF coefficients. Their success depends on the degree of overlapping of areas of clusters. The most appropriate measure of distance is usually the Euclidean distance measure.*

Kľúčové slová: *Shluková analýza, koeficienty, optimální počet shluků, software, R-software*

Keywords: *Cluster analysis, coefficients, optimal number of clusters, software, R-software*

Determinanty chudoby na Slovensku
Determinants of Poverty in Slovakia

Alena Mojsejová
Ekonomická fakulta, Technická univerzita v Košiciach
Faculty of Economics, Technical University of Košice
alena.mojsejova@tuke.sk

Abstrakt: *Cieľom príspevku je na základe údajov zo Zisťovania o príjmoch a životných podmienkach domácností EU SILC 2016 identifikovať faktory ovplyvňujúce výskyt chudoby v slovenských domácnostiach. Na kvantifikáciu vplyvu vybraných faktorov na riziko chudoby a modelovanie pravdepodobnosti bol použitý logistický regresný model. Ukázali sme, že chudoba na Slovensku má svoje regionálne aspekty. Štatisticky významne rozdiely sú podľa premennej kraj aj podľa stupňa urbanizácie. Zistili sme, že premenné ako pohlavie, vek, typ domácností, ekonomická aktivita osoby na čele domácnosti, rodinný stav, vzdelanie, zdravie a vlastnícky status významne ovplyvňujú výskyt chudoby. Na základe vypočítaných koeficientov kontingencie najvyššou mierou bol výskyt chudoby ovplyvnený ekonomickou aktivitou osoby na čele domácností a naopak najnižšou mierou všeobecným zdravím osoby na čele domácnosti.*

Abstract: *The aim of this study is to find out the factors affecting poverty in Slovak households, based on the data from European Study on Income and Living Conditions survey, EU SILC 2016. A logistic regression model was used to quantify the impact of selected factors on the risk of the poverty and for probability modelling. We have shown that poverty in Slovakia has regional aspects. Statistically significant differences are by the region and also by the degree of urbanization. We found that variables such as gender, age, household type, households' economic activity, marital status, education, health, and tenure status significantly affected the occurrence of poverty. According to contingency coefficients, the rate of poverty was at most influenced by the economic activity, on the other hand, the lowest rate was obtained for general health of the person at the head of the household.*

Kľúčové slová: *: EU SILC, hranica rizika chudoby, logistická regresia*

Key words: *EU SILC, at-risk-of-poverty threshold, logistic regression*

Poďakovanie

Tento príspevok bol vypracovaný v rámci projektu VEGA 2/0026/15 Príjmová stratifikácia a perspektívy polarizácie slovenskej spoločnosti do roku 2030.

Použití výstupního systému SAS pro vytváření pracovních listů aplikace Microsoft Excel
Using the SAS Output Delivery System to Create Microsoft Excel Worksheets

Roman Pavelka
Štatistický úrad SR, Miletičova 3, 824 67 Bratislava, Odbor metód štatistických zisťovaní
Statistical Office of Slovak Republic, Miletičova 3, 824 67 Bratislava, Statistical Surveys and Methodology Department
Roman.Pavelka@statistics.sk

Abstrakt: *Pracovní listy sešitu aplikace Microsoft Excel představují nejrozšířenější tabulkový procesor, který efektivně umožňuje uživatelům analyzovat data. Když se zkombinuje výstupní systém programu SAS Output Deliver System (ODS) s funkcemi aplikace Excel, dostáváme výkonnou sadu nástrojů, jejíž pomocí lze snadno využívat výpočetních možností prostředí SAS a manipulovat s daty, včetně formátování či zvýrazňování dat, vkládání vzorců, grafů či vytváření kontingenční tabulky v sešitu Excel. Mimo uvedených funkcionalit výstupní systém programu SAS ODS spolu s možnostmi sešitu Excel poskytuje mnoho způsobů pro zatraktivnění vzhledu vytvářených tabulek a grafů. Předkládaný příspěvek, který je určen nejen pro zkušené programátory v prostředí systému SAS, ale i pro začínajícího analytika, na několika příkladech ilustruje možnosti přesměrování výstupů procedur, grafických výstupů a ostatních programových funkcionalit programu SAS do listů pracovního sešitu Microsoft Excel. Při vytváření pracovních sešitů s přesměrovanými výstupy z programu SAS se přiblíží základní programové příkazy a funkcionality programu SAS, které jsou při přesměrování výstupů do sešitu Microsoft Excel využívány, uživatelské formátování přesměrovaných dat i tvorba komplexních reportů.*

Abstract: *Microsoft Excel worksheets are the most widely used spreadsheets that effectively enable users to analyse data. When the SAS Output Deliver System (ODS) output system combines with Excel, we receive a powerful toolkit that makes it easy to use SAS computing capabilities and data manipulating in an Excel workbook, including formatting or highlighting data, embedding formulas, graphs or creating a PivotTable. In addition to the above features, the SAS ODS output system, along with the Excel workbook, provides many ways to make the tables and charts more productive. The present post, which is designed not only for experienced SAS programmers but also for novice analysts, illustrates on several examples of redirecting outputs of procedures, graphical outputs, and other SAS program functionalities into Microsoft Excel worksheets. When creating workbooks with redirected SAS outputs, the basic program commands and SAS functionalities that are used when redirecting outputs to the Microsoft Excel workbook, user formatting of redirected data, and the creation of complex reports will be approached.*

Kľúčové slová: *list Microsoft Excel, programový systém SAS, tabulkový procesor, výstupní systém programu SAS.*

Key words: *Microsoft Excel worksheets, programming system SAS, SAS Output Deliver System, spreadsheet.*

Identifikácia a použitie metód hĺbkovej analýzy údajov na maximalizáciu ziskovosti retenčnej kampane
Identification and Application of Data Mining Methods in Order to Maximize the Profitability of a Retention Campaign

Iveta Stankovičová, Daniel Gábriš
Univerzita Komenského v Bratislave, Fakulta managementu; Katedra informačných systémov
Comenius University in Bratislava, Faculty of Management
iveta.stankovičová@fm.uniba.sk; daniel.gabris@gmail.com

Abstrakt: *Príspevok je zameraný na identifikáciu a použitie metód hĺbkovej analýzy údajov na optimálnu selekciu zákazníkov do retenčnej kampane. Hlavným cieľom bolo dosiahnuť maximalizáciu celkovej ziskovosti retenčnej kampane. Hĺbková analýza sa týkala portfólia klientov bankového sektora na úrovni individuálneho produktu – spotrebného úveru. Dátový súbor bol poskytnutý nemenovanou bankovou inštitúciou pôsobiacou na slovenskom trhu. V práci sme sa venovali prediktívnemu modelovaniu predčasného splatenia spotrebného úveru. K tomuto účelu sme použili nielen klasické individuálne prediktívne modely, ale aj alternatívne metódy, formou tvorby zložených modelov. Hodnotiacim kritériom jednotlivých modelov bol kumulatívny lift na zvolenom kvantile, odvodený z výpočtu ziskovosti retenčnej kampane. Vytvorené prediktívne modely sme na danom hodnotiacom kritériu porovnali. Pokúsili sme sa aj o interpretáciu zloženého modelu HCES pomocou preskúmania jeho členských modelov. Nakoniec sme ilustrovali potrebu organizácie vyhodnotiť mieru finančnej atraktivity použitia lepších prediktívnych modelov, pomocou svojich špecifických podmienok a faktorov, ktoré sú v každej organizácii individuálne a unikátne. Na výpočty a dosiahnutie vytýčených cieľov sme použili hlavne softvér SAS Enterprise Miner 14.1.*

Abstract: *The paper aims to identify and use data mining methods to optimize customer selection into a retention campaign. The main goal of selection is to maximize the overall profitability of the retention campaign. The in-depth analysis covered the portfolio of retail banking clients at the level of the individual product - consumer credit. An unnamed banking institution operating on the Slovak banking market provided the data set. We have been working on the predictive modelling of early repayment of consumer credit. For this purpose, we used not only classic individual predictive models but also alternative methods, in the form of ensemble models. The assessment criterion for each model was a cumulative lift on the selected quantile, derived from the retention campaign's profitability calculation. All the created predictive models were compared to the given evaluation criterion. We have also attempted to interpret the selective composite HCES model by reviewing its member models. Finally, we have illustrated the organization's need to evaluate the degree of financial attractiveness of using better predictive models, using its specific conditions and factors that are individual and unique in each organization. We mainly used SAS Enterprise Miner 14.1 software to calculate and achieve the goals.*

Kľúčové slová*: hĺbková analýza údajov, marketingová retenčná kampaň, predpoveď odchodovosti, prediktívne zložené modely, metóda hill-climbing ensemble selection (HCES)*

Key words: *data mining, marketing retention campaign, churn prediction, ensemble predictive models, hill-climbing ensemble selection method (HCES)*

Zhlukovanie časových radov dlhodobých úrokových mier
Clustering Time Series of Long-term Interest Rates

Beáta Stehlíková
Fakulta matematiky, fyziky a informatiky, Univerzita Komenského, Mlynská dolina, 842 48 Bratislava
Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia
stehlikova@fmph.uniba.sk

Abstrakt: *V článku sa zaoberáme zhlukovaním časových radov dlhodobých úrokových mier v štátoch Európskej únie. Aplikujeme algoritmus hierarchického zhlukovania s vhodne zvolenými parametrami, pomocou ktorého identifikujeme skupiny štátov s podobným vývojom úrokových mier.*

Abstract: *In the paper we deal with clustering time series of long-term interest rates in the countries of the European Union. We apply a hierarchical clustering algorithm with suitable chosen parameters, by which we identify groups of countries with a similar evolution of the interest rates.*

Kľúčové slová: *časové rady, zhlukovanie, dlhodobé úrokové miery.*

Key words: *time series, clustering, long-term interest rates.*

Poďakovanie

Článok bol vypracovaný za podpory grantu VEGA 1/0251/16.

Analýza doživotných dôchodkov vyplácaných zo starobného dôchodkového sporenia v Slovenskej republike
Analysis of Whole Life Annuities Paid out from the Old-age Pension Scheme in the Slovak Republic

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: *Tento príspevok sa zaoberá výplatnou fázou starobného dôchodkového sporenia v rámci slovenského dôchodkového systému. Cieľom príspevku je skúmať vplyv vybraných rizikových faktorov na výšku mesačných dôchodkov vyplácaných v dekumulačnej fáze starobného dôchodkového sporenia a ilustrovať rozloženie jednorazového poistného. V rámci analýz boli použité úmrtnostné tabuľky Štatistického úradu Slovenskej republiky a dáta z Human Mortality Database v kombinácii s Leeovým-Carterovým modelom dlhovekosti. Pri diskontovaní peňažných tokov sa aplikovala Svenssonova výnosová krivka publikovaná na webovej stránke Európskej centrálnej banky (ECB). Hlavnými výsledkami štúdií sú odhadnuté výšky efektívneho poistného, resp. rizikových a nákladových položiek pri rôznych vekoch odchodu do dôchodku a nastaveniach modelu.*

Abstract: *This contribution deals with the pay-out phase of the old-age pension scheme in Slovakia. The aim of the contribution is to examine impact of selected risk factors on the monthly pension payments in decumulation phase of the old-age pension scheme and illustrate the decomposition of the single premium. In our analyses we applied Life Tables of the Statistical Office of the Slovak Republic and data series from the Human Mortality Database combined with the Lee-Carter model of longevity. For discounting the cash-flow we used Svensson yield curves with parameters published by the European Central Bank (ECB). The main result of our study is an estimation of the effective premium and loading components in case of various retirement ages and model setups.*

Kľúčové slová: *výplatná fáza, starobné dôchodkové sporenie na Slovensku, doživotné dôchodky, rizikové faktory.*

Key words: *pay-out phase, old-age pension scheme in Slovakia, whole life annuities, risk factors.*

Poďakovanie

Tento výskum bol podporený grantmi VEGA č. 1/0251/16 *Kvantitatívna analýza modelov úrokových mier v podmienkach eurozóny a pristupujúcich krajín a jej aplikácie* a VEGA č. 2/0054/18 *Nové štatistické metódy pre špeciálne triedy rozdelení pravdepodobnosti a ich aplikácie*.

Alternatívne prístupy k identifikácii subjektívne chudobných
Alternative Approaches to the Identification of the Subjectively Poor

Tomáš Želinský[a], Martina Mysíková[b]

[a] Ekonomická fakulta, Technická univerzita v Košiciach, Němcovej 32, 040 01 Košice
Faculty of Economics, Technical University of Košice, Němcovej 32, 040 01 Košice, Slovakia

[b] Sociologický ústav AV ČR, v.v.i., Jilská 1, 110 00 Praha 1
Institute of Sociology, Czech Academy of Sciences, Jilská 1, 110 00 Praha 1, Czech Republic

tomas.zelinsky@tuke.sk, martina.mysikova@soc.cas.cz

Abstrakt: *Príjmovú chudobu je možné merať s použitím absolútneho, relatívneho, objektívneho alebo subjektívneho prístupu, pričom v rámci každého prístupu existuje viacero metód. V porovnaní s objektívnymi mierami chudoby, použitie indikátorov subjektívnej chudoby a well-beingu, ako aj ich analýzy možno považovať za výrazne zanedbávané prístupy ako v oficiálnej štatistike, tak aj akademickom výskume. Neexistuje žiadna všeobecne akceptovaná metóda na odhad úrovne subjektívnej chudoby v spoločnosti a v empirickej literatúre možno, z pochopiteľných dôvodov, nájsť rôznorodé výstupy založené na rôznych prístupoch. Na základe možností, ktoré ponúkajú údaje EU-SILC, používame otázku minimálneho príjmu (najnižšia mesačná suma na uspokojenie potrieb) s cieľom odhadnúť subjektívnu hranicu chudoby pre dostupné krajiny a roky. V rámci uvedeného prístupu sú odpovede na otázku minimálneho príjmu konfrontované s aktuálnym príjmom, kontrolujúc o ekonomické a iné charakteristiky domácností. Identifikácia subjektívne chudobných nezávisí len od použitého prístupu, ale aj od požitej metódy odhadu. Cieľom tejto štúdie je porovnať alternatívne metódy odhadu na identifikovanie subjektívne chudobných. Zameriavame sa pritom na porovnanie výsledkov založených použitím nasledovných techník: metóda najmenších štvorcov, kvantilová regresia, polynomická regresia, MARS a mixture of regression models.*

Abstract: *Income poverty can be measured under an absolute, relative, objective, or subjective approach, each of which comprises a variety of methods. Compared to objective measures of poverty, the use of subjective poverty and well-being indicators and its analyses seem to be the most neglected approaches, both in the official statistics and in academic research. No single generally accepted method exists for estimating the overall level of subjective poverty in society and diverse poverty outcomes based on different approaches are usual in the empirical literature for understandable reasons. Based on the possibilities of the EU-SILC data, we apply the MIQ ("Lowest monthly income to make ends meet") to estimate the subjective poverty line for available countries and years. Within this approach, responses to the minimum income question are intersected with the reported actual income, controlling for household and economic variables. Nevertheless, the identification of subjectively poor depends not only on the approach applied but also on the estimation technique. The goal of this study is to compare alternative estimation techniques to the identification of the poor. In particular, we intend to compare the results based on the following estimation techniques: OLS regression, quantile regression, polynomial regression, MARS, and a mixture of regression models.*

Kľúčové slová: *subjektívna chudoba, EU-SILC, regresia.*

Keywords: *subjective poverty, EU-SILC, regression.*

Časť 2

Výpočtová štatistika 2018
Prehliadka prác mladých štatistikov a demografov

Zborník abstraktov a článkov

Príjmová nerovnosť vo vybraných členských krajinách EÚ
Income Inequality in Selected EU Countries

Lucia Janovičová
Fakulta ekonomiky a manažmentu, Slovenská poľnohospodárska univerzita v Nitre, 1. ročník inžinierskeho štúdia
Faculty of Economics and Management, Slovak University of Agriculture in Nitra
xjanovicova@uniag.sk

Abstrakt: *Cieľom príspevku bolo hodnotenie vývoja príjmovej nerovnosti, miery rizika chudoby vo vybraných členských krajinách EÚ v rokoch 2005 - 2017. Použili sme panel ročných údajov 13 ukazovateľov 19 členských krajín EÚ (Eurostat, 2018). Vplyv vybraných faktorov na príjmovú nerovnosť a mieru rizika chudoby sme odhadli ekonometrickými modelmi. Príjmová nerovnosť a miera rizika chudoby v EÚ mali v skúmanom období mierne rastúci trend. Štatisticky preukazný pozitívny vplyv na rast príjmovej nerovnosti a rast miery rizika chudoby v EÚ mali podiel obyvateľov vo veku 65 rokov a viac, miera nezamestnanosti a podiel dospelých žijúcich v domácnostiach bez zamestnania. Rast celkových výdavkov verejnej správy a rast podielu obyvateľov s dosiahnutým stredoškolským vzdelaním štatisticky preukazne znižoval príjmovú nerovnosť aj mieru rizika chudoby. Rast HDP na obyvateľa mal štatisticky preukazný vplyv na pokles príjmovej nerovnosť EÚ.*

Abstract: *The aim of the contribution was to assess development of income inequality, poverty risk rate in the EU Member States (MS) over the years 2005-2017. We used the panel of annual data of 13 indicators of 19 EU Member States (Eurostat, 2018). The impact of selected factors on income inequality and at the risk of poverty rate was estimated by econometric models. The income inequality and at the risk of poverty rate in the EÚ MS were slightly growing. Proportion of population aged 65 years and more, unemployment rate and people aged 18-59 living in jobless household had statistically significant positive effect on income inequality and at the risk of poverty rate growth. Total general government expenditure growth and growing share of population with upper secondary, post-secondary non-tertiary and tertiary education (levels 3-8) had statistically significant negative effect on EÚ income inequality and at the risk of poverty rate growth. GDP per capita growth statistically significantly reduced income inequality.*

Kľúčové slová*: príjmová nerovnosť, miera rizika chudoby, panelové údaje, krajiny EÚ*

Key words: *income inequality, risk of poverty rate, panel data, EU countries*

Predikcia výskytu rakoviny pľúc
Prediction of Lung Cancer Diagnosis

Lucia Plichtová
Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, inžinierske štúdium, 1.ročník
University of Economics in Bratislava, Faculty of Business Informatics
lu.plichtova@gmail.com

Abstrakt: *Cieľom príspevku je pomocou štatistickej analýzy reálnych dát pacientov trpiacich ochorením rakoviny pľúc a metódou logistickej regresie vytvoriť predikčný regresný model v jazyku R a otestovať jeho predikčnú presnosť. Tento model je prepojený s používateľskou aplikáciou v jazyku Java, ktorá na základe informácií od používateľa u neho vyhodnotí pravdepodobnosť výskytu rakoviny pľúc.*

Abstract: *The aim of this paper is to create a predictive regression model in the R language and to test its predictive accuracy with the use of logistic regression and statistical analysis performed on the real data of patients that suffer from lung cancer. This model is connected to an application created in the Java language that takes an input from its user and uses it to calculate their probability of lung cancer.*

Kľúčové slová: *Logistická regresia, imputácia dát, R, predikcia rakoviny pľúc*
Key words: *Logistic regression, data imputation, R language, lung cancer prediction*

1 Úvod

Pľúcny karcinóm, ktorý je známejší pod pomenovaním rakovina pľúc je v súčasnosti už štvrtým najrozšírenejším onkologickým ochorením. Aj v rámci Slovenska patrí medzi najčastejšie príčiny úmrtia. Nevieme nijako ovplyvniť, či nás toto ochorenie postihne alebo nie, avšak v dobe technologického pokroku existujú spôsoby a predikčné metódy, ktoré sú schopné na základe existujúcich štatistík a vzoriek pacientov s rakovinou pľúc zanalyzovať rizikové faktory vplývajúce na toto ochorenie a na základe týchto poznatkov zistiť úroveň možného výskytu ochorenia u iného pacienta. Najprv popíšeme vytvorenie modelu v jazyku R, ktorý je „natrénovaný“ (pomocou metódy logistickej regresie) predpovedať toto ochorenie na základe dát reálnych pacientov, ktorí trpeli týmto ochorením a zistíme tým rizikové faktory, ktoré ovplyvnili vznik ochorenia. Model je možno prepojiť s používateľskou aplikáciou, ktorá je na základe údajov zadaných používateľom schopná u neho predikovať pravdepodobnosť výskytu rakoviny pľúc.

2 Logistická regresia

Metóda logistickej regresie umožňuje odhadnúť pravdepodobnosť nastatia nejakého javu, ktorého výskyt označujeme pomocou hodnoty náhodnej premennej $Y\,\{Y = 1\}$ a to pomocou k známych nezávislých premenných

tzv. regresorov $X_1, \ldots, X_k$, ktorými môžu byť spojité, ale aj kategorické premenné. Výstupom logistickej regresie je hodnota z intervalu $\langle 0,1 \rangle$, teda pravdepodobnosť, s akou nastane náhodný jav ($Y = 1$), ktorého výskyt je podmienený nezávislými premennými. Vzťah na odhad tejto pravdepodobnosti môžeme vyjadriť takto

$$\pi = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j X_j)}{1+\exp(\beta_0 + \sum_{j=1}^{k} \beta_j X_j)} = \frac{1}{1+\exp(-\beta_0 - \sum_{j=1}^{k} \beta_j X_j)} \quad (1)$$

2.1 Dáta a hraničná hodnota t-value

Dáta v logistickej regresii zahŕňajú jeden alebo viac stĺpcov nezávislej premennej a jediný stĺpec závislej premennej. Binárna závislá premenná Y nadobúda hodnoty 0 alebo 1, ktoré indikujú výskyt ($Y = 1$) alebo absenciu sledovaného javu ($Y = 0$) (v našom prípade výskyt alebo absenciu rakoviny pľúc). Model logistickej regresie nepredikuje hodnotu 0 alebo 1 závislej premennej, ale pravdepodobnosť výskytu (absencie) javu, teda hodnoty z intervalu $\langle 0,1 \rangle$. Pre získanie binárneho výsledku sa určuje tzv. hraničná hodnota pravdepodobnosti (threshold value alebo t-value), dosiahnutie a prekročenie ktorej (hodnoty rovné alebo väčšie ako t-value) znamená výskyt hodnoty 1 závislej premennej. V prípade, že je odhadnutá hodnota pravdepodobnosti nižšia ako hraničná hodnota (hodnoty nižšie ako t-value), závislá premenná nadobúda hodnotu 0.

2.2 Hodnotenie úspešnosti modelu

Ako sme už uviedli vyššie, pomocou regresného modelu môžeme u danej osoby indikovať prítomnosť alebo neprítomnosť sledovaného ochorenia – rakoviny pľúc. Je však potrebné položiť si otázku, či daná osoba aj v skutočnosti má alebo nemá sledované ochorenie, resp. do akej miery bola predikcia „úspešná“. Úspešnosť predikcie a regresného modelu je možné ohodnotiť pomocou špecificity, senzitivity a celkovej úspešnosti modelu. Vysvetleniu týchto pojmov sa budeme venovať v tejto podkapitole.

V tabuľke 1 sú jednotlivé prípady rozdelené podľa výsledkov predikcie výskytu daného ochorenia (riadky) a podľa skutočnej prítomnosť ochorenia u danej osoby (stĺpce). Hodnota TP (*true positive*) vyjadruje, koľko výsledkov bolo skutočne pozitívnych, teda koľko osôb bolo správne označených ako „chorý“. Hodnota FP (*false positive*) označuje, koľko výsledkov bolo falošne pozitívnych, teda koľko zdravých osôb bolo chybne označených ako „chorý“. Hodnota FN (*false negative*) vyjadruje počet výsledkov, ktoré boli bolo falošne negatívne, teda koľko chorých osôb bolo chybne označených ako „zdravý“. Hodnota TN (*true negative*) predstavuje počet výsledkov, ktoré boli skutočne negatívne t. j. koľko zdravých osôb bolo správne označených ako „zdravý“.

Tab 1 Výsledky predikcie *(Zdroj: vlastné spracovanie)*

		Skutočná prítomnosť choroby	Skutočná prítomnosť choroby
		Áno	**Nie**
Výsledok predikcie	**Pozitívny**	TP	FP
Výsledok predikcie	**Negatívny**	FN	TN

Senzitivita testu je schopnosť testu rozpoznať skutočne choré osoby. Je to pravdepodobnosť, že test bude pozitívny, keď je osoba skutočne chorá. Senzitivitu testu vypočítame na základe údajov tabuľky 1 pomocou vzťahu TP/(TP+FN).

Špecificita testu je schopnosť testu rozpoznať osoby bez choroby. Je to pravdepodobnosť, že test bude negatívny, keď osoba nie je chorá. Špecificitu testu vypočítame z údajov tabuľky 1 vzťahom $TN/(FP+TN)$. Celkovú správnosť testu možno určiť takto: $(TP+TN)/(TP+FP+FN+TN)$.

3 Výber dát a metodika zberu dát

PLCO je kontrolovaná štúdia Národného onkologického inštitútu (*National Cancer Institute, ďalej NCI*) v USA. NCI je súčasťou Národného inštitútu zdravia, ktorý je jeden z 11 inštitútov tvoriacich Ministerstvo zdravotníctva (*Department of Health and Human Services,* HHS). NCI, založený už v roku 1937 je hlavnou inštitúciou federálnej vlády pre výskum v oblasti rakoviny. V priebehu rokov sa vytvorením Národného programu proti rakovine rozšírila zodpovednosť NCI, a tiež sa pridali nové mandáty na šírenie informácií o rakovine a začlenenie najmodernejších liečebných postupov proti rakovine do klinickej praxe. Štúdia PLCO určuje na základe dotazníkov a snímok, či určité tzv. skríningové vyšetrenia znižujú úmrtnosť na rakovinu prostaty, pľúc, konečníka a vaječníkov. V období od novembra roku 1993 do júla 2001 bolo sledovaných približne 155 000 účastníkov. Boli získané údaje o diagnózach rakoviny a o príčinách úmrtí, ktoré sa vyskytli do 31. decembra 2009.

V štúdii sa zhromaždilo veľké množstvo údajov, ktoré pokrývajú rôzne oblasti zamerania. V našej práci sa zameriavame na predikciu rakoviny pľúc, a preto nás zaujímali údaje zo štúdie zozbierané práve o tomto ochorení. Museli sme si od NCI vyžiadať oficiálny súhlas a prístup k datasetu. Uzatvorili sme dohodu, podľa ktorej nebudeme daný dataset nikde rozširovať a jeho údaje budú použité iba na účely vypracovania tejto práce. Po potvrdení dohody sme so súhlasom NCI získali tzv. „*Lung dataset*“ (dataset týkajúci sa rakoviny pľúc). Je to komplexný súbor údajov, ktorý obsahuje takmer všetky dostupné údaje zo štúdie PLCO, ktoré sú k dispozícii pre analýzu výskytu rakoviny pľúc a úmrtnosti na toto ochorenie.

Súbor údajov obsahuje jeden záznam pre každého z 154 897 účastníkov testu PLCO.

Pre vytvorenie modelu logistickej regresie zo získaných dát a jeho následné natrénovanie sme si zvolili jazyk R.

3.1 Príprava dát v R

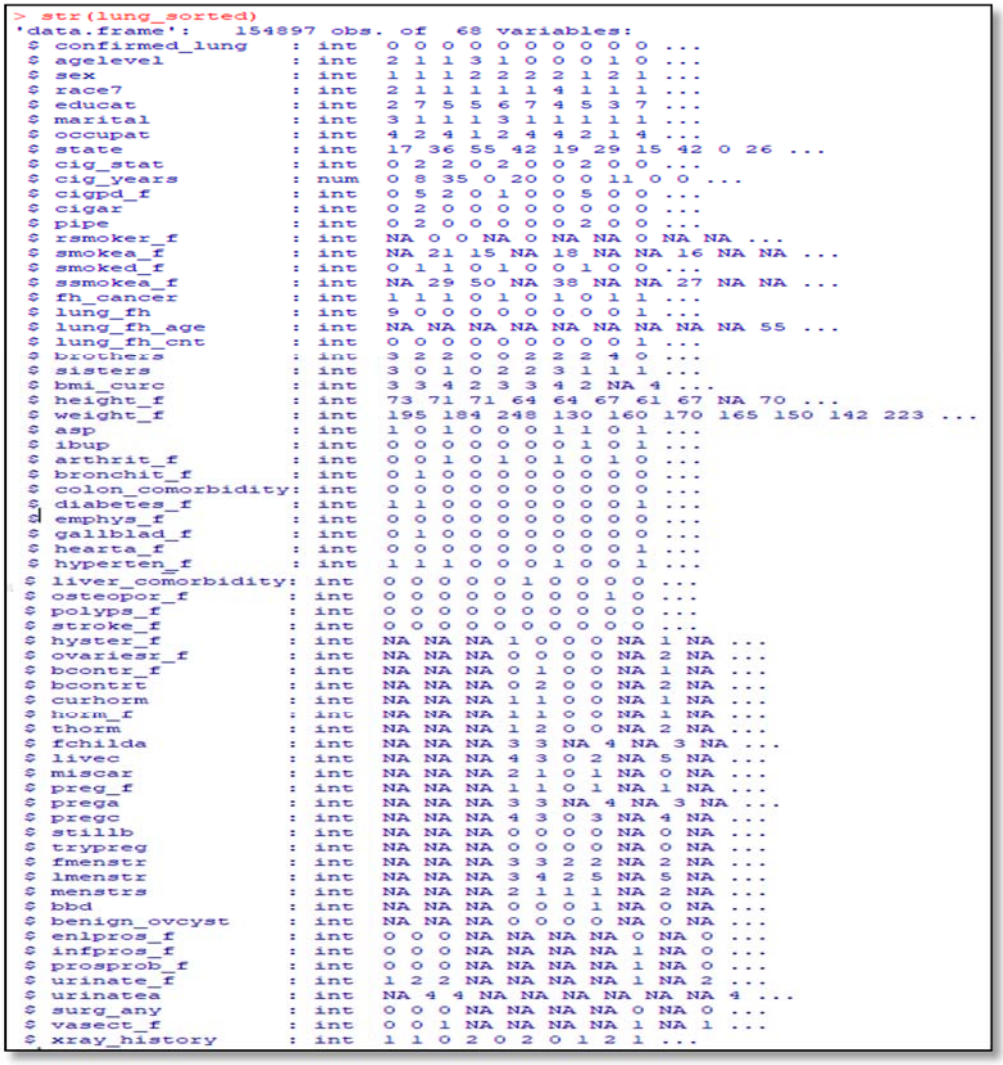
Pred tvorbou modelu logistickej regresie musíme pripraviť dáta do použiteľnej podoby. Ako už bolo spomenuté, dataset získaný od NCI obsahoval dáta od 154 897 ľudí, ktorí odpovedali na približne 206 otázok, čo je skutočne objemný dataset a takéto „hrubé“ dáta sme nemohli hneď vložiť do regresného modelu. Preto sme si museli najprv dataset upraviť do použiteľnej podoby.

Je dôležité podotknúť, že premenné v datasete sú kategorickými premennými, v R sú uvádzané ako faktorové premenné. Faktorové premenné môžu byť nominálne alebo ordinálne. Sú uložené ako celé čísla (*integers*) alebo ako reťazce (*strings*) a nesú znak (*label*) unikátneho celého čísla. Akonáhle je vytvorená faktorová premenná, môže obsahovať len prednastavené hodnoty, známe ako úrovne. Funkčnosť týchto faktorových premenných možno demonštrovať na príklade. Premenná *agelevel* nášho datasetu indikuje, do akej vekovej kategórie patrí daný človek. Môže nadobudnúť len hodnoty 0, 1, 2 alebo 3. Podľa slovníka *Lung cancer dictionary* priloženého k datasetu vieme, že hodnota 0 reprezentuje vekovú kategóriu „menej ako 59 rokov“, hodnota 1 reprezentuje kategóriu „60 až 64 rokov“, hodnota 2 reprezentuje kategóriu „65 až 69 rokov“ a kategória 3 reprezentuje kategóriu „70 a viac rokov“.

Premenné popisujúce už nadobudnuté rakovinové ochorenie, jeho štádiá a iné špecifikácie nie sú pre našu prácu potrebné, keďže nás zaujímajú len možné faktory ovplyvňujúce jeho vznik. Preto sme v R vytvorili nový dataset *lung_sorted*, do ktorého sme vložili všetky premenné nepopisujúce rakovinu (samozrejme okrem premennej *confirmed_lung*, ktorá indikuje výskyt alebo nevýskyt rakoviny pľúc u daného človeka).

Ako si možno všimnúť na obrázku 1, pri niektorých premenných je vypísaná hodnota „NA“ (=*not available*), čo v R značí tzv. chýbajúce dáta. V našom prípade to znamená, že daný respondent sa k danej otázke nevyjadril, a preto v jeho prípade hodnota danej premennej chýba. Chýbajúce dáta sa v datasete, ktorý bude použitý na natrénovanie modelu logistickej regresie nesmú vyskytovať. Tento problém sme mohli riešiť jednoduchou funkciou *na.omit()*, ktorá vytvára nový dataset bez NA hodnôt. Inými slovami, z datasetu odstráni celý riadok (v našom prípade údaje o jednom človeku), v ktorom sa nachádza formula NA. Vzhľadom na veľké množstvo položených otázok a s tým súvisiacu objemnosť datasetu takmer každý človek neodpovedal aspoň na jednu otázku. Funkcia *na.omit()* by tak odstránila každý jeden riadok. Preto sme

museli chýbajúce dáta nahradiť nejakou hodnotou. Na tento úkon sme použili tzv. imputáciu.



```
> str(lung_sorted)
'data.frame':    154897 obs. of  68 variables:
 $ confirmed_lung     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agelevel           : int  2 1 1 3 1 0 0 0 1 0 ...
 $ sex                : int  1 1 1 2 2 2 2 1 2 1 ...
 $ race7              : int  2 1 1 1 1 1 1 4 1 1 1 ...
 $ educat             : int  2 7 5 5 6 7 4 5 3 7 ...
 $ marital            : int  3 1 1 1 3 1 1 1 1 1 ...
 $ occupat            : int  4 2 4 1 2 4 4 2 1 4 ...
 $ state              : int  17 36 55 42 19 29 15 42 0 26 ...
 $ cig_stat           : int  0 2 2 0 2 0 0 2 0 0 ...
 $ cig_years          : num  0 8 35 0 20 0 0 11 0 0 ...
 $ cigpd_f            : int  0 5 2 0 1 0 0 5 0 0 ...
 $ cigar              : int  0 2 0 0 0 0 0 0 0 0 ...
 $ pipe               : int  0 2 0 0 0 0 0 2 0 0 ...
 $ rsmoker_f          : int  NA 0 0 NA 0 NA NA 0 NA NA ...
 $ smokea_f           : int  NA 21 15 NA 18 NA NA 16 NA NA ...
 $ smoked_f           : int  0 1 1 0 1 0 0 1 0 0 ...
 $ ssmokea_f          : int  NA 29 50 NA 38 NA NA 27 NA NA ...
 $ fh_cancer          : int  1 1 1 0 1 0 1 0 1 1 ...
 $ lung_fh            : int  9 0 0 0 0 0 0 0 0 1 ...
 $ lung_fh_age        : int  NA NA NA NA NA NA NA NA NA 55 ...
 $ lung_fh_cnt        : int  0 0 0 0 0 0 0 0 0 1 ...
 $ brothers           : int  3 2 2 0 0 2 2 2 4 0 ...
 $ sisters            : int  3 0 1 0 2 2 3 1 1 1 ...
 $ bmi_curc           : int  3 3 4 2 3 3 4 2 NA 4 ...
 $ height_f           : int  73 71 71 64 64 67 61 67 NA 70 ...
 $ weight_f           : int  195 184 248 130 160 170 165 150 142 223 ...
 $ asp                : int  1 0 1 0 0 0 1 1 0 1 ...
 $ ibup               : int  0 0 0 0 0 0 0 1 0 1 ...
 $ arthrit_f          : int  0 0 1 0 1 0 1 0 1 0 ...
 $ bronchit_f         : int  0 1 0 0 0 0 0 0 0 0 ...
 $ colon_comorbidity: int  0 0 0 0 0 0 0 0 0 0 ...
 $ diabetes_f         : int  1 1 0 0 0 0 0 0 0 1 ...
 $ emphys_f           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ gallblad_f         : int  0 1 0 0 0 0 0 0 0 0 ...
 $ hearta_f           : int  0 0 0 0 0 0 0 0 0 1 ...
 $ hyperten_f         : int  1 1 1 0 0 0 1 0 0 1 ...
 $ liver_comorbidity: int  0 0 0 0 0 0 1 0 0 0 ...
 $ osteopor_f         : int  0 0 0 0 0 0 0 0 1 0 ...
 $ polyps_f           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ stroke_f           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hyster_f           : int  NA NA NA 1 0 0 0 NA 1 NA ...
 $ ovariesr_f         : int  NA NA NA 0 0 0 0 NA 2 NA ...
 $ bcontr_f           : int  NA NA NA 0 1 0 0 NA 1 NA ...
 $ bcontrt            : int  NA NA NA 0 2 0 0 NA 2 NA ...
 $ curhorm            : int  NA NA NA 1 1 0 0 NA 1 NA ...
 $ horm_f             : int  NA NA NA 1 1 0 0 NA 1 NA ...
 $ thorm              : int  NA NA NA 1 2 0 0 NA 2 NA ...
 $ fchilda            : int  NA NA NA 3 3 NA 4 NA 3 NA ...
 $ livec              : int  NA NA NA 4 3 0 2 NA 5 NA ...
 $ miscar             : int  NA NA NA 2 1 0 1 NA 0 NA ...
 $ preg_f             : int  NA NA NA 1 1 0 1 NA 1 NA ...
 $ prega              : int  NA NA NA 3 3 NA 4 NA 3 NA ...
 $ pregc              : int  NA NA NA 4 3 0 3 NA 4 NA ...
 $ stillb             : int  NA NA NA 0 0 0 0 NA 0 NA ...
 $ trypreg            : int  NA NA NA 0 0 0 0 NA 0 NA ...
 $ fmenstr            : int  NA NA NA 3 3 2 2 NA 2 NA ...
 $ lmenstr            : int  NA NA NA 3 4 2 5 NA 5 NA ...
 $ menstrs            : int  NA NA NA 2 1 1 1 NA 2 NA ...
 $ bbd                : int  NA NA NA 0 0 0 1 NA 0 NA ...
 $ benign_ovcyst      : int  NA NA NA 0 0 0 0 NA 0 NA ...
 $ enlpros_f          : int  0 0 0 NA NA NA NA 0 NA 0 ...
 $ infpros_f          : int  0 0 0 NA NA NA NA 1 NA 0 ...
 $ prosprob_f         : int  0 0 0 NA NA NA NA 1 NA 0 ...
 $ urinate_f          : int  1 2 2 NA NA NA NA 1 NA 2 ...
 $ urinatea           : int  NA 4 4 NA NA NA NA NA NA 4 ...
 $ surg_any           : int  0 0 0 NA NA NA NA 0 NA 0 ...
 $ vasect_f           : int  0 0 1 NA NA NA NA 1 NA 1 ...
 $ xray_history       : int  1 1 0 2 0 2 0 1 2 1 ...
```

Obr. 1: Výsledok po zavolaní funkcie *head()* upraveného datasetu *lung_sorted* (*Zdroj: vlastné spracovanie)*

3.2 Imputácia

Imputáciou sa v oblasti štatistiky myslí proces nahrádzania chýbajúcich údajov substituovanými hodnotami. Pretože chýbajúce dáta môžu spôsobiť problémy pri analýze údajov, imputácia nevylúči z analýzy všetky prípady s chýbajúcimi údajmi, ale zachová všetky prípady a nahradí chýbajúce údaje najlepšie odhadovanou hodnotou na základe iných dostupných informácií. Po imputácii všetkých chýbajúcich hodnôt môže byť dataset analyzovaný štandardnými technikami tak, akoby bol kompletný.

Pri analýzach sme použili imputačnú metódu, ktorá využíva tzv. E-M (*Expectation – Maximization* = očakávanie – maximalizácia) algoritmus. Je to iteratívny postup, v ktorom sa používajú ostatné hodnoty na určenie chýbajúcej hodnoty (očakávanie) a následne prebehne kontrola, či táto hodnota je

najpravdepodobnejšia (maximalizácia). Ak nie, postup sa opakuje a imputuje sa pravdepodobnejšia hodnota. E-M imputácia zachováva vzťahy medzi premennými, čo je dôležité pri práci s faktorovými premennými alebo ak chceme robiť regresiu.

R disponuje viacerými balíkmi, ktorými je možné imputovať chýbajúce dáta, my sme vybrali - balík „*mice*“. Tento balík je určený pre spojité, binárne, ale aj kategorické premenné. Vykonáva tzv. viacnásobnú imputáciu reťazcovými rovnicami (*multivariate imputation by chained equations*).

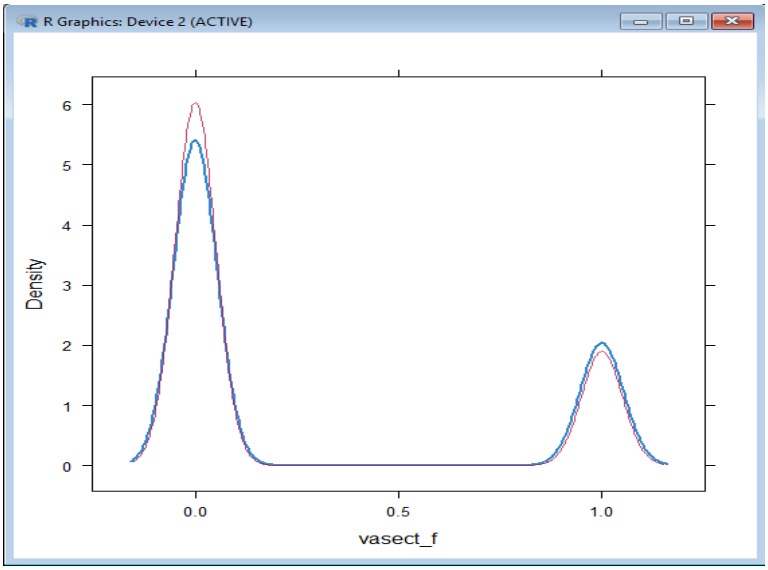
Konečná formula pri volaní funkcie *mice()* bola nasledovná:

imp.lung_sorted <- mice(lung_sorted_testing, m=1, method='pmm').

Po prebehnutí funkcie a vykonaní imputácie bolo treba overiť jej korektnosť. Prvou rýchlou kontrolou bola vizuálna kontrola, kedy sme pomocou funkcie *head*() alebo *summary*() overili, aké údaje boli imputované.

Správnosť prebehnutej imputácie bolo treba overiť samozrejme aj inými spôsobmi. Jedným z nich bolo zavolať funkciu *densityplot()* z balíka „ggplot“.

Na obrázku 2 vidíme výsledok tejto funkcie, ktorým je rozdelenie pravdepodobnosti pôvodných aj imputovaných dát – nami náhodne vybraného parametra *vasect_f*. Na grafe sú červenou krivkou znázornené imputované dáta a modrou krivkou pôvodné dáta. Ako vidno, takmer sa kopírujú, čo je veľmi pozitívne. Keby sa krivky rozchádzali, znamenalo by to, že imputáciou sme narušili pôvodné rozdelenie premenných. Obdobným spôsobom sme otestovali výsledky imputácie aj ďalších premenných.



Obr. 2: Densityplot graf premennej vasect_f. *Zdroj: vlastné spracovanie*

4 Regresný model

Po dôkladnej príprave dát možno pristúpiť ku tvorbe predikčného modelu, ktorý na základe veľkej škály vzoriek, ktoré sme mu poskytli, bude schopný robiť predikciu a identifikovať pravdepodobnosť výskytu rakoviny pľúc.

V našom modeli sme vymedzili závislú premennú *confirmed_lung*, ktorá popisuje výskyt rakoviny pľúc u pacienta. Nezávislými premennými (determinantmi) sú všetky ostatné premenné popisujúcich údaje o pacientoch, ktoré sú spomenuté v kapitole o datasete.

Kvôli validácii modelu sme rozdelili dáta na dve množiny (časti) – trénovaciu a testovaciu. Na množine trénovacích dát sa model „natrénuje“, resp. minimalizuje sa chyba na dátach tejto množiny. Na testovacej množine model testujeme na takých dátach, ktoré predtým model ešte „nevidel”. Testovaciu množinu možno nazvať aj množina ex-post. Táto množina ukazuje predikčnú silu modelu. Je predpoklad, že ak bude mať model dobrú predikčnú silu na dátach množiny ex-post, bude mať podobne dobrú predikčnú silu aj na dátach množiny ex-ante (v našom prípade dáta získané od používateľa našej aplikácie). Neexistuje žiadny predpis, v akom pomere rozdeliť dáta do týchto dvoch množín, avšak štandardne sa delia v pomere 75:25, pričom trénovacia množina tvorí 75 percent dát.

Pre naplnenie modelu sme použili funkciu *glm().* Výsledok funkcie sme vložili do premennej *LungLog.* Funkcia disponuje parametrom „data“, do ktorého sme vložili našu množinu dát *train*. Závislú a nezávislé premenné sme vo funkcii označili tvarom $F \sim x_1 + x_2 + \ldots + x_n$, pričom F je v našom prípade závislá premenná *comfirmed_lung* a x_i sú nezávislé premenné, teda ostatné parametre. Pomocou funkcie *glm()* sme najprv vložili do modelu všetky nezávislé premenné. Pri výbere premenných do modelu sme použili test štatistickej významnosti parametrov modelu. V R sme pre tento účel využili funkciu *summary()*. Z pôvodných 68 premenných nám ostalo 27 štatisticky významných. Informácie o nich, ich názvy a možné nadobúdajúce hodnoty možno nájsť v prílohe článku.

Náš regresný model je úspešne naplnený a napokon je potrebné v R zavolať funkciu *predict()*. Výsledky funkcie sme uložili do premennej *predictTrain.* Formula našej funkcie vyzerá nasledovne*:*

predictTrain<-predict(LungLog, type="response", newdata=train);

pričom LungLog označuje naplnený model dát, „response“ je typ predikcie – predikovaná pravdepodobnosť výskytu javu, a parameter „*newdata*“ označuje voliteľný dataset, v ktorom sa majú hľadať premenné, s ktorými je možné predikovať.

```
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.20324  -0.05927  -0.02858  -0.00288   1.03836

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.0439726  0.0280612  -1.567 0.117125
agelevel       0.0102232  0.0015084   6.777 1.26e-11 ***
educat        -0.0031789  0.0008594  -3.699 0.000217 ***
cig_stat      -0.0335212  0.0039907  -8.400  < 2e-16 ***
cig_years      0.0012391  0.0001196  10.365  < 2e-16 ***
cigpd_f        0.0066361  0.0011412   5.815 6.15e-09 ***
cigar         -0.0058095  0.0018880  -3.077 0.002093 **
fh_cancer      0.0083195  0.0028080   2.963 0.003052 **
lung_fh        0.0031105  0.0008919   3.488 0.000488 ***
lung_fh_cnt    0.0130893  0.0036072   3.629 0.000286 ***
height_f       0.0020399  0.0004402   4.635 3.60e-06 ***
weight_f      -0.0001941  0.0000437  -4.441 9.01e-06 ***
bronchit_f     0.0211079  0.0056440   3.740 0.000185 ***
emphys_f       0.0246214  0.0067432   3.651 0.000262 ***
osteopor_f    -0.0130487  0.0064294  -2.030 0.042418 *
thorm          0.0012857  0.0007503   1.714 0.086631 .
livec          0.0055879  0.0020858   2.679 0.007391 **
miscar         0.0058029  0.0024198   2.398 0.016490 *
pregc         -0.0082226  0.0029317  -2.805 0.005042 **
stillb         0.0080467  0.0042505   1.893 0.058361 .
trypreg        0.0096706  0.0038800   2.492 0.012695 *
fmenstr       -0.0046553  0.0016657  -2.795 0.005199 **
lmenstr       -0.0028604  0.0010632  -2.691 0.007140 **
bbd           -0.0107174  0.0029597  -3.621 0.000294 ***
benign_ovcyst -0.0108630  0.0037950  -2.862 0.004208 **
urinate_f     -0.0060982  0.0014749  -4.135 3.57e-05 ***
urinatea       0.0035931  0.0014013   2.564 0.010351 *
xray_history   0.0046317  0.0015219   3.043 0.002342 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.03325065)

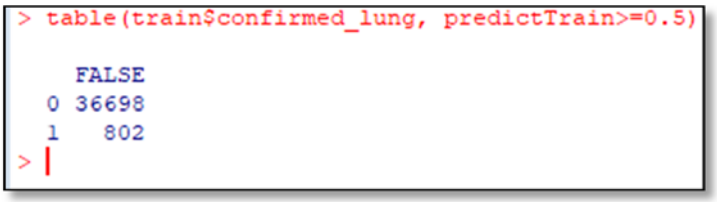
    Null deviance: 683.38  on 19620  degrees of freedom
Residual deviance: 651.48  on 19593  degrees of freedom
  (17879 observations deleted due to missingness)
AIC: -11072
```

Obr.3: Výsledok funkcie summary() po finálnom naplnení modelu
(*Zdroj: vlastné spracovanie)*

Napokon je potrebné položiť si otázku, ako budeme vedieť, ktorá výstupná hodnota – pravdepodobnosť, bude ešte značiť výskyt rakoviny a ktorá už bude označovať absenciu výskytu rakoviny pľúc? Odpoveďou je tzv. *threshold value* alebo t – hodnota, ktorú sme bližšie popisovali v časti 2.1. Ak je výstup logistickej regresie, teda pravdepodobnosť, väčšia alebo rovná ako t-hodnota, bude predpovedaný možný výskyt rakoviny pľúc. Ak je výstup logistickej regresie, teda pravdepodobnosť, nižšia ako t-hodnota, bude predpovedaná absencia výskytu rakoviny pľúc.

Štandardne je t-hodnota nastavená na 0.5, avšak očakáva sa, že v našom prípade to tak nebude. Preto sme sledovali presnosť modelu logistickej regresie s rôznymi t-hodnotami a zvolili sme práve tú t-hodnotu, pri ktorej mal model najvyššiu presnosť.

V jazyku R možno pre tento účel použiť funkciu *table()*, ktorej formula vyzerá nasledovne: *table(train$confirmed_lung, predictTraint>=0.5),* kde porovnávam predpovede s pôvodnými dátami a t-hodnota je štandardná 0.5.



Obr. 4: Výsledok funkcie table() po nastavení t-value na 0.5, *Zdroj: vlastné spracovanie*

Funkcia *table()* by mala po svojom zavolaní vypísať tzv. klasifikačnú maticu, ktorá obsahuje stĺpce FALSE a TRUE. Ako vidno na obrázku 4, po nastavení t-hodnoty na 0.5 stĺpec TRUE úplne absentuje. Preto sme využili aj iné t-hodnoty.

Zistili sme, že pri t-hodnote rovnej 0.075 má náš model presnosť 92.0 %. Avšak treba brať ohľad aj na tzv. špecificitu a senzitivitu. Špecificita vyjadruje schopnosť predikcie presne vybrať prípady, v ktorých choroba nenastáva. Senzitivita (citlivosť) vyjadruje úspešnosť, s akou sa zachytí prítomnosť choroby. Špecificita modelu bola 93,05 %, avšak senzitivita v tomto prípade nadobudla hodnotu len 43.89 %. Preto musíme hľadať presnejšiu t-hodnotu. Skúšame niekoľké, pre prehľadnosť ich uvádzame v tabuľke 2, kde možno vidieť celkovú presnosť, špecificitu a senzitivitu pri jednotlivých t-hodnotách.

Tab.2 Porovnanie presnosti testu pri rôznych t-hodnotách (na trénovacej množine)
Zdroj: vlastné spracovanie

t-hodnota	Celková presnosť	Špecificita	Senzitivita
0.050	82.75%	83.08%	68.95%
0.075	92.00%	93.05%	43.89%
0.060	86.91%	87.48%	60.59%
0.065	88.84%	88.59%	57.48%
0.063	87.93%	88.59%	57.48%
0.068	89.82%	90.65%	51.87%
0.070	90.52%	91.42%	49.37%
0.040	77.84%	77.88%	75.95%
0.045	80.75%	80.63%	73.33%
0.039	77.32%	77.34%	76.30%
0.038	**76.80%**	**76.79%**	**77.05%**

Z tabuľky vidno, že najprijateľnejšia t-hodnota, kde špecificita aj senzitivita sú zhruba na podobnej úrovni a aj celková presnosť je celkom vysoká, je 0.038. Znamená to teda, že v našom modeli môžeme výskyt rakoviny pľúc považovať za možný už od pravdepodobnosti vyššej ako 0.038 so 76.8 percentnou presnosťou.

Napokon je potreba t-value otestovať aj na testovacej množine. Pomocou formuly

predictTest<-predict(LungLog,test,type="response")

zavoláme predikciu aj na testovacích dátach a výsledok uložíme do premennej predictTest. Pomocou formuly

table(test$confirmed_lung, predictTest>=0.038)

následne zisťujeme presnosť testu, jeho špecificitu a senzitivitu s hraničnou hodnotou 0,038 zistenou na trénovacej množine. Celková úspešnosť testu je v tomto prípade 77,18 %, špecificita 77,26 % a senzitivita 73,50 %, čo sú skoro obdobné hodnoty ako na pri trénovacej množine.

5 Aplikácia: Využitie regresného modelu

Pripravený predikčný model rakoviny pľúc v jazyku R však môžu využívať len používatelia, ktorí majú určité znalosti s IT technológiami a najmä s jazykom R alebo s podobným štatistickým prostredím. Ak by takýto používateľ chcel model otestovať na sebe, bolo by to síce možné, ale príliš zložité. Musel by do R vkladať údaje o sebe (pričom by musel prezieravo pracovať s dátami, aby vložil správne kategoriálne hodnoty premenných), znovu volať funkciu *predict()* a inou funkciou vypísať výsledok o sebe. Tento postup je naozaj zdĺhavý a neprehľadný. Taktiež ľudia bez požadovaných znalostí predikčný model nemôžu vyskúšať vôbec. Preto sme sa rozhodli navrhnúť a vytvoriť aplikáciu, ktorá bude jednoduchá, užívateľsky prijateľná a ľahko ovládateľná pre všetky vekové i vzdelanostné kategórie. Aplikácia prevedie používateľa požadovanými otázkami a na konci vyhodnotí, s akou pravdepodobnosťou má používateľ nábeh na rakovinu pľúc. Na jej tvorbu sme využili jazyk Java (v prostredí *NetBeans*) a pomocou knižnice *RCaller* sme prepojili aplikáciu s modelom v R.

Vytvorili sme teda aplikáciu, ktorá by mohla byť nápomocná aj v medicínskom prostredí, kde by sa mohli pacienti testovať a vďaka ktorej by sa mohlo rozšíriť povedomie aj o iných rizikových faktoroch vplývajúcich sa vznik rakoviny pľúc, ktoré sú bežne známe a bežne teoreticky uvádzané.

6 Záver

Cieľom príspevku bolo vytvorenie regresného modelu v jazyku R, ktorý bol metódou logistickej regresie „natrénovaný“ a schopný predpovedať výskyt rakoviny pľúc na základe skutočných dát pacientov a vytvorenie používateľskej aplikácie, ktorá prepojením s týmto modelom vyhodnotí používateľovi pravdepodobnosť výskytu tohto ochorenia. V rámci prípravy dát pred naplnením regresného modelu sme museli zvoliť vhodný postup pre imputáciu chýbajúcich dát. Taktiež sme zisťovali štatistickú významnosť jednotlivých prediktorov vo vytvorenom modeli.

Prínosom práce je okrem iného odhalenie niekoľkých nových rizikových faktorov, ktoré vplývajú na vznik rakoviny pľúc – sú to niektoré prediktory, ktoré sme vyhodnotili ako štatisticky významné a zaradili do nášho predikčného

modelu. Aj bežní používatelia bez akejkoľvek znalosti zo štatistiky alebo logistickej regresie sú schopní použitím našej aplikácie zistiť svoj nábeh na toto ochorenie. V niekoľkých prípadoch by aplikácia dokonca mohla zistiť rizikového človeka a urýchlila by jeho návštevu lekárskeho špecialistu.

7 Literatúra

Budíková, M. (2015*). Aplikovaná statistika II*. [Dostupné na: $https://is.muni.cz/el/1431/jaro2015/MAS02/um/54525918/prednaska12.pdf$]

Janoušková, E. - Dušek, L.(2014). *Analýza dát pro neurovědy*. Brno: Masarykova univerzita v Brne, 2014. [Dostupné na: https://is.muni.cz/el/1411/jaro2014/DSAN01/um/45910621/Analyza_dat_pro_Neurovedy_-_blok_6.pdf]

Gelman, A. - Hill, J.(2006). *Data analysis using regression and multilevel/hierarchical models.*Cambridge University Press, 2006.

Agresti, A.: *Categorical Data Analysis*. New York: John Wiley & Sons, 1990.

Hosmer, D. W. – Lemeshow, S*.: Applied Logistic Regression*. Wiley-Interscience, 2000.

Stankovianska, I.(2018). *Základy štatistiky pre inžinierov.* EDIS Žilinská univerzita, 2018. National Cancer Institute. *Learn About PLCO*. [Dostupné na internete: https://biometry.nci.nih.gov/cdas/learn/plco/]

Príloha: Zoznam významných parametrov modelu (štatistických premenných) a popisy hodnôt, ktoré môžu nadobúdať. *(Zdroj: Vlastné spracovanie)*

Parameter	Popis	Hodnoty, ktoré môže nadobudnúť
agelevel	Veková skupina respondenta	0 =„<=59“ 1=“60-64“ 2=“65-69“ 3=“>=70“
educat	Najvyšší dosiahnutý stupeň vzdelania	1 = „Menej ako 8 rokov štúdia“ 2=“8-11 rokov“ 3=“12 rokov alebo ukončená stredná škola“ 4=“Nadstavba strednej školy(ak nie štúdium na VŠ)“ 5=“Štúdium na VŠ“ 6=“Ukončená VŠ“ 7=“Ukončené postgraduálne štúdium na VŠ“
cig_stat	Stav fajčenia	0=„Nefajčím.“ 1=“Fajčím príležitostne.“ 2=“Fajčím pravidelne.“
cig_years	Počet rokov, počas ktorých respondent fajčí	Číselná hodnota zadaná responendtom
cigpd_f	Počet kusov cigariet, ktoré respondent priemerne spotrebuje za deň	0=“0“ 1=“1-10“ 2=“11-20“ 3=“21-30“ 4=“31-40“ 5=“41-60“ 6=“61-80“ 7=“81+“
cigar	Stav fajčenia cigár	0=“Nikdy“ 1=“Príležitostne“ 2=“Pravidelne“
fh_cancer	Má/mal respondent nejakého blízkeho príslušníka trpiaceho rakovinou?	0=“Nie“ 1=“Áno“
lung_fh	Má/Mal blízky rodinný príslušník respondenta rakovinu pľúc?	0=“Nie“ 1=“Áno“ 2=“Možno, typ rakoviny nebol presne špecifikovaný“
lung_fh_cnt	Počet blízkych rodinných príslušníkov respondenta trpiacich na rakovinu	Číselná hodnota zadaná respondentom
height_f	Výška respondenta	Číselná hodnota zadaná repondentom
weight_f	Váha respondenta	Číselná hodnota zadaná respondentom
bronchit_f	Trpel respondent niekedy chronickou bronchitídou?	0=“Nie“ 1=“Áno“
emphys_f	Trpel respondent niekedy pľúcnym emfyzémom?	0=“Nie“ 1=“Áno“
osteopor_f	Trpel respondent niekedy osteoporózou?	0=“Nie“ 1=“Áno“

thorm	Ak respondentka užíva ženské hormóny, koľko rokov?	0=“Nikdy neužívala“ 1=“10 + rokov“ 2=“6-9 rokov“ 3=“4-5 rokov“ 4=“2-3 roky“ 5=“<= 1 rok“
livec	Koľko úspešných pôrodov respondentka absolvovala(narodilo sa živé dieťa)	0=“0“ 1=“Jeden“ 2=“Dva“ 3=“Tri“ 4=“Štyri“ 5=“Päť a viac“
miscar	Koľko z tehotenstiev respondentky skončilo interrupciou alebo samovoľným potratom?	0=“0“ 1=“Jedno“ 2=“Dva a viac“
pregc	Koľko ráz bola respondentka tehotná? Zahŕňajú sa aj tehotenstvá skončené potratom, či umelé oplodnenie	0=“Ani raz“ 1 = „1“ 2=“2“ 3=“3-4“ 4=“5-9“ 5=“10+“
stilllb	Koľko tehotenstiev skončilo úmrtím plodu po 4. mesiaci tehotenstva?	0=“0“ 1=“1“ 2=“2+“
trypreg	Snažila sa respondentka v minulosti otehotnieť bez úspechu viac ako rok?	0=“Nie“ 1=“Áno“
fmenstr	V akom vekovom rozmedzí dostala respondentka svoju prvú menštruáciu	1=“<10“ 2=“10-11“ 3=“12-13“ 4=“14-15“ 5=“16+“
lmenstr	V akom vekovom rozmedzí vošla respondentka do fázy menopauzy	1=“<40“ 2=“40-44“ 3=“45-49“ 4=“50-54“ 5=“55+“
bbd	Trpela respondentka niekedy fibrocystických ochorením prsníkov?	0=“Nie“ 1=“Áno“
benign_ovcyst	Mala respondentka niekedy nádor alebo cystu na vaječníkoch?	0=“Nie“ 1=“Áno“
urinate_f	Koľko ráz sa respondentovi počas minulého roka stalo, že sa musel ísť v noci vymočiť?	0=“Nikdy“ 1=“Raz“ 2=“Dva razy“ 3=“Tri razy“ 4=“Viac ako tri razy“ 5=“Neviem špecifikovať“

urinatea	Koľko mal respondent rokov, keď nastal jav jeho nočného močenia?	1="<30" 2="30-39" 3="40-49" 4="50-59" 5="60-69" 6="70+"
x_ray_history	Absolvoval respondent v posledných 3 rokoch RTG hrudníka?	0="Nie" 1="Áno, raz" 2="Áno, viac ako raz" 3="Neviem"

Trendy migračných smerov mesta Skalica
Trends of Migration Flows in Skalica Town

Nikoleta Slavíková
Prírodovedecká fakulta Univerzity Komenského v Bratislave, magisterský stupeň, 1. ročník
Faculty of Natural Sciences of the Comenius University in Bratislava
nslavikova3@gmail.com

Abstrakt: *Hlavnou témou tohto príspevku sú migračné smery v meste Skalica v rokoch 1990 – 2017. Smery migrácie sú vyhodnotené na okresnej ako aj obecnej úrovni. Sledované sú migračné smery na úrovni okresov, v Slovenskej aj Českej republike. Osobitná pozornosť je venovaná migračným väzbám s okolitými mestami a vidieckymi obcami, taktiež väzbám na základe špecifík pohraničného mesta. Sťahovanie v rámci mesta Skalica je zhodnotené v krátkej kapitole.*

Abstract: *The main theme of this contribution is the migration flows in the town Skalica in years 1990 – 2017. The migration flows are evaluated at the district level as well as at the municipal level. The migration directions at the level of the districts, both in the Slovak and the Czech Republic, are monitored. Particular attention is paid to migration links with surrounding towns and rural communes, as well as links based on the specifics of the border city. Moving within Skalica is reviewed in a short chapter.*

Kľúčové slová: *mesto Skalica, migračné smery, trendy, sťahovanie v meste*
Key words: *Town Skalica, Migration flows, Trends, Migration in town*

1 Úvod

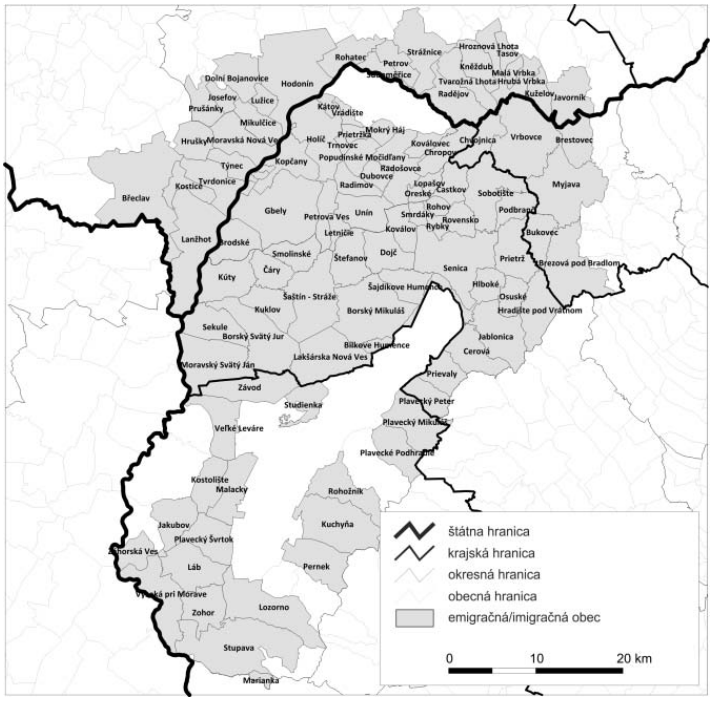
Problematikou migrácie sa zaoberá veľké množstvo autorov, a stala sa teda trendom, ktorý je zaujímavé skúmať a sledovať jeho ďalší vývoj. O vývoj migrácie sa v posledných rokoch začali zaujímať aj orgány samosprávy miest, pretože do značnej miery ovplyvňuje ďalší vývoj a smerovanie mesta, a taktiež ho začali zahŕňať do územného a strategického plánovania miest. V posledných rokoch prebieha v Slovenskej republike trend suburbanizácie, ktorý môžeme sledovať v niektorých okresoch výraznejšie a v niektorých menej výrazne. Taktiež v mestách prebiehajú určité migračné trendy, ktoré sú pre danú oblasť veľmi špecifické, čo platí aj v meste Skalica.

V našom príspevku by sme chceli predstaviť dlhodobé migračné smery a trendy smerov v meste Skalica. Dáta pre celý spracovaný príspevok nám poskytol Mestský úrad v Skalici. Jedná sa o pokračovanie dlhodobého výskumu migračných smerov v nami skúmanom meste (Slavík, et al., 1994, Slavík, 2000). Hlavnou zložkou je zhodnotenie a spracovanie dát migračných smerov mesta Skalica. Na základe spracovaných dát sme vyhotovili niekoľko mapových a grafických výstupov. Dáta nám dokázali, že v meste Skalica začalo prebiehať niekoľko zaujímavých trendov smerov migrácie, ktoré sú detailne vysvetlené a popísané. Špeciálnou kapitolou tejto práce je spracovanie sťahovania v rámci

mesta, ktoré je unikátnym v celej Slovenskej republike. Trendy a fakty, ktoré sme zistili v tomto príspevku sú nápomocné v plánovaní ďalšieho vývoja mesta.

2 Trendy migračných smerov v meste Skalica v rokoch 1990 – 2017

„Migračné trendy sú nový fenomén v migrácii, kedy sa ľudia hromadne sťahujú na určité miesto z určitého dôvodu.“ (Slavíková, 2018). Migrácii sa venuje celý rad článkov v odborných časopisoch a štúdiách, ich zameranie je naozaj pestré. Pre tento príspevok nám boli najbližšie publikácie o migrácii a migračných tokoch (Jurčová, 2007 a 2010) a demografických analýz miest, resp. mestských častí (Bleha, Popjaková, 2007).

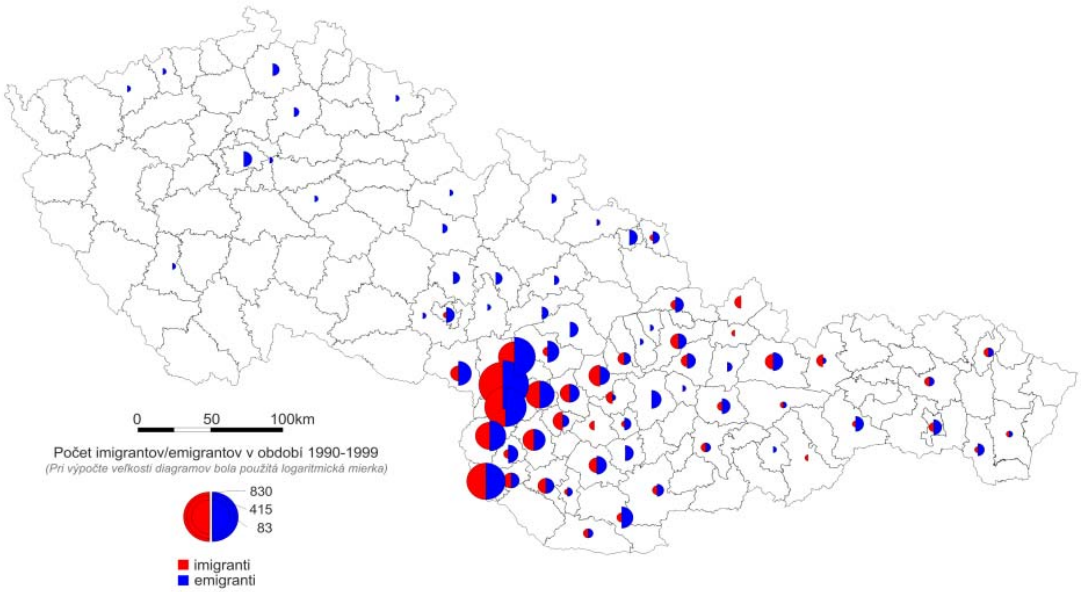


Obr. 1 Názvy obcí v zvolenom výreze najbližšieho okolia mesta Skalica
(*Zdroj dát: ZB GIS 2018*)

V transformačnej etape v rokoch 1990 – 2017 analyzujeme migračné smery na dvoch úrovniach, ide o úroveň okresov a obcí. V prípade Slovenskej republiky ide o okresy podľa územno-správneho usporiadania platného od roku 1996. V prípade okresov Českej republiky musíme poznamenať, že ide len o štatistické jednotky, v systéme verejnej správy boli tieto územné jednotky už zrušené. Celú transformačnú etapu na úrovni okresov analyzujeme najprv po jednotlivých dekádach a na záver je zhodnotená celá etapa, všetky obdobia sú doložené mapami. Migračné smery na úrovni obcí sme zobrazili v tabuľkách a mape a slovne vysvetlili prebiehajúce trendy migrácie za celú etapu. V mapovom výstupe je zachytený len výrez najbližšieho územia (Obrázok 1), kde sú aj obce s názvami, v danej tematickej mape už nie sú, pretože by prekrývali kartodiagramy.

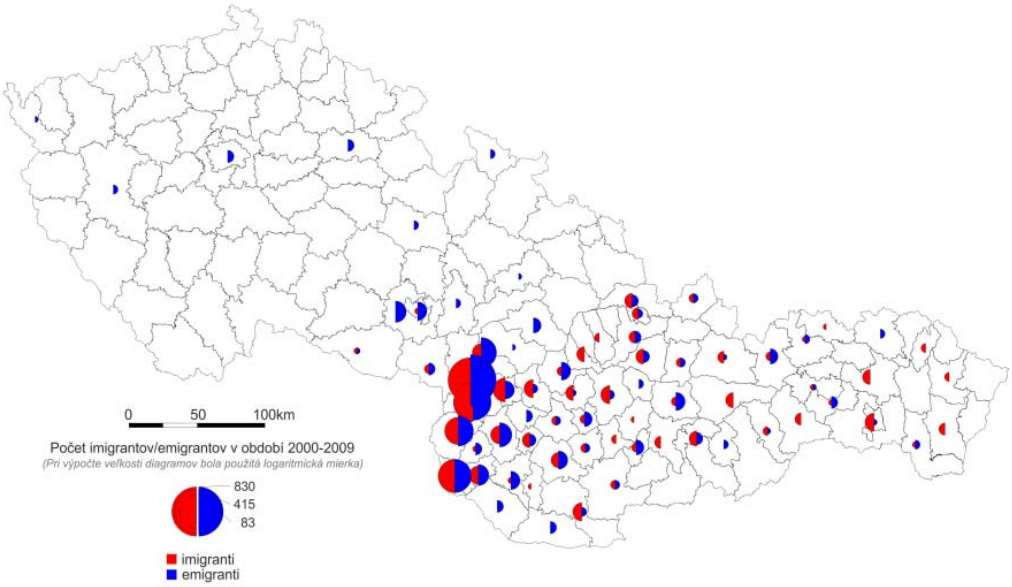
2.1 Úroveň okresov

V prvej transformačnej dekáde, v rokoch 1990 – 1999, nemalo mesto Skalica žiadne migračné väzby len so 14 slovenskými okresmi a 35 českými okresmi. Zo slovenských okresov to boli predovšetkým okresy východného Slovenska. Z hľadiska zastúpenia imigračných aj emigračných väzieb je zásadný rozdiel medzi slovenskými a českými okresmi. V prípade českých okresov dominujú emigračné väzby, u slovenských je vyrovnanejšie zastúpenie imigračných a emigračných väzieb. Táto dekáda je pri úrovni okresov špecifická, pretože v tejto etape máme roky končiaceho Československa a roky začínajúcej Slovenskej republiky a Českej republiky. Nová politická situácia výraznou mierou ovplyvnila aj migračné väzby. Z pohraničného mesta bola zaznamenaná práve v týchto rokoch zvýšená emigrácia do českých okresov. U slovenských okresov sa potvrdili najsilnejšie väzby v rámci okresu Skalica a s najbližšími susednými okresmi. Zo západu na východ sa migračné väzby postupne vytrácajú. V tejto etape je možné sledovať, že Skaličania preferujú sťahovanie do okresov, kde sa nachádza veľké mesto (Bratislava, atď.) (Ravenstein, 1885 a 1889). V rámci Českej republiky výrazne prevažujú emigračné väzby, a to hlavne so susednými okresmi alebo okresmi, kde sa tiež nachádza väčšie mesto (Brno, Praha, atď.). V tejto dekáde sú z daného sledovaného obdobia migračné väzby s Českou republikou najsilnejšie. Ovplyvňuje to fakt, že zvýšený počet obyvateľov Skalice sa po rozdelení republík presťahovalo do Českej republiky (hlavne zmiešané manželstvá).



Obr. 2 Migračné smery mesta Skalica podľa okresov v rokoch 1990 – 1999
(*Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica*)

V dekáde 2000 – 2009 nemalo mesto Skalica migračné väzby so 4 okresmi Slovenskej republiky a 49 okresmi Českej republiky. Zatiaľ čo západná a stredná časť Slovenskej republiky má pomerne vyrovnaný podiel imigrácie a emigrácie, tak na východe republiky môžeme pozorovať zaujímavý zvrat. Oproti minulému desaťročiu začala prevažovať imigrácia nad emigráciou„ *Dochádza k „otváraniu" východnej časti Slovenska a nárastu počtu migrácii do západnej časti."* (Kakaš, 2017). Existencia veľkého zamestnávateľa INA (teraz Schaeffler) pritiahla pracovné sily aj z týchto okresov. Tento trend teda nie je viditeľný len pri okresoch s veľkými okresnými mestami (Bratislava, Trnava), ale začína sa presúvať aj do mesta Skalica. Pri pohľade na Českú republiku sledujeme obdobie stagnácie migrácie. Dominujú emigračné väzby s najbližším okolím.

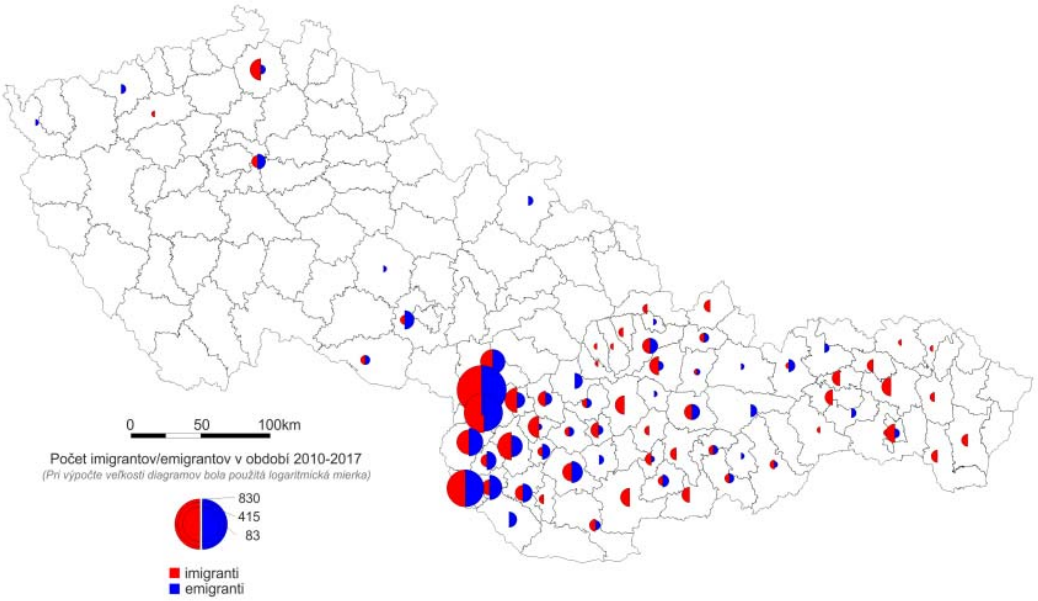


Obr. 3 Migračné smery mesta Skalica podľa okresov v rokoch 2000 – 2009
(*Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica)*

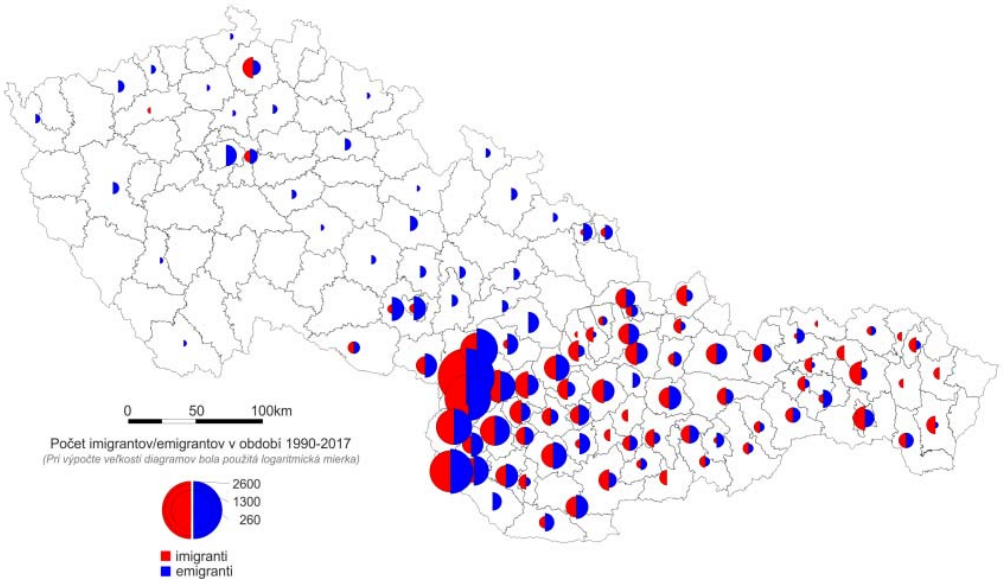
V dekáde 2010 – 2017 nemala Skalica migračné väzby s 5 okresmi Slovenskej republiky a 51 okresmi Českej republiky. Pozorujeme, že imigrácia z východu republiky je ešte výraznejšia ako v predchádzajúcej dekáde a začala výrazne dominovať. Predpokladáme pokračovanie tohto trendu aj v najbližších rokoch. Väzby s okresmi Českej republiky naďalej stagnujú, ale objavili sa tu aj okresy, kde je výraznejšia imigrácia (Česká Lípa).

Sumárna mapa celej etapy prezentuje situáciu v rokoch 1990 – 2017. Mesto Skalica nemalo migračné väzby len s jedným okresom (Medzilaborce) Slovenskej republiky a 23 okresmi Českej republiky. V sumárnej mape celého obdobia je viditeľná prevažujúca emigrácia s Českou republikou. Východ Slovenskej republiky sa vyznačuje dominantnou imigráciou a taktiež sa k nim pridali aj okresy, kde je málo pracovných príležitostí (Námestovo, Levice, atď.). Vyskytujú sa tu aj okresy, kde dominovala imigrácia len kvôli veľmi málo obyvateľom, ktorí

predpokladáme, že sa presťahovali do tohto mesta kvôli uzavretiu manželstva, v menej prípadoch aj kvôli návratu do rodného mesta (rozvod, dožitie, atď.). Na mape je pozorovateľná jednoznačná dominancia migračných väzieb s najbližšími okresmi a Bratislavou.



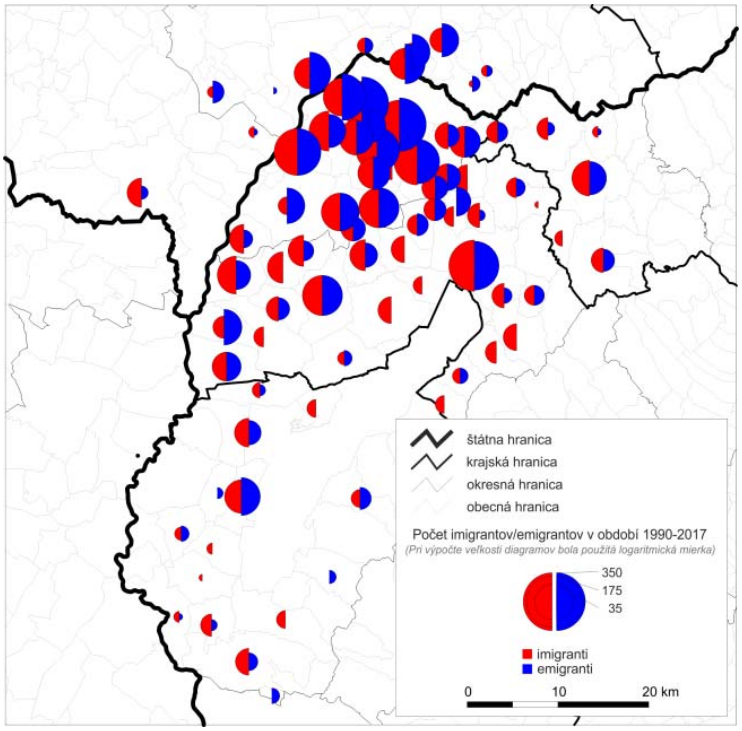
Obr. 4 Migračné smery mesta Skalica podľa okresov v rokoch 2010 – 2017
(*Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica*)



Obr. 5 Migračné smery mesta Skalica podľa okresov v rokoch 1990 – 2017
(*Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica*)

2.2 Obecná úroveň

Za celú etapu 1990 – 2017 na Mape 6 pozorujeme vyrovnanú imigráciu s emigráciou takmer u všetkých obcí. Najsilnejšie migračné väzby pozorujeme hlavne v suburbánnej zóne meste Skalica. *„Väčšina vnútorných migrácii na Slovensku sa realizuje v rozmedzí 5 – 25 km od pôvodného bydliska." (*Kakaš, 2017). Prebiehajúci trend dokazuje aj prvý migračný zákon Ravensteina (1885 a 1889), ktorý hovorí, že väčšina obyvateľstva sa sťahuje na krátke vzdialenosti. Pri menších vidieckych obciach v okrese Senica a Malacky dominuje za celé obdobie len imigrácia. Pravdepodobne išlo o jednorazové presťahovanie odôvodnené manželstvom alebo rozvodom. Keďže je Skalica pohraničným mestom, tak väzby s najbližšími obcami a mestami Českej republiky, sú samozrejmosťou. Dominujú hlavne väzby s najbližším mestom Hodonín (Slavík, 2000).



Obr. 6 Migračné smery mesta Skalica podľa obcí v rokoch 1990 – 2017
(*Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica*)

V Tabuľke 1 môžeme pozorovať, že hodnotu vyše 100 imigrantov dosiahlo 7 obcí okresu Skalica, mesto Senica a hlavné mesto Slovenskej republiky Bratislava. Najvýraznejšiu imigračnú väzbu má mesto Skalica so susedným mestom Holíč, ktorý má dvojnásobok imigrácie oproti Bratislave.

Tab. 1 Top mestá a vidiecke obce s viac ako 100 imigrantami do mesta Skalica v rokoch 1990 – 2017 *(Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica*

P.	MESTO, obec	Okres	1990 – 2017
1.	HOLÍČ	Skalica	742
2.	BRATISLAVA	Bratislava	371
3.	SENICA	Senica	272
4	GBELY	Skalica	185
5	Mokrý Háj	Skalica	179
6	Kopčany	Skalica	152
7	Vrádište	Skalica	140
8	Radošovce	Skalica	132
9	Popudinské Močidlany	Skalica	109
Spolu			**2282**

V Tabuľke 2 sa začali s vyššími počtami emigrácie vyskytovať aj obce v suburbánnej zóne mesta Skalica. Touto tabuľkou teda dokazujeme, že v zázemí mesta Skalica jednoznačne prebieha proces suburbanizácie.

Tab. 2 Top mestá a vidiecke obce s viac ako 100 emigrantami do mesta Skalica v rokoch 1990 – 2017 *(Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica)*

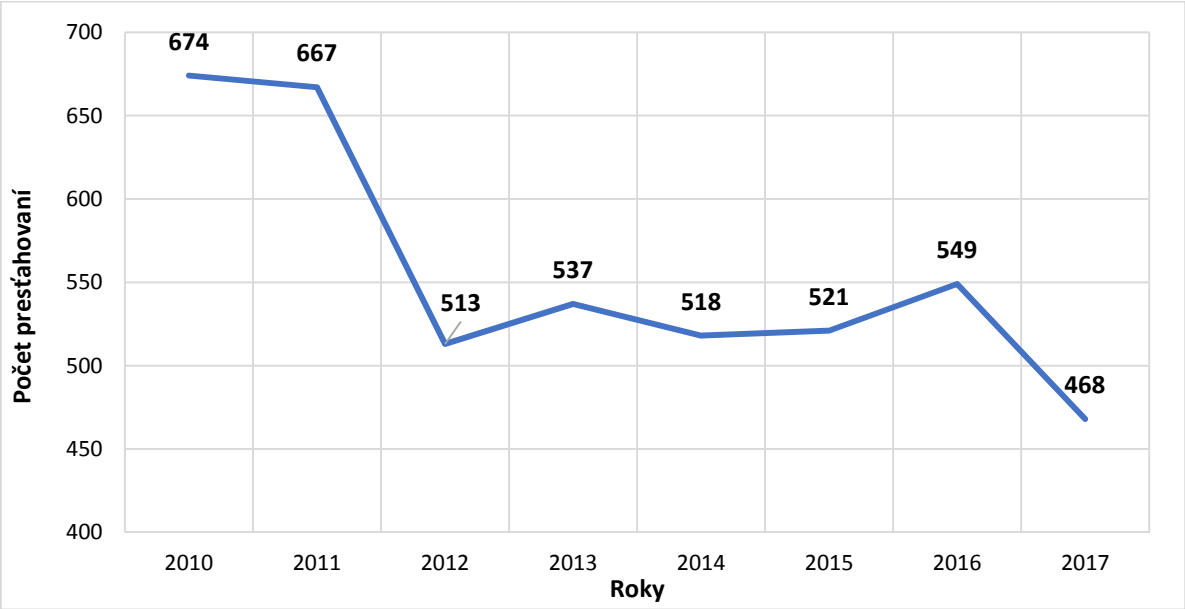
P.	MESTO, Obec	Okres	1990 – 2017
1.	HOLÍČ	Skalica	719
2.	BRATISLAVA	Bratislava	411
3.	Mokrý Háj	Skalica	370
4.	Vrádište	Skalica	294
5.	Kopčany	Skalica	193
6	SENICA	Senica	188
7	GBELY	Skalica	152
8	Popudinské Močidlany	Skalica	130
9	Radošovce	Skalica	121
10	Kátov	Skalica	120
Spolu			**2698**

3 Sťahovanie v rámci mesta Skalica v rokoch 2010 – 2017

Údaje za sťahovanie v rámci mesta sú spravidla málo dostupné, preto sa v odbornej literatúre môžeme stretnúť len s ojedinelými príkladmi analýz sťahovania obyvateľov v rámci uličného systému vybraného mesta (Marszal, 2001). V tejto kapitole sme zosumarizovali základné poznatky o vývoji sťahovania v rámci mesta a hlavné trendy a zaujímavosti za posledné roky.

Vývoj počtu presťahovaní v rokoch 2010 – 2017 v rámci mesta ukazuje Graf 1. Za sledované posledné roky najvyššie počty presťahovaní boli v rokoch 2010 a 2011, kedy počet presťahovaní dokonca presiahol hranicu 650 obyvateľov. V ďalších rokoch mal vývoj klesajúcu tendenciu. Za sledované roky 2010 – 2017 sa celkovo presťahovalo 4 447 ľudí, čo je cca 30% z celkového počtu obyvateľov

mesta. Pri porovnaní počtov presťahovaní v rámci mesta s objemom migrantov sme zistili, že v posledných rokoch počet presťahovaní v meste výrazne prevažuje sumu migrantov.



Graf 1 Vývoj počtu presťahovaní v meste Skalica v rokoch 2010 – 2017
(*Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica*)

Zvýšená intenzita sťahovania v rámci mesta bola spôsobená priaznivou situáciou vo výstavbe rodinných, ale aj bytových domov v meste Skalica. Už dlhšie obdobie sa v meste Skalica realizuje výstavba novej vilovej štvrte, kde má byť celkovo až 450 rodinných domov. Samospráva mesta iniciovala aj výstavbu bytových domov vo vybraných lokalitách, spravidla v doterajších sídliskách resp. na ich okraji.

Tab. 3 Najobývanejšie ulice mesta Skalica v roku 2017
(Zdroj dát: Evidencia obyvateľstva, Mestský úrad Skalica)

Ulica	**Počet obyvateľov**
Dr. Clementisa	1 398
Pod Hájkom	1 336
Pelíškova	995
L. Svobodu	857
Mallého	827
SNP	596
Záhradná	534
Koreszkova	516
Spolu	7 059

Uličný systém mesta Skalica tvorí v súčasnosti 117 ulíc, z nich väčšina je skoncentrovaná v intraviláne mesta. Rozmiestnenie obyvateľstva v týchto uliciach je veľmi rozdielne. V Tabuľke 3 ukazujeme ulice, v ktorých býva viac ako 500 obyvateľov. Takýchto ulíc je až 8, ide o sídliskové ulice. Spolu v nich býva vyše 7 000 obyvateľov, čo je takmer polovica obyvateľov mesta. Medzi nimi vynikajú počtom obyvateľov dve ulice, a to Dr. Clementisa a Pod Hájkom. Vkaždej z nich býva vyše 1 300 obyvateľov. V ďalších 29 uliciach je počet obyvateľov v rozpätí 100 – 499. Menej ako 10 obyvateľov býva v 7 uliciach a niekoľko ulíc je dokonca neobývaných (Dr. Schaefflera – v priemyselnej zóne, bočné uličky v historickom jadre, atď.).

4 Záver

Hlavnou úlohou bolo digitalizovanie a vytvorenie rozsiahlej analýzy migračných smerov, ktorá nám vytvorila jedinečnú databázu migračných smerov mesta. Spracovanie máp migračných smerov so Slovenskou aj Českou republikou je tiež ojedinelé a ich spracovanie za také dlhé obdobie ešte doposiaľ nebolo uskutočnené.

Dokázali sme, že v nami spracovanej etape dochádza v meste Skalica k niekoľkým zaujímavým trendom migračných smerov. Celé toto obdobie bolo sprevádzané obrovským prepadom migračných väzieb s Českou republikou a celkovej stagnácie migrácie. Pri porovnávaní týchto etáp sme objavili zaujímavý trend sťahovania obyvateľstva z východu Slovenskej republiky do mesta Skalica, kde je ako na jedinej ucelenej časti pozorovaná prevyšujúca imigrácia nad emigráciou. V tomto príspevku sme taktiež dokázali, že v suburbánnej zóne mesta Skalica jednoznačne prebieha proces suburbanizácie, ktorý však nie je taký extrémny ako pri niektorých iných mestách Slovenskej republiky. Imigrácia z najbližšieho okolia do mesta Skalica prebieha hlavne z menších obcí susedných okresov. Pri sťahovaní v rámci mesta bola zistená oveľa väčšia intenzita oproti prebiehajúcej imigrácii a emigrácii, ktorá ju v niektorých prípadoch vysoko prevýšila.

5 Literatúra

Bleha, B., & Popjaková, D. (2007). Migrácia ako dôležitý determinant budúceho vývoja na lokálnej úrovni–príklad Petržalky. Geografický časopis, 59(3), 265-291.

Jurčová, D. (2007, March). Migrácia–výzvy, problémy, otázniky. In Forum statisticum Slovacum, 124-128.

Jurčová, D. (2010). Migračné toky v Slovenskej republike. Bratislava, Inštitút informatiky a štatistiky.

Kakaš, A. (2017). Vzdialenosť ako faktor výskumu vnútornej migrácie Slovenska. Geografický časopis, 69(1), 41-60.

Marszal, T., ed. (2001). Atlas of polish emigration in France. Atlas Polonii francuskiej. Lódź : University of Lódź, 73.

Ravenstein, E. G. (1885). The laws of migration. Journal of the Statistical Society of London, 48, 167-235.

Ravenstein, E. G. (1889). The laws of migration. Journal of the Royal Statistical Society, 52, 241-301.

Slavík, V. et al. (1994). Postavenie Skalice ako centra osídlenia (Expertízna analýza). Bratislava, Katedra humánnej geografie a demografie Prírodovedeckej fakulty Univerzity Komenského, 153.

Slavík, V. (2000). Pohraničný región Skalica a cezhraničná spolupráca. Životné prostredie, 34 (1), 25-29.

Slavíková, N. (2018). Trendy migračných smerov mesta Skalica v dlhodobej perspektíve. Bakalárska práca, Katedra humánnej geografie a demografie PRIF UK, 61.

Modelovanie slovenských parlamentných volieb
Modelling of Slovak Parliamentary Elections

Dominika Zorvanová
Univerzita Komenského Bratislava Fakulta matematiky, fyziky a informatiky, magisterské štúdium, prvý ročník
Comenius University Bratislava Faculty of Mathematics, Physics and Informatics
dominika.zorvanova@gmail.com

Abstrakt: *Voľby do Národnej rady Slovenskej republiky možno označiť za kľúčový element slovenského politického systému. Tieto voľby už od roku 1998 využívajú pomerný volebný systém s jedným volebným obvodom, ktorý so sebou prináša ako pozitíva tak aj negatíva. Výhodou celoštátneho obvodu je jeho miera proporcionality, na druhej strane jeho deficitom je neadekvátne zastúpenie regiónov a neúčinné využitie prednostných hlasov. Tieto nedostatky môžeme potlačiť zvýšením počtu volebných obvodov. V článku sa preto zaoberáme modelovaním zmeny počtu volebných obvodov z pôvodného celoštátneho na obvody kopírujúce regionálne a župné členenie Slovenska. Na modelovanie využívame výsledky volieb do Národnej rady Slovenskej republiky z roku 2016. Podstata modelovania údajov spočíva najmä v možnostiach ich praktického uplatnenia. V praxi to znamená, že pri identickom hlasovaní by mohlo nastať odlišné obsadenie parlamentu iba vďaka zmene vymedzenia volebných obvodov.*

Abstract: *Elections to the National Council of the Slovak Republic can be described as a key element of the Slovak political system. Since 1998, these elections have been using a proportional representation system with a single constituency, which brings both positive and negative aspects. The advantage of the nationwide constituency is its degree of proportionality, on the other hand its deficit is the inadequate representation of the regions and the ineffective use of the preferential votes. We can suppress these shortcomings by increasing the number of constituencies. The article therefore deals with modelling of changes in the number of constituencies from the original nationwide to constituencies which copy the regional and counties of Slovakia. For modelling we use the results of elections to the National Council of the Slovak Republic from 2016. The main idea of data modelling lies especially in the possibilities of their practical application. In practice, this means that with the identical vote there could have been different for the parliamentary composition only because of the change of the constituencies.*

Kľúčové slová: *volebná metematika, pomerný volebný systém, volebný obvod, parlamentné voľby.*

Key words: *voting mathematics, proportional representation system, constituencies, parliamentary election.*

Slovenská štatistická a demografická spoločnosť

Prírodovedecká fakulta Univerzity Komenského v Bratislave

Výpočtová štatistika 2018

Analytika očami profesionálov – pásmo prednášok pre študentov

Miesto konania: **Prírodovedecká fakulta Univerzity Komenského v Bratislave, pavilón B1, miestnosť CPS+, Mlynská dolina, Bratislava**

***Dátum konania*: 7. decembra 2018 piatok, 9,00 - 12,30 hod.**

Témy prednášok:

Mgr. Katarína Žiaková: Absolventi na trhu práce

Ing. Ján Kavec: Informačný systém o cene práce (ISCP) a jeho podsystémy

Ing. Marek Bičár: Umelá Inteligencia v Marketingu - Ako zlepšiť zákaznícku skúsenosť?

Informácie o prednášajúcich

Mgr. Katarína Žiaková absolvovala Fakultu matematiky, fyziky a informatiky Univerzity Komenského v Bratislave, odbor aplikovaná matematika. Po skončení štúdia pracovala na Štatistickom úrade SR na sekcii všeobecnej metodiky a registrov. V súčasnosti je členom analytického oddelenia firmy Profesia spol. s.r.o, ktorá prevádzkuje najväčší pracovný portál na Slovensku. Vo svojej práci sa venuje štatistickému spracovaniu údajov, analýze a interpretácii dát týkajúcich sa dopytu a ponuky na trhu práce.

Ing. Ján Kavec pracuje vo výskumno-štatistickej a poradensko-konzultačnej spoločnosti TREXIMA Bratislava, ktorá od roku 1992 vyvíja a úspešne aplikuje nové, efektívne metódy v oblasti výskumu miezd podľa konkrétnych zamestnaní, zamestnanosti, pracovných a sociálnych podmienok zamestnancov v SR a jej jednotlivých regiónov. Súčasťou originálnych aktivít spoločnosti je tvorba a implementácia komplexných sústav povolaní a kvalifikácií, realizácia analyticko-štatistických štúdií a prognózovanie vývoja na trhu práce. Do TREXIMY Bratislava nastúpil po absolvovaní štúdia na Ekonomickej univerzite v Bratislave.

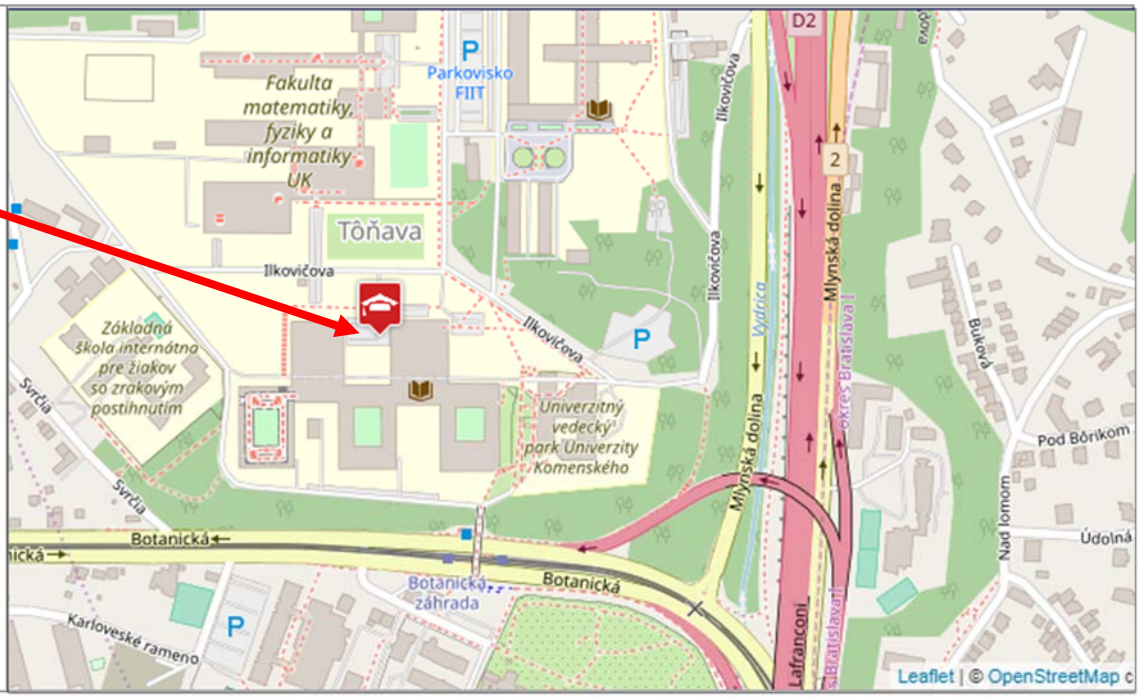
Ing. Marek Bičár je Business Consultant poradenskej a softvérovej spoločnosti SAS Slovakia, s.r.o. od roku 2013 a od začiatku sa venuje problematike Customer Intelligence. Celý svoj profesijný život pracuje v oblasti riadenia vzťahov so zákazníkmi, najprv ako Segment manager v Tatra banke, neskôr v Slovenskej sporiteľni. Marek je absolventom Ekonomickej univerzity v Bratislave.

Informácie o mieste konania:

Prírodovedecká fakulta Univerzity Komenského v Bratislave

Web: http://www.fns.uniba.sk

PRIF UK
Mlynská dolina
Ilkovičova 6
842 15
Bratislava 4



GPS súradnice
šírka: 48°09'00,12" | dĺžka: 17°04'17,85"

Dopravné spojenie v Bratislave www.imhd.sk
zastávka: **Botanická záhrada** pre električky číslo 4, 5, 6, 9 alebo autobusy číslo 29 a 32.

Generálny partner podujatia

sas THE POWER TO KNOW.



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

Výpočtová štatistika 2018

zborník abstraktov z 26. medzinárodného vedeckého seminára **Výpočtová štatistika 2018** a zborník abstraktov a článkov z konferencie **Prehliadka prác mladých štatistikov a demografov 2018** konaných v Bratislave v dňoch 6. – 7. decembra 2018

Editori: Iveta Stankovičová, Janka Medová

Vydala Slovenská štatistická a demografická spoločnosť, Bratislava v roku 2018.
Prvé vydanie

ISBN ISBN 978-80-88946-83-0

ISBN 978-80-88946-83-0