Slovenská štatistická a demografická spoločnosť
Slovak Statistical and Demographic Society

FORUM STATISTICUM SLOVACUM

Recenzovaný vedecký časopis
Scientific peer-reviewed journal

Číslo/Issue: 2/2021
Ročník/Volume: XVII

FORUM STATISTICUM SLOVACUM

recenzovaný vedecký časopis Slovenskej štatistickej a demografickej spoločnosti
scientific peer-reviewed journal of the Slovak Statistical and Demographic Society

Ročník/Volume: XVII (2021) **Číslo/Issue: 2**

Editori / Editors

Iveta Stankovičová (Fakulta managementu Univerzity Komenského v Bratislave, Slovensko)
Martin Boďa (Ekonomická fakulta Univerzity Mateja Bela v Banskej Bystrici, Slovensko)

Redakčná rada / Editorial Board

Branislav Bleha (Prírodovedecká fakulta Univerzity Komenského v Bratislave, Slovensko)
Boris Burcin (Prírodovedecká fakulta Univerzity Karlovej v Prahe, Česko)
Joanna Dębicka (Fakulta manažmentu, informatiky a financií Ekonomickej univerzity vo Wrocławi, Poľsko)
Estefanía Mourelle Espasandín (Fakulta ekonómie a podnikania Univerzity v La Coruña, Španielsko)
Stanislav Katina (Ústav matematiky a štatistiky Masarykovej univerzity v Brne, Česko)
Jana Kubanová (Fakulta ekonomicko-správna Univerzity Pardubice, Česko)
Dagmar Kusendová (Prírodovedecká fakulta Univerzity Komenského v Bratislave, Slovensko)
Viera Labudová (Fakulta hospodárskej informatiky Ekonomickej univerzity v Bratislave, Slovensko)
Jitka Langhamrová (Fakulta informatiky a štatistiky Vysokej školy ekonomickej v Prahe, Česko)
Ivan Lichner (Ekonomický ústav Slovenskej akadémie vied, Slovensko)
Tomáš Löster (Fakulta informatiky a štatistiky Vysokej školy ekonomickej v Prahe, Česko)
Janka Medová (Fakulta prírodných vied Univerzity Konštantína Filozofa v Nitre, Slovensko)
Silvia Megyesiová (Podnikovohospodárska fakulta v Košiciach Ekonomickej univerzity v Bratislave, Slovensko)
Oľga Nanásiová (Fakulta elektrotechniky a informatiky Slovenskej technickej univerzity v Bratislave, Slovensko)
Viliam Páleník (Ekonomický ústav Slovenskej akadémie vied, Slovensko)
Marek Radvanský (Ekonomický ústav Slovenskej akadémie vied, Slovensko)
Hana Řezanková (Fakulta informatiky a štatistiky Vysokej školy ekonomickej v Prahe, Česko)
Ľubica Sipková (Fakulta hospodárskej informatiky Ekonomickej univerzity v Bratislave, Slovensko)
Mária Stachová (Ekonomická fakulta Univerzity Mateja Bela v Banskej Bystrici, Slovensko)
Anna Tirpáková (Fakulta prírodných vied Univerzity Konštantína Filozofa v Nitre, Slovensko)
Vladimír Úradníček (Ekonomická fakulta Univerzity Mateja Bela v Banskej Bystrici, Slovensko)
Mária Vojtková (Fakulta hospodárskej informatiky Ekonomickej univerzity v Bratislave, Slovensko)
Tomáš Želinský (Ekonomická fakulta Technickej univerzity v Košiciach, Slovensko)

Vydavateľ: Slovenská štatistická a demografická spoločnosť, Miletičova 3, 824 67 Bratislava, Slovensko. **Publisher:** Slovak Statistical and Demographic Society, Miletičova 3, 824 67 Bratislava, Slovakia. **Adresa redakcie/Editorial office:** Miletičova 3, 824 67 Bratislava, Slovakia. **IČO/Company ID:** 00178764. **DIČ/Tax ID:** 2021504276. **Mailový kontakt/E-mail contact:** adm.ssds@ssds.sk **Web site/Webové sídlo:** http://www.ssds.sk/

Registráciu vykonalo Ministerstvo kultúry Slovenskej republiky. **Dátum registrácie:** 27. júla 2005. **Evidenčné číslo:** EV 3287/09. **Tematická skupina:** B1. **Periodicita:** minimálne dvakrát ročne. **ISSN:** 1336-7420.
Registered by the Ministry of Culture of the Slovak Republic. **Date of registration:** 27 July 2005. **Registration No:** EV 3287/09. **Topic group:** B1. **Periodicity:** at least two issues per year. **ISSN:** 1336-7420.

FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 2 pp 1-11

Rozcestník moderních přístupů směřujících k automatizované detekci anomálií v časových řadách
A guidepost to modern approaches directed towards automated anomaly detection in time series

Jitka Bartošová, Vladislav Bína, Vladimír Přibyl
Vysoká škola ekonomická v Praze, Fakulta managementu, Jarošovská 1117/II, 377 01 Jindřichův Hradec, Česká republika
University of Economics in Prague, Faculty of Management, Jarošovská 1117/II, 377 01 Jindřichův Hradec, Czech Republic
jitka.bartosova@vse.cz, vladislav.bina@vse.cz, vladimir.pribyl@vse.cz

Abstrakt: *S ohledem k velkému množství dat, která jsou k dispozici vedoucím pracovníkům v takřka každé současné organizaci se automatizované monitorování základních procesů zajišťujících chod organizací postupně stává naprostou nutností. Konečným posuzovatelem a rozhodovatelem musí vždy být manažer s příslušnou zodpovědností, nicméně alespoň pro základní a hrubou přípravu sestav informací o odchylkách (anomáliích, či alertech) v klíčových procesech lze využít širokou škálu matematických a statistických přístupů implementovaných ve vhodném prostředí pro zpracování dat, či ideálně přímo integrovaných v rámci informačního systému. Tento článek přináší alespoň základní přehled přístupů nejčastěji využívaných pro detekci anomálií.*

Abstract: *Given a large amount of data available to executives in almost every contemporary organization, automated monitoring of the basic processes in organizations is gradually becoming an absolute necessity. The final assessor and decision-maker must always be a manager with appropriate responsibilities. However, at least for basic and rough preparation of reports concerning deviations (anomalies or alerts) in key processes, we can use a wide range of mathematical and statistical approaches implemented in an appropriate data processing environment, or ideally integrated directly in the information system. This article provides at least a basic overview of the approaches frequently used for anomaly detection.*

Klíčová slova: *detekce anomálií, časové řady, rozklad, umělá inteligence*

Key words: *anomaly detection, time series, decomposition, artificial intelligence*

1 Úvod

Obrovský rozvoj informačních a komunikačních technologií v prvních dekádách 21. století otevřel podnikatelské a organizační sféře pestrou škálu nových přístupů a výzev. Poslední roky přinesly nástup moderních a slibných technologicky náročných možností v rámci širokých oblastí Průmyslu 4.0 a fenoménu big data. Patří sem přístupy, jakými jsou např. digitální dvojče, průmyslový internet věcí, adaptivní automatizace, robotizace, 3D tisk, chytrá logistika a chytrá výroba (viz např. Evjemo et al., 2020), ale i v době světové pandemie významně posilující oblast e-businessu (Elhrim et al., 2020). Některé z hojně skloňovaných moderních přístupů a zdánlivě slibných oblastí jsou dost

možná prázdnými pojmy a v testu času se neprosadí. Nicméně společným jmenovatelem je stále sílící potřeba zpracování velkého objemu dat, jež se neobejde bez alespoň částečného automatického předzpracování, které zefektivní či přímo umožní rozhodování příslušným manažerům.

Přirozeně s tím souvisí i rozvoj služeb nabízejících zpracování velkých objemů dat na částečně automatizované bázi. Vzhledem k tomu, že v podnikové a organizační sféře se z velké většiny jedná o činnost pravidelnou, nabývá na důležitosti sledování prosté skutečnosti, zda je chod jednotlivých procesů v organizaci v pořádku, či zda nedochází k nějakým zásadním výkyvům představujícím riziko pro fungování organizace. V případě výkyvů, či anomálií je pak klíčové dobré vymezení, hledání souvislostí a případně i kauzálních vztahů.

2 Automatizace hledání anomálií

Ukazuje se, že detekce odlehlých hodnot a anomálií v datech je komplexní a netriviální problém jak z hlediska lidského rozhodovatele, tak z hlediska počítačové podpory manažerského rozhodování (Tabesh et al., 2019). Pro člověka je zpravidla významnou potíží velké množství dat, na druhou stranu ale zpravidla se snadno řekne, která hodnota se jeví „normální“ a která je zřejmou „anomálií“. Do jisté míry komplementární je algoritmický přístup, velké množství dat není potíží, ale spolehlivost detekce a interpretace anomálních hodnot leckdy pokulhávají. Automatizovaná detekce anomálií je oblastí řešenou již několik desetiletí a mezi první dobře využitelné metody týkající se časových řad patřily např. Holt-Wintersovo exponenciální vyrovnávání (Chatfield, 1978) a ARIMA přístupy (Box a Jenkins, 1970).

Narůstající poptávka po řešení z oblasti detekce anomálií způsobila výrazný rozvoj nejen v oblasti teoretické, ale i v oblasti rozvoje komponent BI (business intelligence) řešení. Vlastní přístupy detekce anomálií přinesly velké společnosti jako je Twitter (Kejariwal, 2015), LinkedIn (Luminol, 2018), Facebook (Taylor a Letham, 2018), Google (Brodersen et al., 2015). Zpravidla byly systémy těchto společností volně k dispozici alespoň v nějaké průběžné variantě, mezi slibnými a velmi využívanými zůstává např. Prophet. Kromě toho na automatizované detekci anomálií alespoň částečně postavila svůj byznys řada analytických a BI společností, sem patří například společnost Anodot (Toledano et al., 2018, a https://www.anodot.com), nebo část BI řešení od firmy Microsoft: Microsoft Cognitive Service založená na Spectral Residual (SR) a Convolutional Neural Network (CNN) (viz Ren et al., 2019 a https://azure.microsoft.com/en-us/services/cognitive-services), či např. produkty společností Avora (https://avora.com) a CrunchMetrics (https://www.crunchmetrics.ai).

Ukazuje se, že zdůraznění, případně analýza neobvyklých a zajímavých událostí, které v určitém smyslu vybočují z dosavadního (či očekávaného) průběhu časové řady údajů, se stává velmi důležitou činností ve všech oblastech, kde je zpracováváno velké množství dat. Analýza se napříč oblastmi nasazení soustředí zejména na metriky podporující řízení důležitých podnikových procesů, zákaznických vztahů, či finančních ukazatelů a z toho plynou i vlastnosti časových řad, které obvykle agregují údaje na denní bázi, případně řídčeji a často vykazují týdenní a obvykle i další (např. roční) sezonní složky.

Metody detekce mohou jednak zkoumat časové řady jednotlivě a jednak se mohou zabývat skupinou časových řad jako celkem. V této části se budeme věnovat pouze řadám zahrnujícím jedinou proměnnou, struktury více časových řad budou zmíněny v kapitole 6. Přístupy k detekci anomálií lze dále členit na metody zohledňující časovou informaci a metody, které primárně časovou informaci nezahrnují. Dalším atributem používaných metod je i možnost použití detekce anomálií v reálném čase přímo při zpracování transakcí, což u drtivé většiny metod zahrnuje nutnost nového přepočítání modelu dané veličiny. Zároveň se vlastně jedná o výpočet založený na predikci budoucí hodnoty časové řady, na rozdíl od přístupů, které na základě modelu časové řady identifikují anomálie porovnáním celého reálného průběhu řady s jejím teoretickým modelem.

Při hledání anomálií v jednotlivých časových řadách tak můžeme využívat modelových přístupů (s modelováním hodnot v průběhu řady, či s predikcí budoucí hodnoty) a dále metody založené na hustotě jednotlivých hodnot. Do první kategorie spadají metody založené na regresním modelování, na dekompozici (např. STL, ETS) a také ARIMA přístupy, do druhé kategorie pak např. přístupy izolačních lesů, či přístupy založené na neuronových sítích.

3 Statistické přístupy k detekci anomálií

Chování časových řad zahrnuje řadu typických vzorů a základní přístupy k jejich modelování jsou založeny na rozdělení řad do základních komponent, kterými jsou trend, sezónní složka a cyklická složka. Při rozkladu časové řady se obvykle trendová a cyklická složka spojují do jediné „trend-cyklové“ složky, která se zpravidla jednoduše (a mírně nepřesně) označuje jako trend. Obvyklým způsobem tedy je rozklad na trend (myšleno trend-cyklus), sezónní složku a reziduální složku (tedy část variability dat, kterou model nedokáže vysvětlit). V případě řad s větší frekvencí (nejméně denní), může popis zahrnovat i více sezónních komponent odpovídajících různé periodicitě.

3.1 ETS dekompozice

ETS rozklad vychází z přístupů vytvořených ke konci padesátých let a založených na exponenciálním vyhlazování s využitím vážených průměrů předchozích pozorování. Samotné exponenciální vyrovnávání (Brown, 1959) bylo rozšířeno v přístup zahrnující modelování lineárního trendu (Holt, 1957), který byl dále rozšířen o modelování sezónní složky (Winters, 1960). Postupem času byla rozvinuta široká škála ETS modelů, jejich přehled lze nalézt např. u Hyndmana a Khandakara (2008). K identifikaci anomálií se využívá namodelované hodnoty časové řady a predikčních intervalů s uživatelsky zvolenou hladinou spolehlivosti.

3.2 STL dekompozice

Jedním z univerzálních a poměrně robustních přístupů je dekompozice STL (Cleveland a Cleveland, 1990), která je založena na klasickém rozkladu na sezónní a trendovou složku, přičemž trendová část je modelována s využitím LOESS odhadu (Cleveland, 1979). Na rozdíl od některých dalších přístupů (jako je X11, či SEATS dekompozice), tento rozklad umí pracovat s obecnějšími typy sezónních komponent (ne pouze s čtvrtletními nebo měsíčními) a je poměrně robustní s ohledem k odlehlým hodnotám, tedy jednotlivá pozorování nemají výrazný vliv na odhady složek (Hyndman a Athanasopoulos, 2018). Zároveň se sezónní složka může v čase měnit a rychlost změny spolu s hladkostí trendové složky je možné v některých implementacích měnit. Díky využití LOESS odhadu je metoda velmi úspěšná v popisu trend-cyklové složky, tedy ve chvílích, kdy je zřetelný dlouhodobý trend. Z uvedených důvodů tato a následující metoda patří mezi základní používané přístupy pro detekci anomálií.

3.3 Twitter dekompozice

Jedná se o přístup velmi podobný STL rozkladu, využívá stejný přístup k práci se sezónní složkou a je použit v balíčku AnomalyDetection společnosti Twitter (Hochenbaum et al., 2017). Na rozdíl od STL však trendovou složku modeluje s využitím (lokálního výpočtu) mediánu. Z toho důvodu funguje velmi dobře, pokud je dlouhodobý trend méně významný a dominují krátkodobé sezónní složky.

Dekompozice STL a Twitter lze použít se dvěma přístupy k identifikaci anomálních bodů. A sice s metodami: IQR (Inner Quartile Range) a GESD (Generalized Extreme Studentized Deviate test). Metoda IQR využívá mezikvartilové rozpětí a v základním nastavení jsou anomálie identifikovány za hranicí vzdálenosti trojnásobku od mezikvartilového rozpětí. Jedná se o rychlý přístup, který nevyužívá iterativního výpočtu, jeho slabou stránkou je horší adaptabilita. Flexibilnější GESD přístup dokáže dobře pracovat i s významnými

anomáliemi, které v průběhu výpočtu iterativně odstraňuje. Proto je výpočetně náročnější než IQR přístup, což je patrné zejména v případě většího množství časových řad s velkým množstvím pozorování.

3.4 ARIMA modely

Spolu s přístupem exponenciálního vyrovnávání patří mezi nejpoužívanější metody modelování časových řad. Na rozdíl od dekompozičních přístupů však ARIMA metody nevyužívají popisu trendu a sezónnosti, ale autokorelačních charakteristik modelovaných dat.

Modely Autoregressive Integrated Moving Average zahrnují jednak autoregresní charakter modelu (tedy závislost aktuálně pozorované hodnoty na určeném počtu starších hodnot, jednak integrovanou část spočívající ve využití diferencí tak, aby charakter řady byl stacionární a složku klouzavých průměrů sloužící k modelování aktuálních hodnot jako lineární kombinace chyb několika předchozích hodnot. Přístupy sezonních modelů zahrnují uvedené složky i pro sezónní komponentu časové řady.

Vzhledem k tomu, že plánovaný účel nasazení vyžaduje co nejvyšší míru automatizace, využíváme přístup pojmenovaný jako „automatizovaná ARIMA“ (Hyndman a Khandahar, 2008). Tento přístup využívá testování jednotkového kořene společně s minimalizací AIC a MLE kritérií pro výběr nejvhodnějšího ARIMA modelu. Opět se jedná o univerzální a široce využitelný přístup.

3.5 Prophet

Facebookovský Prophet (Taylor a Letham, 2018) patří rovněž mezi aditivní dekompoziční přístupy. Využívá po částech lineární křivky a logistické přechody k modelování neperiodické složky, dále model zahrnuje sezónní složku a složku zahrnující uživatelem vložené informace o svátcích způsobujících odchylky v časové řadě. Neperiodická složka je odhadována s využitím regresních principů, složka sezónní je modelována s využitím Holt-Wintersova exponenciálního vyrovnávání (Chatfield, 1978). Protože se jedná o přístup k modelování a predikci časových řad, je jeho využití pro detekci anomálií přímočaré. Prophet patří mezi velmi nadějné přístupy, protože se jedná o velmi robustní přístup dobře zvládající chybějící hodnoty, velké sezónní změny, výkyvy v době svátků a prázdnin i výrazné posuny v trendu.

3.6 TBATS modely

Výše uvedené dekompoziční modely v zásadě uvažovaly pouze jednoduchý typ sezónní složky. Mezi přístupy umožňující zahrnout vliv více sezón najednou (v našem případě zpravidla týdenní a roční) patří TBATS modely (De Livera et al., 2011). Pro výpočet používají kombinaci modelu exponenciálního vyhlazování spolu s využitím fourierovských členů k modelování periodických složek a Box-

Coxovy transformace. Na rozdíl od přístupů harmonické regrese TBATS modely umožňují mírnou změnu sezonních komponent s vývojem řady v čase. Tato komplexnost a flexibilita modelů může být velmi užitečná, nicméně využitelnost je výrazně omezena faktem, že v případě delších časových řad je tvorba modelu velmi výpočetně náročná.

4 Metody z oblasti umělé inteligence

Pro detekci anomálií jsou využívány rovněž přístupy zložené na učení bez učitele, jakými jsou například moderní a často používané izolační lesy, či specializované přístupy v rámci umělých neuronových sítí (Recurrent neural networks a vylepšené deep-learningové LSTM autoencoders), nebo googlovský přístup CausalImpact, či využití konvolučních neuronových sítí v řešení společnosti Microsoft. Praktickým úskalím těchto metod může být to, že není k dispozici jejich otevřená implementace vhodná pro zapojení do vyvíjeného systému a nemusí být kompatibilní se zvolenou softwarovou architekturou.

4.1 Izolační lesy

Izolační lesy (Liu et al., 2008) představují přístup učení bez učitele obecně aplikovaný v oblasti detekce anomálií. Princip metody je založen na sestavení souboru izolačních stromů a přes tento soubor je pak průměrováno vypočtené anomální skóre jednotlivých pozorování charakterizující délku cesty k tomuto pozorování. Základní myšlenkou je to, že “normální” body jsou svou hodnotou obdobné jako řada dalších pozorování, tedy mají v sousedství řadu dalších hodnot. K rozlišení takových bodů bude potřeba v drtivé většině vygenerovaných stromů projít velkým množstvím větvení. Naproti tomu se v případě odlehlých hodnot podaří je separovat od ostatních zpravidla průchodem jen několika málo větví. Průměrovaná hloubka průchodu k dané hodnotě stromem tak po použití vhodné normalizace umožňuje identifikovat odlehlé hodnoty.

V případě, že se jedná o identifikaci anomálie v časových řadách, obvykle je potřeba použít vhodné rozšíření této metody tak, aby zohledňovala časovost dat. Jednou z možností je využití klouzavého okna (Ding a Fei, 2013) k zohlednění lokálnosti, resp. časové souvislosti. Nejjednodušší variantou přitom je využití jedné, či několika zpožděných časových řad jako dalších dimenzí pro detekci s využitím základního algoritmu.

4.2 LSTM autoencodery

Neuronové sítě zvané autoencodery reprezentují další přístup učení bez učitele a jsou využívané jako efektivní metody pro kompresi dat. Kromě toho jsou užitečné jako prostředky pro redukci dimenzionality, zobecňování a zpracování obrazů. Detekce anomálií s využitím autoencoderů je založená na tom, že natrénovaný autoencoder dokáže běžná data dobře rekonstruovat, naproti

tomu fakt, že autoencoder v rekonstrukci neuspěje, znamená, že se velmi pravděpodobně jedná o anomální data.

V případě časových řad (a jiných sekvenčních dat) lze využívat LSTM (Long Short-Term Memory) autoencodery (Gers et al., 1999, případně Kulanuwat et al., 2021), které nejsou zatížené nedostatky rekurentních neuronových sítí týkajících se selhání učících algoritmů na mizejících či naopak rostoucích gradientech dlouhých časových řad.

4.3 CausalImpact

CausalImpact (Brodersen et al., 2015) je přístupem využívaným společností Google a je založen na ekonometrických principech Bayesovských strukturálních rovnic pro časové řady. Přístup s využitím časových řad z kontrolní skupiny vytváří syntetickou základní časovou řadu. Anomálie pak nacházíme porovnáním nové, tzv. testové řady s takto vytvořenou syntetickou základní řadou.

5 Moderní implementace důležitých přístupů

V rámci předchozí kapitoly jsme předložili přehled přístupů pro detekci anomálií časových řad a vytipovali sadu dostatečně pokročilých (a tedy i nadějných) přístupů. Alespoň základní možnosti implementace těchto metod v prostředí statistického softwaru R (R Core Team, 2021) se zaměřením na ETS, Arima, STL, Twitter a IF, částečně i na Prophet a TBATS přístupy přinášíme zde.

5.1 Knihovna forecast

Jako základní východisko pro řadu modelů lze využít knihovnu **forecast** (Hyndman et al., 2018). Tato knihovna zahrnuje mimo jiné moderní a robustní implementace přístupů ARIMA (včetně automatizované tvorby modelů přístupem auto.arima), ETS přístupy, STL a Twitter dekompozice i implementaci TBATS modelů a umožňuje na základě vytvořených modelů jednorozměrných časových řad generovat predikční intervaly namodelovaných hodnot, které slouží k identifikaci anomálií.

5.2 Knihovna anomalize

Pro implementaci STL a Twitter přístupu je dobře použitelná knihovna **anomalize** (Dancho a Vaughan, 2020), která umožňuje využívat IQR a GESD přístupy pro detekci anomálních bodů i další parametrizaci.

5.3 Knihovna prophet

Pro využití robustního facebookovského Prophet přístupu dobře zvládajícího sezonní výkyvy, zahrnutí svátků a změny v trendech slouží v prostředí R otevřená knihovna **prophet** (Taylor a Letham, 2020).

5.4 Knihovna solitude jako implementace izolačních lesů

Jako implementace přístupu izolačních lesů může sloužit R balíček **solitude** (Liu et al., 2012). Jedná se o moderní implementaci dobře kompatibilní s pokročilými datovými strukturami v R a využívající paralelizace výpočtu souboru izolačních stromů. Významným úskalím tohoto přístupu je nekorektní zacházení s časovou dimenzí, které lze částečně překonat využitím zpožděné časové řady jako další dimenze. Dalším úskalím je fakt, že není dobře standardizované hraniční skóre oddělující anomální hodnoty. Před případným využitím této metody v oblasti detekce anomálií je tak nezbytné navrhnout vhodnou standardizaci.

5.5 Moderní práce s časovými údaji: lubridate a tsibble

Flexibilní a korektní zpracování různých formátů času a data včetně převodů mezi textovými a datovými formáty, extrakce a nastavování komponent, aritmetických operací a zaokrouhlování, ale i práci s časovými zónami, přestupnými roky a letním časem, to vše umožňuje moderní knihovna **lubridate** (Grolemund a Wickham, 2011).

Moderním přístupem k manipulaci s daty v prostředí R je sbírka balíků **tidyverse** (Wickham et al., 2019, nebo Wickham a Grolemund, 2016). Umožňuje import dat, manipulaci, čištění a úpravu, vizualizaci s vhodným a cíleným využíváním doménově specifických balíčků v této sbírce. Sbírku **tidyverse** rozšiřuje a efektivní práci s časovými daty umožňuje balíček **tsibble** (Wang et al., 2020). Rozšiřuje vlastnosti obecných **tibble** datových struktur, přičemž užívá časové údaje jako klíčové datové sloupce a využívá pořadí daného plynutím času.

6 Základní možnosti výběru významných anomálií a redukce dimenzionality

Výše popsané automatizované přístupy detekce anomálií společně s nasazením na velké množství dat vznikajících a archivovaných v běžné praxi větších firem a organizací představují zdroj detekce velkého množství výkyvů a anomálií, na které může být uživatel systému upozorněn. To samozřejmě přináší velkou pracovní zátěž při kontrole výstupů, či naopak riziko přehlédnutí některých významných alertů mezi spoustou upozornění podstatně méně podstatných. Pro zpřehlednění, vyšší uživatelskou přívětivost a lepší využitelnost systému lze uvažovat následující tři základní přístupy.

6.1 Ekonomické souvislosti

Východiskem pro seskupování na základě ekonomických ukazatelů je řada ekonomických vztahů a souvztažností od jednoduchých po komplexnější a zároveň od deterministických až po stochastické. Velkým úskalím tohoto přístupu je fakt, že jeho důslednější aplikace zpravidla narazí na specifičnost

sestav ukazatelů v různých oblastech a oborech a bude tudíž vyžadovat implementaci diferencovanou pro různá nasazení alertovacího systému. Nicméně typicky nejzákladnější ekonomické souvislosti typu výkyvů ve výši tržeb a s tím zároveň v zásadě zbytečně reportovaným výkyvům v odvedené DPH lze snadno předcházet.

6.2 Metody hledající příbuzné anomálie na základě statistických souvislostí

Statistické přístupy umožňují průběžně vyhledávat úzce související řady, které nebyly zjištěny pomocí známých ekonomických souvislostí. Výhodou přístupu je to, že může být do značné míry automatizován. Časové řady mohou být zjišťovány pomocí korelačních a kointegračních přístupů a případně i pokročilejšími ekonometrickými přístupy, jakými je např. Grangerova analýza (Granger, 1969), nebo Convergent cross mapping (Čenys et al., 1991).

6.3 Hierarchické modely časových řad a sbalování souvisejících anomálií

Ve většině větších společností existuje struktura středisek a poboček a např. z hlediska vyráběného, či prodávaného sortimentu lze velmi často identifikovat zpravidla vícestupňovou hierarchickou strukturu výrobků. Jednoduchou hierarchii nebo i vícenásobné seskupení časových řad lze modelovat přístupy uvedenými v práci Wickramasuriya et al. (2019). Velkou výhodou těchto přístupů je to, že modely těchto časových řad jsou spolu v souladu, to znamená, že např. i součet predikovaných hodnot za nižší jednotky plně odpovídá predikované hodnotě na vyšší úrovni. Při využití těchto přístupů může uživatel procházet hierarchií od vyšší úrovně po nižší a detekované anomálie postupně rozbaluje podle potřeby, přičemž na vyšší úrovni mohou být indikovány např. pouze výrazné anomálie, případně i s indikací, jak velké množství anomálií je na vnořených úrovních.

7 Závěr

Tento článek přináší základní shrnutí metod využitelných pro automatickou detekci anomálií a vytváření varování v rámci informačních systémů jako podkladu pro rozhodování vedoucích pracovníků. V návaznosti na rychlý rozvoj v oblasti dat a informací v podnikové sféře se zejména v posledním desetiletí překotně rozvíjejí i přístupy k predikci v časových řadách a hledání neočekávaných výkyvů v jejich průběhu. Kromě tradičních matematicko-statistických přístupů vznikají moderní přístupy šité na míru internetovému prostředí a prostředí sociálních sítí. Základní přehled uvedený v tomto článku může sloužit prvotní orientaci v této oblasti a lze jej využít v počátcích přípravy automatizovaného systému. Vždy je ale potřeba mít na paměti doménu využití, příslušné ekonomické závislosti, strukturu a hierarchii časových dat i jejich

periodicitu, sezónnost a cykličnost. Tomu je pak zapotřebí přizpůsobit výběr a propojení použitých metod.

8 Literatura

Box, G., Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Brodersen, K. H. et al. (2015). *Inferring causal impact using Bayesian structural time-series models.* Annals of Applied Statistics, 9(1), 247-274.

Brown, R. G. (1959). *Statistical forecasting for inventory control.* McGraw/Hill.

Chatfield, C. (1978). *The Holt-winters forecasting procedure.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 27(3), 264-279.

Cleveland, R. B. et al. (1990). *STL: A seasonal-trend decomposition.* Journal of Official Statistics, 6(1), 3-73.

Cleveland, W. S. (1979). *Robust locally weighted regression and smoothing scatterplots.* Journal of the American Statistical Association, 74(368), 829-836.

Čenys, A., Lasiene, G., Pyragas, K. (1991). *Estimation of interrelation between chaotic observables.* Physica D: Nonlinear Phenomena, 52(2-3), 332-337.

Dancho, M., Vaughan, D. (2020). *anomalize: Tidy Anomaly Detection. R package version 0.2.2.* https://CRAN.R-project.org/package=anomalize.

De Livera, A. M., Hyndman, R. J., Snyder, R. D. (2011). *Forecasting time series with complex seasonal patterns using exponential smoothing.* Journal of the American Statistical Association, 106(496), 1513-1527.

Ding, Z., Fei, M. (2013). *An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window.* IFAC Proceedings Volumes, 46(20), 12-17.

Elrhim, M. A., Elsayed, A. (2020). *The Effect of COVID-19 Spread on the e-commerce market: The case of the 5 largest e-commerce companies in the world.* https://ssrn.com/abstract=3621166.

Evjemo, L. D. et al. (2020). *Trends in Smart Manufacturing: Role of Humans and Industrial Robots in Smart Factories.* Current Robotics Reports, 1(2), 35-41.

Gers, F. A., Schmidhuber, J., Cummins, F. (1999). *Learning to forget: Continual prediction with LSTM.* In 9th International Conference on Artificial Neural Networks: ICANN '99 (pp. 850-855).

Granger, C. W. J. (1969). *Investigating causal relations by econometric models and cross-spectral methods.* Econometrica: Journal of the Econometric Society, 424-438.

Grolemund, G., Wickham, H. (2011). *Dates and Times Made Easy with lubridate.* Journal of Statistical Software, 40(3), 1-25. https://www.jstatsoft.org/v40/i03.

Hochenbaum, J., Vallis, O. S., Kejariwal, A. (2017). *Automatic anomaly detection in the cloud via statistical learning.* https://arxiv.org/abs/1704.07706.

Hyndman, R. J., Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts. https://otexts.com/fpp3.

Hyndman, R. J. et al. (2018). *forecast: Forecasting functions for time series and linear models.* https://pkg.robjhyndman.com/forecast.

Hyndman, R. J., Khandakar, Y. (2008). *Automatic time series forecasting: the forecast package for R.* Journal of Statistical Software, 27(3), 1-22.

Holt, C. C. (1957). *Forecasting Trends and Seasonals by Exponentially Weighted Averages*, vol. 52. Carnegie Institute of Technology, Pittsburgh, Pa, USA.

Kejariwal, A. (2015). *Introducing practical and robust anomaly detection in a time series. Twitter Engineering Blog.* https://blog.twitter.com/engineering/en_us/a/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series.html.

Kulanuwat, L. et al. (2021). *Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series.* Water, 13(13), 1862.

Liu, F. T., Ting, K. M., Zhou, Z. H. (2012). *Isolation-based anomaly detection.* ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), 1-39.

Luminol (2018) *Luminol: LinkedIn's Anomaly Detection and Correlation Library.* https://github.com/linkedin/luminol.

R Core Team (2021). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Ren, H. et al.(2019). *Time-series anomaly detection service at Microsoft.* In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3009-3017).

Tabesh, P., Mousavidin, E., Hasani, S. (2019). *Implementing big data strategies: A managerial perspective.* Business Horizons, 62(3), 347-358.

Taylor, S. J., Letham, B. (2018). *Forecasting at scale.* The American Statistician, 72(1), 37-45.

Toledano, M. et al. (2018). *Real-time anomaly detection system for time series at scale.* In KDD 2017 Workshop on Anomaly Detection in Finance (pp. 56-65).

Wang, E., Cook, D., Hyndman, R. J. (2020). *A new tidy data structure to support exploration and modeling of temporal data.* Journal of Computational and Graphical Statistics, 29(3), 466-478.

Wickham, H. e al. (2019). *Welcome to the tidyverse.* Journal of Open Source Software, 4(43), 1686.

Wickham, H., Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data.* O'Reilly Media, Inc.

Wickramasuriya, S. L., Athanasopoulos, G., Hyndman, R. J. (2019). *Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization.* Journal of the American Statistical Association, 114(526), 804-819.

Winters, P. R. (1960). *Forecasting sales by exponentially weighted moving averages.* Management Science, 6(3), 324-342.

9 Poděkování

Tvorba tohoto článku byla podpořena Technologickou agenturou České republiky v programu TREND v rámci projektu FW01010606: Adaptivní alertovací business intelligence systém.

FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 2 pp 12-27

Aplikácia analýzy nákupného koša v oblasti CRM pomocou softvéru R
Application of market basket analysis in the field of CRM using the R software

Tadeáš Chujac, Martina Jantová, Iveta Stankovičová
Univerzita Komenského v Bratislave, Fakulta managementu, Odbojárov 10, 820 05 Bratislava, Slovenská republika
Comenius University in Bratislava, Faculty of Management, Odbojárov 10, 820 05 Bratislava, Slovak Republic
chujac1@uniba.sk, jantova9@uniba.sk, iveta.stankovicova@fm.uniba.sk

Abstrakt: *Pri rozhodovaní marketingových manažérov je veľmi dôležité vedieť odhadnúť nákupné správanie svojich zákazníkov. Aby informácie o zákazníkoch boli relevantné, je potrebné sa oprieť o dáta a vyťažiť z nich správne informácie, ktoré vedia firme pomôcť v konkurenčnom boji. Jednou z metód data miningu, ktorá dokáže správanie zákazníkov odhadnúť pomocou analýzy transakčných dát je analýza nákupného koša. Cieľom článku je popísanie využitia metódy analýzy nákupného koša v oblasti predaja elektro tovaru. Výsledky tejto analýzy vedia byť užitočné v oblasti CRM a pri rozhodovaní o marketingovej stratégii.*

Abstract: *When making marketing managers' decisions, it is very important to be able to estimate the purchasing behaviour of their customers. To be customer information relevant, it is necessary to rely on data and extract the right information from it, which can help the company in its competition. One of the data mining methods that can estimate customer behaviour using transaction data is market basket analysis. The aim of the article is to describe the use of the market basket analysis method in the field of electronics sales. The results of this analysis can be useful in the field of CRM and in deciding on marketing strategy.*

Kľúčové slová: *analýza nákupného koša, data mining, asociačné pravidlá, apriori algoritmus, transakčné dáta*

Key words: *market basket analysis, data mining, associations rules, apriori algorithm, transaction data*

1 Úvod

V dnešnom modernom svete firmy zhromažďujú veľké množstvá dát a informácií o svojich zákazníkoch, ktoré sú využiteľné v mnohých oblastiach vrátane marketingu či riadenia vzťahu so zákazníkmi (CRM). Pre získanie trhového podielu či zvýšenie tržieb firmy je nevyhnutné tieto dáta využiť čo najefektívnejšie. Na extrakciu skrytých a užitočných informácií z veľkých databáz sa používajú rôzne techniky data miningu. Následná analýza takto získaných informácií je nevyhnutná pre udržanie postavenia spoločnosti v konkurenčnom prostredí.

Analýza nákupného koša je jednou z najpopulárnejších metód hĺbkovej analýzy údajov (data miningu), ktorá sa zameriava na odhalenie „nákupných

vzorcov“ správania zákazníkov. Cieľom je identifikovať z transakčných dát zaujímavé asociácie medzi viacerými produktami, ktoré zákazníci často spolu nakupujú, čiže kladú spolu do nákupného košíka.

Cieľom príspevku je analyzovať transakčné dáta predajcu elektroniky a pomocou analýzy nákupného koša odhaliť, aké produkty si zákazníci kupujú spolu najčastejšie. Tieto informácie vedia byť nápomocné pre marketingových manažérov pri zvyšovaní efektivity marketingových stratégií. Článok je zároveň návodom, ako očistiť, spracovať a analyzovať transakčné dáta v bezplatnom štatistickom programe R.

V prvej časti je zhrnutý prehľad literatúry vo vzťahu k technikám analýzy nákupného koša. Následne je spísaná metodika spracovania príspevku, na ktorú nadväzujú výsledky analýzy nákupného koša na základe reálnych údajov predaja elektro tovaru. Na spracovanie údajov a výpočty bol použitý voľne dostupný štatistický softvér R.

2 Analýza nákupného koša

Cieľom analýzy nákupného koša je nájdenie asociačných pravidiel, teda častých vzorov pri nákupe. Ide o produkty, ktoré zákazníci často kupujú spoločne (Zendulka, 2006). Napríklad pri nákupe nového notebooku si zákazník často kúpi aj novú tašku, či myš, prípadne k mobilu si pravdepodobne dokúpi obal či ochranné sklo. Pri rozhodovaní v marketingu je analýza nákupného koša silným nástrojom na podporu implementácie stratégií krížového predaja. Napríklad, ak nákupný profil konkrétneho zákazníka zapadá do identifikovaného nákupného koša, navrhne sa ďalšia položka (Cavique, 2007). Na základe asociačných pravidiel vieme potom odhadnúť, že ak si zákazník kúpi varnú dosku a chladničku, pravdepodobne bude zariaďovať byt a vieme mu odporučiť aj umývačku riadu a tým zvýšiť tržby obchodu.

Pri využití tejto techniky sa vychádza z takzvaných transakčných dát, ktorých najvýznamnejšími atribútmi sú identifikácia transakcie a identifikácia položiek v transakcii. Položky sú zvyčajne napojené na tabuľku charakteristík produktov. Nákup sa viaže na určitého zákazníka (ID zákazníka), ktorý môže mať aj niekoľko zákaziek a preto sa ID zákazníka v transakčných dátach opakuje na viacerých riadkoch dátovej tabuľky (Stankovičová, 2009).

Táto analýza si našla uplatnenie v rôznych oblastiach manažmentu, najmä marketingu a riadenia vzťahov so zákazníkmi (CRM). Keďže CRM sa zameriava na zvyšovanie spokojnosti zákazníkov ale aj zvyšovanie rozsahu podnikania (Osmanbegovic, 2015), práve analýza nákupného koša vie odporučiť zákazníkovi produkty, o ktoré by mohol mať reálny záujem a tým zvýšiť aj tržby obchodu.

2.1 Asociačné pravidlá

Zďaleka najbežnejšou praxou v analýze nákupného koša je identifikácia asociačných pravidiel (Brijs, 2004). Asociačné pravidlá reprezentujú schémy, resp. vzory v dátach bez toho, aby bola definovaná závislá premenná (Stankovičová, 2009). Tieto pravidlá môžu maloobchodníkom pomôcť pochopiť nákupnú povahu zákazníkov (Manpreet, 2016). Každá dvojica produktov A a B (pričom A a B môže označovať aj viacero položiek) sa hodnotí na základe troch parametrov (Kamakura, 2012):

- ***Support (podpora)*** – spoločná pravdepodobnosť nájdenia páru tovarov A a B vo všetkých košoch, čiže $\mathrm{P}(\mathrm{A} \cap \mathrm{B})$. Ak hodnota vyjde nízka znamená to, že pár A a B nie je relevantný, pretože sa táto kombinácia produktov nenakupuje dostatočne často.
- ***Confidence (spoľahlivosť)*** – vyjadruje pravdepodobnosť, že nákup produktu A povedie k nákupu aj produktu B. Ide o podmienenú pravdepodobnosť, čiže miera sa vypočíta podľa vzťahu: $\mathrm{P}(\mathrm{B}|\mathrm{A}) = \mathrm{P}(\mathrm{A} \cap \mathrm{B})/\mathrm{P}(\mathrm{A})$
- ***Lift (miera závislosti)*** - pomer medzi spoločnou pravdepodobnosťou produktov A a B v košíku a pravdepodobnosťou spoločného výskytu pri ich nezávislosti a vzorec je nasledovný: $\mathrm{P}(\mathrm{A} \cap \mathrm{B})/(\mathrm{P}(\mathrm{A}) \times \mathrm{P}(\mathrm{B}))$. Tento parameter udáva mieru závislosti medzi dvoma produktami A a B. Hodnota tohto parametru hovorí, či je závislosť pozitívna (hodnota lift je väčšia ako 1), negatívna (hodnota lift je menšia ako 1) alebo sa závislosť nevyskytuje (hodnota lift je rovná 0).

Firmám sa oplatí pri nákupe odporučiť produkt B tým zákazníkom, ktorí si práve pridali produkt A do svojho nákupného košíka, ak majú produkty A a B vysokú tendenciu vyskytovať sa v košíku spoločne (t.j. hodnota support je vysoká) a taktiež je vysoká pravdepodobnosť, že nákup produktu A povedie k nákupu produktu B (t.j. vysoká hodnota confidence) (Kamakura, 2012).

V neposlednom rade dôležitým parametrom, ktorý napovie, či sa firme oplatí odporučiť zákazníkovi pri nákupe produktu A aj produkt B, je hodnota lift. Ak vyjde táto hodnota vyššia ako 1, napovie to, že miera závislosti medzi produktami A a B je u daného zákazníka vysoká (napr. hodnota lift = 2 znamená, že u zákazníka, ktorý si kúpil produkt A, je až dvakrát vyššia pravdepodobnosť kúpy aj produktu B ako u iných zákazníkov). Závislosť medzi produktmi však môže byť i negatívna (vtedy hodnota lift je nižšia ako 1), prípadne závislosť nemusí existovať, ak hodnota lift je rovná nule (Stankovičová, 2009).

2.2 Apriori algoritmus

Algoritmus apriori slúži k vyhľadávaniu základných asociačných pravidiel, pričom berie do úvahy všetky transakcie v databáze pri definovaní nákupného

koša. Ako vstup používa tento algoritmus tabuľku s transakciami, pričom každá obsahuje svoju identifikáciu a nakupované položky nasledovne (transakcia, položka) (Cavique, 2007).

Pri tejto metóde vstupujú tri parametre:

- maximálny počet položiek v nákupnom košíku (max_n),
- minimálna podpora (min_sup),
- minimálna spoľahlivosť (min_conf).

Apriori algoritmus postupne generuje množiny, v ktorých sa nachádzajú frekventované položky, ktoré sa často kupujú spolu, pričom algoritmus postupuje od množín s najmenším počtom prvkov k tým najväčším (veľkosti množín sú ohraničené hodnotou max_n). Algoritmus používa apriori vlastnosť frekventovaných množín: Všetky neprázdne podmnožiny frekventovanej množiny položiek sú (musia byť) tiež frekventované. Súbor frekventovaných množín, ktoré majú n prvkov sa označuje L_n. Následne algoritmus odfiltruje množiny, ktoré nesĺňajú minimálnu hodnotu podpory (min_sup). Pokiaľ je to možné, z L_n množín sa generujú L_{n+1} množiny s n+1 prvkami (až kým n nedosiahne max_n). Výstupom algoritmu sú všetky silné n-množiny, z ktorých je možné generovať asociačné pravidlá, ktoré spĺňajú minimálnu hodnotu podpory (min_sup) a spoľahlivosti (min_conf), pričom pre každé z pravidiel vypočíta aj hodnoty support a confidence.

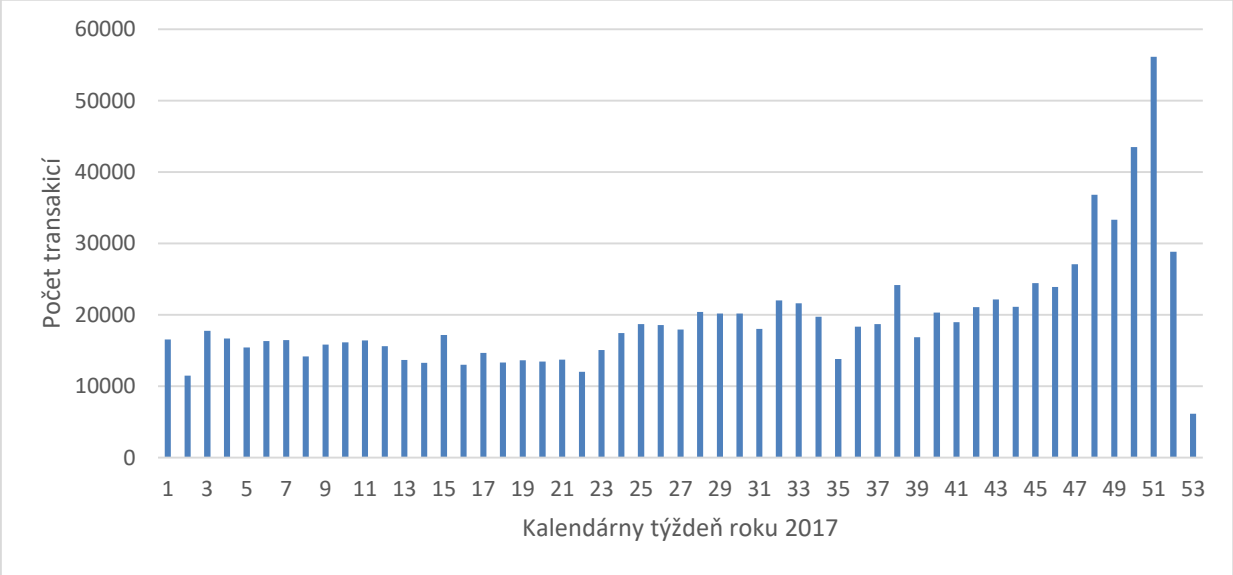
Za výhody Apriori algoritmu možno považovať množstvo vzorov, ktoré sú výstupom a zároveň sú tieto vzory ľahko pochopiteľné a interpretovateľné. Za nevýhodu možno považovať dlhší výpočtový čas.

3 Metodológia a popis dátového súboru

Analyzovaný dátový súbor, na ktorom budeme demonštrovať možnosti využitia analýzy nákupného koša pozostával z reálnych transakcií, resp. nákupov uskutočnených u jedného z najväčších predajcov elektroniky v Slovenskej republike za rok 2017. Dáta aj napriek staršiemu dátumu neznižujú relevantnosť výstupov, keďže nákupné správanie spotrebiteľov, týkajúce sa asociácií produktov, ktoré sú kupované spoločne, sa za ostatné 4 roky výrazne nezmenilo.

Pre analýzu sme spracovali dátový súbor s viac ako 1 milión transakciami, v ktorých viac ako 400 tisíc rôznych zákazníkov s vernostnými kartami zakúpilo takmer 1,7 milióna položiek. Na základe týchto transakčných údajov bolo prostredníctvom tohto výskumu možné simulovať reálne prostredie z praxe, keďže dostupné dáta netvorili vzorku transakcií, ale celú populáciu transakčných údajov za rok 2017 (tab. 1).

Spomínaný obchod s elektro tovarom predáva takmer 16 tisíc unikátnych tovarov. Predajca má vo svojom portfóliu produktov nielen elektroniku, ale aj drogériu, potreby do domácnosti, či športové potreby. Napriek tomu, priemerne sa v jednej transakcii nachádza iba 1,64 položky a dokonca viac ako 62 % transakcií tvoria iba jednopoložkové nákupy (tab. 1). Takúto štruktúru počtu tovarov v jednotlivých transakciách sme však čakali vzhľadom na to, že oblasť elektroniky sa nevyznačuje vysokým počtom položiek v jednej transakcii, keďže väčšina tovarov patrí medzi drahšie tovary, elektroniku si zákazníci kupujú s nižšou frekvenciou ako iné lacnejšie produkty.



Graf 1 Počet transakcií uskutočnených v jednotlivých týždňoch roka 2017 (*Zdroj*: *vlastné spracovanie*)

Tab. 1 Počet tovarov v jednotlivých transakciách *(Zdroj: vlastné spracovanie)*

Počet produktov v transakcii	**Počet transakcií**	**f_i**
1	644605	62,40%
2	236912	22,93%
3	87075	8,43%
4	36452	3,53%
5	14840	1,44%
6	6525	0,63%
7	3092	0,30%
8	1616	0,16%
9	773	0,07%
10 a viac	1201	0,12%
SPOLU	**1033091**	**100,00%**

Keďže predmetom výskumu je analýza nákupného koša, ktorá má za cieľ skúmať vzťahy medzi položkami, ktoré boli do „košíka“ vkladané v rámci jednej

transakcie, pre vyššiu výpovednú hodnotu uskutočnenej analýzy bolo potrebné transakčné dáta upraviť. Po bližšom preskúmaní dátového súboru sme zistili, že v analyzovaných transakčných dátach sa pomerne často vyskytuje skutočnosť, že daný zákazník uskutoční v ten istý deň viacero nákupov. Takmer 100 tisíc zákazníkov urobilo v deň nákupu aj nejaký iný nákup pod odlišným ID nákupu, preto tieto nákupy neregistrujeme v databáze ako jeden (tab. 2).

Tab. 2 Počet transakcií na jedného zákazníka počas jedného dňa *(Zdroj: vlastné spracovanie)*

Počet transakcií zákazníka v jeden deň	n_i
1	819600
2	79421
3	10853
4 a viac	4311

Nákup elektroniky býva často ale spojený aj s dodatočným nákupom rôzneho príslušenstva. Pri mobile si často zákazník uvedomí až dodatočne, že mu napríklad do balenia nepribalili nabíjačku, slúchadlá alebo mobil používa iné zakončenie, ktoré nie je kompatibilné s vybavením, ktoré zákazník vlastní z predchádzajúceho modelu. S tým sú spojené dodatočné nákupy, ktoré obsahujú položky nevyhnutné pre to, aby zakúpený tovar mohol zákazník využívať naplno. Preto sme sa pozreli aj na to, či daný zákazník spravil nejaký nákup aj v rozsahu ďalších 7[1] dní od prvotného nákupu. Nakoniec sa ukázalo, že dokonca až pri viac ako 30 % z celkového počtu transakcií existuje takáto príbuzná transakcia, ktorú urobil daný zákazník do 7 dní. Tieto informácie sú práve pre analýzu nákupného koša veľmi cenné, keďže jedným z využití tejto metódy v praxi je ponúknuť zákazníkovi ďalšie vhodné produkty, ktoré by si mohol k už existujúcemu produktu v nákupnom koši prikúpiť a nedovoliť tak, aby si príslušenstvo zakúpil u konkurenčných predajcov elektroniky. Tieto transakcie, ktoré boli uskutočnené zákazníkov v rozmedzí maximálne 7 dní boli spojené do jednej, čím sme redukovali počet unikátnych transakcií a zvýšili sme počet produktov v jednom nákupe.

Výsledný dátový súbor pozostáva zo 6 premenných: ID_zákazníka, ID_nákupu, Dátum_nákupu, ID_tovaru, Kategória_tovaru (tab. 3). Zahŕňa v sebe viac ako 730 tisíc transakcií, v ktorých bolo zakúpených takmer 1,5 milióna tovarov. Pokles v počte transakcií a produktov je spôsobený spájaním viacerých transakcií do

[1] Voľba 7 dní je subjektívneho charakteru, keďže bolo v úmysle zachytiť produkty, ktoré bezprostredne nasledovali po kúpe iných produktov. Nebol preto zvolený dlhší interval, napr. 30 dní, lebo pri takejto dĺžke intervalu už môžu byť vo väčšej miere zachytené aj také nákupy, ktoré spolu nijakým spôsobom nesúvisia. Na druhej strane nebol zvolený ani kratší interval, keďže pri elektronike, ktorá je zložitejšia na používanie môže zákazníkom trvať aj niekoľko dní, kým si uvedomia, že dané príslušenstvo k nej potrebujú. Preto voľba intervalu 7 dní bola kompromisná voľba, ktorá by mala zachytiť práve tie nákupy, ktoré medzi sebou súvisia.

jednej a v rámci jedného nákupu postačuje, aby sa daný tovar v nej nachádzal iba jedenkrát. Štatistická jednotka výsledného datasetu reprezentuje jeden zakúpený tovar.

Tab. 3 Príklad štruktúry výsledného dátového súboru *(Zdroj: vlastné spracovanie)*

ID_zákazníka	ID_nákupu	Dátum_nákupu	ID_tovaru	Kategória_tovaru
2358304	94417219	2017-09-12	2002144	Hry na herné konzoly
2358304	94417219	2017-09-12	2002142	Herné konzoly
2355514	94417064	2017-09-12	1106806	Pamäťové karty

Prostredníctvom zlučovania transakcií s jedným produktom do početnejších skupín sa podarilo znížiť zastúpenie jednopoložkových transakcií o viac ako 9 %. Priemerný počet produktov na jednu transakciu dosiahol hodnotu 2, hoci mediánový počet produktov v jednej transakcií zostal stále na hodnote 1.

Tab. 4 Počet tovarov v jednotlivých transakciách po úprave datasetu *(Zdroj: vlastné spracovanie)*

Počet produktov v transakcii	Počet transakcií	f_i
1	388690	53,07%
2	180251	24,61%
3	82143	11,22%
4	39303	5,37%
5	18815	2,57%
6	9484	1,29%
7	4992	0,68%
8	2807	0,38%
9	1550	0,21%
10 a viac	4348	0,59%

Celú analýzu sme vykonávali v prostredí bezplatného štatistického nástroja R s využitím knižníc pre manipuláciu s dátami *tidyverse* a *data.table* a knižnicou *arules* určenou v prostredí softvéru R pre analýzu nákupného košíka, manipuláciu s transakčnými dátami a pre tvorbu asociačných pravidiel. Výsledné grafy a tabuľky boli spracované pomocou MS Office.

4 Analýza nákupného koša v R

Analýza nákupného koša v prostredí štatistického softvéru R vyžaduje ešte jeden krok v rámci prípravy. Analyzované dáta je nevyhnutné pretransformovať do formátu transakčných dát, v ktorých dôležitými premennými už budú iba *ID_nákupu* a *ID_tovaru*, resp. kategória tovaru, ak sa analýza nákupného koša robí na viacerých úrovniach databázy produktov. Vyššie spomínaný finálny dátový súbor *„transakcie“* po všetkých úvodných úpravách, ktorý mal pre

potreby úprav štruktúru „data.frame“ sme nasledujúcim príkazom modifikovali na transakčné dáta:

>trans<-as(split(transakcie[,ID_tovaru], transakcie[,ID_nákupu]), 'transactions').

Na rozdiel od finálneho dátového súboru, kde jeden riadok označoval jednu zakúpenú položku, tak v prípade transakčných dát jeden riadok označuje jednu faktúru, resp. jeden nákup a následne zakúpené tovary sú zapísané v množinovej zátvorke s oddeľovačom v podobe čiarky (obr. 1). Takže vieme povedať, že v rámci transakcie číslo 12101449, ktorá má v transakčných dátach poradové číslo 1 boli zakúpené 3 rôzne položky s číslami 1006687, 1020027 a 1020028, ktoré označujú postupne vrecká do vysávača, tablety na vodný kameň a čistiace tablety.

```
> inspect(head(trans, 10))
      items                          transactionID
[1]   {1006687,1020027,1020028}      12101449
[2]   {1076055,1106806}              12101455
[3]   {1065481}                      12101467
[4]   {1104819}                      12101468
[5]   {1075944,2000400}              12101473
[6]   {1108586}                      12101483
[7]   {1049584,1052336}              12101484
[8]   {1020029,1106702}              12101492
[9]   {1114860}                      12101497
[10]  {1033932,1107250}              12101501
```

Obr. 1 Štruktúra transakčných údajov v prostredí softvéru R (*Zdroj: vlastné spracovanie*)

```
> summary(trans)
transactions as itemMatrix in sparse format with
 732383 rows (elements/itemsets/transactions) and
 15919 columns (items) and a density of 0.0001256365

most frequent items:
1025665 1102192 1085679 2009031 1011730 (Other)
  19528   13519   11549    9287    8527 1402361

element (itemset/transaction) length distribution:
sizes
     1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16     17     18     19     20     21     22     23     24
388695 180246  82144  39304  18814   9484   4991   2807   1551    976    623    437    278    236    185    159    117     96     85     75     74     60     59     40
    25     26     27     28     29     30     31     32     33     34     35     36     37     38     39     40     41     42     43     44     45     46     47     48
    37     44     36     35     37     30     28     24     24     23     20     16     19     19     14     18     21     12     18     16     13      7     17     12
    49     50     51     52     53     54     55     56     57     58     59     60     61     62     63     64     65     66     67     68     69     70     71     72
     9      8     12     13     13      5     13     10      8      5      7     11     10      7      3      6      1      6      6      5      4      5      5      3
    73     74     75     76     77     78     79     80     81     82     83     84     85     86     87     88     89     90     91     92     93     94     96     97
     3      6      3      5      7      2      3      3      6      5      2      2      5      1      2      1      3      1      2      1      4      2      1      2
    98    102    103    104    105    106    108    110    111    113    114    115    118    119    124    126    127    130    131    133    135    139    140    142
     3      2      1      2      2      2      1      1      2      1      2      3      1      1      1      2      2      1      1      1      1      1      2      1
   143    144    149    150    151    155    158    160    178    183    186    197    203    207    209    217    229    255    396    598
     1      1      2      1      1      1      1      2      1      2      1      1      1      1      1      1      1      1      1      1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       1       1       2       2     598

includes extended item information - examples:
   labels
1 1000582
2 1000584
3 1000594

includes extended transaction information - examples:
  transactionID
1      12101449
2      12101455
3      12101467
```

Obr. 2 Informácie o transakčných dátach (*Zdroj: vlastné spracovanie*)

Transakčné dáta sú v pamäti uložené prostredníctvom riedkej matice, ktorej počet riadkov označuje počet unikátnych transakcií a počet stĺpcov označuje

počet unikátnych tovarov nachádzajúcich sa v analyzovaných transakciách. Pomocou funkcie *summary()* môžeme v R-ku získať informácie o tom, že v našich transakčných dátach sa nachádza presne 732 383 transakcií obsahujúcich 15 919 jedinečných tovarov. Päť najčastejšie zakúpených položiek tvoria tovary ako Sodastream náplne, vrecká do vysávača, či služba lepenia fólie na smartfón. V dátach sa nachádzajú aj „extrémne" faktúry, ktoré obsahovali viac ako 200 tovarov, či dokonca až 598. Tieto extrémy sú spojené so zlučovaním transakcií a nami nastavené pravidlá zachytili aj niekoľkých veľkoodberateľov, ktorí spôsobujú isté skreslenie dát. Nakoniec súčasťou výstupu sú aj popisné štatistiky, ktoré už boli vyššie popísané ako priemer, či medián (obr. 2).

4.1 Dolovanie asociačných pravidiel z jednoduchých transakčných dát v R

Ďalším krokom po príprave transakčných dát pri analýze nákupného koša v R je ich vloženie do funkcie *apriori().* Apriori algoritmus sa používa na generáciu asociačných pravidiel, ktoré by mali odhaliť skryté vzorce v analyzovaných dátach. Táto funkcia má v R-ku aj svoje základné nastavenie, kedy ako parameter do funkcie stačí vložiť iba transakčné dáta. Ako je možné vidieť na obrázku nižšie (obr. 3), základné nastavenia obsahujú prednastavené hodnoty pre nasledovné parametre (Hahsler, 2006):

- *confidence* – minimálna hodnota confidence (preddefinovaná hodnota = 0.8)
- *smax* – maximálna hodnota support (preddefinovaná hodnota = 1)
- *support* – minimálna hodnota support (preddefinovaná hodnota = 0.1)
- *minlen* – minimálny počet položiek v transakcii (preddefinovaná hodnota = 1)
- *maxlen* – maximálny počet položiek v transakcii (preddefinovaná hodnota = 10)

Najproblematickejším zo základných nastavení, ktoré nedovolilo vygenerovanie ani jedného asociačného pravidla je nastavenie úrovne podpory (support) až na úrovni 0,1. Znamená to, že aby vôbec pravidlo mohlo byť vygenerované, položky sa musia spolu nachádzať aspoň na 10 % transakcií, čo by bolo v našom prípade aspoň 73 238 faktúr. Táto hranica podpory je v prípade tak veľkých databáz produktov, aké máme my k dispozícií na úrovni jednotlivých produktov (takmer 16 000 produktov) veľmi ťažko dosiahnuteľná a tým pádom si to vyžaduje buď zníženie tejto hranice podpory alebo zoskupenie viacerých produktov do kategórií (pozri podkapitola 4.2). Obr. 2 poskytuje základný prehľad o tom, na akú približnú úroveň by sme mali úroveň podpory minimálne posunúť. Keďže vieme, že piaty najpredávanejší produkt sa vyskytoval v transakciách 8 527-krát, support je na úrovni 8527/732383 = 0,01164, čiže vrecká do vysávača boli súčasťou košíka v 1,164 % transakcií.

```
> rules <- apriori(trans)
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.8    0.1    1 none FALSE            TRUE       5     0.1      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 73238

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[15919 item(s), 732383 transaction(s)] done [0.85s].
sorting and recoding items ... [0 item(s)] done [0.02s].
creating transaction tree ... done [0.02s].
checking subsets of size 1 done [0.00s].
writing ... [0 rule(s)] done [0.00s].
creating S4 object  ... done [0.03s].
```

Obr. 3 Výsledky apriori algoritmu so základným nastavením (*Zdroj: vlastné spracovanie*)

Za účelom získania vyššieho počtu asociačných pravidiel sú v ďalšom kroku pozmenené 2 parametre, a to hladiny podpory a spoľahlivosti na hodnoty 0,0001, resp. 0,5. S takýmito vstupnými parametrami apriori algoritmus vygeneroval 190 pravidiel a celkové spracovanie dát aj napriek ich veľkému rozsahu trvalo iba niekoľko sekúnd (obr. 4).

```
> rules2 <- apriori(trans, parameter = list(supp = 0.0001, conf = 0.5))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.5    0.1    1 none FALSE            TRUE       5   1e-04      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 73

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[15919 item(s), 732383 transaction(s)] done [1.49s].
sorting and recoding items ... [4155 item(s)] done [0.05s].
creating transaction tree ... done [1.20s].
checking subsets of size 1 2 3 4 done [0.29s].
writing ... [190 rule(s)] done [0.05s].
creating S4 object  ... done [0.25s].
```

Obr. 4 Výsledky apriori algoritmu s pozmenenými parametrami (*Zdroj: vlastné spracovanie*)

Prostredníctvom funkcie *summary()* je možné zobraziť základné informácie a štatistiky o vygenerovaných asociačných pravidlách. Na základe výstupu tejto funkcie je možné zistiť, že 88 pravidiel je zložených z 2 položiek, čiže aj na ľavej a aj pravej strane je iba 1 položka, 94 pravidiel má na ľavej strane 2 položky a na pravej 1 a 8 pravidiel je zložených až zo 4 položiek, čiže 3 položky na ľavej strane

a 1 na pravej[2]. Dostupné sú na obr. 5 aj popisné štatistiky piatich mier kvality asociačných pravidiel (hladina podpory, spoľahlivosti, hladina podpory ľavej strany (coverage), miera závislosti (lift), počet).

```
> summary(rules2)
set of 190 rules

rule length distribution (lhs + rhs):sizes
 2  3  4
88 94  8

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.579   3.000   4.000

summary of quality measures:
    support           confidence        coverage             lift             count
 Min.   :0.0001010   Min.   :0.5000   Min.   :0.0001024   Min.   :  32.04   Min.   :  74.00
 1st Qu.:0.0001150   1st Qu.:0.5758   1st Qu.:0.0001833   1st Qu.: 157.67   1st Qu.:  84.25
 Median :0.0001700   Median :0.7455   Median :0.0002342   Median : 529.98   Median : 124.50
 Mean   :0.0002706   Mean   :0.7500   Mean   :0.0003792   Mean   :1377.64   Mean   : 198.21
 3rd Qu.:0.0002850   3rd Qu.:0.9226   3rd Qu.:0.0004287   3rd Qu.:1741.92   3rd Qu.: 208.75
 Max.   :0.0032551   Max.   :1.0000   Max.   :0.0063423   Max.   :9508.13   Max.   :2384.00

mining info:
  data ntransactions support confidence
 trans        732383   1e-04        0.5
```

Obr. 5 Popisné štatistiky vygenerovaných asociačných pravidiel (*Zdroj: vlastné spracovanie*)

Jednotlivé asociačné pravidlá je možné zobraziť v prostredí R zoradené podľa jednotlivých mier kvality asociačných pravidiel, napríklad miery závislosti (obr. 6). Prvé a druhé pravidlo hovorí o tom, že ak si niekto kúpi HP azúrovú, (resp. žltú) atramentovú kazetu do tlačiarne, tak má 9508-krát vyššiu pravdepodobnosť, že si taktiež zakúpi HP žltú, (resp. azúrovú) atramentovú kazetu do tlačiarne. Tretie pravidlo sa tiež týkalo náplní do tlačiarní, konkrétne, že ak zákazník nakúpi EPSON čiernu, žltú a purpurovú atramentovú náplň do tlačiarne, tak má až 5126-krát vyššiu pravdepodobnosť, že si zakúpi aj azúrovú náplň rovnakej značky. Tieto 3 pravidlá môžeme všetky zaradiť medzi asociačné pravidlá, ktoré sú samozrejmé, keďže často tlačiarne vyžadujú výmeny všetkých náplní naraz a preto ich zákazníci často nakupujú spoločne v jednom nákupe.

```
> inspect(head(rules2, by = "lift", 3))
    lhs                           rhs       support      confidence coverage     lift     count
[1] {2005638}                  => {2005636} 0.0001010400 0.9866667  0.0001024054 9508.130 74
[2] {2005636}                  => {2005638} 0.0001010400 0.9736842  0.0001037708 9508.130 74
[3] {1056942,1056944,1056945} => {1056943} 0.0001024054 0.9868421  0.0001037708 5125.861 75
```

Obr. 6 Tri najlepšie asociačné pravidlá na základe najvyššej miery závislosti (*Zdroj: vlastné spracovanie*)

[2] Softvér R a funkcia apriori() z knižnice arules generuje iba také asociačné pravidlá, ktoré majú na pravej strane len 1 položku.

Asociačné pravidlá založené na jednoduchých transakčných dátach majú niekoľko nevýhod najmä pri databázach zložených z veľkého množstva produktov. Hodnotu hladiny podpory je potrebné znížiť na veľmi nízke hodnoty, aby aspoň niektoré položky z databázy spĺňali dané kritérium. Vzhľadom na to, že v našich transakčných dátach sa nachádza skoro 10 tisíc produktov, ktoré sa za celý rok predali maximálne 10-krát, tak je potrebné pre lepšiu analýzu nákupného koša vytvoriť hierarchické transakčné dáta.

4.2 Dolovanie asociačných pravidiel z hierarchických transakčných dát v R

Hierarchické transakčné dáta obsahujú nielen ID tovaru a nákupu, ale k ID tovaru je priradená aj kategória, do ktorej tento produkt patrí. V analyzovanej databáze predajcu elektroniky sa nachádzajú 4 preddefinované úrovne kategórií, do ktorých patria jednotlivé produkty. Napr. LED TV patrí do kategórie Čierna elektronika -> TV -> LED TV -> TV LED 52“ + 01. Takže postupne je produkt zaraďovaný do menších a konkrétnejších kategórií. Nevýhodou jednoduchých transakčných údajov v predchádzajúcej podkapitole boli málopočetné, resp. málo zastúpené produkty v transakciách. Tento istý problém nastáva ale aj pri niektorých kategóriách produktov, ktoré nie sú častým artiklom v nákupných košíkoch zákazníkov. Vtedy je potrebné zlučovanie kategórií do väčších, aby sme dosiahli rovnomernejšie prerozdelenie tovarov v jednotlivých kategóriách. Vďaka tejto úprave sa podarilo obmedziť počet kategórií, v ktorých bol zakúpený iba jeden tovar. Navyše touto úpravou sa aj znížil počet jednotlivých kategórií z 488 na 226 (tab. 5).

Tab. 5 Popisné štatistiky početností nákupu pôvodných a novovytvorených kategórií *(Zdroj: vlastné spracovanie)*

Popisná štatistika	Pôvodné	Nové
Priemer	3001,60	6481,34
Medián	589	2732
Modus	1	833
Štandardná odchýlka	8130,44	11646,78
Roztyp	66104073,68	135647376,30
Rozpätie	95760	100634
Minimum	1	73
Maximum	95761	100707
Sum	1464782	1464782
Počet	488	226
5 - ty najväčší	34969	34969
5 - ty najmenší	1	194

Po pretransformovaní do štruktúry transakčných dát, s ktorými pracuje knižnica *arules* je prvým krokom skontrolovanie popisných štatistík analyzovaných dát. Najčastejšie zastúpenými kategóriami v transakciách sú

Batérie, Káva, Vrecká do vysávačov, Mobilné telefóny a Služba poistenie. V transakciách sa najčastejšie nachádzajú produkty iba jednej kategórie, čo je spôsobené najmä tým, že celkovú prevahu v transakciách majú jednopoložkové faktúry.

```
> summary(trans_multi)
transactions as itemMatrix in sparse format with
 732383 rows (elements/itemsets/transactions) and
 226 columns (items) and a density of 0.007804913

most frequent items:
            Batérie               Káva Vrecká do vysávačov    Mobilné telefóny    Služba poistenie             (Other)
              77686              56466                48089               46669               34143            1028805

element (itemset/transaction) length distribution:
sizes
     1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16     17     18     19     20     21
425640 180904  71393  29475  12358   5456   2677   1481    735    455    319    219    165    126     96     75     67     61     54     56     43
    22     23     24     25     26     27     28     29     30     31     32     33     34     35     36     37     38     39     40     41     42
    35     37     36     34     41     34     19     28     17     20     17     15     13     20     14     10     10      9      4      7     11
    43     44     45     46     47     48     49     50     51     53     54     55     56     57     58     59     60     61     62     64     65
     7      2      6      3      7      2      6      4     11      2      4      2      1      5      3      4      1      1      2      1      2
    66     67     68     70     72     73     74     75     79     80     82     84     89     94     99    114    132
     2      1      3      1      1      1      2      1      1      1      1      1      1      1      1      1      1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.764   2.000 132.000

includes extended item information - examples:
                labels
1          Alkoholtestery
2   Atramentové tlačiarne
3 Auto zvukové komponenty

includes extended transaction information - examples:
  transactionID
1      12101449
2      12101455
3      12101467
```

Obr. 7 Popisné štatistiky hierarchických transakčných dát (*Zdroj: vlastné spracovanie*)

Asociačné pravidlá budeme hľadať s pozmenenými parametrami oproti základnému nastaveniu. Najskôr je potrebné znížiť hladinu podpory. Napriek tomu, že zlúčením sme získali početnejšie kategórie, hladina podpory na úrovni 0,1 by znamenala, že by sa daná kategória musela vyskytovať v transakčných dátach aspoň 73 238. Z obr. 7 však vieme, že túto podmienku spĺňa iba kategória batérií. Taktiež je potrebné znížiť aj hladinu spoľahlivosti, keďže z popisných štatistík a charakteru transakčných dát vieme, že väčšinou sa produkty nachádzajú samostatne a iba v menšine sa vyskytujú v analyzovaných košíkoch 2 produkty a viac. Posledným nastaveným parametrom, ktorý budeme využívať bude maximálna dĺžka asociačného pravidla. Apriori algoritmus za účelom nájdenia čo najvyšších hodnôt hladín podpory, spoľahlivosti, či miery závislosti bude hľadať zložité pravidlá zložené z niekoľkých produktov na ľavej strane, ktoré by mali podmieňovať kúpu produktu na pravej strane asociačného pravidla. Avšak, často v praxi marketingoví manažéri nehľadajú zložité vzory, ale jednoduché pravidlá, ktoré sú logické a budú mať za následok zvýšenie predaja. Preto sme maximálnu dĺžku pre účely výskumu znížili na maximálne 3 položky, čiže vo výsledku najdlhšie vydolované asociačné pravidlo bude mať maximálne 2 položky na ľavej strane a jednu na pravej. Apriori algoritmus s takto nastavenými pravidlami vygeneroval 126 pravidiel. Jednotlivé asociačné pravidlá je potom potrebné manuálne prezrieť a overiť, či sa v nich nenachádzajú aj

nejaké nelogické asociácie, ktoré mohli vzniknúť iba náhodným výskytom daných položiek spolu v jednom košíku (obr. 8).

```
> rules_multi <- apriori(trans_multi, parameter = list(supp = 0.001, conf = 0.2, maxlen = 3))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.2    0.1    1 none FALSE            TRUE       5   0.001      1      3  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 732

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[226 item(s), 732383 transaction(s)] done [0.22s].
sorting and recoding items ... [183 item(s)] done [0.02s].
creating transaction tree ... done [0.60s].
checking subsets of size 1 2 3 done [0.02s].
writing ... [126 rule(s)] done [0.00s].
creating S4 object  ... done [0.11s].
```

Obr. 8 Výsledky apriori algoritmu s hierarchickými transakčnými dátami (*Zdroj: vlastné spracovanie*)

Ako môžeme vidieť na obr. 9, tak všetkých 10 najlepších asociačných pravidiel z pohľadu miery závislosti sa týka kuchynských spotrebičov a ich rôznych kombinácií. Všetky v podstate hovoria o tom, že ak si zákazník nakúpi nejaký kuchynský spotrebič, resp. ich kombináciu, tak je vysoko pravdepodobné, že si v tom momente zariaďuje novú kuchyňu a je potrebné mu odporučiť v nákupe ďalšie kuchynské spotrebiče. Napríklad prvé asociačné pravidlo hovorí o tom, že zákazník, ktorý si nakúpi tovar z kategórie odsávače a rúry, tak má až takmer 104-krát vyššiu pravdepodobnosť, že si kúpi aj varnú dosku oproti náhodne vybranému zákazníkovi. Nákupný košík s 3 položkami z týchto kategórií sa nachádzal v transakčných dátach až 1253-krát.

```
> inspect(head(rules_multi, by = "lift", 10))
     lhs                                  rhs                 support     confidence coverage    lift      count
[1]  {Odsávače,Rúry}                   => {Varné dosky}       0.001710853 0.8471941  0.002019435 103.80969 1253
[2]  {Rúry,Umývačky riadu}             => {Varné dosky}       0.001721777 0.8373174  0.002056301 102.59947 1261
[3]  {Chladničky,Rúry}                 => {Varné dosky}       0.001516966 0.8097668  0.001873337  99.22359 1111
[4]  {Mikrovlnné rúry,Rúry}            => {Varné dosky}       0.001581140 0.7882914  0.002005781  96.59213 1158
[5]  {Mikrovlnné rúry,Varné dosky}     => {Rúry}              0.001581140 0.8674157  0.001822817  90.49580 1158
[6]  {Umývačky riadu,Varné dosky}      => {Rúry}              0.001721777 0.8526031  0.002019435  88.95043 1261
[7]  {Chladničky,Varné dosky}          => {Rúry}              0.001516966 0.8157122  0.001859683  85.10167 1111
[8]  {Odsávače,Varné dosky}            => {Rúry}              0.001710853 0.8141650  0.002101360  84.94026 1253
[9]  {Mikrovlnné rúry,Umývačky riadu} => {Rúry}              0.001021324 0.7663934  0.001332636  79.95634  748
[10] {Mikrovlnné rúry,Rúry}            => {Umývačky riadu}   0.001021324 0.5091899  0.002005781  59.22218  748
```

Obr. 9 TOP 10 asociačných pravidiel na základe lift-u (*Zdroj: vlastné spracovanie*)

V ďalších 126 vygenerovaných asociačných pravidlách sa často vyskytovali ešte rôzne kombinácie kuchynských spotrebičov, ale napríklad aj zaujímavé poznatky o tom, že ak si zákazník zakúpi mobil, tak je výhodné ponúknuť mu k nemu puzdro a tvrdené sklo, k vysávaču vrecká do vysávačov, ku práčke odvoz,

k plnoautomatickému espresso kávovaru kávu, či k herným konzolám hry a príslušenstvo k nim v podobe joystickov a podobne. Tieto asociačné pravidlá sa javia byť na prvý pohľad očakávané, ale na druhej strane prinášajú do rozhodovania marketingových manažérov podklady na základe ktorých môžu robiť kvalitné rozhodnutia založené na dátach, v podobe umiestnenia týchto produktov v kamennej predajni v tesnej blízkosti vedľa seba, ponúknutia ďalších tovarov alebo služieb pri vložení daného tovaru do košíka, či správne nastavených akcií a zliav (obr. 10).

[21]	{Herné konzoly}	=>	{Hry na herné konzoly}	0.001269827	0.3445721	0.003685230	39.461880	930
[22]	{Herné konzoly}	=>	{Príslušenstvo herné konzoly}	0.001227500	0.3330863	0.003685230	38.715563	899
[23]	{Púzdra na mobil,Služba inštalácie}	=>	{Tvrdené sklá}	0.005538086	0.7719833	0.007173842	38.561411	4056
[24]	{Mobilné telefóny,Služba inštalácie}	=>	{Tvrdené sklá}	0.008993928	0.7050198	0.012756986	35.216513	6587
[44]	{Púzdra na mobil,Služba poistenie}	=>	{Tvrdené sklá}	0.001054093	0.4225506	0.002494596	21.106868	772
[52]	{Práčky,Služba poistenie}	=>	{Služba odvoz}	0.002456365	0.3279256	0.007490616	17.658051	1799
[53]	{Mobilné telefóny,Služba poistenie}	=>	{Tvrdené sklá}	0.002093167	0.3517669	0.005950439	17.571141	1533
[54]	{Kamery}	=>	{Pamäťové karty}	0.001237058	0.4961665	0.002493231	16.847508	906
[55]	{Mobilné telefóny,Tvrdené sklá}	=>	{Púzdra na mobil}	0.005595433	0.4918387	0.011376561	16.768192	4098
[82]	{Videoprehrávače}	=>	{Káble - HDMI}	0.001112806	0.2196174	0.005067021	11.566519	815
[112]	{Vysávače}	=>	{Vrecká do vysávačov}	0.012016936	0.3916778	0.030680668	5.965151	8801
[114]	{Espresso - plnoautomatické}	=>	{Káva}	0.002648887	0.4259991	0.006218058	5.525352	1940

Obr. 10 Ďalšie asociačné pravidlá (*Zdroj: vlastné spracovanie*)

5 Záver

Analýza nákupného koša je užitočná metóda, ako odhaliť vzory v nákupnom správaní spotrebiteľov a tým predpovedať, o aký typ produktov môžu mať pri nákupe záujem. Takéto informácie sú využiteľné v mnohých oblastiach manažmentu a často sa využívajú v oblasti riadenia vzťahov so zákazníkmi. Po odhalení pravidiel, ktoré produkty si kupujú spotrebitelia spoločne, vedia marketingový manažéri v procese nákupu odporučiť v eshope produkty, u ktorých je vysoká pravdepodobnosť, že si ich kúpia tiež. Taktiež v kamenných predajniach sa oplatí usporiadať regály a tovar v nich tak, aby skupiny produktov, pri ktorých je vysoká pravdepodobnosť že budú kupované spolu, boli v regáloch vedľa seba.

V príspevku bolo na reálnych transakčných dátach z oblasti predaja elektro tovaru popísaný a vysvetlený proces čistenia a prípravy dát a následnej aplikácie analýzy nákupného koša vo voľne dostupnom a bezplatnom štatistickom programe R. Takýto postup môže slúžiť ako návod na použitie tejto metódy v praxi.

Pri využití apriori algoritmu v štatistickom softvéri R sme vygenerovali niekoľko skupín asociačných pravidiel s rôznymi nastaveniami parametrov tohto algoritmu. Významnejšie pre prax sa ukázalo byť použitie hierarchických transakčných dát, ktoré vedia zabezpečiť rovnomernejšie rozdelenie položiek v transakčných dátach. Nakoniec sme sformulovali aj niekoľko prakticky využiteľných asociačných pravidiel, ktoré môžu predajcom zvýšiť tržby, napr. ak by k mobilným telefónom ponúkali priamo aj kúpu puzdra na mobil, tvrdeného

skla a aj službu jeho inštalácie. Asociačné pravidlá odhalili vysokú pravdepodobnosť toho, že keď si zákazník zariaďuje novú kuchyňu, tak často sa vyberie cestou nákupu všetkých potrebných spotrebičov u jedného predajcu. Analýza nákupného košíka sa ukázala byť vhodnou a jednoduchou metódou, ktorej výsledky sú ľahko interpretovateľné a rýchlo implementovateľné do praxe.

6 Literatúra

Aguinis, H., Forcum, L. E., Joo, H. (2013). *Using market basket analysis in management research*. Journal of Management, 39(7), 1799-1824.

Brijs, T. et al. (2004). *Building an association rules framework to improve product assortment decisions*. Data Mining and Knowledge Discovery, 8(1), 7-23.

Cavique, L. (2007). *A scalable algorithm for the market basket analysis*. Journal of Retailing and Consumer Services, 14(6), 400-407.

Chandwani, M. H. (2018). *Market basket analysis using association rule*. International Journal of Advance Research, Ideas and Innovations in Technology.

Hahsler, M., Grün, B., Hornik, K. U. (2006). *Mining association rules and frequent itemsets.* R package version 0.4, 119 s.

Kamakura, W. A. (2012). *Sequential market basket analysis*. Marketing Letters, 23(3), 505-516.

Kaur, M., Kang, S. (2016). *Market basket analysis: Identify the changing trends of market data using association rule mining*. Procedia Computer Science, 85, 78-85.

Osmanbegovic, E., Suljic, M. (2015). *Selection of CRM applications*. International Scientific Conference Economy of Integration, 307-318.

Qisman, M., Rosadi, R., Abdullah, A. S. (2021). *Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data* (case study of Mizan computer retail stores). In Journal of Physics: Conference Series, 1722(1)

Stankovičová, I. (2009). *Možnosti extrakcie asociačných pravidiel z údajov pomocou SAS Enterprise Miner*. Informační a datová bezpečnost ve vazbě na strategické rozhodování ve znalostní společnosti, 1-7.

Trnka, A. (2010). *Market basket analysis with data mining methods*. In 2010 International Conference on Networking and Information Technology (pp. 446-450). IEEE.

Zendulka, J. et al. (2006). *Získávání znalostí z databází, studijní opora*. Brno: FIT VUT.

7 Poďakovanie

Príspevok je čiastkovým výstupom projektu VEGA 1/0737/20 Spotrebiteľská gramotnosť a medzigeneračné zmeny v preferenciách spotrebiteľov pri nákupe slovenských produktov a grantu UK/373/2021 Využitie veľkých objemov údajov v oblasti riadenia vzťahov so zákazníkmi.

FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 2 pp 28-38

Do pharmaceutical companies profit from the Covid19 crisis?†
Profitujú farmaceutické firmy z koronakrízy?†

Peter Pšenák, Michal Páleník
Univerzita Komenského v Bratislave, Fakulta managementu, Odbojárov 10, 820 05 Bratislava, Slovenská republika
Comenius University in Bratislava, Faculty of Management, Odbojárov 10, 820 05 Bratislava, Slovak Republic
peter.psenak@fm.uniba.sk , michal.palenik@fm.uniba.sk

Abstract: *Investing into companies developing vaccines (pharmaceutical companies) showed significantly less profitable than investing into the conservative S&P500 index or Facebook Inc, a company providing a social media platform to spread information. This conclusion is suggested by comparing returns for these types of stock investments, controlling for daily fluctuations of stock prices (which could increase or decrease returns twofold).*

Abstrakt: *Investovanie do firiem vyvíjajúcich vakcíny (farmaceutické firmy) sa ukázalo ako menej ziskové ako investovanie do konzervatívneho S&P500 indexu alebo do Facebooku Inc., spoločnosti poskytujúcej mediálnu platformu pre distribúciu informácií. Tento záver plynie na z porovnania výnosov z investícií do týchto akcieových inštrumentov, kontrolujúc fluktuácie denných hodnôt akcí (fluktuácie by mohli dvojnásobne zvýšiť či znížiť výnosy).*

Key words: *covid19 crisis, pharmaceutical companies, stock prices development*

Kľúčové slová: *covid kríza, farmaceutický priemysel, vývoj cien akcií*

1 Introduction

Covid-19 crisis negatively affected all sectors of economy. One of the few sectors, which could be affected positively by a worldwide pandemic, should be the pharmaceutical sector. There might be a possible increase in pharmacy company revenue, which can stem from the Covid pandemic, including PCR and AG testing, treatment of this illness and vaccination. On the other hand, creating tests, treatments and vaccinations for a new disease requires additional resources for research and development. Further, the worldwide pandemic hinders the number of medical procedures of other illnesses as well as their research process.

In this paper we focus on comparing development of pharmaceutical companies during Covid pandemic in years 2020 and 2021. We do not discuss ramifications resulting from the fact that pharmaceutical sector is oligopolistic sector, which is hugely regulated, subsidized by governments and are the main

† The views expressed in this paper are not necessarily those shared by the Editorial Board of Forum Statisticum Slovacum.

customers are governments. We are discussing whether these companies outperform general market.

2 Misinformation causing delays in vaccination

Corona vaccination rollout was slowed down by misinformation, false claims, and hoaxes. A section of these misinformation[4] had wording like "I do not want to fuel pharmaceutical companies' profits therefore I will not get vaccinated". This includes two claims: "pharmaceutical companies greatly financially profit from vaccines" (which we evaluate in this paper) and "vaccines create more profits than treatments" (for pharmaceutical companies).

Price of Covid vaccines varies greatly, but one dose generally costs $15-30. We assume cost of 20 € per dose plus additional costs for administering the dose (mainly local staff salaries).

Costs of Covid treatments in hospitals again vary. In the case of Slovakia, reported prices average at 2 300 €, reaching well above 10 000 € in case of ICU (Aktuality.sk, 2021). Direct costs of non-hospitalized Covid patient are around 100€ (Darcy Jimenez, 2021).

From February 2020 until November 2021 in Slovakia, there were 63 932 hospitalized Covid patients out of 651 234 tested positive by PCR tests (Inštitút zamestnanosti, 2021; *Koronavírus na Slovensku v číslach*, 2021). So around 12 % of population in Slovakia tested Covid positive and 1.2 % of total population ended in hospital (we ignore the few repeated hospitalizations). New Covid mutations will increase the probability of hospitalization and positivity.

Therefore, average expected costs of hospitalization are 28€, average expected costs of non-hospitalization treatments are 12.3 €, totaling 40€. This estimate is extremely conservative not factoring into account secondary health treatments after overcoming the disease, costs resulting from postponing medical procedures, costs resulting from overwork and burnout of health care employees, further complications resulting from new Covid mutations etc. Also, other economic costs resulting from pandemic are not included (subsidies, lost revenues to the state budget, costs of economic downturn, long term social costs, etc.).

Comparing 40 € costs of two vaccination doses with above mentions costs undoubtedly proves that vaccination is far cheaper than treatment of Covid illness.

[4] We intentionally do not make any references to sources of misinformation. We do not want to increase their citation index. Citation indexes, just like social media, do not distinguish between positive and negative citations.

In the later part of this paper, we focus on development of stock prices of pharmaceutical companies.

3 Development of pharmaceutical sector during the Covid19 crisis

For our study we have chosen a specific subset of pharmaceutical companies which we would like to analyze more deeply. We study mainly the development of these pharmaceutical companies:

PFE - Pfizer Inc.

Pfizer Inc. is an America based multinational biotechnology and pharmaceutical corporation. The company was established in 1849 (Pfizer, n.d.). With more than 100 years of their existence, they are able to produce medicines and vaccines used by the general public. It was included in Dow Jones Industrial Average stock market index (*Exxon Mobil, Pfizer Removed from Dow Jones Industrial Average; Salesforce, Honeywell Added*, 2021).

The market cap of the company at the beginning of 2020 was $204.60 B

MRNA - Moderna, Inc.

Moderna, Inc., is a biotechnological and pharmaceutical company based in Cambridge that focuses on RNA therapeutics, primarily mRNA vaccines (Moderna, 2020). Unlike Pfizer they exist only for a short amount of time being established only in 2010. The company's most notable commercial product is the Moderna COVID-19 vaccine(Kollewe, 2020). They have multiple other vaccines and different treatments in their pipeline(Moderna, 2020).

The market cap of the company at the beginning of 2020 was $41.33 B.

JNJ - Johnson & Johnson

Johnson & Johnson is an American multinational corporation, which develops pharmaceuticals, medical devices and packaged goods. It was established in 1886 and their headquarters are in New Jersey (Jonhson & Johnson, n.d.). The company’s stock is in the Dow Jones Industrial Average and is considered as a creditor with AAA rating (Fortune, 2021; *S&P Revises J&J’s Outlook to Negative after $1.5B Boost to Legal Reserve*, 2020).

J&J has announced in early November 2021 that it is planning to split into two different publicly traded companies. The sole purpose of this decision is, that one company will focus primarily on consumer products and the second will only work with pharmaceuticals (Rilley Griffin, 2021).

The market cap of the company at the beginning of 2020 was $414.30 B.

AZN - AstraZeneca PLC

AstraZeneca plc is a British-Swedish multinational biotechnological and pharmaceutical company. The company was founded in 1999 and it has its

headquarters in Cambridge (*AstraZeneca > GC Powerlist: Sweden Teams 2019*, n.d.; AstraZeneca, 2021). It has a wide portfolio of products for a wide variety of major diseases (*Key Facts - Our Company - AstraZeneca*, 2021). AstraZeneca is listed on the NASDAQ.

The market cap of the company at the beginning of 2020 was $130.98 B.

We compare this group of successful companies (from the view of vaccine development) with other pharmaceutical companies, which have not developed successful Covid vaccine yet. Comparing these two groups will allow for distinguishing the effect of vaccine development of stock prices.

Other group for comparison will be non-pharmaceutical companies represented by general stock market indexes (S&P500).

3.1 Stock prices

We understand company's stock prices as a composite indicator of current situation and expected future development of given company. Our stock prices were downloaded from Yahoo finance (*Yahoo Finance - Stock Market Live, Quotes, Business & Finance News*, 2021), with the help of R (R Core Team, 2021), R-studio and especially the quantmod (Ryan et al., 2017) and PerformanceAnalytics (Peterson et al., 2020) packages. We have chosen Yahoo for its fast API and clean data, which it provides for free (Pšenák & Pšenáková, 2018).

The main question is: Did pharmaceutical stock prices increase during Covid19 pandemic? To be able to answer this question, we will use multiple visualizations and different calculations. In Fig. 1 we show all the prices of every pharmaceutical company in our dataset. As one might immediately realize, the different stock had completely different performances. One of our main analyzed companies, namely Moderna had the biggest jump of their stock prices in the selected years. This might come from the reality, that it is a young company which was between the first companies with usable vaccine for Covid pandemic. We can consider it as an extreme value. Secondly the value of REGN (Regeneron Pharmaceuticals) was extremely high even in the beginning of the pandemic. Most of the company stock values were way below 100$, and except Moderna, no other companies had such a rapid growth of stock price in the analyzed timeframe. Overall, the companies differ in the relative value of stock prices as well as the different stock price changes.

In our second analysis we normalized the values of analyzed companies to 100 $ at the beginning, so we were able to analyze only the price changes. The outcome of this normalization is shown on Figure 2. The relative change of the different stocks is showing us, that the Moderna stock price change was

extremely high in comparison to any other company. As for the other 3 companies of our main interest the changes of prices are close to zero, and there is no implicit correlation between the actual pandemic situation and their market value.

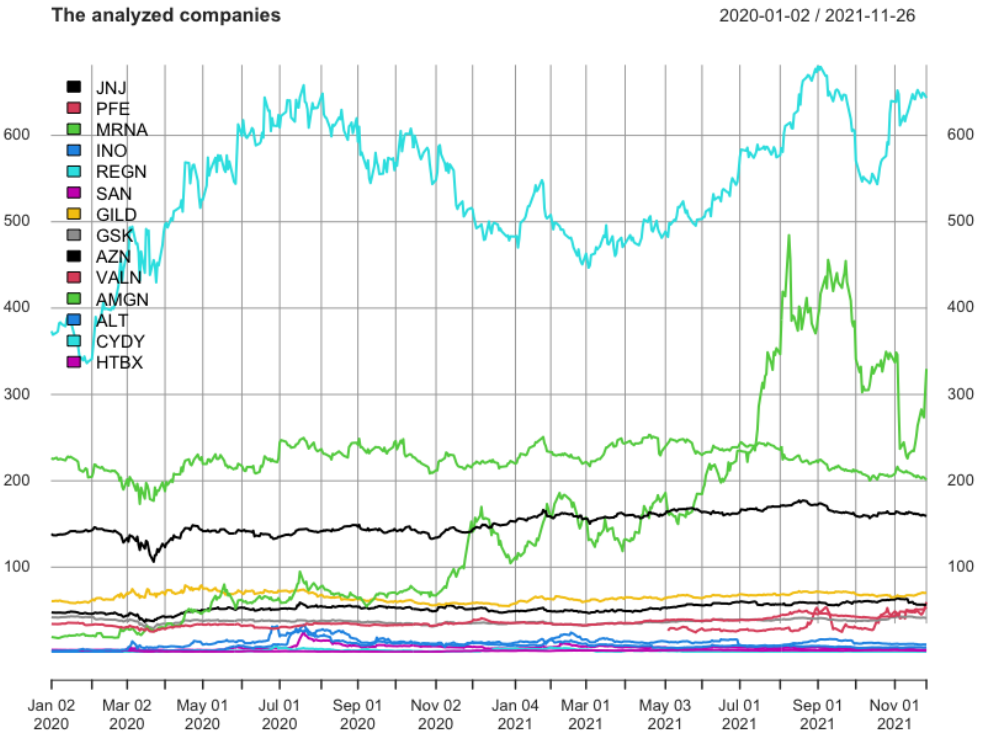


Fig. 1 Stock prices of selected companies, in $ *(Source: the authors)*

For further analysis of the stocks, we calculated some basic descriptive statistics. They helped us to decide, which companies we will analyze further, and which company values changed so extremely, that it would bias our further research. After this analysis we decided to not use MRNA for further analysis, because of it extreme values and changes.

3.2 Stock Prices in comparison to market development

As we described in chapter 3.1, stock prices of pharmaceutical companies increased during the pandemic but so did stock prices of other companies. In this chapter we study development of pharmaceutical companies' stock prices over market development: did these companies outperform general market?

Figure 3 shows traditional approach to comparing stock prices – normalized at beginning of selected period. We can see, that S&P500 (symbol GSPC) yields higher results than other stocks (except for FB).

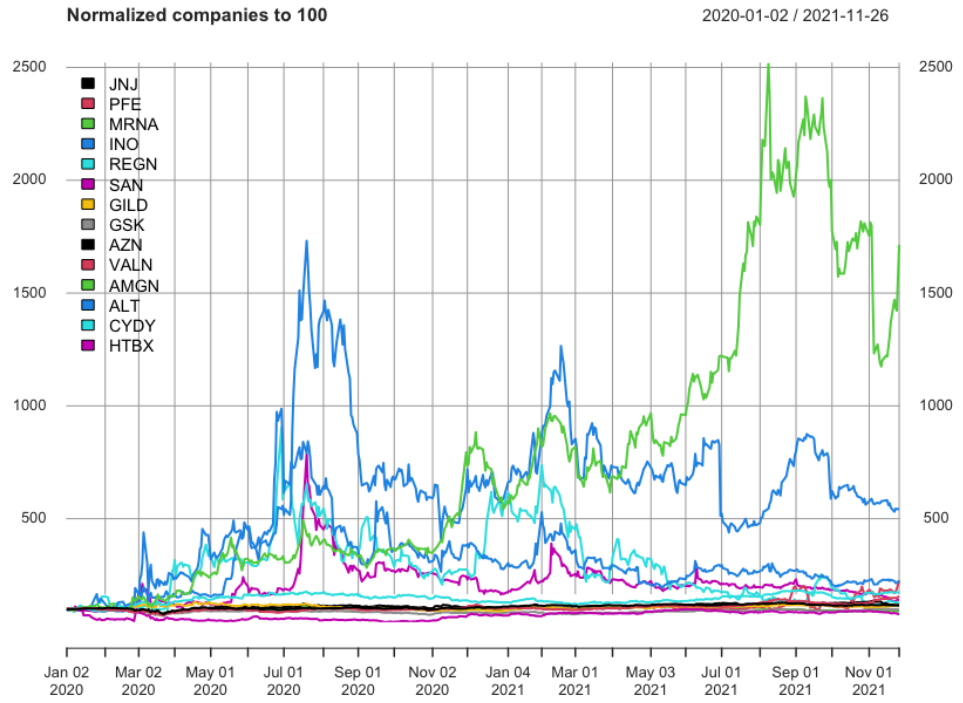


Fig. 2 Stock prices of selected companies, average of January 2020 = 100 *(Source: the authors)*



Fig. 3 Development of stock prices, January 2020 = 1 *(Source: the authors)*

The easiest way to compare performance of two stocks or portfolios is by arbitrary choosing dates for buying and selling this investment. This approach is

extremely sensitive to the choice of date – yield can vary. Table 1 describes stock prices in the month of March 2020. It is not uncommon to see differences of 30% - within one month.

Tab. 1 Statistics of stock prices in March 2020 *(Source: the authors)*

	PFE	MRNA	JNJ	AZN	GSPC
Min	25.25	21.30	106.2	36.49	2237
1st Qu	26.80	25.99	120.1	39.08	2454
Median	28.59	27.93	126.5	40.95	2606
Mean	28.57	27.15	125.2	41.58	2652
3rd Qu	30.2	29.28	130.5	43.88	2848
Max.	32.26	31.58	137.1	46.92	3130

Figure 4 shows the difference of the yearly returns based on date of stock purchase. The first black line shows an investor who would buy a PFE share at the beginning of the year 2020 and the second red line shows an investor who would buy the stock immediately after the march price fall. Both investors would keep their respective stock for the same amount of time however the second one would gain from the changes of price in late 2021, while the first investor would sell his share after a share price drop at the end of the year of 2020.

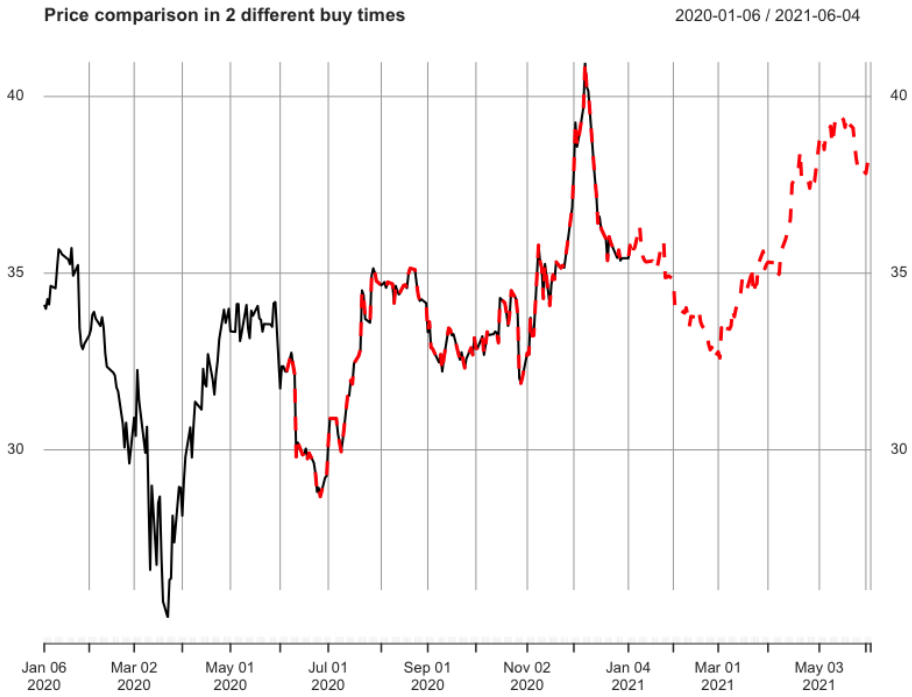


Fig. 4 Graph of yearly returns of PFE based on date of purchase *(Source: the authors)*

In March 2020, Covid was already known, and stock prices were decreasing (Lyócsa et al., 2020). An optimistic investor might expect pharmaceutical companies to increase in value. However, returns of his investments would vary based on exact date of purchase.

To overcome the volatility among dates, we calculate return in % p.a. for investment starting in every day of March 2020 and every day of November 2021.

Each dot in Fig. 5 shows yearly returns of investment of chosen companies. Start date is a day in March 2020, end date is in November 2021.

As we can see from Figure 5 and Tab. 2, the returns from investment for PFE are on average 44 %, AZN 27 %, J&J 18 %. S&P500 index performance was 46 %. Returns to an utterly useless company (at least from a vaccine development point of view) like Facebook were 63 %.

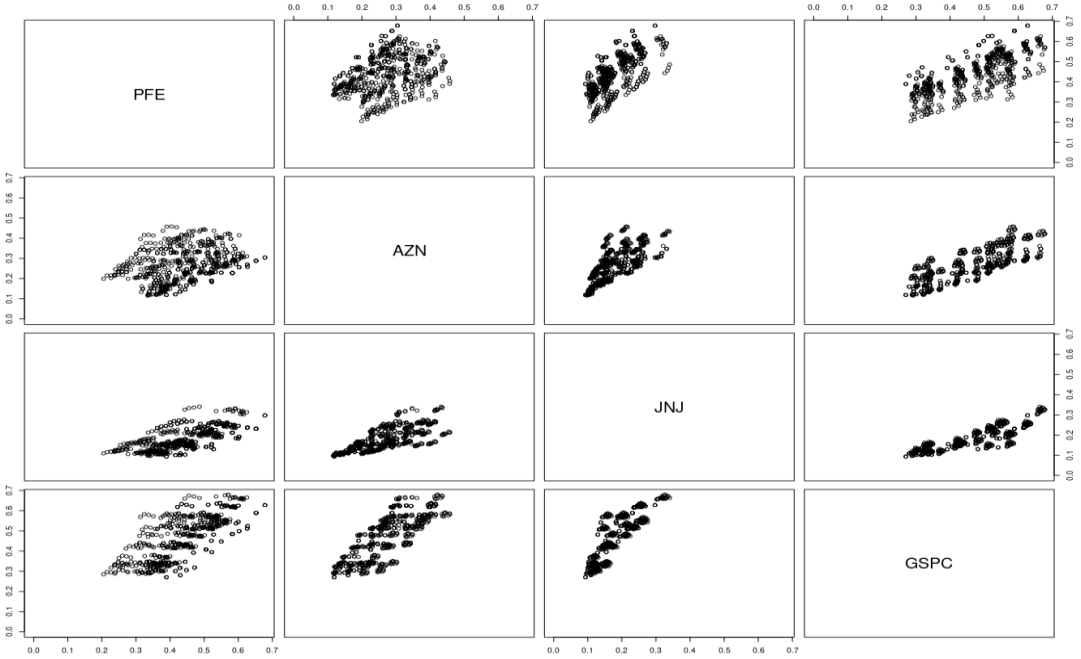


Fig. 5 Returns on investment between March 2020 and November 2021, in % p.a., each dot represents different purchase and selling dates *(Source: the authors)*

Figure 6 visualizes performance differences as well as fluctuations of various assets. As we can see, extremely lucky investor investing into PFE could achieve returns of 67 %, unlucky investor 20 %, averaging at 44.5 % (depending on the choice of dates). Investing into JNJ would yield between 9 % and 34 % with average of 18 %. However, investing into Facebook would give returns between 39 % and 82 % (average 63 %). So an extremely lucky investor into JNJ would be outperformed by anyone investing into the Facebook company, which was accussed of allegedly providing a platform to spread misinformation and hoaxes.

Tab. 2 Returns on investment between March 2020 and November 2021, in % p.a., each dot represents different purchase and selling dates *(Source: the authors)*

	PFE	AZN	JNJ	FB	SP500
Min	20.51	11.63	9.301	39.41	27.02
1st Q	37.46	21.66	13.485	52.62	34.66
Median	44.41	26.39	16.626	64.04	47.80
Mean	44.66	26.67	18.114	63.05	46.52
3rd Q	52.11	32.81	23.027	72.98	55.13
Max	67.81	45.82	34.016	82.73	67.79

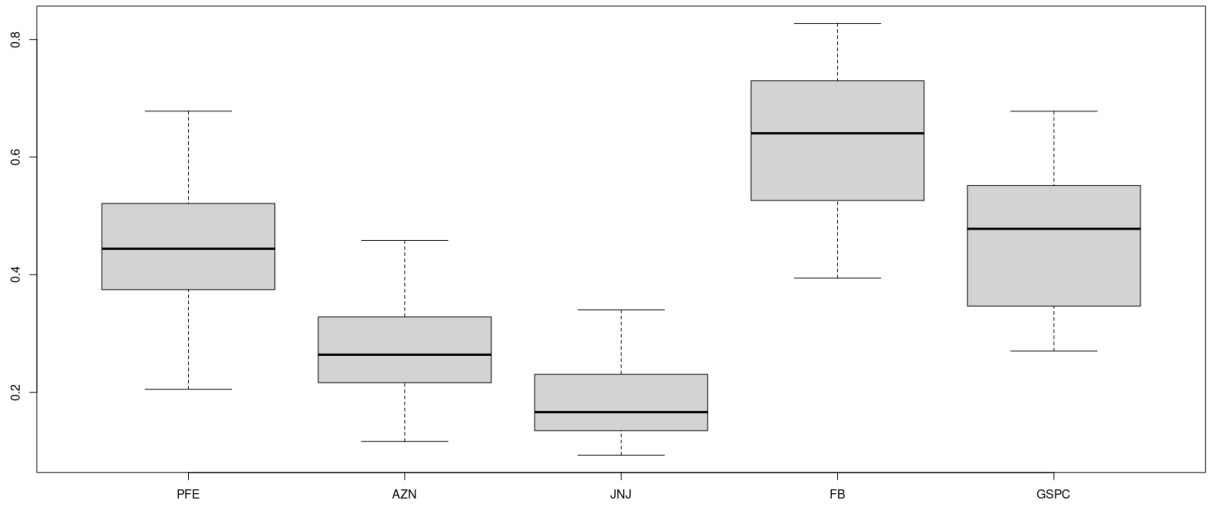


Fig. 6 Returns on investment between March 2020 and November 2021, in % p.a. *(Source: the authors)*

4 Conclusions

In this paper we compared development of stock prices of pharmaceutical companies. We accounted for stock market fluctuations by calculating returns to investment between all days in March 2020 and November 2021. These daily fluctuations in connection to stock market development in March 2020 could increase yearly returns twofold to fourfold. This averaging allowed us to calculate more robust results.

As we saw in chapter 3, pharmaceutical companies with vaccines did not outperform the general stock market. Even if an investor bought pharmaceutical stocks in the lowest prices (March 2020) and sold them in November 2021, investing to general S&P500 index would yield higher returns than Pfizer, 2 times higher p.a. returns than AstraZeneca or Johnson & Johnson. So pharmaceutical companies did not profit from Covid pandemic more than the general stock market.

5 References

Aktuality.sk (2021). *Koronavírus: Poisťovne dali na liečbu pacientov s COVID-19 takmer 60 miliónov eur*. Aktuality.sk. https://www.aktuality.sk/clanok/891228/koronavirus-poistovne-dali-na-liecbu-pacientov-s-covid-19-takmer-60-milionov-eur.

AstraZeneca (2021). *Contact information*. https://www.astrazeneca.com/investor-relations/annual-reports/annual-report-2016/contact-information.html.

AstraZeneca > GC Powerlist: Sweden Teams 2019 (n.d.). [cit. 2019, Nov 30] https:// www.legal500.com/gc-powerlist/sweden-teams-19/astrazeneca-3.

Darcy Jimenez (2021). *Covid-19: Vaccine pricing varies wildly by country and company*. https://www.pharmaceutical-technology.com/analysis/covid-19-vaccine-pricing-varies-country-company.

Exxon Mobil, Pfizer removed from Dow Jones Industrial Average; Salesforce, Honeywell added (2021). USA TODAY. https://www.usatoday.com/story/money/2020/08/25/exxon-pfizer-dow-jones-industrial-average-salesforce-djia/5630916002.

Fortune (2021). *Fortune 500*. Fortune. https://fortune.com/fortune500/2021.

Inštitút zamestnanosti. (2021). *Covid na Slovensku—Inštitút zamestnanosti*. https:// www.iz.sk/korona.

Jonhson & Johnson (n.d.). *Johnson & Johnson our story*. [cit. 2019, Nov 30] https:// ourstory.jnj.com/our-beginning.

Key facts — our company — AstraZeneca (2021). https://www.astrazeneca.com/investor-relations/key-facts.html.

Kollewe, J. (2020, Nov 16). Covid vaccine: Who is behind the Moderna breakthrough? *The Guardian*. https://www.theguardian.com/world/2020/nov/16/covid-vaccine-who-is-behind-the-moderna-breakthrough.

Koronavírus na Slovensku v číslach (2021). Koronavírus a Slovensko. https://korona.gov.sk/koronavirus-na-slovensku-v-cislach.

Lyócsa, Š. et al. (2020). Fear of the coronavirus and the stock markets. *Finance Research Letters*, *36*, 101735.

Moderna (2020). *Moderna. Form 10-K 2020.* https://www.sec.gov/ix?doc=/Archives/edgar/data/0001682852/000168285221000006/mrna-20201231.htm.

Peterson, B. G. et al. (2020). *PerformanceAnalytics: Econometric tools for performance and risk analysis*. R package 2.0.4. https://cran.r-project.org/package=PerformanceAnalytics.

Pfizer (n.d.). *Pfizer. Form 10-K 2020.* (Washington, D.C. 20549).

Pšenák, P., Pšenáková, I. (2018). Az R-rel szebb a statisztika. *InfoDidact 2018: 11*, 169–175.

R Core Team (2021). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Rilley Griffin (2021, Nov 12). Health-care giant J&J to split into drug and consumer compa-nies. *Bloomberg.Com*. https://www.bloomberg.com/news/articles/2021-11-12/johnson-johnson-to-split-into-two-companies-shares-rise.

Ryan, J. A. et al. (2017). *quantmod: Quantitative financial modelling framework*. R package 0.4-12 https://cran.r-project.org/package=quantmod.

S&P revises J&J's outlook to negative after $1.5B boost to legal reserve (2020). https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/ s-p-revises-j-j-s-outlook-to-negative-after-1-5b-boost-to-legal-reserve-60982518.

Yahoo Finance — stock market live, quotes, business & finance news (2021). https://finance.yahoo.com.

FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 2 pp 39-51

Cesta do hlubin DATA kroku programového systému SAS aneb co by měl každý programátor o systému SAS vědět
Journey into the depths of the SAS DATA step or what every programmer should know about SAS

Roman Pavelka
Štatistický úrad Slovenskej republiky, Odbor metód štatistických zisťovaní, Lamačská cesta 3, 840 05 Bratislava, Slovenská republika
Statistical Office of the Slovak Republic, Statistical Surveys and Methodology Department, Lamačská cesta 3, 840 05 Bratislava, Slovak Republic
roman.pavelka@statistics.sk

Abstrakt: *DATA krok je nejmocnějším nástrojem programového systému SAS. Pochopení fungování DATA kroku, co se odehrává a proč, je klíčové pro efektivní zvládnutí kódu a výstupů. V předkládaném příspěvku budou vzpomenuty následující koncepty: (1) co se děje při kompilaci DATA kroku programu, (2) co se děje v době vykonávání DATA kroku programu, (3) programový vektor dat (označovaný zkratkami LPDV nebo PDV), (4) automatické proměnné SAS v DATA kroku a jejich použití, (5) pochopení vnitřních principů zpracování DATA kroku, (6) atributy proměnných a jak se zachycují a ukládají, a (7) zacházení s výchozími hodnotami zpracování DATA kroku a chybějícími hodnotami. Článek se zaměřuje zejména na techniky, které využívají programové možnosti DATA kroku a práci s datovými soubory za platnosti výchozích akcí. Porozumění principů činnosti DATA kroku umožňuje programátorům efektivně ladit své programy a s jistotou interpretovat dosahované výsledky.*

Abstract: *DATA step is the most powerful tool of the SAS programming system. Understanding how the DATA step works, what happens and why, is key to effectively managing code and outputs. In the present paper, the following concepts are addressed: (1) Program Data Vector (abbreviated as LPDV or PDV), (2) SAS automatic variables and their use, (3) understanding the internal principles of DATA step processing, (4) what happens during the program compilation, (5) what happens at program execution time, (6) variable attributes and how they are captured and stored, and (7) handling processing defaults and missing values. In particular, the paper focuses on techniques that exploit the program's DATA step capabilities and the handling of data sets under the validity of default actions. Understanding the principles of DATA step operation allows programmers to efficiently debug their programs and confidently interpret the results obtained.*

Kľúčové slová: *kompilace programu, programový datový vektor, SAS , strojový kód*

Key words: *program compilation, program data vector, machine code, SAS*

1 Úvod

I když jsou procedury SAS velmi výkonné a snadno použitelné, DATA krok nabízí programátorovi nástroj s téměř neomezenými možnostmi. V programátorské praxi se uživatelé programového systému SAS často potýkají s neuspořádanými daty, složitými analytickými úlohami i vysokými nároky na

prezentaci výsledků. DATA krok může všem zvídavým programátorům, jejich programům a jejich uživatelům pomoci lépe fungovat v reálném světě, a to zejména pokud se efektivně využijí dostupné funkce DATA kroku. Článek se zaměřuje na základní techniky, které demonstrují fungování kroku DATA a ilustrují výchozí akce. Kterékoliv z probíraných témat, resp. příkladů v této prezentaci obsahuje více než dost podrobností, zvláštností a výhrad, aby si zasloužily vlastní výukový kurz. Proto jsou uvedeny pouze vybrané základní tipy pro zpracování dat a názorné příklady:

- činnosti v době kompilace instrukcí DATA kroku,
- činnosti v době provádění přeloženého strojového kódu DATA kroku,
- vybrané příkazů data DATA kroku pro vytváření datových souborů SAS,
- automatické proměnné v rámci DATA kroku,
- uchování hodnoty proměnné mezi jednotlivými iteracemi v DATA kroku,
- využití explicitních cyklů v DATA kroku a
- efektivní vytváření datových souborů programového systému SAS.

2 Přehled činnosti DATA kroku při zpracování dat

Činnost DATA kroku při zpracování dat se rozděluje na fázi kompilační a fázi exekuční. Kompilování DATA kroku a vykonání DATA kroku z programu SAS jsou odlišné - nezávislé a sekvenční – činnosti (Howard, 2004). Nejprve je každý datový krok zkompilován a poté vykonán, což se týká i všech programových procedur. Rozhodujícím k dosažení požadovaných výsledků je také pochopení výchozích nastavení jednotlivých činností v DATA kroku. Nastavení jednotlivých činností DATA kroku se dají ovlivnit příkazy DATA kroku. Příkazy v DATA kroku se obecně rozlišují na deklarativní a výkonné. V kompilační fázi jsou použity pouze deklarativní příkazy, které poskytují informace pro logický programový datový vektor (v dalším textu zkratkou LPDV). Deklarativní příkazy lze v rámci kroku DATA umístit v libovolném pořadí. Několik příkladů deklarativních příkazů DATA kroku je uvedeno v tabulce 1.

Do výstupního datového souboru se dostanou proměnné identifikované během kompilace programu. Součástí výstupního datového souboru jsou i tzv. automatické proměnné systému SAS, které se však do výstupního datového souboru za platnosti výchozího nastavení nezapisují. Automatické proměnné jsou systémové proměnné SAS, které se vytvářejí automaticky v DATA kroku. Tyto proměnné (*_N_*, *_ERROR_*, *end=*, *in=*, *point=*, *first.*, *last.*)[5] se přidávají do

[5] Automatická proměnná *_N_*- zachycuje počet iterací (pozorování) v DATA kroku, *_ERROR_* je nastavena na 1, pokud se objevila při exekuci DATA kroku chyba, *end=* detekuje konec datového souboru, *in=* dočasná proměnná rovná 1, pokud soubor vstupních dat vstupuje do aktuálního pozorování v PDV, *point=* dočasná proměnná k přímému přístupu k proměnným v LPDV, *first.* resp.,

datového vektoru programu[6], ale nezobrazují se ve výstupním datovém souboru. Hodnoty automatických proměnných se zachovávají od načtení jednoho pozorování DATA kroku k dalšímu. Proměnné zapisované do výstupního souboru jsou určeny uživatelskými příkazy DROP a/nebo KEEP případě dalšími parametry, přičemž všechny neautomatické proměnné jsou standardně v programovém vektoru LPDV.

Tab. 1 Příklady deklarativních příkazů DATA kroku *(Zdroj: Li, 2013)*

Název příkazu	**Význam příkazu pro činnost DATA kroku**
LENGTH	nastavení vnitřní délky proměnné
FORMAT	nastavení výstupního formátu proměnné
LABEL	definování popisků proměnných
DROP	označuje proměnné, které mají být ve výstupním souboru vynechány
KEEP	označuje proměnné, které mají být do výstupního souboru zahrnuty
RETAIN	označuje proměnné, u které mají být uchovány mezi iteracemi hodnoty

Během kompilace DATA kroku se realizují následující nejvýznamnější činnosti:

- skenování syntaxe napsaného zdrojového kódu SAS,
- přeložení zdrojového kódu SAS do strojového jazyka,
- definují se vstupní a výstupní soubory dat,
- vytvářejí se programové prostředky:
  - vstupní vyrovnávací paměť (a to v případě, že se načítají jiná než nativní datové soubory systému SAS),
  - logický programový datový vektor (LPDV) a
  - informace o deskriptoru datové sady.
- určují se atributy proměnných pro výstupní datovou sadu SAS,
- zaznamenávají se proměnné, které mají být inicializovány jako chybějící.

Ve výchozím nastavení nejsou hodnoty proměnných deklarovaných v příkazu INPUT a uživatelsky definovaných proměnných (např. s příkazem přiřazení) uchovávány během načítání jednotlivých pozorování v DATA kroku, pokud se na ně neodkazuje v příkazu RETAIN. Mezi proměnné, jejichž hodnoty se uchovávají během vykonávání celého DATA kroku, patří:

- všechny speciální automatické proměnné DATA kroku
- všechny proměnné uvedené v příkazu RETAIN

last. jsou proměnné nabývající hodnoty 1 při seskupování proměnných pomocí příkazu BY, kdy proběhne zpracování prvního, resp. posledního pozorování seskupené proměnné.

[6] Programový datový vektor se také nazývá logický programový datový vektor. V anglickém překladu je známý jako Logical Program Data Vector nebo také Program Data Vector (zkráceně LPDV nebo PDV).

- všechny proměnné načtené příkazem SET, MERGE nebo UPDATE
- kumulované proměnné v příkazu SUM.

2.1 Načtení externího textového souboru DATA krokem

Na rozdíl od deklarativních příkazů, na pořadí, v jakém se spustitelné příkazy objevují v kroku DATA, záleží. Je-li například nutné načíst externí textový soubor, DATA krok začíná příkazem INFILE, po kterém následuje příkaz INPUT. Příkaz INFILE slouží k identifikaci umístění externího souboru a příkaz INPUT instruuje systém SAS, jak má číst jednotlivá pozorování. Příkaz INFILE tedy je potřebné umístit před příkaz INPUT, protože systém SAS potřebuje vědět, kde má externí soubor najít, aby jej mohl přečíst.

V exekuční fázi pracuje krok DATA v cyklech, který opakovaně provádí příkazy pro zpracování dat z jednoho pozorování po druhém. Každý cyklus se nazývá iterace a dochází při něm k načítání hodnot pozorování a jejich ukládání do výstupního souboru. Uvedený cyklus se označuje jako implicitní cyklus.

Jak funguje zpracování dat DATA krokem, znázorňuje následující program:

```
data ex3_1;
     infile 'Cesta k souboru.....\example3_1.txt';
     input name $ 1-7 group $ 9 weight 11-12 height 14-16;
     BMI = 700*weight/(height*height);
     output;
run;
```

Program načte data uložená v textovém souboru EXAMPLE3_1.TXT a vytvoří jednu proměnnou, pojmenovanou BMI. Soubor EXAMPLE3_1.TXT obsahuje tři pozorování a 4 proměnné: NAME (sloupce 1-7), GROUP (sloupec 9), HEIGHT (sloupce 11-12) a WEIGHT (sloupce 14-16). Hodnota proměnné HEIGHT je u prvního pozorování zadána jako 12D, což je chyba při zadávání dat. Vzhledem k tomu, že každá proměnná zabírá stejné pozice v každém pozorování a hodnoty těchto proměnných jsou standardní znakové nebo číselné hodnoty, je pro čtení externího souboru textových dat nejvhodnější použít metodu zadávání délek sloupců v příkazu INPUT. Strukturu a obsah souboru EXAMPLE3_1.TXT ilustruje tabulka 2.

Tab. 2 Proměnné a pozorování souboru EXAMPLE3_1.TXT *(Zdroj: Li, 2013)*

Sloupec TXT souboru č.	1–7	9	11–12	14–16
Jméno proměnné	NAME	GROUP	WEIGHT	HEIGHT
Pozorování č. 1	Barbara	A	61	12D
Pozorování č. 2	John	B	62	175
Pozorování č. 3	Alex	B	72	165

2.2 Kompilační fáze činnosti DATA kroku

Na začátku fáze kompilace je vytvořena vstupní vyrovnávací paměť, do které DATA budou načteny jednotlivá pozorování souboru dat. Vstupní vyrovnávací dat slouží k uchovávání načítaných dat. Pokud však DATA krok načítá datový soubor ve formátu SAS[7], potom se vstupní vyrovnávací paměť nevytváří.

V kompilační fázi se také vytváří LPDV. LPDV obsahuje automatické proměnné _N_ a _ERROR_. Automatická proměnná _N_ rovnající se 1 znamená, že se zpracovává první pozorování, _N_ rovnající se 2 znamená, že se zpracovává druhé pozorování atd. Automatická proměnná _ERROR_ je indikátorová proměnná s hodnotami 1 nebo 0. _ERROR_ o hodnotě 1 signalizuje chybu dat aktuálně zpracovávaného pozorování, například čtení dat s nesprávným datovým typem. Kromě těchto dvou automatických proměnných je pro každou z proměnných, které budou vytvořeny z tohoto kroku DATA, vyhrazeno jedno místo v LPDV. Proměnné NAME, GROUP, WEIGHT a HEIGHT jsou načteny z externího souboru, zatímco nově vytvořená proměnná BMI (index tělesné výkonnosti) je odvozena z proměnných WEIGHT a HEIGHT. Vstupní vyrovnávací paměť, ve které je už načteno první pozorování z externího textového souboru, a LPDV jsou schematicky zaznamenány v tabulce 3.

Tab. 3 Vstupní vyrovnávací paměť s načteným pozorováním a LPDV *(Zdroj: Li, 2013)*

Vstupní vyrovnávací paměť	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Načtené pozorování č. 1	**B**	**a**	**r**	**b**	**a**	**r**	**a**		**A**		**6**	**1**		**1**	**2**	**D**

	_N_ (D)	**_ERROR_ (D)**	**NAME (K)**	**GROUP (K)**	**WEIGHT (K)**	**HEIGHT (K)**	**BMI (K)**
LPDV							

Pokud jsou některé proměnné v LPDV jsou označeny písmenem (D), což znamená DROP, a jiné jsou označeny písmenem (K), což znamená KEEP, budou do výstupního souboru dat zapsány pouze proměnné označené písmenem (K). Naproti tomu automatické proměnné jsou vždy označeny symbolem (D), takže se ve výchozím stavu nikdy nevypíší do výstupního souboru. Během fáze kompilace systém SAS kontroluje, zda nedošlo k chybám v syntaxi, jako jsou neplatné názvy proměnných, parametry, interpunkce, chybně napsaná klíčová slova, atd. Programový systém SAS také identifikuje typ a délku nově vytvořených proměnných. Na konci fáze kompilace se vytvoří popisná část datového souboru SAS, tzv. deskriptor dat, který obsahuje název datového souboru, počet pozorování a počet, názvy a atributy proměnných.

[7] Souborem ve formátu SAS se myslí soubor dat s příponou SAS7BDAT.

2.3 Výkonná fáze činnosti DATA kroku

Na začátku exekuční fáze činnosti DATA kroku jsou automatické proměnná _N_ inicializovány na výchozí hodnoty, tj. _N_ na 1 a _ERROR_ na 0 (protože na začátku nedošlo k chybě dat). Neautomatické proměnné jsou nastaveny na chybějící (tzv. MISSING). Jakmile příkaz INFILE identifikuje umístění vstupního souboru, příkaz INPUT zkopíruje první datový řádek do vstupní vyrovnávací paměti. Poté příkaz INPUT přečte hodnoty dat ze záznamu ve vstupní paměti podle instrukcí příkazu INPUT a zapíše je do LPDV. Hodnoty pro proměnné NAME, GROUP, WEIGHT a HEIGHT jsou úspěšně zkopírovány ze vstupní vyrovnávací paměti do LPDV. Hodnota proměnné HEIGHT však činí 12D, což je neplatná číselná hodnota. Tato neplatná číselná hodnota způsobí, že se _ERROR_ nastaví na 1 a hodnota proměnné HEIGHT pro dané pozorování bude chybět. Mezitím je do protokolu SAS odeslána chybová zpráva s uvedením místa, kde došlo k chybě dat (viz protokol z činnosti programu s výpisem v tabulce 4). Dále se provede příkaz přiřazení a BMI zůstane nevyplněna, protože operace nad chybějící hodnotou vedou k chybějící hodnotě.

Tab. 4 Protokol z činnosti programu s výpisem chyb *(Zdroj: vlastní zpracování)*

```
NOTE: Invalid data for height in line 1 14-16.
name=Barbara group=A weight=61 height=. BMI=. _ERROR_=1 _N_=1
NOTE: 3 records were read from the infile '....\example3_1.txt'.
      The minimum record length was 14.
      The maximum record length was 14.
   NOTE: Missing values were generated as a result of performing an
 operation on missing values.
      Each place is given by: (Number of times) at (Line):(Column).
      1 at 34:11

NOTE: The data set WORK.EX3_1 has 3 observations and 5 variables.
```

Po provedení příkazu OUTPUT se do výstupního datového souboru pod názvem EX3_1 kopírují pouze hodnoty těch proměnných z LPDV označených (K) jako jedno pozorování.

Na konci čtení a zápisu jednoho pozorování (iterace) DATA kroku se systém SAS vrátí na začátek DATA kroku, aby zahájil další iteraci (načtení dat). Hodnoty proměnných v LPDV se nastavují na chybějící. Automatická proměnná _N_ se zvýší na 2 a proměnná _ERROR_ se nastaví na 0. Druhý datový řádek se opětovně načte do vstupní vyrovnávací paměti příkazem INPUT. I když se automatické proměnné _N_ a _ERROR_ implicitně nedostávají do výstupního datového souboru, je možné tyto proměnné využít programově pro testování správnosti zápisu dat do LPDV (proměnná _ERROR_), případně pro sledování počtu načítaných pozorování z textového souboru (proměnná _N_).

Na konci druhé iterace DATA kroku se systém SAS opět vrátí na začátek DATA kroku a zahájí další iteraci. Hodnoty proměnných v LPDV se nastavují na chybějící. Automatická proměnná _N_ se zvýší na 3. DATA krok se pokusí přečíst následující pozorování ze vstupního souboru dat. Po načtení posledního pozorování (závěrečná iterace) dosáhne značky konce souboru, což znamená, že již není žádné pozorování, které by bylo možné přečíst. Jakmile dojde ke značce konce souboru, přejde systém SAS k dalšímu kroku DATA nebo proceduře v programu.

2.4 Důležitost příkazu OUTPUT v činnosti DATA kroku

Ve výše uvedené ukázce programu v podkapitole 2.1 se používá příkaz OUTPUT. Tím, že je v programu přímo příkaz uveden, nazývá se jako tzv. explicitní příkaz OUTPUT. Příkaz OUTPUT požaduje po systému SAS, aby okamžitě zapsal aktuální pozorování z LPDV do datového souboru SAS. Ve skutečnosti však není nutné v programu příkaz OUTPUT uvádět, protože obsah LPDV je standardně vypsán tzv. implicitním příkazem OUTPUT (který se v programu neuvádí) na konci DATA kroku. V následující programové sekvenci byl explicitní příkaz OUTPUT vypuštěn, protože implicitní příkaz OUTPUT (v DATA kroku nezaznamenaný) funguje stejně jako jeho explicitní protějšek.

```
data ex3_1;
     infile 'Cesta k souboru.....\example3_1.txt';
     input name $ 1-7 group $ 9 weight 11-12 height 14-16;
     BMI = 700*weight/(height*height);
run;
```

Pokud je do DATA krok umístěn explicitní příkaz OUTPUT, explicitní příkaz OUTPUT přepíše implicitní příkaz OUTPUT; jinými slovy, jakmile je explicitní příkaz OUTPUT použit k zápisu pozorování do výstupní datové sady, na konci DATA kroku již není implicitní příkaz OUTPUT potřebný. Systém SAS přidá pozorování do výstupní datové sady pouze tehdy, když je proveden explicitní příkaz OUTPUT. Kromě toho lze v kroku DATA použít více než jeden příkaz OUTPUT. Příklad použití více příkazů OUTPUT v DATA kroku bude uveden dále.

2.5 Uchování hodnoty proměnné DATA kroku pro kumulované proměnné

V některých situacích je žádoucí si uchovat hodnoty z nově vytvořených proměnných v LPDV po celou dobu provádění kroku DATA. Aby se zabránilo inicializaci nově vytvořených proměnných na začátku každé iterace implicitního cyklu v DATA kroku na chybějící hodnotu, lze použít příkazu RETAIN. V příkazu RETAIN se uvádí název proměnné, jejíž hodnota se má uchovat mezi jednotlivými iteracemi (mezi načítáním jednotlivých pozorování ze souboru dat), a hodnota, která se použije k inicializaci kumulované proměnné, a to pouze při první iteraci provádění DATA kroku. Pokud nebude počáteční hodnota explicitně zadána,

bude kumulovaná proměnná inicializována jako chybějící před prvním vykonáním DATA kroku. Příkaz RETAIN zabrání inicializaci proměnné při každém provedení kroku DATA. Ukázku použití příkazu RETAIN k vytvoření kumulované proměnné TOTAL_WEIGHT (celková hmotnost) ilustruje následující program.

```
data ex3_2;
     infile 'Cesta k souboru.....\example3_1.txt';
     input name $ 1-7 group $ 9 weight 11-12 height 14-16;
     BMI = 700*weight/(height*height);
     retain total_weight 0;
     total_weight = sum(total_weight, weight);
run;
```

Exekuční fáze DATA kroku začíná ihned po dokončení fáze kompilace DATA kroku. Na začátku exekuční fáze jsou proměnné NAME, GROUP, WEIGHT a HEIGHT nastaveny na chybějící hodnoty, proměnná TOTAL_WEIGHT je však z důvodu použití příkazu RETAIN inicializována na 0. Příkaz INPUT zkopíruje první datový řádek do vstupní vyrovnávací paměti. Poté příkaz INPUT přečte hodnoty dat ze záznamu ve vstupní paměti podle instrukcí příkazu INPUT a zapíše je do LPDV. Příkaz RETAIN je příkazem pouze v době kompilace, v exekuční fázi se už neuplatní. Teprve poté se vypočítá proměnná TOTAL_WEIGHT. V programu není explicitní příkaz OUTPUT, a proto zápis pozorování do datové sady způsobí implicitní příkaz OUTPUT na konci DATA kroku. Systém SAS se vrátí na začátek kroku DATA a zahájí další iteraci (načtení dat z externího souboru). Tímto způsobem probíhá zpracování dat DATA krokem až do poslední iterace. Hodnoty nově vytvořené proměnné jsou v důsledku použití příkazu RETAIN uchovávány v průběhu jednotlivých iterací (na rozdíl od ostatních proměnných, které po skončení každé iteraci nastaví na chybějící hodnotu). Uchování hodnot v proměnné TOTAL_WEIGHT proto umožní sumaci proměnné WEIGHT. Stejného výsledku se dosáhne, pokud se v DATA kroku namísto přiřazovacího příkazu „total_weight = sum(total_weight, weight);“ použije příkaz „total_weight + weight;“. Strukturu a obsah výsledného datového souboru s kumulovanou proměnnou TOTAL_WEIGHT zobrazuje tabulka 5. (včetně nově vytvořené proměnné BMI).

Tab. 5 Proměnné a obsah souboru zpracovaného DATA krokem *(Zdroj: vlastní zpracování)*

NAME	GROUP	WEIGHT	HEIGHT	BMI	TOTAL_WEIGHT
Barbara	A	61	12D	.	61
John	B	62	175	31.867845994	123
Alex	B	72	165	22.280092593	195

2.6 Rozdíly mezi zpracováním datového SAS souboru a textového souboru

Pokud DATA krok vytváří výstupní datový soubor ve formátu SAS načítáním textového souboru, inicializuje DATA krok na začátku každé iterace exekuční fáze každou hodnotu proměnné v LPDV na chybějící, s výjimkou automatických proměnných, proměnných pojmenovaných v příkazu RETAIN, proměnných vytvořených příkazem SUM a proměnných vytvořených pomocí příkazu INFILE.

Často se výstupní datový soubor ve formátu SAS namísto načtení externího textového souboru vytváří na základě existujícího datového souboru ve formátu SAS (pomocí příkazu SET, MERGE nebo UPDATE). V uvedeném případě DATA krok nastaví každou proměnnou v LPDV jako chybějící pouze před první iterací (před načtením prvního pozorování). Proměnné si uchovávají své hodnoty v LPDV, dokud nejsou nahrazeny novými hodnotami z načítaného vstupního datového souboru SAS. Proto tyto proměnné existují jak ve vstupním, tak i ve výstupním datovém souboru. Pouze v případě, že při vytváření nového datového souboru na základě existujícího datového souboru SAS bude DATA krok vytvářet nové proměnné na základě již existujících proměnných, jen tyto nové proměnné budou v LPDV na začátku každé iterace provádění nastaveny jako chybějící.

Při zpracování datového souboru ve formátu SAS je DATA krok příkazem SET výrazně efektivnější než při načítání dat z externích datových souborů pomocí jakéhokoliv tvaru příkazu INPUT. Navíc uložení dat do datových souborů nabízí další výhody jako například efektivní využití všech programovacích možností DATA kroku a přímé zpracování dalšími DATA kroky a procedurami programu. V další části článku již jako vstupní soubory budou použity datové soubory ve formátu SAS (načítané příkazem SET).

2.7 Výpočet počtu pozorování a průměrné hodnoty skupiny s využitím automatických proměnných

Nejprve je nutné načítaný vstupní datový soubor uspořádat podle hodnot proměnné, resp. proměnných, na základě kterých se budou vytvářet skupiny. Aby bylo možné použít zpracování po skupinách, je třeba za příkaz SET umístit příkaz BY s jednou nebo více proměnnými BY (podle hodnot proměnné, resp. proměnných se vytvoří skupiny).

Při zpracování skupin vytvoří DATA krok podle příkazu BY pro každou proměnnou v příkaze BY dvě dočasné indikátorové proměnné: FIRST.VARIABLE a LAST.VARIABLE. Proměnná VARIABLE označuje proměnnou, podle které DATA krok vytvoří skupiny pozorování (jsou tedy uvedeny v příkaze BY). Jelikož FIRST.VARIABLE a LAST.VARIABLE jsou dočasné proměnné, nejsou odesílány do výstupní datové sady. Obě proměnné FIRST.VARIABLE a LAST.VARIABLE jsou na začátku provádění kroku DATA inicializovány na hodnotu 1.

V okamžiku, kdy DATA krok načítá první pozorování v každé skupině vytvořené příkazem BY, je v LPDV dočasná proměnná FIRST.VARIABLE nastavena na 1. Při načítání druhého až posledního pozorování v každé skupině podle příkazu BY je tato proměnná nastavena na 0. Dočasná proměnná LAST.VARIABLE se DATA krokem nastaví na 1 při načítání posledního pozorování v každé skupině vymezené příkazem BY a na 0 při načítání těch pozorování, která nejsou posledními ve vymezených skupinách.

Příkladem využití automatických (dočasných) proměnných FIRST.VARIABLE a LAST.VARIABLE může být výpočet skupinového počtu a průměrné hmotnosti ve skupinách A a B. K výpočtům agregovaných hodnot se použije datový soubor ve formátu SAS, který byl vytvořen jako výsledek operací DATA kroku v programu v podkapitole 2.5. Datový soubor ve formátu SAS označený jako ex3_2 obsahuje 6 proměnných se 3 pozorováními. Struktura použitého datového souboru je znázorněna rabulkou 5. Každé pozorování je na základě proměnné GROUP rozděleno do skupiny A anebo B. Výpočet počtu pozorování a průměrné hmotnosti ve skupinách A anebo B se vykoná následujícím programem:

```
data ex3_4_mean (keep = group n total_weight mean_weight);
     set ex3_4;
     by group;
     BMI = 700*weight/(height*height);
     if first.group then
          do;
               total_weight = 0;
               n = 0;
          end;
     total_weight + weight;
     n + 1;
     if last.group then
          do;
               mean_weight = total_weight/n;
               output;
          end;
run;
```

Výsledkem fungování tohoto programu je tabulka skupinových četností a průměrné hmotnosti pro skupiny A anebo B, která je znázorněna v tabulce 6.

Tab. 6 Soubor s počtem a průměrnou hmotností pro skupiny A a B *(Zdroj: vlastní zpracování)*

GROUP	TOTAL_WEIGHT	N	MEAN_WEIGHT
A	61	1	61
B	267	2	133.5

Příkazem BY s proměnnou GROUP (za příkazem SET) v DATA kroku jsou v LPDV vytvořeny dvě automatické proměnné FIRST. GROUP a LAST. GROUP, které jsou v první iteraci nastaveny na hodnotu 1. Při provádění příkazu SET zkopíruje DATA krok první pozorování z uspořádaného datového souboru ex3_2 do LPDV. Protože se jedná o první pozorování pro skupinu A, je dočasná proměnná FIRST.GROUP nastavena na 1. Proměnná LAST.GROUP je nastavena také na 1, protože se jedná současně také o poslední pozorování. Dále je proměnným TOTAL_WEIGHT a N přiřazena hodnota 0, protože se jedná o první pozorování pro skupinu A (příkaz SAS: „if first.group then total_weight = 0;“). Poté příkaz „total_weight + weight“, resp. „n + 1“ kumuluje proměnnou TOTAL_WEIGHT, resp. N. Příkaz „if last.group“ je vyhodnocen jako pravdivý (LAST.GROUP se rovná 1, ve skupině A je pouze jediné pozorování) a obsah LPDV je odeslán do výstupní datové sady a hodnoty proměnných N a MEAN_WEIGHT jsou zapsány do výstupního souboru. Druhá iterace má průběh podobný jako iterace předchozí. Proměnná FIRST.GROUP je nastavena na 1 (jedná se o první pozorování ve skupině B). Splněním podmínek příkazu „if first.group“ se proměnné TOTAL_WEIGHT a N přiřadí hodnota 0, protože se jedná o první pozorování pro skupinu B. Opět dochází ke kumulování hodnot v proměnných TOTAL_WEIGHT a N – nyní však pro skupinu B. Třetí iterace zpracovává poslední pozorování ze skupiny B. Příkaz „if last.group“ je vyhodnocen jako pravdivý (LAST.GROUP se rovná 1, protože se jedná o poslední pozorování ze skupiny B). Obsah LPDV je odeslán do výstupní datové sady a hodnoty proměnných N a MEAN_WEIGHT pro skupinu B jsou zapsány do výstupního souboru. Příkazem KEEP se do výstupního souboru přenesou jen proměnné s požadovanými skupinovými hodnotami. Tím způsobem DATA krok vypočítá požadované skupinové hodnoty.

2.8 Rozdělení vstupního datového souboru do více výstupních souborů

Prostřednictvím DATA kroku je možné také i nasměrovat konkrétní pozorování na základě požadované podmínky do specifického datového souboru. Takový požadavek je v praxi velmi častý. Například, aby pozorování ze skupiny A byla směrována do samostatného výstupního datového souboru, a tak podobně i pro pozorování ze skupiny B. Samozřejmě, že tento požadavek je možné zajistit použitím více DATA kroků, což však přináší požadavky na větší výpočetní kapacity. Z hlediska efektivního programování toto rozdělení vstupní souboru na více souborů výstupních lze velmi efektivně provést pouze jediným DATA krokem.

K rozdělení vstupního souboru na více výstupních při zpracování jediným DATA krokem je vhodné použít explicitní příkaz OUTPUT. K rozdělení vstupního souboru do více výstupních souborů se použije tzv. podmíněný explicitní příkaz OUTPUT. Takový způsob zpracování vstupního datového souboru DATA krok se

nazývá tzv. podmíněným zpracováním DATA krokem (Harrington, 2018). Ukázka podmíněného zpracování souboru DATA krokem je v následujícím programu:

```
data ex3_4_GroupA ex3_4_GroupB;
     set ex3_4(drop = BMI total_weight);
     if group='A' then output ex3_4_GroupA;
     if group='B' then output ex3_4_GroupB;
run;
```

Vstupem do výše uvedeného DATA kroku je vstupní datový soubor ex3_4 s proměnnou GROUP, ve které je uložena informace o příslušnosti jednotlivých pozorování do skupiny A anebo B. Rozdělení vstupního datového souboru se všemi pozorováními do 2 výstupních datových souborů na základě příslušnosti ke skupině A nebo B zajistí podmíněné explicitní příkazy OUTPUT. Patří-li pozorování načtené do LPDV do skupiny A, bude nasměrováno do výstupního souboru ex3_4_GroupA. Patří-li toto pozorování do skupiny B, příkazem OUTPUT se uloží do výstupního datového souboru ex3_4_GroupB. Příkazem DROP jsou ze zpracování DATA kroku vyloučeny pro zpracování nepotřebné proměnné BMI a TOTAL_WEIGHT. Oba výstupní soubory jsou popsány tabulkou 7.

Tab. 7 Rozdělené výstupní datové soubory pro skupiny A a B *(Zdroj:* vlastní zpracování*)*

Soubor ex3_4_GroupA

NAME	GROUP	WEIGHT	HEIGHT
Barbara	A	61	.

Soubor ex3_4_GroupB

NAME	GROUP	WEIGHT	HEIGHT
John	B	62	175
Alex	B	72	165

3 Závěr

Pro vytváření datových souborů, jejich manipulaci s nimi a úspěšnou tvorbu programů v systému SAS je používání DATA krok jednou ze základních programovacích technik. Také některé úlohy související se zpracováním dat lze v systému SAS efektivně provádět pouze prostřednictvím DATA kroku (buď je nelze jiným způsobem provést anebo jsou příliš složité na to, aby je bylo možné realizovat jinými metodami v systému SAS). I když může být zpracování dat v DATA kroku složité, řídí se jasnými pravidly, jejichž zvládnutí umožní převzít kontrolu nad programy a zpracovávanými daty. Článek popisuje některá z těchto pravidel a uvádí příklady, které pomohou lépe pochopit, jak a proč se DATA krok chová a proč způsobuje konkrétní výsledky. Mnoho dalších užitečných poznatků o činnosti DATA kroku jsou obsahem různých odborných dokumentů a učebnic. Další důležité informace nabízí originální dokumentace o principech a činnostech programového systému SAS. Klíčové koncepty a principy, které jsou zde zmíněny, vytvářejí možnosti, jak přimět programový systém SAS, aby v rámci zpracování dat DATA krok udělal přesně to, co je od něj očekáváno.

4 Literatura

Harrington, T. J (2018). Discover the deeper DATA step. In *PharmaSUG 2018*, Paper: EP-16. Cary (NC): SAS Institute, 2018.

Howard, N. (2004). How SAS® thinks… or why the DATA step does what it does. In *NESUG 2004,* Paper: pm20-2004. Cary (NC): SAS Institute, 2004.

Li, A. (2013). *Handbook of SAS DATA step programming.* Boca Raton (FL): CRC Press, 2013. ISBN 978-1-4665-5239-5.

FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 2 pp 52-61

Výsledky sčítaní obyvateľov a historický vývoj časovania sobášnosti na národnej a regionálnej úrovni na Slovensku
Census results and historical development of marriage timing at the national and regional level in Slovakia

Branislav Šprocha[ab], Pavol Tišliar[b]

[a] Centrum spoločenských a psychologických vied Slovenskej akadémie vied, Šancova 56, 811 05 Bratislava, Slovenská republika

[a] Center for Social and Psychological studies of the Slovak Academy of Sciences, Šancova 56, 811 05 Bratislava, Slovak Republic

[b] Univerzita Cyrila a Metoda v Trnave, Filozofická fakulta, Nám. J. Herdu 2, 917 01 Trnava, Slovenská republika

[b] University of Ss. Cyril and Methodius in Trnava, Faculty of Arts, Nám. J. Herdu 2, 917 01 Trnava, Slovak Republic

branislav.sprocha@gmail.com, pavol.tisliar@ucm.sk

Abstrakt: *Demografická štatistika na území Slovenska má bohatú historickú tradíciu a dlhodobo poskytuje kvalitné údaje rôzneho charakteru. Vďaka tomu umožňuje aplikovať celú škálu rôznych nástrojov a konštruovať viaceré indikátory prezentujúce charakteristiky jednotlivých demografických procesov. Cieľom príspevku je ukázať ako je možné využiť výsledky sčítaní obyvateľov pri analýze časovania sobášnosti. Ide o nenahraditeľný zdroj údajov nielen pre historickú analýzu, ale aj identifikáciu regionálnych diferencií a ich prípadných zmien v čase.*

Abstract: *Demographic statistics in Slovakia has a rich historical tradition and has long provided quality data of various kinds. Thanks to this, it allows to apply a whole range of different tools and to construct several indicators presenting the characteristics of individual demographic processes. The aim of the paper is to show how it is possible to use the results of censuses in the analysis of marriage timing. It is an irreplaceable source of data not only for historical analysis, but also for the identification of regional differences and their possible changes over time.*

Kľúčové slová: *sobášnosť, časovanie, sčítanie, Slovensko, regióny*

Key words: *nuptiality, timing, census, Slovakia, regions*

1 Úvod

Časovanie sobášnosti je v dátových podmienkach na Slovensku na národnej, ako aj regionálnej úrovni prezentované prostredníctvom rôznych prierezových ukazovateľov. Asi najčastejšie sú na tieto účely využívané priemerné veky mužov alebo žien pri vstupe do manželstva alebo častejšie, ak pracujeme len so slobodnými osobami, pri vstupe do prvého manželstva. Uvedené ukazovatele sú v najčistejšej podobe odvádzané z vekovo-špecifických mier, ktorých výpočet je nenáročný na vstupné údaje. Konkrétne ide o triedenie počtu sobášov mužov a žien podľa veku a rodinného stavu. Exponovanou populáciou vstupujúcou do

menovateľa je vekové zloženie mužov a žien, pričom najčastejšie sa pracuje s vekom 16 – 49 rokov s ohľadom na konvenčné vymedzenie reprodukčného obdobia.

Pokročilejšie a súčasne však treba doplniť aj menej často využívané analytické nástroje sa opierajú o modely prežívania a to v podobe tabuliek života. Konkrétne ide najmä o tabuľky sobášnosti slobodných osôb. Konštrukčné mechanizmy sú však už ďaleko náročnejšie a aj údaje vstupujúce do výpočtov vyžadujú podrobnejšie triedenia, ktoré sú dostupné len za krátke obdobie a najčastejšie len na národnej úrovni (pozri napr. Rychtaříková, 1984; Pavlík et al., 1986).

V prípade, že je našou snahou analýza dlhodobého vývoja časovania sobášnosti, nemusí byť vhodným ani jeden z uvedených prístupov. Problémom je predovšetkým nedostupnosť potrebných údajov predovšetkým pre najstaršie obdobie existencie modernej demografickej štatistiky na Slovensku (koniec 19. a začiatok 20. storočia). Určitý kvalitatívny posun priniesol až vznik Československa a najmä zavedenie novej formy zberu, spracovania a publikovania demografických údajov v rámci Pohybov obyvateľstva od roku 1925 (Šprocha a Tišliar, 2009).

Okrem evidencie prirodzeného pohybu obyvateľstva však demografická štatistika disponuje aj vyčerpávajúcim zisťovaním v podobe sčítania obyvateľov. V modernej podobe sa na území Slovenska realizuje približne v 10-ročnej periodicite od roku 1869, pričom jeho integrálnou súčasťou sú tiež údaje týkajúce sa veku, pohlavia a rodinného stavu sčítaných osôb. Tie predstavujú dôležitý vstup pre konštrukciu špecifického indikátora časovania sobášnosti. Konkrétne ide o tzv. SMAM (*singulate mean age at marriage)*, ktorý udáva počet rokov, ktoré prežije muž alebo žena ako slobodná. Najčastejšie sa pritom pracuje s hornou hranicou 50 rokov, ktorá konvenčne v demografii vymedzuje koniec reprodukčného obdobia. Uvedený indikátor tak vo svojej podstate predstavuje analógiu priemernému veku pri prvom sobáši u osôb, ktoré vstúpili do manželstva.

Cieľom predloženého príspevku je predstaviť predmetnú metodickú konštrukciu a prakticky ju aplikovať na výsledkoch sčítania obyvateľov na území Slovenska od roku 1900 do roku 2011. Vďaka tomu získavame obraz o dlhodobom vývoji časovania procesu sobášnosti, ktorý iné ukazovatele v takomto časovom vymedzení nemôžu poskytnúť. Okrem toho sa tiež budeme snažiť ukázať ako je možné uvedený indikátor využiť pri priestorových analýzach a identifikácii regionálnych diferencií v časovaní sobášnosti na základe historických údajov na príklade uhorského sčítania z roku 1900 na župách ležiacich, alebo zasahujúcich vo väčšej miere na územie Slovenska.

2 Metodika konštrukcie SMAM

Základom pre konštrukciu indikátora SMAM (bližšie napr. United Nations, 1983) je štruktúra osôb podľa veku, pohlavia a rodinného stavu. Vychádzame pritom z celkového počtu osôb triedených podľa vekových skupín (napr. 1-ročné, 5-ročné a vo všeobecnosti x-ročné) a pohlavia. Druhým primárnym vstupom sú počty slobodných (*never-married*) osôb rovnako triedené podľa pohlavia do príslušných vekových skupín. Praktickú konštrukciu indikátora SMAM (zvlášť pre mužov a ženy) môžeme prezentovať prostredníctvom nasledujúcich na seba nadväzujúcich krokov:

1) Vypočítame podiel slobodných osôb $\boldsymbol{p_x^s}$ pre každú vekovú skupinu,

$$\boldsymbol{p_x^s = \frac{P_x^s}{P_x^{spolu}}}, \qquad (\mathbf{1})$$

kde $\boldsymbol{P_x^s}$ je počet slobodných osôb vo veku $\boldsymbol{x}$ a $\boldsymbol{P_x^{spolu}}$ je počet všetkých osôb vo veku $\boldsymbol{x}$.

2) Vypočítame celkový počet tzv. človekorokov prežitých v slobodnom stave. Ten vychádza z podielu slobodných a šírky vekového intervalu. V prípade, že šírka vekového intervalu je jeden rok, potom počet rokov, ktoré v ňom prežije osoba ako slobodná je rovný podielu slobodných v tomto veku. Predpokladáme rovnomerné rozloženie udalostí počas kalendárneho roka, preto z celkového časového úseku daného intervalu v stave slobodný-slobodná prežijú tú časť, ktorá zodpovedá podielu slobodných. Môžeme zapísať:

$$\boldsymbol{L_x^s = p_x^s \cdot a_x}, \qquad (\mathbf{2})$$

kde $\boldsymbol{L_x^s}$ označuje počet človekorokov, ktoré osoba vo veku $\boldsymbol{x}$ prežije ako slobodná, $\boldsymbol{p_x^s}$ je podiel slobodných osôb vo veku $\boldsymbol{x}$ a $\boldsymbol{a_x}$ je šírka vekového intervalu $\boldsymbol{x}$.

Pre celkový počet človekorokov, ktoré prežije osoba ako slobodná medzi vekmi $\boldsymbol{x_{min}}$ min a $\boldsymbol{x_{max}}$ potom platí:

$$\boldsymbol{L_{x_{min} \to x_{max}}^s = \sum_{x_{min}}^{x_{max}} L_x^s}. \qquad (\mathbf{3})$$

V prípade sobášnosti slobodných má zmysel uvažovať o x_{min} v podobe 16 rokov a x_{max} stanoviť na vek 50 rokov. V prípade x_{min}=16 rokov je potrebné ešte k výrazu pripočítať 16 rokov, ktoré osoba prežije ako slobodná od narodenia do dovŕšenia 16. roku života:

$$L_{16 \to 50}^s = 16 + \sum_{x=16}^{50} L_x^s. \qquad (4)$$

3) Vypočítame podiel aspoň raz ženatých / vydatých a trvalo slobodných. Podiel trvalo slobodných predstavujú osoby, ktoré zostali slobodné na konci sledovaného intervalu - ak budeme analyzovať sobášnosť v reprodukčnom veku (16 – 49 rokov), potom trvalo slobodné osoby budú muži a ženy vo veku 50 rokov. To platí, ak máme k dispozícii jednoročné vekové skupiny. V prípade 5-ročných, resp. $\boldsymbol{x}$ -ročných je potrebné tento podiel odhadnúť jednoduchou interpoláciou:

$$\boldsymbol{p_{50}^{s}} = \frac{(\boldsymbol{s_{45-49}^{s}} + \boldsymbol{s_{50-54}^{s}})}{\mathbf{2}}. \qquad (\mathbf{5})$$

Následne podiel aspoň raz ženatých mužov resp. vydatých žien $\boldsymbol{p_{50}^{ž,v}}$ je možné vyjadriť takto:

$$\boldsymbol{p_{50}^{ž,v}} = \mathbf{1} - \boldsymbol{p_{50}^{s}}. \qquad (\mathbf{6})$$

4) Výpočítame počet človekorokov prežitých trvalo slobodnými osobami v nadväznosti na predchádzajúce kroky, pričom v prípade horného intervalu $\boldsymbol{x_{max}}$ = 49 rokov platí:

$$\boldsymbol{L_{50}^{s}} = \mathbf{50} \cdot \boldsymbol{p_{50}^{s}}. \qquad (\mathbf{7})$$

5) Výslednú konštrukciu indikátora SMAM (pre interval 16 – 49 rokov) je potom možné zapísať v tvare:

$$\boldsymbol{SMAM_{16}^{49}} = \frac{(\mathbf{16} + \boldsymbol{L_{16\to 49}^{s}} - \boldsymbol{L_{50}^{s}})}{\boldsymbol{p_{50}^{ž,v}}} = \frac{(\mathbf{16} + \sum_{16}^{49} \boldsymbol{s_x^s} \cdot \boldsymbol{a_x} - \mathbf{50} \cdot \boldsymbol{p_{50}^{s}})}{\mathbf{1} - \boldsymbol{p_{50}^{s}}}. \qquad (\mathbf{8})$$

3 Výsledky

Dostupnosť údajov o štruktúre populácie podľa pohlavia, veku a rodinného stavu, ako aj štruktúre žien podľa veku a počte narodených detí v dlhých časových radoch, a to pre národnú, a čiastočne aj regionálnu úroveň výraznou mierou rozširuje možnosti analýz historického vývoja časovania procesov sobášnosti a plodnosti.

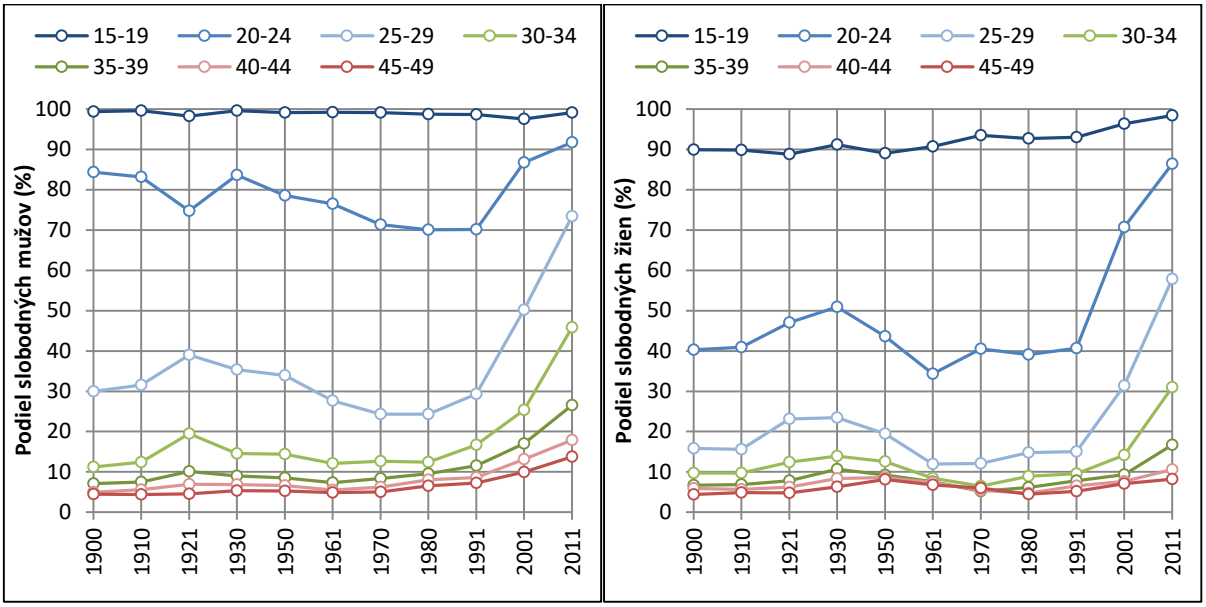
Dlhodobý vývoj podielu slobodných mužov a žien na území Slovenska v 5-ročných vekových skupinách reprodukčného veku prezentované v grafe 1 a 2 poukazujú na viaceré dôležité znaky a tiež zmeny v sobášnom správaní od začiatku 20. storočia.

V mužskej časti populácie môžeme vidieť, že v najmladšom veku do 20 rokov v podstate bez ohľadu na obdobie platilo, že do manželstva vstúpila len minimálna časť z nich. Rovnako vo veku 20 – 24 rokov prekračoval podiel slobodných v najstarších dobách takmer stabilne hranicu 80 %. Výnimkou bolo len povojnová kompenzačná fáza, kedy podiel slobodných mierne klesol pod

uvedenú úroveň. Po druhej svetovej vojne však v dôsledku známej koncentrácie sobášnosti slobodných do mladšieho veku (pozri napr. Šprocha a Tišliar, 2018) dochádza aj v tejto vekovej skupine k významnému poklesu, ktorý vrcholil na začiatku 80. a 90. rokov minulého storočia, keď v predmetnej vekovej skupine bolo slobodných približne 70 % všetkých mužov. Aj napriek tomu je však z grafu 1 zrejmé, že v prvej polovici reprodukčného veku vstupovala do manželstva väčšina slobodných mužov. Vo veku 30 – 34 rokov zostávalo slobodných dlhodobo len niečo viac ako 10 % osôb, pričom smerom k vyšším vekovým skupinám sa ich podiel už menil len minimálne. To svedčí tiež o výraznej vekovej koncentrácii sobášnosti. Na konci reprodukčného obdobia sa v dôsledku toho podiel slobodných pohyboval od začiatku 20. storočia na úrovni 4 – 5 % a až od 70. rokov začal mierne rásť.

U žien sa vo veku do 20 rokov dlhodobo vydávala približne desatina, no približne od 60. rokov začal podiel slobodných v tomto veku mierne rásť. Z grafu 2 je jednoznačne zrejmé, že hlavnú váhu na celkovej sobášnosti môžeme nájsť v nasledujúcej vekovej skupine, keďže už vo veku 25 – 29 zostávala na Slovensku slobodná menej ako pätina žien. Určitou výnimkou bolo len medzivojnové obdobie, kde v dôsledku prvej svetovej vojny vznikol v populácii významný nedostatok vekovo vhodných manželských partnerov. Rovnako ako u mužov sa v druhej polovici reprodukčného obdobia realizovala dlhodobo len veľmi malá časť z celkového počtu sobášov, čo sa prejavovalo na minimálnych zmenách v zastúpení slobodných žien v príslušných vekových skupinách. Vďaka vysokej koncentrácii a dlhodobo pretrvávajúcemu modelu skorej sobášnosti zostával podiel slobodných žien na konci reprodukčného obdobia na úrovni 5 – 8 %.

Politická, spoločenská, kultúrna a hospodárska transformácia, ku ktorej došlo po roku 1989 dramatickým spôsobom ovplyvnila celé rodinné a reprodukčné správanie mladých generácií mužov a žien na Slovensku (bližšie napr. Potančoková 2008; Potančoková et al., 2008; Šprocha a Tišliar, 2016). K ich hlavným znakom patrí predovšetkým výrazné odkladanie a pravdepodobne čiastočne aj odmietanie manželstva, čo sa odráža na pomerne rýchlo rastúcich podieloch slobodných mužov a žien. Ide najmä o vekové skupiny 20 – 24 a 25 – 29, ktoré z historického hľadiska tvorili základ pre realizáciu manželských štartov v populácii Slovenska. Podľa výsledkov zo sčítania 2011 podiel slobodných mužov vo veku 25 – 29 rokov výrazne prekračuje hranicu 70 % a u žien dosahuje takmer 60 %. Postupne sa však zvyšuje aj zastúpenie slobodných vo veku 30 – 39 rokov, čo sa bude odzrkadľovať následne v ďalších sčítaniach aj na rastúcom podiely osôb bez skúseností so životom v manželskom zväzku počas celého reprodukčného obdobia.

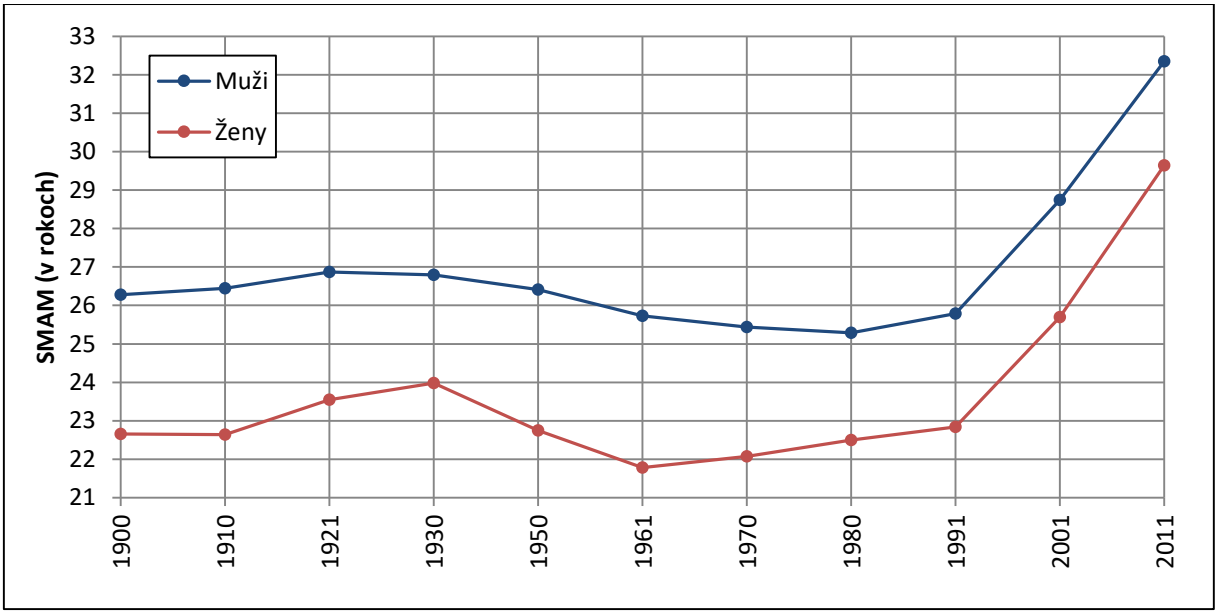


Graf 1 a 2 Podiel slobodných mužov a žien vo vybraných vekových skupinách na Slovensku, sčítania obyvateľov 1900–2011
(*Zdroj: vlastné spracovanie údajov zo sčítaní obyvateľov 1900 – 2011*)

Skoré časovanie sobášnosti na Slovensku v dlhodobej perspektíve potvrdzujú napokon aj výsledky samotného ukazovateľa SMAM. Tie prezentuje graf 3 zvlášť pre mužov a ženy na Slovensku od sčítania 1900 po sčítanie 2011. Pred prvou svetovou vojnou sa priemerný počet rokov prežitých v stave slobodná pohyboval u žien tesne pod hranicou 23 rokov. U mužov boli hodnoty SMAM o niečo viac ako 3 roky vyššie, keďže prekračovali hranicu 26 rokov. Svetový konflikt ovplyvnil nielen intenzitu a časovanie sobášnosti počas svojho trvania (Srb, 2002; Šprocha a Tišliar, 2008), ale sa následne odzrkadlil aj v povojnovom vývoji. Jednak zamedzil vstupu viacerým mladým osobám do manželstva, ktoré sa tak mohli zosobášiť často až po skončení vojny, čím sa u mužov mierne a u žien pomerne výrazne zvýšili hodnoty SMAM. V prípade ženskej populácie išlo tiež o spomínaný vplyv značných nerovnomerností v počte vekovo vhodných manželských partnerov. Nielenže vojnový konflikt spôsobil značné zníženie ich počtu, ale na sobášny trh sa dostali viaceré vdovy, ktoré by za normálnych okolností už na ňom nefigurovali a nezmenšovali už takto okresaný kontingent potenciálnych manželov. SMAM sa u mužov dostal podľa výsledkov prvého Československého sčítania na hodnotu takmer 27 rokov, z ktorej v druhom cenze v roku 1930 mierne klesol. U žien naopak nerovnosti na sobášnom trhu prispievali k pokračujúcemu rastu SMAMu z 23,5 roka (sčítanie 1921) až na 24 rokov (sčítanie 1930).

Druhá svetová vojna také negatívne zásahy do populačného vývoja nepriniesla (s výnimkou posledného roku) a toto obdobie sa nieslo skôr v znamení určitého oživenia sobášnosti (Šprocha a Tišliar, 2018). V kombinácii s povojnovou

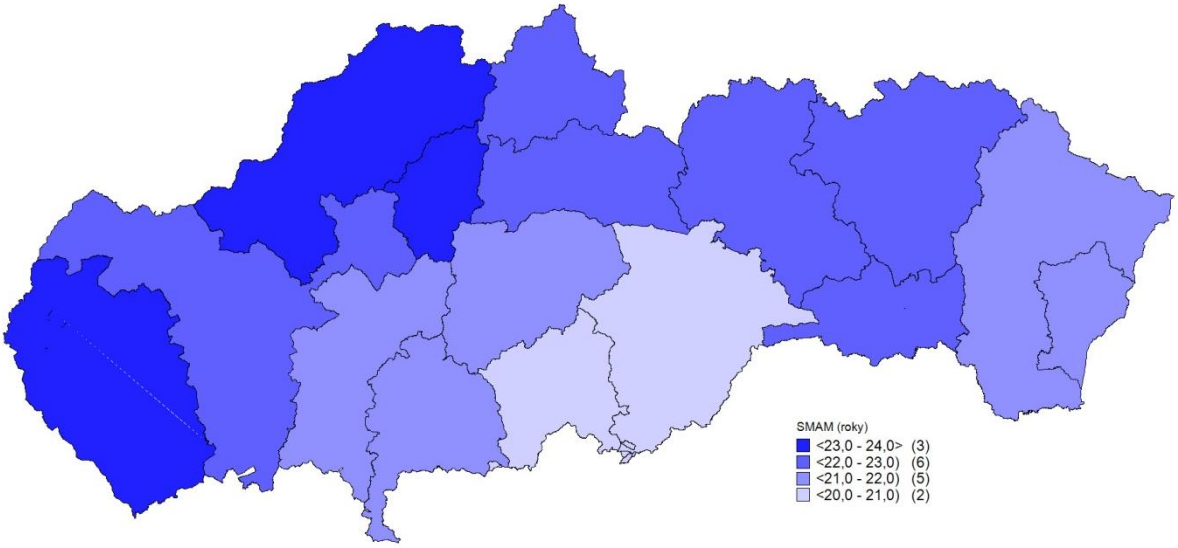
priaznivou populačnou klímou, ktorá sa prejavila v pomerne vysokej intenzite sobášnosti (Šprocha a Tišliar, 2018) sme svedkami pomerne významného poklesu hodnôt SMAM. K tomu prispievalo aj formovanie špecifických podmienok minulého politického režimu, ktoré nevytvárali v procese odkladania manželských a materských štartov (Sobotka, 2003, 2004) vhodný model životných dráh mladých mužov a žien. U mužov sa trend poklesu SMAM niesol v podstate až do roku 1980, keď z takmer 27 rokov klesol na niečo viac ako 25 rokov. V ženskej časti populácie bol pokles spočiatku veľmi dynamický. Medzi sčítaniami 1930 a 1961 sa hodnota SMAM dostala z približne 24 rokov pod hranicu 22 rokov. Nasledujúci vývoj sa však niesol vo veľmi miernom náraste, ktorý medzi sčítaniami 1961 a 1991 prispel k zvýšeniu SMAMu o 1 rok.



Graf 3 Singulate mean age at marriage (SMAM) mužov a žien na Slovensku, sčítania obyvateľov 1900 – 2011
(*Zdroj: vlastné spracovanie údajov zo sčítaní obyvateľov 1900 – 2011*)

Obdobie po roku 1989 sa však nesie v historicky bezprecedentnom a dynamicky sa presadzujúcom modely odkladania vstupov do manželstva čoho dôsledkom je nielen výrazne rastúci počet a podiel slobodných mužov a žien v reprodukčnom období, ale aj samotné rýchle intercenzálne zvyšovanie hodnôt SMAM. Platí to pre obe pohlavia, pričom o niečo dynamickejšie predsa len model skorého sobáša opúšťajú ženy. Kým v roku 1991 sa SMAM u mužov pohyboval na hranici 25,8 rokov, podľa údajov zo sčítania 2011 už dosahuje takmer 32,4 roku. U žien išlo v predmetnom intercenzálnom období o nárast z niečo viac ako 22,8 roka na takmer 29,7 roka. Bilančný vývoj štruktúry obyvateľstva Slovenska podľa veku, pohlavia a rodinného stavu ku koncu kalendárneho roka 2020 však jednoznačne poukazuje na ďalšie zvyšovanie hodnôt tohto ukazovateľa.

Na regionálnej úrovni je možné konštrukciu SMAM využiť nielen pre súčasné obdobie posledných sčítaní (1980 – 2011), za ktoré disponuje ŠÚ SR primárnou databázou a je teda možné v podstate si „vyskladať“ akýkoľvek regionálny celok v akomkoľvek administratívnom členení, ale potrebné vstupné údaje boli publikované aj na úrovni župného zriadenia z uhorských sčítaní ľudu z rokov 1880 – 1910. Ako príklad využitia metodického konštruktu SMAM sme zdigitalizovali a spracovali údaje za ženy z uhorského cenzu z roku 1900, ktorý prezentuje časovanie sobášnosti pred tým, ako tento proces zasiahla prvá svetová vojna.



Obr. 1 Singulate mean age at marriage (SMAM) v župách na Slovensku v roku 1900 (*Zdroj: vlastné spracovanie údajov zo sčítania ľudu 1900*)

Vyššie hodnoty SMAM dosahovali na začiatku 20. storočia 3 župy. Išlo jednak o Bratislavskú župu, kde môžeme predpokladať jednak vplyv mestského prostredia, ale aj väčšie zastúpenie nemeckého elementu, ako aj úzke väzby so západnou časťou monarchie, v ktorej vo všeobecnosti vstup do manželských zväzkov bol orientovaný do vyššieho veku. V prípade Trenčianskej župy môže byť vysvetlením úzke prepojenie na s českými krajinami, ktoré sa tiež vyznačovali neskorším uzatváraním manželstiev. Určitú úlohu môže zohrávať aj častejšia dočasná migrácia za prácou najmä mužského elementu, čo by mohlo odďaľovať sobáš. Turčianska župa je treťou, v ktorej sa hodnota SMAM dostala nad hranicu 23 rokov. Bližšie vysvetlenie tohto javu by sa mohlo skrývať v etnickej a náboženskej štruktúre. Na druhej strane stáli najmä župy na juhu stredného Slovenska, ako aj dve východoslovenské župy. Ukazuje sa, že aj v rámci populácií, ktoré v zmysle Hajnalovej klasifikácie (Hajnal, 1965) zaraďujeme medzi krajiny s tzv. neeurópskym typom sobášneho správania, existovali pomerne značné priestorové rozdiely v časovaní manželských štartov. Bez ďalšej hlbšej analýzy najmä z hľadiska populačných štruktúr a širšieho reprodukčného kontextu však uvedené diferencie nebude možné správne interpretovať.

4 Záver

Demografická štatistika na Slovensku dlhodobo poskytuje rôznorodé, kvalitné údaje, ktoré je možné využiť na konštrukciu celej rady indikátorov časovania jednotlivých demografických procesov. V príspevku sme sa zamerali na možnosti využitia výsledkov sčítaní obyvateľov na výpočet ukazovateľa časovania sobášnosti. Najmä v spojitosti s historickým vývojom tohto ukazovateľa existujú určité problémy s dostupnosťou vhodných vstupných údajov na výpočet konvenčne používaných indikátorov, napr. priemerného veku pri sobáši, priemerného veku pri prvom sobáši a pod. SMAM (*singulate mean age at marriage*) vo svojej povahe predstavuje určitú analógiu k týmto ukazovateľom časovania sobášnosti, ktorú je možné aplikovať na dostupné údaje v podstate už od sčítania 1880.

Na základe praktickej aplikácie výpočtu SMAM pre obdobie rokov 1900 – 2011 bolo možné jednoznačne identifikovať hlavné vývojové zmeny v načasovaní manželských vstupov mužov a žien na Slovensku v dlhodobom kontexte. Získané výsledky potvrdili prevládajúci model skorých sobášov u mužov a najmä žien už pred prvou svetovou vojnou, ako aj významný vplyv vojnového konfliktu na časovanie sobášnosti v medzivojnovom období. Rovnako získané údaje reflektovali zmeny v manželských štartoch, ktoré nastali počas minulého politického režimu a prispeli k prehĺbeniu skorého a takmer univerzálneho manželstva. Obdobie posledných troch desaťročí sa nesie v znamení dynamického odklonu od tohto správania, čo opätovne potvrdil aj samotný vývoj hodnôt SMAM.

V článku bola tiež prezentovaná aj konštrukcia SMAM na subnárodnej úrovni. Išlo o župy zasahujúce na územie Slovenska z uhorského sčítania 1910. Získané výsledky potvrdili existenciu pomerne významných regionálnych rozdielov v časovaní vstupov do manželstva. Tie budú jednak v širšom časovom a regionálnom rámci cieľom nášho ďalšieho výskumu so snahou tiež identifikovať širší kontext a možné vysvetľujúce faktory.

5 Literatúra

Hajnal, J. (1965). European marriage pattern in historical perspective. In Glass, D.V., Eversley, D. E. C. (eds.) *Population in history.* London: Arnold, s. 101–143.

Pavlík, Z., Rychtaříková, J., Šubrtová, A. (1986). *Základy demografie*. Praha: Academia.

Potančoková, M. (2008). *Plodnosť žien na Slovensku v období rokov 1950–2007 v generačnom pohľade*. Bratislava: INFOSTAT.

Potančoková, M. et al.(2008). Slovakia: Fertility between tradition and modernity. *Demographic Research,* 19, Special collection 7, 973-1018.

Rychtaříková, J. (1984). *Tabulky sňatečnosti a metody jejich konstrukce*. Demografie, 26, 110-122.

Sobotka, T. (2003). Re-emerging diversity: Rapid fertility changes in Central and Eastern Europe after the collapse of the Communist regimes. *Population* (English Edition), 58(4-5), 451-485.

Sobotka, T. (2004) *Postponement of childbearing and low fertility in Europe*. Groningen: Rijksuniversiteit Groningen.

Srb, V. (2002). *Obyvateľstvo Slovenska 1918 – 1938.* Bratislava: INFOSTAT.

Šprocha, B., Tišliar, P. (2008). *Náčrt vývoja sobášnosti na Slovensku v rokoch 1919 – 1937*. Bratislava: Stimul.

Šprocha, B., Tišliar, P. (2009). *Údaje o prirodzenej mene obyvateľstva Slovenska publikované v Pohyboch obyvatelstva Československa v rokoch 1919 – 1937*. Bratislava: INFOSTAT.

Šprocha, B., Tišliar, P. (2016). Transformácia plodnosti žien Slovenska v 20. a na začiatku 21. storočia. Bratislava: Centrum pre historickú demografiu a populačný vývoj Slovenska, Filozofická fakulta Univerzity Komenského v Bratislave.

Šprocha, B., Tišliar, P. (2018). *100 rokov obyvateľstva Slovenska: od vzniku Československa po súčasnosť*. Bratislava: Muzeológia a kultúrne dedičstvo.

United Nations (1983). Manual X. Indirect Techniques for Demographic Estimation. *Population Studies*, No. 81, New York: United Nations.

6 Poďakovanie

Článok bol vypracovaný za podpory grantovej agentúry APVV-20-0199 *Transformácia populačného vývoja na Slovensku v regionálnom pohľade od konca 19. do polovice 20. storočia.*

FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 2 pp 62-76

Zmeny plodnosti vydatých žien na Slovensku v časovej a priestorovej perspektíve
Changes in the fertility of married women in Slovakia in a temporal and spatial perspective

Branislav Šprocha[ab], Pavol Tišliar[b]

[a] Centrum spoločenských a psychologických vied Slovenskej akadémie vied, Šancova 56, 811 05 Bratislava, Slovenská republika

[a] Center for Social and Psychological studies of the Slovak Academy of Sciences, Šancova 56, 811 05 Bratislava, Slovak Republic

[b] Univerzita Cyrila a Metoda v Trnave, Filozofická fakulta, Nám. J. Herdu 2, 917 01 Trnava, Slovenská republika

[b] University of Ss. Cyril and Methodius in Trnava, Faculty of Arts, Nám. J. Herdu 2, 917 01 Trnava, Slovak Republic

branislav.sprocha@gmail.com, pavol.tisliar@ucm.sk

Abstrakt: *Plodnosť na Slovensku je aj napriek významným transformačným zmenám po roku 1989 stále v prevažnej miere viazaná na život v manželskom zväzku. Preto je pre pochopenie dlhodobých zmien v jej intenzite, charaktere a priestorových rozdielov nezastupiteľnou analýza plodnosti vydatých žien. Hlavným cieľom príspevku je analýza generačnej plodnosti vydatých žien na Slovensku, ich štruktúry podľa počtu narodených detí, a to v časovej a priestorovej perspektíve. Snažíme sa tak identifikovať nielen medzigeneračné zmeny, ale aj nájsť prípadné regionálne rozdiely v intenzite a paritnej štruktúre vydatých žien. Na tieto účely pracujeme s údajmi zo sčítaní obyvateľov z rokov 1930, 1961, 1991 a 2011, ktoré pokrývajú obdobie transformácie procesu plodnosti v rámci demografickej tranzície, ako aj súčasné prebiehajúce zmeny.*

Abstract: *Fertility in Slovakia, despite significant transformational changes after 1989, is still largely tied to living in a marriage. Therefore, an analysis of fertility of married women is indispensable for understanding long-term changes in its intensity, character and spatial differences. The main goal of the paper is to analyze the cohort fertility of married women in Slovakia, their parity structure, in a temporal and spatial perspective. We try to identify not only inter-cohort changes, but also to find possible regional differences in the intensity and parity structure of married women. For these purposes, we work with census data from 1930, 1961, 1991 and 2011, which cover the period of transformation of the fertility process within the first demographic transition, as well as current changes.*

Kľúčové slová: *plodnosť, vydaté ženy, generácie, okresy, Slovensko*

Key words: *fertility, married women, cohorts, districts, Slovakia*

1 Úvod

Rozvoľňovanie úzkeho prepojenia medzi sobášom, životom v manželskom zväzku a plodnosťou, ktoré identifikujeme na Slovensku v súčasnosti nie je typickou črtou reprodukčného správania a začalo sa formovať len relatívne

nedávno. Na druhej strane údaje z posledných rokov hovoria, že rast podielu detí narodených mimo manželstva sa zastavil a naopak postupne sa začína zvyšovať manželská plodnosť (Šprocha a Tišliar, 2021). Aj preto naďalej platí, že prevažná väčšina detí sa rodí vydatým matkám (približne 60 %), a ich reprodukčné správanie je stále kľúčovým faktorom pre celkový populačný vývoj Slovenska. Samotnej plodnosti vydatých žien je však venovaná pritom vo vedeckej obci pomerne malá pozornosť. Len sporadicky najmä v rámci diferenčných analýz (napr. Šprocha a Tišliar, 2016) sa pracuje aj s procesom rodenia detí v manželskom zväzku.

Dostupné údaje demografickej štatistiky na Slovensku pritom poskytujú niekoľko možných prístupov, a to v dvoch hlavných pohľadoch na čas. Jednak v prierezovom môžeme pracovať s mierami prvej kategórie (čistými mierami plodnosti), konštruovať tabuľky manželskej plodnosti, či hodnotiť proces v kontexte doby uplynulej od uzavretia manželstva (pozri napr. Pavlík et al. 1986). Z longitudinálneho pohľadu sú k dispozícii údaje zo sčítaní obyvateľov, v ktorých sa zisťuje nielen rok narodenia (resp. vek) a počet narodených detí, ale aj rodinný stav ženy. Tie predstavujú základ pre konštrukciu kohortných ukazovateľov plodnosti. Obrovskou výhodou je, že dané premenné sú integrálnou súčasťou sčítaní v podstate už od roku 1930. Tým sa vytvára pomerne široká dátová základňa pre dlhodobú analýzu transformačných zmien. Okrem toho publikované boli z rokov 1930 a 1961 aj údaje za nižšiu ako národnú úroveň, vďaka čomu sa naše analytické možnosti rozširujú smerom k regionálnemu pohľadu. Ani tento aspekt nie je v slovenskom prostredí dostatočne akcentovaný (výnimky napr. Bleha et al. 2014; Šprocha a Tišliar, 2016).

Vzhľadom na uvedené je hlavným cieľom príspevku analýza generačnej úrovne plodnosti vydatých žien na Slovensku a jej zmien v čase. Okrem toho sa zameriame aj na jej vnútorný charakter prostredníctvom vývoja štruktúry vydatých žien podľa počtu narodených detí. Ako už bolo spomenuté, dostupné údaje poskytujú tiež možnosť určitého pohľadu na priestorovú analýzu kohortnej plodnosti vydatých. Budeme sa tak snažiť identifikovať nielen jej vývoj v čase, ale aj existenciu prípadných regionálnych vzorcov s overením ich platnosti naprieč vybranými sčítaniami v 20. a na začiatku 21. storočia.

2 Zdroje údajov a metodika práce

V spojitosti s hlavným zameraním príspevku využívame celkovo výsledky štyroch sčítaní obyvateľov. Prvým je medzivojnové sčítanie ľudu z roku 1930, ktoré prvýkrát v histórii Česka a Slovenska poskytlo údaje potrebné pre kohortnú analýzu plodnosti vydatých žien. Táto premenná sa stala následne integrálnou

súčasťou všetkých nasledujúcich sčítaní s výnimkou roku 1940. Keďže sa snažíme podchytiť jednak zmeny podmienené demografickou revolúciou a tiež transformačnými posunmi posledných troch desaťročí, tomu sme prispôsobili aj výber jednotlivých sčítaní pre našu analýzu. Nemenej dôležitým je aj dostupnosť údajov na regionálnej úrovni. Keďže až od sčítania ľudu 1980 disponujeme anonymizovanou primárnou databázou výsledkov, sme nútení sa obmedziť len na publikované kombinačné tabuľky. Tie zo sčítania 1950 neobsahovali práve údaje z regionálnej úrovne, preto pracujeme až so sčítaním 1961. Určitou výhodou tejto voľby je tiež skutočnosť, že tento posun nám umožnil zohľadniť dlhšie časové obdobie pôsobenia transformačných zmien v rámci demografickej revolúcie, ktorej hlavným znakom okrem iného bolo práve vedomé obmedzovanie manželskej plodnosti (bližšie napr. Coale a Treadway, 1986). Navyše obdobie 50. a 60. rokov je niektorými autormi (Fialová et al., 1990; Vereš, 1986; Šprocha a Tišliar, 2016) vnímané ako časový rámec ukončovania transformačných procesov demografickej revolúcie na národnej i regionálnej úrovni. Rok 1991 predstavuje z hľadiska sčítaní obyvateľov zlomový míľnik na ceste medzi tzv. socialistickým modelom reprodukcie a nástupom dynamicky sa presadzujúcich a v mnohých aspektoch historicky jedinečných transformačných zmien posledných troch desaťročí. Vo svojej podstate tak odzrkadľuje štrukturálne znaky populácie podmienené predchádzajúcimi špecifickými podmienkami tzv. socialistického skleníka (Sobotka, 2011) a predstavuje komparačný rámec zo stavom, ktorý získavame z posledného dostupného sčítania obyvateľov z roku 2011. Ten je už do určitej miery najmä v prípade mladších generácií (bližšie k tejto problematike napr. Potančoková, 2008; Potančoková et al. 2008; Šprocha a Tišliar, 2016) už poznačený ďalšími transformačnými posunmi prebiehajúcimi v populácii Slovenska po roku 1989. V oboch prípadoch pritom dostupné údaje umožňujú konštruovať kohortné ukazovatele plodnosti vydatých žien nielen na národnej, ale aj regionálnej úrovni. Určitou limitáciou v spojitosti s výsledkami sčítaní 1930 a 1961 je odlišné administratívne členenie. Pre rok 1991 bol možný prevod na súčasný stav, no medzi rokmi 1930 a 1961 došlo k pomerne veľkým zmenám a tieto nedokážeme vzhľadom na dostupné údaje upraviť na jednotnú bázu. Preto vzájomné porovnanie výsledkov v čase na regionálnej úrovni je do určitej miery problematické.

Z hľadiska využitia kohortných indikátorov vychádzame predovšetkým z možností poskytovaných publikovanými údajmi zo starších sčítaní obyvateľov. Na národnej úrovni z vybraných cenzov konštruujeme kohortnú plodnosť žien pre jednotlivé skupiny generácií vydatých osôb (v 5-ročných skupinách populačných ročníkov). Tá vyjadruje priemerný počet detí, ktorý sa narodil (v tomto prípade vydatým) ženám do rozhodujúceho okamihu sčítania. Pre každú

generáciu vydatých žien (resp. skupinu generácií G) potom môžeme napísať, že jej kohortná plodnosť KP_G^v je sumou parciálnych konečných kohortných plodností vydatých podľa poradia:

$$KP_G^v = \sum_{i=1}^{i_{max}} KP_G^{i,v}. \quad (1)$$

Samotná kohortná plodnosť parity i je pritom definovaná ako pomer vydatých žien z príslušnej skupiny generácií G, ktorým sa narodilo i a viac detí k celkovému počtu vydatých žien skupiny generácií G:

$$KP_G^{i,v} = \frac{\sum_i^{i_{max}} P_G^{i,v}}{\sum P_G^v}, \quad (2)$$

pričom $\sum_i^{i_{max}} P_G^{i,v}$ je počet vydatých žien skupiny generácií G s počtom detí i a viac, $\sum P_G^v$ reprezentuje celkový počet vydatých žien v skupine generácií G.

Analogicky konštruujeme podiel vydatých žien podľa počtu narodených detí:

$$p_G^{i,v} = \frac{P_G^{i,v}}{\sum P_G^v} \cdot 100. \quad (3)$$

V analýze pritom pracujeme s paritou i = 0, 1, 2 a 3+ detí.

Z hľadiska vývojových zmien intenzity i paritnej štruktúry sú predovšetkým prezentované tzv. konečné úrovne, teda hodnoty za skupiny generácií vydatých žien, ktoré v čase sčítania obyvateľov boli vo veku ukončenej reprodukcie (50 a viac rokov). Na rozdiel od prierezových údajov a rovnako kohortných, ktoré je z nich možné vhodnými úpravami vytvoriť, sú získané výsledky z jednotlivých sčítaní v podstate obrazom poplatným k danému rozhodujúcemu okamihu. To vytvára určité limity pre samotnú analýzu. Predovšetkým v starších generáciách môže dôjsť k efektu selekcie, keďže nedokážeme zachytiť všetky osoby daných populačných ročníkov, ale len tie, ktoré „prežili“ do sčítania. Okrem toho určité problémy môžu byť spojené aj s rodinný stavom. Ten je opätovne poplatný k rozhodujúcemu dátumu sčítania, ale nemusí vypovedať nič o tom, či sa dieťa skutočne narodilo vydatej žene alebo žene v inom rodinnom stave. Výsledky tak len reflektujú odpovede o počte narodených detí žien žijúcich v manželskom zväzku v čase sčítania. Keďže na Slovensku dlhodobo prevažoval scenár, keď sa žena najprv vydala a až potom sa narodilo dieťa (pričom často práve jej tehotenstvo bolo impulzom na sobáš), v sledovaných skupinách generácií pravdepodobne nebude uvedená limita až tak markantne skresľovať získané výsledky.

V prípade analýzy regionálnej úrovne kohortnej plodnosti vydatých žien sú použité indikátory poplatné existujúcim vstupným údajom. Kým v prípade sčítaní obyvateľov 1991 a 2011 nie je problém konštruovať rovnaké ukazovatele prostredníctvom totožných metodických prístupov, ako na národnej úrovni, hlavné limity sa týkajú starších cenzov 1930 a 1961. V oboch prípadoch je však rozsah publikovaných údajov obmedzený na vybrané vekové skupiny, tak aby korešpondoval so staršími sčítaniami. Z hľadiska konečnej kohortnej plodnosti jednotlivých okresov Slovenska tak predmetná hodnota vyjadruje priemerný počet detí, ktorý uviedli vydaté ženy vo veku 50 a viac rokov.

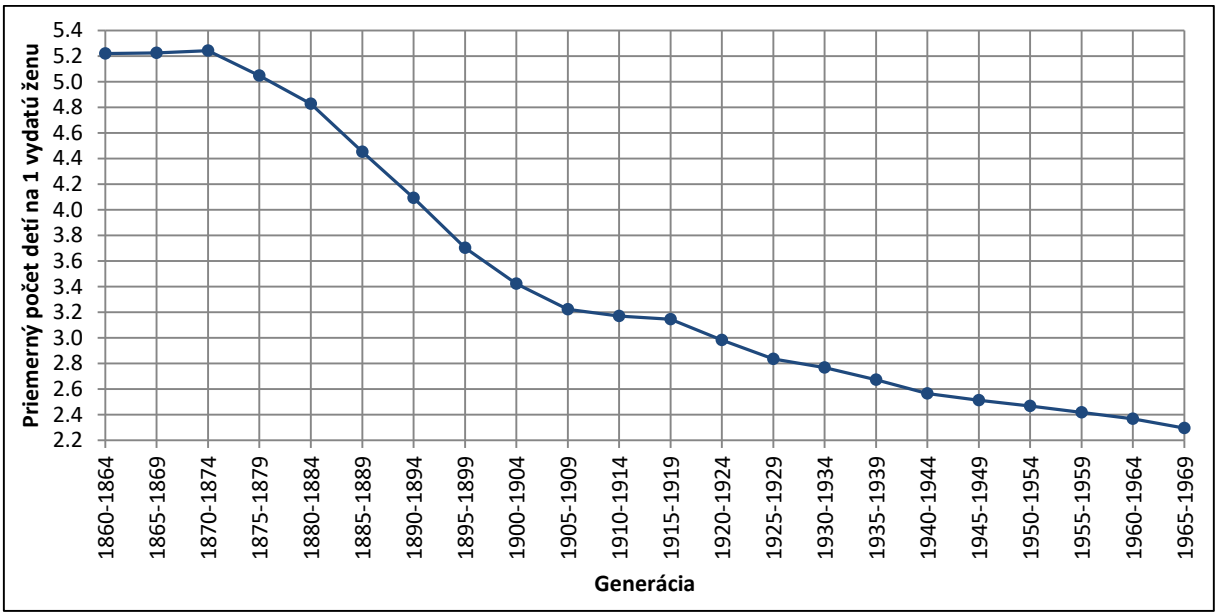
3 Konečná plodnosť a paritná štruktúra vydatých žien na Slovensku

Úzke spojenie manželstva a prokreatívneho správania na Slovensku bolo do značnej miery historicky podmienené a udržiavané prostredníctvom značného vplyvu katolíckej cirkvi a silných vnútorných kontrolných mechanizmoch miestnych (dlho prevažne rurálnych) komunít. Aj keď po druhej svetovej vojne vplyv týchto faktorov nesporne slabol, no tento efekt zdá sa dostatočne vyvážilo pôsobenie špecifických podmienok formujúcich sa počas minulého politického režimu. Preferencia manželskej rodiny tak nielenže pretrvala, ale zdá sa, že sa ešte viac upevnila. O tom svedčí aj skutočnosť, že kým počet a podiel predmanželských koncepcií rástol, podiel detí narodených mimo manželský zväzok dlhodobo dosahoval menej ako desatinu z celkového počtu pôrodov (Šprocha a Tišliar, 2018). Manželský zväzok tak predstavoval z normatívneho hľadiska jediný legitímny priestor pre rodenie detí, pričom nevydaté matky, ako aj ich potomstvo boli vnímané často značne negatívne (bližšie pozri napr. Botíková et al., 1997).

Veľmi dôležitým faktorom bol tiež pretrvávajúci charakter matrimoniálneho správania, ktorý sa obdobne ešte viac upevnil v období minulého politického režimu. Na základe historických analýz (napr. Hajnal, 1965, Livi-Bacii, 2003) môžeme priestor Slovenska zaradiť do skupiny krajín s tzv. neeurópskym typom sobášneho správania. Ten sa vyznačoval predovšetkým veľmi skorým vstupom do manželstva a takmer univerzálnou povahou týchto prechodov v životných dráhach mladých ľudí.

Obdobie po roku 1989 prinieslo významné politické, kultúrne, spoločenské, ekonomické zmeny, ktoré výrazným spôsobom ovplyvnili charakter reprodukčného správania. Viaceré analýzy (napr. Potančoková et al. 2008; Vaňo et al. 2001; Šprocha a Tišliar, 2018) potvrdili, že v nových životných podmienkach dochádza k viacerým, pomerne dynamicky sa presadzujúcim a v mnohých aspektoch historicky jedinečným zmenám prokreatívneho a matrimoniálneho správania. Jedným z nich je už viackrát spomínané rozvoľňovanie spojenia medzi

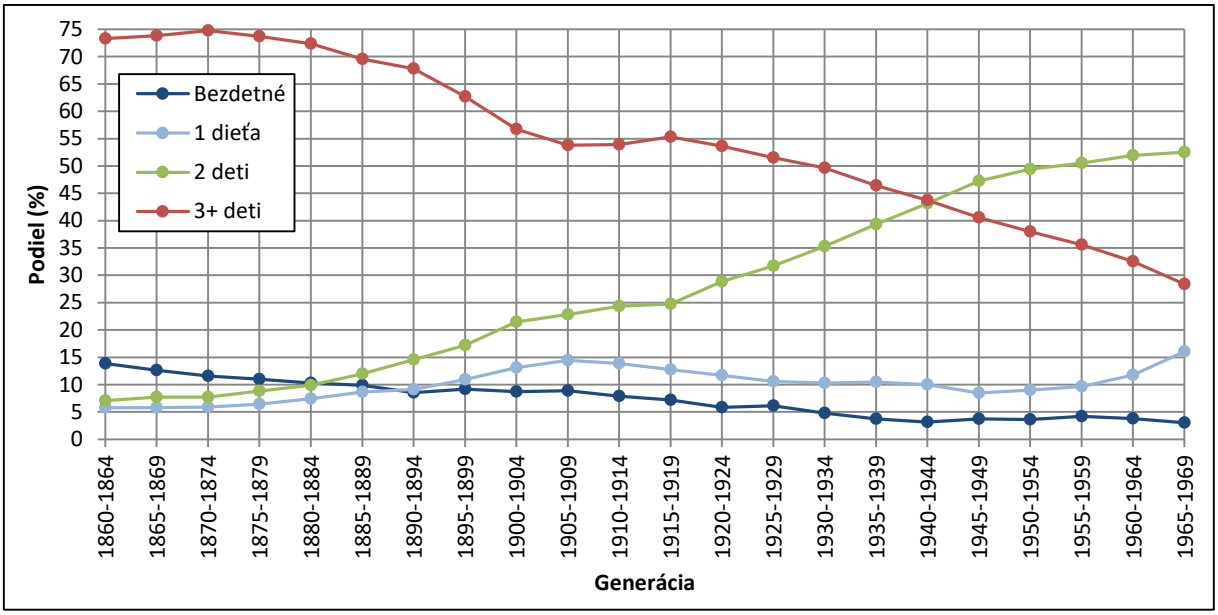
životom v manželstve a rodením detí a tiež znižovanie manželskej plodnosti (pozri napr. Šprocha a Tišliar, 2021). Aj keď v poslednom období identifikujeme v prípade manželskej plodnosti mierne oživenie (Šprocha a Tišliar, 2021), samotný pokles priemerného počtu detí narodených vydatým ženám nie je len doménou posledných troch desaťročí. Ako je možné vidieť z grafu 1, v najstarších generáciách zo 60. a prvej polovice 70. rokov 19. storočia sa konečná plodnosť pohybovala stabilne nad hranicou 5,2 dieťaťa na jednu vydatú ženu. Až u vydatých žien z druhej polovice 70. rokov môžeme identifikovať začiatok kontinuálneho poklesu realizovanej plodnosti. Ten bol pritom pomerne dynamický, keďže už u vydatých žien z prvej polovice 90. rokov 19. storočia klesla konečná plodnosť tesne nad hranicu 4 detí a v nasledujúcej mladšej skupine generácií už priemerný počet detí dosahoval len 3,7 dieťaťa. Práve v prípade žien narodených v druhej polovici 80. a v 90. rokoch bol pokles realizovanej plodnosti historicky najdynamickejší (o 0,37 – 0,39 dieťaťa medzi 5-ročnými za sebou nasledujúcimi populačnými skupinami). Z grafu 1 je tiež zrejmé, že po tomto pomerne rýchlom poklese došlo u vydatých žien narodených v rokoch 1905 – 1919 k medzigeneračnej stabilizácii konečnej plodnosti na hranici približne 3,2 dieťaťa. Tento trend bol do značnej miery pravdepodobne podmienený oživením reprodukcie počas druhej svetovej vojny a po jej skončení (Šprocha a Tišliar 2018).



Graf 2 Konečná plodnosť vydatých žien na Slovensku narodených v rokoch 1860 – 1969
(*Zdroj: vlastné spracovanie na základe údajov zo sčítaní obyvateľov 1930, 1961, 1991 a 2011*)

Ďalší kontinuálny pokles realizovanej plodnosti bol naštartovaný s kohortami žien narodenými v prvej polovici 20. rokov 20. storočia. V podstate tento trend pretrval až do najmladších sledovaných generácií z druhej polovice 60. rokov, pričom predbežné údaje za kohorty zo 70. rokov indikujú, že znižovanie konečnej plodnosti pokračuje ďalej. Tak napríklad z výsledkov sčítania obyvateľov 2011 je zatiaľ možné vidieť, že v generáciách 1970 – 1974 pripadalo v priemere na jednu vydatú ženu približne 2,1 dieťaťa a u žien narodených v rokoch 1975 – 1979 dokonca len okolo 1,8 dieťaťa. Kým v prvej menovanej skupine sa osoby v čase cenzu už nachádzali vekovo nad hranicou 35 rokov, v ktorej plodnosť pomerne prudko s vekom klesá, tak v druhej ide len o prvú polovicu 30. rokov, kde intenzita rodenia detí významne vzrástla. Preto najmä u žien z druhej polovice 70. rokov môžeme očakávať ešte určitý nárast kohortnej plodnosti. Jej definitívnu úroveň však v týchto transformačnými zmenami najviac zasiahnutých generáciách (pozri napr. Potančoková, 2008; Šprocha a Tišliar, 2018) prinesú výsledky nedávno ukončeného sčítania obyvateľov 2021. Ak sa vrátime k vývoju manželskej plodnosti medzi generáciami 1920 – 1969, je zrejmé, že predmetný pokles priemerného počtu narodených detí bol oveľa menej dynamický. V predmetných populačných skupinách klesla konečná plodnosť z hodnoty niečo menej ako 3,2 dieťaťa na jednu vydatú ženu na približne 2,3 dieťaťa. Vnútorné mechanizmy tohto vývoja môžeme detailne sledovať prostredníctvom medzigeneračných zmien v štruktúre vydatých žien podľa počtu narodených detí (graf 2). V najstarších analyzovaných generáciách mali jednoznačnú prevahu vydaté ženy s tromi a viac deťmi. Ich zastúpenie sa pohybovalo v intervale 70 – 75 % až do skupiny osôb narodených v druhej polovici 80. rokov 19. storočia. K viditeľnému poklesu, ktorý podmienil aj samotné znižovanie konečnej plodnosti však začalo dochádzať už v generáciách z konca 70. rokov. Predmetný vývoj sa následne pomerne rýchlo medzigeneračne presadil a k určitej dočasnej stabilizácii došlo až v prípade vyššie spomínaných skupín z rokov 1905 – 1919, keď vydaté ženy s tromi a viac deťmi predstavovali približne 55 % z celého populačného ročníka. Celkom opačný trend v týchto najstarších generáciách malo zastúpenie vydatých žien s dvomi deťmi. Ich podiel rástol z niečo približne 5 % na 25 %. Z grafu 2 je zrejmé, že v najstarších generáciách bol dvojdetný model rodiny na Slovensku v podstate rovnako marginálnym javom ako mať len jedno dieťa. Na druhej strane je zaujímavé, že pomerne nezanedbateľná časť vydatých žien zostávala celoživotne bezdetná. Aj keď váha postupne klesala, v generáciách 1860 – 1884 ich podiel stále prekračoval hranicu 10 % (10,3 – 13,9 %). Je problematické uspokojivo vysvetliť získané výsledky. Do úvahy padá jednak kvalita publikovaných údajov, či v metodickej kapitole spomínaná selekcia u najstarších generácií žien. Okrem toho je možné, že predmetné výsledky

skutočne zachytávajú realitu a poukazujú na častejšiu prítomnosť niektorých rizikových faktorov, ktoré viedli k celoživotnej bezdetnosti.



Graf 2 Štruktúra vydatých žien podľa počtu narodených detí na Slovensku v generáciách 1860 – 1969 (*Zdroj: vlastné spracovanie na základe údajov zo sčítaní obyvateľov 1930, 1961, 1991 a 2011*)

Jednodetnosť predstavovala u vydatých žien na Slovensku dlhodobo najmenej rozšírený model rodiny. Ich zastúpenie sa v najstarších generáciách pohybovalo na hranici 5 % a až so skupinami osôb narodených v druhej polovici 70. a najmä v 80. rokoch je spojený určitý medzigeneračný nárast. Ten však bol do značnej miery len pozvoľný s vrcholom u žien narodených od začiatku 20. storočia do vypuknutia prvej svetovej vojny (graf 2). Uvedený vývoj môžeme vysvetľovať jednak zhoršenými životnými podmienkami počas vojnového konfliktu a krízy v 30. rokoch, ako aj problematickou situáciou žien na sobášnom trhu v medzivojnovom období, keď vznikla značná disproporcia medzi vekovo vhodnými skupinami potenciálnych manželov a manželiek. To sa mohlo spoločne podpísať pod skutočnosť, že v predmetných generáciách jednodetnosť vrcholila až na úrovni takmer 15 %. Smerom k mladším generáciám však podobne ako u bezdetnosti aj zastúpenie žien len s jedným dieťaťom klesá. V prípade bezdetnosti najnižšie hodnoty identifikujeme u vydatých osôb narodených od druhej polovice 30. do konca 60. rokov, keď sa ich podiel pohyboval len na úrovni 3 – 4 %. Pomerne nízka bezdetnosť je zatiaľ signalizovaná aj v predbežných údajoch pre generácie z prvej polovice 70. rokov (niečo viac ako 4 %), no u mladších kohort už pomerne prudko rastie. V ich prípade je však potrebné si počkať na výsledky sčítania obyvateľov 2021. Nárast jednodetnosti je možné pozorovať už skôr, a to po dosiahnutí minimálnej úrovne v generáciách z druhej

polovice 40. a z 50. rokov (menej ako 10 %), keď sa zastúpenie vydatých žien s jedným dieťaťom dostalo v generáciách 1965 – 1969 až nad 15%. Aj v ich prípade predbežné údaje signalizujú ďalší pomerne rýchly medzigeneračný rast. Z uvedeného prehľadu je však zrejmé, že v minulosti (s výnimkou niektorých exponovaných generačných skupín) bola bezdetnosť u vydatých žien na Slovensku výrazne marginalizovaný jav a rovnako stať sa matkou len jedného dieťaťa bol pomerne málo rozšírený reprodukčný model.

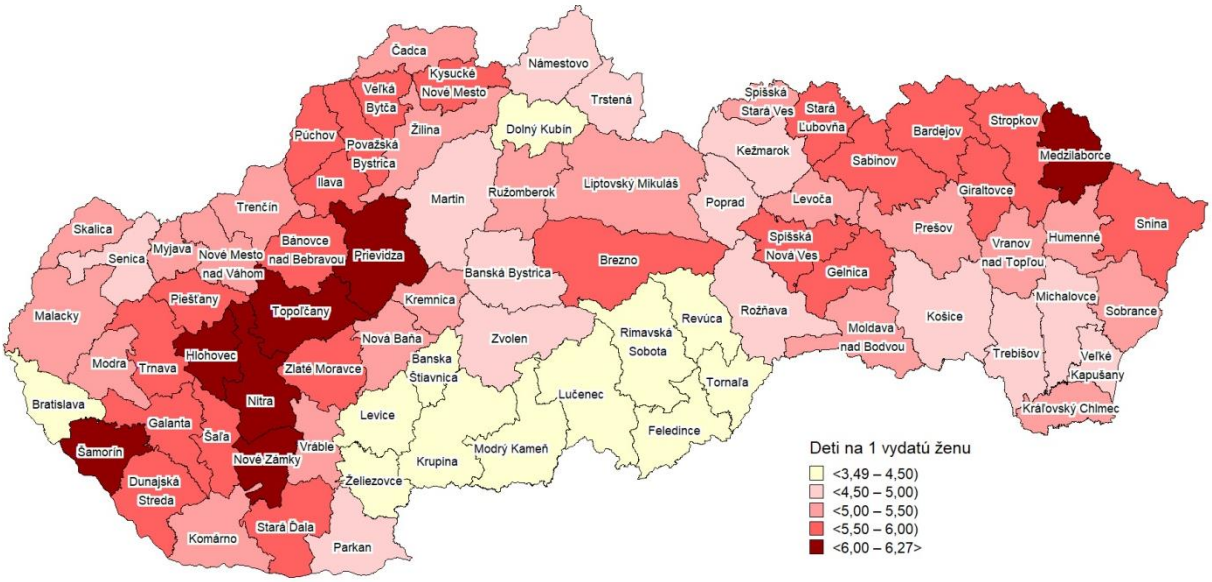
Kým bezdetnosť a jednodetnosť sa v sledovaných populačných ročníkoch menili len v obmedzenom rozsahu, v prípade zastúpenia žien s dvomi, tromi a viac deťmi môžeme identifikovať celkovo rozsiahle a pomerne dynamicky prebiehajúce transformačné posuny. Jednoznačne je pritom zrejmé, že ich hlavným znakom bol medzigeneračný nárast podielu vydatých žien s dvomi deťmi a naopak pokles zastúpenia s tromi a viac deťmi. Uvedený vývoj pritom v podstate bez prerušenia prebiehal od generácií 1915 – 1919 až po najmladšie populačné ročníky z druhej polovice 60. rokov. Kým v prípade podielu žien s tromi a viac deťmi išlo v podstate o rovnomerný pokles z 55 % pod hranicu 30 %, u dvojdetného modelu bola najvyššia dynamika nárastu medzi skupinami vydatých žien z 20. až druhej polovice 40. rokov. V predmetných generáciách z pôvodnej približne jednej štvrtiny vzrástla váha vydatých žien s dvomi deťmi až nad 50 % (graf 2). Smerom k mladším kohortám síce ďalej dochádzalo k zvyšovaniu zastúpenia, no ten už neprebiehal tak dynamicky, a preto v generáciách z druhej polovice 60. rokov podiel dvojdetného modelu dosiahol približne 52,5 %. Predbežné výsledky z generácií vydatých žien zo 70. rokov naznačujú, že predmetná úroveň bude s najväčšou pravdepodobnosťou historickým maximom.

4 Konečná plodnosť a paritná štruktúra vydatých žien v regiónoch Slovenska

Zmeny v intenzite a charaktere reprodukčného správania neprebiehajú v celej populácii naraz, ale vždy je možné identifikovať určité skupiny obyvateľstva (pozri napr. Livi-Bacci, 1986) alebo priestory (napr. Coale a Treadway, 1986; Vereš, 1983, 1986), v ktorých tieto transformačné posuny je možné identifikovať skôr. To je aj prípad Slovenska, ako potvrdili niektoré predchádzajúce výskumy (napr. Fialová et al., 1990; Vereš, 1986). V našom prípade nebudeme vychádzať pri analýze manželskej plodnosti z tzv. Coalových indexov (napr. Coale a Treadway, 1986), ale budeme pracovať priamo s výsledkami sčítaní obyvateľov, ktoré reflektujú realizovanú plodnosť vydatých žien.

Podľa údajov z medzivojnového sčítania ľudu 1930 dosahovali najvyššiu konečnú plodnosť vydaté ženy vo veku 50 a viac rokov v priestore okresov od Prievidze cez Topoľčany, Hlohovec, Nitru až po Nové Zámky, kde sa jej hodnota

pohybovala nad hranicou 6 detí. Vysoký priemerný počet narodených detí nachádzame aj vo väčšine okresov Považia s výnimkou okresov Trenčín, Nové Mesto nad Váhom a Myjava, a to v pomerne súvislom páse od regiónu Žitného ostrova (okresy Šamorín, Dunajská Streda) až po severoslovenské celky Veľká Bytča a Kysucké Nové Mesto (obr. 1). Zaujímavosťou pritom je, že okresy Čadca a najmä oravské celky Námestovo a Trstená, ktoré ešte donedávna boli priestorom s najvyššou plodnosťou na Slovensku (pozri obr. 3) do tejto skupiny nepatrili. Ďalšiu kompaktnejšiu oblasť s konečnou plodnosťou v rozmedzí 5,5 – 6,0 detí na jednu vydatú ženu vytvárali okresy severovýchodného Slovenska tiahnuce sa v prihraničnom páse od Starej Ľubovne po Medzilaborce. Z východného Slovenska sem môžeme ešte zaradiť okresy Spišská Nová Ves, Gelnica a zo stredného okres Brezno (obr. 1).



Obr. 1 Konečná plodnosť vydatých žien vo veku 50 a viac rokov v okresoch Slovenska podľa výsledkov sčítania ľudu 1930
(*Zdroj: vlastné spracovanie na základe údajov zo sčítania ľudu 1930*)

Výsledky sčítania ľudu z roku 1930 tiež potvrdili existenciu pomerne rozsiahleho priestoru s výrazne podpriemernou plodnosťou. Ako nepriamo vo svojich prácach na základe Coalových indexov upozorňovali už Fialová et al. (1990) a Vereš (1983, 1986) išlo predovšetkým o okresy na juhu stredného Slovenska. Najnižšiu konečnú plodnosť vydatých žien náš výskum identifikoval najmä v prihraničných celkoch tiahnucich sa od Želiezoviec, Levíc (spolu s Banskou Štiavnicou) cez Krupinu, Modrý Kameň, Lučenec, Rimavskú Sobotu, Feledince až po Tornaľu a Revúcu. Menej ako 4,5 dieťaťa na jednu vydatú ženu dosahovali tiež okres Bratislava a prekvapivo aj Dolný Kubín. Aj to potvrdzuje skutočnosť, že pásmo oravských okresov nepatrilo medzi priestory

s nadpriemernou intenzitou rodenia detí, ale realita sa zdá bola skôr opačná. Možné príčiny tohto stavu budú určite predmetom nášho ďalšieho výskumu.

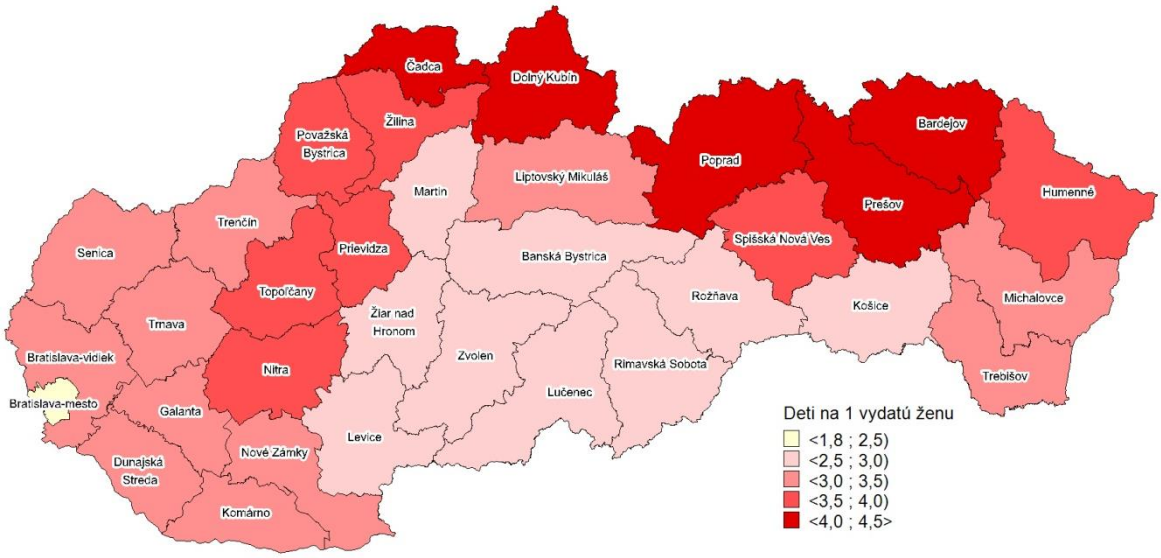
Relatívne nízku plodnosť dosahovali aj niektoré okresy priamo nadväzujúce na identifikovaný priestor veľmi nízkej manželskej plodnosti. Išlo napríklad o okresy stredného a južného Slovenska: Banská Bystrica, Zvolen, Rožňava, Parkan, ale aj niektoré okresy geograficky vzdialené v oblasti východoslovenskej nížiny (Košice, Trebišov, Michalovce, Veľké Kapušany), Turca (okres Martin), Oravy (Námestovo, Trstená) a severného Spiša (Kežmarok, Poprad).

Povojnové sčítanie ľudu z roku 1961 aj napriek výraznej zmene administratívneho členenia potvrdilo významný pokles realizovanej plodnosti vydatých žien aj na regionálnej úrovni. Kým zhruba pred 30 rokmi vo väčšine okresov konečná plodnosť presahovala hranicu 5 detí, na začiatku 60. rokov už ani jeden okres tak vysoký priemerný počet detí na jednu vydatú ženu nezaznamenal. Je pritom zaujímavé, že najvyššiu intenzitu plodnosti dosahovalo 5 okresov severného a severovýchodného Slovenska, kde sa konečná plodnosť dostala nad hranicu 4 detí (obr. 2). Najmä v prípade priestoru Oravy ide o veľmi dôležitý posun v hierarchii. Nadpriemernú konečnú plodnosť tiež môžeme identifikovať v okresoch Nitra, Topoľčany a Prievidza, ktoré sa do určitej miery kryjú s priestorom, v ktorom medzivojnové sčítanie našlo jednoznačne najvyššiu intenzitu manželskej plodnosti na Slovensku. Do tejto skupiny patrili tiež okresy Považská Bystrica a Žilina, ktoré tiež z väčšej časti môžeme vnímať ako pozostatok predchádzajúcej nadpriemernej realizovanej plodnosti vydatých žien. Na východnom Slovensku je obdobný záver možné vyjadriť v spojitosti s okresmi Prešov, Bardejov a Humenné, ktorých väčšina predchodcov v roku 1930 patrila tiež k celkom s vyššou intenzitou rodenia detí v manželskom zväzku.

Celkom opačná situácia bola v súvislom priestore 9 okresov stredného a juhu stredného Slovenska, kde v roku 1961 priemerný počet detí pripadajúcich na jednu vydatú ženu vo veku 50 a viac rokov neprekročil ani hranicu 3 detí. Aj v tomto prípade môžeme hovoriť o historickej previazanosti s celkami vyznačujúcimi sa podpriemernou úrovňou manželskej plodnosti. Navyše je tiež zrejmé, že došlo k ich určitej priestorovej expanzii. Jednoznačne najnižšia konečná plodnosť na začiatku 60. rokov bola identifikovaná u vydatých žien v hlavnom meste. Podľa dostupných údajov zo sčítania 1961 tu priemerný počet detí dosiahol len hodnotu 1,8 dieťaťa.

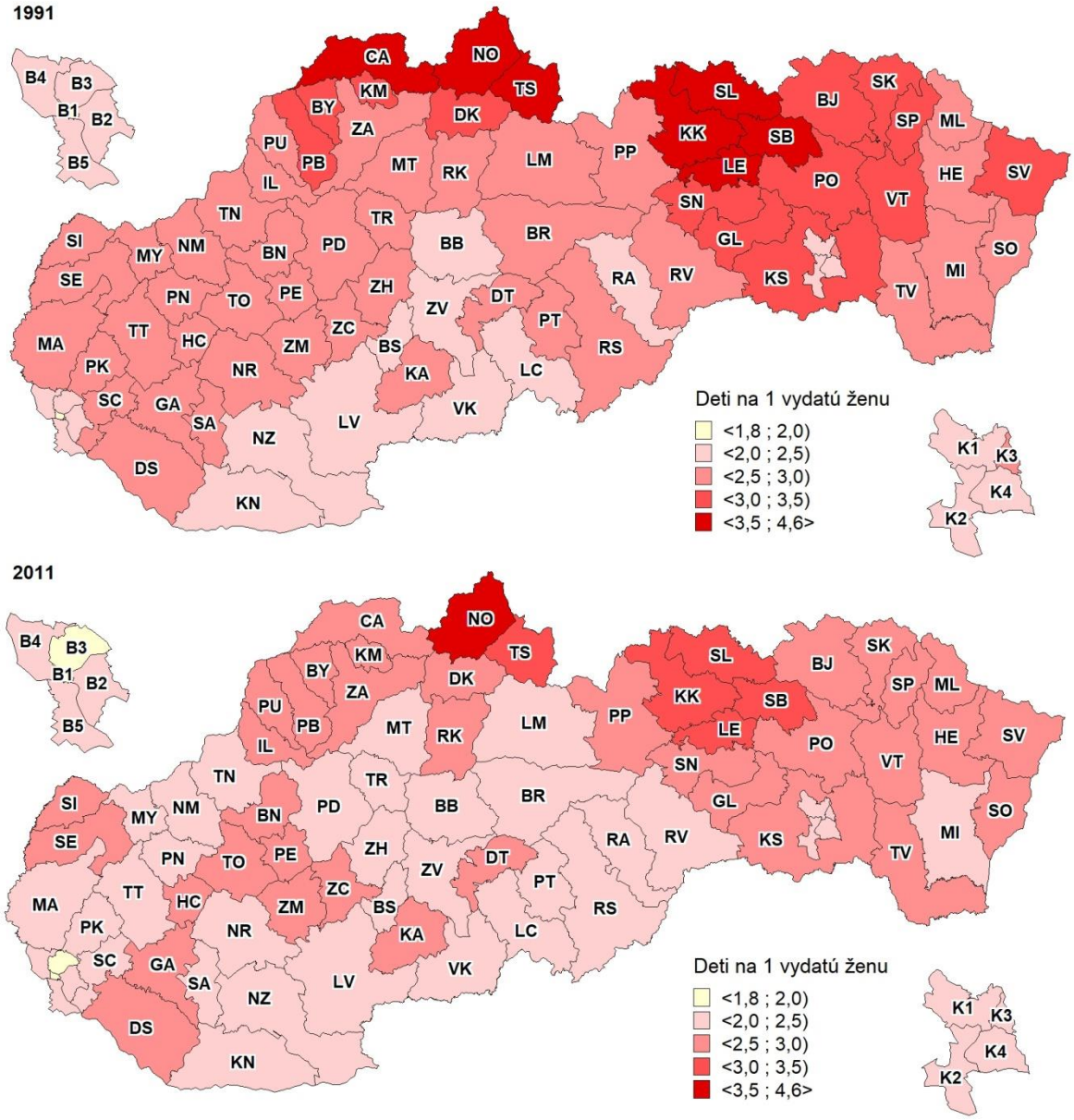
O približne tri desaťročia neskôr pôsobenie špecifických podmienok minulého politického režimu skončilo a výsledky sčítania ľudu z roku 1991 poskytovali naposledy obraz jeho dlhodobého pôsobenia. Vo všeobecnosti sa potvrdil už len mierny pokles konečnej plodnosti, keď vo väčšine okresov pripadalo na jednu vydatú ženu v priemere 2,5 – 3,0 dieťaťa. Súčasne tiež z priestorového hľadiska

v podstate pretrvali niektoré regionálne diferenciácie. Nadpriemerná manželská plodnosť bola spájaná najmä s okresmi severného a východného Slovenska (obr. 3). V ich prípade konečná plodnosť prekračovala hranicu 3, resp. 3,5 dieťaťa na vydatú ženu. Opačnú situáciu identifikujeme v okresoch dvoch najväčších miest a tiež v historickom priestore nízkej manželskej plodnosti na juhu stredného Slovenska. K nim sa pripájali tiež okresy Banská Bystrica a Zvolen (obr. 3)



Obr. 2 Konečná plodnosť vydatých žien vo veku 50 a viac rokov v okresoch Slovenska podľa výsledkov sčítania ľudu 1961
(*Zdroj: vlastné spracovanie na základe údajov zo sčítania ľudu 1961*)

Do roku 2011 sa počet okresov, kde sa konečná plodnosť vydatých žien pohybovala v rozmedzí 2,0 – 2,5 dieťaťa významne zväčšil. S výnimkou skupín okresov v páse od Dunajskej Stredy, Galanty po Bánovce nad Bebravou a Zlaté Moravce so Žiarom nad Hronom pokrýva väčšinu západného Slovenska. Je pritom zaujímavé, že práve tento priestor do určitej miery nadväzuje na historický vzorec okresov s nadpriemernou manželskou plodnosťou (porovnaj obr. 1 – 3). Do tejto skupiny sa tiež zaradila väčšina okresov juhu stredného a stredného Slovenska, čím sa opätovne vytvorila pomerne kompaktná oblasť (rovnako ako tomu bolo v roku 1961). Rovnako z geografického hľadiska aj napriek značnej redukcii z pohľadu početnosti i samotnej úrovne, zostávajú stále niektoré celky na severe a východe Slovenska predstavovať okresy s výrazne nadpriemernou intenzitou plodnosti vydatých žien. Aj v tomto prípade môžeme hovoriť o určitej historickej kontinuite.



Obr. 3 Konečná plodnosť vydatých žien vo veku 50 a viac rokov v okresoch Slovenska podľa výsledkov sčítania ľudu 1991 a 2011
(*Zdroj: vlastné spracovanie na základe údajov zo sčítania ľudu 1991 a 2011*)

5 Záver

Analýza výsledkov vybraných sčítaní obyvateľov potvrdila výrazné a v mnohých aspektoch historicky jedinečné zmeny v intenzite a vnútornom charaktere plodnosti vydatých žien na Slovensku. Predovšetkým umožnili pochopiť rozsah a dynamiku medzigeneračného poklesu realizovanej plodnosti, keď sa konečná plodnosť najprv pomerne rýchlo znížila zo svojej iniciačnej úrovne viac ako 5 detí o približne dve deti na ženu. Po určitej stagnácii sa u žien narodených v 20. rokoch trend znižovania intenzity rodenia detí znovu oživil

a v podstate pokračoval až do najmladších generácií. Priemerný počet detí pripadajúcich na jednu vydatú ženu sa tak u osôb narodených v druhej polovici 60. rokov dostal na hodnotu 2,2 dieťaťa, pričom predbežné údaje o vydatých ženách zo 70. rokov poukazujú na ďalšie pokračovanie tohto trendu.

Z hľadiska paritnej štruktúry je zrejmé, že presadenie sa dvojdetného modelu na úkor početnejšej rodiny pri súčasne marginálnej pozícii jednodetnosti a bezdetnosti bolo hlavným výsledkom transformačných zmien v reprodukčných vzorcoch vydatých žien v medzigeneračnom pohľade. Predbežné údaje za generácie zo 70. rokov, ako aj posledný vývoj v prípade vydatých žien z konca 60. rokov však naznačujú, že tento obraz prechádza a bude prechádzať ďalšími transformačnými posunmi. Môžeme tak aj u vydatých žien predpokladať pokles zastúpenia osôb s dvomi deťmi a naopak rast bezdetnosti a najmä jednodetnosti.

Výsledky sčítaní obyvateľov z rokov 1930, 1961, 1991 a 2011 nielenže potvrdili postupný pokles realizovanej manželskej plodnosti na regionálnej úrovni, ale umožnili tiež identifikovať v čase do určitej miery pretrvávajúci obraz priestorových diferencií konečnej plodnosti vydatých žien. Vo všeobecnosti pri určitej generalizácii môžeme povedať, že priestor juhu stredného a stredného Slovenska je dlhodobo vnímaný ako oblasť podpriemernej intenzity rodenia detí v manželskom zväzku. Sem môžeme tiež zaradiť okresy hlavného mesta. V čase sa tento priestor navyše do určitej miery rozširoval, pričom geografické maximum pozorujeme v poslednom dostupnom sčítaní z roku 2011.

Na druhej strane stoja viaceré okresy východného Slovenska a najmä jeho severnej časti, v ktorých dlhodobo nachádzame výrazne nadpriemernú manželskú plodnosť. V tomto prípade však môžeme vidieť opačný trend, a to zmenšovanie tohto priestoru, ktorý v poslednom cenze predstavovalo len niekoľko celkov. Zaujímavým bol pritom vývoj v severných okresoch stredného Slovenska. Oblasť Oravy a Kysúc sa na začiatku 60. i 90. rokov vyznačovala nadpriemerným počtom narodených detí vydatým ženám. Tento jav však neplatil v medzivojnovom období, kedy najmä v prípade oravských okresov sme mohli pozorovať skôr priemernú až podpriemernú úroveň manželskej plodnosti. To ukazuje, že nie všetky priestorové vzorce vysokej a nízkej plodnosti vydatých žien na Slovensku sú historickým reliktom, ale niektoré sa formovali pravdepodobne až v období minulého politického režimu.

6 Literatúra

Bleha, B., Vaňo, B., Bačík, V. (2014). *Demografický atlas Slovenskej republiky*. Bratislava: GeoGrafika.

Botíková, M., Švecová, S., Jakubíková, K. (1997). *Tradície slovenskej rodiny*. Bratislava: VEDA.

Coale, A.J., Treadway, R. (1986). A summary of the changing of overall fertility, marital fertility, and the proportion married in the provinces of Europe. In Coale, A. J., Watkings, S. C. (eds.) *The decline of fertility in Europe*. Princeton: Princeton University Press, s. 31–181.

Fialová, L., Pavlík, Z., Vereš, P. (1990). Fertility decline in Czechoslovakia during the last two centuries. *Population Studies*, 44(1), 89–106.

Hajnal, J. (1965). European marriage pattern in historical perspective. In Glass, D. V., Eversley, D. E. C. (eds.). *Population in history.* London: Arnold, s. 101–143.

Livi-Bacci, M. (1986). Social-group forerunners of fertility control on Europe. In Coale, A. J., Watkings, S. C. (eds.) *The decline of fertility in Europe*. Princeton: Princeton University Press, s. 182–200.

Livi-Bacci, M. (2003). *Populace v evropské historii*. Praha: Nakladatelství Lidové Noviny.

Pavlík, Z., Rychtaříková, J., Šubrtová, A. (1986). *Základy demografie*. Praha: Academia.

Potančoková, M. (2008). Plodnosť žien na Slovensku v období rokov 1950–2007 v generačnom pohľade. Bratislava: INFOSTAT.

Potančoková, M. et al. (2008). *Slovakia: fertility between tradition and modernity*. *Demographic Research,* 19(7), 973–1018.

Sobotka, T. (2011). Fertility in Central and Eastern Europe after 1989: Collapse and gradual recovery. *Historical Social Research*, 36(2), 246–296.

Šprocha, B., Tišliar, P. (2016). *Transformácia plodnosti žien Slovenska v 20. a na začiatku 21. storočia*. Bratislava: Centrum pre historickú demografiu a populačný vývoj Slovenska, Filozofická fakulta Univerzity Komenského v Bratislave.

Šprocha, B., Tišliar, P. (2018). *100 rokov obyvateľstva Slovenska: od vzniku Československa po súčasnosť*. Bratislava: Muzeológia a kultúrne dedičstvo.

Šprocha, B., Tišliar, P. (2021). Niektoré aspekty nemanželskej plodnosti a detí narodených mimo manželstva na Slovensku. *Sociológia*, 53(4), 339–376.

Vaňo, B. (2001). Populačný vývoj na Slovensku v rokoch 1945 – 2000. Bratislava: INFOSTAT.

Vereš, P. (1983). Vývoj plodnosti na Slovensku v letech 1880 – 1910. *Demografie*, 25(3), 203–208.

Vereš, P. (1986). Regionální vývoj plodnosti na Slovensku v letech 1910 – 1980. *Demografie*, 28(2), 110–117.

7 Poďakovanie

Článok bol vypracovaný za podpory grantovej agentúry APVV-20-0199 Transformácia populačného vývoja na Slovensku v regionálnom pohľade od konca 19. do polovice 20. storočia.

FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 2 pp 77-98

Štatistika v medicíne – úloha štatistiky v príprave medicínskeho fyzika
Statistics in medical physics – the role of statistics in the training of a medical physicist

Iveta Waczulíková[a], Pavol Bokes[a], Radoslav Böhm [a], Milan Zvarík[a], Milan Melicherčík[a], Marcela Morvová[a], Silvia Hnilicová[b], Pavol Vitovič[b]

[a] Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Mlynská dolina F1, 842 48 Bratislava, Slovenská republika
[a] Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Mlynská dolina F1, 842 48 Bratislava, Slovak Republic
[b] Univerzita Komenského v Bratislave, Lekárska fakulta, Špitálska 24, 813 72 Bratislava, Slovenská republika
[b] Comenius University in Bratislava, Faculty of Medicine, Špitálska 24, 813 72 Bratislava, Slovak Republic
iveta.waczulikova@fmph.uniba.sk, pavol.bokes@fmph.uniba.sk, radoslav.bohm@fmph.uniba.sk, milan.zvarik@fmph.uniba.sk, morvova2@uniba.sk, silvia.hnilicova@fmed.uniba.sk, pavol.vitovic@fmed.uniba.sk

Abstrakt: *Význam štatistiky v oblasti medicíny sa viaže častejšie na roly, ktoré plní v epidemiológii a verejnom zdravotníctve. Menej sa zdôrazňuje skutočnosť, že štatistika má nezastupiteľné miesto v medicínskom výskume a jeho translácii do klinickej praxe. Vo využívaní štatistiky v medicíne máme na Slovensku stále veľké rezervy. Pre skvalitnenie prípravy absolventov biofyziky a biomedicínskej fyziky pre pozície vo výskume, vývoji a klinickej praxi, postupne zavádzame zmeny vo výučbe (podporené KEGA 041UK-4/2020). Základný kurz štatistiky sme doplnili prakticky orientovanými predmetmi a simulačným vzdelávaním. Vo výučbe využívame metódy aktívneho vyučovania v tíme (metóda TBL) a témy záverečných prác študentov viažeme na výskumné projekty. V príspevku uvádzame niekoľko príkladov využitia štatistiky a matematického modelovania v oblasti medicínskej fyziky a biomedicíny.*

Abstract: *The importance of medical statistics is more often linked to the roles it plays in epidemiology and public health. Less emphasized is the fact that statistics is one of the crucial parts of medical research and its translation into clinical practice. There are still big gaps in the use of statistics in medicine in Slovakia. In order to improve competences and preparedness of Biomedical physics graduates for positions in research, development and clinical practice, we have introduced changes in teaching (supported by KEGA 041UK-4/2020). We have supplemented the basic course of statistics with practically oriented subjects and simulation training. We use a method of team-based learning (TBL) in teaching and relate the topics of students' theses to research projects. Here we present several examples of the application of statistics and mathematical modeling in Medical Physics and Biomedicine.*

Kľúčové slová: *bioštatistika, štatistická fyzika, medicínska fyzika, simulácia, tímové rozhodovanie.*

Key words: *biostatistics, statistical physics, medical physics, simulation, team decision making.*

1 Úvod

Štatistika zaujíma významné miesto v prakticky vo všetkých vedných oblastiach, medicínu nevynímajúc. Význam štatistiky v medicíne sa však bežne spája s menej alebo viac komplexnými úlohami, ktoré plní v epidemiológii a verejnom zdravotníctve. Menej sa zdôrazňuje skutočnosť, že štatistika má nezastupiteľné miesto v medicínskom výskume a jeho translácii do klinickej praxe. Je dôležitá už pri plánovaní výskumu a má svoje miesto aj pri realizácii výskumu: zbere dát, ich vyhodnocovaní a interpretácii získaných výsledkov. Vo využívaní štatistiky v medicíne však máme na Slovensku stále veľké rezervy. Tento názor si autori článku vytvorili nielen cez osobnú skúsenosť z veľkého počtu vlastných výskumných spoluprác, ale aj vďaka spätnej väzbe od absolventov študijných programov Biofyzika (BF) a Biomedicínska fyzika (BMF)[8] na Fakulte matematiky, fyziky a informatiky Univerzity Komenského (FMFI UK), na ktorých garantovaní a zabezpečovaní sa autori príspevku podieľajú (c.f. Kusendová et al., 2018). Veľká väčšina biomedicínskych a klinických výskumných tímov si uvedomuje, že schopnosť analyzovať dáta, modelovať biologické a fyziologické procesy a rozumieť grafom a tabuľkám predstavuje veľmi dôležitú a užitočnú kompetenciu pre kvalifikované rozhodovanie (Rečková, 2016), a snaží sa preto situáciu (v rámci svojich možností a prostriedkov) riešiť.

Jednou z možností – podľa všetkého najmenej využívanou – je osloviť, prípadne zamestnať v tíme (organizácii) profesionálneho štatistika. Spolupráca je na strane profesionálnych štatistikov limitovaná nedostatkom času a skúseností s problematikou medicínskeho výskumu. Prípadná pozícia v prostredí medicínskeho výskumu naráža na problémy, ktorých rozbor je mimo rámec tohto príspevku. Časť tímov sa snaží zaškoliť a poveriť štatistickými úlohami vlastného člena. Táto možnosť, aj keď je historicky najčastejšia, nesie so sebou nevýhody spojené s nedostatkom základných znalostí štatistiky a skúseností s analýzou (bio)medicínskych a biologických dát, ktoré majú svoje špecifiká. Efektívnou cestou sa ukázalo byť rozšírenie interdisciplinárneho a multidisciplinárneho charakteru medicínskeho výskumu a začlenenie do tímu mladých výskumníkov, ktorí potrebné znalosti a zručnosti získali počas štúdia a môžu sa v tomto smere ďalej rozvíjať.

Od vzniku medziodborového študijného programu BMF vyučovaného v spolupráci s Lekárskou fakultou (LF) UK stále rastie podiel absolventov BMF a BF na poli základného a aplikovaného medicínskeho výskumu. Dôvody zvýšeného záujmu o absolventov BMF vidíme v unikátnej kombinácii predmetov

[8]Facebooková skupina skupina absolventov študijného programu BMF: https://www.facebook.com/groups/biomedicalphysics.

fyziky, matematiky a medicíny[9] v oblasti nadobudnutých znalostí a v priamom zapojení študentov BMF, ale aj BF, do bežiacich výskumných projektov, kde pod odborným vedením fakultných aj externých školiteľov (c.f. Spolupráca s externými pracoviskami uvedená v časti Poďakovanie) vyhodnocujú a interpretujú získané výsledky. Absolventi oboch smerov sa súčasne uplatňujú na pozíciách medicínskeho fyzika, rádiofyzika a klinického fyzika priamo v klinickom prostredí[10].

Pre skvalitnenie prípravy absolventov BF a BMF pre pozície vo výskume, vývoji a klinickej praxi postupne zavádzame zmeny vo výučbe štatistiky a pridružených predmetoch zameraných na radiačnú fyziku, experimentálne metódy a dizajn, výpočtovú techniku a programovanie, spracovanie obrazu ai., o čom sme čiastočne referovali už v príspevku Kusendová et al. (2018). V týchto aktivitách pokračujeme aj v rámci projektu Kultúrnej a edukačnej agentúry KEGA 041UK-4/2020 s názvom *Odborná príprava klinického a biomedicínskeho fyzika - implementácia nových technológií biomedicíny a metód aktívneho učenia.*

Cieľom tohto príspevku je poskytnúť stručný prehľad o výučbe štatistiky a aplikovanej matematiky v študijných programoch BF a BMF v blokoch podľa obsahového zamerania. V prvej časti bloku v krátkosti informujeme o obsahovej náplni predmetov a o prístupoch, ktoré vyučujúci využívajú vo výučbe. V druhej časti bloku uvádzame využitie štatistiky na príkladoch z výskumu. V samostatnej kapitole predstavíme vybrané príspevky našich študentov.

2 Štatistika v predmetoch základného kurzu matematiky a fyziky

2.1 Pravdepodobnosť a štatistika

Výučba

Predmet *Pravdepodobnosť a štatistika* zabezpečuje v druhom roku štúdia doc. Mgr. Pavol Bokes, PhD., z Katedry aplikovanej matematiky a štatistiky FMFI UK. Cvičenia k predmetu vedie doktorandka z tej istej katedry Mgr. Iryna Zabaikina. Predmet je vedený so zreteľom na integráciu teoretických znalostí a technologických zručností. V kontexte výučby štatistiky sa táto integrácia prejavuje využitím štatistických programov na ilustráciu štatistických a pravdepodobnostných metód. Vyučujúcim sa na štatistické výpočty osvedčil sofvér R. Toto prostredie je široko používané na domovskej katedre vyučujúcich. V základnej inštalácii je prostredie pomerne striedme, poskytuje textovú konzolu

[9]Študijný plán študijného programu BMF je dostupný na: https://sluzby.fmph.uniba.sk/infolist/sk/sp_BMF.html.

[10]Príkladom môže byť Oddelenie klinickej fyziky Onkologického ústavu sv. Alžbety: http://www.ousa.sk/sk/klienti-a-pacienti/kliniky-a-lozkove-oddelenia/ustav-klinickej-fyziky-szu-a-ousa.

do ktorej používateľ píše príkazy interpretovaného jazyka. Striedmosť považujú vyučujúci za výhodu – študent si uvedomí, že podstatou štatistickej práce sú algoritmy a dátové štruktúry, zatiaľ čo prípadné užívateľské rozhranie je nadstavbou.

Prostredie R pritom poskytuje širokospektrálnu funkcionalitu, ktorá zahrňuje základné štatistické testy, zobrazovacie metódy, ako aj užitočné implementácie základov teórie pravdepodobnosti. Pre všetky základné rozdelenia pravdepodobnosti preberané v rámci predmetu (binomické, hypergeometrické, normálne, chi-kvadrát, Studentovo, Fisherovo), prostredie poskytuje implementácie hustoty (vzhľadom na sčítaciu resp. Lebesguevu mieru), distribučnej funkcie, kvantilovej funkcie, a (pseudo)náhodnú generáciu premenných.

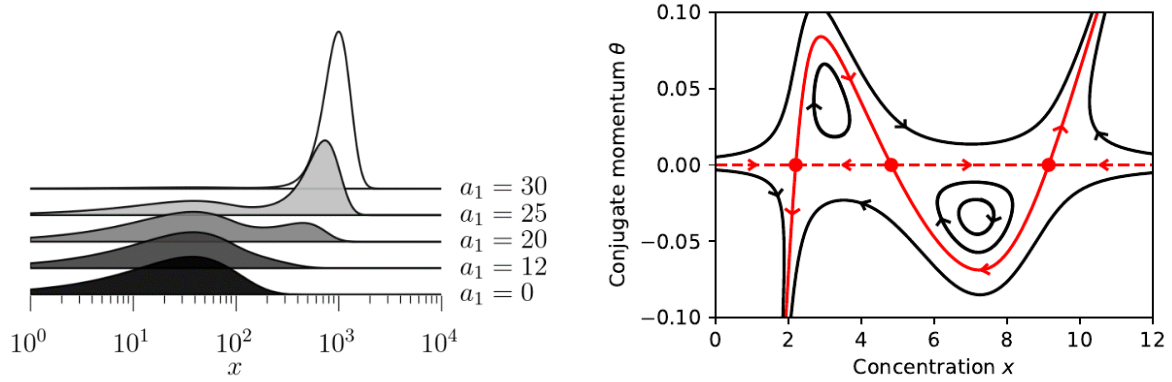
V oblasti štatistiky je použitie technologickej nadstavby nevyhnutné, aby sa študenti dôverne zoznámili s pojmom *p*-hodnoty. Štandardná, učebnicová metodológia je totižto založená na použití kritických hodnôt, ktoré sa dajú stručnejšie tabelovať v učebniciach ako *p*-hodnoty. Oba prístupy sú samozrejme ekvivalentné. Kombinácia klasických a novodobo-technologických prístupov má za cieľ študentom túto ekvivalenciu vysvetliť a dať im dostatočné porozumenie zmyslu *p*-hodnoty nielen pri základných preberaných testoch, ale aj vo všeobecnosti pri diapazóne iných štatistických metodológií.

Čo sa týka metodologického hľadiska prístupu k vyučovanej látke, vyučujúcim sa javí vhodné zdôrazňovať dialektický kontrast medzi diskrétnym a spojitým prípadom. V prípade teórie pravdepodobnosti to zodpovedá známemu rozdeleniu pravdepodobnostných rozdelení podľa toho, či sú absolútne spojité vzhľadom na diskrétnu sčítaciu mieru, alebo spojitú mieru Lebesgue.

Vedecko-výskumné aktivity

V dialektickej otázke diskrétnosti a spojitosti sa pedagogická činnosť vyučujúcich prelína s ich vedeckou prácou. Táto práca je motivovaná predovšetkým biofyzikou génovej expresie na úrovni jednotlivých buniek (c.f. špecializačný predmet *Modelovanie biologických procesov*, kapitola 3.2). V danej oblasti je prirodzené použiť diskrétnu metodiku vzhľadom na atomárnosť jednotlivých biochemických molekúl (DNA, mRNA molekuly, proteíny). Napriek tomu, matematické singulárno-perturbačné metódy vedú k spojitej formulácii takýchto systémov, ktorá je často prehľadnejšia ako pôvodná diskrétna. Napríklad, pre systém pozitívnej regulácie s oneskorením sa ukazuje ako vhodný hamiltonovský popis (obr. 1, vpravo), ktorý je známy z klasickej (spojitej) mechaniky pevného bodu (Bokes et al., 2020). Pre problém s prítomnosťou takzvaných falošných väzobných miest (anglicky *decoy binding sites*) vedie

redukcia k odvodeniu pomerne jednoduchých spojitých molekulárnych rozdelení (obr. 1, vľavo) (Bokes s Singh, 2015). Podobnými systémami sa zaoberá aj cvičiaca predmetu v rámci projektu dizertačnej práce (Zabaikina et al., 2020). Týmto spôsobom dochádza k integrácii vyučovanej tematiky s vedeckou prácou.



Poznámka: Vertikálna os v grafe naľavo predstavuje pravdepodobnostnú hustotu. Horizontálna os v grafe napravo predstavuje množstvo bielkoviny x, vertikálna os zovšeobecnenú hybnosť. Prerušené červené čiary reprezentujú deterministickú dynamiku s dvoma stabilnými bodmi a plné červené čiary možnosť stochastického úniku z deterministicky stabilných stavov prostredníctvom veľkých deviácií

Obr. 1: (vľavo) Teoretické rozdelenia množstva bielkoviny x pri rôznych koncentráciach a_1 aktivátora pri prítomnosti falošných väzobných miest, **(vpravo)** hamiltonovská charakterizácia deviácií v modeli s pozitívnou spätnou väzbou (*vlastné spracovanie v softvéri Matplotlib*).

2.2 Štatistická fyzika

Výučba

Predmety štatistickej fyziky zabezpečuje a koordinuje RNDr. Radoslav Böhm, PhD. Študenti sa s týmto predmetom stretávajú hneď na začiatku svojho štúdia, v prvom ročníku na prednáškach z mechaniky a zo základných matematických metód. Z dlhoročnej skúsenosti vyučujúcich vyplýva, že ideálnou témou na pochopenie štatistiky je popis správania plynov. Študenti s veľkým prekvapením zisťujú, aké obrovské množstvo častíc plyn obsahuje – ak by mali rozdeliť molekuly vzduchu z 1 cm^3 všetkým obyvateľom sveta, každému by sa ušla približne 1 miliarda takýchto častíc. Študenti sami prichádzajú k záveru, že systémy s obrovským počtom členov nemožno (minimálne z praktických dôvodov) popisovať dynamicky a jedinou alternatívou je štatistický prístup. Zoznamujú sa s procesom spriemerovania štatistických súborov, pričom si uvedomia, že je to na úkor straty množstva informácií, ktoré môžu chýbať pri vysvetľovaní niektorých javov. Napríklad, pri odvodzovaní stavovej rovnice sa „ustredňujú“ hybnosti prenášané molekulami na stenu nádoby, s ktorou sa zrážajú, na druhej strane spriemerovaním rýchlostí protónov na Slnku by nebolo

možné vysvetliť termojadrovú fúziu. Protóny, ktoré sa na nej zúčastňujú, pochádzajú z chvosta Maxwellovho rozdelenia a ich rýchlosti sú niekoľkonásobne vyššie ako je hodnota ich strednej rýchlosti.

Štatistika pokračuje štandardným kurzom *Termodynamika a štatistická fyzika* v treťom ročníku (RNDr. Radoslav Böhm, PhD., spolu s RNDr. Pavlom Bartošom, PhD.). Študenti sa oboznamujú s jednotlivými typmi rozdelení: Maxwell-Boltzmannovým, kanonickým, grand-kanonickým, Boseho-Einsteinovým, Fermiho-Diracovým a ich aplikáciami (barometrická formula, efúzia, paramagnetizmus, merná tepelná kapacita tuhých látok, rozšírenie spektrálnych čiar dopplerovým efektom a jeho použitie na meranie teploty v extrémnych podmienkach, Fermiho plyn, žiarenie čierneho telesa a pod). Poznatky zo štatistiky sa v rámci základného kurzu matematiky a fyziky aplikujú aj v kurzoch atómovej a jadrovej fyziky (premena jadier), ako aj kvantovej mechaniky.

Vedecko-výskumné aktivity

Oddelenie radiačnej fyziky na Katedre jadrovej fyziky a biofyziky, kde pôsobí vyučujúci, je (okrem iného) výskumne zamerané na štúdium pravdepodobnosti vzniku rakoviny pľúc inhaláciou radónu a fajčením. Rakovinové ochorenie má monoklonálny pôvod a preto ho možno hodnotiť z bunkovej úrovne. Autor sa výskumne venoval vývoju mikrodozimetrického modelu (Böhm et al., 2014) pre predikciu rizika rakoviny pre jedincov s rôznymi fajčiarskymi návykmi, ako aj rôznymi radónovými expozíciami. Model je založený na predpoklade, že sledovaný biologický účinok (inaktivácia bunky) sa prejaví pri prekročení prahovej hodnoty mernej energie z_0 v terči. Ak pri danej radónovej expozícii W je bunka zasiahnutá a merná energia z neprekročí prah z_0, dochádza k jej subletálnemu poškodeniu. Množstvo týchto buniek je úmerné prírastku relatívneho rizika vzniku rakoviny pľúc. V prípade fajčiara indukuje tabak a jeho komponenty nárast syntézy proteínu nazývaného survivín. Survivín je najmenším členom rodiny inhibítorov apoptózy a spôsobuje tak odolnosť poškodených buniek na podnety (napr. na radiáciu), ktoré za normálnych okolností vedú k ich zániku. To má za následok zvýšenie hodnoty hraničnej energie z_0 a tým aj zvýšenie rizika vzniku rakoviny pľúc. Model sa kalibruje metódou Monte Carlo – simulovaním interakcie alfa častíc s terčovými bunkami pľúcneho tkaniva. Nakalibrovaný model má predikčný potenciál určovať pravdepodobnosť vzniku rakoviny pľúc pre jedincov s rôznymi fajčiarskymi návykmi vystaveným rôznym radónovým expozíciam.

Radón môže slúžiť ako stopovač pľúcnych zmien vyvolaných fajčením (Böhm et al., 2020). S počtom vyfajčených cigariet vzrastá hrúbka mukóznej vrstvy, ktorá chráni terčové bunky pľúcneho tkaniva pred radiáciou, na druhej strane sa fajčením narúša mechanizmus čistenia pľúc od produktov premeny radónu,

dochádza k ich kumulácii na povrchu dýchacích ciest, čo vedie k nárastu radiačnej záťaže. Hyperprodukcia hlienu znižuje bronchiálnu dávku, kumulácia produktov premeny ju zvyšuje. Bronchiálna dávka teda závisí od stupňa poškodenia pľúcneho tkaniva a preto ju možno využiť ako vhodný prostriedok na odhalenie účinkov fajčenia na pľúcny epitel. Stupeň týchto poškodení kvantifikuje pomer dávky fajčiara a nefajčiara ε. V prípade, že u fajčiara prevláda hyperprodukcia hlienu bude hodnota $\varepsilon < 1$, v prípade prepuknutia chronickej obštrukcie $\varepsilon > 1$.

Po skončení fajčenia sa pľúca abstinujúceho fajčiara regenerujú a postupne ozdravujú. V pľúcach opäť začnú rásť riasinky, ktoré znížia riziko infekcie a vyčistia pľúca. Ozdravný proces pľúc je dlhodobý, závisí od stupňa poškodenia pľúc počas fajčenia. Analýzou zmeny pľúcnej dávky počas abstinencie možno zrekonštruovať časový vývoj pľúcnych zmien (Böhm et al., 2019).

3 Štatistika v prakticky zameraných predmetoch

3.1 Výučba aplikovanej štatistiky

Tieto predmety pokrývajú úvod do problematiky databázových systémov, aplikovanej štatistiky a modelovania a slúžia ako základňa pre nadstavbové a špecializačné predmety na bakalárskom a magisterskom stupni štúdia. Absolvovaním nižšie uvedených predmetov získa študent základné vedomosti a prehľad v metódach popisnej a inferenčnej štatistiky a naučí sa rozumieť bioštatistickým dátam, správne s nimi pracovať v každej etape výskumu - od premyslenia a naplánovania výskumu, zberu dát, tvorby a spracovania dátového súboru, až po vyvodenie záverov v súlade s použitými analytickými nástrojmi.

Novoprijatí študenti odboru BMF prichádzajú z rôznych typov škôl a tým aj s rôznou úrovňou vedomosti a praktických skúsenosti so spracovaním dát. Za účelom minimalizovať tieto rozdiely bol vytvorený predmet *Spracovanie textových a* dátových *súborov*, ktorý zabezpečuje RNDr. Milan Zvarík, PhD. Predmet je nasadený v prvom ročníku a jeho absolvovaním sa študenti naučia pracovať so základnými, ale aj pokročilými funkciami tabuľkového procesoru MS Excel, ktorý majú voľne dostupný aj mimo fakulty. Okrem práce so samotným softvérom študenti získavajú základné vedomosti z odborov informatiky (typy dát, syntax funkcie a jej atribúty, cykly) a štatistiky (princíp analýzy variancie ANOVA a lineárnej regresie). Na tento predmet nadväzujú *Úvod do bioštatistiky a Navrhovanie a vyhodnocovanie experimentov s aplikáciami v biomedicíne a biofyzike* (prednášky a cvičenia zabezpečuje doc. RNDr. Iveta Waczulíková, PhD. a cvičenia pre študentov pridruženého bakalárskeho študijného programu Medicínska biológia na Prírodovedeckej fakulte UK vedie RNDr. Ing. Milan Melicherčík, PhD.) Na cvičeniach sa študenti BMF učia používať štatistický program StatsDirect, ktorý bol zakúpený vďaka grantovej podpore KEGA. Na

cvičeniach pre študentov Medicínskej biológie bol zvolený Microsoft Excel, nakoľko nie je k dispozícii dostatok licencií pre všetkých študentov a na väčšinu úloh možnosti Excelu postačujú. Pre pokročilejšie analýzy vyučujúci používajú aj doplnok Excelu BESH Stat (kapitola 5.5), ktorý rozširuje možnosti Excelu a zjednodušuje analýzu štatistických údajov. Študenti majú k dispozícii viacero učebníc, monografií a manuálov, za všetky možno spomenúť zdroje: Pekár a Brabec (2009, 2012), Motulsky (2014), Somorčík a Teplička (2015), Waczulíková a Slezák (2015) a manuály pre R[11].

Vďaka nadstavbovým a špecializačným predmetom, ktorým sa bližšie venuje kapitola 3.2, študenti rozumejú kontextu výskumných a aplikačných cieľov a môžu sa tak s jednotlivými štatistickými metódami oboznamovať v priamej väzbe na návrh (dizajn) reálnych biomedicínskych experimentov a observačných štúdií na vhodných príkladoch z praxe.

Študenti sa pritom učia rozlišovať:

- laboratórne (tzv. pravé) experimenty a klinické experimentálne štúdie, ktorých cieľom je overiť výskumnú hypotézu, a
- štúdie observačného typu, ktorých cieľom je zisťovanie vzťahov a prognózy (typicky štúdie prípadov a kontrol), zhody (jednovýberové štúdie s opakovanými pozorovaniami), prevalencie (typicky prierezové štúdie) alebo incidencie (typicky prospektívne kohortové štúdie).

Po oboznámení sa s experimentom alebo štúdiou realizovanými podľa návrhu (protokolu) študenti pracujú pod dohľadom vyučujúcich na súbore dát[12]. Študenti sa venujú jednoduchým popisným štatistikám a ich sumárnym zobrazeniam v tabuľkách a grafoch. Učia sa odhaľovať chyby v záznamoch, pripraviť údaje na analýzu, urobiť si základnú predstavu o prvých výsledkoch, posúdiť tvar závislosti a vhodnosť rôznych štatistických modelov, prípadne si všimnúť zvláštností, alebo neočakávaných trendov. Následne – podľa typu výskumu – dáta analyzujú individuálne alebo v tímoch.

Vo výučbe nástroje štatistického modelovania kladú vyučujúci dôraz, aby študenti zvládli modely regresného typu, teda modely, ktoré popisujú vzťah medzi jednou až niekoľkými vysvetľujúcimi (nezávislými) premennými a jednou závislou premennou. Poznajú základné vlastnosti regresného modelu, ktorý obsahuje systematickú a náhodnú zložku, v úvodnom predmete začínajú jednoduchým a všeobecným lineárnym modelom (*general linear model* alebo len LM) pre spojitú závislú premennú. Študenti rozumejú významu slov “lineárny”

[11]K dispozícii napr. na https://cran.r-project.org/manuals.html.

[12]Používame hlavne stĺpcový formát, pretože ho vyžadujú štatistické programy, ktoré pri výučbe používame (c.f. Kusendová et al., 2018), a väčšina metód programu R.

a “lineárny prediktor” v kontexte štatistického modelovania, trénujú aplikácie LM na nelineárne vzťahy medzi závislou a vysvetľujúcimi premennými (napr. závislosť odpovede od dávky alebo času), poznajú základné situácie, kedy môžu využiť transformáciu pravej, prípadne oboch strán rovnice, a vedia interpretovať parametre modelu a vykonať základnú diagnostiku modelu.

Študenti sa ďalej trénujú v identifikovaní vhodného základného typu analýzy podľa prítomnosti spojitých a kategoriálnych premenných v lineárnom prediktore. Ak sa v modeli vyskytujú iba spojité premenné, ide o jednoduchú alebo viacnásobnú regresiu, resp. jej analógiu v zovšeobecnenom LM (*generalised linear model,* GLM). V prípade laboratórnych experimentov sa často vyskytujú iba kategoriálne vysvetľujúce premenné tzv. faktory) a výskumný návrh vedie k analýze variancie ANOVA (a jej analógu v GLM). Najčastejšie ide o dva faktory, ktoré ovplyvňujú správanie vybranej spojitej premennej a treba posúdiť mieru ich vplyvu (efekt). V biomedicínskom výskume však nie sú výnimkou ani komplexnejšie dizajny. Študenti sa učia rozlišovať vnútroskupinové a medziskupinové efekty, vykonať detailné porovnanie skupín, prípadne “očistiť” skúmaný vzťah od vplyvu inej premennej. V animálnych experimentoch sú časté faktoriálové dizajny so zahrnutím interakcií, vyskytujú sa aj vnorené (hierarchické) dizajny. V prípade, že je medzi vybranými vysvetľujúcimi premennými aspoň jedna spojitá a aspoň jedna kategoriálna, výskumný návrh vedie k analýze kovariancie ANCOVA (a jej analógu v GLM). V observačných štúdiách sú časté výskumné situácie, keď je závislá premenná kategoriálna (najčastejšie binárna). Naviac, biomedicínsky a klinický výskum je svojou podstatou nerandomizovaným výberom z populácie (s výnimkou veľkých epidemiologických štúdií), kde môže byť vplyv študovaných vysvetľujúcich premenných skreslený inými znakmi (*extraneous variables, confounders*). Preto vyučujúci už na 1. stupni zaradili do výučby úvod do pokročilejších metód štatistického modelovania. Študenti sa v rámci GLM oboznamujú s jednoduchou a viacnásobnou logistickou regresiou pre binárnu a multinomickú závislú premennú[13].

Po absolvovaní týchto predmetov väčšina študentov bez väčších problémov zvládne spracovanie a analýzu dát pri písaní záverečnej práce už na bakalárskom stupni.

[13]Pokročilé metódy matematického modelovania sú tímami v medicínskom výskume málo využívané.

3.2 Výučba nadstavbových a špecializačných predmetov

Na bakalárskom stupni je veľmi dôležitým aspektom koordinácia predmetov orientovaných na výučbu štatistiky s ďalšími nadstavbovými predmetmi, ako napr. *Základy biomedicínskej fyziky* (zabezpečuje RNDr. Marcela Morvová, PhD.) alebo *Medicínske prístroje* (zabezpečujú doc. RNDr. Martin Kopáni, PhD., doc. RNDr. Iveta Waczulíková, PhD., a doc. RNDr. Pavol Vitovič, PhD.). V nich študenti získavajú vedomosti o:

- fyzikálnych zákonitostiach dôležitých pre základné biologické procesy, ktoré sú prístupné experimentálnemu alebo kvalitatívnemu zisťovaniu, a teda generujú dáta;
- pôsobení fyzikálnych faktorov na biologické systémy, ktoré im vyučujúci vhodne spájajú s poznatkami získanými na LF – princípmi fungovania jednotlivých systémov ľudského organizmu a funkčnými vzťahmi medzi systémami na rôznych stupňoch organizácie. Ako príklad môžeme uviesť prístroje používané v medicínskej praxi na snímanie, spracovávanie a vyhodnocovanie signálov, ktoré sú navrhované a vyvíjané v súlade fyzikálnymi zákonitosťami jednotlivých biosystémov;
- využití poznatkov hraničných odborov pri plánovaní a realizácii výskumu, ako aj následnom spracovávaní získaných údajov.

Na magisterskom stupni študenti absolvujú viacero špecializačných predmetov ako napr.:

Zdravotnícka a medicínska informatika (zabezpečuje RNDr. Marcela Morvová, PhD.), kde sa oboznamujú s informačnými technológiami používanými v medicínskom prostredí a so systémom e-Health.

Experimentálne metódy biofyziky a lekárskej fyziky (zabezpečujú RNDr. Marcela Morvová, PhD., RNDr. Milan Zvarík, PhD., a doc. RNDr. Pavol Vitovič, PhD.). Študent po absolvovaní získa znalosti o princípoch, základných schémach a aplikáciách experimentálnych fyzikálnych metód v medicínskej praxi a výskume.

Predmet *Metódy spracovania biosignálov a počítačová grafika je* vyučovaný v tvrtom ročníku (RNDr. Milan Zvarík, PhD.). Časť tohto predmetu je venovaná základnému kurzu programovania v prostredí R, ktoré ponúka vďaka bohatej ponuke knižníc obrovské množstvo štatistických analýz s dobrou dokumentáciou. Prostredie tiež ponúka množstvo možností na grafickú vizualizáciu výsledkov (boxploty, 3D grafy, teplotné mapy ai.). Študentom sú v tomto prostredí demonštrované hlavne princípy spracovania signálov – digitalizácia, de/konvolúcia, Fourierova analýza a štatistické spracovanie signálov.

Ďalšie použitie štatistiky je na predmete *Počítačové modelovanie* (vyučujúci RNDr. Ing. Milan Melicherčík, PhD) pre študentov BMF, kde už študenti samostatne využívajú niektoré štatistické výpočty v programe *GNU Octave*. Ťažiskom predmetu je síce tvorba modelov vybraných javov okolitého sveta, ale výstupy daných modelov (napr. model slnečnej sústavy, systém ideálneho plynu, modely popisujúce systém dravec-korisť a pod.) sa po ich implementácii dajú štatisticky spracovať.

Osnova predmetu *Modelovanie biologických procesov* (vyučujúci doc. Mgr. Pavol Bokes, PhD.) zahŕňa biologické modelovanie s obyčajnými diferenciálnymi rovnicami: princíp hmotnostnej bilancie, pravidlo hmotnostnej akcie, škálovanie a „zbezrozmernenie“ (dimenzionálna analýza), jednozložkové modely (kinetika Michaelisa a Mentenovej, génová autoregulácia), viaczložkové modely (biologické prepínače, oscilátory, epidemiológia). Ďalej modelovanie s diferenciálnymi rovnicami s oneskorením, modely s priestorovou zložkou (reakčno-difúzne systémy, šírenie epidémie, tvorba vzorkovania) a stochastické modely (rovnica bilancie pravdepodobnosti, Gillespieho simulačný algoritmus, modely génovej expresie).

O niečo skrytejšia je rola štatistiky v predmete *Molekulárno-dynamické simulácie* ponúkaného študentom magisterského štúdia v odboroch BMF a BF (RNDr. Ing. Milan Melicherčík, PhD, prof. RNDr. Ján Urban, DrSc.). Predmet je zameraný na teoretické pochopenie simulácie (bio)molekulových systémov a aj praktické realizovanie simulácií. Samotná metóda molekulovej dynamiky je jednoduchá a využíva len klasické newtonovské rovnice popisujúce pohyb častíc. Avšak výsledkom je obrovské množstvo dát o polohe a rýchlosti jednotlivých častíc, ktoré je potrebné spracovať aj za využitia štatistických metód. Nakoľko sa využívajú hotové nástroje používané vo vedeckom výskume, nie je potrebné tieto metódy študentami programovať, ale je snaha, aby si študenti aj napriek tomu uvedomovali, v čom spočíva dotyčná metóda výpočtu, čo môže poskytnúť a aké sú jej výhody a podmienky použitia.

3.3 Simulačná výučba

Simulačnú výučbu našich študentov zabezpečuje doc. RNDr. Pavol Vitovič, PhD., absolvent biofyziky na FMFI a vedúci Ústavu simulačného a virtuálneho medicínskeho vzdelávania (ÚSVMV), ktorý vznikol na LFUK v roku 2013 za účelom modernizovania pregraduálnej výučby študentov.[14] Simulačné vzdelávanie, ktoré sa od roku 2013 postupne implementovalo do širokého spektra teoretických a klinických predmetov, je považované za cennú súčasť

[14]O ústave sa možno viac dozvedieť na stránke: https://www.fmed.uniba.sk/pracoviska/ustav-simulacneho-a-virtualneho-medicinskeho-vzdelavania/zamestnanci.

medicínskeho kurikula a je čoraz preferovanejším trendom vyučovania medicíny. Vzhľadom na to, že na LF UK prebieha aj medziodborové štúdium BMF, môžu sa naši študenti oboznámiť s vybavením ÚSVMV a súčasne čerpať z expertízy vyučujúceho, ktorý má bohaté skúsenosti nielen z biofyzikálneho, ale aj z aplikovaného biomedicínskeho výskumu[15]. Na simulačnej výučbe a projektových úlohách sa podieľa aj MUDr. Silvia Hnilicová, PhD. Vyučujúca sa zaslúžila o zavedenie metód aktívneho vyučovania a tímového učenia (TBL[16]) s využitím moderných vyučovacích prvkov a prístupov (debrífing, kahoot[17] ai.) na LFUK a v rámci projektu KEGA koordinuje aplikáciu a prispôsobenie metódy TBL pre študijné programy BMF a BF.

Simulačné metódy umožňujú využitím umelej reprezentácie reálneho sveta dosiahnuť vzdelávacie ciele prostredníctvom zážitkového učenia. V prípade medicínskej simulácie sa jedná o využitie simulačných postupov s cieľom vzdelávania medicínskych pracovníkov a s cieľom pripraviť ich na situácie, ktorým môžu čeliť v praxi. Navyše, pozitívnym výstupom simulácie je okrem nácviku aj tréning optimalizovania krokov potrebných na stabilizovanie zdravotného stavu pacienta, čo následne zníži počet nehôd a chybných rozhodnutí pri starostlivosti o reálneho pacienta.

Technologickým jadrom ÚSVMV sú pacientske trenažéry a simulátory. Trenažéry sú konštrukčne jednoduchšie a slúžia na nácvik konkrétneho úkonu ako sú napr. paže na nácvik odbery krvi alebo zavádzania katétrov, modely na nácvik spinálnej punkcie, ultrazvukové fantómy a pod. O úroveň vyššie sa nachádzajú pacientske simulátory. Jedná sa o celotelové figuríny, prostredníctvom ktorých je možné okrem partikulárnych úloh nacvičovať najmä komplexnú starostlivosť o pacienta. Okrem nepopierateľnej atraktívnosti pre študentov, a neraz aj pre samotných vyučujúcich, je ich najväčším prínosom realistickosť technického prevedenia umožňujúca v dobrom priblížení naprogramovať fyziologický stav pacienta podľa potrieb vyučovacej hodiny alebo kurzu. Nakoľko sa pri práci so simulátormi vychádza z reálnych klinických prípadov, účastníkom simulovanej výučby je tak umožnené precvičiť si úkony spojené so stabilizáciou pacienta, záchranou jeho života, prípade konkrétnych úkonov, s ktorými sa môžu stretnúť v reálnej medicínskej praxi.

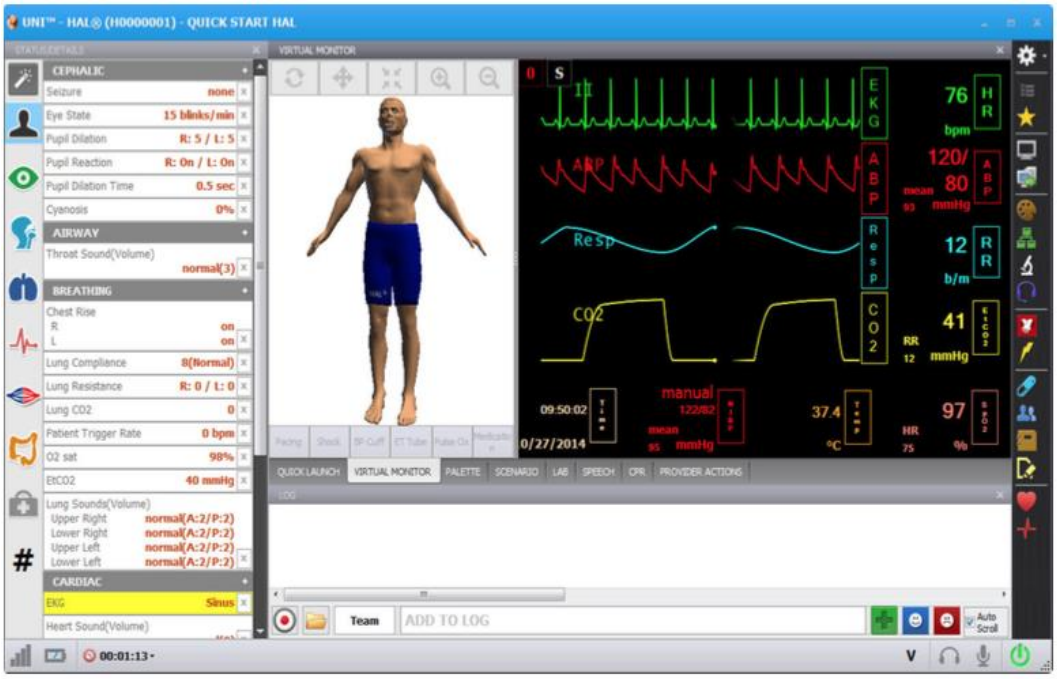
Scenáre, ktoré vyučujúci pripravujú pre študentov, sú dvojakého typu – lineárne alebo rozvetvené. Lineárne scenáre sú jednoduchšie a ich úlohou je vysvetľovať logiku riešenia konkrétneho pacientskeho stavu. Jednotlivé

[15] Simulačnú výučbu medikov zabezpečujú lekári za technickej podpory vedúceho a zamestnancov ÚSVMV.

[16] Pred detaily pozri: https://www.teambasedlearning.org.

[17] Pre detaily pozri: https://kahoot.com.

medzikroky sú prediskutované a vysvetlené vyučujúcim, pričom učiteľ sa krok za krokom dopracuje do finálneho vyriešenia pacientovho stavu. Rozvetvené scenáre sú analógiou rozhodovacích algoritmov, kedy ma účastník simulácie v každom kroku možnosť vybrať si z dvoch alebo viac možností. Na každú voľbu nadväzujú ďalšie možnosti, pričom iba jedna z ciest vedie k správnemu riešeniu. Samozrejmosťou je otvorenosť scenárov, kedy osoba riadiaca simulácie vie meniť podmienky simulácie, prípadne promptne reflektovať na reakcie účastníkov simulácie, ktoré nie sú zahrnuté v scenári a scenár tým v reálnom čase upravovať.



Obr. 2 Ovládací program UNI® pre pacientský simulátor HAL3101 (Gaumard, USA) (*Zdroj: vlastné spracovanie*)

Nevyhnutnou súčasťou simulátora je počítač s nainštalovaným programom, prostredníctvom ktorého sa nastavujú jednotlivé parametre a vytvárajú scenáre (obr. 2). Rastúca výpočtová rýchlosť počítačov umožnila implementovanie matematických modelov fyziológie človeka, t.j. po zadaní základných (bazálnych) hodnôt vstupných premenných dokáže program vypočítať aj ostatné hodnoty, respektíve v prípade dynamických scenárov (napr. môžeme zadať, aby sa dychová frekvencia na danom časovom úseku znížila o 10%) sa podľa modelu menia aj ostatné parametre, pokiaľ to nie je scenárom zadané inak.

Druhou nevyhnutnou súčasťou je pacientský monitor zobrazujúci hodnoty vitálnych funkcií ako sú frekvencia dýchania, srdcová frekvencia, tlak krvi, teplota a pod. (obr. 3). V súčasnosti je štandardom mať ako dotykový monitor dotykový

počítač a manuálne nastavovať parametre, ktoré majú účastníci simulácie vidieť a posúdiť.

Demonštrovanie prednášaných tém na simulátoroch je neoceniteľné v situáciách, kedy je zabezpečenie priamej praktickej skúsenosti limitované napríklad vzácnym typom ochorenia. Opakovaným tréningom na simulátoroch sa tak v účastníkoch simulácie vytvorí pamäťová stopa, ktorá má pozitívny dopad na ich rýchle a správne lekárske rozhodovanie v prípade, že sa v budúcnosti s už natrénovanou situáciou stretnú. Dnes už je dokázané, že retencia vedomostí získaná využitím hravej a intuitívnej simulačnej formy vzdelávania je vyššia v porovnaní s vedomosťami získanými na prednáškach, seminároch alebo samoštúdiom.



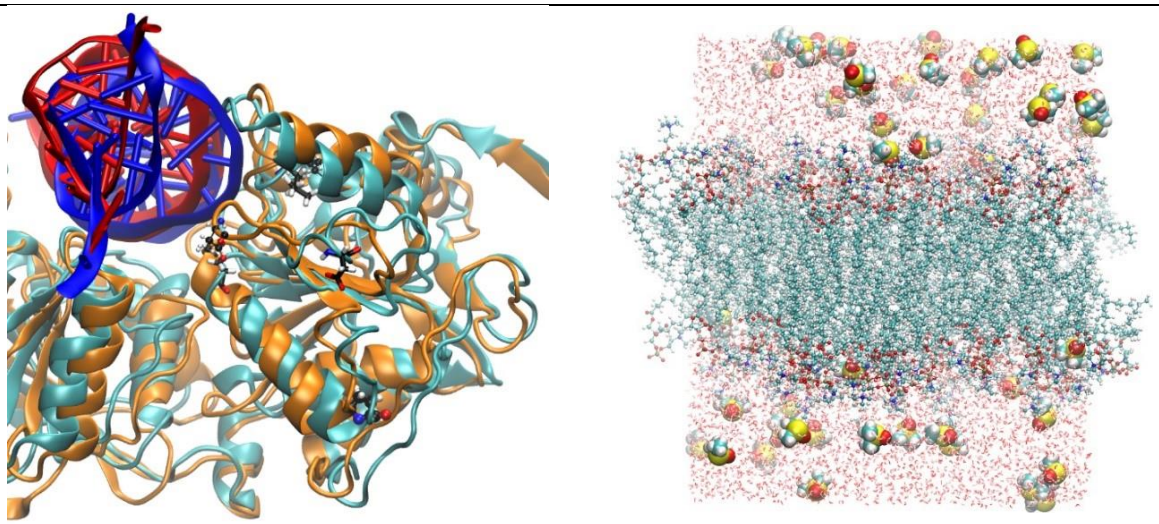
Obr. 3: Pacientský monitor zobrazujúci hodnoty vitálnych funkcií nastavených pre pacientsky simulátor iStan (CAE Healthcare, USA) (*Zdroj: vlastné spracovanie*)

3.4 Vedecko-výskumné aktivity vyučujúcich, ktoré sa premietajú do výučby praktických predmetov a do tém záverečných prác

Aktivity v oblasti chemickej fyziky (RNDr. Ing. Milan Melicherčík, PhD).

Vyššie spomínaná molekulová dynamika sa využíva aj vo vedeckej práci tvorby modelov molekulových systémov a ich následných simuláciách. Pri tom sa používajú dostupné atomárne štruktúry proteínov alebo nukleových kyselín, ktoré sa vyčistia od zvyškov kryštalizačných prímesí, prípadne sa doplnia atómy vodíkov, ktoré niektoré metódy (napr. röntgenová kryštalografia) nedokážu zobraziť. V prípade, že je záujem o porovnanie pôvodného a mutovaného proteínu, vytvoria sa bodové mutácie aminokyselín (alebo sa odstráni časť proteínu – podľa toho, k akej mutácii došlo). Druhú časť modelu tvorí použité silové pole, ktoré určuje interakcie medzi dvojicami atómov. Toto pole zásadným spôsobom vplýva na presnosť a použiteľnosť simulácií, preto sa jeho tvorbe venuje veľká pozornosť. V súčasnosti sa používa niekoľko rodín (*Amber, Gromos, OPLS*), ktoré sa po identifikovaní nejakej nedostatočnosti postupne vylepšujú

(napr. po dlhých simuláciách). Tieto metódy autori použili na simulovanie viacerých biosystémov (Strawn et al., 2010; Pandey et al., 2014). Pri arginínovom represore sa im pomocou týchto simulácií podarilo vysvetliť, že aj keď je systém tvorený šiestimi rovnakými podjednotkami, prvá molekula ligandu sa viaže s inou afinitou, ako zvyšných päť. Ďalšie simulácie rovnakého systému z iných druhov ukázali, že aj keď ide o rovnaký systém, jeho regulácia je zabezpečená úplne iným spôsobom (Pandey et al., 2020).



Poznámka: Graf naľavo zobrazuje štrukturálne zmeny vyvolané mutáciou Ser382Pro v proteíne XPB. Pôvodný stav je znázornený oranžovou a modrou farbou, mutant modrozelenou a červenou farbou. Graf napravo ukazuje stav na konci simulácie membrány z fosfolipidu DMPC vo vodnom roztoku s pridaným rozpúšťadlom DMSO. Niektoré z molekúl DMSO sú naviazané na povrch membrány.

Obr. 4: (vľavo) Štrukturálne zmeny vyvolané mutáciou Ser382Pro v proteíne XPB, **(vpravo)** stav na konci simulácie membrány z fosfolipidu DMPC vo vodnom roztoku s pridaným rozpúšťadlom DMSO (*Zdroj: vlastné spracovanie v programe VMD, Humphrey et al., 1996*)

Iným prípadom použitia je štúdia opravných mechanizmov nukleových kyselín bunky, konkrétne vplyv existujúcich mutácií proteínu XPB (súčasť TFIIH komplexu, ktorý sa zúčastňuje prepisu DNA do RNA, ako aj opráv DNA pomocou vystrihnutia chybných nukleotidov). Tam najvážnejšia mutácia Ser382Pro spôsobila zmenu polohy slučky medzi Ser/Pro382 a Val405 a prerušenie interakcií s molekulou DNA. Na obr. 4 vľavo je znázornený stav na konci 1500ns dlhej simulácie. Veľkosť zmeny je znázornená zobrazením aminokyseliny s najväčšou výchylkou od pôvodného stavu. To spôsobilo oslabenie interakcií medzi proteínom a DNA a zvýšené fluktuácie DNA (článok je v príprave). Rovnako sa MD simulácie použili na objasnenie nameraných údajov vplyvu DMSO na lipidové membrány počas ich ochladzovania (obr. 4 vpravo, článok je v príprave). Podľa koncentrácie DMSO dochádza k rozdielu v množstve naviazaného DMSO na povrch membrány.

*Aktivity v oblasti medicínskej fyziky (*RNDr. Milan Zvarík, PhD., RNDr. Marcela Morvová, PhD., prof. Libuša Šikurová, CSc., doc. RNDr. Iveta Waczulíková, PhD.)

Spektrum vedecko-výskumných aktivít v tejto oblasti je veľmi široké. Časť tímu sa zoberá hľadaním biomarkerov rakovinových ochorení v telových tekutinách s využitím *state-of-the-art* experimentálnych techník. Na analýzu výsledných „veľkých dát" využíva prostredie R. Vstupné dáta do analýz predstavujú získané absorpčné a fluorescenčné charakteristiky telových tekutín. Okrem hodnotenia jednotlivých spektrálnych charakteristík sa tím zameriava na komplexnejší prístup, čo znamená hľadanie akéhosi „podpisu" (*fingerprint*) rakovinového ochorenia v telových tekutinách. Na analýzu tím využíva rôzne metódy jednorozmerných a viacrozmerných regresných analýz ako logistická regresia alebo O/PLS-DA, ktoré budú v blízkej budúcnosti doplnené o modely využívajúce strojové učenie. Veľké dáta sú produkované aj zobrazovacími prístrojmi používanými v klinickej praxi pre určenie diagnózy, terapeutického efektu a prognózy pacienta. Ďalšou výskumnou oblasťou je charakterizovanie fyzikálnych a biologických interakcií syntetických polymérnych nanočastíc testovaných na s cieľom ich využitia v humánnej medicíne. Príspevky študentských vedeckých prác venovaných spomínaným problematikám sú uvedené kapitole 3.5.

Aktivity v oblasti matematickej biológie (doc. Mgr. Pavol Bokes, PhD.)

Tieto aktivity boli opísané v kapitole 2.1.

V nasledujúcej kapitole je bližšie predstavený publikačný výstup doktorandky Iryny Zabaikiny (Zabaikina et al., 2020).

3.5 Vedecko-výskumné výstupy študentov

Analýza a selekcia rádiomických prvkov funkčného obrazu získaného pozitrónovou emisnou tomografiou

Výsledky sú súčasťou úspešne obhájenej a ocenenej bakalárskej práce autorky Bc. Nataše Brisudovej, ktorú vytvorila pod odborným vedením doc. MUDr. Sone Balogovej, PhD., prednostky Kliniky nukleárnej medicíny LF UK a OÚSA, a doc. RNDr. Ivety Waczulíkovej, PhD.[18]

Pri manažmente rakoviny sú pre detekciu nádorov, staging a charakterizáciu používané viaceré zobrazovacie metódy, medzi nimi aj počítačová tomografia (CT) a pozitrónová emisná tomografia (PET). Študentka sa zamerala na

[18] Rozhovor s Natašou Brisudovou bol uverejnený v časopise Naša Univerzita, v rubrike Príbeh študenta pod názvom „Škola mi dodala dôveru v samú seba". Rozhovor je dostupný na: https://uniba.sk/fileadmin/ruk/nasa_univerzita/NU2021-22/NU_202109_WEB_1_.pdf.

rádiologické znaky obrazov PET-CT pacientov so spondylodiscitídou (infekciou stavcov a medzistavcových platničiek, 40 nálezov) a pacientov s kostnými metastázami (40 nálezov). Získavanie a rekonštrukcia obrazu je náročná a komplexná úloha, študentka pre výpočet konvenčných, textúrnych a tvarových prvkov diagnostických obrazov študentka využila *open-source softvér LIFEx*. Na výber vhodných atribútov zvolila algoritmy strojového učenia, funkciu Boruta z knižnice Boruta, kde sa používa iteratívny algoritmus a využíva pritom klasifikátor *Random forest* (tab. 1).

Metóda *Train-test split* (70/30): 70% dát pre natrénovanie a vybudovanie modelu a 30% pre testovanie vyhodnotením diagnostickej výkonnosti (*accuracy*) pre diagnostiku MET a diferenciálnu diagnostiku SD. Plochy pod ROC krivkou (AUC) pre výsledné modely a metódy výberu tréningových a testovacích dát sú uvedené v tab. 2. Modelom s najlepšou prediktívnou hodnotou bol získaný metódou *Random forest* s *Train-test split* metódou výberu tréningových a testovacích dát s presnosťou 83.3 % (AUC = 98.6 %; 95% CI: 95%-100%).

Tab. 1 Metódy algoritmov a metódy výberu tréningových a testovacích dát *(Zdroj: Brisudová, 2021)*

Metóda algoritmov	*Metóda tréningových a testovacích dát*
Random forest	Train test split (70/30)
Logistic regression	Leave-One-Out Cross-Validation
Support Vector Machines	7-fold Cross Validation

Tab. 2 Plochy pod ROC krivkou (AUC) pre výsledné modely *(Zdroj: vlastné spracovanie)*

	Train Test Split	*LOOCV*	*K-fold CV*
Random forest	98.61	92.31	83.33
Logistic regression	84.03	76.25	87.5
Support Vector Machines	87.5	82.5	87.5

Preverenie výsledného *Random forest* modelu s *Train-test split*: na testovacom súbore: cut-off = 0.50; senzitivita (MET) = 98,2%; špecificita (SD) = 86,1 %; plocha pod ROC krivkou AUC = 97% (95% CI: 94,1% - 99,9%)

Študentka výsledky svojej bakalárskej práce prezentovala na československom podujatí Studentský den nukleární medicíny na ČVUT/*Student Day of Nuclear Medicine at CVUT* (Praha, ČR, 10.10.2020) a získala 1. miesto, vďaka ktorému mohla našu fakultu reprezentovať vo februári 2021 aj na medzinárodnej

konferencii *The International Conference for Healthcare and Medical Students* (Dublin, Írsko)[19].

Optimal bang–bang feedback for bursty gene expression

Doktorandka Iryna Zabaikina sa vo svojej práci spolu s kolegami (Zabaikina et al., 2020) venovala problematike stochasticity v génovej expresii a jej významu pre presnú kontrolu bunkových funkcií. Konkrétne práca rieši otázku, ako presne môže stochasticky exprimovaný proteín dosiahnuť danú cieľovú úroveň expresie. Génovú expresiu kontrolujú transkripčné faktory, ktoré sa viažu na úsek DNA pred daným génom a pozitívne alebo negatívne ovplyvňujú prepis génu do RNA. Nedávne experimenty odhalili, že transkripcia a translácia neprebieha spojite, ale v dávkach - akýchsi pulzoch, tzv. *bursts*, ktoré majú nepravidelnú periódu. Proteín, ktorý sa vytvára v dávkach, je schopný ovládať svoju expresiu prostredníctvom negatívnej spätnej slučky. Autori konkrétne uvažujú *bang - bang* typ spätnej väzby, ktorá vypne produkciu proteínu vždy, keď jeho koncentrácia prekročí daný prah. Pomocou krokového deterministického matematického formalizmu autori v práci odvodzujú explicitné výrazy pre pravdepodobnostné rozdelenie koncentrácie proteínu a pre strednú kvadratickú odchýlku od cieľovej úrovne. Použitím kombinácie analytickej a numerickej optimalizácie identifikujú optimálnu hodnotu prahu pre *bang-bang* z hľadiska minimalizácie odchýlky a skúmajú závislosť optimálnej hodnoty na cieľovej úrovni a podprahovej frekvencii dávky. Systematická analýza im umožnila formulovať viacero kvantitatívnych a kvalitatívnych záverov o kontrolovateľnosti expresie génov procesom *burst*. V závere práce autori načrtli smerovanie budúceho výskumu tejto oblasti. Práca Zabaikina et al. (2020) je voľne dostupná v plnom znení.

Príspevky študentov z podujatia 13th European Biophysics Conference – EBSA 2021 24.-28.7.2021 Austria Center Vienna, Austria

Džubinská et al. P166, Velísková et al. P182, Šutý et al. P22, Jacko et al. P272, (Garaiová, Z.), Maťko et al. P14, Oravczová et al. P 335. Abstrakty príspevkov študentov sú publikované v časopise *European Biophysics Journal (2021) 50 (Suppl 1), s. 1-226.*[20]

[19] Zborník ICHAMS Abstract and Conference Book 2021 je dostupný na: https://static1.squarespace .com/static/ 5b2ead4970e80292bad1b812/t/6038e77f707bcd40cb32effb/1614342033006/ICHAMS+ Abstract+and+Conference+Book+2021.pdf

[20] Dostupné na: https://link.springer.com/content/pdf/10.1007/s00249-021-01558-w.pdf.

Príspevky prezentované na XIII. ročníku Interaktívnej konferencie mladých vedcov - PREVEDA 2021

Brisudová et al. (príspevok bol ocenený v sekcii „Omiky“), Velísková et al. (príspevok bol ocenený v sekcii SARS-CoV-2), Džubinská et al., Šutý et al. Príspevky sú publikované na internete.[21]

3.6 Doplnenie analytických nástrojov Excel Add-ins

Pre výučbu a potreby výskumu vyučujúci naďalej využívajú doplnok *MS Excel* s názvom *BESH Stat*, ktorého autorom je bývalý kolega Mgr. Peter Slezák, PhD., spoluriešiteľ predchádzajúceho projektu KEGA 003UK-4/2012.[22] V rámci spomínaného projektu vytvoril v programovacom jazyku *Excel Visual Basic for Application* doplnok tak, aby obsahoval všetky základné štatistické testy a grafické výstupy používané v biomedicínskom výskume. V podpore výučby štatistiky pokračoval aj po odchode z akademického prostredia ako externý lektor povinne voliteľného doktorandského predmetu *Biomedicínska štatistika* na LF UK. V oblasti štatistiky naďalej spolupracuje s garantkou BMF, doc. RNDr. Ivetou Waczulíkovou, PhD., na vybraných výskumných úlohách s medicínskou problematikou. Súčasne pokračoval na rozšírení ponuky analytických nástrojov o neparametrické testy a multivariačné regresné modelovanie (viacnásobná lineárna regresia, binárna a multinomiálna logistická regresia a Coxova regresia proporcionálneho hazardu). Doplnok *BESH Stat* na webovej stránke

4 Záver

Zameranie výučby bioštatistiky a využitia softvéru na riešenie príkladov a situácií z reálneho výskumu a klinickej praxe pomáha študentom vykonať reálne analýzy dát do svojich bakalárskych, diplomových a dizertačných prác. Časť experimentálnych záverečných prác je realizovaná v spolupráci s externými pracoviskami. Pod odborným vedením sa študenti učia formulovať výskumnú otázku v teoretickej rovine a následne ju operacionalizovať v termínoch závislej a nezávislej premennej. V biológii a medicíne je často potrebné experimentálne alebo štatisticky kontrolovať vplyv iných vonkajších alebo skresľujúcich premenných. Vyučujúci preto študentov priamo zapájajú do procesu plánovania a zostavovania návrhov experimentu /štúdie a výberu analytických metód.

[21] Dostupné na: https://www.preveda.sk/conference/viewer_abstract/id=2139, https://www.preveda.sk/conference/viewer_abstract/id=2221, https://www.preveda.sk/conference/viewer_abstract/id=2218, https://www.preveda.sk/conference/viewer_abstract/id=2251.

[22] Doplnok je dostupný na stránke: http://beshstat.eu.

Prepojenie výučby s reálnym výskumom súčasne znižuje u študentov pocit izolovanosti na akademickej pôde a odtrhnutosti vzdelávania od praktického, každodenného života. Študenti získavajú sebavedomie pri prezentovaní svojej experimentálnej práce na odborných fórach[23] a možnosť presadiť sa v tímovej práci. Uvedené skutočnosti priamo vplývajú na zvýšenie ich motivácie a rozvíjanie tvorivého prístupu študentov k výskumnej práci. Znalosti a schopnosti z oblasti analýzy dát môže absolvent využiť aj pre kritické posúdenie odborných článkov a kvalifikované rozhodovanie pri implementácii nových poznatkov do výskumu a klinickej praxe (napr. zlepšenie diagnostických a terapeutických prístupov a pre predikciu rizika a ďalšieho vývoja stavu pacienta). Získané zručnosti im tiež umožňujú skvalitniť pripravované odborné články a zvýšiť úspešnosť a počet prác uverejnených v renomovaných odborných časopisoch.

5 Literatúra

Brisudová, N. (2021). *Rádiomická analýza funkčného obrazu získaného pozitrónovou emisnou tomografiou ako potenciálny biomarker*. Bakalárska práca, Fakulta matematiky, fyziky a informatiky Univerzity Komenského.

Böhm, R. et al. (2014). Use of threshold-specific energy model for the prediction of effects of smoking and radon exposure on the risk of lung cancer. *Radiation Protection Dosimetry*, 160(1-3), 100-103.

Böhm, R. et al. (2019). Lung regeneration in abstaining smokers. *Radiation Protection Dosimetry*, 186(2-3), 397-400.

Böhm, R. et al. (2020). Radon as a tracer of lung changes induced by smoking. *Risk Analysis,* 40(2), 370-384.

Bokes, P. et al. (2020). Mixture distributions in a stochastic gene expression model with delayed feedback: a WKB approximation approach. *Journal of Mathematical Biology,* 81 (1), 343-367.

Bokes, P., Singh, A. (2015). Protein copy number distributions for a self-regulating gene in the presence of decoy binding sites. *PLOS One,* 10 (3).

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of Molecular Graphics,* 14(1), 33–28.

Kollarik, B. et al. (2018). Urinary fluorescence analysis in diagnosis of bladder cancer. *Neoplasma,* 65(2), 234–241.

Kusendová, D. et al. (2018). *Vyučovanie štatistiky a demografie na Univerzite Komenského v Bratislave*. In Stankovičová, I., Medová, J. (eds.) *Zborník abstraktov a príspevkov zo slávnostnej konferencie k 50. výročiu založenia Slovenskej štatistickej a demografickej spoločnosti,* 18. – 20. jún 2018, Častá – Papiernička. ISBN 978-80-88946-82-3, str. 63-80.

[23] Príklady študentských prác uvedené v kapitole 3.5.

Motulsky, H. (2014). *Intuitive biostatistics.* 3. vyd. New York: Oxford University Press, 2014, 540 s. ISBN 987-0-19-994664-8.

Pandey, S. K. et al. (2014). Binding-competent states for L-arginine in E. coli arginine repressor apoprotein. *Journal of Molecular Modeling,* 20(7), 2330.

Pandey, S. K. et al. (2020). Conserved dynamic mechanism of allosteric response to L-arg in divergent bacterial arginine repressors. *Molecules*, 25(9), 2247.

Pekár, S., Brabec, M. (2009). *Moderní analýza biologických dat 1. Zobecněné lineární modely v prostředí R.* Praha: Scientia, 2009, 225 s. ISBN 978-80-86960-44-9.

Pekár, S., Brabec, M. (2012). *Moderní analýza biologických dat 2. Lineární modely s korelacemi v prostředí R.* Brno: Masarykova univerzita, 2009, 225 s. ISBN 978-80-210-5812-5.

Rečková, M. (2016). Štatistika má pre lekársky výskum veľký význam. *Slovenská štatistika a demografia*, 26(2), 69-72.

Somorčík, J., Teplička, I. (2015). *Štatistika zrozumiteľne.* Bratislava: Enigma, 2015, 244 s. ISBN 978-80-813-3042-1.

Strawn, R. et al. (2010). Symmetric allosteric Mechanism of hexameric escherichia coli arginine repressor exploits competition between L-arginine ligands and resident arginine residues. *PLOS Computational Biology* 6(6), e1000801.

Waczulíková, I., Slezák, P. (2015). *Introductory biostatistics.* Bratislava: Univerzita Komenského v Bratislave, 2015, 147 s. ISBN 978-80-223-3938-4.

Zabaikina, I., Bokes, P., Singh, A. (2020). Optimal bang–bang feedback for bursty gene expression. In *European Control Conference* (ECC), 277-282. https://www.biorxiv.org/content/10.1101/793638v1

6 Poďakovanie

Práca je podporená Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ Slovenskej republiky (KEGA), projekt 041UK-4/2020.

Naše poďakovanie patrí aj všetkým kolegom, doktorandom a externým spolupracovníkom, ktorí sa spolu s nami podieľajú na zabezpečení výučby a na vedení záverečných prác študentov biomedicínskej fyziky a biofyziky, ako aj bývalému kolegovi, Mgr. Petrovi Slezákovi, PhD., autorovi doplnku *MS Excel BESH Stat*.

Osobitné poďakovanie patrí našim študentom: Bc. Nataši Brisudovej, Mgr. Martine Velíškovej, Mgr. Daniele Džubinskej, Mgr. Iryne Zabaikine, Mgr. Jurajovi Jackovi a Mgr. Šimonovi Šutému za dosiahnuté výsledky a vynikajúcu reprezentáciu našej fakulty na odborných podujatiach.

Špeciálne oceňujeme výskumnú spoluprácu s tímami nižšie uvedených národných inštitúcií:

- Lekárska fakulta a Jesseniova lekárska fakulta Univerzity Komenského,
- Prírodovedecká fakulta Univerzity Komenského,
- Farmaceutická fakulta Univerzity Komenského,

- Centrum experimentálnej medicíny SAV,
- Centrum biovied a Biomedicínske centrum SAV,
- Onkologický ústav sv. Alžbety v Bratislave (OUSA),
- Národný onkologický ústav (NOÚ) a Národný onkologický inštitút (NOI),
- Národný ústav srdcových a cievnych chorôb (NÚSCH) v Bratislave,
- Premedix Academy,
- Nemocnica sv. Cyrila a Metoda v Bratislave.

Odchod doc. Ing. Vladimíra Úradníčka, Ph.D.
(* 06-09-1963 † 17-09-2021)
Smútočný oznam

Departure of doc. Vladimír Úradníček, Ph.D.
(* 06-09-1963 † 17-09-2021)
Obituary



Štatistickú komunitu na Slovensku a finančnú odbornú verejnosť nepríjemne zaskočila náhla smrť známeho reprezentanta aplikovanej štatistiky a dlhoročného a obľúbeného člena SŠDS, **doc. Ing. Vladimíra Úradníčka, Ph.D**. Vlado Úradníček pôsobil štyri funkčné obdobia v užšom Výbore SŠDS, v rokoch 2006 – 2010 a 2010 – 2014 ako podpredseda pre aplikovanú štatistiku a v rokoch 2015 – 2018 a 2019 – 2021 ako podpredseda pre akademickú štatistiku. Bol autorom a duchovným otcom banskobystrickej akcie SŠDS, konferencie FERNSTAT, ktorej názov bol (aspoň oficiálne) utvorený ako akronym Financie – Ekonomika – Riadenie – Názory – ŠTATistika. Bol tiež iniciátorom a gestorom banskobystrickej regionálnej prehliadky prác mladých štastistikov, na ktorej mali možnosť prezentovať svoje práce študenti Ekonomickej fakulty Univerzity Mateja Bela v Banskej Bystrici, a to predovšetkým študenti študijného programu *financie, bankovníctvo a investovanie* a jeho inžinierskej špecializácie *analytické metódy vo financiách*.

Jeho odchod v piatok 17. septembra 2021 bol nečakaný a smutne nadväzuje na sériu odchodov prominentných štatistických či matematických osobností spätých so životom SŠDS za posledné tri roky: Ján Luha, Belo Riečan, Jozef Chajdiak, Jozef Komorník.

Vlado Úradníček sa narodil 6. septembra 1963 v metropole východného Slovenska, Košiciach, kde prežil prvých osem rokov svojho života. Po presťahovaní sa s rodičmi prežil svoje detstvo a časť svojich štúdií v metropole západného Slovenska, Bratislave. V roku 1982, vo veku deväťnásť rokov začal študovať v metropole stredného Slovenska, Banskej Bystrici, ktorej zostal verný až do konca. Tu v roku 1986 ukončil svoje inžinierske štúdium v študijnom odbore *ekonomika zahraničného obchodu* na pracovisku Obchodnej fakulte elokovanom od Vysokej školy ekonomickej v Bratislave. Diplomová práca svojim zameraním

indikovala jeho inklináciu ku kvantitatívnym metódam. Mala názov *Vybrané metódy prognózovania zahraničných trhov*. Z banskobystrického pracoviska Obchodnej fakulty so vznikom Univerzity Mateja Bela v Banskej Bystrici sa kreovala Ekonomická fakulta a tiež neskoršie aj Fakulta financií Univerzity Mateja Bela v Banskej Bystrici. Doktorské štúdium absolvoval na Ekonomicko-správnej fakulte Masarykovej univerzity v Brne v študijnom programe *ekonomické teórie* a ukončil ho v roku 1999 obhajobou dizertačnej práce nazvanej *Gravitačné modely v cestovnom ruchu*. Jeho habilitácia sa uskutočnila v roku 2010 na Ekonomickej fakulte Univerzity Mateja Bela v Banskej Bystrici v študijnom odbore *ekonomika a manažment podniku.* Kým predošlé práce boli zamerané štatisticky, habilitačná práca s názvom *Alternatívne metódy hodnotenia výkonnosti podnikov* bola zameraná na sféru korporátnej finančnej analýzy.

Dá sa povedať, že Banská Bystrica mu bola osudová – tu začal profesijne pôsobiť ako vysokoškolský pedagóg so zapojením do riadiacich štruktúr akademickej obce na rôznych úrovniach. Ešte pred vznikom Univerzity Mateja Bela v Banskej Bystrici pôsobil v rokoch 1986 – 1992 na banskobystrickej Fakulte ekonomiky služieb a cestového ruchu Vysokej školy ekonomickej v Bratislave a od roku 1992 už na nástupnickej Ekonomickej fakulte Univerzity Mateja Bela v Banskej Bystrici (asistent, odborný asistent). Od roku 1996 do roku 2005 pôsobil na Fakulte financií Univerzity Mateja Bela v Banskej Bystrici (odborný asistent) a od roku 2005 do svojej smrti v roku 2021 na Ekonomickej fakulte Univerzity Mateja Bela v Banskej Bystrici (odborný asistent, docent). Stál pri zrode Univerzity Mateja Bela v Banskej Bystrici a v roku 1992 ako predseda dočasného akademického senátu univerzity dohliadal na voľby prvého rektora univerzity. Rovnako bol jednou z kľúčových osobností Ekonomickej fakulty v 90. rokoch (predseda fakultného akademického senátu) a jednou z formujúcich osobností Fakulty financií. Od roku 2001 pôsobil v riadiacich pozíciách na Fakulte financií (prodekan pre štúdium, prodekan pre pedagogickú činnosť a sociálnu starostlivosť o študentov a štatutárny zástupca dekana) a neskoršie Ekonomickej fakulty (prodekan pre vedeckovýskumnú činnosť a zástupca dekana v rokoch 2015 – 2019) a vlastne aj celej Univerzity Mateja Bela v Banskej Bystrici (prorektor pre pedagogickú činnosť, akreditáciu a vnútorný systém kvality od 1. marca 2019 do 17. septembra 2021).

Vlado Úradníček zohral dôležitú úlohu pri významných medzníkoch vo vysokoškolskej histórii Banskej Bystrice, aktívne sa podieľal na zrode Univerzity Mateja Bela v Banskej Bystrici, vzniku jej Fakulty financií či budovaní vnútorného systému kvality na univerzitnej úrovni v súčasnom období. Riadiace pozície nevyužíval pre svoj osobný prospech či popularitu, postupnými krokmi sa však snažil meniť akademický svet k lepšiemu. Jeho pevnosť, zásadovosť a statočnosť

sa ukázali pri zlomových udalostiach, v novembri 89 či pri preňho veľmi bolestnom procese zlučovania Fakulty financií s Ekonomickou fakultou.

Počas svojho akademického pôsobenia dokazoval nadhľad a víziu. Svojimi myšlienkami sa usiloval vylepšovať veci a vo veciach, ktoré považoval za správne, dokázal byť veľmi neoblomný. Mal zmysel pre precíznosť a pracovitosť. Mal povesť prísneho, nekompromisného a spravodlivého pedagóga a napriek svojej náročnosti na študentov bol mimoriadne obľúbený zásluhou svojich didaktických schopností. Vedel pochopiť nedokonalosť študentských riešení a odpovedí, no vždy ho trápilo, keď odhalil, že študenti nerozumejú veciam podstatným. Absolventi fakulty mu často a nevšedným spôsobom vyjadrovali vďačnosť za to, čo im pre život a kariéru dal. Aj on si na nich rád spomínal, sledoval ich životnú dráhu a s mnohými udržiaval priateľské kontakty aj po skončení ich štúdia. Veľmi mu záležalo, u seba aj u iných, aby pedagogická činnosť nevychádzala nazmar a dôsledne inovoval obsah predmetov a približoval ho nielen študentom, ale tiež potrebám praxe. Obzvlášť to bolo vidno na predmetoch *finančno-ekonomická analýza podniku* a *finančná matematika 1* či *finančná matematika 2*, na výučbu ktorých prichádzal relaxovať.

Vo vedeckovýskumnej činnosti sa fokusoval na aplikáciu kvantitatívnych metód v oblasti financií, podnikovej ekonomiky a riadenia. Jeho vedecké a odborné aktivity sa zameriavali do štyroch kľúčových oblastí: (1.) predikcia finančnej tiesne podnikov, (2.) finančná matematika a finančná analytika, (3.) oceňovanie finančných derivátov, (4.) skúmanie ekonomickej udržateľnosti a konvergencie.

Hýril aktivitou a energiou, ktoré ho čoskoro predurčili aj do významných pozícií v rámci akademickej samosprávy na fakultnej aj univerzitnej úrovni. Žil i pracoval vždy s plným nasadením, využívajúc svoju bystrosť, skvelú pamäť a nevšedný intelekt. Bavilo ho pozorovať okolitý svet a odhaľovať súvislosti faktov a udalostí, veľmi rád sa delil o zaujímavé poznatky a pozorovania so svojimi spolupracovníkmi. Bolo zážitkom počúvať v jeho podaní podrobné príbehy ľudí a udalostí. V pamäti si dlho a presne uchovával dobré i zlé a vedel z nich vyvodiť poučenie pre prítomnosť a budúcnosť. Vďaka tomu dokázal včas rozpoznať možné problémy, a tým napomôcť ich skorému riešeniu a neraz aj predísť ich vzniku.

Jeho pozornému pohľadu neunikli ani slabiny akademického prostredia, v ktorom sa pohyboval. Nebol ľahostajný voči nešvárom zo strany študentov i učiteľov, trápili ho všetky formy akademických podvodov i naše obmedzené možnosti pri ich odhaľovaní a spravodlivom potrestaní. Mal odvahu otvárať aj ťažké témy, a preto – pochopiteľne – nie vždy a každému boli jeho názory a postupy príjemné a nie každý chápal jeho pohľad na svet.

Miloval život a usiloval sa ho žiť naplno. Povestná bola jeho láska k hudobnému umeniu, za ktorým cestoval po Slovensku i do sveta. V banskobystrickej opere bol roky najvernejším divákom a opustil nás pri ceste na predstavenie do košickej opery. Aj pri prechádzkach v prírode rád počúval obľúbené árie. Hoci v súkromí žil sám, mal veľmi rád spoločnosť a veľa času venoval svojim blízkym a priateľom. Pre nich bol mimoriadne pozorným a obľúbeným spoločníkom vo chvíľach radosti aj starostí.

Vo svojej tvorivej práci skúmal finančné zdravie podnikov, no vlastné zdravie svojim životným nasadením nešetril. Odišiel tak ako žil, šokujúco náhle a nečakane, uprostred tvorivých pracovných i svojich obľúbených aktivít. Svojim životom a dielom však vyryl hlbokú a nezmazateľnú stopu v mysliach a srdciach tých, ktorí ho bližšie poznali.

Všetkých nás správa o tom, že Vlado Úradníček už nie je s nami, zaskočila. Všetci by sme boli úprimne radi, keby sme mohli čerpať z jeho optimizmu a dobrej nálady. Život by bol jednoduchší. Jeho pedagogický odkaz a spomienky na neho však s nami zostanú.

Martin Boďa
vedecký tajomník SŠDS

Miroslav Hužvár
podpredseda Akademického senátu Univerzity Mateja Bela v Banskej Bystrici

Dvojkonferencia 20SSK & 18SDK
Informácia o konaní udalosti SŠDS

Bi-conference 20SSK & 18SDK
Information about an event by SSDS

V Banskej Bystrici sa na sklonku leta uskutočnila 9. a 10. septembra 2021 významná tradičná vedecká akcia určená aj pre širšiu odbornú verejnosť, spojená ***20. slovenská štatistická konferencia*** *a* ***18. slovenská demografická konferencia***, ktorá vzhľadom na svoj duálny charakter získala familiárne označenie dvojkonferencia 20SSK & 18SDK. Ide o tradičné konferenčné podujatia organizované Slovenskou štatistickou a demografickou spoločnosťou a treba na ich margo poznamenať, že sú pre spoločnosť významnou súčasťou jej vedeckej identity a zároveň sú jedinými podujatiami, ktoré sa podarilo zrealizovať v roku 2021 na prezenčnej báze. Slovenské štatistické konferencie spoločnosť organizuje od roku 1969 a spočiatku boli na nepravidelnej báze, ale od štvrtého do devätnásteho ročníka v rokoch 1988 – 2018 sa konferencie konali bienálne. Dvadsiaty ročník bol pôvodne plánovaný na rok 2020, ale v dôsledku nepriaznivej spoločenskej situácie bol preložený na rok 2021, čo umožnilo spojiť konanie štatistickej konferencie s demografickou. Prvá slovenská demografická konferencia sa uskutočnila v roku 1975 a od svojho štvrtého ročníka od roku 1993 do súčasnosti je tiež tradičnou konferenciou spoločnosti konanou v pravidelnom dvojročnom cykle, ktorý nebol narušený spoločenskými udalosťami.

Dvojkonferenciu 20SSK & 18SDK zorganizovala spoločnosť v spolupráci so Štatistickým úradom Slovenskej republiky a Ekonomickou fakultou Univerzity Mateja Bela v Banskej Bystrici a konala sa v priestoroch banskobystrického pracoviska Štatistického úradu Slovenskej republiky na Sídlisku. Napriek sprísňujúcemu sa režimu indikovaného tmavnúcim covid automatom sa zásluhou organizačného tímu pod vedením Ing. Zlaty Jakubovie, CSc., generálnej riaditeľky Sekcie zberu a spracovania dát v priemysle a terénnych zisťovaní, podarilo zabezpečiť zdarný priebeh podujatia v priateľskej a nie príliš formálnej atmosfére. Dvojkonferencia mala napriek nepriaznivým podmienkam relatívne vysokú účasť so 72 zúčastnenými.

Konferenčný program bol rozvrhnutý na dva dni. Vo štvrtok 9. septembra 2021 konferenciu otvorila predsedníčka spoločnosti doc. Ing. Iveta Stankovičová, PhD., ktorá privítala účastníkov a odovzdala slovo predsedovi Štatistického úradu Slovenskej republiky Ing. Alexandrovi Ballekovi. Úvodnú sekciu zavŕšila

prezentácia Ing. Zlaty Jakubovie, CSc., ktorá predstavila Banskobystrický kraj cez prizmu štatistiky a čísel.

Vedecký program konferencie bol rozdelený do štyroch samostatných sekcií. Skôr štatisticky tematicky orientované dve sekcie sa uskutočnili vo štvrtok 9. septembra 2021 a v piatok 10. septembra 2021 sa najprv konala sekcia venovaná aktuálnej cenzálnej problematike sčítania obyvateľov, bytov a domov 2021 (SOBD 2021) organizovanej na Slovensku v tomto roku a demograficky zamerané príspevky boli súčasťou poslednej sekcie.

V rámci prvej sekcie odzneli príspevky sledujúce širšie rozpätie tém. Prvý príspevok sa zaoberal kritickým štatistickým zhodnotením tendenčne prezentovanej epidemiologickej situácie (predseda Českej štatistickej spoločnosti Ondřej Vencálek). Druhý príspevok hodnotil vplyv a dosah pandémie koronavírusu v roku 2020 na štruktúru a úroveň úmrtnosti na Slovensku a Česku (Branislav Šprocha, Infostat). Úmrtnosť a príčiny smrti na Slovensku za ostatných päť rokov boli obsahom tretieho príspevku (Veronika Krišková, Štatistický úrad Slovenskej republiky) a koncept aktívneho starnutia bol témou štvrtého príspevku (Alena Kaščáková, Ekonomická fakulta Univerzity Mateja Bela v Banskej Bystrici). Posledný piaty príspevok ukazoval na dopad európskej legislatívy na výberové zisťovania medzi slovenskými domácnosťami (Róbert Vlačuha, Štatistický úrad Slovenskej republiky).

V druhej sekcii boli prezentované štyri príspevky. V prvom príspevku boli predstavené výsledky aplikovanej štúdie z modelovania zákazníckeho potenciálu maloobchodných predajní v regióne Turiec (Pavol Ďurček, Prírodovedecká fakulta Univerzity Komenského v Bratislave). Ďalší príspevok sa týkal odbornej prípravy medicínskeho fyzika a urobil prierez aplikácií štatistiky v medicíne (Iveta Waczulíková, Fakulta matematiky, fyziky a informatiky Univerzity Komenského v Bratislave). Posledné dva príspevky sa zaoberali postupne metodológiou výpočtu nákladov na deti a možnosťami ich pokrytia bez zaťaženia štátneho rozpočtu (Raman Herasimau, Ekonomický ústav Slovenskej akadémie vied) a následne záverečnými výstupmi z projektu aplikovaného výskumu zaberajúceho sa rozvojom podnikateľskej etiky v slovenskom podnikateľskom prostredí (Iveta Stankovičová, predsedníčka Slovenskej štatistickej a demografickej spoločnosti).

Rokovanie vo štvrtok bolo zavŕšené slávnostným rautom a panelovou diskusiou, na ktorej sa mohli obnovovať či utužovať kontakty medzi účastníkmi konferencie z akademickej sféry aj zo sféry štátnej štatistiky, ktoré – ako sa ukázalo – neutrpeli dištancom vynucovaným počas pandemického obdobia.

Ďalšie dve sekcie sa uskutočnili predpoludním v piatok 10. septembra 2021. Rámcový motív tretej sekcie dvojkonferencie boli organizačné záležitosti SOBD

2021 a prezentovali na nej pracovníci Štatistického úradu Slovenskej republiky, ktorí boli do cenzusu zaangažovaní. Prvý príspevok predstavil koncept sčítania a aplikačné rozhrania využívané pri sčítaní (Ľudmila Ivančíková, Lucia Vanišová, Štatistický úrad Slovenskej republiky) a druhý príspevok informoval o spôsobe, akým sa cenzus musel vysporiadať s harmonizáciou údajov o počte obyvateľov z rôznych zdrojov (Martin Kočiš, Štatistický úrad Slovenskej republiky). Zvyšné dva príspevky tejto sekcie postupne sumarizovali skúsenosti z prípravy a realizácie cenzu na krajských pracoviskách (Renáta Dušová, Štatistický úrad Slovenskej republiky) a vyhodnocovali mediálne aktivity a nástroje uplatňované počas projektu sčítania unikátnou formou (Jasmína Stauderová, Štatistický úrad Slovenskej republiky).

Záverečná sekcia bola demografická a počas nej sa realizovalo päť prezentácií. Prvý príspevok predkladal scenár demografického vývoja v slovenskej populácii do roku 2100 (Boris Vaňo, Štatistický úrad Slovenskej republiky). Ďalšie dva príspevky boli na tému reprodukčného správania a analyzovali transformáciu rodinného a reprodukčného správania slovenskej populácie (Branislav Šprocha, Infostat) alebo sledovali demografické aspekty mimomanželskej plodnosti žien v slovenskej populácii (Viera Pilinská, Infostat). Predposledný príspevok konfrontoval trendy rozvodovosti na Slovensku a v Česku pred rokom 2019 a po ňom (Alžeta Garajová, Prírodovedecká fakulta Univerzity Komenského v Bratislave) a posledný príspevok, ktorý vyvolal intenzívny záujem o diskusiu, navrhoval rôzne možnosti pre využitie matematických metód v demografickej analýze (Attila Tóth, Fakulta stredoeurópskych štúdií Univerzity Konštantína Filozofa v Nitre).

Konferenciu uzavrela predsedníčka spoločnosti doc. Ing. Iveta Stankovičová, PhD., a účastníci mali možnosť sa následne zúčastniť fakultatívnej prehliadky historickým centrom Banskej Bystrice, po ktorom ich sprevádzala Lýdia Bulíková.

Na webovej stránke konferencie zriadenej na webovom sídle spoločnosti http://www.ssds.sk/sk/zoznamkonferencii/74/20ssk/ je k dispozícii zborník abstraktov, fotogaléria a pdf prezentácie väčšiny príspevkov.

Možno skonštatovať, aj vzhľadom na tematický priebeh dvojkonferencie, aj vzhľadom na privátne reakcie jej účastníkov, že dvojkonferencia 20SSK & 18SDK bola úspešnou akciou spoločnosti.

Martin Boďa
vedecký tajomník SŠDS

Publikačná etika a vyhlásenie k nekalým praktikám

Redakčná rada Forum Statisticum Slovacum (ďalej len RR FSS) sa v plnom rozsahu stotožňuje s princípmi etiky publikovania deklarovanými Výborom pre pulikačnú etiku. RR FSS prijíma štandardy publikačnej etiky a zavádza opatrenia proti akejkoľvek publikácii s nekalými praktikami. Autori predkladajú svoje práce do časopisu na publikovanie s vyhlásením, že predložené práce sú príspevky autorov a neboli kopírované alebo plagizované z iných prác. Editori posudzujú rukopisy na základe ich intelektuálneho obsahu, bez ohľadu na rasu, pohlavie, sexuálnu orientáciu, náboženské vyznanie, etnický pôvod, štátnu príslušnosť alebo politickú filozofiu autorov. RR FSS očakáva, že všetky strany podieľajúce sa na FSS dodržia publikačnú etiku. RR FSS nebude tolerovať plagiátorstvo alebo iné neetické správanie a neuverejní rukopis, ktorý nespĺňa tieto normy.

Zodpovednosť autorov: Autori potvrdzujú, že ich rukopis je ich pôvodnou prácou, že nebol doteraz uverejnený v rovnakej podobe a zároveň nie je v súčasnej dobe podaný na uverejnenie inde. Autori musia bezodkladne oznámiť RR FSS všetky strety záujmov. Autori musia uviesť všetky zdroje použité pri tvorbe svojho rukopisu. Autori musia bezodkladne nahlásiť všetky chyby, ktoré objavia vo svojom rukopise RR FSS.

Zodpovednosť recenzentov: Recenzenti musia oznámiť RR FSS všetky strety záujmov. Recenzenti musia uchovávať informácie týkajúce sa rukopisu ako dôverné. Recenzent musí upozorniť predsedu RR FSS na informácie, ktoré môžu byť dôvodom na zamietnutie vydania rukopisu. Recenzenti posudzujú rukopisy len na základe ich intelektuálneho obsahu.

Zodpovednosť RR FSS: RR FSS musí uchovávať informácie o podaných rukopisoch ako dôverné. RR FSS musí posúdiť rukopisy len na základe ich intelektuálneho obsahu. RR FSS rozhoduje o zverejnení predložených rukopisov. RR FSS odmietne príspevok, ktorý nie je v súlade s požiadavkami etiky publikovania.

Publication ethics and malpractice statement

The Editorial Board of Forum Statisticum Slovacum (hereinafter abbreviated as EB/FSS) is fully associated with the principles of ethics of publication declared by the Committee of Publication Ethics. The EB/FSS is committed to upholding the highest standards of publication ethics and takes all possible measures against any publication malpractices. Authors submitting their works to the journal for publication as original articles attest that the submitted works represent their authors' contributions and have not been copied or plagiarized in whole or in part from other works. An editor at any time evaluate manuscripts for their intellectual content without regard to race, gender, sexual orientation, religious belief, ethnic origin, citizenship, or political philosophy of the authors. The EB/FSS expects all parties participating in the publication of Forum Statisticum Slovacum commit to these publication ethics. The EB/FSS does not tolerate plagiarism or other unethical behaviour and will remove any manuscript that does not meet these standards.

Author responsibilities: Authors certify that their manuscripts are their original work unpublished previously elsewhere in the same form and not currently being considered for publication elsewhere. Authors must notify the EB/FSS of any conflicts of interest without any delay. Authors must identify all sources used in the creation of their manuscript. Authors must report any errors they discover in their manuscript to the EB/FSS without any delay.

Reviewer responsibilities: Reviewers must notify the EB/FSS of any conflicts of interest. Reviewers must keep information pertaining to the manuscript confidential. Reviewers must bring to the attention of the head of EB/FSS any information that may be reason to reject publication of a manuscript. Reviewers must evaluate manuscripts only for their intellectual content.

Editorial responsibilities: The EB/FSS must keep information pertaining to submitted manuscripts confidential. The EB/FSS must evaluate manuscripts only for their intellectual content. The EB/FSS is responsible for making publication decisions for submitted manuscripts. EB/FSS refuses the manuscript, which is not in accordance with the requirements of publishing ethics.

Obsah / Table of contents

Odborné a vedecké články / Instructive and regular papers

Rozcestník moderních přístupů směřujících k automatizované detekci anomálií v časových řadách
A guidepost to modern approaches directed towards automated anomaly detection in time series
Jitka Bartošová, Vladislav Bína, Vladimír Přibyl 1

Aplikácia analýzy nákupného koša v oblasti CRM pomocou softvéru R
Application of market basket analysis in the field of CRM using the R software
Tadeáš Chujac, Martina Jantová, Iveta Stankovičová 12

Do pharmaceutical companies profit from the Covid19 crisis?
Profitujú farmaceutické firmy z koronakrízy?
Peter Pšenák, Michal Páleník 28

Cesta do hlubin DATA kroku programového systému SAS aneb co by měl každý programátor o systému SAS vědět%
Journey into the depths of the SAS DATA step or what every programmer should know about SAS%
Roman Pavelka 39

Výsledky sčítaní obyvateľov a historický vývoj časovania sobášnosti na národnej a regionálnej úrovni na Slovensku
Census results and historical development of marriage timing at the national and regional level in Slovakia
Branislav Šprocha, Pavol Tišliar 52

Zmeny plodnosti vydatých žien na Slovensku v časovej a priestorovej perspektíve
Changes in fertility of married women in Slovakia in a temporal and spatial perspective
Branislav Šprocha, Pavol Tišliar 62

Štatistika v medicíne – úloha štatistiky v príprave medicínskeho fyzika
Statistics in medical physics – the role of statistics in the training of a medical physicist
Iveta Waczulíková, Pavol Bokes, Radoslav Böhm, Milan Zvarík, Milan Melicherčík, Marcela Morvová, Silvia Hnilicová, Pavol Vitovič 77

% = odborný článok / instructive paper

Zo života SŠDS / From the life of SŠDS

Odchod doc. Ing. Vladimíra Úradníčka, Ph.D. (* 06-09-1963 † 17-09-2021) (smútočný oznam)
Departure of doc. Vladimír Úradníček, Ph.D. (* 06-09-1963 † 17-09-2021) (obituary)
Martin Boďa, Miroslav Hužvár 99

Dvojkonferencia 20SSK & 18SDK (informácia o konaní udalosti SŠDS)
Bi-conference 20SSK & 18SDK (information about an event by SSDS)
Martin Boďa 103

ISSN 1336-7420

Cena / price: 25 €
Ročné predplatné / annual subscription: 50 €
Publikované / published: December 2021