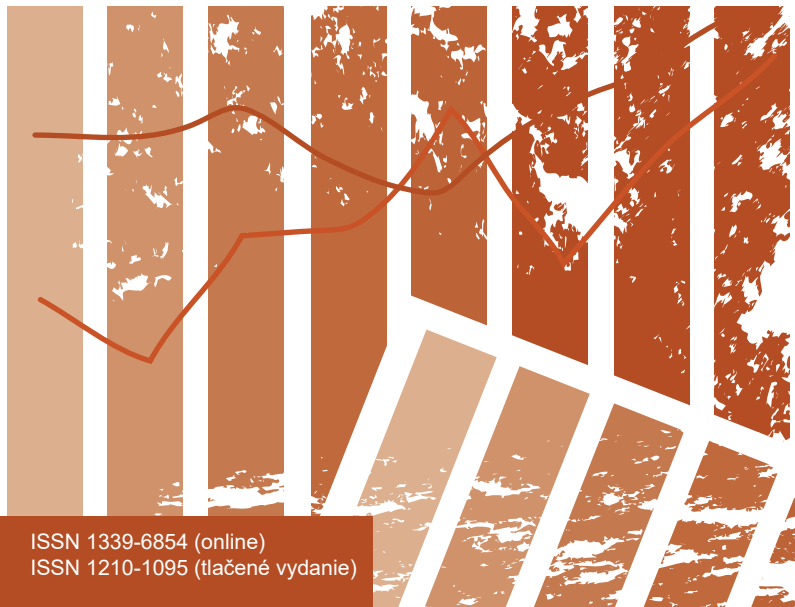
Štatistický úrad Slovenskej republiky
The Statistical Office of the Slovak Republic

SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

vedecký časopis/scientific journal

1/2024
ročník 34



ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY

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

SLOVENSKÁ ŠTATISTIKA A DEMOGRAFIA

Recenzovaný vedecký časopis založený v roku 1991. Jednotlivé čísla časopisu zverejňujeme aj v elektronickej podobe na ssad.statistics.sk a na slovak.statistics.sk. Názory autorov článkov sa nemusia zhodovať s názormi vydavateľa.

SLOVAK STATISTICS AND DEMOGRAPHY

The scientific peer-reviewed journal founded in 1991. Individual copies of the journal are available to readers in electronic form at the websites ssad.statistics.sk and slovak.statistics.sk. The opinions of the authors do not necessarily correlate with the opinions of the publisher.

Adresa redakcie/Address of Editorial Office
Slovenská štatistika a demografia
Štatistický úrad SR
Lamačská cesta 3/C, 840 05 Bratislava 45
Slovenská republika

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

ssad.statistics.sk
www.statistics.sk

OBSAH/CONTENTS

Vážení čitatelia,

rýchlo sa meniaca spoločenská realita sa dotýka takmer každej oblasti, vedecký časopis Slovenská štatistika a demografia nevynímajúc. Po 33 rokoch dochádza k významnej udalosti, keď časopis, ktorý bol doteraz vydávaný Štatistickým úradom SR bude od prvého čísla v roku 2024 vydávaný spoločne so Slovenskou štatistickou a demografickou spoločnosťou.

Naším cieľom a motivátorom spojenia bola a je snaha sústrediť sily na rozvoj tak štatistiky, ako aj demografie na Slovensku. Chceme prispieť k šíreniu znalostí medzi odborníkmi v oboch oblastiach a spájať ich prostredníctvom informácií zo štatistického a demografického sveta. Chceme osloviť mladých autorov – odborníkov a vedcov so spracovaním a zdieľaním inovatívnych tém. Spoločne s predstaviteľmi Slovenskej štatistickej a demografickej spoločnosti diskutujeme aj o ďalších zmenách, napr. vydávanie článkov v anglickom jazyku.

Želám preto redakcii veľa dobrých vedeckých článkov ako výsledok nášho spoločného úsilia.

Ing. Peter PEŤKO, MBA
predseda Štatistického úradu Slovenskej republiky

Milí čitatelia,

Slovenská demografická a štatistická spoločnosť (SDŠS) vznikla v roku 1968 (28. 3. 1968) pri Slovenskej akadémii vied so sídlom v Bratislave. Pod zmeneným názvom Slovenská štatistická a demografická spoločnosť (SŠDS) pôsobí už od roku 1990 (14. 3. 1990) ako dobrovoľné výberové združenie vedeckých a odborných pracovníkov z oblasti štatistiky, demografie a iných príbuzných disciplín. Spoločnosť organizuje vedecké a odborné podujatia s tematikou štatistiky a demografie, ako sú konferencie, semináre, prednášky, workshopy, diskusné popoludnia a podobne. Webové sídlo spoločnosti je dostupné na adrese www.ssds.sk.

Spoločnosť vydávala od roku 2005 recenzovaný vedecký časopis Forum Statisticum Slovacum (FSS). Archív časopisu je možné nájsť na http://fss.ssds.sk/sk_archiv_casopisu.html. Spočiatku boli ročníky časopisu spojené s akciami spoločnosti (ročníky I až XI) a jednotlivé ročníky obsahovali 2 až 8 čísiel. Od XII. ročníka (od roku 2016) časopis začal vychádzať pravidelne dvakrát ročne (v júni a decembri) a nebol už zviazaný s akciami SŠDS. Po XIX. ročníkoch časopis FSS prechádza ďalšou významnou zmenou. Rozhodli sme sa spojiť sily so Štatistickým úradom SR a vydávať jeden slovenský vedecký časopis venovaný štatistike a demografii. Dúfame, že kvantita článkov, ale hlavne ich kvalita týmto spojením vzrastie. Vydávanie nášho vedeckého časopisu Forum Statisticum Slovacum v roku 2024 pozastavíme a naše sily budeme venovať vedeckému časopisu Slovenská štatistika a demografia, ktorý doteraz samostatne vydával Štatistický úrad SR.

doc. Ing. Iveta STANKOVIČOVÁ, PhD.
predsedníčka Slovenskej štatistickej a demografickej spoločnosti

Dear readers,

The rapidly changing social reality affects almost every area, including the scientific journal Slovak Statistics and Demography. After 33 years, a significant event is taking place, when the Journal, which has been published up to now by the Statistical Office of the SR, will be from the No. 1 in 2024 onwards published jointly with the Slovak Statistical and Demographic Society.

The goal and the driving force of this connection was and is the effort to concentrate our forces on the development of both statistics and demography in Slovakia. We want to contribute to the spread of knowledge among experts in both fields and connect them through information from the statistical and demographic world. We want to reach out to the young authors - experts and scientists with processing and sharing of innovative topics. Together with the representatives of the Slovak Statistical and Demographic Society, we are also discussing further changes, e.g. publishing articles in English.

I therefore wish the editors many good scientific articles as a result of our joint efforts.

Ing. Peter PEŤKO, MBA
President of the Statistical Office of the Slovak Republic

Dear readers,

The Slovak Demographic and Statistical Society (SDSS) was established in 1968 (March 28, 1968) at the Slovak Academy of Sciences (SAS) situated in Bratislava. Under the updated name, the Slovak Statistical and Demographic Society (SSDS), has been operating since 1990 (March 14, 1990) as a voluntary selected association of scientific and professional experts in the field of statistics, demography and other related disciplines. The Society organizes scientific and professional events with the theme of statistics and demography, such as conferences, seminars, lectures, workshops, afternoon discussion etc. The Society´s website is available at www.ssds.sk.

Since 2005, the Society has been publishing the peer-reviewed scientific journal Forum Statisticum Slovacum (FSS). The archive of the journal can be found at http://fss.ssds.sk/sk_archiv_casopisu.html. Initially, the volumes of the Journal were associated with the Society's actions (volumes I. to XI.) and individual volumes contained 2 to 8 issues. From the XII. Volume (year 2016), the Journal started to be published regularly twice a year (in June and in December) and was no longer associated with the SSDS´s events. Following the XIX. Volumes, the FSS Journal is undergoing another significant change. We decided to join forces with the Statistical Office of the SR and publish one Slovak scientific journal dedicated to statistics and demography. We do hope that the quantity of articles, but especially their quality, will be enhanced thanks to this connection. From 2024, the publication of our scientific journal Forum Statisticum Slovacum will be suspended and our efforts will be devoted to the scientific journal Slovak Statistics and Demography, which until now has been published independently by the Statistical Office of the SR.

doc. Ing. Iveta STANKOVIČOVÁ, PhD.
President of the Slovak Statistical and Demographic Society

EDITORIÁL

Vážení čitatelia,

PhDr. Ľudmila Ivančíková, PhD.

špeciálne číslo časopisu Slovenská štatistika a demografia je venované projektu s názvom ***Socioekonomické aspekty Big Data***, ktorý Štatistický úrad SR realizoval a ukončil v roku 2023, ako aj jeho výstupom. Príspevky opisujú prácu s 3 novými typmi zdrojov – údajmi od mobilných operátorov, údajmi mýtneho systému a sociálnymi sieťami. Pri práci s nimi sa zároveň používali nové nástroje, predovšetkým machine learning (strojové učenie) a umelá inteligencia. Práve práca s uvedenými zdrojmi, nastavenie prístupu k nim, otázka ochrany údajov a už spomenuté nástroje, vrátane ukladania Big Data a udržateľnosť prístupu k nim boli v centre pozornosti projektu. Dôraz na samotné výstupy je sekundárny, keďže tie majú charakter experimentálnej štatistiky.

Samotný projekt je stručne opísaný v úvodnom príspevku J. Bardyho, kde autor zdôrazňuje, že práca s Big datami *je „meniaca sa oblasť poskytujúca nové zdroje dát, nové možnosti v oblasti hardvérovej a softvérovej infraštruktúry, nové oblasti využitia a potrebu výmeny skúseností doma aj v zahraničí“.*

Existencia Big Data vytvára predpoklad ich využitia v štatistickej produkcii. Často sa spomína najmä aspekt rýchlejšieho prístupu k údajom, ktoré opisujú reálny stav v spoločnosti. Treba však zdôrazniť a projekt sám to potvrdil, že na to, aby sa tento predpoklad naplnil, je potrebné nastaviť jasný a obojstranne prospešný dlhodobý vzťah s vlastníkmi zdrojov a údajov; vytvoriť použiteľnú a robustnú infraštruktúru; mať štatistikov so znalosťami, ktorí budú pracovať aj so „surovými“ údajmi, t. j. údajmi, ktoré nie sú „predpripravené“ vlastníkmi. Len tak bude možné nastaviť rámec hodnotenia kvality zdroja a uviesť zdroj postupne do procesu štatistickej produkcie.

V experimentálnej štatistike je potrebné zohľadniť aj viacročné opakovanie navrhnutých postupov a metód a niektoré jej výstupy nikdy do produkcie zaradené nebudú, resp. navrhnuté metódy nebudú v praxi použiteľné. To však neznamená, že uvedené zdroje by sme mali ignorovať. Súčasná znalosť poznania nám ponúka témy, ktoré nevieme štandardnými zdrojmi pokryť a len pri konkrétnej práci vieme identifikovať ich reálny potenciál. A v tom je prínos spomínaného projektu, ako aj jednotlivých článkov.

PhDr. Ľudmila IVANČÍKOVÁ, PhD.

Autorka je generálnou riaditeľkou sekcie sociálnych štatistík a demografie Štatistického úradu SR.

EDITORIAL

Dear readers,

A special issue of the journal Slovak Statistics and Demography is dedicated to the project entitled ***Socioeconomic Aspects of Big Data***, which was implemented and completed together with its outputs by the Statistical Office of the SR. The articles describe working with 3 new types of sources - data from mobile operators, toll system data and social networks. At the same time, new tools were used while working with them, especially machine learning and artificial intelligence. The project focused on working with the mentioned resources, access adjustment to them, the issue of data protection and the already mentioned tools, including the storage of Big Data and the access sustainability to them. The emphasis on the outputs is secondary, as they have the character of experimental statistics.

The project itself is briefly described in the introductory contribution of J. Bardy, where the author emphasizes that work with Big Data is "*a changing field providing new data sources, new possibilities in hardware and software infrastructure, novel applications and the need to exchange experience at home and abroad*".

The existence of Big Data creates a presumption for their use in the statistical production. The aspect of faster access to data that describes the real situation in society is often mentioned. However, it must be emphasized, and it was confirmed by the project, that in order to fulfill this assumption, it is necessary to set up a clear and mutually beneficial long-term relationship with the data and resource owners; create an applicable and robust infrastructure; to have knowledgeable statisticians who will also work with "raw" data, i.e. data not "pre-prepared" by the owners. Only then will it be possible to set the framework for evaluating the quality of the source and introduce the source gradually into the process of statistical production.

As part of experimental statistics, it is also necessary to take into account multi-annual repetition of the proposed procedures and methods, and some of its outputs will never be included in the production, or the proposed methods will not be applicable in practice. Although this does not mean that the mentioned sources should be ignored. The current state of knowledge gives us topics that cannot be covered with standard sources, and their real potential can only be identified through specific work, which is the benefit of the mentioned project, as well as of the individual articles.

PhDr. Ľudmila IVANČÍKOVÁ, PhD.

The author is the Director of the Social Statistics and Demography Directorate of the Statistical Office of the SR.

Peter KNÍŽAT
Štatistický úrad Slovenskej republiky, Ekonomická univerzita v Bratislave

Dagmar CEĽUCHOVÁ BOŠANSKÁ, Martin JANÍK, Filip NGUYEN
Alistiq, s. r. o.

NOVÉ DÁTOVÉ ZDROJE V ŠTATISTIKE: VPLYV INTENZITY NÁKLADNEJ DOPRAVY NA MAKROEKONOMICKÉ UKAZOVATELE

NEW DATA SOURCES IN STATISTICS: THE EFFECT OF FREIGHT INTENSITY ON MACROECONOMIC INDICATORS

ABSTRAKT

Cieľom tohto článku je overiť možnosti využitia modelov SARIMA a SARIMAX na odhad vývoja HDP Slovenska. V modeli SARIMAX boli údaje o najazdených kilometroch nákladnej dopravy využité ako dodatočná exogénna premenná. Po dekompozícii údajov slovenského elektronického mýtneho systému na jednotlivé zložky bola identifikovaná sezónna zložka a zložka rezíduí, ktoré nemožno vysvetliť ani trendom, ani sezónnymi zložkami, čo naznačuje, že časový rad údajov nie je stacionárny. Pri prispôsobovaní modelov SARIMA/SARIMAX sa na kontrolu stacionarity časových radov použil Augmentovaný Dickeyho-Fullerov test (ADF) a na transformáciu časových radov sa použila metóda kĺzavého priemeru. Primeranosť prispôsobených modelov bola potvrdená pozorovaním výsledku Ljung-Box testu. Navyše bol použitý softvér JDemetra+ na sezónnu analýzu časových radov. Podľa testu stacionárnosti časového radu použitím informačného kritéria Akaike (AIC) bol na prognózu HDP Slovenska najvhodnejší prispôsobený model SARIMAX. Výsledky odmocniny zo strednej kvadratickej percentuálnej chyby (RMSPE) a priemernej absolútnej percentuálnej chyby (MAPE) naznačujú nadradenosť modelu SARIMAX, ktorý bral do úvahy sezónne vzorce a exogénne faktory. Model SARIMAX prekonal SARIMA, pričom predpovedal hodnoty v rámci 95 % intervalu spoľahlivosti s hodnotou RMSPE 8,9 %, zatiaľ čo SARIMA mala RMSPE 17,4 %. Závery tohto príspevku nasvedčujú, že nekonvenčné zdroje údajov môžu mať vysoký potenciál využitia na odhad ekonomických indikátorov, ktoré môžu poskytnúť komplexnejší a aktuálnejší pohľad na hospodársku situáciu.

ABSTRACT

The aim of this article is to verify the possibilities of using the SARIMA and SARIMAX models to estimate the development of Slovakia's GDP. In the SARIMAX model, the data on kilometers travelled in freight transport were used as an additional exogenous variable. After decomposing the Slovak electronic toll system data into its component parts, seasonal component and enough residuals component were identified, indicating that the data time series is not stationary. In fitting the SARIMA/SARIMAX models, the Augmented Dickey-Fuller (ADF) test was used to check the time series stationarity and the Moving Average Method was used for the time series transformation. The adequacy of the fitted models was confirmed by observing the Ljung-Box test result. Moreover, the JDemetra+ software was utilized for seasonal analysis of time series. According to the stationarity test of the time series using the Akaike information criterion (AIC), the fitted SARIMAX model was the most suitable to forecast the GDP of Slovakia. The results from the Root Mean Squared Percentage Error (RMSPE) and the Mean Absolute Percentage Error (MAPE) indicate the

superiority of the SARIMAX model, which took into account the seasonality patterns and exogenous factors. The SARIMAX model outperformed the SARIMA, predicting values within the 95 % confidence interval, with a RMSPE value of 8.9 % while the SARIMA had a RMSPE of 17.4 %. The conclusions of this article indicate that non-conventional data sources can have a high potential of use for the estimation of economic indicators, that can provide a more comprehensive and up-to-date view of the economic situation.

KĽÚČOVÉ SLOVÁ
rýchle odhady, hrubý domáci produkt, nákladná doprava, telemetria, SARIMAX

KEY WORDS
flash estimates, gross domestic product, freight transportation, telemetry, SARIMAX

1. ÚVOD

V januári 2010 Slovenská republika uviedla do prevádzky mýtny systém, ktorý sa po rozšírení na všetky triedy ciest stal najdlhšou sieťou spoplatnených ciest nižšej triedy v Európskej únii. Prostredníctvom satelitnej technológie výberu mýta bolo pokrytých 17 600 km vymedzených úsekov ciest Slovenskej republiky [17]. Vďaka tomu má Slovensko unikátnu pozíciu na využívanie údajov vzniknutých počas prevádzky tohto systému na vytváranie ekonomických štatistík, keďže zachytávajú zásadnú časť cestnej prepravy tovarov.

Tieto údaje vznikajú automatizovaným spôsobom ako inherentná funkcia palubných jednotiek inštalovaných do nákladných vozidiel, ktoré využívajú spoplatnené úseky slovenskej cestnej siete patriace do elektronického mýtneho systému. Palubné jednotky zaznamenávajú aktuálne geografické údaje o vymedzených úsekoch ciest podliehajúcich mýtnej povinnosti (tzv. geo model sleduje polohu vozidla pomocou GPS, ktorá je porovnaná s uloženými údajmi v geo modeli. Ak algoritmus palubnej jednotky zistí, že vozidlo použilo vymedzený úsek podliehajúci úhrade mýta, vytvorí sa v súlade s platnou legislatívou príslušný mýtny záznam o tejto skutočnosti (tzv. mýtna udalosť). Mýtna transakcia je elektronický dátový záznam, ktorý vznikne na základe vyhodnotenia a spracovania jednej alebo kombinácie viacerých mýtnych udalostí. Mýtna transakcia obsahuje dátum a čas mýtnej udalosti, identifikáciu mýtneho úseku, identifikáciu vozidla, výšku mýta, platobný režim a ďalšie údaje. Mýtny úsek je definovaný ako súvislá časť vymedzeného úseku ciest, na ktorej sa vykonáva detekcia mýtnej povinnosti prechádzajúcich vozidiel. Mýtny úsek je spravidla vymedzený od hranice križovatky, ktorá tvorí začiatok vymedzeného úseku ciest, po hranicu križovatky, ktorá tvorí koniec vymedzeného úseku ciest, a naopak. To znamená, že každému vymedzenému úseku ciest prislúchajú spravidla dva mýtne úseky – jeden na smer tam a druhý na smer späť. Každý mýtny úsek je označený jednoznačným identifikátorom, začiatkom úseku, koncom úseku a spoplatnenou dĺžkou úseku [15]. Spoplatnená dĺžka úseku je číselný údaj v kilometroch stanovený vo vyhláške Ministerstva dopravy, výstavby a regionálneho rozvoja Slovenskej republiky č. 228/2020 Z. z., ktorou sa vymedzujú úseky diaľnic, rýchlostných ciest, ciest I. triedy a ciest II. triedy s elektronickým výberom mýta v platnom znení pre príslušný vymedzený úsek ciest a uvádza sa s presnosťou na tri desatinné miesta bez ohľadu na skutočnú fyzickú dĺžku mýtneho úseku. Na Slovensku majú povinnosť platiť mýto všetky motorové vozidlá s najväčšou technicky prípustnou celkovou hmotnosťou nad 3,5 tony alebo jazdnými súpravami s najväčšou technicky prípustnou celkovou

hmotnosťou nad 3,5 tony uvedenými v § 4 ods. 2 písm. b) a c) zákona č. 106/2018 Z. z. o prevádzke vozidiel v cestnej premávke (vozidlá kategórie M a N) okrem motorových vozidiel kategórie M1 a okrem jazdných súprav tvorených motorovým vozidlom kategórie M1 a N1.

Hlavným cieľom tohto článku je analýza nového dátového zdroja v štatistike, ktorý poskytuje veľké množstvo dát a pomocou ktorého je možné získať dáta vo veľmi krátkom časovom horizonte daného analyzovaného obdobia. Tento dátový zdroj by mohol mať v budúcnosti potenciálne využitie na rýchle odhady makroekonomických ukazovateľov. Na empirickú analýzu vplyvu týchto dát z nového zdroja bol vybraný hrubý domáci produkt (HDP) SR. Z ekonomického hľadiska môžeme predpokladať, že diaľničná doprava zachytáva len určitú časť z HDP v SR.

2. POUŽITÉ ÚDAJE

Vstupnými údajmi sú údaje o pohybe nákladných vozidiel v mýtnom systéme za roky 2018 až 2022. Dáta sa zbierajú v reálnom čase, ako sa nákladné vozidlá s hmotnosťou nad 3,5 tony presúvajú po platených úsekoch diaľnic, rýchlostných ciest a ciest I. triedy. Podstatná väčšina (expertný odhad NDS je viac ako 99 %) údajov je dostupná do 10. dňa mesiaca nasledujúceho po mesiaci, v ktorom sa zbierali. Súbor údajov obsahuje:

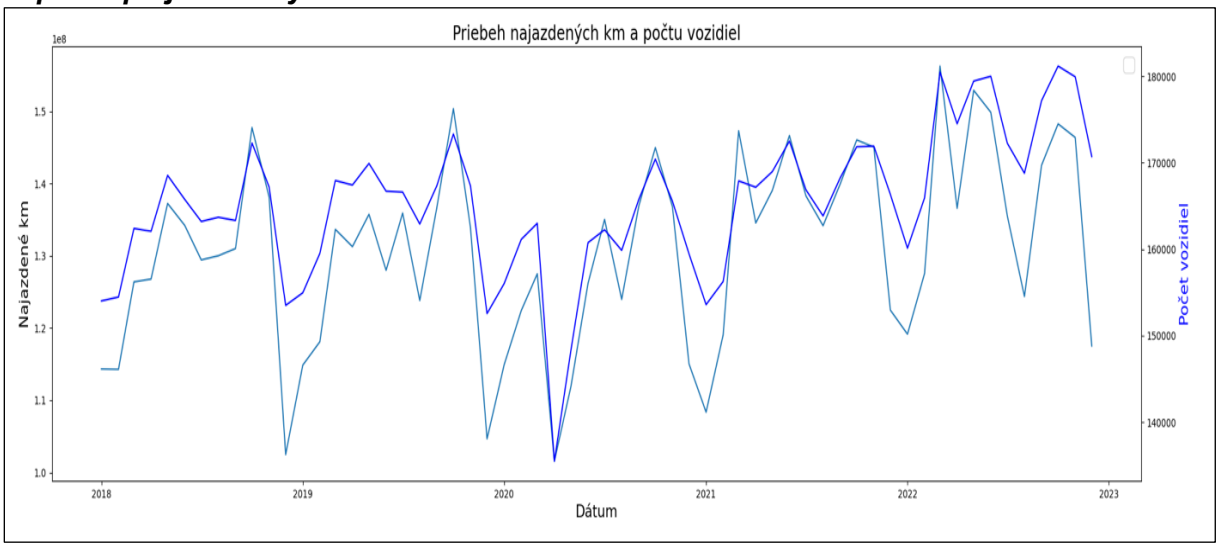
- unikátny identifikátor vozidla, aby bolo možné správne spočítať jeho prejdené kilometre a zrealizované jazdy,
- dátum a čas záznamu polohy vozidla, aby sa kilometre spočítali v správnom časovom okne,
- kategóriu vozidla, aby sa brali do úvahy len vozidlá relevantné pre nákladnú dopravu, a nie napríklad autobusy,
- identifikátor vymedzeného úseku cesty, ktorý určuje polohu vozidla pre daný záznam a spolu so smerom jazdy a ďalšími záznamami o prejdených úsekoch dovoľuje vyskladať celkovú jazdu vozidla od prvého až po posledný záznam v mýtnom systéme,
- prejdená vzdialenosť v kilometroch, ktorú vozidlo prešlo od predchádzajúceho záznamu.

Limitom pre algoritmy bol dátový súbor s odhadovanou premennou HDP v bežných cenách v mil. eur, ktorý vzhľadom na dostupnosť údajov len štvrťročne neobsahoval dostatok dátových bodov v sledovanom období 2018 až 2022. Pri makroekonomických analýzach časových radov je v praxi bežné použitie oveľa dlhšieho časového okna.

3. DÁTOVÁ ANALÝZA A PRIESKUM VSTUPNÝCH ÚDAJOV

Údaje z elektronického mýtneho systému predstavujú súbor dátových bodov zoradených v čase, teda ich možno považovať za časové rady. Obrázok č. 1 ukazuje časový rad najazdených kilometrov a počtu unikátnych vozidiel v nákladnej doprave za jednotlivé mesiace rokov 2018 až 2022. Údaje sú indexované podľa času a agregované na konci každého mesiaca každého roka. Počet najazdených kilometrov a počet vozidiel v priebehu každého roka stúpa a klesá a tento fenomén sa opakuje každý rok, čo predstavuje sezónnosť. Obrázok č. 2 obdobne znázorňuje počet najazdených kilometrov a štvrťročné HDP v bežných cenách v mil. eur, z ktorého však nie je zrejmá podobnosť týchto časových radov, na rozdiel od obrázka č. 1.

Obrázok č. 1: Priebeh počtu najazdených kilometrov a počtu vozidiel v nákladnej doprave po jednotlivých mesiacoch rokov 2018 až 2022



Zdroj: vlastné spracovanie autorov

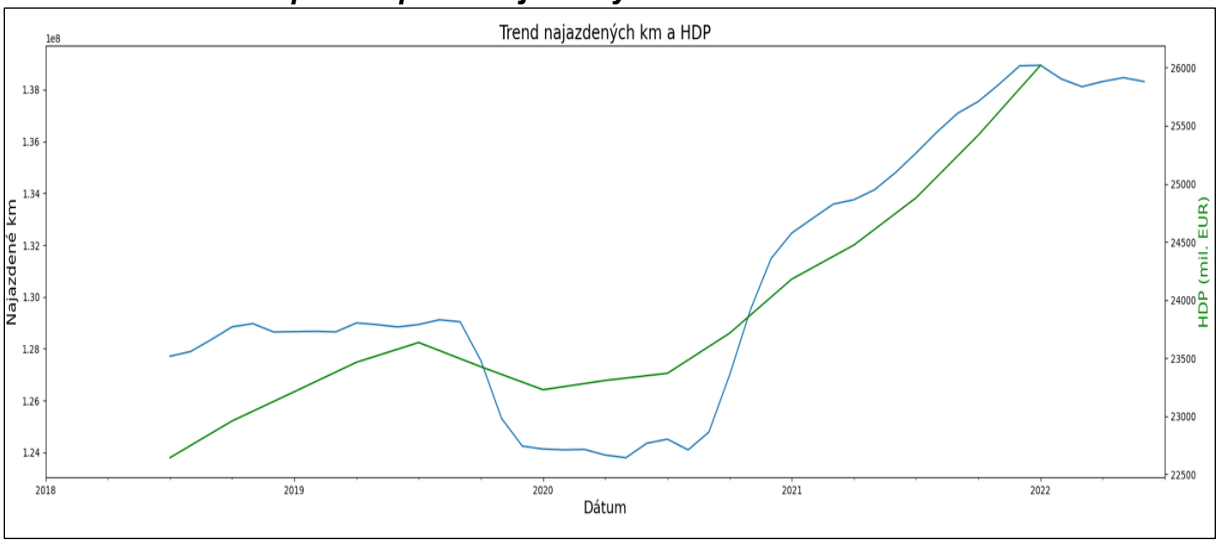
Obrázok č. 2: Priebeh počtu najazdených kilometrov v nákladnej doprave a HDP po jednotlivých mesiacoch a štvrťrokoch rokov 2018 až 2022

Zdroj: vlastné spracovanie autorov podľa údajov [21]

Časové rady možno lepšie pochopiť, keď ich rozložíme na tri zložky: trendovú, sezónnu a reziduálnu. Vizualizácia týchto zložiek časového radu sa nazýva dekompozícia a je definovaná ako štatistická úloha, ktorá rozdeľuje časový rad na jeho jednotlivé zložky [22]. Vizualizácia každej zložky dovoľuje identifikovať trend a sezónny vzorec v údajoch, čo sa nie vždy dá jednoducho rozpoznať len pri pohľade na súbor údajov (obrázok č. 1. a obrázok č. 2).

Dekompozícia mesačného počtu najazdených kilometrov a štvrťročných údajov HDP v bežných cenách v mil. eur, je znázornená na nasledujúcich obrázkoch. Pozorované údaje boli rozdelené na trend (obrázok č. 3), sezónnu zložku (obrázok č. 4) a rezíduá (obrázok č. 5). Výsledkom kombinácie týchto troch zložiek je opäť priebeh počtu najazdených kilometrov a HDP v čase (obrázok č. 2).

Obrázok č. 3: Dekompozícia počtu najazdených kilometrov a HDP – trend



Zdroj: vlastné spracovanie autorov podľa údajov [21]

Obrázok č. 3 znázorňuje celkový pozitívny trend a pokles, ktorý je pravdepodobne spôsobený pandémiou ochorenia COVID-19.

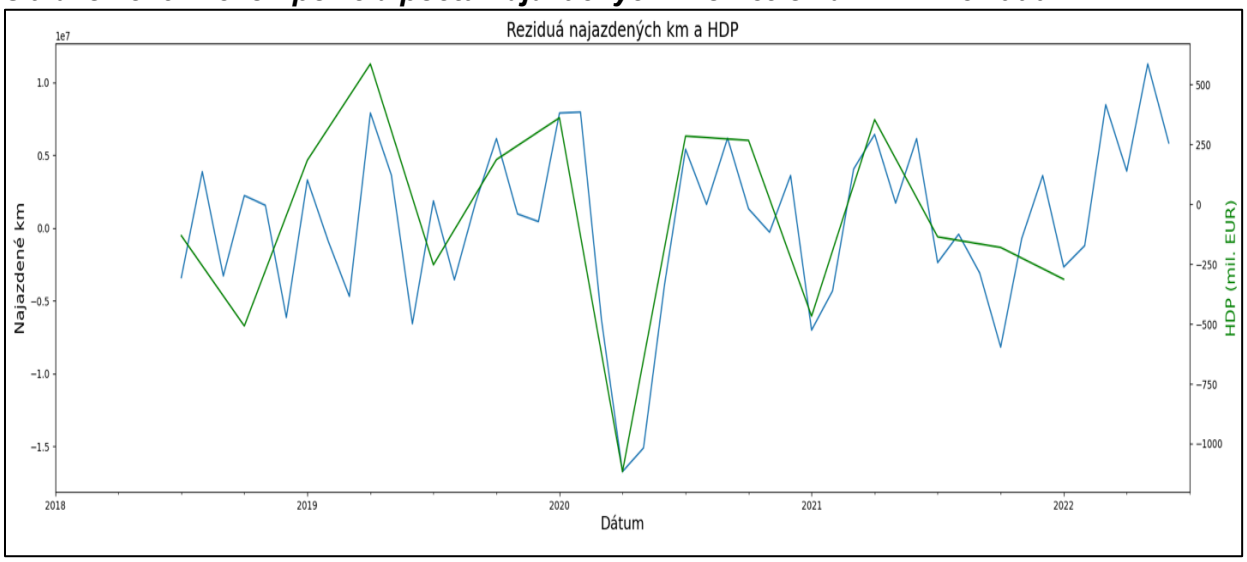
Obrázok č. 4: Dekompozícia počtu najazdených kilometrov a HDP – sezónna zložka

Zdroj: vlastné spracovanie autorov podľa údajov [21]

Sezónna zložka (obrázok č. 4) zachytáva sezónne výkyvy, ktoré predstavujú cyklus, ktorý sa vyskytuje počas kalendárneho roka. V priebehu roka začínajú najazdené kilometre a HDP na nízkej hodnote, následne rastú a na konci roka opäť klesajú.

Obrázok č. 5 zobrazuje rezíduá, ktoré nemožno vysvetliť ani trendom, ani sezónnymi zložkami. Rezíduá sa vypočítajú odčítaním grafu pozorovania (obrázok č. 2) od grafov trendu a sezónnej zložky. Rezíduá zodpovedajú náhodným chybám, označovaným aj ako biely šum, a predstavujú informácie, ktoré nevieme modelovať ani predpovedať kvôli ich náhodnosti [11].

Obrázok č. 5: Dekompozícia počtu najazdených kilometrov a HDP – rezíduum



Zdroj: vlastné spracovanie autorov podľa údajov [21]

4. CHARAKTERISTIKY NA POROVNANIE MODELOV

Na základe podobnosti trendu a sezónnosti vývoja HDP a počtu najazdených kilometrov (obrázok č. 2, obrázok č. 3, obrázok č. 4) možno predpokladať, že práve počet najazdených kilometrov predstavuje tie dodatočné poznatky o aktuálnom stave vývoja HDP, ktoré je možné získať s minimálnym oneskorením rádovo v dňoch, oproti údajom na výpočet HDP, ktoré sa získavajú rádovo v mesiacoch. Preto práve s pomocou hodnoty počtu najazdených kilometrov v danom čase je možné odhadnúť aktuálnu hodnotu HDP alebo aspoň jej vývoj (percentuálny rast, pokles alebo stagnáciu).

Výkonnosť odhadov na testovacej množine bola vyhodnotená pomocou priemernej absolútnej percentuálnej chyby (MAPE), ktorá udáva mieru presnosti odhadu pre prognostické metódy. Výhodou MAPE je ľahká interpretovateľnosť a nízka závislosť od rozsahu údajov. MAPE teda vyjadruje percentuálny podiel toho, ako veľmi sa odhadované hodnoty v priemere odchyľujú od pozorovaných alebo skutočných hodnôt, či už bola predpoveď vyššia, alebo nižšia ako pozorované hodnoty a je vypočítaná nasledovnou rovnicou [4]:

$$MAPE = \frac{1}{n} \times \sum_{t=1}^{n} \left(\frac{|e_t|}{A_t} \right) \times 100 \qquad (1)$$

V tejto rovnici je A_t skutočná hodnota v bode t v čase a e_t je chyba predpovede v bode t v čase, definovaná ako A_t - F_t, pričom F_t je predpovedaná hodnota v bode t v čase, n je jednoducho počet predpovedí. V našom prípade, keďže len odhadujeme aktuálnu hodnotu, n = 1.

Odmocnina zo strednej kvadratickej chyby (RMSE) je najbežnejšou metrikou používanou na meranie presnosti prognostického modelu časového radu. Vypočíta sa pomocou druhej odmocniny zo strednej hodnoty štvorcových rozdielov medzi skutočnými hodnotami a predpoveďami modelu. RMSE sa často uvádza v jednotkách chyba na jednotku, čo umožňuje porovnávať rôzne modely na rovnakom základe. Na porovnávanie odhadov HDP bola použitá odmocnina zo strednej kvadratickej

percentuálnej chyby (RMSPE) ako variant RMSE na percentuálne vyjadrenie, ktorá podobne ako MAPE nezávisí od rozsahu hodnôt údajov [3]:

$$RMSPE = \sqrt{\frac{\sum_{t=1}^{n}\left(\frac{e_t}{A_t}\right)^2}{n}} \times 100 \qquad (2)$$

4.1. Stacionárne procesy a procesy „random walk“

Existujú výnimočné situácie, v ktorých sa časový rad dá predpovedať len pomocou jednoduchých metód. Ide o špeciálne prípady, keď sa proces vyvíja náhodne a nedá sa predpovedať pomocou štatistických metód učenia. To znamená, že ide o proces, ktorý sa nazýva random walk a treba ho vedieť rozpoznať, aby sa dalo naplánovať ďalšie analytické modelovanie. Random walk je proces, pri ktorom je rovnaká šanca, že časový rad bude stúpať alebo klesať o náhodné číslo [22].

4.2. Testovanie stacionarity

Stacionárny časový rad je taký, ktorého štatistické vlastnosti sa v čase nemenia – má konštantný priemer, rozptyl a autokoreláciu a tieto vlastnosti sú nezávislé od času [17]. Mnohé prognostické modely predpokladajú stacionaritu a je možné ich použiť len vtedy, ak je overené, že údaje sú skutočne stacionárne. V opačnom prípade je potrebné časový rad transformovať.

Bežným testom stacionarity je takzvaný augmentovaný Dickeyho-Fullerov test (ADF), ktorý overuje nulovú hypotézu H_0: „v časovom rade existuje jednotkový koreň“. Alternatívna hypotéza znie, že jednotkový koreň neexistuje, a preto je časový rad stacionárny. Výsledkom tohto testu je ADF štatistika, ktorou je záporné číslo. Čím je menšie než nula, tým silnejšie je zamietnutie nulovej hypotézy [5] a pri jeho implementácii v štatistických programoch aj *p*-hodnota. Nulová hypotéza sa zamietne, ak je *p*-hodnota menšia ako 0,05. Testovanie stacionarity časového radu HDP sa realizovalo nasledujúcim spôsobom:

- transformácia časového radu do dátovej štruktúry časového radu,
- prevzorkovanie časového radu HDP zo štvrťročnej frekvencie na mesačnú. Pre chýbajúce hodnoty bol použitý takzvaný „forward filling“ – hodnoty pre daný kvartál sa použili pre každý chýbajúci mesiac v kvartáli[1],
- čistenie časového radu od nečíselných hodnôt,
- ADF testovanie.

Výsledná štatistika testu sa rovná -0,091, *p*-hodnota sa rovná približné 0,95 pri počte použitých oneskorení 12, čiže časový rad nie je stacionárny a bolo ho potrebné v ďalšom kroku transformovať.

4.3. Transformácia časového radu

Transformácia je matematická manipulácia s údajmi, ktorá stabilizuje ich strednú hodnotu a rozptyl, čo odstraňuje alebo znižuje vplyv trendu a sezónnosti a tým sa údaje stávajú stacionárnymi. Najjednoduchšou transformáciou, ktorú možno použiť, je diferenciácia, ktorá zahŕňa výpočet zmeny od jedného časového okamihu k druhému.

[1] Pre forward filling „chýbajúcich hodnôt“ bola využitá funkcia dataframe,ffill() v knižnici pandas v jazyku python.

Pri jednorazovom diferencovaní sa použije diferenciácia prvého rádu. Pri druhom použití by išlo o diferencovanie druhého rádu. Na získanie stacionárneho radu často nie je potrebné diferencovať viac ako dvakrát [17]. V prípade rýchleho odhadu HDP bola vybraná metóda na korekciu pri sezónnych dátach cez odčítanie kĺzavého priemeru za kvartál – teda za štyri oneskorenia.

Po aplikácii transformácie na časový rad sa opakovalo testovanie stacionarity pomocou ADF testu na určenie, či je potrebné aplikovať ďalšiu transformáciu, aby sa časový rad stal stacionárnym. Výsledkom testu bolo konštatovanie stacionarity časového radu.

4.4. Autokorelačná funkcia

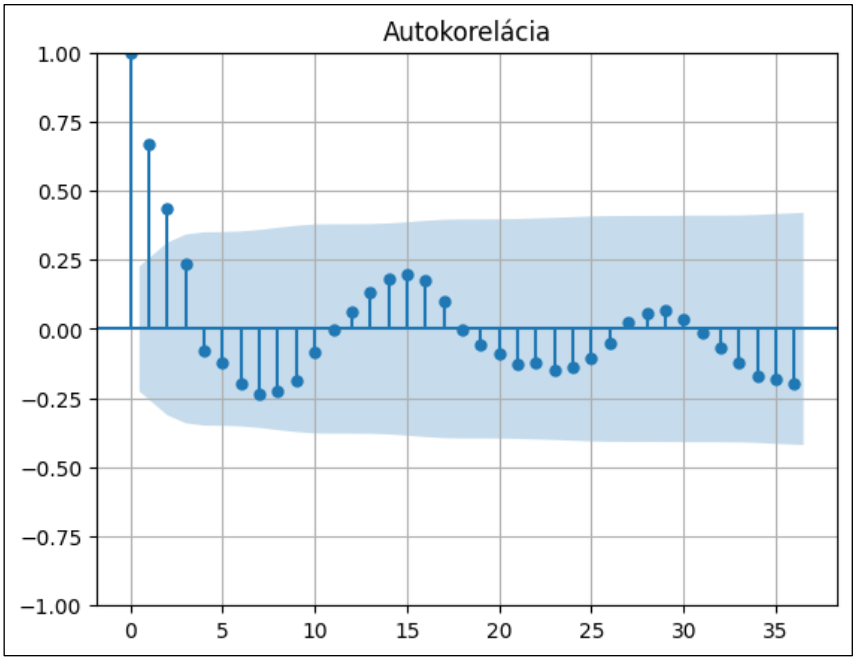
Po potvrdení stacionarity bolo použité vykreslenie autokorelačnej funkcie (ACF) na vylúčenie „random walk“. Autokorelácia meria lineárny vzťah medzi oneskorenými hodnotami časového radu (čiže tej istej premennej v rámci časového radu, ale v inom čase). ACF teda odhaľuje, ako sa korelácia medzi ľubovoľnými dvoma hodnotami mení s rastúcim oneskorením [8].

V časovom rade HDP je oneskorenie jednoducho počet časových krokov, ktoré delia dve hodnoty. Ak neexistuje autokorelácia, tak ide o časový rad, ktorý zodpovedá random walk.

ACF v prípade prítomnosti trendu ukáže, že koeficienty sú vysoké pri krátkych oneskoreniach a s rastúcim oneskorením lineárne klesajú. Ak sú údaje sezónne, tento graf bude takisto zobrazovať cyklické vzory [8].

V časovom rade HDP existuje autokorelácia (obrázok č. 6), keďže s rastúcim oneskorením sa jej hodnota vyvíja v sínusoide, takže nie je nulová. Z toho vyplýva, že sa nejde o random walk. Vysoký koeficient je vo funkcii len pri nulovom oneskorení, pri vyššom oneskorení sa už nenachádzajú žiadne výrazné koeficienty.

Obrázok č. 6: Autokorelácia časového radu HDP v mil. eur



Zdroj: vlastné spracovanie autorov podľa údajov [21]

5. ANALYTICKÉ MODELY

Z výsledkov analýzy časového radu je zrejmé, že ide o časový rad, ktorý je nestacionárny a ktorý obsahuje sezónnosť. Na podobné prognózy slúži model SARIMA a jeho obmena SARIMAX na zakomponovanie externej premennej, ktorá súvisí s modelovaným časovým radom [9].

Na výber optimálneho modelu sa použilo Akaikeho informačné kritérium (AIC). AIC odhaduje kvalitu modelu v porovnaní s ostatnými modelmi. Vzhľadom na to, že pri prispôsobení modelu dôjde k strate určitej informácie, AIC kvantifikuje relatívne množstvo informácie, ktorú model stratil. Čím menej informácií sa stratí, tým nižšia je hodnota AIC a tým lepší je model. **Výber modelu podľa AIC umožňuje udržať rovnováhu medzi zložitosťou modelu a jeho dobrou zhodou s údajmi.** AIC je z definície funkciou počtu odhadovaných parametrov *k* a maximálnej hodnoty funkcie vierohodnosti modelu, ako je uvedené v rovnici: $AIC = 2k - 2\ln(\hat{L})$.
AIC kvantifikuje kvalitu modelu len vo vzťahu k iným modelom. Je to teda relatívna miera kvality [2].

5.1. Model SARIMA

Autoregresný integrovaný kĺzavý priemer (SARIMA) je kombináciou autoregresného procesu AR(*p*), integrácie I(*d*) a procesu kĺzavého priemeru MA(*q*). Rovnako ako proces ARMA, aj proces ARIMA vychádza z predpokladu, že súčasná hodnota závisí od minulých hodnôt, ktoré pochádzajú z časti AR(*p*), a minulých chýb, ktoré pochádzajú z časti MA(*q*). Namiesto pôvodného radu označeného ako y_t však proces ARIMA používa diferencovaný rad označený ako y'_t, ktorý mohol byť diferencovaný viac ako raz. Podobne ako v procese ARMA, rád *p* určuje, koľko oneskorených hodnôt radu je zahrnutých do modelu, zatiaľ čo rád *q* určuje, koľko oneskorených chybových členov je zahrnutých do modelu. Rád *d* je definovaný ako rád integrácie. Integrácia je jednoducho opačný postup ako diferenciácia. Rád integrácie sa teda rovná počtu diferencií, ktoré boli vykonané, aby sa rad stal stacionárnym. Ak rad diferencujeme raz a stane sa stacionárnym, potom *d = 1*. Ak rad diferencujeme dvakrát, aby sa stal stacionárnym, potom *d = 2*.

O časovom rade, ktorý možno urobiť stacionárnym použitím diferencovania, sa hovorí, že je integrovaným radom. V nestacionárnom integrovanom časovom rade môžeme na tvorbu prognóz alebo odhadov použiť model ARIMA(*p, d, q*). Zjednodušene povedané, model ARIMA je jednoducho model ARMA, ktorý možno použiť na nestacionárne časové rady. Zatiaľ čo model ARMA(*p,q*) vyžaduje, aby bol rad pred prispôsobením modelu ARMA(*p, q*) stacionárny, model ARIMA(*p, d, q*) možno použiť na nestacionárne časové rady. Treba však nájsť rád integrácie *d*, ktorý zodpovedá minimálnemu počtu diferencií, ktoré sa musia vykonať, aby sa rad stal stacionárnym. Keď *d = 0*, je model ekvivalentný modelu ARMA(*p, q*). To tiež znamená, že na to, aby boli rady stacionárne, nebolo potrebné ich diferencovať [9].

Pridaním sezónnych javov v časových radoch ako ďalšej vrstvy zložitosti k modelu ARIMA získame model SARIMA. Keďže SARIMA vnáša do modelu sezónnosť ako parameter, je výrazne výkonnejší ako ARIMA pri predpovedaní komplexných časových radov obsahujúcich sezónne cykly. Význam sezónnosti je celkom zrejmý aj pre časový rad HDP či počet najazdených kilometrov a ARIMA túto informáciu implicitne nezachytáva, čo by sa prejavilo na presnosti tohto modelu. Modely SARIMA dokážu túto informáciu zachytiť a zodpovedajúcim spôsobom upraviť prognózy. Napriek tejto

výhode majú modely SARIMA v porovnaní s modelmi ARIMA aj niektoré nevýhody. Jednou z nich je, že vyžadujú viac parametrov na odhad, čo môže zvýšiť zložitosť a výpočtové náklady modelu. Ďalšou nevýhodou je, že nemusia dobre fungovať, keď údaje majú nesezónne trendy alebo štrukturálne zmeny, ako sú zmeny v správaní spotrebiteľov alebo trhových podmienkach. Modely SARIMA predpokladajú, že sezónne cykly sú stabilné a konzistentné v čase, čo nemusí byť v prípade niektorých údajov reálne [6]. V týchto prípadoch môžu byť modely ARIMA flexibilnejšie a robustnejšie.

Sezónny autoregresný integrovaný kĺzavý priemer je model SARIMA $(p,d,q)(P,D,Q)_m$, ktorý pridáva ďalšiu sadu parametrov umožňujúcich zohľadniť periodické zákonitosti pri prognózovaní časového radu, čo nie je vždy možné pri modeli ARIMA(p, d, q). Ide o štyri nové parametre v modeli, pričom prvé tri P,D,Q majú rovnaký význam ako v modeli ARIMA(p, d, q), ale sú to ich sezónne ekvivalenty. Parameter m znamená frekvenciu. V kontexte časového radu je frekvencia definovaná ako počet pozorovaní za cyklus a dĺžka cyklu závisí od súboru údajov. Pri údajoch, ktoré boli zaznamenané každý rok, štvrťrok, mesiac alebo týždeň, sa za dĺžku cyklu považuje jeden rok. Ak sa údaje zaznamenávali ročne, $m = 1$, pretože za rok je len jedno pozorovanie. Ak sa údaje zaznamenávali štvrťročne, $m = 4$, pretože v roku sú štyri štvrťroky, a teda štyri pozorovania za rok. Samozrejme, ak sa údaje zaznamenávali mesačne, $m = 12$. A napokon pri týždenných údajoch je $m = 52$. P je rád sezónneho procesu AR(P), D je rád sezónnej integrácie a Q je rád sezónneho procesu MA(Q). Model SARIMA$(p, d, q)(0,0,0)_m$ je ekvivalentný modelu ARIMA(p, d, q).

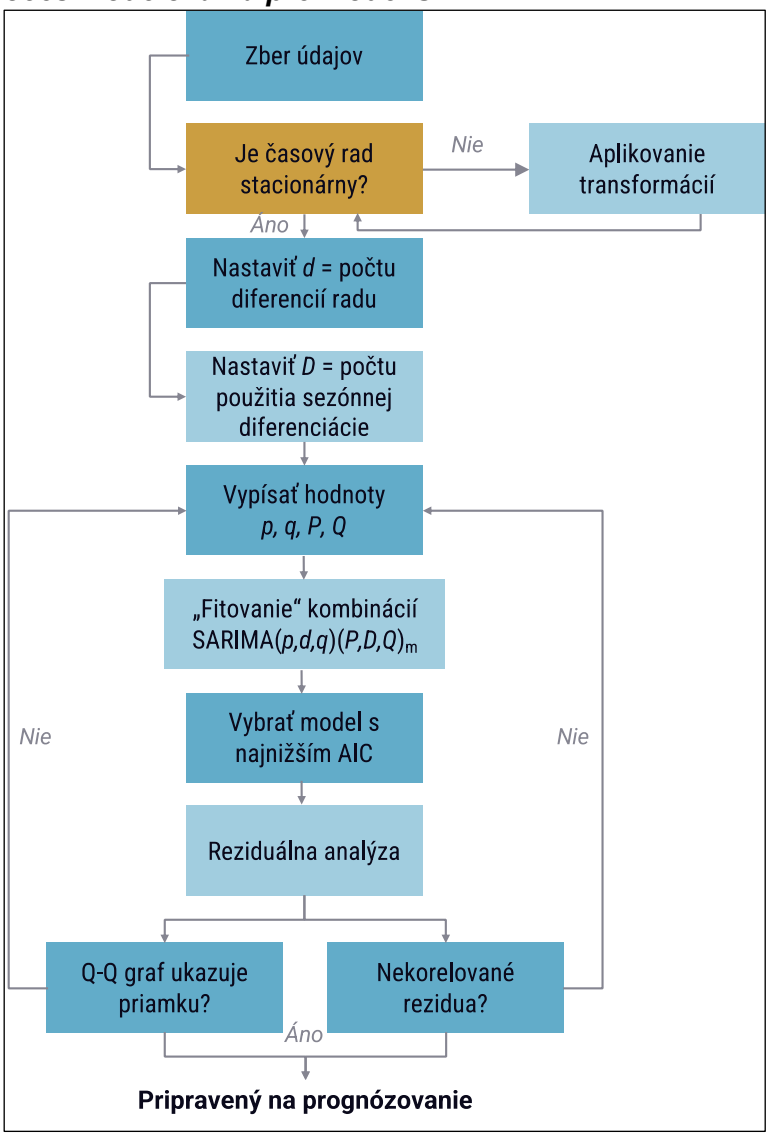
Obrázok č. 7 znázorňuje proces, ktorý je potrebný dodržať pri modelovaní. Prvý krok zberu údajov zostáva nedotknutý. Následne sa kontroluje stacionárnosť a aplikuje transformácia, aby sa stanovil parameter d. Môže sa však vykonať aj sezónna diferenciácia, aby bol rad stacionárny, a D sa bude rovnať minimálnemu počtu aplikovania sezónnej diferenciácie (I v SARIMA).

Potom sa nastaví rozsah možných hodnôt p, q, P a Q, keďže model SARIMA môže zahŕňať aj poradie sezónnych autoregresných a sezónnych kĺzavých priemerov. Pridaním týchto dvoch nových parametrov sa zvýši počet jedinečných kombinácií modelov SARIMA$(p, d, q)(P,D,Q)_m$, ktoré je možné prispôsobovať. Následne sa vyberie model s najnižším AIC a vykoná sa analýza rezíduí pred použitím modelu na prognózovanie [2]. Kvalitatívna časť analýzy rezíduí sa vykonáva pomocou Q-Q grafu. Q-Q graf je graf kvantilov dvoch rozdelení oproti sebe. Pri prognózovaní časových radov sa vykresľuje rozdelenie rezíduí na osi y oproti teoretickému normálnemu rozdeleniu na osi x. Tento grafický nástroj umožňuje posúdiť vhodnosť vybraného modelu. Ak sa rozdelenie rezíduí podobá normálnemu rozdeleniu, Q-Q graf znázorňuje priamku ležiacu na $y = x$. To znamená, že model je dobre prispôsobený, pretože rezíduá sú podobné bielemu šumu. Na druhej strane, ak sa rozdelenie rezíduí líši od normálneho rozdelenia, zobrazí sa na Q-Q grafe zakrivená priamka. Potom je možno konštatovať, že vybraný model nie je dobre prispôsobený, pretože rozdelenie rezíduí sa nepodobá normálnemu rozdeleniu, a preto rezíduá nie sú podobné bielemu šumu [13].

Hoci je Q-Q graf rýchlou metódou na posúdenie kvality vybraného modelu, táto analýza zostáva subjektívna. Preto je vhodné analýzu rezíduí ďalej podporiť

kvantitatívnou metódou použitím Ljungovho-Boxovho testu. Po analýze Q-Q grafu a zistení, že rezíduá sú približne normálne rozdelené, možno použiť Ljungov-Boxov test, aby sa preukázalo, že rezíduá nie sú korelované. Dobrý model má rezíduá, ktoré sú podobné bielemu šumu, takže rezíduá by mali byť normálne rozdelené a nekorelované. Ljungov-Boxov test je štatistický test, ktorý určuje, či sa autokorelácia skupiny údajov významne líši od 0. Pri prognózovaní časových radov je uplatňovaný Ljungov-Boxov test na rezíduá modelu, s cieľom otestovať, či sú podobné bielemu šumu. Nulová hypotéza hovorí, že údaje sú nezávisle rozdelené, čo znamená, že neexistuje autokorelácia. Ak je *p*-hodnota väčšia ako 0,05, nie je možné zamietnuť nulovú hypotézu, čo znamená, že rezíduá sú nezávisle rozdelené. Autokorelácia teda neexistuje, rezíduá sú podobné bielemu šumu a model možno použiť na prognózovanie. Ak je *p*-hodnota menšia ako 0,05, je nulová hypotéza zamietnutá, čo znamená, že rezíduá nie sú nezávisle rozdelené a sú korelované a model nie je možné použiť na prognózovanie [12].

Obrázok č. 7: Proces modelovania pre model SARIMA



Zdroj: vlastné spracovanie autorov

5.2. Model SARIMAX

Model SARIMAX ďalej rozširuje model SARIMA(*p, d, q*)(*P, D, Q*)$_m$ o vplyv exogénnych premenných. V štatistike sa termín exogénny používa na označenie prediktorov alebo vstupných premenných, zatiaľ čo pojem endogénny sa používa na definovanie cieľovej premennej – teda toho, čo sa snažíme predpovedať alebo odhadnúť v prítomnosti (rýchle odhady). To umožňuje modelovať vplyv vonkajších premenných na aktuálnu alebo budúcu hodnotu časového radu. Preto možno súčasnú hodnotu časového radu vyjadriť jednoducho ako model SARIMA(*p, d, q*)(*P, D, Q*)$_m$, ku ktorému pridáme ľubovoľný počet exogénnych premenných, ako je uvedené v nasledujúcej rovnici [16]:

$$y_t = SARIMA(p,d,q)(P,D,Q)_m + \sum_{i=1}^{n} \beta_i X_t^i \quad (3)$$

Po diferencovaní sa časový rad y_t bude označovať y'_t.

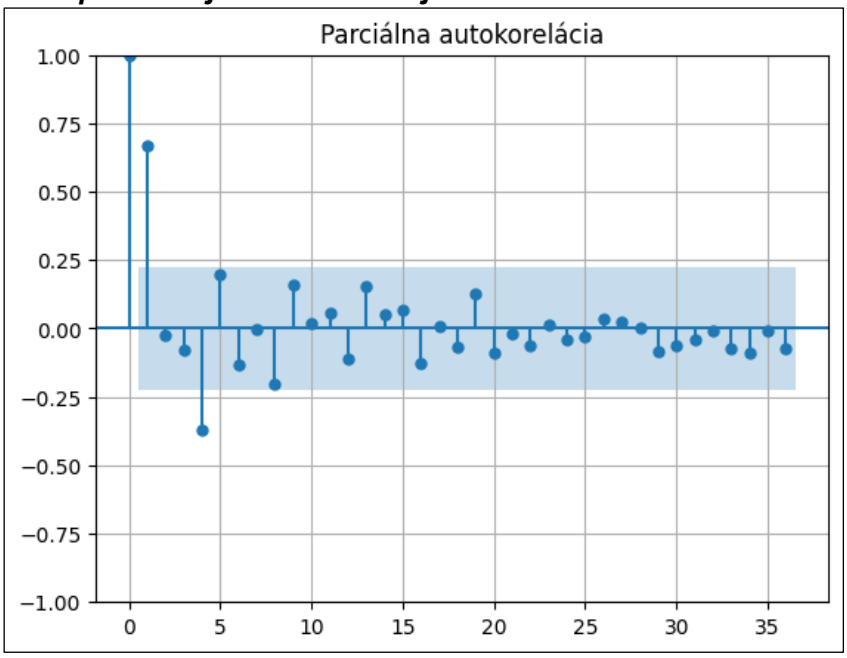
Model SARIMAX je najvšeobecnejší model na predpovedanie časových radov, ktorý umožňuje zohľadniť sezónne vplyvy, autoregresné procesy, nestacionárne časové rady, procesy s kĺzavým priemerom a exogénne premenné v jednom modeli. V dokonalej situácii modelovania sú rezíduá modelu bielym šumom. To znamená, že model zachytil všetky prediktívne informácie a zostala len náhodná fluktuácia, ktorú nemožno modelovať. Rezíduá teda musia byť nekorelované a mať normálne rozdelenie [1].

Analýza rezíduí má dva aspekty: kvalitatívnu analýzu a kvantitatívnu analýzu. Kvalitatívna analýza sa zameriava na štúdium Q-Q grafu, zatiaľ čo kvantitatívna analýza určuje, či sú naše rezíduá nekorelované. Q-Q graf sa vytvorí tak, že sa na os *y* vynesú kvantily rezíduí oproti kvantilom teoretického rozdelenia, v tomto prípade normálneho rozdelenia, na osi *x*. Výsledkom je graf rozptylu. Rozdelenie sa porovná s normálnym rozdelením, pretože sa žiada, aby rezíduá boli podobné bielemu šumu, ktorý je normálne rozdelený. Ak sú obe rozdelenia podobné, čo znamená, že rozdelenie rezíduí je blízke normálnemu rozdeleniu, Q-Q graf zobrazí priamku, ktorá približne leží na *y* = *x*. To zasa znamená, že náš model dobre zodpovedá našim údajom [17].

Parametre výsledného modelu SARIMAX sa vypočítali cez externý nástroj JDemetra+[2], z čoho vyšiel model: **SARIMA(*0, 1, 0*)(*0, 1, 1*)$_4$.** Tieto parametre sme aj čiastočne odvodili zo sezónnych cyklov v dĺžke jedného roka, pričom údaje sa zaznamenávali štvrťročne, teda *m* = *4*. Ďalej bolo potrebné časový rad HDP jedenkrát transformovať na stabilizáciu strednej hodnoty, z čoho vyplýva, že *d* sa rovná 1, a jedenkrát bolo potrebné aplikovať sezónnu diferenciáciu, z čoho takisto vyplýva, že *D* sa rovná 1. Z grafu funkcie autokorelácie (obrázok č. 6) ako aj z grafu parciálnej autokorelačnej funkcie (PACF – obrázok č. 8) vyplýva, že nemožno jednoducho určiť rády *p* a *q*, a treba prispôsobovať model ARMA, ktorý však podľa spomínaného nástroja Jdemetra+ má nulový rád *p* a *q* a pre ekvivalentnú sezónnu zložku ide o autoregresný model prvého rádu, teda *Q* = *1*.

[2] *Zdroj: https://cros-legacy.ec.europa.eu/content/software-jdemetra_en.*

Obrázok č. 8: Graf parciálnej autokorelačnej funkcie časového radu HDP v mil. eur

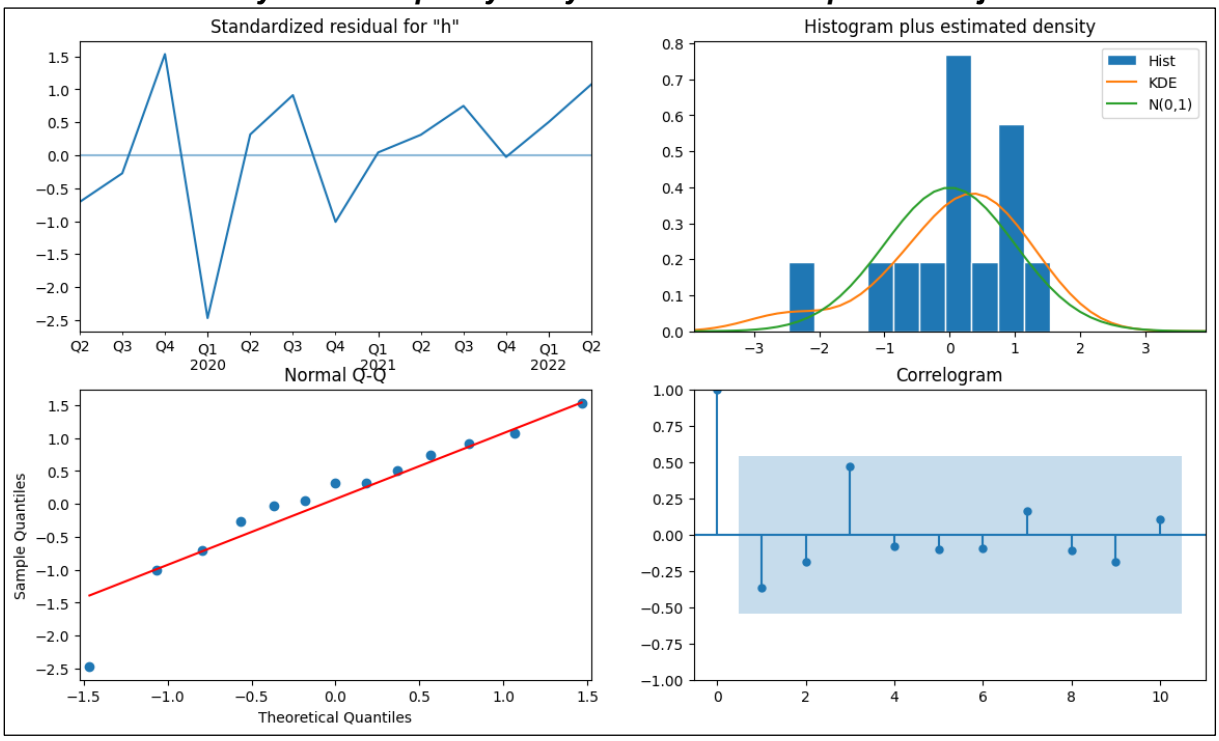


Zdroj: vlastné spracovanie autorov

V tejto analýze možno vidieť *p*-hodnotu spojenú s každým koeficientom každej premennej pre predpoveď v modeli SARIMAX. Často sa *p*-hodnota zneužíva ako spôsob, ako vykonať výber premenných („features“). Mnohí nesprávne interpretujú *p*-hodnotu ako spôsob určenia, či je premenná v modeli korelovaná s cieľom. V skutočnosti *p*-hodnota testuje, či sa koeficient významne líši od 0 alebo nie. Ak je *p*-hodnota menšia ako 0.05, zamietame nulovú hypotézu a usudzujeme, že koeficient je významne odlišný od 0. Neurčuje, či je premenná užitočná na prognózovanie. Preto by sa nemali odstraňovať premenné na základe ich *p*-hodnoty. O tento krok sa postará výber modelu na základe minimalizácie AIC [10].

Q-Q graf pre tento model sa nachádza na obrázku č. 10 v ľavej dolnej časti, kde sú rezíduá normálne rozdelené, s výnimkou nedokonalosti na konci intervalu. Na tomto obrázku možno vidieť, ako `statsmodels` uľahčuje kvalitatívnu analýzu rezíduí. V ľavom hornom grafe sú znázornené rezíduá v celom súbore údajov. Vidno, že neexistuje žiadny trend, hoci priemer sa nezdá stabilný v čase. I keď v rezíduách nie je žiadny trend, zdá sa, že rozptyl nie je konštantný, čo je rozdiel v porovnaní s bielym šumom. Vpravo hore je histogram rezíduí. Teoreticky by mohol byť blízko normálnemu rozdeleniu, avšak nemáme na lepší tvar histogramu dostatok hodnôt. V tejto podobe však naznačuje, že rezíduá nie sú blízke bielemu šumu, keďže biely šum je normálne rozdelený. A napokon, graf vpravo dole ukazuje autokorelačnú funkciu rezíduí. Pri oneskorení 0 je jediný významný vrchol a v ostatných prípadoch je minimum významných koeficientov. To znamená, že rezíduá sú minimálne korelované, čo by mohlo byť ešte vylepšené, ak by sme mali k dispozícii viac hodnôt.

Obrázok č. 9: Analýza rezíduí pre vybraný model SARIMAX podľa zdrojového kódu



Zdroj: vlastné spracovanie autorov

Posledným krokom analýzy rezíduí je použitie Ljungovho-Boxovho testu. Ten umožňuje kvantitatívne posúdiť, či sú rezíduá skutočne nekorelované. Na vykonanie Ljungovho-Boxovho testu na rezíduá bola použitá funkcia acorr_ljungbox zo statsmodels. Funkcia prijíma ako vstup rezíduá, ako aj zoznam oneskorení. V tomto prípade bola vypočítaná Ljungova-Boxova štatistika a *p*-hodnoty pre 10 oneskorení [14].

Výsledný zoznam *p*-hodnôt ukazuje, že každá z nich je vyššia ako 0,05. Preto pri každom oneskorení nie je možné zamietnuť nulovú hypotézu, čo znamená, že rezíduá sú nezávisle rozložené a nekorelované. Z našej analýzy vyvodzujeme záver, že rezíduá sú podobné bielemu šumu.

V ďalšom kroku vyhodnotíme koreláciu medzi exogénnymi premennými, ktoré sú použité v modeli SARIMAX.

Tabuľka č. 1: Korelačná matica exogénnych premenných na odhad premennej hdp_amount modelom SARIMAX

	hdp_ amount	distance	vehicle_ count
hdp_ amount	1,00	0,79	0,78
distance	0,79	1,00	0,91
vehicle_ count	0,78	0,91	1,00

Zdroj: vlastné spracovanie autorov

Je potrebné uviesť, že pri nowcastingu nie je problém multikolinearity medzi exogénnymi premennými významný – respektíve jeho dôsledky model neohrozujú. To okrem iného potvrdzuje aj porovnanie metriky AIC (Tabuľka č. č. 2), ktorá je minimálna práve pre model SARIMAX využívajúci obe exogénne premenné v tabuľke (tabuľka č. 2), ako aj obrázok č. 10 znázorňujúci rôznu kvalitu rýchlych odhadov uvedených

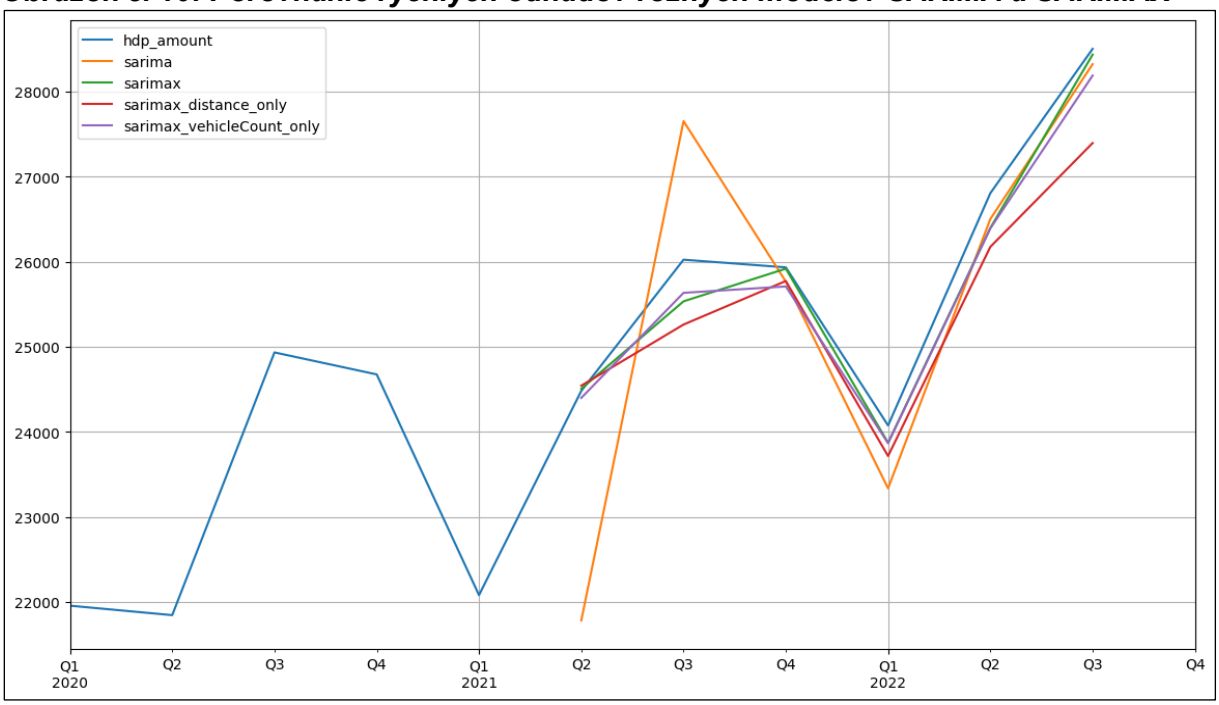
modelov, pričom zelenou je znázornený model SARIMAX s oboma exogénnymi premennými. Na tomto obrázku tiež vidno, že model SARIMA bez exogénnych premenných nedokáže spoľahlivo predpovedať vývoj v roku 2021, ktorý bol ovplyvnený aj pandémiou COVID-19.

Tabuľka č. 2: Porovnanie AIC pre modely SARIMA a SARIMAX

Typ modelu	**AIC**
SARIMA	225,66
SARIMAX s premennou distance	201,72
SARIMAX s premennou vehicle_count	195,13
SARIMAX s premennými distance a vehicle_count	**155,20**

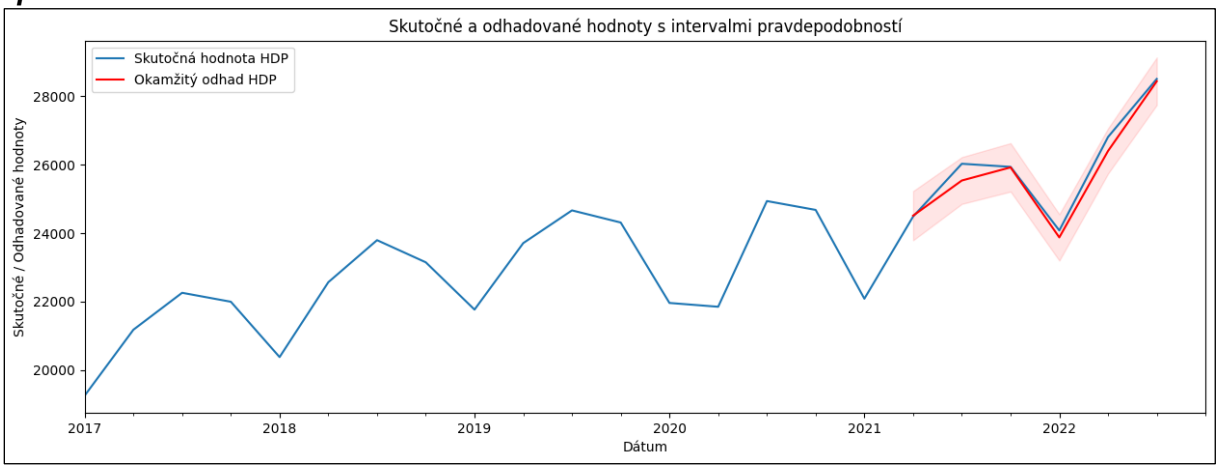
Zdroj: vlastné spracovanie autorov

Obrázok č. 10: Porovnanie rýchlych odhadov rôznych modelov SARIMA a SARIMAX



Zdroj: vlastné spracovanie autorov

Obrázok č. 11: Skutočné a odhadované hodnoty HDP v mil. eur s intervalmi spoľahlivosti



Zdroj: vlastné spracovanie autorov

Navyše obrázok č. 11 znázorňuje veľmi uspokojivý odhad HDP s intervalmi spoľahlivosti.

Pri používaní modelu SARIMAX platí dôležité upozornenie. Zahrnutie externých premenných môže byť potenciálne prospešné, pretože sa dajú nájsť silné prediktory – vstupné premenné do modelu pre cieľovú premennú. Pri prognózovaní viacerých časových krokov do budúcnosti však možno naraziť na problémy. Model SARIMAX používa model SARIMA($p\ d,\ q$)($P,\ D,\ Q$)$_m$ a lineárnu kombináciu exogénnych premenných na predpovedanie jedného časového kroku do budúcnosti. Ale čo ak treba predpovedať dva časové kroky do budúcnosti? Zatiaľ čo s modelom SARIMA je to možné, model SARIMAX vyžaduje, aby sa predpovedali aj exogénne premenné. S týmto problémom sa však pri rýchlych odhadoch nestretneme.

6. ZÁVER

Model SARIMAX nám umožňuje zahrnúť externé premenné, označované aj ako exogénne premenné, do nowcastingu cieľovej premennej – teda aktuálnej hodnoty HDP v mil. eur v bežných cenách. Transformácie sa uplatňujú len na cieľovú premennú, nie na exogénne premenné. Vplyv nového dátového zdroja sa testoval na odhad HDP od Q2 2021 do Q4 2022, kde aj len pozorovaním kriviek na obrázku č. 11 vidno, že model SARIMAX najlepšie kopíruje krivku skutočnej hodnoty HDP v mil. eur v bežných cenách.

Pri aplikácii modelu SARIMAX len na rýchle odhady nemusíme predpovedať viacero časových krokov do budúcnosti, a tak sa nemusia predpovedať ani exogénne premenné. To znamená, že táto vlastnosť modelu SARIMAX nezväčšuje chyby konečného odhadu. Model SARIMAX je vysvetliteľný, nie je príliš komplikovaný ani náročný na výpočet. Využitím modelu prispôsobeného aktuálnej verzii časového radu HDP so všetkými dostupnými hodnotami v modelovanom období minimalizujeme problém chyby prognóz tým, že odhadujeme aktuálnu hodnotu HDP v mil. eur.

Výkonnosť a presnosť modelu SARIMAX potvrdzuje aj tabuľka č. 3, v ktorej sú uvedené pre všetky modely z obrázka č. 11 percentuálne metriky chyby RMSPE a MAPE. Na základe definícií týchto metrík je zrejmé, že čím je nižšia hodnota metriky uvádzaná v percentách, tým o menej percent sa odhad daného modelu líši od skutočnej hodnoty HDP v mil. eur v bežných cenách. Pre model SARIMAX vieme podľa typu metriky dosiahnuť chybu menej ako 9 percent alebo menej ako 0,7 percenta.

Tabuľka č. 3: Porovnanie jednotlivých analytických modelov podľa metrík RMSPE a MAPE

	sarima	sarimax	sarimaxdistance_only	sarimax_vehicle Count _only
RMSPE	17,44	**8,91**	24,92	11,43
MAPE	1,37	**0,67**	2,09	1,10

Zdroj: vlastné spracovanie autorov

Výhodou získaných údajov je ich včasnosť a rozsah. Do systému elektronického mýta sú zahrnuté všetky motorové vozidlá s celkovou hmotnosťou nad 3,5 tony alebo jazdné súpravy s celkovou hmotnosťou nad 3,5 tony. Údaje tiež obsahujú anonymizované identifikátory jednotlivých nákladných vozidiel, z ktorých nie je možné opätovne zistiť či už evidenčné číslo vozidla, majiteľa vozidla alebo identitu vodiča.

Na druhej strane údaje nie sú dostupné priamo vo formáte, ktorý podporuje prelinkovanie, avšak úseky ciest majú svoje identifikátory, pomocou ktorých sa dajú vkresliť do mapy cestnej siete. To možno v budúcnosti využiť napríklad na odhad, aké percento nákladných vozidiel bolo len v tranzite cez Slovenskú republiku. Očistenie dát od tranzitnej nákladnej dopravy by mohlo mať pozitívny vplyv na kvalitu vstupných hodnôt do modelov. Tým by sa mohli eliminovať aj exogénne faktory ovplyvňujúce nákladnú dopravu, ako je geopolitická situácia v okolitých krajinách (vojnový konflikt na Ukrajine) či situácia na hraničných priechodoch (blokáda priechodu Vyšné Nemecké), na ktoré je súčasný model náchylný. Vytvorený model by tiež mohol zefektívniť vypovedaciu schopnosť zahrnutím údajov o preprave tovaru železničnou dopravou, ktorá tvorí okolo 15 % celkovej nákladnej dopravy[3].

LITERATÚRA

[1] ALHARBI, F. R. – CSALA, D.: A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach. In: Inventions, 2022, č. 1, s. 94.
[2] AKAIKE, H.: Factor analysis and AIC. In: Psychometrika, 1987, č. 3, s. 317 – 332.
[3] ALBERTO, A. – DIAMOND, A. – HAINMUELLER, J.: Comparative politics and the synthetic control method. American Journal of Political Science, 2015, č. 2, s. 495–510.
[4] ALLWRIGHT, S. 2023. How to interpret MAPE. [online]. [cit. 11-09-2023]. Dostupné na: https://stephenallwright.com/interpret-mape/.
[5] BAUM, C.: Tests for stationarity of a time series. In: Stata Technical Bulletin, 2000, č. 57, s. 36 – 39.
[6] BOX, G. – JENKINS, M.: Time series analysis: forecasting and control. John Wiley & Sons, 2015.
[7] DURBIN, J. – WATSON, G. S.: Testing for Serial Correlation in Least Squares Regression, II. In: Biometrika, 1951, č. 1 – 2, s. 159 – 179.
[8] HEILBRONNER, R. – Barrett, S. D.: Image Analysis in Earth Sciences: Microstructures and Textures of Earth Materials, Springer-Verlag Berlin Heidelberg, 2014.
[9] HOSSAIN, J.: Comparative Analysis of ARIMA, SARIMAX, and Random Forest Models for Forecasting Future GDP in Relation to Unemployment Rate. 2023.
[10] HYNDMAN, R. Statistical tests for variable selection. [online]. [cit. 11-09-2023]. Dostupné na: https://robjhyndman.com/hyndsight/tests2/.
[11] LANGBEIN, J. – HADLEY J.: Correlated errors in geodetic time series: Implications for time-dependent deformation. In: Journal of Geophysical Research: Solid Earth 102.B1. 1997. s. 591 – 603.
[12] LJUNG, G. M. – BOX, G. E. P.: On a measure of lack of fit in time series models, In: Biometrika, 1978. č. 2, s. 297 – 303.
[13] MEČIAROVÁ, K. Q-Q grafy – Oborový seminár. MFF Univerzita Karlova. 2021. [online]. [cit. 11-09-2023]. Dostupné na: https://www.karlin.mff.cuni.cz/~omelka/Soubory/nmsa401/Q-Q_plots.pdf.
[14] NIST/SEMATECH. Box-Ljung Test. In: e-Handbook of Statistical Methods. [online]. [cit. 11-09-2023]. Dostupné na: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4481.htm.

[3] Štatistický úrad SR. Súhrnné ukazovatele za dopravu [do1003rs].

[15] Najvyšší kontrolný úrad Slovenskej republiky. Elektronický výber mýta: Záverečná správa. 2019.
[16] PERKTOLD, J. – SEABOLD, S. – TAYLOR, J.: statsmodels.tsa.statespace.sarimax.SARIMAX. [online]. [cit. 11-09-2023]. Dostupné na: https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html#statsmodels.tsa.statespace.sarimax.SARIMAX.
[17] RAIMUNDO, B.: Time series Analysis. [online]. [cit. 11-09-2023]. Dostupné na: https://github.com/BernardoRaimundo/Time-Series-Analysis.
[18] SHUMWAY, H. – STOFFER, D.: Time series analysis and its applications. Third edition. New York: Springer, 2011.
[19] Skytoll: Elektronický výber mýta, Slovenská republika. [online]. [cit. 11-09-2023]. Dostupné na: https://www.skytoll.com/elektronicky-mytny-system-sr/.
[20] SONHAO, W.: Multicollinearity in Regression: Why it is a problem? How to check and fix it, In: Towards Data Science. [cit. 11-09-2023]. Dostupné na: https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea.
[21] Štatistický úrad Slovenskej republiky. 2023. Štvrťročné údaje HDP v bežných cenách: [nu0002qs]. [online]. [cit. 11-09-2023]. Dostupné na: http://statdat.statistics.sk/cognosext/cgi-bin/cognos.cgi?b_action=cognosViewer&ui.action=run&ui.object=storeID(%22iA188D5A85FEF4EA9B48BC102C7C66758%22)&ui.name=%C5%A0tvr%C5%A5ro%C4%8Dn%C3%A9%20%C3%BAdaje%20HDP%20v%20be%C5%BEn%C3%BDch%20cen%C3%A1ch%20%5Bnu0002qs%5D&run.outputFormat=&run.prompt=true&cv.header=false&ui.backURL=%2Fcognosext%2Fcps4%2Fportlets%2Fcommon%2Fclose.html&run.outputLocale=sk.
[22] THEODOSIOU, M.: Forecasting monthly and quarterly time series using STL decomposition. In: International Journal of Forecasting, 2011, č. 4, s. 1178 – 1195.

RESUMÉ

V rámci projektu Sociálno-ekonomické aspekty Big Data (SEABD) bol vedený experiment s cieľom preskúmať využitie údajov z elektronického mýtneho systému na vplyv vývoja hrubého domáceho produktu (HDP) ako typ prognózovania. Prognózovanie je postup, ktorý využíva historické údaje ako vstupy na vytvorenie informovaných odhadov, ktoré predpovedajú budúce trendy v dlhšom časovom horizonte. Rýchly odhad oproti tomu predstavuje postup odhadu veľmi nedávnej minulosti, súčasnosti alebo veľmi blízkej budúcnosti stavu ekonomických ukazovateľov. Cieľom projektu bolo na základe údajov o najazdených kilometroch nákladných vozidiel vytvoriť ukazovateľ nákladnej dopravy, ktorý je možné porovnať so štvrťročným ukazovateľom HDP z produkcie Štatistického Úradu SR. Elektronický mýtny systém zbiera údaje automatizovaným systémom s jasne zdokumentovanou metodikou na úsekoch diaľnic podľa platnej vyhlášky. Zozbierané údaje predstavujú agregované hodnoty počtu najazdených kilometrov a počtu unikátnych nákladných vozidiel po mesiacoch a štvrťrokoch. Výhoda tohto súboru údajov je jeho detailnosť, keďže každý najazdený kilometer môže predstavovať konkrétnu ekonomickú transakciu či činnosť. Na hodnotenie presnosti odhadu v rámci nowcastingu sa použili metriky priemernej absolútnej percentuálnej chyby (MAPE) a ich smerodajná odchýlka (RMSPE), ako aj porovnanie so skutočnými hodnotami HDP v danom kvartáli. Výsledky modelu boli veľmi uspokojivé s hodnotami pre MAPE 0,67 % a RMSPE 8,7 %. Možný budúci vývoj a význam pokračujúceho výskumu využívania údajov o najazdených kilometroch nákladných vozidiel na odhady makroekonomických

ukazovateľov spočíva v presnejšom a rýchlejšom určovaní hospodárskej aktivity. Tieto údaje môžu poskytnúť informácie v reálnom čase o pohybe tovaru a obchodu, čo je kritické pre ekonomické rozhodovanie a plánovanie. S rozvojom technológií IT a senzorov v nákladných vozidlách je možné získať ešte presnejšie a aktuálnejšie údaje.

RESUME

As part of the Socio-Economic Aspects of Big Data (SEABD) project, an experiment was conducted to investigate the use of electronic toll system data for the impact of development the gross domestic product (GDP) as a type of forecasting. Forecasting is a practice using historical data as inputs to make informed estimates that predict future trends over a longer time horizon. A flash estimate, on the other hand, represents a process of estimating the very recent past, present or very near future of the state of economic indicators. The goal of the project was to create an indicator of freight transport based on truck mileage data, which can be compared with the quarterly GDP indicator produced by the Statistical Office of the Slovak Republic. The Electronic toll system collects data using an automated system with a clearly documented methodology on sections of motorways. The collected data represents the aggregated values of the number of kilometres travelled and the number of unique trucks by month and quarter. The advantage of this dataset is its detail, as each kilometer driven can represent a specific economic transaction or activity. The mean absolute percentage error (MAPE) and their standard deviation (RMSPE) metrics were used to assess the accuracy of the nowcasting estimate, as well as a comparison with the current GDP values in the given quarter. The results of the model were satisfactory with values for MAPE 0.67% and RMSPE 8.7%. A possible future development and importance of a continued research into the use of truck mileage data for estimation of macroeconomic indicators is to more accurately and rapidly determine economic activity. This data can provide a real-time information on the movement of goods and trade, which is critical for economic decision-making and planning. With the development of the IT technologies and sensors in trucks, even more accurate and up-to-date data can be obtained.

PROFESIJNÝ ŽIVOTOPIS

Peter Knížat, MSc., *je externým študentom doktorandského štúdia na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave. Vyučuje praktické cvičenia: podpora rozhodovacích procesov a viac atribútové rozhodovanie vrátane aplikácie v štatistickom softvéri R. Pracuje ako dátový analytik v sekcii všeobecnej metodiky, registrov a koordinácie národného štatistického systému Štatistického úradu SR, kde je zodpovedný za návrh štatistickej metodiky v cenových štatistikách s využitím webscrapovaných údajov. Predtým pôsobil v medzinárodnej banke ako senior risk a portfóliový manažér, kde viedol vývoj interných modelov používaných v procesoch hodnotenia kreditného rizika.*

Dipl. Ing. Dagmar Ceľuchová Bošanská *je zakladateľkou spoločnosti Alistiq s. r. o. a expertkou na inovácie a digitálnu transformáciu s dlhoročnými skúsenosťami. V roku 2008 absolvovala inžinierske štúdium pre informačné technológie, mobilné komunikácie a štatistické spracovanie signálov na Viedenskej technickej univerzite, kde pôsobila vo vedeckom tíme na vývoji simulátorov technológií pre bezdrôtové siete štvrtej generácie. Od roku 2015 sa venuje vývoju riešenia a návrhu opatrení na zvyšovanie kvality a efektivity využívania údajov vrátane Big Data na sekundárne účely, predovšetkým vo verejnej správe. Aktuálne od roku 2020 pôsobí ako doktorand na Českom vysokom učení technickom v Prahe, kde sa venuje výskumu grafových údajov generovaných z elektronických zdravotných záznamov a ich analýze s využitím strojového učenia a veľkých jazykových modelov.*

Ing. Martin Janík *absolvoval inžinierske štúdium na Fakulte elektrotechniky a informatiky Slovenskej Technickej Univerzity v Bratislave v odbore telekomunikácie, špecializácia bezpečnosť (2008). Už počas štúdia na vysokej škole začal pracovať v súkromnom sektore v oblasti IT, spočiatku v oblasti webových neskôr mobilných technológii ako programátor, analytik a následne softvérový architekt softvérových produktov v oblasti mobilných technológií. Od roku 2022 sa venuje dátovej vede so zameraním na analýzu a návrh grafových dátových štruktúr vo verejnom sektore. Spolupracuje na Centrálnom modeli údajov SR a na analýze a spracovaní Big Data.*

Mgr. Filip Nguyen *absolvoval magisterské štúdium v Ústave pedagogiky a sociálnych štúdií Univerzity Palackého v Olomouci (CZ) v odbore pedagogika – veřejná správa (2019). Od roku 2018 pôsobí v poradenskej spoločnosti Alistiq, s. r. o. ako poradca v oblasti verejného obstarávania a verejných inovačných projektov. Jeho práca sa zameriava na návrh digitálnych služieb štátnej správy a aplikáciu osvedčených postupov PRINCE2 a Agile metodík v projektoch.*

KONTAKT

peter.knizat@statistics.sk
dagmar.bosanska@alistiq.com
martin.janik@alistiq.com
filip.nguyen@alistiq.com

Martin ŠVEDA, Michala SLÁDEKOVÁ MADAJOVÁ
Prírodovedecká fakulta Univerzity Komenského, Geografický ústav SAV, v. v. i.
Pavol HURBÁNEK, Konštantín ROSINA
Geografický ústav SAV, v. v. i.

SPRACOVANIE PASÍVNYCH LOKALIZAČNÝCH ÚDAJOV MOBILNEJ SIETE NA POUŽITIE V EXPERIMENTÁLNEJ POPULAČNEJ ŠTATISTIKE

PROCESSING OF PASSIVE MOBILE POSITIONING DATA FOR USE IN THE EXPERIMENTAL POPULATION STATISTICS

ABSTRAKT
Mobilný telefón sa stal neoddeliteľnou súčasťou každodenného života a unikátnym zdrojom údajov o obyvateľstve, jeho priestorovom rozmiestnení, mobilite a aktivitách. Cieľom príspevku je opísať konceptuálny a metodický prístup spracovania pasívnych lokalizačných údajov mobilnej siete a pokúsiť sa odpovedať na nasledujúce výskumné otázky: Dokážu lokalizačné údaje z mobilnej siete poskytnúť relevantnú informáciu o priestorovom rozmiestnení populácie na Slovensku? Môžu byť tieto údaje relevantným doplnkom ku štatistickým údajom o obyvateľstve a aké sú ich limity? Výsledky spracovania lokalizačných údajov mobilnej siete priniesli robustné výsledky, ktoré môžu prispieť k hlbšiemu poznaniu priestorového rozmiestnenia obyvateľstva. Pri rešpektovaní viacerých limitov dostávame perspektívne dáta, vhodné na rozmanité analýzy na úrovni regiónov alebo lokalít.

ABSTRACT
The mobile phone has become an integral part of everyday life and a unique source of data on the population, including its spatial distribution, mobility, and activities. The aim of this paper is to describe the conceptual and methodological approach of processing passive mobile positioning data and to address the following research questions: Can mobile positioning data provide meaningful information about the spatial distribution of the Slovak population. Can data from the mobile network be a meaningful complement to conventional population statistics, and what their limits are. The results of processing mobile network location data have yielded powerful results that can contribute to a deeper understanding of the spatial distribution of the population. Despite several limitations, we obtain prospective data suitable for diverse analyses at the regional and local level.

KĽÚČOVÉ SLOVÁ: mobilná sieť, lokalizácia, odhad populácie, Slovensko

KEY WORDS: mobile network, localization, population estimates, Slovakia

1. ÚVOD
Poznanie priestorovej distribúcie a mobility obyvateľov je kľúčovým vstupom pre množstvo vedeckých analýz a praktických úloh. Tradičným zdrojom týchto úloh sú údaje z evidencie obyvateľstva a z celoštátneho cenzu. Napriek tomu, že ide o koncepčne a metodicky prepracované zdroje údajov, vývoj spoločnosti prináša nové fenomény, na ktoré konvenčné zdroje údajov nemusia dostatočne spoľahlivo reagovať. Osobitne to platí pre spoločensky orientované vedné disciplíny, ktoré potrebujú aktuálne, presné a detailné údaje o rozmanitých fenoménoch časovo-priestorového správania populácie. Je zrejmé, že ak chceme reagovať na spoločenský

vývoj, potrebujeme viac participovať na možnostiach, ktoré prinášajú informačno-komunikačné technológie. Výsledkom adaptácie výskumných metód na nové podmienky by malo byť predovšetkým využitie informačno-komunikačných technológií, ktoré prinášajú nielen zmenu paradigmy priestorového správania, ale aj prístupu k jeho sledovaniu.

Motívov na použitie iných prístupov je viacero. Predovšetkým pracovné aktivity obyvateľov sa neviažu na jedno miesto tak pevne, ako to bývalo v minulosti a miesto trvalého pobytu, miesto bývania a miesto práce sa čoraz častejšie od seba odlišujú [30, 41]. Pohyby obyvateľstva sa stávajú pestrejšími, nadobúdajú nerutinný a nepravidelný charakter. V súčasnosti nie je neobvyklé bývanie v podnájme, sezónne bývanie na chate či dlhodobé pracovné stáže. Čoraz menšia viazanosť obyvateľstva na jedno miesto sa prejavuje v rôznych spoločenských kontextoch a relativizuje výsledky sociálnych, geografických či ekonomických analýz založených na tradičných štatistických zdrojoch (cenzy, registre obyvateľstva a pod.). Pre časť populácie je zložité jednoznačne odpovedať na otázku miesta bydliska a miesta práce. Túto skutočnosť pritom nemôžeme vnímať ako dočasné „uvoľnenie“ viazanosti na základné priestorové súradnice nášho každodenného života, ale ako dlhodobé trendy prinášajúce zvýšenú fluktuáciu a mobilitu obyvateľstva, ktoré stimulujú viaceré fenomény postmodernej spoločnosti, predovšetkým informačno-komunikačné technológie [8]. Tie uľahčili nielen nástup práce z domu, ale aj ekonomiky postavenej na nezávislých dodávateľoch a externých pracovníkoch (*gig economy*), to všetko s významným vplyvom na spôsob práce a mobility obyvateľov, najmä v metropolitných regiónoch.

Aktuálnym fenoménom, ktorý pravdepodobne prispel k zmene priestorových vzorov správania, je pandémia ochorenia COVID-19. Potreba sociálnej separácie a obmedzenia každodenného života nás prinútili adaptovať sa na nové spôsoby práce, učenia sa a interakcií. Hoci obmedzenia boli len dočasné, môžu mať za následok dlhodobé zmeny v priestorových preferenciách pre prácu, bývanie a každodennú mobilitu [26]. Dnes je však ešte predčasné predpovedať charakter zmien v priestorových prejavoch populácie. Je však ale možné, že stojíme na prahu novej paradigmy, ktorá ich bude definovať najbližšie desaťročia. Je otázne, ako na uvedené zmeny v spoločnosti dokážeme reagovať prostredníctvom konvenčných zdrojov údajov.

V nových zdrojoch údajov, ktoré nám môžu pomôcť porozumieť trendom a fenoménom doby, má osobitné postavenie využitie údajov z mobilnej siete. Mobilný telefón sa stal neoddeliteľnou súčasťou každodenného života a unikátnym zdrojom údajov o obyvateľstve, jeho priestorovom rozmiestnení, mobilite a aktivitách. Myšlienky využitia týchto údajov v priestorových analýzach sa objavili súčasne s masovým rozvojom mobilnej komunikácie na báze GSM [4, 28, 35], avšak až v ostatnom desaťročí sme svedkami skutočne rozsiahleho rozvoja v tejto výskumnej oblasti.

Intenzívny záujem o údaje z mobilnej siete vyvolala nielen prirodzená „zvedavosť“ výskumníkov, ale aj viaceré praktické úlohy, pred ktorými stojí súčasná globalizovaná spoločnosť. Záujem o údaje z mobilnej siete prejavili viaceré medzinárodné organizácie, inštitúcie Európskej únie či národné štatistické úrady. Príkladom sú aktivity *Joint Research Centre* pri Európskej komisii, kde sa podrobne zaoberali využitím údajov z mobilnej siete pre spresnenie priestorového rozmiestnenia populácie

[36]. Spomedzi krajín EÚ sa azda najďalej dostalo Estónsko, kde výskumníci pod vedením R. Ahasa vybudovali komplexnú metodiku pre sledovanie pohybu zahraničných návštevníkov [2].

Lokalizačné údaje mobilnej siete majú v porovnaní s tradičnými údajmi o dochádzke niekoľko pozitívnych vlastností. V prvom rade poskytujú aktuálne informácie, ktoré sú pomerne rýchlo spracovateľné a nevyžadujú si aktívnu participáciu obyvateľov. Nie sú teda poznačené časovým odstupom, ako to spravidla býva v celoplošných cenzoch. Dostupnosť týchto údajov prakticky v akomkoľvek čase umožňuje pružne reagovať na vývoj spoločnosti a aktuálne zmeny v priestorovom rozmiestnení obyvateľstva (výstavba bytov, ciest, koncentrácia novej výroby a pod.). Navyše prostredníctvom mobilnej siete môžeme sledovať zmeny v dennej priestorovej mobilite spôsobené sezónnymi cyklami a flexibilne prispôsobovať rozsah pozorovania spoločenským či výskumným potrebám, respektíve vytvárať dlhodobé pozorovania a sledovať tak reakcie spoločenských procesov na zmeny v intenzite a smerovaní mobility populácie [14, 46]. Skutočnosť, že lokalizačné údaje mobilných zariadení vznikajú na úrovni jednotlivých buniek mobilnej siete, umožňuje agregáciu podľa rozmanitých priestorových schém, ktoré nemusia rešpektovať územnosprávne členenie. To nám umožňuje získavať informácie o rozmiestnení a mobilite populácie na subregionálnej a lokálnej úrovni, mierke bežne nedostupnej s využitím údajov viazaných na administratívne členenie Slovenska. To má význam osobitne pri sledovaní procesov na intraurbánnej úrovni.

Perspektívy tohto nekonvenčného zdroja údajov si uvedomuje aj Štatistický úrad SR. V spolupráci s ním bol realizovaný projekt Socioekonomické aspekty Big data s cieľom preskúmať možnosti využitia údajov z mobilnej siete v experimentálnej populačnej štatistike. Cieľom príspevku je opísať konceptuálny a metodický prístup spracovania lokalizačných údajov mobilnej siete v tomto projekte a pokúsiť sa odpovedať na nasledujúce výskumné otázky:

- Dokážu údaje z mobilnej siete poskytnúť relevantnú informáciu o priestorovom rozmiestnení a mobilite populácie na Slovensku?
- Môžu byť údaje z mobilnej siete relevantným doplnkom ku štatistickým údajom o obyvateľstve a aké sú ich limity?

2. VYUŽITIE ÚDAJOV MOBILNEJ SIETE NA SPRESNENIE PRIESTOROVÉHO ROZMIESTNENIA POPULÁCIE

Nekonzistencia medzi oficiálnym (štatistickým) počtom obyvateľov a skutočným stavom obyvateľstva v priestorových jednotkách rôznej úrovne sa v odborných kruhoch na Slovensku sleduje už približne od polovice 90. rokov minulého storočia (napr. [11, 13, 32, 41]. Napriek pokusom o aproximáciu reálneho počtu obyvateľov využitím zástupných ukazovateľov, ako je napríklad kapacita bytového fondu, musíme konštatovať, že v súčasnosti nepoznáme reálne priestorové rozmiestnenie obyvateľstva na Slovensku. Uvedomenie si tejto triviálnej skutočnosti má pritom kľúčový význam pre sociálno-ekonomické analýzy, hodnotenie transformačných procesov či prognózovanie populačného a hospodárskeho vývoja regiónov, ako aj pre ďalšie praktické opatrenia (rozpočtovanie daní, kreovanie obecných zastupiteľstiev a volebných okrskov a pod.).

Tento problém sa netýka len Slovenska, ale predstavuje globálny fenomén, ktorému sa vo svetovej literatúre venuje značná pozornosť. Súčasná spoločnosť je viac ako

kedykoľvek predtým tvorená tokmi ľudí, tovarov a informácií, avšak dáta, ktoré tieto toky zaznamenávajú, sú pomerne náročné na spracovanie využitím tradičných prístupov spoločenského výskumu [34, 38, 39]. Od prvých štúdií [18, 25, 45] až po najnovšie prístupy (napr. projekt ENACT – [6, 7], sa metódy mapujúce priestorové rozmiestnenie populácie sa postupne zdokonaľovali a poskytovali čoraz presnejší obraz rezidenčnej populácie v rôznych častiach sveta. Vďaka týmto aktivitám mohli byť tradičné mapy znázorňujúce hustotu obyvateľstva v administratívnych jednotkách, nahradené realistickejšími zobrazeniami distribúcie populácie prostredníctvom buniek pravidelných gridov označovaných ako populačné gridy [6]. Vysoké priestorové rozlíšenie populačných gridov (najčastejšie bunky o veľkosti 1 x 1 km) umožnilo širšiu integráciu týchto údajov v geografických informačných systémoch a tieto údaje sa stali nenahraditeľnými pre sociálny a environmentálny výskum či územné plánovanie.

Napriek značnému vedeckému pokroku sú populačné gridy vo svojej podstate naďalej „len“ statickými záznamami rezidenčnej populácie, ktoré zachytávajú rozmiestnenie obyvateľstva počas noci, a to za predpokladu, že väčšina ľudí zotrváva v rámci deklarovaného miesta pobytu. Napriek tomu, že populačné gridy poskytujú rozmanité využitie, opisujú len časť reality. Priestorové rozmiestnenie populácie počas dňa, respektíve v priebehu rôznych častí roka, je prakticky neznáme pre ľubovoľnú priestorovú mierku. Takéto informácie sú však nevyhnutné pre celý rad aplikácií. Distribúcia obyvateľstva počas dňa je determinovaná rozmiestnením hospodárskych, sociálnych a rekreačných zariadení, ktoré priťahujú obyvateľstvo z ich bydliska, predurčujú dochádzkové toky a iné formy dennej mobility. Denné rozmiestnenie populácie sa teda (značne) líši od rezidenčnej (nočnej) populácie. Je preto zrejmé, že modelovanie rozmiestnenia populácie nedokážeme efektívne realizovať len s využitím konvenčných zdrojov údajov.

Jedným z riešení, ako sa vyrovnať s chýbajúcimi údajmi, je využitie údajov z mobilnej siete. So zvyšujúcim sa výskytom mobilných telefónov v populácii sa lokalizácia v mobilnej sieti stáva dôležitým zdrojom údajov, ktorý môže dopĺňať a rozširovať existujúcu populačnú štatistiku. Tento nový zdroj údajov nám umožňuje sledovať, ako sa rozmiestnenie populácie mení nielen v čase, ale aj v závislosti od rôznych cyklov (denný, týždenný, sezónny a pod.) či akejkoľvek udalosti, ktorá ovplyvňuje koncentráciu populácie (športové podujatia, prírodné katastrofy). Uplatnenie týchto údajov tak nachádzame nielen vo vedeckom výskume, ale čoraz častejšie aj v rozhodovacej praxi, kde prispievajú k zvyšovaniu kvality a efektívnosti územného plánovania a manažmentu.

Údaje z mobilnej siete majú potenciál vytvoriť automatizované odhady rozmiestnenia populácie prakticky v reálnom čase. Hoci nedosahujú presnosť a spoľahlivosť cenzov, ich porovnanie s konvenčnými zdrojmi (národnými cenzami či populačnými registrami) prináša pomerne vysoké korelácie [20, 24, 30]. Aktuálne metodické prístupy [16, 17, 27] využívajú lokalizačné údaje z mobilnej siete v úzkej synergii so satelitným snímkovaním a rozmanitými mapovými vrstvami (zastavané územie, využitie zeme, funkčné areály). Môžeme očakávať, že pri spracovaní a analýze časovo-priestorových záznamov z mobilnej siete sa budú čoskoro využívať aj rozšírené možnosti umelej inteligencie [21].

3. LOKALIZÁCIA V MOBILNEJ SIETI

Princíp lokalizácie v mobilnej sieti je prirodzenou nadstavbou základných vlastností mobilnej komunikácie. Z priestorového pohľadu môžeme územie rozdeliť do buniek, ktoré sú obsluhované jednotlivými anténami mobilnej siete. Každá anténa je schopná pokryť určité územie a obslúžiť určitý počet zákazníkov. Identifikačné údaje o aktuálne využívanej anténe, ako aj ďalšie doplnkové informácie sa môžu využiť na určenie približnej polohy mobilného zariadenia. Tieto lokalizačné údaje sú prirodzenou súčasťou mobilnej komunikácie, keďže identifikácia základňových staníc (*base transceiver station*, BTS), s ktorými komunikuje mobilný telefón, je nevyhnutná pre samotné fungovanie mobilnej siete.

Lokalizačné údaje, ktoré môžeme získať prostredníctvom lokalizácie v mobilnej sieti, principiálne rozdeľujeme na aktívne a pasívne [3]. Kým pri pasívnom type ide o využitie existujúcich záznamov v systéme mobilného operátora, pri aktívnom type sa záznam vytvára na základe konkrétneho dopytu a s využitím špecializovaného softvéru. Pasívne lokalizačné údaje sú teda digitálne „stopy“, ktoré zanecháva mobilné zariadenie na infraštruktúre mobilnej siete. Tieto záznamy vznikajú buď pre potreby vyúčtovania hovorov, SMS a pod., alebo sú dôsledkom pravidelných aktualizácií polohy zariadenia v mobilnej sieti. Vzhľadom na skutočnosť, že k údajom o aktivite mobilného zariadenia vieme priradiť nielen približnú polohu (konkrétnu bunku mobilnej siete), ale aj niektoré základné informácie o používateľovi (vek, pohlavie či fakturačná adresa), získavame bohatú databázu, ktorej využitie (pri zachovaní anonymity používateľov, resp. pri čiastočnom agregovaní údajov) prináša nesmierne cenný analytický nástroj.

Lokalizácia mobilného zariadenia na úrovni buniek mobilnej siete umožňuje analyzovať priestorové rozmiestnenie obyvateľstva v relatívne detailnej mierke a v každom okamihu. Tieto vlastnosti predurčujú takýto zdroj údajov na široké multidisciplinárne uplatnenie. Svedčí o tom aj pestrá paleta aplikácií, ktorú nachádzame vo svetovej literatúre [40, 48]. Využitím údajov z mobilnej siete môžeme sledovať nielen statický obraz priestorového rozmiestnenia populácie na úrovni jednotlivých lokalít (napr. obcí, urbanistických obvodov a pod.), ale aj zmeny v koncentrácii obyvateľstva v ľubovoľných časových rezoch (napr. každú hodinu). To má zásadný význam predovšetkým pre lepšie porozumenie denných rytmov lokalít [43] a vo výskume priestorovej mobility [15, 22]. Bezprostredné využitie nachádzajú takéto údaje aj v krízovom manažmente, v ktorom pomáhajú distribuovať záchranné zložky a kapacity na evakuáciu obyvateľstva [9], prípadne pomáhajú predikovať vývoj epidémií [19, 49].

4. SPRACOVANIE PASÍVNYCH LOKALIZAČNÝCH ÚDAJOV MOBILNEJ SIETE

Základným krokom na relevantné využitie mobilnej lokalizácie v množstve priestorovo orientovaných analýz je určenie pravidelnej dennej a nočnej lokality používateliek a používateľov mobilnej siete. Metodický a konceptuálny rámec vychádza z konceptu tzv. kotevných bodov [5]. Kotevné body sa chápu ako hlavné uzly aktivít človeka, ktoré vytvárajú kostru jeho každodenných pohybov. V praxi ide predovšetkým o identifikovanie základných kotevných bodov „domov“ a „práca“. Údaje o koncentrácii denných a nočných kotevných bodov nám umožňujú nielen spresniť priestorovú distribúciu obyvateľstva (napr. koľko ľudí býva v časti mesta), ale aj extrahovať údaje o predpokladaných denných dochádzkových tokoch (napr. koľko ľudí dochádza denne do danej časti mesta).

Originálna metodika využíva rozsiahle údaje z mobilnej siete na odhad rozmiestnenia populácie prostredníctvom identifikovania pravidelnej dennej a nočnej lokality (bunky mobilnej siete) používateľov mobilných zariadení. Cieľom štúdie bolo vypracovať metodický rámec na zhromažďovanie a spracovanie lokalizačných údajov mobilnej siete, ktorý je vhodný na použitie medzi viacerými poskytovateľmi služieb mobilnej komunikácie. Hlavnou výzvou tejto úlohy bolo navrhnúť metodiku, ktorá by bola všeobecne aplikovateľná aj v heterogénnych scenároch, kde niekoľko technických detailov konfigurácie sietí a organizácie dát zostáva špecifických pre jednotlivých poskytovateľov mobilných služieb.

S týmto cieľom sa snažíme o vytvorenie „odolného“ metodického rámca, pričom základný súbor funkcií sa nespolieha na žiadnu neštandardnú konfiguráciu špecifickú pre jednotlivých mobilných operátorov a zároveň je dostatočne flexibilný na využitie prípadných osobitných charakteristík mobilnej siete a/alebo údajov špecifických pre jednotlivých mobilných operátorov. Vďaka tejto flexibilite sa navrhovaná metodika môže v budúcnosti rozšíriť a ďalej zdokonaliť, pričom sa počíta s prirodzenou evolúciou infraštruktúry mobilnej siete (napríklad zmenšovanie plochy buniek a nárast ich počtu).

Vstupné dáta reprezentovali pasívne lokalizačné údaje mobilnej siete od troch najväčších mobilných operátorov na Slovensku (Slovak Telekom, Orange Slovensko, O2 Slovakia), ktoré boli spracované na základe vopred stanovenej metodiky s cieľom identifikovať pravidelnú nočnú a dennú lokalitu mobilných zariadení (SIM kariet) na úrovni buniek mobilnej siete a následne transformovať tieto údaje do siete populačného gridu 1x1 km alebo katastrálneho územia obcí. Súčasťou datasetov sú aj matice sumarizujúce vektory pravidelnej dennej a nočnej lokalizácie individuálnych používateľov mobilnej siete. V nasledujúcom texte predstavíme štruktúru údajov, spôsob extrahovania pravidelných lokalizácií a transformáciu údajov z úrovne buniek mobilnej siete do cieľových priestorových zón.

4.1. Štruktúra pasívnych lokalizačných údajov mobilnej siete

Pasívne lokalizačné údaje z mobilnej siete sa principiálne skladajú z dvoch datasetov. Prvý predstavujú individuálne záznamy mobilného zariadenia na infraštruktúre mobilnej siete. Obsahujú identifikátor telefónneho čísla asociovaného so SIM kartou (MSISDN), dátum, časovú známku a identifikátor bunky mobilnej siete (tabuľka č. 1). Druhý dataset obsahuje priestorovú geometriu vyžarovacích polygónov pre jednotlivé bunky mobilnej siete (tabuľka č. 2). Vyžarovací polygón predstavuje mnohostranný plošný útvar opisujúci pokrytie signálom práve jednej BTS bunky. Vzhľadom na citlivosť obidvoch datasetov tieto záznamy používa len mobilný operátor na generovanie vopred definovaných časových a priestorových agregátov. Hoci sa údaje vzťahujú na individuálnu SIM kartu, pre jednoduchosť budeme v ďalšej časti príspevku uvažovať o ideálnej situácii, keď jedna SIM karta reprezentuje jedného používateľa mobilnej siete. Limitom tohto zjednodušenia sa budeme podrobnejšie venovať v 6. kapitole.

Tabuľka č. 1: Individuálne záznamy v mobilnej sieti

msisdn	dátum	čas	Id bunky mobilnej siete
ID_SIM 001	15.5.2023	18:02:05	BTS_0001

Zdroj: vlastné spracovanie autorov

Tabuľka č. 2: Priestorová geometria vyžarovacích polygónov BTS definovaná ako zlomové body multipolygónu

enodeb_id	Priestorová geometria vyžarovacieho polygónu
BTS_0001	17,0574; 48,1556 ; 17,0566, 48,1776 ;...

Zdroj: vlastné spracovanie autorov

4.2. Extrakcia pravidelnej dennej a nočnej lokalizácie mobilného zariadenia v sieťovej lokalizácii

Základným krokom na relevantné využitie mobilnej lokalizácie pri snahe o zachytenie priestorového rozmiestnenia populácie je určenie pravidelnej dennej a nočnej lokality užívateľov mobilnej siete. V praxi existuje celý rad techník (pozri napr. [5, 50, 51] na extrahovanie relevantných denných (pracovných), nočných (domácich) či iných periodických lokalít z množstva polohových záznamov, ktoré vznikajú v mobilnej sieti. Principiálne môžeme uvažovať o dvoch komplementárnych prístupoch identifikácie periodickej lokalizácie. Pri prvom prístupe určíme tú bunku mobilnej siete, ktorá zaznamenala najväčší počet lokalizácií (záznamov) v danom čase (napr. počas nočných hodín). V druhom prípade je cieľom identifikovať bunku, na ktorej strávilo mobilné zariadenie najdlhší čas. V oboch prípadoch je dôležité v čo najväčšej miere eliminovať náhodné a nerutinné lokalizácie. Na vyprofilovanie rutinných denných, nočných a iných pravidelných lokalizácií je dôležitá dostatočná dĺžka pozorovania, ktorú je potrebné prispôsobiť sledovaným javom, finančným možnostiam, ako aj výpočtovým kapacitám. Na základe predchádzajúcich skúseností [42, 43] boli algoritmy na identifikáciu pravidelných denných a nočných lokalít užívateľov mobilnej siete skonštruované zo 4-týždňového pozorovania takto:

Pravidelná nočná lokalita je územie pokrytia signálom práve jednej bunky mobilnej siete, v ktorom bola SIM karta najčastejšie lokalizovaná počas pracovného týždňa (PO – PIA) v čase od 23:00 do 5:59 hod. Táto lokalita bola identifikovaná zo súboru 20 pracovných dní tak, že za každú hodinu pozorovania sa identifikovala BTS stanica, kde mala SIM karta najväčší počet záznamov. Vzhľadom na 16 nocí pozorovania a 7 hodinových intervalov počas dňa ide o 112 časových intervalov. Bunka mobilnej siete s najväčším počtom lokalizácií (maximálne 112) bola určená ako pravidelná nočná lokalita danej SIM karty.

Pri väčšine používateľov sa pravidelná nočná lokalita extrahovala z rádovo stoviek záznamov. Mohli však nastať situácie, keď sa nočná lokalita určila len z malého počtu záznamov. V realite môže ísť napríklad o používateľov mobilnej siete, ktorí bývajú v zahraničí a na území Slovenska sa vyskytli len počas dňa. Vzhľadom na to sa aplikovala dodatočná podmienka, ktorá ponechala len tých používateľov mobilnej siete (SIM karty), ktorí strávili na infraštruktúre mobilnej siete minimálne 3 hodiny počas pozorovania v pracovnom týždni v čase od 23:00 do 5:59 hod. Túto podmienku bolo potrebné splniť minimálne 9 zo 16 nocí pozorovania.

Analogicky bola identifikovaná pravidelná denná lokalita. S cieľom eliminovať nerutinné vplyvy pondelkového a piatkového režimu sa použilo kratšie obdobie pozorovania, ktoré zahŕňalo UT – ŠT v čase od 9:00 do 14:59 hod. (tabuľka č. 3).

Tabuľka č. 3: Podmienky na extrahovanie pravidelnej nočnej a dennej lokality mobilného zariadenia (SIM karty) z pasívnych lokalizačných údajov mobilnej siete

	Dĺžka pozorovania	**dni počas týždňa**	**čas**	**počet hodinových intervalov**	**Minimálna dĺžka záznamu**
Pravidelná nočná lokalita SIM karty	28 dní	PO – PI	23:00 – 05:59	112	aspoň 3 hodiny počas 9 nocí pozorovania
Pravidelná denná lokalita SIM karty		UT – ŠT	09:00 – 14:59	72	aspoň 3 hodiny počas 7 dní pozorovania

Zdroj: vlastné spracovanie autorov

4.3. Transformácia lokalizačných záznamov z mobilnej siete do cieľových územných rámcov

Na praktické využitie uvedených kategórií v populačnej štatistike bolo potrebné tieto údaje transformovať do cieľových územných zón, v tomto prípade buniek gridu 1 x 1 km (50 661 buniek) a do katastrálnych území obcí (2 927 obcí). Pri transformácii údajov (počtu SIM kariet s pravidelnou nočnou/dennou lokalizáciou) do cieľových priestorových jednotiek sme použili nebinárnu dazymetrickú metódu[1] s využitím pomocnej vrstvy objemu budov. Výber tejto transformačnej metódy bol výsledkom testovania viacerých perspektívnych metód [44]. Pomocná vrstva vychádza zo spracovania polygónov budov databázy ZBGIS (2017) a bola skonštruovaná v dvoch verziách:

1) Na transformáciu údajov o počte SIM kariet s pravidelnou nočnou lokalitou sa použili polygóny rodinných domov a bytových domov. Na spresnenie interpolácie sa použili odlišné váhy pre objem bytových domov a rodinných domov v pomere 3 : 1. Tento pomer je výsledkom kalibrácie modelu na testovacej vzorke údajov, ktoré reprezentovali približne 1/50 územia Slovenska.

2) Na transformáciu údajov z dennej lokalizácie sa použil súbor rezidenčných aj nerezidenčných typov budov, v ktorých môžeme predpokladať prítomnosť obyvateľov počas dňa (napr. školy, administratívne budovy, priemyselné budovy a pod). Objem budov bol odhadnutý prostredníctvom rozlohy zastavanej plochy a maximálnej výšky budovy. Prístup používajúci objem budov je vhodný osobitne vo vysoko

[1] Dazymetrická metóda pracuje na princípe tzv. dazymetrického mapovania, keď sú dáta znázornené prostredníctvom hraníc, ktoré rozdeľujú mapované územie do zón relatívnej homogenity. Umožňuje nám teda jemnejšie prerozdelenie údajov v zdrojovej zóne, a to prostredníctvom rôznych pomocných informácií (ancillary data). Tieto pomocné informácie sa týkajú študovaného územia a zvyčajne sú to vrstvy využitia zeme či krajinnej pokrývky. Prekryvom zdrojových jednotiek (bunky mobilnej siete) a vrstvami pomocných informácií (napr. vrstva zastavaných areálov) dostávame tzv. dazymetrické zóny, ktoré môžeme prepojiť na cieľovú zónu (napr. katastrálne územia obcí).

urbanizovanom území, kde je vertikálna dimenzia a objem budov kľúčovým atribútom, ktorý determinuje priestorovú distribúciu obyvateľov [12, 23, 47].

4.4. Vyhodnotenie modelu

Presnosť výsledkov sme hodnotili porovnaním s referenčnými údajmi z národného cenzu (SODB 2021) a pomocou niekoľkých korelačných mier, vrátane Pearsonovej korelácie (r), Spearmanovej korelácie (s), Kendallovej korelácie ($\boldsymbol{\tau}$), strednej kvadratickej odchýlky (RMSE) a relatívnej celkovej absolútnej chyby (RTAE). Použitie RTAE ponúka možnosť relativizovať výslednú celkovú absolútnu chybu k najlepšiemu dostupnému odhadu populácie v záujmovom území, ktorý môžeme vyjadriť ako súčet referenčných hodnôt všetkých buniek v sledovanom území (S) podľa vzťahu (1):

$$S = \sum_{i=1}^{n} P_i \tag{1}$$

Výsledkom tejto relativizácie je relatívna celková absolútna chyba (RTAE) vyjadrená ako podiel nesprávne alokovanej populácie vzhľadom na počet obyvateľov v území:

$$RTAE = \frac{\sum_{i=1}^{n} \left| X_i - P_i \right|}{\sum_{i=1}^{n} P_i} \times 100\ (\%) = \frac{\sum_{i=1}^{n} \left| X_i - P_i \right|}{S} \times 100\ (\%) \tag{2}$$

kde P_i je veľkosť referenčnej populácie, teda počet obyvateľov (počet obyvateľov evidovaných na súčasný pobyt v zóne i podľa údajov z cenzu 2021) a X_i je modelovaný počet obyvateľov v zóne *i*, definovaný podľa vzťahu (3) ako:

$$X_i = k \times U_i \tag{3}$$

kde U_i je odhadovaný počet používateľov mobilnej siete v zóne i a k je koeficient zabezpečujúci, že:

$$\sum_{i=1}^{n} P_i = \sum_{i=1}^{n} X_i = k \times \sum_{i=1}^{n} U_i \tag{4}$$

4.5. Anonymizácia

Kvôli citlivosti jednotlivých záznamov a dodržiavaniu platných zákonov o ochrane osobných údajov boli údaje anonymizované a agregované prostredníctvom infraštruktúry prevádzkovateľa mobilnej siete. Aj po zoskupení obsahovali niektoré zóny (bunky gridu/obce) len malý počet alokovaných používateľov. Aby sme predišli riziku práce s malými číslami, zónam s 1 až 3 používateľmi sme priradili priemernú hodnotu podľa ich štatistickej distribúcie.

5. VÝSLEDKY

Skôr ako sa pozrieme na výsledky modelu, je potrebné uviesť dôležitú interpretačnú poznámku. Spracované údaje z mobilnej siete tvorí počet alokovaných SIM kariet v priestorových jednotkách (grid, obce). Hoci počet SIM kariet nemôžeme spoľahlivo

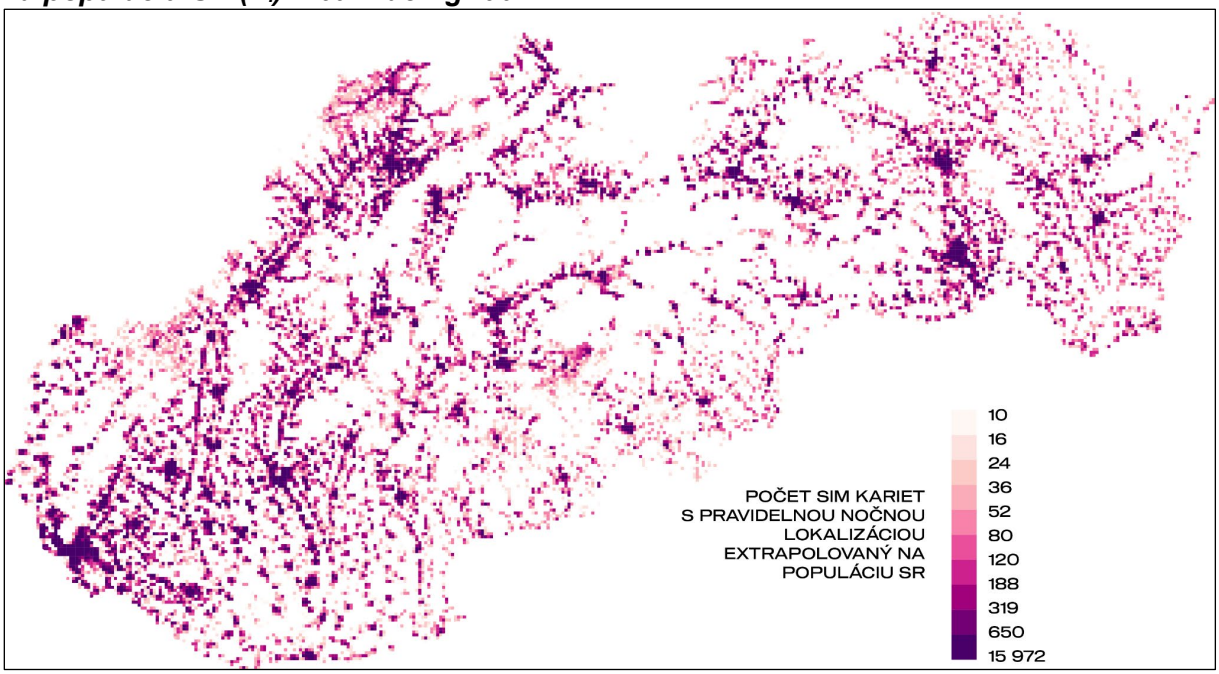
stotožniť s počtom obyvateľov, na účely vyhodnotenie globálnej štatistiky a presnosti modelu prijmeme predpoklad, že jedna SIM karta reprezentuje jedného obyvateľa.

5.1. Vyhodnotenie validity údajov pravidelnej nočnej lokalizácie SIM kariet na úrovni gridu 1 x 1 km

V rámci zvolenej metodiky sme hľadali pravidelné nočné a denné lokalizácie individuálnych používateľov mobilnej siete (SIM kariet), ktoré by mohli slúžiť na odhad prítomnej populácie. V prvom priblížení sme alokovali pravidelnú dennú alebo nočnú polohu mobilného zariadenia do populačného gridu 1 x 1 km, ktorý predstavuje mriežku zloženú z pravidelnej siete štvorcových buniek obsahujúcich počet obyvateľov v Európskom štatistickom systéme. Ako zachytáva obrázok č. 1, rozmiestnenie používateľov mobilnej siete (SIM kariet) prináša očakávaný priestorový obraz s koncentráciou populácie do urbanizovaného územia Slovenska.

Na vyhodnotenie presnosti použitého prístupu sme skonštruovali lineárny regresný model na predikciu počtu obyvateľov, kde počet používateľov s pravidelnou nočnou lokalizáciou v obci je nezávislou premennou a počet obyvateľov zo sčítania obyvateľstva je závislou premennou. Použili sme údaje z posledného cenzu z roku 2021 a z dvoch dostupných kategórií sme zvolili kategóriu súčasný pobyt, ktorý by mal lepšie opisovať skutočné priestorové rozmiestnenie obyvateľstva. Všetky testované datasety priniesli štatisticky významné výsledky ($p < 0,001$) a potvrdili očakávania o schopnostiach odhadnúť priestorové rozmiestnenie populácie. Tesnosť závislosti medzi modelovanými a referenčnými údajmi zachytáva tabuľka č. 4.

Obrázok č. 1: Počet SIM kariet s pravidelnou nočnou lokalizáciou extrapolovaný na populáciu SR (X_i) v bunkách gridu 1 x 1 km



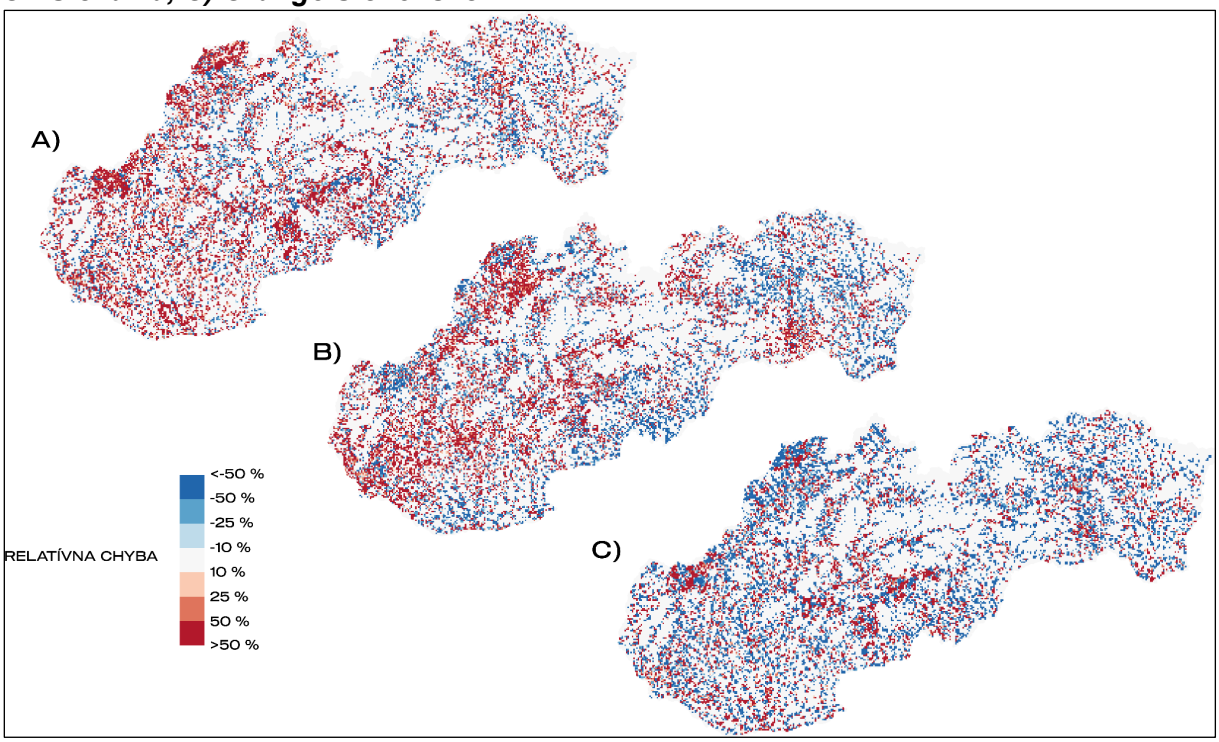
Zdroj údajov: Slovak Telekom (2023)

Tabuľka č. 4: Lineárna regresia medzi modelovanými údajmi (počet SIM kariet s pravidelnou nočnou lokalizáciou) a referenčnými údajmi (súčasný pobyt) v bunkách gridu 1 x 1 km

Dataset	*r*	*s*	τ	RMSE	Relatívna celková absolútna chyba [%]
Slovak Telekom	0,92	0,83	0,77	198	43,63
O2 Slovakia	0,95	0,88	0,82	163	36,95
Orange Slovensko	0,89	0,82	0,75	242	62,50

Zdroj údajov: Slovak Telekom (2023), O2 Slovakia (2023), Orange Slovensko (2023), SODB (2021)

Obrázok č. 2: Relatívna chyba medzi modelovaným počtom obyvateľov (X_i) a referenčným počtom obyvateľov (P_i) v bunkách gridu 1 x 1 km. A) Slovak Telekom, B) O2 Slovakia, C) Orange Slovensko



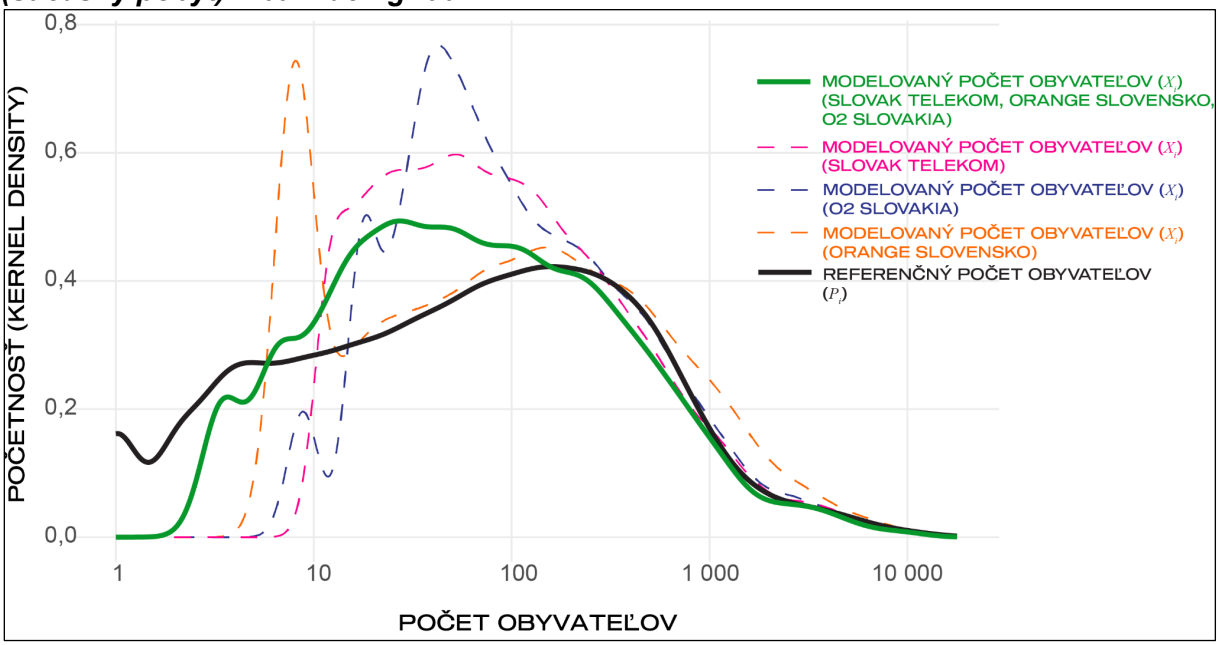
Zdroj údajov: Slovak Telekom (2023), O2 Slovakia (2023), Orange Slovensko (2023), SODB (2021)

Na detailnejšie vyhodnotenie odhadu populácie na úrovni populačného gridu je potrebné venovať pozornosť aj priestorovej lokalizácii buniek s významným nadhodnotením alebo podhodnotením modelovaného počtu rezidentov. Priestorový priemet relatívnej chyby (obrázok č. 2) zachytáva rozdielne oblasti podhodnotenia a nadhodnotenia rozličných operátorov. Táto skutočnosť je pravdepodobne dôsledkom priestorovo diferencovaných trhových podielov mobilných operátorov. Početné nadhodnotené bunky signalizujú nesprávne alokovaných používateľov mobilnej siete do buniek gridu bez evidovanej populácie. Detailný pohľad na nadhodnotené bunky ukazuje, že prevažná väčšina z nich sa nachádza mimo urbanizovaného územia. Priradenie používateľov s pravidelnou nočnou lokalitou do buniek populačného gridu, ktoré evidujú malý alebo žiadny počet obyvateľov, reflektuje limity použitého prístupu, ktorý prerozdeľuje rezidentov podľa objemu rodinných a bytových domov. Ak sa v bunke nachádza čo len jedna rezidenčná budova, model prerozdelí časť používateľov

mobilnej siete. V praxi môže ísť o osamelé budovy, napr. časti hospodárskych dvorov. V niektorých prípadoch však nadhodnotené bunky indikujú lokality, ktoré koncentrujú väčší počet obyvateľov než udávajú referenčné dáta. V praxi sú to napr. lokality novej rezidenčnej výstavby, chalupárske či rekreačné lokality.

Pri celorepublikovom pohľade prostredníctvom spojitého odhadu početnosti (obrázok č. 3) môžeme pozorovať vyšší počet málo saturovaných buniek gridu, ktoré zodpovedajú spomínanej vlastnosti modelu prerozdeľovať používateľov mobilnej siete do rezidenčných budov (bez ohľadu na to, či sú obývané). V prípade údajov zo Slovak Telekomu (2023) je počet nenulových buniek do 10 obyvateľov 5298, naproti tomu buniek s počtom rezidentov evidovaných na súčasný pobyt je len 4184. Výsledkom integrácie údajov od všetkých troch operátorov je rozdelenie početnosti bližšie k referenčným údajom. V grafe môžeme pozorovať podobný priebeh krivky najmä v populačne stredne veľkých a veľkých buniek gridu.

Obrázok č. 3: Spojitý odhad početnosti (kernel density estimation) údajov z mobilnej siete (pravidelná nočná lokalita SIM kariet) a referenčných údajov z národného cenzu (súčasný pobyt) v bunkách gridu 1 x 1 km



Zdroj údajov: Slovak Telekom (2023), O2 Slovakia (2023), Orange Slovensko (2023), SODB (2021)

5.2. Vyhodnotenie validity údajov pravidelnej nočnej lokalizácie SIM kariet na úrovni obcí

Pri spracovaní a vyhodnotení odhadu rezidenčnej populácie na úrovni obcí sme postupovali analogicky ako v prípade gridu. Vzhľadom na nižší počet cieľových zón (2 927 obcí) než buniek gridu (50 661) sú výsledky globálnej štatistiky ešte spoľahlivejšie, než v prípade gridu (tabuľka č. 5).

Tabuľka č. 5: Lineárna regresia medzi modelovanými údajmi (počet SIM kariet s pravidelnou nočnou lokalizáciou) a referenčnými údajmi (súčasný pobyt) v obciach SR

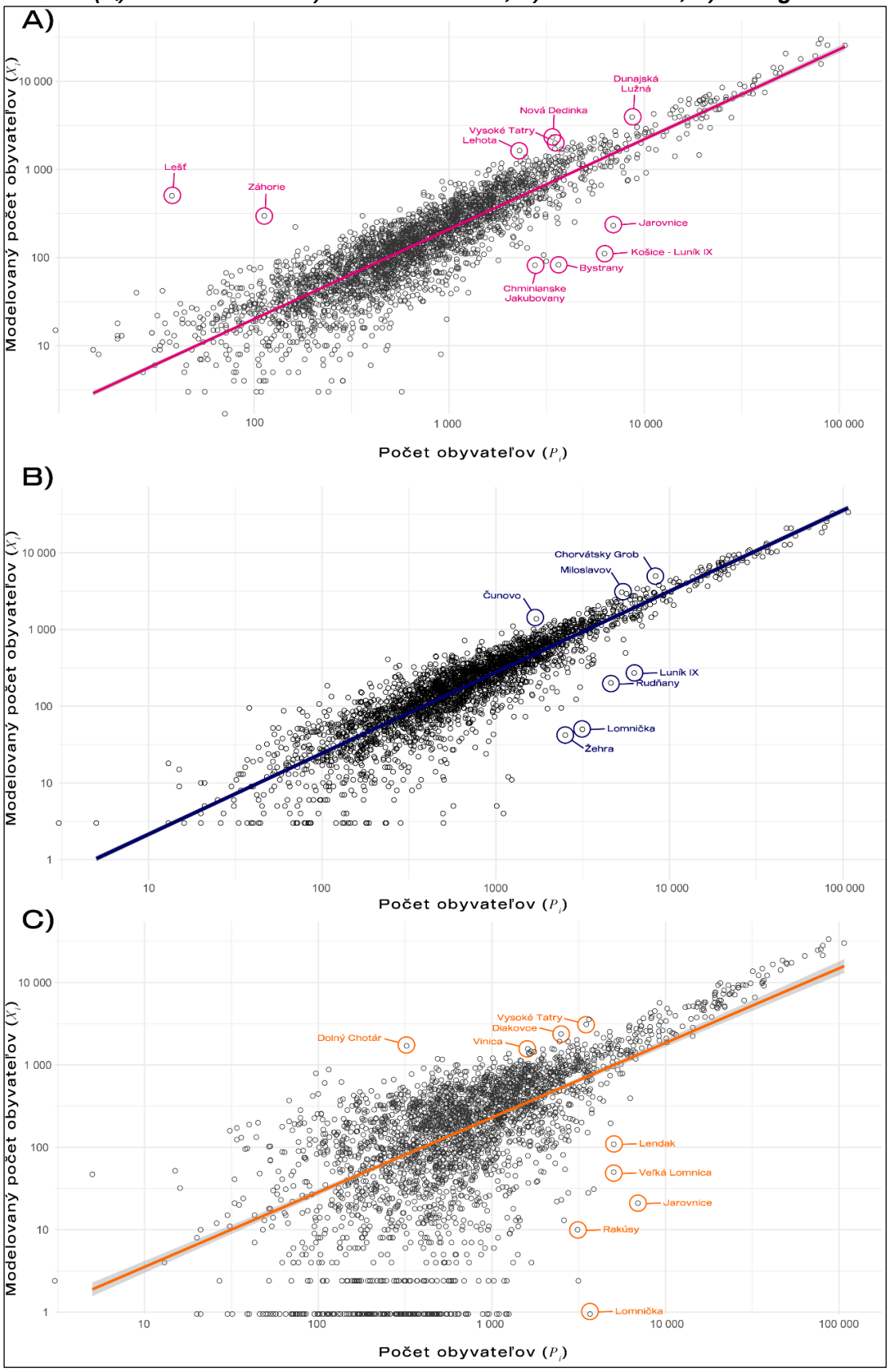
Dataset	*r*	*s*	τ	RMSE	Relatívna celková absolútna chyba [%]
Slovak Telekom	0,97	0,73	0,90	1576	28,74
O2 Slovakia	0,98	0,74	0,90	1036	22,49
Orange Slovensko	0,96	0,49	0,67	1288	37,86

Zdroj údajov: Slovak Telekom (2023), O2 Slovakia (2023), Orange Slovensko (2023), SODB (2021)

Hoci regresia priniesla veľmi uspokojivé výsledky, na hlbšie pochopenie použitého modelu odhadu populácie je potrebné preskúmať odľahlé hodnoty. Pri ich interpretácii budeme pozornosť venovať len stredne veľkým a veľkým obciam, keďže pri malých obciach je väčší rozptyl prirodzený a spôsobený nerovnomerným pokrytím územia mobilným signálom, ako aj rozkolísaným trhovým podielom mobilných operátorov v malých územných jednotkách. Ako zachytáva obrázok č. 4, v podhodnotených obciach môžeme pozorovať niekoľko spoločných znakov. V početných prípadoch ide o obce s marginalizovanými rómskymi komunitami (Lomnička, Stráne pod Tatrami, Jarovnice, Chminianske Jakubovany a pod.). Podhodnotenie modelovanej populácie je pravdepodobne spôsobené špecifickým využívaním služieb mobilnej siete (predplatené karty), nižšou penetráciou mobilných telefónov, ako aj štruktúrou populácie s početnou detskou zložkou (deti, ktoré mobilný telefón nemajú). Svoju úlohu môže zohrať aj skutočnosť, že segregované rómske osady sú často vytvorené z budov, ktoré nemusia byť evidované v mapových vrstvách, prípadne nemajú atribút rodinného domu. Pri transformácii údajov z vyžarovacích polygónov sa používatelia neprerozdelia do týchto území, keďže tieto budovy neboli súčasťou pomocných vrstiev. Závažnosť uvedených limitov nie je možné spoľahlivo vyhodnotiť bez podrobnej hĺbkovej analýzy.

V nadhodnotených obciach nie je možné identifikovať jednoznačného spoločného menovateľa. Ukazuje sa však, že nadhodnotené obce sa často nachádzajú v blízkosti Bratislavy (Dunajská Lužná, Nová Dedinka, Kvetoslavov), kde môžeme predpokladať vplyv väčšej skupiny neprihlásených rezidentov na trvalý (resp. súčasný) pobyt. Ďalšiu skupinu tvoria obce s dominantou rekreačnou funkciou (Vysoké Tatry, Demänovská dolina), kde je väčší počet pravidelne nocujúcich SIM kariet spôsobený nezanedbateľným počtom ľudí pracujúci v turizme. Svoju úlohu môžu zohrať i rezidenti dlhodobo bývajúci v rekreačných objektoch. Medzi nadhodnotenými obcami nachádzame aj viaceré nešpecifické obce, ktorých vyšší počet trvalo nocujúcich používateľov mobilnej siete môže byť výsledkom rozmanitých dôvodov, predovšetkým však nadpriemerného trhového podielu daného mobilného operátora.

Obrázok č. 4: Porovnanie modelovaného počtu obyvateľov (X_i) s referenčným počtom obyvateľov (P_i) v obciach SR. A) Slovak Telekom, B) O2 Slovakia, C) Orange Slovensko



Zdroj údajov: Slovak Telekom (2023), O2 Slovakia (2023), Orange Slovensko (2023), SODB (2021)

5.3. Vyhodnotenie validity integrovaných údajov pravidelnej nočnej lokalizácie na úrovni obcí

Výsledky regresie integrovaných údajov ponúka tabuľka č. 6. Dosiahnuté skóre relatívnej celkovej absolútnej chyby na úrovni 21 % môžeme považovať za veľmi dobrý výsledok, na ktorom sa významne podieľa výber vhodnej pomocnej vrstvy na interpoláciu údajov, ako aj skutočnosť, že údaje od jednotlivých operátorov sa vzájomne vhodne dopĺňajú, čím prispievajú k zvýšeniu presnosti a spoľahlivosti modelu nočnej populácie. Domnievame sa však, že existuje ešte nezanedbateľný priestor na zlepšenie presnosti modelu. Rezervy vidíme najmä v konštrukcii spoľahlivejšej pomocnej vrstvy, ako aj v odstránení „šumu“ vo vstupných údajoch. Podrobnejšie sa tejto téme venujeme v 6. kapitole.

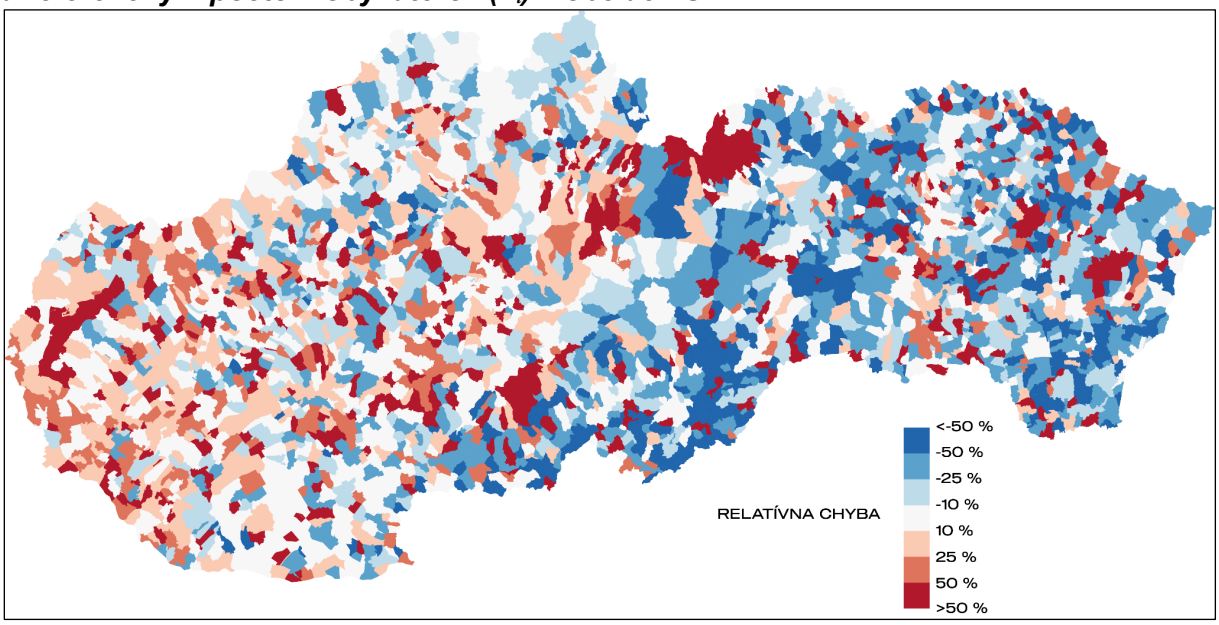
Tabuľka č. 6: Lineárna regresia medzi modelovanými údajmi (počet SIM kariet s pravidelnou nočnou lokalizáciou) a referenčnými údajmi (súčasný pobyt) v obciach SR

Dataset	*r*	*s*	τ	RMSE	Relatívna celková absolútna chyba [%]
Slovak Telekom + O2 Slovakia + Orange Slovensko	0,98	0,73	0,89	1078	21,23

Zdroj údajov: Slovak Telekom (2023), O2 Slovakia (2023), Orange Slovensko (2023), SODB (2021)

Výsledný pohľad na priestorovú diferenciáciu nadhodnotených a podhodnotených obcí z hľadiska modelu nočnej populácie (obrázok č. 5) prináša známy obraz o Slovensku, ktorý poznáme z analýzy rozmanitých sociálno-ekonomických ukazovateľov. Interpretácia však nie je jednoduchá, keďže do výsledku môže vstupovať množstvo rozličných faktorov, ktorých povahu a vplyv nie vždy vieme spoľahlivo vyhodnotiť. K interpretačnej zdržanlivosti nás nútia aj početné limity použitého modelu nočnej populácie. Napriek uvedenému sa pokúsime aspoň naznačiť niektoré súvislosti. Východiskom interpretácie môže byť analýza diferencií medzi súčasným a trvalým pobytom a z toho vyplývajúcich disproporcií medzi evidovaným a reálnym priestorovým rozmiestnením obyvateľstva (podrobnejšie sa téme venuje [30]. Vysokú mieru prisťahovaných na súčasný pobyt zaznamenali najmä obce v zázemí Bratislavy a Košíc, v celoslovenskom pohľade však môžeme jednoznačne identifikovať východo-západný gradient, ktorý pôsobí v prospech vyššej koncentrácie obyvateľov na západnom Slovensku. Výsledky modelu nočnej populácie by v tomto kontexte mohli naznačovať podobný trend. Pod vplyvom suburbanizácie a rozmanitých metropolizačných efektov sa môže dosahovať v Bratislave a jej zázemí ešte vyššia koncentrácia populácie, než zachytávajú oficiálne údaje. V prospech tejto argumentácie pôsobí aj veková štruktúra obyvateľstva, ktorá do suburbanizovaných obcí prináša najmä mladé rodiny [29]. Protiargumentom je však skutočnosť, že v tomto sociálne a ekonomicky silnom regióne môžeme očakávať vyššiu penetráciu mobilných zariadení než vo zvyšku Slovenska a teda aj početné duplicity spôsobené vlastníctvom dvoch či viacerých mobilných telefónov jednej osoby. Uvedené argumenty nie je možné v súčasnosti spoľahlivo vyhodnotiť a v ďalšom postupe pri spracovaní týchto unikátnych časopriestorových údajov by sme odporúčali podrobne analyzovať možné vplyvy vybraných sociálnych, ekonomických či sídelných faktorov na identifikované diferencie medzi evidovaným počtom obyvateľov a počtom používateľov mobilnej siete s pravidelnou nočnou lokalizáciou.

Obrázok č. 5: Relatívna chyba medzi modelovaným počtom obyvateľov (X_i) a referenčným počtom obyvateľov (P_i) v obciach SR



Zdroj údajov: Slovak Telekom (2023), O2 Slovakia (2023), Orange Slovensko (2023), SODB (2021)

6. ZHODNOTENIE PREDNOSTÍ A LIMITOV MODELU

Lokalizačné údaje mobilných telefónov prinášajú nový zdroj údajov o populácii, ktorý môže poskytovať významné informácie k už existujúcim zdrojom údajov, akými sú celoplošné cenzy a výberové zisťovania. Na rozdiel od týchto „tradičných" zdrojov však prinášajú viaceré výhody, ktoré charakterizuje veľkosť vzorky, rýchlosť a frekvencia zberu údajov, ako aj finančné náklady potrebné na ich obstaranie.

6.1. Prednosti modelu

Metodika vyvinutá v tejto štúdii prináša niekoľko dôležitých inovácií, ktoré rozvíjajú oblasť využitia údajov z mobilnej siete aj v širšom (európskom) kontexte. Ide o nasledujúce charakteristiky:

- Model je založený na spracovaní signalizačných údajov z mobilnej siete. Ide o automaticky generované záznamy, ktoré produkuje mobilná sieť pri pravidelných kontrolách pripojených zariadení. Vzhľadom na veľký počet takýchto záznamov je ich spracovanie pomerne náročné a ich využitie v obdobných analýzach bolo dosiaľ iba sporadické. Napriek tomu ich význam v priestorových analýzach bude pravdepodobne narastať. Dôvodom sú zmeny vo využívaní mobilného telefónu. Kým v minulosti sme ho využívali najmä na volania a SMS správy, v súčasnosti toto využitie ustupuje v prospech dátových služieb (sociálne siete, odosielanie a prijímanie okamžitých správ a pod.) a pasívneho využívania telefónu (napr. rôzne notifikácie v aplikáciách si nevyžadujú aktívnu interakciu).
- Model pracuje s podrobnou topológiou mobilnej siete, kde využíva detailný model vyžarovacích polygónov. Dosiaľ bežne používaným prístupom bolo použitie polohy antény mobilnej siete, alebo využitie jednoduchej teselácie (vyplnenie roviny pomocou jedného alebo viacerých geometrických útvarov bez prekrývania a medzier) založenej na Voronoiových polygónoch (napr. [10, 31]. Na transformáciu údajov z vyžarovacích polygónov buniek mobilnej siete

do cieľových priestorových rámcov (obce, grid) model využíva dazymetrickú interpoláciu mapujúcu alokovaných používateľov s pravidelnou nočnou/dennou lokalitou do objemu budov. Tým sa zásadne spresňuje odhad priestorového rozmiestnenia populácie, keďže koncentrácia obyvateľov je prirodzene viazaná na sídelnú zástavbu.

- Cieľovou priestorovou zónou modelu nie sú len katastrálne územia obcí, ale aj bunky gridu 1x1 km. Odhad populácie v tomto detaile nie je bežný a prináša mnohé úskalia. Napriek tomu model priniesol veľmi uspokojivé výsledky aj v tejto mierke. Konštrukcia vektorov medzi pravidelnou nočnou lokalizáciou a pravidelnou dennou lokalizáciou používateľa mobilnej siete prináša unikátny pohľad najmä na intraurbánnu mobilitu. Tento typ informácií bol doposiaľ možný len prostredníctvom nákladných prieskumov mobility a výberových zisťovaní.

Hoci lokalizačné údaje z mobilnej siete prinášajú bezprecedentný rozsah informácií o pobyte a pohybe veľkej časti populácie, je potrebné si uvedomiť, že ich charakter a kvalitu ovplyvňuje použitá technológia, kontext ich vzniku a spôsob formalizovania dátového modelu (entity, kategórie, atribúty, väzby). Práca s týmito údajmi si navyše vyžaduje opatrné prehodnocovanie postavené na kontextuálnych znalostiach s ohľadom na charakter analýzy a spôsob interpretácie.

6.2. Limity modelu

Pri interpretácii pasívnych lokalizačných údajov mobilných zariadení v sieťovej lokalizácii je dôležité si uvedomiť viaceré limity, ktoré vyplývajú z princípov prevádzky mobilnej siete a z právnych podmienok spracovania údajov jej používateľov. Vo všeobecnosti ide najmä o obmedzenia spojené s ochranou súkromia používateľov mobilnej siete, s limitovanými socio-demografickými štruktúrami, ako aj so špecifikami mobilnej komunikácie. V stručnosti sa pokúsime upozorniť na tieto kľúčové limity:

- Počet SIM kariet nemožno spoľahlivo stotožniť s počtom osôb (individuálnych používateľov). Dôvodom je skutočnosť, že nemôžeme predpokladať, že každý obyvateľ disponuje mobilným zariadením. Takisto nedokážeme spoľahlivo vylúčiť osoby využívajúce viaceré SIM karty.
- Presnosť lokalizácie závisí od architektúry mobilnej infraštruktúry, typu antén (2G/3G/4G/5G) a reliéfu. V mestskom prostredí sa pohybuje rádovo v stovkách metrov, vo voľnej krajine dosahuje rádovo kilometre. Presnosť lokalizácie ovplyvňuje aj spoľahlivosť modelu vyžarovacích polygónov. Ich tvar nemusí presne zodpovedať skutočnému pokrytiu mobilným signálom a v čase sa môže meniť (napr. vplyvom počasia). Prístupy, akým sa modeluje topológia mobilnej siete sú pri jednotlivých prevádzkovateľoch mobilnej siete rozdielne. Príkladom je model vyžarovacích polygónov Orange Slovensko, ktorý využíva polohu BTS, azimut (smer) a silu signálu. Výsledkom je vyžarovací polygón v tvare kruhovej výseče. Je zrejmé, že takto skonštruovaný polygón len veľmi voľne opisuje skutočný tvar územia, ktorý pokrýva anténa mobilnej siete.
- Mobilné zariadenia sa spravidla pripájajú na najbližšiu základňovú stanicu. V systéme priradenia konkrétnej bunky k telefónu však zohrávajú úlohu desiatky faktorov (napr. intenzita signálu, atmosférické podmienky, zaťaženie infraštruktúry, plán údržby a pod.). Je preto možné, že používateľ na rovnakom mieste, ktorý uskutoční napr. tri hovory, sa pripojí k trom rôznym anténam a teda k fyzickému presunu používateľa nedôjde, ale v lokalizačných záznamoch pohyb nastane. Táto skutočnosť môže ovplyvniť aj identifikáciu pravidelnej dennej/nočnej lokality SIM karty.

- Pasívne lokalizačné údaje mobilnej siete nedokážu poskytnúť informácie o type a trvaní konkrétnych aktivít. Príkladom je odhad dochádzkových tokov medzi pravidelnou dennou a nočnou lokalitou používateľa mobilnej siete. Je potrebné poznamenať, že nemusí ísť nevyhnutne o dochádzku do zamestnania či školy. V realite môže takáto väzba pokrývať veľmi rozdielne situácie. Pri použití údajov o využití zeme či type urbanizovaného prostredia sa môžeme pokúsiť niektoré aktivity zmysluplne odhadnúť.
- Jednotliví operátori mobilnej siete môžu byť zacielení na špecifickú skupinu používateľov či používateliek (napr. na mladších používateľov, podnikateľov, seniorov a pod.). Extrapolácia údajov na celú populáciu je tak značne obmedzená. Pri spracovaní údajov od viacerých operátorov, ktorí spoločne pokrývajú väčšinu populácie, sa však tento problém do značnej miery eliminuje. Napriek tomu niektoré vekové kategórie ostávajú poddimenzované (najmä malé deti a seniori). Ako sme naznačili v predchádzajúcich častiach príspevku, reprezentatívnosť údajov mobilnej siete je problematická najmä v obciach s marginalizovanými rómskymi komunitami. V týchto skupinách obyvateľstva môžeme predpokladať nielen nižšiu penetráciu mobilných zariadení (štruktúra populácie s veľkou detskou zložkou), ale aj špecifické využívanie mobilnej siete (predplatené karty minoritných prevádzkovateľov mobilnej komunikácie).
- Na zabezpečenie anonymity a bezpečnosti údajov je potrebná eliminácia záznamov pod určitú hranicu. V predkladanej analýze to bola hranica na úrovni 3 SIM kariet v danej územnej jednotky alebo toku. Hoci interpretačný dôraz sa kladie najmä na veľké koncentrácie používateľov alebo veľké toky SIM kariet v rámci infraštruktúry mobilnej siete, vynechaním málo početných lokalizácií sa môžu závažne skresliť výsledky, najmä pokiaľ je zámerom extrapolovať počty používateľov mobilnej siete na celú populáciu. Takýto prístup nežiaduco skresľuje aj vektorové dáta, ktoré majú prirodzene výrazné zošikmenie distribúcie (prevažujú málopočetné toky).
- Algoritmus na identifikáciu pravidelnej nočnej a dennej lokality nezohľadňuje rozdielnu hustotu záznamov jednotlivých mobilných operátorov. Tento limit bol identifikovaný až počas implementačnej fázy a vzhľadom na vopred stanovenú metodiku (postavenú na jednotnom algoritme) nebolo možné vykonať dodatočné zmeny. Ako perspektívna možnosť sa javí individuálne „vyladenie“ algoritmu vzhľadom na špecifiká jednotlivých mobilných operátorov.
- Použitá metodika spracovania lokalizačných údajov z mobilnej siete je založená na spracovaní surových záznamov (signalingových, CDR). Vzhľadom na dlhšie obdobie pozorovania však pri niektorých mobilných operátoroch došlo k čiastočnej agregácii starších údajov. Výsledkom tejto semiagregácie je „zriedenie“ záznamov individuálnych používateľov, čo môže mať vplyv na spoľahlivosť identifikovania pravidelnej nočnej a dennej lokality SIM karty (a z toho odvodených vektorov dochádzky). Vzhľadom na to, že nám neboli známe detaily uvedenej semiagregácie, nebolo možné dodatočne prispôsobiť algoritmy.
- Pri transformácii údajov z vyžarovacích polygónov mobilnej siete do cieľových priestorových jednotiek (obec/grid) sme použili dazymetrickú transformáciu s využitím objemu budov. Hoci táto metóda priniesla veľmi uspokojivé výsledky, má nepochybne aj limity. Tým prvým je skutočnosť, že použitá vrstva budov pochádza z databázy ZBGIS (2017), takže neodráža aktuálny stav urbanizovaného územia. V praxi ide najmä o lokality s dynamickou výstavbou okolí veľkých miest. V databáze sú aj početné chyby, resp. neúplné

atribúty o stave a type budov. Druhým limitom je použitie len niektorých typov budov na konštrukciu pomocnej vrstvy. Na nočnú lokalizáciu to boli rodinné domy, bytové domy a polyfunkčné domy. V praxi však množstvo ľudí trvalo obýva aj iné typy budov, napr. chaty. Na túto skutočnosť je vhodné prihliadať najmä v oblastiach s rozptýleným osídlením, v obciach s marginalizovanými rómskymi komunitami a v záhradkárskych či rekreačných lokalitách.

Z uvedených dôvodov je zrejmé, že môžu existovať značné rozdiely medzi kvantitatívnymi charakteristikami populácie vypočítanými na základe oficiálnych dát (napr. počet dochádzajúcich do obce) a údajmi z mobilnej siete. Vo všeobecnosti je potrebné sa vyhnúť práci na úrovni absolútnych hodnôt a preferovať skôr prístup využívajúci relativizované údaje (napr. podiel dochádzajúcich do obce namiesto počtu dochádzajúcich do obce).

Napriek uvedeným limitom priniesli výsledky spracovania lokalizačných údajov mobilnej siete veľmi solídne výsledky, ktoré môžu prispieť k hlbšiemu poznaniu priestorového rozmiestnenia a dynamiky obyvateľstva. Pri rešpektovaní uvedených limitov dostávame perspektívne dáta, vhodné na rozmanité analýzy na úrovni regiónov (obce) alebo lokalít (grid).

Hoci navrhnutá metodika priniesla uspokojivé výsledky, z nášho pohľadu ide len o východiskový bod, na základe ktorého chceme stimulovať ďalšiu odbornú diskusiu medzi poskytovateľmi služieb mobilnej komunikácie a verejnými inštitúciami, s cieľom zosúladiť predstavy a možnosti implementácie týchto unikátnych údajov. Predpokladáme, že navrhnutá metodika budú ďalej zdokonalená a prinesie ešte spoľahlivejšie výsledky.

7. ZÁVER

Lokalizačné údaje mobilnej siete predstavujú perspektívny zdroj údajov o priestorovom rozmiestnení a mobilite obyvateľov a sľubný výskumný smer s bohatým využitím, osobitne v riadiacej a plánovacej praxi. Stále sme však ďaleko od toho, aby sa tieto údaje stali bežnou súčasťou priestorových analýz. Je však zrejmé, že rozvoj v tejto oblasti výskumu dospel dostatočne ďaleko, aby prekročil prah experimentálneho používania a stal sa súčasťou aplikovanej praxe. Ako dokumentuje narastajúci počet štúdií, údaje z mobilnej siete poskytujú porovnateľnú presnosť a spoľahlivosť ako štatistické údaje o obyvateľstve [1, 37, 42]. Konceptuálny a metodický aparát, ktorý vyprodukovala geografia a iné vedné disciplíny v oblasti spracovania údajov z mobilnej siete, je skutočne rozsiahly a prináša praktické riešenia a významné výsledky. Najväčšou prekážkou pre rozvoj tejto perspektívnej oblasti výskumu je v súčasnosti dostupnosť údajov, najmä možnosť analýzy na úrovni buniek mobilnej siete a so zohľadnením doplnkových údajov potrebných na korektnú interpoláciu (tvar vyžarovacích polygónov) a extrapoláciu (poznanie regionálne diferencovaného trhového podielu) údajov. Tento typ údajov však nemôžu poskytnúť spoločnosti spracúvajúce údaje mobilnej siete pre koncových zákazníkov. Principiálne ide o citlivé údaje, a tak je zrejmé, že ďalší rozvoj analýz založených na údajoch mobilnej siete sa nezaobíde bez úzkej spolupráce s mobilnými operátormi. Kľúčovou úlohou bude nájsť taký model spolupráce, ktorý minimalizuje riziká spojené so spracovaním mikroúdajov a zároveň prináša potenciálne inovácie do spracovania údajov mobilnej siete. Je teda dôležité, aby akademický či verejný sektor prichádzal s takými návrhmi spolupráce, ktoré neohrozia kredibilitu mobilných operátorov,

a súčasne prinesú inovatívne metódy spracovania údajov, ktoré v konečnom dôsledku môžu byť prínosné aj pre mobilných operátorov. Prekonanie bariér bude nepochybne výhodné pre obe strany. Ak sa nájde model dlhodobej spolupráce, môžu verejné či výskumné inštitúcie získať spoľahlivé údaje na pravidelnej báze (mesiac/rok). Mobilní operátori zas môžu využiť príležitosť ako monetizovať „vedľajší produkt“ fungovania mobilnej siete.

Lokalizačné údaje z mobilnej siete budú nepochybne dôležitým zdrojom informácií na prijímanie kvalifikovaných rozhodnutí v početných krízach, ktorým čelí a bude čeliť naša spoločnosť. Bolo by chybou tento cenný zdroj priestorových údajov nevyužiť.

LITERATÚRA

1] AASA, A. – KAMENJUK, P. – SALUVEER, E. – ŠIMBERA, J. – RAUN, J.: Spatial interpolation of mobile positioning data for population statistics. In: Journal of Location Based Services, 2021, č. 4, p. 239 – 260.
[2] AHAS, R. – AASA, A. – MARK, Ü. – PAE, T. – KULL, A.: Seasonal tourism spaces in Estonia: Case study with mobile positioning data. In: Tourism Management, 2007a, č. 3, s. 898 – 910.
[3] AHAS, R. – AASA, A. – SILM, S. – AUNAP, R. – KALLE, H. – MARK, Ü.: Mobile positioning in space–time behaviour studies: social positioning method experiments in Estonia. In: Cartography and Geographic Information Science, 2007b, č. 4, s. 259 – 273.
[4] AHAS, R. – MARK, Ü: Location based services – new challenges for planning and public administration? In: Futures, 2005, č. 6, s. 547 – 561.
[5] AHAS, R. – SILM, S. – JÄRV, O. – SALUVEER, E.– TIRU, M.: Using mobile positioning data to model locations meaningful to users of mobile phones. In: Journal of Urban Technology, 2010, č. 1, s. 3 – 27.
[6] BATISTA E SILVA, F. – CRAGLIA, M. – FREIRE, S. – ROSINA, K. – LAVALLE, C. – MARIN, M. – SCHIAVINA, M.: Enhancing activity and population mapping. European Commission: Joint Research Centre, 2016.
[7] BATISTA E SILVA, F. – FREIRE, S. – SCHIAVINA, M. – ROSINA, K. – MARÍN-HERRERA, M. A. – ZIEMBA, L. – CRAGLIA, M. – KOOMEN, E. – LAVALLE, C.: Uncovering temporal changes in Europe’s population density patterns using a data fusion approach. In: Nature Communications, 2020, č. 1, s. 1 –11.
[8] BATTY, M.: The new science of cities. MIT press, 2013
[9] BENGTSSON, L. – LU, X. – THORSON, A. – GARFIELD, R. – VON SCHREEB, J.: Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. In: PLoS Medicine, 2011, č. 8, e1001083.
[10] BERGROTH, C. – JÄRV, O. – TENKANEN, H. – MANNINEN, M. – TOIVONEN, T.: A 24-hour population distribution dataset based on mobile phone data from Helsinki Metropolitan Area, Finland. In: Scientific data, 2022, č. 1, s. 39.
[11] BEZÁK, A.: Vnútorné migrácie na Slovensku: súčasné trendy a priestorové vzorce. In. Geografický časopis, 2006, č. 1, s. 15 – 44.
[12] BILJECKI, F. – ARROYO OHORI, K. – LEDOUX, H. – PETERS, R. – STOTER, J.: Population estimation using a 3D city model: A multi-scale country-wide study in the Netherlands. In. PloS one, 2016, č. 6, e0156808.
[13] BLEHA, B.– POPJAKOVÁ, D.: Migrácia ako dôležitý determinant budúceho vývoja na lokálnej úrovni–príklad Petržalky. In: Geografický časopis, 2007, č. 3, s. 265 – 291.

[14] CASTELLS, M. – FERNANDEZ-ARDEVOL, M. – QIU, J. L. – SEY, A.: Mobile communication and society: A global perspective. MIT Press, 2009.
[15] CSÁJI, B. C. – BROWET, A. – TRAAG, V. A. – DELVENNE, J. C. – HUENS, E. – VAN DOOREN, P. – SMOREDA, Z. – BLONDEL, V. D.: Exploring the mobility of mobile phone users. In: Physica A: Statistical Mechanics and its Applications, 2013, č. 6, s. 1459 – 1473.
[16] DEVILLE, P. – LINARD, C. – MARTIN, S. – GILBERT, M. – STEVENS, F. R. – GAUGHAN, A. E. – BLONDEL, V. D. – TATEM, A. J.: Dynamic population mapping using mobile phone data. In: Proceedings of the National Academy of Sciences, 2014, č. 45, s. 15888 – 15893.
[17] DOUGLASS, R. W. – MEYER, D. A. – RAM, M. – RIDEOUT, D. – SONG, D.: High resolution population estimates from telecommunications data. In: EPJ Data Science, 2015, č. 4, s. 1 – 13.
[18] GOODCHILD, M. F., ANSELIN, L., DEICHMANN, U.: A framework for the areal interpolation of socioeconomic data. In: Environment and planning A: Economy and Space, 1993, č. 3, s. 383 – 397.
[19] GRANTZ, K. H. – MEREDITH, H. R. – CUMMINGS, D. A. – METCALF, C. J. E. – GRENFELL, B. T. – GILES, J. R. – MEHTA, S. – SOLOMON, S. – LABRIQUE, A. – KISHORE, N. – BUCKEE, C. – WESOLOWSKI, A.: The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. In: Nature Communications, 2020, č. 1, s. 1 – 8.
[20] HORANONT, T.: A Study on Urban Mobility and Dynamic Population Estimation by Using Aggregate Mobile Phone Sources. CSIS Discussion Paper No. 115, 2012.
[21] CHEN, M. – CLARAMUNT, C. – ÇÖLTEKIN, A. – LIU, X. – PENG, P. – ROBINSON, A. C. – LÜ, G.: Artificial intelligence and visual analytics in geographical space and cyberspace: Research opportunities and challenges. Earth-Science Reviews, 2023,104438.
[22] JÄRV, O. – AHAS, R. – WITLOX, F.: Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. In: Transportation Research Part C: Emerging Technologies, 2014, s. 122 – 135.
[23] JÄRV, O., H. – TENKANEN, T. TOIVONEN.:“Enhancing Spatial Accuracy of Mobile Phone Data Using Multi-temporal Dasymetric Interpolation.” In: International Journal of Geographical Information Science, 2017, č. 8, s. 1630 – 1651. Taylor & Francis
[24] KRINGS, G. – CALABRESE, F. – RATTI, C. – BLONDEL, V. D.: Urban gravity: a model for inter-city telecommunication flows. In. Journal of Statistical Mechanics: Theory and Experiment, 2009, L07003.
[25] LANGFORD, M. – MAGUIRE, D. J. – UNWIN, D. J.: The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In. Masser, I., Blakemore M. B. (Eds.). Handling geographic information. Essex (Longman Scientific & Technical), 1991, s. 55 – 77.
[26] LONG, A. – CARNEY, F. – KANDT, J.: Who is returning to public transport for non-work trips after COVID-19? Evidence from older citizens' smart cards in the UK's second largest city region. In: Journal of Transport Geography, 2023, 103529.
[27] MARTIN, D. – COCKINGS, S. – LEUNG, S.: Developing a flexible framework for spatiotemporal population modeling. In: Annals of the Association of American Geographers, 2015, č. 4, s. 754 – 772.
[28] MOUNTAIN, D. – RAPER, J.: Modelling human spatio-temporal behaviour: a challenge for location-based services, In: *Proceedings of the Sixth International*

Conference on GeoComputation, University of Queensland, Brisbane, Australia, 24 – 26 September, 2001.
[29] NOVOTNÝ, L. – PREGI, L.: Selective migration of population subgroups by educational attainment in the urban region of Bratislava. In: Geografický časopis, 2017, č. 1, s. 21 – 39.
[30] ÓVÁRI, K. – KOČIŠ, M.: Priestorová diferenciácia rozmiestnenia obyvateľstva v kontexte súčasného pobytu v Slovenskej republike k 1. 1. 2021. In: Slovenská štatistika a demografia, 2023, č. 2, s. 5 – 29.
[31] PENG, Z. – WANG, R. – LIU, L. – WU, H.: Fine-Scale Dasymetric Population Mapping with Mobile Phone and Building Use Data Based on Grid Voronoi Method. In: ISPRS International Journal of Geo-Information, 2020, roč. 9, č. 6, 344.
[32] PODOLÁK, P.: Centre and Hinterland-Migration Relations. In: Folia Geographica, 2002, 5, s. 143 – 145.
[33] RATTI, C. – FRENCHMAN, D. – PULSELLI, R. M. – WILLIAMS, S.: Mobile landscapes: using location data from cell phones for urban analysis. In: Environment and planning B: Planning and Design, 2006, č. 5, s. 727 – 748.
[34] READES, J. – CALABRESE, F. – RATTI, C.: Eigenplaces: analysing cities using the space–time structure of the mobile phone network. In: Environment and Planning B: Planning and Design, 2009, č. 5, s. 824 – 836.
[35] READES, J. – CALABRESE, F. – SEVTSUK, A. – RATTI, C.: Cellular census: Explorations in urban data collection. In: IEEE Pervasive computing, 2007, č. 3, s. 30 – 38.
[36] RICCIATO, F. – WIDHALM, P. – CRAGLIA, M. – PANTISANO, F.: Estimating population density distribution from network-based mobile phone data. Luxembourg (Publications Office of the European Union), 2015.
[37] RICCIATO, F. – WIDHALM, P. – PANTISANO, F. – CRAGLIA, M.: Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. In: Pervasive and Mobile Computing, 2017, s. 65 – 82.
[38] SHELLER, M.: Mobile publics: beyond the network perspective. In: Environment and Planning D: Society and Space, 2004, č. 1, s. 39 – 52.
[39] SHOVAL, N.: Sensing human society. In: Environment and Planning B: Planning and Design, 2007, č. 2, s. 191 – 195.
[40] STEENBRUGGEN, J. – BORZACCHIELLO, M. T. – NIJKAMP, P. – SCHOLTEN, H.: Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. In. Geo Journal, 2013, č. 2, s. 223 – 243.
[41] ŠVEDA, M. – PODOLÁK, P.: Fenomén neúplnej evidencie migrácie v suburbánnej zóne (na príklade zázemia Bratislavy). In: Geografický časopis, 2014, č. 2, s. 115 – 132.
[42] ŠVEDA, M. – SLÁDEKOVÁ MADAJOVÁ, M.: Estimating distance decay of intra-urban trips using mobile phone data: The case of Bratislava, Slovakia. In: Journal of Transport Geography, 2023, 103552.
[43] ŠVEDA, M. – SLÁDEKOVÁ MADAJOVÁ, M. – BARLÍK, P. – KRIŽAN, F. – ŠUŠKA, P.: Mobile phone data in studying urban rhythms: towards an analytical framework. In: Moravian Geographical Reports, 2020, č. 4, s. 248 – 258.
[44] ŠVEDA, M. – SLÁDEKOVÁ MADAJOVÁ, M. – ROSINA, K. – HURBÁNEK, P. – FORSTL, F. – ZÁBOJ, P. – VÝBOŠŤOK, J.: When spatial interpolation matters: seeking an appropriate data transformation from the mobile network for population estimates, In: Computers, Environment and Urban Systems, 2023, rukopis.

[45] TOBLER, W. – DEICHMANN, U. – GOTTSEGEN, J. – MALOY, K.: World population in a grid of spherical quadrilaterals. In: International Journal of Population Geography, 1997, č. 3, s. 203 – 225.
[46] TOWNSEND, A.: Life in the real-time city: mobile telephones and urban metabolism. In. Journal of Urban Technology*,* 2000, s. 85 – 104.
[47] URAL, S. – HUSSAIN, E. – SHAN, J.: Building population mapping with aerial imagery and GISdata. In: International Journal of Applied Earth Observation and Geoinformation, 2011, č. 6, s. 841 – 852.
[48] WANG, Z. – HE, S. Y. – LEUNG, Y.: Applying mobile phone data to travel behaviour research: A literature review. In. Travel Behaviour and Society, 2018, s. 141 – 155.
[49] WESOLOWSKI, A. – EAGLE, N. – TATEM, A. J. – SMITH, D. L. – NOOR, A. M. – SNOW, R. W. – BUCKEE, C. O.: Quantifying the impact of human mobility on malaria. In. Science, 2012, č. 6112, s. 267 – 270.
[50] YANG, Y. – XIONG, C. – ZHUO, J. – CAI, M.: Detecting home and work locations from mobile phone cellular signaling data. In: Mobile Information Systems, 2021, s. 1 – 13.
[51] YOO, B. – KANG, S. – CHON,K. – KIM, S.: Origin–destination estimation using cellular phone BS information. In: Journal of the Eastern Asia Society for Transportation Studies, 2005, s. 2574 – 2588.

RESUMÉ

Súčasné trendy, ktoré stimuluje postmoderná spoločnosť a informačno-komunikačné technológie, kladú nové požiadavky na získavanie aktuálnych a detailných údajov o časovo-priestorovom správaní populácie. Príspevok sa zaoberá možnosťami využitia lokalizačných údajov z mobilnej siete na poznanie priestorovej distribúcie a mobility obyvateľov. Projekt Socioekonomické aspekty Big data Štatistického úradu SR reflektuje túto potrebu a sústreďuje sa na využitie údajov z mobilnej siete v experimentálnej populačnej štatistike. Kľúčovou otázkou je, či údaje z mobilnej siete môžu poskytnúť relevantnú informáciu o priestorovom rozmiestnení a mobilite populácie na Slovensku. Okrem toho sa skúma aký je prínos a limity údajov z mobilných sietí v porovnaní s bežnými štatistickými údajmi o obyvateľstve.
Príspevok prezentuje originálnu metodiku, ktorá bola použitá na extrahovanie pravidelných denných a nočných lokalít používateľov mobilnej siete (SIM kariet). Výsledky naznačujú vysokú koreláciu medzi údajmi získanými z mobilnej siete a referenčnými údajmi o počte obyvateľov z národného cenzu. Prostredníctvom údajov z mobilnej siete tak dostávame relevantnú informáciu o priestorovom rozmiestnení a mobilite populácie, hoci pri interpretácii výsledkov je dôležité si uvedomiť viaceré limity, ktoré vyplývajú z princípov prevádzky mobilnej siete a z právnych podmienok spracovania údajov jej používateľov. Vo všeobecnosti ide najmä o špecifiká vyplývajúce z princípov mobilnej komunikácie, priestorovo diferencovaných trhových podielov mobilných operátorov a nedostatočného pokrytia niektorých sociálno-demografických štruktúr populácie. V praxi ide napríklad o marginalizované rómske komunity. Napriek početným limitom priniesli výsledky spracovania lokalizačných údajov mobilnej siete pomerne spoľahlivé výsledky, ktoré môžu prispieť k hlbšiemu poznaniu priestorového rozmiestnenia a dynamiky obyvateľstva. Pri rešpektovaní limitov tak dostávame perspektívne dáta, vhodné na rozmanité analýzy na úrovni regiónov alebo obcí.

RESUME

Current trends stimulated by postmodern society and information and communication technologies impose new requirements for obtaining up-to-date and detailed data on the spatiotemporal behavior of the population. The contribution focuses on the possibilities of utilizing location data from the mobile network to understand the spatial distribution and mobility of the population. The project Socioeconomic Aspects of Big Data by the Statistical Office of the Slovak Republic reflects this need and concentrates on the use of mobile network data in experimental population statistics. The key question is whether mobile network data can provide meaningful information about the spatial distribution and mobility of the population in Slovakia. Additionally, we aim to explore the benefits and limits of mobile network data compared to conventional population statistics.

The contribution presents an original methodology used to extract regular day and night locations of mobile phone users (SIM cards). The results suggest a high correlation between the data obtained from the mobile network and reference data on population from the national census. Through mobile network data, we thus obtain meaningful information about the spatial distribution and mobility of the population. However, when interpreting the results, it is important to be aware of several limitations that resulted from the principles of mobile network operation and the legal conditions for processing user data. In general, these limitations stem from the specifics of mobile communication principles, spatially differentiated market shares of operators, and underrepresentation of certain socio-demographic population structures. In practice, this includes, for example, the marginalized Roma communities. Despite numerous limitations, the results of processing mobile network location data have yielded relatively powerful results that can contribute to a deeper understanding of the spatial distribution and dynamics of the population. By respecting these limitations, we obtain prospective data suitable for diverse analyses at the regional or municipal level.

PROFESIJNÝ ŽIVOTOPIS

***Mgr. Martin Šveda, PhD.**, absolvoval magisterské štúdium v odbore geografia a kartografia (2007) a doktorandské štúdium v odbore regionálna geografia na Prírodovedeckej fakulte Univerzity Komenského v Bratislave (2011). Od roku 2017 pôsobí ako odborný asistent na Katedre regionálnej geografie a rozvoja regiónov Prírodovedeckej fakulty Univerzity Komenského v Bratislave. Súčasne je samostatným vedeckým pracovníkom na Geografickom ústave SAV. Vo svojej výskumnej činnosti sa zameriava predovšetkým na procesy suburbanizácie a ich dopady na transformáciu prímestských sídiel. Venuje sa aj sledovaniu časovo-priestorových vzorov správania obyvateľov prostredníctvom lokalizačných údajov mobilnej siete.*

***Mgr. Michala Sládeková Madajová, PhD.**, absolvovala magisterské štúdium v odbore geografia a kartografia (2006) a doktorandské štúdium v odbore regionálna geografia na Prírodovedeckej fakulte Univerzity Komenského v Bratislave (2010). V súčasnosti pôsobí ako samostatná vedecká pracovníčka v oddelení humánnej a regionálnej geografie Geografického ústavu SAV a od roku 2019 vyučuje na Katedre regionálnej geografie a rozvoja regiónov Prírodovedeckej fakulty Univerzity Komenského v Bratislave. Venuje sa rôznym témam z oblasti humánnej a regionálnej geografie, ale predovšetkým problematike harmonizácie geografických dát a možnostiam využitia štatistických metód v geografii.*

***Mgr. Pavol Hurbánek, PhD.**, absolvoval magisterské štúdium v odbore geografia a kartografia (2000) a doktorandské štúdium v odbore humánna geografia na Prírodovedeckej fakulte Univerzity Komenského v Bratislave (2007). Pôsobil ako odborný asistent na Katedre humánnej geografie a demogeografie na Prírodovedeckej fakulte Univerzity Komenského v Bratislave (2005 až 2008) a na Katedre geografie na Pedagogickej fakulte Katolíckej univerzity v Ružomberku (2011 až 2022). Od roku 2016 je vedeckým pracovníkom*

Geografického ústavu SAV v Bratislave. Venuje sa výskumu tematickej presnosti mapovania zastavaného územia na odhad počtu obyvateľov, vymedzeniu mestských a vidieckych oblastí s využitím sídelných sietí či regionálnej taxonómii.
***Mgr. Konštantín Rosina, PhD.**, absolvoval magisterské štúdium na Prírodovedeckej fakulte Univerzity Komenského v Bratislave v odbore geografia (2010). V roku 2015 ukončil doktorandské štúdium v oddelení geoinformatiky Geografického ústavu SAV. V rokoch 2016 – 2018 pôsobil ako výskumný pracovník Spoločného výskumného centra Európskej Komisie v Ispre (IT). Od roku 2019 pracuje ako špecialista na diaľkový prieskum Zeme v spoločnosti Solargis s. r. o. Od roku 2020 je aj výskumným pracovníkom Geografického ústavu SAV.*

KONTAKT

martin.sveda@uniba.sk
geogmada@savba.sk
pavolhurbanek@gmail.com
konstantin.rosina@savba.sk

Informatívny článok/Informative article

Juraj BÁRDY
Alistiq, s. r. o.

PROJEKT SOCIOEKONOMICKÉ ASPEKTY BIG DATA

SOCIOECONOMIC ASPECTS OF BIG DATA PROJECT

ABSTRAKT
Článok opisuje projekt Socioekonomické aspekty Big Data a jeho 3 podprojekty, ktoré sa realizovali ako štatistické experimenty v Štatistickom úrade Slovenskej republiky. Jednotlivé podprojekty sú opísané z hľadiska ich cieľa, vstupných údajov, výsledku a možného prínosu pre produkciu. Podprojekty sa zaoberali rýchlym odhadom hrubého domáceho produktu, odhadom priestorového rozmiestnenia a mobility obyvateľstva s využitím lokalizačných údajov mobilnej siete, monitorovaním sociálneho napätia z príspevkov na sociálnej sieti Facebook.

ABSTRACT
The article describes the project Socioeconomic Aspects of Big Data in Statistics and its 3 subprojects, which were implemented as statistical experiments at the Statistical Office of the Slovak Republic. Individual sub-projects are described in terms of their goal, input data, outcome and the possible contribution to production. The sub-projects dealt with flash estimation of the gross domestic product, estimation of the spatial distribution and population mobility using the mobile phone network localization data, monitoring of social tension from posts on the social networking website Facebook.

KĽÚČOVÉ SLOVÁ
Big Data, rýchly odhad HDP, lokalizačné údaje mobilnej siete, analýza sentimentu, socioekonomická štatistika

KEY WORDS
Big Data, flash estimates of GDP, mobile phone network localization data, sentiment analysis, socioeconomic statistics

1. ÚVOD
Tvorcovia oficiálnych štatistík sa už tradične spoliehajú na vlastný zber údajov, využívajúc elektronický zber údajov, osobné a telefonické rozhovory alebo v posledných rokoch aj spôsob zberu údajov dostupných online. Nové zdroje údajov otvárajú nové možnosti modernizácie oficiálnej štatistiky vďaka využitiu metód na spracovanie veľkého množstva údajov, ktoré vznikajú ako vedľajší produkt v rámci digitalizovanej spoločnosti a ekonomiky. V posledných rokoch sa objavujú aj zahraničné skúsenosti vytvárania nových štatistických produktov založených na spracúvaní Big Data.

Hoci koncept Big Data je ťažké presne definovať, jeho fundamentálne charakteristiky sú pomerne ľahko rozpoznateľné a odlíšiteľné od tradičných zdrojov údajov. Big Data sú zvyčajne definované z hľadiska 3V: objem (angl. volume), rozmanitosť (angl. variety) a rýchlosť (angl. velocity). Veľké údaje nie sú jednoducho

definované objemom, ide aj o ich zložitosť. Mnohé malé súbory údajov, ktoré sa považujú za Big Data, nezaberajú veľa fyzického priestoru, ale sú svojou povahou obzvlášť zložité. Zároveň veľké súbory údajov, ktoré vyžadujú značný fyzický priestor, nemusia byť dostatočne zložité na to, aby sa mohli považovať za Big Data. Rozmanitosť odkazuje na rôzne typy štruktúrovaných a neštruktúrovaných údajov, ako sú údaje na úrovni transakcií, videa a zvuku, alebo môže ísť aj o textové a protokolové súbory. Rýchlosť je údaj o tom, ako rýchlo údaje vznikajú, prípadne sa upravujú [2].

Štatistická komunita oficiálne uznala potenciál veľkých dát, keď sa v EÚ v priebehu roka 2013 viedli diskusie na globálnej úrovni o identifikácii možností, ktoré Big Data prinášajú oficiálnej štatistike, a zároveň o hlavných strategických a metodických problémoch, ktoré Big Data predstavujú pre oficiálnu štatistiku. Záverom týchto debát bolo Scheveningenské memorandum. V marci 2014 Štatistická komisia OSN zriadila globálnu pracovnú skupinu (UN Global Working Group on Big Data for Official Statistics) s cieľom : „*poskytnúť strategickú víziu, smerovanie a globálny program o veľkých údajoch pre oficiálnu štatistiku, podporiť praktické využitie zdrojov Big Data pre oficiálnu štatistiku pri hľadaní riešení ich výziev a podporiť budovanie kapacít a vymieňanie skúseností v tejto oblasti*“. Použitie Big Data umožňuje generovanie štatistických produktov v reálnom čase, zatiaľ čo oficiálne štatistiky prinášajú hĺbku detailov a reprezentáciu prostredníctvom overených štatistických zisťovaní. Najlepšie výsledky môže priniesť spojenie týchto dvoch prístupov. Odhalenie prínosov však vôbec nie je jednoduché. Momentálne sa predpokladá, že využitie Big Data nedokáže nahradiť štandardné metódy a oficiálnu štatistiku, ale môže byť prínosným doplnkom.

Používanie Big Data v štatistike mení kontext a organizačné zabezpečenie štatistickej produkcie. Je potrebná zvýšená spolupráca medzi Štatistickým úradom SR, organizáciami, ktoré generujú Big Data (zdroje) a akademickým sektorom. V budúcnosti by to mohlo znamenať posun úlohy Štatistického úradu SR, pokiaľ ide o poskytovanie vysoko kvalitných štatistických informácií v reálnom čase. Pri navrhovaní modelov budúcej spolupráce je dôležité sústrediť sa na optimálne využitie silných stránok zainteresovaných strán. Medzi tradičné silné stránky Štatistického úradu SR patrí na jednej strane schopnosť zbierať údaje a kombinovať zdroje údajov a na druhej strane jeho zameranie na kvalitu, transparentnosť a spoľahlivú metodiku. Štatistický úrad SR má tiež jedinečné znalosti o oficiálnych metódach tvorby štatistiky. Zároveň je ho možné považovať za nestrannú a apolitickú tretiu stranu.

2. PROJEKT SOCIOEKONOMICKÉ ASPEKTY BIG DATA

Projekt Socioekonomické aspekty Big Data (SEABD) bol projekt Štatistického úradu Slovenskej republiky (Štatistický úrad) s cieľom preskúmať nové zdroje údajov, metódy spracovania Big Data a ich použiteľnosť pre socioekonomickú štatistiku. Projekt sa realizoval v rokoch 2022 až 2023 (1.4.2022 – 29.9.2023) ako súčasť výziev operačného programu Integrovaná infraštruktúra (OPII) Európskych štrukturálnych a investičných fondov (EŠIF). Projekt sa osobitne zameriaval na analýzu dostupnosti a identifikáciu požiadaviek na zdrojové údaje, vytvorenie infraštruktúry na ukladanie a spracúvanie Big Data a na vytvorenie metodických materiálov na spracovanie a overenie kvality výstupov vrátane čiastkových výstupov, monitorovanie výkonnosti modelov hlavne z pohľadu presnosti a správnosti a publikovanie výsledkov. Dôležitým aspektom počas riadenia všetkých etáp projektu bolo zabezpečenie etického, bezpečného a dôveryhodného spracovania všetkých

údajov, vrátane osobných údajov. Do projektu SEABD boli zapojení zamestnanci Štatistického úradu, ktorí získali skúsenosti s prácou vo vybudovanom prostredí SAS Viya a Amazon Web Services S3, s prácou v jazyku Python, trénovaním a anotáciou modelov strojového učenia. Praktická časť implementácie projektu bola rozdelená do podprojektov podľa troch vybraných oblastí:

- rýchle odhady hrubého domáceho produktu podľa intenzity nákladnej dopravy,
- odhad priestorového rozmiestnenia a mobility obyvateľstva s využitím lokalizačných údajov mobilnej siete,
- monitorovanie sociálneho napätia z príspevkov na sociálnej sieti Facebook.

Modely v skúmaných oblastiach boli vytvorené na základe analýzy zahraničných skúseností a prípadov použitia Big Data pre potreby štatistiky krajín EÚ.

2.1. Rýchle odhady hrubého domáceho produktu

Cieľom podprojektu *Rýchle odhady hrubého domáceho produktu* bolo overenie štatistického vzťahu medzi indexom počítaným z údajov mýtneho systému a štatistikou vývoja hrubého domáceho produktu (HDP) Slovenska. Cieľom výstupného modelu bolo poskytnúť rýchly odhad aktuálneho ekonomického vývoja na Slovensku vyjadreného pomocou kľúčového ukazovateľa – HDP. Vstupné údaje tvoril súbor údajov získaný zberom údajov o presnej polohe registrovaných nákladných vozidiel v mýtnom systéme, prevádzkovanom Národnou diaľničnou spoločnosťou. Transakcie v mýtnom systéme sa ukladajú v reálnom čase a zahŕňajú nákladné vozidlá nad 3,5 tony, ktoré sa pohybujú po platených úsekoch diaľnic, rýchlostných ciest a ciest I. triedy. Tieto údaje sú dostupné v priebehu 10 dní nasledujúceho mesiaca, v ktorom sa zhromažďujú. Vytvorený model autoregresívnej analýzy časových radov bol navrhnutý tak, aby zohľadňoval sezónne vzorce a exogénne faktory, čím sa zabezpečila spoľahlivá metóda, ktorá umožňuje odhadnúť hodnotu štvrťročného HDP na Slovensku do 10. dňa nasledujúceho mesiaca, čo je oveľa skorší odhad ako štandardné štvrťročné oneskorenie. Z toho vyplýva, že nekonvenčné zdroje údajov majú veľký potenciál na využitie pri prognóze ekonomických indikátorov a tým poskytnúť komplexnejší a aktuálnejší pohľad na hospodársku situáciu, čo je veľmi užitočné pri rozhodovacích procesoch.

Model ponúka vysokú nákladovú efektívnosť, pretože spracovanie surových dát môže prebiehať v infraštruktúre Národnej diaľničnej spoločnosti (NDS) a údaje sa primárne zbierajú na efektívny výber elektronického mýta. Je však dôležité sledovať vplyv zmien cestnej siete a správania nákladných vozidiel na spoľahlivosť prognóz. Taktiež kvalita údajov závisí od infraštruktúry elektronického mýtneho systému NDS, pričom momentálne minimálna hodnota efektívnosti výberu mýta je stanovená na 98,91%. Je preto nevyhnutné sledovať, či aj prípadný nový dodávateľ bude plniť rovnakú efektívnosť a s ňou spojenú presnosť a včasnosť zberu údajov. Ďalšou silnou stránkou modelu je, že produkcia vykazuje nízku náročnosť na personálne kapacity úradu, čo zjednodušuje nasadenie a prináša širokú škálu príležitostí na rozšírenie modelu, napríklad o údaje z ciest ostatných tried. Údaje o najazdených kilometroch nákladných vozidiel navyše ponúkajú nové možnosti na prognózovanie aj iných indikátorov, ako je index priemyselnej výroby.

Jedným z možných úskalí je obmedzený prístup k údajom, ktorý môže byť ovplyvnený rozhodnutím dodávateľov údajov. Toto môže viesť k neschopnosti úradu získať potrebné informácie na realizáciu projektu. Štatistický úrad navyše získava

údaje od NDS na základe zmluvy, ktorej platnosť sa končí. Toto predstavuje potenciálne riziko, pretože udržateľnosť výstupov projektu závisí od ochoty NDS uzatvoriť novú dohodu o poskytovaní údajov. Ak sa nedosiahne dohoda, môže to mať negatívny vplyv na celkový úspech, udržateľnosť projektu a možné využitie v praxi.

Model má ambíciu slúžiť pre tvorcov politík k prijímaniu informovanejších rozhodnutí, pretože dokáže poskytnúť včasnú indikáciu vývoja ekonomickej aktivity a v budúcnosti by mohol umožniť identifikovať nové trendy a vzorce v rôznych sektoroch alebo časových obdobiach. Tieto informácie by mohli byť užitočné pre fiškálne plánovanie, alokáciu zdrojov a tvorbu postupov a stratégií. Skoré informácie o ekonomických trendoch navyše predstavujú cenný zdroj pre aktivity výskumníkov, akademickej obce a občianskej spoločnosti, ktorá by získala lepšie prostriedky na pochopenie celkovej ekonomickej situácie a zvýšilo by sa tak všeobecné povedomie o týchto otázkach.

2.2. Odhad priestorového rozmiestnenia a mobility obyvateľstva s využitím lokalizačných údajov mobilnej siete

Podprojekt *Odhad priestorového rozmiestnenia a mobility obyvateľstva s využitím lokalizačných údajov mobilnej siete* vychádzal z potreby preskúmať hodnotu údajov o polohe SIM kariet zákazníkov mobilných operátorov pre potreby demografických štatistík a štatistík o pohybe obyvateľov. Operátori mobilných sietí tieto údaje vytvárajú pre potrebu priradiť mobilný telefón k bázovej stanici a údaje sú tak vedľajším produktom ich hlavnej funkcie prevádzky mobilnej siete. Vstupné údaje do modelovania tvorili tzv. signalizačné údaje agregované do matice dochádzkových tokov medzi dvoma sledovanými územnými jednotkami, z ktorých nie je možné identifikovať jednotlivca, keďže tieto údaje so sebou neprenášajú identifikátory držiteľov SIM kariet a v datasetoch o dochádzkových tokoch boli toky menšie ako 3 nahradené fixnou hodnotou. Signalizačné údaje chápeme ako automaticky generované záznamy, ktoré produkuje mobilná sieť pri pravidelných kontrolách pripojených zariadení. Tiež ich možno opísať ako údaje obsahujúce prakticky všetky udalosti, ktoré zahŕňajú komplexnú komunikáciu medzi zariadením a sieťou, a preto sa odporúčajú na štatistické spracovanie [1]. Cieľom podprojektu bolo vytvorenie robustného a škálovateľného modelu, ktorý je možné dopĺňať údajmi zo sčítania obyvateľov, domov a bytov a prípadne z iných administratívnych zdrojov. Výsledný model preukázal schopnosť spoľahlivo odhadovať dennú populáciu a hlavné vzorce mobility obyvateľstva, čo umožňuje sledovať vzory priestorového rozmiestnenia obyvateľstva a vytvárať funkčné regióny. Dennú populáciu definujeme ako počet ľudí prítomných v nejakom určenom území počas denného času. Tento údaj zohľadňuje nielen trvalo registrovaných obyvateľov, ale aj osoby, ktoré do daného územia dochádzajú za prácou, školou alebo inými dennými aktivitami. Závery podprojektu potvrdili, že štatistiky získané z projektu sa môžu široko využiť v oficiálnych štatistikách vrátane sledovania cestovného ruchu a dennej populácii.

V rámci podprojektu vznikol aj efektívny komunikačný kanál s dodávateľmi údajov, čo zabezpečuje rýchle a efektívne získavanie vstupných údajov. Obmedzenia týchto údajov súvisia najmä s nastavenou ochranou súkromia používateľov mobilnej siete, so zastúpením vekových skupín a závislosťou kvality údajov od infraštruktúry mobilných sietí, ktorá sa môže líšiť medzi dodávateľmi. Tieto náklady je však možné minimalizovať legislatívnym ukotvením poskytovania údajov o polohe SIM kariet pre potreby štatistiky a vytvorením stáleho interného tímu na prácu s týmto druhom údajov.

2.3. Monitorovanie sociálneho napätia z príspevkov na sociálnej sieti Facebook

Základnou ideou podprojektu *Monitorovanie sociálneho napätia z príspevkov na sociálnej sieti Facebook* bolo sledovanie spoločnosti prostredníctvom štatistického spracovania údajov zo sociálnych sietí. Ako skúmaný koncept sa určilo napätie v spoločnosti. Na tento účel sa využili údaje z najvýznamnejšej sociálnej siete, kde sa diskutuje o politických, sociálnych a iných témach, hoci je dôležité mať na pamäti, že správanie používateľov na sociálnych sieťach sa dynamicky vyvíja a líši sa medzi generáciami. Vytvorený model preukázal schopnosť kvantifikovať sentiment verejných príspevkov na rôzne témy a identifikovať sociálne napätie. Úspech modelu predovšetkým súvisí s použitím popredného jazykového modelu XLM-RoBERTa Large, založeného na hĺbkovom strojovom učení cez neurónové siete. Tento model poskytuje základ na ďalšiu prácu, pričom ho možno ho relatívne jednoducho dotrénovať na ďalšie klasifikácie a úlohy. Okrem toho využitie údajov zo sociálnych sietí umožňuje prístup k dostatočne veľkej vzorke spoločnosti.

Komplexný prístup umožňuje celistvú analýzu a poskytuje ucelený obraz o aktuálnych náladách a postojoch v spoločnosti a prináša viaceré príležitosti na využitie. Model môže byť napríklad prospešný pri riešení úloh v štátnej správe, pretože dokáže identifikovať verejnú mienku, čo môže pomôcť pri rozhodovaní a riešení rôznych problémov. Okrem toho, model poskytuje možnosť identifikovať včasné varovné signály potenciálnych konfliktov alebo spoločenských problémov, čo môže prispieť k prevencii a lepšiemu riadeniu situácií.

Výsledný analytický model má aj svoje slabé stránky, medzi ktoré patrí predovšetkým náročnosť analýzy na výpočtovú infraštruktúru. Dodávateľ, ktorý zabezpečuje zber údajov zo sociálnych sietí, nesie kľúčovú zodpovednosť za kvalitu vstupných údajov. S tým však prichádza potenciálne riziko nespoľahlivosti údajov, čo môže spôsobiť nepresnosť výsledkov. Skreslenie výsledkov vzniká aj z nejasnej reprezentácie vzorky (používatelia vybranej sociálnej siete a ich interakcie) a chýbajúcich demografických informácií, vrátane nedostatočného zohľadnenia zastúpenia vekových skupín. Paralelne s týmto problémom sa stretávame s nevysvetliteľnosťou modelu a jeho rozhodnutí pri klasifikácii, čo môže pôsobiť nejasne pre niektorých používateľov výsledkov.

3. ZÁVER

Jedným z trendov produkcie národných štatistických úradov aj v Európskom štatistickom systéme je používanie nových zdrojov údajov, ktoré predstavujú niekedy extrémne veľké objemy dát s vysokou frekvenciou, ako napríklad údaje získané zo senzorov alebo údaje o polohe SIM kariet od mobilných operátorov. Práve pomocou takýchto dát, ktoré označujeme ako Big Data je možné efektívne merať a opisovať vývoj v spoločnosti prostredníctvom štatistík v reálnom čase. Projekt preukázal potenciál a schopnosť Big Data obohatiť doterajšie poznanie a poskytnúť pridanú hodnotu v štatistike. V rámci projektu bol tiež overený experimentálny priestor AWS S3 a výsledné modely boli nasadené na novovybudovanú infraštruktúru SAS Viya, ktorú možno použiť na ďalšie skúmanie a produkciu. Nemenej dôležitým výsledkom SEABD sú analytické a metodické materiály pre prácu s Big Data pomocou nástrojov dátovej vedy, strojového učenia a umelej inteligencie. Práca s Big Data je meniaca sa oblasť poskytujúca nové zdroje dát, nové možnosti v oblasti hardvérovej a softvérovej infraštruktúry, nové oblasti využitia a potrebu výmeny skúseností doma aj v zahraničí.

LITERATÚRA

[1] ESKO, S.: What is mobile phone data? Overview of data generated by mobile communication technologies. Positium, 2019.
[2] KITCHIN, R.: Big Data and Human Geography: Opportunities, challenges and risks. In: Dialogues in Human Geography, 2013, č. 3, s. 262 – 267.

RESUMÉ

Bez údajov by štatistika nebola možná, a preto sú inovácie v metódach, technikách a prístupoch rozvíjajúcich prácu s údajmi dôležité pre napredovanie slovenskej štatistiky. Projekty podobné tým, aké sú opísané v texte, sú nevyhnutné na overenie možnosti využitia alternatívnych zdrojov údajov pre oficiálnu štatistiku. Avšak cesta od identifikácie nového zdroja údajov po jeho zavedenie do produkcie a diseminácie štatistík je dlhá a náročná. Preto je potrebné mať štatistické projekty, ktoré pomôžu zefektívniť, zmodernizovať a zlepšiť štatistické procesy. Je dôležité hľadať zdroje, ktoré podporia tieto iniciatívy a prinesú dlhodobý rozvoj do oblasti štatistiky. Príspevok opisuje projekt Socioekonomické aspekty Big Data a jeho 3 podprojekty, ktoré sa realizovali v Štatistickom úrade Slovenskej republiky

RESUME

Statistics is not possible without data and therefore the innovations in methods, techniques and approaches for developing working with data are important for the advancement of Slovak statistics. Projects similar to those described in the text are necessary to verify the possibility of using alternative data sources for official statistics. However, the road from the identification of a new data source towards its introduction into the production and dissemination of statistics is long and arduous. Therefore, there is a need for statistical projects that will help to streamline, modernize and improve statistical processes. It is important to look for resources that will support these initiatives and bring long-term development to the field of statistics. The paper describes the project Socioeconomic Aspects of Big Data in Statistics and its 3 subprojects, which were implemented at the Statistical Office of the Slovak Republic.

PROFESIJNÝ ŽIVOTOPIS

***Ing. Juraj Bárdy** absolvoval magisterské štúdium na Fakulte riadenia a informatiky Žilinskej univerzity v Žiline v odbore informačné a riadiace systémy (2006). Venuje sa inováciám verejných služieb a politík a digitálnej transformácii vo verejnej správe, so zameraním na využitie strojového učenia a lepší manažment údajov. Podieľal sa na návrhu Národnej koncepcie verejnej správy (2016) a príprave Stratégie digitálnej transformácie SR (2019). Je partnerom v konzultačnej spoločnosti Alistiq s .r. o.*

KONTAKT

juraj.bardy@alistiq.com

Informatívny článok/Informative article

Peter ĎURIŠ
Go SMART, s. r. o.

METODICKÝ RÁMEC NA PODPORU POUŽÍVANIA BIG DATA V ŠTATISTIKE

METHODOLOGICAL FRAMEWORK TO SUPPORT THE USE OF BIG DATA IN STATISTICS

ABSTRAKT
Článok sa zaoberá definovaním vhodných postupov pre adopciu Big Data[1] do procesu tvorby štatistických produktov. Realizáciou projektu Socioekonomické aspekty Big Data Štatistický úrad Slovenskej republiky demonštruje, že existujú modely a postupy ako využiť Big Data, ktoré je možné štandardizovať do metodických usmernení pre túto oblasť. Cieľom príspevku je predstavenie metodologického rámca na štandardizáciu procesu zberu, spracovania a vyhodnocovania Big Data.

ABSTRACT
The text deals with the definition of appropriate procedures for the adoption of Big Data in the process of creating statistical products. By implementing a project Socio-economic aspects of Big Data, the Statistical Office of Slovak Republic demonstrates that there are models and procedures for using Big Data and that these can be standardized into methodological guidelines for this field. The aim of the presentation of the methodology framework for the standardization of the process of collecting, processing and evaluating Big Data.

KĽÚČOVÉ SLOVÁ
veľké údaje, experimentálna štatistika, životný cyklus analytického modelu, GSBPM, získavanie Big Data, spracovanie Big Data, scenáre použitia Big Data

KEY WORDS
Big Data, experimental statistics, analytical model lifecycle, GSBPM, Big Data mining, processing of Big Data, Big Data usage scenarios

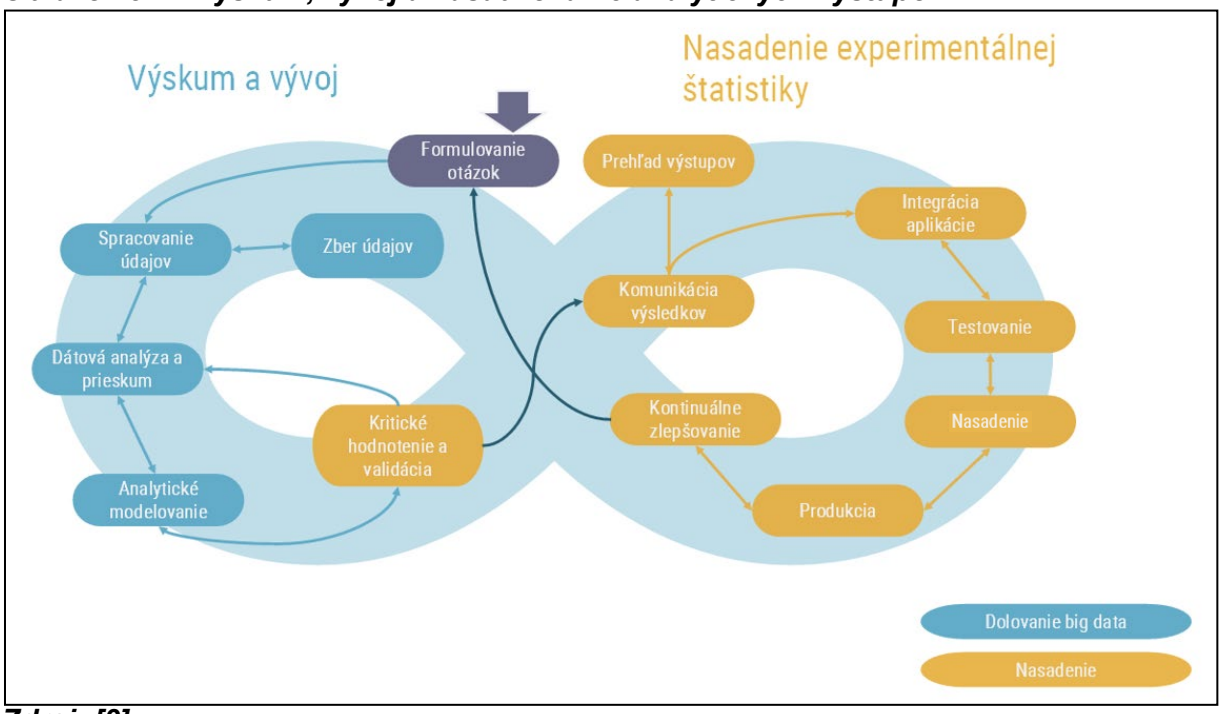
1. ÚVOD
Všeobecná definícia termínu Big Data od Tama a Clarkea [5] charakterizuje Big Data ako zdroje štatistických údajov zahŕňajúce tradičné zdroje aj nové zdroje, ktoré sa stávajú dostupnými z „webu všetkého“[2]. Európska komisia [2] definovala Big Data ako „*veľké množstvá údajov vyprodukovaných veľmi rýchlo veľkým počtom rôznych zdrojov*“. D. Laney [3] poskytol definíciu pomocou 3V: „*Údaje o veľkom objeme, vo veľkej rýchlosti a z rôznorodých zdrojov, ktoré si vyžadujú nákladovo efektívne, inovatívne formy spracovania informácií, ktoré uľahčujú lepší prehľad, rozhodovanie a automatizáciu procesov*“.

Využitie Big Data pre štatistické úrady si vyžaduje neustálu inováciu a zvládnutie nových postupov. Pracovné postupy v tejto oblasti ešte nie sú jednotné a môžu sa líšiť

[1] Definícia Big Data sa vzťahuje na veľké, variabilné dáta, ktorých spracovanie a analýza sú mimoriadne cenné, pretože vďaka týmto procesom sa získavajú nové, veľmi cenné informácie.
[2] Z anglického pojmu Web of everything.

medzi jednotlivými úradmi, preto bolo potrebné v pilotnom projekte určiť jasné metodické prístupy k celému životnému cyklu: od ich získavania, spracovania a následného používania. Štatistický úrad Slovenskej republiky so zámerom preskúmať možnosti práce s Big Data realizoval projekt Socioekonomické aspekty Big Data (SEABD), ktorého hlavným cieľom bolo overiť možnosti využívania Big Data pri tvorbe inovatívnych štatistických produktov. V rámci tohto projektu vznikol upravený procesný model, inšpirovaný všeobecným štatistickým modelom obchodného procesu podľa Loisona a Kuonena [4]. Pri určení metodologického rámca bolo nevyhnutné zohľadniť inovačný charakter hĺbkovej analýzy Big Data a prispôsobiť postupy konkrétnym modelom a dátovým zdrojom. Na nasledujúcej schéme (obrázok č. 1) je znázornený prístup k využívaniu Big Data v oblasti štatistiky.

Obrázok č. 1: Výskum, vývoj a nasadzovanie analytických výstupov



Zdroj: [3]

Fáza výskumu a vývoja je neoddeliteľnou súčasťou životného cyklu analytických modelov. V tejto fáze je možné využiť hĺbkovú analýzu Big Data a dodržať overené procesy dátovej vedy. Pri tomto prístupe je dôležité kombinovať induktívne a deduktívne uvažovanie a prispôsobiť ich potrebám konkrétneho modelu. V konečnom dôsledku je potrebné, aby pracovné postupy pre štatistické úrady využívajúce Big Data boli štandardizované a zohľadňovali špecifiká jednotlivých modelov a dátových zdrojov. Kľúčom k úspechu je aj neustále sledovanie a adaptácia nových technológií a postupov v tejto oblasti.

2. POSTUP VYUŽIVANIA BIG DATA PROSTREDNÍCTVOM ŽIVOTNÉHO CYKLU ANALYTICKÝCH MODELOV

Životný cyklus analytických modelov pozostáva z nasledujúcich fáz:

1. posúdenie scenára – od idey k projektovému zámeru,
2. výskum a vývoj,
3. vyhodnotenie užitočnosti,
4. nasadenie a prevádzka analytického modelu.

2.1. Posúdenie scenára

Cesta využívania Big Data sa môže začať rôznymi spôsobmi. V niektorých prípadoch projekt iniciujú zvedaví a angažovaní jednotlivci, ktorí chcú zlepšiť súčasný stav. Môže sa začať pokynmi od vyššieho manažmentu alebo ako čistý výskumný projekt bez konkrétneho plánu na uvedenie riešenia do produkcie. Bez ohľadu na dôvod vzniku iniciatívy je potrebné správne pochopiť potreby organizácie a navrhnúť projektový zámer. Táto etapa spočíva v posúdení možného scenára. Big Data a súvisiace techniky hĺbkovej analýzy údajov možno vo všeobecnosti použiť v dvoch oblastiach relevantných pre oficiálne štatistiky:

- na automatizáciu čiastkových úloh pri tvorbe oficiálnych štatistík,
- na tvorbu nových oficiálnych štatistík s využitím nových zdrojov Big Data.

Projekty a idey v oblasti Big Data by mali mať svoju jasnú štruktúru a určené prínosy, dosah a prípady použitia v reálnom čase. Preto je potrebné pri koncipovaní projektového zámeru nájsť odpovede na základné otázky, ktoré obsahujú najdôležitejšie prvky budúceho projektu z rôznych uhlov pohľadu. Ide o minimálny súbor otázok, na ktoré by mal vlastník procesov a kľúčový používateľ[3] odpovedať pred samotným návrhom projektu v oblasti využitia Big Data.

Kontext realizácie zamýšľaného projektu vytvára náročnosť zodpovedania daných otázok. Preto je potrebné venovať dostatočnú pozornosť pri plánovaní projektu a vyhodnotení jeho kontextu, komplexnosti, dosahu, procesnej zložitosti, rizikovosti realizácie. To sa dá dosiahnuť na základe odpovede na súbor otázok, ktoré určujú základné parametre postupu scenára využitia Big Data:

- účel: Prečo je riešenie potrebné a aké výsledky má priniesť?
- využitie: V akých procesoch a okolnostiach je vhodné projekt/riešenia využiť?
- dosah: Aké dôsledky (dobré aj zlé) má realizácia riešenia na spoločnosť?
- predpoklad: Na akých predpokladoch je riešenie postavené a aké sú limity a bariéry použitia?
- údaje: Na akých zdrojoch dát bude riešenie postavené a aké sú limity a bariéry využitia a získavania údajov?
- vstupy: Aké nové údaje sú potrebné na riešenie?
- mitigácia: Aké aktivity musia byť prijaté na zníženie negatívnych dopadov, ktoré vyplývajú z limitov a bariér využitia?
- etika riešenia: Aké hodnotenie etiky využitia riešenia bolo zrealizované (napríklad ochrana osobných údajov) ?
- výhľad: Do akej miery je potrebný ľudský úsudok pred algoritmom a kto je zodpovedný za jeho správne používanie?
- hodnotenie: Ako a na základe akých kritérií kvality bude riešenie hodnotené?

2.2. Výskum a vývoj (proof-of-concept)

Výskumno-vývojová fáza slúži na vývoj analytického modelu v plnom rozsahu, aby sme získali konkrétnu predstavu, či je riešenie daného problému a údajov uskutočniteľné. Analytický model v tejto fáze vývoja sa nazýva aj Proof-of-Concept (PoC). Výskum analytického modelu poskytuje príležitosť na získanie kvantitatívnych výsledkov, ktoré sa použijú na podporu rozhodovania o užitočnosti riešenia, ako aj na objavenie a identifikovanie neočakávaných problémov. Na meranie výkonnosti

[3] *Tí, ktorí zodpovedajú za príslušnú oblasť, v ktorej sa bude metodika BIG DATA aplikovať.*

výskumného dátového modelu sa stanovia podrobné a kvantifikovateľné kritériá kvality, ako je presnosť, včasnosť a nákladová efektívnosť. Výber metriky kvality by mal brať do úvahy potreby a celkový kontext projektu.

V tejto fáze sa aplikujú poznatky dátovej vedy. Dátová veda poskytuje nové metodologické a technologické postupy na analýzu Big Data kombinovaním prístupov z rôznych vedných odborov, ako matematika, štatistika, informatika a v neposlednom rade aj z odboru, z ktorého sú samotné údaje na analýzu zozbierané. Pomocou dátovej vedy hľadáme funkciu mapovania vstupu na výstup.

Obrázok č. 2: Proces dátovej vedy

Zdroj: vlastné spracovanie autora

Proces dátovej vedy znázornený na obrázku č. 2 vychádza z aktívneho výskumu a vývoja – teda z prístupu, pri ktorom sa navrhne a zrealizuje experiment s cieľom získať a zanalyzovať správne údaje na efektívne zodpovedanie danej otázky alebo vyriešenie daného problému. Príprava výskumného analytického modelu pozostáva z nasledujúcich krokov (kroky 2 až 5 predstavujú proces hĺbkovej analýzy veľkých objemov údajov – Big Data):

1. **Správne formulovanie otázok alebo problému**: Čo a ako sa snažíme zanalyzovať pomocou Big Data? So správne formulovanými otázkami súvisí aj samotný návrh experimentu. Ide o najnáročnejšiu fázu, v ktorej treba nájsť odpovede na tieto a iné otázky: Aký je najlepší spôsob, ako zodpovedať danú otázku alebo vyriešiť daný problém? Aké sú tie správne údaje, ktoré nám k tomu dopomôžu? Ako možno tieto údaje zozbierať? Aké nástroje a knižnice sa budú dať použiť na analýzu? Aké stratégie a metódy zvoliť, aby sa dalo predísť prípadným pochybeniam v návrhu experimentu alebo pri zbere údajov?

2. **Zber údajov**, ktorý sa realizuje až po získaní odpovedí na uvedené otázky. Treba mať teda na pamäti, že **nesprávne údaje vedú k nesprávnej analýze a tá vedie k zlým rozhodnutiam**. Zabezpečenie kvalitných a dostatočných údajov je preto kľúčovým krokom. Je potrebné jasne špecifikovať požiadavky na údaje a dohodnúť podmienky ich zabezpečenia na účely výskumu a následnej produkcie. Je málo pravdepodobné, že by potreby oficiálnej štatistiky naštartovali celý nový proces zberu Big Data, preto sa predpokladá, že sa bude využívať existujúci zdroj. Súbor údajov v štádiu výskumu nemusí byť skutočným súborom údajov, ale aj syntetickými údajmi, verejne dostupnými údajmi alebo malou podmnožinou skutočných údajov. Po fáze zberu sa údaje často ukladajú do databázy alebo všeobecnejšie do dátového skladu na spracovanie.

3. **Spracovanie údajov:** Údaje sa analyzujú, čistia (napríklad detekcia chýb, spracovanie chýbajúcich hodnôt, extrémne hodnoty), vizualizujú a transformujú predtým, ako sú vložené do algoritmov pre spracovanie Big Data ako je napríklad strojové učenie. Tento krok sa opäť stáva súčasťou systematického manažmentu údajov. Prvú fázu spracovania predstavuje predspracovanie údajov. Ide hlavne

o extrakciu vstupných premenných a čistenie údajov, keďže zozbierané údaje spravidla nie sú vo forme, ktorá je vhodná na spracovanie. Aby boli údaje vhodné na spracovanie, je nevyhnutné ich transformovať do formátu, ktorý je vhodný pre algoritmy získavania Big Data. Fáza extrakcie vstupných premenných sa často vykonáva paralelne s čistením údajov, keď sa chýbajúce a chybné časti údajov odhadujú alebo opravujú. V mnohých prípadoch môžu byť údaje extrahované z viacerých zdrojov a je potrebné ich integrovať do jednotného formátu na spracovanie. Konečným výsledkom tohto postupu je štruktúrovaný súbor údajov, ktorý môže počítačový program efektívne využiť. Po fáze predspracovania môžu byť údaje opäť uložené v databáze na spracovanie.

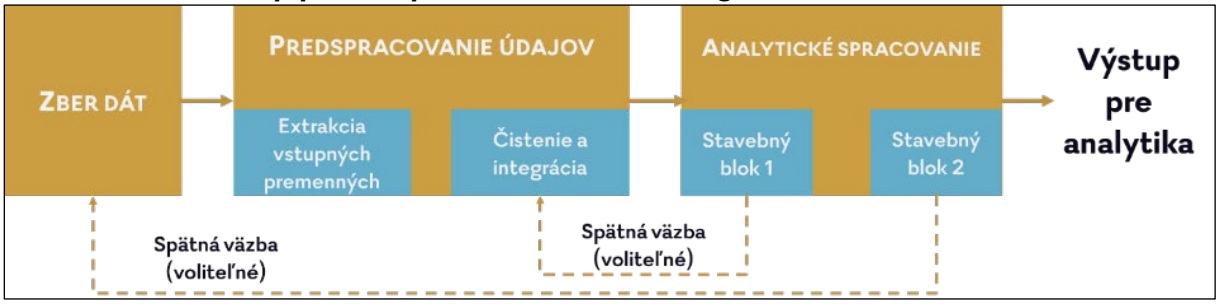
4. **Dátová analýza a prieskum** slúži na overenie predpokladov a odpovedí na otázky z kroku 1, na overenie kvality spracovania údajov z kroku 2 a na hlbšie porozumenie Big Data, aby bolo možné identifikovať vhodné algoritmy hĺbkovej analýzy Big Data na prípravu analytického modelu. V mnohých prípadoch sa nebude dať priamo použiť štandardný algoritmus hĺbkovej analýzy Big Data, no **často je možné rozdeliť analytické spracovanie na stavebné bloky využívajúce štandardné algoritmy**.

5. **Analytické modelovanie:** V tomto kroku ide predovšetkým o trénovanie analytického modelu, keď sa rôzne modely získavania Big Data trénujú na súbore pripravenom v predchádzajúcom kroku. Aby sa predišlo problémom s tzv. overfittingom, údaje sa ideálne rozdelia na trénovaciu, validačnú a testovaciu množinu. V tejto fáze sa používa iba prvá a druhá množina údajov, aby sa v ďalšej fáze mohol model testovať na nezávislom súbore údajov, ktorému nebol vystavený, čiže na tretej testovacej množine údajov.

6. **Kritické zhodnotenie výstupov:** V tejto fáze sa uskutočňuje hlavne testovanie vytvoreného analytického modelu a zhodnotenie jeho výstupov predovšetkým na základe presnosti. Ak výsledky nie sú uspokojivé, je potrebné model ďalej optimalizovať alebo zvoliť iný, v najhoršom prípade je potrebné sa vrátiť aj k ďalším predtým vykonaným krokom. Keď sa dosiahne želaná presnosť odhadu, možno prejsť do ďalšej fázy životného cyklu, ktorou je vyhodnotenie užitočnosti.

7. **Komunikácia výsledkov:** Ide o posledný krok procesu dátovej vedy. Forma komunikácie závisí od fázy, v ktorej sa experiment realizoval.

Obrázok č. 3: Postup práce v procese získavania Big Data



Zdroj: [1]

Kľúčovou súčasťou procesu dátovej vedy je hĺbková analýza Big Data. Celkový proces hĺbkovej analýzy údajov je znázornený na obrázku č. 3 (ide o kroky 2 až 5

z celkového procesu dátovej vedy). Blok analytického spracovania na obrázku č. 3 znázorňuje viacero stavebných blokov predstavujúcich návrh riešenia pre konkrétny prípad použitia. Táto časť algoritmického dizajnu závisí od zručnosti analytika.

2.3. Vyhodnotenie užitočnosti

Podľa rámca na posúdenie kvality štatistík sa vyhodnotí kvalita vstupov a kvalita výstupov výskumného analytického modelu. V tejto fáze je dôležité posúdiť najmä presnosť odhadov a relevantnosť výstupov modelu, ako aj technické obmedzenia a možnosti získavania údajov zo zdroja počas produkcie. V prípade, že bude kvalita vstupov alebo výstupov nevyhovujúca, proces sa vráti na začiatok fázy výskum a vývoj, keďže bude potrebné zmeniť prístup a prehodnotiť vstupné parametre na jednej z úrovní analytického modelu. O výsledkoch výskumu sa vypracuje správa, v ktorej sa popíšu dosiahnuté výsledky.

Posúdenie technických požiadaviek a obmedzení je v tejto fáze kľúčovým prvkom hodnotenia. Je dôležité, aby použité softvérové nástroje (napríklad Python alebo R) boli kompatibilné s produkčným prostredím alebo aby existovala cesta, ako model do produkčného prostredia presunúť. Na záver sa určí, či sa oplatí investovať ďalšie zdroje a či sa výstup môže použiť ako experimentálna štatistika.

2.4. Nasadenie a prevádzka analytického modelu

Nasadenie analytického modelu je procesom integrácie modelu do existujúceho prostredia a procesov tak, aby jeho výsledky boli dostupné používateľom. Model je možné nasadiť vo fáze výskumu a vývoja (čisto na účel testovania a ladenia parametrov) a ako experimentálnu štatistiku. Ak sa experimentálna štatistika osvedčí, možno rozhodnúť o jej nasadení ako produkčnej štatistiky (stane sa oficiálnym štatistickým produktom). Analytický model je možné nasadiť rôznymi spôsobmi:

- **API:** napríklad keď sa výstupy modelu vkladajú ako vstup do iného produktu alebo služby plne automatizovaným spôsobom. API postavené okolo modelu môže stačiť na uľahčenie interakcie medzi modelom hĺbkovej analýzy Big Data a inými pripojenými službami.
- **Poloautomatický proces:** ak sa model používa na automatizáciu procesu klasifikácie tým, že pomáha ľudskému personálu, je vhodné vybudovať aj servisnú aplikáciu s používateľsky prívetivým rozhraním.
- **Štatistický produkt:** dátový model sa môže používať priamo na odhad konečných štatistík v takom prípade je často vhodné publikovať pre verejnosť konečný štatistický produkt.
- **Webová interaktívna aplikácia:** front-end pre používateľov, ktorí majú záujem experimentovať s modelom, vkladať doň vlastné údaje a pracovať s výsledkami prognóz.

Súčasťou nasadenia musí byť zavedenie pracovného postupu, ktorý zabezpečí kontinuálne strojové učenie modelu z nových historických údajov počas prevádzky a zlepšovanie jeho odhadov. Znamená to zahrnutie tém ako spätná väzba na odhady modelu, dopĺňanie klasifikácie a dopĺňanie nových údajov dedikovaným tímom expertov.

3. ZÁVER

Realizovaný projekt v oblasti zberu a spracovania Big Data na použitie v štatistike ukázal, že existujú metodiky a prístupy, ktoré umožňujú pracovať s Big Data pri tvorbe

nových analytických modelov. Výstupy získané spracovaním a analýzou údajov nakoniec nemusia byť len štatistickým výstupom, ale môžu poukázať na potenciál ich využitia v iných odvetviach, ako je napr. využitie v oblasti pohybu obyvateľstva na tvorbu funkčných regiónov, alebo model nálad na zvýšenie porozumenia verejnej mienky, identifikácie sociálneho napätia a merania sentimentov na rôzne politiky a problémy a podobne.

LITERATÚRA

[1] AGGARWAL, C. C. et al.: Data mining: The Textbook. New York: Springer, 2015.
[2] European Commision, Big data. [online]. [cit. 13-11-2023]. Dostupné na: https://digital-strategy.ec.europa.eu/en/policies/big-data.
[3] LANEY, D.: 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group: Application Delivery Strategies, file 949. 2001.
[4] LOISON B. – KUONEN D.: Are Current Frameworks in the Official Statistical Production Appropriate for the Usage of Big Data and Trusted Smart Statistics? 2018.
[5] TAM, S. M. – CLARKE, F.: Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. Methodology and Data Management Division. Australian Bureau of Statistics, In: International Statistical Review, 2015, č. 3, s. 436 – 438.

RESUMÉ

Využívanie Big Data v štatistike je čoraz častejším javom. Potenciál využitia sa preukázal vo viacerých projektoch, ktoré sú realizované aj na medzinárodnej úrovni. Podstatným je aby proces využívania údajov mal jasne definované pravidlá, ktoré umožnia spracovanie akýchkoľvek Big Data za účelom hľadania ich využitia vo výskume, experimentálnej štatistike alebo v následnej produkcii. Za týmto cieľom je nevyhnutné neustále preskúmavať možnosti využívania Big Data, ako aj sprostredkúvať know how z realizovaných projektov s partnermi na úrovni národných štatistických úradov alebo inými potenciálnymi partnermi.

RESUME

The use of Big Data in statistics is an increasingly wide-spread phenomenon. Its potential has been demonstrated in several projects that are implemented at the international level. It is essential that the process of using data has clearly defined the rules that will enable the processing of any Big Data for the purpose of finding their utilization in research, experimental statistics or in post-production. Therefore, it is essential to constantly explore the possibility of using Big Data, as well as to share the know-how from the implemented projects with the partners at the level of statistical offices or other potential partners.

PROFESIJNÝ ŽIVOTOPIS

***Peter Ďuriš** je absolventom Ekonomickej univerzity v Bratislave. Profesionálne začínal ako procesný analytik v konzultačnej spoločnosti Centire, kde mal na starosti riadenie analytických tímov, ako aj riadenie projektov. Od roku 2012 je konateľom spoločnosti Go SMART, s. r. o., v ktorej okrem iného pôsobí ako konzultant v oblasti využívania nových zdrojov údajov a metód v oficiálnej štatistike. Rovnako sa zaoberá prípravou projektov a ich realizáciou v štátnej, vo verejnej ale aj v súkromnej sfére.*

KONTAKT

peter.duris@gosmart.consulting

Informatívny článok/Informative article

Dagmar CEĽUCHOVÁ BOŠANSKÁ, Juraj BÁRDY
Alistiq, s. r. o.

POUŽITIE BIG DATA V ŠTATISTIKE

USE OF BIG DATA IN STATISTICS

ABSTRAKT
Využitie veľkých údajov, takzvaných Big Data, na doplnenie oficiálnych štatistík predstavuje zaujímavú príležitosť. Tento typ údajov možno získavať zo sociálnych sietí, obchodných systémov a prepojených inteligentných zariadení, ktoré sa nazývajú aj internet vecí, čo môže viesť k rýchlejšiemu vytváraniu štatistík takmer v reálnom čase. Článok sa zameriava na kategórie Big Data a ich vlastnosti, ako napríklad údaje bez závislostí (kategoriálne, kvantitatívne, textové) a údaje so závislosťami (časové, sieťové, priestorové). Big Data nachádzajú uplatnenie v oblastiach ako cestovný ruch, demografia, vývoj cien, priemyselná produkcia a v mnohých iných. Napriek príležitostiam prináša aj riziká a problémy, ktoré je potrebné zohľadniť.

ABSTRACT
The use of Big Data to complement official statistics represents an intriguing opportunity. This type of data can be obtained from social networks, business systems, and interconnected smart devices, also referred to as the Internet of Things, enabling near real-time production of statistics. The article focuses on categories of Big Data and their characteristics, such as data without dependencies (categorical, quantitative, textual) and data with dependencies (temporal, networks, spatial). Big Data are widely used in areas like tourism, demography, price development, and industrial production. Despite the opportunities, it also brings risks and challenges that need to be taken into consideration.

KĽÚČOVÉ SLOVÁ
Big Data v štatistike, nové zdroje údajov, údaje bez závislosti, údaje zo závislosťami, strojové učenie

KEY WORDS
Big Data in statistics, new data sources, data without dependencies, data with dependencies, machine learning

1. ÚVOD
Rýchly pokrok v oblasti technológií dnes kladie nové nároky na oficiálne štatistiky, pričom vlády, podniky a občania očakávajú presné údaje ideálne v reálnom čase. Udalosti, ako nedávna pandémia ochorenia COVID-19 ukazujú potrebu včasných a detailných informácií s cieľom rýchlej reakcie, a zároveň odhaľujú nedostatky existujúcich systémov štatistickej produkcie.

Počet obyvateľov, ktorí vykonávajú svoje každodenné aktivity online a vlastnia mobilné telefóny, každý deň prispievajú k obrovskému množstvu údajov v digitálnej ekonomike. Ak by sa tieto údaje spracovali etickým spôsobom a zaručila sa ochrana

súkromia, mohli by sme ich spracovaním získať podstatný príspevok k oficiálnym štatistikám.

S technologickým pokrokom a narastajúcou potrebou časovo aktuálnych a spoľahlivých údajov sa tak vytvára dopyt po revízii prístupov k produkcii oficiálnych štatistík ako aj k redefinícii úloh národných štatistických úradov. Štatistické úrady v tejto súvislosti potrebujú integrovať nové zdroje údajov udržateľným spôsobom. To zahŕňa budovanie partnerstiev so zainteresovanými stranami (zdrojmi Big Data), investovanie do novej infraštruktúry a kompetencií a prispôsobenie rámcov na spracovanie údajov tak, aby reflektovali kľúčové charakteristiky Big Data.

R. Kitchin [7] sumarizuje kľúčové charakteristiky zdrojov údajov označovaných ako Big Data takto:

- veľký objem údajov,
- vznikajúce v reálnom čase a vo vysokej frekvencii,
- rozmanité v charaktere, zahŕňajúce štruktúrované aj neštruktúrované údaje,
- rozsiahlej mierky, snažiace sa zachytiť celé populácie alebo systémy,
- s jemným rozlíšením umožňujúcim individuálnu indexáciu,
- umožňujúce prepájať rôzne databázy,
- flexibilné, umožňujúce jednoduchú rozšíriteľnosť (pridávať nové polia) a škálovateľnosť (meniť veľkosť, záber).

Podľa autora môžeme zdroje Big Data rozdeliť na tri kategórie:
1) riadené, 2) automatizované a 3) údaje postavené na aktivitách dobrovoľníkov.

Riadené údaje sú generované prostredníctvom digitálnych foriem sledovania, pri ktorom je technológia smerovaná na miesto alebo človeka prostredníctvom ľudského operátora.

Automatizované údaje sa niekedy označujú aj pojmom strojové, sú generované ako inherentná funkcia zariadenia alebo systému a obsahujú tzv. stopy, ktoré zanechávajú digitálne zariadenia. Príkladom sú záznamy mobilných zariadení v infraštruktúre mobilného operátora, interakcie v internetovej sieti či pohyby používateľov pri prehliadaní webovej stránky. Často sa používajú i automatizované nástroje na prechádzanie a prehľadávanie webových stránok na zhromažďovanie informácií (Web Crawlers, Web Scrapers). Môžeme sem zaradiť aj údaje rôznych senzorov, ktoré zaznamenávajú teplotu, tlak, polohu, rýchlosť a podobne a sú súčasťou konkrétneho zariadenia alebo prostredia. Takéto údaje sú vhodné najmä na sledovanie sp*rávani*a spotrebiteľov, prípadne určenej skupiny populácie. S nárastom využívania moderných technológií v každodennom živote ich množstvo exponenciálne narastá.

Tretiu skupinu tvoria ***údaje postavené na aktivitách dobrovoľníkov***.
Vznikajú ako výsledok interakcie v sociálnych sieťach alebo prostredníctvom crowdsourcingu[1] údajov, keď skupiny dobrovoľníkov prispievajú k vzniku spoločnej

[1] *Crowdsourcing je metóda získavania informácií, riešenia problémov alebo vykonávania úloh prostredníctvom využitia skupiny ľudí. Ide o proces, keď organizácia, spoločnosť alebo iná entita deleguje úlohy alebo otázky verejnosti, pričom sa spolieha na múdrosť a schopnosti veľkého počtu jednotlivcov alebo skupín. Účastníci crowdsourcingu zvyčajne prispievajú svojimi nápadmi, znalosťami alebo schopnosťami, čo umožňuje rýchlejšie a efektívnejšie dosahovanie cieľov organizácie. Táto*

dátovej platformy, napríklad ako to je v projekte OpenStreetMap. Fenoménom posledného desaťročia sú sociálne siete, na ktorých vzniká enormné množstvo obsahu generovaného používateľmi (príspevky, komentáre, fotky, videá, reakcie). Spoločnosti využívajú sociálne siete ako jeden z marketingových nástrojov. Analýzou týchto údajov zisťujú správanie spotrebiteľov, ich sentiment vzhľadom na produkty či služby. Takéto porozumenie potom pomáha v rozhodovacom procese [7].

2. BIG DATA V OFICIÁLNEJ ŠTATISTIKE

Big Data otvárajú nové možnosti modernizácie oficiálnej štatistiky vďaka využitiu metód na spracovanie veľkého množstva údajov, ktoré vznikajú ako vedľajší produkt v digitalizovanej spoločnosti a ekonomike. Dopĺňajú tradičné zdroje oficiálnych štatistík ako sú údaje zo štatistického zisťovania a administratívne zdroje. Big Data v oficiálnej štatistike predstavujú externé údaje generované digitálnymi aktivitami, ktoré sa spracúvajú na sekundárne účely štatistiky. Môže ísť o údaje o používaní mobilného telefónu, aktivite na sociálnych sieťach, využívaní digitálnych peňazí alebo o údaje generované senzormi a podobne [3].

Vďaka spracovaniu a využívaniu Big Data je možné vytvárať nielen nové štatistické produkty, ktoré doteraz s danými parametrami kvality neboli možné, ale aj automatizovať niektoré náročné manuálne úlohy pri spracúvaní údajov ako ich kategorizácia alebo dopĺňanie chýbajúcich hodnôt.

Využitie Big Data v oficiálnej štatistike prináša so sebou niekoľko výziev, ktoré treba zohľadniť pri implementácii a spracovaní riešení. Medzi hlavné výzvy, ktoré treba adresovať patrí regulačný rámec; otázky súkromia, etiky a dôvery; modely partnerstva a náklady na zabezpečenie prístupu k zdrojom údajov [6].

Regulačný rámec: Použitie Big Data v oficiálnej štatistike ešte nie je dostatočne legislatívne podchytené. Neexistujú štandardy pre jednotlivé typy Big Data. Častým problémom je aj zabezpečenie súladu s právnymi predpismi týkajúcimi sa spracovania a ochrany údajov. Potrebná je preto úprava predpisov, aby mal Štatistický úrad SR garantovaný prístup k potrebným zdrojom údajov na účel experimentovania a následnej produkcie. Dôležité je najmä nastavenie pravidiel na prístup k Big Data (dobrovoľné sprístupnenie, zabezpečenie povinného sprístupnenia, nákup údajov).

Súkromie, etika a dôvera: Veľké množstvo údajov z nových zdrojov môže obsahovať citlivé informácie, čo predstavuje výzvu v oblasti ochrany osobných údajov. Štatistický úrad SR musí verejnosti zaručiť, že aplikuje etické postupy pri spracúvaní Big Data. Postupy týkajúce sa anonymizácie a pseudonymizácie je potrebné prehodnotiť vzhľadom na rozsah a bohatstvo údajov.

Prístup k Big Data a modely partnerstva: Získanie prístupu k Big Data môže byť výzvou, pretože informácie sú väčšinou v súkromnom vlastníctve alebo pod kontrolou súkromných spoločností. Budovanie efektívnych partnerstiev so súkromným sektorom je preto dôležité na zabezpečenie prístupu k relevantným informáciám ako i technológiám na ich spracovanie.

---

metóda umožňuje využiť rozsiahle množstvo zdrojov a perspektív a otvára priestor na spoluprácu a inovácie vo vyriešení problémov alebo v dosiahnutí výsledkov.

Náklady a verejné obstarávanie: Spracovanie a analýza Big Data býva finančne náročná, čo predstavuje výzvu pre rozpočet. Je potrebné zabezpečiť správne technológie spracovanis Big Data a tiež pravidelný prístup k zdrojom údajov [2].

3. KATEGÓRIE BIG DATA

Jedným zo zaujímavých aspektov nových zdrojov Big Data je široká škála typov údajov, ktoré sú k dispozícii na analýzu. V procesoch spracovania Big Data existujú dva typy údajov rôznej zložitosti:

1. **Údaje bez závislostí:** Zvyčajne sa týkajú jednoduchých typov údajov, ako sú viacrozmerné údaje alebo textové údaje. Tieto typy údajov sú najjednoduchšie a najčastejšie sa s nimi stretávame. V týchto prípadoch záznamy údajov nemajú žiadne špecifikované závislosti medzi údajovými položkami alebo atribútmi (premennými). Príkladom je súbor záznamov o jednotlivcoch, ktoré obsahujú ich vek, pohlavie a PSČ.

2. **Údaje so závislosťami**: V týchto prípadoch môžu medzi údajovými položkami existovať implicitné alebo explicitné vzťahy. Napríklad dátový súbor sociálnej siete obsahuje množinu uzlov (údajových položiek), ktoré sú navzájom spojené množinou hrán (vzťahov). Typickým príkladom sú časové rady, kde medzi jednotlivými položkami (údajmi) radu v čase existujú implicitné závislosti [1].

Vo všeobecnosti údaje so závislosťami sú náročnejšie z dôvodu zložitosti spôsobenej už existujúcimi vzťahmi medzi údajovými položkami. Takéto závislosti medzi údajovými položkami sa musia začleniť priamo do analytického procesu, aby sa získali kontextovo zmysluplné výsledky.

3.1. Údaje bez závislostí

Táto forma údajov je najjednoduchšia a zvyčajne sa vzťahuje na viacrozmerné údaje. Tieto údaje zvyčajne obsahujú množinu záznamov. Záznam sa tiež označuje ako dátový bod, inštancia, príklad, transakcia, entita, objekt alebo vektor vstupných premenných v závislosti od daného prípadu použitia. Každý záznam obsahuje množinu polí, ktoré sa označujú aj ako vstupné premenné, dimenzie alebo charakteristické znaky. Najčastejšie v tomto článku budeme používať výraz (vstupné) premenné. Tieto polia opisujú rôzne vlastnosti daného záznamu.

3.2. Údaje so závislosťami

V praxi môžu byť rôzne hodnoty údajov (implicitne) navzájom prepojené časovo, priestorovo alebo prostredníctvom explicitných prepojení sieťových vzťahov medzi údajovými položkami. Znalosť už existujúcich závislostí výrazne mení proces hĺbkovej analýzy Big Data, pretože hĺbková analýza údajov sa týka predovšetkým hľadania vzťahov medzi dátovými položkami. Existuje niekoľko typov závislostí, ktoré môžu byť implicitné alebo explicitné:

1. **Implicitné závislosti**: V tomto prípade závislosti medzi údajovými položkami nie sú výslovne špecifikované, ale je známe, že „typicky“ existujú v danej doméne. Napríklad po sebe idúce hodnoty teploty zhromaždené senzorom budú pravdepodobne navzájom veľmi podobné. Preto ak sa hodnota teploty zaznamenaná senzorom v určitom čase výrazne líši od hodnoty zaznamenanej v nasledujúcom okamihu, potom je to mimoriadne nezvyčajné a môže to byť

zaujímavé pre proces hĺbkovej analýzy údajov. Táto situácia sa líši od viacrozmerných dátových súborov, kde sa s každým dátovým záznamom zaobchádza ako s nezávislou entitou.

2. **Explicitné závislosti**: Zvyčajne sa týkajú grafických alebo sieťových údajov, v ktorých sa hrany používajú na určenie explicitných vzťahov. Grafy sú veľmi silnou abstrakciou, ktorá sa často používa ako prechodná reprezentácia na riešenie problémov s hĺbkovou analýzou Big Data v kontexte iných typov údajov [8].

Nasledujúca tabuľka č. 1 sumarizuje konkrétnejšie typy údajov v tejto skupine.

Tabuľka č. 1: Prehľad konkrétnych typov údajov so závislosťami

Typ údajov so závislosťami	**Opis**
Údaje časových radov	Údaje časových radov obsahujú hodnoty, ktoré sa zvyčajne generujú kontinuálnym meraním v čase. Napríklad environmentálny senzor bude nepretržite merať teplotu. Takéto údaje majú zvyčajne implicitné závislosti zabudované do hodnôt prijatých v priebehu času. Napríklad susedné hodnoty zaznamenané snímačom teploty sa budú zvyčajne v priebehu času plynulo meniť a tento faktor je potrebné výslovne použiť v procese hĺbkovej analýzy Big Data. Povaha časovej závislosti sa môže v závislosti od prípadu použitia výrazne líšiť. Napríklad niektoré formy údajov zo snímačov môžu vykazovať periodické vzorce meraného atribútu v priebehu času. Dôležitým aspektom hĺbkovej analýzy časových radov je extrakcia takýchto závislostí v údajoch. Na formalizáciu otázky závislostí spôsobených časovou koreláciou sú atribúty rozdelené do dvoch typov: Kontextové atribúty: Sú to atribúty, ktoré definujú kontext, na základe ktorého sa implicitné závislosti vyskytujú v údajoch. Napríklad v prípade údajov zo snímačov sa za kontextový atribút môže považovať časová pečiatka, pri ktorej sa hodnota merania odčítala. Iné typy údajov môžu mať viac ako jeden kontextový atribút. Atribúty správania: Predstavujú hodnoty, ktoré sa merajú v konkrétnom kontexte. Pri príklade senzora je teplota hodnotou atribútu správania. Je možné mať viac ako jeden atribút správania. Ak napríklad viaceré senzory zaznamenávajú údaje pri synchronizovaných časových pečiatkach, výsledkom je viacrozmerný dataset časových radov. Kontextové atribúty majú zvyčajne silný vplyv na závislosti medzi hodnotami atribútov správania v údajoch. Údaje časových radov sú relatívne bežné v mnohých senzorových aplikáciách, prognózach a analýzach finančného trhu.
Diskrétne sekvencie a reťazce	Diskrétne sekvencie možno považovať za kategoriálnu analógiu údajov časových radov. Rovnako ako v prípade údajov časových radov je kontextovým atribútom časová pečiatka alebo index pozície v poradí. Atribút správania je kategoriálna hodnota, preto sú diskrétne sekvenčné údaje definované podobným spôsobom ako údaje časových radov. Príkladom diskrétnej sekevencie môže byť postupnosť webových prístupov, kde sa pre 100 rôznych prístupov zhromažďuje adresa webovej stránky a pôvodná IP adresa žiadosti. To predstavuje diskrétnu postupnosť dĺžky $n = 100$ a dimenzionality $d = 2$. Obzvlášť častým prípadom v sekvenčných údajoch je jednorozmerný scenár, v ktorom hodnota d je 1. Takéto sekvenčné údaje sa tiež označujú ako reťazce. Diskrétne sekvencie sú pre algoritmy hĺbkovej analýzy často náročnejšie, pretože nemajú plynulú hodnotovú kontinuitu ako údaje časových radov.
Priestorové údaje	V priestorových údajoch sa meria mnoho nepriestorových atribútov (napríklad teplota, tlak). Napríklad meteorológovia často zhromažďujú údaje o teplote morskej hladiny, aby predpovedali výskyt hurikánov.

	V takýchto prípadoch priestorové súradnice zodpovedajú kontextovým atribútom, zatiaľ čo atribúty, ako je teplota, zodpovedajú atribútom správania. Zvyčajne existujú dva priestorové atribúty. Rovnako ako pri údajoch časových radov je možné mať viacero atribútov správania. Napríklad pri aplikácii teploty morskej hladiny je možné merať aj iné atribúty správania, ako je tlak. Hĺbkováanalýza priestorových údajov úzko súvisí s hĺbkovou analýzou údajov v časových radoch, pretože atribúty správania v najčastejšie študovaných priestorových prípadoch použitia sú kontinuálne, hoci niektoré prípady použitia môžu používať aj kategoriálnu atribúty. Preto sa kontinuita hodnôt pozoruje vo všetkých súvislých priestorových lokalitách, rovnako ako sa pozoruje kontinuita hodnôt v súvislých časových pečiatkach v údajoch časových radov.
Časopriestorové údaje	Osobitnou formou priestorových údajov sú časopriestorové údaje, ktoré obsahujú priestorové aj časové atribúty. Presná povaha údajov závisí aj od toho, ktoré z atribútov sú kontextové a ktoré sú behaviorálne. Najbežnejšie sú dva druhy časopriestorových údajov: Priestorové aj časové atribúty sú kontextové: Tento druh údajov možno považovať za priame zovšeobecnenie priestorových aj časových údajov. Je obzvlášť užitočný, keď sa súčasne meria priestorová a časová dynamika konkrétnych atribútov správania. Napríklad keď je potrebné merať zmeny teploty povrchu mora v priebehu času. V takýchto prípadoch je teplota atribútom správania, zatiaľ čo priestorové a časové atribúty sú kontextové. Časový atribút je kontextový, priestorové atribúty sú behaviorálne: Tento druh údajov možno takisto považovať za údaje časových radov. Priestorová povaha atribútov správania však v mnohých scenároch poskytuje aj lepšiu interpretovateľnosť a cielenejšiu analýzu. Najbežnejšia forma týchto údajov vzniká v kontexte analýzy trajektórie. Treba zdôrazniť, že akékoľvek 2- alebo 3-dimenzionálne údaje časových radov možno mapovať na trajektórii. Je to užitočná transformácia, pretože to znamená, že algoritmy hĺbkovej analýzy trajektórie sa môžu použiť aj pre 2- alebo 3-dimenzionálne údaje časových radov.
Sieťové a grafové údaje	Pri sieťových a grafových údajoch môžu údajové hodnoty zodpovedať uzlom v sieti, zatiaľ čo vzťahy medzi údajovými hodnotami môžu zodpovedať hranám v sieti. V niektorých prípadoch môžu byť atribúty priradené k uzlom v sieti. Aj keď je možné priradiť atribúty k hranám v sieti, je to oveľa menej bežné. Hrana môže byť nasmerovaná alebo nenasmerovaná, v závislosti od použitia. Napríklad webový graf môže obsahovať nasmerované hrany zodpovedajúce smerom hypertextových odkazov medzi stránkami, zatiaľ čo priateľstvá v sociálnej sieti Facebook sú nenasmerované. Niektoré príklady údajov, ktoré sú znázornené ako grafy: Webový graf: Uzly zodpovedajú webovým stránkam a hrany zodpovedajú hypertextovým odkazom. Uzly majú textové atribúty zodpovedajúce obsahu na stránke. Sociálne siete: Uzly zodpovedajú aktérom sociálnych sietí, zatiaľ čo hrany zodpovedajú priateľským väzbám. Uzly môžu mať atribúty zodpovedajúce obsahu sociálnej stránky. V niektorých špecializovaných formách sociálnych sietí, ako sú e-mailové siete alebo siete chatovacích aplikácií, môžu mať hrany obsah, ktorý je s nimi spojený. Tento obsah zodpovedá komunikácii medzi rôznymi uzlami.

Zdroj: [1]

4. PRÍKLADY POUŽITIA BIG DATA

Ak už poznáme druhy údajov, poďme sa pozrieť na konkrétne možnosti a príklady použitia Big Data v štatistike. Využitie Big Data v oblasti oficiálnej štatistiky otvára nové príležitosti na získavanie hodnotných informácií o rôznych aspektoch spoločnosti. Nové zdroje údajov môžu poskytnúť komplexnejší a aktuálnejší pohľad

na ekonomické, sociálne, environmentálne a ďalšie javy. Príklady je možné identifikovať v mnohých oblastiach, od zdravotnej starostlivosti cez komunikáciu po spoločenské vedy. Nasledujúca tabuľka č. 2 prenáša prehľad najdôležitejších zdrojov Big Data, ktoré v súčasnosti skúmajú národné štatistické úrady.

Tabuľka č. 2: Príklady domén oficiálnej štatistiky pre jednotlivé zdroje BD

Zdroj	Typ údajov	Príklad štatistickej domény
Telekomunikačná sieť	Údaje mobilných telefónov	Cestovný ruch, populácia
Internet	Vyhľadávanie na webe	Zamestnanosť, migrácia
	e-commerce stránky	Vývoj cien
	Stránky zamestnávateľov	Obchodné registre, indikátory kybernetickej bezpečnosti
	Inzercia zamestnaní	Zamestnanosť
	Inzercia realít	Vývoj cien (realít)
	Sociálne siete	Wellbeing, spotrebiteľská dôvera, HDP
	Interakcie so spravodajskými médiami	Wellbeing, demokratizácia
	Vnútroštátne platformy ubytovania alebo vstupeniek	Cestovný ruch, kultúra
Internet vecí	Senzory v doprave	Nowcasting, Mi
	Meteostanice	Životné prostredie, doprava
	Inteligentné merače	Spotreba energií, produktivita priemyslu
	Satelitné snímkovanie	Využitie pôdy, poľnohospodárstvo, životné prostredie
Generované transakcie	Záznamy leteckej dopravy	Migrácia, cestovný ruch, emisie
	Údaje o maloobchode (supermarkety)	Vývoj cien, spotreby domácností
	Lekárske záznamy	Epidemiológia
	Transakcie bánk, bankových kariet	HDP, ekonomické štatistiky
	Transakcie na burzovom trhu	HDP, ekonomické štatistiky
Crowdsourcing	Dobrovoľné webové stránky s geografickými informáciami (OpenStreetMap, Wikimapia, Geowiki)	Využitie pôdy, poľnohospodárstvo, životné prostredie
	Komunitné zbierky obrázkov (flickr, Instagram, Panoramio)	Wellbeing, spotrebiteľská dôvera, HDP

Zdroj: vlastné spracovanie autorov

5. RIZIKÁ SPOJENÉ S VYUŽÍVANÍM BIG DATA

Ako každý nový prístup, aj využitie Big Data pre oficiálnu štatistiku prináša riziká a problémy, ktoré je potrebné pomenovať. Jedným z hlavných rizík spojených s využívaním Big Data v oficiálnej štatistike je to, že metodológie na tvorbu štatistiky nebudú správne aplikované. Veľa zdrojov Big Data ako napríklad sociálne siete obsahujú údaje z pozorovaní, ktoré neboli zámerne navrhnuté na údajovú analýzu, a preto nemajú dobre definovanú cieľovú populáciu, štruktúru a kvalitu. Na takéto údaje sa nedajú aplikovať tradičné štatistické metódy založené na teórii vzorkovania. Pre veľa zdrojov Big Data interpretácia údajov a ich vzťahu s daným sociálnym fenoménom nie je ani zďaleka taká očividná a tak môže byť chybná, alebo minimálne nepresná, či nejasná.

Ďalším rizikom je nepresné porovnávanie trendov v indexoch alebo štatistikách vypočítaných z Big Data. Napríklad populácia zložená z používateľov vybranej

sociálnej siete sa môže časom zásadne meniť (napríklad pri sieti Facebook je známy trend, že sa mení v čase vekové zloženie používateľov [5]. Pri Big Data sa zvykne kompenzovať nepresnosť údajov ich objemom. Rizikom je, ako tento fakt ovplyvní presnosť oficiálnych štatistík a aká je prijateľná miera zníženej presnosti za cenu napríklad štatistík, ktoré sú včasnejšie alebo podrobnejšie.

Rizikom je tiež porušenie súkromia používateľov pri práci s Big Data, či niektorých právnych predpisov ohľadom vlastníctva a autorského práva, keďže tieto oblasti sú často oveľa menej jasné ako pri práci s iným typom údajov. Kľúčové je nenarušiť dôveryhodnosť Štatistického úradu SR ako inštitúcie. Preto je dôležité dôsledne aplikovať pravidlá transparentnosti pri získavaní a využívaní Big Data a súvisiacich modelov. Tiež treba mať na pamäti, že Big Data musia byť zozbierané od používateľov s ich súhlasom na daný účel. Alebo musí ísť o verejne dostupné Big Data, ako sú napríklad verejné príspevky používateľov na sociálnych sieťach či na webových stránkach [2].

Rizikom sú tiež zvýšené nároky na prenos, ukladanie a spracovanie Big Data, ktoré môžu vyvolať privysoké dodatočné náklady. Využitie efektívnych cloudových služieb a moderných open source nástrojov môže pomôcť do veľkej miery s týmto problémom.

Ďalším rizikom je nestálosť zdrojov Big Data – ak daný zdroj prestane byť dostupný, naruší sa kontinuita štatistických produktov v čase, ktoré tento zdroj využívali. Pre veľa používateľov štatistických produktov je táto kontinuita v čase zásadná.

V neposlednom rade je rizikom aj nájdenie a udržanie dostatočných personálnych kapacít, ktoré vedia pracovať s Big Data a aplikovať moderné metódy spracovania Big Data [7].

6. ZÁVER

Rozvoj nových metód spracovania údajov označovaných ako techniky strojového učenia prináša tiež nové možnosti ako opisovať svet. V období rastúcej dostupnosti údajov je preto potrebné zaoberať sa novými zdrojmi údajov a prinášať rýchle informácie o spoločnosti, hospodárstve alebo zdraví. Používatelia štatistík často potrebujú poznať informácie o aktuálnych trendoch ako aj vstupy do vlastných modelov pre predikciu – analýzy čo bude, analýzy čo by bolo, keby a kontrafaktuálne analýzy. Na tieto účely nie je vždy nevyhnutné mať reprezentatívne údaje a tak Big Data môžu byť veľmi cenné a užitočné. Avšak pri práci s Big Data je stále nevyhnutné mať jasné povedomie o kvalite údajov a procese ich tvorby, čo samo osebe predstavuje výzvu.

Okrem zdrojov údajov je potrebné dôkladne a transparentne opísať modely a ich spracovanie. Akékoľvek využitie modelov strojového učenia pri spracovaní Big Data by malo byť explicitné, zdokumentované a transparentné pre používateľov. Preto by každý model mal byť postavený na skutočných pozorovaných údajoch za relevantné obdobie, ktoré sa týka ekonomických a sociálnych javov, ktoré sa snažíme štatisticky opísať. Tvorba štatistických produktov musí byť založená na experimentálnom vyhodnotení výsledkov.

Využívanie Big Data na účely štatistiky možno považovať za revolučné. Nie všetky národné štatistické úrady sú na takéto zmeny pripravené. Pripraviť zavedenie

systematického využívania Big Data znamená zásadným spôsobom prebudovať myslenie a personálne obsadenie aspoň v kľúčových útvaroch úradu. V Štatistickom úrade SR dnes na inovatívnych projektoch Big Data robia pracovníci popri svojej bežnej činnosti a nie je to ich hlavná pracovná náplň, alebo sú zapojení výhradne do projektov (v rokoch 2022 až 2023 sa realizovali dva projekty: SEABD – Socioekonomické aspekty Big Data v štatistike a DCM – Dynamický cenový model). V rámci projektov sa okrem toho podarilo vybudovať kapacity aj z pohľadu IT (sekcie informačných systémov). Bol vytvorený priestor na výskum a vývoj Big Data, ktorý bude flexibilne podporovať základné nástroje na spracovanie Big Data. Vytvorilo sa aj produkčné prostredie pre experimentálne štatistiky. Dôležité je, aby sa výskumné prostredie rýchlo a bezpečne sprístupnilo riešiteľským tímom.

Existujú dva základné prístupy, ako dosiahnuť využívanie Big Data a vytvoriť potrebné personálne a technické zázemie. Prvým prístupom je vytvorenie nového centra, orientovaného na využitie Big Data a inovácie v štatistike. Týmto smerom išiel holandský štatistický úrad, ktorý vytvoril Centrum pre štatistiku Big Data (CBDS – Center for Big Data Statistics). Misiou CBDS je proaktívne hľadať inovatívne využitie Big Data v oficiálnej štatistike a spolupracovať s partnermi, ktorí môžu predstavovať producentov údajov a používateľov výstupov pre jednotlivé prípady použitia [4].

Druhým prístupom je posilnenie súčasných útvarov v samotnom Štatistickom úrade SR. V súčasných podmienkach Štatistického úradu SR sa ako vhodnejší javí tento druhý prístup. Zabráni sa tak vzniku paralelného centra a lepšie sa prepoja nové inovatívne metódy so súčasnou praxou. Na zvládnutie tejto zmeny možno navrhnúť nasledujúce opatrenia:

1. Vytvoriť novú organizačnú zložku v Štatistického úradu SR, ktorá bude mať na starosti inovácie a Big Data. Mohla by mať na starosti prípravu a aktualizáciu metodík, výber scenárov na riešenie, programový manažment Big Data, koordináciu ostatných zapojených útvarov, posudzovanie kvality a publikovanie výsledkov.
2. Kapacitne treba posilniť odborné sekcie, aby mali vyčlenené tímy expertov na prácu s Big Data, ktorí by sa špecializovali na problematiku a zaradili by sa do práce na projektoch.
3. Je vhodné zaviesť mechanizmus na výber nového prípadu použitia riešenia a posudzovania kvality štatistík Big Data. Navrhnúť plán tém na realizáciu. Takýto plán poskytuje štruktúrovaný prehľad o témach, ktoré treba zahrnúť, minimalizuje riziko prehliadnutia dôležitých aspektov a umožňuje efektívne a účinné fungovanie procesu. Pomocou neho pracovníci majú jasné usmernenie a postup pri analyzovaní a vyhodnocovaní dát, čo prispieva k spoľahlivým a presným výsledkom pri hodnotení a zlepšovaní týchto štatistík. Celkovo navrhnutie plánu tém zabezpečuje systematický a kvalitný prístup k riešeniu problémov v oblasti štatistík Big Data.
4. V neposlednom rade odporúčame vytvoriť program vzdelávania zamestnancov v oblasti Big Data (metódy, nástroje, použitie) a zabezpečiť pravidelné školenia.
5. Z pohľadu marketingu je vhodné nastaviť komunikáciu projektov a výsledkov Big Data.

LITERATÚRA

[1] AGGARWAL, C. C.: Data Mining: The Textbook. New York: Springer, 2015. 727 s. ISBN 978-3-319-14142-8 (eBook).
[2] BORMIDA, M. D.: The Big Data World: Benefits, Threats and Ethical Challenges. In: Ethical Issues in Covert, Security and Surveillance Research. Emerald Publishing Limited, 2021, s. 71 – 91. ISBN 978-1-80262-414-4.
[3] BRAAKSMA, B – ZEELENBERG, K.: Big Data in Official Statistics. Discussion paper, Centraal Bureau voor de Statistiek, 2020. 23 s.
[4] DE BROE, S. et al.: Big Data to improve policy and decision making: The Experience of Statistics Netherlands. In: Conference on Big Data in Social Sciences & Public Policies, Colmex, Mexico City, 2019.
[5] ENBERG, J.: Facebook can't shake its teen problem, but its user base is getting younger. Insider Intelligence, [online]. [cit. 24-11-2023]. Dostupné na: https://www.insiderintelligence.com/content/facebook-teen-problem
[6] HOWE, E. – ELENBERG, F.: Ethical challenges posed by big data. In: Innovations in clinical neuroscience, 2020, č. 17, s. 24 – 30.
[7] KITCHIN, R.: Big data and human geography: Opportunities, challenges and risks. In: Dialogues in Human Geography, 2013, č. 3, s. 262 – 267.
[8] YANG, CH. C. et al.: Identifying implicit and explicit relationships through user activities in social media. In: International Journal of Electronic Commerce, 2013, č. 18, s.73 – 96.

RESUMÉ

V súčasnosti sa využitie Big Data ako nového zdroja na doplnenie oficiálnych štatistík javí ako veľmi zaujímavá príležitosť. Big Data umožňujú produkovať štatistiky využitím metód na spracovanie obrovského množstva údajov, ktoré vznikajú ako vedľajší produkt v rámci digitalizovanej spoločnosti a ekonomiky. Dopĺňajú tradičné zdroje oficiálnych štatistík ako sú údaje zo štatistického zisťovania a administratívne zdroje. Big Data v štatistike predstavujú externé informácie z digitálnych aktivít, ako je používanie mobilného telefónu, aktivity na sociálnych sieťach, digitálne transakcie alebo údaje generované senzormi a podobne.
Spracovaním Big Data je možné vytvárať nové štatistické produkty a automatizovať náročné manuálne úlohy, ako je kategorizácia alebo dopĺňanie chýbajúcich hodnôt. Využívanie Big Data v oficiálnych štatistikách však prináša výzvy vrátane regulačného rámca, otázok súkromia, etiky a dôvery, modelov partnerstva a nákladov na zabezpečenie prístupu k zdrojom údajov.
V článku sa sústredíme najmä na prehľad jednotlivých kategórii Big Data a na opis ich vlastností a charakteristík. Dôležité kategórie údajov sú údaje bez závislostí, ako napríklad kategoriálne, kvantitívne a textové údaje, a údaje so závislosťami, ako napríklad časové rády, siete a priestorové údaje. Big Data je možné použiť v doménach, ako cestovný ruch, demografia, vývoj cien alebo sledovanie priemyselnej produkcie.
Ako každý nový prístup, aj využitie Big Data za účelom štatistiky prináša riziká a problémy, ktoré je potrebné adresovať. Využívanie Big Data v oficiálnej štatistike prináša riziká spojené s nesprávnym použitím metodológií a komplikovanou interpretáciou. Nepresnosť údajov sa často kompenzuje ich objemom, čo môže ovplyvniť presnosť oficiálnych štatistík. Riziká zahŕňajú aj porušenie súkromia, právne nejasnosti a náklady na spracovanie. Dôležité je udržať dôveryhodnosť inštitúcie a transparentne aplikovať pravidlá na získavanie a využívanie Big Data.

Pripraviť zavedenie systematického využívania Big Data znamená zásadným spôsobom prebudovať myslenie a personálne obsadenie aspoň v kľúčových útvaroch štatistických úradov. Pre Štatistický úrad SR to znamená vytvoriť jednotku pre inovácie a Big Data, zabezpečiť expertov a koordináciu projektov a posilniť tak kapacitu odborných tímov pre Big Data. Potom bude možné zaviesť mechanizmus na výber prípadov použitia a hodnotenie kvality Big Data štatistík s cestovným plánom tém. Dôležité je aj realizovať program vzdelávania zamestnancov v oblasti Big Data, a zabezpečiť pravidelné školenia a efektívne komunikovať projekty a výsledky Big Data.

RESUME

Currently, the use of Big Data as a new source to complement official statistics is seen as a highly interesting opportunity. Big Data enables the production of statistics using methods for processing vast amounts of data generated as a by-product in a digitized society and economy. These data complement traditional sources of official statistics, such as survey data and administrative sources. In the realm of statistics, Big Data represent external information from digital activities, including smartphone usage, social media activities, digital transactions, or sensor-generated data.

Processing Big Data allows the creation of new statistical products and the automation of demanding manual tasks like categorization or completing the missing values. However, the use of Big Data in official statistics poses challenges, including regulatory frameworks, privacy, ethical and trust issues, partnership models, and the costs associated with providing access to data sources.

This article primarily focuses on an overview of different categories of Big Data and describes their properties and characteristics. Significant data categories include data without dependencies, such as categorical, quantitative, and textual data, and data with dependencies, such as time series, networks, and spatial data. Big Data can be widely used in fields such as tourism, demography, price development, or monitoring industrial production.

Similarly as any new approach, the use of Big Data for statistical purposes brings risks and issues that need to be addressed. Leveraging Big Data in official statistics entails risks associated with the misue of methodologies and complicated interpretation. Data inaccuracies are often compensated for by their volume, but this can impact the accuracy of official statistics. Risks also include privacy breaches, legal uncertainties, and processing costs. It is crucial to maintain the institution's credibility and transparently apply rules for obtaining and utilizing Big Data.

Introducing the systematic use of Big Data signifies a fundamental change in the way of thinking and in staffing within the key departments of statistical offices. For the Statistical Office of the Slovak Republic, this entails establishing an innovation unit for Big Data, ensuring experts and project coordination, thus reinforcing the capacity of expert teams for Big Data. This approach enables the implementation of a mechanism for selecting the cases of use and assessing the quality of Big Data statistics, with a roadmap of topics. It is also crucial to implement an employee education program in the field of Big Data, ensure regular training, and effectively communicate the projects and the results of Big Data.

PROFESIJNÝ ŽIVOTOPIS

Dipl. Ing. Dagmar Ceľuchová Bošanská *je zakladateľkou spoločnosti Alistiq s. r. o. a expertkou na inovácie a digitálnu transformáciu s dlhoročnými skúsenosťami. V roku 2008 absolvovala inžinierske štúdium pre informačné technológie, mobilné komunikácie a štatistické spracovanie signálov na Viedenskej technickej univerzite, kde pôsobila vo vedeckom tíme*

na vývoji simulátorov technológií pre bezdrôtové siete štvrtej generácie. Od roku 2015 sa venuje vývoju riešenia a návrhu opatrení na zvyšovanie kvality a efektivity využívania údajov vrátane Big Data na sekundárne účely, predovšetkým vo verejnej správe. Aktuálne od roku 2020 pôsobí ako doktorand na Českom vysokom učení technickom v Prahe, kde sa venuje výskumu grafových údajov generovaných z elektronických zdravotných záznamov a ich analýze s využitím strojového učenia a veľkých jazykových modelov.
***Ing. Juraj Bárdy** absolvoval magisterské štúdium na Fakulte riadenia a informatiky Žilinskej univerzity v odbore informačné a riadiace systémy (2006). Venuje sa inováciám verejných služieb a politík a digitálnej transformácii vo verejnej správe, so zameraním na využitie strojového učenia a lepší manažment údajov. Podieľal sa na návrhu Národnej koncepcie verejnej správy (2016) a príprave Stratégie digitálnej transformácie SR (2019). Je partnerom v konzultačnej spoločnosti Alistiq s. r. o.*

KONTAKT
dagmar.bosanska@alistiq.com
juraj.bardy@alistiq.com

Informatívny článok/Informative article

Martin ŠVEDA
Prírodovedecká fakulta Univerzity Komenského, Geografický ústav SAV v. v. i.

MOBILNÁ SIEŤ AKO PERSPEKTÍVNY ZDROJ INFORMÁCIÍ O PRIESTOROVOM ROZMIESTNENÍ A MOBILITE POPULÁCIE

MOBILE NETWORK AS A PROSPECTIVE SOURCE OF INFORMATION ON THE SPATIAL DISTRIBUTION AND MOBILITY OF THE POPULATION

ABSTRAKT
Mobilné zariadenia sa stali všadeprítomnou a neoddeliteľnou súčasťou nášho každodenného života. Ich prostredníctvom zanechávame čoraz presnejšie a rozmanitejšie digitálne stopy. Príspevok predstavuje základné vlastnosti údajov z mobilnej siete v kontexte ich spracovania pre potreby sledovania rozmiestnenia a mobility populácie. Principiálne môžeme údaje z mobilnej siete rozdeliť do dvoch skupín, na pasívne a aktívne. Práve prvé menované majú značný potenciál pre exploratívny výskum v oblasti časovo-priestorového správania populácie, keďže pri použití vhodného spracovania dokážeme z miliónov záznamov v mobilnej sieti vyťažiť významnú informáciu o pravidelných lokalizáciách značnej časti populácie.

ABSTRACT
Mobile devices have become a ubiquitous and integral part of our daily lives. Through them, we leave increasingly precise and diverse digital footprints. This paper presents the fundamental characteristics of mobile network data during their processing for the purpose of analyzing population distribution and mobility. As a matter of principle, mobile network data can be categorized into two groups: passive and active. The former has significant potential for exploratory research in the field of spatio-temporal behavior of the population, as by means of appropriate processing we can extract meaningful information about the regular localizations of a substantial part of the population from millions of records in the mobile network.

KĽÚČOVÉ SLOVÁ
mobilná sieť, lokalizácia, odhad populácie

KEY WORDS
mobile network, localization, population estimates

1. ÚVOD
Súčasná spoločnosť je viac ako kedykoľvek predtým tvorená tokmi ľudí, tovarov a informácií, avšak dáta, ktoré tieto toky zaznamenávajú, sú pre výskumníkov a výskumníčky často nedostupné. Príkladom sú údaje z mobilnej siete, ktoré majú jednu dôležitú predispozíciu: umožňujú analyzovať časovo-priestorové trajektórie prakticky na individuálnej úrovni a vo vysokom časovom a priestorovom detaile. Význam intrapersonálnej perspektívy pri štúdiu správania ľudí v čase a priestore sa v posledných rokoch zvýšil v dôsledku globalizácie, individualizácie a rozvoja mobilných technológií. Tieto zmeny zvýšili komplexnosť individuálnych časovo-

priestorových trajektórií [11, 16, 20]. Vo vyspelých krajinách už cesty súvisiace s prácou predstavujú približne pätinu všetkých ciest a približne štvrtinu celkovej vzdialenosti, ktorú ľudia precestujú, zatiaľ čo najväčší podiel ciest súvisí s voľným časom [28]. Ak chceme porozumieť fenoménom postmoderného sveta, potrebujeme porozumieť tokom, ktoré tento svet vytvárajú [4]. Je zrejmé, že konvenčné zdroje údajov nám ponúkajú len obmedzený pohľad a naše poznatky o časovom a priestorovom rozmiestnení obyvateľstva na subregionálnej úrovni sú stále veľmi obmedzené [26]. Vo veľkej miere sa spoliehame na desaťročné sčítanie obyvateľstva, ako aj na prieskumy, ktoré sa sporadicky vykonávajú v medzicenzovom období. Rozmiestnenie obyvateľstva je však veľmi dynamické so zásadnými zmenami počas dňa, ale aj v sezónnych cykloch. Tento nedostatok primeraného časového a geografického rozlíšenia konvenčných údajov môže viesť k nežiadúcim skresleniam, podhodnoteniam javov, alebo k neefektívnemu riadeniu [25]. V dôsledku toho sú potrebné nové prístupy k mapovaniu obyvateľstva, ktoré by dokázali rozšíriť naše poznanie o rozmiestnení a mobilite populácie.

Pre výskumníkov a výskumníčky v oblasti priestorových analýz je mobilná sieť zaujímavým zdrojom údajov predovšetkým vďaka možnosti poskytovania údajov o priestorovej a časovej lokalizácii veľkého množstva používateľov. Pre pochopenie toho, ako tieto údaje vznikajú a ako sa dajú analyticky využiť sa v príspevku oboznámime so základnými princípmi prevádzky mobilnej siete a možnosťami lokalizácie mobilných zariadení.

2. MOBILNÁ SIEŤ

Mobilná celulárna (bunková) sieť (ďalej len „mobilná sieť“) je rádiová telekomunikačná sieť, ktorá poskytuje bezdrôtovú konektivitu na rozsiahlom území a je zabezpečovaná veľkým počtom základňových staníc (antén), ktoré poskytujú pokrytie signálom komunikačným zariadeniam, najčastejšie mobilným telefónom (ale aj tabletom, autám či iným prístrojom). Za uplynulé tri desaťročia sa technológia mobilnej komunikácie progresívne rozvíjala podľa rôznych medzinárodných štandardov, ktoré neboli vždy kompatibilné vo všetkých regiónoch sveta. Prvé generácie mobilných sietí vznikli v 80. rokoch v rámci národných systémov (Japonsko, USA) s obmedzenou medzinárodnou kompatibilitou. Skutočne celosvetový rozvoj zaznamenala mobilná komunikácia až v 90. rokoch so zavedením štandardu GSM *(Global System for Mobile Communications*) vyvinutým Európskym inštitútom pre telekomunikačné normy (ETSI). GSM tak predstavoval už druhú generáciu (2G) mobilných sietí a od svojich predchodcov sa výrazne líšil v tom, že obidva kanály, signalizačný aj hlasový, sú digitálne. Univerzálnosť tohto technologického štandardu prispela k jeho masívnemu rozšíreniu do celého sveta. GSM nasledoval ďalší celosvetový štandard UMTS (*Universal Mobile Telecommunications System*), ktorý predstavoval už tretiu generáciu (3G) a umožňoval využívanie mobilnej siete na multimediálne služby (internet, video-stream). Štvrtá generácia (LTE – *Long Term Evolution*) bola v Európe zavedená od roku 2011 s cieľom splniť požiadavky nových konceptov komunikačných sietí, vrátane internetu vecí (*internet of things*), inteligentného riadenia miest (*smart cities*) či dopravných sietí. V súčasnosti sme svedkami postupného zavádzania už piatej (5G) generácie mobilnej siete, ktorá reaguje na čoraz väčší dopyt po rýchlom prenose veľkých objemov dát.

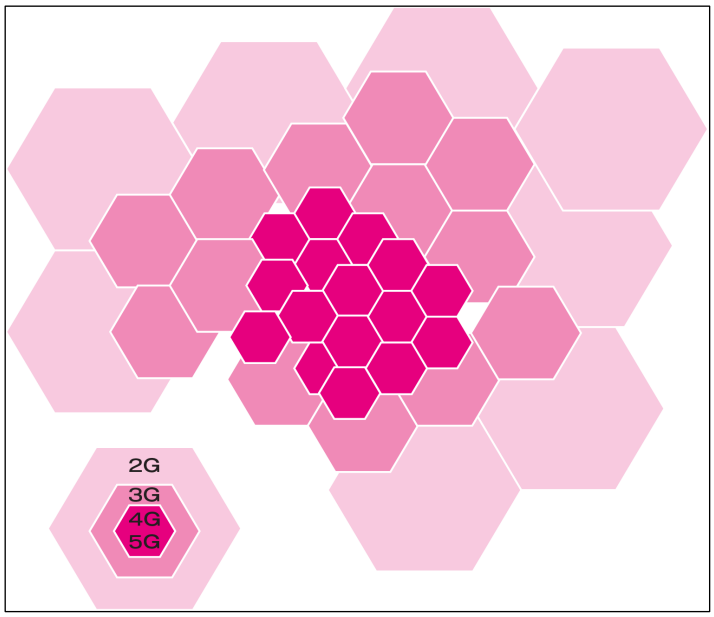
Mobilná sieť je štruktúrovaná prostredníctvom základňových staníc (antén) do jednotlivých buniek (odtiaľ *cellular network*, *cell-phone* a pod.), ktoré sú v priestore

rozmiestnené na základe koncentrácie používateľov (prevažne dennej koncentrácie) a s ohľadom na členitosť reliéfu a infraštruktúru (najmä nadradenú cestnú sieť alebo železnice). Základňová stanica, angl. *Base Transceiver Station* (BTS), je teda základnou konštrukčnou jednotkou mobilnej siete a na území Slovenska sa ich nachádzajú tisíce.

Územie pokrytia signálom reprezentuje vyžarovací polygón, ktorý predstavuje mnohostranný plošný útvar opisujúci pokrytie signálom práve jednej BTS bunky. Vyžarovací polygón spravidla nevytvára spojitú a kompaktnú plochu, ktorú by sme mohli definovať prostredníctvom jednoduchej geometrie. V realite ide o zložitý polygón (prípadne multipolygón), ktorého podobu ovplyvňuje výkon vysielača, smerovanie, druh použitej technológie (frekvencie) a prekážky obmedzujúce šírenie signálu. V praxi ide najmä o reliéf (pohoria, údolia) a charakter zástavby.

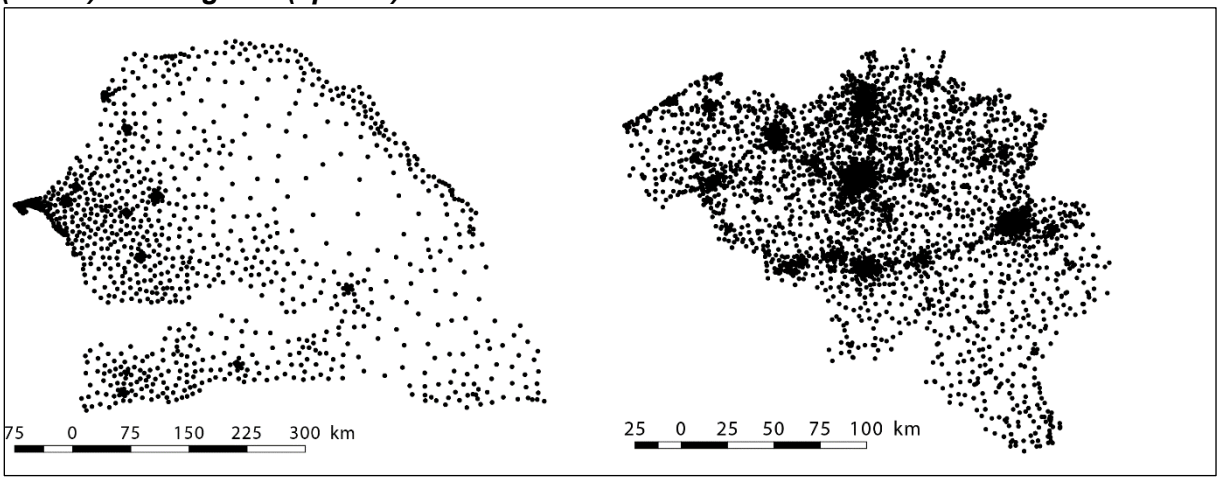
Každá BTS dokáže obslúžiť obmedzený počet používateľov mobilnej siete. Z dôvodu bezproblémovej prevádzky sa bunky mobilnej siete prekrývajú, aby v prípade potreby dokázali poskytnúť kapacitu väčšiemu počtu používateľov. Kapacita úzko súvisí s technológiou, ktorá sa na danej BTS používa. Jednotlivé technológie sa označujú zjednodušene ako 2G, 3G, 4G, 5G a pod. Písmeno G označuje pojem „generácia“. Mobilnú sieť tak tvorí viacero sietí rôznych generácií. Vo všeobecnosti pritom môžeme povedať, že vysoko urbanizované časti územia sú pokryté prevažne menšími bunkami novších generácií mobilnej siete (4G/5G), pričom s narastajúcou vzdialenosťou od miest sa hustota buniek mobilnej siete znižuje spolu s narastajúcou veľkosťou ich vyžarovacích polygónov (obrázok č. 1). Priestorová granularita mobilnej siete sa teda pohybuje od desiatok metrov (v nákupných centrách, letiskách a pod.), cez stovky metrov (v mestských oblastiach) až po desiatky kilometrov (na vidieku), v závislosti predovšetkým od hustoty zaľudnenia, ako aj od intenzity dopravy. Na architektúru mobilnej siete však vplýva aj ekonomická vyspelosť krajiny a rôzne regionálne či národné špecifiká (obrázok č. 2).

Obrázok č. 1: Schéma pokrytia územia rôznymi generáciami mobilnej siete v mestskom regióne



Zdroj: vlastné spracovanie autora

Obrázok č. 2: Priestorové rozmiestnenie základňových staníc mobilnej siete v Senegale (vľavo) a v Belgicku (vpravo)



Zdroj: [17]

3. PRIESTOROVÁ LOKALIZÁCIA S VYUŽITÍM ÚDAJOV MOBILNEJ SIETE

Na základnej úrovni môžeme rozlíšiť dva spôsoby priestorovej lokalizácie s využitím mobilnej siete: 1) *sieťová lokalizácia* je výsledkom základných technických daností prevádzky mobilnej siete a je založená na granularite infraštruktúry (veľkosť buniek vyžarovacích polygónov) a potreby zbierania údajov o aktivite mobilného zariadenia. Lokalizácia mobilného telefónu vzniká ako „vedľajší produkt" fungovania mobilnej komunikácie, keďže údaje sú generované primárne pre potreby prevádzky mobilnej siete; 2) *lokalizácia prostredníctvom služieb priestorovej lokalizácie* je založená na funkcionalite mobilných aplikácií, ktoré používajú polohu mobilného telefónu (napr. prostredníctvom lokalizácie *Assisted-GPS*). Okrem zabezpečenia funkcionality danej aplikácie (napr. navigácia) sa tieto údaje využívajú aj na marketingové účely (ponuka služieb zohľadňujúca polohu používateľa).

Pri druhom spôsobe je informácia o polohe explicitnou súčasťou údajov a ich využiteľnosť na geografické analýzy je tak prirodzene priamočiara. Pri využití *sieťovej lokalizácie* je extrahovanie polohy používateľa mobilnej siete náročnejšie a prináša viaceré metodické a konceptuálne výzvy, na ktoré doposiaľ odborný diskurz nepriniesol jednoznačné riešenia. V ďalšej časti sa budeme venovať výlučne využitiu údajov v sieťovej lokalizácii, keďže ich spracovanie pre potreby priestorových analýz vytvára mnohé technické a metodické výzvy, no zároveň prináša potenciálne veľmi vysoký prínos pre spoločensko-vedný rozvoj.

Princíp sieťovej lokalizácie v mobilnej sieti je prirodzenou nadstavbou základných vlastností mobilnej komunikácie. Z priestorového pohľadu môžeme územie rozdeliť do buniek, ktoré obsluhujú jednotlivé antény mobilnej siete. Každá anténa je schopná pokryť určité územie a obslúžiť určitý počet zákazníkov. Identifikačné údaje o aktuálne využívanej anténe, ako aj ďalšie doplnkové informácie, môžu byť využité pre určenie približnej polohy používateľa mobilného zariadenia. Tieto lokalizačné údaje sú prirodzenou súčasťou mobilnej komunikácie, keďže identifikácia základňových staníc (BTS), s ktorými komunikuje mobilný telefón, je nevyhnutná pre samotné fungovanie mobilnej siete. Lokalizačné údaje, ktoré môžeme získať prostredníctvom sieťovej lokalizácie, principiálne rozdeľujeme na aktívne a pasívne [2]. Kým pri pasívnom type ide o využitie existujúcich záznamov v systéme mobilného operátora, pri aktívnom type

sa záznam vytvára na základe konkrétneho dopytu a s využitím špecializovaného softvéru.

Pasívne lokalizačné údaje sú digitálne „stopy“, ktoré zanecháva mobilné zariadenie na infraštruktúre mobilnej siete. Tieto záznamy vznikajú buď pre potreby vyúčtovania hovorov, SMS a pod., alebo sú dôsledkom pravidelných aktualizácií polohy zariadenia v mobilnej sieti. Vzhľadom na skutočnosť, že k údajom o aktivite mobilného zariadenia vieme priradiť nielen približnú polohu (konkrétnu bunku mobilnej siete), ale aj niektoré základné informácie o používateľovi (vek, pohlavie či fakturačná adresa), získavame bohatú databázu, ktorej využitie (pri zachovaní anonymity používateľov, resp. pri čiastočnom agregovaní údajov) prináša nesmierne cenný zdroj údajov. Príkladom môže byť využitie pravidelnej dennej a nočnej lokalizácie používateľov, ktoré vytvárajú základnú kostru ich každodenných aktivít [3]. Údaje o koncentrácii dennej a nočnej lokalizácie nám umožňujú nielen spresniť priestorovú distribúciu obyvateľstva (napr. koľko ľudí býva v danom meste/v blízkosti nákupného centra a pod.), ale aj extrahovať údaje o predpokladanej priestorovej mobilite (napr. koľko ľudí dochádza do mesta/do nákupného centra). Pasívne lokalizačné údaje sa tak stávajú cenným zdrojom informácií pre dopravné modely [19], urbánne plánovanie [5, 29] či v manažmente krízových udalostí [15].

Principiálne rozoznávame dva typy pasívnych lokalizačných údajov z mobilnej siete: 1) *signalizačné dáta* slúžia na zabezpečenie prevádzky mobilnej siete; 2) záznamy zachytávajúce aktivitu mobilného zariadenia (*call-detail-records*) slúžia primárne na vyúčtovanie služieb používateľovi mobilnej siete.

Signalizačné dáta (*signalling data*, *ping data*) sú automaticky generované záznamy, ktoré produkuje mobilná sieť pri pravidelných kontrolách pripojených zariadení. Jednotlivé BTS stanice v pravidelných intervaloch (definovaných oblasťou/operátorom) vysielajú signál na všetky dostupné a prihlásené zariadenia. Na základe ich polohy a vyťaženosti siete sa určí, ktorá BTS stanica bude konkrétnemu zariadeniu poskytovať signál, čím sa zabezpečuje čo najlepšie pokrytie pre každé zariadenie v dosahu. Prednosť má pritom pripojenie k anténe s novšou technológiou prenosu pred intenzitou signálu. Tieto dáta vznikajú bez interakcie používateľa či používateľky so zariadením. SIM karta vygeneruje záznam pri každej zmene BTS stanice, čím sa vygenerujú desiatky až stovky záznamov denne. Vzhľadom na veľký počet takýchto záznamov a rozličné technické špecifiká je ich spracovanie pomerne náročné. Keďže sa tieto údaje priebežne prepisujú, je ich zber potrebné realizovať v reálnom čase a na pozadí sieťových operácií [26]. Ich praktické využitie v analýzach bolo dosiaľ obmedzené, hoci sa už objavili prvé pionierske počiny [31, 32]. Môžeme však predpokladať, že ich význam v priestorových analýzach bude narastať. Dôvodom sú zmeny vo využívaní mobilného telefónu. Kým v minulosti sme ho využívali najmä na volania a SMS správy, v súčasnosti toto využitie ustupuje v prospech dátových služieb (sociálne siete, okamžitá komunikácia a pod.) a pasívnemu využívaniu telefónu (napr. rôzne notifikácie v aplikáciách si nevyžadujú aktívnu interakciu).

CDR dáta (*call-detail-records*) sú záznamy, ktoré vznikajú pri akejkoľvek aktivite mobilného telefónu (SIM karty) s mobilnou sieťou. Za takéto aktivity považujeme uskutočnený/prijatý (neprijatý) hovor, doručenie/zaslanie SMS či dátový prenos. CDR záznam teda nevzniká automaticky, ale výlučne pri aktivite telefónu (SIM karty).

Takýchto záznamov preto existuje principiálne menej než pri signalizačných dátach. Počas bežného dňa SIM karta vygeneruje rádovo desiatky CDR záznamov. Vďaka relatívne jednoduchému spracovaniu (anonymizácia) a dostupnosti (vznikajú pre potreby vyúčtovania používateľov) patria tieto údaje k najčastejším zdrojom v rozmanitých priestorových analýzach [26]). Literatúra poskytuje viacero príkladov úspešnej aplikácie CDR záznamov. Napr. Deville a kol. [10] preukázali, že údaje z mobilnej siete môžu poskytovať presné a finančne dostupné mapy rozmiestnenia populácie na národnej úrovni. Podobne Csáji a kol. [9] dokumentovali silnú koreláciu medzi údajmi z cenzu a CDR dátami na regionálnej úrovni. Rozsiahly prehľad literatúry k využitiu CDR údajov ponúkajú Blondel a kol. [6].

Kým pri pasívnej lokalizácii vychádzame z databázovej infraštruktúry prevádzkovateľa mobilnej siete (ide o anonymizované či agregované dáta), pri aktívnej lokalizácii vznikajú záznamy o polohe individuálneho mobilného zariadenia na základe cieleného lokalizačného dopytu a s využitím špecializovaného softvéru. Nevyhnutnou podmienkou záznamu je informovaný súhlas používateľa mobilného zariadenia a presne stanovené podmienky charakteru a dĺžky záznamu. Právne a etické zásady spojené s aktívnou lokalizáciou diskutuje Ahas a kol. [2], Dufková a kol. [12] či Novák [24], ktorí priniesli aj inšpiratívne pilotné sondy využívajúce aktívnu mobilnú lokalizáciu. Tento spôsob prekonáva základný nedostatok prevádzkových údajov mobilných sietí – záznamov o aktivite mobilného zariadenia, ktorých časové rozostupy (frekvencia záznamov) neumožňujú presné určenie polohy používateľa počas dňa. Pomocou aktívnej lokalizácie je možné doplniť polohu používateľa aj o konkrétne aktivity, ktoré v lokalite vykonáva. Vznikajú tak nesmierne cenné údaje, ktoré do veľkej miery môžu nahradiť tradičné zdroje údajov (dotazníky, časovo-priestorové denníky) na behaviorálny výskum či rôzne prístupy v rámci geografie času (podrobnejšie [30]). Je však potrebné zdôrazniť, že spracovanie údajov z aktívnej lokalizácie si vyžaduje špecifické výpočtové (hardvérové) kapacity, ako aj finančné náklady. V súčasnosti je aktívna lokalizácia skôr marginalizovaným spôsobom využitia mobilnej infraštruktúry. Príkladom je štúdia realizovaná v metropolitnom území mesta Boston na vzorke 1 mil. používateliek a používateľov mobilnej siete [7]. Pri každej aktivite mobilného zariadenia bola na základe triangulácie odhadnutá poloha používateľa v bunke mobilnej siete. Tým sa dosiahla vyššia priestorová presnosť, vďaka čomu bolo možné sledovať mobilitu obyvateľov v porovnateľnom rozlíšení, aké umožňujú tradičné metódy zberu údajov (napr. prostredníctvom cenzu).

4. URČENIE POLOHY MOBILNÉHO ZARIADENIA S VYUŽITÍM SIEŤOVEJ LOKALIZÁCIE

Na využití údajov z mobilnej siete výskumníkov a výskumníčky asi najviac priťahuje práve ich schopnosť poskytnúť údaje o (približnej) lokalizácii používateľov. Tento atribút má ďalekosiahle možnosti využitia, je však potrebné si uvedomiť, že pokiaľ nepoužívame údaje z aplikácií v samotnom mobilnom zariadení (využívajúce GPS lokalizáciu), záznam o polohe nie je štandardnou súčasťou surových údajov z mobilnej siete a môžeme sa k nemu dopracovať len prostredníctvom aproximácie.

Určenie polohy mobilného zariadenia v sieťovej lokalizácii prebieha takto. Zariadenie s aktívnou SIM kartou sa pripája na BTS stanicu. Zvyčajne ide o najbližší vysielač, nemusí to však byť pravidlom. Niektoré BTS stanice totiž slúžia na prenos dát (napr. 4G bunky), iné na prenos hovorov (2G/3G). Sieť zároveň autonómne distribuuje signál všetkým pripojeným zariadeniam podľa svojho vyťaženia. Ak je teda

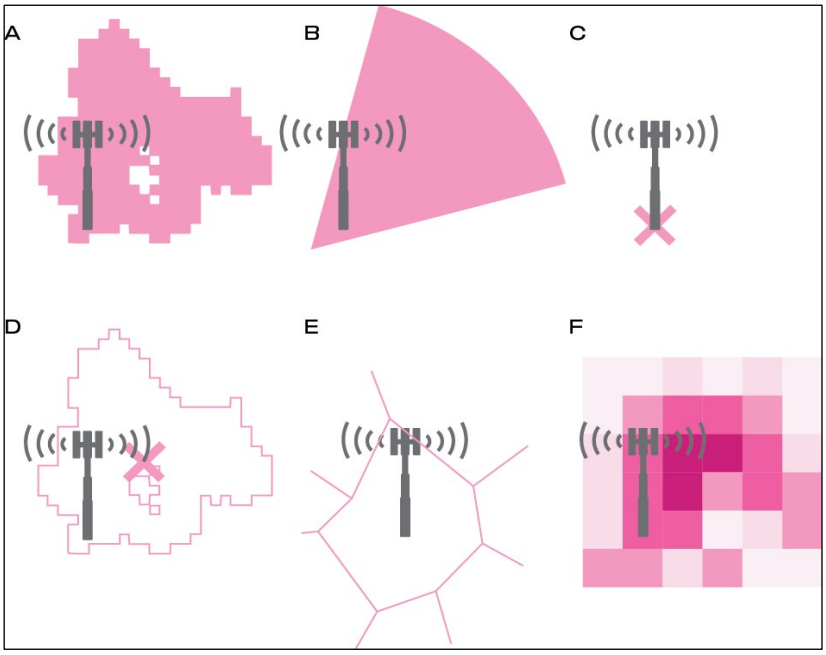
niektorá bunka preťažená, pošle časť zariadení na inú bunku. Práve organické zmeny buniek mobilnej siete (tzv. *handover*) môžu vytvárať zdanlivý pohyb SIM karty, hoci jej poloha je statická. Pri časových agregátoch tento problém zaniká, keďže sa pre danú časovú jednotku (napr. hodinu) vyberá bunka s najväčším počtom záznamov, resp. bunka, kde mobilné zariadenie strávilo najdlhší čas.

Každá bunka distribuuje signál rôznej intenzity a rôznymi smermi. Plochu, ktorú ožaruje nazývame vyžarovací polygón *(cell coverage area, service area*) a vždy ide o nepravidelný útvar, ktorého tvar je výsledkom smerovania a intenzity signálu, ako aj charakteru prírodných podmienok (najmä reliéfu, zástavby, vegetácie a iných prekážok). Z uvedených vlastností základných stavebných prvkov mobilnej siete vyplýva, že tvar a plocha záberu vyžarovacieho polygónu sa môže meniť v rôznych podmienkach. Ako teda pracovať s priestorovým ukotvením vyžarovacieho polygónu? V praxi máme niekoľko možností (obrázok č. 3):

a) Použitie plochy vyžarovacieho polygónu. – Mobilný operátor pozná predpokladanú oblasť pokrytia každej bunky mobilnej siete. Len zriedka je však možné pracovať s presným tvarom vyžarovacieho polygónu, keďže v konkurenčnom prostredí viacerých mobilných operátorov ide spravidla o citlivú informáciu.
b) Zjednodušený tvar vyžarovacieho polygónu. – Ďalšou možnosťou je odhadnúť oblasť pokrytia bunky mobilnej siete prostredníctvom základných parametrov konfigurácie antény, ako je výška antény, výkon a uhol pokrytia. Ani tento typ geo-referencovaných údajov však nebýva bežne k dispozícii. Ako však preukázali Ricciato a kol. [26], vďaka podrobnejším informáciám o topológii mobilnej siete (použitie približných oblastí pokrytia namiesto jednoduchej polohy BTS) môžeme výrazne zvýšiť presnosť modelov založených na CDR údajoch.
c) Poloha BTS. – Častejšou možnosťou je ignorovanie tvaru vyžarovacieho polygónu a využitie len polohy základňovej stanice. Tá je presne daná a v čase stabilná. Nevýhodou však je, že poloha antény sa spravidla nenachádza v priestorovom ťažisku vyžarovacieho polygónu. Dôvodom je skutočnosť, že výkon vyžarovacieho polygónu je často smerovaný jedným smerom, takže výsledný tvar má skôr charakter kruhového výseku, než kruhu či elipsy.
d) Ťažisko vyžarovacieho polygónu. – Presnejšiu informáciu o polohe mobilného zariadenia nám môže poskytnúť ťažisko *(centre of gravity*) vyžarovacieho polygónu. Pri zjednodušení polygónu do bodu síce strácame časť informácie, no takýto prístup je optimálny nielen z hľadiska práce s citlivými polohovými údajmi, ale aj z dôvodu jednoduchšieho spracovania a výpočtov.
e) Voroniove polygóny. – Často používaným riešením je zjednodušenie buniek mobilnej siete prostredníctvom *Voroniových* (*Thiessenových*) polygónov. V tejto metóde sa z bodových údajov polohy BTS vytvorí sieť polygónov, ktoré pravidelne pokrývajú územie. Nevýhodou je skutočnosť, že tvar výsledných polygónov sa môže zásadne líšiť od skutočného tvaru vyžarovacích polygónov, čo môže mať zásadný vplyv na presnosť lokalizácie mobilného zariadenia. Vo výsledku tak môžu byť odhady prítomného obyvateľstva (používateľov mobilných zariadení) značne pod- alebo nad-hodnotené [21].
f) Pravdepodobnostný model pokrytia. – Poloha mobilného zariadenia v mobilnej sieti môže byť vyjadrená aj prostredníctvom pravdepodobnostného modelu, ktorý zohľadňuje vzdialenosť od jednotlivých BTS staníc (čím bližšie, tým je

väčšia pravdepodobnosť lokalizácie), ako aj prekryv jednotlivých buniek mobilnej siete.

Obrázok č. 3: Ilustrácia tvaru vyžarovacieho polygónu a rôzne možnosti jeho generalizácie



Poznámka: a) skutočný tvar vyžarovacieho polygónu, b) zjednodušenie vyžarovacieho polygónu do kruhového výseku, c) redukcia vyžarovacieho polygónu do bodu lokalizácie základňovej stanice (BTS), d) redukcia vyžarovacieho polygónu do jeho priestorového ťažiska, e) zjednodušenie buniek mobilnej siete prostredníctvom Voroniových (Thiessenových) polygónov, f) nahradenie tvaru vyžarovacích polygónov pravdepodobnostným modelom.

Zdroj: vlastné spracovanie autora

Poloha mobilného zariadenia (používateľa mobilnej siete) je limitovaná veľkosťou vyžarovacieho polygónu. Vieme teda, že ak je daná SIM karta pripojená ku konkrétnej BTS bunke, nachádza sa v danom vyžarovacom polygóne. Keďže intenzita mobilného signálu klesá so štvorcom vzdialenosti, môžeme predpokladať, že mobilné zariadenie bude pripojené k najbližšej BTS. Vzhľadom na veľkosť buniek mobilnej siete (stovky metrov až kilometre) je takáto presnosť lokalizácie často nedostatočná, najmä pri štúdiách v mikromierke lokalít. Pre väčšinu aplikácií pracujúcich v regionálnej mierke (napr. na úrovni miest, okresov či krajov) nie sú uvedené limity prekážkou.

Ako sme uviedli, jednotlivé bunky majú rôzny dosah v závislosti od použitej technológie mobilnej siete. Polygóny 2G buniek majú dosah v kilometroch, polygóny 4G buniek už len rádovo stovky metrov. Pri snahe o dosiahnutie vyššej presnosti priestorovej lokalizácie je možné z datasetu vylúčiť práve 2G bunky, ktoré zvyčajne slúžia iba na prenos hovorov. Takýto prístup je vhodný napríklad v urbanizovaných územiach s vysokým počtom BTS buniek novších generácií. Pri analýze vidieckych oblastí by však vyradenie 2G buniek mohlo zásadne redukovať objem záznamov, keďže vidiecka a prírodná krajina je pokrytá oveľa nižším počtom plošne väčších buniek. Riešením je rozdelenie územia do pravidelnej siete, kde je pre každú jednotku definovaný podiel pokrytia jednotlivých typov BTS staníc. Na jeho základe sa následne alokujú údaje o používateľoch a používateľkách mobilnej siete do uvedenej pravidelnej siete. Takýto postup prináša jednoduchý priestorový systém záznamov, ktoré môžeme agregovať do ľubovoľných väčších priestorových jednotiek. Popri skladobnosti je

výhodou aj prispôsobenie týchto záznamov iným zdrojom priestorových údajov (napr. údaje zo sčítania obyvateľstva agregované do priestorovej mriežky s veľkosťou 1 x 1 km). Rizikom tohto prístupu je použitie nevhodne zvolenej metódy interpolácie bodových (prípadne plošných) údajov z mobilnej siete do cieľových priestorových jednotiek. Nesprávne zvolená metóda interpolácie môže priniesť nežiaduce skreslenia, najmä v prípade menej urbanizovaného územia (podrobnejšie [1, 18]).

Poznanie distribúcie mobilných zariadení (používateľov mobilnej siete) v priestore a čase je základnou schopnosťou lokalizačných údajov sieťovej lokalizácie. Ďalším krokom spracovania týchto údajov je extrahovanie periodicky navštívených lokalít používateľov mobilnej siete a následná konštrukcia tokov medzi nimi.

5. KLASIFIKÁCIA POUŽÍVATEĽOV MOBILNÝCH ZARIADENÍ NA ZÁKLADE PRAVIDELNOSTI POBYTU V MOBILNEJ SIETI

Základnou informáciou, ktorú poskytuje lokalizácia v mobilnej sieti, je sledovanie statickej lokalizácie (pobytu) používateľov v priestore. Principiálne môžeme rozlíšiť dva typy pobytu:

2) Jednorazový pobyt mobilného zariadenia v lokalite je ohraničený časovým intervalom vstupu do danej bunky mobilnej siete (vytvorením záznamu) a výstupu (vytvorením záznamu v inej bunke mobilnej siete). Dĺžka pobytu v bunke mobilnej siete pritom nemusí zodpovedať skutočnej dĺžke pobytu používateľa.
3) Pravidelný (opakovaný) pobyt mobilného zariadenia v rámci lokality. Sledovanie opakovaných lokalizácií mobilného zariadenia v jednej bunke mobilnej siete (prípadne v lokalite vytvárajúcich viacero buniek) umožňuje identifikovať významné lokalizácie z hľadiska periodickej mobility populácie a filtrovať nerutinné (nepravidelné) lokalizácie v mobilnej sieti. Práve tento typ záznamu má osobitný význam pre spoločensko-vedné analýzy, keďže umožňuje identifikovať kotevné body a priradiť im predpokladaný význam z hľadiska každodenného života obyvateľov.

Pri identifikácii kotevných bodov z mobilnej siete môžeme rozlíšiť nasledujúce kategórie [3, upravené]:

- *Respondent:* Používateľ mobilného telefónu (anonymizovaný) vykonávajúci aktivity na mobilnej sieti.
- *Bunka pravidelnej lokalizácie:* Bunka mobilnej siete (vyžarovací polygón BTS), v ktorej sa respondent pripája na mobilnú sieť opakovane počas určitého obdobia (napr. 30 dní).
- *Bunka náhodnej lokalizácie:* Bunka mobilnej siete (vyžarovací polygón BTS), v ktorej mal respondent lokalizačný záznam iba raz za určité časové obdobie (napr. 30 dní).
- *Významné miesto:* Bunka pravidelnej lokalizácie, ktorej vieme priradiť určitý význam v každodennom živote respondenta.
- *Každodenný kotevný bod:* Bunka pravidelnej lokalizácie, v ktorej respondent vytvoril záznamy väčšinu dní pozorovania.
- *Kotevný bod domov/práca:* Každodenný kotevný bod, ktorý spĺňa zadané kritéria na určenie pravidelnej dennej/nočnej lokalizácie (napr. najväčší počet lokalizácií v čase od 9:00 do 15:00 hod.).
- *Multifunkčný kotevný bod:* Každodenný kotevný bod, ktorý spĺňa kritériá na identifikovanie kotevného bodu domov i kotevného bodu práca.

Uvedené kategórie sú len veľmi všeobecne definované a je potrebné ich spresniť vzhľadom na charakter údajov (CDR, signalizačné), priestorovú mierku a ciele analýzy. Často je potrebné uplatniť ďalšie kritériá, ktoré z datasetu vylúčia tých respondentov, ktorí majú malý alebo veľký počet záznamov (lokalizácií v bunkách mobilnej siete). V praxi môže ísť napr. o mobilné zariadenia vo firmách (call centrá, zákaznícka podpora a pod.) alebo krátkodobých návštevníkov daného územia.

Identifikovanie pravidelnej (dennej a nočnej) lokalizácie je súčasťou hľadania časovo-priestorových vzorcov pohybu používateľov mobilnej siete. Ide o širokú problematiku s množstvom prístupov (napr. v behaviorálnej geografii a geografii času) a aplikácií. Na rozdiel od tradičných nástrojov zberu údajov na zachytenie časovo-priestorových vzorcov pohybu obyvateľov (časovo-priestorové rozpisy – diáre, GPS trackovanie a pod.) lokalizácia s využitím mobilnej siete neposkytuje informácie o aktivitách, použitých dopravných prostriedkoch či sociálnom kontexte. Napriek tomu môžeme identifikovať niekoľko základných kategórií s predpokladanou interpretáciou.

Rezident: Jedno miesto pravidelnej lokalizácie v nočných hodinách môže identifikovať bydlisko používateľa mobilného zariadenia. Môže však ísť napríklad aj o pracovisko nočnej zmeny (najmä pri lokalizácii v priemyselných areáloch).

Pracujúci: Jedno miesto pravidelnej lokalizácie v denných hodinách môže identifikovať pracovisko používateľa mobilného zariadenia. Na rozdiel od rezidenčnej lokality miesto práce podlieha oveľa väčšej variabilite. Problémom môže byť už samotné určenie časti dňa, ktorá by mala špecifikovať pracovnú činnosť (práca na zmeny, na skrátený pracovný čas a pod.), nehovoriac o pracovných aktivitách, ktoré nie sú viazané na jedno miesto (poštové doručovateľky, vodiči, živnostníci a pod.).

Identifikácia rezidentov a pracujúcich je zvyčajne prvým a nevyhnutným krokom na určenie ďalších kategórií používateľov mobilnej siete:

Dochádzajúci do práce (školy): Pokiaľ sa denná lokalita nezhoduje s nočnou, môžeme daného používateľa identifikovať ako dochádzajúceho do práce alebo školy. Ak používateľ nie je v danom regióne identifikovaný ako rezident, môžeme uvažovať o interregionálnej alebo cezhraničnej dochádzke.

Odchádzajúci do práce (školy): Ak rezidentovi nevieme priradiť dennú lokalitu (ani jedna bunka mobilnej siete nemá dostatočný počet záznamov počas určenej časti dňa), je pravdepodobné, že svoje pracovné alebo vzdelávacie aktivity realizuje mimo sledovaného regiónu.

Tranzitujúci: Ak mobilnému zariadeniu nie je možné priradiť rezidenčnú či pracovnú lokalitu (nespĺňa stanovené podmienky, napr. počtu záznamov alebo dĺžky zotrvania v lokalite), no zároveň sa daná SIM karta viacnásobne objavuje v mobilnej sieti sledovaného územia, môžeme uvažovať o kategórii tranzitujúcich (napr. vodiči) či krátkodobých, no pravidelných, návštevníkov (napr. nakupujúci z iných regiónov).

Nezaradení používatelia mobilnej siete sú takí, ktorých mobilné zariadenie nesplnilo kritériá ani jednej z predchádzajúcich kategórií. V praxi tieto záznamy väčšinou nevstupujú do analýz, keďže ich relevantná interpretácia je prakticky nemožná.

Odstránenie tohto dátového „šumu“ je častou, a vo svojej podstate nevyhnutnou súčasťou práce s údajmi typu *big data*.

Okrem uvedených základných kategórií môžeme uvažovať o ďalších kategóriách, ktoré by zodpovedali očakávaným priestorovým vzorcom správania a analytickým zámerom. Príkladom je snaha o identifikovanie používateľov mobilnej siete dochádzajúcich za službami či inými aktivitami (napr. voľný čas). Pri tomto zámere je potrebné definovať nielen čas lokalizácie (napr. poobedné hodiny), ale aj bližšie špecifikovať očakávané charakteristiky lokality (nákupné centrum, lesopark). Je zrejmé, že lokalita dochádzky za službami by mala byť rozdielna od dennej (pracovnej) lokality.

Údaje pochádzajúce z mobilnej siete sú relatívne sémanticky chudobné a poskytujú len obmedzené informácie o používateľoch, ktoré však nemusia zodpovedať skutočnému používateľovi mobilného zariadenia. Väčšinu doplnkových informácií môžeme identifikovať len prostredníctvom časových a priestorových vzorcov správania a využitím doplnkových podkladov (napr. krajinná pokrývka, dopravná infraštruktúra, funkčná štruktúra).

Mobilní operátori v praxi disponujú aj ďalšími doplnkovými charakteristikami používateľov a používateliek mobilnej siete, ktoré umožňujú (v anonymizovanej a agregovanej podobe) štruktúrovať lokalizačné záznamy napríklad podľa fakturačnej adresy, pohlavia, veku, typu mobilného zariadenia či ďalších údajov, ktoré pochádzajú zo zmluvného vzťahu medzi používateľom a poskytovateľom telekomunikačných služieb (tabuľka č. 1). Napríklad spotreba mobilných dát alebo typ mobilného telefónu umožňuje zacieliť analýzu na špecifickú skupinu používateľov s očakávaným správaním v mobilnej sieti. Môžeme totiž predpokladať, že najnovšie mobilné zariadenia s vysokou spotrebou údajov bude využívať špecifická skupina používateľov (napr. mladí pracujúci s vyšším vzdelaním). Iným príkladom môže byť identifikácia sociálno-patologických javov prostredníctvom výberu tých používateľov mobilnej siete, ktorí neplatia načas faktúry a aktívne využívajú stávkovanie cez mobilný telefón.

Tabuľka č. 1: Štruktúra vybraných doplnkových údajov k lokalizačným údajom mobilnej siete

Údaje o používateľovi	vek, pohlavie, fakturačná adresa, typ užívateľa (privátny/biznis), štátna príslušnosť (na základe unikátneho čísla IMSI (*International Mobile Subscriber Identity)* obsahujúceho kód krajiny a kód mobilného operátora)
Využívanie mobilnej siete	typ mobilného zariadenia, spotreba údajov, typ využívanej služby (paušálu)
Využívanie rozmanitých služieb	využívanie tiesňových liniek, charitatívne príspevky, stávkovanie, hry, poistenie, parkovanie, verejná doprava a pod.
Finančné ukazovatele	forma platby za telekomunikačné služby, index bonity, index finančnej zodpovednosti
Cestovné návyky	počet dní strávených v zahraničí, navštívené krajiny (pripojenie cez roaming)

Zdroj: www.marketlocator.sk

6. POHYB V MOBILNEJ SIETI

Sieťová lokalizácia neumožňuje zaznamenávať reálny priestorový pohyb (trajektóriu) mobilného zariadenia. Nepriamo však môžeme pohyb odvodiť chronologickým pospájaním lokalizácií v jednotlivých bunkách mobilnej siete. Takto vytvorená trajektória však nemusí zodpovedať skutočnému pohybu používateľa. S využitím cestnej siete a modelovaním predpokladaných trajektórií sa môžeme viac priblížiť ku skutočnému pohybu používateľov, prípadne môžeme využiť informácie z reprezentatívnych prieskumov v aktívnej lokalizácii.

Pri sledovaní pohybu mobilného zariadenia je dôležité v čo najväčšej miere odlíšiť skutočné pohyby od fiktívnych zmien polohy. Nereálne pohyby vznikajú viacerými spôsobmi. Jedno miesto býva spravidla pokryté signálom viacerých základňových staníc mobilnej siete (BTS). Pri pravidelnej aktualizácii polohy mobilného zariadenia, alebo pri jeho aktivite, môže dôjsť k zmene BTS bez toho, aby používateľ vykonal pohyb v priestore. Dôvodom môže byť zmena aktivity (hlasový hovor, dátový tok), naplnenie kapacity danej BTS alebo prepnutie na vysielač so silnejším signálom. V blízkosti štátnych hraníc nastáva často prepojenie mobilného telefónu na sieť mobilného operátora vo vedľajšej krajine. Výsledkom môže byť viacnásobný záznam na infraštruktúre mobilnej siete (minimálne jeden záznam pri zmene BTS), hoci priestorová poloha mobilného zariadenia ostala nezmenená (napr. pri pobyte na pracovisku, v domácnosti). Eliminácia týchto fiktívnych pohybov používateľa v mobilnej sieti nie je vôbec triviálna úloha. Odporúčaným riešením je použitie dlhšieho pozorovacieho času, vďaka ktorému sa fiktívne pohyby „vyhladia“.

7. OCHRANA OSOBNÝCH ÚDAJOV

Pri spracovaní údajov z mobilnej siete má ochrana osobných údajov veľmi dôležité postavenie. Znalosť polohy individuálneho používateľa alebo identifikácia jeho pravidelne navštevovaných lokalít (domov, práca) predstavujú neakceptovateľný zásah do súkromia [3, 8]. Nejde však len o poznanie priestorovej lokalizácie. Analýzou smerovania hovorov možno skonštruovať sociálnu sieť jednotlivca, či dokonca analyzovať jeho osobnosť [23].

Na ochranu súkromia používateľov mobilnej siete sú údaje vždy anonymizované, t . j. všetky osobné údaje, ako je meno, adresa, telefónne číslo atď., sú buď odstránené z databázy, alebo nahradené náhodne generovaným číslom, aby sa predišlo identifikácii. Aj v prípade, ak sú individuálne údaje agregované, je potrebné venovať zvýšenú pozornosť vždy, keď sa tieto údaje analyzujú a vizualizujú mimo prostredia prevádzkovateľa mobilnej siete. Kľúčovým právnym rámcom v Európskej únii je nariadenie o ochrane osobných údajov (*General Data Protection Regulation* – GDPR). Hoci prácu s individuálnymi záznamami vylučujú právne normy, samotní prevádzkovatelia mobilnej komunikácie majú často vnútorné smernice nastavené ešte prísnejšie. Strata dôvery zákazníkov zvyčajne oveľa prevyšuje potenciálny profit vyplývajúci z monetizácie lokalizačných údajov.

Ochrana osobných údajov prináša pre výskumníkov a výskumníčky analyzujúcich údaje z mobilnej siete viacero výziev. Kľúčovým problémom je nájdenie takého prístupu, ktorý umožňuje analyzovať napr. priestorové rozmiestnenie populácie alebo dochádzkové toky bez toho, aby bolo možné identifikovať akékoľvek individuálne údaje. V praxi ide najmä o prácu s malými sídlami (s nízkym počtom rezidentov) alebo s malými („jednotkovými“) tokmi. Pokiaľ nechceme o tieto údaje prísť, jedným z riešení

je využitie diferenciálnej anonymizácie, ktorá pridáva k pôvodným hodnotám náhodný šum (*random noise*). Potrebné množstvo „šumu“ je riadené parametrom epsilon, ktorý možno odvodiť z pravdepodobnosti, či je možné identifikovať pridanie alebo vynechanie akýchkoľvek individuálnych údajov z datasetu. Podrobnejšie sa tejto problematike venujú Dwork a Roth [13] a Ruggles a kol [27].

Dôslednú anonymizáciu údajov z mobilnej siete nemôžeme podceňovať. Hoci by sa mohlo zdať, že z miliónov záznamov nie je možné identifikovať konkrétnu osobu, nie je to tak. Pri štúdii, ktorá spracovala CDR záznamy 15-mesačného pozorovania [23], sa preukázalo, že prostredníctvom štyroch náhodne zvolených záznamov konkrétneho používateľa dokážeme identifikovať až 95 % používateľov mobilnej siete. Dôsledné rešpektovanie ochrany súkromia je nevyhnutným krokom pri budovaní partnerstiev medzi mobilnými operátormi a spracovateľmi údajov z ich infraštruktúry (akademický, štátny a súkromný sektor). Rovnako je však dôležité aj pri budovaní dôvery zo strany verejnosti.

8. ZÁVER

Rozvoj technológií mobilnej komunikácie prispel k prudkému nárastu jej používateľov. Mobilný telefón sa stal neoddeliteľnou súčasťou každodenného života a unikátnym zdrojom údajov o obyvateľstve, ich priestorovom rozmiestnení, pohybe a aktivitách. Vďaka vysokej penetrácii mobilných telefónov v populácii a schopnosti sledovať ich pohyb na úrovni základňových staníc mobilnej siete môžeme prekonať viaceré limity, ktoré sa spájajú s tradičnými údajmi o populácii, najmä čo sa týka frekvencie zisťovania populačných dát, rýchlosti ich spracovania a v neposlednom rade i ochoty obyvateľov poskytovať presné údaje v cenze či inom plošnom zisťovaní. V roku 2021 bolo v krajinách EÚ registrovaných 1 054 SIM kariet s aktivovanou hlasovou alebo dátovou službou mobilnej siete na 1 000 obyvateľov [14]. Tento bezprecedentný rozsah pokrytia populácie vytvára unikátny predpoklad pre holistické prístupy sledovania miest a regiónov. Všadeprítomnosť a štandardizovaný charakter mobilnej infraštruktúry vytvára dlhodobý rámec na sledovanie priestorovej variability populácie vo vysokom časovom a priestorovom rozlíšení. Ak však chceme s týmito unikátnymi údajmi pracovať, potrebujeme poznať princípy ich vzniku a možnosti ich spracovania, ktoré sme si v stručnosti predstavili v tomto príspevku.

LITERATÚRA

[1] AASA, A. – KAMENJUK, P. – SALUVEER, E. – ŠIMBERA, J. – RAUN, J.: Spatial interpolation of mobile positioning data for population statistics. In: Journal of Location Based Services, 2021, č. 15, s. 239 – 260.
[2] AHAS, R. – AASA, A. – SILM, S. – AUNAP, R. – KALLE, H. – MARK, Ü.: Mobile positioning in spacetime behavior studies: social positioning method experiments in Estonia. In: Cartography and Geographic Information Science, 2007, č. 34, s. 259 – 273.
[3] AHAS, R. – SILM, S. – JÄRV, O. – SALUVEER, E. – TIRU, M.: Using mobile positioning data to model locations meaningful to users of mobile phones. In: Journal of Urban Technology, 2010, č. 17, s. 3 – 27.
[4] BATTY, M.: The new science of cities. MIT press, 2013.
[5] BECKER, R. A. – CACERES, R. – HANSON, K. – LOH, J. M. – URBANEK, S. – VARSHAVSKY, A. – VOLINSKY, C.: A tale of one city: Using cellular network data for urban planning. In: IEEE Pervasive Computing, 2011, č. 10, s. 18 – 26.

[6] BLONDEL, V. D. – DECUYPER, A. – KRINGS, G.: A survey of results on mobile phone datasets analysis. In: EPJ data science, 2015, č. 4, s.1 – 55.
[7] CALABRESE, F. – DI LORENZO, G. – LIU, L. – RATTI, C.: Estimating origin-destination flows using mobile phone location data. In: IEEE Pervasive Computing, 2011, č. 10, s. 36 – 44.
[8] CALABRESE, F. – DIAO, M. – DI LORENZO, G. – FERREIRA JR. J. – RATTI, C.: Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. In: Transportation Research Part C: Emerging Technologies, 2013, č. 26, s. 301 – 313.
[9] CSÁJI, B. C. – BROWET, A. – TRAAG, V. A. – DELVENNE, J. C. – HUENS, E. – VAN DOOREN, P. – SMOREDA, Z. – BLONDEL, V. D.: Exploring the mobility of mobile phone users. In: Physica A: Statistical Mechanics and its Applications, 2013, č. 6, s. 1459 – 1473.
[10] DEVILLE, P. – LINARD, C. – MARTIN, S. – GILBERT, M. – STEVENS, F. R. – GAUGHAN, A. E. – BLONDEL, V. D. – TATEM, A. J.: Dynamic population mapping using mobile phone data. In: Proceedings of the National Academy of Sciences, 2014, č.45, s. 15 888 – 15 893.
[11] DOHERTY, S. T.: Should we abandon activity type analysis? Redefining activities by their salient attributes. In: Transportation, 2006, č. 33, s. 517 – 536.
[12] DUFKOVÁ, K. – FICEK, M. – KENCL, L. – NOVÁK, J. – KOUBA, J. – GREGOR, I. – DANIHELKA, J.: Active GSM cell-id tracking: Where Did You Disappear? In: MELT 2008: Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments. San Francisco, 2008, s. 7 – 12.
[13] DWORK, C. – ROTH, A.: The algorithmic foundations of differential privacy. In: Foundations and Trends in Theoretical Computer Science, 2014, č. 9, s. 211 – 407.
[14] EUROSTAT: European Neighbourhood Policy - East - statistics on science, technology and digital society. [online]. [cit. 22-12-2023]. Dostupné na: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=European_Neighbourhood_Policy_-_East_-_statistics_on_science,_technology_and_digital_society.
[15] GETHING, P. W. – TATEM, A. J.: Can mobile phone data improve emergency response to natural disasters? In: PLoS medicine, 2011, č. 8, e1001085.
[16] GONZÁLEZ, M. C. – HIDALGO, C. A. – BARABÁSI, A. L.: Understanding individual human mobility patterns. In: Nature, 2008, č. 7196, s. 779 – 782.
[17] JACQUES, D. C. Mobile phone metadata for development. Technical Report. 2018, arXiv: 1806.03086.
[18] JÄRV, O. – TENKANEN, H. – TOIVONEN, T.: Enhancing Spatial Accuracy of Mobile Phone Data Using Multi-temporal Dasymetric Interpolation. In: International Journal of Geographical Information Science, 2017, č. 31, s. 1630 – 1651.
[19] LIU, F. – JANSSENS, D. – CUI, J. – WANG, Y. – WETS, G. – COOLS, M.: Building a validation measure for activity-based transportation models based on mobile phone data. In: Expert Systems with Applications, 2014, č. 41, s. 6174 – 6189.
[20] MOKHTARIAN, P. L. – SALOMON, I. – HANDY, S. L.: The impacts of ICT on leisure activities and travel: a conceptual exploration. In: Transportation, 2006, č. 33, s. 263 – 289.
[21] MOLINARI, M. – FIDA, M. R. – MARINA, M. K. – PESCAPE, A.: Spatial interpolation based cellular coverage prediction with crowdsourced measurements. In: Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data, 2015, s. 33 – 38.

[22] MONTJOYE, Y. A. – QUOIDBACH, J. – ROBIC, F. – PENTLAND, A. S.: Predicting personality using novel mobile phone-based metrics. In: Greenberg, A. – Kennedy, W. – Bos, N. (eds.): Social computing, behavioral-cultural modeling, and prediction. Berlin: Springer, 2013a, s. 48 – 55.
[23] MONTJOYE, Y. A. – HIDALGO, C. A. – VERLEYSEN, M. – BLONDEL, V. D.: Unique in the crowd: The privacy bounds of human mobility. In: Scientific Reports, 2013b, č. 3, s.1 – 5.
[24] NOVÁK, J.: Lokalizační data mobilních telefonů: možnosti využití v geografickém výzkumu. [Dizertačná práca]. Praha: Univerzita Karlova v Prahe, 2010.
[25] RICCIATO, F. – WIDHALM, P. – CRAGLIA, M. – PANTISANO, F.: Estimating population density distribution from network-based mobile phone data. Luxembourg: Publications Office of the European Union, 2015.
[26] RICCIATO, F. – WIDHALM, P. – PANTISANO, F. – CRAGLIA, M.: Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. In: Pervasive and Mobile Computing, 2017, č. 35, s. 65 – 82.
[27] RUGGLES, S. – FITCH, C. – MAGNUSON, D. – SCHROEDER, J.: Differential privacy and census data: Implications for social and economic research. In: AEA Papers and Proceedings, 2019, s. 403 – 408.
[28] SCHLICH, R. – SCHÖNFELDER, S. – HANSON, S. – AXHAUSEN, K. W.: Structures of leisure travel: temporal and spatial variability. In: Transport Reviews, 2004, č. 24, s. 219 – 237.
[29] STEENBRUGGEN, J. – TRANOS, E. – NIJKAMP, P.: Data from mobile phone operators: A tool for smarter cities? In: Telecommunications Policy, 2015, č. 39, s. 335 – 346.
[30] ŠVEDA, M. – MADAJOVÁ, M.: Merging diaries and GPS records: The method of data collection for spatio-temporal research. In: Moravian Geographical Reports, 2015, č. 2, s. 12 – 25.
[31] ŠVEDA, M. – SLÁDEKOVÁ MADAJOVÁ, M.: Estimating distance decay of intra-urban trips using mobile phone data: The case of Bratislava, Slovakia. In: Journal of Transport Geography, 2023, 103552.
[32] YANG, J. – SHI, Y. – YU, C. – CAO, S. J.: Challenges of using mobile phone signalling data to estimate urban population density: towards smart cities and sustainable urban development. In: Indoor and Built Environment, 2020, č. 2, s. 147 – 150.

RESUMÉ

Príspevok predstavuje údaje z mobilnej siete ako nekonvenčný zdroj údajov na sledovanie priestorového rozmiestnenia a mobility populácie. Vďaka vysokej penetrácie mobilných zariadení v populácii máme možnosť získať údaje s charakterom plošných zisťovaní. V úvodnej časti sa zaoberá architektúrou mobilnej siete a jej diferenciáciou na jednotlivé vývojové časti – generácie (2G, 3G, 4G, 5G). V ďalšej časti sa zaoberá základnými možnosťami lokalizácie v mobilnej sieti, kde rozlišujeme pasívnu a aktívnu lokalizáciu. Kým využitie aktívnej lokalizácie limitujú uvedomelé súhlasy používateľov mobilnej siete, pri pasívnej lokalizácii môžeme vyťažiť množstvo polohových informácií zo záznamov, ktoré necháva mobilné zariadenia pri využívaní mobilnej siete (digitálne stopy). Hoci údaj o polohe SIM karty sa explicitne nenachádza v údajoch mobilnej siete, dokážeme ho významne aproximovať využitím rôznych prístupov, ktoré v príspevku predstavujeme. Pravidelnosť lokalizácií v mobilnej sieti potom umožňuje extrahovať niektoré signifikantné lokality, ako napr. pravidelnú dennú

alebo nočnú lokalitu. Lokalizácia s využitím mobilnej siete neposkytuje informácie o aktivitách, použitých dopravných prostriedkoch či sociálnom kontexte. Napriek tomu môžeme identifikovať niekoľko základných kategórií s predpokladanou interpretáciou, ako napr. rezident, pracujúci, dochádzajúci, tranzitujúci a pod. Príspevok uzatvára téma ochrany osobných údajov a potreba anonymizácie a agregácie týchto unikátnych údajov.

RESUME

The contribution presents data from a mobile network as an unconventional data source for monitoring spatial distribution and mobility of population. Given the widespread use of mobile devices among the population, we have an opportunity to obtain data with the character of large nationwide surveys. The introductory part of the contribution addresses the architecture of the mobile network and different generations of mobile networks (e.g., 2G, 3G, 4G, 5G). The next part deals with essential possibilities of localization in the mobile network, distinguishing between passive and active localization. While the use of active localization is limited by the informed consent of mobile network users, by means of passive localization, we can extract the location information from records left by mobile devices while using the mobile network (digital traces). Although the location data of the SIM card is not explicitly present in the mobile network data, we can meaningfully approximate it using various approaches presented in the contribution. The regularity of localizations in the mobile network then allows us to extract some significant locations, such as regular day or night locations. Localization using the mobile network does not provide information about activities, modes of transportation, or social context. Nevertheless, we can identify several basic categories with presumed interpretations, such as residents, workers, commuters, transit users, and others. The contribution concludes with the topic of personal data protection and the need for anonymization and aggregation of these unique data.

PROFESIJNÝ ŽIVOTOPIS

***Mgr. Martin Šveda, PhD.**, absolvoval magisterské štúdium v odbore geografia a kartografia (2007) a doktorandské štúdium v odbore regionálna geografia na Prírodovedeckej fakulte Univerzity Komenského v Bratislave (2011). Od roku 2017 pôsobí ako odborný asistent na Katedre regionálnej geografie a rozvoja regiónov Prírodovedeckej fakulty Univerzity Komenského v Bratislave. Súčasne pracuje ako samostatný vedecký pracovník v Geografickom ústave SAV. Vo svojej výskumnej činnosti sa zameriava predovšetkým na procesy suburbanizácie a ich vplyvy na transformáciu prímestských sídiel. Venuje sa aj sledovaniu časovo-priestorových vzorcov správania obyvateľov prostredníctvom lokalizačných údajov mobilnej siete.*

KONTAKT

martin.sveda@uniba.sk

Informatívny článok/Informative article

Dagmar CEĽUCHOVÁ BOŠANSKÁ, Martin JANÍK, Filip NGUYEN
Alistiq, s. r. o.

POKUS O MONITOROVANIE SOCIÁLNEHO NAPÄTIA Z PRÍSPEVKOV NA SOCIÁLNEJ SIETI FACEBOOK

ATTEMPT OF MONITORING OF SOCIAL TENSION FROM POSTS ON THE FACEBOOK SOCIAL NETWORKING WEBSITE

ABSTRAKT
Sociálne siete sú čoraz populárnejším miestom, kde ľudia zdieľajú svoje pocity a názory na rôzne udalosti a tematicky súvisiace správy. Táto popularita vytvára veľké množstvo údajov týkajúcich sa nálady a sociálneho napätia ľudí. Cieľom bolo analyzovať sentiment príspevkov na sociálnej sieti Facebook v slovenčine a použiť tieto údaje na získanie informácií o emocionálnom stave ľudí. Použitá metóda využíva model na spracovanie prirodzeného jazyka XLM-RoBERTa-large a umožňuje korelovať sociálne napätie s udalosťami zo skutočného sveta. Analýza sentimentu príspevkov na sociálnych sieťach môže byť užitočná pre tvorcov politík a sociológov, a v budúcnosti sa môže rozšíriť na ďalšie platformy a jazyky.

ABSTRACT
Social networks are an increasingly popular place where people share their feelings and opinions on various events and thematically related topics. This popularity creates a large amount of data regarding people's mood and social tension. The aim was to analyze the sentiment of the posts on the social networking website Facebook in Slovak and to use this data to obtain information about the emotional state of people. This method uses the natural language processing model XLM-RoBERTa-large and allows social tensions to be correlated with real-world events. Sentiment analysis of social media posts can be useful for policymakers and sociologists, and may be extended to other platforms and languages in the future.

KĽÚČOVÉ SLOVÁ
analýza sentimentu, sociálne siete, strojové učenie, sociálne napätie, hĺbková analýza dát

KEY WORDS
sentiment analysis, social networks, machine learning, social tension, data mining

1. ÚVOD
Používatelia online sociálnych sietí sú prepojení prostredníctvom vzťahov a interakcií podobne ako v reálnom živote. Vzťahy sú stabilné spojenia medzi dvoma alebo viacerými používateľmi. Existuje viacero typov vzťahov, napríklad priateľstvo alebo príslušnosť k rovnakej skupine. Niektoré typy vzťahov sú vzájomné, ako napríklad príbuzenstvo, zatiaľ čo iné sú riadené alebo asymetrické, ako napríklad vzťah sledovania populárnej osobnosti na sociálnej sieti. Interakcie sa objavujú, keď používateľ komunikuje s ostatnými. Typy interakcií zahŕňajú zverejnené statusy, priame správy, odpovede a zmienky používateľov. Väčšina týchto typov zahŕňa aj tvorbu obsahu. Keď používateľ vytvorí alebo zverejní nový obsah, vytvorí sa autorský

vzťah medzi používateľom a obsahom. Nový obsah môže súvisieť aj s existujúcim obsahom (napr. ako odpoveď alebo zmienka) alebo s inými používateľmi (napr. používateľ je spomenutý v obsahu). Používatelia potom môžu interagovať s novovytvoreným obsahom tak, že naň odpovedia, napríklad že sa im páči, alebo si ho uložia.

Interakcie na sociálnych sieťach je možné vnímať ako objemné množstvo údajov neštruktúrovaného textu ako sú články, blogové príspevky, statusy a komentáre. Dajú sa ľahko analyzovať pomocou niekoľkých techník dátovej vedy, pričom ako vhodná metóda analýzy neštruktúrovaných textových dát sa javí sémantická sieťová analýza, ktorá umožňuje sledovať fenomény, ako je sociálne napätie, dôvera alebo polarizácia [26].

Vzostup sociálnych sietí ako prostriedkov na vyjadrenie názorov a emócií možno pripísať zvyšujúcej sa dostupnosti technológií. S rozšírením smartfónov a internetového pripojenia majú ľudia neustály prístup k platformám sociálnych médií, čo umožňuje jednotlivcom vyjadrovať svoje názory a emócie v pohybe, kedykoľvek a z akéhokoľvek miesta. Najvyužívanejšou sociálnou sieťou u nás (ale aj globálne) je Facebook. Podľa prieskumu Go4insight ju [9] používa až 76 % obyvateľov Slovenska aspoň raz do mesiaca. Denné využívanie je na úrovni 55 %. Keďže popularita týchto platforiem stále rastie, rastie aj dôraz na dôležitosť porozumenia interakcií v týchto digitálnych priestoroch.

1.1. Analýza sentimentu na sociálnej sieti pohľadom modernej sociológie

Emócie predstavujú dôležitý spoločenský signál pre ostatných – informujú ich o rôznych spôsoboch interakcie s ohľadom na vlastnú motiváciu a ciele. Svoje emócie ľudia pravidelne zdieľajú s ostatnými ľuďmi aj na sociálnych sieťach vzhľadom na ich hodnotnú spoločenskú funkciu a ovplyvňujú emócie ostatných. Napríklad šťastie sa môže rozšíriť cez sociálnu sieť a spôsobiť vznik klastrov šťastných a nešťastných ľudí. Prístupov, ako nazerať a skúmať emócie, je celý rad, svoj pohľad na emócie ponúkajú antropológovia, neurofyziológovia, psychológovia a mnohí ďalší. Český psychológ Milan Nakonečný vymedzuje pojem emócie ako prežívanie jednoty cítenia a telesných zmien s odkazom na ich pôvodný biologický význam, ktorým bola adaptívna reakcia na dôležité životné situácie. V tomto zmysle uvádza emócie ako procesy hodnotenia životného významu situácií, ktoré sú súčasne spojené s mobilizáciou energie (aktiváciou organizmu) nutné na vytvorenie účelného správania. Podľa Nakonečného boli primárne emócie súčasťou inštinktov [19].

V zahraničí sa vo vzťahu k analýze sentimentu uznáva Plutchikova teória primárnych emócií a ich zmiešanín. Podľa Plutchika vychádzajú emócie z genetického základu a vznikli selekciou životne úspešných adaptívnych mechanizmov a ich postupným zdokonaľovaním, a ako také sú formou adaptácie. Plutchik uvádza osem primárnych emócií, ktoré sú vrodené, ostatné vznikajú ich miešaním. Primárne emócie sú akceptovanie, strach, prekvapenie, smútok, hnus, hnev, očakávania, a radosť. Sekundárne emócie sú tzv. dyády, triády a ďalšie zmiešaniny vybraných emócií [23].

V klasickej sociologickej literatúre sa pojem sociálne napätie používa pri analýze procesov pretrhávania sociálnych väzieb, straty hodnôt a zvýšenia anómie. Sociológ Neil Joseph Smelser definuje „napätie“ ako tendenciu k nerovnováhe výmeny medzi dvoma alebo viacerými zložkami systém [27]. Sociálne napätie je možné opísať aj ako

špecifickú podmienku sociálneho vedomia a sociálnych emócií. Tento stav je charakterizovaný hromadením duševnej únavy a podráždenosti, frustrácie a deprivácie, agresivity a depresie významnej časti spoločnosti. K sociálnemu napätiu je možné pristúpiť aj tak, že verejné inštitúcie môžu fungovať ako „bezpečnostné ventily“, tie môžu slúžiť na prerušenie pocitov nepriateľstva, môžu fungovať ako bleskozvody, ale nemôžu zabrániť opakovanému nahromadeniu napätia [5].

Prvým krokom v analýze sentimentu je zvyčajne definovanie viet, ktoré vyjadrujú sentiment. V minulosti sa tieto vety kategorizovali buď ako subjektívne alebo objektívne, pričom iba prvé z nich boli spojené so sentimentom. Objektívne vety prezentovali fakty, zatiaľ čo subjektívne vety vyjadrovali osobné pocity alebo presvedčenie. V súčasnosti sa však všeobecne uznáva, že subjektivita by sa nemala vždy stotožňovať s názorom. Napríklad, subjektívne vety nemusia vždy vyjadrovať sentiment, ako napr.: „*nepoznám farmára, ktorému by aspoň raz zviera neušlo*“. Na druhej strane, objektívne vety môžu skutočne vyjadrovať sentiment, ako napr. „*Slúchadlá sa rozbili za dva dni*“. Okrem toho vety často obsahujú viacero polarít, takže je náročné klasifikovať vetu ako čisto pozitívnu alebo negatívnu [16]. Sentiment príspevkov a komentárov sa preto hodnotí vždy z pohľadu autora, t. j. zisťuje sa emocionálny podtón textu z pohľadu, či v prejave autora prevláda pozitívna alebo negatívna nálada. Inými slovami, cieľom analýzy sentimentu je určiť postoj alebo emocionálny stav autora príspevku vrátane primárnych emócií.

Odkedy si online sociálne médiá získali celosvetovú popularitu, objavil sa široký prúd výskumu extrakcie a štruktúrovania dostupných informácií [7]. Indikátory sentimentu sa vytvárajú s cieľom reflektovať nálady, názory alebo očakávania skupiny respondentov. Môže ísť o rôzne skupiny jednotlivcov, domácností alebo podnikov. Indikátory sentimentu sú väčšinou založené na kvalitatívnych prieskumoch, v ktorých sa kladú otázky týkajúce sa názorov na minulé alebo aktuálne dianie, či na vývoj do budúcna. Podľa Hartla [10] sa nálada všeobecne definuje ako emocionálny stav, ktorý v priebehu určitého času sprevádza prežívanie a činnosť človeka. Ako uvádza Nakonečný kvalita cítenia zotrvávajúca v čase sa nazýva nálada, a je to teda citový stav, ktorý tvorí akési pozadie duševného života jedinca [19].

Najpopulárnejšími sociálnymi sieťami sú X (predtým známa ako Twitter) a Facebook. X je obzvlášť populárny medzi zahraničnými vedcami kvôli dostupnosti veľkého množstva údajov, ktoré je možné zhromaždiť. Tieto sociálne siete sú ideálnym zdrojom údajov na analýzu sentimentu, pretože poskytujú jednotlivcom platformu na vyjadrenie ich myšlienok, pocitov a názorov na rôzne témy, od aktuálnych udalostí a spoločenských problémov až po osobné skúsenosti a záujmy. Schopnosť slobodne sa vyjadrovať a byť vypočutý ostatnými prispieva k rastúcej príťažlivosti sociálnych sietí. Vďaka možnosti sledovať, páčiť sa a komentovať príspevky ostatných sa jednotlivci môžu zapojiť do konverzácií a diskusií. Táto interakcia umožňuje výmenu rôznych názorov a emócií, výsledkom čoho je podnetné online prostredie.

Vedecké práce v sociológii zamerané na analýzu sentimentu príspevkov na sieťach Facebook a X sa zaoberali najmä reakciou verejnosti na extrémne udalosti [15, 21]. modelovaním ľudského správania [1] a predpovedaním akcií [20], postojov [17] či dokonca výsledkov volieb [28].

Analýza sentimentu sociálnych sietí je okrem uvedených príkladov aj cenným nástrojom pre marketing, pretože pomáha výskumníkom a podnikom získať prehľad o tom, ako zákazníci vnímajú ich produkty, značku alebo celkovú zákaznícku skúsenosť. Analýzou a pochopením pocitov vyjadrených v recenziách zákazníkov, príspevkoch na sociálnych sieťach a iných formách spätnej väzby online môžu marketéri efektívne prispôsobiť svoje marketingové stratégie a správy. Analýza pozitívneho sentimentu sa môže použiť na identifikáciu spokojných zákazníkov a influencerov, na ktorých sa môžu zamerať [11].

2. POUŽITÉ ÚDAJE

V podprojekte *Monitorovanie sociálneho napätia z príspevkov na sociálnej sieti Facebook* sa pri analýze sentimentu využili verejné príspevky na Facebooku, pozostávajúce zo statusov, komentárov, odpovedí a reakcií (Páči sa mi) za obdobie jedného roka, od 1. septembra 2022 do 31. augusta 2023. Zdrojom zhromažďovania týchto údajov bola sociálna sieť Facebook spoločnosti Meta, pričom technika použitá na kompiláciu údajov z tejto platformy bola Facebook GraphAPI. Zhromažďovanie údajov sa realizovalo v prísnom súlade s etickými normami s dôrazom na anonymizáciu akýchkoľvek osobných údajov, čím sa zabezpečila neutralita a integrita zhromaždených údajov. Získavali sa len verejne zdieľané príspevky a komentáre bez informácie o profiloch používateľov, čiže anonymizované údaje.

Zdrojové údaje predstavujú súbor neštruktúrovaných údajov a obsahujú celkove 31 755 230 príspevkov, ktoré majú svoj identifikátor a časovú pečiatku, čo umožňuje zoradiť ich v čase a agregovať po jednotlivých dňoch. Komentáre sú prelinkované s príspevkom, ktorého sa týkajú a komentáre na komentáre aj s komentárom, na ktorý reagujú. Každý príspevok a komentár má navyše pripísaný počet reakcií cez emotikony ako *Páči sa mi* a *srdiečko*. Momentálne sa však ešte pre takýto typ prepojení a informácií nenašlo využitie.

Na analýzu sentimentu sa využili príspevky (POST) a komentáre (COMMENT) od používateľov na skupinových stránkach (FB groups), dedikovaných stránkach organizácií a firiem (FB pages) alebo profiloch verejne známych osobností. Zdroje – teda profily, skupiny a stránky – sa vyberali kombináciou algoritmov strojového učenia a ľudskej expertízy. Pokročilé neurónové siete najprv zmapovali široký terén informácií v online priestore a identifikovali hlavné a okrajové zdroje na sociálnych sieťach, v blogoch, na fórach, webových stránkach, platformách na zasielanie správ, v skupinách a ďalších zdrojoch. Na identifikáciu ďalších zdrojov vrátane subkultúrnych a okrajových skupín sa používali pokročilé algoritmy založené na grafoch. Potom odborní analytici s dlhoročnými skúsenosťami v oblasti bezpečnosti, dezinformácií a politickej analýzy tieto zdroje preskúmali, aby vytvorili vyvážený súbor tisícov najrelevantnejších zdrojov otázok, ako je extrémizmus, verejná mienka a propaganda. Tento hybridný prístup zabezpečuje široké pokrytie vplyvných zdrojov zo všetkých kútov online priestoru. Patria sem najvplyvnejší mediálni a politickí aktéri (strany aj jednotliví politici), relevantní influenceri a alternatívne mediálne projekty. Zoznam je doplnený o ďalšie zdroje z hľadiska informačnej bezpečnosti vrátane dezinformačných a extrémistických kanálov a skupín. Konkrétne ide o: politikov, politické strany, mainstream médiá, alternatívne médiá, vplyvné médiá, štátne inštitúcie/orgány, predstaviteľov vlády, mimovládne organizácie, komunity.

Z vytvoreného zoznamu zdrojov sa následne vybrali tie, ktoré sa nachádzajú na sociálnej sieti Facebook, a sledovali sa rôzne metriky ich relevantnosti, ako napríklad počet príspevkov za deň. Zdroje sa automaticky pridávajú – to znamená, že ak niekto v príspevku spomenie iný profil alebo stránku, pridá sa do zoznamu. Vytvorený zoznam zdrojov pravidelne kontrolujú a aktualizujú analytici spoločnosti Gerulata. Obdobie aktualizácie zoznamu nie je dlhšie ako 3 mesiace. Pre tento prípad použitia sú zo zoznamu zdrojov identifikované len tie, ktoré produkujú obsah v slovenčine, a to pomocou modelu s otvoreným zdrojovým kódom[1], natrénovaným na klasifikáciu textu podľa jazyka, v ktorom je napísaný.

Cieľom analytického spracovania zdrojových údajov bolo identifikovanie tematického zamerania príspevku alebo komentára, a určenie sentimentu v danom príspevku alebo komentára, čím sa myslí kategorizovanie textu ako pozitívneho, negatívneho alebo neutrálneho. V prvom kroku bolo dôležité získať odhad alebo predpoveď, ktorá by slúžila ako základný model na porovnanie s komplexnejšími modelmi. Základný model je triviálne riešenie, ktoré na výpočet predpovede používa heuristiku alebo jednoduchú štatistiku. Pri každom modelovaní je dôležité mať základný model, pretože ho možno použiť na porovnanie výkonnosti zložitejších finálnych modelov. Jediný spôsob, ako zistiť, či je daný model dobrý alebo výkonný, je porovnať ho so základným modelom [12].

3. VYTVORENIE ZÁKLADNÉHO MODELU

Na vytvorenie základného modelu na určovanie sentimentu alebo témy bol zvolený binárny klasifikátor založený na technike „Support Vector Machines“ (SVM). SVM sú novou technikou vhodnou na úlohy binárnej klasifikácie, ktorá obsahuje prvky neparametrickej aplikovanej štatistiky, neurónových sietí a strojového učenia. SVM model môže byť lineárny alebo nelineárny [14].

SVM klasifikátor slúži na rozdelenie príspevkov hyperrovinou na negatívne (hodnota klasifikátora -1) a pozitívne (hodnota klasifikátora +1). Cieľom SVM je nájsť oddeľujúcu hyperrovinu (rozhodovaciu hranicu), ktorá je maximálne vzdialená od akéhokoľvek bodu v trénovacích údajoch. V tomto prípade sa aplikuje lineárny SVM algoritmus na binárny klasifikačný problém, kde rozhodovacia hranica je lineárna. Klasifikačný problém predstavuje priradenie sentimentu (pozitívneho alebo negatívneho) súboru vstupných vektorov $\vec{x}$, získaných z príspevkov a komentárov zo sociálnej siete. Lineárny binárny klasifikátor $f(\vec{x})$ založený na SVM možno získať pomocou nasledujúcej rovnice: $f(\vec{x}) = sign\,(\vec{w}^T\vec{x} + b)$, kde $\vec{w}^T$ je transpozícia váhových vektorov, znamienko („sign“) označuje rozhodnutie na základe skóre $\vec{w}^T\vec{x} + b$. Dve možné rozhodnutia sú $+1$ a -1, ktoré označujú pozitívny a negatívny sentiment. Absolútny člen b sa používa na určenie hyperrovín, ktoré sú kolmé na normalizovaný vektor [8].

4. POUŽITÝ MODEL ZALOŽENÝ NA NEURÓNOVÝCH SIEŤACH A HĹBKOVOM UČENÍ

Na klasifikáciu získaných údajov zo sociálnej siete Facebook bol použitý oveľa komplexnejší model XLM-RoBERTa-large, založený na neurónových sieťach, ktoré sú aktuálne nepísaným štandardom v oblasti spracovania prirodzeného jazyka. Tento model je založený na modeli transformátora predstavujúcom neurónovú sieť, ktorá sa

[1] *Zdroj: https://huggingface.co/facebook/fasttext-language-identification, dátum referencie: 04.09.2023.*

učí kontext a tým aj význam slov sledovaním vzťahov v sekvenčných údajoch, ako sú slová vo vete. Modely transformátora používajú vyvíjajúci sa súbor matematických techník, nazývaných pozornosť alebo sebapozornosť, na detekciu prvkov, dokonca aj vzdialených dátových prvkov, ktoré sa v sérii navzájom ovplyvňujú a navzájom so sebou súvisia [1].

XLM-RoBERTa („A Robustly Optimized BERT Pretraining Approach“) je nový model od Facebook AI a je inšpirovaný modelom BERT („Bidirectional Encoder Representations from Transformers“) z roku 2018 od spoločnosti Google [4]. Hoci je komplexné porovnanie medzi rôznymi metódami zložité, príprava tohto jazykového modelu výrazne zlepšila výkonnosť.

XLM-RoBERTa sa líši od XLM tým, že trénuje model RoBERTa na obrovskom viacjazyčnom súbore údajov[2], ale vyhýba sa cieľu prekladového jazykového modelu – XLM-RoBERTa je trénovaný iba s cieľom modelu na maskovanie jazyka [6]. XLM-RoBERTa je preto viacjazyčný maskovaný jazykový model založený na transformátore, ktorý je predtrénovaný na texte veľkosti 2,5 TB v 100 jazykoch, a tým dosahuje uspokojivý výkon v oblasti medzijazykovej klasifikácie, sekvenčného označovania a odpovedí na otázky. Dosahuje tiež nevídaný výkon v medzijazykovom porozumení, čo je úloha, pri ktorej sa model trénuje v jednom jazyku a potom sa používa s inými jazykmi bez ďalších trénovacích údajov. Text použitý na trénovanie pochádza z otvoreného repozitára Common Crawl, ktorý obsahuje texty získané nástrojmi crawlers z verejne dostupných webových stránok. Model XLM-RoBERTa je dostupný v dvoch veľkostiach, ako uvádza tabuľka č. 1. V tabuľke je na porovnanie pridaný model SlovakBERT [22], ktorý je takisto založený na architektúre RoBERTa, ktorý ale bol trénovaný na slovenský jazyk.

Tabuľka č. 1: Porovnanie relevantných parametrov dvoch modelov XLM-RoBERTa rôznej veľkosti a modelu SlovakBERT

Relevantné premenné charakterizujúce model	XLM-RoBERTa-base	XLM-RoBERTa-large	SlovakBERT
Počet vrstiev (blokov transformátora)	12	24	12
Počet skrytých vrstiev neurónových sieti	768	1024	768
Počet „heads“ pre mechanizmus sebapozornosti	12	16	12
Celkový počet parametrov modelu	278 miliónov	560 miliónov	125 miliónov
Počet podporovaných jazykov	100 (vrátane slovenského jazyka)		1
Veľkosť trénovacieho datasetu (v tokenoch)	$167x10^9$		$4,6\ x10^9$
Veľkosť slovenského datasetu (v tokenoch)	$3,2\ x10^9$		$4,6\ x10^9$
Veľkosť slovníka	250 000		50 000

Zdroj: https://huggingface.co/xlm-roberta-base/, https://huggingface.co/xml-roberta-large, https://huggingface.co/gerulata/slovakbert

Na základe expertného hodnotenia, ktoré zahŕňalo kritériá ako presnosť, adaptabilitu a efektívnosť, bolo rozhodnuté pokračovať s XLM-RoBERTa-large. Podľa spomínaných kritérií sa hľadal model, ktorý bude najviac vyhovovať klasifikácii sentimentu a kategorizácii z pohľadu tém.

[2] *Model BERT trénovaný na približne 100 jazykoch.*

4.1. Klasifikácia príspevkov z pohľadu témy

Predtrénovaný model s hĺbkovým učením XLM-RoBERTa-large sa použil na klasifikáciu textu do tém na základe neustále aktualizovanej taxonómie IPTC Media Topic Newscodes so zameraním na kategorizáciu textu.

Tabuľka č. 2: Témy taxonómie IPTC Media Topic Newscodes

Názov témy (MediaTopic ID)	Definícia
Umenie, kultúra, zábava a médiá (medtop:01000000)	Všetky formy umenia, zábavy, kultúrneho dedičstva a médií
Konflikt, vojna a mier (medtop:16000000)	Akty sociálne alebo politicky motivovaného protestu alebo násilia, vojenské činnosti, geopolitické konflikty, ako aj úsilie o riešenie problémov
Zločin, právo a spravodlivosť (medtop:02000000)	Stanovenie a / alebo vyhlásenie o pravidlách správania v spoločnosti, presadzovanie týchto pravidiel, porušovanie pravidiel, trestanie páchateľov a organizácie a orgány zapojené do týchto činností
Katastrofa, nehoda a mimoriadny incident (medtop:03000000)	Človek alebo prírodná udalosť, ktorá má za následok stratu na životoch alebo zranenie živých tvorov a / alebo poškodenie neživých predmetov alebo majetku
Hospodárstvo, podnikanie a financie (medtop:04000000)	Všetky záležitosti týkajúce sa plánovania, výroby a výmeny bohatstva.
Vzdelávanie (medtop:05000000)	Všetky aspekty formálneho alebo neformálneho zvyšovania vedomostí
Životné prostredie (medtop:06000000)	Všetky aspekty ochrany, poškodenia a stavu ekosystému planéty Zem a jej okolia.
Zdravie (medtop:07000000)	Všetky aspekty fyzického a duševného blaha
Ľudský záujem (medtop:08000000)	Položka, zahŕňajúca emocionálne diskusie o jednotlivcoch, skupinách, zvieratách, rastlinách alebo iných objektoch
Práca (medtop:09000000)	Sociálne aspekty, organizácie, pravidlá a podmienky ovplyvňujúce tvorbu bohatstva a poskytovanie služieb a ekonomickú podporu nezamestnaným.
Životný štýl a voľný čas (medtop:10000000)	Činnosti vykonávané na potešenie, relaxáciu alebo rekreáciu mimo plateného zamestnania vrátane stravovania a cestovania.
Politika (medtop:11000000)	Miestne, regionálne, národné a medzinárodné vykonávanie moci alebo boj o moc a vzťahy medzi riadiacimi orgánmi a štátmi.
Náboženstvo (medtop:12000000)	Systémy viery, inštitúcie a ľudia, ktorí poskytujú morálne vedenie nasledovníkom
Veda a technika (medtop:13000000)	Všetky aspekty týkajúce sa ľudského porozumenia, ako aj metodického štúdia a výskumu prírodných, formálnych a spoločenských vied, ako sú astronómia, lingvistika alebo ekonómia
Spoločnosť (medtop:14000000)	Obavy, problémy, záležitosti a inštitúcie súvisiace s ľudskými sociálnymi interakciami, problémami a blahobytom, ako sú chudoba, ľudské práva a plánovanie rodiny
Šport (medtop:15000000)	Športová činnosť alebo zručnosti, ktoré zahŕňajú fyzické a/alebo duševné úsilie a organizácie a orgány zapojené do týchto činností
Počasie (medtop:17000000)	Štúdia, predpoveď a hlásenie meteorologických javov

Zdroj:https://www.iptc.org/std/NewsCodes/treeview/mediatopic/mediatopic-en-GB.html

Na trénovanie modelu slúžili datasety zostavené použitím kombinácie reálnych a umelo vytvorených dát v slovenskom jazyku. Dáta boli spracované a optimalizované prostredníctvom rôznych prístupov vrátane multilingválnych a spoločne anotovaných údajov, aby sa maximalizovala ich účinnosť a presnosť. Anotovaných bolo 1000 inštancií textov – komentárov, pričom každý text anotovali štyria anotátori.

Na trénovanie a validáciu modelu sa využili len tie dáta, pri ktorých sa zhodli aspoň traja anotátori. Takto sa zabezpečilo vytvorenie tzv. zlatého datasetu.

Trénovanie bolo rozdelené do viacerých iterácií. Každou iteráciou sa zvyšovala správnosť modelu – z pôvodných 79 % na 87 %. Dosiahnutý výsledok pri komplexnejších úlohách a jazykových modeloch sa považuje za pomerne vysoko kompetentný.

Ďalej sú uvedené postupné fázy trénovania s akcentom na experimenty s rôznymi kombináciami reálnych a umelo vytvorených dát. Tento proces umožnil vyvinúť optimalizovaný model, ktorý mohol lepšie riešiť problémy s rozpoznávaním emotikonov a zabraňovať chybným klasifikáciám. Ukázalo sa, že v tomto prípade ladenia dochádza navyše k prenesenému učeniu (transfer learning), teda trénovacie dáta v inom jazyku pozitívne ovplyvňujú výsledky v slovenskom jazyku. Išlo o nasledovné iterácie trénovania, pričom sa použili nasledujúce postupy a dosiahnutá uvedená presnosť modelu:

- reálne dáta v slovenskom jazyku (správnosť 79 %),
- reálne + Umelo vytvorené dáta v slovenskom jazyku (správnosť 83 %),
- reálne + Umelo vytvorené dáta v slovenskom jazyku + Spoločne anotované dáta (správnosť 81 %),
- reálne + umelo vytvorené dáta v slovenskom jazyku + spoločne anotované dáta + multilingválne reálne dáta a umelo vytvorené dáta (správnosť 86 %),
- dáta na obmedzenie nesprávnej klasifikácie na základe emotikonov (k vetám s negatívnym sentimentom boli priradené niektoré emotikony, ktoré v takých prípadoch nenesú sentiment, prípadne sú často využívané pri vyjadrení sarkazmu). Model trénovaný na tomto datasete je v produkcii (správnosť 87 %).

Výstupom bol súbor vo formáte JSON, ktorý každému textu priraďuje tému zo spomínanej taxonómie ako aj skóre (pravdepodobnosť), že sa tejto témy skutočne týka. Z pohľadu témy boli klasifikované iba príspevky typu POST (to znamená, že komentáre sa neklasifikovali), ku každému príspevku bola priložená celá odpoveď klasifikátora, aby bolo možné pri výpočte indexu sociálneho napätia pracovať s nastavením požadovanej hranice pravdepodobnosti aj s množinou tém, ktoré reálne majú vstupovať do výpočtu. Pre príspevky typu COMMENT (komentár) bola téma priradená na základe POSTu, na ktorý sa COMMENT vzťahuje. Teda všetky komentáre k príspevku majú spoločnú tému príspevku.

Vyladený model XLM-RoBERTa-large bol schopný efektívne kategorizovať tematické oblasti podľa IPTC štandardu do hĺbky prvej úrovne, ukazujúc jeho robustnosť a spoľahlivosť v analýze a klasifikácii textových dát. Pri vybraných kategóriách IPTC štandardu model podporuje aj hĺbku druhej úrovne. V rámci dostupných tém taxonómie IPTC sme zvolili podmnožinu tém, ktorá dáva význam z pohľadu sledovania nálad v spoločnosti. Konkrétne sme sa zamerali na témy prvej a druhej úrovne. Prvou témou je Umenie, kultúra, zábava a médiá ((medtop:01000000)). Druhou témou je Konflikt, vojna a mier ((medtop:16000000)), ktorá sa delí na podtémy ako Teroristický čin (medtop:20000053), Ozbrojený konflikt (medtop:20000056), Občianska nepokoj (medtop:20000065), Štátny prevrat (medtop:20000070), Kybernetická vojna (medtop:20001361), Masaker (medtop:20000071), Mierový proces (medtop:20000073), Povojnová obnova (medtop:20000077) a Vojnoví zajatci (medtop:20000080). Ďalšou témou je Zločin,

právo a spravodlivosť (medtop:02000000), ktorá zahŕňa Zločin (medtop:20000082), Súdnictvo (medtop:20000106), Právo (medtop:20000121) a Vymožiteľnosť práva (medtop:20000129). Nasleduje téma Katastrofa, nehoda a mimoriadny incident (medtop:03000000). Ďalej boli zvolené témy Hospodárstvo, podnikanie a financie (medtop:04000000), Životné prostredie (medtop:06000000), Zdravie (medtop:07000000), Ľudský záujem (medtop:08000000), Práca (medtop:09000000) a Politika (medtop:11000000), ktorá sa delí na Voľby (medtop:20000574), Základné práva (medtop:20000587), Vláda (medtop:20000593), Vládna politika (medtop:20000621), Medzinárodné vzťahy (medtop:20000638), Mimovládna organizácia (medtop:20000646), Politická kríza (medtop:20000647), Politický disent medtop:20000648 a Politický proces (medtop:20000649). Ďalšie zvolené témy boli Náboženstvo (medtop:12000000), Veda a technika (medtop:13000000), Spoločnosť (medtop:14000000), Šport (medtop:15000000), Počasie (medtop:17000000).

4.2. Ladenie správnosti modelov pomocou anotátorov

Na klasifikáciu textu z pohľadu sentimentu bol použitý rovnaký predtrénovaný model na ladenie. Cieľom klasifikácie bolo určiť, či daný text vyjadruje pozitívny, negatívny alebo neutrálny postoj autora k diskutovanej téme. Hoci sentiment predstavuje v modeli len tri triedy (pozitívny, negatívny alebo neutrálny), bolo potrebné vytvoriť dodatočné datasety na ladenie správnosti modelov pomocou anotátorov, najmä pre prípady, keď bolo ťažké jednoznačne určiť sentiment, napríklad, ak išlo o nepochopenú iróniu.

Anotátori sa pri určovaní pozitívneho, negatívneho a neutrálneho sentimentu riadili obecným postupom, ktorý bol vytvorený ako súčasť výstupov projektu. Všeobecný postup určovania sentimentu obsahoval všeobecné prípady pozitívnych (tabuľka č. 3), negatívnych (tabuľka č. 4) alebo neutrálnych (tabuľka č. 5) príspevkov a konkrétne rozhodnutia týkajúce sa anotácie v špeciálnych prípadoch.

Tabuľka č. 3: Prípady pozitívneho sentimentu v príspevkoch

Prípad	**Príklad**
Príspevok prejavuje pozitívne nálady, postoj alebo emocionálny stav autora príspevku.	„*Joj krásavec krásny*.“
Príspevok obsahuje nejaké kladné slová.	*„Krásne foto.“*
Príspevok obsahuje pozitívny ilokučný akt (zložka rečového aktu, ktorou hovoriaci vyjadruje svoj vzťah k oznamovanému obsahu, napr.: wow, blahoželám, top).	*„Amen.“*

Zdroj: [2, 18, 25], vlastné spracovanie autorov

Tabuľka č. 4: Prípady negatívneho sentimentu v príspevkoch

Prípad	**Príklad**
Príspevok prejavuje negatívne nálady, postoj alebo emocionálny stav autora príspevku.	*„Tvoj čas konči hurá.“*
Príspevok obsahuje nejaké negatívne slová.	*„To hej to by už vyzerala ozaj jak sliepka“.*
Príspevok obsahuje urážlivé výrazy, slová, frázy alebo výrazy, ktoré sú neúctivé alebo škodlivé voči jednotlivcom alebo skupinám. Zahŕňa používanie jazyka, ktorého cieľom je zastrašiť, ponížiť alebo znevážiť ostatných, často s cieľom spôsobiť emocionálnu alebo psychickú ujmu. Urážlivý jazyk môže mať rôzne podoby vrátane urážok, hanlivých výrazov, nadávok, vyhrážok alebo vulgarizmov.	*„slnieckar jak vysity-este aj ten ksicht...“*

Príspevok obsahuje negáciu, slová alebo frázy, ktoré naznačujú negatívny alebo protichodný význam. Zahŕňa odmietnutie alebo negáciu vyhlásenia, myšlienky alebo návrhu.	*„Len to nie.“*

Zdroj: [2, 18, 25], vlastné spracovanie autorov

Tabuľka č. 5: Prípady neutrálneho sentimentu v príspevkoch

Prípad	**Príklad**
Príspevok neprejavuje pozitívne alebo negatívne nálady, postoj alebo emocionálny stav autora príspevku.	„Ja dávam korenie, soľ a polohrubú muku.“
Príspevok predkladá objektívne informácie, obsahuje len výroky a citácie bez prepojenia na nálady, postoj alebo emocionálny stav autora príspevku.	„Bezplatný zber elektroodpadu aj v septembri.“
Príspevok obsahuje reklamný text, vrátane reklamného textu, ktorý obsahuje pozitívny alebo negatívny sentiment (tzv. clickbait).	„Viac než dve tretiny voličov, ktorí sa zúčastnili na utorňajších voľbách do Kongresu USA, nechcú, aby sa súčasný prezident Spojených štátov Joe Biden v roku 2024 opätovne uchádzal o post šéfa Bieleho domu. #zosveta #webta3“.
Príspevok obsahuje pozitívne aj negatívne slová a zároveň nie je možné jednoznačne určiť pozitívnu alebo negatívnu náladu, postoj alebo emocionálny stav autora príspevku.	„Juraj Blanár si myslí = Fico si myslí... Podľa mňa štandardná debata, každý mlel to svoje. Vrátane Kovačiča.... Takže za mňa ok, nič nové..“

Zdroj: [2, 18, 25], vlastné spracovanie autorov

Všeobecný postup určovania sentimentu sa uplatňoval aj pri špeciálnych prípadoch na základe dostupnej literatúry [2, 18, 25], ktoré nie sú inherentné v uvedených tabuľkách. Cieľom nebolo ponúknuť anotátorom podrobné a komplikované pokyny – tie môžu byť dokonca kontraproduktívne, pretože anotátori nemusia rozumieť príslušným jemným odchýlkam alebo nemusia mať sklon pochopiť ich. Všeobecný postup určovania sentimentu sa zhrnul do šiestich bodov A až F:

A. Emocionálny stav autora príspevku: Emocionálny stav autora môže, ale nemusí mať rovnakú polaritu ako jeho vyjadrený názor. Napríklad príspevok politika môže naznačovať negatívny názor na minulú indiskrétnosť súpera a pozitívny emocionálny stav, pretože správy nepriaznivo ovplyvnia súpera. – Odporúčanie pre anotátora: Ak nie je zrejmý emocionálny stav autora príspevku, hodnotíme príspevok ako neutrálny.
B. Úspech alebo neúspech jednej strany: Vety často opisujú úspech alebo zlyhanie jednej strany vo vzťahu k druhej. Vnímanie týchto udalostí ako pozitívne alebo negatívne závisí od podpory príslušných zainteresovaných strán. Napríklad, keď Fínsko porazilo Rusko v hokeji, táto udalosť bola prevažne koncipovaná ako „Rusko prehralo s Fínskom“ a nie „Fínsko porazilo Rusko“. Toto rozlíšenie nebolo spôsobené negatívnym názorom na ruský tím, ale skôr preto, že Rusko ako hostiteľský štát získalo väčšiu pozornosť a ich hokejové tímy boli tradične silné. – Odporúčanie pre anotátora: Podobné príspevky hodnotíme ako neutrálne.
C. Sarkazmus: Sarkazmus môže často sprostredkovať pozitívny emocionálny stav rečníka, odvodený od aktu zosmiešňovania niekoho alebo niečoho. Tento pozitívny emocionálny stav však nemusí nutne znamenať pozitívny postoj k zosmiešňovanému subjektu. – Odporúčanie pre anotátora: Ak zo sarkazmu nie je možné odvodiť emocionálny stav autora (POZ alebo NEG), je potrebné hodnotiť sarkazmus ako neutrálny sentiment.

D. Prosby a žiadosti: Niektoré príspevky vyjadrujú pozitívne prosby k Bohu alebo pozitívne prosby ľuďom v kontexte (zvyčajne) negatívnej situácie. Napríklad: „Nech Boh pomáha tým, ktorí sú vysídlení vojnou.“ – Odporúčanie pre anotátora: Prosby a žiadosti v kontexte negatívnej situácie hodnotíme ako negatívny sentiment.
E. Rečnícke otázky: Rétorické otázky často nesú silný emocionálny podtón. Môžu sa použiť na zdôraznenie určitého bodu alebo na vyvolanie emocionálnej reakcie čitateľa alebo poslucháča. Rétorické otázky sa bežne používajú aj na vyjadrenie sarkazmu alebo irónie a vyjadrený sentiment môže byť v rozpore s doslovným významom otázky. – Odporúčanie pre anotátora: S rétorickými otázkami možno zaobchádzať jednoducho ako s otázkami (a teda neutrálne) alebo ako s výrokmi, ktoré prezrádzajú emocionálny stav hovoriaceho.
F. Emodži: Vplyv emotikonov na vnímaný sentiment je ovplyvnený individuálnou interpretáciou, kultúrnymi normami, kontextovými podnetmi a špecifickou kombináciou emotikonov a textu. Emodži sa niekedy tiež používajú na neutralizáciu alebo zmiernenie sentimentu vyjadreného v texte. Napríklad pridanie emodži žmurkajúcej tváre po potenciálne negatívnom vyhlásení môže zmierniť vnímaný sentiment a naznačiť, že to bolo myslené ako vtip alebo odľahčený komentár, a nie ako vyjadrenie skutočnej negativity. – Odporúčanie pre anotátora: vplyv emodži na celkový sentiment je na posúdenie anotátora. Emodži by nemalo mať väčšiu váhu ako samotný text.

4.3. Index sociálneho napätia

Kombinácia reálnych a umelo vytvorených dát v slovenskom jazyku spolu s multilingválnymi a spoločne anotovanými údajmi a zahrnutie analýzy emotikonov viedli k vytvoreniu modelu s vysokou úrovňou správnosti (87 %). Prístup neurónovej siete dosahuje na prvý pohľad horšie výsledky ako základný model SVM (87 % verzus 90 %). Avšak takto jednoducho nemožno dva modely porovnávať, pretože neboli vytvorené a aplikované na rovnakom datasete. Aj model založený na neurónovej sieti XLM-RoBERTa-large je schopný rozlišovať aj tretiu kategóriu neutrálneho príspevku, čo je zložitejšie, ako len binárny klasifikátor.

Kombináciou modelu na klasifikáciu príspevkov z pohľadu témy a modelu na klasifikáciu príspevkov z pohľadu sentimentu bol získaný textový dataset obsahujúci príspevky a komentáre v jednotlivých témach prevedené na číselný časový rad, čo dovoľuje vytvoriť index sociálneho napätia nasledujúcim postupom:

– Krok 1: Skóre pravdepodobnosti (prahovú hodnotu) určujúce či sa príspevok týka témy, bolo stanovené na základe pozorovania menšej vzorky referenčných dát na 0,7. Pri všetkých príspevkoch (POST) bol zaznačený zoznam tém ako výstup z klasifikátora za predpokladu, že príspevok patril do vybranej podmnožiny tém a súčasne daná téma dosiahla skóre vyššie ako 0,7. Komentáre (COMMENT) boli priradené do istej témy ako príspevky, ku ktorým patrili. Ostatné príspevky, ktoré boli neisté, sa v ďalších analýzach nepoužili. Jeden príspevok však mohol patriť naraz aj do viacerých tém, pokiaľ splnil podmienky výberu (išlo o „mutli-label classification“).

– Krok 2: Pre každú zo zvolených kategórii sa spočítal denný priemer sentimentu. Každému príspevku alebo komentáru sa priradila číselná hodnota na základe sentimentu, ktorý priradil klasifikátor, pričom pozitívny sentiment mal hodnotu 1, negatívny -1 a neutrálny 0. Tieto hodnoty sentimentu sa v každej vybranej kategórii

spočítali za daný deň a vydelili denným počtom príspevkov v danej kategórii. Rovnaký výpočet za každý deň sa potom vykonal celkovo pre všetky príspevky, ktoré boli aspoň v jednej zo zvolených kategórií. Pokiaľ sa príspevok nachádzal vo viacerých zvolených kategóriách, do výpočtu vstupoval iba raz.

Postupom opísaným v uvedených krokoch sa získal celkový číselný denný priemer sentimentu ako aj číselný denný priemer sentimentu vybraných kategórií. Tieto číselné denné priemery sentimentu označujeme ako index sociálneho napätia, ktorý môže dosahovať hodnoty od 1 (dokonalo pozitívne naladená spoločnosť) do -1 (úplne negatívne naladená spoločnosť). Teda ide o číselný časový rad tohto indexu za sledované obdobie od 1. 9. 2022 do 31. 8. 2023, kde premenná času je v rozčlenení na dni s konkrétnym označením dátumu. Pre detekciu náhlych zmien indexu sentimentu, ktoré môžu indikovať významné spoločenské udalosti, bola určená prahová hodnota ako zmena sentimentu v danom dni, ktorá je väčšia ako 25 % 14-dňového plávajúceho priemeru. 14-dňové obdobie priemerovania sa zvolilo ako kompromis. Kratšie obdobie citlivejšie reagovalo na zmeny, no častejšie označovalo „návrat“ do normálu aj ako významnú zmenu. Dlhšie obdobie lepšie eliminovalo „návraty“, ale bolo menej citlivé na drobnejšie výkyvy.

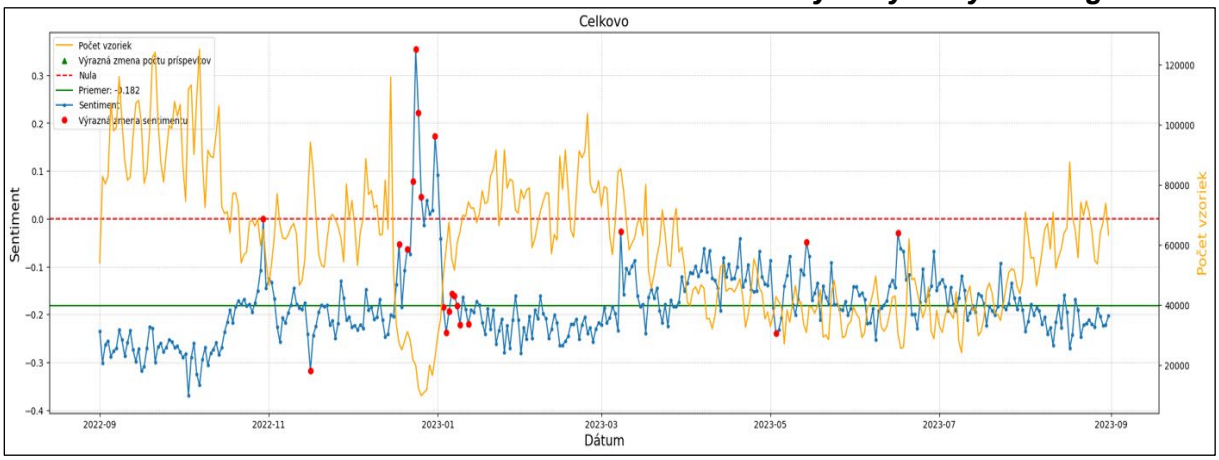
5. ANALÝZA VÝSLEDKOV

Obrázok č. 1 znázorňuje priebeh indexu sentimentu a počtu vzoriek (príspevkov a komentárov) za sledované obdobie od 1. 9. 2022 do 31. 8. 2023 vo všetkých vybraných kategóriách. Priemerný počet príspevkov a komentárov, ktoré boli zohľadnené v tejto analýze, je **59 831** za deň. Celkový analyzovaný počet príspevkov a komentárov predstavuje skoro 32 miliónov príspevkov. Na tomto obrázku ako aj na ďalších obrázkoch vidno aj zelenú vodorovnú čiaru znázorňujúcu priemernú hodnotu indexu sociálneho napätia, ako aj červenú prerušovanú vodorovnú čiaru, ktorá znázorňuje neutrálny sentiment. Zelený trojuholník znázorňuje výraznú zmenu počtu príspevkov, červený kruh zas výraznú zmenu v indexe sociálneho napätia. Len obrázok č. 1 zachytáva isté obdobie okolo vianočných sviatkov, keď je index nad nulou, teda pozitívny. Môže k tomu prispievať pozitívna sviatočná nálada, avšak tento vzostup preukazuje aj najviac zastúpená kategória politika (obrázok č. 2). Dňa 23. 12. 2022 skončil minister financií Igor Matovič vo svojej funkcii ministra. Vo všetkých obrázkoch však vidíme výrazný pokles počtu príspevkov vo vianočnom období pred ich výrazným vzostupom 24. 12. 2022.

Obrázok č. 3 ukazuje situáciu zo začiatku októbra 2022, keď sa stala udalosť usmrtenia ľudí na zastávke Zochova vodičom pod vplyvom alkoholu a neskôr nato teroristický čin pred podnikom Tepláreň. Model zachytil výrazné zmeny v počte príspevkov aj poklese indexu sociálneho napätia hlboko pod priemerom.

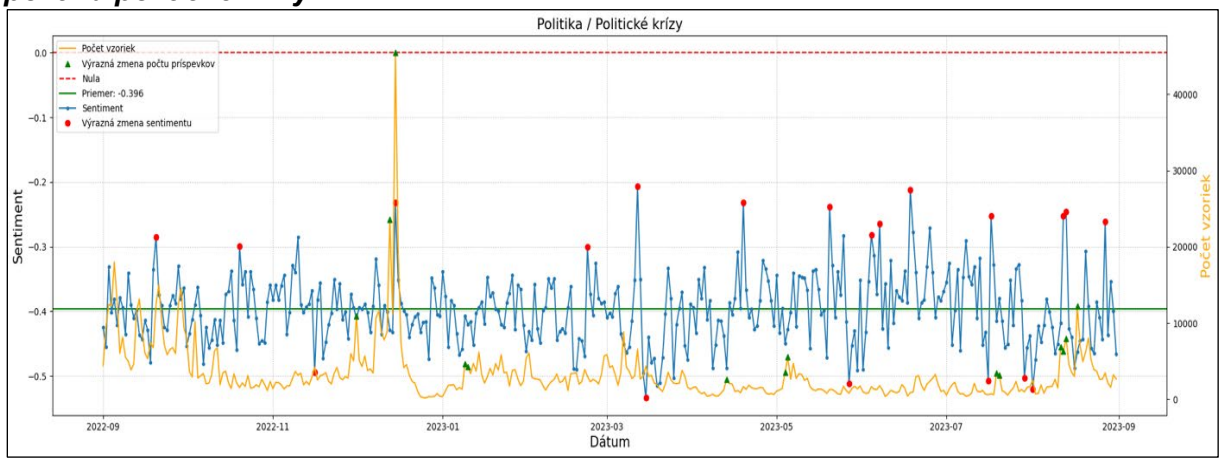
Obrázok č. 4 indikuje výročie vojny na Ukrajine, keď 24. 2. 2023 sledujeme výrazný nárast počtu príspevkov na vyše 13 400, pritom bežne ich je okolo 3- až 5-tisíc. Index sa však už po roku konfliktu drží blízko svojej priemernej hodnoty bez zásadnej zmeny.

Obrázok č. 1: Priebeh indexu sentimentu a počtu vzoriek (príspevkov a komentárov) za sledované obdobie od 1. 9. 2022 do 31. 8. 2023 vo všetkých vybraných kategóriách



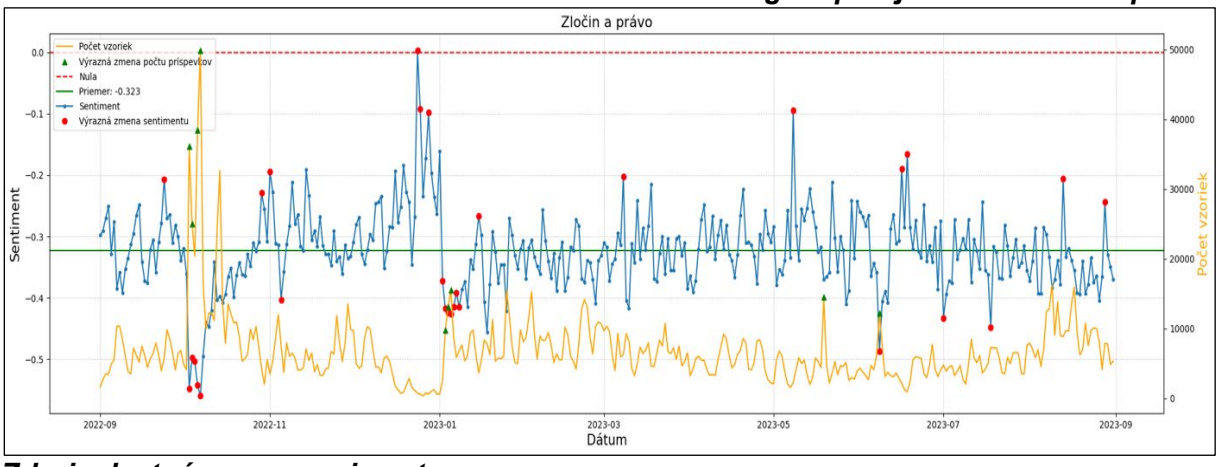
Zdroj: vlastné spracovanie autorov

Obrázok č. 2: Priebeh indexu sentimentu a počtu vzoriek (príspevkov a komentárov) za sledované obdobie od 1. 9. 2022 do 31. 8. 2023 v kategórii druhej úrovne politika/politické krízy



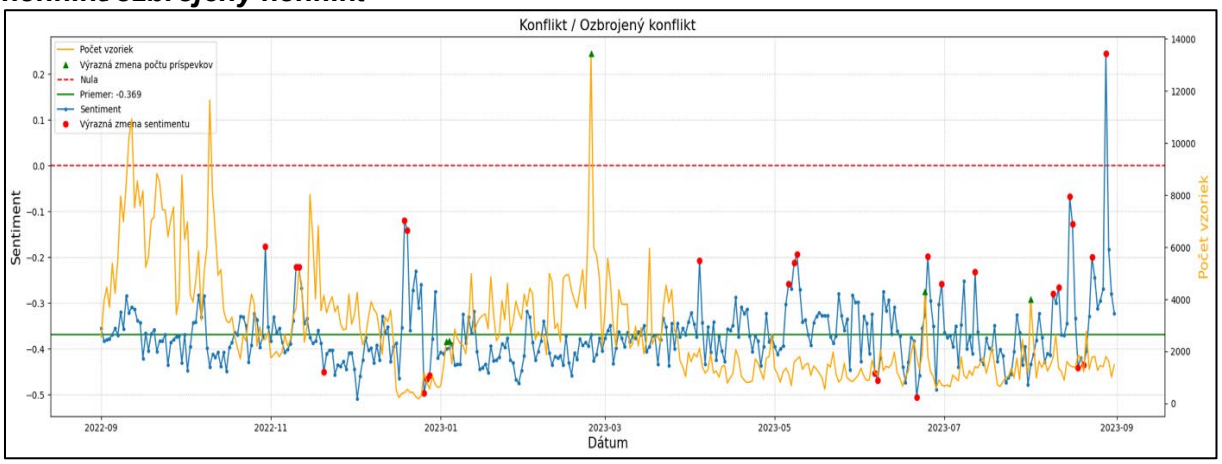
Zdroj: vlastné spracovanie autorov

Obrázok č. 3: Priebeh indexu sentimentu a počtu vzoriek (príspevkov a komentárov) za sledované obdobie od 1. 9. 2022 do 31. 8. 2023 v kategórii prvej úrovne zločin a právo



Zdroj: vlastné spracovanie autorov

Obrázok č. 4: Priebeh indexu sentimentu a počtu vzoriek (príspevkov a komentárov) za sledované obdobie od 1. 9. 2022 do 31. 8. 2023 v kategórii druhej úrovne konflikt/ozbrojený konflikt



zdroj: vlastné spracovanie autorov

Napriek tomu, že využitý model na klasifikáciu nevieme úplne priamo porovnať s alternatívnymi modelmi v literatúre, je nepísaným konsenzom, že správnosť nad 85 % (v našom prípade 87 %) je veľmi dobrým výsledkom. Jazykové modely založené na transformátoroch s miliónmi parametrov, akým je použitý model XLM-RoBERTa, sú aktuálne štandardom v súkromnej sfére pre spracovanie prirodzeného jazyka vďaka svojej multijazyčnosti a relatívnej jednoduchosti pre ich dotrénovanie na špecializované úlohy.

Výsledky spracované v tejto kapitole potvrdzujú použiteľnosť verejne dostupných príspevkov na sociálnych sieťach na monitorovanie nálad a názorov na zásadné udalosti v spoločnosti. Aktuálna definícia indexu sociálneho napätia a sledovanie jeho zmeny umožnili identifikovať zásadné udalosti, ktoré vychýlili náladu v spoločnosti z priemeru. Pre hlbšie porozumenie a predpovedanie zintenzívňovania sociálneho napätia a jeho prejavov vo fyzickom svete sú tieto výsledky len prvým, nevyhnutným krokom.

6. ZÁVER

Sociálne siete všeobecne poskytujú údaje o pomerne veľkej vzorke obyvateľstva rôznorodej demografie v pomerne širokom geografickom pokrytí takmer v reálnom čase. Analýza sociálneho sentimentu z príspevkov zo sociálnych sietí preto môže pomôcť zachytiť dynamické zmeny, pochopiť rôzne demografické perspektívy a identifikovať regionálne trendy, ktoré môžu byť užitočné napríklad ako dodatočný zdroj informácií pre oficiálnu štatistiku o kvalite života.

Výberom modelu XLM-ROBERTA-large sme získali výhody výkonnejšieho a flexibilnejšieho modelovania, ktoré nám umožnilo úspešne adaptovať model na špecifické úlohy a jazykové nuansy. Multilingválny model, umožnil integráciu a analýzu dát z rôznych jazykových prostredí, čo bolo kľúčové pri začleňovaní multilingválnych datasetov do experimentov. Keďže textov a datasetov v rôznych jazykoch je dostupných viac, je možné týmto multilingválnym prístupom zvyšovať objektivitu modelu. Metódy trénovania zahŕňali použitie reálnych, umelých a multilingválnych datasetov, ktoré umožnili preskúmať a identifikovať optimálne

kombinácie pre tréningový proces. To zase viedlo k vytvoreniu modelu s vysokou úrovňou presnosti a robustnosti.

Aj vzhľadom na výsledky je správnosť výstupov dostatočná – na úrovni 87 percent. Je však ešte možné ju ďalej ladiť a vylepšovať cez lepšie nastavenie prahových hodnôt na priradenie daného textu do vybranej triedy, a to postupom cez ROC krivku a metriku AUC. Priestor na zlepšenia tiež ostáva v otázke, ako interpretovať populáciu, ktorej sa daný štatistický produkt týka, a ako ho využívať v praxi.

Projekt overil, že modely založené na modeli RoBERTa vynikajú v úlohách vyžadujúcich hlboké porozumenie kontextu, ako je analýza sentimentu, odpovedanie na otázky a rozpoznávanie pomenovaných entít. Ide o predtrénovaný model s otvoreným zdrojovým kódom, ktorý by sa dal nasadiť a ladiť aj v prostredí Štatistického úradu SR alebo štátnej správy. Model XLM-ROBERTA-large preukázal schopnosť zachytiť jemné jazykové nuansy, najmä pri analýze sentimentu, kde je identifikácia sarkazmu a komplexných emócií kritická.

Na druhej strane monitorovanie sentimentu verejnosti na sociálnych médiách pomocou algoritmov strojového učenia predstavuje niekoľko rizík. Trottier [24] argumentuje, že vznikajú obavy o súkromie, pretože jednotlivci môžu mať pocit, že ich sloboda prejavu a právo na súkromie sa porušujú, keď vládne inštitúcie a úrady monitorujú ich aktivity na sociálnych sieťach. To môže viesť k autocenzúre a negatívnemu vplyvu na slobodu prejavu. Nesprávna interpretácia sentimentu je ďalším rizikom, pretože algoritmy môžu mať problémy s určením sentimentu niektorých typov príspevkov, ako napríklad sarkazmus, rečnícke otázky a emodži [12], čo môže viesť ku skreslenej interpretácii a nesprávnym rozhodnutiam. Zaujaté rozhodovanie v oblasti algoritmov sa navyše môže neprimerane zameriavať na určité skupiny alebo komunity [29], čím sa udržiava diskriminácia a prehlbujú sa sociálne nerovnosti. Nedostatočná transparentnosť tieto riziká ešte viac znásobuje, keďže vlády nemusia zverejniť rozsah a účel monitorovacích aktivít, čo sťažuje posúdenie spravodlivosti a etiky.

LITERATÚRA

[1] ANTELMI, A. Towards an exhaustive framework for online social networks user behaviour modelling. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization. 2019, s. 349 – 352.
[2] BAI, Q. – DAN, Q. – MU, Z. – YANG, M.: A Systematic Review of Emoji: Current Research and Future Perspectives. In: Frontiers in Psychology. 2019, č. 10.
[3] BAO, H. – DONG, L. – WANG, W. et al.: Fine-tuning pretrained transformer encoders for sequence-to-sequence learning. In: International Journal of Machine Learning and Cybernetics. 2023, č. 11.
[4] CONNEAU, A. – KHANDELWAL, K. – GOYAL, N. – CHAUDHARY, V. – WENZEK, G., GUZMÁN, F. et al.: Unsupervised cross-lingual representation learning at scale. In: Annual Meeting of the Association for Computational Linguistics, 2019.
[5] DAHRENDORF, R.: Toward a theory of social conflict. In: Journal of Conflict Resolution. 1958, č. 2. s. 170 – 183.
[6] EO, S. – PARK, C. – MOON, H. – SEO, J. – LIM, H.: Comparative Analysis of Current Approaches to Quality Estimation for Neural Machine Translation. In: Applied Sciences. 2021, č. 14.

[7] FEIZOLLAH, A. et al.: Halal products on Twitter: Data extraction and sentiment analysis using stack of deep learning algorithms. In: IEEE Access. 2019, č. 7. s. 83354 – 83362.
[8] GAZZAH, S. – ESSOUKRI BEN AMARA, N. Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script. In: International Arab Journal of Information Technology. 2008, č. 1, s. 92 – 101.
[9] Go4insight: Koľko Slovákov je na sociálnych sieťach v roku 2022? [online]. [cit. 7-11-2023]. Dostupné na: https://www.go4insight.com/post/ko%C4%BEko-slov%C3%A1kov-je-na-soci%C3%A1lnych-sie%C5%A5ach-v-roku-2022
[10] HARTL, P.: Psychologický slovník. Jiří Budka, Praha 1993, ISBN 80-90 15 49-0-5.
[11] CHAMLERTWAT, W. et al.: Discovering Consumer Insight from Twitter via Sentiment Analysis. In: Journal of Universal Computer Science. 2012, č. 18. s. 973 – 992.
[12] SINGH, K. – DEEPAK T. – ARUN S. Sentiment analysis: a review and comparative analysis over social media. In: Journal of Ambient Intelligence and Humanized Computing. 2020, č. 11. s. 97 – 117.
[13] KORA, R. – MOHAMMED, A.: An enhanced approach for sentiment analysis based on meta-ensemble deep learning. In: Social Network Analysis and Mining. 2023, č. 38.
[14] LAURA, A. – MORO, R. A., Support Vector Machines (SVM) as a Technique for Solvency Analysis. In: DIW Berlin Discussion Paper. 2008, č. 811, s. 2.
[15] LEE, J. et al.: Sentiment analysis of Twitter users over time: the case of the Boston bombing tragedy. In: E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life: 15th Workshop on e-Business, WEB 2015, Fort Worth, Texas, USA, December 12, 2015, Revised Selected Papers. 2015.
[16] LIU, B.: Sentiment analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, 2015.
[17] MAHMUD, J. et al.: Predicting attitude and actions of Twitter users. In: Proceedings of the 21st International Conference on Intelligent User Interfaces. 2016, s. 2 – 6.
[18] MOHAMMAD, S.: A Practical Guide to Sentiment Analysis. In: Socio-Affective Computing. 2017, č. 5.
[19] NAKONEČNÝ, M.: Emoce. Praha: Triton, 2012. ISBN: 978-80-7387-614-2.
[20] NASEER, K. et al. Travel behaviour prediction amid covid-19 underlaying situational awareness theory and health belief model. In: Behaviour & Information Technology. 2022, č. 15. s. 3318 – 3328.
[21] NEPPALLI, K. et al.: Sentiment analysis during Hurricane Sandy in emergency response. In: International journal of disaster risk reduction. 2017, č. 21. s. 213 – 222.
[22] PIKULIAK, M. – GRIVALSKÝ, Š. – KONÔPKA, M. – BLŠTÁK, M. – TAMAJKA, M. – BACHRATÝ, V. – UHLÁRIK, F. et al.: SlovakBERT: Slovak Masked Language Model. In: Findings of the Association for Computational Linguistics: EMNLP 2022. 2022. s. 7156 – 7168.
[23] PLUTCHIK, R - KELLERMAN, H. (ed.): Theories of emotion. Academic press, 2013.
[24] TROTTIER, D.: Social media as surveillance: Rethinking visibility in a converging world. Routledge, 2016.
[25] QURESHI, M. – ASIF, M. – HASSAN, M. et al.: A Novel Auto-Annotation Technique for Aspect Level Sentiment Analysis. In: Computers, Materials & Continua. 2022, č. 3. s. 4987-5004.

[26] RADICIONI, T. – SARACCO, F. – PAVAN, E. et al.: Analysing Twitter semantic networks: the case of 2018 Italian elections. In: Scientific Reports. 2021, č. 11.
[27] SMELSER, N.J.: The Theory of Collective Behavior. New York: Free Press, 1962. ISBN 9781136277900.
[28] YAVARI, A. et al. Election prediction based on sentiment analysis using twitter data. In: International Journal of Engineering. 2022, č. 2. s. 372 – 379.
[29] ZEITZOFF, T.: How social media is changing conflict. In: Journal of Conflict Resolution. 2017, č. 9. s. 1970 – 1991.

RESUMÉ
V dnešnej dobe sú sociálne siete čoraz populárnejšie pri vyjadrovaní názorov a emócií ľudí, ktorí často verejne zdieľajú svoje pocity a názory na rôzne udalosti a tematicky súvisiace správy. Táto popularita sociálnych médií ako miesta, kde sa ľudia vyjadrujú, vytvára veľké množstvo dát týkajúcich sa nálady ľudí a sociálneho napätia. Cieľom tohto experimentu bola práca s verejnými príspevkami v slovenčine na sociálnej sieti Facebook a analýza sentimentu týchto príspevkov k rôznym témam. Sledovanie sentimentu môže poskytnúť základné pochopenie emocionálneho stavu ľudí, čo sa môže ďalej použiť v rôznych oblastiach, vrátane štatistiky o kvalite života. Použitá metóda analýzy príspevkov s cieľom identifikovať ich kategóriu a získať kvantitatívnu hodnotu sentimentu obsiahnutého v textoch využíva multilingválny model XLM-ROBERTA-large pre spracovanie prirodzeného jazyka. Vytvorený model dovoľuje korelovať identifikované sociálne napätie s udalosťami zo skutočného sveta. Napríklad po teroristickom útoku na Zámockej ulici v Bratislave sa v komentároch objavuje zvýšený negatívny sentiment, ktorý signalizuje zvýšenie napätia. Analýza sentimentu príspevkov na sociálnych sieťach poskytuje dodatočné a rýchle zdroje informácií o sociálnom napätí verejnosti, čo môže mať význam napríklad pre tvorcov politík a sociológov. Potenciál pre budúci výskum zahŕňa možnosť rozšírenia tejto analýzy sentimentu na iné platformy a na iné jazyky.

RESUME
Nowadays, social networks are increasingly popular for expressing the opinions and emotions of people, who often publicly share their feelings and opinions about various events and thematically related news. This popularity of social media as a place for people to express themselves creates a large amount of data regarding people's moods and social tensions. The aim of this experiment was to deal with the public posts in Slovak on the social networking website Facebook and to analyze the sentiment of these posts on various topics. Sentiment monitoring can provide a basic understanding of people's emotional state, which can be further used in various fields, including quality of life statistics. The method applied for analyzing the posts in order to identify their category and obtain a quantitative value of the sentiment in the texts uses the XLM-ROBERTA-large multilingual model for natural language processing. The created model allows to correlate the identified social tension with real-life events. For example, after the terrorist attack on Zámocká street in Bratislava, an increased negative sentiment appears in the comments, indicating a higher social tension. The sentiment analysis of social media posts provides additional and rapid sources of information about public social tensions, which may be relevant to policymakers and sociologists, for example. Potential for a future research includes broadening this sentiment analysis to other platforms and languages.

PROFESIJNÝ ŽIVOTOPIS

Dipl. Ing. Dagmar Ceľuchová Bošanská *je zakladateľkou spoločnosti Alistiq s. r. o. a expertkou na inovácie a digitálnu transformáciu s dlhoročnými skúsenosťami. V roku 2008 absolvovala inžinierske štúdium pre informačné technológie, mobilné komunikácie a štatistické spracovanie signálov na Viedenskej technickej univerzite, kde pôsobila aj vo vedeckom tíme na vývoji simulátorov technológií pre bezdrôtové siete štvrtej generácie. Od roku 2015 sa venuje vývoju riešenia a návrhu opatrení na zvyšovanie kvality a efektivity využívania údajov vrátane Big Data na sekundárne účely, predovšetkým vo verejnej správe. Aktuálne od roku 2020 pôsobí ako doktorand na Českom vysokom učení technickom v Prahe, kde sa venuje výskumu grafových údajov generovaných z elektronických zdravotných záznamov a ich analýze s využitím strojového učenia a veľkých jazykových modelov.*

Ing. Martin Janík *absolvoval inžinierske štúdium na Fakulte elektrotechniky a informatiky Slovenskej technickej univerzity v Bratislave v odbore telekomunikácie, špecializácia tezpečnosť (2008). Už počas štúdia na vysokej škole začal pracovať v súkromnom sektore v oblasti IT, spočiatku v oblasti webových, neskôr mobilných technológií ako programátor, analytik a následne softvérový architekt softvérových produktov v oblasti mobilných technológií. Od roku 2022 sa venuje dátovej vede so zameraním na analýzu a návrh grafových dátových štruktúr vo verejnom sektore. Spolupracuje na centrálnom modeli údajov SR a na analýze a spracovaní Big Data.*

Mgr. Filip Nguyen *absolvoval magisterské štúdium v Ústave pedagogiky a sociálnych štúdií Univerzity Palackého v Olomouci v odbore pedagogika – veřejná správa (2019). Od roku 2018 pôsobí v poradenskej spoločnosti Alistiq, s. r. o. ako poradca v oblasti verejného obstarávania a verejných inovačných projektov. Jeho práca sa zameriava na návrh digitálnych služieb štátnej správy a aplikáciu osvedčených postupov PRINCE2 a Agile metodík v projektoch.*

KONTAKT

dagmar.bosanska@alistiq.com
martin.janik@alistiq.com
filip.nguyen@alistiq.com

Informácia/Information

KONFERENCIA A PREDSTAVENIE PRÍRUČKY O POSILŇOVANÍ KAPACITY ÚDAJOV PRE DETI V POHYBE

CONFERENCE AND PRESENTATION OF THE MANUAL ON CHILD-SPECIFIC DATA CAPACITY-STRENGTHENING ON CHILDREN ON THE MOVE

V dňoch 20. – 21. 11. 2023 metropola Grécka – Atény pod záštitou Gréckeho štatistického úradu ELSTAT usporiadala podujatie pri príležitosti vydania príručky o posilnení zberu a spracovania špecifických údajov o deťoch v pohybe (*Manual on Child-Specific Data Capacity-Strengthening on Children on the Move*). Príručka je výsledkom spoločného úsilia viacerých členov patriacich do Medzinárodnej aliancie údajov pre deti v pohybe (IDAC – International Data Alliance for Children on the Move). Práve IDAC sa svojím pôsobením v medzinárodnom meradle snaží o to, aby sa vysídlené deti a deti migrantov štandardne nielen započítavali do oficiálnych štatistík, ale aby sa s touto zraniteľnou skupinou počítalo pri nastavovaní verejných politík v jednotlivých krajinách. V tomto duchu sa niesli obidva dni konferencie, ktorú svojou prítomnosťou poctili aj predsedovia či riaditelia niektorých štatistických úradov EÚ (Bulharsko, Chorvátsko, Estónsko, Lotyšsko, Malta, Slovensko), zástupca UNICEF-u v Grécku, či ďalší zástupcovia z medzinárodných organizácií ako UNHCR, IOM, OECD, UNICEF alebo ústredných vládnych orgánov Grécka. Okrem nosnej témy predstavenia príručky bola ďalšou významnou témou otázka dôležitosti oficiálnych štatistík v spoločnosti a výzvy, ktoré prináša dezagregácia údajov a ich ochrana v národných štatistických systémoch. V rámci tejto problematiky sa predseda Štatistického úradu SR Peter Peťko aktívne zúčastnil panelovej diskusii.

Konferenciu otvoril predseda Gréckeho štatistického úradu Athanasios C. Thanopoulos s príhovorom, v ktorom zdôraznil, že migrácia a vysídľovanie ovplyvňuje celosvetovo podľa odhadov až 35,5 milióna detí. Poukázal tiež na to, že údaje z oficiálnych zdrojov ako sú štatistické úrady, by mali zohrávať kľúčovú úlohu v plnení medzinárodných záväzkov, ktoré chránia práva detí a nastavujú ich ochranu a integráciu. V programe konferencie bola odprezentovaná úloha aliancie IDAC v systéme zberu a spracovania údajov o migrujúcich deťoch. IDAC je medzisektorová koalícia, ktorej členmi sú Eurostat a niektoré členské krajiny Eurostatu, Medzinárodná organizácia pre migráciu (IOM), Organizácia pre hospodársku spoluprácu a rozvoj (OECD), Detský fond OSN (UNICEF) a ďalšie. IDAC poskytuje širokú škálu iniciatív vrátane tvorby príručiek, usmernení či webinárov. V spolupráci s gréckym úradom bol postavený do čela pracovnej skupiny, ktorá stála za vytvorením príručky. Za účasti

Zdroj: Manual on Child-Specific Data Capacity-Strengthening on Children on the Move

ostatných relevantných členov tejto pracovnej skupiny (zástupcovia UNICEF-u, IOM, UNHCR a niektorých ústredných orgánov štátnej správy v Grécku – ministerstvo ekonomiky a financií, ministerstvo pre migráciu a azyl) spoločne uviedli základné informácie o príručke a načrtli problémy, pre ktoré je potrebný spoločný postup či metodika pre krajiny, ktorých sa daný typ migrácie dotýka najviac. Príručka podľa ich slov ponúka analytické usmernenia pre štatistické úrady a ostatných aktérov o tom, ako pracovať s údajmi o deťoch v pohybe v súlade so základnými princípmi oficiálnych štatistík podľa OSN. Podľa odprezentovaných informácií, mnohé krajiny nezbierajú údaje o migračných tokoch a rozčlenené údaje o migračných tokoch podľa veku a pohlavia sú dokonca veľmi zriedkavé. Napríklad 3 z 10 krajín neprodukujú údaje o migrantoch rozčlenené podľa veku a 9 z 10 krajín, ktorých sa dotýka vnútorne vysídlenie osôb v súvislosti s konfliktmi, nie sú členené podľa veku. Príčinou sú často rozlične zbierané údaje rôznymi inštitúciami a za použitia rôznych metodík. Predkladatelia príručky za jednu z kľúčových úloh preto pokladajú úlohu, nastaviť jednotný rámec na zber týchto údajov. Predstavenie príručky vrátane príkladov postupu jednotlivých krajín odzneli postupne počas dvoch dní konferencie, svoje skúsenosti prezentovali zástupcovia z Čile, Maroka, Talianska a Grécka.

V nadväznosti na danú problematiku sa počas konferencie konali dve kolá panelovej diskusie, na tému *Dôležitosť oficiálnych štatistík a výzvy v oblasti dezagregácie údajov v národných štatistických systémoch*. V prvom dni sa na diskusii zúčastnili predsedovia úradov z Litvy, Chorvátska a Bulharska, druhý deň to boli predsedovia z Estónska, Malty, opäť Bulharska a Slovenska.

Predseda Štatistického úradu SR Peter Peťko sa spolu s ostatnými kolegami zhodli na tom, že oficiálne štatistiky sú dôležitým nástrojom v procese tvorby oficiálnych politík a mali by pomôcť identifikovať oblasti, v ktorých je potrebná pomoc a podpora. Predseda Peťko hovoril aj o stratégii, ktorá by mala byť prítomná medzi jednotlivými aktérmi poskytujúcimi štatistiky, od národných úradov až po neziskové organizácie. Medzi jednotlivými aktérmi musí fungovať vzájomná spolupráca a zdieľanie

Zdroj: vlastné fotografie autorov

vedomostí, metodík a v neposlednom rade aj samotných údajov a to všetko so zachovaním ich dôvernosti. Poukázal na skutočnosť, že v SR je dostupnosť štatistík týkajúcich sa zahraničnej migrácie rozdrobená medzi viacero rezortov (Štatistický úrad SR, Migračný úrad Ministerstva vnútra SR, Úrad hraničnej a cudzineckej polície Prezídia Policajného zboru, Ministerstvo zahraničných vecí a európskych záležitosti SR, sekcia verejnej správy Ministerstva vnútra SR, Ministerstvo školstva, vedy výskumu a športu SR, Úrad práce, sociálnych vecí a rodiny SR), v dôsledku čoho je koordinácia zberu údajov už aj na národnej úrovni veľmi náročná. Diskutujúci sa tiež zhodli na chýbajúcich štandardoch v klasifikáciách na jednotlivých úrovniach a potrebe vyššej miery harmonizácie.

Závery konferencie a predstavenie manuálu by mohli efektívne zvýšiť povedomie o problematike migrácie detí a podporiť výmenu nápadov, poznatkov a metód na vylepšenie spracúvania údajov o deťoch v kontexte migrácie a vysídľovania. Konferencia poskytla priestor na diskusiu o tom, ako by štatistické úrady mohli efektívnejšie spolupracovať, pričom skúsenosti krajín a ich osvedčené postupy v zbere a analýze údajov o migrácii detí boli cenným príspevkom k tejto problematike.

Ing. Peter PEŤKO, MBA
Autor je predseda Štatistického úradu SR.

Mgr. Martin KOČIŠ
Autor pôsobí v sekcii sociálnych štatistík a demografie Štatistického úradu SR.

PRIPRAVUJEME/COMING SOON

Boris VAŇO
VEK ŽIEN PRI PRVOM PÔRODE NA SLOVENSKU
AGE OF WOMEN AT FIRST BIRTH IN SLOVAKIA

Roman PAVELKA
STATISTICKÁ ANALÝZA CHYBĚJÍCÍCH DAT
STATISTICAL ANALYSIS OF MISSING DATA

* * *

ONLINE VERZIA ČÍSLA 1/2024 SLOVENSKEJ ŠTATISTIKY A DEMOGRAFIE JE VEREJNE DOSTUPNÁ *na internetovej stránke slovak.statistics.sk a ssad.statistics.sk od* ***15. JANUÁRA 2024****.*

THE ONLINE VERSION OF THE JOURNAL SLOVAK STATISTICS AND DEMOGRAPHY No 1 (2024) IS PUBLICLY BE AVAILABLE *at the website slovak.statistics.sk and ssad.statistics.sk from* ***JANUARY 15, 2024.***

INFORMÁCIE PRE PRISPIEVATEĽOV

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

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

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

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

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

Redakcia si vyhradzuje právo zverejniť články schválené redakčnou radou v tlačenej a elektronickej podobe na ssad.statistics.sk.

INFORMATION FOR AUTHORS

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

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

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

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

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

The editorial office reserves the right to publish articles approved by the editorial board in printed and electronic form at the website ssad.statistics.sk.

SLOVENSKÁ ŠTATISTIKA A DEMOGRAFIA

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

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

Vydáva:
Štatistický úrad SR

Identifikačné číslo vydavateľa:
IČO 00166197

Vychádza:
Štyrikrát ročne

Dátum vydania:
15. január 2024

Tlač:
Reprografické stredisko
Štatistického úradu SR

Predplatné:
20 € (na rok)
5 € (za jeden výtlačok)

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

Evidenčné číslo/Evidence number 272/08

SLOVAK STATISTICS AND DEMOGRAPHY

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

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

Issued by:
Statistical Office of the SR

Company registration number:
00166197

Published:
Four times a year

Date of issue:
15th January 2024

Press:
Reprographic centre of the
Statistical Office of the SR

Subscription:
€20 (per year)
€5 (for one copy)

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