SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

1/2024
ročník/volume 34

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

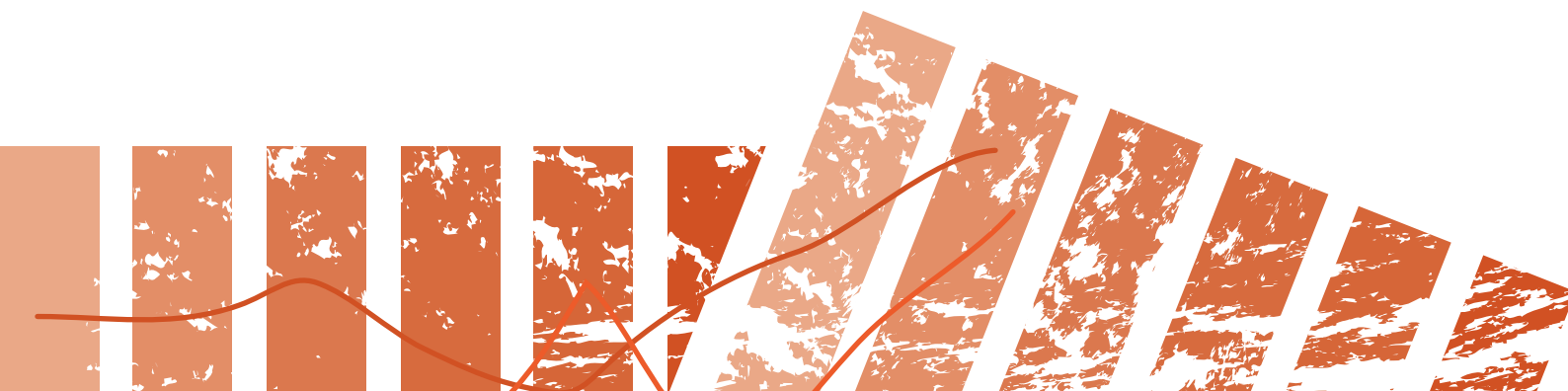
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 5

Typ článku/Type of article: informatívny článok/informative article

Strany/Pages: 52 – 57

Dátum vydania/Publication date: 15. január 2024/January 15, 2024



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