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: 7

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

Strany/Pages: 65 – 76

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

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