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: 6

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

Strany/Pages: 58 – 64

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

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

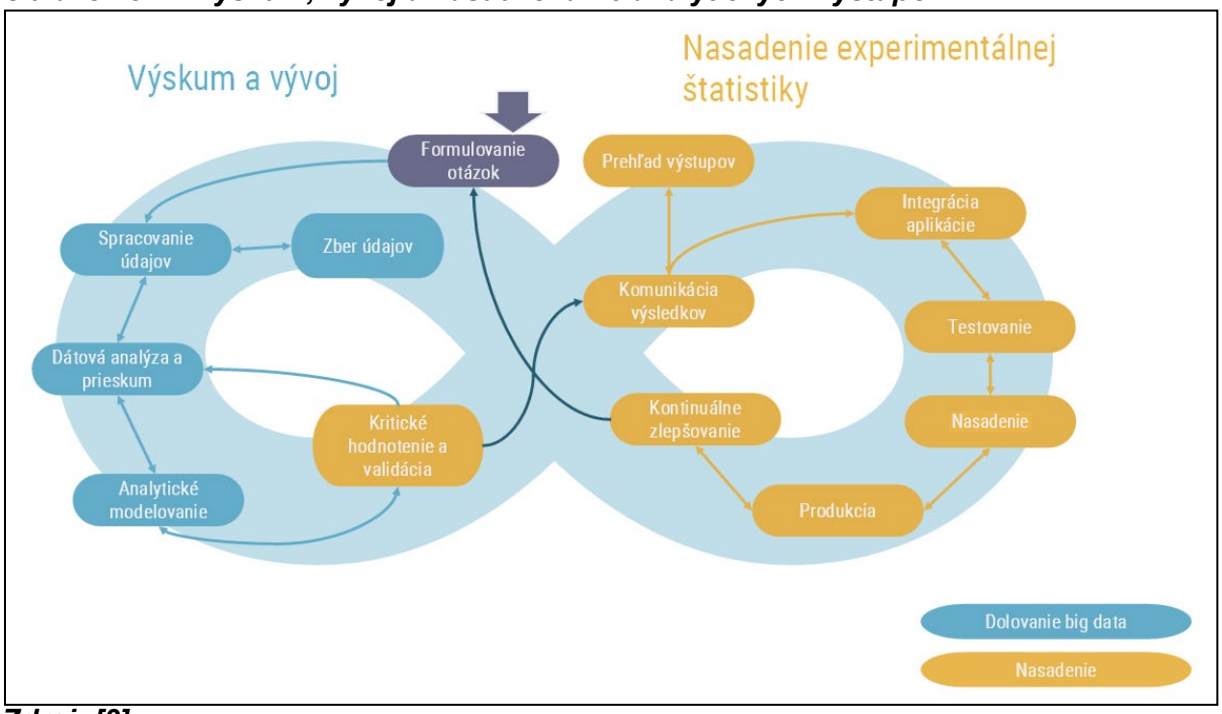
1. ÚVOD
Všeobecná definícia termínu Big Data od Tama a Clarkea [5] charakterizuje Big Data ako zdroje štatistických údajov zahŕňajúce tradičné zdroje aj nové zdroje, ktoré sa stávajú dostupnými z „webu všetkého“[2]. Európska komisia [2] definovala Big Data ako „*veľké množstvá údajov vyprodukovaných 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.

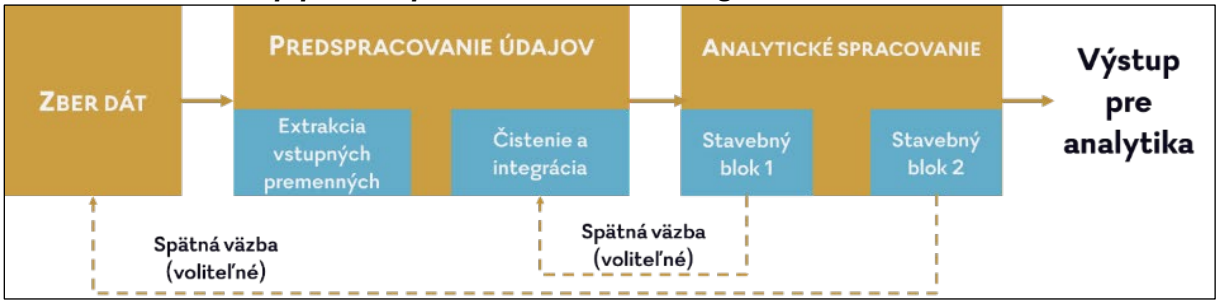
4. **Dátová analýza a prieskum** slúži na overenie predpokladov a odpovedí na otázky z kroku 1, na overenie kvality spracovania údajov z kroku 2 a na hlbšie porozumenie Big Data, aby bolo možné identifikovať vhodné algoritmy hĺbkovej analýzy Big Data na prípravu analytického modelu. V mnohých prípadoch 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