SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

3/2021
ročník/volume 31

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

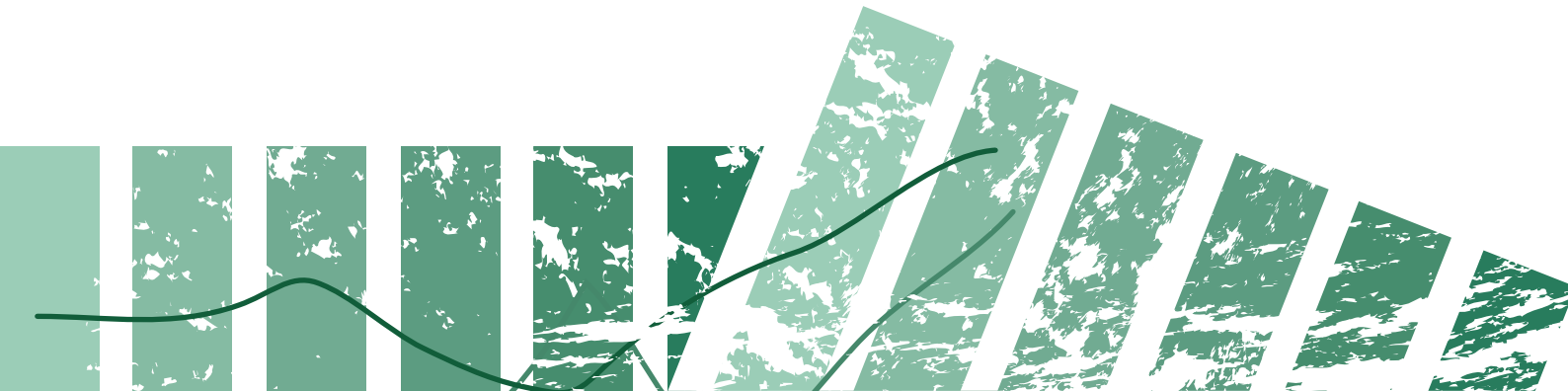
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 4

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 57 – 71

Dátum vydania/Publication date: 15. júl 2021/July 15, 2021



Ľudmila IVANČÍKOVÁ
Štatistický úrad Slovenskej republiky
Boris VAŇO
INFOSTAT – Výskumné demografické centrum

TEORETICKO-METODOLOGICKÉ ASPEKTY HODNOTENIA KVALITY ADMINISTRATÍVNYCH ZDROJOV ÚDAJOV VYUŽÍVANÝCH NA ŠTATISTICKÉ ÚČELY

THEORETICAL AND METHODOLOGICAL ASPECTS OF QUALITY ASSESSMENT OF ADMINISTRATIVE DATA SOURCES USED FOR STATISTICAL PURPOSES

ABSTRAKT
Článok sa zameriava na teoretické a metodologické aspekty hodnotenia administratívnych zdrojov údajov, ktoré sa využívajú na štatistické účely. Vytvorený teoreticko-metodologický rámec bude slúžiť ako základ na hodnotenie všetkých administratívnych údajov vstupujúcich do štatistického systému.

ABSTRACT
The article focuses on the theoretical and methodological aspects of the assessment of administrative data sources that are used for statistical purposes. The established theoretical and methodological framework will serve as a basis for the evaluation of all the existing administrative data entering the statistical system.

KĽÚČOVÉ SLOVÁ
administratívne zdroje údajov, kvalita, hodnotiaca schéma, indikátory kvality

KEY WORDS
administrative data sources, quality, evaluation framework, quality indicators

1. ÚVOD
Administratívne zdroje údajov (AZÚ) obsahujú informácie zbierané prioritne na administratívne účely. Čoraz častejšie sa však administratívne údaje využívajú aj na iné účely, než boli vytvorené a to predovšetkým na účely oficiálnej (alebo štátnej) štatistiky.

Hlavným benefitom ich využitia je redukcia nákladov, zníženie záťaže respondentov (fyzických aj právnických osôb), zvýšenie kvality údajov, zlepšenie včasnosti údajov, ako aj vyššia miera flexibility uspokojovania požiadaviek na podrobnejšie informácie (napr. z územného hľadiska).

Tvorba administratívnych zdrojov údajov je často mimo pôsobnosť štatistických úradov. To v prvom rade znamená, že štatistický úrad potrebuje informácie o zdroji a kvalite týchto údajov, ktoré využíva. Ani dobrá kvalita administratívnych zdrojov údajov však nemusí znamenať, že jeho údaje sú vhodné na štatistické využitie. Preto pod kvalitou administratívnych zdrojov údajov treba rozumieť predovšetkým vhodnosť ich použitia na príslušný štatistický účel.

Pri vyhodnotení kvality administratívnych údajov je nevyhnutná spolupráca zo strany vlastníka alebo správcu administratívneho zdroja dát. Pokiaľ takáto spolupráca nie je v dostatočnej miere zabezpečená, nie je možné naplniť niektoré aspekty hodnotenia kvality administratívnych zdrojov údajov.

Cieľom článku je zhrnúť všeobecne platné teoretické a metodologické aspekty hodnotenia administratívnych zdrojov údajov, ktoré sa využívajú na štatistické účely. Tie sa následne využili na hodnotenie kvality na národnej úrovni, konkrétne pri príprave Sčítania obyvateľov, domov a bytov 2021 (ďalej aj ako „SODB 2021“) ktoré je založené na kombinácii údajov z AZÚ a od obyvateľov.

Základom na sumarizáciu teoretických a metodologických aspektov hodnotenia kvality administratívnych údajov využívaných na štatistické účely sú odporúčania, smernice a štandardy Eurostatu, ako aj poznatky a skúsenosti štatistických úradov viacerých krajín (hlavne Holandska, Talianska a Švédska), ktoré majú s touto problematikou dlhoročné skúsenosti.

2. VÝCHODISKÁ POUŽITIA ADMINISTRATÍVNYCH ZDROJOV ÚDAJOV NA ŠTATISTICKÉ ÚČELY

Vo všeobecnosti sú identifikované dva základné spôsoby využitia administratívnych údajov – administratívne a externé [4]. Administratívne využitie je využitie administratívnych údajov organizáciou, ktorá ich vlastní, obvykle na účely, na ktoré ich vytvorila. Údaje sa môžu využívať napríklad na sledovanie aktivít organizácie alebo na získavanie informácií na rozhodovaciu činnosť manažmentu alebo na hodnotenie výkonnosti organizácie. Pri externom využití získavajú administratívne údaje rôzne organizácie alebo inštitúcie (štátne, verejné, neziskové, firmy), ktoré ich využívajú na iné účely, než na ktoré pôvodne vytvorené. Do tejto časti patrí aj využitie administratívnych zdrojov údajov na štatistický účel s cieľom vytvárať štatistiky.

Administratívne údaje sa dodávajú externým používateľom zvyčajne v agregovanej podobe. Mikroúdaje sa sprístupňujú len v špeciálnych prípadoch, ktoré sú obvykle špecifikované v zákone. Externé použitie administratívnych údajov musí spĺňať kritériá dôvernosti a ochrany súkromia, ktoré sú taktiež ustanovené zákonom. V prípade využitia administratívnych údajov na štatistické účely majú štatistické úrady zákonom umožnený prístup k príslušným údajom a to isté platí aj o využívaní týchto údajov[1].

Administratívne údaje určené na štatistické využitie musia spĺňať viaceré požiadavky, ktoré sú prepojené priamo s účelom použitia a ktoré vymedzujú indikátory na ich hodnotenie a na hodnotenie ich kvality. Ako hlavné požiadavky na použitie administratívnych zdrojov údajov a samotných údajov boli identifikované nasledujúce požiadavky [4]:

- Administratívne údaje musia byť vhodné na **účel**, na ktorý sa majú použiť. Napríklad nie je možné využiť administratívny register ako oporu výberu, pokiaľ neobsahuje informáciu, ktorá by umožňovala lokalizovať jednotky v základnom súbore.
- Administratívne údaje musia vyhovovať **koncepcii** zisťovania, na ktoré sa majú využívať. Napríklad výrazné rozdiely v definícii premenných môžu spôsobiť, že

[1] *Zákon č. 540/2001 Z. z. – zákon o štátnej štatistike.*

administratívne údaje nie sú použiteľné na príslušný štatistický účel. Rozdiely sa budú pravdepodobne zväčšovať s narastajúcim počtom administratívnych databáz využívaných na jeden účel, pretože definícia sa musí zhodovať pri všetkých použitých zdrojoch.

- Pre každú administratívnu databázu musia byť k dispozícii **metadáta** popisujúce obsah databázy. Metadáta musia popisovať okrem iného aj administratívne procedúry, ktorými boli údaje vytvorené, všetky dôležité administratívne udalosti, ktoré sa vzťahujú na údaje, a definície pojmov, premenných a súborov, ktorých sa týkajú. Vyhodnotenie kvality založenej na metaúdajoch sa realizuje pri nových, ale aj pri opakovane použitých zdrojoch a je často súčasťou hodnotenia kvality samotného zdroja.
- Dôležitou požiadavkou je, aby sa **referenčný dátum** administratívnych údajov zhodoval s referenčným dátumom zisťovania, na ktoré sa majú administratívne údaje využívať. Keďže táto požiadavka často nie je splnená, musí byť aspoň možné spraviť vhodný prepočet.
- Administratívne údaje musia poskytnúť **adekvátne pokrytie** zisťovaných jednotiek v základnom súbore a nemali by obsahovať duplicity a neúplné údaje. Nesplnenie tejto podmienky môže mať za následok problémy s reprezentatívnosťou.
- Administratívne údaje musia byť **presné**, aby neznižovali presnosť výsledného štatistického produktu. To nemusí nevyhnutne znamenať presnosť na mikroúrovni. Dôležitá je presnosť na tej úrovni, na ktorej sa administratívne údaje využívajú.
- Administratívne údaje z jedného zdroja musia byť **stabilné** v čase vo všetkých ohľadoch (štruktúra súboru, obsiahnuté premenné, definície, atď.). V ideálnom prípade sa v čase menia len hodnoty premenných.
- Musí byť možné získať administratívne údaje vo vhodnej **forme**, ktorá zodpovedá účelu ich použitia. Problémy obvykle nastávajú aj vtedy, keď správca administratívneho zdroja údajov a štatistický úrad používajú vysoko nekompatibilné databázové systémy, ktoré komplikujú výmenu údajov.
- **Štruktúra** administratívnych databáz musí v prípade potreby umožňovať efektívne spájanie s údajmi zo zisťovania alebo z iných databáz. To znamená, že každá databáza musí obsahovať premennú alebo kombináciu premenných, ktoré jednoznačne identifikujú každú jednotku v základnom súbore. Najlepšou zárukou úspešného spájania údajov je rovnaký jednoznačný identifikátor vo všetkých databázach.

Ako sme spomínali, pri hodnotení kvality administratívneho zdroja je potrebné zohľadniť aj účel použitia v rámci štatistickej produkcie. Môže ísť napr. o návrh a plánovanie zisťovania, zber údajov, rozširovanie reprezentatívnosti, verifikáciu údajov, editovanie a imputovanie údajov, tvorbu pomocných premenných a tvorbu štatistických registrov.

Pri návrhu zisťovania sa administratívne zdroje údajov môžu použiť ako opora výberu. Administratívne údaje poskytujú tiež užitočné informácie využiteľné pri stratifikácii. Pomocou analýzy uskutočnenej na administratívnych údajoch je možné získať informácie o vzťahoch medzi premennými. Pri príprave zisťovaní napríklad pomáhajú pri špecifikácii požadovaného počtu respondentov a ich rozmiestnenia.

V určitých prípadoch administratívne údaje nahrádzajú štatistické zisťovania a umožňujú tak štatistickým úradom šetriť finančné prostriedky a znižovať zaťaženie respondentov. Niekedy sa údaje potrebné pre štatistický produkt čerpajú výlučne

z administratívnych zdrojov, v niektorých prípadoch ide o kombináciu administratívnych údajov a štatistického zisťovania (časť otázok sa nahrádza údajmi z administratívnych registrov alebo niektoré časti zisťovania sa získavajú z administratívnych zdrojov, alebo v prípade neodpovedí sa môžu využívať údaje z administratívnych zdrojov). Príkladom je systém výberových zisťovaní v severských krajinách alebo realizácia kombinovaného či plne registrovaného sčítania obyvateľov, domov a bytov.

Neaktuálna alebo nepresná opora výberu môže spôsobiť mnohé problémy pri zisťovaní. Administratívne údaje môžu predovšetkým poskytnúť správnu lokáciu jednotiek v základnom súbore. To znamená, že prispievajú k znižovaniu počtu nezastihnutých osôb a zároveň môžu odhaliť existenciu takých údajov, ktoré neboli zaradené do opory výberu.

Údaje zo štatistického zisťovania je možno verifikovať a podrobiť krížovej kontrole, pokiaľ rovnaké premenné týkajúce sa rovnakých jednotiek v základnom súbore existujú aj v databázach vytvorených na administratívne účely.

Administratívne dáta môžu poskytnúť potrebné údaje o pomocných premenných, ktoré sa využili na výpočet váh a v štatistických metódach ako poststratifikácia, regresia alebo kalibrácia.

Pri editovaní sa preveruje, či údaje získané od jednotlivých respondentov vyhovujú určitým podmienkam. Niektoré z týchto podmienok zahŕňajú vzťahy medzi premennými, ktoré musia byť dodržané pri každom respondentovi. Tieto vzťahy sa niekedy odhadujú na základe administratívnych údajov. Administratívne údaje sa často využívajú aj pri imputácii, konkrétne na odhad modelovaných hodnôt, ktorými sa nahrádzajú chybné alebo chýbajúce hodnoty z pôvodného zisťovania.

Administratívne údaje sa môžu využiť aj na tvorbu štatistických registrov. Ide o registre, v ktorých sú uložené údaje za všetky sledované jednotky. Administratívne údaje predstavujú hlavný vstup do týchto registrov, zvyšok údajov pochádza zo štatistických zisťovaní. Registre sa kontrolujú a dopĺňajú na pravidelnej alebo nepravidelnej báze. Štatistické registre sa dajú využívať namiesto pôvodných administratívnych údajov. Ich výhodou je, že zodpovedajú štatistickým požiadavkám, čo nie je vždy prípad administratívnych údajov. Ďalšou výhodou štatistických registrov je, že ich spravujú štatistické úrady, ktoré tak majú väčší dosah na ich kvalitu.

Práca s administratívnymi zdrojmi údajov sa môže plne realizovať v súlade so štatistickými procesmi popísanými v generickom modeli štatistického biznis procesu (GSBPM) [7].

Údaje, ktoré potrebuje štatistický úrad na konkrétny účel, sa môžu nachádzať vo viacerých administratívnych databázach. Sú prípady, keď niektorý zdroj nie je vhodné použiť na príslušný účel alebo jeden zdroj nedokáže pokryť všetky požiadavky a je potrebné kombinovať údaje z viacerých administratívnych zdrojov. Štatistický úrad musí preto identifikovať všetky potenciálne zdroje údajov a preskúmať ich vhodnosť, aby bolo možné rozhodnúť, ktorý zdroj sa použije.

Štatistický úrad často potrebuje administratívne údaje o jednotkách v základnom súbore, za ktoré už má určité dáta. Alebo potrebuje získavať informácie o jednotkách v základnom súbore z viacerých administratívnych zdrojov. V takýchto prípadoch je potrebné identifikovať vhodné vety v administratívnych databázach a spojiť ich navzájom alebo so správnou jednotkou v základnom súbore. Základným predpokladom pre úspešného spájania viet je už spomínaný jednoznačný identifikátor, ktorý je rovnako konštruovaný vo všetkých databázach – štatistických aj administratívnych.

3. KVALITA ADMINISTRATÍVNYCH ZDROJOV ÚDAJOV

Hodnotenie kvality štatistických údajov je viacrozmerný proces, ktorým sa zisťuje, ako dobre zodpovedá príslušný štatistický produkt svojmu účelu. V európskom štatistickom systéme (EŠS) je kvalita manažovaná prostredníctvom Kódexu postupov pre európsku štatistiku (kódex postupov), ktorý stanovuje štandardy pre vývoj, tvorbu a distribúciu európskych štatistík [5], [6]. Hodnotenie kvality zahŕňa viaceré aspekty týkajúce sa všetkých typov štatistických procesov[2]. My sa ďalej budeme zaoberať len kvalitou administratívnych zdrojov údajov.

V súlade s kódexom postupov EŠS sa kvalita administratívnych zdrojov údajov hodnotí na základe kritérií, ktoré sa môžu týkať vstupov, procesov, výstupov a inštitucionálneho prostredia. Vzhľadom na potrebu hodnotenia možného použitia administratívnych zdrojov údajov na sčítanie obyvateľov, domov a bytov 2021 sme sa v príspevku zamerali najmä na hodnotenie kvality na vstupe.

Kritériami hodnotiacimi **kvalitu výstupov** sú relevantnosť (ako výstupy spĺňajú požiadavky používateľa), presnosť a spoľahlivosť (ako výstupy zodpovedajú realite), včasnosť (či sú výstupy dodávané v dohodnutom čase), koherentnosť a porovnateľnosť (toto kritérium zahŕňa vnútornú súdržnosť výstupov a ich porovnateľnosť v čase a z regionálneho pohľadu), dostupnosť a jasnosť (či sú výstupy prezentované jasne a zrozumiteľne, dodávané vo vhodnom formáte, dostupné na nestrannej báze vrátane metadát a komentárov).

Kvalita výstupov sa vždy dosahuje prostredníctvom **kvality procesov**. Vo všeobecnosti kvalita procesov má dva široké aspekty. Je to účinnosť, ktorá zabezpečuje dobrú kvalitu výstupov a efektívnosť, ktorá zabezpečuje dosahovanie dobrej kvality výstupov pri čo najmenších nákladoch. V koncepcii EŠS a v súlade s princípmi kódexu postupov existujú štyri kritériá kvality štatistických procesov. Niektoré z týchto kritérií sa týkajú aj inštitucionálneho prostredia, majú teda dvojité využitie. Prvým kritériom je používaná metodológia (vrátane adekvátnych nástrojov, procedúr a expertíz, ktoré podporujú kvalitu štatistických produktov). Druhým kritériom sú vhodné štatistické procedúry, implementované počnúc zberom údajov a končiac ich validáciou. Tretím kritériom hodnotenia kvality štatistických procesov je zaťaženie respondentov, ktoré nemôže byť neúmerné a štatistický úrad ho musí priebežne kontrolovať a prehodnocovať. Posledným kritériom je nákladová efektívnosť, ktorá hodnotí efektívnosť využívania zdrojov.

[2] V rámci Európskeho štatistického systému existuje šesť základných typov štatistických procesov – výberové zisťovania, cenzy, administratívne zdroje údajov, viaczdrojové štatistiky, cenové a ekonomické indexy, štatistické kompilácie.

Inštitucionálne prostredie predstavuje celý kontext, v ktorom štatistický úrad pôsobí a v rámci ktorého prebiehajú štatistické procesy. Niektoré kritériá hodnotiace kvalitu inštitucionálneho prostredia sa týkajú aj štatistických procesov. Prvým kritériom je profesionálna nezávislosť. Ide o nezávislosť štatistického úradu od politických, regulačných a administratívnych subjektov, ako aj od súkromných firiem. Druhým kritériom je mandát na zber údajov. Štatistický úrad musí mať jasný, nespochybniteľný a právne podložený mandát na zbieranie údajov pre potreby európskej štatistiky. Štátna a verejná správa, podniky, domácnosti a široká verejnosť by mali byť právne zaviazané v prípade potreby poskytnúť údaje pre potreby európskej štatistiky. Tretím kritériom je adekvátnosť zdrojov, ktoré musia byť postačujúce na plnenie požiadaviek európskej štatistiky. Štvrtým kritériom je dôraz na kvalitu. Štatistické úrady musia pravidelne a systematicky identifikovať problémy a zlepšovať kvalitu štatistických procesov a produktov. Posledným kritériom kvality, ktoré sa týka inštitucionálneho prostredia, je štatistická dôvernosť. Štatistický úrad musí garantovať súkromie poskytovateľov údajov, dôvernosť poskytovaných informácií a ich využitie výlučne na štatistické účely.

3.1 FAKTORY OVPLYVŇUJÚCE KVALITU ADMINISTRATÍVNYCH ÚDAJOV

Pokiaľ chceme zlepšovať kvalitu administratívnych údajov, je dôležité poznať faktory, ktoré ju ovplyvňujú. Do procesu tvorby administratívnych databáz a ich využitia je zapojených viacero subjektov – respondenti, dodávatelia administratívnych zdrojov údajov, štatistické úrady ako špecifickí používatelia administratívnych údajov a koncoví používatelia štatistických údajov. A práve s týmito subjektami sa spájajú aj jednotlivé faktory ovplyvňujúce kvalitu administratívnych databáz. Znalosť faktorov umožňuje viesť diskusiu medzi dodávateľom administratívnych údajov a štatistickým úradom, čím sa vytvárajú predpoklady na zvýšenie kvality administratívnych údajov. Nasledujúce faktory sú vymedzené na základe prác Daasa a Van Nederpelta [3].

Na začiatku procesu tvorby údajov aj v prípade administratívnych zdrojov stoja respondenti. Sú to osoby (skupiny osôb) alebo organizácie, ktoré dodávajú údaje vlastníkom alebo správcom databáz. S respondentmi sa spájajú predovšetkým faktory ako motivácia a schopnosť poskytnúť korektné a reálne údaje a motivácia a schopnosť poskytnúť údaje včas. Kvalitu, včasnosť a kompletnosť poskytnutých údajov ovplyvňuje aj stratégia a koncepcia prístupu k respondentovi zo strany zadávateľa (vlastníka alebo správcu administratívneho zdroja údajov). Ide o zrozumiteľnosť otázok, ich prehľadné usporiadanie, rozumný rozsah námahy a úsilia, ktoré treba vynaložiť pri poskytovaní údajov.

Pod systémom rozumieme nástroj na automatický zber údajov od respondenta a ich dodanie vlastníkovi alebo správcovi administratívneho zdroja údajov. Do tejto skupiny faktorov patrí dostupnosť systému (napr. vzhľadom na poruchy alebo údržbu) a jeho správne fungovanie (korektný zber údajov a ich dodanie).

Najviac faktorov, ktoré ovplyvňujú kvalitu administratívnych údajov, sa týka ich dodávateľov. Pod dodávateľom administratívnych údajov rozumieme subjekt, ktorý zbiera údaje od respondentov s využitím systému a spracované údaje dodáva štatistickému úradu. Faktory, ktoré svedčia o kvalite dodávateľa administratívnych údajov sú napr. kontinuita (vykonávanie činnosti počas dlhšieho obdobia), reputácia (na základe referencií), motivácia a schopnosť dodávať korektné údaje v stanovenom čase, ochota spolupracovať, motivácia a schopnosť monitorovať okolie, motivácia

a schopnosť meniť procesy (ak je to potrebné). Výhodou je, keď dodávateľ dokáže vnímať a prípadne aj zohľadňovať záujmy a potreby štatistického úradu. Ďalšími faktormi, ktoré ovplyvňujú kvalitu administratívnych údajov a ktoré môže ovplyvniť ich dodávateľ, sú spôsob registrácie údajov, úplnosť údajov, stabilita (zmeny v registri), spôsob dodania údajov štatistickému úradu, efektívnosť, dodanie údajov v stanovenom termíne. Všetky informácie týkajúce sa administratívnych údajov, ktoré dodávateľ poskytuje štatistickému úradu, sú veľmi dôležitými faktormi. Ide o opis produkčných procesov, informácie o údajoch, informácie o zmenách. Úlohu zohráva aj kvalita personálu, ktorý má dodávateľ k dispozícii a ktorý spolupracuje so štatistickým úradom.

Využiteľnosť administratívnych údajov na štatistické účely ovplyvňuje spolupráca dodávateľa administratívnych údajov a štatistického úradu. To znamená, do akej miery je štatistický úrad zapojený do procesu tvorby administratívnych údajov a do akej miery môže a dokáže tento proces ovplyvňovať. Dôležitý je aj spôsob komunikácie zo strany štatistického úradu smerom so správcom administratívneho zdroja údajov, ktorý je ovplyvnený zaužívanými postupmi, ako aj personálnymi kapacitami a ich kvalitou.

Poslednú skupinu tvoria faktory týkajúce sa dohôd a právnych úprav. Právne normy riadia, resp. ovplyvňujú proces tvorby administratívnych údajov a ich dodania štatistickému úradu. Právne normy by mali zabezpečovať potreby štatistického úradu, mali by byť aktuálne (aktualizované), zrozumiteľné a nespochybniteľné. V nadväznosti na existujúcu legislatívu by mala existovať dohoda medzi dodávateľom administratívnych údajov a štatistickým úradom, ktorá by tiež mala byť korektná, zrozumiteľná, jednoznačná a aktuálna.

3.2 METÓDY A NÁSTROJE NA HODNOTENIE KVALITY ADMINISTRATÍVNYCH ÚDAJOV

Ak chceme posudzovať kvalitu administratívnych údajov, je potrebné mať okrem presných a podrobných informácií o týchto údajoch aj vhodné nástroje na meranie a hodnotenie. Nástrojom sú indikátory, ktoré v prípade administratívnych zdrojov údajov rozdeľujeme na:

- vstupné indikátory kvality, ktoré hodnotia samotný administratívny zdroj údajov a definujú kvalitu administratívnych údajov využívaných na štatistické účely na úrovni vstupných premenných,
- procesné indikátory kvality, ktoré merajú kvalitu súvisiacu s produkčnými procesmi, v ktorých sa využívajú administratívne údaje,
- výstupné indikátory kvality, ktoré merajú kvalitu štatistických výstupov, ktoré vznikli s využitím administratívnych údajov, pričom zohľadňujú aj kvalitu vstupov a procesov).

Ďalej sa budeme zaoberať len vstupnými indikátormi kvality, ktoré umožňujú posúdiť kvalitu administratívnych databáz, vstupujúcich do procesu tvorby štatistických údajov na úrovni samotného zdroja a na úrovni údajov.

Meranie kvality administratívnych údajov používaných na štatistické účely nie je rovnaké ako meranie kvality štatistických zisťovaní[3]. Aby bolo možné korektne a vyčerpávajúco vyhodnotiť všetky aspekty kvality administratívnych údajov,

[3] Meranie kvality štatistických produktov je zadefinované Eurostatom a má šesť dimenzií – relevantnosť, presnosť, včasnosť, dostupnosť a jasnosť, porovnateľnosť, koherentnosť.

vychádzali sme pre účely hodnotenia a kvality AZÚ v podmienkach Slovenska (s osobitným dôrazom na účel využitia AZÚ pri SODB 2021) z prehodnotenia existujúceho rámca hodnotenia kvality štatistických zisťovaní a opierali sme sa o hierarchický a multidimenzionálny prístup holandských štatistikov [2].

Z hierarchického hľadiska ide o štyri úrovne. Najvyššou úrovňou sú hyperdimenzie (zvyčajne sa označujú aj ako úrovne alebo kategórie). Každá hyperdimenzia sa skladá z viacerých dimenzií a každá dimenzia obsahuje niekoľko indikátorov kvality. Poslednou úrovňou sú metódy merania. Na výpočet každého indikátora kvality existuje jedna alebo viac kvalitatívnych alebo kvantitatívnych metód merania.

Multidimenzionalita prístupu vyplýva zo skutočnosti, že ku kvalite administratívneho zdroja údajov sa pristupuje ako k celku a nie len vzhľadom na údaje. Zatiaľ čo hyperdimenzie, dimenzie aj indikátory kvality sú stabilné, metódy merania sú flexibilné. Tento prístup umožňuje vybrať pre každý indikátor kvality najvhodnejšie metódy merania, čo na druhej strane umožňuje flexibilné hodnotenie administratívnych databáz bez ohľadu na ich typ, na oblasť štatistiky, v ktorej sa využívajú, a na spôsob, akým sa využívajú.

Vzhľadom na hodnotenie vstupnej kvality administratívnych údajov je možné vymedziť tri relevantné hyperdimenzie. Ide o **zdroj údajov**, **metaúdaje** a **údaje.** Každá hyperdimenzia umožňuje špecifické hodnotenie použiteľnosti administratívnych údajov na štatistické účely.

V hyperdimenzii *zdroj údajov* sa skúma zdroj ako celok, správca administratívneho zdroja, ako aj dodanie údajov štatistickému úradu. Skladá sa z piatich dimenzií – dodávateľ, relevantnosť, súkromie a bezpečnosť, dodanie a procedúry. Indikátory kvality (vyznačené kurzívou) pre všetky dimenzie spolu s metódami na ich meranie (v zátvorke) sú nasledujúce:

Pri dodávateľovi je potrebné sledovať *kontakt* (názov AZÚ, kontaktné informácie na správcu AZÚ, kontaktná osoba pre štatistický úrad) a *dôvod* (dôvod využívania AZÚ na štatistické účely).

Hlavnými indikátormi relevantnosti sú *užitočnosť* (dôležitosť AZÚ pre štatistický úrad), *možné využitie* (potenciálne štatistické využitie AZÚ), *dopyt* (uspokojuje AZÚ dopyt po informáciách?) a *zaťaženie respondentov* (vplyv využitia AZÚ na zaťaženie respondentov).

Pri dimenzii súkromie a bezpečnosť sú identifikované indikátory *právny predpis* (právny základ pre existenciu AZÚ), *dôvernosť* (použitie zákona o ochrane osobných údajov) a *bezpečnosť* (spôsob doručenia AZÚ na štatistickému úradu a bezpečnostné opatrenia týkajúce sa hardvéru a softvéru).

Dimenzia dodanie sa hodnotí prostredníctvom indikátorov popisujúcich *náklady* (náklady spojené s využívaním AZÚ), *dohody* (termíny dodania AZÚ a frekvencia dodávok AZÚ), *včasnosť* (včasnosť dodania AZÚ, rýchlosť informovania o výnimkách a rýchlosť, s akou sú údaje ukladá správca databázy), *formát* (formát, v ktorom sa údaje dodávajú) a *výber* (Aké údaje boli dodané? Vyhovujú dodané údaje požiadavkám štatistického úradu?).

Poslednú dimenziu môžeme definovať prostredníctvom *zberu údajov* (informácie o spôsobe zberu údajov), *plánovaných zmien* (informácie o plánovaných zmenách v zdroji údajov, komunikácia o zmenách medzi správcom a štatistickým úradom), *spätnej väzby* (kontaktné údaje správcu zdroja údajov pre prípad problémov, určenie prípadov na spätnú väzbu) a *núdzového scenára* (riziko závislosti štatistického úradu a bezpečnostné opatrenia, ak zdroj údajov nie je dodaný podľa dohody).

Druhá hyperdimenzia sa týka *metaúdajov*. Jasnosť definícií a kompletnosť metainformácií patria medzi jej hlavné kvalitatívne aspekty. Hyperdimenzia metaúdaje sa skladá zo štyroch dimenzií – jasnosť, porovnateľnosť, jednoznačný identifikátor a práca s údajmi (zo strany správcu AZÚ). Indikátory kvality (vyznačené kurzívou) pre všetky dimenzie spolu s metódami merania (v zátvorke) sú uvedené v nasledujúcom texte:

Pre dimenziu jasnosť sa identifikovali indikátory, ako *definícia zisťovanej jednotky* (stupnica podľa definície), *definícia klasifikačnej premennej* (stupnica podľa definície), *definícia početnosti* (stupnica podľa definície), *časová dimenzia* (stupnica podľa definície) a *definícia zmien* (oboznámenosť s uskutočnenými zmenami).

Pri dimenzii porovnateľnosť ide o *porovnanie definície zisťovanej jednotky* (porovnanie s definíciou štatistického úradu, *porovnanie definície klasifikačnej premennej* s definíciou štatistického úradu, *porovnanie definície početnosti* s definíciou štatistického úradu a porovnanie *časových rozdielov* s obdobím vykazovania v štatistickom úrade.

Pri jednoznačnom identifikátori je potrebné sledovať samotný *identifikátor* (existencia jednoznačného identifikátora a porovnanie s jednoznačnými identifikátormi používanými v štatistickom úrade), ako aj *jednoznačnú kombináciu premenných (*existencia použiteľnej kombinácie premenných).

Pri práci s údajmi sú dôležité *kontroly* (kontrola zisťovaných jednotiek, kontrola premenných, kontrola kombinácie premenných, kontrola extrémnych hodnôt) a *modifikácie* (oboznámenosť s modifikáciou údajov, označenie modifikovaných údajov, oboznámenosť s použitím difoltných hodnôt).

Hyperdimenzia *údaje* je zameraná na hodnotenie kvality údajov (budúcich premenných), ktoré sa nachádzajú v databáze. Kvalitatívne aspekty tejto hyperdimenzie sú prevažne spojené s presnosťou. Hyperdimenzia údaje sa skladá z piatich dimenzií – technické kontroly, integrovateľnosť, presnosť, úplnosť, časové hľadisko. Indikátory kvality pre všetky dimenzie spolu s popisom sú uvedené v tabuľke č. 1.

Na hodnotenie kvality administratívnych údajov sa odporúča využívať rovnocenne všetky tri hyperdimenzie. Kvalitatívne aspekty hyperdimenzie *zdroj údajov* sú všeobecnejšie, v ostatných dvoch hyperdimenziách sú menej všeobecné, pričom najkonkrétnejšie sú v hyperdimenzii údaje. V prvých dvoch hyperdimenziách sa kvalita hodnotí skoro výlučne pomocou kvalitatívnych indikátorov, v hyperdimenzii údaje výrazne prevláda hodnotenie kvality prostredníctvom kvantitatívnych údajov [1].

Z uvedeného vyplýva, že formálna stránka hodnotenia nemôže byť vo všetkých hyperdimenziách rovnaká. Na vyhodnotenie hyperdimenzií zdroj údajov a metaúdaje je vhodné použiť kontrolný dotazník, vypracovaný špeciálne na tento účel. Pre hyperdimenziu údaje nie je možné použiť dotazník, a to kvôli veľkému množstvu výpočtov, ktoré treba spraviť pri kvantifikácii indikátorov kvality v tejto hyperdimenzii [1].

Okrem štandardných kontrolných procedúr obsiahnutých v troch hyperdimenziách, môže používateľ v prípade záujmu spraviť špeciálne kontroly kvality. Najčastejšie ide o porovnanie údajov získaných z administratívneho zdroja s rovnakými údajmi získanými zo štatistických zisťovaní.

4. ZÁVER

Cieľom článku bolo spracovať teoretické podklady k problematike hodnotenia administratívnych zdrojov údajov používaných na štatistické účely s osobitným dôrazom na hodnotenie kvality zdroja na vstupe, aby sa tak potvrdila relevantnosť využitia konkrétnych administratívnych zdrojov a registrov na štatistické účely, v praxi na účely sčítania obyvateľov, domov a bytov 2021.

Výsledná hodnotiaca schéma bola automatizovaná a použitá pri hodnotení AZÚ a ich kvality s dosahom na využitie konkrétnych zdrojov údajov pri sčítaní domov a bytov. Hodnotiaca schéma bola otestovaná aj pri hodnotení kvality vybraných administratívnych databáz použitých pri sčítaní obyvateľov v rokoch 2018 – 2020.

Konkrétne použitie bude popísané v osobitnom príspevku k hodnoteniu kvality administratívnych zdrojov údajov použitých na integrované sčítanie obyvateľov, domov a bytov.

Tabuľka č. 1: Hyperdimenzia dáta – dimenzie, indikátory kvality, metódy merania

Dimenzia	Indikátor kvality	Metódy merania
Technické kontroly	Čitateľnosť	Podiel chybných alebo neznámych súborov
		Podiel nečitateľných súborov
	Konvertibilita	Podiel objektov s chybou v dekódovaní alebo poškodené údaje
	Dodržiavanie deklarácie súboru	Podiel premenných, ktoré sa líšia od dohodnutej špecifikácie
Integrovateľnosť	Porovnateľnosť objektov	Podiel identických objektov = počet objektov s presne rovnakou analyzovanou jednotkou a rovnakou definíciou akú má ŠÚ / počet všetkých relevantných objektov v AZÚ
		Podiel zodpovedajúcich objektov = počet objektov, ktoré po harmonizácii budú zodpovedať požiadavkám ŠÚ / počet všetkých relevantných objektov v AZÚ
		Podiel neporovnateľných objektov = počet objektov, ktoré ani po harmonizácii nebudú zodpovedať požiadavkám ŠÚ / počet všetkých relevantných objektov v AZÚ
	Prepojenie objektov	Podiel identických priradených objektov = počet objektov v referenčnom štatistickom súbore s presne rovnakou analyzovanou jednotkou a rovnakou definíciou ako je v AZÚ / celkový počet relevantných objektov v referenčnom štatistickom súbore
		Podiel zodpovedajúcich priradených objektov = počet objektov v referenčnom štatistickom súbore, ktoré po harmonizácii zodpovedajú jednotkám alebo častiam jednotiek v AZÚ / celkový počet relevantných objektov v referenčnom štatistickom súbore
		Podiel nepriradených objektov = počet objektov v referenčnom štatistickom súbore, ktoré ani po harmonizácii nemôžu byť priradené k niektorej jednotke v zdroji dát / celkový počet relevantných objektov v referenčnom štatistickom súbore
		Podiel nepriradených agregovaných objektov = časť objektov na agregovanej úrovni v zdroji 1, ktoré nemožno priradiť + časť objektov na rovnakej agregovanej úrovni v zdroji 2, ktoré nemožno priradiť
	Prepojenie premenných	Podiel objektov s neprepojenými premennými = počet objektov v zdroji dát bez prepojenej premennej / celkový počet objektov v AZÚ
		Podiel objektov s prepojenými premennými, ktoré sú rozdielne od premenných používaných ŠÚ = počet objektov v AZÚ s prepojenými premennými rozdielnych od premenných používaných ŠÚ / celkový počet objektov s prepojenými premennými v AZÚ
		Podiel objektov s korektne konvertovateľnými prepojenými premennými = počet objektov v AZÚ pre ktoré pôvodne prepojené premenné môžu byť prekonvertované na premennú používanú ŠÚ / celkový počet objektov s prepojenou premennou v AZÚ

Integrovateľnosť	Porovnateľnosť premenných	Použitie inšpekčných metód pre štatistické údaje na porovnanie súčtov zoskupenia špecifických objektov pre premenné v obidvoch zdrojoch. Grafické metódy, ktoré je možné použiť, sú stĺpcový a bodový diagram. Môže sa porovnávať aj rozdelenie hodnôt.
		Stredná absolútna percentuálna chyba, ktorá meria priemer absolútnych percentuálnych chýb
		Metóda odvodená od chí-kvadrát testu, ktorá vyhodnocuje rozdelenie numerických hodnôt v obidvoch databázach. Pre kvalitatívne údaje možno použiť koeficient Cramerovo V
		Podiel objektov s identickými hodnotami premenných = počet objektov v zdroji 1 a 2 s presne rovnakými hodnotami sledovaných premenných / celkový počet relevantných objektov v obidvoch zdrojoch
Presnosť	Autenticita	Podiel objektov so syntakticky nesprávnym identifikátorom
		Podiel objektov pre ktoré AZÚ údajov obsahuje kontradiktívnu informáciu k informácii uvedenej v referenčnom zozname
		Kontaktovať vlastníka AZÚ v súvise s podielom neautentických objektov v AZÚ
	Nekonzistentné objekty	Podiel objektov zapojených do nelogických vzťahov s inými objektami
	Pochybné objekty	Podiel objektov zapojených do nepravdepodobných avšak nie nevyhnutne nekorektných vzťahov s inými objektami
	Chyby merania	Podiel neoznačených hodnôt v AZÚ pre každú premennú (označené hodnoty neobsahujú chybu merania – označuje správca AZÚ)
		Kontaktovať vlastníka AZÚ a položiť nasledujúce otázky týkajúce sa kvality údajov
		Používa sa dizajn v procese zberu údajov?
		Kontrolujú sa údaje počas fázy reportingu?
		Používajú sa štandardy pre niektoré premenné?
		Používajú sa kontroly vstupu údajov?
		Používajú sa kontroly údajov počas spracovania alebo uchovávania údajov?
	Nekonzistentné premenné	Podiel objektov s nekonzistentnými hodnotami premenných (mimo intervalu) alebo objektov, ktorých kombinácie hodnôt premenných nie sú logické
	Pochybné premenné	Podiel objektov s pochybnými hodnotami premenných alebo objektov(kombinácie sú nepravdepodobné avšak nie nevyhnutne nesprávne)
Úplnosť	Podhodnotenie	Podiel objektov z referenčného zoznamu chýbajúcich v AZÚ
	Nadhodnotenie	Podiel objektov v AZÚ nezahrnutých do referenčného súboru
	Selektívnosť	Použitie inšpekčných metód pre štatistické údaje (napr. histogramy) na porovnanie premenných pre objekty v AZÚ a v referenčnom súbore Použitie zložitejších grafických metód ako napr. tabuľkový diagram

Úplnosť	Selektívnosť	Vypočítať ukazovateľ reprezentatívnosti pre objekty v AZÚ (napr. R-indikátor)
	Nadbytočnosť	Podiel duplicitných objektov v AZÚ (s rovnakým identifikátorom)
		Podiel duplicitných objektov v AZÚ (s rovnakou hodnotou pre vybrané údaje)
		Podiel duplicitných objektov v AZÚ (s rovnakou hodnotou pre všetky premenné)
	Chýbajúce hodnoty	Podiel objektov s chýbajúcou hodnotou pre príslušnú premennú
		Podiel objektov so všetkými chýbajúcimi hodnotami pre vybrané premenné
	Imputované hodnoty	Podiel imputovaných hodnôt na premennú v AZÚ
		Kontaktovať vlastníka AZÚ a spýtať sa na podiel imputovaných hodnôt na premennú
Časové hľadisko	Včasnosť	Časový rozdiel (dni) = dátum doručenia do ŠÚ - dátum konca referenčného obdobia
		Časový rozdiel (dni) = dátum doručenia používateľovi – dátum konca referenčného obdobia
	Dochvíľnosť	Časový rozdiel (dni) = dátum doručenia do ŠÚ - dohodnutý dátum, ktorý je stanovený v zmluve
	Celkový čas oneskorenia	Celkový časový rozdiel (dni) = predpokladaný dátum, keď bude môcť ŠÚ využívať údaje z AZÚ – dátum konca referenčného obdobia
	Oneskorenie	Kontaktovať vlastníka AZÚ, aby poskytol informáciu o oneskorení registrácie
		Časový rozdiel (dni) = dátum zachytenia zmeny v AZÚ zo strany správcu – dátum vzniku zmeny
	Dynamika objektov	Podiel vzniknutých objektov v čase t = vzniknuté objekty v čase t / všetky objekty v čase t = vzniknuté objekty v čase t / (vzniknuté objekty v čase t + existujúce objekty v čase t)
		Podiel zaniknutých objektov v čase t = zaniknuté objekty v čase t / všetky objekty v čase t = zaniknuté objekty v čase t / (vzniknuté objekty v čase t + zaniknuté objekty v čase t)
		Podiel zaniknutých objektov v čase t-1 = zaniknuté objekty v čase t / všetky objekty v čase t-1 = zaniknuté objekty v čase t /(vzniknuté objekty v čase t + existujúce objekty v čase t)
	Stabilita premenných	Použitie inšpekčných metód na porovnanie hodnôt špecifických premenných pre pretrvávajúce objekty v rôznych dodaniach súboru.
		Podiel zmien = počet objektov so zmenenými hodnotami / celkový počet pretrvávajúcich objektov s vyplnenou hodnotou pre sledovanú premennú
		Môžu sa použiť korelačné štatistické metódy a to na určenie v akom rozsahu sa hodnoty zmenili v rovnakom smere pre rôzne objekty. Pre kvalitatívne údaje je možné použiť metódy ako Cramerovo V.

Zdroj: [1], vlastné spracovanie

LITERATÚRA

[1] CERRONI, F. – DI BELLA, G. – GALIÉ, L.: Evaluating administrative data quality as input of the statistical production process. Rivista di Statistica Ufficiale, 2014, č. 1 – 2, s. 117 – 146.
[2] DAAS, P. – OSSEN, S.: Metadata quality evaluation of secondary data sources. International Journal for Qualiry research, 2011, Vol. 5, No. 2, p. 57 – 66.
[3] DAAS, P. – VAN NEDERPELT, P.: 49 factors that influence the quality of secondary data sources.2012. The Hague/Heerlen, Statistics Netherlands, 21 s.
[4] Eurostat.: Quality assessment of administrative data for statistical purposes. Luxembourg, 2003. 22 s.
[5] Eurostat.: Kódex postupov pre európsku štatistiku. Luxemburg, 2011. 8 s.
[6] Eurostat.: European Statistics Code of Practice, revised edition 2017. [online]. [cit. 11-03-2021]. Dostupné na: https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice
[7] GSBPM: Generic Statistical Business Process Model [online]. [cit. 22-02-2021]. Dostupné na: https://ec.europa.eu/eurostat/cros/content/gsbpm-generic-statistical-business-process-model-theme_en

RESUMÉ

Administratívne zdroje údajov obsahujú informácie zbierané prioritne na administratívne účely. Čoraz častejšie sa však administratívne údaje využívajú aj na iné účely, než boli vytvorené a to predovšetkým účely oficiálnej (alebo štátnej) štatistiky. Hlavným prínosom ich využitia je redukcia nákladov, zníženie záťaže respondentov, zvýšenie kvality údajov, zlepšenie včasnosti údajov, ako aj vyššia miera flexibility uspokojovania požiadaviek na podrobnejšie informácie (napr. z územného hľadiska).
Tvorba administratívnych zdrojov údajov je často mimo pôsobnosti štatistických úradov. To v prvom rade znamená, že štatistický úrad potrebuje informácie o zdroji a kvalite týchto údajov.
Cieľom článku je zhrnúť všeobecne platné teoretické a metodologické aspekty hodnotenia administratívnych zdrojov údajov, ktoré sa využívajú na štatistické účely. Tie sa následne využili na tvorbu teoreticko-metodologického rámca na hodnotenie administratívnych údajov, ktoré vstupujú do štatistického systému.
Meranie kvality administratívnych údajov používaných na štatistické účely sa líši od merania kvality štatistických zisťovaní. Existujúci rámec hodnotenia kvality štatistických zisťovaní bolo preto potrebné prehodnotiť. Využil sa pritom hierarchický a multidimenzionálny prístup.
Z hierarchického hľadiska ide o štyri úrovne. Najvyššou úrovňou sú hyperdimenzie Každá hyperdimenzia sa skladá z viacerých dimenzií a každá dimenzia obsahuje niekoľko indikátorov kvality. Najnižšou úrovňou sú metódy merania. Na výpočet každého indikátora kvality existuje jedna alebo viac kvalitatívnych alebo kvantitatívnych metód merania.
Multidimenzionalita prístupu vyplýva zo skutočnosti, že ku kvalite administratívneho zdroja údajov sa pristupuje ako k celku a nie len vzhľadom na údaje. Zatiaľ čo hyperdimenzie, dimenzie aj indikátory kvality sú stabilné, metódy merania sú flexibilné. Tento prístup umožňuje vybrať pre každý indikátor kvality najvhodnejšie metódy merania, čo na druhej strane umožňuje flexibilné hodnotenie administratívnych databáz bez ohľadu na ich typ, na oblasť štatistiky, v ktorej sa využívajú a na spôsob, akým sa využívajú.

RESUME

Administrative data sources contain information collected primarily for administrative purposes. Increasingly, however, administrative data are also used for the purposes other than those for which they were generated, in particular for official statistics. The main benefits of their use are reduced costs, reduced burden on respondents, increased data quality, improved data timeliness, as well as a greater degree of flexibility in satisfying requests for more detailed information (e.g. from territorial terms). The production of administrative data sources is often outside the scope of statistical offices. This means, first and foremost, that the statistical office needs information on the source and the quality of these data.

The aim of the article is to summarize the generally valid theoretical and methodological aspects of the evaluation of administrative data sources used for statistical purposes. These were then used to establish a theoretical and methodological framework for the evaluation of administrative data entering the statistical system.

Measuring the quality of administrative data used for statistical purposes differs from measuring the quality of statistical surveys. Therefore, the existing framework for assessing the quality of statistical surveys needs to be revised. This was done by means of a hierarchical and multidimensional approach.

From a hierarchical perspective, there are four levels. The highest level is hyperdimension. Each hyperdimension consists of several dimensions, and each dimension contains several quality indicators. The lowest level is the measurement method. There are one or more qualitative or quantitative measurement methods for the calculation of each quality indicator.

The multidimensionality of the approach stems from the fact that the quality of an administrative data source is approached as a whole and not only with regard to the data. While hyperdimensions, dimensions and quality indicators are stable, measurement methods are flexible. This approach enables to select the most appropriate measurement methods for each quality indicator, which in turn allows flexible evaluation of administrative databases, regardless of their type, the field of statistics in which they are used and the method of their use.

PROFESIJNÝ ŽIVOTOPIS

***PhDr. Ľudmila Ivančíková, PhD.**, vyštudovala sociológiu na Filozofickej fakulte UK v Bratislave. Od roku 1987 pracuje v Štatistickom úrade SR. Prešla viacerými pozíciami od expertky, vedúcej oddelenia až po súčasnú funkciu generálnej riaditeľky sekcie sociálnych štatistík a demografie. V minulosti sa zaoberala problematikou terénnych zisťovaní a zisťovaní zameraných na meranie životných podmienok. Ako medzinárodná expertka pôsobila v oblasti výberových zisťovaní v sociálnych štatistikách. V centre jej pozornosti je oblasť chudoby, sociálnej inklúzie, životnej úrovne a kvality života.*

***Ing. Boris Vaňo** vyštudoval Vysokú školu ekonomickú v Bratislave, následne absolvoval postgraduálne štúdium z demografie na Karlovej Univerzite v Prahe. Od roku 1980 pracuje v Inštitúte informatiky a štatistiky ako výskumný pracovník v oblasti demografie. V rokoch 2000 – 2014 bol vedúcim Výskumného demografického centra. V období rokov 2006 – 2010 pôsobil ako podpredseda Slovenskej štatistickej a demografickej spoločnosti pre demografiu. Špecializuje sa na hodnotenie populačného vývoja, demografické prognózy a populačnú politiku.*

KONTAKTY

ludmila.ivancikova@statistics.sk
vano@infostat.sk