SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

3/2014
ročník/volume 24

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

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

Článok/Article: 5

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

Strany/Pages: 37 – 48

Dátum vydania/Publication date: 15. júl 2014/July 15, 2014

Boris FRANKOVIČ
Štatistický úrad SR

OCHRANA DÔVERNÝCH ŠTATISTICKÝCH ÚDAJOV V SČÍTANÍ OBYVATEĽOV, DOMOV A BYTOV 2011

STATISTICAL DISCLOSURE CONTROL IN THE 2011 POPULATION AND HOUSING CENSUS

ABSTRAKT
Článok rozpracúva problematiku ochrany dôverných údajov v Sčítaní obyvateľov, domov a bytov 2011 (ďalej „SODB 2011“). Upozorňuje na možné riziká ohrozenia dôvernosti vo frekvenčných tabuľkách, ktoré predstavujú základné výstupy, a navrhuje vhodné riešenia na minimalizáciu týchto rizík. Článok predstavuje metódy ochrany údajov použité pri publikovaní výsledkov SODB 2011 zamerané na zachovanie dôvernosti údajov získaných od jednotlivých respondentov.

ABSTRACT
Article elaborates on the issue of confidentiality in 2011 Population and Housing Census (hereinafter referred to as “SODB 2011”). It calls attention to possible risks threatening the frequency tables that make up the basic outcomes and offers appropriate solutions to minimize them. It represents protection methods used in publishing the results of SODB 2011 designed to maintain the confidentiality of individual respondents.

KĽÚČOVÉ SLOVÁ
ochrana dôverných údajov, sčítanie, frekvenčné tabuľky, odhalenie, swapping

KEY WORDS
statistical disclosure control, census, frequency tables, disclosure, swapping

1. ÚVOD
Sčítanie obyvateľov, domov a bytov je najrozsiahlejším štatistickým zisťovaním. Jeho výsledky sú súčasťou mnohých medzicenzových analýz. Využitie dát zo sčítania je širokospektrálne. Údaje z cenzu o. i. pomáhajú štatistickým orgánom pri výberových zisťovaniach, pretože presne opisujú základný súbor. Záujem zo strany verejnosti je obrovský. Ako bežní občania, tak aj vedecká obec požadujú rôzne výstupy na rôzne účely. Zatiaľ čo výskumníci majú záujem najmä o mikroúdaje, bežní občania, samosprávy a mnohí ďalší žiadajú frekvenčné tabuľky. Tie znázorňujú počet jednotlivcov (prípadne bytov, domov či iných štatistických jednotiek) patriacich do danej oblasti v určitom triedení. Na prvý pohľad by sa zdalo, že takto vytvorené tabuľky nemôžu predstavovať žiadne riziko odhalenia dôverných údajov. Opak je však pravdou.

Pod *dôverným údajom* rozumieme údaj týkajúci sa spravodajskej jednotky a predstavujúci informáciu, ktorá o nej nie je zrejmá ani verejne dostupná. Rôzne premenné majú rôzny stupeň dôvernosti. Zatiaľ čo pohlavie sa nepovažuje za dôverný údaj, najvyššie dosiahnuté vzdelanie či náboženské vierovyznanie nimi bezpochyby sú. Hoci frekvenčné tabuľky nepredstavujú z pohľadu dôvernosti také

riziko, ako napríklad samotné mikroúdaje (obsahujúce všetky hodnoty každej spravodajskej jednotky), pri ich publikovaní je namieste zvýšená opatrnosť. Štatistický úrad sa preto rozhoduje pre optimálnu stratégiu ochrany, aby riziko odhalenia dôverných údajov nepresiahlo prípustnú úroveň za súčasného zachovania informačnej hodnoty údajov.

2. POTREBA OCHRANY

Azda najčastejšou otázkou, ktorá vznikne pri diskusii o ochrane dôverných štatistických údajov, je jej potreba. Je ochrana skutočne nevyhnutná? Odpovede by sa dali hľadať vo viacerých rovinách. Z pohľadu Štatistického úradu SR ide o rozličné aspekty. Najpodstatnejším je zákon. Tak zákon o štátnej štatistike č. 540/2001 Z. z. v znení neskorších predpisov, ako aj nariadenie č. 223/2009 o európskej štatistike vymedzujú povinnosť ochraňovať údaje a ich dôvernosť. Tie môžu byť publikované iba vo forme štatistických informácií, ktoré vznikli sumarizáciou dôverných údajov a neumožňujú priamu ani nepriamu identifikáciu. Je zrejmé, že frekvenčné tabuľky nepredstavujú žiadne riziko priamej identifikácie, keďže neobsahujú žiadne identifikačné údaje. Nepriama identifikácia však v určitých prípadoch môže nastať, a práve to sú situácie, ktorým by štatistické úrady mali primeranými opatreniami zabrániť.

Druhou rovinou je pohľad respondenta. Ten si je vedomý, že údaje v štatistickom zisťovaní poskytol výhradne na štatistické účely, a niektoré svoje údaje, ktoré považuje za citlivé, by nechcel odhaliť verejnosti. Ak má pocit, že takáto situácia nemôže nastať, zvyšuje sa jeho dôvera v štatistický úrad, a tým stúpa pravdepodobnosť, že mu riadne poskytne svoje údaje. To je nepochybne aj v záujme samotného úradu, ktorý sa snaží maximalizovať mieru návratnosti i vierohodnosť poskytnutých údajov.

Na základe toho úrad vie, že aplikácia ochrany je nevyhnutná. Prílišnou ochranou však docieli opak, a to negatívny ohlas verejnosti, ktorá nadobudne pocit, že poskytnuté údaje nespĺňajú svoj účel. Hľadanie prípustnej hranice, kedy publikované údaje nepredstavujú riziko odhalenia údajov, ale ich informačná hodnota je vysoká, je preto najdôležitejším krokom.

Situácia okolo ochrany dôverných štatistických údajov čoraz viac rezonuje aj v zahraničí, osobitne v Eurostate. Prijatie nariadenia č. 223/2009 o európskej štatistike implikovalo nutnosť novelizácie nariadenia č. 831/2002 o prístupe k dôverným údajom na vedecké účely. Celosvetovo sa podporuje poskytovanie údajov pre verejnosť, pretože s rozvojom moderných technológií výrazne rastie objem dát a dopyt po nich. So silnejúcou podporou dátovej infraštruktúry však narastá i potreba ich ochrany. Riziko odhalenia citlivých údajov rastie úmerne s objemom dostupných dát. V prostredí EÚ sa preto tejto problematike venuje čoraz väčšia pozornosť. Eurostat riadi Pracovnú skupinu pre štatistickú dôvernosť, ktorá sa zaoberá štatistickou dôvernosťou, i expertnú skupinu na ochranu dôverných štatistických údajov (Štatistický úrad SR je členom oboch). Najmä prvá je zložená zo zástupcov všetkých členských krajín Európskej únie. Prerokúva a schvaľuje nové metódy a postupy v ochrane dát. Na medzinárodnej úrovni sa pravidelne uskutočňujú školenia o ochrane údajov. S počtom zúčastnených krajín však stúpa aj počet protichodných názorov na jednotlivé metódy ochrany dát. Základné otázky ochrany, medzi ktoré aktuálne patria úvahy o tom, či je ochrana údajov ako taká vôbec

potrebná, nespochybňuje žiadna členská krajina EÚ. Aplikujú ju všetky v závislosti od vlastných možností, kapacít, zákonov i požiadaviek. Sčítanie obyvateľov, domov a bytov 2011 nie je výnimkou.

Základnými výstupmi z SODB 2011 sú frekvenčné tabuľky. Ich štruktúra neumožňuje plnohodnotne vyhovieť cieľom výskumných analýz, preto sa vedecká obec pravidelne obracia na Štatistický úrad SR so žiadosťou o poskytnutie dôverných mikroúdajov z SODB 2011 na výskumné účely. Štatistický úrad SR každú žiadosť osobitne posúdi a rozhodne o vyhovení, resp. nevyhovení žiadosti. V prípade, že mikroúdaje sa výskumnej inštitúcii poskytnú, tá môže voliť spôsob prístupu k nim dvomi spôsobmi. Prvým je práca v „safe centre úradu“. Nachádza sa v budove Štatistického úradu SR na Miletičovej ulici v Bratislave. V ňom majú výskumníci prístup k dôverným mikroúdajom, na ktoré bola aplikovaná iba minimálna miera ochrany (odstránenie priamych identifikátorov). Počas práce v „safe centre úradu“ sú monitorovaní a výsledky ich analýz musia prejsť kontrolou zamestnancami Štatistického úradu SR. Druhou možnosťou je priame zaslanie anonymizovaných mikroúdajov, na ktoré bola aplikovaná vyššia miera ochrany. Okrem odstránenia priamych identifikátorov sa vykonajú zmeny v mikroúdajoch tak, aby nebol poškodený ich výskumný potenciál, ale minimalizované riziko nepriamej identifikácie. Výskumná inštitúcia, ktorá získa prístup k dôverným mikroúdajom na vedecké účely, podpíše so štatistickým úradom zmluvu špecifikujúcu podmienky prístupu k nim, opatrenia na ich ochranu, ako i sankcie v prípade porušenia ustanovení zmluvy.

3. RIZIKÁ ODHALENIA VO FREKVENČNÝCH TABUĽKÁCH

Frekvenčná tabuľka je tabuľka, ktorá predstavuje počty spravodajských jednotiek patriacich do daných oblastí v danom triedení. V prípade SODB 2011 sú to počty obyvateľov (resp. domov, bytov) s určitými charakteristikami v jednotlivých oblastiach, napríklad počet vysokoškolsky vzdelaných obyvateľov v okrese Prievidza, prípadne počet katolíčok v Košiciach. Jednotlivé tabuľky sa líšia samotným triedením a regionálnou úrovňou. So stúpajúcou geografickou presnosťou (a stúpajúcim počtom dimenzií tabuľky) stúpa aj riziko odhalenia dôverných údajov, ktoré je určite vyššie v tabuľke za obec Rakoľuby než v tabuľke za okres Nové Mesto nad Váhom. V tejto časti si ukážeme možné riziká spojené s frekvenčnými tabuľkami, ako sa píše v [3].

a) Identifikácia

Identifikácia nastáva vtedy, ak votrelec (osoba, ktorá má tabuľku k dispozícii a snaží sa identifikovať dôverné údaje, teda použiť tabuľku na iný účel, ako je určená) nadobudne vedomosť o malom počte respondentov v danom triedení. Najčastejšie tak identifikácia nastáva vtedy, ak je v tabuľke prítomná hodnota 1. Tá nám hovorí, že s danými vlastnosťami je v danom území prítomný iba jeden človek. Bez ďalších okolností identifikácia nepredstavuje žiadne riziko. V spojitosti s inými tabuľkami, externými informáciami alebo pri špecifickej situácii však môže byť nebezpečná. Štandardne sa hodnoty 1 a 2 jednotne považujú za citlivé, čo môže spôsobiť neúmernú informačnú stratu. Je preto potrebné rozlišovať jednotlivé prípady a pristupovať k tabuľkám obozretne. V prípade hodnoty 2 nastáva identifikácia vtedy, ak jeden z dvoch je v pozícii votrelca (scenár „zvedavého suseda“). Získava tak informáciu, že okrem neho je v danej oblasti už iba jeden respondent s danými charakteristikami. Ilustrujme to na jednoduchom príklade.

Tabuľka č. 1 nám okrem iného hovorí, že v danom regióne je iba jedna slobodná žena so základným vzdelaním, čím nastáva identifikácia. Táto je napriek tomu bezriziková. Informácia totiž nie je natoľko bohatá, aby nám povedala, ktorá žena je slobodná a iba so základným vzdelaním. V regióne sa nachádza 18 žien so základným vzdelaním, 9 slobodných. Aby sme teda dokázali rozpoznať konkrétnu, museli by sme ju osobne poznať a v tom prípade nám tabuľka č. 1 neodhalí dôverný údaj. Ak by sa jedna z rozvedených žien so základným vzdelaním našla v tabuľke, získala by vedomosť o jedinej inej žene s touto charakteristikou. K odhaleniu dôverných údajov by však na základe predošlej úvahy nedošlo.

Tabuľka č. 1: Ilustrácia možnosti identifikácie

Vzdelanie Rodinný stav	Muži			Ženy			Spolu
	základné	stredné	vysokoškolské	základné	stredné	vysokoškolské	
Ženatý/ vydatá	30	6	4	15	3	3	61
Rozvedený/ rozvedená	3	4	0	2	4	3	16
Slobodný/ slobodná	3	0	3	1	4	4	15
Spolu	36	10	7	18	11	10	92

b) Odhalenie individuálnej vlastnosti

Toto riziko je prirodzeným pokračovaním identifikácie, ktorá je pre tento typ odhalenia nutná. Nastáva vtedy, keď z publikovanej tabuľky možno zistiť citlivý údaj o konkrétnom respondentovi. To sa môže udiať buď získaním údajov z viacerých tabuliek, prípadne z iných zdrojov, alebo keď sumárna hodnota v riadku alebo stĺpci tabuľky je 1 a tento riadok či stĺpec vyjadruje úplný rozklad citlivej premennej (napríklad pri národnosti). Pod úplným rozkladom rozumieme obsiahnutie všetkých kategórií, ktoré pre danú premennú môžu nastať. Situáciu ilustrujeme na nasledujúcich príkladoch.

Tabuľka č. 2: Bezpečná tabuľka z pohľadu odhalenia individuálnej vlastnosti

Národnosť Rodinný stav	Muži			Ženy			Spolu
	slovenská	maďarská	ukrajinská	slovenská	maďarská	ukrajinská	
Ženatý/ vydatá	30	6	2	14	3	2	57
Rozvedený/ rozvedená	3	4	0	1	0	1	9
Slobodný/ slobodná	0	0	1	2	1	1	5
Spolu	33	10	3	17	4	4	71

Tabuľka č. 3: Príklad odhalenia individuálnej vlastnosti

Národnosť Rodinný stav	Muži			Ženy			Spolu
	slovenská	rómska	iná	slovenská	rómska	iná	
Ženatý/ vydatá	30	6	2	14	3	2	57
Rozvedený/ rozvedená	3	4	0	1	1	0	9
Slobodný/ slobodná	0	1	0	2	1	1	5
Spolu	33	11	2	17	5	3	71

Tabuľka č. 2 nepredstavuje riziko, pretože citlivá premenná národnosť netvorí úplný rozklad. Vidíme jediného slobodného muža ukrajinskej národnosti, avšak z charakteru tabuľky vyplýva, že v danom regióne môžu byť aj ďalší muži iných národností. V prípade tabuľky č. 3 však už k odhaleniu individuálnej vlastnosti dochádza. Jediný slobodný muž v oblasti má rómsku národnosť. Podobne rozvedená Slovenka vie, že druhá žena v oblasti je Rómka. Táto situácia sa nazýva scenár „zvedavého suseda“ a je, pochopiteľne, oveľa menej pravdepodobná ako prvá, preto predstavuje aj menšie riziko.

Uvažujme tabuľky. Ide o tabuľky publikované z toho istého zdroja, znázorňujúce počty obyvateľov obcí v závislosti od vzdelania a veku. Z tabuľky č. 4 identifikujeme muža z obce Rovná, ktorý má viac ako 80 rokov, čo však samo osebe nie je odhalenie dôverného údaja. Porovnaním údajov z tabuľky č. 5 dokážeme zistiť, že tento muž má iba základné vzdelanie. Ktokoľvek, kto má vedomosť o najstaršom občanovi obce Rovná, dokáže odhaliť úroveň jeho vzdelania.

Tabuľka č. 4: Odhalenie jednotlivca spojením tabuliek (1. časť)

Vek Obec	Muži			Ženy			Spolu
	menej ako 60	60 – 80	80+	menej ako 60	60 – 80	80+	
Dlhá	40	5	3	32	8	5	93
Rovná	25	4	1	27	6	0	63
Široká	35	7	2	39	10	5	98
Spolu	100	16	6	98	24	10	254

Zdroj: [1]

Takýto typ odhalenia môže vzniknúť pri publikovaní veľkého množstva tabuliek z jedného zdroja, čo je aj prípad SODB 2011. Jeho vyčerpávajúci charakter iba zvyšuje riziko tohto typu odhalenia. Jedným z riešení, ako mu zamedziť, je nedovoliť prítomnosť hodnoty 1 v žiadnom z výstupov, čo však, ako sme už napísali, spôsobuje príliš vysokú informačnú stratu.

Tabuľka č. 5: Odhalenie jednotlivca spojením tabuliek (2. časť)

Vzdelanie Obec	Základné			Stredné			Spolu
	menej ako 60	60 – 80	80+	menej ako 60	60 – 80	80+	
Dlhá	10	3	0	20	5	4	42
Rovná	8	1	1	15	4	0	29
Široká	11	3	0	22	6	3	45
Spolu	29	7	1	57	15	7	254

Zdroj: [1]

c) Odhalenie skupinovej vlastnosti

Ide o veľmi podceňovaný typ odhalenia, keď všetci respondenti patria do jednej kategórie citlivej premennej, ktorá tvorí úplný rozklad. Odhalenie skupinovej vlastnosti nevyžaduje identifikáciu, preto sa naň často zabúda. Zvýšená pozornosť venovaná práve tomuto riziku je preto namieste. Uvažujme o obsahu tabuľky č. 6.

Tabuľka č. 6: Príklad odhalenia skupiny

Obec Náboženstvo	Nižné	Vyšné	Spolu
Katolícke	0	20	20
Evanjelické	0	15	15
Jehovovi svedkovia	5	22	27
Iné	0	20	20
Spolu	5	77	82

Všetci obyvatelia obce Nižné patria k Jehovovým svedkom. V tomto prípade nebola potrebná identifikácia. Pozitívom je, že takáto situácia nastáva iba zriedka, oveľa častejšie je odhalenie individuálnej vlastnosti. Pri zvážení scenára „zvedavého suseda“ môže nastať situácia ilustrovaná v tabuľke č. 7, keď obyvateľ s katolíckou vierou získava informáciu, že všetci ostatní sú evanjelici.

Tabuľka č. 7: Príklad odhalenia skupiny jednotlivcom

Obec Náboženstvo	Nižné	Vyšné	Spolu
Katolícke	1	20	21
Evanjelické	6	15	21
Jehovovi svedkovia	0	22	22
Iné	0	20	20
Spolu	7	77	84

Tabuľka č. 8 ilustruje situáciu, keď skupina respondentov dominuje v citlivej kategórii. Jeden z dvoch katolíkov obce Nižné vie, že zo zvyšných 17 je 16 evanjelikov.

Tabuľka č. 8: Príklad dominujúcej citlivej skupiny

Obec Náboženstvo	Nižné	Vyšné	Spolu
Katolícke	2	20	22
Evanjelické	16	15	31
Jehovovi svedkovia	0	22	22
Iné	0	20	20
Spolu	18	77	95

d) Odhalenie odčítaním

Tento typ odhalenia nastáva vtedy, ak viaceré publikované tabuľky majú rovnakú štruktúru, pričom jedna je podmnožinou druhej. Vtedy je možné odčítať jednotlivé hodnoty a dostať tak novú tabuľku umožňujúcu niektoré z predošlých typov odhalení (identifikácia, individuálne alebo skupinové odhalenie). Príklad je znázornený v tabuľkách č. 9 a 10, kde votrelec dokáže pomocou odčítania vytvoriť tabuľku za časť obce Lančár (hodnoty nie sú skutočné).

Tabuľka č. 9: Príklad odhalenia odčítaním (1. časť)

	Kočín-Lančár		
Pohlavie Rodinný stav	muži	ženy	spolu
Ženatý/ vydatá	15	20	35
Slobodný/ slobodná	20	15	35
Rozvedený/ rozvedená	28	22	50
Iný	16	20	36
Spolu	79	77	156

Tabuľka č. 10: Príklad odhalenia odčítaním (2. časť)

	Kočín		
Pohlavie Rodinný stav	muži	ženy	spolu
Ženatý/ vydatá	9	20	29
Slobodný/ slobodná	15	15	30
Rozvedený/ rozvedená	20	21	41
Iný	10	20	30
Spolu	54	76	130

e) Prepájanie tabuliek

Ide o najsofistikovanejší typ odhalenia, keď votrelec dokáže úsudkom, poznaním vzťahov medzi premennými a tabuľkami odvodiť nové tabuľky, ktoré umožnia odhalenie predošlých typov. Proti takémuto riziku sa môže úrad brániť len veľmi ťažko. V okamihu sprístupnenia údajov nedokáže predvídať všetky možnosti, ktoré poskytne prepájanie už zverejnených tabuliek s údajmi a tabuliek, ku ktorým sa verejnosť dostane v budúcnosti. Prijateľným riešením je štandardizovaná diseminácia výstupov, t. j. na rovnakej geografickej úrovni, s rovnakými kategóriami a s rovnakými

pravidlami na ochranu. Prípad SODB 2011, ktoré tvorí podklad na rozsiahlu publikačnú činnosť, toto riziko iba zvyšuje.

4. METÓDY OCHRANY FREKVENČNÝCH TABULIEK

V tejto kapitole vysvetlíme jednotlivé metódy, ktoré dokážu minimalizovať riziká odhalenia dôverných údajov pri maximálnom možnom zachovaní ich využitia. Metódy ochrany frekvenčných tabuliek rozdeľujeme do troch základných skupín podľa ich konkrétneho využitia, a to pred vytvorením tabuliek, počas vytvárania tabuliek (ich prestavbou) a po vytvorení tabuliek. Každá z týchto metód má svoje výhody i nevýhody. Ich kladné a záporné stránky budeme hodnotiť najmä z pohľadu užitočnosti pre potreby ochrany dát z SODB 2011.

a) Pred vytvorením tabuliek

Princíp spočíva v aplikácii ochrany na úrovni mikroúdajov, z ktorých sa zostavujú jednotlivé tabuľky. Tie sa do určitej miery pozmenia, aby modifikácia neuškodila štatistickému účelu, ktorý v sebe majú obsahovať. Výhody tejto metódy prevažujú nad nevýhodami. Medzi záporné stránky patrí zložitejšia implementácia a nižšia miera zrozumiteľnosti pre používateľov. Pozitíva sú však obrovské. Všetky vytvorené tabuľky spĺňajú aditívnosť (jednotlivé sumy v nich obsiahnuté sú skutočnými súčtami hodnôt riadkov a stĺpcov), konzistentnosť (identická bunka má v každej tabuľke rovnakú hodnotu) a poskytujú dostatočnú mieru ochrany. Tá nie je smerovaná na všetky citlivé bunky, ale iba na niektoré z nich, čím je vytvorená istá miera neistoty. Votrelec tak nevie, či jeho údajná identifikácia jednotlivca je skutočná alebo iba fiktívna. Takýto systém je veľmi vhodný pri diseminácii veľkého množstva tabuliek z toho istého zdroja, keďže po jednorazovej ochrane kompletnej databázy mikroúdajov nie je pri tvorbe frekvenčných tabuliek potrebná žiadna ďalšia kontrola, čo výrazným spôsobom uľahčuje prácu úradu a zvyšuje komfort používateľov, ktorí svoje požadované výstupy môžu dostať vo veľmi krátkom čase. Metóda, ktorá sa javí ako najvhodnejšia, je *targeted record swapping*, navrhnutá v Office for National Statistics (UK) ako základná metóda na ochranu dát z cenzu 2011 vo Veľkej Británii.

Jej postup je uvedený v [2], pričom základným kameňom je výmena hodnôt geografických premenných, ktoré sa považujú za najidentifikujúcejšie prvky. Ako sa píše v [1], „v prvom kroku sa v mikroúdajoch identifikuje rizikový jednotlivec, a to vypočítaním jeho rizikového skóre. Domácnosť, ktorej je členom, sa označí ako riziková. Rizikové skóre sa vypočíta ako priemer pravdepodobnosti odhalenia vzhľadom na každú rizikovú premennú pre každého jednotlivca na každej geografickej úrovni zvlášť. To, ktoré premenné sa označia ako rizikové, zvolí úrad v rámci procesu ochrany a táto informácia sa nezverejní. Pre každý kraj sa určí *swap miera*, čo je percentuálny podiel bytov v kraji, ktoré sa medzi sebou vymenia. Tá závisí od miery neodpovedí a imputácií v danom kraji, pretože vysoká miera imputácií sama osebe spôsobí veľkú informačnú neistotu. Počet domácností, ktoré budú podliehať *swappingu*, sa pre každú obec určí v závislosti od jej veľkosti (inverzne) a podielu rizikových domácností v nej. Najviac sa tak ochrana dotkne malých obcí s veľkým počtom rizikových domácností. V každej obci sa následne náhodným výberom s pravdepodobnosťou výberu úmernou rizikovému skóre vyberie množina domácností, ktoré sa ochránia. Ku každej domácnosti vo výbere sa nájde iná domácnosť v inej obci v tom istom okrese spĺňajúca vekovo-pohlavnú štruktúru spolu s niektorými ďalšími základnými ukazovateľmi a tieto dve obce si vymenia

adresy.“[1] Tento postup sa aplikuje na všetky databázy SODB 2011 (bytové domácnosti, cenzové domácnosti, hospodáriace domácnosti) zvlášť. Priamo úmerné *swap miere* sú zmeny v mikroúdajoch. Ak je vysoká, môže sa stať, že štatistiky na obecnej úrovni môžu byť do istej miery skreslené, najmä v prípade premenných, ktoré netvoria podklad na hľadanie vhodných domácností. Vzťahy medzi jednotlivými premennými však nikdy nebudú narušené, keďže jediný typ aplikovanej zmeny spočíva vo výmene adries. Ak v danom okrese vhodná domácnosť neexistuje, postupujeme viacerými spôsobmi. Najskôr sa môžeme pokúsiť nájsť prislúchajúcu domácnosť v rámci toho istého kraja, prípadne na úrovni republiky. Ak taká domácnosť neexistuje, hľadáme najpodobnejšiu domácnosť, pričom mieru podobnosti si stanovíme priradením váh dotknutým premenným. Zvážiac úroveň zmien a časovú náročnosť, ktorú predstavujú ostatné metódy ochrany, považujeme túto metódu za optimálnu.

Aplikáciou metódy tak dostávame ochránený mikrodátový súbor, z ktorého možno publikovať ľubovoľné štandardné i neštandardné výstupy. Jej náročnosť je však nezanedbateľnou stránkou; použitie tohto postupu výrazne predĺži čas medzi zberom údajov a prvými výstupmi.

b) Prestavba tabuliek

Tabuľkám určeným na publikáciu sa pri tomto postupe agregujú triediace premenné na vyššiu úroveň (napríklad z obce na okres). Vznikajú tak výstupy s menším počtom citlivých buniek, ale aj s nižšou úrovňou podrobností , ktorá môže byť v niektorých prípadoch neakceptovateľná. Výhodami sú však jednoduchá aplikácia, ponechanie skutočných hodnôt bez skreslenia, zrozumiteľnosť pre používateľov, aditivita v rámci tabuliek a konzistencia medzi nimi. Metóda je zobrazená na príklade v tabuľkách č. 11 a 12.

Tabuľka č. 11: Príklad prestavby tabuliek (1. časť)

Národnosť Obec	Slovenská	Maďarská	Rómska	Česká	Iná	Spolu
Dlhá	50	10	4	2	12	78
Rovná	60	11	1	2	4	78
Široká	70	8	2	1	2	83
Spolu	180	29	7	5	18	239

[1] *[1] FRANKOVIČ, B. 2014. Ochrana dôverných štatistických údajov v mikrodátových súboroch. In: Zborník príspevkov z medzinárodnej vedeckej konferencie Štatistického úradu SR Potrebuje ešte Slovensko po sčítaní 2011 ďalší cenzus? Bratislava: Štatistický úrad SR, 2014. 128 s. ISBN 978-80-8121-364-9.*

Tabuľka č. 12: Príklad prestavby tabuliek (2. časť)

Národnosť Obec	Slovenská	Maďarská	Iná	Spolu
Dlhá	50	10	18	78
Rovná	60	11	7	78
Široká	70	8	5	83
Spolu	180	29	30	239

Ak je prestavba tabuliek nedostačujúca, pristupujeme k metódam po vytvorení tabuľky.

c) Po vytvorení tabuliek

Najprirodzenejšou metódou ochrany citlivých hodnôt v už vytvorených tabuľkách je ich zakrytie. Takáto bunka sa však dá vypočítať pomocou sumárnych hodnôt v riadkoch a stĺpcoch, zakrytie ďalších buniek je preto nevyhnutnosťou. Tie sa volia tak, aby vznikla čo najmenšia informačná strata. Je ľahko predstaviteľné, že takto ochránená tabuľka implikuje značnú informačnú stratu, pretože nie sú zverejnené ani bunky, ktoré nevytvárajú žiadne riziko. V prípade veľkého počtu výstupov je táto metóda veľmi ťažko implementovateľná, keďže konzistentnosť medzi tabuľkami (zakrytá bunka v jednej tabuľke bude zakrytá i v druhej) je z dôvodu časového spektra (publikácia tabuliek v čase) nedosiahnuteľná. Aj napriek svojej zrozumiteľnosti vzhľadom na používateľa je tak zakrytie buniek vhodné iba na lokálne použitie, nie ako hlavná metóda pre rozsiahly systém tabuliek.

Ďalšou možnosťou je zaokrúhľovanie hodnôt buniek. Zvolí sa báza (malé číslo, napríklad 3) a každá hodnota v tabuľke sa zaokrúhli na jej násobok. V závislosti od variácie metódy zaokrúhľuje sa buď na *najbližší násobok*, *náhodný násobok*, alebo *kontrolovane*. Zatiaľ čo prvé dve metódy nie sú aditívne, pri kontrolovanom zaokrúhľovaní sa priradenie násobkov bázy rieši v zmysle výsledných súm riadkov i stĺpcov. Spoločným menovateľom všetkých typov zaokrúhľovania je však nekonzistentnosť vedúca k prelomiteľnej ochrane. Spolu s nadpriemerne vysokou informačnou stratou sa tak táto metóda javí ako nie najvhodnejšia a v prostredí Štatistického úradu SR sa neaplikuje. Príklad zaokrúhlenia na najbližší násobok bázy je uvedený v tabuľkách č. 13 a 14, pričom za základnú bázu je stanovená hodnota 5. Hodnoty, ktoré sú násobkami bázy, sa nemenia a hodnoty menšie ako 5 sa zaokrúhlia na hodnotu 5 (v opačnom prípade by hodnota 0 implikovala pôvodnú hodnotu menšiu ako 3).

Tabuľka č. 13: Príklad zaokrúhlenia hodnôt (1. časť)

Národnosť Obec	Slovenská	Maďarská	Iná	Spolu
Dlhá	51	12	18	81
Rovná	63	11	3	77
Široká	68	1	5	74
Spolu	182	24	26	232

Tabuľka č. 14: Príklad zaokrúhlenia hodnôt (2. časť)

Národnosť Obec	Slovenská	Maďarská	Iná	Spolu
Dlhá	50	10	20	80
Rovná	65	10	5	75
Široká	70	5	5	75
Spolu	180	25	25	230

5. ZÁVER

Sčítanie obyvateľov, domov a bytov predstavuje výrazný míľnik v oblasti štatistiky. Jeho výstupy sú kľúčové pre celú spoločnosť. Presne opisujú celú populáciu a umožňujú na jej základe vytvárať relevantné závery či štúdie. Štatistický úrad SR ich poskytuje verejnosti v pokiaľ možno najširšom objeme. Táto snaha však naráža na prekážku ochrany dôvernosti štatistických údajov. Scenáre odhalenia ilustrované v 3. kapitole sú v prípade SODB 2011 oveľa väčšou hrozbou ako v prípade výberových zisťovaní, preto sa im musí venovať značná pozornosť. Zachovať citlivé údaje občanov utajené je podstatnou a neoddeliteľnou súčasťou každodennej publikačnej činnosti. Odbor metód štatistických zisťovaní, ako aj informačný servis Štatistického úradu SR venujú tejto oblasti značnú pozornosť a výsledkom je primeraná ochrana dôverných štatistických údajov vo výsledkoch a výstupoch z SODB 2011. Jednotliví respondenti tak nemusia mať akékoľvek pochybnosti o prípadnom zneužití ich dôverných údajov.

Vzhľadom na náročnosť implementácie metódy *targeted record swapping* a zámer nepredlžovať obdobie čakania na prvé výsledky Štatistický úrad SR aplikoval na ochranu výsledkov SODB 2011 metódu prestavby tabuliek. V určitých prípadoch, keď sa nenaruší konzistencia tabuliek, pristupuje aj k lokálnemu zakrytiu buniek. Jednotlivé tabuľky hodnotí štatistický úrad z hľadiska ich citlivosti a možného rizika, čím udržuje mieru ochrany na optimálnej úrovni pri maximalizácii informačnej hodnoty tabuliek. Ochrana dôverných štatistických údajov je tak vo výstupoch z SODB 2011 riadne zabezpečená, keďže je v nich minimalizované riziko zneužitia dôverných údajov.

Jedným z najväčších výstupov SODB 2011 je Census Hub, elektronické rozhranie určené na prezeranie výsledkov sčítania 2011 v jednotlivých členských štátoch Únie a na Islande, v Lichtenštajnsku, Nórsku a vo Švajčiarsku prostredníctvom hyperkociek. Vzhľadom na to, že štruktúra hyperkociek bola daná nariadeniami, aplikácia prestavby tabuliek nebola možná. Preto pristúpil Štatistický úrad SR k zakrytiu citlivých buniek a niektorých ďalších tak, aby sa citlivé bunky nemohli vypočítať z celkových súčtov. Zameral sa pritom iba na tie hyperkocky, ktoré predstavovali riziko odhalenia dôverných údajov, pričom sa kládol dôraz najmä na informačnú hodnotu hyperkociek. Priemerná informačná strata dotknutých hyperkociek sa tak pohybuje na úrovni cca 10 %, čo je vzhľadom na malú populáciu Slovenskej republiky prijateľná hodnota. Ako nástroj na riešenie bol zvolený softvér *tau-Argus*, ktorý vznikol v rámci projektu ESSNet[2] na ochranu dôverných štatistických údajov. Tomu sa podarilo efektívne zvládnuť hierarchickú štruktúru hyperkociek

[2] *European Statistical System Network.*

a primárne zvolené citlivé hodnoty doplnil ďalšími zakrytiami. Tie členské krajiny, ktoré neaplikovali *record swapping* a zaokrúhľovanie ako primárnu metódu na ochranu, sa taktiež priklonili k zakrývaniu buniek hyperkociek, čo je vzhľadom na ich presne stanovenú štruktúru a konečný počet prijateľné riešenie.

Na nasledujúce sčítanie, ktoré sa uskutoční v roku 2021, otestuje Štatistický úrad SR možnosti použitia *targeted record swapping* ako alternatívnej metódy k aktuálne vykonávanému spôsobu ochrany. Na údaje z administratívnych zdrojov, ktorých využitie sa predpokladá, sa budú dať aplikovať rovnaké metódy ochrany ako na údaje zo zisťovania, keďže tie nerozlišujú medzi pôvodom údajov, ale zohľadňujú iba ich dôvernosť.

LITERATÚRA

[1] FRANKOVIČ, B. 2014. Ochrana dôverných štatistických údajov v mikrodátových súboroch. In: Zborník príspevkov z medzinárodnej vedeckej konferencie Štatistického úradu SR Potrebuje ešte Slovensko po sčítaní 2011 ďalší cenzus? Bratislava: Štatistický úrad SR, 2014, s. 121 – 129. ISBN 978-80-8121-364-9.
[2] FREND, J. et al. 2011. Statistical Disclosure Control for Communal Establishments in the UK 2011 Census. Tarragona: Joint UNECE/Eurostat work session on statistical data confidentiality. Dostupné na http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/15_UK.pdf
[3] HUNDEPOOL, A. et al. 2010. Handbook on Statistical Disclosure Control. ESSNet SDC. Dostupné na http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf

RESUMÉ

V článku sme sa venovali ochrane dôverných štatistických údajov vo frekvenčných tabuľkách, ktoré tvoria základnú formu výstupov zo Sčítania obyvateľov, domov a bytov 2011. Vysvetlili sme, aké možné hrozby tieto tabuľky obsahujú, a spomenuli sme tiež metódy určené na minimalizáciu rizika odhalenia citlivých údajov. V záverečnej časti článku sme predstavili postupy, ktoré používa Štatistický úrad SR pri aplikácii ochrany dôvernosti vo výstupoch SODB 2011, čím zabezpečuje nezverejnenie citlivých údajov obyvateľov.

RESUME

In this article we dealt with the statistical disclosure control in frequency tables, which form the basic form of the output from the Population and Housing Census 2011. We explained what possible threats impose these tables, as well as methods designed to minimize the risk of disclosure of sensitive data. In the final section, we introduced procedures that SOSR applies to protect confidentiality in 2011 Census outputs, ensuring that sensitive data of citizens are not disclosed.

PROFESIJNÝ ŽIVOTOPIS

Boris Frankovič *vyštudoval odbor pravdepodobnosť a štatistika na Fakulte matematiky, fyziky a informatiky UK v Bratislave. Po štúdiu sa zamestnal v Štatistickom úrade SR v odbore metód štatistických zisťovaní. Medzi jeho pracovné oblasti patrí ochrana dôvernosti štatistických údajov, kalibrácia váh štatistických zisťovaní, ako i celkový záber na štatistickú metodiku. Zastupuje Štatistický úrad SR vo viacerých pracovných skupinách Eurostatu.*

KONTAKT

boris.frankovic@statistics.sk