STATISTICKÁ OCHRANA DŮVĚRNOSTI PRO MIKRODATA ZE SČÍTÁNÍ LIDU, DOMŮ A BYTŮ 2011

STATISTICAL DISCLOSURE CONTROL FOR MICRODATA FROM POPULATION AND HOUSING CENSUS 2011

Jiří Novák

Abstract

For quality scientific research, it is essential to have quality data in as much detail as possible. In the area of population statistics, one of the essential sources in the Czech Republic is the population census. The detailed values for the collected socio-demographic variables are the subject of personal data protection, and their dissemination has to be carefully controlled. A promising method that will enable the dissemination of microdata with a multidimensional structure is the synthetic simulation of microdata, in which a new dataset with a similar structure to the original data is created. Thanks to this method, microdata that would otherwise remain hidden will be able to be disseminated. The contribution presents the results of comparing selected synthetic simulation models with the original dataset from the Population and housing census 2011.

Key words: population census, microdata, statistical disclosure control, synthetic, confidentiality

JEL Code: C13, Cl8, C80

Úvod

Pro kvalitní vědecký výzkum je nezbytné mít dostupná data na co nejpodrobnější úrovni. V oblasti statistiky obyvatelstva je jedním z nejdůležitějších zdrojů v České republice nabízejících tento detail populační census. Podrobné hodnoty sociodemografických proměnných, které jsou shromažďovány, jsou ale předmětem zvýšeného zájmu z hlediska ochrany osobních údajů a jejich poskytování veřejnosti nebo i vědecké obci musí být pečlivě kontrolováno. Slibnou metodou, která by mohla umožnit šíření mikrodat s vícerozměrnou strukturou, je syntetická simulace mikrodat, při které vzniká nový datový soubor s podobnou strukturou jako měla původní data. Metoda simulace syntetických mikrodat je přístup, který by

mohl umožnit statistickým úřadům a agenturám publikovat mikrodata, která by jinak musela zůstat skrytá a chráněná.

Mikrodata jsou datové soubory poskytující informace o sadě proměnných pro každého jednotlivého respondenta. Tyto údaje slouží primárně pro účely vědeckého výzkumu a je nezbytné, aby byla zachována důvěrnost citlivých údajů a data byla upravena do stavu, ve kterém není možné identifikovat konkrétního respondenta. Mikrodata z populačního censu představují speciální případ oproti běžným šetřením, jelikož případný narušitel je přirozeně obeznámen se skutečností, že všichni respondenti by měli být zahrnuti v datovém souboru.

Mikrodata mohou být zveřejňována několika způsoby, přičemž každý způsob vyžaduje jiný přístup a odlišné technicko-materiální nároky. Mezi základní metody zveřejňování mikrodat patří například soubory pro veřejné použití (Public Use Files), soubory pro vědecké použití (Scientific Use Files), chráněné výzkumné centrum, vzdálené spouštění či vzdálený přístup.

V současné době jsou mikrodata z Českého statistického úřadu (ČSÚ) publikována ve formě Scientific Use Files v chráněném výzkumném centru, tzv. SafeCentrum ČSÚ, kde jsou mikrodata poskytována specificky jen pro vědecké nebo výzkumné účely na základě zvláštní smlouvy, která má zajistit dodržení podmínek výzkumu, jež neumožní přímé určení zpravodajské jednotky. Mezi mikrodaty nabízenými pro odbornou veřejnost se ovšem nenachází data pocházející z populačních censů.

Tento příspěvek je zaměřen na možnost zveřejňování mikrodat pocházejících z populačních censů ve formě Scientific Use Files prostřednictvím chráněného výzkumného centra, přičemž k tvorbě daných mikrodat by bylo využito metody syntetické simulace mikrodat. Tato metoda vytváří z původního datového souboru nová/umělá/syntetická mikrodata, která neobsahují původní hodnoty, ale zachovává vztahy mezi proměnnými a hierarchickou strukturu obsaženou v datech.

Zpřístupnění těchto mikrodat pomocí metody simulace syntetických dat by umožnilo publikování dat velmi blízkých realitě za současného zachování velmi nízké úrovně rizika prozrazení individuálních údajů o respondentech. Zpřístupněná data pak mohou sloužit vědcům a výzkumníkům k rozvoji a vývoji metod na datech co nejbližších realitě, či lektorům, kteří mají zájem vyučovat pomocí datových souborů blízkých reálné populaci.

V minulosti statistické úřady publikovaly pouze výstupy v agregované podobě, jako jsou například tabulky, mapy či grafy. Stále více jsou ale žádána detailní mikrodata a je nezbytné, aby se stala standardním výstupem statistického úřadu.

2 Statistická ochrana důvěrnosti

Pokud statistický úřad publikuje citlivé údaje, musí je zabezpečit před takzvaným prozrazením (v angličtině je používán technický termín *disclosure*). K tomuto prozrazení dochází, pokud cizí osoba nebo organizace rozpozná nebo se dozví něco nového, co o jiné osobě nebo organizaci doposud nevěděla. Soubor metod ke snižování rizika prozrazení informací o jednotlivcích, podnicích nebo jiných organizacích se nazývá Ochrana důvěrnosti statistických dat (v angličtině *Statistical disclosure control* – SDC).

V současné době již existuje poměrně velký výběr metod ochrany zveřejňovaných dat. Metody navržené pro ochranu mikrodat jsou podrobně popsány v (Hundepool et al., 2012). SDC metody rozlišujeme na neperturbativní metody, perturbativní metody a simulace syntetických dat.

Neperturbativní metody snižují množství informací obsažených ve zveřejněném souboru prostřednictvím zmenšování detailu v datech. Typickými neperturbativními metodami jsou potlačení buněk (anglicky *cell suppression*) a globální překódování (anglicky *global recoding*). Výhoda těchto metod spočívá v tom, že nemění data (neperturbují), místo toho buňky potlačí v rámci tabulky nebo sloučí některé kategorie proměnných do jedné velké kategorie.

Perturbativní metody se snaží zachovat většinu původně shromážděných informací, ale ochrana dat spočívá ve změnách (perturbacích) vybraných hodnot. Tyto metody tedy záměrně lehce mění data. Dle (Antal, Enderle and Giessing, 2017) je ale ztráta informací, která je způsobena perturbativními metodami často na nižší úrovni, než jaká by byla vytvořena neperturbativními metodami. Pro případné uživatele to znamená, že data by pro ně měla být užitečnější. Další výhodou perturbativních metod je to, že většinou nemění strukturu dat. Typickými příklady perturbativních metod jsou výměna záznamů (anglicky *record swapping*) nebo přidávání náhodného šumu (anglicky *adding random noise*). Tyto metody jsou používány na tabulární výstupy z populačních censů na základě doporučení Eurostatu.

Simulace syntetických dat je metoda, která vytváří nová data na základě původních mikrodat. Výhodou této metody je, že lze simulovat datové sady blízké realitě za současného zachování velmi nízkého rizika zveřejnění. Tato metoda byla vybrána na ČSÚ pro ochranu mikrodat z populačního censu z důvodu její schopnosti bezpečně ochránit publikované výstupy vícerozměrných mikrodat.

Nový syntetický dataset musí být dle (Templ, 2017) realistický, což znamená statisticky ekvivalentní skutečné původní populaci. Měl by tedy splňovat následující charakteristiky.

Rozdělení syntetické populace podle regionu nebo vrstvy (straty) by mělo být kvazi-identické s rozdělením původní populace. Marginální rozdělení a korelační struktura mezi proměnnými by měly být co nejpřesněji reprezentovány. Měla by být zachována heterogenita mezi podskupinami, zejména v regionálním aspektu. Nové záznamy syntetické populace by neměly vzniknout čistou replikací z původních dat.

Pro vlastní simulaci mikrodat je využíván postup založený na modelu navržený (Alfons et al., 2011) a (Templ and Filzmoser, 2014). V počátečním kroku je replikována struktura populace dle struktury domácnosti podle věku a pohlaví. Následně jsou simulovány kategorické proměnné pomocí multinomiální logistické regrese či metod založených na rozhodovacích stromech, kdy je prováděn náhodný výběr z pozorovaných podmíněných rozdělení v rámci každé kombinace straty, věkové skupiny a pohlaví. V dalším kroku jsou pak generovány spojité a polospojité proměnné. U nich jsou navrhovány dva přístupy. V prvním přístupu se provádí imputace kategorizované spojité proměnné pomocí multinomiální logistické regrese následované náhodnými výběry z rovnoměrného rozdělení v rámci vytvořených kategorií, přičemž pro největší kategorie provádíme výběr ze zobecněného Paretova rozdělení. Druhý přístup pak spočívá v použití dvoukrokového regresního modelu využívajícího náhodné reziduální členy. V případě potřeby lze některé spojité proměnné rozdělit na složky (komponenty) pomocí přístupu založeného na podmíněném převzorkování, což je například typické pro proměnné příjmu.

Obecný rámec simulace je tedy pak podle (Templ, 2017) následující:

1) Nastavení struktury domácnosti
2) Simulace kategoriálních proměnných
3) Simulace spojitých proměnných
4) Rozdělení spojitých proměnných na komponenty

V kroku 1) se nastavuje základní struktura domácnosti, což znamená replikaci stávající struktury originálních mikrodat. Proměnné, kterými je tato struktura definována, jsou identifikační čísla osob a domácností, velikost domácnosti a dále také proměnná věku a pohlaví respondentů. Jako proměnná vrstvy (straty) se používá proměnná regionu. Doporučuje se jako tuto základní strukturu nastavit pouze několik proměnných, protože riziko prozrazení je pak nižší. Struktura je simulována samostatně pro každou kombinaci vrstvy a velikosti domácnosti, přičemž počet primárních jednotek se odhaduje pomocí Horvitz-Thompsonova estimátoru (Horvitz and Thompson, 1952).

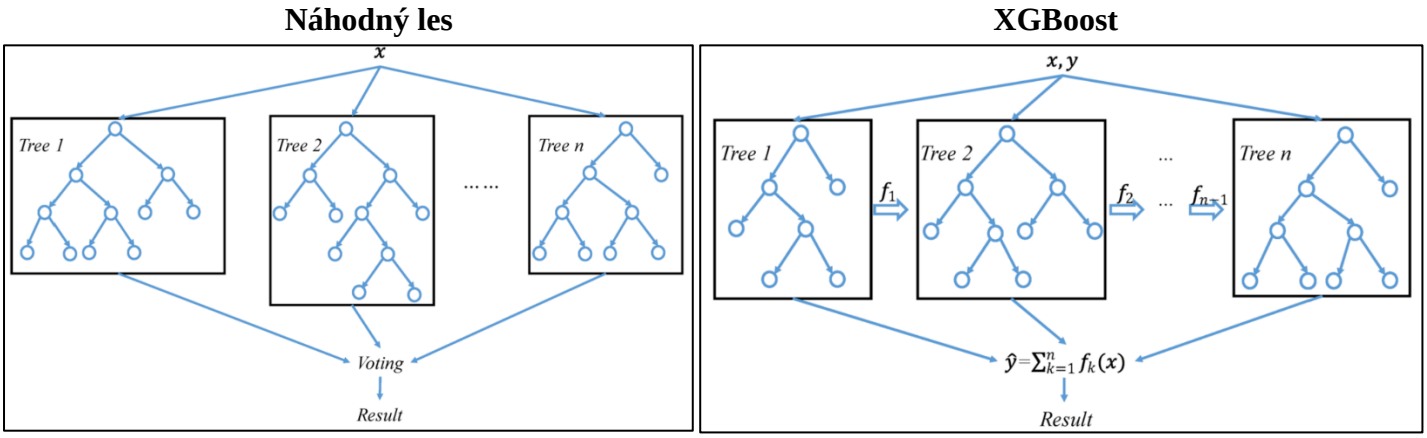
V kroku 2) jsou postupně použity tři různé metody. Těmito metodami jsou multinomická logistická regrese a metody založené na klasifikačním stromu: náhodný les a algoritmus XGBoost.

Pomocí multinomické logistické regrese (Agresti, 2007) jsou odhadována podmíněná rozdělení, přičemž každá vrstva (strata) je pomocí těchto modelů vyrovnávána separátně. Vysvětlovanou proměnnou v regresním modelu je proměnná, která má být simulována a vysvětlující proměnné jsou proměnné věk a pohlaví respondentů, velikost domácnosti a již vygenerované kategorické proměnné. Po odhadu podmíněného rozdělení vysvětlované proměnné se hodnoty simulované proměnné vyplní náhodným výběrem z tohoto podmíněného rozdělení.

Do rodiny metod založených na klasifikačních stromech pak patří náhodný les a algoritmus XGBoost. V rámci klasifikačních stromů se dle (Řezanková, 2020) vytváří stromová struktura pravidel, která umožnuje odhad podmíněných hodnot vysvětlované proměnné. Postupně se vytváří hierarchická stromová struktura, která původní datový soubor rozděluje do stále detailnějších podsouborů. Principem využívaným pro vytváření podmnožin objektů je minimalizace vnitroskupinové variability. Jelikož se jedná o kategoriální proměnné, jako míra variability je zde využívána entropie. Větvení je iterativní proces, kdy je v každém kroku vybírána vysvětlující proměnná, pomocí které se vytvoří podmnožiny s nejmenší vnitroskupinovou variabilitou. V metodě náhodného lesa se dle (Wang et al., 2019) nejdříve provede bootstraping populace pro získání *n* náhodných populací. Následně je vytvořen pro každou novou populaci klasifikační strom. Finální pravidla pro tvorbu stromu jsou založena na většinovém hlasování mezi jednotlivými klasifikačními stromy. XGBoost je iterativní algoritmus pro klasifikační stromy, ve kterém jsou postupně vytvářeny nové klasifikační stromy, přičemž každý strom se učí z chyb/odchylek všech předchozích stromů na základě ztrátové funkce. Obrázek 1 na další straně zobrazuje princip těchto metod na jednoduchém diagramu.

Krok 3) Simulace spojitých proměnných a krok 4) Rozdělení spojitých proměnných na komponenty nebyly v rámci tohoto příspěvku využity z důvodu kategoriálního charakteru mikrodat.

Obr. 1: Princip náhodného lesa a algoritmu XGBoost



Zdroj: (Wang et al., 2019)

Poslední fází analýzy je zhodnocení užitečnosti simulovaných dat a míry informační ztráty. Hodnocení užitečnosti nových syntetických dat záleží na jejich účelu a cíli, pro který jsou vytvářena. Obecně se uvádí, že motivace pro simulaci nových syntetických mikrodat je poskytnout data, která by za normální situace měla zůstat skrytá, ale je ve veřejném zájmu poskytnout tyto údaje například pro výzkumné pracovníky, pro použití v nejrůznějších studiích či pro lektory statistických kurzů jako data blízká realitě, na kterých lze trénovat použití statistických metod. Podle toho, pro koho jsou mikrodata určena se pak bude rozhodovat, jestli jsou pro koncového uživatele užitečná či nikoliv. Vzhledem k tomu, že data pochází z populačního censu, musí být riziko prozrazení velmi nízké, limitně blížící se nule. Tento předpoklad je zaručen výběrem metody simulace plně syntetických populačních dat, kdy všechny hodnoty, které by byly publikovány, budou syntetického charakteru a dataset pak vůbec neobsahuje původní reálné hodnoty. Avšak riziko prozrazení není ani v případě syntetického datasetu nulové, protože se může teoreticky stát, že vytvořený model by byl až moc dobrý a vytvořen by byl identický dataset, jen se syntetickými hodnotami.

Pro zhodnocení informační ztráty se porovnávají vybrané statistiky vypočítané z původních a chráněných datasetů. Dle (Domingo-Ferrer and Torra, 2001) lze pro měření informační ztráty použít přímé srovnání kategorických hodnot, porovnání kontingenčních tabulek, či opatření založená na entropii. V této práci je použito pro zhodnocení informační ztráty porovnání kontingenčních tabulek. Dle doporučení (Antal, Enderle and Giessing, 2017) o metodách doporučených pro použití v populačním censu pro členské země Evropské unie budou k porovnání využity absolutní odchylka, relativní odchylka a Hellingerova vzdálenost mezi původními a syntetickými četnostmi v tabulkách. Souhrnné statistiky, které poté budou

použity pro celkové porovnání jsou průměrná absolutní odchylka, suma relativních vzdáleností a Hellingerova vzdálenost, která je založena na rozdílech mezi odmocninami.

Vizuálním nástrojem pro kontrolu kategoriálních proměnných jsou pak mozaikové grafy (Hartigan & Kleiner, 1981). Mozaikové grafy jsou grafická znázornění vícerozměrných kontingenčních tabulek, která příjemným vizuálním způsobem zobrazují strukturu v kategorických datech. V případě, že by v datasetu byly přítomny spojité proměnné, bylo by pro vizualizaci využito grafů kumulativních distribučních funkcí.

K porovnání rozdílnosti vícerozměrné struktury původního datasetu se syntetickým datasetem je využito metody vícenásobné korespondenční analýzy. Jak popisuje (Hendl, 2012), korespondenční analýza je metoda analogická k metodě hlavních komponent, přičemž je přizpůsobena pro kategoriální charakter dat. V rámci korespondenční analýzy je vypočítáno jak jsou proměnné mezi sebou asociovány a jestli mezi proměnnými existuje nějaký vztah. Nástrojem pro vizualizaci korespondenční analýzy je pak korespondenční mapa, která zobrazuje vztahy a variabilitu v korespondenční tabulce. Obecně platí, že čím blíže jsou řádkové či sloupcové profily v korespondenční mapě, tím podobnější jsou si jednotlivé profily kategorií pro danou proměnnou.

3 Analýza

Analýza prozkoumává metody založené na simulaci syntetických dat, ze kterých byla vybrána metoda simulace plně syntetických populačních dat na základě modelu. Z modelů, které byly vybrány pro generování kategoriální dat, jsou analyzovány výsledky multinomické logistické regrese, náhodného lesa a algoritmu XGBoost. Primárním důvodem pro tento typ ochrany je původ dat, protože u dat ze sčítání lidu je nezbytným cílem zabezpečení maximální ochrany důvěrnosti mikrodat. Veškerá analýza byla provedena v programovacím jazyce R a k syntetické simulaci bylo využito balíčku simPop (Templ et al., 2017).

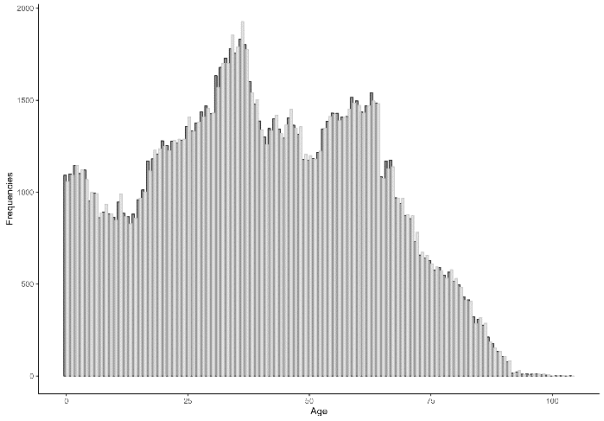
Po simulaci syntetických mikrodat je nezbytné zkontrolovat úspěšnost simulace struktury populace a to, zda informační ztráta, která byla vytvořena, jestli je na přijatelné úrovni. Z důvodu omezeného rozsahu příspěvku jsou dále prezentovány pouze vybrané výsledky pro ilustraci analýzy komparace hodnot na všech proměnných.

V prvním kroku probíhá analýza jednorozměrných proměnných pomocí odchylek vzdáleností absolutní odchylky, relativní odchylky a Hellingerovy vzdálenosti, a vizuální komparace pomocí sloupcových grafů a histogramů. Obrázek 2 zobrazuje porovnání replikovaných hodnost pro domácnosti a jednoleté věkové intervaly. Tmavě šedé sloupečky

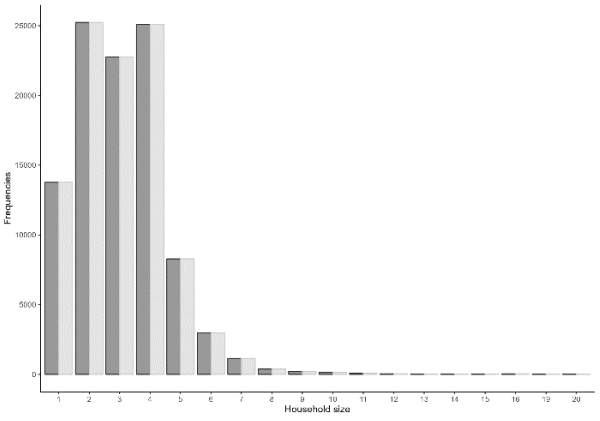
reprezentují hodnoty originálního datasetu a světle šedé sloupečky patří simulovaným hodnotám. Nedochází zde k významným odchylkám a struktura domácností je následně použita pro modelování ostatních proměnných. U jednorozměrné struktury neměl žádný z modelů problém v simulaci proměnných, přičemž ale nejnižších hodnot pro míry odchylek vzdáleností bylo dosaženo u algoritmu XGBoost.

Obr. 2: Věk a počet členů domácnosti

Jednoleté věkové intervaly



Počet členů domácnosti

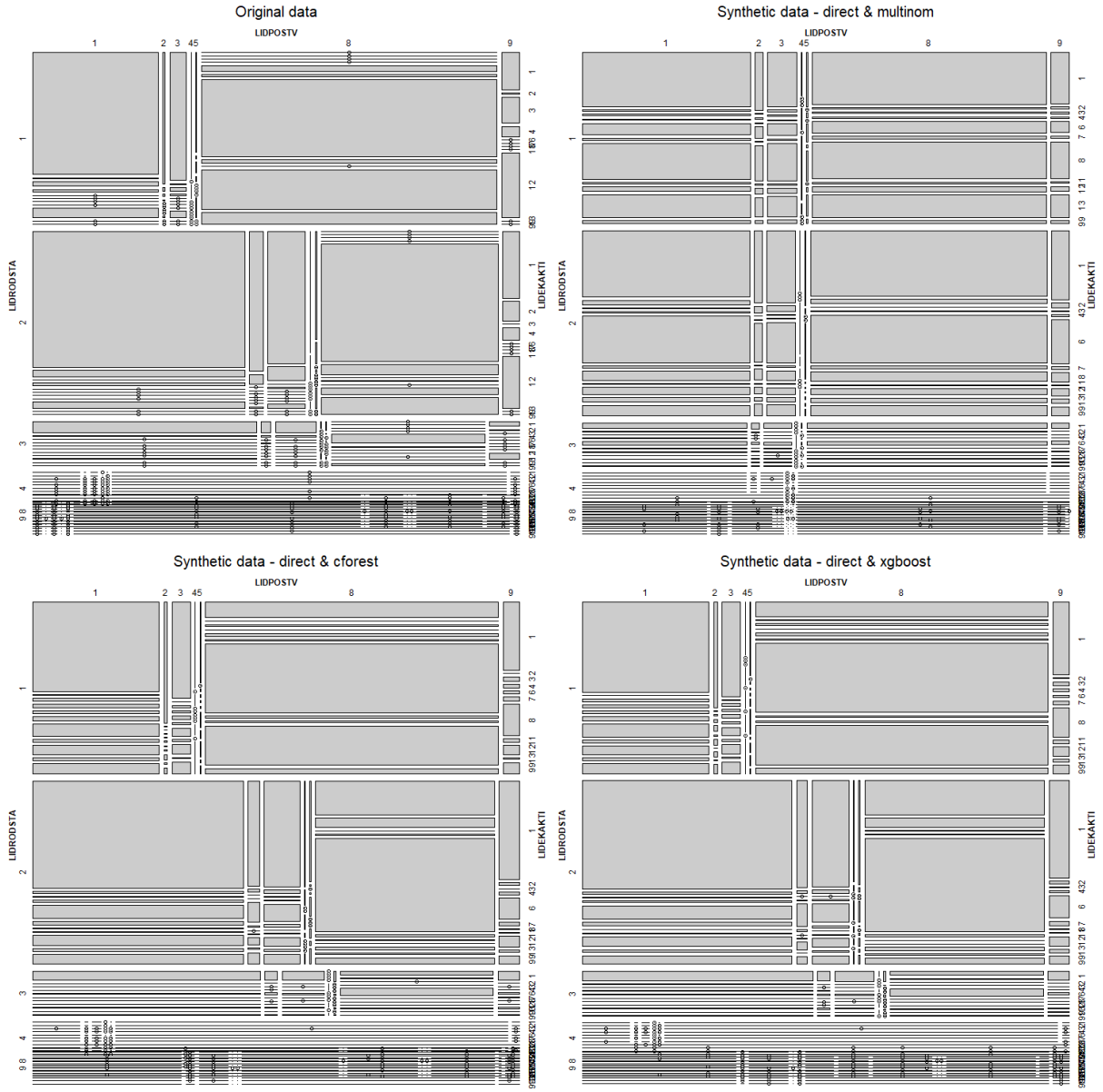


Zdroj: Vlastní zpracování

V dalším kroku probíhá analýza vícerozměrné struktury kontingenčních tabulek, k čemuž je využito mozaikových grafů. Ty představují vizuální nástroj pro komparaci velikosti hodnot daných kombinací proměnných. Obrázek 3 zobrazuje vybrané porovnání pro proměnné Rodinný stav, Postavení v zaměstnání a Typ ekonomické aktivity. V porovnání zde nejhůře dopadly hodnoty nasimulované pomocí multinomické logistické regrese, jelikož velikostí jednotlivých kombinací se nejvíce odlišují od originálních hodnot. Významně lepších hodnot je zde dosaženo u náhodného lesa a algoritmu XGBoost, kterým se tuto kombinaci proměnných podařilo zachytit na dobré úrovni. Nejlepších výsledků v analýze mozaikových grafů bylo dosaženo u algoritmu XGBoost.

Posledním krokem je prozkoumání vybraných výsledků vícerozměrné analýzy podobnosti proměnných, k čemuž bylo využito korespondenční analýzy a z ní vycházející korespondenční mapy, která umožňuje vizuální komparaci rozdílnosti vícerozměrné struktury původního datasetu mikrodat s nasimulovanými syntetickými mikrodaty.

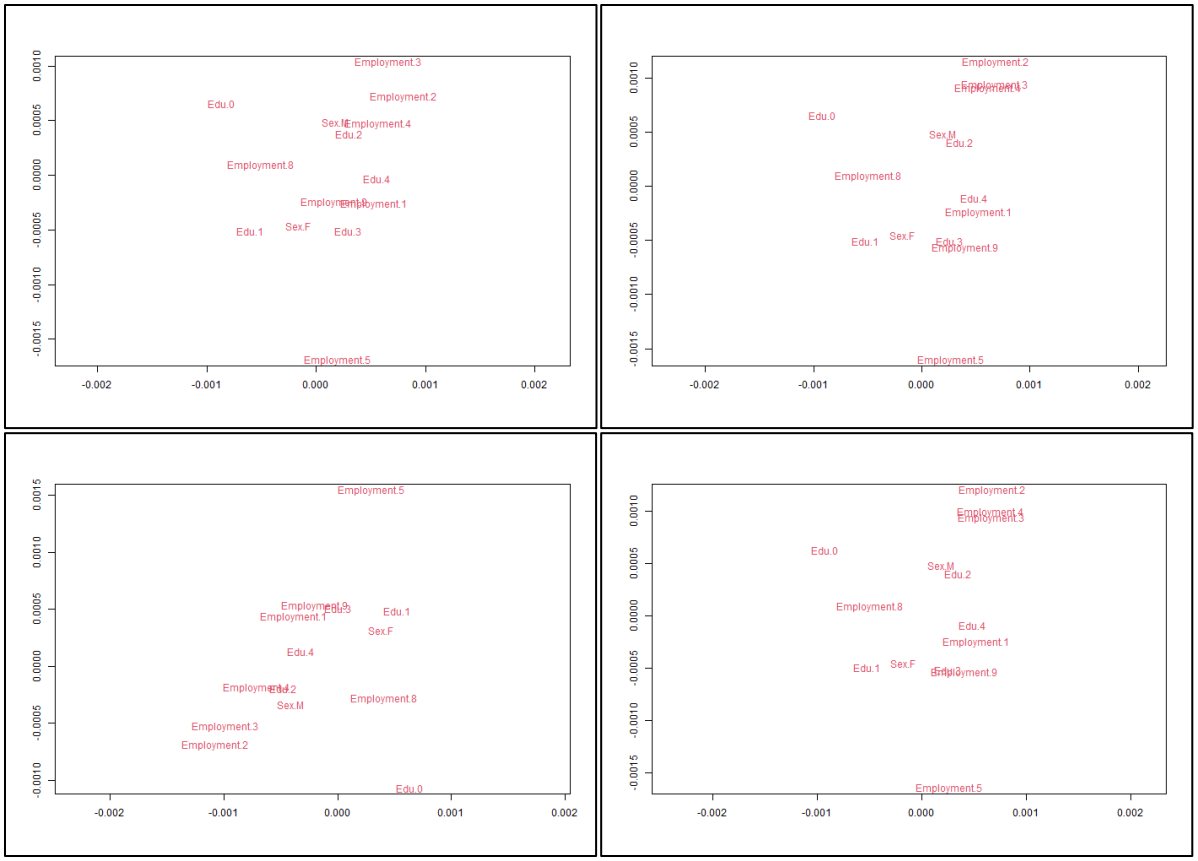
Obr. 3: Mozaikové grafy pro proměnné: Rodinný stav, Postavení v zaměstnání, Typ ekonomické aktivity



Zdroj: Vlastní zpracování

Obrázek 4 zobrazuje porovnání vícerozměrné struktury pro proměnné Typ zaměstnání, Vzdělání a Pohlaví. Nejhůře z porovnání vychází metoda náhodného lesa, jejíž řádkové a sloupcové profily se významně odlišují od původního tvaru sloupcových a řádkových profilů originálních mikrodat. Metoda multinomické regrese a algoritmu XGBoost zde zobrazuje podobné výsledky, které lze považovat za srovnatelné se strukturou původních mikrodat. Odchylky jsou očekávatelné, ale tvar struktury a vzdálenosti profilů jsou přijatelné.

Obr. 4: Korespondenční mapy pro proměnné: Typ zaměstnání, Vzdělání, Pohlaví

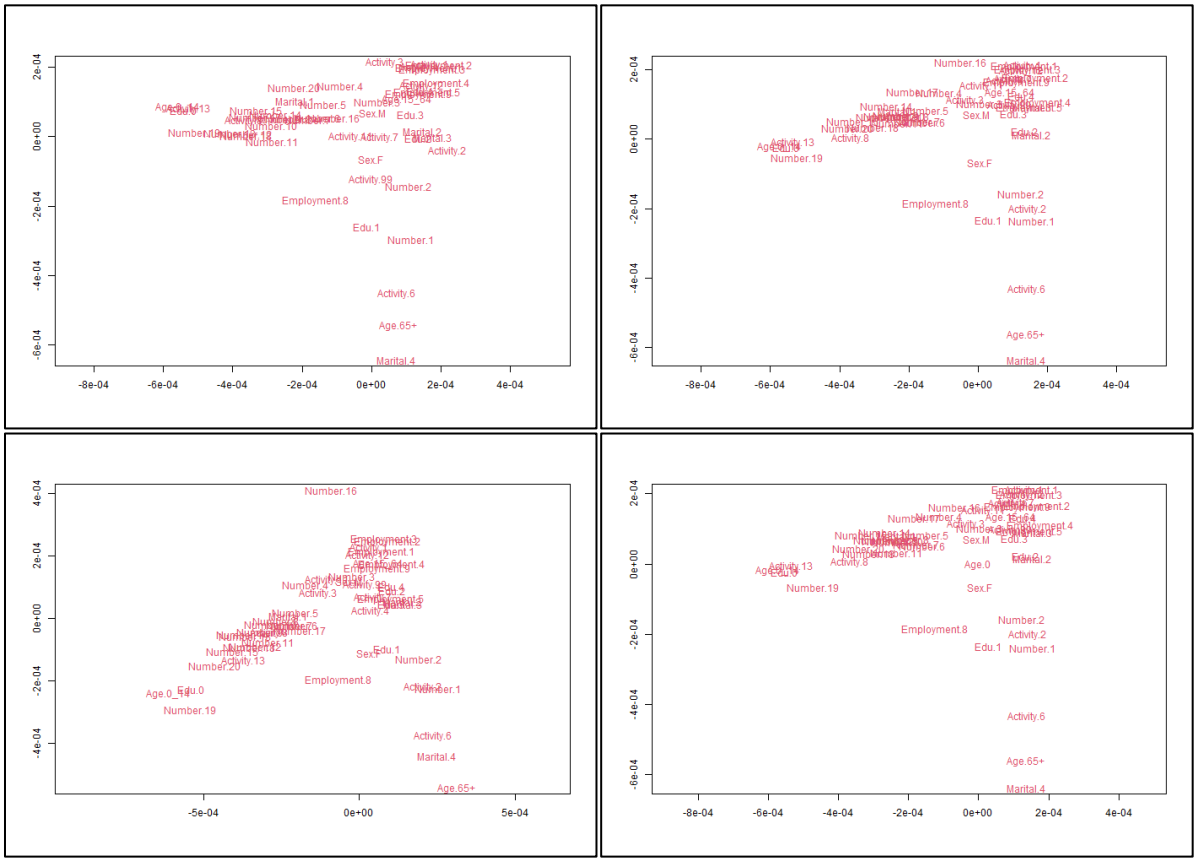


Zdroj: Vlastní zpracování

Obrázek 5 pak zobrazuje porovnání vícerozměrné struktury pro proměnné Počet členů domácnosti, Věk, Pohlaví, Rodinný stav, Postavení v zaměstnání, Ekonomická aktivita a Vzdělání. Nejhorší výsledek lze opět pozorovat u metody náhodného lesa. Lepších výsledků je dosaženo u multinomické regrese a algoritmu XGBoost, které velice obstojně nasimulovaly strukturu dané kombinace proměnných. Řádkové a sloupcové profily zde velice dobře kopírují strukturu, kterou můžeme pozorovat v originálních mikrodatech. Jedná se o mimořádný úspěch obou metod.

Po prozkoumání ostatních výsledků se jako nejlepší model pro simulací syntetických mikrodat jeví využití klasifikačního stromu s algoritmem XGBoost.

Obr. 5: Korespondenční mapy pro proměnné: Počet členů domácnosti, Věk, Pohlaví, Rodinný stav, Postavení v zaměstnání, Ekonomická aktivita, Vzdělání



Zdroj: Vlastní zpracování

Finální rozhodnutí o úspěšnosti simulace závisí na účelu dat, pro analytické účely je větší disproporce syntetických dat oproti originálním datům negativním jevem a data nelze prohlásit za analyticky platná a zajímavá. Na druhou stranu, pokud by účelem dat bylo je poskytnout jen jako dummy dataset, například pro přípravu skriptu později aplikovaného na reálnější data v zabezpečeném Safecentru, pro výukové účely nebo jako prostředek pro vzdálený přístup k datům, pak jsou co největší rozdíly žádané, aby data zachycovala jen technickou strukturu datasetu.

Závěr

Sčítání lidu představuje jeden z nejdůležitějších zdrojů populačních statistik a podrobná data v co největším detailu jsou ve vědecké komunitě velice žádaná. Data tohoto charakteru jsou velice citlivá, a nejen u mikrodat ale i u publikovaných tabulek postupují statistické úřady velice opatrně s jejich diseminací. Mikrodata je nezbytné chránit pomocí těch nejmodernější metod

a tento příspěvek představuje simulaci syntetických mikrodat pomocí metody simulace založené na modelu. Jsou analyzovány následující modely: multinomická logistická regrese, náhodný les a klasifikační strom s algoritmem XGBoost.

Cílem daného příspěvku bylo prozkoumat možnosti simulace syntetických mikrodat z populačního censu a zhodnotit úspěšnost vybraných metod. Tyto metody byly vybrány, kvůli charakteru dat, jelikož u diseminace dat ze sčítání lidu je primárním cílem zabezpečení maximální ochrany důvěrnosti mikrodat. Veškerá analýza byla provedena v programovacím jazyce R a k simulaci bylo využito balíčku simPop. Data, jež byla použita na analýzu, pochází ze Sčítání lidu, domů a bytů z roku 2011.

Na základě analýzy informační ztráty, kdy byly využity vybrané míry odchylek vzdáleností, mozaikové grafy a vícenásobná korespondenční analýza, byl vybrán jako nejvhodnějších model pro simulaci klasifikační strom s algoritmem XGBoost.

Do budoucna bude dále pokračovat zkoumání a analyzování dalších modelů, které mohou být vhodné pro bezpečné publikování mikrodat z populačních censů. Nezbytným krokem pak bude porovnání těchto modelů mezi sebou a následovat bude rozhodnutí o vhodné míře šumu a informační ztráty v mikrodatech, tak aby byla data bezpečná a důvěrná, ale přesto užitečná pro výzkumníky v jejich budoucím výzkumu.

Acknowledgment

This paper has been prepared with the support of a project of the Prague University of Economics and Business – Internal Grant Agency, project No. F4/50/2021.

References

Agresti, A. (2007). *An introduction to categorical data analysis*. John Wiley & Sons.

Alfons, A., Kraft, S., Templ, M., & Filzmoser, P. (2011). *Simulation of close-to-reality population data for household surveys with application to EU-SILC*. Statistical Methods & Applications, 20(3), 383–407. https://doi.org/10.1007/s10260-011-0163-2

Antal, L., Enderle, T. and Giessing, S. (2017) '*Harmonised protection of census data in the ESS*'. Centre of Excellence on Statistical Disclosure Control.

Domingo-Ferrer, J., & Torra, V. (2004). Disclosure risk assessment in statistical data protection. Journal of Computational and Applied Mathematics, 164-165, 285–293. https://doi.org/10.1016/s0377-0427(03)00643-5

Hartigan, J. A., Kleiner, B. (1981). Mosaics for contingency tables. Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, 268–273. https://doi.org/10.1007/978-1-4613-9464-8_37

Hendl, J. (2012). Korespondenční analýza. In Přehled statistických metod zpracování dat: Analýza a metaanalýza dat (Vol. 4, pp. 588–591). chapter, Portál.

Horvitz, D. G., & Thompson, D. J. (1952). *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association, 47(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., de Wolf, P.-P. (2012). *Statistical disclosure control*. John Wiley & Sons Inc.

Řezanková, H. (2020). Methods of selecting explanatory variables in classification trees. Slovak Statistics and Demography, 3(2020), 40–53.

Templ, M., & Filzmoser, P. (2013). Simulation and quality of a synthetic close-to-reality employer–employee population. Journal of Applied Statistics, 41(5), 1053–1072. https://doi.org/10.1080/02664763.2013.859237

Templ, M., Meindl, B., Kowarik, A., Dupriez, O. (2017). Simulation of synthetic complex data: The r package simpop. Journal of Statistical Software, 79(10). https://doi.org/10.18637/jss.v079.i10

Templ, M. (2017) *Statistical disclosure control for microdata*. New York, NY: Springer Berlin Heidelberg.

Wang, Y., Pan, Z., Zheng, J., Qian, L., & Li, M. (2019). *A hybrid ensemble method for pulsar candidate classification*. Astrophysics and Space Science, 364(8). https://doi.org/10.1007/s10509-019-3602-4

Contact

Jiří Novák

Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Economic Statistics

nám. W. Churchilla 1938/4

130 67 Praha 3

Czech Republic

jiri.novak.kest@vse.cz

Czech Statistical Office, Demography and social statistics section, Population Statistics Department,

Na padesátém 3268/81

100 82 Praha 10

Czech Republic

jiri.novak@czso.cz