Tento projekt je financován se státní podporou Technologické agentury ČR v rámci programu BETA2
www.tacr.cz
Výzkum užitečný pro společnost

Metodika a softwarová podpora odhadů variability indikátorů sociální statistiky

Konečný uživatel výsledků: **Český statistický úřad**
Na padesátém 3268/81
100 82 Praha 10

Název projektu: Metodika a softwarová podpora odhadů variability indikátorů sociální statistiky

Číslo projektu: TITACSU025

Řešitel projektu: Technická univerzita v Liberci

Doba řešení: 1. 6. 2021 – 31. 5. 2022

Důvěrnost a dostupnost: veřejně přístupný

Informace o autorském týmu:

doc. Ing. Jan Šembera, Ph.D. – hlavní řešitel

Ing. Vratislav Žabka, Ph.D.

Ing. Miloslav Nechyba

Únor 2022

Tento projekt je financován se státní podporou Technologické agentury ČR v rámci programu BETA2

www.tacr.cz
Výzkum užitečný pro společnost

ČESKÝ STATISTICKÝ ÚŘAD

CERTIFIKOVANÁ METODIKA

Metodika a softwarová podpora odhadů variability indikátorů sociální statistiky (TITACSU025)

Únor 2022

Technická univerzita v Liberci

SEZNAM POJMŮ A ZKRATEK

ČR	Česká republika
ČSÚ	Český statistický úřad
ECB	Evropská centrální banka
EU	Evropská Unie
Eurostat	Statistický úřad Evropské unie
Nmet	Certifikovaná metodika
Open source	Počítačový software s otevřeným zdrojovým kódem
R	Programovací jazyk a prostředí určené pro statistickou analýzu dat
Software	Programové vybavení

OSNOVA

OBSAH

1 ÚVOD

Certifikovaná metodika byla zpracována v rámci realizace výzkumného projektu „Metodika a softwarová podpora odhadů variability indikátorů sociální statistiky" (TITACSU025) financovaného z programu aplikovaného výzkumu a inovací pro potřeby státní správy BETA2. Realizace výzkumné aktivity byla prováděna za účelem kvalitnějšího a efektivního výkonu státní správy, konkrétně s požadavkem na výzkum a vývoj v oblasti implementace nových poznatků, metodik a nástrojů pro tvorbu makroekonomických, produkčních, environmentálních, demografických a sociálních statistik. Cílem projektu bylo podpořit aplikaci statistických metod pro odhady variability a intervalů spolehlivosti pro indikátory sociální statistiky založené na výběrových šetřeních u domácností.

V rámci realizace projektu byly vytvořeny tři aplikované výstupy. První z nich byla v kategorii Ostatní „Úvodní studie k implementaci odhadů variability statistických odhadů se zaměřením na podmínky jejich implementace pro výběrová šetření". Druhým výstupem z kategorie Nmet je Certifikovaná metodika k implementaci odhadů variability statistických odhadů se zaměřením na podmínky jejich implementace pro výběrová šetření. Posledním výstupem, z kategorie software, je „Aplikace pro výpočet odhadů variability statistických odhadů se zaměřením na podmínky jejich implementace
pro výběrová šetření".

2 CÍL METODIKY

Cílem metodiky je vývoj a implementace nových postupů a metod, které podpoří aplikaci statistických metod pro odhady variability a intervalů spolehlivosti pro indikátory sociální statistiky založené na výběrových šetřeních u domácností. Metodika obsahuje doporučené postupy a popis jejich aplikace v praxi vč. jejich využití v rámci software aplikace.

2.1 ÚČEL METODIKY

Hlavním účelem navržené a certifikované metodiky je implementace odhadů variability statistických odhadů se zaměřením na podmínky jejich implementace pro výběrová šetření. Data z výběrových šetření u domácností tvoří důležitý datový zdroj pro

plánování a vyhodnocování politik a související národních strategií a akčních plánů v oblastech jakými jsou sociální začleňování, politika stárnutí a důchodová reforma, trh práce a podpora zaměstnanosti, celoživotní vzdělávání nebo integrace osob se zdravotním postižením. Certifikovaná metodika je jedním z hlavních výstupů výzkumné aktivity za účelem kvalitnějšího a efektivního výkonu státní správy pomocí implementace nových poznatků, metodik a nástrojů pro tvorbu makroekonomických, produkčních, environmentálních, demografických a sociálních statistik.

Metodika obsahuje návrh postupů (alternativních postupů) pro výpočty odhadů variability (tzv. "standard errors" - standardní chyba odhadu; "confidence intervals" - konstrukce intervalů spolehlivosti) pro měření statistické spolehlivosti indikátorů sociální statistiky z výběrových šetření Českého statistického úřadu.

Vzhledem k tomu, že nově navržené postupy a metody budou v praxi realizovány prostřednictvím software aplikace v jazyce R, je součástí tohoto dokumentu i popis této aplikace a s ní související postupy. Aplikace pro výpočet odhadů variability umožňuje pokrýt širokou paletu statistických indikátorů a parametrizaci vstupů pokud jde
o velikost a způsob provedení výběrů pro dané šetření - minimálně v rozsahu zobecněného modelu dvoustupňového stratifikovaného výběru koncových jednotek výběru (bytů nebo osob) - prostřednictvím uživatelského rozhraní.

2.2 IDENTIFIKACE PROBLÉMŮ

Odhady variability a intervalů spolehlivosti indikátorů a statistických výstupů vzniklých na základě výběrových šetření u domácností jsou se stále vyšší intenzitou poptávaným výstupem a patří ke standardním součástem dokumentace k těmto výstupům a zpráv o kvalitě k jednotlivým šetřením. Český statistický úřad v současné době používá aproximační algoritmy naprogramované pro omezený soubor základních indikátorů. Revize příslušné metodiky se zohledněním současného stavu vědeckého poznání v této oblasti, návrh příslušné metodiky pro postupy se zohledněním konkrétní situace těchto šetření a příprava obecnějšího softwarového modulu pro aplikaci této metodiky na širší množinu indikátorů a výstupů výrazně rozšíří možnosti a nabídku informací uživatelům statistických dat.

Realizace projektu a návrh metodiky byl vyvíjen na základě zadání a zájmu ČSÚ, a týkal se následujících výběrových šetření u domácností:

SILC - Životní podmínky

Šetření je prováděno ve 4 ročních vlnách. Dvoustupňovým výběrem je vybráno 4750 bytů, z nichž je v první vlně šetření získáno přibližně 2400 odpovědí, v dalších třech vlnách je z tohoto výběru získáno dalších přibližně 6200 odpovědí. Šetření probíhá primárně na výběru domácností, jsou však sbírána také data o osobách. Pro zajištění reprezentativnosti výběru jsou odpovědi váženy integrovanými vahami (stejné váhy pro osoby i pro domácnosti), které jsou kalibrované dle pohlaví, věkové skupiny, počtu obyvatel krajů, počtů osob dle velikostní skupiny obce a počtů osob dle ekonomické aktivity v jednotlivých krajích – vše za populaci žijící v bytech. Výstupy z šetření jsou publikovány jak za domácnosti, tak za osoby. Součástí výstupů jsou nejen úhrny a z nich odvozené podílové ukazatele (průměr, struktura), ale též složité indikátory, jejichž definice vychází ze 3 i více ukazatelů (náhodných veličin). Pro odhad variability některých ukazatelů šetření je v ČSÚ používán skript využívající Package survey v programu R. Pro většinu ukazatelů není variabilita odhadována.

Statistika rodinných účtů

Toto šetření je integrováno ve 3. a 4. vlně šetření SILC, je prováděno ve dvou vlnách na podvýběru domácností z šetření SILC. V první vlně je vybráno přibližně 2200 domácností, z nichž je získáno asi 900 odpovědí, ve druhé vlně je osloveno přibližně 1000 domácností a získáno asi 900 odpovědí. Reprezentativnost výběru je zajištěna dopočtem na úhrny SILC s vahami získanými poststratifikací. Výstupy jsou publikovány za domácnosti. Výstupy z šetření jsou vesměs průměry, struktury a meziroční indexy, které se zpracovávají jako klouzavé odhady z dat dvou po sobě následujících let. Variabilita ukazatelů není odhadována.

FSD - Finanční situace domácností

Toto šetření je integrováno do 4. vlny šetření SILC, je prováděno najednou na podvýběru domácností z šetření SILC. Pro šetření je vybráno přibližně 2200 domácností, z nichž je získáno asi 1500 odpovědí. Dopočty a váhy budou teprve řešeny ve spolupráci s ČNB, zřejmě však budou rovněž založeny na úhrnech ze SILC. Výstupy jsou publikovány za domácnosti. Součástí výstupů jsou nejen úhrny a z nich odvozené podílové ukazatele (průměr, struktura), ale též složité indikátory, jejichž definice vychází ze tří i více ukazatelů (náhodných veličin). Předpokládá se zpracování

ve formě klouzavých odhadů z dat dvou po sobě následujících let. Variabilita ukazatelů není odhadována.

VŠPS - Výběrové šetření pracovní sil

Šetření je prováděno v rotačním panelu v 5 čtvrtletních vlnách. Dvoustupňovým výběrem odlišným od výběru pro SILC je vybráno v každé vlně 6780 bytů, rozsah šetření tak zahrnuje bezmála 34000 bytů (až 65000 osob). Výstupy jsou publikovány za osoby. Reprezentativnost výběru je zajišťována vážením, které je primárně konstruováno pro osoby. Dopočty na celou populaci jsou provedeny jednoduchou poststratifikací dle pohlaví, věkové skupiny a okresu. Váhy za domácnosti (používají se výjimečně) se tvoří jako průměr vah osob. Výstupní ukazatele jsou vesměs úhrny a z nich odvozené podílové ukazatele složené ze dvou náhodných veličin. Odhad variability tří základních ukazatelů je vyjadřován formou intervalů spolehlivosti, pro ostatní ukazatele není variabilita odhadována.

VŠCR - Výběrové šetření cestovního ruchu

Šetření probíhá celoročně. Výběr (byty) je identický s výběrem pro VŠPS ze 4. čtvrtletí předchozího roku. Zahrnuje 12000-13000 osob. Reprezentativnost výběru je zajištěna vahami dopočítanými za osoby na celkovou populaci v kategoriích počet osob v jednotlivých krajích, počet mužů a žen v ČR, celkové počty osob dle věkových skupin. Výstupy jsou publikované jednak za osoby a jednak za cesty. Výstupní ukazatele jsou vesměs úhrny a z nich odvozené podílové ukazatele (průměr, struktura) složené ze dvou náhodných veličin. Ve zprávě o kvalitě jsou požadovány variační koeficienty počtu různých druhů cest na osobu za rok. Zatím nejsou odhadovány.

VŠIT - Výběrové šetření informačních technologií

Šetření probíhá ve 2. čtvrtletí roku. Výběr odpovídá výběru páté vlny VŠPS a zahrnuje 4000-4600 domácností. Pro zajištění reprezentativnosti výběru se používají integrované váhy, které stojí na dopočtech výběrových dat na celkovou populaci v kategoriích počet osob v jednotlivých krajích dále tříděné podle ekonomické aktivity, počet mužů a žen v ČR, celkové počty osob dle věkových skupin, počet bytů v ČR.

Výstupy jsou publikovány jak za osoby, tak za domácnosti. Výstupními ukazateli jsou vesměs úhrny a z nich odvozené podílové ukazatele složené ze dvou náhodných veličin. Variabilita ukazatelů není odhadována.

EHIS - European Health Interwiev Survey

Šetření probíhá jednou za pět let. Výběrem bytů jsou některé z vln VŠPS. Výběr jedné až dvou osob v rámci bytu je prováděn losem. Výběr zahrnuje 6700-8000 osob. Reprezentativnost výběru je zajišťována vážením za osoby, které vychází z počtu osob 15+ v krajích a počtů osob podle pohlaví a věkových skupin. Data jsou doplňována proxy šetřením mimo provedený výběr osob, pro které jsou odvozovány speciální váhy. Výstupy jsou publikovány za osoby. Výstupními indikátory jsou vesměs úhrny a z nich odvozené podílové ukazatele složené ze dvou náhodných veličin. Variabilita ukazatelů není odhadována.

Ostatní jednorázová šetření (ENERGO, AES, VŠPO, ...)

Opakují se v periodě cca pět let. Výběr je vždy odvozen od výběru VŠPS. Rozsah se liší podle typu šetření, vesměs nad 10 tis. jednotek (bytů/osob). Výstupy jsou publikovány pro osoby nebo byty podle typu šetření. Variabilita ukazatelů není odhadována.

Metoda výběrových šetření je široce rozšířená a velmi často používaná sociálněvědní metoda, která umožňuje získávat a analyzovat objektivní i subjektivní proměnné o chování a postojích lidí ve zkoumané populaci. Proměnné, v kontextu řešeného projektu nazývané jako indikátory sociální statistiky, přinášejí do jisté míry přesné a objektivní údaje o zkoumané populaci, avšak také obsahují chyby.

Účelem této metodiky je hodnocení variability indikátorů s ohledem na tyto chyby, zejména na výběrové neboli statistické chyby, které jsou zohledněním faktu, že data pocházejí z výběru a nikoliv z celé populace. Metodika předpokládá, že zpracovávaná data obsahují kromě odpovědí respondentů také váhy, které korigují reprezentativitu výběrového souboru a náhodnou distribuci jednotek v souboru vyplývající z použitého výběrového schématu a zejména z neúplné návratnosti dotazníků v rámci šetření.

Metodika se vůbec nezabývá hodnocením vlivu případných chyb měření způsobených náhodně nebo záměrně chybnými odpověďmi respondentů.

2.3 Přehled hlavních cílů metodiky

Hlavními cíli metodiky jsou návrhy postupů (alternativních postupů) pro výpočty odhadů variability (tzv. "standard errors" - standardní chyba odhadu; "confidence intervals" - konstrukce intervalů spolehlivosti) pro měření statistické spolehlivosti indikátorů sociální statistiky z výběrových šetření Českého statistického úřadu, a jejich implementace do praxe. Certifikovaná metodika popisuje vhodné postupy výpočtu odhadů variability pro měření statistické spolehlivosti různých typů indikátorů sociální statistiky z výběrových šetření Českého statistického úřadu. Metodika rozlišuje několik kategorií indikátorů
a popisuje vhodné postupy pro každou kategorii zvlášť. Vybrané indikátory, které ČSÚ publikuje, zařadí do těchto kategorií. Kromě metodiky existuje i softwarový nástroj pro výpočet variability vybranými metodami. Nástroj je naprogramován v jazyku R, běží na běžném PC (s případnými minimálními požadavky na výkon procesoru, vnitřní paměť ad.) a obsahuje uživatelské rozhraní umožňující vstup datových souborů v běžném formátu a výstup výsledků formou datových souborů a reportů. Soubor metod implementovaných v softwarovém nástroji pokrývá dostatečný rozsah indikátorů publikovaných ČSÚ a umožňuje zpracování dalších nových indikátorů v případě dodržení dobře specifikovaných podmínek pro předchozí zpracování dat.

3 CHARAKTERISTIKA NOVOSTI POSTUPŮ

Z pohledu novosti se jedná o jiné, dosud nezpracované poznatky a metody, které jsou výsledkem řešení projektu aplikovaného výzkumu, a budou využity pro potřeby zadavatele v jeho činnosti. Výsledky vznikly v rámci realizace projektu TITACSU025 financovaného z programu aplikovaného výzkumu a inovací pro potřeby státní správy BETA2 v souladu se zákonem č. 130/2002 Sb., Zákon o podpoře výzkumu a vývoje z veřejných prostředků a o změně některých souvisejících zákonů (zákon o podpoře výzkumu a vývoje).

Zadavatel ani zpracovatelé metodiky si nejsou vědomi ke dni její certifikace existence obdobné metodiky, proto je možné ji považovat z tohoto pohledu za novou, inovativní. Obdobná problematika odhadů variability indikátorů sociální

statistiky v takto zamýšleném rozsahu a s takto pojatými výstupy v podobě inovace poskytované veřejné služby nebyla dříve v českém prostředí zkoumána. Projekt pokrývá tedy oblast, která je v prostředí České republiky zcela nová. Český statistický úřad v současné době používá aproximační algoritmy naprogramované pouze pro omezený soubor základních indikátorů. Tento stav neodpovídá odborným požadavkům ze strany uživatelů statistických dat.

Metodický postup představuje komplexní aplikovatelný postup pro řešení aktuálních požadavků zadavatele, který je schopen opakovaně a s vyšší přesností dosáhnout požadovaných odhadů variability indikátorů sociální statistiky.

Tato metodika, byla vytvořena na základě synergie různých oblastí výzkumu a vývoje, pomocí multidisciplinárního přístupu. Pro efektivní hodnocení nebyly pouze aplikovány vzorce ze statistiky či matematiky, ale využity i poznatky sociologů a informatiků.

Pro aplikační přínos je klíčové propojení této metodiky se softwarovou aplikací, která umožní zpracovávat potřebné množství dat a rychle reagovat na případnou změnu struktury vstupů. Proto je návrh software zpracován v takové formě
a struktuře, aby do něj bylo možné v budoucnu kdykoliv dle potřeby promítnout potřebné požadavky na změny indikátorů během hodnocení.

Novost postupu metodiky spočívá v univerzalitě řešení pro hodnocení variability indikátorů ve výběrových šetřeních ČSÚ pro všechny zpracovávané indikátory s určitou mírou rozšiřitelnosti na případné další indikátory. Metodika navrhuje využívat ověřené metody odpovídajícím způsobem tak, aby podpořila srovnatelnost výsledků s výsledky jiných agentur.

Implementace certifikované metodiky přinese zejména:

- Návrh nových postupů (alternativních postupů) pro výpočty odhadů variability zapracovaných v rámci jednoho uceleného certifikovaného metodického dokumentu
- Vyšší míru efektivity metod a počítačového zpracování dané dílčí problematiky

- Nově vyvinutý software na open source platformě se zapracovanými inovovanými metodami pro dlouhodobé využití zaměstnanci

4 POPIS UPLATNĚNÍ CERTIFIKOVANÉ METODIKY

Realizátorem projektu a beneficientem výstupu Certifikovaná metodika je Český statistický úřad. Koneční uživatelé výsledků projektu budou zaměstnanci Českého statistického úřadu. Aplikaci výsledků do praxe, konkrétně využití nových postupů zpracovaných v rámci certifikované metodiky a využívání softwarové aplikace, bude realizována prostřednictvím zaměstnanců ze „6 Sekce demografie a sociálních statistik", v rámci odborů „61 Odbor statistiky obyvatelstva", „62 Odbor šetření v domácnostech", „63 Odbor statistik rozvoje společnosti" a „64 Odbor statistiky trhu práce a rovných příležitostí".

K úspěšné realizaci výzkumného záměru a jeho implementaci do praxe bylo třeba u technických výstupů dodržet určité podmínky a sledovat kontinuálně jejich naplnění. Konkrétně se jedná o softwarovou aplikaci pro výpočet odhadů variability statistických odhadů se zaměřením na podmínky jejich implementace pro výběrová šetření. Vzhledem k požadavkům na pokrytí široké palety statistických indikátorů a parametrizací vstupů (velikost a způsob provedení výběrů pro dané šetření) bude nutné definovat požadavky na prostředí
a hardware, ve kterém budou spouštěny výše uvedené výpočty, a zajistit z technického, personální a finančního pohledu provoz a případný budoucí rozvoj aplikace.

Finanční podmínky

Zajištění finančních prostředků pro financování provozních nákladů souvisejících s provozem aplikace (upgrade a update prostředí), případně prostředky na její rozvoj (nové funkcionality, ukládání a zobrazení výsledků). Uvedené aktivity je možné realizovat i rámci stávajících aktivit rozpočtu organizace.

Technické podmínky

Zajištění technické infrastruktury – zařízení pro provoz softwarové aplikace. Konkrétní požadavky na hardware budou obecně popsány v části Technické a technologické vybavení této studie, detailnější popis bude uveden v rámci technické dokumentace.

Požadované funkcionality

Vkládání vstupů (dat) do aplikace, zajištění funkčního hodnocení a intepretace výsledků s průběžnou aktualizací potřeb a požadavků na funkcionality, případně úpravy související s legislativními změnami.

Personální požadavky

Zaškolení personálu v oblasti práce s prostředím jazyka R a s ním souvisejících prostředí využívaných pro zpracování dat a interpretaci výsledků.

Časové požadavky

Implementace výstupu software bude zahájena v prostředí ČSÚ ihned po předání finální verze aplikace, která bude minimálně po dobu 30 dnů ověřena v testovacím provozu. Aplikace musí být připravena z technického hlediska pro ostrý provoz na zařízení uživatelů nejpozději k 1. 6. 2022.

Dokumentace software

Podmínkou uvedení metodiky do praxe a aplikace do ostrého provozu bude zpracování technické a uživatelské dokumentace. Technická dokumentace bude zpracována pro účely správy software (práva, role, zabezpečení), podmínek a doporučení upgrade a update, případně požadavků na instalace konkrétních opravných či rozšiřujících „balíčků". Uživatelská dokumentace bude sloužit pro seznámení se s prostředím a funkcionalitami aplikaci.

5 ZPŮSOB A HARMONOGRAM IMPLEMENTACE

Způsob a harmonogram implementace přímo navazuje na realizační část projektu a souvisí se dvěma výstupy projektu, certifikovanou metodikou a software aplikací.

Realizační část

V rámci realizační části byly během 9 měsíců realizovány aktivity spojené s návrhem, zpracováním a certifikací nově vyvinuté metodiky. Realizační část byla zahájena 1. 9. 2021 a ukončena 31. 5. 2022.

Období Září 2021 - Listopad 2021

Návrh metodického postupu pro výpočty odhadů variability z šetření u domácností prováděných Českým statistickým úřadem.

Období Prosinec 2021 - Únor 2022

Dokončený návrh metodického postupu pro výpočty odhadů variability z šetření u domácností prováděných Českým statistickým úřadem a jeho ověření.

Období Březen 2022 - Květen 2022

Finální verze metodiky se zapracovanými připomínkami a provedenou certifikací.

Implementační část

Implementační část výsledků „Nmet - Certifikovaná metodika" a „R – Sofware" do praxe z pohledu jejich reálného využití bude zahájena ihned po ukončení výzkumného projektu, tedy 1. 6. 2022.

Skutečný termín dokončení bude záviset na zavedení výsledků v rámci organizace a interních procesů.

Implementaci lze považovat za realizovanou a ukončenou v momentě, kdy dojde k reálnému využívání postupů z certifikované metodiky při práci zaměstnanců ČSÚ, a k reálnému využívání software aplikace při práci zaměstnanců ČSÚ. Předpokládaným termínem zahájení reálného využití je podzim roku 2022.

6 METODICKÝ POSTUP

Metodický postup vzniká spolu se softwarovou aplikací, jejíž základní parametry jsou popsány v kapitole 7 a je jednak podkladem k vývoji aplikace, jednak metodickým návodem, jak provádět vyhodnocení variability indikátorů s jejím použitím. Podrobnější návod práce s aplikací bude součástí dokumentace softwarové aplikace po jejím dokončení.

6.1 Kategorizace indikátorů

Pro účely odhadu variability indikátorů rozdělíme indikátory do tří skupin:

1) Indikátory jednoduché
2) Indikátory jednoduše odvozené
3) Specifické indikátory

Indikátory jednoduchými rozumíme indikátory, jejichž hodnota je přímo sbírána v dotazníku šetření (např. Nejvyšší dokončené vzdělání nebo Hrubé příjmy z podnikání).

Indikátory jednoduše odvozenými rozumíme indikátory, jejichž hodnota je přímo odvozena z hodnot sbíraných v dotazníku šetření (např. Celkové hrubé příjmy, Míra materiální deprivace ad.).

Indikátory specifickými rozumíme indikátory, k jejichž vyhodnocení je potřeba statistika jiných indikátorů. Zejména jde o tzv. laekenské indikátory, mezi něž patří např. Hranice příjmové chudoby, Relativní propad příjmů, Koeficient příjmové nerovnosti S80/S20 a Giniho koeficient.

6.2 Popis metod

Odhad variability indikátorů bude prováděn ve všech případech přímo z dat bez zkoumání předpokladu binomického rozdělení veličiny. Pro jednoduché indikátory bude použit výpočet podle vzorce (1) s využitím centrální limitní věty, která umožňuje stanovit α% interval spolehlivosti pro odhad h ukazatele H pomocí obecného vztahu:

$$h \mp u_{1-\alpha/2} . s_h, \quad (1)$$

h je odhad ukazatele H a

s_h je směrodatná odchylka odhadu h

$u_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení.

Tento vzorec je převzat z metodických vysvětlivek ČSÚ [1].

Do výsledku je ještě nutno promítnout vliv tzv. design efektu, což je vliv skutečnosti, že výběrovým schématem nebyl prostý náhodný výběr, jak standardní vzorce předpokládají, ale výběr stratifikovaný na úrovni kraje a 4 velikostních skupin obcí, a dále dvoustupňový (nejprve náhodný výběr sčítacích obvodů a následně bytů v každém z nich). Ten se obecně vyjadřuje podle vzorce:

$$\text{deff}(h) = s_h^2 / s_h^2\{\text{pnv}\}, \quad (2)$$

s_h^2 je rozptyl proměnné h při skutečném výběrovém schématu

$s_h^2\{\text{pnv}\}$ je rozptyl proměnné h při prostém náhodném výběru.

Tento vliv zohledňují implementované metody, pro jeho správné zohlednění musí být odpovídajícím způsobem zvoleny parametry pro funkci svydesign (viz kapitola 6.4).

Variabilita indikátorů jednoduše odvozených jako lineární kombinace náhodných veličin nebo jako kategorická nebo hodnotová podmínka bude odhadována tak, že pro každý vzorek ve výběru bude odvozen indikátor jako jedna realizace náhodné veličiny, a takto získané hodnoty budou zpracovány stejným způsobem jako jednoduchý indikátor. Tento postup bude odpovídat

tomu, jako by datový soubor byl rozšířen o nový sloupec obsahující hodnotu jednoduše odvozené náhodné veličiny.

Je-li jednoduše odvozený indikátor podílem dvou náhodných veličin (jednoduchých indikátorů nebo jejich lineárních kombinací, případně jednoduše odvozených indikátorů ve formě kategorické nebo hodnotové podmínky), je třeba provést odhad variability podle vzorce (3) odvozeného pro podíl dvou náhodných veličin y a x za předpokladu prostého náhodného výběru bez vracení při vážení výběrových dat vahami:

$$\frac{y_w}{x_w} \pm \frac{u_{1-\alpha/2}}{x_w} \sqrt{\left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{n}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n}\left[w_i\left(y_i-\frac{y_w}{x_w}x_i\right)^2\right]} \qquad (3)$$

$u_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení,

n je počet prvků ve výběru,

x_w resp. y_w jsou vážené výběrové úhrny $x_w = \frac{n}{\sum w_i}\sum_{i=1}^{n} w_i x_i$ resp. $y_w = \frac{n}{\sum w_i}\sum_{i=1}^{n} w_i y_i$.

Tento vzorec je převzat z metodických vysvětlivek ČSÚ [1].

Stejně jako v případě jednoduchých indikátorů musí být v každém případě do výsledku promítnut vliv tzv. design efektu. Tento vliv zohledňují implementované metody, pro jeho správné zohlednění musí být odpovídajícím způsobem zvoleny parametry pro funkci svydesign (viz kapitola 6.4).

Pro odhad variability specifických indikátorů nejsou předchozí popsané postupy dostatečné a je třeba použít metody založené na postupu replikace výběru nebo její formě využívající linearizaci na základě Taylorova rozvoje, jejichž odhady jsou podle publikovaných výsledků ve [2] a [3] vzájemně srovnatelné. Přitom výpočty metodou založenou na linearizaci jsou výrazně méně časově náročné
a nevyžadují složitou přípravu dat. Omezení použitelnosti všech uvedených metod spočívá v potřebě dostatečného počtu dat v každém hodnoceném stratu. Zatímco v případě metod založených na replikaci výběru (bootstrap a jacknife)

nedostatek vzorků vede ke zhoršení přesnosti a zhoršení konvergence metod, v případě linearizace nedostatek vzorků snižuje přesnost. Proto byla pro implementaci v softwarové aplikaci vybrána linearizační metoda [3], [4], [5].

Pro správný odhad variability specifických indikátorů je klíčové zohlednění výběrového schématu. I zde se jeho parametry předávají výpočetním funkcím volbou parametrů pro funkci svydesign (viz kapitola 6.4).

Hranice příjmové chudoby a Míra chudoby

Hranice příjmové chudoby je hranice příjmu, jejíž podkročení definuje obyvatele, jejichž životní úroveň je nízká vzhledem obecné životní úrovni. Míra chudoby určuje podíl osob s příjmem nižším než je hranice příjmové chudoby. Použitá metoda odhadu těchto indikátorů a jejich variability založená na linearizaci s použitím Horvitzova-Thompsonova estimátoru je odvozena a diskutována v [5].

Relativní propad příjmů

Určuje relativní rozdíl mezi mediánem příjmu osob, které mají příjem nižší než je hranice příjmové chudoby a samotnou hranicí příjmové chudoby. Použitá metoda odhadu relativního propadu příjmů a jeho variability založená na linearizaci s použitím Horvitzova-Thompsonova estimátoru je odvozena a diskutována v [5].

Koeficient příjmové nerovnosti S80/S20

Jde o míru nerovnosti distribuce příjmů. Její hodnotou je podíl úhrnného příjmu nejbohatší pětiny obyvatel a úhrnného příjmu nejchudší pětiny obyvatel. Použitá metoda odhadu tohoto indikátoru a jeho variability založená na linearizaci s použitím Horvitzova-Thompsonova estimátoru je odvozena a diskutována v [5].

Giniho koeficient

Giniho index je pokusem o vyjádření nerovnosti prezentované Lorenzovou křivkou v jednom čísle. Je to dvojnásobek plochy uzavřené mezi křivkou rovnosti a skutečnou Lorenzovou křivkou. K jeho výpočtu je použita linearizovaná forma vzorce navrženého Osierem ve článku [3].

6.3 Použitelnost metod

Použitelnost a přesnost použitých metod je omezena zejména dvěma faktory – počtem zpracovávaných statistických vzorků a správným popisem výběrového schématu. Technické minimum pro výpočet variability indikátoru ze statistického výběru je existence alespoň dvou hodnot v každém vyhodnocovaném stratu. Pro zajištění rozumné vypovídací hodnoty je třeba, aby byly výběry větší. Nejmenší efektivní velikost vzorku (velikost výběru opravenou o vliv výběrového schématu), se kterou úspěšně pracovaly autorky studie [2], byla 11,13.

Promítnutí vlivu výběrového schématu musí být při hodnocení variability všech indikátorů zahrnuto správnou volbou parametrů pro funkci svydesign (viz kapitola 6.4). Nesprávná volba parametrů způsobí v každém případě chybu v určení odhadu variability.

Softwarová aplikace vyvíjená pro implementaci této metodiky bude v případě nevyhovujících podmínek výpočtu ohlašovat srozumitelným způsobem důvody, proč výpočet nemohl být proveden, nebo případná omezení věrohodnosti výsledku, pokud výpočet proběhl.

6.4 Tvorba software

Výpočty středních hodnot a intervalů spolehlivosti jednoduchých a jednoduše odvozených indikátorů jsou v rámci aplikace realizovány následujícími funkcemi z knihovny survey:

- svytotal
- svymean
- confint
- svyciprop
- svyratio

Výpočty středních hodnot a intervalů spolehlivosti jednotlivých specifických indikátorů jsou v rámci aplikace realizovány následujícími funkcemi z knihoven convey a survey:

- Hranice příjmové chudoby – funkce svyarpt, confint
- Míra chudoby – funkce svyarpr, confint

- Relativní propad příjmů – funkce svyrmpg, confint
- Koeficient příjmové nerovnosti S80/S20 – funkce svyqsr, confint
- Giniho koeficient – funkce svygini, confint

Popisy algoritmů a použití uvedených metod jsou zdokumentovány v [6].

Následují vysvětlivky pro jednotlivé parametry nastavení procedury svydesign tak, aby byl správným způsobem zohledněn design effect.

Parametr weights

Parametr weights odkazuje na sloupce s daty obsahujícími pravděpodobnostní váhy, tj. inverze pravděpodobnosti, že položka bude zahrnuta do výběru na základě použitého výběrového schématu.

Parametr strata

Parametr strata odkazuje na sloupce se stratifikací položek, tj. začleněním do zvolených strat (skupin podle demografických a dalších parametrů). Na základě jeho volby je výběr rozdělen na jednotlivá strata, která jsou vyhodnocována nezávisle na sobě. Ve většině případů je třeba, aby součástí každého strata byly alespoň dvě PSU. Stratifikace obvykle snižuje odhad hodnoty variability, protože v rámci strata je obvykle rozptyl hodnot náhodných veličin menší než v celém náhodném výběru.

Parametr id

Parametrem id se nastavují datové sloupce obsahující zařazení vzorku do primárních vzorkovacích jednotek, případně do vzorkovacích jednotek vyšších stupňů. Primární vzorkovací jednotka (Primary Sampling Unit, PSU) je největší jednotkou, která je vzorkovaná v rámci výběrového schématu, tedy jednotka v první úrovni členění. Správné určení PSU v rámci hodnocení dat zpřesňuje (a také typicky zvyšuje) odhad variability indikátorů, protože prostý náhodný výběr vede typicky k nižší variabilitě indikátorů. Tento parametr je tedy klíčový ke správnému započítání tzv. design effectu do odhadu variability.,

Parametr FPC

Faktor opravy pro konečnou populaci (Finite Population Correction Factor, FPC) se užívá v případě, kdy je velikost výběru relativně velká vzhledem k velikosti populace. V případě, že výběr je malý vzhledem k populaci, je jeho velikost blízká jedné a lze jej bezpečně opominout. V případě výběrů vzhledem

k populaci, nebo v případě zpracování dat velkých podvýběrů vzhledem k hodnocenému výběru, je nutné FPC zahrnout do výpočtu.

Příklady nastavení parametrů výběrového schématu

Uživatel má možnost nastavit přímo parametry id, strata, weights a fpc vybráním jednoho či více sloupců z některé z tabulek nahraných do softwaru. Druhou možností je kompletní definice výběrového schématu pomocí textového vstupu, kde může uživatel nastavit i všechny další varianty funkce svydesign uvedené
v dokumentaci. V obou případech aplikace hlídá, jestli je výběrové schéma nastaveno úplně. Pokud ano, vypíše vybranou variantu výběrového schématu i s její definovanou formulí. V opačném případě vypíše chybové hlášení s návodem, jak ho má uživatel opravit.

Následují ukázky tří variant zadání výběrového schématu:

Stratifikovaný výběr

```
## Stratified Independent Sampling design
## svydesign(id = ~1, strata = ~stype, weights = ~pw, data = apistrat,
##     fpc = ~fpc)
```

Jednoúrovňové výběrové schéma

```
## 1 - level Cluster Sampling design
## With (15) clusters.
## svydesign(id = ~dnum, weights = ~pw, data = apiclus1, fpc = ~fpc)
```

Dvouúrovňové výběrové schéma s vahami vypočtenými z velikosti populací

```
## 2 - level Cluster Sampling design
## With (40, 126) clusters.
## svydesign(id = ~dnum + snum, fpc = ~fpc1 + fpc2, data = apiclus2)
```

7 POPIS SOFTWARE APLIKACE

Spolu s touto metodikou vznikla také aplikace VariabiliatoR, která umožňuje přímé praktické využití popsané metodiky v podmínkách činnosti ČSÚ. Tato kapitola obsahuje stručný popis hlavních parametrů aplikace. Podrobnější popis aplikace a způsobu jejího použití je obsahem dokumentace k softwaru.

7.1 Návrh použitých technologií

Návrh použitých technologií vychází z požadavku ČSÚ realizovat software na open source platformě R (bez užití komponent vyžadujících licence) pracující na osobních počítačích s operačním systémem Windows 10. Návrh technologií je tak omezen na doporučenou verzi software R a doporučené balíčky.

Minimální a doporučená konfigurace pro práci s aplikací je uvedena v bodě 9. Technické a technologické zajištění.

Přehled použitých technologií:

- R version 4 a vyšší
- library(shiny)
- library(readxl)
- library(dplyr)
- library(lubridate)
- library(ggplot2)
- library(survey)
- library(convey)
- library(DT)

7.2 Popis architektury a základní vymezení aplikace

Aplikace je rozdělena na dvě hlavní částí UI (user interface) a server (funkcionality ovládacích prvků), což odpovídá dvěma základním textovým souborům ui.R a server.R. Pro spuštění aplikace musí být tyto dva soubory v jedné složce a z R se jejich spuštění volá příkazem runApp(), jehož parametrem je název této složky. Alternativním způsobem spuštění aplikace je metoda volaná pomocí tlačítka Run App přímo v RStudiu. Aplikace se otevře a běží v okně prohlížeče.

Vizuálně bude aplikace tvořena několika záložkami ve formátu navbarPage s ovládáním v levé části pomocí sidebarLayout.

7.3 Správa aplikace

Správa aplikace bude probíhat v rámci správy osobních počítačů jednotlivých pracovníků včetně aktualizací programu R a příslušných balíčků. Vzhledem

k charakteru aplikace nejsou předpokládány významné budoucí úpravy v souvislosti a aktualizací výše uvedených software prostředí.

7.4 Práce s daty a datová rozhraní

Software aplikace pracující na osobních počítačích s operačním systémem Windows 10 bude provádět požadovaná hodnocení na základě dat načtených (importovaných) ze zvolených souborů nebo z dat získaných přímo z databáze Českého statistického úřadu.

Data z načtených souborů jsou myšlena data načtená ve formátu csv případně xlsx (xls). Jedná se o formát dat, který je standardně používán při práci zaměstnanců v rámci organizace.

Získávání dat pro další hodnocení bude realizováno buď předchozím stažením dat z databáze do formátu csv jiným nástrojem, nebo napojením k databázi přes ovladač ODBC coby standardizované softwarové API pro přístup k databázovým systémům. Způsob připojení bude určen v průběhu implementace softwarové aplikace na základě podmínek přístupu k databázím ČSÚ.

8 DOPORUČENÍ V OBLASTI ORGANIZAČNÍ A PERSONÁLNÍ

Koneční uživatelé výsledků projektu budou zaměstnanci Českého statistického úřadu. Aplikace výsledků do praxe, konkrétně využití nových postupů zpracovaných v rámci certifikované metodiky a využívání softwarové aplikace, bude realizována prostřednictvím zaměstnanců ze Sekce demografie a sociálních statistik.

Aplikace certifikované metody ani software aplikace nevyžadují žádné dodatečné lidské kapacity. Není ani potřeba zapracovat do plánu školení stávajících zaměstnanců speciální vzdělávací program týkající se jejich znalostí a dovedností.

9 TECHNICKÉ A TECHNOLOGICKÉ ZAJIŠTĚNÍ

Vzhledem k charakteru instituce a její hlavní činnosti disponuje Český statistický úřad technickým a technologickým zázemím pro zpracování velkého množství dat na vysoké úrovni. Pro potřeby zpracování části dat a problematiky hodnocení variability indikátorů sociální statistiky však využívají zaměstnanci pouze osobní počítače. Stejně tak by tomu mělo být i nadále po realizace této zakázky, kdy nově vyvinutá softwarová aplikace pro výpočet odhadů variability statistických odhadů se zaměřením na podmínky jejich implementace pro výběrová šetření poběží pouze na osobních počítačích, nikoliv v serverovém prostředí.

Aplikace certifikované metody ani software aplikace nevyžadují žádné dodatečné technologické infrastrukturní kapacity.

Doporučená konfigurace pro práci s aplikací je ke dni certifikace metodiky následující:

- OS: Windows 10, 11 (64-bit)
- CPU: Intel Core i7 nebo AMD Ryzen 7
- RAM: 16 GB +
- HDD: SSD
- Grafická karta s DirectX 9 nebo novější s ovladačem WDDM 1.0
- Rozlišení obrazovky 800×600 a vyšší
- R version 4 a vyšší
- Prohlížeč internetu: Edge, Mozilla, Chrome
- Přehled balíčků:
  - library(shiny)
  - library(readxl)
  - library(dplyr)
  - library(lubridate)
  - library(ggplot2)
  - library(survey)
  - library(convey)
  - library(DT)

Minimální konfigurace pro práci s aplikací je ke dni certifikace metodiky následující:

- OS: Windows 10, 11 (64-bit)

- CPU: Intel Core i5 nebo AMD Ryzen 5
- RAM: 4 GB
- HDD: SSD
- Grafická karta s DirectX 9 nebo novější s ovladačem WDDM 1.0
- Rozlišení obrazovky 800×600 a vyšší
- R version 4 a vyšší
- Prohlížeč internetu: Edge, Mozilla, Chrome
- Přehled balíčků:
  - library(shiny)
  - library(readxl)
  - library(dplyr)
  - library(lubridate)
  - library(ggplot2)
  - library(survey)
  - library(convey)
  - library(DT)

10 LEGISLATIVNÍ RÁMEC

Český statistický úřad (ČSÚ) je ústředním orgánem státní správy České republiky. Byl zřízen dne 8. ledna 1969 zákonem č. 2/1969 Sb., o zřízení ministerstev a jiných ústředních orgánů státní správy. V rámci své činnosti vychází z platné legislativy, konkrétně:

- Zákon o státní statistické službě – Zákon č. 89/1995 Sb., o státní statistické službě, ve znění pozdějších předpisů
- Základní právní rámec EU v oblasti statistiky
- Kodex evropské statistiky (Code of Practice)
- Závazek o důvěryhodnosti statistiky v ČR
- Zákon o sčítání lidu, domů a bytů v roce 2021 - Zákon č. 332/2020 Sb., o sčítání lidu, domů a bytů v roce 2021 a o změně zákona č. 89/1995 Sb., o státní statistické službě, ve znění pozdějších předpisů
- Vyhláška č. 490/2020 Sb. - Vyhláška č. 490/2020 Sb., kterou se provádějí některá ustanovení zákona č. 332/2020 Sb., o sčítání lidu, domů a bytů v roce 2021 a o změně zákona č. 89/1995 Sb., o státní statistické službě, ve znění pozdějších předpisů.

- Vyhláška, kterou se mění vyhláška č. 490/2020 Sb. - Vyhláška, kterou se mění vyhláška č. 490/2020 Sb., kterou se provádějí některá ustanovení zákona č. 332/2020 Sb., o sčítání lidu, domů a bytů v roce 2021 a o změně zákona č. 89/1995 Sb., o státní statistické službě, ve znění pozdějších předpisů.

Mezi nejvýznamnější strategické dokumenty patří:

- Strategický plán ČSÚ na období 2022–2026
- Etický kodex zaměstnance Českého statistického úřadu
- Prioritní úkoly ČSÚ na rok 2021
- Politika diseminace Českého statistického úřadu (duben 2013)
- Politika revizí ČSÚ platná k 3. dubnu 2020
- Závazek kvality ČSÚ
- Politika kvality ČSÚ
- Rámcová bezpečnostní politika ČSÚ
- Politika interní komunikace ČSÚ
- Politika řízení změn ČSÚ

Proces certifikace ani vlastní metodika nebudou mít dopad do stávajících legislativních i nelegislativních předpisů, ani nevyžadují zapracování jakýkoliv změn.

11 EKONOMICKÉ ASPEKTY

Ekonomické dopady nově implementované certifikované jsou z pohledu budoucích nákladů neutrální. Nepředpokládá se, že by si nově implementované metody do praxe vyžádaly náklady na straně nových pracovních míst, náklady na školení stávajících míst, či náklady na pořízení technologií potřebných pro realizaci výpočtů a hodnocení. Rovněž software prostředí mající charakter open source, v rámci něhož je vyvinuta aplikace, nevyžaduje žádné investiční či provozní prostředky pro vlastní budoucí provoz. Software aplikace bude možné spustit a využívat na běžném kancelářském počítači.

Postupy z metodiky budou aplikovány do stávající pracovní náplně zaměstnanců za účelem zvýšení kvality hodnocení dat a inovace postupů. Nepředpokládá se zavedení nových služeb, které by generovaly nové příjmy a měly ekonomický dopad pro ČSÚ. Nefinanční přínosy jsou očekávány zejména v kvalitnějším a efektivním výkonu státní správy.

12 ZÁVĚR

Implementace této certifikované metodiky a využití spolu s ní vytvořeného specializovaného softwarového nástroje VariabiliatoR pro provádění odhadů variability implementující postupy výše uvedené metodiky umožní Českému statistickému úřadu inovovat stávající poskytované služby. Implementace nových poznatků, metodik a nástrojů pro tvorbu makroekonomických, produkčních, environmentálních, demografických a sociálních statistik získaných během výzkumného projektu docílí kvalitnějšího a efektivního výkonu státní správy.

LITERATURA

[1] Český statistický úřad, Metodické vysvětlivky. [Online]
https://www.czso.cz/documents/10180/142681148/16002121mc.pdf/1629f5c3-8949-41f3-8d5b-c7f9b705536c?version=1.3

[2] T. Laureti a I. Benedetti. Measuring and communicating uncertainty of poverty indicators at regional level. 2020 edition. EUROSTAT. [Online]
https://ec.europa.eu/eurostat/documents/3888793/12137895/KS-TC-20-010-EN-N.pdf/6745684c-c989-b3e5-33ae-7bd7dd89bf92

[3] G. Osier, Variance estimation for complex indicators of poverty and inequality using linearization techniques, Survey Research Methods (2009) Vol.3, No.3, pp. 167-195, ISSN 1864-3361.
http://ojs.ub.uni-konstanz.de/srm/article/view/369

[4] Y. Berger, G. Osier a T. Goedemé, Standard error estimation and related sampling issues, in A. B. Atkinson, A.-C. Guio a E. Marlier (eds.) Monitoring social inclusion in Europe, Luxembourg: Publication Office of the EU. 2017, str. 465-478.

[5] J.-C. Deville, Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques. Survey Methodology (1999) 25 (2): 193–203.
http://www.statcan.gc.ca/pub/12-001-x/1999002/article/4882-eng.pdf

[6] G. Jacob, A. Damico a D. Pessoa, Poverty and Inequality with Complex Survey Data. [Online]
https://guilhermejacob.github.io/context/index.html

Program veřejných zakázek v aplikovaném výzkumu a inovacích pro potřeby státní správy BETA2 byl schválen usnesením vlády České republiky č. 278 ze dne 30. 3. 2016 a je zaměřen na podporu aplikovaného výzkumu a inovací pro potřeby orgánů státní správy. Poskytovatelem finančních prostředků je Technologická agentura ČR.