

WORKING PAPERS IER SAS

Pracovné listy EÚ SAV



Tomáš Miklošovič

Vplyv vstupných mikroúdajov na výsledky
populačného mikrosimulačného modelovania



Financované
Európskou úniou
NextGenerationEU

PLÁN [OBNOVY]

eúsav
Ekonomický ústav SAV
Institute of Economic Research SAS

AUTHOR

Tomáš Miklošovič, Institute of Economic Research SAS, Bratislava, Slovakia 

Financované EÚ NextGenerationEU prostredníctvom Plánu obnovy a odolnosti SR v rámci projektu č. 09I03-03-V04-00536.

Vplyv vstupných mikroúdajov na výsledky populačného mikrosimulačného modelovania

ABSTRAKT

Kvalita vstupných mikroúdajov predstavuje jeden z rozhodujúcich faktorov ovplyvňujúcich spoľahlivosť populačných mikrosimulačných modelov. Napriek rozsiahlemu rozvoju mikrosimulačného modelovania sa odborná literatúra sústreďuje predovšetkým na vývoj simulačných algoritmov, zatiaľ čo otázka vhodnosti rôznych zdrojov vstupných mikroúdajov zostáva relatívne málo preskúmaná. Cieľom štúdie je analyzovať význam vstupných mikroúdajov v populačnom mikrosimulačnom modelovaní a experimentálne posúdiť vplyv troch oficiálnych štatistických zdrojov – Labour Force Survey (LFS), European Union Statistics on Income and Living Conditions (EU-SILC) a populačného CENZUSu na výsledky identického populačného mikrosimulačného modelu.

Výskum je založený na kontrolovanom experimentálnom dizajne, v ktorom zostávajú všetky komponenty modelu vrátane simulačných algoritmov, parametrov, demografických procesov, časového horizontu a validačných postupov nezmenené. Jedinou experimentálne menenou premennou je zdroj vstupných mikroúdajov, prípadne parametrov vychádzajúcich z týchto údajov. Takto navrhnutý prístup umožňuje izolovať vplyv vstupného dátového základu od ostatných faktorov ovplyvňujúcich výsledky mikrosimulácie.

Výsledky ukazujú, že voľba vstupných mikroúdajov má iba obmedzený vplyv na agregované populačné ukazovatele, ako sú celkový počet obyvateľov alebo počet narodených detí, pričom všetky analyzované databázy vedú k veľmi podobným dlhodobým trendom populačného vývoja. Naopak, pri projekciách ekonomicky aktívneho obyvateľstva, jeho vekovej a vzdelanostnej štruktúry sa rozdiely medzi jednotlivými vstupnými databázami postupne kumulujú a vedú k odlišným výsledkom. Experiment potvrdil, že citlivosť populačného mikrosimulačného modelu na voľbu vstupných mikroúdajov závisí od typu analyzovaného výstupného ukazovateľa.

Hlavným prínosom štúdie je návrh experimentálneho metodologického rámca umožňujúceho systematicky hodnotiť vplyv rôznych zdrojov mikroúdajov pri zachovaní identickej štruktúry mikrosimulačného modelu. Výsledky zdôrazňujú potrebu venovať zvýšenú pozornosť výberu a validácii vstupných mikroúdajov, najmä pri modelovaní štruktúrnych charakteristík populácie, a poskytujú metodické odporúčania pre tvorbu populačných mikrosimulačných modelov využívajúcich oficiálne štatistické údaje.

KEYWORDS: populačné mikrosimulačné modelovanie, mikroúdaje, Labour Force Survey, EU-SILC, CENZUS

JEL CLASSIFICATION: C81, C53, J11

The Impact of Input Microdata on the Results of Population Microsimulation Modeling

ABSTRACT

The quality of input microdata is one of the key factors influencing the reliability of population-based microsimulation models. Despite the extensive development of microsimulation modeling, the professional literature focuses primarily on the development of simulation algorithms, while the issue of the suitability of various sources of input microdata remains relatively under-explored. The aim of this study is to analyze the importance of input microdata in population microsimulation modeling and to experimentally assess the impact of three official statistical sources—the Labor Force Survey (LFS), the European Union Statistics on Income and Living Conditions (EU-SILC), and the population census—on the results of an identical population microsimulation model.

The research is based on a controlled experimental design in which all components of the model—including simulation algorithms, parameters, demographic processes, the time horizon, and validation procedures—remain unchanged. The only experimentally varied variable is the source of the input microdata, or the parameters derived from these data. This approach allows us to isolate the impact of the input data set from other factors influencing the results of the microsimulation.

The results show that the choice of input microdata has only a limited impact on aggregated population indicators, such as the total population or the number of births, with all analyzed databases yielding very similar long-term trends in population development. Conversely, in projections of the economically active population and its age and educational structure, differences between individual input databases gradually accumulate and lead to divergent results. The experiment confirmed that the sensitivity of the population microsimulation model to the choice of input microdata depends on the type of output indicator analyzed.

The main contribution of the study is the proposal of an experimental methodological framework that allows for the systematic evaluation of the impact of different microdata sources while maintaining an identical microsimulation model structure. The results highlight the need to pay closer attention to the selection and validation of input microdata, particularly when modeling the structural characteristics of the population, and provide methodological recommendations for the development of population microsimulation models using official statistical data.

KLÚČOVÉ SLOVÁ: population microsimulation modeling, microdata, Labor Force Survey, EU-SILC, CENSUS

JEL CLASSIFICATION: C81, C53, J11

The WORKING PAPER SERIES is intended to convey preliminary, partial results of ongoing research achieved by fellows or research groups of the Institute of Economic Research which can be prepared for later publications.

The views expressed in the WP and the language revision is those of the authors.

LAYOUT BY: Tomáš Miklošovič

Institute of Economic Research SAS
Ekonomický ústav SAV, v.v.i.
Šancová 56, 811 05 Bratislava
www.ekonom.sav.sk

CONTACT / KONTAKT: ekonedra@savba.sk

© Institute of Economic Research SAS/Ekonomický ústav SAV, v.v.i., Bratislava 2026

Obsah

1. Úvod	5
2. Vývoj mikrosimulačného modelovania a rastúci význam vstupných mikroúdajov.....	6
3. Požiadavky na vstupné mikroúdaje v populačnom mikrosimulačnom modelovaní	7
4. Metodika výskumu.....	11
4.1. Výskumný dizajn.....	11
4.2. Harmonizácia vstupných údajov	12
4.3. Porovnanie vstupných údajov	12
4.4. Mikrosimulačný model SLAMM.....	21
5. Hodnotenie výsledkov	24
6. Diskusia	31
7. Záver	33
Literatúra	36

1. Úvod

V posledných desaťročiach sa mikrosimulačné modelovanie etablovalo ako jeden z najvýznamnejších analytických prístupov využívaných pri hodnotení verejných politík, populačných projekcií a modelovaní sociálno-ekonomických procesov. Na rozdiel od tradičných makroekonomických modelov pracujú mikrosimulačné modely na úrovni jednotlivcov alebo domácností, čím umožňujú zachytiť heterogenitu populácie, analyzovať distribučné účinky navrhovaných opatrení a simulovať správanie rôznych populačných skupín v meniacich sa ekonomických a demografických podmienkach. Vďaka tejto vlastnosti nachádzajú uplatnenie v širokom spektre oblastí vrátane daňovo-dávkových systémov, trhu práce, dôchodkových systémov, zdravotníctva, školstva, regionálneho rozvoja a populačných projekcií (Orcutt, 1957; Harding, 1996; Figari et al., 2015).

Rozvoj mikrosimulačných modelov bol podmienený nielen rastúcim výpočtovým výkonom, ale predovšetkým dostupnosťou kvalitných mikroúdajov. Zatiaľ čo prvé modely boli vytvárané na relatívne malých experimentálnych databázach, súčasné mikrosimulácie využívajú rozsiahle harmonizované databázy oficiálnej štatistiky, administratívne registre alebo syntetické populácie vytvorené kombináciou viacerých dátových zdrojov. Vývoj mikrosimulačného modelovania tak postupne presunul pozornosť z otázky konštrukcie samotného modelu na otázku kvality vstupných údajov, ich reprezentatívnosti, priestorovej detailnosti a schopnosti zachytiť relevantné charakteristiky modelovanej populácie (Chapuis et al., 2022; Kokosi et al., 2022). Výber vstupných údajov dnes predstavuje jedno z najdôležitejších metodologických rozhodnutí pri tvorbe mikrosimulačného modelu. Každý dátový zdroj bol vytvorený na odlišný účel a odlišným spôsobom zachytáva sociálnu, ekonomickú alebo demografickú štruktúru populácie. Výberové zisťovania poskytujú podrobné individuálne charakteristiky, avšak ich využitie limituje veľkosť výberovej vzorky. Populačné cenzy zabezpečujú takmer úplné pokrytie populácie a vysokú priestorovú detailnosť, obsahujú však podstatne menej ekonomických premenných. Administratívne registre ponúkajú vysokú presnosť a aktuálnosť údajov, avšak vznikajú na administratívne, nie výskumné účely. Žiadny z uvedených zdrojov preto samostatne neposkytuje všetky informácie potrebné pre komplexné populačné mikrosimulačné modelovanie (Ballas et al., 2005; UNECE, 2018; Lovelace & Dumont, 2016), čo vysvetľuje rastúci význam ich vzájomnej integrácie prostredníctvom štatistického prepájania dát a tvorby syntetických populácií.

Napriek intenzívnemu metodologickému rozvoju mikrosimulačných modelov zostáva jedna otázka v odbornej literatúre prekvapivo málo preskúmaná. Väčšina publikovaných štúdií sa sústreďuje na vývoj simulačných algoritmov, validáciu modelov alebo aplikáciu konkrétnych databáz, pričom samotný výber vstupných mikroúdajov býva spravidla považovaný za daný predpoklad modelovania. Len obmedzený počet prác systematicky analyzuje, ako sa výsledky identického mikrosimulačného modelu menia pri použití rôznych oficiálnych štatistických zdrojov. Tento nedostatok je o to významnejší, že charakter vstupných údajov ovplyvňuje reprezentatívnosť modelova-

nej populácie, dostupnosť vysvetľujúcich premenných, možnosti priestorovej dezagregácie aj celkovú interpretáciu výsledkov. Predkladaná štúdia reaguje práve na túto metodologickú medzeru. Jej cieľom je analyzovať postavenie troch najvýznamnejších oficiálnych zdrojov mikroúdajov využívaných v európskom prostredí – Labour Force Survey (LFS), European Union Statistics on Income and Living Conditions (EU-SILC) a populačného CENZUSu z pohľadu ich vhodnosti pre populačné mikrosimulačné modelovanie. Na rozdiel od existujúcich prehľadových prác, ktoré jednotlivé databázy opisujú prevažne izolovane, táto štúdia vytvára jednotný metodologický rámec ich komparácie a identifikuje ich silné stránky, obmedzenia a potenciál pre tvorbu populačných mikrosimulačných modelov.

Teoretická časť článku systematicky sumarizuje vývoj mikrosimulačného modelovania, analyzuje vlastnosti jednotlivých oficiálnych zdrojov mikroúdajov, diskutuje význam syntetických populácií a kriticky porovnáva ich využitie v súčasnej literatúre. Na tento prehľad nadväzuje empirická časť štúdie, v ktorej bude vytvorený identický populačný mikrosimulačný model aplikovaný na tri rozdielne vstupné databázy – LFS, EU-SILC a CENZUS. Nakoľko je cieľ práce zameraný na porovnanie výsledkov mikrosimulačného modelu pre rôzne typy vstupných databáz, vybrali sme rok 2021 ako obdobie, v ktorom vznikli databázy všetkých troch typov. Takto navrhnutý experimentálny prístup umožní izolovať vplyv samotného výberu vstupných údajov na výsledky mikrosimulácie a formulovať metodické odporúčania pre výber vhodného dátového základu pri populačnom mikrosimulačnom modelovaní.

2. Vývoj mikrosimulačného modelovania a rastúci význam vstupných mikroúdajov

Mikrosimulačné modelovanie predstavuje jednu z najvýznamnejších metodologických inovácií v oblasti aplikovanej ekonómie, demografie a analýzy verejných politík. Jeho základným princípom je simulácia správania jednotlivcov alebo domácností pri zachovaní heterogenity populácie, čo umožňuje analyzovať dôsledky alternatívnych scenárov na úrovni konkrétnych populačných skupín. Na rozdiel od agregovaných makroekonomických modelov pracujú mikrosimulačné modely s individuálnymi mikroúdajmi, pričom výsledné agregované ukazovatele vznikajú až ako dôsledok simulovaného správania jednotlivých jednotiek. Práve schopnosť zachytiť individuálnu heterogenitu patrí medzi ich najvýznamnejšie metodologické prednosti. Prvá generácia mikrosimulačných modelov bola orientovaná predovšetkým na analýzu daňových systémov, príjmovej redistribúcie a sociálnych transferov (Orcutt, 1957; Harding, 1996; Orcutt et al., 1986). Modely vznikali prevažne vo výskumných inštitúciách a ich spoločným znakom bolo využívanie relatívne malých mikro-databáz, ktorých štruktúra bola prispôbena konkrétnemu analytickému cieľu. Vývoj modelov sa preto sústreďoval najmä na formuláciu simulačných algoritmov a ekonomických vzťahov, zatiaľ čo otázka kvality vstupných údajov zostávala do značnej miery v úzadí.

Významný rozvoj mikrosimulačného modelovania nastal od 80. rokov 20. storočia, keď sa tento prístup začal systematicky uplatňovať pri hodnotení verejných politík v Európe a Austrálii. V tom-

to období vznikali prvé národné daňovo-dávkové mikrosimulačné modely, ktoré sa postupne rozšírili o modelovanie dôchodkových systémov, trhu práce a populačného vývoja. Súčasne sa rozvíjali dynamické mikrosimulačné modely umožňujúce simulovať životné dráhy jednotlivcov prostredníctvom opakovaného modelovania demografických, ekonomických a sociálnych udalostí v čase (Zaidi & Rake, 2001; Harding, 1996; Figari et al., 2015). Významným medzníkom vo vývoji európskeho mikrosimulačného modelovania bol vznik modelu EUROMOD (Sutherland & Figari, 2013), ktorý vytvoril jednotné prostredie na analýzu daňovo-dávkových systémov členských štátov Európskej únie. Okrem harmonizovanej implementácie legislatívnych pravidiel zdôraznil aj význam kvalitných, harmonizovaných mikroúdajov a jednotných validačných postupov ako základného predpokladu porovnateľných a spoľahlivých výsledkov. Súčasný vývoj mikrosimulačného modelovania sa čoraz viac orientuje na integráciu viacerých dátových zdrojov, prepájanie administratívnych registrov s výberovými zisťovaniami a tvorbu syntetických populácií. Pozornosť sa tak presúva od samotnej konštrukcie modelov k zabezpečeniu kvality vstupných mikroúdajov, ktoré významne ovplyvňujú presnosť a interpretáciu modelových výstupov (Chapuis et al., 2022; Hradec et al. 2022).

S rastúcou komplexnosťou mikrosimulačných modelov narastali aj požiadavky na rozsah a kvalitu vstupných mikroúdajov. Zatiaľ čo statické mikrosimulácie bolo možné budovať prevažne na reprezentatívnych výberových zisťovaniach, dynamické modely vyžadujú podrobné informácie o demografických, ekonomických a sociálnych charakteristikách jednotlivcov počas ich životného cyklu. Súčasne rastie význam prepájania rôznych dátových zdrojov a tvorby syntetických populácií, ktoré umožňujú prekonať obmedzenia jednotlivých databáz a lepšie zachytiť heterogenitu populácie. V súčasnom mikrosimulačnom modelovaní sa preto kvalita, reprezentatívnosť a vhodnosť vstupných mikroúdajov považujú za jeden zo základných predpokladov spoľahlivých modelových výstupov.

Vývoj mikrosimulačného modelovania tak možno charakterizovať ako postupný prechod od orientácie na modelové algoritmy k orientácii na kvalitu dátového základu. Kým prvé generácie modelov riešili predovšetkým otázku, ako simulovať správanie jednotlivcov, súčasný výskum čoraz intenzívnejšie skúma otázku, na základe akých údajov by mala byť takáto simulácia realizovaná. Práve tento posun vytvára metodologické východisko predkladanej štúdie, ktorej cieľom nie je porovnávať samotné mikrosimulačné modely, ale systematicky analyzovať vhodnosť rôznych oficiálnych zdrojov mikroúdajov ako ich vstupného dátového základu.

3. Požiadavky na vstupné mikroúdaje v populačnom mikrosimulačnom modelovaní

Úspešnosť mikrosimulačného modelovania je podmienená nielen správnou špecifikáciou modelových algoritmov, ale aj kvalitou vstupných mikroúdajov. S rastúcou dostupnosťou výberových zisťovaní, administratívnych registrov a harmonizovaných medzinárodných databáz sa výber vhodného dátového zdroja stal jedným z rozhodujúcich metodologických krokov pri tvorbe mik-

rosimulačných modelov. Vhodnosť jednotlivých databáz závisí od účelu modelovania, rozsahu sledovaných premenných, reprezentatívnosti populácie a požadovanej úrovne priestorovej alebo tematickej detailnosti. Preto databáza vhodná na analýzu daňovo-dávkových systémov nemusí byť optimálna na populačné projekcie alebo modelovanie trhu práce (Sutherland & Figari, 2013; Chapuis et al., 2022).

Z dostupnej literatúry možno identifikovať niekoľko základných charakteristík, ktoré určujú využitelnosť mikroúdajov pre populačné mikrosimulačné modelovanie. Ide predovšetkým o reprezentatívnosť populácie, rozsah dostupných premenných, priestorovú úroveň údajov, časovú aktualnosť, možnosť kalibrácie a validácie, kontinuitu časových radov a možnosti prepájania s ďalšími dátovými zdrojmi. Tieto charakteristiky sa vo väčšine publikovaných štúdií objavujú opakovane, avšak len zriedkavo sú systematicky využité ako jednotný rámec na porovnanie jednotlivých oficiálnych štatistických zdrojov. Práve vytvorenie takéhoto rámca predstavuje jeden z metodologických prínosov tejto štúdie.

Reprezentatívnosť populácie

Reprezentatívnosť patrí medzi základné kritériá kvality vstupných mikroúdajov v mikrosimulačnom modelovaní. Keďže modely simulujú správanie jednotlivcov alebo domácností, vstupná databáza musí čo najvernejšie reprodukovать štruktúru cieľovej populácie vrátane vzťahov medzi demografickými, ekonomickými a sociálnymi charakteristikami. Nedostatočná reprezentatívnosť alebo nevhodná kalibrácia vstupných údajov môže viesť k systematickému skresleniu modelových výstupov, ktoré nemožno úplne odstrániť následným modelovaním alebo agregáciou výsledkov (Deville & Särndal, 1992; Chapuis et al., 2022).

Rozsah a obsah premenných

Druhým rozhodujúcim kritériom kvality vstupných mikroúdajov je rozsah dostupných informácií. Požiadavky na vstupnú databázu sa líšia podľa typu mikrosimulačného modelu – modely trhu práce vyžadujú detailné údaje o ekonomickej aktivite, populačné modely predovšetkým demografické charakteristiky a daňovo-dávkové modely podrobné informácie o príjmoch, sociálnych transferoch a domácnostiach. Keďže žiadny z bežne využívaných zdrojov mikroúdajov neposkytuje všetky potrebné premenné v požadovanom rozsahu, pri tvorbe mikrosimulačných modelov sa čoraz častejšie využíva kombinácia výberových zisťovaní, populačných cenзов, administratívnych registrov alebo syntetických populácií. Výber vhodného dátového základu preto závisí od analytického účelu modelu a od charakteru výstupov, ktoré majú byť prostredníctvom mikrosimulácie analyzované (Li & O'Donoghue, 2013; Sutherland & Figari, 2013).

Priestorová úroveň údajov

S rastúcim významom regionálne orientovaných verejných politík narastajú aj požiadavky na priestorovú detailnosť vstupných mikroúdajov. Národná reprezentatívnosť už sama osebe nepostačuje, keďže mnohé mikrosimulačné modely sa využívajú na analýzu regionálnych rozdielov,

plánovanie verejných služieb alebo tvorbu lokálnych populačných projekcií. Významným kritériom pri výbere vstupnej databázy sa preto stáva dostupnosť údajov na úrovni regiónov, okresov alebo obcí, ako aj schopnosť zachytiť priestorovú heterogenitu populácie. Práve v tejto oblasti sa najvýraznejšie prejavujú rozdiely medzi populačnými cenami, výberovými zisťovaniami a administratívnymi registrami (Tanton & Edwards, 2013; Lovelace & Dumont, 2016).

Časová aktuálnosť a kontinuita údajov

Presnosť mikrosimulačných modelov závisí aj od aktuálnosti vstupných mikroúdajov. Časový odstup medzi referenčným obdobím vstupnej databázy a obdobím simulácie môže viesť k systematickým odchýlkam výsledkov, najmä pri modelovaní dynamických ekonomických a demografických procesov. Z tohto dôvodu sa v posledných rokoch rozvíjajú metodické prístupy založené na pravidelnej aktualizácii vstupných údajov, prepájaní administratívnych registrov s výberovými zisťovaniami a využívaní nowcastingových techník na zlepšenie aktuálnosti mikrosimulačných analýz (O'Donoghue & Sologon, 2023).

Kalibrácia a validácia

Spôľahlivosť mikrosimulačných modelov úzko súvisí s kalibráciou vstupných mikroúdajov a validáciou modelových výstupov. Moderné mikrosimulačné modely využívajú kalibračné postupy založené na úprave váh respondentov alebo na prepájaní viacerých dátových zdrojov s cieľom zabezpečiť súlad vstupných údajov s oficiálnymi populačnými štatistikami. Rovnako dôležitá je validácia modelových výstupov prostredníctvom nezávislých štatistických alebo administratívnych údajov, ktorá umožňuje posúdiť presnosť a konzistentnosť simulovaných výsledkov. Kalibrácia a validácia preto predstavujú základné metodologické kroky pri tvorbe moderných mikrosimulačných modelov (Deville & Särndal, 1992; Li & O'Donoghue, 2013).

Možnosti integrácie dátových zdrojov

V poslednom desaťročí sa medzi významné kritériá kvality vstupných mikroúdajov zaradila aj možnosť ich prepájania s ďalšími dátovými zdrojmi. Rozvoj syntetických populácií, štatistického prepájania databáz a využívania administratívnych registrov rozšíril možnosti konštrukcie mikrosimulačných modelov a umožnil efektívnejšie kombinovať prednosti jednotlivých zdrojov údajov. Pri výbere vstupnej databázy preto zohráva významnú úlohu nielen jej obsah a reprezentatívnosť, ale aj schopnosť integrácie s ďalšími štatistickými alebo administratívnymi údajmi (Chapuis et al., 2022; Kokosi et al., 2022; Hradec et al. 2022).

Z uvedených kritérií vyplýva, že neexistuje univerzálny zdroj mikroúdajov vhodný pre všetky typy mikrosimulačných modelov. Jednotlivé databázy sa odlišujú svojou reprezentatívnosťou, rozsahom premenných, priestorovou detailnosťou, časovou aktuálnosťou aj možnosťami integrácie s ďalšími zdrojmi údajov. Namiesto hľadania „najlepšej“ databázy je preto metodologicky vhodnejšie analyzovať ich vlastnosti vo vzťahu ku konkrétnemu účelu modelovania. V nadväznosti na tento prístup sa ďalšia časť štúdie venuje systematickej analýze štyroch najvýznamnejších zdrojov

oficiálnych mikroúdajov využívaných v populačnom mikrosimulačnom modelovaní: LFS, EU-SILC, CENZUS a administratívnych registrov, pričom ich hodnotenie vychádza z kritérií formulovaných v tejto kapitole.

Kvalita vstupných mikroúdajov predstavuje jeden z rozhodujúcich faktorov ovplyvňujúcich výsledky mikrosimulačného modelovania. Napriek významnému metodologickému rozvoju mikrosimulačných modelov sa však odborná literatúra sústreďuje predovšetkým na vývoj simulačných algoritmov, kalibračných postupov alebo validáciu modelových výstupov, zatiaľ čo samotný výber vstupného dátového zdroja býva vo väčšine prípadov považovaný za exogénny predpoklad modelovania. Vo väčšine publikovaných štúdií je preto konkrétny zdroj mikroúdajov zvolený ešte pred samotnou konštrukciou modelu a jeho vhodnosť nebýva predmetom systematického skúmania. Takýto prístup je pochopiteľný z pohľadu aplikačne orientovaných štúdií, ktorých cieľom je riešenie konkrétneho analytického problému. Z metodologického hľadiska však vytvára významné obmedzenie. Ak súčasne dochádza k zmene vstupných údajov, modelových predpokladov aj samotného algoritmu, nie je možné jednoznačne určiť, akú časť rozdielov vo výsledkoch spôsobuje samotný výber vstupnej databázy. Výsledky publikovaných mikrosimulačných štúdií preto len výnimočne umožňujú oddeliť vplyv modelovej štruktúry od vplyvu dátového základu.

Predkladaná štúdia vychádza z odlišného metodologického princípu. Namiesto porovnávania rôznych mikrosimulačných modelov alebo rôznych simulačných algoritmov sa sústreďuje na jedinú premennú – voľbu vstupných mikroúdajov. Základným princípom navrhovaného výskumu je zachovanie identickej štruktúry populačného mikrosimulačného modelu pri súčasnej systematickej zmene vstupnej databázy. Všetky ostatné komponenty modelu, vrátane simulačných algoritmov, demografických predpokladov a spôsobu agregácie výsledkov, zostanú nezmenené. Jediným experimentálne meneným prvkom bude zdroj mikroúdajov využitý na inicializáciu modelovanej populácie, prípadne nastavenie parametrov vychádzajúcich priamo zo vstupných mikroúdajov. Takto navrhnutý experiment umožňuje izolovať vplyv vstupných údajov od ostatných faktorov ovplyvňujúcich výsledky mikrosimulácie. Z metodologického hľadiska ide o obdobu citlivostnej analýzy, ktorá sa v modelovaní tradične využíva na hodnotenie vplyvu parametrov modelu na výsledné výstupy. V predkladanej štúdii však nebude predmetom citlivostnej analýzy modelový parameter, ale samotný dátový základ modelu. Pozornosť sa tak presúva od otázky ako model simulovať k otázke na základe akých údajov model simulovať. Tento prístup sa v literatúre objavuje len okrajovo, hoci význam vstupných údajov je opakovane zdôrazňovaný v metodických prácach venujúcich sa mikrosimulačnému modelovaniu.

Experiment bude realizovaný na troch najvýznamnejších zdrojoch oficiálnych mikroúdajov využívaných v európskom štatistickom systéme: LFS, EU-SILC a CENZUS. Výber týchto databáz nie je náhodný. Každá z nich predstavuje odlišnú filozofiu zberu štatistických údajov a súčasne poskytuje rozdielny rozsah informácií o modelovanej populácii. LFS je orientovaný predovšetkým na ekonomickú aktivitu a charakteristiky trhu práce, EU-SILC na príjmovú situáciu a životné podmienky domácností a populačný CENZUS na demografickú a priestorovú štruktúru populácie. Každý z

týchto zdrojov preto disponuje špecifickými prednosťami aj obmedzeniami, ktoré môžu ovplyvniť výsledky populačného mikrosimulačného modelovania. Pre účely článku sme ako verzie vstupných databáz vybrali rok 2021, v ktorom sa uskutočnil zber dát pre všetky tri datasety. Koncepčný rámec štúdie predpokladá, že rozdiely medzi jednotlivými databázami sa neprejavujú iba v rozsahu dostupných premenných, ale aj v samotnej reprezentácii populácie. Rozdielna veľkosť výberových súborov, odlišná periodicita zberu údajov, rozdielna úroveň priestorovej detailnosti, dostupnosť ekonomických charakteristík alebo možnosti kalibrácie môžu ovplyvniť presnosť populačných projekcií, stabilitu modelových odhadov aj interpretáciu výsledkov. Cieľom štúdie preto nie je identifikovať univerzálne najlepší zdroj údajov, ale analyzovať, za akých podmienok je konkrétny zdroj vhodnejší pre populačné mikrosimulačné modelovanie. Na základe uvedených východísk možno formulovať centrálnu výskumnú otázku štúdie:

Do akej miery ovplyvňuje voľba oficiálneho zdroja mikroúdajov výsledky populačného mikrosimulačného modelu pri zachovaní identickej modelovej štruktúry?

Takto koncipovaný výskum predstavuje posun od tradičného porovnávania jednotlivých databáz k experimentálnemu hodnoteniu ich vplyvu na výsledky mikrosimulačného modelovania. Výstupom štúdie preto nebude iba opis silných a slabých stránok jednotlivých zdrojov mikroúdajov, ale metodicky podložené odporúčania pre ich využitie pri konštrukcii populačných mikrosimulačných modelov.

4. Metodika výskumu

4.1. Výskumný dizajn

Cieľom empirickej časti štúdie je experimentálne posúdiť, do akej miery ovplyvňuje výber vstupných mikroúdajov výsledky populačného mikrosimulačného modelovania. Na rozdiel od väčšiny existujúcich štúdií, ktoré porovnávajú odlišné modelové prístupy alebo alternatívne simulačné algoritmy, je predmetom tejto analýzy výlučne vplyv vstupného dátového základu. Výskumný dizajn preto vychádza z princípu kontrolovaného experimentu, v ktorom zostávajú všetky komponenty mikrosimulačného modelu nezmenené a jedinou systematicky menenou premennou je zdroj mikroúdajov použitý na inicializáciu modelovanej populácie a vybrané parametre pochádzajúce práve zo vstupnej databázy. Takto navrhnutý prístup umožňuje oddeliť vplyv vstupných údajov od ostatných faktorov ovplyvňujúcich výsledky mikrosimulácie. Rozdiely medzi jednotlivými simuláciami preto nemožno interpretovať ako dôsledok rozdielnych modelových predpokladov alebo algoritmov, ale ako dôsledok odlišných vlastností použitých mikroúdajov. Experiment možno z metodologického hľadiska chápať ako citlivosťnú analýzu zameranú na dátový základ modelu, pričom analyzovaným faktorom nie je hodnota modelového parametra, ale samotný zdroj vstupných údajov.

4.2. Harmonizácia vstupných údajov

Keďže jednotlivé databázy vznikajú na odlišné štatistické účely, ich priame porovnanie nie je možné bez predchádzajúcej harmonizácie. Pred samotnou simuláciou budú preto všetky vstupné údaje transformované do jednotnej dátovej štruktúry. Každá vstupná databáza obsahuje spolu len 5 premenných, ktoré sa budú využívať pri mikrosimulácií. Tento obmedzený a relatívne malý počet premenných je dostatočne obsiahly na účel porovnania výsledkov simulácií za predpokladu využitia rovnakého modelu. Medzi tieto premenné patria:

- Pohlavie – rozdelená na dve kategórie muž/žena.
- Vek – premenná určujúca presný vek respondenta v členení po jednom roku. V prípade LFS je vek v vstupnom datasete definovaný na 5 ročných kohortách. V tomto prípade sme pristúpili k dezagregácii 5 ročných kohort na jednoročné s využitím administratívnych údajov pochádzajúcich zo Štatistického úradu.
- Najvyššie dosiahnuté vzdelanie – agregované na 4 skupiny: základné, stredné bez maturity, stredné s maturitou a vysokoškolské.
- Status – popisuje ekonomický status respondenta. Táto premenná je rozdelená na 5 skupín: pracujúci, nezamestnaný, študent, dôchodca, inak neaktívny. Takto definovaný status jednotlivca následne môžeme rozdeliť na dve skupiny: ekonomicky aktívna, alebo ekonomicky neaktívna populácia.
- Krajina narodenia – premenná slúžiaca na identifikáciu osoby pochádzajúcej mimo územia Slovenska.

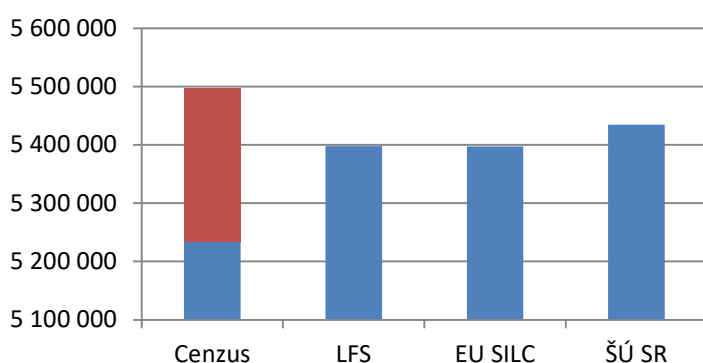
Cieľom harmonizácie nie je odstrániť prirodzené rozdiely medzi databázami, ale zabezpečiť, aby porovnanie výsledkov nebolo ovplyvnené technickými rozdielmi vo formáte alebo definíciách premenných.

4.3. Porovnanie vstupných údajov

Porovnanie vstupných údajov poukazuje na rozdiely zachytené v jednotlivých dátových zdrojoch. Referenčné údaje Štatistického úradu Slovenskej republiky uvádzajú na konci roka 2021 celkový počet obyvateľov Slovenska 5 434 712. Po aplikovaní príslušných populačných váh poskytujú výberové zisťovania LFS a EU-SILC veľmi podobné odhady veľkosti populácie. V prípade LFS predstavuje odhadovaný počet obyvateľov 5 397 603 osôb a v prípade EU-SILC 5 397 337 osôb. Rozdiel medzi týmito dvoma výberovými zisťovaniami dosahuje iba 266 osôb, čo poukazuje na takmer identickú reprodukciu celkovej veľkosti populácie. V porovnaní s referenčným údajom Štatistického úradu SR je však odhad populácie v LFS a EU-SILC nižší o viac ako 37 tisíc osôb, teda približne o 0,7 %. Rozdiely sú z hľadiska celkovej populácie relatívne malé a naznačujú, že populačné váhy oboch výberových zisťovaní umožňujú reprodukovať agregovanú veľkosť populácie s vysokou mierou zhody. Súčasne však potvrdzujú, že vážené údaje z výberových zisťovaní nepredstavujú úplnú reprodukciu administratívne alebo demograficky evidovaného počtu obyvateľov. Aj relatívne malá odchýlka v počiatočnej populácii môže pri dlhodobom mikrosimulačnom modelovaní ovplyvniť absolútnu úroveň výsledných populačných projekcií.

Výraznejší rozdiel možno identifikovať pri údajoch populačného cenzu. Vstupná populácia odvodená z cenzových údajov predstavuje 5 233 612 osôb, čo je o 201 100 osôb menej než referenčný počet ŠÚ SR. Relatívna odchýlka dosahuje približne 3,70 %. Nižší počet obyvateľov však nemožno automaticky interpretovať ako nepresnosť cenzových údajov. V analyzovanom súbore je počet obyvateľov vymedzený podľa súčasného pobytu na území Slovenskej republiky, pričom počet osôb s trvalým pobytom na Slovensku, ktorých súčasný pobyt bol v čase sčítania mimo územia krajiny je na úrovni 264 269 osôb. Po započítaní týchto osôb do datasetu cenzu dostávame celkový počet osôb na úrovni 5 497 881 osôb, čo je o 63 169 viac v porovnaní s údajmi ŠÚ SR. Relatívna odchýlka predstavuje 1,2 %. Rozdiel preto vyplýva najmä z odlišného vymedzenia cieľovej populácie a nie nevyhnutne z rozdielnej kvality dátového zdroja.

Graf 1: Veľkosť populácie

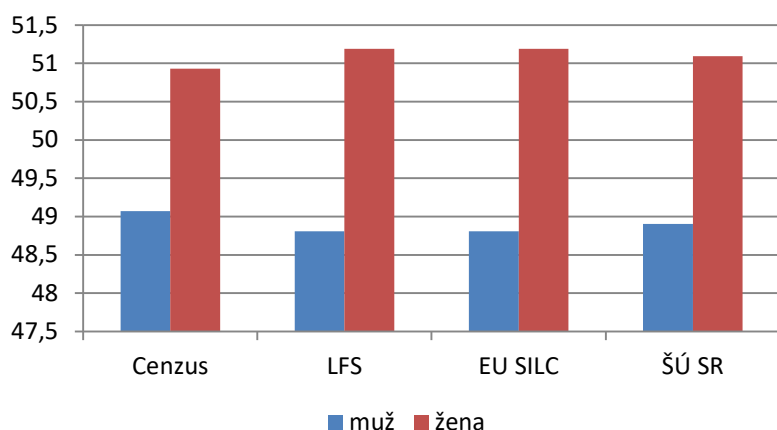


Zdroj: CENZUS 2021, LFS 2021, EÚ SILC 2021 a ŠÚ SR

Uvedené zistenie poukazuje na význam harmonizácie populačných konceptov pred aplikáciou jednotlivých databáz v mikrosimulačnom modeli. Ak vstupné súbory nevychádzajú z rovnakej definície obvykle bývajúceho, súčasného alebo trvalo bývajúceho obyvateľstva, rozdiely vo výsledkoch simulácie môžu odrážať rozdielne vymedzenie populácie namiesto samotných vlastností dátového zdroja. Pri interpretácii výsledkov je preto potrebné rozlišovať medzi rozdielmi spôsobenými výberovou variabilitou, systémom populačných váh a rozdielmi vyplývajúcimi z definície cieľovej populácie. V nasledujúcich krokoch sme sa rozhodli zahrnúť do celkovej populácie aj osoby, ktoré majú trvalý pobyt na území Slovenska ale súčasný mimo územia. Pre tento krok sme sa rozhodli predovšetkým kvôli obtiažnej identifikácii populácie, ktorej návrat na Slovensko je možný ako napríklad: sezónny pracovníci, študenti v zahraničí a iní.

Z hľadiska štruktúry populácie podľa pohlavia vykazujú všetky analyzované zdroje vysokú mieru zhody. Podľa referenčných údajov Štatistického úradu SR tvoria muži 48,91 % a ženy 51,09 % populácie. LFS aj EU-SILC reprodukovujú rovnakú štruktúru, pričom podiel mužov dosahuje 48,81 % a podiel žien 51,19 %. Odchýlka od referenčných údajov tak predstavuje iba približne 0,10 percentuálneho bodu. Výsledky naznačujú, že napriek výberovému charakteru oboch zisťovaní ich populačné váhy veľmi presne reprodukovujú základné rozdelenie obyvateľstva podľa pohlavia.

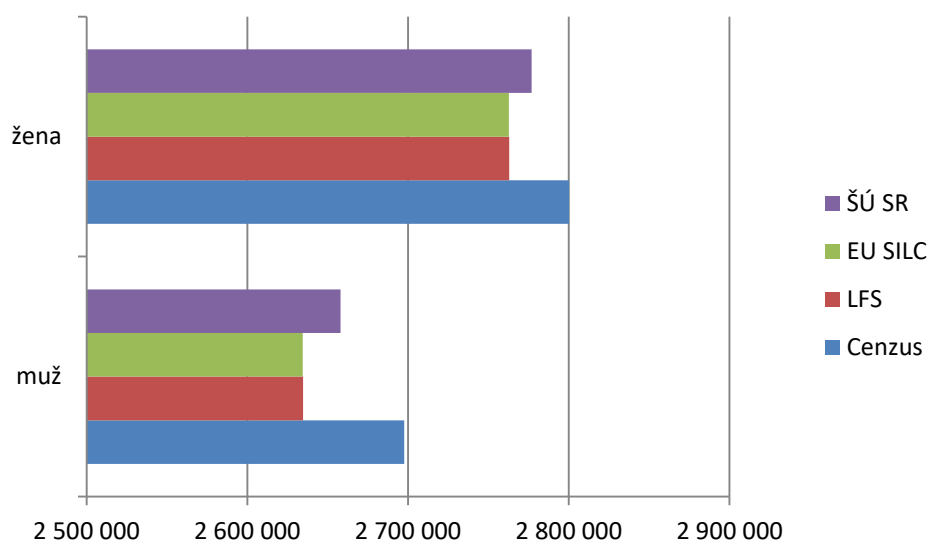
Graf 2: Percentuálne zloženie obyvateľstva podľa pohlavia



Zdroj: CENZUS 2021, LFS 2021, EÚ SILC 2021 a ŠÚ SR

Cenzové údaje vykazujú mierne vyšší podiel mužov. Muži predstavujú 49,07 % a ženy 50,93 % analyzovanej populácie. V porovnaní s údajmi Štatistického úradu SR je podiel mužov vyšší približne o 0,16 percentuálneho bodu a podiel žien o rovnakú hodnotu nižší. Aj táto odchýlka je však veľmi malá a z hľadiska celkovej štruktúry populácie nepredstavuje zásadný rozdiel. Všetky tri vstupné databázy konzistentne zachytávajú mierne početnú prevahu žien v populácii Slovenska. Pri porovnaní absolútnych počtov sú rozdiely výraznejšie, keďže odrážajú predovšetkým rozdielnu celkovú veľkosť vstupnej populácie. Rozdiely medzi LFS a EU-SILC sú minimálne. Obe výberové zisťovania preto poskytujú takmer identickú reprodukciu nielen relatívneho rozdelenia podľa pohlavia, ale aj vážených absolútnych početností.

Graf 3: Zloženie obyvateľstva podľa pohlavia



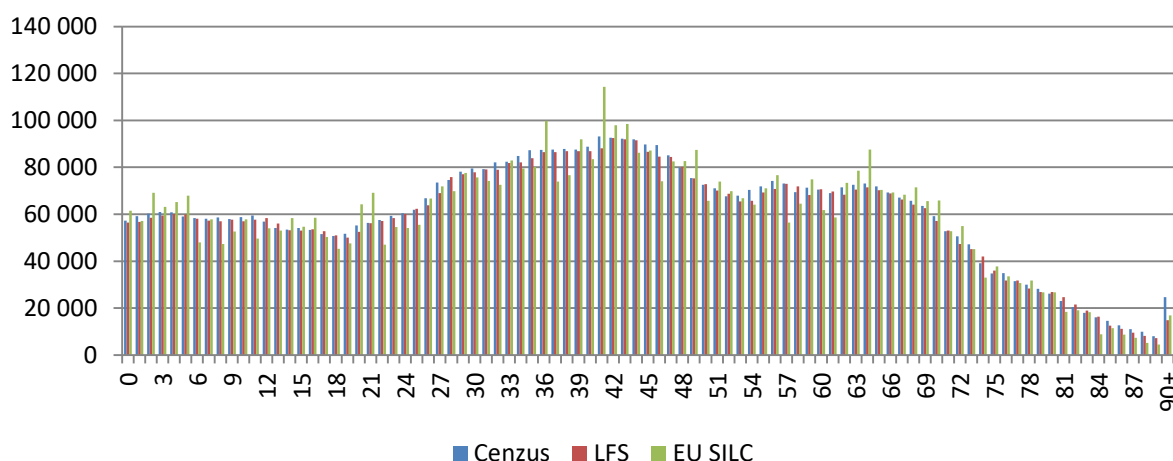
Zdroj: CENZUS 2021, LFS 2021, EÚ SILC 2021 a ŠÚ SR

Výsledky celkovo ukazujú, že rozdiely medzi vstupnými zdrojmi sa prejavujú výraznejšie v celkovom počte zahrnutých obyvateľov než v ich základnej štruktúre podľa pohlavia. LFS a EU-SILC poskytujú po aplikovaní populačných váh takmer identické výsledky a veľmi tesne sa približujú

referenčným údajom Štatistického úradu SR. Cenzové údaje vykazujú nižší celkový počet obyvateľov, pričom podstatná časť rozdielu súvisí s odlišným vymedzením analyzovanej populácie podľa miesta súčasného pobytu. Napriek rozdielom v absolútnej veľkosti populácie zostáva jej štruktúra podľa pohlavia vo všetkých zdrojoch veľmi podobná. Z pohľadu následného mikrosimulačného experimentu možno predpokladať, že rozdiely v zastúpení mužov a žien nebudú samy osebe významným zdrojom odlišností medzi výsledkami jednotlivých scenárov. Väčší vplyv môže mať rozdielna počiatočná veľkosť populácie, najmä ak budú modelové výstupy prezentované v absolútnych počtoch. Pri hodnotení vplyvu jednotlivých dátových zdrojov bude preto potrebné oddeliť efekt rozdielnej veľkosti vstupnej populácie od efektu jej vnútornej demografickej štruktúry.

Porovnanie vekového zloženia populácie poukazuje na rozdiely medzi údajmi CENZUS, LFS a EU-SILC. Napriek rozdielnemu charakteru a metodike jednotlivých zdrojov zachytávajú všetky analyzované súbory základný priebeh vekovej štruktúry obyvateľstva. Vo všetkých prípadoch možno identifikovať nižšie zastúpenie mladších vekových skupín, nárast početnosti populácie vo veku približne od 25 do 40 rokov a následný postupný pokles počtu osôb vo vyšších vekových kategóriách. Výraznejšie zastúpenie populačne silnejších ročníkov je viditeľné najmä v strednom a vyššom produktívnom veku.

Graf 4: Vekové zloženie obyvateľstva SR



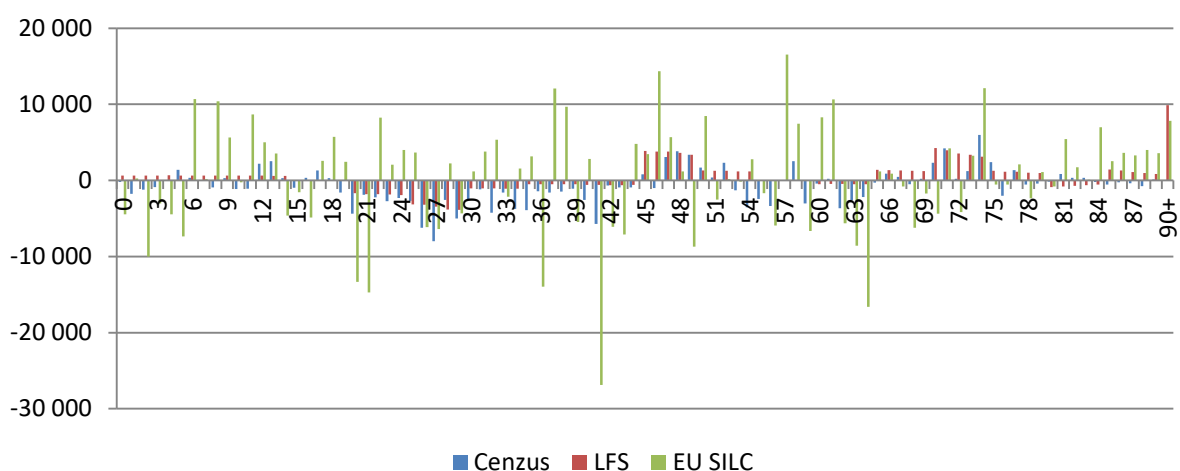
Zdroj: CENZUS 2021, LFS 2021 a EÚ SILC 2021

Údaje z CENZUSu vykazujú v porovnaní s ostatnými zdrojmi relatívne plynulý priebeh vekovej štruktúry. Je to dôsledok populačného charakteru sčítania, ktoré zahŕňa prakticky celé obyvateľstvo a nie je založené na výberovej vzorke. Hodnoty za jednotlivé vekové skupiny preto menej podliehajú náhodným výkyvom a vo väčšej miere zodpovedajú referenčnému vekovému rozloženiu populácie. Odchýlky údajov CENZUSu od údajov Štatistického úradu SR (viď graf 5) sú prevažne systematické a vo väčšine vekových kategórií nadobúdajú záporné hodnoty. Výraznejšie rozdiely možno pozorovať najmä pri niektorých mladších vekových skupinách a vo vyššom veku.

LFS vo všeobecnosti reprodukuje základný profil vekovej štruktúry populácie, v jednotlivých vekových kategóriách však vykazuje väčšiu variabilitu ako CENZUS. Rozdiely oproti referenčným úda-

jom Štatistického úradu SR sa pohybujú v kladnom aj zápornom smere, pričom výraznejšie odchýlky možno identifikovať najmä pri niektorých vekových skupinách produktívneho veku a medzi staršími osobami. Tieto rozdiely môžu súvisieť s výberovým charakterom zisťovania, veľkosťou vzorky a následným použitím váh pri prepočte výsledkov na celkovú populáciu. Napriek lokálnym odchýlkam údaje LFS zachovávajú celkový priebeh vekového zloženia obyvateľstva aj vďaka dezagregácii 5 ročných kohort na jednorokové s využitím práve údajov ŠÚ SR. Tým pádom sa potenciálne veľké odchýlky vyhladili a celkový populačný profil je podobný s údajmi pochádzajúcimi zo ŠÚ SR. Na druhej strane použitím metódy delenia 5 ročných vekových kohort na jednorokové dochádza k značnému skresleniu ďalších informácií počas života jedinca, ako je napríklad ekonomický status (odchod do dôchodku).

Graf 5: Rozdiely v populačných kohortách voči údajom zo ŠÚ SR.



Zdroj: CENZUS 2021, LFS 2021, EÚ SILC 2021 a ŠÚ SR

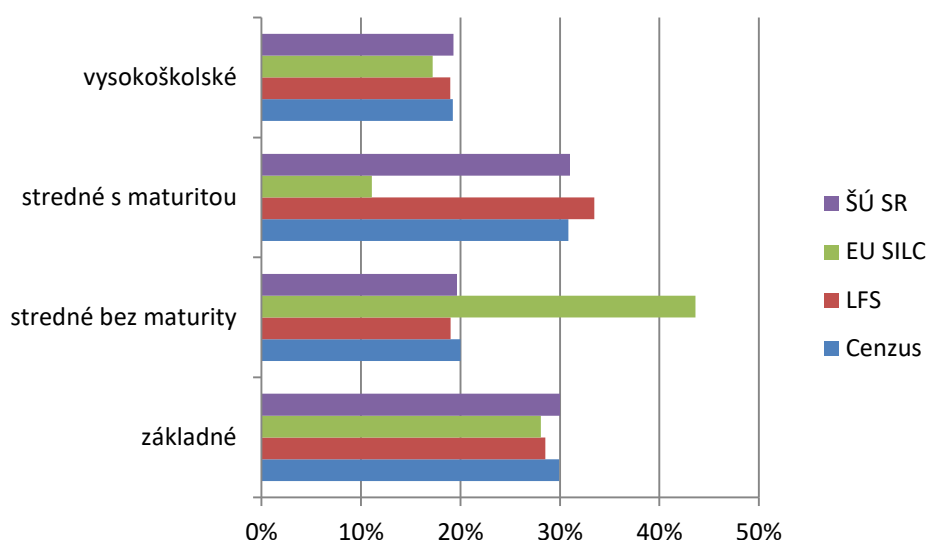
Najvýraznejšiu variabilitu vykazujú údaje zo zisťovania EU-SILC. Pri viacerých jednotlivých vekoch možno pozorovať výraznejšie kladné alebo záporné odchýlky od referenčných údajov. Väčšie rozdiely sa vyskytujú najmä pri najmladších vekových kategóriách, v strednom produktívnom veku a pri vybraných skupinách staršieho obyvateľstva. V niektorých vekových kategóriách dosahujú rozdiely rádovo desiatky tisíc osôb. Zvýšená variabilita môže byť spôsobená menším rozsahom výberového súboru, keďže zisťovanie EU-SILC je primárne zamerané na analýzu príjmov, životných podmienok a sociálneho postavenia domácností. Pri členení údajov podľa jednotlivých rokov veku sa preto môže výraznejšie prejavovať výberová chyba a spôsob váženía pozorovaní. Výsledky porovnania naznačujú, že všetky analyzované zdroje zachytávajú hlavné demografické charakteristiky vekového zloženia populácie, odlišujú sa však mierou variability na úrovni jednotlivých rokov veku.

Vzdelanostná štruktúra patrí medzi najvýznamnejšie charakteristiky vstupnej populácie, keďže ovplyvňuje široké spektrum demografických, ekonomických a sociálnych procesov modelovaných v mikrosimulačných modeloch. Dosiahnuté vzdelanie je v realite prepojené s ekonomickou aktivitou, príjmovou situáciou, plodnosťou, úmrtnosťou aj migráciou. Rozdiely vo vzdelanostnom zložení vstupnej populácie preto môžu významne ovplyvniť výsledky mikrosimulačných projekcií aj v

případe, že ostatné charakteristiky populácie zostávajú porovnateľné. Porovnanie vzdelanostnej štruktúry celej populácie ukazuje, že všetky analyzované databázy reprodukovujú základné rozdelenie obyvateľstva podľa najvyššieho dosiahnutého vzdelania, avšak medzi jednotlivými zdrojmi existujú rozdiely v zastúpení jednotlivých vzdelanostných skupín. Najväčšia zhoda je pozorovateľná pri podiele obyvateľov so základným vzdelaním, kde sa odhady jednotlivých databáz odlišujú len minimálne. Podobne vysokú mieru zhody možno identifikovať aj pri podiele vysokoškolsky vzdelanej populácie, pričom rozdiely medzi jednotlivými zdrojmi nepresahujú niekoľko percentuálnych bodov. Z hľadiska populačného mikrosimulačného modelovania možno tieto rozdiely považovať za relatívne malé.

Výraznejšie rozdiely sa prejavujú pri obyvateľoch so stredoškolským vzdelaním. Najmä kategória stredného vzdelania bez maturity vykazuje podstatne vyššiu variabilitu medzi jednotlivými zdrojmi údajov. Údaje EU-SILC indikujú výrazne vyšší podiel tejto skupiny v porovnaní s ostatnými databázami, zatiaľ čo Labour Force Survey, CENZUS aj referenčné údaje Štatistického úradu SR vykazujú podstatne nižšie a navzájom porovnateľné zastúpenie. Tento nesúlad je spôsobený predovšetkým inou kategorizáciou dosiahnutého vzdelania v prípade zisťovania EU-SILC. Zatiaľ čo pre populáciu 16-34 je možné jednoznačne rozdeliť kategóriu strednej školy na maturitu/bez maturity, tak v prípade staršej vekovej kohorty to nie je možné. Preto boli v tomto prípade staršie vekové kohorty zaradené do kategórie stredné školy bez maturity. Vyššie zastúpenie osôb s dosiahnutým vzdelaním bez maturity v databáze EU-SILC sa súčasne prejavuje nižším podielom osôb so stredoškolským vzdelaním ukončeným maturitou.

Graf 6: Vzdelanostné zloženie populácie Slovenska



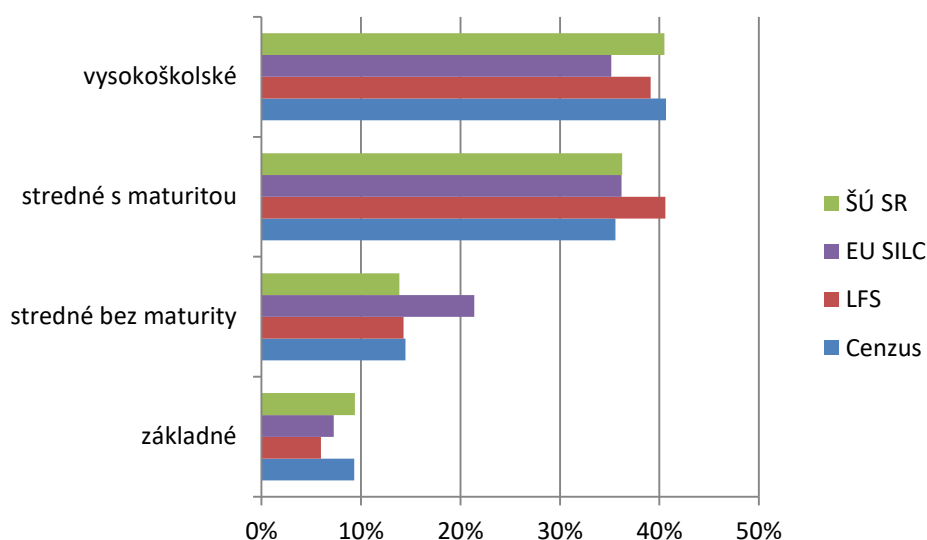
Zdroj: CENZUS 2021, LFS 2021, EÚ SILC 2021 a ŠÚ SR

Tieto rozdiely pravdepodobne súvisia s odlišnou metodikou zberu údajov a klasifikáciou vzdelania v jednotlivých zisťovaniach. LFS je primárne orientovaný na charakteristiky trhu práce, zatiaľ čo EU-SILC vznikol predovšetkým na analýzu príjmov a životných podmienok domácností. Odlišné zameranie zisťovaní môže ovplyvniť spôsob zisťovania najvyššieho dosiahnutého vzdelania, ako

aj následnú harmonizáciu jednotlivých kategórií podľa medzinárodnej klasifikácie ISCED. Pri populačnom CENZUSE je vzdelanie zisťované plošne pre celú populáciu, čo znižuje výberovú variabilitu a poskytuje stabilnejší obraz o vzdelanostnej štruktúre obyvateľstva.

Menej významné rozdiely medzi databázami sa prejavujú pri analýze mladších vekových kohort. V populácii vo veku 30 až 35 rokov dochádza k očakávanému poklesu podielu osôb so základným vzdelaním a súčasne k výraznému nárastu zastúpenia vysokoškolsky vzdelanej populácie. Tento trend zachytávajú všetky analyzované zdroje údajov, čo potvrdzuje rastúcu úroveň vzdelania mladších generácií. Najvýraznejšie rozdiely možno opäť identifikovať pri osobách so stredoškolským vzdelaním bez maturity v prípade zisťovania EU-SILC. Zvyšné databázy dosahujú podobný podiel s týmto vzdelaním. Pri osobách so stredným vzdelaním s maturitou na druhej strane dosahuje zisťovanie LFS vyšší podiel zastúpenia v porovnaní s ostatnými databázami. Tieto vyššie podiely v prípade stredného vzdelania sa kompenzujú nižšími hodnotami v prípade vysokoškolského a základného vzdelania, kde vyššie hodnoty dosahuje CENZUS a referenčné údaje Štatistického úradu SR.

Graf 7: Vzdelanostné zloženie obyvateľov Slovenska vo veku od 30 do 35 rokov



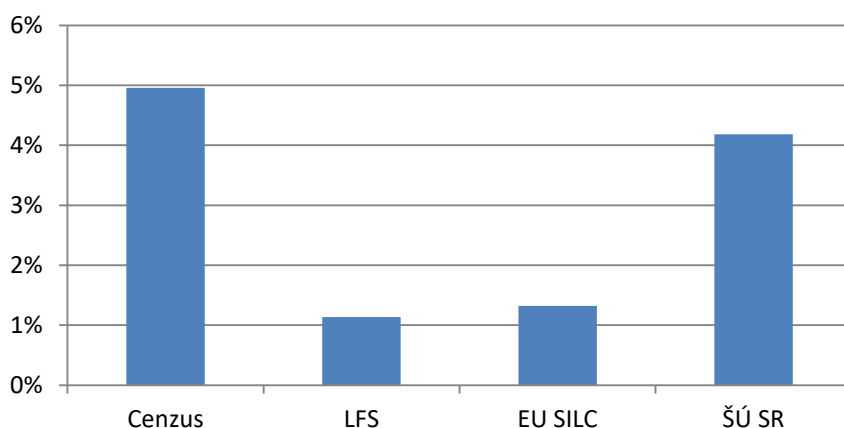
Zdroj: CENZUS 2021, LFS 2021, EÚ SILC 2021 a ŠÚ SR

Z metodologického hľadiska je dôležité, že rozdiely medzi databázami v prípade vzdelanostnej štruktúry obyvateľstva vo veku 30 až 35 rokov nie sú veľké. Práve táto skupina je referenčná pri tvorbe vzdelanostnej štruktúry v modelovacej časti. Vo vzdelávacom module mikrosimulačného modelu sa presne toto zastúpenie zachováva pri simulovaní najvyššieho dosiahnutého vzdelania populácie ďalšom období.

V prípade premennej týkajúcej sa pôvodu narodenia dosahujú databázy rozdielne hodnoty. Zatiaľ čo CENZUS a údaje zo ŠÚ SR uvádzajú podiel obyvateľov narodených mimo územia Slovenska 4 %, resp. 5 %, výberové zisťovania dosahujú tento ukazovateľ podstatne nižší. Len niečo viac ako 1 % obyvateľov v oboch prípadoch zisťovania uvádza, že sa narodili mimo územia Slovenska. Tento nižší podiel je spôsobený reprezentatívnym náhodným zisťovaním, pri ktorom sa systematicky

uvádzajú nižšie percento ľudí pochádzajúcich mimo Slovenska. Z metodologického hľadiska tento rozdielny podiel nemá zásadný vplyv na mikro simulácie, nakoľko predpoklady o migrácií pochádzajú z externého zdroja, prognózy Europop. Využitie informácií o migrantoch sa v modeli využíva len na kopírovaní jednotlivých znakov nových imigrantov, ktorí prichádzajú na Slovensko. Zahrnutie tohto atribútu ako súčasť datasetov vychádza aj z motivácie poukázať na obmedzenia, ktoré vychádzajú z konceptu výberového zisťovania.

Graf 8: Percentuálne zastúpenie populácie, ktorá sa narodila mimo územia Slovenska



Zdroj: CENZUS 2021, LFS 2021, EÚ SILC 2021 a ŠÚ SR

Ekonomický status predstavuje jednu z kľúčových charakteristík vstupnej populácie v populačných mikrosimulačných modeloch. Zaradenie jednotlivcov medzi pracujúcich, nezamestnaných, študentov, dôchodcov alebo ekonomicky neaktívne osoby priamo ovplyvňuje modelovanie trhu práce a v pokročilejších modeloch aj príjem domácností, sociálne transfery a demografické procesy. Správna reprodukcia ekonomickej štruktúry populácie preto patrí medzi základné predpoklady spoľahlivých výsledkov mikrosimulácie.

Porovnanie troch analyzovaných databáz ukazuje, že všetky zdroje zachytávajú rovnakú základnú štruktúru ekonomickeho statusu populácie, avšak medzi jednotlivými kategóriami existujú rozdiely v ich početnom zastúpení. Najväčšiu skupinu vo všetkých databázach tvoria ekonomicky aktívne pracujúce osoby. Populačný CENZUS eviduje približne 2,52 milióna pracujúcich, zatiaľ čo LFS a EU-SILC odhadujú približne rovnakú hodnotu 2,46 milióna. Rozdiel predstavuje približne 60 osôb, čo zodpovedá približne 2 % populácie pracujúcich. Ide o relatívne malú odchýlku, ktorá naznačuje vysokú mieru konzistentnosti medzi jednotlivými zdrojmi pri zachytení dominantnej ekonomickej skupiny. Výraznejšie rozdiely sa prejavujú pri nezamestnaných osobách. LFS identifikuje približne 265 tisíc nezamestnaných, čo predstavuje najvyšší počet zo všetkých analyzovaných zdrojov. Naopak, EU-SILC eviduje približne 214 tisíc nezamestnaných, zatiaľ čo populačný CENZUS približne 232 tisíc osôb. Rozdiel medzi LFS a EU-SILC presahuje 50 tisíc osôb, čo predstavuje viac než pätinu populácie nezamestnaných. Takáto odchýlka je metodologicky očakávateľná, keďže LFS je koncipovaný primárne na meranie situácie na trhu práce a využíva medzinárodne harmonizovanú definíciu nezamestnanosti podľa metodiky Medzinárodnej organizácie práce

(ILO). EU-SILC je orientovaný predovšetkým na príjmy a životné podmienky domácností, pričom ekonomický status predstavuje jednu z viacerých sledovaných charakteristík.

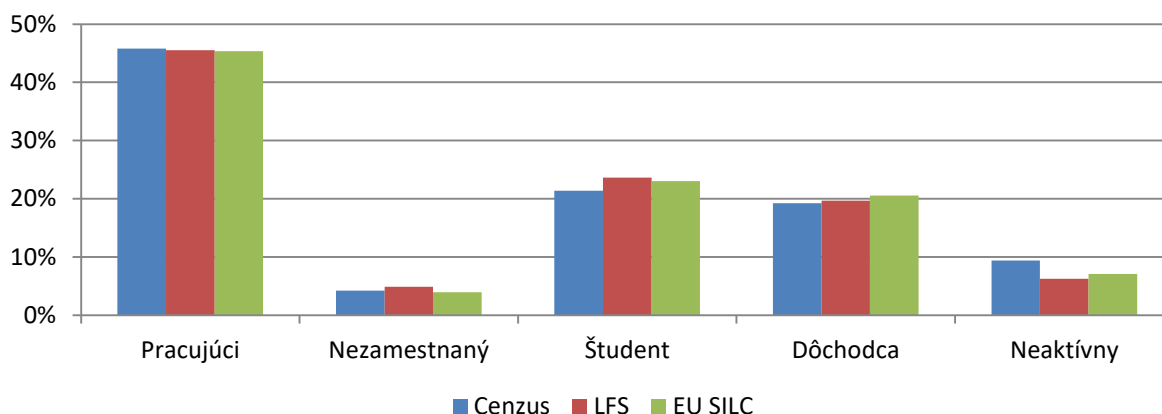
Tabuľka 1: Rozdelenie populácie podľa ekonomického statusu v osobách

	Pracujúci	Nezamestnaný	Študent	Dôchodca	Neaktívny
CENZUS	2 517 136	231 826	1 175 722	1 057 297	515 900
LFS	2 456 702	264 837	1 275 171	1 061 703	339 190
EU SILC	2 448 683	214 294	1 243 100	1 108 527	382 733

Zdroj: CENZUS 2021, LFS 2021 a EÚ SILC 2021

Rozdiely možno pozorovať aj pri populácii študentov. Najvyšší počet študentov zachytáva LFS, nasledovaný EU-SILC, zatiaľ čo CENZUS vykazuje približne o 100 tisíc nižší počet. Tieto rozdiely môžu byť dôsledkom odlišného referenčného obdobia zisťovania, spôsobu klasifikácie ekonomického statusu alebo rozdielov vo vymedzení hlavnej ekonomickej aktivity respondentov. V neposlednom rade je rozdiel spôsobený aj zaradením všetkých mladých ročníkov medzi študentov, čo v prípade rozdielnej vekovej štruktúry predovšetkým u mladších kohort zvýrazňuje rozdiely medzi datasetmi. Pri populácii dôchodcov vykazujú všetky tri databázy pomerne vysokú mieru zhody. Počet dôchodcov sa pohybuje približne medzi 1,06 a 1,11 milióna osôb, pričom najvyšší počet eviduje EU-SILC a najnižší CENZUS. Rozdiely sú relatívne malé a pravdepodobne nebudú predstavovať významný zdroj variability populačných projekcií. Keďže dôchodkový status je vo väčšine prípadov jednoznačne definovaný, možno očakávať vyššiu konzistentnosť medzi jednotlivými zdrojmi údajov.

Graf 9: Percentuálne rozdelenie populácie podľa ekonomického statusu



Zdroj: CENZUS 2021, LFS 2021 a EÚ SILC 2021

Najväčšie relatívne rozdiely možno identifikovať pri ekonomicky neaktívnej populácii. CENZUS eviduje približne 516 tisíc ekonomicky neaktívnych osôb, zatiaľ čo LFS približne 339 tisíc a EU-SILC približne 383 tisíc. Rozdiel medzi CENZUSom a výberovými zisťovaniami presahuje 130 tisíc osôb. Táto skutočnosť pravdepodobne súvisí s odlišným spôsobom klasifikácie respondentov, keďže ekonomická neaktivita predstavuje reziduálnu kategóriu zahŕňajúcu osoby, ktoré nie sú ani pracujúce, ani nezamestnané, študenti alebo dôchodcovia. Rozdielne metodické prístupy k zarade-

niu osôb na rodičovskej dovolenke, osôb v domácnosti alebo dlhodobo neaktívnych môžu viesť k výraznejším rozdielom medzi jednotlivými databázami.

Z pohľadu populačného mikrosimulačného modelovania majú uvedené rozdiely odlišný význam. Relatívne malé rozdiely v počte pracujúcich alebo dôchodcov pravdepodobne nebudú zásadne ovplyvňovať dlhodobé projekcie populácie. Naopak, výraznejšie rozdiely v počte nezamestnaných, študentov a ekonomicky neaktívnych osôb môžu ovplyvniť modelovanie prechodov medzi jednotlivými ekonomickými stavmi, najmä keď ekonomický status vstupuje do modelu. V prípade statusu neaktívny je to jeden z kľúčových parametrov, ktorý mikrosimulačný model zachováva ako percentuálny podiel populácie v prognózovanom období. Zistené rozdiely naznačujú, že ekonomický status predstavuje jednu z charakteristík, pri ktorých sa metodické rozdiely medzi vstupnými databázami prejavujú výraznejšie než pri pohlaví alebo základnej vekovej štruktúre. Ekonomický status vstupuje do výpočtu pravdepodobností socioekonomických udalostí, čo znamená, že odlišná počiatočná štruktúra ekonomickej aktivity sa premietne aj do výsledkov mikrosimulačných scenárov. Empirická časť štúdie preto overí, do akej miery možno tieto rozdiely eliminovať kalibráciou vstupných údajov a do akej miery predstavujú inherentnú vlastnosť jednotlivých dátových zdrojov.

Výsledky porovnania zároveň naznačujú, že rozdiely medzi databázami nevyplývajú iba z výberovej variability, ale aj z odlišného účelu jednotlivých štatistických zisťovaní. LFS poskytuje najdetailnejší obraz trhu práce, EU-SILC lepšie zachytáva sociálno-ekonomickú situáciu domácností a CENZUS ponúka najkomplexnejšie pokrytie celej populácie. Výber vhodného dátového zdroja preto závisí od analytického zamerania mikrosimulačného modelu a od významu, ktorý ekonomický status zohráva pri simulovaných procesoch.

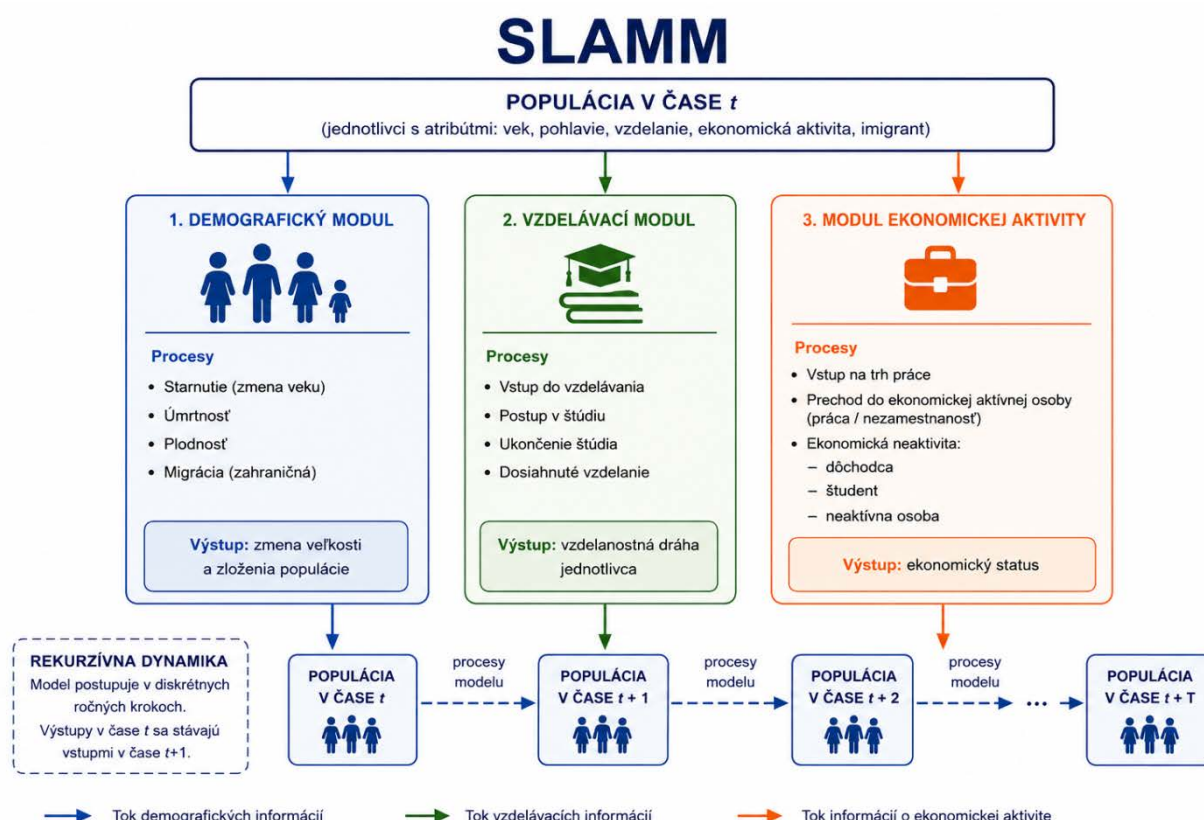
4.4. Mikrosimulačný model SLAMM

Empirická časť využíva populačný mikrosimulačný model SLAMM vytvorený na individuálnej úrovni, v ktorom základnú jednotku predstavuje osoba. Model simuluje vývoj populácie prostredníctvom aplikácie pravdepodobnostných procesov reprezentujúcich základné demografické udalosti. Každá jednotka je charakterizovaná súborom individuálnych atribútov, ktoré určujú pravdepodobnosť jednotlivých životných udalostí počas simulovaného obdobia. Model je determinovaný identickou štruktúrou vo všetkých experimentálnych scenároch. Výsledné rozdiely medzi jednotlivými scenármi preto možno pripísať výlučne odlišným charakteristikám vstupných mikroúdajov. V každej simulácii je použitý identický mikrosimulačný model, pričom sa mení iba zdroj vstupných údajov. Prvý scenár využíva údaje z LFS 2021, druhý scenár databázu EU-SILC 2021 a tretí scenár vychádza z údajov CENZUS 2021. Výber týchto troch databáz vychádza z ich dominantného postavenia v európskej oficiálnej štatistike a z ich odlišného informačného obsahu.

Mikrosimulačný model SLAMM predstavuje rekurzívne dynamický model, ktorý obsahuje v tomto prípade 3 moduly: demografický, vzdelávací a ekonomickej aktivity. V demografickom module sa

simulujú procesy starnutie, narodenie, úmrtie a migrácia. Vo vzdelávacom module sa simuluje proces dosiahnutia najvyššieho vzdelania a dĺžka samotného štúdia. V poslednom module ekonomickej aktivity sa simulujú procesy odchodu do dôchodku, predčasného dôchodku, pracujúceho dôchodcu a stav neaktivity. Všetky tieto procesy sa simulujú rekurzívne, znamená to, že prebiehajú následne v čase t a potom sa znova opakujú v čase $t+1$. Časový horizont simulácie začína v roku 2022 a končí v roku 2070. Model je skonštruovaný v softvéri LIAM2, ktorý je určený na mikrosimulačné modelovanie. Viac o modeli SLAMM je možné nájsť aj v predchádzajúcich publikáciách (Štefánik & Miklošovič; 2020). Schéma mikrosimulačného modelu SLAMM je znázornená na nasledujúcom obrázku.

Obrázok 1: Schéma mikrosimulačného modelu SLAMM



Zdroj: Autor

Okrem vstupných databáz (LFS, EU-SILC a CENZUS) do modelu vstupujú aj ďalšie exogénne parametre, ktoré vychádzajú priamo z týchto databáz. Medzi tieto parametre zahrňujeme:

- pravdepodobnostné rozdelenie najvyššieho dosiahnutého vzdelania
- podiel neaktívnych obyvateľov v jednotlivých vekových kohortách
- podiel predčasných dôchodcov v jednotlivých vekových kohortách
- podiel študentov po dovŕšení štandardnej dĺžky štúdia
- podiel pracujúcich dôchodcov

Pravdepodobnostné rozdelenie najvyššieho dosiahnutého vzdelania

Jeden z kľúčových prvkov v rámci vzdelávacieho modulu v mikrosimulačnom modeli SLAMM je simulovanie dosiahnutia najvyššieho vzdelania. Študentom je priradený najvyššie dosiahnutí stupeň vzdelania počas štúdia, pričom tento najvyšší stupeň je priradený náhodne tak, aby podiel jednotlivých stupňov vzdelania kopíroval údaje zo vstupnej databázy. Tento proces je nastavený na zachovaní vzdelanostnej štruktúry obyvateľstva vo vekovej kohorte od 30 do 35 rokov vrátane, ktorá pochádza zo vstupných dát. Z toho vyplýva, že samotná vstupná databáza určuje vzdelanostnú štruktúru obyvateľstva v simulovanom období.

Podiel neaktívnych obyvateľov v jednotlivých vekových kohortách

Pri simulovaní obyvateľstva so statusom neaktívny na trhu práce, mikrosimulačný model zachováva podiel s týmto statusom pre každú vekovú kohortu. Znamená to, že vstupná databáza priamo ovplyvňuje veľkosť neaktívneho obyvateľstva v prognózovanom období. Status neaktívny môže jedinec nadobudnúť od skončenia školy po dosiahnutí dôchodkového veku. Tento parameter je kalibrovaný špecificky pre obe pohlavia.

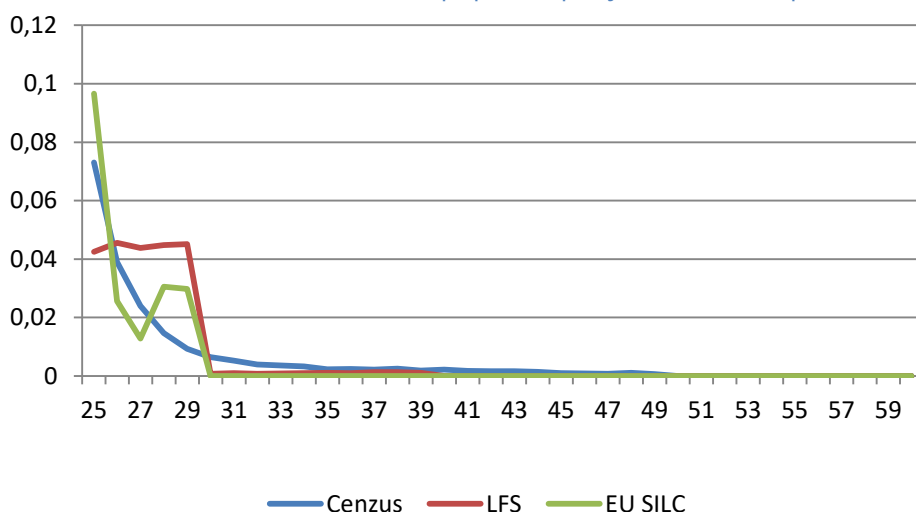
Podiel predčasných dôchodcov v jednotlivých vekových kohortách

Prognózovanie predčasných dôchodcov taktiež vychádza z pozorovaných situácií, ktoré pochádzajú zo vstupnej databázy. Jedinec sa môže stať predčasným dôchodcom ešte pred dovŕšením dôchodkového veku. Po dovŕšení dôchodkového veku sa z neho stáva starobný dôchodca. Status predčasného dôchodcu môže jedinec získať už po skončení školy, avšak tento úkaz nie je pozorovaný v žiadnej zo vstupných databáz. Tento parameter je kalibrovaný špecificky pre obe pohlavia.

Podiel študentov po dovŕšení štandardnej dĺžky štúdia

V rámci simulácie má jedinec status študent počas celého obdobia štúdia. Ukončenie tohto obdobia je parametricky definované na základe stupňa vzdelania. Avšak aj po presiahnutí tohto veku môže mať časť populácie status študenta. Počet študentov v nadštandardnej dĺžke štúdie sa určuje na základe podielov študentov v rámci jednotlivých vekových skupín do dovŕšenia dôchodkového veku. Tento podiel je určený na základe pozorovaných udalostí zo vstupnej databázy. Tento parameter je kalibrovaný špecificky pre obe pohlavia.

Graf 10: Podiel študentov (mužov) v populácii pre jednotlivé vstupné databázy



Zdroj: autor

Podiel pracujúcich dôchodcov

Aj po dosiahnutí dôchodkového veku môže časť populácie mať status pracujúci, resp. patriť do skupiny ekonomicky aktívneho obyvateľstva. Status pracujúci po dosiahnutí dôchodkového veku je možné získať na základe pozorovaných skutočností zo vstupnej databázy. Vypočítaný podiel špecificky pre jednotlivé vekové kategórie sa zachováva aj prognózovanom období. Tento parameter je kalibrovaný pre obe pohlavia.

Nezávisle od použitého zdroja vstupných mikroúdajov zostávajú všetky ostatné komponenty mikrosimulačného modelu identické. Vo všetkých experimentálnych scenároch sú použité rovnaké algoritmy simulácie, ktoré zabezpečujú jednotný spôsob modelovania demografických, vzdelávacích a ekonomických procesov. Zachovanie identických algoritmov eliminuje možnosť, aby boli prípadné rozdiely vo výsledkoch spôsobené odlišnou implementáciou modelu. Vo všetkých simuláciách je zároveň použitý rovnaký časový horizont projekcie. Model preto simuluje vývoj populácie počas identického obdobia a pri rovnakom počte simulačných krokov, čo umožňuje priame porovnanie výsledkov jednotlivých scenárov v rovnakých časových bodoch. Posledný rok simulácie zostáva nastavený na 2070. Nezmenená zostáva aj definícia všetkých modelovaných procesov. Okrem skôr uvedených parametrov pochádzajúcich zo vstupných databáz, pravdepodobnosti úmrtnosti, plodnosti, a migrácie sú vo všetkých scenároch definované rovnakým spôsobom a vychádzajú z rovnakých metodických predpokladov. Tým sa zabezpečuje, že rozdiely vo výsledkoch nemožno pripísať odlišnej špecifikácii modelových procesov.

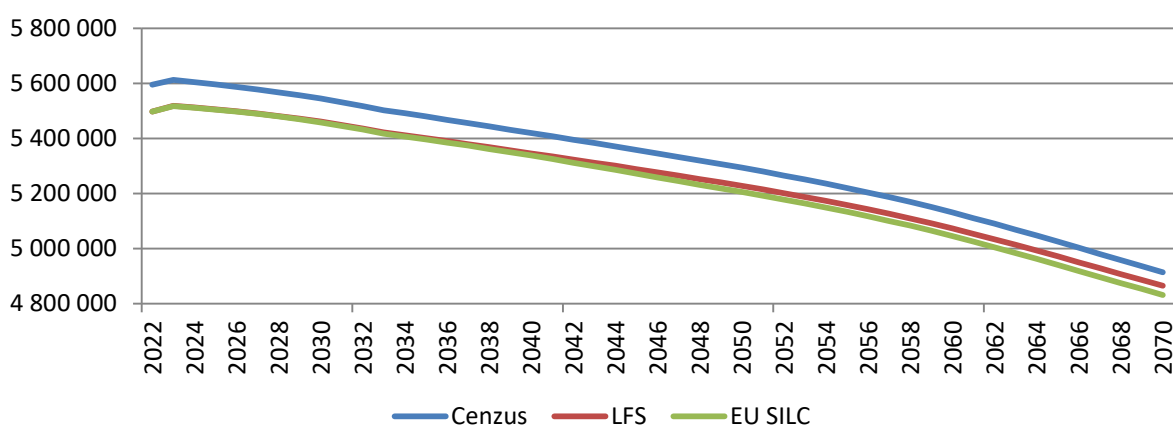
5. Hodnotenie výsledkov

Prvým hodnoteným ukazovateľom experimentu bol vývoj celkového počtu obyvateľov v období rokov 2022 až 2070 pri použití troch rôznych zdrojov vstupných mikroúdajov. Výsledky ukazujú,

že všetky tri scenáre vedú k rovnakému základnému trendu populačného vývoja. Model vo všetkých prípadoch predpokladá postupný pokles počtu obyvateľov Slovenska počas celého sledovaného obdobia. Tento vývoj je v súlade s dlhodobými demografickými prognózami Slovenskej republiky a odráža kombinovaný účinok nízkej plodnosti, starnutia populácie a migračnej bilancie.

Napriek zhodnému trendu sa medzi jednotlivými scenármi prejavujú rozdiely v absolútnej úrovni projekcie. Najvyšší počet obyvateľov počas celého simulačného obdobia dosahuje scenár založený na údajoch CENZUSu. Naopak, najnižšie hodnoty produkuje scenár využívajúci údaje EU-SILC, pričom výsledky založené na LFS sa nachádzajú medzi týmito dvoma krajnými scenármi. Charakteristické je, že poradie jednotlivých scenárov zostáva počas celej projekcie prakticky nezmenené a rozdiely medzi nimi nevykazujú výraznú tendenciu k zblížovaniu ani k ďalšiemu zväčšovaniu. Takýto priebeh naznačuje, že voľba vstupnej databázy ovplyvňuje predovšetkým počiatočný stav modelovanej populácie, zatiaľ čo samotná dynamika populačného vývoja zostáva vzhľadom na identickú špecifikáciu modelu veľmi podobná. Keďže všetky simulačné algoritmy, parametre modelu aj pravdepodobnosti jednotlivých demografických udalostí boli vo všetkých scenároch totožné, rozdiely vo výsledkoch možno pripísať predovšetkým rozdielnej veľkosti a štruktúre vstupných mikroúdajov. Tieto zistenia potvrdzujú, že aj relatívne malé rozdiely vo vstupnej populácii identifikované v predchádzajúcej analýze sa prenášajú do dlhodobých populačných projekcií. Hoci sa rozdiely medzi scenármi pohybujú iba v rozsahu niekoľkých desiatok tisíc obyvateľov, pri dlhodobých simuláciách zostávajú zachované počas celého modelovaného obdobia.

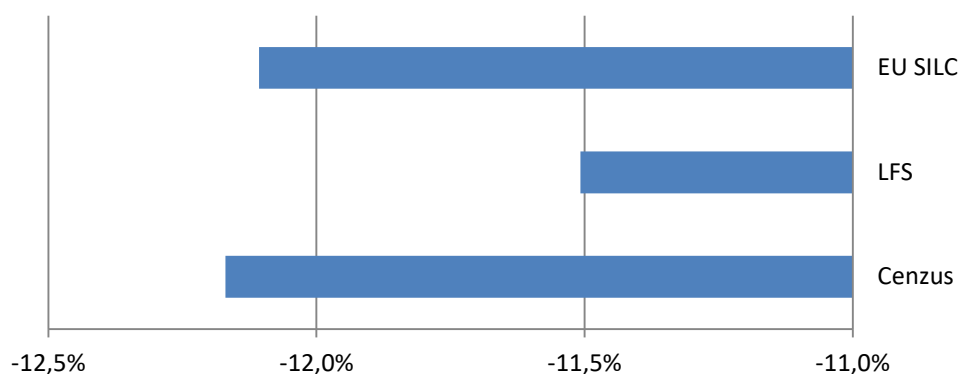
Graf 11: Predikcia počtu obyvateľstva Slovenska



Zdroj: autor

Pre lepšie porovnanie výsledkov bola analyzovaná aj relatívna zmena celkového počtu obyvateľov medzi rokmi 2022 a 2070. Výsledky ukazujú, že všetky tri scenáre predpokladajú veľmi podobnú intenzitu populačného poklesu. Najmenší relatívny pokles populácie bol zaznamenaný pri scenári využívajúcom údaje LFS, kde celkový počet obyvateľov klesol približne o 11,5 %. Scenár založený na EU-SILC vykazuje pokles približne 12,1 %, zatiaľ čo pri použití údajov CENZUS dosahuje pokles približne 12,2 %. Rozdiel medzi najvyšším a najnižším odhadom predstavuje menej než jeden percentuálny bod.

Graf 12: Percentuálny vývoj obyvateľstva Slovenska medzi rokmi 2022-2070ň



Zdroj: autor

Tieto výsledky naznačujú, že výber vstupných mikroúdajov má na odhad celkového populačného poklesu iba obmedzený vplyv. Hoci jednotlivé databázy vykazujú rozdiely vo veľkosti vstupnej populácie aj vo vybraných demografických charakteristikách, základný smer populačného vývoja zostáva vo všetkých scenároch prakticky identický. Mikrosimulačný model preto vykazuje vysokú robustnosť vo vzťahu k voľbe vstupného dátového zdroja pri hodnotení agregovaného populačného vývoja.

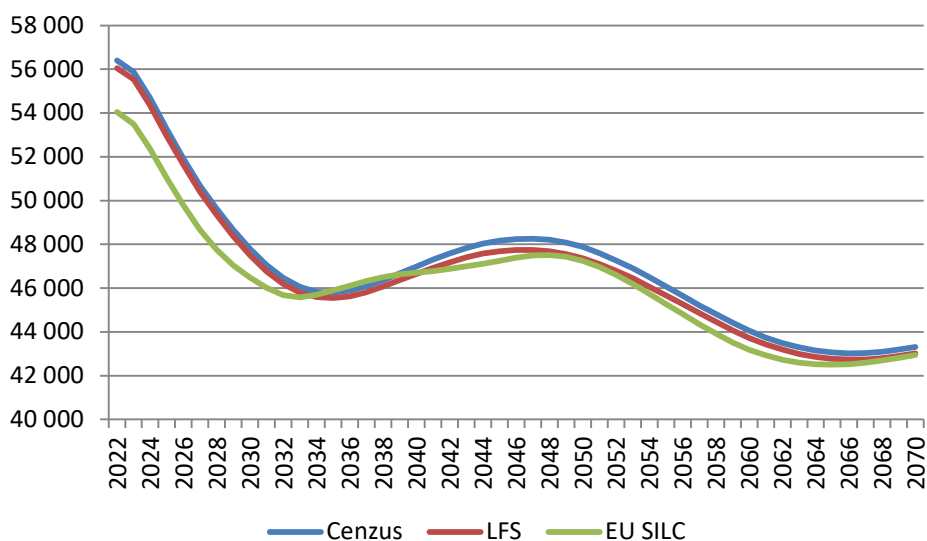
Počet narodených detí predstavuje jeden z dôležitejších výstupov populačného mikrosimulačného modelu, keďže priamo ovplyvňuje budúci vývoj populačnej veľkosti aj vekovej štruktúry. Na rozdiel od agregovaného počtu obyvateľov je počet narodených detí výsledkom individuálnych rozhodnutí simulovaných na úrovni jednotlivcov, pričom je determinovaný najmä vekovou štruktúrou ženskej populácie, jej reprodukčným správaním a pravdepodobnosťami plodnosti implementovanými v modeli.

Hoci všetky scenáre vykazujú prakticky identickú dynamiku vývoja, medzi jednotlivými vstupnými databázami existujú systematické rozdiely v absolútnom počte narodených detí. Najvyšší počet narodených počas celého simulačného obdobia produkuje scenár založený na údajoch CENZUS. Naopak, najnižšie hodnoty sú vo väčšine rokov zaznamenané pri použití údajov EU-SILC, zatiaľ čo výsledky založené na LFS sa nachádzajú medzi oboma krajnými scenármi. Rozdiely medzi scenármi sú však relatívne malé a počas celej projekcie si zachovávajú podobnú veľkosť. Keďže vo všetkých experimentálnych scenároch boli použité identické vekovo špecifické miery plodnosti aj rovnaké simulačné algoritmy, nemožno rozdiely vo výsledkoch pripísať odlišnej implementácii modelu. Ich zdrojom je predovšetkým rozdielna veková a demografická štruktúra žien reprodukčného veku vo vstupných mikroúdajoch. Aj malé rozdiely v zastúpení jednotlivých vekových kohort (pozri graf 4.) sa pri simulácii plodnosti premietajú do rozdielného počtu narodených detí.

Výsledky zároveň naznačujú vysokú robustnosť mikrosimulačného modelu. Napriek rozdielom vo vstupných databázach zostáva dlhodobý trend vývoja pôrodnosti vo všetkých scenároch prakticky identický. Voľba vstupného dátového zdroja preto nemení smer populačného vývoja, ale ovplyvňuje predovšetkým jeho absolútnu úroveň. Tento záver je v súlade s výsledkami predchá-

dzajúcej analýzy celkového počtu obyvateľov, podľa ktorej rozdiely medzi databázami zostávajú počas celej projekcie relatívne stabilné.

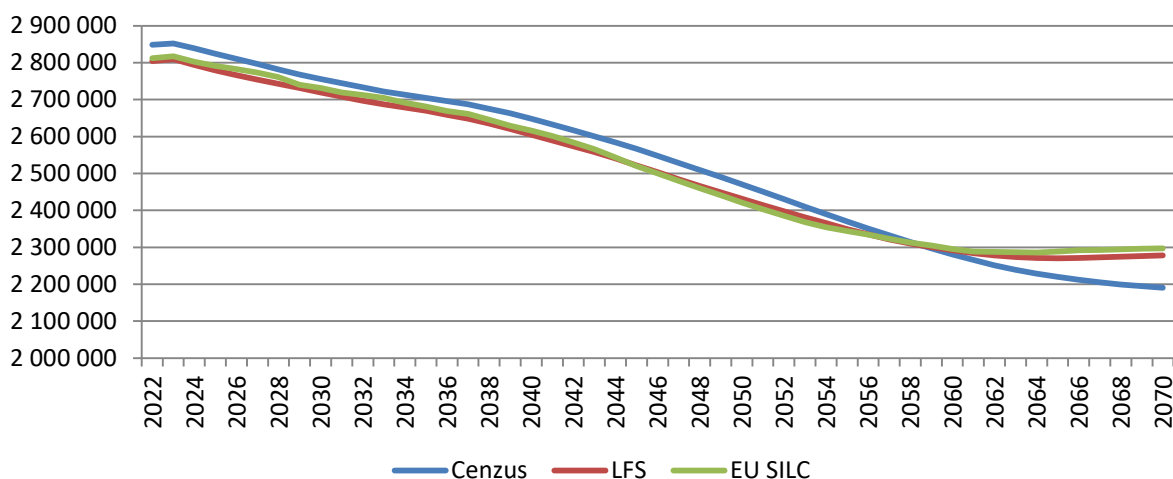
Graf 13: Predikcia počtu narodených detí na Slovensku



Zdroj: autor

Ako ďalší z ukazovateľov, ktorý budeme analyzovať v rámci nášho experimentu je predikcia ekonomicky aktívneho obyvateľstva. Na prvý pohľad by sa mohlo zdať, že všetky tri scenáre vedú k veľmi podobným výsledkom, keďže projekcie ekonomicky aktívneho obyvateľstva vykazujú porovnateľný dlhodobý trend poklesu. Podrobnejšia analýza vekovej štruktúry ekonomicky aktívnej populácie však ukazuje, že medzi jednotlivými scenármi existujú významné rozdiely, ktoré nie sú zrejmé z pohľadu samotných absolútnych početností.

Graf 14: Projekcia ekonomicky aktívneho obyvateľstva

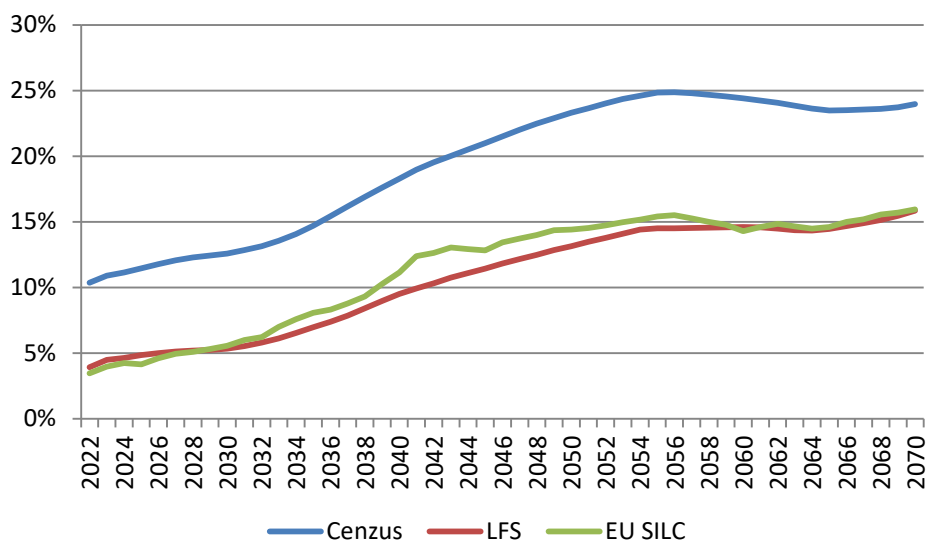


Zdroj: autor

Vývoj podielu ekonomicky aktívnych osôb vo veku nad 60 rokov poukazuje na výrazne odlišné starnutie pracovnej sily v jednotlivých scenároch. Zatiaľ čo pri simulácii založenej na CENZUSE podiel ekonomicky aktívnych osôb vo veku 61 a viac rokov postupne rastie z približne 10 % na

takmer 25 %, projekcie založené na údajoch LFS a EU-SILC dosahujú v závere simulácie iba približne 16 %. Rozdiel medzi scenármi tak predstavuje približne deväť percentuálnych bodov, čo je z metodologického hľadiska výrazne významnejšia odchýlka než rozdiely pozorované pri projekcii celkového počtu ekonomicky aktívnych osôb.

Graf 15: Predikcia podielu ekonomicky aktívneho obyvateľstva vo veku viac ako 60 rokov



Zdroj: autor

Tieto výsledky naznačujú, že voľba vstupnej databázy ovplyvňuje nielen veľkosť ekonomicky aktívnej populácie, ale predovšetkým jej vekové zloženie. Vyšší podiel ekonomicky aktívnych osôb vo vyššom veku pri scenári založenom na populačnom CENZUSE odráža odlišnú počiatočnú vekovú štruktúru ekonomicky aktívneho obyvateľstva, ktorá sa počas rekurzívnej simulácie postupne prenáša do ďalších simulačných krokov. Keďže model pracuje v diskrétnych ročných intervaloch a výstupy z jedného roku predstavujú vstupy pre nasledujúci rok, aj relatívne malé rozdiely vo vstupnej populácii sa v priebehu dlhodobých projekcií kumulujú. Výsledky zároveň potvrdzujú, že agregované ukazovatele môžu zakrývať podstatné rozdiely vo vnútornej štruktúre modelovanej populácie. Hoci všetky tri scenáre produkujú veľmi podobný počet ekonomicky aktívnych osôb, ich vekové zloženie sa postupne diverzifikuje. Z pohľadu hodnotenia trhu práce, udržateľnosti dôchodkového systému alebo budúcej ponuky pracovnej sily je pritom veková štruktúra ekonomicky aktívnej populácie často dôležitejšia než jej celková početnosť.

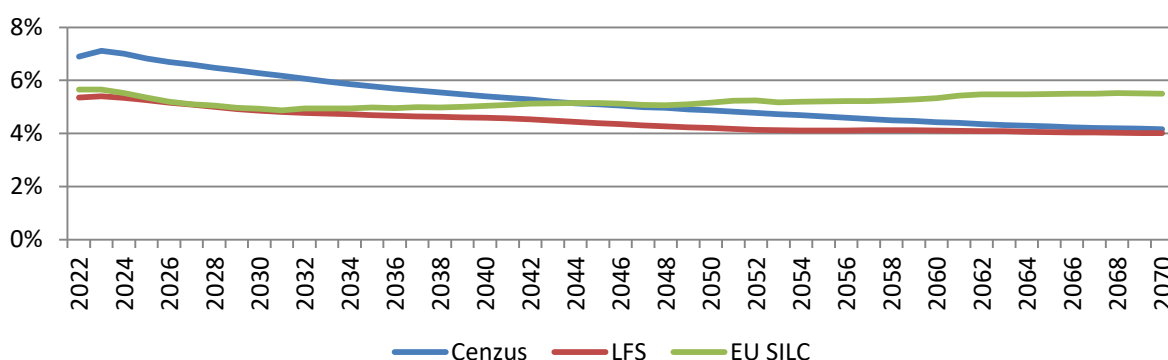
Predložený experiment preto potvrdzuje hlavnú hypotézu štúdie, že citlivosť mikrosimulačného modelu na voľbu vstupných mikroúdajov závisí od typu analyzovaného výstupného ukazovateľa. Kým pri projekcii celkového počtu obyvateľov alebo ekonomicky aktívnych osôb zostávajú rozdiely medzi jednotlivými scenármi relatívne malé, pri ukazovateľoch opisujúcich vnútornú štruktúru populácie sa rozdiely medzi vstupnými databázami výrazne zosilňujú. Z metodologického hľadiska to znamená, že pri populačných mikrosimulačných modeloch nemožno kvalitu vstupných údajov hodnotiť iba podľa schopnosti reprodukovat agregované populačné ukazovatele. Rovnako

dôležitá je schopnosť databázy zachytiť vekovú, vzdelanostnú a ekonomickú štruktúru populácie, pretože práve tieto charakteristiky určujú dlhodobý vývoj modelovaných procesov.

Na posúdenie vplyvu vstupných mikroúdajov na kvalitatívnu štruktúru pracovnej sily bola analyzovaná taktiež projekcia ekonomicky aktívneho obyvateľstva podľa najvyššieho dosiahnutého vzdelania. Výsledky sú prezentované ako podiel jednotlivých vzdelanostných skupín na celkovom počte ekonomicky aktívnych osôb v jednotlivých rokoch simulácie. Takýto prístup umožňuje eliminovať vplyv rozdielnej veľkosti ekonomicky aktívnej populácie a sústrediť sa výlučne na zmeny jej vzdelanostného zloženia.

Osoby so základným vzdelaním

Graf 16: Podiel ekonomicky aktívneho obyvateľstva so základným vzdelaním

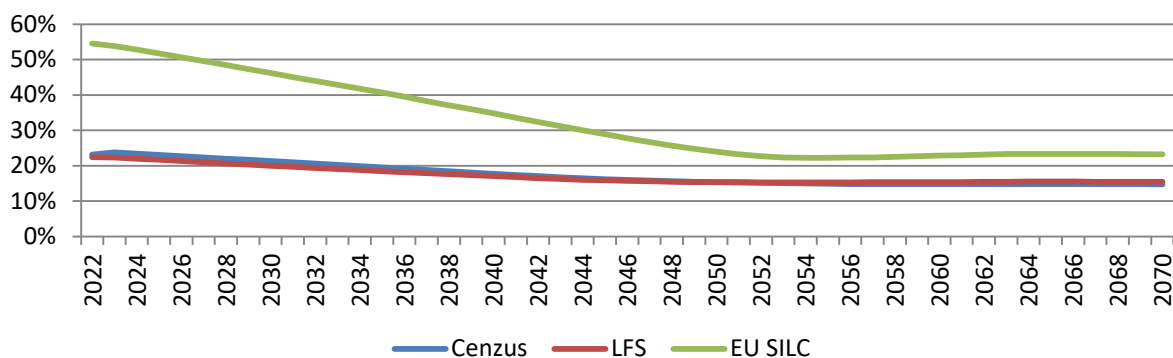


Zdroj: autor

Vo všetkých troch scenároch dochádza k dlhodobému poklesu podielu ekonomicky aktívnych osôb so základným vzdelaním. Tento vývoj odráža generačnú obmenu populácie, keď staršie kohorty s nižším dosiahnutým vzdelaním postupne opúšťajú trh práce a sú nahrádzané mladšími generáciami s vyššou úrovňou vzdelania. Tieto rozdiely naznačujú, že zastúpenie osôb s najnižším stupňom vzdelania je citlivé na počiatočnú vzdelanostnú štruktúru vstupnej populácie a na spôsob, akým jednotlivé databázy reprezentujú ekonomicky aktívne osoby s nízkou kvalifikáciou.

Osoby so stredným vzdelaním bez maturity

Graf 17: Podiel ekonomicky aktívneho obyvateľstva so stredným vzdelaním bez maturity



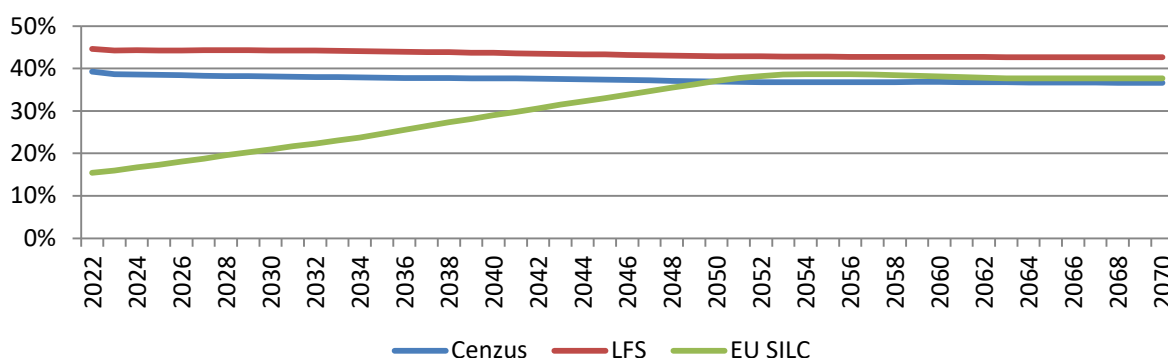
Zdroj: autor

Najvýraznejšie rozdiely medzi jednotlivými scenármi možno identifikovať pri ekonomicky aktívnych osobách so stredným odborným vzdelaním bez maturity. Scenáre založené na populačnom CENZUse a LFS vykazujú veľmi podobný vývoj, pričom podiel tejto skupiny postupne klesá z približne 23 % na približne 15 % ekonomicky aktívnej populácie. Naopak, scenár využívajúci údaje EU-SILC sa od ostatných dvoch výrazne odlišuje už od začiatku simulácie (pozri graf 6). Počiatočný podiel tejto skupiny presahuje 50 % ekonomicky aktívneho obyvateľstva a počas prvých troch desaťročí simulácie dochádza k jeho postupnému poklesu približne na 22 %. Následne sa podiel stabilizuje na úrovni približne 23 %. Takto výrazný rozdiel možno vysvetliť samotnou dynamikou modelu, keďže všetky scenáre využívajú identické simulačné algoritmy, avšak simulujú dosiahnuté vzdelanie u mladšej generácie parametricky na základe pozorovaných hodnôt (pozri graf 7).

Osoby so stredným vzdelaním s maturitou

Odlišný vývoj možno pozorovať pri ekonomicky aktívnych osobách so stredoškolským vzdelaním ukončeným maturitou. V scenároch založených na populačnom CENZUse a LFS zostáva podiel tejto skupiny počas celej simulácie relatívne stabilný. V prípade cenзовých údajov sa pohybuje približne medzi 36 % a 39 %, zatiaľ čo LFS dlhodobo vykazuje mierne vyššie zastúpenie približne na úrovni 43 %.

Graf 18: Podiel ekonomicky aktívneho obyvateľstva so stredným vzdelaním s maturitou



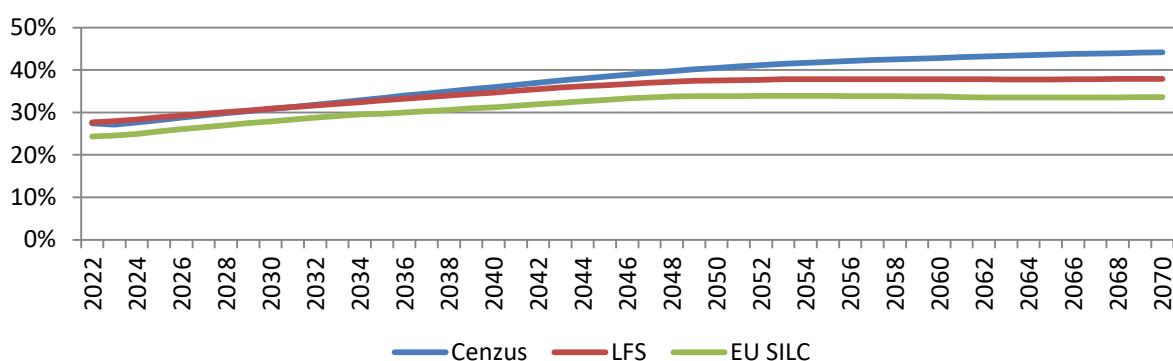
Zdroj: autor

Výrazne odlišný priebeh opäť vykazuje scenár založený na údajoch EU-SILC. Počiatočný podiel ekonomicky aktívnych osôb s maturitou predstavuje iba približne 15 %, avšak počas celej simulácie kontinuálne rastie a približne v polovici projekčného obdobia dosahuje úroveň porovnateľnú s ostatnými scenármi. V závere simulácie sa stabilizuje približne na úrovni 38 %. Vývoj tejto kategórie je do značnej miery komplementárny s predchádzajúcou skupinou osôb so stredným vzdelaním bez maturity. Pokles jednej skupiny je sprevádzaný rastom druhej, čo naznačuje, že rozdiely medzi databázami vyplývajú najmä z odlišného rozdelenia ekonomicky aktívnych osôb medzi tieto dve vzdelanostné kategórie.

Osoby s vysokoškolským vzdelaním

Vo všetkých troch scenároch dochádza k postupnému rastu podielu ekonomicky aktívnych osôb s vysokoškolským vzdelaním. Tento trend odráža pokračujúci rast vzdelanostnej úrovne populácie a vyššie zastúpenie vysokoškolského vzdelania v mladších generáciách vstupujúcich na trh práce.

Graf 19: Podiel ekonomicky aktívneho obyvateľstva s vysokoškolským vzdelaním



Zdroj: autor

Najvyšší podiel vysokoškolsky vzdelaných ekonomicky aktívnych osôb vykazuje počas celej simulácie scenár založený na populačnom CENZUSE. Podiel tejto skupiny rastie približne z 27 % na viac ako 44 % ekonomicky aktívnej populácie. LFS vykazuje podobný trend, avšak s nižšou intenzitou rastu, pričom v závere simulácie dosahuje približne 38 %. Najnižšie hodnoty sú zaznamenané pri databáze EU-SILC, kde sa podiel vysokoškolsky vzdelaných ekonomicky aktívnych osôb stabilizuje približne na úrovni jednej tretiny ekonomicky aktívnej populácie. Rozdiely medzi scenármi sa počas simulácie postupne zväčšujú, čo naznačuje, že vstupná vzdelanostná štruktúra významne ovplyvňuje dlhodobý vývoj kvalifikačného zloženia pracovnej sily.

Výsledky ukazujú, že vzdelanostná štruktúra ekonomicky aktívneho obyvateľstva patrí medzi ukazovatele s najvyššou citlivosťou na voľbu vstupných mikroúdajov. Na rozdiel od projekcie celkového počtu obyvateľov alebo celkového počtu ekonomicky aktívnych osôb sa rozdiely medzi jednotlivými databázami prejavujú výrazne už v prvých rokoch simulácie a v niektorých vzdelanostných skupinách sa počas projekcie ďalej kumulujú. Tieto výsledky naznačujú, že pri populačných mikrosimulačných modeloch zameraných na analýzu trhu práce alebo ľudského kapitálu zohráva voľba vstupného dátového zdroja podstatne významnejšiu úlohu než pri projekcii agregovaných populačných ukazovateľov. Vzdelanostná štruktúra pracovnej sily je výsledkom dlhodobého pôsobenia demografických a vzdelávacích procesov, a preto citlivejšie reaguje na rozdiely vo vstupných mikroúdajoch než samotná veľkosť ekonomicky aktívnej populácie.

6. Diskusia

Predložená štúdia sa zamerala na experimentálne posúdenie vplyvu rôznych zdrojov oficiálnych mikroúdajov na výsledky populačného mikrosimulačného modelovania. Na rozdiel od väčšiny publikovaných prác, ktoré analyzujú vplyv modelových predpokladov alebo simulačných algoritmov, bol v tomto prípade experiment navrhnutý tak, aby jedinou systematicky menenou pre-

mennou bol zdroj vstupných mikroúdajov. Takýto prístup umožnil izolovať vplyv vstupnej databázy od ostatných komponentov modelu a experimentálne posúdiť citlivosť výsledkov na voľbu dátového základu.

Robustnosť agregovaných populačných ukazovateľov

Prvým významným zistením je vysoká robustnosť modelu pri projekcii agregovaných populačných ukazovateľov. Bez ohľadu na to, či boli ako vstup použité údaje populačného CENZUSu, LFS alebo EU-SILC, všetky scenáre viedli k veľmi podobnému vývoju celkového počtu obyvateľov aj počtu narodených detí. Rozdiely medzi jednotlivými projekciami sa prejavovali predovšetkým v absolútnej úrovni ukazovateľov, zatiaľ čo ich dlhodobý trend zostával prakticky identický.

Tento výsledok naznačuje, že pri populačných mikrosimulačných modeloch, ktorých cieľom je predovšetkým projekcia celkovej veľkosti populácie, nemusí byť výber medzi analyzovanými oficiálnymi dátovými zdrojmi rozhodujúcim faktorom. Pokiaľ sú vstupné databázy vhodne harmonizované a kalibrované, model vykazuje vysokú stabilitu voči rozdielom vo vstupných údajoch. Z metodologického hľadiska ide o dôležité zistenie, pretože potvrdzuje, že populačné mikrosimulačné modely dokážu pri agregovaných ukazovateľoch čiastočne kompenzovať počiatočné rozdiely medzi jednotlivými databázami prostredníctvom rekurzívnej dynamiky modelu.

Rastúca citlivosť pri štruktúrnych ukazovateľoch

Druhé zistenie predstavuje výrazne vyššia citlivosť modelu pri analýze štruktúrnych charakteristík populácie. Tá sa prejavila najmä pri projekcii ekonomicky aktívneho obyvateľstva, vekovej štruktúry pracovnej sily a jej vzdelanostného zloženia. Výsledky ukázali, že rozdiely medzi vstupnými databázami sa počas simulácie postupne kumulujú. Kým projekcie celkového počtu ekonomicky aktívnych osôb zostávali relatívne podobné, ich veková štruktúra sa začala výrazne diferencovať. Najvýraznejšie rozdiely boli zaznamenané pri podiele ekonomicky aktívnych osôb vo veku 61 rokov a viac, kde scenár založený na populačnom CENZUSE vykazoval podstatne vyššie hodnoty než scenáre využívajúce údaje LFS alebo EU-SILC. Ešte výraznejšie rozdiely boli identifikované pri vzdelanostnej štruktúre ekonomicky aktívnej populácie. Zatiaľ čo podiel osôb so základným vzdelaním vykazoval vo všetkých scenároch podobný vývoj, pri stredoškolskom a vysokoškolskom vzdelaní sa jednotlivé projekcie postupne odlišovali. Rozdiely sa prejavili najmä medzi kategóriami stredoškolského vzdelania s maturitou a bez maturity, pričom scenár založený na údajoch EU-SILC vykazoval odlišné zastúpenie týchto skupín už na začiatku simulácie. Tento výsledok naznačuje, že štruktúrne charakteristiky populácie sú podstatne citlivejšie na vlastnosti vstupných mikroúdajov než agregované populačné ukazovatele.

Význam počiatočnej štruktúry populácie

Tretím významným zistením je skutočnosť, že dlhodobý vývoj modelovaných charakteristík je výrazne ovplyvnený počiatočnou štruktúrou vstupnej populácie. Keďže použitý mikrosimulačný model pracuje rekurzívne, výstupy každého simulačného kroku predstavujú vstup pre krok na-

sledujúci. Rozdiely prítomné vo vstupnej populácii sa preto počas simulácie postupne prenášajú do ďalších rokov a môžu sa kumulatívne prejaviť vo výsledkoch aj napriek tomu, že všetky modelové algoritmy a parametre zostávajú identické. Experiment ukázal, že samotná harmonizácia vstupných databáz nedokáže úplne eliminovať rozdiely vyplývajúce z ich odlišnej populačnej štruktúry. Rozdiely v zastúpení jednotlivých vekových alebo vzdelanostných skupín sa počas simulácie transformujú do rozdielov vo vekovej štruktúre ekonomicky aktívnej populácie a následne ovplyvňujú aj ďalšie modelované procesy. Výber vstupnej databázy preto nemožno chápať iba ako technický krok pri príprave údajov, ale ako metodologické rozhodnutie, ktoré môže ovplyvniť interpretáciu výsledkov.

Metodologické implikácie

Výsledky štúdie zároveň naznačujú, že vhodnosť konkrétneho zdroja mikroúdajov závisí od typu analyzovaného problému. Ak je cieľom modelu projekcia celkovej veľkosti populácie alebo základných demografických ukazovateľov, analyzované databázy poskytujú porovnateľné výsledky. Ak však model slúži na analýzu trhu práce, ekonomickej aktivity alebo kvalifikačnej štruktúry pracovnej sily, rozdiely medzi jednotlivými zdrojmi údajov sa stávajú výrazne významnejšími. Výber vstupnej databázy by preto mal vychádzať z analytického účelu modelu a z charakteru výstupných ukazovateľov, ktoré majú byť predmetom interpretácie. Výsledky zároveň poukazujú na potrebu venovať väčšiu pozornosť validácii vstupných mikroúdajov ešte pred samotnou konštrukciou mikrosimulačného modelu. V odbornej literatúre je validácia najčastejšie zameraná na porovnanie modelových výstupov s referenčnými populačnými štatistikami. Predložený experiment však ukazuje, že rovnako dôležitá je validácia samotnej vstupnej databázy, najmä jej vekovej, vzdelanostnej a ekonomickej štruktúry. Kvalita vstupných údajov totiž ovplyvňuje nielen počiatkový stav modelu, ale prostredníctvom rekurzívnej dynamiky aj jeho dlhodobé správanie.

Limity štúdie

Pri interpretácii výsledkov je potrebné zohľadniť aj niekoľko obmedzení výskumu. Experiment bol realizovaný na troch oficiálnych štatistických zdrojoch využívaných na Slovensku a jeho závery preto nemožno bez ďalšieho overenia zovšeobecniť na administratívne registre alebo syntetické populácie. Zároveň boli vo všetkých scenároch použité identické modelové parametre a pravdepodobnostné funkcie, čo umožnilo izolovať vplyv vstupných údajov, avšak neumožňuje posúdiť interakciu medzi kvalitou dát a alternatívnymi špecifikáciami modelu. Budúci výskum by preto mohol rozšíriť experiment aj o ďalšie zdroje mikroúdajov, prípadne analyzovať kombinované využitie výberových zisťovaní a administratívnych registrov.

7. Záver

Predkladaná štúdia sa zamerala na analýzu významu vstupných mikroúdajov v populačnom mikrosimulačnom modelovaní. Hlavným cieľom bolo experimentálne posúdiť, do akej miery ovplyvňuje voľba oficiálneho zdroja mikroúdajov výsledky populačného mikrosimulačného modelu pri

zachovaní identickej modelovej štruktúry. Na rozdiel od väčšiny publikovaných prác, ktoré sa sústreďujú na vývoj simulačných algoritmov alebo validáciu modelových výstupov, bola pozornosť tejto štúdie zameraná na vstupný dátový základ modelu. Experiment bol navrhnutý ako kontrolované porovnanie troch najvýznamnejších zdrojov oficiálnych mikroúdajov – populačného CENZU-Su, LFS a EU-SILC – pri použití identického populačného mikrosimulačného modelu.

Teoretická časť štúdie ukázala, že kvalita vstupných mikroúdajov predstavuje jeden z rozhodujúcich predpokladov spoľahlivého mikrosimulačného modelovania. Súčasná odborná literatúra venuje značnú pozornosť vývoju simulačných algoritmov, harmonizácii dát a validačným postupom, avšak systematické porovnanie rôznych oficiálnych zdrojov mikroúdajov zostáva relatívne málo preskúmanou oblasťou. Predložená štúdia preto vytvorila jednotný metodologický rámec na hodnotenie vstupných databáz z pohľadu ich využiteľnosti v populačných mikrosimulačných modeloch a následne tento rámec overila prostredníctvom experimentálnej aplikácie.

Výsledky experimentu potvrdili, že voľba vstupných mikroúdajov ovplyvňuje výsledky populačného mikrosimulačného modelovania, avšak intenzita tohto vplyvu závisí od charakteru analyzovaného výstupného ukazovateľa. Pri agregovaných populačných charakteristikách, ako je celkový počet obyvateľov alebo počet narodených detí, boli rozdiely medzi jednotlivými scenármi relatívne malé a všetky tri vstupné databázy viedli k veľmi podobným dlhodobým trendom populačného vývoja. Tieto výsledky naznačujú, že pri modelovaní základných demografických procesov vykazuje populačný mikrosimulačný model vysokú mieru robustnosti voči voľbe vstupného dátového zdroja.

Odlišná situácia bola identifikovaná pri analýze štruktúrnych charakteristík populácie. Projekcie ekonomicky aktívneho obyvateľstva, vekovej štruktúry pracovnej sily a jej vzdelanostného zloženia ukázali, že rozdiely medzi jednotlivými vstupnými databázami sa počas simulácie postupne kumulujú a vedú k odlišným výsledkom. Najvýraznejšie rozdiely boli zaznamenané pri vekovom a vzdelanostnom zložení ekonomicky aktívnej populácie, kde sa prejavil význam počiatkovej štruktúry vstupných mikroúdajov. Experiment tak potvrdil, že vlastnosti vstupnej databázy môžu významne ovplyvniť projekcie ukazovateľov opisujúcich vnútornú štruktúru populácie, aj keď agregované populačné charakteristiky zostávajú veľmi podobné.

Významným metodologickým prínosom štúdie je aplikácia experimentálneho prístupu založeného na zachovaní identickej štruktúry mikrosimulačného modelu pri systematickej zmene jediného faktora – vstupného dátového zdroja. Takto navrhnutý experiment umožnil oddeliť vplyv vstupných mikroúdajov od vplyvu simulačných algoritmov a ostatných modelových predpokladov. Výsledky ukázali, že vstupné údaje nemožno chápať iba ako technický predpoklad modelovania, ale ako samostatný metodologický faktor ovplyvňujúci výsledky populačných mikrosimulácií. Na základe získaných poznatkov možno konštatovať, že neexistuje univerzálne najvhodnejší zdroj mikroúdajov pre všetky typy populačných mikrosimulačných modelov. Vhodnosť konkrétnej databázy závisí predovšetkým od analytického účelu modelu a od typu výstupných ukazovateľov. Ak je cieľom modelovania projekcia celkovej veľkosti populácie alebo základných demografických cha-

rakteristík, analyzované zdroje poskytujú porovnateľné výsledky. Pri modelovaní ekonomickej aktivity, kvalifikačnej štruktúry pracovnej sily alebo ďalších detailných charakteristík populácie však rozdiely medzi vstupnými databázami nadobúdajú podstatne väčší význam a mali by byť zohľadnené už pri návrhu modelu.

Výsledky štúdie zároveň poukazujú na potrebu venovať väčšiu pozornosť príprave a validácii vstupných mikroúdajov. Harmonizácia premenných a kalibrácia databáz predstavujú nevyhnutné kroky pri tvorbe populačných mikrosimulačných modelov, avšak samy osebe nedokážu úplne eliminovať rozdiely vyplývajúce z odlišnej populačnej štruktúry jednotlivých dátových zdrojov. Budúci výskum by sa preto mal zamerať na možnosti kombinácie oficiálnych štatistických zisťovaní s administratívnymi registrami, na tvorbu syntetických populácií a na hodnotenie ich vplyvu na výsledky populačných mikrosimulačných modelov.

Predložená štúdia vytvára metodologické východisko pre ďalší výskum v oblasti populačného mikrosimulačného modelovania. Navrhnutý experimentálny prístup možno aplikovať aj na ďalšie typy vstupných mikroúdajov a rozšíriť ho o hodnotenie administratívnych databáz, registrových údajov alebo syntetických populácií. Súčasne poskytuje praktické odporúčania pre výskumníkov a tvorcov populačných mikrosimulačných modelov, ktorí stoja pred rozhodnutím o výbere vhodného dátového základu pre svoje analytické aplikácie.

Literatúra

- Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). *SimBritain: A spatial microsimulation approach to population dynamics*. Population, Space and Place, 11(1), 13–34. DOI: <https://doi.org/10.1002/psp.351>
- Chapuis, K., Taillandier, P., & Drogoul, A. (2022). *Generation of Synthetic Populations in Social Simulations: A Review of Methods and Practices*. Journal of Artificial Societies and Social Simulation, 25(2), 6. DOI: <https://doi.org/10.18564/jasss.4762>
- Deville, J.-C., & Särndal, C.-E. (1992). *Calibration Estimators in Survey Sampling*. Journal of the American Statistical Association, 87(418), 376–382. DOI: <https://doi.org/10.1080/01621459.1992.10475217>
- Figari, F., Paulus, A., & Sutherland, H. (2015). *Microsimulation and Policy Analysis*. In Handbook of Income Distribution (Vol. 2, pp. 2141–2221). Elsevier. DOI: <https://doi.org/10.1016/B978-0-444-59429-7.00025-X>
- Harding, A. (Ed.). (1996). *Microsimulation and Public Policy*. Amsterdam: North-Holland. Dostupné na: <https://www.jasss.org/1/3/review3.html>
- Hradec, J., Craglia, M., Di Leo, M., De Nigris, S., Ostlaender, N., & Nicholson, N. (2022). *Multipurpose synthetic population for policy applications*. Publications Office of the European Union. DOI: <https://doi.org/10.2760/50072>
- Kokosi, T., De Stavola, B., Mitra, R., Frayling, L., Doherty, A., Dove, I., Sonnenberg, P., & Harron, K. (2022). *An overview of synthetic administrative data for research*. International Journal of Population Data Science. DOI: <https://doi.org/10.23889/ijpds.v7i1.1727>
- Li, J., & O'Donoghue, C. (2013). *A survey of dynamic microsimulation models: Uses, model structure and methodology*. International Journal of Microsimulation, 6(2), 3–55. DOI: <https://doi.org/10.34196/ijm.00082>
- Lovelace, R., & Dumont, M. (2016). *Spatial Microsimulation with R*. CRC Press. Dostupné na: <https://spatial-microsim-book.robinlovelace.net/>
- O'Donoghue, C., & Sologon, D. M. (2023). *The Transformation of Public Policy Analysis in Times of Crisis – A Microsimulation-Nowcasting Method Using Big Data*. IZA Discussion Paper No. 15937. Institute of Labor Economics (IZA). Dostupné na: <https://www.iza.org/publications/dp/15937/the-transformation-of-public-policy-analysis-in-times-of-crisis-a-microsimulation-nowcasting-method-using-big-data>
- Orcutt, G. H. (1957). *A new type of socio-economic system*. The Review of Economics and Statistics, 39(2), 116–123. DOI: <https://doi.org/10.2307/1928528>
- Orcutt, G. H., Merz, J., & Quinke, H. (1986). *Microanalytic Simulation Models to Support Social and Financial Policy*. Economics, Political Science
- Sutherland, H., & Figari, F. (2013). *EUROMOD: The European Union tax-benefit microsimulation model*. International Journal of Microsimulation, 6(1), 4–26. DOI: <https://doi.org/10.34196/ijm.00075>
- Štefánik, M. & Miklošovič, T. (2020). *Modelling foreign labour inflows using a dynamic microsimulation model of an ageing country - Slovakia*. In International Journal of Microsimulation, vol. 13, no. 2, pp. 102–113. DOI: <https://doi.org/10.34196/ijm.00220>
- Tanton, R., & Edwards, K. L. (Eds.). (2013). *Spatial Microsimulation: A Reference Guide for Users*. Springer. DOI: <https://doi.org/10.1007/978-94-007-4623-7>

United Nations Economic Commission for Europe. (2018). *Guidelines on the use of registers and administrative data for population and housing censuses*. United Nations. Dostupné na: <https://unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20184.pdf>

Zaidi, A., & Rake, K. (2001). *Dynamic microsimulation models: A review and some lessons for SAGE* (SAGE Discussion Paper No. 2). London School of Economics, ESRC Research Group Simulating Social Policy in an Ageing Society (SAGE). Dostupné na: https://www.researchgate.net/publication/265280929_Dynamic_Microsimulation_Models_A_Review_and_Some_Lessons_for_SAGE