MASARYKOVA
UNIVERZITA

Přírodovědecká fakulta

Diplomová práce

Eliška Křapová

BRNO 2026

MASARYKOVA
UNIVERZITA

Přírodovědecká fakulta

Možnosti tvorby syntetické populace v národním prostředí

Diplomová práce

Eliška Křapová

Vedoucí práce: prof. RNDr. Petr Kubíček, CSc.

Geografický ústav

BRNO 2026

Bibliografický záznam

Autor/Autorka:	Eliška Křapová Přírodovědecká fakulta, Masarykova univerzita Geografický ústav
Název práce:	Možnosti tvorby syntetické populace v národním prostředí
Studijní program:	PřF N-GKG Geografická kartografie a geoinformatika
Studijní obor:	PřF N-GKG Geografická kartografie a geoinformatika
Vedoucí práce:	prof. RNDr. Petr Kubíček, CSc.
Akademický rok:	2025/2026
Počet stran:	96
Klíčová slova:	Syntetická populace, Brno, IPF, Python, ČSÚ, Alokace, ArcGIS Experience Builder, AMOS

Bibliografic Entry

Author:	Eliška Křapová Faculty of Science, Masaryk University Department of Geography
Title of Thesis:	Possibilities of a synthetic population development in the national environment
Degree Programme:	PřF GKGI Geografická kartografie a geoinformatika
Field of Study:	PřF GKGI Geografická kartografie a geoinformatika
Supervisor:	prof. RNDr. Petr Kubíček, CSc.
Academic Year:	2025/2026
Number of Pages:	96
Keywords:	Synthetic population, Brno, IPF, Python, ČSÚ, Alocation, ArcGIS Experience Builder, AMOS

Abstrakt

Tato diplomová práce se zabývá metodikou tvorby a prostorového ukotvení syntetické populace v prostředí České republiky, konkrétně na území statutárního města Brna. Syntetická populace představuje stěžejní vstupní datovou vrstvu pro pokročilé mikrosimulační a agentní modely (např. v dopravním modelování), jelikož kombinuje vysokou míru demografického detailu s naprostou ochranou osobních údajů. Cílem práce bylo adaptovat stávající generovací algoritmy na specifika českých datových sad. Práce analyzuje a integruje dostupná data z národních registrů (Sčítání lidu, domů a bytů 2021, Registr sčítacích obvodů a budov) i lokálních geoportálů (data.brno.cz). Hlavní přínos spočívá v identifikaci a algoritmickém řešení datových anomálií a v úpravě procesu formování domácností tak, aby odpovídal reálným biologickým a demografickým předpokladům. Závěrečnou fází byla prostorová alokace vygenerovaných domácností na konkrétní adresní body pomocí identifikátorů RÚIAN. Výstupem práce je validovaná syntetická populace města Brna a interaktivní webová aplikace postavená v prostředí ArcGIS Experience Builder, která umožňuje detailní hierarchické prohlížení dat na úrovni budov, podlaží a jednotlivých domácností.

Abstract

This master's thesis focuses on the methodology of creating and spatially anchoring a synthetic population within the Czech Republic, specifically in the territory of the statutory city of Brno. A synthetic population represents a crucial input data layer for advanced microsimulation and agent-based models (e.g., in transport modeling), as it combines a high degree of demographic detail with the complete protection of personal data. The aim of the thesis was to adapt existing generation algorithms to the specific characteristics of Czech datasets. The work analyzes and integrates available data from national registers (the 2021 Census of Population, Houses and Dwellings, and the Register of Census Districts and Buildings) as well as local geoportals (data.brno.cz). The main contribution lies in the identification and algorithmic resolution of data anomalies, as well as in modifying the household formation process to align with realistic biological and demographic assumptions. The final phase involved spatially allocating the generated households to specific address points using RÚIAN identifiers. The thesis's output is a validated synthetic population of the city of Brno and an interactive web application built in the ArcGIS Experience Builder environment, which enables detailed, hierarchical viewing of the data at the building, floor, and individual household levels.

MUNI
SCI

Masarykova univerzita

Přírodovědecká fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: **Eliška Křapová**
Studijní program: **Geografická kartografie a geoinformatika**

Ředitel Geografického ústavu PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s tématem:

Možnosti tvorby syntetické populace v národním prostředí

Possibilities of a synthetic population development in the national environment

Zásady pro vypracování:

Snaha o využívání dat o populaci v detailním prostorové rozlišení se často dostává do konfliktu s principy ochrany osobních údajů, respektive s dostupností určitého typu informace v detailním měřítku. Uvedený problém je částečně řešitelný využíváním tzv. syntetické populace. Syntetická populace je konstrukt, který je vytvořen na základě reálných dat (primárně statistických) a empirických znalostí o skutečné populaci. Cílem práce je ověřit vybrané přístupy konstrucke syntetické populace v rámci České republiky a na zvoleném vzorku dokumentovat možnosti vizualizace a dalšího využití.
Struktura práce bude zahrnovat následující kapitoly:
1) Úvod a rešerši současného stavu problematiky.
2) Analýzu metodických přístupů k tvorbě syntetické populace ve světě a v Evropě.
3) Analýza datové základny pro tvorbu syntetické populace.
4) Návrh možností tvroby syntetické populace v ČR a jeho realizace na vybraném vzorku.
5) Diskuze výsledků a návrh možných scénářů využití.

Rozsah grafických prací: podle potřeby

Rozsah průvodní zprávy: cca 60 až 80 stran

Seznam odborné literatury:

ADIGA, A., AGASHE, A., ARIFUZZAMAN, S., BARRETT, C. L., BECKMAN, R., BISSET, K., CHEN, J., CHUNGBAEK, Y., EUBANK, S., GUPTA, S., KHAN, M., KUHLMAN, C. J., LOFGREN, E., LEWIS, B., MARATHE, A., MARATHE, M. V., MORTVEIT, H. S., NORDBERG, E., RIVERS, C., STRETZ, P., SWARUP, S., WILSON, A., XIE, D. (2015): Generating a synthetic population of the United States. https://arifuzzaman.faculty.unlv.edu/paper/synth_popu15.pdf

HRADEC, J., CRAGLIA, M., DI, L. M., DE, N. S., OSTLAENDER, N., NICHOLSON, N. (2022): Multipurpose synthetic population for policy applications, JRC Publications Repository. https://doi.org/10.2760/50072

HRADEC, J., LEO, M. D. [2024]: Synthetic populations for personalized policy. https://indico.cern.ch/event/1059494/contributions/4532812/attachments/2310904/3937091/DiLeo_HTC_workshop_21.pdf

LAATABI, A., AIT BABRAM, M., MOULAY LHASSAN, H. (2015): Generating and mapping a synthetic population of Marrakesh. doi: 10.1109/ICoCS.2015.7483306

LIN, Y., XIAO, N. (2023): Developing Synthetic Individual-level Population Datasets: the Case of Contextualizing Maps of Privacy- Preserving Census Data. https://doi.org/10.48550/arXiv.2206.04766

LIU, J., MA, X., ZHU, Y., LI, J., HE, Z., YE, S. (2021): Generating and Visualizing Spatially Disaggregated Synthetic Population Using a Web-Based Geospatial Service. Sustainability. 13, 1587. doi: 10.3390/su13031587

RUBINYI, S., VERSCHUUR, J., GOLDBLATT, R., GUSSENBAUER, J., KOWARIK, A., MANNIX, J., BOTTOMS, B., HALL, J. (2022): High-resolution synthetic population mapping for quantifying disparities in disaster impacts: An application in the Bangladesh Coastal Zone. Frontiers in Environmental Science, 10. doi: 10.3389/fenvs.2022.1033579

THOMSON, D., KOOLS, L., JOCHEM, W. (2018): Linking Synthetic Populations to Household Geolocations: a Demonstration in Namibia. Data, 3, 30. doi: 10.3390/data3030030

Jazyk závěrečné práce: cze

Vedoucí diplomové práce: prof. RNDr. Petr Kubíček, CSc.

Konzultant diplomové práce: Mgr. Jiří Malý, Ph.D.

Datum zadání diplomové práce: listopad 2024
Datum odevzdání diplomové práce: podle harmonogramu

RNDr. Vladimír Herber, CSc.
pedagogický zástupce ředitele ústavu

Poděkování

Ráda bych ocenila všechny, kteří mě při studiu a zpracování diplomové práce vytrvale podporovali. Mé díky si jmenovitě zaslouží prof. RNDr. Petr Kubíček, CSc. za trpělivost a vstřícný přístup, Mgr. Yunshuo Tang za pomocnou ruku při analýze postupů a Mgr. Robert Šanda, Ph.D. za poskytnutí potřebných dat využitých v praktické části. Ovšem to největší si zaslouží Mgr. Dajana Snopková, Ph.D., která byla neustále připravena mi pomoci úplně s čímkoliv a předávala potřebné know-how. V neposlední řadě bych chtěla vyjádřit vděčnost své rodině a příteli za jejich neomezenou podporu, trpělivost a zázemí, které mi po celou dobu studia poskytovali.

Prohlášení

Prohlašuji, že jsem svoji bakalářskou/diplomovou práci vypracovala samostatně pod vedením ..... a s využitím informačních zdrojů, které jsou v práci citovány.

Prohlašuji, že jsem nástroje AI využila v souladu s principy akademické integrity a že na využití těchto nástrojů v práci vhodným způsobem odkazuji.

Brno, 16. 4. 2026

.....

Jméno Příjmení

Obsah

1 Úvod ... 11

2 Základní pojmy ... 12
2.1 Syntetická data ... 12
2.2 Syntetická populace ... 14
2.3 Mikrosimulace ... 15
2.3.1 Prostorová mikrosimulace ... 15
2.4 Model založený na agentech (Agent-based model, ABM) ... 16

3 Metody generování syntetické populace ... 18
3.1 Syntetická rekonstrukce (SR) ... 18
3.1.1 Algoritmus IPF ... 20
3.1.2 Algoritmus IPU – příklad ... 22
3.1.3 Porovnání výsledků algoritmů IPF a IPU ... 23
3.2 Kombinatorická optimalizace (CO) ... 25
3.2.1 Porovnání IPF a CO ... 26
3.3 Statistické učení (SL) ... 28
3.4 Velké jazykové modely (LLM) ... 30
3.5 Porovnání metod přístupů ... 34

4 Syntetická populace v prostoru ... 37

5 Případové studie ... 41
5.1 Pokročilý proces syntézy poptávky po dopravě pro vytvoření modelu aktivity MATSim: Případ Ústí nad Labem ... 41
5.2 Investigativní studie změny energetické politiky v Amsterdamu ... 45
5.2.1 Deskriptivní statistika ... 45
5.2.2 Strojové učení ... 45
5.2.3 Syntetická populace ... 47
5.2.4 Závěr ... 48
5.3 GenSynthPop ... 49

6 Dostupná vstupní data pro syntetickou populaci v České republice ... 52
6.1 Český statistický úřad ... 52
6.1.1 Sčítání lidu, domů a bytů ... 52
6.2 Ministerstva ... 55
6.2.1 Ministerstvo dopravy ... 55
6.2.2 Ministerstvo školství a tělovýchovy ... 57
6.2.3 Ministerstvo práce a sociálních věcí ... 57
6.2.4 Ministerstvo zdravotnictví ... 58
6.3 Lokální průzkumy ... 58
6.4 Zákon o svobodném přístupu k informacím ... 58

7 Prostorová data využitelná pro syntetickou populaci ... 60
7.1 Datový portál GIS ... 60
7.1.1 Vchody do budovy s TEP (RSO) ... 60
7.2 Lokální datové portály ... 61
7.2.1 Průzkum budov v Brně ... 62

8 Generování syntetické populace ... 63
8.1 Metodika ... 64
8.2 Identifikace a řešení chyb (tvorba syntetické populace) ... 68
8.2.1 Logické nedostatky vstupních datových sad ... 68
8.2.2 Přidávání ekonomické aktivity ve věkové skupině 0-14 let ... 69
8.2.3 Proces přidávání atributů do syntetické populace ... 70
8.2.4 Dojížďka ... 71
8.2.5 Přidávání na základě velkého množství atributů ... 72
8.2.6 Metoda eliminace nezjištěných hodnot atributů ... 72
8.2.7 Household Grouper ... 73
8.2.8 Přiřazení atributu typ budovy a patro budovy ... 76
8.3 Výstup ... 76

8.4 Prostorové rozmístění 78
8.4.1 Postup zpracování prostorových dat 78
8.5 Návrh vizualizace 82
8.5.1 Popis aplikace 82

9 Diskuze 87

10 Závěr 89

Seznam použité literatury 90

Seznam příloh 96

1 Úvod

V současné éře digitalizace a rozvoje konceptů Smart Cities narůstá potřeba přesného modelování městských systémů, ať už se jedná o plánování dopravy, optimalizaci veřejných služeb či analýzu šíření epidemií. Tyto pokročilé simulační nástroje, zejména agentní modely (Agent-Based Models – ABM), vyžadují vysoce detailní vstupní data o obyvatelstvu. Na straně druhé však stojí oprávněný a legislativně ukotvený požadavek na ochranu osobních údajů, který znemožňuje využití reálných, neagregovaných dat o konkrétních jednotlivcích. Ideálním řešením tohoto konfliktu je tvorba tzv. syntetické populace. Jak zdůrazňuje výzkumná zpráva Společného výzkumného střediska Evropské komise (HRADEC A KOL. 2022), syntetické populace představují zcela klíčový a perspektivní nástroj. Generují totiž depersonalizované databáze fiktivních obyvatel a domácností, které nepodléhají restrikcím GDPR, avšak jejichž statistické rozložení, socioekonomické charakteristiky a prostorové ukotvení věrně zrcadlí reálný stav daného území a poskytují tak nezbytný detail pro moderní politické a urbanistické rozhodování.

Předložená diplomová práce se zaměřuje na proces tvorby syntetické populace v národním prostředí České republiky, a to na příkladu statutárního města Brna. Zatímco v zahraničí je tato problematika, i s ohledem na zmíněné evropské iniciativy, metodicky a datově silně podporována, aplikace stávajících zahraničních algoritmů na české prostředí naráží na specifické překážky. Ty spočívají především v odlišné struktuře agregovaných dat z Českého statistického úřadu (ČSÚ), v anomáliích obsažených v otevřených datech a v absenci některých detailních registrů.

Hlavním cílem této práce je navrhnout, implementovat a kriticky zhodnotit metodický postup pro vygenerování a následné prostorové ukotvení syntetické populace města Brna tak, aby mohla sloužit jako validní vstup pro navazující dopravní mikrosimulace (např. v rámci projektu AMOS). Jako výchozí nástroj pro praktickou realizaci byla zvolena softwarová knihovna GenSynthPop. Dosažení tohoto cíle vyžaduje nejen integraci dat z celostátních a lokálních geoportálů, ale především nutnost zásahů do vybraného generovacího modelu. Práce se proto detailně věnuje řešení logických rozporů v datech a optimalizaci procesu shlukování jednotlivců do realistických domácností, což vede k maximálnímu možnému zpřesnění výsledného modelu.

V teoretické části práce jsou představeny základní koncepty syntetických populací a přehled dostupných datových zdrojů. Praktická část se následně zaměřuje na samotnou metodiku generování, čištění datových anomálií a prostorovou alokaci obyvatel na konkrétní adresní body a podlaží. Nedílnou součástí výstupů je také návrh a realizace interaktivní webové vizualizace v prostředí ArcGIS Experience Builder, která umožňuje plynulé prohlížení vygenerovaných dat. V závěru práce jsou kriticky diskutovány inherentní limity zvolených přístupů a navrženy směry pro budoucí rozvoj tohoto výzkumu.

2 Základní pojmy

V této kapitole jsou popsány základní termíny, které se vztahují k oblasti generování syntetické populace.

2.1 Syntetická data

Syntetická data představují uměle vytvořená data, jejichž cílem je věrně reflektovat strukturu a statistické charakteristiky původních (reálných) dat (EVROPSKÝ INSPEKTOR OCHRANY OSOBNÍCH ÚDAJŮ 2025). Podle Jordon a kol. (2022) zatím v odborné komunitě absentuje jednotná definice, načež autoři navrhují vymezení: „Syntetická data jsou data, která byla generována pomocí speciálně vytvořeného matematického modelu nebo algoritmu s cílem vyřešit (soubor) úkolů v oblasti datové vědy.“ Při následných statistických analýzách by měly obě datové sady poskytovat velmi podobné výsledky. Míra podobnosti datových sad přímo indikuje užitečnost zvolené metody a modelu. (EVROPSKÝ INSPEKTOR OCHRANY OSOBNÍCH ÚDAJŮ 2025)

Generování, respektive syntéza, může být prováděna pomocí různých technik jako je iterační proporcionální prokládání (IPF, *Iterative Proportional Fitting*), rozhodovací stromy, metody hlubokého učení (*deep learning*) anebo obecné systémy umělé inteligence (AI). Na základě originálních dat, která byla využita v procesu syntézy, lze syntetická data rozdělit do tří skupin: syntéza založena přímo na reálných datech (mikrodatech/vzorku), syntéza vycházející z agregovaných statistických výstupů (okrajové/marginální rozdělení, podmíněné distribuce, průměr…) a syntéza využívající kombinaci reálných dat a statistik. (EVROPSKÝ INSPEKTOR OCHRANY OSOBNÍCH ÚDAJŮ 2025)

Implementace syntetických dat představuje zásadní posun v metodice práce s citlivými nebo důvěrnými daty, protože zachovávají bezpečnost, ale zároveň poskytují cenné informace k rozhodování. Syntéza dat je finančně i časově méně náročná než tradiční metody sběru dat, jako jsou terénní průzkumy. Syntetická data se efektivněji ukládají a umožňují snadnější manipulaci. Používají se pro testování systémů nebo rychlá prototypování. Dále mohou lépe odpovídat požadovaným specifikacím na formát nebo kvalitu.

Zásadním analytickým přínosem je také možnost aktivně potlačovat tzv. systematické zkreslení v datech (*data bias*). Zatímco reálné datové sady často inherentně odrážejí existující společenské nerovnosti či diskriminační vzorce, syntetická data lze algoritmicky vyvažovat. Tím se minimalizuje riziko přenosu těchto nepřesností do následně trénovaných modelů strojového učení, což vede nejen ke zvýšení jejich přesnosti, ale především k zajištění větší objektivity výstupů. Vzhledem ke své plně depersonalizované povaze jsou navíc tato data bezpečně distribuovatelná v rámci odborných spoluprací, což ve výsledku přispívá k vyšší transparentnosti a reprodukovatelnosti vědeckého výzkumu. (LAMBERTI 2023, EVROPSKÝ INSPEKTOR OCHRANY OSOBNÍCH ÚDAJŮ 2025)

V závislosti na použitém modelu a kvalitě vstupních dat se mohou syntetická data v různé míře odchylovat od skutečnosti. Pokud model nezachytí kritické detaily, může to vést ke špatné předpovědi a následným neadekvátním rozhodnutím. Generování syntetických dat může být výpočetně náročný proces, zejména v případě nestrukturovaných dat (obrázky nebo text). Je velice obtížné validovat podobnost syntetických a reálných dat (které ani nemusí v takovém rozsahu existovat), proto neexistuje záruka, že bude model používající syntetická data vždy přesný. Další nevýhodou je závislost na kvalitě skutečných dat (sloužících jako vstupy do syntézy), veškeré chyby a zkreslení v podkladových datech se v procesu syntézy propagují do výsledku, pokud nejsou zachyceny. Problematický je i možný únik podkladových dat při generování. (LAMBERTI 2023, EVROPSKÝ INSPEKTOR OCHRANY OSOBNÍCH ÚDAJŮ 2025)

Pro dosažení optimálního výsledku je vhodné generovat rozmanitá data reflektující heterogenitu populace (nejen jednu izolovanou skupinu). Dále je potřeba syntetická data popsat metrikami kvality (úplnost, přesnost, preciznost) a otestovat je. Pro dlouhodobé využívání je nezbytné sledovat trendy v reálných datech a na základě nich aktualizovat ta syntetická. (LAMBERTI 2023)

V odborné literatuře se pro hodnocení užitečnosti syntetických dat navrhuje využívat již existující způsoby (viz Tab. 1).

Tab. 1 Hodnocení užitečnosti syntetických dat (přeloženo: EL EMAM 2020)

Přístup k hodnocení užitečnosti	**Vysvětlení**	**Použitelnost**
Strukturální podobnost	To je velmi důležité. Pokud data nejsou strukturálně podobná, pak to analytikům jen ztěžuje jejich použití.	Provést pro každou sadu dat
Obecné metriky užitečnosti	To je velmi důležité. Každý soubor dat musí splňovat minimální sadu metrik užitečnosti. To je relativně snadné, protože to lze do značné míry automatizovat.	Provést pro každou sadu dat
Replikace studií	Replikace je přesvědčivý způsob, jak prokázat, že se na metodu syntetických dat lze spolehnout. Jedná se o časově náročný proces, který vyžaduje odborné znalosti v dané oblasti.	Vyhodnocení metodiky
Subjektivní hodnocení odborníky na danou oblast	Tento typ hodnocení metodiky syntézy může být také poměrně přesvědčivý. Jedná se o náročnější hodnocení.	Provést pro každou sadu dat
Posouzení zkreslení a stability	Jedná se o obecně užitečný typ posouzení, které je třeba provést u každého vydání syntetických dat. Váha důkazů, které přidává k užitečnosti souboru syntetických dat, je však menší než u ostatních přístupů.	Vyhodnocení metodiky
Srovnání s veřejnými souhrnnými údaji	Srovnání s veřejnými údaji, jsou-li k dispozici, zvýší důvěru v metodiku syntézy.	Vyhodnocení metodiky
Srovnání s jinými technologiemi zvyšující ochranu soukromí	Tento typ hodnocení je užitečné v určitém okamžiku provést, aby pomohl rozhodovacím orgánům rozhodnout o relativních silných a slabých stránkách jednotlivých PET pro poskytování přístupu k datům.	Vyhodnocení metodiky

Aby byla syntetická data strukturálně podobná podkladovým datům, je nutné, aby splňovala stejné typy a formáty proměnných, názvy proměnných, metadata a formáty souborů. To umožňuje aplikovat identický analytický kód na obě datové sady (který může sloužit pro testování této metriky).

Obecné metriky užitečnosti porovnávají jednoduché popisné statistiky (průměr, rozpětí…), rozdělení hodnot (histogram) a vztahy mezi proměnnými (korelační matice). Patří sem i rozlišitelnost mezi reálnými a syntetickými daty. Se snižující se rozlišitelností, roste podobnost, a tak i užitečnost. Tyto metriky mohou být automatizované.

Jestliže je provedena identická analýza na syntetických datech se stejným výsledkem jako na reálných, jsou ta syntetická vysoce užitečná podle metriky replikace studií. Analýza musí být smysluplná vzhledem k účelu, za kterým byla syntetizována. Provedení jedné analýzy je pouze částečně informativní o dalších možných.

Subjektivní hodnocení odborníky probíhá posouzení věrohodnosti a reality v datech na jejich vzorku. K vyhodnocení se používají metriky přesnosti jako F-skóre. Nízká přesnost odpovídá vysoké podobnosti syntetických a reálných dat. (EL EMAM 2020)

Podle El Emam (2020) je syntéza stochastický proces, tudíž při každé iteraci vznikne odlišná datová sada. Posouzení zkreslení (*bias*) se provádí z velkého množství datových sad pocházejících ze stejného modelu a stejných vstupů do modelu a následném výpočtu průměrných obecných metrik užitečnosti. Variabilita těchto parametrů určuje jejich stabilitu. Zkreslení a stabilitu lze určovat i z replikací.

Srovnáním veřejných souhrnných údajů se syntetickými agregovanými údaji se také zjišťuje užitečnost, při lepší shodě je užitečnost vyšší.

Pseudonymizace, deidentifikace, federovaná analýza a protokoly založené na homomorfním šifrování jsou další způsoby anonymizace dat. Porovnáním datasetů těchto metod a syntetického datasetu je získáno relativní užitečnosti syntézy dat. Může to sloužit jako faktor při výběru metody anonymizace při poskytování dalším stranám. (EL EMAM 2020)

Syntetická data se nacházejí uplatnění v dopravě při simulacích a testování bezpečnosti autonomních vozidel. Ve finančním sektoru napomáhají odhalovat podvody u kreditních karet a pojistných událostí u domácností. Ve zdravotnictví umožňují simulovat dynamiku šíření nemocí nebo urychlovat vývoj farmaceutik. (IBM 2023)

2.2 Syntetická populace

Syntetická populace představuje specifickou podmnožinu syntetických dat. Je to uměle vytvořený model populace na vymezeném území, konstruován pomocí reálných dat a statistických metod tak, aby se podobal té skutečné. Vysoká míra podobnosti se skutečnou populací umožňují simulaci různých procesů, mezi které patří urbanistické procesy, dopravní poptávka, spotřeba energií,

hodnocení dopadů různých politik, prostorová dynamika šíření epidemií… Koncept není omezen pouze na lidskou populaci, může jít třeba o zvířecí nebo soubory ekonomických subjektů, například firmy.

Teoretické základy tohoto koncept vycházejí z pomezí statistiky a informatiky. Jedno z prvních využití nalezla populace v plánování dopravy už v roce 1996. Model Richard J. Beckman a jeho kolektivu byl uveden i do praxe jako součást projektu Transportation Analysis and Simulation System (TRANSIMS). Model využíval algoritmu IPF. Syntetizoval silnice, veřejnou dopravu i lidskou populaci ve zkoumaných oblastech. Populace jednotlivců s různými demografickými údaji a simulace jejich každodenního života byly schopny zlepšit dopravu a její vliv na kvalitu ovzduší, spotřebu energie a emise CO_2.

V současné době se pro syntézu používají pokročilejší modely, které poskytují uspokojivější a přesnější výsledky a snazší přístup ke kvalitnějším vstupním údajům. (THE DECISION LAB 2025)

2.3 Mikrosimulace

Mikrosimulace představují specifickou modelovací techniku založenou na interakci mikrojednotek, která se projeví na makroúrovni. Slouží k pochopení vlivu různých rozhodnutí a změn na budoucí populaci. Čím větší počet vzájemně závislých procesům a faktorů vstupujících do simulace, tím je model a pochopení dílčích vlivů náročnější. Mikrosimulace lze rozdělit na statické a dynamické.

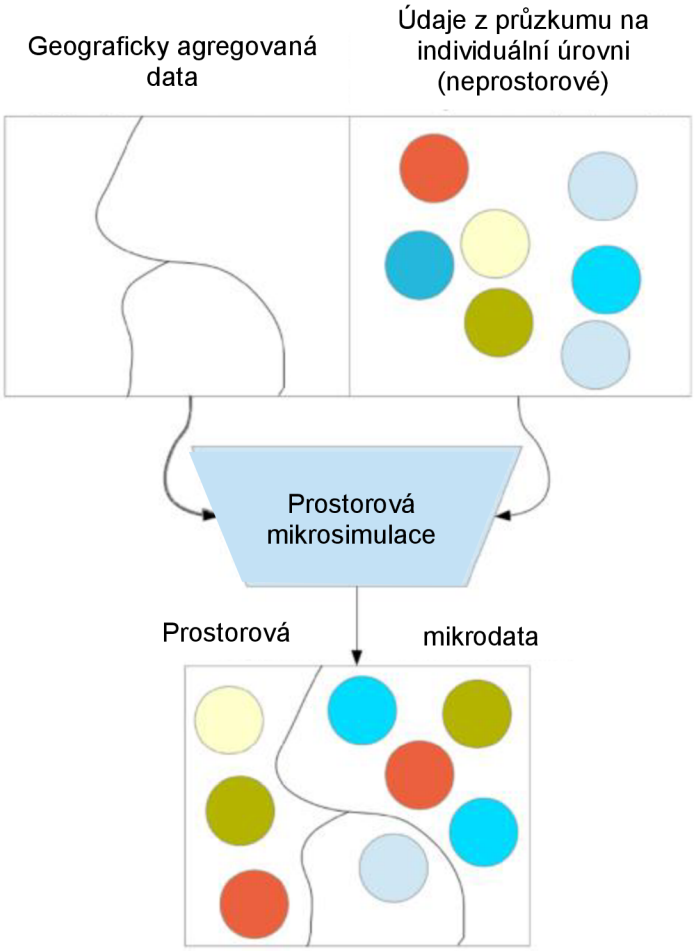
Statické mikrosimulace se používají ke studiu krátkodobého účinku změny politik. Takový model se skládá ze dvou částí: základní databáze (jednotlivci, domácnosti, vstupem může být syntetická populace) a soubor pravidel, která postihují novou politiku. Příkladem statickým modelem je EUROMOD, který umožňuje srovnání mezi jednotlivými zeměmi mající k němu přístup.

Dynamické modely umožňují sledovat jednotlivce v průběhu celého života. Obecně modely postihují dva typy chování, demografické události a změny způsobené vnějšími faktory (změnou politik). (SPIELAUER 2011)

2.3.1 Prostorová mikrosimulace

Podle příručky autorů Lovelace a Dumont (2018) lze prostorovou mikrosimulaci chápat jako specifický přístup nebo techniku. Může se jednat o metodu generování prostorových mikrodat, při které dochází k přiřazování jednotlivců do geografických zón (přičemž jsou vyžadovány minimálně dvě zóny, jinak se nejedná o prostorovou mikrosimulaci, ale pouze o mikrosimulaci), nebo o nástroj k pochopení jevů založených na prostorových mikrodatech. Definice termínu „prostorová mikrosimulace“ se může lišit v závislosti na kontextu a zaměření uživatele.

Na Obr. 1 je vyjádřen rozdíl mezi prostorovými mikrodaty a běžnějšími, ale oficiálně poskytovanými daty (v horní části).



Obr. 1 Prostorová mikrodata (přeloženo: DUMONT, LOVELACE 2018)

Metoda prostorové mikrosimulace zahrnuje syntézu dat (generování syntetické populace je prvním krokem mikrosimulace), proto se hodí i pro analýzy, ke kterým jsou dostupná pouze omezená data. Proces nekončí samotnou syntézou dat, neboť hlavním cílem je následná analýza vytvořených dat. V rámci syntézy se replikují jednotlivci (nevytváří se noví), čímž se nezvyšuje rozmanitost. K tomu se využívají metody syntetické rekonstrukce a kombinatorické optimalizace (nikoliv metody statistického učení, vytvářejí nové jednotlivce). (DUMONT, LOVELACE 2018)

2.4 Model založený na agentech (Agent-based model, ABM)

ABM jsou často považovány za odlišnou metodu, než představují mikrosimulace (SPIELAUER 2011). Neexistuje však jasně vymezená hranice, neboť ze statistického a výpočetního hlediska se často jedná o analogické principy. Mikrosimulace se primárně zabývají analýzou politik a jejich dopadů, zatímco ABM jsou více orientované na teorii a slouží k testování výzkumných hypotéz či ke studiu chování vyplývajícího z interakcí. Mikrosimulace obecně využívají přístupu částečné rovnováhy (sledují změnu v jednom trhu či v jedné politice, zbytek zůstává konstantní), oproti tomu ABM zase uzavřený systém (žádné vstupy nebo výstupy do vnějšího prostředí). (CEMPA 2025)

Prostorová mikrosimulace na rozdíl od ABM nemusí nutně zahrnovat analýzu interakcí mezi jednotlivci, vytváří jednotlivce a přiřazuje charakteristiky v prostoru v rámci jednodušších scénářů (zpravidla pro několik časových momentů). Naproti tomu ABM disponují větším prostorovým a časové rozlišení, což umožňuje modelovat chování a interakce jednotlivců v prostoru. Z toho vyplývá, že ABM je daleko náročnější na výpočetní kapacitu při stejném počtu jednotlivců. Z tohoto úhlu pohledu je možné prostorovou mikrosimulaci považovat za hierarchicky nižší úroveň ABM. (DUMONT, LOVELACE 2018) ABM využívají přístup zdola nahoru (bottom-up), protože se snaží odhalovat globální vzorce pomocí interakcí jednotlivců, zatímco mikrosimulace často využívají odlišný přístup (top-down), kdy se vyhodnocují globální změny a sledují se pomocí různých ukazatelů (ZAGHENI 2015).

3 Metody generování syntetické populace

Metody generování syntetické populace lze klasifikovat do tří hlavních kategorií: syntetická rekonstrukce (SR), kombinatorická optimalizace (CO) a statistické učení (SL) (viz Tab. 2). Dosud však neexistuje unifikovaná metodika pro výběr mezi těmito kategoriemi (YAMEOGO A KOL. 2020). S ohledem na dynamický vývoj v oblasti umělé inteligence v posledních letech je však nezbytné tuto klasifikaci rozšířit o velké jazykové modely (LLM).

Tab. 2 Základní přehled metod generování syntetické populace (přeloženo: YAMEOGO A KOL. 2020)

Kategorie	Náhodnost	Princip tvorby populace	Vybrané metody
SR	Deterministická	Kopie jednotlivců	Hierarchical Iterative Proportional Fitting, Iterative Proportional Update, Generalized Raking, Entropy maximization
CO	Stochastická	Optimální kombinace jednotlivců	Genetic algorithm, Greedy heuristic, Hill-climbing, Simulated annealing
SL	Stochastická	Odhad společného rozdělení	Hierarchical Markov Chain Monte Carlo, Hierarchical Mixture, Deep Generative Modeling, Bayesian Network
LLM	Stochastická	Sémantické modelování a autoregresivní syntéza sekvencí atributů	GReaT (Generation of Realistic Tabular data)

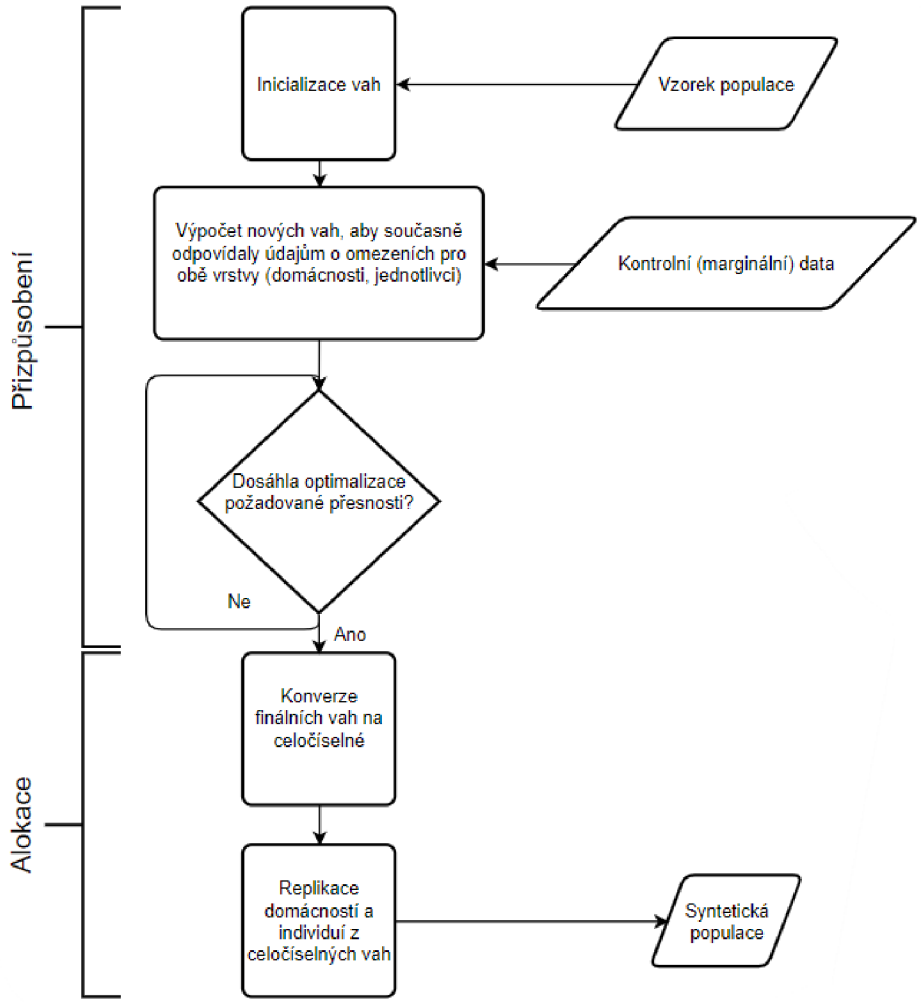
3.1 Syntetická rekonstrukce (SR)

Syntetická rekonstrukce je považována za tradiční přístup k modelování syntetické populace, který vyžaduje vstupní vzorek obsahující všechny atributy požadované ve výsledné populaci. Tento vzorek je následně zpracován pomocí IPF nebo jiného algoritmu za účel odhadu společného rozdělení tak, aby výsledné hodnoty odpovídaly známým marginálním součtům jednotlivých atributů.

Vzhledem k tomu, že detailní mikrodatový vzorek reálné populace není vždy dostupný, byl vyvinut alternativní přístup v rámci SR, který se obejde bez detailních mikrodat, ale vyžaduje dostatečně kvalitní agregovaná data. Na základě studií Barthelemy, Toint (2013) a Lenormand, Deffuant (2013) (in PELLEGRINO A KOL. 2023) jsou výsledky generované tradičním přístupem i touto modifikací srovnatelně uspokojivé. Hlavním negativem je častá ignorace vnitřní závislostí mezi úrovní domácností a jednotlivců (YAMEOGO A KOL. 2020).

Metody SR mají základní dva kroky: přizpůsobení se společnému rozdělení a alokace. V prvním kroku dochází k přiřazení vah jednotlivcům nebo domácnostem, aby součet vah odpovídal marginálním součtům dané oblasti. Váhy jsou vyjádřeny obvykle kladnými desetinnými

čísly. Při alokačním kroku jsou tyto váhy převedeny na celá čísla a provede se replikace podle váhy. Detailnější popis metody znázorňuje schéma (Obr. 2). Většina modelů v této kategorii vyžaduje reprezentativní vzorek, v němž je zachycena reálná korelace mezi atributy, a marginální součty. Klíčovou podmínkou je, aby vzorek obsahoval alespoň jedno pozorování pro každou kombinaci atributů; v opačném případě nemůže být jedinec s touto kombinací vygenerován (tzv. problém „nulových buněk“).



Obr. 2 Zjednodušený vývojový diagram metody syntetické rekonstrukce generování dvouúrovňové syntetické populace (přeloženo: YAMEOGO A KOL. 2020)

Algoritmus IPF je běžně standardně využívaný nástroj v rámci modelů syntetické rekonstrukce. Princip spočívá v iterativní úpravě buněk vícerozměrné kontingenční tabulky tak, aby výsledné hodnoty korespondovaly s marginálními součty cílových atributů. Metoda IPF je ceněna pro svou výpočetní nenáročnost, rychlost, a schopnost zachovat vnitřní korelační strukturu mezi atributy. Limitací jeho základní podoby je však neschopnost simultánně odhadovat charakteristiky na dvou hierarchických úrovních (detailní popis v kapitole 3.1.1)

Pro generování dvouúrovňové syntetické populace se proto využívají pokročilejší algoritmy. IPU (*Iterative Proportional Update*, iterativní proporcionální aktualizace) upravuje váhy pro domácnosti a následně pro osoby, dokud není dosaženo požadované přesnost, přičemž tyto dvě úrovně pojímá jako vzájemně provázanou kombinaci (příklad v kapitole 3.1.2). HIPF (*Hierarchical Iterative Proportional Fitting*, hierarchické iterativní proporcionální přizpůsobení) na rozdíl od IPU explicitně respektuje hierarchickou strukturu (domácnost je výš než jednotlivec). Algoritmus střídavě aktualizuje váhy obou úrovní tak, aby se výsledné řešení v každém kroku blížilo předcházející iteraci.

Alternativami využitelnými pro generování dvouvrstvé syntetické populace jsou metody entropické optimalizace, respektive kalibrační odhad (*Generalized Raking – GR*). GR představuje rámec různých algoritmů požadujících vstupní váhy, které jsou upravovány na základě minimalizace statistické vzdálenosti mezi vstupními (počátečními) a novými (konečnými) vahami. Entropická optimalizace je pouze jedním z algoritmů GR, konkrétně se snaží zmenšit logaritmickou vzdálenost.

3.1.1 Algoritmus IPF

IPF představuje nejrozšířenější metodu pro syntézu populací v rámci syntetické rekonstrukce. Vstupní data tvoří počáteční matice (*seed matrix*, viz Tab. 3) a marginální součet atributů (barevné pole tabulky) roven. Je nutné, aby byl součet sloupců (zelená pole tabulky) a řádků (modrá pole tabulky). Základní podmínkou je shoda celkových součtů marginálií v obou dimenzích (součet barevného řádku a součet barevného sloupce musí být roven). Buňky vstupní matice by neměly obsahovat nulové hodnoty, pokud se nejedná o strukturální nuly, přičemž platí, že čím reprezentativnější je poměr hodnot ve vstupním vzorku vůči marginálním omezením, tím vyšší je pravděpodobnost dosažení kvalitního výsledku. Taktéž cílové marginální hodnoty nesmí být nulové, v praxi se případná 0 nahrazuje velmi malým kladným číslem (např. 0.001). Jelikož IPF v průběhu výpočtu pracuje s reálnými čísly, výsledné hodnoty buněk jsou zpravidla desetinné. Pro získání celočíselné populace, která přesně odpovídá marginálním údajům, je nutné aplikovat dodatečné zaokrouhlovací procedury, neboť samotné IPF k exaktní shodě s celočíselnými okrajovými údaji zpravidla nedospěje. (HUNSINGER 2008)

Tab. 3 Vstupní hodnoty do algoritmu IPF (2 proměnné). Obarvené hodnoty v záhlaví tabulky jsou součty sloupce/řádku. (převzato: HUNSINGER 2008)

	35	40	25
20	6	6	3
30	8	10	10
35	9	10	9
15	3	14	8

Jedna iterace obsahuje vždy dva kroky, konkrétně úprava řádků a úprava sloupců. Pro lichý krok se používají hodnoty řádků, pro sudé hodnoty sloupců. Algoritmus skončí, když hodnoty součtu řádků a sloupců odpovídají (se nejvíce blíží) vstupním.

V prvním kroku (3.1) se vezme hodnota buňky ze vstupní (poslední) tabulky ($x_{i,j}$). Daná buňka je podělena součtem řádku ($\Sigma_i x_{i,j}$) ze vstupní tabulky a vynásobena marginálií pro daný řádek (R_i).

$$x_{i,j}^{(2k-1)} = x_{i,j}^{(2k-2)} \cdot \frac{R_i}{\sum_i x_{i,j}^{(2k-2)}} \qquad 3.1$$

Pro první buňku (35, 20): $6 \cdot 20/(6+6+3) = 8$

V druhém kroku (3.2) se zpracovávají hodnoty z poslední předchozí tabulky, daná buňka ($x_{i,j}$) podělena součtem sloupce ($\Sigma_j x_{i,j}$) z předchozí tabulky a vynásobena marginálií pro daný sloupec (C_i) (je nutné vycházet z Tab. 4 – iterace 1, krok 1).

$$x_{i,j}^{(2k)} = x_{i,j}^{(2k-1)} \cdot \frac{C_i}{\sum_j x_{i,j}^{(2k-1)}} \qquad 3.2$$

Pro první buňku (35, 20 → 29.62, 20): $8 \cdot 35/(8+8.57+11.25+1.8) = 9.45$

Tab. 4 Postup výpočtu algoritmu IPF (převzato: HUNSINGER 2008)

Úprava řádků

Iterace 1, krok 1

	29.62	39.61	30.76
20	8.00	8.00	4.00
30	8.57	10.71	10.71
35	11.25	12.50	11.25
15	1.80	8.40	4.80

Iterace 2, krok 3

	34.81	40.09	25.10
20	9.10	7.77	3.13
30	10.25	10.95	8.81
35	13.27	12.60	9.13
15	2.20	8.77	4.03

Iterace 3, krok 5

	34.99	40.00	25.00
20	9.14	7.75	3.11
30	10.30	10.92	8.78
35	13.34	12.57	9.09
15	2.21	8.76	4.02

Úprava sloupců

Iterace 1, krok 2

	35	40	25
20.78	9.45	8.08	3.25
29.65	10.13	10.82	8.71
35.06	13.29	12.62	9.14
14.51	2.13	8.48	3.90

Iterace 2, krok 4

	35	40	25
20.02	9.15	7.76	3.12
30.00	10.30	10.92	8.77
35.01	13.34	12.57	9.09
14.98	2.21	8.75	4.02

Iterace 3, krok 6

	35	40	25
20.00	**9.14**	**7.75**	**3.11**
30.00	**10.30**	**10.92**	**8.77**
35.00	**13.34**	**12.57**	**9.09**
15.00	**2.21**	**8.76**	**4.02**

3.1.2 Algoritmus IPU – příklad

IPU byl poprvé představen v roce 2009. Poskytuje lepší výsledky zejména při generování dvouvrstvé populace než IPF (JAIN, RONALD, WINTER 2015).

Obr. 3 obsahuje zjednodušený vzorek: typ domácnosti (1, 2) a typ jednotlivce (1, 2, 3) a počet domácností/jednotlivců daného typu v domácnosti. Všechny typy domácností a jednotlivců mají stejnou počáteční váhu (sloupec Weight). Pod tabulkou domácností a osob jsou parametry. Vážená suma (Weighted Sum) před začátkem IPU odpovídá součtu domácností/jednotlivců ve vstupním vzorku. Dalším známým údajem je marginální údaj *„Constraints“*, což je celkový známý počet domácností/jednotlivců daného typu. Jako ukazatel neshody je použita δ, čím menší, tím lepší shoda.

$$\delta = \frac{weighted\ sum - constraints}{weighted\ sum_0} \qquad 3.3$$

Iterace začíná úpravou vah podle domácnosti prvního typu, podle rovnice níže, kde i značí krok, který je počítán ($i = 0 \rightarrow$ iniciální stav)

$$weight_i = \frac{constraints}{weighted\ sum_{i-1}} \cdot weight_{i-1} \qquad 3.4$$

Po dosazení:

$$weight_1 = \frac{constraints}{weighted\ sum} \cdot weight = \frac{35}{3} \cdot 1 = 11.67$$

Následně se vypočítá *Weighted Sum 1* jako vážený počet typů domácností/jednotlivců. První iterace skončí po doplnění přizpůsobení jednotlivce typu 3 (*Weighted Sum 5*). Po jedné iteraci klesla průměrná neshoda o téměř 90 %. Po 638. iteraci je výsledek ideální, součet finálních vah odpovídá marginálním údajům. (YE A KOL. 2009)

Variable	Weight	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Weight 1	Weight 2	Weight 3	Weight 4	Weight 5	Final Weight
Household ID												
1	1	1	0	1	1	1	11.67	11.67	9.51	8.05	12.37	1.36
2	1	1	0	1	0	1	11.67	11.67	9.51	9.51	14.61	25.66
3	1	1	0	2	1	0	11.67	11.67	9.51	8.05	8.05	7.98
4	1	0	1	1	0	2	1.00	13.00	10.59	10.59	16.28	27.79
5	1	0	1	0	2	1	1.00	13.00	13.00	11.00	16.91	18.45
6	1	0	1	1	1	0	1.00	13.00	10.59	8.97	8.97	8.64
7	1	0	1	2	1	2	1.00	13.00	10.59	8.97	13.78	1.47
8	1	0	1	1	1	0	1.00	13.00	10.59	8.97	8.97	8.64
Output measure												
Weighted sum		3.00	5.00	9.00	7.00	7.00						
Constraints		35.00	65.00	91.00	65.00	104.00						
δ_b^a		0.9143	0.9231	0.9011	0.8923	0.9327						
Weighted Sum 1		**35.00**	5.00	51.67	28.33	28.33						
Weighted Sum 2		35.00	**65.00**	111.67	88.33	88.33						
Weighted Sum 3		28.52	55.38	**91.00**	76.80	74.39						
Weighted Sum 4		25.60	48.50	80.11	**65.00**	67.68						
Weighted Sum 5		35.02	64.90	104.84	85.94	**104.00**						
δ_a^b		0.0006	0.0015	0.1521	0.3222	0.0000						
Final weighted sum		35.00	65.00	91.00	65.00	104.00						

NOTE: δ = deviation measure. Text in bold signifies that the weighted sum for a control variable has been matched with the corresponding constraint.
[a]Average of $\delta_b = 0.9127$.
[b]Average of $\delta_a = 0.0954$.

Obr. 3 Generování syntetické populace pomocí IPU (převzato: KONDURI A KOL. 2016; YE A KOL. 2009)

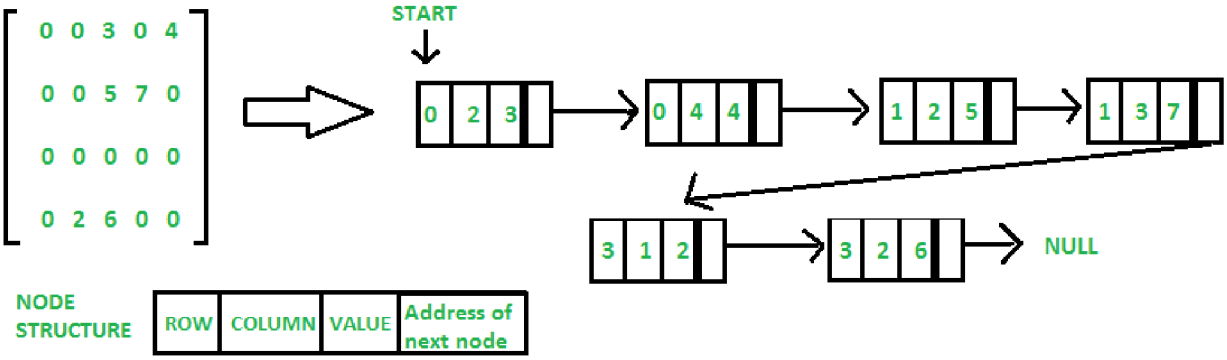
3.1.3 Porovnání výsledků algoritmů IPF a IPU

V roce 2015 provedli Jain, Ronald a Winter porovnávací studie dvou softwarových nástrojů pro tvorbu syntetické populace v australském Melbourne na základě dat ze sčítání lidu v roce 2011. Analýza se zaměřila na přesnost a výkonnost obou nástrojů při generování populací na úrovni domácností i jednotlivců s využitím metod syntetické rekonstrukce.

PopSynWin byl vyvinut Illinoiskou univerzitou v Chicagu v roce 2008. Využívá algoritmu IPF. V současné době software není veřejně dostupný, pravděpodobně byl technologicky zastaralý.

PopGen z Arizonské státní univerzity byl poprvé vydaný v roce 2009, poslední verze je z roku 2016. V současné době má dvě verze: první 1.1 s uživatelským rozhraním a druhé verze 2.0, která pracuje s CMD (příkazovým řádkem) a je 3-4krát rychlejší díky novějším standardům (PopGen 2024). PopGen pracuje na principu jako IPU (*Iterative Proportion Updating*).

IPU (na rozdíl od IPF) nevyužívá k ukládání dat klasické kontingenční tabulky, ale strukturu založenou na řídkých seznamech (*the sparse list-based data structure*), což je jeden z přepisů řídké matice (viz Obr. 4) (MÜLLER, AXHAUSEN 2010). Matice je definována jako řídká, pokud obsahuje méně než 5 % nenulových prvků a pokud její zpracování pomocí specializovaných technik vede k výrazné úspoře paměti a času při řešení úloh (TOMŠÍK 2012). Kontingenční tabulka s přibývajícími atributy roste exponenciálně, kdežto řídký seznam roste jenom lineárně. Proto se u IPU předpokládá i rychlejší zpracování. Dalším rozdílem je to, že IPF je vhodná pro data jenom s jednou úrovní, zatímco IPU dokáže efektivně pracovat s více hierarchickými úrovněmi současně.

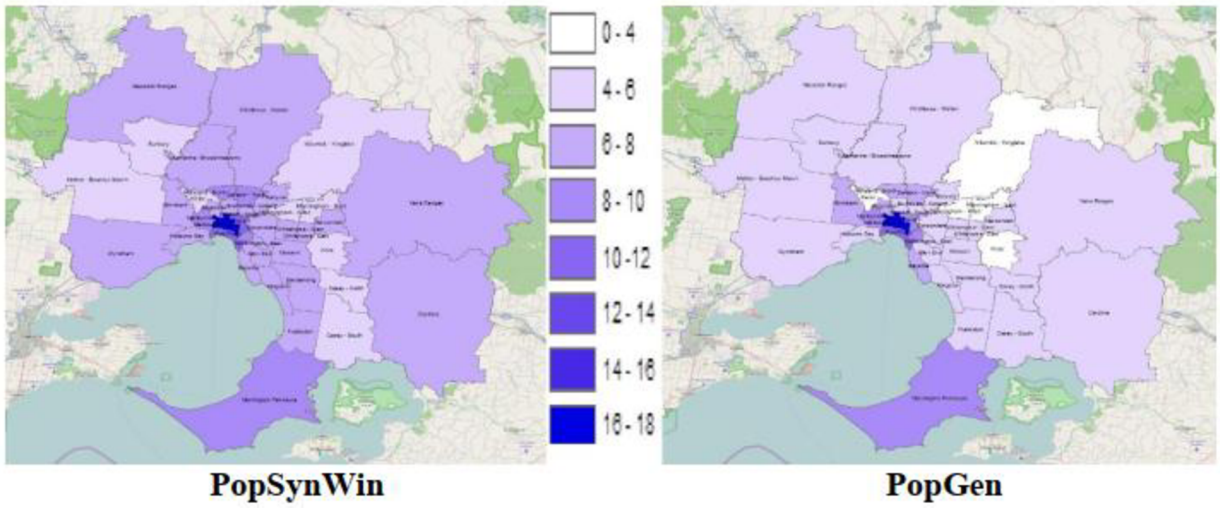


Obr. 4 Možný přepis řídké matice formou propojeného listu (zapisování nenulového prvku (node structure) pomocí indexu řádku (0-3), sloupce (0-4), nenulové hodnoty a adresa další nenulové hodnoty) (převzato: GEEKS FOR GEEKS 2023)

Po vygenerování populací byly výsledky agregovány za účelem validace kontrolních proměnných. Na úrovni domácností se jednalo o obytnou strukturu, počet osob obvykle bydlících v daném typu obydlí, počet motorových vozidel na domácnost. u jednotlivých osob bylo sledováno pohlaví, věk a ekonomická aktivita. Dále byly výsledky zkoumány i podle regionů, menších územních jednotek města Melbourne. (JAIN, RONALD, WINTER 2015)

Tab. 5 Vyhodnocení generování syntetických populací (přeloženo: JAIN, RONALD, WINTER 2015)

	Skutečná	PopSynWin	% rozdíl	PopGen	% rozdíl
Domácnosti	1 430 581	1 430 387	-0.01	1 430 581	0.00
Jednotlivci	3 995 665	3 708 065	-7.20	3 763 258	-5.82



Obr. 5 Procentuální rozdíl v počtu generovaných osob ve srovnání se skutečným počtem obyvatel (převzato: JAIN, RONALD, WINTER 2015)

V souladu s předpoklady autorů software PopGen poskytnul syntetickou populaci s lepšími výsledky (viz Tab. 5). Z vizuální reprezentace (Obr. 5) plyne, že přesnost syntézy se v různých

regionech studované oblasti liší. Tato variabilita je způsobena odlišnými charakteristikami regionů oproti celkové studované oblasti. Tento problém by bylo možné eliminovat generování populací pro menší území s homogennějšími charakteristikami.

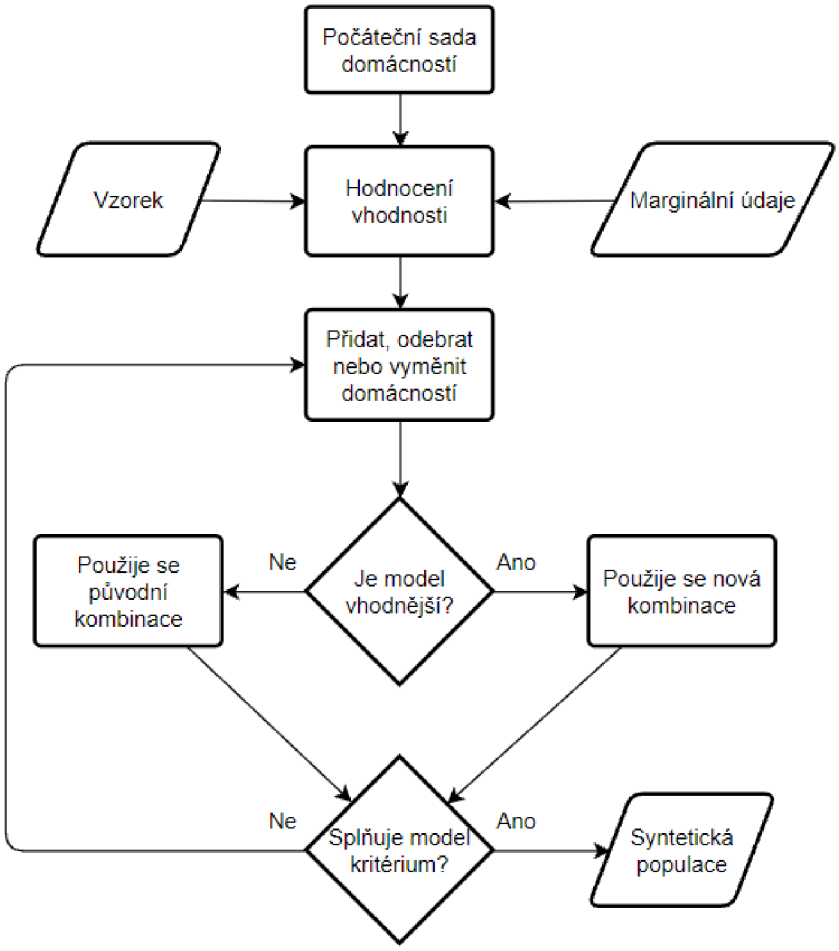
3.2 Kombinatorická optimalizace (CO)

Jako u většiny modelů SR tvoří základní vstupy pro modely CO reprezentativní vzorek populace a marginální data. Významnou předností modelů CO je schopnost generovat hierarchické úrovně (domácnosti i jednotlivce) současně v rámci jednoho výpočetního procesu. Metoda replikuje konkrétní jednotlivce ze vzorku, aniž by vyžadovala explicitní definici společného rozdělení atributů. Na rozdíl od SR, která v úvodní fázi přiřazuje jednotlivců desetinné váhy, pracuje kombinatorická optimalizace od počátku s vahami celočíselnými. Výhodou tohoto přístupu jsou méně striktní požadavky na vstupní data, nevýhodou je delší výpočetní čas a skutečnost, že u rozsáhlých populací nelze vždy garantovat dosažení globálního optima.

V úvodním kroku procesu se zájmové území rozdělí na menší celky, pro které jsou dostupná marginální data. Pro každou oblast je ze vzorku vybrán soubor domácností (včetně informací o příslušných členech) a následně je vyhodnocena míra shody (fitness) vzhledem k marginálním součtům. Algoritmus následně provádí iterativní proces, při kterém je domácnost v modelu přidána, odebrána nebo nahrazena jinou entitou ze vzorku a posléze se opětovně testuje shoda. Jedna domácnost se přidá nebo nahradí za jinou a znovu se zjistí shoda. Pokud se shoda zlepší, je nová domácnost v modelu ponechána; v opačném případě je změna zamítnuta. Tento cyklus se opakuje do naplnění zadaného počtu iterací nebo do dosažení uspokojivé úrovně shody (schéma principu metody vystihuje Obr. 6).

Shoda může být vyhodnocována například pomocí relativního součtu čtverců Z-skóre. K minimalizaci této hodnoty a dosažení lepší shody se využívají optimalizační algoritmy, jako je genetický algoritmus, greedy heuristika, hill-climbing nebo simulované žíhání.

Modely založené na kombinatorické optimalizaci jsou v praxi méně častější než modely syntetické rekonstrukce a ve většině případů použití ve studiích jsou jimi generovány populace s nižší počtem atributů (oproti modelům SR) a často se omezují jenom na jednoúrovňové populace (buď jednotlivci, nebo domácnosti). (YAMEOGO A KOL. 2020)



Obr. 6 Zjednodušený diagram kombinatorické optimalizace (přeloženo: YAMEOGO A KOL. 2020)

3.2.1 Porovnání IPF a CO

V rámci studie bylo provedeno porovnání kombinatorické optimalizace (CO) a syntetické rekonstrukce (konkrétně IPF) na modelové populaci firem. Byly zkoumány dva vlivy: velikost vzorku a úroveň detailu vstupních marginálních dat.

Byly zkoumány 1%, 2.5%, 5%, 7.5%, 10%, 20%, 50% a 100% vzorky dat a 3 úrovně detailu:

Úroveň detailu a – 2 tabulky (nízká složitost)

1. Kategorie počtu zaměstnanců (EmpCat), lokace (CT – census tract)
2. Detailní typ průmyslu (SIC-2d), lokace (CT)

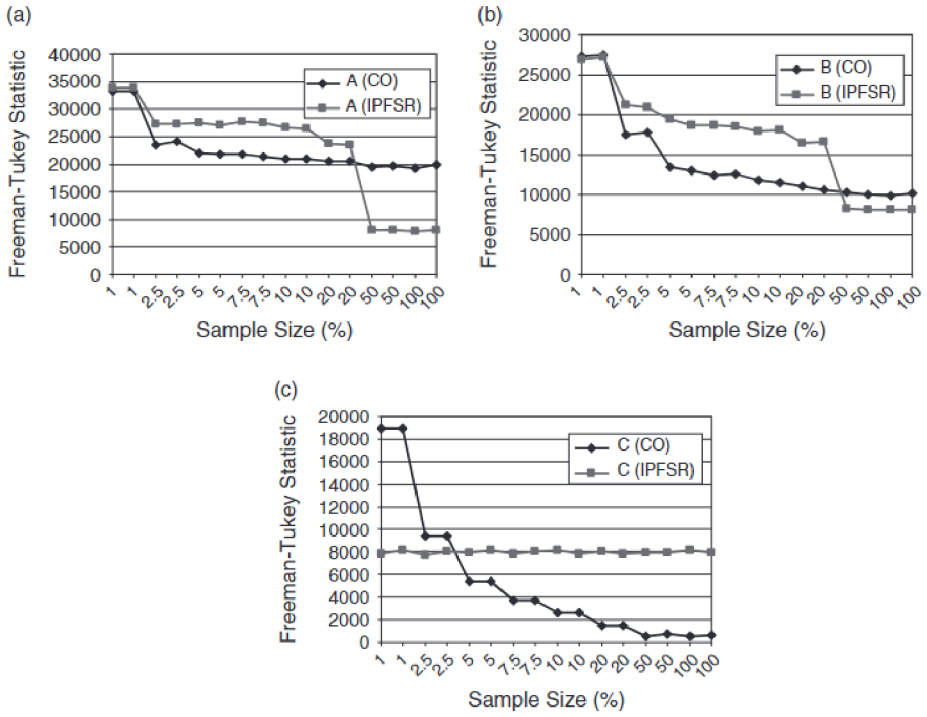
Úroveň detailu B – 2 tabulky (střední složitost)

1. Kategorie počtu zaměstnanců (EmpCat), méně detailní typ průmyslu (SIC-E), lokace (CT – census tract)
2. Detailní typ průmyslu (SIC-2d), lokace (CT)

Úroveň detailu C – 1 tabulka (vysoká složitost)

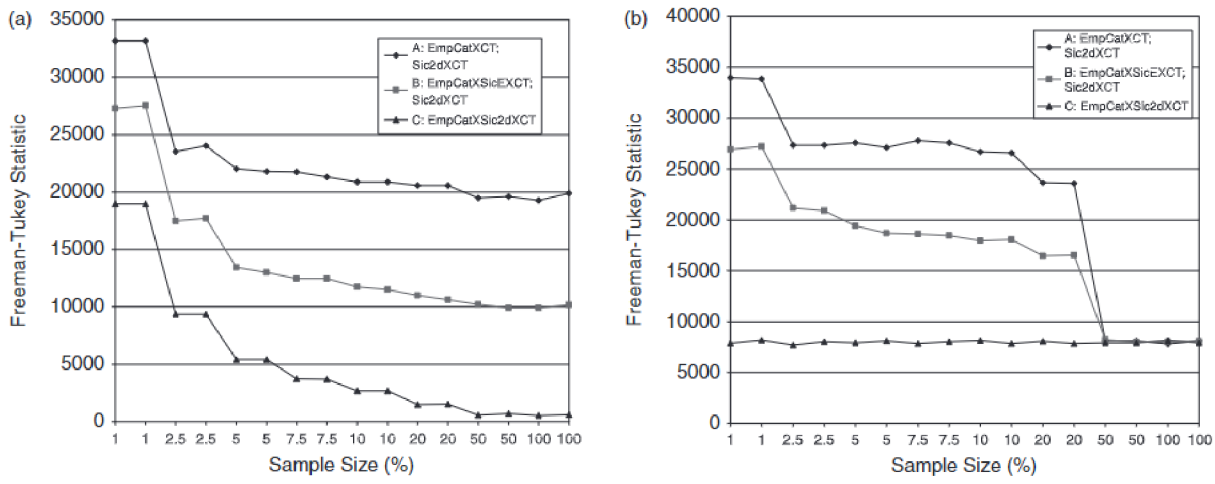
1. Kategorie počtu zaměstnanců (EmpCat), detailní typ průmyslu (SIC-2d), lokace (CT)

Bylo zjištěno, že syntetická populace při použití 100% vzorku a úrovně detailu a vykazovala horší shodu s reálnou populací než při použití 1% vzorku úrovně detailu C. Toto zjištění potvrzuje, že interakce mezi atributy zachycená v marginálních datech má na přesnost modelu zásadnější vliv než samotný objem mikrodat. Dále byl potvrzen předpokládaný kladný vztah mezi velikostí vzorku a mírou podobnosti syntetické populace s reálnou předlohou.



Obr. 7 Závislost nepodobnosti syntetické a reálné populace (Freeman-Tukeyho statistiky) na velikosti vzorku pro 3 úrovně detailu a 2 porovnávané modely (převzato: RYAN, MAOH, KANAROGLOU 2009)

Na základě výsledků autoři navrhují použití metody IPF pouze v případech nízkého až středního detailu omezujících dat (úrovně A, B) za předpokladu, že je k dispozici vzorek větší než 20 %. Metoda CO je doporučována pro scénáře s maximální úrovní detailu (úroveň C) nebo pro situace s nižším detailem dat, pokud je dostupný vzorek menší než 20 %.



Obr. 8 Závislost nepodobnosti syntetické populace s reálnou (Freeman-Tukeyho statistiky) a velikostí vzorku pro a) CO a b) IPFSR (převzato: RYAN, MAOH, KANAROGLOU 2009)

Hlavními limity tohoto srovnání jsou relativně malý rozsah testované populace (řádově tisíce subjektů) a omezený počet sledovaných atributů přiřazovaných jednotlivým firmám. (RYAN, MAOH, KANAROGLOU 2009)

3.3 Statistické učení (SL)

Statistické učení představuje přístup založený na simulaci, při které dochází k odhadu pravděpodobnosti kombinace atributů ve společném rozdělení všech sledovaných atributů proměnných ve vstupním vzorku. Pravděpodobnost je odhadována i pro kombinace, které ve vzorku přímo pozorovány nebyly. Hlavními výhodami této metody jsou skutečnosti, že jako vstupní data nejsou nezbytně vyžadovány souhrnné (marginální) hodnoty, takže stačí jenom reprezentativní vzorek, a že model dokáže generovat jednotlivce s kombinacemi atributů, které se ve vstupním vzorku nevyskytují. Díky tomu metody SL poskytují kvalitnější výsledky při použití malého (méně reprezentativního) vzorku než metody SR a CO. V případě potřeby kombinace marginálních údajů a vzorku je však nutný post-processing, kdy se doporučuje využít hybridní přístup. Pomocí SL se vytvoří dostatečně kvalitní syntetický vzorek, který vstupuje do SR spolu s marginálními údaji o populaci. Zjednodušený proces je znázorněn diagramem na Obr. 9.

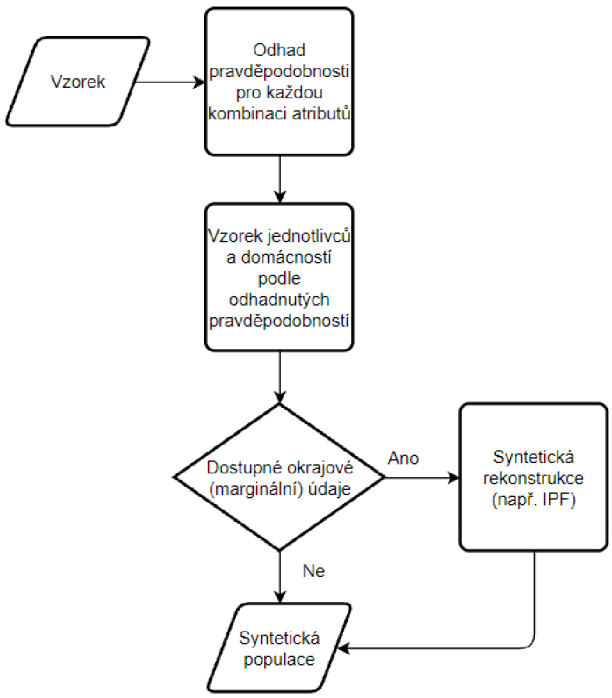
Metoda Markov Chain Monte Carlo (MCMC) představuje pokročilou třídu algoritmů určených k algoritmickému vzorkování z komplexních, vícerozměrných pravděpodobnostních rozdělení. Jak již název napovídá, spojuje dva matematické koncepty: metodu Monte Carlo, která využívá opakovaného náhodného vzorkování k odhadu výsledků, a tzv. markovské řetězce (Markov Chains), pro něž je charakteristická bezpaměťovost, každý následující stav systému závisí výlučně na stavu bezprostředně předcházejícím. Spojením těchto přístupů algoritmus iterativně prochází prostorem všech možných řešení a postupně konverguje k cílovému rozdělení pravděpodobnosti.

V oblasti populační syntézy se tento princip uplatňuje především pro modelování vícerozměrných vztahů mezi proměnnými (např. věk, vzdělání, příjem). Na rozdíl od jednodušších metod dokáže MCMC efektivně zpracovat provázanost těchto atributů ve vstupních datech tak, aby vygenerovaná fiktivní populace statisticky co nejvěrněji zrcadlila realitu. Standardní modely MCMC však typicky operují pouze na jedné úrovni, což znamená, že generují buď izolované jednotlivce, nebo pouze agregované domácnosti. Pro specifické účely generování tzv. dvouvrstvé populace, která vyžaduje současnou simulaci charakteristik jednotlivců i formování jejich pevných vazeb v rámci rodinných jednotek, je nezbytné tento přístup rozšířit. Z tohoto důvodu se pro danou úlohu jeví jako nejvhodnější využití hierarchické varianty (hMCMC).

V rámci procesu hMCMC probíhá tvorba domácností sekvenčně na základě pravděpodobnostního vzorku. Nejprve je vybrána tzv. „hlava domácnosti“ (např. osoba s nejvyšším příjmem) a následně jsou do struktury doplňováni další členové v definovaném pořadí, zpravidla partner, děti a ostatní osoby (např. prarodiče či spolubydlící). Výběr každého dalšího člena je přitom přímo podmíněn vlastnostmi osob, které již byly do dané domácnosti zařazeny.

Hlavní předností tohoto přístupu je schopnost generovat obě úrovně populace simultánně v rámci jednoho výpočetního kroku. Jak však upozorňují Casati a kol. (2015 in YAMEOGO A KOL. 2020), nevýhodou metody hMCMC zůstává vysoká míra subjektivity při výběru podmiňujících parametrů. Při generování dalších členů domácnosti totiž není vždy zřejmé, které atributy jednotlivce by měly sloužit jako primární podklad pro výběr vhodných partnerů či dětí, přičemž model nedokáže o nejvhodnější kombinaci těchto proměnných rozhodnout automaticky.

Dalšími metodami statistického učení jsou Bayesova síť (BN). Jejich zásadní výhoda oproti tradičním modelům neuronových sítí spočívá v tom, že neposkytují pouze deterministický bodový odhad (jednu konkrétní hodnotu), ale pracují s celým pravděpodobnostním rozdělením (Staff 2025). Vedle Bayesovských sítí se v současnosti prosazují pokročilé přístupy hlubokého generativního modelování (*Deep Generative Modeling* – DGM), kam spadají například generativní adverziální sítě (GAN) nebo variační autoenkodéry (VAE). Tyto modely jsou schopny se naučit komplexní strukturu reálných dat a následně generovat zcela nové, syntetické jednotky, které vykazují identické statistické vlastnosti jako originální vzorek. Poslední významnou kategorií jsou modely hierarchických směsí (*Hierarchical Mixture* – HM), které umožňují efektivně zachytit heterogenitu populace tím, že ji dekomponují na dílčí, statisticky homogennější podskupiny. (YAMEOGO A KOL. 2020)



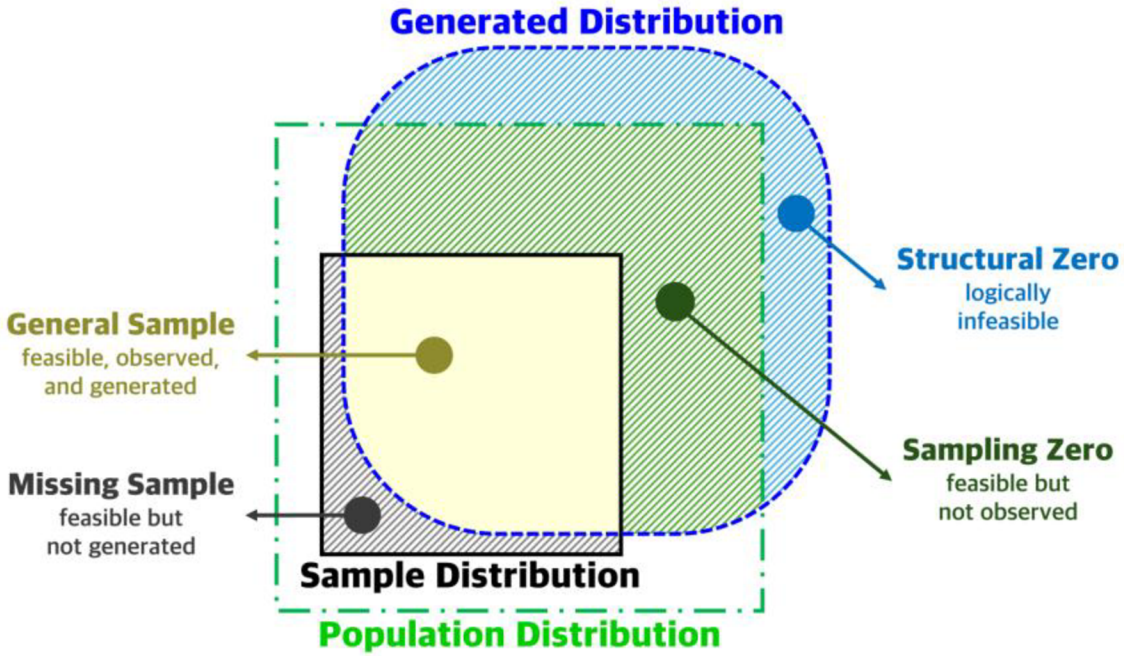
Obr. 9 Zjednodušený diagram procesu generování syntetické populace metodou statistického učení (přeloženo: YAMEOGO A KOL. 2020)

3.4 Velké jazykové modely (LLM)

Velké jazykové modely (LLM) představují v současnosti univerzální technologii pro produkci různorodého syntetického obsahu napříč mnoha odvětvími, od běžných administrativních úloh, jako je formulace e-mailů a shrnování dokumentů, až po komplexní interakce v rámci virtuálních asistentů a pokročilých překladatelských nástrojů. Zásadní vlastností moderních LLM je jejich narůstající multimodalita, která těmto modelům umožňuje v rámci jednoho systému současně interpretovat a generovat různé datové typy, jako je text, obraz, zvuk či video. Právě schopnost sémantického chápání světa a práce s rozsáhlým kontextem umožňuje těmto modelům přesahovat rámec čistého textu a nacházet uplatnění i při modelování strukturovaných dat, což představuje základní kámen pro moderní metody populační syntézy.

Významnou metodou generování dat pomocí LLM je GReaT (*Generation of Realistic Tabular data*), která využívá architekturu transforméru k modelování distribucí v tabulkových souborech. LLM modely jsou primárně autoregresivní (generují výstup jenom zleva doprava) a nejsou přizpůsobeny pro práci s 2D daty (čtení tabulek), proto GReaT využívá proces serializace. V rámci serializace byly řádky tabulky převedeny na textové sekvence (věk je 30, pohlaví je žena, povolání je učitel). Aby byl model schopen flexibilně chápat vztahy mezi proměnnými nezávisle na jejich pořadí v textu, využívá metoda při tréninku tzv. permutace atributů (náhodné míchání pořadí ve větě). (BORISOV A KOL. 2023)

Oproti algoritmům strojového učení LLM rozeznává sémantické vztahy mezi atributy a využívá znalosti načtené z rozsáhlých textových korpusů (*large-scale text corpora*), proto jsou vhodné pro minimalizaci strukturálních nul vyskytujících se v syntetické populaci (viz Obr. 10).



Obr. 10 Koncepční rámec strukturálních nul a vzorkovacích nul v populační syntéze (převzato: LIM A KOL. 2025)

Při generování syntetické populace je klíčové rozlišovat mezi dvěma typy nulových četností v kontingenčních tabulkách. Strukturální nuly *(Structural Zero*) představují logicky nepřípustné kombinace atributů (věk 6 let a řidičský průkaz sk. B). Vzorkovací nuly (*Sampling Zero*) jsou kombinace atributů nepozorované ve vzorku, ale jsou logicky v pořádku, mohou se tedy vyskytnout ve skutečné populaci. Ideálním případem při generování syntetické populace je maximalizace počtu vzorkovacích nul a minimalizace strukturálních. Každý generativní model založený na algoritmech strojového učení (*Deep Generative Models* – DGM) vytváří oba typy nul, s větším počtem vzorkovacích roste pravděpodobnost výskytu strukturálních. A právě tenhle nedostatek může být překonán využitím modelu, který bude vnímat sémantické vztahy (chápe význam generovaných hodnot).

Běžně dostupné proprietární LLM měly problém s rozmanitostí populace nebo naopak generují příliš mnoho strukturálních nul. Jako efektivní řešení byl navržen přístup využívající jednodušší open-source LLM, který je specificky natrénován pro tento účel. Taktéž využíval Bayesovu síť, která obsahovala vztahy mezi atributy, pravděpodobnost jejich kombinací a pořadí generování atributů. V prvním kroku bylo natrénováno LLM na vzorku reálné populace, důležité bylo nepřetrénovat, které by vedlo k pouhému kopírování namísto učení se vnitřní logice. Spojením sémantické kapacity LLM a statistické přípustnosti Bayesovských sítí vzniká rozmanitá populace s minimální počtem strukturálních nul. (LIM A KOL. 2025)

Na Obr. 11 je srovnáno 6 modelů:

- Prototypický agent (prototypical agent): Deterministická základna, která replikuje marginální distribuce z 5% vzorku (pravděpodobně metoda SR).
- DGM-VAE: Variační autoenkodér trénovaný k rekonstrukci 5% vzorku.
- DGM-WGAN: Wasserstein GAN trénovaný k rekonstrukci 5% vzorku.
- LLM-Few-shot: Proprietární LLM (GPT-4o) s few-shot učením na 5% vzorku.
- LLM-Random: Jednodušší LLM (GPT-2) vyladěný na 5% vzorku pomocí náhodných sekvencí atributů.
- LLM-BN: Jednodušší LLM (GPT-2) doladěný na 5% vzorku pomocí sekvencí atributů řízených BN.

Byla sledována distribuční podobnost (*distribution similarity*), která indikovala nepodobnost statistického rozdělení syntetické a reálné populace. Marginální SMRSE (*Standardized Mean Root Squared Error*, standardizovaná průměrná odmocnina kvadratické chyby) sleduje celkové počty v jednotlivých kategoriích, kdežto bivariantní kontroluje vztahy mezi dvojicemi atributů. V obou případech menší SMRSE ukazuje lepší shodu mezi syntetickou a reálnou populací. Druhým kritériem byla rozmanitost (diversity) jako počet unikátních kombinací *(# of combination*) a procentuální podíl zachycených kombinací modelem na kombinacích ve skutečné populaci (recall). Poslední zkoumanou vlastností byla logická přípustnost (*feasibility / precision*), která představuje procento vygenerovaných agentů, kteří jsou logicky v pořádku (zbytek tvoří strukturální nuly). Souhrnným ukazatelem je celková kvalita (*overal quality / F1 score*), která kombinuje předchozí kritéria.

LLM modely mají oproti ostatním větší SMRSE, natrénované jednodušší modely LLM poskytují více kombinací obsažených ve skutečné populaci a méně strukturálních chyb než proprietární LLM. Velmi dobrou celkovou kvalitu poskytují také modely SL (DGM-VAE a DGM-WGAN). Naopak proprietární LLM má celkovou kvalitu horší než využití metody syntetické rekonstrukce.

Model	Distributional Similarity		Diversity		Feasibility (Precision)	Overall Quality (F1 Score)
	Marg. SMRSE	Bivar. SMRSE	# of combinations	Recall		
Prototypical agent	0.008	0.020	30,837	56.4%	100.0%	72.1%
DGM-VAE	0.037	0.089	331,747	**81.5%**	73.6%	77.3%
DGM-WGAN	0.032	0.094	263,925	80.8%	81.4%	81.1%
LLM-Few-shot	0.220	0.617	40,666	50.2%	84.6%	63.0%
LLM-Random	0.165	0.394	180,504	**80.3%**	**90.8%**	**85.2%**
LLM-BN	0.249	0.604	120,541	76.0%	**95.3%**	**84.6%**

Obr. 11 Srovnání výkonnosti modelů pro generování syntetické populace (převzato: LIM A KOL. 2025)

LLM má významný potenciál pro predikování chování syntetických jednotlivců. Ovšem dotazování se na rozhodnutí jednotlivých agentů rozsáhlých syntetických populací se stává neproveditelným z výpočetního hlediska. Tento problém je možné řešit pomocí dotazů jen na jednotlivé reprezentativní archetypy (*prototypical agents*). Namísto simulace každého jednotlivce jsou dotazovány pouze vybrané typy agentů zastupující specifické socioekonomické kohorty, což umožňuje škálovat model i pro masivní simulace. (CHOPRA A KOL. 2024)

Výzvou zůstává generovat agenty nezápadní kultury. V současném době ve výzkumu dominují tzv. WEIRD populace (*Western, Educated, Industrialized, Rich and Democratic* – západní, vzdělané, industrializované, bohaté a demografické), které tvoří zhruba jen 13 % světové populace. Metodika generování „kulturních agentů“ vyžaduje nejprve extrakci specifických kulturních profilů, na jejichž základě jsou následně vytvářeny persony agentů. Tito syntetičtí kulturní agenti jsou podrobování klasickým behaviorálním experimentům. Výsledky byly porovnány s reálným lidským chováním dané kultury (pokud byli dostupné) a vykazují kvalitativní shodu v chování. Takto vytvořená populace je vhodný nástroj pro ladění výzkumných otázek nebo metodiky v sociálních vědách. (GONZALEZ-BONORINO, CAPRA, PANTOJA 2025)

Problémem je také paměť LLM, která je značně malá. Pokročilé systémy využívají architekturu externího úložiště vzpomínek (*Memory Stream*), kam si agenti zapisují vše, co se jim stalo. Každé události přiřadí LLM 3 hodnoty: důležitost, čerstvost a relevantnost k tématu (dynamická hodnota podle aktuálního jednání agenta). Agent je schopen se zamyslet, aby vyvodil závěry ze svých vzpomínek, nebo také plánovat a dynamicky plán měnit. Díky těmto schopnostem je možné pozorovat na virtuální syntetické populaci realistické (nenaprogramované) chování. (PARK A KOL. 2023)

Kvalitu simulace ovlivňuje také proces „zarovnání“ (*Alignment*) modelů. Standardní chatovací modely (např. Mixtral) jsou cenzurované a mají být nápomocné, neškodné a etické, čímž dochází pozitivnímu zkreslení (*positivity bias*) a k potlačení negativních vlastností. Následkem

toho je deformace distribuce osobností v syntetické populaci. Oproti tomu základní modely (např. DarkLlama), které nepodléhají cenzuře, dokáží vygenerovat širší spektrum lidských charakterů, včetně těch negativních, byť jsou náročnější na přesné zadání instrukcí. (CASORIA A KOL. 2025)

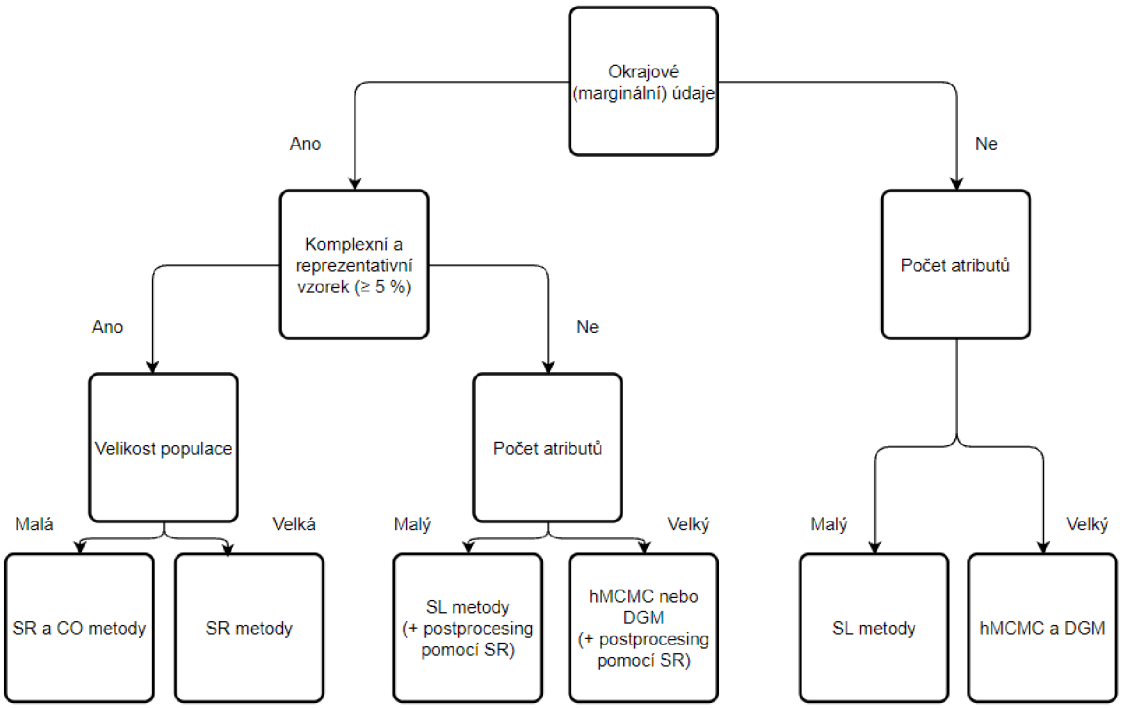
3.5 Porovnání metod přístupů

V Tab. 6 je pomocí generativní AI vytvořena tabulka podle Yameogo a kol. 2020 a kapitoly 3.4. Z porovnání plyne, že LLM představují alternativní větev k SL.

Tab. 6 Porovnání metod přístupů (OPEN AI 2025A)

Kritérium	SR (Synthetic Reconstruction)	CO (Combinatorial Optimization)	SL (Statistical Learning)	LLM (Large Language Models)
Přístup	Deterministický (fitting + alokace)	Stochastický (iterativní výběr + výměna)	Pravděpodobnostní (učení rozdělení)	Sémantický / Generativní (autoregresivní modelování sekvencí)
Vstupní data	Vzorek + marginály	Vzorek + marginály	Jen vzorek (může fungovat bez marginálů)	Extenzivní předtrénované korpusy textů + vzorek (few-shot / fine-tuning)
Schopnost fitovat marginály	Ano (přesně)	Ano (přesně)	Ne — nutný postprocessing SR metodami	Ne — vyšší chybovost (SMRSE), vyžaduje integraci např. s Bayesovými sítěmi
Vhodné pro malý vzorek	Ne — potřeba velkého reprezentativního vzorku	Ne — nutný dostatečný vzorek	Ano — může fungovat i s malým vzorkem	Ano — využívá obrovskou globální bázi znalostí z předtrénování
Počet atributů (škálovatelnost)	Vysoká — zvládá mnoho atributů	Nízká — naráží na výpočetní náročnost	Omezená — např. BN těžko škáluje s více atributy	Vysoká — limitováno pouze velikostí kontextového okna a procesem serializace
Zero-cell problém	Ano — nelze generovat kombinace mimo vzorek	Ano — podobně jako SR	Ne — umí interpolovat i nepozorované kombinace	Ne — exceluje v minimalizaci strukturálních nul díky chápání reálné logiky světa
Spojení jednotlivců a domácností	Někdy odděleně (HIPF, IPU to řeší lépe)	Ano – výběr celých domácností s osobami	Ano – přirozené v hierarchických modelech (hMCMC)	Lze definovat — závisí na struktuře promptu a schopnosti modelu formátovat výstup (např. JSON)
Složitost implementace	Nízká až střední	Střední	Vysoká (např. hMCMC, BN, DGM vyžadují pokročilé modely)	Velmi vysoká — nároky na hardware, složitý prompt engineering a alignment
Reprodukce vzácných kombinací	Ne – omezeno na vzorek	Ne – omezeno na vzorek	Ano – modeluje kombinace i mimo vzorek	Vysoce rozmanité — dokáže generovat nové reálné vzorce díky sémantické flexibilitě
Použití v praxi	Běžně používané (např. IPF, IPU, PopSyn)	Málo používané (náročné na výpočet)	Zatím spíš výzkumné prototypy	Nová experimentální sféra (cutting-edge výzkum)

Podle dostupných dat a parametrů výsledné populace byl sestaven rozhodovací strom pro nalezení ideální metody pro generování dvouvrstvé syntetické populace, tak aby byl výsledek, co nejlepší. (YAMEOGO A KOL. 2020)



Obr. 12 Rozhodovací strom pro výběr metody generování syntetické populace (přeloženo: YAMEOGO A KOL. 2020)

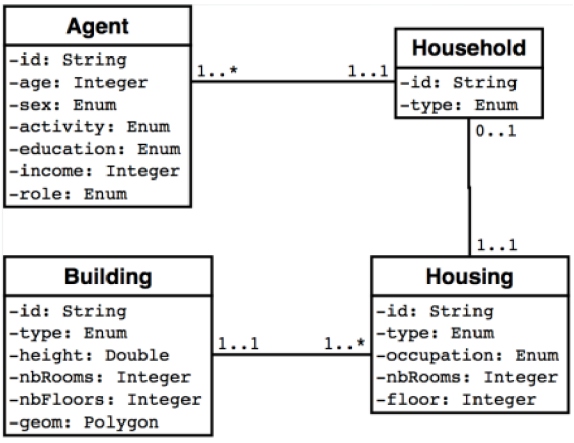
4 Syntetická populace v prostoru

Syntetická populace se obvykle pojí s určitým (většinou administrativním) územím, pro které je generována. V mnoha aplikacích, jako je studium dojížďky za prací nebo simulace šíření epidemií, je nezbytné doplnit k atributům i údaje o poloze, přičemž primárním bodem zájmu je místo bydliště. Pro přiřazení konkrétních souřadnic neexistuje zavedená unifikovaná metoda, protože rozložení objektů je velmi nerovnoměrné, čímž mohou vzniknout velké prázdné prostory (LENTI A KOL. 2025).

Nejjednodušším mechanismem je náhodné umisťování domácností do budov v rámci daného území tak, aby byly respektována kapacita a účel budovy (PEREIRA A KOL. 2022; MEISTER A KOL. 2008). Tento způsob je nenáročný na vstupní data, ale často opomíjí socioekonomické odlišnosti.

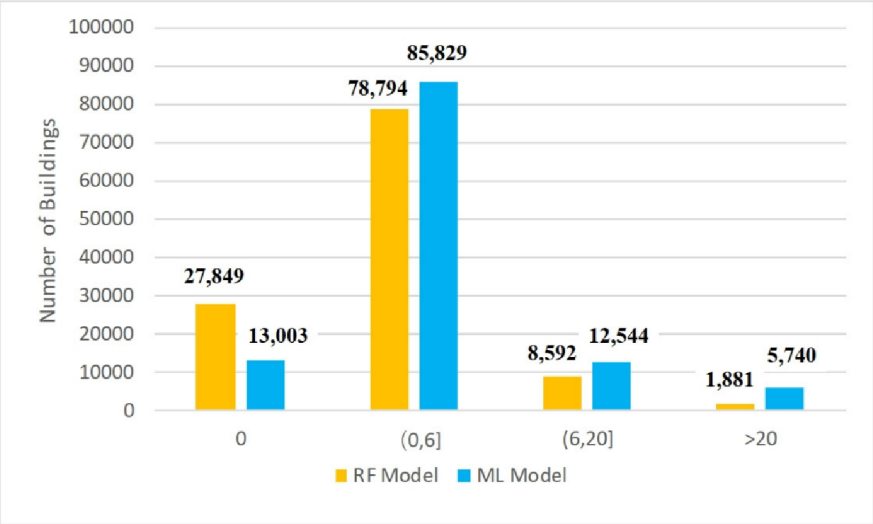
Pro irskou populaci byl vyvinut přístup (model SMILE), kdy se pro každou domácnost vypočte pravděpodobnost umístění v bytě bytového domu (na základě počtu osob v domácnosti, příjmu, věku a pohlaví hlavy domácnosti a náhodného čísla logistického modelu). Domácnosti jsou následně seřazeny podle této pravděpodobnosti, vybrány ty s největší pravděpodobností a umístěny do bytových domů v obydlených rezidenčních oblastech pocházejících z GeoDirectory, databáze obsahující všechny byty a domy. Konkrétní byt byl vybrán náhodně. Po naplnění byly zbývající domácnosti přiřazeny obdobnou logikou do rodinných domů. (CULLINAN 2010)

Další přístup byl vyvinut v rámci MobiSim, francouzského agentního LUTI (*Land Use and Transportation Interaction*) modelu, který využívá podmíněné pravděpodobnosti. Generuje jednotlivce, kteří jsou seskupeni do domácností, následně každou domácnost na základě diskrétního pravděpodobnostního zákona zařadí do bytové jednotky (bytového domu nebo rodinného) na základě rozdílu počtu jednotlivců v domácnosti a počtu pokojů bytu (které jsou výsledkem náhodného výběru na základě rozlohy bytu). Budovy jsou rozděleny na bytové a rodinné podle plochy a výšky. Ze sumy objemů každé budovy lze vypočítat průměrný objem pokoje (počet pokojů je znám pro každou domácnost). Pro každou budovu je potom vypočítán počet pokojů a na základě toho odvozen počet podlaží. Poté dochází k přiřazení bytu s domácností do budovy počtu pokojů v bytě a domě, dochází tedy k propojení informací z cenzu a geografických informací. Vzhledem k tomu, že vstupy do modelu jsou známé pro celou Francii, pocházejí z cenzu a IGN geografické databáze, je možné generovat jakoukoliv oblast IRIS (oblast čítající 1 800 – 5 000 obyvatel). Schéma jednotlivých datových sad je na Obr. 13. (ANTONI, KLEIN 2017)



Obr. 13 Vstupní datové sady a vztahy mezi nimi. Jednotlivci (agent) jsou charakterizováni věkem (age), pohlavím (sex), ekonomickou aktivitou (activity), vzděláním (education), příjmem (income) a postavením v domácnosti (role). Domácnost (household) je popsána poze typem (type). Byt (housing) má typ (type), obsazenost (occupation), počet pokojů (nbRooms), podlaží (floor). Budova (building) se vyznačuje typem (type), výškou (height), počtem pokojů (nbRooms), počtem podlaží (nbFloors) a geometrií (geom). (převzato: ANTONI, KLEIN 2017)

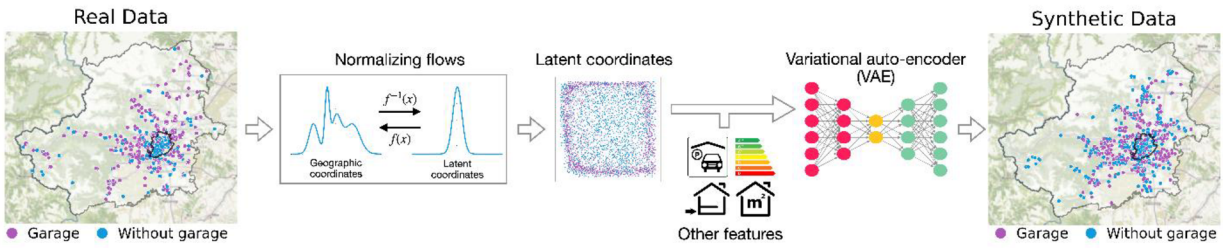
Alternativní metodou je využití algoritmu Random Forest (RF), který je schopen predikovat počet obyvatel v konkrétní budově na základě plochy budovy, vzdálenosti od silnic, řek a bodů zájmu, land-use a míry veřejného osvětlení (odvození z nočních satelitních snímků). Na základě takového vzorku je možné odhadovat počet lidí v jiných budovách. Výkonnost algoritmu Random Forest byla porovnána s lineární regresí založenou na strojovém učení, kde byl každý parametr vážen koeficientem tak, aby jejich suma predikovala výsledný počet obyvatel budovy. Výsledky studie potvrzují vyšší přesnost RF. Obr. 14 znázorňuje distribuci počtu budov v závislosti na chybě v absolutním počtu obyvatel. (WANG A KOL. 2022)



Obr. 14 Počet budov v závislosti na chybě v absolutním počtu osob v budově (převzato: WANG A KOL. 2022)

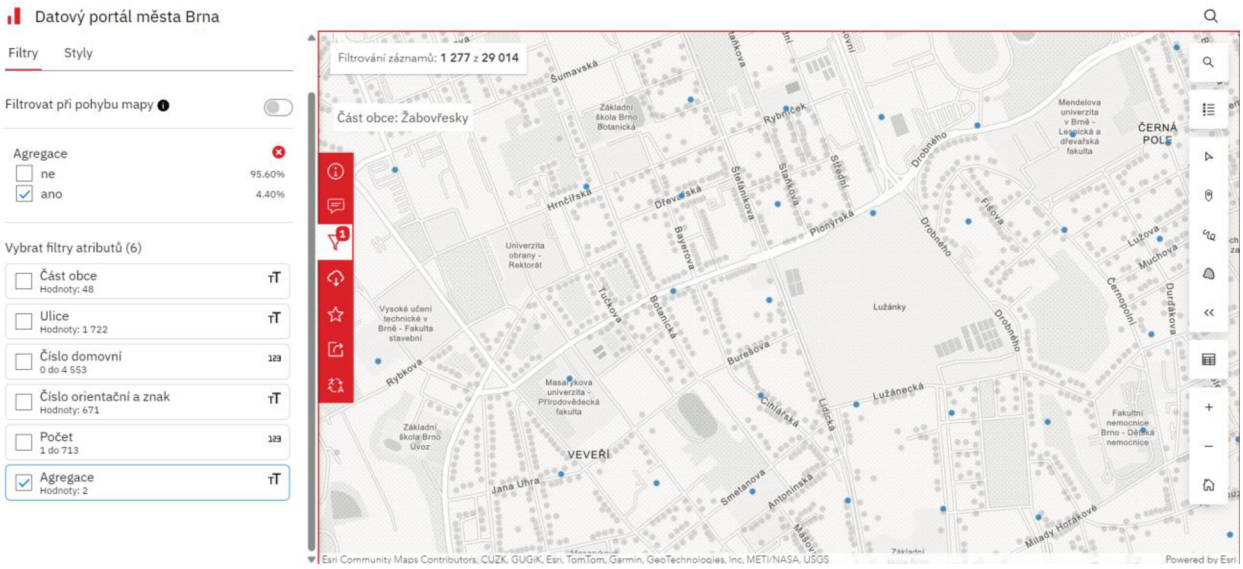
Další možností v rámci moderních přístupů je využití geolokalizovaných dat jako přímého vstupního vzorku. Pro tento účel byl navržen trojfázový algoritmus (viz Obr. 15), který k syntéze

využívá normalizační toky (*Normalizing Flows*). V první fázi tyto toky transformují vybraného území do pravidelného latentního prostoru (odstraní neobyvatelné části a upraví ho tak, aby bylo odpovídalo známému teoretické rozdělení). Následuje aplikace variační autoenkodéru, který se v latentním prostoru učí vztah mezi geografickou polohou a socioekonomickými charakteristikami populace. Následně probíhá inverzní normalizační krok, který mapuje vygenerovaná syntetická data zpět do reálného geografického prostoru. Hlavní odlišností tohoto přístupu je, že nealokuje entity do konkrétních budov, ale generuje pravděpodobné místo výskytu domácnosti. (LENTI A KOL. 2025)



Obr. 15 Schéma fungování trojfázového algoritmu (LENTI A KOL. 2025)

V českém národním prostřední je při prostorové distribuci nezbytné pracovat primárně na úrovni budov (pokud populace nebude generována na souřadnice), protože v současné době neexistuje ucelená databáze či registr bytů (ŠANDA 2022). Pro město Brno existují volně dostupné datové sady „Průzkum budov v Brně“, která obsahuje údaje o počtu pater a funkční využití jednotlivých pater, a „Počet osob na adresních místech“, které by mohly stanovit kapacitu pro konkrétní budovy. V této datové sadě proběhla anonymizace, pokud byly v budově registrovani méně než tři osoby, byl vytvořen centroid gridu s hranou 250 metrů s agregovanou hodnotou (viz Obr. 16).



Obr. 16 Centroidy gridu o hraně 250 metrů s agregovanou hodnotou počtu osob (převzato: STATUTÁRNÍ MĚSTO BRNO 2025)

Díky údajům ze SLDB, kdy se získávají také informace o patru a typu domu, by se dala teoreticky každá budova spárovat s 0 až *n* domácnostmi. Při praktické implementaci je nutné vyřešit nesoulady. Data jsou z různých zdrojů a odlišného stáří, tudíž s vysokou pravděpodobností se nebude shodovat počet obyvatel v datové sadě „Počet osob na adresních místech“ se syntetickou populací vygenerovanou podle dat ČSÚ. Bude nutné vyřešit umístění agregovaných údajů v podobě centroidů (cca 19 000 lidí). V neposlední řadě je potřeba přesně lokalizovat neprotínající se adresní body s příslušnou budovou.

Další vhodnou datovou sadou jsou geodata „Vchody do budovy s TEP (RSO)“ z ČSÚ 2026b, které poskytují adresní body budovy s technicko-ekonomickými parametry budovy. Využitelnými atributy jsou: způsob využití budovy dle ISKN (párovatelné na typ budovy podle číselníku (ČÚZK 2026)), počet všech podlaží budovy, počet bytů v budově, vchodu, počet obvykle bydlících obyvatel dle SLDB 2021 a identifikátor stavebního objektu v RÚIAN. Datová sada je dostupná pro celou ČR.

5 Případové studie

Pro podrobnější demonstraci praktického využití syntetické populace byly vybrány tři případové studie: jedna z českého prostředí zaměřená na dynamické modelování pohybu, druhá statická, určená k analýze dopadů politik, a poslední, která se týká modelu využitého jako základ pro praktickou část této práce.

5.1 Pokročilý proces syntézy poptávky po dopravě pro vytvoření modelu aktivity MATSim: Případ Ústí nad Labem

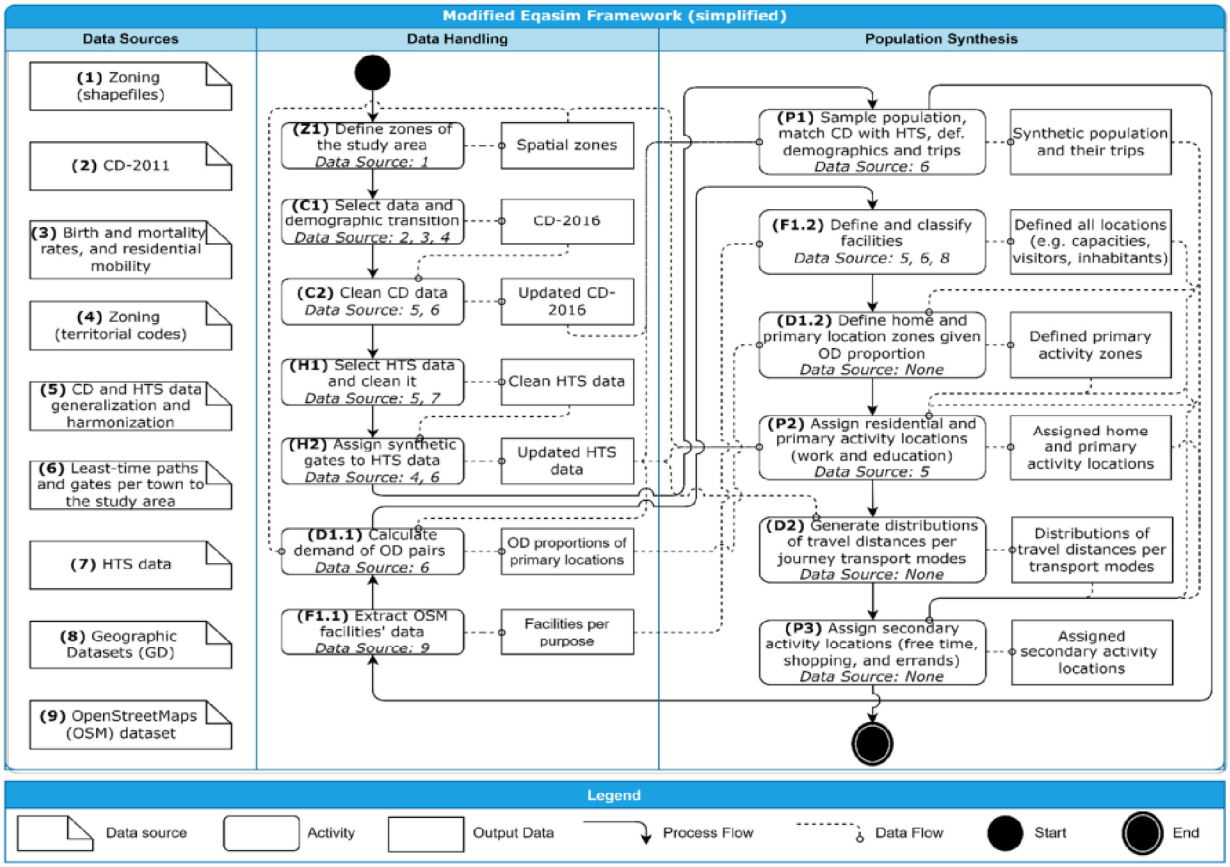
Tento model vznikl jako součást projektu *„Smart City—Smart Region—Smart Community"* a představuje první aplikaci syntetické populace v českém prostředí, konkrétně ve funkčním regionu Ústí nad Labem (zahrnujícím krajské město a přilehlé obce tvořící s ním provázaný celek) (PEREIRA A KOL. 2022). Celkově se jedná o 116 916 obyvatel v regionu, který v posledních dekádách prošel transformací z ekonomiky orientované na průmysl směrem k terciárnímu sektoru.

Syntetická populace zahrnující cestovní poptávku byla vygenerována s pomocí upraveného modelu Eqasim, který byl původně navržen pro São Paulo. Vzhledem k dostatečné reprezentativnosti vstupních dat byl aplikován základní scénář. Samotný výpočetní kód byl rozčleněn na menší modulární segmenty, aby byla usnadněna následná manipulace a úpravy systému.

Vstupní databázi pro generování syntetických jednotlivců tvořily údaje ČSÚ a data z výběrových šetření o mobilitě domácností. Jelikož poslední dostupná data ze sčítání lidu pocházela roku 2011 (*CD-2011*), byla tato data transformována pomocí demografického algoritmu (*Select data and demographic transition*), aby se vymodelovala struktura obyvatelstva odpovídající roku 2016. Data z průzkumu mobility domácností (*HTS data*) poskytla vzorek obyvatel obsahující informace o preferovaném dopravním módu, účelu, vzdálenosti a čase cesty. V modelu byla kombinována dvě šetření – národní průzkum zpracovaný Centrem dopravního výzkumu a lokální šetření na úrovni města, které zachycovalo cesty začínající nebo končící na území Ústí nad Labem. Národní průzkum byl vyfiltrován, aby pokrýval ostatní cesty po 23 obcích v okolí. Pro zvýšení vypovídající hodnoty byl využit *„EU Survey on Issues Related to Transport and Mobility"*, který posloužil jako odhad pravděpodobnosti pobytu jednotlivců v domácím prostředí v závislosti na jejich ekonomické aktivitě. Geografický kontext modelu zajistily prostorové datové sady (*Geographic Datasets*) zahrnující funkční využití území (*land use*), budovy a vybavenost, které byly extrahovány z registru sčítacích obvodů a budov. Všechna demografická data byla harmonizována, aby si kategorie vzájemně odpovídaly. Pro upřesnění prostorových informací, jako jsou otevírací doby, byla využita data z projektu OpenStreetMap (*OpenStreetMap dataset*) a Rejstřík škol a školských zařízení.

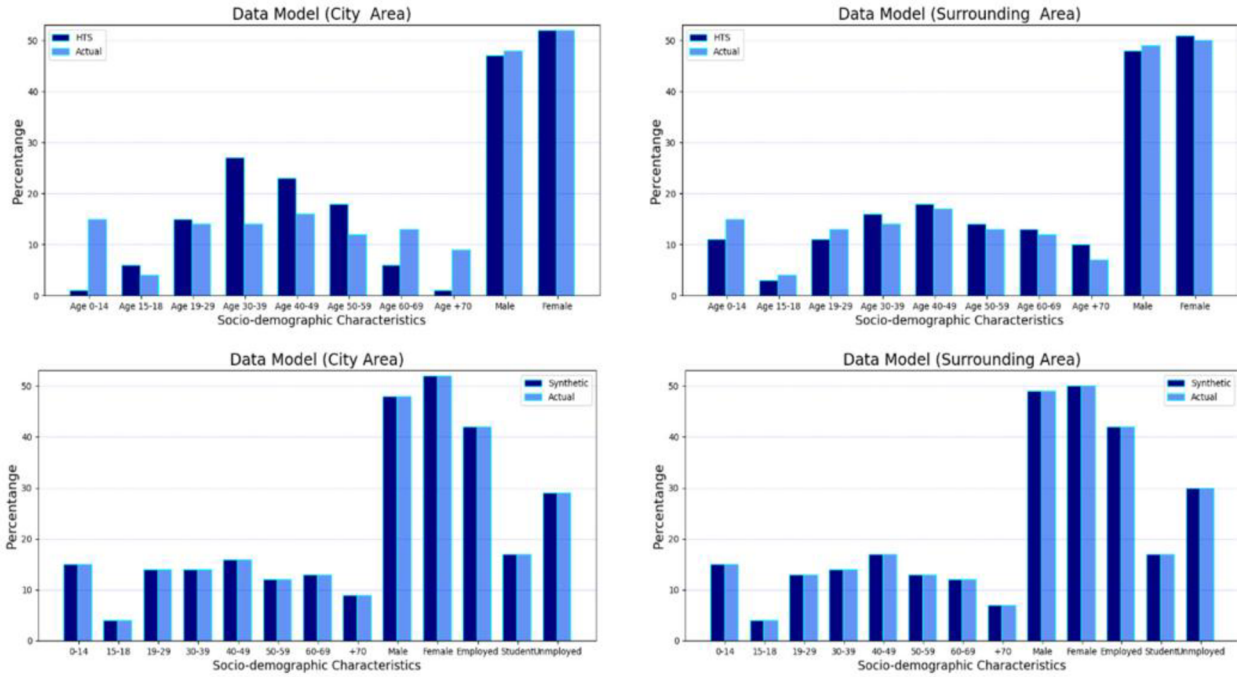
Procesní postup (Obr. 17) zahrnoval nejprve definování zájmového území (*Define zones of the study area*) na základě územního členění ČR (*Zoning – shapefiles*). Následně proběhla

modelace demografického vývoje (*Select data and demographic transition*), která zohlednila natalitu, mortalitu a migraci (*Birth and mortality rates, and residential mobility*) v rámci daných územních kódů (*Zoning – territorial codes*). Data ze sčítání lidu byla pro rok 2016 očištěna (*Clean CD data*), přičemž došlo k filtraci jednotlivců ve studovaném území a k harmonizaci kategorií mezi cenzem a dopravními průzkumy (*CD and HTS data generalization and harmonization*). V rámci výběru a filtrace dat mobility (*Select HTS data and clean it*) byly odstraněny nekonzistentní záznamy s chybějícími klíčovými údaji. Vzhledem k tomu, že celostátní šetření neposkytovalo dostatečný vzorek pro studovanou lokalitu, byly definovány syntetické brány pro tranzitní, příjezdové a odjezdové cesty (*Assign synthetic gates to HTS data*). Tyto brány byly vytvořeny na základě průměrných dob cestování a populace v cílových obcích. Následně byla cenzální data spárována s daty z průzkumů (*Sample population, match CD with HTS, def. demographics and trips*) na základě podobnosti sociodemografických atributů, čímž byly k syntetickým entitám přiřazeny konkrétní cestovní vzorce. Z databáze OSM byly vyextrahovány budovy a plochy dle klíčových slov definujících jejich využití (*Extract OSM facilities' data*). Podle vzorku respondentů byly vypočítány podíly cest za primární aktivitou, tedy prací či vzděláním (*Calculate demand of OD pairs*). Extrahovaná prostorová data byla klasifikována a doplněna o externí registry, například Registr škol a školských zařízení (*Define and classify facilities*). Pro každého jednotlivce byl definován domov a lokalita primární aktivity (*Define home and primary location zones given OD proportion*), které byly následně přiřazeny pomocí dvou algoritmů (*Assign residential and primary activity locations*). Na základě průzkumů domácností byla vygenerována rozložení vzdáleností pro každý dopravní mód (*Generate distributions of travel distance per journey transport modes*). V závěrečné fázi byly s využitím relaxačně-diskretizačního algoritmu přiřazeny lokality pro sekundární aktivity (volný čas, nákupy či vyřizování záležitostí), přičemž systém vybíral zařízení s volnou kapacitou a porovnával vypočtenou vzdálenost s reálnými daty z průzkumů (*Assign secondary activity locations*).



Obr. 17 Schéma upraveného rámce Eqasim (Modified Eqasim Framework) pro modelování regionu Ústí nad Labem (převzato: PEREIRA A KOL. 2022)

Porovnání dat z průzkumů mobility (*HTS*) s modifikovanými daty ze sčítání (*Actual*) ukazuje shodu v poměru pohlaví, avšak odhaluje výrazné odchylky ve věkové struktuře, zejména v centrální části města. Tento nesoulad je způsoben omezenou velikostí vzorku lokálního šetření v Ústí nad Labem oproti robustnějšímu národnímu průzkumu využitému pro okolní obce. Výsledky potvrzují nezbytnost využívat cenzální data jako kontrolní a podpůrný mechanismus pro vážení výběrových šetření. Spodní část grafů (*Synthetic vs. Actual*) však prokazuje, že syntetická populace věrně replikuje reálné charakteristiky území, čímž úspěšně koriguje chyby vstupních průzkumů.



Obr. 18 Socio-demografické charakteristiky populace pro centrální město (City Area) a okolní oblasti (Surrounding Area) (převzato: PEREIRA A KOL. 2022)

Analýza mobility ukazuje, že veřejná doprava dominuje u primárních aktivit (cesty do práce a škol), zatímco u sekundárních aktivit (nákupy, volný čas) výrazně narůstá preference individuální automobilové dopravy. Sekundární cesty vykazují vyšší míru prostorové a časové homogenity. Nejvyšší četnost cest u primárních aktivit vykazuje věková skupina ≥ 40 let, což reflektuje její dominantní zastoupení v populaci. Vyšší míra využívání veřejné dopravy u žen je v souladu s etablovanou literaturou v oblasti transportního inženýrství. (PEREIRA A KOL. 2022)

5.2 Investigativní studie změny energetické politiky v Amsterdamu

Tato kapitola hodnotí dotační programy zaměřené na zvyšování energetické účinnosti budov v Amsterdamu. V rámci studie Hradec a kol. 2022 byly zkoumány a komparovány tři metodické přístupy: deskriptivní statistika, strojové učení (*Machine Learning*) a tvorba syntetické populace.

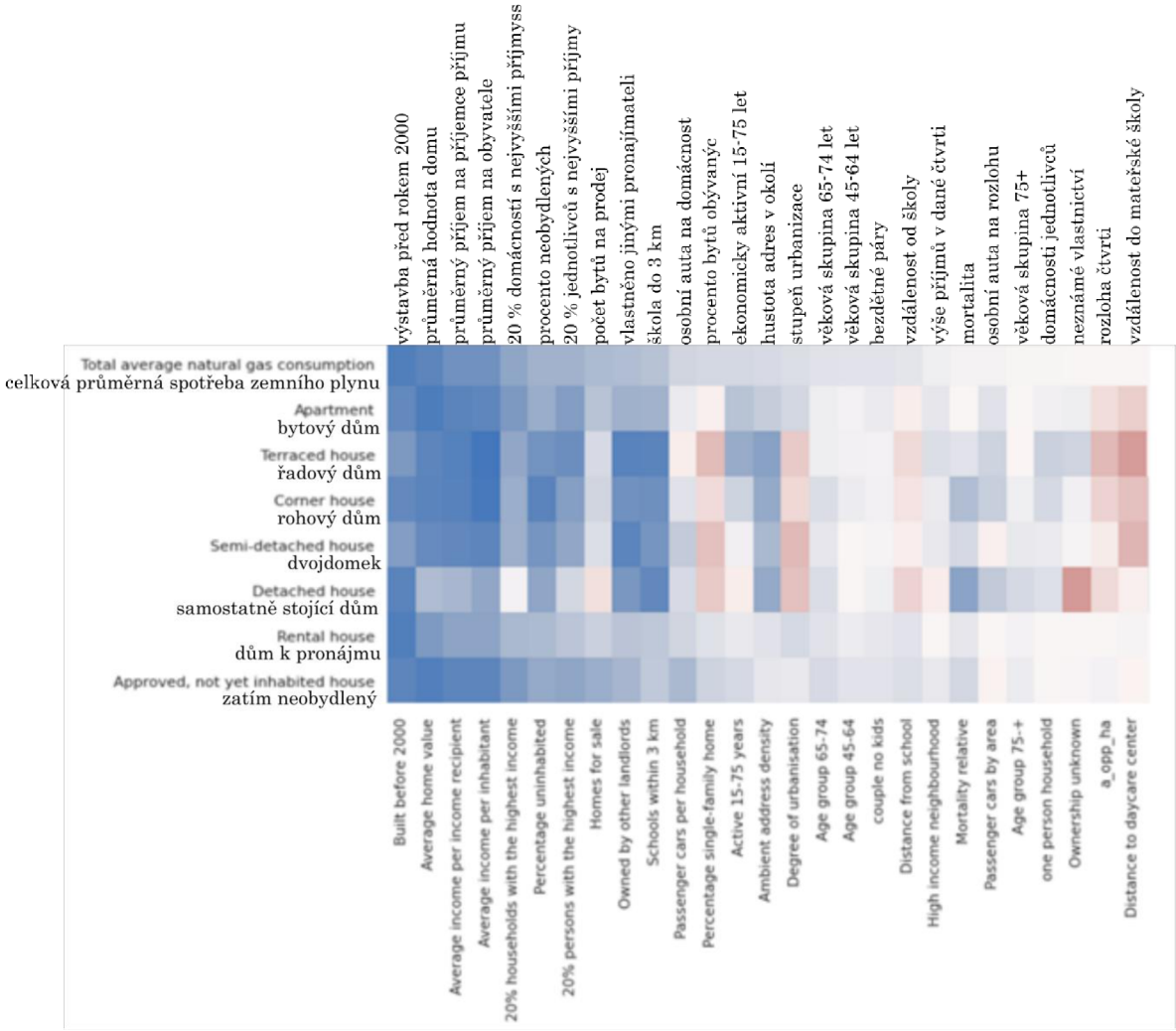
5.2.1 Deskriptivní statistika

Statistický přístup představuje metodicky nejsnazší cestu, neboť se opírá o porovnávání marginálních a průměrných hodnot. Zjištěný trend rostoucí spotřeby tepelné energie v Amsterdamu naznačuje, že navrhovaná politika úspor je relevantním nástrojem. Na základě údajů o počtu domácností a jejich průměrném příjmu se energetická půjčka s minimálním úrokem jeví jako vhodné a široce dostupné řešení. Ačkoliv se jedná o velmi jednoduchý a pro politické rozhodování přesvědčivý nástroj, jeho limitací je přílišné zjednodušení komplexních sociodemografických vazeb. (HRADEC A KOL. 2022)

5.2.2 Strojové učení

Tento analytický nástroj pracuje s podrobnými údaji o jednotlivých budovách (adresa, spotřeba energie, výška a typ objektu) v kombinaci s charakteristikami obyvatel, jako je věk a příjem. Data byla vizualizována pomocí vícerozměrné projekce UMAP (*Uniform Manifold Approximation and Projection*), která slouží k identifikaci shluků domácností s podobnými vlastnostmi. Jelikož docházelo ke shlukování, byly vytvořeny další grafy znázorňující spotřebu plynu podle typu domácností v závislosti na průměrné plošné spotřebě plynu. Korelační vztahy byly analyzovány pomocí Pearsonova korelačního koeficientu mezi typem budovy a ostatních charakteristik budov. Z koeficientů plyne, že rodinné domy mají menší energetickou stopu, díky vysokým soukromým investicím.

Další korelační koeficienty byly vypočteny mezi demografickým složením a spotřebou plynu podle typu bydlení, rodinným stavem a spotřebou plynu podle typu bydlení, příjmem a spotřebou plynu podle typu bydlení. Z těchto hodnot byly nalezeny a seskupeny nejdůležitější vlivy na spotřebu plynu (viz Obr. 19 Korelační matice socioekonomických, technických a environmentálních atributů: modrá kladná korelace, bílá nulová korelace, červená záporná; intenzita vyjadřuje sílu korelace (přeloženo: HRADEC A KOL. 2022)Obr. 19).



Obr. 19 Korelační matice socioekonomických, technických a environmentálních atributů: modrá kladná korelace, bílá nulová korelace, červená záporná; intenzita vyjadřuje sílu korelace (přeloženo: HRADEC A KOL. 2022)

Pomocí algoritmu SelectKBest s testem chí-kvadrát bylo identifikováno 10 nejvlivnějších faktorů, přičemž jako dominantní determinant spotřeby plynu byl určen příjem domácnosti (viz Tab. 7).

Tab. 7 Faktory určující spotřebu plynu (přeloženo: HRADEC A KOL. 2022)

Faktor	Spotřeba plynu	Skóre
Rodina s vysokým příjmem	vysoká	133 679
Rodina s nízkým příjmem	nízká	88 963
Procento lidí nezápadního původu	nízká	72 374
Procento lidí na sociálních dávkách	nízká	68 887
Samostatně stojící dům	nízká	60 918
Procento domů postavených v letech 1965-1974	vysoká	2 752
Procento domů postavených po 2010	nízká	2 217
Procento domů postavených v letech 1992-2009	nízká	2 115
Procento domů postavených v letech 1946-1964	vysoká	2 105
Procento domů postavených v letech 1975-1991	vysoká	1 678

Distribuce věku obyvatel ve vztahu k typu budovy byla analyzována pomocí odhadu hustoty jádra (*Kernel Density Estimation* – KDE), což odhalilo, že starší generace často obývá energeticky neefektivní domy postavené v 60. a 70. letech. Při zkoumání rozdělení spotřeby v oblastech podle poštovních směrovacích čísel byla kvůli vysoké šikmosti dat aplikována logaritmická transformace. Výsledkem této metody je zjištění, že stávající politické nástroje řeší spíše absolutní čísla než podstatu problému; potřebu renovace domu totiž nelze spolehlivě určit pouze na základě aktuální spotřeby energie, která může být u nízkopříjmových skupin uměle nízká kvůli úspornému chování z donucení. (HRADEC A KOL. 2022)

5.2.3 Syntetická populace

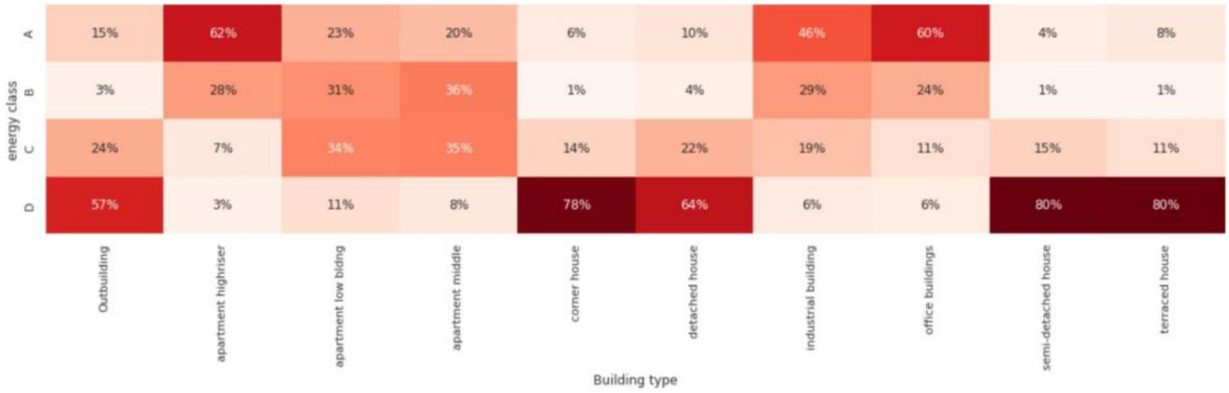
Při agregaci dat dochází k nevyhnutelné ztrátě strukturních informací, čemuž předchází metoda syntetické populace, která umožňuje mapovat a aproximovat složité nelineární funkce. Cílem bylo ověřit hypotézy, že nízkopříjmové rodiny s dětmi postrádají kapitál na zateplení a senioři nemají zájem o dlouhodobé úvěrové závazky.

Pomocí algoritmu IPF byla vygenerována rozmanitá populace s vysokou mírou entropie (rozptýlenosti), aby se předešlo nadměrné koncentraci kolem středních hodnot. Jednotlivci byly zatříděni do domácností a jednotlivé domácnosti následně do budov.

Domácnosti byly seřazeny podle spotřeby plynu normalizované na plochu bytu a rozděleny do decilů. V rámci nich byla provedena frekvenční analýza k vytvoření deseti reprezentativních profilů. Těchto deset skupin bylo následně podrobeno shlukové analýze na základě podobnosti atributů, čímž bylo výsledných deset decilů redukováno na pět klíčových person (profilů). Z nich pak byly identifikovány tři zcela unikátní behaviorální vzorce, zatímco zbylé dva vykazovaly vysokou míru korelace s již definovanými profily.

Součástí analýzy bylo také rozdělení budov do čtyř umělých energetických tříd (A–D) na základě normalizované spotřeby plynu na m². Výsledné rozložení (viz Obr. 20) odhalilo zásadní rozdíly mezi stavebními typologiemi. Zatímco moderní vysoké bytové domy (*apartment*

highriser) a administrativní budovy (*office buildings*) spadají v 60–62 % případů do nejúspornější třídy A, individuální formy bydlení vykazují opačný trend. Samostatně stojící domy (*detached house*), rohové domy (*corner house*) a zejména domy řadové (*terraced house*) se drtivou většinou (64–80 %) nacházejí v nejméně úsporné třídě D.



Obr. 20 Typ budovy v závislosti na energetické třídě; outbuilding = venkovní přístavek, apartment low building = nízké bytové domy, apartment middle = středně-vysoké bytové domy, industrial buildings = průmyslové objekty, semi-detached house = dvojdomek (převzato: HRADEC A KOL. 2022)

Analýza prostorového rozložení ukázala, že nejvyšší energetickou náročnost vykazuje starší nízkopodlažní zástavba v centru města. Směrem k novějším typologiím spotřeba lineárně klesá; středně vysoké bytové domy vykazují přibližně o třetinu nižší spotřebu plynu na m² a u moderních výškových budov je tento rozdíl ještě markantnější (pokles o další třetinu). Specifickou kategorii tvoří (polo)samostatně stojící domy, které sice představují pouze 3,9 % domovního fondu a obývá je 1,4 % populace, avšak díky vysokým příjmům jejich majitelů disponují tito agenti dostatečným kapitálem pro investice do zateplení bez nutnosti čerpání úvěrových programů.

Syntetická populace dále odhalila kritické signály v chování ohrožených skupin. Domácnosti závislé na sociálních dávkách vykazují navzdory delšímu času strávenému v domácnosti o 20 % nižší spotřebu energie, což indikuje přítomnost energetické chudoby a úsporného chování z donucení. Nejnižší spotřebu na obyvatele bez ohledu na typ bydlení vykazují osoby starší 75 let a neúplné rodiny. Na opačné straně spektra stojí populace s vyšším dosaženým vzděláním, u které přímá korelace s vyššími příjmy vede k signifikantně vyšší celkové spotřebě energie. (HRADEC A KOL. 2022)

5.2.4 Závěr

Závěrečná komparace potvrzuje, že deskriptivní statistika (kapitola 5.2.1), ačkoliv je díky své časové a výpočetní nenáročnosti nejčastěji využívaným nástrojem, poskytuje v kontextu energetické politiky často zavádějící výsledky. Agregovaná data totiž maskují individuální extrémy a vedou k plošným řešením, která neřeší podstatu problému.

Analýza cenzálních dat pomocí strojového učení (kapitola 5.2.2) sice nabízí detailní pohled na jednotlivé proměnné, avšak nedokáže zachovat komplexní vztahy a interakce mezi nimi.

V těchto datech se sice podařilo identifikovat určité trendy, jednalo se však pouze o slabé signály, které vyžadovaly další náročnou verifikaci.

Jako nejrobustnější nástroj se profiluje syntetická populace (kapitola 5.2.3). Přestože je založena na pravděpodobnostních odhadech distribucí, poskytuje nejjasnější analytické signály a nejlépe ze všech zmíněných metod vysvětluje kauzální mechanismy v chování jednotlivců. Právě schopnost zachovat strukturní integritu dat a vazby mezi atributy z ní činí klíčový nástroj pro simulaci dopadů politických rozhodnutí na úrovni mikrodat. (HRADEC A KOL. 2022)

5.3 GenSynthPop

Případová studie od Pellegrino a kol. 2023 pokrývá 14 sousedství ve městě Haag (provincie Jižní Holandsko). Autoři se museli vypořádat se situací, kdy některé vstupní tabulky byly dostupné pouze za celou obec či větší administrativní celek, avšak bylo možné je podmínit atributy (tzv. podmiňující atributy), které byly známé i na nižší prostorové úrovni jednotlivých sousedství.

Proces generování syntetické populace primárně vychází z tvorby jednotlivců. V prvním kroku je pro každé sousedství vygenerován soubor datových záznamů (*n-tic*), jehož velikost přesně odpovídá celkovému počtu obyvatel dané územní jednotky. Základním atributem každého záznamu je identifikátor sousedství, ke kterému je jednotlivec prostorově přiřazen. Jelikož v kontingenčních tabulkách plní roli podmiňujících atributů nejčastěji věková skupina a pohlaví, byly tyto dvě charakteristiky k záznamům přiřazeny jako první. Následně byl doplněn i konkrétní věk, a to na základě celoměstské marginální distribuce věku uvnitř daných věkových kohort. Vzhledem k tomu, že pohlaví a věková skupina jsou k dispozici na úrovni sousedství, ale detailnější vazba (pohlaví podmíněné věkovou skupinou) existuje pouze na úrovni města, byl využit algoritmus iterativního proporcionálního přizpůsobování (IPF). Ten nejprve odhadl sdružené rozdělení věku a pohlaví pro celou syntetickou populaci a následně jej aplikoval pro každou čtvrť zvlášť. Na základě těchto proporcí byl každému syntetickému obyvateli přiřazen finální atribut pohlaví. Identický postup byl uplatněn pro přiřazení migračního původu (kde byla zdrojová data dostupná pouze na úrovni obce), přičemž podmiňujícími atributy byly opět věk a pohlaví.

Komplexnější výzvu představovalo modelování vzdělání. Z důvodu předpokládané vzájemné korelace se autoři rozhodli implementovat dva oddělené atributy: současné studium a nejvyšší dosažené vzdělání. K tomuto účelu byly složitě sloučeny tři různé zdrojové tabulky do dvou výsledných kontingenčních tabulek, kde podmiňujícími atributy byly věk, pohlaví a migrační původ. Na závěr profilování jednotlivců byly do záznamu přidány atributy definující vlastnictví řidičského oprávnění (pro automobil, motocykl a moped).

Následná fáze spočívala ve sdružování vygenerovaných syntetických jednotlivců do domácností prostřednictvím navazujícího skriptu. Data o složení domácností (např. sezdané/nesezdané páry s dětmi či bez dětí, domácnosti jednotlivců) byla dostupná na celoměstské

úrovni a byla podmíněna věkem a pohlavím. Počet dětí v domácnosti byl odvozen pomocí rovnice 5.2 a děti byly rozděleny podle počtu sourozenců na základě rovnice 5.1. Na nižší úrovni sousedství byla k dispozici marginální data pouze o třech základních typech domácností (jednotlivci, s dětmi, bez dětí). Tato data následně sloužila jako podmiňující atributy ke stávajícímu věku a pohlaví. Z takto vzniklé kontingenční tabulky byl syntetickému jednotlivci přiřazen atribut „postavení v domácnosti“ (např. rodič z neúplné rodiny se dvěma dětmi, dítě se třemi sourozenci).

Rovnice 5.1 vyjadřuje transformovanou pravděpodobnost, že náhodně vybrané dítě pochází z domácnosti o velikosti *c* dětí:

$$P'(C = c) = \frac{P(C = c) \cdot c}{\sum_{c' \in C} P(C = c') \cdot c'} \qquad 5.1$$

kde:

- $P'(C = c)$ je transformovaná pravděpodobnost, že náhodně vybrané dítě pochází z domácnosti s přesně *c* dětmi,
- $P(C = c)$ je původní pravděpodobnost (relativní frekvence) existence domácnosti s přesně *c* dětmi,
- c je konkrétní počet dětí

Tato rovnice de facto váží původní podíl domácností počtem dětí v nich žijících. Matematicky tak kompenzuje skutečnost, že ačkoliv může být jednodětných a vícečetných rodin v populaci stejný počet, vícečetné rodiny generují do celkového souboru syntetických dětí absolutně vyšší počet jedinců.

Na základě takto transformované pravděpodobnosti algoritmus následně vypočítá přesný počet domácností, které je nutné v prostorovém modelu fyzicky vytvořit, aby bylo možné ubytovat všech *n* syntetických dětí. Tento krok definuje rovnice:

$$h(c) = \left\lfloor \frac{n \cdot P'(C = c)}{c} \right\rfloor \qquad 5.2$$

kde:

- $h(c)$ je cílový počet domácností s přesně *c* dětmi,
- n je celkový počet vygenerovaných syntetických dětí v populaci

Samotné formování domácností probíhalo pomocí párování na základě předem definovaných parametrů vhodnosti (tzv. *fitness*). Algoritmus dbal na to, aby věkový rozdíl mezi sourozenci byl realistický (blízký, avšak nikoliv totožný, nejednalo-li se o dvojčata), partneři byli k sobě přiřazováni s ohledem na věk a pohlaví, a párování rodičů s dětmi respektovalo biologicky možný věkový odstup. Pokud algoritmus vygeneroval přebytek rodičů vůči počtu dětí, byly hůře

hodnocené páry ponechány bezdětné. V opačném případě (nedostatek rodičů) byly zbývající děti alokovány i mezi páry či jednotlivce původně zamýšlené jako bezdětné.

Po zformování rodinných jednotek byl každé domácnosti přiřazen atribut standardizované příjmové skupiny, který byl namodelován pro celé město s podmíněním podle tří základních typů domácnosti. Obdobným způsobem byl připojen údaj o vlastnictví a počtu automobilů (vycházející z dat na národní úrovni), který byl podmíněn příjmovou skupinou a strukturou domácnosti. Proces generování byl završen náhodným přidělením poštovního směrovacího čísla (PSČ) v závislosti na sousedství, do kterého byla domácnost ukotvena.

V metodice autoři důrazně obhajují preferenci využití podmíněných pravděpodobností (relativních četností) namísto reálných absolutních počtů. V praxi totiž dochází k situacím, kdy se agregovaná data pocházející z různých zdrojů či administrativních úrovní (např. data za město vs. data za sousedství) vzájemně rozcházejí a matematicky si odporují. Přímé použití absolutních počtů by v takovém případě vedlo k neřešitelným konfliktům v algoritmu, zatímco převod na relativní pravděpodobnosti umožňuje modelu tyto datové nekonzistence elegantně překlenout. (PELLEGRINO A KOL. 2023)

6 Dostupná vstupní data pro syntetickou populaci v České republice

Před zahájením procesu generování syntetické populace je nezbytné definovat dostupnost a granularitu vstupních informací. Na základě těchto parametrů lze následně zvolit konkrétní metodu generování, případně vybrat již existující model. Tato kapitola představuje klíčové instituce a datové sady, které lze při tvorbě syntetické populace využít. Závěrečná část kapitoly se věnuje zákonu o svobodném přístupu k informacím, který představuje nástroj pro posílení otevřenosti dat nad rámec standardně poskytovaných statistických výstupů.

6.1 Český statistický úřad

Český statistický úřad (ČSÚ) je ústředním orgánem státní správy, jehož úkolem je vytvářet objektivní a komplexní obraz o ekonomickém, sociálním a demografickém vývoji státu. ČSÚ organizuje Sčítání lidu, domů a bytů (SLDB), které představuje nejrozsáhlejší statistické šetření v České republice, a v současnosti je nejvýznamnějším editorem otevřených dat o populaci v tuzemsku.

6.1.1 Sčítání lidu, domů a bytů

Sčítání lidu probíhá na českém území od roku 1869, zpravidla v desetiletých cyklech. Poslední cenzus se uskutečnil na jaře roku 2021, přičemž rozhodným okamžikem byla stanovena půlnoc z 26. na 27. března. Dominantním způsobem sběru dat byly online dotazníky, které využilo přibližně 86 % respondentů. Jednalo se o v pořadí druhé kombinované sčítání v historii ČR, kdy byly vedle dotazníkového šetření v maximální míře využity také administrativní zdroje dat (registry). Cílem zapojení registrů je dosažení úplnějších a kvalitnějších výsledků při současném snížení zátěže respondentů. Zatímco v roce 2011 byl využit pouze Informační systém evidence obyvatel (ISEO), při sčítání v roce 2021 bylo integrováno více než deset různých registrů. Tento vývoj potvrzuje postupný přechod od tradičních metod sčítání přes kombinované až k čistě administrativním, která nevyžadují terénní šetření a spoléhají výhradně na existující databáze státní správy. Vzhledem k vysoké míře pokrytí populace představuje cenzus univerzální a klíčový zdroj dat pro tvorbu syntetických populací. (ŠANDA 2023)

6.1.1.1 SLDB 2011

Při realizaci SLDB 2011 byl v českém prostředí učiněn první významný krok k modelu kombinovaného cenzu skrze integraci Informačního systému evidence obyvatel (ISEO). Administrativní data z tohoto registru, zahrnující primárně jména, příjmení a adresy, sloužila k personalizaci a předvyplnění sčítacích formulářů. Tento postup měl usnadnit práci sčítacím komisařům, kteří následně vyhledávali konkrétní osoby na nahlášených adresách trvalého pobytu. Logistickým úskalím této metody se však stala skutečnost, že značné množství předvyplněných dotazníků zůstalo nevyužito, pokud se adresáti v místě registrace nezdržovali. V případech, kdy

byli respondenti zastiženi mimo nahlášenou adresu, jim byly distribuovány prázdné, nepředvyplněné formuláře.

Metodicky bylo ISEO využíváno jako alternativní a doplňkový zdroj. Ačkoliv sčítací archy obsahovaly otázky na údaje, které již byly v registrech evidovány, v rámci zpracování byla (až na specifické výjimky) prioritizována data vyplněná respondenty v dotazníku. K administrativním zdrojům se přistupovalo teprve v okamžiku, kdy údaj v papírovém archu absentoval, nebo pokud se jednalo o proměnné, které byly z ISEO přebírány přímo bez dotazování. Klíčovým přínosem této integrace byla možnost validace a následného doplnění celkového počtu obyvatel; populace sečtená pomocí dotazníků byla v konečných výstupech rozšířena o osoby evidované v ISEO, které se nepodařilo zkontaktovat v rámci terénního šetření, čímž bylo minimalizováno riziko podhodnocení celkového počtu obyvatel ČR. (ŠANDA 2022)

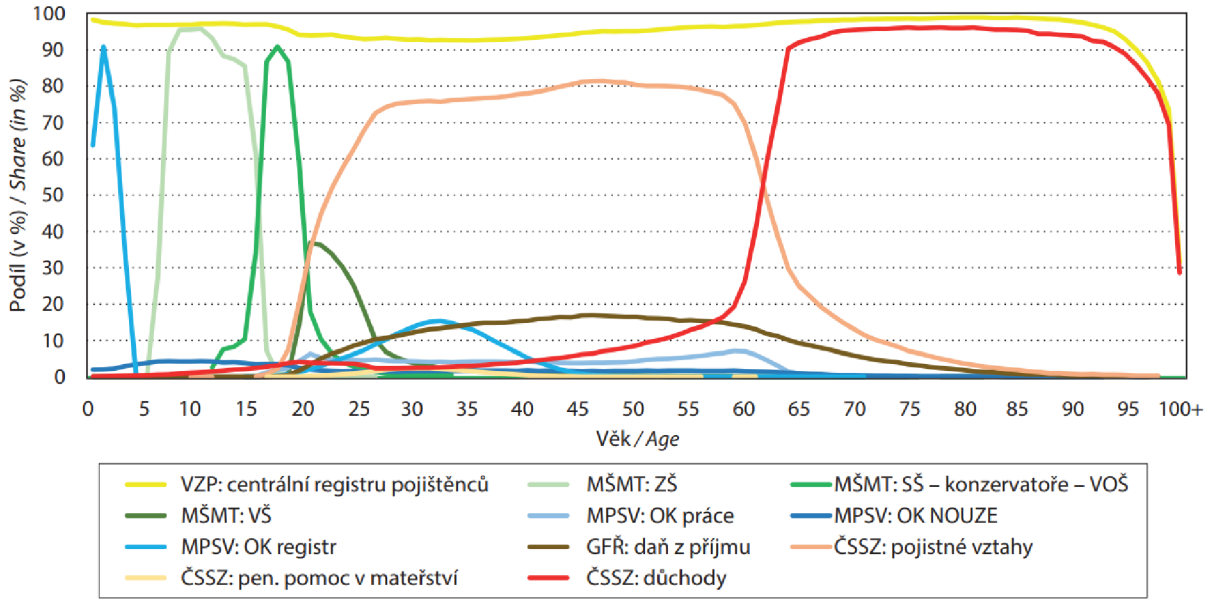
6.1.1.2 SLDB 2021

Sčítání lidu, domů a bytů 2021 (SLDB 2021) představuje prozatím poslední cenzus realizovaný na území České republiky, přičemž jeho rozhodným okamžikem byla půlnoc z 26. na 27. března. Stejně jako v roce 2011 proběhlo šetření kombinovaným způsobem, avšak s výrazně širším zapojením administrativních zdrojů dat. Tento posun byl umožněn zásadními změnami v evropské legislativě, zejména přijetím nařízení Evropského parlamentu a Rady (EU) 2015/759. Tato norma udělila statistickým úřadům právo na bezplatný a rychlý přístup ke všem relevantním administrativním záznamům za účelem tvorby státní statistiky. Vedle tradičních dotazníků tak byly integrovány údaje ze Základního registru obyvatel (ROB), Centrálního registru pojištěnců (CRP), subsystémů České správy sociálního zabezpečení, databází regionálního školství, Sdružených informací matrik studentů (SIMS), systémů Ministerstva práce a sociálních věcí (MPSV) a údajů z daňových přiznání fyzických osob.

Základní registr obyvatel poskytl klíčové identifikační údaje, adresy pobytu a demografické charakteristiky nejen občanů ČR, ale i registrovaných cizinců z EU, EFTA a osob s uděleným azylem či doplňkovou ochranou. Centrální registr pojištěnců, jakožto druhý nejrozsáhlejší zdroj, umožnil prostřednictvím analýzy přerušení zdravotního pojištění identifikovat osoby dlouhodobě pobývající v zahraničí. Z informačního systému sociálního zabezpečení byly využity subsystémy důchodového pojištění, peněžité pomoci v mateřství a pojistných vztahů, což umožnilo přesné zařazení osob do kategorií ekonomické aktivity. Podobně data z rezortu školství posloužila k identifikaci studentů a částečně k verifikaci dosažené úrovně vzdělání. Systémy MPSV (zejména aplikace OK práce, OK nouze a OK registr) poskytly informace o uchazečích o zaměstnání a příjemcích sociálních dávek. Údaje z daňových přiznání byly využity spíše komplementárně, vzhledem k časovému nesouladu dostupných dat za rok 2020.

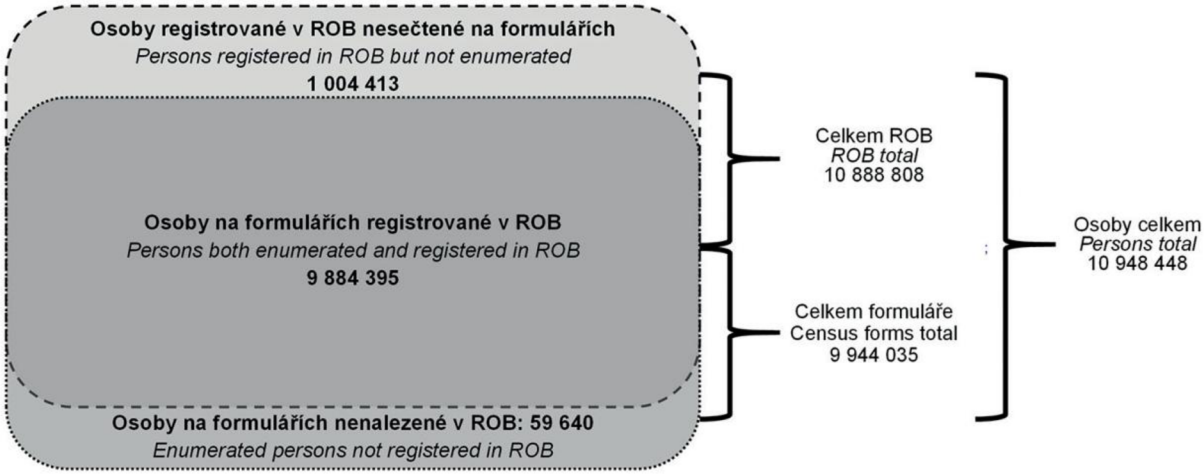
Většina těchto registrů, na rozdíl od ROB a CRP, pokrývá pouze specifické populační skupiny, což se projevuje v rozdílné míře postižení jednotlivých věkových kategorií. Obr. 21 ukazuje procentuální postihnutí věkové skupiny v rámci jednotlivých registrů, které byly využity

při sčítání, oproti ROB (tvoří 100 %). ROB sloužil jako primární referenční rámec, podle něhož byly finální údaje kalibrovány.



Obr. 21 Procentuální podíl postihnutí populace v závislosti na věku ve vybraných registrech v ČR (převzato: ŠANDA 2023)

Zásadním krokem v procesu zpracování bylo párování dotazníků s ROB. Vzhledem k častým chybám v identifikačních údajích vyplněných respondenty byly aplikovány pokročilé deterministické a pravděpodobnostní algoritmy. Potvrdil se dlouhodobý trend klesajícího počtu sečtených osob pomocí formulářů, což zvyšuje význam administrativních zdrojů.



Obr. 22 Schéma sečtení pomocí ROB a sčítacích formulářů (převzato: ŠANDA 2023)

Skupinu osob sečtených na formulářích, ale nenalezených v ROB, představují převážně cizinci, kteří na území státu pobývají, ale nejsou zde přihlášeni ke sčítání. Toto číslo bylo zrevidováno na zhruba 36 000 poté, co byly vyškrtnuty dotazníky osob nesplňujících podmínky pro zařazení do obvykle bydlícího obyvatelstva (např. osoby s neúplnými či chybně vyplněnými

identifikačními údaji). Většina populace byla sečtena prostřednictvím formulářů a zároveň registrována v ROB; necelých 100 tisíc osob však bylo vyřazeno, protože uvedly obvyklý pobyt v zahraničí.

Přibližně milion osob registrovaných v ROB, které se sčítání nezúčastnily, tvoří lidé, kteří v ČR fakticky nežijí (pobývají v zahraničí bez ukončení pobytu v ČR) nebo jde o nezaevidovaná úmrtí. Pro identifikaci osob, které se již na území ČR nenacházejí, byla provedena analýza *„sign-of-life“*, která zkoumala aktivitu daných osob v jiných administrativních registrech. Po provedení této analýzy bylo z 1 004 413 osob do sčítání dodatečně zařazeno 702 214; odstraněné záznamy představovaly tzv. přesah pokrytí (*over-coverage*), který vyjadřuje nesoulad mezi administrativními daty a reálným stavem. Celkový přesah pokrytí činil 3,7 %, což je v mezinárodním srovnání vysoká hodnota.

Míra účasti na sčítání byla územně diferencovaná. Nejvíce nesečtených osob doplňovaných z ROB bylo v severozápadních Čechách (Karlovarský a Ústecký kraj), nejméně naopak v kraji Vysočina. Patrná byla souvislost mezi mírou participace a některými socioekonomickými charakteristikami či religiozitou. Podobný prostorový vzorec jako nesečtení (dodatečně zařazení) vykazuje i skupina sečtených, avšak následně vyřazených osob. V severních Čechách se v tomto případě jednalo především o občany Německa. Kromě severozápadních Čech byla výrazná míra vyřazování zaznamenána také v Brně u cizinců registrovaných na adresách ohlašoven pobytu.

Výsledný počet osob po SLDB 2021 (10 524 167) se lišil od dosavadní intercenzální bilance k 1. 1. 2021 (10 701 777) o přibližně 178 000. Na základě těchto výsledků byl stav populace k počátku roku zrevidován na 10,5 milionu a vzniklý statistický rozdíl („skok“) byl promítnut do revize k začátku roku. (ŠANDA 2023)

6.2 Ministerstva

6.2.1 Ministerstvo dopravy

Ministerstvo dopravy disponuje přístupem k Centrálnímu registru silničních vozidel a eviduje informace o řidičích, řidičských průkazech, přestupcích či trestných činech. Vzhledem k tomu, že se velká část syntetických populací využívá k modelování dopravy, je vhodné zahrnout údaje o řidičských oprávněních, případně informace o dostupnosti vozidla v domácnosti, jako modelované charakteristiky. Úřad na svých webových stránkách pravidelně publikuje statistické přehledy (MINISTERSTVO DOPRAVY 2026A), které mohou sloužit k následné validaci věrohodnosti syntetické populace. Cenným zdrojem jsou rovněž informace poskytnuté na základě žádostí dle zákona č. 106/1999 Sb., o svobodném přístupu k informacím (MINISTERSTVO DOPRAVY 2026B). Tyto veřejně dostupné odpovědi často obsahují specifické a detailní datové sady, které byly úřadem zpracovány nad rámec standardně zveřejňovaných otevřených dat.

6.2.1.1 Česko v pohybu

V letech 2017 až 2019 byl realizován historicky první celorepublikový průzkum domácností zaměřený na dopravní chování obyvatel. Šetření provedlo Centrum dopravního výzkumu, v. v. i. (jehož zřizovatelem je Ministerstvo dopravy). Cílem projektu bylo získat data pro celostátní model poptávky po osobní dopravě a definovat charakteristiky dopravního chování pro potřeby efektivního plánování rozvoje infrastruktury. Výsledkem je výběrový dataset zahrnující přibližně 10 000 domácností.

Sběr dat se zaměřil na čtyři základní jednotky: domácnost jako celek, jednotlivé členy domácnosti, všechna dostupná vozidla v domácnosti a veškeré vykonané cesty během předem určeného rozhodného dne. Informace byly strukturovány do dvou hlavních sekcí: dotazníku pro domácnost (zahrnujícího i socioekonomické údaje za jednotlivce) a cestovního deníku, který vyplňoval každý člen domácnosti samostatně.

Původní výběrový soubor čítal 17 822 adres. Vlivem nezastižení respondentů nebo odmítnutí účasti byl soubor zúžen na 9 419 domácností (čistý výběrový soubor). Z tohoto počtu bylo následně vyřazeno dalších 401 záznamů z důvodu jejich nepoužitelnosti, která vyplývala z nedostatečné kvality či neúplnosti vyplnění. Tab. 8 uvádí strukturu výběrového souboru včetně počtu validních a vyřazených záznamů. (CENTRUM DOPRAVNÍHO VÝZKUMU, V. V. I. 2022)

Tab. 8 Vyhodnocení průzkumu domácností z hlediska použitelnosti (převzato: CENTRUM DOPRAVNÍHO VÝZKUMU, V. V. I. 2022)

Použitelnost	Domácnosti		Automobily		Osoby		Cesty	
	n	%	n	%	n	%	n	%
Použitelné	9 018	95.74	9 095	97.95	21 076	95.27	51 396	99.93
Nepoužitelné	401	4.26	190	2.05	1 046	4.73	38	0.07
Čistý výběrový soubor	9 419	100.00	9 285	100.00	22 122	100.00	51 434	100.00

Data průzkumu byla využita například v případě případové studie v Ústí nad Labem (5.1).

6.2.2 Ministerstvo školství a tělovýchovy

Statistické agendy Ministerstva školství, mládeže a tělovýchovy (MŠMT) spadají do kompetence Odboru školské statistiky a analýz. Tento odbor spravuje klíčové administrativní zdroje, zejména Rejstřík škol a školských zařízení a matriky studentů, jako je systém SIMS (Sdružené informace matrik studentů) pro vysoké školy.

Význam těchto agend zásadně vzrostl při realizaci Sčítání lidu, domů a bytů 2021, které bylo koncipováno jako kombinované sčítání s prioritním využitím administrativních datových zdrojů státu. Právě data z rezortu školství byla Českým statistickým úřadem (ČSÚ) využita jako jeden z primárních pilířů pro vymezení obvykle bydlícího obyvatelstva a validaci jeho charakteristik. Propojení sčítacích formulářů s registry MŠMT umožnilo zpřesnit údaje o nejvyšším dosaženém vzdělání a aktuální školní docházce, což eliminovalo nepřesnosti vznikající při subjektivním vyplňování dotazníků respondenty.

Data z matrik MŠMT lze efektivně využít při generování syntetické populace, a to zejména pro přesné určení ekonomické aktivity (identifikace statusu studenta) nebo pro detailnější modelování prostorového rozložení žáků a studentů v rámci školské sítě. (MŠMT ČR 2025)

6.2.3 Ministerstvo práce a sociálních věcí

Ministerstvo práce a sociálních věcí (MPSV) monitoruje a analyzuje trh práce z hlediska nabídky, poptávky, ceny práce a struktury uchazečů o zaměstnání, včetně specifických skupin, jako jsou absolventi škol. Vzhledem k podrobnému sledování nezaměstnanosti představují agendy ministerstva významný zdroj dat pro modelování ekonomické aktivity jednotlivců v rámci syntetické populace. Úřad tato data publikuje prostřednictvím svého datového portálu, který nabízí detailní statistiky nezaměstnanosti (MINISTERSTVO PRÁCE A SOCIÁLNÍCH VĚCÍ 2026A) v čase i specifické pololetní přehledy o situaci absolventů na trhu práce (MINISTERSTVO PRÁCE A SOCIÁLNÍCH VĚCÍ 2026B). Tyto výstupy umožňují přesnější kalibraci pravděpodobnostních modelů přiřazování zaměstnanosti v rámci mikrodat.

6.2.4 Ministerstvo zdravotnictví

Ministerstvo zdravotnictví spolu s Ústavem zdravotnických informací a statistiky ČR (ÚZIS) spravuje Národní zdravotnický informační portál (NZIP). Na tomto portálu jsou publikována data z různých lékařských oborů a o poskytování zdravotní péče. Významným zdrojem jsou záznamy o přirozeném pohybu obyvatel, které lze využít k predikci budoucí velikosti a struktury populace. (ÚZIS 2025A)

Vzhledem k vysoké citlivosti spravovaných údajů nemůže ÚZIS poskytovat primární data odborné veřejnosti v jejich surové podobě. Z tohoto důvodu jsou generována syntetická data, která svou strukturou věrně odpovídají datům primárním. Tento specifický přístup umožňuje analytikům připravit výpočetní skripty nad těmito syntetickými vzorky; po jejich zaslání do ÚZIS je možné danou analýzu provést přímo nad primárními daty a tazateli vrátit pouze výsledné validované výstupy. (ÚZIS 2025B)

6.3 Lokální průzkumy

Vedle plošných šetření a celostátních administrativních registrů představují významný zdroj dat pro tvorbu a kalibraci syntetických populací také specializované lokální průzkumy. Ty jsou obvykle realizovány samosprávami, městskými organizacemi či univerzitními pracovišti a zaměřují se na detailní zmapování specifických vzorců chování v konkrétním území. Ačkoliv tyto průzkumy disponují menším vzorkem respondentů oproti národním databázím, jejich přínos spočívá ve vysoké prostorové granularitě a schopnosti podchytit lokální anomálie a specifika, která agregovaná národní data nedokážou zohlednit.

Příkladem takového šetření je rozsáhlý Průzkum dopravního chování v Brně a okolí, který v roce 2022 iniciovala Kancelář architekta města Brna (KAM) společně s Odborem dopravy Magistrátu města Brna. Cílem tohoto průzkumu bylo získat detailní a aktuální informace o mobilitních návycích obyvatel brněnské metropolitní oblasti (Brněnské aglomerace), nezbytné pro strategické plánování dopravy, nastavení parkovacích politik a rozvoj veřejného prostoru. (KAM BRNO 2022)

6.4 Zákon o svobodném přístupu k informacím

Zákon č. 106/1999 Sb., o svobodném přístupu k informacím, představuje klíčový právní rámec pro transparentnost veřejné správy v České republice. Do českého právního řádu implementuje relevantní normy Evropské unie a definuje tzv. povinné subjekty – zejména orgány státní správy, územní samosprávné celky a veřejné instituce – přičemž se vztahuje i na další subjekty rozhodující o právech či povinnostech osob ve veřejnoprávní sféře. Zákon však výslovně vylučuje poskytování údajů, jejichž režim je upraven zvláštními právními předpisy (např. centrální evidence účtů či ochrana průmyslového vlastnictví). Právo na informace se rovněž nevztahuje na dotazy na názory, budoucí rozhodnutí ani na povinnost vytvářet nové informace (tzv. analýzy na zakázku).

Důležitým prvkem je definice veřejného podniku jako právnické osoby ovládané povinným subjektem, která vykonává činnosti ve veřejném zájmu (např. v oblasti veřejné dopravy či služeb), i když sama není státní institucí. Celkově zákon č. 106/1999 Sb. posiluje princip otevřenosti, zakotvuje procesní nástroje pro žadatele a definuje jasné limity pro odmítnutí žádosti, čímž tvoří základní pilíř ochrany veřejnosti před netransparentním rozhodováním. (OPEN AI 2025B)

7 Prostorová data využitelná pro syntetickou populaci

Základním předpokladem pro praktické využití syntetické populace v dopravním či urbanistickém modelování je její přesné prostorové ukotvení. Zatímco předchozí kapitola se věnovala generování demografických a socioekonomických atributů (tedy definování kdo jsou aktéři modelu), tato kapitola se zaměřuje na prostorovou dimenzi (tedy kde tito aktéři žijí), případně jak je možné alokovat vygenerované syntetické domácnosti na konkrétní adresní body a stavební objekty v rámci zájmového území.

7.1 Datový portál GIS

Zásadním posunem v dostupnosti prostorových informací pro demografické a urbanistické modelování se stalo nedávné rozšíření Statistického geoportálu Českého statistického úřadu (ČSÚ) o tzv. Datový portál GIS. Tato platforma, spuštěná v rámci rozsáhlejší digitalizace a tvorby nových otevřených datových sad, představuje centralizovaný katalog geografických dat (Open Data) přímo navázaných na oficiální statistiky.

Z metodického i praktického hlediska přináší tento portál výrazné zjednodušení práce s prostorovými daty. Umožňuje interaktivní prohlížení prostorových sad v mapovém klientovi, a především jejich přímé stahování ve standardizovaných a strojově čitelných formátech (např. Esri Shapefile, GeoJSON či GeoPackage). Pro účely prostorového ukotvení syntetické populace představuje tento katalog klíčový a vysoce spolehlivý zdroj.

Datový portál GIS poskytuje nejen přesné vektorové vrstvy administrativního a statistického členění území (hranice městských částí, základní sídelní jednotky, sčítací obvody či územně technické jednotky), ale umožňuje i provázání těchto polygonových vrstev s konkrétními agregovanými výsledky Sčítání lidu, domů a bytů (SLDB 2021). Přístup k datům formou Open Data tak eliminuje dřívější nutnost složitého dolování informací z nesourodých tabulek a výrazně zefektivňuje integraci demografických údajů do prostředí geografických informačních systémů (GIS) za účelem další analytické práce.

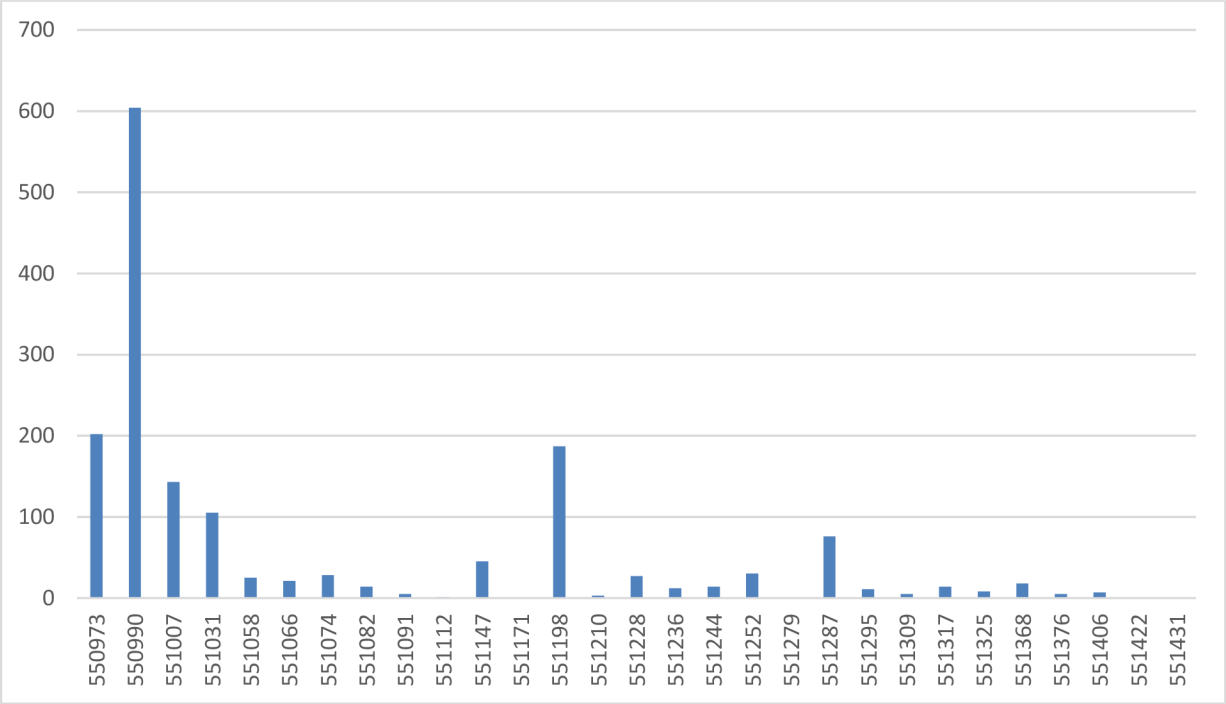
7.1.1 Vchody do budovy s TEP (RSO)

Tato datová sada, referující ke stavu k 1. 1. 2026, poskytuje v mnoha ohledech aktuálnější informace než stávající lokální registry města Brna. Provedená datová analýza však odhalila několik logických a strukturálních anomálií.

U některých objektů nabýval identifikátor RÚIAN nulové hodnoty. V rámci zájmového území Brna se jednalo například o rozsáhlý areál Fakultní nemocnice u sv. Anny či o neidentifikovanou stavbu v Králově Poli (bez přiřazeného čísla popisného i orientačního). Dále byly v sadě detekovány objekty vykazující nulový počet podlaží. Namátkovou prostorovou kontrolou byla potvrzena chyba například u objektu na ulici Třešňová 331 byl zjištěn markantní nesoulad, kdy databáze chybně evidovala 16 podlaží namísto skutečných 4.

Významným problémem byla rovněž zjištěná demografická nekonzistence. Celkový počet obyvatel alokovaných k budovám měl teoreticky přesně korespondovat se zdrojovými tabulkami syntetické populace, neboť oba zdroje vycházejí z téhož cenzu. Při porovnání byl však detekován deficit 1 610 prostorově umístěných obyvatel (Obr. 23). Ačkoliv evidovaný počet bytů v této sadě převyšoval počet generovaných domácností, vzhledem k průměrně menší velikosti syntetických rodinných jednotek nebyl tento rozdíl klasifikován jako kritický parametr.

Jelikož zdrojová data neobsahovala atribut příslušnosti k městské části, bylo nutné provést prostorové sjednocení (spatial join) s polygonovou vrstvou brněnských městských částí. I v tomto kroku byla identifikována topologická nepřesnost, jeden záznam z geodat ČSÚ se nacházel zcela mimo vymezené polygony Brna. Objekt, který vykazoval nulový počet obyvatel, byl v datech ponechán, avšak algoritmicky k němu nebyly přiřazeny žádné syntetické domácnosti.



Obr. 23 Počet nelokalizovaných obyvatel v jednotlivých městských částech Brna

7.2 Lokální datové portály

Lokální datové portály, vzhledem ke svému přirozenému zaměření na konkrétní administrativní území, disponují výrazně menším počtem jednotlivých záznamů než ty národní. Tato prostorová redukce dat pro uživatele představuje zásadní metodickou výhodu v podobě vyšší relevance informací a mnohem efektivnější orientace v databázi. Z tohoto důvodu byla jako primární zdroj pro praktickou část diplomové práce zvolena platforma data.brno.cz.

Budování vlastních specializovaných katalogů otevřených dat (Open Data) přitom není výsadou pouze statutárního města Brna; podobnými platformami disponují i další velká města (např. Praha, Ostrava či Plzeň). Tento trend v decentralizovaném publikování dat se v posledních

letech úspěšně rozšiřuje i na úroveň krajských samospráv, což významně přispívá k transparentnosti a lepší dostupnosti regionálních informací pro odbornou i laickou veřejnost.

7.2.1 Průzkum budov v Brně

Tato polygonová vrstva představuje výsledky historicky prvního plošného průzkumu budov realizovaného v letech 2018–2020. Atributová tabulka detailně zachycuje počet podlaží a primární využití pro jednotlivá patra (od 1. do 28. podlaží). Míra detailu postupně klesá s výškou budovy: pro 1. patro lze evidovat až 5 různých funkcí, pro 2. patro maximálně 3 funkce a pro každé další vyšší podlaží pouze jedinou funkci.

Zvolený databázový model této sady maximalizuje deskriptivní detail na úkor efektivity údržby, datové velikosti a logické konzistence. Výsledkem je naddimenzovaná atributová tabulka s enormním podílem prázdných buněk. Extrémním příkladem je sloupec definující primární využití 28. patra, kterým v celé datové sadě disponují pouze dvě části téže budovy (se shodným identifikátorem RÚIAN).

Tento formát značně komplikuje standardní analytickou práci. Manuální filtrace v prostředí ArcGIS Pro, například vyhledání budov disponujících alespoň jedním podlažím s obytnou funkcí, se ukázala jako uživatelsky vysoce neefektivní kvůli systémovému limitu přednastaveného filtrování (omezení na max. 5 podmínek). (STATUTÁRNÍ MĚSTO BRNO 2021)

8 Generování syntetické populace

Hlavními požadavky na výběr vhodného modelu byla jeho veřejná dostupnost prostřednictvím open-source repozitáře, schopnost generování populace bez nutnosti vstupního mikrodatového vzorku (tzv. sample-free přístup), aktuálnost zdrojového kódu a implementace v jazyce Python, případně přítomnost grafického uživatelského rozhraní (GUI). Dále bylo žádoucí tematické zaměření korespondující s projektem AMOS a dostupnost vzorových dat pro prvotní testování.

V rámci rešerše byly hodnoceny a prakticky testovány tři konkrétní open-source modely, avšak u všech se projevily zásadní technické či metodické překážky. Prvním testovaným nástrojem byl model UrbanSim (MAURER 2020), u kterého se objevily nepřekonatelné problémy se softwarovými závislostmi. Model tak nebylo možné úspěšně nainstalovat ani spustit prostřednictvím standardního správce balíčků v jazyce Python. Podobně problematické bylo nasazení modelu SPENSER (LOMAX, ARCHER 2020), kde sestavení funkčního virtuálního prostředí znemožnily zastaralé softwarové knihovny. Samotný repozitář na platformě GitHub navíc vykazuje známky dlouhodobé neaktivity, o čemž svědčí i přítomnost nevyřešených problémů z roku 2020. Třetí hodnocená platforma, SimMobility (MASSACHUSETTS INSTITUTE OF TECHNOLOGY 2022), kombinuje jazyky Python a C++, což při lokálním nasazení vyžaduje náročnou kompilaci zdrojových kódů. Pro účely prvotního testování navíc chyběla vzorová testovací data a k dispozici byla pouze dokumentace popisující jejich požadovanou strukturu.

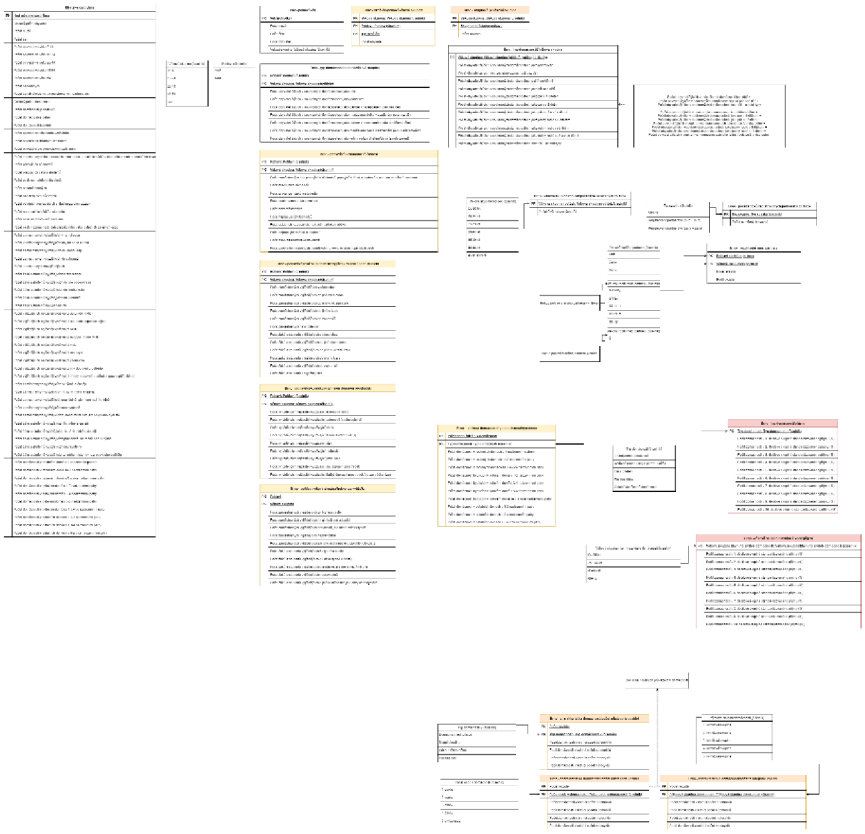
Zásadním zjištěním při bližším metodickém průzkumu však bylo, že všechny výše jmenované modely ve svém jádru využívají algoritmy, které ke svému běhu vyžadují vstupní vzorek mikrodat. Přestože by teoreticky bylo možné o příslušná anonymizovaná mikrodata zažádat Český statistický úřad (ČSÚ), v praxi je tento postup značně problematický. Z důvodu striktní ochrany soukromí obyvatel a přísných interních předpisů přistupuje úřad k poskytování detailních mikrodatových sad externím subjektům s velkou zdrženlivostí a vysoce restriktivně. Tento fakt definitivně vyřadil zmíněné modely z výběru a potvrdil nutnost nalezení striktně „sample-free“ generátoru, který je nezávislý na obtížně dostupné institucionální podpoře.

Na základě výše zmíněných technických a datových limitů existujících platforem byl pro účely této práce nakonec zvolen otevřený generátor syntetické populace vyvinutý Pellegrino a kol. 2023, který autoři představili pod názvem GenSynthPop (kapitola 5.3). Výběr tohoto metodického přístupu byl motivován třemi klíčovými faktory. Prvním z nich je striktně sample-free charakter modelu, který umožňuje generování populace výhradně pomocí agregovaných kontingenčních tabulek a marginálních distribucí, čímž zcela eliminuje závislost na mikrodatech z ČSÚ. Druhým faktorem je jeho původní tematické zaměření – model byl primárně vyvinut v rámci projektu zkoumajícího mobilitu obyvatel a volbu dopravního prostředku (konkrétně v kontextu budoucí autobusové dopravy), což koresponduje s cíli projektu AMOS. Třetím a neméně důležitým důvodem je architektura modelu, která umožňuje flexibilně kombinovat vstupní data o různé prostorové granularitě (např. propojení dat z celoměstské úrovně s daty za jednotlivá sousedství).

Čtvrtým, vysoce pragmatickým benefitem je pak celková jednoduchost přístupu k nástroji a aktivní dosažitelnost samotných autorů modelu. Možnost přímé komunikace a absence zbytečných technických bariér představují výraznou přidanou hodnotu při řešení implementačních výzev, což tvoří ostrý kontrast s dříve testovanými repozitáři. S tím se pojí i využití nejrozšířenějšího programovacího jazyka, Pythonu.

8.1 Metodika

V souladu s tematickým a prostorovým vymezením projektu AMOS bylo jako zájmové území definováno statutární město Brno a jeho městské části. Na základě původní struktury modelu byly namapovány a následně modifikovány vstupní datové sady (Obr. 24) tak, aby vyhovovaly specifickým potřebám projektu. Do struktury dat byly nově integrovány atributy ekonomické aktivity a vyjížďky (cílové místo, frekvence, zvolený dopravní prostředek). Na úrovni domácností byla přidána charakteristika obydlí (typ budovy a podlažnost), jejímž účelem je umožnit následné párování syntetické populace s datovými sadami z průzkumu budov a s údaji o počtu osob na adresních místech, jež poskytuje datový portál města Brna. Naopak byl z původního modelu odstraněn atribut „migrační pozadí“, a to z důvodu vysoké obtížnosti jeho mapování na česká data. V rámci SLDB je totiž možné na otázky týkající se národnosti či mateřského jazyka uvést i dvě hodnoty současně, což by představovalo zásadní komplikaci při agregaci a tvorbě jednoznačných kategorií. Z důvodu vysoké datové a výpočetní náročnosti byl rovněž odebrán atribut „současné studium“. Výsledné schéma požadovaných dat bylo předloženo Českému statistickému úřadu (ČSÚ), který následně poskytnul agregované kontingenční tabulky, primárně ze zdrojů SLDB, v požadovaném formátu. Úřad však nebyl schopný neposkytnout údaje o vlastnictví specifických typů řidičských oprávnění, počtu vozidel v domácnosti, příjmech domácností a počtech sňatků či registrovaných partnerství, které se v původním modelu vyskytovaly a byly relevantní pro účely projektu. (ČSÚ 2021B, Příl. 2)



Obr. 24 Náhled na namapované vstupy, (zdroj dat: PELLEGRINO, MOOIJ 2024), více na Vstupní tabulky 2026 nebo v Příl. 1

Absence datové sady pro sňatky a registrovaná partnerství si vyžádala doplnění těchto informací z externích zdrojů do původní (nizozemské) struktury tabulky. Údaje o registrovaných partnerstvích nejsou standardně monitorovány v intercenzální bilanci ČSÚ (k jejich zavedení dochází až v aktuálním roce), proto byla využita tisková zpráva Sociologického ústavu Akademie věd ČR z roku 2020 (SOCIOLOGICKÝ ÚSTAV AKADEMIE VĚD ČR 2020). Ta sumarizuje celorepublikové počty uzavřených svazků za období 2006–2019. Pro účely modelu byla využita data za rok 2019, neboť novější údaje bližší referenčnímu roku sčítání nebyly k dispozici. Z důvodu zachování časové konzistence byly pro stejný rok (2019) převzaty z ČSÚ celorepublikové statistiky o sňatečnosti (ČSÚ 2026A).

Většina poskytnutých tabulek s podmíněnými distribucemi neobsahovala přímé textové hodnoty, nýbrž kódové označení dle číselníků dostupných v dedikovaném sešitu MS Excel. Pro zajištění strojové čitelnosti byly jednotlivé listy číselníků vyexportovány do formátu .csv. Následná dešifrace (překlad) kódů probíhala automatizovaně prostřednictvím vlastního skriptu, který byl integrován přímo do výpočetního běhu modelu.

V rámci přípravy modelu byly rovněž modifikovány funkce pro čtení souhrnných údajů a pro zpracování podmíněných distribucí (věk, věková skupina, pohlaví). Prostřednictvím těchto funkcí byly každému záznamu (jednotlivci) přiřazeny čtyři základní prostorově-demografické atributy:

městská část, věková skupina, pohlaví a konkrétní věk. Členění do věkových skupin bylo zachováno v identické podobě jako v původní nizozemské případové studii (Tab. 9).

Tab. 9 Věkové skupiny

Věková skupina	Popis
0-14	děti, typicky bez vzdělání a zaměstnání
15-24	mládež a mladí dospělí, typicky ukončení studia
25-44	střední dospělost, typicky reprodukční věk
45-64	zralá dospělost
65+	důchodový věk

Zpracování atributu nejvyššího dosaženého vzdělání vyžadovalo sjednocení rozdílné granularity vstupních dat. Zatímco podmíněné distribuce definovaly vzdělání v detailních podskupinách (Tab. 10), marginální součty operovaly pouze se základním dělením. Pro zachování datové konzistence bylo proto primárně aplikováno agregované členění skládající se ze čtyř nadřazených kategorií: nezjištěno, bez vzdělání či základní, střední a vysokoškolské.

Za účelem bezproblémového propojení byl původní datový číselník rozšířen o nový mapovací sloupec, který tyto úrovně propojoval. Do profilu syntetických obyvatel byla v prvním kroku přiřazena tato agregovaná (hrubá) kategorie vzdělání. Následně byl ke každému záznamu, s podmíněním na již přidělenou nadřazenou kategorii, připojen druhý atribut, který specifikoval detailní stupeň vzdělání přesně dle originální klasifikace SLDB.

Tab. 10 Číselník nejvyššího dosaženého vzdělaní (zdroj: ČSÚ 2021B, Příl. 2)

Kód	Text	Kategorie
1	bez vzdělání	Žádné nebo základní
2	nedokončené základní vzdělání	Žádné nebo základní
3	základní vzdělání	Žádné nebo základní
4	střední nebo vyučení (bez maturity)	Středoškolské
5	úplné střední všeobecné (s maturitou)	Středoškolské
6	úplné střední odborné (s maturitou)	Středoškolské
7	nástavbové/zkrácené studium, absolvování dvou a více oborů středních škol	Středoškolské
8	pomaturitní studium	Středoškolské
9	konzervatoř	Středoškolské
10	vyšší odborné vzdělání	Středoškolské
11	bakalářské	Vysokoškolské
12	magisterské	Vysokoškolské
13	doktorské	Vysokoškolské
88	nedefinováno - osoby ve věku 0–14 let	Žádné nebo základní
99	nezjištěno	Nedefinováno

Následně byl vytvořen skript pro čtení a přiřazování atributu ekonomické aktivity. I v tomto kroku byly kategorie dostupné v marginálních a podmíněných datech nejprve porovnány. Bylo

zjištěno, že v marginálních datech byly kategorie „děti předškolního věku“ a „ostatní závislí“ sloučeny do jedné agregované skupiny (která byla dostupná v datech podmíněných). Tato datová nekonzistence byla vyřešena přidáním nového mapovacího sloupce do číselníku, obdobně jako při zpracování atributu vzdělání.

Pro definování vyjížďky bylo nutné vytvořit odvozené pomocné atributy vycházející z hodnoty ekonomické aktivity. Místo vyjížďky a její frekvence totiž nebyly podmíněny pouze pohlavím a věkovou skupinou, ale právě i typem ekonomické aktivity. Z tohoto důvodu byly primární kategorie ekonomické aktivity reklasifikovány do čtyř základních skupin: pracující, studující, nezjištěno a ostatní. Volba dopravního prostředku byla taktéž podmíněna ekonomickou aktivitou, avšak vstupní data nerozlišovala mezi dojížďkou do zaměstnání a do školy; proto byl vytvořen druhý, specifičtější pomocný atribut.

Zpracování atributu „místo vyjížďky“ vyžadovalo rozdělení dat do dvou samostatných tabulek: pro pracující a pro studující. Do každé tabulky byl implementován nový sloupec, jehož název korespondoval s příslušným pomocným atributem (s hodnotou „škola“ pro tabulku studujících a „práce“ pro tabulku pracujících). Následně byly obě tabulky sloučeny zpět do jednoho celku. Tento dataset byl vzápětí obohacen o údaje z podmíněných distribucí ekonomické aktivity, zahrnující skupinu ostatní (nepracující senioři, děti předškolního věku, nezaměstnaní atd.) a jednotlivce s nezjištěnou ekonomickou aktivitou. Osobám ze skupiny ostatní byla přiřazena hodnota definující absenci dojížďky (nepohybují se), zatímco osobám s nezjištěnou ekonomickou aktivitou byla explicitně přiřazena hodnota „nezjištěno“ pro umožnění případného dalšího datového zpracování. Výsledná sloučená tabulka tak pokrývala 100 % obyvatel Brna a obsahovala jejich věkovou skupinu, pohlaví, pomocný atribut ekonomické aktivity, místo vyjížďky a počet obyvatel v dané kategorii. Pro kalibraci modelu byly jako marginální údaje použity upravené atributy o dojížďce doplněné o vybrané kategorie ekonomické aktivity tak, aby se jejich celkový součet přesně shodoval s podmíněnými distribucemi.

Obdobným algoritmickým postupem byly zpracovány i dílčí zdrojové tabulky definující frekvenci vyjížďky. Podmíněné distribuce byly sloučeny a doplněny o celkovou tabulku podmíněných součtů, přičemž byly definovány jak příslušné hodnoty frekvence, tak marginálie vázané na ekonomickou aktivitu. Téměř identický postup byl aplikován i u podmíněné distribuce dopravního prostředku; vzhledem k chybějícímu rozlišení mezi pracujícími a studujícími byl opět využit druhý odvozený pomocný atribut ekonomické aktivity.

Přiřazování pozice jednotlivce v domácnosti bylo z výpočetních důvodů částečně předsunuto mimo hlavní model. Pomocí samostatného skriptu v prostředí Jupyter Notebook byla nejprve ze dvou dílčích vstupních sad vygenerována jedna komplexní podmíněná distribuce. Dále byly na základě shodných kódů městských částí sloučeny tabulky okrajových součtů za jednotlivce a domácnosti. V rámci samotného běhu hlavního modelu byl následně syntetickým jednotlivcům přidán atribut pětileté věkové skupiny a finální pozice v domácnosti byla přiřazena na základě

výstupů z předzpracování v Jupyter Notebooku a definovaných okrajových součtů typů domácností.

V dalším kroku byly jednotlivci spojeni v domácnosti na základě věku, pohlaví a pozice domácnosti. A následně přiřazen typ budovy a rozsahem podlažím na základě typu domácnosti.

8.2 Identifikace a řešení chyb (tvorba syntetické populace)

Vzhledem k rozdílné povaze původních nizozemských dat od nových českých vznikaly chyby při generování. Pro ladění skriptů bylo využito AI.

8.2.1 Logické nedostatky vstupních datových sad

Při detailní analýze vstupních dat agregovaných Českým statistickým úřadem (ČSÚ) byly zjištěny logické nedostatky, které se týkaly především věkové struktury matek a věku uživatelů u atributu hlavní dopravní prostředek (Tab. 11).

Tab. 11 Výňatek ze vstupní tabulky závislosti dopravního prostředku na pohlaví a věkové skupině (zdroj: ČSÚ 2021B, Příl. 2)

Věková skupina	Hlavní dopravní prostředek	Pohlaví	Počet
0-14	automobil - řidič	muž	31
0-14	automobil - řidič	žena	29

Na území ČR bylo v době konání sčítání (2021) možné legálně řídit motorové vozidlo skupiny „automobil“ nejdříve od 16 let (specificky vozidla přestavěná na tříkolky, např. Fiat 500 Ellenator), případně standardní osobní automobil od 18 let. V poskytnutých datech se nicméně vyskytovalo 60 řidičů (31 mužů a 29 žen) spadajících do věkové skupiny 0–14 let. Korekci této datové anomálie bylo nutné algoritmicky ošetřit před samotným spuštěním modelu. Z časových důvodů nebyl ČSÚ zpětně kontaktován se žádostí o doplňující informace a oprava proběhla v rámci zpracování dat. Eliminace neexistující skupiny 15letých řidičů nebyla vzhledem k agregaci dat nutná, jelikož kohorta končila 14. rokem.

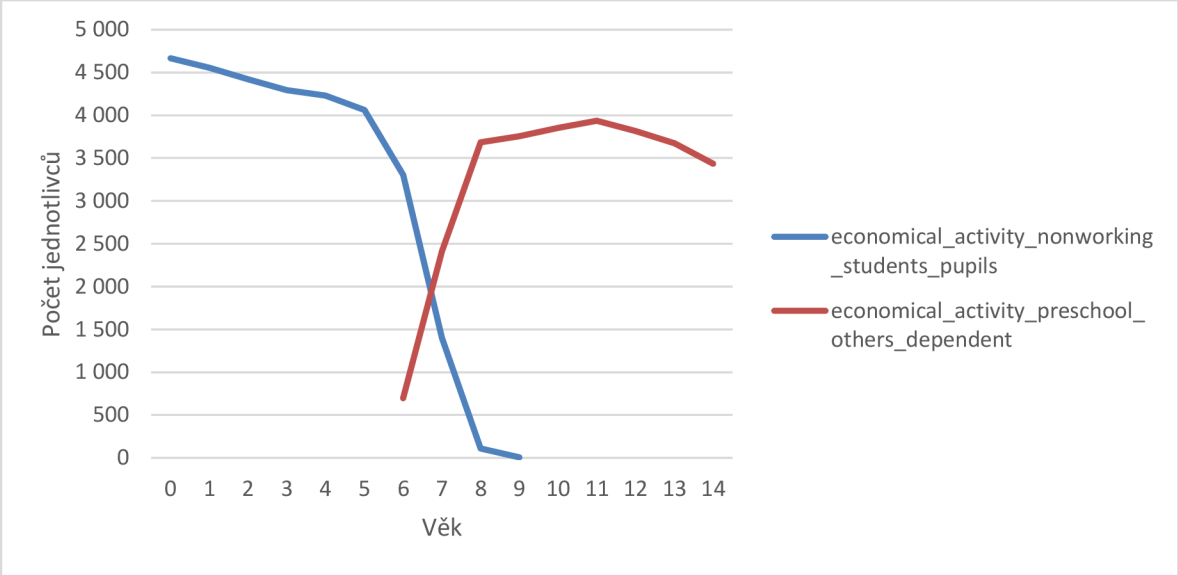
Tab. 12 Počet živě narozených v závislosti na věkové skupině matky v celé ČR v roce 2021 (zdroj: ČSÚ 2021B, Příl.2)

Věková skupina matky	Počet živě narozených	Procentuální podíl
10-14	17	0.015
15-19	1 923	1.720
20-24	10 954	9.798
25-29	33 325	29.810
30-34	40 478	36.208
35-39	20 268	18.130
40-44	4 469	3.998
45-49	348	0.311
50-54	9	0.008
55-59	2	0.002

Další zjištěná nesrovnalost se týkala dat o porodnosti. Podle vstupních údajů se v roce 2021 narodilo 17 dětí matkám ve věkové skupině 10–14 let (viz Tab. 12). Z tohoto důvodu bylo nutné přímo zasáhnout do funkce, která přiřazuje jednotlivce do rodinných domácností (kapitola 8.2.7).

8.2.2 Přidávání ekonomické aktivity ve věkové skupině 0-14 let

K výrazné datové nekonzistenci došlo při mapování ekonomické aktivity u nejmladší populace. V rámci věkové kategorie 0–14 let se ve vstupních datech vyskytovaly dva odlišné typy: nepracující studenti a žáci a děti předškolního věku a ostatní závislí (Obr. 25).

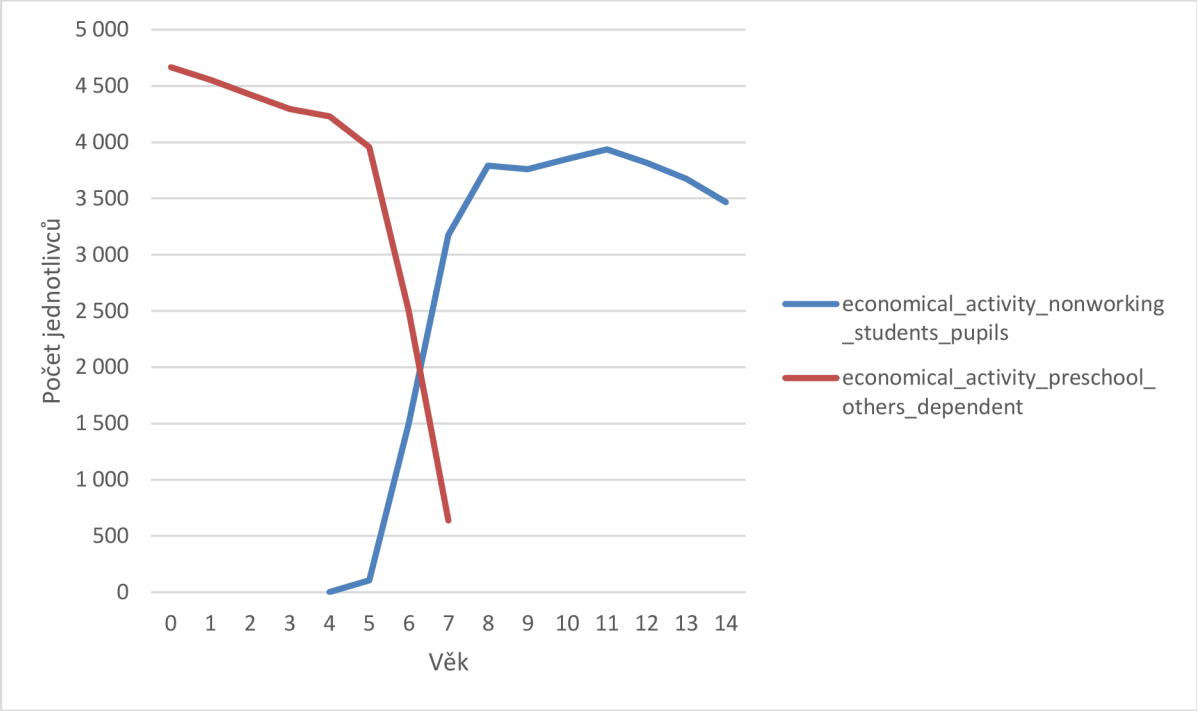


Obr. 25 Rozložení skupin nepracujících studentů a žáků (modře) a předškolních dětí a ostatních závislých (červeně) ve věkové skupině 0-14

Pro vyřešení tohoto rozporu se nabízela tři možná řešení:

1. Sloučení obou skupin do zcela nového typu ekonomické aktivity.
2. Relokace typů v rámci věkové skupiny odděleně pro jednotlivé městské části (nejvhodnější přístup bez přímého zásahu do celkových dat).
3. Pevné věkové vymezení (0–4 roky jako předškolní věk, 8–14 let jako žák a hodnoty pro 5–7 let dopočítané podle marginálií).

Jako metodicky nejméně invazivní byl zvolen přístup č. 2. Výpočetní rozdíl mezi odhadovanou pozicí v kategorii (získanou odečtením předpokládaných marginálií) a reálným počtem v syntetické populaci se u jednotlivých městských částí blížil nule (s maximální odchylkou 2 osoby). Příznivým vedlejším efektem tohoto postupu byla kompletní eliminace neznámých hodnot atributu ekonomické aktivity v této věkové skupině. Z Obr. 26 zobrazujícího výsledek funkce je patrné, že v syntetické populaci existují i 4letí a 5letí žáci. Byť je to v reálné populaci málo pravděpodobné, ukazuje to, že aplikace přístupu č. 3 by byla příliš invazivní, uměle by ovlivnila výsledky modelu a zkreslila statistiky úspěšnosti.

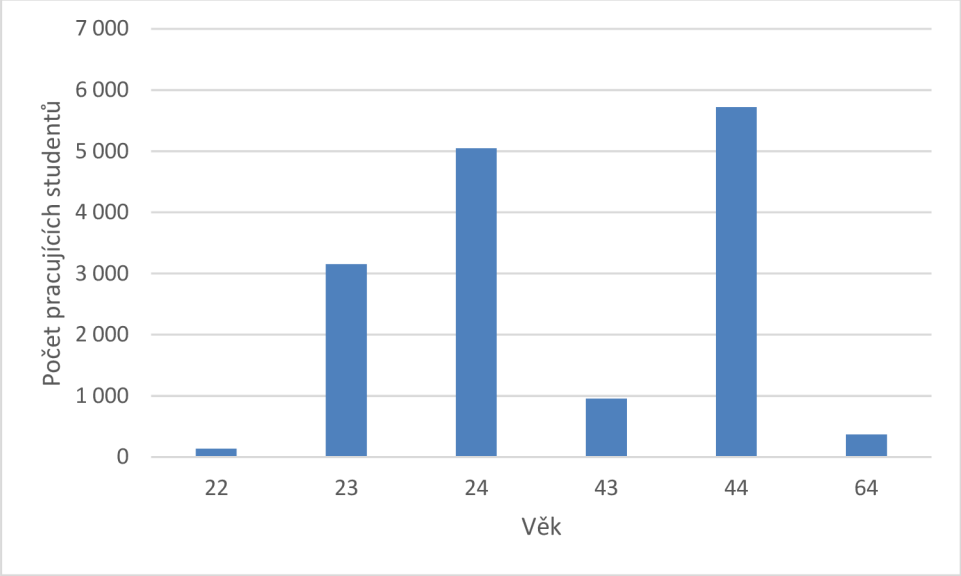


Obr. 26 Rozložení nepracujících studentů a žáků a předškolních dětí a ostatních závislých ve věkové skupině 0-14 po korekci

Obdobným způsobem byla navržena korekční funkce pro atribut hlavního dopravního prostředku, která řešila přítomnost 60 nezletilých řidičů automobilu (ve věku 0–14 let). Model identifikoval nejstarší jednotlivce v této skupině (14 let) se zachováním pohlaví a příslušnosti k městské části. Způsob dopravy byl následně algoritmicky prohozen mezi původně vyfiltrovaným (chybným) záznamem a tímto 14letým jedincem. Tuto korekci bylo nezbytné aplikovat až v samotném závěru po eliminaci nezjištěných hodnot, aby nedošlo k situaci, kdy by byl způsob dopravy „řidič automobilu“ náhodně přiřazen jedinci ve věku 0–14 let v rámci nahrazování hodnoty „nezjištěno“ (viz kapitola 8.2.6).

8.2.3 Proces přidávání atributů do syntetické populace

Během generování bylo pozorováno nepřirozené shlukování vlastností, například všichni pracující studenti byli přiřazeni jen k velmi úzkému spektru konkrétních věkových ročníků (Obr. 27). Pro zajištění rovnoměrnější distribuce napříč celou populací byl do algoritmu zaveden proces automatického náhodného míchání (*shuffling*) pole jednotlivců, který se inicializoval vždy po přidání nového atributu.



Obr. 27 Rozložení pracujících studentů ve věku 15-64

8.2.4 Dojížďka

Při bližší analýze dat o dojížďce vyplynulo, že atribut „frekvence vyjížďky“ je silně podmíněn samotným „místem vyjížďky“ (Tab. 13, kód 88), přičemž obdobná závislost existovala i u volby dopravního prostředku. Pořadí generování atributů dojížďky proto muselo být restrukturalizováno tak, aby logicky kopírovalo sčítací arch: nejprve místo, poté frekvence, a nakonec dopravní prostředek (ČSÚ 2021A).

Do syntetické populace byl k tomuto účelu zaveden odvozený pomocný atribut pro místo vyjížďky. Ten nabýval hodnot: „známé místo vyjížďky“; „necestující“ (pro pracující/studující z domova, osoby bez stálého pracoviště a pro nepracující osoby, jako jsou senioři či děti); a „nezjištěno“ (pro jednotlivce s nezjištěnou ekonomickou aktivitou či nezjištěným místem dojížďky). Během tohoto procesu došlo k metodickému přesunu kódu pro osoby s nezjištěným místem dojížďky z hodnoty 88 na 99, aby byla usnadněna algoritmická imputace (nahrazování) nezjištěných hodnot.

Tab. 13 Číselník frekvence dojížďky (zdroj: ČSÚ 2021B, Příl. 2)

Kód	Text
1	5x týdně a častěji
2	1x – 4x týdně
3	Pravidelně, ale méně než 1x týdně
4	Dojíždím/docházím zcela nepravidelně
5	Docházka z jiného místa než z obvyklého bydliště
88	Nedefinováno - osoby nepracující a nestudující, s nezjištěným místem dojížďky, pracujícím/studujícím doma a bez stálého místa pracoviště/školy
99	Nezjištěno

Tabulka podmíněných pravděpodobností pro frekvenci dojížďky byla přebudována (včetně indexace škola/práce, sloučení a párování pomocného atributu). Následně byly z tabulky místa vyjížďky vyfiltrovány záznamy o nepracujících/nestudujících jednotlivcích a o osobách s nezjištěnou ekonomickou aktivitou a přeneseny do upravované tabulky frekvence. Záznamům byla logicky přiřazena hodnota pomocného atributu „nevyjíždějící“ a „nezjištěno“ (podle Tab. 14). Zavedení pomocného atributu frekvence tak zajistilo, aby necestujícím osobám nebyl v dalším kroku nesmyslně přidělen dopravní prostředek. Zpracování dopravního prostředku probíhalo obdobně, avšak bez počátečního rozlišení na pracující a studující. Výsledné atributy dojížďky byly tedy do syntetické populace přidávány s podmíněním na čtyři až pět proměnných (pohlaví, věková skupina, ekonomická aktivita a místo, resp. frekvence vyjížďky).

Tab. 14 Matice kombinací ekonomické aktivity, místa vyjížďky a pomocných atributů

Ekonomická aktivita	**Pomocný atribut ekonomické aktivity (1)**	**Místo vyjížďky**	**Pomocný atribut místa vyjížďky**
nezaměstnaní, ve starobním/invalidním důchodu, s jiným vlastním zdrojem obživy děti předškolního věku a ostatní bez vlastního příjmu, na rodičovské dovolené	nevyjíždějící	nevyjíždějící	nevyjíždějící
nezjištěno	nezjištěno	nezjištěno	nezjištěno
zaměstnanci, podnikatelé, OSVČ…, pracující důchodci, pracující studenti, na mateřské dovolené	pracující	v rámci obce, do jiné obce v okresu, jiný okres, jiný kraj, zahraničí	známé
		nevyjíždějící, bez stálého místa pracoviště	nevyjíždějící
		nezjištěno	nezjištěno
nepracující studenti a žáci	studující	v rámci obce, do jiné obce v okresu, jiný okres, jiný kraj, zahraničí	známé
		nevyjíždějící	nevyjíždějící
		nezjištěno	nezjištěno

8.2.5 Přidávání na základě velkého množství atributů

Při vysokém počtu podmiňujících atributů exponenciálně roste riziko tzv. prázdných množin/situací, kdy zdrojové tabulky neobsahují všechny teoreticky možné kombinace. Pokud tyto kombinace v algoritmu zcela chybí, dochází k chybám běhu (*crash*) modelu. Pro zajištění stability bylo nezbytné tyto kombinace programově definovat: logicky přípustným, avšak v datech se nevyskytujícím kombinacím byla přiřazena marginální pravděpodobnost (např. 0,00001), zatímco kombinacím logicky nemožným byla explicitně dosazena pravděpodobnost nulová.

8.2.6 Metoda eliminace nezjištěných hodnot atributů

Vstupní data obsahovala u řady atributů záznamy typu „nezjištěno“, které bylo pro účely přesného modelování nutné nahradit reálnými hodnotami. V rámci post-processingu byla implementována

funkce pro probabilistické nahrazování (imputaci). Funkce seskupila jednotlivce podle shodných podmiňujících atributů (věk, pohlaví apod.), načež z nich vypočítala vnitřní pravděpodobnostní rozdělení daného znaku (při ignorování nezjištěných hodnot). Chybějící hodnoty byly následně nahrazeny náhodným losováním z tohoto odvozeného rozdělení.

8.2.7 Household Grouper

Sdružování jednotlivců do domácností zajišťoval algoritmus Household Grouper, který je standardní součástí původní knihovny GenSynthPop. Jeho úkolem je formovat domácnosti (jednotlivce, sezdané/nesezdané páry a rodiny s různým počtem dětí) na základě věku, pohlaví a pozice jednotlivce. Původní algoritmus využíval jednoduchou heuristickou strategii bez iterací, což sice zajišťovalo rychlost, avšak v českém datovém prostředí to vedlo k logickým chybám. Původní skript v rámci uzavřené knihovny navíc neumožňoval standardní ladění (*debugging*), proto byl jeho zdrojový kód vyjmut a přepsán přímo do těla modelu.

Hlavní problém původní funkce spočíval v deterministickém přiřazování dětí k rodičům (postupně od nejstaršího dítěte k ideální matce). Pokud vhodná matka nebyla k dispozici, algoritmus slevoval z kritérií, což v praxi běžně vedlo k absurdním situacím, kdy nejstarším párům byly systematicky přiřazovány nejmladší děti. V případě vyčerpání volných rodičů navíc proces skončil kritickou chybou (Obr. 28). Tato systémová selhání vznikala jako přímý důsledek metodiky předzpracování, konkrétně během fáze, kdy byl jednotlivcům přiřazován věk a specifická pozice v domácnosti ještě před samotným formováním rodinných jednotek. Z těchto důvodů bylo nezbytné zcela přepracovat logiku dvou hlavních funkcí: metodiku shlukování sourozenců a přiřazování dětí k rodičům. Základním limitem upraveného modelu ovšem nadále zůstává jeho strukturální omezení, nedokáže spolehlivě generovat domácnosti se třetími dospělými osobami, vícegenerační soužití nebo rodiny se čtyřmi a více dětmi. V důsledku toho jsou vygenerované syntetické domácnosti průměrně menší než domácnosti reálné, zjištěné během SLDB 2021.

```
Exception has occurred: ValueError ×
Child age gap too small: 76 - 69 = 7
  File "C:\Users\Eliska\Documents\takuzkonecne\diplomka\generate_households.py", line 210, in partition_households
    return hh_grouper.run()
  File "C:\Users\Eliska\Documents\takuzkonecne\diplomka\generate_households.py", line 229, in perform_stage
    df_synth_pop, df_synth_households = action(df_synth_pop, df_synth_households)
  File "C:\Users\Eliska\Documents\takuzkonecne\diplomka\generate_households.py", line 508, in <module>
    df_synth_pop_iteration, df_synth_household_iteration = perform_stage(
ValueError: Child age gap too small: 76 - 69 = 7
```

Obr. 28 Kritická chyba původního skriptu, která zapříčila ukončení generování (rozdíl mezi věkem matky a dítě činil 7 let)

8.2.7.1 Metodika shlukování sourozenců

Nově navržený přístup opouští strategii tzv. hladového algoritmu (*greedy algorithm*) a posouvá ho k iterativní optimalizaci nad celými skupinami sourozenců. Na rozdíl od původního řešení, které vybíralo sourozence sekvenčně na základě lokálního minima (tj. minimalizace věkového rozdílu

vůči již vybraným dětem, konkrétně vůči věkově nejbližšímu), je nová metoda založena na hodnocení kvality celé skupiny jako celku.

Proces začíná náhodným rozdělením všech dostupných dětí do skupin o požadované velikosti. Toto počáteční řešení slouží jako výchozí stav pro následnou optimalizaci. Každé skupině je přiřazena penalizační funkce, která vyjadřuje míru biologické pravděpodobnosti věkového rozpětí mezi sourozenci. Na rozdíl od původního lineárního přístupu je zde použita penalizační funkce založena na expertním odhadu, protože empirická data nebyla k dispozici. Nízké nebo nulové penalizace jsou přiřazeny skupinám s malým věkovým rozpětím, zatímco s rostoucím rozdílem věku penalizace roste nelineárně. Specifickým případem je identický věk, který je penalizován samostatně, a extrémní věkové rozdíly, které jsou hodnoceny velmi vysokou penalizací jako biologicky nepravděpodobné (Tab. 15).

Tab. 15 Použitá penalizace při tvorbě sourozenců vytvořena na základě expertního odhadu

Rozsah	**Interpretace**	**Penalizace**
0	dvojčata	20
1–4 roky	ideální sourozenci	0
5–10	ok	2 × rozdíl
11-30	rychleji penalizuje	10 × rozdíl
>30	biologický nesmysl	5000

Samotná optimalizace probíhá iterativně prostřednictvím lokálních záměn mezi dvojicemi skupin. V každé iteraci jsou identifikovány skupiny s nenulovou penalizací, které jsou považovány za problematické, a u nichž se algoritmus snaží nalézt zlepšení. Pro každou takovou skupinu jsou testovány možné záměny jednotlivých dětí s dětmi z jiných skupin stejné velikosti. Každá potenciální záměna je vyhodnocena pomocí penalizační funkce aplikované na obě dotčené sourozenecké skupiny.

Záměna je přijata pouze v případě, že vede ke snížení penalizace u problematické skupiny a zároveň nezpůsobí zhoršení druhé skupiny. Tím je zajištěno, že kvalita již dobře sestavených skupin není narušena. Optimalizační proces je tedy monotónní ve smyslu nezhoršování řešení a postupně směřuje ke stabilní konfiguraci. Algoritmus je ukončen v okamžiku, kdy již nelze nalézt žádnou záměnu vedoucí ke zlepšení, případně po dosažení předem stanoveného maximálního počtu iterací.

Navržený přístup umožňuje lépe kontrolovat globální vlastnosti výsledného rozdělení a současně zachovat výpočetní efektivitu, jelikož pracuje pouze s lokálními úpravami existujícího řešení. Nevýhodou zůstává skutečnost, že parametry penalizační funkce jsou v současné fázi založeny na expertním odhadu, což může ovlivnit výslednou kvalitu generovaných skupin. V budoucnu by proto bylo vhodné tyto parametry kalibrovat na základě reálných dat, případně nahradit pevně dané prahy adaptivním přístupem. Limitem funkce taky zůstává uvíznutí v lokálním minimu, není možné udělat vícenásobnou výměnu.

8.2.7.2 Přiřazování rodič-dítě

Původní přístup k přiřazování rodičů k dětem byl založen na sekvenčním zpracování s využitím externě definovaného rozdělení věkových rozdílů mezi rodiči a dětmi. Pro každou skupinu dětí byl postupně vybírán nejvhodnější rodičovský pár na základě minimalizace rozdílu mezi požadovaným a skutečným věkovým odstupem. Výběr probíhal deterministicky z množiny dosud nepřiřazených rodičů, přičemž vhodnost byla určena funkcí penalizující odchylku od předem definovaných intervalů. Tento přístup však optimalizoval přiřazení pouze lokálně pro jednotlivé domácnosti a nebral v úvahu globální kvalitu výsledného rozdělení.

Nově navržený algoritmus rozšiřuje tento princip o vícekrokovou optimalizaci nad celou množinou domácností. V první fázi dochází k inicializačnímu přiřazení, které kombinuje náhodnost s jednoduchými biologickými omezeními. Děti i rodiče jsou nejprve náhodně permutováni, čímž se eliminuje systematické zkreslení způsobené pořadím vstupních dat. Následně jsou skupiny dětí přiřazovány rodičům tak, aby byl splněn základní biologický předpoklad minimálního a maximálního věkového rozdílu. Pokud takové přiřazení není nalezeno, je použito náhradní řešení bez ohledu na kvalitu, aby bylo zajištěno úplné pokrytí populace.

Kvalita formovaných domácností je algoritmicky hodnocena pomocí penalizační funkce, která kvantifikuje věkové rozdíly mezi rodiči a nejstarším, respektive nejmladším dítětem ve skupině. Tato funkce je matematicky navržena jako po částech definovaná a plně reflektuje reálná biologická a demografická omezení. Nulová hodnota penalizace je přiřazena takovému věkovému rozdílu, který odpovídá realistickému věku matky při narození dítěte (Tab. 12). Konkrétně je tento tolerovaný interval stanoven na 15 až 50 let věkového rozdílu, čímž je s vysokou mírou spolehlivosti pokryto 99,975 % všech reálných případů v populaci.

Hlavní přínos nové metody spočívá v zavedení iterativní optimalizace, která kombinuje prvky náhodného prohledávání a lokálního zlepšování. V každém cyklu dochází nejprve k náhodnému „narušení“ aktuálního řešení, kdy jsou mezi vybranými domácnostmi prohozeny skupiny dětí. Tento krok umožňuje opustit lokální minima a prozkoumat širší prostor možných konfigurací. Následně je aplikována cílená lokální optimalizace, která systematicky testuje záměny mezi dvojicemi domácností. Záměna je přijata v případě, že vede ke snížení celkové penalizace posuzované dvojice a zároveň nevytváří extrémně nevhodné kombinace.

Součástí algoritmu je také mechanismus uchovávání nejlepšího nalezeného řešení v průběhu výpočtu. Po každé iteraci je vyhodnocen počet problematických domácností a celková výše penalizace. Pokud dojde ke zlepšení oproti dosud nejlepšímu stavu, je aktuální konfigurace uložena. Tento přístup zajišťuje, že výsledné řešení odpovídá nejkvalitnější nalezené konfiguraci, i v případě, že během optimalizace dojde k dočasnému zhoršení.

Optimalizační proces je ukončen při dosažení stabilního stavu, kdy již nedochází ke zlepšení, případně po dosažení maximálního počtu iterací. Výsledkem je přiřazení rodičů a dětí do domácností, které lépe odpovídá biologickým a demografickým předpokladům než původní

sekvenční metoda. Přestože je algoritmus výpočetně náročnější, zachovává přijatelnou efektivitu díky omezení operací na lokální záměny a současně výrazně zlepšuje globální kvalitu výsledného rozdělení.

8.2.8 Přiřazení atributu typ budovy a patro budovy

Pro původní atributy vztahující se k domácnostem nebyly v českém prostředí nalezeny ekvivalentní zdroje. Pro lepší výsledek přiřazování domácností na adresní body byly k jednotlivým domácnostem připojen typ budovy (rodinný, bytový a ostatní) a rozmezí pater, ve kterém se domácnost nachází. Rozmezí pater bylo stanoveno kvůli redukci počtu kategorií pro jednotlivé typy budov zvlášť na základě dostupných dat. Přiřazování bylo provedeno pomocí stejných funkcích jako atributy u jednotlivců.

8.3 Výstup

Model (Příl. 3) ukládá jednotlivé verze syntetické populace jako .csv vždy po přidání atributu nebo agregaci v domácnosti. V Tab. 16 se nachází názvy atributů využitých v tabulkách s významem.

Tab. 16 Vysvětlivky jednotlivých názvů atributů (číselníky k vybraným atributům jsou dostupné v modelu /datasources/code_lists)

Tabulka	Název atributu	Význam
Jednotlivci	agent_id	jednoznačný identifikátor jednotlivce
Jednotlivci, domácnosti	neighb_code	kód městské části
Jednotlivci	age_group	věková skupina
Jednotlivci	gender	pohlaví
Jednotlivci	age	věk
Jednotlivci	education	základní členění vzdělání
Jednotlivci	education_specific	konkrétní vzdělání
Jednotlivci	economical_activity	ekonomická aktivita
Jednotlivci	ea_school_work	pomocný atribut založený na ekonomické aktivitě (1) - hodnoty: práce, škola, nevyjíždějící
Jednotlivci	ea_school_work_2	pomocný atribut založený na ekonomické aktivitě (2) - hodnoty: práce+škola, nevyjíždějící
Jednotlivci	place_activity	místo vyjížďky
Jednotlivci	category_place_activity	pomocný atribut založený na místě vyjížďky – hodnoty: známé, neznámé (místo vyjížďky)
Jednotlivci	frequency_activity	frekvence vyjížďky
Jednotlivci	category_f_activity	pomocný atribut založený na frekvenci vyjížďky – hodnoty: nevyjíždějící, ostatní

Tab. 16 Pokračování

Tabulka	Název atributu	Význam
Jednotlivci	vehicle_activity	hlavní dopravní prostředek
Jednotlivci	small_age_group	5letá věková skupina
Jednotlivci	household_position	pozice v domácnosti
Jednotlivci, domácnosti	household_id	jednoznačný identifikátor domácnosti
Domácnosti	hh_type	typ domácnosti
Domácnosti	hh_size	počet členů domácnosti
Domácnosti	small_hh_type	obecný typ domácnosti
Domácnosti	building_floor	typ budovy + interval podlaží

Model automaticky hodnotí i svoji úspěšnost pomocí testu Chí-kvadrát. Během samotného procesu generování populace a sekvenčního přidávání atributů může docházet k dílčím statistickým odchylkám, na které hodnotící funkce algoritmu průběžně upozorňují varovnými hlášeními (poklesem p-hodnoty pod hladinu významnosti $\alpha = 0.05$). Jednou z hlavních příčin tohoto jevu je prostorová heterogenita dat. Algoritmus je nucen aplikovat agregované pravděpodobnosti z celoměstské úrovně na lokální úroveň jednotlivých sousedství, čímž dochází ke ztrátě informací o specifických prostorových vzorcích. S tím úzce souvisí i další faktor, a to vliv opomenutých proměnných (tzv. nepozorovaná heterogenita). Zkoumaný atribut totiž může být v realitě silně podmíněn jinou socio-demografickou charakteristikou (např. volba dopravního prostředku závisí nejen na věku, ale i na příjmu či rodinném stavu). Pokud tato specifická vazba není z důvodu datových či výpočetních limitů explicitně zahrnuta v podmiňující matici, algoritmus distribuuje vlastnosti na základě zprůměrovaných trendů. V důsledku těchto lokálních i strukturálních kompromisů se následné součty přirozeně odchylují od dokonalé shody. Je však nutné zdůraznit, že tyto odchylky jsou objektivním a očekávaným limitem vstupních dat. Absolutní chyba se v kontextu celého města pohybuje v řádech desetin procenta, a celková reprezentativnost i validita vygenerované syntetické populace tak zůstává pro účely modelování plně zachována.

S rostoucím počtem podmiňujících proměnných u komplexních atributů dochází v modelu k logickému a očekávanému zhoršení statistické shody (měřené testem Chí-kvadrát). S každou další přidanou podmiňující dimenzí (např. při společné kombinaci věku, pohlaví, ekonomické aktivity a dojížďky) roste počet možných kategorií exponenciální řadou. V takto fragmentované vícerozměrné matici klesá očekávaná absolutní četnost v jednotlivých buňkách na velmi nízké hodnoty, často až pod úroveň jednotek osob.

Nutnost převodu teoretických spojitých pravděpodobností z IPF algoritmu na reálné diskrétní agenty (celá čísla) následně generuje v těchto řídkých buňkách výrazné relativní odchylky. Zatímco chyba o jednoho agenta v buňce s očekávanou četností tisíc osob je statisticky zcela zanedbatelná, stejná absolutní odchylka v okrajové buňce s očekávanou četností dvou osob představuje disproporci 50 %. Test Chí-kvadrát je svou matematickou podstatou na tyto relativní

nepřesnosti u velmi nízkých četností extrémně citlivý. Výsledný pokles p-hodnot u hluboce podmíněných atributů je tedy odrazem strukturálního limitu vstupních, nikoliv indikátorem selhání optimalizačního algoritmu.

Závěrečné statistické zhodnocení kvality vygenerované syntetické populace je v rámci modelu realizováno prostřednictvím sady specializovaných evaluačních funkcí. Cílem těchto analytických nástrojů je systematické porovnání pozorovaných četností (rozložení atributů ve finální syntetické populaci) s četnostmi očekávanými (odpovídajícími původním marginálním statistikám, respektive teoretickým vícerozměrným maticím odvozeným pomocí algoritmu IPF). Každá z těchto funkcí hodnotí cílový atribut napříč různými úrovněmi podmiňujících proměnných, jako příklad je na Obr. 29 výřez ekonomické aktivity v závislosti na věkové skupině a pohlaví (*male* – muž; *female* – žena, *selfsufficient* – s jiným vlastním zdrojem obživy, *undefined* – nedefinováno, *unemployed* – nezaměstnaný, *working retired* – pracující důchodce, *working student* – pracující student, *employed* – zaměstnaný).

age_group	gender	economical_activity	synthetic_population_counts	fitted_sourc
25-44	female	economical_activity_selfsufficient	108	105.7486
25-44	female	economical_activity_undefined	1179	1150.398
25-44	female	economical_activity_unemployed	2795	2774.695
25-44	female	economical_activity_working_retired	978	983.8441
25-44	female	economical_activity_working_student	3279	3275.909
25-44	male	economical_activity_employed	56124	56263.59

Obr. 29 Porovnání pozorovaných (synthetic population counts) a očekávaných (fitted source data counts) četností

8.4 Prostorové rozmístění

Prostorová alokace syntetické populace v rámci zájmového území vychází z kombinace celostátních a lokálních geodat. Konkrétně se jedná o propojení národní sady „Vchody do budovy s TEP (RSO)“, pocházející z Datového portálu GIS (ČSÚ), a lokální vrstvy „Průzkum budov v Brně“, získané z městského geoportálu data.brno.cz. Obě klíčové datové sady byly vzájemně ztotožněny prostřednictvím unikátních identifikátorů stavebních objektů v systému RÚIAN. Metodický popis původu těchto zdrojů a jejich analýza jsou součástí kapitoly 7.

8.4.1 Postup zpracování prostorových dat

Datové sady (popsány v kapitolách 7.1.1, 7.2.1) byly v ArcGIS Pro sloučeny na principu jeden vchod k první budově na základě RÚIAN identifikátoru se zachováním všech prvků z bodové vrstvy vchodů. Takto vytvořená vrstva byla uložena jako geopackage.

Počet podlaží byl vzat z průzkumu budov, protože se jedná o důvěryhodnější zdroj, pokud informace chyběla, bylo automaticky rozhodnuto, že budova má 1 podlaží.

Byl vytvořen skript přiřazující domácnosti do budov. Vstupem algoritmu jsou dvě hlavní datové struktury. První představuje soubor syntetických domácností obsahující informace

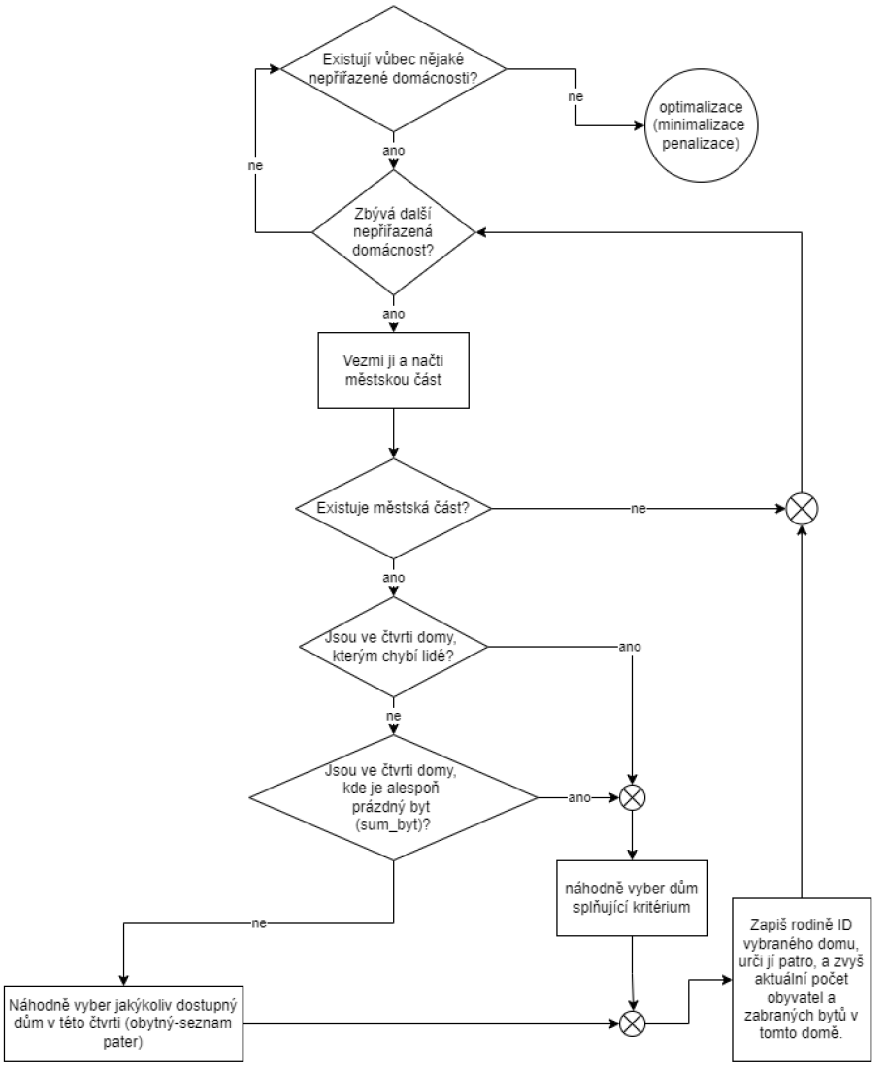
o velikosti domácnosti, typu budovy a preferovaném rozmezí podlaží. Druhou tvoří tabulka budov získaná z geoprostorových dat, která obsahuje identifikátory budov, počet bytových jednotek, počet obyvatel dle sčítání a informace o počtu podlaží. V rámci předzpracování dochází k očištění dat (Příl. 4) a k odvození seznamu obytných podlaží pro každou budovu (Příl. 5). Pokud tato informace není dostupná, je generována na základě existence obyvatel či bytových jednotek, čímž je zajištěno, že každá relevantní budova má definovaný alespoň minimální rozsah obytných pater.

Samotný proces přiřazení je rozdělen do tří na sebe navazujících fází. V první fázi dochází k primárnímu přiřazení domácností k budovám na základě lokální optimalizace v rámci jednotlivých územních jednotek. Domácnosti jsou nejprve seskupeny podle územního kódu a typu budovy, což umožňuje preferovat odpovídající typy zástavby (Příl. 6). Pro každou budovu jsou následně iterativně vybírány domácnosti tak dlouho, dokud není dosaženo cílového počtu obyvatel. Kapacita budovy je řízena hodnotou počtu obyvatel ze sčítání lidu (Příl. 7).

Výběr konkrétní domácnosti je realizován pomocí penalizační funkce, která hodnotí změnu odchylky mezi aktuálním a cílovým počtem obyvatel po hypotetickém přiřazení domácnosti. Přeplnění budovy je penalizováno výrazně silněji než její nedostatečné obsazení. Doplňkově je zohledněna kompatibilita mezi požadovaným typem podlaží domácnosti a dostupnými obytnými podlažími v budově. Výsledkem je výběr domácnosti, která nejlépe přispívá k dosažení cílové populace při zachování základní prostorové konzistence.

Druhá fáze řeší situace (Obr. 30), kdy některé domácnosti zůstaly nepřiřazené. Tyto případy vznikají zejména v důsledku lokálních omezení během první fáze. Algoritmus v tomto kroku upřednostňuje budovy, které dosud nedosahují cílového počtu obyvatel. Pokud takové budovy nejsou k dispozici, jsou využity budovy s volnou kapacitou bytových jednotek, a teprve v krajním případě dochází k náhodnému přiřazení v rámci dané územní jednotky. Tento postup zajišťuje úplné pokrytí všech domácností a současně minimalizuje odchylky od referenčních hodnot.

Ve třetí fázi je aplikována globální optimalizace pomocí iterativních záměn mezi dvojicemi domácností. V každé iteraci jsou náhodně vybrány dvě domácnosti v rámci stejné územní jednotky a je testována jejich výměna mezi budovami. Kritériem pro přijetí záměny je snížení kvadratické chyby mezi aktuálním a cílovým počtem obyvatel v obou dotčených budovách. Tento postup umožňuje postupné vyrovnávání odchylek a vede ke zlepšení celkové shody modelu s referenčními daty. Současně je při každé změně aktualizováno přiřazení podlaží tak, aby odpovídalo charakteristikám cílové budovy.



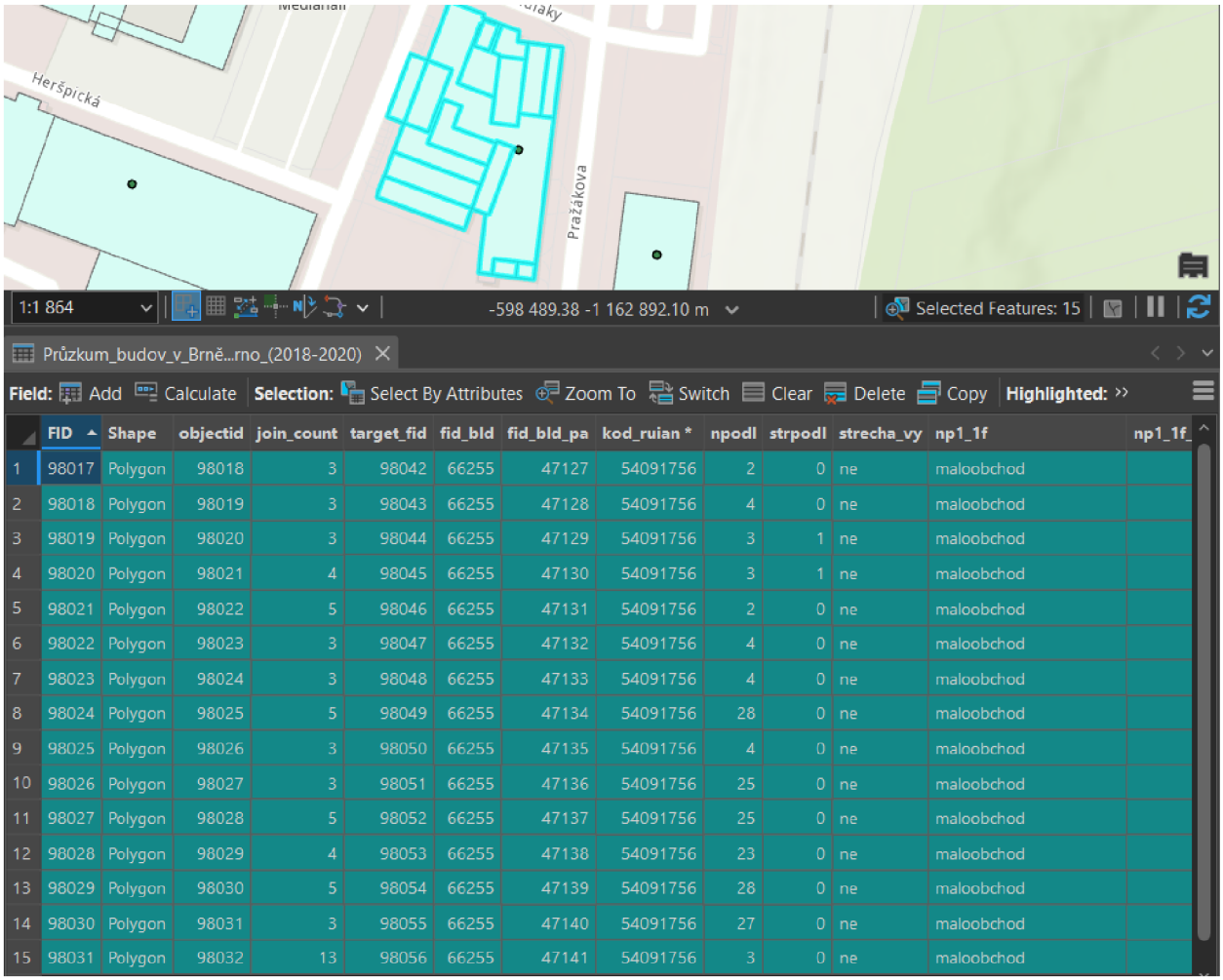
Obr. 30 Druhá fáze algoritmu, více na Zpracování prostorových dat 2026 nebo Příl. 4-7

Celý algoritmus kombinuje deterministické i stochastické prvky. Náhodnost je využita při výběru kandidátů i během optimalizačního procesu, což umožňuje efektivnější prohledávání prostoru řešení a snižuje riziko uvíznutí v lokálním minimu. Rozhodování je však řízeno penalizační funkcí, která zajišťuje konzistenci s dostupnými daty a základními prostorovými omezeními.

Výsledkem je syntetická populace, ve které jsou domácnosti přiřazeny ke konkrétním budovám tak, aby co nejlépe odpovídaly známému rozložení počtu obyvatel. Omezením přístupu zůstává závislost na kvalitě vstupních dat a na nastavení penalizační funkce, která je založena na expertním odhadu.

Konkrétním limitem je způsob spojení prostorových dat v režimu 1 : 1. V průzkumu budov se vyskytují případy, kdy jedna budova je rozdělena na několik částí (Obr. 31). V důsledku toho dochází k vybrání pouze prvního záznamu, který může obsahovat neúplné, zkreslené nebo nereprezentativní informace. Typickým případem je situace, kdy vybraný záznam reprezentuje pouze část budovy s nízkým počtem podlaží (např. 2 podlaží namísto skutečných 28) a současně

neobsahuje funkci bydlení. To následně negativně ovlivňuje přesnost přiřazení domácností, zejména ve vztahu k určení odpovídajícího podlaží.



	FID	Shape	objectid	join_count	target_fid	fid_bld	fid_bld_pa	kod_ruian *	npodl	strpodl	strecha_vy	np1_1f	np1_1f_
1	98017	Polygon	98018	3	98042	66255	47127	54091756	2	0	ne	maloobchod	
2	98018	Polygon	98019	3	98043	66255	47128	54091756	4	0	ne	maloobchod	
3	98019	Polygon	98020	3	98044	66255	47129	54091756	3	1	ne	maloobchod	
4	98020	Polygon	98021	4	98045	66255	47130	54091756	3	1	ne	maloobchod	
5	98021	Polygon	98022	5	98046	66255	47131	54091756	2	0	ne	maloobchod	
6	98022	Polygon	98023	3	98047	66255	47132	54091756	4	0	ne	maloobchod	
7	98023	Polygon	98024	3	98048	66255	47133	54091756	4	0	ne	maloobchod	
8	98024	Polygon	98025	5	98049	66255	47134	54091756	28	0	ne	maloobchod	
9	98025	Polygon	98026	3	98050	66255	47135	54091756	4	0	ne	maloobchod	
10	98026	Polygon	98027	3	98051	66255	47136	54091756	25	0	ne	maloobchod	
11	98027	Polygon	98028	5	98052	66255	47137	54091756	25	0	ne	maloobchod	
12	98028	Polygon	98029	4	98053	66255	47138	54091756	23	0	ne	maloobchod	
13	98029	Polygon	98030	5	98054	66255	47139	54091756	28	0	ne	maloobchod	
14	98030	Polygon	98031	3	98055	66255	47140	54091756	27	0	ne	maloobchod	
15	98031	Polygon	98032	13	98056	66255	47141	54091756	3	0	ne	maloobchod	

Obr. 31 Průzkum budov Brna, části budovy AZ Tower

Další je problém interpretace patra v případě vícepatrových bytů typicky vyskytujících se v rodinných domech. Ve vstupních datech jednoznačně definováno, jestli se hodnota vztahuje k počtu podlaží obývaných domácností nebo pořadovému číslu konkrétního podlaží a zároveň není zřejmé, zda tato hodnota reprezentuje například první, poslední nebo průměrné podlaží v rámci bytu. Tato nejednoznačnost vede k nutnému zjednodušení v modelu, kdy je každé domácnosti přiřazeno pouze jedno reprezentativní podlaží, což nemusí plně odpovídat reálnému způsobu využívání prostoru.

Možným vylepšením je využití počtu podlaží z vchodů do budov pro záznamy, kde neproběhlo sjednocení, ale vzhledem k neznámé kvalitě dat datové sady nebylo toto řešení implementováno. Významný potenciál pro budoucí rozvoj modelu a další zpřesnění prostorové alokace pak představuje plánované zavedení registru bytů.

8.5 Návrh vizualizace

Prvotním návrhem byla interaktivní aplikace v ArcGIS Online pro případy kontroly vygenerovaných prostorově umístěných domácností. Sloučená prostorová data budov s vchody byla nahrána do programu Excel, kde byly zduplikovány řádky na základě počtu podlaží, ke každému podlaží byla přiřazena jeho funkce, v případě hodnot *null* bylo podlaží doplněno na hodnotu 1 a funkce hodnotou 0. Následně byl sloučen unikátní identifikátor (ne RÚIAN kvůli opakujícím se nulovým hodnotám) pro každý vchod s konkrétním podlažím jako unikátní identifikátor podlaží. Tabulka domácností byla doplněna unikátní identifikátor podlaží, aby bylo možné spárovat tabulku podlaží a domácností. Tabulka jednotlivců byla přeložena a vyčištěna.

Tabulky podlaží, domácností, jednotlivců a bodová vrstva vchodů spolu s budovami byla nahrána do cloudového prostředí ArcGIS Online a v Experience Builderu vytvořena aplikace. Byly nastaveny prokliky mezi jednotlivými tabulkami.

Při analýze možností vizualizace syntetické populace je nutné zohlednit, že dostupné přístupy jsou do určité míry omezené. Klíčové je proto využít přidanou hodnotu syntetické populace oproti veřejně dostupným okrajovým údajům. Jednou z hlavních výhod je možnost analyzovat vztahy mezi netradičními kombinacemi proměnných, které nejsou v oficiálních statistikách běžně publikovány. Další možností je zobrazování dat na jemnější prostorové úrovni, než pro jakou jsou dostupná agregovaná data, přičemž tyto výstupy mohou být doplněny vhodnými grafickými reprezentacemi.

Vzhledem k tomu, že potenciální využití syntetické populace je velmi široké, lze očekávat potřebu uživatelsky definovaných vstupů pro vizualizaci, které umožní přizpůsobit výstupy konkrétním požadavkům. Různí uživatelé mohou sledovat odlišné jevy a vztahy, a proto je vhodné uvažovat flexibilní přístup k tvorbě vizualizací.

Současně je třeba zdůraznit, že syntetická populace v této podobě představuje statická data, která jsou typicky využívána jako vstup do agentních modelů (ABM). Z tohoto důvodu nebývá jejich vizualizace primárním cílem analýzy a je často upozaděna ve prospěch následného modelování a simulací.

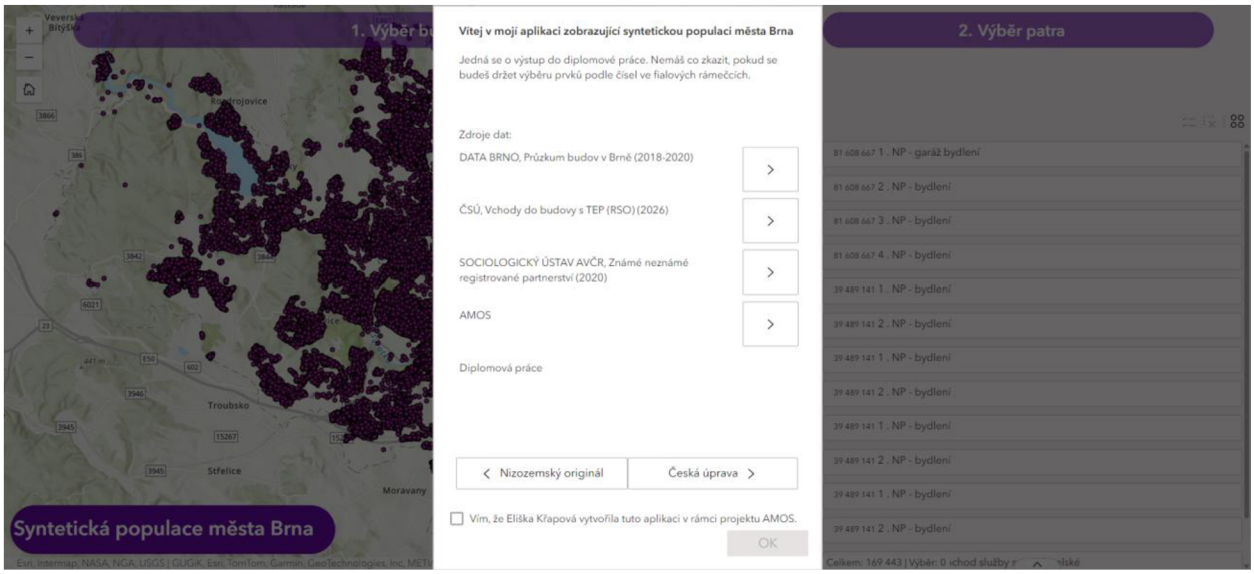
8.5.1 Popis aplikace

Výsledná interaktivní aplikace (Finální_apka 2026 nebo také Příl. 8) plně reflektuje výše popsanou relační datovou strukturu. Uživatelské rozhraní je koncipováno hierarchicky tak, aby umožňovalo plynulý průchod od nejvyšší prostorové úrovně až k nejmenšímu demografickému detailu.

8.5.1.1 Úvodní obrazovka a navigace:

Při prvotním spuštění aplikace (nebo obnovení okna prohlížeče) je uživatel seznámen s informačním panelem (Obr. 32), který shrnuje metadatový kontext a použité zdroje. Pro přístup k mapě je nezbytné tyto informace potvrdit. K panelu se lze následně kdykoliv vrátit pomocí

tlačítka „Zdroje“ v pravém dolním rohu mapy (Obr. 33). Základní pohyb v mapovém poli je řešen standardně, přibližování pomocí kolečka myši a posun tažením se stisknutým levým tlačítkem.



Obr. 32 Informační panel

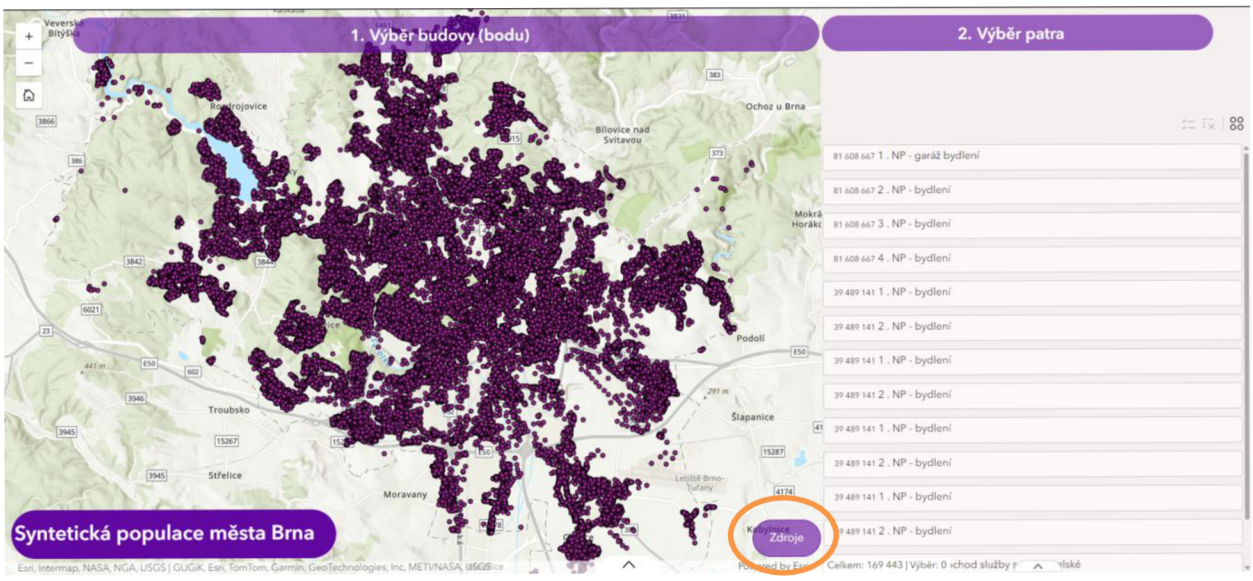
8.5.1.2 Hierarchická filtrace dat

Proces prohlížení prostorově ukotvené populace probíhá ve třech logických krocích, které těží z předem definovaných vazeb mezi tabulkami:

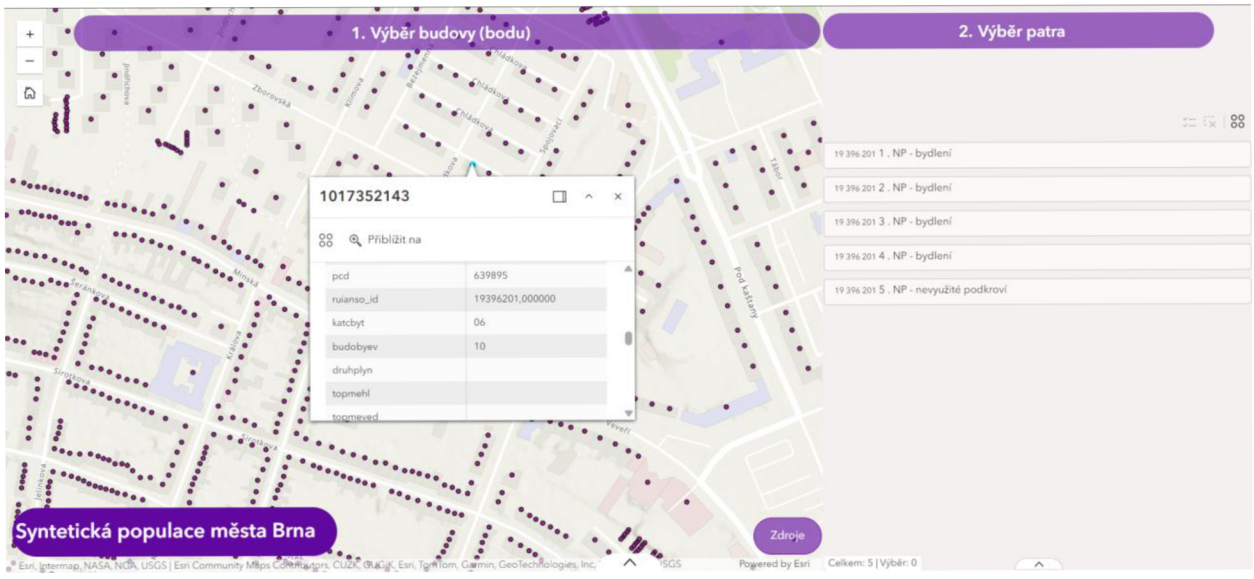
1. Výběr adresního bodu: Vstupním bodem je volba konkrétního vchodu na mapě (fialový kruhový symbol). Po kliknutí je prvek zvýrazněn tyrkysovým ohraničením a rozevře se vyskakovací okno (pop-up) se základními informacemi o budově. Tento úkon na pozadí aktivuje první úroveň prostorového filtru (funguje jenom pro patra a domácnosti, jednotlivci nemají přímou vazbu na budovu) (Obr. 34).

2. Výběr podlaží: v postranním panelu uživatelského rozhraní se dynamicky vygeneruje identifikátor RÚIAN, číslo objektu a seznam všech dostupných podlaží vybrané budovy včetně jejich funkcí. Volba konkrétního patra (indikována fialovým orámováním) využívá připravený unikátní identifikátor podlaží ke spárování s navazující tabulkou (Obr. 35).

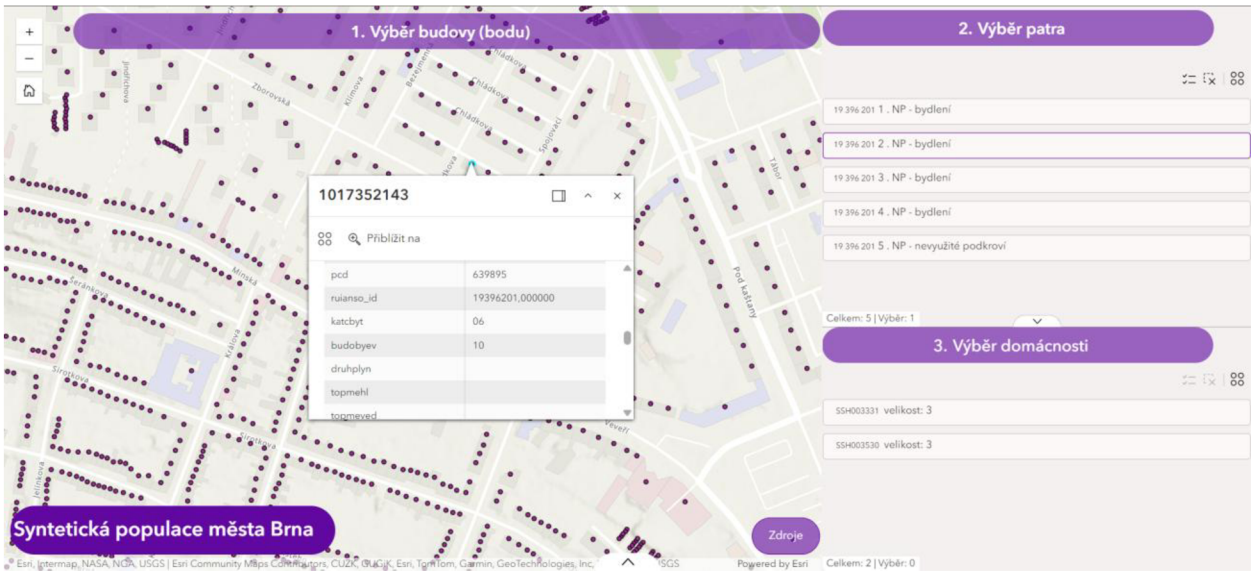
3. Výběr domácnosti: Na základě zvoleného patra se ve spodní sekci panelu vyfiltrují a zobrazí alokované domácnosti (identifikovatelné pomocí ID a počtu členů). Pokud patro neobsahuje žádné obyvatele, systém zobrazí hlášení „žádná data“ (Obr. 36). Ačkoliv je výskyt domácností nejpravděpodobnější v patrech s obytnou funkcí, systém vizualizuje reálný výstup modelu, a tedy i případné domácnosti alokované v nebytových patrech (např. AZ Tower – Obr. 37).



Obr. 33 Počáteční obrazovka aplikace

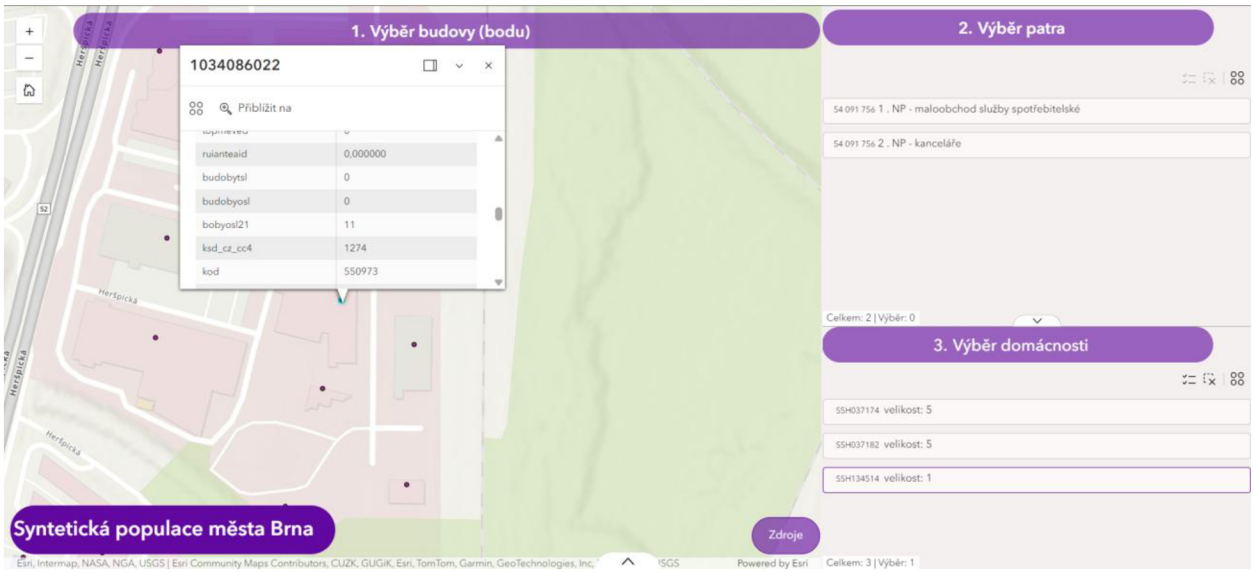


Obr. 34 Výběr bodu (zaznačen tyrkysově), pop-up a filtrace patra a domácností

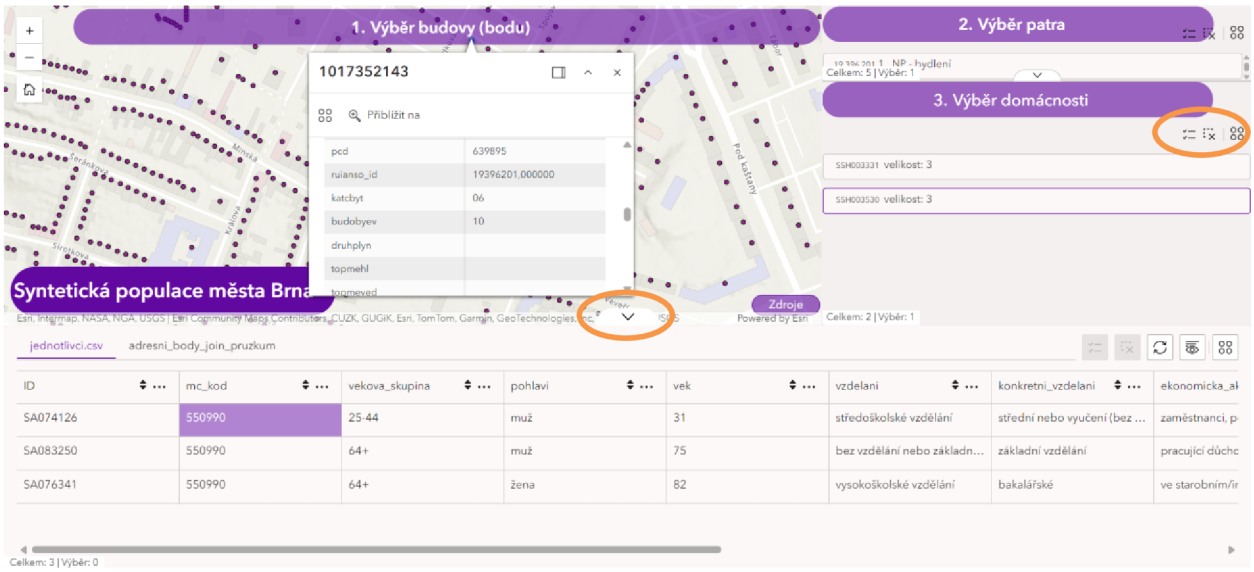


Obr. 35 Výběr patra a filtrování domácností v daném patře budovy

Obr. 36 Výběr patra bez domácností



Obr. 37 AZ Tower má dvě patra s neobytnou funkcí, ale přesto jsou k němu přiřazeny domácnosti (není vybrané patro, proto se zobrazují všechny domácnosti budovy)



Obr. 38 Vybrání domácnosti otevřelo skrytou atributovou tabulku jednotlivců daně domácnosti,

8.5.1.3 Detail jednotlivců a pokročilé nástroje

Finálním krokem je výběr konkrétní domácnosti, který ve spodní části obrazovky inicializuje zobrazení atributové tabulky. Zde se načítají přeložená a vyčištěná data o charakteristikách všech jednotlivců tvořících vybranou rodinnou jednotku (Obr. 38).

Pro maximalizaci uživatelského komfortu a efektivity práce disponuje rozhraní doplňkovými ovládacími prvky. Zrušení výběrů lze provádět buď odznačením konkrétní položky (např. opětovným klikem na patro/domácnost), zavřením vyskakovacího okna na mapě, nebo plošně pomocí ikony vymazání (čtverec s křížkem) u příslušných sekcí. Pokročilé kontextové menu (ikona čtyř bodů) pak nabízí funkce jako export dat pro externí analýzy. Posledním tlačítkem jsou dvě fajfky s pomlčkami, které filtrují vybrané prvky (skryjí ty, co jsou nevybrané). Panely s výběry domácností i atributovou tabulku lze kdykoliv minimalizovat pomocí šipek směřujících dolů (tlačítka jsou zaznačena na Obr. 38). Filtraci jednotlivců ovlivňuje jenom výběr domácnosti.

9 Diskuze

Kvalita výstupů modelu je bezprostředně závislá na charakteru vstupních dat, jejich vzájemných vazbách a inherentních omezeních. Analýza ukázala, že nejpřesnějších výsledků při generování syntetické populace je dosahováno v geografických celcích, které vykazují podobné procentuální zastoupení klíčových atributů, a při důsledném využití lokálních okrajových (marginálních) podmínek.

Zásadním úskalím procesu se ukázala být surovost vstupních dat ze sčítání lidu, která obsahují zjevné anomálie. Tyto odchylky jsou pravděpodobně způsobeny chybovostí na straně respondentů, ať už v důsledku nepochopení metodiky dotazníku, nebo záměrného uvedení nepravdivých či zkreslených údajů. Typickým příkladem zachyceným v datech jsou osoby ve věku 0–14 let evidované jako řidiči osobních automobilů, případně matky ve stejné věkové kategorii. Jelikož použitý model nedokáže tyto logické rozpory automaticky detekovat a korigovat, vyvstává nutnost rozsáhlé předběžné analýzy a manuálního či algoritmického čištění dat. Tento krok je navíc ztížen absencí detailnější metodické podpory ze strany poskytovatelů dat.

V této souvislosti se naplno projevuje limitace české datové základny ve srovnání se zahraničím. Na rozdíl od původní nizozemské případové studie, která těží z vysoce rozvinuté infrastruktury, v České republice prozatím absentují komplexní a snadno dostupné registry pro tyto účely (např. registr bytů, centrální evidence příjmů, statistiky počtu vozidel v domácnostech či kontinuální sledování registrovaných partnerství). Tento deficit výrazně komplikuje tvorbu robustních demografických modelů.

Z metodologického hlediska představuje největší slabinu modelu samotný proces shlukování jednotlivců do domácností. Současný algoritmus operuje s velmi omezenou typologií rodinných struktur, neumožňuje například generování domácností se čtyřmi a více dětmi, opomíjí přítomnost třetích osob a nedokáže adekvátně modelovat vícegenerační soužití. Důsledkem této simplifikace je systematické podhodnocení průměrné velikosti syntetických domácností oproti reálnému stavu, což následně vede k umělému nadhodnocení celkového počtu generovaných domácností v populaci.

Tato chyba se následně propaguje do fáze prostorové alokace. Kvalita prostorových dat (jak geometrická pro účely prostorového spojení, tak atributová) je pro výsledný model kritická. Zásadní komplikace přineslo slučování datasetů s rozdílným datem aktualizace, konkrétně párování definičních bodů vchodů s polygony budov prostřednictvím identifikátoru RÚIAN. V praxi je nezbytné data validovat a konfrontovat se skutečností, neboť datové sady se často rozcházejí v počtu podlaží, případně obsahují neplatné hodnoty (např. počet podlaží roven nule). Tyto defekty musí být ošetřeny již ve fázi předzpracování dat, případně přímo saturovány v logice kódu.

Samotné přiřazování domácností k podlažím naráží na problém výpočtu kapacit. Algoritmus primárně vychází z kapacity budov definované počtem obyvatel dle cenzu. Kvůli vyššímu počtu syntetických domácností je nevyhnutelným důsledkem přeplnění určitých budov nad jejich nominální kapacitu. Situaci nelze uspokojivě řešit ani alokací na základě počtu bytů, jelikož ten nekoresponduje ani s údaji z cenzu, ani s počtem syntetických domácností. Specifický problém představují objekty nebytového charakteru či hromadného ubytování (domovy pro seniory, budovy veřejné správy). Protože tyto objekty nejsou ve vstupních datech jednoznačně klasifikovány a vyděleny, model k nim přistupuje jako ke standardnímu rezidenčnímu bydlení. Výsledkem jsou nelogické prostorové alokace, kdy je například do objektu domova pro seniory umístěna čtyřčlenná rodina.

Přes výše zmíněné limity splňuje vytvořená syntetická populace svůj účel. Není primárně určena k plošné vizualizaci, nýbrž slouží jako detailní datový vstup pro navazující analytické nástroje. Vzhledem k vysoké míře detailu v parametrech dojížďky se vytvořený model profiluje jako vysoce vhodný nástroj pro modelování dopravního chování obyvatelstva v řešeném území města Brna. K plnému využití v oblasti mobility by však bylo žádoucí model do budoucna obohatit o atributy vlastnictví řidičského oprávnění a počtu vozidel na domácnost, které tvořily součást původní nizozemské metodiky.

S ohledem na to, že se model v současnosti spoléhá na data z cenzu prováděného v desetiletých cyklech, nabízí se jako směr budoucího výzkumu vývoj intercenzálního demografického algoritmu. Takový nástroj by dokázal reagovat na meziroční populační dynamiku a průběžně syntetickou populaci aktualizovat. Dalším zásadním krokem pro budoucí rozvoj je restrukturalizace logiky tvorby domácností – uvažovat lze například o inverzním běhu modelu, kdy by primárně probíhalo generování struktury domácností a teprve do nich by byli následně dosazováni jednotlivci. Zásadním posunem pro budoucí prostorovou alokaci by bylo využití chystaného registru bytů a integrace dalších socioekonomických proměnných. To by umožnilo modelovat logiku realitního trhu a přiřazovat syntetické domácnosti ke konkrétním bytovým jednotkám na základě multidimenzionální shody (např. korelace příjmu a velikosti domácnosti s počtem pokojů či celkovou výměrou bytu). Závěrečný prostor pro celkovou optimalizaci modelu pak spočívá v hlubší metodické revizi stávajících podmiňujících atributů a v pokročilejším využití jejich vzájemných statistických korelací již v úvodních fázích generování populace.

10 Závěr

Hlavním cílem předložené diplomové práce bylo ověřit vybrané přístupy konstrukce syntetické populace v podmínkách České republiky a na zvoleném vzorku dokumentovat možnosti její vizualizace a dalšího analytického využití. Práce v úvodních fázích zmapovala a analyzovala dostupné metodické přístupy a následně je kriticky zhodnotila z hlediska aplikovatelnosti na dostupnou tuzemskou datovou základnu.

Jako výchozí pro tvorbu modelu posloužila agregovaná a marginální data ze Sčítání lidu, domů a bytů poskytnutá Českým statistickým úřadem, doplněná o prostorové databáze RÚIAN a lokální průzkumy budov. Během zpracování se jasně ukázalo, že surovost, nekonzistence a specifická omezení českých datových sad představují pro přesné mikrosimulační modelování značnou výzvu. I přes tyto deficity a absenci klíčových evidencí (např. registru bytů) se však podařilo navrhnout a realizovat funkční postup pro generování populace na úrovni jednotlivců i domácností.

Stěžejním výstupem práce je optimalizace a implementace vlastního algoritmu v prostředí jazyka Python, aplikovaná na případovou studii města Brna. Oproti původním sekvenčním metodám (např. v modulu Household grouper) byl navržen a otestován přístup iterativní optimalizace s využitím expertně definovaných penalizačních funkcí. Tento model prokázal schopnost efektivněji shlukovat jednotlivce do logických rodinných struktur s respektováním biologických a demografických limitů a následně tyto domácnosti prostorově alokovat k odpovídajícím obytným budovám v zájmovém území.

Vytvořená syntetická populace představuje cenný datový konstrukt, který elegantně řeší existující konflikt mezi potřebou detailního prostorového rozlišení demografických dat a striktními požadavky na ochranu osobních údajů. Jak bylo v práci demonstrováno, ačkoliv lze tato data interaktivně vizualizovat, jejich hlavní aplikační potenciál spočívá ve využití jakožto vstupů do navazujících agentních modelů (ABM). Výsledný prostorový dataset, obohacený o detailní atributy dojížďky, dopravy a ekonomické aktivity, otevírá široké možnosti využití především v oblasti modelování dopravního chování a urbánní mobility v Brně.

Do budoucna by další rozvoj tohoto přístupu mohl vést k vytvoření intercenzálního demografického algoritmu reagujícího na meziroční dynamiku. Za předpokladu integrace chystaných celostátních registrů a využití inverzního modelování domácností má tato metoda potenciál stát se standardizovaným a robustním nástrojem pro efektivní městské plánování a rozhodovací procesy ve veřejné správě.

Seznam použité literatury

ANTONI, J.-P., KLEIN, O. (2017): Generating a Located Synthetic Population of Individuals, Households, and Dwellings. SSRN Electronic Journal, doi: 10.2139/ssrn.2972615

BORISOV, V., SESSLER, K., LEEMANN, T., PAWELCZYK, M., KASNECI, G. (2023): Language Models are Realistic Tabular Data Generators. arXiv.

CASORIA, L., NERONI, P., SABATUCCI, L., AUGELLO, A., CAGGIANESE, G. (2025): Evaluating LLMs for Synthetic Personas Generation: A Comparative Analysis of Personality Representation and Censorship Effects. In: Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter. Association for Computing Machinery, New York, NY, USA, 1–9.

CEMPA (2025): AB models and microsimulations, Centre for Microsimulation and Policy Analysis, https://www.microsimulation.ac.uk/jas-mine/resources/focus/abm-and-microsimulations/ (31. 7. 2025).

CENTRUM DOPRAVNÍHO VÝZKUMU, V. V. I. (2022): Zpráva z průzkumu | Česko v pohybu, https://www.ceskovpohybu.cz/zprava/ (12. 9. 2025).

CULLINAN, J. (2010): Developing a Continuous Space Representation of a Simulated Population. Spatial Economic Analysis, 5, 317–338. doi: 10.1080/17421772.2010.493954

ČSÚ (2026b): Datový portál - ODS - externí, https://geodata.csu.gov.cz/as/dp-gis/?sort=name&sortType=asc&tab=dj&col=dataSet&locale=cs&ft=tep&idDset=0b0594d4-bcfc-44be-85c7-804440044fd1 (17. 5. 2026).

ČSÚ (2021a): Podrobná nápověda k otázkám, Sčítání 2021, https://scitani.gov.cz/rozsirena-napoveda-k-lsf (28. 4. 2026).

ČSÚ (2026a): Sňatky, rozvody, Statistika, https://csu.gov.cz/snatky-rozvody (28. 4. 2026).

ČSÚ (2021b): Vstupní datové sady, Google Drive, https://drive.google.com/drive/folders/157uC8NzElVOvqy75rQgbDxna8X9t_lCi (28. 4. 2026).

ČÚZK (2026): Způsob využití stavby, https://cuzk.gov.cz/Katastr-nemovitosti/Poskytovani-udaju-z-KN/Ciselniky-ISKN/Ciselniky-k-nemovitosti/Zpusob-vyuziti-stavby.aspx (28. 4. 2026).

DUMONT, M., LOVELACE, R. (2018): Spatial Microsimulation with R. https://spatial-microsim-book.robinlovelace.net (2. 8. 2025).

EL EMAM, K. (2020): Seven Ways to Evaluate the Utility of Synthetic Data. IEEE Security & Privacy, 18, 4, 56–59. doi: 10.1109/MSEC.2020.2992821

EVROPSKÝ INSPEKTOR OCHRANY OSOBNÍCH ÚDAJŮ (2025): Syntetická data | Evropský inspektor ochrany osobních údajů, https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data (28. 7. 2025).

Finální_apka (2026): https://experience.arcgis.com/experience/aee106bbd759433fb8f22784adeebabe (28. 4. 2026).

GEEKS FOR GEEKS (2023): Sparse Matrix and its representations | Set 1 (Using Arrays and Linked Lists), GeeksforGeeks, https://www.geeksforgeeks.org/dsa/sparse-matrix-representation/ (17. 7. 2025).

GONZALEZ-BONORINO, A., CAPRA, M., PANTOJA, E. (2025): LLMs Model Non-WEIRD Populations: Experiments with Synthetic Cultural Agents. arXiv.

HRADEC, J., CRAGLIA, M., DI, L. M., DE, N. S., OSTLAENDER, N., NICHOLSON, N. (2022): Multipurpose synthetic population for policy applications, JRC Publications Repository, https://doi.org/10.2760/50072 (24. 10. 2024).

HUNSINGER, E. (2008): Iterative Proportional Fitting For A Two-Dimensional Table. https://edyhsgr.github.io/IPFDescription/AKDOLWDIPFTWOD.pdf (16. červenec 2025)

CHOPRA, A., KUMAR, S., GIRAY-KURU, N., RASKAR, R., QUERA-BOFARULL, A. (2024): On the limits of agency in agent-based models. arXiv.

IBM (2023): What Is Synthetic Data? | IBM, https://www.ibm.com/think/topics/synthetic-data (29. 7. 2025).

JAIN, S., RONALD, N., WINTER, S. (2015): Creating a Synthetic Population: A Comparison of Tools.

JORDON, J., HOUSSIAU, F., CHERUBIN, G., COHEN, S. N., SZPRUCH, L., BOTTARELLI, M. (2022): Synthetic Data - what, why and how? https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Synthetic_Data_Survey-24.pdf

KAM BRNO (2022): Průzkum dopravního chování, https://kambrno.cz/pruzkum-dopravniho-chovani/ (28. 4. 2026).

KONDURI, K., YOU, D., GARIKAPATI, V., PENDYALA, R. (2016): Enhanced Synthetic Population Generator That Accommodates Control Variables at Multiple Geographic Resolutions. Transportation Research Record: Journal of the Transportation Research Board, 2563, 40–50. doi: 10.3141/2563-08

LAMBERTI, A. (2023): The benefits and limitations of generating synthetic data, https://syntheticus.ai/blog/the-benefits-and-limitations-of-generating-synthetic-data (28. 7. 2025).

LENTI, J., COSTANTINI, L., FOSCH, A., MONTICELLI, A., SCALA, D., PANGALLO, M. (2025): Population synthesis with geographic coordinates. arXiv.

LIM, S. Y., YUN, H., BANSAL, P., KIM, D.-K., KIM, E.-J. (2025): A Large Language Model for Feasible and Diverse Population Synthesis. arXiv.

LOMAX, N., ARCHER, L. (2020): Alan Turing Institute-Vivarium Population SPENSER. The Alan Turing Institute.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY (2022): SimMobility. Singapore-MIT Alliance for Research and Technology.

MAURER, S. (2020): UDST/urbansim: Platform for building statistical models of cities and regions, https://github.com/UDST/urbansim (28. 4. 2026).

MEISTER, K., RIESER, M., CIARI, F., HORNI, A., BALMER, M., AXHAUSEN, K. (2008): Anwendung eines agentenbasierten Modells der Verkehrsnachfrage auf die Schweiz.

MINISTERSTVO DOPRAVY, T. (2026b): Poskytnuté informace, https://md.gov.cz/Ministerstvo/Zadost-o-poskytnuti-informace-(1)/Poskytnute-informace (28. 4. 2026).

MINISTERSTVO DOPRAVY, T. (2026a): Statistiky, https://md.gov.cz/Statistiky (28. 4. 2026).

MINISTERSTVO PRÁCE A SOCIÁLNÍCH VĚCÍ (2026a): Otevřená Data MPSV, https://data.mpsv.cz/portal/sestavy/statistiky/trh-prace/nezamestnanost/statistiky-nezamestnanosti (28. 4. 2026).

MINISTERSTVO PRÁCE A SOCIÁLNÍCH VĚCÍ (2026b): Pololetní statistiky absolventů škol a mladistvých v evidenci ÚP ČR | MPSV, https://mpsv.gov.cz/pololetni-statistiky-absolventu (28. 4. 2026).

MŠMT ČR (2025): Statistické výstupy a analýzy, https://msmt.gov.cz/vzdelavani/skolstvi-v-cr/statistika-skolstvi/statisticke-vystupy-a-analyzy (13. 9. 2025).

MÜLLER, K., AXHAUSEN, K. W. (2010): Population synthesis for microsimulation: State of the art. https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/152146/eth-1650-01.pdf

OPEN AI (2025a): Tabulka shrnutí metod generování syntetické populace, ChatGPT, https://chatgpt.com/?locale=cs-CZ (27. 8. 2025).

OPEN AI (2025b): Zákon č. 106/1999 Sb., ChatGPT, https://chatgpt.com/?locale=cs-CZ (27. 8. 2025).

PARK, J. S., O'BRIEN, J. C., CAI, C. J., MORRIS, M. R., LIANG, P., BERNSTEIN, M. S. (2023): Generative Agents: Interactive Simulacra of Human Behavior. arXiv.

PELLEGRINO, M., MOOIJ, J. de (2024): A Practical Agent Programming Language/Synthetic Population The Hague South West: A Synthetic Population for the South-West District of The Hague, The Netherlands, constructed using GenSynthPop-Python, https://github.com/A-Practical-Agent-Programming-Language/Synthetic-Population-The-Hague-South-West (28. 4. 2026).

PELLEGRINO, M., MOOIJ, J. de, SONNENSCHEIN, T., DASTANI, M., ETTEMA, D., LOGAN, B., VERSTEGEN, J. A. (2023): GenSynthPop: Generating a Spatially Explicit Synthetic Population of Agents and Households from Aggregated Data. Research Square.

PEREIRA, A. M., DINGIL, A. E., PŘIBYL, O., MYŠKA, V., VOREL, J., KŘÍŽ, M. (2022): An Advanced Travel Demand Synthesis Process for Creating a MATSim Activity Model: The Case of Ústí nad Labem. Applied Sciences, 12, 19, 10032. doi: 10.3390/app121910032

PopGen (2024): MARG - Mobility Analytics Research Group, https://www.mobilityanalytics.org/popgen.html (9. 11. 2024).

RYAN, J., MAOH, H., KANAROGLOU, P. (2009): Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. Geographical Analysis, 41, 2, 181–203. doi: 10.1111/j.1538-4632.2009.00750.x

SOCIOLOGICKÝ ÚSTAV AKADEMIE VĚD ČR (2020): Známé neznámé registrované partnerství, https://www.soc.cas.cz/images/drupal/soubory/tz_20200626_zname_nezname_registrovane_partnerstvi_aktualizovano.pdf (28. 2. 2026).

SPIELAUER, M. (2011): What is Social Science Microsimulation? Social Science Computer Review, 29, 1, 9–20. doi: 10.1177/0894439310370085

STAFF, C. (2025): A Beginner's Guide to the Bayesian Neural Network, Coursera, https://www.coursera.org/articles/bayesian-neural-network (25. 7. 2025).

STATUTÁRNÍ MĚSTO BRNO (2021): Průzkum budov v Brně / Buildings research in Brno (2018-2020), https://data.brno.cz/datasets/d3afcb7538e04a258353a56e9e94b9cd_0/about (17. 5. 2026).

STATUTÁRNÍ MĚSTO BRNO (2025): Počet osob na adresních místech / ODAE T ROB | data.Brno, https://data.brno.cz/datasets/1a80454f0cf5431183054a62d0244501_0/explore?filters=eyJhZ3JlZ2FjZSI6WyJhbm8iXX0%3D&location=49.206798%2C16.604903%2C14 (18. 2. 2026).

ŠANDA, R. (2022): Administrativní zdroje dat ve sčítáních lidu se zaměřením na sčítání 2011 v Česku, Statistika, https://csu.gov.cz/administrativni-zdroje-dat-ve-scitanich-lidu-se-zamerenim-na-scitani-2011-v-cesku (8. 9. 2025).

ŠANDA, R. (2023): Využití administrativních zdrojů dat při vymezení obyvatelstva ve sčítání lidu 2021 v Česku, Statistika, https://csu.gov.cz/vyuziti-administrativnich-zdroju-dat-pri-vymezeni-obyvatelstva-ve-scitani-lidu-2021-v-cesku (8. 9. 2025).

THE DECISION LAB (2026): The Decision Lab - Behavioral Science, Applied., The Decision Lab, https://thedecisionlab.com/reference-guide/statistics/synthetic-population (30. 7. 2025).

TOMŠÍK, J. (2012): Řídké matice a jejich použití. Masarykova univerzita, Přírodovědecká fakulta. https://is.muni.cz/th/tn3j0/ (16. 7. 2025).

ÚZIS (2025a): První krok ke zdraví: Zdravotnická data | NZIP, NZIP.cz, https://www.nzip.cz/zdravotnicka-data (13. 9. 2025).

ÚZIS (2025b): Syntetická data NZIS, NZIP.cz, https://www.nzip.cz/clanek/2230-synteticka-data-nzis (13. 9. 2025).

Vstupní tabulky (2026): https://drive.google.com/file/d/1WPIUqrSC34hrmfC64LsPN3LPyh28X_5q/view?usp=sharing (28. 4. 2026).

WANG, M., WANG, Y., LI, B., CAI, Z., KANG, M. (2022): A Population Spatialization Model at the Building Scale Using Random Forest. Remote Sensing, 14, 8, 1811. doi: 10.3390/rs14081811

YAMEOGO, B. F., GASTINEAU, P., HANKACH, P., VANDANJON, P. O. (2020): Comparing Methods for Generating a Two-Layered Synthetic Population. Transportation Research Record, 2675, 1, 136–147. doi: 10.1177/0361198120964734

YE, X., KONDURI, K., PENDYALA, R., SANA, B. (2009): Methodology to match distributions of both household and person attributes in generation of synthetic populations.

ZAGHENI, E. (2015): Microsimulation in Demographic Research. International Encyclopedia of the Social & Behavioral Sciences, doi: 10.1016/B978-0-08-097086-8.31018-2

Zpracování prostorových dat (2026): Google Docs, https://drive.google.com/file/d/1e83O8gda0eAxIh5RcTkY9V1thfffe4Wr/view?usp=sharing&usp=embed_facebook (28. 4. 2026).

PŘÍLOHY

SEZNAM PŘÍLOH

Přílohy jsou dostupné v Informačním systému Masarykovy univerzity:

Příl. 1 Struktura vstupních tabulek

Příl. 2 Vstupní tabulky a dokument

Příl. 3 QR kód na github upraveného modelu

Příl. 4 Diagram zpracování prostorových dat

Příl. 5 Diagram tvorby datasetu pater

Příl. 6 Diagram prvotního přiřazení domácností do pater budov

Příl. 7 Diagram doplnění budov domácnostmi

Příl. 8 QR kód na aplikaci