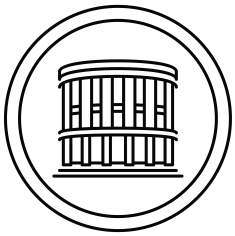
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



IMPUTÁCIA VÝDAVKOV ZO ZISŤOVANIA ŠTATISTIKY RODINNÝCH ÚČTOV DO ZISŤOVANIA EU-SILC POMOCOU PRAVDEPODOBNOSTNÝCH METÓD

DIPLOMOVÁ PRÁCA

2025 Bc. Lucia Bakošová

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Imputácia výdavkov zo zisťovania štatistiky rodinných účtov do zisťovania EU-SILC pomocou pravdepodobnostných metód

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko - finančná matematika a modelovanie
Študijný odbor:	1113 Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Norbert Švarda, PhD.

Bratislava 2025 **Bc. Lucia BAKOŠOVÁ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Lucia Bakošová
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Imputácia výdavkov zo zisťovania štatistiky rodinných účtov do zisťovania EU-SILC pomocou pravdepodobnostných metód
Statistical matching of the Household Budget Survey with EU SILC

Anotácia: Mikrosimulačné modely pre odhad daní, odvodov a dávok sú založené na individuálnych dátach. Ako podkladové dáta sa najčastejšie používajú výberové zisťovania. Rôzne výberové zisťovania však zbierajú údaje z rozličných oblastí. Mikrosimulačné modely však pre čo najpresnejšie výsledky potrebujú mať všetky informácie spojené. Pre spojenie rozdielnych výberových zisťovaní zbieraných na rovnakej populácii existuje viacero parametrických aj neparametrických štatistických metód. Úlohou je zistiť, ktoré metódy by boli najvhodnejšie pre spojenie dvoch výberových zisťovaní, konkrétne HBS (Štatistika o rodinných účtoch) a EU-SILC.

Vedúci: Mgr. Norbert Švarda, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 11.01.2024

Dátum schválenia: 25.01.2024 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou sa chcem poďakovať svojmu vedúcemu diplomovej práce Mgr. Norbertovi Švardovi, PhD. za trpezlivosť, ochotu, pomoc, odborné rady a podnetné pripomienky, ktoré mi pomohli pri písaní tejto práce. Taktiež by som sa chcela poďakovať aj Mgr. Andrei Grancovej za jazykovú korektúru textu.

Abstrakt v štátnom jazyku

BAKOŠOVÁ, Lucia: Imputácia výdavkov zo zisťovania štatistiky rodinných účtov do zisťovania EU–SILC pomocou pravdepodobnostných metód [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Norbert Švarda, PhD., Bratislava, 2025, 76s

V tejto diplomovej práci sa zaoberáme imputáciou údajov o výdavkoch z výberového zisťovania rodinných účtov (RÚ) do výberového zisťovania o príjmoch a životných podmienkach (EU-SILC). Imputáciu vykonávame za účelom spojenia týchto dvoch výberových zisťovaní do jednej databázy, ktorá by slúžila ako podkladová databáza pre už existujúci slovenský daňovo - odvodový mikrosimulačný model SIMTASK. Na začiatku práce si detailne popisujeme obe vstupné výberové zisťovania a ich štruktúru dát a premenných. Následne vysvetľujeme teoretické základy imputačného procesu, predpoklady a predstavujeme tri najpoužívanejšie metódy na imputáciu výdavkov a to konkrétne lineárnu regresiu, metódu hot deck matching a predictive mean matching. Praktická časť diplomovej práce sa zaoberá imputáciou výdavkov použitím neparametrickej Hot deck matching metódy. V práci uvažujeme viacero imputačných alternatív v závislosti od výberu vysvetľujúcich premenných. Imputácia prebieha v softvéri STATA. Výsledný dátovy súbor EU-SILC s naimputovanými výdavkami validujeme voči agregovaným výdavkom z databázy RÚ použitím váh domácnosti.

Kľúčové slová: rodinné účty, EU-SILC, imputácia, výdavky, hot deck matching

Abstract

BAKOŠOVÁ, Lucia: Statistical matching of the Household Budget Survey with EU SILC [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Norbert Švarda, PhD., Bratislava, 2025, 76p

In this thesis, we focus on statistical matching of Household Budget Survey (HBS) and Survey on Income and Living Conditions (EU-SILC). The purpose of statistical matching is to combine these two surveys into a single database, which would serve as an underlying database for the existing Slovak tax-benefit microsimulation model SIMTASK. At the beginning of the thesis, we provide a detailed description of both surveys, including their data structures and variables. Secondly, we explain the theoretical foundations of statistical matching, its assumptions, and introduce the three most commonly used methods for expenditure imputation: linear regression, Hot deck matching and Predictive mean matching. Practical part of the thesis focuses on implementation of imputation of expenditure data using a non-parametric Hot deck matching method. We propose several imputation alternatives depending on the choice of explanatory variables. Matching procedure is performed using STATA software. The final EU - SILC dataset with imputed expenditure data is validated against aggregated expenditure figures from HBS, using household weights to ensure comparability.

Keywords: household budget survey, EU-SILC, imputation, expenditure, hot deck matching

Obsah

Úvod **9**

1 Prehľad literatúry **11**

2 Výberové zisťovania **14**

2.1 Výberové zisťovanie rodinných účtov 14

2.2 Výberové zisťovanie o príjmoch a životných podmienkach domácnosti EU - SILC 16

3 Imputácia dát **18**

3.1 Imputačný proces 18

3.2 Príprava dát a predpoklady 18

3.2.1 Cramerov V koeficient 19

3.2.2 Predpoklad o podmienenej nezávislosti 20

3.3 Imputačné metódy 21

3.3.1 Lineárna regresia 22

3.3.2 Dvojkroková lineárna regresia 22

3.3.3 Hot deck matching 23

3.3.4 Predictive mean matching 24

4 Príprava dát a overenie predpokladov **26**

4.1 Tvorba dodatočných premenných 26

4.2 Overenie predpokladov a výber spoločných premenných 30

5 Praktická implementácia imputácie **36**

5.1 Prenos výdavkov z RÚ do EU - SILC 36

5.2 Validácia imputačnej metódy 38

5.2.1 Validácia voči databáze RÚ 38

5.2.2 Validácia pomocou porovnania distribúcie 46

5.2.3 Validácia pomocou analýzy zmeny korelačnej štruktúry 47

Záver **50**

Zoznam použitej literatúry **52**

A Príloha **56**

B Príloha **57**

C Príloha **64**

D Príloha **67**

E Príloha **72**

F Príloha **75**

G Príloha **76**

Úvod

V súčasnosti vzniká neustála potreba sledovania a analyzovania rôznych ekonomických dát. Jednou skupinou z nich sú aj dáta o správaní sa domácnosti, a to konkrétne ich výdavky a príjmy. Okrem ich súčasného správania nás však zaujímajú aj rôzne hypotetické scenáre, ktoré môžu nastať pri zmene iných faktorov, ktoré sa ich priamo týkajú, akými sú napríklad dane, odvody, sociálne dávky a iné legislatívne zmeny.

Na simulovanie správania sa domácnosti pri zmene externých podmienok nám slúžia takzvané mikrosimulačné modely. Sú to analytické nástroje, ktoré simulujú správanie jednotlivcov resp. domácnosti ešte predtým, než sa zmeny zavedú, prípadne predtým než sa jednotlivci stihnú týmto zmenám prispôsobiť, čo nám dáva možnosť včas reagovať a optimalizovať opatrenia. Jedným takým mikrosimulačným modelom je aj slovenský daňovo - odvodový model SIMTASK, ktorý bol vyvinutý analytikmi Kancelárie Rady pre rozpočtovú zodpovednosť. Jeho vývoj bol inšpirovaný mikrosimulačným modelom EUROMOD, ktorý simuluje daňovo odvodový a sociálny systém európskych krajín. Simulácie, ktoré poskytuje SIMTASK slúžia na vyhodnotenie dopadov zmien v daniach, odvodoch a sociálnych dávkach pre jednotlivcov a domácnosti [14].

Mikrosimulačný model je založený na individuálnych dátach, na mikroúrovni. Ako podkladové dáta sa používajú výberové zisťovania, pričom každé výberové zisťovanie sleduje údaje z rôznych oblastí, no stále na tej istej populácii. Aby mikrosimulačný model mohol poskytnúť čo najpresnejšie a najobjektívnejšie výsledky, je potrebné, aby boli všetky informácie spojené. Inými slovami, vstupom pre mikrosimulačný model musí byť len jeden dátový súbor. Z tohto dôvodu je potrebné nájsť spôsob, ako z viacerých výberových zisťovaní vytvoriť jednu databázu, ktorá obsahuje všetky požadované údaje a je dostatočne reprezentatívna pre skúmanú populáciu. Takýto proces spájania dvoch dátových súborov sa nazýva imputácia dát.

V súčasnosti existujú rôzne parametrické a neparametrické imputačné metódy, ktoré slúžia na spojenie dvoch výberových zisťovaní. Cieľom tejto diplomovej práce je vyskúšať neparametrický prístup imputovania výdavkov a to konkrétne tak, že budeme imputovať údaje o výdavkoch zo zisťovania rodinných účtov (RÚ) do zisťovania o príjmoch a životných podmienkach domácnosti (EU - SILC).

Prvá kapitola tejto diplomovej práce sa zameriava na stručný popis témy imputácie dát

a predstavuje súčasnú literatúriu zaoberajúcu sa touto problematikou. V druhej kapitole práce charakterizujeme obe vstupné výberové zisťovania. V tretej kapitole sa venujeme teoretickým základom celého imputačného procesu, vrátane jeho predpokladov a postupu predprípravy údajov. Ďalej predstavujeme 3 najznámejšie metódy, ktoré sa v súčasnosti používajú na imputáciu výdavkov a to konkrétne lineárnu regresiu, metódu Hot deck matching a metódu Predictive mean matching. V štvrtej kapitole si predstavujeme prípravu dát a tvorbu dodatočných premenných. V poslednej kapitole vyhodnocujeme kvalitu výsledkov dosiahnutých pomocou Hot deck matching metódy.

1 Prehľad literatúry

Na úvod diplomovej práce si predstavíme problematiku spajania rôznych dátových súborov a poskytneme prehľad súčasnej literatúry, ktorá sa zaoberá témou imputácie dát a konkrétne aj imputáciou výdavkov. V práci vychádzame z veľkej väčšiny zo zdrojov, ktoré sú úradnými publikáciami Eurostatu [12], [8], Spoločného výskumného centra Európskej komisie, [2] alebo pochádzajú od autorov pôsobiacich v štatistickej inštitúcii inej európskej krajiny [4].

V posledných rokoch začala vznikať potreba modernizácie sledovania a zbierania rôznych štatisticko - ekonomických dát, ktorá by poskytla lepší pohľad na meranie nielen stavu ekonomiky ale aj toho, ako sa aktuálna ekonomická situácia premieta do každodenného života domácnosti a jednotlivcov. Problémom je však to, že väčšina prieskumov a výberových zisťovaní nevie pokryť širokú škálu všetkých požadovaných údajov. Z tohto dôvodu v roku 2010 začal Eurostat zostrojovať metodologický postup imputovania dát (*angl. statistical matching*) za účelom spájania viacerých dátových súborov do jedného [12].

Publikácia [12] od autorov Leulescu (Eurostat) a Agafitei (Eurostat) z roku 2013 poskytuje ucelený prehľad a teóriu imputovania dát. V práci sa definuje základny pojem imputovania dát, princípy imputačného procesu a identifikujú sa štatistické predpoklady, ktoré musia byť pri imputácii dát splnené. Práca stručne predstavuje hlavné skupiny imputačných metód, ktorými su parametrické (regresné), neparametrické a zmiešane metódy. Publikácia obsahuje aj dve konkrétne prípadové štúdie, kde sa autori zameriavajú na imputáciu údajov pomocou neparametrickej metódy Hot deck matching. V prvej štúdii sa spájajú údaje z databázy EU-SILC s údajmi z databázy európského prieskumu kvality života - EQLS (*European Quality of Life Surveys*). Druhá prípadova štúdia vykonáva imputáciu údajov o mzdách z EU-SILC do databázy prieskumu o pracovnej sile - LFS (*Labour force survey*).

V závere tejto publikácie autori uvádzajú, že imputácia údajov predstavuje vcelku efektívny nástroj na tvorbu jedeného integrovaného dátového súboru. Ich účinnosť však závisí od konzistencie a predikčnej sily premenných v oboch súboroch a voľby konkrétnej imputačnej metódy. Napriek tomu, že sa autori nezameriavajú priamo na imputáciu výdavkov, práca poskytuje dobrý teoretický rámec imputovania dát, ktorý sa dá uplatniť aj

pri imputovaní výdavkov.

Iným prínosným obsahom k tejto téme je dokument [8] z roku 2013, ktorý bol rovnako ako predchádzajúca publikácia zverejnený Eurostatom. Na rozdiel od [12], tento dokument sa venuje už priamo praktickej imputácii výdavkov z databázy rodinných účtov (*angl. Houshold budget survey - HBS*) do databázy EU-SILC a to pomocou metódy Hot deck matching. Proces imputácie výdavkov je rozdelený na prípravnú fázu, počas ktorej sa kontroluje vhodnosť dát na imputáciu a na validačnú fazu, kde sa validuje správnosť, presnosť a výkonnosť imputačnej metódy.

Ďalším článkom, ktorý sa venuje výlučne imputácii výdavkov zo zisťovania rodinných účtov do zisťovania EU-SILC je článok [4], ktorý bol v roku 2014 uverejnený v časopise *International Journal of Economic Sciences*. Článok sa zameriava na imputáciu údajov o výdavkoch do zisťovania EU-SILC v Taliansku taktiež pomocou Hot deck matching metódy. V článku je podobne ako v [12] vysvetlený základný princíp imputačného procesu a štatistické predpoklady imputácie. V závere sa autori venujú validácii tejto imputačnej metódy.

Posledným kľúčovým zdrojom, z ktorého pri tejto diplomovej práci vychádzame, je [2]. Ide o publikáciu zo Spoločného výskumného centra Európskej komisie z roku 2020 od autorov Akoğuz, Capéau, Decoster, De Sadeleer, Güner D, Manios, Paulus a Vanheukelom. Táto publikácia predstavuje kompletný teoretický výklad imputácie výdavkov zo zisťovania rodinných účtov do zisťovania EU-SILC a to pre účely mikrosimulačného modelu EUROMOD, ktorý simuluje dopad zmien v daniach, odvodoch a sociálnych dávkach na domácností v Európskej únii [25]. Keďže EU-SILC je podkladová databáza pre EUROMOD, je potrebné do nej doplniť údaje o výdavkoch prostredníctvom imputácie [2]. Dokument [2] poskytuje na rozdiel od predchádzajúcej literatúry aj detailnejší a algoritmický postup jednotlivých imputačných techník. Taktiež spomína rôzne problémy a výzvy, ktoré sa zvyknú v priebehu imputácie výdavkov vyskytovať. Následne vykonáva imputáciu výdavkov pre vybrané krajiny EÚ. V závere uvádza rôzne prístupy k validácii, vrátane validácie voči databáze rodinných učtov, čo je pre účely tejto diplomovej práce kľúčové.

V prípade Slovenska bola imputácia výdavkov zo zisťovania rodinných účtov do zisťovania EU-SILC vykonaná ako súčasť vývinu mikrosimulačného modelu SIMTASK, ktorý

podobne ako EUROMOD simuluje dopad zmien v daniach, odvodoch a sociálnych dávkach. V tomto prípade bola imputácia vykonaná pomocou parametrickej metódy, konkrétne pomocou dvojkrokovej lineárnej regresie. Pri validácii tejto metódy sa však ukázalo, že regresný prístup neposkytol dostatočný prenos výdavkov pre všetky výdavkové kategórie [16]. Z tohto dôvodu vzniká potreba implementovania inej imputačnej techniky, ktorá by priniesla lepšie výsledky.

2 Výberové zisťovania

V tejto diplomovej práci budeme pracovať s dvoma výberovými zisťovaniami a to s výberovým zisťovaním rodinných účtov (*ozn.* RÚ) a výberovým zisťovaním o príjmoch a životných podmienkach domácností (*ozn.* EU - SILC). Pri charakteristikách výberových zisťovaní v tejto kapitole vychádzame z [9], [21], [23], [24].

Definícia 2.1. *Výberové zisťovanie je metodika získavania informácií o skúmanej populácii, pri ktorom sa prieskum realizuje na relatívne malej vzorke respondentov; v ideálnom prípade táto vzorka je náhodne zostavená a reprezentatívna z hľadiska skúmaných hlavných ukazovateľov [21].*

Definícia 2.2. *Výberový súbor je podmnožina základného súboru, ktorá sa vyberá podľa vopred stanovených pravidiel a postupu; nazýva sa aj ako výskumná vzorka [21].*

Inými slovami, výberové zisťovanie neprebieha na celej populácii, ale len na určitej podmnožine skúmanej populácie. Výberové zisťovanie teda poskytuje presné charakteristiky len pre výberový súbor. Základným súborom je v tomto prípade obyvateľstvo Slovenskej republiky.

2.1 Výberové zisťovanie rodinných účtov

Cieľom výberového zisťovania RÚ je zisťovať informácie o štruktúre, výške a vývoji výdavkov, spotreby a príjmov fyzických osôb súkromných hospodáriacich domácností. Hlavným cieľom zisťovania sú síce výdavky, avšak v rámci zisťovania sú sledované aj všetky peňažné a nepeňažné príjmy súkromných hospodáriacich domácností. V rámci výdavkov sú zisťované čisté peňažné výdavky, a to v súlade s detailným členením podľa medzinárodnej klasifikácie individuálnej spotreby COICOP [1]. Okrem príjmov a výdavkov sa sledujú aj základné demografické črty domácnosti, jej veľkosť podľa počtu členov, typy domácností, počet závislých detí, a tiež vybavenie domácnosti predmetmi dlhodobej spotreby.

Výberové zisťovanie rodinných účtov je pod správou Štatistického úradu Slovenskej republiky a zároveň je to jedno z najstarších zisťovaní. Vykonáva sa už od polovice 50-tych

[1] Classification of individual consumption by purpose je medzinárodne uznávaná klasifikácia, ktorá sa používa na kategorizáciu spotrebných výdavkov domácností podľa účelu tovaru. Bola vytvorená štatistickou divíziou OSN. Používa sa napríklad aj pri výpočte CPI indexu [9].

rokov 20. storočia. Údaje sú sledované na mesačnej báze a finálna databáza je zverejňovaná raz za rok. [24].

Databáza RÚ poskytuje údaje na úrovni domácnosti a zároveň na úrovni jednotlivcov. Výdavky v databáze sú reportované na ročnej báze. Štruktúru premenných v databáze RÚ môžeme pozorovať v Tabuľke 1 a v Tabuľke 2.

Poznámka 2.3. Pod pojmom domácnosť rozumieme súkromne hospodariacu domácnosť, ktorú tvoria osoby, ktoré bývajú spolu v tom istom byte (obydlí) a spoločne sa podieľajú na výdavkoch, najmä na bývaní a stravovaní [24].

Kód premennej	Názov cieľovej premennej	Hodnoty	Poznámka
HA04	Identifikácia domácnosti		
Kraj	Kraj	1 2 3 4 5 6 7 8	Bratislavský kraj Trnavský kraj Trenčiansky kraj Nitriansky kraj Žilinský kraj Banskobystrický kraj Prešovský kraj Košický kraj
⋮	⋮	⋮	⋮
HE01_p	Potraviny a nealkoholické nápoje	EUR	Ročný výdavok domácnosti
HE011_p	Potraviny	EUR	Ročný výdavok domácnosti
HE0111_p	Chlieb a obilniny	EUR	Ročný výdavok domácnosti
HE01111_p	Ryža	EUR	Ročný výdavok domácnosti
HE01112_p	Múka a ostatné obilniny	EUR	Ročný výdavok domácnosti
⋮	⋮	⋮	⋮

Tabuľka 1: Ukážka štruktúry premenných v databáze RÚ (na úrovni domácnosti), Zdroj: Databáza RÚ, Rada pre rozpočtovú zodpovednosť

Kód premennej	Názov cieľovej premennej	Hodnoty	Poznámka
MA04A	Identifikačné číslo domácnosti		
MB01A	Krajina narodenia	kód krajiny	
MB011A	Hlavné štátne občianstvo	kód krajiny	
MB04	Rodinný stav osoby	1	Slobodný/á
		2	Ženatý/vydatá
		3	Vdovec/vdova
		4	Rozvedený/rozvedená
		9	Neuvedené
⋮	⋮	⋮	⋮

Tabuľka 2: Ukážka štruktúry premenných v databáze RÚ (na úrovni osôb), Zdroj: Databáza RÚ, Rada pre rozpočtovú zodpovednosť

2.2 Výberové zisťovanie o príjmoch a životných podmienkach domácnosti EU - SILC

Cieľom výberového zisťovania EU- SILC je poskytovať informácie o príjmoch, úrovni a štruktúre chudoby a sociálnom vylúčení domácnosti na Slovensku. Zisťované sú údaje týkajúce sa príjmov, kvality bývania a finančnej situácie hospodáriacich domácnosti. Taktiež sa sledujú aj demografické charakteristiky členov domácností, vzdelanie, pracovná aktivita, vnímanie zdravia, subjektívne hodnotenie chudoby, materiálnej deprivácie a tiež hodnotenie kvality bývania.

Zisťovanie je pod správou Štatistického úradu Slovenskej republiky, no taktiež je súčasťou Európskeho štatistického programu. Je realizované na základe Nariadenia Európskeho parlamentu a Rady (EÚ) č. 2019/1700 a delegovaných nariadení na príslušný rok, čo znamená, že Štatistický úrad Slovenskej republiky k tomu zaväzuje aj európska legislatíva. Tá určuje konkrétne metodologické postupy, ktoré musia byť pri zisťovaní splnené.

Zisťovanie sa realizuje pravidelne ročne. Na Slovensku sa zisťovanie EU - SILC realizuje od roku 2005 [23]. Ukážku štruktúry dát v databáze EU - SILC môžeme pozorovať v Tabuľke 3 a v Tabuľke 4

Kód premennej	Názov cieľovej premennej	Hodnoty	Poznámka
DB030	Identifikácia domácnosti		Identifikačné číslo domácnosti
KRAJ	Kraj	1	Bratislavský kraj
		2	Trnavský kraj
		3	Trenčiasnky kraj
		4	Nitriansky kraj
		5	Žilinský kraj
		6	Banskobystrický kraj
		7	Prešovský kraj
		8	Košický kraj
⋮	⋮	⋮	⋮

Tabuľka 3: Ukážka štruktúry premenných v databáze EU-SILC (na úrovni domácnosti), zdroj: Databáza EU-SILC, Rada pre rozpočtovú zodpovednosť

Kód premennej	Názov cieľovej premennej	Hodnoty	Poznámka
RB030	Osobná identifikácia		Identifikačné číslo osoby
RB040	Identifikácia domácnosti		Identifikačné číslo domácnosti
RB080	Rok narodenia	1890 - 2023	Rok
⋮	⋮	⋮	⋮

Tabuľka 4: Ukážka štruktúry premenných v databáze EU-SILC (na úrovni osôb), Zdroj: Databáza EU-SILC, Rada pre rozpočtovú zodpovednosť

3 Imputácia dát

Imputácia dát je proces, pri ktorom imputujeme dáta z jedného dátového súboru do druhého väčšinou za účelom chýbajúcich hodnôt. V tejto diplomovej práci budeme vykonávať imputáciu za účelom spojenia dvoch dátových súborov do jedného.

V tejto kapitole si vysvetlíme princíp imputácie dát a jej rôzne metódy a techniky, ktoré sú v súčasnosti známe. Obzvlášť sa budeme venovať tým, ktoré v súčasnosti už boli použité na imputáciu výdavkov z RÚ do EU - SILC na Slovensku alebo v iných krajinách. Pri charakterizácii imputácie dát vychádzame z [2], [4] a [12].

3.1 Imputačný proces

Predpokladajme, že máme k dispozícii dva dátové súbory - súbor A a súbor B. V súbore A sú pozorované premenné (X, Y), ktoré sú reprezentatívne pre určitú populáciu a v súbore B sú pozorované premenné (X, Z), ktoré sú tiež reprezentatívne pre tú istú populáciu. Cieľom imputácie dát je vytvoriť jeden dátový súbor, ktorý obsahuje premenné (X, Y, Z) [4]. Premenná X, ktorá je sledovaná v oboch dátových súboroch sa nazýva spoločná premenná. Súbor z ktorého dáta imputujeme sa nazýva *darca* a súbor do ktorého dáta imputujeme sa nazýva *príjemca* [2].

V tejto diplomovej práci bude databáza RÚ darcom a EU - SILC príjemcom. Cieľom je dáta naimputovať tak, aby naimputované hodnoty boli podobné tomu, čo môžeme pre ne očakávať na základe informácie o združenom rozdelení spoločných premenných. Ak sú obe zisťovania reprezentatívne pre tú istú populáciu a na základe spoločných a imputovaných premenných v súbore príjemcu získame isté charakteristiky pre obyvateľstvo, tak tieto charakteristiky by mali byť zhodné s charakteristikami ktoré vieme vyčítať zo súboru darcu. Inými slovami, darca a príjemca (už s naimputovanými hodnotami) by mali vyzerať tak, akoby pochádzali z rovnákeho data - generujúceho procesu, resp. rovnakého rozdelenia [2].

3.2 Príprava dát a predpoklady

Rovnako ako pri každej štatistickej metóde, aj pri imputácii dát je najprv potrebné oboznámiť sa s predpokladmi metódy a podniknúť pred samotnou implementáciou určité

kroky. Z tohto dôvodu je potrebné najprv zaistiť, aby platili nasledovné princípy:

1. Oba dátové súbory musia byť navzájom koherentné. Súbory musia byť reprezentatívne pre tú istú populáciu, sledované za rovnaké časové obdobie a spoločné premenné musia byť rovnako definované, merané a identicky kategorizované, v opačnom prípade je potrebné ich rekategorizovať [4], [12].

2. Spoločné premenné v oboch súboroch musia pochádzať z aspoň približne rovnakého empirického rozdelenia. Spôsobov ako kvantifikovať ich mieru podobnosti je viacero. Pri imputácii výdavkov sa však štandardne využíva Hellingerova vzdialenosť v tvare

$$HD(V, V') = \sqrt{\frac{1}{2}\sum_{i=1}^{K}\left(\sqrt{P(V=i)} - \sqrt{P(V'=i)}\right)^2} = \sqrt{\frac{1}{2}\sum_{i=1}^{K}\left(\sqrt{\frac{n_{Di}}{N_D}} - \sqrt{\frac{n_{Pi}}{N_P}}\right)^2} \tag{1}$$

   kde V je rozdelenie v súbore darcu a V' je rozdelenie v súbore príjemcu. D označuje súbor darcu a P označuje súbor príjemcu. K je celkový počet možných nadobudnutých hodnôt v súbore darcu a príjemcu. n_{Di} je početnosť hodnoty i v súbore darcu D a n_{Pi} je početnosť hodnoty i v súbore príjemcu P. N je celkový počet pozorovaní v danom súbore. Výsledok vzorca (1) môže nadobúdať hodnoty od 0 po 1, pričom 0 indikuje, že rozdelenia sú zhodné, zatiaľ čo 1 hovorí, že rozdelenia sú úplne odlišné. Následne je možné nastaviť hranicu, po ktorú považujeme rozdelenia za rovnaké. V literatúre sa pri imputácii výdavkov štandardne dve rozdelenia považujú za dostatočne zhodné, ak ich Hellingerova vzdialenosť nie je väčšia než 0.05. [12]

3. Spoločné premenné, ktoré zvolíme na imputáciu, musia byť čo najkvalitnejšie v tom zmysle, že vedia dobre predikovať úroveň výdavkov. Konkrétna premenná by mala byť silne korelovaná s hodnotami výdavkov resp. mala by byť medzi nimi signifikantná asociácia. Toto je možné overiť napríklad pomocou Rao - Scottovho testu alebo Cramerovho V - koeficientu. [12]

3.2.1 Cramerov V koeficient

V tejto časti si predstavíme Cramerov V koeficient resp. Cramerovo V, pričom vychádzame z [1] a [17]. Cramerovo V slúži na vyjadrenie miery asociácie medzi dvoma kategorickými

premennými. V tejto diplomovej práci ho budeme neskôr využívať na kvantifikáciu závislosti medzi výdavkami a zvyšnými premennými. Cramerovo V sa počíta pomocou vzorca

$$V = \frac{\chi^2}{n(k-1)} \tag{2}$$

kde χ^2 je hodnota testovej štatistiky Chí kvadrát testu nezávislosti, n je počet pozorovaní a k je menší z počtu riadkov a stĺpcov v kontingenčnej tabuľke. Hodnota Cramerovho V koeficientu nadobúda hodnoty od 0 po 1 a má nasledovnú interpretáciu:

- $V \in [0, 0.05]$ - žiadna resp. veľmi slabá asociácia
- $V \in (0.05, 0.1]$ - slabá asociácia
- $V \in (0.1, 0.15]$ - mierna asociácia
- $V \in (0.15, 0.25]$ - silná asociácia
- $V \in (0.25, 1]$ - veľmi silná asociácia

3.2.2 Predpoklad o podmienenej nezávislosti

Ak sme prešli bodmi 1 až 3 z časti 3.2, potom je možné aplikovať konkrétnu imputačnú metódu. V súčasnosti existujú rôzne parametrické a neparametrické metódy imputácie dát, ktoré si predstavíme v časti 3.3. Väčšina imputačných techník, ktorú ponúka súčasná literatúra predpokladá dva kľúčove predpoklady:

1. Pozorovania sú nezávislé a rovnomerné rozdelené
2. Platí podmienená nezávislosť medzi premennými Y a Z pri podmieňovaní na premennú X (angl. *Conditional Independence Assumption (CIA)*[2])

Podľa [6] sa CIA predpoklad dá matematicky zapísať ako

$$P(X, Y, Z) = P(Y|X)P(Z|X)P(X) \tag{3}$$

Pokiaľ CIA predpoklad platí, tak predpokladáme, že vzťah medzi Y a Z vieme kompletne vysvetliť pomocou premennej X. Pokiaľ je predpoklad o CIA splnený, tak imputačný proces prebehne v poriadku a výsledky budú presné. Ak tento predpoklad splnený

[2]Pre jednoduchosť budeme tento predpoklad ďalej v diplomovej práci označovať už len ako CIA

nie je, výsledky môžu byť skresľujúce. Predpoklad v reálnych dátach málokedy platí no na druhej strane, vopred tento predpoklad nevieme ani nijak overiť, kedže v prípade imputácie výdavkov nemáme údaje o výdavkoch v jednom zo súborov dostupné. [4],[12].

Súčasna literatúra [4], [12] ponúka dve hlavné riešenia na zmiernenie problému vyvolaného kvôli CIA a to konkrétne:

- Využitie pomocnej informácie
- Analýza neistoty

Pomocnou informáciou v tejto súvislosti môže byť napríklad dodatočný dátový súbor, kde by boli pozorované všetky premenné naraz (X, Y, Z), ktorý môže slúžiť na validáciu výsledkov imputácie [4]. Alternatívnym riešením je zavedenie takzvanej proxy premennej, ktorá je odvodená z dostupných premenných a zároveň má dobrú schopnosť predikcie [12]. Takou premennou môže byť podľa [8] napríklad vytvorenie príjmových kvintilov.

Pod pojmom analýza neistoty rozumieme zostavenie intervalov, ktoré nám určujú mieru neistoty rozdelenia premenných Y, Z za podmienky daného X. Intervaly je možné zostaviť pomocou Frechetových ohraničení:

$$max(0, \theta_{y.|x} + \theta_{.z|x} - 1) \leq \theta_{yz|x} \leq min(\theta_{y.|x}, \theta_{.z|x}) \tag{4}$$

kde θ je pravdepodobnosť a bodka symbolizuje podmieňovanie [4]. Frechetove hranice nám poskytujú všetky možné hodnoty združenej distribúcie premennych Y, Z (za podmienky X), za predpokladu, že poznáme distribúcie Y (za podmienky X) a Z (za podmienky X). V prípade, že tieto intervaly vyjdú príliš široké, symbolizuje to vysokú mieru neistoty. Ak je interval úzky, tak takéto rozdelenie môže byť považované za vhodné na imputáciu [4].

3.3 Imputačné metódy

V tejto časti si predstavíme 3 metódy, ktoré sa využívajú pri imputovaní výdavkov z RÚ do EU-SILC. Pri ich charakterizácii vychádzame primárne z [2] a [12]. Ide o metódy, ktoré boli už v minulosti vyskúšané v iných krajinách Európy. V prípade Slovenska, imputácia výdavkov bola už v minulosti vykonaná a to pomocou dvojkrokovej lineárnej regresie, no úspešne sa podarilo naimputovať iba 11 z 13 výdavkových kategórií [16]. Z tohto dôvodu vzniká potreba implementácie iných imputačných metód.

3.3.1 Lineárna regresia

Prvou imputačnou metódou, ktorú si predstavíme je imputácia pomocou lineárnej regresie. Je to parametrická imputačná metóda a jej prístup spočíva v tom, že v súbore darcu sa odhadne lineárny regresný model v tvare

$$z_{dh} = \beta^T x_{dh} + \varepsilon_h \tag{5}$$

kde x_{dh} je vektor spoločných premenných domácnosti h, pozorovaný v súbore darcu d a z_{dh} je vektor premenných domácnosti h v súbore darcu d, ktoré budeme chcieť neskôr imputovať do súboru príjemcu.

Spoločné premenné sú teda vysvetľujúce premenné, pomocou ktorých chceme modelovať vysvetľované premenné, ktorými sú výdavky, ktoré potrebujeme naimputovať zo súboru darcu do súboru príjemcu. Model (5) je tým pádom zostavený na základe dát v súbore darcu. Následne sa použije na predikovanie hodnôt v súbore príjemcu [2].

3.3.2 Dvojkroková lineárna regresia

Veľmi častým a známym problémom pri imputácii výdavkov je imputácia nulových výdavkov. Jedná sa o také tovary a služby, ktoré sú využívané iba určitou časťou populácie a naopak vykazujú nulovú hodnotu pre veľký podiel domácností (napr. výdavky na tabak, alkohol, vzdelanie atď.). V takom prípade je Model (5) nepostačujúci, pretože predikovaná hodnota by bola nenulová. Pomocou Modelu (5) je prakticky nemožné identifikovať domácnosti, v ktorých je úroveň výdavku na danom tovare nulová. Riešením tohto problému je použitie dvojkrokovej lineárnej regresie. V prvom kroku sa predikuje či je podiel výdavkov na danom tovare z celkového disponibilného príjmu kladný a následne v druhom kroku sa modeluje výška podielu výdavku (za predpokladu, že je kladný).

Označme e_{dhi} ako výdavok domácnosti h v súbore darcu d na tovare i. Hodnotu výdavku e_{dhi} vyjadríme ako podiel z celkového disponibilného príjmu y_{dhi}

$$w_{dhi} = \frac{e_{dhi}}{y_{dhi}}, \quad i \in \mathcal{N}, \tag{6}$$

kde $\mathcal{N}$ je množina indexov pre jednotlivé tovary. Jednotlivé podiely výdavkov zlúčime do širších kategórií. Množinu kategórií označme ako $M = \{A, B \dots\}$. Prvky množiny M tým pádom určujú neprázdne a neprekrývajúce sa podmnožiny $\mathcal{N}_A, \mathcal{N}_B \dots$ a platí, že ich

zjednotenie je $\mathcal{N}$. Potom podiel výdavku kategórie X na disponibilnom príjme je

$$W_{dhX} = \sum_{i \in \mathcal{N}_X} w_{dhi}.$$

Po takejto agregáci výdavkov do širších kategórií následne modelujeme pravdepodobnosť, že domácnosť vykazuje kladnú hodnotu W_{dhX}, konkrétne pomocou probit modelu

$$P(W_{dhX} > 0) = 1 - \phi(-\gamma_X^T x_{dh}) = \phi(\gamma_X^T x_{dh}), \tag{7}$$

kde ϕ je distribučná funkcia normalizovaného normálneho rozdelenia, x_{dh} je vektor spoločných premenných v súbore darcu a γ_X je vektor odhadnutých parametrov pre kategóriu X.

Následne sa lineárnou regresiou odhadne model pre W_{dhX}, ktorý je analogický k Modelu (5)

$$W_{dhX} = \beta_X^T x_{dh} + \varepsilon_{hX}, \qquad W_{dhX} > 0. \tag{8}$$

Posledným krokom je imputovanie predikovaných hodnôt

$$\widehat{W}_{dhX} = \phi(\hat{\gamma}_X^T x_{dh}) \hat{\beta}^T x_{dh} \tag{9}$$

do súboru príjemcu, kde $\hat{\gamma}_X$ sú odhadnuté parametre pre Model (7) a $\hat{\beta}_X$ sú odhadnuté parametre pre Model (8). [2]

3.3.3 Hot deck matching

Ďalšou alternatívnou metódou, ktorou je možné imptutovať dáta o výdavkoch z RÚ do EU - SILC je prístup Hot deck matching. Pojem Hot deck matching (HDM) označuje triedu neparametrických metód, ktoré spočívajú v tom, že každému pozorovaniu zo súboru príjemcu je priradené jedno pozorovanie zo súboru darcu. Údaje o výdavkoch zo súboru darcu sú následne priradené ku spárovanému pozorovaniu v súbore príjemcu [2]. Hlavným znakom týchto metód je, že napríklad na rozdiel od lineárnej regresie, sa v tomto prípade imputujú "živé"hodnoty, ktoré skutočne existujú v súbore darcu [12].

Jednotlivé typy HDM metódy sa od seba navzájom líšia spôsobom párovania jednotlivých pozorovaní. V tejto diplomovej práci sa budeme zaoberať metódou, ktorá je z nich

najrozšírenejšia a to Distance Hot deck matching [3]. Tento variant meria vzdialenosť medzi dvoma pozorovaniami, pričom berie do úvahy spoločné premenné.

Spôsobov merania vzdialenosti medzi dvoma pozorovaniami je taktiež viacero. Predpokladajme, že by spoločná premenná bola iba jedna. V takom prípade, bude výsledná vzdialenosť medzi dvoma pozorovaniami absolútnou hodnotou rozdielu medzi spoločnou premennou týchto dvoch pozorovaní, bez ohľadu na to, akú metriku zvolíme. Avšak ak máme spoločných premenných viac, potom už výber typu vzdialenosti zaváži. Pre účely imputácie sa môže použiť Manhattanská vzdialenosť alebo Euklidovská vzdialenosť. Najčastejšie sa však používa Mahalanobisova vzdialenosť v tvare:

$$d(x_{pg}, x_{dh}) = \sqrt{(x_{pg} - x_{dh})^T \Sigma^{-1} (x_{pg} - x_{dh})}, \tag{10}$$

kde x_{pg} je vektor spoločných premenných domácnosti g v súbore príjemcu p a x_{dh} je vektor spoločných premenných domácnosti h v súbore darcu d. Σ je kovariančná matica spoločných premenných. Hlavnou výhodou Mahalanobisovej vzdialenosti je, že zohľadňuje variancie jednotlivých premenných a kovariancie medzi nimi [2]. Zo vzťahu (10) pre Mahalanobisovu vzdialenosť je vidieť, že ak by premenné neboli korelované a všetky by mali jednotkovú disperziu, tak potom by platilo, že Σ by bola matica identity a zo vzťahu (10) by sme dostali obyčajnú euklidovskú vzdialenosť [13]. Mahalanobisova vzdialenosť tým pádom meria vzdialenosť medzi dvoma pozorovaniami, pričom zohľadňuje aj štatistické správanie dát. Vzťah (10) sa zvykne označovať aj ako dištančná funkcia.

Hlavným nedostatkom HDM metódy je, že jedno pozorovanie zo súboru darcu môže byť priradené viacerým pozorovaniam v súbore príjemcu a na druhej strane, niektoré pozorovania sa nepoužijú ani raz. Ak by sme jednu hodnotu naimputovali veľakrát, mohlo by to viesť ku skresleniu výsledkov, pretože rozdelenie imputovanej premennej v súbore príjemcu by mohlo byť odlišné od rozdelenia danej premennej v súbore darcu. [12].

3.3.4 Predictive mean matching

Poslednou metódou, ktorú si predstavíme v tejto diplomovej práci je metóda Predictive mean matching (PMM). Je to prístup, ktorý kombinuje metódu lineárnej regresie a HDM

[3] Pre prehľadnosť budeme ďalej v diplomovej práci túto metódu označovať len ako Hot deck matching resp. skrátene HDM.

metódy. PMM sa preto označuje za parametricko - neparametrickú metódu. V nasledujúcich krokoch uvádzame stručný algoritmus jej jednorozmernej verzie (imputácia len jedného výdavku):

1. Odhadneme lineárny regresný model v súbore darcu rovnako ako v modeli (5).

2. Odhadnutý model z časti 1. sa použije na predikovanie hodnoty výdavku v súbore darcu a zároveň aj v súbore príjemcu.

3. Zrátame vzdialenosti medzi predikovanými hodnotami výdavkov v súbore darcu a v súbore príjemcu. V tomto prípade je len absolútna hodnota rozdielu predikovaných hodnôt mierou rozdielu medzi domácnosťami v oboch súboroch. Pozorovanie zo súboru darcu, ktoré je v tomto zmysle najbližšie k pozorovaniu v súbore príjemcu s ním vytvorí pár.

4. Chýbajúcu hodnotu výdavku v súbore príjemcu naimputujeme so skutočnou pozorovanou hodnotou výdavku v súbore darcu.

Odhadnutie lineárneho regresného modelu sa pri PMM metóde prakticky využije len pre účel zmerania vzdialenosti. PMM metódu môžeme preto chápať ako špecifický typ HDM metódy, ktorá využíva taký typ metriky, kde sa premenným vstupujúcim do výpočtu vzdialenosti priraďujú regresné koeficienty ako váhy. Vzhľadom na to, že PMM metóda obsahuje regresný prvok, nie je vhodná pre imptutáciu výdavkov pri ich detailnom členení. Preto sa pri PMM odporúča združiť výdavky do širších kategórií a prípadne použiť dvojkrokovú lineárnu regresiu, rovnako ako sme to popísali v časti 3.3.2 [2].

4 Príprava dát a overenie predpokladov

Na základe teoretických znalostí, ktoré sme predstavili v druhej kapitole sme praktickú časť tejto práce začali harmonizáciou dát. V prípade niektorých premenných sa stalo, že v jednej databáze bolo iné kódovanie ako v tej druhej. Vzhľadom na to, že EU - SILC nám slúži ako súbor príjemcu, prispôsobili sme kódovanie premenných tejto databáze. Premenné, ktoré museli byť takýmto spôsobom prekódované v RÚ sú Hlavná ekonomická aktivita (samo-definovanie) a Rodinný stav referenčnej osoby [4].

4.1 Tvorba dodatočných premenných

Prvým nevyhnutným krokom bolo zjednotenie definície disponibilného príjmu v oboch databázach, pretože v RÚ sa príjmy agregujú na menších úrovniach ako v EU - SILC a zároveň obe databázy obsahujú premenné rôzneho variantu príjmu. V RÚ je premenná Čistý peňažný príjem, ktorá sa vypočítava z hrubého peňažného príjmu odpočítaním dane z príjmov a povinného osobného poistenia. *Hrubé peňažné príjmy predstavujú sumu všetkých peňažných príjmov každého člena domácnosti. Ide o súhrn peňažných príjmov zo zamestnania (vrátane dane z príjmov a povinných príspevkov do Sociálnej poisťovne a zdravotných poisťovní), peňažných čiastok, ktoré vyčlenila samostatne zárobkovo činná osoba zo svojho podnikateľského príjmu pre potreby domácnosti, sociálnych príjmov, peňažných darov, náhrad od poisťovní, výhier v lotérii, výberov pôžičiek a pod."* (Štatistický úrad Slovenskej republiky). V EU - SILC bolo preto potrebné vytvoriť premennú, ktorá by bola ekvivalentná tejto premennej z RÚ. To sme dosiahli odpočítaním nasledujúcich premenných od premennej Celkový disponibilný príjem domácnosti: Nepeňažný príjem zo služobného auta, Dôchodok z individuálnych súkromných dôchodkových fondov, Pravidelné prijaté peňažné transfery medzi domácnosťami, Úroky, dividendy, zisk z kapitálových investícií do neregistrovaného podniku. Následne sa ešte pripočítala premenná Pravidelné platené peňažné transfery medzi domácnosťami. Výsledkom bola novovytvorená premenná HY0020_alt.

Pre účely imputácie sa však neodporúča pracovať s absolútnymi hodnotami príjmov a výdavkov [8] a preto sme vytvorili dodatočné premenné, ktoré príjem a výdavky škálujú

[4] Pre jednoduchosť tu spomíname už len premenné, ktoré sme využili pri imputácii

na menšie hodnoty. Ide o zlogaritmovaný príjem, príjmový decil, príjmový ventil[5], výdavkový decil a výdavkový ventil. V prípade decilu bolo vytvorených 10 príjmových resp. výdavkových skupín, v prípade ventilu 20 skupín, pričom sa zohľadňovala aj váha domácnosti. Hodnoty príjmových decilov pre obe databázy môžeme pozorovať v Tabuľke 5. Hodnoty príjmových ventilov pre obe databázy vidíme v Tabuľke 6. Hodnoty výdavkových decilov a ventilov môžeme pozorovať v Tabuľke 7 a v Tabuľke 8. Vytvorenie príjmových skupín navyše pomáha uvoľňovať CIA predpoklad a tým pádom vytvoriť lepší model [8]. Logaritmickú funkciu sme taktiež aplikovali aj na premennú Rok narodenia.

Decil	Interval príjmu (RÚ)	Interval príjmu (EU - SILC)
1	(0, 8 147]	(0, 7 049.72]
2	(8 147, 11 492]	(7 049.72, 10 514.62]
3	(11 492, 13 782]	(10 514.62, 12 761.5]
4	(13 782, 15 779]	(12 761.5, 15 144.23]
5	(15 779, 17 950]	(15 144.23, 17 067.86]
6	(17 950, 19 671]	(17 067.86, 19 117.01]
7	(19 671, 22 007]	(19 117.01, 21 405.12]
8	(22 007, 24 575]	(21 405.12, 24 527.95]
9	(24 575, 29 251]	(24 527.95, 31 198.87]
10	viac ako 29 251	viac ako 31 198.87

Tabuľka 5: Hodnoty príjmových decilov v databáze RÚ a EU-SILC

Ventil	Interval príjmu (RÚ)	Interval príjmu (EU-SILC)
1	(0, 6 286]	(0.5, 5 087.25]
2	(6 286, 8 147]	(5 087.72, 7 049.72]
3	(8 147, 10 005]	(7 049.72, 8 768]
4	(10 005, 11 492]	(8 768, 10 514.62]
5	(11 492, 12 647]	(10 514.62, 11 710.89]
6	(12 647, 13 782]	(11 710.89, 12 761.5]

[5]Decily a ventily boli vypočítané použitím váh domácnosti so zaokrúhlením na celé číslo.

Ventil	Interval príjmu (RÚ)	Interval príjmu (EU-SILC)
7	(13 782, 14 675]	(12 761.5, 13 963]
8	(14 675, 15 779]	(13 963, 15 144.23]
9	(15 779, 17 013]	(15 144.23, 16 013.48]
10	(17 013, 17 950]	(16 013.48, 17 067.86]
11	(17 950, 18 885]	(17 067.86, 18 113.47]
12	(18 885, 19 671]	(18 113.47, 19 117.01]
13	(19 671, 20 747]	(19 117.01, 20 277.95]
14	(20 747, 22 007]	(20 277.95, 21 405.12]
15	(22 007, 23 282]	(21 405.12, 22 807.36]
16	(23 282, 24 575]	(22 807.36, 24 527.95]
17	(24 575, 26 230]	(24 527.95, 27 352.54]
18	(26 230, 29 051]	(27 352.54, 31 198.87]
19	(29 051, 33 710]	(31 198.87, 36 864.1]
20	viac ako 33 710	viac ako 36 864.1

Tabuľka 6: Hodnoty príjmových ventilov v databáze RÚ a EU-SILC

Decil	Interval výdavku	Decil	Interval výdavku
1	(0, 6 244.93]	6	(11 669.79, 13 180.66]
2	(6 244.93, 7 838.42]	7	(13 180.66, 14 853.11]
3	(7 838.42, 9 054.55]	8	(14 853.11, 17 329.2]
4	(9 504.55, 10 377.31]	9	(17 329.2, 21 911.05]
5	(10 377.31, 11 669.79]	10	viac ako 21 911.05

Tabuľka 7: Hodnoty výdavkových decilov - RÚ

Ventil	Interval výdavku	Ventil	Interval výdavku
1	(0, 5 221.2]	11	(11 669.79, 12 346.66]
2	(5 221.2, 6 244.93]	12	(12 346.66, 13 180.66]
3	(6 244.93, 7 192.04]	13	(13 180.66, 13 827.91]
4	(7 192.04, 7 838.42]	14	(13 827.91, 14 853.11]
5	(7 838.42, 8 504.2]	15	(14 853.11, 15 804.28]
6	(8 504.2, 9 054.55]	16	(15 804.28, 17 329.2]
7	(9 054.55, 9 769.26]	17	(17 329.2, 19 042.26]
8	(9 769.26, 10 377.31]	80	(19 042.26, 21 911.05]
9	(10 377.31, 11 012.61]	19	(21 911.05, 27 095.14]
10	(11 012.61, 11 669.79]	20	viac ako 27 095.14

Tabuľka 8: Hodnoty výdavkových ventilov - RÚ

Ďalšou dodatočnou premennou ktorú sme vytvorili bola premenná Typ domácnosti. V RÚ aj v EU-SILC sa premenná s takýmto názvom síce už nachádza, no obsahuje iba 7 kategórií a to Jednočlenná domácnosť, Osamelý rodič s aspoň jedným dieťaťom vo veku menej ako 25 rokov, Osamelý rodič so všetkými deťmi vo veku 25 a viac, Pár bez detí, Pár s aspoň jedným dieťaťom vo veku menej ako 25 rokov, Pár so všetkými deťmi vo veku 25 a viac a Iný typ domácnosti. Ak však chceme nájsť čo najpodobnejšiu domácnosť, je lepšie poznať detailnejšiu štruktúru domácnosti. Prvou alternatívou je využiť premennú Ekvivalentná veľkosť domácnosti, ktorá je počítaná pomocou modifikovanej OECD škály, konkrétne podľa vzorca:

$$\text{Ekvivalentná veľkosť domácnosti} = 1 + (0.7(A-1)) + (0.5B) \tag{11}$$

kde A je počet členov domácnosti vo veku 14 a viac rokov a B je počet členov domácnosti vo veku menej ako 14 rokov [6]. Premenná nám udáva modifikovanú veľkosť domácnosti, vďaka ktorej môžeme porovnávať domácnosti rôznej veľkosti a zloženia.

Inou alternatívou je vytvoriť premennú, ktorá zachytáva priamo počet členov domácnosti s dôrazom na počet závislých detí, keďže to môže byť pri predikovaní úrovne výdavkov smerodajné. Pomocou pôvodnej premennej Typ domácnosti, počtu dospelých osôb

[6] zdroj: Databáza RÚ, Rada pre rozpočtovú zodpovednosť

a počtu závislých detí, sme zostavili dodatočnú premennú Typ domácnosti s deviatimi úrovňami, ktoré môžeme pozorovať v Tabuľke 9.

Hodnota premennej	Počet dospelých	počet detí
1	1	0
2	2	0
3	3 a viac	0
4	1	1
5	2	1
6	1	2
7	2	2
8	1	3 a viac
9	2	3 a viac
10	3 a viac	3 a viac

Tabuľka 9: Encoding premennej Typ domácnosti (odvodená premenná)

Aby sa dala táto premenná použiť na hľadanie najbližších domácností, je potrebné aby malo kódovanie správne poradie odzrkadľujúce istú hierarchiu. Z Tabuľky 9 je možné pozorovať, že kódovanie premennej je určené podľa narastajúceho počtu dospelých členov a počtu závislých detí v domácnosti.

4.2 Overenie predpokladov a výber spoločných premenných

Základnou podmienkou pri imputácii dát je, aby spoločné premenné zo súboru darcu a súboru príjemcu pochádzali z približne rovnakého rozdelenia resp. vykazovali podobnú štruktúru. V tejto časti si zhrnieme základné štatistické charakteristiky spoločných premenných. V nasledujúcich tabuľkách uvádzame údaje o premenných bez využitia váh. Vážené údaje nám síce poskytujú obraz o celej populácii Slovenska, no v tejto časti sa zameriavame na kontrolu štatistického predpokladu, ktorým je zhodnosť rozdelení spoločných premenných. Vzhľadom na to, že pri imputácii sa váhy nevyužívajú, má preto zmysel v tejto súvislosti uvažovať nevážené údaje.

V Tabuľke 10 môžeme pozorovať základné štatistické charakteristiky numerických spo-

ločných premenných pre datbázu RÚ aj EU - SILC.

	Priemer		Štandardná odchýlka		Medián	
	RÚ	EU - SILC	RÚ	EU - SILC	RÚ	EU - SILC
Rok narodenia referenčnej osoby	1 963.88	1 962.75	15.07	14.6271	1 962	1 961
Disponibilný príjem	15 178.5	14 597.88	7 817.97	8 097.95	13 837.5	13 304.57
Ekvivalentná veľkosť domácnosti	1.6	1.64	0.56	0.6	1.5	1.5

Tabuľka 10: Základné charakteristiky spoločných premenných (bez zohľadnenia váh domácností), zdroj: Databáza RÚ, Databáza EU-SILC, Rada pre rozpočtovú zodpovednosť

Z Tabuľky 10 môžeme pozorovať, že uvedené premenné v oboch databázach vykazujú dobrú kvalitu z hľadiska podobného správania, čo naznačuje že dáta sú vhodné na imputáciu.

Pre zvyšné kategorické premenné sme si zostavili ešte tabuľky početností pre jednotlivé kategórie, ktoré môžeme pozorovať v Tabuľke 11 až 15.

	Percentuálny podiel v databáze	
Kraj	RÚ	EU - SILC
Bratislavský	14.93 %	15.90 %
Trnavský	10.85 %	9.57 %
Trenčiansky	11.53 %	11.42 %
Nitriansky	13.23 %	12.71 %
Žilinský	12.14 %	12.22 %
Banskobystrický	12.04 %	12.95 %
Prešovský	12.31 %	13.07 %
Košický	12.98 %	12.17 %

Tabuľka 11: Porovnanie početnosti jednotlivých krajov medzi databázou RÚ a databázou EU - SILC (bez zohľadnenia váh domácností), zdroj: Databáza RÚ, Databáza EU-SILC, Rada pre rozpočtovú zodpovednosť

	Percentuálny podiel v databáze	
Stupeň urbanizácie	RÚ	EU - SILC
Územie s hustým osídlením	42.78 %	25.99 %
Územie s priemerne hustým osídlením	24.38 %	36.53 %
Územie s riedkym osídlením	32.83 %	37.48 %

Tabuľka 12: Porovnanie početnosti premennej Stupeň urbanizácie medzi databázou RÚ a databázou EU - SILC (bez zohľadnenia váh domácností), zdroj: Databáza RÚ, Databáza EU-SILC, Rada pre rozpočtovú zodpovednosť

	Percentuálny podiel v databáze	
Rodinný stav	RÚ	EU - SILC
Slobodný / slobodná	18.81 %	14.77 %
Ženatý / vydatá	50.98%	52.63 %
Vdovec / vdova	15.83 %	18.08%
Rozvedený / rozvedená	14.38 %	14.52%

Tabuľka 13: Porovnanie početnosti premennej Rodinný stav medzi databázou RÚ a databázou EU - SILC (bez zohľadnenia váh domácností), zdroj: Databáza RÚ, Databáza EU-SILC, Rada pre rozpočtovú zodpovednosť

	Percentuálny podiel v databáze	
Typ domácnosti	RÚ	EU - SILC
1 dospelý	29.27 %	28.63 %
2 dospelí	35.36 %	33.60 %
viac ako 3 dospelí	9.54 %	10.75 %
1 dospelý a 1 dieťa	2.24 %	1.85 %
2 dospelí a 1 dieťa	7.30 %	7.70 %
1 dospelý a 2 deti	1.07%	0.89 %
2 dospelí a 2 deti	7.38 %	7.63 %
1 dospelý a viac ako 3 deti	0.14 %	0.16 %
2 dospelí a viac ako 3 deti	2.07 %	2.37 %
viac ako 3 dospelí a viac ako 3 deti	5.64 %	6.42 %

Tabuľka 14: Porovnanie početnosti premennej Typ domácnosti medzi databázou RÚ a databázou EU - SILC (bez zohľadnenia váh domácností), zdroj: Databáza RÚ, Databáza EU-SILC, Rada pre rozpočtovú zodpovednosť

	Percentuálny podiel v databáze	
Ekonomická aktivita referenčnej osoby	RÚ	EU - SILC
Zamestnaný	53.71 %	46.95 %
Nezamestnaný	0.92 %	2.43 %
Starobný dôchodca	42.66 %	46.67%
Invalidný dôchodca	2.09 %	2.95 %
Študent	0.18 %	0.05%
Osoba v domácnosti	0.12 %	0.19 %
Iné	0.31 %	0.75%

Tabuľka 15: Porovnanie početnosti premennej Ekonomická aktivita referenčnej osoby medzi databázou RÚ a databázou EU - SILC (bez zohľadnenia váh domácností), zdroj: Databáza RÚ, Databáza EU-SILC, Rada pre rozpočtovú zodpovednosť

Z Tabuľky 11 až 15 je opäť vidieť, že premenné sa v databáze EU - SILC správajú podobne ako v databáze RÚ okrem premennej Stupeň urbanizácie. Aby sme túto podobnosť vedeli presne kvantifikovať, vypočítali sme si pre každú spoločnú premennú hodnotu Hellingerovej vzdialenosti, konkrétne podľa vzorca (1) z časti 3.2. Porovnanie Hellingerových vzdialeností jednotlivých spoločných premenných môžeme pozorovať v Tabuľke 16.

Z Tabuľky 16 je vidieť, že všetky premenné vykazujú vcelku podobné rozdelenie v oboch databázach. Premenné Kraj, Príjmový decil, Rodinnný stav a Typ domácnosti dosahuju hodnotu Hellingerovej vzdialenosti nižšiu ako 0.05. Z tohto dôvodu môžeme tieto premenné považovať za najvhodnejšie premenné na imputovanie výdavkov. Na druhej strane, premenná Stupeň urbanizácie sa ukázala ako najmenej vhodná premenná pre imputáciu.

Ďalším krokom bolo zisťovanie, ktoré premenné majú najlepšiu schopnosť dobre predikovať úroveň výdavkov. Toto sme overili pomocou Cramerovho V koeficientu, ktorý sme si predstavili v časti 3.2.1. Vzhľadom na to, že Cramerovo V predpokladá kategorické premenné, v prípade výdavkov sme brali do úvahy výdavkové decily a ventily. Výsledné hodnoty koeficientov je možné pozorovať v Tabuľke 17.

Premenná	Hellingerova vzdialenosť
Kraj	0.0279
Stupeň urbanizácie	0.1313
Rok narodenia referenčnej osoby	0.0746
Príjmový decil	0.0488
Príjmový ventil	0.0585
Rodinný stav	0.0410
Typ domácnosti	0.0255
Ekvivalentná veľkosť domácnosti (OECD)	0.0515
Ekonomická aktivita	0.0668

Tabuľka 16: Hellingerove vzdialeností spoločných premenných (bez zohľadnenia váh domácností)

Premenná	Cramerovo V (výdavkové decily)	Cramerovo V (výdavkové ventily)
Kraj	0.0578	0.0736
Stupeň urbanizácie	0.0499	0.0705
Rok narodenia referenčnej osoby	0.2140	0.1728
Príjmový decil	0.2316	0.2397
Príjmový ventil	0.2396	0.1733
Rodinný stav	0.2330	0.2416
Typ domácnosti	0.2304	0.2381
Ekvivalentná veľkosť domácnosti	0.2406	0.1780
Ekonomická aktivita	0.1901	0.1955

Tabuľka 17: Cramerov V koeficient medzi výdavkami a spoločnými premennými

Z Tabuľky 17 môžeme vidieť že najvyššiu asociáciu s výdavkami dosahujú premenné Rok narodenia referenčnej osoby, Príjmový decil, Príjmový ventil, Rodinný stav, Typ domácnosti a Ekvivalentná veľkosť domácnosti. Relatívne vysoké hodnoty Cramerových V koeficientov indikujú, že na základe úrovne týchto premenných vieme získať informáciu o tom, ako sa približne správajú výdavky v závislosti od konkrétnej kategórie premennej.

V prípade výdavkov a príjmov, ktoré sú číselnými premennými, je možné ešte zohľadniť ich korelačné koeficienty, ktoré môžeme pozorovať v Tabuľke 18.

Pearsonov korelačný koeficient	Spearmanov korelačný koeficient
0.3547	0.4466

Tabuľka 18: Hodnoty korelačných koeficientov medzi príjmami a výdavkami v RÚ (so zohľadnením váh domácností)

Zvoliť premenné ktoré sa využijú pri imputácii znamenalo nájsť kompromis medzi vysokou hladinou Cramerovho V a hodnotou Hellingerovej vzdialenosti do 5 %. Premenné ktoré spĺňali tieto charakteristiky sú Kraj, Príjmový decil, Rodinný stav, Typ domácnosti. Pri tvorbe rôznych imputačných variantov sme však zaradili do imputačného procesu aj premenné, ktoré síce z predpokladov nevyšli ako vhodné premenné, no po ekonomickej stránke malo zmysel ich zahrnúť. Na základe validácie všetkých variantov, sa neskôr ukázalo, či tieto premenné priniesli modelu väčšiu presnosť, alebo nie.

5 Praktická implementácia imputácie

V tejto kapitole si predstavíme praktickú implementáciu imputácie výdavkov z RÚ do EU - SILC pomocou neparametrickej metódy Hot deck matching. V kapitole č. 4 sme vytvorili 5 dodatočných premenných a to Zlogaritmovaný rok narodenia, Zlogaritmovaný príjem, Príjmový decil, Príjmový ventil a Typ domácnosti. V nasledujúcej časti 5.1 uvedieme, o akých rôzynch alternatívach imputácie výdavkov v závislosti od kombinácie týchto premenných sme uvažovali. Pri detailnom validovaní výsledkov sa zameriavame iba na jeden finálny variant, ktorý sme zvolili na základe dobrej ekonomickej interpretácie. Oba dátove súbory s ktorými sme pracovali, boli za rok 2023.

Vzhľadom na podstatu HDM metódy, ktorou je nájdenie najbližšej vzdialenosti, a na vzorec pre výpočet Mahalanobisovej vzdialenosti, má zmysel medzi spoločné premenné zahrnúť iba numerické premenné respektíve kategorické premenné, ktoré majú hierarchické kódovanie, takzvané ordinálne premenné. Jediná premenná, ktorá nespĺňala toto kritérium bola premenná Ekonomická aktivita referenčnej osoby. Avšak, keďže táto premenná má dobrú schopnosť predikovania výdavkov domácnosti, bolo nutné ju pri imputácii zahrnúť do vektora spoločných premenných. To sme docielili tým, že sme oba dátové súbory (RÚ a EU-SILC) rozčlenili na menšie dátové súbory, a to podľa kategórie premennej Ekonomická aktivita referenčnej osoby. Získali sme tak v každej databáze 7 menších dátových súborov zvlášť pre kategóriu Zamestnaný, Nezamestnaný, Starobný dôchodca, Invalidný dôchodca, Študent, Osoba v domácnosti a kategóriu Iné. Imputačný proces následne prebehol v každom z týchto súborov zvlášť.

5.1 Prenos výdavkov z RÚ do EU - SILC

Proces programovania HDM metódy spočíval v zostrojení matice vzdialenosti medzi domácnosťami z RÚ a EU-SILC, a to zvlášť pre každú kategóriu premennej Ekonomická aktivita referenčnej osoby. Z oboch databáz sme odstránili tie pozorovania, ktoré nemali vyplnenú aspoň jednu premennú. Celkovo sme pracovali s 5 729 pozorovaniami z EU - SILC a 4 876 pozorovaniami z RÚ.

Označme k ako kategóriu premennej Ekonomická aktivita referenčnej osoby[7], n_{kr} ako

[7]Jej slovný popis môžeme pozorovať v Tabuľke 15

počet pozorovaní tejto kategórie v databáze RÚ a n_{ks} počet pozorovaní tejto kategórie v databáze EU - SILC. Matica vzdialeností MD_k pre kategóriu k premennej Ekonomická aktivita referenčnej osoby mala tvar

$$MD_k = \begin{pmatrix} d_{11k} & d_{12k} & \dots & d_{1jk} \\ d_{21k} & \dots & \dots & d_{2jk} \\ \vdots & \ddots & \ddots & \vdots \\ d_{i1k} & \dots & \dots & d_{ijk} \end{pmatrix} \quad \begin{array}{l} \forall \quad i = 1, 2 \dots n_{ks} \\ \qquad j = 1, 2 \dots n_{kr} \\ \qquad k = 1, 2, 3, 4, 5, 6, 7 \end{array} \tag{12}$$

Riadky reprezentujú domácnosti z EU - SILC a stĺpce domácnosti z RÚ. Hodnota d_{ijk} označuje hodnotu Mahalanobisovej vzdialenosti medzi i-tou domácnosťou z EU - SILC a j - tou domácnosťou z RÚ kategórie k. Za najbližšiu domácnosť z RÚ ku každej domácnosti z EU - SILC sa zvolila domácnosť s najkratšou vzdialenosťou. Imputáciu sme vykonávali na úrovni najmenšej agregácie výdavkov, čo znamená, že sme separátne imputovali 301 rôznych typov výdavkov. Pre každú najbližšiu domácnosť d z RÚ sme pre každý výdavok i vypočítali jeho podiel na celkovom disponibilnom príjme domácnosti v tvare

$$w_{di} = \frac{e_{di}}{y_d}, \quad \forall \quad i = 1{,}2 \ ... \ 301 \quad \forall \quad d = 1{,}2 \ ... \ 4\ 876 \tag{13}$$

kde e_{di} je výška výdavku i domácnosti d a y_d je disponibilný príjem domácnosti d. Podiely (13) sme následne aplikovali na disponibilné príjmy spárovaných domácností z EU - SILC, čím sme vypočítali imputované výdavky pre 301 kategórií výdavkov.

Celkovo sme vyskúšali imputáciu výdavkov pomocou štrnástich rôznych variantov v závislosti od výberu spoločných premenných. Ich zoznam spolu s využitými spoločnými premennými je uvedený v Prílohe A.

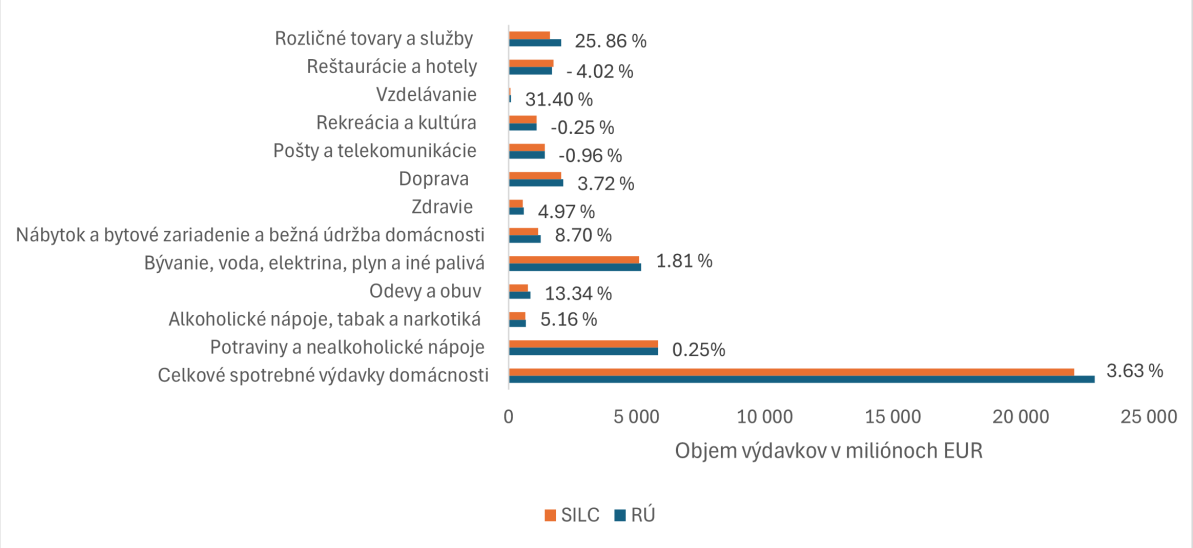
5.2 Validácia imputačnej metódy

Hlavné kritérium na základe ktorého sme z dostupných variantov zvolili jeden najpresnejší, bola úspešnosť prenosu výdavkov do databázy EU - SILC a to vzhľadom na celkový počet výdavkov vypočítaných pomocou váh domácnosti a počet výdavkov v jednotlivých COICOP kategóriách.

Najlepším imputačným variantom bol v tomto smere Variant č. 2. Z tohto dôvodu si v tejto časti zhrnieme výstupy len tohto Variantu. Výstupy zo zvyšných trinástich variantov sú uvedené v Prílohe B.

5.2.1 Validácia voči databáze RÚ

V tejto časti budeme validovať výsledky voči agregátnym štatistikám z databázy RÚ. Pokiaľ sú obe zisťovania sledované pre tú istú populáciu, tak agregované výdavky v databáze EU-SILC by mali vykazovať podobné správanie ako v RÚ. Na Obrázku 1 môžeme pozorovať porovnanie celkového súčtu výdavkov v oboch databázach pre Variant č.2. Porovnanie výdavkov agregovaných do užších 115 skupín môžeme pozorovať v Prílohe C.



Obr. 1: Porovnanie celkového objemu výdavkov v Databáze RÚ a EU - SILC pre COICOP kategórie, vypočítaných využitím váh domácnosti. Percento v grafe označuje percentuálnu odchýlku výdavkov v databáze RÚ voči SILC.

Z Obrázka 1 je vidieť, že imputácia výdavkov spôsobila odchýlku na celkových výdavkoch len 3.63 %. Kategórie Potraviny a nealkoholické nápoje a Bývanie, voda, elektrina

plyn a iné palivá, ktoré tvoria z hľadiska objemu výdavkov domácnosti najväčšie kategórie, mali odchýku na úrovni 0.25 % a 1.81 % čo taktiež indikuje dobrú kvalitu imputačnej metódy. Na druhej strane kategórie Odevy a obuv, Vzdelávanie a Rozličné tovary a služby dosahovali najvyššiu odchýlku. Zároveň sú to však najmenšie kategórie, ktoré sa nepodarilo úspešne preniesť ani pomocou iných imputačných variantov.

Pre lepšie priblíženie správania sa výdavkov sme si zhrnuli vývoj priemerov a mediánov pre obe databázy, taktiež pre 12 COICOP kategórií. Ich porovnanie môžeme sledovať v Tabuľke 19.

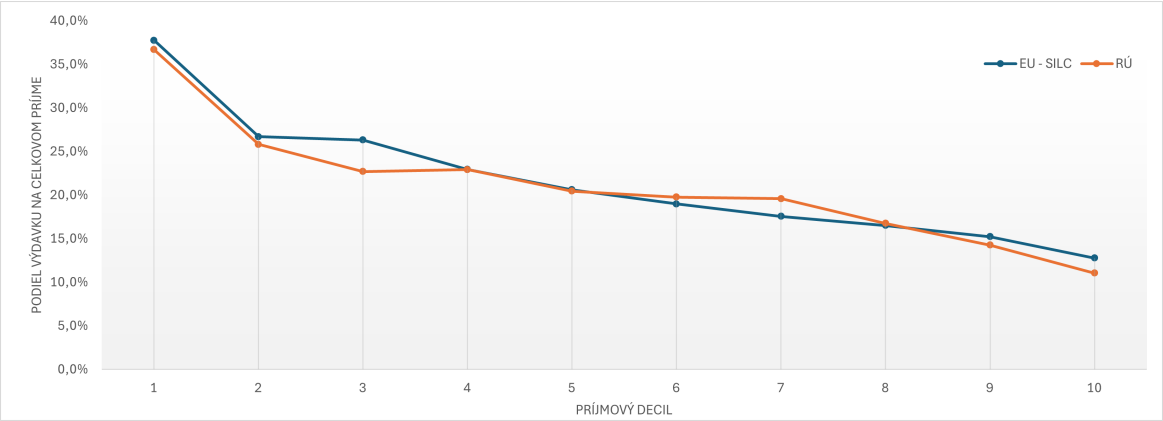
	Priemer		Medián	
Typ výdavku	RÚ	EU - SILC	RÚ	EU - SILC
Celkové spotrebné výdavky domácnosti	13 255.01	12 751.13	11 669.79	10 885.79
Potraviny a nealkoholické nápoje	3 384.96	3 366	3 097.92	2 957.98
Alkoholické nápoje. tabak a narkotiká	394.50	374	201.84	203,9397
Odevy a obuv	491.90	432.67	276.44	209.98
Bývanie. voda. elektrina. plyn a iné palivá	3 004.016	2 941.74	2 750.16	2 557.12
Nábytok a bytové zariadenie a bežná údržba domácnosti	725.11	665.04	336.15	284.70
Zdravie	343.52	326.26	223.15	186.37
Doprava	1 234.48	1 186.55	470	368.63
Pošty a telekomunikácie	815.37	820.76	736.14	707.88
Rekreácia a kultúra	637.56	637.24	281.92	263.35
Vzdelávanie	59.49	45	0	0
Reštaurácie a hotely	978.52	1016.38	495.6	464
Rozličné tovary a služby	1 185.57	937	867.9	526

Tabuľka 19: Porovnanie priemerného výdavku domácnosti a mediánu výdavku domácnosti u 12 výdavkových COICOP kategórií medzi databázou RÚ a EU-SILC (s využitím váh domácností).

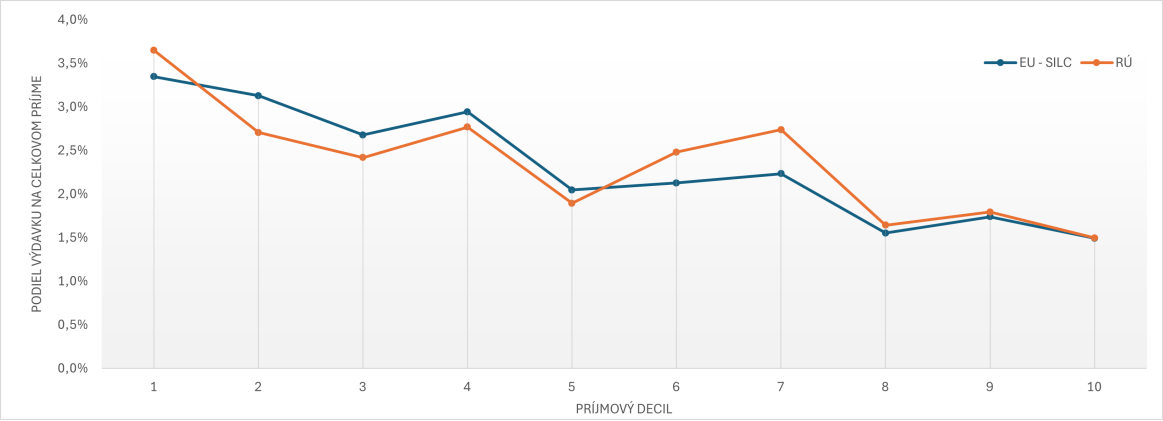
Tabuľka 19 môže slúžiť nie len na číselne porovnanie, ale aj na získanie informácie o rozdelení jednotlivých skupín výdavkov. V prípade že je medián výrazne odlišný od priemeru, naznačuje to určitú mieru šikmosti rozdelenia, čo vidíme najmä pri skupinách Odevy a obuv, Nábytok a bytové zariadenie a bežná údržba domácnosti, Doprava, Rekreácia a kultúra, Vzdelanie a Reštaurácie a hotely. Výrazné rozdiely medzi priemerom a mediánom boli následne zachované aj v databáze EU-SILC.

Ďalej sme sa pozreli, či bola imputačná metóda úspešná aj vzhľadom na jednotlivé príjmové decily. Zisťovali sme, koľko percent výdavkov z celkového príjmu danej kategórie

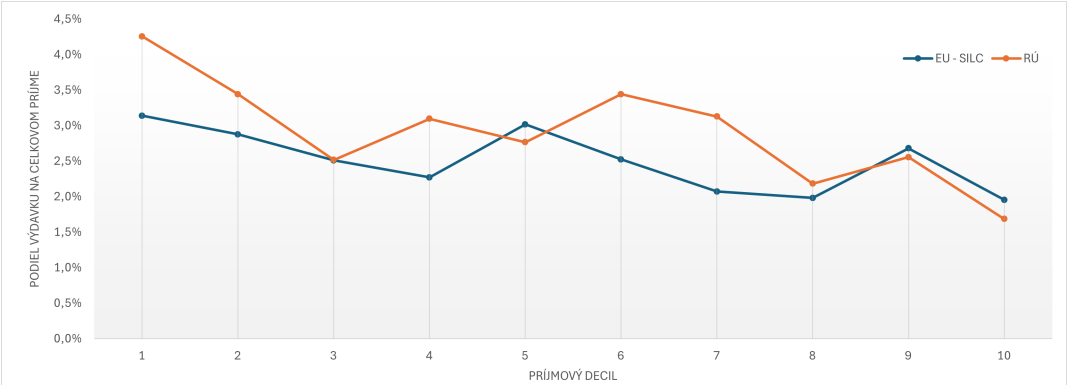
míňajú domácnosti z jednotlivých príjmových decilov v databáze RÚ a či sa toto správanie zachovalo aj po imputácii, v databáze EU - SILC. Na obrázkoch 2 až 13 môžeme pozorovať vývoj podielov výdavkov konkrétnej kategórie na celkovom príjme pre jednotlivé príjmové decily.



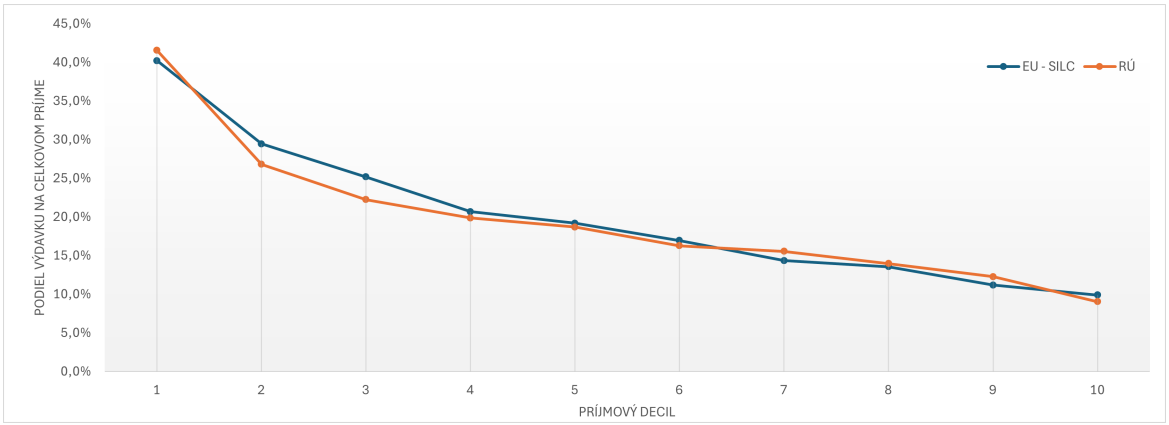
Obr. 2: Vývoj podielov výdavkov kategórie Potraviny a nealkoholické nápoje na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



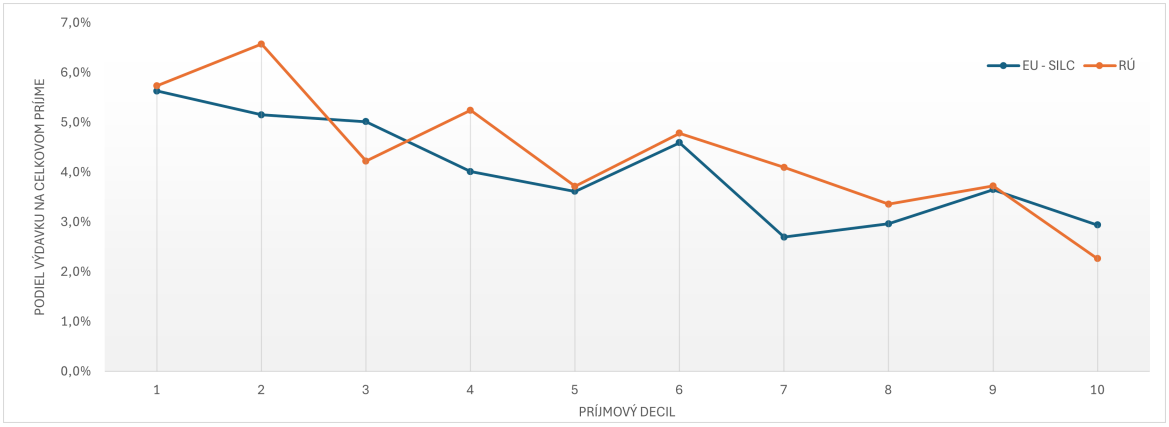
Obr. 3: Vývoj podielov výdavkov kategórie Alkoholické nápoje, tabak a narkotiká na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



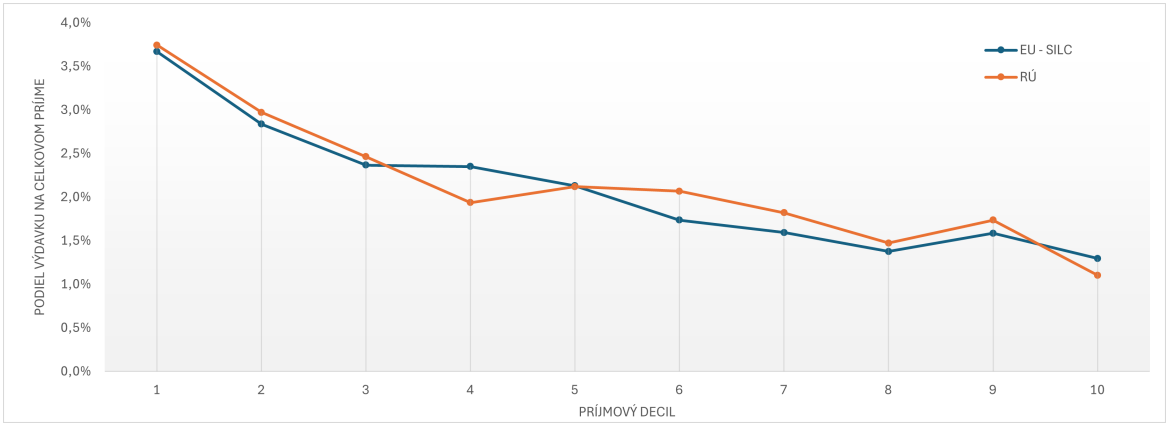
Obr. 4: Vývoj podielov výdavkov kategórie Odevy a obuv na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



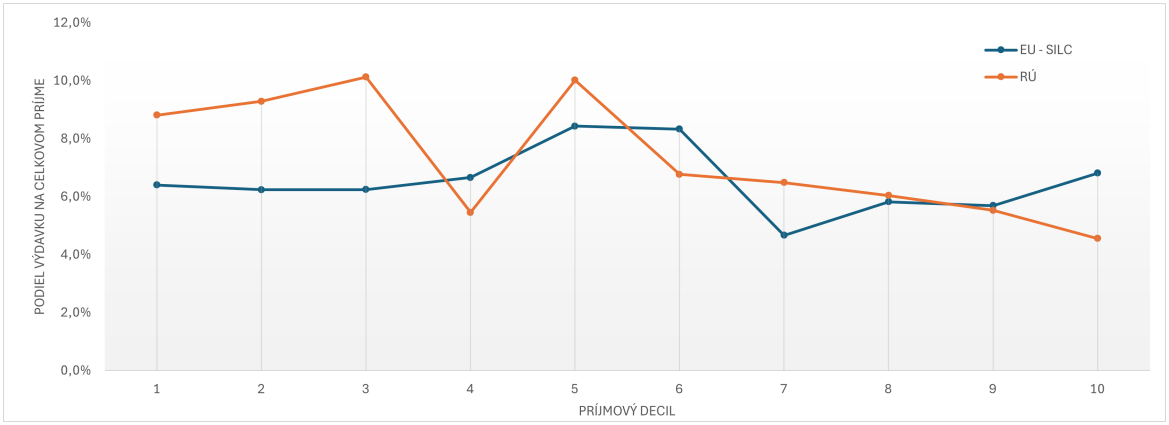
Obr. 5: Vývoj podielov výdavkov kategórie Bývanie, voda, elektrina, plyn a iné palivá na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



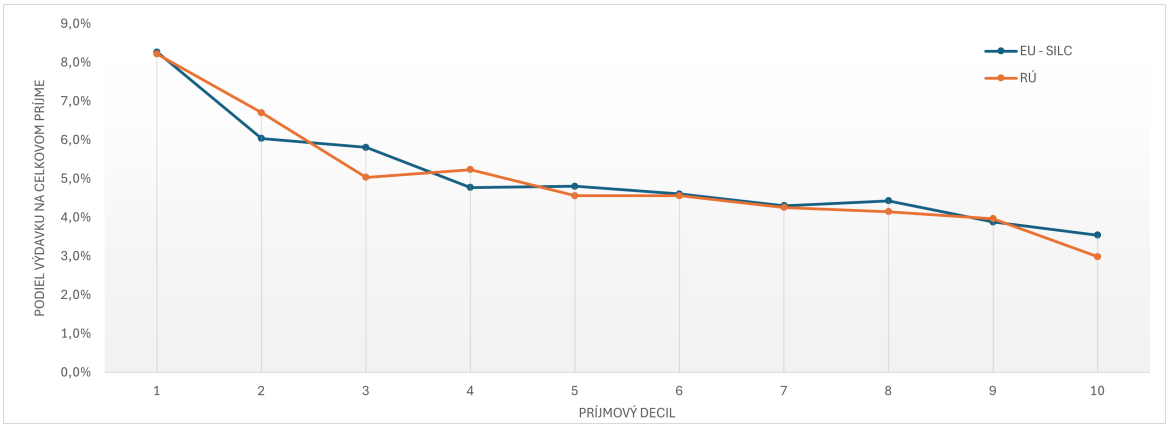
Obr. 6: Vývoj podielov výdavkov kategórie Nábytok a bytové zariadenie a bežná údržba domácnosti na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



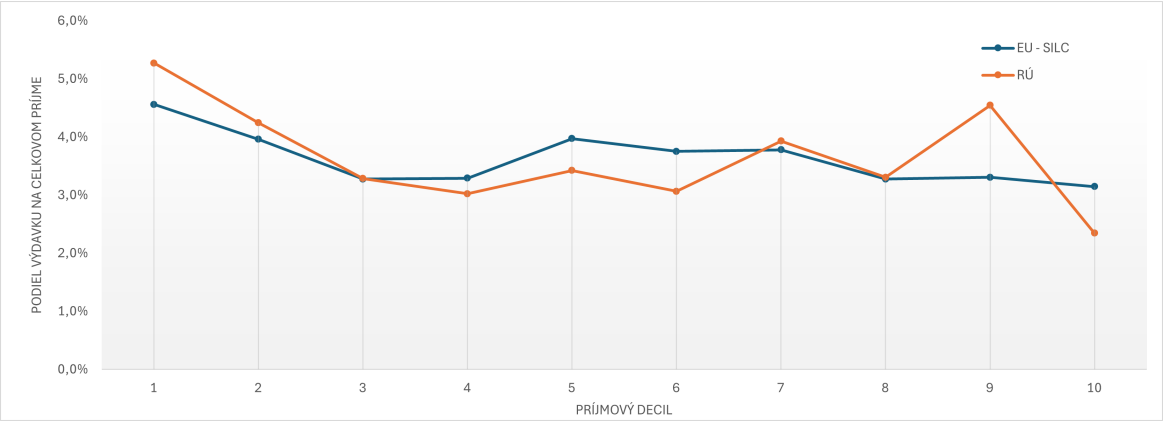
Obr. 7: Vývoj podielov výdavkov kategórie Zdravie na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



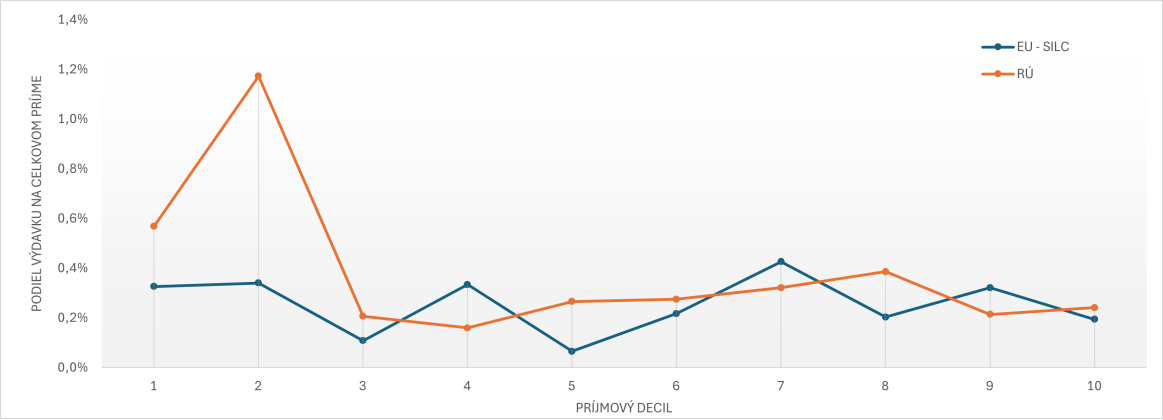
Obr. 8: Vývoj podielov výdavkov kategórie Doprava na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



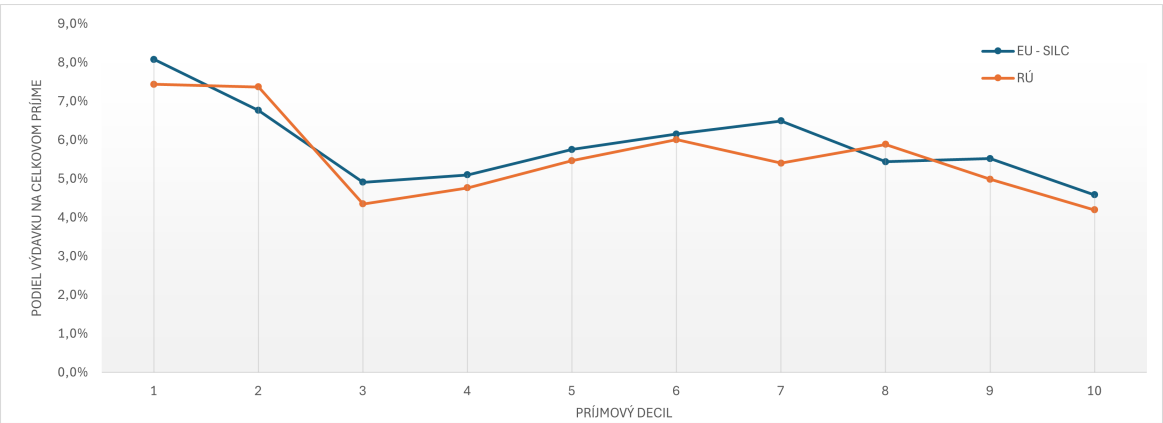
Obr. 9: Vývoj podielov výdavkov kategórie Pošty a telekomunikácie na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



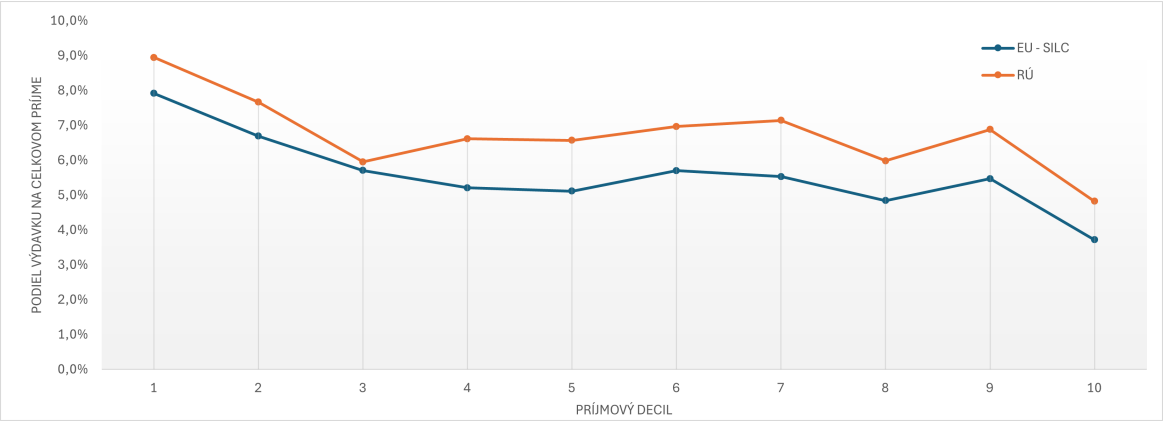
Obr. 10: Vývoj podielov výdavkov kategórie Rekreácia a kultúra na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



Obr. 11: Vývoj podielov výdavkov kategórie Vzdelávanie na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



Obr. 12: Vývoj podielov výdavkov kategórie Reštaurácie a hotely na celkovom príjme pre jednotlivé príjmové decily (s využitím váh)



Obr. 13: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 1. príjmový decil (s využitím váh)

Z obrázkov 2 až 13 je vidieť, že imputačnej metóde sa darilo dobre aj vo vnútri jednotlivých príjmových decilov, obzvlášť pri kategórii Potraviny a nealkoholické nápoje (Obrázok 2) a kategórii Bývanie, voda, elektrina, plyn a iné palivá (Obrázok 5), kde krivka podielov z EU - SILC dostatočne kopíruje krivku z RÚ.

V prípade že chceme pozorovať rozdelenie podielov výdavkov medzi výdavkové kategórie pre konkrétny príjmový decil, môžeme ich vidieť na grafoch v Prílohe D.

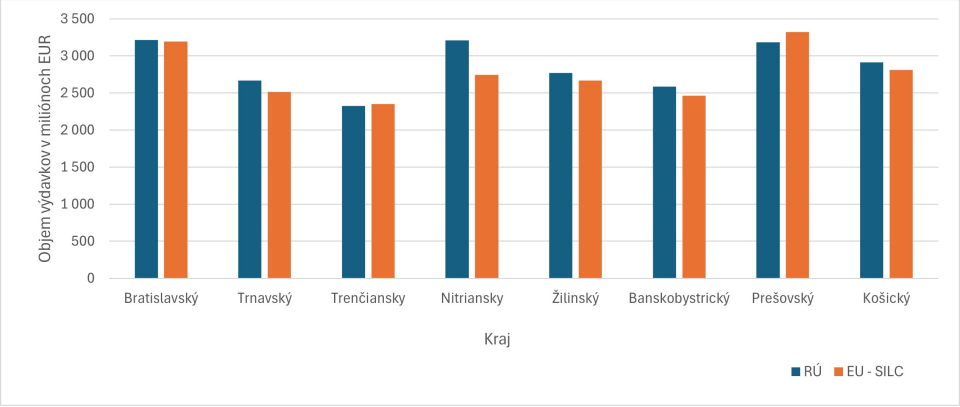
Veľmi častým problémom pri imputovaní výdavkov je imputácia výdavkov, ktoré sú nulové. Imputovať tieto výdavky pomocou parametrických metód ako je lineárna regresia sa podľa [2] javí ako menej spoľahlivé, no HDM metóda je schopná si s týmto problémom poradiť lepšie, pretože v tomto prípade sa imputujú hodnoty, ktoré skutočne existujú v databáze RÚ. V rámci validácie HDM metódy sme skúmali podiel domácností s nulovými výdavkami v databáze EU - SILC v porovnaní s pôvodnou databázou RÚ. Porovnanie podielu domácnosti s nulovými výdavkami pre COICOP kategórie môžeme pozorovať v Tabuľke 20.

Ketgória	RÚ	EU-SILC
Potraviny a nealkoholické nápoje	0 %	0 %
Alkoholické nápoje, tabak a narkotiká	18,21 %	17,15 %
Odevy a obuv	14,8 %	15,42 %
Bývanie, voda, elektrina, plyn a iné palivá	0 %	0 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	3,09 %	3,24 %
Zdravie	9,82 %	9,76 %
Doprava	17,38 %	16,02 %
Pošty a telekomunikácie	0,4 %	0,23 %
Rekreácia a kultúra	5,15 %	4,12 %
Vzdelávanie	83,23 %	85,1 %
Reštaurácie a hotely	30,73 %	30,61 %
Rozličné tovary a služby	1,01 %	0,95 %

Tabuľka 20: Porovnanie % domácnosti s nulovými výdavkami danej kategórie medzi databázou RÚ a EU-SILC (s využitím váh domácností).

Z Tabuľky 20 je možné pozorovať, že podiel domácnosti s nulovými výdavkami v závislosti od 12 COICOP kategórií sa po imputácii podarilo úspešne zachovať. Teória o tom, že HDM metóda prenáša nulové výdavky efektívnejšie sa v tomto prípade potvrdila. Úspešnosť zachovania početnosti domácnosti s nulovými výdavkami je preto najväčšiou výhodou oproti dosposiaľ vyskúšanej imputačnej metóde. Porovnanie podielu domácnosti s nulovými výdavkami pri uvažovaní o 115 výdavkových kategóriách je uvedené v Prílohe E.

Nakoniec sme si ešte vykreslili rozdiel medzi výdavkami vzhľadom na jednotlivé kraje, ktoré môžeme pozorovať na Obrázku 14.

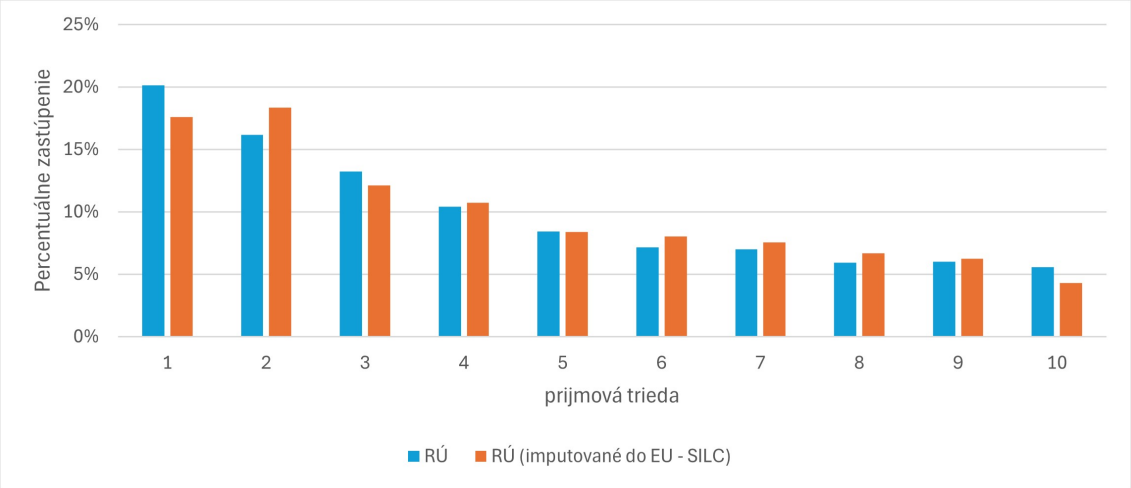


Obr. 14: Porovnanie celkového objemu výdavkov pre jednotlivé kraje (s využitím váh domácností)

Z Obrázka 14 je opäť vidieť podobné správanie ako v databáze RÚ, okrem Nitrianskeho kraja. Tento nedostatok sa však podobne prejavoval aj v iných imputačných variantoch. Vieme, že celkovo sa do EU - SILC prenieslo menej výdavkov ako je v RÚ a z pohľadu krajov sa to najviac prejavilo v Nitrianskom kraji.

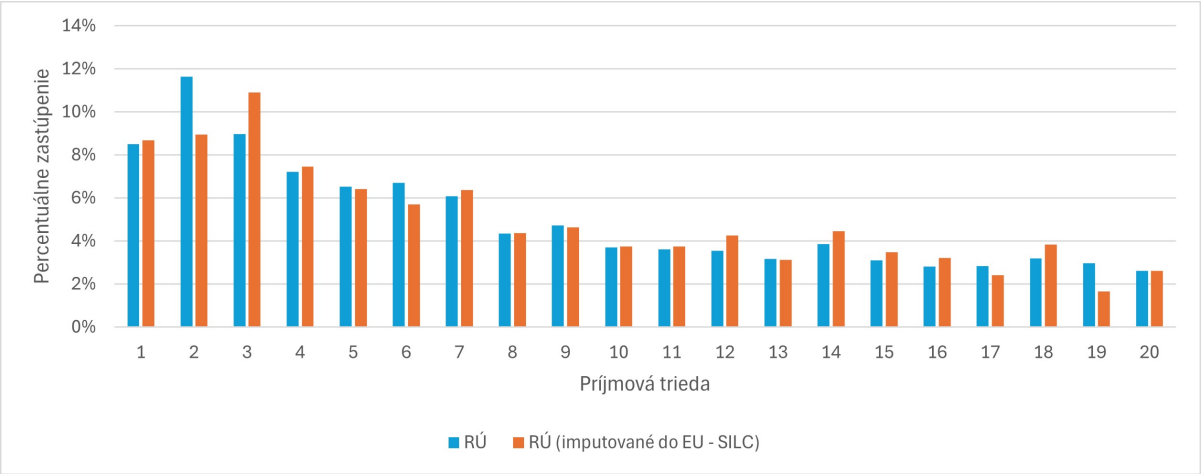
5.2.2 Validácia pomocou porovnania distribúcie

Ďalším prístupom, ktorý nám povie o priebehu imputačnej metódy je porovnávanie distribúcie príjmových decilov v pôvodnej databáze RÚ a v domácnostiach, ktoré boli využité na imputáciu. Porovnanie môžeme vidieť na Obrázku 15.



Obr. 15: Porovnanie distribúcie príjmových decilov v pôvodnej databáze RÚ a imputovaných domácností (bez využitia váh)

Na Obrázku 15 je modrou farbou znázornená distribúcia jednotlivých príjmových decilov v databáze RÚ. Keďže decily boli počítané využitím váh domácnosti, tento graf nám hovorí o tom, že v databáze RÚ sa nachádza viac nízkopríjmových domácností a naopak menej vysokopríjmových domácností. Oranžovou farbou je znázornená distribúcia tých domácností z RÚ, ktoré sa využili pri imputácii, teda ich výdavky sa naimputovali do EU - SILC. V ideálnom prípade by mali mať oba stĺpce pri jednotlivom decile rovnakú výšku, no vzhľadom na to, že v databáze RÚ máme menej pozorovaní ako v EU - SILC, niektoré domácnosti z databázy RÚ museli byť na imputáciu využité viackrát. Analogický graf pre príjmové ventily môžeme pozorovať na Obrázku 16.



Obr. 16: Porovnanie distribúcie príjmových ventilov v pôvodnej databáze RÚ a imputovaných domácností (bez využitia váh)

V procese imputácie dát je snaha o to, aby sa v imputovanom súbore zachovalo rovnaké, alebo aspoň približne podobné rozdelenie ako v pôvodnom súbore [2]. Z tohto dôvodu sme si vypočítali Hellingerovu vzdialenosť medzi združeným rozdelením príjmových decilov a výdavkových decilov medzi súborom RÚ a súborom EU - SILC, a to so zohľadnením váh domácností. Výsledkom bola hodnota Hellingerovej vzdialenosti na úrovni 0.1134, čo indikovalo mierne zhoršenie podobnosti rozdelení. Tabuľky s jednotlivými distribúciami môžeme pozorovať v Prílohe F.

5.2.3 Validácia pomocou analýzy zmeny korelačnej štruktúry

Ďalším spôsobom ako vyhodnotiť kvalitu imputačnej metódy je pozrieť sa, či sa zachovala úroveň korelačného koeficientu medzi príjmami a výdavkami, ktorý bol nameraný v RÚ. Porovnanie môžeme pozorovať v Tabuľke 21.

	Pearsonov korelačný koeficient	Spearmanov korelačný koeficient
RÚ	0.3547	0.4466
EU - SILC	0.4701	0.5437

Tabuľka 21: Porovnanie korelačných koeficientov medzi databázou RÚ a EU - SILC (s využitím váh)

V našom prípade sa oba typy korelačných koeficientov v databáze EU - SILC posunuli smerom nahor o približne 10 percentuálnych bodov oproti databáze RÚ. Hoci je pri

imputovaní výdavkov snaha o zachovanie korelačnej štruktúry, autori v [2] vysvetľujú, že zmena korelácií je pri HDM metóde prirodzeným javom a to práve z toho dôvodu, že niektoré domácnosti z databázy RÚ môžu byť použité viackrát a niektoré ani raz.

Podobným prístupom je zostrojiť korelačnú maticu v súbore RÚ a EU-SILC. Premenné vstupujúce do matice sú Kraj, Stupeň urbanizácie, Rok narodenia referenčnej osoby, Rodinný stav, Ekvivalentná veľkosť domácnosti, a podiely výdavkov na celkovom príjme (pri uvažovaní 12 COICOP kategórií). Následne je možné zostrojiť maticu

$$\Delta C(x, w) = |C_{R\acute{U}}(x, w) - C_{EU-SILC}(x, w)| \tag{14}$$

kde C je korelačná matica, x je vektor spoločných premenných a w je vektor podielov 12 kategórií výdavkov na celkovom príjme. Ak sú oba súbory reprezentatívne pre tú istú populáciu, potom matica $\Delta C(x, w)$ by mala byť blízka nulovej matici [2]. Matica (14) má v našom prípade rozmery 17x17, z ktorej vieme vyčítať 3 rôzne typy korelácií. Ľavá horná podmatica rozmerov 5x5 nám nevraví nič o kvalite imputačnej metódy. Určuje nám len to, ako veľmi podobná je korelačná štruktúra medzi demografickými premennými na základe databázy RÚ a na základe databázy EU-SILC. Pravá horná podmatica s rozmermi 5x12 nám poskytuje údaj o tom, nakoľko bola zachovaná korelačná štruktúra medzi spoločnými premennými a výdavkovými podielmi. Pravá dolná podmatica s rozmermi 12x12 nám určuje, ako sa po imputácii zachovala korelačná štruktúra medzi jednotlivými výdavkovými podielmi [2]. Maticu $\Delta C(x, w)$ môžeme pozorovať v Prílohe G. Ak by sme chceli z toho vyvodiť stručnejší záver, môžeme sa zamerať na priemerné rozdiely v koreláciách medzi súborom RÚ a EU-SILC, rovnako ako autori v [2] a [7] a to pre nasledujúce skupiny premenných:

- spoločné premenné
- výdavkové podiely
- spoločné premenné a výdavkové podiely

Výsledky priemerných rozdielov medzi koreláciami pre tieto 3 skupiny premenných uvádzame v Tabuľke 22. Ako sme už spomenuli vyššie, o kvalite imputačnej metódy nám najviac napovie druhý riadok Tabuľky 22. Priemerný rozdiel v koreláciách na úrovni 0.0282 naznačuje, že korelácie sa po imputácii až tak výrazne nezmenili.

Premenné	Rozdiel
Spoločné premenné	0.0393
Spoločné premenné a výdavkové podiely	0.0282
Výdavkové podiely	0.1289

Tabuľka 22: Priemerný rozdiel korelačých koeficientov jednotlivých premenných medzi databázou RÚ a EU-SILC (s využitím váh domácností).

Záver

Hlavným cieľom tejto diplomovej práce bolo naimputovať údaje o výdavkoch z výberového zisťovania rodinných účtov (RÚ) do výberového zisťovania o príjmoch a životných podmienkach (EU - SILC), ktoré tvorí podkladovú databázu pre už existujúci mikrosimulačný model SIMTASK, vytvorený analytikmi Kancelárie Rady pre rozpočtovú zodpovednosť. Motiváciou pre túto imputáciu bolo modifikovať spomínaný model, a to konkrétne pomocou použitia neparametrickej imputačnej metódy, nakoľko doposiaľ použitá parametrická metóda neumožnila prenos všetkých výdavkových tried z RÚ do EU - SILC. Zaujímalo nás, aký výkonný bol v imputovaní výdavkov neparametrický prístup.

Prvá kapitola sa sústreďovala na prehľad súčasnej literatúry, ktorá sa venuje téme imputácie dát. Objasnili sme, z akých zdrojov sme pri tejto práci primárne vychádzali a vysvetlili sme, aká bola hlavná motivácia imputovania výdavkov do EU-SILC na Slovensku.

V druhej kapitole sme si predstavili obe výberové zisťovania s ktorými sme neskôr pracovali. Zadefinovali sme si pojem výberového zisťovania a detailne popísali priebeh zbierania dát, sledovaných údajov a frekvenciu výberových zisťovaní. Nakoniec sme uviedli aj ukážku dát, aby sme identifikovali v akej štruktúre sú sledované premenné.

V tretej kapitole sme sa venovali teoretickej stránke imputačného procesu. Vysvetlili sme, čo všetko predchádza implementovaniu konkrétnej imputačnej metódy a aké sú predpoklady imputácie výdavkov. Ďalej sme predstavili konkrétne tri imputačné metódy, a to imputáciu pomocou lineárnej regresie, Hot deck matching metódu a Predictive mean matching metódu.

V štvrtej kapitole sme vytvorili ďalšie potrebné premenné, ktorými sú Príjmový decil, Príjmový ventil a Typ domácnosti. Taktiež sme sa zamerali na overovanie predpokladov použitia imputačnej metódy. Overili sme podobnosť marginálnych rozdelení spoločných premenných pomocou Hellingerovej vzdialenosti. Ďalej sme skúmali mieru asociácie medzi úrovňou výdavkov a ostatnými spoločnými premennými a napokon sme zadefinovali spoločné premenné, ktoré sme neskôr využili na imputáciu výdavkov. Uviedli sme základné štatistické charkteristiky spoločných premenných v oboch dátových súboroch, čo nám poskytlo lepší pohľad na to, ako sa spoločné premenné správajú v oboch databázach.

V poslednej kapitole sme sa venovali praktickej časti diplomovej práce. Venovali sme

sa implementácii metódy Hot deck matching v softvéri STATA. Jadrom tejto imputačnej metódy bolo zostrojenie matice vzdialeností medzi domácnosťami zo Zisťovania rodinných účtov a domácnosťami z EU - SILC. Ku každej domácnosti z databázy EU - SILC sa následne priradila práve jedna domácnosť z databázy RÚ, ktorej výdavky sa naimputovali k spárovanej domácnosti z databázy EU-SILC. Neimputovala sa však absolútna úroveň výdavku, ale percentuálny podiel daného výdavku na celkovom disponibilnom príjme domácnosti. Tento podiel sa následne aplikoval na domácnosť z databázy EU-SILC čím sa vypočítala peňažná úroveň výdavku.

Spolu sme si predstavili 14 rôznych variantov HDM metódy v závislosti od voľby rôznych kombinácií spoločných premenných. Z nich sme vybrali jeden na základe toho, ako veľmi sa celkový objem imputovaných výdavkov v databáze RÚ podobal výdavkom z databázy RÚ, a to aj po rozčlenení výdavkov na 12 COICOP kategórií. Nami zvolený variant vykazoval odchýlku 3.63 % v celkových výdavkoch, pričom v kategórii Potraviny a nealkoholické nápoje bola odchýlka na úrovni 0.25 % a v kategórii Bývanie, voda, elektrina, plyn a iné palivá bola odchýlka na úrovni 1.81 %, čo sme vyhodnotili ako najoptimálnejšie rozloženie. Ďalej sme sa pozreli na to, ako sa zmenila korelačná štruktúra medzi príjmami a výdavkami po imputácii. Zmena Pearsonovho aj Spearmanovho korelačného koeficientu bola približne na úrovni 10 %, čo znamenalo miernu zmenu, ktorá bola však očakávaná vzhľadom na priebeh HDM metódy. Rovnako sa mierne zmenilo aj združené rozdelenie príjmových a výdavkových decilov, čo je tiež pri HDM metóde predpokladané [2]. Analyzovali sme taktiež aj to, ako si HDM metóda poradila s nulovými výdavkami, ktorých imputácia je väčšinou problematická, no ukázalo sa, že pri HDM metóde sa relatívne podarilo zachovať ich proporciu.

Hlavným prínosom tejto diplomovej práce je imputácia výdavkov z výberového zisťovania rodinných účtov do zisťovania EU-SILC pomocou neparametrickej Hot deck matching metódy. Táto metóda umožnila prenos výdavkov v detailnejšej štruktúre než existujúce parametrické metódy doposiaľ využívané na Slovensku.

Zoznam použitej literatúry

[1] Akoglu H., *User's guide to correlation coefficients*, review article, Turkish Journal of Emergency Medicine 18 (2018) 91-93, dostupné na internete (30.3.2025): https://www.researchgate.net/publication/326885374_User's_guide_to_correlation_coefficients

[2] Akoğuz, E.C. a kol.: *A new indirect tax tool for EUROMOD Final Report*, Spoločné výskumné centrum Európskej komisie, 2020, dostupné na internete (16.1.2025): https://euromod-web.jrc.ec.europa.eu/sites/default/files/2021-03/A%20new%20indirect%20tax%20tool%20for%20EUROMOD%20Final%20Report.pdf

[3] Conti, P., L., Marella, D., Scaunt, M.: *Uncertainity Analysis in Statistical Matching*, Journal of Official Statistics, 28, (2012), 69-88, dostupné na internete: (16.1.2025): https://www.researchgate.net/publication/244993934_Uncertainty_Analysis_in_Statistical_Matching

[4] Donatiello, G. a kol: *Statistical Matching of Income and Consuption Expenditures*, International Journal of Economic Sciences Vol. 3, No. 3, 2014, dostupné na internete (16.1.2025): https://www.iises.net/download/Soubory/soubory-puvodni/pp50-65-ijoes-V3N3.pdf

[5] D'Orazio, M., Di Zio, M., Scanu, M.: *Old and new approaches in statistical matching when samples are drawn with complex survey designs*, Istat, Rím, dostupné na internete (16.1.2025): https://www.istat.it/it/files/2013/12/Old-and-new-approaches-in-statistical-matching.pdf

[6] D'Orazio, M.: *Statistical Matching: Methodological issues and practice with R-StatMatch*, Italian National Institute of Statistics, 2013, dostupné na internete (16.1.2025): https://www.eustat.eus/productosServicios/datos/Seminario_55_Slides.pdf

[7] Dreoni, I. a kol.: *Statistical imputation and validation of consumption microdata for EUROMOD*, European Commision, Seville, 2025, dostupné na internete (23.4.2025): https://publications.jrc.ec.europa.eu/repository/handle/JRC141513

[8] European Commission, Eurostat: *4.1 Data matching - Final report ESTAT study, Case study on income and wealth (SILC - HFCS)*, Luxembourg, 2013

[9] Eurostat, Glossary, dostupné na internete (13.11.2024): https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Classification_of_individual_consumption_by_purpose_(COICOP)

[10] Janková, K., Pazmán, A.: *Pravdepodobnosť a štatistika*, učebné texty, FMFI UK, Bratislava, 2001, ISBN 978-80-223-2931-6

[11] Klevmarken A.: *Microsimulation. A Tool for Economic Analysis*,1Department of Economics, Uppsala University, Uppsala, Švédsko, 2022, dostupné na internete (12.1.2024): https://microsimulation.pub/articles/00246

[12] Leulescu, L., Agafitei, M.: *Statistical matching: a model based approach for data integration*, Eurostat, 2013, ISBN 978-92-79-30355-5, dostupné na internete (16.1.2024): https://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF.pdf/477dd541-92ee-4259-95d4-1c42fcf2ef34?t=1414780333000

[13] McLachlan, J. G.: *Mahalanobis distance*, Resonance 1999, dostupné na internete (19.12.2024): https://www.ias.ac.in/article/fulltext/reso/004/06/0020-0026

[14] *Mikrosimulačný model SIMTASK, Rada pre rozpočtovú zodpovednosť*, dostupné na internete (15.12.2024): https://simtask.rrz.sk/

[15] Siebertová, Z. a kol: *Expenditure imputation and microsimulation of VAT, Technická*, Discussion Paper No. 1/2018, Rada pre rozpočtovú zodpovednosť, Bratislava, 2018 dostupné na interenete (7.5.2025) : https://www.rrz.sk/en/expenditure-imputation-and-microsimulation-of-vat/

[16] Siebertová, Z.: *Simulačný model SIMTASK - DPH*, technická dokumentácia, RRZ, Bratislava, 2023, dostupné na interete (15.12.2024): https://www.rrz.sk/simtask-dph/

[17] Sheskin D. J.: *Parametric and nonparametric statistical procedures second edition*, CRC, Western Connecticut State University, 2000, dostupné na internete (30.3.2025): https://dl.icdst.org/pdfs/files3/22a131fac452ed75639ed5b0680761ac.pdf

[18] Somorčík, J.: *Ekonometria*, prednášky, FMFI UK, Bratislava, 2023

[19] Somorčík, J.: *Štatistické metódy*, prednášky, FMFI UK, Bratislava, 2022

[20] Somorčík, J.: *Počítačová štatistika*, prednášky, FMFI UK, Bratislava, 2022

[21] Szűcs, G.: *Prednášky z demografickej štatistiky*, učebné texty, FMFI UK, Bratislava 2023

[22] Štatistický úrad Slovenskej republiky: *Príjmy, výdavky a životné podmienky*, dostupné na internete (5.1.2025): https://slovak.statistics.sk/wps/portal/ext/themes/demography/income/metadata/!ut/p/z1/jdCxDoIwEAbgZ_EJetBC27EUKY0NglDALobJNFF0MD6_hLA4iNx2yfdfLj9yqEduHN7-Orz8Yxxu03528aU1JUuSQICiTQC6KGrbsgoTHqNuBhXVC8gtm0ACSnETkIQityUvlcgJNQDMqAi0yO2JVxiDwNvy8GMEbMuvALd-vkNuJmsNfAN2kHICUZNGbYHLPVlAVpeh4ETJ9HTMQDcyZLWJQ4BgAWsl_XvzebfW9uC1F7sP3572sA!!/dz/d5/L2dJQSEvUUt3QS80TmxFL1o2X1E3SThCQjFBMDhCVjIwSTdOUjFLUVFHNUcw/

[23] Zber štatistických údajov: *Zisťovania v domácnostiach, Zisťovanie o príjmoch a životných podmienkach domácnosti*, dostupné na internete (13.11.2024): https://zber.statistics.sk/zis%C5%A5ovanie-o-pr%C3%ADjmoch-a-%C5%BEivotn%C3%BDch-podmienkach-dom%C3%A1cnost%C3%AD-eu-silc

[24] Zber štatistických údajov: *Zisťovania v domácnostiach, Zisťovanie rodinných účtov*, dostupné na internete (13.11.2024): https://zber.statistics.sk/zis%C5%A5ovanie-rodinn%C3%BDch-%C3%BA%C4%8Dtov

[25] *What is EUROMOD ?*, Európska komisia, dostupné na internete (8.5.2025): https://euromod-web.jrc.ec.europa.eu/overview/what-is-euromod

Variant	Kraj	Stupeň urbanizácie	Rodinný stav	Ekonomická aktivita referenčnej osoby	Ekvivalentná veľkosť domácnosti	Typ domácnosti	Rok narodenia ref. osoby	Zlogaritmovaný rok narodenia ref. osoby	Zlogaritmovaný príjem	Príjmový decil	Príjmový ventil
Variant 1	×	×	×	×	×		×		×		
Variant 2	×	×	×	×	×		×			×	
Variant 3	×	×	×	×	×		×				×
Variant 4	×	×	×	×	×			×	×		
Variant 5	×	×	×	×	×			×		×	
Variant 6	×	×	×	×	×			×			×
Variant 7	×	×	×	×		×	×		×		
Variant 8	×	×	×	×		×	×			×	
Variant 9	×	×	×	×		×	×				×
Variant 10	×	×	×	×		×		×	×		
Variant 11	×	×	×	×		×		×		×	
Variant 12	×	×	×	×		×		×			×
Variant 13	×		×	×	×		×			×	
Variant 14	×			×	×		×			×	

Tabuľka 23: Prehľad 14 imputačných variantov s využitými spoločnými premennými. Symbol × označuje, že premenná bola použitá v konkrétnom variante.

B Príloha

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	22 229	2.85 %
Potraviny a nealkoholické nápoje	5 839	5 966	-2.14 %
Alkoholické nápoje. tabak a narkotiká	680	675	0.82 %
Odevy a obuv	848	755	12.33 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	5 096	1.68 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 094	14.34 %
Zdravie	593	584	1.50 %
Doprava	2 129	2 037	4.55 %
Pošty a telekomunikácie	1 406	1 413	-0.44 %
Rekreácia a kultúra	1 100	1 121	-1.87 %
Vzdelávanie	103	71	44.47 %
Reštaurácie a hotely	1 688	1 796	-6.02 %
Rozličné tovary a služby	2 045	1 622	26.08 %

Tabuľka 24: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imputácii) so zohľadnením váh domácností pre Variant č. 1

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 930	4.25 %
Potraviny a nealkoholické nápoje	5 839	5 848	-0.16 %
Alkoholické nápoje. tabak a narkotiká	680	658	3.45 %
Odevy a obuv	848	746	13.68 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	5 046	2.68 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 074	16.47 %
Zdravie	593	564	5.01 %
Doprava	2 129	2 022	5.32 %
Pošty a telekomunikácie	1 406	1 414	-0.53 %
Rekreácia a kultúra	1 100	1 115	-1.37 %
Vzdelávanie	103	73	41.28 %
Reštaurácie a hotely	1 688	1 767	-4.47 %
Rozličné tovary a služby	2 045	1 602	27.49 %

Tabuľka 25: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 3

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	22 246	2.77 %
Potraviny a nealkoholické nápoje	5 839	5 963	-2.08%
Alkoholické nápoje. tabak a narkotiká	680	675	0.82 %
Odevy a obuv	848	756	12.29 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	5 094	1.71 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 097	13.98 %
Zdravie	593	585	1.37 %
Doprava	2 129	2 042	4.26 %
Pošty a telekomunikácie	1 406	1 413	-0.47 %
Rekreácia a kultúra	1 100	1 126	-2.35%
Vzdelávanie	103	72	42.93 %
Reštaurácie a hotely	1 688	1 799	-6.19 %
Rozličné tovary a služby	2 045	1 624	25.89%

Tabuľka 26: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 4

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	22 006	3.89%
Potraviny a nealkoholické nápoje	5 839	5 823	0.26%
Alkoholické nápoje. tabak a narkotiká	680	648	5.03 %
Odevy a obuv	848	746	13.67 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	5 086	1.87 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 131	10.59%
Zdravie	593	564	5.02%
Doprava	2 129	2 026	5.11%
Pošty a telekomunikácie	1 406	1 418	-0.82%
Rekreácia a kultúra	1 100	1 101	–0.09%
Vzdelávanie	103	78	31.49%
Reštaurácie a hotely	1 688	1 762	-4.19 %
Rozličné tovary a služby	2 045	1 623	25.98 %

Tabuľka 27: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 5

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 936	4.23 %
Potraviny a nealkoholické nápoje	5 839	5 849	-0.17 %
Alkoholické nápoje. tabak a narkotiká	680	657	3.49 %
Odevy a obuv	848	745	13.84 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	5 046	2.68 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 073	16.54 %
Zdravie	593	564	5.03 %
Doprava	2 129	2 023	5.23 %
Pošty a telekomunikácie	1 406	1 414	-0.56 %
Rekreácia a kultúra	1 100	1 116	-1.42 %
Vzdelávanie	103	73	41.28 %
Reštaurácie a hotely	1 688	1 771	-4.69 %
Rozličné tovary a služby	2 045	1 604	27.48 %

Tabuľka 28: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 6

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 490	6.39 %
Potraviny a nealkoholické nápoje	5 839	5 806	0.56 %
Alkoholické nápoje. tabak a narkotiká	680	674	0.90 %
Odevy a obuv	848	711	19.35 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	4 987	3.90 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 065	17.49 %
Zdravie	593	586	1.15 %
Doprava	2 129	1 803	18.11 %
Pošty a telekomunikácie	1 406	1 407	-0.03 %
Rekreácia a kultúra	1 100	1 061	3.66 %
Vzdelávanie	103	71	45.28 %
Reštaurácie a hotely	1 688	1 770	-4.66 %
Rozličné tovary a služby	2 045	1 550	31.95%

Tabuľka 29: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 7

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 471	6.48 %
Potraviny a nealkoholické nápoje	5 839	5 654	3.27 %
Alkoholické nápoje. tabak a narkotiká	680	639	6.54 %
Odevy a obuv	848	709	19.74 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	4 982	4.01 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 098	13.86 %
Zdravie	593	567	4.57 %
Doprava	2 129	1 935	10.05 %
Pošty a telekomunikácie	1 406	1 390	1.16 %
Rekreácia a kultúra	1 100	1 083	1.57 %
Vzdelávanie	103	68	50.04 %
Reštaurácie a hotely	1 688	1 704	-0.95 %
Rozličné tovary a služby	2 045	1 642	24.49 %

Tabuľka 30: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 8

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 409	6.79 %
Potraviny a nealkoholické nápoje	5 839	5 635	3.62 %
Alkoholické nápoje. tabak a narkotiká	680	642	6.01%
Odevy a obuv	848	699	21.41 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	4 952	4.63 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 070	16.88 %
Zdravie	593	569	4.21 %
Doprava	2 129	1 919	10.95 %
Pošty a telekomunikácie	1 406	1 396	0.73 %
Rekreácia a kultúra	1 100	1 113	-1.21 %
Vzdelávanie	103	69	47.80 %
Reštaurácie a hotely	1 688	1 742	-3.11 %
Rozličné tovary a služby	2 045	1 603	27.54 %

Tabuľka 31: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 9

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 496	6.36 %
Potraviny a nealkoholické nápoje	5 839	5 810	0.50 %
Alkoholické nápoje. tabak a narkotiká	680	675	0.85 %
Odevy a obuv	848	709	19.61 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	4 986	3.92 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 066	17.31%
Zdravie	593	585	1.22 %
Doprava	2 129	1 803	18.09 %
Pošty a telekomunikácie	1 406	1 406	0.01 %
Rekreácia a kultúra	1 100	1 062	3.53 %
Vzdelávanie	103	71	45.23 %
Reštaurácie a hotely	1 688	1 771	-4.69 %
Rozličné tovary a služby	2 045	1 551	31.82 %

Tabuľka 32: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 10

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 472	6.48 %
Potraviny a nealkoholické nápoje	5 839	5 664	3.08 %
Alkoholické nápoje. tabak a narkotiká	680	642	5.96%
Odevy a obuv	848	709	19.64 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	4 973	4.19 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 099	13.84 %
Zdravie	593	570	3.89 %
Doprava	2 129	1 928	10.44 %
Pošty a telekomunikácie	1 406	1 388	1.33 %
Rekreácia a kultúra	1 100	1 086	1.22 %
Vzdelávanie	103	68	50.28 %
Reštaurácie a hotely	1 688	1 710	-1.31 %
Rozličné tovary a služby	2 045	1 634	25.15 %

Tabuľka 33: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 11

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 399	6.84 %
Potraviny a nealkoholické nápoje	5 839	5 632	3.66 %
Alkoholické nápoje. tabak a narkotiká	680	642	6.02 %
Odevy a obuv	848	698	21.61 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	4 951	4.66 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 064	17.52 %
Zdravie	593	568	4.31 %
Doprava	2 129	1 919	10.96 %
Pošty a telekomunikácie	1 406	1 396	0.71 %
Rekreácia a kultúra	1 100	1 113	-1.21 %
Vzdelávanie	103	69	47.86 %
Reštaurácie a hotely	1 688	1 745	-3.29 %
Rozličné tovary a služby	2 045	1 601	27.71 %

Tabuľka 34: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 12

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	21 945	4.18%
Potraviny a nealkoholické nápoje	5 839	5 803	0.62%
Alkoholické nápoje. tabak a narkotiká	680	644	5.72%
Odevy a obuv	848	785	8.12%
Bývanie. voda. elektrina. plyn a iné palivá	5 181	5 167	0.27%
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 039	20.33%
Zdravie	593	595	-0.41%
Doprava	2 129	1 802	18.13%
Pošty a telekomunikácie	1 406	1 450	-3%
Rekreácia a kultúra	1 100	1 119	-1.69 %
Vzdelávanie	103	80	27.65%
Reštaurácie a hotely	1 688	1 768	-4.53%
Rozličné tovary a služby	2 045	1 693	20.77%

Tabuľka 35: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) so zohľadnením váh domácností pre Variant č. 13

Premenná	RÚ	EU - SILC	Odchýlka
Celkové spotrebné výdavky domácnosti	22 863	22 006	4.18 %
Potraviny a nealkoholické nápoje	5 839	5 772	1.16 %
Alkoholické nápoje. tabak a narkotiká	680	684	-0.57 %
Odevy a obuv	848	810	4.79 %
Bývanie. voda. elektrina. plyn a iné palivá	5 181	5 144	0.72 %
Nábytok a bytové zariadenie a bežná údržba domácnosti	1 251	1 062	17.73 %
Zdravie	593	641	-7.59 %
Doprava	2 129	1 839	15.80 %
Pošty a telekomunikácie	1 406	1 425	-1.28 %
Rekreácia a kultúra	1 100	1 088	1.03 %
Vzdelávanie	103	95	8.11 %
Reštaurácie a hotely	1 688	1 734	-2.69 %
Rozličné tovary a služby	2 045	1 711	19.51 %

Tabuľka 36: Porovnanie objemu agregovaných výdavkov v miliónoch EUR v RÚ a EU - SILC (po imptutácii) s využitím váh domácností pre Variant č. 14

C Príloha

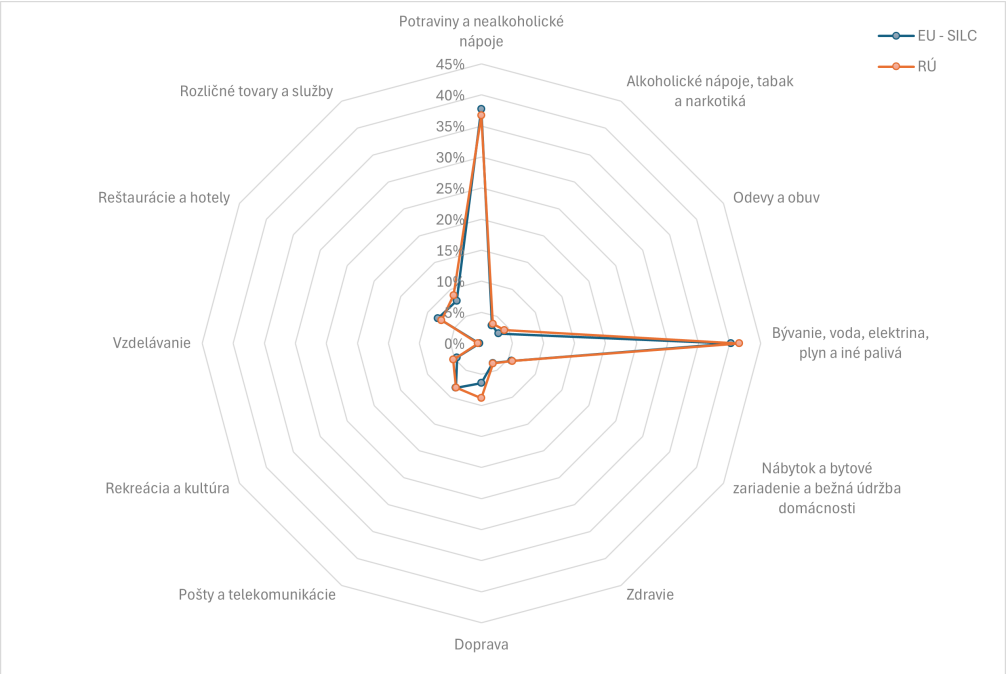
Premenná	RÚ	EU - SILC	Odchýlka
Chlieb a obilniny	1012	1019	-0.75 %
Mäso	1305	1295	0.75 %
Ryby a morské plody	163	157	3.94 %
Mlieko, syry a vajcia	939	937	0.22 %
Oleje a tuky	257	256	0.32 %
Ovocie	424	408	4.07 %
Zelenina	560	554	1.07 %
Cukor, džem, med, čokoláda a cukrovinky	379	394	-3.74 %
Potravinárske výrobky in	312	310	0.39 %
Káva. čaj a kakao	209	203	3.06 %
Minerálne vody, nealkoholické nápoje, ovocné a zeleninové šťavy	278	290	-4.1 %
Liehoviny	146	136	7.28 %
Víno	89	88	0.27 %
Pivo	149	134	11.53 %
Tabak	296	289	2.71 %
Drogy	0	0	0 %
Odevné materiály	1	1	-1.68 %
Hotové odevy	629	545	15.5 %
Ostatné časti odevov a doplnky k odevom	16	15	4.18 %
Čistenie. oprava a požičiavanie odevov	4	5	-27.16 %
Topánky a iná obuv	198	182	8.64 %
Oprava a požičiavanie obuvi	1	1	46.03 %
Skutočné nájomné platené nájomníkmi	270	240	12.24 %
Ostatné nájomné	7	4	81.27 %
Imputované nájomné vlastníkov - obyvateľov	0	0	0 %
Ostatné imputované nájomné	0	0	0 %
Materiály na údržbu a opravu bytu	377	349	8.08 %
Služby na údržbu a opravu bytu	208	244	-14.85 %
Dodávka vody	236	249	-5.12 %
Zber smetí	332	405	-17.95 %
Odvádzanie odpadových vôd	244	257	-5.13 %
Ostatné služby týkajúce sa obydlia in	531	412	28.97 %
Elektrina	1165	1144	1.76 %
Plyn	894	957	-6.57 %
Kvapalné palivá	8	7	22.2 %
Pevné palivá	209	224	-6.63 %
Tepelná energia	701	598	17.19 %
Nábytok a bytové zariadenie	333	298	11.6 %
Koberce a iné podlahové krytiny	18	17	9 %
Oprava nábytku a bytového zariadenia a podlahových krytín	2	1	70.5 %
Bytové textílie	29	31	-5.65 %
Veľké domáce spotrebiče elektrické aj neelektrické	273	285	-4.22 %
Malé domáce elektrické spotrebiče	44	44	0.6 %

Premenná	**RÚ**	**EU - SILC**	**Odchýlka**
Oprava domácich spotrebičov	2	1	64.68 %
Výrobky zo skla, stolný riad a potreby pre domácnosť	17	19	-10.81 %
Veľké nástroje a náradie	39	42	-5.56 %
Malé náčinie a rôzne príslušenstvo	138	119	16.64 %
Tovary pre domácnosť na krátkodobú spotrebu	345	286	20.38 %
Bytové služby a služby pre domácnosť	9	7	28.04 %
Farmaceutické výrobky	343	318	7.99 %
Ostatné zdravotnícke výrobky	14	13	7.16 %
Terapeutické pomôcky a zariadenia	50	49	1.88 %
Lekárske služby	16	9	75.22 %
Služby zubných lekárov	132	151	-12.41 %
Služby stredného zdravotníckeho personálu	32	21	49.71 %
Nemocničné služby	4	3	69.43 %
Motorové vozidlá	580	647	-10.27 %
Motocykle	27	32	-16.08 %
Bicykle	36	39	-5.52 %
Dopravné prostriedky ťahané zvieratami	0	0	0 %
Náhradné diely a príslušenstvo pre prostriedky osobnej dopravy	41	39	5.14 %
Palivá a mazadlá pre prostriedky osobnej dopravy	1088	953	14.12 %
Údržba a oprava prostriedkov osobnej dopravy	185	182	1.7 %
Ostatné služby týkajúce sa prostriedkov osobnej dopravy	65	61	6.46 %
Železničná osobná preprava	12	13	-6.46 %
Cestná osobná doprava	62	57	7.3 %
Osobná letecká doprava	9	12	-30.98 %
Námorná a vnútrozemská vodná osobná doprava	0	0	443.67 %
Kombinovaná osobná doprava	23	16	42.95 %
Ostatné zakúpené dopravné služby	1	1	-9.37 %
Poštové služby	6	5	30.21 %
Telefónne a faxové zariadenia	333	467	-28.66 %
Telefónne a faxové služby	1067	948	12.53 %
Zariadenia na príjem. záznam a reprodukciu zvuku a obrazu	66	54	21.46 %
Fotografické a filmové zariadenia a optické prístroje	4	3	17.92 %
Zariadenia na spracovanie informácií	72	79	-9.14 %
Nahrávacie médiá	1	1	41.31 %
Oprava audiovizuálnej a fotografickej techniky a zariadení na spracovanie informácií	1	1	8.4 %
Významnejšie tovary dlhodobej spotreby na rekreáciu vonku	1	1	44.82 %
Hudobné nástroje a veľké zariadenia dlhodobej spotreby na rekreáciu	4	4	24.34 %
Údržba a oprava ostatných významnejších tovarov dlhodobej spotreby na rekreáciu a kultúru	0	0	34.89 %
Hry, hračky a záľuby	47	31	50 %
Pomôcky na šport. kempovanie a rekreáciu v prírode	31	35	-10.1 %
Záhrady. rastliny a kvety	26	31	-14.47 %
Domáce zvieratá a s nimi súvisiace výrobky	88	83	6.87 %

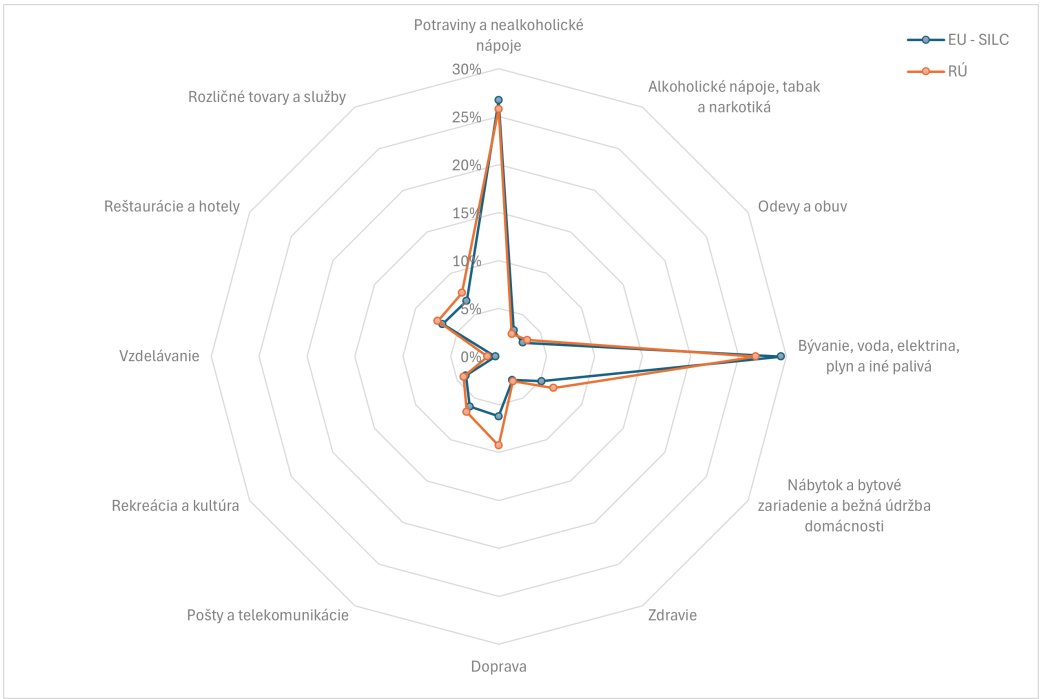
Premenná	**RÚ**	**EU - SILC**	**Odchýlka**
Veterinárne a ostatné služby pre domáce zvieratá	22	25	-13.3 %
Rekreačné a športové služby	54	69	-22.01 %
Kultúra	186	203	-8.61 %
Hazardné hry	2	1	51.93 %
Knihy	30	29	3.05 %
Noviny a periodiká	44	39	11.04 %
Rôzne tlačoviny	6	5	3.58 %
Kancelárske potreby a potreby na kreslenie	67	61	10.09 %
Organizované dovolenkové zájazdy	348	347	0.25 %
Predprimárne a primárne vzdelávanie	29	18	56.94 %
Sekundárne vzdelávanie	6	5	25.75 %
Post-sekundárne vzdelávanie (nezahrňované do terciárneho)	0	0	17.79 %
Vysokoškolské vzdelávanie	34	20	70.26 %
Vzdelávanie nedefinovateľné podľa úrovňou	33	35	-3.62 %
Reštaurácie. kaviarne a podobné zariadenia	1106	1153	-4.06 %
Závodné jedálne	378	414	-8.63 %
Ubytovacie služby	204	192	6.21 %
Kadernícke služby a skrášľovacie služby	142	138	3.41 %
Elektrické spotrebiče na osobnú starostlivosť	5	7	-30.94 %
Ostatné pomôcky a výrobky pre osobnú starostlivosť	409	331	23.52 %
Prostitúcia	0	0	0 %
Šperky. hodiny a náramkové hodinky	25	35	-29.54 %
Ostatné osobné predmety	69	49	42.05 %
Sociálna ochrana	2	4	-38 %
Poistenie v súvislosti s bývaním	79	88	-9.31 %
Zdravotné poistenie	1007	694	45.03 %
Poistenie v súvislosti s dopravou	237	226	4.85 %
Ostatné poistenie	9	6	61.24 %
Ostatné finančné služby i n	7	6	14.06 %
Ostatné služby i n	53	42	26.78 %

Tabuľka 37: Prehľad odchýliek medzi databázou RÚ a databázou EU - SILC pri agregovaní do 115 výdavkových skupín (s využitím váh domácností). Sumy výdavkov sú v miliónoch EUR

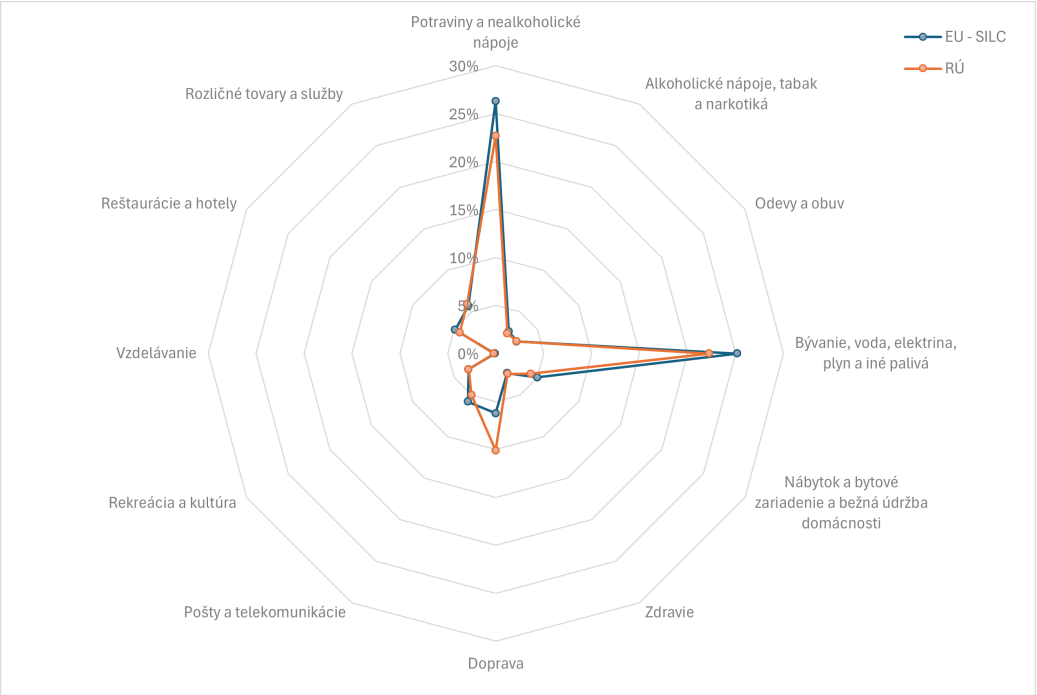
D Príloha



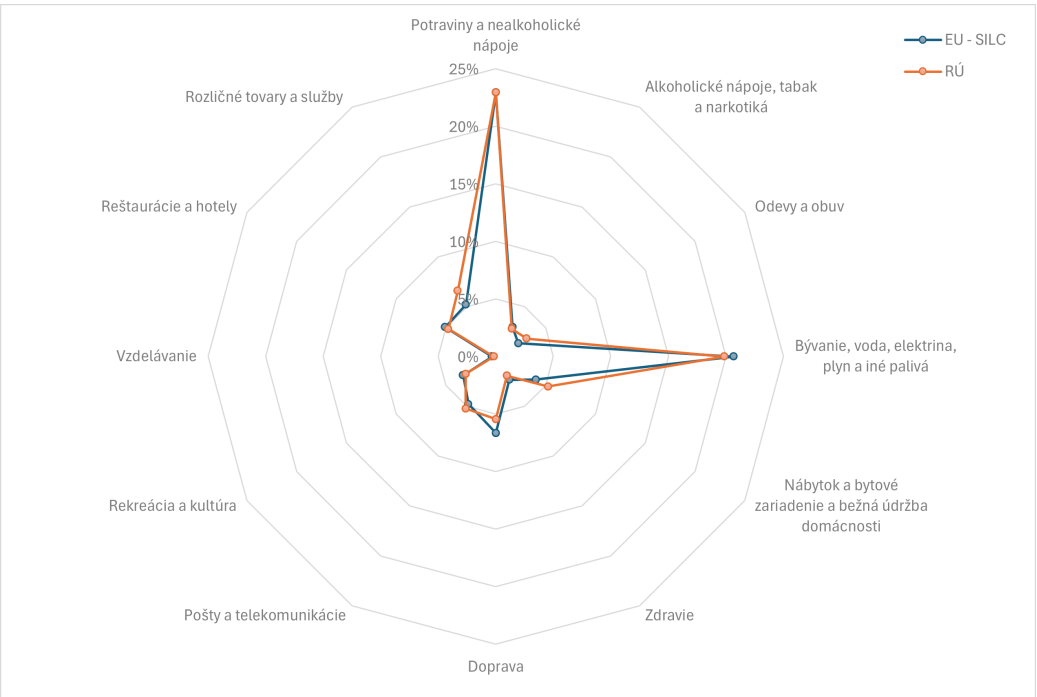
Obr. 17: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 1. príjmový decil (s využitím váh)



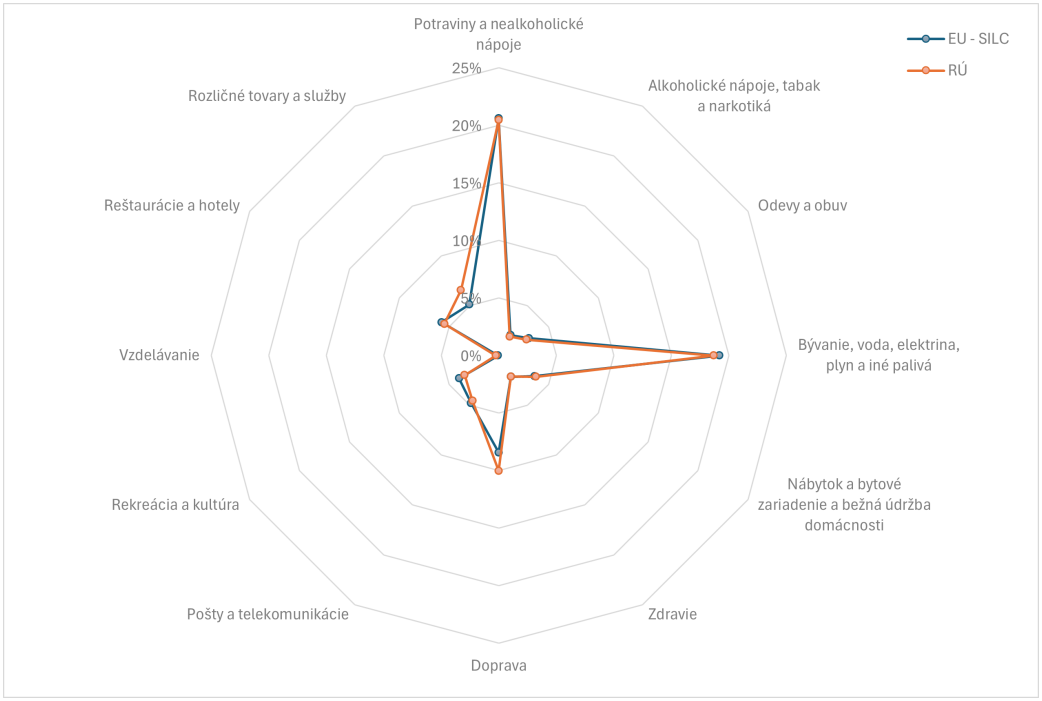
Obr. 18: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 2. príjmový decil (s využitím váh)



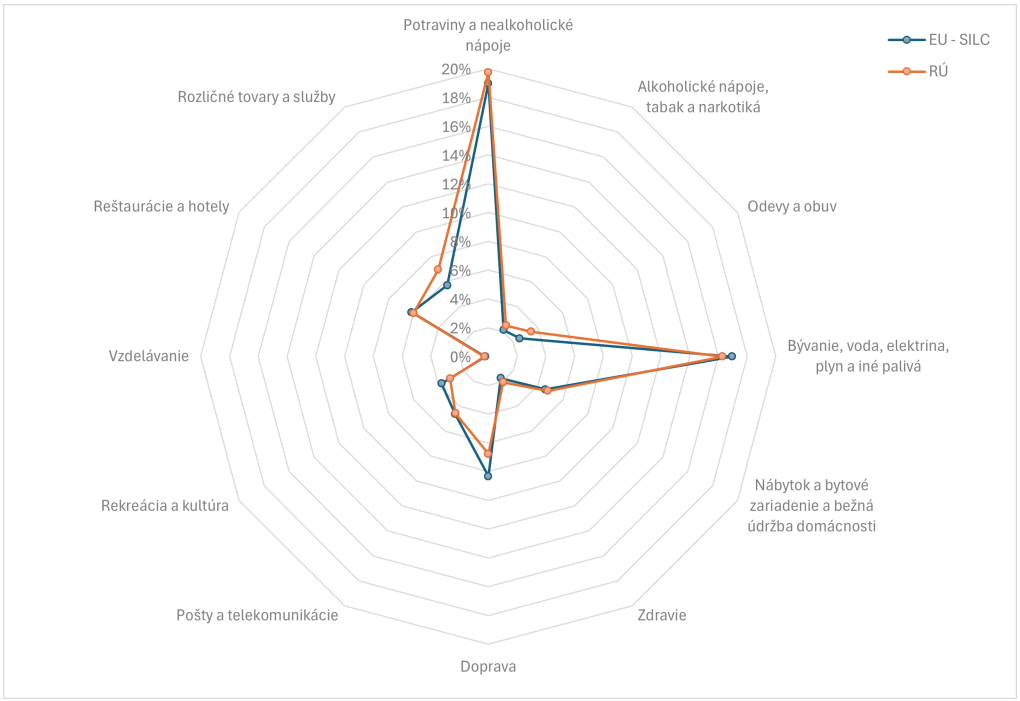
Obr. 19: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 3. príjmový decil (s využitím váh)



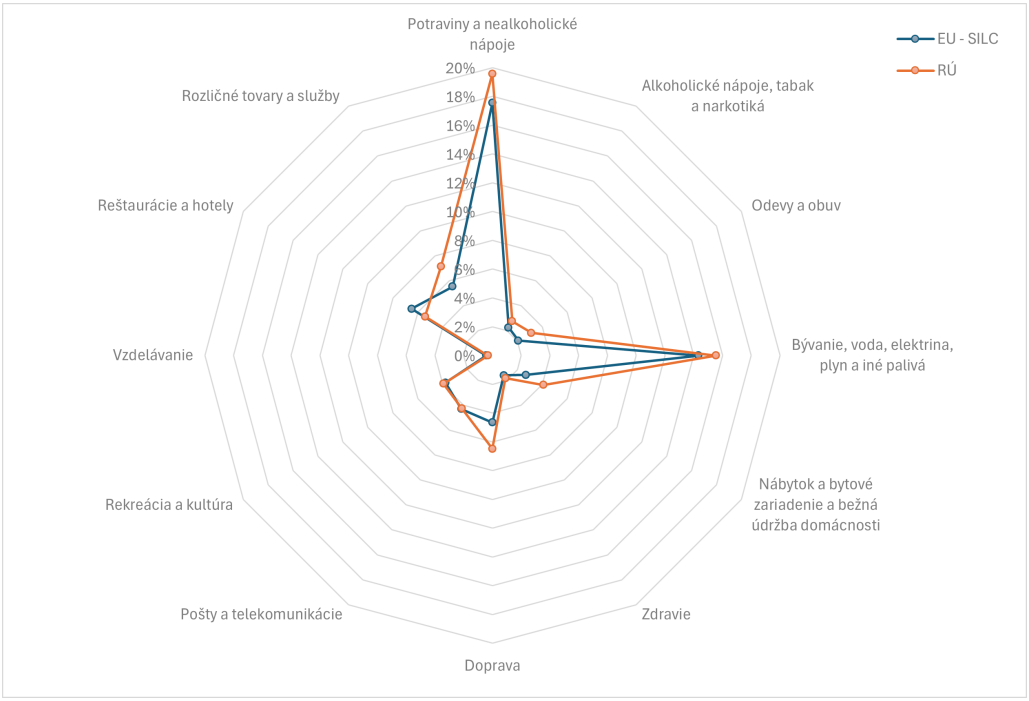
Obr. 20: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 4. príjmový decil (s využitím váh)



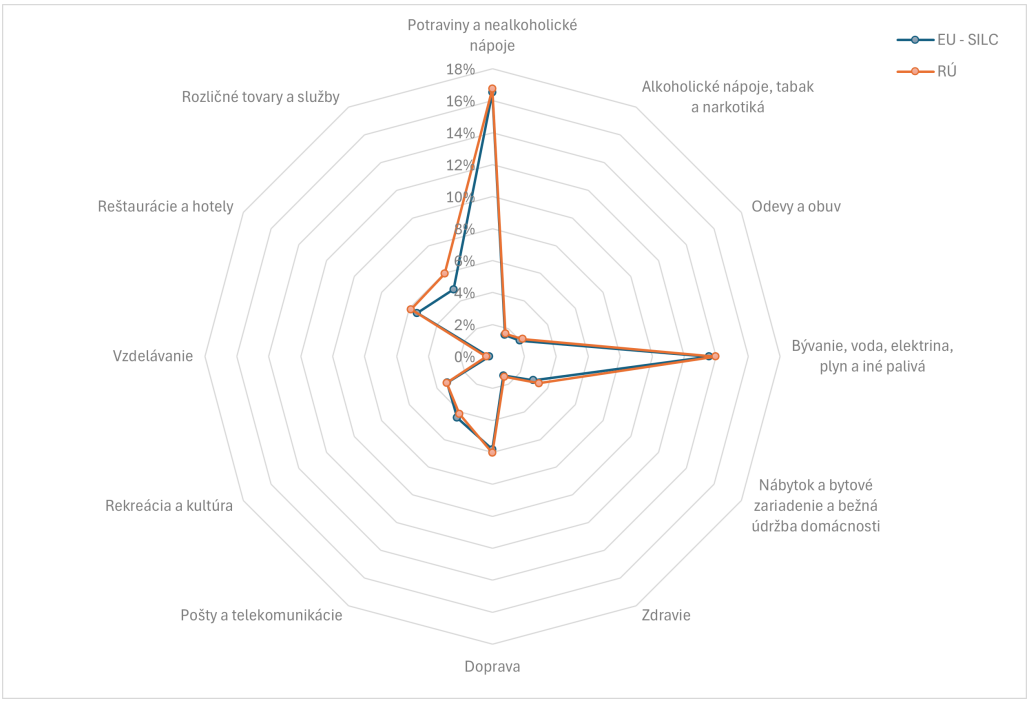
Obr. 21: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 5. príjmový decil (s využitím váh)



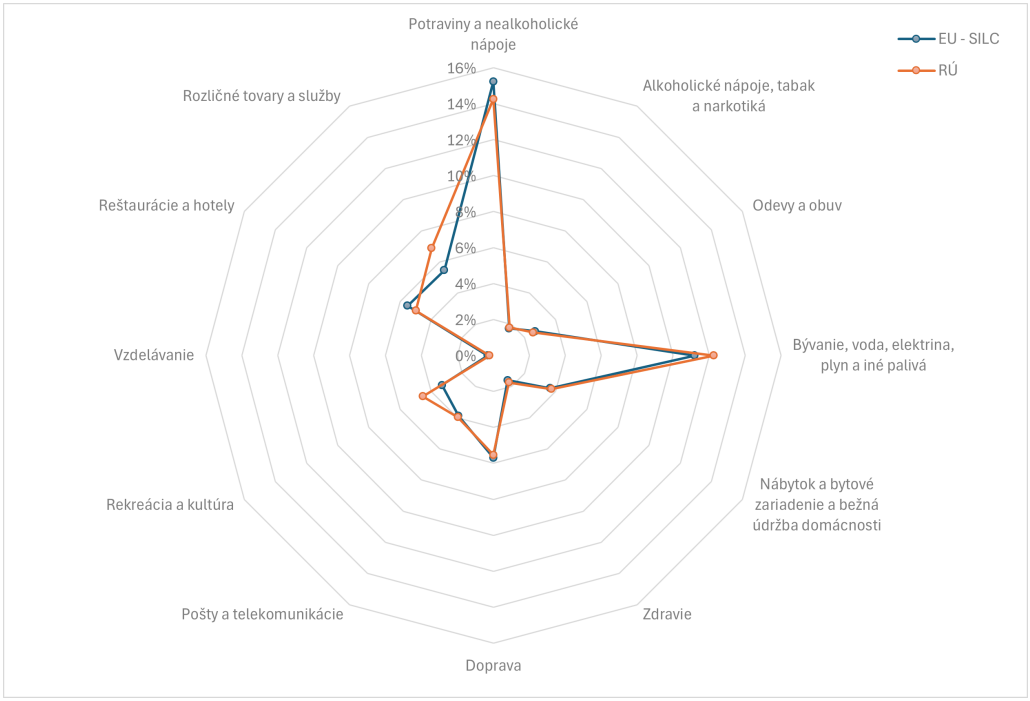
Obr. 22: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 6. príjmový decil (s využitím váh)



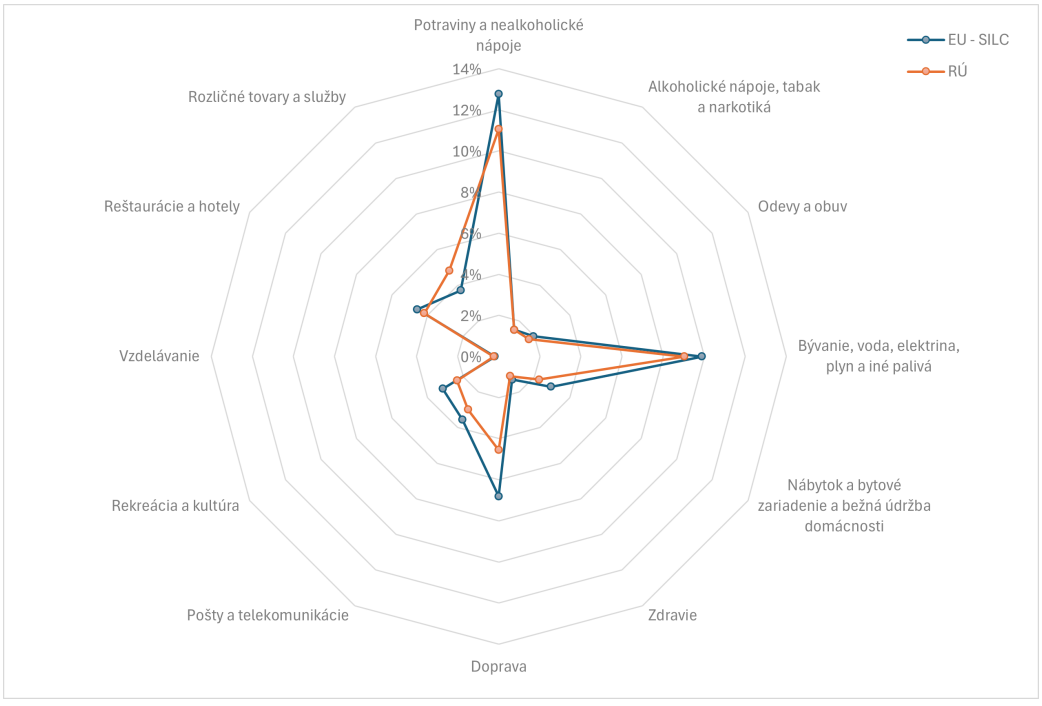
Obr. 23: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 7. príjmový decil (s využitím váh)



Obr. 24: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 8. príjmový decil (s využitím váh)



Obr. 25: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 9. príjmový decil (s využitím váh)



Obr. 26: Vývoj podielov COICOP výdavkových kategórii na celkovom príjme pre 10. príjmový decil (s využitím váh)

E Príloha

Premenná	**RÚ**	**EU - SILC**
Chlieb a obilniny	0.01 %	0.1 %
Mäso	0.62 %	0.45 %
Ryby a morské plody	22.63 %	22.95 %
Mlieko, syry a vajcia	0.18 %	0.21 %
Oleje a tuky	6.91 %	6.81 %
Ovocie	2.12 %	3.05 %
Zelenina	0.62 %	1.14 %
Cukor, džem, med, čokoláda a cukrovinky	4.44 %	3.57 %
Potravinárske výrobky in	4.27 %	4.23 %
Káva. čaj a kakao	14.66 %	14 %
Minerálne vody, nealkoholické nápoje, ovocné a zeleninové šťavy	8.4 %	7.47 %
Liehoviny	62.57 %	61.97 %
Víno	62.26 %	61.61 %
Pivo	38.36 %	39.56 %
Tabak	73.46 %	72.11 %
Drogy	100 %	100 %
Odevné materiály	99.59 %	99.42 %
Hotové odevy	18.42 %	19.09 %
Ostatné časti odevov a doplnky k odevom	87.4 %	87.12 %
Čistenie. oprava a požičiavanie odevov	98.01 %	98.15 %
Topánky a iná obuv	41.95 %	42 %
Oprava a požičiavanie obuvi	98.57 %	98.89 %
Skutočné nájomné platené nájomníkmi	94.37 %	95.29 %
Ostatné nájomné	99.36 %	99.46 %
Imputované nájomné vlastníkov - obyvateľov	100 %	100 %
Ostatné imputované nájomné	100 %	100 %
Materiály na údržbu a opravu bytu	58.8 %	55.3 %
Služby na údržbu a opravu bytu	69.66 %	64.48 %
Dodávka vody	5.25 %	6.03 %
Zber smetí	1.78 %	2.13 %
Odvádzanie odpadových vôd	5.12 %	5.76 %
Ostatné služby týkajúce sa obydlia in	48.34 %	54.18 %
Elektrina	1.31 %	1.61 %
Plyn	19.28 %	19.31 %
Kvapalné palivá	98.7 %	98.87 %
Pevné palivá	78.31 %	76.09 %
Tepelná energia	50.8 %	55.36 %
Nábytok a bytové zariadenie	72.27 %	70.49 %
Koberce a iné podlahové krytiny	91.21 %	91.09 %
Oprava nábytku a bytového zariadenia a podlahových krytín	99.46 %	99.17 %
Bytové textílie	59.09 %	59.19 %
Veľké domáce spotrebiče elektrické aj neelektrické	75.07 %	74.1 %
Malé domáce elektrické spotrebiče	79.4 %	79.22 %

Premenná	**RÚ**	**EU - SILC**
Oprava domácich spotrebičov	98.56 %	98.52 %
Výrobky zo skla, stolný riad a potreby pre domácnosť	66.55 %	64.86 %
Veľké nástroje a náradie	91.5 %	92.29 %
Malé náčinie a rôzne príslušenstvo	67.12 %	65.58 %
Tovary pre domácnosť na krátkodobú spotrebu	9.59 %	10.41 %
Bytové služby a služby pre domácnosť	96.8 %	96.19 %
Farmaceutické výrobky	14.73 %	15.04 %
Ostatné zdravotnícke výrobky	85.59 %	86.04 %
Terapeutické pomôcky a zariadenia	87.22 %	85.9 %
Lekárske služby	94.25 %	95.65 %
Služby zubných lekárov	66.56 %	64.94 %
Služby stredného zdravotníckeho personálu	92.58 %	93.23 %
Nemocničné služby	98.36 %	98.19 %
Motorové vozidlá	96.77 %	96.45 %
Motocykle	99.18 %	99.29 %
Bicykle	96.35 %	96.52 %
Dopravné prostriedky ťahané zvieratami	100 %	100 %
Náhradné diely a príslušenstvo pre prostriedky osobnej dopravy	87.03 %	87.55 %
Palivá a mazadlá pre prostriedky osobnej dopravy	31.65 %	29.27 %
Údržba a oprava prostriedkov osobnej dopravy	79.31 %	78.67 %
Ostatné služby týkajúce sa prostriedkov osobnej dopravy	65.93 %	64.46 %
Železničná osobná preprava	94.59 %	94.56 %
Cestná osobná doprava	80.3 %	78.93 %
Osobná letecká doprava	98.87 %	98.6 %
Námorná a vnútrozemská vodná osobná doprava	99.98 %	99.99 %
Kombinovaná osobná doprava	93.71 %	94.76 %
Ostatné zakúpené dopravné služby	99.49 %	99.49 %
Poštové služby	93.7 %	94.18 %
Telefónne a faxové zariadenia	62.78 %	52.63 %
Telefónne a faxové služby	4.07 %	4.11 %
Zariadenia na príjem, záznam a reprodukciu zvuku a obrazu	88.7 %	89.61 %
Fotografické a filmové zariadenia a optické prístroje	99.15 %	98.93 %
Zariadenia na spracovanie informácií	88.81 %	88.04 %
Nahrávacie médiá	97.37 %	98.09 %
Oprava audiovizuálnej a fotografickej techniky a zariadení na spracovanie informácií	99.25 %	99.33 %
Významnejšie tovary dlhodobej spotreby na rekreáciu vonku	99.62 %	99.49 %
Hudobné nástroje a veľké zariadenia dlhodobej spotreby na rekreáciu	99.21 %	99.13 %
Údržba a oprava ostatných významnejších tovarov dlhodobej spotreby na rekreáciu a kultúru	99.73 %	99.77 %
Hry. hračky a záľuby	75.98 %	78.3 %
Pomôcky na šport. kempovanie a rekreáciu v prírode	89.24 %	88.22 %
Záhrady. rastliny a kvety	70.1 %	65.53 %
Domáce zvieratá a s nimi súvisiace výrobky	66.77 %	64.74 %
Veterinárne a ostatné služby pre domáce zvieratá	89.17 %	88.49 %

Premenná	**RÚ**	**EU - SILC**
Rekreačné a športové služby	82.6 %	81.53 %
Kultúra	21.94 %	23.97 %
Hazardné hry	96.77 %	97.86 %
Knihy	79.71 %	79.42 %
Noviny a periodiká	70.69 %	69.08 %
Rôzne tlačoviny	93.86 %	93.53 %
Kancelárske potreby a potreby na kreslenie	67.06 %	64.27 %
Organizované dovolenkové zájazdy	84.1 %	82.03 %
Predprimárne a primárne vzdelávanie	92.04 %	93.36 %
Sekundárne vzdelávanie	95.52 %	95.76 %
Post-sekundárne vzdelávanie (nezahrňované do terciárneho)	99.76 %	99.87 %
Vysokoškolské vzdelávanie	97.53 %	98.12 %
Vzdelávanie nedefinovateľné podľa úrovňou	93.53 %	93.45 %
Reštaurácie, kaviarne a podobné zariadenia	38.56 %	37.76 %
Závodné jedálne	78.05 %	77.39 %
Ubytovacie služby	82.06 %	82.69 %
Kadernícke služby a skrášľovacie služby	60.8 %	57.75 %
Elektrické spotrebiče na osobnú starostlivosť	95.37 %	94.66 %
Ostatné pomôcky a výrobky pre osobnú starostlivosť	5.52 %	4.25 %
Prostitúcia	100 %	100 %
Šperky. hodiny a náramkové hodinky	91.69 %	90.78 %
Ostatné osobné predmety	77.05 %	77.73 %
Sociálna ochrana	99.83 %	99.56 %
Poistenie v súvislosti s bývaním	65.05 %	61.55 %
Zdravotné poistenie	42.88 %	56.83 %
Poistenie v súvislosti s dopravou	39.67 %	38.46 %
Ostatné poistenie	96.8 %	97.52 %
Ostatné finančné služby i n	94.63 %	93.81 %
Ostatné služby i n	92.67 %	92.87 %

Tabuľka 38: Porovnanie % domácností s nulovými výdavkami v RÚ a EU - SILC pri agregovaní do 115 výdavkových skupín (s využitím váh domácností)

F Príloha

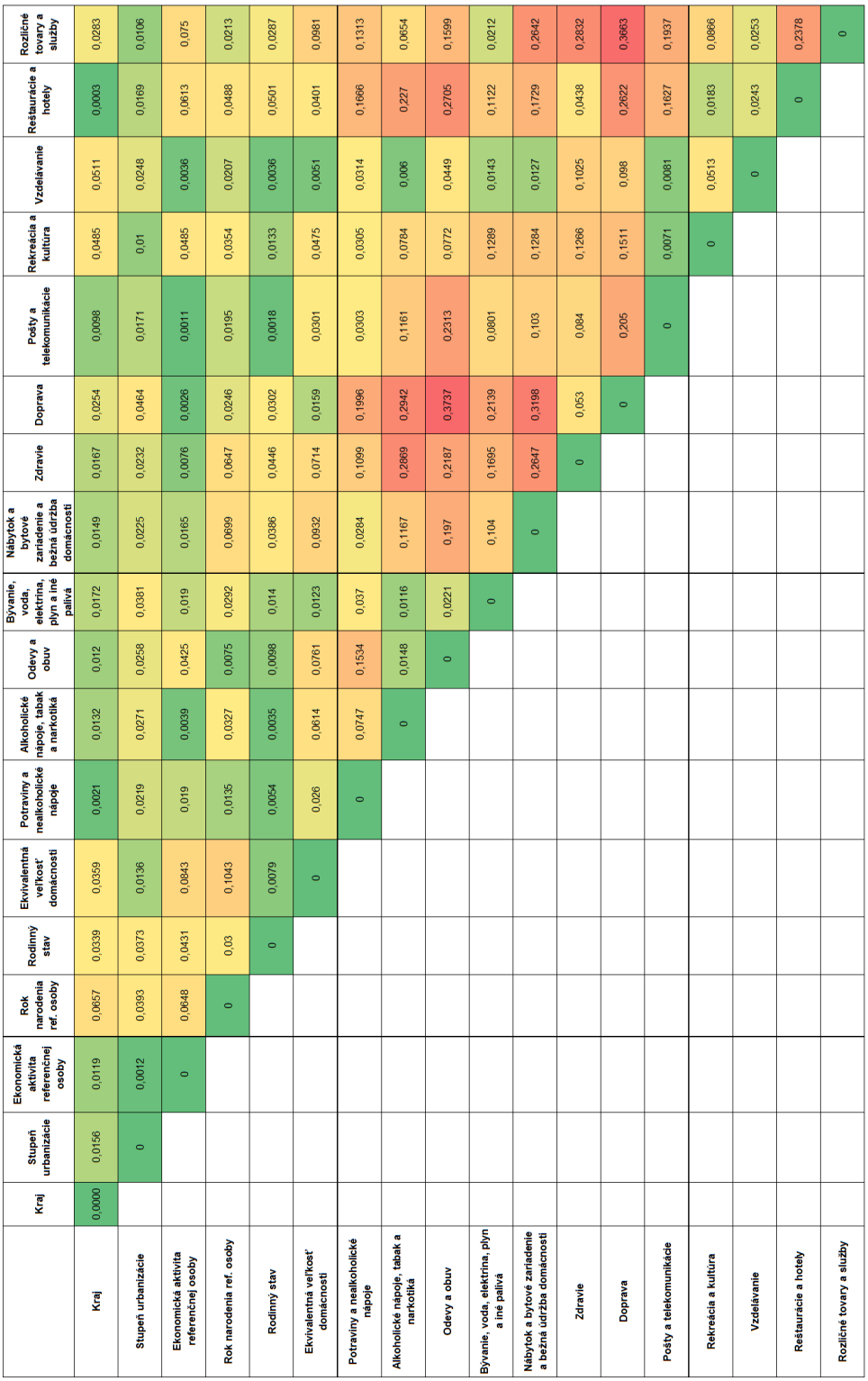
	Výdavkový decil									
Príjmový decil	1	2	3	4	5	6	7	8	9	10
1	4.39 %	1.67 %	0.95 %	0.91 %	0.4 %	0.23 %	0.77 %	0.41 %	0.13 %	0.1 %
2	2.19 %	1.99 %	1.56 %	1.03 %	0.77 %	0.61 %	0.47 %	0.36 %	0.28 %	0.75 %
3	1.15 %	1.86 %	1.56 %	1.58 %	0.98 %	0.94 %	0.82 %	0.46 %	0.44 %	0.35 %
4	0.67 %	1.19 %	1.22 %	1.55 %	0.96 %	1.33 %	0.66 %	0.86 %	1 %	0.49 %
5	0.67 %	0.51 %	0.96 %	1.26 %	1.07 %	1.58 %	1.16 %	1.02 %	0.79 %	1 %
6	0.32 %	0.6 %	0.75 %	0.9 %	1.62 %	1.16 %	1.24 %	1.06 %	1.24 %	1.08 %
7	0.04 %	0.57 %	0.62 %	0.57 %	1.04 %	1.2 %	1.39 %	1.65 %	1.67 %	1.24 %
8	0.23 %	0.48 %	1.27 %	0.48 %	0.91 %	1.24 %	1.23 %	1.72 %	0.99 %	1.48 %
9	0.04 %	0.46 %	0.56 %	1.06 %	1.68 %	0.77 %	1.1 %	0.96 %	1.81 %	1.57 %
10	0.34 %	0.66 %	0.65 %	0.59 %	0.58 %	0.94 %	1.17 %	1.59 %	1.54 %	1.93 %

Tabuľka 39: Združená distribúcia príjmových a výdavkových decilov databáze RÚ (s využitím váh domácností)

	Výdavkový decil									
Príjmový decil	1	2	3	4	5	6	7	8	9	10
1	5.43 %	1.58 %	0.85 %	0.73 %	0.6 %	0.08 %	0.35 %	0.15 %	0.13 %	0.11 %
2	2.33 %	2.42 %	1.4 %	1.12 %	0.71 %	0.41 %	0.45 %	0.31 %	0.42 %	0.42 %
3	0.61 %	1.68 %	1.62 %	1.79 %	1.05 %	0.92 %	1.07 %	0.72 %	0.28 %	0.27 %
4	0.48 %	1.26 %	1.49 %	1.02 %	1.21 %	1.45 %	0.81 %	1.17 %	0.75 %	0.37 %
5	0.44 %	0.58 %	1.01 %	1.45 %	1.26 %	1.54 %	0.97 %	1.29 %	0.69 %	0.75 %
6	0.32 %	0.52 %	1.06 %	0.83 %	1.41 %	1.55 %	1.06 %	1.06 %	1.19 %	1.01 %
7	0.15 %	0.48 %	0.67 %	1.41 %	0.88 %	1.09 %	1.36 %	1.46 %	2.02 %	0.48 %
8	0.15 %	0.58 %	1.09 %	0.56 %	0.91 %	1.31 %	1.36 %	1.68 %	1.09 %	1.25 %
9	0.09 %	0.38 %	0.7 %	0.45 %	0.99 %	0.78 %	1.38 %	1.51 %	1.69 %	2.07 %
10	0.02 %	0.51 %	0.13 %	0.62 %	0.98 %	0.88 %	1.19 %	0.64 %	1.75 %	3.25 %

Tabuľka 40: Združená distribúcia príjmových a výdavkových decilov databáze EU - SILC (s využitím váh domácností)

G Príloha



	Kraj	Stupeň urbanizácie	Ekonomická aktivita referenčnej osoby	Rok narodenia ref. osoby	Rodinný stav	Ekvivalentná veľkosť domácnosti	Potraviny a nealkoholické nápoje	Alkoholické nápoje, tabak a narkotiká	Odevy a obuv	Bývanie, voda, elektrina, plyn a iné palivá	Nábytok a bytové zariadenie a bežná údržba domácnosti	Zdravie	Doprava	Pošty a telekomunikácie	Rekreácia a kultúra	Vzdelávanie	Reštaurácie a hotely	Rozličné tovary a služby
Kraj	0,0000	0,0156	0,0119	0,0657	0,0339	0,0359	0,0021	0,0132	0,012	0,0172	0,0149	0,0167	0,0254	0,0098	0,0485	0,0511	0,0003	0,0283
Stupeň urbanizácie		0	0,0012	0,0393	0,0373	0,0136	0,0219	0,0271	0,0258	0,0381	0,0225	0,0232	0,0464	0,0171	0,01	0,0248	0,0169	0,0106
Ekonomická aktivita referenčnej osoby			0	0,0648	0,0431	0,0843	0,019	0,0039	0,0425	0,019	0,0165	0,0076	0,0026	0,0011	0,0485	0,0036	0,0613	0,075
Rok narodenia ref. osoby				0	0,03	0,1043	0,0135	0,0327	0,0075	0,0292	0,0699	0,0647	0,0246	0,0195	0,0354	0,0207	0,0488	0,0213
Rodinný stav					0	0,0079	0,0054	0,0035	0,0098	0,014	0,0386	0,0446	0,0302	0,0018	0,0133	0,0036	0,0501	0,0287
Ekvivalentná veľkosť domácnosti						0	0,026	0,0614	0,0761	0,0123	0,0932	0,0714	0,0159	0,0301	0,0475	0,0051	0,0401	0,0981
Potraviny a nealkoholické nápoje							0	0,0747	0,1534	0,037	0,0284	0,1099	0,1996	0,0303	0,0305	0,0314	0,1666	0,1313
Alkoholické nápoje, tabak a narkotiká								0	0,0148	0,0116	0,1167	0,2869	0,2942	0,1161	0,0784	0,006	0,227	0,0654
Odevy a obuv									0	0,0221	0,197	0,2187	0,3737	0,2313	0,0772	0,0449	0,2705	0,1599
Bývanie, voda, elektrina, plyn a iné palivá										0	0,104	0,1695	0,2139	0,0801	0,1289	0,0143	0,1122	0,0212
Nábytok a bytové zariadenie a bežná údržba domácnosti											0	0,2647	0,3198	0,103	0,1284	0,0127	0,1729	0,2642
Zdravie												0	0,053	0,084	0,1266	0,1025	0,0438	0,2832
Doprava													0	0,205	0,1511	0,098	0,2622	0,3663
Pošty a telekomunikácie														0	0,0071	0,0081	0,1627	0,1937
Rekreácia a kultúra															0	0,0513	0,0183	0,0866
Vzdelávanie																0	0,0243	0,0253
Reštaurácie a hotely																	0	0,2378
Rozličné tovary a služby																		0

Obr. 27: Rozdiel medzi korelačnou maticou z RÚ a EU-SILC ($\Delta C(x, w)$) (s využitím váh domácností). Názvy výdavkových premenných symbolizujú podiel daných premenných na celkovom príjme.