UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Struktura demografických dat

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Lucie Kawecká**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2019

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Lucie Kawecká

Název práce: Struktura demografických dat

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2019

Abstrakt: Stárnutí populace, migrace a další demografické procesy jsou jedním z důležitých témat současnosti. Cílem bakalářské práce je provést analýzu relativní struktury vybraných demografických dat užitím hierarchického shlukování a provést též odpovídající interpretaci výsledků v kontextu aktuálního vývoje.

Klíčová slova: populační pyramida, symbolická analýza dat, metrika, vzdálenost, shluková analýza

Počet stran: 41

Počet příloh: 2

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Lucie Kawecká

Title: Structure of demographic data

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2019

Abstract: Aging of population, migration, and other demographic processes are one of the most important themes of these days. The aim of bachelor's thesis is to analyse relative structure of demographic data using hierarchical clustering in frame of symbolic data analysis and make relevant interpretation of results in context actual development.

Key words: population pyramid, symbolic data analysis, metric, distance, cluster analysis

Number of pages: 41

Number of appendices: 2

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením pana doc. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych poděkovala vedoucímu bakalářské práce doc. RNDr. Karlu Hronovi, Ph.D. za spolupráci, ochotu a čas, který mi věnoval při konzultacích. Také bych chtěla poděkovat své rodině a přátelům, kteří mě po celou dobu studia podporovali.

Úvod

Demografické jevy a procesy jsou důležitým tématem současné doby a často bývají vyjádřené ve formě populačních pyramid. Obsahem mé práce je navržení postupu statistického zpracování populačních pyramid jako relativní demografické struktury a aplikaci na data týkající se relativní struktury výskytu žen a mužů v jednotlivých krajích České republiky.

V první kapitole se seznámíme s demografií jako s vědním oborem, jejími základními pojmy, se strukturou obyvatelstva a s propojením demografie a statistiky. Dále bude následovat převedení populačních pyramid do kontextu symbolické analýzy, kde zadefinuji modální proměnnou a popíši postup úpravy dat. Třetí kapitola popisuje metriky vhodné pro symbolickou analýzu. Zde se snažím čtenáři přiblížit i méně známé metriky, jejich pozitiva a negativa a aplikaci na příkladech. V předposlední kapitole je představen teoretický základ ke shlukové analýze, která byla posledním krokem mého zpracování. V páté kapitole uvedu grafy a výsledky analýzy na svých datech.

Téma jsem si vybrala díky praktickému zaměření. Demografická data mě nadchla a myslím si, že jejich zpracování a výsledky jsou jak zajímavé, tak velmi užitečné.

Kapitola 1

Demografie a statistika

Demografie je vědní obor, který studuje proces reprodukce lidských populací. Ty jsou chápané „*jako přirozená obnova stavu obyvatelstva prostřednictvím biosociálních procesů porodnosti a úmrtnosti a jednak jako celková obnova obyvatelstva, zahrnující i obnovu obyvatelstva jeho stěhováním.*“ [14, kap. 1]. Název tohoto vědního oboru pochází z řečtiny a je složen ze slov *démos* (dříve překládáno jako „obec“, nyní jako „lid“) a *grafein* („psáti“).

Při tvorbě této kapitoly bylo čerpáno ze zdrojů [2], [10], [13], [14], [16].

1.1. Demografická statistika

Demografie ke svému výzkumu získává konkrétní informace ze statistiky obyvatelstva. Ty jsou pro ni zásadní, neboť jde o empirický materiál, bez kterého nemůže existovat. Demografické informace jsou dvojího druhu:

1. informují o stavu,

2. informují o pohybu.

Stavové informace popisují rozsah populace (tj. počet členů populace) a její strukturu podle zajímavých rysů, které se vážou k určitému, předem domluvenému časovému okamžiku. Zjišťují se *soupisem obyvatelstva* nebo *sčítáním lidu*.

Pohybem rozumíme události, které se bezprostředně vztahují k reprodukci populace. Jde o narození, sňatek, rozvod, přestěhování se a úmrtí.

Záznamy evidující narození, sňatek, rozvod a úmrtí se nazývají *evidence přirozené měny*. Stěhování je zaznamenáváno v *evidenci migrace*.

1.2. Základní pojmy

V této kapitole uvedeme základní pojmy z oboru demografie, které budou v následujícím textu použité.

- **Populace** je soubor osob nacházejících se na daném území v určitém čase.
- **Populační struktura** popisuje okamžikový stav populace. Zvlášť se vymezuje pro ženské a mužské pohlaví kvůli značným odlišnostem. Při nerozlišení by došlo ke značnému zkreslení dat.
- **Kohorta** „*Označuje skupinu osob, jichž se během určitého časového intervalu týkala nějaká konkrétní událost; tou událostí může být samozřejmě i narození, takže termín kohorta zahrnuje i generaci.*“ [10]
- **Natalita** (porodnost) vyjadřuje počet narozených jedinců. Její hrubá míra je podíl narozených jedinců z dané skupiny ku celkovému počtu obyvatel v dané skupině za určité období. Je ovlivněna velikostí časového intervalu a velikostí sledované populace. Uvádí se v promilích, takže v přepočtu na 1 000 jedinců.
- **Mortalita** (úmrtnost) vyjadřuje počet zemřelých jedinců. Její hrubá míra je podíl zemřelých jedinců z dané skupiny ku celkovému počtu obyvatel v dané skupině za určité období. S rostoucím věkem populace se mortalita významně zvyšuje. Uvádí se v promilích.
- **Fertilita** (plodnost) vyjadřuje počet živě narozených jedinců. Její obecná míra je dána podílem živě narozených jedinců ku počtu žen v reprodukčním věku (15-49 let).

1.3. Struktura obyvatelstva

Na stavu populace můžeme sledovat spoustu aspektů, tj. můžeme studovat různé struktury. Jedná se o studii okamžikového stavu populace, tedy o demografickou statiku. V následujícím textu se budeme zabývat strukturou podle pohlaví a věku, která obvykle nese jednotný název demografická struktura. Další pro demografii zajímavé struktury jsou:

- struktura podle rodinného stavu a typu domácnosti,
- ekonomická struktura (třídění na ekonomicky aktivní a neaktivní),
- geografická struktura,
- ostatní struktury, např. podle vzdělání, náboženství, národnosti...

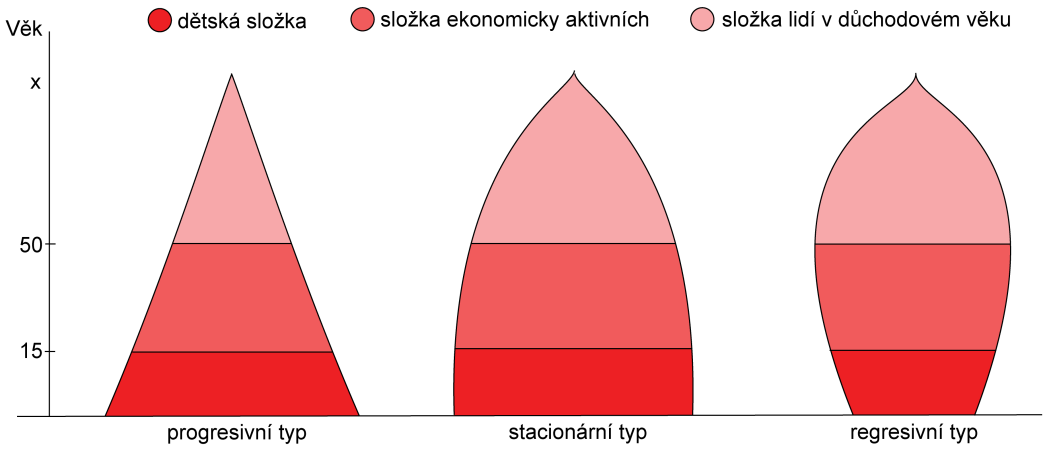
1.3.1. Demografická struktura

Studie demografické struktury obyvatelstva je založena na třídění podle pohlaví a věku. To je důsledkem demografických procesů, které se vyvíjí v populaci po dobu předešlých desetiletí a stejně tak do určité míry předesílá vývoj populace v nadcházejících dekádách.

Grafickým znázorněním demografické struktury je **věková pyramida**. Věková pyramida popisuje demografické rozložení populace. Může naznačovat události minulosti, současnou situaci a budoucí demografické trendy, například pohlavní a věkové rozdílnosti, reproduktivní schopnosti a další demografické znaky. Šířka základny svědčí o natalitě. Široká základna vypovídá o vysoké natalitě, zatímco úzká základna o nízké. Tvar jejích stran popisuje mortalitu. Konkávní tvar strany vypovídá o vysoké mortalitě, zatímco konvexní tvar vypovídá o nízké míře mortality. Zkoumání populačního rozložení mezi předreproduktivní (0 - 14 let), reproduktivní (15 - 49 let) a postreproduktivní (50 let a výše) věkovou skupinou vypovídá o demografických zátěžích a poměrech závislosti mezi nimi. Hrboly nebo náhlé lomy po stranách odhalují určité anomálie jako je například rychlý nárůst nebo pokles populace (tato kolísání způsobují například baby boomy, masová

migrace, války, epidemie...). Asymetrie mezi stranami naznačuje disproporce mezi mužskou a ženskou populací ve stejné věkové kohortě (může jít například o výsledky odlišností délky života nebo mužské masové migrace za prací).

Rysy ve struktuře populačních pyramid poprvé popsal švédský demograf a statistik Axel Gustav Sundbärg. Své poznatky vydal v roce 1900. Progresivní model představuje typicky mladou a rozrůstající se populaci, většinou s rozvíjející se ekonomikou. Jde například o země třetího světa. Stacionární model se svým zvonovým tvarem popisuje stabilní demografickou situaci bez přírůstku ani poklesu populace. Tento tvar je typický pro severské státy. Regresivní tvar pyramidy se podobá tvaru urny. Je typický pro starší a zmenšující se populaci.



Obrázek 1.1: Typy věkových pyramid [18]

Pro svou práci jsme vybrali data ve formě populačních pyramid 14 krajů České republiky, protože jejich tvar plyne z fertility, mortality a migrace. Migrační proudy plynou z různých pull and push faktorů. Push faktory jsou takové, které nutí migranty k opuštění svého dosavadního domova (např. nezaměstnanost, nízká kvalita života apod.). Pull faktory jsou přitažlivé, tzn. že vedou migranty k přesunutí se do určitých cílových zemí (např. vysoká kvalita života, příležitost zaměstnání, možnost seberealizace apod.).

Fertilita s mortalitou plynou například ze sociálních priorit a volby životního stylu, ekonomických a životních podmínek, přírodních procesů, přístupu ke kva-

litní zdravotní péči. Proto na tvar regionálních populačních pyramid mohou působit faktory sociální, ekonomické, politické včetně faktoru životního prostředí. Avšak demografické studie ukazují, že populační struktura v Evropě plyne hlavně z ekonomických faktorů [13]. Navíc, regionální populační pyramidy mohou být ovlivňovány jinými faktory než ty popisující rozložení celých států nebo ještě větších území. Když porovnáme regiony, měli bychom mít lepší pochopení pro demografické procesy v České republice.

1.4. Data

Analýzu jsme prováděli na datech zahrnujících 14 krajů ČR: Hlavní město Praha, Středočeský kraj, Jihočeský kraj, Plzeňský kraj, Karlovarský kraj, Ústecký kraj, Liberecký kraj, Královehradecký kraj, Pardubický kraj, Kraj Vysočina, Jihomoravský kraj, Olomoucký kraj, Moravskoslezský kraj a Zlínský kraj. Zároveň se jedná o NUTS 3 [1].

Dataset tedy zahrnuje NUTS populaci s 18 věkovými skupinami, které mají vždy pětileté rozpětí ((0,5); [5,10); [10,15);...;[85,∞)) třízené podle pohlaví k 1. lednu 2015. Empirická data pochází z Eurostatu. Ochotně mi je k práci poskytla dr. Justyna Wilk z Univerzity Adama Mickiewicze v Poznani (Polsko), spoluautorka článku [2].

Poznamenejme, že Eurostat při sběru dat pracuje s trvalým bydlištěm osob. Trvalé bydliště znamená místo, kde člověk tráví denní dobu, bez ohledu na dočasnou absenci za účelem odpočinku, dovolené, práce, lékařských potřeb apod.

[1] NUTS (Nomenclature of territorial units for statistics) jsou územní celky vytvořené pro statistické účely Eurostatu. Pro NUTS 3 jsou českým ekvivalentem kraje.

Kapitola 2

Symbolická analýza dat

Symbolická analýza dat (SDA) je rozšířením standardní analýzy. Využívá se hlavně k analýze velkých datových souborů, protože svým přístupem neztrácí při sumarizaci nepřiměřeně mnoho informace v datech obsažené. Zpracování dat funguje na principu vzniku nových proměnných, tzv. symbolických proměnných. Speciálním případem symbolických dat mohou být i tzv. kompoziční data, která zmíníme dále v práci.

Cílem této kapitoly je seznámit čtenáře se symbolickým přístupem k datům a zároveň s postupem, který jsme využili k analýze demografických dat. Ve své práci budeme následovat přístup Rogera S. Bivanda, Justyny Wilkové a Tomasze Kossowského, který popsali ve svém článku „*Spatial association of population pyramids across Europe: The application of symbolic data, cluster analysis and join-count tests*“ [2]. Při tvorbě této kapitoly jsem čerpala ze zdrojů [1], [2], [9].

2.1. Populační pyramidy ve smyslu symbolických dat

V této kapitole zaveďme značení potřebné k dalšímu postupu.

Mějme soubor $\mathbf{E}$ o n jednotkách. Každá jednotka je popsána p proměnnými $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p$. Pro každou proměnnou uvažujeme zobrazení $\mathbf{Y}_i : \mathbf{E} \to \mathbf{y}_i$, kde množina možných hodnot $\mathbf{y}_i$ tvoří definiční obor proměnné $\mathbf{Y}_i$. Pro každou jednotku $j \in \mathbf{E}$ se proměnná $\mathbf{Y}_i$ realizuje pouze v jedné hodnotě $\mathbf{y}_i$, označeno jako

$\mathbf{y}_{ji} = \mathbf{Y}_i(j)$. Hodnoty definičního oboru $\mathbf{y}_i$ mohou být také kvalitativní nebo kvantitativní. Finálním výsledkem přístupu je matice $\mathbf{Y} = [\mathbf{y}_{ji}]$, kde složkami jsou realizace p proměnných ze souboru $\mathbf{E}$.

Uvažujeme-li proměnnou $\mathbf{Y}$ s definičním oborem $\mathbf{y}$, potom množinu všech možných podmnožin $\mathbf{y}$ nazveme $\mathbf{P}(\mathbf{y})$. Pro další úvahy zavedeme systém

$$\boldsymbol{B} = \boldsymbol{P}(\boldsymbol{y})\backslash\{\emptyset\} = \{\boldsymbol{U} : \boldsymbol{U} \subseteq \boldsymbol{y}, \boldsymbol{U} \neq \emptyset\}.$$

Jde tedy pouze o systém všech možných podmnožin $\mathbf{P}(\mathbf{y})$ bez prázdné množiny.

V našem případě si tohle značení můžeme představit následovně: máme datový soubor $\mathbf{E}$ o n jednotkách - krajích. Každá jednotka je popsána p proměnnými - u nás jde o jednotlivé věkové skupiny dělené ještě podle pohlaví. Hodnoty našeho definičního oboru jsou kvantitativního typu, jde o frekvence výskytu pozorování v jednotlivých věkových skupinách.

Pomocí následujících definic zavedeme tzv. *modální proměnnou*, která bude potřeba pro další postup.

Definice 2.1. Zobrazení $\mathbf{Y} : \mathbf{E} \to \mathbf{B}$ nazveme *množinová proměnná* s definičním oborem $\mathbf{y}$. Pro každé j přiřazuje hodnotu $\mathbf{Y}(j) = \mathbf{U} \in \mathbf{B}$. Jestliže pro každé $j \in \mathbf{E}$ platí $\|\mathbf{Y}(j)\| = 1$, potom dostáváme *jednohodnotovou proměnnou*.

Definice 2.2. Zobrazení $\mathbf{Y} : \mathbf{E} \to \mathbf{B}$ nazveme *vícehodnotová proměnná* , jestliže je *množinovou proměnnou* a pro její hodnoty platí $|\mathbf{Y}(j)| < \infty$ (je konečná podmnožina $\mathbf{y}$) pro $\forall j \in \mathbf{E}$.

Pro další definici uvažujme, že definiční obor $\mathbf{y}$ nabývá pouze reálných hodnot nebo žádné přímo uspořádané hodnoty s pořadím $\alpha \prec \beta$.

Definice 2.3. Pokud hodnota proměnné $\mathbf{Y}(j) = \mathbf{U} \in \mathbf{B}$ pro každé $j \in \mathbf{E}$, kde $\mathbf{B}$ je množina intervalů v $\mathbb{R}$ s omezením $\alpha, \beta, \alpha \leq \beta$, nebo hodnota proměnné $\mathbf{Y}(j) = [\alpha, \beta] \in \mathbf{B}$ pro každé $j \in \mathbf{E}$, kde $\mathbf{B}$ je množina intervalů v $\mathbf{y}$ s ohledem na dané pořadí $\prec$ v $\mathbf{y}$, a $\alpha \prec \beta$, potom se $\mathbf{Y}$ nazývá *intervalová proměnná*.

Pro další potřeby zavedeme nezápornou míru π na $\mathbf{y}$ za předpokladu $\mathbf{U} \in \mathbf{B}$, $\mathbf{U} \subseteq \mathbf{y}$.

Definice 2.4. Zobrazení $\mathbf{Y} : \mathbf{E} \to (\mathbf{B}, \pi)$, kde pro každé $j \in \mathbf{E}$ máme $\mathbf{Y}(j) = (\mathbf{U}(j), \pi_j)$ se nazývá *modální proměnná* s definičním oborem $\mathbf{y}$ možných hodnot.

Definice 2.5. Zobrazení $\mathbf{Y} : \mathbf{E} \to \mathbf{B}$, kde $\mathbf{B} = \mathbf{M}(\mathbf{y})$ patří do skupiny nezáporných hodnot π v definičním oboru s hodnotami $\mathbf{Y}(j) = \pi_j$, je nazývána modální proměnnou s definičním oborem $\mathbf{y}$.

Popsaný koncept modálních proměnných aplikujeme na populační pyramidy našich 14 krajů. Populační pyramida je graf sestávající se z 2 přidružených dimenzí (věk a pohlaví). Strany tohoto grafu reprezentují mužské a ženské rozložení společnosti. Každá ze stran je dále rozložena na 18 pětiletých skupin. Z toho plyne, že populační pyramida $\mathbf{A}_j$ pro všechny jednotky $j \in \mathbf{E}$ je rozdělena na ženy a muže, což může být zapsáno jako rozložení množiny $\mathbf{B}$ do 18 intervalů:

$$f(x) = \begin{cases} (0, 5), & k = 1, \\ [5(k-1), 5k), & k = 2, \ldots, 17, \\ [85, \infty), & k = 18. \end{cases} \tag{2.1}$$

Můžeme vidět, že $\mathbf{U}_l \cap \mathbf{U}_m = \emptyset$ pro každé $l, m = 1, \ldots, 18, l \neq m$, a $\mathbf{U}_k \in \mathbf{B}$, $\bigcup_k \mathbf{U}_k \subseteq \mathbf{B}$.

Dále definujeme dvě zobrazení, $\mathbf{M} : \mathbf{E} \to \mathbf{B}$ a $\mathbf{F} : \mathbf{E} \to \mathbf{B}$, $\mathbf{M}$ pro populaci mužů a $\mathbf{F}$ pro populaci žen. Takto definované proměnné považujeme za modální proměnné s definičním oborem $\mathbf{B}$ popsaným v definici 2.1.

Nakonec formalizujeme populační pyramidy. Míru π interpretujeme jako proporci mužské nebo ženské populace v každé věkové skupině ve vztahu k celé populaci jednotlivého kraje. Transformujeme obě proměnné $\mathbf{M}$ a $\mathbf{F}$ ve smyslu definice 2.4 do tvaru modálních proměnných a zjednodušíme pomocí definice 2.5.

Následující příklad reprezentuje 2 populační pyramidy rozložené ve smyslu symbolických dat. Množina frekvencí má 18 věkových kategorií zvlášť pro ženy a pro muže. Součet frekvencí pro každý kraj je roven 1. Rozložené populační pyramidy mohou být zapsány následovně:

Praha = [Males: {0.0278 (0,5), 0.0252 [5,10), 0.0176 [10,15), 0.0174 [15,20), 0.0264 [20,25), 0.0353 [25,30), 0.042 [30,35), 0.0484 [35,40), 0.0393 [40,45), 0.0318 [45,50),0.0308 [50,55), 0.0307 [55,60), 0.0349 [60,65), 0.0352 [65,70), 0.0265 [70,75), 0.0165 [75,80), 0.015 [80,85), 0.01452 [85,∞)}] $\wedge$ [Females: {0.0293 (0,5),0.0264 [5,10), 0.0187 [10,15), 0.0179 [15,20), 0.0258 [20,25), 0.0355 [25,30), 0.0427 [30,35), 0.0506 [35,40), 0.0419 [40,45), 0.0326 [45,50),0.0303 [50,55), 0.028 [55,60), 0.0302 [60,65), 0.0281 [65,70), 0.0203 [70,75), 0.0111 [75,80), 0.0087 [80,85), 0.0065 [85,∞)}]

Olomoucký kraj = [Males: {0.0247 (0,5), 0.0264 [5,10), 0.0219 [10,15), 0.0219 [15,20), 0.0300 [20,25), 0.0323 [25,30), 0.0331 [30,35), 0.041 [35,40), 0.038 [40,45), 0.0318 [45,50),0.0316 [50,55), 0.034 [55,60), 0.0364 [60,65), 0.0349 [65,70), 0.0265 [70,75), 0.0187 [75,80), 0.0152 [80,85), 0.0129 [85,∞)}] $\wedge$ [Females: {0.0259 (0,5), 0.0274 [5,10), 0.0231 [10,15), 0.0228 [15,20), 0.0314 [20,25), 0.03356 [25,30), 0.0351 [30,35), 0.044 [35,40), 0.0398 [40,45), 0.0332 [45,50),0.0322 [50,55), 0.0326 [55,60), 0.0335 [60,65), 0.0288 [65,70), 0.0199 [70,75),0.0121 [75,80), 0.0082 [80,85), 0.0052 [85,∞)}].

Kapitola 3

Metriky vhodné pro symbolickou analýzu dat

Tato kapitola se zabývá metrikami vhodnými pro použití na datech, která byla zpracovaná symbolickým přístupem. Zároveň jsme tyto metriky využili pro tvorbu matic vzdáleností při shlukové analýze. V této kapitole jsem čerpala ze zdrojů [2], [3], [4], [5], [8], [11], [12], [15], [17], [19].

3.1. Kullback-Leiblerova vzdálenost

Kullback-Leiblerova divergence (také KL divergence) byla představena Solomonem Kullbackem a Richardem Leiblerem v roce 1951. Je založena na rozdílu mezi dvěma rozděleními pravděpodobnosti na stejném oboru hodnot. Můžeme ji uvažovat pro spojitou i diskrétní náhodnou veličinu.

$D_{KL}(X \,\|\, Y)$ čteme jako Kullback-Leiblerova divergence od X k Y. Pro spojité rozdělení pravděpodobnosti veličin X a Y, kde $p(x)$ a $q(y)$ jsou jejich hustoty, definujeme KL divergenci jako

$$D_{KL}(X \parallel Y) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \ .$$

Pro diskrétní rozdělení pravděpodobnosti veličin X a Y, kde $p(x)$ a $q(y)$ jsou pravděpodobnostní funkce těchto veličin, definujeme KL divergenci jako

$$D_{KL}(X \parallel Y) = \sum_{x \in M} p(x) \ln \frac{p(x)}{q(x)},$$

kde M značí obor hodnot. Nulová je právě tehdy, když rozdělení X a Y jsou totožná.

Kullback-Leiblerova divergence je zvláštní případ širší třídy odchylek nazývaných f-divergence. Ačkoli je často chápána jako metoda měření vzdálenosti mezi rozděleními pravděpodobnosti, nejde o skutečnou metriku, protože porušuje axiom symetrie, tzn. $D_{KL}(X \parallel Y)$ není to samé jako $D_{KL}(Y \parallel X)$.

Její symetrickou formu, kterou již považujeme ze metriku, zavádíme vztahem

$$d_{KL}(X \parallel Y) = \frac{D_{KL}(X \parallel Y) + D_{KL}(Y \parallel X)}{2}. \tag{3.1}$$

V našem případě uvažujeme zvlášť rozdělení pravděpodobnosti pro každý kraj pro dané pohlaví. Na naší datové sadě si to můžeme představit následovně: X reprezentuje rozdělení pravděpodobnosti výskytu žen v jednom kraji a Y reprezentuje rozdělení pravděpodobnosti výskytu žen v jiném kraji. Vzdálenost mezi nimi vypočítáme pomocí vztahu 3.1.

Na konci početního procesu je výstupem zvlášť matice vzdáleností mezi jednotlivými kraji pro ženy a zvlášť matice vzdáleností mezi jednotlivými kraji pro muže. Tyto dvě matice následně použijeme pro potřeby shlukové analýzy.

3.2. Wassersteinova vzdálenost

Wassersteinova vzdálenost vychází z tzv. optimální dopravní teorie. Jejím cílem je poskytnout způsob, jak porovnávat dvě rozdělení pravděpodobnosti.

Jde o metriku počítanou mezi dvěma pravděpodobnostními rozděleními veličin X a Y, kde $F(x), F(y)$ jsou jejich distribuční funkce a $F^{-1}(x), F^{-1}(y)$ jsou kvantilové funkce. Obecný vztah pro tuto vzdálenost vypadá následovně:

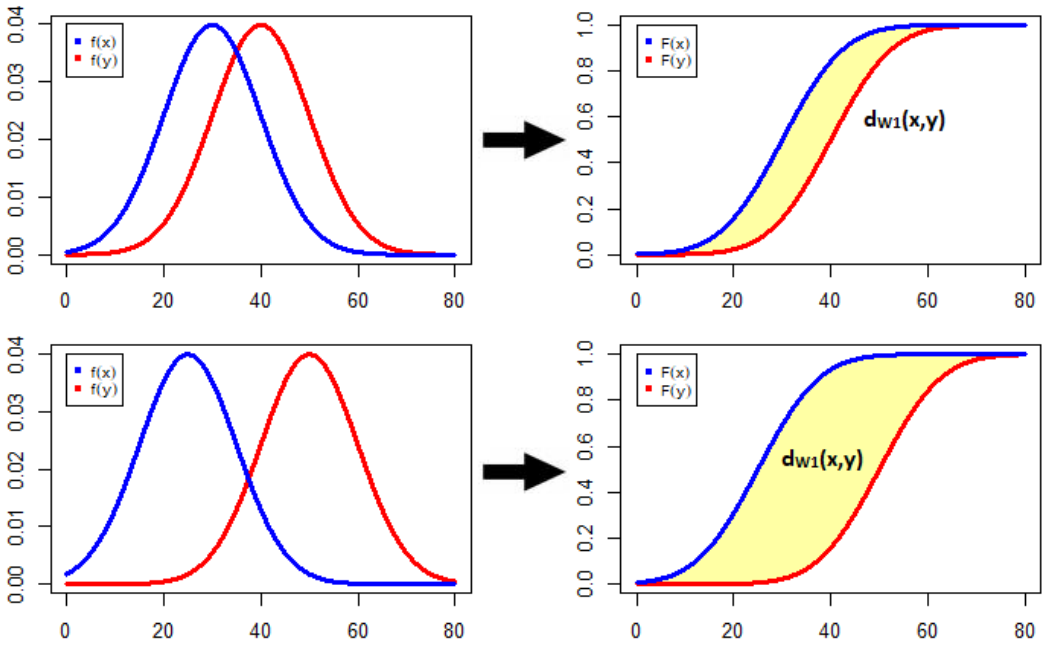
$$d_{W_p}(X, Y) = \left(\int_0^1 |F^{-1}(x) - F^{-1}(y)|^p \, dt \right)^{\frac{1}{p}} \tag{3.2}$$

za předpokladu konečného p-tého obecného momentu.

Ve své práci jsem tuto vzdálenost počítala s druhým konečným momentem, tedy $p = 2$. Vztah pak počítá vzdálenost v prostoru funkcí L^2 a vypadá následovně:

$$d_{W_2}(x, y) = \left(\int_0^1 |F^{-1}(x) - F^{-1}(y)|^2 \, dt \right)^{\frac{1}{2}}, \tag{3.3}$$

V tomto případě tedy počítáme plochu rozdílu mezi kvantilovými funkcemi. Na distribučních funkcích (odpovídající zde spojitým veličinám) to můžeme ilustrovat na obrázku 3.1.

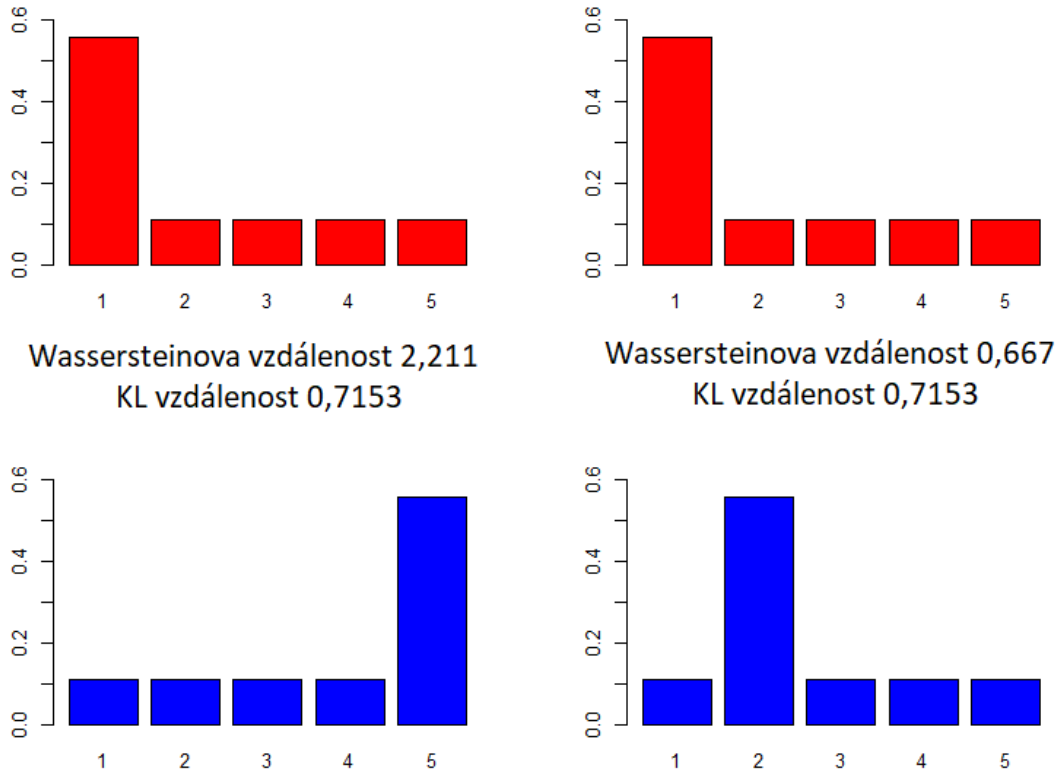


Obrázek 3.1: Rozdíl distribučních funkcí

Koncept uvedení vzdálenosti představil v roce 1969 ruský matematik Leonid Vaserštein, byla ale pojmenována až o rok později R.L. Dobrushinem.

Na uvažovaných datech bylo věkové rozdělení populace považováno za rozdělení pravděpodobnosti a jednotlivé četnosti za váhy (pravděpodobnosti) realizací náhodné veličiny.

Příklad 1. Na následujícím příkladu nastíníme rozdíl mezi Wassersteinovou a KL vzdáleností. Jedním z nejdůležitějších faktů je, že na rozdíl od KL (a mnoha dalších měr) bere Wassersteinova vzdálenost v úvahu metrický prostor. Význam v méně abstraktním slova smyslu vysvětlíme na příkladu.



Obrázek 3.2: Příklad rozdílu KL a Wassersteinovy metriky

Na obrázku 3.2 porovnáváme vždy červený a modrý histogram nad sebou, které zastupují určité rozdělení pravděpodobnosti. KL vzdálenost vychází u obou rozdělení stejně na rozdíl od Wassersteinovy. Ta, jak jsme již uvedli, vychází z transportního problému. Díky tomu uvažuje práci potřebnou k přepravě pravděpodobnostního množství z červeného stavu do modrého pomocí osy x jako „silnice“. Na našem obrázku si to můžeme představit jako přesunutí červeného nejpočetnějšího pole na místo modrého nejpočetnějšího. Wassersteinova vzdálenost je vyšší v prvním případě, protože zde musí urazit delší cestu než v druhém.

Rozdíl bude vyšší, čím vyšší bude ono množství pravděpodobnosti.

Příklad 2. Srovnejme Wassersteinovu vzdálenost pro normální rozdělení pravděpodobnosti s různými středními hodnotami a různými rozptyly. Vektory $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{y}_1$, $\mathbf{y}_2$ vznikly diskretizací normálního rozdělení pomocí histogramů. Wassersteinova vzdálenost byla počítána mezi hodnotami odpovídajícími středům jednotlivých sloupců s váhami odpovídajícím výšce těchto sloupců. Zvolili jsme poměrně hustou diskretizaci, kterou lze ovšem pro účel uvedeného příkladu považovat za dostačující.

V souladu s tím řekněme, že:
$\mathbf{x}_1$ je vektor 5 hodnot odpovídající středům intervalů v histogramu z dat náhodně generovaných z N(40, 10) a $\mathbf{v}_1$ je vektor vah k němu příslušný, $\mathbf{x}_2$ je vektor 5 hodnot odpovídající středům intervalů v histogramu z dat náhodně generovaných z N(40, 15) a $\mathbf{v}_2$ je vektor vah k němu příslušný. Tato normální rozdělení mají stejnou střední hodnotu, ale odlišné rozptyly.

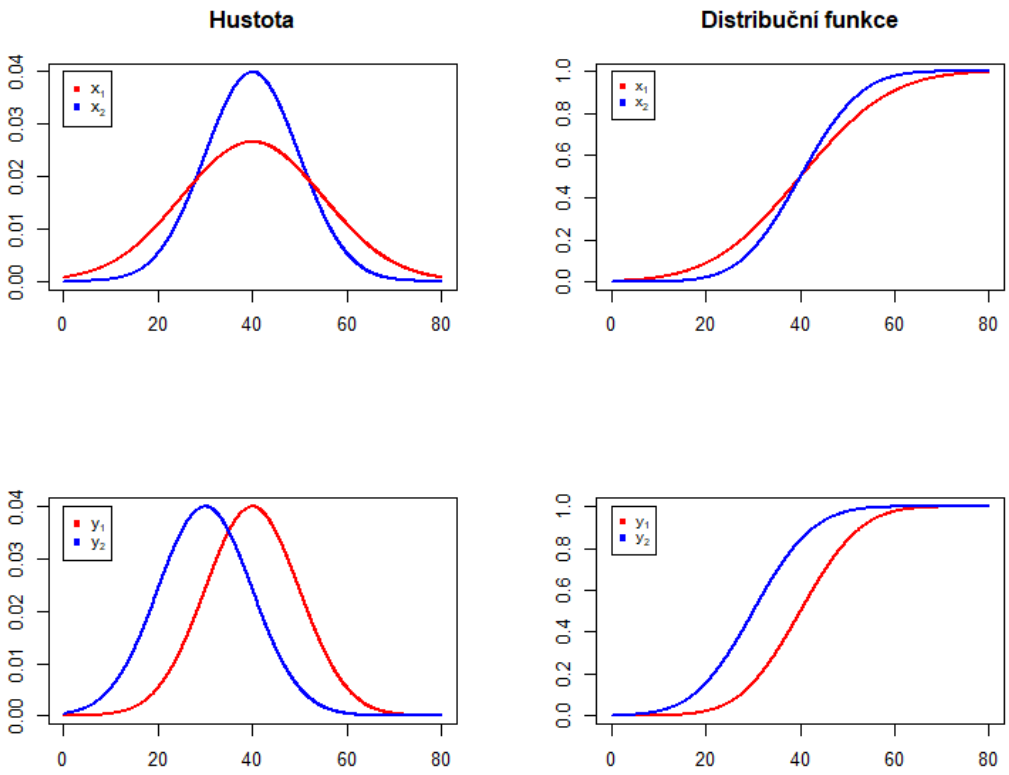
Dále, $\mathbf{y}_1$ je vektor 5 hodnot odpovídající středům intervalů v histogramu z dat náhodně generovaných z N(40, 10) a $\mathbf{w}_1$ je vektor vah k němu příslušný, $\mathbf{y}_2$ je vektor 5 hodnot odpovídající středům intervalů v histogramu z dat náhodně generovaných z N(30, 10) a $\mathbf{w}_2$ je vektor vah k němu příslušný. Tato normální rozdělení mají odlišnou střední hodnotu a stejné rozptyly.

Pro vektory $\mathbf{x}_1$ a $\mathbf{x}_2$ je vzdálenost rovna 12,3 a pro vektory $\mathbf{y}_1$ a $\mathbf{y}_2$ je to 17,6. Wassersteinova vzdálenost počítá rozdíl mezi distribučními funkcemi, proto na ni má větší vliv rozdílná střední hodnota, která distribuční funkce posune, na rozdíl od odlišného rozptylu, ten distribuční funkce pouze „klopí“, viz obrázek 3.3.

3.3. Aitchisonova vzdálenost

Aitchisonova vzdálenost je třetí a zároveň poslední, kterou jsem využila při tvorbě matic vzdáleností. Od předchozích dvou se liší svým běžným využitím na kompozičních datech.

Kompoziční data jsou definována jako vektory s kladnými složkami, kde zdro-



Obrázek 3.3: Hustota, distribuční funkce $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{y}_1$ a $\mathbf{y}_2$

jem relevantní informace jsou podíly mezi nimi. Hodnoty jednotlivých složek tak vždy zastupují části na nějakém celku. To znamená, že nesou pouze relativní informaci, ta může být reprezentována například procentuálním podílem (konstantní součet je roven 100) nebo proporcionálními částmi celku (konstantní součet je roven 1).

Na data s takovými vlastnostmi nelze použít standardní statistické postupy, protože vyžadují při statistickém zpracování odlišný přístup než standardní mnohorozměrná data s informací absolutní. Pokud bychom použili běžné metody, mohli bychom znehodnotit výsledky nebo dojít ke špatné interpretaci.

Těmto datům tedy též odpovídá „relativní metrika“. Tu si můžeme dovolit použít, protože máme data zastoupena (relativní) frekvencí četností výskytu obyvatelstva po celé ČR. Hodnoty (váhy) věkových intervalů v každém regionu pro

každé pohlaví mají konstantní součet 1. Proto na ně můžeme aplikovat nástroje z analýzy kompozičních dat.

Místo euklidovské geometrie, která pracuje s absolutními hodnotami proměnných, se v případě kompozičních dat používá tzv. Aitchisonova geometrie na simplexu. Simplex je výběrovým prostorem reprezentací kompozičních dat a p-složkový simplex je definován jako

$$S^p = \left\{ \mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p_+ \mid x_i > 0,\ \sum_{i=1}^{p} x_i = k \right\}, \tag{3.4}$$

kde k je konstantní součet složek vektoru $\mathbf{x}$.

Nyní uvažujme dvě kompozice $\mathbf{x}$ a $\mathbf{y}$ ze simplexu S^p. Potom Aitchisonova vzdálenost je definována jako

$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2p} \sum_{i=1}^{p} \sum_{j=1}^{p} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}. \tag{3.5}$$

Příklad 3. Srovnejme aplikaci Aitchisonovy a euklidovské vzdálenosti. Uvažujme vektory: $\mathbf{x}_1 = (10, 70, 20),\ \mathbf{x}_2 = (15, 65, 20),\ \mathbf{y}_1 = (55, 25, 20)$ a $\mathbf{y}_2 = (60, 20, 20)$.

Intuitivně tušíme, že rozdíl mezi $\mathbf{x}_1$ a $\mathbf{x}_2$ není stejný jako rozdíl mezi $\mathbf{y}_1$ a $\mathbf{y}_2$. Euklidovská vzdálenost je mezi nimi stejná $(7,071)$, protože mezi prvními dvěma příslušnými složkami vektorů je rozdíl 5 jednotek, třetí složka je vždy stejná. Ale v prvním případě, mezi prvními složkami je proporcionální rozdíl dvě třetiny, zatímco v druhém to je jedenáct dvanáctin.

Tento relativní rozdíl se zdá být vhodněji popsán pomocí kompoziční variability. Aitchisonova vzdálenost aplikovaná na obě dvojice vektorů proto stejná není. Pro první rozdíl je rovna 0,365, pro druhý 0,226. Vektory $\mathbf{y}_1$ a $\mathbf{y}_2$ jsou si uvážením jejich relativního měřítka bližší.

Příklad 4. Vraťme se zpět k příkladu 1. Označíme-li P jako proměnnou s rozdělením relativních četností vektoru (1,1,1,1,1,2,3,4,5) (v grafu 3.2 reprezentováno červeným histogramem), Q_1 jako proměnnou s rozdělením relativních četností

vektoru (1,2,3,4,5,5,5,5,5) (levý modrý histogram), Q_2 jako proměnnou s rozdělením relativních četností vektoru (1,2,2,2,2,2,3,4,5) (pravý modrý histogram) a dopočítáme-li i Aitchisonovu vzdálenost, můžeme výsledky zapsat do tabulky:

	KL vzdálenost	**Wassersteinova vzdálenost**	**Aitchisonova vzdálenost**
vzdálenost mezi P a Q_1	0,7153	2,211	2,276
vzdálenost mezi P a Q_2	0,7153	0,667	2,276

Vidíme, že jediná Wassersteinova metrika bere v úvahu uspořádání realizací příslušných proměnných. Kullback-Leiblerova a Aitchisonova vzdálenost tuto skutečnost v potaz neberou.

Kapitola 4

Shluková analýza populačních pyramid

Cílem shlukové analýzy je najít v množině objektů její podmnožiny neboli shluky objektů takovým způsobem, aby si prvky v jednom shluku byly dostatečně podobné, ale zároveň si nebyly příliš podobné s prvky jiného shluku. Shlukovat můžeme objekty i proměnné.

Uplatnění nachází zejména tam, kde se množina objektů reálně rozpadá do tříd (objekty se seskupují do přirozených shluků, např. věk, pohlaví,...). Příklady použití shlukové analýzy:

- Marketing: pomáhá k vyvíjení cílených marketingových programů pro jednotlivé skupiny potencionálních nakupujících.
- Pojištění: identifikace skupin držitelů pojistných smluv s vysokými nároky na pojistné plnění.
- Plánování města: identifikace skupin domů podle jejich typu, hodnoty a zeměpisné polohy.

V této kapitole jsem čerpala ze zdrojů [2], [6], [7].

4.1. Shluková analýza

Uvažujme matici $\mathbf{E}$ typu $n \times p$ (n = počet objektů (pozorování), p = počet proměnných v tomto případě odpovídá věkovým skupinám mužů a žen). Dále

uvažujme různé rozklady $\mathbf{B}^{(k)} = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_k\}$ množiny n objektů do k shluků, kde $\mathbf{U}_i$ značí i-tý shluk. Tyto rozklady jsou množiny disjunktních podmnožin, které dohromady tvoří $\mathbf{E}$. To můžeme zapsat následovně:

$$\mathbf{U}_l \cap \mathbf{U}_m = \emptyset,\ \mathbf{U}_1 \cup \mathbf{U}_2 \cup \ldots \cup \mathbf{U}_k = \mathbf{E}.$$

Úlohu shlukové analýzy klasifikujeme trojím způsobem:

1. Úloha má na svém začátku počet shluků již zadaný.
2. Sami musíme určit vhodný počet shluků.
3. Hierarchické shlukování, kde každý následující rozklad je zjemněním předchozího. V této práci se budeme věnovat této možnosti.

4.2. Míry vzdálenosti a podobnosti

Po výběru proměnných, které uspokojivě charakterizují vlastnosti shlukovaných objektů a zjištění jejich hodnot rozhodujeme o způsobu hodnocení vzdáleností či podobností objektů.

Výsledkem výpočtu příslušných měr pro všechny páry objektů je matice $\mathbf{D}$, která se nazývá matice vzdáleností. Je typu $n \times n$ a na její diagonále se nachází nuly. Pokud jsou na diagonále 1, jde o matici podobností $\mathbf{A}$.

Úlohu můžeme klasifikovat podle podobnosti na podobnost dvou objektů, podobnost objektu a shluku, nebo podobnost dvou shluků. Při značení podobnosti používáme 0 pro maximální rozdílnost a 1 pro totožnost.

4.2.1. Měření vzdálenosti

Pro kvantitativní data $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n$ se pro sestavení matice $\mathbf{D}$ využívají tři základní vzdálenosti:

- **Manhattanská vzdálenost** (pojmenovaná podle pravoúhlého systému ulic na Manhattanu). Je definovaná vztahem

$$d_H(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \sum_{l=1}^{p} |x_{il} - x_{i'l}|.$$

- **Euklidovská vzdálenost** je definovaná vztahem

$$d_E(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{i'l})^2}.$$

- **Čebyševova vzdálenost** je definovaná vztahem

$$d_C(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \max_l |x_{il} - x_{i'l}|.$$

Pro měření podobnosti proměnných můžeme využít výběrového korelačního koeficientu. Výsledkem bude výběrová korelační matice typu $p \times p$ a na její diagonále budou 1. Ta se případně může upravit na matici nepodobností. Shlukování proměnných se využívá například k redukci dimenze a tím následnému zjednodušení analýzy např. v psychologii.

4.3. Hierarchické shlukování

Hierarchické shlukování je charakterizováno posloupností vnořených rozkladů $\mathbf{B}^{(k)}$, která na jedné straně začíná triviálním rozkladem, kdy každý shluk znamená jeden prvek (objekt) a končí jedním shlukem obsahujícím všechny prvky (objekty). Používá se pro rozpoznání datové struktury.

Hierarchické shlukování dělíme podle směru postupu na aglomerativní a divizivní. Aglomerativní metoda popisuje spojování nejpodobnějších objektů ze kterých se tvoří shluky. S těmi nadále pracuje jako se samostatnými objekty. Takto se postupuje až do chvíle, kdy zůstane pouze jeden velký shluk.

Divizivní metoda pracuje naopak. Vychází z jednoho shluku, který dělí nejčastěji na dvě části. Každou z nich považuje za samostatný shluk, který nadále dělí.

Nevýhodou hierarchického typu shlukování může být skutečnost, že se v každém kroku snaží dosáhnout nejlepšího řešení v daném čase a pozici, tudíž nebere ohled na další postup. Aglomerativní metoda nerozdělí již vzniklé shluky a divizivní metoda rozdělené shluky znovu nespojí, což by případně mohlo vlastnosti rozkladu vylepšit.

Grafickým znázorněním hierarchického shlukování je dendrogram (stromový graf). Uzly tohoto grafu představují jednotlivé shluky. Horizontální řezy představují rozklady v posloupnosti shluků. Vertikální směr představuje vzdálenosti mezi shluky (mezi jednotlivými rozklady), např. obrázek 5.1.

4.3.1. Aglomerativní a divizivní postup

V této podkapitole nastíníme postupy, které se při hierarchickém shlukování používají. Aglomerativní hierarchický postup vypadá následovně:

1. Vypočteme matici vzdáleností $\mathbf{D}$.
2. Proces začneme od rozkladu $\mathbf{B}^{(n)}$, tj. od n shluků, z nichž každý obsahuje 1 prvek.
3. Prohledáme $\mathbf{D}$ a najdeme dva shluky (i-tý a j-tý), jejichž vzdálenost d_{ij} je minimální.
4. Spojíme i-tý a j-tý shluk do g-tého shluku. V $\mathbf{D}$ nahradíme i-tý a j-tý řádek a sloupec řádkem a sloupcem pro nový shluk.
5. Zaznamenáme pořadí cyklu.
6. Znovu pokračujeme krokem 3, dokud proces neskončí spojením do $\mathbf{B}^{(1)}$.

Divizivní metoda, na rozdíl od aglomerativní, vytváří systém rozkladů postupným rozdělováním shluků. Při aplikaci divizivního přístupu postupujeme tak, že postupně rozdělujeme každý z existujících shluků na dva nové, aby výsledný rozklad tohoto shluku byl optimální vzhledem k nějakému kritériu.

4.3.2. Přehled vybraných aglomerativních postupů

Vzdálenosti mezi shluky odvozujeme ze vzdáleností mezi objekty daných shluků.

- **Metoda nejbližšího souseda** (Simple linkage)
  Tato metoda vytváří shluky z jednotlivých objektů (shluků), které mají

mezi sebou nejkratší vzdálenost v porovnání s ostatními objekty (shluky). Nevýhodou je, že při existenci objektů s totožnou vzdáleností od existujících shluků může dojít k řetězení. Další nevýhodou je, že i značně vzdálené objekty se mohou sejít ve stejném shluku. Může tedy dojít k chybnému závěru.

- **Metoda nejvzdálenějšího souseda** (Complete linkage)
  Jak už název napovídá, tato metoda shlukuje objekty (shluky), které jsou v porovnání od ostatních od sebe nejdále. Vezme tu největší ze vzdáleností každých dvou objektů ze dvou různých shluků. Z takto vypočítané vzdálenosti bere v rámci dané úrovně shlukování tu nejmenší a shlukuje příslušné dva objekty (shluky). Tato metoda má tendenci tvořit nepříliš velké kompaktní shluky.

- **Metoda průměrné vazby** (Average linkage)
  Metoda průměrné vazby je podobná metodě nejbližšího souseda s tím rozdílem, že vzdálenost mezi shluky se vypočítá jako průměr vzdáleností mezi každými dvěma objekty z dvou různých shluků. Z takto vypočítaných vzdáleností bere tu nejkratší průměrnou. Protože průměry vzdáleností mezi objekty bývají často unikátní, k nejednoznačným výsledkům dochází méně často.

- **Wardova metoda**
  Wardova metoda vychází z analýzy rozptylu. Vzdálenosti objektů se měří čtvercovou euklidovskou vzdáleností a slučují se takové shluky, kde je minimální součet čtverců. Je velmi používaná, protože má tendenci vytvářet kompaktní, poměrně malé shluky zhruba stejné velikosti.

V této práci provedeme metody nejvzdálenějšího souseda a Wardovy metody, které jsou v praxi asi nejužívanější.

Kapitola 5

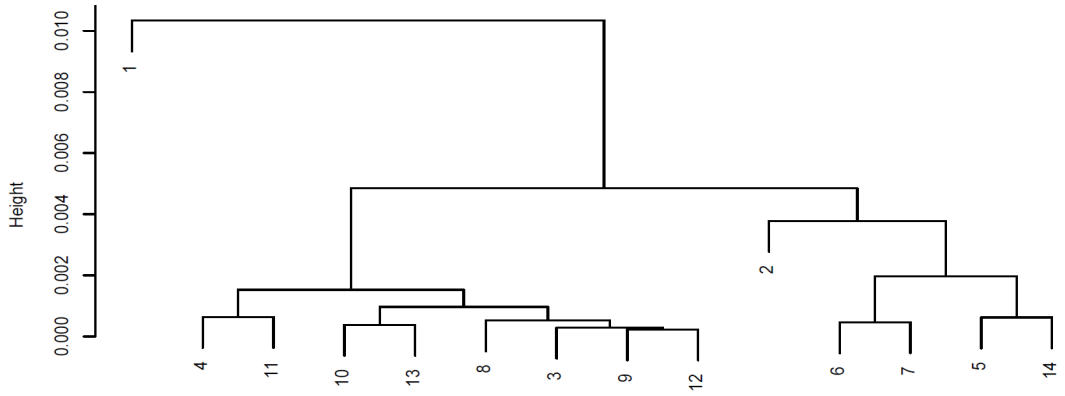
Aplikace postupu na demografická data

V této kapitole aplikujeme nastíněný postup na demografická data. Podoba dat je popsána v kapitole 1.4. Data již byla podle nástrojů symbolické analýzy upravena do stavu, ve kterém jsme je získali. Tyto úpravy jsme popsali v kapitole 2.1; výsledná data jsou k dispozici v přílohách A a B.

Pro takto upravený datový soubor jsme pro zjištění vzdáleností mezi symbolickými objekty použili dvoufázový postup. V prvním kroku jsme pro každé pohlaví napočítali matice vzdáleností mezi relativními četnostmi výskytu objektů v daných věkových skupinách pomocí symetrické Kullback-Leiblerovy, Aitchisonovy a Wassersteinovy vzdálenosti. Tímto způsobem můžeme zkoumat podobnost mezi jednotlivými jednotkami vzhledem k modální proměnné. Vznikají nám dvě matice vzdáleností $\mathbf{D}^M = (d_{ij}^M)_{i,j=1}^n$ a $\mathbf{D}^F = (d_{ij}^F)_{i,j=1}^n$ pro každou metriku.

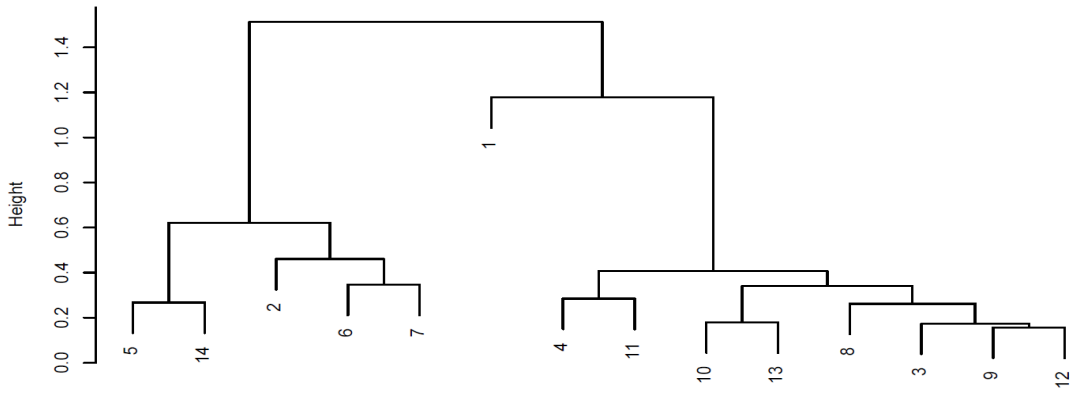
V druhém kroku zjistíme celkovou podobnost mezi pohlavími pomocí euklidovské vzdálenosti, tedy vztahem $d_{ij} = \sqrt{(d_{ij}^M)^2 + (d_{ij}^F)^2}$, $i, j = 1, \ldots, n$. Čím nižší hodnota d_{ij}, tím větší podobnost daného páru jednotek.

Následně jsme pomocí hierarchického shlukování vytvořili dendrogramy popisující strukturu dat pro jednotlivé vzdálenosti. Pro zajímavost jsme, jak již bylo uvedeno, použili dvě metody shlukování - Wardovu metodu a metodu nejvzdálenějšího souseda. Obě metody jsou popsány v kapitole 4.3.2.

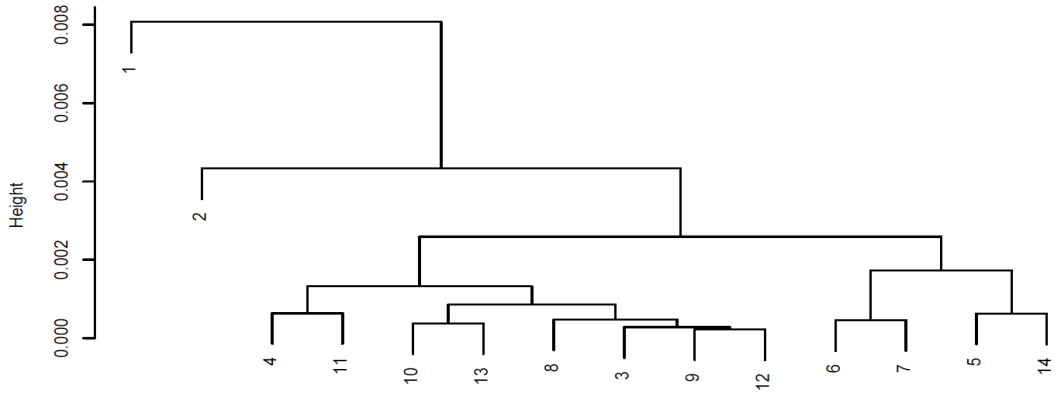


Obrázek 5.1: Dendrogram pomocí Kullback-Leiblerovy vzdálenosti, Wardova metoda

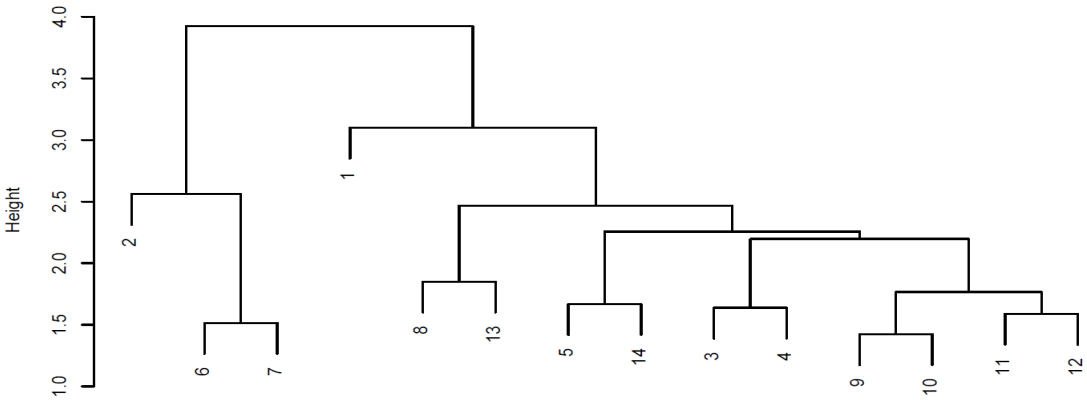
Obrázek 5.2: Dendrogram pomocí Wassersteinovy vzdálenosti, Wardova metoda



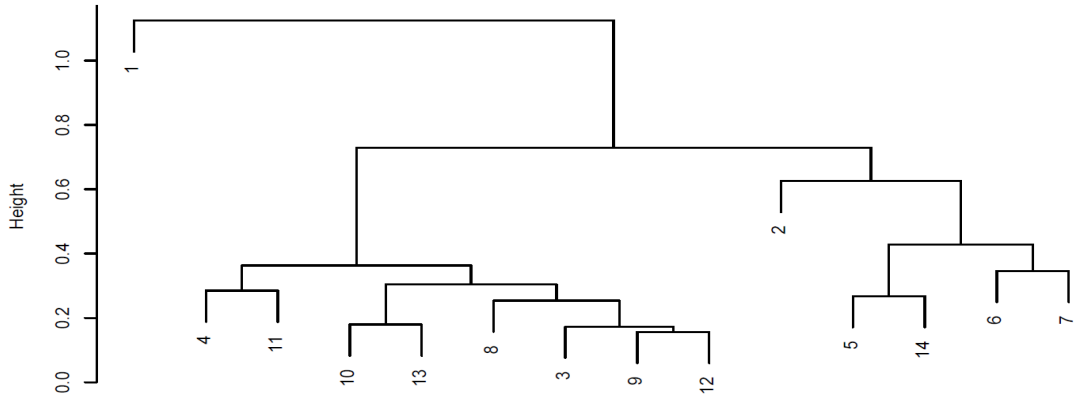
Obrázek 5.3: Dendrogram pomocí Aitchisonovy vzdálenosti, Wardova metoda



Obrázek 5.4: Dendrogram pomocí Kullback-Leiblerovy vzdálenosti, complete linkage



Obrázek 5.5: Dendrogram pomocí Wassersteinovy vzdálenosti, complete linkage



Obrázek 5.6: Dendrogram pomocí Aitchisonovy vzdálenosti, complete linkage

Tabulka 5.1: Přiřazení číselného označení jednotlivých krajů.

Značení	**Název kraje**
1	Hlavní město Praha
2	Středočeský kraj
3	Jihočeský kraj
4	Plzeňský kraj
5	Karlovarský kraj
6	Ústecký kraj
7	Liberecký kraj
8	Královehradecký kraj
9	Pardubický kraj
10	kraj Vysočina
11	Jihomoravský kraj
12	Olomoucký kraj
13	Moravskoslezský kraj
14	Zlínský kraj

Dendrogramy vzniklé pomocí Kullback-Leiblerovy vzdálenosti (obrázky 5.1, 5.4) napovídají, že pokud bychom graf řezali např. na hodnotě 0,002, kde vidíme stále ještě velké rozdíly mezi shluky, potom by se kraje přirozeně dělily do čtyř shluků. Dva jednoprvkové shluky by byly tvořeny krajem Středočeským a Prahou. Třetí shluk by byl tvořen krajem Karlovarským, Ústeckým, Libereckým a Moravskoslezským. Poslední shluk je tvořen zbylými kraji. Jediný rozdíl mezi těmito dvěma grafy je ve Středočeském kraji. Wardova metoda by ho dříve přiřadila k třetímu shluku, zatímco metoda nejvzdálenějšího souseda (MNS) by dříve spojila třetí a čtvrtý shluk.

Dendrogramy vzniklé pomocí Wassersteinovy vzdálenosti (obrázky 5.2, 5.5) se z trojice použitých vzdáleností na první pohled liší nejvíce. Praha, která v předchozích grafech nešla přehlédnout a shlukovala se až v posledním shluku se nyní shlukuje s početnou skupinou všech krajů kromě Středočeského, Ústeckého a Libereckého. Tyto tři tvoří samostatný shluk a přidávají se až v posledním kroku analýzy. Intuitivně bychom graf řezali přibližně v hodnotě 2,9. Vzniknou nám tak tři shluky. Tím prvním je již zmíněná trojice - Středočeský, Ústecký

a Liberecký kraj. Druhým je Praha a třetím jsou zbylé kraje. Rozdíl mezi grafy pro Wardovu metodu a MNS je nepatrný.

Dendrogramy vzniklé pomocí Aitchisonovy vzdálenosti jsou rozdílné. Ten, který vznikl Wardovou metodou (obrázek 5.3) je podobný oběma grafům vzniklým pomocí Wassersteinovy vzdálenosti. Přirozeně se rozpadá do tří shluků. Samozřejmě Praha jako samostatný shluk. Ke Středočeskému, Ústeckému a Libereckému kraji se připojily kraje Karlovarský a Moravskoslezský (tak jsme to viděli už u Kullback-Leiblerovy vzdálenosti, viz obrázky 5.1, 5.4). Třetí shluk je tvořen zbytkem republiky. Dendrogram vzniklý MNS (obrázek 5.6) je v zásadě totožný s dendrogramem Kullback-Leiblerovy vzdálenosti na obrázku 5.1, liší se pouze měřítkem.

Obecným závěrem může být dělení krajů do 3 shluků. Prvním je Praha jako jednoprvkový shluk. Druhý obsahuje kraje Středočeský, Ústecký a Liberecký. Třetí je tvořen zbytkem krajů. Pozornost si zaslouží dvojice krajů Karlovarský a Moravskoslezský. Mají tendenci řadit se do druhého i třetího shluku dle použité vzdálenosti a metody shlukování.

Závěr

Cílem bakalářské práce bylo analyzovat relativní věkovou strukturu populace vyjádřenou pomocí populačních pyramid ve smyslu symbolické analýzy dat a porovnat metriky vhodné pro takto zpracovaná data – Kulllback-Leiblerovu a Wassersteinovu. K tomu jsem připojila i Aitchisonovu vzdálenost, která se běžně využívá na datech relativní povahy, tzv. kompozičních datech. Výsledkem bylo šest dendrogramů prezentujících přirozené shlukování krajů na základě frekvence výskytu žen a mužů v jednotlivých věkových kategoriích. Dendrogramy jsem porovnala a interpretovala. Česká republika se dle toho přirozeně rozpadá na tři shluky.

Práce pro mě byla velmi přínosná. Zjistila jsem, že i data z populačních pyramid (které většina lidí zná pouze jako typ grafu) se dají po určité úpravě statisticky zpracovat. Dále, že existuje mnoho jiných metrik kromě těch základních, běžně vyučovaných a je zajímavé, ale nesnadné, je mezi sebou porovnávat. V neposlední řadě jsem se naučila sázet text a matematiku v typografickém systému LaTeX.

Příloha A

Relativní četnosti výskytu žen dle regionů a věku v ČR k 1. lednu 2015.

Věk	**Praha**	**Středočeský kraj**	**Jihočeský kraj**	**Plzeňský kraj**	**Karlovarský kraj**	**Ústecký kraj**	**Liberecký kraj**
$(0, 5)$	0,0293	0,0306	0,0266	0,0257	0,025	0,0266	0,0274
$[5, 10)$	0,0264	0,0314	0,0279	0,0276	0,0281	0,0292	0,0286
$[10, 15)$	0,0187	0,0243	0,0235	0,0223	0,0234	0,0249	0,0243
$[15, 20)$	0,0179	0,0222	0,0233	0,0217	0,0234	0,0237	0,0237
$[20, 25)$	0,0258	0,0284	0,031	0,0293	0,0311	0,0316	0,0309
$[25, 30)$	0,0355	0,0314	0,0328	0,0338	0,0345	0,0343	0,0334
$[30, 35)$	0,0427	0,0362	0,0349	0,0372	0,0341	0,0356	0,0353
$[35, 40)$	0,0506	0,0475	0,0425	0,0441	0,0419	0,0439	0,0443
$[40, 45)$	0,0418	0,0436	0,0394	0,041	0,0412	0,0421	0,0415
$[45, 50)$	0,0326	0,0339	0,033	0,0339	0,0351	0,0341	0,0326
$[50, 55)$	0,0303	0,0309	0,0324	0,0323	0,0339	0,0322	0,0305
$[55, 60)$	0,028	0,0304	0,0344	0,033	0,0334	0,0317	0,0303
$[60, 65)$	0,0302	0,0324	0,035	0,0346	0,0354	0,0342	0,0344
$[65, 70)$	0,0281	0,0293	0,0307	0,0309	0,0303	0,0315	0,0317
$[70, 75)$	0,0203	0,0187	0,0199	0,0212	0,0193	0,0191	0,0193
$[75, 80)$	0,0112	0,0106	0,0118	0,0127	0,012	0,0106	0,0106
$[80, 85)$	0,0087	0,0074	0,0087	0,0086	0,0075	0,0066	0,0074
$[85, \infty)$	0,0065	0,0045	0,005	0,0051	0,004	0,0036	0,0045

Věk	Královehradecký kraj	Pardubický kraj	kraj Vysočina	Jihomoravský kraj	Olomoucký kraj	Zlínský kraj	Moravskoslezský kraj
$(0, 5)$	0,0259	0,0268	0,026	0,0276	0,0259	0,0249	0,0251
$[5, 10)$	0,0279	0,0277	0,0271	0,0271	0,0274	0,0267	0,0271
$[10, 15)$	0,0234	0,0239	0,0238	0,0221	0,0231	0,0232	0,0233
$[15, 20)$	0,0239	0,0244	0,0247	0,0218	0,0228	0,0232	0,0239
$[20, 25)$	0,0307	0,0319	0,033	0,0302	0,0314	0,0316	0,0325
$[25, 30)$	0,0328	0,034	0,0344	0,0342	0,0336	0,0341	0,0352
$[30, 35)$	0,0339	0,0356	0,0356	0,0384	0,0351	0,0354	0,0349
$[35, 40)$	0,0426	0,0442	0,0421	0,045	0,044	0,0426	0,0416
$[40, 45)$	0,0402	0,0402	0,0393	0,04	0,0398	0,04	0,0399
$[45, 50)$	0,0329	0,0324	0,0337	0,0335	0,0332	0,0337	0,0353
$[50, 55)$	0,0310	0,0313	0,0336	0,0318	0,0322	0,0338	0,0347
$[55, 60)$	0,0324	0,0324	0,0345	0,032	0,0326	0,0338	0,0328
$[60, 65)$	0,0347	0,0339	0,0326	0,0319	0,0335	0,033	0,0329
$[65, 70)$	0,0311	0,0294	0,0292	0,028	0,0288	0,028	0,0273
$[70, 75)$	0,0207	0,0197	0,02	0,0205	0,0199	0,0196	0,0203
$[75, 80)$	0,0124	0,0121	0,0129	0,0119	0,0121	0,0124	0,0119
$[80, 85)$	0,0093	0,0089	0,0089	0,0082	0,0082	0,0084	0,0071
$[85, \infty)$	0,0057	0,0054	0,0053	0,0054	0,0052	0,005	0,0041

Příloha B

Relativní četnosti výskytu mužů dle regionů a věku v ČR k 1. lednu 2015.

Věk	**Praha**	**Středočeský kraj**	**Jihočeský kraj**	**Plzeňský kraj**	**Karlovarský kraj**	**Ústecký kraj**	**Liberecký kraj**
$(0, 5)$	0,0278	0,0291	0,0251	0,025	0,0236	0,025	0,0259
$[5, 10)$	0,0252	0,0297	0,0266	0,0262	0,0265	0,0279	0,0277
$[10, 15)$	0,0176	0,0228	0,0223	0,0215	0,022	0,0235	0,0228
$[15, 20)$	0,0174	0,0212	0,0219	0,0205	0,0222	0,0229	0,0226
$[20, 25)$	0,0264	0,0272	0,0293	0,0278	0,0292	0,0294	0,0292
$[25, 30)$	0,0353	0,0304	0,0315	0,0322	0,0321	0,0313	0,0318
$[30, 35)$	0,042	0,0358	0,0331	0,0342	0,0316	0,0323	0,0335
$[35, 40)$	0,0484	0,0466	0,041	0,0417	0,0405	0,0407	0,0419
$[40, 45)$	0,0393	0,0405	0,0385	0,0384	0,0391	0,0396	0,0397
$[45, 50)$	0,0318	0,0317	0,0321	0,0316	0,0335	0,0322	0,0316
$[50, 55)$	0,0308	0,0295	0,0318	0,0313	0,0324	0,0305	0,0305
$[55, 60)$	0,0307	0,0307	0,0341	0,0334	0,0342	0,0322	0,0315
$[60, 65)$	0,0349	0,0346	0,0361	0,0361	0,0378	0,0376	0,038
$[65, 70)$	0,0352	0,033	0,0341	0,0347	0,0353	0,0363	0,0365
$[70, 75)$	0,0265	0,0235	0,0251	0,026	0,0254	0,0247	0,0247
$[75, 80)$	0,0165	0,0159	0,0181	0,018	0,0183	0,0162	0,0162
$[80, 85)$	0,015	0,0133	0,0147	0,0149	0,0124	0,0121	0,0132
$[85, \infty)$	0,0145	0,0109	0,0118	0,0114	0,0101	0,01	0,012

Věk	**Královehradecký kraj**	**Pardubický kraj**	**kraj Vysočina**	**Jihomoravský kraj**	**Olomoucký kraj**	**Zlínský kraj**	**Moravskoslezský kraj**
$(0, 5)$	0,0246	0,0256	0,0244	0,0263	0,0247	0,0235	0,024
$[5, 10)$	0,0265	0,0263	0,0255	0,026	0,0264	0,0252	0,0259
$[10, 15)$	0,0219	0,0221	0,0224	0,0209	0,0219	0,0217	0,022
$[15, 20)$	0,022	0,0227	0,0233	0,0209	0,0219	0,0222	0,0228
$[20, 25)$	0,0292	0,0297	0,0311	0,0289	0,03	0,0301	0,0309
$[25, 30)$	0,0308	0,032	0,0317	0,0327	0,0323	0,0318	0,0331
$[30, 35)$	0,0314	0,0323	0,0319	0,0362	0,0331	0,0324	0,0318
$[35, 40)$	0,0401	0,0409	0,0388	0,0425	0,041	0,0402	0,0389
$[40, 45)$	0,0382	0,038	0,0373	0,0379	0,038	0,0376	0,038
$[45, 50)$	0,0316	0,0314	0,0319	0,0316	0,0319	0,0323	0,0337
$[50, 55)$	0,0306	0,0307	0,032	0,0309	0,0316	0,0333	0,0336
$[55, 60)$	0,0329	0,0329	0,0332	0,0329	0,034	0,034	0,0343
$[60, 65)$	0,0377	0,0364	0,034	0,035	0,0364	0,0366	0,0366
$[65, 70)$	0,0363	0,0336	0,0331	0,0338	0,0349	0,0336	0,0339
$[70, 75)$	0,0272	0,0251	0,0254	0,0265	0,0265	0,0265	0,0279
$[75, 80)$	0,0183	0,0183	0,0193	0,0183	0,0187	0,0202	0,0177
$[80, 85)$	0,0156	0,0156	0,0156	0,0152	0,0152	0,0166	0,0136
$[85, \infty)$	0,0134	0,0122	0,0126	0,0139	0,0129	0,013	0,0116

Literatura

[1] Andrášiková, A.: *Vícerozměrná analýza symbolických dat* [online]. 2017, [cit. 2018-11-16]. dostupné z: https://theses.cz/id/ybu8a0.

[2] Bivand, R. S., Wilk, J., Kossowski, T. : *Spatial association of population pyramids across Europe: The application of symbolic data, cluster analysis and join-count tests* [online]. 2017, [cit. 2018-09-23]. Dostupné z: https://doi.org/10.1016/j.spasta.2017.03.003.

[3] Comas-Cufí M.: *Package 'coda.base'* [online]. 2019, [cit. 2019-03-03]. Dostupné z: https://cran.r-project.org/web/packages/coda.base/index.html.

[4] Dudek A.: *Package 'SymbolicDA'* [online]. 2018, [cit. 2019-03-03]. Dostupné z: https://cran.r-project.org/web/packages/symbolicDA/index.html.

[5] Filzmoser, P., Hron K., Templ M.: *Applied compositional data analysis: With worked examples in R.* Springer, Cham, 2018. ISBN 9783319964201.

[6] Hebák, P. : *Vícerozměrné statistické metody 3.* Informatorium, Praha, 2005. ISBN 80-7333-039-3.

[7] Kelbel, J., Šilhán, D.: *Shluková analýza* [online]. [cit. 2018-11-23]. Dostupné z: http://cmp.felk.cvut.cz/cmp/courses/recognition/zapis-prednasky/zapis-02/13/shlukovani.pdf.

[8] Kolouri S., Rohde K. G.: *Optimal Transport: A Crash Course* [online]. 2008, [cit. 2019-1-31].Obrázek ve formátu PNG. Dostupné z: http://faculty.virginia.edu/rohde/transport/OTCrashCourse.pdf.

[9] Korenjak-Černe, S., Kejžar, N., Batagelj, V: *A weighted clustering of population pyramids for the world's countries, 1996, 2001, 2006* [online]. 2015, [cit. 2018-11-11]. Dostupné z: https://doi.org/10.1080/00324728.2014.954597.

[10] Koschin, F.: *Demografie poprvé.* Vyd. 2., přeprac. Oeconomica, Praha, 2005. ISBN 80-245-0859-1.

[11] Kracík, J.: *Kullback-Leiblerova divergence* [online]. [cit. 2018-11-16]. Dostupné z: https://homel.vsb.cz/ kra0220/sta3/KLD.pdf.

[12] Pawlowsky-Glahn V., Egozcue J. J, Tolosana-Delgado R.: *Modeling and analysis of compositional data*. John Wiley & Sons, Chichester, West Sussex, 2015. ISBN 1118443063.

[13] Prskawetz, A., Lindh, T.: *The Relationship Between Demographic Change and Economic Growth in the EU* [online]. 2007, [cit. 2018-11-11]. Dostupné z: https://www.oeaw.ac.at/fileadmin/subsites/Institute/VID/PDF/Publications/Forschungsberichte/FB32.pdf.

[14] Roubíček, V.: *Úvod do demografie*. Codex Bohemia, Praha, 1997. ISBN 80-85963-43-4.

[15] Schuhmacher D.: *Package 'transport'* [online]. 2019, [cit. 2019-03-03]. Dostupné z: https://cran.r-project.org/web/packages/transport/index.html.

[16] Slovník cizích slov – Domovská stránka [online]. [cit. 2018-11-11]. Dostupné z: https://slovnik-cizich-slov.abz.cz.

[17] Sueur J.: *Package 'seewave'* [online]. 2018, [cit. 2019-03-03]. Dostupné z: https://cran.r-project.org/web/packages/seewave/index.html.

[18] *Typy věkových pyramid* [online]. 2008, [cit. 2018-08-08].Obrázek ve formátu PNG. Dostupné z: http://cs.wikipedia.org/wiki/Soubor:Typy vekovych pyramid.png.

[19] *What is the advantages of Wasserstein metric compared to Kullback-Leibler divergence?* In: Stackexchange [online]. 2018, [cit. 2019-02-07]. Dostupné z: https://stats.stackexchange.com.