UNIVERZITA PALACKÉHO V OLOMOUCI
PRÍRODOVEDECKÁ FAKULTA

BAKALÁRSKA PRÁCA

Zhlukovanie pomocou nehierarchických metód

Katedra matematickej analýzy a aplikácií matematiky
Vedúci bakalárskej práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Paulína Jašková**
Študijný program: B1103 Aplikovaná matematika
Študijný obor: Aplikovaná štatistika
Forma studia: prezenčná
Rok odovzdania: 2018

BIBLIOGRAFICKÁ IDENTIFIKÁCIA

Autor: Paulína Jašková

Názov práce: Zhlukovanie pomocou nehierarchických metód

Typ práce: Bakalárska práca

Pracovisko: Katedra matematickej analýzy a aplikácií matematiky

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

Rok obhajoby práce: 2018

Abstrakt: Demografický vývoj významne ovplyvňuje fungovanie spoločnosti, preto sa štúdiu demografických procesov venuje veľká pozornosť. Táto bakalárska práca sa zaoberá aplikovaním viacrozmernej štatistickej metódy, zhlukovej analýzy na demografické údaje týkajúce sa žien v Slovenskej republike. Úlohou bolo identifikovať podobnosti medzi analyzovanými demografickými údajmi a geografickým rozložením Slovenskej republiky. Predvolený počet zhlukov bol navrhnutý pomocou metódy hierarchického zhlukovania. V rámci samotného nehierarchického zhlukovania potom boli využité metódy K-priemerov, K-medoidov a tiež modelové zhlukovanie, ktoré predstavujú teoreticky najprepracovanejší prístup. Výstupy všetkých metód boli interpretované a vzájomne porovnané.

Kľúčové slová: štatistické metódy, zhluková analýza, demografické data

Počet strán: 56

Počet príloh: 0

Jazyk: slovenský

BIBLIOGRAPHICAL IDENTIFICATION

Author: Paulína Jašková

Title: Clustering using nonhierarchical methods

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Applications of Mathematics

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

The year of presentation: 2018

Abstract: Demographic developments significantly affects the operation of the company, therefore, the study of demographic processes devoted much attention. This Bachelor thesis deals with the application of multivariate statistical methods, cluster analysis, to demographic data relating to women in the Slovak Republic. The task was to identify the similarities between the analysed demographic data and geographic layout of the Slovak Republic, the default number of clusters has been proposed using hierarchical methods together. In the context of non-hierarchical clustering itself, methods like K-means, K-medoids and also a model-based clustering were used. The last one represent the most sophisticated approach in cluster analysis. The outputs of all the methods were interpreted and compared.

Key words: statistical methods, cluster analysis, demographic data

Number of pages: 56

Number of appendices: 0

Language: Slovak

Prehlásenie

Prehlasujem, že som bakalársku prácu spracovala samostatne pod vedením pána doc. RNDr. Karla Hrona, Ph.D a všetky použité zdroje som uviedla v zozname literatúry.

V Olomouci dňa ....

podpis

Obsah

Poďakovanie

Rada by som poďakovala doc. RNDr. Karlovi Hronovi, Ph.D. za všetok čas, rady a pripomienky, ktorými prispel k tvorbe tejto práce.

Úvod

Cieľom bakalárskej práce je analýza okresov Slovenskej republiky podľa niektorých demografických ukazovateľov a porovnanie troch metód nehierarchického zhlukovania. Chceme zistiť či sú okresy Slovenskej republiky nejakým spôsobom podobné, či nepodobné a na základe toho vytvoriť zhluky. K tomu je v práci využitá zhluková analýza, ktorá je jednou z techník viacrozmernej analýzy údajov. Zhluková analýza slúži k triedeniu objektov do zhlukov tak, aby si objekty zaradené do rovnakého zhluku boli čo najviac podobné a naopak čo najviac nepodobné s objektami v ostatných zhlukoch. V práci sme vychádzali predovšetkým z literatúry Rona Wehrensa „Chemometrics with R". V tejto práci su podrobne popísané postupy hierarchického a hlavne potom nehierarchického zhlukovania.

Pred aplikovaním samotnej zhlukovej analýzy sú údaje najskôr overené pomocou prieskumovej analýzy údajov. Cieľom analýzy bolo preskúmať štruktúru údajov, prítomnosť extrémnych pozorovaní, či údaje pochádzajú z normálneho rozdelenia alebo ako sú medzi sebou premenné korelované. V práci je následne analyzovaných celkom dvanásť premenných, demografických ukazovateľov, súvisiacich so ženskou populáciou v Slovenskej republike.

Následne je v teoretickej časti hierarchickou zhlukovou analýzou s aglomeratívnym prístupom odhadnutý vhodný počet zhlukov. Tento odhad je využitý pri nehierarchických zhlukovacích metódach, ktoré sú hlavným cieľom predloženej bakalárskej práce.

V tejto práci sú aplikované a porovnávané tri metódy nehierarchického zhlukovania. Ide o metódu K-priemerov, Fuzzy zhlukovania a K-medoidov. Rôzne metódy nehierarchického zhlukovania môžu viesť k rôznym výsledkom zhlukov

okresov. Z tohto dôvodu je vhodné výsledky zhlukovania porovnať. Tomu je venovaná záverečná časť práce. Spracovanie údajov v celej práci je pomocou štatistického softwaru R. V celej práci sme použili knižnice $cluster$, $NbClust$, $mclust$, $flexclust$, $factoextra$ a $tidyverse$ [10].

Kapitola 1

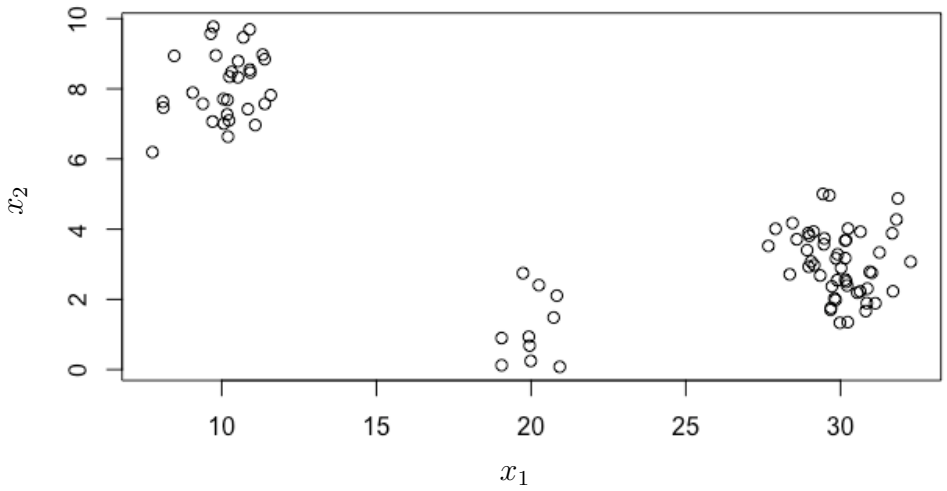
Úvod do zhlukovej analýzy

Prvá systematická zmienka o zhlukovej analýze sa spája s americkým psychológom R. C. Tryonom. V roku 1939 prvýkrát vykreslil metódy zhlukovej analýzy a neskôr pokračoval v ich rozvinutí. Zhluková analýza je v súčastnosti veľmi populárna metóda mnohorozmernej štatistiky.

Čo to vlastne zhluková analýza znamená? Zhluková analýza slúži na zaradenie objektov do skupín na základe podobnosti týchto sledovaných objektov. Znamená to, že objekty, ktoré sú podobnejšie, budú v jednej skupine a budú menej podobné s objektmi z ostatných skupín. Tieto skupiny sa v zhlukovej analýze nazývajú zhluky alebo klastre. Metóda zhlukovej analýzy sa používa pre viacrozmerné dátové súbory a má za úlohu z veľkého počtu údajov vybrať len potrebné informácie, s ktorými sa nám bude ľahšie ďalej pracovať. Využitie zhlukovej analýzy je veľmi široké. Uplatnenie možno nájsť napríklad v medicíne, ekonómii, biológii a ďalších vedeckých odboroch a praktických aplikáciach. Pri analýze demografických procesov sa možno tiež stretnúť s aplikáciou zhlukovej analýzy.

Pojem zhluk si lepšie vysvetlíme na nasledujúcom grafe. Pre pojem zhluk pritom neexistuje všeobecná definícia, preto výsledkom tejto metódy môže byť vždy rozdielna štruktúra zhlukov. Zhlukom môžeme rozumieť skupinu pozorovaní, ktoré sú v určitom zmysle blízke nejakému reprezentantovi danej skupine.

Na obr. 1.1 máme dátový súbor náhodne vygenerovaný z normálneho rozdelenia, ktorý je tvorený dvojrozmernými pozorovaniami (x_1, x_2).



Obr. 1.1: Zobrazenie zhlukov

Z grafu je zrejmé, že sa nám sledované údaje zoskupili do troch viditeľných zhlukov. Môžeme teda povedať, že objekty v zhluku sú si viac podobné medzi sebou ako s objektami z ostatných zhlukov.

Výber metódy na tvorbu zhlukov záleží aj na tom ako sa zhluky tvoria a ako počítame podobnosť objektov v zhluku. Pri zhlukovaní objektov sa najčastejšie používajú dve skupiny metód, heuristické metódy a metódy založené na modely.

Heuristické metódy sa ďalej delia na hierarchicé a nehierarchické metódy. Rozdelenie je podľa toho, akým systémom sú tvorené výsledné zhluky.

Pri zpracovaní kapitoly o zhlukovej analýze, hierarchickom aj nehierarchickom zhlukovaní boli využité predovšetkým zdroje [3],[6] a [11].

1.1. Analýza vstupných údajov

V tejto časti priblížime demografické údaje, ktoré budú v ďalších častiach analyzované metódami zhlukovej analýzy. Pomocou týchto údajov bude ozrejmená problematika zhlukovej analýzy v priebehu celej práce.

V Slovenskej republike je v súčasnosti podľa medzinárodnej klasifikácie štátnych územných celkov $NUTS$ 4 celkom 8 krajov, 79 okresov a 2890 obcí a miest.

Údaje sú demografické procesy v rôznych okresoch Slovenskej republiky z roku 2016. Okresy sú územné jednotky stredného stupňa. Vybraným rokom analýzy je rok 2016 nakoľko v dobe prípravy tejto práce neboli všetky údaje za rok 2017

publikované. V práci je analyzovaných celkom 12 ukazovateľov, ktoré pochádzajú z verejnej databázy Slovenského štatistického úradu [1]. Všetkých 79 okresov môžeme vidieť v následujúcej tabuľke:

Bratislava I	Nové Zámky	Rimavská Sobota
Bratislava II	Šaľa	Stará Ľubovňa
Bratislava III	Topoľčany	Stropkov
Bratislava IV	Zlaté Moravce	Svidník
Bratislava V	Tvrdošín	Vranov nad Topľou
Malacky	Žilina	Bardejov
Pezinok	Bytča	Humenné
Senec	Čadca	Kežmarok
Dunajská Streda	Dolný Kubín	Levoča
Galanta	Kysucké Nové Mesto	Medzilaborce
Hlohovec	Liptovský Mikuláš	Poprad
Piešťany	Martin	Prešov
Senica	Námestovo	Sabinov
Skalica	Ružomberok	Snina
Trnava	Turčianske Teplice	Spišská Nová Ves
Bánovce nad Bebravou	Veľký Krtíš	Trebišov
Ilava	Zvolen	Gelnica
Myjava	Žarnovica	Košice I
Nové Mesto nad Váhom	Žiar nad Hronom	Košice II
Partizánske	Banská Bystrica	Košice III
Považská Bystrica	Banská Štiavnica	Košice IV
Prievidza	Brezno	Košice - okolie
Púchov	Detva	Michalovce
Trenčín	Krupina	Rožňava
Komárno	Lučenec	Sobrance
Levice	Poltár	
Nitra	Revúca	

Podkladom pre zhlukovú analýzu je následne zdrojová matica typu $n \times p$. V p stĺpcoch zdrojovej matice sa nachádzajú premenné, demografické ukazovateľe. V n riadkoch sú objekty, okresy SR, na ktorých sú tieto premenné namerané. Prvky matice označujeme x_{ij}, ktoré predstavujú hodnoty j-tej premennej, $j = 1, 2, \ldots, p$, ktorá bola zistená u i-tého objektu, $i = 1, 2, \ldots, n$. V našom prípade $p = 11$ a $n = 79$.

2016 ženy	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Okres Bratislava I	33,77	86,50	10,43	33,24	34,36	82,77	13,99	16,60	1,70	43,60	36,35	27,17
Okres Bratislava II	33,16	84,93	12,12	31,96	32,71	79,02	10,01	13,40	1,40	41,40	29,46	25,06
Okres Bratislava III	33,65	81,38	13,12	32,16	32,69	80,48	12,76	13,50	1,50	41,40	38,12	25,79
Okres Bratislava IV	32,43	86,40	11,25	31,94	32,29	77,45	7,88	15,60	1,50	42,20	26,61	23,06
Okres Bratislava V	33,87	85,34	12,43	32,03	32,83	74,13	7,02	16,90	1,40	43,80	21,96	27,54
Okres Malacky	32,19	83,89	11,74	30,57	31,53	77,19	8,20	14,30	1,50	40,10	21,38	9,94
Okres Pezinok	31,61	84,66	11,43	31,35	32,44	77,91	8,75	14,00	1,50	41,00	22,95	10,83
Okres Senec	31,65	84,14	13,73	31,76	33,05	76,53	7,34	13,70	1,50	40,20	51,94	14,80
Okres Dunajská Streda	32,07	80,54	9,69	29,85	31,74	77,16	9,76	13,80	1,40	39,60	11,59	5,78
Okres Galanta	31,95	81,29	8,58	29,44	31,28	76,76	10,17	17,60	1,50	41,60	9,56	10,77
Okres Hlohovec	31,83	84,82	8,95	29,62	31,98	78,58	10,86	17,90	1,60	42,20	8,56	10,42
Okres Piešťany	31,08	86,07	8,43	30,64	31,53	79,80	10,03	17,10	1,50	42,30	9,42	9,66
Okres Senica	31,20	82,65	9,27	29,89	31,46	76,55	9,17	13,80	1,50	38,80	8,52	10,18
Okres Skalica	31,28	83,58	9,16	30,02	31,40	78,52	9,96	16,50	1,40	41,00	6,64	7,27
Okres Trnava	31,82	83,12	10,23	31,21	32,51	77,88	8,51	14,90	1,50	40,30	10,67	6,68
Okres Bánovce nad Bebravou	31,14	90,73	10,22	29,49	30,50	76,51	8,99	16,50	1,50	41,50	7,01	8,19
Okres Ilava	31,28	86,59	8,43	30,34	31,03	77,31	8,85	14,50	1,50	40,00	7,48	10,68
Okres Myjava	31,52	83,46	7,05	30,17	30,85	78,62	13,52	16,70	1,40	42,20	9,67	9,96
Okres Nové Mesto nad Váhom	31,72	81,82	9,53	29,73	31,67	79,11	11,00	14,90	1,50	40,30	8,59	8,33
Okres Partizánske	31,18	84,27	8,02	29,46	31,01	78,93	9,00	17,80	1,40	42,60	7,47	10,44
Okres Považská Bystrica	30,63	89,09	9,38	30,09	31,13	77,84	8,75	15,80	1,50	40,40	4,06	8,72
Okres Prievidza	32,21	81,53	8,86	29,72	31,89	77,02	8,99	15,70	1,60	40,80	4,32	6,84
Okres Púchov	29,72	89,21	8,26	30,05	30,65	78,97	10,82	16,70	1,60	40,80	6,54	7,24
Okres Trenčín	31,22	86,98	8,36	31,08	32,38	79,00	9,35	15,90	1,50	41,60	8,72	7,40
Okres Komárno	32,50	80,34	8,17	29,00	31,54	76,07	11,38	15,20	1,50	40,50	4,86	5,55
Okres Levice	31,93	80,75	8,42	29,08	31,13	78,19	10,79	16,20	1,50	40,60	6,14	8,44
Okres Nitra	31,59	84,82	9,81	30,20	31,52	78,02	9,98	15,90	1,50	40,20	8,44	7,83
Okres Nové Zámky	31,65	82,42	7,82	29,80	31,51	78,29	11,79	15,60	1,50	41,20	6,57	7,57
Okres Šaľa	31,99	82,23	8,66	29,42	30,70	75,43	8,84	14,30	1,50	38,80	7,95	9,96
Okres Topoľčany	31,45	83,20	7,89	30,22	31,34	78,97	9,60	13,00	1,60	38,80	7,70	10,15
Okres Zlaté Moravce	30,66	87,00	9,57	28,90	31,28	78,23	11,87	15,20	1,50	40,30	8,28	11,10
Okres Tvrdošín	27,88	93,91	11,44	29,95	30,63	76,74	5,16	16,80	1,90	38,90	4,05	8,61
Okres Žilina	30,36	88,79	10,28	30,56	31,40	77,49	9,40	16,00	1,40	41,10	6,83	6,33
Okres Bytča	27,97	93,68	9,84	29,34	30,30	78,06	10,61	15,60	1,70	39,00	10,16	8,29
Okres Čadca	29,17	89,61	9,18	29,29	30,60	76,88	8,86	15,60	1,50	39,50	2,58	5,16
Okres Dolný Kubín	30,13	88,26	10,36	29,95	32,14	75,85	7,66	14,60	1,70	39,30	5,60	9,31
Okres Kysucké Nové Mesto	29,90	92,23	10,51	29,74	32,51	75,63	9,79	15,80	1,50	40,00	6,69	7,70
Okres Liptovský Mikuláš	31,09	83,21	7,68	29,49	30,93	78,39	9,13	15,50	1,60	41,20	6,16	5,70
Okres Martin	32,10	83,94	9,50	30,87	31,23	78,10	8,50	15,40	1,40	41,10	6,35	7,05
Okres Námestovo	26,30	94,71	15,53	29,37	30,60	74,76	6,28	12,00	1,60	35,20	3,39	5,76
Okres Ružomberok	30,20	88,00	9,14	30,03	31,32	78,73	10,50	15,70	1,60	41,30	6,28	8,09
Okres Turčianske Teplice	31,28	88,16	7,05	29,00	34,14	79,12	12,40	17,30	1,20	42,90	10,09	11,30
Okres Veľký Krtíš	31,27	85,78	8,50	28,06	28,28	77,37	10,77	17,00	1,50	40,90	5,96	11,39
Okres Zvolen	31,86	85,58	8,83	30,38	31,88	78,13	9,73	15,80	1,50	41,60	10,20	11,01
Okres Žarnovica	30,47	87,58	8,13	29,87	31,19	79,72	10,29	17,20	1,50	40,40	9,40	11,26
Okres Žiar nad Hronom	32,14	85,58	7,57	29,13	30,44	78,74	10,94	16,40	1,60	41,80	7,24	9,99
Okres Banská Bystrica	32,84	84,22	9,07	30,87	32,12	77,80	8,37	16,80	1,60	42,80	8,56	8,31
Okres Banská Štiavnica	29,79	86,59	6,93	29,48	31,70	76,61	10,87	11,30	1,40	38,40	11,58	11,82
Okres Brezno	30,76	85,99	8,83	28,48	29,21	77,31	10,47	15,70	1,40	39,80	5,13	7,89
Okres Detva	30,28	89,93	8,17	29,40	30,86	76,87	9,07	15,90	1,60	39,60	8,53	10,15
Okres Krupina	29,44	91,89	10,13	28,68	30,12	79,29	11,78	16,90	1,70	39,60	8,75	13,77
Okres Lučenec	32,15	78,71	8,73	28,15	29,21	76,04	10,84	17,10	1,60	41,10	7,38	7,80
Okres Poltár	32,36	79,17	9,52	29,19	29,64	75,62	11,94	17,40	1,20	42,00	7,54	14,55
Okres Revúca	30,55	83,43	10,15	26,87	29,05	74,59	10,69	18,20	1,70	42,40	7,26	9,22
Okres Rimavská Sobota	29,13	85,88	10,47	26,51	29,40	76,71	9,75	16,20	1,80	39,70	5,12	6,94
Okres Stará Ľubovňa	27,43	94,93	13,44	27,50	29,74	75,51	6,46	13,70	1,60	37,60	2,82	7,10
Okres Stropkov	27,45	92,25	9,79	28,44	30,06	80,27	8,04	17,30	1,70	41,70	8,43	9,98
Okres Svidník	27,29	95,14	8,17	28,15	30,56	78,94	8,59	16,90	1,80	39,90	6,07	10,94
Okres Vranov nad Topľou	26,97	92,47	13,06	27,20	28,46	76,27	8,69	14,40	1,70	37,40	4,42	8,35
Okres Bardejov	26,54	95,70	10,43	28,11	30,09	79,08	8,95	15,60	2,00	37,90	4,04	7,82
Okres Humenné	29,32	88,80	7,48	28,83	31,30	77,60	7,58	16,40	1,60	40,10	4,69	10,80
Okres Kežmarok	26,60	95,29	15,93	27,19	29,35	73,60	6,61	16,40	1,90	39,60	6,31	7,16
Okres Levoča	27,78	90,28	11,20	27,95	29,77	75,44	6,97	14,60	1,50	38,50	8,93	11,02
Okres Medzilaborce	28,40	89,66	6,23	27,18	31,41	78,62	13,44	15,10	1,50	37,50	10,16	10,82
Okres Poprad	29,73	89,88	10,92	28,92	30,12	76,20	7,25	16,10	1,50	41,30	6,32	8,33
Okres Prešov	28,94	92,26	11,77	29,01	29,44	78,55	7,06	15,20	1,60	39,90	6,75	6,21
Okres Sabinov	26,87	95,36	15,63	26,86	30,83	76,68	7,29	17,00	1,80	40,30	4,19	6,96
Okres Snina	29,39	89,38	8,84	28,89	30,47	77,06	9,74	15,60	1,60	38,80	4,50	10,64
Okres Spišská Nová Ves	29,29	91,02	11,95	27,15	29,51	76,57	6,94	16,10	1,70	40,20	4,15	7,98
Okres Trebišov	29,56	89,80	10,92	27,04	29,07	75,95	9,74	14,90	1,50	38,40	5,55	8,62
Okres Gelnica	28,87	88,19	14,94	26,03	28,17	76,81	8,35	16,80	1,60	41,10	6,46	9,41
Okres Košice I	32,77	80,23	8,24	31,07	31,84	78,78	8,89	15,40	1,40	40,80	21,09	20,44
Okres Košice II	30,72	88,29	9,84	29,73	32,08	73,95	6,65	17,50	1,40	42,90	15,64	21,22
Okres Košice III	34,28	79,59	9,46	31,31	33,46	69,60	6,53	14,40	1,50	40,70	17,66	30,65
Okres Košice IV	32,59	83,13	9,02	31,41	32,21	76,81	10,59	16,00	1,60	42,60	23,35	19,70
Okres Košice - okolie	28,49	89,65	13,13	27,50	28,86	74,55	8,34	15,60	1,70	39,50	15,01	9,02
Okres Michalovce	29,77	88,79	9,88	27,86	29,85	75,43	8,82	15,30	1,40	39,50	6,26	8,20
Okres Rožňava	32,58	80,78	10,51	27,00	29,80	77,75	10,30	15,90	1,60	41,10	4,21	7,49
Okres Sobrance	27,63	92,25	9,24	27,44	30,17	77,90	11,15	15,90	1,40	39,60	7,49	11,68

Obr. 1.2: Ukážka demografických dat

Premenné sú:

x_1 – priemerný vek snúbencov (žien) pri sobáši,

x_2 – podieľ snúbencov (žien), ktoré vstupujú do manželstva ako slobodné,

x_3 – počet narodených žien na počet obyvateľov,

x_4 – priemerný vek ženy pri pôrode,

x_5 – priemerný vek ženy pri potrate,

x_6 – priemerný vek ženy pri úmrtí,

x_7 – počet zomrelých žien na počet obyvateľov,

x_8 – priemerná dĺžka trvania rozvedeného manželstva,

x_9 – priemerný počet maloletých detí v rozvedenom manželstve,

x_{10} – priemerný vek manželov pri rozvode,

x_{11} – počet prisťahovaných žien na počet obyvateľov,

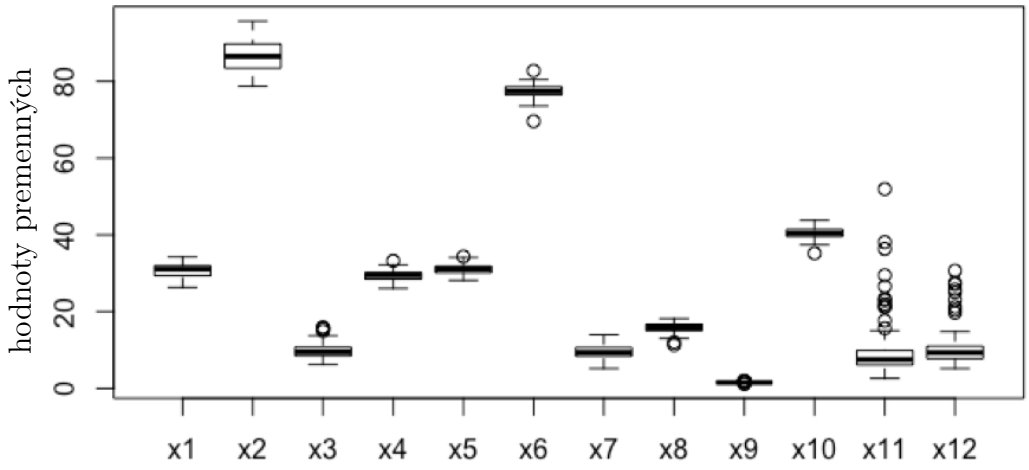
x_{12} – počet vysťahovaných žien na počet obyvateľov.

Pri použití softwaru R na vysvetlenie metód budeme tento dátový súbor označovať ako X.

1.1.1. Prieskumová analýza údajov

Pred aplikovaním samotnej zhlukovej analýzy je nutné zdrojové údaje podrobne preskúmať. K tomu sa využíva prieskumová (exploračná) analýza údajov. Jej cieľom je identifikácia „nedostatkov" v údajoch a následne nájdenie dôvodu ich existencie a prípadne potlačiť ich vplyv na ďalšie štatistické spracovanie. Prieskumovú analýzu údajov tvorí skupina štatistických techník využívajúcich grafické znázorňovacie metódy. Medzi základné prvky exploračnej analýzy údajov patrí vizualizácia údajov pomocou zhodných grafov. V práci je k tomuto účelu využívaný krabicový graf (boxplot). Pomocou grafu možno porovnať centrálne tendencie údajov, ich rozptýlenosť, zošikmenie aj prítomnosť odľahlých hodnôt. Pre zabezpečenie určitej homogenity analyzovaných údajov bola na začiatku uskutočnená ich štandartizácia. Normovanie sme zabezpečili klasickým spôsobom tak, že sme odčítali od hodnoty priemernú hodnotu príslušnej premennej a podelili smerodajnou odchýlkou. Na obr. 1.3 vidíme spomínané krabicové grafy

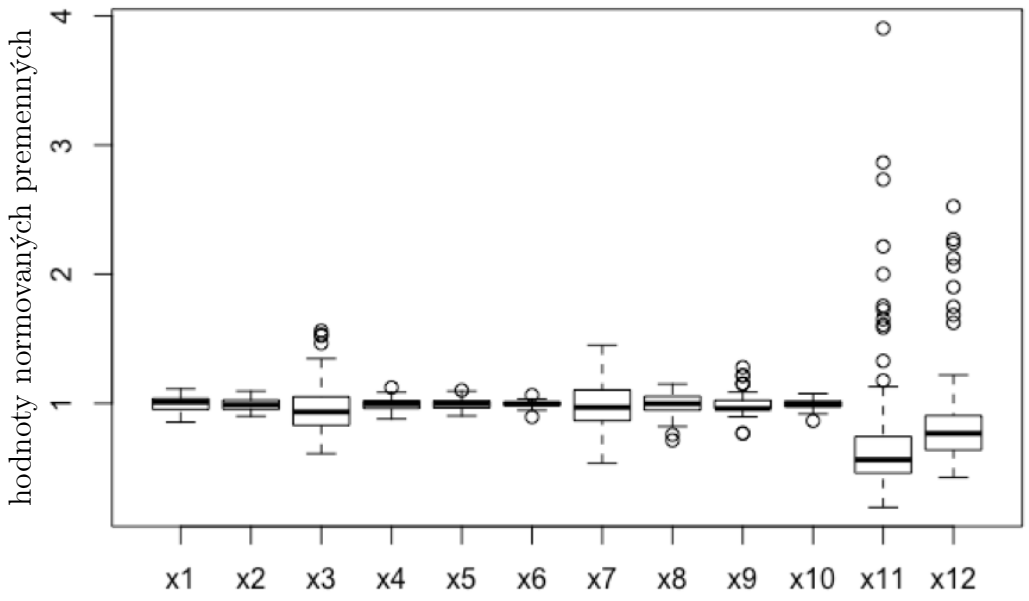
pre jednotlivé premenné pred normovaním.



Obr. 1.3: Boxploty pred normovaním

Z obr. 1.3 môžeme skonštatovať, že skoro každá premenná má niekoľko odľahlých hodnôt a taktiež mierky jednotlivých premenných sú veľmi rozdielne. Z tohto dôvodu sú pred samotným zhlukovaním údaje normalizované.

Na obr. 1.4 sú krabicové grafy pre premenné už po ich štandardizácií pomocou príkazu *scale*.



Obr. 1.4: Boxploty po normovaní

Na grafe je vidieť, že v premennej x_{11}, počet prisťahovaných žien na počet obyvateľov, je niekoľko vyšších odľahlých hodnôt. Tieto vysoké hodnoty sú v okresoch Bratislavy a okresoch Košíc. Ich existencia je logická, rozhodli sme sa ich neodstraňovať, nakoľko je sťahovanie obyvateľov do najväčších miest krajiny prirodzené. Tak isto vidíme na grafe, že aj v premennej x_{12}, počet vysťahovaných žien na počet obyvateľov, je niekoľko vyšších odľahlých hodnôt. Tieto vysoké hodnoty sú zistené tiež v okresoch Bratislavy a Košíc. Hodnoty sme sa tiež rohodli neodstraňovať, keďže je logické že pri väčších mestách bude väčšia vysťahovanosť na dediny a do menších miest, aj medzi aglomeráciami navzájom.

1.2. Miery podobností

Zhluková analýza zaraďuje objekty do zhlukov na základe podobností objektov. Zhluky, ktoré sú si najviac podobné, zaradí do rovnakého zhluku. Podobnosť je miera toho, nakoľko je možné považovať objekty za rovnaké. Často miery podobnosti predstavujú vzdialenosti medzi jednotlivými objektmi. Definovaných je mnoho mier podobností na základe ktorých rozhodujeme, do ktorého zhluku objekt patrí. Môžme ich rozdeliť do štyroch skupín:

1. miery vzdialeností,
2. asociačné koeficienty,
3. korelačné koeficienty,
4. pravdepodobnostné miery podobnosti.

Najčastejšie sa používajú miery podobností, založené práve na vzdialenosti. Čím menšia vzdialenosť medzi dvomi pozorovaniami, tým rastie tendencia zaradiť ich do rovnakého zhluku. Prvý krok pri zhlukovej analýze je vytvorenie matice podobnosti. Je to matica, kde v riadkoch a stĺpcoch sú zoradené objekty a vnútri sú miery podobnosti medzi každou jednou dvojicou objektov.

Pre intervalové veličiny sa najčastejšie používa euklidovská vzdialenosť, štvorcová euklidovská vzdialenosť, Hammingova vzdialenosť a mnoho ďalších. V prípade početností pracujeme hlavne s mierami chí-kvadrátu a fí-kvadrátu. Podrobnejšie si popíšeme euklidovskú metriku.

Euklidovská vzdialenosť medzi i-tým a j-tým pozorovaním, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ a $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})^T$ je definovaná vzťahom

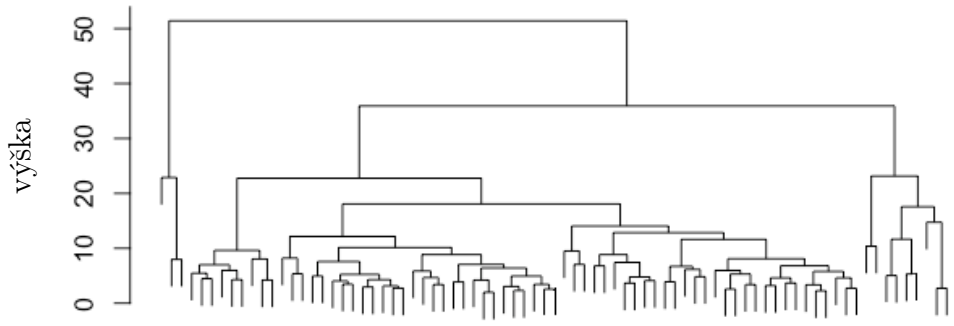
$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

kde x_{ik} je hodnota k-tej premennej pre i-tý objekt a x_{jk} je hodnota k-tej premennej pre j-tý objekt. Táto metrika je bežná v reálnom svete a vieme pomocou nej merať vzdialenosti v rovine alebo aj v priestore. Avšak musíme počítať s tým, že euklidovská vzdialenosť nie je invariantná na zmenu mierky, čo práve implikuje nutnosť predchádzajúceho normovania premenných.

Kapitola 2

Hierarchická zhluková analýza

Hierarchické zhlukovacie postupy sú založené na hierarchickom usporiadaní objektov a ich zhlukov. Hierarchické zhlukovacie metódy vychádzajú z jednotlivých objektov, kde každý objekt tvorí prvotný zhluk. Ich spájaním sa v každom kroku počet zhlukov postupne zmenšuje až sa nakoniec všetky zhluky spoja do jedného celku (postup môže byť aj opačný). Hierarchické metódy vedú k stromovej štruktúre, ktorá sa graficky zobrazuje ako dendrogram. Objekty sú v dendrograme radené tak (v našom prípade horizontálna os), aby bolo možné sledovať postupné spájanie objektov do zhlukov. Teda algoritmus hierarchického zhlukovania je ukončený až keď dostaneme jeden konečný veľký zhluk. Príklad dendrogramu reprezentujúceho dátový súbor analyzovaných demografických ukazovateľov je na obr. 2.1.



Obr. 2.1: Dendrogram

Na začiatku zhlukovania predstavujú všetky objekty (79 okresov) samostatné zhluky. Postupne možno sledovať vytváranie jednotlivých zoskupení. Na x-ovej osi sú pozorovania a na y-ovej osi sú vzdialenosti medzi pozorovaniami, respektívne medzi zhlukmi.

Na vyjadrenie hierarchickej štruktúry v softwari R sme využili príkaz $hierarch = hclust(d)$, kde d je matica vzdialeností vypočítaná pomocou príkazu $dist$. Na vykreslenie dendogramu sme následne použili funkciu $plot(hierarch)$.

Výhodou hierarchického zhlukovania je, že nepotrebujeme poznať výsledný počet zhlukov. Nevýhodou však je, že objekt ktorý sme už raz zaradili do zhluku, nemôže byť z neho vyradený a pridaný do iného zhluku. Táto metóda nie je vhodná pre veľmi veľké dátové súbory. Hierarchické zhlukovanie rozdeľujeme na aglomeratívne a divízne.

2.1. Aglomeratívne zhlukovanie

Aglomeratívne zhlukovanie je založené na princípe, že na začiatku každé pozorovanie tvorí jeden zhluk a ďalej sa po dvoch spájajú, pokiaľ nevytvoria až jeden veľký zhluk. Každý objekt je teda považovaný za jeden zhluk. Následne dva zhluky medzi ktorými je najmenšia vzdialenosť sa spoja do jedného nového zhluku. K tomuto novému zhluku sa buď pripojí ďalší zhluk, ktorý má od týchto dvoch najmenšiu vzdialenosť. Ak má vzdialenosť od iného zhluku menšiu, tak sa s ním spojí a tieto dva objekty vytvoria ďalší spoločný zhluk. Algoritmus sa opakuje, buď sa pridá nový zhluk k existujúcemu zhluku alebo sa vytvorí úplne nový zhluk alebo sa kombinujú spolu už existujúce zhluky, až kým dosiahneme požadovaný počet zhlukov. V poslednom kroku všetky objekty vytvoria jeden veľký zhluk. Spomenutý algoritmus aglomeratívneho zhlukovania možno popísať nasledujúcimi bodmi:

1. Vypočítame všetky párové (euklidovské) vzdialenosti a tieto párové vzdialenosti usporiadame do symetrickej matice $\mathbf{D} = (d_{ij})$.
2. Nájdeme najmenšiu vzdialenosť (napríklad medzi i-tym a j-tym pozorova-

ním) a vytvoríme zhluk, do ktorého dáme objekty s najmenšiou vzdialenosťou.

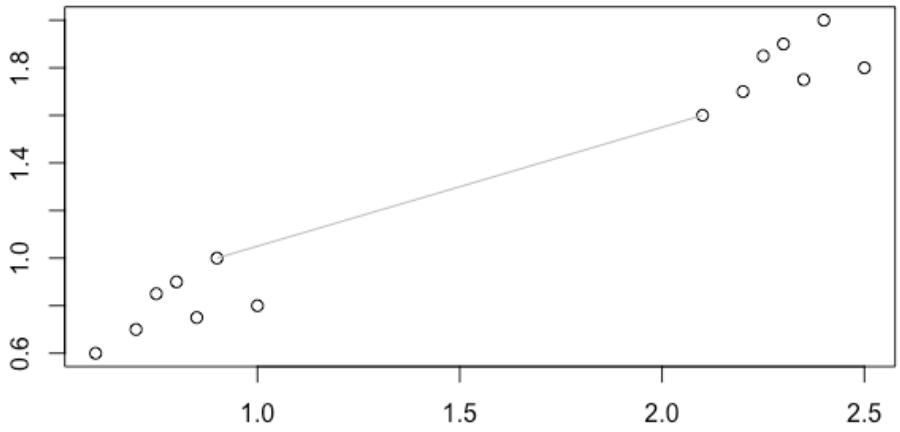
3. Vypočítame nové vzdialenosti medzi novým zhlukom a ostatnými pozorovaniami pomocou jednej z nižšie uvedených metód:
   - metóda najbližšieho suseda,
   - metóda najvzdialenejšieho suseda,
   - metóda priemernej väzby suseda,
   - Wardová metóda.
4. Zostrojíme novú maticu $\mathbf{D}_2$ o rozmeroch $(n-1)\times(n-1)$ vynechaním i-tého a j-tého riadku a stĺpca, a pridáme nový riadok a stĺpec ij s vypočítanými vzdialenostami $d_{ij,k}$ zhluku obsahujúceho i-té a j-té pozorovanie od k-tého pozorovania.
5. Krok 3 a 4 opakujeme $(n-1)$-krát až kým nedostaneme len jeden zhluk.

Teraz si bližšie vysvetlíme metódy použité v kroku 3. Každú s týchto metód aplikujeme na naše demografické údaje.

Single linkage/Metóda najbližšieho suseda - metóda najbližšieho suseda je jednoduchá metóda, ktorá spája zhluk A a B medzi ktorými je vzdialenosť určená ako vzdialenosť medzi dvoma najbližšími objektmi, to znamená že

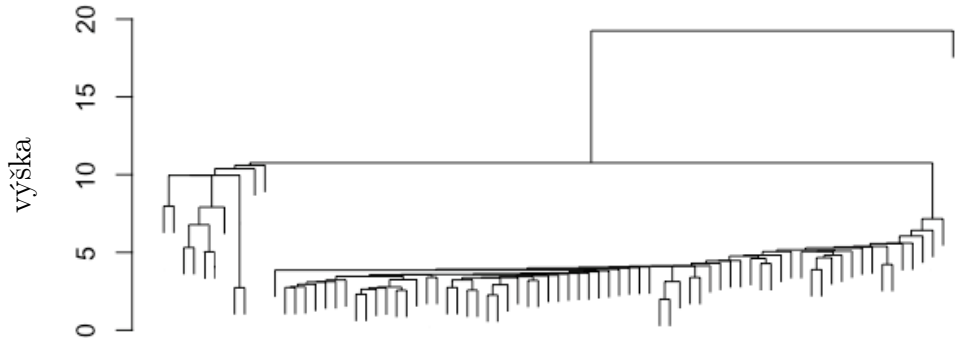
$$d(A,B) = \min d_{ij},$$

kde i-tý objekt patrí do zhluku A a j-tý objekt patrí do zhluku B. Na obr. 2.2 je ilustrovaný princíp tejto hierarchickej aglomeratívnej úlohy.



Obr. 2.2: Metóda najbližšieho suseda

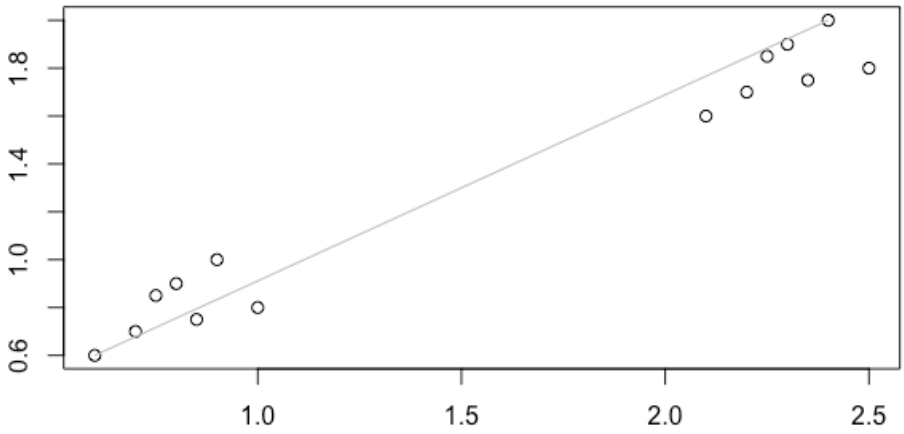
Ako je možné vidieť z výsledného dendrogramu tohto procesu zhlukovania, nie je vždy táto metóda najvhodnejšia a neprináša požadované praktické výsledky.



Obr. 2.3: Metóda najbližšieho suseda-dendrogram

Na obr. 2.3 môžeme vidieť nevhodnosť tejto metódy. Napríklad jeden z nevyhovujúcich faktorov je, že jedno pozorovanie sa s ostatnými zhluklo až v poslednom kroku. Je to spôsobené extrémnou hodnotou premennej x_{11} – počet prisťahovaných žien na počet obyvateľov pre okres Senec. Rozhodli sme sa túto hodnotu neodstrániť zo súboru, nakoľko je prirodzené, že bude počet prisťahovaných v okrese Senec extrémny. Je to z dôvodu, že tento okres je satelitným okresom pre hlavné mesto Bratislavu.

Complete linkage/Metóda najvzdialenejšieho suseda - metóda najvzdialenejšieho suseda je založená na presne opačnom princípe ako metóda najbližšieho suseda (obr. 2.4).

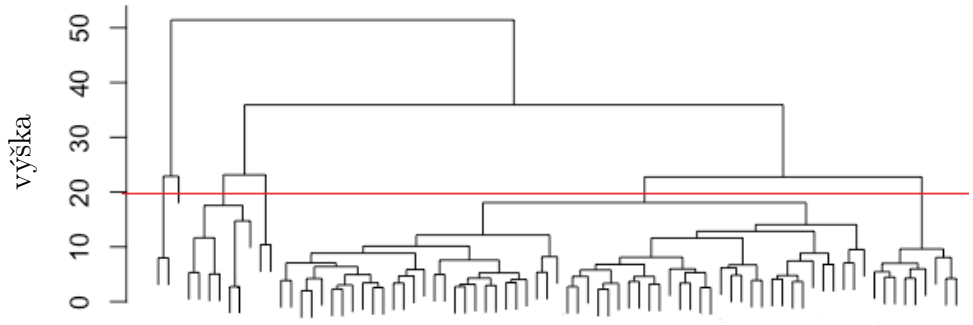


Obr. 2.4: Metóda najvzdialenejšieho suseda

V tomto prípade je vzdialenosť medzi dvoma zhlukmi A a B určená ako vzdialenosť medzi dvoma najvzdialenejšími objektami v jednotlivých zhlukoch. To znamená

$$d(A, B) = \max d_{ij}.$$

Výhoda tejto metódy je, že vytvára menej početné zhluky, ktoré vieme od seba dobre odlíšiť. V nasledujúcom dendrograme (obr. 2.5) je pomocou červenej čiary naznačený možný vhodný počet zhlukov z demografických údajov analyzovaných v bakalárskej práci.



Obr. 2.5: Metóda najvzdialenejšieho suseda-dendrogram

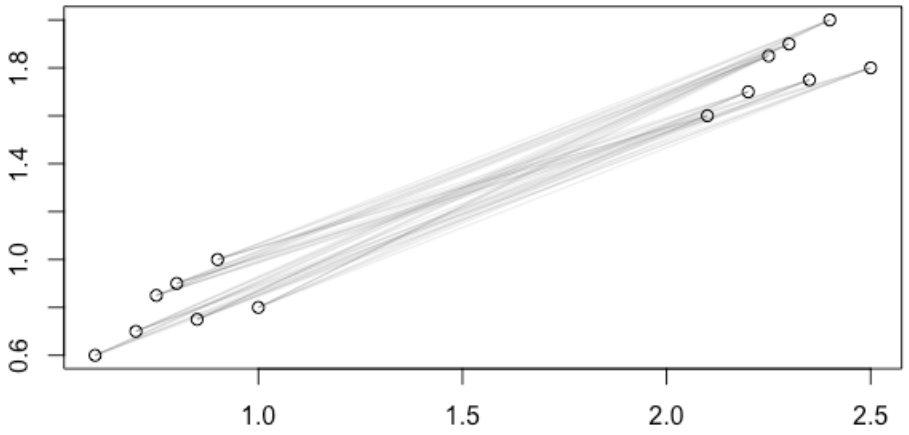
V dendrograme je zrejmé, že podľa tejto metódy by sme vybrali pre následujúce použitie šesť zhlukov. Sofistikovanejšie metódy sa pre tento účel využívajú len zriedka.

Average linkage/Metóda priemernej väzby suseda - kritérium vzniku zhluku je priemerná vzdialenosť všetkých objektov v jednom zhluku ku všetkým objektom v druhom zhluku,

$$d(A, B) = \frac{\sum_{\mathbf{x}_i \in A} \sum_{\mathbf{x}_j \in B} d_{ij}}{n_A \cdot n_B},$$

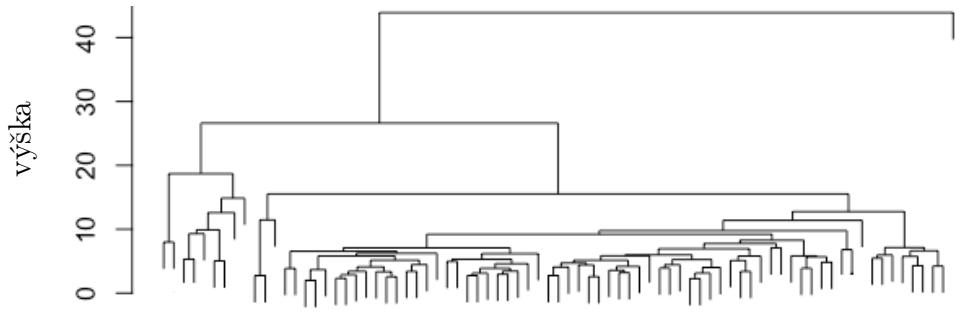
kde n_A a n_B sú počty pozorovaní v stĺpcoch A a B.

Vznik zhluku závisí na všetkých objektoch zhluku a nie len na jedinom páre dvoch extrémnych objektov. Princíp zhlukovania pomocou tejto metódy môžeme vidieť na obr. 2.6.



Obr. 2.6: Metóda priemernej väzby suseda

Na nasledujúcom dendrograme máme aplikovanú metódu priemernej väzby na demografické údaje analyzovené v bakalárskej práci.



Obr. 2.7: Metóda priemernej väzby suseda-dendrogram

Z postupu zhlukovania v dendrograme možno vidieť, že táto metóda nie je

najvhodnejšia na odhad počtu zhlukov. Stretávame sa tu s rovnakým problémom ako pri metóde najbližšieho suseda a to že okres Senec sa s ostatnými spája až v poslednom kroku zhlukovania.

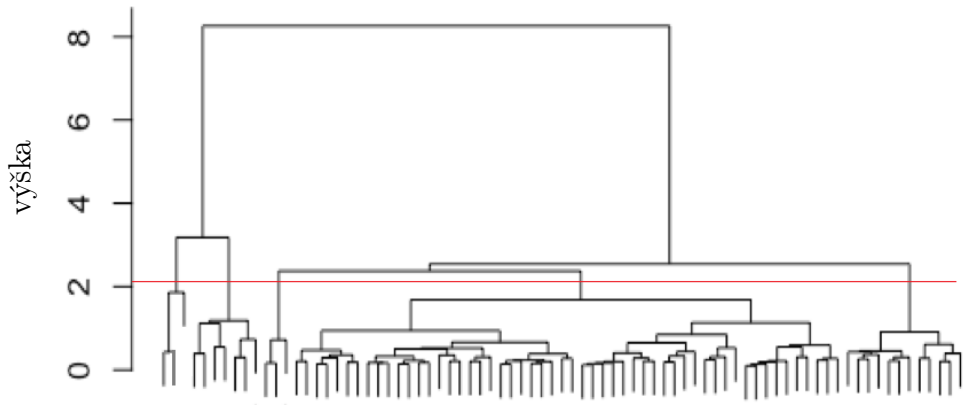
Wardová metóda - pri tejto metóde sa podobnosť objektov, respektíve zhlukov, meria ako suma štvorcov medzi objektami z dvoch zhlukov sčítaná cez všetky atribúty daných objektov. Jedinečnosť tejto metódy spočíva v minimalizácii sumy rozptylov cez všetky novovytvorené zhluky. V každom kroku sa pre všetky dvojice spočíta prírastok súčtu štvorcov odchýlok, ktorý vznikol ich zlúčením a potom sa spoja tie zhluky, ktorým odpovedá minimálna hodnota tohto prírastku. V prípade, že zhluk tvorí k objektov, ktoré sú charakterizované p znakmi, je k dispozícii matica $k \times p$ s prvkami x_{ij}, kde x_{ij} je hodnota j-tého znaku pre k-ty objekt. Variabilitu pozorovaní, ktorú tvoria riadky tejto matice, určíme pomocou zhlukovej variability WSS (vnútrozhluková suma štvorcových chýb) ktorá je daná vzťahom

$$WSS = \sum_{j=1}^{p} \sum_{i=1}^{k} (x_{ij} - \bar{x}_j)^2,$$

kde

$$\bar{x}_j = \frac{1}{k} \sum_{i=1}^{k} x_{ij}.$$

Pridávaním ďalších zhlukov s k_1 objektami sa zväčší počet riadkov pôvodnej matice na $k + k_1$ a WSS sa počíta pre väčší počet objektov. Keď začíname od jednoprvkového zhluku, bude pôvodná $WSS = 0$. Všeobecne platí, že je táto metóda považovaná za veľmi efektívnu, avšak má tendenciu kombinovať zhluky s malým počtom objektov. Na nasledujúcom obrázku môžeme vidieť dendrogram pre Wardovu metódu aplikovaný na dané údaje.



Obr. 2.8: Wardova metóda-dendrogram

V dendrograme je znázornená červená čiara, ktorá charakterizuje vhodný počet zhlukov vybraných aplikovaním tejto metódy. Na základe toho by sme zvolili päť zhlukov pre všetky údaje.

Na zobrazenie grafov pre rôzne aglomeratívne metódy sme použili príkaz $aglom = agnes(d, method = "single")$, kde sme parameter *method* menili podľa metódy na "*average*", "*complete*" alebo "*Ward*". Následne sme pre vykreslenie dendogramov pre jednotlivé metódy použili funkciu $plot(aglom)$.

Pre každý dendrogram je vypočítaná hodnota tzv. aglomeratívneho koeficientu *ac*. Tento koeficient je daný nasledujúcim vzťahom:

$$ac = \frac{1}{n} \sum_{i=1}^{n} (1 - m_i),$$

kde n je počet objektov v sledovanom súbore (v našom prípade $n = 79$) a m_i je pomer nepodobnosti prvého zhluku, ku ktorému je daný objekt i pridaný a nepodobnosti konečného zhluku (po ktorom zostáva už len jeden zhluk).

Tento koeficient v software *R* získame pomocou funkcie $aglom\$ac$ a vypíše sa nám presná honota aglomeratívneho koeficientu pre danú metódu.

Aglomeratívne koeficienty, vypočítané pre jednotlivé metódy sú v nasledujúcej tabuľke 2.1.

Metóda	hodnota ac
Najbližšieho suseda	$0,7758248$
Najvzdialenejšieho suseda	$0,9056961$
Priemernej väzby	$0,8824351$
Wardová	$0,9557056$

Tabuľka 2.1: Počet zhlukov

Čím je hodnota aglomeratívneho koeficientu vyššia, tým sa aplikovaná metóda javí vhodnejšia [11]. Z toho vyplýva, že Wardova metóda je najvhodnejšia pre analyzované demografické údaje.

Hierarchické zhlukovanie bolo uvedené z dôvodu využitia jeho výsledkov pri nasledujúcej metóde, nehierarchickej zhlukovej analýze. Z dendrogramu pre Wardovú metódu sme zvolili predpokladaný počet hlavných zhlukov, ktorého zadanie nehierarchické metódy vyžadujú. Treba podotknúť, že takto sa v praxi postupuje veľmi často. V nasledujúcej časti tak budeme pre všetky nehierarchické metódy uvažovať vždy v úvode päť zhlukov.

Kapitola 3

Nehierarchická zhluková analýza

Nehierarchické zhlukovacie metódy sa od hierarchických líšia tým, že tieto metódy vyžadujú na vstupe informáciu, aký počet zhlukov budeme vyžadovať vo výsledku. Nevytvárajú hierarchickú štruktúru, rozkladajú danú množinu do podmnožín podľa nejakého predom zvoleného kritéria. Prvý rozklad na podmnožiny sa ďalej nedelí. Upravuje sa tak, aby sa optimalizovala vzájomná vzdialenosť a odlišnosť zhlukov a aby boli v nich objekty rovnomerne rozložené. Nájdenie najlepšieho rozkladu vyskúšaním všetkých možností je veľmi zdĺhavé, v niektorých prípadoch neuskutočniteľné. Preto väčšina metód začína práve tým, že si predom určíme, na koľko zhlukov chceme rozklad previesť a potom hľadáme optimálny rozklad. Optimálny rozklad znamená najlepší rozklad, ktorý je najlepší vzhľadom k danému kritériu. Optimálnym rozkladom získame z údajov významné poznatky, ktoré budú ľahko interpretovatelné a vďaka ním získame požadovanú informáciu. Nevýhodou týchto metód je, že väčšinou končia len s lokálne optimalizovaným rozkladom.

Na vstupe týchto metód máme dátovú množinu s n objektami a počet výsledných zhlukov určených číslom K. Nehierarchické metódy zoskupujú objekty z dátovej množiny do K zhlukov. Zhluky sú sformované na základe nejakého kritéria, napríklad pomocou mier vzdialeností. Pred začatím samotnej analýzy dát je nutné vedieť požadovaný optimálny počet zhlukov. Podľa tohto môžme nehierarchické metódy rozdeliť na metódy s pevne daným počtom zhlukov a na metódy ktoré menia počet zhlukov počas analýzy.

Pri metodách s konštantným počtom zhlukov sa triedenie spraví niekoľkokrát pre rôzne počty zhlukov a z výsledkov sa vyberie najlepší. Ďalej je nutné určiť počiatočný rozklad. K tomuto najčastejšie používame dva postupy. Prvý, že rozklad je vytvorený náhodným priraďovaním objektov do zhlukov alebo sa využije nejaká hierarchická metóda. Potom sa každý jednotlivý objekt priradí postupne do všetkých zhlukov a vypočíta sa hodnota daného kritéria. Vyberie sa usporiadanie, pre ktorý je táto hodnota kritéria najlepšia. Algoritmus končí v tej chvíli, keď sa hodnoty kritéria už nelepšia, čo znamená, že najlepšej kvality sme už dosiahli. Druhý postup, ktorý môžme použiť, je vytváranie vzorových množín alebo vzorových bodov. Na začiatku prevedieme rozklad, potom sa podľa nejakého, predom určeného kritéria, vyberie z každého zhluku vzorová množina objektov. Potom vypočítame vzdialenosti jednotlivých objektov od objektov v týchto vzorových množinách a tam, kde vzdialenosť bude najmenšia daný objekt priradíme. Následne sa vypočítajú nové vzorové množiny alebo objekty. Tento postup opakujeme, pokiaľ sa už v krokoch po sebe nič nezmení.

Metódy ktoré používame pri nehierarchickom zhlukovaní sú:

- metóda K-priemerov,
- fuzzy zhlukovanie,
- metóda K-medoidov,
- kohonenové mapy (samoorganizujúce mapy).

Okrem poslednej z nich ich následne v práci predstavíme. Kľúčovým problémom všetkých nehierarchických procedúr je voľba zhlukových zárodkov (počtu zhlukov K). Existuje niekoľko heuristických postupov zadávania počiatočného počtu zhlukov, napr. sekvenčný prah alebo paralélny prah [7].

Ďalšiu z možností poskytuje výpočet koeficientu CCC (Cubic clustering criterion). Toto kritérium navrhol v roku 1983 W.S. Sarle [9] a bolo empiricky odvodené z metód Monte Carlo. Kritérium je využívané k odhadu výsledného počtu zhlukov u Wardovej metódy minimalizácie rozptylu, u metódy K- priemerov, či iných metód využívajúcich minimalizácie medziskupinového súčtu štvorcov. Výpočet koeficientu v programe R bol uskutočnený pomocou nasledujúceho príkazu

$CCC = NbClust(data = X, method = "ward.D")$ z knižinice $library(NbClust)$. Vhodný počet zhlukov indikuje $CCC > 3$. Vo všeobecnosti čím vyššie je CCC, tým lepší počet zhlukov získame. Pri hierarchickom postupe zhlukovania možno pozorovať niekoľko lokálnych vysokých hodnôt CCC. Pri nehierarchickom zhlukovaní je pozorované veľmi odlišné globálne maximum CCC v závislosti od počiatočného určenia počtu zhlukov. V našom prípade tak vyšiel optimálny počet zhlukov 5, čo odpovedá naším predchádzajúcim úvahám a konkrétna hodnota koeficientu CCC vyšla 21.2197. Podľa tohto koeficienta by sa pozorovania mali rozdeliť do piatich zhlukov.

Na základe výsledkov z hierarchického zhlukovania pomocou Wardovej metódy, CCC koeficientu a subjektívneho posúdenia analyzovaných demografických údajov sme zvolili počiatočný počet zhlukov $K = 5$. Prvé dva zhluky sú tvorené územnými celkami dvoch najväčších miest na Slovensku. Tieto okresy sú si podobné nie len analyzovanými demografickými údajmi, ale aj niektorými ekonomickými a socialnými ukazovateľmi. V týchto okresoch je najvyššia hodnota Regionálneho HDP na obyvateľa a najnižšia nezamestnanosť. V treťom zhluku sú tri satelitné okresy miest Bratislava a Košice. Štvrtý zhluk tvoria okresy severu Slovenska, ktoré majú najnižšie hodnoty ukazovateľov ekonimockej prosperity regiónov. Piaty zhluk je tvorený najväčšiou skupinou okresov. V tomto zhluku rôzne socialno-ekonomické ukazovatele vykazujú hodnoty podobné celoslovenskému priemeru.

3.1. Metóda K-priemerov

Pri vytváraní malého počtu zhlukov z veľkého počtu objektov je najvhodnejšia metóda K-priemerov. Algoritmus tejto metódy navrhol Stuart Lloyd v roku 1957 ako techniku pulznej modulácie, aj keď algoritmus bol publikovaný až v roku 1982. Metóda K-priemerov vyžaduje spojité premenné bez odľahlých pozorovaní. Objekty sú spájané do zhlukov podľa zhlukovacích premenných na základe minimalizácie medzizhlukovej sumy štvorcov. Nehierarchická zhluková metóda K-priemerov hľadá optimálne rozloženie zhlukov z údajov na základe minimalizácie

vhodného kritéria, typickým je v tomto ohľade kritérium súčtu štvorcov chýb (ESS). To je definované nasledovne,

$$ESS = \sum_{k=1}^{K} \sum_{c(i)=k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T (\mathbf{x}_i - \bar{\mathbf{x}}_k),$$

kde $c(i)$ je zhluk obsahujúci pozorovanie $\mathbf{x}_i$, $\bar{\mathbf{x}}_k$ sú priemery jednotlivých zhlukov pre všetky $k = 1, \ldots, K$.

Na rozdiel od hierarchického zhlukovania si táto metóda vyžaduje zadať na úvod požadovaný počet zhlukov. V rámci toho potom považujeme za optimálne riešenie také, kde ak presunieme jeden objekt z jedného zhluku do druhého, tak nám to nezníži súčet štvorcov vzdialeností v zhluku. Algoritmus môžme opakovať viackrát s rozdielnými počiatočnými konfiguráciami. Z niekoľkých výsledných riešení vyberieme to najlepšie. Táto metóda je jednoduchá a pozostáva z troch krokov. Postup je nasledovný:

1. Vyberieme K objektov z údajovej množiny ako ťažiská zhlukov.
2. Priradíme objekt zhluku, ktorému je objekt najpodobnejší na základe priemernej hodnoty objektov v zhluku. Vypočítame nové ťažiská zhlukov a to tak, že v každom zhluku vypočítame priemernú hodnotu z objektov.
3. Postup opakujeme, pokiaľ nezaznamenávame zmeny v zhlukoch.

Metóda K-priemerov je jedna zo základných zhlukovacích metód aj vďaka jednoduchému výpočtu. Avšak ako každá metóda, má svoje nevýhody. Nehodí sa na odhaľovanie zhlukov s nekonvexným tvarom, čo znamená že objekty nevieme spojiť jedným ťahom tak, aby celá spojnica ležala v oblasti vytvorenej daným zhlukom. Ak by sa v dátach nachádzali zle namerané hodnoty, mohli by podstatne ovplyvniť ťažiská zhlukov. Problém chýbajucích hodnôt možno riešiť pomocou nastavenia percenta chýbajúcich parametrov. Potom pozorovania s vyššiou hodnotou percenta chýbajúcich hodnôt než je stanovená sa ignorujú.

Aplikovaním metódy K-priemerov na analyzované demografické údaje bolo rozdelených 79 objektov do piatich zhlukov následujúcim spôsobom:

zhluk	1	2	3	4	5
počet okresov	3	24	16	9	27

Tabuľka 3.1: Počet zhlukov

Pomocou softwaru R sme dostali presné rozdelenie okresov do zhlukov príkazom $cl = kmeans(X, 5)$. V nasledujúcej tabuľke môžeme vidieť maticu zhlukových centier :

Zhluk	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Zhluk 1	33.02	84.01	12.43	32.39	33.37	79.93	11.37	14.60	1.57	41.73	42.14	22.59
Zhluk 2	28.47	91.88	11.36	28.37	30.17	76.77	8.19	15.60	1.63	39.39	5.68	8.42
Zhluk 3	30.24	88.16	8.66	29.49	31.32	78.13	10.27	15.79	1.52	40.44	8.87	10.03
Zhluk 4	32.62	84.05	10.61	31.26	32.38	76.09	8.28	15.28	1.47	41.72	22.24	20.94
Zhluk 5	31.66	82.87	8.91	29.31	30.85	77.50	10.19	15.98	1.51	40.87	7.34	8.61

Tabuľka 3.2: Matica zhlukových centier

V tabuľke 3.2 sú vypočítané priemerné hodnoty analyzovaných premenných pre každý zhluk. Pre okresy zhluku 1 je napríklad primerná hodnota veku snúbencov (premenná x_1) vyššia skoro o 5 rokov, ako v zhluku 2. Je všeobecne známe, že mladí ľudia, žijúci vo veľkomestách vstupujú do manželstva neskôr ako ľudia žijúci v malých mestách.

Pomocou metódy K-priemerov sme dostali päť zhlukov s nasledujúcou štruktúrou:

Zhluk 1 (ružová)
Bratislava I
Bratislava III
Senec

Zhluk 2 (modrá)			
Bánovce nad Bebravou	Kysucké Nové Mesto	Bardejov	Snina
Považská Bystrica	Námestovo	Kežmarok	Spišská Nová Ves
Tvrdošín	Stará Ľubovňa	Levoča	Trebišov
Bytča	Stropkov	Poprad	Gelnica
Čadca	Svidník	Prešov	Michalovce
Dolný Kubín	Vranov nad Topľou	Sabinov	Sobrance

Zhluk 3 (žltá)			
Piešťany	Zlaté Moravce	Zvolen	Krupina
Ilava	Žilina	Žarnovica	Humenné
Púchov	Ružomberok	Bánska Štiavnica	Medzilaborce
Trenčín	Turčianske Teplice	Detva	Košice - okolie

Zhluk 4 (červená)		
Bratislava II	Malacky	Košice II
Bratislava IV	Pezinok	Košice III
Bratislava V	Košice I	Košice IV

Zhluk 5 (zelená)			
Dunajská Streda	Nové Mesto nad Váhom	Šaľa	Brezno
Galanta	Partizánske	Topoľčany	Lúčenec
Hlohovec	Prievidza	Liptovský Mikuláš	Poltár
Senica	Komárno	Martin	Revúca
Skalica	Levice	Veľký Krtíš	Rimavská Sobota
Trnava	Nitra	Žiar nad Hronom	Rožňava
Myjava	Nové Zámky	Banská Bystrica	

Podrobnejšie si interpretujeme vzniknuté zhluky.

Prvý a štvrtý zhluk tvoria okresy dvoch najväčších miest na Slovensku, Bratislava a Košice. Sú to okresy s najvyššim Regionálnym HDP na obyvateľa, najnižšiou nezamestnanosťou, najvyššiou vzdelanosťou a najvyšším príjmom.

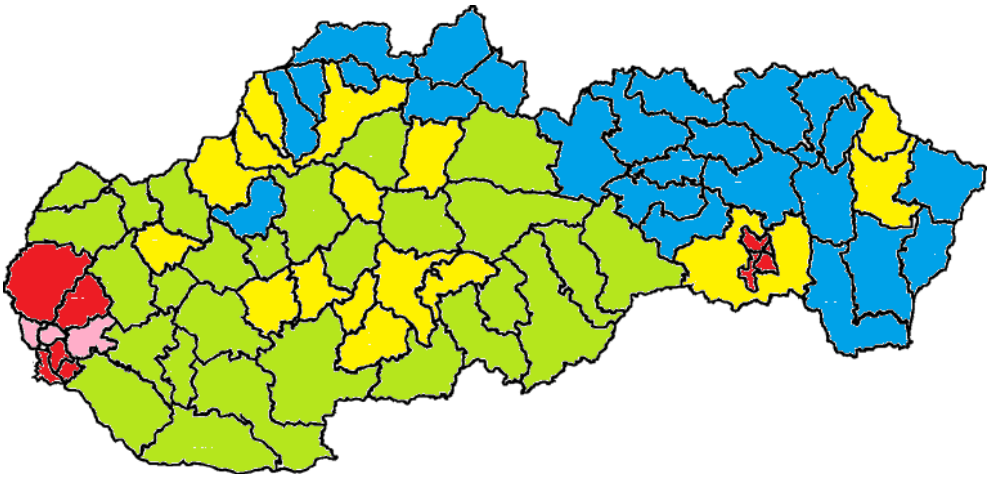
Prvý zhluk je tvorený tromi okresmi Bratislavského kraja. V týchto okresoch väčšina z analyzovaných demografických ukazovateľov nadobúda maximálnu hodnotu. V tomto zhluku je najvyššia hodnota priemerného veku žien, vstupujúcich do manželstva. Toto je charakteristický jav veľkých aglomerácií. S tým súvisí aj najvyššia hodnota priemerného veku ženy pri pôrode.

Zhluk 2 tvoria okresy, kde ženy vstupujú do manželstva v najnižšiom veku, je tu najnižší priemerný vek ženy pri pôrode i pri potrate. Sú to okresy, ktoré majú najnižší Regionálny HDP na obyvateľa v SR. V týchto okresoch sa postupne utlmuje napríklad textilný priemysel. Možno konštatovať, že tu existuje najnižšia možnosť profesionálnej realizácie žien na Slovensku. Taktiež je tu vysoké percento obyvateľstva rómskej etniky. Tento zhluk je tiež zaujímavý najnižšou ženskou úmrtnosťou. To opäť môže súvisieť s tým, že v týchto regiónoch sa ženy viac venujú životu v domácnosti ako profesijnej kariére.

Zhluk 3 sa výrazne odlišuje len v najnižšom počte narodených žien na počet obyvateľov,čo je však ukazovateľ, ktorým nemožno nijako ovplyvniť a nezávisí so sociálno-ekonomickými podmienkami okresov. Zaradené sú tu okresy, ktoré z ekonomického hľadiska zaznamenávajú v poslednom období pokrok. Ostatné hodnoty v tomto zhluku sú priemerné.

Zhluk 4 obsahuje okresy v okolí Bratislavy a Košíc. Na rozdiel od prvého zhluku, tento zhluk nemá žiadne významne rozdielne hodnoty od ostatných. Dokonca čo je zaujímavé, tento zhluk sa vyznačuje tým že obsahuje okresy s najnižším priemerným vekom žien pri úmrtí a najnižším počtom maloletých detí v rozvedenom manželstve. Je to celkom zaujímave, keďže v tomto zhluku by sme očakávali skôr vysoké hodnoty.

Zhluk 5 ktorý obsahuje 34% okresov má priemerné hodnoty všetkých demografických ukazovateľov, až na nízky podiel žien, ktoré vstupujú do manželstva ako slobodné. Z dlhodobého prieskumu je známe, že tieto okresy obsahujú dosť veľké percento rómskeho etnika. Ženy sa tu často vydávajú veľmi mladé a nie je to považované za nič nezvyčajné.

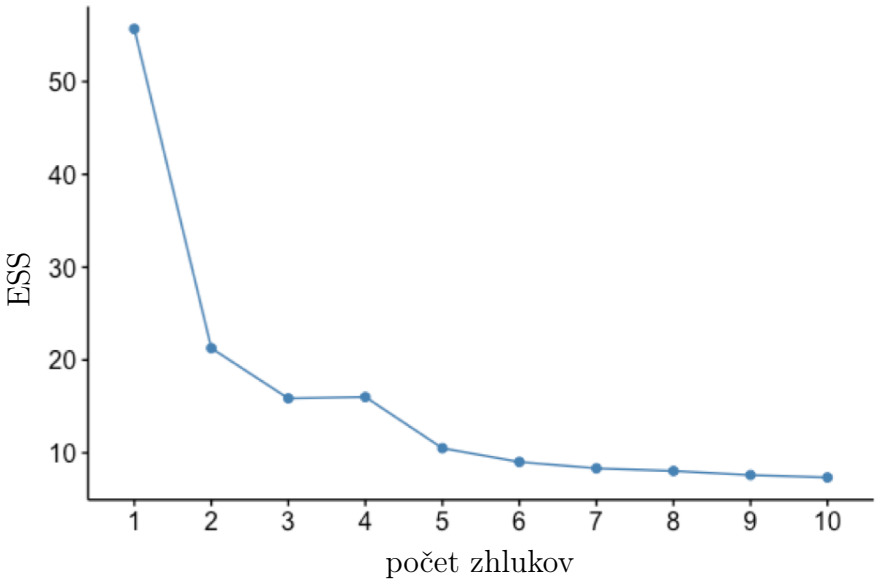


Obr. 3.2: Mapa zhlukov pre K-priemerov

Na obr. 3.2 môžeme vidieť farebne odlíšené zhluky získané pomocou metódy K-priemerov. Až na okresy v zhluku 3 (žltá) sa všetky ostatné zhluky javia homogénne.

Pri zhlukovej analýze je vždy diskutabilný zvolený počet zhlukov.

Na nasledujúcom grafe je zobrazená závislosť hodnoty ESS od počtu zhlukov. Zdá sa, že voľba piatich zhlukov je najvhodnejšia nakoľko zlom krivky ESS sa javí pri čísle 5.



Obr. 3.3: Optimálny počet zhlukov v závislosti na ESS

Zlom krivky môžeme vidieť aj pri čísle 2, čo by znamenalo že aj voľba dvoch

zhlukov by mohla byť vhodná. Avšak zaraďovanie okresov len do dvoch zhlukov dostatočne nevystihuje štruktúru populácie na Slovensku. Preto volíme pre lepšiu interpretáciu päť zhlukov.

Tento graf sme získali pomocou softwaru R, kde sme nainštalovali knižnice $NbClust$, $cluster$ a $factoextra$.

Potom sme na naše údaje použili funkciu $fviz\_nbclust(X, kmeans, method =$ "wss"), ktorá nám vykreslila daný graf.

3.2. Metóda K-medoidov

Ako vieme, cieľom zhlukovej analýzy je rozdeliť súbor objektov do dvoch alebo viacerých zhlukov tak, aby objekty v rámci jedného zhluku boli podobné a objekty medzi rôznymi zhlukmi naopak zas rozdielne. Metóda K-medoidov, ktorú teraz budeme prezentovať, sa pokúša tento cieľ dosiahnuť nájdením súboru reprezentatívnych objektov nazývaných medoidy. Medoid zhluku je definovaný ako objekt, pre ktorý je priemerná vzdialenosť voči ostatným objektom v klastri minimálna. Ak je vyžadovaných K zhlukov, tak nájdeme aj K medoidov. Akonáhle nájdeme medoidy v dátach, objekty sú zaradené do zhluku podľa najbližšieho medoidu. Máme k dispozícií dva algoritmy na vykonanie zhlukovania. Prvý, ktorý prezentoval Spatha v roku 1985, používa na začiatku náhodnú konfiguráciu klastrov. Druhý je od Kaufmana a Rousseeuwa (rok 1990), ktorý má špeciálne použitie siluetovej štatistiky, ktorá nám pomôže správne určiť počet zhlukov [8].

Ako prvý si vysvetlíme algoritmus popísaný Spathom, keďže bol prezentovaný v roku 1985, niektoré literatúry ho uvádzajú ako prvý K-medoidový algoritmus. Spomínaná metóda minimalizuje kritérium optimality tým, že postupne presúva objekty z jedného do druhého zhluku. Algoritmus začína pri náhodnej štartovacej konfigurácií, potom prechádza na nájdenie lokálneho minima inteligentným pohybom objektov z jedného zhluku do druhého. Ak už pohyb objektov nevedie k zníženiu hodnoty kritéria optimality, algoritmus ukončujeme. Nanešťastie, toto lokálne minimum nemusí byť vždy nutne aj globálnym minimum. Spomenuté kritérium optimality D je celková vzdialenosť medzi objektami v rámci zhlukov. Matematicky je vyjadrená nasledovným vzťahom:

$$D = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in C_k} d_{ij},$$

kde K je počet zhlukov, d_{ij} je zvolená vzdialenosť medzi objektami i a j, a C_k je súbor všetkých objektov v zhluku k, $k = 1, \ldots, K$.

V metóde K-priemerov sú stredy klastrov dané ako priemer objektov v danom zhluku. Avšak priemery sú veľmi náchylné k odľahlým hodnotám. Napríklad sa

môžme dostať do situácie, keď budeme mať len pár objektov v zhluku. Ak by náhodou ešte aj tie boli odľahlé, tak by sme potom veľmi ťažko mohli interpretovať výsledok. Jedno riešenie ako môžme takúto situáciu vyriešiť je, že použijeme silný algoritmus, kde vplyv odľahlých hodnôt bude minimálny.

Jeden z príkladov takéhoto algoritmu je algoritmus, ktorý sa nazýva Partitioning Around Medoids, v tejto práci ho budeme nazývať algoritmus PAM. Predstavuje druhý zo spomínaných algoritmov. Algoritmus PAM je založený na hľadaní K reprezentatívnych objektov alebo medoidov medzi pozorvaniami. Tieto pozorovania potom predstavujú štruktúru údajov. Po nájdení množiny K medoidov sa následne zhluky zostavujú tak, že každé pozorovanie priradíme k najbližšiemu medoidu. Cieľom je nájsť K reprezentatívnych objektov, ktoré minimalizujú súčet nepodobností objektov k ich najbližšiemu reprezentatívnemu objektu.

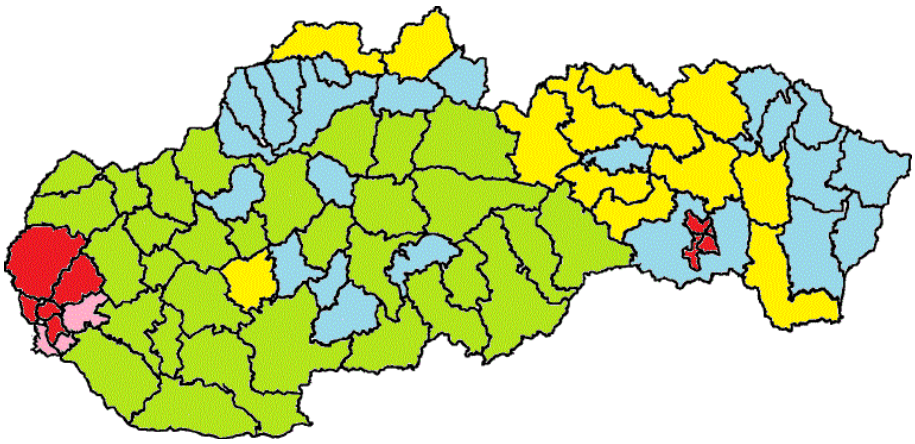
Na začiatku algoritmu, keď ešte nemáme špecifikované medoidy, algoritmus najskôr hľadá dobrú počiatočnú sadu medoidov. Táto fáza sa nazýva fáza vytvárania. V druhej fáze tento algoritmus skúša všetky možné výmeny medoidov až pokiaľ nenastáva žiadne zlepšenie. Táto fáza sa nazýva fáza výmeny. Keď sú medoidy špecifikované, na ich poradí nezáleží. Vo všeobecnosti bol tento algoritmus navrhnutý tak aby nezáležalo na poradí pozorovaní. V software R sa najskôr musí nainštalovať knižnica $flexclust$ a potom použiť príkaz $dpk = pam(d, k = 5)$.

Opísaný algoritmus je použitý v nasledujúcej časti na klasifikáciu analyzovaných demografických údajov. Aplikovaním algoritmu PAM bolo odhadnutých päť zhlukov. V tabuľke 3.4. je uvedené jednotlivé zaradenie objektov do týchto zhlukov. Ďalej je tu uvedená maximálna vzdialenosť medzi objektami v zhluku a priemerná vzdialenosť. Túto tabuľku dostaneme pomocou príkazu $dpk\$clusinfo$.

zhluk	1	2	3	4	5
počet okresov	3	9	31	23	13
maximálna vzdialenosť	19.25	15.62	9.62	9.21	9.62
priemerná vzdialenosť	9.07	8.82	4.79	5.28	5.38

Tabuľka 3.3: Rozdelenie zhlukov

Namiesto nájdenia centrov zhluku v optimálnych pozíciach má metóda K-medoidov za cieľ nájsť K reprezentatívnych objektov v rámci dátového súboru. Typicky najskôr minimalizujeme súčet vzdialeností skôr ako súčet štvorcových vzdialeností, čo nám zaručuje zníženie významnosti veľkých vzdialeností. V našom prípade celková priemerná vzdialenosť po druhej fáze sa znížila približne o 0.01 z pôvodnej hodnoty 0.3471236 na hodnotu 0.3394799.



Obr. 3.5: Mapa zhlukov pre metódu K-medoidov

Na obr. 3.5 je ilustrovaný výsledok popísaného zhlukovania. Vidíme, že objekty sú do zhlukov zaradené podľa prirodzenej geografickej štruktúry regiónov Slovenskej republiky.

Konkrétne zaradenie jednotlivých okresov do zhlukov je v nasledujúcej tabuľke.

Zhluk 1 (ružová)
Bratislava I
Bratislava III
Senec

Zhluk 2 (červená)		
Bratislava II	Malacky	Košice II
Bratislava IV	Pezinok	Košice III
Bratislava V	Košice I	Košice IV

Zhluk 3 (zelená)			
Levice	Rožňava	Trnava	Zvolen
Nové Zámky	Liptovský Mikuláš	Myjava	Piešťany
Nové Mesto nad Váhom	Martin	Partizánske	Brezno
Lučenec	Bánska Bystrica	Poltár	Veľký Krtíš
Skalica	Dunajská Streda	Topoľčany	Trenčín
Prievidza	Šaľa	Hlohovec	Rimavská Sobota
Komárno	Senica	Revúca	Ružomberok
Galanta	Nitra	Žiar nad Hronom	

Zhluk 4 (modrá)			
Detva	Bánovce nad Bebravou	Košice-okolie	Banská Štiavnica
Krupina	Púchov	Žilina	Svidník
Sobrance	Považská Bystrica	Michalovce	Turčianske Teplice
Medzilaborce	Bytča	Žarnovica	Zlaté Moravce
Humenné	Dolný Kubín	Levoča	Ilava
Snina	Stropkov	Kysucké Nové Mesto	

Zhluk 5 (žltá)			
Stará Ľubovňa	Námestovo	Prešov	Poprad
Sabinov	Tvrdošín	Gelnica	
Kežmarok	Spišská Nová Ves	Čadca	
Vranov nad Topľou	Bardejov	Trebišov	

Zaradenie okresov do zhlukov pri tejto metóde je podobné ako pri aplikovaní metódy K-priemerov. Na rozdiel od metódy K-priemerov sa nám zaradenie okresov do zhlukov javí menej homogénne. Preto môžeme skonštatovať, že metóda K-medoidov je v tomto prípade menej efektívnejšia. Priemerné hodnoty ukazovateľov v jednotlivých zhlukov sú v tabuľke 3.4.

Zhluk	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Zhluk 1	33.02	84.01	12.43	32.39	33.37	79.93	11.37	14.60	1.57	41.73	42.14	22.59
Zhluk 2	32.63	84.05	10.61	31.26	32.38	76.09	8.28	15.28	1.47	41.72	22.23	20.94
Zhluk 3	31.56	83.49	8.92	29.43	30.92	77.70	10.19	16.06	1.52	40.92	7.54	8.85
Zhluk 4	29.5	89.82	9.18	29.01	30.93	77.54	9.73	15.74	1.54	39.93	7.94	9.96
Zhluk 5	28.01	92.55	12.79	27.97	29.77	76.43	7.51	15.43	1.68	39.02	4.69	7.59

Tabuľka 3.4: Matica zhlukových centier

Prvý a druhý zhluk sú okresy Bratislavy a Košíc. Môžeme vidieť že tieto dva zhluky sa úplne rovnaké ako zhluk 1 a 3 pri metóde K-priemerov.

Zhluk 1 obsahuje tri okresy, ktoré ležia v okolí Bratisavy. Preto je prirodzené, že v tomto zhluku máme extrémne hodnoty niektorých premenných. Napríklad sa vyznačuje vysokou priemernou hodnotou pre vek ženy pri sobáši aj pri úmrtí, vysokým počtom zomrelých žien na počet obyvateľov, vysokým priemerným vekom manželov pri rozvode a taktiež vysokou hodnotou v počte prisťahovalých a vysťahovalých žien.

V zhluku 2 sú okresy Bratislavy a Košíc. Tento zhluk má na prekvapenie skoro všetky hodnoty priemerné. Dokonca v tomto zhluku môžeme spozorovať najnižší priemerný vek ženy pri úmrtí a najnižší priemerný počet maloletých detí v rozvedenom manželstve. Môže to súvisieť s lokalitou týchto okresov pri dvoch najväčších mestách SR. Obyvateľstvo tu žije vo veľkých mestských sídliskách bez možnosti prírodného, sociálneho a kultúrneho života.

V zhluku 3 sa nachádza väčšina okresov stredného a západného Slovenska. V tomto zhluku je zaujímavá vysoká hodnota priemernej dĺžky trvania rozvedeného manželstva. Sú to okresy, ktoré sú charakterizované klasickým pohľadom na rodinu so silným kresťanským vplyvom.

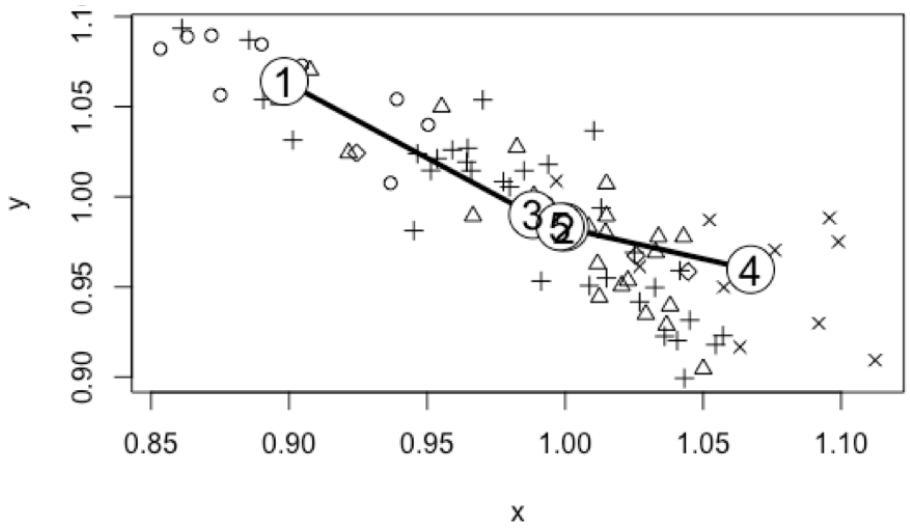
Priemerné hodnoty zhluku 4 sú blízke priemerným hodnotám sledovaných ukazovateľov v Slovenskej republike.

Okresy zaradené do piateho zhluku tvoria z ekonomicko-sociálneho hľadiska najchudobnejšie okresy Slovenska. Významné rozdiely medzi sledovanými demografickými ukazovateľmi vidíme skoro vo všetkých premenných. Zrejmé je to napríklad v prípade počtu prisťahovaných aj vysťahovaných žien, kde toto číslo je

veľmi malé na rozdiel od ostatných zhlukov. Naopak v tomto zhluku je výrazne najvyšší počet žien, ktoré vstupujú do manželstva ako slobodné. To opäť môže súvisieť s nízkym počtom kariérnych príležitostí ženskej populácie.

Jedným z ďalších grafických výstupov pre zhlukovanie je graf okolia (susedstva) alebo v literatúre známy ako „Neighbourhood graph". V tomto grafe máme zhluky očíslované, toto číslo je vždy na mieste centroidu zhluku. Centroidy sú spájané rôzne hrubými priamkami. V tomto grafe sa dva centroidy zhlukov spoja vtedy, ak existuje aspoň jedno pozrovanie, pre ktoré sú to dva najbližšie centroidy. Ukážku takéhoto grafu môžeme vidieť na Obr. 3.7, kde je analyzovaných 79 objektov.

Tento graf sme získali pomocou príkazov $k = cclust(X, k = 5, save.data = TRUE)$ a príkazu $plot(k, hull = FALSE, col = rep("black", 5))$.



Obr. 3.6: Zhlukovanie pomocou metódy K-medoidov

Je zrejmé, že bolo identifikovaných päť zhlukov, ktoré sú si dosť blízke. Dokonca tri z nich sú si tak blízke až sa prekrývajú. Na základe tohto grafu môžeme skonštatovať, že by mohli byť aj tri zhluky ako vhodný počet pre zhlukovanie. Vzhľadom na charakter analyzovaných demografických údajov sme sa rozhodli zoskupiť okresy do piatich zhlukov.

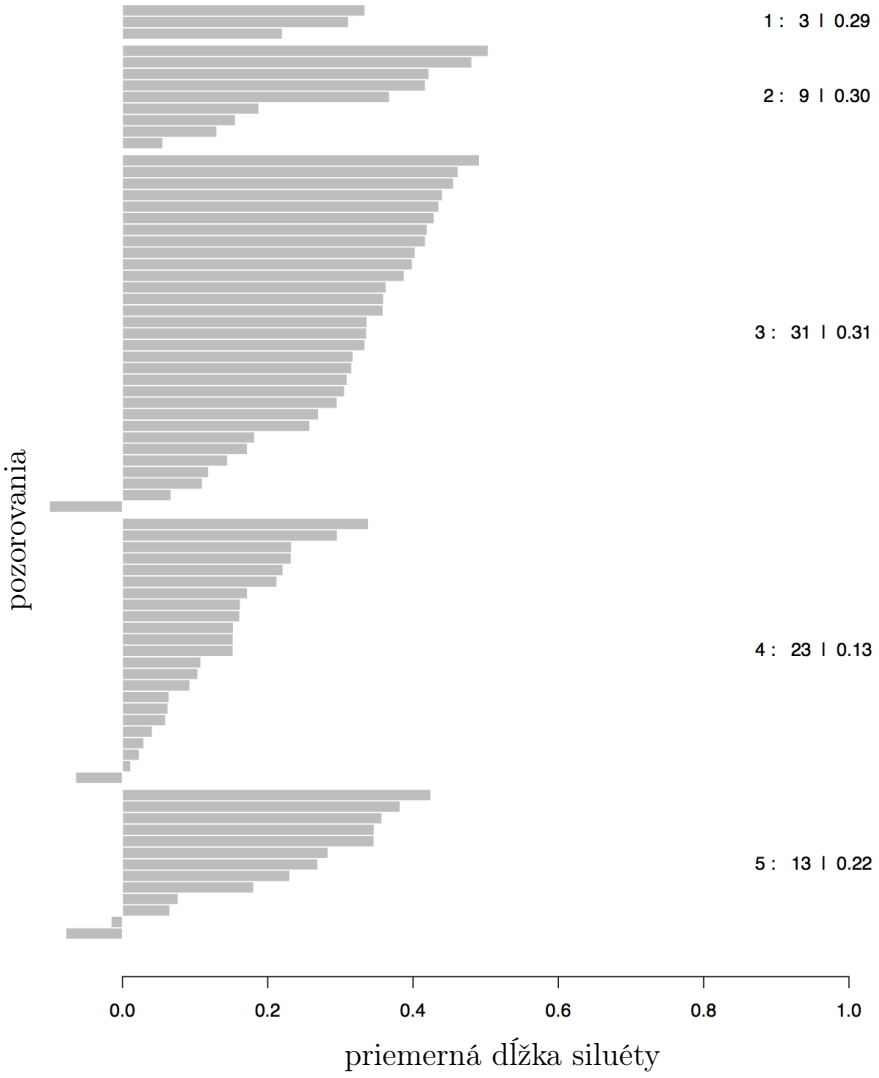
3.2.1. Siluetový graf

Jeden z užitočných výstupov nehierarchických metód je siluetový graf. Často sa stretávame aj s názvom obrysový graf alebo po anglicky „Silhouette plot". Tento graf nás informuje ako sa nám rozdelili jednotlivé pozorovania do zhlukov. Znázorňuje aj kvalitu rozdelenia pre individuálne zhluky. To znamená, že objekty s veľmi dlhou siluetov (blízko jednotky) sú veľmi dobre zaradené do zhluku. Zatiaľ čo objekty s hodnotou blízko nuly ležia medzi dvomi a viacerými zhlukmi. Objekt s negatívnou hodnotou môže byť tiež zaradený v zlom zhluku. Dĺžka siluéty s_i objektu i je daná následujúcim vzťahom

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)},$$

kde a_i je priemerná vzdialenosť objektu i od ostatných objektov v danom zhluku a b_i je najmenšia vzdialenosť objektu i k ďalšiemu zhluku. Takže maximálnu hodnotu dĺžky siluety dostaneme vtedy, keď vnútrozhluková vzdialenosť a_i má oveľa menšiu hodnotu ako medzizhluková vzdialenosť b_i.

Pre dané demografické údaje môžme vidieť siluetový graf na následujúcom obrázku. Tento siluétový graf sme dostali príkazom navezujúcim na metódu PAM, $plot(dpk, main = "Silhouette\ plot")$.



Obr. 3.7: Siluetový graf pre metódu K-medoidov aplikovanú na dané demografické dáta.

Situáciu z výstupu analýzy v krátkosti zhodnotíme v nasledujúcej tabuľke.

zhluk	1	2	3	4	5
počet okresov	3	9	31	23	13
priemerná dĺžka siluety	0,29	0,30	0,31	0,13	0,22

Tabuľka 3.5: Rozdelenie zhlukov

Z relatívne menších priemerných dĺžok siluét v rámci jednotlivých zhlukov (viď tabuľka 3.5) je možné posúdiť, že priradenie do zhlukov nie je vo väčšine prípadov celkom jednoznačné, čo zodpovedá charakteru demografických údajov, skôr menšími rozdielmi v (normovaných) hodnotách premenných.

Môžme vidieť, že v treťom, štvrtom a piatom zhluku je niekoľko zle zaradených objektov. V tomto prípade je hodnota siluety záporná. V treťom zhluku je to pozorovanie číslo 41, čo reprezentuje okres Ružomberok. V tomto regióne sa hodnoty jednotlivých ukazovateľov výrazne odlišujú od priemerných hodnôt týchto ukazovateľov zhluku 3. Napríklad priemerný vek ženy vstupujúcej do manželstva v zhluku 3 je 31.56 rokov, ale v okrese Ružomberok má tento ukazovateľ hodnotu 33.54. Táto hodnota je dokonca vyššia ako najväčšia priemerná hodnota.

Vo štvrtom zhluku je zle zaradené pozorovanie číslo 17, okres Ilava. Okres Ilava sa z historického a geografického hľadiska dlhodobo javí ako umelo vytvorený na základe politickej požiadavky vtedajšieho obdobia. V tomto prípade je to podobné ako s okresom Ružomberka, kde sa hodnoty pre okres Ilava výrazne líšia od priemerných hodnôt štvrtého zhluku, do ktorého bol zaradený.

V piatom zhluku máme až dve zle zaradené pozorovania, pozorovanie číslo 70, okres Trebišov a 65, okres Poprad. Okres Poprad sa výrazne líši v šiestich ukazovateľoch. Najvýraznejšie je to v ukazovateli x_2, podiel žien ktoré vstupujú do manželstva ako slobodné a x_{10}, priemerný vek manželov pri rozvode. V týchto ukazovateľoch sú hodnoty výrazne nižšie ako v ostatných okresoch, zaradených do piateho zhluku. Okres Trebišov sa výrazne líši v ukazovateli x_2 a x_7, počet zomrelých žien na počet obyvateľov. Tieto hodnoty sú v porovnaní s ostatnými okresmi vyrazne nižšie. Môže to opäť súvisieť s výrazným percentuálnym zastúpením rómskej menšiny v okrese Trebišov.

3.3. Modelové zhlukovanie

Aglomeratívne zhlukovanie a zhlukovanie pomocou metódy K-priemerov a K-medoidov, ktoré sme spomínali v predchádzajúcej časti, majú spoločný problém. Tieto metódy nie sú založené na formálnom modele na zhlukovanie objektov. Tieto metódy tak neumožňujú efektívne riešiť niektoré problémy ako napríklad rozhodovanie medzi modelmi, odhad počtu zhlukov, atď. To je najväčší problém, že bez formálneho modelu nedokážeme robiť zrozumiteľné závery. Avšak nie je to problém na základe ktorého by sme usúdili, že tieto metódy sú úplne zlé, pretože zhluková analýza je veľmi často využívaná ako nástroj na preskúmanie dát. Ak ale pomocou inej metódy nájdeme oveľa zrozumiteľnejší (prijateľnejší) model pre údajovú štruktúru, tak ho rozhodne uprednosníme. Jedna z alternatívnych zhlukovacích metód, k tým predchádzajúcim, je zhluková analýza založená na predpoklade štatistického modelu, ktorá nám prináša oveľa presvedčivejšie riešenia.

V tejto kapitole si modelové zhlukovanie podrobnejšie vysvetlíme. Predpokladáme formálny štatistický model pre populáciu, z ktorej máme vzorku našich údajov. Taký model, v ktorom predpokladáme, že daná populácia sa skladá z viacerých subpopulácií („zhlukov“) a kde každá táto subpopulácia má premenné s rozdielnou mnohorozmernou pravdepodobnostnou hustotou. Agregaciou týchto hustôt subpopulácií vzniká hustota pre populáciu, ktorú nazývame *konečná zmiešaná hustota*. Konečná zmiešaná hustota veľmi často poskytuje rozumný štatistický model pre zhlukovanie a zhluková analýza, ktorá je založená na tejto hustote sa nazýva *modelové zhlukovanie*. Pri zpracovaní kapitoly modelového zhlukovania boli využité zdroje [5], [2] a [12].

3.3.1. Konečná zmiešaná hustota

Konečná zmiešaná hustota je trieda viacerých funkcií hustôt pravdepodobnosti a je zadaná v následujúcom tvare

$$f(\mathbf{x}; \mathbf{p}, \boldsymbol{\theta}) = \sum_{j=1}^{c} p_j g_j(\mathbf{x}; \boldsymbol{\theta}_j), \tag{3.1}$$

kde $\mathbf{x}$ predstavuje realizácie p-rozmernej náhodnej premennej, $\mathbf{p} = (p_1, p_2, \ldots, p_c)^T$, ktorá obsahuje proporcionálne vplyvy jednotlivých zložiek hustôt g_j na celkovej hustote f ($p_j > 0$, $\sum_{j=1}^{c} p_j = 1$) a zložky $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_c^T)^T$ potom predstavujú parametrizáciu hustôt g_j. Počet zložiek v zmesi (predpokladaný počet zhlukov) je tak daný číslom c.

Konečná zmiešaná hustota poskytuje vhodný model zhlukovania práve vtedy, ak platí, že každá skupina pozorovaní, v rámci populácie rozdelení, má rôzne rozdelenie pravdepodobnosti. Môžu patriť do rovnakej triedy, ale budú sa líšiť hodnotami parametrov rozdelenia. Príkladom tohto princípu su naše údaje, ktoré tu analyzujeme, nakoľko majú normálne rozdelenie s rôznymi vektormi priemerov a kovariančnými maticami. Ak máme odhadnuté parametre predpokladanej zmiešanej hustoty, pozorovania môžeme zaraďovať do konkrétnych zhlukov na základe maximálnych hodnôt odhadovanej podmienenej pravdepodobnosti

$$\widehat{P}(zhlukj|\mathbf{x}_i) = \frac{\widehat{p}_j g_j(\mathbf{x}_i; \widehat{\theta}_j)}{f(\mathbf{x}_i; \widehat{\mathbf{p}}, \widehat{\theta})}, j = 1, \ldots, c.$$

3.3.2. Odhad parametrov v modele zmiešaných hustôt

Na základe pozorovaní $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ dosadených do vzťahu (3.1) definujeme logaritmickú vierohodnostnú funkciu l,

$$l(\mathbf{p}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(\mathbf{x}_i; \mathbf{p}, \boldsymbol{\theta}).$$

Odhady parametrov použitých v hustote pomocou metódy maximálnej vierohodnosti môžeme dostať pomocou vyriešenia následujúcej rovnice

$$\frac{\partial l(\boldsymbol{\varphi})}{\partial(\boldsymbol{\varphi})} = 0,$$

kde $\boldsymbol{\varphi} = (\mathbf{p}^T, \boldsymbol{\theta}^T)^T$. Avšak v prípade konečnej zmiešanej hustoty je táto hustota obvykle dosť zložitá na to, aby sa použili bežné metódy na riešenie výslednej sústavy rovníc. Preto maximálne odhady parametrov v modele s konečnou zmiešanou hustotou musíme vypočítať iným spôsobom. V prípade keď máme zmes,

v ktorej má hustota j-tej zložky viacnásobné normálne rozdelenie s vektorom stredných hodnôt $\boldsymbol{\mu}_j$ a kovariančnou maticou $\boldsymbol{\Sigma}_j$, odhadneme parametre pri ich vhodnej počiatočnej volbe iteratívne nasledujúcim spôsobom:

$$\widehat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \widehat{P}(j|\mathbf{x}_i),$$

$$\widehat{\boldsymbol{\mu}}_j = \frac{1}{n\widehat{p}_j}\sum_{i=1}^{n} \mathbf{x}_i \widehat{P}(j|\mathbf{x}_i),$$

$$\widehat{\boldsymbol{\Sigma}}_j = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T \widehat{P}(j|\mathbf{x}_i),$$

kde $\widehat{P}(j|\mathbf{x}_i)$ je spomínaná podmienená pravdepodobnosť. Po jej nasledovnom novom odhade pomocou $\widehat{p}_j$, $\widehat{\boldsymbol{\mu}}_j$ a $\widehat{\boldsymbol{\Sigma}}_j$ sa opäť vraciame na začiatok iteračného cyklu až do stabilizácie uvedených odhadov na základe vhodného kritéria konvergencie. Poprípade sa tento iteračný proces tiež adaptuje podľa obmedzenia na kovariančnú maticu subpopulácií.

V [4] bola vyvinutá sekcia modelov konečnej zmiešanej hustoty s viacrozmerným normálnym rozdeleným zložiek v modele, v ktorom umožnujú niektoré, ale nie všetky vlastnosti rozdelenia vychádzajúcich z kovariančnej matice (orientácia, veľkosť a tvar) sa meniť medzi zhlukmi. Zároveň obmedzovali ostatné, aby boli rovnaké. Tieto nové kritéria vyplývaju z úvahy reparametrizácie kovariančnej matice $\boldsymbol{\Sigma}_j$ z hľadiska spektrálneho rozkladu

$$\boldsymbol{\Sigma}_j = \mathbf{D}_j \boldsymbol{\Lambda}_j \mathbf{D}_j^T,$$

kde $\mathbf{D}_j$ je matica vlastných vektorov a $\boldsymbol{\Lambda}_j$ je diagonálna matica s vlastnými číslami $\boldsymbol{\Sigma}_j$ na diagonále. Orientácia osí elipsoidu určeného maticou $\boldsymbol{\Sigma}_j$ je daná pomocou $\mathbf{D}_j$, zatiaľ čo $\boldsymbol{\Sigma}_j$ definuje veľkosť a tvar kontúry hustoty. Špeciálne môžeme písať

$$\boldsymbol{\Lambda}_j = \lambda_j \mathbf{A}_j,$$

kde λ_j je najväčšia vlastná hodnota $\boldsymbol{\Sigma}_j$ a $\mathbf{A}_j = diag(1, \alpha_2, \ldots, \alpha_p)$ obsahuje pomery vlastných čísiel po rozdelení λ_j. Preto λ_j kontroluje veľkosť (objem)

j-tého zhluku a $\mathbf{A}_j$ jeho tvar. V dvojrozmernom prípade by parametre odrážali pre každý zhluk koreláciu medzi dvomi premennými a veľkosť ich štandartných odchyliek.

V následujúcej tabuľke vidíme sériu modelov zodpovedajúcich rôznym obmedzeniam uložených na kovariančnej matici.

Model	Distribúcia	Objem	Tvar	Orientácia
EII	sférická	rovnaký	rovnaký	—
VII	sférická	rôzny	rovnaký	—
EEI	diagonálna	rovnaký	rovnaký	podľa súradnicových osí
VEI	diagonálna	rôzny	rovnaký	podľa súradnicových osí
EVI	diagonálna	rovnaký	rôzny	podľa súradnicových osí
VVI	diagonálna	rôzny	rôzny	podľa súradnicových osí
EEE	eliptická	rovnaký	rovnaký	rovnaký
EVE	eliptická	rovnaký	rôzny	rovnaký
VEE	eliptická	rôzny	rovnaký	rovnaký
VVE	eliptická	rôzny	rôzny	rovnaký
EEV	eliptická	rovnaký	rovnaký	rôzny
VEV	eliptická	rôzny	rovnaký	rôzny
EVV	eliptická	rovnaký	rôzny	rôzny
VVV	eliptická	rôzny	rôzny	rôzny

Tabuľka 3.6: Model

Označenia modelov opisujú obmedzenia modelu na objeme λ_j, tvare $\mathbf{A}_j$ a orientácií $\mathbf{D}_j$. Jednotlivé písmená v názve modelu znamenajú V = premenlivý medzi subpopuláciami (zhlukmi), E = rovnaký pre všetky zhluky a I = kovariančná matica je rovná jednotkovej matici pri všetkých zhlukoch.

Výber modelu je kombinácia výberu vhodného zhlukovacieho modelu pre populáciu, z ktorej sme vybrali n pozorovaní a optimálny počet zhlukov. Optimálny počet zhlukov sa dá zistiť pomocou nasledujúcich metód, ktoré nazývame Akaikovo informačné kritérium (AIC) a Bayesovo informačné kritérium (BIC). AIC kritérium je definované vzťahom

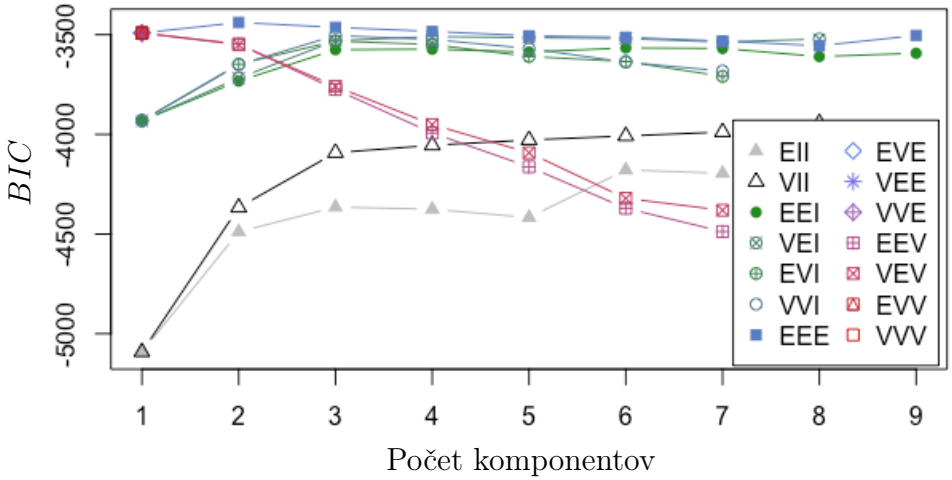
$$AIC = -2l + 2p,$$

kde l je a p je počet parametrov v modele. Podobne je definované aj BIC krité-

rium

$$BIC = -2l + p \ln n.$$

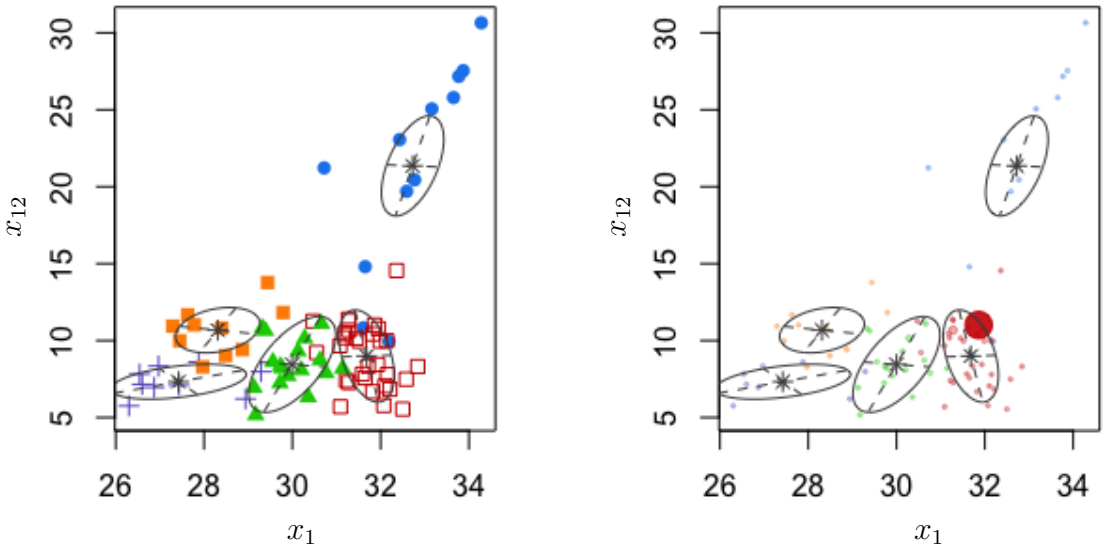
Je každopádne podstatné, že obidve kritéria zahrňujú tiež vplyv počtu parametrov, ktoré pri použití všeobecnejšieho modelu rýchle narastá. To potom býva pri voľbe takéhoto modelu limitujúce, i keď vedú k lepšiemu odhadu parametrov. Optimálny model nájdeme pomocou minimálnej hodnoty kritéria AIC alebo BIC (v praxi väčšinou berieme opačné hodnoty kritéria, keď hľadáme naopak maximálnu hodnotu). Podľa [11] je pomocou kritéria BIC možno odhadnúť lepšie modely ako pomocou AIC. Preto na ďalšom obr. 3.8 budeme brať do úvahy všetkých 10 hlavných modelov z tabuľky 3.3. Pripomenieme, že E v modele znamená rovnosť pre všetky zhluky, V variabilitu a I identitu. Teda napr. model EEI znamená diagonálnu kovariančnú maticu s rovnakým objemom a tvarom. Značka VEV indikuje elipsoidný model s rovnakým tvarom pre všetky zhluky, ale s premenlivou orientáciou a veľkosťou. Na základe tohto môžeme konštatovať, že viac obmedzené modely dosiahnú oveľa väčšie BIC hodnoty pre vyšší počet zhlukov, pretože neobmedzené modely sú silnejšie penalizované pri odhade toľkých parametrov.



Obr. 3.8: BIC graf

Obr. 3.8 sme dostali pomocou príkazu $k = cclust(X, k = 5)$ a následne $plot(k, hull = FALSE)$. Na grafe môžeme vidieť, že aj keď sa obmedzené modely rovnako nezhodujú pre rovnaký počet zhlukov, sú menej penalizované a dosahujú

vyššie hodnoty BIC pre väčší počet zhlukov. Podľa BIC kritéria sa najvhodnejší javí model EEE, teda elipsoidný model, ktorý má rovnaký tvar,veľkosť aj distribúciu. To zrejme čiastočne odpovedá tomu, že pracujeme s normovanými premennými. Aj keď sa javí ako optimálne využitie len dvoch zhlukov, na základe predchádzajúcich zkúseností sa opäť prikláňame ku $K = 5$.



Obr. 3.9: Zaradenie zhlukov

Na obr. 3.5 je na ľavej strane graf, ktorý zobrazuje ako by sa zaradili dané objekty a na pravej strane graf, na ktorom môžeme vidieť neistoty zaradenia. Podmienená pravdepodobnosť $\widehat{P}(zhlukj|\mathbf{x}_i)$, ktorú si označíme ako z_{ik}, môžeme brať aj ako indikátor neistoty zaradenia. Čím väčšie $z_{i,\max}$ (maximálna hodnota zo všetkých z_{ik} pre objekt i) tým je istejšie zaradenie objektu do zhluku. Hodnota $1 - z_{i,\max}$ vyjadruje neistotu zaradenia. Na grafe neistoty zaradenia vidieť jednu veľkú červenú bodku. Je to tým, že v tomto grafe majú menšie body z-hodnotu väčšiu ako $0,95$ a väčšie body majú z-hodnotu menšiu ako $0,75$. Ostatné zhluky majú z-hodnotu medzi týmito dvomi hodnotami. Čím väčšia je z-hodnota, tým môžeme zkonštatovať väčšiu neistotu zaradenia. V našom prípade máme len jeden takýto bod, ktorý je nesprávne zaradený do zhluku a môže byť reprezentantom viacerých objektov. Rozdelenie do zhlukov pomcou modelového zhlukovania zariadí príkaz $m = Mclust(X, G = 5)$, kde nasledujúci príkaz $head(m\$classification)$ nám vyjadrí presné zaradenie každej premennej. Predtým bolo nutné nainštalovať

knižnicu *mclust*. Na základe modelového zhlukovania sa nám objekty zaradili do zhlukov nasledovne :

Zhluk 1 (ružová)
Bratislava I
Bratislava III
Senec

Zhluk 2 (červená)		
Bratislava II	Malacky	Košice II
Bratislava IV	Pezinok	Košice III
Bratislava V	Košice I	Košice IV

Zhluk 3 (zelená)			
Dunajská Streda	Púchov	Dolný Kubín	Poltár
Galanta	Trenčín	Kysucké Nové Mesto	Revúca
Hlohovec	Komárno	Liptovský Mikuláš	Rimavská Sobota
Piešťany	Levice	Ružomberok	Stropkov
Senica	Nitra	Turčianske Teplice	Svidník
Skalica	Nové Zámky	Zvolen	Humenné
Trnava	Topoľčany	Žarnovica	Levoča
Myjava	Zlaté Moravce	Banská Bystrica	Medzilaborce
Nové Mesto nad Váhom	Žilina	Banská Štiavnica	Snina
Partizánske	Bytča	Krupina	Košice-okolie
Prievidza	Čadca	Lučenec	Sobrance

Zhluk 4 (modrá)			
Bánovce nad Bebravou	Martin	Detva	Spišská Nová Ves
Ilava	Veľký Krtíš	Stará Ľubovňa	Trebišov
Považská Bystrica	Žiar nad Hronom	Poprad	Michalovce
Šaľa	Brezno	Prešov	Rožňava

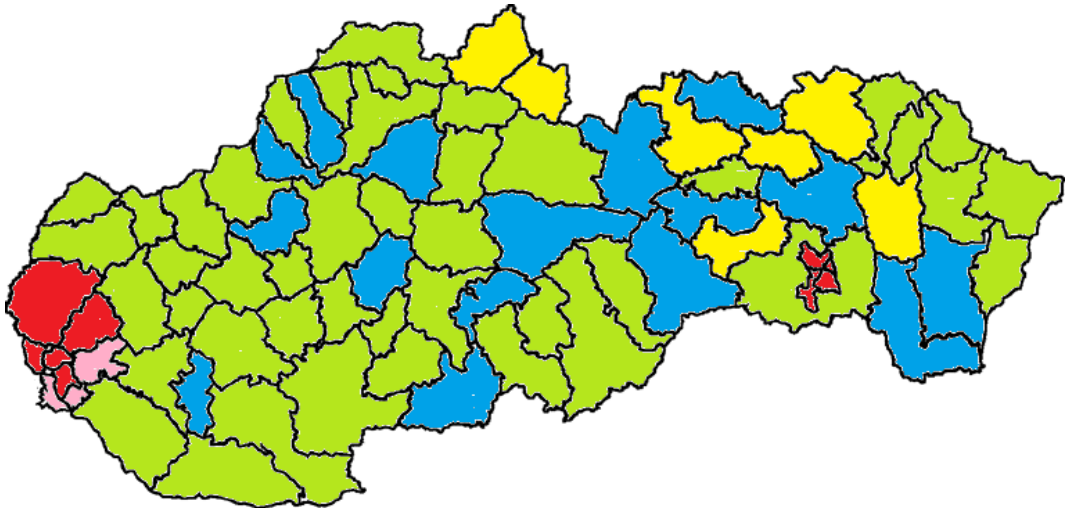
Zhluk 5 (žltá)	
Tvrdošín	Kežmarok
Námestovo	Sabinov
Vranov nad Topľou	Gelnica
Bardejov	

Priemerné hodnoty analyzovaných demografických ukazovateľov pre jednotlivé zhluky sú v nasledujúcej tabuľke:

Zhluk	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Zhluk 1	33.02	84.01	12.43	32.39	33.37	79.93	11.37	14.60	1.57	41.73	42.14	22.59
Zhluk 2	32.62	84.05	10.61	31.26	32.38	76.09	8.28	15.28	1.47	41.72	22.23	20.94
Zhluk 3	30.52	86.14	9.00	29.32	31.11	77.72	10.07	15.93	1.53	40.54	7.90	9.16
Zhluk 4	30.56	87.96	9.92	28.74	30.06	76.97	8.86	15.54	1.53	40.12	5.99	8.62
Zhluk 5	27.15	93.66	13.85	27.82	29.73	76.28	7.33	15.57	1.79	38.63	4.69	7.72

Tabuľka 3.7: Matica zhlukových centier

Na kartograme na obr. 3.10 je znázornená štruktúra zaradenia okresov do piatich zhlukov.



Obr. 3.10: Mapa zhlukov pre modelové zhlukovanie

V zhluku 1 sú zaradené dva okresy Bratislavy a okres Senec. Toto zaradenie je rovnaké ako aj v predchádzajúcich metódach. Tento zhluk je zaujímavý tým, že všetky jeho priemerne hodnoty jednotlivých ukazovateľov sú extrémne. Buď sú najnižšie alebo najvyššie. V tomto zhluku je nízky podel žien, ktoré vstupujú do

manželstva ako slobodné a krátka priemerná dĺžka trvania rozvedeného manželstva. Môžeme konštatovať, že v týchto okresoch veľa žien vstupuje do manželstva až po rozvode a veľa manželstiev sa po krátkej dobe rozpadne. V tomto zhluku je zaujímavé, že priemerná hodnota pre ukazovateľ x_6, primerný vek ženy pri úmrtí, je o 3 roky a viac vyššia ako u ostatných zhlukov. Pravdepodobne to súvisí s kvalitnejšou zdravotnou starostlivosťou vyšším vzdelaním tejto populácie a tým aj vyšším povedomímo starostlivosti o zdravie.

V zhluku 2 sú zaradené ostatné okresy Bratislavy, okresy Košíc a nejaké satelitné okresy Bratislavy. To, čo je prekvapujúce na tomto zhluku je, že nie je hodnotovo nijako výrazný ako by sme čakali keďže Košice a Bratislava sú najväčšie mestá na Slovensku. Tento zhluk má priemerné hodnoty vo všetkých ukazovateľoch. Vzhľadom na zhoršené životné prostredie by sme mohli očakávať, že v tomto zhluku bude vek ženy pri úmrtí najvyšší. Rozhodujúce v tomto ohľade sa však pravdepodobne javí vynikajúce sociálno-ekonomické prostredie.

V zhluku 3, ktorý tvorí 55.7% okresov, vidíme extrémne málo narodených žien na počet obyvateľov. Avšak naopak trvanie manželstva pred rozvodom je dlhšie ako u ostatných zhlukov. Ako bolo už spomenuté, patria tu regióny, v ktorých populácia uznáva kresťanské princípy.

V zhluku 4 sú zaradené okresy ktoré nevykazujú extrémne priemerné hodnoty sledovaných ukazovateľov. Tieto regióny sa nachádzajú v lokalitách, kde dominuje poľnohospodárstvo nad priemyslom. Hustota obyvateľstva je tu nižšia v porovnaní s ostatnými okresmi SR.

V zhluku 5 máme len 7 okresov. Skoro všetky priemerné hodnoty ukazovateľov sú extrémne. Pre tieto okresy je príznačné, že majú vysokú hodnotu podielu žien, ktoré vstupujú do manželstva ako slobodné, veľa narodených žien napočet obyvateľov a vysokú hodnotu priemerného počtu maloletých detí v rozvedenom manželstve. Zo sociálno-ekonomického aspektu patria medzi najchudobnejšie okresy SR. Ako môžeme vidieť podla kartogramu táto metóda okresy vôbec nerozdelila dobre na rozdiel od predchádzajúcich metód. Modelové zhlukovanie spravilo jeden veľký zhluk so 44 okresmi, ostatné zhluky neboli vrámci SR jasne ohraničené.

Kapitola 4

Porovnanie aplikovaných metód

Jedným z cieľov bakalárskej práce bolo posúdiť, ktorá z uvedených nehierarchických metód zhlukovej analýzy je pre analyzované demografické údaje najlepšia. Treba zdôrazniť, že výber najlepšej metódy sa nedá potvrdiť objektívnym kritériom. To môže byť spôsobené tým, že hodnotíme údaje, pri ktorých nepoznáme príslušnosť do zhlukov. Najlepšiu metódu teda vyberieme podľa subjektívneho kritéria. Nechceme, aby sa okresy v zhlukoch reťazili a v neposlednom rade chceme výsledné zhluky interpretovať.

V práci sme aplikovali tri rôzne metódy nehierarchického zhlukovania. Vybrali sme metódy K-priemerov, K-medoidov a modelové zhlukovanie. Pri všetkých týchto metódach je nutné poznať požadovaný počet zhlukov. Tento počet bol navrhnutý využitím hierarchických aglomeratívnych metód zhlukovania. Z nich bola pomocou aglomeratívneho koeficientu ac určená najvhodnejšia metóda, Wardová metóda. Pomocou Wardovej metódy sme identifikovali päť zhlukov.

Metódou K-priemerov boli okresy Slovenskej republiky klasifikované do piatich, približne homogénnych zhlukov. Podľa výsledného kartogramu Slovenskej republiky sa nám táto metóda javila ako najlepšia, z dôvodu rozdelenia objektov do homogénnych zhlukov. Zhluky, vytvorené aplikovaním metódy K-priemerov, boli dobre impretovateľné. Môžeme konštatovať, že táto metóda zaradila okresy do zhlukov aj podľa sociálno-ekonomických aspektov regiónov.

Druhá metóda, ktorá bola aplikovaná na demografické ukazovatele, bola metóda K-medoidov. Pomocou tejto metódy sme získali zaujímavé výsledky, avšak

môžeme povedať že metóda K-medoidov horšie zatriedila zhluky ako predchádzajúca. Pomocou siluétového grafu sme zaznamenali viacero zle zaradených okresov.

Ďalšou metódou, pomocou ktorej sme klasifikovali objekty, bola metóda modelového zhlukovania. V tejto metóde sme využili pri rozhodovaní o najvhodnejšiom modele kritérium BIC. Z grafu BIC bol zistený najvhodnejší model EEE. Je to elipsoidný model ktorý ma rovnaký tvar, veľkosť aj distribúciu. V tomto prípade bolo zhlukovanie nie celkom jednoznačné, nakoľko sa vopred zvolených päť zhlukov čiastočne prekrývalo. Taktiež boli identifikované objekty, ktoré boli nesprávne klasifikované do zhluku. Aj po analýze výsledného kartogramu sme mohli zhodnotiť, že metóda modelového zhlukovania vhodne nezaradila okresy do zhlukov. Môže to byť dôsledkom, že výsledný model nereflektoval dobre štruktúru náhod.

Aj keď sme čakali že najvhodnejší model získame pomocou metódy modelového zhlukovania, tak sa nám najvhodnejší model javil pomocou metódy K-priemerov. Zatriedené okresy v jednotlivých zhlukoch vykazujú aj približnú geografickú príslušnosť.

Záver

Pri prieskume metód zhlukovej analýzy, z ktorých som vyberela, som bola prekvapená ich využiteľnosťou v praxi a vynikajúcou interpretáciou. Získaný prehľad je pre mňa veľkým prínosom pri štúdiu rozdielnych viacrozmerných štatistických metód.

Za prínos tiež považujem prácu so softwarom R, v ktorom som spracovávala reálne údaje pomocou rozdielnych príkazov a knižníc. Verím že moja práca bude prínosom aj pre ostatných záujemcov, ktorí sa budú chcieť dozvedieť niečo viac o problematike zhlukovej analýzy a o jej použití v software R.

Najťažšia časť tejto práce bola interpretácia výsledných grafov, popísanie a správne porozumenie jednotlivých metód. Vývoj populácie z pohľadu demografických ukazovateľov je veľmi zložitý, dlhodobo prebiehajúci proces. Ovplyvnený je celkovým charakterom sociálno-ekonomického vývoja spoločnosti v Slovenskej republike. Správna interpretácia výsledkov aplikovaných metód by si určite vyžadovala podrobnejšiu analýzu, s pohľadu ekonomických, sociálnych, geografických a regionálnych disciplín.

Podľa môjho názoru bol teda cieľ práce, vytvoriť teoretický prehľad o jednotlivých metódach nehierarchického zhlukovania a túto teóriu aplikovať na reálne dáta pomocou softwaru R, splnený.

Literatúra

[1] Datacube – Domovská stránka [online]. Dostupné z: `http://datacube.statistics.sk`

[2] Day N. E.: Estimating the Components of a Mixture of Normal Distributions, *Biometrika*, Vol. 56, 1969.

[3] Everitt , B., Hothorn , T.: *An introduction to Applied Multivariate Analysis with R.* Springer,Heidelberg, 2011.

[4] Fraley, C., Raftery, A. E. :Model-based clustering, discriminant analysis, and density estimation *Journal of the American Statistical Association*, 2002.

[5] Hasselblad, V.: Estimation of Parameters for a Mixture of Normal Distributions,*Technometrics* ,Vol. 8, 1966.

[6] Izenman, A.J.: *Modern Multivariate Statistical Techniques.* Springer, Heidelberg, 2008.

[7] Meloun, M., Militký, J., Hill, M.: *Statistická analýza vícerozměrných dat v příkladech.* Academia, Praha, 2012.

[8] *NCSS Statistical Sowtware.* Dostupné z: `https://www.ncss.com/software/ncss/upgrade/`

[9] Sarle W.S.: *Cubic Clustering Criterion*, SAS Technical Report, Cary, NC: SAS Institute Inc., 1983.

[10] *The R Project for Statistical Computing.* Dostupné z: `http://www.r-project.org`

[11] Wehrens, R.: *Chemometrics with R.* Multivariate Data Analysis in the Natural Sciences and Life Sciences, Springer, Heidelberg, 2011.

[12] Wolfe, J.: Pattern clustering by multivariate mixture analysis, *Multivariate Behavioral Research*,Vol. 5, 1970.