UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Shluková analýza výsledků parlamentních voleb

Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **Mgr. Kamila Fačevicová, Ph.D.**
Vypracoval(a): **Jakub Kotala**
Studijní program: Aplikovaná matematika
Studijní obor: Aplikovaná matematika - specializace Data Science
Forma studia: prezenční
Rok odevzdání: 2023

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Jakub Kotala

Název práce: Shluková analýza výsledků parlamentních voleb

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Kamila Fačevicová, Ph.D.

Rok obhajoby práce: 2023

Abstrakt: Bakalářská práce se zabývá aplikací shlukovacích metod na výsledky voleb do Národní rady Slovenské republiky. V teoretické části jsou popsány shlukovací metody a problematika kompozičních dat. V praktické části jsou tyto znalosti využity při analýze volebních výsledků.

Klíčová slova: Kompoziční data, shluková analýza, parlamentní volby

Počet stran: 65

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Jakub Kotala

Title: Cluster analysis of electoral data

Type of thesis: Bachelor's

Department: Department of mathematical analysis and applications of mathematics

Supervisor: Mgr. Kamila Fačevicová, Ph.D.

The year of presentation: 2023

Abstract: The bachelor thesis aims to apply cluster analysis methods on election results from Slovakia. In the theoretical part of thesis cluster methods and basics of compositional data analysis are described. This knowledge is used to analyse election results in the practical part of thesis.

Key words: Compositional data, cluster analysis, electoral data

Number of pages: 65

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně pod vedením paní Mgr. Kamily Fačevicové, Ph.D. a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Poděkování

Rád bych poděkoval paní Mgr. Kamile Façevicové, Ph.D. za odborné vedení práce, věnovaný čas a rady, které mi pomohly při tvorbě této práce.

Úvod

Tématem práce je shluková analýza výsledků parlamentních voleb. Na tyto výsledky bude nahlíženo jako na data nesoucí relativní informaci.

Tato práce má tři základní cíle. Prvním je seznámit čtenáře se shlukovou analýzou. Druhým je popsání základních principů kompozičních dat. Posledním cílem je ukázat výsledky jednotlivých shlukovacích metod na volebních datech, které popisují, jak dopadly volby do Národní rady Slovenské republiky v letech 2016 a 2020. V celé práci je kladen důraz na 5 úspěšných stran, které se v obou letech dostaly přes požadovanou hranici zisku hlasů. Zbylé strany budou označovány jako Ostatní.

Práce se skládá ze čtyř kapitol. V úvodní části je popsána problematika kompozičních dat. Následuje kapitola o shlukovací analýze, ve které jsou představeny její jednotlivé metody. Jako součást druhé kapitoly jsou uvedeny postupy, které vysvětlují, jakým způsobem se dají použít shlukovací metody na kompoziční data. Ve třetí kapitole jsou stručně popsány volby do Národní rady Slovenské republiky a také volební programy politických stran. Jsou zde mimojiné uvedeny volební výsledky. V praktické části jsou rozebrány výsledky jednotlivých shlukovacích metod. U metody k-průměrů je přidána i interpretace vzhledem k demografickým údajům.

Při analýze byl použit software R. Kód pro praktickou část je přiložen v příloze.

Kapitola 1

Kompoziční data

V praktické části se budeme zabývat volebními daty. Pro naše zkoumání nebudou však důležité absolutní informace v datech, zajímat nás budou naopak data z relativního pohledu. Je důležité zmínit, že budeme sledovat výsledky na úrovni okresů.

1.1 Motivace

Slovensko se dělí na 79 okresů. Největším okresem jsou Levice a nejmenším Banská Štiavnica [20]. Pokud bychom chtěli zjistit, ve kterém z těchto dvou okresů se ve volbách v roce 2020 více dařilo například politickému subjektu SaS, absolutní počty by nám daly zkreslenou informaci, protože závisí na počtů voličů v daném okrese. V okresu Levice získala SaS 2640 hlasů, v Banské Štiavnici pro tuto politickou stranu hlasovalo 435 voličů, přesto nemůžeme s jistotou tvrdit, že se straně dařilo více v Levicích. Problémem je rozdílná velikost okresů a s tím související i rozdílný počet odevzdaných hlasů.

Příklad 1.1 V prvním příkladě si ukážeme výsledky dvou výše zmíněných okresů. Vektor **Levice** znázorňuje výsledky v okrese Levice, vektor **BS** zase výsledky v okrese Banská Štiavnica. Přičemž hodnoty odpovídají těmto stranám: SaS, SME RODINA, OĽANO, SMER SD, ĽSNS, Ostatní. Pojmem Ostatní označujeme všechny strany kromě pěti výše zmíněných.

- **Levice** = (2640, 4943, 14099, 8766, 4225, 21616)
- **BS** = (435, 901, 1895, 1829, 1018, 2550)

Na první pohled to vypadá, že se všem stranám dařilo více v okrese Levice. Sas zde dokonce získala několikrát více hlasů než v okrese Banská Štiavnica. Na základě tohoto pozorování se může zdát, že se SaS dařilo více v Levicích. Tato úvaha však může být chybná, protože jsme nebrali v potaz rozdílný počet odevzdaných hlasů v těchto okresech, který přirozeně souvisí s velikostí okresu a tedy počtem potenciálních voličů. V okrese Levice bylo odevzdáno zhruba 6-krát více hlasů (56299) než v okrese Banská Štiavnica (8628), a proto je vhodné zkusit jiný úhel pohledu.

Lepší informace bychom dostali, pokud bychom se podívali na procentuální zisky stran v těchto oblastech. Informace, které vyčteme z absolutních a relativních počtů můžou být značně rozdílné, ne-li protichůdné.

Příklad 1.2 Tento příklad je velice podobný předchozímu příkladu. Jediný rozdíl je v tom, že ve vektorech **Levice** a **BS** budeme mít místo absolutních počtů procentuální zisky stran.

- **Levice** = (0.05, 0.09, 0.25, 0.16, 0.07, 0.38)
- **BS** = (0.05, 0.1, 0.22, 0.21, 0.12, 0.3)

Hned si můžeme povšimnout, že první hodnota v obou vektorech je stejná. To znamená, že v obou okresech si SaS vedla stejně a získala pouze 5 % všech hlasů. Tento fakt však odporuje zjištění z prvního příkladu, kde to vypadalo, že si tato strana vedla v Levicích mnohem lépe.

Účel těchto dvou příkladů je nás namotivovat, abychom se na volební data dívali z relativního pohledu, respektive abychom s nimi pracovali jako s daty kompozičními. Zřetelně jsme viděli, že sledování absolutního počtu hlasů nám dá zkreslenou informaci.

Při práci s daty s relativní strukturou se namísto rozdílu sleduje podíl mezi dvěma složkami, protože rozdíl je ovlivněn typem reprezentace. Podíly však na reprezentaci nezáleží. Fakt, že strana SME RODINA je v obou okresech přibližně 2krát lepší než SaS, vyčteme jak z absolutní, tak i z relativní reprezentace.

1.2 Základní informace

V této kapitole si představíme, co to vlastně kompozice jsou a seznámíme se s jejich nejdůležitějšími vlastnostmi. Tyto nabyté znalosti o kompozičních datech poté využijeme v praktické části.

Definice 1 *Řádkový vektor* $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)$ *se definuje jako kompozice s* D *složkami, pokud jsou všechny složky vektoru kladná reálná čísla nesoucí relativní informaci* [17].

Příklad 1.3 Jako příklad kompozice si můžeme uvést procentualní zisk politických stran ve volbách v roce 2020 v okresu Levice. Jedná se o kompozici s šesti částmi.

- **Levice** = (0.05, 0.09, 0.25, 0.16, 0.07, 0.38)

Tradičně se předpokládá, že kompozice mají pevně daný součet složek roven konstantě κ. V příkladu výše byla κ rovna jedničce. Kdybychom měli $\kappa = 100$, hodnoty pozorování by byly vyjádřeny v procentech.

Při práci s kompozičními daty jsou důležité tři základní principy [5]:

- **Invariance vůči změně měřítka:** Pří práci s kompozičními daty nezáleží v jakých jednotkách jsou jednotlivé části kompozice vyjádřeny. Pokud bychom měli jiný součet κ, informace o poměrech mezi jednotlivými částmi kompozice by se nezměnila, a tedy i výsledek analýzy by byl stejný.

- **Invariance vůči pořadí:** Při analýze dat nezáleží na pořádí jednotlivých složek v kompozici. Pořadí však musí být stejné napříč všemi pozorováními a může ovlivnit interpretaci, např. v regresi.

- **Podkompoziční koherence:** Vzdálenost mezi dvěma kompozicemi je větší nebo rovna vzdálenosti mezi dvěma libovolnými podkompozicemi těchto kompozic.

Definice 2 Výběrovým prostorem kompozičních dat je simplex, který definujeme jako

$$S^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}_+^D, \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Tato definice poslední dobou ztratila na své oblíbenosti kvůli předpokladu pevného součtu. Abychom mohli zavést upravenou verzi definice simplexu, musíme nejprve zavést pojem uzávěr kompozice.

Definice 3 Pro libovolný vektor s D reálnými kladnými částmi

$$\mathbf{x} = (\, x_1, x_2, \ldots, x_D) \, \in \mathbb{R}_+^D$$

se uzávěr kompozice $\mathbf{x}$ definuje jako

$$C_\kappa(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right).$$

Výsledný vektor $\mathbf{x}$ odpovídá původnímu vektoru $\mathbf{x}$ až na přeškálování, po kterém součet částí výsledného vektoru $\mathbf{x}$ je roven κ [17].

Definice simplexu, kterou jsme si uvedli výše, má jednu nevýhodu. Předpokládá, že součet všech kompozic je roven konstantě κ. Jako příklad si můžeme uvést naše volební data, v nichž bude součet hlasů v jednotlivých okresech různý. Po definování pojmu uzávěr kompozice jsme schopni zavést upravenou definici simplexu, která bere v potaz různé hodnoty konstanty κ a využívá faktu, že na její hodnotě při práci s kompozičními daty nezáleží. Za výběrový prostor kompozičních dat budeme místo S^D z Definice 2 považovat simplex $\tilde{S^D}$ definovaný jako

$$\tilde{S^D} = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}_+^D, \forall \kappa > 0 \; \exists! \lambda > 0 : \mathbf{x} = \lambda C_\kappa(\mathbf{x}) \right\}.$$

Kompoziční data nejsou konzistentní s klasickou Euklidovskou geometrií [7]. Z tohoto důvodu se pro práci s kompozicemi vytvořil nový geometrický koncept nazvaný Aitchisonova geometrie definovaná na simplexu.

1.3 Aitchisonova geometrie

Poprvé byly principy Aitchisonovy geometrie představeny v roce 1986 J. Aitschisonem v [1]. V euklidovské geometrii pracujeme s operacemi jako sčítání vektorů, násobení vektoru skalárem atd. Pro práci s kompozičními daty tyto operace nemůžeme použít, proto definujeme operace nové. Následující definice operací jsou převzaty z [7].

Definice 4 Perturbace kompozice $\mathbf{x} \in \tilde{S^D}$ kompozicí $\mathbf{y} \in \tilde{S^D}$ je definována jako

$$\mathbf{x} \oplus \mathbf{y} = (x_1 y_1, x_2 y_2, \ldots, x_D y_D).$$

Definice 5 Mocnění kompozice $\mathbf{x} \in \tilde{S^D}$ konstantou $\alpha \in \mathbb{R}$ je definováno jako

$$\alpha \odot \mathbf{x} = (x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha).$$

Definice 6 Aitchisonův skalární součin kompozic $\mathbf{x}$, $\mathbf{y} \in \tilde{S^D}$ je definovaný vztahem

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

Definice 7 Aitchisonova norma kompozice $\mathbf{x} \in \tilde{S^D}$ je definovaná vztahem

$$\|\mathbf{x}\|_A = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2}.$$

Definice 8 Aitchisonova vzdálenost mezi kompozicemi $\mathbf{x}$ a $\mathbf{y} \in \tilde{S^D}$ je definována vztahem

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

1.4 Souřadnice

Ke kompozičním datům je výhodné přistupovat jiným způsobem než k datům nekompozičním. Často se pro jejich analýzu používají klasické statistické metody, nicméně je důležité před použitím těchto metod, data nejprve vhodně transformovat. Kompoziční data se transformují tak, aby následně respektovala zákonitosti Euklidovské geometrie [7].

Souřadnicových transformací vhodných pro práci s kompozičními daty je mnoho. Nejznámější jsou logpodílové souřadnicové transformace, a to například aditivní (alr), centrované (clr) a isometrické souřadnice (ilr). Mezi ilr se řadí také pivotové souřadnice. Pro naši práci jsou nejdůležitější clr a ilr souřadnice, proto je v následujících podkapitolách detailněji popíšeme.

1.4.1 Centrované logpodílové souřadnice

Začneme tím jednodušším, a to clr souřadnicemi. Idea této transformace je založená na využití geometrického průměru. Kompozici $\mathbf{x} \in \tilde{S^D}$ můžeme vyjádřit ve tvaru:

$$\mathbf{y} = \mathrm{clr}(\mathbf{x}) = (y_1, \ldots, y_D) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{k=1}^{D} x_k}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{k=1}^{D} x_k}} \right).$$

Toto vyjádření kompozice nazýváme centrované logpodílové souřadnice, v některé literatuře také centrované logpodílové koeficienty. V čitateli je hodnota dané

složky, kterou dělíme geometrickým průměrem hodnot všech složek a celý tento výraz logaritmujeme.

Použití geometrického průměru nám nezvýhodňuje žádnou proměnnou a logaritmus přináší symetrii. Zároveň z clr souřadnic můžeme ihned vyčíst zajímavé informace.

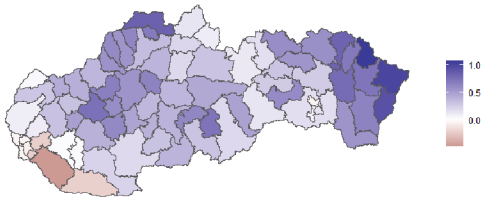
Příklad 1.4 V příkladě 1.3 obsahoval vektor **Levice** procentuální zisk stran v okrese Levice. Po transformaci hodnot na clr souřadnice dostaneme tento nový vektor:

- $\mathbf{Levice}_{CLR}$ = (-0.98, -0.39, 0.63, 0.19, -0.51, 1.05)

Ihned můžeme vyčíst, že první strana, jejíž clr souřadnice je rovna -0.98, si oproti ostatním stranám vedla špatně, protože její clr souřadnice je výrazně záporná. Naopak třetí strana dopadla v daném okresu poměrně slušně, jelikož její clr souřadnice je kladná. V případě, že by byla clr souřadnice rovna 0, tak by si strana vedla stejně jako ostatní strany v průměru (včetně té uvažované strany).

V R-ku se na výpočet centrovaných logpodílových souřadnic může použít funkce cenLR() z knihovny robCompositions.

Příklad 1.5 Dalším možným využitím clr souřadnic, je výpočet všech clr souřadnic jedné strany. V tomto příkladě si to ukážeme na straně SMER SD a jejích výsledcích ve volbách v roce 2020. Na obrázku 1.1 vidíme hodnoty clr souřadnic zakreslené do mapy ve všech okresech. Na základě tohoto obrázku můžeme tvrdit, že straně SMER SD se na její poměry dařilo ve východní části Slovenské republiky. Na jihozápadu země už její výsledky nebyly tak dobré.



Obrázek 1.1: CLR souřadnice, SMER SD, 2020

Možností, jak zapsat vztah transformované a původní proměnné, je hned několik. Pro praktické potřeby je především důležitý vztah:

$$y_1 = \frac{1}{D}\left(\ln\frac{x_1}{x_2} + \cdots + \ln\frac{x_1}{x_D}\right).$$

Z tohoto vzorce můžeme vyčíst, že hodnotu clr souřadnice můžeme interpretovat jako průměrnou dominanci složky vůči složkám ostatním.

Oproti jiným souřadnicovým transformacím dostaneme po transformaci D souřadnic. Je zřejmé, že se clr souřadnice pro jedno pozorování nasčítají na nulu. Právě kvůli této vlastnosti se o nich někdy mluví spíše jako o koeficientech než jako o souřadnicích [7].

V situaci kdybychom se chtěli vrátit zpět k původním hodnotám kompozice $\mathbf{x}$, použijeme exponencionální funkci na hodnoty $\mathbf{y}$,

$$x_j = \exp(y_j) \qquad \text{pro } \; j = 1, \ldots, D\,.$$

Důležité je však myslet na to, že získáme pouze jen jednu z reprezentací původní kompozice. Součet složek se může lišit.

V [7] jsou ukázány další vlastnosti platící pro clr koeficienty, které jsou po souřanicových systémech vyžadovány, abychom s nimi mohli dále pracovat v mnohorozměrných metodách. Pro kompozice $\mathbf{x}_1$ a $\mathbf{x}_2 \in \tilde{S^D}$ a c $\in \mathbb{R}$ platí:

$$\mathrm{clr}(\mathbf{x}_1 \oplus \mathbf{x}_2) = \mathrm{clr}(\mathbf{x}_1) + \mathrm{clr}(\mathbf{x}_2), \quad \mathrm{clr}(c \odot \mathbf{x}_1) = c \cdot \mathrm{clr}(\mathbf{x}_1),$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A = \langle \mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2) \rangle, \quad \|\mathbf{x}_1\|_A = \|\mathrm{clr}(\mathbf{x}_1)\|,$$

$$d_A(\mathbf{x}_1, \mathbf{x}_2) = d(\mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2)).$$

Díky těmto vlastnostem můžeme tvrdit, že clr transformace zachovávají všechny metrické vlastnosti, jedná se o izometrii [7] a převádí Aitchisonovu geometrii na Euklidovskou. Jak jsme si již uvedli výše, tak počet clr souřadnic je vždy D. Některé statistické metody využívají varianční matici souřadnic. V případě clr souřadnic nastane problém, jelikož sloupce matice s clr souřadnicemi jsou lineárně závislé, a proto bude varianční matice singulární. Tento fakt je velkou nevýhodou centrovaných logpodílových souřadnic a také jeden z důvodů, proč si zavedeme isometrické logpodílové souřadnice.

1.4.2 Isometrické logpodílové souřadnice

Další možností transformace je použití isometrických logpodílových souřadnic. Výhodou těchto souřadnic je fakt, že jich je po transformaci pouze D - 1 a netrpí tak nulovým součtem. Kompozici $\mathbf{x} \in \tilde{S^D}$ můžeme vyjádřit ve tvaru:

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1}).$$

Při popisu konstrukce ilr souřadnic byly čerpány informace z [6]. Myšlenka těchto souřadnic spočívá v nalezení systému D - 1 ortonormálních bázových vektorů $\mathbf{e}_i \in S^D$, pomocí kterých získáme transformované souřadnice $\mathbf{z} = (z_1, \ldots, z_{D-1}) \in \mathbb{R}^{D-1}$ dle vztahu:

$$z_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_A \quad \text{pro } \; i = 1, \ldots, D-1.$$

Každou z D - 1 ilr souřadnic můžeme alternativně zapsat jako lineární kombinaci:

$$z_i = \xi_{i1} \ln x_1 + \cdots + \xi_{iD} \ln x_D \quad \text{kde} \quad \sum_{i=1}^{D} \xi_{iD} = 0. \tag{1.1}$$

Jednou z možností, jak získat vektory $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iD})$ a vytvořit ilr souřadnice, je postupné binární dělení (SBP). Tento postup je založený na dělení kompozice na podkompozice tak dlouho, než jednotlivé podkompozice jsou tvořeny pouze jednou částí.

Před použítím SBP má kompozice D částí, proto proces dělení skončí po D - 1 krocích. V této metodě se využívá značení +, - a 0. Znaménky + a - budeme rozlišovat složky z jednotlivých podkompozic, 0 bude značit, že se složkou v daném kroku nepracujeme.

V každém kroku SBP dostaneme vektor logkontrastových koeficitentů $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iD})$. V i-tém kroku SBP odpovídá složkám z první podkompozice, ozn. +, log-kontrastový koeficient $\xi_{i+} = \sqrt{\frac{s_i}{r_i(r_i+s_i)}}$, složkám z druhé podkompozice, ozn. -, odpovídá log-kontrastový koeficient $\xi_{i-} = -\sqrt{\frac{r_i}{s_i(r_i+s_i)}}$. Číslo r_i je rovno počtu složek z první podkompozice (počet znamének +) a s_i počtu složek z druhé podkompozice (počet znamének -). Zbylé koeficienty budou mít v tomto kroce hodnotu 0.

K bázovým vektorům $\mathbf{e}_i$ se dostaneme jednoduše pomocí vztahu:

$$\mathbf{e}_i = \exp(\boldsymbol{\xi}_i).$$

Díky možnosti zapsat ilr souřadnice jako lineární kombinaci (viz vzorec 1.1) není potřeba počítat vektory $\mathbf{e}_i$ a i-tou souřadnici kompozice vypočteme pomocí vzorce:

$$z_i = \sum_{j=1}^{D} \xi_{ij} \ln(x_j) \qquad \text{pro} \;\; i = 1, \ldots, D-1.$$

Tím pádem můžeme ilr souřadnice chápat jako lineární kombinací zlogaritmovaných původních proměnných neboli logkontrasty.

Další možností je maticový zápis. Mějme matici $\mathbf{V}$ s rozměry (D - 1) $\times$ D a jejíž řádky jsou tvořeny vektory $\boldsymbol{\xi}_i$. Pro vztah mezi ilr reprezentací $\mathbf{z}$, clr reprezentací a kompozičním vektorem $\mathbf{x}$ pak platí vztah:

$$\mathbf{z} = \ln(\mathbf{x})\mathbf{V}^{\mathrm{T}} = \mathrm{clr(x)}\mathbf{V}^{\mathrm{T}}.$$

V případě, že se chceme dostat z ilr reprezentace $\mathbf{z}$ zpět na kompoziční vektor $\mathbf{x}$, tak využijeme vztah:

$$\mathbf{x} = \exp(\mathbf{z}\mathbf{V}).$$

Důležité je však upozornit na fakt, že tato rovnost platí až na uzávěr kompozice.

Z předchozích úvah vyplývá, že ilr souřadnice můžeme vyjádřit dle vztahu:

$$z_i = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ln\left(\frac{g(x_{i_1}, \ldots, x_{i_{r_i}})}{g(x_{i'_1}, \ldots, x_{i'_{s_i}})}\right) \qquad \text{pro } \; i = 1, \ldots, D-1.$$

Kde g(.) značí geometrický průměr a indexy $i_1, \ldots, i_{r_i}$ identifikují složky první podkompozice se znaménkem + a hodnoty $i'_1, \ldots, i'_{s_i}$ identifikují složky druhé podkompozice se znaménkem -. Takto zkonstruované souřadnice tedy mají podobu bilancí mezi dvěma skupinami složek, kdy je každá reprezentovaná svým geometrickým průměrem.

Příklad 1.6 Ilr souřadnice jsou flexibilní a můžeme je volit s ohledem na interpretaci. Například v případě volebních výsledků můžeme bilance volit takto, viz Tabulka 1.1. Průběh sekvenčního dělení je inspirovaný výsledkem Q-mode shlukování stran z roku 2020 z podkapitoly 2.3. Nejprve si oddělíme strany SMER SD a SME Rodina od zbylých stran, protože při Q-mode shlukování společně tvořili jednu větev dendrogramu. V dalším kroku tyto dvě strany rozdělíme.
Ve zbývajících krocích rozkládáme kompozici se zbylými čtyřmi stranami. V sedmém a osmém sloupci tabulky jsou hodnoty r_i a s_i, pomocí kterých poté získáme ilr souřadnice. Uvažujeme 6 stran (D = 6) a po transformaci proto dostaneme 5 (= D - 1) souřadnic.

Ostatní	SMER SD	ĽSNS	OĽANO	SaS	SME RO-DINA	r_i	s_i
-	+	-	-	-	+	2	4
0	+	0	0	0	-	1	1
-	0	+	-	-	0	1	3
-	0	0	-	+	0	1	2
-	0	0	+	0	0	1	1

Tabulka 1.1: Ilr souřadnice

Vektor ilr souřadnic pro okres Levice bude vypadat následovně:

- **Levice**$_{ILR}$ = (-0.17, 0.4, -0.68, -1.54, -0.3)

Z hodnoty -0.17 můžeme vyčíst, že stranám SMER SD a SME RODINA se v tomto okrese dařilo méně než zbylým stranám. Druhá hodnota je rovna 0.4, tím pádem strana SMER SD v tomto okrese předčila SME RODINU. Předposlední hodnota -1.54 značí, že OĽANO a Ostatní jasně předčily SaS.

Specifickou variantou ilr souřadnic jsou tzv. pivotové souřadnice, které představují "automatickou" volbu, nemáme-li konkrétní představu o interpretaci (viz Příklad 1.6). Vytvoří se podobným postupem, akorát při dělení první subkompozici tvoří pouze jedna složka a druhou subkompozici tvoří zbylé složky uvažované v daném kroce SBP. Tímto způsobem pokračujeme tak dlouho, dokud neprovedeme D - 1 dělení. Proměnnou, která v i-tém kroku sama tvoří první subkompozici, nazýváme pivot.

Příklad 1.7 V tomto příkladě si ukážeme, jakým způsobem se můžou vytvořit pivotové souřadnice. Postup je znázorněn v Tabulce 1.2.

SMER SD	Ostatní	ĽSNS	OĽANO	SaS	SME RO-DINA	r_i	s_i
+	-	-	-	-	-	1	5
0	+	-	-	-	-	1	4
0	0	+	-	-	-	1	3
0	0	0	+	-	-	1	2
0	0	0	0	+	-	1	1

Tabulka 1.2: Pivotové souřadnice

Vektor pivotových souřadnic pro okres Levice vypadá následovně:

- **Levice**$_{PIVOT}$ = (0.21, 1.26, -0.26, 1.11, -0.44)

Z těchto hodnot můžeme vyčíst, že SMER SD si v roce 2020 v Levicích oproti ostatním stranám vedl mírně lépe (hodnota souřadnice 0.21). Z poslední hodnoty je jasné, že straně SaS se dařilo méně než SME Rodina.

Obecný zápis pivotových souřadnic je uvedený v [7], kde j-tá souřadnice je definována vztahem:

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}} \qquad \text{pro } \; j = 1, \ldots, D-1.$$

Stejně jako clr souřadnice, tak i ilr souřadnice reprezentují izometrii [4]. Při použití ilr transformace jsou zachovány všechny metrické vlastnosti a pro kompozice $\mathbf{x}_1$ a $\mathbf{x}_2 \in \tilde{S^D}$ a c $\in \mathbb{R}$ platí tyto vztahy:

$$\text{ilr}(\mathbf{x}_1 \oplus \mathbf{x}_2) = \text{ilr}(\mathbf{x}_1) + \text{ilr}(\mathbf{x}_2), \quad \text{ilr}(c \odot \mathbf{x}_1) = c \cdot \text{ilr}(\mathbf{x}_1),$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A = \langle \text{ilr}(\mathbf{x}_1), \text{ilr}(\mathbf{x}_2) \rangle, \quad \|\mathbf{x}_1\|_A = \|\text{ilr}(\mathbf{x}_1)\|,$$

$$d_A(\mathbf{x}_1, \mathbf{x}_2) = d(\text{ilr}(\mathbf{x}_1), \text{ilr}(\mathbf{x}_2)).$$

Pro konstrukci pivotových souřadnic lze v R-ku využít funkci pivotCoord() z balíčku robCompositions. Pokud chceme ovlivnit, jak bude probíhat sekvenční dělení, můžeme použít funkci coordinates() z balíčku coda.base a pomocí parametru basis zvolíme, jak bude probíhat postupné dělení při tvorbě souřadnic. Stejnou funkci můžeme použít i pro obecné ilr souřadnice.

Kapitola 2

Shluková analýza

V praktické části bychom rádi zjistili, ve kterých okresech byly výsledky voleb podobné a ve kterých se naopak výrazně lišily. K tomuto účelu je ideální použít jednu z metod shlukové analýzy. Shluková analýza je označení pro mnohorozměrné metody patřící k průzkumové analýze dat. Cílem shlukovacích metod je vytvořit nepřekrývající se shluky s podobnými pozorováními. Pozorování z různých shluků by naopak měly být co nejrozdílnější.

Metod shlukové analýzy je mnoho. Nejprve si představíme metody hierarchického shlukování. U těchto metod není nutné dopředu vědět, do kolika shluků chceme pozorování rozdělit. Poté se budeme věnovat dělícím metodám, které vyžadují jako jeden ze svých vstupních parametrů počet shluků. Obě tyto obsáhlé skupiny metod se věnují shlukování pozorování. Další možností je tzv. Q-mode shlukování, které nevytváří shluky pozorování, ale vytváří shluky proměnných.

Kromě představení jednotlivých metod je také důležité rozeznat, které řešení nejlépe (optimálně) popisuje strukturu dat. Proto si ukážeme způsoby, jak mezi sebou shlukovací metody srovnávat.

Pro tuto kapitolu si zavedeme několik označení:

- n je počet pozorování,
- k je počet shluků,
- $\mathbf{x}_i^{(j)}$ označuje i-té pozorování přiřazené j-tému shluku,

- n_j je počet pozorování ve shluku j,
- $\mathbf{c}_j$ je centroid shluku j, který je definován jako aritmetický průměr všech pozorování z j-tého shluku, matematicky zapsáno jako: $\mathbf{c}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{x}_i^{(j)}$,
- $\overline{\mathbf{v}}$ je průměr centroidů všech shluků, matematicky zapsáno jako: $\overline{\mathbf{v}} = \frac{1}{k} \sum_{j=1}^{k} \mathbf{c}_j$,
- $\mathbf{m}_j$ je medoid shluku j, tento pojem si blíže představíme v podkapitole 2.2.2,
- C_k je množina indexů pozorování ze shluku k, neboli $C_k = 1, \ldots, n_k$.

2.1 Hierarchické shlukování

Shlukovací metody jsou založeny na shlukování podobných pozorování. U hierarchických shlukovacích metod podobnost chápeme jako vzdálenost mezi pozorováními nebo jako vzdálenost mezi jednotlivými shluky. Čím menší je vzdálenost dvou objektů, tím více jsou si podobné. Tato podkapitola vychází z [24]. Pokud nebude řečeno jinak, budeme pracovat s klasickou euklidovskou vzdáleností.

Vektory $\mathbf{x}$ a $\mathbf{y}$ mají oba n prvků. Euklidovskou vzdálenost těchto dvou vektorů budeme označovat jako d($\mathbf{x}$, $\mathbf{y}$) a platí pro ni tento vztah:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Metody hierarchického shlukování můžeme rozdělit do dvou skupin na metody aglomerativní a divisivní. Divisivní jsou výpočetně náročnější, a proto si představíme pouze ty aglomerativní.

Po celou kapitolu o shlukování budeme předpokládat, že máme n p-rozměrných pozorování, které tvoří matici $\mathbf{X}$ s rozměry n $\times$ p. V našem případě máme 79 okresů a 6 stran. Všechny proměnné jsou spojité.

Na počátku tvoří každé pozorování jeden shluk, tím pádem máme celkem n shluků. Poté si spočítáme vzdálenost mezi objekty a objekty s nejmenší vzdáleností sloučíme do nového shluku. Vzdálenosti mezi objekty (pozorováními a

shluky) přepočítáme. Tento postup opakujeme tak dlouho, dokud nemáme jeden shluk obsahující všechna pozorování.

Z tohoto postupu je zřejmé, že v průběhu výpočtu dostaneme n různých možností, jak vytvořit shluky. Přičemž první možnost (každé pozorování tvoří jeden shluk) a poslední možnost (všechna pozorování tvoří jeden shluk) nám o podobnosti pozorování nedávají žádnou relevantní informaci.

Princip aglomerativního shlukování je jednoduchý, důležité je si však ujasnit, jakým způsobem vypočteme vzdálenost mezi objekty.

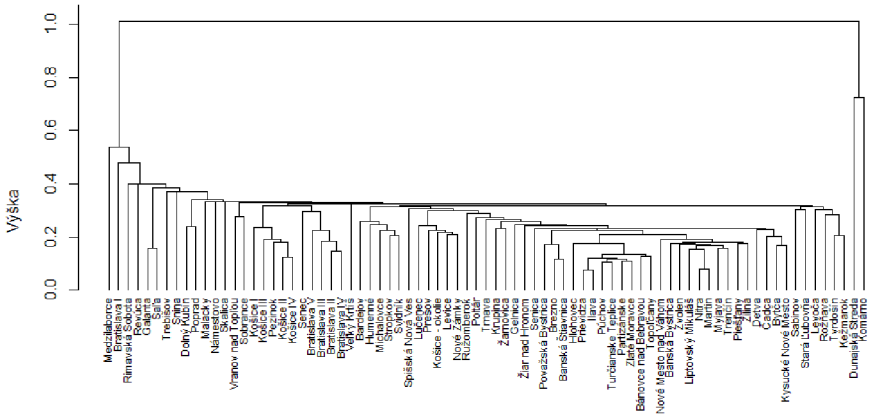
Mějme shluk S, respektive shluk P obsahující pozorování $x_i^{(s)}, \ldots, x_{n_s}^{(s)}$, respektive $x_j^{(p)}, \ldots, x_{n_p}^{(p)}$. Způsobů, jak vypočíst vzdálenost mezi těmito shluky, je mnoho a ty nejznámější si představíme na následujích řádcích.

- Metoda nejvzdálenějšího souseda (Complete linkage):
  $\max\limits_{\substack{i=1,\ldots,n_s \\ j=1,\ldots,n_p}} d(\mathbf{x}_i^{(s)}, \mathbf{x}_j^{(p)})$
- Metoda nejbližšího souseda (Single linkage): $\min\limits_{\substack{i=1,\ldots,n_s \\ j=1,\ldots,n_p}} d(\mathbf{x}_i^{(s)}, \mathbf{x}_j^{(p)})$
- Metoda průměrné vazby (average linkage): $\operatorname{average}\limits_{\substack{i=1,\ldots,n_s \\ j=1,\ldots,n_p}} d(\mathbf{x}_i^{(s)}, \mathbf{x}_j^{(p)})$
- Metoda vzdálenosti dvou středů (centroid linkage): $d(\mathbf{c}_S, \mathbf{c}_P)$
- Wardova metoda: $d(\mathbf{c}_S, \mathbf{c}_P) \cdot \sqrt{\frac{2n_s n_p}{n_s + n_p}}$

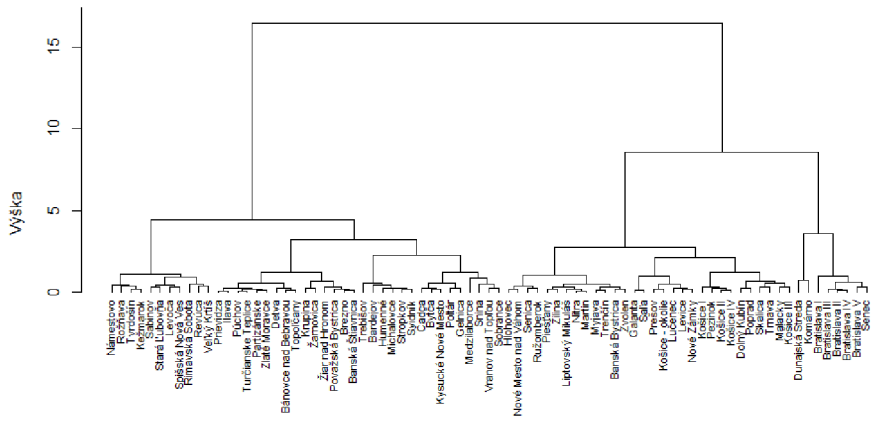
Při metodě nejvzdálenějšího souseda vypočteme vzdálenosti pro všechny možnosti dvojic pozorování ze shluků P a J. A za vzdálenost těchto dvou shluků vezmeme tu největší. Při metodě nejbližšího souseda je postup podobný, musíme však vzít naopak nejmenší vzdálenost. U metody průměrné vazby vzdálenosti zprůměrujeme a u metody vzdálenosti dvou středů počítáme vzdálenost mezi centroidy shluků. Wardova metoda je opět založena na vzdálenosti mezi centroidy shluků, kterou ještě přenásobíme faktorem zohledňujícím velikost shluků.

Jednou z možností jak graficky znázornit výsledek hierarchického shlukování je použití dendrogramu. Z tohoto grafu můžeme vyčíst, jakým způsobem se shluky vytvářely a na svislé ose najdeme informaci o vzdálenosti mezi dvěma shluky v okamžiku, kdy se spojily v jeden nový. Na obrázku 2.1 vidíme příklad dendrogramu. U tohoto příkladu byla použita metoda nejbližšího souseda. Nevýhodou této metody je řetězení malých shluků [7]. Tento úkaz se vyskytuje i u našeho příkladu. V případě, že použijeme Wardovu metodu, dostaneme samozřejmě dendrogram jiného tvaru, viz obrázek 2.2. Obecně Wardova metoda vede na kompaktnější shluky.

Naše volební data jsou relativní povahy, a proto na ně nemůžeme hned použít klasické metody. U metod hierarchického shlukování jsou dvě možnosti, jak dále postupovat. První možnost je výpočet vzdáleností s využitím Aitchisonovy vzdálenosti (viz vzorec 8) místo klasické Euklidovské vzdálenosti. Tento postup byl použit při tvorbě obrázků 2.2 a 2.1. Druhou možností je aplikování metod na data, která jsou vyjádřena maticí obsahující ilr souřadnice. Oba postupy by nám měly dát velice podobné nebo stejné výsledky.



Obrázek 2.1: Hierarchické shlukování, Metoda nejbližšího souseda, 2020



Obrázek 2.2: Hierarchické shlukování, Wardova metoda, 2020

V R-ku se pro hierarchické shlukování používá funkce hclust() z knihovny stats.

2.1.1 Volba optimálního počtu shluků

Dendrogram je skvělý způsob, jak zobrazit průběh procesu shlukování. V mnoha situacích je však požadován výstup v podobě několika shluků. Jak provést řez dendrogramu je jasné, otázkou však je, kudy řez provedeme, respektive jaký počet shluků je optimální.

Nejjednodušší a zároveň žádnou teorií podložená metoda, je metoda tzv. "kouknu a vidím". V některých případech může tento způsob dát smysluplné výsledky. Lepší by bylo mít však metody, které se opírají o strukturu dat jednotlivých shluků.

Možností jak uříznout dendrogram je mnoho, jednou z nich je využití Calinski-Harabasz indexu, který je představen v [7]. Tento index má využití i při evaluaci jiných než hierarchických shlukovacích metod. Calinski-Harabsz index má v [7] tvar:

$$CH_k = \frac{B_k/(k-1)}{W_k/(n-k)},$$

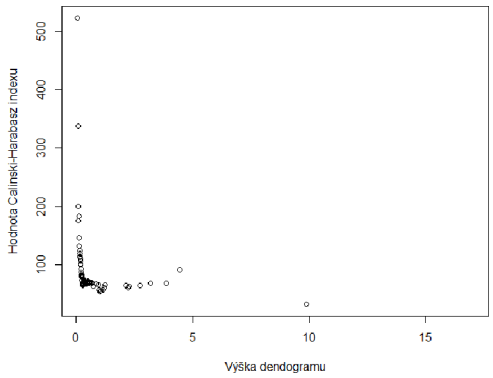
kde B_k a W_k jsou definovány následovně:

$$B_k = \sum_{j=1}^{k} d^2(\mathbf{c}_j, \overline{\mathbf{v}}), \tag{2.1}$$

$$W_k = \sum_{j=1}^{k} \sum_{i \in C_j} d^2(\mathbf{x}_i^{(j)}, \mathbf{c}_j). \tag{2.2}$$

Kritérium B_k počítá vzdálenosti mezi jednotlivými centroidy a průměrem všech centroidů. W_k se dívá na vzdálenosti mezi pozorováními a centroidy z daného shluku. Požadujeme co nejvyšší hodnotu Calinski-Harabasz indexu, protože chceme, aby jednotlivé shluky byly rozdílné (velké B_k) a zároveň aby pozorování ve shlucích byla podobná (malé W_k).

V kontextu volby optimálního počtu shluků lze využít tento index následujícím způsobem. Pro různé výšky dendrogram uřízneme a u vzniklých shluků vypočteme CH_k. Tedy pro ten počet shluků k, který vede na největší hodnotu CH_k, náš dendrogram uřežeme a se vzniklými shluky dále pracujeme. Výstup tohoto postupu jde vidět na obrázku 2.3. Podle grafu na obrázku bychom dendrogram měli uříznout ve výšce blízké nule. Pokud bychom tak udělali, výsledkem by bylo mnoho shluků s žádnou pro nás relevantní informací. Po strmém sestupu následuje mírný růst CH_k, a proto jednou z variant je zvolení výšky blízké 5.



Obrázek 2.3: Volba optimálního počtu shluků, Calinski-Harabasz index, 2020

Na výpočet Calinski-Harabasz indexu se v Rku může použít funkce calinhara() z balíčku fpc. Pokud chceme uříznout dendrogram, tak jedna z možností je funkce cutree() z balíčku stats. Jeden argumentů této funkce je výška, v které chceme dendrogram uříznout. Pokud se v této výšce spojují dva shluky, tak ve výstupu funkce budou tvořit pouze jeden shluk.

2.2 Dělící metody

Druhou rozsáhlou skupinu v rámci shlukovacích metod tvoří tzv. dělící metody. Podstatným rozdílem mezi dělícími metodami a metodami hierarchickými je počet shluků, který u dělících metod musí být dopředu zadán. Počet shluků může být menší nebo roven počtu pozorování [9]. Nejznámější dělící shlukovací metodou je metoda k-průměrů. Dále si představíme metodu k-medoidů, která je robustnější vůči odhlehlým pozorováním [7].

2.2.1 Metoda k-průměrů

Při popisu této metody bylo čerpáno z [7]. Tato metoda stejně jako metody hierarchického shlukování využívá vzdálenosti mezi objekty. Centroid je pojem, který je pro tuto metodu klíčový. Už z názvu nám může být zřejmé, že centroidů a shluků bude k. Toto číslo je vstupní parametr algoritmů inspirovaných touto

metodou. Metoda k-průměrů je založena na minimalizačním procesu, při kterém minimalizujeme součet čtverců uvnitř shluků. To znamená, že se při výpočtu snažíme minimalizovat součet vzdáleností mezi pozorováními a centroidy. Nepočítáme však vzdálenost pozorování ke všem centroidům, ale jenom k tomu centroidu, který patří ke stejnému shluku jako pozorování, viz následující vzorec.

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} d^2(\mathbf{x}_i^{(j)}, \mathbf{c}_j) \longrightarrow \min \tag{2.3}$$

Jeden z algoritmů řešících tento minimalizační problém si představíme. Jedná se o jednu z variant EM algoritmu. EM je zkratka pro anglické výrazy expectation (očekávání) a maximization (maximalizace). Je složen ze 4 kroků [24]:

1. Po zvolení hodnoty parametru k musíme inicializovat k centroidů. Za centroidy můžeme náhodně vybrat k pozorování.

2. Přiřazení pozorování k nejbližšímu centroidu. To znamená, že si pro každé pozorování spočteme $d^2(\mathbf{x}_i^{(j)}, \mathbf{c}_j)$ pro j=1, ..., k. Pozorování $\mathbf{x}_i^{(j)}$ přiřadíme do shluku, jehož centroid byl k pozorování nejblíž. (M - krok)

3. Vypočtení centroidů všech shluků s využitím vztahu $\mathbf{c}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{x}_i^{(j)}$ pro j=1, ..., k. (E - krok)

4. Pokud není splněno zastavovací kritérium, vrátíme se na krok 2.

Tento algoritmus je iterativní, a proto je důležité zvolit vhodné zastavovací kritérium. Jednu z možností volby zastavovacího kritéria si nyní ukážeme. Hodnotu minimalizované funkce ze vzorce 2.3 si označíme písmenem F. Dolní index bude označovat iteraci algoritmu. |.| je označení pro absolutní hodnotu. Zastavovací kritérium může mít tvar:

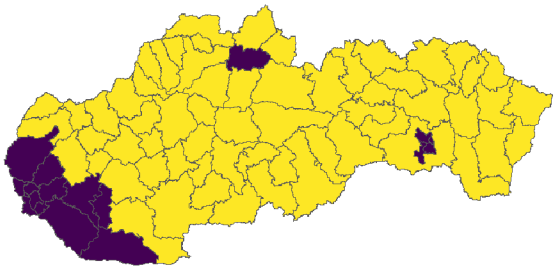
$$|F_s - F_{s-1}|.$$

Pokud hodnota zastavovacího kritéria bude menší než zadané ϵ, tak algoritmus zastavíme. Neboli sledujeme, jestli se hodnota minimalizovaná funkce dostatečně zmenšuje. Algoritmus ve většině situací zkonverguje, nalezení globálního minima však není zajištěno [24]. V prvním kroku je vybrání pozorování náhodné, na čemž závisí konečný výsledek. Nevýhodou této metody je náchylnost na odlehlá pozorování. Nejen z tohoto důvodu si představíme ještě další metodu, a to metodu k-medoidů.

Na obrázku 2.4 je příklad použití metody k-průměrů. Parametr k je roven 2. Zajímavé je, že výsledné shluky jsou souvislé, i když jsme se nezabývali geografickými vlastnostmi okresů.

Existuje několik možností jak zvolit parametr k. Můžeme využít Calinski-Harabasz kritérium představené v předchozí podkapitole. S dalšími možnostmi se setkáme v podkapitole o evaluaci metod, jelikož postupy, které se používají při hodnocení shlukovacích metod, tak můžeme použít i při volbě parametru k. U metody k-medoidů můžeme použít stejné postupy. V praktické části se na tento problém podíváme detailněji.

Při použití dělících metod na kompoziční data opět musíme myslet na specifické vlastnosti dat relativní povahy. Opět můžeme použít Aitchisonovu vzdálenost nebo pracovat s daty, které jsou reprezentované ilr souřadnicemi. Při tvorbě obrázků 2.4 a 2.5 byl použit druhý postup.



Obrázek 2.4: Metoda k-průměrů, k = 2, 2020

Obrázek 2.4 byl vytvořen pomocí funkce kmeans() z balíčku stats. Argumentem této funkce byla matice obecných ilr souřadnic vytvořených pomocí funkce coordinates() bez upřesnění báze.

2.2.2 Metoda k-medoidů

Při psaní této podkapitoly bylo čerpáno z [9]. Metoda k-medoidů je variantou známější metody k-průměrů. I pro algoritmy této metody je dopředu nutné znát hodnotu k. Nepracujeme zde s centroidy, ale s medoidy. Medoidy tentokrát budou po celý běh algoritmu některá pozorování a to taková, která nejlépe reprezentují daný shluk. Na základě jakého kritéria určíme, která pozorování jsou medoidy, si vysvětlíme později. Tato metoda se opět zakládá na minimalizačním procesu, tentokrát má funkce, kterou minimalizujeme, jiný tvar a to:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j} d^2(\mathbf{x}_i^{(j)}, \mathbf{m}_j) \longrightarrow \min. \tag{2.4}$$

Důležité je si říct, jakým způsobem přijdeme na pozorování, které bude medoid daného shluku. Medoid shluku je to pozorování, jehož součet vzdáleností k ostatním pozorováním v daném shluku je nejmenší. Mějme shluk j s n_j pozorováními. Spočteme si hodnotu výrazu pro každé pozorování, to znamená pro i=1,...,n_j:

$$\sum_{k=1}^{n_j} d^2(\mathbf{x}_i^{(j)}, \mathbf{x}_k^{(j)}).$$

I-té pozorování s nejmenší hodnotou se stává medoidem daného shluku.

Jeden z algoritmů této metody se nazývá PAM. Tato zkratka je pro anglický výraz partitioning around medoids [9], což by se dalo přeložit jako dělení kolem medoidů. Algoritmus se skládá se ze 4 kroků:

1. Po zvolení hodnoty parametru k musíme inicializovat k medoidů. Za medoidy náhodně vybereme k pozorování.

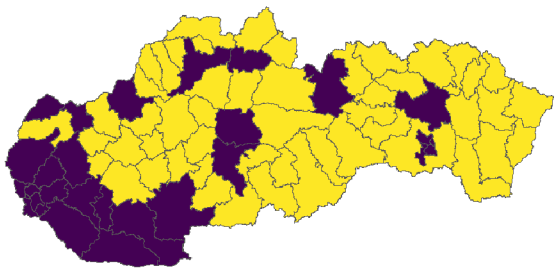
2. Přiřazení pozorování k nejbližšímu medoidu. To znamená, že si pro každé pozorování spočteme $d^2(\mathbf{x}_i^j, \mathbf{m}_j)$ pro j=1, ..., k. Pozorování $\mathbf{x}_i^j$ přiřadíme do shluku, jehož medoid byl k pozorování nejblíž.

3. Vypočtení medoidů všech shluků. Postupujume dle odstavce o výpočtu medoidu, tento postup aplikujeme na každý shluk, tedy pro j = 1, ..., k.

4. Každý medoid a každé pozorování, které není medoid, zkusíme vzájemně prohodit a sledujeme, o kolik se změní hodnota minimalizované funkce (2.4). Prohodíme medoid a pozorování, u kterého při prohození hodnota minimalizované funkce nejvíce klesla, a následně se vrátíme na krok 2. Pokud žádné prohození nezmenší hodnotu minimalizované funkce, algoritmus zastavíme.

Nevýhodou této metody je vysoká časová náročnost a to z toho důvodu, že se při čtvrtém kroku algoritmu vypočte vzdálenost každého medoidu a každého pozorování, které není medoid, musíme tedy projít k·(n-k) možností.

Na obrázku 2.5 vidíme výsledky metody k-medoidů při k = 2. Jde vidět, že oproti obrázku 2.4 nejsou výsledné shluky na mapě tak souvislé. Nicméně u obou obrázků je určitá podobnost jasně patrná.



Obrázek 2.5: Metoda k-medoidů, k = 2, 2020

Obrázek 2.5 byl vytvořen pomocí funkce pam() z balíčku cluster. Datovou matici tvořily stejně jako u metody k-průměrů obecné ilr souřadnice.

2.3 Q-mode shlukování

U hierarchického shlukování a dělících metod máme za cíl vytvořit shluky podobných pozorování. Co když nás ale zajímá, které proměnné jsou si podobné? Metody Q-mode shlukování řeší právě tento problém. Pro naše data bude zajímavé tyto metody použít. Zjistíme, které strany měly podobnou strukturu voličů a které odlišnou. Výsledkem tohoto typu shlukování je opět dendrogram. Podobně jako u hierarchického shlukování je nutné určit, jakým způsobem vypočteme vzdálenost mezi objekty.

Následující odstavec představuje variační matici, přičemž x_{i1} je označení pro první složku i-tého pozorování.

Nechť $\mathbf{X}$ *je kompoziční matice s* D *složkami a* n *pozorováními, kde* $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})^{/}$ *pro* $i = 1, \ldots, n$. *Potom prvky* t_{jk} *variační matice s rozměry* $D \times D$ *jsou definovány jako:*

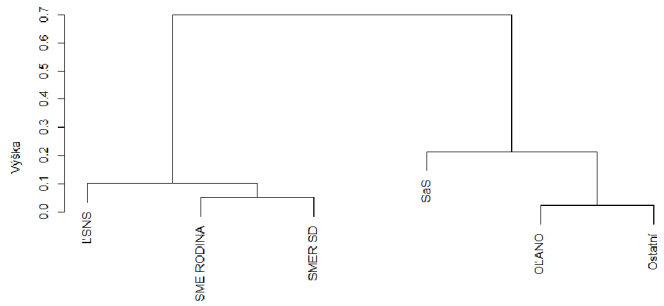
$$t_{jk} = var\left[ln\left(\frac{x_{1j}}{x_{1k}}\right), ln\left(\frac{x_{2j}}{x_{2k}}\right), \ldots, ln\left(\frac{x_{nj}}{x_{nk}}\right)\right],$$

kde j,k = 1, . . . , D a "var"je označení pro rozptyl [7].

Variační matice představená výše je symetrická a její diagonální prvky jsou rovny 0, a proto tato matice může být použita jako matice podobností pro shlukování [7].

Při použití Q-mode shlukovacích metod je důležité se rozhodnout, jakým způsobem budeme odhadovat rozptyl. Jednou z možností je použít klasický výběrový rozptyl, nebo můžeme zvolit nějaký robustní odhad, např. MCD.

Na obrázku 2.6 je výsledek Q-mode shlukování stran z roku 2020, přičemž na výpočet vzdáleností mezi objekty byla použita Wardova metoda.



Obrázek 2.6: Q-mode shlukování, Wardova metoda, 2020

V R-ku existuje příkaz variation() z knihovny robCompositions, který slouží k výpočtu variační matice. Poté se postupuje stejným způsobem jako u hierarchického shlukování, vypočte se matice vzdáleností a použije se funkce hclust() z knihovny stats.

2.4 Evaluace metod

Shlukovacích metod je nepřeberné množství. My jsme se detailněji podívali na ty nejznámější a na ty, které by nám mohli dát zajímavou informaci o našich datech. Na první pohled není jednoduché říct, která metoda je nejlepší. Obecně neexistuje metoda, která se hodí nejvíce pro všechny situace. Můžeme však zjistit, která je optimální v konkrétním případě. V této podkapitole si ukážeme hned několik způsobů, jak zjistit kvalitu našeho shlukování.

Cílem shlukovacích metod je vytvořit shluky, které budou navzájem rozdílné a zároveň pozorování uvnitř jednotlivých shluků budou podobné.

První kritérium, které si představíme, je kritérium W_k, s kterým jsme se seznámili již při popisu Calinski-Harabasz kritéria. Toto kritérium sleduje homogenitu uvnitř shluků, jeho hodnota je rovna součtu vzdáleností mezi pozorováními a centroidy a dostaneme ji pomocí vztahu:

$$W_k = \sum_{j=1}^{k} \sum_{i \in C_j} d^2(\mathbf{x}_i^{(j)}, \mathbf{c}_j). \tag{2.5}$$

Podle způsobu výpočtu je jasné, že chceme aby kritérium W_k bylo co nejmenší.

Další kritérium je kritérium B_{jl}, které sleduje rozdílnost mezi shluky j a l. Podobá se B_k z Calinski-Harabasz indexu, ale není stejné. V tomto kritériu pracujeme pouze s centroidy a jeho hodnota se vypočte jako:

$$B_{jl} = d^2(\mathbf{c}_j, \mathbf{c}_l).$$

Protože chceme, aby shluky byly co nejrozdílnější, tak požadujeme, aby vzdálenost jejich centroidů byla co největší a kritérium B_{jl} taktéž co největší.

Obě kritéria můžeme zkombinovat dohromady a dostaneme nové kritérium $V(k)$ [24]. Toto kritérium nemá žádný speciální název, je závislé na počtu shluků a sleduje jak homogenitu shluků, tak i heterogenitu všech párů shluků.

$$V(k) = \frac{\sum_{j=1}^{k} W_j}{\sum_{j<l=1}^{k} B_{jl}}$$

Z předchozích úvah víme, že chceme aby výraz v čitateli, byl co nejmenší a výraz ve jmenovateli co největší, takže celkově je pro nás nejlepší malá hodnota kritéria $V(k)$.

Další možností je tzv. průměrná šířka siluety [7]. Toto kritérium je založené na porovnání průměrné rozdílnosti pozorování od zbytku pozorování v daném shluku a průměrné rozdílnosti pozorování od všech pozorování z jiného shluku. Rozdílnost je vlastně vzdálenost, ale při popisu výpočtu kritéria budeme používat označení rozdílnost.

Průměrná rozdílnost pozorování od zbytku pozorování v daném shluku se označuje jako d_{i,C_k} a je definována jako:

$$d_{i,C_k} = \frac{1}{n_k - 1} \sum_{i,j \in C_k, i \neq j} d^2(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(k)}).$$

Průměrná rozdílnost pozorování od všech pozorování z jiného shluku si označíme jako d_{i,C_l} a je definována jako:

$$d_{i,C_l} = \frac{1}{n_l} \sum_{j \in C_l} d^2(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(l)}).$$

Nejmenší z těchto hodnot si označíme jako $d_{i,C}$.

$$d_{i,C} = \min_l d_{i,C_l}$$

Toto číslo nám ukazuje, ke kterému shluku má i-té pozorování nejblíže. Poté vypočteme tzv. siluetovou hodnotu

$$s_i = \frac{d_{i,C} - d_{i,C_k}}{max(d_{i,C_k}, d_{i,C})}.$$

Hodnoty s_i jsou z intervalu [-1, 1]. Číslo blízké 1 znamená, že i-té pozorování je dobře zařazeno, 0 znamená, že pozorování se nachází někde mezi dvěma shluky a -1 znamená špatné zařazení.

Pro porovnání různých shlukovacích metod se používá výše zmíněná průměrná šířka siluety, kterou dostaneme pomocí vztahu:

$$\frac{1}{n} \sum_{i=1}^{n} s_i.$$

Čím vyšší je hodnota, tím lepšího shlukování jsme dosáhli. Průměrnou šířku siluety můžeme taktéž použít při volbě optimálního počtu shluků.

Kapitola 3

Volební data

V praktické části se budeme zabývat daty z voleb do Národní rady Slovenské republiky v letech 2016 a 2020. Cílem této práce není podrobně rozebrat politickou situaci na Slovensku, ale naopak studovat strukturu volebních výsledků z pohledu statistiky. Pro lepší pochopení výsledků voleb jsou zde však také v krátkosti představena politická uskupení a jejich programy.

3.1 Národní rada Slovenské republiky

Národní rada Slovenské republiky je jediný ústavodárný a zákonodárný orgán Slovenské republiky. Je to orgán státní moci a od jeho primárního postavení v republice je odvozené postavení ostatních státních orgánů. Jedná se o volený orgán a reprezentuje suverenitu státu a lidu. Plní závažnou úlohu při budování Slovenské republiky jako moderního a demokratického státu a při zavádění sociální a ekologicky orientované tržní ekonomiky. Poslanci národní rady jsou voleni ve všeobecných, rovných a přímých volbách tajným hlasováním. Poslanců je 150 a jejich funkční období trvá čtyři roky [16].

Do Národní rady Slovenské republiky vstupují ta uskupení, které ve volbách získají více než 5 % platných hlasů [10]. Existuje pouze 5 politických uskupení, které v obou volbách (rok 2016 a 2020) překročily požadovanou hranici. Jedná se o SMER-SD, SaS, OĽaNO, ĽSNS a SME RODINA. V dalším textu pro nás

bude těchto pět stran nejdůležitějších. V průběhu let se jména uskupení více či méně měnila. Pro přehlednost budeme v textu používat pouze názvy subjektů z roku 2020.

3.2 Politická uskupení

Našich pět důležitých stran si teď popíšeme a představíme jejich volební program. Nepůjdeme moc do hloubky, důležité pro nás budou fakta, která by nám potom mohla pomoct s analýzou. Při popisu stran a jejich programů bylo čerpáno z: [19], [15], [21], [22], [12], [8], [13], [14] , [18] a [3].

- SMER SD - předseda Robert Fico, sociálně demokratická levicová strana, dlouhodobě velmi úspěšná strana, předseda Robert Fico považuje migraci za hrozbu pro Evropu, dříve se SMER-SD vymezoval proti stranám podporující maďarskou menšinu, při své vládě prosadili například cestování vlakem zdarma pro znevyhodněné skupiny obyvatelstva
  - před volbami v roce 2016 vládli bez koaličního partnera
  - po volbách v roce 2016 opět ve vládě
- SaS -předseda Richard Sulík, pravicová liberální strana, propagují malou roli státu v ekonomice, vymezují se proti SMER-SD, vystupují proti vstřícné migrační politice
  - před volbami v roce 2016 v opozici
  - po volbách v roce 2016 opět v opozici
- OĽaNO - předseda Igor Matovič, který se často dostává do sporů s ostatními politiky, strana působí jako hnutí, někteří členové v minulosti působili v SaS, prezentuje se jako pravicová strana, chtějí bojovat proti korupci a za zavedení hmotné odpovědnosti politiků, prozápadní orientace
  - před volbami v roce 2016 v opozici

  - po volbách v roce 2016 opět v opozici

- ĽSNS - předsedou je Marian Kotleba, extremistická strana, pravicová strana, požaduje vystoupení Slovenské republiky z NATO a EU, lidi jiné rasy nebo původu považují za hrozbu, odvolávají se na křesťanské tradice
  - před volbami v roce 2016 v opozici
  - po volbách v roce 2016 opět v opozici

- SME RODINA - tato strana vznikla až v roce 2015, předsedou je Boris Kollár, který byl již před vstupem do politiky známou osobností, chtějí chránit rodiny na Slovensku, ostře vystupují proti LGBT komunitě a sňatkům homosexuálů, odmítají přijímání imigrantů, odmítají se politicky zařadit
  - před volbami v roce 2016 v opozici, antimigrační rétorika, považují rodinu za to nejdůležitější, jedno z jejich hesel: *Nejsem politik, mi můžete věřit.*
  - po volbách v roce 2016 opět v opozici, volební program v roce 2020 pojmenovali Pomoc rodinám, chtějí zavést různé druhy finanční podpory pro rodiny s dětmi, exekuční amnestii

V následujících dvou podkapitolách se podíváme na samotné výsledky voleb. Zjistíme kolik hlasů jednotlivé strany obdržely, jaký byl jejich procentuální zisk a kolik mandátů získaly.

3.3 Výsledky voleb 2016

Volby do Národní rady Slovenské republiky se konaly v roce 2016 na počátku března. Při těchto volbách bylo odevzdáno 2 607 750 platných hlasů. Celkem odevzdalo svůj hlas 58.92 % oprávněných voličů. Drtivé vítězství si připsal SMER-SD, který získal přes 28 % všech hlasů a téměř třetinu mandátů. Na druhém místě se umístilo SaS s 12.1 % a na třetím OĽaNO s 11.02 %. Výsledky byly čerpány

z [10]. V tabulce 3.1 je kromě procentuálního zisku stran, uveden také celkový počet získaných hlasů a mandátů.

Strana	Počet hlasů	%	Mandáty
SMER-SD	737 481	28.28	49
SaS	315 558	12.10	21
OĽANO	287 611	11.02	19
ĽSNS	209 779	8.04	14
SME RODINA	172 860	6.62	11
Ostatní	884 461	33.94	36

Tabulka 3.1: Výsledné zisky politických uskupení v roce 2016

Po volbách začaly jednání o vytvoření vládnoucí koalice. Nakonec koalici tvořily strany: SMER-SD, SNS, MOST-HÍD a Sieť [10].

3.4 Výsledky voleb 2020

Při volbách do Národní rady Slovenské republiky v roce 2020 bylo odevzdáno 2 881 511 platných hlasů. Volební účast byla 65.8 %. Oproti přechozím volbám se změnilo vítězné uskupení, vítěz z roku 2016 SMER-SD získal pouze 18.29 % hlasů. Naopak OĽANO si výrazně polepšilo a získalo přes čtvrtinu všech hlasů. Téměř o polovinu méně hlasů dostalo SaS, SME Rodina si lehce polepšila a zisk ĽSNS zůstal téměř stejný. Výsledky byly čerpány z [25]. V tabulce 3.2 je kromě procentuálního zisku stran, uveden také celkový počet získaných hlasů a mandátů.

Strana	Počet hlasů	%	Mandáty
SMER-SD	527 172	18.29	38
SaS	179 246	6.22	13
OĽANO	721 166	25.02	53
ĽSNS	229 660	7.97	17
SME RODINA	237 531	8.24	17
Ostatní	986 736	34.26	12

Tabulka 3.2: Výsledné zisky politických uskupení v roce 2020

Jednání o vytvoření koalice nakonec dopadly následujícím způsobem. Koalici budou tvořit OĽaNO, SME RODINA, SaS a Za ľudí [11].

Kapitola 4

Praktická část

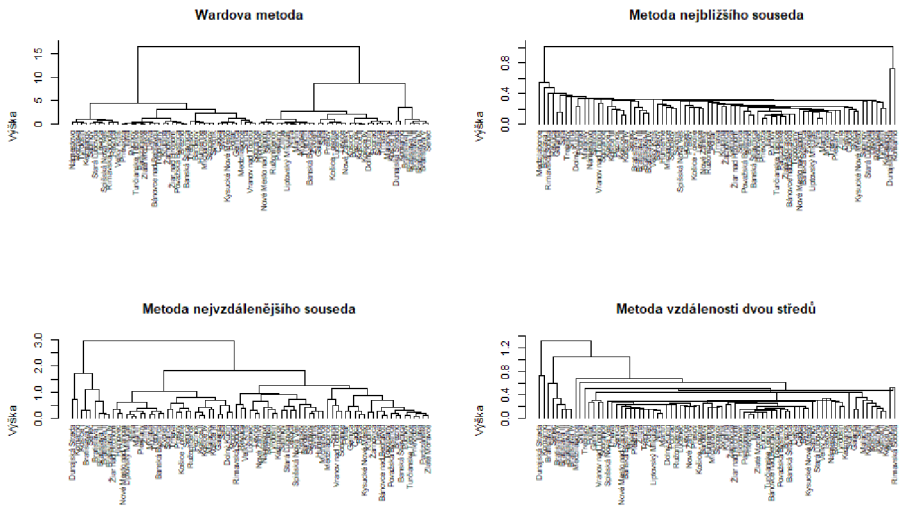
Jeden z cílů naší práce je provést analýzu volebních dat. V druhé kapitole o shlukování jsme si představili hned několik shlukovacích metod. Tyto znalosti využijeme a ukážeme si, jaké výsledky u jednotlivých metod dostaneme. U metody k-průměrů navíc zkusíme výsledné hodnoty interpretovat pomocí demografických a politických informací.

Výsledky voleb nás zajímají na úrovni okresů. Slovensko jich má 79. Při aplikaci shlukovacích metod zanedbáme počet hlasů, které strany získaly v zahraničí. V roce 2016 bylo v zahraničí odevzdáno 1044 platných hlasů a o čtyři roky později 3825 hlasů.

4.1 Výsledek hierarchického shlukování

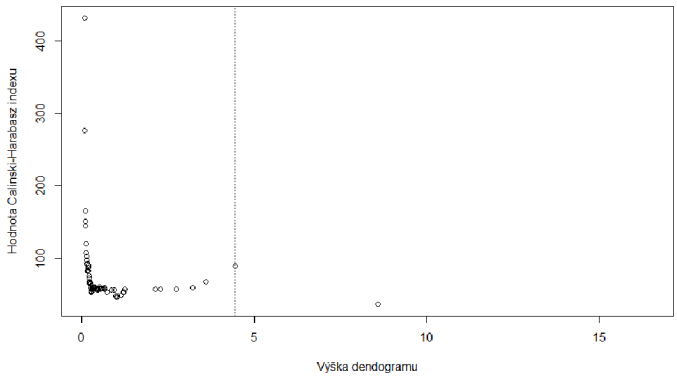
V teoretické části jsme se nejdříve zabývali hierarchickým shlukováním.
Na obrázku 4.1 vidíme výsledný dendrogram pro hierarchické shlukování s využitím čtyř různých metod pro výpočet vzdálenosti mezi shluky.

Z obrázku je patrné, že výsledky metod nejbližšího souseda a metod vzdálenosti dvou středů mají společný rys, a to řetězení shluků. Lepší informaci nám dají zbylé dvě metody.



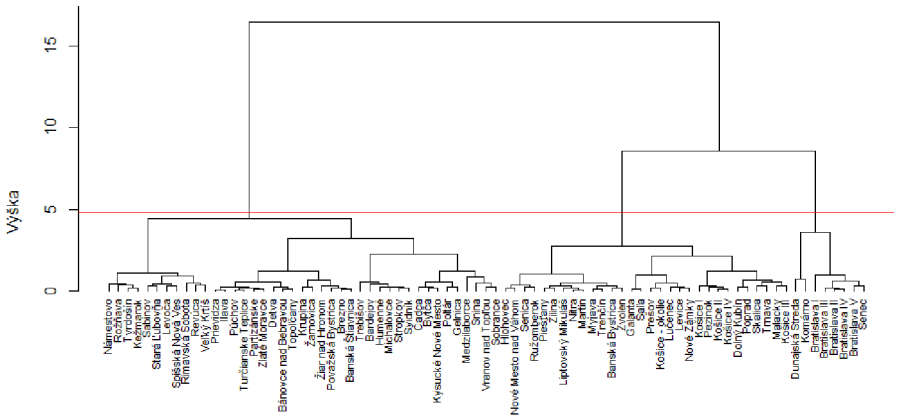
Obrázek 4.1: Hierarchické shlukování, 2020

S dendrogramem úzce souvisí problém s volbou optimálního počtu shluků. V teoretické části jsme si představili Calinski-Harabasz kritérium, které se dá použít jak při evaluaci shlukovacích metod, tak při volbě optimálního počtu shluků, čehož teď využijeme a zkusíme si to na výsledku Wardovy metody. Podle grafu na obrázku 4.2 bychom dendrogram měli uříznout ve výšce blízké nule. Pokud bychom tak udělali, výsledkem by bylo mnoho shluků s žádnou pro nás relevantní informací. Po strmém sestupu následuje mírný růst CH_k, a proto zkusíme dendrogram uříznout ve výšce blízké 5.



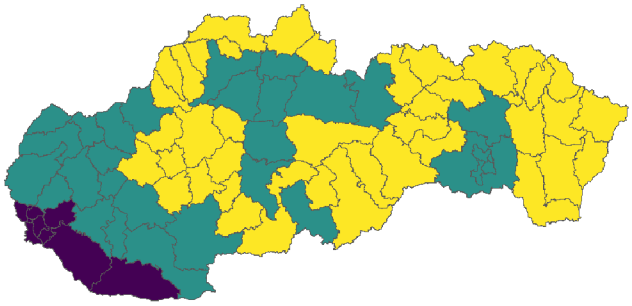
Obrázek 4.2: Volba optimálního počtu shluků, Calinski-Harabasz index, 2020

Na obrázku 4.3 je červenou čarou označen řez dendrogramu. Tento řez naše data rozdělí do třech shluků.



Obrázek 4.3: Hierarchické shlukování, volba optimálního počtu shluků, 2020

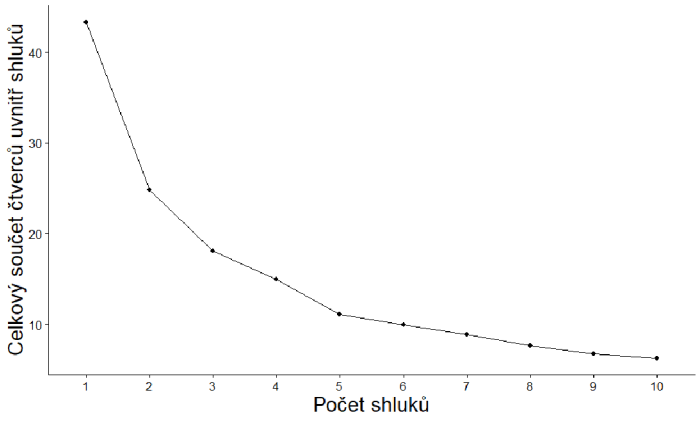
Výsledné tři shluky zakreslíme do mapy a výsledek můžeme vidět na obrázku 4.4. Tmavě modrý shluk tvoří 9 okresů, včetně 5 bratislavských. Zbylé dva shluky mají 30 a 41 okresů.



Obrázek 4.4: Hierarchické shlukování, 3 shluky, 2020

4.2 Výsledek metody k-medoidů

V podkapitole o dělících metodách jsme si představili dvě metody, tou méně známější je metoda k-medoidů. Před zobrazením výsledků je potřeba se rozhodnout, jakou hodnotu bude mít parametr k. V podkapitole 2.4 o evaluaci shlukovacích metod jsme si však představili několik způsobů, jak mezi sebou porovnat jednotlivé metody. Stejné postupy můžeme využít i nyní při volbě parametru k.

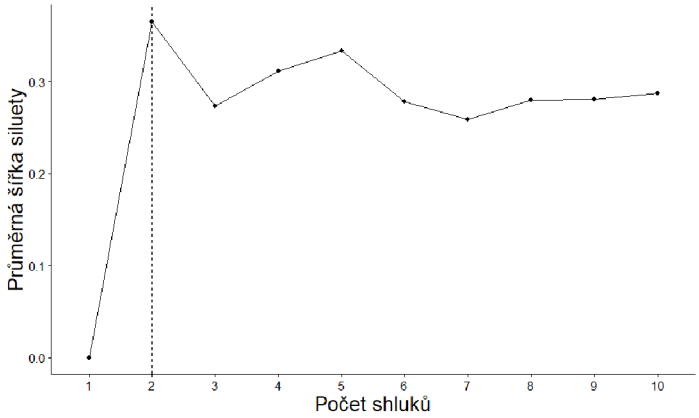


Obrázek 4.5: Metoda k-medoidů, kritérium W, 2020

Jednou z možností je využití kritéria W_j, viz vzorec 2.5, které sleduje celkovou homogenitu uvnitř shluků. Na obrázku 4.5 vidíme hodnoty tohoto kritéria

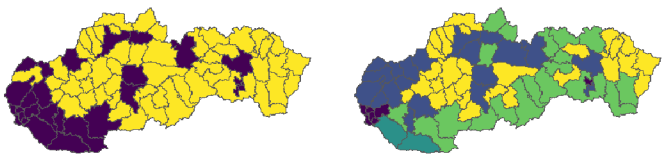
při různém počtu shluků. Pokud bychom použili tzv. pravidlo lokte, tak bychom dále pracovali s 5 shluky.

Další možností je využití průměrné šířky siluety. Z obrázku 4.6 můžeme usoudit, že dva shluky jsou nejlepší.



Obrázek 4.6: Metoda k-medoidů, průměrná šířka siluety, 2020

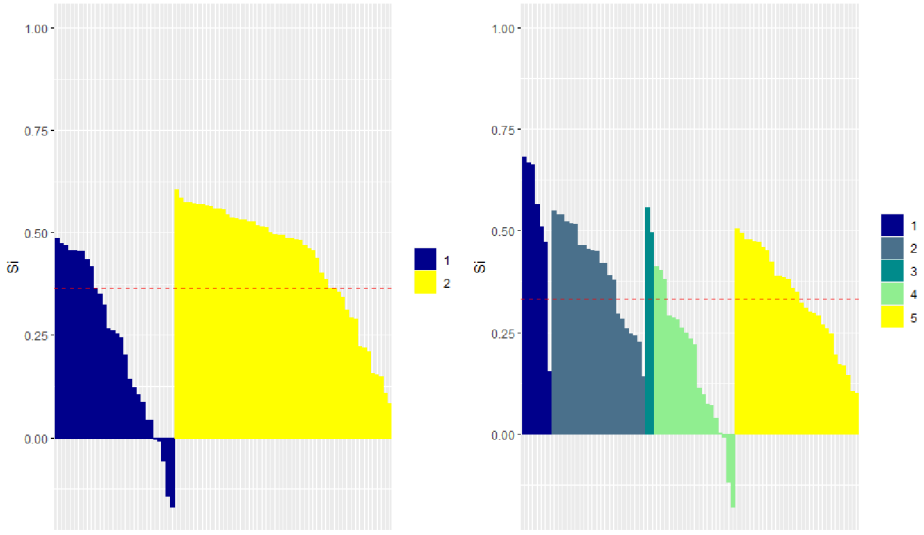
Jedna metoda nám radí se věnovat pěti shlukům, druhá pouze dvěma. Na obrázku 4.7 vidíme výsledné mapy pro k = 2 a k = 5. Z pohledu interpretace by bylo jednodušší dále pracovat s mapou při k = 2.



Obrázek 4.7: Metoda k-medoidů, k = 2 a k = 5, 2020

Dále bychom rádi srovnali, který z těchto výsledků provedl shlukování kvalit-

něji. Na obrázku 4.8 vidíme siluetové grafy. Pro k = 2 byla průměrná šířka siluety 0.36 (na grafu znázorněna červenou přerušovanou čarou), pro k = 5 byla hodnota tohoto kritéria trochu nižší, a to 0.32. Na obou grafech vidíme, že některé hodnoty s_i jsou záporné, tím pádem by se dané pozorování hodily více do jiného shluku, než do kterého byly zařazeny.

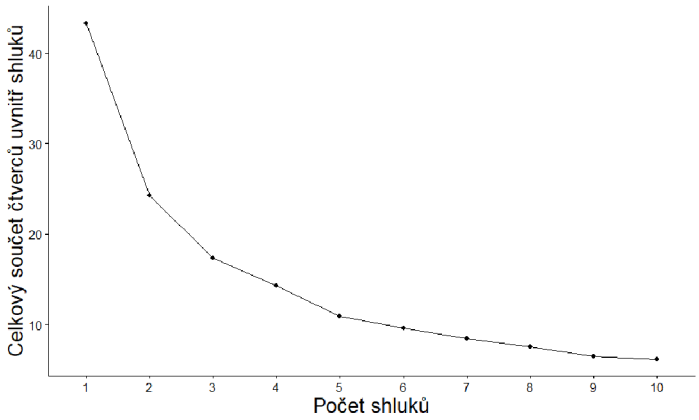


Obrázek 4.8: Siluetový graf, k = 2 a k = 5, 2020

4.3 Výsledek metody k-průměrů

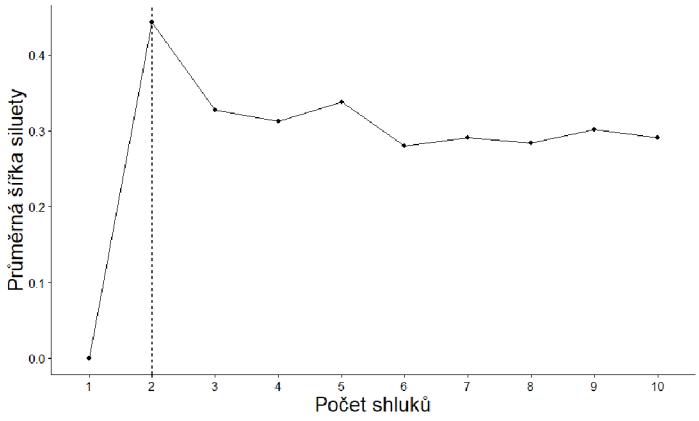
Další metoda, jejíž výsledky si představíme, je metoda k-průměrů. Nejprve si opět představíme postup při volbě parametru k. Následně si výsledky vykreslíme a zkusíme je interpretovat pomocí demografických údajů o Slovensku.

Podobně jak u kapitoly 4.2 o výsledcích metody k-medoidů tak i tentokrát využijeme kritérium W_j a průměrnou šířku siluety. Na obrázku 4.9 vidíme hodnoty W_j při různém počtu shluků. Jednou z možností je dále pracovat s pěti shluky.



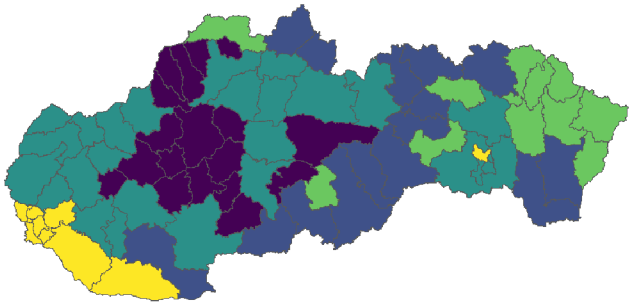
Obrázek 4.9: Metoda k-průměrů, kritérium W, 2020

Následně se podíváme i na průměrnou šířkou siluety, na obrázku 4.10 vidíme, že dva shluky by byly nejlepší.



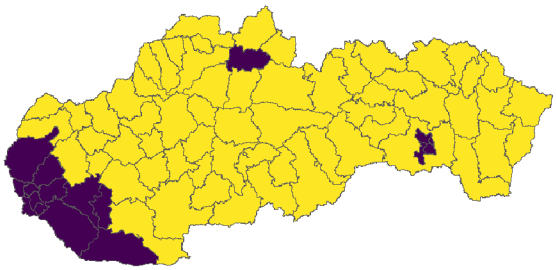
Obrázek 4.10: Metoda k-průměrů, průměrná šířka siluety, 2020

Jedna metoda nám radí se věnovat pěti shlukům, druhá pouze dvěma.
Z pohledu interpretace je pro nás lepší se věnovat shlukům dvěma. V případě pěti shluků by bylo velmi těžké něco vyčíst z výsledné mapy, viz obrázek 4.11. Dále tedy budeme pracovat s výsledkem při k=2.



Obrázek 4.11: Metoda k-průměrů, k = 5, 2020

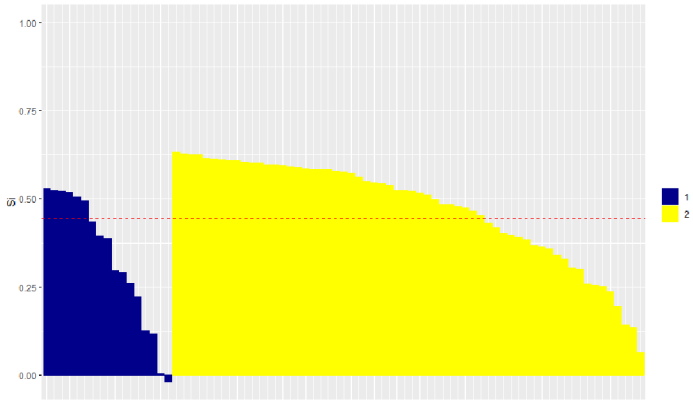
Výsledek metody k-průměrů při k=2 vidíme na obrázku 4.12.



Obrázek 4.12: Metoda k-průměrů, k = 2, 2020

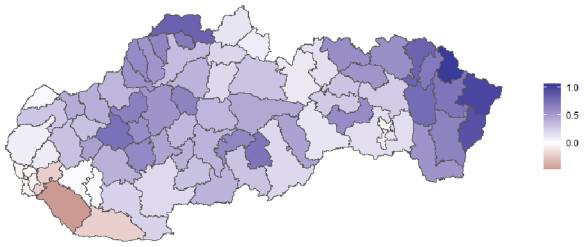
Fialový shluk tvoří 17 okresů včetně všech devíti městských okresů v Bratislavě a Košicích. Samostatný fialový okres na severu Slovenska je Dolný Kubín. Dá se nějakým způsobem toto rozdělení do shluků vysvětlit?

Nejprve se podíváme na okres Dolný Kubín. Tento okres je zcela ohraničen okresy patřící do žlutého shluku. Na obrázku 4.13 vidíme, že pouze jeden okres měl zápornou hodnotu s_i. Shodou okolností to je právě Dolný Kubín, jehož s_i je rovno -0.02. To znamená, že by měl spíše patřit do žlutého shluku. Je třeba však upozoronit, že hodnota 0.02 není velká.

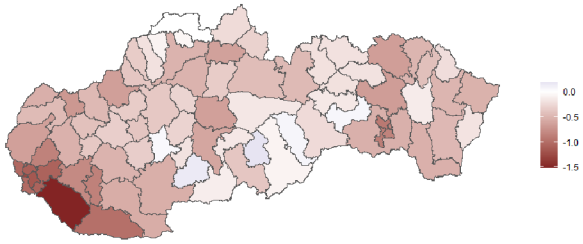


Obrázek 4.13: Metoda k-průměrů, siluetový graf, k = 2, 2020

Důležitější pro nás však bude, zkusit zjistit, které strany mohly ovlivnit toto rozdělení do dvou shluků. Zajímavou informaci nám dají clr souřadnice pro strany ĽSNS a SMER-SD. Z obrázků 4.14 a 4.15 můžeme vyčíst, že oběma stranám se na jihozápadě moc nedařilo, taktéž v Košicích nedosahovaly na své poměry žádných oslnivých výsledků. V okrese Dolný Kubín dosáhly průměrných výsledků.

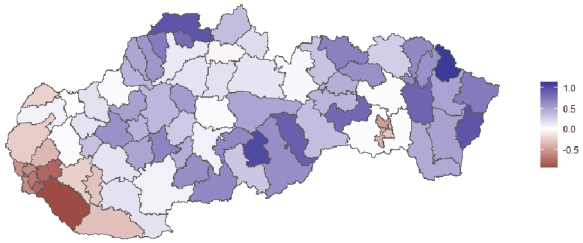


Obrázek 4.14: CLR souřadnice, SMER-SD, 2020



Obrázek 4.15: CLR souřadnice, ĽSNS, 2020

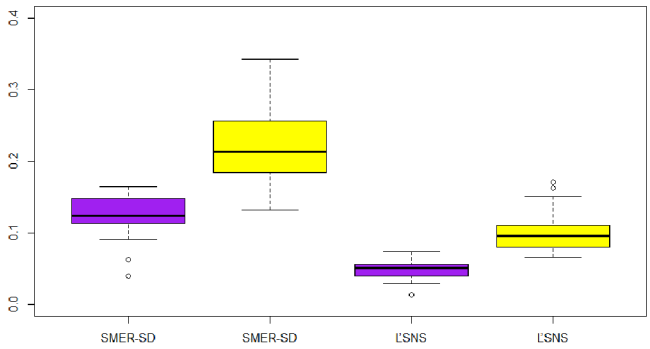
Vedle clr souřadnic jsme si v teoretické části představili taktéž ilr souřadnice. Z obrázků 4.14 a 4.15 můžeme usoudit, že výsledky stran SMER SD a ĽSNS měly podobné rysy. Toho bychom mohli využít a zkonstruovat ilr souřadnici, kdy jednu podkompozici budou tvořit tyto dvě strany a druhou podkompozici zbylé strany. Vykreslení hodnot této ilr souřadnice můžeme vidět na obrázku 4.16. Jde dobře vidět, že stranám SMER-SD a ĽSNS se oproti zbývajícím stranám na západě Slovenska nedařilo. Také v košických městských okresech zaostávali.



Obrázek 4.16: ILR souřadnice, ĽSNS a SMER-SD vs. OĽANO, SaS a SME Rodina, 2020

Další možností je přímo se podívat, jak se těmto dvěma stranám dařilo ve "fi-

alových"a "žlutých"okresech. Na obrázku 4.17 vidíme čtyři boxploty. Obě strany měly v "žlutých"okresech vyšší zisk hlasů.

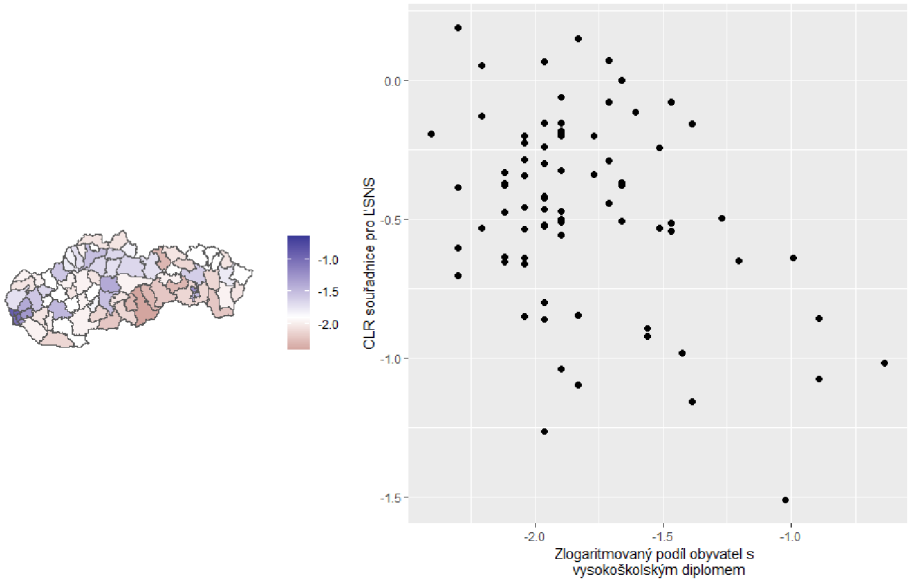


Obrázek 4.17: Boxplot procentuálního zisku v jednotlivých shlucích, ĽSNS a SMER-SD, 2020

Z úvah výše můžeme usoudit, že rozdělení okresů do dvou shluků mohlo být ovlivněno výsledky stran SMER-SD a ĽSNS. Pokud připustíme tuto možnost, měli bychom si položit otázku, proč výsledky těchto stran vypadají právě tímto způsobem? Nesouvisí to s demografickými údaji v těchto okresech?

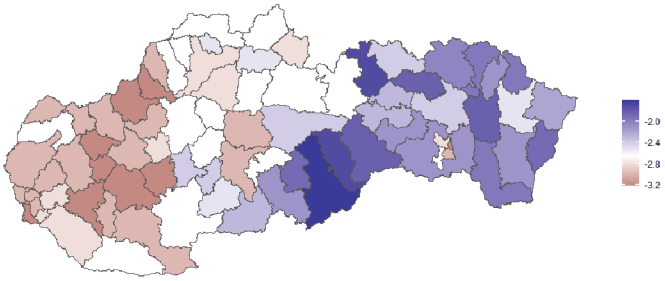
O straně ĽSNS víme, že je to extremistické hnutí a mohlo by nás napadnout, že jejich stoupenci můžou být občané s rebelujícími sklony. Může to znamenat, že ji volí občané s nižším stupněm dosaženého vzdělání? Proto se podíváme na data ohledně počtu pobyvatel s vysokoškolským diplomem v jednotlivých okresech.

Data s demografickými informacemi byla čerpána z [2] a údaje o vzdělání jsou z roku 2021. Předpokládámě však, že tyto údaje se během pár let drasticky nemění. Při zkoumání obrázku 4.18 si můžeme povšimnout určitých znaků, které by mohli nahrávat teorii, že straně ĽSNS se moc nedařilo v okresech s vyšším podílem obyvatel s vysokoškolským vzděláním (Košice, Bratislava). V bodovém grafu si můžeme povšimnout mírného trendu, kde se vzrůstajícím podílem obyvatel s vysokoškolským diplomem klesá hodnota CLR souřadnice pro ĽSNS.



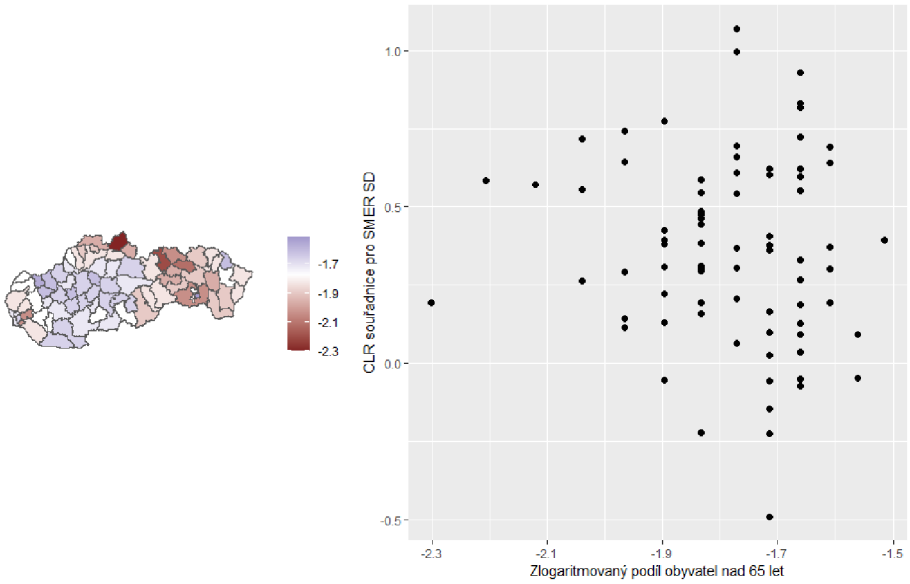
Obrázek 4.18: Zlogaritmovaný podíl obyvatel s vysoškolským vzděláním v okresech, 2021

Dalším ukazatelem, kterým se budeme zabývat je nezaměstnanost. Na obrázku 4.19 vidíme, že Slovensko je z pohledu nezaměstnanosti rozdělené na dvě poloviny. Výjimkou jsou Košice, kde nezaměstnanost je podobně nízká jako na západě země. Jestli tato informace nějakým způsobem ovlivnila výsledky stran ĽSNS a SMER-SD, to nelze s jistotou říci.



Obrázek 4.19: Zlogaritmovaný podíl nezaměstnaných obyvatel v okresech, 2020

V [23] je napsané, že SMER-SD podporuje důchodce. Obrázek 4.20 s podílem obyvatel nad 65 let je poměrně rozmanitý. Jasnou spojitost s výsledky strany SMER-SD tam těžko najdeme, taktéž v bodovém grafu není vidět nějaký trend. Ale například v středozápadní části Slovenska, kde žilo poměrně hodně obyvatel nad 65 let, byla tato strana relativně úspěšná. Ale naopak na východě země, kde se straně dařilo nejvíce, tam tolik starších obyvatel nežilo.



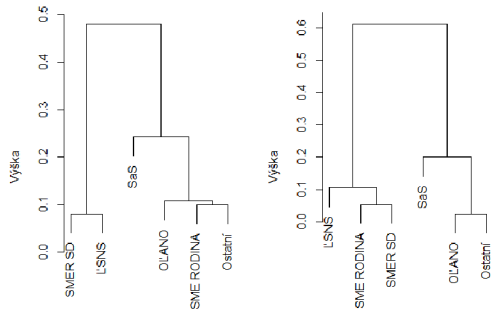
Obrázek 4.20: Podíl obyvatel nad 65 let v okresech, 2020

Výsledek metody k-průměrů pro k = 2 byl ovlivněn výsledky stran ĽSNS a SMER-SD. Zisky stran v rámci Slovenska se mohly odvíjet od demografických údajů jednotlivých okresů, např. obyvatelé s vysokoškolským diplomem nebo počet obyvatel nad 65 let.

4.4 Výsledek Q-mode shlukování

Kromě pozorování, můžeme shlukovat i proměnné. V podkapitole o Q-mode shlukování jsme si ukázali výsledek této metody pro rok 2020. Nyní se podíváme i na výsledek z roku 2016 a zkusíme je porovnat.

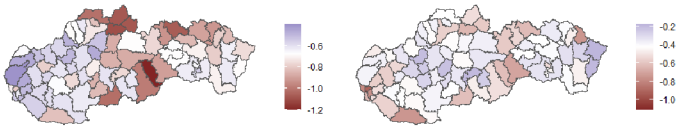
Na obrázku 4.21 vidíme, že levý dendrogram má dvě větve. Jednu tvoří dvojice SMER SD a ĽSNS. Zbylé strany tvoří druhou větev. Zajímavé je, že v roce 2020 se k této dvojici stran připojila strana SME RODINA.



Obrázek 4.21: Q-mode shlukování, Wardova metoda, 2016 (vlevo) a 2020 (vpravo)

Bylo to způsobené odlivem voličů v některých okresech? Nebo změnili svůj volební program?

Pokud porovnáme hodnoty clr souřadnic z obou let, tak na obrázku 4.22 můžeme vidět, že obě mapy nejsou rozhodně stejné, ale ani naopak nejsou výrazně jiné.



Obrázek 4.22: CLR souřadnice, SME Rodina, 2016 a 2020

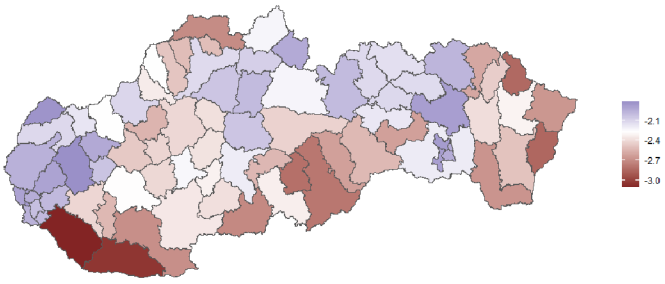
V podkapitole 3.2 o politických uskupeních jsme si blíže představili i tuto stranu. SME RODINA je založená na pomoci rodinám, vymezuje se proti migraci a její předseda je známá osoba. Na první pohled není zřejmá výrazná změna v jejich volebním programu.

Na otázku, zda SME RODINA změnila svůj volební program, neumíme odpovědět. Výstup shlukování proměnných nám ukázal, že nejenom shlukování pozorování nám může dát zajímavou informaci.

4.5 Vzestup OĽANA

Kromě výsledků shlukovacích metod se ještě podíváme na vzestup strany OĽANO. Ve volbách v roce 2016 získala 11.02 % všech hlasů. Kromě vítězné strany SMER-SD, byla i SaS v tomto roce lepší. O čtyři roky později však OĽANO jasně zvítězilo se ziskem 25.02 %. Jaké okresy způsobily tento vzestup? A stalo se někde, že si naopak strana OĽANO pohoršila oproti předchozím volbám?

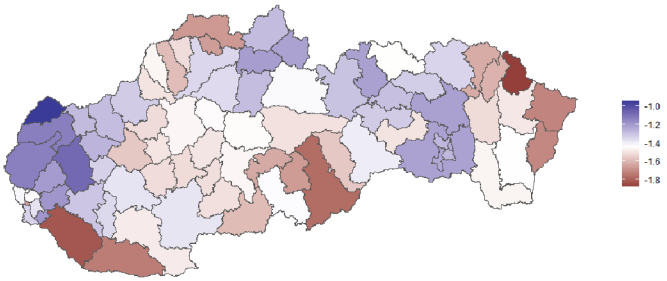
Nejprve se můžeme podívat na procentuální zisky této strany v roce 2016. Na obrázku 4.23 vidíme, že OĽANO si nejlépe vedlo na severozápadě a středovýchodě země. Ve většině okresů na jihu Slovenska se jim už tolik nedařilo, taktéž u hranic s Ukrajinou měli nízké zisky.



Obrázek 4.23: Zlogaritmovaný procentuální zisk, OĽANO, 2016

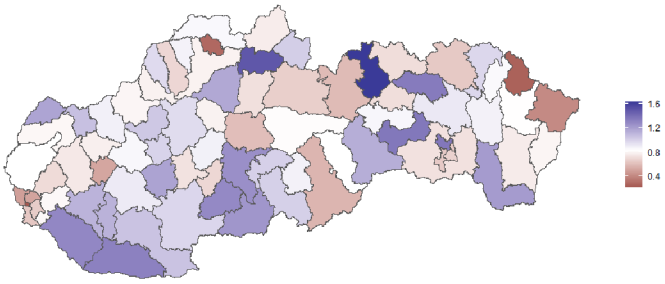
Pokud si vykreslíme výsledky této strany z voleb z roku 2020, tak vidíme,

viz obrázek 4.24, že mapa je podobně zbarvená jako na předchozím obrázku. Vypadá to, že v okresech, kde se jim tolik nedařilo v roce 2016, tak ani v roce 2020 to nebyly jejich nejlepší okresy.



Obrázek 4.24: Zlogaritmovaný procentuální zisk, OĽANO, 2020

Třetí možností je vykreslit rozdíl mezi zisky z roku 2020 a 2016. Protože naše volební data bereme jako data kompoziční, tak místo rozdílu použijeme podíl, který ještě zlogaritmujeme. Tyto hodnoty jsou vykresleny na obrázku 4.25.
Na první pohled je zřejmé, že si ve všech okresech polepšili (hodnoty jsou kladné). Největší zlepšení prokázali na jihozápadě země, kde se jim v roce 2016 nedařilo. Naopak na východě země a v bratislavských okresech nebylo zlepšení až tak výrazné.



Obrázek 4.25: Procentuální zisk, OĽANO, rozdíl mezi zlogaritmovanými zisky z let 2020 a 2016

Strana OĽANO si ve všech okresech polepšila, nutno podotknout, že největší nárust byl v okresech na jihozápadě země. Okresy z této oblasti sice nepatřily v obou letech mezi její nejsilnější, přesto v nich byl nárůst zisků znatelný. Existují však i oblasti, kde v obou letech moc podpory nezískali a progres v nich tak výrazný nebyl, např. okresy ve východní části země.

Závěr

Ukazáli jsme si, proč je výhodné se na volební výsledky dívat jako na data nesoucí relativní informaci. Při aplikaci shlukovacích metod na kompoziční data jsme mysleli na jejich vlastnosti a zjistili jsme, že v softwaru R se dá dobře pracovat i s tímto typem dat.

Shlukovacích metod existuje velké množství, my jsme si představili ty nejznámější. Při volbě počtu shluků je důležité myslet i na následnou interpretaci. Čím více shluků máme, tím je interpretace výsledků náročnější.

Výsledné shluky u metody k-průměrů byly v mapě souvislé i přesto, že jsme se vůbec nazajímali o prostorové vlastnosti okresů. Výsledek této metody mohl být ovlivněn zisky stran SMER SD a ĽSNS a také demografickými údaji, např. podílem obyvatel s vysokoškolským diplomem v jednotlivých okresech. Dále jsme si ukázali, že i shlukování proměnných může přinést zajímavé výsledky.

Přínos své práce vidím v praktické části, kde uvádím výsledky jednotlivých shlukovacích metod a přikládám svůj kód v softwaru R. Dále jsem rád, že jsem mohl spojit dvě témata (kompoziční data a shlukovou analýzu) a ukázat, že je důležité myslet na relativní povahu volebních dat.

Literatura

[1] Aitchison, J. *The statistical analysis of compositional data.* Chapman and Hall, London, 1986.

[2] *DATAcube* [online]. © Štatistický úrad Slovenskej republiky [cit. 25.2.2023]. Dostupné z:
https://datacube.statistics.sk/

[3] Dohnal, M. Nejsem politik, hlásá extravagantní milionář Kollár. Má šanci uspět. In: *deník.cz* [online]. 5.3.2016 [cit. 12.3.2023]. Dostupné z:
https://www.denik.cz/ze_sveta/nejsem-politik-hlasa-extravagantni-milionar-kollar-ma-sanci-uspet-20160305.html

[4] Egozcue J.J., Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric logratio transformations for compositional data analysis. *Mathematical Geology.* 2003, 35(3):279–300.

[5] Egozcue, J.J. Reply to "On the Harker Variation Diagrams; ..." by J.A. Cortés. *Mathematical Geosciences.* 2009, 41, 829–834.

[6] Fačevicová, K., Filzmoser, P., Hron, K. Compositional cubes: a new concept for multi-factorial compositions. *Statistical Papers.* 2022.

[7] Filzmoser, P., Hron, K., Templ, M. *Applied Compositional Data Analysis. With Worked Examples in R.* Springer Series in Statistics, 2018.

[8] Chalupková, P. (2017). *NOVÉ POLITICKÉ STRANY A VOLEBNÁ PODPORA NA SLOVENSKU V ROKU 2016.* Brno, 2017. Bakalářská práce.

Masarykova univerzita. Fakulta sociálních studií. Dostupné z: https://is.muni.cz/th/wf66j/

[9] Kaufman, L., Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.

[10] Kerekeš D., Pink M., Šedo J. Slovenské stranické zemětřesení 2016. Pomohla by malá volební reforma? [online] *Sociológia - Slovak Sociological Review* [cit. 13.3.2023]. Bratislava: Slovenská akadémia vied, 2019, roč. 51, č. 1, s. 64-83. Dostupné z: https://www.sav.sk/journals/uploads/02191108Kerekes,%20Pink,%20Sedo%201-2019.pdf

[11] Konzervatívny denník postoj. Matovič, Kollár, Sulík a Kiska majú za sebou prvé rokovanie. In: *postoj.sk* [online]. 5.3.2020 [cit. 23.2.2023]. Dostupné z: https://www.postoj.sk/52285/matovic-rokuje-s-kiskom-popoludni-zasadne-predsednictvo-za-ludi-11

[12] *ĽSNS* [online]. KOTLEBA ĽUDOVÁ STRANA NAŠE SLOVENSKO [cit. 12.3.2023]. Dostupné z: http://www.lsnaseslovensko.sk/

[13] Mareš, J. *Případová studie strany Borise Kollára z hlediska konceptu strany – obchodní firmy*. Brno, 2017. Bakalářská práce. Masarykova univerzita. Fakulta sociálních studií. Dostupné z: https://is.muni.cz/th/ad68p/

[14] Novák, M., Chripák, D. Bod zlomu na Slovensku: Suverén Fico je v ohrožení, volby slibují nevídané drama. In: *Aktuálně.cz* [online]. 27. 2. 2020 [cit. 23.2.2023]. Dostupné z: https://zpravy.aktualne.cz/zahranici/slovenske-volby-2020-pruzkumy-profily-stran/r bc4c0aa8587811eaaa180cc47ab5f122/

[15] *Odvážne už 10 rokov* [online]. OBYČAJNÍ ĽUDIA a nezávislé osobnosti [cit. 23.2.2023]. Dostupné z:
https://www.obycajniludia.sk/sme-obycajni-ludia/

[16] *Postavenie a právomoci* [online]. Národna rada Slovenskej republiky [cit. 23.2.2023]. Dostupné z:
https://www.nrsr.sk/web/?sid=nrsr/poslanie

[17] Pawlowsky-Glahn, V., Egozcue, J.J., Delgado, R.T. *Lecture Notes on Compositional Data Analysis*. 2007. [online] [cit. 12.1.2023]. Dostupné z:
http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf

[18] *PARLAMENTNÉ VOĽBY 2016 - POLITICKÉ STRANY* [online]. aktuality.sk. [cit. 23.2.2023]. Dostupné z:
https://www.aktuality.sk/volby-2016/politicke-strany/

[19] *Sloboda a Solidarita* [online]. SaS [cit. 12.3.2023]. Dostupné z:
https://www.sas.sk/

[20] *Slovensko regionální geografie* [online]. Katedra geografie Přírodovědecká fakulta UP [cit. 15.2.2023]. Dostupné z:
https://geography.upol.cz/soubory/lide/smolova/RGSR/ucebnice/index.html

[21] *SMER* [online]. SMER SLOVENSKÁ SOCIÁLNA DEMOKRACIA [cit. 12.3.2023]. Dostupné z:
https://www.strana-smer.sk/

[22] *SME RODINA* [online]. SME RODINA Boris Kollár [cit. 12.3.2023]. Dostupné z:
https://hnutie-smerodina.sk/

[23] Sviták M. Kto je kto. Průvodce slovenskými politickými stranami. In: *ceskatelevize.cz* [online]. 1.3.2020 [cit. 23.2.2023]. Dostupné z:
https://ct24.ceskatelevize.cz/svet/3051390-kdo-je-kdo-pruvodce-slovenskymi-politickymi-stranami

[24] Varmuza, K., Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics (1st ed.)*. CRC Press, 2009.

[25] *Výsledky parlamentných volieb 2020* [online]. volby.sme.sk [cit. 23.2.2023]. Dostupné z: https://volby.sme.sk/parlamentne-volby/2020/vysledky