UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Analýza archetypov, jej aplikácie a meranie polarizácie v dátach

Diplomová PRÁCA

Študijný program:	Ekonomická matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. RNDr. Beáta Stehlíková, PhD.

Bratislava 2024 **Bc. Vendelín Vincze**



29226119

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

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

Názov: Analýza archetypov, jej aplikácie a meranie polarizácie v dátach
Archetype analysis, its applications and measuring polarization in data

Anotácia: Cieľom analýzy archetypov je vyjadriť pozorovania ako konvexnú kombináciu malého počtu extrémnych bodov, tzv. archetypov. V prvej časti práce sú vysvetlené zákadné princípy, ktoré sa potom demonštrujú pri analýze vybraných reálnych dát pomocou balíka archetypes, dostupného pre jazyk R. Archetypy sú následne využité aj pri návrhu spôsobu merania polarizácie v dátach, ktorý je ďalej modifikovaný pre prípad vážených pozorovaní.

Vedúci: doc. RNDr. Beáta Stehlíková, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 10.01.2023

Dátum schválenia: 11.01.2023

prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Na tomto mieste by som sa rád poďakoval mojej vedúcej, doc. RNDr. Beáte Stehlíkovej, PhD za ochotu, ústretovosť a množstvo užitočných rád a podnetov, vďaka ktorým má táto práca podobu, akú má.

Abstrakt v štátnom jazyku

Vincze, VENDELÍN: Analýza archetypov, jej aplikácie a meranie polarizácie v dátach [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2024, 66 s.

V tejto práci predstavíme archetypálnu analýzu ako nový prístup k analýze dát. Popíšeme základnú teóriu a metódy, ktoré potom aplikujeme v rámci vzorovej analýzy reálneho datasetu, popisujúceho výsledky volieb do NR SR v rokoch 2016, 2020 a 2023. Nástroje archetypálnej analýzy sú potom použité v kombinácii s ďalšími štatistickými metódami (K-means) na získanie zmysluplného spôsobu merania polarizácie v dátach. Tento spôsob ďalej vylepšujeme a zovšeobecňujeme zahrnutím váh pozorovaní, aby sme dokázali lepšie zachytiť dynamiku dát. Základná aj všetky vylepšené verzie merania sú ďalej aplikované na skôr spomenuté dáta o výsledkoch volieb. Pri písaní tejto práce sme využívali predprogramované funkcie, zahrnuté v jazyku R a jeho balíkoch, doplnené o naše vlastné numerické výpočty.

Kľúčové slová: Archetyp, Najmenšie štvorce, Koeficient determinácie, Polarizácia, K-means

Abstract

VINCZE, VENDELÍN: Archetype analysis, its applications and measuring polarization in data [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2024, 66 p

In this thesis, Archetypal analysis is introduced as a new approach to data analysis. Basic theory and methods are described and then applied in an exemplary analysis of a real-world dataset, concerning the results of the parlamentiary elections in Slovakia in the years 2016, 2020 and 2023. Tools of Archetypal analysis are then used in combination with other statistical methods (K-means) to provide a meaningful way of meassuring polarization in data. We then further enhance and generalize this meassurment by allowing the inclusion of weights of observations, to better capture the dynamics of the data. Default and all the improved versions of the maessurment are then applied to the aforementioned election data. While working on this thesis, we use preprogrammed functions included in *R* and its packages as well as out own numerical computations.

Keywords: Archetype, Least squares, Coefficient of determination, Polarization, K-means

Obsah

Úvod **10**

1 Základy archetypálnej analýzy **12**

1.1 Motivácia a formulácia základnej optimalizačnej úlohy 12

1.2 Poloha archetypov 16

1.3 Archetypálny algoritmus a striedavá optimalizácia 16

2 Aplikácia na reálne dáta **19**

2.1 Dáta - kraje 19

2.1.1 Voľby v rokoch 2020 a 2023 26

2.2 Dáta - okresy 27

2.2.1 Voľby v rokoch 2020 a 2023 30

3 Vážené archetypy **33**

4 Meranie polarizácie v dátach **39**

4.1 Porovnanie polarizácie v rokoch 2016, 2020 a 2023 44

4.2 Zahrnutie váh 46

4.2.1 Váhy pri výpočte R^2 pre zhlukovanie 47

4.2.2 Váhy pri hľadaní zhlukovania - vážené ohraničené K-means 52

4.3 Meranie polarizácie, využívajúce vážené archetypy 59

4.4 Zhrnutie 61

Záver **63**

Zoznam použitej literatúry **65**

Úvod

Analýza dát je v súčasnosti jednou z najvýznamnejších oblastí aplikácií matematiky. Medzi najčastejšie používané metódy patria napríklad K-means (analýza zhlukov) alebo PCA (redukcia dimenzie dát). Rôzne prístupy so sebou prinášajú rôzne výhody, ale aj nevýhody, a preto sú vhodné na použitie v rozdielnych situáciách. Archetypálna analýza (ďalej AA) prináša nový pohľad na problematiku, kombinujúci niektoré z výhod PCA a podobných metód (flexibilita modelu, zachovanie informácie) s výhodami metód analýzy zhlukov (interpretovateľnosť).

Jednotlivé pozorovania sú v AA aproximované ako konvexné kombinácie malého počtu vektorov rovnakého rozmeru (archetypov), od ktorých požadujeme, aby boli samé osebe konvexnými kombináciami pozorovaní. Základným problémom, riešeným pri fitovaní modelu AA, je teda hľadanie konvexných kombinácií (koeficientov), popisujúcich samotné archetypy, ako aj tých, ktoré reprezentujú aproximácie pozorovaní, takých, aby bola minimalizovaná strata informácie. Práve podmienky, kladené na archetypy aj aproximácie pozorovaní, zabezpečujú vysokú mieru interpretovateľnosti výsledkov.

Základy AA boli položené v [2], kde sú aj veľmi zrozumiteľne vysvetlené najdôležitejšie poznatky a postupy. Tieto sú tiež kvalitne implementované v softwari R v balíku *archetypes*, detailne popísanom v [4]. Hlbší pohľad ponúka napr. [7]. Rozšírenia či úpravy základného modelu možno nájsť v [5] a [1]. Napriek množstvu práce, ktoré už bolo odvedené, sa však stále jedná o pomerne novú oblasť, výskum v ktorej aj naďalej prebieha a prináša nové poznatky.

Práca pozostáva zo 4 kapitol. Prvá bude predstavovať všeobecný úvod do problematiky, v rámci ktorého uvedieme základné poznatky a postupy AA, ktoré budeme využívať v ďalších častiach. V druhej kapitole tieto poznatky a postupy aplikujeme v rámci vzorovej analýzy reálnych dát, popisujúcich výsledky volieb do NR SR z rokov 2016, 2020 a 2023. Na tie isté dáta ďalej v tretej kapitole aplikujeme aj modifikovanú verziu AA, umožňujúcu zahrnúť do modelu váhy jednotlivých pozorovaní. Z týchto analýz vzíde otázka, ako merať polarizáciu v dátach, ktorej zodpovedaniu sa budeme venovať vo štvrtej kapitole. Uvedieme vlastný spôsob merania, ktorý budeme ďalej vylepšovať a zovšeobecňovať pridaním možnosti zahrnutia váh jednotlivých pozorovaní. Výstupy základného spôsobu merania polarizácie ako aj jeho vylepšených variantov na záver porovnáme.

Našim cieľom bude predstaviť AA ako novú a dosiaľ málo využívanú metódu analýzy dát a predstaviť možnosti jej aplikácie. Okrem toho ju využijeme na zostrojenie a postupné zdokonalovanie spôsobu merania polarizácie v dátach, v rámci ktorého ju budeme kombinovať aj s ďalšími metódami (ohraničené K-means, ohraničené vážené K-means...).

1 Základy archetypálnej analýzy

V tejto kapitole predstavíme základné poznatky zo známej teórie, týkajúcej sa archetypálnej analýzy (ďalej AA), ktoré potom využijeme v ďalších častiach práce. Uvedieme situácie, v ktorých je vhodné aplikovať popisované metódy, formalizujeme optimalizačnú úlohu, ktorú je potrebné vyriešiť, predstavíme algoritmus striedavej optimalizácie a venovať sa budeme aj polohe archetypov. Budeme tu vychádzať najmä z [2] a [4].

1.1 Motivácia a formulácia základnej optimalizačnej úlohy

Majme dataset, reprezentovaný maticou X, definovanou ako

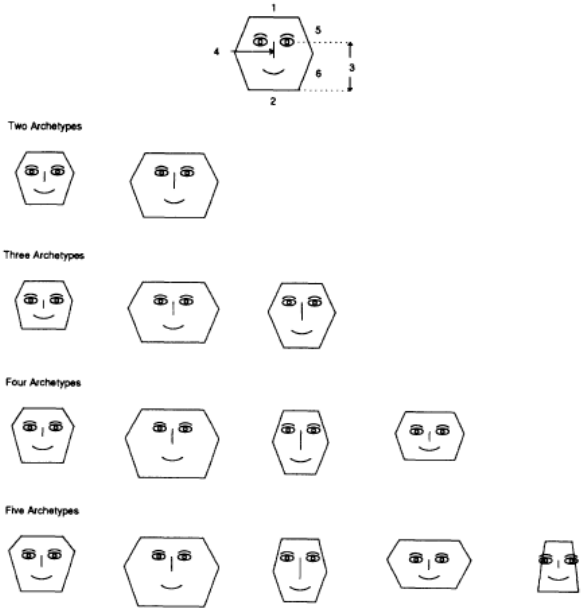
$$X = \begin{bmatrix} -x_1^T- \\ -x_2^T- \\ \vdots \\ -x_n^T- \end{bmatrix}, \tag{1}$$

kde $\{x_i \in \mathbb{R}^m,\, i = 1, ..., n\}$ sú jednotlivé pozorovania. Našim cieľom v AA je pre vopred zvolené číslo p nájsť vektory $z_1, ..., z_p \in \mathbb{R}^m$, od ktorých budeme požadovať, aby dobre popisovali dáta, a tiež, aby boli samé osebe realistické a dobre interpretovateľné. Tieto budeme nazývať archetypmi. Medzi prvými, ktorí sa zaoberali podobnou problematikou, boli Flury a Riedwyl, ktorí pracovali s datasetom, obsahujúcim 6 rozmerov tvárí 200 vojakov švajčiarskej armády. Ich snahou bolo nájsť typické, „archetypálne“ tváre, ktoré by reprezentovali extrémnych, no nie nerealistických jedincov. Merané rozmery a archetypy, nájdené pre rôzne p pomocou metód, ktoré popíšeme neskôr, možno vidieť na obr. 1, využitom aj v [2]. Všimnime si, že, hoci je každá z nájdených tvárí svojim spôsobom jedinečná a výrazne odlišná od priemeru, žiadna z nich nepredstavuje objekt (jedinca), o ktorom by sme už pri prvom pohľade vyhlásili, že nemôže existovať.

Jedným z prístupov k riešeniu takto voľne formulovaného problému je hľadať $z_1, ..., z_p$ ako vektory minimalizujúce

$$\sum_{i=1}^{n} \left|\left| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right|\right|^2 .$$

Centrovaním pozorovaní a pridaním požiadavky na ortonormalitu $z_1, ..., z_p$ dostávame úlohu, ktorej riešením je rozklad na hlavné komponenty [2]. Takýto prístup síce spĺňa požiadavku dobrého popisovania dát, ale problematická je realistickosť nájdených vektorov.



Obr. 1: Archetypy tvárí, nájdené pre rôzne p. Zdroj: [2]

Pri aplikovaní na skôr spomínané dáta, popisujúce rozmery tvárí, sa napríklad objavili také objekty, pri ktorých boli niektoré rozmery záporné. Formuláciu optimalizačnej úlohy teda bude potrebné ďalej modifikovať.

Aby sme získali archetypy s požadovanými vlastnosťami, budeme od nich požadovať, aby boli konvexnými kombináciami pozorovaní $x_1, ..., x_n$. Okrem toho budú aj samotné pozorovania aproximované ako konvexné kombinácie archetypov. Získavame tak nasledovnú optimalizačnú úlohu:

$$\min_{\alpha,\beta} \quad \sum_{i=1}^{n} \left|\left| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right|\right|^2 \tag{2}$$

$$z_k = \sum_{j=1}^{n} \beta_{kj} x_j \tag{3}$$

$$\sum_{k=1}^{p} \alpha_{ik} = 1 \tag{4}$$

$$\alpha_{ik} \geq 0 \tag{5}$$

$$\sum_{j=1}^{n} \beta_{kj} = 1 \tag{6}$$

$$\beta_{kj} \geq 0 \tag{7}$$

Cieľom je získať čo najpresnejšiu aproximáciu pôvodných dát. Ohraničenia (3), (6) a (7) zabezpečujú, že nájdené archetypy budú konvexnými kombináciami pozorovaní, zatiaľ čo pozorovania sú vďaka ohraničeniam (4) a (5) a tvaru účelovej funkcie aproximované

konvexnými kombináciami archetypov (toto nám dáva novú, p-rozmernú reprezentáciu pozorovaní, využívajúcu koeficienty α, AA teda môže slúžiť aj ako nástroj na redukciu dimenzie dát).

Podľa [1] možno takýto model (s drobnými úpravami, nakoľko v článku sa uvažovali dáta s pozorovaniami reprezentovanými stĺpcami, nie riadkami) ekvivalentne popísať aj maticovo.

$$X \approx ABX = AZ \tag{8}$$

$$b_{ij} \geq 0 \tag{9}$$

$$\sum_{j=1}^{n} b_{ij} = 1 \tag{10}$$

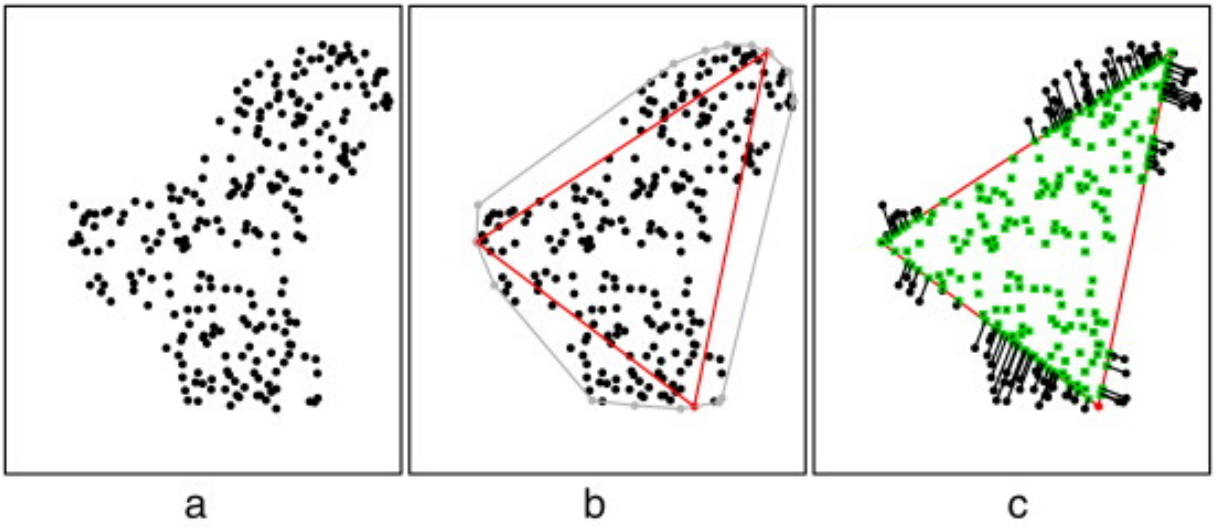
$$a_{ij} \geq 0 \tag{11}$$

$$\sum_{j=1}^{p} a_{ij} = 1 \tag{12}$$

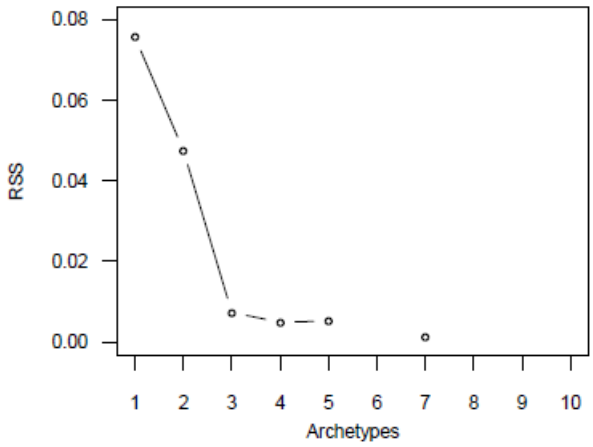
Základná myšlienka AA je dobre ilustrovaná na Obr.2, pochádzajúcom z [4]. Využíva sa tu jednoduchý 2-rozmerný modelový dataset toy, dostupný v balíku *archetypes* [4]. Na strednom obrázku možno vidieť konvexný obal pozorovaní a takisto máme červenou farbou zobrazené 3 nájdené archetypy a ich konvexný obal. Posledný obrázok potom zobrazuje aproximácie jednotlivých pozorovaní. Všimnime si, že pozorovania, nachádzajúce sa vo vnútri konvexného obalu trojice archetypov, sú aproximované presne, kým ostatné sú projektované na konvexný obal.

Ak by sme takto definovanú metódu chceli porovnať s inými metódami dekompozície a analýzy dát, v skratke by sme mohli povedať, že sa mnohými vlastnosťami nachádza na pomedzí medzi analýzou zhlukov a PCA. Hoci sa totiž interpretovateľnosťou blíži k metódam analýzy zhlukov, vďaka vyššej flexibilite modelu zachováva viac informácie (hoci stále menej ako PCA).

Jednou z najzásadnejších otázok, ktoré je pri aplikácii AA potrebné zodpovedať, je voľba p - počtu hľadaných archetypov. V niektorých prípadoch nám môže pomôcť znalosť skúmanej problematiky alebo vhodná vizualizácia dát ako napríklad pri už spomínaných dátach toy z Obr. 2, kde je z tvaru dát už na prvý pohľad jasné, že vhodný počet archetypov bude 3. Nakoľko však často pracujeme s viacrozmernými dátami, kde takáto vizualizácia nie je možná, je nutné postupovať inak.



Obr. 2: Vľavo: dáta toy; V strede: Konvexný obal dát, archetypy a ich konvexný obal; Vpravo: Projekcie dát na konvexný obal množiny archetypov. Zdroj: [4]



Obr. 3: Lakťový diagram na určenie vhodného počtu archetypov - dáta toy

Podobne ako pri analýze zhlukov (K-means, K-medoids) alebo PCA sa tu využíva tzv. lakťový diagram (Obr. 3). V tomto diagrame je na x-ovej osi znázornený počet archetypov a na y-ovej dosiahnutá hodnota RSS (2). Vhodný počet archetypov je taký, pre ktorý sa krivka láme a prestáva strmo klesať. Takéto správanie nám totiž hovorí, že ďalším zvyšovaním komplexity modelu (zvyšovanie počtu parametrov) sa nedosiahne výrazné skvalitnenie aproximácie. Všimnime si, že na Obr. 3, zobrazujúcom lakťový diagram pre dáta toy, takýto zlom nastáva pre $p = 3$, čo sa zhoduje s intuíciou, ktorú možno nadobudnúť už pri prvom pohľade na tvar týchto dát.

1.2 Poloha archetypov

V tejto časti uvedieme vetu, pojednávajúcu o polohe archetypov vo vzťahu k datasetu.

Veta 1.1. *Nech C je konvexný obal množiny pozorovaní $x_1, ..., x_n$. Nech S je množina pozorovaní nachádzajúcich sa na hranici C. Nech N je počet prvkov S. Potom*

1. *ak $p = 1$, položením $z_1 = \bar{x}$ sa minimalizuje RSS.*
2. *ak $1 < p < N$, existuje množina archetypov $\{z_1, ..., z_p\}$, nachádzajúcich sa na hranici C, ktorá minimalizuje RSS.*
3. *ak $p = N$, položením $\{z_1, ..., z_p\} = S$ sa dosiahne $RSS = 0$.*

Vetu aj jej dôkaz možno nájsť v [2].

1.3 Archetypálny algoritmus a striedavá optimalizácia

V tejto časti popíšeme proces hľadania archetypov, vedúci k striedavej optimalizácii. Našim cieľom je pre vopred zvolené p nájsť vektory $z_1, ..., z_p$, ktoré sú riešením optimalizačnej úlohy (2) - (7), v ktorej minimalizujeme

$$RSS = \sum_{i=1}^{n} \left|\left| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right|\right|^2 .$$

Toto vieme pomocou ohraničenia (3) prepísať do tvaru

$$RSS = \sum_{i=1}^{n} \left|\left| x_i - \sum_{k=1}^{p} \alpha_{ik} \sum_{j=1}^{n} \beta_{kj} x_j \right|\right|^2 . \tag{13}$$

Cieľom je RSS minimalizovať cez koeficienty α a β, spĺňajúce vyššie spomenuté ohraničenia. Takúto úlohu možno riešiť bežnými algoritmami určenými na riešenie problémov ohraničených nelineárnych najmenších štvorcov, avšak takýto postup by bol priveľmi neefektívny na to, aby bolo praktické ho aplikovať na problémy s väčším počtom premenných alebo pozorovaní. Z tohto dôvodu autori v [2] navrhli využiť tzv. striedavú optimalizáciu.

Hlavná myšlienka striedavej optimalizácie spočíva v tom, že namiesto hľadania optimálnych hodnôt koeficientov α a β v jednom kroku sa strieda medzi hľadaním najlepších hodnôt α pre dané archetypy (popísané pomocou daných koeficientov β) a hľadaním najlepších archetypov (hodnôt koeficientov β) pre dané α. Ukážeme (podľa [2]), že v každom

kroku sa potom pre dané u a $t_1, ..., t_q$ riešia problémy konvexných najmenších štvorcov (CLS) tvaru

$$\min_{w_1,...,w_q} \quad \left|\left| u - \sum_{k=1}^{q} w_k t_k \right|\right|^2 \tag{14}$$

$$w_k \geq 0 \tag{15}$$

$$\sum_{k=1}^{q} w_k = 1. \tag{16}$$

Pozrime sa najprv na hľadanie najlepších hodnôt koeficientov α pre dané vektory $z_1, ..., z_p$. V tomto kroku je potrebné nájsť vhodnú aproximáciu (konvexnú kombináciu vektorov $z_1, ..., z_p$) každého pozorovania, preto sa rieši nasledovných n problémov CLS:

$$\min_{\alpha_{i1},...,\alpha_{ip}} \quad \left|\left| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right|\right|^2 \tag{17}$$

$$\alpha_{ik} \geq 0 \tag{18}$$

$$\sum_{k=1}^{p} \alpha_{ik} = 1. \tag{19}$$

Každý z týchto problémov má m pozorovaní (jednotlivé zložky x_i) a p premenných (koeficienty pre jednotlivé vektory $z_1, ..., z_p$).

Pri hľadaní najlepších vektorov $z_1, ..., z_p$ (teda vhodných koeficientov β) pre dané α je situácia o niečo zložitejšia, avšak dá sa ukázať, že stále vedie na problémy CLS. Považujme všetky vektory $z_1, ..., z_p$ okrem jedného (z_l) za konštantné (preto bude potrebné riešiť p optimalizačných úloh) a označme

$$\begin{aligned} v_i &= \frac{x_i - \sum_{k \neq l} \alpha_{ik} z_k}{\alpha_{il}} \\ \bar{v} &= \frac{\sum_{i=1}^{n} \alpha_{il}^2 v_i}{\sum_{i=1}^{n} \alpha_{il}^2}. \end{aligned}$$

Potom môžeme napísať

$$\begin{aligned} RSS &= \sum_{i=1}^{n} \left|\left| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right|\right|^2 = \sum_{i=1}^{n} \left|\left| x_i - \sum_{k \neq l} \alpha_{ik} z_k - \alpha_{il} z_l \right|\right|^2 \\ &= \sum_{i=1}^{n} \alpha_{il}^2 \left|\left| \frac{x_i - \sum_{k \neq l} \alpha_{ik} z_k}{\alpha_{il}} - z_l \right|\right|^2 = \sum_{i=1}^{n} \alpha_{il}^2 \left|\left| v_i - z_l \right|\right|^2 \\ &= \sum_{i=1}^{n} \alpha_{il}^2 \left|\left| v_i - \bar{v} \right|\right|^2 + \sum_{i=1}^{n} \alpha_{il}^2 \left|\left| \bar{v} - z_l \right|\right|^2. \end{aligned}$$

Prvý člen nezávisí od z_l a s α_{il} pracujeme v tomto kroku ako s konštantou, vďaka čomu

dostávame, využijúc (3), novú, ekvivalentnú optimalizačnú úlohu

$$\min_{\alpha_{i1},\ldots,\alpha_{ip}} \quad \left\| \bar{v} - \sum_{j=1}^{n} \beta_{lj} x_j \right\|^2 \tag{20}$$

$$\beta_{lj} \geq 0 \tag{21}$$

$$\sum_{j=1}^{n} \beta_l j = 1, \tag{22}$$

ktorá už má tvar problému CLS. V tejto úlohe vystupuje m pozorovaní (zložky $\bar{v}$) a n premenných (koeficienty pre jednotlivé pozorovania).

Namiesto súčasného hľadania koeficientov α a β teda postupujeme iteračne, pričom v každej iterácii riešime najprv n úloh, hľadajúcich pre dané β vhodné α, a potom p úloh, zodpovedajúcich opačnej situácii. Každá z týchto úloh je však v tvare problému CLS, na riešenie ktorých existujú mnohé známe algoritmy, napríklad penalizovaná verzia NNLS algoritmu Lawsona a Hansona, popísaná v [2].

V každom kroku sa nájdením vhodnejších vektorov $z_1, ..., z_p$, resp. vhodnejších aproximácii jednotlivých pozorovaní znižuje hodnota RSS a archetypálny algoritmus pokračuje, až kým sa nedosiahne dostatočné zlepšenie.

2 Aplikácia na reálne dáta

V tejto kapitole popísané metódy aplikujeme na reálne dáta, dostupné na stránke Štatistického úradu SR [13], popisujúce výsledky volieb do NR SR v roku 2016 vzhľadom na jednotlivé kraje, resp. okresy. Používať budeme jazyk *R* [11], a balíky *archetypes* [4], *vcd* [10] a *Ternary* [12] (vizualizácia), *stringr* [17] (práca so stringmi) a *HSAUR* [6] (vizualizácia pomocou mnohorozmerného škálovania).

2.1 Dáta - kraje

V tejto časti budeme pracovať s dátami, zobrazenými v Tab. 1. Tieto popisujú výsledky volieb do NR SR v roku 2016 z hľadiska rozloženia podpory jednotlivých strán v rámci krajov (a zahraničia). Ako pozorovania tu vystupujú jednotlivé strany a ako premenné pomery, vyjadrujúce, koľkonásobne vyššie/nižšie preferencie dosiahla tá-ktorá strana v jednotlivých krajoch oproti svojim celkovým preferenciám. Napríklad číslo v ľavom hornom rohu teda vyjadruje, že preferencie strany SIEŤ boli v Bratislavskom kraji približne 1,16-násobne vyššie ako jej výsledné preferencie v rámci celej SR. V týchto dátach máme 23 pozorovaní o 9 premenných.

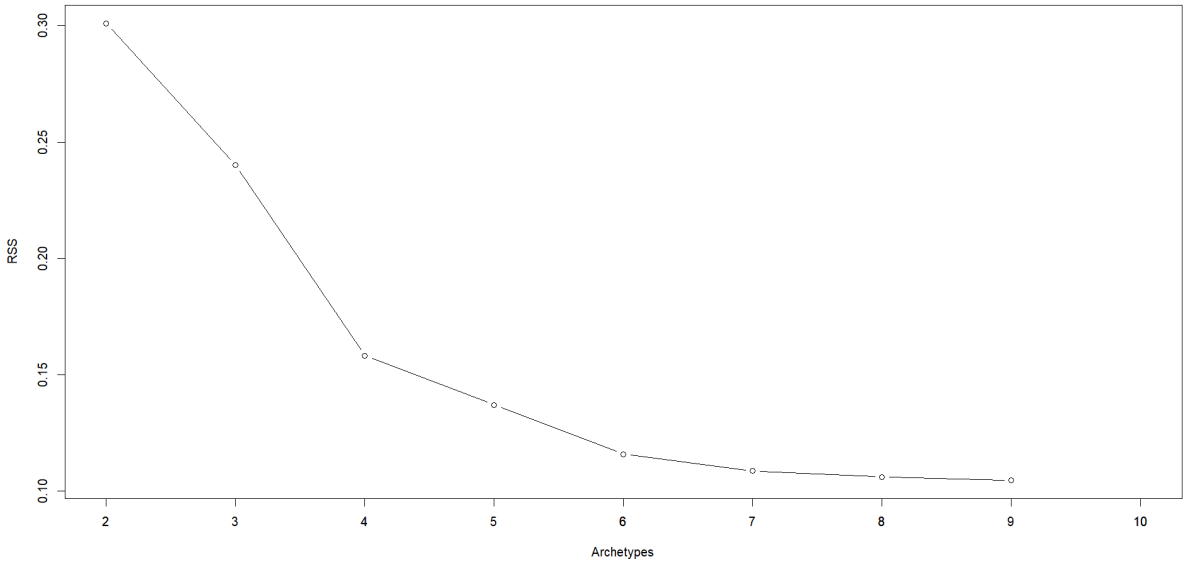
Na začiatku nájdeme najlepšie modely pre rôzne p (počet archetypov) a na Obr. 4 zobrazíme lakťový diagram. Najvýraznejšie zlomy nastávajú pre $p = 4$ a $p = 6$, čo sú zrejme najvhodnejšie počty archetypov.

My na začiatku položíme $p = 3$, hoci sa takýto počet archetypov pre korektnú analýzu ukazuje ako nevhodný. Umožňuje však použiť užitočný nástroj na vizualizáciu výsledkov, využívajúci schopnosť AA poskytnúť reprezentáciu dát s nižšou dimenziou.

Archetypy na takýchto dátach budú virtuálnymi objektami (stranami), od ktorých očakávame, že budú realistické, ale extrémne, podobne ako tváre na Obr. 1. Nájdené archetypy možno vidieť v Tab. 2. Pozrime sa teraz na každý z nich a pokúsme sa o jeho interpretáciu. Pri prvom archetype si všimneme najmä veľmi nízke úrovne podpory pre Trenčiansky, Žilinský a Prešovský kraj a naopak vysoké pre Trnavský, Nitriansky a v menšej miere aj Banskobystrický a Košický. V skratke, jedná sa o stranu s nízkou podporou na severe Slovenska a vysokou na juhu. Toto nasvedčuje tomu, že sa jedná o archetyp maďarskej strany, o čom sa presvedčíme neskôr, keď nahliadneme na hodnoty koeficientov

	BA	TT	TN	NR	ZA	BB	PO	KE	Zahraničie
#SIEŤ	1.16	0.82	1.01	0.73	1.14	0.86	1.03	1.17	2.84
Demokrati Slovenska - Ľudo Kaník	0.71	1.00	1.00	0.86	0.86	1.14	1.29	1.29	1.29
Kresťanskodemokratické hnutie	0.80	0.79	0.86	0.67	1.22	0.61	1.83	1.07	1.03
Spoločne za Slovensko	1.17	1.00	1.00	0.67	1.00	1.00	1.17	1.17	0.00
Komunistická strana Slovenska	0.58	0.82	1.31	0.74	0.92	1.35	1.27	1.06	0.15
MKA - MKS	0.33	1.67	0.00	2.56	0.11	1.56	0.11	1.78	0.00
MOST - HÍD	1.27	1.79	0.30	1.91	0.27	1.03	0.40	1.19	2.48
OĽANO - NOVA	1.29	1.06	0.89	0.75	1.01	0.86	1.10	1.00	1.16
Odvaha	0.69	0.46	0.77	0.62	0.85	1.15	1.69	1.54	2.92
PRIAMA DEMOKRACIA	1.31	0.62	0.85	0.77	0.92	1.15	1.23	1.23	0.69
SDKÚ - DS	1.65	1.38	0.81	0.50	0.77	0.92	1.04	0.92	1.81
Slovenská občianska koalícia	1.58	0.76	0.95	0.72	0.87	1.04	0.89	1.10	1.61
SME RODINA - Boris Kollár	1.13	1.02	1.16	0.93	0.94	0.90	0.90	1.03	0.24
SMER - sociálna demokracia	0.69	0.81	1.26	0.95	1.11	1.03	1.17	0.99	0.21
SMK - MKP	0.25	2.85	0.00	3.04	0.00	1.20	0.01	1.18	0.43
Slovenská národná strana	0.70	0.70	1.28	0.89	1.47	1.11	1.03	0.79	0.22
SMS	0.41	0.71	0.82	0.65	0.76	1.53	1.82	1.24	0.53
Strana zelených Slovenska	1.07	0.93	1.21	0.87	0.99	1.10	0.91	0.91	1.00
Sloboda a Solidarita	2.02	0.87	0.94	0.73	0.90	0.87	0.68	0.84	2.47
Strana TIP	0.39	0.61	0.71	0.82	0.74	1.58	1.54	1.61	0.00
VZDOR - strana práce	0.58	0.75	1.00	0.75	1.17	1.25	1.42	1.00	0.75
Kotleba - ĽSNS	0.59	0.85	1.09	0.85	1.21	1.30	1.17	0.97	0.65
ŠANCA	0.80	0.68	0.64	0.64	0.68	1.84	1.24	1.36	1.88

Tabuľka 1: Relatívne podpory kandidujúcich strán v porovnaní s ich podporami v rámci celej SR



Obr. 4: Lakťový diagram - kraje - 2016

β. Pri druhom archetype stoja za zmienku najmä vysoké podpory v Bratislavskom kraji a v zahraničí a nízke v Trnavskom, Nitrianskom, Banskobystrickom a Košickom. Môžeme sa teda domnievať, že sa jedná o archetyp strany mestského voliča. Tretí archetyp sa v mnohom správa ako opak druhého. V Bratislavskom kraji a v zahraničí dosahuje nízku podporu, zatiaľ čo najvyššie preferencie zaznamenáva v Banskobystrickom kraji a na východe slovenska. Jedná sa zrejme o archetyp vidieckej či regionálnej strany. Niektoré z vlastností archetypov sa ukážu jasnejšie v ďalšej časti práce, kde namiesto krajov vezmeme jednotlivé okresy a budeme tak schopný lepšie zahrnúť napríklad vplyv miest ako Košice.

	BA	TT	TN	NR	ZA	BB	PO	KE	Zahraničie
1	0.28	2.46	0.00	2.88	0.04	1.32	0.04	1.38	0.29
2	1.48	0.85	1.00	0.76	1.05	0.86	0.98	0.90	1.59
3	0.40	0.65	0.76	0.75	0.75	1.56	1.66	1.45	0.22

Tabuľka 2: Nájdené archetypy - 2016

Pozrime sa ďalej na hodnoty koeficientov α a β, ktoré sú zároveň zložkami matíc A, resp. B z maticového zápisu modelu (8). Koeficienty α vyjadrujú, akými konvexnými kombináciami archetypov aproximujeme jednotlivé pozorovania. Ich hodnoty pre naše

dáta môžeme vidieť v Tab. 3.

Keď sa pozrieme na 1. archetyp (1. stĺpec), podľa očakávania vidíme, že tento mal najvyššie zastúpenie pri aproximácii maďarských strán. Zaujať môže aj to, že z trojice kandidujúcich maďarských strán bol tento archetyp najmenej zastúpený pri strane MOST - HÍD. Toto však korešponduje s realitou, nakoľko na rozdiel od zvyšných dvoch maďarských strán bývala táto vnímaná o čosi menej ako striktne národnostná strana a vyznačovala sa o niečo rozmanitejšou voličskou základňou, nesúcou podobnosti napríklad s voličskou základňou SaS.

2. archetyp je najviac zastúpený pri aproximácii strán SaS, SDKÚ - DS a SIEŤ. Toto sa zhoduje s našou interpretáciou tohto archetypu ako archetypu strany mestského voliča. Všimnime si tiež, že je tento archetyp významnejšie zastúpený pri aproximácii výrazne vyššieho počtu strán ako archetyp 1, ktorý popisuje oveľa špecifickejší typ strany. Zároveň dostávame 1. možnosť nahliadnuť na jednu z hlavných nevýhod AA, ktorou je vysoká citlivosť na outliery, spôsobená snahou čo najlepšie zachytiť extrémy v dátach. Keď sa pri 1. archetype pozrieme na hodnoty pre Trenčiansky, Žilinský a Prešovský kraj, vidíme, že sú skutočne veľmi nízke, čo spôsobuje, že je tento archetyp veľmi vzdialený od väčšiny pozorovaní, a preto je slabo zastúpený pri ich aproximácii.

Keď sa na záver pozrieme na 3. archetyp, vidíme, že najväčšiu úlohu zohráva pri aproximácii Strany TIP a SMS. Interpretácia je tu sťažená tým, že sa jedná o menšie a menej známe strany (citlivosť na outliery). Keď však nahliadneme na rozloženie ich preferencií (Tab. 1), u oboch si môžeme všimnúť nízku podporu v Bratislavskom kraji a v zahraničí a, naopak, vysokú v Banskobystrickom, Prešovskom a Košickom kraji. Toto sa zhoduje s našou prvotnou interpretáciou, že sa jedná o archetyp regionálnej strany.

V súvislosti s koeficientami α považujeme na tomto mieste za vhodné spomenúť možnosť využitia AA na redukciu dimenzie dát. Nakoľko máme každé pozorovanie aproximované konvexnou kombináciou archetypov ($x_i \approx \sum_{k=1}^{p} \alpha_{ik} z_k$), vieme ho reprezentovať pomocou príslušných koeficientov $\alpha_{i1}, ..., \alpha_{ip}$. V našom prípade ($p = 3$) teda namiesto 9-rozmerného pozorovania dostaneme jeho 3-rozmernú reprezentáciu. AA je v tomto podobná napríklad PCA, hoci pridaním ohraničení na koeficienty α a β sa zvýšila interpretovateľnosť na úkor zachovanej informácie.

Pozrime sa ďalej na hodnoty koeficientov β. Tieto vyjadrujú, akými konvexnými kom-

	1	2	3
#SIEŤ	0.00	0.91	0.09
Demokrati Slovenska - Ľudo Kaník	0.05	0.46	0.48
Kresťanskodemokratické hnutie	0.00	0.72	0.28
Spoločne za Slovensko	0.02	0.62	0.36
Komunistická strana Slovenska	0.00	0.41	0.59
MKA - MKS	0.79	0.00	0.21
MOST - HÍD	0.52	0.48	0.00
OĽANO - NOVA	0.05	0.85	0.10
Odvaha	0.00	0.38	0.62
PRIAMA DEMOKRACIA	0.00	0.58	0.42
SDKÚ - DS	0.08	0.92	0.00
Slovenská občianska koalícia	0.03	0.85	0.12
SME RODINA - Boris Kollár	0.08	0.76	0.16
SMER - sociálna demokracia	0.00	0.62	0.38
SMK - MKP	1.00	0.00	0.00
Slovenská národná strana	0.00	0.74	0.26
SMS	0.00	0.11	0.89
Strana zelených Slovenska	0.03	0.78	0.18
Sloboda a Solidarita	0.00	1.00	0.00
Strana TIP	0.02	0.00	0.98
VZDOR - strana práce	0.00	0.49	0.51
Kotleba - ĽSNS	0.00	0.53	0.47
ŠANCA	0.05	0.16	0.79

Tabuľka 3: Hodnoty koeficientov α

binániami pozorovaní sú tvorené jednotlivé archetypy. Ich hodnoty pre naše dáta možno (v transponovanej podobe) vidieť v Tab. 4. Ako prvé si všimnime to, koľko z nich je na rozdiel od koeficientov α rovných 0. Toto je v súlade s teóriou, popísanou v kap. 1.2, kde Veta 1.1 hovorila o polohe archetypov pre rôzne p. Naša situácia zodpovedá druhej možnosti z vety, archetypy z_1, z_2, z_3 sa teda musia nachádzať na hranici konvexného obalu, čo znamená, že musia byť konvexnými kombináciami pozorovaní z tejto hranice.

Pre archetyp 1 máme nenulové hodnoty koeficientov β iba pri 2 pozorovaniach, pričom v obidvoch prípadoch sa jedná o strany, ktoré boli aproximované prevažne pomocou tohto archetypu. Zaujať môže nulový príspevok strany MOST - HÍD, hoci aj pri aproximácii tej bol s najväčšou váhou zastúpený práve archetyp 1.

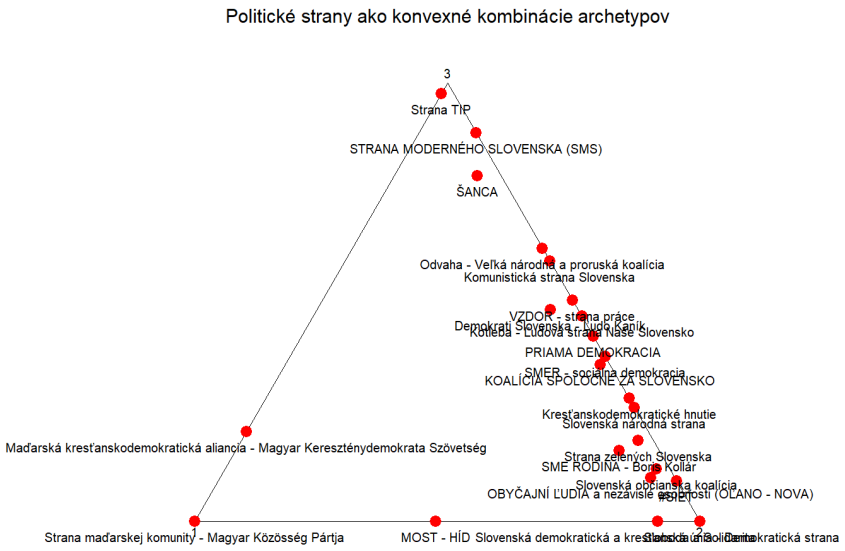
Archetyp 2 je tvorený konvexnou kombináciou 4 pozorovaní, pričom najväčšia váha pripadá strane SaS, ktorej aproximácia bola tvorená takmer výlučne iba týmto archetypom. Zvláštna sa môže zdať nulová hodnota pre stranu SIEŤ, pri aproximácii ktorej mal archetyp 2 váhu viac ako 0,9. Jedná sa o podobnú situáciu ako pri archetype 1 a strane MOST - HÍD. Toto súvisí s rôznym významom koeficientov α a β a najmä s polohou archetypov na hranici konvexného obalu množiny pozorovaní.

Archetyp 3 je konvexnou kombináciou 2 pozorovaní, konkrétne strán TIP a SMS. Práve do aproximácie tých vstupoval s najvyššími váhami. Aj tu, podobne ako pri archetypoch 1 a 2, existuje viacero strán, pre ktoré sú príslušné hodnoty koeficientov β nulové, hoci archetyp 3 zohral nemalú úlohu pri ich aproximácii.

Spomenuli sme, že pre $p = 3$ poskytuje AA užitočný nástroj na vizualizáciu dát. Jedná sa o tzv. ternary plot [2], ktorý možno pre naše dáta vidieť na Obr. 5. Tento funguje na princípe nasledovného mapovania. i-te pozorovanie je aproximované ako $x_i \approx \alpha_{i1}z_1 + \alpha_{i2}z_2 + \alpha_{i3}z_3$. Nech μ_1, μ_2 a μ_3 sú vrcholy rovnostranného trojuholníka. x_i potom zobrazíme do roviny ako $\alpha_{i1}\mu_1 + \alpha_{i2}\mu_2 + \alpha_{i3}\mu_3$. Ľavý dolný roh zodpovedá archetypu 1, teda archetypu maďarskej strany, pravý dolný roh archetypu 2, archetypu mestskej strany, a horný roh zodpovedá archetypu 3, archetypu regionálnej strany. Všimnime si, že strany SMK, SaS a TIP, ktoré boli aproximované takmer výlučne 1 archetypom (v poradí archetypmi 1, 2 a 3) sa prirodzene nachádzajú v rohoch obrázka, zatiaľ čo napríklad strana MOST - HÍD, reprezentovaná ako kombinácia archetypov 1 a 2 (s váhami blízkymi 0,5), sa nachádza takmer presne medzi rohmi, zodpovedajúcimi týmto 2 archetypom.

	1	2	3
#SIEŤ	0.00	0.00	0.00
Demokrati Slovenska - Ľudo Kaník	0.00	0.00	0.00
Kresťanskodemokratické hnutie	0.00	0.20	0.00
Spoločne za Slovensko	0.00	0.00	0.00
Komunistická strana Slovenska	0.00	0.00	0.00
MKA - MKS	0.33	0.00	0.00
MOST - HÍD	0.00	0.00	0.00
OĽANO - NOVA	0.00	0.00	0.00
Odvaha	0.00	0.00	0.00
PRIAMA DEMOKRACIA	0.00	0.00	0.00
SDKÚ - DS	0.00	0.00	0.00
Slovenská občianska koalícia	0.00	0.00	0.00
SME RODINA - Boris Kollár	0.00	0.12	0.00
SMER - sociálna demokracia	0.00	0.00	0.00
SMK - MKP	0.67	0.00	0.00
Slovenská národná strana	0.00	0.15	0.00
SMS	0.00	0.00	0.42
Strana zelených Slovenska	0.00	0.00	0.00
Sloboda a Solidarita	0.00	0.54	0.00
Strana TIP	0.00	0.00	0.58
VZDOR - strana práce	0.00	0.00	0.00
Kotleba - ĽSNS	0.00	0.00	0.00
ŠANCA	0.00	0.00	0.00

Tabuľka 4: Hodnoty koeficientov β



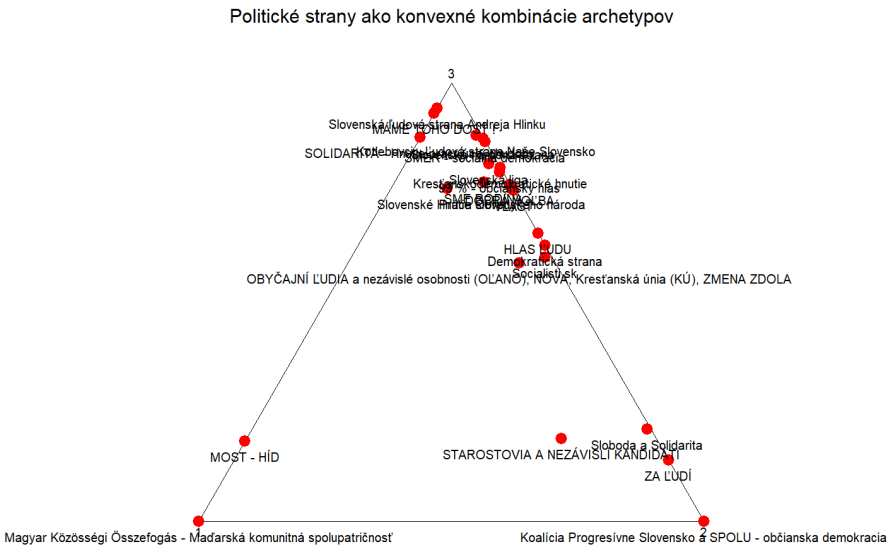
Obr. 5: Ternary plot - vizualizácia dát pre 3 archetypy

2.1.1 Voľby v rokoch 2020 a 2023

Analýzu na podobných dátach sme uskutočnili aj pre voľby v rokoch 2020 a 2023 (dáta: [14], [15]), kvôli stručnosti ju tu však nebudeme celú opakovať. Spomenieme dva najzásadnejšie závery, ktoré z týchto analýz vzišli.

Pre všetky tri roky sme dostali veľmi podobné trojice archetypov. Najkonzistentnejší bol archetyp maďarskej strany s nadpriemernými preferenciami v Trnavskom, Nitrianskom, Banskobystrickom, a Košickom kraji a naopak podpriemernými v Bratislavskom, Trenčianskom, Žilinskom a Prešovskom kraji. Takisto sa pre každý rok objavil archetyp mestskej strany s vysokou podporou v Bratislavskom kraji a v zahraničí a nízkou v Nitrianskom, Banskobystrickom, Prešovskom a Košickom kraji. Tretím, opäť pomerne konzistentným archetypom bol pre všetky roky archetyp regionálnej strany s vysokými preferenciami v Banskobystrickom kraji a na východe Slovenska a nízkymi v Bratislavskom, Trnavskom a Nitrianskom kraji a takisto v zahraničí.

Čo bolo medzi jednotlivými rokmi výrazne rozdielne, bolo priestorové rozloženie aproximácií jednotlivých pozorovaní, ktoré sme si všimli vďaka zobrazeniu jednotlivých ternary plotov (Obr. 5, 6 a 7). Všimnime si najmä kvalitatívny rozdiel medzi ternary plotmi pre roky 2016 a 2020, špecificky rozloženie aproximácií medzi archetypmi 2 a 3. Kým v roku 2016 bolo toto rozloženie pomerne rovnomerné, v roku 2020 došlo ku vzniku dvoch výrazne

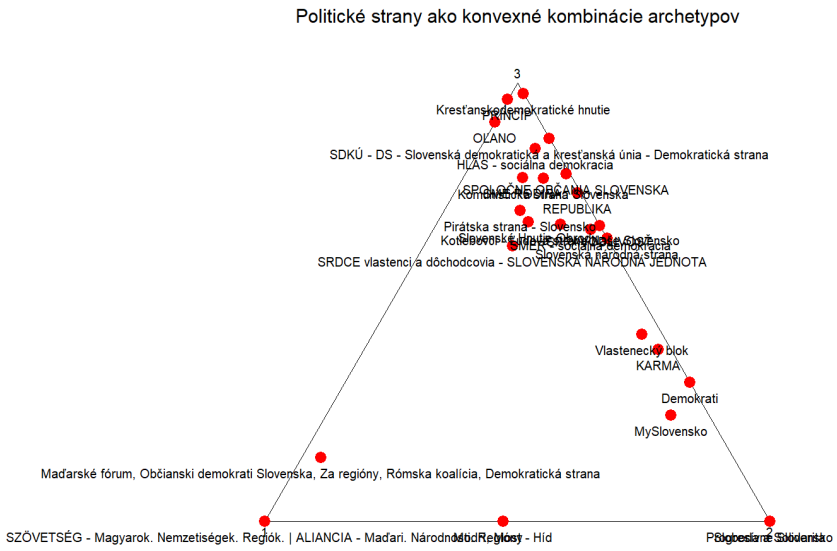


Obr. 6: Ternary plot pre rok 2020

oddelených zhlukov okolo rohov, zodpovedajúcich archetypom 2 a 3. Tento jav do určitej miery pretrval aj pri dátach z roku 2023, hoci tam došlo ku vzniku ďalšieho zhluku 4 strán, nachádzajúceho sa medzi dvoma, ktoré sme už spomínali. Nakoľko k tejto kvalitatívnej zmene správania dát došlo medzi rokmi 2016 a 2020, uvažovali sme, či nemôžu súvisieť so spoločenskými udalosťami, ktoré sa v tomto období odohrali, špecificky udalosťami z roku 2018 (vražda novinára Jána Kuciaka, mohutné občianske protesty a rekonštrukcia vlády) a polarizáciou spoločnosti, ktorá s nimi bola úzko spojená a ktorá sa práve v tom čase začala výrazne prehlbovať. Zodpovedanie tejto otázky ale nie je predmetom práce a my sme sa preto namiesto nej pozreli na problém hľadania spôsobu, ako polarizáciu v dátach exaktne merať. Výsledky, ku ktorým sme sa dopracovali, možno nájsť v kap. 4.

2.2 Dáta - okresy

V tejto podkapitole uvedieme podobnú analýzu ako v kap. 2.1, avšak ako premenné v našich dátach nebudú vystupovať relatívne podpory strán v rámci slovenských krajov (a zahraničia), ale ich relatívne podpory v rámci okresov (a zahraničia). Hodnoty premenných budú opäť vyjadrovať, koľkonásobne vyššie alebo nižšie preferencie dosiahla daná strana v tom-ktorom okrese. Dáta, s ktorými pracujeme, tu kvôli ich rozmerom nebudeme vypisovať, ale je možné ich nájsť na https://github.com/vincze16/name/tree/main/AA.



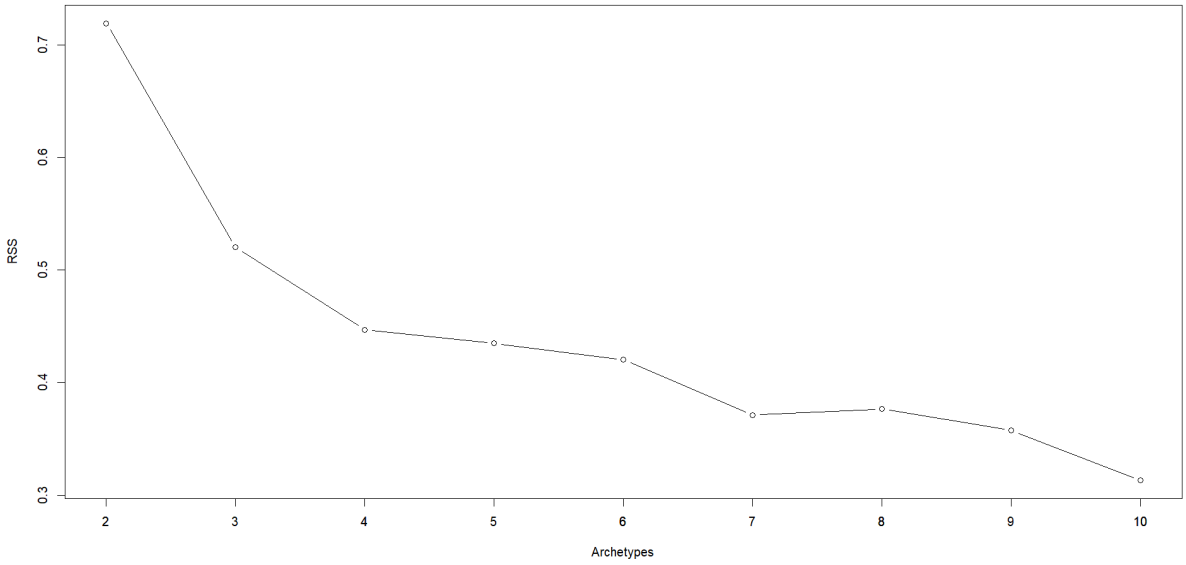
Obr. 7: Ternary plot pre rok 2023

Našu modelovú analýzu uskutočníme pre rok 2016. V tom sa volieb do NR SR zúčastnilo 23 strán, dataset teda obsahuje 23 pozorovaní, každé z nich 80-rozmerné (79 okresov a zahraničie).

V prvom kroku potrebujeme zvoliť vhodné p, na čo využijeme opäť lakťový diagram (Obr. 8). Ako vhodné počty archetypov sa javia $p = 4$ a $p = 7$, my budeme pracovať so 4 archetypmi.

Výstupy (https://github.com/vincze16/name/tree/main/AA) sú v mnohom podobné tým, ktoré sme dostali pri dátach o podpore strán naprieč krajmi ($p = 3$). Zo 4 nájdených archetypov sú 2 také, s ktorých podobnými sme sa už stretli. Jedná sa o archetyp maďarskej strany s vysokou podporou v južných okresoch a, naopak, nízkou na severe Slovenska. Tento archetyp zohrával dôležitú úlohu (vysoké hodnoty príslušných koeficientov α) najmä pri aproximáciách relatívnych podpôr strán MKDA - MKDSZ a SMK - MKP. O niečo nižšia, stále však vysoká bola hodnota príslušného koeficientu α pri strane MOST - HÍD. Pokiaľ ide o koeficienty β, tento archetyp je výlučne konvexnou kombináciou strán MKDA - MKDSZ a SMK - MKP.

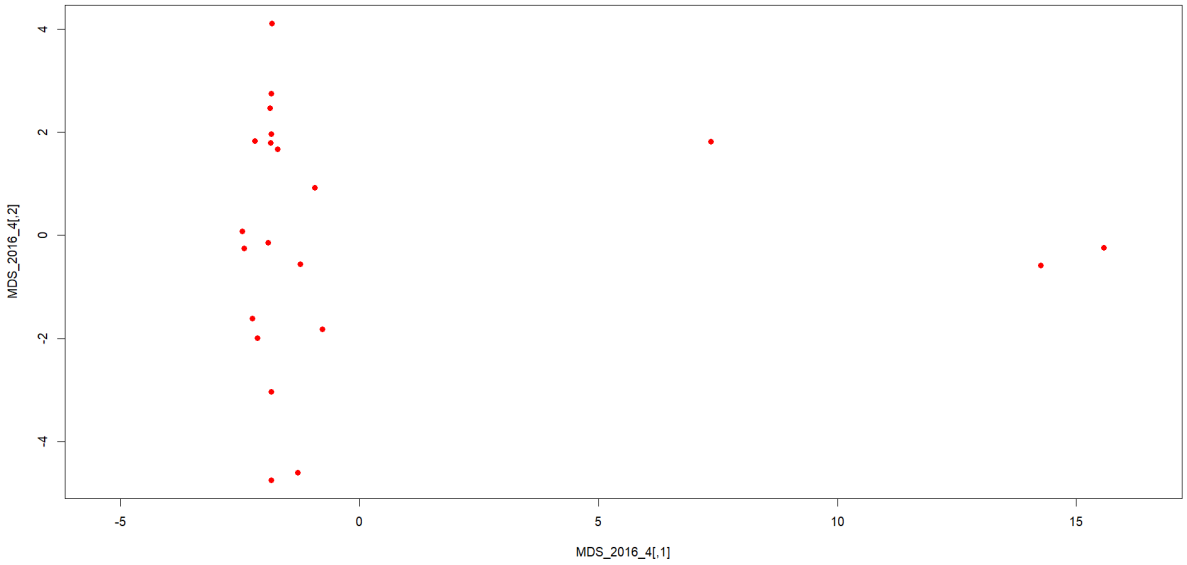
Druhým archetypom, s ktorému podobným sme sa stretli v kap. 2.1, je archetyp mestskej strany. Tento sa vyznačoval najmä vysokou podporou v bratislavských okresoch (a priľahlých okresoch ako Malacky, Pezinok či Senec) a zahraničí, ale aj v Košiciach, a



Obr. 8: Lakťový diagram - okresy - 2016

nižšou v menej rozvinutých okresoch južného a východného Slovenska. Pokiaľ ide o koeficienty α, tento archetyp vystupoval najmä pri aproximáciách relatívnych podpôr strán SaS, SKOK! či SDKÚ - DS, z parlamentných strán to potom boli v o niečo menšej miere strany OĽANO a SIEŤ. Na základe hodnôt koeficientov β možno vidieť, že tento archetyp je tvorený najmä stranou SaS, v menšej miere potom aj stranami SKOK! a SDKÚ - DS.

Pri posledných 2 archetypoch je potom situácia o niečo zložitejšia. Dá sa povedať, že tieto vznikli rozdelením archetypu regionálnej strany, s ktorým sme sa stretli v kap. 2.1. Oba s ním zdieľajú nízku podporu v Bratislave a v zahraničí, no zároveň medzi nimi možno pozorovať aj výrazné rozdiely. Kým prvý archetyp dosahoval vysokú podporu najmä v severných okresoch, v prípade druhého to boli skôr okresy na juhu a východe krajiny. K určitej zhode potom došlo na severovýchode Slovenska, kde sa vysokou podporou vyznačovali oba archetypy. Možno si tiež všimnúť, že v prípade prvého archetypu bola oproti druhému vysoká podpora dosahovaná v okresoch, ktoré sú vo všeobecnosti (s určitými výnimkami) považované za rozvinutejšie, zároveň však tradične konzervatívne. Kým pomocou prvého archetypu boli aproximované najmä strany ako KDH, SMER - SD, SNS či ĽS - NS, druhý zohral najväčšiu úlohu pri aproximácií strán SMS, TIP a Odvaha - Veľká národná a proruská koalícia. Veľmi podobná je aj situácia pri pohľade na koeficienty β, teda tieto 2 archetypy vznikli ako konvexné kombinácie strán, pri aproximácii ktorých



Obr. 9: Okresy (2016) - výstupy, zobrazené pomocou klasického mnorozmerného škálovania

potom samé vystupovali. Prvý archetyp by sme v tomto bode na základe toho, čo sme zistili, nazvali archetypom konzervatívnej strany. Pomenovať podobným spôsobom druhý archetyp nie je jednoduché. Zdá sa, že by ho bolo možné nazvať archetypom strany zastupujúcej málo rozvinuté regióny, tu však treba postupovať obzvlášť opatrne. Uvedomme si totiž, že všetky strany, ktoré sme v súvislosti s týmto archetypom spomínali (v spojitosti s koeficientami α aj β), boli strany s veľmi nízkymi preferenciami. Pri takýchto stranách je z dôvodu nízkeho počtu voličov oveľa pravdepodobnejšie, že premenné nadobudnú extrémne hodnoty (ktorých reálna výpovedná hodnota je však pomerne nízka). Práve tieto ale v najväčšej miere vplývajú na výstupy AA. Z týchto dôvodov je možné, že prítomnosť druhého archetypu, ako sme ho popísali, je dôsledkom vysokej citlivosti AA na outliery a pri jeho interpretácii je potrebné byť obzvlášť opatrný.

Nakoľko sme zvolili $p > 3$, na zobrazenie výstupov už nemôžeme využiť ternary plot. Namiesto toho na vizualizáciu výstupov využijeme klasické mnohorozmerné škálovanie (Obr. 9).

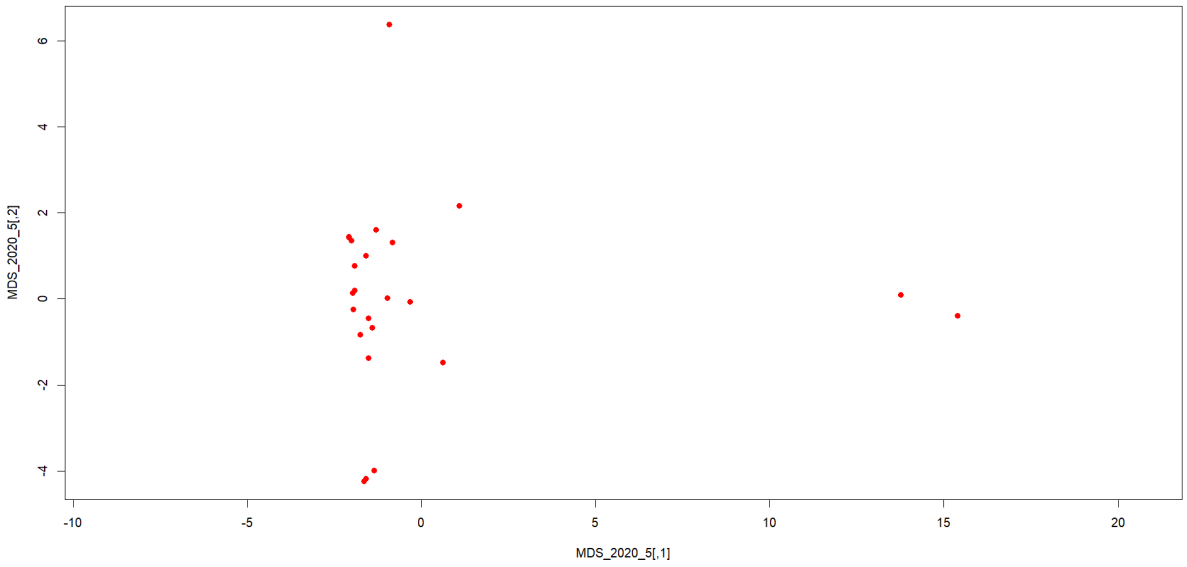
2.2.1 Voľby v rokoch 2020 a 2023

V roku 2020 sa ako najvhodnejšie zdalo zvoliť $p = 5$. Medzi archetypmi sa opäť objavili archetypy maďarskej strany a strany mestského voliča. Taktiež sa objavil archetyp, ktorý

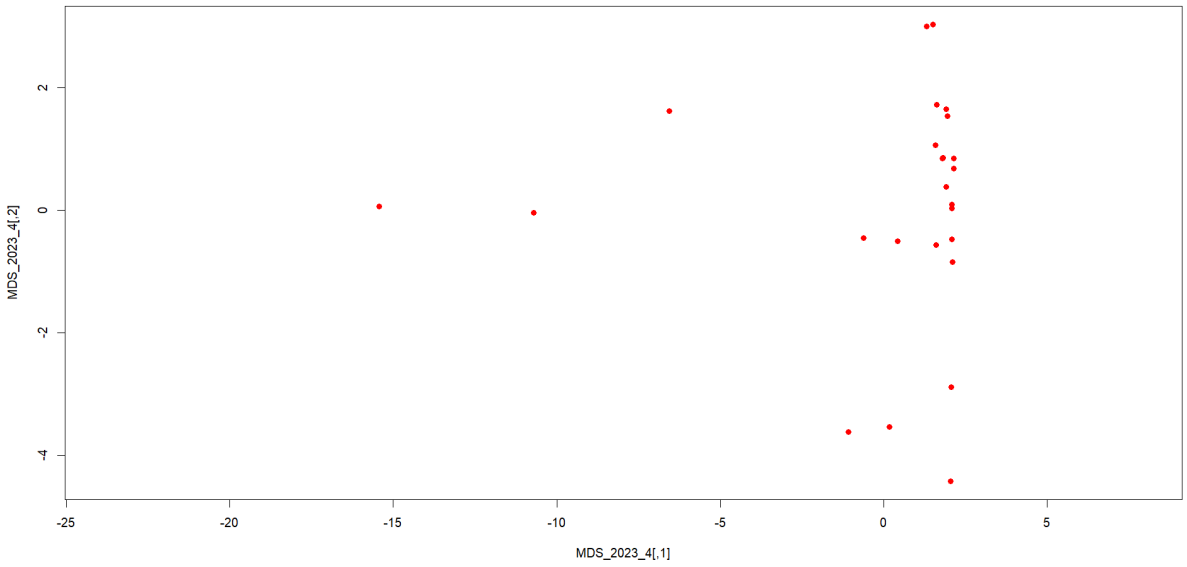
sme nazývali archetypom konzervatívnej strany, reprezentovaný podobnými stranami ako v roku 2016. Pokiaľ ide o rozloženie podpory v rámci tohto archetypu, aj tentokrát sa jednalo najmä o vysokú podporu na severe Slovenska, ale pribudli k nej aj mnohé menej rozvinuté okresy na juhu a východe krajiny. Zdá sa teda, že v rámci tohto archetypu splynuli archetypy 1 a 2 z roku 2016. Ostávajúce 2 archetypy sú potom o niečo ťažšie popísateľné, bez dobre viditeľných vzorcov. Vidíme pri nich napríklad výrazne odlišné (niekedy viac ako 100-násobne) miery podpory v rámci jednotlivých bratislavských okresov. Nakoľko ale oba tieto archetypy súvisia so stranami s nízkymi preferenciami (počty hlasov niektorých z nich sa v niektorých okresoch počítali rádovo nanajvýš v desiatkach), môžeme sa opäť domnievať, že ich prítomnosť súvisí s vysokou citlivosťou AA na outliery.

Pre rok 2023 boli najvhodnejšími počtami archetypov $p = 4$ a $p = 6$, my najmä pre možnosť názornejšieho porovnania s rokom 2016 zvolíme $p = 4$. Opäť sa objavil archetyp maďarskej strany a strany mestského voliča. Zvyšné 2 archetypy súvisia s archetypmi 1 a 2 v roku 2016. Kým vtedy sa však jeden archetyp vyznačoval najmä vysokou podporou v severných okresoch a druhý v južných a východných, v roku 2023 sa toto rozloženie zmenilo. Pre jeden z týchto archetypov bola typická podpora na strednom slovensku a jeho najvýznamnejším predstaviteľom bola strana ĽS - NS. Druhý bol potom silný na východnom Slovensku a na Orave a zastupovaný bol napríklad stranami OĽANO a KDH.

Výstupy pre roky 2020 a 2023 možno vidieť vizualizované pomocou klasického mnohorozmerného škálovania na Obr. 10 a 11.



Obr. 10: Okresy (2020) - výstupy, zobrazené pomocou klasického mnorozmerného škálovania



Obr. 11: Okresy (2023) - výstupy, zobrazené pomocou klasického mnorozmerného škálovania

3 Vážené archetypy

Archetypálna analýza, ako je popísaná v kap. 1, sa snaží dáta popísať pomocou archetypov - extrémnych vzorov, zvyčajne sa nachádzajúcich na hranici konvexného obalu datasetu (veta 1.1). Dôsledkom toho je jedna z jej hlavných nevýhod - vysoká citlivosť na outliery. Jej prejavy sme mohli vidieť aj pri našej vzorovej analýze v kap. 2.2, kde dochádzalo ku vzniku archetypov, súvisiacich prevažne s malými stranami, ktorých podpora v rámci slovenska bola vplyvom delenia malého počtu hlasov medzi veľký počet okresov náchylná na veľmi neštandardné správanie.

Spôsobov, ako sa vysporiadať s týmto problémom, je viacero. Patria medzi ne napríklad vážené archetypy alebo robustné archetypy, obe popísané v [5]. My na tomto mieste priblížime prvý menovaný spôsob.

Riešená optimalizačná úloha pri ňom nadobudne tvar

$$\min_{a,b} \quad ||W(X - AZ)||_2 = \tag{23}$$

$$Z = BX \tag{24}$$

$$b_{ij} \geq 0 \tag{25}$$

$$\sum_{j=1}^{n} b_{ij} = 1 \tag{26}$$

$$a_{ij} \geq 0 \tag{27}$$

$$\sum_{j=1}^{p} a_{ij} = 1. \tag{28}$$

Všimnime si podobnosť s reprezentáciou (8) - (12). Jediným rozdielom je prítomnosť W - matice váh, ktorá je tvaru $n \times n$. Táto môže mať rôzne významy, avšak najnázornejšia je jej interpretácia v prípade, že je diagonálna, kedy hodnoty v nej predstavujú váhy (dôležitosť) jednotlivých pozorovaní. Cieľom metódy je v takomto prípade čo najlepšie aproximovať dôležitejšie pozorovania (vyššie váhy), zatiaľ čo odchýlky pri tých menej dôležitých až takú úlohu nezohrávajú. Obrátiac sa späť k príkladu dát, s ktorými sme pracovali v kap. 2, ak by sme ako váhy pozorovaní (strán) vzali ich preferencie, dosiahli by sme práve zmenšenie vplyvu malých, na výkyvy náchylných strán, ktorých prítomnosť bola dôvodom zle interpretovateľných výsledkov.

Podľa [5] tiež vieme, že platí

$$W(X - AZ) = WX - ABWX \tag{29}$$

a, ak označíme $\tilde{X} = WX$ a $\tilde{Z} = B\tilde{X}$, môžeme napísať

$$RSS = \left|\left|\tilde{X} - A\tilde{Z}\right|\right|_2,$$

vďaka čomu je optimalizačná úloha ekvivalentná pôvodnej úlohe na upravených vážených dátach $\tilde{X}$, a teda je možné ju riešiť rovnakým algoritmom.

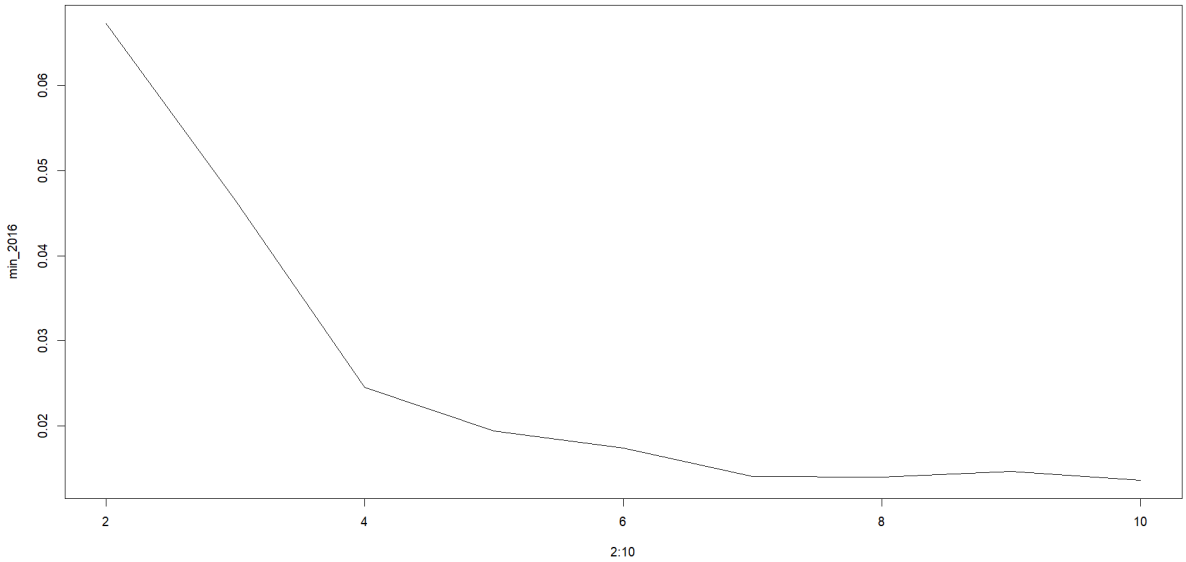
Keďže je teda hľadanie vážených archetypov ekvivalentné hľadaniu štandardných archetypov na transformovaných dátach, je nutné mať na pamäti niekoľko vecí.

Prvou z nich je to, že dáta je potrebné centrovať, nakoľko prenásobenie dátovej matice X maticou váh W môže mať za následok zmršťovanie (prípadne roztahovanie) datasetu, a teda aj jeho konvexného obalu, čo môže pri necentrovaných dátach viesť k nájdeniu archetypov, ktoré vôbec neležia v konvexnom obale pôvodných dát. Tým sa stráca ich zmysluplnosť v kontexte konkrétnej aplikácie, a teda aj interpretovateľnosť, ktorá je považovaná za hlavnú výhodu AA.

Z rovnakého dôvodu je nutné váhy škálovať tak, aby boli z intervalu $[0, 1]$. V opačnom prípade by totiž konvexný obal transformovaných dát (a teda potenciálne aj nájdené archetypy) opäť mohol obsahovať aj body nepatriace do konvexného obalu pôvodných dát (posunutých). Takýmto škálovaním sa dosiahne, že konvexný obal transformovaných dát ($\tilde{X}$) je podmnožinou konvexného obalu pôvodných dát (posunutých), čím bude zabezpečená dobrá interpretovateľnosť.

Vážené archetypy opäť aplikujeme na naše dáta o voľbách do NR SR (začínajúc rokom 2016), využívajúc funkciu *weightedArchetypes*, ktorá je súčasťou balíka archetypes [4]. Ako váhy zvolíme celkové preferencie jednotlivých strán, predelené preferenciami víťaznej strany, najvyššia váha teda dosiahne hodnotu 1. Dáta následne vycentrujeme, aby zmršťovaním konvexného obalu datasetu nedošlo ku vzniku nezmyselných objektov. Na takto upravené dáta už môžeme aplikovať funkciu *weightedArchetypes*. Pomocou lakťového diagramu zvolíme vhodné p (Obr. 12). Ako najvhodnejšie sa zdá položiť $p = 4$.

Archetypy, získané v rámci nami zvoleného modelu, následne spätne posunieme (vracajúc sa tak späť k dátam pred centrovaním). Ako presne nájdené archetypy vyzerajú, možno vidieť v Tab. 5. Dva z nich sa podobajú tým, s ktorými sme sa stretli už v kap. 2. Archetyp 1 je archetypom maďarskej strany a archetyp 2 je archetypom strany mestského voliča. Všimnime si však, že hodnoty jednotlivých premenných sú výrazne bližšie k 1, a teda menej extrémne, ako tomu bolo pri štandardných archetypoch. Toto svedčí



Obr. 12: Lakťový diagram (vážené archetypy) - kraje - 2016

o nižšej citlivosti na extrémne pozorovania, pokiaľ sú tieto pomocou váh označené ako menej dôležité. Pre interpretáciu archetypov 3 a 4 sa pozrieme na hodnoty koeficientov α a β.

Pri pohľade na koeficienty α (Tab. 6) vidíme, že kým archetyp 3 zohrával úlohu najmä pri aproximácii strán SMER - SD, KSS a ĽSNS, v prípade archetypu 4 to boli najmä strany Odvaha, TIP a ŠANCA. Pokiaľ ide o koeficienty β, kým do archetypu 3 vstupovala hlavne strana SMER - SD, archetyp 4 vznikol s pomerne rovnomerným zastúpením viacerých veľmi rôznorodých strán vrátane KDH, SNS, SMK a SaS. S najvyššími váhami však k nemu prispievajú strany KOALÍCIA SPOLOČNE ZA SLOVENSKO a KSS.

Kým archetyp 3 by sa teda dal nazvať archetypom konzervatívne a ľavicovo orientovanej strany, pre archetyp 4 takéto jednoduché pomenovanie nájsť nedokážeme a môžeme vidieť, že napriek využívaniu váh, tento archetyp zachytáva najmä informáciu z menších, menej významných strán.

Pre rok 2020 boli výsledky pomerne podobné. Opäť sa vyskytli archetypy maďarskej strany, strany mestského voliča aj strany konzervatívneho a ľavicovo orientovaného voliča. Zaujímavý bol posledný archetyp, vyznačujúci sa nadpriemernou podporou v Bratislavskom a Trnavskom kraji a na východe Slovenska, ktorý bol prostredníctvom koeficientov α aj β úzko prepojený so stranou OĽANO. Nakoľko práve táto strana v daných voľbách

	BA	TT	TN	NR	ZA	BB	PO	KE	Zahraničie
1	0.87	1.24	0.74	1.26	0.74	1.15	0.93	1.16	1.07
2	1.39	0.95	0.89	0.87	0.88	1.03	0.91	1.02	1.66
3	0.70	0.82	1.24	0.95	1.11	1.04	1.16	0.99	0.25
4	0.91	0.99	0.86	0.97	0.87	1.15	1.10	1.16	1.07

Tabuľka 5: Nájdené archetypy - 2016

zvíťazila (a teda do modelu vstupuje s najväčšou váhou), nie je až také prekvapivé, že sa ju model pokúsil aproximovať čo najlepšie, a to až do tej miery, že jej bol takpovediac dedikovaný jeden zo 4 archetypov.

Porovnať výsledky s tými z roku 2023 je o niečo zložitejšie, nakoľko pre tento rok sa ako najvhodnejšie p javí 6 alebo 8. Ak položíme $P = 6$, opäť sa vyskytnú archetypy maďarskej strany a strany mestského voliča. Taktiež dôjde k vzniku archetypu, ktorý hrá úlohu najmä pri aproximácii strany OĽANO, hoci tiež súvisí aj s KDH. Jeden archetyp bol úzko spojený so stranou SMER - SD (skrz koeficienty α aj β, archetyp a strana sú teda takmer totožné), kým ďalším bola aproximovaná strana HLAS - SD (hoci tá samotná doň nevstupovala skrz koeficienty β). Posledný archetyp Potom opäť súvisel prevažne s rôznorodými malými stranami, a preto nie je až tak dobre interpretovateľný.

Dáta aj niektoré výstupy možno nájsť na https://github.com/vincze16/name/tree/main/AA.

	1	2	3	4	Váha
#SIEŤ	0.00	1.00	0.00	0.00	0.20
Demokrati Slovenska - Ľudo Kaník	0.00	0.00	0.03	0.97	0.00
Kresťanskodemokratické hnutie	0.00	0.20	0.67	0.13	0.17
KOALÍCIA SPOLOČNE ZA SLOVENSKO	0.00	0.14	0.50	0.36	0.00
Komunistická strana Slovenska	0.00	0.00	0.84	0.16	0.02
MKDA - MKDS	1.00	0.00	0.00	0.00	0.00
MOST - HÍD	1.00	0.00	0.00	0.00	0.23
OĽANO - NOVA	0.00	0.76	0.24	0.00	0.39
Odvaha - Veľká národná a proruská koalícia	0.00	0.00	0.00	1.00	0.00
PRIAMA DEMOKRACIA	0.00	0.21	0.05	0.74	0.00
SDKÚ - DS	0.00	1.00	0.00	0.00	0.01
Slovenská občianska koalícia	0.00	1.00	0.00	0.00	0.03
SME RODINA - Boris Kollár	0.01	0.37	0.62	0.00	0.23
SMER - sociálna demokracia	0.00	0.00	1.00	0.00	1.00
SMK - MKP	1.00	0.00	0.00	0.00	0.14
Slovenská národná strana	0.00	0.00	1.00	0.00	0.31
STRANA MODERNÉHO SLOVENSKA (SMS)	0.00	0.00	0.11	0.89	0.01
Strana zelených Slovenska	0.00	0.49	0.51	0.00	0.02
Sloboda a Solidarita	0.00	1.00	0.00	0.00	0.43
Strana TIP	0.00	0.00	0.00	1.00	0.03
VZDOR - strana práce	0.00	0.00	0.66	0.34	0.00
Kotleba - Ľudová strana Naše Slovensko	0.00	0.00	0.75	0.25	0.28
ŠANCA	0.00	0.00	0.00	1.00	0.01

Tabuľka 6: Hodnoty koeficientov α pri vážených archetypoch (kraje - 2016) a príslušné váhy pozorovaní

	1	2	3	4	Váha
#SIEŤ	0.10	0.00	0.00	0.00	0.20
Demokrati Slovenska - Ľudo Kaník	0.00	0.16	0.00	0.00	0.00
Kresťanskodemokratické hnutie	0.13	0.03	0.01	0.09	0.17
KOALÍCIA SPOLOČNE ZA SLOVENSKO	0.05	0.00	0.00	0.17	0.00
Komunistická strana Slovenska	0.18	0.00	0.02	0.15	0.02
MKDA - MKDS	0.06	0.07	0.00	0.05	0.00
MOST - HÍD	0.00	0.00	0.00	0.00	0.23
OĽANO - NOVA	0.00	0.00	0.00	0.00	0.39
Odvaha - Veľká národná a proruská koalícia	0.00	0.00	0.01	0.13	0.00
PRIAMA DEMOKRACIA	0.06	0.00	0.01	0.00	0.00
SDKÚ - DS	0.00	0.05	0.00	0.01	0.01
Slovenská občianska koalícia	0.00	0.00	0.00	0.00	0.03
SME RODINA - Boris Kollár	0.00	0.01	0.01	0.00	0.23
SMER - sociálna demokracia	0.00	0.00	0.91	0.00	1.00
SMK - MKP	0.19	0.00	0.00	0.08	0.14
Slovenská národná strana	0.00	0.09	0.03	0.10	0.31
STRANA MODERNÉHO SLOVENSKA (SMS)	0.00	0.08	0.00	0.00	0.01
Strana zelených Slovenska	0.00	0.00	0.00	0.00	0.02
Sloboda a Solidarita	0.13	0.52	0.00	0.13	0.43
Strana TIP	0.00	0.00	0.00	0.00	0.03
VZDOR - strana práce	0.00	0.00	0.00	0.00	0.00
Kotleba - Ľudová strana Naše Slovensko	0.00	0.00	0.00	0.00	0.28
ŠANCA	0.10	0.00	0.00	0.09	0.01

Tabuľka 7: Hodnoty koeficientov β pri vážených archetypoch (kraje - 2016) a príslušné váhy pozorovaní

4 Meranie polarizácie v dátach

V kap. 2.1 nás po pohľade na ternary ploty, vizualizujúce dáta z rokov 2016 a 2020 (Obr. 5 a 6) zaujal kvalitatívny rozdiel medzi nimi. Kým pri dátach z roku 2016 boli aproximácie pozorovaní, nachádzajúce sa okolo archetypov 2 a 3 a medzi nimi, pomerne rovnomerne rozložené, pre rok 2020 došlo ku vzniku 2 výrazne oddelených zhlukov, sústredených okolo týchto archetypov. Táto zmena môže súvisieť so spoločenskými udalosťami medzi rokmi 2016 a 2020, zodpovedanie tejto otázky však nie je predmetom našej práce. My sa pokúsime vyriešiť problém, na ktorý by akýkoľvek ďalší výskum v tejto oblasti narazil hneď na svojom začiatku. Kvalitatívnu zmenu medzi výsledkami z rokov 2016 a 2020 sme totiž dosiaľ popisovali len na základe vizuálneho porovnania 2 grafických výstupov. My v tejto časti práce predstavíme spôsob ako takúto „zhlučenosť dát okolo extrémnych pozícií“ (ktorú nazveme polarizáciou) exaktne merať. Náš postup budeme demonštrovať na dátach, s ktorými sme pracovali v kap. 2.1.

Vzhľadom na to, ako sme polarizáciu (nateraz voľne) definovali, je prirodzené pozrieť sa na metódy analýzy zhlukov a ďalej stavať na nástrojoch, ktoré nám poskytnú. Špecificky budeme pracovať s metódou K-means, ako je popísaná v [9]. Zaujímať nás budú najmä sumy štvorcov, s ktorými sa pri nej pracuje. Prvou z nich je minimalizovaná vnútrozhluková suma štvorcov (within-cluster sum of squares (WSS)), definovaná ako

$$WSS = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2 , \tag{30}$$

kde

$$\mu_k = \frac{1}{n_k} \sum_{i \in C_k} x_i$$

a C_k je množina indexov pozorovaní, zaradených do k-teho zhluku, a n_k je veľkosť tejto množiny. Táto suma štvorcov popisuje tú časť variancie v dátach, ktorú model nedokáže popísať (ekvivalent RSS pri lineárnej regresii). Okrem WSS pre nás bude dôležitá celková suma štvorcov (total sum of squares (TSS)), daná ako

$$TSS = \sum_{i=1}^{n} ||x_i - \bar{x}||^2 , \tag{31}$$

kde $\bar{x}$ značí výberový priemer. Táto suma štvorcov vyjadruje celkovú varianciu v dátach. Správanie WSS a TSS popíšeme pomocou nasledovnej lemmy. Táto hovorí o všeobecne

známych vlastnostiach K-means, no my ju aj napriek tomu uvedieme aj s dôkazom, nakoľko ten pre nás bude užitočný v ďalších častiach práce.

Lema 4.1. *Nech* $\{C_1, ..., C_{K_1}\}$ *je zhlukovanie, využívajúce metódu K-means, nad datasetom, daným ako* $X = \{x_1, ..., x_n\}$*, s vnútrozhlukovou sumou štvorcov* WSS*. Nech* $\left\{\tilde{C}_1, ..., \tilde{C}_{K_2}\right\}$ *je zhlukovanie (takisto využívajúce K-means) nad tým istým datasetom s vnútrozhlukovou sumou štvorcov* $\widetilde{WSS}$ *a* $K_2 > K_1$*. Potom*

1. *ak* $K_1 = 1$*, tak* $WSS = TSS$*.*

2. *ak* $K_1 = n$*, tak* $WSS = 0$*.*

3. $\widetilde{WSS} \leq WSS$*.*

Dôkaz. Dokážeme postupne všetky 3 časti lemmy.

1. Ak $K_1 = 1$, potom máme iba 1 zhluk, obsahujúci všetky pozorovania, ktorého priemer je rovný výberovému priemeru. Preto

$$WSS = \sum_{k=1}^{K_1} \sum_{i \in C_k} ||x_i - \mu_k||^2 = \sum_{i=1}^{n} ||x_i - \bar{x}||^2 = TSS.$$

2. Ak $K_1 = n$, potom máme n zhlukov, pričom každý z nich obsahuje práve 1 pozorovanie, teda $\mu_k = x_k$. Preto

$$WSS = \sum_{k=1}^{K_1} \sum_{i \in C_k} ||x_i - \mu_k||^2 = \sum_{i=1}^{n} ||x_i - x_i||^2 = 0. \tag{32}$$

3. Nech $K_2 > K_1$. Vezmime $\hat{K} = K_1 + 1$ a oddelením jedného pozorovania od niektorého zo zhlukov pôvodného zhlukovania vytvorme nové zhlukovanie $\left\{\hat{C}_1, ..., \hat{C}_{\hat{K}}\right\}$ s vnútrozhlukovou sumou štvorcov označenou ako $\widehat{WSS}$. Bez ujmy na všeobecnosti, nech takto rozdelíme zhluk, daný množinou C_{K_1} (teda $C_1 = \hat{C}_1, ..., C_{K_1-1} = \hat{C}_{K_1-1}$)

a nech je oddeleným pozorovaním x_j. Potom platí

$$\begin{aligned}
WSS &= \sum_{k=1}^{K_1}\sum_{i\in C_k}||x_i-\mu_k||^2 = \sum_{k=1}^{K_1-1}\sum_{i\in C_k}||x_i-\mu_k||^2 + \sum_{i\in C_{K_1}}||x_i-\mu_{K_1}||^2 \\
&= \sum_{k=1}^{K_1-1}\sum_{i\in C_k}||x_i-\mu_k||^2 + \sum_{i\in \hat{C}_{K_1}}||x_i-\mu_{K_1}||^2 + ||x_j-\mu_{K_1}||^2 \\
&\geq \sum_{k=1}^{K_1-1}\sum_{i\in C_k}||x_i-\mu_k||^2 + \sum_{i\in \hat{C}_{K_1}}||x_i-\mu_{K_1}||^2 \\
&= \sum_{k=1}^{K_1-1}\sum_{i\in C_k}||x_i-\mu_k||^2 + \sum_{i\in \hat{C}_{K_1}}||x_i-\mu_{K_1}||^2 + ||x_j-\hat{\mu}_{\hat{K}}||^2 \\
&\geq \sum_{k=1}^{K_1-1}\sum_{i\in C_k}||x_i-\mu_k||^2 + \sum_{i\in \hat{C}_{K_1}}||x_i-\hat{\mu}_{K_1}||^2 + ||x_j-\hat{\mu}_{\hat{K}}||^2 \\
&= \widehat{WSS}.
\end{aligned}$$

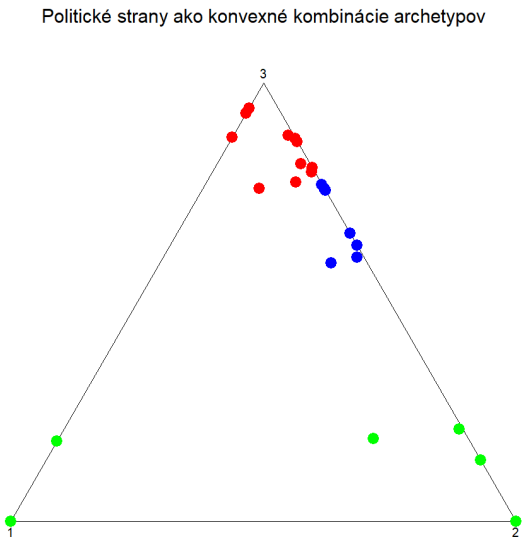
Pre $\hat{K} = K+1$ sme teda našli zhlukovanie, ktorého WSS je menšie alebo rovné ako bolo pri pôvodnom zhlukovaní s K_1 zhlukmi. Takýmto postupným oddeľovaním po 1 pozorovaní sa vieme postupne dostať ku K_2 zhlukom, pričom v žiadnom kroku nebude potrebné zvýšiť hodnotu WSS. Zhlukovanie, ku ktorému sa takto dopracujeme, navyše ešte nemusí byť optimálne v zmysle K-means, teda hodnotu účelovej funkcie (WSS) môže byť možné ďalej znižovať. Preto $\widetilde{WSS} \leq \widehat{WSS} \leq WSS$.

□

Pomocou WSS a TSS sa definuje koeficient determinácie

$$R^2 = 1 - \frac{WSS}{TSS}. \tag{33}$$

Z lemmy 4.1 vyplýva, že $0 \leq WSS \leq TSS$, a teda $0 \leq R^2 \leq 1$. Takýto ukazovateľ preto vyjadruje, akú pomernú časť variancie dát dokáže vysvetliť model K-means, presnejšie, ako dobre sú jednotlivé pozorovania rozdeliteľné do zhlukov. Toto už niektorými vlastnosťami nápadne pripomína to, čo hľadáme. My však od polarizácie požadujeme, aby bola mierou „zhlučenosti dát okolo extrémnych pozícií“ a nie iba mierou zhlučenosti vo všeobecnosti. Pokiaľ by sme na naše dáta aplikovali iba štandardné K-means, výsledné zhlukovanie by mohlo vyzerať napríklad ako to, ktoré máme znázornené na Obr. 13. Toto však nie je vhodné zhlukovanie, nakoľko modrý zhluk sa nenachádza v žiadnej z extrémnych pozícií,



Obr. 13: Nevhodné zhlukovanie pre meranie polarizácie

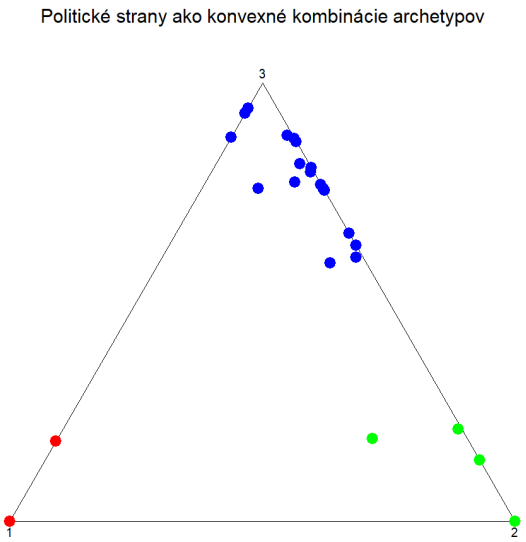
prináležiacich 3 nájdeným archetypom, zatiaľ čo zelený zhluk sa rozkladá až v dvoch. Náš postup preto budeme musieť ďalej modifikovať.

Hlavnou zmenou bude nahradenie štandardného K-means ohraničeným, bližšie popísaným v [3], pri implementácii využívajúc balík conclust [8]. V ňom zahrnutá funkcia *ckmeans* nám dovoľuje okrem požadovaného počtu zhlukov zadať aj ohraničenia typu *mustlink*, požadujúce, aby boli nami špecifikované množiny objektov umiestnené do rovnakých zhlukov, a *cantlink*, vďaka ktorým nami špecifikované dvojice objektov nebudú umiestnené do rovnakého zhluku. Pre nás budú kľúčové ohraničenia typu *cantlink*, pomocou ktorých dosiahneme, že pozorovania, nachádzajúce sa v najextrémnejších pozíciách, nebudú môcť byť umiestnené po dvojiciach do toho istého zhluku. Vďaka tomu dosiahneme zhlukovanie okolo nájdených p archetypov (Obr. 14).

Namiesto pôvodných dát budeme používať ich aproximácie pomocou konvexných kombinácií archetypov (reprezentovaných koeficientami α). Postup bude potom nasledovný:

1. Nájsť p „najarchetypálnejších“ objektov, teda pre každé $k = 1, ..., p$ nájsť j také, že $\alpha_{jk} \geq \alpha_{ik} \quad \forall\, i = 1, ..., n.$

2. Na dáta, reprezentované pomocou koeficientov α (matica A z (8)), aplikujeme ohraničené K-means, pričom ohraničeniami *cantlink* zabezpečíme, aby v predchá-



Obr. 14: Vhodné zhlukovanie pre meranie polarizácie

dzajúcom bode nájdené „najarchetypálnejšie“ objekty neboli umiestnené po dvojiciach do rovnakého zhluku.

3. Pre takto získané zhlukovanie spočítame R^2 podľa (33). Takto získanú hodnotu nazveme polarizáciou.

Keď takýto spôsob merania postupne aplikujeme na dáta z rokov 2016, 2020 a 2023 pre $p = 3$ (Obr. 5, 6 a 7), dostaneme hodnoty 0,77 pre rok 2016, 0,88 pre rok 2020 a 0,82 pre rok 2023, čo sa zhoduje s intuíciou, ktorú možno nadobudnúť pri pohľade na obrázky. Najnižšia hodnota bola dosiahnutá pre rok 2016, kedy boli aproximácie veľkej časti pozorovaní pomerne rovnomerne rozložené medzi archetypmi 2 a 3. Pri dátach z roku 2020 došlo k vzniku 2 výrazne oddelených zhlukov okolo tých istých archetypov, preto neprekvapí, že práve pri tomto roku bola dosiahnutá najvyššia hodnota. V roku 2023 bola situácia podobná, avšak s tým rozdielom, že došlo k vzniku ďalšieho zhluku strán, nachádzajúceho sa medzi archetypmi 2 a 3, a preto sa dá očakávať, že sa bude výsledná polarizácia nachádzať medzi hodnotami z rokov 2016 a 2020.

Miera polarizácie, ako sme ju definovali, sa môže meniť v závislosti od zvoleného počtu archetypov v modeli. Porovnanie medzi jednotlivými rokmi, ktoré sme urobili v predchádzajúcom odseku, preto nemusí byť celkom korektné, nakoľko sme na Obr. 4 videli, že $p = 3$ zrejme nebude vhodný počet archetypov pre rok 2016, a podobná situácia nastala

aj pre zvyšné roky. Toto porovnanie bolo užitočné na ilustráciu želaných vlastností nášho spôsobu merania polarizácie a jeho súhlasu so skôr nadobudnutou intuíciou, avšak na získanie relevantnejších výsledkov uskutočníme v kap. 4.1 rozsiahlejšiu analýzu.

4.1 Porovnanie polarizácie v rokoch 2016, 2020 a 2023

Prvým krokom našej analýzy bude výber vhodného počtu archetypov. Nakoľko tu nepracujeme s 1 datasetom, ale s 3, pričom pre každý budeme fitovať osobitný model, jednoduchý lakťový diagram nám tu, žiaľ, nepomôže. Budeme využívať koeficient determinácie pre model AA, ktorý označíme štandardne ako R^2 a spočíta sa ako

$$R^2 = 1 - \frac{RSS}{TSS}, \tag{34}$$

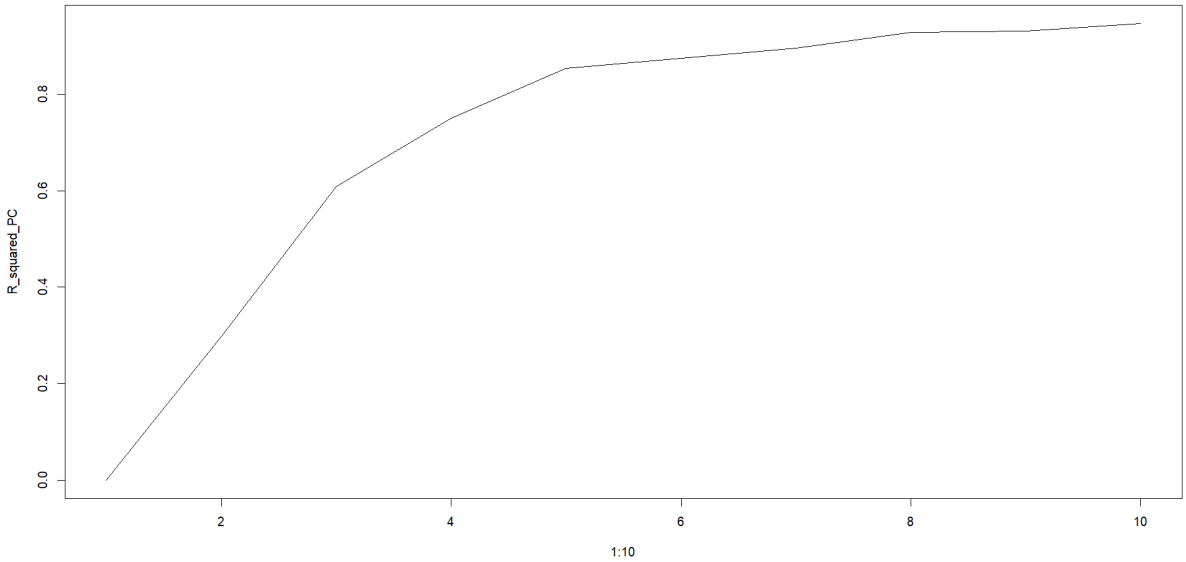
pričom RSS je minimalizovaná suma štvorcov z (2) a TSS sa definuje rovnako pri K-means. Pre RSS pri AA sformulujeme podobnú lemmu ako pre WSS pri K-means.

Lema 4.2. *Nech RSS je reziduálna suma štvorcov pre model AA s p_1 archetypmi. Nech $\widetilde{RSS}$ je reziduálna suma štvorcov pre model AA s p_2 archetypmi. Nech $p_2 > p_1$. Potom $RSS \geq \widetilde{RSS}$.*

Dôkaz. Vo veľkej miere analogický s dôkazom 3. časti lemmy 4.1. Namiesto oddelenia jedného z pozorovaní od jedného zo zhlukov však pridáme nový archetyp - jedno z pozorovaní, patriacich do konvexného obalu datasetu, ktoré ešte nepatrí do konvexného obalu množiny dosiaľ nájdených archetypov (ak také existuje). Ak také neexistuje, konvexný obal už nájdených archetypov pokrýva celý konvexný obal datasetu a nastáva rovnosť. □

Okrem toho z Vety 1.1 vyplýva, že, ak $p = 1$, potom $RSS = TSS$ a, ak $p \geq N$ (počet pozorovaní, ležiacich na hranici konvexného obalu datasetu), potom $RSS = 0$. Z toho ľahko vyplynie, že pre R^2 - koeficient determinácie pre model AA - tiež platí $0 \leq R^2 \leq 1$. Opäť teda možno povedať, že sa jedná o pomerné vyjadrenie toho, akú časť variancie dát model popisuje.

Pre každý z rokov 2016, 2020 a 2023 nájdeme najlepší model AA pre $p = 2, ..., 10$ a spočítame preň R^2. Takto získame sériu 9 trojíc hodnôt, na základe ktorých potrebujeme zvoliť vhodný počet archetypov, ktorý použijeme pre všetky roky. Takto získané trojice potom využijeme ako vstupy funkcie užitočnosti (Cobb-Douglasova funkcia s rôznymi



Obr. 15: Výber vhodného počtu archetypov pre všetky datasety - funkcia dokonale komplementárnych vstupov

hodnotami parametrov α, β a γ, funkcia dokonale zameniteľných a dokonale substitučných vstupov). Pre každú hodnotu p tak dostaneme 1 hodnotu namiesto 3 a pre takto vzniknutú postupnosť už môžeme použiť postup podobný lakťovému diagramu pre výber vhodného spoločného počtu archetypov. Na x-ovej osi budú podobne ako predtým vystupovať počty archetypov, zatiaľ čo na y-ovej výstupy niektorej z funkcií užitočnosti. Rovnako ako pri lakťovom diagrame, aj tu hľadáme počet, pri ktorom sa vzniknutá krivka láme. Takýto zlom naznačuje, že sa nám neoplatí ďalej zvyšovať zložitosť modelov, lebo to vzhľadom na vybranú funkciu užitočnosti prestáva viesť k výraznému skvalitneniu aproximácie.

Na Obr. 15 vidím, ako môže takáto krivka vyzerať pre funkciu dokonale komplementárnych vstupov. Vidíme, že najvýraznejšie zlomy nastávajú pre $p = 3$ a $p = 5$, my zvolíme $p = 5$, nakoľko pre tento počet archetypov vznikli pomerne výrazné zlomy aj pri použití iných druhov funkcie užitočnosti.

Po voľbe p je už zvyšok postupu rovnaký ako v prvej časti kapitoly. Pozorovania máme pre každý rok aproximované ako konvexné kombinácie 5 archetypov, teda každému zodpovedá 5 koeficientov α. Pre každý rok nájdeme 5 „najarchetypálnejších“ objektov a na aproximácie dát (reprezentované maticou A) potom aplikujeme ohraničené K-means. Pre takto vzniknuté zhlukovanie následne pre každý rok spočítame hodnotu R^2. Výsledné

hodnoty sú približne 0,77 pre rok 2016, 0,80 pre rok 2020 a 0,73 pre rok 2023. Najvyššia polarizácia vyšla opäť pre rok 2020, avšak prekvapiť môže fakt, že hodnota pre rok 2023 je nižšia ako pre rok 2016. Tento výsledok ilustruje dôležitosť voľby počtu archetypov pri meraní polarizácie a tiež nás varuje pred vyslovovaním príliš odvážnych záverov na základe (grafických) výstupov modelov, ktoré boli fitované pre nevhodné p.

4.2 Zahrnutie váh

V tejto časti práce budeme postup 4 ďalej vylepšovať a zovšeobecňovať zahrnutím váh pre jednotlivé pozorovania. Ak totiž budeme všetky pozorovania považovať za rovnocenné, môžeme zanedbávať dôležitú informáciu a dospieť tak k výsledkom, ktoré nezodpovedajú realite. Pre ilustráciu sa pozrime na aproximácie dát z roku 2016 a Obr. 5. Našimi „najarchetypálnejšími“ stranami sú SMK, SaS a TIP. Uvažujme nasledovné 2 scenáre. V prvom sú preferencie spomenutých 3 strán po 30%, kým zvyšné strany si medzi sebou delia ostávajúcich 10%. V druhom by naopak podpora tejto trojice strán bola nízka a väčšinu hlasov by si medzi sebou delili strany nachádzajúce sa v menej extrémnych pozíciách, medzi archetypmi 2 a 3 (napr. PRIAMA DEMOKRACIA, SMER...). Ak sa cez naše dáta chceme pozrieť na polarizáciu spoločnosti, tieto scenáre popisujú 2 výrazne odlišné situácie. Ak si totiž uvedomíme, že každá strana reprezentuje určitú časť populácie, potom v prvom scenári sa väčšina populácie (resp. strany, ktoré ju reprezentujú) nachádza v extrémnych pozíciách, kým v druhom tomu tak nie je a týmto pozíciám prináleží len veľmi malá časť spoločnosti. Hoci by sme viac polarizovanou nazvali spoločnosť v prvom scenári, postup 4 by medzi nimi nerozlišoval a obom by prisúdil rovnakú hodnotu polarizácie. Ako prirodzené riešenie tohto problému sa ponúka zahrnúť do nášho postupu váhy pozorovaní (v tomto prípade preferencie jednotlivých strán), pričom bude našou snahou, aby mali na finálny výsledok väčší vplyv pozorovania s vyššou váhou (strany s vyššími preferenciami). Aby sme toto dosiahli, do postupu 4 zahrnieme 2 modifikácie, ktoré popíšeme v nasledujúcich častiach práce.

Pre naše váhy budeme používať značenie $w_1, ..., w_n$, pričom w_i je váha zodpovedajúca pozorovaniu x_i.

4.2.1 Váhy pri výpočte R^2 pre zhlukovanie

Prvou (intuitívnou) modifikáciou je zahrnutie váh pri výpočte R^2 pre zhlukovanie, nájdené v rámci postupu 4. Presnejšie, nahradíme sumy štvorcov z (30) a (31) ich váženými verziami. Dostaneme

$$WSS_w = \sum_{k=1}^{K} \sum_{i \in C_k} w_i \left|\left|x_i - \mu_k\right|\right|^2 \tag{35}$$

$$TSS_w = \sum_{i=1}^{n} w_i \left|\left|x_i - \mu\right|\right|^2 . \tag{36}$$

Vážená verzia koeficientu determinácie sa potom bude počítať rovnako ako predtým, teda

$$R_w^2 = 1 - \frac{WSS_w}{TSS_w}. \tag{37}$$

Celý postup potom vyzerá nasledovne:

1. Nájsť p „najarchetypálnejších“ objektov, teda pre každé $k = 1, ..., p$ nájsť j také, že $\alpha_{jk} \geq \alpha_{ik} \quad \forall\, i = 1, ..., n.$

2. Na dáta, reprezentované pomocou koeficientov α (matica a z (8)), aplikujeme ohraničené K-means, pričom ohraničeniami *cantlink* zabezpečíme, aby v predchádzajúcom bode nájdené „najarchetypálnejšie“ objekty neboli umiestnené po dvojiciach do rovnakého zhluku.

3. Pre takto získané zhlukovanie spočítame R_w^2 podľa (37). Takto získanú hodnotu nazveme polarizáciou.

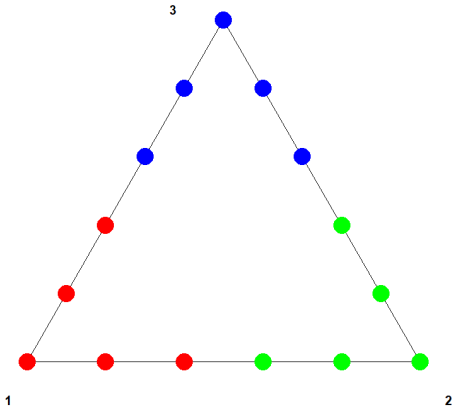
Je dôležité si uvedomiť, že pre WSS_w, ako je definované v (35), neplatí vážený ekvivalent lemmy 4.1, konkrétne jej tretia časť. Platnosť prvej a druhej časti by sa ukázala analogicky a pre tretiu ukážeme jednoduchý kontrapríklad. Majme 1-rozmerné dáta a ich váhy, popísané v Tab. 8. Pre $K = 1$ platí $WSS_w = TSS_w = 2108$. Ak by sme potom K zvýšili na 2, algoritmus K-means (pre $K = 2$) by dataset rozdelil na zhluk obsahujúci prvých 5 pozorovaní (-5 až -1) a zhluk obsahujúci druhých 5 pozorovaní (1 až 5). WSS_w potom stúpne na 8012, hoci sa zvýšilo K. Toto nás neprekvapí, pretože, kým porovnávané sumy štvorcov už zahŕňajú váhy, algoritmus hľadajúci optimálne zhlukovanie na ne ohľad neberie.

i	x_i	w_i
1	-5	1
2	-4	1
3	-3	1
4	-2	1
5	-1	1000
6	1	1000
7	2	1
8	3	1
9	4	1
10	5	1

Tabuľka 8: Modelové jednorozmerné dáta, ilustrujúce neplatnosť niektorých vlastností po nahradení štandardných súm štvorcov ich váženými verziami

Ešte dôležitejším je pre nás pozorovanie, že WSS_w môže byť väčšie ako TSS_w. V dôsledku toho vieme váženú verziu koeficientu determinácie ohraničiť iba zhora, a to $R_w^2 \leq 1$. Tento ukazovateľ teda môže teoreticky nadobudnúť aj záporné hodnoty a nedá sa tak korektne povedať, že by vyjadroval pomernú časť variancie, popísanej modelom. Zároveň však považujeme za nutné zdôrazniť, že dáta, ktorými sme to ilustrovali, boli umelé, zvolené zámerne tak, aby nastala práve takáto situácia. Pri našej práci s takto zostrojeným ukazovateľom (na umelých aj reálnych dátach) podobný prípad nenastal.

Postup 4.2.1 aplikujeme najprv na umelé dáta a demonštrujeme tak, že skutočne želaným spôsobom zlepšuje správanie pôvodného algoritmu. Bude sa jednať o dáta, simulujúce situáciu s 3 archetypmi, aby sme mali k dispozícii zobrazenie pomocou ternary plotu. Dáta možno vidieť znázornené na Obr. 16, pričom sú už pomocou ohraničeného K-means rozdelené do 3 zhlukov okolo „najarchetypálnejších“ objektov a patrične farebne odlíšené. V prvom prípade budeme uvažovať rovnaké (jednotkové) váhy, teda situáciu ekvivalentnú s postupom 4. Takýmto spôsobom dostávame hodnotu polarizácie, rovnú približne 0,79. Uvažujme ďalej situáciu, kde váhy pre všetky pozorovania rovnaké nie sú, konkrétne priradíme vyššie váhy pozorovaniam, nachádzajúcim sa medzi archetypmi, v menej extrémnych pozíciách. Túto situáciu zobrazuje Obr. 17, pričom váhy jednotlivých pozorovaní sú zná-

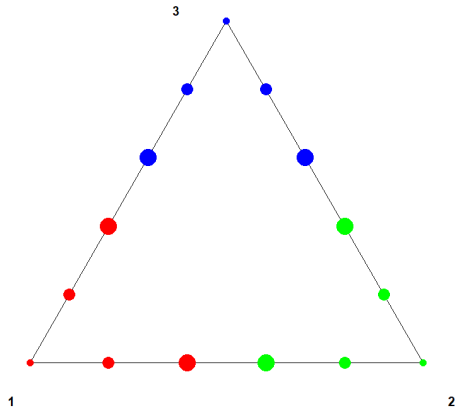


Obr. 16: Umelé dáta - rovnaké váhy všetkých pozorovaní

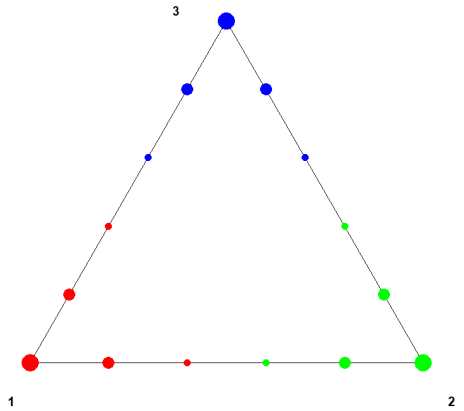
zornené veľkosťou jednotlivých bodov. Od nášho spôsobu merania polarizácie požadujeme, aby bol výsledok nižší ako v predchádzajúcom prípade, nakoľko je váha koncentrovaná v menej extrémnych pozíciách. Aplikovaním postupu 4.2.1 získame hodnotu približne 0,73, teda skutočne došlo k zníženiu polarizácie. Keď sa potom pozrieme na Obr. 18, vidíme opačnú situáciu, teda pozorovania s vyššími váhami v extrémnejších pozíciách, preto očakávame aj vyššiu mieru polarizácie. Postupom 4.2.1 získavame hodnotu približne 0,84, čo je viac ako pôvodných 0,79, získaných bez zahrnutia váh.

Ukážme si teraz ešte 1 extrémnejší prípad. Na Obr. 19 vidíme situáciu, kde máme prevažnú väčšinu váhy sústredenú v strede, v najmenej extrémnych pozíciách. Preto by sme očakávali nízku mieru polarizácie, ktorú aj dostávame (0,37). Zdalo by sa teda, že náš spôsob merania sa správa práve tak, ako od neho požadujeme, no toto nanešťastie súvisí s problémom neplatnosti váženej verzie lemmy 4.1, o ktorom sme pojednávali skôr v tejto podkapitole. Ak by sme totiž pokračovali so zvyšovaním váh 6 pozorovaní, nachádzajúcich sa v strede, dospeli by sme (po asi 3-násobnom zvýšení) do bodu, v ktorom R^2 začne nadobúdať záporné hodnoty. Opäť ale podotýkame, že sa jedná o umelé dáta, konštruované špeciálne za účelom ilustrácie tohto problému.

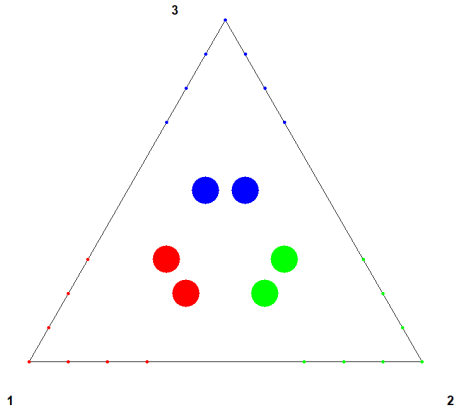
Takto upravený spôsob merania polarizácie teraz aplikujeme na naše dáta o voľbách do NR SR, postupujúc podobne ako v kap. 4.1.



Obr. 17: Umelé dáta - veľkosti bodov zodpovedajú váham pozorovaní.



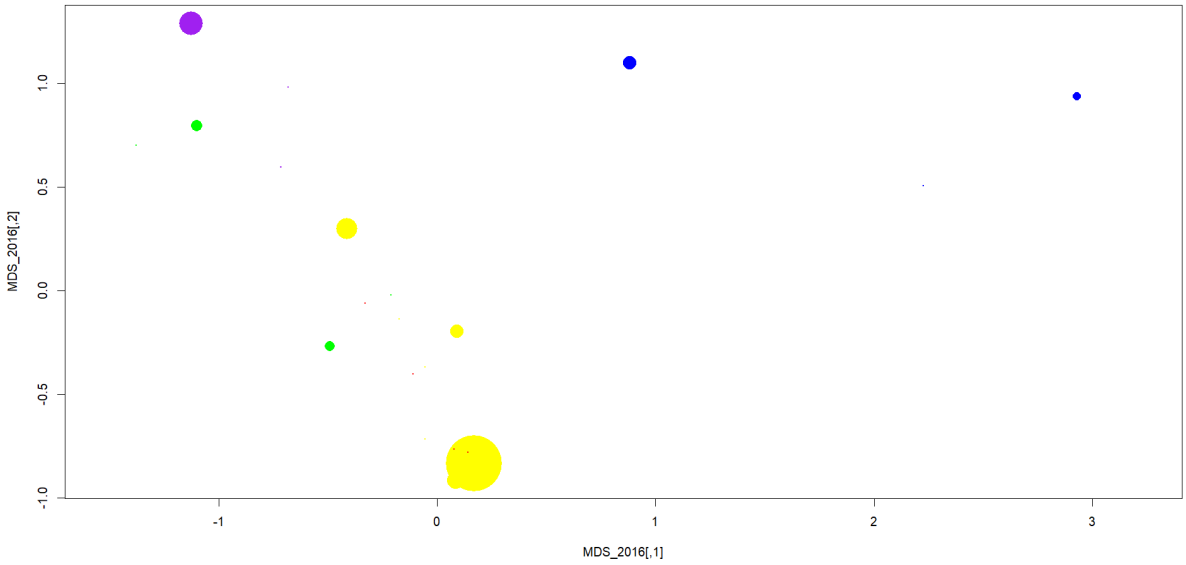
Obr. 18: Umelé dáta - veľkosti bodov zodpovedajú váham pozorovaní.



Obr. 19: Umelé dáta - veľkosti bodov zodpovedajú váham pozorovaní.

Najprv pomocou R^2 pre model AA a funkcií užitočnosti zvolíme najvhodnejší počet archetypov, ktorý potom použijeme pre dáta z rokov 2016, 2020 a 2023. Táto časť postupu je nezmenená a používame tie isté dáta ako predtým, preto opäť zvolíme $p = 5$ (Obr. 15). Na dáta pre každý rok (aproximované pomocou koeficientov α ako konvexné kombinácie 5 archetypov) aplikujeme postup 4.2.1, ktorý oproti postupu 4 nahrádza WSS a TSS pri výpočte R^2 pre zhlukovanie ich váženými verziami WSS_w (35) a TSS_w (36). Dostávame hodnoty 0,76 pre rok 2016, 0,84 pre rok 2020 a 0,89 pre rok 2023. Pri pohľade do Tab. 10 vidíme, že hodnota pre rok 2016 sa takmer nezmenila, hodnota pre rok 2020 mierne narástla, zatiaľ čo pre rok 2023 došlo k oveľa výraznejšiemu nárastu. Na to, prečo nastala takáto situácia, sa pokúsime nahliadnuť pomocou vizualizácie aproximácií dát, využívajúcej klasické mnohorozmerné škálovanie (Obr. 20, 21, 22).

Pozrime sa na každom z obrázkov na jednotlivé farebne odlíšené zhluky. Uvedomme si, že pre každý zhluk predstavuje jedno doň patriace pozorovanie vybratý „najarchetypálnejší“ objekt, zodpovedajúci jednej z p najextrémnejších pozícií v dátach. Cieľom nami využívaného klasického mnohorozmerného škálovania je nájsť menejrozmernú reprezentáciu dát, čo najvernejšie zachovávajúcu vzájomné vzdialenosti pozorovaní. Rozptýlenosti objektov patriacich do jednotlivých zhlukov nám teda dávajú určitú informáciu o vzdialenostiach týchto objektov od „najarchetypálnejšieho“ objektu (aproximácie) z daného

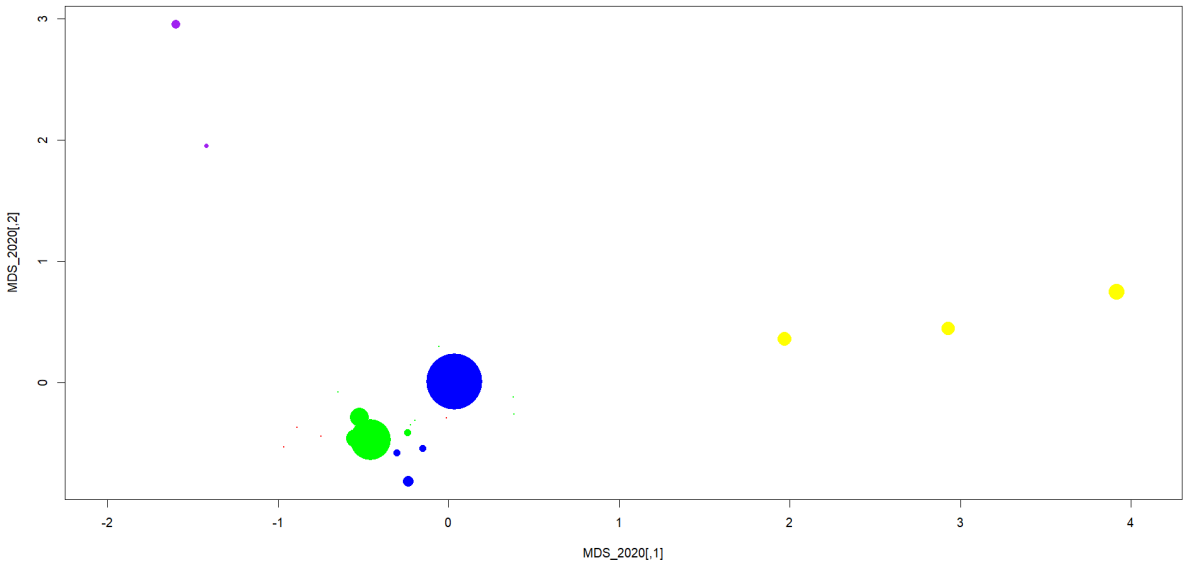


Obr. 20: Aproximácie dát (2016), vizualizované pomocou klasického mnohorozmerného škálovania. Váhy pozorovaní sú znázornené veľkosťami bodov.

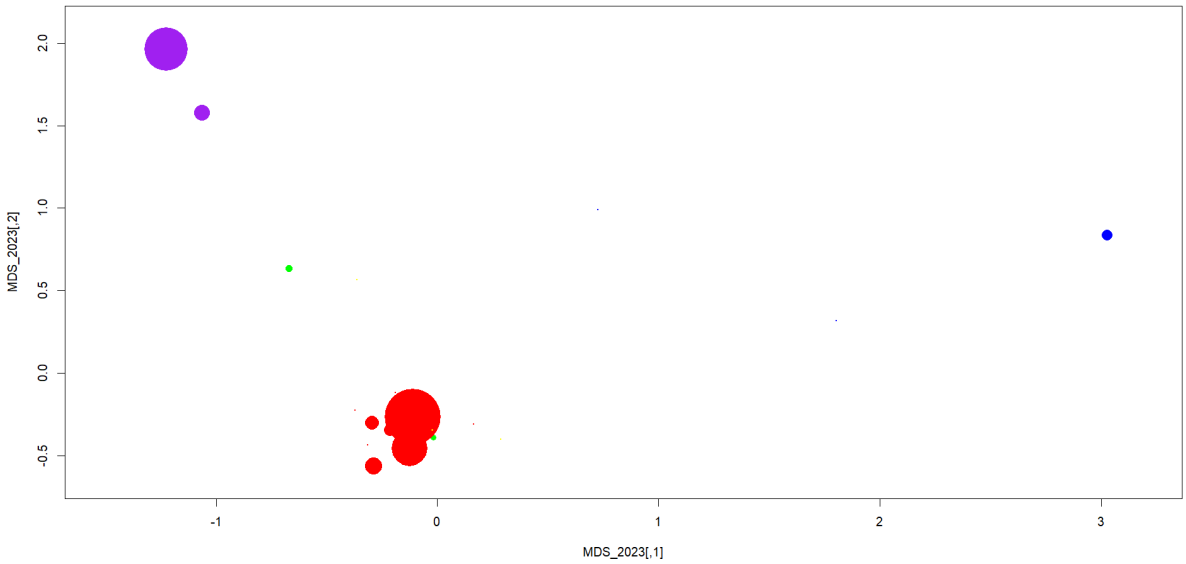
zhluku, a teda aj o extrémnosti ich samotných. Keď sa pozrieme na Obr. 20, vidíme, že zhluk predstavujúci najväčšiu časť váhy (žltý) je pomerne rozptýlený, zatiaľ čo na Obr. 21 sú takéto dva zhluky (modrý a zelený) oveľa kompaktnejšie. Tento jav je ešte výraznejší na Obr. 22 (červený a fialový zhluk). Toto by mohlo do určitej miery vysvetliť, prečo zahrnutie váh spôsobilo zvýšenie polarizácie pri dátach z roku 2020, ale ešte viac pre rok 2023. Zároveň si uvedomujeme, že aplikovaním mnohorozmerného škálovania strácame informáciu, a takúto analýzu preto považujeme skôr za prvý náhľad ako za pokus o definitívne zodpovedanie položenej otázky.

4.2.2 Váhy pri hľadaní zhlukovania - vážené ohraničené K-means

V kap. 4.2.1 sme sumy štvorcov nahradili ich váženými verziami ((35), (36)), avšak samotné zhlukovanie sme hľadali pomocou štandardného (ohraničeného) K-means. Takáto modifikácia mala 2 významné dôsledky. Prvým z nich bolo to, že nebolo zaručené splnenie ohraničenia $0 \leq R_w^2 \leq 1$, presnejšie takto definované R_w^2 mohlo byť záporné, čo nepovažujeme za želanú vlastnosť. Okrem toho váhy neboli žiadnym spôsobom zahrnuté pri hľadaní zhlukovania, ktoré tak nemuselo zodpovedať skutočnej dynamike v dátach. Keďže štandardné K-means pristupuje k pozorovaniam ako k rovnocenným, mohlo do-



Obr. 21: Aproximácie dát (2020), vizualizované pomocou klasického mnohorozmerného škálovania. Váhy pozorovaní sú znázornené veľkosťami bodov.



Obr. 22: Aproximácie dát (2023), vizualizované pomocou klasického mnohorozmerného škálovania. Váhy pozorovaní sú znázornené veľkosťami bodov.

chádzať ku vzniku zhlukov obsahujúcich objekty s malými váhami (strany s nízkymi preferenciami, reprezentujúce malú časť populácie), kým objekty s veľkými váhami (strany, predstavujúce väčšinu voličov) ostali nerozdelené. Z hľadiska merania polarizácie môže byť takéto zhlukovanie skresľujúce a vedúce k neuspokojivým výsledkom. Oba tieto problémy nám pomôže vyriešiť nahradenie štandardného (ohraničeného) K-means váženým (ohraničeným) K-means, ako je popísané v [18] (prípadne menej formálne, ale na pochopenie hlavnej myšlienky potenciálne jednoduchšie v [16]).

Na popísanie hlavnej myšlienky váženého (ohraničeného) K-means využijeme dáta o voľbách do NR SR, s ktorými sme pracovali doteraz, presnejšie aproximácie jednotlivých pozorovaní, využívajúce koeficienty α. Predstavme si, že sa na jednotlivé strany pozeráme ako na množiny ich voličov, ktorých by sme považovali za homogénnych. Namiesto 1 pozorovania by sme ich tak pre každú stranu dostali toľko (identických), koľko ľudí danú stranu volilo. Aplikovaním K-means na takéto dáta by sme dostali rozdelenie spoločnosti, nehľadiace na strany, ale na jednotlivcov. Práve takéto zhlukovanie je pre nás pri meraní polarizácie žiaduce.

Intuitívny postup, pomocou ktorého by sa dalo nájsť také zhlukovanie, by zahŕňal nakopírovanie pozorovaní (strán) podľa ich váh (počtu voličov) a aplikáciu štandardného K-means. Tento postup je však ťažkopádny a pri väčších datasetoch výpočtovo príliš náročný. Preto namiesto neho využijeme váženú verziu K-means, ktorá pracuje s pôvodnými pozorovaniami v nezmenenom počte, priamo však zahŕňa ich váhy.

Táto je podobná štandardnému K-means, ako je popísané v [9], avšak optimalizovaná je vážená verzia vnútrozhlukovej sumy štvorcov, definovaná nasledovne:

$$WSS_{\tilde{w}} = \sum_{k=1}^{K} \sum_{i \in C_k} w_i \left|\left|x_i - \mu_k^w\right|\right|^2, \tag{38}$$

kde

$$\mu_k^w = \frac{1}{\sum_{i \in C_k} w_i} \sum_{i \in C_k} w_i x_i. \tag{39}$$

Podobným spôsobom definujeme aj upravenú verziu celkovej sumy štvorcov

$$TSS_{\tilde{w}} = \sum_{i=1}^{n} w_i \left|\left|x_i - \mu^w\right|\right|^2, \tag{40}$$

kde

$$\mu^w = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i x_i, \tag{41}$$

hoci zvyčajne sa uvažuje situácia, kde $\sum_{i=1}^{n} w_i = 1$. Všimnime si, že oproti váženým sumám štvorcov (35) a (36), využívaných v postupe 4.2.1, v týchto vystupuje namiesto štandardného aritmetického priemeru vážený.

Pred pokračovaním v práci si uvedieme a dokážeme lemmu, ktorú budeme neskôr využívať.

Lema 4.3. *Majme dataset, daný ako* $X = \{x_1, ..., x_n\}$ *a funkciu* $J : \mathbb{R}^m \to \mathbb{R}$*, definovanú ako*

$$J(x) = \sum_{i=1}^{n} w_i \left|\left|x_i - x\right|\right|^2.$$

Potom účelová hodnota tejto funkcie je minimalizovaná pre $x = \mu^w$*.*

Dôkaz. Bez ujmy na všeobecnosti budeme uvažovať $\sum_{i=1}^{n} w_i = 1$.

$$\begin{aligned}
J(x) &= \sum_{i=1}^{n} w_i \left|\left|x_i - x\right|\right|^2 = \sum_{i=1}^{n} w_i \left|\left|(x_i - \mu^w) - (x - \mu^w)\right|\right|^2 \\
&= \sum_{i=1}^{n} w_i \left|\left|x_i - \mu^w\right|\right|^2 - 2\sum_{i=1}^{n} w_i (x_i - \mu^w)^T (x - \mu^w) + \sum_{i=1}^{n} w_i \left|\left|x - \mu^w\right|\right|^2 \\
&= \sum_{i=1}^{n} w_i \left|\left|x_i - \mu^w\right|\right|^2 - 2(x - \mu^w)^T \left(\sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i \mu^w\right) + \sum_{i=1}^{n} w_i \left|\left|x - \mu^w\right|\right|^2 \\
&= \sum_{i=1}^{n} w_i \left|\left|x_i - \mu^w\right|\right|^2 - 2(x - \mu^w)^T (\mu^w - \mu^w) + \sum_{i=1}^{n} w_i \left|\left|x - \mu^w\right|\right|^2 \\
&= \sum_{i=1}^{n} w_i \left|\left|x_i - \mu^w\right|\right|^2 + \sum_{i=1}^{n} w_i \left|\left|x - \mu^w\right|\right|^2 \geq \sum_{i=1}^{n} w_i \left|\left|x_i - \mu^w\right|\right|^2,
\end{aligned}$$

pričom rovnosť nastáva práve pre $x = \mu^w$. □

Pre takto modifikovanú metódu K-means a príslušné sumy štvorcov si uvedieme upravenú verziu lemmy 4.1.

Lema 4.4. *Nech* $w_1, ..., w_n$ *sú váhy. Nech* $\{C_1, ..., C_{K_1}\}$ *je zhlukovanie, využívajúce metódu váženého K-means, nad datasetom, daným ako* $X = \{x_1, ..., x_n\}$*, s váženou vnútrozhlukovou sumou štvorcov* $WSS_{\tilde{w}}$*. Nech* $\left\{\tilde{C}_1, ..., \tilde{C}_{K_2}\right\}$ *je zhlukovanie (takisto využívajúce vážené K-means) nad tým istým datasetom s váženou vnútrozhlukovou sumou štvorcov* $\widetilde{WSS_{\tilde{w}}}$ *a* $K_2 > K_1$*. Potom*

1. *ak* $K_1 = 1$*, tak* $WSS_{\tilde{w}} = TSS_{\tilde{w}}$*.*

2. *ak* $K_1 = n$*, tak* $WSS_{\tilde{w}} = 0$*.*

3. $\widetilde{WSS_{\tilde{w}}} \leq WSS_{\tilde{w}}$.

Dôkaz. Analogický s dôkazom lemmy 4.1, pričom v časti 3 pri poslednej nerovnosti sa využije lemma 4.3, zabezpečujúca, že vážený priemer minimalizuje danú sumu štvorcov. □

Využijúc takto definované vážené sumy štvorcov ((38), (39), (40), (41)) definujeme pre model váženého K-means koeficient determinácie ako

$$R^2_{\tilde{w}} = 1 - \frac{WSS_{\tilde{w}}}{TSS_{\tilde{w}}}. \tag{42}$$

Vďaka lemme 4.4 vieme, že platí

$$0 \leq WSS_{\tilde{w}} \leq TSS_{\tilde{w}},$$

a teda aj

$$0 \leq R^2_{\tilde{w}} \leq 1, \tag{43}$$

teda koeficient determinácie môžeme chápať ako pomernú časť „váženej“ variancie, prítomnej v dátach, ktorú sa modelu podarilo vysvetliť. Alternatívne ho možno interpretovať ako pomernú časť vysvetlenej variancie v modeli, ktorý by namiesto zahrnutia váh pracoval s kopírovaním pozorovaní (voliči strán namiesto strán samotných).

S takouto prípravou môžeme uviesť novú modifikáciu nášho spôsobu merania polarizácie. Postup bude nasledovný:

1. Nájsť p „najarchetypálnejších“ objektov, teda pre každé $k = 1, ..., p$ nájsť j také, že $\alpha_{jk} \geq \alpha_{ik} \quad \forall\, i = 1, ..., n.$

2. Na dáta, reprezentované pomocou koeficientov α (matica a z (8)), aplikujeme vážené ohraničené K-means, pričom ohraničeniami *cantlink* zabezpečíme, aby v predchádzajúcom bode nájdené „najarchetypálnejšie“ objekty neboli umiestnené po dvojiciach do rovnakého zhluku.

3. Pre takto získané zhlukovanie spočítame $R^2_{\tilde{w}}$ podľa (42). Takto získanú hodnotu nazveme polarizáciou.

Táto modifikácia má oproti postupu 4.2.1 niekoľko výhod. Prvou je to, že vďaka (43) nemôže nastať situácia, že polarizácia klesne pod 0. Okrem toho sú výstupy takéhoto

Spôsob zahrnutia váh / situácia	1	2	3	4
Nezahrnuté	0,79	0,79	0,79	0,82
Zahrnuté pri výpočte R^2	0,79	0,73	0,84	0,37
Zahrnuté pri výpočte R^2 aj pri hľadaní zhlukovania	0,79	0,74	0,85	0,68

Tabuľka 9: Porovnanie postupov 4, 4.2.1 a 4.2.2 na umelých dátach. Riadky zodpovedajú jednotlivým postupom, kým stĺpce zodpovedajú kombináciám dát a váh (situáciám).

spôsobu merania vďaka ohraničenosti vhodnejšie na interpretáciu. Druhou výhodou je to, že takýto spôsob merania lepšie využíva informáciu, ktorú máme k dispozícii. Pristupovať totiž pri hľadaní zhlukovania k jednotlivým pozorovaniam ako k rovnocenným neraz príliš zjednodušuje situáciu, čoho ilustráciou sú napríklad naše dáta o voľbách do NR SR, kde rôzne strany reprezentujú výrazne odlišné počty ľudí. Nakoľko sa snažíme nahliadnuť na rozdelenie spoločnosti, je rozumné hľadať zhlukovanie, pozerajúce sa prostredníctvom váh na ľudí, z ktorých táto spoločnosť pozostáva.

Postup 4.2.2 podobne ako postup 4.2.1 najprv odtestujeme na modelových umelých dátach a váhach. Pozrieme sa na situácie z Obr. 16 (situácia 1), 17 (situácia 2), 18 (situácia 3) a 19 (situácia 4). Výsledné hodnoty polarizácie možno vidieť v Tab. 9. Všimnime si, že pri prvých 3 situáciách sa jednotlivé postupy správajú takmer identicky a ku vzniku väčších rozdielov dochádza až pri (výrazne extrémnejšej) situácii 4. Kým pri použití postupu 4.2.1 klesla miera polarizácie na menej ako polovicu (pripomeňme, že ďalším zvyšovaním váh pozorovaní v strede by sa dostala do záporných hodnôt), v prípade postupu 4.2.2 bol tento pokles výrazne nižší. Môžeme sa na to pozrieť tak, že postup 4.2.2 je ovplyvňovaný najmä pozorovaniami s vysokými váhami (ktoré sú jasne oddelené do 3 zhlukov) a zvyšné pozorovania vníma ako akýsi „šum“, ktorého prítomnosť nerobí pozorovania v strede menej extrémnymi (čo súvisí s nahradením štandardného aritmetického priemeru váženým).

Pomocou postupu 4.2.2 budeme opäť analyzovať aj polarizáciu v našich dátach o voľbách do NR SR. Podobne ako v predchádzajúcich častiach práce, porovnáme výstupy pre dáta z jednotlivých rokov a na záver túto modifikáciu nášho spôsobu merania polarizácie porovnáme s predošlými.

Na hľadanie zhlukovania pomocou váženého ohraničeného K-means použijeme našu

Spôsob zahrnutia váh	2016	2020	2023
Nezahrnuté	0,77	0,80	0,73
Zahrnuté pri výpočte R^2	0,76	0,84	0,89
Zahrnuté pri výpočte R^2 aj pri hľadaní zhlukovania	0,86	0,90	0,95

Tabuľka 10: Porovnanie polarizácie pre roky 2016, 2020 a 2023 pri rôznych spôsoboch zahrnutia váh (dáta - kraje)

vlastnú funkciu, ktorá vznikla úpravou funkcie *conclust* (ohraničené K-means) z rovnomenného balíka [8], v ktorej umožníme zahrnúť váhy jednotlivých pozorovaní. Váhy zohľadňujúci K-means algoritmus, ktorý sme použili, je popísaný napríklad v [16]. Je podobný štandardnému Lloydovmu algoritmu, avšak pri hľadaní nových centroidov sa namiesto obyčajného priemeru počíta vážený.

Medziročné porovnanie začneme opäť výberom najvhodnejšieho počtu archetypov pre roky 2016, 2020 a 2023. Postup je rovnaký ako v kap. 4.1, využívajúci tie isté dáta, preto opäť volíme $p = 5$ (Obr. 15). Pre každý z rokov potom nájdeme archetypy a jednotlivé pozorovania aproximujeme ako ich konvexné kombinácie pomocou koeficientov α. Na tieto aproximácie následne aplikujeme postup 4.2.2. Výsledné hodnoty možno vidieť v poslednom riadku Tab. 10 (pripájame aj Tab. 12, obsahujúcu rovnaké porovnanie pre okresy). Pri pohľade na obe tabuľky nás môže zaujať, že pri prechode z postupu 4.2.1 na postup 4.2.2 došlo pri umelých aj reálnych dátach k zachovaniu alebo zvýšeniu miery polarizácie. Toto môže súvisieť s faktom, že pri postupe 4.2.2 využívame zhlukovanie, ktoré minimalizuje príslušnú vnútrozhlukovú sumu štvorcov (a teda aj pomocou nej dopočítané $R^2_{\tilde{w}}$), zatiaľ čo pri postupe 4.2.1 takéto priame prepojenie neexistuje.

Zároveň podotknime, že takéto správanie vo všeobecnosti nemusí nastať. Pozrime sa na umelé dáta z Tab. 11, pomocou ternary plotu znázornené na Obr. 23. Tieto sú rozdelené do 3 zhlukov, pričom toto rozdelenie je rovnaké pri použití váženej aj neváženej verzie ohraničeného K-means. Ďalej sa pozrime na príslušné sumy štvorcov, využívané pri postupoch 4.2.1 a 4.2.2. V tomto prípade platí

i	α_i^1	α_i^2	α_i^3	w_i
1	1	0	0	1
2	0,9	0,05	0,05	1
3	0	1	0	1
4	0	0	1	10

Tabuľka 11: Umelé dáta, využité na porovnanie postupov 4.2.1 a 4.2.2

$$\begin{aligned}
WSS_w &= \sum_{k=1}^{K} \sum_{i \in C_k} w_i \left|\left|x_i - \mu_k\right|\right|^2 \\
&= w_1 \left|\left|x_1 - \mu_1\right|\right|^2 + w_2 \left|\left|x_2 - \mu_1\right|\right|^2 + w_3 \left|\left|x_3 - \mu_2\right|\right|^2 + w_4 \left|\left|x_4 - \mu_3\right|\right|^2 \\
&= w_1 \left|\left|x_1 - \mu_1\right|\right|^2 + w_2 \left|\left|x_2 - \mu_1\right|\right|^2 + w_3 \left|\left|x_3 - x_3\right|\right|^2 + w_4 \left|\left|x_4 - x_4\right|\right|^2 \\
&= w_1 \left|\left|x_1 - \mu_1^w\right|\right|^2 + w_2 \left|\left|x_2 - \mu_1^w\right|\right|^2 + w_3 \left|\left|x_3 - \mu_2^w\right|\right|^2 + w_4 \left|\left|x_4 - \mu_3^w\right|\right|^2 \\
&= WSS_{\tilde{w}} \\
TSS_w &= \sum_{i=1}^{4} w_i \left|\left|x_i - \mu\right|\right|^2 \geq \sum_{i=1}^{4} w_i \left|\left|x_i - \mu^w\right|\right|^2 = TSS_{\tilde{w}},
\end{aligned}$$

pričom posledná nerovnosť je dôsledkom lemy 4.3. Pre naše dáta táto nerovnosť platí ostro, a preto

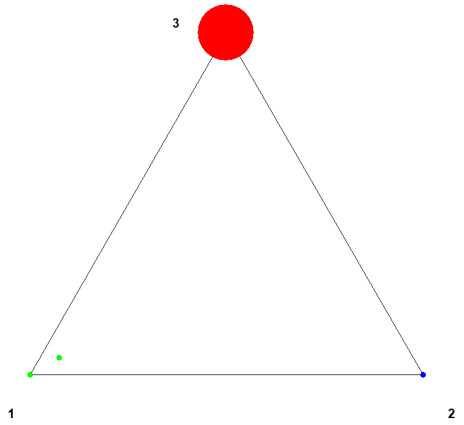
$$R_w^2 = 1 - \frac{WSS_w}{TSS_w} > 1 - \frac{WSS_{\tilde{w}}}{TSS_{\tilde{w}}} = R_{\tilde{w}}^2.$$

Spôsob zahrnutia váh	2016	2020	2023
Nezahrnuté	0,63	0,78	0,87
Zahrnuté pri výpočte R^2	0,69	0,83	0,94
Zahrnuté pri výpočte R^2 aj pri hľadaní zhlukovania	0,78	0,88	0,96

Tabuľka 12: Porovnanie polarizácie pre roky 2016, 2020 a 2023 pri rôznych spôsoboch zahrnutia váh (dáta - okresy)

4.3 Meranie polarizácie, využívajúce vážené archetypy

V kap. 3 sme stručne predstavili koncept vážených archetypov. Ten môžeme prirodzene skombinovať s nami popísanými spôsobmi na meranie polarizácie v dátach (postupy 4,



Obr. 23: Umelé dáta, využité na porovnanie postupov 4.2.1 a 4.2.2 - vizualizácia

4.2.1 a 4.2.2). V prvom kroku v rámci každého z nich sme dáta aproximovali pomocou koeficientov α ako konvexné kombinácie nájdených p archetypov. Tento krok zopakujeme, avšak použitými archetypmi budú vážené archetypy. Zvyšné kroky jednotlivých postupov ostávajú rovnaké. Myšlienka postupov sa nemení, avšak popis kroku, v ktorom sa hľadajú „najarchetypálnejšie“ objekty, by sme preformulovali na hľadanie „najarchetypálnejších relevantných“ objektov, nakoľko sa zahrnutím váh znižuje vplyv menej dôležitých pozorovaní a metóda sa snaží zachovať najmä informáciu, súvisiacu s tými dôležitými (čoho dôsledkom je, ako je ukázané v kap. 3, to, že vážené archetypy sa hľadajú v rámci množiny, ktorá je podmnožinou konvexného obalu pôvodných dát).

Takto modifikované postupy sme podobným spôsobom ako v predchádzajúcich častiach kapitoly aplikovali na naše dáta, popisujúce výsledky volieb do NR SR. Výsledky možno vidieť a porovnať v Tab. 13 a 14. Pokiaľ tieto tabuľky porovnáme s Tab. 10 a 12, môžeme si všimnúť, že pri využití vážených archetypov má polarizácia tendenciu dosahovať o niečo vyššie hodnoty (pri porovnaní príslušných dvojíc hodnôt sa tak stalo v 14 z 19 prípadov). Na základe našich výsledkov nemôžeme s určitosťou tvrdiť, že sa jedná o pravidlo, ale takýto jav je v súlade s teóriou z kap. 4, kde sa ukázalo, že hľadanie vážených archetypov je ekvivalentné hľadaniu štandardných archetypov na transformovaných dátach ($\tilde{X} = WX$). Ak potom vezmeme diagonálnu maticu váh W so zložkami menšími

Spôsob zahrnutia váh	2016	2020	2023
Nezahrnuté	0,92	0,83	0,85
Zahrnuté pri výpočte R^2	0,89	0,95	0,90
Zahrnuté pri výpočte R^2 aj pri hľadaní zhlukovania	0,93	0,98	0,95

Tabuľka 13: Porovnanie polarizácie, meranej pomocou vážených archetypov, pre roky 2016, 2020 a 2023 pri rôznych spôsoboch zahrnutia váh (dáta - kraje)

Spôsob zahrnutia váh	2016	2020	2023
Nezahrnuté	0,87	0,77	0,72
Zahrnuté pri výpočte R^2	0,92	0,95	0,74
Zahrnuté pri výpočte R^2 aj pri hľadaní zhlukovania	0,93	0,95	0,95

Tabuľka 14: Porovnanie polarizácie, meranej pomocou vážených archetypov, pre roky 2016, 2020 a 2023 pri rôznych spôsoboch zahrnutia váh (dáta - okresy)

alebo rovnými ako 1, budú sa hľadať objekty na množine, ktorá bude oproti konvexnému obalu pôvodných dát zmrštená, teda bude ich podmnožinou. Je potom prirodzené, že počet pozorovaní projektovaných na hranicu konvexného obalu množiny vážených archetypov bude spravidla vyšší, ako tomu bolo u štandardných archetypov. Aproximácie týchto pozorovaní sa potom ale budú nachádzať v pozíciách, ktoré budú v kontexte nájdených archetypov vzdialenejšie od stredu a bližšie k pozíciám, ktoré považujeme za extrémne. Keďže nami navrhnutý spôsob merania polarizácie posudzuje „mieru zhlučenia“ aproximácií pozorovaní okolo „najarchetypálnejších“ objektov, neprekvapí nás, že v takejto situácii dostávame vyššie hodnoty.

4.4 Zhrnutie

V tejto kapitole sme predstavili 3 postupy na meranie polarizácie v dátach. Najzákladnejší postup 4 je najjednoduchší, spĺňa $0 \leq R^2 \leq 1$, vďaka čomu sú jeho výstupy dobre interpretovateľné, a je vhodné ho použiť v prípadoch, keď nemáme k dispozícii váhy jednotlivých pozorovaní, prípadne ich pri meraní polarizácie nepovažujeme za dôležité.

Postup 4.2.1 nahrádza štandardné sumy štvorcov ich váženými verziami. Nakoľko pri našej implementácii využíva takmer identické funkcie ako postup 4, podobná je aj jeho

efektivita. Nakoľko sa však pri hľadaní zhlukovania využíva štandardné, neväžené K-means a centroidy datasetu a jednotlivých zhlukov nie sú nahradené ich váženými verziami, prechodom k tomuto postupu prichádzame o niektoré želané vlastnosti. Problematická je najmä neplatnosť ohraničenia $0 \leq R_w^2 \leq 1$, keďže R_W^2 môže dosahovať aj záporné hodnoty. Toto spolu s faktom, že spôsob hľadania zhlukovania nie je konzistentný s použitými sumami štvorcov, má za následok horšiu interpretovateľnosť výstupov.

Postup 4.2.2 takisto umožňuje zahrnutie váh a odstraňuje hlavné nedostatky postupu 4.2.1. Na hľadanie zhlukovania je využité vážené K-means a upravené sú aj využité sumy štvorcov, vďaka čomu opäť platí $0 \leq R_w^2 \leq 1$ a výstupy sú tak porovnateľne interpretovateľné ako pri postupe 4. Pri našej implementácii je ale tento postup o niečo pomalší.

Všetky tri postupy sa potom dajú uplatniť aj v situáciách, keď je výhodné štandardné archetypy nahradiť váženými.

Nami navrhnuté spôsoby merania polarizácie sme demonštrovali na umelých dátach a tiež na dátach popisujúcich výsledky volieb do NR SR v rokoch 2016, 2020 a 2023. Veríme ale, že tieto postupy sú aplikovateľné aj v iných situáciách (najmä v spoločenských vedách). Predstavme si napríklad situáciu, kde by nám boli expertmi v príslušnej oblasti poskytnuté (dvojrozmerné) dáta, popisujúce umiestnenie politikov (poslancov NR SR, členov kongresu USA...) v rámci štandardného politického spektra (1. os: pravica - ľavica, 2. os: liberalizmus - konzervativizmus). Pomocou postupov 4, 4.2.1 alebo 4.2.2 by sa dala merať tendencia politikov hlásiť sa skôr k extrémnym alebo naopak umierneným pozíciám, čo by mohlo byť cennou informáciou pre výskumníkov v danej oblasti. Nakoľko by sa podobných príkladov dalo vymyslieť viac, sme presvedčení, že výstupy tejto kapitoly by sa pre mnohých ľudí mohli stať užitočnými nástrojmi.

Záver

Hlavným cieľom tejto práce bolo predstaviť archetypálnu analýzu ako nový prístup k analýze dát a uviesť konkrétne príklady jej aplikácií. Neskôr k našim cieľom pribudlo nájsť spôsob merania polarizácie v dátach, využívajúci metódy AA.

V kap. 1 sme najprv priblížili situáciu v ktorej by bolo vhodné použiť AA, a popísali sme motiváciu, ktorá viedla k jej vzniku. Následne sme uviedli najdôležitejšie zo známych poznatkov a postupov AA. V kap. 1.1 sme predstavili základnú optimalizačnú úlohu ((2), (3), (4), (5), (6) a (7)) ako aj maticový zápis modelu ((8), (9), (10), (11) a (12)) a tiež sme predstavili najpopulárnejší spôsob výberu vhodného počtu archetypov (Obr. 3). Následne sme v kap. 1.2 uviedli vetu 1.1, popisujúcu polohu archetypov v rámci datasetu. V kap. 1.3 sme potom popísali základy archetypálneho algoritmu a striedavej optimalizácie. Pri písaní tejto kapitoly sme vychádzali najmä z [2], pričom v časti o striedavej optimalizácii sme niektoré kroky popisovali o niečo detailnejšie, ako boli uvedené v článku.

V kap. 2 sme uskutočnili vzorové analýzy dát ([13], [14], [15]) popisujúcich výsledky volieb do NR SR z hľadiska geografického rozloženia podpory strán v rámci Slovenska. Prvá analýza (kap. 2.1) využívala dáta, kde ako premenné vystupovali relatívne (oproti celkovým preferenciám danej strany) podpory strán v rámci jednotlivých krajov Slovenska. Jednalo sa o modelovú analýzu, v ktorej sme z dôvodu jednoduchosti a možnosti využitia zobrazovacej techniky ternary plot zvolili nižší (pre reálnu analýzu nie celkom vhodný) počet archetypov. Následne sme v kap. 2.2 uskutočnili podobnú analýzu, avšak namiesto krajov nás zaujímali relatívne podpory jednotlivých strán v rámci okresov, teda sme pracovali s dátami, v ktorých je rozmer pozorovaní vyšší ako ich počet. Porovnanie ternary plotov pre roky 2016, 2020 a 2023 (Obr. 5, 6 a 7) nás priviedlo k otázke merania polarizácie v dátach, ktorej sme sa bližšie venovali v kap. 4.

V kap. 3 sme predstavili vážené archetypy ako modifikovanú verziu AA, popísanú v [5]. Jej význam je najmä v tom, že znižuje citlivosť pôvodnej metódy na outliery. Tá bola problémom, na ktorý sme narazili už v kap. 2. Vážené archetypy sme následne aplikovali na rovnaké dáta ako v kap. 2. Pri tejto aplikácii sa ukázalo, že skutočne do istej miery znižujú citlivosť na outliery, hoci táto je stále vážnym nedostatkom.

V kap. 4 sme hľadali vhodný spôsob merania polarizácie v dátach, pričom okrem AA bolo našim nástrojom aj ohraničené K-means. Detailne sme popísali jednotlivé kroky, kto-

rými sme sa uberali, ktorých výsledkom je postup 4. Tento sa, ako sa ukázalo, zhodoval s intuíciou z kap. 2.1. Následne sme s jeho pomocou (na vhodnú voľbu počtu archetypov využívajúc koeficient determinácie pre model AA (34) a funkcie užitočnosti) porovnali z hľadiska polarizácie dáta z rokov 2016, 2020 a 2023. Tento spôsob merania sme potom v kap. 4.2 ďalej vylepšovali a zovšeobecňovali pridávaním možnosti zahrnúť váhy a tak presnejšie popísať dynamiku dát. Váhy vstupovali jednak do výpočtu koeficientu determinácie pre zhlukovanie (postup 4.2.1), využívaného pri výpočte polarizácie, ale takisto aj do hľadania zhlukovania samotného prostredníctvom nahradenia ohraničeného K-means váženým ohraničeným K-means (postup 4.2.2). Všetky tri postupy sme testovali na umelých aj reálnych dátach a výstupy tohto testovania ako aj ich teoretické vlastnosti sme zhrnuli v kap. 4.4.

Za hlavný prínos tejto práce považujeme prehľadné spracovanie problematiky AA, o ktorej, pokiaľ vieme, až dosiaľ neexistovali zdroje v Slovenskom jazyku, a uvedenie príkladu jej vhodnej aplikácie. Okrem toho sme ju využili ako nástroj pri hľadaní spôsobu merania polarizácie v dátach a jeho ďalších úpravách (zahrnutie váh), ktorý by mohol byť užitočným pri výskume najmä v spoločenskovedných oblastiach (politológia, sociológia...). Okrem toho na https://github.com/vincze16/name/tree/main/AA poskytujeme kódy všetkých nami naprogramovaných funkcií, ktoré sme pri písaní tejto práce používali.

Nakoľko je AA pomerne mladou oblasťou analýzy dát, priestor na ďalší výskum je stále široký. Okrem toho vnímame priestor aj na výrazne častejšie aplikovanie tejto metódy, ktorá je dodnes relatívne málo známa, hoci, ako sme sa snažili ukázať aj my v tejto práci, disponuje mnohými výhodami, ktoré ju robia cenným nástrojom v arzenáli človeka zaoberajúceho sa analýzou dát.

Zoznam použitej literatúry

[1] Bauckhage Ch., Manshaei K.: *Kernel Archetypal Analysis for Clustering Web Search Frequency Time Series*, 22nd International Conference on Pattern Recognition (2014), 1544–1549

[2] Breinman L., Cutler A.: *Archetypal Analysis*, Technometrics 36 (1994), 338–347

[3] Cardie C., Rogers S., Schrödl S., Wagstaff K.: *Constrained K-means Clustering with Background Knowledge*, Proceedings of the Eighteenth International Conference on Machine Learning (2001), 577–584

[4] Eugster M. J. A, Leisch F.: *From Spider-Man to Hero - Archetypal Analysis in R*, Journal of Statistical Software 30 (2009), 1–23

[5] Eugster M. J. A., Leisch F.: *Weighted and Robust Archetypal Analysis*, Technical Report N. 082 (2010), Department of Statistics, University of Munich

[6] Everitt B. S., Hothorn T (2022): *HSAUR: A Handbook of Statistical Analyses Using R (1st Edition), R package version 1.3-10*, dostupné na internete (26.12.2023): https://CRAN.R-project.org/package=HSAUR

[7] Hansen L. K., Mørup M.: *Archetypal analysis for machine learning and data mining*, Neurocomputing 80 (2012), 54–63

[8] Hiep T. K., Duc N. M. (2016): *conclust: Pairwise Constraints Clustering, R package version 1.1*, dostupné na internete (26.12.2023): https://CRAN.R-project.org/package=conclust

[9] Ikotun A. M., Ezugwu A. E., Abualigah L., Abuahija B., Heming J.: *K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data*, Information Sciences 622 (2023), 178–210

[10] Meyer D, Zeileis A, Hornik K (2023): *vcd: Visualizing Categorical Data, R package version 1.4-11*, dostupné na internete (26.12.2023): https://CRAN.R-project.org/package=vcd

[11] R Core Team (2023): *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Viedeň, Rakúsko, dostupné na internete (26.12.2023): www.R-project.org

[12] Smith (2017): *Ternary: An R Package for Creating Ternary Plots, R package version 2.3.0*, dostupné na internete (2.2.2024): https://cran.r-project.org/web/packages/Ternary/vignettes/Ternary.html

[13] Štatistický úrad SR: *Voľby do Národnej rady Slovenskej republiky 2016 - Súbory na stiahnutie*, dostupné na internete (4.1.2024): https://volby.statistics.sk/nrsr/nrsr2016/sk/download.html

[14] Štatistický úrad SR: *Voľby do Národnej rady Slovenskej republiky 2020 - Súbory na stiahnutie*, dostupné na internete (4.1.2024): https://volby.statistics.sk/nrsr/nrsr2020/sk/download.html

[15] Štatistický úrad SR: *Voľby do Národnej rady Slovenskej republiky 2023 - Súbory na stiahnutie*, dostupné na internete (4.1.2024): https://volby.statistics.sk/nrsr/nrsr2023/sk/subory_na_stiahnutie.html

[16] Using Weighted K-Means Clustering to Determine Distribution Centres Locations, dostupné na internete (29.1.2024): https://towardsdatascience.com/using-weighted-k-means-clustering-to-determine-distribution-centres-locations-2567646fc31d

[17] Wickham H (2022): *stringr: Simple, Consistent Wrappers for Common String Operations, R package version 1.5.0*, dostupné na internete (26.12.2023): https://CRAN.R-project.org/package=stringr

[18] Zhang W., Wangwu J., Lin Z.: *Statistical Modeling in Biomedical Research*, Springer, 2020