UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Struktura kriminálních činů pohledem logpodílové metodiky

Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **prof. RNDr. Karel Hron, Ph.D.**
Vypracoval: **Bc. David Vranešic**
Studijní program: N0541A170026 Aplikovaná matematika
Studijní obor: Aplikovaná matematika
Forma studia: prezenční
Rok odevzdání: 2024

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. David Vranešic

Název práce: Struktura kriminálních činů pohledem logpodílové metodiky

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2024

Abstrakt: Data týkající se struktury kriminálních činů jsou typickým příkladem dat nesoucích relativní informaci, tzv. kompozičních dat, pro jejichž statistickou analýzu je vhodné jejich vyjádření v logpodílových souřadnicích. Cílem diplomové práce bude zpracovat konkrétní datový soubor tohoto typu užitím pokročilých metod mnohorozměrné statistiky ve vhodných souřadnicových reprezentacích, které povedou k jednoduché interpretaci získaných výsledků.

Klíčová slova: Kompoziční data, metoda hlavních komponent, korespondenční analýza, trojfaktorová analýza, PARAFAC, kompoziční tabulky, trojrozměrná korespondenční analýza.

Počet stran: 64

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. David Vranešic

Title: Crime structure analysis using the logratio methodology

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

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

The year of presentation: 2024

Abstract: The data concerning the structure of criminal acts are a typical example of data carrying relative information, so-called compositional data, for the statistical analysis of which it is appropriate to express them in log-ratio coordinates. The aim of the thesis will be to process an empirical dataset by using advanced methods of multivariate statistics in appropriate coordinate representations that will lead to a simple interpretation of the obtained results.

Key words: Compositional data, principal component analysis, correspondence analysis, threeway analysis, PARAFAC, compositional tables, threeway correspondence analysis.

Number of pages: 64

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně pod vedením pana prof. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Jako prvnímu bych rád poděkoval vedoucímu mé diplomové práce panu prof. RNDr. Karlu Hronovi, Ph.D., a to nejen za jeho odborné rady v rámci tvorby této práce, ale zejména za velmi lidský a přátelský přístup, díky kterému jsem věřil, že práci dotáhneme včas do zdárného konce. Další velké poděkování patří mé přítelkyni, která mi byla a je nejen při studiu oporou. Na místě je i poděkovat spolužákům, díky kterým bylo studium skrze vzájemnou pomoc výrazně příjemnější.

Úvod

V mé diplomové práci s názvem Struktura kriminálních činů pohledem logpodílové metodiky budu pracovat s datovým souborem o počtech kriminálních činů v České republice za roky 2016 až 2022, který poskytuje Policie ČR. Tento typ dat jsem si vybral z důvodu mého zalíbení v True Crime televizních pořadech a podcastech. V této práci budeme s těmito daty pracovat jako s kompozičními, tj. budeme pracovat s relativní informací v nich ukrytou, čili je bude nutné vyjádřit ve vhodných logpodílových souřadnicích. Cílem diplomové práce bude tento datový soubor vyjádřený ve vhodných souřadnicích analyzovat užitím metod mnohorozměrné statistiky a dosáhnout tím jednoduché interpretace výsledků. Jednotlivé metody budou v rámci kapitol stručně uvedeny a aplikovány na data v softwaru `R`.

V první kapitole si popíšeme základní poznatky potřebné pro čtení práce. Dozvíme se zde o tom, co jsou to vícerozměrná data, kompoziční data a představíme si podrobněji datovou sadu, se kterou budeme pracovat. Jako první metodu aplikujeme na data za rok 2022 metodu hlavních komponent v rámci druhé kapitoly. Budeme pracovat pouze s pozorováními ve formě krajů ČR a proměnnými, kterými budou jednotlivé kategorie trestných činů. V závěru kapitoly se pokusíme dosáhnout obdobného výsledku jako v případě metody hlavních komponent pomocí korespondenční analýzy.

Ve třetí kapitole nám přibude faktor času a budeme data analyzovat pomocí metody PARAFAC za roky 2016-2022. V kapitole čtvrté budeme opět analyzovat rok 2022, pouze si přidáme faktor objasněnosti. Představíme si práci s kompozičními tabulkami a jejich rozklad na interakční a nezávislou část. V poslední páté kapitole si představíme a aplikujeme trojrozměrnou korespondenční analýzu na data za rok 2022 rozdělená dle faktoru objasněnosti.

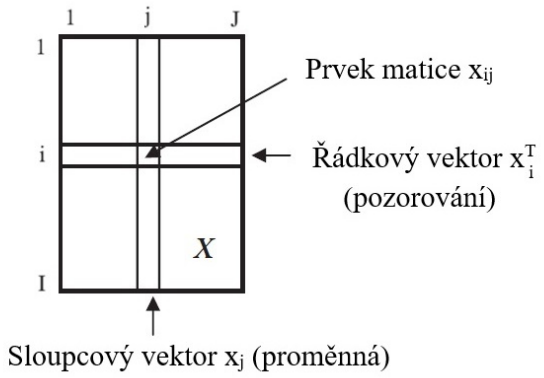
1. Základní poznatky

V této kapitole si představíme základní pojmy, které pro nás budou v diplomové práci užitečné. Řekneme si, co to vlastně kompoziční data jsou a představíme si datový soubor, s nímž budeme v rámci této práce laborovat. V této kapitole budeme vycházet ze zdrojů [3], [5], [8] a [9].

1.1. Vícerozměrná data

Vícerozměrná data vznikají, když při průzkumu zaznamenáváme hodnoty několika náhodných veličin (statistických znaků) na řadě subjektů, což vede k vektorovému pozorování u každého z nich. Můžeme je zapsat buď do podoby datové matice, nebo datového kvádru.

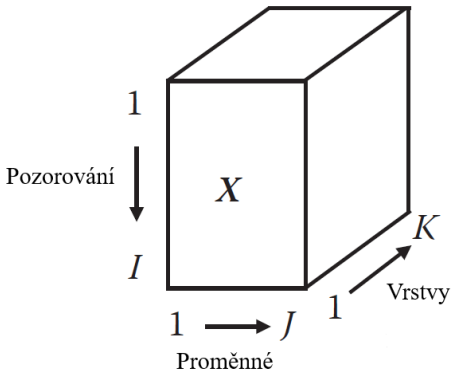
Datová matice je vhodná v případě, kdy máme pouze dva faktory. Graficky je možné ji znázornit jako zde na obrázku 1.



Obrázek 1: Datová matice

Řádky $\mathbf{x}_i^T$, $i = 1, \ldots, I$, datové matice $\mathbf{X}$ představují pozorování. Proměnné $\mathbf{x}_j$, $j = 1, \ldots, J$, jsou zapisovány do sloupců.

Mohou nastat ale situace, kdy si nevystačíme pouze s dvěma faktory a bude potřeba přidat faktor třetí. Budeme mít například tu naši datovou matici, ale budeme ji mít pro několik roků, tudíž nám vzniknou vrstvy. To znamená, že k pozorováním a proměnným přibude faktor času. Graficky nám datový kvádr znázorní obrázek 2.



Obrázek 2: Datový kvádr.

1.2. Kompoziční data

Jsou situace, ve kterých nemusí být pohled na data z absolutního hlediska vše vypovídající. Dejme si ihned zkraje příklad, kdy tomu tak je. Značka automobilů ABC měla 2,5% podíl na celkových prodejích automobilů v ČR, zatímco prodeje značky RST se v témže roce pohybovaly na úrovni 40 %. Během dekády obě automobilky navýšily své prodeje o 2,5 %. Z absolutního hlediska to pro nás znamená přesně to, jak to zní – polepšily si o 2,5 %. Zajímavější pohled je ale ten, že ABC si dvojnásobně polepšila, zatímco pro RST to nebyl nikterak markantní nárůst. Takovou informaci v datech nazýváme relativní, když je relevantní informace mezi proměnnými obsažena v jejich podílech. Vícerozměrná data, která nesou relativní informaci, označujeme

jako kompoziční. Průkopníkem teorie kompozičních dat byl John Aitchison, který tento typ dat nejen charakterizoval, ale také navrhl možnost užití log-podílových transformací k jejich statistické analýze. V této podkapitole budu vycházet ze zdrojů [3] a [5].

Kompoziční data popisují tedy nějakou část celku, přičemž se s nimi můžeme nejčastěji setkat jako s vektory proporcí či procent. Reprezentace kompozic konstantním součtem κ vede při vektoru proporcí na $\kappa = 1$, v druhém případě na $\kappa = 100$. Jako první pojem je potřeba zmínit D-složkovou kompozici, pod kterou si představme vektor s kladnými reálnými složkami $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$. Zmíněnou relativní informaci obsahují právě podíly mezi těmito složkami a právě složky tohoto vektoru se nasčítají na součet κ. Množina všech D-složkových kompozic se nazývá výběrový prostor kompozičních dat S^D, neboli simplex

$$S^D = \{\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}} \in \mathbb{R}^{\mathbb{D}} \mid x_i > 0, i = 1, \ldots, D, \sum_{i=1}^{D} x_i = \kappa\}. \qquad (1)$$

Realita je však taková, že požadavek na součet všech složek kompozice na zvolenou konstantu κ není častokrát splněn. Kladný D-složkový vektor $\mathbf{x}$ lze převést na daný součet κ pomocí operace uzávěr, přičemž matematický zápis uzávěru kompozice $C(\mathbf{x})$ je následovný

$$C(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right)^{\mathrm{T}}. \qquad (2)$$

Nicméně informace, která nás u kompozic zajímá je obsažená v podílech mezi složkami. Tudíž při zpracování dat pro nás nehraje konstanta κ roli.

Jak dále pracovat s kompozičními daty? Euklidovská geometrie není vyhovující, jelikož nedokáže pracovat s relativní informací v nich obsaženou. Pro práci s kompozicemi je tedy potřeba mít jinou geometrii, která je pro to

vhodná. Budeme pracovat na simplexu, na kterém byla definovaná geometrie pojmenována po Johnu Aitchisonovi - Aitchisonova geometrie.

Při využití operace uzávěru C můžeme pro $\mathbf{x}$, $\mathbf{y} \in S^D$ a $\alpha \in \mathbb{R}$ zavést postupně základní operace jakými jsou perturbace

$$\mathbf{x} \oplus \mathbf{y} = C\left(x_1 y_1, \ldots, x_D y_D\right) \in S^D \tag{3}$$

a mocninná transformace

$$\alpha \odot \mathbf{x} = C\left(x_1^{\alpha}, \ldots, x_D^{\alpha}\right) \in S^D. \tag{4}$$

Operace $(\oplus, \odot)$ nám na simplexu S^D společně určují vektorový prostor. Pro definování euklidovského lineárního vektorového prostoru je potřeba dodefinovat pro kompozice $\mathbf{x}, \mathbf{y} \in S^D$ následující tři operace.
Aitchisonův skalární součin $\langle \mathbf{x}, \mathbf{y} \rangle_{\alpha}$

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\alpha} = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}. \tag{5}$$

Aitchisonovu normu $\|\mathbf{x}\|_{\alpha}$

$$\|\mathbf{x}\|_{\alpha} = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=i}^{D} \left(\ln \frac{x_i}{x_j}\right)^2}. \tag{6}$$

Aitchisonovu vzdálenost $d_{\alpha}(\mathbf{x}, \mathbf{y})$

$$d_{\alpha}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\right)^2}. \tag{7}$$

Pro aplikaci standardních metod určených pro analýzu mnohorozměrných

dat je potřeba vyjádřit kompoziční data ve vhodných reálných souřadnicích. Souřadnicové reprezentace, které nám umožní tento převod jsou například aditivní log-podílové souřadnice (alr), centrované log-podílové koeficienty (clr), nebo izometrické log-podílové souřadnice (ilr). V této práci budeme pro analýzu používat pouze clr koeficienty.

Pomocí clr koeficientů tedy dokážeme D-složkovou kompozici $\mathbf{x}$ vyjádřit jako D-složkový reálný vektor $\mathbf{y}$ následovně,

$$\mathbf{y} = clr\left(\mathbf{x}\right) = \left(y_1, \ldots, y_D\right)^{\mathrm{T}} = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{k=1}^{D} x_k}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{k=1}^{D} x_k}} \right)^{\mathrm{T}}, \quad (8)$$

přičemž jmenovatel $\sqrt[D]{\prod_{k=1}^{D} x_k}$ představuje geometrický průměr složek kompozice.

Pro matici kompozičních dat $\mathbf{X}$ o rozměrech $n \times D$ můžeme poskládat matici clr koeficientů $\mathbf{Y}$ tímto způsobem

$$\mathbf{y}_i^{\mathrm{T}} = \left(clr\left(x_i\right)\right)^{\mathrm{T}} = \left(\ln \frac{x_{i1}}{\sqrt[D]{\prod_{k=1}^{D} x_{ik}}}, \ldots, \ln \frac{x_{iD}}{\sqrt[D]{\prod_{k=1}^{D} x_{ik}}} \right), \quad (9)$$

kde $\mathbf{x}_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{iD})$ jsou řádky matice $\mathbf{X}$ pro $i = 1, \ldots, n$. Zde vidíme, že geometrický průměr je počítán zvlášť pro každé pozorování. Větší detail a popis dalších transformací včetně vztahů mezi nimi lze dohledat v knize [3] v kapitole 3.3.

Pro interpretaci mají log-podíly důležitou a pěknou vlastnost, jelikož nám symetrizují role čitatele a jmenovatele ve zlomku kolem nuly. Vezmeme-li podíl dvou reálných čísel $\frac{a}{b}$, kde $a > b$, tak čitatel dominuje jmenovatele a budeme mít výsledek větší než jedna. Naopak, když $a < b$, tj. jmenovatel bude dominovat čitatele, tak dostanu číslo mezi nulou a jedničkou. Což není symet-

rické. Vezmeme-li log-podíl $\log \frac{a}{b}$, tak dominanci jmenovatele bude odpovídat číslo z intervalu $(-\infty, 0)$ a dominanci čitatele číslo z intervalu $(0, \infty)$. Rovnováze bude odpovídat nula. Log-podíly navíc převádí multiplikativní vztahy na aditivní, což je velmi dobrá vlastnost.

Jednou z důležitých vlastností kompozičních dat je podkompoziční koherence (soudržnost), která nám zjednodušeně říká to, že informace získaná z analýzy d-složkové kompozice, kde $d < D$, nemá být rozdílná oproti informaci získané z analýzy provedené na původní D-složkové kompozici. Prostřednictvím Aitchisonovy geometrie máme splnění této vlastnosti pro analýzu kompozičních dat zajištěnu.

1.3. Data o kriminálních činech

V této diplomové práci budu pracovat s datovou sadou týkající se struktury kriminálních činů v České republice dle krajů za období mezi lety 2016 až 2022. Data jsou veřejně dostupná na stránkách Policie ČR [8], kde jsou každý měsíc aktualizována. Jednotlivé přestupky proti zákonu jsou v původní datové sadě rozděleny do sedmi sumarizačních kategorií – kriminalita násilná, mravnostní, majetková, ostatní, zbývající, hospodářská a vojenské a protiústavní činy. Co pod sebou skrývá násilná, či majetková kriminalita může být leckomu jasné, nicméně co spadá do zbývající nebo ostatní kriminality už tak průzračné být nemusí. V následující tabulce 1 vidíme u každé kategorie pro představu několik příkladů.

Kriminalita	Obsah
Násilná	Vraždy, týrání, vydírání,...
Mravnostní	Znásilnění, dětská pornografie, kuplířství,...
Majetková	Krádeže, zpronevěra, poškození cizí věci,...
Ostatní	Výtržnosti, sprejerství, podávání alkoholu dítěti,...
Zbývající	Pomluva, lichva, nadržování,...
Hospodářská	Zkrácení daně, pojistné podvody,...
Vojenská	Vlastizrada, teroristický útok, financování terorismu,...

Tabulka 1: Příklady jednotlivých trestných činů v daných kategoriích.

Já jsem ve své analýze skrze velmi nízké jednotky vojenských a protiústavních činů udělal jejich sumarizaci s ostatní kriminalitou, přičemž toto spojení pro nás bude výhodné při analýze datového souboru z hlediska počtu proměnných, kterých bude šest, případně dvanáct. Oproti tomu roli pozorování bude hrát čtrnáct krajů České republiky.

Pro další postup a jednoznačné porozumění datům je na místě uvést krátký přehled zkratek v následující tabulce 2, který se bude prolínat celou prací.

Zkratka	Význam	Zkratka	Význam
PRA	Praha	Nas	Násilná kriminalita
STR	Středočeský kraj	Mrav	Mravnostní kriminalita
JC	Jihočeský kraj	Maj	Majetková kriminalita
PLZ	Plzeňský kraj	Ost	Ostatní kriminalita
UST	Ústecký kraj	Zb	Zbývající kriminalita
KH	Královehradecký kraj	Hosp	Hospodářská kriminalita
JM	Jihomoravský kraj		
MSL	Moravskoslezský kraj		
OL	Olomoucký kraj		
ZL	Zlínský kraj		
VYS	kraj Vysočina		
PAR	Pardubický kraj		
LIB	Liberecký kraj		
KAR	Karlovarský kraj		

Tabulka 2: Přehled základních zkratek užívaných v diplomové práci.

Přehled o celkovém počtu registrovaných spáchaných kriminálních činů v roce 2022 nabízí tabulka 3. Počty trestných činů jsem v jednotlivých krajích normoval na 100 tisíc obyvatel. Nicméně v logpodílové metodice toto normování nehraje žádnou roli, jelikož kompozice jsou invariantní na změnu měřítka.

	Nas	Mrav	Maj	Ost	Zb	Hosp
PRA	137	34	2139	383	199	234
STR	95	22	759	234	213	74
JC	140	33	608	249	235	145
PLZ	119	27	950	260	249	149
UST	188	52	991	355	323	165
KH	106	31	555	228	211	105
JM	108	32	903	223	185	102
MSL	151	32	1081	287	225	105
OL	134	27	657	241	265	146
ZL	105	25	443	173	215	98
VYS	103	26	474	192	177	83
PAR	68	27	469	180	212	67
LIB	154	39	1007	305	266	158
KAR	149	32	841	285	249	203

Tabulka 3: Počty registrovaných spáchaných trestných činů dle jejich klasifikace v roce 2022.

Tabulku 3 využijeme zejména v kapitole 2 v metodě hlavních komponent. Stejný typ tabulky, pouze za roky 2016 až 2022, použijeme v kapitole 3 v metodě PARAFAC. Pro další kapitoly bude ale potřeba přidat další faktor a tím je objasněnost. To znamená, že v tabulce 3 rozdělíme proměnné na objasněno a neobjasněno, přičemž objasněno znamená, kolik bylo z daných kriminálních činů registrovaných v daném roce zároveň i vyřešeno. Objasněné počty nalezneme v tabulce 4, neobjasněné v tabulce 5.

V této podkapitole jsme mohli vidět pouze rok 2022. Tabulkové přehledy za roky 2016 až 2021 jsou k dispozici k náhledu v příloze této práce.

	Nas	Mrav	Maj	Ost	Zb	Hosp
PRA	74	18	287	223	136	59
STR	48	11	157	177	159	20
JC	92	25	231	206	203	95
PLZ	63	13	236	192	181	76
UST	130	32	440	290	239	106
KH	70	21	201	181	175	57
JM	61	17	200	136	137	33
MSL	95	21	422	223	175	45
OL	94	15	213	184	221	66
ZL	78	17	165	141	184	60
VYS	83	16	145	157	152	37
PAR	45	21	175	148	174	31
LIB	97	25	307	240	192	75
KAR	96	17	298	215	177	146

Tabulka 4: Počty objasněných registrovaných trestných činů dle jejich klasifikace v roce 2022.

	Nas	Mrav	Maj	Ost	Zb	Hosp
PRA	63	16	1852	160	63	175
STR	47	11	602	57	54	54
JC	48	8	377	43	32	50
PLZ	56	14	714	68	68	73
UST	58	20	551	65	84	59
KH	36	10	354	47	36	48
JM	47	15	703	87	137	33
MSL	56	11	659	64	50	60
OL	40	12	444	57	44	80
ZL	27	8	278	32	31	38
VYS	20	10	329	35	25	46
PAR	23	6	294	32	38	36
LIB	57	14	700	65	74	83
KAR	53	15	543	70	72	57

Tabulka 5: Počty neobjasněných registrovaných trestných činů dle jejich klasifikace v roce 2022.

2. Metoda hlavních komponent

Metoda hlavních komponent je bezesporu jednou z nejpoužívanějších mnohorozměrných metod. V rámci této kapitoly si představíme koncept a způsob řešení, aplikujeme metodu na datovou sadu a provedeme srovnání s korespondenční analýzou, přičemž budeme vycházet ze zdrojů [2] a [9].

2.1. Koncept metody

Primárním cílem metody hlavních komponent (PCA) je redukce dimenze dat. Máme datovou matici $\mathbf{X}_{(I\times J)}$, tj. máme I pozorování a J proměnných. Tato metoda nám shrne informaci z těchto J dimenzí ideálně do dvou, aby bylo možné ji následně graficky zobrazit. Konstrukce nových proměnných proběhne tak, že ty nové proměnné budou lineární kombinací těch původních proměnných, přičemž již několik prvních nových proměnných, hlavních komponent, bude schopných vysvětlit většinu informace, která se v datech vyskytuje. Informací ve smyslu PCA rozumíme variabilitu. Máme tedy variabilitu v jednotlivých proměnných a tu když sečteme, tak dostaneme celkovou variabilitu toho datového souboru. Pomocí oné redukce dimenze budeme chtít ukrojit co nejvíce z té celkové variability, to znamená co nejvíce celkové variability v daném datovém souboru vysvětlit již pomocí prvních několika nových proměnných - hlavních komponent.

První hlavní komponenta PC_1 představuje lineární kombinaci s_1 původních proměnných $x_j, j = 1, \ldots, J$ takovou, že rozptyl této první komponenty bude maximální a to za podmínky, že neznámý vektor koeficientů, zátěžových vektorů, $\mathbf{z}_1 = (z_{11}, \ldots, z_{J1})^{\mathrm{T}}$ je normovaný. Máme tedy lineární kombinaci proměnných

$$s_1 = x_1 z_{11} + \cdots + x_J z_{J1} \tag{10}$$

s neznámým vektorem zátěží

$$\mathbf{z}_1 = (z_{11}, \ldots, z_{J1})^{\mathrm{T}}, \tag{11}$$

přičemž první hlavní komponentu PC_1 hledáme tak, aby nám napříč všemi hlavními komponentami vysvětlila největší množství informace, to znamená $\mathrm{var}(s_1) \to \max$ za podmínky $\mathbf{z}_1^T \mathbf{z}_1 = 1$, která nám říká, že jsou tyto neznámé koeficienty normované.

Pro druhou hlavní komponentu PC_2 nám přibude podmínka na ortogonalitu zátěžových vektorů, tj. $\mathbf{z}_1^{\mathrm{T}} \mathbf{z}_2 = 0$, tudíž PC_2 získáme opět pomocí maximalizace variability s_2, tj. $\mathrm{var}(s_2) \to \max$ za podmínek $\mathbf{z}_2^{\mathrm{T}} \mathbf{z}_2 = 1$ a $\mathbf{z}_1^{\mathrm{T}} \mathbf{z}_2 = 0$, kde s_2 je

$$s_2 = x_1 z_{12} + \cdots + x_J z_{J2} \tag{12}$$

a vektor neznámých koeficientů vypadá následovně

$$\mathbf{z}_2 = (z_{12}, \ldots, z_{J2})^{\mathrm{T}}. \tag{13}$$

Obecně tedy pro j-tou hlavní komponentu PC_j

$$\mathrm{var}(s_j) = \mathrm{var}(x_1 z_{1j} + \cdots + x_J z_{Jj}) = \mathbf{z}_j^{\mathrm{T}} \mathrm{var}(x_1, \ldots, x_J) \mathbf{z}_j = \mathbf{z}_j^{\mathrm{T}} \mathbf{\Sigma} \mathbf{z}_j, \tag{14}$$

pro $j = 1, \ldots, J$ za podmínky $\mathbf{Z}^{\mathrm{T}} \mathbf{Z} = \mathbf{I}$. Symbol $\mathbf{\Sigma}$ značí varianční matici.

Naši podmíněnou optimalizační úlohu lze přepsat pomocí Lagrangeových multiplikátorů $\lambda_j, j = 1, \ldots, J$, následujcím vztahem

$$\varphi_j = \mathbf{z}_j^{\mathrm{T}} \mathbf{\Sigma} \mathbf{z}_j - \lambda_j (\mathbf{z}_j^{\mathrm{T}} \mathbf{z}_j - 1), \tag{15}$$

jejíž řešení nalezneme standardně derivací daného výrazu dle $\mathbf{z}_j$ a derivovaný

výraz položíme roven nule.

To nám pro $j = 1, \ldots, J$ dává rovnost známou z úlohy hledání vlastních vektorů a vlastních čísel,

$$\mathbf{\Sigma}\mathbf{z}_j = \lambda_j \mathbf{z}_j. \tag{16}$$

Z výše uvedeného nám vyplývá, že za řešení neznámých parametrů $\mathbf{z}_j$ vezmeme vlastní vektory matice $\mathbf{\Sigma}$ a odpovídající vlastní čísla této matice budou odpovídat vlastním číslům λ_j. Pak dle (14) je rozptyl

$$\mathrm{var}(s_j) = \mathbf{z}_j^{\mathrm{T}} \mathbf{\Sigma} \mathbf{z}_j = \mathbf{z}_j^{\mathrm{T}} \lambda_j \mathbf{z}_j = \lambda_j \tag{17}$$

a jelikož vlastní vektory a jim odpovídající vlastní čísla jsou seřazena v sestupném pořadí, tak také rozptyly jednotlivých hlavních komponent klesají s vyšším pořadím komponenty. Tím pádem matice $\mathbf{Z}$ s koeficienty pro lineární kombinaci je maticí vlastních vektorů, nebo také maticí zátěží. Shrneme-li to, tak sloupce matice $\mathbf{S}$, jež se nazývá matice skórů, tak opovídají souřadnicím nových proměnných v prostoru hlavních komponent. Sloupce matice zátěží $\mathbf{Z}$ odpovídají vektorům vah původních proměnných, což jsou naše vektory zátěží.

2.2. Singulární rozklad matice

Přístupů k PCA je několik, zde si představíme velmi užitečný singulární rozklad matice (SVD - Singular value decomposition). Tento postup je výhodný zejména díky tomu, že se dá aplikovat i na situace, kdy máme více proměnných než pozorování. Principiálně jde o to, že dokážeme rozložit každou centrovanou matici $\mathbf{X}_{(I \times J)}$ na součin tří matic následujícím způsobem

$$\mathbf{X} = \mathbf{S}_0 \cdot \mathbf{D} \cdot \mathbf{Z}^{\mathrm{T}}, \tag{18}$$

kde $\mathbf{S}_0$ je rozměru $I \times J$, diagonální matice $\mathbf{D}$ je rozměru $J \times J$ a $\mathbf{Z}^{\mathrm{T}}$ je transponovaná čtvercová matice zátěží rozměru $J \times J$.

Matice $\mathbf{S_0}$ obsahuje skóry hlavních komponent, které jsou normované na délku 1, respektive mají jednotkový rozptyl. Provedeme-li součin této matice s maticí $\mathbf{D}$ obsahující na diagonále singulární hodnoty, jež jsou rovny $\sqrt{\lambda_j}$, tj. směrodatným odchylkám skórů, tak dostáváme matici skórů $\mathbf{S}$ ve smyslu, v jakém jsme si ji definovali výše,

$$\mathbf{S} = \mathbf{S}_0 \cdot \mathbf{D}. \tag{19}$$

Komplikace může nastat za situace, kdy bychom měli vysoce-dimenzionální data. V takovém případě na to nepůjdeme přes rozklad matice $\mathbf{X}$, ale matici $\mathbf{Z}$ získám z vlastních vektorů matice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ a matici $\mathbf{S}_0$ z vlastních vektorů matice $\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Obě matice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ a $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ mají stejná vlastní čísla. Matice, která bude většího rozměru, tak bude mít zbytek vlastních čísel roven nule. Potom na základě rozkladu zmíněného ve vztahu (18) se dá matice zátěží $\mathbf{Z}$ napočítat následovně

$$\mathbf{Z} = \mathbf{X}^{\mathrm{T}} \cdot \mathbf{S}_0 \cdot \mathbf{D}^{-1} = \mathbf{X}^{\mathrm{T}} \cdot \mathbf{S} \cdot \mathbf{D}^{-2}. \tag{20}$$

Singulární rozklad nám ale neumožňuje robustifikaci, takže se dá použít pouze pro klasickou verzi PCA. Nicméně v této práci se robustní verzí PCA zabývat nebudeme.

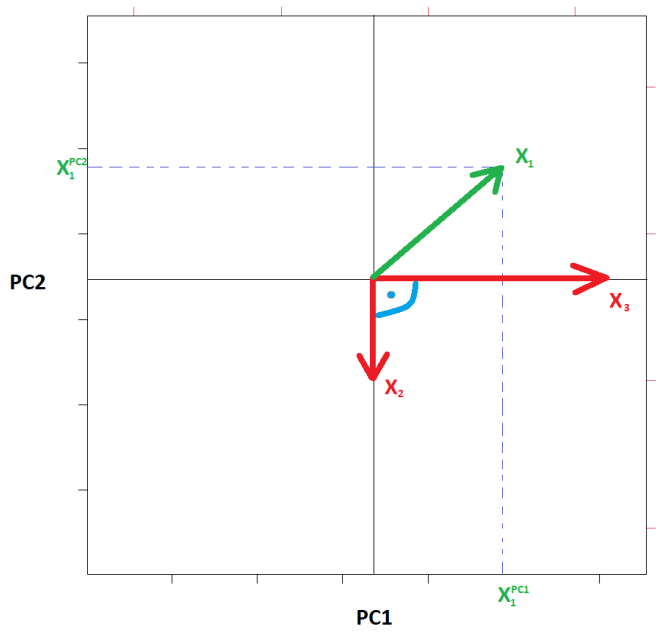
2.3. Počet komponent a biplot

Pro stanovení počtu hlavních komponent neexistují žádné testy operující s p-hodnotami, což bychom možná mohli očekávat od statistické metody. Platí zde spíše zásada, že vezmeme takový počet hlavních komponent, který

mi vysvětlí dostatečné množství variability. Dostatečné množství zní docela vágně, ale dle [3] pro dobrou představu o datové struktuře vícedimenzionálních dat ve dvou dimenzích stačí 70 % vysvětlené variability. Pokud budeme mít málo proměnných, jako v našem případě šest, tak nám úplně vystačí dvě komponenty, jinak by ta redukce dimenze neměla smysl. Mimo toto číselné doporučení existuje i grafický nástroj pro určení počtu komponent a tím je síťový graf (scree plot). Nicméně se stále jedná o subjektivní záležitost, jelikož doporučení zní tak, že zvolíme počet komponent podle toho, kde se nám v síťovém grafu při grafickém znázornění klesající variability hlavních komponent vyskytne 'loket'. Nalezení takového 'zlomu' v grafu ale také nemusí být tak jednoznačné a držíme se u dat s několika málo proměnnými spíše oné procentuální zásady.

Od začátku této kapitoly jsem několikrát zmínil, že chceme, aby již první hlavní komponenty popsaly co nejvíce variability, která se nám v datech vyskytuje. Ideálně chceme, aby nám PCA zredukovala informaci z mnohorozměrných dat do dvou dimenzí. Nyní se dostáváme k tomu, proč tomu tak je. Ve dvou dimenzích totiž dokážeme naše data názorně graficky zobrazit pomocí biplotu, což je vlastně graf dvou hlavních komponent.

Abychom co nejlépe vysvětlili tvorbu biplotu, tak to vezmeme pěkně košatě od počátku, kdy máme datovou matici $\mathbf{X}_{(I\times J)}$, matici skórů $\mathbf{S}_{(I\times J)}$ a matici zátěží $\mathbf{Z}_{(J\times J)}$. Řádky matice $\mathbf{S}$ představují jednotlivá pozorování vyjádřená v novém souřadnicovém systému. Řádky matice $\mathbf{Z}$ představují příspěvky jednotlivých původních proměnných ke komponentám - například prvek matice zátěží na pozici odpovídající prvnímu řádku a druhému sloupci říká, jak moc první původní proměnná přispívá ke konstrukci druhé hlavní komponenty. Nyní se podívejme na následující obrázek 3.



Obrázek 3: Ukázka biplotu.

Délka šipky biplotu odpovídá variabilitě dané proměnné, tj. $\text{var}(x_j)$, pro $j = 1, \ldots, J$. Úhel, který svírají jednotlivé šipky mezi sebou aproximuje korelaci mezi danými proměnnými $\text{cor}(x_l, x_j)$, kde $l, j = 1, \ldots, J, l \neq j$. V případě, že šipky mezi sebou svírají pravý úhel, ten v obrázku 3 svírají šipky pro x_2 a x_3, jsou dané proměnné nekorelované[1]. Dále vidíme, že šipka pro x_3 je rovnoběžná s $PC1$. To znamená, že první hlavní komponenta nese hodně informace o třetí proměnné. Pokud se nám tvoří shluky pozorování v biplotu, tak se podíváme ve směru jaké šipky tomu tak je, tedy jaká proměnná v těchto pozorováních dominuje.

Budeme aplikovat metodu hlavních komponent na data transformovaná do clr koeficientů. Tím se nám ale změní interpretace uvedená v předchozím odstavci, která platí pro klasický biplot. Délka šipky nyní aproximuje variabi-

[1]Nekorelovanost x_2 a x_3 znamená, že $\text{cor}(x_2, x_3) = 0$

litu proměnné v clr koeficientech, to znamená, že délka šipky $\approx \text{var}(\text{clr}(x_i))$. Délka spojnice vrcholů dvou šipek aproximuje varibilitu logpodílu dvou proměnných, tedy tato délka $\approx \text{var}(\ln \frac{x_i}{x_j})$. Poloha bodů ve směru šipek nám nyní bude hovořit o relativní dominanci dané proměnné v rámci toho pozorování, tj. jak moc je dané pozorování danou proměnnou ovlivněno, bereme-li do úvahy logpodíly této proměnné s ostatními proměnnými.

2.4. Aplikace na data

Data z tabulky 3, v `R` souboru pod názvem CRbyKN2022_CLK[2], převedeme do clr koeficientů pomocí funkce `cenLR` z knihovny `robCompositions` následujícím způsobem.

```
>setwd("Vaše umístění datového souboru")
>load("KRIMI_DATA.rda")
>library('robCompositions')
>CCRbyKN2022_CLK=cenLR(CRbyKN2022_CLK)$x.clr
```

Nyní můžeme provést PCA buď pomocí příkazu `princomp`, nebo `prcomp`. Funkce `prcomp` je založená na singulárním rozkladu matice, který jsme si vysvětlili v podkapitole 2.2, tudíž na data aplikujeme právě `prcomp`.

```
>PCA_clr_2022<-prcomp(CCRbyKN2022_CLK, center = T)
>summary(PCA_clr_2022)
```

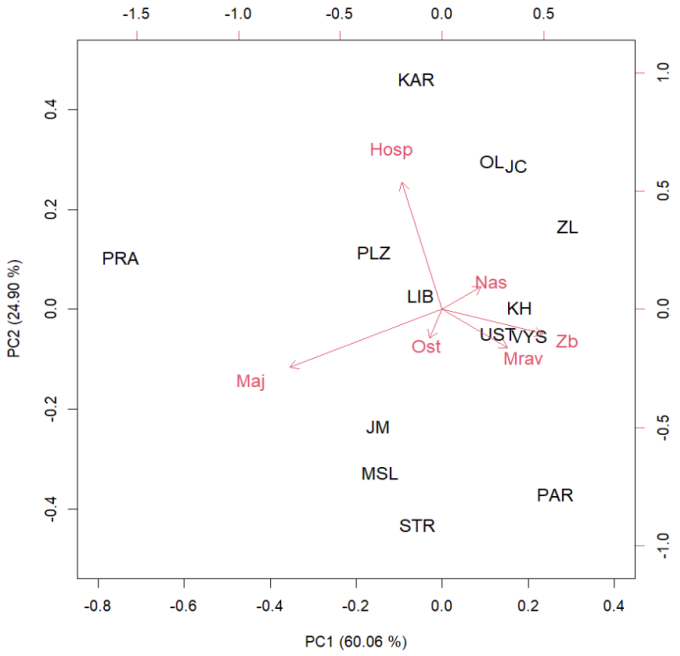
V software po příkazu `summary` lze vidět, že nám již prvních pět hlavních komponent vysvětlilo 100 % variability, což je důsledek konstrukce clr koeficientů, které mají díky nulovému součtu singulární varianční matici. Důležité pro nás je ale to, že již první dvě hlavní komponenty vysvětlily 84.96 % informace obsažené v datovém souboru, tudíž můžeme použít pro grafické

[2]CR (Česká republika), by (podle), K (krajů), N (normované), 2022 (rok), CLK (registrovaný počet trestných činů celkem)

znázornění biplot. Pro jeho konstrukci využijeme funkci `biplot`. V ní si nastavíme rozsah os x a y, přidáme popisky os a následně si v biplotu zvětšíme text v zobrazení.

```
>biplot(PCA_clr_2022, xlab="PC1(60.06 %)", ylab="PC2(24.90 %)",
xlim = c(-0.8, 0.4), ylim = c(-0.5, 0.5), cex=1.2)
```

V software `R` můžeme provést PCA pro ověření jak na normovaných datech na 100 tisíc obyvatel, tak na těch nenormovaných a opravdu vyjdou dva identické výsledky. Na biplot PCA provedené na tabulce 3 transformované do clr koeficientů se můžeme podívat níže na obrázku 4.



Obrázek 4: PCA clr biplot trestné činnosti v ČR v roce 2022.

Čím dále ve směru dané proměnné se nachází naše pozorování, tak tím více tato proměnná dominuje v daném kraji. Biplot PCA nám aproximuje strukturu datového souboru a z něj můžeme vyčíst například to, že Liberecký kraj je průměrný ve všech proměnných, jelikož v něm žádný trestný čin nedominuje. Hospodářské a majetkové prohřešky proti zákonu se ve velké míře objevují v Praze. Hospodářská kriminalita dominuje jednoznačně v Karlovarském kraji a násilná trestná činnost ve Zlínském kraji.

2.5. PCA vs. CA

Korespondenční analýza (CA) analyzuje vztahy dvou kategoriálních proměnných uspořádaných do kontigenční tabulky a snaží se obdobně jako metoda hlavních komponent o redukci dimenze. Cílem tedy je vysvětlení co největšího množství informace obsažené v původních proměnných v rámci již několika prvních nových proměnných, komponent. Dle článku [5] se lze na základě volby mocninné transformace dat a následné aplikace CA dostat k obdobným výsledkům jako v případě PCA v log-podílové metodice. V rámci této podkapitoly se pokusíme srovnat výsledky těchto dvou metod.

Máme datovou matici $\mathbf{X}_{(I\times J)} = (x_{ij})_{i=1,\dots,I,j=1,\dots,J}$, kde I odpovídá počtu kategorií u první proměnné, to budou v našem případě kraje ČR a J odpovídá počtu kategorií u druhé proměnné, to budou kriminální činy. Postupně definujme celkový součet

$$x_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}, \tag{21}$$

částečný řádkový součet

$$x_{i+} = \sum_{j=1}^{J} x_{ij} \tag{22}$$

a částečný sloupcový součet

$$x_{+j} = \sum_{i=1}^{I} x_{ij}. \tag{23}$$

Korespondenční matici $\mathbf{P}_{(I\times J)}$ získáme pomocí proporcionální reprezentace

$$\mathbf{P} = \frac{1}{x_{++}}\mathbf{X}. \tag{24}$$

Prvky r_i vektoru řádkových četností $\mathbf{r}$ vypočteme podílem

$$r_i = \frac{x_{i+}}{x_{++}} \tag{25}$$

a prvky c_j vektoru sloupcových četností $\mathbf{c}$ obdobně

$$c_j = \frac{x_{+j}}{x_{++}}. \tag{26}$$

Dále mějme k dispozici diagonální matice četností $\mathbf{D}_r$ a $\mathbf{D}_c$,

$$\mathbf{D}_r = \mathrm{diag}(r_1, r_2, \ldots, r_I), \tag{27}$$

$$\mathbf{D}_c = \mathrm{diag}(c_1, c_2, \ldots, c_J). \tag{28}$$

Nyní máme vše potřebné k výpočtu matice, na které budeme následně provádět singulární rozklad. Ta je uvedena v následujícím vztahu

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathbf{T}})\mathbf{D}_c^{-\frac{1}{2}}. \tag{29}$$

Dalším krokem korespondenční analýzy je výpočet řádkových a sloupcových souřadnic. Ty získáme právě aplikací singulárního rozkladu na matici $\mathbf{S}$.

Pro zisk obdobného výsledku jako v případě PCA je potřeba vzít data ve formě proporcí, tj. vezmeme si tabulku 3 vyjádřenou v proporcích a na tu aplikujeme mocninnou transformaci. Nejpopulárnější mocninou transformací je Box-Coxova mocninná transformace, jež je s mocninným paramatrem α definována následovně

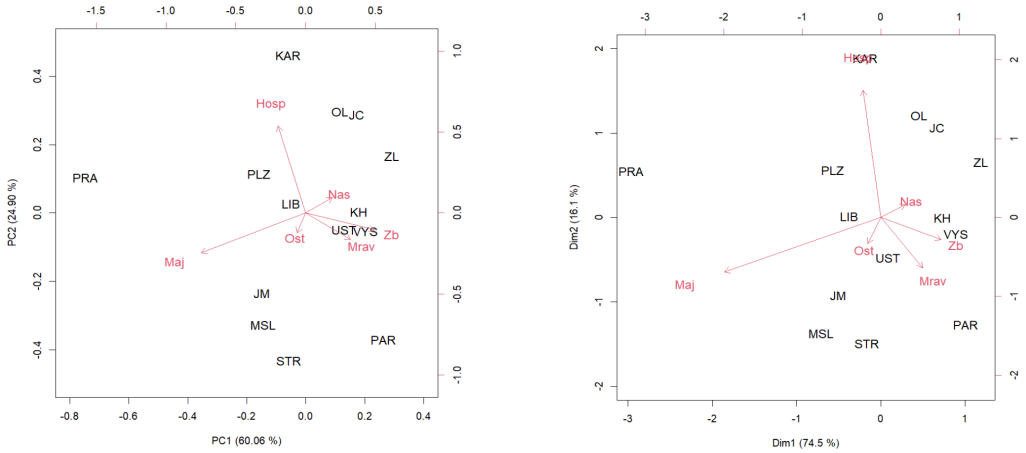
$$f(x) = \begin{cases} (\frac{1}{\alpha})(x^{\alpha} - 1), & \alpha > 0, \\ \ln(x), & \alpha = 0. \end{cases} \tag{30}$$

Přičemž log-transformace je limitním případem Box-Coxovy mocninné transformace v případě, že α se blíží k nule, tj. $f(x) \to \ln(x)$, když $\alpha \to 0$.

Data z tabulky 3 pomocí příkazu `prop.table` převedeme na tabulku proporcí, tu log-transformujeme pomocí funkce `log` a následně z balíčku `ca` na ni aplikujeme funkci `ca`. Poté vytvoříme biplot.

```
>library("ca")
>PCRbyKN2022_CLK<-prop.table(CRbyKN2022_CLK)
>LPCRbyKN2022_CLK<-log(PCRbyKN2022_CLK)
>CA_2022 <- ca(LPCRbyKN2022_CLK)
>CA_2022$colcoord <- CA_2022$colcoord*(-1)
>biplot(CA_2022$rowcoord, CA_2022$colcoord, xlab="Dim1(74.5 %)",
ylab="Dim2 (16.1 %)", ylim = c(-2, 2), cex=1.2)
```

Pro lepší srovnání v následujícím obrázku 5 máme jak biplot PCA, tak CA.



Obrázek 5: Vlevo je biplot získaný pomocí PCA, vpravo potom jeho protějšek jako výsledek CA pro data o relativní struktuře kriminality v roce 2022.

Při srovnání lze pozorovat, že výsledné biploty jsou až na drobné odchylky identické, čímž je argumentováno i v článku [5]. Při užití clr transformace odečítáme od každého prvku geometrický průměr všech prvků na řádku, čímž vlastně data vycentrujeme. Následně, aplikací PCA vycentrujeme sloupce tabulky, což znamená, že centrujeme dvojitě. Při CA odečítáme součin částečných řádkových a sloupcových četností, čímž data symetrizujeme kolem očekávaných hodnot. Navíc, před aplikací CA data log-transformujeme, což nás v podstatě přibližuje k podobné situaci jako v případě PCA, kde pracujeme s clr koeficienty. Tato teoretická souvislost mezi log-podílovou metodikou a korespondenční analýzou má několik praktických důsledků. Například log-podílová metodika má vlastnost podkompoziční koherence, zatímco korespondenční analýza ne. Tudíž pokud provedeme CA na data, která byla mocninně transformována, přičemž parametr $\alpha \to 0$, tak se CA blíží k tomu, aby měla vlastnost podkompoziční koherence.

3. Trojfaktorová analýza

V metodě hlavních komponent jsme uvažovali pouze dva faktory (pozorování vs. proměnné), nyní nám ale přibude faktor třetí. Můžeme tak říci, že trojfaktorová analýza je zobecněním PCA. Mezi její nejpopulárnější metody patří Tucker3 a PARAFAC. Jelikož PARAFAC je řekněme speciálním případem Tucker3, tak se prvně podíváme právě na Tucker3 a poté na PARAFAC, který aplikujeme na kriminální data za roky 2016 až 2022. V této kapitole budeme vycházet ze zdrojů [7] a [9].

Dosud jsme měli mód A, který představuje pozorování a mód B představující proměnné. To znamená, že jsme měli data uspořádaná do matice $\mathbf{X}$ o rozměrech $(I \times J)$, kde I je počet pozorování a J je počet proměnných. Prvky této matice jsou x_{ij}, kde $i \in \{1, \ldots, I\}$ a $j \in \{1, \ldots, J\}$. Nyní nám přibude třetí index $k \in \{1, \ldots, K\}$, mód C, představující vrstvy v datech. To si můžeme v našem případě představit tak, že pozorování jsou kraje, proměnné trestné činny a údaje za jednotlivé roky vnímáme jako vrstvy. Takto uspořádaná data charakterizujeme jako datový kvádr, který označíme $\underline{\mathbf{X}}$.

Datový kvádr $\underline{\mathbf{X}}$ si tedy představme jako sadu K matic o rozměru $(I \times J)$. Mějme $\mathbf{X}_A = [\mathbf{X}_{..1} \cdots \mathbf{X}_{..k} \cdots \mathbf{X}_{..K}]$, kde $\mathbf{X}_{..k}$ představuje k-tý člen módu C. Matice $\mathbf{X}_A$ obsahuje I řádků odpovídajících módu A a JK sloupců odpovídajících všem možným kombinancím módů B a C. Získáme ji tak, že poskládáme vedle sebe čelní řezy kvádrem $\underline{\mathbf{X}}$.

3.1. Tucker3 model

Model Tucker3 (T3) můžeme vnímat jako multilineární model, který se snaží popsat informaci obsaženou v $\underline{\mathbf{X}}$ (resp. redukovat její dimenzi) tak, že pro každý mód můžeme vzít jiný počet komponent. Tento model lze při volbě

P komponent pro mód A, Q komponent pro mód B a R komponent pro mód C s komponentními maticemi $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ zapsat jako

$$\mathbf{X}_A = \mathbf{A}\mathbf{G}_A(\mathbf{C}^T \otimes \mathbf{B}^T) + \mathbf{E}_A. \tag{31}$$

Prvek a_{ip} matice $\mathbf{A}$ představuje i-tý prvek p-té komponenty pro mód A. Podobně prvek b_{jq} matice $\mathbf{B}$ představuje j-tou entitu ke q-té komponentě a prvek c_{kr} komponentní matice $\mathbf{C}$ je k-tým subjektem k r-té komponentě. $\mathbf{G}_A$ představuje transformaci jádra $\underline{\mathbf{G}}$ rozměru $(P \times Q \times R)$ do dvouindexového schématu. Prvky g_{pqr} jádra $\underline{\mathbf{G}}$ přitom vyjadřují interakce napříč našimi třemi módy. $\mathbf{E}_A$ značí chybovou matici. Symbol $\otimes$ značí Kroneckerův součin dvou matic, který můžeme pro dvě matice $\mathbf{U}$ a $\mathbf{V}$ o rozměrech $(I \times J)$ a $(K \times L)$ zapsat následovně

$$\mathbf{U} \otimes \mathbf{V} = \begin{bmatrix} u_{11}\mathbf{V} & \cdots & u_{1J}\mathbf{V} \\ \vdots & \ddots & \vdots \\ u_{I1}\mathbf{V} & \cdots & u_{IJ}\mathbf{V} \end{bmatrix}. \tag{32}$$

Model T3 ze vztahu (31) můžeme formulovat skalárním zápisem takto

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk}. \tag{33}$$

V případě modelu T3 se snažíme o redukci v rámci všech třech módů, ale jsou situace, kdy je výhodnější redukovat pouze dva módy (T2 model), případně pouze jeden (T1 model). Při redukci pouze módů A a B lze zapsat model T2 (T2-AB) následujícím způsobem

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}b_{jq}g_{pqr} + e_{ijk}. \tag{34}$$

Pokud bychom chtěli redukovat například pouze mód A, tak by nám

ze vztahu (34) modelu T2 vypadla suma $\sum_{q=1}^{Q}$, prvky matice komponent $\mathbf{B}$ a získali bychom tak model T1 (T1-A).

Optimální řešení těchto modelů lze nalézt minimalizací čtverce normy chybové matice

$$\|\mathbf{E}_A\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} e_{ijk}^2. \tag{35}$$

Minimalizaci vztahu (35) lze pro T3 a T2 provést pomocí algoritmu alternujících nejmenších čtverců (ALS - Alternating Least Squares). Jedná se o iterativní algoritmus, kdy na jeho začátku můžeme například náhodně zvolit hodnoty čtveřice matic $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ a $\mathbf{G}$ a následně vždy využijeme tři z objektů k odhadu toho čtvrtého, tj. odhad $\mathbf{G}$ provedeme pomocí $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, pomocí metody nejmenších čtverců. Následně již se získaným odhadem matice $\mathbf{G}$, odhadneme $\mathbf{A}$ pomocí $\mathbf{G}$, $\mathbf{B}$, $\mathbf{C}$. Iterujeme dokud se hodnoty ztrátové funkce nestabilizují, to znamená dokud hodnoty ve dvou po sobě jdoucích iteracích se liší maximálně o předem stanovenou toleranci - prahovou hodnotu. Dle [7] je dokázáno, že algoritmus konverguje v konečném počtu iterací alespoň k lokálnímu minimu. Abychom risk 'spadnutí' do lokálního minima minimalizovali, tak algoritmus můžeme realizovat vícekrát s různými náhodnými počátečními hodnotami matic. Jelikož T1 je ekvivalentem PCA na $\mathbf{X}_A$, tak řešení lze získat pomocí singulárního rozkladu matice $\mathbf{X}_A$.

Nicméně výhodou i nevýhodou T3 současně je možnost rotace faktorů. Nevýhodou proto, že díky tomu získané řešení není jednoznačné. To znamená, že můžeme získat ekvivalentní řešení při uvážení ortogonálních matic $\mathbf{S}$, $\mathbf{T}$ a $\mathbf{U}$ jako $\tilde{\mathbf{A}} = \mathbf{AS}$, $\tilde{\mathbf{B}} = \mathbf{BT}$, $\tilde{\mathbf{C}} = \mathbf{CU}$ a $\tilde{\mathbf{G}}_A = \mathbf{S}^{-1}\mathbf{G}_A[(\mathbf{U}^T)^{-1} \otimes (\mathbf{T}^T)^{-1}]$. Přitom rotace komponentních matic se projeví změnou vztahů v jádře. Výhodou takové rotace je možnost zjednodušení struktury řešení a tím pádem získání lepší interpretovatelnosti výsledků analýzy. To s sebou samozřejmě

přináší i další nevýhodu, a to takovou, že do výsledků se může propsat statistikova subjektivita.

3.2. Parafac model

Model Parafac, známý též jako Candecomp model (CP), cílí na redukci kvádru $\underline{\mathbf{X}}$ tak, že pro každý z módů A, B, C vezme stejný počet komponent, řekněme S. Skalárně model zapíšeme následovně,

$$x_{ijk} = \sum_{s=1}^{S} a_{is} b_{js} c_{ks} + e_{ijk}. \tag{36}$$

CP model můžeme vnímat jako speciální případ modelu T3 za předpokladu, že $P = Q = R$ a

$$g_{pqr} = \begin{cases} 1, & p = q = r \\ 0, & \text{jinak.} \end{cases} \tag{37}$$

Zapíšeme-li model CP maticově, tak nám to pomůže porozumět lépe vtahu mezi modely. Mějme tedy maticový zápis CP

$$\mathbf{X}_A = \mathbf{A}\mathbf{I}_A(\mathbf{C}^T \otimes \mathbf{B}^T) + \mathbf{E}_A, \tag{38}$$

kde $\mathbf{I}_A$ je maticovou verzí kvádru $\underline{\mathbf{I}}$, jehož prvky i_{pqr} jsou

$$i_{pqr} = \begin{cases} 1, & p = q = r \\ 0, & \text{jinak.} \end{cases} \tag{39}$$

Srovnáme-li nyní vztahy (31) a (38), tak můžeme vypozorovat, že model Parafac získáme jako model T3, kde $\mathbf{G}_A = \mathbf{I}_A$. Z toho nám vyplývá i jednoznačnost řešení CP, jelikož zde není možná žádná rotace z důvodu

jednotkového jádra - rotaci v rámci komponentích matic není kde vykompenzovat.

3.3. Aplikace na data

Máme k dispozici údaje o kriminálních činech za roky 2016 až 2022, tedy 14 pozorování v módu A, 6 proměnných v módu B a 7 vrstev v módu C. Pro analýzu využijeme balíček `ThreeWay`. Pro zahájení analýzy pomocí modelu CP spustíme funkci `CP`. Před zahájením ale musíme převést data za jednotlivé roky do clr koeficientů a pomocí příkazu `cbind` poskládáme transformovaná data vedle sebe. Dále si musíme uložit názvy pro jednotlivé módy. Poté již můžeme přejít ke spuštění samotné funkce pro analýzu pomocí modelu CP.

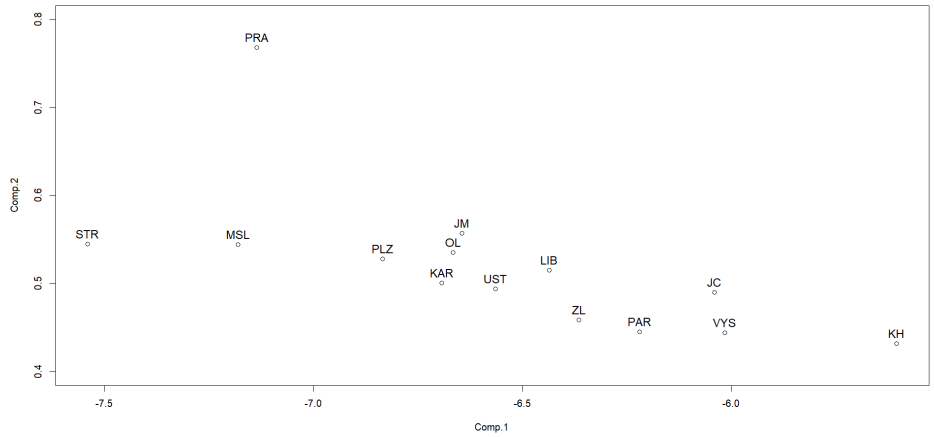
```
>CCRbyKN2016_CLK <- cenLR(CRbyKN2016_CLK)$x.clr
>CCRbyKN2017_CLK <- cenLR(CRbyKN2017_CLK)$x.clr
>CCRbyKN2018_CLK <- cenLR(CRbyKN2018_CLK)$x.clr
>CCRbyKN2019_CLK <- cenLR(CRbyKN2019_CLK)$x.clr
>CCRbyKN2020_CLK <- cenLR(CRbyKN2020_CLK)$x.clr
>CCRbyKN2021_CLK <- cenLR(CRbyKN2021_CLK)$x.clr
>CCRbyKN2022_CLK <- cenLR(CRbyKN2022_CLK)$x.clr
>AK <- cbind(CCRbyKN2016_CLK,CCRbyKN2017_CLK,CCRbyKN2018_CLK,CC
RbyKN2019_CLK,CCRbyKN2020_CLK,CCRbyKN2021_CLK,CCRbyKN2022_CLK)
>laba <- rownames(CCRbyKN2022_CLK)
>labb <- colnames(CCRbyKN2022_CLK)
>labc <- c("2016","2017","2018","2019","2020","2021","2022")
>AKCP <- CP(AK,laba,labb,labc)
```

Otevře se nám prostředí funkce, které se nás zeptá na několik otázek. Vyplníme do něj počet pozorování, proměnných a vrstev v jednotlivých módech a následně stanovíme počet komponent pro každý mód - zde vezmeme

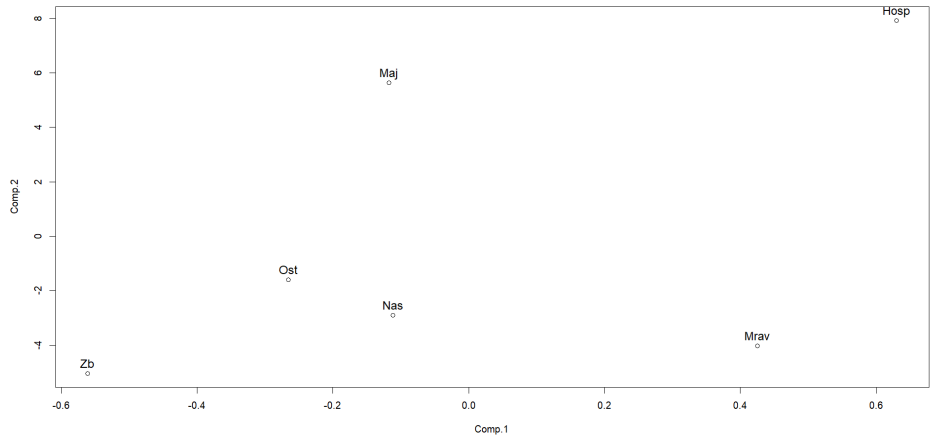
2 komponenty. Po projití celým prostředím dostaneme výsledek. Musíme přitom také uvážit, že podobně jako u PCA je řešení v jednotlivých módech jednoznačné až na orientaci (násobek (-1)), což využijeme u módu B pro odpovídající interpretaci a vystihnutí struktury objevující se v datech v průběhu let.

```
>AKCP$B <- AKCP$B*(-1)
>plot(AKCP$A, main="PARAFAC - mod A", ylim = c(0.4, 0.8))
>text(AKCP$A, labels=AKCP$laba, pos=3, cex=1.2)
>plot(AKCP$B, main="PARAFAC - Mod B")
>text(AKCP$B, labels=AKCP$labb, pos=3, cex=1.2)
>plot(AKCP$C, main="PARAFAC - mod C")
>text(AKCP$C, labels=AKCP$labc, pos=3, cex=1.2)
```

Graficky znázorněné výsledky můžeme vidět v obrázcích 6, 7 a 8.



Obrázek 6: Metoda PARAFAC 2016-2022 - mód A.

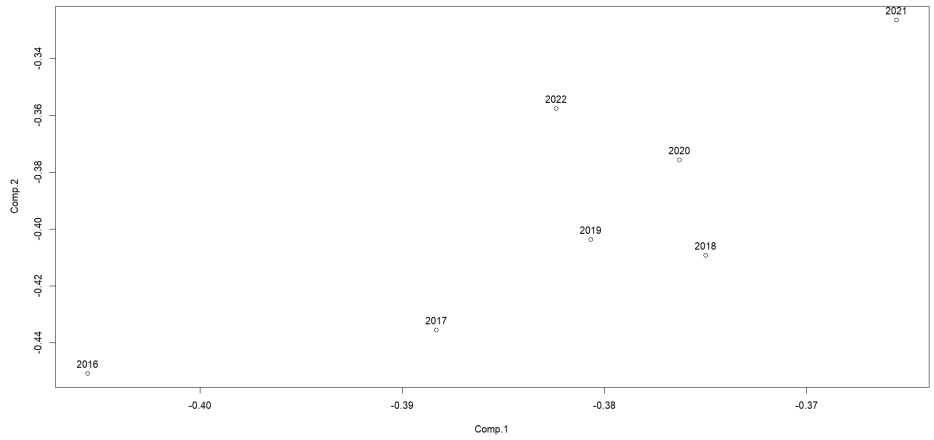


Obrázek 7: Metoda PARAFAC 2016-2022 - mod B.

Pro pochopení výsledků Parafac analýzy je potřeba pracovat při interpretaci s jednotlivými módy současně. Vezmeme-li mód A a B, srovnáme si je, tak z módu B vyplývá, že kraje, které se budou blížit k pravému dolnímu rohu, tak v nich bude dominovat nějakým způsobem, pozitivně či negativně, mravnostní kriminalita. Ta v sedmiletém časovém úseku dominuje v Jihočeském, Pardubickém, Královehradeckém kraji a na Vysočině. Vidíme, že i přes přidání časové dimenze, tak v Praze jednoznačně převažují majetkové trestné činy. Pro kraje situované uprostřed dominuje ostatní a násilná kriminalita. To, že ostatní kriminalita dominuje ve Středočeském a Moravskoslezském kraji jsme pozorovali i v biplotu PCA v obrázku 4 a bylo tomu tak i v předcházejících letech. Čím blíže ale budeme levému dolnímu rohu, tak tím spíše bude převažovat zbývající kriminalita. To, že by hospodářská kriminalita v dlouhodobém měřítku někde dominovala, se na základě módu A a B úplně říci nedá. Nicméně, při pohledu na původní data, ze struktury v jednotlivých letech se ve směru hospodářské kriminality střídají různé kraje, přičemž stálicí je nějakým způsobem Praha, u které je ale výrazná dominance majetkové kriminality.

V případě módu A je zajímavé, jak se k sobě různě napárovaly kraje se společnými hranicemi. Přirozeně se dalo očekávat, že kraje spolu sousedící budou mít podobnou strukturu kriminality. Vidíme tak spolu například Jihomoravský a Olomoucký kraj. Zároveň je zde hezká 'posloupnost' pohraničních krajů - Libereckého, Ústeckého, Karlovarského a Plzeňského. Případně Jihočeského, Vysočiny a Pardubického. Provedli jsme i PCA na jednotlivé roky 2016 až 2022 pro porovnání struktury, která se v jednotlivých letech vyskytovala, abychom dokázali posoudit smysluplnost výsledků Parafacu. Dle tohoto pozorování došlo k velmi dobrému vystihnutí časového aspektu a výsledky vyplývající ze srovnání jednotlivých módů dávají pěkný smysl.

Na následujícím obrázku 8 vyobrazujícím mód C lze vysledovat různorodost struktury kriminality v průběhu let 2016 až 2023.



Obrázek 8: Metoda PARAFAC 2016-2022 - mód C.

Pozorujeme odlišnosti v struktuře v letech 2016 a 2021, každopádně o důvodech můžeme pouze polemizovat. Žádný relevantní důvod pro efekt odhlehlosti v roce 2016 se bohužel nepodařilo dohledat, nicméně rok 2021 byl rokem silné vlny nemoci COVID-19. Dá se tak očekávat, že při všech různých opatřeních a počtu nakažených v daném roce, menšímu počtu zahraničních turistů, to zřejmě nějaký vliv na strukturu krimiality mělo. Zároveň ale v roce

2020, kdy nemoc v ČR vypukla, byla opatření drakoničtější, ale daný rok strukturou zapadl mezi ostatní.

4. Kompoziční tabulky

V kapitole 2 jsme brali kraje jako pozorování, trestné činy jako proměnné. V navazující kapitole 3 jsme přidali faktor času a podívali se na strukuru za roky 2016 až 2022. V této kapitole se opět zaměříme na rok 2022 s tím, že nyní nám přibude faktor objasněnosti. To znamená, že jednotlivé trestné činy budou rozděleny na objasněno a neobjasněno. V rámci kapitoly si řekneme, co to vlastně ty kompoziční tabulky jsou, jak se s nimi může pracovat a na interakční části zkusíme provést Parafac pro zařazení časového aspektu. Zdrojem informací o kompozičních tabukách v této práci jsou [2] a [3].

Kompoziční tabulka představuje datovou tabulku, která je uspořádaná vzhledem ke dvěma faktorům a popisuje relativní příspěvky k danému celku, jenž je rozdělen na tyto dva faktory. Jedná se vlastně o dvoufaktorové rozšíření kompozičního vektoru, který v sobě zahrnuje relativní informaci mezi danými faktory, které se nám nyní v tabulce vyskytují v řádcích a ve sloupcích. Kompoziční tabulku $\mathbf{x}$ si lze představit následovně,

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}, x_{ij} > 0, i = 1, \ldots, I, j = 1, \ldots, J. \tag{40}$$

Jelikož kompoziční tabulka představuje pouze určité rozšíření kompozičního vektoru, tak zůstávají v platnosti operace zavedené v kapitole 1.2. Pro reprezentaci kompoziční tabulky $\mathbf{x}$ v IJ-složkovém simplexu S^{IJ} vektorizovaných tabulek $\text{vec}(\mathbf{x}) = (x_{11}, \ldots, x_{I1}, \ldots, x_{IJ})$ je operace uzávěru $C(\mathbf{x})$

definována takto

$$C(\mathbf{x}) = \begin{pmatrix} \frac{\kappa x_{11}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{1J}}{\sum_{i,j} x_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\kappa x_{I1}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{IJ}}{\sum_{i,j} x_{ij}} \end{pmatrix}. \tag{41}$$

Pro kompoziční tabulky $\mathbf{x}$, $\mathbf{y}$ a $\alpha \in \mathbb{R}$ můžeme postupně definovat nám známé operace, jakými jsou perturbace $\mathbf{x} \oplus \mathbf{y}$,

$$\mathbf{x} \oplus \mathbf{y} = \begin{pmatrix} x_{11}y_{11} & \cdots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \cdots & x_{IJ}y_{IJ} \end{pmatrix}, \tag{42}$$

mocninná transformace $\alpha \odot \mathbf{x}$,

$$\alpha \odot \mathbf{x} = \begin{pmatrix} x_{11}^{\alpha} & \cdots & x_{1J}^{\alpha} \\ \vdots & \ddots & \vdots \\ x_{I1}^{\alpha} & \cdots & x_{IJ}^{\alpha} \end{pmatrix}, \tag{43}$$

a Aitchisonův skalární součin $\langle \mathbf{x}, \mathbf{y} \rangle_A$

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\alpha} = \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}}. \tag{44}$$

Analyzovat původní kompoziční tabulku $\mathbf{x}$ jako takovou pro nás nemusí být dostatečně vypovídající. Při užití ortogonální dekompozice ji lze rozdělit na interakční $\mathbf{x}_{int}$ a nezávislou část $\mathbf{x}_{ind}$, což nám může dát lepší vhled do původní tabulky. Dekompozici můžeme zapsat následovně,

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int}. \tag{45}$$

Konstrukce nezávislé části $\mathbf{x}_{ind}$ kompoziční tabulky $\mathbf{x}$ je provedena tak, aby extrahovala veškerou relativní informaci o sloupcovém a řádkovém fak-

toru za podmínky, že původní kompoziční tabulka $\mathbf{x}$ je součinem svých řádkových a sloupcových marginálních vektorů, ovšem v tomto případě tvořených geometrickými průměry jednotlivých řádků, respektive sloupců. Informace o vztazích mezi těmito dvěma faktory je obsažena v interakční části $\mathbf{x}_{int}$ a snaží se je charakterizovat při odchýlení od nezávislosti mezi faktory. Pokud by mezi danými faktory panovala plná nezávislost, tak by si jednotlivé prvky v interakční části byly rovny, jelikož po dekompozici by pro interakční část již žádná informace nezbývala.

4.1. Pivotové souřadnice

Pro práci s kompozičními tabulkami v software `R` budeme používat funkci `tabCoordWrapper`, která pracuje s pivotovými souřadnicemi. Tudíž si v této podkapitole řekneme, co to pivotové souřadnice jsou a jakým způsobem se z nich dostaneme zpět do clr koeficientů. Pro souřadnicovou reprezentaci kompozičních tabulek tedy využijeme pivotové souřadnice, jež představují speciální případ izometrických logpodílových souřadnic $\mathbf{z} \in \mathbb{R}^{D-1}$ reprezentující ortonormální souřadnice kompozice $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$ vzhledem k Aitchisonově geometrii, a které lze vypočíst jako

$$\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}^1 \rangle_A, \langle \mathbf{x}, \mathbf{e}^2 \rangle_A \ldots, \langle \mathbf{x}, \mathbf{e}^{D-1} \rangle_A)^{\mathrm{T}}, \tag{46}$$

kde D-složkové kompozice $\mathbf{e}^i = C(e_1^i, e_2^i, \ldots, e_D^i), i = 1, \ldots, D-1$, tvoří ortonormální bázi na simplexu.

Ve srovnání s clr koeficienty je u ilr koeficientů problém s interpretací v takovém smyslu, že způsobů jejich konstrukce je teoreticky nekonečně mnoho v závislosti na volbě bázových vektorů $\mathbf{e}^i, i = 1, \ldots, D-1$. K takovéto volbě se například užívá sekvenční (postupné) binární dělení (SBP), nicméně vy-

žaduje významnou znalost kompozic a výsledkem jsou souřadnice zvané bilance. Výhoda pivotových souřadnic je taková, že jejich využití je vhodné v situacích, kdy nám chybí znalost k provedení SBP. Pivotové souřadnice lze vypočítat jako

$$z_i^{(p)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(p)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(p)}}}, \tag{47}$$

kde $x_i^{(p)}$ představuje i-tou část přeskládané kompozice $(x_p, x_1, \ldots, x_{p-1}, x_{p+1}, \ldots, x_D)$. Musíme provést permutaci kompozičních složek v každém z D souřadnicových systémů tak, aby p-tá složka $\mathbf{x}$ byla na první pozici. Je to z toho důvodu, že v každém systému první pivotová souřadnice $z_1^{(p)}$ vysvětluje veškerou relativní informaci o složce x_1 a navíc, je proporcionální k příslušnému clr koeficientu,

$$\mathrm{clr}(x_1^{(p)}) = \sqrt{\frac{D}{D-1}} z_1^{(p)}. \tag{48}$$

Pro další postup je důležitá existence lineární transformace pro přecházení mezi clr a ilr koeficienty pomocí matice $\mathbf{V}$ rozměru $D \times (D-1)$, která obsahuje clr reprezentace ilr bázových vektorů. Následující vztah je platný i pro pivotové souřadnice,

$$\mathrm{clr}(\mathbf{x}) = \mathbf{V}\mathbf{z} = [\mathrm{clr}(\mathbf{e}_1)^{\mathrm{T}}, \mathrm{clr}(\mathbf{e}_2)^{\mathrm{T}}, \ldots, \mathrm{clr}(\mathbf{e}_{D-1})^{\mathrm{T}}] \cdot \mathrm{ilr}(\mathbf{x}). \tag{49}$$

Dle [2] vyplývá z ortogonality procesu dekompozice kompoziční tabulky, že dimenze interakční části $\mathbf{x}_{int}$ klesne na $(I-1)(J-1)$ a nezávislé části $\mathbf{x}_{ind}$ na $I+J-2$. Vhodnou souřadnicovou reprezentací ctící tyto dimenze tabulek jsou právě pivotové souřadnice. Jak již bylo zmíněno v úvodu podkapitoly, tak právě v pivotových souřadnicích budou tabulky vyjádřené i při užití

funkce `tabCoordWrapper`, se kterou budeme pracovat dále. To nám ale nutně nevadí, jelikož máme vztah (49), tudíž můžeme velmi jednoduše přejít do clr koeficientů, které nám jsou intepretačně bližší.

Pivotové souřadnice dělíme obecně na tři typy, přičemž z_i^c odpovídá sloupcům, z_i^r řádkům a z_{rs}^{OR} 'poměru šancí' dělení kompoziční tabulky. Nezávislá část je pak reprezentována pomocí souřadnic z_i^r a z_i^c, a interakční část pomocí souřadnic z_{rs}^{OR}. Dohromady tvoří souřadnicovou reprezentaci původní tabulky $\mathbf{x}$.

Vypočet řádkového typu souřadnic proběhne tak, že vezmeme celý první řádek jako pivotový prvek a oddělíme ho od ostatních. V dalším kroce vezmeme následující řádek jako nový pivotový element a oddělíme od zbývajících, a tak dále až do chvíle, než budeme mít určeno všech $I + J - 2$ pivotových souřadnic. Přitom výpočet sloupcových souřadnic probíhá principiálně stejně, pouze místo řádků uvažujeme sloupce. Vztahy pro výpočet z_i^r a z_i^c mají následující podobu,

$$z_i^r = \sqrt{\frac{(I-i)J}{1+I-i}} \ln \frac{g(\mathbf{x}_{i\bullet})}{[g(\mathbf{x}_{i+1\bullet}), \ldots, g(\mathbf{x}_{I\bullet})]^{1/(I-i)}}, \quad i = 1, \ldots, I-1, \qquad (50)$$

$$z_j^c = \sqrt{\frac{I(J-j)}{1+J-j}} \ln \frac{g(\mathbf{x}_{\bullet j})}{[g(\mathbf{x}_{\bullet j+1}), \ldots, g(\mathbf{x}_{\bullet J})]^{1/(J-j)}}, \quad j = 1, \ldots, J-1, \qquad (51)$$

kde $g(\mathbf{x}_{i\bullet})$ je geometrický průměr i-tého řádku a $g(\mathbf{x}_{\bullet j})$ geometrický průměr j-tého sloupce.

Výpočet zbylých $(I-1)(J-1)$ souřadnic je postaven na rozdělení původní kompoziční tabulky na 4 bloky,

$$\begin{pmatrix} A & \vdots & B \\ \cdots & \cdots & \cdots \\ C & \vdots & D \end{pmatrix},$$

kde blok A obsahuje vždy pouze jednu buňku, pivot, s indexem rs. Vzorec pro výpočet z_{rs}^{OR} obsahuje v logpodílu v čitateli prvky bloků A a D, ve jmenovateli prvky bloků B a C

$$z_{rs}^{OR} = \sqrt{\frac{1}{(I-r)(J-s)(I-r+1)(J-s+1)}} \ln \prod_{i=r+1}^{I} \prod_{j=s+1}^{J} \frac{x_{ij}x_{rs}}{x_{is}x_{rj}}. \quad (52)$$

Abychom dokázali spočítat všechny souřadnice z^{OR} v pořadí odpovídajícím souřadnicím z^r a z^c, tak je nutné, abychom pivotové buňky brali nejprve podle řádků se zafixovaným prvním sloupcem, $r = 1, \ldots, I-1$, pak vezmeme sloupec po sloupci se zafixovaným posledním řádkem, $s = 1, \ldots, J-1$. Následně je pozice řádku snížena o jedničku a pozici sloupců $s = 1, \ldots, J-1$, opět měníme pro daný řádek, dokud nepokryje plný rozměr $r \times s$ tabulky.

Každopádně je nutné podotknout, že můžeme clr koeficienty interakční a nezávislé části napočítat i napřímo z původní kompoziční tabulky $\mathbf{x}$ pomocí následujících dvou vztahů (53) a (54).

$$\mathrm{clr}(\mathbf{x}_{ind})_{ij} = \ln \frac{g(\mathbf{x}_{i\bullet})g(\mathbf{x}_{\bullet j})}{g(\mathbf{x}_{\bullet\bullet})^2}, \quad (53)$$

$$\mathrm{clr}(\mathbf{x}_{int})_{ij} = \ln \frac{x_{ij}g(\mathbf{x}_{\bullet\bullet})}{g(\mathbf{x}_{i\bullet})g(\mathbf{x}_{\bullet j})}, \quad (54)$$

kde $g(\mathbf{x}_{i\bullet})$ je geometrický průměr i-tého řádku, $g(\mathbf{x}_{\bullet j})$ geometrický průměr j-tého sloupce a $g(\mathbf{x}_{\bullet\bullet})$ geometrický průměr prvků celé kompoziční tabulky.

4.2. Aplikace na data

V této podkapitole aplikujeme představenou teorii na data o kriminiálních činech. Využijeme pro to v software `R` knihovny `robCompositions` a `MASS`, přičemž klíčovou pro nás bude funkce ukrývající se v první z těchto knihoven,

a to funkce `tabCoordWrapper`. Do této funkce ale nemůžeme vložit data ve tvaru jako v předchozích metodách, ale musíme je vkládat ve formátu, jaký nám ukazuje následující tabulka 6. Jedná se vlastně o kombinaci tabulek 4 a 5.

Kraj	Trestný čin	Stav	Počet
PRA	Nas	O	74
STR	Nas	O	48
JC	Nas	O	92
PLZ	Nas	O	63
UST	Nas	O	130
KH	Nas	O	70
JM	Nas	O	61
MSL	Nas	O	95
OL	Nas	O	94
ZL	Nas	O	78
VYS	Nas	O	82
PAR	Nas	O	45
LIB	Nas	O	97
KAR	Nas	O	96
PRA	Nas	N	63
⋮	⋮	⋮	⋮
KAR	Hosp	N	57

Tabulka 6: Přehled kriminálních činů v ČR 2022 dle krajů. Vstup pro funkci `tabCoordWrapper`.

V tabulce 6 vidíme, že máme čtyři sloupce, kdy jeden odpovídá názvu kraje, což je pro tuto funkci ID pozorování, trestný čin je sloupcový faktor a stav je řádkový faktor. Počet představuje hodnotu kombinace těchto dvou faktorů. Funkce `tabCoordWrapper` nám vytvoří kompoziční tabulku pro každý kraj jako v tabulce 7 pro Olomoucký kraj.

	Hosp	Maj	Mrav	Nas	Ost	Zb
N	80	444	12	40	57	44
O	66	213	15	94	184	221

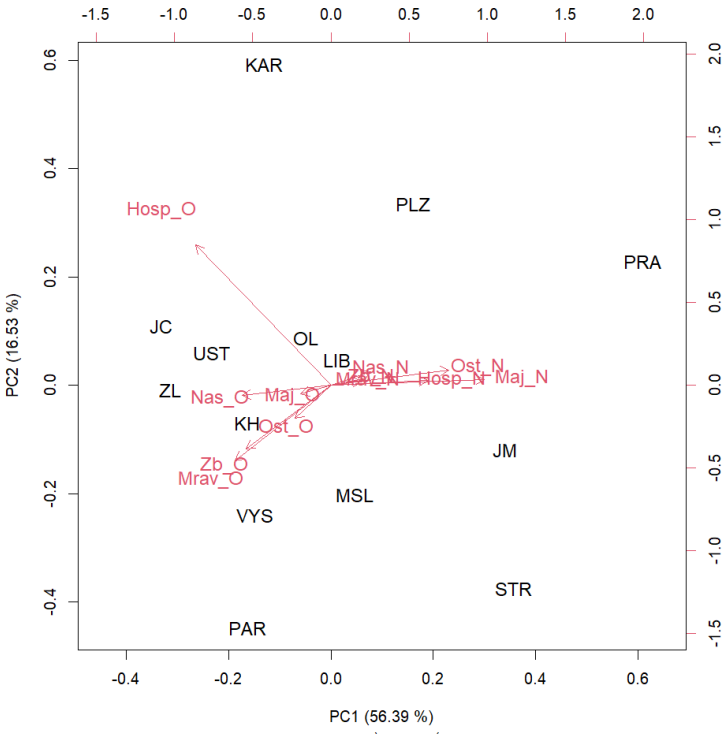
Tabulka 7: Kompoziční tabulka pro Olomoucký kraj pro rok 2022.

Zároveň vypočte pro naše data pivotové souřadnice. Jelikož máme kompoziční tabulky rozměru 2×6, tak budeme mít jednu řádkovou pivotovou souřadnici z_1^r, pět souřadnic odpovídajících sloupcům $z_1^c, z_2^c, \ldots, z_5^c$ a pět souřadnic $z_{11}^{OR}, z_{12}^{OR}, \ldots, z_{15}^{OR}$. Navíc dostaneme z této funkce i kontrastní matici, která nám poslouží k převodu na clr koeficienty. Vyjádříme si odpovídající část kontrastní matice pro souřadnice reprezentující interakční část, pro souřadnice nezávislé části, a jejich součiny s příslušnými souřadnicemi získáme vyjádření těchto dvou částí v clr koeficientech. Na takto upravených datech již můžeme postupně provést metodu hlavních komponent, kdy prvně aplikujeme PCA na původní tabulky, následně na nezávislou část a na závěr na interakční část. Kód pro provedení doposud vysvětleného postupu v této podkapitole je vložen pod tímto odstavcem.

```
>X22<-tabCoordWrapper(CRbyKN2022_KT,obs.ID='Kraj',row.factor =
'Stav', col.factor = 'Trestny_cin', value='Pocet')
>## Vyjádříme souřadnice pro celou tabulku
>X22_C<-X22$Coordinates
>## Vyjádříme souřadnice pro interakční tabulku
>X22_INT<-X22$Int.coord
>## Vyjádříme souřadnice pro nezávislou část
>X22_IND<-X22$Ind.coord
# Vyjádříme kontrastní matici
>X22_CM<-X22$Contrast.matrix
>X22_clr<-X22_C%*%X22_CM
>colnames(X22_clr)<-c("Hosp_N","Maj_N","Mrav_N","Nas_N","Ost_N",
"Zb_N","Hosp_O","Maj_O","Mrav_O","Nas_O","Ost_O","Zb_O")
>## PCA celé tabulky v clr
>PCA22<-princomp(X22_clr)
```

```
>biplot(PCA22, xlab="PC1 (56.39 %)", ylab="PC2 (16.53 %)",xlim
= c(-0.45, 0.65),cex=1.2) # Obrázek 9
>## Nezávislá část
>## Vyjádříme z CM to, co odpovídá nezávislé části
>X22_C_IND<-X22_CM[1:6,]
>X22_IND_clr<-X22_IND%*%X22_C_IND # nezávislá část v clr
>colnames(X22_IND_clr)<-c("Hosp_N","Maj_N","Mrav_N","Nas_N",
"Ost_N","Zb_N","Hosp_O","Maj_O","Mrav_O","Nas_O","Ost_O","Zb_O")
>PCA22_ind<-princomp(X22_IND_clr)
>summary(PCA22_ind)
>biplot(PCA22_ind, xlab="PC1 (72.01 %)", ylab="PC2 (14.60 %)",
xlim=c(-0.45, 0.65), ylim=c(-0.55,0.45),cex=1.2) # Obrázek 10
>## Interakční část
>X22_C_INT<-X22_CM[7:11,]
>X22_INT_clr<-X22_INT%*%X22_C_INT #interakční část v clr
>colnames(X22_INT_clr)<-c("Hosp_N","Maj_N","Mrav_N","Nas_N",
"Ost_N","Zb_N","Hosp_O","Maj_O","Mrav_O","Nas_O","Ost_O","Zb_O")
>PCA22_int<-princomp(X22_INT_clr)
>summary(PCA22_int)
>biplot(PCA22_int, xlab="PC1 (56.98 %)", ylab="PC2 (20.40 %)",
xlim = c(-0.65, 0.4), cex=1.2) # Obrázek 11
```

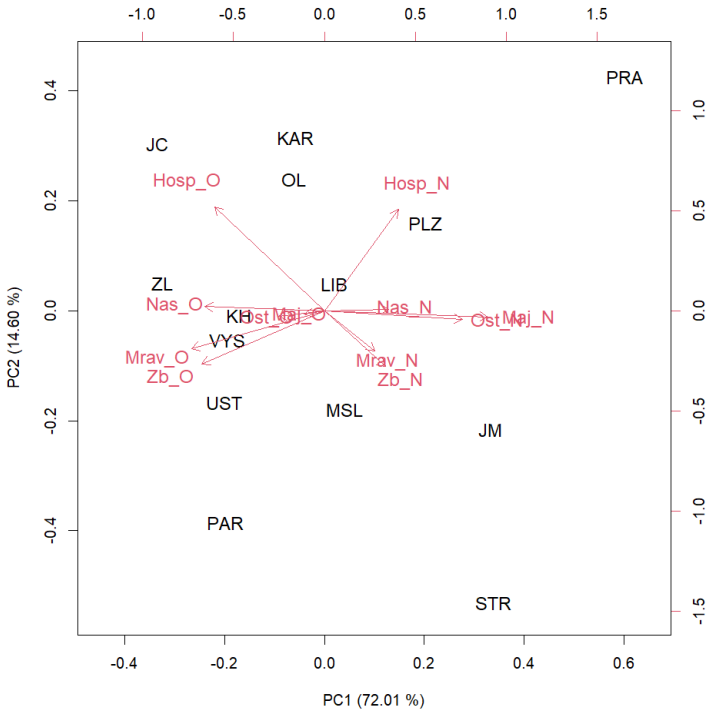
Výsledky jednotlivých biplotů můžeme prozkoumat na následujících obrázcích 9, 10 a 11.



Obrázek 9: Biplot PCA kompoziční tabulky pro data z roku 2022.

Způsob interpretace biplotu kompoziční tabulky je obdobný biplotu v kapitole 2. V našem případě jsme v obrázku 4 vypozorovali, že majetkové a hospodářské trestné činy jsou dominantní v rámci relativní struktury trestných činů Prahy, zatímco v Karlovarském kraji to byly především hospodářské prohřešky proti zákonu. Zde ale vidíme při přidání faktoru objasněnosti, že danou dominanci způsobují objasněné hospodářské trestné činy, v Praze naopak neobjasněné hospodářské a majetkové. V absolutních čísel v tabulkách 4 a 5 vidíme, že objasněnost majetkových trestných činů je velmi nízká a onen vysoký počet neobjasněných majetkových trestných činů se promítá následně i do relativní struktury. Důvodem tohoto efektu může být dle mého názoru turismus, kdy různé drobné krádeže spáchají cizinci a šance na dopadení pachatele je velmi nízká. Liberecký kraj se nám stejně jako před přidáním

faktoru objasněnosti nachází uprostřed biplotu a můžeme tak opět jen konstatovat, že v jeho struktuře nedominuje žádná složka. Biplot nezávislé části, jenž nám má představovat situaci ideálního světa, to jest situaci, kdy jsou dané faktory nezávislé, nám představuje obrázek 10.

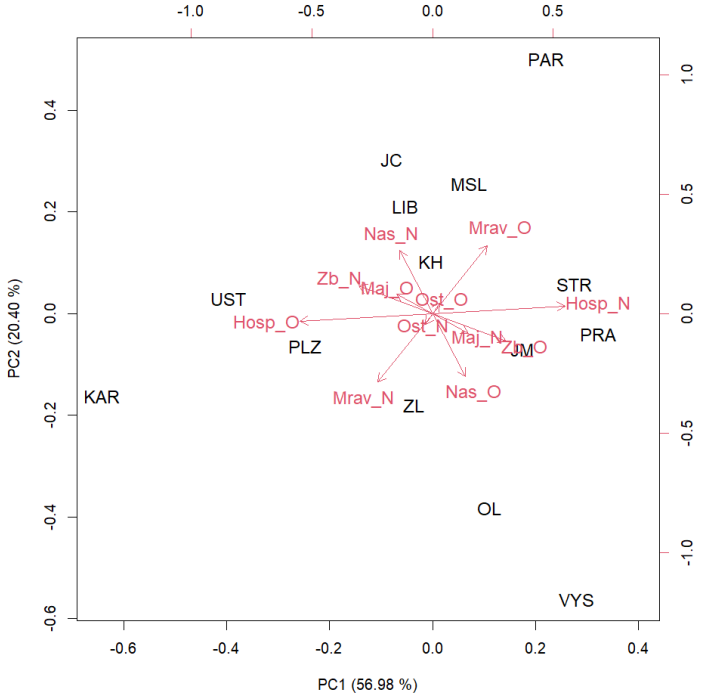


Obrázek 10: Biplot PCA na nezávislé části kompoziční tabulky za rok 2022.

Biplot pracující s nezávislou částí kompozičních tabulek nám tedy zachytává hypotetickou rovnováhu ve smyslu kombinace daných faktorů v případě nezávislosti. To znamená, že se nám vztah mezi faktorem objasněnosti a typem kriminálního činu vyfiltruje. Biplot se hezky rozdělil na dvě části, kde v levé polovině pozorujeme relativní dominanci objasněných, a naopak v pravé části neobjasněných trestných činů. Liberecký kraj se stále nachází uprostřed. Hospodářské činy dominovaly v Praze a Karlovarském kraji. Neobjasněnost těchto činů v Praze jednoznačně dominuje, zatímco Karlovarský

kraj se nachází uprostřed mezi objasněností a neobjasněností. To znamená, že ano, hospodářská kriminalita v Karlovarském kraji dominuje, ale nedá se říct, zdali spíš v objasněném či neobjasněném slova smyslu. Například u Ústeckého a Moravskoslezského kraje, kde primárně můžeme dohledat mravnostní prohřešky, vidíme, že v Ústeckém kraji končily spíše úspěšným objasněním a v Moravskoslezském kraji hříšníkům spíše jejich prohřešky v rámci daného roku prošly.

Interakční část nám naopak předkládá situaci při porušení nezávislosti (respektive míru odchýlení od nezávislosti), tedy ukazuje nám, ve kterých regionech je toto odchýlení od předpokladu nezávislosti vyšší než u jiných.



Obrázek 11: Biplot PCA na interakční části kompoziční tabulky za rok 2022.

Biplot stále interpretujeme ve smyslu dominance. Tato dominance ale neznamená, že když budeme mít ve směru šipky hospodářská neobjasněno Prahu, tak že v Praze je nejvíce neobjasněné hospodářské trestné činnosti. Pouze tam tato kombinace faktorů způsobuje odchýlení od nezávislosti. Každopádně zřejmě právě neobjasněná hospodářská kriminalita způsobuje v Praze odchýlení od nezávislosti nejvíce. Je třeba mít tedy na paměti, že interakční tabulka sama o sobě nám nedává jasnou informaci o dominanci jednotlivých kombinací obou faktorů. To, že byla správně provedena dekompozici kompoziční tabulky na interakční a nazávislou část, nám napoví právě biplot interakční části. Když se podíváme na jednotlivé proměnné, tak šipky (ne)objasněnosti daného trestného činu jsou přímo naproti sobě. Této 'protilehlé' struktury dosáhneme vždy, pokud u daného faktoru budeme mít pouze dvě kategorie a provedeme rozklad správně. Může to tedy v takovémto případě sloužit jako kontrolní aspekt správně provedeného rozkladu.

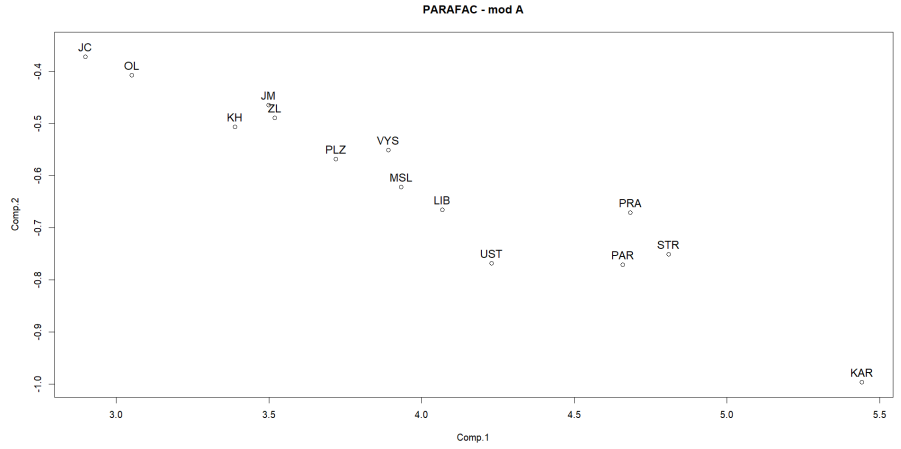
4.3. Interakční část KT a PARAFAC

Zajímavou myšlenkou může být pokus o přidání časového aspektu do kompozičních tabulek, vzít jejich interakční část za roky 2016 až 2022 a aplikovat metodu PARAFAC. Zajímavé to může být z toho důvodu, jelikož chceme vidět, zdali udrží v dlouhodobém aspektu právě onu 'protilehlou' strukturu. Metodu PARAFAC provedeme stejným způsobem jako v kapitole 3.3. Vypočteme interakční část naší kompoziční tabulky za roky 2016 až 2022, spojíme je pomocí `cbind` a následně na vzniklou datovou matici v clr koeficientech aplikujeme funkci `CP`.

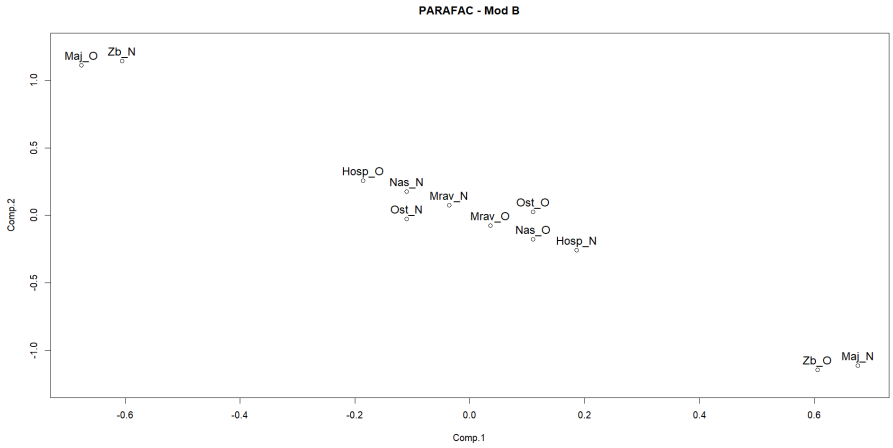
```
>INT_PAR<-cbind(X16_INT_clr,X17_INT_clr,X18_INT_clr,X19_INT_clr,
X20_INT_clr,X21_INT_clr,X22_INT_clr)
>Alab<-rownames(X22_INT_clr)
```

```
>Blab<-colnames(X22_INT_clr)
>Clab<-c("2016","2017","2018","2019","2020","2021","2022")
>INT_CP<-CP(INT_PAR,Alab,Blab,Clab)
># Parafac analysis with 2 components, gave a fit of 93.97 %
>plot(INT_CP$A, main="PARAFAC - mod A", ylim = c(-1.0, -0.35))
>text(INT_CP$A, labels=INT_CP$laba, pos=3, cex=1.2)
>INT_CP$B<-INT_CP$B*(-1)
>plot(INT_CP$B, main="PARAFAC - Mod B", ylim = c(-1.25, 1.25))
>text(INT_CP$B, labels=INT_CP$labb, pos=3, cex=1.2)
>plot(INT_CP$C, main="PARAFAC - mod C")
>text(INT_CP$C, labels=INT_CP$labc, pos=3, cex=1.2)
```

Grafické znázornění módu A a módu B si můžeme prohlédnout na následujících dvou obrázcích 12 a 13.

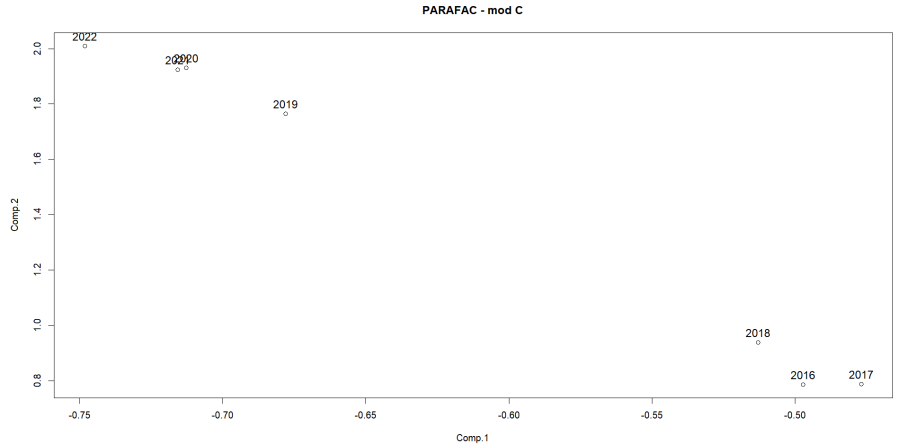


Obrázek 12: Metoda PARAFAC na $\mathbf{x}_{int}$ za roky 2016-2022 - mód A.



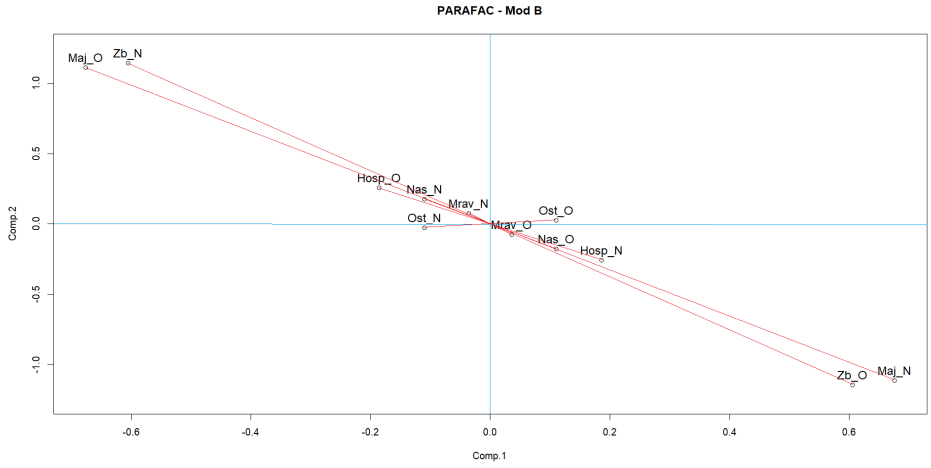
Obrázek 13: Metoda PARAFAC na $\mathbf{x}_{int}$ za roky 2016-2022 - mód B.

Pokud půjdeme od středu grafů do pravého dolního rohu, tak nám budou způsobovat majetkové neobjasněné a zbývající objasněné trestné činy větší odchýlení od toho ideálního stavu, tedy nezávislosti našich dvou faktorů objasněnosti a typu trestné činnosti. V Praze takové odchýlení způsobují násilné objasněné, hospodářské neobjasněné a ostatní objasněné trestné činy. V Olomouckém a Jihočeském kraji majetková objasněná a zbývající neobjasněná trestná činnost. Ná následujícím obrázku 14 vidíme strukturu módu C. Tam můžeme vypozorovat při cestě očima zprava doleva nějaký systematický trend ve vývoji v rámci let, kdy roky postupují zprava doleva téměř v řádném pořadí, narozdíl od situace, kterou jsme pozorovali na obrázku 8.



Obrázek 14: Metoda PARAFAC na $\mathbf{x}_{int}$ za roky 2016-2022 - mód C.

Abychom se ale vrátili k původní tezi této podkapitoly, tak se podívejme na obrázek 15. Chtěli jsme, respektive, doufali jsme, že bude zachována struktura interakční tabulky. To znamená, že jednotlivé trestné činy budou spolu s faktorem objasněnosti středově souměrné. V obrázku 15 jsme naznačili spojnice mezi těmito dvojicemi a lze vidět, že tato struktura zůstala i při přidání časového aspektu zachována.



Obrázek 15: Metoda PARAFAC na $\mathbf{x}_{int}$ za roky 2016-2022 - mód B. Naznačení zachování struktury interakční části při přidání časového aspektu.

5. Trojrozměrná korespondenční analýza

Stejně jako tomu bylo v kapitole 3, tak i zde budeme zkoumat počty trestných činů dle krajů, typu trestné činnosti a objasněnosti, tj. budeme uvažovat tři proměnné. Stejně jako jsme měli zobecnění PCA do trojrozměrného prostoru v podobě metod PARAFAC/Tucker3, tak trojrozměrná korespondenční analýza (CA3) je zobecněním CA. Cílem této metody je rozklad kontigenční tabulky takovým způsobem, že maximální množství informace o vztazích mezi proměnnými bude znázorněno v nízkodimenzionální vizualizaci. V této kapitole se pokusíme o stručné uvedení této metody a její aplikaci na naše data, přičemž informace budeme čerpat z publikace [6].

Existuje několik přístupů k CA3. My zvolíme ten klasický, kterým je symetrická trojrozměrná korespondenční analýza. Symetričnost je zde myšlena v takovém smyslu, že postavení všech tří proměnných je rovnocenné, neuvažujeme žádný kauzální vztah. V nesymetrické verzi by jedna proměnná byla uvažovaná jako vysvětlovaná a zbylé dvě jako vysvětlující. Symetrická CA3 je postavena na rozkladu Pearsonovy trojrozměrné $\mathbf{\Phi}^2$ statistiky a Tucker3 dekompozici $\mathbf{\Pi}_p$.

5.1. Pearsonova trojrozměrná statistika

Máme data reprezentována trojrozměrnou kontingenční tabulkou obsahující I řádků, J sloupců a K vrstev, přičemž každá buňka této tabulky představuje hodnotu (průsečíku) těchto tří proměnných.

Označme $\underline{\mathbf{N}}$ kontingenční tabulku o rozměru $I \times J \times K$ patřící do prostoru $\mathbb{R}^{I \times J \times K}$ jejímž ijk-tým prvkem je n_{ijk} pro $i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$. Dále označme $\underline{\mathbf{P}}$ tabulku relativních četností $\underline{\mathbf{N}}$, jejíž ijk-tý prvek

získáme jednoduchým dělením

$$p_{ijk} = \frac{n_{ijk}}{n}, \tag{55}$$

přičemž $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk} = 1$. Postupně definujme jednorozměrné částečné relativní četnosti

$$p_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}, \quad p_{\bullet j\bullet} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ijk}, \quad p_{\bullet\bullet k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}, \tag{56}$$

a dvourozměrné částečné relativní četnosti

$$p_{ij\bullet} = \sum_{k=1}^{K} p_{ijk}, \quad p_{i\bullet k} = \sum_{j=1}^{J} p_{ijk}, \quad p_{\bullet jk} = \sum_{i=1}^{I} p_{ijk}. \tag{57}$$

Dále definujme jednotkovou matici $\mathbf{I}_I$ řádu I z prostoru $\mathbb{R}^I$ a diagonální matice $\mathbf{D}_I \in \mathbb{R}^I$, $\mathbf{D}_J \in \mathbb{R}^J$ a $\mathbf{D}_K \in \mathbb{R}^I$ obsahující jednorozměrné částečné relativní četnosti $p_{i\bullet\bullet}$, $p_{\bullet j\bullet}$ a $p_{\bullet\bullet k}$.

Uvažujeme-li tedy proměnné jako symetrické, tj. bez kauzálních vztahů, tak můžeme vztahy mezi nimi analyzovat užitím Pearsonovy trojrozměrné $\mathbf{\Phi}^2$ statistiky,

$$\begin{aligned}
\mathbf{\Phi}^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk} - p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} - 1\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\pi_{p_{ijk}}\right)^2.
\end{aligned} \tag{58}$$

Za předpokladu nezávislosti lze $\mathbf{\Phi}^2$ rozdělit dle [6] následujícím způsobem

$$\begin{aligned}\mathbf{\Phi}^2 = &\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet\bullet}p_{\bullet j\bullet}\left(\frac{p_{ij\bullet}-p_{i\bullet\bullet}p_{\bullet j\bullet}}{p_{i\bullet\bullet}p_{\bullet j\bullet}}\right)^2 \\ &+\sum_{i=1}^{I}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet\bullet k}\left(\frac{p_{i\bullet k}-p_{i\bullet\bullet}p_{\bullet\bullet k}}{p_{i\bullet\bullet}p_{\bullet\bullet k}}\right)^2 \\ &+\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{\bullet jk}-p_{\bullet j\bullet}p_{\bullet\bullet k}}{p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2 \\ &+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk}-{}_{\alpha}p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2,\end{aligned} \tag{59}$$

kde

$${}_{\alpha}\hat{p}_{ijk} = \hat{p}_{ij\bullet}\hat{p}_{\bullet\bullet k} + \hat{p}_{i\bullet k}\hat{p}_{\bullet j\bullet} + \hat{p}_{\bullet jk}\hat{p}_{i\bullet\bullet} - 2\hat{p}_{i\bullet\bullet}\hat{p}_{\bullet j\bullet}\hat{p}_{\bullet\bullet k}. \tag{60}$$

Pro detailní rozbor odkazujeme na [6], v důsledků dostáváme následující rozklad

$$\mathbf{\Phi}^2 = \mathbf{\Phi}^2_{IJ} + \mathbf{\Phi}^2_{IK} + \mathbf{\Phi}^2_{JK} + \mathbf{\Phi}^2_{IJK}, \tag{61}$$

jehož členy charakterizují vztahy mezi dvojicemi faktorů, respektive mezi všemi faktory současně.

5.2. Dekompozice

V symetrické verzi trojrozměrné korespondenční analýzy jsou prvky trojrozměrné Pearsonovy $\mathbf{\Phi}^2$ statistiky rozloženy pomocí Tucker3 dekompozice. Zopakujeme-li velmi stručně poznatky o Tucker3 z kapitoly 3.1, tak trojrozměrnou matici $\underline{\mathbf{X}}$ s prvky x_{ijk} lze rozložit tak, jak tomu bylo ve vztahu (33), přičemž maticově lze Tucker3 dekompozici vyjádřit pomocí vztahu (31).

Dekompozici Tucker3 aplikujeme na trojrozměrné pole $\underline{\mathbf{\Pi}_p}$ s prvky

$$\pi_{p_{ijk}} = \frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} - 1, \tag{62}$$

přičemž se předpokládá, že matice komponent $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ jsou ortogonální vzhledem k diagonálním maticím obsahujícím jednorozměrné částečné relativní četnosti, tj.

$$\mathbf{A}^{\mathrm{T}}\mathbf{D}_I\mathbf{A} = \mathbf{I}_P, \quad \mathbf{B}^{\mathrm{T}}\mathbf{D}_J\mathbf{B} = \mathbf{I}_Q, \quad \mathbf{C}^{\mathrm{T}}\mathbf{D}_K\mathbf{C} = \mathbf{I}_R. \tag{63}$$

Můžeme si povšimnout, že Pearsonova trojrozměrná $\mathbf{\Phi}^2$ statistika (59) je vlastně váženým součtem čtverců prvků $\pi_{p_{ijk}}$. Lze tedy říci, že symetrická verze CA3 odpovídá minimalizaci váženého rozdílu čtverců mezi standardizovanými odchylkami nezávislosti v trojrozměrném poli (trojindexovém schématu) s aproximovanými hodnotami při užití modelu Tucker3, tj. minimalizujeme následující výraz

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\pi_{p_{ijk}} - \hat{\pi}_{p_{ijk}}\right)^2, \tag{64}$$

přičemž pro zvolené hodnoty $P \leq I$, $Q \leq J$ a $R \leq K$

$$\hat{\pi}_{p_{ijk}} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}a_{ip}b_{jq}c_{kr}. \tag{65}$$

Závěrem této podkapitoly je potřeba podotknout, že tento přístup k CA3 je přímým rozšířením klasické dvourozměrné CA, o které jsme se bavili již v kapitole 2.5.

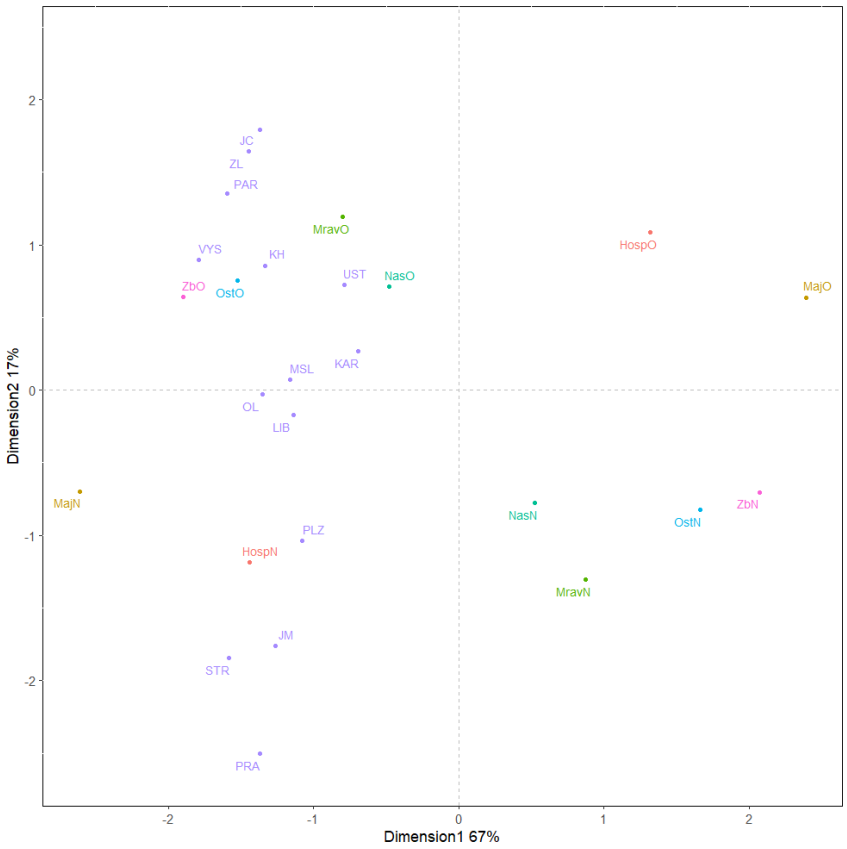
5.3. Aplikace na data

Po instalaci a načtení knihovny `CA3variants` zadáme datový vstup, což jsou v tomto případě tabulky 3 a 4. Pro vytvoření formátu tabulek požadovaného funkcí `CA3variants` využijeme funkci `structure`. V příkazu funkce si data ještě log-transformujeme (viz. kapitola 2.5), zvolíme počet komponent pro jednotlivé proměnné, tj. zde $2 \times 2 \times 1$ a typ zvolíme CA3, což odpovídá symetrické variantě. Kód výše popsaného následuje po tomto odstavci.

```
>install.packages("CA3variants")
>library(CA3variants)
>TC_2022<-structure(c(74,48,92,63,130,70,61,95,94,78,83,45,97,
96, 18,11,25,13,32,21,17,21,15,17,16,21,25,17,
287,157,231,236,440,201,200,422,213,165,145,175,307,298,
223,177,206,192,290,181,136,223,184,141,157,148,240,215,
136,159,203,181,239,175,137,175,221,184,152,174,192,177,
59,20,95,76,106,57,33,45,66,60,37,31,75,146,
63,47,48,56,58,36,47,56,40,27,20,23,57,53,
16,11,8,14,20,10,15,11,12,8,10,6,14,15,
1852,602,377,714,551,354,703,659,444,278,329,294,700,543,
160,57,43,68,65,47,87,64,57,32,35,32,65,70,
63,54,32,68,84,36,48,50,44,31,25,38,74,72,
175,54,50,73,59,48,69,60,80,38,46,36,83,57),
.Dim = c(14, 6, 2), .Dimnames=list(c("PRA","STR","JC","PLZ",
"UST", "KH","JM","MSL","OL","ZL","VYS","PAR","LIB","KAR"),
c("Nas","Mrav","Maj","Ost","Zb","Hosp"),
c("O","N")))
>krimi<-CA3variants(log(TC_2022), dims = c(p = 2, q = 2, r = 1),
ca3type = "CA3")
```

```
>plot(krimi,cex=1.5,addlines=F)
```

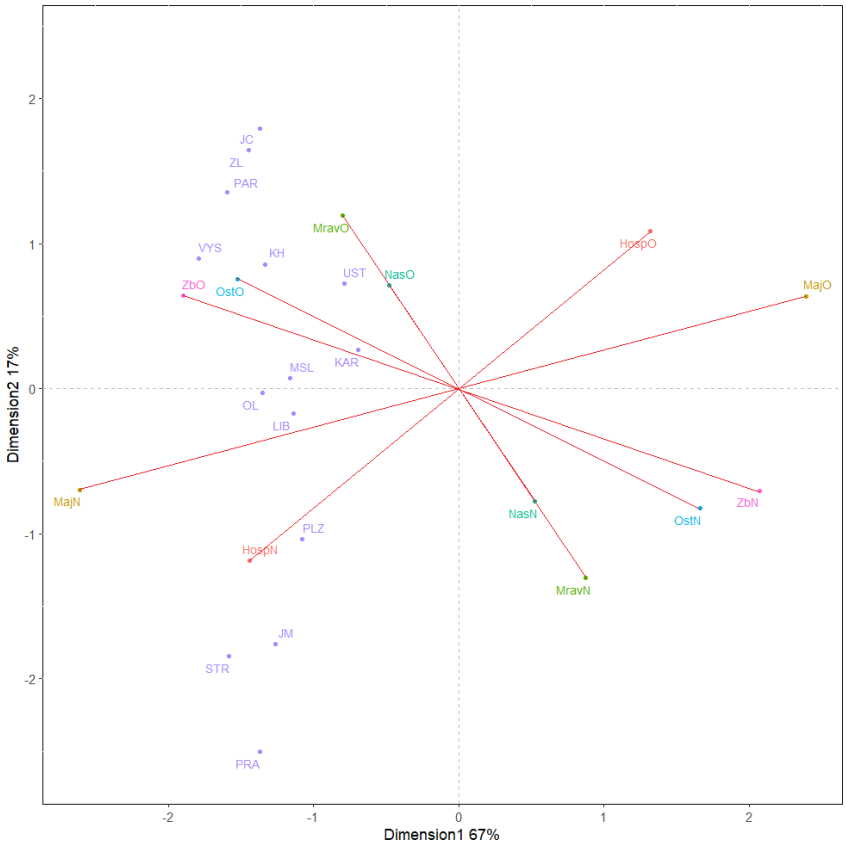
Pro zobrazení vztahů mezi proměnnými se využívá interaktivní biplot. Ten vezme jednu proměnnou jako referenční, v tomto případě to jsou kraje, kterou zobrazí společně se všemi párovými kombinacemi zbývajících dvou proměnných, tj. kombinaci objasněnosti s druhem trestné činnosti. Uvedené zobrazení tedy odpovídá interaktivním souřadnicím řádek-sloupec. Interaktivnost spočívá v tom, že v rámci příkazu `plot` můžu ještě zvolit interaktivní souřadnice řádek-vrstva, sloupec-vrstva. Interaktivní biplot naší analýzy můžeme vidět na obrázku 16.



Obrázek 16: Trojrozměrná korespondenční analýza aplikovaná na log-transformovaná data o kriminálních činech v ČR v roce 2022.

V symetrické verzi CA3 využíváme Pearsonovu trojrozměrnou $\mathbf{\Phi}^2$ statistiku, která je vhodná pro zkoumání odchýlení od trojrozměrné nezávislosti v případě symetrie proměnných. Právě zkoumání odchýlení od nezávislé struktury nám připomíná cíl interakční části u kompozičních tabulek. Zajímavé na první pohled v obrázku 16 je rozdělení činů na objasněné a neobjasněné, kdy objasněné trestné činy se nachází v horní polovině grafu a neobjasněné v dolní. Co je ale ještě zajímavější, tak je podoba s biplotem interakční části kompoziční tabulky, protože vidíme, jak jsou jednotlivé trestné činy středově souměrnné podle objasněnosti. Dále můžeme pozorovat rozložení krajů v oblastech okolo jednotlivých činů. V okolí kombinace faktorů hospodářské neobjasněné vidíme Plzeňský, Jihomoravský, Středočeský kraj a Prahu. Ústecký kraj je blízko objasněné násilné činnosti. Pardubický kraj je zase v blízkosti mravnostní kriminality objasněné.

Relevantním protějškem pro CA3 je PCA na interakční části kompoziční tabulky, přičemž příslušný biplot nalezneme v obrázku 11. Každopádně se nám zde vytrácí výsledková konzistence v rámci jednotlivých metod, což se dalo i očekávat, jelikož zde zpracováváme informaci jiným způsobem. Zatímco v případě kompozičních tabulek zachycujeme dva faktory a místo třetího pracujeme s výběrem tabulek, tak v případě CA3 máme dvě tabulky, kde každá z těchto tabulek reprezentuje jednu vrstvu, faktor objasněnosti, a dohromady takto vznikne datový kvádr. Přesto určitý efekt, který nám spojitost s interakční částí kompoziční tabulky potvrzuje, můžeme v obrázku 16 nalézt - toto propojení jsme naznačili v následujícím obrázku 17.



Obrázek 17: Trojrozměrná korespondenční analýza aplikovaná na log-transformovaná data o kriminálních činech v ČR v roce 2022 - naznačení struktury

To je velmi zajímavé a může to jen více a více poukazovat na spojitost mezi korespondenční a kompoziční analýzou, ke které jsme došli i v podkapitole 2.5.

Závěr

Cílem diplomové práce byla aplikace mnohorozměrných metod na kompoziční data vyjádřená ve vhodných souřadnicích, které povedou k jednoduché interpretaci. První kapitola nás seznámila s kompozičními daty a představili jsme si datovou sadu o struktuře kriminálních činů v České republice. Jako první jsme vzali v naší analýze data za rok 2022 v podobě, kdy jsme měli jednotlivé kraje jako pozorování a druhy trestné činnosti představovali proměnné. Na data v takovém formátu v clr koeficientech jsme aplikovali v druhé kapitole metodu hlavních komponent. Již první dvě hlavní komponenty nám vysvětlily 84.96 % informace ukrývající se v datech a po grafickém znázornění ve formě biplotu jsme tak z příslušného grafu například vyčetli, že v Libereckém kraji nedominuje žádný kriminální čin, v Karlovarském kraji dominuje hospodářská kriminalita, která společně s majetkovou kriminalitou dominuje i v Praze. Dále jsme zjistili, že pokud si převedeme původní tabulku na proporce a výslednou tabulku log-transformujeme, tak dostaneme aplikací korespondenční analýzy velmi podobné výsledky jako v případě clr biplotu.

Při průzkumu této datové sady jsme postupně přistupovali k jejímu rozvětvení, kdy jsme přidali jako další faktor objasněnost, to znamená, že jsme jednotlivé trestné činy v datové sadě za rok 2022 rozdělili na objasněno/neobjasněno. Ve čtvrté kapitole jsme aplikovali na taková data ve formě kompoziční tabulky rozklad na interakční a nezávislou část. Zjistili jsme například to, že nestabilitu v Praze způsobuje sice hospodářská a majetková kriminalita, ale v neobjasněném smyslu. Liberecký kraj nám stejně jako u PCA z druhé kapitoly zůstal uprostřed biplotu, což jen potvrzovalo to, že žádná složka v tomto kraji nedominovala. U interakční části jsme si mohli všimnout, že se formuje do hezké středově souměrné struktury, kdy trestný čin uvedený

s přívětkem objasněno byl středově souměrný se sebou samým s přívětkem neobjasněno. Další aplikovanou metodou byla trojrozměrná korespondenční analýza, u které se nám právě interakční část kompoziční tabulky nabízela jako vhodný protějšek pro srovnání. Při srovnání výsledků jsme sice nedošli ze zřejmých důvodů ke shodě, ale vypozorovali jsme to, že CA3 nám vytvořila onu středově souměrnou strukturu jako tomu tak bývá u interakční části kompoziční tabulky, což je velmi zajímavá skutečnost.

Dalším logickým rozšířením datové tabulky užité v kapitole 2 bylo přidání časového aspektu. Zatímco v druhé kapitole jsme využili pouze rok 2022, tak nyní jsme vzali období za roky 2016 až 2022. Typickou metodou pro analýzu dat tohoto typu je metoda PARAFAC, kdy jsme došli k závěrům, že i v dlouhodobém aspektu stále v Praze dominuje majetková kriminalita. Na Vysočině a v Pardubickém, Jihočeském, Královehradeckém kraji dominovala mravnostní kriminalita. Hezké zde bylo to, že díky delšímu časovému úseku se nám hezky k sobě začaly skládat kraje se společnými hranicemi. Metodu PARAFAC jsme ale využili navíc i v rámci kapitoly 4, kdy jsme se podívali na interakční část kompoziční tabulky, kterou jsme si taktéž vzali za období 2016 až 2022 a zjistili jsme v módu B, že zůstala zachována středově souměrná struktura, jež jsme vypozorovali v biplotu interakční části kompoziční tabulky.

Dle mého názoru bylo vytyčeného cíle dosaženo a v rámci této práce jsem vyzkoušel několik metod, které daly mně i čtenářům vhled do strukury kriminálních činů v ČR. Nejvíce mě zaujala kapitola o trojrozměrné korespondenční analýze. Přestože to bylo jen velmi stručné uvedení do problematiky, tak nám ukázala nějaké určité propojení s kompoziční analýzou. Právě vztah mezi kompoziční a korespondenční analýzou je velmi zajímavý a rozhodně se jedná o oblast, která by si do budoucna zasloužila hlubší bádání.

Literatura

[1] CUADRAS, Carles M. a GREENACRE, Michael. A short history of statistical association: From correlation to correspondence analysis to copulas. Journal of Multivariate Analysis. 2022, roč. 188. ISSN 0047-259X. Dostupné z: https://doi.org/10.1016/j.jmva.2021.104901

[2] DE SOUSA, Julie; HRON, Karel; FAČEVICOVÁ, Kamila a FILZMOSER, Peter. Robust principal component analysis for compositional tables. Journal of Applied Statistics. 2020, roč. 48, č. 2, s. 214-233. Dostupné z: https://doi.org/10.1080/02664763.2020.1722078.

[3] FILZMOSER, Peter; HRON, Karel a TEMPL, Matthias. Applied compositional data analysis. Cham: Springer, 2018. ISBN 978-3-319-96420-1.

[4] GREENACRE, Michael. Log-ratio analysis is a limiting case of correspondence analysis. Math Geosci. 2010, roč. 42, 129–134. Dostupné z: https://doi.org/10.1007/s11004-008-9212-2

[5] HRON, Karel. Elementy statistické analýzy kompozičních dat. Informační bulletin České statistické společnosti. 2010, roč. 21, č. 3, s. 41-48. ISSN 1210-8022.

[6] LOMBARDO, Rosaria; VELDEN, Michel van de a BEH, Eric J. Three-way correspondence analysis in R. The R Journal. 2023, roč. 15, č. 2, s. 237-262. ISSN 2073-4859. Dostupné z: http://dx.doi.org/10.32614/RJ-2023-049

[7] PAOLO, Giordani; HENK A. L., Kiers a MARIA ANTONIETTA, Del Ferraro. Three-way component analysis using the R package ThreeWay. Journal of Statistical Software. 2014, roč. 57, č. 7, s. 1-23. Dostupné z: https://doi.org/10.18637/jss.v057.i07

[8] POLICIE ČR. Kriminalita. Online. Kriminalita - Policie České republiky. 2024. Dostupné z: https://www.policie.cz/statistiky-kriminalita.aspx [cit. 2024-03-21].

[9] VARMUZA, Kurt a FILZMOSER, Peter. Introduction to multivariate statistical analysis in Chemometrics. Online. CRC Press, 2009. ISBN 9780429145049. Dostupné z: https://doi.org/10.1201/9781420059496. [cit. 2024-03-21].