UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Detekce odlehlých pozorování v kompozičních datech

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **prof. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Eliška Kremeňová**
Studijní program: B1101 Matematika
Studijní obor: Matematika a její aplikace
Forma studia: prezenční
Rok odevzdání: 2022

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Eliška Kremeňová

Název práce: Detekce odlehlých pozorování v kompozičních datech

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2022

Abstrakt: Při analýze datového souboru je častou úlohou odhalit extrémní (odlehlá) pozorování či jednotlivé odlehlé buňky v datech. U kompozičních dat se k tomu navíc přidává potřeba vzít v potaz jejich specifickou relativní povahu. Cílem bakalářské práce bude popsat, případně rozvinout metody, které jsou k tomuto účelu vhodné, a aplikovat je na data z aplikací.

Klíčová slova: detekce odlehlých pozorování, kompoziční data, extrémní hodnoty, relativní data, logpodíly

Počet stran: 44

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Eliška Kremeňová

Title: Outlier detection in compositional data

Type of thesis: Bachelor's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

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

The year of presentation: 2022

Abstract: When analyzing a dataset, often the task is to detect extreme (outlying) observations or individual outlying cells in the data. In addition for compositional data, there is a need to take into account their specific relative nature. The aim of the bachelor thesis will be to describe, respectively develop methods that are suitable for this purpose and apply them on data from applications.

Key words: outlier detection, compositional data, extreme values, relative data, logratios

Number of pages: 44

Number of appendices: 1

Language: Czech

Prohlášení

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

Olomouc ....

podpis

Obsah

Úvod **9**

1 Seznámení s odlehlými hodnotami a jejich detekcí **10**
1.1 Ilustrace odlehlých hodnot na příkladu 12

2 Kompoziční data a logpodíly **14**
2.1 Kompoziční data 14
2.2 Absolutní vs. relativní data 15
2.3 Vyjádření v logpodílech 16
2.4 Korelace mezi kompozičními složkami 17
2.5 Logpodílová metodika kompozičních dat 19
2.5.1 Souřadnicový systém kompozičních dat 20

3 Detekce prvkových odlehlých hodnot v kompozičních datech **22**
3.1 Seznámení s modelem LR-DDC 22
3.2 Popis algoritmu LR-DDC 24

4 Simulační studie **29**

5 Použití modelu LR-DDC na data z aplikací **33**
5.1 Výdaje států Evropské unie 33
5.2 Geochemická data 36
5.2.1 Odlehlé hodnoty zkoumané dle lokalit 37
5.2.2 Odlehlé hodnoty zkoumané dle typu půdy 39

Závěr **41**

Literatura **43**

Seznam obrázků

1.1 Detekce odlehlých pozorování řádkově vs. prvkově 11
1.2 Ilustrace odlehlých hodnot na simulovaných datech 12

3.1 Porovnání detekce odlehlých hodnot na původních datech (vlevo) a na logpodílech (vpravo) 23

5.1 Aplikování LR-DDC na výdaje domácností ve státech EU 34
5.2 Aplikace LR-DDC na naměřené hodnoty v Dobšicích dle typu půdy 37
5.3 Aplikace LR-DDC na naměřené hodnoty v Ivani dle typu půdy . . 38
5.4 Aplikace LR-DDC na naměřené hodnoty v Držovicích dle typu půdy 38
5.5 Aplikace DDC-LR na naměřené hodnoty hnědozemě dle lokalit . . 39
5.6 Aplikace LR-DDC na naměřené hodnoty černozemě dle lokalit . . 39
5.7 Aplikace LR-DDC na naměřené hodnoty parahnědozemě dle lokalit 40

Seznam tabulek

2.1 Výdaje domácností v EUR 16
2.2 Počet studentů doktorského studia dle oborového zaměření ve vybraných zemích Evropy, Japonska a USA 17

4.1 Výsledky pro 15 pozorování 30
4.2 Výsledky pro 30 pozorování 31
4.3 Výsledky pro 100 pozorování 31

Poděkování

Ráda bych poděkovala panu profesoru Karlu Hronovi za veškerou jeho pomoc a trpělivost při tvorbě bakalářské práce. Velice si cením jeho entusiasmu a nadšení, s kterým mě poslední čtyři semetry provázel, a taky našich konzultací, ať už prezenčních či distančních. Věnoval mi svůj čas a znalosti a provedl mě tématem detekce odlehlých pozorování. Také bych ráda poděkovala panu doktorovi Danu Šimíčkovi za poskytnutá data a za možnost pracovat na článku, kde byla moje metoda využita.

Úvod

Během získávání a zpracování dat bývá častým problémem, že se v nich vyskytují odlehlé hodnoty. Při zpracování je pak potřeba se s odlehlými hodnotami vypořádat, jelikož by nadále mohly ovlivnit budoucí analýzu či odhady parametrů modelů aplikovaných na tato data.

V tomto kontextu je důležité zjistit, kde se v datech odlehlé hodnoty nachází ať již na úrovni buněk, či celých pozorování. A to proto, abychom měli informaci, v jaké míře se v datovém souboru odlehlé hodnoty vyskytují, a zda je tedy potřeba použít vhodné metody, tzv. robustní metody, které vliv těchto hodnot na celkové výsledky analýzy potlačí.

Někdy ale odlehlé hodnoty nemusí signalizovat problém, ale mohou obsahovat užitečnou informaci týkající se dat. V případě kompozičních dat, u nichž se relativní informace nachází mezi podíly jednotlivých složek, se při podrobnější analýze odlehlých hodnot dají získat zajímavé informace obohacující následnou analýzu.

Pro tento účel jsme navrhli model LR-DDC (Logratio Deviating Data Cells) vzniklý adaptací původního modelu DDC, který takovéto odlehlé hodnoty detekuje a sám, na základě predikovaných hodnot, určí, zda je hodnota signifikantně vyšší či nižší než její zjištěné predikce.

Volba parametrů, které model využívá, je optimalizována na simulovaných datech a následně je model s těmito parametry aplikován na soubory reálných dat.

Kapitola 1

Seznámení s odlehlými hodnotami a jejich detekcí

Data se skládají z n pozorování a d proměnných, přičemž může nastat i situace, kdy máme více proměnných než pozorování, tj. $d > n$. S výhodou můžeme data uchovávat v matici o rozměrech $n \times d$, kde d odkazuje na počet sloupců a n na počet řádků v matici.

V praxi bývá častým problémem, že pozorování v našich datech jsou „obohacena“ odlehlými hodnotami, které lze definovat následovně [15]: „Ve statistice definujeme odlehlou hodnotu v pozorování jako hodnotu, která se signifikantně liší od těch zbývajících. Může být zapříčiněna variabilitou měření nebo indikovat nějakou chybu v daném pozorování.“

Ke vzniku odlehlých hodnot může tedy dojít z různých důvodů. Může jít o chybu měření nebo o určitou formu kontaminace dat. Bohužel odlehlé hodnoty mohou výrazně ovlivnit odhady parametrů zkoumaného modelu, a tedy vést k nespolehlivým výsledkům. Proto je důležité odlehlé hodnoty detekovat a dále se jimi zabývat, přičemž ve vysoké dimenzi d může být takováto detekce velmi náročná.

Při analýze dat se zpravidla za odlehlou hodnotu považuje celé jedno pozorování, přičemž je třeba předpokládat, že většina pozorování neobsahuje odlehlé hodnoty. Takováto detekce se nazývá řádková, což odkazuje na fakt, že pozorování jsou obvykle uložená v řádcích matice, kdežto sloupce odkazují na proměnné.

Toto téma je zkoumáno již od 60. let 20. století v rámci tzv. robustní statistiky.

Cílem je nalézt metodu pro detekci odlehlých hodnot, která nebude tolik citlivá na odlehlá řádková pozorování.

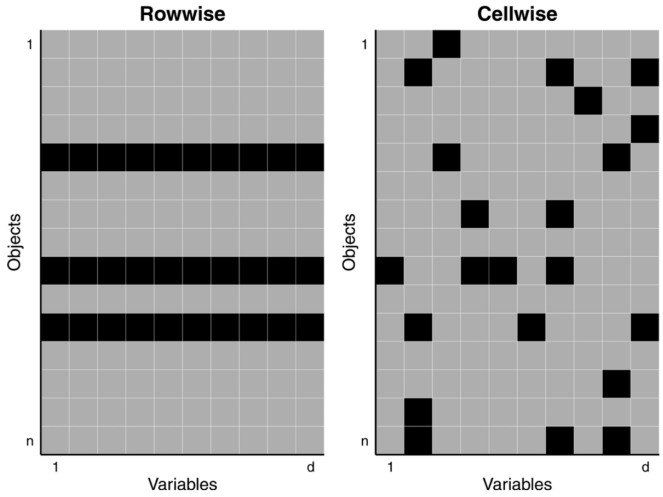
V článku [2] je popsána propagace odlehlých hodnot následujícím způsobem: „Mějme náhodně rozmístěný podíl kontaminovaných buněk, které si označíme ε. Potom očekávaný podíl kontaminovaných řádků, který je dán vztahem

$$1 - (1 - \varepsilon)^d \tag{1.1}$$

rychle přesáhne 50 % pro rostoucí ε nebo rostoucí dimenzi d.“

Proto v dnešní době, kdy často zkoumáme data s velkým počtem proměnných, již není řádková detekce postačující. Navíc pozorování často obsahují odlehlé hodnoty pouze vzhledem k jedné či k malé podmnožině proměnných. Považovat celá pozorování za odlehlá může tedy vést ke ztrátě užitečné informace.

Na obrázku 1.1 si lze všimnout, že v případě označení celého řádku nelze poznat, které hodnoty se liší od ostatních a které ne. Je potřeba potlačit vliv všech označených řádků a spokojit se jen s těmi zbývajícími. Proto je výhodnější zaměřit se na prvkovou detekci (detekci na úrovni buněk), která detekuje pouze jednotlivé odlehlé prvky, tedy neoznačí celý řádek a zachová si tak mnohem více informací o datech.



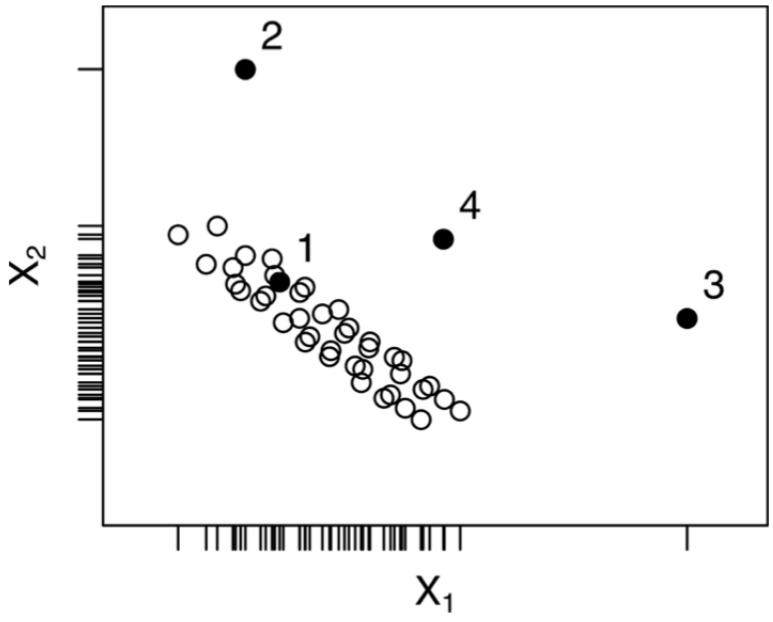
Obrázek 1.1: Detekce odlehlých pozorování řádkově vs. prvkově

1.1. Ilustrace odlehlých hodnot na příkladu

Náročnost detekce odlehlých hodnot si ukážeme na simulovaných datech o dvou proměnných x_{1i}, x_{2i} a N pozorováních, tj. $i = 1, \ldots, N$ a $d = 2$.

Na obrázku 1.2 jsou čárkami vyznačeny projekce pozorování vzhledem k proměnným x_1 a x_2. Je zřejmé, že pozorování 1 o souřadnicích x_{11} a x_{21} nijak nevybočuje z dat. Taktéž souřadnice x_{12} a x_{23} zapadají do lineárního trendu daného zbylými pozorováními. Naopak souřadnice x_{22} a x_{13} v trendu nejsou, vybočují, a proto obě příslušná pozorování označíme za odlehlá. Ale co pozorování 4? Obě jeho souřadnice x_{14} a x_{24} spadají do oblasti výskytu většiny souřadnic a tedy za odlehlost je zodpovědná specifická kombinace hodnot obou proměnných. Museli bychom označit jako odlehlé celé pozorování.

Nyní však předpokládejme, že získáme dalších pět proměnných k našim pozorováním, korelujících s proměnnými, které již máme. Pak může vyjít najevo, že pozorování x_{24} zapadá mezi nově obdržené hodnoty, zatímco x_{14} nikoli. Závěrem by bylo, že prvek x_{14} je odlehlý a není třeba označovat celý řádek.



Obrázek 1.2: Ilustrace odlehlých hodnot na simulovaných datech

Z příkladu vyplývá nezvyklá vlastnost, že s rostoucí dimenzí (počtem proměnných) může být detekce odlehlých hodnot přesnější a jednodušší. V tomto kontextu se jedná o pozorování, kde jejich „odlehlost“ nevidíme při projekci na jednotlivé proměnné, ale až v mnohorozměrném kontextu daném v tom nejjednodušším případě vztahy mezi dvojicemi proměnných.

Kapitola 2

Kompoziční data a logpodíly

2.1. Kompoziční data

Kompozičními daty rozumíme data relativní povahy, tedy proměnné nesoucí relativní informaci. Přesnější definici kompozičních dat zavádí John Aitchison v práci [1]: „Tyto proměnné nesoucí relativní informaci se považují za vnitřně propojené složky, tzv. kompozice či kompoziční složky, a jejich pozorování se obecně označují jako kompoziční data neboli komponenty."

Kompoziční proměnné jsou často generovány nějakou formou zpracování signálu v chemometrii, biologii či v environmentálních vědách. Vyjadřují se v jednotkách, jakými jsou procenta, parts per million (ppm), mg/dm^3, mmol/mol a podobně a typicky představují části nějakého celku. Příkladem mohou být vícerozměrná měření koncentrací těkavých látek či látek znečišťující ovzduší, nutriční složení potravin nebo relativní zastoupení určitého druhu. Kompoziční proměnné se taktéž objevují ve společenských vědách nebo v ekonomice. Příkladem jsou investiční portfolia, rozpočty domácností nebo rozpočty, které lze dát do souvislosti s ukazatelem produktivity či ziskovosti.

Správné statistické zpracování kompozičních dat, tedy zohlednění jejich specifické povahy, je klíčovým předpokladem jak pro získání interpretovatelných výsledků, tak k celkové validitě statistické analýzy.

Pokud se vrátíme ke zkoumání odlehlých hodnot, tak při práci s kompozicemi je potřeba vzít v potaz jejich relativní povahu, tedy že veškeré relativní informace

o kompozičních složkách jsou obsaženy v podílech mezi nimi.

2.2. Absolutní vs. relativní data

Na příkladu si uvedeme, jakou výhodu nám dává brát v potaz relativní povahu dat. Budeme zkoumat simulovaná data reprezentována tabulkou 2.1 představující měsíční výdaje tří domácností v EUR. Údaje obsahují pouze některé výdaje a další nejsou uvedeny. Proměnné, jako například zdraví či oblečení, které mohou rovněž tvořit významné položky měsíčního rozpočtu, nejsou k dispozici. Nemáme tedy k dispozici všechny proměnné, které by tvořily celkové výdaje, což ovšem nevadí, protože tak jako tak je součet výdajů (z hlediska jejich relativní struktury) pro dané pozorování irelevantní. Dostáváme tak bez ztráty (relativní) informace odpovídající procentuální údaje, které jsou v tomto případě pro všechna tři pozorování (domácnosti) stejné.

Jak původní hodnoty, tak jejich procentuální vyjádření znamenají příspěvek jednotlivých položek k celkovým výdajům. Nicméně absolutní údaje uvedené v EUR ukazují zřetelný rozdíl ve výši jednotlivých položek ve všech pozorováních (domácnostech).

Relativní informace by nyní mohla být reprezentována procentuálními údaji, tedy čtyřmi hodnotami pro každé pozorování. Na druhou stranu se jedná pouze o jednu možnou reprezentaci relativní informace. Z tohoto hlediska se jeví jako výhodnější zaměřit se na analýzu podílů mezi jednotlivými složkami. Například $\frac{1710}{570} = \frac{540}{180} = \frac{900}{300} = 3$, tedy všechny tři domácnosti utratí třikrát více za bydlení než za dopravu. Celkem můžeme pro každou domácnost vyjádřit $\binom{4}{2} = 6$ podílů. Z příkladu vyplývá, že podíly obsahují mnohem podrobnější informaci než pouhé procentuální údaje a zachovávají si stejnou hodnotu i pro jinak reprezentovaná data (po jejich vynásobení zvolenou kladnou konstantou), například pokud vezmeme jinou měnu než EUR.

Poznamenejme, že pro nás mohou mít větší váhu absolutní informace o datech. Například pokud by cílem analýzy bylo zkoumání úrovně bohatství v domácnostech, má smysl pro další statistickou analýzu uvažovat přímo data

v EUR.

Typ	Domácnost	*Bydlení*	*Potraviny*	*Doprava*	*Komunikace*	Součet
Absolutní	1	1710	950	570	570	3800
informace	2	540	300	180	180	1200
v EUR	3	900	500	300	300	2000
Relativní	1	45	25	15	15	100
informace	2	45	25	15	15	100
v EUR	3	45	25	15	15	100

Tabulka 2.1: Výdaje domácností v EUR

2.3. Vyjádření v logpodílech

Na příkladu jsme ukázali, že relativní data reprezentována podíly mezi jednotlivými složkami nám dávají relevantnější informace o datech, než pokud bychom je zkoumali pouze z hlediska jejich absolutní hodnoty.

Z důvodu asymetrického chování podílů ale není ani z matematického, ani z interpretačního hlediska vhodné pracovat s původními podíly mezi jednotlivými proměnnými. Podíly mohou nabývat hodnot z intervalu $(0, +\infty)$, kde 1 znamená dokonalou rovnováhu mezi oběma proměnnými, jako je tomu u dopravy a komunikace v tabulce 2.1. Interval $(1, +\infty)$ tak odpovídá dominující proměnné v čitateli oproti proměnné ve jmenovateli. Naopak pokud vezmeme interval $(0, 1)$, je proměnná ve jmenovateli dominantní, jako je tomu u dopravy a bydlení, kde dostáváme podíl ve tvaru $\frac{1}{3}$.

Pro symetrizaci interpretace podílů je první volbou použít logaritmickou transformaci podílů, a to z následujících důvodů. Oborem hodnot logpodílů (logaritmů podílů) je celá reálná přímka od $-\infty$ do $+\infty$, kde rovnováhu proměnných představuje 0. Pro obě možnosti, kdy jedna proměnná dominuje nad druhou, je vyhrazena polopřímka $(-\infty, 0)$ a $(0, +\infty)$, navíc logaritmus podílu a jeho reciproké hodnoty se liší pouze znaménkem. S logaritmy se také lépe pracuje z matematického hlediska, protože logaritmus podílu dvou proměnných lze vyjádřit jako jejich rozdíl. Jsou-li známy všechny logpodíly mezi složkami, lze odvodit libovolnou reprezentaci původních kompozičních složek a naopak.

2.4. Korelace mezi kompozičními složkami

Problém korelační analýzy mezi složkami kompozičních dat při jejich různých reprezentacích lze ukázat na příkladu relativní struktury oborového zaměření studentů doktorského studia ve vybraných zemích Evropy, Japonska a USA, který prezentuje tabulka 2.2. Údaje v této tabulce jsou volně dostupné v Eurostatu [6], počty studentů jsou uváděny pro různé studijní skupiny.

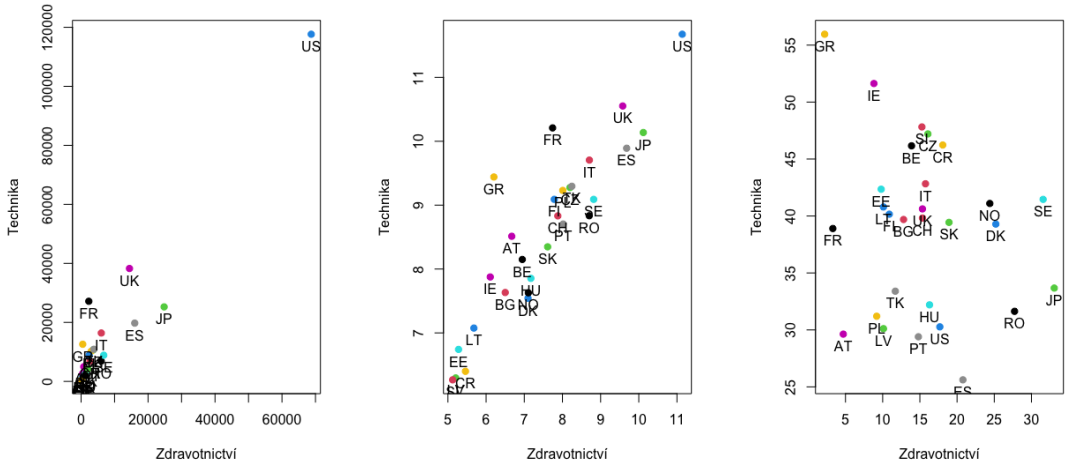
	Celkem	Technika	Soc-eko-právo	Humanitní	Zdravotnické	Zemědělské
BE	7500	3462	1469	997	1041	532
BG	5200	2064	1102	1170	666	198
CZ	2260	10668	3748	3518	3633	1035
DK	4800	1886	614	696	1210	394
EE	2000	847	424	420	196	112
IE	5100	2633	787	1124	450	107
GR	22500	12590	3941	5090	495	383
ES	77100	19751	20704	18885	16026	1733
FR	69800	27152	21429	18846	2303	70
IT	38300	16403	7621	5803	6035	2437
LV	1800	542	603	434	182	40
LT	2900	1183	916	400	293	107
HU	8000	2576	1648	1992	1304	480
AT	16800	4978	6374	4103	790	555
PL	32700	10202	7881	9974	3008	1635
PT	20500	6027	6191	4879	3034	369
RO	21700	6864	3801	3323	6017	1694
SI	1100	526	174	189	168	43
SK	10700	4220	2121	1971	2024	364
FI	22100	8875	4990	5365	2406	464
SE	21400	8872	2651	2694	6756	428
UK	94200	38266	19747	20408	14456	1323
CR	1300	601	94	286	235	84
TK	32600	10888	7922	7335	3814	2641
NO	5000	2055	870	635	1220	220
CH	17200	6849	4537	2691	2640	483
JP	75000	25255	10102	10408	24796	4439
US	388700	117658	104456	94748	68731	3106

Tabulka 2.2: Počet studentů doktorského studia dle oborového zaměření ve vybraných zemích Evropy, Japonska a USA

Korelace dvou proměnných z uvedené tabulky, například studentů technického a zdravotnického zaměření, z vybraných zemí Evropy, Japonska a USA je znázorněna na obrázku 2.1.

Zejména kvůli vysokým absolutním hodnotám studentů v USA vykazuje klasická míra korelace prostřednictvím Pearsonova korelačního koeficientu vysokou míru kladné korelace, což ukazuje obrázek 2.1a. Tato odlehlá hodnota bude velmi pravděpodobně dominovat i v ostatních statistických metodách a povede ke zkresleným výsledkům. Z důvodu relativního měřítka dat provedeme jejich logaritmickou transformaci. Získaný obrázek 2.1b ukazuje, že data vykazují pozitivní lineární trend, či dokonce, že rozdělení dvou proměnných se zdá být blízké dvourozměrnému normálnímu rozdělení.

Původní data lze také převést na procenta, tj. vydělit hodnoty všech proměnných jejich součtem a vynásobit 100. Výsledek pro obě uvažované proměnné je uveden na obrázku 2.1c. Dříve pozorovaná silně pozitivní korelace v tomto případě zaniká.



(a) absolutní počet studentů (b) počet studentů po logaritmické transformaci (c) procentuální počet studentů

Obrázek 2.1: Korelace proměnných vyjadřujících technické a zdravotnické zaměření studentů doktorského studia z vybraných zemí Evropy, Japonska a USA

Ze tří různých výsledků není jasné, jak s nimi naložit. K tomu, abychom mohli výsledky „sjednotit“, využijeme opět logpodílů, jelikož logpodíly poskytují stejné výsledky nezávisle na tom, zda jsou počítány z absolutních hodnot nebo

z procent. Logpodíly jsou tedy klíčem k analýze dat tam, kde je důležitá relativní informace.

2.5. Logpodílová metodika kompozičních dat

Při detekci odlehlých hodnot v kompozičních datech se chceme zaměřit na logpodíly mezi jednotlivými složkami, které zohledňují relativní povahu dat a zároveň sílu vztahu mezi nimi, vyjádřenou v tomto případě pomocí jejich proporcionality.

Jediný problém týkající se kompozičních dat, který je třeba vyřešit pro správnou statistickou analýzu, je nalezení vhodného souřadnicového systému, ve kterém budeme mnohorozměrnou informaci obsaženou v kompozičních datech reprezentovat. Pojďme se na předchozí úvahy podívat trochu formálněji.

Definice 2.1. *D-složkovou kompozici definujeme jako náhodný vektor* $\boldsymbol{X} = (X_1, X_2, \ldots, X_D)'$ *s kladnými komponentami (kompozičními složkami) obsahující relativní informaci.*

V souladu s touto definicí chápeme kompoziční data jako vícerozměrná pozorování, u nichž je relevantní informace obsažená v podílech mezi komponentami.

Při práci s kompozicemi nepracujeme s podíly, ale s logpodíly, kde pro dvojici složek X_j a X_k platí, že

$$\ln \frac{X_j}{X_k} = -\ln \frac{X_k}{X_j}. \tag{2.1}$$

Z tohoto vztahu vyplývá, že pro účely detekce odlehlých prvkových hodnot je třeba uvažovat pouze $\frac{D(D-1)}{2}$ logpodílů místo D^2 podílů.

Mějme pozorování $\mathbf{x} = (x_1, x_2, \ldots, x_N)'$ náhodné D-složkové kompozice $\mathbf{X} = (X_1, X_2, \ldots, X_D)'$. Je zřejmé, že pokud „forma kontaminace“ generuje odlehlou hodnotu ve složce x_j, ovlivní to všechny logpodíly, v nichž je x_j obsažena. Na druhou stranu kontaminace dat, která generuje pouze jeden odlehlý logpodíl $\ln \frac{x_j}{x_k}$, mohla vzniknout ze dvou odlehlých komponent, konkrétně x_j a x_k. Tyto předpoklady je třeba vzít v úvahu při vývoji metody detekce odlehlých prvků v kontextu analýzy kompozičních dat.

2.5.1. Souřadnicový systém kompozičních dat

Kompoziční data se formálně řídí takzvanou Aitchisonovou geometrií o dimenzi $D-1$ na výběrovém prostoru kompozičních dat, tzn. na třídách ekvivalence proporcionálních vektorů. Z geometrického hlediska je tedy zkonstruován nový souřadnicový systém vzhledem k Aitchisonově geometrii. Pro náš účel jsou výhodné tzv. izometrické logpodílové souřadnice, které umožňují vyjádřit kompozice v ortonormálním souřadnicovém systému.

Reprezentace kompozic pomocí izometrických logpodílových souřadnic umožňuje zachovat vzdálenosti mezi body z původní Aitchisonovy geometrie v reálné euklidovské geometrii dimenze $\mathbb{R}^{D-1}$. Konkrétně volíme tzv. pivotové souřadnice, v nichž je zvýrazněna role samostatné komponenty vůči ostatním. Tímto způsobem pro D-složkovou kompozici $\mathbf{X} = (X_1, ..., X_D)'$ získáme reálný vektor $\mathbf{Z} = (Z_1, ..., Z_{D-1})'$ se složkami

$$Z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{X_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} X_k}}, \qquad j = 1, \ldots, D-1. \tag{2.2}$$

Veškerá relativní informace o X_1 vzhledem k (geometrickému) průměru zbývajících složek je tedy obsažena v první pivotové souřadnici Z_1, přičemž Z_1 lze vyjádřit následovně

$$Z_1 = \frac{1}{\sqrt{D(D-1)}} \left[\ln \frac{X_1}{X_2} + \cdots + \ln \frac{X_1}{X_D} \right], \tag{2.3}$$

tj. jako normovaný součet všech logpodílů s komponentou X_1 v čitateli. Permutováním složek kompozice $\mathbf{X} = (X_1, \ldots, X_D)'$ tak, že na první pozici umístíme pokaždé jinou složku, získáme D různých ortonormálních souřadnicových systémů, které jsou navzájem ortogonálními rotacemi. Každá z nich zdůrazňuje roli příslušné komponenty na první pozici.

Výraz v (2.2) pak zobecníme tak, že označíme $\mathbf{X}^{(l)} = (X_1^{(l)}, \ldots, X_D^{(l)})' = (X_l, X_1, X_2, \ldots, X_{l-1}, X_{l+1}, \ldots, X_D)'$, $\mathbf{Z} = (Z_1^{(l)}, \ldots, Z_{D-1}^{(l)})$, čímž dostáváme

$$Z_j^{(l)} = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{X_j^{(l)}}{\sqrt[D-j]{\prod_{k=j+1}^{D} X_k^{(l)}}}, \tag{2.4}$$

kde $j = 1, \ldots, D-1, \quad l = 1, \ldots, D$.

Veškerá relativní informace o libovolné složce X_l, $l = 1, \ldots, D$ (tj. příslušné logpodíly s touto složkou) je obsažena v odpovídající první pivotové souřadnici $Z_1^{(l)}$. Pro zpětnou transformaci $\mathbf{Z} = (Z_1^{(l)}, \ldots, Z_{D-1}^{(l)})$ na $\mathbf{X}^{(l)} = (X_1^{(l)}, \ldots, X_D^{(l)})'$ použijeme inverzní zobrazení,

$$
\begin{aligned}
X_1^{(l)} &= \exp\left(\sqrt{\frac{D-1}{D}} Z_1^{(l)}\right), \\
X_j^{(l)} &= \exp\left(-\sum_{k=1}^{j-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} Z_k^{(l)} + \sqrt{\frac{D-j}{D-j+1}} Z_j^{(l)}\right), \\
&\qquad j = 2, \ldots, D-1, \\
X_D^{(l)} &= \exp\left(-\sum_{k=1}^{D-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} Z_k^{(l)}\right).
\end{aligned}
\tag{2.5}
$$

Toto inverzní zobrazení umožňuje vyjádřit výstupy ze statistického zpracování dat v reálném prostoru zpět v původním výběrovém prostoru kompozičních dat s využitím libovolné proporcionální reprezentace v rámci třídy ekvivalence podle zvoleného součtu složek.

Kapitola 3

Detekce prvkových odlehlých hodnot v kompozičních datech

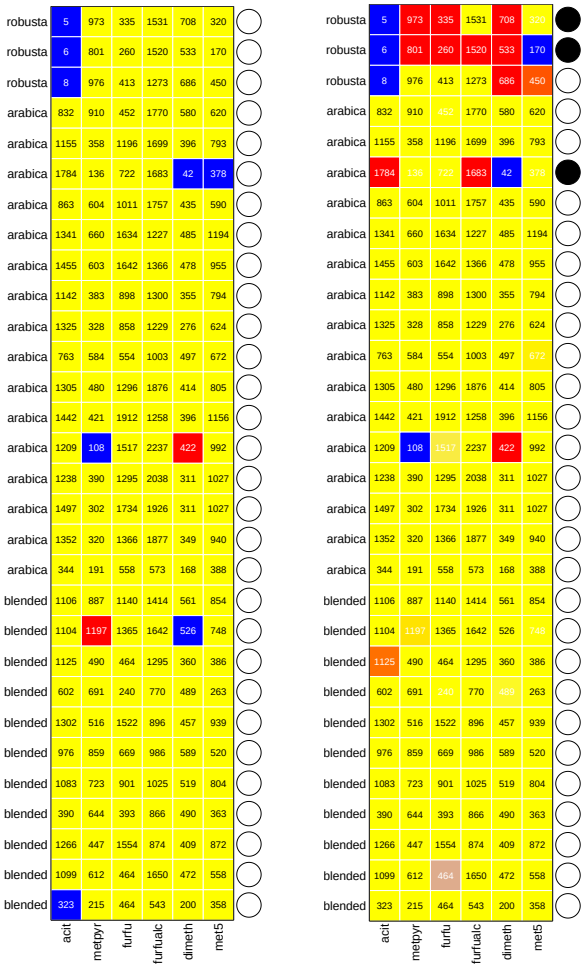
3.1. Seznámení s modelem LR-DDC

Pro detekci prvkových i řádkových odlehlých hodnot, která bere v potaz korelaci mezi proměnnými a v případě kompozičních dat také relativní informaci obsaženou v jejich logpodílech, byla vyvinuta metoda Logratio Deviating Data Cells, zkráceně LR-DDC, která vznikla adaptací metody DDC z článku [8].

Oproti existujícím metodám detekce odlehlých hodnot metoda DDC neklade omezení na to, kolik procent řádků musí být bez odlehlých hodnot. Z předpokladu plyne, že alespoň 50 %. Další výhodou této metody je, že se dokáže vypořádat s vysokou dimenzí d a že poskytuje predikci odlehlých prvků. Navíc je schopná na základě predikce doplnit chybějící data.

LR-DDC má na vstupu kompoziční data, neboli kladná data relativní povahy. Z těchto dat v prvé řadě vytvoří $\frac{D(D-1)}{2}$ logpodílů namísto D^2 podílů, jelikož navzájem inverzní logpodíly poskytují stejnou informaci až na znaménko, tedy platí $\ln \frac{X_j}{X_k} = -\ln \frac{X_k}{X_j}$. Po celou dobu pak pracujeme s logpodíly, na něž aplikujeme původní algoritmus DDC pro detekci odlehlých hodnot, který je jako knihovna k dispozici ve statistickém softwaru R [3].

Modifikovaný algoritmus DDC (LR-DDC) tedy místo původních dat pracuje s jejich logpodíly, protože uvažovat korelaci mezi původními proměnnými by v tomto případě nedávalo smysl (viz kapitola 2.4). Rozdíl mezi původním algoritmem a jeho modifikací vystihuje obrázek 3.1, který srovnává aplikování algoritmu DDC na původních datech a na logpodílech.



(a) DDC (b) DDC-LR

Obrázek 3.1: Porovnání detekce odlehlých hodnot na původních datech (vlevo) a na logpodílech (vpravo)

Data, na která byly oba modely použity, jsou volně k dispozici ve statistickém softwaru R v knihovně `robCompositions` [3]. Konkrétně se jedná o soubor dat s názvem `coffee` popisující pomocí 6 proměnných 30 různých káv, které se dělí na kávy typu arabica, robusta či na kávy složené smíchaným těchto dvou typů v různém poměru. Proměnné popisují výši acidity, metpyru, furfuralu, furfuryl alkoholu, 2,6 dimethylpyrazinu a 5-methylfurfuralu v jednotlivých kávách.

LR-DDC je schopen odhalit více odlehlých hodnot tím, že bere v potaz korelaci mezi proměnnými (v tomto případě mezi logpodíly). Zachycuje například odlišnost mezi kávou typu arabica a robusta. Nízká acidita u kávy typu robusta je zaznamenána jak metodou DDC, tak metodou LR-DDC, ale pouze LR-DDC

zachycuje také výrazně vysoké hodnoty ve zbývajích proměnných. DDC tyto hodnoty nezaznačil, jelikož samy o sobě nijak nevybočují. Až pokud vezmeme v potaz proměnnou acidity, tj. kyselost kávy, zjistíme, že vzhledem k tak nízkým hodnotám by i zbylé hodnoty měly být nízké.

Aplikací původní metody DDC na podíly po logaritmické transformaci nalezneme logpodíly vybočující vzhledem k celkové datové struktuře a určíme také, které složky v daném pozorování (kompozici) se signifikantně odlišují od zbývajících. Přesnější popis algoritmu LR-DDC bude uveden v následující kapitole.

Metoda LR-DDC je tedy schopna detekovat jak řádkové, tak prvkové odlehlé hodnoty, což je výhodné, jelikož se v datech mohou vyskytovat oba tyto typy odlehlých hodnot současně.

3.2. Popis algoritmu LR-DDC

Algoritmus na začátku zpracuje kompoziční data tak, že vytvoří $\frac{D(D-1)}{2}$ logpodílů (sloupců), přičemž tvoří logpodíly postupně vzhledem k jedné proměnné vůči všem zbývajícím.

Na takto vytvořená data aplikuje původní DDC algoritmus s konkrétní volbou parametrů $alpha = 0.95$ a $pOutLR = 0.3$. Parametr $pOutLR$ určuje, kolik logpodílů je třeba označit za odlehlé, aby daný prvek v původních datech byl také označen za odlehlý. Parametr $alpha$ určuje hranici pro označení hodnoty v modelu za odlehlou. Čím je hodnota vyšší, tím méně prvků model označí, jelikož požadujeme větší jistotu v označení prvku za odlehlý.

Aplikováním DDC na logpodíly obdržíme prostřednictvím funkce `DDC` z balíčku `robCompositions` statistického softwaru R [3] hned několik výstupů. Mezi nejzajímavější výstupy patří

1. matice X_{imp} obsahující naše původní data, jejichž NA hodnoty jsou nahrazeny predikovanými hodnotami,

2. seznam odlehlých řádků a jednotlivých odlehlých prvků (logpodílů) v rámci pozorování,

3. matice X_{est} s predikcemi hodnot matice X,

4. matice *stdResid* obsahující residua spočtená jako rozdíl původních hodnot od jejich predikcí standardizovaných podle sloupce.

Z výše zmíněných výstupů použijeme seznam odlehlých prvků k určení odlehlých složek v původních kompozičních datech. Ty nalezneme tak, že postupně v rámci daného pozorování sdružíme odlehlé hodnoty $D-1$ logpodílů vzhledem k jednotlivým proměnným $1, \ldots, D$. Sečteme, kolikrát se daná složka vyskytovala v odlehlých logpodílech, a pokud byla obsažena v alespoň $pOutLR * (D-1)$, potom ji v rámci tohoto pozorování označíme za odlehlou.

***Příklad**: Pokud máme $D = 6$ proměnných, dostáváme ke každé kompoziční složce $D - 1 = 5$ logpodílů, kde zkoumaná složka je buď v čitateli, nebo jmenovateli logpodílu. Zvolme $pOutLR = 0.4$, tedy chceme, aby alespoň 40 % logpodílů označilo složku za odlehlou. Dostáváme, že pokud složka vzhledem k proměnné x_i, $i = 1, \ldots D$, byla nalezena alespoň ve $0.4 * 5 = 2$ logpodílech, označíme ji za odlehlou.*

Pokud je navíc v některém pozorování více než $pOutROW * D$ odlehlých složek (prvků datové matice), označíme celý řádek za odlehlý. Takto získáme odlehlé řádky, které sjednotíme se seznamem odlehlých řádků získaným algoritmem DDC.

Nakonec vytvoříme vlastní matici standardizovaných residuí, která nám určí, jak moc se naše odlehlá hodnota od predikované hodnoty odlišuje. Pokud je mnohem vyšší, než by měla být, označí model hodnotu oranžově až červeně, pokud je nižší než predikovaná hodnota, tak fialově až modře (viz příklad tabulky na obrázku 3.1). Matici residuí tvoříme na základě standardizovaných residuí v jednotlivých logpodílech, přičemž k sobě sdružíme logpodíly se zkoumanou proměnnou v čitateli, z nichž pak vezmeme průměr (pro jednotlivá pozorování). Pokud se zkoumaná proměnná vyskytuje ve jmenovateli logpodílu, vezmeme opačnou hodnotu jejího rezidua, jelikož platí $\ln \frac{x_i}{x_j} = -\ln \frac{x_j}{x_i}$.

Podrobnější popis algoritmu DDC

Podívejme se na algoritmus DDC podrobněji a budeme přitom opět čerpat z článku DDC [8], kde se také nachází detailní vyjádření všech následně zmíněných robustních odhadů. Algoritmus DDC nejprve standardizuje vstupní data tak, že

pro každý sloupec j vstupní matice X vypočítá

$$m_j = robLoc_i(x_{ij}) \qquad \text{a} \qquad s_j = robScale_i(x_{ij} - m_j), \tag{3.1}$$

kde *robLoc* je robustní odhad polohy a *robScale* je robustní odhad měřítka. Dále standardizuje matici X na matici Z pomocí následujícího postupu

$$z_{ij} = \frac{x_{ij} - m_j}{s_j} \tag{3.2}$$

a definuje novou matici U, která slouží k použití jednorozměrné detekce odlehlých hodnot pro všechny proměnné

$$u_{ij} = \begin{cases} z_{ij} & \text{if } |z_{ij}| \leq c \\ NA & \text{if } |z_{ij}| > c. \end{cases} \tag{3.3}$$

Díky standardizaci v (3.2) slouží vztah (3.3) k detekci odlehlých hodnot v jednotlivých sloupcích. Mezní hodnota c se bere jako

$$c = \sqrt{\chi^2_{1,alpha}}, \tag{3.4}$$

kde $\sqrt{\chi^2_{1,alpha}}$ je *alpha*-tý kvantil chí-kvadrát rozdělení s jedním stupněm volnosti, přičemž pravděpodobnost *alpha* je standardně 99 %, takže za předpokladu normálního rozdělení dat je označeno pouze 1 % prvků. Na základě simulační studie však pro náš model volíme $alpha = 95$ %.

Pro libovolné dvě proměnné $h \neq j$ následně vypočítáme jejich korelaci

$$cor_{jh} = robCorr_i(u_{ij}, u_{ih}), \tag{3.5}$$

kde *robCorr* je robustní míra korelace. Výpočet probíhá nad všemi i, pro které u_{ij} ani u_{jh} není NA. Dále hledáme vztah mezi proměnnými j a h pouze tehdy, když

$$|cor_{jh}| \geq corrlim, \tag{3.6}$$

v němž je *corrlim* implicitně nastavena na 0.5. Proměnné j, které splňují (3.6) pro některé $h \neq j$, se budou nazývat korelované a budou o sobě navzájem obsahovat užitečné informace. Ostatní se nazývají samostatné proměnné (nekorelované). Pro dvojice (j, h) splňující (3.6) také vypočítáme

$$b_{jh} = robSlope_i(u_{ij}|u_{jh}), \tag{3.7}$$

kde *robSlope* počítá směrnice robustní regresní přímky bez absolutního členu, která predikuje proměnnou j na základě proměnné h. Tyto směrnice slouží k predikcím pro korelované proměnné.

Na základě směrnic regresní přímky algoritmus předpoví hodnoty $\hat{z}_{ij}$ pro všechny prvky. Pro každou proměnnou j uvažuje množinu H_j sestávající ze všech proměnných h splňujících (3.6), včetně samotné proměnné j. Pro $\forall i = 1, \ldots, n$ pak stanoví

$$\hat{z}_{ij} = G(\{b_{jh} u_{ih} : h \text{ in } H_j\}), \tag{3.8}$$

kde G je kombinační pravidlo aplikované na tato čísla, které volíme jako vážený průměr s váhami $w_{jh} = |cor_{jh}|$ po vynechání NA hodnot.

Predikce daná vztahem (3.8) má tendenci zmenšovat variabilitu proměnných, což je nežádoucí. Za tímto účelem nahradíme hodnoty $\hat{z}_{ij}$ hodnotami $a_j \hat{z}_{ij}$ pro všechna i a j, kde

$$a_j := robSlope_i(z_{i'j} | \hat{z}_{i'j}) \tag{3.9}$$

pochází z regrese pozorovaného $z_{.j}$ na (zmenšeném) predikovaném $\hat{z}_{.j}$. V předešlých krocích jsou pro všechny prvky vypočítané predikované hodnoty $\hat{z}_{ij}$. Dále algoritmus vypočítá standardizovaná rezidua buněk

$$r_{ij} = \frac{z_{ij} - \hat{z}_{ij}}{robScale_{i'}(z_{i'j} - \hat{z}_{i'j})}. \tag{3.10}$$

V každém sloupci j pak označí všechny buňky s $|r_{ij}| > c$ jako odlehlé, kde c bylo dáno v (3.4). Sestaví také matici Z_{imp}, která se rovná Z s tím rozdílem, že odlehlé buňky z_{ij} a NA nahradí jejich predikovanými hodnotami $\hat{z}_{ij}$. Neoznačené buňky zůstávají v původní podobě. Při rozhodování, zda označit řádek i jako odlehlý, se využívá toho, že za předpokladu mnohorozměrného normálního rozdělení vstupních dat je rozdělení r_{ij} blízké normovanému normálnímu rozdělení, takže distribuční funkce r_{ij}^2 je přibližně rovna distribuční funkci χ_1^2. To vede ke kritériu

$$T_i = \frac{1}{d} \sum_{j=1}^{d} F(r_{ij}^2). \tag{3.11}$$

Poté algoritmus standardizuje T_i podle (3.2) a označí řádky i, pro které standardizované T_i přesahuje mezní hodnotu c z (3.4). Nakonec změní imputovanou

matici Z_{imp} na imputovanou matici X_{imp} inverzním zobrazením ke standardizaci v (3.2).

Největší výhodou DDC je predikce jednotlivých buněk. Tu lze charakterizovat jako lokálně lineární fit, kde „lokálně“ není myšleno v obvyklém smyslu (euklidovské vzdálenosti), ale místo toho se vztahuje k prostoru proměnných vybavených určitou metrikou založenou na korelaci.

Metoda DDC má přirozené afinně ekvivarianční vlastnosti. Pokud ke všem hodnotám ve sloupci matice X přičteme konstantu nebo některý sloupec vynásobíme nenulovým faktorem nebo změníme pořadí řádků či sloupců, výsledek se změní podle očekávání.

Kapitola 4

Simulační studie

Před aplikací modelu LR-DDC na reálná data je nejprve nutné optimalizovat nejvhodnější parametry. Těmito parametry byly *alpha* a *pOutLR*. Parametr *pOutLR* určuje, kolik logpodílů je třeba označit za odlehlé, aby daný prvek v původních datech byl označen také za odlehlý. Parametr *alpha* určuje hranici pro označení hodnoty v modelu za odlehlou. Čím je hodnota vyšší, tím méně prvků model označí, jelikož požadujeme větší jistotu v označení prvku za odlehlý.

Simulaci provádíme tak, že si nejprve vygenerujeme pozitivně definitní matici o velikosti $d \times d$, kterou si označíme symbolem Σ. Hodnota d je zvolena 17, a to podle počtu proměnných v následně analyzovaných geochemických datech. Volíme jednoduchou strukturu varianční matice jako pozitivně definitní matici s hodnotou 1 na hlavní diagonále a mimo ni volíme postupně hodnoty 0.2, 0.5, 0.8 a 0.95. Poté si nagenerujeme data z mnohorozměrného normálního rozdělení s nulovou střední hodnotou a varianční maticí Σ.

Na základě těchto reálných dat můžeme použít funkci `pivotCoordInv` z balíčku `robCompositions` [3], čímž vytvoříme pomocí inverzního zobrazení (2.5) kompoziční data. V této fázi si nasimulujeme odlehlé hodnoty a budeme požadovat, aby je model LR-DDC správně detekoval. Vybereme náhodně 10 % prvků z kompozičních dat, které jsme vytvořili, a přenásobíme je konstantou c, kterou volíme 10. Na data nyní aplikujeme model LR-DDC, čímž obdržíme seznam odlehlých prvků a řádků. Z těchto dat určíme, jak dobře si model vedl. Jelikož máme data s převážně standardními (neodlehlými) hodnotami, model správně označí velké množství hodnot právě za neodlehlé. Abychom zamezili zkreslení výsledků, použijeme pro zhodnocení modelu *F-score* definované následovně [4]:

$$F\text{-}score = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{4.1}$$

kde

$$Precision = \frac{True\_positives}{True\_positives + False\_positives}, \tag{4.2}$$

$$Recall = \frac{True\_positives}{True\_positives + False\_negatives}. \tag{4.3}$$

True_positives jsou modelem správně odhalené odlehlé hodnoty, *False_positives* jsou hodnoty, které model označil za odlehlé, přestože nijak nevybočovaly, a *False_negatives* jsou hodnoty, které byly odlehlé, ale nebyly modelem označené. *F-score* tedy může nabývat hodnot z intervalu $[0, 1]$, kde 1 znamená, že model správně detekoval všechny odlehlé hodnoty bez jakékoliv chyby, a naopak 0 znamená, že neoznačil správně jediný odlehlý prvek.

Při pevně zvoleném počtu simulací 300 a dané velikosti varianční matice 17×17 jsme zkoumali hodnoty parametrů *alpha* a $pOutLR$ pro různý počet pozorování a pro různé mimodiagonální hodnoty varianční matice. Výsledné hodnoty *F-score* jsou uvedeny v následujících tabulkách.

Parametry	**Max**	**Průměr**	**Medián**	**Min**	**Σ hodnoty**
$alpha = 0.99, pOutLR = 0.4$	0.540	0.289	0.279	0.053	0.2
	0.622	0.366	0.368	0.098	0.5
	0.750	0.531	0.542	0.239	0.8
	0.947	0.742	0.755	0.390	0.95
$alpha = 0.99, pOutLR = 0.3$	0.622	0.350	0.348	0.094	0.2
	0.642	0.405	0.407	0.171	0.5
	0.764	0.569	0.577	0.304	0.8
	0.915	0.726	0.738	0.390	0.95
$alpha = 0.95, pOutLR = 0.4$	0.667	0.362	0.369	0.105	0.2
	0.625	0.425	0.425	0.209	0.5
	0.828	0.557	0.557	0.308	0.8
	0.915	0.675	0.682	0.309	0.95
$alpha = 0.95, pOutLR = 0.3$	0.625	0.392	0.393	0.211	0.2
	0.635	0.430	0.437	0.191	0.5
	0.754	0.554	0.559	0.310	0.8
	0.857	0.633	0.639	0.318	0.95

Tabulka 4.1: Výsledky pro 15 pozorování

Parametry	**Max**	**Průměr**	**Medián**	**Min**	**Σ hodnoty**
$alpha = 0.99, pOutLR = 0.4$	0.404	0.232	0.235	0.063	0.2
	0.457	0.292	0.290	0.129	0.5
	0.674	0.490	0.486	0.319	0.8
	0.904	0.783	0.790	0.599	0.95
$alpha = 0.99, pOutLR = 0.3$	0.468	0.293	0.289	0.145	0.2
	0.538	0.365	0.373	0.143	0.5
	0.752	0.579	0.589	0.369	0.8
	0.891	0.784	0.794	0.610	0.95
$alpha = 0.95, pOutLR = 0.4$	0.512	0.369	0.373	0.202	0.2
	0.587	0.437	0.434	0.228	0.5
	0.845	0.619	0.614	0.376	0.8
	0.893	0.728	0.734	0.480	0.95
$alpha = 0.95, pOutLR = 0.3$	0.545	0.403	0.407	0.220	0.2
	0.660	0.479	0.479	0.313	0.5
	0.850	0.643	0.652	0.416	0.8
	0.841	0.694	0.700	0.520	0.95

Tabulka 4.2: Výsledky pro 30 pozorování

Parametry	**Max**	**Průměr**	**Medián**	**Min**	**Σ hodnoty**
$alpha = 0.99, pOutLR = 0.4$	0.289	0.176	0.176	0.074	0.2
	0.306	0.212	0.212	0.112	0.5
	0.584	0.461	0.459	0.304	0.8
	0.865	0.808	0.811	0.700	0.95
$alpha = 0.99, pOutLR = 0.3$	0.337	0.231	0.231	0.135	0.2
	0.402	0.282	0.283	0.164	0.5
	0.673	0.552	0.553	0.419	0.8
	0.851	0.798	0.799	0.720	0.95
$alpha = 0.95, pOutLR = 0.4$	0.456	0.377	0.378	0.280	0.2
	0.555	0.460	0.460	0.375	0.5
	0.779	0.683	0.680	0.540	0.8
	0.835	0.745	0.744	0.627	0.95
$alpha = 0.95, pOutLR = 0.3$	0.505	0.422	0.423	0.334	0.2
	0.624	0.513	0.514	0.377	0.5
	0.780	0.699	0.700	0.574	0.8
	0.805	0.712	0.712	0.606	0.95

Tabulka 4.3: Výsledky pro 100 pozorování

Na základě těchto výsledků při následujících aplikacích volíme parametry $alpha = 0.95$ a $pOutLR = 0.3$, jelikož jak na datech o malém, tak na datech o velkém počtu pozorování vykazoval model nejlepší hodnoty, nezávisle na volbě hodnot varianční matice Σ. Pouze v případě, kdy mimodiagonální hodnoty varianční matice jsou rovny 0.95, se jeví lepší zvolit $alpha = 0.99$ a $pOutLR = 0.4$. Ale takováto vysoká hodnota značí silný vztah mezi všemi proměnnými, který v reálných datech nebývá tak častý.

Kapitola 5

Použití modelu LR-DDC na data z aplikací

Model LR-DDC detekující odlehlé hodnoty pomocí logpodílů v této kapitole aplikujeme na reálná data. Parametry, které nejvíce ovlivňovaly výsledky analýzy, jsou *alpha* a *pOutLR*. Proto jsou právě tyto dva parametry voleny na základě simulační studie uvedené v kapitole 4.

Prvními daty jsou průměrné roční výdaje domácností ve státech Evropské Unie z knihovny `RobComposition` statistického softwaru R [8]. Data jsou k dispozici na stránkách Eurostatu [6]. Druhými daty jsou geochemická data získaná z reálných měření ve spolupráci s Mgr. Danielem Šimíčkem, PhD., z Katedry geologie Přírodovědecké fakulty Univerzity Palackého v Olomouci. Podrobné výsledky analýzy geochemických dat jsou obsaženy v článku [11].

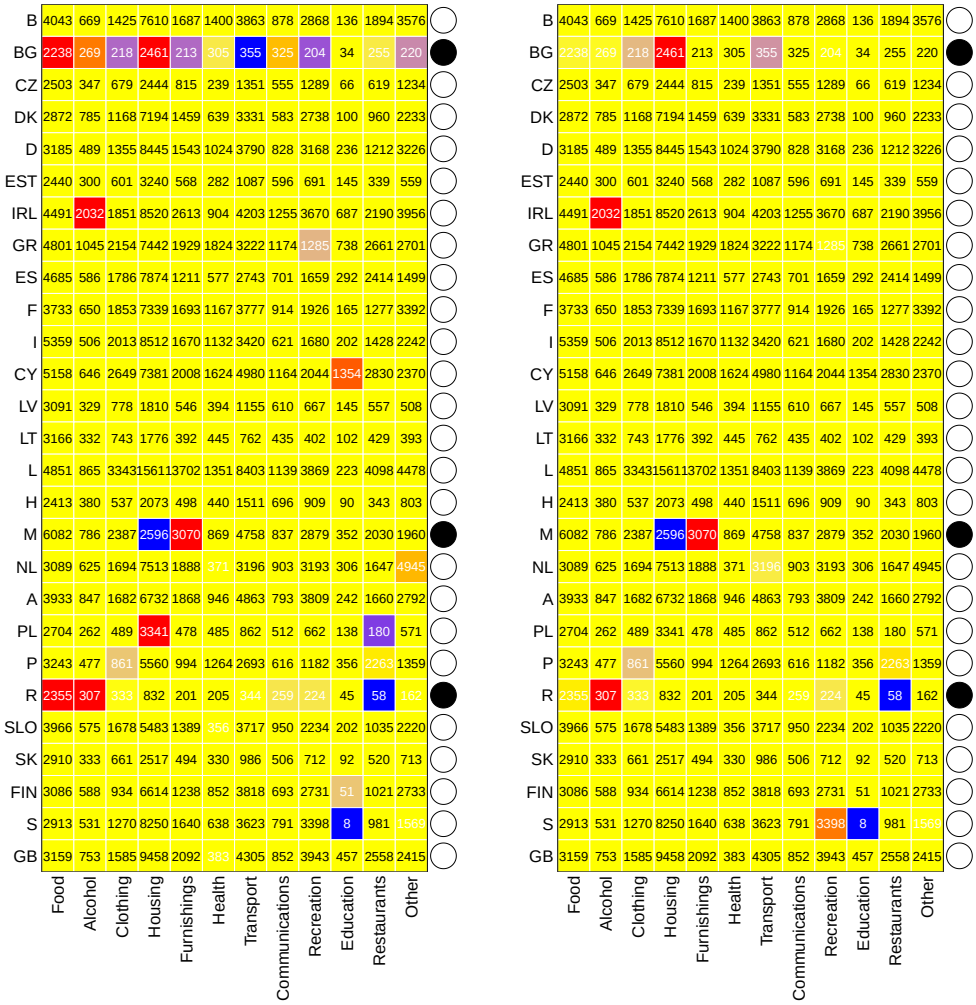
5.1. Výdaje států Evropské unie

Jedná se o průměrné roční výdaje domácností na spotřebu v EUR v jednotlivých státech EU. Výdaje domácností zahrnují veškeré domácí výdaje, ať už rezidentů či nerezidentů, na individuální potřeby. Data obsahují 27 pozorování, stejně jako je států Evropské unie (ve stavu před přijetím Chorvatska a naopak ještě s Velkou Británií jako členským státem), a 12 proměnných představujících výdaje např. za jídlo, alkohol, oblečení, zdraví atd.

Na data jsme aplikovali LR-DDC s různou volbou parametru *alpha*. Výsledky jsou zobrazeny na obrázku 5.1, z něhož vyplývá, že hodnota $alpha = 0.99$ (vpravo) je, co se týče označení odlehlých hodnot, přísnější, jelikož označených prvků není

tolik jako pro $alpha = 0.95$ (vlevo).

Dále ještě zmíníme parametr $pOutROW$, který jsme ponechali na hodnotě určené algoritmem DDC, a to 0.75. Parametr určuje minimální podíl logpodílů k označení celého řádku za odlehlý. Odlehlé řádky jsou na obrázku zaznačeny černou tečkou na okraji tabulky.



(a) ExpendituresEU, $alpha = 0.95$ (b) ExpendituresEU, $alpha = 0.99$

Obrázek 5.1: Aplikování LR-DDC na výdaje domácností ve státech EU

Model LR-DDC označil na obrázku 5.1 hned několik prvků i řádků za odlehlé. Podrobnou analýzu dat budeme dále provádět jenom pro hodnotu $alpha = 0.95$ (vlevo), protože jsme na základě simulační studie zjistili, že poskytuje lepší výsledky (tj. detekuje rozumně velký počet odlehlých hodnot).

Prvním odlehlým řádkem je Bulharsko, které bychom označili za výrazně vybočující, jelikož všechny hodnoty, až na jednu, se v daných proměnných výrazně liší. To odpovídá současné ekonomické situaci Bulharska, patřícího k slabším evropským ekonomikám a nejslabším v rámci EU. Hospodářství této země [13] se drasticky propadlo po zhroucení RVHP a zavedení volného trhu, a to především proto, že bulharský zahraniční obchod byl orientován především na SSSR. Sovětská odbytiště však roku 1991 skončila. Dalším z faktorů, které bulharské hospodářství těžce poznamenalo, bylo uvalení sankcí na země jako Irák a Jugoslávie, kam Bulharsko vyváželo významnou část své produkce.

Další země označené jako odlehlé jsou Malta a Rumunsko. Malta, jako jedna z vyspělejších zemí, má zaznačené pouze dva prvky, a to nízké výdaje na bydlení a vyšší výdaje na vybavení domácností. Rumunsko má prvků zaznačených více. Má zvýšené výdaje na jídlo a alkohol a nižší výdaje na všechny zbylé položky. To odpovídá současné ekonomice Rumunska [14], kde investování do jídla a alkoholu je prioritou. Výdaje na jiné věci jsou deprioritizované.

Rumunsko totiž bylo jednou z chudších zemí RVHP a dodnes se řadí k méně rozvinutým zemím Evropy. Mezi zeměmi Evropské unie je jeho HDP na obyvatele třetí nejnižší, po Bulharsku a Chorvatsku. Může za to především dlouholetá autoritativní vláda Nicolae Ceaușesca, která naprosto znemožnila rozvoj rumunské ekonomiky.

Na základě dosavadní analýzy můžeme usoudit, že chudší země jsou charakteristické vyššími relativními výdaji za jídlo, ubytování či alkohol. Naopak relativní výdaje na jiné služby, jako jsou rekreace, oblečení či transport, jsou vzhledem k jejich ekonomice nižší. Takovýto úsudek je v souladu s naší intuicí.

Dále se ukazuje, jaká je výhoda používat prvkovou detekci spolu s řádkovou namísto detekce pouze řádkové. Při samotné řádkové detekci bychom pouze označili všechny tři země a přitom neměli k dispozici spoustu dalších cenných informací, které nám prvková detekce poskytla.

Posledními zeměmi, které zahrneme do analýzy, jsou Irsko a Švédsko. Obyvatelé Irska ve velké míře konzumují alkohol, a proto do něj investují nemalé finanční prostředky. Takováto informace naprosto odpovídá jak naší představě, tak záznamům o konzumaci alkoholu napříč všemi zeměmi EU. Konkrétně v posledním desetiletí se spotřeba alkoholu ve většině zemí EU snížila [5]. V Irsku je přesto nadále spotřeba na obyvatele vyšší než v celém evropském regionu, a to i v absolutním kontextu (bez uvážení spotřeby alkoholu v rámci relativní struk-

tury výdajů domácností). To je dáno vysokou cenou alkoholu, která je po Finsku druhá nejvyšší v rámci EU [9] a také svou roční spotřebou 11 litrů na obyvatele staršího 15 let [5] korespondující osmé nejvyšší spotřebě za alkohol z 27 členských států EU.

Malá investice domácností ve Švédsku do vzdělání také není nijak překvapující. Vzhledem ke štědrému sociálnímu systému [10] ve Švédsku a dalších skandinávských zemích je vzdělávání v rámci veřejného školského systému bezplatné.

5.2. Geochemická data

Spraše-paleosolové sekvence představují důležité terestrické archivy kvartérních klimatických změn. Elementární složení spraše a paleosolu je významným zdrojem paleoklimatologických a paleoenvironmentálních proxy dat. Ke konečnému složení sprašovo-paleosolových sekvencí přispívá původ eolického detritálního materiálu a následné sprašové a pedogenní procesy. Z geochemických dat lze tedy odvodit vlivy globálního kvartérního klimatického cyklu a lokální charakteristiky (geologie zdrojové oblasti, topografie, srážky atd.). K dešifrování komplexních informací v kompozičních souborech dat je třeba využít vhodný statistický nástroj.

Detekce odlehlých hodnot na úrovni prvků je vhodná pro rozlišení geochemické variability mezi spraší a paleosolem. Nejvíce odlehlých hodnot lze pozorovat v půdních vrstvách svědčících o vyplavování uhličitanů během pedogeneze nebo při ovlivnění některých půdotvorných procesů jako jsou podzolizace nebo lesivace. Podzolizace je chemická migrace Al, Fe a případně organických látek způsobující relativní zvýšení obsahu Si a lesivace je mechanická migrace malých minerálních částí z A do B horizontu.

Prvková detekce odlehlých hodnot byla provedena na souboru dat za účelem zjištění prvků, které jsou zodpovědné za litologickou/pedologickou variabilitu na každé lokalitě. Tato variabilita je graficky znázorněna v tabulkách, které jsou vytvořeny pomocí modelu LR-DDC.

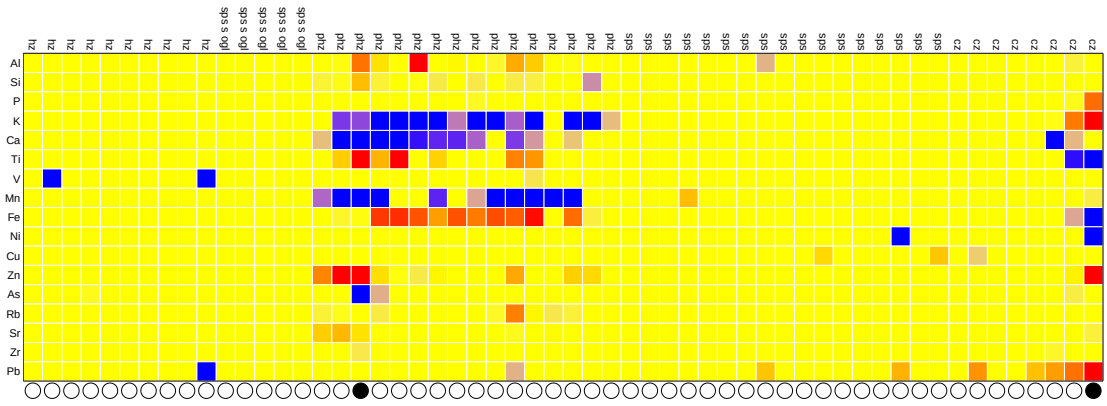
Pro model LR-DDC jsme měli k dispozici soubor reálných dat obsahující naměřené hodnoty chemických prvků vyskytujících se u jednotlivých půdních typů. Ty byly naměřeny v pěti lokalitách, a to v Dobšicích, Ivani, Držovicích, Rozvadovicích a Klopotovicích. Tato data zkoumáme nejprve dle lokalit a poté dle

typu půdy, přičemž půdy, které budeme takto zkoumat, jsou černozem, hnědozem, spraš, spraš se znaky oglejnění, štěrk, parahnědozem či spraš/černozem solifikace. Odlehlé hodnoty hledáme vzhledem k naměřeným hodnotám chemickým prvkům Al, Si, P, K, Ca, Ti, V, Mn, Fe, Ni, Cu, Zn, As, Rb, Sr, Zr, Pb. Po celou dobu bude model pracovat s hodnotami parametrů $alpha = 0.95$ a $pOutLR = 0.3$ zvolenými na základě simulační studie.

K představení v této práci byly vybrány vždy tři datové soubory obsahující největší množství významně odlišných hodnot od příslušných predikcí dle modelu LR-DDC, v rámci barevných kategorií, které jsme představili dříve.

5.2.1. Odlehlé hodnoty zkoumané dle lokalit

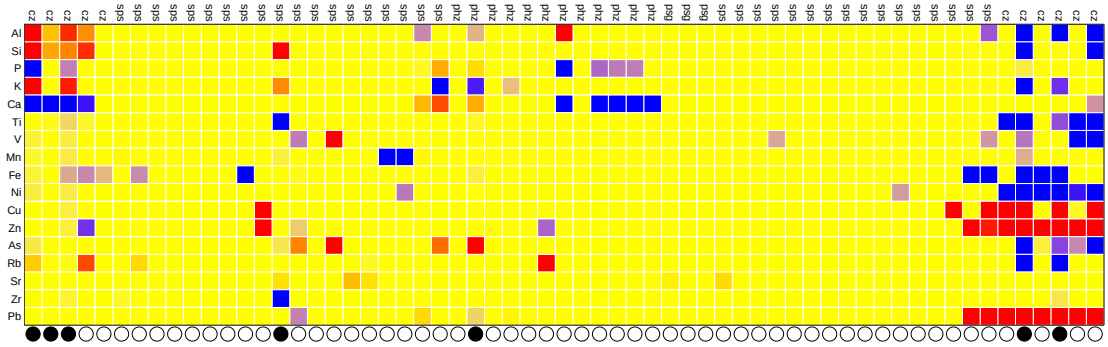
Výsledky po aplikování modelu LR-DDC rozebíráme dle jednotlivých lokalit. V Dobšicích v tabulce na obrázku 5.2 jsou zaznamenány signifikantně odlišné hodnoty pro parahnědozem a černozem. V případě parahnědozemě je výskyt draslíku, vápníku a hořčíku v Dobšicích výrazně nižší než u zbývajících prvků. Oproti tomu železo se zde vyskytovalo v hojném množství. V případě černozemě si můžeme povšimnout například vyšších hodnot u olova a nižších u vápníku. Spraš se ve všech proměnných chová normálně, stejně tak i hnědozem, která měla jenom pár nižších hodnot.



Obrázek 5.2: Aplikace LR-DDC na naměřené hodnoty v Dobšicích dle typu půdy

V obci Ivaň (viz tabulka na obrázku 5.3) převládají jak vyšší, tak nižší hodnoty u černozemě. Parahnědozem obsahuje opět několik nižších hodnot u vápníku, ale není jich tolik jako v Dobšicích. Dále je pár vyšších i nižších hodnot u spraše. Převážná většina odlehlých hodnot se vyskytuje u černozemě. Zde nám model

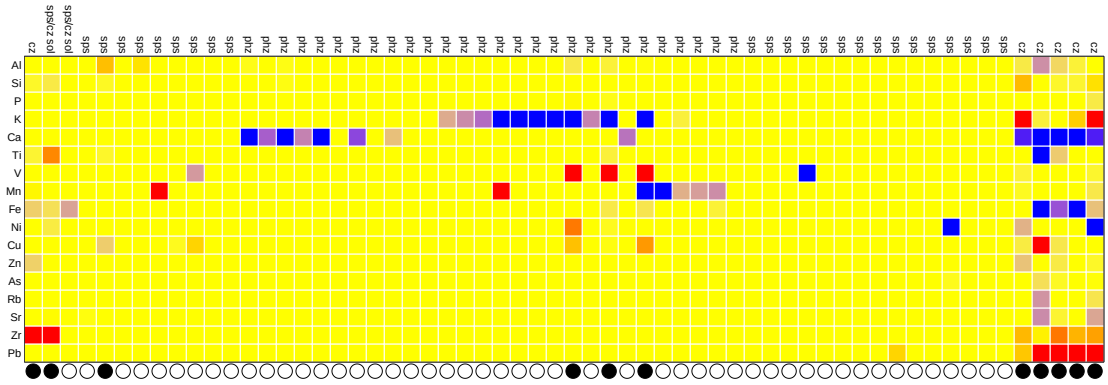
poskytuje informaci, že se v Ivani v černozemi vyskytuje mnohem méně niklu a železa než u jiných půd a naopak je zde mnohem více olova, zinku a mědi.



Obrázek 5.3: Aplikace LR-DDC na naměřené hodnoty v Ivani dle typu půdy

V obci Držovice (viz tabulka na obrázku 5.4) se opět vyskytují nějaké nižší hodnoty u parahnědozemě, konkrétně u draslíku a vápníku. Zbývající odlehlé hodnoty se vyskytují u černozemě, konkrétně jde o výrazně vyšší hodnoty olova a zirkonia či opět nízké hodnoty vápníku.

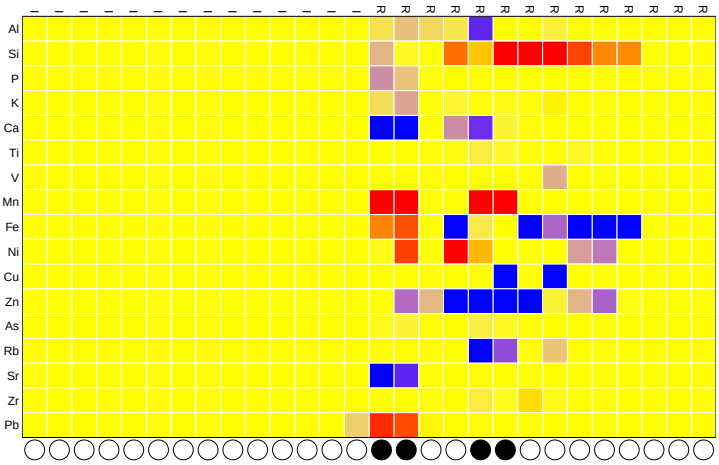
Nadhodnoty olova ve svrchních vrstvách půdy u všech výše zmíněných lokalit indikují znečištění lidskou činností.



Obrázek 5.4: Aplikace LR-DDC na naměřené hodnoty v Držovicích dle typu půdy

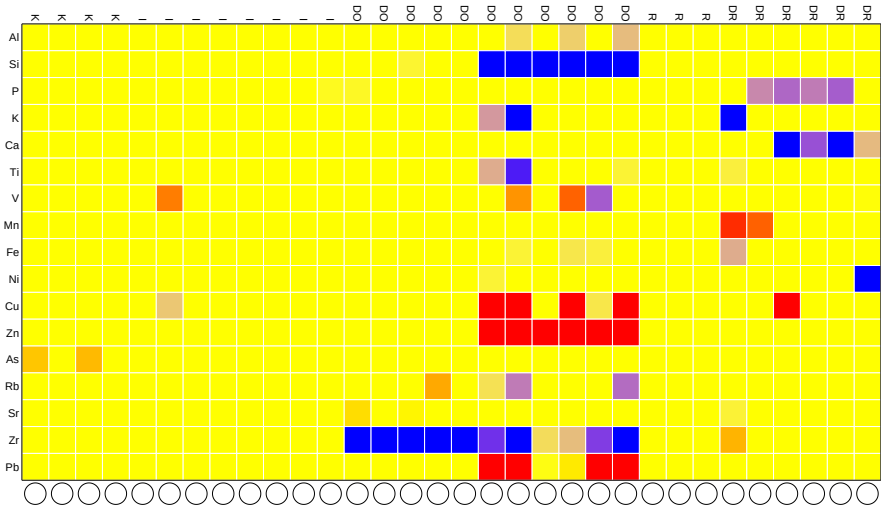
5.2.2. Odlehlé hodnoty zkoumané dle typu půdy

Výsledky po aplikaci modelu LR-DDC na následujících obrázcích lze rozebrat dle typu půdy. Na obrázku 5.5 je možné vidět, že naměřené hodnoty v Ivani se u žádného prvku nijak neodlišují, kdežto v Rozvadovicích se vyskytují u mnoha pozorování odlehlé hodnoty. Říkají nám, že v Rozvadovicích hnědozem obsahuje více křemíku než v Ivani a méně hořčíku, niklu, mědi či zinku.



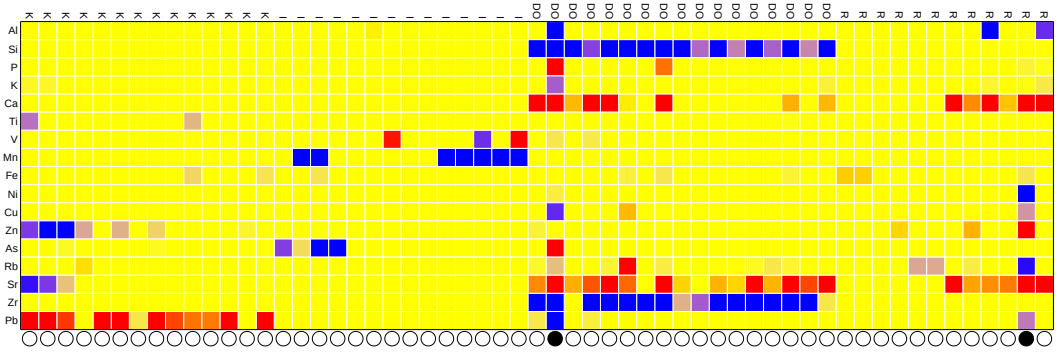
Obrázek 5.5: Aplikace DDC-LR na naměřené hodnoty hnědozemě dle lokalit

Černozem vykazuje odchylující se hodnoty pouze vzhledem k jediné obci, a to k Dobšicím, viz obrázek 5.6. Zde dostáváme informaci o nižším obsahu křemíku a zirkonia a vyšším obsahu olova, zinku či mědi.



Obrázek 5.6: Aplikace LR-DDC na naměřené hodnoty černozemě dle lokalit

Nejbohatší záznam vykazuje parahnědozem, jelikož informaci o naměřených hodnotách máme k dispozici ze všech obcí, jak je patrné z obrázku 5.7. Držovice a Ivaň nevykazují žádné odchylky. V Klopotovicích obsahuje parahnědozem mnohem více olova, v Dobšicích méně křemíku a zirkonia, ale více stroncia než v ostatních obcích. V Rozvadovicích se žádné odlehlé hodnoty nevyskytují.



Obrázek 5.7: Aplikace LR-DDC na naměřené hodnoty parahnědozemě dle lokalit

Závěr

Cílem bakalářské práce bylo vyvinout model, který bude schopen odhalit odlehlé hodnoty a určit, zda nám neposkytují nějakou doplňující informaci o zkoumaném souboru dat. Zároveň bylo potřeba zavést předpoklady, které tento model musí splňovat.

Jelikož se v datech mohou vyskytovat jak prvkové, tak řádkové odlehlé hodnoty, musel model zvládat zpracovat oba tyto druhy chybějících hodnot současně. Důraz byl kladen především na prvkovou detekci, jelikož pouhá řádková detekce je limitována maximálním počtem řádků, které mohou odlehlé hodnoty obsahovat . Řádková detekce proto sloužila jako vedlejší nástroj pro obohacení prvkové detekce.

Ukázalo se, že při práci s kompozičními daty je navíc nutné vzít v potaz jejich specifickou relativní povahu, tedy že veškerá relevantní informace o kompozičních složkách je obsažena v podílech mezi nimi. Správné statistické zpracování kompozičních dat se pak stalo jedním z klíčových předpokladů pro získání interpretovatelných výsledků a posloužilo k celkové validitě statistické analýzy.

Jelikož práce s absolutními či proporcionálními daty je nedostačující k zachycení vztahu korelace mezi jednotlivými proměnnými, byla tedy pozornost zaměřena na reprezentaci kompozic v logpodílech. Jejich výhodou je symetričnost a nezávislost na měřítku dat, ať už na původních (absolutních) proměnných či například proporcionální reprezentaci.

Na základě těchto utvořených předpokladů pro detekci odlehlých hodnot byl navržen model Logratio Deviating Data Cells, zkráceně LR-DDC, beroucí v potaz korelaci mezi proměnnými a také relativní informaci obsaženou v logpodílech mezi nimi.

Parametry, které model využívá, byly optimalizovány na simulovaných datech a následně byl model s těmito parametry aplikován na soubory reálných dat. Výsledky ukázaly, že tento algoritmus vskutku poskytuje užitečné informace o datové struktuře, které mohou sloužit jako podklad pro případné vyloučení

některých pozorování nebo proměnných z další analýzy, popř. jako jeden z argumentů pro použití robustních metod při následném statistickém zpracování dat. Jelikož se model vytvořený na základě této bakalářské práce osvědčil i při analýze reálných kompozičních dat z aplikací, bude přidán do knihovny `robCompositions` statistického softwaru R pro veřejné využití.

Literatura

[1] Aitchison J.: *The Statistical Analysis of Compositional Data*. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 2/44 (1982), s. 139–177.
Dostupné z: https://doi.org/10.1111%2Fj.2517-6161.1982.tb01195.x.

[2] Alqallaf F., Van Aelst S., Yohai V. J., Zamar R. H.: *Propagation of Outliers in Multivariate Data*. Annals of Statistics, Vol. 37 (2009), s. 311-331.
Dostupné z: https://arxiv.org/abs/0903.0447.

[3] CRAN.R. - robCompositions: Compositional Data Analysis [online]. [cit. 2021-10-13].
Dostupné z: https://CRAN.R-project.org/package=robCompositions.

[4] DeepAI - What is the F-score [online]. [cit. 2021-10-21].
Dostupné z: https://deepai.org/machine-learning-glossary-and-terms/f-score.

[5] Drinkaware - Alcohol consumption in Ireland [online]. [cit. 2021-12-26 26.12].
Dostupné z: https://drinkaware.ie/research/alcohol-consumption-in-ireland/?a=problem-alcohol-use-in-ireland.

[6] Eurostat - Domovská stránka [online]. [cit. 2021-12-15].
Dostupné z: https://ec.europa.eu/eurostat.

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

[8] Rousseeuw P. J., Bossche W. V. D.: *Detecting Deviating Data Cells*. Technometrics 2/60 (2018), s. 135-145.
Dostupné z:
https://www.tandfonline.com/doi/full/10.1080/00401706.2017.1340909.

[9] Statista - Price level index for alcoholic beverages in the European Union (EU-28) and other European countries in 2018 [online]. [cit. 2022-03-07].
Dostupné z: https://www.statista.com/statistics/1114418/price-level-index-for-alcoholic-beverages-european-union/.

[10] StudyPortals Masters - Study in Sweden: Tuition Fees and Living Costs in 2022 [online]. [cit. 2021-01-15].
Dostupné z: https://www.mastersportal.com/articles/1661/study-in-sweden-tuition-fees-and-living-costs-in-2022.html.

[11] Šimíček D., Hron K., Vrtková A., Kremeňová E., Bábek O., Sipos G.: *Classification using random forests and detection of deviating cells in compositional data as tools for interpretation of geochemistry of the loess-paleosol sequences; Central European loess belt, Czech Republic*. Rukopis v recenzi.

[12] Štefelová N., Alfons A., Palarea-Albaladejo J., Filzmoser P., Hron K.: *Robust regression with compositional covariates including cellwise outliers*. Advances in Data Analysis and Classification 15 (2021), s. 869-909.
Dostupné z: https://doi.org/10.1007/s11634-021-00436-9.

[13] Wikipedie - Ekonomika Bulharska [online]. [cit. 2021-12-26].
Dostupné z: https://cs.wikipedia.org/wiki/Ekonomika_Bulharska.

[14] Wikipedie - Ekonomika Rumunska [online]. [cit 2021-12-26].
Dostupné z: https://cs.wikipedia.org/wiki/Ekonomika_Rumunska.

[15] Wikipedie - Outlier [online]. [cit. 2021-10-13].
Dostupné z: https://en.wikipedia.org/wiki/Outlier.

[16] Wikipedie - Rumunsko [online]. [cit. 2021-12-26].
Dostupné z: https://cs.wikipedia.org/wiki/Rumunsko.