UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Statistická analýza kompozičních dat pomocí knihovny „compositions“ softwaru R

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Klára Juráňová**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Matematika–ekonomie se zaměřením na bankovnictví/pojišťovnictví
Forma studia: prezenční
Rok odevzdání: 2017

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Klára Juráňová

Název práce: Statistická analýza kompozičních dat pomocí knihovny „compositions“ softwaru R

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2017

Abstrakt: Cílem bakalářské práce je představit knihovnu „compositions“ softwaru R pro statistickou analýzu kompozičních dat. První část popisuje kompoziční data, jejich charakteristické vlastnosti a geometrii. Také zmiňuje, jak řešit problémy spojené s analýzou kompozičních dat. Druhá část se zaměřuje na samotnou knihovnu, její funkce a použití. Práce je doplněna praktickými příklady ze sociologie s kódem z R a interpretací výsledků.

Klíčová slova: knihovna „compositions“, software R, kompoziční data, statistická analýza dat

Počet stran: 55

Počet příloh: 2

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Klára Juráňová

Title: Statistical analysis of compositional data using R package "compositions"

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

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

The year of presentation: 2017

Abstract: The aim of this bachelor thesis is to introduce the R package "compositions" for the statistical analysis of compositional data. The first part describes characteristic features of compositional data and their geometry. It also mentions how to deal with problems which relate to compositional data analysis. The second part is focused on the package itself, its functions and usage. The thesis is completed by practical examples from sociology with the R code and the interpretation of the results.

Key words: "compositions" package, software R, compositional data, statistical data analysis

Number of pages: 55

Number of appendices: 2

Language: Czech

Prohlášení

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

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych na tomto místě poděkovala svému vedoucímu bakalářské práce panu doc. RNDr. Karlu Hronovi, Ph.D. za jeho ochotu, trpělivost, motivaci a čas strávený při konzultacích.

Úvod

Tématem této bakalářské práce je statistická analýza kompozičních dat pomocí knihovny „compositions“ softwaru R. Hlavním cílem práce je seznámit čtenáře s knihovnou pro zpracování kompozičních dat a ukázat její použití na konkrétních datech. Knihovnu „compositions“ softwaru R vytvořili v roce 2009 K. Gerald van den Boogaart, Raimon Tolosana a Matevz Bren, kteří jsou zároveň i autory stěžejní literatury bakalářské práce [3].

První kapitola se věnuje uvedení do problematiky *kompozičních dat*. Jedná se o speciální typ dat nesoucí pouze relativní informaci. Setkáváme se s nimi v mnoha oblastech lidské činnosti, např. v geologii, chemii, ekonomii a dalších odvětvích. Důvodem jejich zkoumání je fakt, že použití standardních metod na kompoziční data by mohlo vést k chybným výsledkům. Je proto třeba s nimi zacházet jinak než s běžnými vícerozměrnými daty. Základy statistické analýzy kompozičních dat položil v roce 1986 John Aitchison [1], proto je specifická geometrie kompozičních dat označována jako Aitchisonova. Základním přístupem pro zpracování těchto dat je jejich převedení na *logpodílové souřadnice*, což jsou již reálné vektory, které lze zpracovávat klasickými metodami.

Druhá kapitola se zaměřuje na samotné zpracování dat v softwaru R. Popisuje funkce knihovny „compositions“, možnosti grafického zobrazení kompozičních dat a metody popisné statistiky na vybraných datových souborech ze sociologie. Vše je doplněno kódem z R a interpretací výsledků. Předpokladem je základní znalost softwaru R.

1 Základní myšlenky kompoziční analýzy dat

Při tvorbě následující kapitoly bylo čerpáno zejména z [3], [4] a [7].

1.1 Kompoziční data

Datový soubor nazveme kompozičním (jednotlivá pozorování pak též kompozičními vektory nebo jednoduše *kompozicemi*), pokud data vyjadřují části z nějakého celku. Jako příklad si můžeme uvést procenta pracovníků pracující v různých sektorech, koncentrace různých látek v nápoji nebo podíly hlasů politických stran ve volbách. Každá složka kompozice má svou hodnotu, která reprezentuje její příspěvek na celku. Kompoziční data mohou být uvedena v absolutním měřítku v jednotkách jako peníze, množství či čas nebo v relativním měřítku, např. v procentech. O tom, zda se skutečně jedná o kompoziční data, následně rozhodne cíl statistické analýzy, v tomto případě zkoumání relativní struktury dat.

Datový soubor může být kompoziční jedině v případě, jestliže má alespoň dvě složky. Informace o tom, kolik hlasů získala jedna politická strana ve volbách, by nám nic neřekla, pokud bychom nevěděli, kolik hlasů získaly ostatní strany nebo alespoň kolik lidí hlasovalo celkem. Kompoziční data jsou tedy přirozeně vícerozměrná. Na rozdíl od standardních vícerozměrných dat nemůžeme ale analyzovat každou složku zvlášť, protože bez znalosti ostatních složek nemá jedna složka žádný význam.

V literatuře [3] je kompozice definována takto:

Definice 1.1 *Kompozicí nazveme vektor D kladných reálných složek $\boldsymbol{x} = [x_1, \ldots, x_D]$, jejichž součet je konstantní a je roven κ.*

Pro kompoziční data je typické, že na velikosti celkového součtu κ nezáleží. Buď je irelevantní (počet zaměstnanců organizace při zkoumání rozložení pracovníků v sektorech) nebo předdefinovaný (objem láhve při zjišťování koncentrace látek

v nápoji). Kromě toho se nám často nepodaří získat kompletní informaci a součet složek neodpovídá možné dosažitelné celkové hodnotě. Navíc, pokud bychom pracovali s absolutními celkovými součty složek, nemohli bychom v mnoha případech porovnávat kompozice mezi sebou. Zkoumáme-li například rozdělení obyvatelstva různých států podle věku, nemůžeme státy mezi sebou v absolutním měřítku porovnávat, neboť každý stát má jiný počet obyvatel. Proto někdy upravujeme definici kompozičních dat [7] a uvažujeme irelevantní celkový součet složek místo konstantního. To odpovídá představě, že kompoziční data obsahují relativní informaci, tj. podstatné jsou vzájemné podíly mezi složkami. Proto často kompozice převádíme tak, aby $\kappa = 1$ (proporcionální části na celku) nebo $\kappa = 100$ (procentuální podíly).

Důvodem zkoumání kompozičních dat je fakt, že při použití standardních metod bychom došli k zavádějícím výsledkům, a to hned v několika směrech:

1. Spojením nezávislých složek do jednoho vektoru a převedením na předepsaný konstantní součet, tzv. *uzavřením* (formálně viz definice 1.2) dostaneme mezi složkami zápornou korelaci, což neodpovídá obvyklé představě, kdy nezávislost znamená nulovou korelaci. Hovoříme o tzv. *negativním biasu (vychýlení)*.

2. U kompozičních dat záleží kovariance mezi dvěma složkami na tom, jaké další složky jsou obsaženy v datech, což popírá klasickou definici kovariance jako vztahu mezi dvěma složkami. Tento problém se nazývá tzv. *falešná korelace*.

3. Varianční matice je vždy singulární kvůli omezení na konstantní součet. To znemožňuje použití metod, které jsou založeny na regulární varianční matici.

4. Složky nemohou být normálně rozděleny kvůli omezené škále jejich hodnot, což nám zabraňuje použít mnoho metod založených na mnohorozměrném normálním rozdělení.

Proto musely být vyvinuty vlastní metody analýzy kompozičních dat, jejichž základy položil John Aitchison (1986) [1].

1.2 Software pro kompoziční analýzu

Statistická analýza se neobejde bez vhodného softwaru. Pro kompoziční data bylo vytvořeno několik knihoven:

- *CODA* – Aitchisonova knihovna pro mikropočítač (1986), která je ale již zastaralá.
- *CoDaPack3D* – knihovna založená na Microsoft Excel a Visual Basic (2003) vhodná pro uživatele preferující grafické uživatelské prostředí.
- *compositions* – knihovna softwaru R (2009), která je v současnosti zřejmě nejpropracovanější variantou pro analýzu kompozičních dat.
- *robCompositions* – knihovna softwaru R (2011) určená především pro robustní statistickou analýzu kompozičních dat.
- *zCompositions* – knihovna softwaru R (2015) zabývající se problémem nulových hodnot v kompozičních datech.

1.2.1 Práce se softwarem R

V této práci budeme všechny příklady počítat pomocí softwaru R, který je volně dostupný z `http://cran.r-project.org`. Na této adrese lze nalézt i instrukce k instalaci a manuály, neboť zde předpokládáme jistou základní znalost jazyka R. Kromě instalace samotného softwaru R je nutné instalovat i potřebné knihovny, v našem případě knihovnu „compositions“.

Samotná instalace knihovny ale nestačí, dále je potřeba ji před začátkem práce v R načíst z paměti. K tomu slouží následující příkaz:

```
> require(compositions)
```

Po jeho zadání jsou upraveny některé standardní příkazy speciálně pro analýzu kompozičních dat.

Některé obecné příkazy se chovají odlišně, pokud je aplikujeme na různé typy objektů v R. Můžeme se například podívat, jak je počítán rozptyl `var` pro kompoziční soubor dat. Nejprve se podíváme na metody dostupné pro výpočet rozptylu:

```
> methods(var)
  [1]  var.acomp  var.aplus  var.default  var.lm      var.mlm
  [6]  var.rcomp  var.rmult  var.rplus    var.tensor  var.test
```

Následně si můžeme zobrazit nápovědu pro rozptyl třídy objektů `acomp` (takto jsou ukládány kompozice, zkratka pro *Aitchison's composition*), ve které se dozvíme detaily výpočtu.

```
> ? var.acomp
```

1.3 Principy kompoziční analýzy dat

1.3.1 Podkompozice a operace uzávěru

Pokud je kompozice A pouze částí z jiné kompozice B, nazývá se A *podkompozicí* B. Ve skutečnosti totiž často kompoziční data neobsahují všechny možné složky. Ve volbách se zajímáme pouze o výsledky nejdůležitějších stran, zanedbáváme menšinové strany. Do analýzy vody nezahrneme „nedůležité“ stopové prvky.

Pro další práci s kompozicí nejprve uzavřeme hodnoty složek, které nás zajímají tak, aby dávaly součet 100 %, popř. 1. Toho snadno docílíme pomocí příkazu `clo` (jako zavřít z anglického *close*).

Nyní si to zkusíme na příkladu, který nás bude provázet celou prací. Jedná se o datový soubor `household`, který popisuje rozdělení typů domácností ve státech Evropy podle počtu dospělých a dětí. Nejsou zde zahrnuty instituce jako třeba domovy důchodců či internáty. Budeme rozlišovat 6 typů domácností:

x_1	1 dospělý s dětmi	`Single_yes`
x_2	1 dospělý bez dětí	`Single_no`
x_3	Pár s dětmi	`Couple_yes`
x_4	Pár bez dětí	`Couple_no`
x_5	Ostatní typy domácností s dětmi	`Other_yes`
x_6	Ostatní typy domácností bez dětí	`Other_no`

Pro názornost vezměme první tři státy tabulky, v tabulce jsou počty jednotlivých typů domácností v tisících. Celou již upravenou tabulku nalezneme na straně 54 v příloze.

Stát	Celkem	x_1	x_2	x_3	x_4	x_5	x_6
Belgie	4 651,8	264,7	1 315,8	1 038,7	1 271,4	240,3	521,0
Bulharsko	2 759,9	78,2	732,7	481,4	703,3	234,4	529,9
ČR	4 606,9	216,2	1 382,0	1 051,4	1 222,8	210,5	523,9

Dejme tomu, že budeme chtít zanedbat ostatní typy domácností a vybereme pouze podkompozici složenou z prvních čtyřech složek (pomocí parametru `parts`) a uzavřeme ji na součet 100 (parametr `total`).

```
> clo(household, parts=
                c("Single_yes", "Single_no",
                  "Couple_yes", "Couple_no"),
                total=100)
```

Stát	x_1	x_2	x_3	x_4
Belgie	6,80	33,82	26,70	32,68
Bulharsko	3,92	36,72	24,12	35,24
ČR	5,58	35,69	27,15	31,58

Nyní už lze porovnávat státy mezi sebou. Kupříkladu můžeme konstatovat, že mezi těmito 3 státy je v Bulharsku největší podíl bezdětných párů žijících ve společné domácnosti.

Parametr `parts` je nepovinný, takže kdybychom jej nezadali, automaticky se vyberou všechny složky kompozice. Při následném statistickém zpracování podkompozic musí být dodržena *podkompoziční soudržnost*. Tj. každá relevantní analýza podkompozic musí být provedena tak, abychom dostali stejné výsledky, ať

už bereme v úvahu celou kompozici nebo jen její podkompozici. Pomocí parametru `total` zadáváme nové κ, implicitní hodnotou je 1. Následující definice uzávěr kompozice formalizuje.

Definice 1.2 *Uzávěrem D-složkové kompozice* $\boldsymbol{x} = [x_1, \ldots, x_D]$ *je vektor*

$$\mathcal{C}[x_1, \ldots, x_D] = \left[\frac{x_1 \cdot \kappa}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D \cdot \kappa}{\sum_{i=1}^{D} x_i} \right],$$

kde $\mathcal{C}$ *je operace uzávěru na konstantu* κ.

1.3.2 Kompozice jako třídy ekvivalence

Jelikož složky kompozice reprezentují relativní příspěvky na celku a na celkovém součtu složek nezáleží, můžeme kompoziční data libovolně přeškálovat. Při vynásobení kompozice kladnou hodnotou se totiž informace obsažena v datech nezmění. Dva vektory tak můžeme označit za kompozičně ekvivalentní, jestliže je jeden jednoduše násobkem druhého:

$$\boldsymbol{a} =_A \boldsymbol{b} \Leftrightarrow \exists\, s > 0 \quad \forall\, i : a_i = s \cdot b_i.$$

Tedy vektory $\boldsymbol{a} = [\frac{2}{7}, \frac{4}{7}, \frac{1}{7}]$, $\boldsymbol{b} = [2, 4, 1]$ a $\boldsymbol{c} = [60, 120, 30]$ znázorňují tutéž kompozici se stejnými poměry mezi složkami. Tato vlastnost se označuje jako *invariance* (neměnnost) *vůči změně měřítka*. Aitchison [1] navíc ukázal, že všechny funkce kompozic, které jsou invariantní vůči změně měřítka, se dají vyjádřit jako funkce logaritmů podílů $\ln(x_i/x_j)$.

1.3.3 Perturbace jako změna jednotek

Kompoziční data mohou být dána v mnoha různých jednotkách. Někdy je potřeba jednotky převést. Uvažujme kompozici $\boldsymbol{u}$ složení látek v pokrmu v gramech, které chceme převést na jednotky energie kJ. K provedení je třeba znát vektor množství kJ/g pro jednotlivé látky, označme jej $\boldsymbol{v}$. Součinem složek kompozice

a příslušných složek vektoru $\boldsymbol{v}$ získáme nový vektor, který po uzavření tvoří požadovanou kompozici v jednotkách kJ. Právě jsme provedli *perturbaci vektorů* $\boldsymbol{u}$ a $\boldsymbol{v}$, kterou značíme $\oplus$, tj. $\boldsymbol{u} \oplus \boldsymbol{v}$ (viz definice 1.4).

Kompoziční data jsou *invariantní vůči perturbaci*. I když různí analytici pracují v jiných jednotkách, a dojdou proto k různým hodnotám, jedná se pořád o tutéž kompozici. Kvalitativní výsledky budou stejné, pokud lze jednotky mezi sebou převádět, tzn. pokud kompozice obsahují stejnou informaci.

1.3.4 Amalgamace

Další typickou operací pro analýzu kompozičních dat je *amalgamace*, což je spojení (součet) některých složek do jedné. Před provedením amalgamace je ale nutné si dobře promyslet, jestli je vhodná pro danou situaci, jelikož vždy dojde ke ztrátě informace. Kupříkladu sloučení složek „jablka“ a „hrušky“ je oprávněné, pokud zkoumáme stravovací návyky jednotlivých populací (a zajímá nás podíl ovoce v potravě), ale může být chybné, pokud sledujeme důležitost jednotlivých druhů ovoce v různých kulturách.

Vraťme se ke kompozicím popisujícím rozdělení typů domácností ve státech Evropy. Řekněme, že bychom chtěli zkoumat pouze rozdělení mezi jednotlivce, páry a ostatní typy domácností, tj. počet dětí by byl pro náš problém irelevantní. Je třeba sloučit složky `Single_yes` a `Single_no`, `Couple_yes` a `Couple_no`, `Other_yes` a `Other_no`. Pro sloučení domácností s jedním dospělým do sloupce `Single` zadáme příkaz:

```
> household$Single = totals(aplus(household,
                        c("Single_yes","Single_no")))
```

Amalgamaci vypočítáme pomocí funkce `totals` z neuzavřeného souboru ze třídy objektů `rplus` nebo `aplus` (blíže viz podkapitola 1.4). Pro tyto třídy objektů je hodnota celkového součtu složek stále podstatná. Sloučení složek `Single_yes` a `Single_no` znázorňuje následující tabulka:

Stát	Single	Single_yes	Single_no
Belgie	1580,5	264,7	1315,8
Bulharsko	810,9	78,2	732,7
ČR	1598,2	216,2	1382,0

1.3.5 Invariance vůči permutaci

Již jsme si ukázali tři ze čtyř principů invariance, které tvoří základní pilíře analýzy kompozičních dat: invariance vůči změně měřítka, invariance vůči perturbaci a podkompoziční soudržnost. Nyní zbývá zmínit zbývající vlastnost kompozičních dat, a to *invarianci vůči permutaci*.

Výsledky žádné analýzy by neměly záviset na pořadí, ve kterém jsou složky uspořádány v datovém souboru. Zdá se to být evidentní, ale je překvapivé, kolik metod tuto vlastnost nemá. Například euklidovská vzdálenost dále zmíněných *alr souřadnic* kompozičních dat (v podkapitole 1.5.3) není invariantní vůči permutaci. Proto je třeba pro výpočet vzdálenosti použít jiný typ souřadnic.

1.4 Mnohorozměrná měřítka

Jednou ze základních vlastností každé proměnné je její měřítko, které popisuje hodnoty, jakých může proměnná nabývat, a jak je interpretovat. S tím souvisí i matematické operace, jež je možno použít. Ve statistice data obvykle rozdělujeme podle měřítka do několika skupin:

1. Nominální data – nabývají jedné z několika možných hodnot, které nelze porovnávat.

2. Ordinální data – nabývají jedné z několika možných hodnot, které lze seřadit.

3. Intervalová data – nabývají hodnot z množiny reálných čísel, absolutní rozdíly mezi hodnotami mají smysl.

4. Poměrová data – nabývají hodnot z množiny kladných reálných čísel, relativní poměry mezi hodnotami mají smysl.

Každé měřítko je spojeno s typickým statistickým modelem jako třeba multinomické rozdělení pro nominální data a normální rozdělení pro reálná intervalová data. U kompozic není mnohorozměrné měřítko jen zřetězením měřítek pro jednotlivé proměnné. Kompozice jsou speciální objekty s novým vícerozměrným měřítkem, které má některé společné vlastnosti: skládá se z kladných čísel, má součet 1 (popř. 100 %) a podíly mezi složkami jsou relevantní.

V minulosti byl výběr měřítka pro kompoziční data diskutovanou otázkou. Proto knihovna „compositions“ obsahuje větší množství variant:

- `rmult` – reálné vícerozměrné měřítko

  Data pocházejí z $\mathbb{R}^D$ a statistickým modelem je mnohorozměrné normální rozdělení.

- `rplus` – reálné intervalové měřítko z $\mathbb{R}_+^D$

  Data jsou kladná s absolutní geometrií, příslušným rozdělením je seříznuté (angl. *truncated*) mnohorozměrné normální rozdělení.

- `aplus` – Aitchisonovo (tzn. poměrové) měřítko z $\mathbb{R}_+^D$

  Data jsou kladná s relativní geometrií. Je běžné je analyzovat až po převedení na log-souřadnice, příslušné rozdělení je typicky logaritmicko-normální.

- `rcomp` – reálné (tzn. intervalové) kompoziční měřítko

  Data jsou považována za standardní reálná data s konstantním součtem složek.

- `acomp` – Aitchisonovo (poměrové) kompoziční měřítko

  Uvažuje kompoziční data tak, jako jsme je doposud popsali. Obsahují relativní informaci a mají konstantní součet složek. Statistickým modelem je normální rozdělení na *simplexu* (viz podkapitola 1.5.1).

- `ccomp` – celočíselné kompoziční měřítko

  Speciální případ předchozího měřítka, kdy z podstaty problému plyne celočíselnost dat.

Preferovat bychom ale měli Aitchisonovo kompoziční měřítko `acomp` [3]. Proto před každou analýzou přidělíme datovému souboru `x` toto měřítko pomocí příkazu `acomp(x)`. Knihovna „compositions“ pak dále automaticky pracuje s daty podle zvoleného měřítka.

V návaznosti na výzkum Johna Aitchisona bylo na přelomu 20. a 21. století ukázáno, že základní principy pro kompoziční analýzu dat (invariance vůči změně měřítka, perturbaci a permutaci a podkompoziční soudržnost) implikují relativní geometrii, ve které záleží jen na podílech mezi složkami [7]. Podle něj byla pak pojmenována jako Aitchisonova geometrie (acomp), která bude podrobně popsána v následující části.

1.5 Aitchisonova geometrie na simplexu

Po zavedení Aitchisonovy geometrie pro kompoziční data bylo dokázáno, že tato struktura tvoří euklidovský vektorový prostor, který je izometricky ekvivalentní s $\mathbb{R}^{D-1}$. Logickým důsledkem této ekvivalence je převedení problému obsahujícího kompozice o D složkách na klasický vícerozměrný problém zahrnující reálné vektory o $D-1$ souřadnicích. Pokud se dá vybrané měřítko popsat pomocí euklidovského vektorového prostoru, lze statisticky analyzovat kompozice v souřadnicích vzhledem k ortonormální bázi místo původních dat a dané měřítko bude zachováno.

1.5.1 Simplex a operace na simplexu

Aitchisonova geometrie mívá přívlastek „na simplexu“, neboť simplex je výběrovým prostorem reprezentací kompozičních dat s konstantním součtem složek. Kompozice, jejíž složky mají konstantní součet, se nazývá uzavřená a množina těchto kompozic tvoří právě simplex.

Definice 1.3 *Množinu všech možných uzavřených kompozic*

$$\mathbb{S}^D := \left\{ \boldsymbol{x} = [x_1, \cdots, x_D] : x_i \geq 0, \sum_{i=1}^{D} x_i = 1 \right\}$$

nazveme D-složkovým simplexem.

Tudíž dvousložkovou kompozici můžeme zobrazit jako bod na intervalu $(0, 1)$, resp. $(0, \kappa)$. Simplex pro $D = 3$ složky zobrazíme rovinným ternárním diagramem, pro $D = 4$ pak trojrozměrným pravidelným čtyřstěnem, kde každá z možných trojsložkových podkompozic je reprezentována jednou stěnou čtyřstěnu. Chceme-li vybrat trojsložkovou podkompozici ze čtyřsložkové kompozice, promítneme body ve čtyřstěnu na požadovanou stěnu tak, že obraz leží na spojnici vzoru a protilehlého vrcholu čtyřstěnu (tento vrchol odpovídá právě odstraněné složce kompozice). Zároveň tato projekce odpovídá uzávěru trojsložkové podkompozice pocházející ze čtyřsložkové kompozice. Pro více složek pak dostaneme vícerozměrný simplex, který funguje opět na stejných principech.

Již jsme zmínili, že Aitchisonova geometrie na simplexu tvoří vektorový prostor. Budeme tak uvažovat analogické operace k těm, které známe z euklidovské geometrie. Jak víme, množina všech reálných vektorů spolu s operacemi sčítání vektorů a násobení vektoru skalárem tvoří vektorový prostor. Tyto operace ale nejsou invariantní vůči změně měřítka ani podkompozičně soudržné. Představme si tedy perturbaci, zmíněnou v podkapitole 1.3.3 a formalizovanou definicí 1.4, jako kompoziční součet vektorů a nově zavedenou mocninnou transformaci jako kompoziční násobení vektoru skalárem (viz definice 1.5).

Definice 1.4 *Perturbaci $\boldsymbol{z}$ dvou D-složkových kompozic $\boldsymbol{x}$ a $\boldsymbol{y}$ z $\mathbb{S}^D$ definujeme jako*

$$\boldsymbol{z} = \boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}[x_1 \cdot y_1, \ldots, x_D \cdot y_D].$$

S knihovnou „compositions“ můžeme vektory `x` a `y` perturbovat dvěma způsoby. Buď zadáme `perturbe(x,y)` nebo jednoduše sečteme (resp. odečteme) dva vektory třídy `acomp`.

Snadno vidíme, že perturbace tvoří spolu se simplexem komutativní grupu [7], která má neutrální prvek $\mathbf{1} = [1/D, \cdots, 1/D]$ a inverzní prvek

$$\ominus \boldsymbol{x} = \mathcal{C}[1/x_1, \cdots, 1/x_D].$$

Tedy platí vztahy

$$\boldsymbol{x} \oplus \mathbf{1} = \boldsymbol{x},$$

$$\boldsymbol{x} \oplus (\ominus \boldsymbol{x}) = \mathbf{1}.$$

Perturbace inverzním (opačným) prvkem hraje roli odečítání, proto se značí $\ominus$ a platí

$$\boldsymbol{x} \ominus \boldsymbol{y} := \boldsymbol{x} \oplus (\ominus \boldsymbol{y}).$$

Při statistické analýze je často nutné perturbovat všechny kompozice z datového souboru. Proto zavádíme velký operátor $\oplus$ pro N kompozic:

$$\bigoplus_{i=1}^{N} \boldsymbol{x_i} := \boldsymbol{x_1} \oplus \cdots \oplus \boldsymbol{x_N}.$$

Definice 1.5 *Mocninnou transformaci $\boldsymbol{z}$ D-složkové kompozice $\boldsymbol{x}$ z $\mathbb{S}^D$ skalárem (reálným číslem) λ definujeme jako*

$$\boldsymbol{z} = \lambda \odot \boldsymbol{x} = \mathcal{C}[x_1^{\lambda}, \ldots, x_D^{\lambda}].$$

Knihovna opět umožňuje dva způsoby zápisu mocninné transformace. Kompozici `x` mocninně transformujeme skalárem `lambda` příkazem `power(x,lambda)` nebo násobením skaláru vektorem třídy `acomp`.

Mocninnou transformaci číslem -1 můžeme použít pro definici opačného prvku: $(-1) \odot \boldsymbol{x} = \ominus \boldsymbol{x}$.

1.5.2 Kompoziční skalární součin, norma a vzdálenost

Je možné dokázat, že trojice $(\mathbb{S}^D, \oplus, \odot)$ tvoří vektorový prostor [7]. Doplněním Aitchisonova skalárního součinu dostaneme euklidovský vektorový prostor, tudíž budeme moci definovat normu kompozice či vzdálenost dvou kompozic.

Definice 1.6 *Aitchisonův skalární součin kompozic* $\boldsymbol{x} = [x_1, \cdots, x_D]$ *a* $\boldsymbol{y} = [y_1, \cdots, y_D]$ *z* $\mathbb{S}^D$ *definujeme jako*

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

Aitchisonův skalární součin získáme zadáním `scalar(x,y)`, pokud jsou vektory `x` a `y` třídy `acomp`.

Normou kompozice, tedy její délkou, rozumíme číslo $\|\boldsymbol{x}\|_A = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_A}$ a v R ji vypočteme příkazem `norm(x)`, kde `x` je třídy `acomp`. Pokud chceme kompozici `x` normovat (tzn. její délka bude jednotková), zadáme `normalize(x)`.

Definice 1.7 *Aitchisonovu vzdálenost kompozic* $\boldsymbol{x} = [x_1, \cdots, x_D]$ *a* $\boldsymbol{y} = [y_1, \cdots, y_D]$ *z* $\mathbb{S}^D$ *definujeme jako*

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

Vzdálenost kompozic `x` a `y` přímo spočteme příkazem `norm(x-y)`.

1.5.3 Souřadnice

Přidáním operace Aitchisonova skalárního součinu ke trojici $(\mathbb{S}^D, \oplus, \odot)$ získáváme $(D-1)$-rozměrný euklidovský vektorový prostor na simplexu. To znamená, že můžeme kompozice převádět na reálné vektory pomocí tzv. logpodílových souřadnic, neboť euklidovský vektorový prostor je vždy ekvivalentní s reálným. Této ekvivalence je dosaženo díky izometrii, transformaci ze simplexu do reálného prostoru, která zachovává úhly a vzdálenosti.

Centrované logpodílové souřadnice (clr)

Prvními izometrickými souřadnicemi kompozice $\boldsymbol{x}$ jsou *centrované logpodílové (clr) souřadnice*

$$clr(\boldsymbol{x}) = \left(\ln \frac{x_i}{g(\boldsymbol{x})} \right)_{i=1,\ldots,D}, \quad \text{kde } g(\boldsymbol{x}) = \sqrt[D]{x_1 \cdot x_2 \cdots x_D}.$$

Pokud $\boldsymbol{x}^* = clr(\boldsymbol{x})$, pak kompozici $\boldsymbol{x}$ dostaneme zpětně jako $\mathcal{C}[\exp(\boldsymbol{x}^*)]$, kde je aplikace exponenciální funkce provedena po složkách. Podle definice dávají clr souřadnice součet nula. Znázorněním clr souřadnic je nadrovina (a také vektorový podprostor značený $\mathbb{H}$, nazývaný clr-rovina) reálného prostoru $\mathbb{H} \subset \mathbb{R}^D$ ortogonální k vektoru $\mathbf{1} = [1, \ldots, 1]$. Varianční matice clr souřadnic kompozice je singulární, což následně vedlo k zavedení jiných vhodnějších souřadnic pro statistické metody, které vyžadují regularitu varianční matice. Clr souřadnice kompozice spočteme zadáním příkazu `clr` a zpět na kompozici je převedeme pomocí `clrInv`.

Izometrické logpodílové souřadnice (ilr)

Kompozici vyjádříme v *izometrických logpodílových (ilr) souřadnicích* pomocí izometrického lineárního zobrazení mezi simplexem a $\mathbb{R}^{D-1}$. Ilr souřadnice konstruujeme následovně. Nechť je dána ortonormální báze $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{D-1}\}$ simplexu $\mathbb{S}^D$ a uvažujme $D \times (D-1)$-rozměrnou matici $\boldsymbol{V}$, jejíž sloupce jsou $\mathrm{clr}(\boldsymbol{e}_i)$. Pro ortonormální bázi platí, že $\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle_A = \delta_{ij}$ (δ_{ij} je Kroneckerovo delta, které je nula pro $i \neq j$ a jedna pro $i = j$). Platí $\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle_A = \langle clr(\boldsymbol{e}_i), clr(\boldsymbol{e}_j) \rangle = \delta_{ij}$, z čehož pro matici $\boldsymbol{V}$ plyne:

$$\boldsymbol{V}^t \cdot \boldsymbol{V} = \boldsymbol{I}_{D-1} \quad \text{a} \quad \boldsymbol{V} \cdot \boldsymbol{V}^t = \boldsymbol{I}_D - \frac{1}{D}\, \mathbf{1}_{D \times D},$$

kde $\boldsymbol{I}_D$ je jednotková matice řádu D a $\mathbf{1}_{D \times D}$ je čtvercová matice řádu D plná jedniček. Sloupce matice $\boldsymbol{V}$ mají součet 0, neboť reprezentují vektory clr-roviny.

Díky těmto vlastnostem můžeme najít jednoduché vztahy mezi ilr souřadnicemi ξ a kompozicí $\boldsymbol{x}$:

$$clr(\boldsymbol{x}) \cdot \boldsymbol{V}^t = \ln \boldsymbol{x} \cdot \boldsymbol{V}^t =: ilr(\boldsymbol{x}) = \xi,$$

$$ilr(\boldsymbol{x}) \cdot \boldsymbol{V} = clr(\boldsymbol{x}) \rightarrow \boldsymbol{x} = \mathcal{C}[\exp \xi \cdot \boldsymbol{V}].$$

Těmito vztahy jsme tedy zadefinovali ilr souřadnice, což jsou souřadnice kompozice vzhledem k dané ortonormální bázi. Existuje nekonečně mnoho různých ilr souřadnicových systémů, jako máme možných ortonormálních bází splňujících

dané podmínky. Jak vybrat správnou bázi tak, aby byly výsledné ilr souřadnice dobře interpretovatelné, se dozvíme v podkapitole 2.2.7 o bilancích.

Ilr souřadnice kompozice $\boldsymbol{x}$ získáme zadáním `ilr(x)`. Bázi získáme pomocí příkazu `ilrBase(x,z,D)` s parametry kompozice `x`, ilr souřadnice `z` a počet složek `D` (v příkazu můžeme použít pouze jeden z parametrů). Další užitečné příkazy jsou `ilrInv` pro výpočet kompozice příslušné k daným souřadnicím a `ilr2clr`, popř. `clr2ilr` pro převod mezi ilr a clr souřadnicemi.

Aditivní logpodílové souřadnice (alr)

Za zmínku stojí i původní Aitchisonův přístup založený na *aditivních logpodílových (alr) souřadnicích*, přestože se již nepoužívá,

$$alr(\boldsymbol{x}) = (\ln (x_i/x_D)_{i=1,\dots,D-1}).$$

Jedním z důvodů je kromě absence izometrie s Aitchisonovou geometrií i porušení invariance vůči permutaci složek kompozice, zmíněné v podkapitole 1.3.5.

Porovnání jednotlivých typů souřadnic

Důvodem ke vzniku tolika logpodílových souřadnicových systémů je, že každý má svoje specifické vlastnosti vhodné pro různé statistické metody. Pro všechny tři druhy souřadnic platí, že převedou perturbaci a mocninnou transformaci na klasický součet a násobení vektoru skalárem:

$$clr(\boldsymbol{x} \oplus \boldsymbol{y}) = clr(\boldsymbol{x}) + clr(\boldsymbol{y}), \quad clr(\lambda \odot \boldsymbol{x}) = \lambda \cdot clr(\boldsymbol{x}),$$
$$ilr(\boldsymbol{x} \oplus \boldsymbol{y}) = ilr(\boldsymbol{x}) + ilr(\boldsymbol{y}), \quad ilr(\lambda \odot \boldsymbol{x}) = \lambda \cdot ilr(\boldsymbol{x}),$$
$$alr(\boldsymbol{x} \oplus \boldsymbol{y}) = alr(\boldsymbol{x}) + alr(\boldsymbol{y}) \quad alr(\lambda \odot \boldsymbol{x}) = \lambda \cdot alr(\boldsymbol{x}).$$

Clr a ilr souřadnice jsou izometrické, tzn. že zachovávají skalární součin. Alr souřadnice tuto vlastnost nemají, tudíž nemohou být použity v případě, který zahrnuje vzdálenosti, úhly nebo tvary, protože je deformuje. Nevýhodou clr souřadnic je, že vedou k singulární varianční matici. Pokud bychom potřebovali inverzní matici, museli bychom použít obecnější Moore-Penroseovu inverzní matici. Na druhou stranu jsou clr souřadnice snáze interpretovatelnější než ilr souřadnice,

u kterých musíme pro dobrou interpretaci vhodně zvolit ortonormální bázi, k čemuž je třeba znát podstatu zkoumaného problému.

Při analýze kompozičních dat tedy nejprve spočteme nové souřadnice kompozice a dále již s nimi pracujeme tak, jako by se jednalo o běžná data. Použijeme standardní statistické metody, protože souřadnice kompozic jsou reálné neomezené hodnoty, je obvykle pouze zapotřebí přihlédnout k interpretaci použitých souřadnic. U některých statistických metod se navíc ukazuje [3], že výsledky nezávisí na volbě báze, kterou použijeme k výpočtu souřadnic.

2 Analýza kompozičních dat v R

Základní literaturou pro psaní druhé kapitoly byly knihy [3], [7] a článek [6]. Budeme zde již více pracovat s konkrétními daty, která byla vybrána z webových stránek Eurostatu. Jedná se o datový soubor `household` [8], který jsme již představili v předchozí kapitole, a soubor `satisfaction` [9]. Kompletní datové soubory nalezneme v přílohách (strana 54 a 55). První šestisložkovou kompozici obsahující rozdělení domácností ve vybraných státech Evropy podle počtu dospělých a dětí v domácnosti již známe. Druhá trojsložková kompozice popisuje spokojenost se životem v některých evropských státech. Spokojenost byla bodována na škále od 0 do 10 a rozdělena do tří kategorií na malou (0–5 bodů), střední (6–8 bodů) a velkou (9–10 bodů). V softwaru R jsme složky označili jako `Low`, `Medium` a `High`. Pro názornost zobrazíme první tři kompozice souboru dat:

Stát	Malá	Střední	Velká
Belgie	9,2	69,9	20,9
Bulharsko	64,2	29,8	5,9
ČR	25,4	53,3	21,3

Všechny kompozice byly upraveny pomocí operace uzávěru na celkový součet 100 (viz 1.3.1). Jsou tedy vyjádřeny v procentech pro jednotlivé složky.

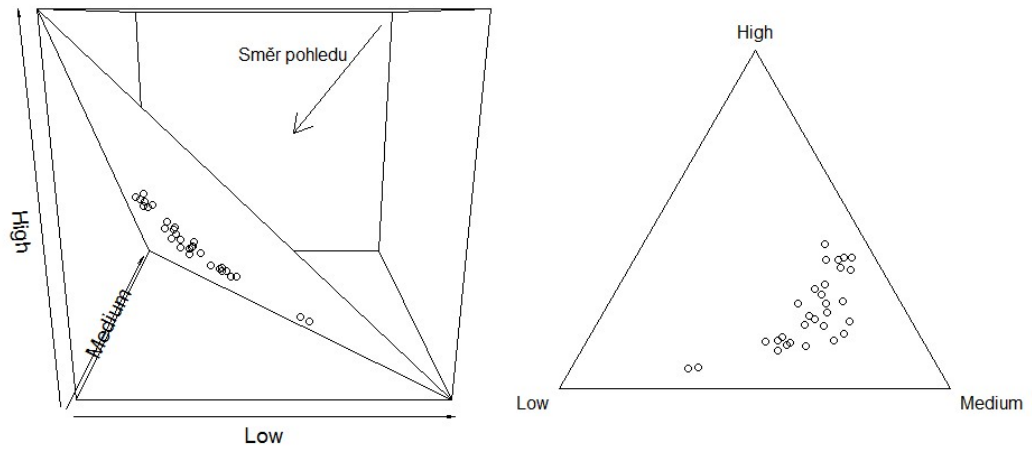
2.1 Základní grafické zobrazení

Existuje několik způsobů, jak zobrazit kompozice. Nejčastěji používáme ternární diagram pro trojsložkové kompozice, bodový graf logaritmů podílů několika složek nebo skupinu sloupcových či koláčových grafů.

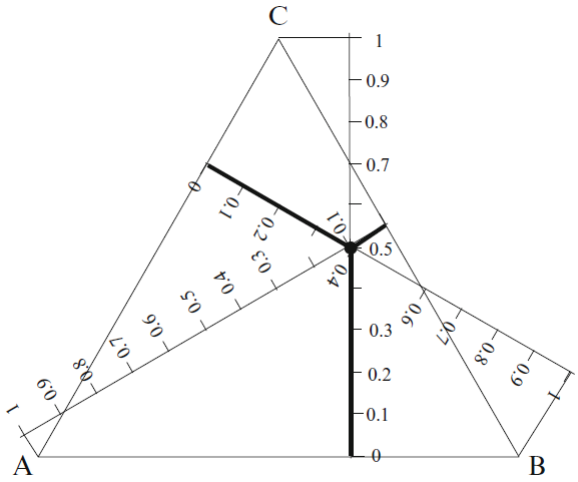
2.1.1 Ternární diagram

Ternárním diagramem zobrazujeme uzavřené trojsložkové kompozice (resp. podkompozice). Kdybychom chtěli zobrazit takovou kompozici v trojrozměrném bodovém grafu, vždy by ležela v trojúhelníku s vrcholy $(1, 0, 0)$, $(0, 1, 0)$ a $(0, 0, 1)$.

Ternárním diagramem je právě tento trojúhelník s vrcholy označenými názvem osy, na které leží (viz obrázek 2.1). Body leží právě ve zmíněném trojúhelníku proto, že uzavřená trojsložková kompozice má dva stupně volnosti.

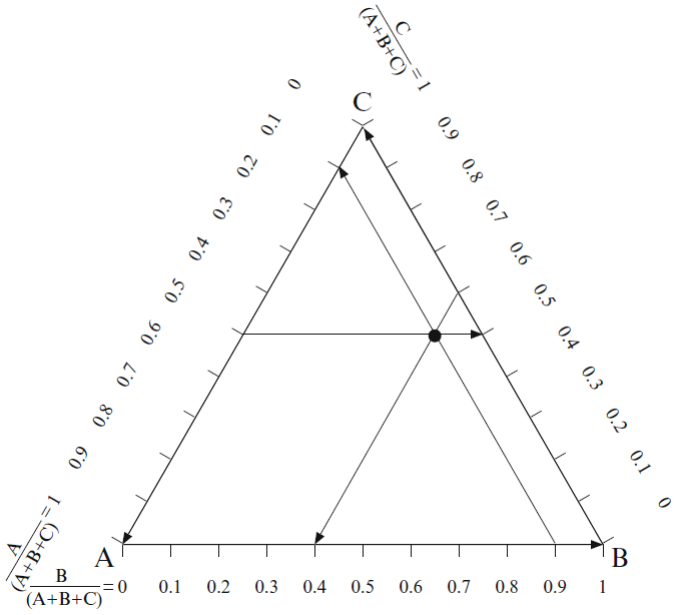


Obrázek 2.1: Ternární diagram kompozičních dat `satisfaction` umístěný v trojrozměrném prostoru



Obrázek 2.2: Zobrazení kompozice $[A, B, C] = [0,1; 0,4; 0,5]$ v ternárním diagramu její interpretace (převzato z [3])

Při interpretaci využijeme vlastnosti, že úsečky spojující libovolný bod uprostřed trojúhelníku kolmo se třemi stranami mají v součtu vždy stejnou délku. Délky jednotlivých částí znázorňují podíly složek kompozice na celku (stejně jako vidíme na obrázku 2.2).



Obrázek 2.3: Další interpretace kompozice $[A, B, C] = [0,1; 0,4; 0,5]$ v ternárním diagramu (převzato z [3])

Další možností je vytvořit souřadnicové osy z jednotlivých stran trojúhelníku. Chceme-li vyčíst některou ze složek kompozice $[A, B, C]$, zobrazíme daný bod na příslušnou osu tak, že ho posuneme podle předchozí osy (při zobrazování na osu B posunujeme podle osy A apod., viz obrázek 2.3).

Pro vykreslení ternárního diagramu kompozice `x` zadáme `plot(x)`, kde `x` je třídy `acomp`, popř. `rcomp`. Pro přidání kompozic do již existujícího ternárního diagramu nastavíme parametr `add` na `TRUE`: `plot(x, add=TRUE)`.

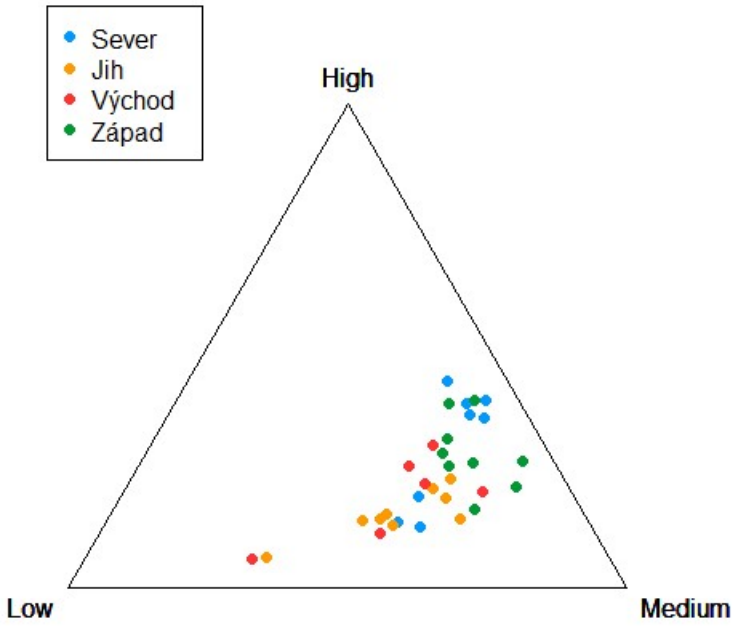
Ternární diagram kompozičních dat `satisfaction` z pravé části obrázku 2.1 získáme takto:

```
> satisfaction=acomp(satisfaction)
> plot(satisfaction)
```

V prvním kroku převedeme kompozice na třídu `acomp` a poté ji vykreslíme v ternárním diagramu. Nyní podle něj určíme, v jaké části Evropy jsou lidé nejspokojenější. Tuto informaci pak můžeme využít při rozhodování, kam vyrazit za prací. Rozdělíme si státy na čtyři skupiny podle světových stran, v R si příslušné řádky přiřadíme do proměnných `north`, `south`, `east` a `west`:

Sever	Jih	Východ	Západ
`north`	`south`	`east`	`west`
Dánsko	Chorvatsko	Bulharsko	Belgie
Island	Itálie	ČR	Francie
Estonsko	Kypr	Maďarsko	Irsko
Finsko	Malta	Polsko	Lucembursko
Litva	Portugalsko	Rumunsko	Německo
Lotyšsko	Řecko	Slovensko	Nizozemsko
Norsko	Slovinsko		Rakousko
Švédsko	Srbsko		Švýcarsko
	Španělsko		VB

Vyznačíme si jednotlivé části v ternárním diagramu barevně (obrázek 2.4), čehož dosáhneme pomocí následujícího kódu.



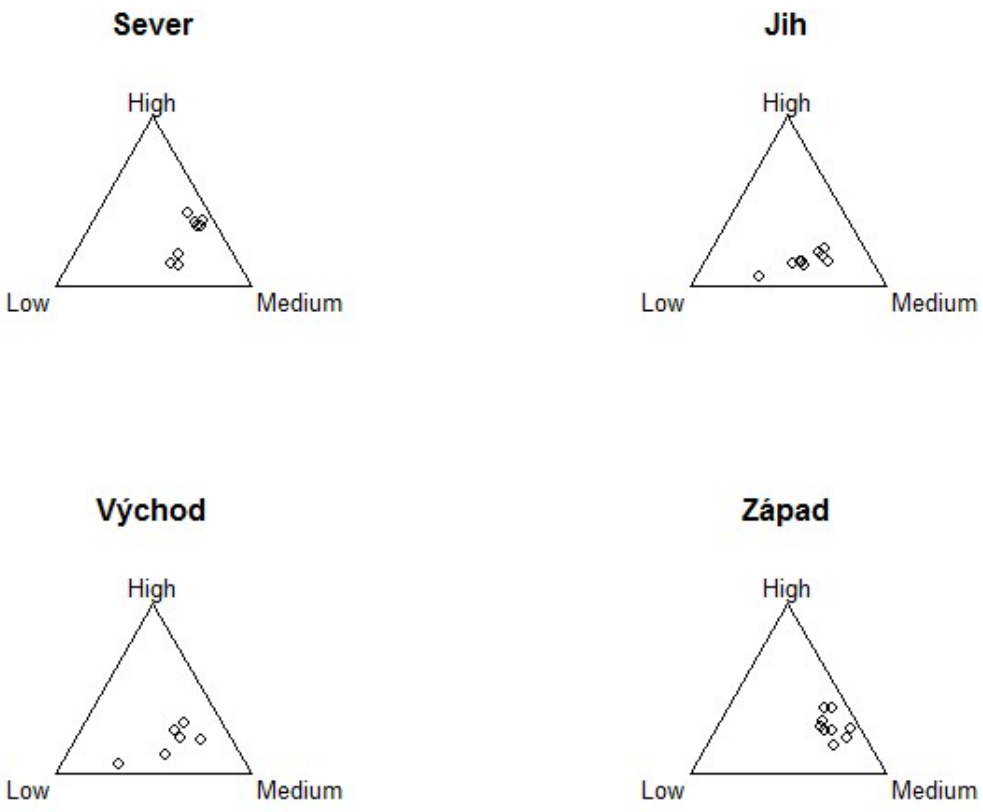
Obrázek 2.4: Spokojenost se životem v různých částech Evropy

```
> plot(north, col="#0099ff", pch=16)
# Graf kompozic north, barva modrá (col), značka puntík (pch)
> par(new=TRUE)              # Přidání dalších kompozic do grafu
> plot(south, col="#ff9900", pch=16)
> par(new=TRUE)
```

```
> plot(east, col="#ff3333", pch=16)
> par(new=TRUE)
> plot(west, col="#009933", pch=16)
> legend("topleft",                              # Umístění legendy
        c("Sever","Jih","Východ","Západ"),    # Položky legendy
         pch=16,col=c("#0099ff","#ff9900","#ff3333","#009933"))
```

Dále si vytvoříme matici ternárních diagramů pro jednotlivé části Evropy (obrázek 2.5). Matici zadefinujeme příkazem `par(mfrow=c(2, 2))`, potom postupně zadáváme ternární diagramy částí Evropy. Např. pro severní Evropu píšeme:

```
> plot(north, title(main = "Sever"))
```



Obrázek 2.5: Spokojenost se životem v různých částech Evropy

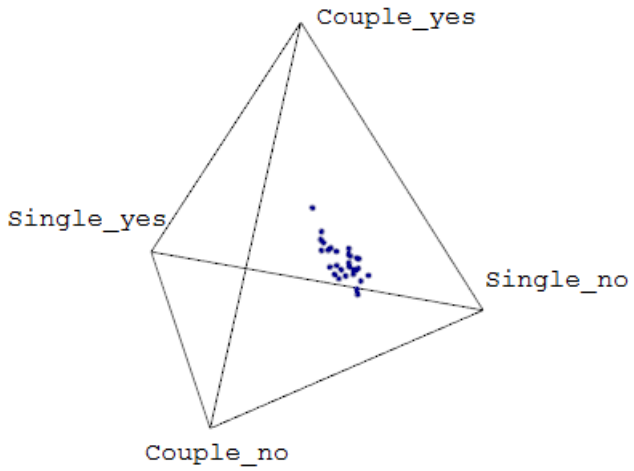
Již na první pohled vidíme, že jsou lidé spokojenější v západní a severní Evropě. U severu si navíc můžeme všimnout, že jsou zde 2 skupiny. První skupinou

jsou Norsko, Švédsko, Finsko, Dánsko a Island, u kterých hodnoty patří k nejlepším mezi všemi státy. Druhou skupinu tvoří Estonsko, Lotyšsko a Litva, bývalí členové SSSR, které bychom podle hodnot zařadili spíše mezi východní Evropu, ačkoliv bývají spojovány s Evropou severní.

2.1.2 Tetraedr

Obdobou ternárního diagramu pro čtyři složky je pravidelný čtyřstěn (tetraedr). V R ho zobrazíme s pomocí knihovny „rgl“ (a samozřejmě i knihovny „compositions“) tímto příkazem:

```
> plot3D(acomp(household),
        cex=3.5,col="darkblue",        # Velikost a barva bodů
        lwd=4, axis.col="black")            # Šířka a barva os
```



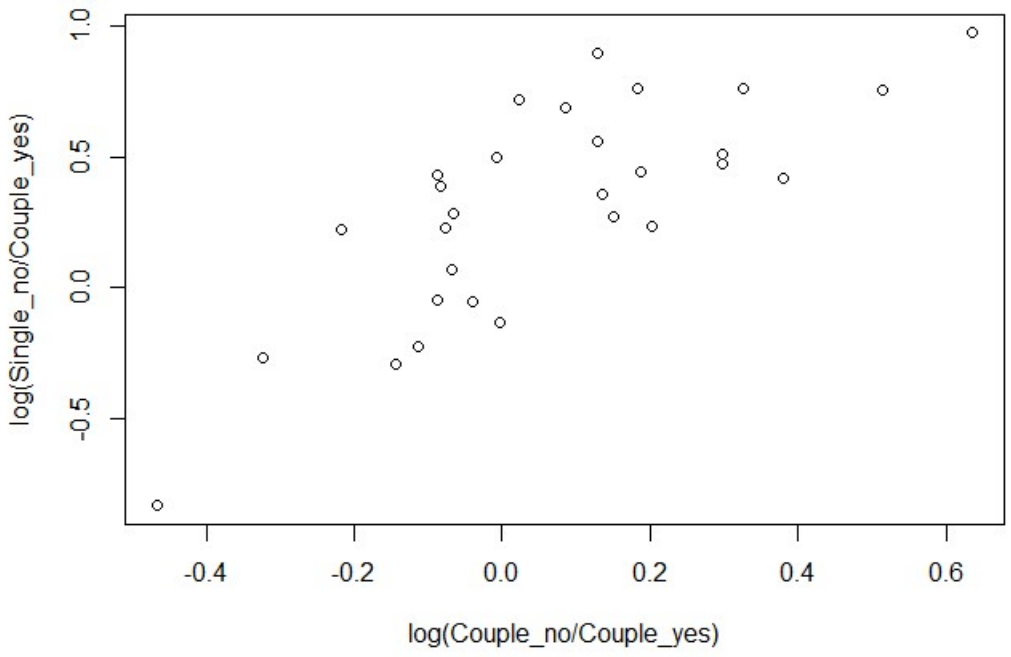
Obrázek 2.6: Zobrazení čtyřsložkové podkompozice vybrané z datového souboru `household`

Graf se otevře v interaktivním prostředí, které jím umožňuje otáčet. Jelikož se jedná o graf pro čtyři složky, automaticky se vybraly první čtyři. Výběr složek můžeme nastavit pomocí parametru `parts`. Při pohledu na obrázek 2.6 však musíme připustit, že je tetraedr oproti ternárnímu diagramu nepřehledný. Nezískáme z něj bližší představu o datech.

2.1.3 Bodový logpodílový graf

Použití klasického bodového (nebo také korelačního) grafu, který vyjadřuje závislost mezi dvěma proměnnými, není pro kompoziční data vhodné. Tento graf totiž nesplňuje základní principy pro nakládání s kompozičními daty (invarianci vůči změně měřítka, perturbaci a podkompoziční soudržnost). Proto se seznámíme s bodovým grafem logaritmu podílu dvou složek proti logaritmu podílu dalších dvou (jiných) složek. Taková reprezentace splňuje hlavní principy práce s kompozicemi a nezávisí na konkrétním měřítku v němž byla kompoziční data vyjádřena. Navíc ji lze považovat (po normalizaci $1/\sqrt{2}$) za graf dvou ortonormálních souřadnic. Někdy volíme do jmenovatele zlomků stejnou složku, pak se jedná o bodový graf alr souřadnic (podkapitola 1.5.3).

Pro vytvoření konkrétního bodového logpodílového grafu využijeme datový soubor `household`. Do grafu 2.7 vykreslíme pomocí funkce `plot` logaritmus podílu složek `Single_no` a `Couple_yes` v závislosti na logaritmu podílu složek `Couple_no` a `Couple_yes`.



Obrázek 2.7: Logpodílový graf složek z datového souboru `household`

Je třeba zadat:

```
> Single_no = household[,"Single_no"]      # Označení proměnných
> Couple_yes = household[,"Couple_yes"]
> Couple_no = household[,"Couple_no"]
> plot(log(Single_no/Couple_yes)~log(Couple_no/Couple_yes))
```

Z grafu 2.7 je zřejmé, že podíl bezdětných jedinců žijících o samotě ku párům s dětmi souvisí s příslušným podílem s bezdětnými páry v daném státě.

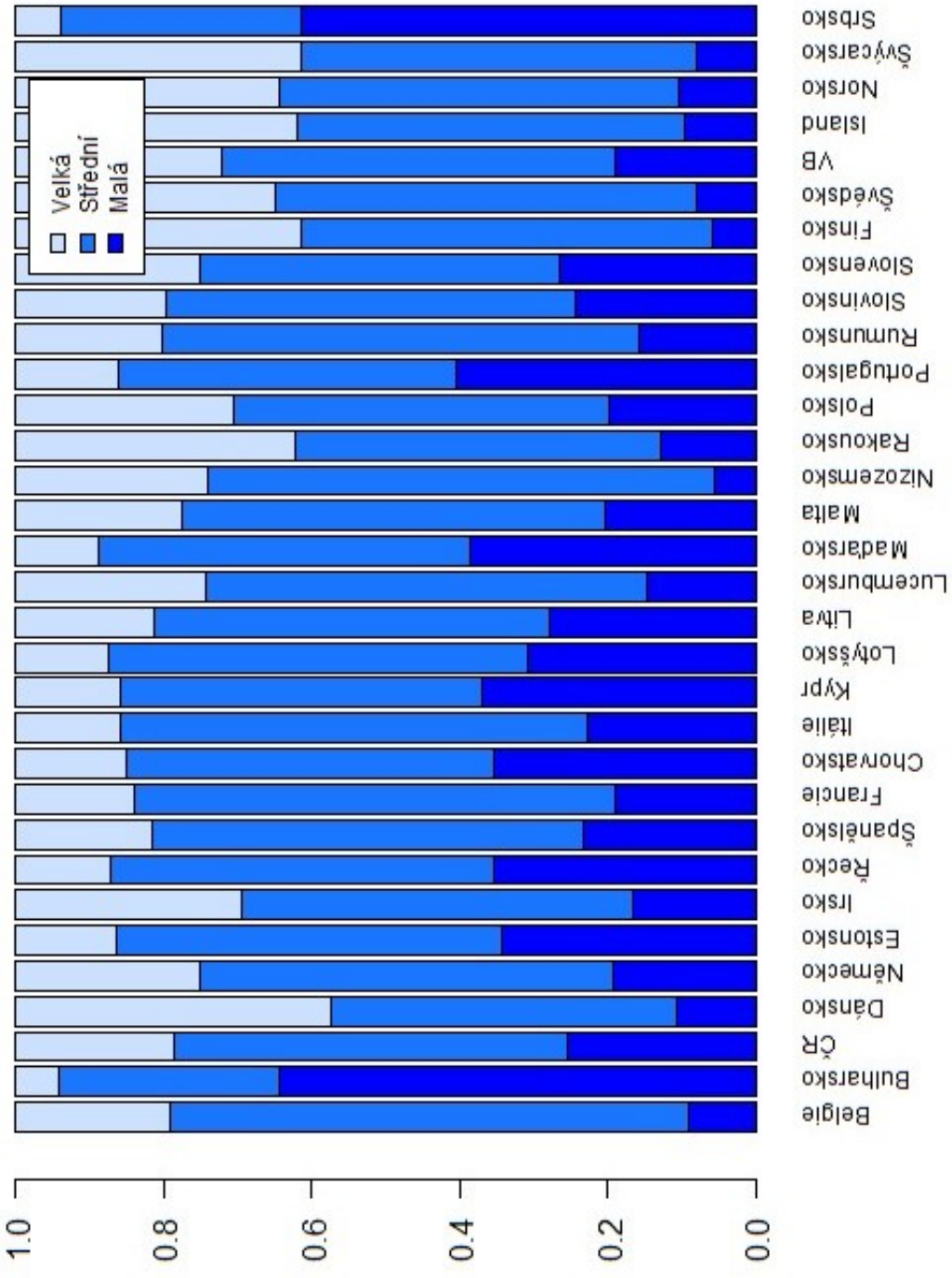
2.1.4 Sloupcové a koláčové grafy

Další možností pro zobrazení kompozic je sloupcový graf, což si názorně ukážeme na datovém souboru `satisfaction`. Každý stát je reprezentován jedním sloupečkem, který se skládá ze tří částí, jejichž velikosti jsou dány podílem daných složek na celku (který je roven 1, neboť jsme použili příkaz `acomp()`). Graf z obrázku 2.8 vytvoříme příkazem `barplot` :

```
> barplot(acomp(satisfaction),                    # Datový soubor
          col=c("#0000ff","#1a75ff","#cce0ff"),          # Barvy
          las=2,                              # Umístění popisků os
          cex.names=0.75,                 # Velikost popisků osy x
          xlim=c(0,35),                                # Šířka osy x
          legend.text=c("Low", "Medium", "High"),      # Legenda
          args.legend=list(cex=0.75))          # Velikost legendy
```

Jednotlivé kompozice můžeme znázornit také koláčovým grafem. Koláčové grafy ale nejsou doporučeny pro kompozice o více než dvou složkách, neboť lidské oko špatně porovnává úhly v těchto grafech. Zkusíme si zobrazit koláčový graf pro Českou republiku týkající se spokojenosti se životem (obrázek 2.9), k čemuž užijeme příkaz `pie()`.

```
> pie(acomp(satisfaction["ČR",]),             # Výběr kompozice ČR
      main="Spokojenost se životem v ČR",                # Nadpis
      col=c("#0000ff","#1a75ff","#cce0ff")                # Barvy
```



Obrázek 2.8: Sloupcový graf datového souboru `satisfaction`

Obrázek 2.9: Koláčový graf spokojenosti se životem v České republice

2.2 Popisná statistika

V klasických aplikacích provádíme popisnou analýzu vícerozměrných dat jako jednorozměrný nebo dvourozměrný popis marginálních proměnných. Složky kompozičních dat jsou ale mezi sebou svázány, tudíž musíme nejdříve kompozici vyjádřit v souřadnicích. Dále pak provádíme klasickou jednorozměrnou analýzu jednotlivých souřadnic.

2.2.1 Centrum

Aitchison [1] upozornil, že smysluplné charakteristiky polohy, variability a kovariance by měly splňovat některé vlastnosti:

1. Posunutím datového souboru (perturbací konstantní kompozicí) by se neměla změnit variabilita nebo kovariance mezi složkami. Mělo by dojít pouze k posunutí centra distribuce.

2. Pokud u datového souboru změníme měřítko (u kompozic pomocí operace mocninná transformace), tak jeho centrum získáme rovněž jen stejnou změ-

nou měřítka. Zatímco u variability a kovariance se změna měřítka projeví v druhé mocnině.

Definice 2.1 *Centrem nebo také kompozičním průměrem datového souboru* $\boldsymbol{X}$ *o* N *pozorováních a* D *složkách je kompozice*

$$\overline{\boldsymbol{x}} = \frac{1}{N} \odot \bigoplus_{n=1}^{N} \boldsymbol{x}_n = clr^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} clr(\boldsymbol{x}_n)\right) = \mathcal{C}\left[\exp\left(\frac{1}{N}\sum_{n=1}^{N}\ln(\boldsymbol{x}_n)\right)\right],$$

kde $\boldsymbol{x}_n$ *je n-té pozorování, v n-tém řádku* $\boldsymbol{X}$*.*

Centrum spočteme příkazem `mean(x)`, přičemž `x` je třídy `acomp`. Zkusme spočítat centrum kompozičních dat `satisfaction`:

```
> satisfaction = acomp(satisfaction)
> mean(satisfaction)
    Low  Medium    High
 0.2106  0.5683  0.2211
attr(,"class")
[1] acomp
```

Vidíme, že průměrná evropská země má něco málo přes polovinu středně spokojených lidí. Kolem dvacítky se pohybuje procento málo a velmi spokojených lidí.

Při analýze reálných vícerozměrných dat je typické centrovat data odečtením jejich průměru. U kompozičních dat to provedeme pomocí perturbace inverzí centra, tj. $\boldsymbol{X}^* = \boldsymbol{X} \ominus \overline{\boldsymbol{x}}$.

```
> mean(satisfaction-mean(satisfaction))
    Low  Medium    High
 0.3333  0.3333  0.3333
attr(,"class")
[1] acomp
```

Průměrem centrovaného datového souboru je neutrální prvek simplexu $\mathbf{1}$, vektor stejných hodnot pro každou složku.

2.2.2 Metrický rozptyl a směrodatná odchylka

Existují nejrůznější charakteristiky variability pro kompoziční data. Jako míru celkové variability souboru můžeme použít např. metrický rozptyl (také známý jako celkový rozptyl):

$$\mathrm{mvar}(\boldsymbol{X}) = \frac{1}{N-1}\sum_{n=1}^{N} d_A^2(\boldsymbol{x}_n, \overline{\boldsymbol{x}}),$$

tj. průměrný čtverec vzdálenosti od centra, s upravenými stupni volnosti stejně jako u konvenčního rozptylu. Celkový rozptyl získáme příkazem `mvar(x)`, kde `x` je třídy `acomp`.

Celkový rozptyl datového souboru `satisfaction` je:

```
> mvar(satisfaction)
[1] 0.6550
```

Pro kvantitativní interpretaci výsledků je lepší si ještě spočítat metrickou směrodatnou odchylku, neboť je vyjádřena ve stejných jednotkách jako sledovaná data:

$$\mathrm{msd}(\boldsymbol{X}) = \sqrt{\frac{1}{D-1}\mathrm{mvar}(\boldsymbol{X})}.$$

```
> msd(satisfaction)
[1] 0.5723
```

Výsledku rozumíme jako určité průměrné variabilitě dat. Kdybychom měli k dispozici více datových souborů, mohli bychom pomocí hodnoty metrické směrodatné odchylky porovnávat jejich variabilitu.

Na základě podkompoziční soudržnosti nesmí být metrický rozptyl podkompozice větší než je u kompozice původní, ze které byla podkompozice vybrána.

2.2.3 Variační a varianční matice

Metrický rozptyl nedává žádnou informaci o vztazích mezi složkami. Nemůžeme ale použít klasickou kovarianci (popř. korelaci) kvůli problému tzv. falešné korelace (viz podkapitola 1.1). Proto definujeme variační matici:

Definice 2.2 *Variační maticí (náhodné) kompozice* $\boldsymbol{x} = [x_1, \ldots, x_D]$ *z* $\mathbb{S}^D$ *nazveme matici* D^2 *prvků, definovaných jako*

$$\tau_{ij} = \mathrm{var}\left(\ln \frac{x_i}{x_j}\right)$$

a odhadovaných z výběru kompozice $\boldsymbol{x} = [x_{n1}, \ldots, x_{nD}], \quad n = 1, \ldots, N$ *hodnotou*

$$\widehat{\tau}_{ij} = \left[\frac{1}{N-1}\sum_{n=1}^{N} \ln^2 \frac{x_{ni}}{x_{nj}}\right] - \ln^2 \frac{\overline{x}_i}{\overline{x}_j},$$

kde N *je počet pozorování.*

Odhad variační matice datového souboru `x` dostaneme příkazem `variation(x)`. Jedná se o symetrickou matici, neboť $\ln(a/b) = -\ln(b/a)$ a $\mathrm{var}(-c) = \mathrm{var}(c)$ pro kladná reálná čísla a, b a náhodnou veličinu c. Malé τ_{ij} znamená malý rozptyl $\ln(x_i/x_j)$, a tedy i těsnou závislost (proporcionalitu) mezi x_i a x_j. Čím více se blíží hodnota nule, tím těsnější je závislost mezi danými složkami. Aitchison [2] navrhuje transformovat variační matici:

$$\rho_{ij} = \exp(-\tau_{ij}^2/2),$$

hodnoty se pak mají interpretovat jako korelační koeficient, tzn. že hodnoty se budou pohybovat na intervalu $\langle 0, 1\rangle$. Charakteristika ale na rozdíl od korelačního koeficientu nevyjadřuje míru těsnosti lineárního vztahu mezi složkami [5], nýbrž, jak již bylo zmíněno, jejich proporcionalitu.

Podívejme se na variační matici datového souboru `satisfaction`.

```
> variation(satisfaction)
            Low  Medium   High
Low     0.000   0.566  1.201
Medium  0.566   0.000  0.198
High    1.201   0.198  0.000
```

Největší variabilitu můžeme spatřit mezi složkami `Low` a `High`, tedy mezi velkou a malou spokojeností. Je to zřejmě způsobeno rozdíly mezi vyspělými a rozvojovějšími zeměmi. Dále pak zkusme použít zmíněnou transformaci na variační matici souboru `household`:

```
> variation_household = variation(household)
> exp(-variation_household^2/2)
```

	Single_yes	Single_no	Couple_yes	Couple_no	Other_yes	Other_no
Single_yes	1.000	0.982	0.974	0.978	0.690	0.736
Single_no	0.982	1.000	0.986	0.997	0.705	0.784
Couple_yes	0.974	0.986	1.000	0.998	0.927	0.947
Couple_no	0.978	0.997	0.998	1.000	0.841	0.897
Other_yes	0.690	0.705	0.927	0.841	1.000	0.997
Other_no	0.736	0.784	0.947	0.897	0.997	1.000

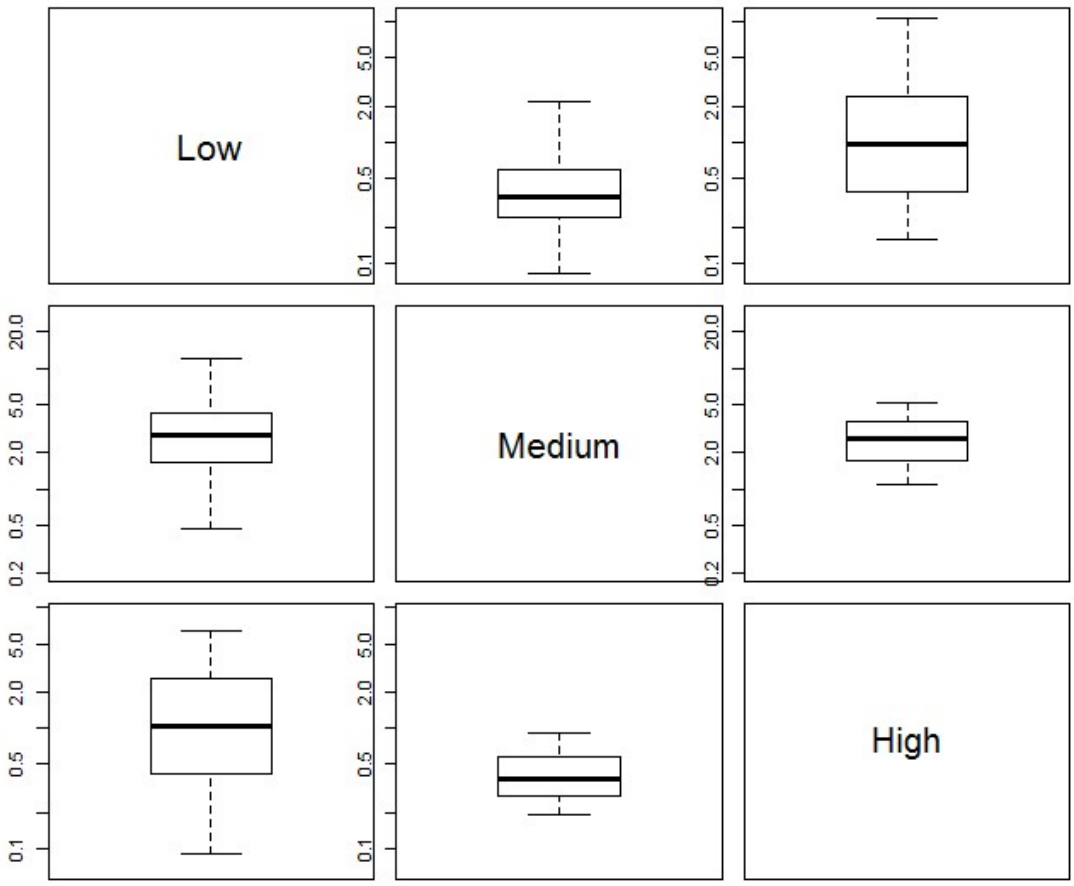
Tentokrát znamená těsnou závislost x_i a x_j hodnota, která se blíží 1. Vidíme, že variabilita se projevuje spíše mezi jednotlivými formami vztahů, než mezi tím, jestli má domácnost děti. Disperze se pak ukazuje zvláště v podílech mezi typem soužití `Other` a zbývajícími typy. Proto se zamyslíme, jaké typy domácností sem můžou patřit, neboť skupina „Ostatní“ je dosti nekonkrétní. Jako příklad si můžeme uvést domácnosti, ve kterých spolu bydlí 3 generace dohromady, nebo skupinu alespoň 2 dospělých lidí, kteří sdílejí domácnost, ale netvoří pár.

Další možností je grafické zobrazení variability dat v podobě matice boxplotů pro jednotlivé logaritmy podílů složek. Stačí zadat `boxplot(x)`, kde `x` je třídy `acomp`. Na obrázku 2.10 se můžeme podívat na boxploty pro datový soubor `satisfaction`.

Variační matice se velmi dobře interpretuje, někdy se ale potřebujeme více přiblížit klasickému popisu variability v podobě varianční matice. Pro takové případy zavedeme *clr varianční matici*, což je varianční matice clr souřadnic kompozice. V R ji získáme pomocí funkce `var` aplikované na clr souřadnice. Prvky varianční matice $\widehat{\Sigma}_{ij}$ kompozice $\boldsymbol{x} = [x_1, \ldots, x_D]$ z $\mathbb{S}^D$ tedy definujeme jako

$$\widehat{\Sigma}_{ij} = cov(clr_i(\boldsymbol{x}), clr_j(\boldsymbol{x})) = \left[\frac{1}{N}\sum_{n=1}^{N} \ln\frac{x_{ni}}{g(\boldsymbol{x}_n)} \cdot \ln\frac{x_{nj}}{g(\boldsymbol{x}_n)}\right] - \ln\frac{\bar{x}_i}{g(\bar{\boldsymbol{x}})} \cdot \ln\frac{\bar{x}_j}{g(\bar{\boldsymbol{x}})},$$

kde funkce g značí geometrický průměr, $\bar{\boldsymbol{x}} = [\bar{x}_1, \cdots, \bar{x}_D]$ je centrum (definice 2.1) a N označuje počet pozorování. Protože složky clr souřadnic dávají součet 0, tak také jednotlivé řádky a sloupce varianční matice mají součet 0. Každý prvek matice zahrnuje ve svém výpočtu všechny složky kompozice (neboť součástí vztahů pro clr souřadnice je i $g(\boldsymbol{x})$). Variační matice slouží pouze jako matematický nástroj. Nemůže být přímo interpretována nebo přepočtena na korelační matici, kvůli negativnímu vychýlení (viz podkapitola 1.1).



Obrázek 2.10: Boxploty logaritmů podílů mezi složkami datového souboru `satisfaction`

2.2.4 Standardizace

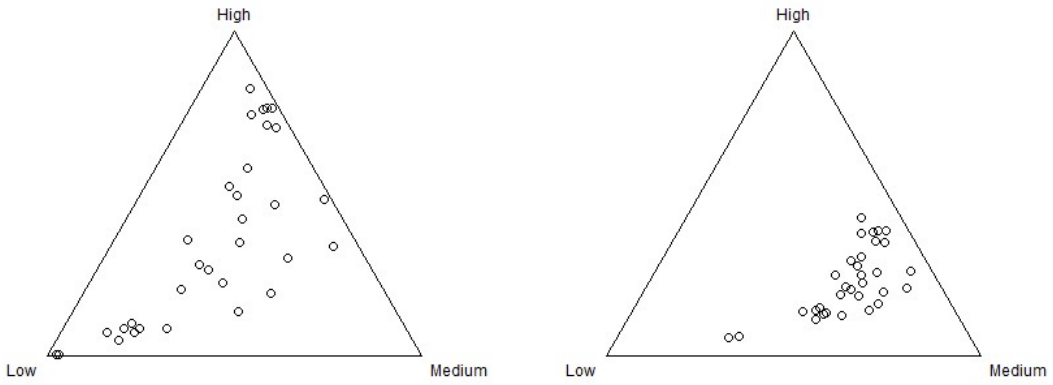
Často v aplikacích provádíme standardizaci dat tím, že od nich odečteme průměr a poté je vydělíme směrodatnou odchylkou. Činíme tak, abychom získali bezrozměrné veličiny, které můžeme srovnávat mezi sebou i mezi různými datovými soubory. Je zřejmé, že u kompozičních dat by tento postup neměl smysl. Za prvé jsou kompoziční data přirozeně bezrozměrná. Za druhé bychom odečtením průměru mohli dostat záporné hodnoty, a tím ztratit základní interpretaci kompozičních dat. Proto místo součtu vektorů a násobení vektoru skalárem použijeme opět perturbaci a mocninnou transformaci:

$$\boldsymbol{Z} = \frac{1}{\sqrt{\mathrm{mvar}(\boldsymbol{X})}} \odot (\boldsymbol{X} \ominus \bar{\boldsymbol{x}}).$$

V R kompozici `x` standardizujeme příkazem:

`> scale(x, center = TRUE, scale = TRUE)`.

Nejčastěji je nutné standardizovat data, pokud jsou koncentrována na kraji nebo v rohu ternárního diagramu. Dostaneme pak lepší grafickou představu o datech. Jako příklad jsme standardizovali datový soubor satisfaction (obrázek 2.11).



Obrázek 2.11: Datový soubor `satisfaction` po standardizaci (vlevo) a před standardizací (vpravo)

2.2.5 Oblast spolehlivosti a predikční oblast

Nyní si zobrazíme centrum a disperzi dat graficky. Centrum `x` třídy `acomp` přidáme do ternárního diagramu příkazem `plot(x, add=TRUE)`. Variabilitu znázorníme pomocí funkce `ellipses(mean, var, r)`, kde za `var` volíme clr varianční matici. Elipsa může odpovídat oblasti spolehlivosti pro centrum se spolehlivostí $p = 1 - \alpha$, pokud zvolíme poloměr `r = r1` podle Fisherova $\mathscr{F}$ rozdělení o $D - 1$ a $N - D + 1$ stupních volnosti:

$$r_1 = \sqrt{\frac{D-1}{N-D+1} \cdot \mathscr{F}_p(D-1, N-D+1)}.$$

Platí to ale pouze tehdy, pokud data mají normální rozdělení na simplexu (tzn. normální rozdělení v libovolných logpodílových souřadnicích) nebo pokud je počet pozorování N vysoký (pak se totiž uplatní centrální limitní věta). Dále zkonstruujeme predikční oblast se spolehlivostí $p = 1 - \alpha$ za předpokladu, že data se řídí normálním rozdělením na simplexu se známou clr varianční maticí

(v praxi ovšem přesto typicky určenou přímo z dat). Potom má elipsa poloměr r_2 daný χ^2 rozdělením o $D-1$ stupních volnosti:

$$r_2 = \sqrt{\chi^2_p(D-1)}.$$

Obvykle volíme $\alpha = 0{,}05$, tedy konstruujeme oblasti se spolehlivostí 95 %. Nakonec ještě demonstrujeme konstrukci oblastí na datovém souboru `satisfaction`.

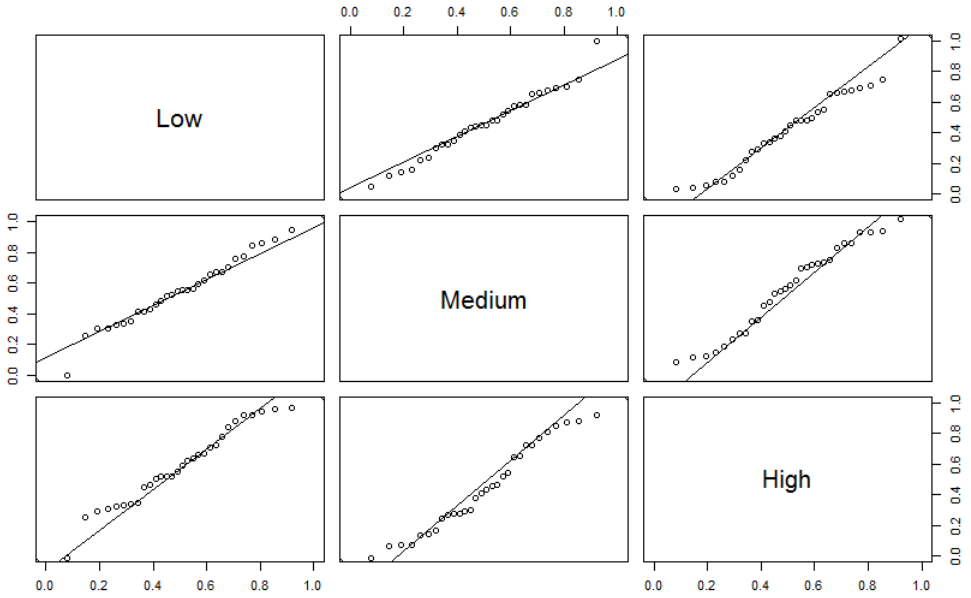
Nejprve však otestujeme normalitu, abychom splnili předpoklady pro obě oblasti. Budeme testovat nulovou hypotézu, že se datový soubor řídí normálním rozdělením na simplexu. Jak již bylo zmíněno dříve, to je ekvivalentní s hypotézou, že se datový soubor vyjádřený v logpodílových souřadnicích řídí mnohorozměrným normálním rozdělením. Narazíme však na problém, že neexistuje ucelený test normality pro vícerozměrná data. Jsme vždy schopni normalitu ověřit jen v některých směrech. Často proto bereme vícerozměrná data za „dostatečně normální“, pokud splňují testy normality pro jednotlivé marginální proměnné. U kompozic ale musíme uvažovat alespoň dvojici (původních) proměnných, protože samostatná proměnná nemá z podstaty kompozičních dat smysl.

Testujme tedy normalitu logaritmů podílů jednotlivých složek. Jednoduše toho dosáhneme voláním funkce `qqnorm.acomp(X, alpha)` (kde `X` je kompoziční soubor dat třídy `acomp`), která vykreslí Q-Q graf pro logaritmy podílů jednotlivých dvojic složek. Q-Q graf porovnává teoretické kvantily normálního rozdělení se skutečnými kvantily z dat. Jestliže výsledné body mají alespoň přibližně lineární trend, považujeme rozdělení dat za dobře aproximovatelné normálním rozdělením. Pokud zadáme za nepovinný parametr `alpha` hladinu významnosti, provede se i *Shapirův-Wilkův test normality* pro jednotlivé logaritmy podílů na hladině významnosti `alpha` vydělené číslem $D(D-1)/2$ (což je tzv. *Bonferroniho korekce*). Zamítnutí nulové hypotézy je vyznačeno červeným vykřičníkem nad Q-Q grafem.

Pro datový soubor `satisfaction` vidíme Q-Q grafy na obrázku 2.12, za parametr `alpha` jsme zvolili hodnotu 0,05. Protože v celé matici grafů není žádný červený vykřičník, nulové hypotézy o normalitě logaritmů podílů jednotlivých složek nemůžeme zamítnout. Do následující tabulky si ještě zobrazíme p-hodnoty z Shapirova-Wilkova testu pro jednotlivé proměnné (tedy pro logaritmy podílů příslušných složek). Protože žádná z nulových hypotéz nebyla zamítnuta, jsou p-hodnoty větší než daná hladina významnosti.

`Low`	$0,8523$	$0,3429$
$0,8523$	`Medium`	$0,2164$
$0,3429$	$0,2164$	`High`

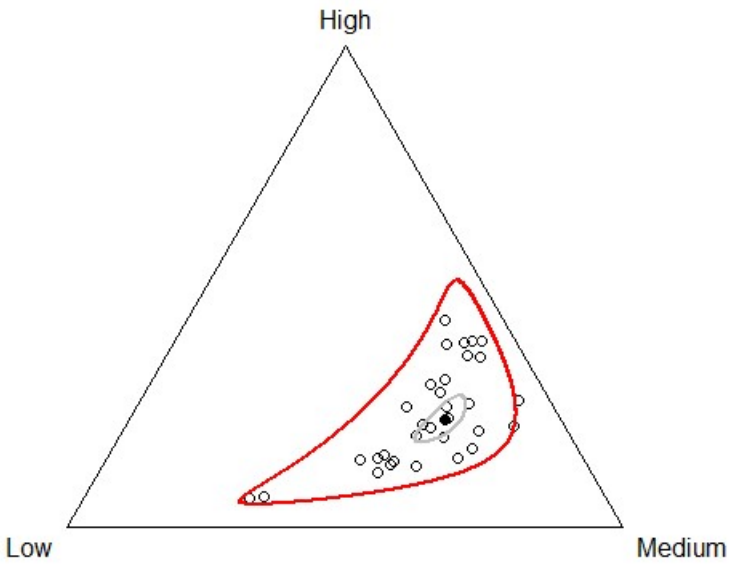
Tabulka p-hodnot z Shapirova-Wilkova testu pro logaritmy podílů jednotlivých složek



Obrázek 2.12: Q-Q graf pro datový soubor `satisfaction` s provedením Shapirova-Wilkova testu na hladině významnosti $\alpha = 0{,}05/3$

Nyní už můžeme přejít k samotné konstrukci oblastí. Obrázek 2.13 se zobrazením šedé oblasti spolehlivosti a červené predikční oblasti získáme následujícím kódem:

```
> satisfaction = acomp(satisfaction)              # Třída acomp
> mn = mean(satisfaction)                              # Centrum
> vr = var(satisfaction)                  # Clr varianční matice
> df1 = ncol(satisfaction)-1                    # Stupně volnosti
> df2 = nrow(satisfaction)-ncol(satisfaction)+1
> r1 = sqrt( qf(p=0.95, df1,df2)*df1/df2 )     # r1 podle vzorce
> r2 = sqrt( qchisq(p=0.95, df=df1) )          # r2 podle vzorce
> plot(satisfaction)                          # Ternární diagram
> plot(mn, pch=19,add=TRUE)           # Centrum jako plná tečka
> ellipses(mean=mn,var=vr,r=r1,col="gray",lwd=2)    # 1. elipsa
```



Obrázek 2.13: 95% oblast spolehlivosti (šedě) pro centrum (plný puntík) a 95 % predikční oblast pro celý statistický soubor (červeně) v ternárním diagramu

```
> ellipses(mean=mn,var=vr,r=r2,col="red",lwd=2)      # 2. elipsa
```

2.2.6 Marginály

Jelikož nejsme schopni vidět více než tři dimenze, vyskytla se otázka, jak zobrazit vícesložkové kompozice. U běžných dat se situace řeší maticí bodových grafů pro každou dvojici proměnných. Nicméně jsme již zmínili (viz podkapitola 2.1.3), že použití standardního bodového grafu pro kompoziční data není vhodné. Řešením je tedy matice ternárních diagramů zahrnujících vždy dvojici proměnných podle daného řádku a sloupce. Chybí ale třetí proměnná, kterou je třeba dodefinovat, budeme ji nazývat *marginálem*. Matici ternárních diagramů kompozičního datového souboru `x` zobrazíme příkazem `plot(x)` a marginál definujeme nepovinným parametrem `margin`. Při nastavování tohoto parametru máme tři možnosti:

1. `margin = "acomp"` počítá třetí složku jako geometrický průměr všech složek kromě daných dvou, které jsou určeny řádkem a sloupcem matice. Jedná se o implicitní hodnotu.

2. `margin = "rcomp"` počítá třetí složku jako amalgamaci všech složek kromě daných dvou, které jsou určeny řádkem a sloupcem matice.

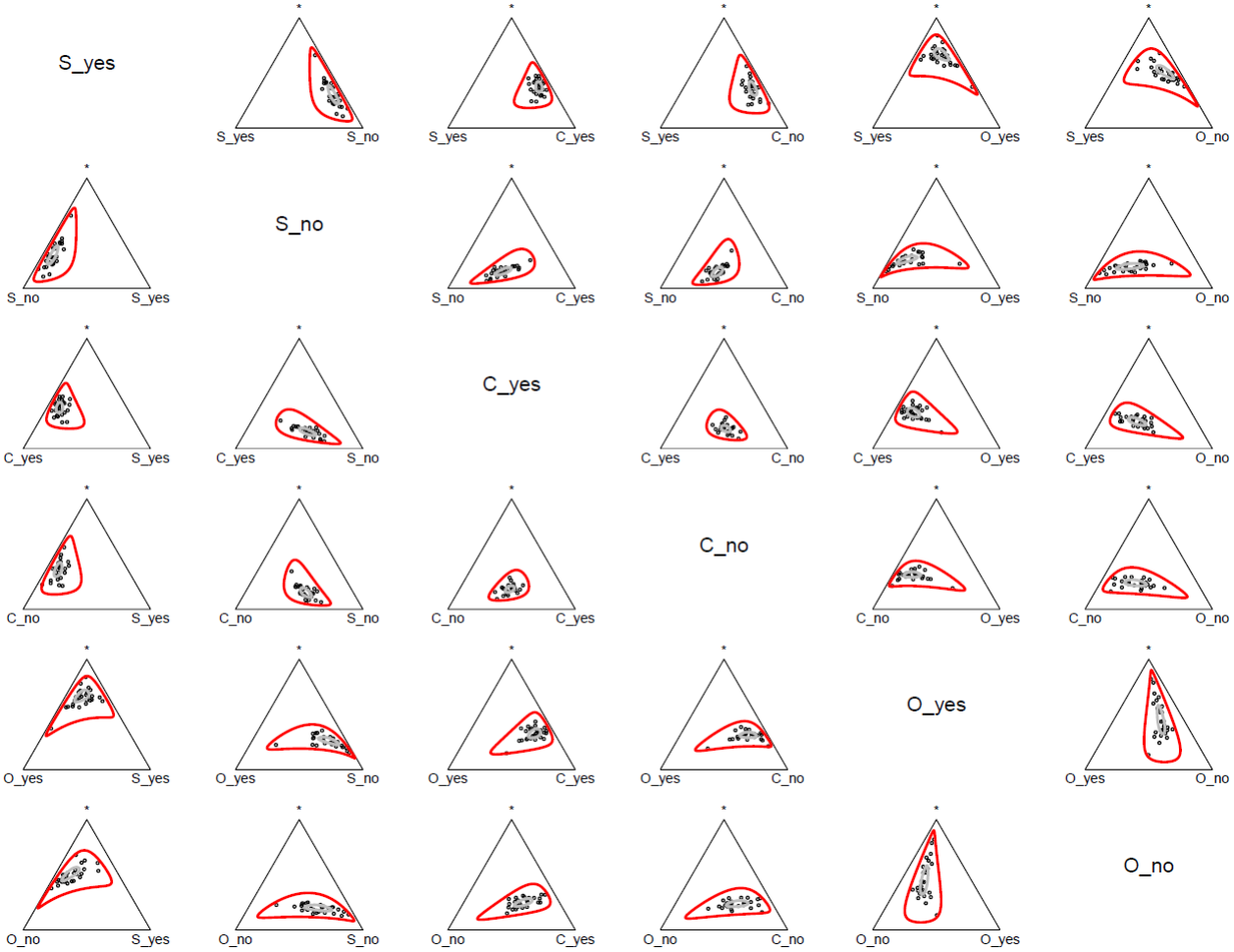
3. Poslední možností je položit fixně parametr `margin` roven jedné ze složek kompozice. Matice pak bude mít $D - 1$ řádků a sloupců.

Výhodou první možnosti, kterou dostaneme i bez specifikace parametru, je její konzistence s Aitchisonovou geometrií na simplexu. Můžeme proto do grafu přidat další informace (jako například centrum či oblast spolehlivosti). Je třeba dbát opatrnosti při použití amalgamace, neboť je nelineární operací v Aitchisonově geometrii a porušuje principy kompoziční analýzy dat.

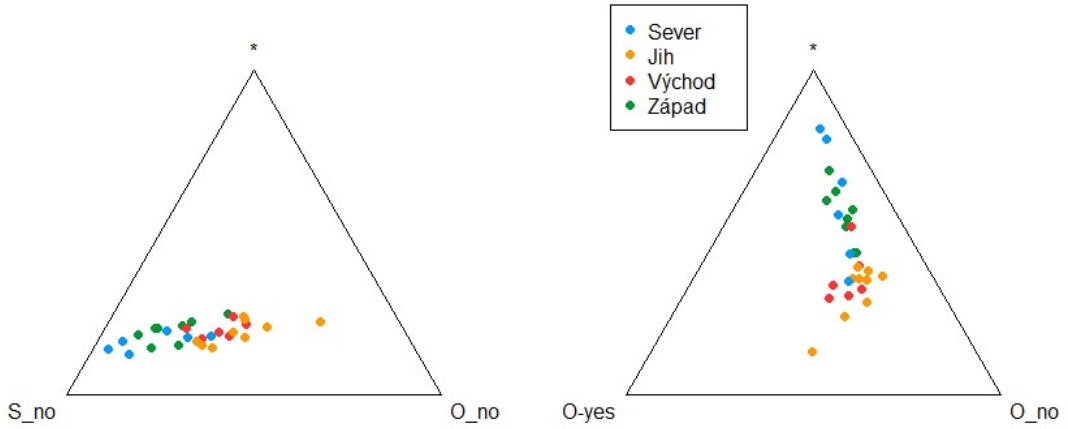
Lepší představu o tom, jak matice ternárních diagramů vypadá, si uděláme při pohledu na obrázek 2.14 pro datový soubor `household`. Kvůli velikosti matice jsme pro označení typů domácností v proměnných použili pouze počáteční písmena (tj. `S` jako `Single`, `C` jako `Couple` a `O` jako `Other`). Zvolili jsme implicitní hodnotu parametru `margin`, proto jsme si mohli dovolit zobrazit i oblast spolehlivosti pro centrum a predikční oblast pro všechna data. Normalitu dat jsme ověřili podobně jako v podkapitole 2.2.5.

Matice je rozsáhlá, zaměříme se tedy na některé detaily. Opět vidíme malou variabilitu (silnou proporcionalitu) mezi složkami `C_yes` a `C_no` (příp. `S_yes` a `S_no`), což znamená, že nejsou z hlediska podílů mezi složkami rozdíly v tom, jestli mají jednotlivé typy domácností děti či ne. Při pohledu na protáhlý tvar elipsy mezi složkami `S_no` a `C_yes` vidíme, že existují rozdíly mezi tradičními státy (upřednostňujícími život v rodině s dětmi) a státy ovlivněnými novým trendem života „single“, kdy člověk dává přednost kariéře před rodinným životem. Větší variabilitu vidíme hlavně v posledním sloupci matice, pojďme se tedy blíže podívat na dva konkrétní ternární diagramy (obrázek 2.15), do kterých jsme si navíc barevně vyznačili jednotlivé oblasti Evropy.

Z obrázku 2.15 je zřejmé, že lidé na jihu Evropy dávají přednost bydlení ve společných domácnostech (více generací nebo více nepříbuzných lidí v jedné domácnosti). Opačným pólem jsou severané, kteří tyto typy bydlení vyhledávají mnohem méně a nejvíce ze všech skupin preferují bydlení o samotě bez dětí (typ `Single_no`).



Obrázek 2.14: Matice ternárních diagramů pro datový soubor `household` pro parametr `margin = "acomp"` se zobrazenými oblastmi spolehlivosti a predikčními oblastmi



Obrázek 2.15: Vybrané ternární diagramy z matice na obrázku 2.14 s barevným vyznačením oblastí Evropy

2.2.7 Bilance a CoDa-dendrogram

V podkapitole 1.5.3 jsme u ilr souřadnic zmínili, že lze vybrat ortonormální bázi tak, aby byly souřadnice dobře interpretovatelné. K tomu používáme postup nazývaný *postupné binární dělení (PBD)* a získáme při něm $D - 1$ souřadnic označovaných jako *bilance* (neboli *rovnováhy*). Nyní si ukážeme, jak se PBD provede. Na začátku uvažujeme všechny složky kompozice a rozdělíme je do dvou skupin. Poté vezmeme jednu z těchto skupin a opět ji rozdělíme na dvě podskupiny. Tento postup aplikujeme rekurzivně, dokud neobsahují všechny skupiny jen jedinou složku. Abychom získali dobře interpretovatelné souřadnice, je třeba složky dělit do skupin podle jejich vzájemné příbuznosti a společných vlastností. K správnému určení dělícího kritéria je třeba mít jistou zkušenost a znalost studovaného problému. Zkusme sestavit bilance pro datový soubor `household`.

Označení proměnných je možné si připomenout v příloze na straně 54. Pro jednoduchost budeme v této kapitole používat označení složek kompozice x_1 až x_6. Nejprve rozdělíme složky na skupinu ostatní (složky x_5 a x_6) a zbytek necháme v druhé skupině. Poté rozdělíme větší skupinu na jednotlivce (složky x_1 a x_2) a páry (složky x_3 a x_4). Pokračujeme rozdělením skupiny ostatních na domácnosti s dětmi a bez dětí. V posledních dvou krocích provedeme totéž pro jednotlivce a páry. Při každém dělení jednu skupinu označíme znaménkem $+$, druhou $-$. Zapisujeme do tabulky:

	Složka					
Bilance	x_1	x_2	x_3	x_4	x_5	x_6
z_1	+	+	+	+	−	−
z_2	+	+	−	−		
z_3					+	−
z_4	+	−				
z_5			+	−		

Bilance z_i mezi dvěma skupinami složek označenými + a − je míra celkové relativní významnosti jedné skupiny oproti druhé. Získáme ji podle vztahu

$$z_i = \sqrt{\frac{rs}{r+s}} \ln \frac{(\prod_+ x_j)^{1/r}}{(\prod_- x_k)^{1/s}} \qquad i = 1, \cdots, D-1,$$

kde r je počet složek skupiny + a s je počet složek skupiny −.

Pro sestavené PBD kompozic `household` vypadají bilance následovně:

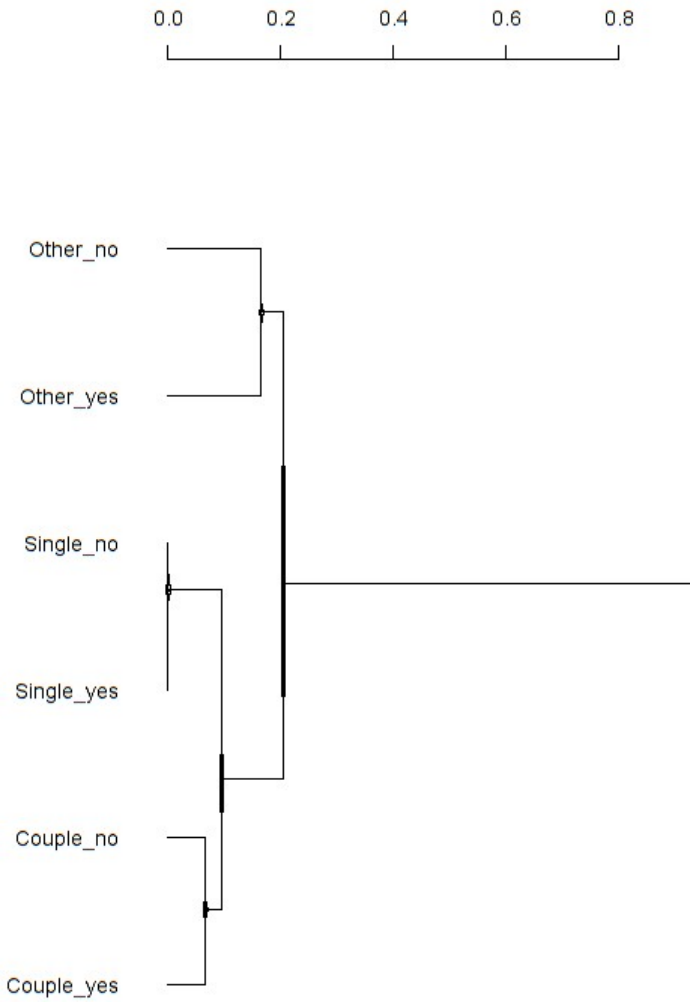
$$z_1 = \sqrt{\tfrac{8}{6}} \ln \frac{\sqrt[4]{x_1 x_2 x_3 x_4}}{\sqrt{x_5 x_6}}, \qquad z_2 = \ln \frac{\sqrt{x_1 x_2}}{\sqrt{x_3 x_4}}, \qquad z_3 = \tfrac{1}{\sqrt{2}} \ln \tfrac{x_5}{x_6},$$

$$z_4 = \tfrac{1}{\sqrt{2}} \ln \tfrac{x_1}{x_2}, \qquad z_5 = \tfrac{1}{\sqrt{2}} \ln \tfrac{x_3}{x_4}.$$

Nástrojem pro zobrazení bilancí je CoDa-dendrogram, akronym CoDa značí *compositional data* neboli kompoziční data. CoDa-dendrogram získáme zadáním příkazu `CoDaDendrogram = (X, signary)`, kde za `X` volíme datový soubor typu `acomp`, za parametr `signary` pak dosadíme znaménkovou matici určenou postupným binárním dělením. Vytvoříme ji z tabulky znamének PBD, když místo znaménka + napíšeme 1, znaménko − nahradíme číslem (−1) a prázdná místa nulou. Navíc uvažujeme sloupcové vektory. Pro vytvořené PBD kompozičních dat `household` dostaneme tuto znaménkovou matici:

```
> Signary = t(matrix( c( 1, 1, 1, 1,-1,-1,
                         1, 1,-1,-1, 0, 0,
                         0, 0, 0, 0, 1,-1,
                         1,-1, 0, 0, 0, 0,
                         0, 0, 1,-1, 0, 0 ),
                ncol=6,nrow=5,byrow=TRUE))
> CoDaDendrogram(X = household, signary = Signary)
```

Příkazem `CoDaDendrogram` se vykreslí CoDa-dendrogram z obrázku 2.16. Graf znázorňuje hierarchii, jak jsou složky spojeny do skupin podle PBD. Vertikální



Obrázek 2.16: Coda-dendrogram kompozic `household` daný bilancemi z_1 až z_5

úsečka spojující dvojici skupin představuje osu pro odpovídající bilanci. Implicitně každá osa zahrnuje interval $(-4, 4)$, ale je možno jej upravit změnou hodnoty parametru `range`. Průsečík vertikální a horizontální osy představuje průměr bilance. Variabilitu bilance zjistíme podle boxplotu umístěného na vertikální ose. Délka horizontální úsečky vycházející z boxplotu určuje, jakou část z celkové variability vysvětluje daná bilance. Podrobnější popis CoDa-dendrogramu najdeme například v [6]. Z obrázku 2.16 jasně vidíme, že největší část variability vysvětluje první bilance z_1 mezi skupinou ostatních typů domácností a zbytkem.

Další možností pro zobrazení bilancí je graf hodnot jedné konkrétní bilance pro jednotlivé evropské státy. Tentokrát zvolíme jiné PBD a vykreslíme sloupcový graf

pro bilanci b_2 (obrázek 2.17), která vysvětluje míru celkové relativní významnosti skupiny `Single` oproti skupině `Other`:

$$b_2 = \ln \frac{\sqrt{x_1 x_2}}{\sqrt{x_5 x_6}}$$

Podle obrázku 2.17 můžeme evropské státy rozdělit na dvě skupiny. První skupinou jsou státy severní a západní Evropy, které jsou nejvíce ovlivněny trendem života „single“. Jedná se v největší míře o Švédsko, Dánsko, Nizozemsko a Francii. Na druhé straně máme jižní státy, státy více tradiční a křesťanské, které preferují život ve společných domácnostech. Jako příklad můžeme uvést Makedonii, Chorvatsko či Maltu.

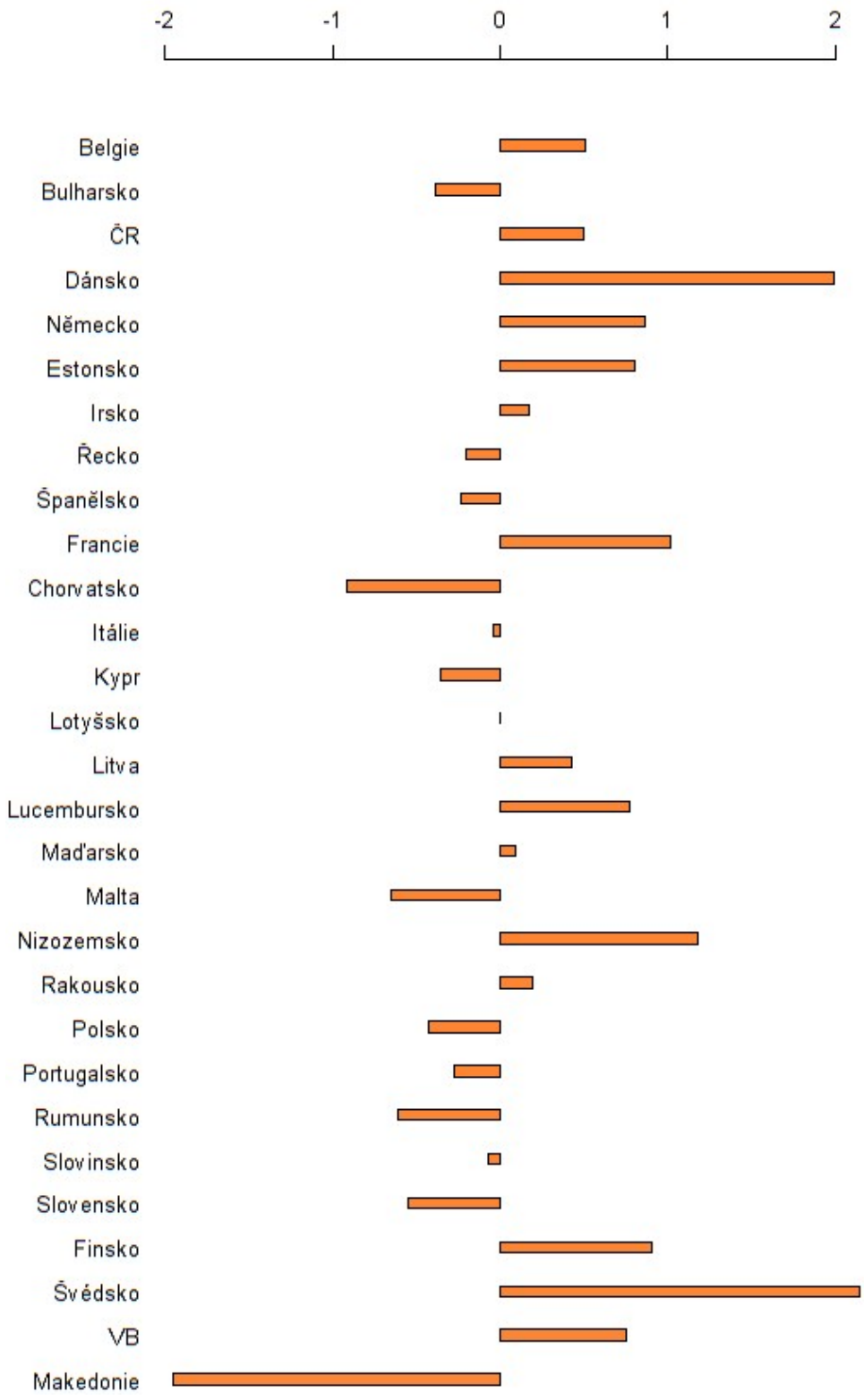
2.2.8 Shlukový dendrogram

Pokud není zřejmé, jak rozdělit pozorování do skupin, můžeme použít shlukovou analýzu. Pomocí hierarchického shlukování jsou statistické jednotky postupně spojovány do skupin podle vzájemné podobnosti. Postupujeme ve dvou krocích:

1. Získáme matici vzdáleností mezi objekty příkazem `dist(X, method)`, kde `X` je kompoziční datová matice obsahující po řádcích jednotlivá pozorování v clr souřadnicích. Implicitní hodnotou parametru `method` je `''euclidean''`, která počítá vzdálenost podle definice 1.7. K dispozici jsou i Minkowského, Manhattanská a maximová metoda.

2. Aplikujeme shlukovací techniku. Na začátku jsou prvky samostatně, poté postupně spojujeme vždy dva nejbližší prvky do té doby, než tvoří jen jedinou skupinu. Použijeme příkaz `hclust(d, method)`, kde za první parametr volíme výstup z předchozího bodu. Druhý pak určuje, jakým způsobem počítáme vzdálenost mezi skupinami:

   - `''complete''` počítá maximální vzdálenost mezi prvky dvou skupin a je implicitní hodnotou.
   - `''average''` počítá průměrnou vzdálenost mezi prvky dvou skupin.
   - `''single''` počítá minimální vzdálenost mezi prvky dvou skupin.
   - `''ward''` počítá vzdálenost mezi váženými průměry dvou skupin.



Obrázek 2.17: Hodnoty bilance b_2 pro jednotlivé státy Evropy zobrazené sloupcovým grafem

Pro datový soubor `household` si shlukový dendrogram (obrázek 2.18) znázorníme takto:

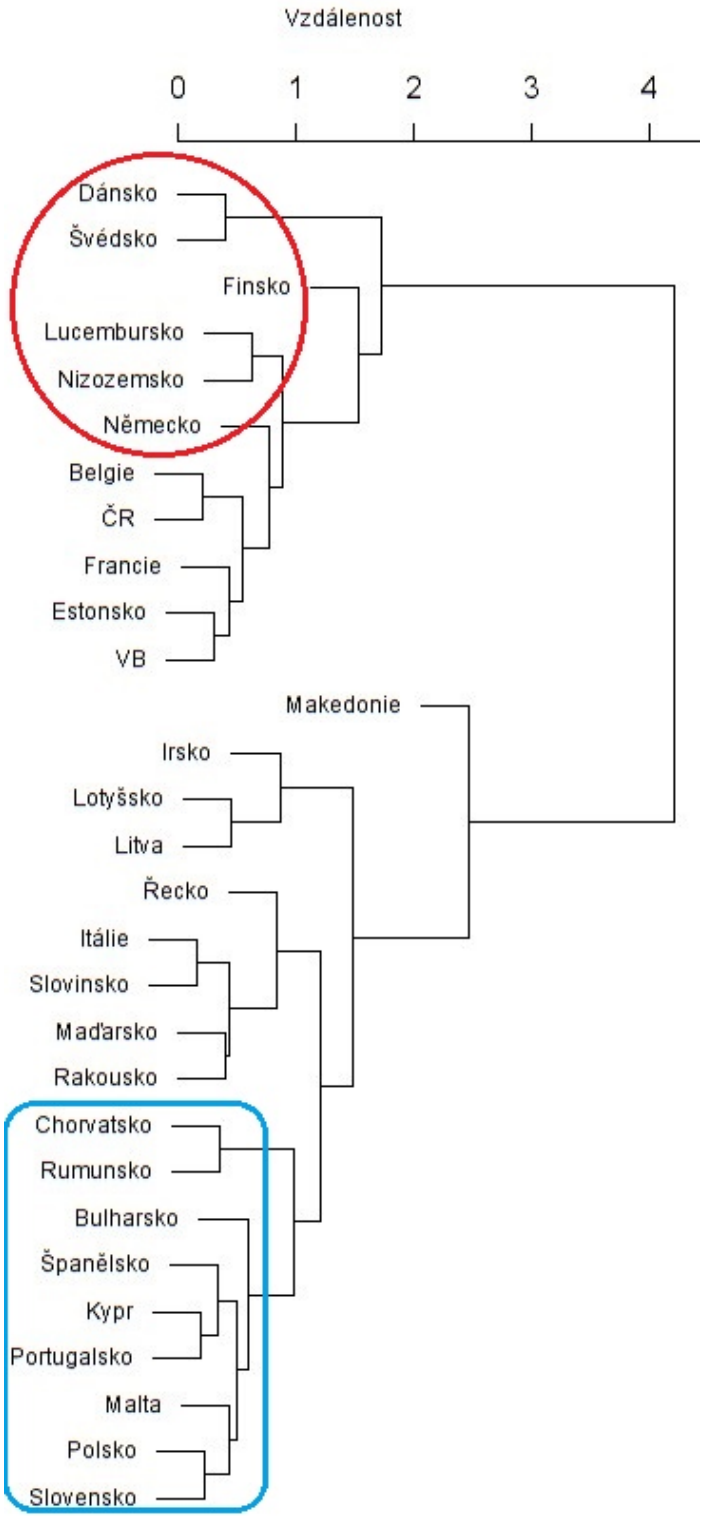
```
> plot(hclust(dist(household)))
```

Opět vidíme podobné rozdělení jako u předchozího obrázku na tradiční státy (zakroužkovány modře) a státy s výskytem moderních společenských trendů (červeně). Ostatní státy tvoří přechod mezi těmito skupinami.

2.3 Shrnutí výsledků kompoziční analýzy

První datový soubor `satisfaction` vyjadřuje spokojenost obyvatel vybraných evropských států se životem. U průměrného evropského státu je 21 % občanů málo spokojených, 57 % středně spokojených a 22 % velmi spokojených. Největší variabilitu spatřujeme mezi složkami `Low` a `High` (tedy malou a velkou spokojeností). Přisuzujeme to rozdílům mezi vyspělými a rozvojovějšími zeměmi. Nejspokojenější je populace severní a západní Evropy. Velmi pozitivně svůj život hodnotí lidé ze skandinávských zemí, Dánska a Islandu. Naopak ve východní a jižní Evropě je obyvatelstvo spokojené méně. Nejhůř svůj život bodují Srbové, Bulhaři, Maďaři a Portugalci.

Druhý kompoziční datový soubor `household` člení domácnosti evropských zemí do šesti kategorií podle počtu dětí a dospělých. Zjišťujeme, že se variabilita příliš neprojevuje mezi složkami lišícími se pouze přítomností dětí. Naopak ji vidíme mezi složkami „ostatních“ domácností (typ `Other`) a zbývajícími složkami. Zatímco lidé na jihu Evropy mají největší podíl společných domácností více generací nebo nepříbuzných lidí, tak v severní Evropě vynikají v počtu obyvatel žijících o samotě bez dětí (tzv. „single“). Trend života „single“ se projevuje rovněž v západní Evropě, např. v Nizozemsku či Německu. Stále ale existují i tradiční státy preferující rodinný život s dětmi. Sem můžeme zařadit kupříkladu Makedonii, Chorvatsko či Maltu.



Obrázek 2.18: Shlukový dendrogram pro datový soubor `household`

Závěr

V bakalářské práci jsme se seznámili s knihovnou „compositions“ softwaru R, která slouží ke zpracování kompozičních dat. Nejprve jsme uvedli problematiku kompozičních dat a jejich geometrie. Naučili jsme se, jak správně s kompozičními daty zacházet, tedy jak je převádět pomocí různých typů logpodílových souřadnic na reálné vektory. Uvedli jsme, jaké knihovny jsou v současné době pro analýzu těchto speciálních dat k dispozici. Dále jsme se zaměřili na práci s knihovnou „compositions“, vyzkoušeli jsme některé její funkce na sociologických datech. Ukázali jsme možnosti grafického znázornění kompozičních dat a předvedli jsme interpretaci konkrétních dat. Nakonec jsme se věnovali popisné statistice, mírám polohy, variability i vzájemným vztahům mezi složkami.

Zjistili jsme, že knihovna „compositions“ představuje vhodný nástroj k analýze kompozičních dat. Pracuje se s ní pohodlně a rychle: díky několika krátkým příkazům získáme o datech řadu informací a můžeme vyvozovat příslušné závěry. Jelikož potřebujeme kompoziční data zpracovávat v mnoha oborech, je nezbytné je i nadále studovat a rozvíjet knihovny pro jejich analýzu.

Literatura

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

[2] Aitchison, J.: *The one-hour course in compositional data analysis or compositional data analysis is simple.* In V. Pawlowsky-Glahn (Ed.), Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology, Volume I, II and addendum. International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 1997.

[3] van den Boogaart, K. G., Tolosana-Delgado, R.: *Analyzing compositional data with R.* Springer, Heidelberg, 2013.

[4] Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V.: *Compositional data analysis in the geosciences: From theory to practice.* Geological Society, London, 2006.

[5] Hron, K.: *Elementy statistické analýzy kompozičních dat.* Informační bulletin České statistické společnosti, č. 3/2010, 50–56.

[6] Pawlowsky-Glahn, V. and Egozcue, J. J.: *Exploring compositional data with the CoDa-dendrogram.* Austrian Journal of Statistics, č. 1 a 2/2011, 103–113.

[7] Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R.: *Lecture notes on compositional data analysis.* Universitat de Girona, 2007 [online].[cit. 2016-10-23]. Dostupné z `http://dugi-doc.udg.edu/bitstream/handle/10256/297/CoDa-book.pdf?sequence=1`.

[8] Statistics Explained [online]. [cit. 2016-06-20]. Dostupné z: `http://ec.europa.eu/eurostat/statistics-explained/index.php/Household_composition_statistics`.

[9] Statistics Explained [online]. [cit. 2016-06-20]. Dostupné z: `http://ec.europa.eu/eurostat/statistics-explained/index.php/Quality_of_life_in_Europe_-_facts_and_views_-_overall_life_satisfaction`.

Příloha A

Rozdělení typů domácností ve státech Evropy podle počtu dětí a dospělých, 2014 (v % jednotlivých typů domácností) [8]

	Typ domácnosti	Označení proměnné
x_1	1 dospělý s dětmi	`Single_yes`
x_2	1 dospělý bez dětí	`Single_no`
x_3	Pár s dětmi	`Couple_yes`
x_4	Pár bez dětí	`Couple_no`
x_5	Ostatní typy domácností s dětmi	`Other_yes`
x_6	Ostatní typy domácností bez dětí	`Other_no`

Stát	x_1	x_2	x_3	x_4	x_5	x_6
Belgie	5,7	28,3	22,3	27,3	5,2	11,2
Bulharsko	2,8	26,5	17,4	25,5	8,5	19,2
ČR	4,7	30,0	22,8	26,5	4,6	11,4
Dánsko	8,3	42,5	19,9	23,9	1,8	3,7
Německo	3,8	40,3	15,2	28,7	3,1	8,9
Estonsko	6,8	35,6	21,5	21,4	5,1	9,6
Irsko	6,3	22,1	28,9	20,9	6,3	15,4
Řecko	1,8	30,2	21,1	24,2	4,7	18,0
Španělsko	3,5	25,0	23,3	21,8	7,3	19,1
Francie	6,1	34,3	22,0	26,5	3,7	7,4
Chorvatsko	2,0	24,5	19,5	18,0	14,0	21,9
Itálie	2,8	33,1	21,5	19,7	6,0	16,9
Kypr	3,7	20,0	26,8	23,2	8,8	17,4
Lotyšsko	5,7	32,2	16,2	17,7	10,0	18,3
Litva	6,6	36,2	17,6	18,0	7,1	14,5
Lucembursko	4,5	34,3	27,4	22,1	4,8	6,9
Maďarsko	4,1	32,7	18,7	21,3	6,9	16,3
Malta	3,2	19,0	23,8	21,3	9,8	22,9
Nizozemsko	4,3	36,5	21,8	29,4	2,9	5,2
Rakousko	3,0	37,0	17,2	23,9	5,7	13,1
Polsko	3,7	22,2	23,3	22,4	11,8	16,5
Portugalsko	4,3	20,9	23,9	23,8	8,4	18,7
Rumunsko	2,5	27,4	20,6	19,3	13,2	17,0
Slovinsko	2,7	32,8	22,2	20,4	6,6	15,2
Slovensko	3,4	21,8	22,8	20,9	11,2	19,8
Finsko	1,6	40,3	19,0	31,7	2,0	5,4
Švédsko	6,3	47,9	19,5	22,2	1,7	2,5
VB	7,2	31,4	19,5	26,3	4,6	10,9
Makedonie	1,5	9,6	22,1	13,9	26,6	26,2

Příloha B

Spokojenost se životem ve státech Evropy, 2013 (v %) [9]

`Low`	Malá	0-5 bodů
`Medium`	Střední	6-8 bodů
`High`	Velká	9-10 bodů

Stát	Malá	Střední	Velká
Belgie	9,2	69,9	20,9
Bulharsko	64,2	29,8	5,9
ČR	25,4	53,3	21,3
Dánsko	10,6	46,6	42,7
Německo	19,2	55,8	25,0
Estonsko	34,4	52,1	13,5
Irsko	16,7	52,7	30,6
Řecko	35,5	51,8	12,8
Španělsko	23,2	58,4	18,4
Francie	19,1	64,8	16,1
Chorvatsko	35,4	49,5	15,0
Itálie	22,7	63,0	14,2
Kypr	37,0	48,8	14,2
Lotyšsko	30,8	56,6	12,6
Litva	27,9	53,3	18,8
Lucembursko	14,8	59,5	25,7
Maďarsko	38,5	50,2	11,3
Malta	20,3	57,2	22,5
Nizozemsko	5,6	68,3	26,1
Rakousko	12,9	49,3	37,9
Polsko	19,9	50,7	29,4
Portugalsko	40,5	45,7	13,8
Rumunsko	15,9	64,2	19,9
Slovinsko	24,4	55,2	20,4
Slovensko	26,4	48,6	25,0
Finsko	6,0	55,5	38,6
Švédsko	8,1	56,8	35,1
VB	19,1	53,2	27,8
Island	9,5	52,4	38,1
Norsko	10,3	54,1	35,6
Švýcarsko	8,0	53,5	38,5
Srbsko	61,4	32,4	6,2