UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Imputace chybějících hodnot pro kompoziční data

Vedoucí diplomové práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2011

Vypracoval:
Tereza Rychlá
AME, II. ročník

Prohlášení

Prohlašuji, že jsem vytvořila tuto diplomovou práci samostatně pod vedením RNDr. Karla Hrona, Ph.D. a že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 31. 3. 2011

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce RNDr. Karlovi Hronovi, Ph.D. za jeho spolupráci i za čas, který mi věnoval při konzultacích. Dále bych chtěla poděkovat rodině a přátelům za podporu při mém studiu.

Obsah

Úvod

Pro diplomovou práci jsem si vybrala téma imputace chybějících hodnot pro kompoziční data. V prvé řadě mě zajímalo, čím jsou kompoziční data tak specifická, jelikož se jedná o relativně nový obor statistiky, který se neustále vyvíjí. Na druhou stranu jsem se chtěla více dozvědet o imputačních metodách pro chybějící údaje v datových souborech. V běžném životě se při různých průzkumech a anketách mohou chybějící hodnoty vyskytovat. Zajímalo mě, z jakých důvodů jsou data nepřítomna a také, jak je co nejefektivněji nahradit, abychom datový soubor mohli použít k další například statistické analýze, kterou je možno aplikovat vždy pouze pro úplný datový soubor.

Problematice imputace chybějících hodnot je věnováno velké množství prací, jež se věnují i různým specifickým problemům chybějících hodnot. Například můžeme pro missings, jak jsou někdy chybějící data označována podle anglického názvu, rozlišovat různé modely jejich výskytu. Zkoumáme mechanismy jejich vzniku a také rozhodujeme, jestli jsme je schopni nějakým způsobem nahradit právě pomocí imputačních metod.

Diplomová práce je rozdělena do pěti kapitol. První kapitola pojednává o teoretických poznatcích, o které se kompoziční data opírají. Jsou zde zmíněny základní definice, principy práce s kompozicemi, Aitchisonova geometrie nebo vyjádření kompozic v reálném prostoru. Druhá kapitola s názvem Imputace chybějících hodnot nejprve stručně pojednává právě o mechanismech vzniku chybějících hodnot. Především jsem rozlišila mechanismy podle náhodnosti vzniku missings a jejich závislosti na naměřených hodnotách v datovém souboru. Aby čtenář danou problematiku lépe pochopil, vše je vysvětleno na příkladech z běžného života. Také jsem se zmínila o regresní imputaci, jako zřejmě nejužívanější metodě náhrady chybějících hodnot, zejména o teoretickém aparátu regresní analýzy.

Třetí kapitola zahrnuje popis metod používaných k imputaci složek kompozic. V rozsahu textu jsem popsala dvě užívané imputační metody: algoritmus k nejbližších sousedů a iterativní regresní imputaci, a provedla jejich kvalita-

tivní ohodnocení. Obě dvě metody jsou vvysvětleny podrobně na ilustrativním příkladu.

Čtvrtá kapitola stručně popisuje práci s kompozice ve statistickém softwaru R. Uvedla jsem příkazy pro imputaci pomocí iterativní regrese. Kapitola je doplněna krátkým sledem příkazů, týkajících se iterativní regrese pro ilustrativní příklad ze třetí kapitoly. V poslední, páté kapitole, jsem zpracovávala reálná data získaná přímo z Českého statistického úřadu. Potřebné výpočty byly provedeny v softwaru R.

Závěrem bych chtěla poznamenat, že cílem práce je obě metody pro imputaci chybějících hodnot popsat na příkladu a provést jejich zhodnocení v souvislosti s kvalitou imputace.

1 Kompoziční data

Datové soubory jsou obvykle uvedeny jako vícerozměrná pozorování s kvantitativním charakterem složek. Složky mohou být vyjádřeny jako podíly na celku, například v procentech nebo proporcích. Například procentuální vyjádření zastoupení politických stran v Parlamentu České republiky, procentuální podíl skupin prodaných výrobků na celkovém zisku nebo věkové zastoupení obyvatel státu v procentech. Tento charakter vede k jejich speciální interpretaci, protože podíly složek na celku nám dávají jedinou relevantní informaci. Uvedené úvahy vedly v 80. letech 20. století Johna Aitchisona k zavedení pojmu *kompoziční data* neboli *kompozice* [1].

Naneštěstí standardní statistické metody vytvořené na základě vlastností euklidovské geometrie, vedou při aplikaci na kompozice často k nerozumným výsledkům. Příklad takto problematické chování můžeme získat i při použití obyčejného korelačního koeficientu [16]. Pro kompoziční data totiž uvažujeme simplex jako jejich výběrový prostor, na kterém lze při zavedení operací mocninná transformace, perturbace a Aitchisonova skalárního součinu definovat tzv. Aitchisonovu geometrii.

Kompoziční data se vyskytují ve velkém množství rozdílných praktických úloh z mnoha oblastí aplikací. Například se využívají v archeologii, geologii, medicíně, ekonomii a v dalších oblastech.

1.1 Základní definice

Definice 1: Řádkový vektor $\mathbf{x}= (x_1, \ldots, x_D)^T$ je definován jako *D-složková kompozice*, pokud jsou všechny její složky ryze kladná reálná čísla a nesou pouze relativní informaci [15].

Relativní informace je obsažena v podílech mezi složkami kompozice, tyto vždy představují části nějakého celku, jako je například jejich hmotnost nebo procentuální podíl na určitém vzorku. Ve většině příkladů mají kompozice kon-

stantní součet k. Často k volíme rovno 1 (proporce) nebo 100 (procenta) podle charakteru úlohy. V některých případech nelze konstantní součet určit, protože složky kompozice uvažujeme v jednotkách koncentrace (např. mg/L).

Definice 2: Výběrový prostor kompozičních dat nazýváme *simplex* a definujeme jej jako

$$S^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)^T | x_i > 0, i = 1, \ldots, D; \sum_{i=1}^{D} x_i = k \right\}.$$

Simplex je $(D-1)$-rozměrný podprostor R^D. Tuto definici však nelze aplikovat na data měřená v jednotkách koncentrace, a proto definujeme následující operaci.

Definice 3: Pro každý vektor s D kladnými reálnými složkami $\mathbf{z}= (z_1, \ldots, z_D)^T \in \mathrm{R}_+^D, z_i > 0$ pro každé $i = 1, \ldots, D$, lze definovat *uzávěr kompozice* vzhledem ke konstantnímu součtu k,

$$C(z) = \left(\frac{k \cdot z_1}{\sum_{i=1}^{D} z_i}, \ldots, \frac{k \cdot z_D}{\sum_{i=1}^{D} x_i} \right).$$

Operace uzávěr kompozice transformuje součet složek kompozice na zvolenou konstantu k bez ztráty informace. Výsledný vektor je tak vyjádřený v jednotkách, které můžeme snáze interpretovat (například %).

Definice 4: Nechť je dána kompozice $\mathbf{x}$. *Subkompozice* $\mathbf{x}_s$ o s složkách při použití operace uzávěr kompozice je subvektor $(x_{i_1}, \ldots, x_{i_s})^T$ vzniklý z kompozice $\mathbf{x}$. Subindexy $i_1, \ldots, i_s,\ 1 \leq i_1 < \ldots < i_s \leq D$, označují složky zahrnuté do subkompozice, nemusí to být nutně prvních s složek.

1.2 Principy práce s kompozičními daty

Principy statistické analýzy kompozičních dat zavedl John Aitchinson v knize [1]. Byly definovány spíše na základě zkušeností v mnoha oblastech vědy (např.

geologie), než na základě nějakých statistických výpočtů. Můžeme je shrnout do tří základních principů, které popisují jejich podstatu.

1.2.1 Invariance vůči poměru

Nejdůležitější vlastností kompozičních dat je, že nesou pouze relativní informaci. Informace v kompozici tak nezávisí na jednotkách, ve kterých je kompozice vyjádřena. S tím souvisí i následující definice [15].

Definice 5: Dva vektory $\mathbf{x}$, $\mathbf{y} \in \mathrm{R}_+^D$, $x_i, y_i > 0$ pro každé $i = 1, \ldots, D$, s D kladnými reálnými složkami jsou *kompozičně ekvivalentní*, pokud existuje skalár $\lambda \in \mathrm{R}^+$ takový, že

$$\mathbf{x} = \lambda \cdot \mathbf{y}$$

a ekvivalentně

$$C(\mathbf{x}) = C(\mathbf{y}).$$

Této vlastnosti kompozic můžeme využít, pokud analýzy vedou na stejné výsledky nezávislé na hodnotách λ. Podle [1] ji nazýváme *invariance vůči poměru*.

Definice 6: Funkce $f(\cdot)$ je *poměrově invariantní*, jestliže pro každé reálné $\lambda \in \mathrm{R}^+$ a pro každou kompozici $\mathbf{x} \in S^D$ funkce splňuje vztah

$$f(\lambda \mathbf{x}) = f(\mathbf{x}),$$

tj. poskytuje stejný výsledek pro všechny kompozičně ekvivalentní vektory.

Tohoto může být dosaženo pouze tehdy, je-li funkce logaritmickou funkcí složek kompozice $\mathbf{x}$.

1.2.2 Invariance vůči změně pořadí

Funkce je *invariantní vůči změně pořadí*, jestliže poskytuje stejné výsledky, i když změníme uspořádání složek v kompozici. Permutace složek kompozice tak nezmění informaci zprostředkovanou pomocí vektoru kompozice [3].

1.2.3 Soudržnost subkompozice

Budeme-li uvažovat subkompozice, tak vzdálenost měřená mezi dvěmi plnohodnotnými kompozicemi musí být větší (nebo přinejmenším stejná) než vzdálenost mezi jejich subkompozicemi. To znamená, je-li $\Delta_n(\cdot,\cdot)$ libovolná vzdálenost mezi kompozicemi, potom

$$\Delta_D(\mathbf{x},\mathbf{y}) \geq \Delta_d(\mathbf{x}_d,\mathbf{y}_d),$$

kde $\mathbf{x}$, $\mathbf{y}$ jsou D-složkové kompozice a $\mathbf{x}_d$, $\mathbf{y}_d$ jsou subkompozice o d složkách, $d \leq D$, vytvořené z předešlých kompozic $\mathbf{x}$ a $\mathbf{y}$ [3].

Toto specifické chování vzdálenosti se nazývá *subkompoziční dominance*. Na příkladu 2.4 uvedeném v [15] lze ukázat, že euklidovská vzdálenost mezi dvěma vektory tuto vlastnost nesplňuje, a tudíž není vhodná pro počítání vzdáleností mezi kompozicemi.

1.3 Aitchisonova geometrie

V reálném euklidovském prostoru jsme zvyklí sčítat vektory, násobit je skalárem, analyzovat vlastnosti jako ortogonalita nebo počítat vzdálenosti mezi prvky prostoru. Standardní euklidovskou geometrii však nelze použít pro kompoziční data. Proto je pro potřeby analýzy kompozičních dat požadována rozumná geometrie, kterou budeme dále nazývat Aitchisonovou geometrií. Nejprve budeme definovat dvě operace, které simplexu dodávají strukturu vektorového prostoru. První z nich je perturbace a je analogická ke sčítání vektorů v euklidovské geometrii. Druhá se nazývá mocninná transformace a je obdobou násobení skalárem. Obě tyto operace požadují v definici operaci uzávěr uvedenou v Definici 3, protože uzávěr je projekce vektoru s kladnými složkami na simplex [15].

Definice 7: *Perturbace kompozice* $\mathbf{x} \in S^D$ kompozicí $\mathbf{y} \in S^D$ je definována jako

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

Definice 8: *Mocninná transformace* kompozice $\mathbf{x} \in S^D$ konstantou $\alpha \in \mathrm{R}$ je

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

Trojice simplex, perturbace a mocninná transformace $(S^D, \oplus, \odot)$ tvoří vektorový prostor. Abychom dostali euklidovský vektorový prostor, uvažujeme následující skalární součin se související normou a vzdáleností.

Definice 9: *Skalární součin* kompozic $\mathbf{x}$, $\mathbf{y} \in S^D$ je definován

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

Definice 10: *Norma kompozice* $\mathbf{x} \in S^D$ je

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

Definice 11: *Vzdálenost kompozic* $\mathbf{x}$, $\mathbf{y} \in S^D$ se definuje

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

Aitchisonova vzdálenost nahrazuje u kompozic euklidovskou vzdálenost, protože simplex má jinou geometrickou strukturu než standardní euklidovský prostor.

S odvoláním se na vlastnosti prostoru $(S^D, \oplus, \odot, \langle \cdot, \cdot \rangle_A)$ jako na euklidovský vektorový prostor, můžeme hovořit celkově o *Aitchisonově geometrii* na simplexu a speciálně o *Aitchisonově vzdálenosti*, *normě* a *skalárním součinu* [15].

Algebraicko-geometrická struktura simplexu S^D splňuje základní vlastnosti jako zaměnitelnost vzdálenosti s perturbací a mocninnou transformací, tj.

$$d_A(\mathbf{p} \oplus \mathbf{x}, \mathbf{p} \oplus \mathbf{y}) = d_A(\mathbf{x}, \mathbf{y}),$$

$$d_A(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{y}) = |\alpha| d_A(\mathbf{x}, \mathbf{y}),$$

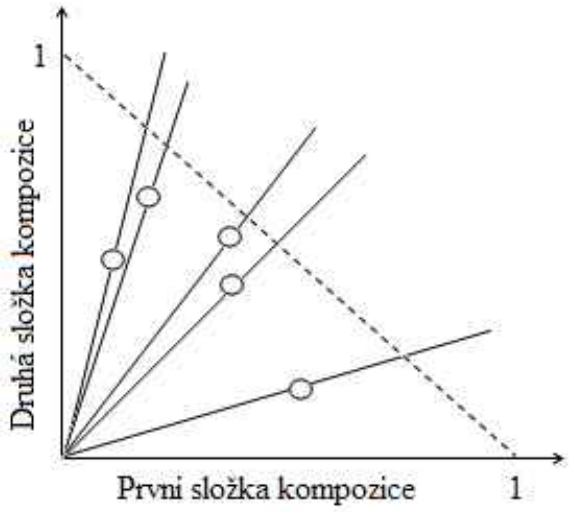
pro každou kompozici $\mathbf{x}$, $\mathbf{y}$, $\mathbf{p} \in S^D$ a $\alpha \in \mathrm{R}$.

Kompoziční data jsou charakterizována konstantním součtem k. Kompozice $\mathbf{x} = (x_1, \ldots, x_D)^T$ s naměřenými hodnotami je definována jako prvek třídy ekvivalence

$$\underline{\mathbf{x}} = \left\{ c\mathbf{x}, c \in \mathrm{R}^+ \right\}$$

odpovídající kompozici $\mathbf{x}$. Jinými slovy, dvě kompozice, které jsou prvky stejné třídy ekvivalence $\underline{\mathbf{x}}$, obsahují stejnou informaci a můžeme je nazývat *kompozičně ekvivalentní* [8].

Budeme například uvažovat kompozici se dvěma složkami. Dále stanovíme konstantní součet $k = 1$, v grafu je označen čerchovanou čarou. Součet složek je pro každou kompozici menší než k. Každý prvek třídy ekvivalence můžeme posunout libovolně po přímce jdoucí od počátku k čerchované čáře, aniž bychom změnili podíl mezi dvěma složkami kompozice. Popsanou situaci ukazuje následující obrázek [8].



Obr. 1

1.4 Kompozice v reálném prostoru

Standardní observace (s euklidovskou metrikou) lze obvykle vyjádřit v souřadnicích vzhledem ke kanonické bázi $\{\vec{\mathbf{e}_1}, \ldots, \vec{\mathbf{e}_D}\} \in \mathrm{R}^D$, tedy při souřadnicích $(x_1, \ldots, x_D)$ obdržíme

$$\mathbf{x} = x_1(1, 0 \ldots, 0) + x_2(0, 1, \ldots, 0) + \ldots + x_D(0, 0, \ldots, 1) = \sum_{i=1}^{D} x_i \cdot \vec{\mathbf{e}_i}.$$

Problémem je, že vektory $\{\vec{\mathbf{e}_1}, \ldots, \vec{\mathbf{e}_D}\}$ nejsou bází s ohledem na vektorovou strukturu simplexu S^D, protože ne každá kombinace koeficientů dává prvek simplexu.

1.4.1 Generující systém

Abychom mohli správně definovat ortonormální bázi na simplexu, musíme najít generující systém vzhledem k Aitchisonově geometrii. Ten se následně užívá k vytvoření báze. Obvyklý způsob nalezení takového systému je, že vezmeme v úvahu množinu kompozic $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$ takovou, že

$$\mathbf{w}_i = C(\exp(\vec{\mathbf{e}_i})) = C(1, 1, \ldots, e, \ldots, 1),$$

kde pro každou kompozici $\mathbf{w}_i$ je číslo e na i-té pozici, $i = 1, \ldots, D$. Tuto rovnici můžeme zapsat také ve tvaru

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln(x_i) \odot \mathbf{w}_i = \ln(x_1) \odot (e, 1, \ldots, 1) \oplus \ldots \oplus \ln(x_D) \odot (1, 1, \ldots, e).$$

Uvedený generující systém lze využít při definici tzv. centred logratio (clr) transformace.

Definice 12: Pro kompozici $\mathbf{x} \in S^D$ jsou koeficienty *clr transformace* složky jediného vektoru $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_D) = clr(\mathbf{x})$, splňující

$$\mathbf{x} = clr^{-1}(\boldsymbol{\xi}) = C(\exp(\xi_1), \ldots, \exp(\xi_D)), \qquad \sum_{i=1}^{D} \xi_i = 0,$$

pro i-tý koeficient clr transformace [15] platí

$$\xi_i = \frac{\ln(x_i)}{g(\mathbf{x})},$$

kde $g(\mathbf{x})$ je geometrický průměr složek kompozice $\mathbf{x}$.

Centrální logratio transformace je pro složky kompozice symetrická, ale na druhou stranu musí být součet složek roven nule.

1.4.2 Ortonormální souřadnice

Ortonormální báze musí být v souladu s geometrickou strukturou, ve které pracujeme. V našem případě ji můžeme získat z výše uvedeného generujícího systému například použitím Gram-Schmidtovy procedury. Takto získaná báze je jen jednou z nekonečně mnoha ortonormálních bází, které můžeme na simplexu vzhledem k Aitchisonově geometrii definovat. Literatura [15] se podrobněji zabývá jejich obecnými vlastnostmi. Vybrali jsme jednu ortonormální bázi $\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}$. Kompozice $\mathbf{x} \in S^D$ je tedy vyjádřena ve tvaru

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i,$$

$$x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_A ,$$

kde $\mathbf{x}^* = (x_1^*, \ldots, x_{D-1}^*)^T$ je vektor souřadnic kompozice $\mathbf{x}$ s ohledem na vybranou bázi.

Funkce $ilr : S^D \to \mathrm{R}^{D-1}$ přiřazující kompozici $\mathbf{x}$ souřadnice $\mathbf{x}^*$, se nazývá *ilr transformace*, v angličtině označovaná isometric log-ratio transformation. Někdy se pro zjednodušení označuje ilr transformace jako h, tj. $ilr \equiv h$. Tato transformace se jeví jako nejvhodnější pro vytvoření ortonormální báze na simplexu. Představuje totiž dobré teoretické i praktické vlastnosti. Mezi nimi je tou klíčovou zejména izomerie, neboli, například Aitchisonova vzdálenost dvou kompozic $\mathbf{x}, \mathbf{y} \in S^D$ je stejná jako euklidovská vzdálenost odpovídající jejich ilr obrazům $\mathbf{x}^*$ a $\mathbf{y}^*$. Tedy ilr transformace umožňuje reprezentovat kompoziční data v podmínkách

standardního euklidovského prostoru, a tudíž mohou být následně použity standardní statistické metody.

Pro určení ortonormální báze na simplexu existuje několik způsobů. Jedním z nich je *postupné (sekvenční) binární dělení*. Umožňuje interpretaci ve smyslu skupin složek kompozice. Souřadnice zde nazýváme bilance nebo také rovnováhy. Při konstrukci souřadnic postupujeme následujícím způsobem. V prvním kroku jsou všechny složky rozděleny do dvou skupin označených +1 a -1. Dostaneme souřadnici, která vyjadřuje rovnováhu mezi těmito dvěma skupinami. Tato souřadnice zastupuje podíly mezi jednotlivými složkami uvedených dvou skupin. V dalších krocích je každá předešlá skupina rozdělena opět do dvou skupin, označených +1 a -1. Získané souřadnice v každém kroku opět vyjadřují podíly mezi jednotlivými složkami podle rozdělení do skupin. Proces se opakuje, dokud všechny skupiny neobsahují jedinou složku. K tomu je potřeba právě $D-1$ kroků, což je rovno dimenzi simplexu [6].

Pro každý krok dělení můžeme určit bilanci mezi dvěmi podskupinami. Označme +1 r složek první podskupiny, tj. $i_1, \ldots, i_r$, a -1 s složek druhé podskupiny, tj. $j_1, \ldots, j_s$. Bilance je definována

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{(x_{i1} x_{i2} \cdots x_{ir})^{1/r}}{(x_{j1} x_{j2} \cdots x_{js})^{1/s}} = \ln \frac{(x_{i1} x_{i2} \cdots x_{ir})^{a_+}}{(x_{j1} x_{j2} \cdots x_{js})^{a_-}},$$

kde

$$a_+ = +\frac{1}{r}\sqrt{\frac{rs}{r+s}}, \quad a_- = -\frac{1}{s}\sqrt{\frac{rs}{r+s}}.$$

Celý proces výpočtu báze se zapisuje do tabulky, jejíž názornou ukázku můžete vidět v následujícím příkladu.

Příklad 1: Uvažujeme kompozici $\mathbf{x}= (x_1, x_2, x_3, x_4, x_5)^T$, která představuje zastoupení pěti politických stran v Poslanecké sněmovně Parlamentu České republiky. Pro tuto kompozici chceme nalézt ortonormální bázi použitím metody postupného binárního dělení.

Řešení:

V prvním kroku politické strany rozdělíme na strany pravicové (ODS, TOP09, VV) a strany levicové (ČSSD, KSČM). Dále budeme dělit pravicové strany na strany ryze pravicové (ODS, TOP09) a strany umístěné spíše ve středu politického spektra (VV). Tím dostaneme jednu jednoprvkovou skupinu stran. V dalším kroku oddělíme dvě levicové strany do samostatných skupin a tudíž získáme dvě jednoprvkové skupiny. V posledním kroku od sebe oddělíme dvě ryze pravicové strany a opět získáme dvě jednoprvkové skupiny politických stran. Následuje tabulka, která názorně ukazuje výše uvedené dělení.

	x_1 = ODS	x_2 = ČSSD	x_3 = KSČM	x_4=VV	x_5 = TOP09
z_1	+	−	−	+	+
z_2	+			−	+
z_3		+	−		
z_4	+				−

Bilance po dosazení do výše uvedeného vztahu jsou tavru

$$z_1 = \sqrt{\frac{3 \cdot 2}{3+2}} \ln \frac{(x_1 x_4 x_5)^{\frac{1}{3}}}{(x_2 x_3)^{\frac{1}{2}}} = \sqrt{\frac{6}{5}} \ln \frac{(x_1 x_4 x_5)^{\frac{1}{3}}}{(x_2 x_3)^{\frac{1}{2}}},$$

$$z_2 = \sqrt{\frac{2}{3}} \ln \frac{(x_1 x_5)^{\frac{1}{2}}}{(x_4)},$$

$$z_3 = \ln \frac{(x_2)}{(x_3)},$$

$$z_4 = \ln \frac{(x_1)}{(x_5)}.$$

Poznamenejme, že postupné binární dělení lze provést i jiným způsobem při zachování smysluplné interpretace výsledných souřadnic. Různé ilr transformace téže kompozice jsou totiž vzájemně převoditelné pomocí ortogonální transformace.

Ilr transformace pomocí postupného binárního dělení se může použít v souvislosti s odhady chybějících hodnot. Bude-li například nejvíce chybějících hodnot obsaženo v první složce kompozice, můžeme vyjádřit její ilr transformaci jako

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

kde

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{\sqrt[D-j]{\prod_{i=j+1}^{D} x_i}}{x_j}$$

pro $j = 1, \ldots, D-1$. Potom tato volba bilancí odděluje v z_1 veškerou relativní informaci složky x_1 od zbývajících složek $x_2, \ldots, x_D$. Jinak řečeno, bilance z_1 obsahuje veškerou relativní informaci o x_1 vzhledem k ostatním hodnotám $x_2, \ldots, x_D$, tj. obsahuje veškerou informaci o podílech x_1 vzhledem ke zbylým složkám kompozice $\mathbf{x}$. Princip je naznačen v následující tabulce.

	x_1	x_2	x_3	$\ldots$	x_{D-1}	x_D
z_1	$+$	$-$	$-$		$-$	$-$
z_2		$+$	$-$		$-$	$-$
$\vdots$			$+$		$-$	$-$
z_{D-1}					$+$	$-$

Pro bilance $z_1, \ldots, z_{D-1}$ můžeme určit jejich rozptyly [4], vyjádřené pomocí rozptylů logaritmů podílů jednotlivých složek kompozice,

$$\operatorname{var}(z_i) = \frac{1}{D-i+1} \sum_{p=i+1}^{D} \operatorname{var}\left(\ln \frac{x_i}{x_p}\right) - \frac{1}{2(D-i)(D-i+1)} \sum_{p=i+1}^{D} \sum_{q=i+1}^{D} \operatorname{var}\left(\ln \frac{x_p}{x_q}\right).$$

Výše uvedený vztah je podrobně odvozen v [4], odkud jsme ho pro potřeby práce převzali. Skládá se ze dvou částí. První část tvoří rozptyly pro logaritmy podílů složek $x_i, \ldots, x_D$, k nimž se vztahuje bilance z_i. Protože se tyto rozptyly ve $\operatorname{var}(z_i)$ vyskytují v kladném tvaru, potvrzují výše zmíněnou interpretaci bilancí se smyslu vysvětlení daných podílů mezi složkami kompozice. Druhá část vyjadřuje rozptyly pro logaritmy podílů mezi ostatními složkami navzájem a je dána způsobem konstrukce bilancí.

Vrátíme-li se k příkladu 1, dostaneme po dosazení rozptyly složek z_i ve tvaru

$$\begin{aligned}
\mathrm{var}(z_1) &= \frac{1}{5-1+1}\sum_{p=1+1}^{5}\mathrm{var}\left(\ln\frac{x_1}{x_p}\right) - \frac{1}{2(5-1)(5-1+1)}\sum_{p=1+1}^{5}\sum_{q=1+1}^{5}\mathrm{var}\left(\ln\frac{x_p}{x_q}\right) = \\
&= \frac{1}{5}\left[\mathrm{var}\left(\ln\frac{x_1}{x_2}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_5}\right)\right] - \\
&-\frac{1}{2\cdot 4\cdot 5}\sum_{p=2}^{5}\left[\mathrm{var}\left(\ln\frac{x_p}{x_2}\right) + \mathrm{var}\left(\ln\frac{x_p}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_p}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_p}{x_5}\right)\right] = \\
&= \frac{1}{5}\left[\mathrm{var}\left(\ln\frac{x_1}{x_2}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_5}\right)\right] - \\
&-\frac{1}{20}\left[\mathrm{var}\left(\ln\frac{x_3}{x_2}\right) + \mathrm{var}\left(\ln\frac{x_4}{x_2}\right) + \mathrm{var}\left(\ln\frac{x_5}{x_2}\right) + \mathrm{var}\left(\ln\frac{x_4}{x_3}\right) + \right. \\
&\left. +\mathrm{var}\left(\ln\frac{x_5}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_5}{x_4}\right)\right], \\
\mathrm{var}(z_2) &= \frac{1}{5-2+1}\sum_{p=3}^{5}\mathrm{var}\left(\ln\frac{x_2}{x_p}\right) - \frac{1}{2(5-2)(5-2+1)}\sum_{p=3}^{5}\sum_{q=3}^{5}\mathrm{var}\left(\ln\frac{x_p}{x_q}\right) = \\
&= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_2}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_2}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_2}{x_5}\right)\right] - \\
&-\frac{1}{12}\left[\mathrm{var}\left(\ln\frac{x_4}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_5}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_5}{x_4}\right)\right], \\
\mathrm{var}(z_3) &= \frac{1}{5-3+1}\sum_{p=4}^{5}\mathrm{var}\left(\ln\frac{x_3}{x_p}\right) - \frac{1}{2(5-3)(5-3+1)}\sum_{p=4}^{5}\sum_{q=4}^{5}\mathrm{var}\left(\ln\frac{x_p}{x_q}\right) = \\
&= \frac{1}{3}\left[\mathrm{var}\left(\ln\frac{x_3}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_3}{x_5}\right)\right] - \frac{1}{6}\mathrm{var}\left(\ln\frac{x_5}{x_4}\right), \\
\mathrm{var}(z_4) &= \frac{1}{5-4+1}\mathrm{var}\left(\ln\frac{x_4}{x_5}\right) - \frac{1}{2(5-4)(5-4+1)}\mathrm{var}\left(\ln\frac{x_5}{x_5}\right) = \frac{1}{2}\mathrm{var}\left(\ln\frac{x_4}{x_5}\right).
\end{aligned}$$

Ve smyslu výše uvedeného by potom postupné binární dělení v příkladu 1 představovala tabulka

	x_1 =ODS	x_2 =ČSSD	x_3 =KSČM	x_4 =VV	x_5 =TOP09
z_1	+	−	−	−	−
z_2		+	−	−	−
z_3			+	−	−
z_4				+	−

2 Imputace chybějících hodnot

2.1 Mechanismy vzniku chybějících hodnot

Tak jako nás zajímají modely chybějících hodnot, chceme se blíže zaměřit na mechanismy vedoucí k jejich vzniku. Především nás zajímá, jestli chybějící hodnoty nějakým způsobem souvisí s hodnotami daných proměnných v datovém souboru. Vlastnosti metod použitých pro imputaci chybějících hodnot totiž závisí na charakteru závislosti proměnných v mechanismu.

Můžeme vyjmenovat několik příkladů mechanismů vzniku chybějících hodnot. Prvním příkladem může být situace v dotazníku o výši platu dotazovaných. Tady platí, že šance, že bude otázka o výši platu nezodpovězena, závisí právě na výši platu dotazovaného. Další možností je případ, kdy jsou položky dotazníku nezodpovězeny kvůli nepřítomnosti samotného účastníka v místě bydliště, protože je například v zaměstnání. Další dva příklady můžeme spojit s pokusy. První z nich nastává, když šance, že subjekt opustí klinický pokus, závisí na jeho reakci na ošetření. Druhý případ nastane, když subjekt může být vyřazen z experimentu, jestliže je jeho stav nedostatečně kontrolován [14].

Mějme kompletní datovou matici $\mathbf{Y}= (y_{ij})$ a matici indikující chybějící hodnoty $\mathbf{M}= (m_{ij})$, kde

$$m_{ij} = \begin{cases} 1 & y_{ij} \text{ chybí} \\ 0 & y_{ij} \text{ je přítomna} \end{cases} .$$

Mechanismus vzniku chybějících hodnot je charakterizován podmíněným rozdělením pravděpodobnosti matice $\mathbf{M}$ za podmínky $\mathbf{Y}$, tj.

$$f\left(\mathbf{M}|\mathbf{Y}, \phi\right),$$

kde ϕ označuje neznámé parametry [11].

2.1.1 MCAR - Missing Completely At Random

Česky můžeme říci, že pozorování v tomto mechanismu chybějí zcela náhodně. Tento případ nastává, jestliže chybějící hodnoty nezávisí na hodnotách proměnných

v $\mathbf{Y}$, ať pozorovaných nebo chybějících. Vyjádřeno pomocí podmíněného rozdělení pravděpodobnosti

$$f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\phi)$$

pro každou složku $\mathbf{Y}$ a ϕ [11]. Neboli, pravěpodobnost, že pozorování Y_{ij} je chybějící, nesouvisí se samotnou hodnotou Y_{ij} ani s ostatními hodnotami proměnných. Neznamená to však, že by byl náhodný model chybějících hodnot jako takový. Náhodný je pouze charakter chybějících hodnot v závislosti na datovém souboru.

Důvodem vzniku MCAR dat může být podle povahy problému například, že dané zařízení selhalo, bylo škaredé počasí nebo data nebyla správně zaznamenána. Podle toho můžeme příkladů takto vzniklých dat nalézt hned několik. Data MCAR mohou vzniknout, jestliže účastník není schopen se vrátit k pokračujícímu dotazování z důvodu nesouvisejícím s danou studií jako je nepříznivé počasí, nemoc nebo smrt v rodině. Nebo jako další uveďme případ, kdy porucha počítače zabrání znovuzískání jen určitých hodnot dat účastníků a ne ostatních. Znamená to tak, že ony okolnosti vznikly zcela náhodně, tj. nelze nalézt jakýkoli vztah mezi předmětem, modelem nebo měřeními studie [2].

Na druhou stranu [5] vysvětluje podrobněji situaci, kdy se o MCAR nejedná. Například chybějící data o rodinném příjmu nebudou považována za MCAR, pokud osoby s nízkými příjmy budou méně pravděpodobněji odpovídat na otázky ohledně rodinného příjmu než osoby s příjmy vyššími. Nebo podobně, jestliže běloši pravděpodobněji vynechají otázku o příjmu než afroameričané, opět nemůžeme považovat chybějící data za zcela náhodná. V tomto případě je nepřítomnost dat ve vzájemném vztahu s etnickým původem dotazovaného.

2.1.2 MAR - Missing At Random

Mechanismus MAR představuje situaci, která stojí uprostřed mezi dvěmi krajními mechanismy. Nyní označme $\mathbf{Y}_{obs}$ jako napozorované složky datové matice $\mathbf{Y}$ a $\mathbf{Y}_{mis}$ jako složky chybějící. O mechanismu MAR hovoříme, jsouli chybějící hodnoty závislé pouze na pozorovaných složkách $\mathbf{Y}_{obs}$ matice $\mathbf{Y}$ a ne

na složkách, které v matici $\mathbf{Y}$ chybí, tj.

$$f\left(\mathbf{M}|\mathbf{Y},\phi\right) = f\left(\mathbf{M}|\mathbf{Y}_{obs},\phi\right)$$

pro každou složku $\mathbf{Y}_{mis}$, ϕ [11]. Tento mechanismus vzniku chybějících hodnot generuje data chybějící náhodně. Víme, že při MAR pravděpodobnost chybějících hodnot závisí na pozorovaných hodnotách. Toto však nekoresponduje s intuitivním chápáním náhodnosti. Důležitou myšlenkou tedy je, že MAR může být vyjádřen výhradně z hlediska měření, která jsou napozorována [14].

Příkladem může být situace popsaná v [2]. Uvažujme, že členové jednoho pohlaví (P1) prozradí svou hmotnost (H) v dotazníku méně pravděpodobněji než druzí. Pravděpodobnost, že hodnota H je chybějící, závisí na těch hodnotách P1, pro která jsou data k dispozici. V tomto případě jsou hodnoty H považovány za MAR, protože nepřítomnost hodnoty H určuje, jestli se jedná o muže nebo ženu a ne to kolik skutečně jednotliví dotazovaní váží. Uvažujme podle [5], že by deprimovaní lidé mohli mít větší sklon k nezodpovězení otázky o jejich příjmu. Tudíž nahlášený příjem bude ve vztahu s depresemi. Deprimovaní lidé by také mohli mít obecně nižší příjmy, čili máme-li velký výskyt chybějících hodnot mezi deprimovanými jedinci, stávající průměrný příjem by měl být nižší než by tomu bylo bez chybějících hodnot. Nicméně, jestliže mezi deprimovanými pacienty pravděpodobnost nenahlášeného příjmu nebyla ve vztahu s úrovní příjmu, potom bychom data považovali za MAR.

2.1.3 NMAR - Not Missing At Random

Pokud neplatí ani MCAR ani MAR, nastává třetí varianta mechanismu vzniku chybějících hodnot, a to NMAR. Znamená to tedy, že rozdělení $\mathbf{M}$ závisí na chybějících hodnotách v matici $\mathbf{Y}$. Tento mechanismus je v mnoha situacích tím správným a nejpravděpodobnějším. Bohužel představuje nejvíce problematický model. Jinými slovy NMAR nastává, jestliže je pravděpodobnost chybějících hodnot ve vztahu s hodnotami, které jsou samy o sobě nepřítomny [2], [14].

Například lidé s vyšší hmotností, mohou méně pravděpodobněji odhalit svou hmotnost v průzkumu. Tedy je-li nezveřejnění váhy účastníka výzkumu spojeno

s jeho hmotností a není v souvislosti s žádnou jinou pozorovanou proměnnou v analýze, potom pravděpodobnost, že informace o hmotnosti dotazovaného je chybějící, závisí výhradně právě na jeho vyšší hmotnosti. Údaj o hmotnosti však není k dispozici, protože je onou chybějící hodnotou [2]. Na tomto místě uvažujme opět příklad s osobami s pokleslou náladou. Studujeme duševní zdraví osob. Právě jedinci, jimž byly diagnostikovány depresivní nálady, méně pravděpodobněji nahlásí svůj duševní stav než ostatní lidé. Tudíž data jsou NMAR [5].

2.2 Regresní imputace chybějících hodnot

Než čtenáři v následujícím představíme jednu z neujžívanějších imputačních metod, zmíníme se stručně o podstatě regresní analýzy, o kterou se daná metoda opírá. Teoretické poznatky jsou převzaty z [7].

Při regresi uvažujeme dva typy proměnných. První z nich je závisle proměnná neboli vysvětlovaná, která je určena nezávisle proměnnou neboli vysvětlující. Závisle proměnnou můžeme uvažovat jako náhodnou veličinu Y, která má při dané hodnotě vysvětlující veličiny x určité rozdělení pravděpodobnosti. V praktických úlohách je nutné ověřit nenáhodnost vysvětlující veličiny v souladu s daným problémem, aby bylo dosaženo předpokladů metody.

Z mnoha úloh z praxe vyplývá, že změny závisle proměnné nemůžeme vysvětlit pouze jedinou vysvětlující proměnnou. Zvyšujeme tedy počet nenáhodných vysvětlujících proměnných a tím pádem můžeme hovořit o *vícenásobné regresi*. Vytvoříme regresní funkci, která určuje podmíněnou střední hodnotu veličiny Y za podmínky $x_1, \ldots, x_p$. Toutu funkcí je lineární funkce proměnných $x_1, \ldots, x_p$,

$$\mathrm{E}(Y|(x_1, x_2, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

Potom pro výsledek i-tého pozorování platí

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde $\beta_0, \ldots, \beta_p$ jsou neznámé parametry, $x_1, \ldots, x_p$ jsou vysvětlující proměnné, $x_{1i}, \ldots, x_{pi}$ i-tá pozorování celkem p proměnných a ε_i je náhodná chyba při i-tém

pozorování. Uvedenou situaci můžeme přepsat do regresního modelu

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

za předpokladu

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{o}, \qquad \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n,$$

kde $\mathbf{o}$ označuje sloupcový nulový vektor a $\mathbf{I}_n$ je jednotková matice řádu n.

Odhad vektorového parametru $\boldsymbol{\beta}$ vypočítáme za předpokladu plné sloupcové hodnosti matice $\mathbf{X}$ podle známého vztahu

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^T.$$

Regresní imputace se obyčejně používá k nahrazení nezodpovězených položek, máme-li k dispozici pomocná data. V běžné praxi pak počítáme odhady chybějících hodnot tak, že s dříve imputovanými hodnotami zacházíme jako s napozorovanými daty (hrají roli vysvětlujících proměnných) a využíváme výše uvedeného vztahu pro odhady neznámých parametrů (a následné imputaci) metodou nejmenších čtverců. O dané problematice pojednává [11] a [13].

Pro vysvětlení principu této metody uvažujeme nejprve jednorozměrnou neúplnou proměnnou s plně napozorovanými $Y_1, \dots, Y_{k-1}$ a Y_k pozorovanou pro prvních r případů a chybějící pro posledních $n - r$ pozorování. Regresní imputace počítá regresi proměnné Y_k na $Y_1, \dots, Y_{k-1}$ založenou na r kompletních případech a následně chybějící hodnoty vyplňuje předpovědmi z regrese [11].

Při regresní imputaci ve vícerozměrném modelu můžeme aplikovat metody imputace iterativně na proměnné s chybějícími hodnotami v datovém souboru. Uvažujeme chybějící hodnoty uspořádané do matice $\mathbf{Y}$ se sloupci odpovídajícími

proměnným $Y_1, \ldots, Y_k$ a dále uvažujeme plně napozorované prediktory $\mathbf{X}$. Nejprve budeme imputovat všechny chybějící hodnoty v $\mathbf{Y}$ vyžitím nějakého robustnějšího přístupu, například jednotlivé proměnné zvlášť s použitím predikotů uspořádaných do matice $\mathbf{X}$. Získané odhady chybějících hodnot pak budou sloužit jako iniciační pro další průběh imputace. V něm budeme nejprve imputovat Y_1 danými $Y_2, \ldots, Y_k$ a $\mathbf{X}$, potom Y_2 pomocí Y_1, $Y_3 \ldots, Y_k$ a $\mathbf{X}$, kde využijeme nově imputovaných hodnot pro Y_1; Y_3 pomocí Y_1, Y_2, Y_4,... a $\mathbf{X}$ s imputacemi v Y_1 a Y_2, atd. Tak pokračujeme dále ve vytváření každé proměnné, dokud není aproximativně dosaženo konvergence [13].

Iterativní regresní imputace má oproti jiným metodám výhodu, že soubor jednotlivých regresních modelů (jeden pro každou proměnnou matice $\mathbf{Y}$) je lépe prezentovatelný. Tedy dovoluje sestavit možný a rozumný model v každém kroku.

3 Imputace pro kompoziční data

Vzhledem ke specifickému charakteru kompozičních dat, popsaném v kapitole 1, musíme uvažovat jisté odlišnosti i při metodách imputace chybějících hodnot v kompozicích.

Existují dva přístupy, jak odhadovat chybějící hodnoty v kompozicích. Určitý postup byl například navržen v [12], tento je ovšem silně vázán na konkrétní součet složek kompozice, představující pouze zvolenou reprezentaci observace. V [11] je navržen jiný postup v návaznosti na definici kompozice s úvahou pouze o podílech mezi složkami. Kompozice obsahují informace právě v podílech a v mnoha praktických úlohách není součet složek konstantní. Pokud je omezení konstantním součtem požadováno, můžeme provést operaci uzávěr kompozice.

Nejjednodušší možnost imputace chybějících hodnot v kompozici je nahradit je geometrickým průměrem ze všech dostupných hodnot v dané složce. Následně můžeme příslušnou hodnotu geometrického průměru přenásobit podílem součtu známých složek z nekompletního pozorování a součtem hodnot odpovídajících složkám vektoru geometrických průměrů. Takto provedená imputace je ovšem velmi nekvalitní a zcela ignoruje mnohorozměrnou strukturu dat. Dále vás proto podrobněji seznámíme se dvěma metodami, které využívají vícerozměrné informace dat pro imputaci. Čerpaly jsme z [8].

3.1 Algoritmus k nejbližších sousedů

Algoritmus k-nn neboli *k-nearest neighbor algoritmus* je jedním z jednodušších metod imputace chybějících hodnot [18]. Počátky algoritmu můžeme najít v 50. letech 20. století ve Spojených státech. V 70. letech byly některé formální vlastnosti přepracovány a to dalo vznik algoritmu v češtině označovaném jako *k-nejbližších sousedů* [9].

Algoritmus k-nn je učící se algoritmus, kterého se využívá v mnoha aplikacích při různých vyznamech dat jako rozpoznání statistického modelu, zpracování obrazu a jiných. Mezi úspěšné praktické aplikace zahrnujeme rozpoznávání ru-

kopisu, rozpoznání satelitních snímků nebo EKG obrazu. Algoritmus je velmi jednoduchý a patří mezi nejjednodušší mechanické výukové algoritmy. Používá se ke klasifikaci objektů, ale také právě k imputaci chybějících hodnot za užití k "nejbližších" pozorování v datovém souboru. V praktických úlohách je k typicky malé kladné přirozené číslo, udávané v jednotkách nebo desítkách spíš než ve stovkách nebo tisících (v podmínkách klasifikace). Je zřejmé, že náročnost výpočtů roste s rostoucím k. Avšak výhodou je, že vyšší hodnoty k obvykle poskytují lepší výsledky imputace [10].

Obecně pro standardní mnohorozměrná data platí, že nejprve určíme k jako parametr tohoto algoritmu. Další krok představuje nalezení k nejbližších sousedů. Jako nejbližší sousedy uvažujeme všechna pozorování, je-li jejich vzdálenost (například euklidovská v případě standardních dat) k požadovanému objektu menší nebo rovna k-té nejmenší vzdálenosti.

I u kompozičních dat princip algoritmu spočívá ve využití míry vzdálenosti k nalezení k nejvíce podobných pozorování v datovém souboru ke konkrétní kompozici s chybějícími složkami. Chceme takto nahradit chybějící hodnoty pomocí dostupných informací o hodnotách složek k nejbližších sousedních pozorování. Protože pracujeme s kompozicemi, musíme k tomuto účelu použít příslušnou Aitchisonovu vzdálenost. Následující odstavce jsou převzaty z [8].

Uvažujme situaci, kdy kompozice obsahuje chybějící hodnoty v několika složkách. Potom může být imputace prováděna dvěma způsoby:

1. Nejprve můžeme imputovat hodnoty současně pro všechny složky tak,
    a) že hledáme k nejbližších sousedů mezi všemi kompletními pozorováními,
    b) že hledáme k nejbližších sousedů mezi pozorováními, která mohou být nekompletní, ale kde je informace ve složkách na místě imputace a kde jsou k dispozici informace i v některých dalších složkách.
2. Hodnoty imputujeme postupně, přičemž
    a) budeme hledat k nejbližších sousedů mezi pozorováními, která sice nemusí být kompletní, ale kde je k dispozici informace související s ne-

chybějícími složkami a navíc je obsažena informace v proměnné určené k imputaci,

b) hledáme sousedy mezi pozorováními, kde vedle proměnné k imputaci nechybí i některé další složky.

V prvním způsobu odhadu chybějících složek kompozice je využito informací o stejných k pozorováních, zatímco u druhého způsobu se k pozorování může měnit během sekvenční imputace. Je to z důvodu, že obecně více sousedů uvažovaných pro imputaci povede na jistější výsledky imputace. Jednodušší a oblíbenou možností je hledat k nejbližších sousedů mezi všemi kompletními kompozicemi, což odpovídá metodě 1a.

Pro imputaci chybějících složek kompozice využíváme mediánu z odpovídajících složek k nejbližších sousedů. Nicméně musíme nejprve přizpůsobit složky shodně s celkovým součtem složek kompozice. Toto nebylo nutné pro nalezení k nejbližších sousedů, protože Aitchisonova vzdálenost je stejná pro každou kompozici $\mathbf{x}$ a $\mathbf{y}$ patřící do ekvivalentní třídy $\underline{\mathbf{x}}$ a $\underline{\mathbf{y}}$, viz kapitola 1.

Uvažujme kompozici $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})^T$, $i = 1, \ldots, n$, o n pozorováních a nechť $\mathrm{M}_i \subset \{1, \ldots, D\}$ označuje množinu indexů odkazující na chybějící složky kompozice $\mathbf{x}_i$. Potom $\mathrm{O}_i = \{1, \ldots, D\} \setminus \mathrm{M}_i$ odkazuje na pozorovanou část $\mathbf{x}_i$. Pro imputování chybějících složek x_{ij}, pro každé $j \in \mathrm{M}_i$, uvažujeme mezi všemi zbývajícími kompozicemi ty, které mají nechybějící složku na pozici j a O_i, a počítáme k nejbližších sousedů $x_{i1}, \ldots, x_{ik}$ ke kompozici $\mathbf{x}_i$ s využitím Aitchisonovy vzdálenosti (metoda 1b). Tedy pro imputaci nás zajímají j-té složky všech k nejbližších sousedů.

Nejprve musíme upravit tyto složky pomocí koeficientu srovnávající velikost složek v O_i. Upravující koeficient můžeme uvažovat ve tvaru

$$f_{ii_l} = \frac{\sum_{o \in \mathrm{O}_i} x_{io}}{\sum_{o \in \mathrm{O}_i} x_{i_l o}}, \qquad \text{pro } l = 1, \ldots, k. \tag{1}$$

Použitím těchto koeficientů jako vah pozorování učiníme k nejbližší sousedy po-

rovnavatelnými. Imputovaná hodnota nahrazující chybějící buňku x_{ij} je pak

$$x^*_{ij} = \text{med}\{f_{ii_1}x_{i_1j}, \ldots, f_{ii_k}x_{i_kj}\}, \tag{2}$$

kde med značí medián příslušných hodnot v argumentu funkce. Použitím právě mediánu získáme určitou robustnost vůči odlehlým hodnotám v j-té složce nejbližších sousedů. I když je volba upravovacího koeficientu v souladu s definicí kompozice, můžeme preferovat jeho robustnější verzi. Jednou z možností poskytující stabilnější výsledky je

$$f^*_{ii_l} = \frac{\text{med}_{o\in \text{O}_i}\ x_i}{\text{med}_{o\in \text{O}_i}\ x_{i_lo}}, \qquad \text{pro } l = 1, \ldots, k. \tag{3}$$

Imputace pomocí k-nn je numericky stabilní metoda, ale má svá omezení. Prvním z nich je, že musí být určen optimální počet k nejbližších sousedů. Ideálně může být toto číslo k nalezeno v rámci simulace náhodného nastavení pozorovaných hodnot za chybějící, odhadnutím těchto chybějících hodnot založených na různých volbách k a pozorováním chyb mezi imputovanými a prvotně naměřenými hodnotami. Za optimální považujeme ta k, která způsobují nejmenší chybu mezi imputovanými a naměřenými hodnotami. Další omezení se týká malého rozsahu výběru. Může se stát, že pomocí Aitchisonovy vzdálenosti dostaneme nejbližší sousedy obsahující horší informaci pro odhadování chybějících hodnot než data, která leží dále od nich. Nastěstí má v praxi většina datových souborů rozumnou velikost. Nakonec imputace pomocí k nejbližších sousedů nezohledňuje mnohorozměrné vztahy mezi složkami kompozice. Ty jsou uvažovány pouze nepřímo, když nejbližší sousedy hledáme. Z tohoto pohledu může být kvalita imputace zlepšena, použijeme-li upravenou metodu regresní imputace, popsanou v další kapitole.

Přesný princip výpočtu pomocí metody k-nn ukážeme na následujícím příkladu. Při výpočtu budeme postupovat podle výše uvedného postupu.

Příklad 2: Mějme k dispozici data z [1], str. 395, v originále nazývající se Household expenditures data. Datový soubor popisuje výdaje domácností na pět

různých skupin komodit u 20 samostatně žijících mužů. Hodnoty jsou uváděny v bývalých hong-kongských dolarech. První sloupec v tabulce dále označený jako "housing" zahrnuje náklady na bydlení včetně tepla a elektřiny. Druhá složka "foodstuff" reprezentuje výdaje na potraviny, třetí "alcohol&tobacco" výdaje za alkohol a tabákové výrobky. Čtvrtá složka pojmenovaná "others" označuje výdaje za jiné zboží jako oblečení, obuv a další dlouhodobé zboží. Poslední složka "services" představuje výdaje na služby a zahrnuje náklady na přepravu a dopravní prostředky.

Z původního datového souboru jsme vynechali třetí řádek, protože tyto hodnoty se značně liší od ostatních v datovém souboru a tyto odlehlé hodnoty by mohly výrazně ovlivnit výsledky imputace. Abychom mohli imputovat chybějící hodnoty, musíme některé hodnoty uměle odstranit. Tím v datovém souboru dostaneme požadované chybějící hodnoty; vybrali jsme x_{72}, x_{75} a $x_{14,2}$. V závěru pak porovnáme námi imputované hodnoty pomocí algoritmu k-nn s původními hodnotami, které jsme odstranili. Původní hodnoty v datovém souboru jsou rovny $x_{72} = 305$, $x_{75} = 112$ a $x_{14,2} = 386$. Ty jsme pro naše účely označili NA (Not Available). Následuje tabulka hodnot s již chybějícími hodnotami.

i	housing	foodstuff	alcohol & tobacco	others	services
1	640	328	147	169	196
2	1800	484	515	2291	912
3	616	331	126	117	149
4	875	368	191	290	275
5	770	364	196	242	236
6	990	415	284	588	420
7	414	**NA**	94	68	**NA**
8	1394	440	393	1161	636
9	1285	374	363	785	487
10	1102	469	243	496	388
11	1717	452	452	1977	832
12	1549	454	424	1345	676
13	838	386	155	208	222
14	845	**NA**	211	317	280
15	1130	394	271	490	386
16	1765	466	524	2133	822
17	1195	443	329	974	523
18	2180	521	553	2781	1010
19	1017	410	225	419	345

Řešení:

1. Podle výše popsaného postupu určíme množiny M_i a O_i podle výskytu chybějících hodnot v jednotlivých kompozicích rozdělených do řádků tabulky. Hodnoty NA se nachází v sedmé a čtrnácté kompozici, tj. $M_7=\{2,5\}$ a $O_7=\{1,3,4\}$, $M_{14}=\{2\}$ a $O_{14}=\{1,3,4,5\}$.

2. V závislosti na tom, v jaké kompozici se nachází chybějící hodnota, budeme počítat Aitchisonovu vzdálenost mezi subkompozicí odpovídající kompozici s hodnotou NA a ostatními subkompozicemi se složkami přítomnými na místech jako v řádku s NA (metoda 1b).

   Nejprve budeme počítat imputaci pro kompozici s menším počtem chybějících hodnot, tj. pro kompozici ve čtrnáctém řádku datového souboru. Abychom mohli imputovat hodnotu $x_{14,2}$, počítáme Aitchisonovy vzdálenosti mezi subkompozicí složek ze čtrnáctého řádku $(x_{14,1}, x_{14,3}, x_{14,4}, x_{14,5})$ a všemi ostatními subkompozicemi, které mají hodnoty přítomny na stejných místech

jako čtrnáctá kompozice. Při výpočtu vzdáleností vynecháme sedmý řádek, protože zde v pátém sloupci hodnota chybí a tudíž by nebyly subkompozice určeny stejnými složkami. Pro lepší orientaci vyznačíme uvažovanou subkompozici v tabulce tučně a vynecháme druhou složku všech kompozic. Vypočítané vzdálenosti doplníme do tabulky do sedmého sloupce s označením $d_A(\mathrm{x}_i, \mathrm{x}_{14})$, kde $i = 1, \ldots, 19,\ i \neq 7, 14$.

i	housing	foodstuff	alcohol & tobacco	others	services	$d_A = (\mathrm{x}_i, \mathrm{x}_{14})$
1	640		147	169	196	0.2657
2	1800		515	2291	912	0.9471
3	616		126	117	149	0.4952
4	875		191	290	275	0.1096
5	770		196	242	236	0.1546
6	990		284	588	420	0.3358
7	414	NA	94	68	NA	-
8	1394		393	1161	636	0.6079
9	1285		363	785	487	0.3635
10	1102		243	496	388	0.2209
11	1717		452	1977	832	0.8949
12	1549		424	1345	676	0.6517
13	838		155	208	222	0.3021
14	**845**	**NA**	**211**	**317**	**280**	-
15	1130		271	490	386	0.1378
16	1765		524	2133	822	0.8977
17	1195		329	974	523	0.5978
18	2180		553	2781	1010	0.9953
19	1017		225	419	345	0.1548

Pro lepší vysvětlení postupu čtenáři ukážeme výpočet první Aitchisonovy vzdálenosti v tabulce označené $d_A = (\mathrm{x}_1, \mathrm{x}_{14})$.

Nejprve vyjádříme subkompozice obecně s označením a poté konkrétně se složkami s vynecháním právě druhé složky,

$\mathrm{x}_1 = (x_{11}, x_{13}, x_{14}, x_{15}) = (640, 147, 169, 196)$,

$\mathrm{x}_{14} = (x_{14,1}, x_{14,3}, x_{14,4}, x_{14,5}) = (845, 211, 317, 280)$.

Nyní dosadíme do vzorce pro výpočet Aitchisonovy vzdálenosti a dostaneme vzdálenost pro první subkompozici v datovém souboru.

$$d_A(\mathbf{x}_1, \mathbf{x}_{14}) = \left\{ \frac{1}{2 \cdot 4} \sum_{i=1}^{4} \sum_{j=1}^{4} \left(\ln \frac{x_{1i}}{x_{1j}} - \ln \frac{x_{14,i}}{x_{14,j}} \right)^2 \right\}^{\frac{1}{2}} = \left\{ \frac{1}{8} \sum_{i=1}^{4} \left[\left(\ln \frac{x_{1i}}{x_{11}} - \ln \frac{x_{14,i}}{x_{14,1}} \right)^2 + \right. \right.$$

$$\left. \left. + \left(\ln \frac{x_{1i}}{x_{13}} - \ln \frac{x_{14,i}}{x_{14,3}} \right)^2 + \left(\ln \frac{x_{1i}}{x_{14}} - \ln \frac{x_{14,i}}{x_{14,4}} \right)^2 + \left(\ln \frac{x_{1i}}{x_{15}} - \ln \frac{x_{14,i}}{x_{14,5}} \right)^2 \right] \right\}^{\frac{1}{2}} =$$

$$= \left\{ \frac{1}{8} \sum_{i=1}^{4} \left[\left(\ln \frac{x_{1i}}{640} - \ln \frac{x_{14,i}}{845} \right)^2 + \left(\ln \frac{x_{1i}}{147} - \ln \frac{x_{14,i}}{211} \right)^2 + \left(\ln \frac{x_{1i}}{169} - \ln \frac{x_{14,i}}{317} \right)^2 + \right. \right.$$

$$\left. \left. + \left(\ln \frac{x_{1i}}{196} - \ln \frac{x_{14,i}}{280} \right)^2 \right] \right\}^{\frac{1}{2}} = \left\{ \frac{1}{8} \left[\left(\ln \frac{640}{640} - \ln \frac{845}{845} \right)^2 + \left(\ln \frac{147}{640} - \ln \frac{211}{845} \right)^2 + \right. \right.$$

$$+ \left(\ln \frac{169}{640} - \ln \frac{317}{845} \right)^2 + \left(\ln \frac{196}{640} - \ln \frac{280}{845} \right)^2 + \left(\ln \frac{640}{147} - \ln \frac{845}{211} \right)^2 +$$

$$+ \left(\ln \frac{147}{147} - \ln \frac{211}{211} \right)^2 + \left(\ln \frac{169}{147} - \ln \frac{317}{211} \right)^2 + \left(\ln \frac{196}{147} - \ln \frac{280}{211} \right)^2 +$$

$$+ \left(\ln \frac{640}{169} - \ln \frac{845}{317} \right)^2 + \left(\ln \frac{147}{169} - \ln \frac{211}{317} \right)^2 + \left(\ln \frac{169}{169} - \ln \frac{317}{317} \right)^2 +$$

$$+ \left(\ln \frac{196}{169} - \ln \frac{280}{317} \right)^2 + \left(\ln \frac{640}{196} - \ln \frac{845}{280} \right)^2 + \left(\ln \frac{147}{196} - \ln \frac{211}{280} \right)^2 +$$

$$\left. \left. + \left(\ln \frac{169}{196} - \ln \frac{317}{280} \right)^2 + \left(\ln \frac{196}{196} - \ln \frac{280}{280} \right)^2 \right] \right\}^{\frac{1}{2}} = \mathbf{0.2657}$$

3. Nyní podle volby $k = 1, 2, 3, 4, 5$ budeme rozlišovat kolik nejbližších sousedů zahrneme do výpočtu imputované hodnoty. V závislosti na tomto čísle se budou měnit výsledky imputace. Můžeme proto rozhodnout o nejvhodnější volbě čísla k. Odpovídající hodnoty poté zahrneme do upravujícího koeficientu, jež počítáme podle vzorce (1). Následuje tabulka k nejbližších sousedů podle volby k.

pořadí	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
1.	0.1096	0.1096	0.1096	0.1096	0.1096
2.		0.1378	0.1378	0.1378	0.1378
3.			0.1546	0.1546	0.1546
4.				0.1548	0.1548
5.					0.2209

4. Na tomto místě provedeme výpočet upravujících koeficientů pro chybějící hodnotu ve čtrnácté kompozici podle vzorce

$$f_{ii_l} = \frac{\sum_{o \in \mathrm{O}_i} x_{io}}{\sum_{o \in \mathrm{O}_i} x_{i_l o}}, \qquad \text{pro } l = 1, \ldots, k.$$

Do tabulky můžeme shrnout hodnoty složek subkompozic pro jednotlivé nejbližší sousedy podle zvoleného k, které poté dosadíme do vzorce pro výpočet upravujícího koeficientu.

k-tý soused	i	$d_A = (\mathrm{x}_i, \mathrm{x}_{14})$	x_{i1}	x_{i3}	x_{i4}	x_{i5}
1.	4	0.1096	875	191	290	275
2.	15	0.1378	1130	271	490	386
3.	5	0.1546	770	196	242	236
4.	19	0.1548	1017	225	419	345
5.	10	0.2209	1102	243	496	388

Výpočet se mění v závislosti na čísle k následovně.

Pro $k = 1$,

$$f_{14,14_1} = \frac{x_{14,1} + x_{14,3} + x_{14,4} + x_{14,5}}{x_{14_1,1} + x_{14_1,3} + x_{14_1,4} + x_{14_1,5}} = \frac{845 + 211 + 317 + 280}{875 + 191 + 290 + 275} = \mathbf{1.0135},$$

pro $k = 2$,

$$f_{14,14_2} = \frac{x_{14,1} + x_{14,3} + x_{14,4} + x_{14,5}}{x_{14_2,1} + x_{14_2,3} + x_{14_2,4} + x_{14_2,5}} = \frac{845 + 211 + 317 + 280}{1130 + 271 + 490 + 386} = \mathbf{0.7259},$$

pro $k = 3$,

$$f_{14,14_3} = \frac{x_{14,1} + x_{14,3} + x_{14,4} + x_{14,5}}{x_{14_3,1} + x_{14_3,3} + x_{14_3,4} + x_{14_3,5}} = \frac{845 + 211 + 317 + 280}{770 + 196 + 242 + 236} = \mathbf{1.1447},$$

pro $k = 4$,

$$f_{14,14_4} = \frac{x_{14,1} + x_{14,3} + x_{14,4} + x_{14,5}}{x_{14_4,1} + x_{14_4,3} + x_{14_4,4} + x_{14_4,5}} = \frac{845 + 211 + 317 + 280}{1017 + 225 + 419 + 345} = \mathbf{0.8240},$$

pro $k = 5$,

$$f_{14,14_5} = \frac{x_{14,1} + x_{14,3} + x_{14,4} + x_{14,5}}{x_{14_5,1} + x_{14_5,3} + x_{14_5,4} + x_{14_5,5}} = \frac{845 + 211 + 317 + 280}{1102 + 243 + 496 + 388} = \mathbf{0.7416}.$$

5. Na základě těchto upravujících koeficientů můžeme vypočítat imputaci chybějící hodnoty $x_{14,2}$ podle vztahu

$$x^*_{ij} = \text{med}\{f_{ii_1}x_{i_1j},\ \ldots\ ,\ f_{ii_k}x_{i_kj}\}.$$

Imputované hodnoty dostaneme následujícím postupem podle zvoleného k.

Pro $k = 1$,

$$x^*_{14,2} = \text{med}\left\{f_{14,14_1}x_{14_1,2}\right\} = \text{med}\left\{1.0135 \cdot 368\right\} = \text{med}\left\{372.968\right\} = \mathbf{372.968},$$

pro $k = 2$,

$$\begin{aligned} x^*_{14,2} &= \text{med}\left\{f_{14,14_1}x_{14_1,2},\ f_{14,14_2}x_{14_2,2}\right\} = \text{med}\left\{1.0135 \cdot 368, 0.7259 \cdot 394\right\} = \\ &= \text{med}\left\{372.968, 286.0046\right\} = \mathbf{372.968}, \end{aligned}$$

pro $k = 3$,

$$\begin{aligned} x^*_{14,2} &= \text{med}\left\{f_{14,14_1}x_{14_1,2},\ f_{14,14_2}x_{14_2,2},\ f_{14,14_3}x_{14_3,2}\right\} = \text{med}\left\{1.0135 \cdot 368,\right. \\ &\left. 0.7259 \cdot 394, 1.1447 \cdot 364\right\} = \text{med}\left\{372.968, 286.0046, 416.6708\right\} = \\ &= \mathbf{372.968}. \end{aligned}$$

pro $k = 4$,

$$\begin{aligned} x^*_{14,2} &= \text{med}\{f_{14,14_1}x_{14_1,2},\ f_{14,14_2}x_{14_2,2},\ f_{14,14_3}x_{14_3,2},\ f_{14,14_4}x_{14_4,2}\} = \\ &= \text{med}\{1.0135 \cdot 368, 0.7259 \cdot 3941.1447 \cdot 364, 0.8240 \cdot 410\} = \\ &= \text{med}\{372.968, 286.0046, 416.6708, 337.84\} = \mathbf{355.404}, \end{aligned}$$

pro $k = 5$,

$$\begin{aligned} x^*_{14,2} &= \text{med}\{f_{14,14_1} x_{14_1,2},\ f_{14,14_2} x_{14_2,2},\ f_{14,14_3} x_{14_3,2},\ f_{14,14_4} x_{14_4,2},\ f_{14,14_5} x_{14_5,2}\} = \\ &= \text{med}\{1.0135 \cdot 368, 0.7259 \cdot 394, 1.1447 \cdot 364, 0.8240 \cdot 410, 0.7416 \cdot 469\} = \\ &= \text{med}\{372.968, 286.0046, 416.6708, 337.84, 347.8104\} = \mathbf{347.8104}. \end{aligned}$$

Původní hodnota z datového souboru je rovna 386. Z výsledků imputace je patrné, že nejlepší imputace této chybějící hodnoty jsme dosáhli pro první tří nejbližší sousedy, tj. pro $k = 1, 2, 3$ a imputovaná hodnota je po zaokrouhlení rovna 373.

Nyní následuje výpočet imputace zbývajících chybějících hodnot x_{72} a x_{75}. Budeme imputovat obě hodnoty naráz. To tedy znamená, že uvažujeme trojsložkové subkompozice, ze kterých budeme počítat Aitchisonovy vzdálenosti. Upravující faktory budou taktéž pro obě chybějící hodnoty stejné, rozdíl nastane při výpočtu mediánu. Ten budeme počítat pomocí upravujících faktorů a hodnot na pozicích v řádku podle k-tého nejbližšího souseda a ve sloupci s chybějící hodnotou. Následuje tabulka se subkompozicemi, pro které budeme počítat Aitchisonovy vzdálenosti. Ty jsou uvedeny opět v sedmém sloupci datové tabulky.

i	housing	foodstuff	alcohol & tobacco	others	services	$d_A = (\mathrm{x}_i, \mathrm{x}_7)$
1	640		147	169		0.3830
2	1800		515	2291		1.5859
3	616		126	117		0.1773
4	875		191	290		0.5899
5	770		196	242		0.4899
6	990		284	588		0.9683
7	**414**	**NA**	**94**	**68**	**NA**	-
8	1394		393	1161		1.2466
9	1285		363	785		0.9954
10	1102		243	496		0.8353
11	1717		452	1977		1.5332
12	1549		424	1345		1.2901
13	838		155	208		0.4451
14	845		211	317		0.6391
15	1130		271	490		0.7713
16	1765		524	2133		1.5318
17	1195		329	974		1.2368
18	2180		553	2781		1.6303
19	1017		225	419		0.7618

V dalším kroku vybereme na základě výpočtených vzdáleností k nejbližších sousedů. Tyto spolu s hodnotami složek subkompozice odpovídající k-tému nejbližšímu sousedu vypíšeme do tabulky uvedené níže. Právě tyto hodnoty nám poslouží k výpočtu upravujících koeficientu pro zvolené k.

k-tý soused	i	$d_A = (\mathrm{x}_i, \mathrm{x}_7)$	x_{i1}	x_{i3}	x_{i4}
1.	3	0.1773	616	126	117
2.	1	0.3830	640	147	169
3.	13	0.4451	838	155	208
4.	5	0.4899	770	196	242
5.	4	0.5899	875	191	290

Nyní dopočítáme upravující koeficienty, které jsou pro obě chybějící hodnoty totožné. Hodnoty podvektoru $x_{7o} = (414, 94, 68),\ o \in \{1, 3, 4\}$ budeme dosazovat do čitatele, do jmenovatele budeme postupně dosazovat složky subkompozic jednotlivých nejbližších sousedů v závislosti na volbě k. Upravující koeficienty vypadají pro imputaci hodnot x_{72} a x_{75} takto

$$f_{77_1} = \frac{(414+94+68)}{(616+126+117)} = \mathbf{0.6706},$$

$$f_{77_2} = \frac{(414+94+68)}{(640+147+169)} = \mathbf{0.6025},$$

$$f_{77_3} = \frac{(414+94+68)}{(838+155+208)} = \mathbf{0.4796},$$

$$f_{77_4} = \frac{(414+94+68)}{(770+196+242)} = \mathbf{0.4768},$$

$$f_{77_5} = \frac{(414+94+68)}{(875+191+290)} = \mathbf{0.4248},$$

Posledním krokem algoritmu je výpočet mediánu z hodnot násobených upravujícím koeficientem. Použijeme stejné upravující koeficienty, ale ty budeme násobit hodnotami složek ve sloupci, ve kterém byla hodnota NA, tj. hodnotami odpovídajícími druhé a páté složce podle k-tého nejbližšího souseda. Opět uvažujeme různé volby k.

Pro $k = 1$,

$$x^*_{72} = \text{med}\,\{0.6706 \cdot 331\} = \mathbf{221.9686},$$

$$x^*_{75} = \text{med}\,\{0.6706 \cdot 149\} = \mathbf{99.9194},$$

pro $k = 2$,

$$x^*_{72} = \text{med}\,\{0.6706 \cdot 331, 0.6025 \cdot 328\} = \text{med}\,\{221.9686, 197.62\} = \mathbf{209.7943},$$

$$x^*_{75} = \text{med}\,\{0.6706 \cdot 149, 0.6025 \cdot 196\} = \text{med}\,\{99.9194, 118.09\} = \mathbf{109.0047},$$

pro $k = 3$,

$$x^*_{72} = \text{med}\,\{0.6706 \cdot 331, 0.6025 \cdot 328, 0.4796 \cdot 386\} = \text{med}\,\{221.9686, 197.62, 185.1256\} = \mathbf{197.62},$$

$$x^*_{75} = \text{med}\,\{0.6706 \cdot 149, 0.6025 \cdot 196, 0.4796 \cdot 222\} = \text{med}\,\{99.9194, 118.09, 106.4712\} = \mathbf{106.4712},$$

pro $k = 4$,

$$\begin{aligned} x^*_{72} &= \text{med}\,\{0.6706 \cdot 331, 0.6025 \cdot 328, 0.4796 \cdot 386, 0.4768 \cdot 364\} = \\ &= \text{med}\,\{221.9686, 197.62, 185.1256, 173.5552\} = \mathbf{191.3728}, \\ x^*_{75} &= \text{med}\,\{0.6706 \cdot 149, 0.6025 \cdot 196, 0.4796 \cdot 222, 0.4768 \cdot 236\} = \\ &= \text{med}\,\{99.9194, 118.09, 106.4712, 112.5248\} = \mathbf{109.498}, \end{aligned}$$

$k = 5$

$$\begin{aligned} x^*_{72} &= \text{med}\,\{0.6706 \cdot 331, 0.6025 \cdot 328, 0.4796 \cdot 386, 0.4768 \cdot 364, 0.4248 \cdot 368\} = \\ &= \text{med}\,\{221.9686, 197.62, 185.1256, 173.5552, 156.3264\} = \mathbf{185.1256}, \\ x^*_{75} &= \text{med}\,\{0.6706 \cdot 149, 0.6025 \cdot 196, 0.4796 \cdot 222, 0.4768 \cdot 236, 0.4248 \cdot 275\} = \\ &= \text{med}\,\{99.9194, 118.09, 106.4712, 112.5248, 116.82\} = \mathbf{112.5248} \end{aligned}$$

Z výsledků je patrné, že se nám povedla naimputovat hodnota x_{75}. Původní hodnota byla 112 a hodnota získaná imputací pro $k = 5$ je rovna po zaokrouhlení 113. Na druhou stranu jsme s imputací nedosáhli dobrých výsledků pro hodnotu x_{72}, protože původní hodnota byla 305. Nám se nepodařilo ani pro jedno zvolené k dostatečně přiblížit k této hodnotě. Nejbližší hodnoty jsme dosáhli pro $k = 1$, avšak po zaokrouhlení je hodnota 222 na číselné ose od 305 ještě docela hodně vzdálená. Příčinou tohoto poněkud paradoxního výsledku je zřejmě malý rozsah datového souboru a atypičnost kompozice, ve které imputujeme.

Závěrem tedy shrňme, že se nám povedlo naimputovat hodnoty x_{75} a $x_{14,2}$ kvalitním způsobem, tj. původní hodnoty jsou blízké těm imputovaným. Pro chybějící hodnotu x_{72} jsme ve srovnání s původní z datového souboru nedosáhli uspokojivých výsledků.

3.2 Iterativní regresní imputace

V této metodě se využívá principu regresní imputace, o které jsme se zmínili již dříve. V každém kroku iterace se jedna proměnná používá jako závislá a zbývající proměnné využijeme jako regresory. Tedy pro imputaci závisle proměnné je využito vícerozměrné informace. Protože pracujeme s kompozicemi, tak nemůžeme přímo použít původní data v regresi, ale musíme pracovat v transformavaném prostoru. Pro tento účel si vybereme ilr transformaci a využijeme výhodných interpretačních vlastností bilancí. Avšak pro sestavení bilancí potřebujeme kompletní datovou matici, narozdíl od popsané situace se standardními daty. Dalším problémem může být fakt, že pro sestavení bilancí potřebujeme některé nebo dokonce všechny proměnné. Čili musíme bilance vybrat obezřetně. Poznatky této problematiky jsou převzaty z [8].

Uvažujme datovou matici s n pozorováními a D složkami. Složky uvažované ilr transformace, můžeme pro i-tou kompozici $\mathbf{x}_i = (x_{i1}, \dots x_{iD})^T$, $i = 1, \dots, n$ přepsat jako $ilr(\mathbf{x}_i) = \mathbf{z}_i = (z_{i1}, \dots, z_{i,D-1})^T$, kde

$$z_{ij} = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{\sqrt[D-j]{\prod_{l=j+1}^{D} x_{il}}}{x_{ij}}, \qquad \text{pro } j = 1, \dots, D-1. \tag{4}$$

Povšimněme si, se tyto bilance shodují s uvedenými na konci kapitoly 1.4.2. Odpovídající inverzní transformace je $ilr^{-1}(\mathbf{z}_i) = \mathbf{x}_i = (x_{i1}, \dots, x_{iD})^T$ s prvky

$$x_{i1} = \exp\left(-\sqrt{\frac{D-1}{D}} z_{i1}\right), \tag{5}$$

$$x_{ij} = \exp\left(\sum_{l=1}^{j-1} \frac{1}{\sqrt{(D-l+1)(D-l)}} z_{il} - \sqrt{\frac{D-j}{D-j+1}}\ z_{ij}\right), \text{ pro } j = 2, \dots, D-1, \tag{6}$$

$$x_{iD} = \exp\left(\sum_{l=1}^{D-1} \frac{1}{\sqrt{(D-l+1)(D-l)}}\ z_{il}\right). \tag{7}$$

Iterativní algoritmus můžeme shrnout do následujících kroků:

1. Nejprve nalezneme chybějící hodnoty pomocí algoritmu k-nn, který je založen na Aitchisonově vzdálenosti.

2. Uspořádáme proměnné podle počtu chybějících hodnot. Abychom se vyvarovali komplikovanému zápisu, budeme předpokládat, že proměnné už jsou uspořádány, tj. $\mathcal{M}(\mathbf{x}_1) \geq \mathcal{M}(\mathbf{x}_2) \geq \ldots \geq \mathcal{M}(\mathbf{x}_D)$, kde $\mathcal{M}_j$ označuje počet chybějících složek proměnné $\mathbf{x}_j$. Všimněme si, že $\mathbf{x}_j$ nyní označuje j-tý sloupec datové matice.

3. Nechť $l = 1$.

4. Použijeme ilr transformaci na datový soubor kompozice.

5. Označíme $m_l \subset \{1, \ldots, n\}$ indexy těch pozorování, které byly prvotně u proměnné $\mathbf{x}_l$ chybějící. Dále označíme $o_l = \{1, \ldots, n\} \setminus m_l$ indexy odpovídající pozorovaným složkám proměnné $\mathbf{x}_l$. Kromě toho $\mathbf{z}_l^{o_l}$ a $\mathbf{z}_l^{m_l}$ označuje l-tou bilanci s napozorovanými a chybějícími hodnotami. Nechť $\mathbf{Z}_l^{o_l}$ a $\mathbf{Z}_l^{m_l}$, v tomto pořadí, označuje matice se zbývajícími bilancemi korespondujícími s pozorovanými a chybějícícmi buňkami $\mathbf{x}_l$. Navíc, první sloupec matic $\mathbf{Z}_l^{o_l}$ a $\mathbf{Z}_l^{m_l}$ obsahuje jedničky, abychom v následujícím regresním modelu

$$\mathbf{z}_l^{o_l} = \mathbf{Z}_l^{o_l}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{8}$$

   s neznámým regresním parametrem $\boldsymbol{\beta}$ a chybou $\boldsymbol{\varepsilon}$ zachytili též absolutní člen regresní funkce.

6. Odhadneme regresní koeficienty $\boldsymbol{\beta}$ z regrese uvedené výše a použijeme odhadnuté regresní koeficienty $\widehat{\boldsymbol{\beta}}$, abychom nahradili chybějící složky $\mathbf{z}_l^{m_l}$

$$\widehat{\mathbf{z}}_l^{m_l} = \mathbf{Z}_l^{m_l}\widehat{\boldsymbol{\beta}}. \tag{9}$$

7. Použijeme přepočítané bilance pro zpětnou transformaci na simplex podle výše uvedených vzorců (5), (6), (7). V důsledku toho jsou hodnoty, které byly původně chybějící v buňkách m_l proměnné $\mathbf{x}_l$, aktualizovány.

8. Kroky 4.-7. provedeme postupně pro každé $l = 2, \ldots, D$.

9. Opakujeme kroky 3.-8., dokud není euklidovská vzdálenost mezi výběrovými variančními maticemi, počítanými z ilr-transformovaných dat podle vztahu (4) z předešlé a následující iterace menší než stanovená hranice.

Ačkoli nebyl proveden důkaz konvergence této metody (který je ostatně obtížně proveditelný i u její "standardní" podoby), výpočty s reálnými a simulovanými daty (provedeno v [8]) ukázaly, že algoritmus obvykle konverguje po pár iteracích a že po druhé iteraci již není dosaženo zvlášť významného zlepšení.

Následuje příklad popisující krok za krokem regresní imputaci podle výše uvedeného postupu. Pro výpočet použijeme data z příkladu 2, protože pro ně máme k dispozci již naimputované hodnoty pomocí k-nn. Navíc je našim záměrem pro obě provést shrnutí kvality imputace.

Příklad 3: Máme k dispozici datovou matici sestávající se z 19 kompozic o 5 složkách, tj. $n = 19$, $D = 5$. V závislosti na přiřazení proměnných jednotlivým sloupcům datové matice, budeme rozlišovat první a druhou iteraci. Nejprve popíšeme postup imputace v první iteraci, tzn. chceme imputovat dvě složky x_{72} a $x_{14,2}$.

1. Chybějící hodnoty, které jsme z původního datového souboru uměle odstranili, jsme naimputovali pomocí algoritmu k-nn (postup popsán v příkladu 2). Jedná se o hodnoty s označením $x_{72} = 222$, $x_{75} = 113$ a $x_{14,2} = 373$.

2. Nyní uspořádáme proměnné podle počtu chybějících hodnot. Pro další výpočty musíme podotknout, že $\mathbf{x}_j$, $j = 1, \ldots, 5$, dále označuje sloupec odpovídající j-té složce kompozice.

   Nejvíce chybějících hodnot se před imputací pomocí k-nn vyskytovalo ve druhém sloupci původní datové matice, tudíž za $\mathbf{x}_1$ volíme sloupec nazvaný foodstuff. V pátém sloupci původní matice se vyskytovala pouze jediná chybějící hodnota, tzn. $\mathbf{x}_2$ bude dále označovat sloupec services.

Protože v datové matici již žádné chybějící hodnoty nenalezneme, přiřadíme proměnné $\mathbf{x}_3$, $\mathbf{x}_4$ a $\mathbf{x}_5$ zbývajícím sloupcům po řadě zleva. Pro lepší orientaci přiřazení zapišme přehledněji: foodstuff $\rightarrow \mathbf{x}_1$, services $\rightarrow \mathbf{x}_2$, housing $\rightarrow \mathbf{x}_3$, alcohol & tobacco $\rightarrow \mathbf{x}_4$, others $\rightarrow \mathbf{x}_5$. Následuje tabulka hodnot již s uspořádanými sloupci, kde zvýrazněné hodnoty představují imputované hodnoty pomocí k-nn.

i	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
1	328	196	640	147	169
2	484	912	1800	515	2291
3	331	149	616	126	117
4	368	275	875	191	290
5	364	236	770	196	242
6	415	420	990	284	588
7	**222**	**113**	414	94	68
8	440	636	1394	393	1161
9	374	487	1285	363	785
10	469	388	1102	243	496
11	452	832	1717	452	1977
12	454	676	1549	424	1345
13	386	222	838	155	208
14	**373**	280	845	211	317
15	394	386	1130	271	490
16	466	822	1765	524	2133
17	443	523	1195	329	974
18	521	1010	2180	553	2781
19	410	345	1017	225	419

3. Nechť $l = 1$.

4. Ilr transformace pro původní hodnoty složek vypočítáme podle vzorce (4). Pro názornost vyjádříme první čtyři ilr-transformované složky i s dosazením

do příslušného vzorce,

$$z_{11} = \sqrt{\frac{5-1}{5-1+1}}\ \ln\frac{\sqrt[5-1]{x_{12}\cdot x_{13}\cdot x_{14}\cdot x_{15}}}{x_{11}} = \sqrt{\frac{4}{5}}\ \ln\frac{\sqrt[4]{196\cdot 640\cdot 147\cdot 169}}{328} =$$
$$= \mathbf{-0.2934},$$
$$z_{12} = \sqrt{\frac{5-2}{5-2+1}}\ \ln\frac{\sqrt[5-2]{x_{13}\cdot x_{14}\cdot x_{15}}}{x_{12}} = \sqrt{\frac{3}{4}}\ \ln\frac{\sqrt[3]{640\cdot 147\cdot 169}}{196} = \mathbf{0.2158},$$
$$z_{13} = \sqrt{\frac{5-3}{5-3+1}}\ \ln\frac{\sqrt[5-3]{x_{14}\cdot x_{15}}}{x_{13}} = \sqrt{\frac{2}{3}}\ \ln\frac{\sqrt{147\cdot 169}}{640} = \mathbf{-1.1442},$$
$$z_{14} = \sqrt{\frac{5-4}{5-4+1}}\ \ln\frac{x_{15}}{x_{14}} = \sqrt{\frac{1}{2}}\ \ln\frac{169}{147} = \mathbf{0.0986}.$$

Nyní uvádíme tabulku zobrazující všechna transformovaná data.

i	$\mathbf{z}_1$	$\mathbf{z}_2$	$\mathbf{z}_3$	$\mathbf{z}_4$
1	-0.2934	0.2158	-1.1442	0.0986
2	0.7969	0.2972	-0.4124	1.0554
3	-0.4881	0.2915	-1.3260	-0.0524
4	-0.0714	0.2442	-1.0722	0.2953
5	-0.1591	0.2950	-1.0311	0.1491
6	0.1902	0.2317	-0.7225	0.5146
7	-0.4684	0.1751	-1.3427	-0.2290
8	0.5319	0.2613	-0.5916	0.7660
9	0.4941	0.3331	-0.7173	0.5454
10	0.0141	0.2371	-0.9431	0.5045
11	0.7648	0.2829	-0.4873	1.0434
12	0.5910	0.3033	-0.5866	0.8163
13	-0.2926	0.2609	-1.2578	0.2080
14	-0.0450	0.2730	-0.9667	0.2878
15	0.1961	0.2768	-0.9240	0.4188
16	0.7910	0.3659	-0.4185	0.9926
17	0.3687	0.2842	-0.6101	0.7675
18	0.8559	0.3406	-0.4606	1.1421
19	0.0352	0.2448	-0.9779	0.4397

5. Označme množiny indexů odkazující na chybějící a pozorované složky, v tomto pořadí, $m_1 = \{7, 14\}$, $o_1 = \{1, \ldots, 6, 8, \ldots, 13, 15, \ldots, 19\}$.

Dále označme vektory a matice bilancí odpovídající množinám m_1 a o_1,

$$\mathbf{z}_1^{m_1} = \begin{pmatrix} -0.4684 \\ -0.0450 \end{pmatrix}, \qquad \mathbf{Z}_1^{m_1} = \begin{pmatrix} 1 & 0.1751 & -1.3427 & -0.2290 \\ 1 & 0.2730 & -0.9667 & 0.2878 \end{pmatrix},$$

$$\mathbf{z}_1^{o_1} = \begin{pmatrix} -0.2934 \\ 0.7969 \\ -0.4881 \\ -0.0714 \\ -0.1591 \\ 0.1902 \\ 0.5319 \\ 0.4941 \\ 0.0141 \\ 0.7648 \\ 0.5910 \\ -0.2926 \\ 0.1961 \\ 0.7910 \\ 0.3687 \\ 0.8559 \\ 0.0352 \end{pmatrix}, \qquad \mathbf{Z}_1^{o_1} = \begin{pmatrix} 1 & 0.2158 & -1.1442 & 0.0986 \\ 1 & 0.2972 & -0.4124 & 1.0554 \\ 1 & 0.2915 & -1.3260 & -0.0524 \\ 1 & 0.2442 & -1.0722 & 0.2953 \\ 1 & 0.2950 & -1.0311 & 0.1491 \\ 1 & 0.2317 & -0.7225 & 0.5146 \\ 1 & 0.2613 & -0.5916 & 0.7660 \\ 1 & 0.3331 & -0.7173 & 0.5454 \\ 1 & 0.2371 & -0.9431 & 0.5045 \\ 1 & 0.2829 & -0.4873 & 1.0434 \\ 1 & 0.3033 & -0.5866 & 0.8163 \\ 1 & 0.2609 & -1.2578 & 0.2080 \\ 1 & 0.2768 & -0.9240 & 0.4188 \\ 1 & 0.3659 & -0.4185 & 0.9926 \\ 1 & 0.2842 & -0.6101 & 0.7675 \\ 1 & 0.3406 & -0.4606 & 1.1421 \\ 1 & 0.2448 & -0.9779 & 0.4397 \end{pmatrix}.$$

Regresní model

$$\mathbf{z}_1^{o_1} = \mathbf{Z}_1^{o_1}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

je po dosazení tvaru

$$\begin{pmatrix} -0.2934 \\ 0.7969 \\ -0.4881 \\ -0.0714 \\ -0.1591 \\ 0.1902 \\ 0.5319 \\ 0.4941 \\ 0.0141 \\ 0.7648 \\ 0.5910 \\ -0.2926 \\ 0.1961 \\ 0.7910 \\ 0.3687 \\ 0.8559 \\ 0.0352 \end{pmatrix} = \begin{pmatrix} 1 & 0.2158 & -1.1442 & 0.0986 \\ 1 & 0.2972 & -0.4124 & 1.0554 \\ 1 & 0.2915 & -1.3260 & -0.0524 \\ 1 & 0.2442 & -1.0722 & 0.2953 \\ 1 & 0.2950 & -1.0311 & 0.1491 \\ 1 & 0.2317 & -0.7225 & 0.5146 \\ 1 & 0.2613 & -0.5916 & 0.7660 \\ 1 & 0.3331 & -0.7173 & 0.5454 \\ 1 & 0.2371 & -0.9431 & 0.5045 \\ 1 & 0.2829 & -0.4873 & 1.0434 \\ 1 & 0.3033 & -0.5866 & 0.8163 \\ 1 & 0.2609 & -1.2578 & 0.2080 \\ 1 & 0.2768 & -0.9240 & 0.4188 \\ 1 & 0.3659 & -0.4185 & 0.9926 \\ 1 & 0.2842 & -0.6101 & 0.7675 \\ 1 & 0.3406 & -0.4606 & 1.1421 \\ 1 & 0.2448 & -0.9779 & 0.4397 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{17} \\ \varepsilon_{18} \\ \varepsilon_{19} \end{pmatrix}.$$

6. Pro odhad regresního parametru $\boldsymbol{\beta}$ dosadíme do známého vztahu a obdržíme výsledek

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{Z}_1^{o_1 T}\mathbf{Z}_1^{o_1}\right)^{-1}\mathbf{Z}_1^{o_1 T}\mathbf{z}_1^{o_1} = \begin{pmatrix} 0.2587 \\ 1.1473 \\ 0.7519 \\ 0.4891 \end{pmatrix}.$$

Abychom nahradili chybějící složky $\mathbf{z}_1^{m_1}$, dosadíme odhadnuté regresní parametry do rovnice

$$\widehat{\mathbf{z}}_1^{m_1} = \mathbf{Z}_1^{m_1}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 0.1751 & -1.3427 & -0.2290 \\ 1 & 0.2730 & -0.9667 & 0.2878 \end{pmatrix}\begin{pmatrix} 0.2587 \\ 1.1473 \\ 0.7519 \\ 0.4891 \end{pmatrix} = \begin{pmatrix} -0.6619 \\ -0.0141 \end{pmatrix}.$$

7. Inverzní transformaci provedeme v několika krocích. Nejprve odhadnuté hodnoty $\widehat{\mathbf{z}}_1^{m_1}$ dosadíme místo těch před odhadem v matici ilr-transformovaných dat na pozice z_{71} a $z_{14,1}$. Pak provedeme inverzní transformaci pomocí vzorců (5), (6) a (7). Abychom nakonec dostali imputovanou hodnotu

v původních jednotkách datového souboru, musíme hodnotu po inverzní transformaci v prvním sloupci v sedmém a čtrnáctém řádku (nynější pozice hodnot NA při prohozených sloupcích) vynásobit pořadě součtem složek sedmé a čtrnácté kompozice z datového souboru po k-nn imputaci bez NA (tabulka na straně 41). Vše ukážeme i s dosazením.

Data z inverzní ilr transformace označme pro další použití x_{ij}^{ilr}, vypadají takto:

i	$\mathbf{x}_1^{ilr}$	$\mathbf{x}_2^{ilr}$	$\mathbf{x}_3^{ilr}$	$\mathbf{x}_4^{ilr}$	$\mathbf{x}_5^{ilr}$
1	0.2216	0.1324	0.4324	0.0993	0.1142
2	0.0806	0.1520	0.2999	0.0858	0.3817
3	0.2472	0.1113	0.4600	0.0941	0.0874
4	0.1841	0.1376	0.4377	0.0955	0.1451
5	0.2013	0.1305	0.4259	0.1084	0.1339
6	0.1539	0.1557	0.3671	0.1053	0.2180
7	0.2857	0.1171	0.4292	0.0975	0.0705
8	0.1093	0.1580	0.3464	0.0977	0.2885
9	0.1135	0.1478	0.3901	0.1102	0.2383
10	0.1738	0.1438	0.4084	0.0901	0.1838
11	0.0832	0.1532	0.3162	0.0832	0.3641
12	0.1021	0.1520	0.3482	0.0953	0.3024
13	0.2134	0.1227	0.4632	0.0857	0.1150
14	0.1790	0.1391	0.4197	0.1048	0.1574
15	0.1475	0.1445	0.4231	0.1015	0.1835
16	0.0816	0.1440	0.3091	0.0918	0.3735
17	0.1279	0.1510	0.3450	0.0950	0.2812
18	0.0740	0.1434	0.3094	0.0785	0.3947
19	0.1697	0.1428	0.4210	0.0931	0.1734

Abychom dostali imputované hodnoty x_{72}^* a x_{75}^* na původních pozicích v datovém souboru, budeme násobit hodnoty inverzně ilr-transformovaných dat na pozicích x_{17} a $x_{14,1}$ z předešlé tabulky součtem složek kompozic v odpovídajících řádcích původní matice ze strany 41, tedy

$$x^*_{72} = x^{ilr}_{71} \sum_{j=1}^{5} x_{7j} = 0.2857 \cdot (414 + 222 + 94 + 68 + 113) = \mathbf{260},$$

$$x^*_{14,2} = x^{ilr}_{14,1} \sum_{j=1}^{5} x_{14,j} = 0.1790 \cdot (845 + 373 + 211 + 317 + 280) = \mathbf{363}.$$

Závěrem můžeme říci, že se nám povedlo pomocí iterativní regrese naimputovat hodnotu x_{72} kvalitněji než tomu bylo u k-nn, poněvadž pro k-nn je hodnota rovna 222 oproti 260 z iterativní regrese. Na druhou stranu se imputovanou hodnotu pomocí iterativní regrese 363 ve srovnání s hodnotou v k-nn 373 nepodařilo vylepšit. Ale vzhledem k tomu, že rozdíl je pouze o deset jednotek, můžeme usuzovat, že regresní imputace byla celkově úspěšnější než metoda k nejbližších sousedů.

Jelikož se v pátém sloupci původního datového souboru z příkladu 2 nachází ještě jedna chybějící hodnota, budeme postup regresní imputace opakovat. Provedeme tím pádem druhou iteraci této metody pro imputování hodnoty x_{75} v původním datovém souboru s využitím již naimputovaných hodnot v proměnné $\mathbf{x}_1$ z první iterace. Pro lepší názornost přehodíme pořadí první a druhé proměnné, protože je chybějící hodnota nyní již jen u $\mathbf{x}_2$, ostatní proměnné ponecháme stejně. Tímto vznikne nová tabulka hodnot, ve druhém sloupci jsou zvýrazněny již imputované hodnoty z první iterace a hodnota určená k imputaci ve sloupci prvním.

i	$\mathbf{x}_2$	$\mathbf{x}_1$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
1	196	328	640	147	169
2	912	484	1800	515	2291
3	149	331	616	126	117
4	275	368	875	191	290
5	236	364	770	196	242
6	420	415	990	284	588
7	**113**	**260**	414	94	68
8	636	440	1394	393	1161
9	487	374	1285	363	785
10	388	469	1102	243	496
11	832	452	1717	452	1977
12	676	454	1549	424	1345
13	222	386	838	155	208
14	280	**363**	845	211	317
15	386	394	1130	271	490
16	822	466	1765	524	2133
17	523	443	1195	329	974
18	1010	521	2180	553	2781
19	345	410	1017	225	419

Poněvadž jsme změnili pořadí proměnných $\mathbf{x}_1$ a $\mathbf{x}_2$, označme $l = 2$. V postupu budeme pokračovat krokem 4, tzn. ilr transformací dat. Pro první čtyři pozorování v datové tabulce vypočítáme hodnoty ilr-transformace i s dosazením,

$$z_{11} = \sqrt{\frac{4}{5}} \ln \frac{\sqrt[4]{328 \cdot 640 \cdot 147 \cdot 169}}{196} = \mathbf{0.2823},$$

$$z_{12} = \sqrt{\frac{3}{4}} \ln \frac{\sqrt[3]{640 \cdot 147 \cdot 169}}{328} = -\mathbf{0.2301},$$

$$z_{13} = \sqrt{\frac{2}{3}} \ln \frac{\sqrt{147 \cdot 169}}{640} = -\mathbf{1.1442},$$

$$z_{14} = \sqrt{\frac{1}{2}} \ln \frac{169}{147} = \mathbf{0.0986}.$$

Všimněme si, že v porovnání s hodnotami v tabulce na straně 42, se první dvě nové proměnné liší z důvodu jejich výměny. Hodnoty zbylých dvou složek jsou totožné s hodnotami z předcházející iterace.

K množinám $m_2 = \{7\}$ a $o_2 = \{1, \ldots, 6, 8, \ldots, 19\}$ vytvořme příslušné vektory

a matice bilancí,

$$\mathbf{z}_2^{m_2} = 0.3219, \qquad \mathbf{Z}_2^{m_2} = \begin{pmatrix} 1 & -0.5466 & -1.3427 & -0.2290 \end{pmatrix},$$

$$\mathbf{z}_2^{o_2} = \begin{pmatrix} 0.2823 \\ 0.0885 \\ 0.4043 \\ 0.2543 \\ 0.3254 \\ 0.1768 \\ 0.1200 \\ 0.1990 \\ 0.2261 \\ 0.0827 \\ 0.1459 \\ 0.3258 \\ 0.2695 \\ 0.2190 \\ 0.1565 \\ 0.1830 \\ 0.1158 \\ 0.2282 \end{pmatrix}, \qquad \mathbf{Z}_2^{o_2} = \begin{pmatrix} 1 & -0.2301 & -1.1442 & 0.0986 \\ 1 & 0.8459 & -0.4124 & 1.0554 \\ 1 & -0.3997 & -1.3260 & -0.0524 \\ 1 & -0.0080 & -1.0722 & 0.2953 \\ 1 & -0.0803 & -1.0311 & 0.1491 \\ 1 & 0.2421 & -0.7225 & 0.5146 \\ 1 & 0.5804 & -0.5916 & 0.7660 \\ 1 & 0.5617 & -0.7173 & 0.5454 \\ 1 & 0.0730 & -0.9431 & 0.5045 \\ 1 & 0.8113 & -0.4873 & 1.0434 \\ 1 & 0.6481 & -0.5866 & 0.8163 \\ 1 & -0.2181 & -1.2578 & 0.2080 \\ 1 & 0.0482 & -0.9667 & 0.2878 \\ 1 & 0.2591 & -0.9240 & 0.4188 \\ 1 & 0.8574 & -0.4185 & 0.9926 \\ 1 & 0.4280 & -0.6101 & 0.7675 \\ 1 & 0.9139 & -0.4606 & 1.1421 \\ 1 & 0.0953 & -0.9779 & 0.4397 \end{pmatrix}.$$

Regresní model

$$\mathbf{z}_2^{o_2} = \mathbf{Z}_2^{o_2}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

je po dosazení tvaru

$$\begin{pmatrix} 0.2823 \\ 0.0885 \\ 0.4043 \\ 0.2543 \\ 0.3254 \\ 0.1768 \\ 0.1200 \\ 0.1990 \\ 0.2261 \\ 0.0827 \\ 0.1459 \\ 0.3258 \\ 0.2695 \\ 0.2190 \\ 0.1565 \\ 0.1830 \\ 0.1158 \\ 0.2282 \end{pmatrix} = \begin{pmatrix} 1 & -0.2301 & -1.1442 & 0.0986 \\ 1 & 0.8459 & -0.4124 & 1.0554 \\ 1 & -0.3997 & -1.3260 & -0.0524 \\ 1 & -0.0080 & -1.0722 & 0.2953 \\ 1 & -0.0803 & -1.0311 & 0.1491 \\ 1 & 0.2421 & -0.7225 & 0.5146 \\ 1 & 0.5804 & -0.5916 & 0.7660 \\ 1 & 0.5617 & -0.7173 & 0.5454 \\ 1 & 0.0730 & -0.9431 & 0.5045 \\ 1 & 0.8113 & -0.4873 & 1.0434 \\ 1 & 0.6481 & -0.5866 & 0.8163 \\ 1 & -0.2181 & -1.2578 & 0.2080 \\ 1 & 0.0482 & -0.9667 & 0.2878 \\ 1 & 0.2591 & -0.9240 & 0.4188 \\ 1 & 0.8574 & -0.4185 & 0.9926 \\ 1 & 0.4280 & -0.6101 & 0.7675 \\ 1 & 0.9139 & -0.4606 & 1.1421 \\ 1 & 0.0953 & -0.9779 & 0.4397 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{17} \\ \varepsilon_{18} \\ \varepsilon_{19} \end{pmatrix}.$$

Výpočet odhadu regresního parametru je

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{Z}_2^{o_2 T}\mathbf{Z}_2^{o_2}\right)^{-1}\mathbf{Z}_2^{o_2 T}\mathbf{z}_2^{o_2} = \begin{pmatrix} 0.1597 \\ 0.0105 \\ -0.1465 \\ -0.1275 \end{pmatrix},$$

výpočet odhadu chybějící složky $\mathbf{z}_2^{m_2}$ pomocí odhadnutého regresního parametru je

$$\widehat{\mathbf{z}}_2^{m_2} = \mathbf{Z}_2^{m_2}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1 - 0.5466 - 1.3427 - 0.2290 \end{pmatrix} \begin{pmatrix} 0.1597 \\ 0.0105 \\ -0.1465 \\ -0.1275 \end{pmatrix} = \mathbf{0.3799}.$$

Tuto odhadnutou hodnotu dosadíme do matice ilr-transformovaných dat na pozici původní hodnoty $z_{71} = 0.3219$. Spočítáme inverzní ilr transformaci složek a tím dostaneme hodnoty vypsané tabulkou

i	$\mathbf{x}_1^{ilr}$	$\mathbf{x}_2^{ilr}$	$\mathbf{x}_3^{ilr}$	$\mathbf{x}_4^{ilr}$	$\mathbf{x}_5^{ilr}$
1	0.1324	0.2216	0.4324	0.0993	0.1142
2	0.1520	0.0806	0.2999	0.0858	0.3817
3	0.1113	0.2472	0.4600	0.0941	0.0874
4	0.1376	0.1841	0.4377	0.0955	0.1451
5	0.1305	0.2013	0.4259	0.1084	0.1339
6	0.1557	0.1539	0.3671	0.1053	0.2180
7	0.1124	0.2760	0.4395	0.0998	0.0722
8	0.1581	0.1093	0.3464	0.0977	0.2885
9	0.1478	0.1135	0.3901	0.1102	0.2383
10	0.1438	0.1738	0.4085	0.0901	0.1838
11	0.1532	0.0832	0.3162	0.0832	0.3641
12	0.1520	0.1021	0.3482	0.0953	0.3024
13	0.1227	0.2134	0.4632	0.0857	0.1150
14	0.1389	0.1801	0.4191	0.1047	0.1572
15	0.1445	0.1475	0.4231	0.1015	0.1835
16	0.1440	0.0816	0.3091	0.0918	0.3735
17	0.1510	0.1279	0.3450	0.0950	0.2812
18	0.1434	0.0740	0.3094	0.0785	0.3947
19	0.1428	0.1697	0.4210	0.0931	0.1734

Imputaci chybějící hodnoty x_{75} z původního datového souboru dostaneme po dosazení do vztahu

$$x_{75}^{*} = x_{71}^{ilr} \sum_{j=1}^{5} x_{7j} = 0.1124 \cdot (414 + 260 + 94 + 68 + 113) = \mathbf{106}.$$

Chybějící hodnota se nám povedla naimputovat pomocí iterativního regresního přístupu v podstatě shodně s metodou k-nn, protože obě imputované hodnoty se od původní 112 liší pouze v jednotkách.

Závěrem podotkněme, že iterativní regresní imputací jsme pouze jednu ze tří chybějících hodnot doplnili kvalitnějším způsobem než u metody k-nn, zato v tomto případě došlo k viditelnému zkvalitnění imputované hodnoty. Uvedené chování ovšem mohl způsobit fakt, že rozsah výběru není příliš velký a tudíž jsou metody méně přesné. Nemohou se proto plně projevit jejich vlastnosti.

Původní a imputovaná data můžeme v dalším postupu porovnat pomocí dvou kritérií [8]. Prvním z nich je *kompoziční rozptyl chyb*. Nechť M$\subset \{1, \ldots, n\}$ označuje množinu indexů označující pozorování zahrnující nejméně jednu chybějící složku, a $n_M = |\mathrm{M}|$ označuje počet těchto pozorování (kompozic v datovém souboru s alespoň jednou chybějící hodnotou). Potom je kompoziční rozptyl chyb definován jako

$$\frac{1}{n_M} \sum_{i \in \mathrm{M}} \mathrm{d}_A^2(\mathbf{x}_i, \hat{\mathbf{x}}_i),$$

kde $\mathbf{x}_i$ označuje původní kompozici (ve smyslu před nastavením některých buněk chybějícími) a $\hat{\mathbf{x}}_i$ představuje kompozici s imputovanými chybějícími složkami.

Druhým kritériem k posouzení kvality imputace je *rozdíl v kovariační struktuře*. Označme $\mathbf{S}= (s_{ij})$ jako výběrovou varianční matici ilr-transformovaných pozorování z_{ij}. Můžeme využít například transformaci podle vztahu (4), použití jiné ilr báze by ovšem toto kritérium stejně nezměnilo. Dále označme $\hat{\mathbf{S}} = (\hat{s}_{ij})$ jako výběrovou variační matici počítanou pro stejná ilr-transformovaná data, kde jsou však všechny chybějící složky datové matice imputovány. Potom toto kritérium je založeno na euklidovské vzdálenosti mezi oběma odhady variančních matic, tj.

$$\frac{1}{D-1} \sqrt{\sum_{i=1}^{D-1} \sum_{j=1}^{D-1} (s_{ij} - \hat{s}_{ij})^2} = \frac{1}{D-1} \left\| \mathbf{S} - \hat{\mathbf{S}} \right\|.$$

Poznamenejme přitom, že kompoziční rozptyl chyb měří blízkost/přesnost imputovaných hodnot v Aitchisonově geometrii, zatímco rozdíl v kovarianční struktuře vyjadřuje vliv imputace na mnohorozměrnou kompoziční datovou strukturu.

Vrátíme-li se k příkladu 2 a 3, můžeme dopočítat kompoziční rozptyl chyb a následně zhodnotit kvalitativní charakter metod. U obou metod je množina

$\mathrm{M} = \{7, 14\}$ a $n_M = 2$. Hodnota kritéria je pro k-nn rovna

$$\begin{aligned}
&\frac{1}{2}\left[d_A^2\left(\mathbf{x}_7, \hat{\mathbf{x}}_7\right) + d_A^2\left(\mathbf{x}_{14}, \hat{\mathbf{x}}_{14}\right)\right] = \\
&= \frac{1}{2}\left\{d_A^2\left[(414, 305, 94, 68, 112)\,; (414, 222, 94, 68, 113)\right] + \right. \\
&\left. + \; d_A^2\left[(845, 386, 211, 317, 280)\,; (845, 373, 211, 317, 280)\right]\right\} = \mathbf{0.0414},
\end{aligned}$$

a pro iterativní regresní imputaci

$$\begin{aligned}
&\frac{1}{2}\left[d_A^2\left(\mathbf{x}_7, \hat{\mathbf{x}}_7\right) + d_A^2\left(\mathbf{x}_{14}, \hat{\mathbf{x}}_{14}\right)\right] = \\
&= \frac{1}{2}\left\{d_A^2\left[(414, 305, 94, 68, 112)\,; (414, 260, 94, 68, 106)\right] + \right. \\
&\left. + \; d_A^2\left[(845, 386, 211, 317, 280)\,; (845, 363, 211, 317, 280)\right]\right\} = \mathbf{0.0112}.
\end{aligned}$$

Na základě těchto výsledků můžeme tvrdit, že proces imputace pomocí iterativního přístupu je kvalitnější, protože kompoziční rozptyl chyb je menší než kompoziční rozptyl chyb u algoritmu k nejbližších sousedů.

4 Imputace pro kompoziční data v R

Software R je volně dostupný statistický program pro výpočty různých statistických ukazatelů a charakteristik. Také se běžně používá ke grafickému zobrazení dat ve statistice. Tento program je volně dostupný na **http://cran.r-project.org**. Uživatel si program může "vylepšit" doinstalováním knihoven podle jeho potřeby. Tyto knihovny obsahují argumenty inputů i outputů a vysvětlení jednotlivých funkcí i s příklady.

Pro kompoziční data existují v rámci R dvě speciální knihovny `compositions` a `robCompositions`. Druhý z nich obsahuje nástroje k výpočtům v rámci statistické analýzy kompozic a jejich vizualizaci pomocí grafických zobrazení. Mimoto uživateli poskytuje dvě metody pro imputaci chybějících hodnot. Jedna je založena na metodě k-nn, druhá na iterativním modelu odhadu aplikovaném na ilr-transformované kompozice [17].

Knihovny lze jednoduchým způsobem ke stávajícím doinstalovat v menu programu R pod záložkou Packages volbou položky seznamu Install package(s). Nejprve je nutné zvolit zemi a poté již stačí vybrat danou knihovnu (v našem případě `compositions` a `robCompositions`). Instalace je dokončena zobrazením stavového hlášení do konzoly programu. Pokud chceme knihovny dále používat k výpočtům, musíme je načíst. Toho dosáhneme, když v menu programu zvolíme Packages a dále Load packages. V konzole se ukáže taktéž protokol o načtení knihovny. Po tomto úkonu můžeme již zadávat funkce v nich nadefinované. Konkrétně úplný přehled funkcí knihovny `robCompositions` můžeme dostat, pokud do programu R zadáme příkaz

```
> help(package='robCompositions').
```

V rámci této knihovny jsou uživateli k dispozici datové soubory, které je možné použít pro další výpočty statistických metod. Jejich přehled nalezneme v R po zadání příkazů

```
> require (robCompostions),
> data (package =‘ robCompositions'),
```

kde první příkaz znamená načtení knihovny, druhý představuje samotný seznam datových souborů. Zadáme-li příkaz

```
> help(expenditures)
```

zobrazí se soubor s popisem datového souboru Household expenditures, který jsme použili pro demonstraci použití imputačních metod v předchozí kapitole.

Knihovna `robCompositions` současně závisí na pěti jiných knihovnách, speciálně na `utils, robustbase, rrcov, car a MASS`, ze kterých je importováno několik funkcí.

V příslušné knihovně existují tři možnosti, jak vyjádřit kompozice pomocí souřadnic: alr transformace, clr transformace a ilr transformace. Pro ilr transformaci dat stačí zadat do příkazové řádky

```
> z=ilr(expenditures).
```

Samozřejmě je implementována i inverzní transformace. Provede se příkazem

```
> invilr(z).
```

Oba příkazy pro výpočet transformací ve svých parametrech připouštějí nastavení specifických parametrů.

4.1 Imputace chybějících hodnot

Ve třetí kapitole byly podrobně popsány dvě metody pro imputaci kompozic, na kterou nyní navážeme aplikací výpočtu v softwaru R. První metoda využívá principů k nejbližších sousedů. Druhá metoda se označuje jako iterativní regresní imputace. Ve svém výpočtu využívá vypočítané hodnoty z algoritmu k-nn a dále postupuje podle výše vysvětleného postupu. V této podkapitole na příkladu ukážeme výpočet iterativní regresní imputace v programu R. K tomuto účelu jsme využili nadefinovaných funkcí knihovny `robCompositions`. Celá procedura přitom, pokračuje dokud se imputované hodnoty neustálí nebo dokud není dosaženo maximálního počtu iterací.

Ilustrativní imputaci pomocí softwaru R jsme provedli pro data z příkladu 2 uvedeném ve třetí kapitole. Jedná se původně o soubor z knihovny `robCompositions` s názvem Household expenditures. Protože jsme z datového souboru odstranili třetí řádek, načteme do programu R již takto upravený datový soubor s názvem "expPriklad2.txt" pomocí příkazu `read.table`. Nově vzniklou tabulku hodnot jsme pojmenovali jednoduše "vydaje". Protože jsme chtěli imputovat chybějící hodnoty, museli jsme je v tomto souboru vytvořit. To jsme provedli pomocí příkazů v R k tomu určených. Poté jsme již naimputovali hodnoty pomocí příkazů pro iterativní regresní imputaci. Nyní následuje sled příkazů definovaných pro iterativní regresní imputaci.

```
> vydaje=read.table("expPriklad2.txt")
> vydaje[7,2]<- vydaje[7,5]<- vydaje[14,2]<- NA

# imputace pomocí iterativní regrese bez transformace dat
> imp2=impCoda(vydaje, method="lm", closed=TRUE, k=5)
> imp2
 ---------------------------------------
[1] "3 missing vales were imputed"
[1] "1 iteration was needed"
[1] "the last change was 0.1078"
 ---------------------------------------
> imp2$xImp[7,2]   # zobrazení imputované hodnoty na dané pozici
[1] 246.4926
> imp2$xImp[7,5]
[1] 89.52897
> imp2$xImp[14,2]
[1] 351.3506
```

Je vidět, že rozdíly mezi imputovanými hodnotami pomocí softwaru R a původními jsou značné. K tomuto jevu zřejmě došlo právě v důsledku malého rozsahu

datového souboru. Navíc, typické vlastnosti metod se proto nemohou plně projevit; také poznamenejme, že při konvergenci iteračního algoritmu může (například vlivem odlehlých hodnot či nekvalitně stanovené startovací imputace metodou k nejbližších sousedů) docházet k nabalování chyb.

5 Praktický příklad

Tato kapitola pojednává o imputaci chybějících hodnot pro reálná data, jež jsme získali od Českého statistického úřadu; z důvodu jejich rozsahu jsou umístěna na přiloženém CD. Nutno dodat, že data pocházejí z fiktivní populace. Soubor představuje náklady na bydlení a obsahuje vybrané údaje za domácnosti, které bydlí v bytových domech a platí za příslušné složky nákladů na bydlení v Kč. Rozsah výběru o 3970 pozorováních není ještě dost velký, nicméně musel být brán ohled na určitou homogenitu vzorku. Optimálně by měl datový soubor čítat alespoň přibližně deset tisíc pozorování.

Soubor na přiloženém CD je zpracován v MS Excelu. Kromě datových tabulek obsahuje první list souboru stručný popis dat. Na druhém listu může čtenář nalézt kompletní datový soubor se všemi hodnotami plně napozorovanými a datové tabulky modifikované v závislosti na procentu položkové non-response, tj. kdy data považujeme za nezodpovězená. Nejprve uvažujeme případ, kdy u každé složky je nasimulováno 10 % non-response, druhou možností je, že v každé položce je přítomno právě 20 % chybějících hodnot a poslední je situace se složkovou non-response u jednotlivých složek od 5% do 25%. Pro každou variantu procentuálního zastoupení chybějících hodnot jsou určeny dva listy. První z nich vždy obsahuje imputované chybějící hodnoty v kompletních datových tabulkách, tzn. obsahuje kompletní tabulky s imputovanými hodnotami pro jednotlivé metody. Druhý list čtenáři poskytuje porovnání pouze imputovaných hodnot pro jednotlivé metody rozdělených podle složek kompozice. Především nás zajímalo, jak bude imputace úspěšná v závislosti na zvoleném procentu nezodpovězených položek datového souboru a hlavně na zvolené metodě.

Jak již bylo poznamenáno, jedná se o náklady na bydlení celkem uváděné v Kč. Protože nás zajímají podíly mezi složkami kompozice, nemusíme v dalším údaje o celkových nákladech uvažovat. V datovém souboru jsou složky reprezentovány označením

najem	nájemné, úhrada za užívání bytu,
elektr	elektřina,
plyn	plyn,
vodne	vodné a stočné,
ost_sluz	ostatní služby spojené s užíváním bytu.

Pro představu čtenáři na tomto místě poskytneme datové tabulky, se kterými jsme v Excelu pracovali. Nejprve uvádíme tabulku s kompletně napozorovanými hodnotami

najem	elektr	plyn	vodne	ost_sluz
800	600	200	250	298
3800	900	570	140	627
2000	600	1300	200	457
549	748	807	150	72
4002	850	1750	1316	782
1998	2720	2931	723	328
3705	1200	1500	1452	843
6111	1500	1500	1500	1389
4872	1000	1500	1641	487
1400	1400	1200	1000	242
⋮	⋮	⋮	⋮	⋮

a tabulku s 20% položkovou non-response

najem	elektr	plyn	vodne	ost_sluz
800	600	200	250	298
0	900	570	140	627
0	600	1300	200	0
0	748	807	0	72
4002	850	1750	1316	0
0	2720	0	723	0
0	1200	0	1452	843
6111	1500	1500	1500	1389
4872	1000	0	1641	487
1400	1400	0	1000	242
⋮	⋮	⋮	⋮	⋮

Na každém listu popisujícím srovnání imputace chybějících hodnot (vždy druhý list pro každou variantou non-response) jsme provedli porovnání celkového

počtu chybějících hodnot a počtu imputovaných hodnot, které byly lépe nahrazeny pomocí iterativní regrese pro jednotlivé složky nákladů na bydlení. Výpočty byly provedeny pomocí příkazů softwaru R popsaných v kapitole 4.1. Jejich znění je taktéž zkopírováno na přiloženém CD ve formě skriptu softwaru R. Pro přehlednost vše shrnuje následující tabulka.

data s % nonresponse	složky nákladů	celkový počet non-response	imputace k-nn	it. regresní imputace
data_10	najem	398	176	222
	elektr	385	176	209
	plyn	397	232	165
	vodne	405	190	215
	ost_sluzby	378	176	202
data_20	najem	767	526	241
	elektr	797	303	494
	plyn	792	390	402
	vodne	777	348	429
	ost_sluzby	797	305	492
data_5_25	najem	191	83	108
	elektr	374	154	220
	plyn	594	312	282
	vodne	985	457	528
	ost_sluzby	1199	538	661

Z tabulky je patrné, že iterativní regresní imputace byla pouze ve třech případech neúspěšná, když hovoříme o lepším nahrazení chybějících hodnot oproti metodě k-nn. Výhodou zde bylo, že jsme měli původní data k dispozici a tudíž jsme mohli srovnání provést. Na základě výsledků lze říci, že se potvrdila naše domněnka, že iterativní regresní imputace provádí nahrazení chybějících hodnot kvalitněji, i když zřejmě většinou nedochází k zásadnímu vylepšení imputace oproti metodě k nejlepších sousedů.

Závěr

V této diplomové jsem nejprve shrnula teoretické poznatky týkající se kompozičních dat. Ty jsou důležité pro pochopení speciálních vlstností kompozic a také pro další práci s kompozičními daty. Po představení mechanismů a stručném seznámení s regresní imputací jsem popsala detailně dvě nejužívanější metody imputace chybějících hodnot u kompozičních dat. Obě dvě metody jsou vysvětleny na ilustrativním příkladu. Ten může čtenáři poskytnout návod, jak pomocí těchto metody chybějící hodnoty imputovat, protože obsahuje jednotlivé mezivýpočty i s vysvětlením. Protože jsem chtěla v práci ukázat použití obou metod na reálných datech, popsané metody jsem aplikovala na data získaná z Českého statistického úřadu. Jedná se o datový soubor popisující náklady na bydlení v Kč u fiktivní populace o rozsahu 3970 pozorovávní. Na základě provedených výpočtů lze říci, že iterativní regresní imputace poskytuje kvalitnější imputaci chybějících hodnot než metoda k nejbližších sousedů.

Ke zpracování dat jsem používala statistický software R a jeho knihovny `compositions` a `robCompositions`, určené pro práci s kompozicemi. Po prvních neúspěších jsem si software oblíbila a mohu říci, že je pro specifickou práci s kompozičními daty velmi dobře vybavený, zejména pokud uživatel potřebuje spočítat specifické operace v oboru kompozic.

Závěrem bych chtěla podotknout, že tato práce mi poskytla možnost se dozvědět nové poznatky o oblastech statistiky, které nejsou běžnou náplní studijních plánů. Zároveň jsem měla příležitost se zdokonalit v práci se softwarem R a konečně i Microsoft Excel, ve kterém mi byly poskytnuty datové soubory od ČSÚ.

Literatura

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

[2] Buhi, E. R., Goodson, P., Neilands, T. B., *Out of sight, not out of mind: strategies for handling missing data* [online], dostupné z: http://findarticles.com/p/articles/mi_7414/is_1_32/ai_n32056906/, [citováno 15.3.2011].

[3] Egozcue, J. J., *Reply to "On the Harker Variation Diagrams; . . ." by J.A Cortés*. Mathematical Geosciences, **41** (7), 829-834 (2009).

[4] Fišerová E., Hron, K., *On interpretation of orthonormal coordinates for compositional data*. Mathematical Geosciences, DOI:10.1007/s11004-011-9333-x.

[5] Howell, D. C., *Treatment of Missing Data* [online], dostupné z: http://www.uvm.edu/ dhowell/StatPages/More_Stuff/Missing_Data/Missing.html, [citováno 15.3.2011].

[6] Hron, K., *Elementy statistické analýzy kompozičních dat* [online], dostupné z: http://www.statspol.cz/bulletiny/ib-2010-3.pdf, [citováno 28.10.2010].

[7] Hron, K., Kunderová, P., *Regresní analýza* (učební text). Interní materiál KMAAM, PřF UP Olomouc, 2010.

[8] Hron, K., Templ, M., Filzmoser, P., *Imputation of missing values for compositional data using classical a robust methods*. Computational Statistics and Data Analysis, **54** (12), 3095-3107 (2010).

[9] *K-nearest neighbor* [online], dostupné z: http://www.scholarpedia.org/article/K-nearest_neighbor, [citováno 13.12.2011].

[10] *K nearest neighbors tutorial* [online], dostupné z: http://people.revoledu.com/kardi/tutorial/KNN/index.html, [citováno 5.12.2010].

[11] Little, R., J.A., Rubin, D.B., *Statistical analysis with missing data*, 2nd edition. John Wiley & Sons, New York, 2002.

[12] Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V., *Dealing with zeros and missing values in compositional data sets using nonparametric imputation*. Mathematical Geology, **35** (3), 253-278 (2003).

[13] *Missing-data imputation* [online], dostupné z: http://www.stat.columbia.edu/g̃elman/arm/missing.pdf, [citováno 10.12.2011].

[14] *Missing value mechanism* [online], dostupné z: http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=74:missing-value-mechanism&catid=40:missingness-mechanisms&Itemid=96, [citováno dne 10.12.2010].

[15] Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, J., *Lecture notes on compositional data analysis*, The University of Girona, 2007 [online], dostupné z: http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf, [citováno 28. 10. 2010].

[16] Petera, Martin, *Korelační analýza pro kompoziční data*, Olomouc: PřF UP Olomouc, 2010, 37 s.

[17] Templ, M., Hron, K., Filzmoser, P., *robComposition : An R-package for robust statistical analysis of compositional data*. useR! 2010, Gaithersburg, Maryland, USA.

[18] Templ, M., Filzmoser, P., Hron, K.,: *Imputation of item non-responses in compositional data using robust methods*. UNECE 2009, Neuchatel, Switzerland.

[19] Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays*. Bioinformatics, **17** (6), 520-525 (2001).