# STŘEDOŠKOLSKÁ ODBORNÁ ČINNOST

*Obor 01 - Matematika a statistika*

## Korelační analýza a kompoziční data


| | |
|---|---|
| Autoři: | Klára Švarcová |
| | Milan Barančík |
| Škola: | Slovanské gymnázium Olomouc, |
| | tř. Jiřího z Poděbrad 13, |
| | Olomouc, 771 10 |
| Studijní obor: | všeobecný se zaměřením na přírodní vědy |
| | 3. ročník |
| Konzultant: | RNDr. Karel Hron, Ph.D. |
| | Přírodovědecká fakulta UP Olomouc, |
| | Katedra matematické analýzy a aplikací matematiky |


Olomouc, 2010

**Čestné prohlášení:**
Prohlašujeme tímto, že jsme soutěžní práci vypracovali samostatně pod vedením pana RNDr. Karla Hrona, Ph.D. a použili jsme pouze podklady (literaturu, SW atd.) uvedené v přiloženém seznamu. Nemáme závažný důvod proti zpřístupňování této práce v souladu se zákonem č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) v platném znění.

V Olomouci dne Podpisy:

**Poděkování:**
Chtěli bychom především poděkovat panu profesoru Hronovi, který jakožto hlavní vedoucí naší práce s námi strávil mnoho času vysvětlováním a odtajňováním krás matematiky. Rozšířil naše znalosti a ukázal nám, že tato věda v žádném případě není nudná a nezajímavá a že má mnoho aplikací v reálném životě.

# Obsah

# Úvod

Možná by někoho mohla napadnout otázka, proč jsme si vybrali právě statistiku, vědu, o které se nechvalně říká, že není moc zajímavá, zřejmě pro svoji složitost. V našem případě to bude především proto, že nás oba baví matematika, a statistika je vlastně matematika, která může popisovat a zkoumat téměř jakýkoliv problém v reálném životě. A právě tato její pestrost nás velmi zaujala, společně s možností zabývat se reálnými daty a z nich následně vyvodit příslušné závěry, o což se pokusíme i v naší práci.

Statistika má mnoho zajímavých "odvětví"a právě jedno z nich jsme si vybrali. Teorie kompozičních dat jako teorie pozorování nesoucích pouze relativní informaci je poměrně mladou oblastí statistiky, alespoň co se z hlediska vývoje matematiky týče, a tudíž ještě není zdaleka tak prozkoumána. Proto se s ní musíme seznámit a proniknout do její problematiky natolik hluboko, abychom byli schopni s tímto druhem dat pracovat a následně interpretovat naše výsledky, které budou, doufáme i pro čtenáře, zajímavé.

Hlavním cílem naší práce bude totiž zkoumat vztahy mezi proměnnými (statistickými znaky). V případě běžných pozorování nabízí vyhovující řešení korelační analýza, jak si shrneme v první kapitole. Pro data nesoucí relativní informaci bude ovšem potřeba hledat jinou alternativu, čemuž se podrobně věnují následující dvě kapitoly. Na závěr práce použijeme získané vědomosti k analýze ekonomicky orientovaného problému, získaného z databáze Eurostatu.

# 1 Korelační analýza

Korelační analýza je metoda popisné statistiky, která umožňuje zkoumat vztahy mezi statistickými znaky [2]. Jako typickou situaci v tomto smyslu můžeme uvažovat vztah mezi výškou studentů třetího ročníku gymnázia v cm (statistický znak $X$) a jejich váhou v kg (statistický znak $Y$). Provedením $n$ měření získáme pro znak $X$ hodnoty $x_1, \ldots, x_n$, resp. $y_1, \ldots, y_n$ pro znak $Y$.

Pro výpočet korelačního koeficientu nejprve vypočítáme aritmetické průměry a rozptyly,

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) \ , \ s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$\bar{y} = \frac{1}{n}(y_1 + \ldots + y_n) \ , \ s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

pro jednotlivé statistické znaky.

Odmocněním rozptylů získáme směrodatné odchylky pro znak $X$ , resp. $Y$,

$$s_X = \sqrt{s_X^2} \ , \ s_Y = \sqrt{s_Y^2}.$$

Před samotným výpočtem korelačního koeficientu je potřeba určit kovarianci jako charakteristiku závislosti mezi statistickými znaky,

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Takto se dostáváme k následující definici. *Korelačním koeficientem* rozumíme výraz

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Shrňme si základní vlastnosti korelačního koeficientu. Vyjadřuje v nejjednodušším smyslu vzájemný lineární vztah mezi statistickými znaky $X$ a $Y$. Tento vztah může být kladný, pokud (přibližně) platí $Y = kX + q$, nebo záporný ($Y = -kX + q$), kde $k > 0$ a $q$ je libovolné reálné číslo. Korelační koeficient

je definován jen v případě, jestliže obě směrodatné odchylky jsou konečné a nenulové.

Konkrétněji, korelační koeficient určuje míru lineární závislosti statistických znaků $X$ a $Y$. Je to přitom bezrozměrná charakteristika; nabývá hodnot od $-1$ až po $+1$ ($-1 \leq r_{XY} \leq 1$) a pracuje s oběma znaky symetricky, tj. $r_{XY} = r_{YX}$. Hodnota korelačního koeficientu $-1$ značí nepřímou úměrnost, tedy čím jsou větší hodnoty prvního znaku, tím více se naopak zmenší hodnoty druhého znaku. Všechny naměřené dvojice hodnot tak přitom leží na výše uvedené přímce $Y = -kX + q$. Hodnota korelačního koeficientu $+1$ analogicky značí přímou úměrnost. Čím je hodnota korelačního koeficientu blíže 1 anebo $-1$, tím je lineární závislost mezi znaky těsnější a body $(x_i, y_i)$, $i = 1, \ldots, n$, budou mít tendenci ležet na dané přímce. Jeho kladná (záporná) hodnota tak vlastně odpovídá celkové rostoucí (klesající) lineární závislosti mezi $X$ a $Y$. Pokud je korelační koeficient roven 0, pak mezi znaky není žádná statisticky zjistitelná lineární závislost. I při nulovém korelačním koeficientu na sobě ovšem znaky mohou záviset, pouze tento vztah nelze vyjádřit lineární funkcí, a to ani přibližně.

Máme ale situace, resp. data, pro které se koncept korelačního koeficientu ukazuje jako zcela nevhodný. Jako ukázku si uvedeme následující příklad [1].

**Příklad:** Mějme dva geology, A a B, kteří zkoumají proporce a vztahy mezi jednotlivými složkami v půdních vzorcích. Byly odebrány tři vzorky půdy, obsahující rostlinnou, živočišnou, jílovitou (neživou) a vodní složku, a zkoumány nezávisle oběma geology. Takto tedy vědec A získal vzorky se čtyřmi složkami, zatímco vědec B nejprve vzorky vysušil a obdržel tak pouze první tři složky. Za předpokladu absence chyb měření dostaneme následující tabulku. Data přitom pro snadnou interpretovatelnost vyjádřili oba geologové již v proporcích, tedy jako pozorování se součtem složek rovným jedné.

| | geolog A | | | | geolog B | | |
|---|---|---|---|---|---|---|---|
| vzorek | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x'_1$ | $x'_2$ | $x'_3$ |
| 1 | 0.1 | 0.2 | 0.1 | 0.6 | 0.25 | 0.50 | 0.25 |
| 2 | 0.2 | 0.1 | 0.2 | 0.5 | 0.40 | 0.20 | 0.40 |
| 3 | 0.3 | 0.3 | 0.1 | 0.3 | 0.43 | 0.43 | 0.14 |

Při spočítání příslušných korelačních koeficientů mezi jednotlivými statistickými znaky (půdními složkami) obdržíme následující matice.

| Cor A | $x_1$ | $x_2$ | $x_3$ | $x_4$ | Cor B | $x'_1$ | $x'_2$ | $x'_3$ |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1.00 | **0.50** | 0.00 | **-0.98** | $x'_1$ | 1.00 | **-0.57** | **-0.05** |
| $x_2$ | | 1.00 | **-0.87** | **-0.65** | $x'_2$ | | 1.00 | **-0.79** |
| $x_3$ | | | 1.00 | 0.19 | $x'_3$ | | | 1.00 |
| $x_4$ | | | | 1.00 | | | | |

Hodnoty pod hlavní diagonálou v maticích jsou vynechámy z důvodu symetrie korelačního koeficientu.

Při srovnání tabulek korelačních koeficientů zjistíme několik závažných výsledků. První z nich se týká vztahu mezi složkami $x_1$, $x_2$, resp. $x'_1$ a $x'_2$ (rostlinná a živočišná sožka), kde geolog A dospěl k hodnotě $0,5$, kdežto geolog B k hodnotě $-0,57$. Z toho by první usuzoval na kladný lineární vztah mezi statistickými znaky, ale druhý na záporný lineární vztah. Takto by tedy oba vědci dospěli k protichůdným závěrům!

Dalším nepříznivým jevem je výskyt záporných korelací, který souvisí s kovarianční strukturou dat s konstantním součtem složek.

K výše uvedeným rozporuplným výsledkům jsme dospěli tím, že data týkající se prvních tří složek u geologa A jsme u geologa B přeškálovali. Tím se samozřejmě nezmění podíly mezi jednotlivými složkami, změní se ovšem absolutní hodnoty u jednotlivých statistických znaků, které vedou k různým hodnotám korelačních koeficientů. Přitom výsledky měření jak u geologa A, tak u geologa B jsou smysluplné: zajímají nás totiž nikoliv absolutní hodnoty, ale pouze podíly mezi jednotlivými složkami, jinak řečeno, součet složek je v tomto případě irelevantní.

Uvedené vlastnosti jsou charakteristické pro data, která nazýváme *kompoziční* [1].

Takto tedy můžeme vyslovit definici, že *D-složkovou kompozicí* nazveme vektor $\mathbf{x} = (x_1, ..., x_D)$ s kladnými složkami, kde jediná relevantní (smysluplná) informace je obsažena v podílech mezi složkami těchto dat.

Jinak řečeno, jedná se vlastně o mnohorozměrná data, jejichž složky kvantitativně vyjadřují podíly daných částí na celku. Pokud bychom se na zavedená data podívali ještě z jiného pohledu, platí, že pro libovolné kladné reálné číslo $a$ obsahují vektory $(x_1, ..., x_D)$ a $(ax_1, ..., ax_D)$ tutéž informaci.

Z toho je zřejmé, že můžeme libovolně měnit součet složek kompozice (vektoru) tak, aby byla tato snadno interpretovatelná. V našem případě předepíšeme daný součet jako $\kappa$, kdy $\kappa$ nabývá nejčastěji hodnot 1 nebo 100, a potom budou složky kompozice reprezentovat proporce, resp. procentuální podíly.

Kompoziční data se ale samozřejmě nevyskytují jenom v geologii. Budeme si to ilustrovat následujícím příkladem.

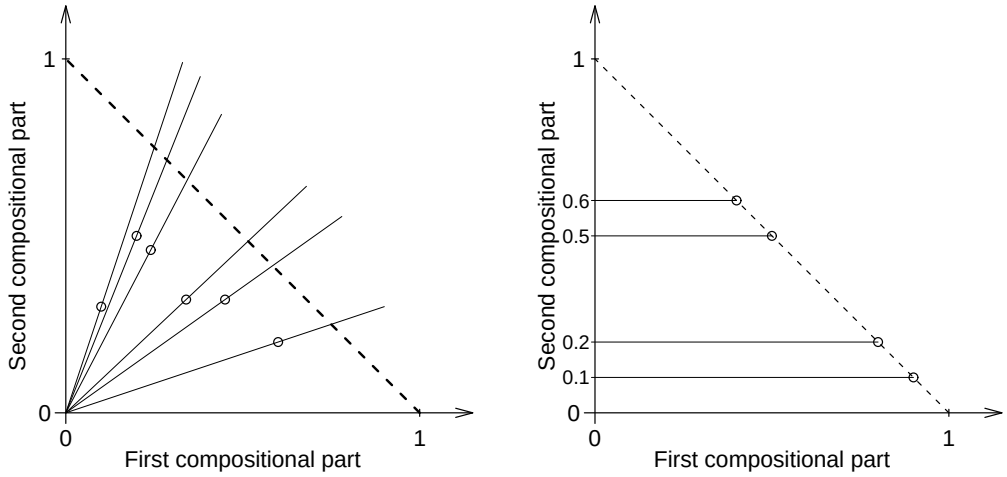
**Příklad:** Předpokládejme, že výdaje v domácnosti můžeme rozdělit na tři části. Na výdaje související s bydlením, na základní výdaje jako jídlo a oblečení a na ostatní výdaje (doprava, vzdělání, kultura apod.). Takto bychom pro konkrétní domácnost obdrželi příslušnou trojsložkovou kompozici, např. $\mathbf{x} = (10000, 8000, 7000)$.

Spíše než absolutní hodnoty nás ale zajímá procentuální (proporcionální) rozložení výdajů na jednotlivé položky, vlastně podíly jednotlivých složek na celkových výdajích. Tedy danou kompozici můžeme vyjádřit např. ve formě proporcí,

$$\mathbf{x}' = \left(\tfrac{10000}{25000}, \tfrac{8000}{25000}, \tfrac{7000}{25000}\right) = (0.4, 0.32, 0.28);$$

je přitom zřejmé, že $\mathbf{x}$ i $\mathbf{x}'$ pro nás obsahují tutéž informaci.

Na konci této kapitoly si ještě v krátkosti řekneme, jak uvedené vlastnosti kompozičních dat interpretovat geometricky. Na obrázku 1, který je převzatý z článku [6], jsou znázorněna dvousložková kompoziční data. Na levém grafu si můžeme všimnout, že podíly složek všech kompozic, které leží na dané




Obrázek 1: Geometrie kompozičních dat.

polopřímce vycházející z počátku, jsou stejné, tím pádem obsahují totožnou informaci. Průsečíky polopřímek s čárkovanou úsečkou vyjadřují takové reprezentace těchto kompozic, které mají součet složek rovné jedné (složky tedy vyjadřují proporcionální podíly na celku).

Na pravém grafu je znázorněna další charakteristická vlastnost kompozičních dat, totiž že jejich škála je relativní. Na tomto grafu se jeví vzdálenost mezi oběma dvojicemi kompozic jako stejná. Ovšem v prvním případě se proporce druhé složky (např. nějakého chemického prvku) zvýšila z $0,1$ na $0,2$, tedy na dvojnásobek, v druhém případě z $0,5$ na $0,6$, tedy pouze o jednu pětinu. Je zřejmé, že naše intuitivní vzdálenost první dvojice kompozice je tak mnohem větší než druhé dvojice. Toto by měla odpovídající geometrie kompozic respektovat. Bohužel z tohoto pohledu se standardní euklidovská geometrie nejeví pro kompoziční data jako vhodná.

Protože euklidovská geometrie je též základem většiny statistických metod (včetně rozptylu či korelačního koeficientu), ukázali jsme na příkladu s geologickými daty, že její použití u kompozičních dat není smysluplné ani v případě

jejich statistického zpracování a může vést ke zcela zavádějícím výsledkům.

# 2 Geometrie kompozičních dat

Jak jsme si uvedli výše, není euklidovská geometrie pro kompoziční data vhodná, na druhé straně pouze v ní může probíhat statistické zpracování dat. V této kapitole si tedy zavedeme geometrii, která plně odpovídá charakteru kompozic a zároveň si ukážeme způsob, jak z ní kompoziční data vyjádřit jako standardní pozorování, která se řídí euklidovskou geometrií (tzn. data nesoucí absolutní informaci).

Na úvod si shrňme základní vlastnosti standardní vektorové algebry. Pro vektory $\mathbf{x} = (x_1, ..., x_D)$, $\mathbf{y} = (y_1, ..., y_D)$ a číslo $c \in R$ definujeme součet vektorů jako

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

a násobek vektoru číslem jako

$$c\mathbf{x} = (cx_1, \ldots, cx_D).$$

Navíc zavádíme též vzdálenost dvou vektorů,

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_D - y_D)^2},$$

velikost vektoru,

$$\| \mathbf{x} \| = \sqrt{x_1^2 + \cdots + x_D^2},$$

a skalární součin,

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + \cdots + x_n y_n.$$

Z tohoto pohledu platí, že

$$\| \mathbf{x} \| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle},\ d(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \| = \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle .$$

Geometrie kompozičních dat, která se v současné době nejčastěji nazývá Aitchisonova, bude mít stejnou strukturu jako euklidovská geometrie, nyní ale již s přihlédnutím k přirozenosti těchto pozorování [4,5]. Aby byl zápis co možná nejobecnější, zavedeme si operaci uzávěr, která danou kompozici $\mathbf{x} = (x_1, \ldots, x_D)$ upraví na předepsaný součet složek $\kappa$, například 100 v případě procentuálních podílů,

$$\mathcal{C}(\mathbf{x}) = \left( \frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right).$$

Množinou všech $D$-složkových kompozic rozumíme tzv. simplex, tedy množinu reprezentací kompozic při daném součtu $\kappa$,

$$S^D = \left\{ (x_1, \ldots, x_D) \ , \ x_i > 0 \ , \ \sum_{i=1}^{D} = \kappa \right\}.$$

Nejprve zavedeme operace, které budou odpovídat sčítání vektorů a násobení vektoru reálným číslem, operace *perturbace* a *mocninná transformace*.

Mějme kompozice $\mathbf{x} = (x_1, ..., x_D), \mathbf{y} = (y_1, ..., y_D)$ a $c \in R$. Potom perturbací dvou kompozic $\mathbf{x}$ a $\mathbf{y}$ rozumíme kompozici

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

mocninnou transformaci kompozice $\mathbf{x}$ reálným číslem $c$ potom definujeme jako kompozici

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

Než si zavedeme pojmy vzdálenost dvou kompozic, velikost kompozice a skalární součin dvou kompozic, zastavme se na chvíli a ukažme si použití standardní euklidovské vzdálenosti pro kompoziční data. Obecně by totiž mělo platit, že vzdálenost dvou vektorů je vždy větší nebo rovna vzdálenosti podvektorů, speciálně pak složek těchto vektorů. Například pro dva (euklidovské) vektory $\mathbf{x} = (x_1, x_2) = (1, 1)$ a $\mathbf{y} = (y_1, y_2) = (2, 3)$ obdržíme

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{1^2 + 2^2} = \sqrt{5} \ , \ d(x_1, y_1) = \mid 2 - 1 \mid = 1,$$

z čehož plyne, že $d(\mathbf{x}, \mathbf{y}) \geq d(x_1, y_1)$.

Mějme ale nyní $\mathbf{x} = (0,65; 0,3; 0,05)$, $\mathbf{y} = (0.2, 0.7, 0.1)$, kde $\mathbf{x}$ a $\mathbf{y}$ jsou dvě kompozice; jejich euklidovská vzdálenost je zřejmě $d(\mathbf{x}, \mathbf{y}) = 0.604$. Uvažujme následně kompozice, které se budou skládat pouze z prvních dvou složek kompozic $\mathbf{x}$ a $\mathbf{y}$. Protože nám na součtu těchto složek nezáleží, můžeme je vyjádřit jako $\mathbf{x}' = (0.684, 0.316)$, $\mathbf{y}' = (0.222, 0.778)$ se vzdáleností $d(\mathbf{x}', \mathbf{y}') = 0.653 \geq 0,604 = d(\mathbf{x}, \mathbf{y})$, což je zřejmě ve sporu s výše uvedenou vlastností, kterou by měla každá vzdálenost splňovat.

Této nežádoucí situaci se vyhneme, pokud zavedeme pro kompozice $\mathbf{x}$ a $\mathbf{y}$ *Aitchisonovu vzdálenost* definovanou jako

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

Když dosadíme do tohoto vzorce výše uvedené kompozice $\mathbf{x}$ a $\mathbf{y}$, resp. $\mathbf{x}'$ a $\mathbf{y}'$, dostaneme

$$d(\mathbf{x}, \mathbf{y}) = \left( \frac{1}{6} \left[ \left( \ln \frac{x_1}{x_1} - \ln \frac{y_1}{y_1} \right)^2 + \left( \ln \frac{x_1}{x_2} - \ln \frac{y_1}{y_2} \right)^2 + \left( \ln \frac{x_1}{x_3} - \ln \frac{y_1}{y_3} \right)^2 + \right.\right.$$

$$+ \left( \ln \frac{x_2}{x_1} - \ln \frac{y_2}{y_1} \right)^2 + \left( \ln \frac{x_2}{x_2} - \ln \frac{y_2}{y_2} \right)^2 + \left( \ln \frac{x_2}{x_3} - \ln \frac{y_2}{y_3} \right)^2 + \left( \ln \frac{x_3}{x_1} - \ln \frac{y_3}{y_1} \right)^2 +$$

$$\left.\left. + \left( \ln \frac{x_3}{x_2} - \ln \frac{y_3}{y_2} \right)^2 + \left( \ln \frac{x_3}{x_3} - \ln \frac{y_3}{y_3} \right)^2 \right] \right)^{\frac{1}{2}} = 1.595 \ ,$$

$$d(\mathbf{x}', \mathbf{y}') = \left( \frac{1}{4} \left[ \left( \ln \frac{x_1}{x_1} - \ln \frac{y_1}{y_1} \right)^2 + \left( \ln \frac{x_1}{x_2} - \ln \frac{y_1}{y_2} \right)^2 + \right.\right.$$

$$\left.\left. + \left( \ln \frac{x_2}{x_1} - \ln \frac{y_2}{y_1} \right)^2 + \left( \ln \frac{x_2}{x_2} - \ln \frac{y_2}{y_2} \right)^2 \right] \right)^{\frac{1}{2}} = 1.432.$$

V tomto případě už tedy obdržíme $d_A(\mathbf{x},\mathbf{y}) = 1.595 \geq d_A(\mathbf{x}',\mathbf{y}') = 1.432$. Obdobně pak definujeme *Aitchisonovu normu* (velikost kompozice) pro kompozici $\mathbf{x} \in S^D$ jako

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

a *Aitchisonův skalární součin* jako

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

Zřejmě platí

$$\| \mathbf{x} \|_A = \sqrt{\langle \mathbf{x}, \mathbf{y} \rangle_A},\ d(\mathbf{x},\mathbf{y})_A = \| \mathbf{x} \ominus \mathbf{y} \|_A = \langle \mathbf{x} \ominus \mathbf{y}, \mathbf{x} \ominus \mathbf{y} \rangle_A,$$

kde $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$.

Doposud jsme si ukázali, že použití standardních statistických metod (speciálně korelačního koeficientu) pro kompoziční data nevede k rozumným výsledkům. To je důsledkem odlišné geometrie, která je těmto datům přirozená (Aitchisonova geometrie). Abychom mohli kompoziční data statisticky zpracovat, je potřeba data převést ze simplexu do reálného prostoru, tedy vlastně nahradit Aitchisonovu geometrii geometrií euklidovskou.

Hledáme tedy transformaci $h$, pro kterou bude platit, že

a) perturbaci a mocninnou transformaci kompozic nahradí sčítáním vektorů a násobením vektoru reálným číslem $c$,

$$h(\mathbf{x} \oplus \mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y})\ ,\ h(c \odot \mathbf{x}) = c \cdot h(\mathbf{x});$$

b) nahradí Aitchisonův skalární součin, normu (velikost) a vzdálenost příslušnými euklidovskými protějšky,

$$d_A(\mathbf{x},\mathbf{y}) = d(h(\mathbf{x}), h(\mathbf{y}))\ ,\ \| \mathbf{x} \|_A = \| h(\mathbf{x}) \|\ ,\ \langle \mathbf{x}, \mathbf{y} \rangle_A = \langle h(\mathbf{x}), h(\mathbf{y}) \rangle\ .$$

Třída takových transformací, které tyto vlastnosti splňují, se nazývá izometrické logratio (ilr) transformace [4], které $D$-složkovou kompozici zobrazí jako $(D-1)-$rozměrný reálný vektor; tedy pro $\mathbf{x} = (x_1, \ldots, x_D) \in S^D$ a jednu konkrétní volbu takové transformace obdržíme $\mathbf{z} = h(\mathbf{x}) = (z_1, \ldots, z_{D-1})$, kde

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

Přitom $\prod_{i=1}^{m} a_i = a_1 \cdot \ldots \cdot a_m$ pro $a_i \in R,\ i = 1, \ldots, m$.

Například pro trojsložkovou kompozici, tedy $D = 3$, takto obdržíme

$$z_1 = \sqrt{\frac{2}{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}\ ,\ z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}.$$

Interpretace těchto nových souřadnic obecně není vůbec snadná. Pro dvousložkovou kompozici, tedy $\mathbf{x} = (x_1, x_2)$, je ale vše velmi intuitivní. Potom totiž obdržíme jedinou souřadnici $z = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$, která vysvětluje informaci, obsaženou v podílu mezi oběma složkami.

Nyní ověříme platnost výše uvedených vlastností ilr transformace právě pro dvousložkovou kompozici. Pro perturbaci (a $\kappa = 1$):

$$h(\mathbf{x} \oplus \mathbf{y}) = h\left(\mathcal{C}(x_1 y_1, x_2 y_2)\right) = h\left(\frac{x_1 y_1}{x_1 y_1 + x_2 y_2}, \frac{x_2 y_2}{x_1 y_1 + x_2 y_2}\right) = \frac{1}{\sqrt{2}} \ln \frac{x_1 y_1}{x_2 y_2},$$

$$h(\mathbf{x}) + h(\mathbf{y}) = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2} + \frac{1}{\sqrt{2}} \ln \frac{y_1}{y_2} = \frac{1}{\sqrt{2}} \ln \frac{x_1 y_1}{x_2 y_2}.$$

Pro mocninnou trasformaci:

$$h(c \odot \mathbf{x}) = h\left(\mathcal{C}(x_1^c, x_2^c)\right) = h\left(\frac{x_1^c}{x_1^c + x_2^c}, \frac{x_2^c}{x_1^c + x_2^c}\right)$$

$$= \frac{1}{\sqrt{2}} \ln \frac{x_1^c}{x_2^c} = \frac{1}{\sqrt{2}} \ln \left(\frac{x_1}{x_2}\right)^c = \frac{c}{\sqrt{2}} \ln \frac{x_1}{x_2},$$

$$c \cdot h(\mathbf{x}) = c \cdot \left(\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}\right) = \frac{c}{\sqrt{2}} \ln \frac{x_1}{x_2}.$$

Také můžeme ověřit platnost převodního vztahu pro vzdálenost:

$$d_A(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2} =$$

$$\sqrt{\frac{1}{4}\left[\left(\ln\frac{x_1}{x_1} - \ln\frac{y_1}{y_1}\right)^2 + \left(\ln\frac{x_1}{x_2} - \ln\frac{y_1}{y_2}\right)^2 + \left(\ln\frac{x_2}{x_1} - \ln\frac{y_2}{y_1}\right)^2 + \left(\ln\frac{2_1}{x_2} - \ln\frac{y_2}{y_2}\right)^2\right]} =$$

$$\sqrt{\frac{1}{4}\left[\left(\ln\frac{x_1}{x_2} - \ln\frac{y_1}{y_2}\right)^2 + \left(\ln\frac{x_2}{x_1} - \ln\frac{y_2}{y_1}\right)^2\right]} = \sqrt{\frac{1}{4}\cdot 2\left(\ln\frac{x_1}{x_2} - \ln\frac{y_1}{y_2}\right)^2} =$$

$$= \sqrt{\frac{1}{2}\left(\ln\frac{x_1}{x_2} - \ln\frac{y_1}{y_2}\right)^2},$$

$$d(h(\mathbf{x}),h(\mathbf{y})) = d\left(\frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2}, \frac{1}{\sqrt{2}}\ln\frac{y_1}{y_2}\right) = \sqrt{\left(\frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2} - \frac{1}{\sqrt{2}}\ln\frac{y_1}{y_2}\right)^2} =$$

$$= \sqrt{\frac{1}{2}\left(\ln\frac{x_1}{x_2} - \ln\frac{y_1}{y_2}\right)^2}.$$

Jako poslední pojem, o kterém se musíme v souvislosti s kompozičními daty zmínit, je tzv. *amalgamace*, kterou označujeme součet složek kompozice. Tato operace totiž nutně vede ke vzniku nové složky s novou interpretací.

Uvažujme čtyřsložkovou kompozici $\mathbf{x} = (x_1, x_2, x_3, x_4)$, kde složka $x_1$ představuje výdaje za jídlo v korunách, $x_2$ výdaje za bydlení, $x_3$ výdaje za dopravu a $x_4$ výdaje za kulturu. Pokud sečteme složky $x_1$ a $x_2$, dostaneme složku, kterou bychom mohli nazvat základní výdaje. Naopak součtem složek $x_3$ a $x_4$ obdržíme výdaje za služby. Je zřejmé, že amalgamace může v reálné situaci usnadnit interpretaci dat a zmenšit celkový počet složek zkoumané kompozice, v takovém případě však ztratíme informaci o původních složkách, kterou již nelze zpětně obnovit. To znamená, že amalgamaci provádíme pouze v případě, kdy součet složek bude jasně definovatelný a interpretovatelný.

# 3 Charakteristika závislosti pro kompoziční data

Uvažujme znovu kompozici $\mathbf{x} = (x_1, \ldots, x_D)$. Na počátku této práce jsme si ukázali, že se korelační koeficient bohužel nedá jako charakteristika závislosti mezi dvěma složkami $x_i$ a $x_j$, $i, j = 1, \ldots, D$, použít. Budeme tedy hledat nějaký jiný způsob, jak charakterizovat těsnost vztahu mezi složkami. Zřejmě se nebudeme v tomto případě opírat o absolutní hodnoty, ale odpovídající míra bude nějakým způsobem souviset s podílem obou složek.

Jakákoliv takováto charakteristika ovšem nemůže být vypočtena přímo pro $x_i$ a $x_j$, ale pouze pro příslušnou ilr souřadnici $z_{ij} = \frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}$. Pokud tuto souřadnici budeme ve statistickém souboru kompozičních dat uvažovat jako statistický znak $Z$ (pro libovolně zvolenou, ale pevnou dvojici složek $x_i$ a $x_j$), můžeme například vypočítat příslušný rozptyl znaku $s_Z^2$. Tento rozptyl lze přitom interpretovat ve smyslu původních složek $x_i$ a $x_j$.

Jestliže totiž obdržíme malou hodnotu $s_Z^2$, bude to značit, že hodnoty statistického znaku $Z$ jsou koncentrovány kolem aritmetického průměru $\bar{z}$, což ale bude zároveň (z tvaru příslušné souřadnice) znamenat, že podíl mezi odpovídajícími složkami $x_i$ a $x_j$ bude stabilní. Takto můžeme říci, že rozptyl $s_Z^2$ vlastně představuje "míru závislosti"mezi složkami kompozice $x_i$ a $x_j$. Je potřeba si ovšem uvědomit, že se nejedná o závislost ve smyslu korelačního koeficientu, protože nevyjadřuje sílu lineárního vztahu mezi statistickými znaky. Uvedená charakteristika se také někdy vyjadřuje ve tvaru $\exp(-s_Z^2)$, který nám umožňuje obdržené hodnoty vyjádřit v intervalu $(0; 1)$. Přitom čím se bude hodnota více přibližovat jedné, tím bude podíl mezi $x_i$ a $x_j$ ve statistickém souboru stabilnější. Ze zkušenosti lze o stabilitě podílu mezi složkami hovořit přibližně od hodnoty 0,8, záleží ovšem vždy na konkrétních datech.

Pokud výše uvedené rozptyly pro všechny dvojice složek $x_i$ a $x_j$ uspořádáme do matice, obdržíme tzv. *variační matici*,

$$\mathbf{T}^* = \begin{pmatrix} t^*_{11} & t^*_{12} & \dots & t^*_{1D} \\ t^*_{21} & t^*_{22} & \dots & t^*_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D1} & t^*_{D2} & \dots & t^*_{DD} \end{pmatrix},$$

kde prvky $t^*_{ij}$ představují rozptyl statistického souboru $\{z^k_{ij} = \frac{1}{\sqrt{2}} \ln \frac{x_{ik}}{x_{jk}}, k = 1, \dots, n\}$ o rozsahu $n$, $z_{ij} = 1, \dots, D$. Pro lepší ínterpretovatelnost výsledků pak užíváme tzv. *normovanou variační matici*,

$$\exp(-\mathbf{T}^*) = \begin{pmatrix} \exp(-t^*_{11}) & \exp(-t^*_{12}) & \dots & \exp(-t^*_{1D}) \\ \exp(-t^*_{21}) & \exp(-t^*_{22}) & \dots & \exp(-t^*_{2D}) \\ \vdots & \vdots & \ddots & \vdots \\ \exp(-t^*_{D1}) & \exp(-t^*_{D2}) & \dots & \exp(-t^*_{DD}) \end{pmatrix}.$$

# 4 Praktický příklad z ekonomiky

V poslední části naší práce budeme uvedené teoretické úvahy aplikovat na reálná data. Pro tento účel jsme zvolili datový soubor Mean consumption expenditure of households z databáze statistického úřadu Eurostat [3], který vyjadřuje průměrné roční výdaje domácností (v eurech) na jednotlivé komodity ve všech 27 státech Evropské unie. Příslušné výpočty byly provedeny pomocí statistického softwaru R (*www.r-project.org*) a jeho knihovny *compositions*.

Takto byly sledovány výdaje za potraviny, alkohol a tabák, oblečení, bydlení, nábytek a vybavení domu, zdravotnictví, dopravu, komunikaci, rekreaci, vzdělávání, restaurace a hotely a na ostatní zboží a služby, které jsou vyjádřeny pro Belgii, Bulharsko, Českou republiku, Dánsko, Estonsko, Finsko, Francii, Irsko, Itálii, Kypr, Litvu, Lotyšsko, Lucembursko, Maďarsko, Maltu, Německo, Nizozemsko, Polsko, Portugalsko, Rakousko, Rumunsko, Řecko, Slovensko, Slovinsko, Španělsko, Švédsko a Velkou Británii (viz tabulka 1).

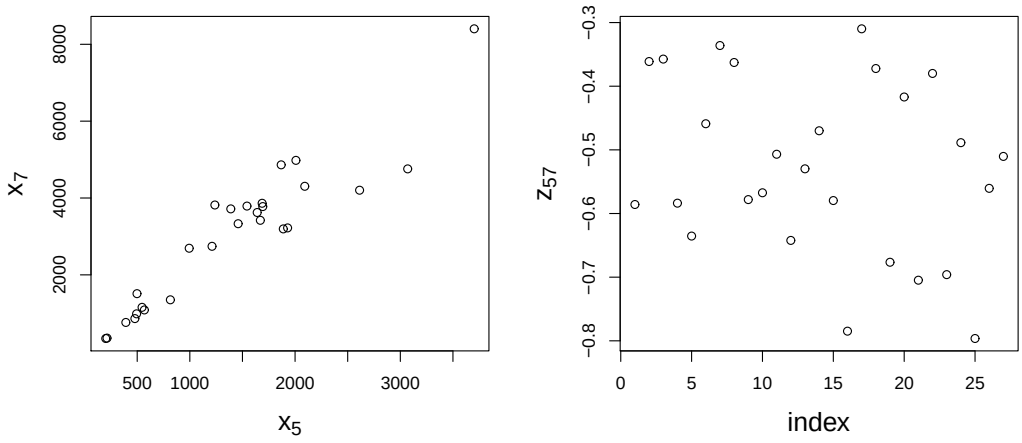
V centru našeho zájmu budou zřejmě podíly jednotlivých výdajů na výdajích celkových, protože se jedná o výdaje z celkových příjmů domácností. Například, pokud utratíme více peněz za jednu komoditu, logicky nám zbude méně peněz

Tabulka 1: Průměrné roční výdaje domácností ve státech EU (v eurech).

| Státy | potraviny | alkohol a kouření | oblečení a obuv | bydlení | vybavení domácnosti | zdraví | doprava | komunikace | kultura a rekreace | vzdělání | restaurace a hotely | ostatní |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Belgie | 4043 | 669 | 1425 | 7610 | 1687 | 1400 | 3863 | 878 | 2868 | 136 | 1894 | 3576 |
| Bulharsko | 2238 | 269 | 218 | 2461 | 213 | 305 | 355 | 325 | 204 | 34 | 255 | 220 |
| Česká Republika | 2503 | 347 | 679 | 2444 | 815 | 239 | 1351 | 555 | 1289 | 66 | 619 | 1234 |
| Dánsko | 2872 | 785 | 1168 | 7194 | 1459 | 639 | 3331 | 583 | 2738 | 100 | 960 | 2233 |
| Německo | 3185 | 489 | 1355 | 8445 | 1543 | 1024 | 3790 | 828 | 3168 | 236 | 1212 | 3226 |
| Estonsko | 2440 | 300 | 601 | 3240 | 568 | 282 | 1087 | 596 | 691 | 145 | 339 | 559 |
| Irsko | 4491 | 2032 | 1851 | 8520 | 2613 | 904 | 4203 | 1255 | 3670 | 687 | 2190 | 3956 |
| Řecko | 4801 | 1045 | 2154 | 7442 | 1929 | 1824 | 3222 | 1174 | 1285 | 738 | 2661 | 2701 |
| Španělsko | 4685 | 586 | 1786 | 7874 | 1211 | 577 | 2743 | 701 | 1659 | 292 | 2414 | 1499 |
| Francie | 3733 | 650 | 1853 | 7339 | 1693 | 1167 | 3777 | 914 | 1926 | 165 | 1277 | 3392 |
| Itálie | 5359 | 506 | 2013 | 8512 | 1670 | 1132 | 3420 | 621 | 1680 | 202 | 1428 | 2242 |
| Kypr | 5158 | 646 | 2649 | 7381 | 2008 | 1624 | 4980 | 1164 | 2044 | 1354 | 2830 | 2370 |
| Lotyšsko | 3091 | 329 | 778 | 1810 | 546 | 394 | 1155 | 610 | 667 | 145 | 557 | 508 |
| Litva | 3166 | 332 | 743 | 1776 | 392 | 445 | 762 | 435 | 402 | 102 | 429 | 393 |
| Lucembursko | 4851 | 865 | 3343 | 15611 | 3702 | 1351 | 8403 | 1139 | 3869 | 223 | 4098 | 4478 |
| Maďarsko | 2413 | 380 | 537 | 2073 | 498 | 440 | 1511 | 696 | 909 | 90 | 343 | 803 |
| Malta | 6082 | 786 | 2387 | 2596 | 3070 | 869 | 4758 | 837 | 2879 | 352 | 2030 | 1960 |
| Nizozemsko | 3089 | 625 | 1694 | 7513 | 1888 | 371 | 3196 | 903 | 3193 | 306 | 1647 | 4945 |
| Rakousko | 3933 | 847 | 1682 | 6732 | 1868 | 946 | 4863 | 793 | 3809 | 242 | 1660 | 2792 |
| Polsko | 2704 | 262 | 489 | 3341 | 478 | 485 | 862 | 512 | 662 | 138 | 180 | 571 |
| Portugalsko | 3243 | 477 | 861 | 5560 | 994 | 1264 | 2693 | 616 | 1182 | 356 | 2263 | 1359 |
| Rumunsko | 2355 | 307 | 333 | 832 | 201 | 205 | 344 | 259 | 224 | 45 | 58 | 162 |
| Slovinsko | 3966 | 575 | 1678 | 5483 | 1389 | 356 | 3717 | 950 | 2234 | 202 | 1035 | 2220 |
| Slovensko | 2910 | 333 | 661 | 2517 | 494 | 330 | 986 | 506 | 712 | 92 | 520 | 713 |
| Finsko | 3086 | 588 | 934 | 6614 | 1238 | 852 | 3818 | 693 | 2731 | 51 | 1021 | 2733 |
| Švédsko | 2913 | 531 | 1270 | 8250 | 1640 | 638 | 3623 | 791 | 3398 | 8 | 981 | 1569 |
| Velká Británie | 3159 | 753 | 1585 | 9458 | 2092 | 383 | 4305 | 852 | 3943 | 457 | 2558 | 2415 |
| Zkratka | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ |

Tabulka 2: Procentuální zastoupení jednotlivých komodit na průměrných ročních výdajích domácností v zemích EU.

| Státy | potraviny | alkohol a kouření | oblečení a obuv | bydlení | vybavení domácnosti | zdraví | doprava | komunikace | kultura a rekreace | vzdělání | restaurace a hotely | ostatní |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Belgie | 13.45 | 2.23 | 4.74 | 25.33 | 5.61 | 4.66 | 12.86 | 2.92 | 9.54 | 0.45 | 6.30 | 11.90 |
| Bulharsko | 31.53 | 3.79 | 3.07 | 34.68 | 3.00 | 4.30 | 5.00 | 4.58 | 2.87 | 0.48 | 3.59 | 3.10 |
| Česká Republika | 20.62 | 2.86 | 5.59 | 20.13 | 6.71 | 1.97 | 11.13 | 4.57 | 10.62 | 0.54 | 5.01 | 10.16 |
| Dánsko | 11.94 | 3.26 | 4.85 | 29.90 | 6.06 | 2.66 | 13.84 | 2.42 | 11.38 | 0.42 | 3.99 | 9.28 |
| Německo | 11.18 | 1.72 | 4.75 | 29.63 | 5.41 | 3.60 | 13.30 | 2.91 | 11.12 | 0.83 | 4.25 | 11.32 |
| Estonsko | 22.49 | 2.76 | 5.54 | 29.87 | 5.24 | 2.60 | 10.02 | 5.49 | 6.37 | 1.34 | 3.13 | 5.15 |
| Irsko | 12.35 | 5.59 | 5.09 | 23.42 | 7.18 | 2.49 | 11.56 | 3.45 | 10.09 | 1.89 | 6.02 | 10.88 |
| Řecko | 15.50 | 3.37 | 6.95 | 24.03 | 6.23 | 5.89 | 10.40 | 3.79 | 4.15 | 2.38 | 8.60 | 8.72 |
| Španělsko | 18.00 | 2.25 | 6.86 | 30.25 | 4.65 | 2.22 | 10.54 | 2.69 | 6.37 | 1.12 | 9.28 | 5.76 |
| Francie | 13.39 | 2.33 | 6.64 | 26.32 | 6.07 | 4.18 | 13.54 | 3.28 | 6.91 | 0.59 | 4.58 | 12.16 |
| Itálie | 18.62 | 1.76 | 7.00 | 29.57 | 5.80 | 3.93 | 11.88 | 2.16 | 5.84 | 0.70 | 4.96 | 7.79 |
| Kypr | 15.08 | 1.89 | 7.74 | 21.58 | 5.87 | 4.75 | 14.56 | 3.40 | 5.98 | 3.96 | 8.27 | 6.93 |
| Lotyšsko | 29.19 | 3.11 | 7.35 | 17.09 | 5.16 | 3.72 | 10.91 | 5.76 | 6.30 | 1.37 | 5.26 | 4.80 |
| Litva | 33.76 | 3.54 | 7.92 | 18.94 | 4.18 | 4.75 | 8.13 | 4.64 | 4.29 | 1.09 | 4.58 | 4.19 |
| Lucembursko | 9.34 | 1.67 | 6.44 | 30.06 | 7.13 | 2.60 | 16.18 | 2.19 | 7.45 | 0.43 | 7.89 | 8.62 |
| Maďarsko | 22.57 | 3.55 | 5.02 | 19.39 | 4.66 | 4.11 | 14.13 | 6.51 | 8.50 | 0.84 | 3.21 | 7.51 |
| Malta | 21.26 | 2.75 | 8.34 | 9.08 | 10.73 | 3.04 | 16.63 | 2.93 | 10.06 | 1.23 | 7.10 | 6.85 |
| Nizozemsko | 10.52 | 2.13 | 5.77 | 25.58 | 6.43 | 1.26 | 10.88 | 3.07 | 10.87 | 1.04 | 5.61 | 16.84 |
| Rakousko | 13.04 | 2.81 | 5.58 | 22.32 | 6.19 | 3.14 | 16.12 | 2.63 | 12.63 | 0.80 | 5.50 | 9.24 |
| Polsko | 25.31 | 2.45 | 4.58 | 31.27 | 4.47 | 4.50 | 48.07 | 4.79 | 6.20 | 1.29 | 1.68 | 5.34 |
| Portugalsko | 15.54 | 2.29 | 4.13 | 26.64 | 4.76 | 6.06 | 12.90 | 2.95 | 5.66 | 1.71 | 10.84 | 6.51 |
| Rumunsko | 44.23 | 5.77 | 6.25 | 15.62 | 3.77 | 3.85 | 6.46 | 4.86 | 4.21 | 0.85 | 1.09 | 3.04 |
| Slovinsko | 16.66 | 2.42 | 7.05 | 23.03 | 5.83 | 1.50 | 15.61 | 3.99 | 9.38 | 0.85 | 4.35 | 9.33 |
| Slovensko | 27.00 | 3.09 | 6.14 | 23.36 | 4.59 | 3.06 | 9.15 | 4.70 | 6.61 | 0.85 | 4.83 | 6.62 |
| Finsko | 12.67 | 2.41 | 3.83 | 27.15 | 5.08 | 3.50 | 15.67 | 2.84 | 11.21 | 0.21 | 4.19 | 11.22 |
| Švédsko | 11.37 | 2.07 | 4.96 | 32.21 | 6.40 | 2.49 | 14.15 | 3.09 | 13.27 | 0.03 | 3.83 | 6.13 |
| Velká Británie | 9.88 | 2.36 | 4.96 | 29.60 | 6.55 | 1.20 | 13.47 | 2.66 | 12.34 | 1.43 | 8.00 | 7.56 |
| Evropská unie | 18.22 | 2.85 | 6.02 | 25.67 | 5.86 | 3.34 | 12.38 | 3.72 | 8.04 | 0.85 | 5.17 | 7.88 |
| Zkratka | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ |




Obrázek 2: Zobrazení původních složek $x_5$ a $x_7$ (vlevo) a statistického znaku $z_{57}$ vzniklého jejich ilr transformací.

na ostatní, z čehož vyplývá, že ne absolutní hodnoty, ale pouze jejich poměry jsou informativní. Uvedená data jsou tedy zřejmě opravdu kompoziční data. Ilustrujme si toto obrázku 2, kde jsou vlevo zobrazeny původní složky $x_5$ (vybavení domácnosti) a $x_7$ (doprava) v eurech a vpravo statistický znak $z_{57}$ vzniklý ilr transformací těchto složek. Takto vidíme zcela jinou konfiguraci pozorování (např. "zmizí"odlehlé pozorování z levého obrázku).

Na úvod vypočítáme průměrné výdaje na jednotlivé komodity v rámci celé Evropské unie, tedy spočítáme geometrické průměry jednotlivých sloupců a tyto výdaje znázorníme jako procentuální podíly na celku v tabulce 2. Dalo by se dokázat, že tato charakteristika je pro použití v případě kompozičních dat vhodnější, než kdybychom použili vektor aritmetických průměrů.

Po výpočtu takto obdržíme, že obyvatel státu, který je součástí Evropské unie, průměrně vydá za jeden rok 18.22 procent ze svých příjmů za potraviny, 2.85 procent za alkohol a kouření, 6.02 procent za oblečení a obuv, 25.67 procent na bydlení, 5.86 procent za domácí vybavení, 3.34 procent pro udržení

zdraví, 12.38 procent za dopravu, 3.72 procent za komunikaci, 8.04 procent za kulturní vyžití, 0.85 procent na vzdělání, 5.17 procent za restaurace, hotely a podobné služby a 7.88 procent ročně ze svých příjmů za ostatní služby.

Je zřejmé, že první a druhá tabulka obsahují tutéž informaci o příslušných datech, avšak data v první tabulce vyjadřují pouze absolutní hodnoty, které mohou poskytovat zkreslené informace. To je dáno zřejmě například tím, že kolísá cenová hladina v jednotlivých státech, tedy ačkoliv mají lidé v bohatších státech obecně vyšší příjmy, roste tím většinou také nutnost vyšších výdajů, protože na velikost příjmů reagují obchodníci vyššími cenami svých produktů.

Jestliže nás zajímají závislosti mezi jednotlivými složkami, musíme se důsledně práci s absolutními hodnotami vyvarovat, takto by byly námi získané údaje značně nespolehlivé. Tohle tvrzení dokazuje fakt, že výsledek ovlivní už jen to, zda-li pro výpočet korelačních koeficientů použijeme naměřené hodnoty v eurech či procentech.

V příloze 1 jsou spočítány korelační koeficienty mezi jednotlivými komoditami, které jsou vyjádřeny v eurech (tabulka 1). V příloze 2 jsou pak spočítány korelační koeficienty pro tatáž data, která ovšem byla vyjádřena pomocí procentuálních podílů (tabulka 2). Na první pohled je zřejmé, že korelační koeficienty spočítané z absolutních hodnot se zjevně liší od korelačních koeficientů spočítaných z procentuálních podílů, a tudíž mezi nimi neexistuje žádná souvislost, i když vlastně oba způsoby interpretace dat vyjadřují tutéž informaci. Tento nesoulad je především zapříčiněn tím, že jsme naše data přeškálovali a následně nepoužili odpovídající nástroje pro analýzu kompozičních dat.

Podíváme-li se na tuto situaci podrobněji, rozdíl lze pěkně ilustrovat na korelačním koeficientu mezi výdaji za oblečení a obuv a výdaji za komunikaci. V příloze 1 totiž nabývá korelační koeficient hodnoty 0.82, což znamená silnou (kladnou) závislost. Naopak v případě výpočtu z procentuálních dat (příloha 2) nabývá korelační koeficient hodnoty 0.02, což znamená naprosto žádnou lineární závislost. Stejně tak je tomu u výdajů za bydlení a za restaurace a hotely. Korelační koeficient v první matici nabývá hodnoty 0.8 a ve druhé hodnoty $-0.01$.

Proto je pro zpracování dat nutné nalézt takovou metodu, u které nám nebude záležet na absolutních hodnotách, ale právě na relativních proporcích. Tyto získáme vypočítáním podílů mezi jednotlivými složkami. V tomto případě nebude mít totiž celkový součet složek na výsledek vliv, takže můžeme absolutní hodnoty libovolně přeškálovat a výsledná informace zůstane zachována.

Míru vztahu závislosti pro složky kompozice, která není založena na absolutních hodnotách (jak je tomu u korelačního koeficientu), nýbrž na podílech mezi jednotlivými složkami, nám nabízí variační matice (tabulka 3) nebo její lépe interpretovatelná podoba - normovaná variační matice (tabulka 4). Teprve tyto nám tak poskytnou žádanou plnohodnotnou interpretaci vztahů mezi jednotlivými výdaji.

Z normované variační matice zjistíme, že existují dvě skupiny složek, jejichž podíly jsou navzájem velmi stabilní. První skupinu lze označit jako osobní výdaje a zahrnujeme do ní výdaje na potraviny, alkohol a kouření a komunikaci (příslušné hodnoty jsou vyznačeny zeleně). Druhá skupina by se dala označit jako externí výdaje a patří do ní výdaje na oblečení a obuv, vybavení domácnosti, dopravu, kulturu a rekreaci a ostatní výdaje (zvýrazněno modrou barvou).

Dále, v posledním řádku tabulky 4 jsou vypočteny průměrné hodnoty obdržených charakteristik. Z nich lze vysledovat, že podíly výdajů na vybavení domácnosti jsou vůči ostatním výdajům stabilní, příslušná průměrná procentuální hodnota 5,86 (tabulka 2) tak bude poměrně přesně charakterizovat tyto výdaje ve všech státech EU. To samé platí pro výdaje na oblečení a obuv a pro výdaje na dopravu, kde odpovídající podíl celkových výdajů vyjádřených procentuálně na začátku tohoto příkladu bude opět velmi stabilní. Ostatní podíly sice už takovou stabilitu nevykazují, ale jejich hodnoty jsou i tak vyrovnané. Tohle tvrzení se ovšem netýká vzdělání, u kterého je tomu přesně naopak. Na vzdělání jsou totiž výdaje z celkových příjmů v různých státech proporciálně značně odlišné. To je způsobeno tím, že v některých státech jsou výdaje na vzdělání několikanásobně vyšší než v jiných. Například na Kypru je to 3,96 procent a ve Švédsku 0,03 procent (viz tabulka 2), což znamená, že

na Kypru je procentuální část celkových příjmů domácností vydaná na vzdělání 132-krát vyšší než ve Švédsku.

Je zřejmé, že právě aritmetické průměry sloupců, ev. řádků, normované variační matice představují taktéž klíčovou informaci. Udávají totiž stability podílů jednotlivých složek vzhledem ke všem ostatním složkám v kompozici. To může být v praxi snadněji interpretovatelné než jednotlivé prvky matice, které zase narozdíl od průměrů poskytují o stabilitě podílů složek v kompozici úplnou informaci.

Tabulka 3: Variační matice.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ostatní | 0.60 | 0.37 | 0.22 | 0.21 | 0.10 | 0.46 | 0.09 | 0.38 | 0.10 | 1.01 | 0.25 | 0.00 | |
| restaurace a hotely | 0.63 | 0.47 | 0.25 | 0.31 | 0.19 | 0.46 | 0.19 | 0.49 | 0.33 | 0.81 | 0.00 | 0.25 | |
| vzdělání | 0.80 | 0.78 | 0.69 | 1.02 | 0.84 | 0.85 | 0.94 | 0.76 | 1.20 | 0.00 | 0.81 | 1.01 | |
| kultura a rekreace | 0.55 | 0.33 | 0.23 | 0.23 | 0.09 | 0.54 | 0.07 | 0.35 | 0.00 | 1.20 | 0.33 | 0.10 | |
| komunikace | 0.07 | 0.08 | 0.14 | 0.22 | 0.22 | 0.22 | 0.27 | 0.00 | 0.35 | 0.76 | 0.49 | 0.38 | |
| doprava | 0.42 | 0.27 | 0.11 | 0.18 | 0.03 | 0.32 | 0.00 | 0.27 | 0.07 | 0.94 | 0.19 | 0.46 | |
| zdraví | 0.21 | 0.24 | 0.25 | 0.28 | 0.31 | 0.00 | 0.32 | 0.22 | 0.54 | 0.85 | 0.46 | 0.09 | |
| vybavení domácnosti | 0.34 | 0.20 | 0.06 | 0.18 | 0.00 | 0.31 | 0.03 | 0.22 | 0.09 | 0.84 | 0.19 | 0.10 | |
| bydlení | 0.34 | 0.24 | 0.21 | 0.00 | 0.18 | 0.28 | 0.18 | 0.22 | 0.23 | 0.69 | 0.25 | 0.00 | |
| oblečení a obuv | 0.19 | 0.16 | 0.00 | 0.21 | 0.06 | 0.25 | 0.11 | 0.14 | 0.23 | 0.69 | 0.25 | 0.22 | |
| alkohol a kouření | 0.12 | 0.00 | 0.16 | 0.24 | 0.20 | 0.24 | 0.27 | 0.08 | 0.33 | 0.78 | 0.47 | 0.37 | |
| potraviny | 0.00 | 0.12 | 0.19 | 0.34 | 0.34 | 0.21 | 0.42 | 0.07 | 0.55 | 0.80 | 0.63 | 0.60 | |
| | potraviny | alkohol a kouření | oblečení a obuv | bydlení | domácí vybavení | zdraví | doprava | komunikace | kultura a rekreace | vzdělání | restaurace a hotely | ostatní | |

Tabulka 4: Normovaná variační matice.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ostatní | 0.54 | 0.68 | 0.79 | 0.80 | 0.89 | 0.62 | 0.91 | 0.68 | 0.90 | 0.36 | 1.00 | 0.72 | |
| restaurace a hotely | 0.52 | 0.62 | 0.77 | 0.72 | 0.82 | 0.62 | 0.82 | 0.61 | 0.71 | 0.44 | 1.00 | 0.77 | |
| vzdělání | 0.44 | 0.45 | 0.49 | 0.35 | 0.42 | 0.42 | 0.38 | 0.46 | 0.29 | 1.00 | 0.44 | 0.36 | 0.41 |
| kultura a rekreace | 0.57 | 0.71 | 0.79 | 0.78 | 0.91 | 0.58 | 0.93 | 0.70 | 1.00 | 0.29 | 0.71 | 0.90 | 0.72 |
| komunikace | 0.92 | 0.91 | 0.86 | 0.80 | 0.80 | 0.80 | 0.76 | 1.00 | 0.70 | 0.46 | 0.61 | 0.68 | 0.76 |
| doprava | 0.65 | 0.75 | 0.89 | 0.83 | 0.96 | 0.72 | 1.00 | 0.76 | 0.93 | 0.38 | 0.82 | 0.91 | 0.78 |
| zdraví | 0.80 | 0.78 | 0.77 | 1.00 | 0.72 | 1.00 | 0.72 | 0.80 | 0.58 | 0.42 | 0.62 | 0.62 | 0.69 |
| vybavení domácnosti | 0.70 | 0.81 | 0.93 | 0.83 | 1.00 | 0.72 | 0.96 | 0.80 | 0.91 | 0.42 | 0.82 | 0.89 | 0.80 |
| bydlení | 0.70 | 0.78 | 0.80 | 1.00 | 0.83 | 0.75 | 0.83 | 0.80 | 0.78 | 0.35 | 0.72 | 0.80 | 0.74 |
| oblečení a obuv | 0.82 | 0.84 | 1.00 | 0.80 | 0.93 | 0.77 | 0.89 | 0.86 | 0.79 | 0.49 | 0.77 | 0.80 | 0.80 |
| alkohol a kouření | 0.88 | 1.00 | 0.84 | 0.78 | 0.81 | 0.78 | 0.75 | 0.91 | 0.71 | 0.45 | 0.62 | 0.68 | 0.75 |
| potraviny | 1.00 | 0.88 | 0.82 | 0.70 | 0.70 | 0.80 | 0.65 | 0.92 | 0.57 | 0.44 | 0.52 | 0.54 | 0.69 |
| | potraviny | alkohol a kouření | oblečení a obuv | bydlení | domácí vybavení | zdraví | doprava | komunikace | kultura a rekreace | vzdělání | restaurace a hotely | ostatní | průměr |

# Závěr

Během práce na naší SOČ jsme se dozvěděli mnoho poznatků o kompozičních datech, zjistili jsme především, že mohou existovat data nesoucí relativní informace, se kterými nemůžeme pracovat v euklidovské geometrii, ale musíme použít jinou, Aitchisonovu. Také jsme zjistili, jakým způsobem se dostaneme z Aitchisonovy geometrie do geometrie euklidovské, abychom následně mohli použít příslušných statistických metod, díky kterým lze charakterizovat vztahy mezi složkami kompozic.

Na příkladu z oblasti ekonomiky jsme si ukázali, že odlišný způsob práce s kompozičními daty je nezbytný a nesmí se přehlížet. Jak jsme uvedli už na začátku naší práce, kompoziční data se přitom nevyskytují pouze v ekonomii, ale i v mnoha dalších běžných oborech, jako jsou geologie, chemie nebo medicína. I to je jeden z důvodů, proč jsme se zabývali právě kompozičními daty. Mohli jsme si tak totiž vyzkoušet práci s reálnými daty na vlastní kůži, jinak řečeno jsme měli možnost uvést teorii v praxi, čehož se na gymnáziu obecně moc nedostává.

V uvedeném příkladu je též obsažen hlavní přínos naší práce. V dostupné literatuře byla totiž kompoziční data dosud aplikována především na data z oblasti geologie, tedy takto podrobná analýza ekonomického problému (spolu s odpovídající interpretací výsledků) je z tohoto pohledu inovativní. Při řešení tohoto příkladu jsme navíc nově použili aritmetické průměry sloupců (ev. možno i řádků) normované variační matice, které umožní provést úsudek o stabilitě jednotlivých složek vzhledem k ostatním složkám v kompozici. Ukazuje se přitom, že tyto dodatečné charakteristiky dále zhodnocují informaci obsaženou v normované variační matici. Ta totiž nevypovídá o celkové stabilitě složek, ale pouze o stabilitě podílů jednotlivých dvojic složek kompozice.

Již nyní můžeme říci, že nám práce na této SOČ byla neocenitelnou zkušeností. Dozvěděli jsme se mnoho užitečných informací a získali nové zkušenosti, kterých si nesmírně vážíme, a které nás posunuly zase o kousek výš. Přitom to, že jsme měli možnost pracovat i s reálnými daty a takto "viděli matematiku v praxi", utužilo náš zájem o ni a vědu jako takovou.

# Literatura

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

[2] BUDÍKOVÁ, M., MIKOLÁŠ, Š., OSECKÝ, P. Popisná statistika. Brno: Přírodovědecká fakulta MU, 2002.

[3] Eurostat (2008). Mean consumption expenditures (in euro) of households on 12 domestic year costs in all 27 member states of the european union (2005) [online]. [Cit. 1.4.1999]. Dostupné z URL: http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Household_consumption_expenditure

[4] EGOZCUE, JJ., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G., BARCELÓ-VIDAL, C. Isometric logratio transformations for compositional data analysis. Mathematical Geology, 2003, vol. 35, p. 279 – 300.

[5] HRON, K. Elementy statistické analýzy kompozičních dat. Sborník konference Robust 2010, odesláno.

[6] HRON, K., TEMPL, M., FILZMOSER, P. Imputation of missing values for compositional data using classical and robust methods. Computation Statistics and Data Analysis, v tisku.

Příloha 1: Korelační matice pro původní data (tabulka 1).

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ostatní | 0.47 | 0.64 | 0.73 | 0.80 | 0.80 | 0.54 | 0.80 | 0.78 | 0.81 | 0.33 | 0.67 | 1.00 | |
| restaurace a hotely | 0.72 | 0.58 | 0.87 | 0.80 | 0.84 | 0.70 | 0.84 | 0.77 | 0.62 | 0.62 | 1.00 | 0.67 | |
| vzdělání | 0.57 | 0.49 | 0.58 | 0.31 | 0.45 | 0.58 | 0.40 | 0.67 | 0.22 | 1.00 | 0.62 | 0.33 | |
| kultura a rekreace | 0.40 | 0.62 | 0.64 | 0.76 | 0.82 | 0.33 | 0.84 | 0.67 | 1.00 | 0.22 | 0.62 | 0.81 | |
| komunikace | 0.64 | 0.74 | 0.82 | 0.69 | 0.81 | 0.66 | 1.00 | 0.67 | 0.67 | 0.67 | 0.77 | 0.78 | |
| doprava | 0.67 | 0.57 | 0.89 | 0.85 | 0.93 | 1.00 | 0.78 | 0.84 | 0.84 | 0.40 | 0.84 | 0.80 | |
| zdraví | 0.67 | 0.44 | 0.67 | 0.58 | 1.00 | 0.63 | 0.66 | 0.33 | 0.58 | 0.58 | 0.70 | 0.54 | |
| vybavení domácnosti | 0.75 | 0.69 | 0.91 | 0.77 | 1.00 | 0.58 | 0.81 | 0.82 | 0.45 | 0.45 | 0.84 | 0.80 | |
| bydlení | 0.47 | 0.52 | 0.76 | 1.00 | 0.77 | 0.58 | 0.85 | 0.69 | 0.76 | 0.31 | 0.80 | 0.80 | |
| oblečení a obuv | 0.86 | 0.56 | 1.00 | 0.76 | 0.91 | 0.67 | 0.89 | 0.82 | 0.64 | 0.58 | 0.87 | 0.73 | |
| alkohol a kouření | 0.51 | 1.00 | 0.56 | 0.52 | 0.69 | 0.44 | 0.57 | 0.74 | 0.62 | 0.49 | 0.58 | 0.64 | |
| potraviny | 1.00 | 0.51 | 0.86 | 0.47 | 0.75 | 0.67 | 0.67 | 0.64 | 0.40 | 0.57 | 0.72 | 0.47 | |
| | potraviny | alkohol a kouření | oblečení a obuv | bydlení | domácí vybavení | zdraví | doprava | komunikace | kultura a rekreace | vzdělání | restaurace a hotely | ostatní | |

Příloha 2: Korelační matice pro data vyjádřená v procentech (tabulka 2).

| | potraviny | alkohol a kouření | oblečení a obuv | domácí vybavení | bydlení | domácí vybavení | zdraví | doprava | kultura a rekreace | vzdělání | restaurace a hotely | ostatní |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| potraviny | 1.00 | 0.58 | 0.18 | -0.39 | -0.50 | 0.28 | -0.72 | 0.69 | -0.65 | -0.01 | -0.48 | -0.71 |
| alkohol a kouření | 0.58 | 1.00 | -0.06 | -0.38 | -0.20 | 0.08 | -0.49 | 0.46 | -0.26 | 0.05 | -0.34 | -0.30 |
| oblečení a obuv | 0.18 | -0.06 | 1.00 | -0.60 | 0.38 | -0.01 | 0.14 | 0.01 | -0.24 | 0.37 | 0.22 | -0.12 |
| domácí vybavení | -0.39 | -0.38 | -0.60 | 1.00 | -0.35 | -0.08 | -0.12 | -0.32 | 0.04 | -0.24 | -0.01 | 0.07 |
| bydlení | -0.50 | -0.20 | 0.38 | -0.35 | 1.00 | -0.35 | 0.65 | -0.44 | 0.53 | 0.09 | 0.35 | 0.40 |
| domácí vybavení | 0.28 | 0.08 | -0.018 | -0.08 | -0.35 | 1.00 | -0.23 | 0.17 | -0.6 | 0.32 | 0.11 | -0.29 |
| zdraví | -0.72 | -0.49 | 0.14 | -0.12 | 0.65 | -0.23 | 1.00 | -0.51 | 0.66 | -0.02 | 0.35 | 0.46 |
| doprava | 0.69 | 0.46 | .019 | -0.32 | -0.44 | 0.17 | -0.51 | 1.00 | -0.41 | 0.09 | -0.50 | -0.44 |
| kultura a rekreace | -0.65 | -0.26 | -0.24 | 0.04 | 0.53 | -0.61 | 0.66 | -0.41 | 1.00 | -0.32 | -0.01 | 0.57 |
| vzdělání | -0.01 | 0.05 | 0.37 | -0.24 | 0.09 | 0.32 | -0.02 | 0.09 | -0.32 | 1.00 | 0.46 | -0.13 |
| restaurace a hotely | -0.48 | -0.34 | 0.22 | -0.01 | 0.35 | 0.11 | 0.35 | - 0.50 | -0.01 | 0.46 | 1.00 | 0.15 |
| ostatní | -0.71 | -0.30 | -0.12 | 0.07 | 0.40 | -0.29 | 0.46 | -0.44 | 0.57 | -0.13 | 0.15 | 1.00 |

# Abstrakt

Cílem naší práce bylo navržení postupu pro analýzu závislosti mezi pro- měnnými pro data, zvaná kompoziční, která obsahují pouze relativní informaci, tedy například pro data vyjadřující procentuální podíly. V takovém případě se totiž ukazuje, že nelze použit standardní korelační koeficient, ale je ho potřeba nahradit jinou charakteristikou.

V první části práce jsme uvedli základy popisné statistiky včetně korelačního koeficientu a seznámili jsme se s kompozičními daty. Také jsme zjistili, že euklidovská geometrie není vhodná pro práci s tímto typem pozorování, což se následně odráží i při jejich statistickém zpracování.

Druhá část je věnována speciální geometrii kompozičních dat na simplexu s odpovídajícími operacemi, které jsou analogiemi operací ve standardní vektorové algebře. Také jsme odvodili vztahy pro Aitchisonovu vzdálenost, normu a skalární součin a způsob vyjádření kompozic v euklidovské geometrii.

Nakonec je v naší práci uvedena charakteristika závislosti mezi složkami kompozičních dat, kterou jsme následně aplikovali na praktickém příkladě z ekonomiky a snažili se najít možné souvislosti mezi statistickými znaky, výdaji za různé komodity z celkových výdajů domácností ve státech EU. Tedy zajímalo nás, které z podílů mezi jednotlivými složkami jsou stabilní, které naopak nikoliv, a podle toho jsme se pokusili vyvodit příslušné závěry.

V uvedeném příkladu je též obsažen hlavní přínos naší práce. V dostupné literatuře byla totiž kompoziční data dosud aplikována především na data z oblasti geologie, tedy takto podrobná analýza ekonomického problému (spolu s odpovídající interpretací výsledků) je z tohoto pohledu inovativní. Při řešení tohoto příkladu jsme navíc nově použili aritmetické průměry sloupců (ev. řádků) normované variační matice, které umožní provést úsudek o stabilitě jednotlivých složek vzhledem k ostatním složkám v kompozici.

**Klíčová slova**

Korelační koeficient, kompoziční data, relativní informace, Aitchisonova geometrie na simplexu, variační matice.