# Možnosti jednorozměrné statistické analýzy kompozičních dat

K. Hron[1], P. Filzmoser[2], E. Fišerová[1], P. de Caritat[3], A. Gardlo[1]

[1]Katedra matematické analýzy a aplikací matematiky - Univerzita Palackého
[2]Institut für Stochastik und Wirtschaftsmathematik - Technische Universität Wien, Österreich
[3]Geoscience Australia

Robust 2016, 12. září 2016

# Kompoziční data

= *D-složkové vektory, popisující kvantitativně části nějakého celku, nesoucí relativní informaci o složkách* (Aitchison, 1986; Pawlowsky-Glahn a kol., 2015)

# Kompoziční data

= *D-složkové vektory, popisující kvantitativně části nějakého celku, nesoucí relativní informaci o složkách* (Aitchison, 1986; Pawlowsky-Glahn a kol., 2015)

- **obvyklé jednotky měření**: procenta, mg/kg (*konstantní součet složek*), mg/l (*konstantní součet se nevyskytuje*)

# Kompoziční data

= *D-složkové vektory, popisující kvantitativně části nějakého celku, nesoucí relativní informaci o složkách* (Aitchison, 1986; Pawlowsky-Glahn a kol., 2015)

- **obvyklé jednotky měření**: procenta, mg/kg (*konstantní součet složek*), mg/l (*konstantní součet se nevyskytuje*)

- **příklady**: geochemická data - proporce minerálů v hornině; koncentrace fenolických kyselin ve víně (mg/l); výdaje domácností na různé položky (jídlo, bydlení, ošacení) apod.

# Kompoziční data

= *D-složkové vektory, popisující kvantitativně části nějakého celku, nesoucí relativní informaci o složkách* (Aitchison, 1986; Pawlowsky-Glahn a kol., 2015)

- **obvyklé jednotky měření**: procenta, mg/kg (*konstantní součet složek*), mg/l (*konstantní součet se nevyskytuje*)
- **příklady**: geochemická data - proporce minerálů v hornině; koncentrace fenolických kyselin ve víně (mg/l); výdaje domácností na různé položky (jídlo, bydlení, ošacení) apod.
- konstantní součet složek (1, 100) = *vhodná reprezentace kompozic*

# Geometrické aspekty analýzy kompozičních dat

- požadavky na relevantní analýzu kompozic: *invariance na změnu měřítka*, *podkompoziční soudržnost*, *zachování relativního měřítka* $\Rightarrow$ **Aitchisonova geometrie** (AG; EVP dimenze $D-1$)

- většina statistických metod se opírá o předpoklad euklidovské geometrie (Eaton, 1983)

# Geometrické aspekty analýzy kompozičních dat

- požadavky na relevantní analýzu kompozic: *invariance na změnu měřítka*, *podkompoziční soudržnost*, *zachování relativního měřítka* $\Rightarrow$ **Aitchisonova geometrie** (AG; EVP dimenze $D-1$)

- většina statistických metod se opírá o předpoklad euklidovské geometrie (Eaton, 1983)

$\Rightarrow$ vyjádřit kompoziční data v souřadnicích vzhledem k ortonormální bázi na simplexu (Egozcue a kol., 2003) $\rightarrow$ statistická analýza, interpretace (*bilance, absence standardních (kartézských) souřadnic*)

- **log-podílová (log-ratio) analýza** kompozičních dat (Aitchison, 1986)

# Problém

- **výchozí pozice:** kompoziční data jsou z definice mnohorozměrná, libovolná složka kompozice nemůže být uvažována nezávisle na ostatních (veškerá relevantní informace je obsažena v podílech mezi složkami)

- **dáno:** $D$-složková kompozice $\mathbf{x} = (x_1, \ldots, x_D)'$, vyjádřená v $D - 1$ ortonormálních souřadnicích vzhledem k AG

# Problém

- **výchozí pozice:** kompoziční data jsou z definice mnohorozměrná, libovolná složka kompozice nemůže být uvažována nezávisle na ostatních (veškerá relevantní informace je obsažena v podílech mezi složkami)

- **dáno:** $D$-složková kompozice $\mathbf{x} = (x_1, \ldots, x_D)'$, vyjádřená v $D - 1$ ortonormálních souřadnicích vzhledem k AG

- **cíl:** zachytit relativní informaci o **konkrétní složce**, řekněme $x_1$, pomocí jedné ze souřadnic

# Problém

- **výchozí pozice:** kompoziční data jsou z definice mnohorozměrná, libovolná složka kompozice nemůže být uvažována nezávisle na ostatních (veškerá relevantní informace je obsažena v podílech mezi složkami)

- **dáno:** $D$-složková kompozice $\mathbf{x} = (x_1, \ldots, x_D)'$, vyjádřená v $D - 1$ ortonormálních souřadnicích vzhledem k AG

- **cíl:** zachytit relativní informaci o **konkrétní složce**, řekněme $x_1$, pomocí jedné ze souřadnic

- **otázka:** které složky musíme uvažovat, např. při tvorbě geochemické mapy pro složku $x_1$?

# Problém

- **výchozí pozice:** kompoziční data jsou z definice mnohorozměrná, libovolná složka kompozice nemůže být uvažována nezávisle na ostatních (veškerá relevantní informace je obsažena v podílech mezi složkami)

- **dáno:** $D$-složková kompozice $\mathbf{x} = (x_1, \ldots, x_D)'$, vyjádřená v $D - 1$ ortonormálních souřadnicích vzhledem k AG

- **cíl:** zachytit relativní informaci o **konkrétní složce**, řekněme $x_1$, pomocí jedné ze souřadnic

- **otázka:** které složky musíme uvažovat, např. při tvorbě geochemické mapy pro složku $x_1$?

- **problém:** některé složky mohou být zatíženy chybou měření, což může ovlivnit relativní informaci o $x_1$

# Souřadnicová reprezentace

- **veškerá relativní informace** o složce $x_1$ může být reprezentována souřadnicí $z_1$, lze zkonstruovat ortonormální souřadnice (pivotové bilance) $\mathbf{z} = (z_1, \ldots, z_{D-1})'$ k $z_1$ jako

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{k=i+1}^{D} x_k}}, \quad i = 1, \ldots, D-1$$

  (Fišerová a Hron, 2011)

- složka $x_1$ je obsažena pouze v $z_1$, která může být taktéž vyjádřena jako normovaný součet $\ln(x_1/x_2) + \cdots + \ln(x_1/x_D)$

- zbývající souřadnice $(z_2, \ldots, z_{D-1})$ reprezentují podkompozici o složkách $x_2, \ldots, x_D$

# Vážené pivotové bilance

- **vážený protějšek** k $z_1$ (Filzmoser a Hron, 2015): vezmeme vážený součet párových log-podílů k $x_1$:

$$\alpha_2 \ln \frac{x_1}{x_2} + \ldots + \alpha_D \ln \frac{x_1}{x_D}, \quad \alpha_k > 0,\ k = 1, \ldots, D;\ \alpha_2 + \cdots + \alpha_D = 1$$

- s normující konstantou (pro dosažení ortonormality) obdržíme souřadnici

$$w_1 = \frac{1}{\sqrt{1 + \sum_{k=2}^{D} \alpha_k^2}} \ln \frac{x_1}{\prod_{k=2}^{D} x_k^{\alpha_k}}$$

- zbývajících $D - 2$ souřadnic získáme řešením posloupnosti vhodně zvolených lineárních homogenních systémů

# Vážené pivotové bilance: konstrukce

1) vyjádříme váženou pivotovou bilanci jako vektor logkontrastových koeficientů, $w_1 = \mathbf{w}_1' \ln \mathbf{x}$

$$\mathbf{w}_1 = \frac{1}{\sqrt{1 + \sum_{k=2}^{D} \alpha_k^2}} (1, -\alpha_2, \ldots, -\alpha_D)'$$

2) vektory koeficientů $\mathbf{w}_2, \ldots, \mathbf{w}_{D-1}$ získáme postupně při dodržení požadavku ortonormality $\mathbf{w}_i' \mathbf{w}_j = \delta_{ij}$, $i, j = 1, \ldots, D-1$ a logkontrastové podmínky $\mathbf{w}_i' \mathbf{1} = 0$

3) volné parametry v lineárních systémech jsou užity pro zajištění toho, že je informace o $x_1$ obsažena pouze v $w_1$ a $w_{D-1}$

# Vážené pivotové bilance: konstrukce

1) vyjádříme váženou pivotovou bilanci jako vektor logkontrastových koeficientů, $w_1 = \mathbf{w}_1' \ln \mathbf{x}$

$$\mathbf{w}_1 = \frac{1}{\sqrt{1 + \sum_{k=2}^{D} \alpha_k^2}} (1, -\alpha_2, \ldots, -\alpha_D)'$$

2) vektory koeficientů $\mathbf{w}_2, \ldots, \mathbf{w}_{D-1}$ získáme postupně při dodržení požadavku ortonormality $\mathbf{w}_i' \mathbf{w}_j = \delta_{ij}$, $i, j = 1, \ldots, D-1$ a logkontrastové podmínky $\mathbf{w}_i' \mathbf{1} = 0$

3) volné parametry v lineárních systémech jsou užity pro zajištění toho, že je informace o $x_1$ obsažena pouze v $w_1$ a $w_{D-1}$

$\Rightarrow$ lze předpokládat, $w_1$ obsahuje relevantní relativní informaci o $x_1$ a $w_{D-1}$ tu zbývající (redundantní)

# Vážené pivotové bilance: obecné vztahy

- $\mathbf{w}_2 = (\underbrace{0, \ldots, 0}_{D-3}, \alpha_{D-1} - \alpha_D, \alpha_D - \alpha_{D-2}, \alpha_{D-2} - \alpha_D)'$,
- $\mathbf{w}_p = (\underbrace{0, \ldots, 0}_{D-p-1}, w_{p,D-p}, w_{p,D-p+1} \ldots, w_{p,D})'$,

  $p = 3, \ldots, D-2$:

$$w_{p,D-p} = 2 \sum_{i=D-p+1}^{D-1} \sum_{j=i+1}^{D} \alpha_i \alpha_j - (p-1) \sum_{i=D-p+1}^{D} \alpha_i^2,$$

$$w_{p,k} = \sum_{\substack{i=D-p+1 \\ i \neq k}}^{D} \alpha_i^2 - (\alpha_{D-p} + \alpha_k) \sum_{\substack{i=D-p+1 \\ i \neq k}}^{D} \alpha_i + (p-1)\alpha_k \alpha_{(D-p)},$$

$$k = D-p+1, D-p+2, \ldots, D$$

# Vážené pivotové bilance: obecné vztahy

- pro $\mathbf{w}_{D-1} = (w_{D-1,1}, w_{D-1,2}, \ldots, w_{D-1,D})'$ obdržíme následující koeficienty

$$w_{D-1,1} = 2\sum_{i=2}^{D-1}\sum_{j=i+1}^{D}\alpha_i\alpha_j - (D-2)\sum_{i=2}^{D}\alpha_i^2,$$

$$w_{D-1,k} = (D-2)\sum_{\substack{i=2\\ i\neq k}}^{D}\alpha_i^2 + (1-\alpha_k)\sum_{\substack{i=2\\ i\neq k}}^{D}\alpha_i - (D-2)\alpha_k,\ \ k = 2,\ldots,D$$

- specifická role $x_1$ v souřadnici $w_{D-1}$ je zřejmá
- následně lze určit normující konstanty (pro obdržení *ortonormálních* souřadnic)

# Vážené pivotové bilance: případ $D = 3$

- **případ $D = 3$:**

$$w_1 = \frac{1}{\sqrt{2(1-\alpha_2\alpha_3)}} \ln \frac{x_1}{x_2^{\alpha_2} x_3^{\alpha_3}}$$

$$w_2 = \frac{1}{\sqrt{6(1-\alpha_2\alpha_3)}} \ln x_1^{\alpha_3-\alpha_2} x_2^{-(1+\alpha_3)} x_3^{1+\alpha_2}$$

- poznamenejme, že v závislosti na $\alpha_2$ a $\alpha_3$ souřadnice $w_2$ taktéž obsahuje relativní informaci o $x_1$; **speciální případ:** $\alpha_2 = \alpha_3 = \frac{1}{2}$:

$$w_1 = z_1 = \sqrt{\frac{2}{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}} \quad \text{and} \quad w_2 = z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}$$

# Volba vah

- variační(!) matice je základní charakteristikou kompoziční variability pro $\mathbf{x} = (x_1, \ldots, x_D)'$:

$$\mathbf{T} = \left\{ \mathrm{var}\left( \ln \frac{x_i}{x_j} \right) \right\}_{i,j=1}^{D}$$

- předpokládejme, že složka, která nás zajímá, je $x_1$, tedy uvažujeme první řádek (sloupec) matice $\mathbf{T}$:

$$\mathbf{t}_1 = (t_{11}, \ldots, t_{1D}) = \left( \mathrm{var}\left( \ln \frac{x_1}{x_1} \right), \mathrm{var}\left( \ln \frac{x_1}{x_2} \right), \ldots, \mathrm{var}\left( \ln \frac{x_1}{x_D} \right) \right)$$

- definujeme váhy: $\tilde{\alpha}_i = \dfrac{1}{t_{1i}^2}$, pro $i = 2, \ldots, D$, které přiřazují převrácené čtvercové hodnoty rozptylů párových log-podílů s $x_1$ (dále normované pro získání $\alpha_2, \ldots, \alpha_D$)

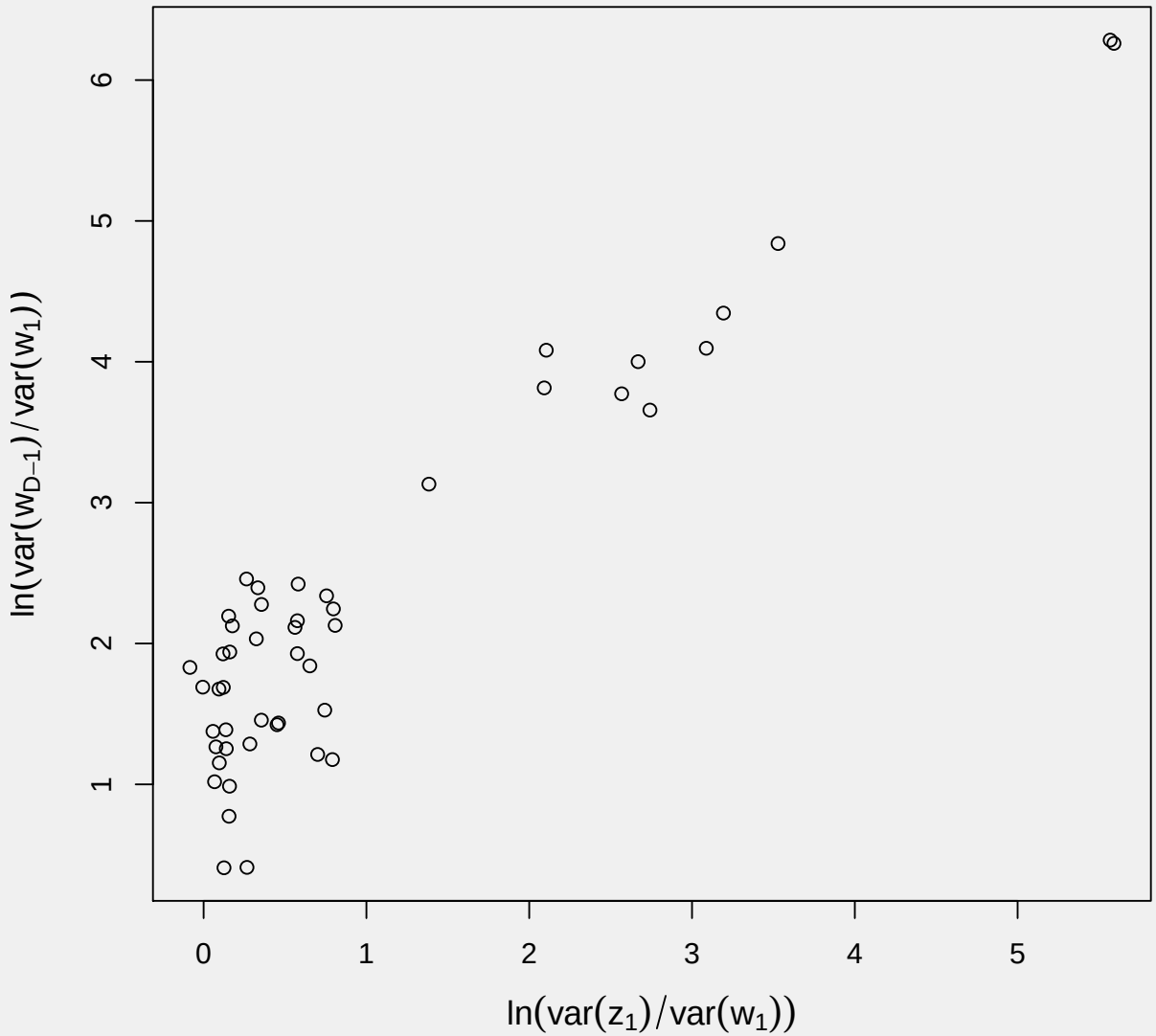
# Koncentrace prvků v sedimentech (National Geochemical Survey of Australia)

- projekt National Geochemical Survey of Australia (NGSA) shromáždil 1315 vzorků pokrývajících přes 80 % Austrálie (Caritat a Cooper, 2011)
- výsledek rozsáhlého geochemického výzkumu pokrývajícího přes 6 million $km^2$, provedeného Geoscience Australia a geologickými službami jednotlivých států Austrálie
- pro analýzu bylo vybráno 49 proměnných (koncentrací chemických prvků)

# NGSA data: výsledky

- simulační studie (se všemi složkami): ukázalo se, že vážené bilance jsou efektivním prostředkem k zachycení jednorozměrné (relativní) informace v geochemických mapách díky potlačení „redundantních“ log-podílů skrze vhodnou volbu vah

# NSGA data: simulace



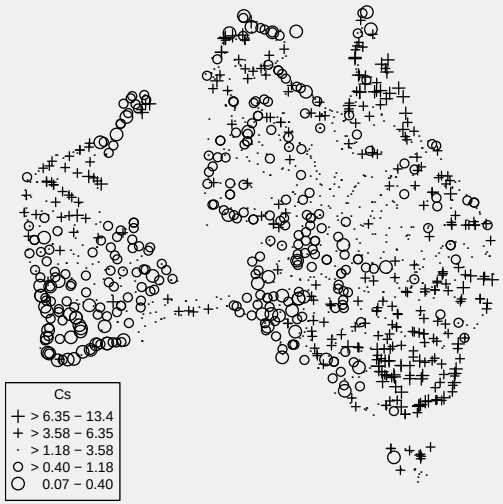
# NGSA data: výsledky

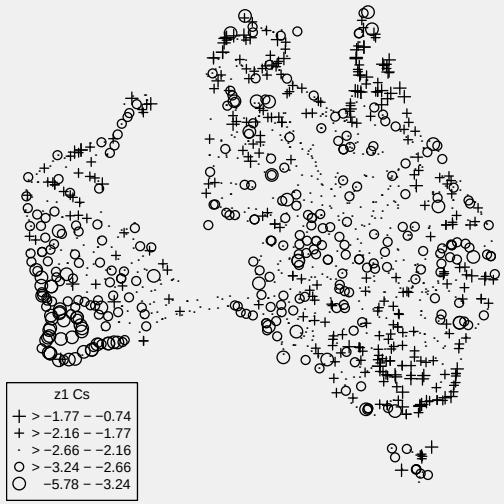
- simulační studie (se všemi složkami): ukázalo se, že vážené bilance jsou efektivním prostředkem k zachycení jednorozměrné (relativní) informace v geochemických mapách díky potlačení „redundantních“ log-podílů skrze vhodnou volbu vah

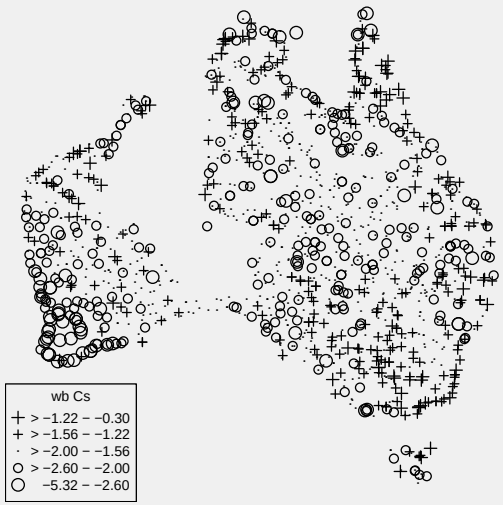
- eliminace tzv. **pobřežního efektu** (v oblastech jeho výskytu), zejména pro stopové prvky

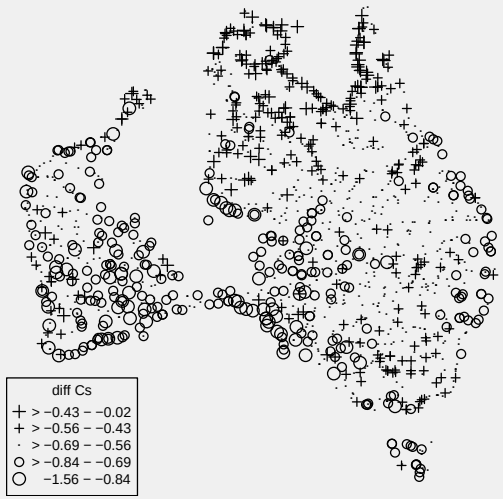
→ demonstrováno pro případ cesia (Cs)

# NSGA data: srovnání pro Cs



# Závěry

- vážené pivotové bilance představují alternativu k „automatické“ konstrukci pivotových bilancí

# Závěry

- vážené pivotové bilance představují alternativu k „automatické“ konstrukci pivotových bilancí
- získané ortonormální souřadnice mohou být následně využity pro statistickou analýzu (korelační analýza v souřadnicích, regresní analýza)

# Závěry

- vážené pivotové bilance představují alternativu k „automatické“ konstrukci pivotových bilancí
- získané ortonormální souřadnice mohou být následně využity pro statistickou analýzu (korelační analýza v souřadnicích, regresní analýza)
- $w_1$ a $w_{D-1}$ jsou konstruovány tak, že obsahují veškerou relativní informaci o vybrané složce; $w_1$ obsahuje „relevantní“ informaci, $w_{D-1}$ obsahuje zbývající informaci

# Závěry

- vážené pivotové bilance představují alternativu k „automatické“ konstrukci pivotových bilancí
- získané ortonormální souřadnice mohou být následně využity pro statistickou analýzu (korelační analýza v souřadnicích, regresní analýza)
- $w_1$ a $w_{D-1}$ jsou konstruovány tak, že obsahují veškerou relativní informaci o vybrané složce; $w_1$ obsahuje „relevantní“ informaci, $w_{D-1}$ obsahuje zbývající informaci
- váhy mohou být určeny pomocí variační matice, nebo na základě expertní znalosti (např. kvality měření)

# Literatura



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


Caritat, P. de, Cooper, M.: *National Geochemical Survey of Australia: The Geochemical Atlas of Australia*. Geoscience Australia Record, 2011/20 (2 Volumes)


Eaton, M.L.: *Multivariate statistics: A vector space approach*. Wiley, New York, 1983.


Egozcue, J.J., Pawlowsky-Glahn, V.: *Groups of parts and their balances in compositional data analysis*. Mathematical Geology 37, 795-828, 2005.


Filzmoser, P., Hron, K.: *Robust coordinates for compositional data using weighted balances*. In Nordhausen, K., Taskinen, S., editors, *Modern nonparametric, robust and multivariate methods*. Springer, Heidelberg, 2015.


Fišerová, E., Hron, K.: *On interpretation of orthonormal coordinates for compositional data*. Mathematical Geosciences 43, 455-468, 2011.


Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: *Modeling and analysis of compositional data*. Wiley, Chichester, 2015.


Prezentace je k dispozici na http://compositions.sweb.cz/.