Předzpracování kompozičních dat

Karel Hron

Katedra matematické analýzy a aplikací matematiky,
Přírodovědecká fakulta Univerzity Palackého, Olomouc

Robust 2012, Němčičky, 10. září 2012

Úvod

Poděkování: P. Filzmoser (*TU Wien*), M. Templ (*TU Wien*), J.A. Martín-Fernández (*University of Girona*), E. Fišerová (*UP Olomouc*)

1. Kompoziční data a jejich geometrie
2. Imputační metody pro kompozice
3. Nahrazení nulových hodnot v kompozičních datech

Kompoziční data

= *D-rozměrné vektory, jejichž složky kvantitativně popisují části nějakého celku, nesoucí výhradně relativní informaci*
(Pawlowsky-Glahn a Buccianti, 2011)

Kompoziční data

= *D-rozměrné vektory, jejichž složky kvantitativně popisují části nějakého celku, nesoucí výhradně relativní informaci* (Pawlowsky-Glahn a Buccianti, 2011)

- **obvyklé jednotky měření**: procenta, mg/kg (*konstantní součet složek*), mg na litr (*součet není konstantní*)

Kompoziční data

= *D-rozměrné vektory, jejichž složky kvantitativně popisují části nějakého celku, nesoucí výhradně relativní informaci* (Pawlowsky-Glahn a Buccianti, 2011)

- **obvyklé jednotky měření**: procenta, mg/kg (*konstantní součet složek*), mg na litr (*součet není konstantní*)
- **příklady**: geochemická data - proporcionální zastoupení minerálů v hornině; koncentrace fenolických kyselin ve víně (mg/l); výdaje domácností na konečnou spotřebu (jídlo, ubytování, ošacení) a další

Kompoziční data

= *D-rozměrné vektory, jejichž složky kvantitativně popisují části nějakého celku, nesoucí výhradně relativní informaci* (Pawlowsky-Glahn a Buccianti, 2011)

- **obvyklé jednotky měření**: procenta, mg/kg (*konstantní součet složek*), mg na litr (*součet není konstantní*)
- **příklady**: geochemická data - proporcionální zastoupení minerálů v hornině; koncentrace fenolických kyselin ve víně (mg/l); výdaje domácností na konečnou spotřebu (jídlo, ubytování, ošacení) a další
- problém zpracování dat s ***konstantním součtem*** byl v minulosti řešen pomocí **standardních statistických metod**, tedy za předpokladu **euklidovské geometrie v reálném prostoru**
- konstantní součet složek (1, 100) = *vhodná reprezentace kompozic*

Geometrické aspekty analýzy kompozičních dat

$\Rightarrow$ simplex = *výběrový prostor reprezentací kompozic*

- předpoklady relevantní analýzy kompozic: *invariantnost na změnu škály*, *podkompoziční soudržnost*, *respektování relativní škály* $\Rightarrow$ **Aitchisonova geometrie** (EVP dimenze $D-1$)
- drtivá většina statistických metod konstruována za předpokladu euklidovské geometrie (Eaton, 1983)

Geometrické aspekty analýzy kompozičních dat

$\Rightarrow$ simplex = *výběrový prostor reprezentací kompozic*

- předpoklady relevantní analýzy kompozic: *invariantnost na změnu škály*, *podkompoziční soudržnost*, *respektování relativní škály* $\Rightarrow$ **Aitchisonova geometrie** (EVP dimenze $D-1$)
- drtivá většina statistických metod konstruována za předpokladu euklidovské geometrie (Eaton, 1983)

$\Rightarrow$ vyjádřit kompoziční data v souřadnicích vzhledem k ortonormální bázi na simplexu $\rightarrow$ statistická analýza, interpretace (*bilance*)

- **log-ratio analýza** kompozičních dat (Aitchison, 1986; Egozcue a kol., 2003) - *alr*, *clr*, *ilr transformace*

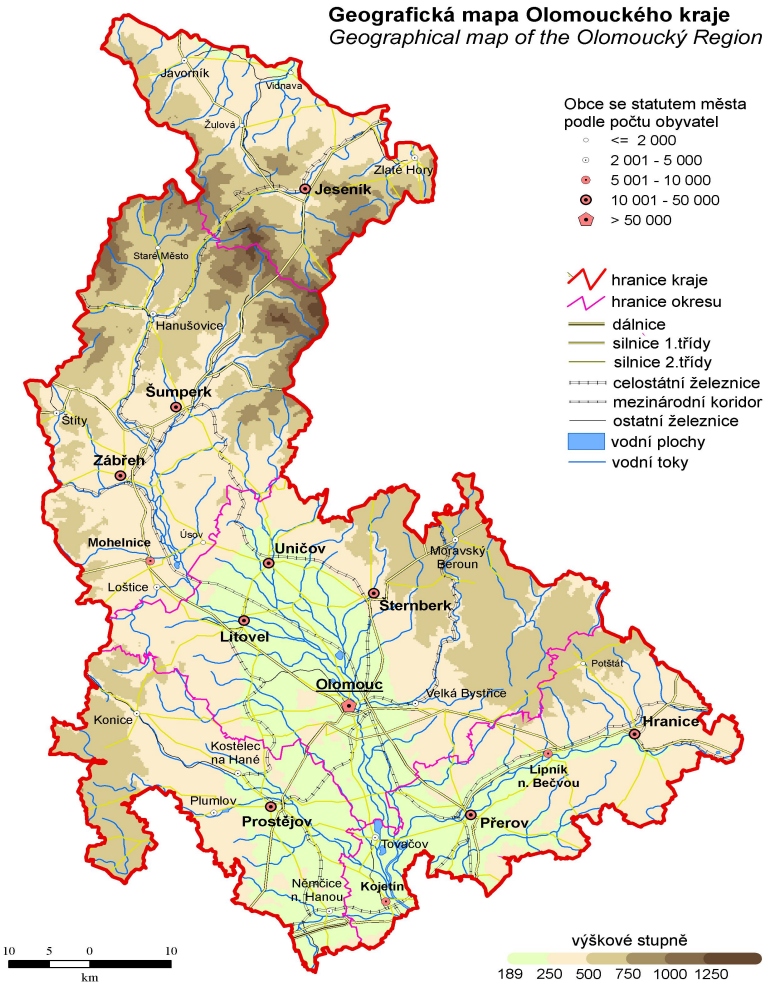
Příklad: Vegetační struktura v 11 bioregionech (Olomoucký kraj, původní data v hektarech)

	x_1	x_2	x_3	bioregion
1	19.80	47.90	43.40	prostěj.
2	237.80	0.60	111.50	litovel.
3	3.00	1.60	4.90	svitav.
4	229.10	70.30	331.90	drahan.
5	649.70	31.70	3198.40	šumper.
6	1090.70	29.80	758.90	n.-jes.
7	159.30	0.50	3297.90	jesen.
8	562.30	0.40	302.20	vidnav.
9	1557.45	53.59	747.79	hranic.
10	59.70	0.80	13.30	podbes.
11	237.80	0.60	108.90	kojetín.

mapový zdroj: Statistická ročenka Olomouckého kraje

Příklad: Vegetační struktura v 11 bioregionech (Olomoucký kraj, původní data v hektarech)

	x_1	x_2	x_3	bioregion
1	19.80	47.90	43.40	prostěj.
2	237.80	0.60	111.50	litovel.
3	3.00	1.60	4.90	svitav.
4	229.10	70.30	331.90	drahan.
5	649.70	31.70	3198.40	šumper.
6	1090.70	29.80	758.90	n.-jes.
7	159.30	0.50	3297.90	jesen.
8	562.30	0.40	302.20	vidnav.
9	1557.45	53.59	747.79	hranic.
10	59.70	0.80	13.30	podbes.
11	237.80	0.60	108.90	kojetín.



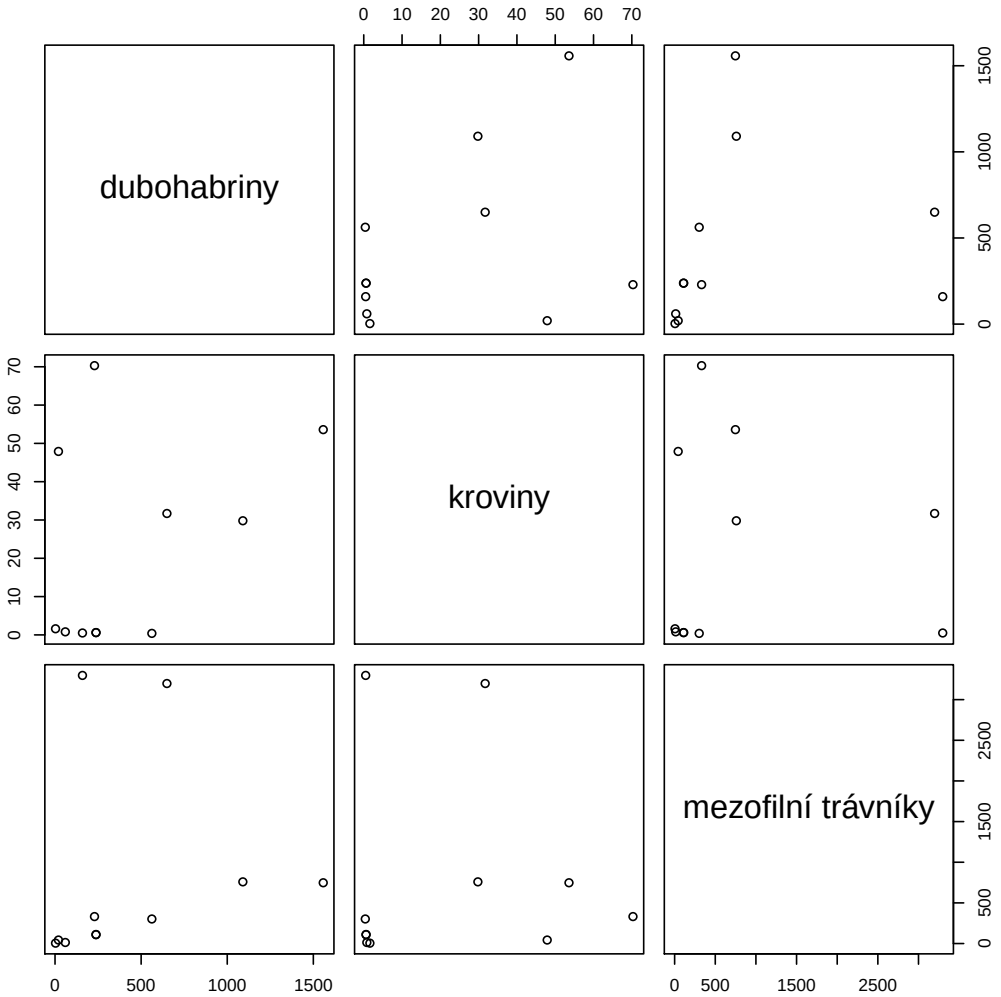
mapový zdroj: Statistická ročenka Olomouckého kraje

Příklad: Vegetační struktura v 11 bioregionech (Olomoucký kraj, původní data v hektarech)

	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	111.50
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	1090.70	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90

Příklad: Vegetační struktura v 11 bioregionech (Olomoucký kraj, původní data v hektarech)

	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	111.50
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	1090.70	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90

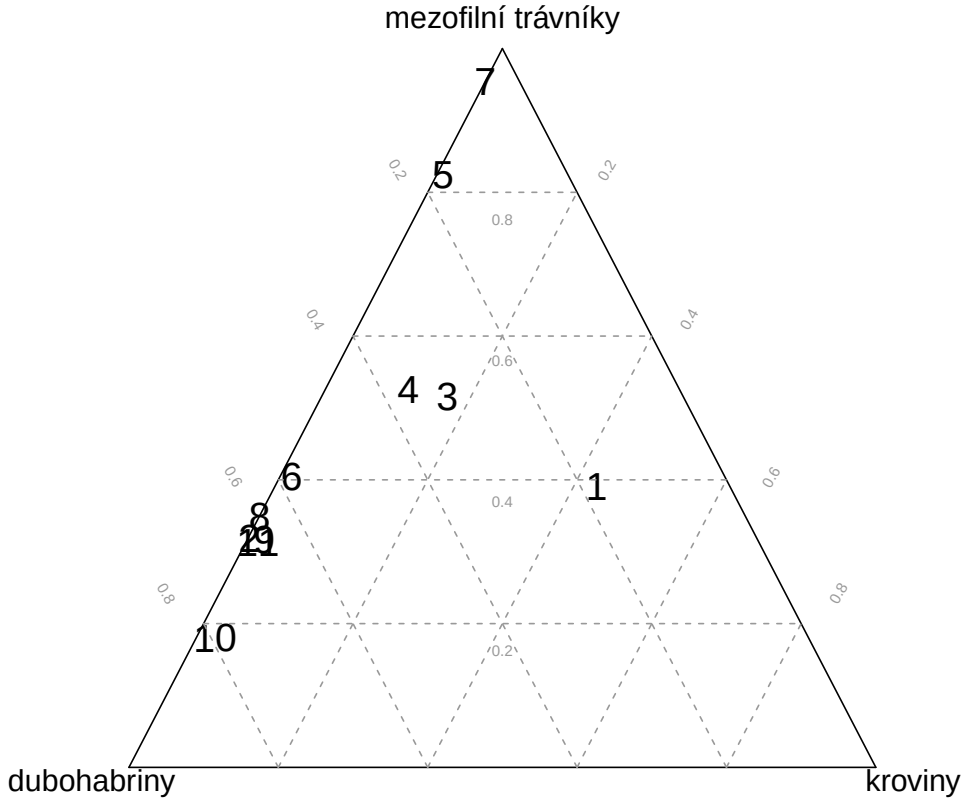


Příklad: Vegetační struktura v 11 bioregionech (v procentech)

	x_1	x_2	x_3
1	17.82	43.11	39.06
2	67.96	0.17	31.87
3	31.58	16.84	51.58
4	36.29	11.14	52.57
5	16.75	0.82	82.44
6	58.03	1.59	40.38
7	4.61	0.01	95.38
8	65.01	0.05	34.94
9	66.03	2.27	31.70
10	80.89	1.08	18.02
11	68.47	0.17	31.36

Příklad: Vegetační struktura v 11 bioregionech (v procentech)

	x_1	x_2	x_3
1	17.82	43.11	39.06
2	67.96	0.17	31.87
3	31.58	16.84	51.58
4	36.29	11.14	52.57
5	16.75	0.82	82.44
6	58.03	1.59	40.38
7	4.61	0.01	95.38
8	65.01	0.05	34.94
9	66.03	2.27	31.70
10	80.89	1.08	18.02
11	68.47	0.17	31.36



Ortonormální souřadnice a jejich interpretace

- D-složková kompozice $\mathbf{x} = (x_1, \ldots, x_D) \rightarrow$ souřadnice (bilance) $\mathbf{z} = (z_1, \ldots, z_{D-1})$
- **jediná ortonormální souřadnice** (z_1) **obsahuje všechnu relativní informaci** (vysvětluje všechny podíly) **o jedné složce**
- nechť x_1 je takovou složkou $\Rightarrow$

$$z_k = \sqrt{\frac{D-k}{D-k+1}} \ln \frac{x_k}{\sqrt[D-k]{\prod_{j=k+1}^{D} x_k}}, \quad k = 1, \ldots, D-1. \qquad (1)$$

s rozptyly

$$\mathrm{var}(z_k) = \frac{1}{D-k+1} \sum_{p=k+1}^{D} \mathrm{var}\left(\ln \frac{x_k}{x_p}\right) - \frac{1}{2(D-k)(D-k+1)} \sum_{p=k+1}^{D} \sum_{q=k+1}^{D} \mathrm{var}\left(\ln \frac{x_p}{x_q}\right)$$

Ortonormální souřadnice a jejich interpretace

$\Rightarrow$ konstrukce jiné ortonormální báze, kde **první souřadnice vysvětluje danou kompoziční složku** (x_1 v předchozím snímku):

$$(x_1, \ldots, x_D) \rightarrow$$

$$(x_I, x_1, \ldots, x_{I-1}, x_{I+1}, \ldots, x_D) =: (x_1^{(I)}, x_2^{(I)}, \ldots, x_I^{(I)}, x_{I+1}^{(I)}, \ldots, x_D^{(I)});$$

$$z_k^{(I)} = \sqrt{\frac{D-k}{D-k+1}} \ln \frac{\sqrt[D-k]{\prod_{j=k+1}^{D} x_j^{(I)}}}{x_k^{(I)}}, \quad k = 1, \ldots, D-1, \qquad (2)$$

zřejmě $z_k^{(1)} = z_k$ z (1) pro $k = 1, \ldots, D-1$

Ortonormální souřadnice a jejich interpretace

$\Rightarrow$ konstrukce jiné ortonormální báze, kde **první souřadnice vysvětluje danou kompoziční složku** (x_1 v předchozím snímku):

$$(x_1, \ldots, x_D) \rightarrow$$

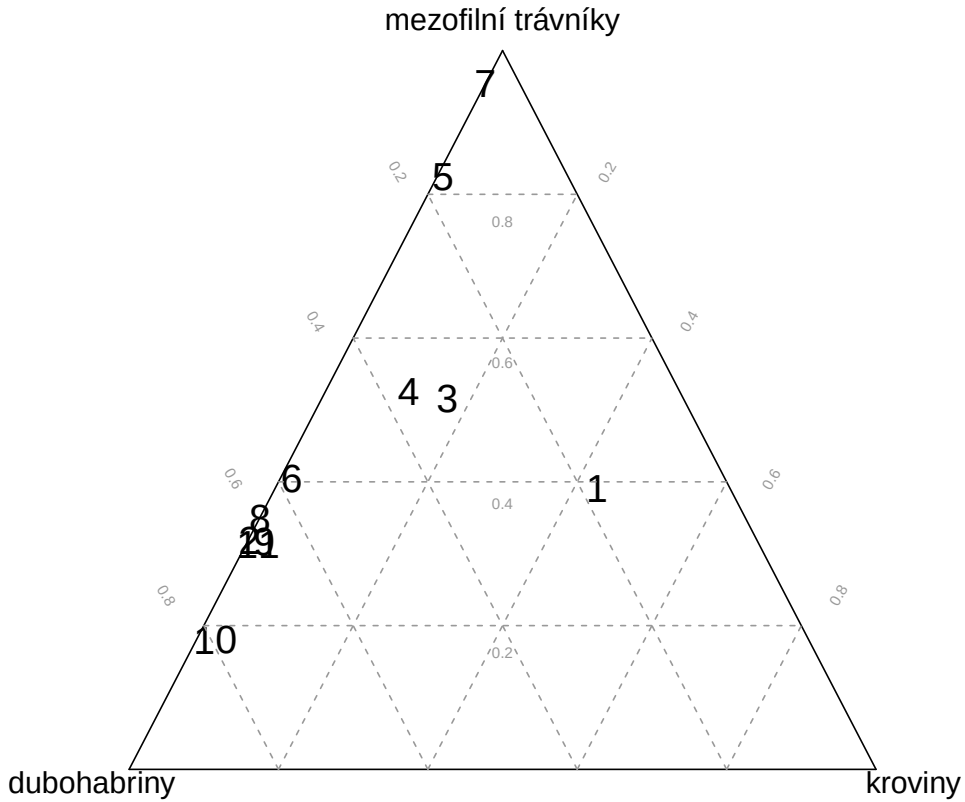
$$(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D) =: (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)});$$

$$z_k^{(l)} = \sqrt{\frac{D-k}{D-k+1}} \ln \frac{\sqrt[D-k]{\prod_{j=k+1}^{D} x_j^{(l)}}}{x_k^{(l)}}, \quad k = 1, \ldots, D-1, \qquad (2)$$

zřejmě $z_k^{(1)} = z_k$ z (1) pro $k = 1, \ldots, D-1$

- takové souřadnice jsou zejména užitečné, když **se zajímáme o jedinou kompoziční složku** (imputace chybějících hodnot)

Příklad: Vegetační struktura v 11 bioregionech (v ortonormálních souřadnicích)



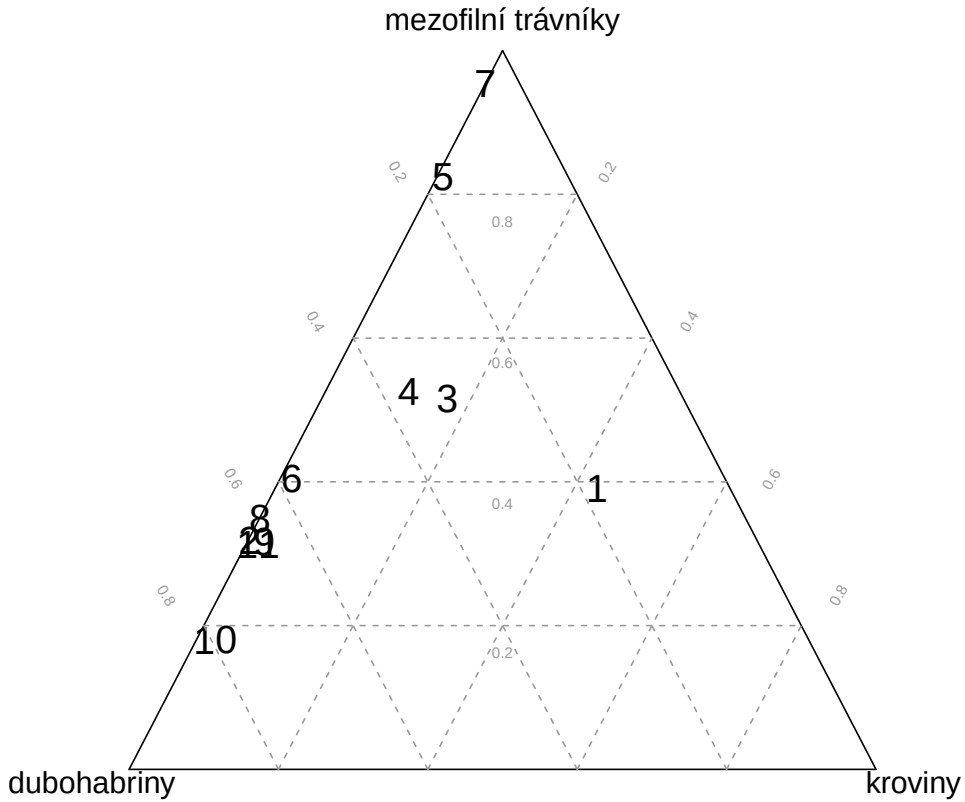
Příklad: Vegetační struktura v 11 bioregionech (v ortonormálních souřadnicích)

- obdržíme dvě souřadnice; např. zvolíme

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

$$z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}.$$

- odhalení skutečné struktury dat
- dvě přirozené odlehlé hodnoty (5, 7)

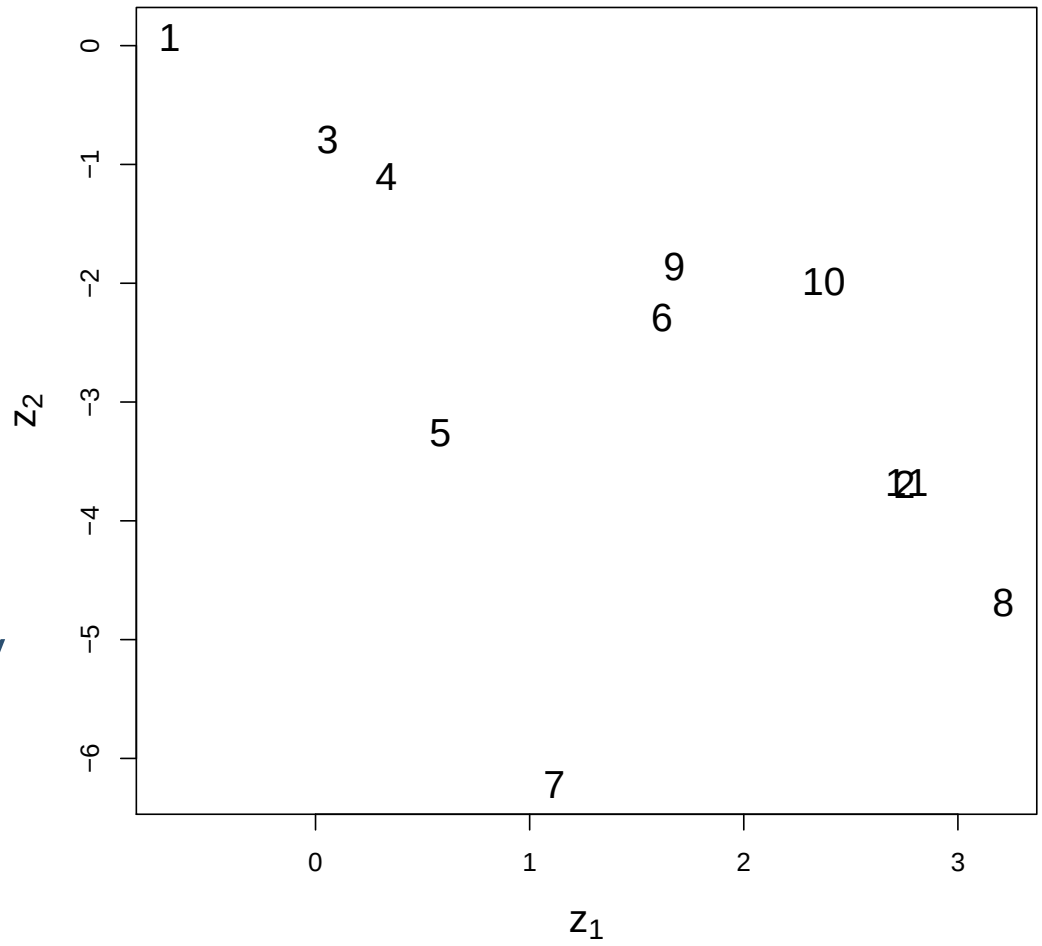


Příklad: Vegetační struktura v 11 bioregionech (v ortonormálních souřadnicích)

- obdržíme dvě souřadnice; např. zvolíme

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

- odhalení skutečné struktury dat
- dvě přirozené odlehlé hodnoty (5, 7)



Kompoziční data a chybějící informace

- chybějící hodnoty se vyskytují v mnoha reálných (kompozičních) datových souborech
- většinu statistických metod nelze aplikovat přímo na data s chybějícími hodnotami
- prosté vynechání pozorování s chybějícími hodnotami by způsobilo **ztrátu informace**

Kompoziční data a chybějící informace

- chybějící hodnoty se vyskytují v mnoha reálných (kompozičních) datových souborech
- většinu statistických metod nelze aplikovat přímo na data s chybějícími hodnotami
- prosté vynechání pozorování s chybějícími hodnotami by způsobilo **ztrátu informace**

$\Rightarrow$ **doplnit chybějící údaje vhodnými hodnotami** a pokračovat se ve statistické analýze

Kompoziční data a chybějící informace

- chybějící hodnoty se vyskytují v mnoha reálných (kompozičních) datových souborech
- většinu statistických metod nelze aplikovat přímo na data s chybějícími hodnotami
- prosté vynechání pozorování s chybějícími hodnotami by způsobilo **ztrátu informace**

$\Rightarrow$ **doplnit chybějící údaje vhodnými hodnotami** a pokračovat se ve statistické analýze

- potřeba speciálního přístupu k imputaci kompozičních dat (jediná relevantní informace je obsažena v podílech)
- při výskytu odlehlých hodnot selhávají klasické imputační metody $\Rightarrow$ **užití robustních metod**

Imputace pomocí k nejbližších sousedů

Když imputujeme jednu chybějící hodnotu

- užijeme **Aitchisonovu vzdálenost** $d_A(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{D} \sum_{j=1}^{D} (\ln(x_i/x_j) - \ln(y_i/x_j))^2$ k nalezení k nejbližších sousedů
- **upravíme** odpovídající hodnoty vzhledem k celkové velikosti složek

Imputace pomocí k nejbližších sousedů

Když imputujeme jednu chybějící hodnotu

- užijeme **Aitchisonovu vzdálenost**
  $d_A(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{D} \sum_{j=1}^{D} (\ln(x_i/x_j) - \ln(y_i/x_j))^2$ k nalezení k nejbližších sousedů
- **upravíme** odpovídající hodnoty vzhledem k celkové velikosti složek

nejbližší soused	80	30	50	100
hodnota k imputaci	20	NA	12.5	25

$\longrightarrow$ upravíme 30 vzhledem k informaci v ostatních složkách

Imputace pomocí k nejbližších sousedů

Když imputujeme jednu chybějící hodnotu

- užijeme **Aitchisonovu vzdálenost**
  $d_A(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{D} \sum_{j=1}^{D} (\ln(x_i/x_j) - \ln(y_i/x_j))^2$ k nalezení k nejbližších sousedů
- **upravíme** odpovídající hodnoty vzhledem k celkové velikosti složek

nejbližší soused	80	30	50	100
hodnota k imputaci	20	NA	12.5	25

  $\longrightarrow$ upravíme 30 vzhledem k informaci v ostatních složkách

imputace pomocí metody k nejbližších sousedů nebere plně do úvahy vztahy mezi kompozičními složkami
$\Rightarrow$ **modelová imputace**

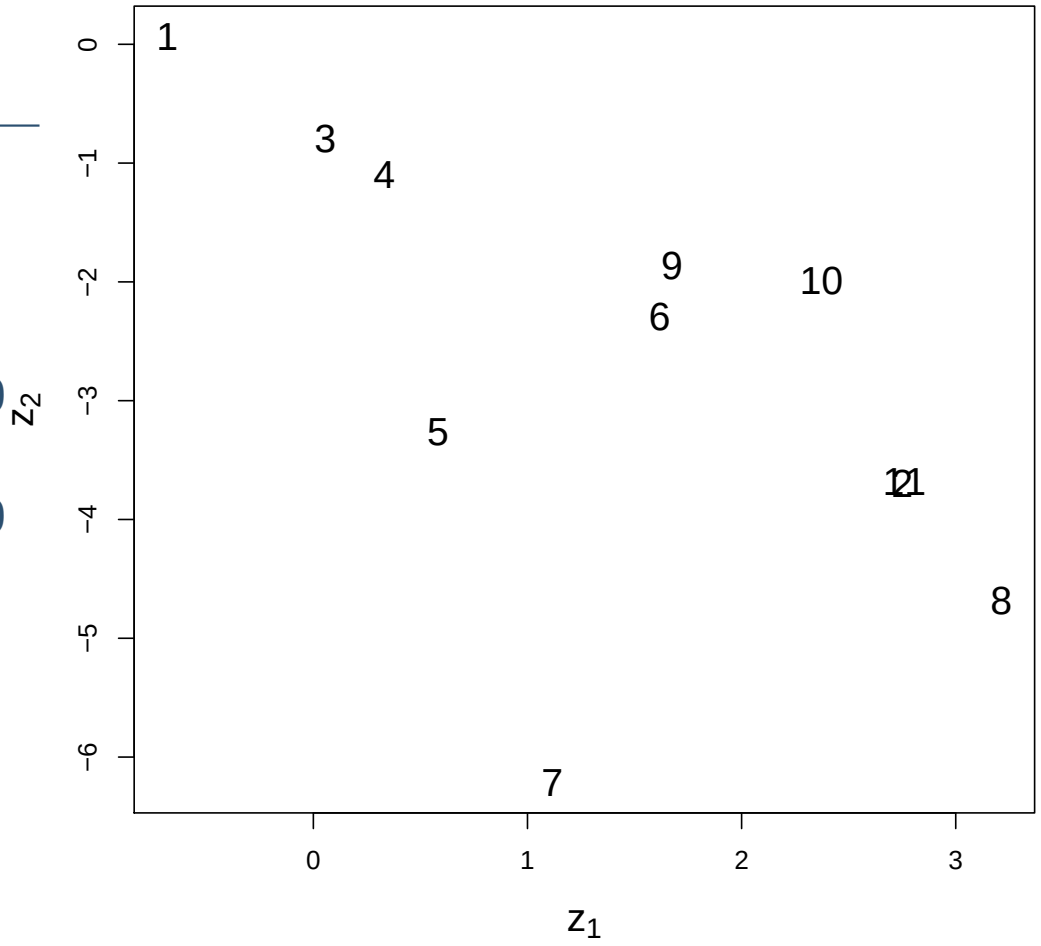
Iterativní modelová imputace

- inicializace chybějících hodnot výsledky metody k nejbližších sousedů
- uspořádat kompoziční složky podle klesajícího počtu jejich chybějících hodnot
- do dosažení konvergence: pro $l = 1, \ldots, D$
  - provést ilr transformaci (2) dat
  - provést robustní (klasickou) regresi $z_1^{(l)}$ na $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$
  - nahradit (původně) chybějící hodnoty a zpětná transformace dat
- přeuspořádat složky dle původního pořadí

$\rightarrow$ **obvykle zlepšuje výsledky metody k nejbližších sousedů**

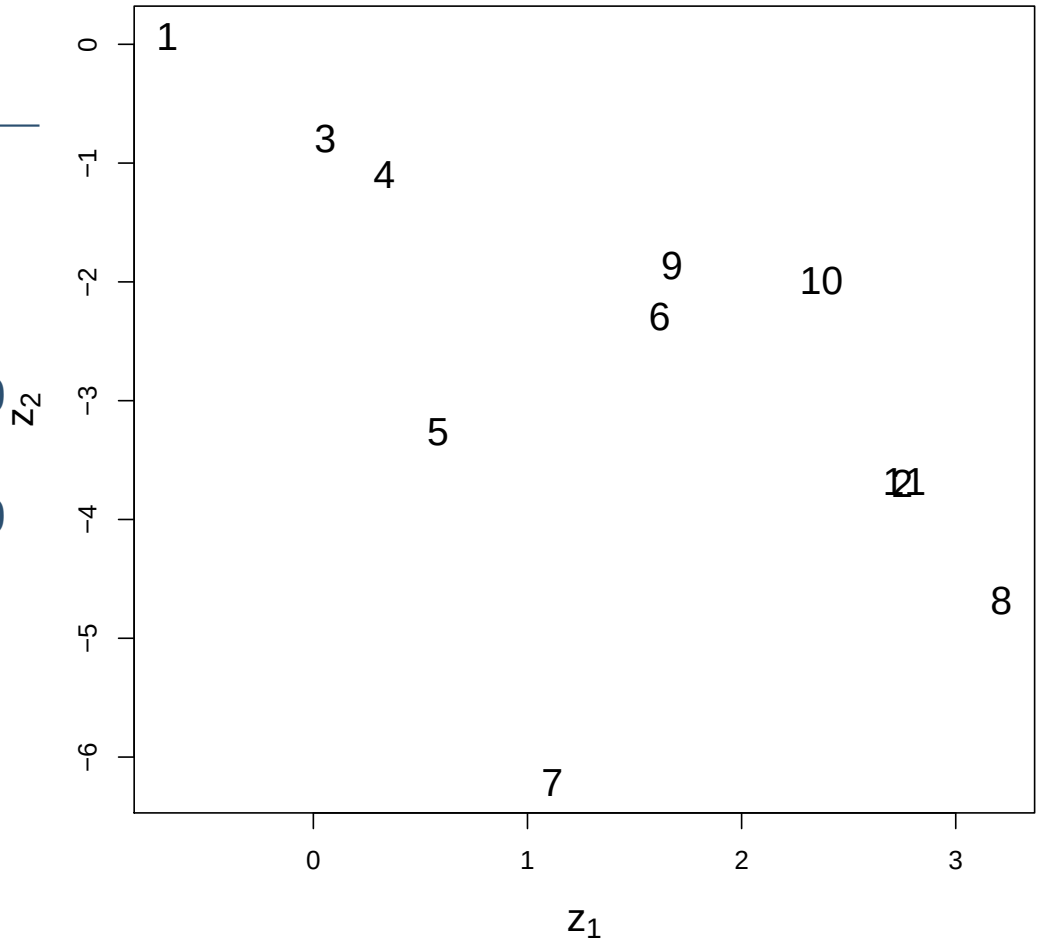
Příklad: Vegetační struktura v 11 bioregionech

	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	111.50
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	1090.70	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90



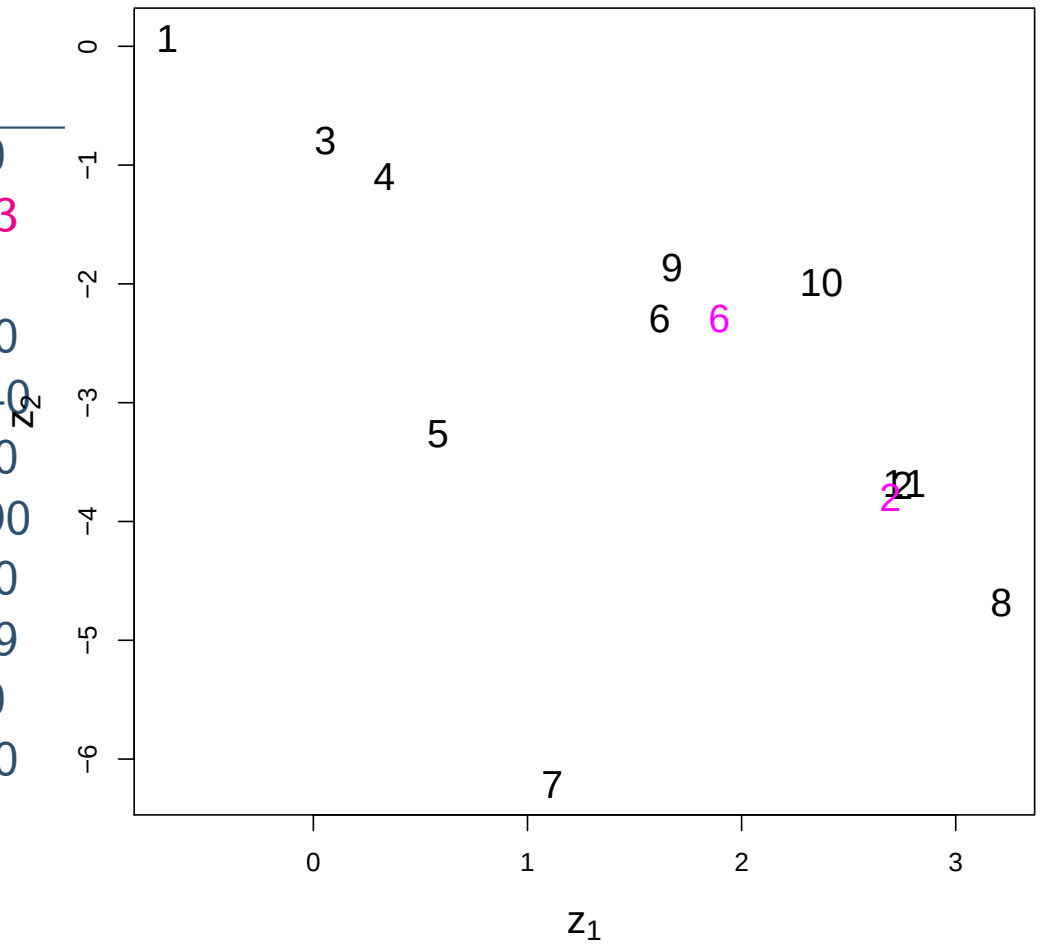
Příklad: Vegetační struktura v 11 bioregionech (NA)

	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	NA
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	NA	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90



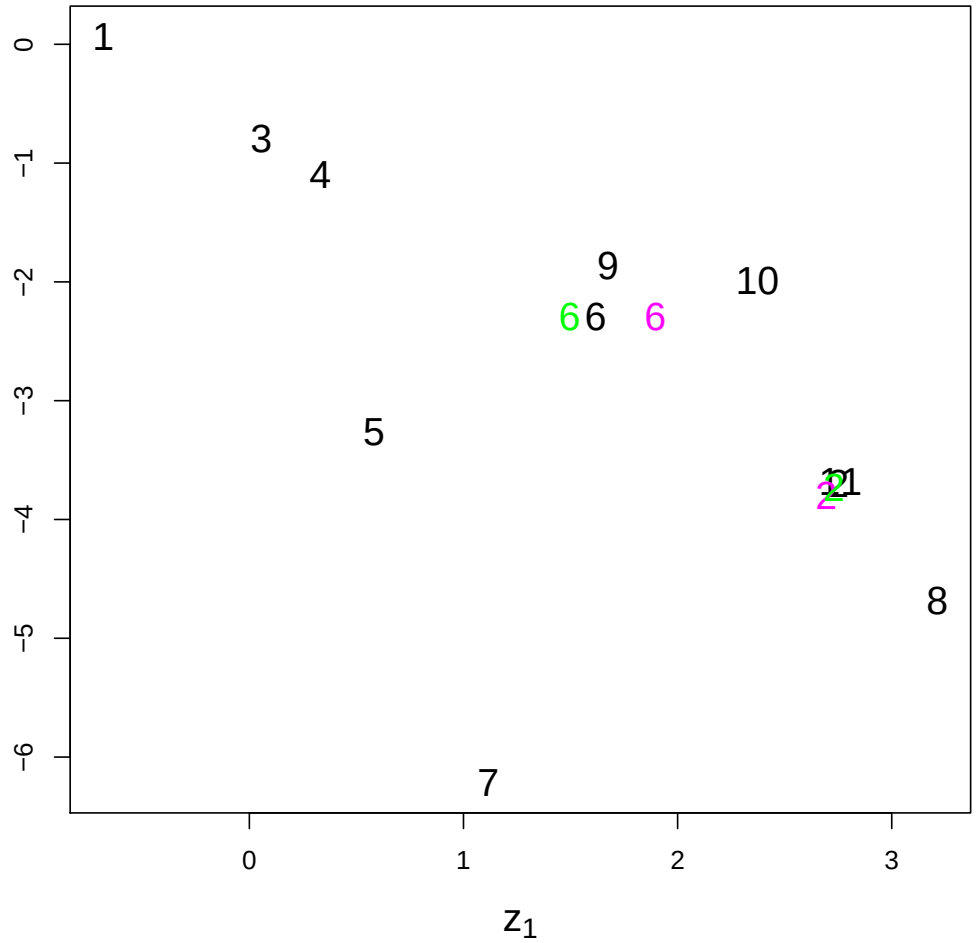
Příklad: imputace pomocí k nejbližších sousedů

	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	128.03
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	1532.807	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90



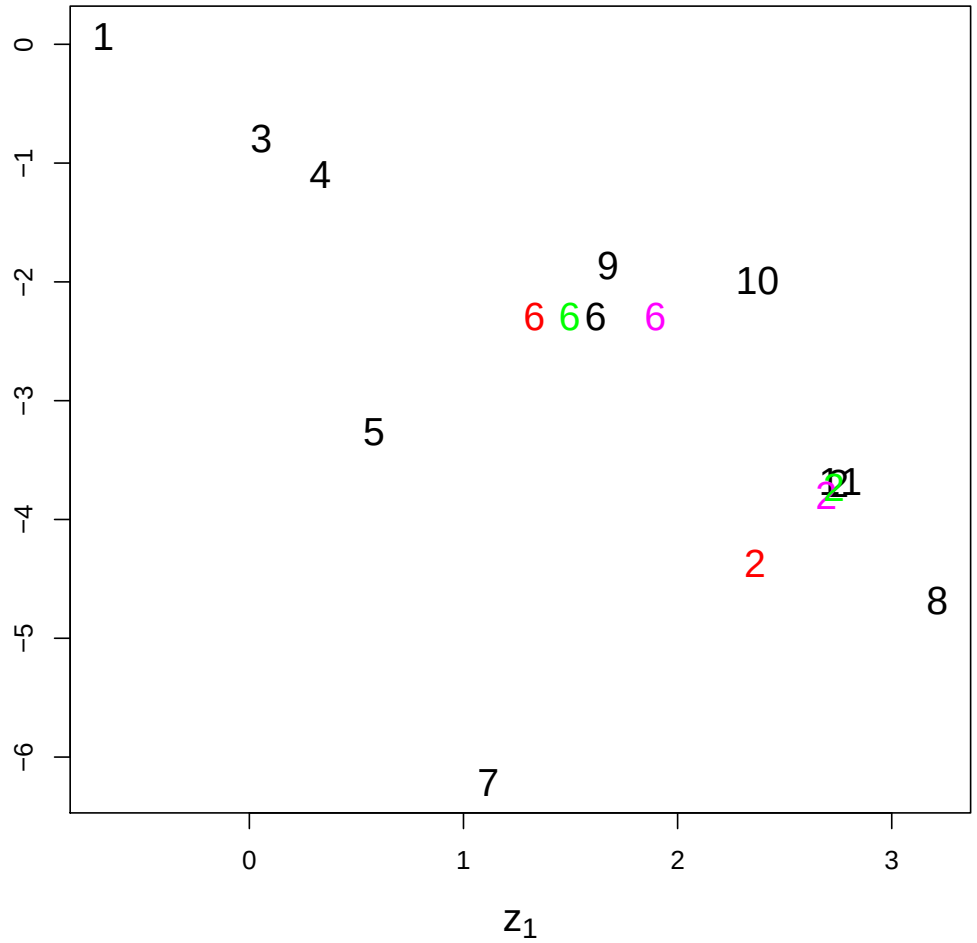
Příklad: imputace pomocí robustního iterativního postupu

	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	117.55
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	940.54	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90

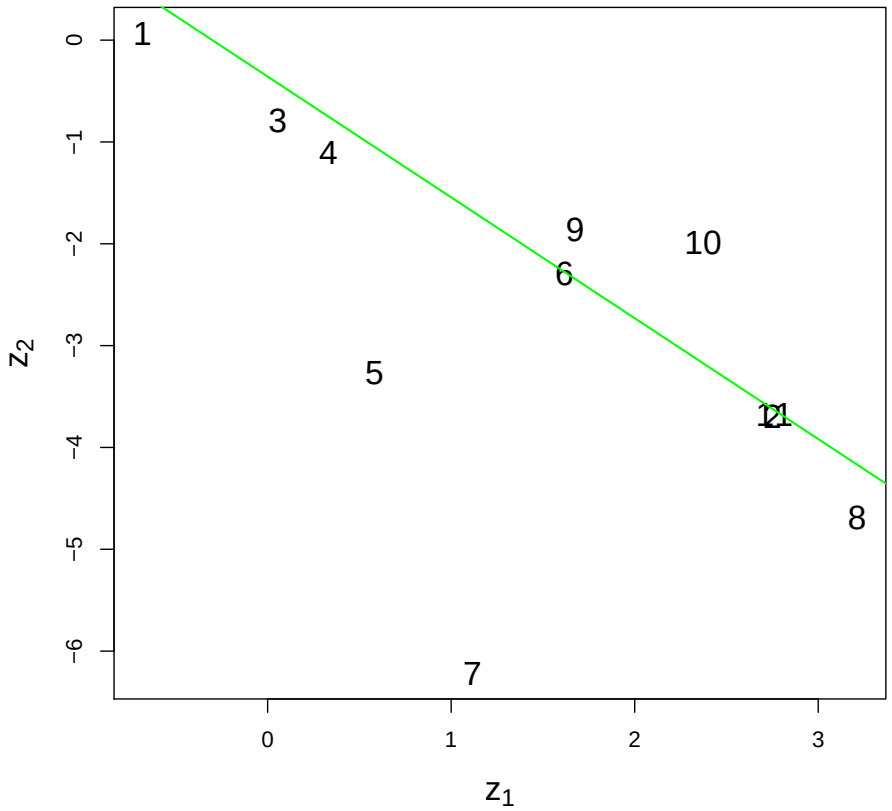


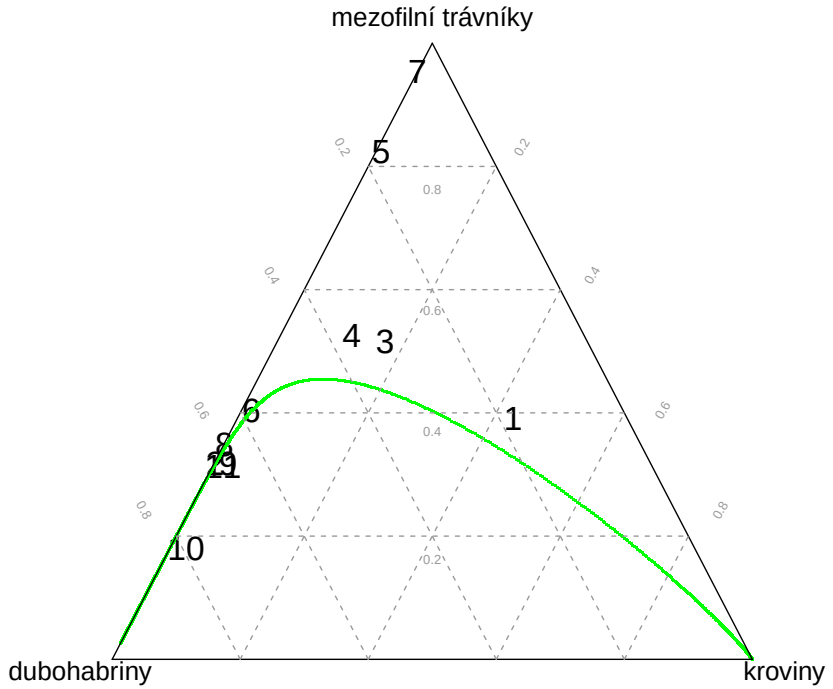
Příklad: Imputace pomocí klasického iterativního postupu

	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	289.99
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	767.96	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90

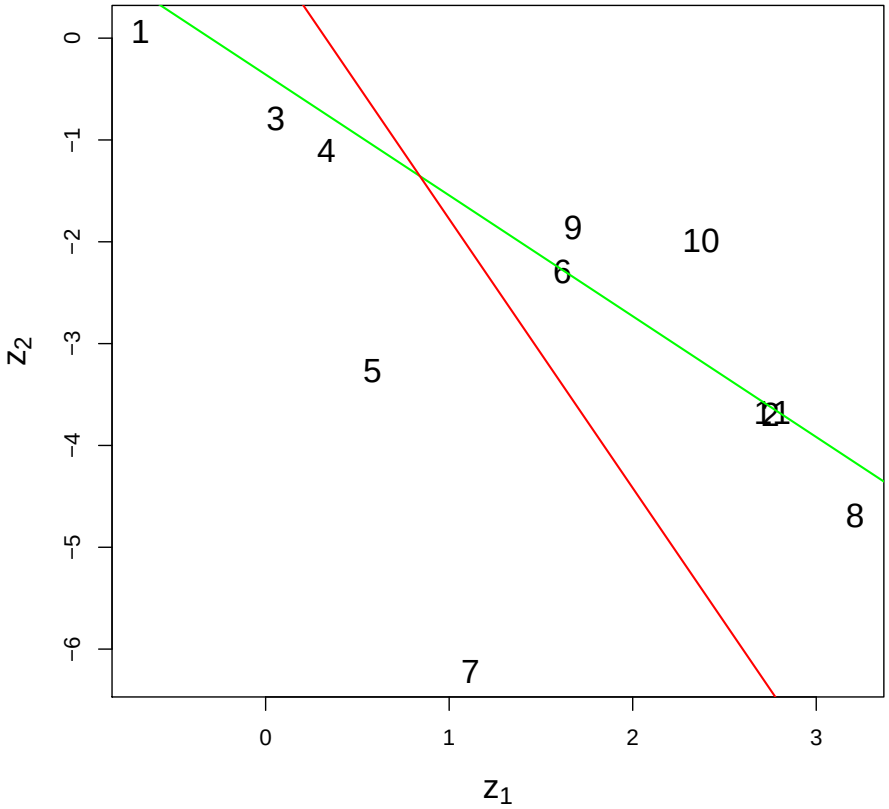


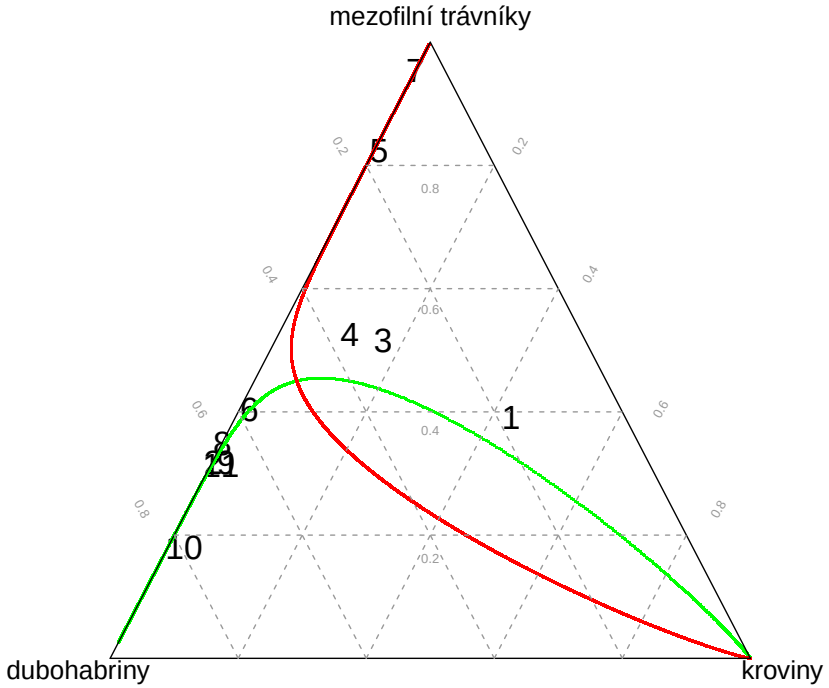
Příklad: Robustní (LTS) regrese z_1 na z_2





Příklad: Problém klasické regrese





Výskyt nul v kompozičních datech

- **log-ratio analýza** kompozičních dat **nemůže být použita pro kompozice s nulovými hodnotami složek**
- **přirozený požadavek**, protože u kompozice je veškerá relevantní informace obsažena v *podílech* mezi kompozičními složkami

Výskyt nul v kompozičních datech

- **log-ratio analýza** kompozičních dat **nemůže být použita pro kompozice s nulovými hodnotami složek**
- **přirozený požadavek**, protože u kompozice je veškerá relevantní informace obsažena v *podílech* mezi kompozičními složkami
- dva základní typy nul:
  - strukturní nuly - výsledek nějakého strukturního procesu (výdaje za alkohol v domácnostech abstinentů), **řešení**: *několik přístupů, např. analýza podkompozic, další výzkum nutný*
  - **nuly vzniklé zaokrouhlením** - výsledek nepřesného měření stopových prvků v kompozici (geochemická měření, environmentální data), **řešení**: nahradit malou nenulovou hodnotou (např. 2/3 detekčního limitu nebo užít nějaký **parametrický přístup**)

Modelové nahrazení zaokroulovacích nul

- algoritmus modelové imutace doplněný o tobitovou regresi (inicializace = 2/3 detekčního limitu)
- pro $l = 1, \ldots, D$, $i = 1, \ldots, n$, $\mathbf{Z}^{(l)} = [\mathbf{z}_1^{(l)}, \mathbf{Z}_{-1}^{(l)}]$:

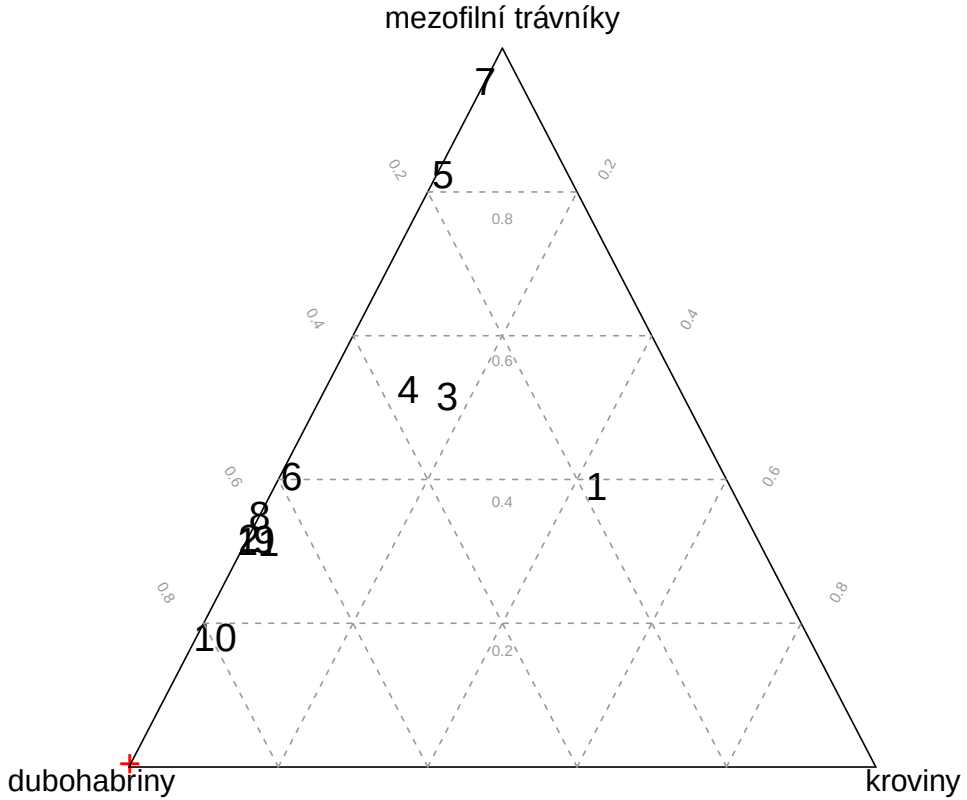
$$\psi_{i1}^{(l)} = \sqrt{\frac{D-1}{D}} \ln \frac{e_{i1}^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}^{(l)}}} \text{ souř. pro detekční limit } e_{i1}^{(l)} \equiv e_{il};$$

$$\hat{z}_{i1}^{(l)} = \mathbf{z}^{(l)t}_{i,-1} \cdot \hat{\boldsymbol{\beta}}^{(l)} - \hat{\sigma}^{(l)} \frac{\phi\left(\frac{\psi_{i1}^{(l)} - \mathbf{z}^{(l)t}_{i,-1} \cdot \hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}}\right)}{\Phi\left(\frac{\psi_{i1}^{(l)} - \mathbf{z}^{(l)t}_{i,-1} \cdot \hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}}\right)}, \tag{3}$$

náhrada neznámých hodnot v $\mathbf{z}_1^{(l)}$ podmíněnou střední hodnotou $\mathrm{E}[\mathbf{z}_1^{(l)} | \mathbf{Z}_{-1}^{(l)}, \mathbf{z}_1^{(l)} < \boldsymbol{\psi}_1^{(l)}]$

Příklad: Výskyt nulové hodnoty (DL=0.1)

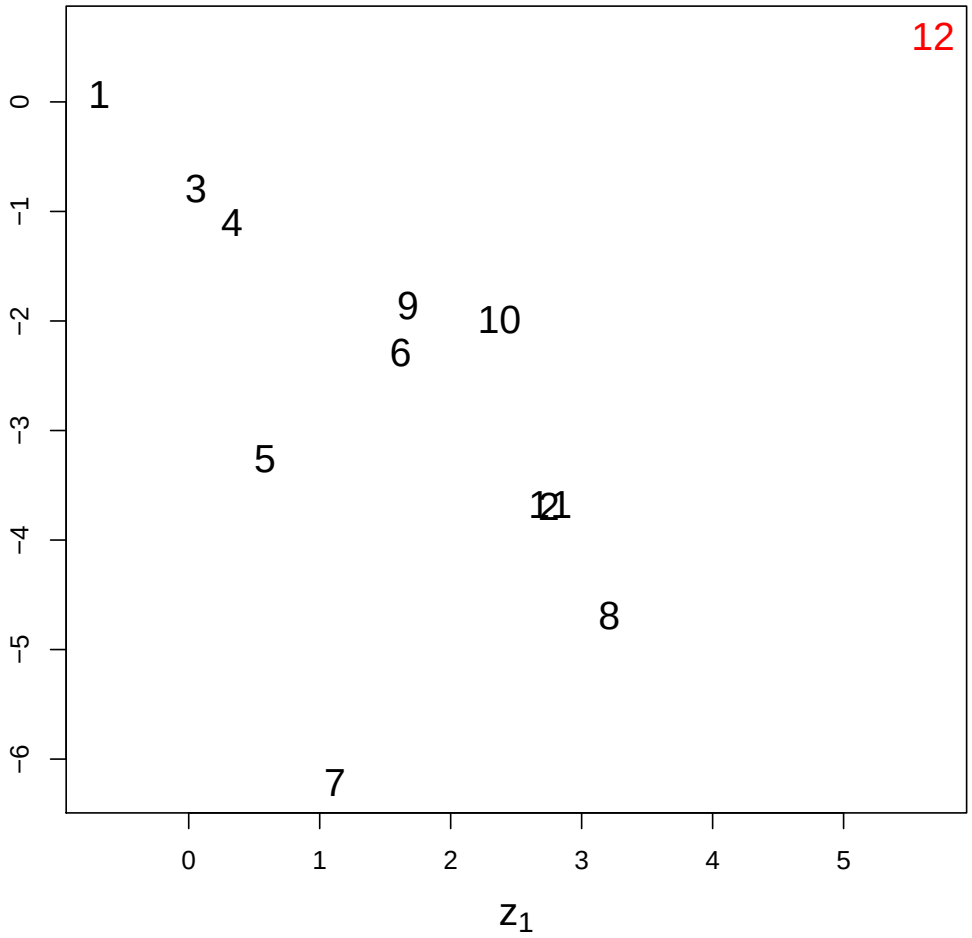
	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	111.50
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	1090.70	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90
12	69.00	0.10	0.00



ždánicko-litenčický bioregion

Příklad: Nahrazení nulové hodnoty (DL=0.1)

	x_1	x_2	x_3
1	19.80	47.90	43.40
2	237.80	0.60	111.50
3	3.00	1.60	4.90
4	229.10	70.30	331.90
5	649.70	31.70	3198.40
6	1090.70	29.80	758.90
7	159.30	0.50	3297.90
8	562.30	0.40	302.20
9	1557.45	53.59	747.79
10	59.70	0.80	13.30
11	237.80	0.60	108.90
12	69.00	0.10	**0.0427**



Závěr

- kompozice vyžadují specifický přístup k předzpracování datového souboru

Závěr

- kompozice vyžadují specifický přístup k předzpracování datového souboru
- **představené metody překonávají existující metody ve většině datových konfigurací** (např. u imputace NA testováno s více než 20 imputačními technikami při různých mechanismech (MCAR, MAR))

Závěr

- kompozice vyžadují specifický přístup k předzpracování datového souboru
- **představené metody překonávají existující metody ve většině datových konfigurací** (např. u imputace NA testováno s více než 20 imputačními technikami při různých mechanismech (MCAR, MAR))
- k dispozici v R-knihovně `robCompositions`; tato prezentace na `http://compositions.sweb.cz/`

Závěr

- kompozice vyžadují specifický přístup k předzpracování datového souboru
- **představené metody překonávají existující metody ve většině datových konfigurací** (např. u imputace NA testováno s více než 20 imputačními technikami při různých mechanismech (MCAR, MAR))
- k dispozici v R-knihovně `robCompositions`; tato prezentace na `http://compositions.sweb.cz/`
- **v přípravě: algoritmus pro práci se strukturními nulami**

Literatura (kompoziční data)

- Aitchison, J. (1986) The statistical analysis of compositional data. Chapman and Hall, London.
- Eaton, M. (1983) Multivariate statistics. A vector space approach. John Wiley and Sons, New York.
- Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C. (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35 (3), 279–300.
- **Pawlowsky-Glahn, V., Buccianti, A., eds. (2011) Compositional data analysis: Theory and applications. Wiley, Chichester.**

Literatura (téma prezentace)

- Fišerová, E., Hron, K. (2011) On interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences, 43 (4), 455-468.
- **Hron, K., Templ, M., Filzmoser, P. (2010) Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis, 54 (12), 3095-3107**
- **Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J. (2012) Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. Computational Statistics and Data Analysis, 56 (9), 2688-2704**