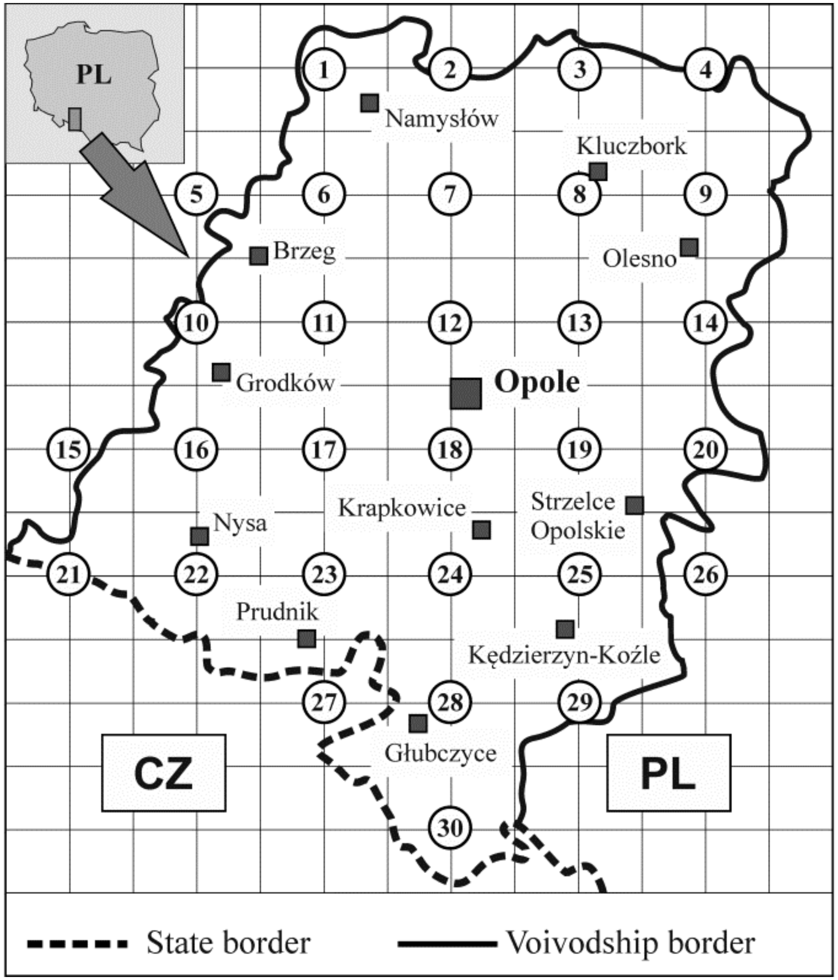
Z. Ziembik[1], A. Dołhańczuk-Śródka[1], A. Kłos[1], Yu. A. Aleksiayenak[2], M. Rajfur[1], M. Wacławek[1], M.V Frontasyeva[2]

[1]Opole University, Opole, Poland
[2]FLNP JINR, Dubna, Russia

Spatial patterns of element distributions in moss in the area of the Opole province (Poland)

Sampling sites



The neutron activation analysis procedure was used to determine the concentration of 42 elements: Na, Mg, Al, Cl, K, Ca, Sc, V, Cr, Mn, Fe, Ni, Co, Zn, As, Se, Br, Rb, Sr, Zr, Nb, Mo, I, Ag, Cd, Sb, Ba, Cs, La, Ce, Nd, Sm, Eu, Tb, Yb, Hf, Ta, W, Au, Hg, Th, and U accumulated in mosses sampled for testing in September and October 2011 in the Opole Province (Southern Poland).

Kłos A., Aleksiayenak Yu.A., Ziembik Z., Rajfur M., Jerz D., Wacławek M., Frontasyeva M.V. (2013). The Use of Neutron Activation Analysis in the Biomonitoring of Trace Element Deposition in the Opole Province (Poland), Ecological Chemistry and Engineering S, 20(4), 677-687

The sample space

Sample space – „ ... a set of convenient symbols which can be identified through a one-to-one correspondence with the possible outcomes of the observational or experimental process”[1].

For compositional data the appropriate choice is

simplex.

[1] Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press).

Compositiona Data (CoDa)

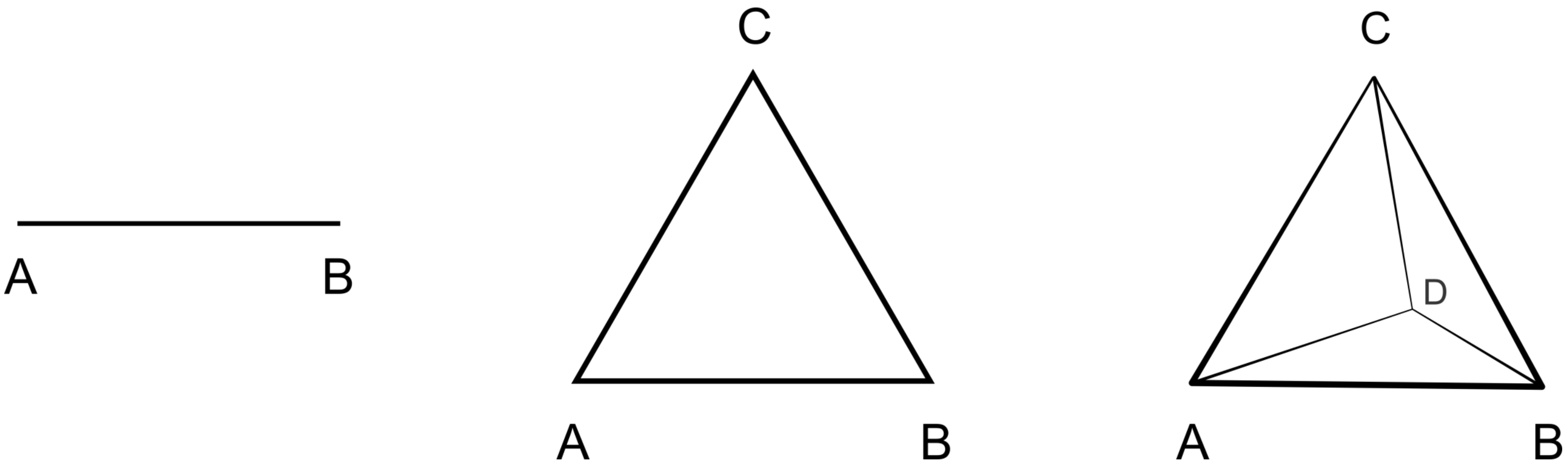
Compositiona data – any vector $\boldsymbol{x}$ with non-negative elements x_1, …, x_D representing proportions of some whole, subjected to the constraint:

$$x_1 + x_2 + \ldots + x_{D-1} + x_D = 1$$

The closure operation C applied on quantity $\mathbf{w}$:

$$\mathrm{C}(\mathbf{w}) = \left[\frac{w_1}{\sum_{j=1}^{D} w_j}, \frac{w_2}{\sum_{j=1}^{D} w_j}, \ldots, \frac{w_{D-1}}{\sum_{j=1}^{D} w_j}, \frac{w_D}{\sum_{j=1}^{D} w_j} \right] = \left[x_1, x_2, \ldots, x_{D-1}, x_D \right] = \mathbf{x}$$

Simplex sample space for the data elaboration.



Simplex spaces representing 2 components system (A, B – line segment), 3 components system (A, B, C – equilateral triangle) and 4 components system (A, B, C, D – tetrahedron).

Basic properties of the simplex space

Consider 2 compositional points, **A** and **B**.

The inner product

$$\left\langle \mathbf{x}_{\mathrm{A}}, \mathbf{x}_{\mathrm{B}} \right\rangle_{\mathrm{a}} = \frac{1}{D} \sum_{j<l} \ln \frac{x_{\mathrm{A}j}}{x_{\mathrm{A}l}} \ln \frac{x_{\mathrm{B}j}}{x_{\mathrm{B}l}} = \sum_{j=1}^{D} \ln \frac{x_{\mathrm{A}j}}{\mathrm{g}(\mathbf{x}_{\mathrm{A}})} \ln \frac{x_{\mathrm{B}j}}{\mathrm{g}(\mathbf{x}_{\mathrm{B}})}$$

where: $$\mathrm{g}(\mathbf{x}) = \sqrt[D]{\prod_{j=1}^{D} x_j}$$

The distance between **A** and **B**

$$d_{\mathrm{a}}(\mathbf{x}_{\mathrm{A}}, \mathbf{x}_{\mathrm{B}}) = \sqrt{\frac{1}{D} \sum_{j<l} \left(\ln \frac{x_{\mathrm{A}j}}{x_{\mathrm{A}l}} - \ln \frac{x_{\mathrm{B}j}}{x_{\mathrm{B}l}} \right)^2} = \sqrt{\sum_{j=1}^{D} \left(\ln \frac{x_{\mathrm{A}j}}{\mathrm{g}(\mathbf{x}_{\mathrm{A}})} - \ln \frac{x_{\mathrm{B}j}}{\mathrm{g}(\mathbf{x}_{\mathrm{B}})} \right)^2}$$

From simplex to Euclidean space

The centered logratio clr transformation

$$\mathrm{clr}(\mathbf{x}) = \left[\frac{x_1}{\mathrm{g}(\mathbf{x})}, \ldots, \frac{x_D}{\mathrm{g}(\mathbf{x})} \right]$$

$$d_{\mathrm{a}}(\mathbf{x}_{\mathrm{A}}, \mathbf{x}_{\mathrm{B}}) = \sqrt{\sum_{j=1}^{D} \left(\mathrm{clr}_j x_{\mathrm{A}} - \mathrm{clr}_j x_{\mathrm{B}} \right)^2}$$

As a result of transformation of the raw data (concentrations), the orthogonal base of coordinate vectors can be constructed:

- isometric logratio (ilr)

$$y_k = \sqrt{\frac{k}{k+1}} \ln\left(\frac{\mathrm{g}\left(x_{1,} \ldots, x_k\right)}{x_{k+1}}\right), \quad k = 1,2,\ldots, D-1$$

- balance

$$z_l = \sqrt{\frac{rs}{r+s}} \ln \frac{\left(\Pi_+\right)^{1/r}}{\left(\Pi_-\right)^{1/s}}, \quad l = 1,2\ldots, D-1$$

where Π_+ is product of *r* selected components (concentrations) and Π_- is product of *s* selected components, different than the ones included in Π_+.

Operations and linear process in simplex space

Perturbation

$$\mathrm{x} \oplus \mathrm{p} = \mathrm{C}\left[x_1 p_1, \ldots, x_D p_D\right]$$

Power

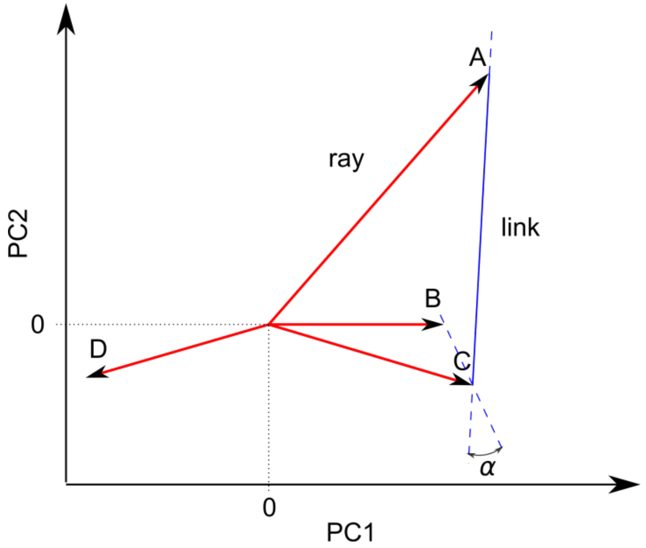
$$t \otimes \mathbf{x} = \mathrm{C}\left[x_1^t, \ldots, x_D^t\right]$$

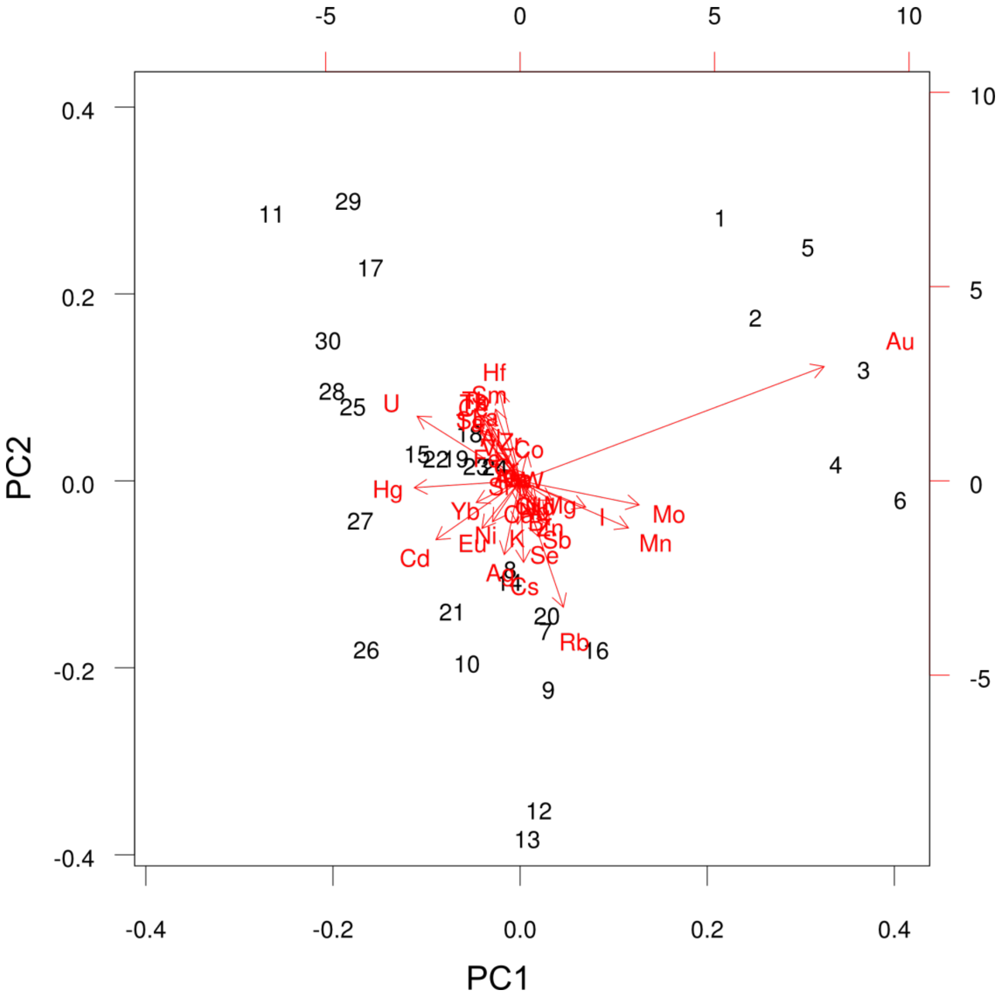
The linear process

$$\mathbf{x}(t) = \mathbf{x_0} \oplus (t \otimes \mathbf{v})$$

The data exploration

The results overview - Principal Component Analysis of the clr transformed variables

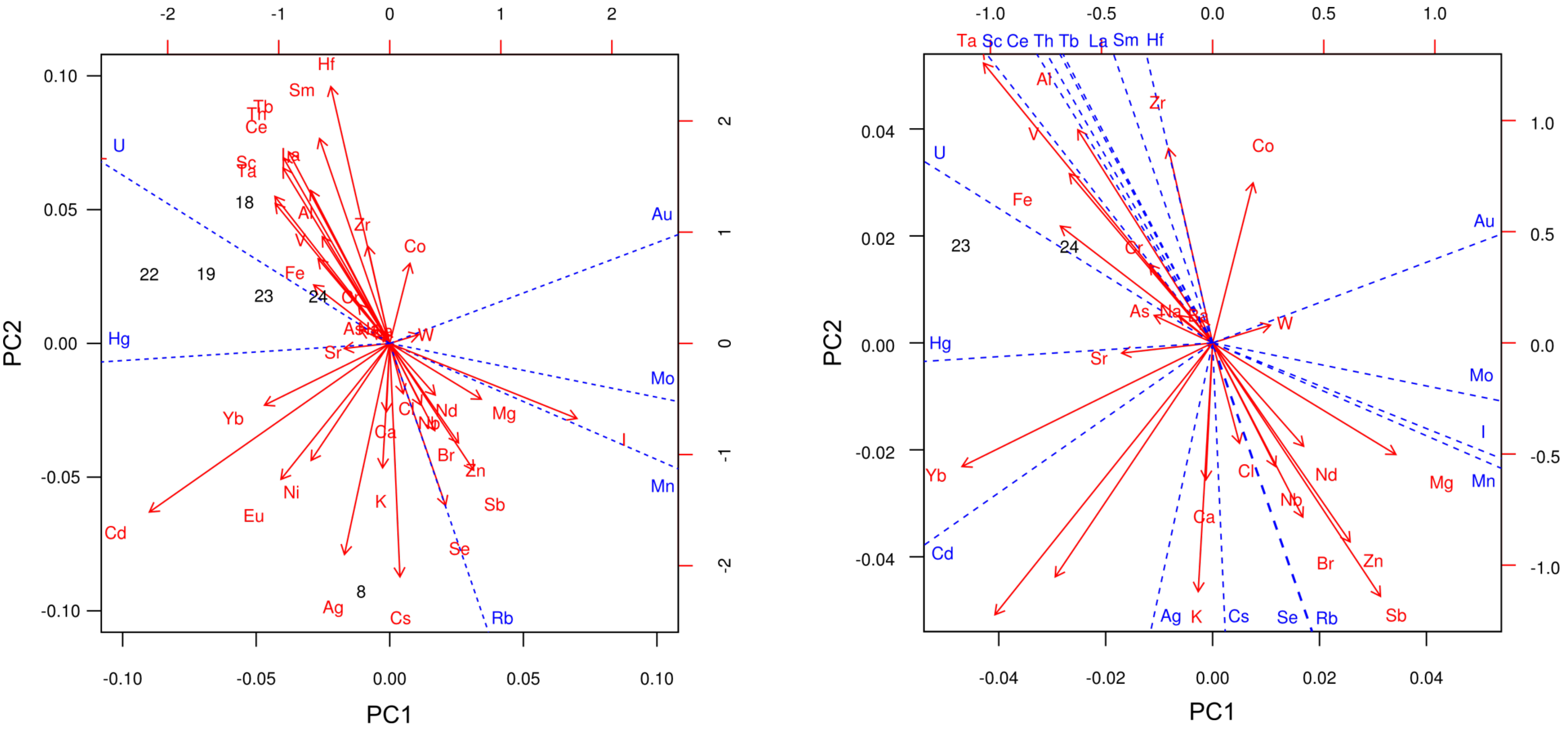




- Length of the ray estimates standard deviation of the transformed variable
- Length of the link estimates h_{jl}.
- $\cos\alpha$ estimates correlation between logratios

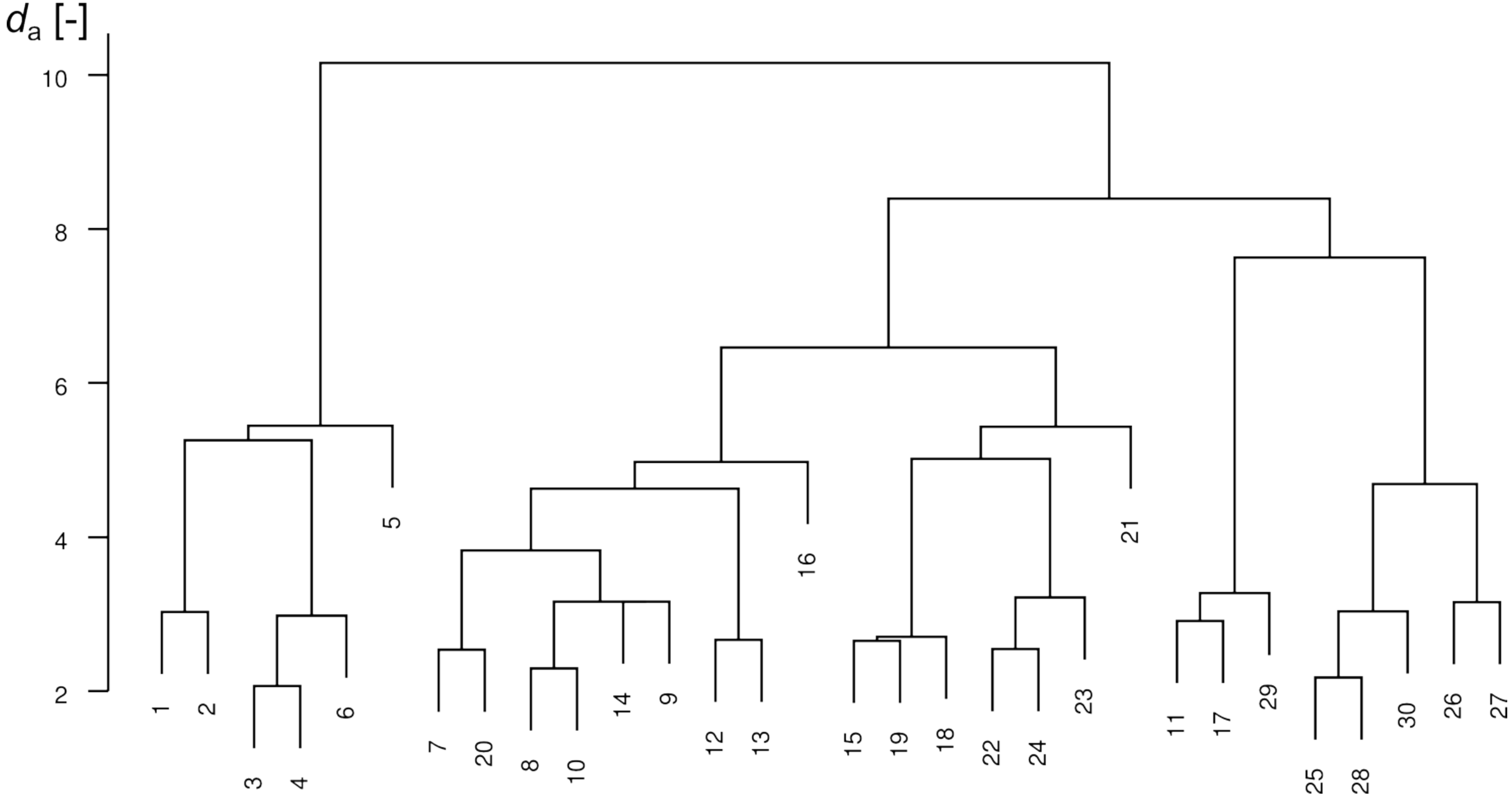
PC1 35.8 % of the total variance
PC2 22.0 % of the total variance

Aitchison, John & Greenacre, M., 2002. Biplots of compositional data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 51(4), pp.375-392.


PC1
PC2
-2
-1
0
1
2
0.10
0.05
0.00
-0.05
-0.10
U
Au
Hg
Mo
Mn
Rb
Hf
Sm
Tb
Th
Ce
Sc
Ta
La
18
Al
V
Zr
Co
22
19
Fe
23
24
Cr
As
Sr
W
Yb
Cl
Nd
Mg
Nb
Ca
Br
Zn
I
Ni
K
Sb
Eu
Cd
Se
8
Ag
Cs
-1.0
-0.5
0.0
0.5
1.0
0.04
0.02
0.00
-0.02
-0.04
Na
Ba
K
-0.5
-1.0

Cluster analysis.
Based on distances between points in simplex space

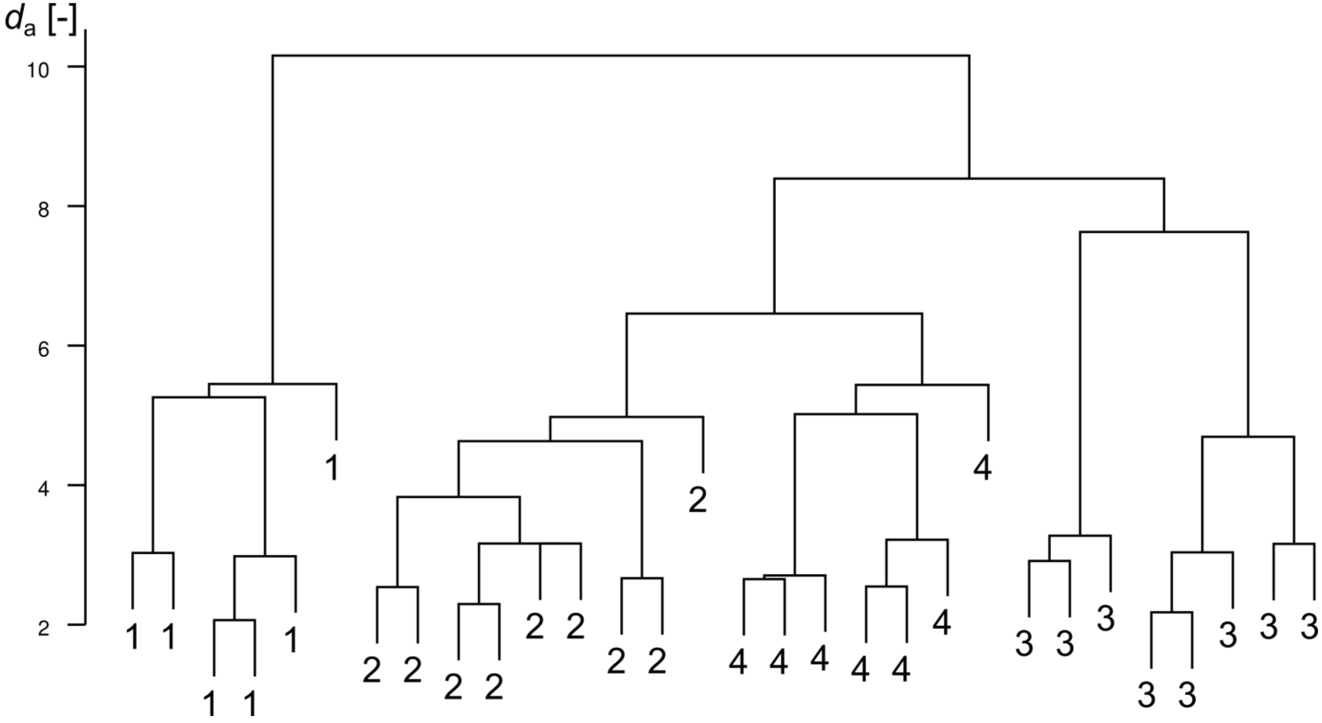
Divisive algorithm



4 clusters

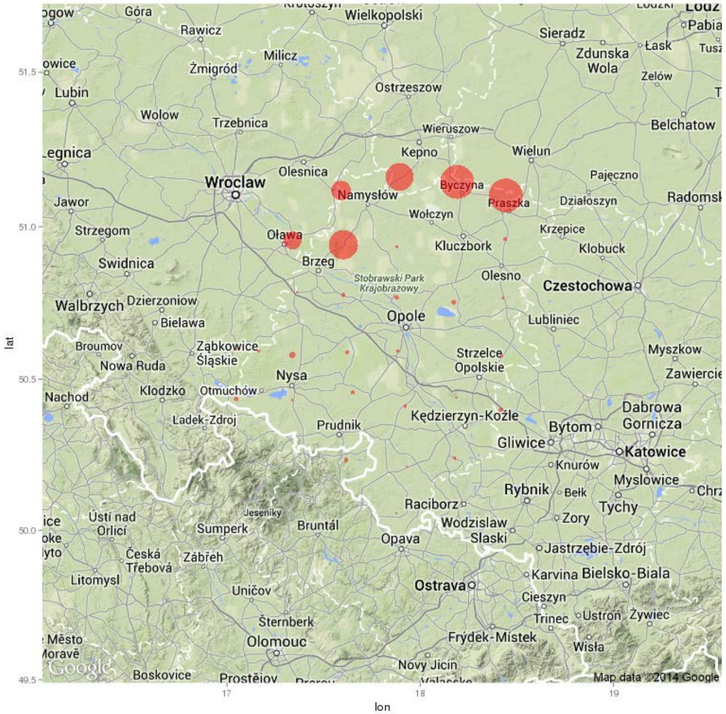
Fuzzy clustering

Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all-or-nothing) but "fuzzy". In a fuzzy clustering, each observation is “spread out” over the various clusters. The memberships are nonnegative, and for a fixed observation they sum to 1.

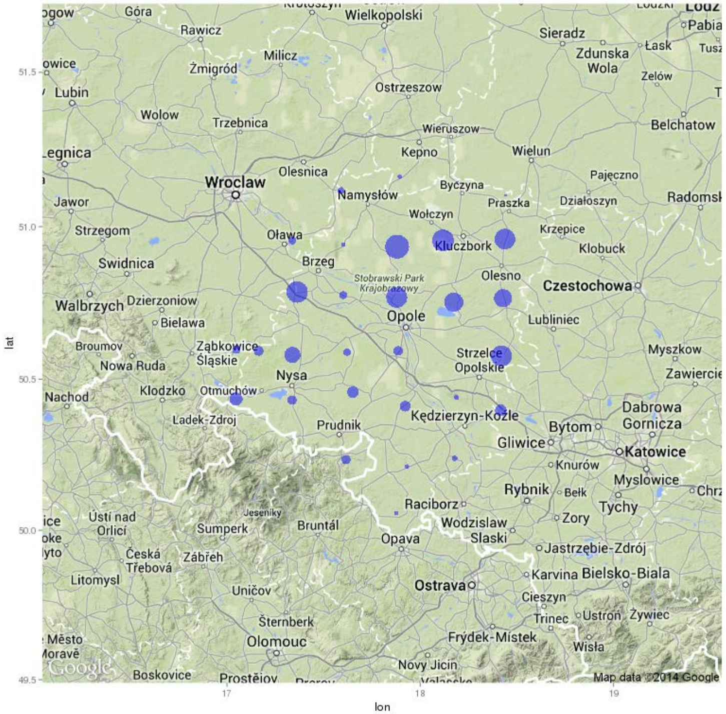


Result of “hard clustering”

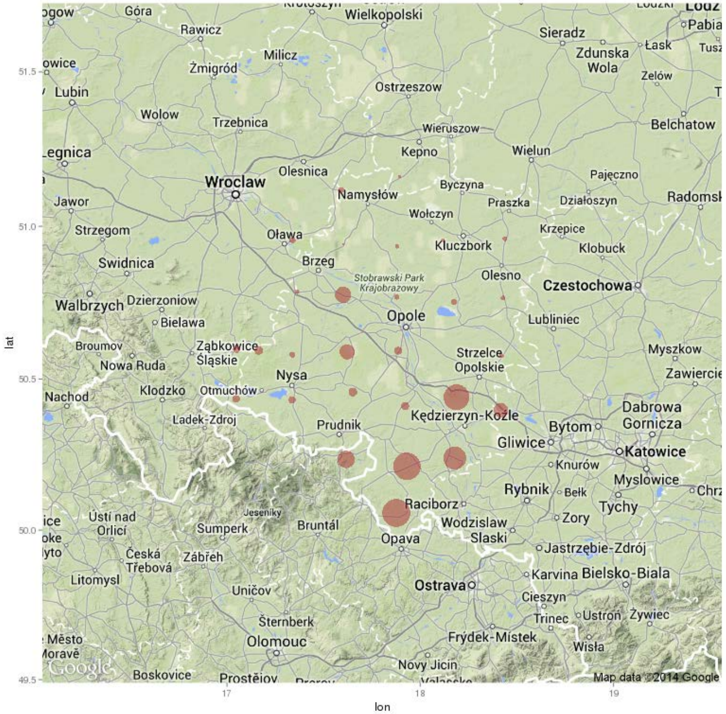
cluster 1



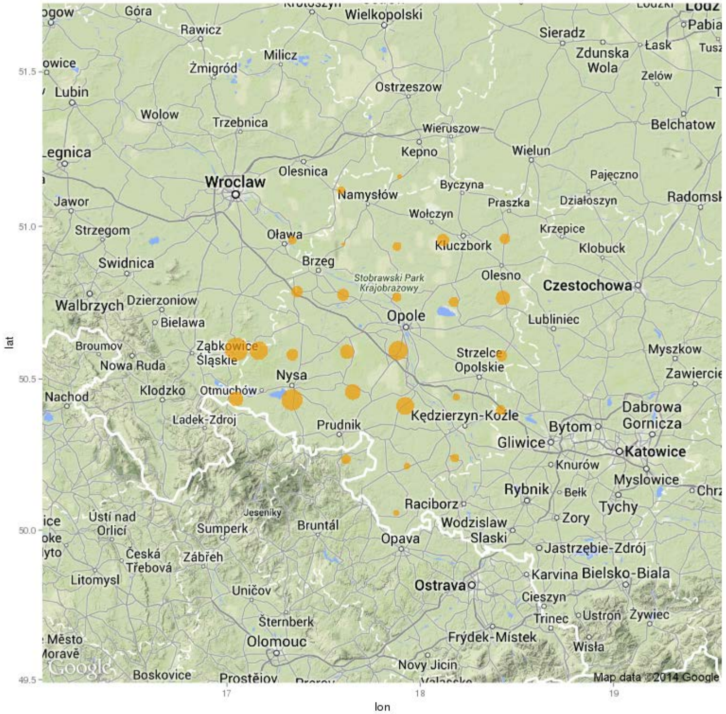
cluster 2



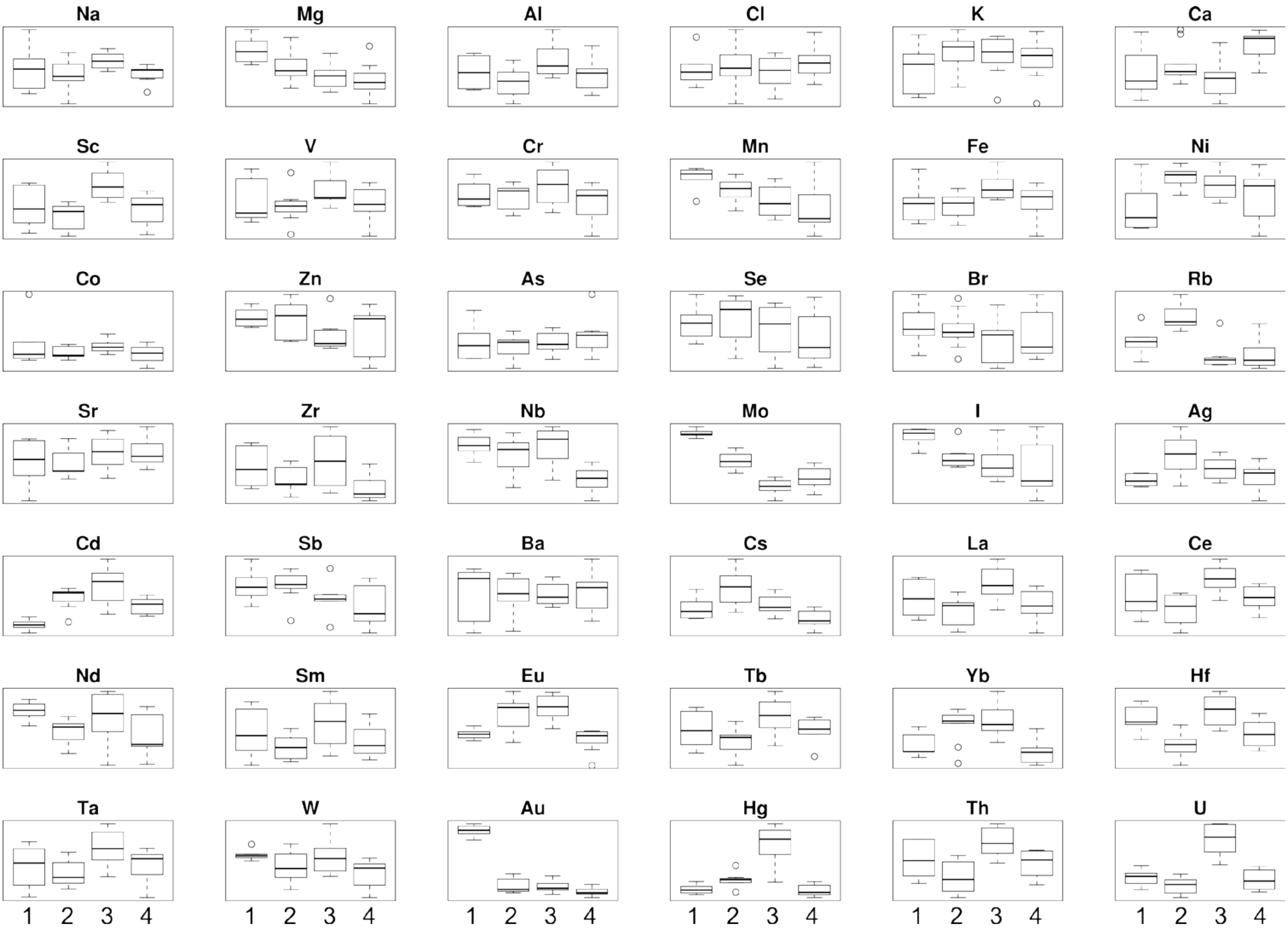
cluster 3



cluster 4



Distribution of logarithms of elements concentration in clusters



Differences in clusters compositions

cluster	Mn	Ni	Mo	I	Ag	Cd	Nd	Eu	Au	Hg	U
1	+	-	+	+	-	-	0	-	+	0	0
2	0	0	0	0	+	0	-	+	0	0	0
3	0	0	0	0	+	+	0	+	0	+	+
4	-	0	0	0	-	0	0	-	0	0	0

Covariability of elements concentration in clusters

Low $h_{E1,E2}$ values

Total number of pairs = 861 pairs × 4 clusters

E1	E2	h(1)
Mo	Eu	0.00
La	Th	0.00
Sc	Th	0.00
Sc	Tb	0.01
Al	La	0.01
Sc	La	0.01
Tb	Th	0.01
Mo	Cd	0.01
Mo	Nd	0.02
Nd	Eu	0.02

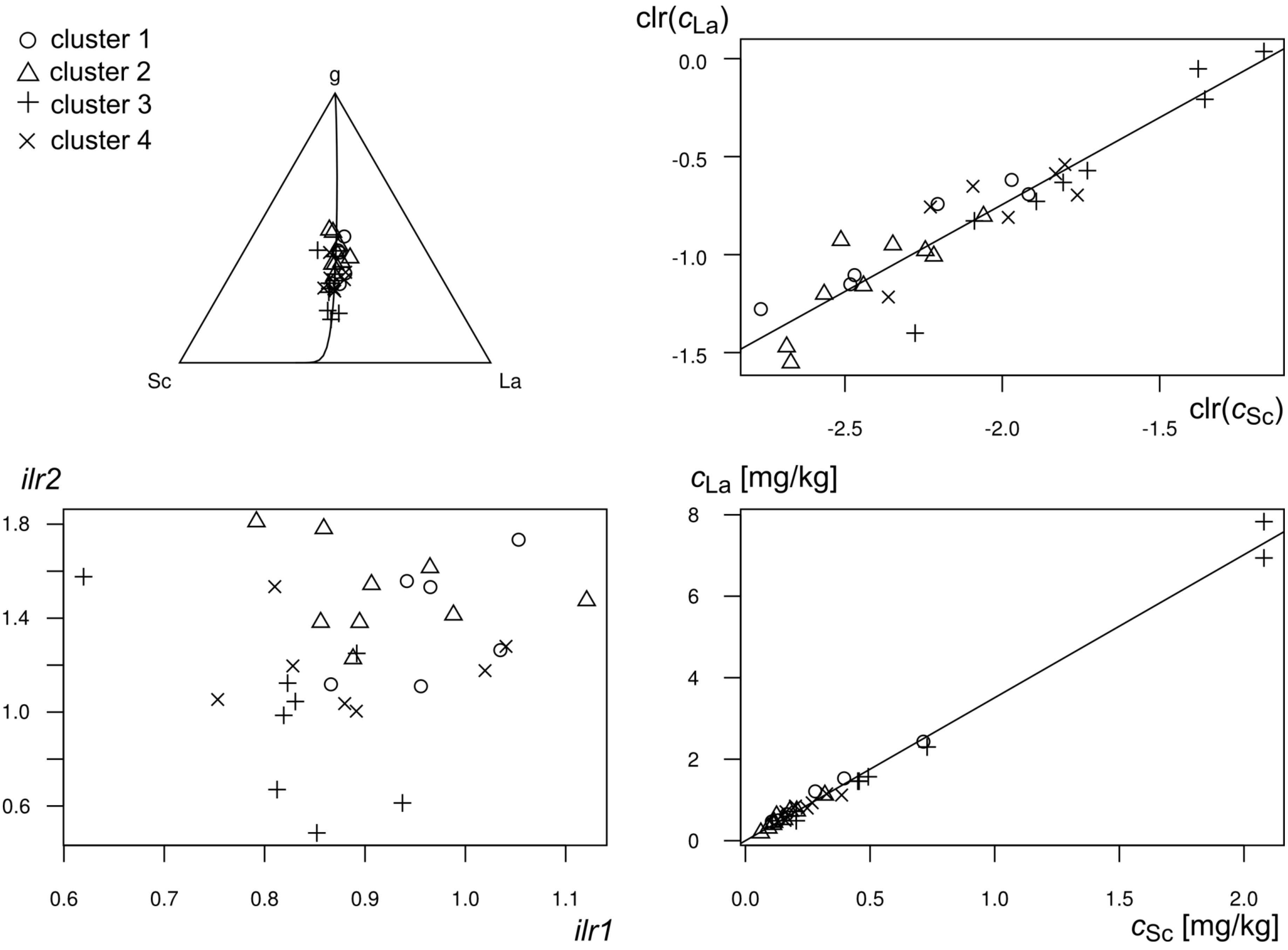
E1	E2	h(2)
Sm	U	0.01
Sc	La	0.02
Sc	U	0.02
Al	Sc	0.03
Sc	As	0.03
Al	U	0.03
La	U	0.03
Hf	Ta	0.03
Zr	Mo	0.03
Sc	Zn	0.03

E1	E2	h(3)
La	Ce	0.01
Sc	Th	0.01
La	Ta	0.02
Sc	La	0.02
Ta	Th	0.02
Sc	Fe	0.02
Fe	Ba	0.02
Al	Sc	0.02
Al	V	0.02
La	Th	0.02

E1	E2	h(4)
Sc	Co	0.01
La	Ce	0.01
Ce	Th	0.01
Mo	Ce	0.01
Fe	La	0.01
Nd	Hg	0.01
Sc	Fe	0.02
Th	Sc	0.02
Fe	Ce	0.02
Sc	Ce	0.02

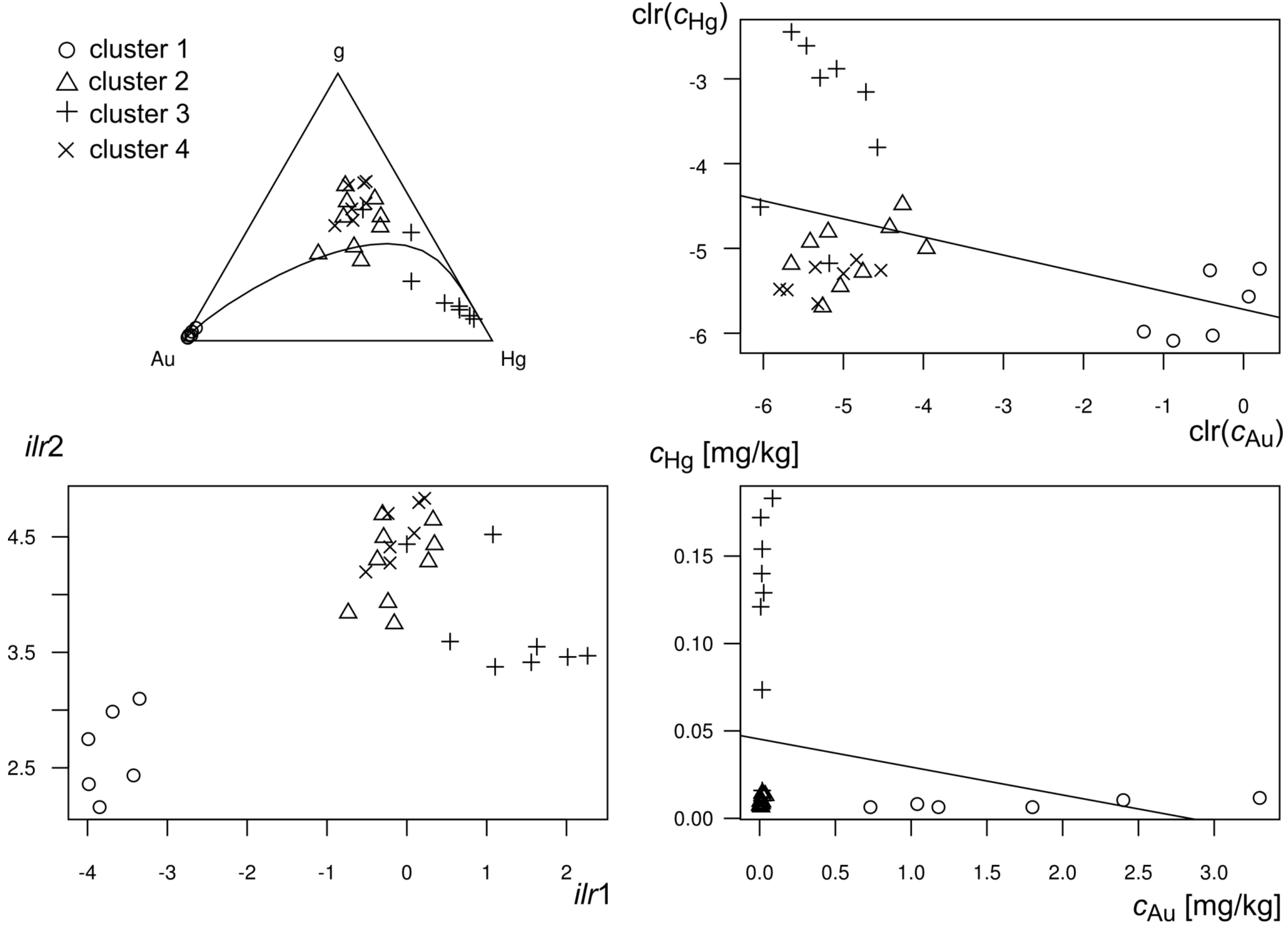
Relationship between concentrations of Sc and La

h=0.021, ***t***=[0.18, 0.19, 0.63]



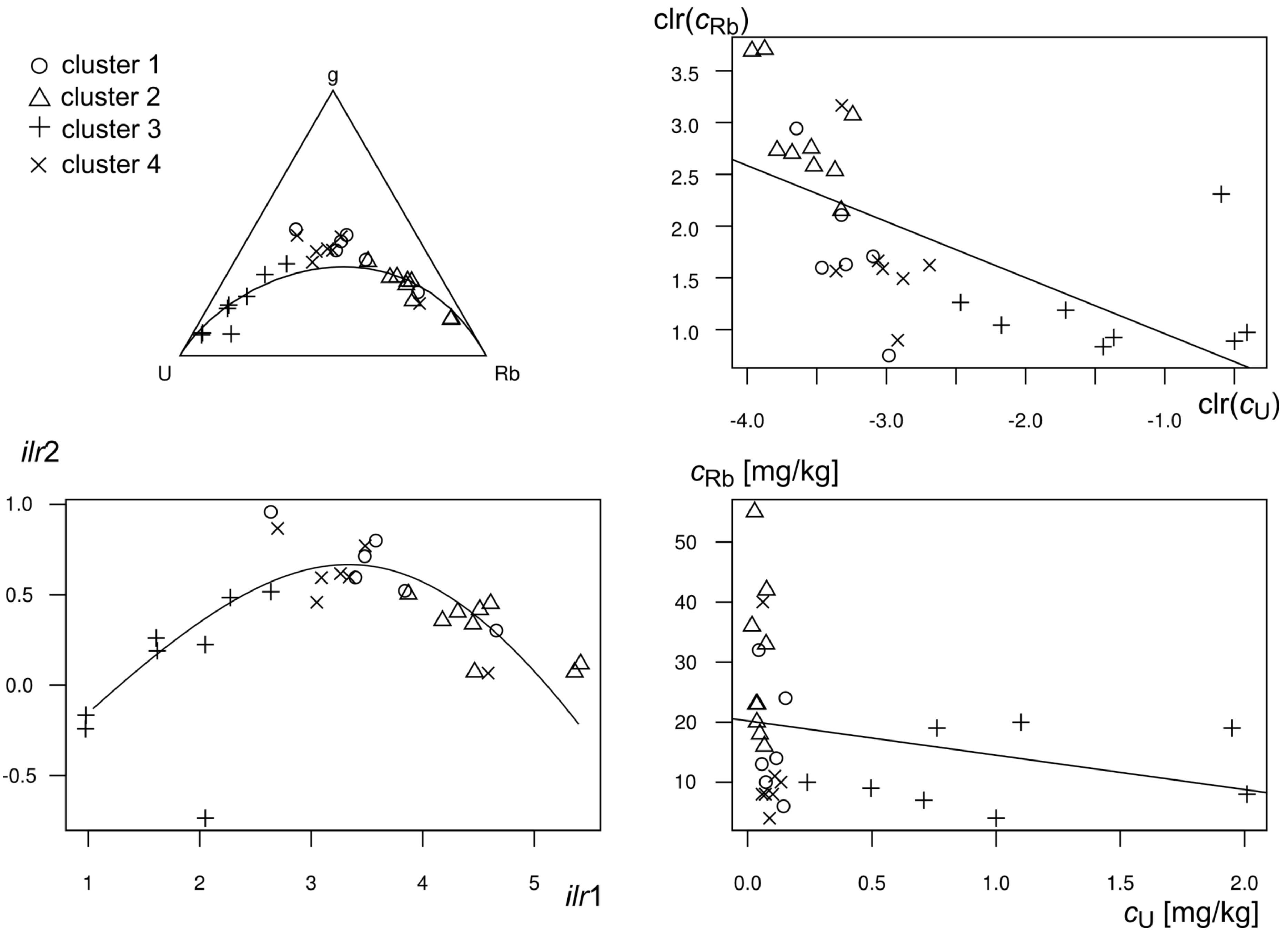
Relationship between concentrations of Au and Hg

h=6.6, ***t***=[0.62, 0.16, 0.22]



Relationship between concentrations of U and Rb

h=2.9, ***t***=[0.59, 0.14, 0.27]



Conclusions

- Compositional data require appropriate sample space to avoid erroneous or delusive statistical inference.
- Joint application of different statistical methods may provide a deep insight in the data structure.
- The obtained results constituted a solid base for further investigation regarding sources of matter, directions of spread, analysis of possible threats for environment.

Todo:
To ensure precision of inferences in the data analysis the expert knowledge regarding pedology, geology, industrial technology, history, and economy should be considered.