Weierstraß-Institut für
Angewandte Analysis und Stochastik

Finding Groups in Compositional Data – Some Experiments

Hans-Joachim Mucha, Tatjana Mirjam Gluhak

AG DANK Meeting, Berlin, WIAS, November 19, 2016

Outline

- Introduction
- Motivation
- Clustering of profiles
- Applications to archaeometry
- Validation of stability of clusters
- Summary

Introduction

In archaeometry or geology, the chemical composition of oxides of objects is measured, and often the results are presented in percentages. Then, so-called “compositional data analysis” (Aitchison 1986) should be applied as the only one valid statistical analysis.

Nowadays, besides oxides, a much greater number of trace elements can be measured by new innovative technical equipment. Usually, these measurements are in ppm (parts per million) or ppb (parts per billion).

The question arises: Can we find groups in such (mixed) data by applying “usual” statistical clustering?

Introduction

First, the talk is concerned with finding groups (clusters) in (strict) compositional data, that is nonnegative data with row sums equal to a constant, usually 1 in case of proportions or 100 in case of percentages.

Without loss of generality, the cluster analysis of observations of compositional data is considered, where the row profiles contains parts of some whole.

Distance functions between profiles and appropriate clustering methods are recommended. Finally, applications to archaeometry are presented.

Motivation

In archaeology, the aim of clustering is to find groups in data such as proveniences of glass objects or pottery.

The motivating example is taken from Baxter & Freestone (2006) where the complete original data matrix **Z** is published as it is analyzed hereafter.

Each object is characterized by $J = 11$ variables, the contents of oxides in %. The sum in each row is 100%.

This dataset of colorless Romano-British vessel glass contains two groups. Group 1 consists of 40 cast bowls with high amounts of Fe_2O_3. Group 2 also consists of 40 objects: this is a collection of facet-cut beakers with low Al_2O_3.

Motivation

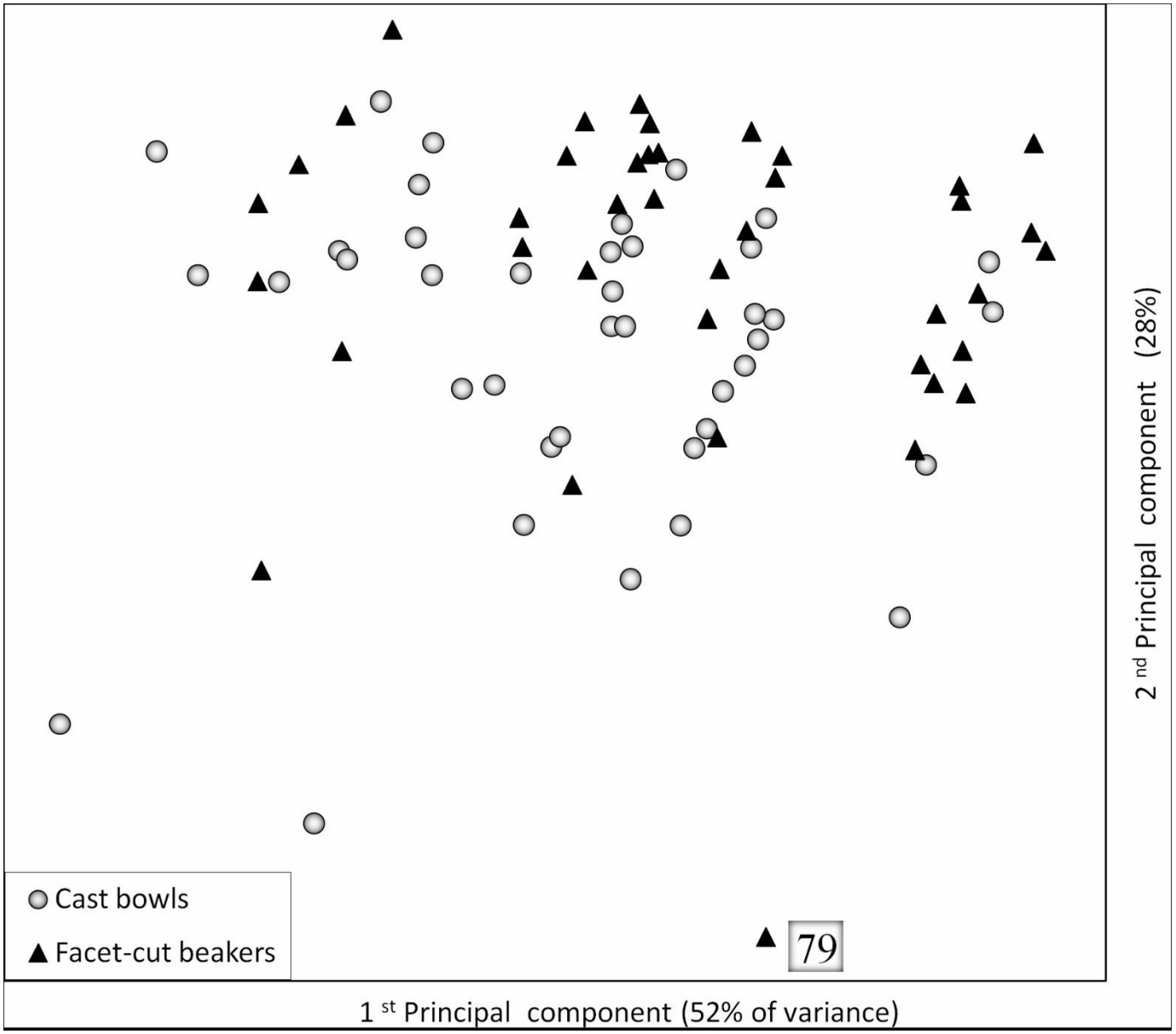
Log-ratio compositional data analysis (Aitchison 1986): use **X** with elements

$$x_{ij} = \log\left(z_{ij} / g(\mathbf{z}_i)\right)$$

instead of **Z**, where

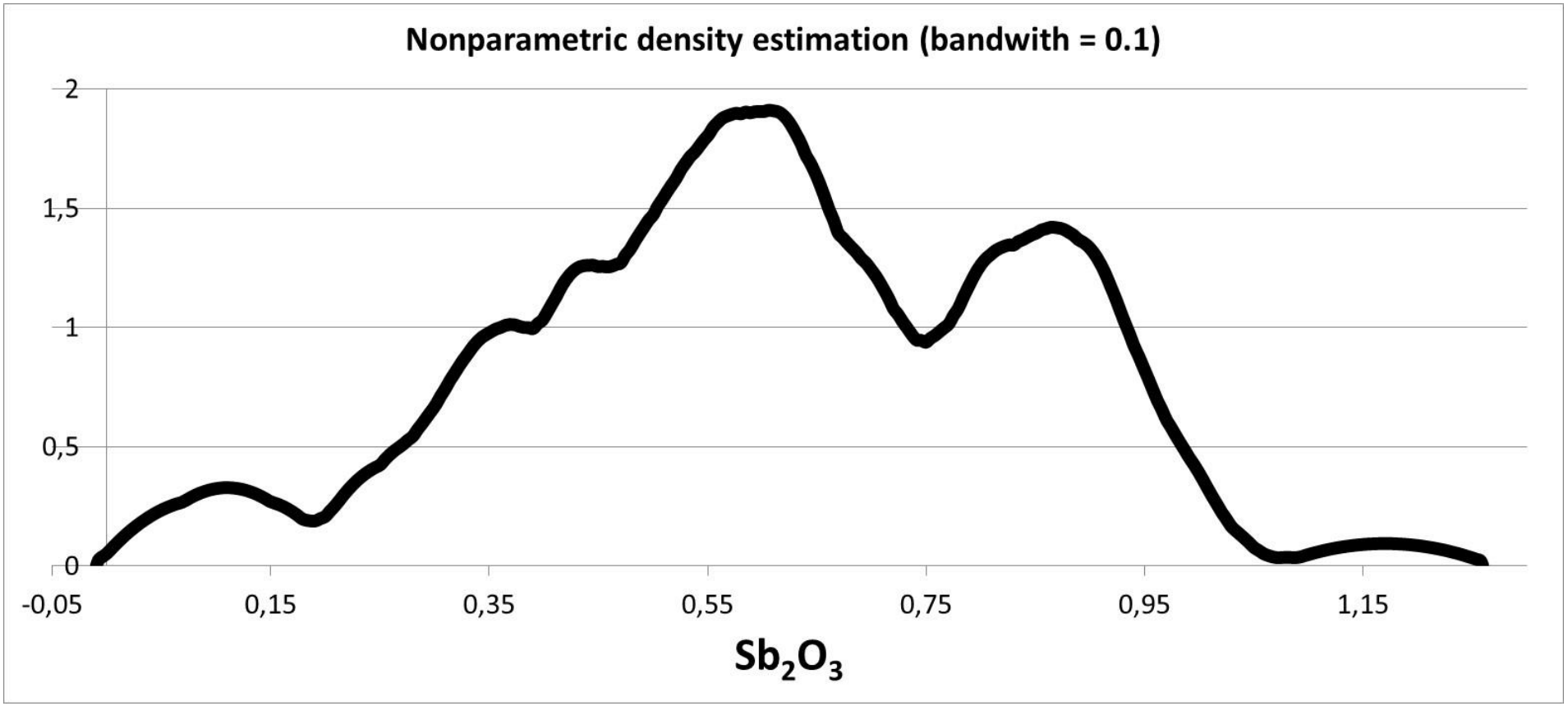
$$g(z_i) = (z_{i1} z_{i2} \ldots z_{iJ})^{1/J}$$

is the geometric mean of object i.



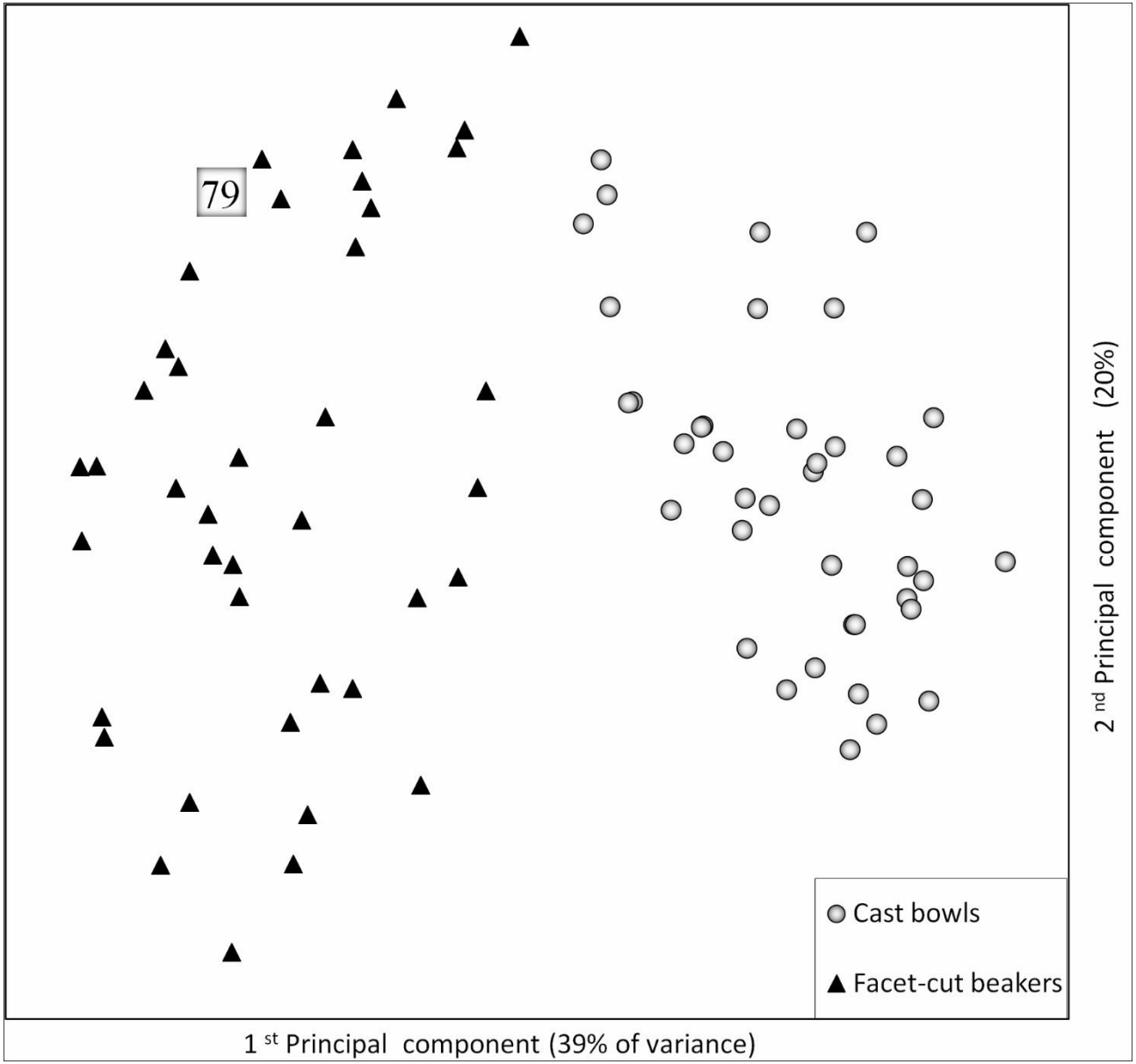
Motivation

Log-ratio compositional data analysis produces outliers, see object 79. Moreover, if one changes the value of Sb_2O_3 quite slightly, say from 0.08 to 0.0008, object 79 “drives” far away in the PCA plot.



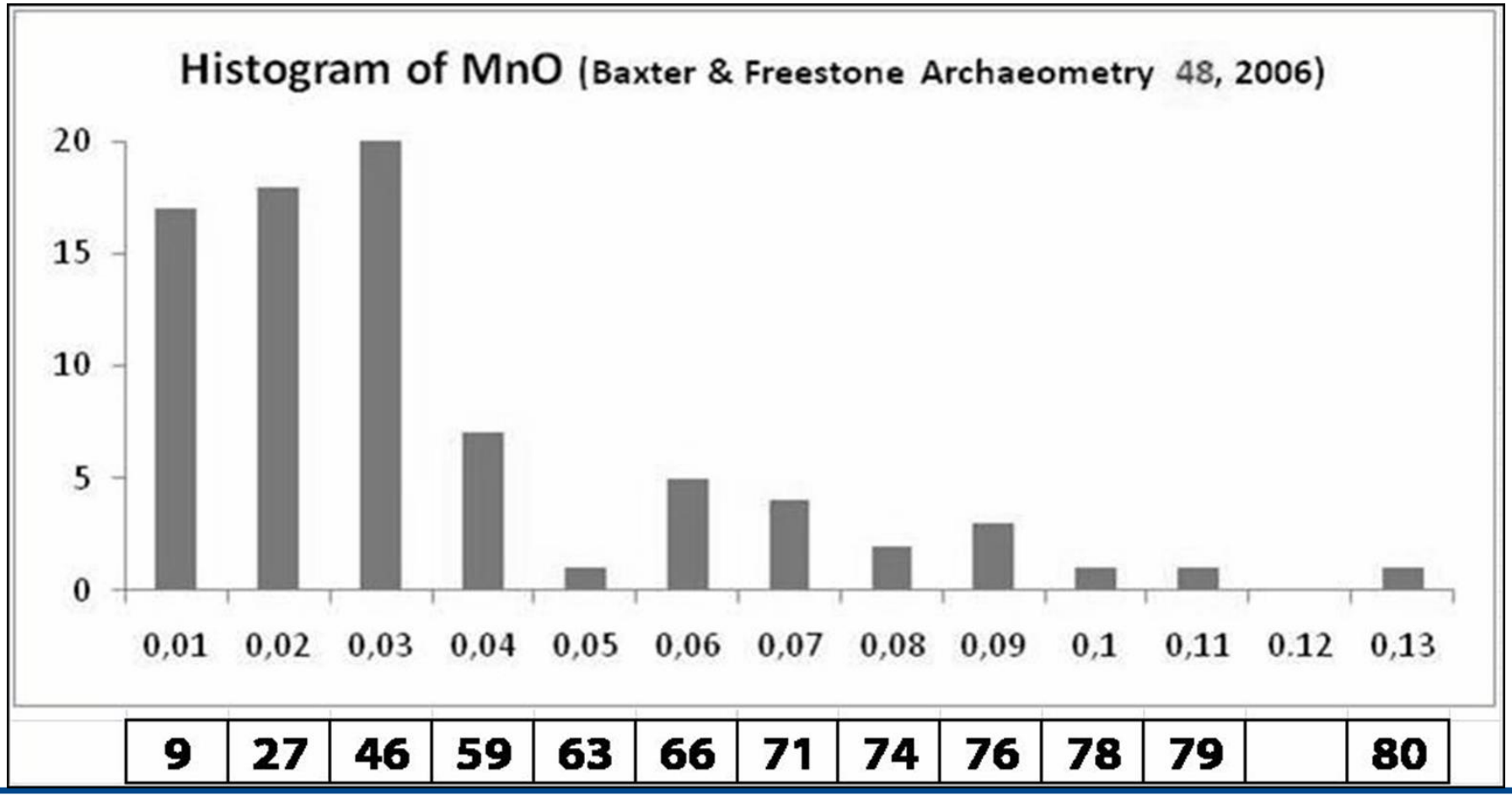
Motivation

In robust statistics, a common approach is to transfer the values for every variable to their rank values. PCA plot of groups of Romano-British vessel glass based on ranks.



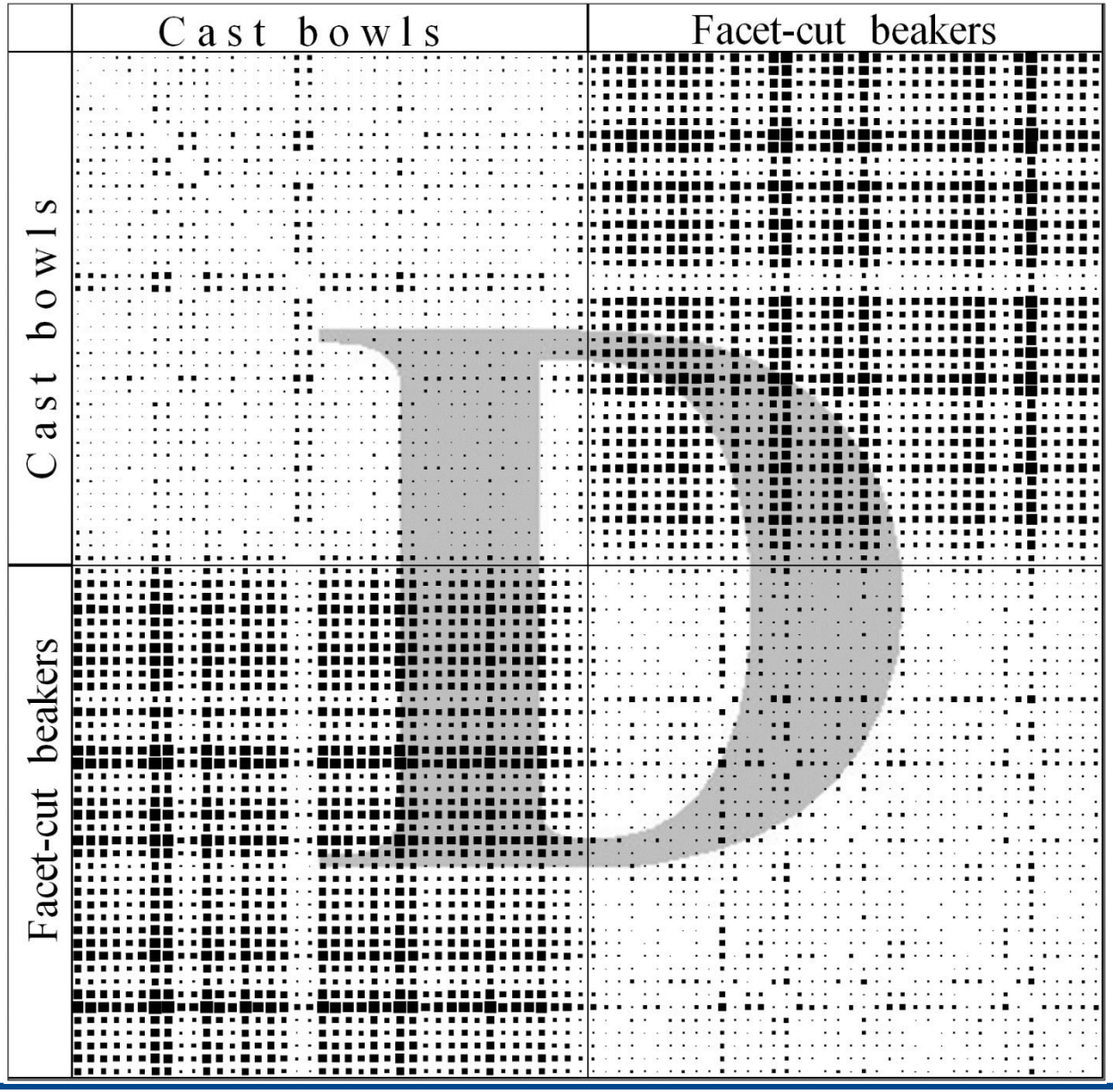
Motivation

Transformation into ranks solves the problems of different scales, skewness, and (univariate) outliers. Alternatives are concave transformations such as the logarithmic transformation (see later).



Motivation

What happens if one applies “usual” clustering? Heatmap of the Euclidean (80 × 80) distance matrix **D** of the vessel glass objects. The square of a cell is proportional to the distance value. Both, Ward and *k*-means find the true classes without error.



Clustering of profiles

Let's consider a **compositional** data matrix $\mathbf{Z} = (z_{ij})$ consisting of I rows and J columns (variables). The values of each row sums up to a constant c.

To be general, let us work with the matrix $\mathbf{X} = (x_{ij})$ of profiles of proportions $x_{ij} = z_{ij} / c$ (or $x_{ij} = z_{ij} / z_{i+}$).

Further, let $C = \{ \mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_I \}$ denote the finite set of the I observations (shortly: $C = \{ 1, ..., i, ..., I \}$).

In order to find groups in compositional data some kind of appropriate (dis)similarity measure is needed such as the Minkowski distance between profiles

$$d_{il} = (|x_{i1} - x_{l1}|^p + |x_{i2} - x_{l2}|^p + ... + |x_{iJ} - x_{lJ}|^p)^{1/p}$$

where p is a real number larger than or equal to 1.

Clustering of profiles

The usual task of clustering is finding a partition of C in K non-empty clusters (subsets) C_k, $k = 1,2,...,K$.

The most general model-based Gaussian clustering is when the covariance matrix $\mathbf{\Sigma}_k$ of each cluster k is allowed to vary completely.

Then $V_K = \sum_{k=1}^{K} n_k \log \left| \frac{\mathbf{W}_k}{n_k} \right|$ has to minimized, where

$\mathbf{W}_k = \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T$ is the sample cross-product matrix, n_k the cardinality, and $\bar{\mathbf{x}}_k$ the mean profile cluster k.

Clearly, $\mathbf{W}_k$ is singular in the case of compositional data.

Clustering of profiles

However, the sum of squares (SS) criterion $V_K = \mathrm{tr}(\sum_{k=1}^{K} \mathbf{W}_k)$ looks fit for clustering of profiles.

This is equivalent to minimizing $V_K = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in C_k} \sum_{\substack{l \in C_k \\ l>i}} d_{il}$,

where $d_{il} = \|\mathbf{x}_i - \mathbf{x}_l\|^2$ is the squared Euclidean distance. The latter can be generalized to

$$d_{\mathbf{Q}}(\mathbf{x}_i, \mathbf{x}_l) = (\mathbf{x}_i - \mathbf{x}_l)^T \mathbf{Q} (\mathbf{x}_i - \mathbf{x}_l) = \|\mathbf{x}_i - \mathbf{x}_l\|_{\mathbf{Q}}^2 \quad ,$$

where $\mathbf{Q}$ is diagonal with $q_{jj} = q_j > 0$. Examples are

$\mathbf{Q} = \mathrm{diag}(\frac{x_{++}}{x_{+1}}, \frac{x_{++}}{x_{+2}}, ..., \frac{x_{++}}{x_{+J}})$ (to get Chi-squared distance),

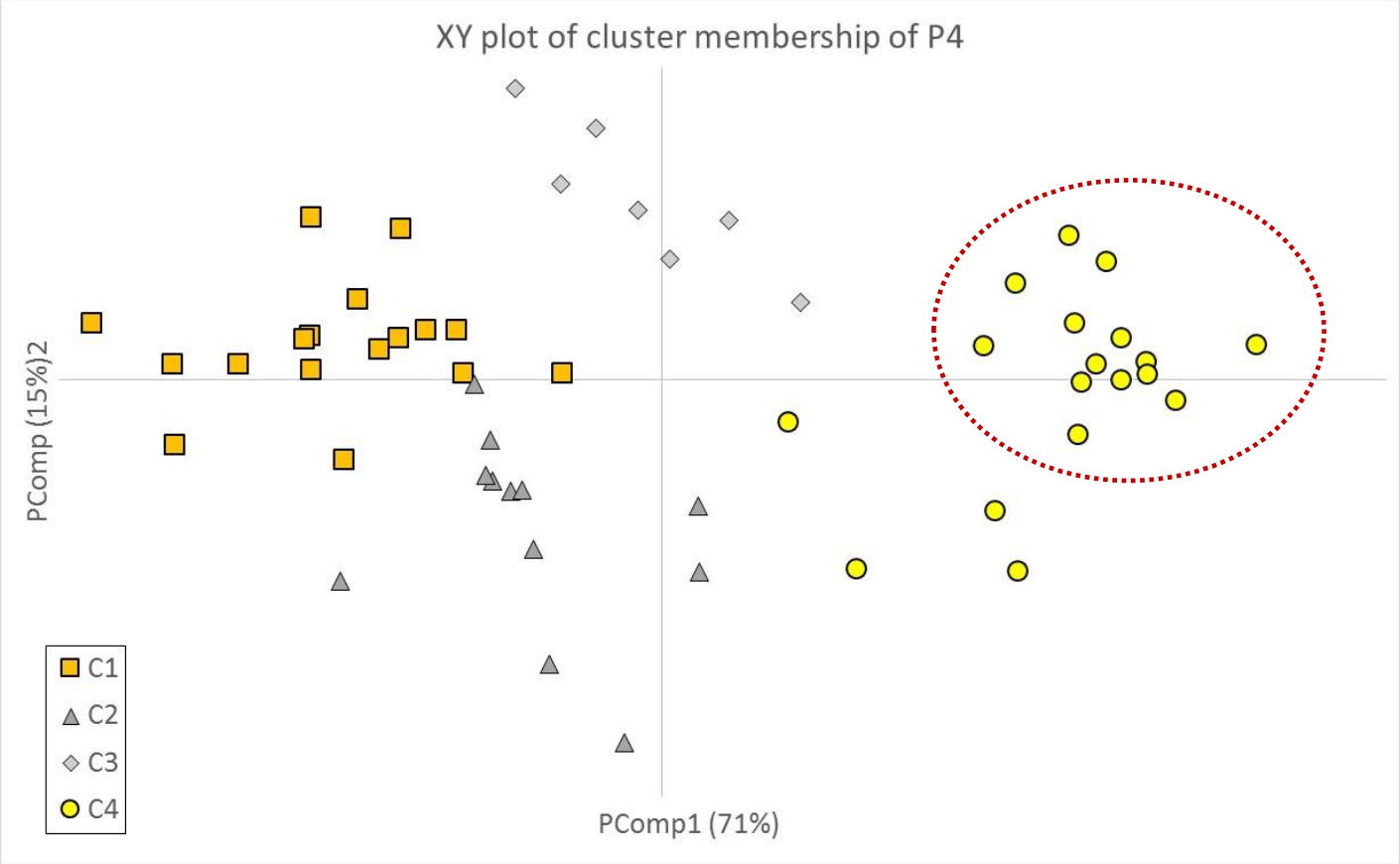
$\mathbf{Q} = (\mathrm{diag}(\mathbf{S}))^{-1}$ (**S**: estimate of Σ)

$\mathbf{Q} = (\mathrm{diag}(\bar{x}_1, \bar{x}_2, ..., \bar{x}_J))^{-1/2}$ (transformation to mean 1) …

Applications to archaeometry

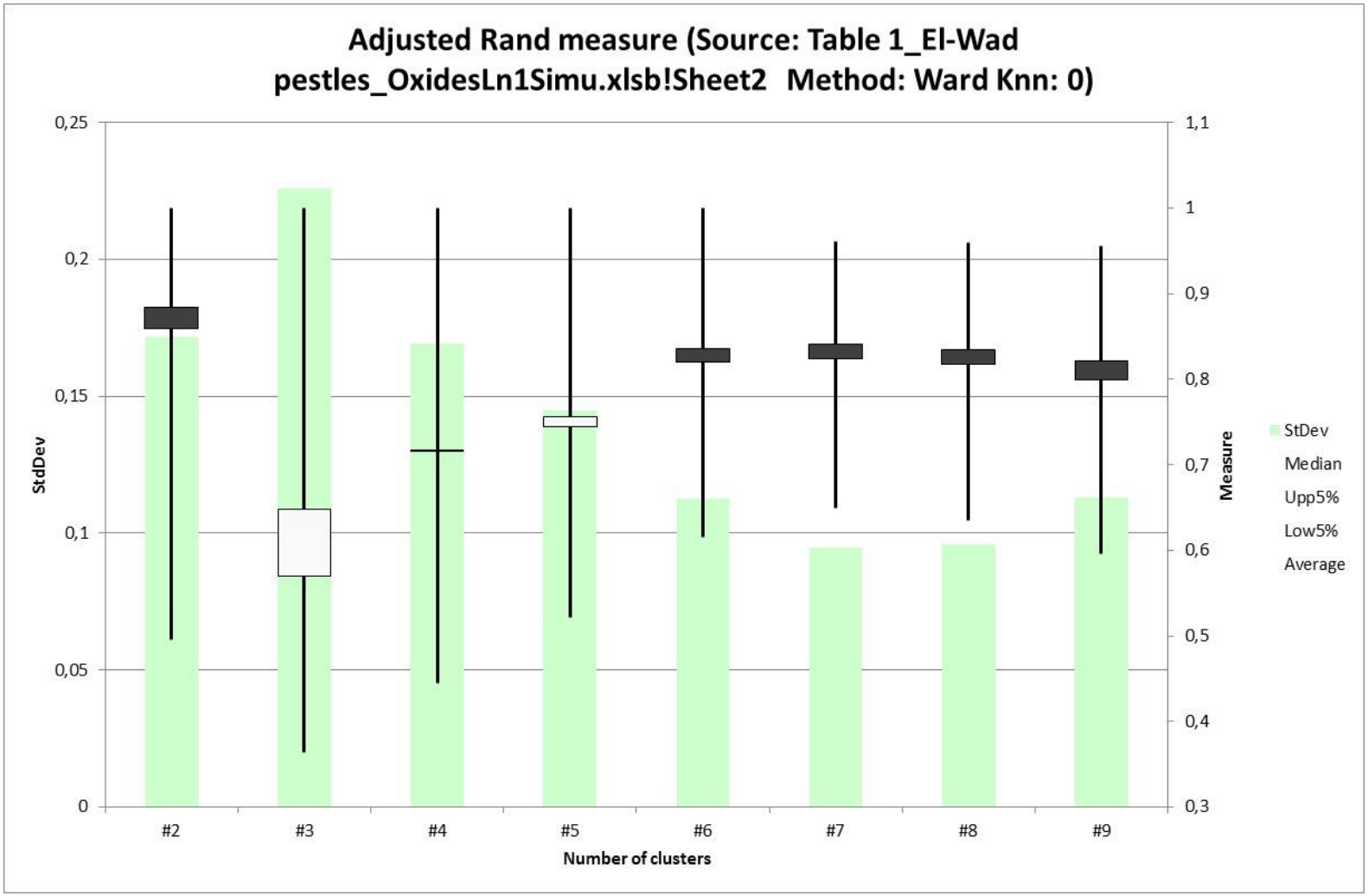
PCA plot of cluster membership of 54 basalt mortars from El-Wad based on logarithmic transformed data $x_{ij} = \ln(z_{ij} + 1)$. Ward's hierarchical method is applied. 11 oxides (28 trace elements not used).

Most stable region (see validation below): 103, 105, 107, 111, 122, 124, 136, 137, 139, 25, 40, 48, 53, 83).



Applications to archaeometry

Investigation of stability of Ward's clustering via bootstrap resampling technique. The ARI votes for K=2 clusters.



Applications to archaeometry

Investigation of individual cluster stability of Ward's clustering. The stability of clusters look quite different.

Jaccard measure and averaged Jaccard (bottom)

Cluster	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
1	0,889	0,878	0,859	0,84	0,761	0,71101	0,72555	**0,945**
2	0,467	0,37	0,81	0,777	0,677	0,67024	0,47474	**0,905**
3	0,847	0,831	0,691	0,747	0,737	0,75238	**0,914**	
4	0,681	0,685	0,379	0,873	0,866	**0,901**		
5	0,506	0,449	0,851	0,952	0,864			
6	0,686	0,782	0,963	0,623				
7	0,723	0,948	0,746					
8	0,9	0,777						
9	0,795							
Total	0,797	0,811	0,823	0,831	0,796	0,77075	0,756	**0,931**

Introduction

Investigation of stability (reliability) of cluster membership of each observation (partial view of the table).

	Partitions			Reliability (#Simul: 250)		
Name	P2	P3	P4	P2 %	P3 %	P4 %
101	1	1	1	100	92,72	98,68
102	1	1	2	100	82,94	57,06
103	2	3	4	100	100	100
105	2	3	4	100	100	100
106	2	3	4	66,45	49,68	37,42
107	2	3	4	100	100	100
110	1	1	2	97,39	59,48	99,35
111	2	3	4	100	100	100
114	1	1	2	85,35	42,68	80,89
115	1	2	3	97,18	98,59	98,59
116	1	1	1	100	88,05	81,76
118	1	2	3	95,24	100	100
120	1	1	1	100	78,82	57,06
121	1	1	1	100	92,26	99,35
122	2	3	4	100	100	100
124	2	3	4	100	100	100
125	1	1	1	100	94,04	99,34
128	1	1	2	98,18	60,61	99,39
133	1	1	1	100	78,67	60,67
136	2	3	4	100	100	100
138	1	1	1	100	95,97	99,33
139	2	3	4	100	100	100
140	2	3	4	93,88	85,03	75,51
187	1	1	1	100	77,24	54,48
188*	1	1	2	100	77,92	88,96
189	1	1	1	100	94,05	97,02
19	1	1	1	100	77,91	59,88
200	1	1	1	100	84,3	58,72
22	1	1	2	100	81,44	82,63
25	2	3	4	100	100	100

Summary

Log-ratio compositional data analysis is problematic (zero values not allowed, generates outliers, geometric mean seems to be inappropriate,…).

The most difficult problems in clustering remain:

- model selection (variable selection),
- validation (investigation of stability) of clustering results,
- appropriate data preprocessing such as logarithmic transformation (for instance, $x_{ij} = \ln(z_{ij} + 1)$)

Thank you very much for your attention!