Weierstraß-Institut für
Angewandte Analysis und Stochastik

Visualisation and Cluster Analysis

Hans-Joachim Mucha

Outline

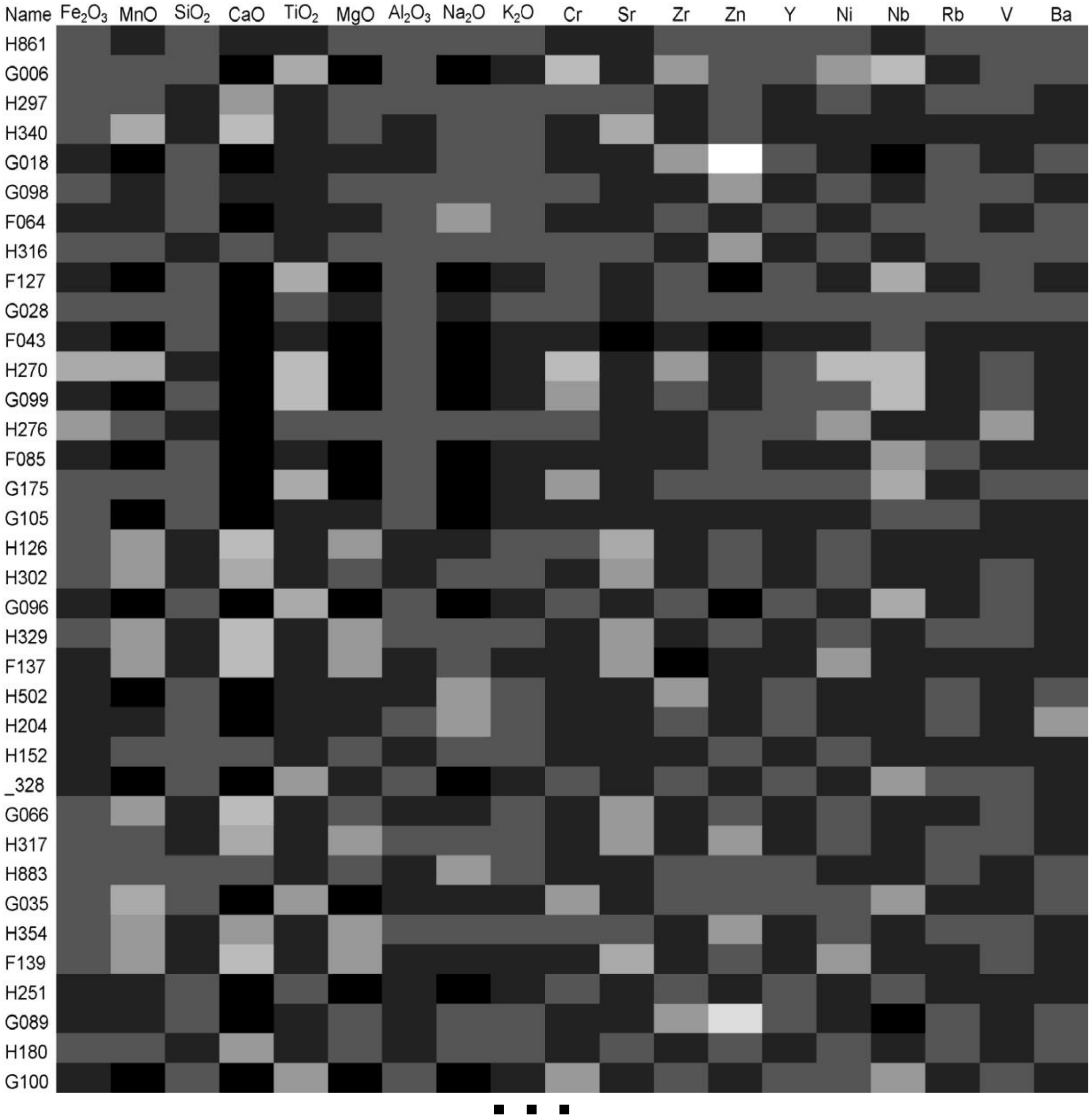
- Introduction
- Hierarchical and partitional clustering
- Multivariate projection
- Dendrograms
- Mapping findspots
- Summary

Introduction

Usually, the starting point for clustering is a I x J data matrix $\mathbf{X}=(x_{ij})$ with I observations and J variables.

Application to archaeometry

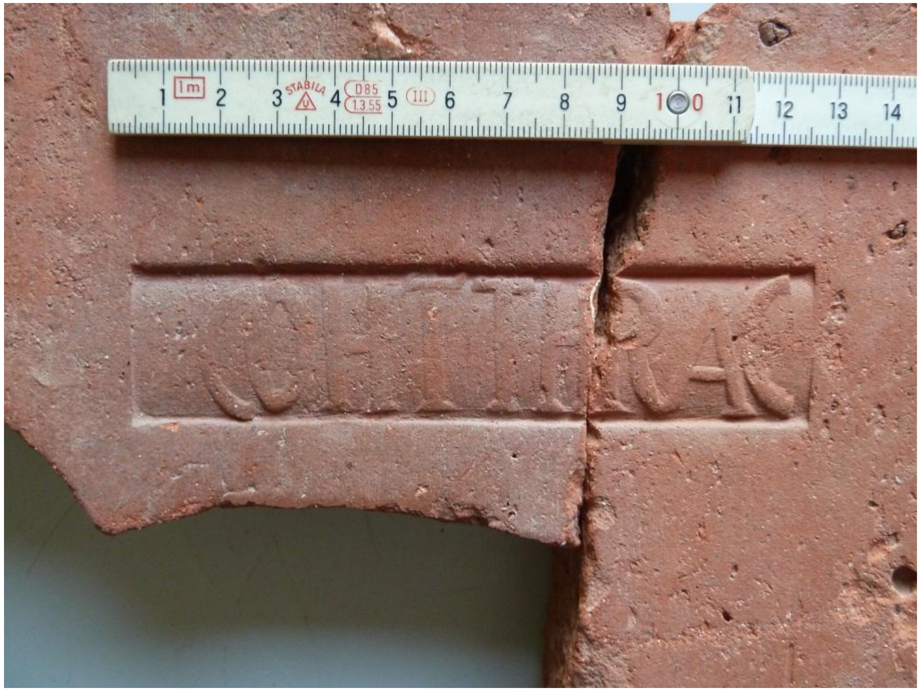
Snapshot of a fingerprint of the data as it comes: measurements of $J=19$ oxides and elements of $I=613$ Roman tiles from locations in *Germania Superior*. The darker the gray the higher is the measurement value.

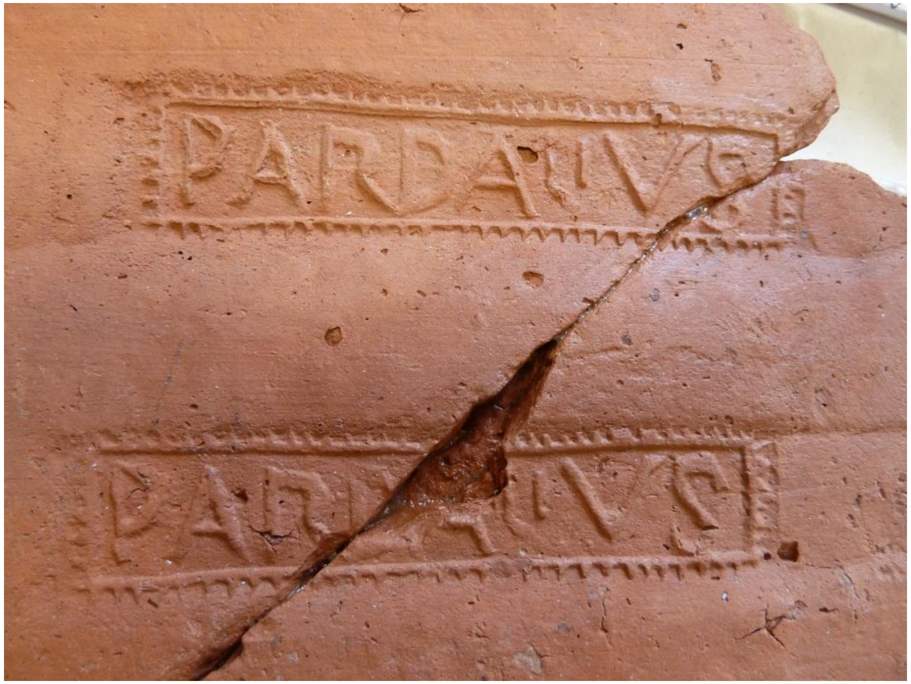


Introduction

Roman tiles (tegula) with a stamp from different locations in *Germania Superior*.

(More about Roman tiles and their multivariate statistical analysis at the website **http://www.ziegelforschung.de** .)

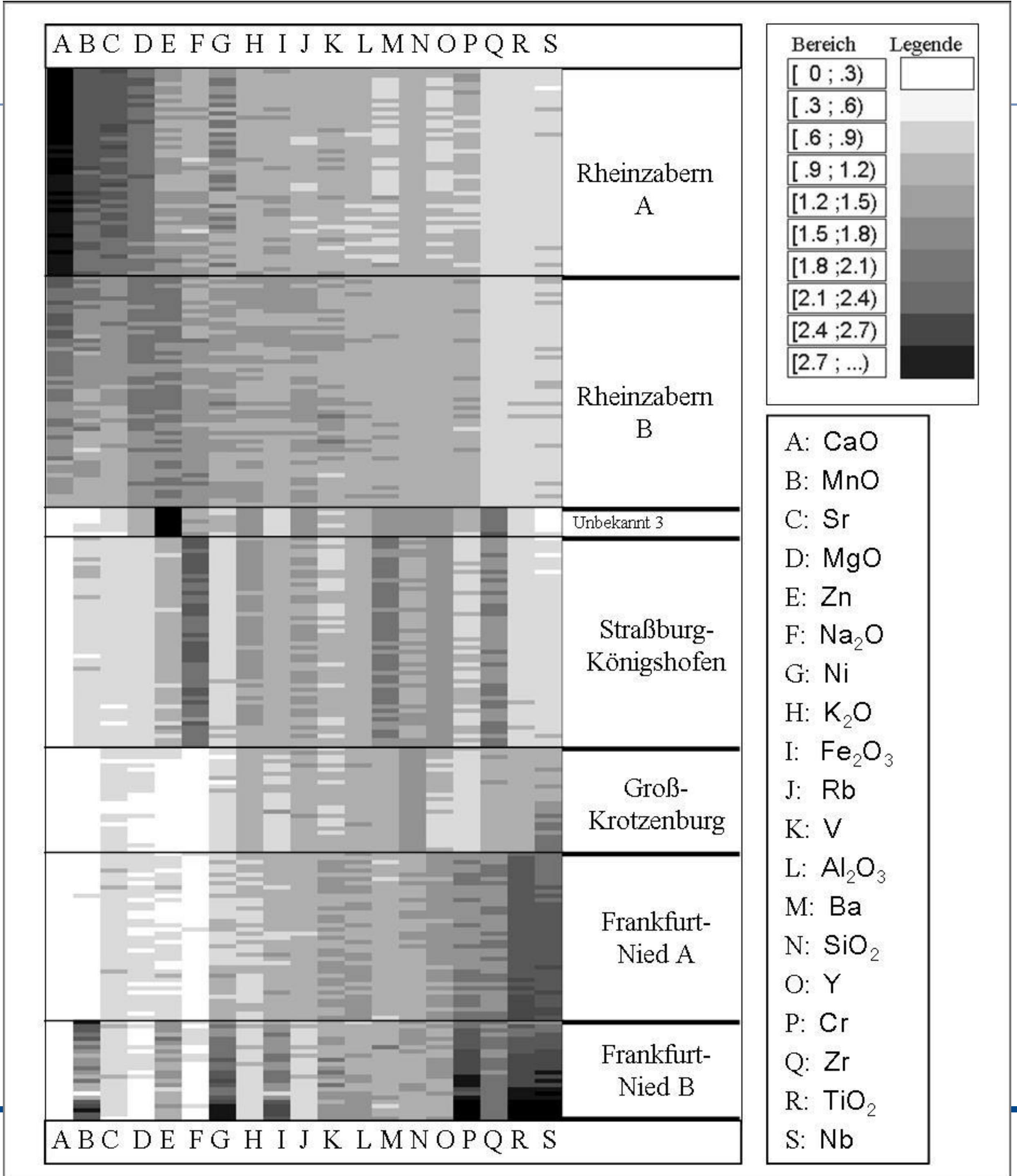




Introduction

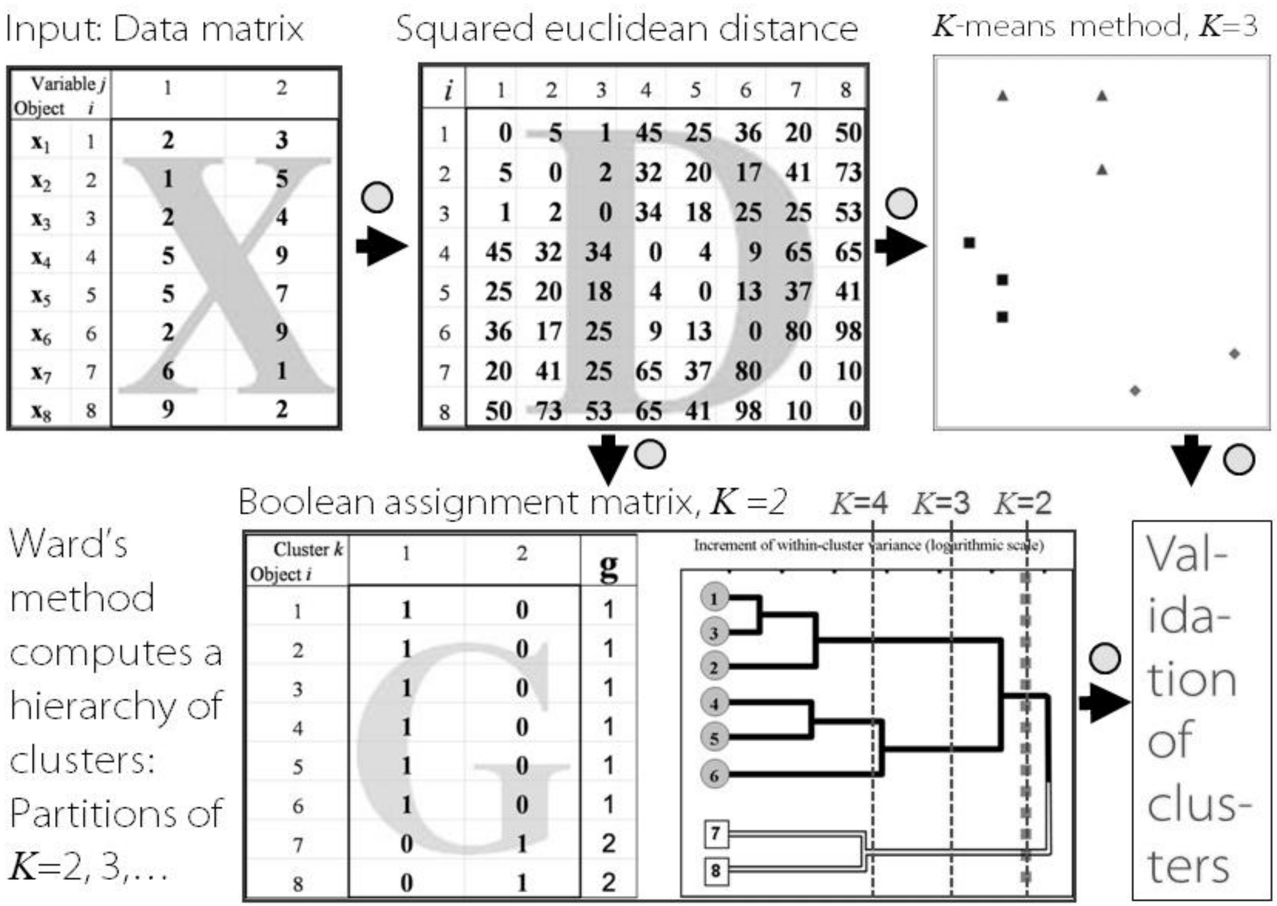
Aim: Clustering is used to group objects such that similar ones are collected in the same cluster.

Tiles (same data as before): Here, the final aim of clustering is to find proveniences. Then the data can be ordered by both the clusters and some projection scores such as the first principal component to make an ordered fingerprint.



Introduction

Sometimes, both hierarchical and partitional clustering can be formulated as pairwise data clustering, where instead of $\mathbf{X}=(x_{ij})$ a distance (proximity) matrix $\mathbf{D}=(d_{il})$ is used.



Input: Data matrix

Variable j Object i		1	2
$\mathbf{x}_1$	1	2	3
$\mathbf{x}_2$	2	1	5
$\mathbf{x}_3$	3	2	4
$\mathbf{x}_4$	4	5	9
$\mathbf{x}_5$	5	5	7
$\mathbf{x}_6$	6	2	9
$\mathbf{x}_7$	7	6	1
$\mathbf{x}_8$	8	9	2

Squared euclidean distance

i	1	2	3	4	5	6	7	8
1	0	5	1	45	25	36	20	50
2	5	0	2	32	20	17	41	73
3	1	2	0	34	18	25	25	53
4	45	32	34	0	4	9	65	65
5	25	20	18	4	0	13	37	41
6	36	17	25	9	13	0	80	98
7	20	41	25	65	37	80	0	10
8	50	73	53	65	41	98	10	0

K-means method, *K*=3

Ward's method computes a hierarchy of clusters: Partitions of *K*=2, 3,…

Boolean assignment matrix, *K* =2

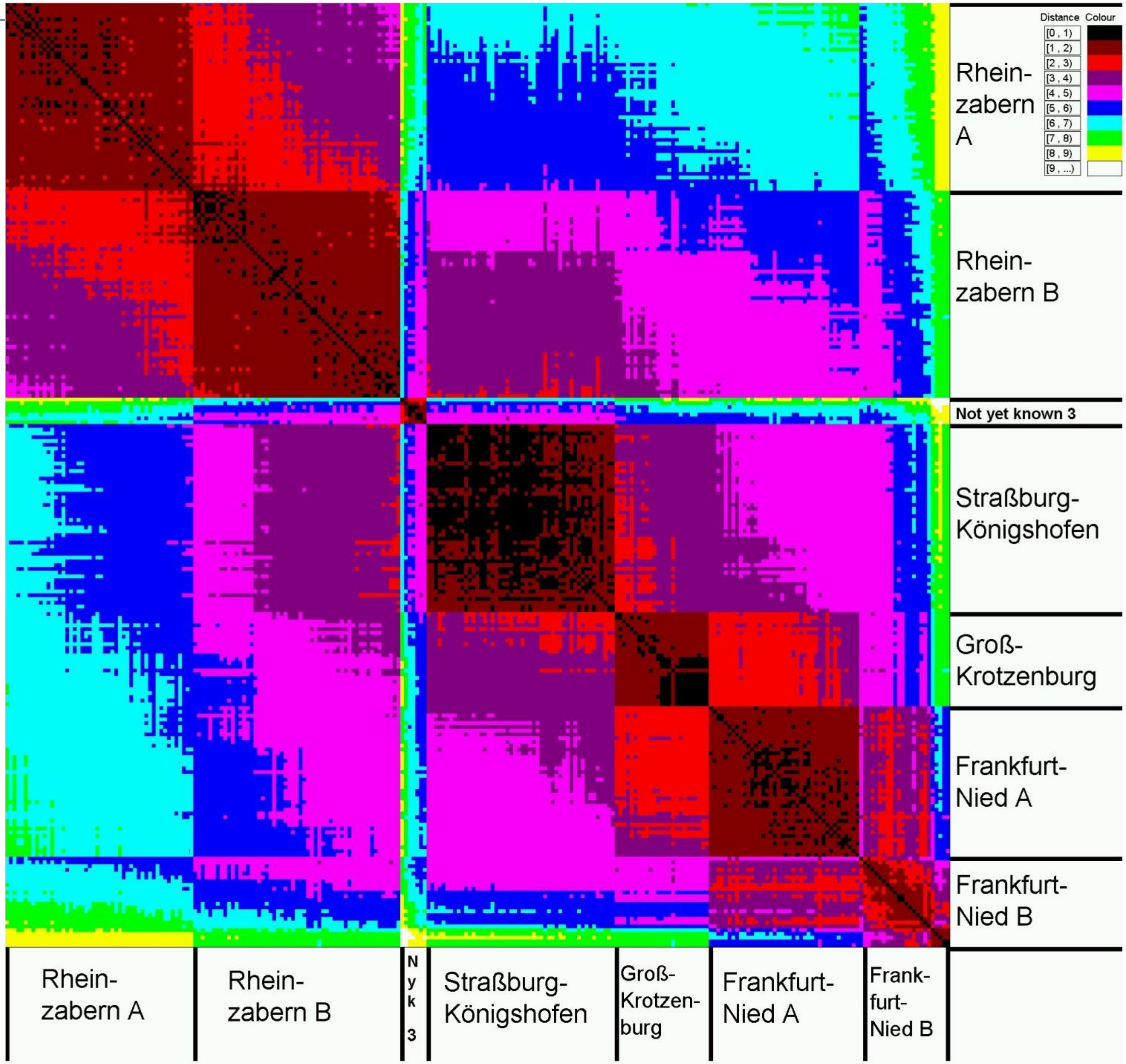
Cluster k Object i	1	2	$\mathbf{g}$
1	1	0	1
2	1	0	1
3	1	0	1
4	1	0	1
5	1	0	1
6	1	0	1
7	0	1	2
8	0	1	2

Validation of clusters

Introduction

A heatmap of pairwise distances allows two-dimensional visualisations of arbitrary high-dimensional data.

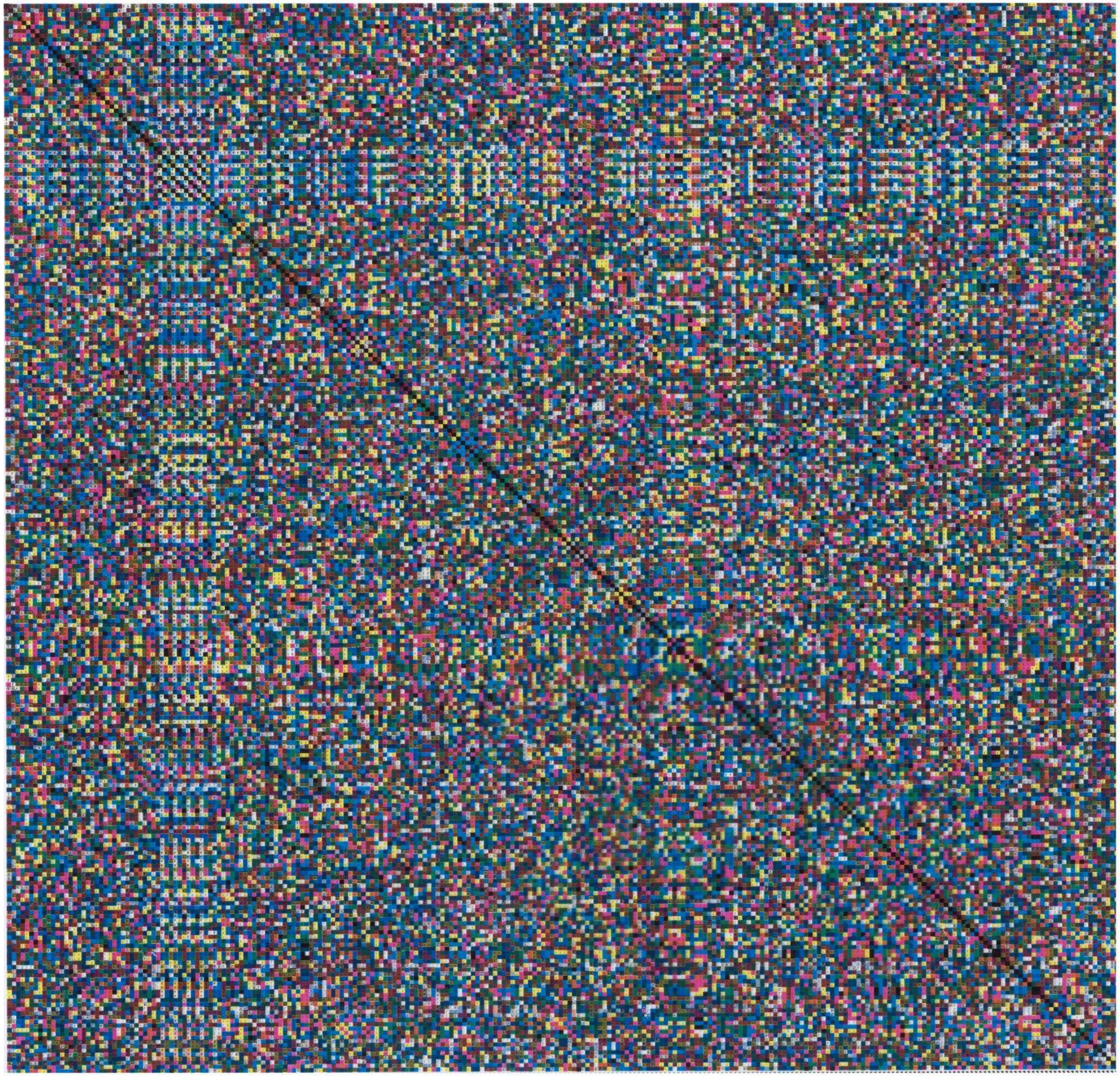
Tiles (same data as before): Ordered heatmap of the Euclidean distance matrix $\mathbf{D}$ between the observations.



Introduction

A heatmap of the Euclidean distances of random generated 20 dimensional Gaussian data.

A built-in analysis function of Excel 2003 was used here to generate the Gaussian data. Obviously, these generated data are not random.



Hierarchical and partitional clustering

To handle both ***hierarchical and partitional*** methods at the same time in an unique fashion, we would like to focus on Gaussian model-based cluster analysis in its simplest setting. Concretely, the ***sum of squares*** (SS) has to be minimised. (By the way, a similar formulation can be derived for the logarithmic SS criterion.)

Starting point is a distance matrix $\mathbf{D}_{\mathrm{Q}} = (d_{il})$ with pairwise ***(weigthed) squared euclidean distances*** as elements:

$$d_{\mathbf{Q}}(\mathbf{x}_i, \mathbf{x}_l) = (\mathbf{x}_i - \mathbf{x}_l)^T \mathbf{Q}(\mathbf{x}_i - \mathbf{x}_l).$$

Here the matrix $\mathbf{Q}$ is diagonal with weights q_{jj} >0. The SS criterion V_k based on $\mathbf{D}_{\mathrm{Q}}$ has to be minimised with respect to a fixed number of clusters K:

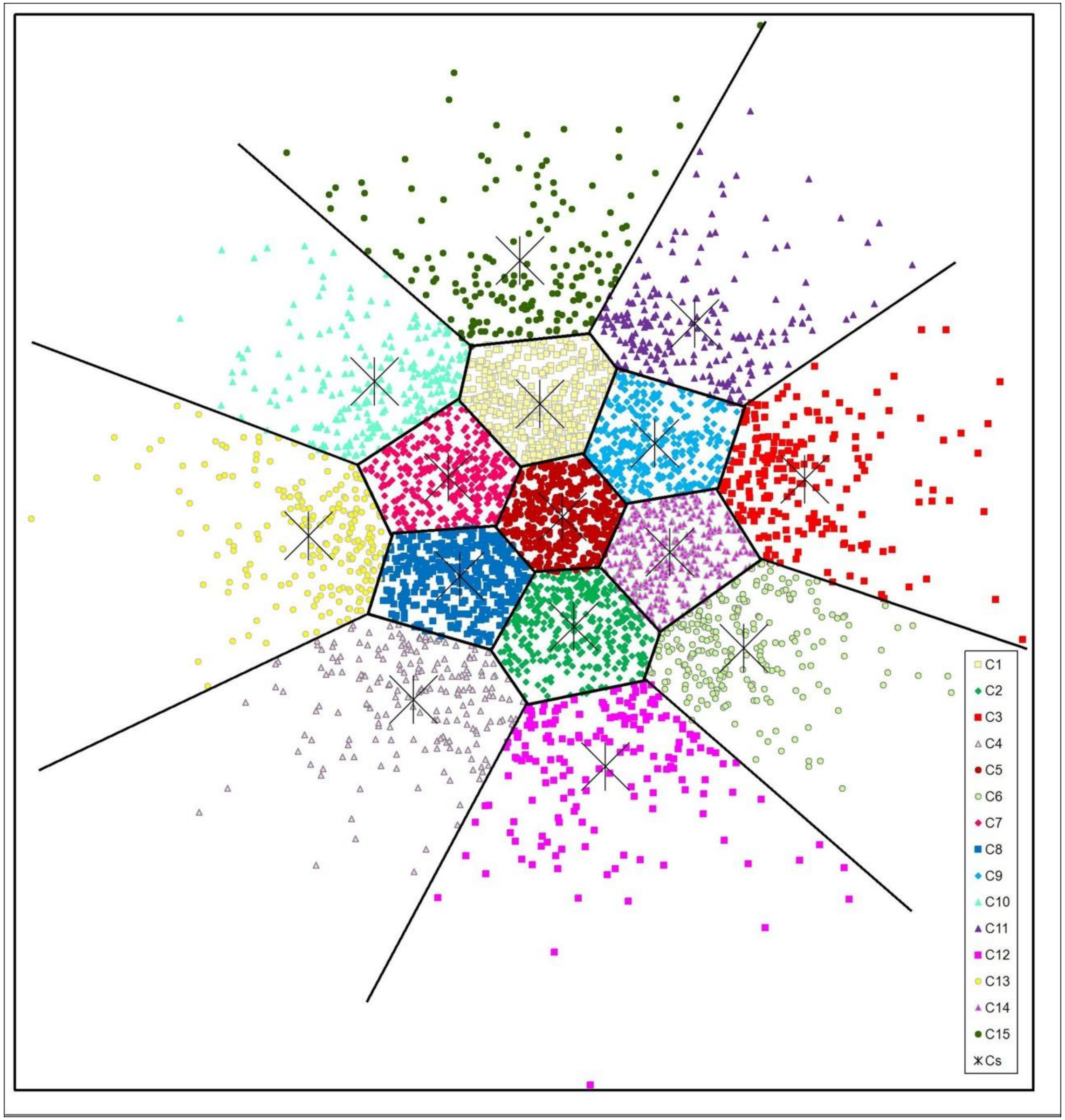
$$\boxed{V_K = \sum_{k=1}^{K} \frac{1}{M_k} \sum_{i \in C_k} m_i \sum_{\substack{l \in C_k \\ l > i}} m_l d_{\mathbf{Q}}(\mathbf{x}_i, \mathbf{x}_l)}$$

(m_i and M_k : weight of object i and cluster C_k, respectively) .

Hierarchical and partitional clustering

Scatterplot of the result of the partitional K means clustering of bivariate Gaussian data without a cluster structure into $K = 15$ clusters. The result is a so-called Voronoi (or Dirichlet) tessellation. In the plot, additionally, the cluster centres are marked by stars.

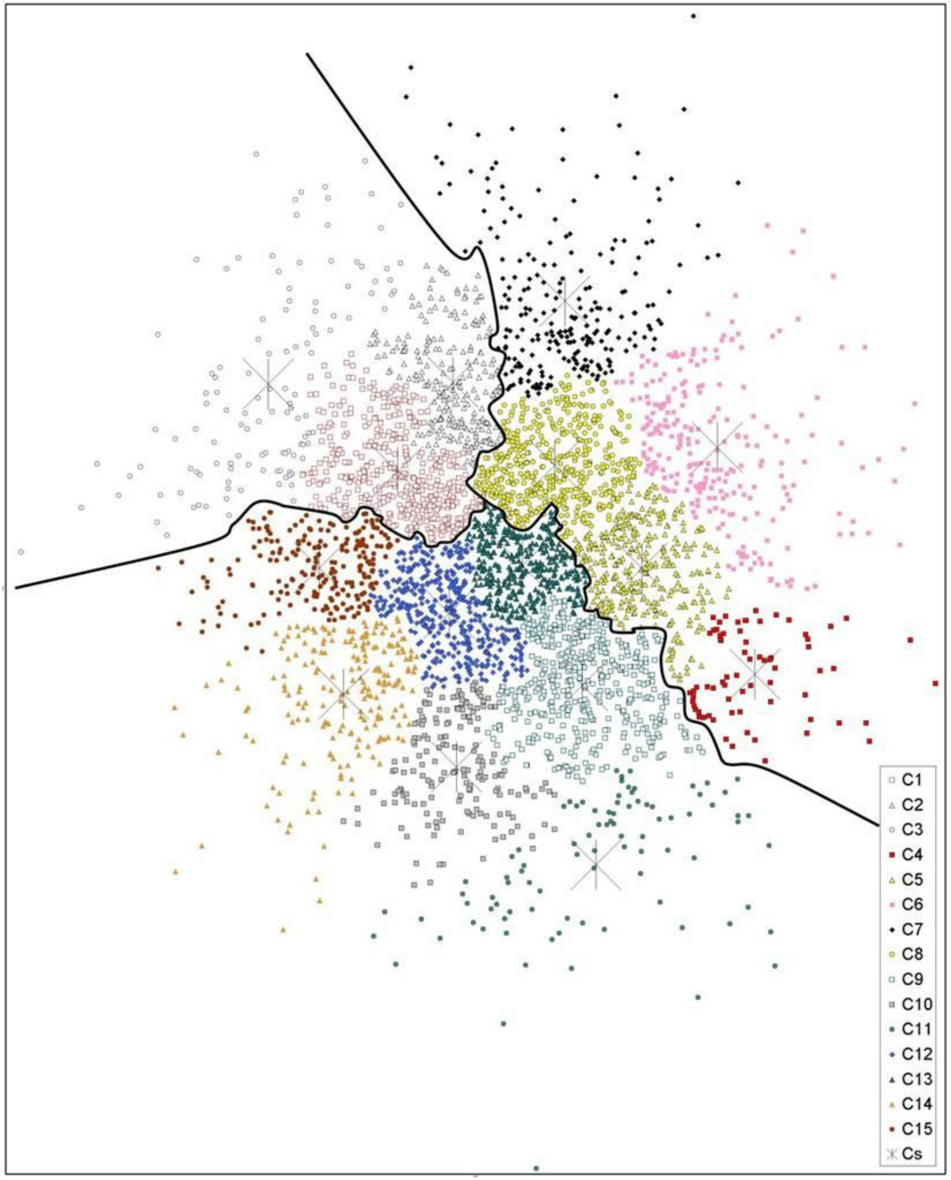
(Criterion used for clustering: minimum sum of squares.)



Hierarchical and partitional clustering

Scatterplot of two results of the hierarchical ***Ward*** method of bivariate Gaussian data without a cluster structure (same data as before). The solution for $K = 15$ clusters is marked by colour, the three cluster solution by freehand lines. Obviously, quite “different kinds” of clusters occur compared to the K means clustering.

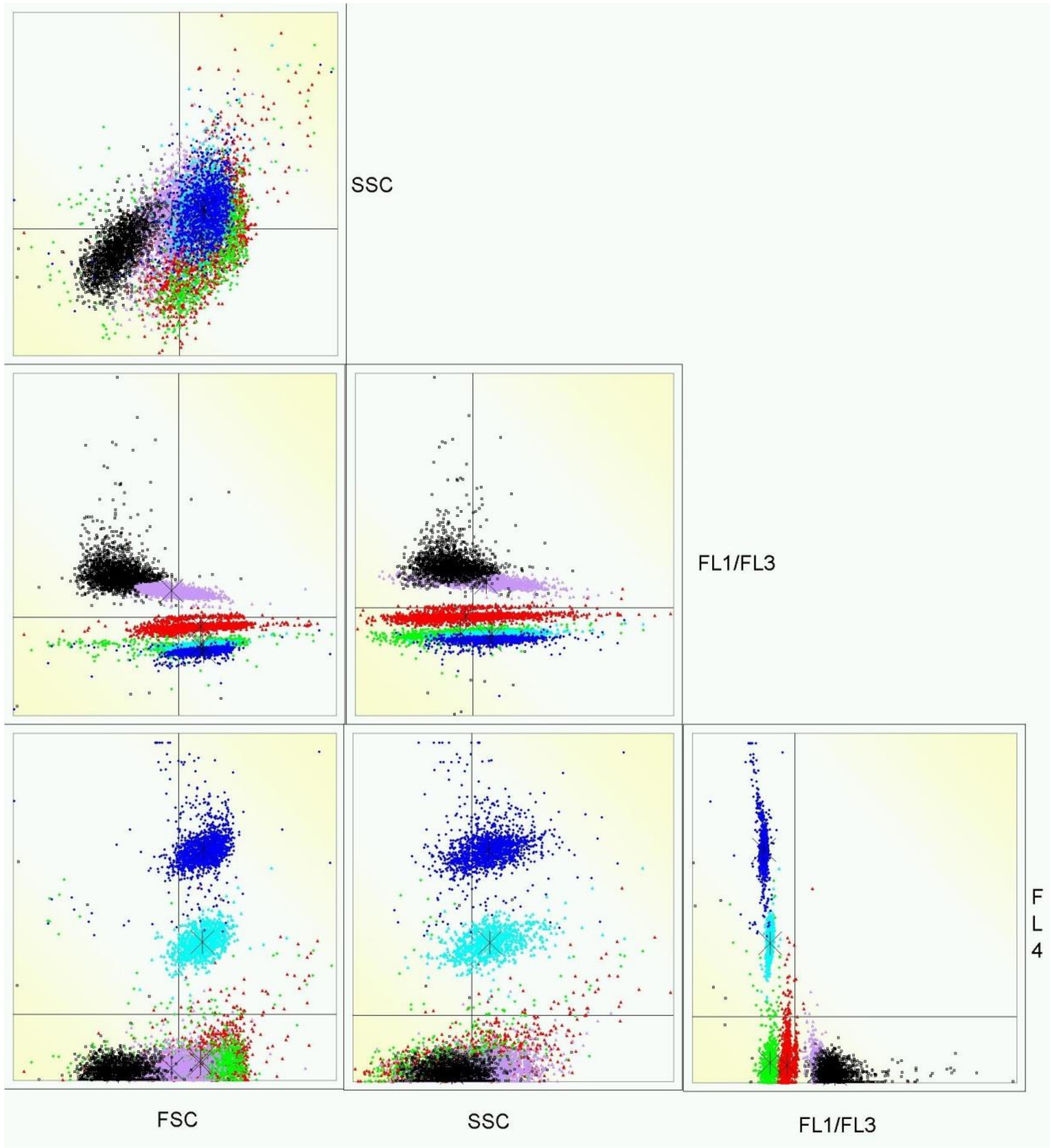
(Same criterion as before: minimum sum of squares.)



Multivariate projection

A scatterplot matrix can be recommended when the number of variables ranges from 3 to low numbers such as 20.

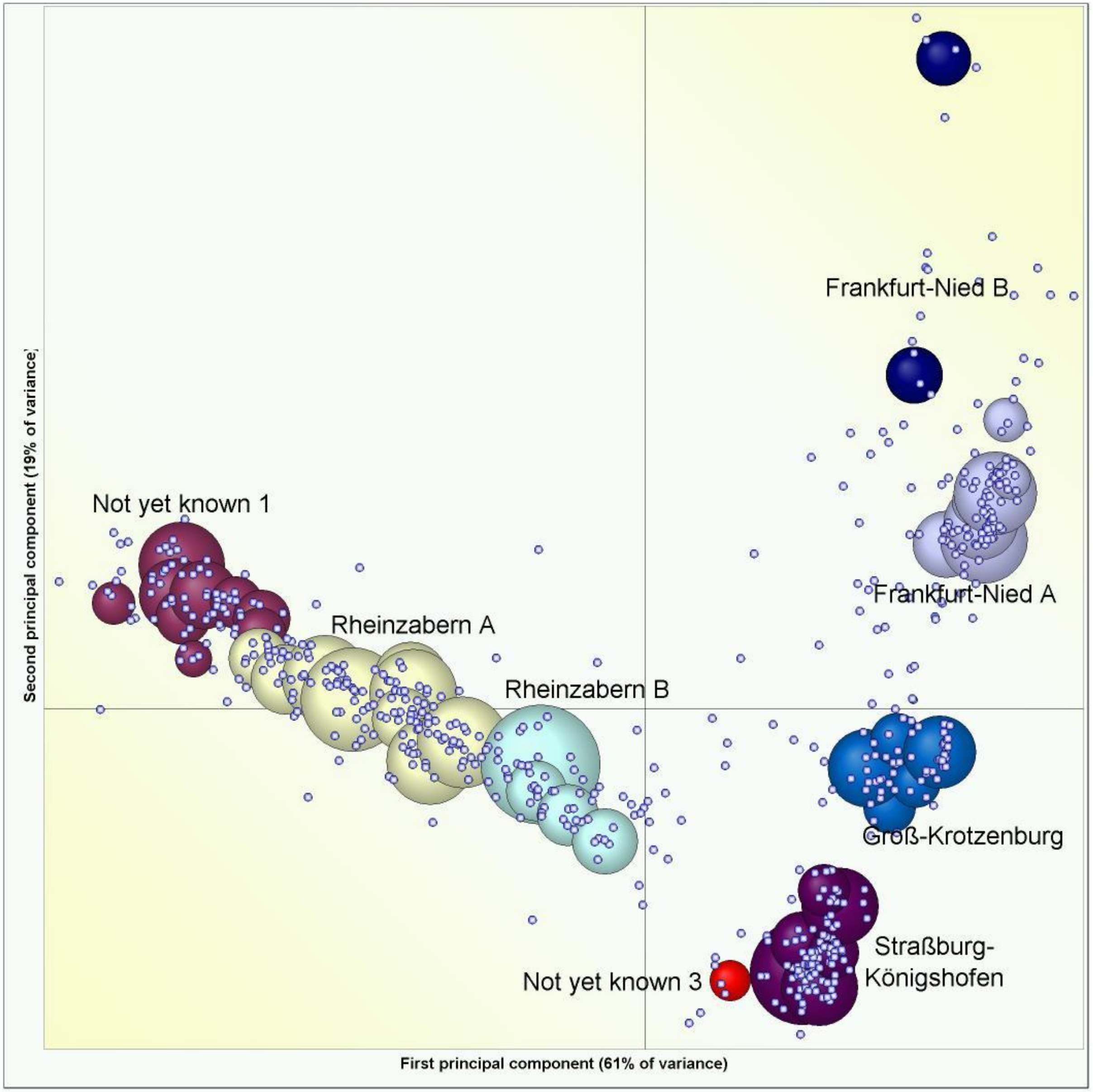
Application to ecology: Visualisation of particles (in water) based on measurements of their fluorescence properties. Their cluster membership is marked by colour (Mucha et al. 2002: WIAS Technical Report **5**).



Multivariate projection

Principal components analysis (PCA) and linear discriminant analysis are well-known multivariate projection techniques. The latter requires information about classes.

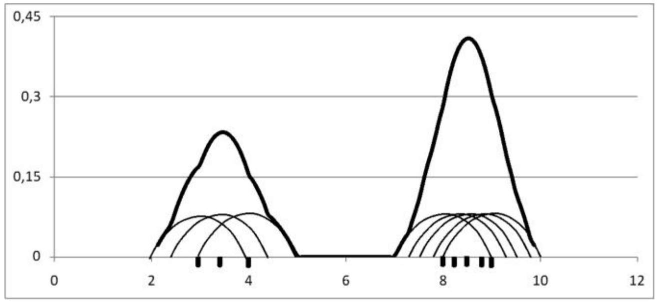
Tiles: Simultaneous PCA plot of observations (marked by points), dense regions (bubbles), and the final clusters (colour).

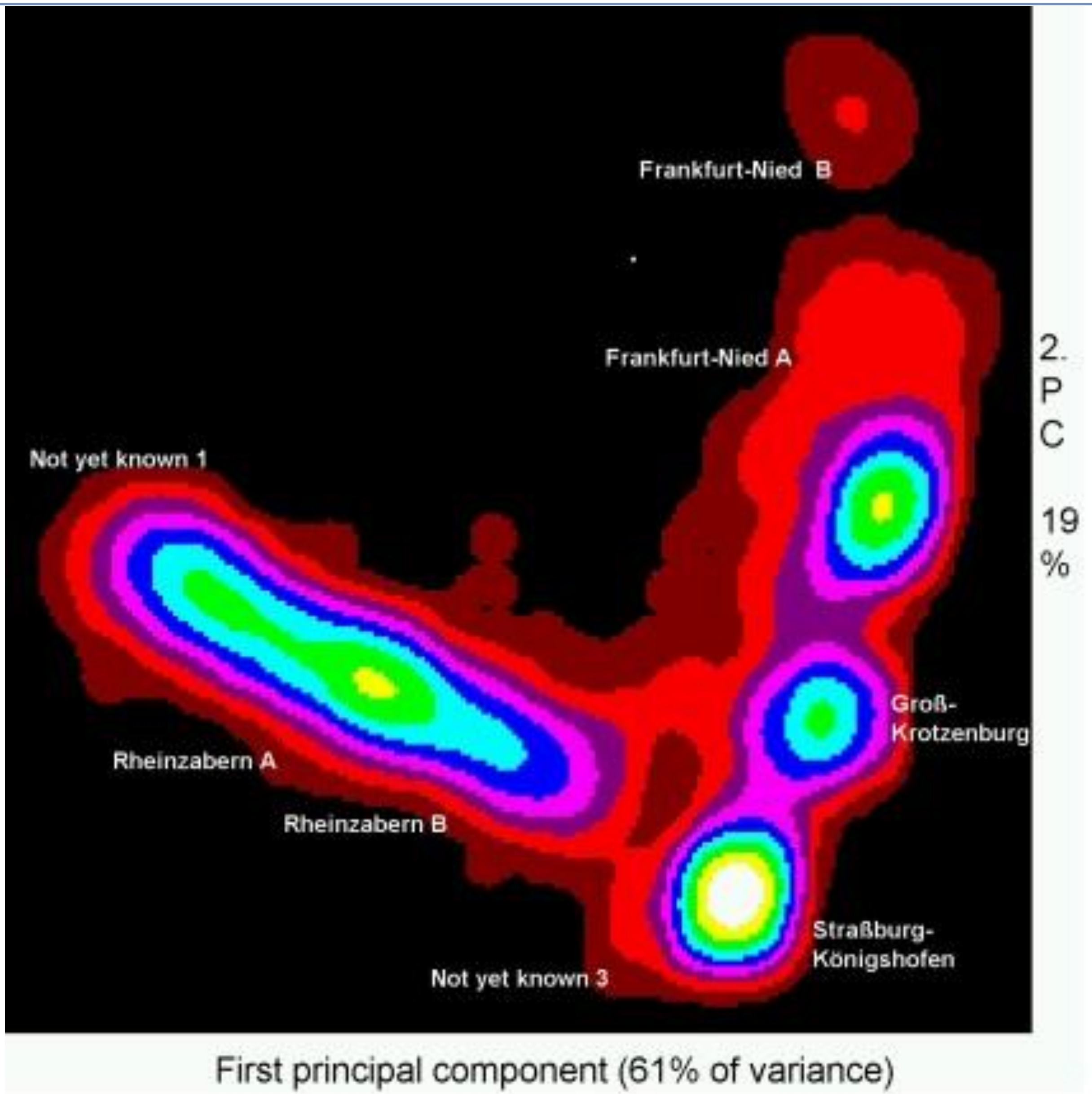


Multivariate projection

Tiles: Several cuts of a non-parametric density estimate in the plane of the first two principal components (PC): this is another (smooth) view at the data.

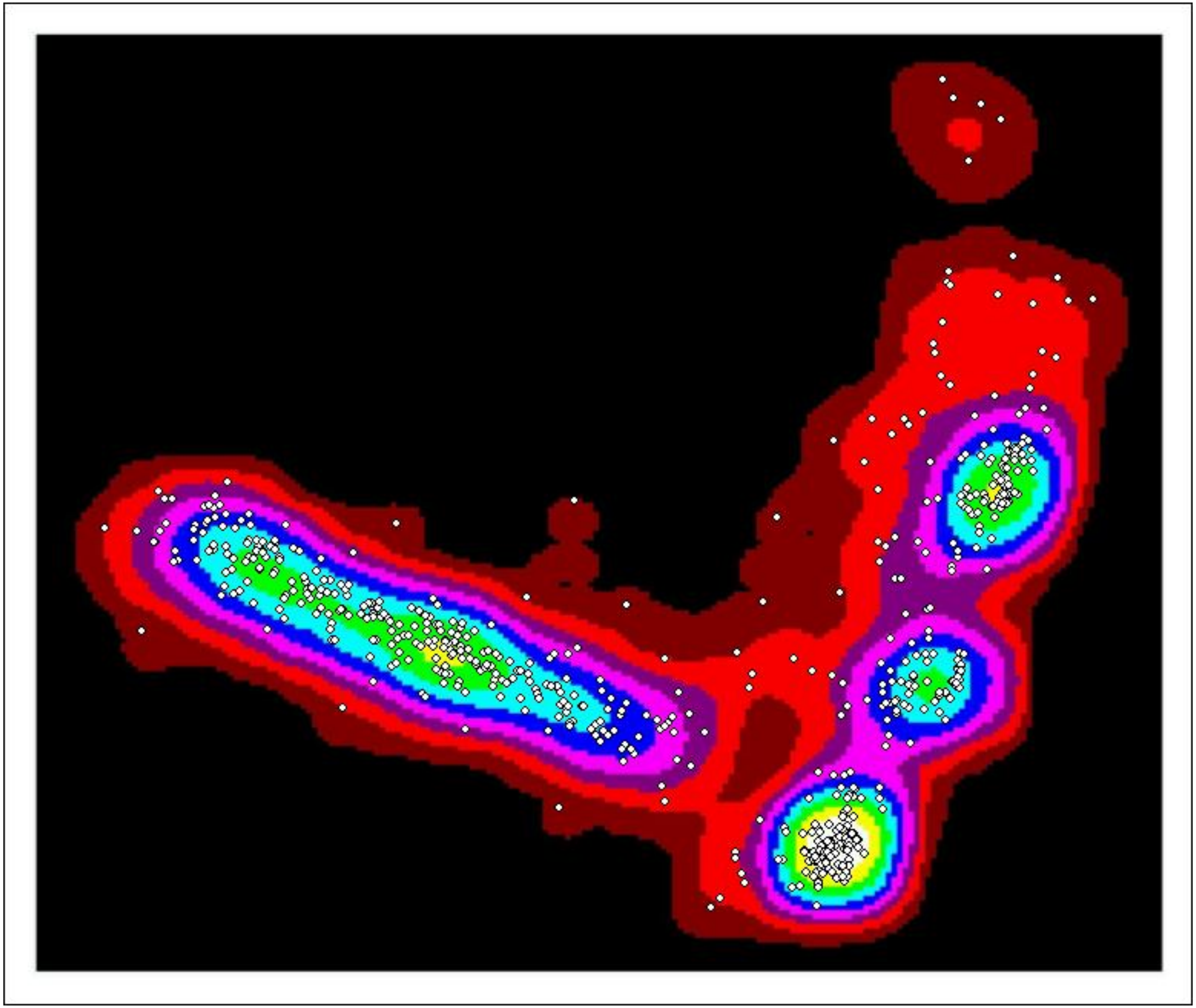
The idea of density estimation (smoothing) is the superposition of elementary kernels:





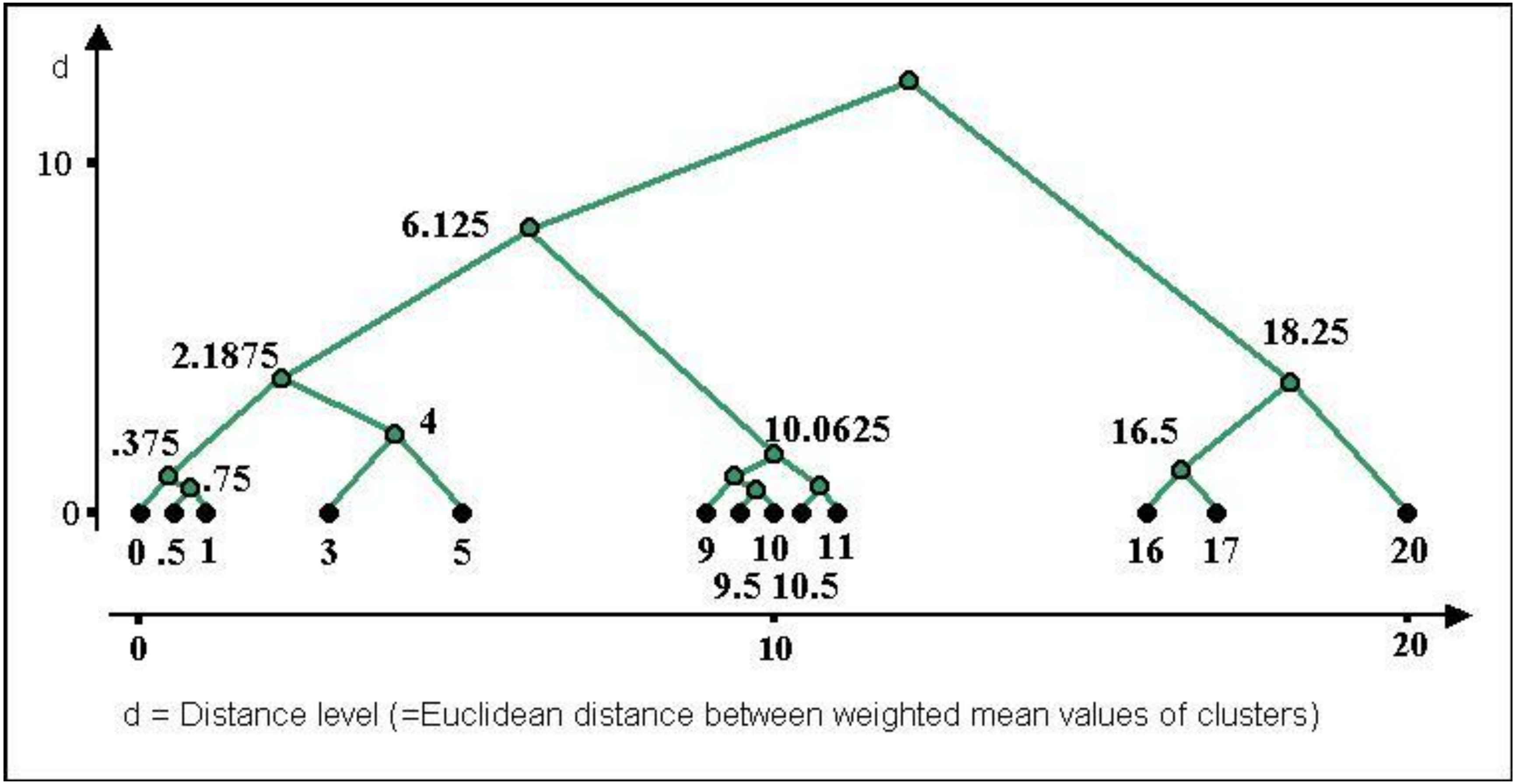
Bivariate density estimation

Tiles: from individual observations (points) to a smooth bivariate non-parametric density. The latter was cut at different levels.



Dendrograms

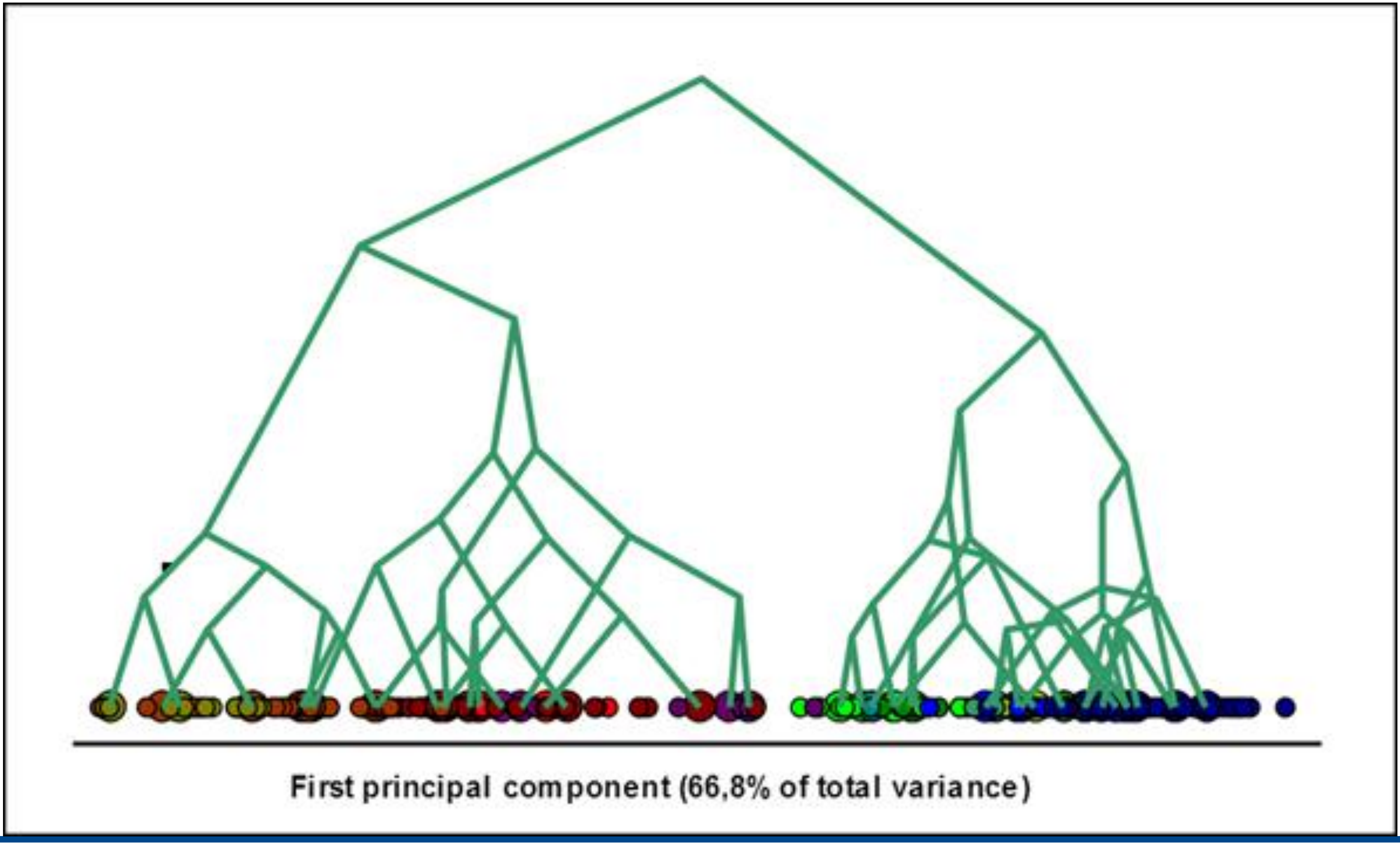
Dendrograms report about the process of hierarchical clustering.
Example: Non-equidistant dendrogram of Centroid clustering of 20 data point {0, 0.5,1, 3, 5, …, 20} on the real line (abscissa).



d = Distance level (=Euclidean distance between weighted mean values of clusters)

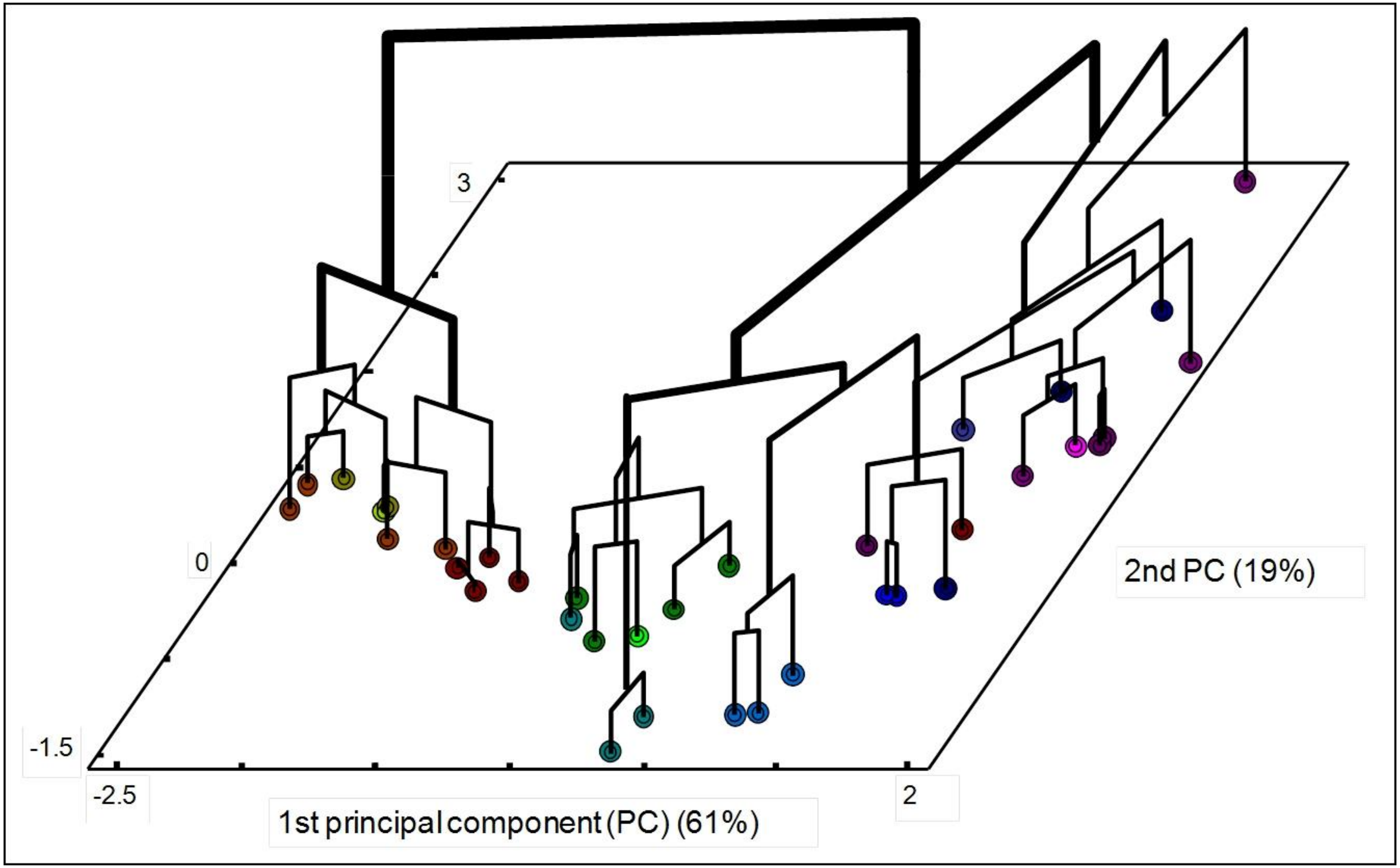
Dendrograms

Example: Non-equidistant dendrogram of the hierarchical ***Ward*** clustering of 200 Swiss bank notes based on 6 measurements. The genuine bank notes on the right hand side look more homogeneous than the forged ones.



Dendrograms

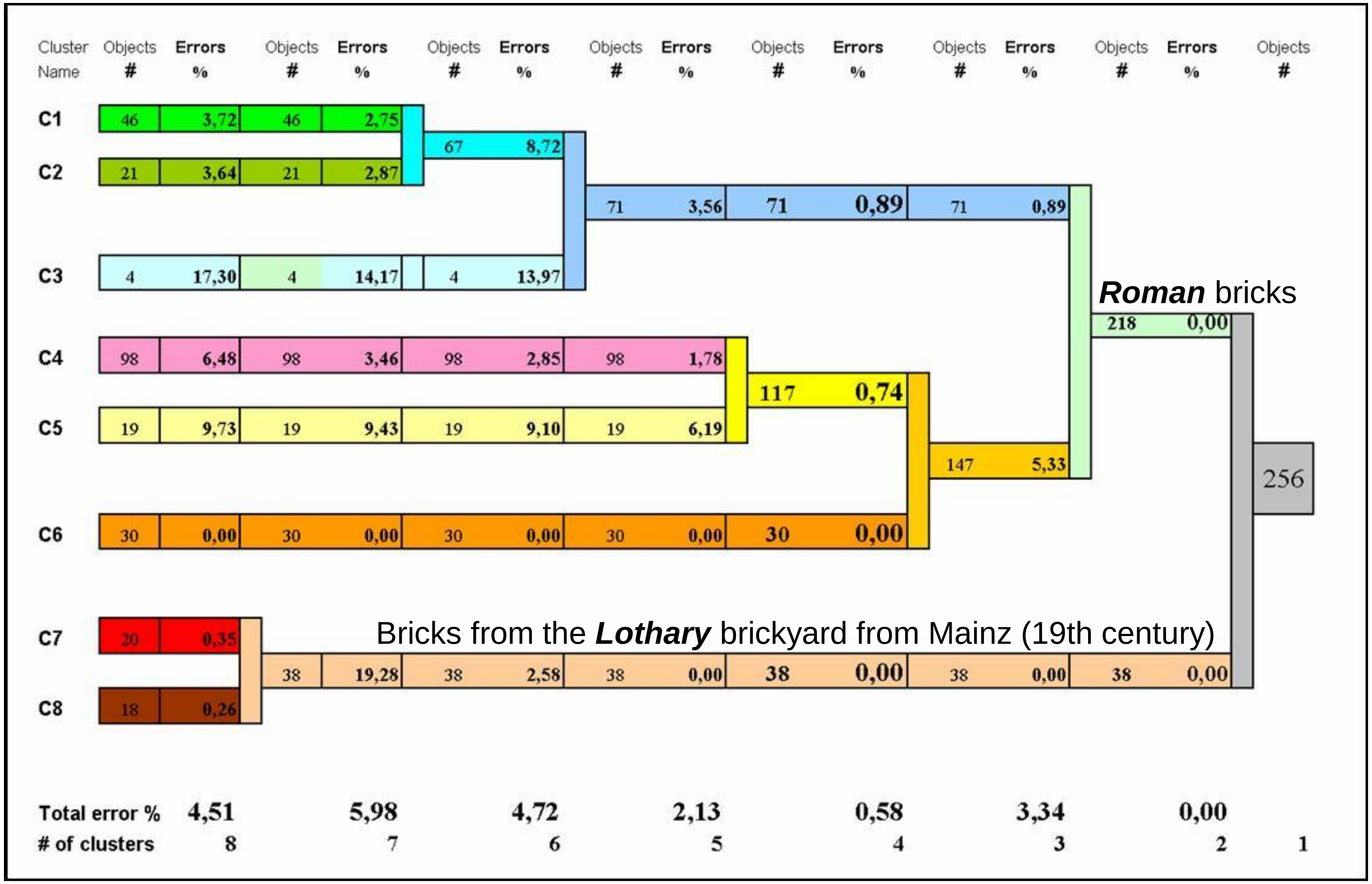
Plot-dendrogram of ***Ward*** ’s clustering of Roman tiles. The (reduced) dendrogram is projected onto the plane of the first two PC.



Dendrograms

Informative dendrograms are ordered binary trees that show additional information such as *Jaccard*'s stability values. To define stability with respect to the individual clusters, the following measures of similarity between two clusters (sets) E and F are used: $\gamma(E,F)=\frac{|E\cap F|}{|E\cup F|}$ (Jaccard), $\tau(E,F)=\frac{2|E\cap F|}{|E|+|F|}$ (Dice), and $\eta(E,F)=\frac{|E\cap F|}{|E|}$.

Roman bricks versus modern ones. Ward's hierarchical clustering clearly separates Roman bricks from modern ones. But, what about the stability of the clusters? In the dendrogram, the results of validation are presented using the "rate of recovery" η. (η is related to the "errors" by (1- η)*100%.) Here, 250 bootstrap clusterings were compared with the original clustering.



Cluster Name	Objects #	Errors %	Objects #	Errors %	Objects #	Errors %	Objects #	Errors %	Objects #	Errors %	Objects #	Errors %	Objects #	Errors %	Objects #
C1	46	3,72	46	2,75	67	8,72	71	3,56	71	0,89	71	0,89	218	0,00	256
C2	21	3,64	21	2,87											
C3	4	17,30	4	14,17	4	13,97									
C4	98	6,48	98	3,46	98	2,85	98	1,78	117	0,74	147	5,33			
C5	19	9,73	19	9,43	19	9,10	19	6,19							
C6	30	0,00	30	0,00	30	0,00	30	0,00	30	0,00					
C7	20	0,35	38	19,28	38	2,58	38	0,00	38	0,00	38	0,00	38	0,00	
C8	18	0,26													
Total error %	4,51		5,98		4,72		2,13		0,58		3,34		0,00		
# of clusters	8		7		6		5		4		3		2		1

Roman bricks
Bricks from the *Lothary* brickyard from Mainz (19th century)

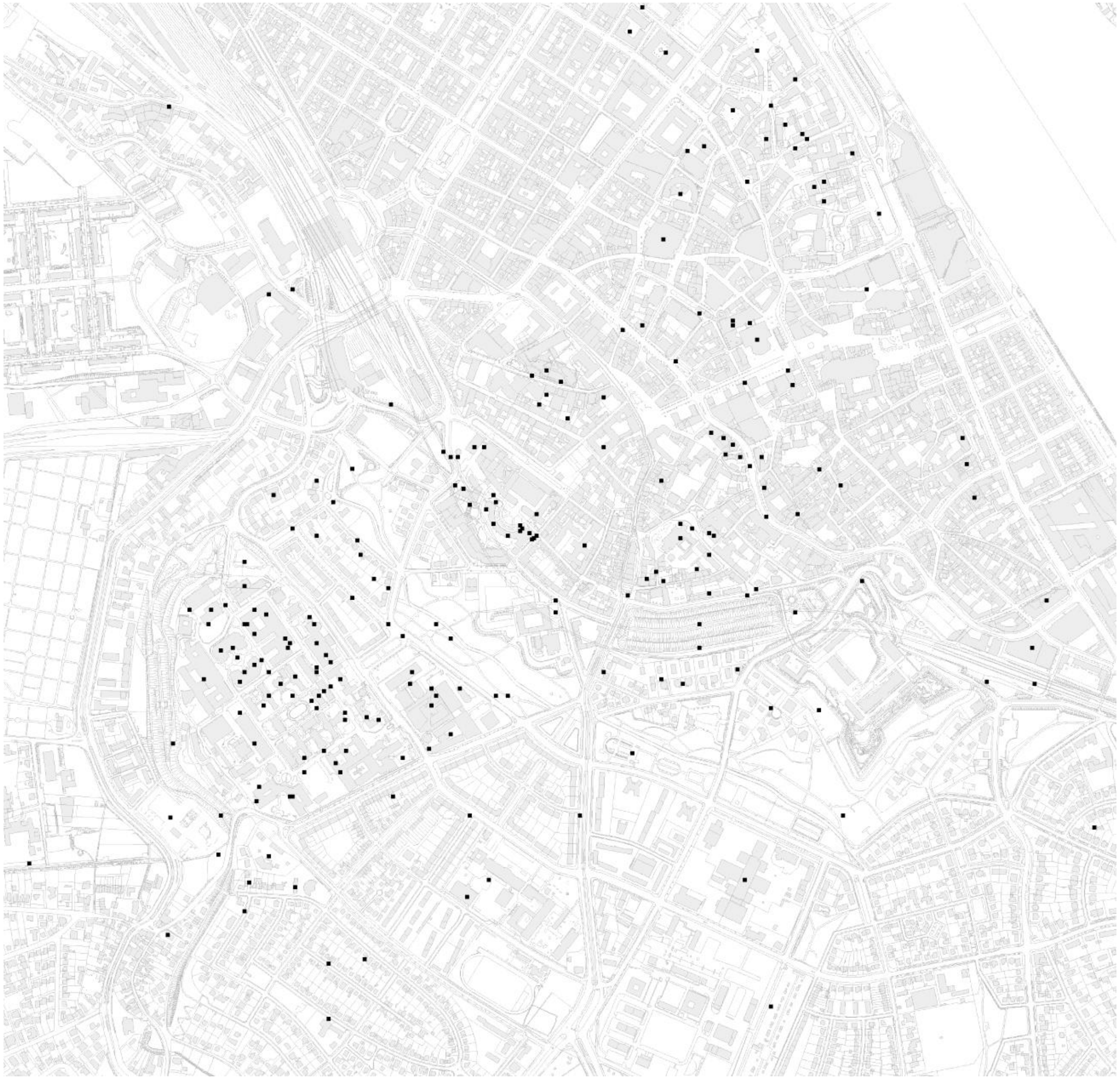
Mapping findspots

Spatial and statistical analysis of findspots of Roman military brickstamps in Mainz. The background:

- The findings come from the first four centuries A.D. The Roman bricks and tiles have been classified based on their stamp. In addition, new types of stamps have been defined.
- The database itself is under development yet: many hundreds of findings are not yet identified by their coordinates of location.
- Mainz has been a great urban area for thousands of years. A planned excavation is not possible. Therefore most of the places of discovery are found by "accidents" such as the excavation around the main railway station.

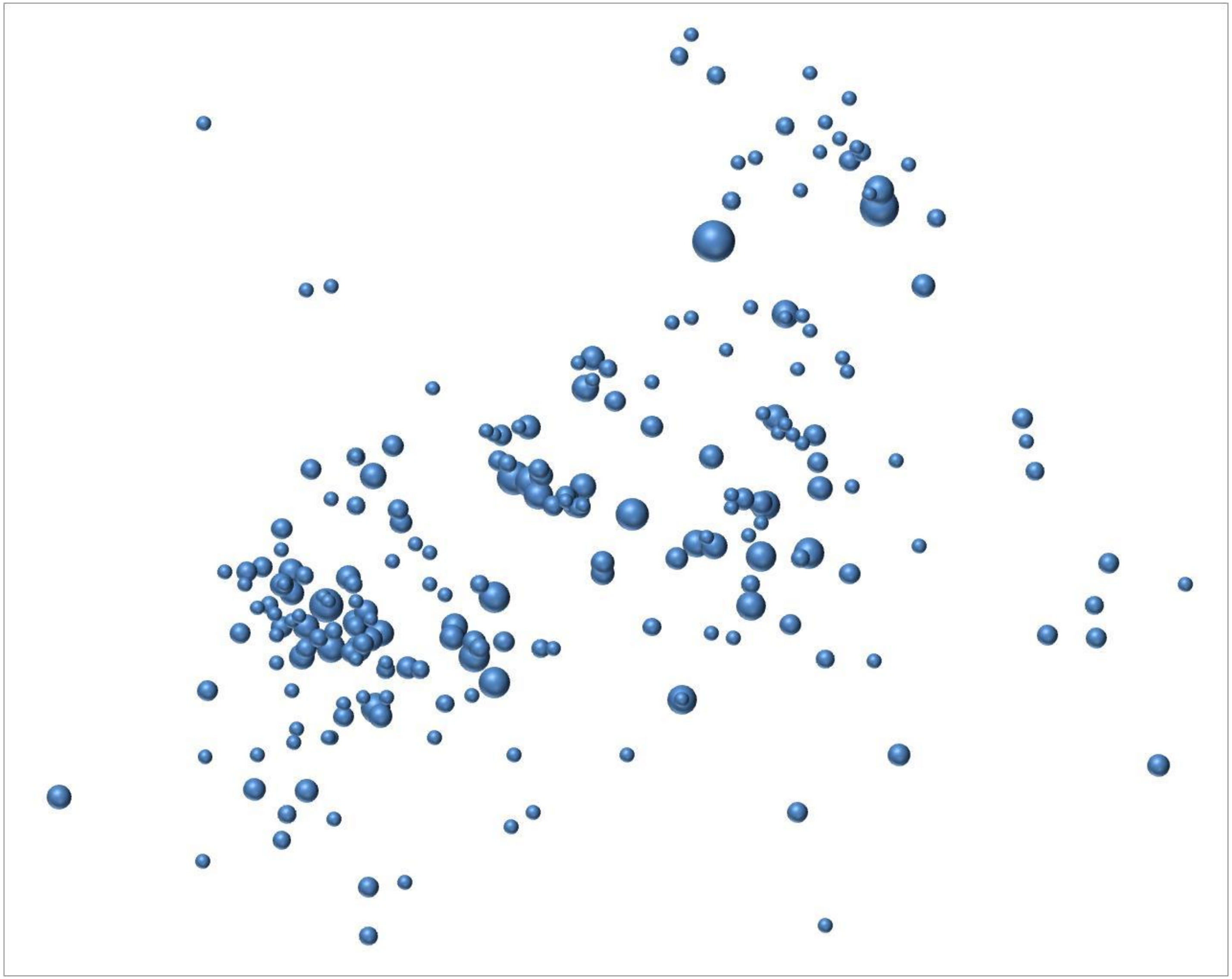
Mapping findspots

Example of a mapping of a zoomed area of Mainz (city) with locations of findspots. The river Rhine is located in the right upper corner. All in all there are 246 sites until now.



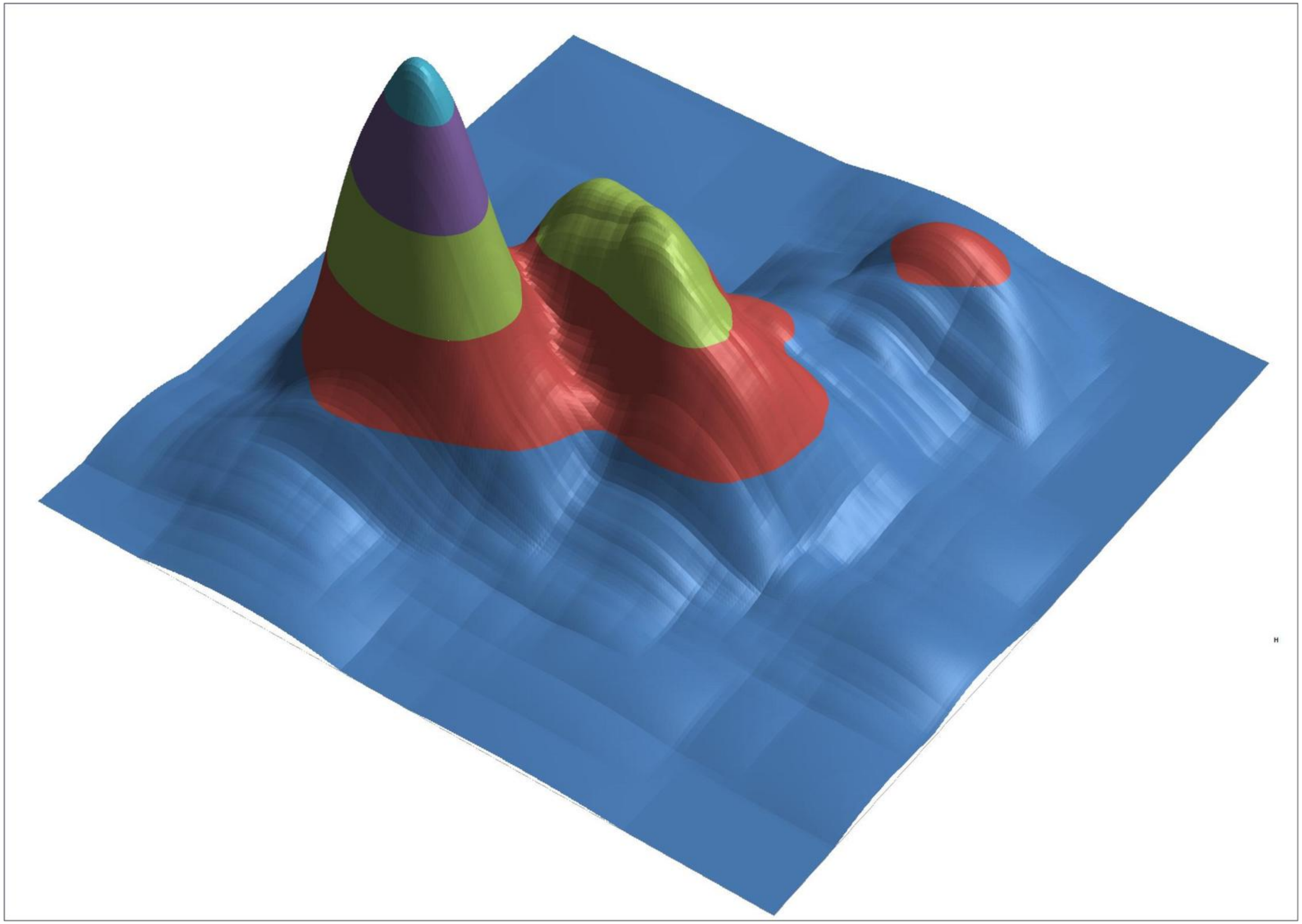
Mapping findspots

The locations of findspots are marked by bubbles. The size of a bubble is proportional to the number of findings at the location. So the plot becomes more informative.



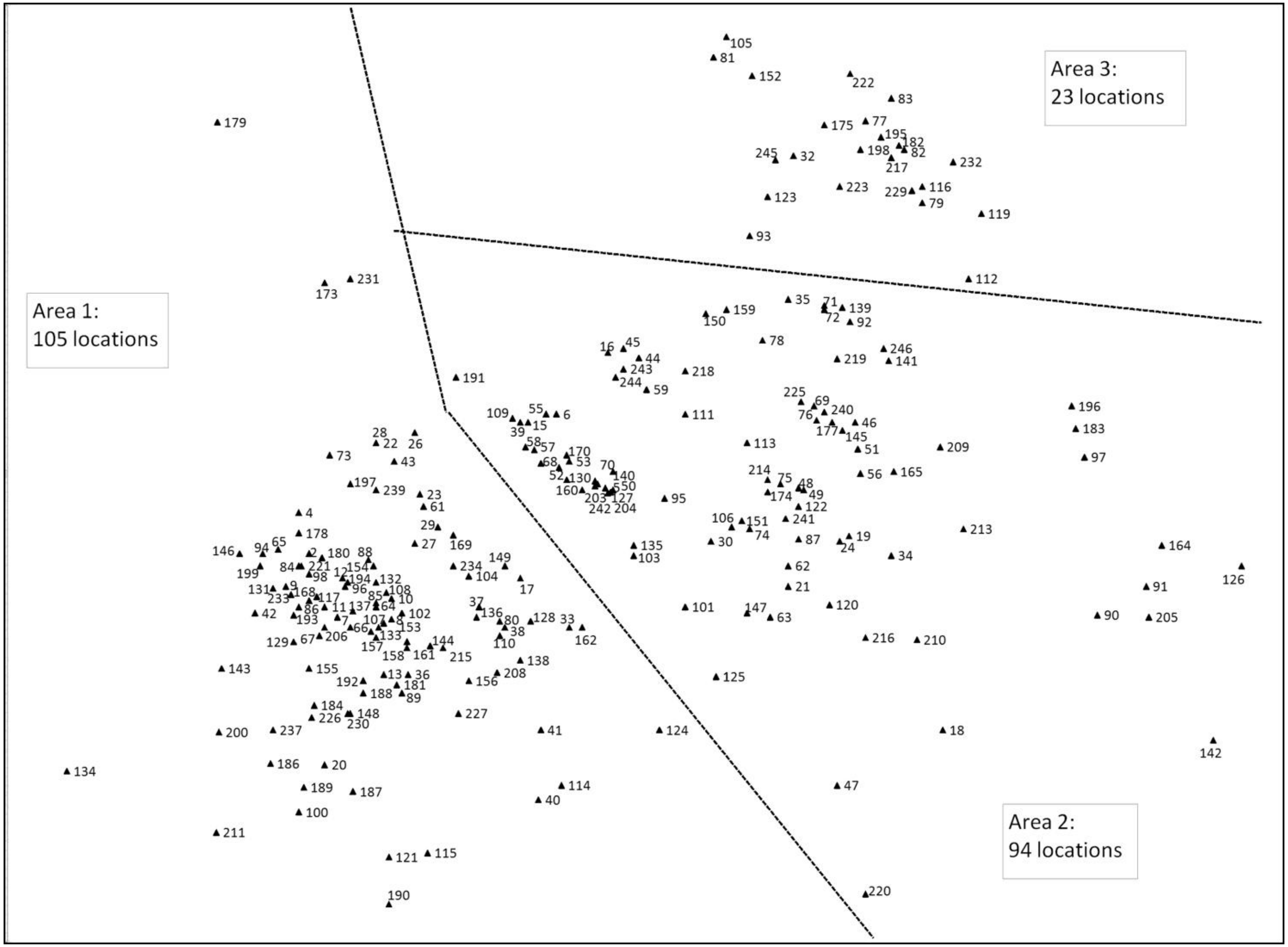
Mapping findspots

Bivariate density estimation of all locations. Herein they are weighted by the logarithm of the number of findings plus 1.



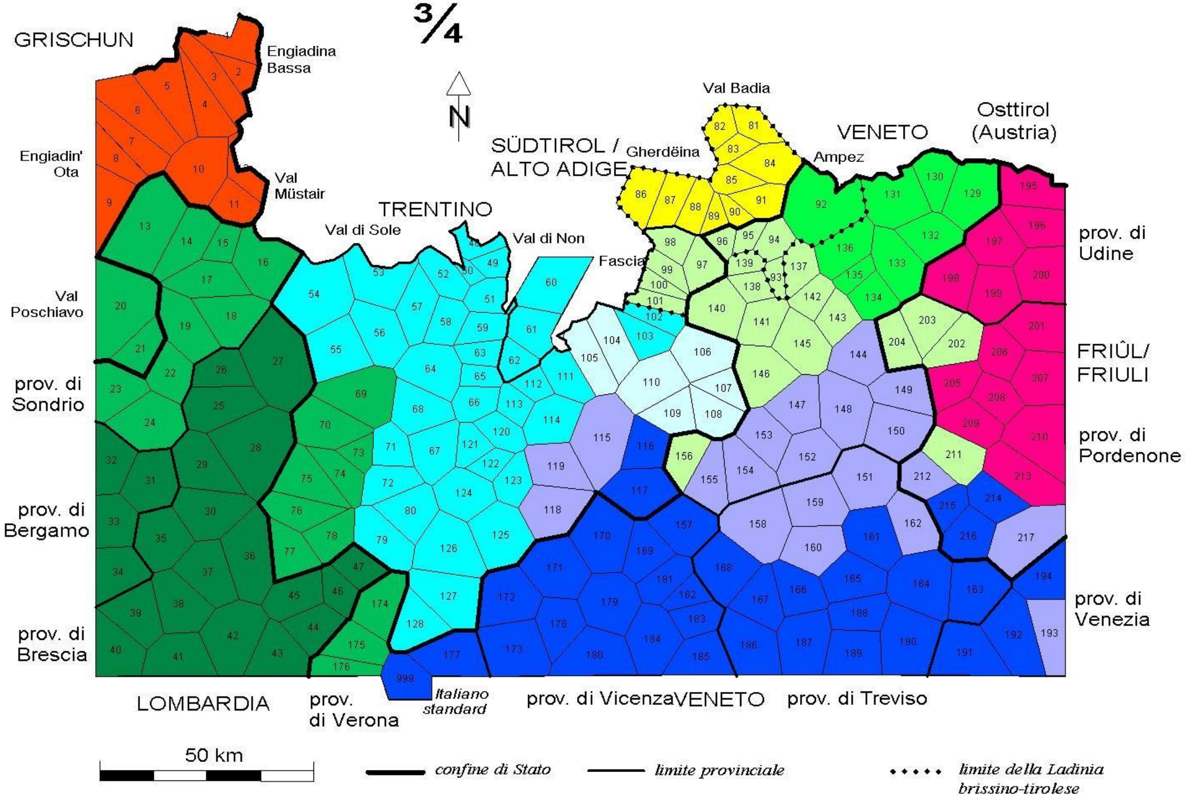
Mapping findspots

Archaeological dissection of the ancient Mainz into the three main spatial regions *Area 1*, *Area 2*, and *Area 3*. It turns out that these regions coincide with three time periods significantly.



Dialectometry: Mapping of cluster analysis results

The visualisation of the reliability of the cluster membership of each individual observation is of special interest in the field of geographical mapping. The reliability can be assessed based on the investigation of stability of individual clusters. The figure shows an application to quantitative linguistics where significant regions (95% level) in North Italy are in dark colour (Mucha and Haimerl 2005).



Summary

Visualisation is essential for a better understanding of data analysis and cluster analysis.

Thank you for your attention!

Bibliography

Mucha, H.-J., Bartel, H.-G. and Dolata, J. (2005): Techniques of rearrangements in binary trees (dendrograms) and applications. MATCH Commun. Math. Comput. Chem., 54, pp. 561--582.

Mucha, H.-J. and Ritter, G. (2009): Classification and clustering: Models, software and applications. WIAS Report No. 26, WIAS, Berlin.