Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Report ISSN 0946 – 8838

Classification and Clustering: Models, Software and Applications

Hans–Joachim Mucha[1] and Gunter Ritter[2] (Eds.)

submitted: September 17, 2009

[1] Weierstraß-Institut für Angewandte Analysis und Stochastik
Mohrenstr. 39
10 117 Berlin
E-Mail: mucha@wias-berlin.de

[2] Fakultät für Informatik und Mathematik
Universität Passau
Innstr. 33
94 032 Passau
E-Mail: ritter@fim.uni-passau.de

No. 26
Berlin 2009

1991 *Mathematics Subject Classification.* 62-07, 62H30, 62H25, 62P10, 90-08.

Key words and phrases. classification, cluster analysis, dimension reduction, spectral clustering, visualization.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Foreword and Acknowledgments

We are pleased to present the report on the 30th Fall Meeting of the working group "Data Analysis and Numerical Classification" (AG–DANK) of the German Classification Society. The meeting took place at the Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Berlin, from Friday Nov. 14 till Saturday Nov. 15, 2008. Already 12 years ago, WIAS had hosted a traditional Fall Meeting with special focus on classification and multivariate graphics (Mucha and Bock, 1996). This time, the special topics were stability of clustering and classification, mixture decomposition, visualization, and statistical software.

The working group AG–DANK of the German Classification Society ("Gesellschaft für Klassifikation," GfKl) deals with all statistical, mathematical, and computational aspects of data analysis and classification problems (clustering, discriminant analysis, supervised/unsupervised classification, pattern recognition, data mining) and with their applications in the sciences, the economy, engineering, archaeometry, and the administration. The GfKl, founded in the year 1977, is a transdisciplinary scientific society that aims at promoting methods of classification and data analysis in theory and application.

The editors and the working group AG–DANK would like to thank all who have contributed to this report. Our special thanks go to the head of WIAS for their active support and thorough preparation of the event.

Gunter Ritter
Chair of AG-DANK

Hans–Joachim Mucha
Local organizer

References

MUCHA, H.-J. and BOCK, H.-H. (Eds.) (1996): Classification and multivariate graphics: models, software and applications. Report no. 10, WIAS, Berlin.

Preface

During the 30th Fall Meeting of the working group AG-DANK a dozen talks were presented. Among them, three discussion papers dealt with statistical analyses of special data sets issued in advance. All in all 16 participants contributed to the meeting with talks and discussions, see the Appendix (Part III below). Here follows the list of the talks and analyses.

Part I: Talks

- Peter Kurz, TNS Infratest, München: Stability of "clustering on clusters"-methods in the field of marketing
- Hans–Joachim Mucha, WIAS Berlin: `ClusCorr98`® for Excel 2007: Clustering, Multivariate Visualization, and Validation
- Gunter Ritter, Uni Passau: Resolving ambiguity in segmentation problems by the method of variants
- Hans–Georg Bartel, Humboldt-Universität zu Berlin: Archäometrische Daten römischer Ziegel aus *Germania Superior*
- Marcus Weber, ZIB Berlin: Spectral Clustering
- Susanna Röblitz, ZIB Berlin: Clustering of high-dimensional data by domain decomposition methods
- Christian Hennig, University College London: Merging Gaussian mixture components - an overview
- Florian Meyer, Uni Marburg: Interpretable models of leptokurtic distributions
- Anne Spickenheuer, BGFA an der Ruhr-Universität Bochum: Classification of workers with different exposure levels to fumes of bitumen

Part II: Data Analyses

- Roman Tiles Data Set
  - Gerhard Pöppel and Reinhard Schachtner, Infineon, Regensburg: Analysis I by Projection Pursuit
  - Gunter Ritter: Analysis II by Model Based Clustering
  - Susanna Röblitz and Marcus Weber: Analysis III by Spectral Clustering
  - Hans–Georg Bartel, Hans–Joachim Mucha and Jens Dolata: Comparison of Results for the the Tiles Data

- Synthetic Data Set 1
  - Gunter Ritter: Structure of the synthetic data set Berlin08_synth1
  - Gerhard Pöppel: Analysis A by MCLUST
  - Susanna Röblitz and Marcus Weber: Analysis B by Spectral Clustering
  - Gunter Ritter and Hans–Joachim Mucha: Comparison of Results
- Synthetic Data Set 2
  - Gunter Ritter: Structure of the synthetic data set Berlin08_synth2
  - Gerhard Pöppel and Reinhard Schachtner: Analysis by Projection Pursuit
  - Gunter Ritter and Hans–Joachim Mucha: Comparison of Results

The present publication does not cover all talks presented at the Fall Meeting, but only a selection of contributions. They are listed in the order of their presentation at WIAS.

Contents

List of Figures

List of Tables

Part I
Papers of Talks

1 ClusCorr98® for Excel 2007: Clustering, Multivariate Visualization, and Validation

Hans–Joachim Mucha
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39, D-10117 Berlin
mucha@wias-berlin.de

Abstract

Really, in Excel 2007, you get much more because it now supports 1 048 576 rows and 16 384 columns. Such an Excel 2007 "Big Grid" spreadsheet is both the distinguished repository for data/distances and the perfect plotting board for multivariate graphics that can be composed in VBA-code (examples are dendrograms, plot-dendrograms, scatterplot matrices, density plots, principal components analysis plots,...), or can be obtained by "playing" with the properties of cells in cell-based graphics (pixel graphics) like the heat plot of a proximity matrix. Pairwise data clustering, and furthermore restricted in terms of the two simplest Gaussian model-based cluster analyses, will be considered in more detail here. That means, the usual estimation of expectation values of clusters is no more necessary. By using special randomized weights one can perform built-in validations of cluster analysis results by bootstrapping techniques [13].

Keywords: Cluster analysis, dimension reduction, statistical software, dendrogram, Excel, Visual Basic for Applications

Introduction

The Excel 2007 "Big Grid": spreadsheets with 64 times more columns (= 16 536) than before and about 1 million rows. Now a huge $I \times I$ proximity matrix can be stored directly in an Excel sheet for further usage (clustering, visualization,...). (Proximities are a general term for pairwise distances, similarities,...) Thus, the huge number of cells (= "pixels" in cell-based graphics) is up to now often far beyond the technical limits of screens and printers. Therefore one has to navigate inside the big pictures to look for interesting areas.

Concerning cluster analysis, the main focus here is on simple Gaussian model-based methods that can be generalized by (adaptive) weighting of the variables and/or the observations. It should be noted that the simple models are essential with respect to the subsequent use of complex Gaussian models because of an appropriate initialization of parameters and partitions [19]. In the following, pairwise distances are the starting point for cluster analysis, but also for appropriate multivariate

graphics of data and clusters by applying methods of projection, such as principal components analysis and multidimensional scaling.

Another main focus is on validation. The proposed built-in validation techniques can verify the results of the two most important families of methods, the hierarchical and partitional cluster analysis. The finding of the appropriate number of clusters, as the main task of model selection, is the ultimate aim here. The built-in validation evaluates additionally the stability of each cluster and the degree of membership of each observation to its cluster.

This paper outlines the complete life cycle of cluster analysis starting from raw data up to ending with both the validated results and the multivariate graphics. Some of the most successful applications come from quite different fields such as archaeometry (see below in Sect. 3 and Sect. 7 in this report, [16], [17]), computational linguistics [18], ecology ([19], [21]), economics [13], chemometrics [15], and demography (see Fig. 18 below).

1.1 The Statistical Software `ClusCorr98`®

`ClusCorr98`® is based on the programming environment Visual Basic for Application (VBA) in the host application Excel. One does not forget about the "A": sorting, removing duplicates, pivot tables, SQL for querying a database, working with ranges (= matrices, e.g. "$\mathbf{X}$ = selection.value" gets you a data matrix $\mathbf{X}$, and "selection.value = $\mathbf{D}$" saves a distance matrix $\mathbf{D}$),... would be much faster than any VB code or C++ code anyone could write. `ClusCorr98`® is much more than can be presented here, that is, this paper is designed as a kind of introduction only.

The statistical software performs exploratory data analysis mainly by using adaptive methods of cluster analysis, classification, and multivariate visualization. The main focus is on simple, stable models accompanied by appropriate multivariate (graphical) methods like principal components plots and informative dendrograms (binary trees). The "Big Grid" sheets are also a perfect plotting board for scatterplot matrices whose support often has been of value for the search of an appropriate cluster analysis model. Cell-based graphic is another kind of visualization in spreadsheets. Examples are heat plots (see Fig. 3 below), colored dendrograms (Fig. 16), or informative dendrograms (Fig. 20) that show the results of validation of clusters.

The software `ClusCorr98`® comes in an Excel 2003 format. However it can make use of the "Big Grid" spreadsheets of Excel 2007 by the VBA programming language. Moreover, there are new or improved built-in graphic tools of Excel 2007 which can do a lot for you, for a better understanding of the data at hand.

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

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

Figure 1: From a data matrix $\mathbf{X}$ to a distance matrix $\mathbf{D}$.

1.2 Pairwise Distances

Let us introduce into the problem of finding clusters (groups, subsets) of observations (objects) based on pairwise distances. Pairwise distances play also an important role in Sect. 3, Sect. 4, and Sect. 5 in this report. Without loss of generality the focus here is on clustering the observations. Clustering the variables can be done often in a similar way, for instance, in the case of binary data or contingency tables (see Fig. 10 below).

Let a sample of I independent observations (objects) is given in R^J and denote by $\mathbf{X} = (x_{ij})$ the corresponding data matrix consisting of I rows and J columns (variables), where the element x_{ij} provides a value for the jth variable describing the ith object. Objects can be species of plants, animals, individuals, archaeological findings, countries, enterprises, and so on. Further, let $\mathcal{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_I\}$ denote the finite set of the I objects (see at the outer left hand side of Fig. 1). Alternatively, let us write shortly $\mathcal{C} = \{1, \ldots, i, \ldots, I\}$.

Generally, the clusters that we are looking for should be as homogeneous as possible in some sense. Usually, cluster analysis techniques are applied in order to reach this aim. Fig. 1 presents two dimensional toy data and introduces both the data matrix $\mathbf{X} = (x_{ij})$ and the corresponding distance matrix $\mathbf{D} = (d_{ih})$ with the element $d_{ih} = (x_{i1} - x_{h1})^2 + (x_{i2} - x_{h2})^2$ (that is the usual squared Euclidean distance in R^2). In some applications the distance matrix $\mathbf{D}$ (or a proximity matrix) may arise directly. Therefore, and because of its more general meaning, a distance matrix will be our preferable starting point for practical cluster analysis. Usually, $\mathbf{D}$ is symmetric with $d_{ih} = d_{hi}$.

For a distance matrix, filled with real numbers and stored in an Excel spreadsheet, a so-called fingerprint (or heat plot) can be created easily. Fig. 2 shows such a heat plot of a 250×250 distance matrix $\mathbf{D}$. One can clearly see the pixels that become a color that is dependent on the distance value of the corresponding cell

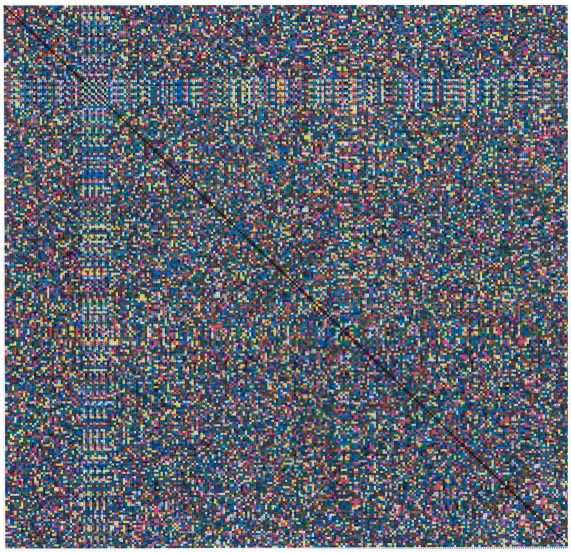


Figure 2: Heat plot of a distance matrix of randomly generated multivariate data.

in the spreadsheet. Here the distances are computed from a multivariate randomly generated data set $\mathbf{X}$, where the number of dimensions is $J = 20$. Obviously the random number generator does not do a good job because the result looks systematic in some regions of $\mathbf{D}$. Hence, such a heat plot seems to be an easy way to verify your random number generator, especially in the case of very high dimensions. Whatever the number of original variables J is (10, 100, 1000, or much more) the size of the corresponding distance matrix remains the same. Of course, it can be become very dangerous to compute one single real number d_{ih} (i.e., the distance value) from two "arbitrary" high-dimensional observations $\mathbf{x}_i$ and $\mathbf{x}_h$ because the true statistical distance is unknown in practice usually. From this point of view, the "appropriate" computation of distances seems to be the most critical point in cluster analysis.

For the next figures, Fig. 3 and Fig. 5, Excel 2007 is required. They show two fragments of the heat plot of a huge 4000 $\times$ 4000 distance matrix $\mathbf{D}$ that contains Euclidean distances: see Fig. 4 for the corresponding schematic view of $\mathbf{D}$ (grey background color) and the legend of relations of color and distance. Here, at the left hand side and at bottom, the clusters are indicated by their number. The lower left rectangle in Fig. 3 presents the pairwise distances between observations of class 3 (last 1300 rows) and class 1 (first 1100 columns). Great distances are located at the lower left corner (in blue, light blue, green to yellow). These are distances between pairs of objects coming from cluster 3 and cluster 1, respectively. The smaller the distances the darker become the red and brown color. For instance, the distances inside class 2 (see at the right top in Fig. 3 and at left and middle in Fig. 5 (1585 rows and first 1585 columns)) seem to be very small compared with the inter-class distances mentioned above. The heat plots suggest that there may be a structure in the data. The entire heat plot cannot be presented because of technical limitations of both the screen and the printer. Each figure is composed of 3 591 610 (= 1585 $\times$ 2266) pixels. The data behind $\mathbf{D}$ is a 4000 $\times$ 2 data matrix $\mathbf{X}$ consisting of the

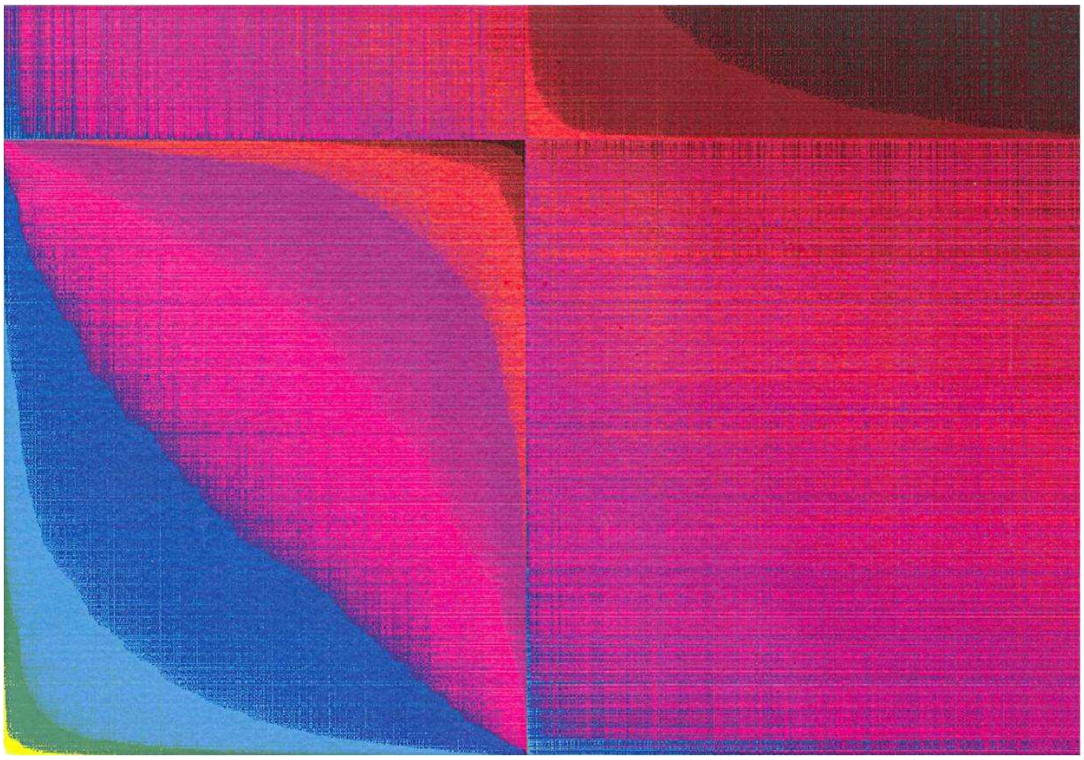


Figure 3: Heat plot of a distance matrix (lower left subarea) of the three-class data
.

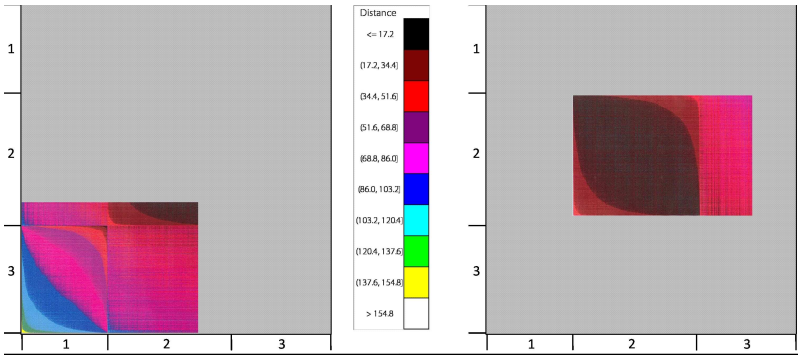


Figure 4: Location of the subareas that are presented in Fig. 3 and Fig. 5.

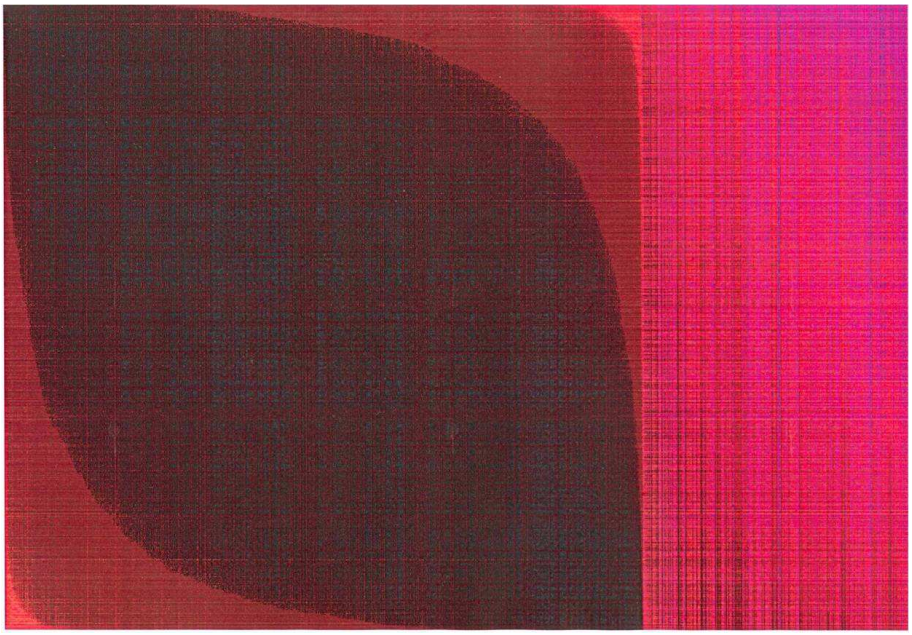


Figure 5: Heat plot of a distance matrix (middle part) of the three-class data.

following three spherical, randomly generated Gaussian classes of sizes 1100, 1600, and 1300, respectively, with the following different mean values $(-3, 3)$, $(0, 0)$, $(3, 3)$ and different standard deviations $(1, 1)$, $(0.7, 0.7)$, and $(1.2, 1.2)$. The values (x_{ij}) were generated simply by using the rand() function of VBA that delivers randomly generated uniformly distributed values r_1 and r_2 coming from [0,1):

$$x_{ij} = \sigma_j \sqrt{-2 \ln r_1} \cos(2\pi r_2) + \mu_j \quad .$$

Here σ_j and μ_j are the standard deviations and mean values of variable j, respectively. This data set is presented in Fig. 6 (scatterplot).

However, more information about the classes is given by histograms as shown in Fig. 7 (scatterplot-histogram). The univariate histogram of the first variable (that is projected onto the wall at the left hand side and estimates the corresponding marginal distribution) reflects the three-class structure much better than the one of the second variable. Also, one can see that cluster $\mathcal{C}_2$ at the right hand side has much less deviations from mean as compared to the other two clusters. Thus, the cluster $\mathcal{C}_2$ is most compact. A smooth and much more impressive version of Fig. 7 can be obtained by non-parametric density estimation. Fig. 8 (density-plot) shows such a bivariate density surface of the two-dimensional three-class data. It is obvious to expect that cluster analysis should be successful in dividing (decomposing) into smaller subsets (clusters) here.

As one might suppose correctly, the ordering of objects plays an important role in heat plots, see for instance [16]. Really, the observations in Fig. 3 to Fig. 5 are ordered first by classes 1, 2, and 3 and then by the scores of the first principal component.

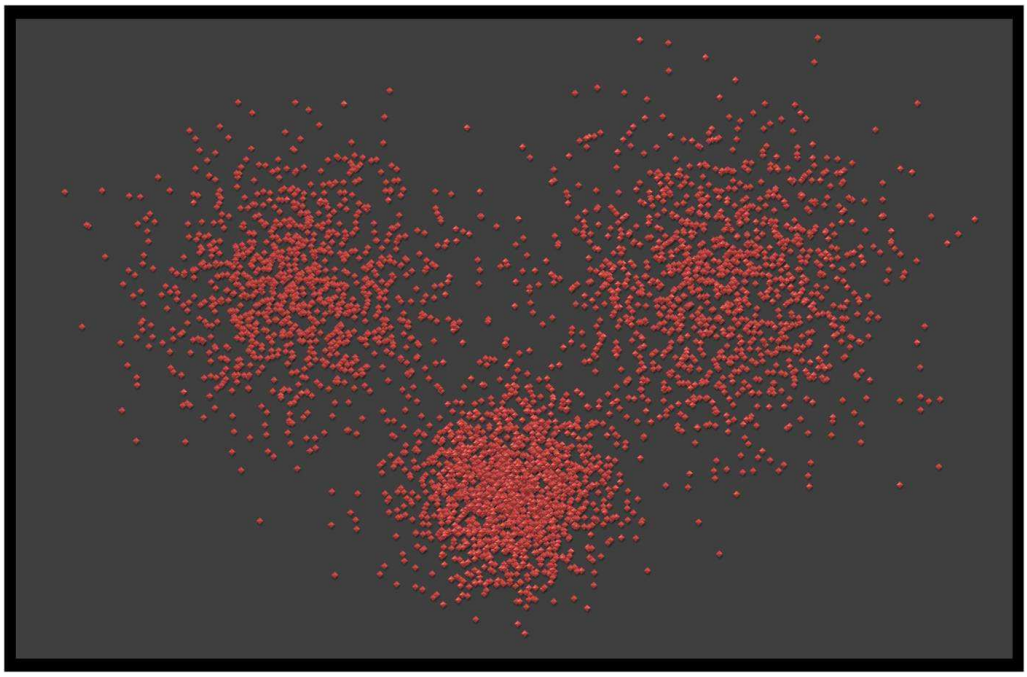


Figure 6: Scatterplot of the three-class data.

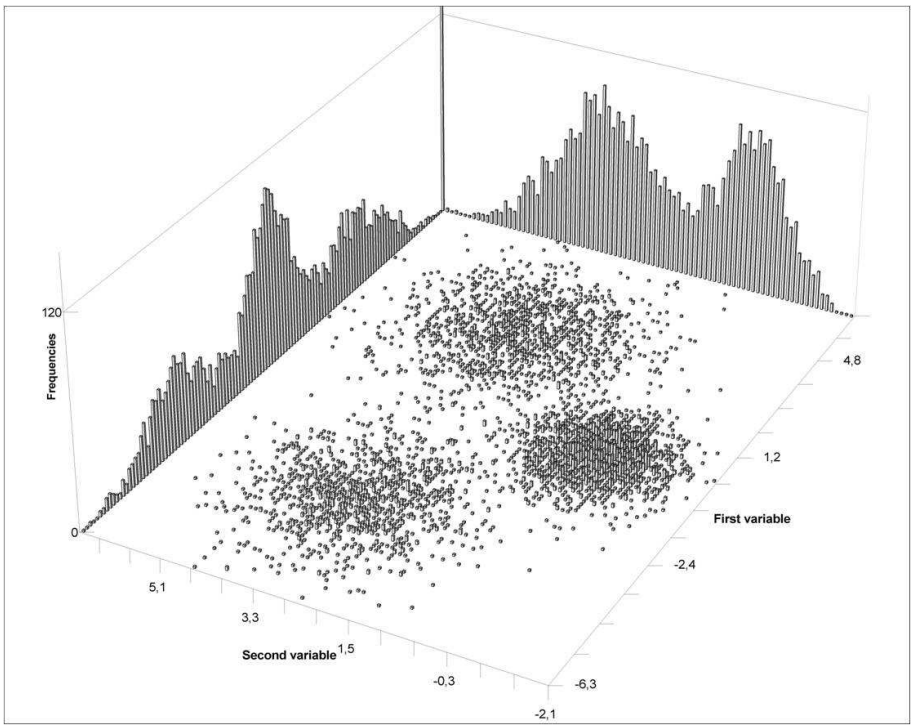


Figure 7: Plot of several histograms of the three-class data.

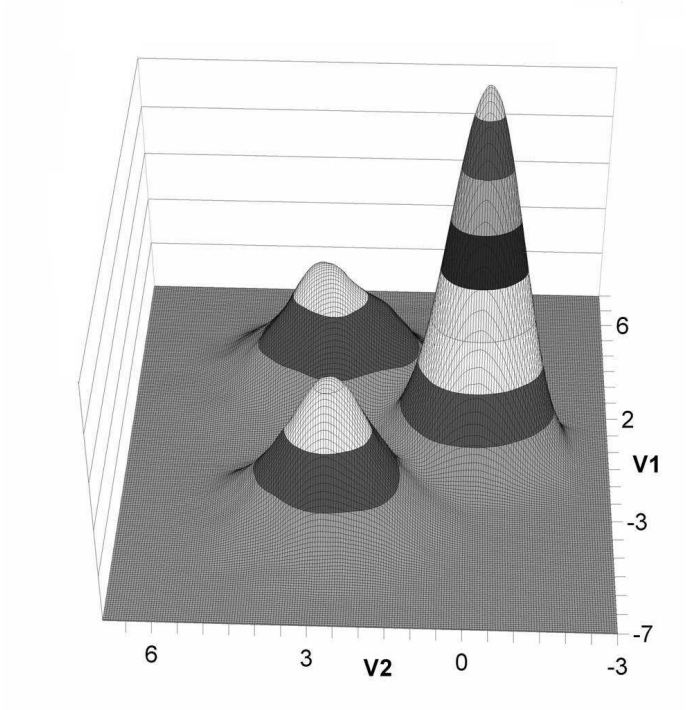


Figure 8: Density plot of the three-class data.

Heat plots offer an easy way to detect structures, also in other matrices such as a data matrix $\mathbf{X}$ (see Fig. 46 in Sect. 3) or a correlation matrix. However, the elements of the matrices should be ordered in an optimum way before.

1.3 From Distances to Partitions and Hierarchies

But now to our basic problem: finding clusters. In the following, the focus is on model-based Gaussian clustering of observations in its simplest setting that results in the sum of squares and logarithmic sum of squares methods. It should be mentioned that both methods can be extended to adaptive techniques ([11], [12]). These simple methods can become a little bit flexible by weighting objects and/or variables, and thus they get more practical relevance. The general model-based Gaussian clustering approach was described first in [1] in all its glory. Because we will make use only of the simple model-based Gaussian clustering based on pairwise distances in this paper, we briefly introduce first some general underlying facts and notations.

Above we introduced the starting points of cluster analysis. Now let us formalize the simplest (elementary) solution to the clustering problem with a fixed number of clusters K: the Boolean assignment matrix $\mathbf{G} \in \{0,1\}^{I \times K}$ (that is, $\mathbf{G} = (g_{ik})$) with the restriction of uniqueness and exhaustive assignment (completeness) $\sum_{k=1}^{K} g_{ik} = 1$ for every object i. Formally, the mapping is,

$$G : \mathcal{C} \times \{1, 2, \ldots, K\} \longrightarrow \{0, 1\}$$

Cluster k / Object i	1	2	**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

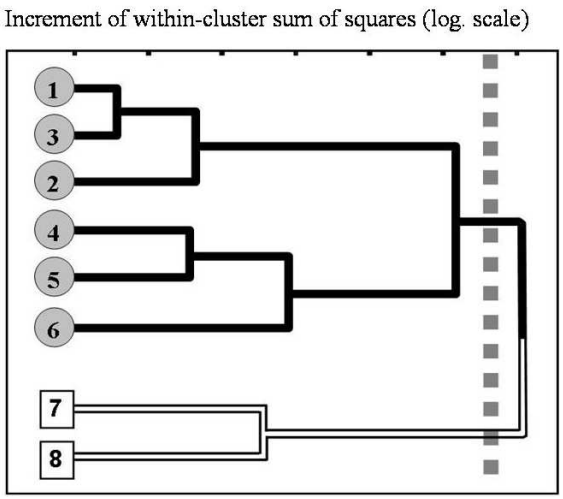


Figure 9: The Boolean assign matrix $\mathbf{G}$, the vector $\mathbf{g}$ of cluster labels, and the corresponding partition in the dendrogram.

with

$$g_{ik} = \begin{cases} 1 & \text{if } i \text{ comes from the cluster (subset) } \mathcal{C}_k \\ 0 & \text{otherwise.} \end{cases}$$

Indeed, the cluster mapping G induces a partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ of $\mathcal{C}$. Hereby

$$\bigcup_{k=1}^{K} = \mathcal{C}$$

and

$$\mathcal{C}_k \cap \mathcal{C}_l = \emptyset$$

for every pair of clusters $\mathcal{C}_k$ and $\mathcal{C}_l$, $k, l = 1, 2, \ldots, K, k \neq l$. This cluster mapping yields exactly K clusters (subsets), where the numbering of the clusters is arbitrary because it usually depends on the applied clustering algorithm itself. Alternatively, let $\mathbf{g} = (g_1, \ldots, g_I)^T$ denote the identifying labels for the clustering and thus for the cluster mapping G, where $g_i = k$ if the ith object $\mathbf{x}_i$ comes from the kth cluster. One can understand $\mathbf{g}$ as a categorical variable or partition variable with K different nominal states $\{1, 2, \ldots, K\}$. Formally, $\mathbf{g} = \mathbf{Ge}$, where the vector $\mathbf{e} = (1, 2, 3, \ldots, K)^T$ has K entities.

Fig. 9 (at the left hand side) illustrates both an example of the Boolean assignment matrix $\mathbf{G} = (g_{ik})$ that maps the I object into $K = 2$ clusters and the corresponding partition $\mathbf{g}$ of the data that is given above in Fig. 1. Here $\mathbf{G}$ gives the optimum K-means cluster analysis solution that is shown in Fig. 11. Typically, a partition is the result of a partitional cluster analysis method like K-means. Such a partition can be obtained also by cutting a dendrogram at a certain level of cluster distance (as it is indicated by the dashed vertical line at the right hand side of Fig. 9). By the way, here the hierarchical cluster analysis method finds the same optimum partition into $K = 2$ clusters as the partitional $D_{ih}Ex$ method does (see below for details on

the methods). This dendrogram is the result of hierarchical clustering by Ward's incremental sum of squares method based on the distance matrix **D** in Fig. 1. By cutting a dendrogram at several different levels of cluster distances one gets a set of partitions.

Banfield and Raftery [1] developed a model-based framework for clustering by parameterizing the covariance matrix in terms of its eigenvalue decomposition. In the following the focus is on a special assumption about the covariance structure. When the covariance matrix is constrained to be diagonal and uniform across all K assumed clusters, the sum of within-clusters sum of squares criterion (shortly: sum of squares = SS)

$$W_K(\mathbf{G}) = \sum_{k=1}^{K} \mathrm{tr}(\mathbf{W}_k) \tag{1}$$

has to be minimized with respect to $\mathbf{G}$ for a fixed K. Herein

$$\mathbf{W}_k = \sum_{i=1}^{I} g_{ik}(\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \tag{2}$$

is the sample cross-product matrix for the kth cluster $\mathcal{C}_k$, and

$$\overline{\mathbf{x}}_k = \frac{1}{g_{.k}} \sum_{i=1}^{I} g_{ik}\mathbf{x}_i \tag{3}$$

is the usual maximum likelihood estimate of expectation values in cluster $\mathcal{C}_k$. Further, $g_{.k}$ is the cardinality of cluster $\mathcal{C}_k$, that is, $g_{.k} = \sum_i g_{ik}$. The SS is fundamental for inferential statistics and descriptive statistics, not only in cases $K > 1$. When scaling the SS (or more precisely the sum of the squared deviations) and the cross product matrix for the number of degrees of freedom in the case $K = 1$, it becomes the variance and the covariance matrix, respectively. In (1), no pairwise distances occur directly in the case of Gaussian distribution, but indirectly they are introduced via the corresponding density function. It is well known that criterion (1) can be written in the following equivalent form without the explicit specification of cluster centers (centroids) $\overline{\mathbf{x}}_k$

$$W_K(\mathbf{G}) = \sum_{k=1}^{K} \frac{1}{2g_{.k}} \sum_{i=1}^{I} \sum_{h=1}^{I} g_{ik}g_{hk}d_{ih} \ , \tag{4}$$

and

$$d_{ih} = d(\mathbf{x}_i, \mathbf{x}_h) = (\mathbf{x}_i - \mathbf{x}_h)^T(\mathbf{x}_i - \mathbf{x}_h) = \|\mathbf{x}_i - \mathbf{x}_h\|^2 \tag{5}$$

is the squared Euclidean distance between two objects i and h. It is also well known that this criterion is dependent on the scales of the variables. Different scales can be formalized by introducing weights of variables. Behind this special use, the variables can be weighted generally by giving important variables more weight (i.e., to gain

in importance). Taking into account weights of the variables the squared weighted Euclidean distance

$$d_{ih} = d_Q(\mathbf{x}_i, \mathbf{x}_h) = (\mathbf{x}_i - \mathbf{x}_h)^T \mathbf{Q} (\mathbf{x}_i - \mathbf{x}_h) \ , \tag{6}$$

generalizes formulae (4), where the $J \times J$ matrix $\mathbf{Q}$ is restricted to be diagonal. With $\mathbf{Q} = \mathrm{diag}(q_1, q_2, \ldots, q_J)$, where q_j $(=q_{jj})$ denotes the weight of the jth variable, we can write simply

$$d_{ih} = \sum_{j=1}^{J} q_j (x_{ij} - x_{hj})^2 \ .$$

In doing so, at least scaling problems can be handled fashionably without any data preprocessing step such as the standardization of variables. Moreover adaptive weights of variables can be used that are estimated during the iteration process of clustering. (For details, also in the frame of principal components analysis (PCA) and in terms of the sample cross-product matrices (2), see [12].) Another approach of assigning weights to variables is clustering of objects with regard to subsets of variables (see Friedman and Meulman [5]). Of course, the statistical distance (6) with a positive definite matrix $\mathbf{Q}$ can be generalized further to cluster specific statistical distances, where instead of $\mathbf{Q}$ the K inverse within-cluster covariance matrices $\mathbf{Q_k}$ are used [22].

Now we are able to forget (2) and thus the corresponding estimates (3). Keep in mind, pairwise distances $\mathbf{D}$ are more general as starting point for (exploratory) cluster analysis and data analysis than a data matrix $\mathbf{X}$. However, the criterion (4) presents practical problems of storage and computation time for increasing I because of their quadratic increase, as Späth [22] pointed out. Meanwhile, a new generation of computers can deal easily with both problems also for $I > 10\,000$. And, the Excel 2007 "Big Grid" spreadsheet is coming in the nick of time.

In order to cluster a practically unlimited number of objects based on criterion (4), let us generalize further by introducing positive weights of objects $u_i, i = 1, 2, \ldots, I$ that will be called here also masses. Instead of dealing with millions of objects directly in (4), their appropriate representatives are clustered subsequent to a preprocessing step of data aggregation. Usually, an aggregation is like smoothing and it has a stabilizing effect. Especially the influence of outliers can be handled in this way to some degree. Obviously, the estimates (3) are affected by masses, but the distances (6) are independent of masses. That means, from the computational point of view, that distances are most suitable for simulation studies (see next Section) because they need to be figured out only ones. The criterion (4) becomes the generalized form

$$W_K^*(\mathbf{G}) = \sum_{k=1}^{K} \frac{1}{2U_k} \sum_{i=1}^{I} u_i \sum_{h=1}^{I} g_{ik} g_{hk} u_h d_{ih} \tag{7}$$

that has to be minimized by incorporating positive masses (weights of objects) and weighted distances (6), where $U_k = \sum_i g_{ik} u_i$ and u_i denote the mass of cluster $\mathcal{C}_k$

and the mass of object i, respectively. In the case of weighted observations, the sample cross-product matrix for the kth cluster $\mathcal{C}_k$ (2) becomes

$$\mathbf{W}_k^* = \sum_{i=1}^{I} g_{ik} u_i (\mathbf{x}_i - \overline{\mathbf{x}}_k^*)(\mathbf{x}_i - \overline{\mathbf{x}}_k^*)^T$$

with

$$\overline{\mathbf{x}}_k^* = \frac{1}{U_k} \sum_{i=1}^{I} g_{ik} u_i \mathbf{x}_i \ .$$

As already mentioned above, the principle of weighting the observations is a key idea for handling cores (representatives) and outliers. In the case of outliers one has to downweight them in some way in order to reduce their influence. In the case of representatives of cores, one has to weight them, for example, proportionally to the cardinality of the cores (for details and applications see [19]). Moreover, in [19], concerning the *K-means* algorithm, one will find conditions of exchange of an observation i from cluster k into cluster g that has to be fulfilled for minimizing (7).

Fig. 10 give you an impression about the flexibility of the simplest model-based criterion (7) when using both special weights of rows u_i and special weights of columns q_j in (6). In doing so, the decomposition of the chi-square statistic of a contingency table is obtained that is of special interest (see [6] for details). In Fig. 10, the data at hand counts the world's largest merchant fleets by country of owner, i.e. all self-propelled oceangoing vessels 1,000 gross tons and greater (as of July 1, 2003, published by CIA World Factbook [3]). The data matrix consists of 20 observations (countries) with the three variables Full Container (abbr.: Cont), Dry Bulk (Bulk), and Tanker (Tank).

By the way, going from pairwise squared Euclidean distances d_{ih} to within-cluster sum of squares w of the two objects i and h means generally

$$w\{i,h\} = \frac{u_i u_h}{u_i + u_h} d_{ih} \ ,$$

and in particular

$$w\{i,h\} = \frac{1}{2} d_{ih}$$

in the case of unit masses $u_i = u_h = 1$. This way is correct in the case of the assumption of the simplest Gaussian model. The advantage of distances are that they are fixed forever independent on weighting the corresponding objects. This is in contrast to the sample cross product matrices (2) and the cluster centers (3) that are affected by changing the weights of objects. Therefore, "soft bootstrapping" by random weighting the objects or subsampling can be performed with an unchanged distance matrix $\mathbf{D}$. For example, one can think about weighting the objects like doubling the sample ($u_i = 2, i = 1, 2, \ldots, I$). In this special case, where the distances become the sum of squares of the set consisting of a pair of objects, the result of clustering should be unchanged. SPSS gives you the opportunity to double a sample easily by a click. However afterwards, SPSS cluster analysis comes up with different

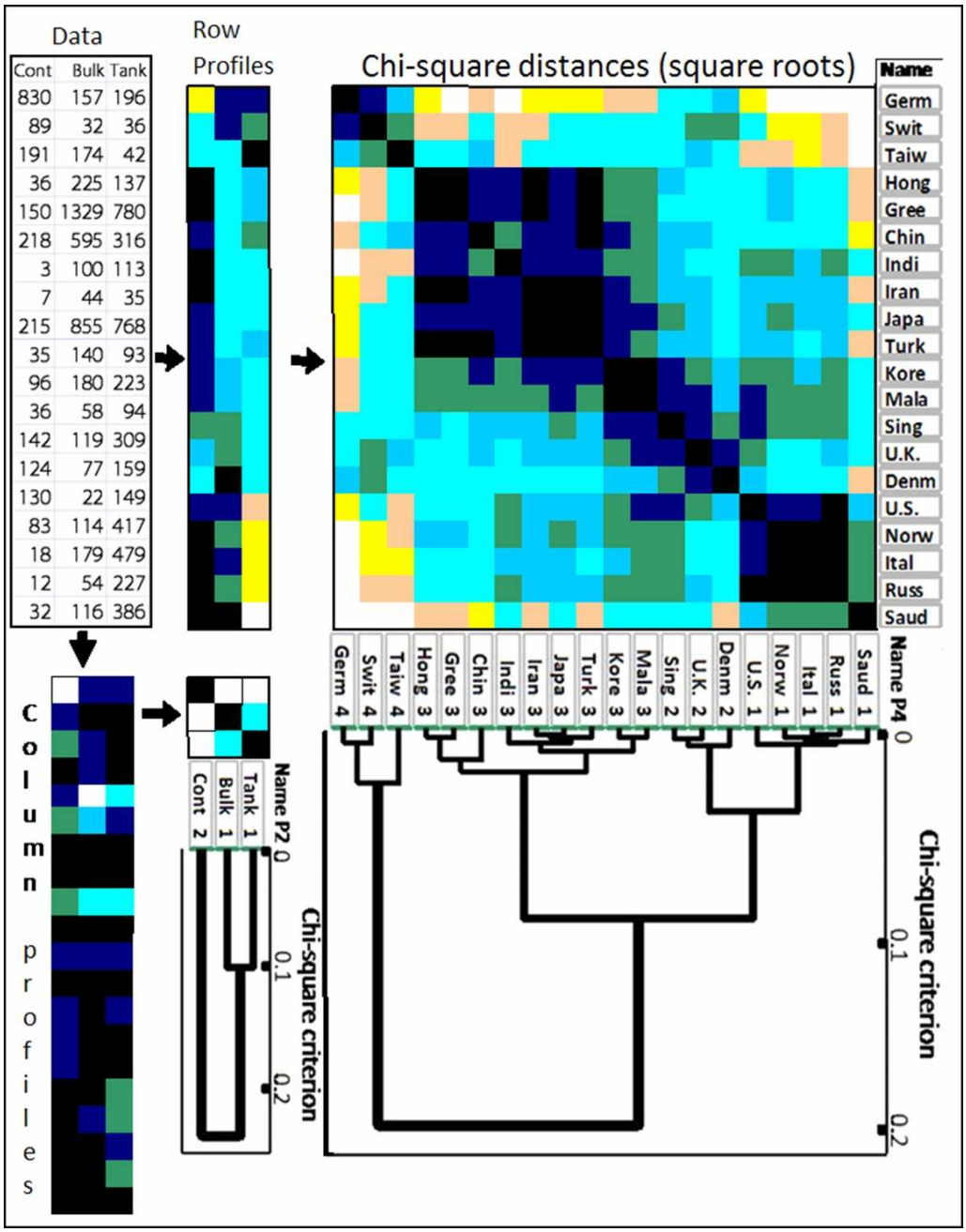


Figure 10: From distances to hierarchies: cluster analysis of a contingency table.

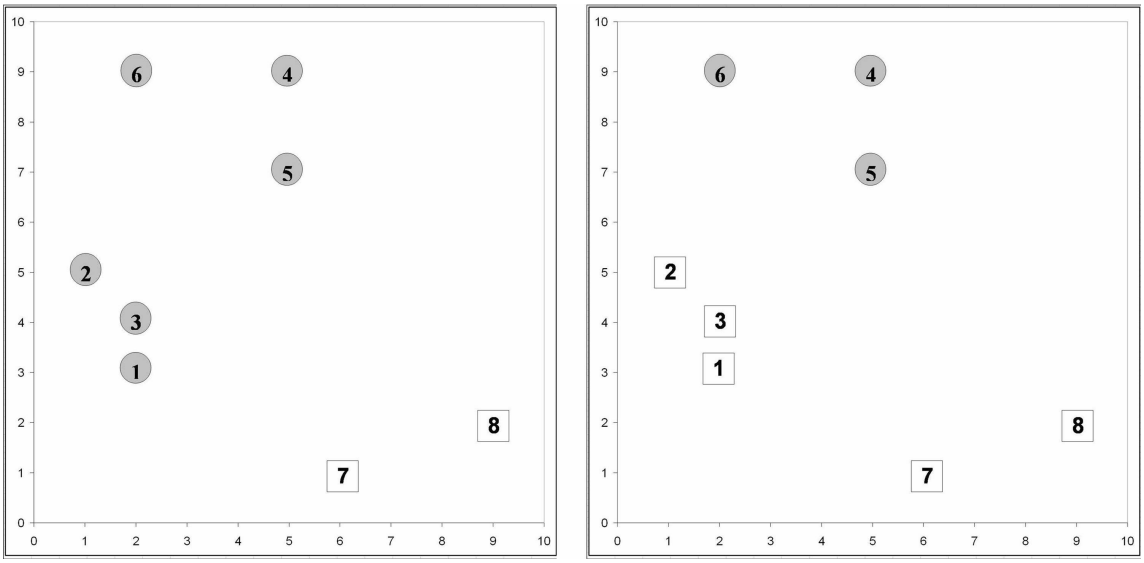


Figure 11: Clustering into two clusters by $D_{ih}Ex$ method (optimum solution) and *Quickcluster* of SPSS (at the right hand side).

results! Correct statistical analysis of weighted objects seems to be a serious problem for some software ...

There are at least two well-known clustering techniques for minimizing the sum of squares criterion based on pairwise distances: the partitional $D_{ih}Ex$ method minimizes the criterion (4) for a single partition $\mathbf{G}$ by exchanging objects between clusters ([22], [2]), and the hierarchical *Ward*'s clustering minimizing (4) in a stepwise manner by agglomerative grouping [10]. The well-known K-means method becomes a special case of the $D_{ih}Ex$ method in the framework of pairwise clustering based on squared Euclidean distances without using centroids anymore. Also one usual definition out of many possible definitions of the most typical object (MTO) of a cluster can be: a MTO is that object that is the most similar one to the centroid, becomes more general in pairwise clustering (see also [10]). Here the typical object of a cluster is the one that minimizes the sum of the (pairwise) distances to the other members of the cluster. Of course, in the case of Gaussian normals, this general most typical object is located usually nearby the expectation value, i.e. the centroid, of the cluster.

In Fig. 11 two different results of clustering are given that divide the toy data of Fig. 1 into two clusters. The plot at the left hand side shows the optimum result with regard to minimum sum of within-cluster sum of squares (= 52.667, see below (1)) that should be easy to find by appropriate methods like the well-known K-means clustering which looks for optimum K centroids (means). However, this result here is obtained by the $D_{ih}Ex$ method (speak Dihex, it is based on the exchange algorithm, for details see [2]) that is a partitional clustering method (based on the pairwise distances $\mathbf{D} = (d_{ih})$. In our statistical software `ClusCorr98`®, a generalized method [2] is used that goes back to Späth [22] who called it TIHEXM. Furthermore,

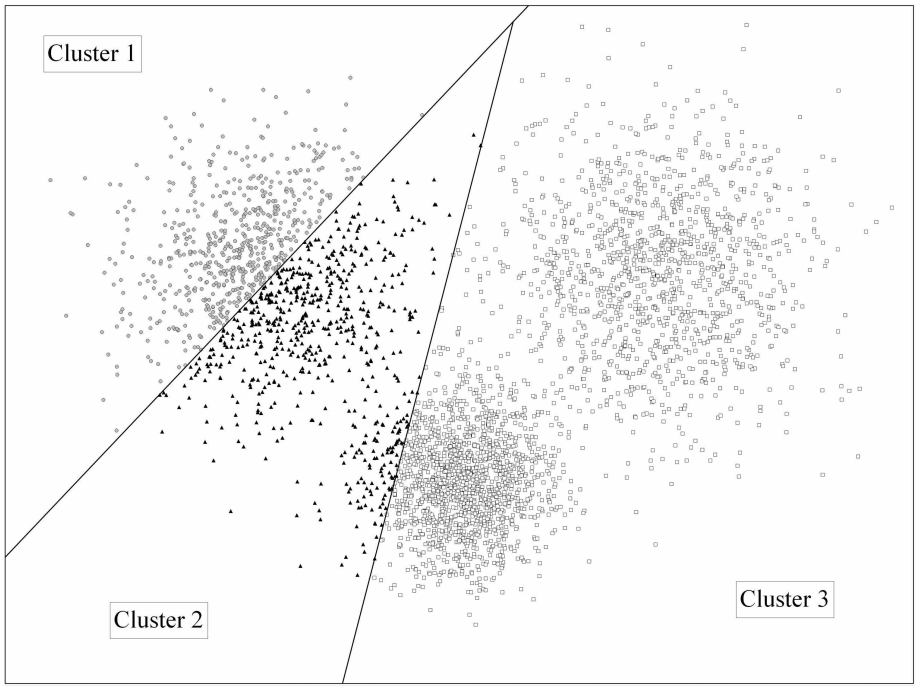


Figure 12: Result of *Quickcluster* of SPSS (same data as in Fig. 7).

our algorithms work with random access to the observations. At the right hand of Fig. 11 a bad sub-optimum solution (= 64.667, see the plot) is found by the K-means clustering procedure *Quickcluster* of SPSS [1] using the running means option. This option means nothing more than exchange method. Moreover, the *Quickcluster* - iterations failed to converge! That is to say, the iterations stopped only when the maximum number of iterations was performed. If one repeats the run, the result remains the very same.

Fig 12 shows a quite bad result of clustering the three randomly generated normal subpopulations. Here the procedure *Quickcluster* of SPSS is applied again with the option running means: the clusters are updated after each object is assigned to a new cluster. The main reason for this failure is that the underlying algorithm is depends on the sequence of objects. What, if one can not check the validity of cluster analysis results visually by eye? (Here in R^2 a graphical inspection is both easy to do and very powerful.) There is a need for a validation approach that works in almost all situations (see the next subsection).

Therefore, let us generalize the validation by comparing two partitions $\mathbf{f}$ and $\mathbf{g}$ into M and K clusters (categories), respectively. Well-known measures of correspondence between such categorical variables are based on the contingency table $\mathbf{N} = (n_{mk})$ that is obtained by crossing the vectors $\mathbf{f}$ and $\mathbf{g}$. A contingency table, also referred to as pivot table, can be established easily in Excel. Alternatively, such a contingency

[1] Here release SPSS 17 for Windows is used for the tests (see also Fig. 12). Some of the problems which have occurred will be mentioned below. Almost all have been well known for a long time.

	Cluster k / Class m	Partition **g** 1	2	3	Sum
Partition **f** (true)	1	**596**	**501**	**3**	1100
	2	**0**	**130**	**1470**	1600
	3	**0**	**8**	**1292**	1300
	Sum	596	639	2765	4000

g (*Ward*'s method) 1	2	3
1089	9	2
7	**1587**	6
5	21	**1274**
1101	1617	1282

g (**D**$_{ih}$Ex method) 1	2	3
5	0	**1095**
1595	0	5
31	**1255**	14
1631	1255	1114

Figure 13: A contingency table **N** that is obtained by crossing two partitions **f** and **g**, and two other tables of results.

table **N** can be formulated by simple matrix notation

$$\mathbf{N} = \mathbf{F}^T\mathbf{G}$$

based on the corresponding two Boolean assignment matrices **F** and **G**. Fig. 13 shows at the left hand side the contingency table that comes from crossing the true partition of the simulated data set of 4000 objects with the cluster analysis result of *Quickcluster* of SPSS (see Fig. 12). At the right hand side two other contingency tables are presented that come from *Ward*'s clustering and $D_{ih}Ex$ clustering (outside right), respectively. The last two results are based on pairwise clustering using the squared Euclidean distances (5). A comparison of the performance of the three methods on clustering the three-class data says: *Ward*'s method performs slightly better (50 errors only) than the $D_{ih}Ex$ clustering with 55 misclassifications. It has to be mentioned that the criterion (1) that is minimized here by the three methods is not the most appropriate one for this data. The right one is the logarithmic sum of squares (see below).

Almost all (exploratory) clustering techniques detect clusters, even on data without any cluster structure. Often, clustering techniques are applied for finding (practical useful) segmentations of data such as vector quantization by using K-means clustering. Fig. 14 shows the cluster membership of 4000 points into 15 clusters that is obtained by the hierarchical Ward method. This is an example of a data set without any cluster structure. The randomly generated data in R^2 come from a bivariate Gaussian distribution $N_2(\mu, \mathbf{\Sigma})$ with parameters μ (mean vector) and $\mathbf{\Sigma}$ (covariance matrix). Here one standard normal population is generated with $\mu = 0$ and $\mathbf{\Sigma} = \mathbf{I}_J$, where $\mathbf{I}_J$ is the $J \times J$ identity matrix. The centers of the clusters "C1", ..., "C15" are marked by an asterisk ("Cs" in the legend). Fig. 15 shows the clusters that are obtained by the $D_{ih}Ex$ method, that is, by the K-means method.

Even though both clustering techniques, the hierarchical Ward and the partitional K-means method, have the same underlying statistical model and minimize the same criterion (4), but in another way, the results are usually different. The last one

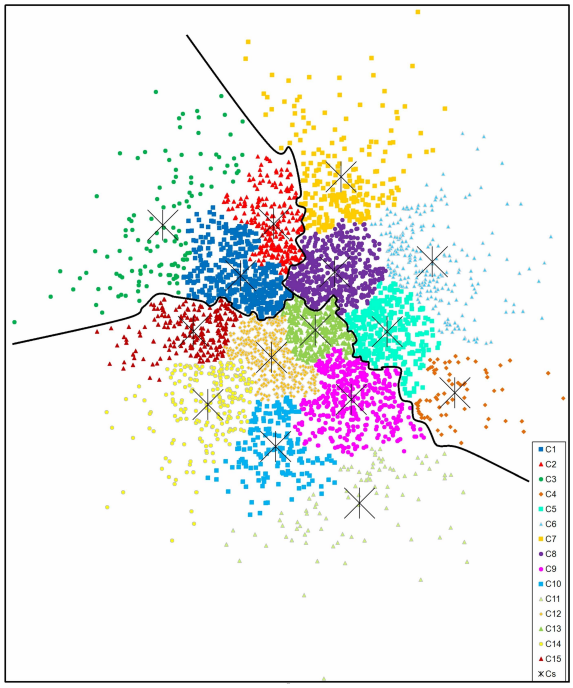


Figure 14: Ward's clustering of no-structure data into 15 and 3 clusters, respectively.

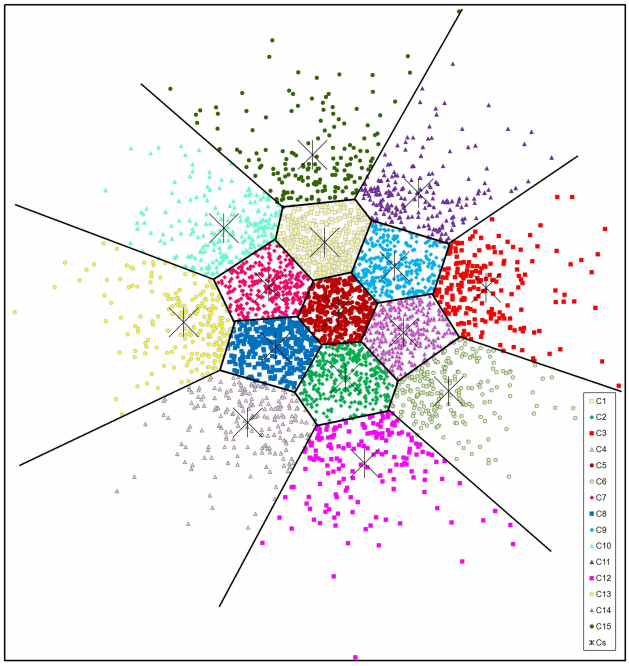


Figure 15: $D_{ih}Ex$ clustering of no-structure data into 15 clusters.

is leading to the well-known Voronoi tessellation, where the objects have minimum distance to their centroid and thus, the borderlines between clusters are hyperplanes. By contrast, the Ward method does not necessarily have to creates hyperplanes as borderlines between clusters, as Fig. 14 shows for the three-cluster solution.

When the covariance matrix of each cluster C_k $(k = 1, 2, \ldots, K)$ is constrained to be diagonal, but otherwise allowed to vary between groups, the logarithmic sum-of-squares criterion

$$V_K(\mathbf{G}) = \sum_{k=1}^{K} g_{.k} \log \operatorname{tr} \frac{\mathbf{W}_k}{g_{.k}} \ . \tag{8}$$

Once again the following equivalent formulation can be derived

$$V_K(\mathbf{G}) = \sum_{k=1}^{K} g_{.k} \log(\sum_{i=1}^{I} \sum_{h=1}^{I} \frac{g_{ik} g_{hk}}{2 g_{.k}^2} d_{il}). \tag{9}$$

Considering formulae (9) (and (8) in the case of formulation with sample cross product matrices, respectively) and weights of observations u_i, the logarithmic sum-of-squares criterion can be generalized to

$$V_K^*(\mathbf{G}) = \sum_{k=1}^{K} U_k \log(\sum_{i=1}^{I} \sum_{h=1}^{I} \frac{u_i u_h}{2 U_k^2} g_{ik} g_{hk} d_{il}). \tag{10}$$

According to this logarithmic sum-of-squares criterion, the partitional *K-means*-like clustering algorithm is also referred to as *Log-K-means* and the hierarchical *Ward*-like agglomerative method as *LogWard*, respectively [19]. Concerning the hierarchical algorithms there are special treatments of observations with low weights in use (see, for example, [19]). Such special tricks are essential because the original Ward's hierarchical agglomerative clustering is based on minimum incremental of sum of squares, and therefore all observations with zero (or quasi-zero) weight would be merged together into one cluster, whatever the level of distance values may be. By the way, *K-means* and *Log-K-means* based on pair-wise distances (6) are also more general because they never require an $(I \times J)$-data matrix $\mathbf{X}$. The pixel graphic of Fig. 16 shows the result of hierarchical clustering based on the criterion (10) when using both special weights of rows u_i and special weights of columns q_j in (6).

1.4 Built-in Validation of Cluster Analysis Results

As already shown above, more often than not clustering techniques always detect clusters. Moreover, hierarchical clustering presents all the clusters that are established during the agglomeration or the division process. In Fig. 17 the amalgamation process of 13 points is illustrated. The points are located at the real line and their values can be taken from the picture. Each point is a terminal node in the tree. It is marked by a dark circle with its value given below. Here the *weighted pair-group method using centroids* (called also *Median method*) is applied [11]. Each non-trivial

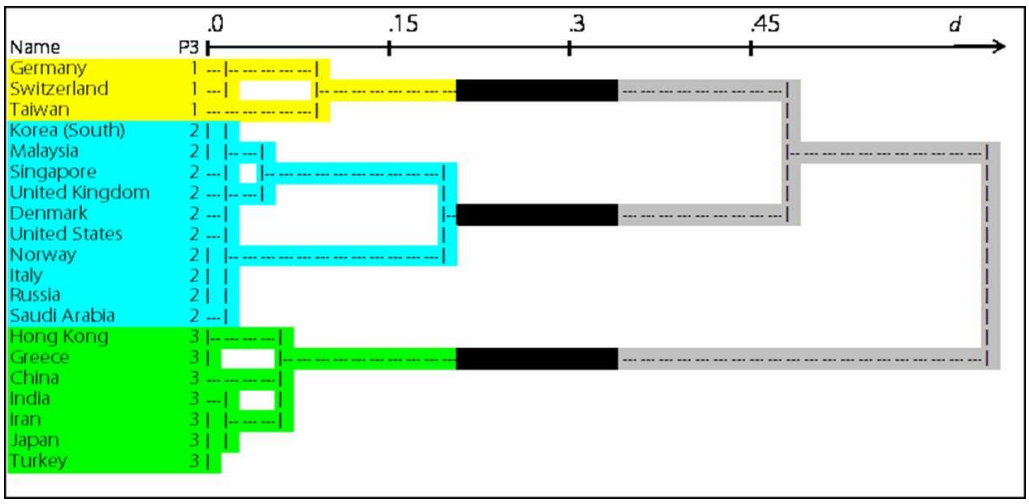


Figure 16: Colored dendrogram of *LogWard*-clustering of the table of Fig. 10.

cluster (non-terminal node) is marked by a light circle. Most of all clusters are characterized by their centroids additionally. Obviously, in every agglomeration step an increasing amount of information is lost. A better visual impression can be obtained by a plot-dendrogram, that is a dendrogram projected onto a plane. Fig. 18 shows a hierarchy of 227 countries based on demographic variables. The main question is: how many clusters are there? Or, in other words, when should the agglomeration process stops? Or, in terms of the density estimation in Fig. 19, what is the right cut-off density level for fixing clusters? Another outfit of this figure is presented below in Sect. 3 in Fig. 44.

There are so many different clustering algorithms and new ones occur daily in the literature. More often than not they do their job and usually, they present a solution in almost all cases. As from now let us suppose that they do a good (accurate) job (because otherwise the validation can give the right answer to the wrong question). Then the main question arises: is there really a cluster structure in the data? Therefore, in this section a validation of clustering results based on resampling techniques is highly recommended that can be considered as a three level assessment of stability. The first, most general level is decision making about the appropriate number of clusters. Second, the stability of each individual cluster is assessed based on measures of similarity between sets. From many applications it is known that it makes sense to investigate the specific stability of clusters. In the third and most detailed level of validation, the reliability of the cluster membership of each individual object can be assessed.

In any case, it is highly recommended that the stability of the obtained clusters has to be assessed by using validation techniques ([9], [11], [7]). Concretely, here a built-in validation of clustering results based on resampling or subsampling techniques (bootstrapping) is highly recommended that can be considered as a three level assessment of stability. An alternative way beside bootstrapping is disturbing the data by randomly generated errors (noise). The first and most general level is

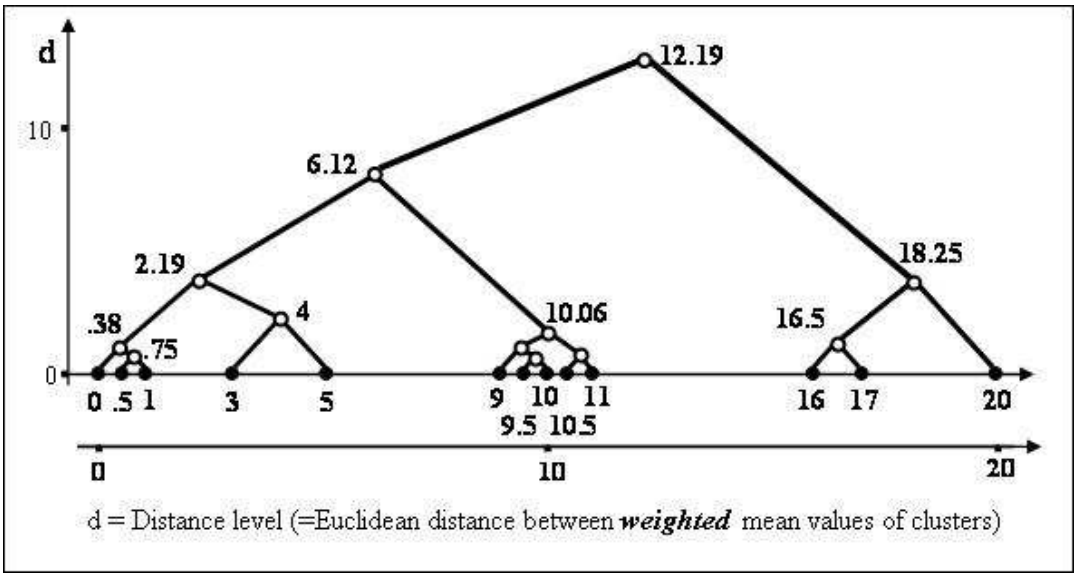


Figure 17: A non-equidistant dendrogram of 13 points located on the real line.

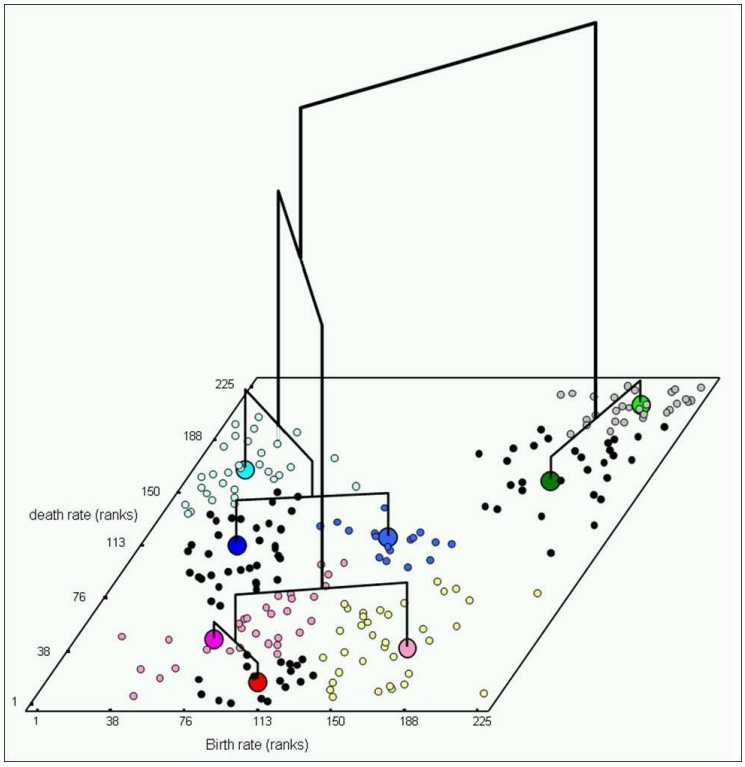


Figure 18: A so-called plot-dendrogram based on demographic data of 227 countries.

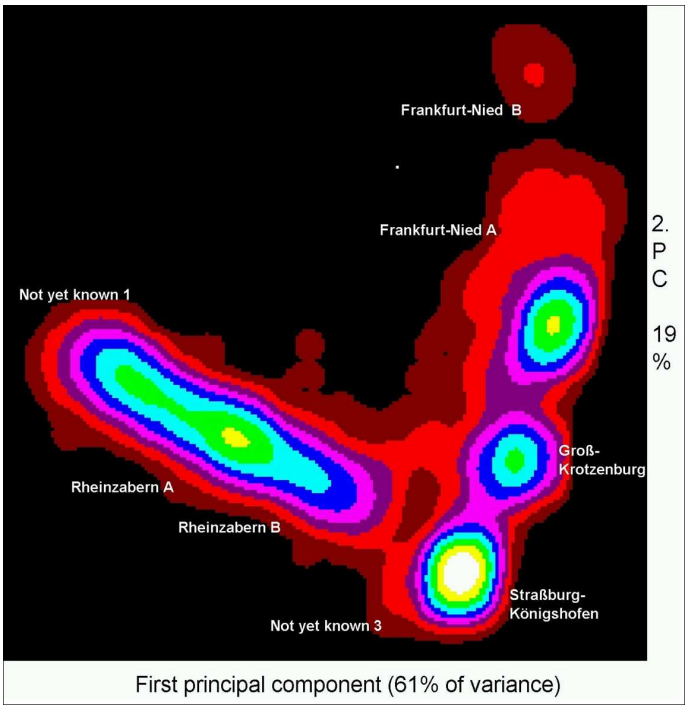


Figure 19: Cuts at several levels of the bivariate nonparametric density estimate.

decision making about the appropriate number of clusters. This decision is based on such well-known measures of correspondence between partitions like the *Rand*'s index ([20]), the *adjusted Rand*'s index of Hubert and Arabie [8], and the index of Fowlkes and Mallows [4]. Second, the stability of each individual cluster is assessed based on measures of similarity between sets, e.g., the asymmetric measure of cluster agreement or the symmetric *Jaccard* measure. It should be mentioned that it makes sense to investigate the (often quite different) specific stability of clusters of the same clustering on the same data. Often one can observe that the clusters have a quite different stability. Some of them are very stable. Thus, they can be reproduced and confirmed to a high degree, for instance, by bootstrap simulations. They are both homogeneous inside and well separated from each other. Moreover, sometimes they are located far away from the main body of the data like outliers. On the other side, hidden and tight neighboring clusters are more difficult to detect and they cannot be reproduced to a high degree. In the third and most detailed level of validation, the reliability of the cluster membership of each individual object will be assessed.

Here we don't consider special properties like compactness and isolation as it is done in [9]. A general purpose technology for validation is recommended that works well especially in highdimensional settings. In low dimensional cases and in cases where projection methods result in good approximations into R^2 or R^3, graphical methods are often the better and more efficient choice for validation. What are stable clusters from a general statistical point of view? These clusters can be confirmed and reproduced to a high degree. To define stability with respect to the individual

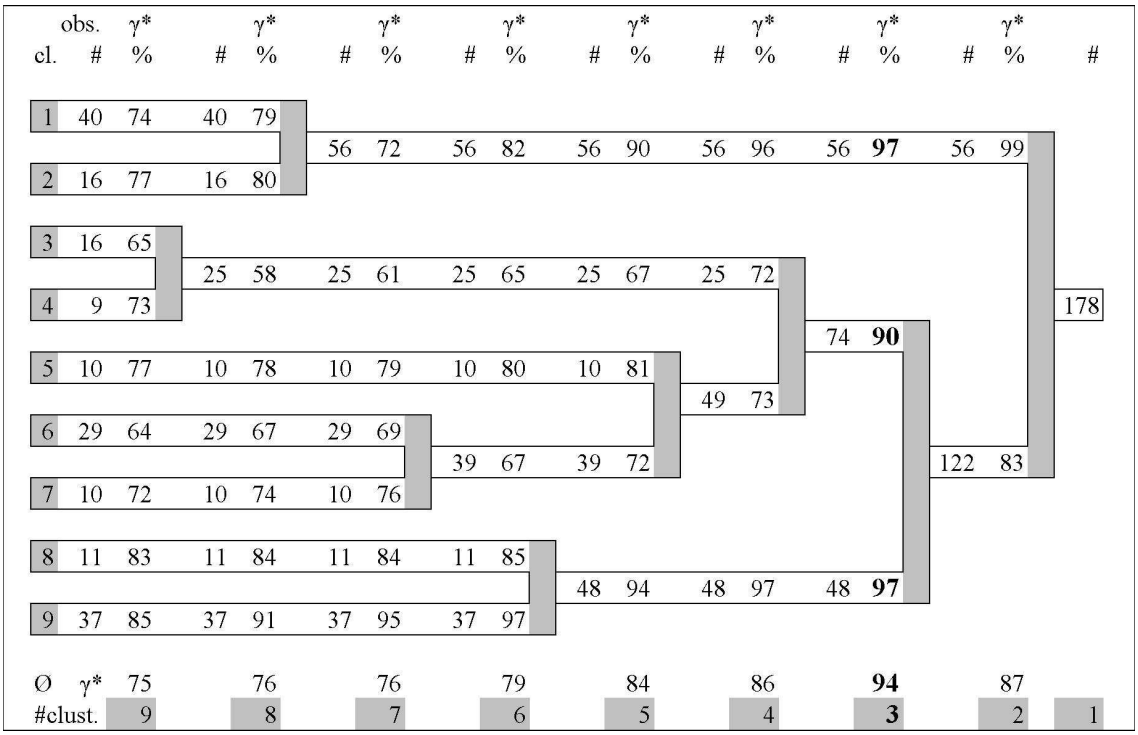


Figure 20: Dendrogram (extract) of the wine dataset with assessment of stability of nodes based on measure (13).

clusters, measures of correspondence between a cluster $\mathcal{E}$ and a cluster $\mathcal{H}$ such as

$$\tau(\mathcal{E},\mathcal{H}) = \frac{|\mathcal{E} \cap \mathcal{H}|}{|\mathcal{E} \cup \mathcal{H}|} \tag{11}$$

or

$$\gamma(\mathcal{E},\mathcal{H}) = \frac{|\mathcal{E} \cap \mathcal{H}|}{|\mathcal{E}|} \tag{12}$$

have to be defined. ($\mathcal{E}$ and $\mathcal{H}$ are nonempty subsets of some finite set.) Hennig [7] suggests the *Jaccard* coefficient τ (11). This measure is symmetric and it attains its minimum 0 only for disjoint sets and its maximum 1 only for equal ones. The asymmetric measure γ assesses the rate of recovery of subset $\mathcal{E}$ by the subset $\mathcal{H}$. It attains its minimum 0 only for disjoint sets and its maximum 1 only if $\mathcal{E} \subseteq \mathcal{H}$ holds. We obviously have $\tau \leq \gamma$.

Now suppose, a clustering of a set of entities $\mathcal{C} = \{1, \ldots, i, \ldots, I\}$ into a collection of K subsets $\{\mathcal{C}_1, \ldots, \mathcal{C}_k, \ldots, \mathcal{C}_K\}$ of $\mathcal{C}$ has to be investigated. Let $\mathcal{C}_k$ be one individual cluster whose stability has to be assessed. To investigate the stability, validation techniques based on random resampling are recommended. Let us consider one simulation step: Clustering of a randomly drawn sample of the set of entities $\mathcal{C}$ into a collection of K clusters $\{\mathcal{H}_1, \ldots, \mathcal{H}_K\}$ in the same way as clustering the whole set $\mathcal{C}$. The definition of stability of cluster $\mathcal{C}_k$ using measure γ (12) is based on the most similar cluster

$$\gamma_k^* = \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \ldots, \mathcal{H}_K\}} \gamma(\mathcal{C}_k, \mathcal{H}_i) \ . \tag{13}$$

By repeating resampling and clustering many times, the stability of the cluster $\mathcal{C}_k$ can be assessed, for instance, by computing the median or the average of the corresponding values of γ_k^*. Let us denote such an estimate $\hat{\gamma}_k^*$.

Illustrating this let us have a look at the hierarchical clustering of a real data set: the wine recognition data[2] (for details see [15]). Altogether there are 178 Italian wines that are described by 13 constituents (variables). Here we worked with ranks instead of the original values that come from scales that are not comparable one with each other. Fig. 20 shows the schematic dendrogram of Ward's clustering for up to 9 clusters. Each node (cluster) of the binary tree is denoted by both the corresponding number of objects (symbol #) and the average rate of recovery (13) in %. The three-cluster solution is emphasized in bold type. While looking for stable clusters and for an outstanding number of clusters you should keep in mind that a hierarchy is a set of nested partitions. Therefore it is recommended to walk step by step through the binary tree (dendrogram) from the right hand side (that corresponds to the root of the tree) to the left. At each step $K-1$ clusters remain unchanged and one cluster is divided only into two parts. Usually, the higher the number of clusters K becomes during the trip through the dendrogram the smaller amount of changes of the averaged measures of stability can be expected that are given at the bottom of the figure. Some of the clusters remain unchanged during many steps such as the cluster of 56 observations at the top of Fig. 20. However, the value of stability of this cluster decreases from 99% for the partition into 2 clusters to 72% only for the partition into 7 clusters because of the altering clusters in its neighborhood.

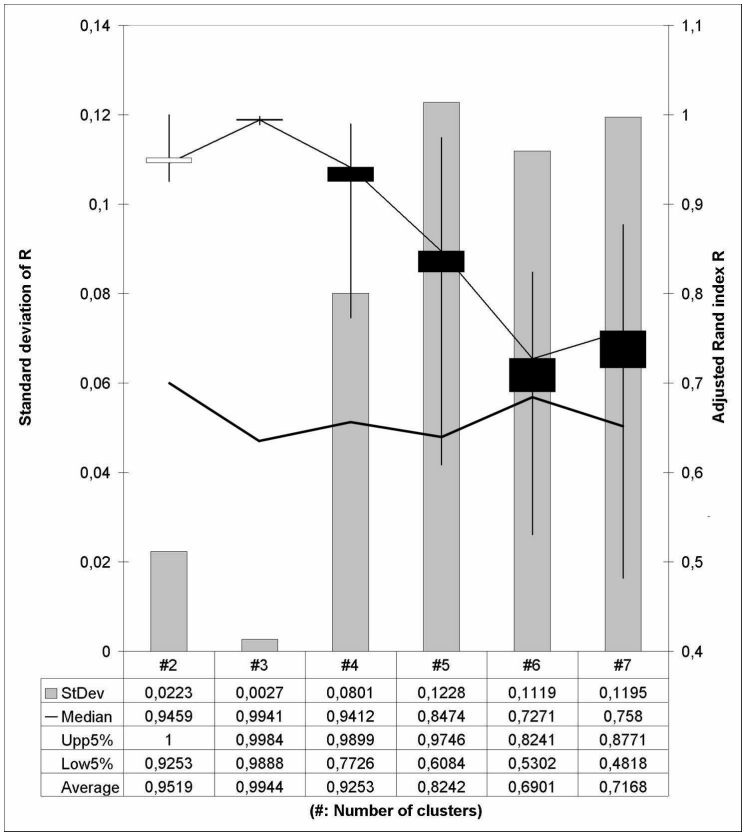
It is difficult to fix an appropriate threshold to consider a cluster as stable. To support the decision about stable regions, the clusters can often be visualized in low dimensional projections by applying methods like discriminant analysis (DA), PCA, and multidimensional scaling (MDS). The simulation itself is computationally expensive.

Now forget that the classes are known beforehand in the case of the three-class data that was presented above (for instance, see Fig. 6 or Fig. 7). Is it possible to confirm by simulations that there are three clusters? The simulation results concerning the determination of the number of clusters are given in Fig. 21. $D_{ih}Ex$ cluster analysis applied (pairwise clustering by the exchange method, see the result at the right hand side of Fig. 13). The simulation results are based purely on clustering of data by resampling techniques. Hubert and Arabie [8] recommended the adjusted *Rand* index R based under the assumption of the generalized hypergeometric model:

$$R = \frac{\sum_{k=1}^{K}\sum_{m=1}^{M}\binom{n_{km}}{2} - [\sum_{k=1}^{K}\binom{n_{k.}}{2}\sum_{m=1}^{M}\binom{n_{.m}}{2}]/\binom{n_{..}}{2}}{\frac{1}{2}[\sum_{k=1}^{K}\binom{n_{k.}}{2} + \sum_{m=1}^{M}\binom{n_{.m}}{2}] - [\sum_{k=1}^{K}\binom{n_{k.}}{2}\sum_{m=1}^{M}\binom{n_{.m}}{2}]/\binom{n_{..}}{2}}. \qquad (14)$$

For the notations concerning $\mathbf{N}$ see above. This measure is appropriate for the decision about the number of clusters K because it takes the value 0 when the Rand index equals its expected value for each $k, k = 2, 3, ..., K$. The median of the *adjusted*

[2] `http://www.ics.uci.edu/~mlearn/MLSummary.html`



	#2	#3	#4	#5	#6	#7
StDev	0,0223	0,0027	0,0801	0,1228	0,1119	0,1195
Median	0,9459	0,9941	0,9412	0,8474	0,7271	0,758
Upp5%	1	0,9984	0,9899	0,9746	0,8241	0,8771
Low5%	0,9253	0,9888	0,7726	0,6084	0,5302	0,4818
Average	0,9519	0,9944	0,9253	0,8242	0,6901	0,7168

Figure 21: Statistics of the *adjusted Rand*'s index versus number of clusters.

Rand's values (the scale is at the right hand side of the plot) achieved its maximum value at three clusters. Moreover, the standard deviation of the 250 *adjusted Rand*'s values for each partition into $K = 2, 3, \ldots$ cluster has its minimum value also at three clusters. This supports the decision for the solution into three clusters. More than three clusters are less likely because the clusters can not be confirmed to a high degree and their stability decreases rapidly. Further statistics like the average over 250 *adjusted Rand*'s values are given at the bottom of the table.

Additionally, for reasons of comparison, the thick line in Fig 21 represents a so-called reference curve for the median of *adjusted Rand*'s values that are obtained from randomly generated datasets without a class structure like the one shown in Fig. 14. The reference curve depends on the dimensionality J. It tends to zero if J increases. In R^{20}, for instance, the reference values are nearly equal to 0.1 for $K = 2, 3, \ldots$. It should be kept in mind that a stable cluster solution in K classes, say $K = 3$ in our case, affects at least the stability of cluster analysis into $K - 1$ and $K + 1$ clusters to some degree. Or to a high degree as documented in Fig 21, where both the two-cluster and four-cluster solution have *adjusted Rand*'s values that are far from the reference curve. Obviously, as also indicated in Fig 21, the influence of the true solution should become less important by going to solutions of $K + 2, K + 3, \ldots$ clusters. In hierarchical cluster analysis, the influence of stable clusters can be much higher usually because they can remain unchanged during many steps of amalgamation (see Fig. 20).

Concerning adaptive weighting of the variables, for example, the aim is to estimate in automatic mode what counts and what doesn't count for finding clusters. Often the performance and stability of these methods can be improved by using them in a local fashion [14]. The improvement of stability can be measured by simulation studies such as described above.

References

[1] BANFIELD, J.D. and RAFTERY, A.E. (1993): Model-Based Gaussian and non-Gaussian Clustering. *Biometrics, 49, 803–821.*

[2] BARTEL, H.-G., MUCHA, H.-J. and DOLATA, J. (2003): Über eine Modifikation eines graphentheoretisch basierten partitionierenden Verfahrens der Clusteranalyse. *Match 48, 209–223.*

[3] CIA World Factbook. World's largest merchant fleets by country of owner. `http://www.geographic.org`, 2003.

[4] FOWLKES E.B. and MALLOWS, C.L. (1983): A Method for Comparing two Hierarchical Clusterings. *JASA 78, 553–569.*

[5] FRIEDMAN, J.H. and MEULMAN, J.J. (2002): Clustering Objects on Subsets of Attributes. `http://www-stat.stanford.edu/~jhf/ftp/cosa.pdf`. Depart-

ment of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford.

[6] GREENACRE, M.J. (1988): Clustering the Rows and Columns of a Contingency Table, *Journal of Classification, 5, 39–52*.

[7] HENNIG, C. (2004): A General Robustness and Stability Theory for Cluster Analysis. *Preprint, 7,*, Universität Hamburg.

[8] HUBERT, L.J. and ARABIE, P. (1985): Comparing Partitions. *Journal of Classification, 2, 193–218*.

[9] JAIN, A.K. and DUBES, R.C. (1988): *Algorithms for Clustering Data*. Prentice Hall, New Jersey.

[10] KAUFMAN, L. and ROUSSEEUW, P.J. (1990): *Finding Groups in Data*. Wiley, New York.

[11] MUCHA, H.-J. (1992): *Clusteranalyse mit Mikrocomputern*. Akademie Verlag, Berlin.

[12] MUCHA, H.-J. (1995). XClust: Clustering in an Interactive Way. In: W. Härdle, S. Klinke, and B.A. Turlach (Eds.): *XploRe: An Interactive Statistical Computing Environment*. Springer, New York, 141–168.

[13] MUCHA, H.-J. (2004): Automatic Validation of Hierarchical Clustering. In: J. Antoch (Ed.): *Proceedings in Computational Statistics, COMPSTAT 2004, 16th Symposium*. Physica-Verlag, Heidelberg, 1535–1542.

[14] MUCHA, H.-J. (2006): Finding Meaningful and Stable Clusters Using Local Cluster Analysis. In: V. Batagelj, H.-H. Bock, A. Ferligoj and A. Ziberna (Eds.): *Data Science and Classification*, Springer, Berlin, 101–108.

[15] MUCHA, H.-J. (2007): On Validation of Hierarchical Clustering. In: R. Decker and H.-J. Lenz (Eds.): *Advances in Data Analysis*. Springer, Berlin, 115–122.

[16] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2005): Techniques of Rearrangements in Binary Trees (Dendrograms) and Applications. *Match 54 (3), 561–582*.

[17] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2008): Effects of Data Transformation on Cluster Analysis of Archaeological Data. In: C. Preisnach, H. Burkhardt, L. Schmidt-Thieme and R. Decker (Eds.): *Data Analysis, Machine Learning and Applications*, Springer, Berlin, 681–688.

[18] MUCHA, H.-J. and HAIMERL, E. (2005): Automatic Validation of Hierarchical Cluster Analysis with Application in Dialectometry. In: C. Weihs and W. Gaul (Eds.): *Classification - The Ubiquitous Challenge*, Springer, Berlin, 513–520.

[19] MUCHA, H.-J., SIMON, U. and BRÜGGEMANN, R. (2002): Model-based Cluster Analysis Applied to Flow Cytometry Data of Phytoplankton. *Weierstrass Institute for Applied Analysis and Stochastic, Technical Report No. 5.* `http://www.wias-berlin.de/`.

[20] RAND, W.M. (1971): Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association, 66, 846–850.*

[21] SIMON, U., MUCHA, H.-J. and BRÜGGEMANN, R. (2004): Model-Based Cluster Analysis Applied to Flow Cytometry Data. In: D. Baier and K.-D. Wernecke (Eds.): *Innovations in Classification, Data Science, and Information Systems.* Springer, Berlin, 69–76.

[22] SPÄTH, H. (1985): *Cluster Dissection and Analysis.* Ellis Horwood, Chichester.

[23] WARD, J.H. (1963): Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58, 236–244.

2 Resolving ambiguity in segmentation problems by the method of variants

Gunter Ritter
Fakultät für Informatik und Mathematik
Universität Passau
ritter@fim.uni-passau.de

Abstract

Often feature extraction from objects in pattern recognition, such as images or acoustic signals, is ambiguous. Ambiguity occurs, in particular, in segmentation problems. In order to resolve ambiguities, the statistical method of variants has been developed in the last decade. The method is applied here to the segmentation of random, cyclic processes.

Keywords: segmentation problems, ambiguity, variant analysis, parameter estimation, sun spots

2.1 Segmentation and ambiguity

Segmentation is the decomposition of complex objects such as images or acoustic signals in simpler components. The components are subsequently analyzed and classified. Well-known examples are speech processing, see Rabiner (1989), optical character recognition, see Casey and Lecolinet (1996), gene finding in functional genetics, see Majoros (2007), www.geneprediction.org, and Fig. 22, and the chromosome classification problem, see Ritter and Gao (2008) and Fig. 23.

One encounters substantial ambiguity during the segmentation process when the object allows more than one interpretation so that it is not immediately clear where the object should be cut. In such situations it is beneficial to offer more than one solution and to postpone the resolution of the ambiguity to a later statistical analysis. The data extracted from various possible solutions for the same object are called *variants* of the object under study, see Ritter (2000), Ritter and Gallegos (2000), Ritter and Gallegos (2002). The variants make up an *ambiguous data set*. Whereas, in a classical data set, each object is represented by a single line, it may occupy several lines in an ambiguous data set, see. Fig. 24. It is there where the method of variants catches the ambiguity. Each line labelled with the same object corresponds to some interpretation, the one that comes from the correct interpretation is the (unknown) *regular variant*. It may happen that the correct interpretation is not available or has not been found. Then there is no regular variant and the object must be considered an outlier.

AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAA
AAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAAT
TAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAG
CGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTA
GCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCT
GACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCT
TTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAAGT

Figure 22: The initial section of the genome of E. coli. Possible start and stop codons of genes, ATG and TGA, are indicated. Not every ATG initiates and not every TGA terminates a gene which gives rise to ambiguity.

Figure 23: A human metaphase (left) and the associated karyogram. Automatic segmentation of the metaphase in its 46 chromosomes displayed in the karyogram is not an easy task since the touchings and overlappings may allow several interpretations of the image thus giving rise to ambiguities.

obj1	1.23	2.34
obj2	4.26	3.00
obj3	7.28	7.42
obj4	1.91	2.84
obj5	4.02	3.04
obj6	1.02	2.04

obj1	1.23	2.34
obj1	4.26	3.00
obj2	7.28	7.42
obj2	1.91	2.84
obj2	4.02	3.04
obj3	1.02	2.04

Figure 24: Classical (left) and ambiguous data set

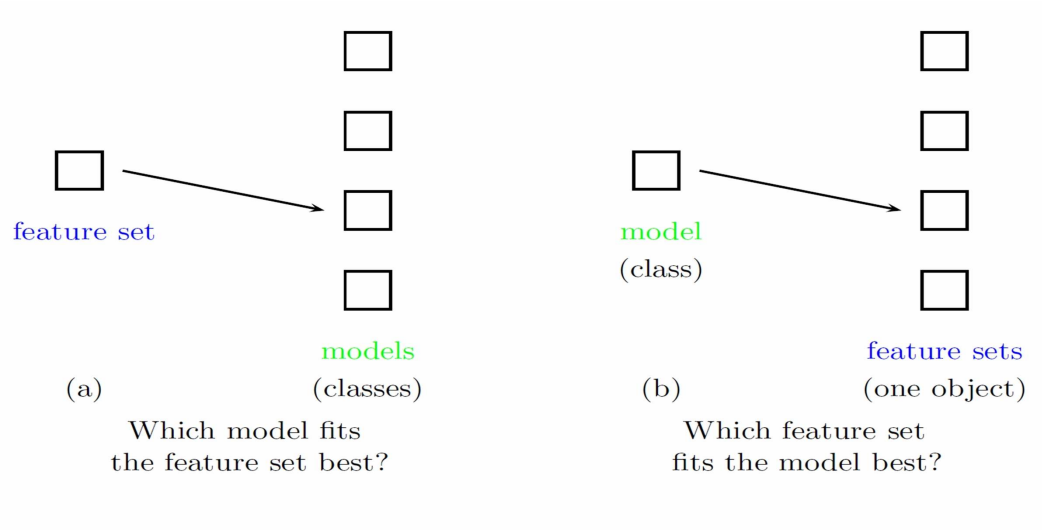


Figure 25: Comparison of classification (left) and variant selection

Of course, ambiguity occurs also in domains other than segmentation. The Fall meeting was in part dedicated to the study of a data set consisting of ancient Roman tiles, see these proceedings. The features of the data set are the contents of nineteen minerals and metals determined by a chemical analysis of probes taken from the tiles. However, the tiles are inhomogeneous so that the probes of some tiles might not be typical, see Dolata and Werr (1998/99). Since this causes errors it would be beneficial to analyze probes taken from several sites on the same tile thus creating variants. The ambiguity can be resolved in the subsequent analysis.

The simplest question about an ambiguous data set is this: given variants of one object, find its regular variant, i.e., discover the correct interpretation given some information on it. It was the subject matter of Ritter and Gallegos (2000). This question is in some sense dual to discriminant analysis, see Fig. 25. All statistical questions that arise in the study of classical data sets can also be asked for ambiguous data sets – parameter estimation, discriminant analysis, clustering, ... Parameter estimation in ambiguous data sets was treated in Ritter and Gallegos (2006), discriminant analysis in Ritter and Gallegos (2000) and in Ritter and Pesch (2001). Variant analysis was applied in several contexts to image analysis, in particular to the problem of chromosome classification, see Ritter and Schreib (2000, 2001), Ritter and Pesch (2001), and Ritter and Gao (2008).

2.2 Parameter estimation in ambiguous data sets

The application studied in Sect 2.3 needs parameter estimation in ambiguous data sets, the subject matter of Gallegos and Ritter (2006). In this section, I review some of the results referring the interested reader to the paper for more details and for proofs.

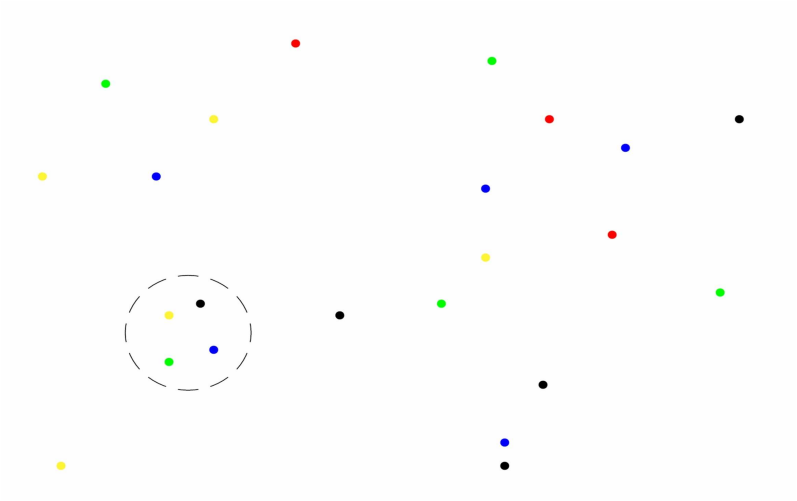


Figure 26: Scatter plot of an ambiguous data set of five objects including an outlier. Each color stands for an object, the regular variants are encircled, and the outlier is plotted in red.

Let E be a sample space. We are given an ambiguous data set $(x_i)_{i=1}^n$ of n objects, object i being observed by b_i variants $x_{i,h} \in E$, $1 \leq h \leq b_i$, i.e., $x_i = (x_{i,1}, \ldots, x_{i,b_i})$. The data set may contain (gross) outliers. Fig. 26 visualizes an ambiguous data set with an outlier. The task is to estimate the positions of the regular variants in the data set and the outliers. If this has been successfully performed then it is a classical task to estimate the parameters of the regular variants. However, it is not possible to estimate these positions without knowing the model and vice versa. As a way out of this deadlock we estimate both simultaneously. To this end, we establish a statistical model of the ambiguous data set with outliers. We assume that the number of regular elements is at least r ($\leq n$). There are several models, the simplest is the *spurious variants model* which we use here. It postulates that the irregular variants of an object contain no information on its regular variant. Denoting the (unknown) set of regular objects by $R \subseteq 1..n$, we assume here for simplicity $\#R = r$ and start with the basic *ordered* model of an object $i \in R$ with (observed) number b_i of variants,

$$Z_i = (Z_{i,1}, \ldots, Z_{i,b_i}) : \Omega \to E^{b_i}.$$

Here, $Z_{i,1}$ stands for the regular variant of the regular object i. It is distributed according to some probability with density function f_γ on E, $\gamma \in \Gamma$, and the family $(Z_{i,1})_{i \in R}$ of regular variants is assumed to be statistically independent. A simple way of dealing with spuriousness is to assume that the $(b_i - 1)$-tuple of irregular variants is "flat" given the regular. Moreover, a spurious *outlier* $i \in 1..n \setminus R$ is an object that lacks a regular variant and we assume that Z_i is altogether "flat."

Since we do not know the position of the regular variant of object i, we observe Z_i only in disorder, i.e., we observe $X_{i,k} = Z_{i,T_i(k)}$ for some random permutation T_i in

$\mathcal{S}_{b_i}$. The items to be estimated from these observations are the parameter $\gamma \in \Gamma$ of the regular population, the set of regular elements R, and the *variant selection* $\mathbf{h} = (h_i)_{i \in R}$, i.e., the regular variant h_i of all $i \in R$. The variant selection $\mathbf{h}$ is a partial function on $1..n$ with support $\operatorname{supp} \mathbf{h}$. It is shown in Gallegos and Ritter (2006) that, under some natural conditions of independence, the logarithm of the *trimmed likelihood function* of the present model is

$$\log f_{\mathbf{h},\gamma}(x_1, \ldots, x_n) = \sum_{i \in R} \log f_\gamma(x_{i,h_i}). \tag{15}$$

The likelihood function is the criterion to be optimized w.r.t. all $\gamma \in \Gamma$, all r–element subsets $R \subseteq 1..n$, and all r–tuples $(h_i)_{i \in R}$. This looks like a formidable task, but there is the following proposition which justifies an efficient, alternating algorithm. The proposition detects an improvement of the criterion before the new parameters are computed.

Proposition. Let $\mathbf{h}$ and $\mathbf{h}_{\text{new}}$ be two variant selections s. th.

$$\sum_{i \in \operatorname{supp}(\mathbf{h}_{\text{new}})} \log f_{\gamma(\mathbf{h})}(x_{i,h_{\text{new},i}}) > \sum_{i \in \operatorname{supp}(\mathbf{h})} \log f_{\gamma(\mathbf{h})}(x_{i,h_i}). \tag{16}$$

Then the parameters computed from the "new" variant selection strictly increase the Criterion (15).

The proposition suggests the following

Reduction step.

Input: a selection $\mathbf{h}$;
Output: a selection with larger Criterion (15) *or* the signal "stop."

(i) Compute the ML–estimate $\gamma(\mathbf{h})$ for the $\mathbf{h}$–regular observations;

(ii) compute the log–density of each variant (i, k), $i \in 1..n$, $k \in 1..b_i$, w.r.t. $\gamma(\mathbf{h})$;

(iii) for each object i, determine the variant with the maximum value;

(iv) the r largest maxima determine R_{new} and the selection $\mathbf{h}_{\text{new}}$;

(v) *if* $\mathbf{h}_{\text{new}}$ satisfies (16), return $\mathbf{h}_{\text{new}}$; *else* "stop."

There is the following theorem.

Theorem. If the reduction step does not output the "stop" signal then it improves the Criterion (15).

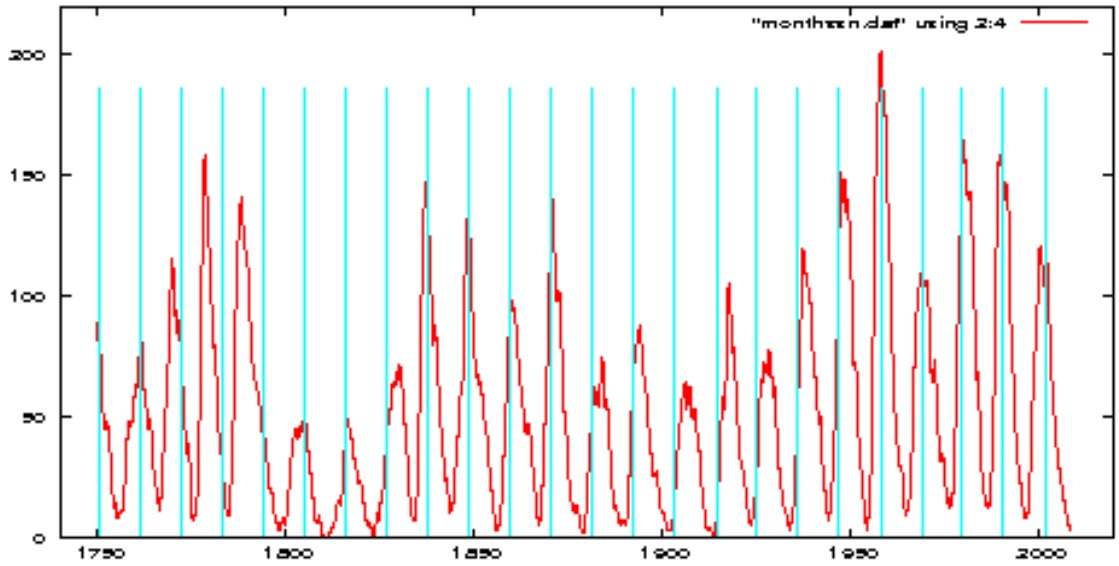


Figure 27: The time series of the smoothed numbers of sunspots. The equidistant bars show its non-periodicity.

The reduction step alternates parameter estimation, variant selection, and trimming thereby improving the criterion. In an overall algorithm one iterates reduction steps. Since there are only finitely many variant selections $\mathbf{h}$, the iteration must come to a standstill with the "stop"–signal at a selection–parameter pair that fit each other. The selection may be called a *minimum–distance selection* (MDS). The minimum of the criterion is an MDS but there are many others so that the algorithm has to be replicated, possibly many times, with random or purposefully chosen initial selections in order to attain a high value of the criterion.

All universal optimization paradigms, such as local ascent, the Metropolis algorithm, Gibbs–sampling, and genetic algorithms, too, may be applied to the present problem of optimizing Criterion (15).

Of course, classical data sets (with or without gross outliers) are contained in the present set-up by way of $b_i = 1$ for all i. In the normal case, Criterion (15) extends Rousseeuw's (1985) minimum covariance determinant estimator (MCD) for robust estimation of the covariance matrix in classical data sets. Still in this case, the reduction step extends Rousseeuw and Van Driessen's (1999) for robust parameter estimation.

2.3 Application: segmentation of a random cyclic process

The foregoing theory may be applied to the segmentation of a random cyclic process: the smoothed, monthly values of the numbers of sunspots observed since the year 1749. They are found under the URL `www.sidc.be/` . The time series is cyclic

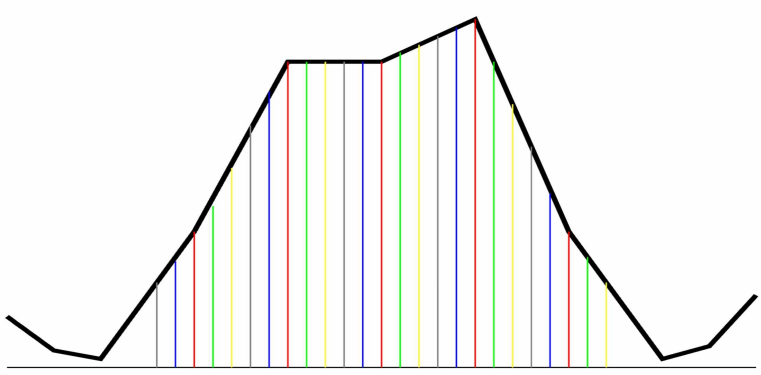


Figure 28: Variant extraction. In this graphic, a sunspot cycle is sampled equidistantly in five ways at the locations shown in different colors. Each color corresponds to a variant.

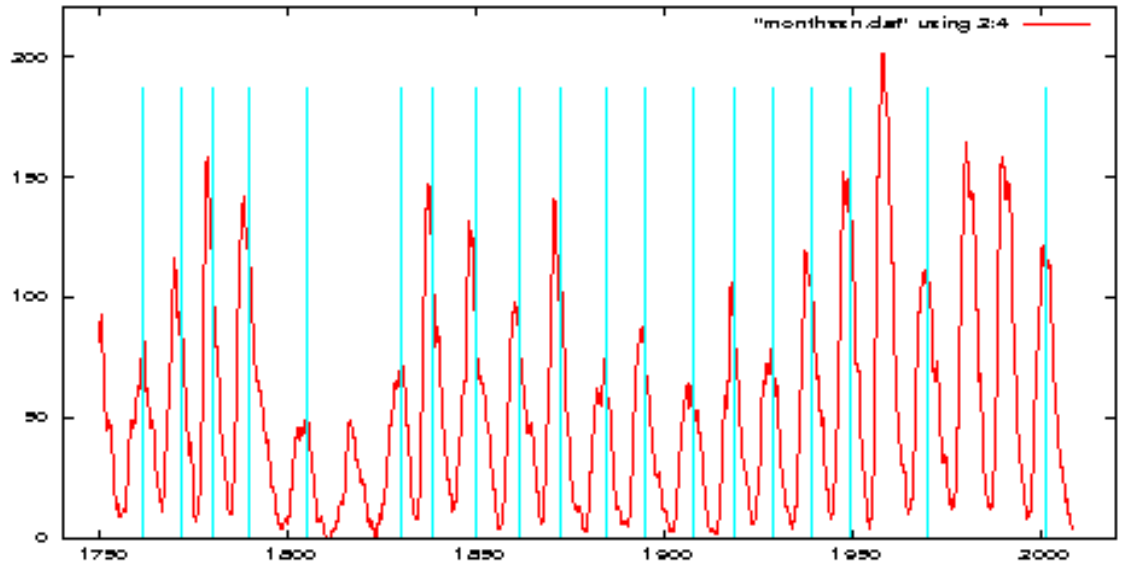


Figure 29: The sunspot cycles determined by variant analysis with four outliers.

with a cycle time of about eleven years but not quite periodic as is seen from the positions of the equidistant bars in Fig. 27. Some of them hit the peaks, others are right in between. The aim is to synchronize the 23 full cycles. To this end, we consider them the objects of our analysis. We sample each cycle at all points of the time pattern -48 -36 -24 -15 15 24 36 48 months around each peak to obtain an 8D-observation. This pattern is shifted by $-35, -30, -25, \ldots, 35$ months in order to generate fifteen variants of each cycle, see a different pattern with shifts indicated by different colorings in Fig. 28. In this way each spike is represented by fifteen 8-dimensional variants and the ambiguity about the cycle phases is caught in this ambiguous data set. Variant selection and parameter estimation as described in Sect. 2.2 matches the most typical cycle phases thereby synchronizing the cycles. The algorithm was run with four discarded objects and the result is presented in Fig. 29. Interestingly, three high, good-looking cycles are discarded – although they look complete they are atypical. It is rather broken spikes that are the rule. Also a small cycle with a thin peak is discarded.

Acknowledgment. I thank Dr. H.-G. Bartel for pointing out Dolata and Werr (1998/99) to me.

References

CASEY, R.G. and LECOLINET, E. (1996): A Survey of Methods and Strategies in Character Segmentation. *IEEE Trans. Patt. Anal. Mach. Int., 18, 690–706.*

DOLATA, J. and WERR, U. (1998/1999): Wie gleich ist derselbe? - Homogenität eines römischen Ziegels und Aussagegrenzen geochemischer Analytik aufgrund von Messtechnik und Materialvarietät. *Mainzer Archäologische Zeitschrift 5/6, 129-147.*

GALLEGOS, M.T. and RITTER, G. (2006): Parameter estimation under ambiguity and contamination with the spurious model. *J. Multivariate Analysis, 97, 1221–1250.*

MAJOROS, W.H. (2007): Methods of Computational Gene Prediction. Cambridge University Press.

RABINER, L.R. (1989): A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE, 77, 257–286.*

RITTER, G. (2000): Classification and clustering of objects with variants. In: Wolfgang Gaul, Otto Opitz, and Martin Schader (Eds.): *Data Analysis, Scientific Modeling and Practical Application.* Springer, Berlin-Heidelberg, 41–50.

RITTER, G. and GALLEGOS, M.T. (2000): A Bayesian approach to object identification in pattern recognition. In: A. Sanfeliu et al. (Eds.): *Proceedings of the 15th International Conference on Pattern Recognition vol. 2.* Barcelona, 418–421.

RITTER, G. and GALLEGOS, M.T. (2002): Bayesian object identification: variants. *Journal of Multivariate Analysis, 81, 301–334.*

RITTER, G. and GAO, L. (2008): Automatic segmentation of metaphase cells based on global context and variant analysis. *Pattern Recognition, 41, 38–55*

RITTER, G. and PESCH, C. (2001): Polarity–free automatic classification of chromosomes. *Computational Statistics and Data Analysis, 35, 351–372*

RITTER, G. and SCHREIB, G. (2000): Profile and feature extraction from chromosomes. In: A. Sanfeliu et al. (Eds.): *Proceedings of the 15th International Conference on Pattern Recognition vol. 2.* Barcelona, 287–290.

RITTER, G. and SCHREIB, G. (2001): Using dominant points and variants for profile extraction from chromosomes. *Pattern Recognition, 34, 923–938.*

ROUSSEEUW, P.J. (1985): Multivariate estimation with high breakdown point. In: Grossmann, W. et al. (Eds.): *Mathematical Statistics and Applications vol. 8B* Dordrecht etc., 283–297.

ROUSSEEUW, P.J. and VAN DRIESSEN, K. (1999): A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41, 212–223.*

3 Archäometrische Daten römischer Ziegel aus *Germania Superior*

Hans–Georg Bartel
Institut für Chemie, Humboldt-Universität zu Berlin
Brook-Taylor-Straße 2, D-12489 Berlin
hg.bartel@yahoo.de

Abstract

Roman stamped bricks and tiles are under investigation coming from different findspots in *Germania Superior*. Their chemical composition was measured by X-ray fluorescence analysis. Here we describe both the archaeological background and the research aims. A data set on the chemical composition of 613 bricks and tiles was carried out using methods of cluster analysis, and the results obtained were interpreted archaeologically.

Keywords: archaeometry, X-ray fluorescence analysis, cluster analysis

3.1 Zum Inhalt der Wissenschaftsdisziplin Archäometrie

Es ist bei der Besprechung archäometrischer Daten sicher nicht ohne Wert, einleitend den Inhalt und das Anliegen der Archäometrie darzulegen. Der Name selbst – obwohl von den altgriechischen Wörtern *archaîos* (alt, altehrwürdig) und *metréō* (messen) ableitbar – ist nicht antik, vielmehr geht er auf das erste Organ dieses Wissenschaftszweiges zurück: die Oxforder Zeitschrift »*Archaeometry*«, die im März 1958 erstmals erschien und sich – und damit zugleich den durch sie repräsentierten Wissenschaftszweig – heute mit *"Archaeometry is an international research journal covering the application of the physical and biological sciences to archaeology and the history of art. The topics covered include dating methods, artifact studies, mathematical methods, remote sensing techniques, conservation science, environmental reconstruction, biological anthropology and archaeological theory."* [13] definiert. Hierbei müssen unter den *physical sciences* neben der Physik auch die Chemie, Mineralogie, Geologie, Astronomie und andere, von den biologischen und medizinischen Richtungen verschiedene Naturwissenschaften verstanden werden. Weiterhin kommen in der Archäometrie viele Technikwissenschaften zur Anwendung, und auf der Seite der Geisteswissenschaften ist in dieser Hinsicht außer der Mathematik zumindest noch die Philologie zu erwähnen. Fig. 30 zeigt vereinfacht die „Stützen" der Archäometrie, die sich in naturwissenschaftliche (einschließlich technische), geisteswissenschaftliche und zwischen diesen beiden einzuordnende unterscheiden lassen.

Die Archäometrie im modernen, derzeitigen Sinne hat sich erst in den 1950er Jahren herausgebildet und etabliert, ihre Anfänge gehen aber bereits auf das letzte Jahrzehnt

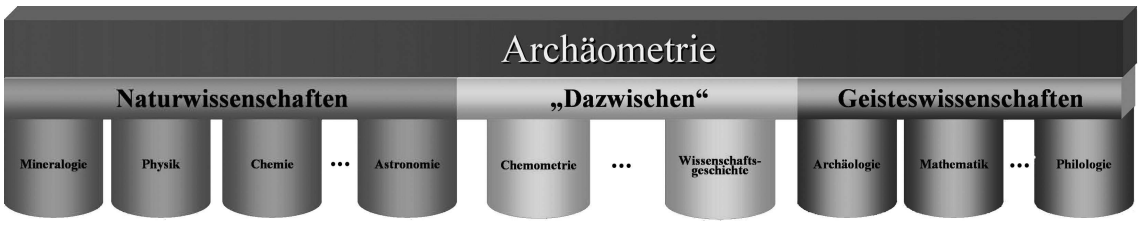
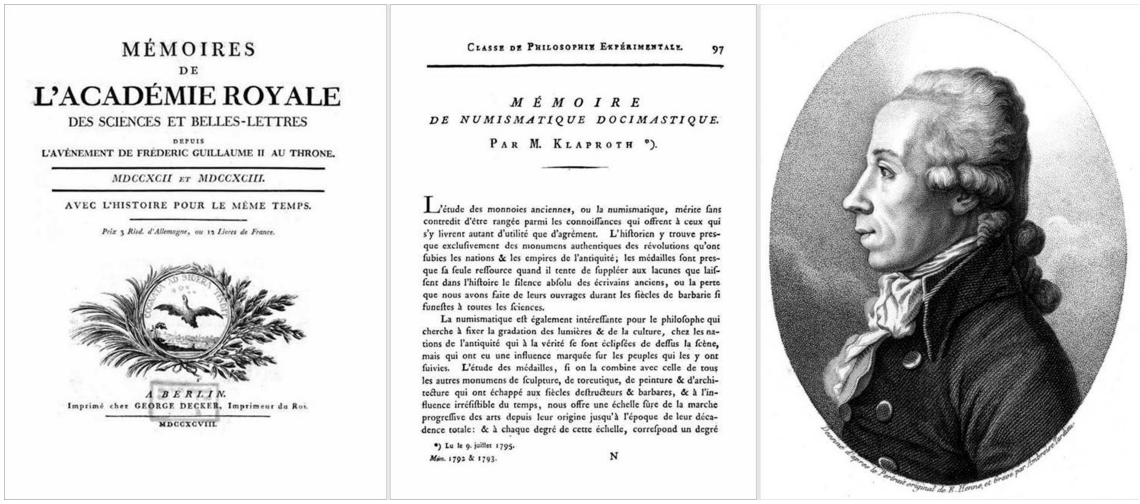


Figure 30: Die „Stützen“ der Archäometrie


MÉMOIRES
DE
L'ACADÉMIE ROYALE
DES SCIENCES ET BELLES-LETTRES
DEPUIS
L'AVÉNEMENT DE FREDERIC GUILLAUME II AU THRONE.
MDCCXCII ET MDCCXCIII.
AVEC L'HISTOIRE POUR LE MÊME TEMPS.
A BERLIN.
Imprimé chez GEORGE DECKER, Imprimeur du Roi.
MDCCXCVIII.

CLASSE DE PHILOSOPHIE EXPÉRIMENTALE. 97

MÉMOIRE
DE NUMISMATIQUE DOCIMASTIQUE.
PAR M. KLAPROTH *).

L'étude des monnoies anciennes, ou la numismatique, mérite sans contredit d'être rangée parmi les connoissances qui offrent à ceux qui s'y livrent autant d'utilité que d'agrément. L'historien y trouve presque exclusivement des monumens authentiques des révolutions qu'ont subies les nations & les empires de l'antiquité; les médailles sont presque sa seule ressource quand il tente de suppléer aux lacunes que laissent dans l'histoire le silence absolu des écrivains anciens, ou la perte que nous avons faite de leurs ouvrages durant les siècles de barbarie si funestes à toutes les sciences.

La numismatique est également intéressante pour le philosophe qui cherche à fixer la gradation des lumières & de la culture, chez les nations de l'antiquité qui à la vérité se sont éclipsées de dessus la scène, mais qui ont eu une influence marquée sur les peuples qui les y ont suivies. L'étude des médailles, si on la combine avec celle de tous les autres monumens de sculpture, de toreutique, de peinture & d'architecture qui ont échappé aux siècles destructeurs & barbares, & à l'influence irrésistible du temps, nous offre une échelle sûre de la marche progressive des arts depuis leur origine jusqu'à l'époque de leur décadence totale: & à chaque degré de cette échelle, correspond un degré

*) Lu le 9. juillet 1795.
Mém. 1792 & 1793. N

Figure 31: Martin Heinrich Klaproth und die erste Publikation archäometrischen Inhalts

des 18. Jahrhunderts zurück. Es handelt sich dabei um chemisch-analytische Untersuchungen von griechischen und römischen Münzen, welche der Chemiker und Apotheker Martin Heinrich Klaproth (1743–1817) (Fig. 31) in seinem Berliner Privatlaboratorium vorgenommen hatte. An der Königlich Preußischen Akademie der Wissenschaften zu Berlin, deren ordentliches Mitglied Klaproth seit 1788 war, stellte er 1795 seine diesbezüglichen Ergebnisse vor. Dieser erste Vortrag archäometrischen Inhalts, «*lu le 9. juillet 1795*», ist 1798 in französischer [3] (Fig. 31) und ein Jahr später in deutscher Sprache [4] publiziert worden.

Von den zitierten, in [13] genannten *“topics”* der Archäometrie muss Klaproths erste und nachfolgende Arbeiten sowie eine Anzahl weiterer anderer Forscher im 19. Jahrhundert zu den *“artifact studies”* gerechnet werden, die man erweiternd als Gebiet der archäometrischen Materialuntersuchungen verallgemeinern kann. Diese Richtung wird für die hier zu schildernden Untersuchungen von Bedeutung sein. Von spezielleren Arbeitsrichtungen wie der Archäoastronomie abgesehen, lässt sich die Archäometrie neben den Materialuntersuchungen in zwei weitere Hauptrichtungen untergliedern, die hier nur genannt, in den weiteren Ausführungen aber keine Rolle spielen werden. Hierher gehören die naturwissenschaftliche Datierung bzw.

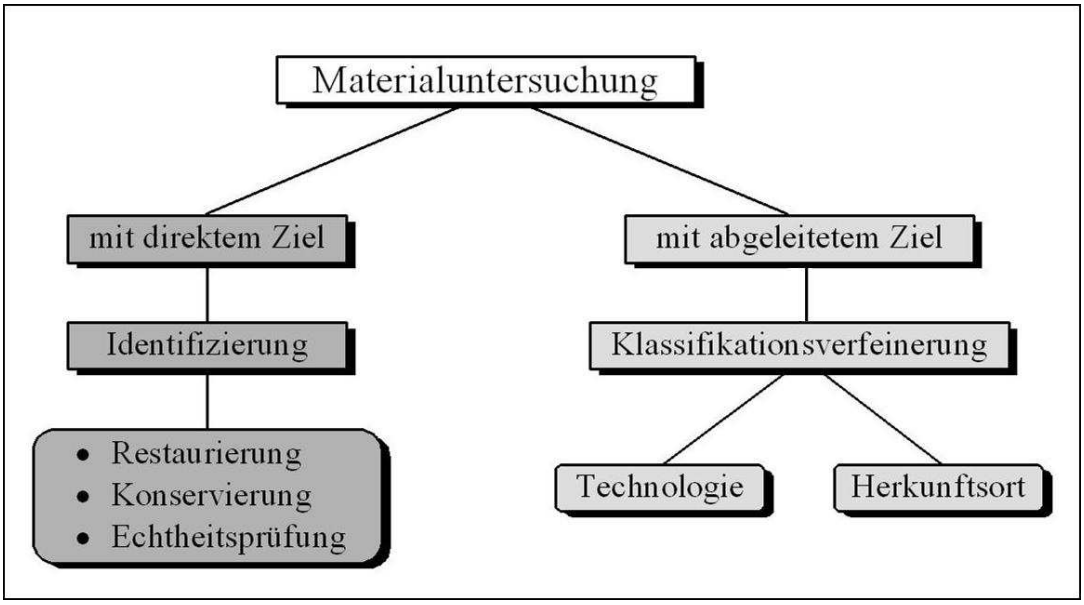


Figure 32: Ziele der archäometrischen Materialuntersuchung

Altersbestimmung – etwa vermittels ^{14}C-, Thermolumineszenz-Methode u.a. – und die Prospektion (lat. *prospicio* = voraussehen, ausschauen – also die Erkundung und Erfassung archäologischer Stätten eines Gebietes), wobei neben anderen an die Bio- und geophysikalische Prospektion (Sondierung von Unbekanntem unter der Erdoberfläche) zu denken ist.

3.2 Ziele archäometrischer Materialuntersuchungen

„Das Gebiet der Materialuntersuchungen ist das am weitesten verzweigte und aufgefächerte Teilgebiet der Archäometrie.“ stellte Hans Mommsen fest ([6], S. 65). Fig. 32, in der die beiden wesentlichen Ziele dieser Richtung aufgezeigt werden, stellt gewissermaßen graphisch verdichtete Aussagen Mommsens dar ([6], S. 12–14, 66).

Materialuntersuchungen mit der direkten Zielsetzung der Identifizierung sind im Zusammenhang mit dem hier Darzustellenden ohne Interesse. Die zu beschreibenden Untersuchungen von Ziegeln auf der Grundlage ihrer Materialanalyse und deren mathematisch-statistische Auswertung verfolgt das abgeleitete Ziel einer verfeinerten Klassifizierung. Dabei sollte die Verfeinerung darin bestehen, die bekannten Klassifikationsmerkmale der betrachteten Ziegel wie beispielsweise ‘grobe Tonkeramik’, ‘römisch’ oder ‘geziegelt von Legion Nr. . . . ’ durch die Angabe ihrer Provenienz, d.h. ihres Herstellungsortes zu präzisieren. Die Frage nach der Herstellungsart bzw. der Technologie der Ziegelproduktion spielte hier keine Rolle.

Rohstoff	Ein *anorganisch-nichtmetallisches pulvriges* Material
Prozess I (Formung)	wird *plastisch* geformt
Prozess II (Brennen)	und anschließend durch *Wärmeeinwirkung* (Hitze)
Produkt	in einen *irreversibel verfestigten* Werkstoff umgewandelt.

Figure 33: Definition des Werkstoffs Keramik

3.3 Zum Werkstoff Keramik

Da die zugrunde liegende Materialart die Auswahl und Anwendung der heranzuziehenden Untersuchungs- oder Analysenmethode bzw. -methoden und in gewissem Umfang auch der benutzten mathematisch-statistischen Verfahren der Auswertung beeinflusst, sei eine kurze Betrachtung des Materials Keramik, aus dem (gebrannte) Ziegel bestehen, vorgenommen. Die Definition eines Keramikwerkstückes lässt sich durch Charakterisierung des Ausgangsstoffes, der beiden grundsätzlichen Prozess-Schritte bei der Herstellung und des Endproduktes vornehmen, wie es in der Übersicht (Fig. 33) entnommen werden kann. Dabei sind die wesentlichen Aussagen kursiv gesetzt.

Bildet das Lockergestein Ton (plastische Bestandteile: verschiedene Tonmineralien, nichtplastische Bestandteile: Quarz, Feldspäte, Calcit, Dolomit, Glimmer u.a.) den wichtigen Bestandteil des anorganischen Rohstoffpulvers, so heißt der erzeugte Werkstoff Tonkeramik oder auch einfach nur Keramik. Gebrannte Ziegel, gefertigt aus Ton, Lehm, Löß oder Letten, gehören also zu dieser Werkstoffklasse. Andere Keramikarten, wie etwa die Oxidkeramiken, zu welchen die bereits in der Antike gefertigte Quarzkeramik zählt, brauchen hier folglich nicht betrachtet zu werden.

Das Material der Tonkeramik, der Scherben, ist inhomogen aus verschiedenen kristallinen und amorphen (glasigen) Bestandteilen zusammengesetzt. Sind alle Inhomogenitäten mit bloßem Auge nicht erkennbar (Korngrößen $< 0{,}2$ mm), so spricht man von Feinkeramik, anderenfalls von Grobkeramik. Die nach [2] gestaltete Übersicht (Fig. 34) zur Einteilung der tonkeramischen Werkstoffe verdeutlicht, dass es sich bei gebrannten Ziegeln um einen porösen grobkeramischen Werkstoff handelt.

Diese Tatsache muss beispielsweise bei den Probenahmen beachtet werden, da ein zu kleiner Bereich, aus welchem Material entnommen wird bzw. welcher einer analytischen Untersuchung unterzogen wird, nicht repräsentativ für die interessierende mittlere chemische Zusammensetung oder andere Eigenschaften des Werkstücks zu sein braucht.

Dieser Inhomogenität der Ziegel ist eine Arbeit von J. Dolata und U. Werr [1] gewidmet, in welcher auf einem einzigen *later* an zehn verschiedenen Stellen relativ große Probemengen entnommen (Fig. 35) und analysiert wurden. Zu dieser Publikation [1] siehe auch den Beitrag von Gunter Ritter in diesem Report (Sect. 2).

Tonkeramische Werkstoffe	**grob**	**porös** (WAF > 6%)	**Ziegel** (gebrannt), ...
		dicht (WAF ≤ 6%)	Klinker, ...
	fein	porös (WAF > 2%): Tongut	Irdengut
			Steingut
		dicht (WAF ≤ 2%): Tonzeug	Steinzeug
			Porzellan

Figure 34: Einteilung der keramischen Werkstoffe (WAF: Wasseraufnahmefähigkeit)

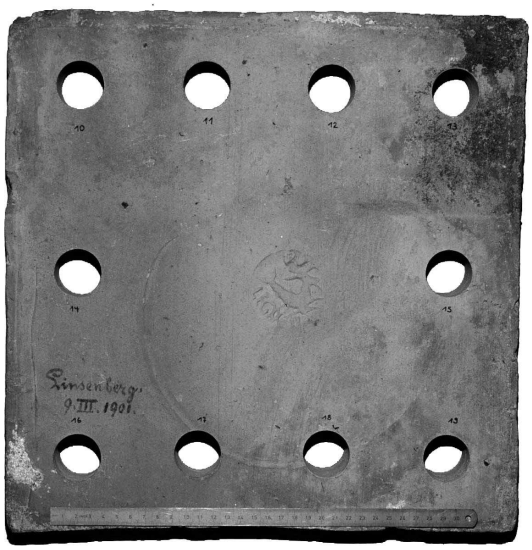


Figure 35: Auf Homogenität bzw. Inhomogenität beprobter Ziegel [1] (*later* der *LEG*(*egionis*) *XXII P*(*rimigeniae*) *P*(*iae*) *F*(*idelis*), hadrianisch, Herstellungs-Provenienz: Frankfurt-Nied, 38 x 38 x (4,5-5) cm)

3.4 Archäologische Fragestellung und Bestimmung der archäometrischen Daten

Die untersuchten römischen Ziegel (*tegulae*, *lateres*, *imbrices*, *tubuli* etc.) stammen aus Fundorten, die auf dem Gebiet der römischen Provinz *Germania Superior* (Obergermanien) liegen. Wie es in den Provinzen üblich war, wurden die Ziegel in Heeresziegeleien produziert. Die damit beauftragten militärischen Einheiten (Legionen, Kohorten) versahen eine größere Anzahl ihrer Erzeugnisse mit ihrem Stempel. Die hier in Betracht gezogenen Ziegel trugen eine solche Kennzeichnung der jeweiligen Militäreinheit, aus der sich mehrere archäologische Informationen ablesen lassen.

Da der Fundort eines Ziegels im Allgemeinen nicht mit seinem Herstellungsort übereinstimmt, bestand die an die Archäometrie gestellte Frage darin, mit naturwissenschaftlichen und mathematischen Methoden Aussagen zur Lokalisierung von Heeresziegeleien in Obergermanien zu erarbeiten. Aus der Kenntnis dieser Provenienzen und der ihnen zugeordneten Ziegel bzw. Stempeltypen etc. sind weitere archäologische und historische Feststellungen zu erwarten.

Zur Lösung dieser Aufgabe wurden 613 Ziegel beprobt und deren chemische Zusammensetzung ermittelt. Da man davon ausgehen kann, dass eine Tonlagerstätte ein typisches chemisches Zusammensetzungsmuster (CZM) besitzt, das sich von dem anderer unterscheidet und dass diese Lagerstätte die Rohstoffquelle einer gesuchten, in ihrer Nähe liegenden Heeresziegelei ist, sollte es möglich sein mit Hilfe von Referenzmaterial und durch Einsatz von Verfahren der automatischen Klassifikation Gruppierungen der Ziegel zu erhalten, die einer in ihrer Lokalität bekannten oder auch unbekannten Ziegelei zugeordnet werden können. Mit anderen Worten: Es gilt die Annahme, dass eine Rohstoffquelle/Tongrube mit typischem, durch den Produktionsprozess unverändertem CZM mit einem Ziegeleiort korrespondiert.

Als Methode für die chemische Analyse wurde die wellenlängendispersive Röntgenfluoreszenzanalyse (WD-RFA) ([6], S. 99–107) benutzt. Hier wird die Erscheinung der Fluoreszenz ausgenutzt, deren Zustandekommen in einem nach dem polnischen Physiker Alexander Jablonski (1898–1980) benannten Diagramm sehr schematisch erklärt wird (Fig. 36): Die Absorption von elektromagnetischer Strahlung $h\nu_A$ (h: Plancksches Wirkungsquantum, ν: Frequenz) – im Falle der RFA Röntgenlicht – bewirkt eine Anregung von Elektronen aus Niveaus niedrigere Energie in solche höherer Energie. Die Rückkehr in den energetisch bevorzugten Zustand geschieht bei der Fluoreszenz in zwei Schritten, einem strahlungslosen Übergang in ein Niveau niedrigerer Energie und der nachfolgenden Rückkehr in den Ausgangszustand unter Emission der Fluoreszenzstrahlung $h\nu_F$ mit $\nu_F < \nu_A$.

Da eine bestimmte Atomsorte durch eine typische Menge von Energieniveaus $E_i = h\nu_i = \frac{hc}{\lambda_i}$ (c: Lichtgeschwindigkeit, λ: Wellenlänge) charakterisiert ist, lässt sich beim Auftreten bestimmter Frequenzen bzw. Wellenlängen bei der Fluoreszenz (charakteristische Röntgenlinien) auf das Vorhandensein entsprechender Atomsorten bzw. chemischer Elemente schließen (qualitative Analyse). Die Intensität der jeweilig

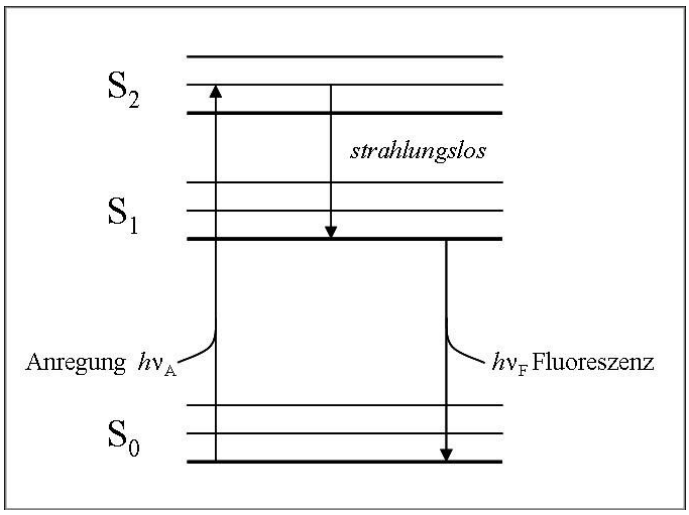


Figure 36: Schematisches Jablonski-Diagramm zur Erläuterung der Fluoreszenz

zugehörigen Strahlung ist der Anzahl der Atome des Elements proportional (quantitative Analyse). In einem Röntgen-Fluoreszenzspektrum (Abszisse/Qualität: Energie, Frequenz, Wellenlänge; Ordinate/Quantität: Intensität) somit die Informationen über die in einer untersuchten Probe vorhandenen chemischen Elemente und deren Menge (Gehalt) enthalten.

Das Schema einer WD-RFA-Messanordnung, mit der solche Spektren aufgenommen werden können, wobei die Empfindlichkeit und Selektivität größer ist als der alternativen energiedispersiven RFA, zeigt Fig. 37. Die Wirkung des Analysator- oder Bragg-Kristalls beruht auf der Anwendung der nach William Henry Bragg (1862–1942) und William Lawrence Bragg (1890–1971) benannten Gleichung $n\lambda = 2d \sin\theta$, wobei d der Abstand der Gitterebenen des Kristalls und n die (ganzzahlige) Interferenzordnung sind. Damit ist die Wellenlängen λ und somit die Energie mit dem durch Drehung des Bragg-Kristalls einstellbaren Glanzwinkel θ in direkte Verbindung gebracht. Der gemäß dem Reflexionsgesetz nachzustellende Detektor misst die auftreffende Menge der Röntgenstrahlung.

Die Vermessung der 613 Ziegel-Proben mit der WD-RFA wurde im Laboratorium von Gerwulf Schneider an der Freien Universität Berlin durchgeführt.

Entsprechend der im vorhergehenden Absatz erwähnten Inhomogenität wurden den untersuchten Ziegeln verhältnismäßig große zylindrische Proben entnommen (vgl. Fig. 35). Um eine gewisse Homogenisierung des Materials zu erhalten, wurden diese anschließend pulverisiert. Unter Verwendung von Lithiumborat wurden aus dem Pulver die zu vermessenden Schmelztabletten hergestellt. (Die im Lithiumborat enthaltenen Elemente Lithium, Bor und Sauerstoff werden von der RFA nicht erfasst.)

Die ermittelten Masse-Anteile von 19 chemischen Elementen wurden für diese Auswertung benutzt. Dabei handelt es sich um

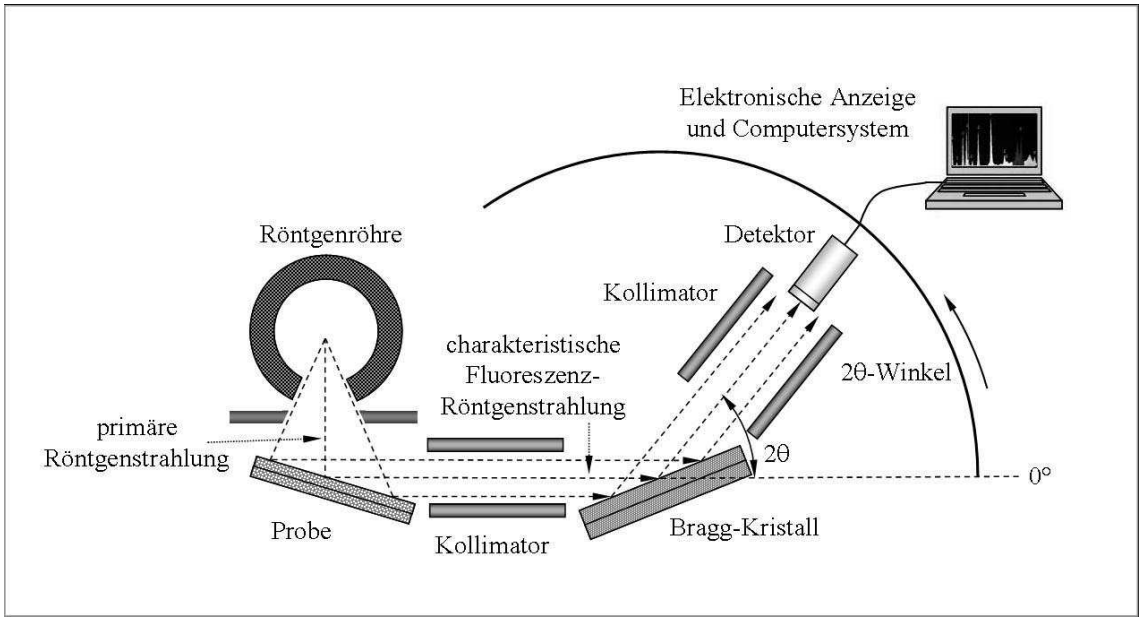


Figure 37: Schema einer Messanordnung der wellenlängendispersiven Röntgenfluoreszenzanalyse

- neun Hauptelemente (Angabe als Oxid, Maßangabe: %): Natrium (Na_2O), Magnesium (MgO), Aluminium (Al_2O_3), Silizium (SiO_2), Kalium (K_2O), Calcium (CaO), Titan (TiO_2), Mangan (MnO), Eisen (Fe_2O_3) sowie
- zehn Spurenelemente (Angabe als Element, Maßangabe: ppm): Vanadium (V), Chrom (Cr), Nickel (Ni), Zink (Zn), Rubidium (Rb), Strontium (Sr), Yttrium (Y), Zirkonium (Zr), Niob (Nb), Barium (Ba).

Einige weitere ebenfalls bestimmte Gehalte wie beispielsweise diejenigen von Phosphor (P_2O_5), Chlor (Cl), Kupfer (Cu) und Zinn (Sn) wurden im Weiteren nicht berücksichtig, da entweder die Messwerte apparativ bedingt zu ungenau sind oder – wie beim Phosphor – eine Kontamination von phosphorhaltigen Substanzen (z.B. Phosphaten) während der Bodenlagerung der Ziegel nicht auszuschließen ist.

Das Ergebnis der archäometrischen Messungen bestand somit in einer 613 x 19-Datenmatrix $\mathbf{X}^{(0)} = (x_{ij}^{(0)})$, die unter Anwendung von Verfahren der multivariaten Statistik, insbesondere der Clusteranalyse ausgewertet werden musste.

3.5 Datenaufbereitung und -auswertung – Clusteranalyse

Beim Betrachten der folgenden Übersicht der (gerundeten) Minimal-, Maximal- und Mittelwerte der 19 Variablen (Fig. 38) stellt man fest, dass der Wertebereich zahlenmäßig, d.h. ohne Berücksichtigung der Maßangabe fast fünf Zehnerpotenzen von 10^{-2} (Minimum MnO) bis $9{,}14 \cdot 10^2$ (Maximum Ba) umfasst. Ähnlich verhält

Variable	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O
Maßangabe	%								
Minimum	51,14	0,43	9,00	2,58	0,01	0,49	0,18	0,11	1,14
Maximum	81,10	2,52	25,10	9,97	0,14	4,14	21,50	1,30	6,07
Mittelwert	68,05	0,90	15,28	4,61	0,06	1,74	5,85	0,61	2,74

Variable	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
Maßangabe	*ppm*									
Minimum	44	47	3	13	69	67	17	103	1	256
Maximum	147	323	123	361	193	474	52	421	65	914
Mittelwert	82	109	49	73	132	201	30	215	21	477

Figure 38: Übersicht über den Wertebereich und die Mittelwerte der 19 Variablen

es sich mit den Mittelwerten: $6{\cdot}10^{-2}$ (MnO) bis $4{,}77{\cdot}10^{2}$ (Ba). Eine Umrechnung der Angaben von Prozent in *ppm* ($1\ \% = 10^4$ *ppm*) oder umgekehrt würde den Zahlenumfang noch erweitern: 1 *ppm* (Minimum Nb) bis $8{,}11{\cdot}10^{5}$ *ppm* (Maximum SiO_2).

Die Werte der einzelnen Variablen mussten also hinsichtlich ihrer Größenordnungen für die Clusteranalyse vergleichbar gemacht werden. Als besonders geeignet hat sich dafür erwiesen, die Werte $x_{ij}^{(0)}$ jeder der 19 Variablen jeweils durch deren Mittelwert $\overline{x}_j^{(0)}$ zu dividieren, d.h., die folgende Transformation

$$\mathbf{X}^{(0)} = (x_{ij}^{(0)}) \longrightarrow \mathbf{X}^{(v)} = (x_{ij}^{(v)})$$

mit

$$x_{ij}^{(0)} \longrightarrow x_{ij}^{(v)} = \frac{x_{ij}^{(0)}}{\overline{x}_j^{(0)}}$$

und

$$\overline{x}_j^{(0)} = n^{-1} \sum_{q=1}^{n} x_{qj}^{(0)} \ (j = 1, 2, \ldots, 19)$$

durchzuführen. Dabei ist n die Anzahl der betrachteten Objekte (Proben), für den gesamten Datensatz beträgt also $n = 613$.

Bei dieser Transformation ist die Standardabweichung der transformierten Variablen gleich dem Variationskoeffizienten der ursprünglichen, d.h. gemessenen. Alle transformierten Variablen haben denselben Mittelwert $\overline{x}_j^{(v)} = 1$.

Zur Lösung der archäologischen Fragestellung nach der Provenienz der Heeresziegeleien wurde der Datensatz $\mathbf{X}^{(v)} = (x_{ij}^{(v)})$ mit $n \leq 613$ Objekten vermittels hierarchischer und partitionierender Verfahren der Clusteranalyse in Klassen zerlegt. Hinsichtlich der jeweils zugrunde liegenden Theorie siehe beispielsweise [5], [7–10]

und [12]. Eine beschreibende Übersicht über die mit den Ziegeluntersuchungen herangezogenen Verfahren ist in [A5] zu finden.

Im Einzelnen wurden hauptsächlich benutzt:

(a) das hierarchische Verfahren nach J.H. Ward [12], basierend auf dem Varianzkriterium

$$V_k = \mathrm{tr}(\sum_{\alpha=1}^{k} \mathbf{W}_\alpha) \longrightarrow \textit{Minimum}$$

$$(\mathbf{W}_\alpha = \sum_{i \in C_\alpha} (\mathbf{x}_i - \overline{\mathbf{x}}_\alpha)(\mathbf{x}_i - \overline{\mathbf{x}}_\alpha)^T\text{: Produktsummenmatrix für die } \alpha\text{-te Klasse } C_\alpha).$$

Das Ward-Verfahren wurde für die Datenanalyse im Rahmen der archäometrischen Ziegeluntersuchungen ausgewählt, da es zu den wenigen hierarchischen Verfahren gehört, die sich aus einem statistischen, auf Verteilungsannahmen über die Daten basierenden Modell ableiten lässt.

(b) das modifizierte Ward-Verfahren mit dem Zielkriterium

$$V_k^{\log} = \sum_{\alpha=1}^{k} n_\alpha \log \mathrm{tr}(\frac{\mathbf{W}_\alpha}{n_\alpha}) \longrightarrow \textit{Minimum}$$

(logarithmiertes gemitteltes Varianzkriterium), wobei n_α die Masse der Klasse C_α bedeutet.

Dieses Verfahren ist geeignet, wenn in einigen Klassen bei großer Variablenanzahl nur geringe Objektanzahl vorhanden ist. Das ist bei den römischen Ziegelproben tatsächlich der Fall.

(c) das auf John B. MacQueen [5] zurückgehende partitionierende k-Means-Verfahren

Dieses Verfahren sucht nach lokalen Minima des unter (a) genannten Varianzkriterium V_k, das die Fehlerterme zu den Klassenmitteln (Zentren) summiert.

(d) die zentrenfreie Variante zum k-Means-Verfahren

Hier werden die unter (a) bzw. (b) genannten Varianzkriteria zentrenfrei formuliert:

$$V_k = \mathrm{tr}(\sum_{\alpha=1}^{k} \mathbf{W}_\alpha) = \sum_{\alpha=1}^{k} \frac{1}{n_\alpha} \sum_{x_i \in C_\alpha} \sum_{x_h \in C_\alpha, h>i} d_{ih}^{E^2}$$

bzw.

$$V_k^{\log} = \sum_{\alpha=1}^{k} n_\alpha \log \mathrm{tr}(\frac{\mathbf{W}_\alpha}{n_\alpha}) = \sum_{\alpha=1}^{k} n_\alpha \log(\sum_{x_i \in C_\alpha} \sum_{x_h \in C_\alpha, h>i} \frac{1}{n_\alpha^2} d_{ih}^{E^2}),$$

wobei

$$d_{ih}^{E^2} = \|\mathbf{x}_i - \mathbf{x}_h\|^2 = \sum_{j=1}^{l} (x_{ij} - x_{hj})^2$$

die quadrierte euklidische Distanz zwischen den Beobachtungen $\mathbf{x}_i$ und $\mathbf{x}_h$ und l die Variablenanzahl sind.

(e) die adaptive Variante des k-Means-Verfahrens

Statt der quadrierten euklidischen Distanz $d_{ih}^{E^2}$ wird die gewichtete quadrierte euklidische Distanz

$$d_{\mathbf{Q}}^2 = \|\mathbf{x}_i - \mathbf{x}_h\|_{\mathbf{Q}}^2 = \sum_{j=1}^{l} q_j |x_{ij} - x_{hj}|^2$$

zwischen den Beobachtungen $\mathbf{x}_i$ und $\mathbf{x}_h$ benutzt, wobei die q_j nichtnegative Gewichte sind, die im Iterationsprozess adaptiv bestimmt werden.

3.6 Auswertung und archäologische Interpretation

In einem Anhang (s.u.) wurden diejenigen Publikationen aufgelistet, die sich mit der Anwendung mathematisch-statistischer Methoden bei Untersuchungen obergermanisch-römischer Ziegel beschäftigen. Bei den letzteren handelt es sich in der Mehrzahl der Fälle um solche, deren archäometrische Daten in der oben beschriebenen Matrix erfasst sind.

In diesem Abschnitt sollen einige der wichtigsten Aussagen und Resultate dargestellt werden, welche aus der multivariaten statischen Analyse dieses Datensatzes ablesbar sind.

Die Clusteranalyse mit dem ursprünglichen Verfahren nach Ward ließ insbesondere zwei Gesichtspunkte deutlich werden: (A) Die gewonnene optimale Zerlegung in acht Klassen entspricht in großen Zügen der archäologischen Erfahrung. (B) Die ermittelten Klassen sind in ihrer Stärke sehr unterschiedlich, so dass es – wie im vorhergehenden Abschnitt erwähnt – sinnvoll ist, das modifizierte Ward-Verfahren zur Anwendung zu bringen. Obwohl das erhaltene Klassifikationsergebnis als eine positive Hypothesenstütze zu werten ist, wurde es im Hinblick auf die Komplexität der archäologischen Fragestellung als *„nur eine Vor-Interpretation von Daten“* aufgefasst und berücksichtigt, dass in diesem Sinne Verfahren der automatischen Klassifikation *„den Charakter von numerischen Experimenten haben, die Korrelationen aufzeigen können aber nicht Kausalitäten.“* [11]. Daher wurde das Resultat durch die archäologische Erfahrung leicht ‘überformt’ [A5], [A9], [A23], wie es der nicht einheitlich durchgehende („schiefe“) Schnitt im Dendrogramm zeigt (Fig. 39).

Auf diese Weise ließen sich insgesamt sieben Orte von Heeresziegeleien erkennen,

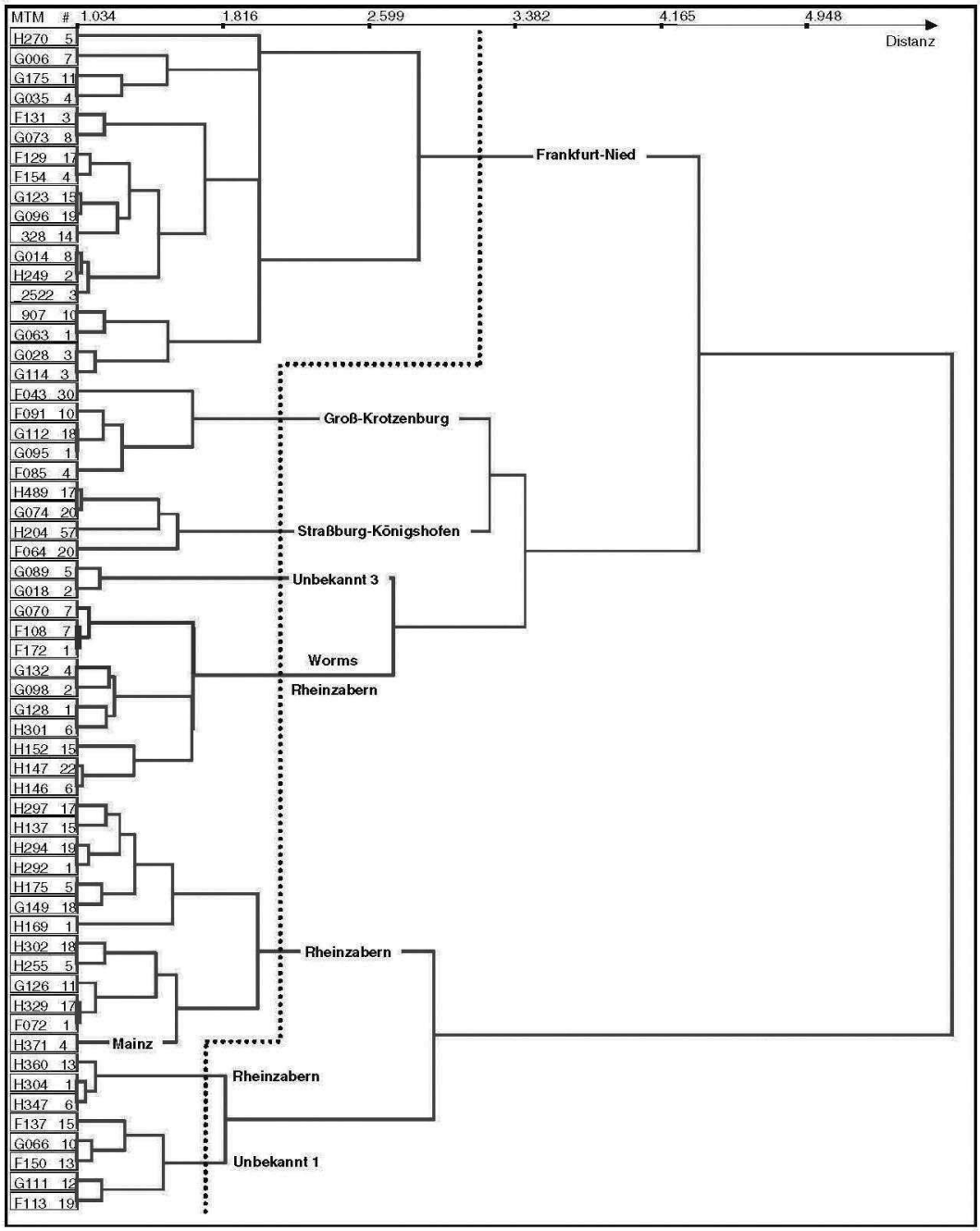


Figure 39: Rechter Teil des Dendrogramms (Ward-Verfahren) mit schiefem Schnitt unter Einbeziehung archäologischen Erkenntnisstandes (MTM: *Most Typical Member*)

Heeresziegelei-Ort	Lateinischer Name	Heutiges Land	Abkürzung	Objektanzahl
Rheinzabern	*Tabernae*	Rheinland-Pfalz	Rh	192
(Frankfurt-)Nied	*Nidda*	Hessen	FN	137
Straßburg-Königshofen	*Argentorate*	Elsass (Frankreich)	SK	113
Unbekannt 1			U1	67
Groß-Krotzenburg		Hessen	GK	63
Worms (ehemals 'Unbekannt 2')	*Borbetomagnus*	Rheinland-Pfalz	Wo	19
Unbekannt 3			U3	7

Figure 40: Zur Lokalisierung von Heeresziegeleien

von denen fünf benennbar und zwei hinsichtlich ihrer Provenienz bisher noch nicht festlegbar sind. Die Übersicht (Fig. 40) präzisiert diese Aussage.

Weiterhin gibt es eine kleine Referenzgruppe mit vier Proben von neuzeitlichen Ziegeln aus Mainz und eine Sonderklasse von elf nicht referenzfähigen Objekten.

Das archäologisch überformte Ergebnis der Clusteranalyse nach Ward ist in Form eines Hauptkomponentenplots in Fig. 41 dargestellt.

Bei der Anwendung des modifizierten Ward-Verfahrens konnten wiederum acht Klassen ('Cluster Nr. 1' bis 'Cluster Nr. 8') als optimale Zerlegung des Datensatzes abgeleitet werden (Ellenbogentest). Fig. 42 verdeutlicht dieses Ergebnis.

In der Pivot-Tabelle (Fig. 43) sind die beiden Clusteranalysen-Resultate gegenübergestellt worden.

Aus dieser Übersicht ist ersichtlich, dass die Klassen 'Straßburg-Könighofen', 'Groß-Krotzenburg' und 'Unbekannt 3' als stabil bezeichnet werden können. Die der Provenienz Frankfurt-Nied zugeordneten Proben lassen sich in zwei Klassen aufspalten. Ähnlich verhält es sich mit der Provenienz Rheinzabern. Allerdings enthält 'Rheinzabern A' die ursprünglich 'Unbekannt 1' zugeordneten Ziegel und 'Rheinzabern B' diejenigen einer Wormser Heeresziegelei. Es konnte aber in einigen Arbeiten, die jeweils den relevanten Teildatensatz analysierten, gezeigt werden, dass die Frankfurt-Nied zugeordneten Ziegel signifikant in zwei Klassen aufgeteilt werden können ([A6], [A17]) und dass sich ebenso sowohl 'Worms' [A10] als auch 'Unbekannt 1' [A12] von den Rheinzabern-Klassen unterscheiden lassen. Schließlich erwies sich die Aufteilung der Ziegel mit der Provenienz Rheinzabern auf zwei Klassen als annehmbar. Diese Aussagen sind auch mit der in Fig. 44 wiedergegebenen zweidimensionalen Dichteschätzung vereinbar.

Um das Ergebnis der Klassifikation mit dem modifizierten Ward-Verfahren zu demonstrieren, wurde die Distanzmatrix $\mathbf{D} = (d_{ij})$ in der Weise graphisch dargestellt, dass bestimmte Intervalle der Distanzwerte durch eine Grauwertstufe wiedergegeben werden. Hier wurde die Festlegung getroffen, steigende Unähnlichkeit in zehn Stufen von Schwarz bis Weiß zu unterscheiden. In Fig. 45 ist die Distanzmatrix auf diese Weise in einem repräsentativen Auszug (251 Objekte) dargestellt, wobei Zeilen und Spalten nach der 1. Hauptachse (Fig. 42) sortiert wurden.

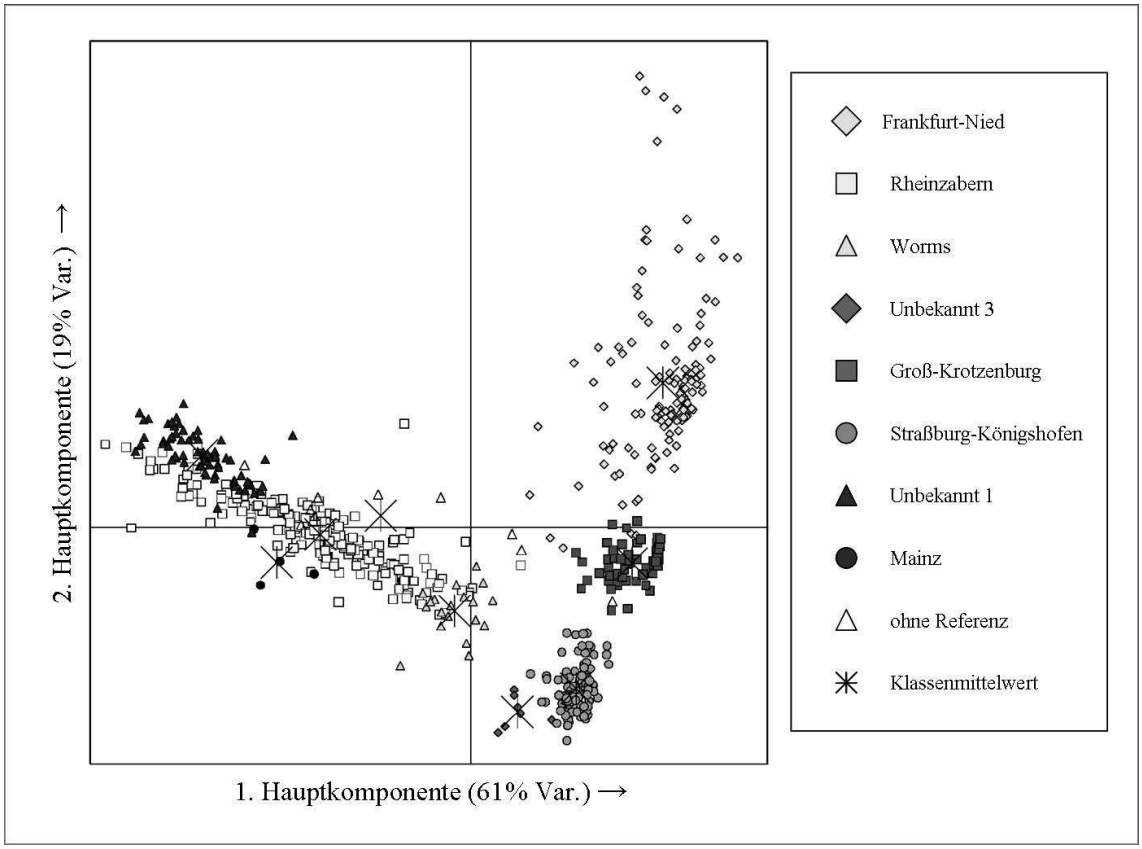


Figure 41: Auftragung der 1. und 2. Hauptkomponente mit Einfärbung der Objekte nach dem archäologisch überformten Klassifikationsergebnis der Clusteranalyse nach Ward (s. Fig. 39)

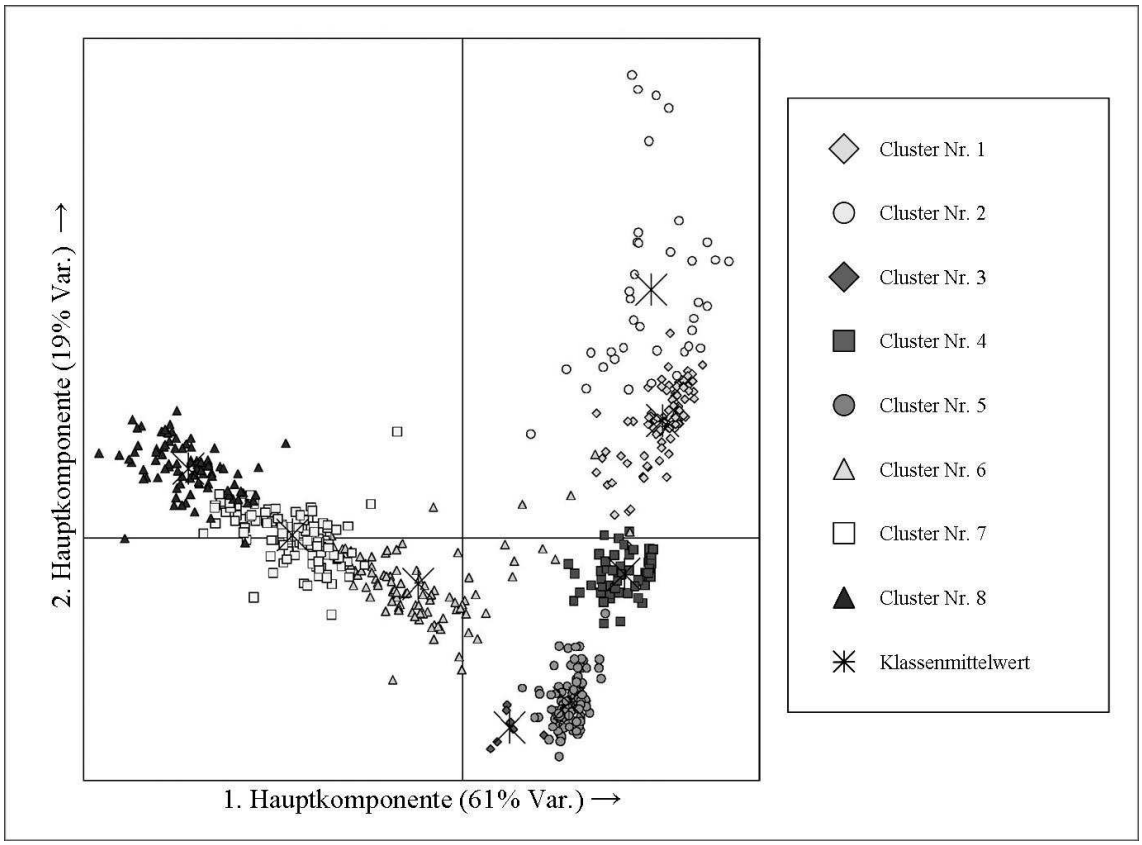


Figure 42: Auftragung der 1. und 2. Hauptkomponente mit Einfärbung der Objekte nach dem mit dem modifizierten Ward-Verfahren erhaltenen Klassifikationsergebnis

			Ward-Verfahren & archäologische Überformung									
			FN	Rh	Wo	U3	GK	SK	Mainz	U1	o. Ref.	Total
modifiziertes Ward-Verfahren	Frankfurt-Nied A	Cluster 1	92									92
	Frankfurt-Nied B	Cluster 2	38									38
	Unbekannt 3	Cluster 3				7						7
	Groß-Krotzenburg	Cluster 4	1				63					64
	Straßburg-Königshofen	Cluster 5						113			1	114
	Rheinzabern B	Cluster 6	6	66	19						3	94
	Rheinzabern A	Cluster 7		106					4		5	115
		Cluster 8		20						67	2	89
	Total		137	192	19	7	63	113	4	67	11	613

Figure 43: Vergleich zweier Klassifikationsresulte

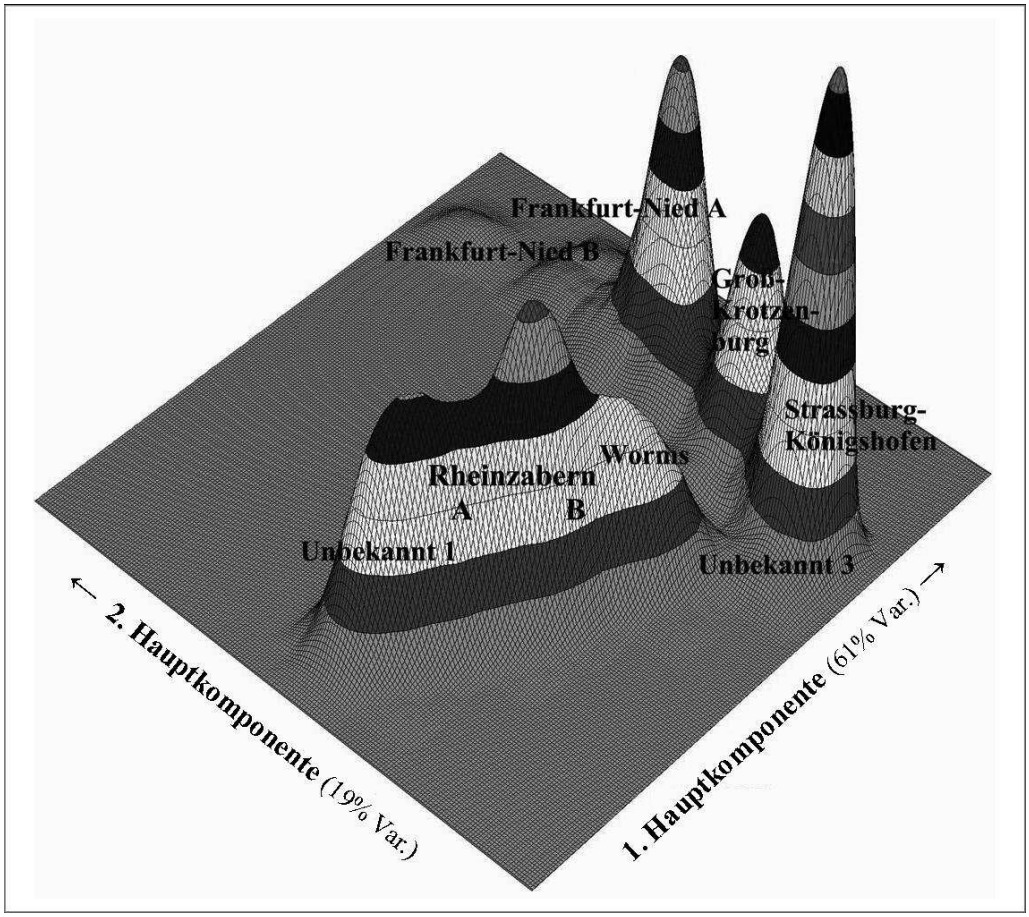


Figure 44: Zweidimensionale Dichteschätzung über der von der 1. und 2. Hauptachse aufgespannten Ebene

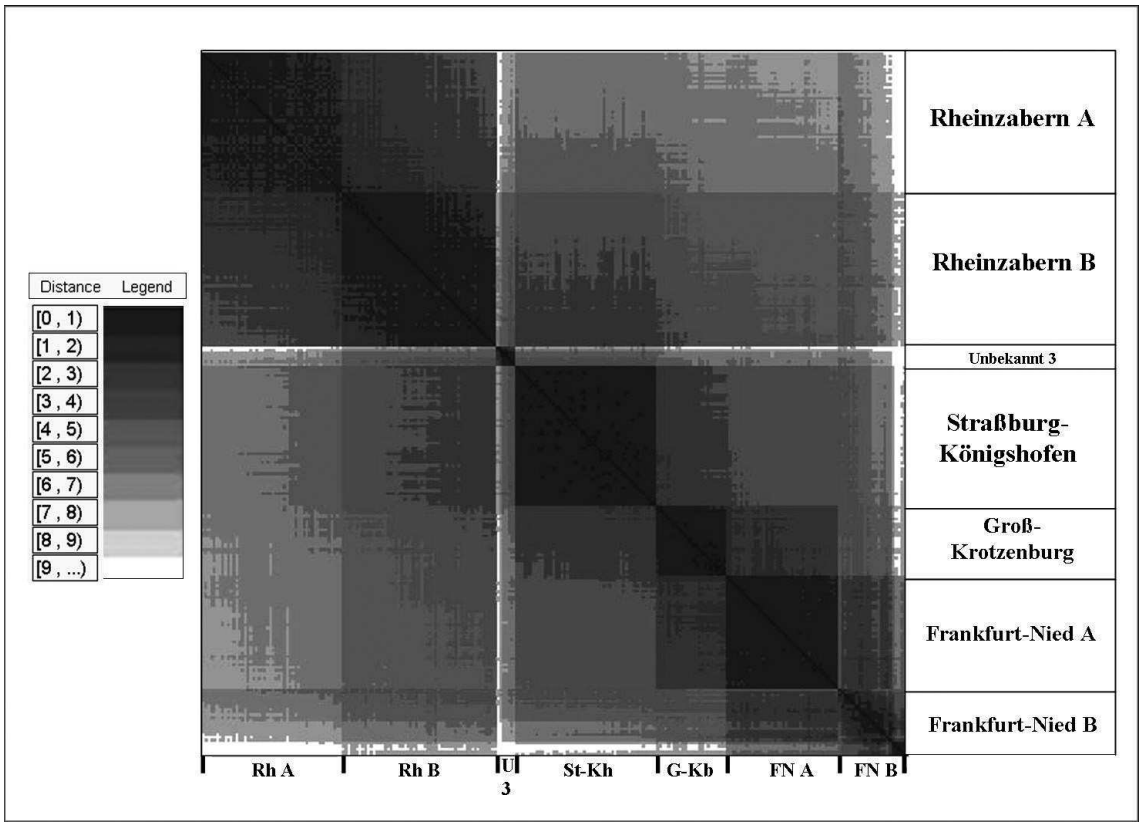


Figure 45: Darstellung der Distanzmatrix mit nach der 1. Hauptachse (Fig. 42) sortierten Zeilen und Spalten

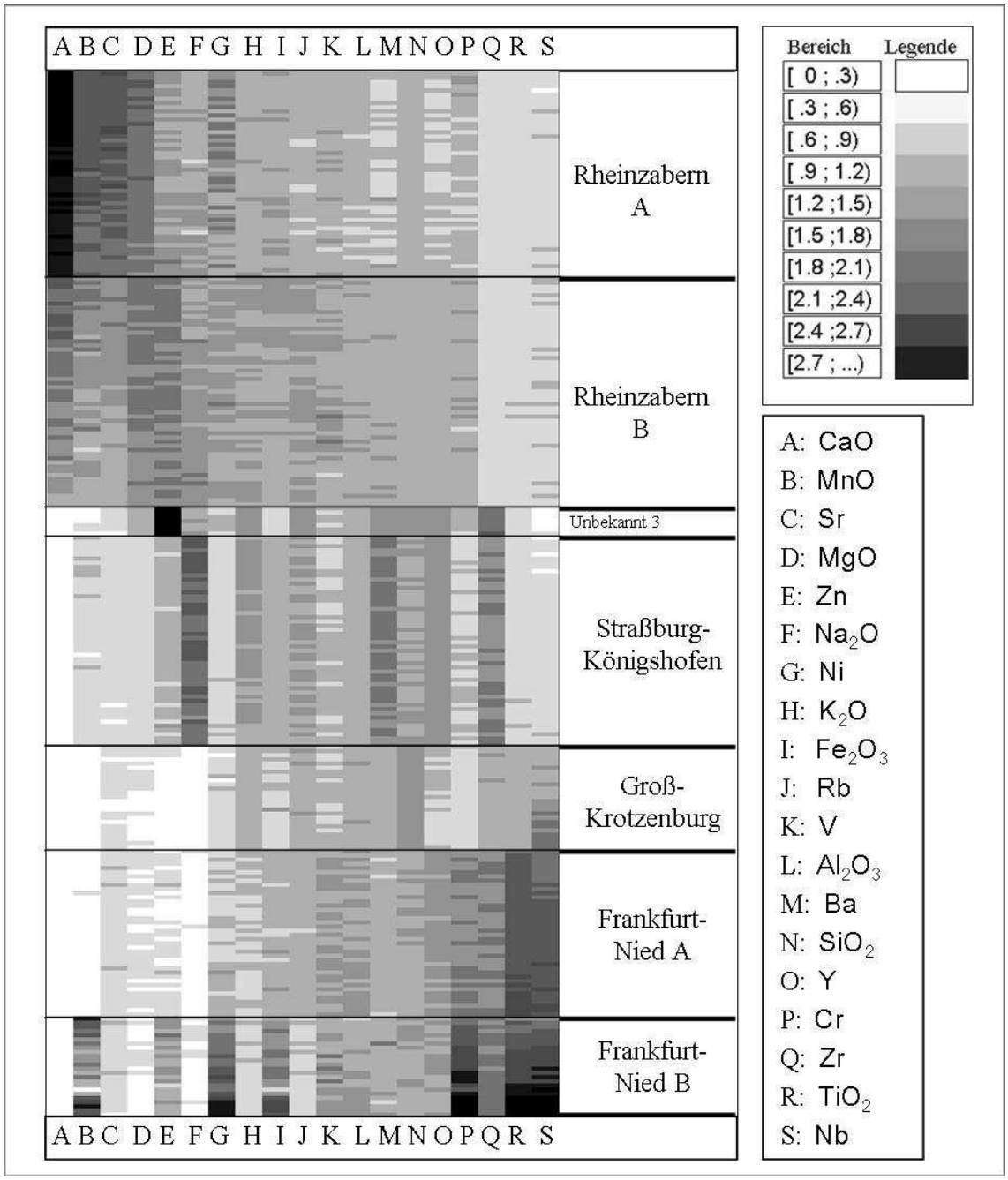


Figure 46: Sortieren der Oxid- und Spurenelementkonzentrationen (transformierte Werte) nach der 1. Hauptachse

In Fig. 45 werden die Klassen ebenso wie ihre interne Differenzierung gut erkenntlich. Offenbar ist 'Rheinzabern A' nicht so einheitlich wie 'Rheinzabern B'. Analoges kann beim Vergleich der Klassen 'Frankfurt-Nied A' und 'Frankfurt-Nied B' festgestellt werden, wobei hier die letztere die geringere Einheitlichkeit zeigt.

Zur Charakterisierung der Klassen hinsichtlich ihrer chemischen Zusammensetzung kann die Darstellung der Fig. 46 dienen. Dabei wurden die Muster $\mathbf{x}_j^{(v)}$ ($j = 1, \ldots, 19$) der transformierten Variablen für eine repräsentative Teilmenge von 251 Objekten nach der 1. Hauptachse sortiert. Die einzelnen Werte der Variablen werden in zehn Graustufen von Weiß bis Schwarz verdeutlicht, wobei ein Wert umso größer ist, je dunkler er dargestellt wird.

Aus Fig. 46 lassen sich somit bestimmte Besonderheiten der aufgeführten Klassen bzw. Provenienzen von Heeresziegeleien erkennen. So unterscheiden sich die beiden Rheinzabern-Klassen durch ihre extrem hohen (Gruppe A) und sehr hohen (Gruppe B) CaO-Gehalte von allen anderen Klassen. Die Proben von 'Unbekannt 3' besitzen die größten Zn-Gehalte, die für große Distanzen zu den anderen Klassen bzw. Objekten verantwortlich sind und auf diese Weise für die Stabilität dieser recht kleinen und heterogenen Klasse sorgen. 'Straßburg-Königshofen' ist charakterisiert durch hohe Gehalte an Na_2O, Ba sowie Zr und 'Frankfurt-Nied' (besonders die Gruppe B) durch große bis größte Gehalte an TiO_2, Nb und Cr, während die Groß-Krotzenburg zugeordneten Ziegel ein relativ ausgeglichenes Elementmuster aufweisen, etc.

References

[1] DOLATA, J. and WERR, U. (1998/1999): Wie gleich ist derselbe? – Homogenität eines römischen Ziegels und Aussagegrenzen geochemischer Analytik aufgrund von Meßtechnik und Materialvarietät. *Mainzer Archäologische Zeitschrift* 5/6, 129–147.

[2] GEBAUER, W. (1983): *Kunsthandwerkliche Keramik*. Fachbuchverlag, Leipzig, S. 10.

[3] KLAPROTH, M.H. (1798): Mémoire de numismatique docimastique. *Mémoires de l'Académie Royale des Sciences et Belles-lettres* [...], *MDCCXCII et MDCCXCIII, Classe de Philosophie Expérimental*, 97–113.

[4] KLAPROTH, M.H. (1799): Beitrag zur numismatischen Docimasie. *Sammlung der deutschen Abhandlungen, welche in der Königlichen Akademie der Wissenschaften zu Berlin vorgelesen worden in den Jahren 1792 – 1797, Abhandlungen der Königlichen Akademie der Wissenschaften und Schönen Künste, Experimental-Philosophie*, 3–14.

[5] MACQUEEN, J.B. (1967): Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics und Probability* 1, 281–297.

[6] MOMMSEN, H. (1986): *Archäometrie - Neuere naturwissenschaftliche Methoden und Erfolge in der Archäologie.* Teubner, Stuttgart.

[7] MUCHA, H.-J. (1992): *Clusteranalyse mit Microcomputern.* Akademie Verlag, Berlin, S. 28–31, 119–126, 141–142.

[8] PAPAGEORGIOU, I., BAXTER, M.A. and CAU, M.J. (2001): Model-based Cluster Analysis of Artefact Compositional Data. *Archaeometry* 43, 571–588.

[9] SPÄTH, H. (1980): *Cluster Analysis Algorithms for Data Reduction and Classification of Objects.* Ellis Horwood, Chichester.

[10] SPÄTH, H. (1985): *Cluster Dissection and Analysis, FORTRAN Programs, Examples.* Ellis Horwood, Chichester.

[11] VARMUZA, K. (1984): Chemometrie. In: Ziegler, E. (Hrsg.): *Computer in der Chemie – Praxisorientierte Einführung.* Springer-Verlag, Berlin, S. 148.

[12] WARD, J.H. (1963): Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58, 236–244.

[13] www.wiley.com/bw/journal.asp?ref=0003-813X

Anhang

Publikationen aus den Jahren 2000 – 2009
(in chronologischer Reihenfolge nach dem Jahr des Erscheinens)

zur Anwendung mathematischer Methoden bei der Untersuchung obergermanisch-römischer Ziegel aus dem Weierstraß-Institut für Angewandte Analysis und Stochastik, Berlin (Hans-Joachim Mucha), dem Institut für Chemie der Humboldt-Universität zu Berlin (Hans-Georg Bartel) und der Generaldirektion Kulturelles Erbe Rheinland-Pfalz, Mainz (Jens Dolata)

[A1] BARTEL, H.-G., DOLATA, J. and MUCHA, H.-J. (2000): Klassifikation gestempelter römischer Ziegel aus Obergermanien. *Archäometrie und Denkmalpflege*, 86–88.

[A2] BARTEL, H.-G., MUCHA, H.-J. and DOLATA, J. (2001): Parametrische und nichtparametrische Identifikationsmethoden, dargestellt am Beispiel römischer Baukeramik aus Obergermanien. *Archäometrie und Denkmalpflege*, 104–106.

[A3] DOLATA, J., MUCHA, H.-J. and BARTEL, H.-G. (2001): Untersuchungsperspektiven zum kombinierten archäologischen und materialanalytischen Nachweis römischer Ziegelherstellung in Mainz und Worms. *Archäometrie und Denkmalpflege*, 107–109.

[A4] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2002): Exploring Roman Brick and Tile by Cluster Analysis with Validation of Results. In: Gaul, W. and Ritter, G. (Eds.): *Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation*, Passau, 471–478.

[A5] BARTEL, H.-G., MUCHA, H.-J. and DOLATA, J. (2002): Automatische Klassifikation in der Archäometrie: Berliner und Mainzer Arbeiten zu oberrheinischen Ziegeleien in römischer Zeit. *Berliner Beiträge zur Archäometrie* 19, 31–62.

[A6] DOLATA, J., BARTEL, H.-G. and MUCHA, H.-J. (2003): Statistische Untersuchung zur Aufklärung der Binnenstruktur römischer Ziegelproduktion von Frankfurt-Nied. *Archäometrie und Denkmalpflege*, 40–42.

[A7] BARTEL, H.-G., MUCHA, H.-J. and DOLATA, J. (2003): Über eine Modifikation eines graphentheoretisch basierten partitionierenden Verfahrens der Clusteranalyse. *Match* 48, 209–223.

[A8] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2003): Core-Based Clustering Techniques: Methods, Software, and Applications. In: Schader, M., Gaul, W. and Vichi, M. (Eds.): *Proceedings of the 26th Annual Conference of the Gesellschaft für Klassifikation*, Mannheim, 74–82.

[A9] DOLATA, J., MUCHA, H.-J. and BARTEL, H.-G. (2003): Archäologische und mathematisch-statistische Neuordnung der Orte römischer Baukeramikherstellung im nördlichen Obergermanien. *Xantener Berichte* 13, 381–409.

[A10] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2003): Modellbasierte Clusteranalyse römischer Ziegel aus Worms und Rheinzabern. *Archäologische Informationen* 26/2, 471–480.

[A11] SWART, C., PAZ, B., DOLATA, J., SCHNEIDER, G., SIMON, J., BARTEL, H.-G. and MUCHA, H.-J. (2004): Analyse römischer Ziegel mit ICP-MS/ -OES: Methodenvergleich zwischen RFA und ICP. *Archäometrie und Denkmalpflege*, 25–27.

[A12] DOLATA, J., MUCHA, H.-J. and BARTEL, H.-G. (2004): Eine Anwendung der hierarchischen Clusteranalyse: „Unbekannt 1“ als Provenienz einer bekannten obergermanischen Heeresziegelei? *Archäometrie und Denkmalpflege*, 28–30.

[A13] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2004): Model-based Cluster Analysis of Roman Bricks and Tiles from Worms and Rheinzabern. In: Weihs, C. and Gaul, W. (Eds.): *Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation*, Dortmund, 317–324.

[A14] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2005): Techniques of Rearrangements in Binary Trees (Dendrograms) and Applications. *Match* 54, 561–582.

[A15] DOLATA, J., BARTEL, H.-G. and MUCHA, H.-J. (2006): Provenienz von Ziegeln aus dem römischen Theater in Mainz: Archäologische Bewertung von clusteranalytischen Resultaten. *Archäometrie und Denkmalpflege*, 150–152.

[A16] MUCHA, H.-J., BARTEL, H.-G., DOLATA, J., SWART, C. and GRAUBNER, K. (2006): Zur Ausweisung einer Ziegelreferenz „Mainz“ durch Bewertung der Stabilität von Klassen- und Objektzuweisungen. *Archäometrie und Denkmalpflege*, 153–155.

[A17] DOLATA, J., MUCHA, H.-J. and BARTEL, H.-G. (2007): Uncovering the Internal Structure of the Roman Brick and Tile Making in Frankfurt-Nied by Cluster Validation. In: Decker, R. and Lenz, H.-J. (Eds.): *Proceedings of the 30^{th} Annual Conference of the Gesellschaft für Klassifikation*, Berlin, 663–670.

[A18] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2007): Zur Clusteranalyse und Hauptkomponentenanalyse archäometrischer Daten auf Grundlage von Rankinformationen. *Archäometrie und Denkmalpflege*, 14–16.

[A19] DOLATA, J., BARTEL, H.-G. and MUCHA, H.-J. (2007): Archäologisch-historische Auswertung älterer und neuer Materialanalysen oberrheinischer Ziegel – Zusammenschau der Messungen verschiedener Arbeitsgruppen anlässlich der Ziegelstempelvorlage von Oedenburg bei Biesheim im Oberelsaß. *Archäometrie und Denkmalpflege*, 86–88.

[A20] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2008): Effects of Data Transformation on Cluster Analysis of Archaeological Data. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L. and Decker, R. (Eds.): *Proceedings of the 31^{st} Annual Conference of the German Classification Society*, Freiburg i. Br., 681–688.

[A21] BARTEL, H.-G., MUCHA, H.-J. and DOLATA, J. (2008): Über Identifikationsmethoden, dargestellt am Beispiel römischer Baukeramik aus Obergermanien. *Berliner Beiträge zur Archäometrie* 21, 115–132.

[A22] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2008): Finding Roman Brickyards in Germania Superior by Model-based Cluster Analysis of Archaeometric Data. In: Posluschny, A., Lambers, K. and Herzog, I. (Eds.): *Proceedings of the 35^{th} International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, Berlin, 360–365.

[A23] BARTEL, H.-G., DOLATA, J. and MUCHA, H.-J. (2009): Klassifikation von 613 Proben als Referenzen für die Herstellungsprovenienzen römischer Baukeramik im nördlichen Obergermanien. *Mainzer Archäologische Zeitschrift* 8, 51–76.

[A24] DOLATA, J., BARTEL, H.-G. and MUCHA, H.-J. (2009): Geochemische und statistische Erkundung der Herstellungsorte von Ziegeln der Legio XXI Rapax. In: Reddé, M. (Ed.): *Oedenburg – Fouilles françaises, allemandes et suisses*

à Biesheim et Kunheim, Haut-Rhin, France. Vol. 1: Les camps militaires Julio-Claudiens. (Monographien des Römisch-Germanisches Zentralmuseums, Band 79,1). Verlag Römisch-Germanisches Zentralmuseum, Mainz, 355–364.

[A25] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2009): Zur Klassifikation römischer Ziegel von Fundorten im südlichen Obergermanien. *Archäometrie und Denkmalpflege* (= *Metalla (Bochum)*, Sonderheft 2), 144–146.

[A26] DOLATA, J., MUCHA, H.-J. and BARTEL, H.-G. (2009): Mapping Findspots of Roman Military Brickstamps in Mogontiacum (Mainz) and Archaeometrical Analysis. In: Fink, A., Lausen, B., Seidel, W. and Ultsch, A. (Eds.): *Proceedings of the 32nd Annual Conference of the Gesellschaft für Klassifikation*, Hamburg, im Druck.

4 Fuzzy Spectral Clustering by PCCA+

Susanna Röblitz and Marcus Weber
Zuse Institute Berlin (ZIB)
Takustraße 7, 14195 Berlin
susanna.roeblitz@zib.de, weber@zib.de

Abstract

Given pairwise similarities between data points, spectral clustering makes use of the eigenvectors of the corresponding similarity matrix to perform dimensionality reduction for clustering in fewer dimensions. One example from this class of algorithms is the Robust Perron Cluster Analysis (PCCA+). In contrast to other algorithms, PCCA+ does not assign each object to exactly one cluster but it results in a fuzzy clustering where every object belongs to all clusters with certain membership values. The present article explains the main ideas of PCCA+ and presents the results of clustering the data sets provided by the organizers of the 30th Fall Meeting 2008 of the AG-DANK.

Keywords: spectral clustering, graph Laplacian, Perron cluster, PCCA+

4.1 Introduction

Clustering deals with the problem of separating data objects in different clusters according to their similarities. A partition of n objects $o_1, \ldots, o_n$ into k clusters $C_1, \ldots, C_k$ can be represented by an indicator matrix $\chi \in I\!R^{n \times k}$ with

$$\chi(i,j) = \begin{cases} 1, & \text{if } o_i \in C_j \\ 0, & \text{else} \end{cases} .$$

The idea of fuzzy clustering is to perform a relaxation by discarding the condition on the discrete values for χ and instead allow χ to take values in the interval $[0,1]$ such that

$$0 \leq \chi(i,j) \leq 1, \quad \sum_{j=1}^{k} \chi(i,j) = 1 \quad \forall\, i = 1, \ldots, n.$$

The entry $\chi(i,j)$ can be interpreted as membership value of object i with respect to cluster j. The matrix χ is therefore denoted as *membership matrix*.

The method we apply to obtain the membership matrix χ is called *Robust Perron Cluster Analysis* (PCCA+) [1, 7]. This method belongs to the class of *spectral clustering* algorithms. In recent years, a number of articles shed light on spectral clustering. We especially recommend [4], which also contains an extensive list of references to this topic.

The article is organized as follows. First, we summarize the basics of spectral clustering in Sect. 4.2, before we give a short introduction to PCCA+ in Sect. 4.3. In Sect. 4.4, we briefly discuss the choice of an appropriate similarity measure.

4.2 Spectral Clustering

The starting point for spectral clustering are objects $o_1, \ldots, o_n$ with pairwise similarities $s_{ij} > 0$, $i, j = 1, \ldots, n$. The data can be represented in form of an undirected *similarity graph* $G = (V, E)$, where the vertices $V = \{v_1, \ldots, v_n\}$ represent the objects o_i. Each edge between two vertices v_i and v_j carries a weight $w_{ij} \geq 0$, which enters the *adjacency matrix* $W = (w_{ij})_{i,j=1,\ldots,n}$. The *degree matrix* D is defined as the diagonal matrix with entries

$$d_i = \sum_{j=1}^{n} w_{ij}$$

on the diagonal.

There are several possibilities to construct the similarity graph, for example the ε-neighborhood graph or k-nearest neighbor graphs, which are sparse representations of the data. As long as the number n of objects is of moderate size (about ≤ 2000), we prefer the *fully connected graph* with $w_{ij} = s_{ij}$.

Given a set $o_1, \ldots, o_n$ of objects, it is mostly more intuitive to compute *pairwise distances* d_{ij} instead of similarities. A popular way to transform these distances into similarities is the *Gaussian similarity function*

$$s_{ij} = \exp(-\beta\, d_{ij}^2), \quad \beta = \frac{1}{2\sigma^2}. \tag{17}$$

The parameter σ controls the width of the neighborhoods.

Clustering is now equivalent to finding a partition of the graph such that edges between different clusters have a low weight and edges within a cluster have high weight. The most common spectral clustering algorithms have the following form:

1. Construct a similarity graph with weighted adjacency matrix W.
2. Compute a graph Laplacian L.
3. Compute the first k eigenvectors $X = [x_1, \ldots, x_k]$ of L.
4. For $i = 1, \ldots, n$ let $y_i \in I\!R^k$ be the ith row of X. Cluster the points $(y_i)_{i=1,\ldots,n}$ into clusters $C_1, \ldots, C_k$.

Spectral clustering requires only the computation of a few eigenvectors, which is quite easy with standard numerical software like MATLAB.

The different algorithms differ in the computation of the graph Laplacian and the clustering of the rows of X. For properties of graph Laplacians, the reader is referred to [4]. For example, in normalized spectral clustering according to [5], the Laplacian is computed by

$$L = I - D^{-1}W.$$

There are several reasons why this Laplacian should be favored over other constructions, see [4].

The matrix $P = D^{-1}W$ is a row-stochastic matrix and can be interpreted as transition matrix of a random walk which jumps from vertex to vertex. The transition probability of jumping in one step from vertex i to vertex j is given by $p_{ij} = w_{ij}/d_i$. If the graph is connected and non-bipartite, then the random walk possesses a unique stationary distribution $\pi = [\pi_1, \ldots, \pi_n]^\top$ given by $\pi_i = d_i / \sum_j d_j$. Spectral clustering corresponds to finding a partition of the graph such that the random walk stays long within the same cluster and seldom jumps between clusters. Since P and $L = I - P$ have the same eigenvectors, spectral clustering on L is equivalent to spectral clustering on P.

The clustering of points $(y_i)_{i=1,\ldots,n}$ in step 4 is usually done by the k-means algorithm, but any other method could be used instead. One possible choice is PCCA+, which will be explained in the following Section.

4.3 Robust Perron Cluster Analysis (PCCA+)

The transition probability matrix $P = D^{-1}W$ represents a Markov chain on the state space $S = \{o_1, \ldots, o_n\}$. In case of a decomposable Markov chain or, equivalently, a disconnected similarity graph, an appropriate permutation of objects according to their connectedness results in a block-diagonal matrix P with k blocks. This matrix has a k-fold eigenvalue $\lambda = 1$. The corresponding eigenvectors $X = [x_1, \ldots, x_k]$ are piecewise constant on the blocks and can thus be used to identify the clusters. In fact, the rows of X can be considered as vertices of a $(k-1)$-dimensional simplex. Every object can be assigned to one of the k vertices and thus to one of the k clusters.

Generally, the matrix P constructed from practical data is not decomposable. However, if there are k hidden clusters, P has a cluster of eigenvalues $1 = \lambda_1 > \lambda_2 > \ldots > \lambda_k > 1 - \varepsilon$ near the Perron eigenvalue $\lambda_1 = 1$. The rows y_i of the corresponding eigenvectors still nearly form a simplex. the first eigenvector is always constant, the rows can be considered

The goal of PCCA+ is to identify the vertices of a simplex σ_{k-1} such that all points y_i are located within the simplex. Then every point y_i can be assigned to one of the k vertices and thus to one of the k clusters by a certain membership vector $\chi(i,:) = [\chi(i,1), \ldots, \chi(i,k)]$.

The identification of such a simplex is equivalent to finding a non-singular transfor-

mation matrix $\mathcal{A}$ such that

$$\chi = X\mathcal{A}$$

and

(1a) $\chi(i,j) \geq 0 \quad \forall i \in \{1,\ldots,n\},\, j \in \{1,\ldots,k\}$ (positivity),

(1b) $\sum_{j=1}^{k} \chi(i,j) = 1 \quad \forall i \in \{1,\ldots,n\}$ (partition of unity).

Among the feasible transformation matrices we search for a matrix $\mathcal{A}$ such that the resulting membership vectors $\chi(i,:)$ are as crisp as possible, i.e. the columns $\chi(:,j)_{j=1,\ldots,n}$ should be as close to indicator vectors as possible. This can be achieved by maximizing the objective function [10]

$$I(\mathcal{A}; X, \pi) = \sum_{i=1}^{k} \frac{\langle \chi_i, \chi_i \rangle_\pi}{\langle \chi_i, e \rangle_\pi} \leq k, \tag{18}$$

where e denotes the vector with all entries equal to 1. To summarize, PCCA+ aims at

$$\text{maximizing} \quad I(\mathcal{A}; X, \pi)$$

subject to

(1) $\chi(i,j) \in \sigma_{k-1} \quad \forall i \in \{1,\ldots,n\}$ (simplex),

(2) $\chi = X\mathcal{A}$, $\mathcal{A}$ non-singular (invariance).

Constraint (1) is just a compact formulation of conditions (1a) and (1b).

One has to maximize a convex function with linear constraints, which is not a trivial task. However, the optimization problem can be solved by the Nelder-Mead algorithm provided that a good initial guess for $\mathcal{A}$ is available. This starting guess is obtained by the *inner simplex algorithm* as described in [8].

Finally, in order to obtain a partition of data points into clusters, the real-valued solution matrix χ needs to be re-transformed into a discrete indicator matrix by

$$o_i \in C_k \quad \text{if} \quad \chi(i,k) = \max_j \chi(i,j).$$

Number of clusters

Since the number of clusters k is unknown in advance, it is recommended to run the cluster algorithm several times with different input values for k and to choose the "best" solution. In order to evaluate the quality of the solution, several criteria can be used:

- The spectral gap. If there are k well-separated clusters, there will be a significant gap between the eigenvalues λ_k and λ_{k+1}.

- The condition of the invariant subspace X [6]. In case of k well-separated clusters, the corresponding invariant subspace X spanned by the first k eigenvectors of P or L is well-conditioned. In case of a reversible Markov chain or a symmetric similarity matrix S, respectively, this criterion is equivalent to the spectral-gap criterion.

- The *minChi*-criterion [9]. In general, the initial guess for $\mathcal{A}$ is infeasible, i.e. it leads to a membership matrix χ with negative entries. However, if there exist well separated clusters, the value

$$\text{minChi} = \min_i \min_j \chi(i,j)$$

  will be close to zero. Thus, one can decide for the number k that maximizes the minChi-value.

- Optimality of the solution. Since $I(\mathcal{A}; X, \pi) \leq k$, one could choose the number k for which $I(\mathcal{A}; X, \pi)/k$ is maximal.

In general, if there are k well-separated clusters, all proposed criteria will favor this number.

4.4 Similarity Graph

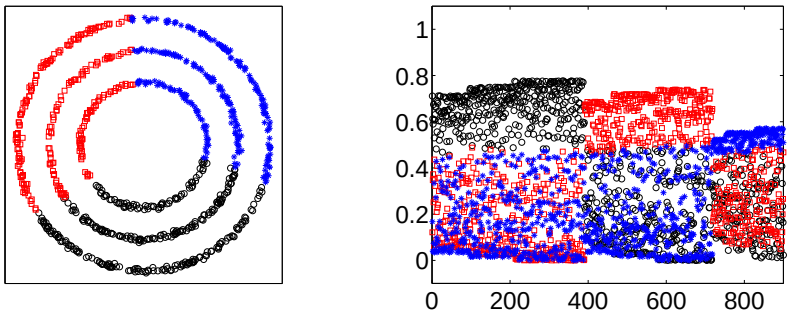
We compute pairwise similarities s_{ij} by the Gaussian similarity function (17). However, the results of spectral clustering are quite sensitive to the choice of the distance function $d(o_i, o_j)$. Although the Euclidean distance

$$d_{ij} = \|o_i - o_j\|_2$$

is a natural choice, it is not always appropriate, as illustrated in Fig. 47. Clustering based on the Euclidean distance assumes that the data points are grouped to "compact" clusters where the objects within one cluster are either mutually similar to each other or they are similar with respect to a common representative or centroid. cluster is larger different clusters, spectral clustering distance will fail. Whenever data subsets occupy elongated regions like spiral arms or circles, an alternative distance function based on connectedness of data points is required. Such a distance function has been introduced in [2, 3] in conjunction with path based clustering. Path based clustering assigns two objects to the same cluster if they are connected by a path with high similarity between adjacent objects on the path. The *effective distance* between two objects is calculated as the minimum over all path distances,

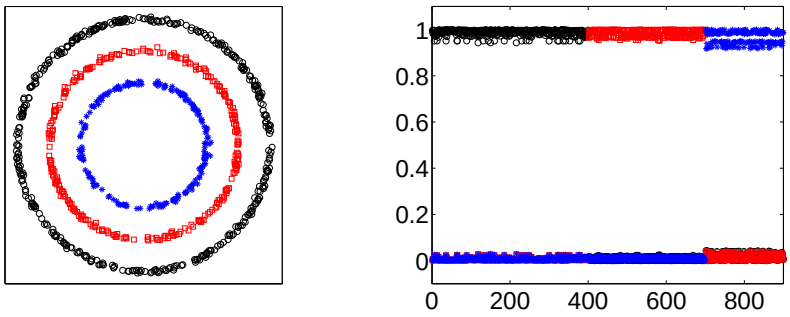
$$d_{ij}^{\text{eff}} = \min_{p \in \mathcal{P}_{ij}(E)} \left\{ \max_{1 \leq k < |p|} d_{p[k]p[k+1]} \right\}.$$

Here, $\mathcal{P}_{ij}(E)$ denotes the set of all paths from object i to object j and $d_{p[k]p[k+1]}$ denotes the Euclidean distance between the kth and $(k+1)$th object on the path. The



(a) Clusters (b) Membership values

Figure 47: Example for a point set where the Euclidean distance is inappropriate to identify clusters. The separation between the clusters is quite weak, which is illustrated by the small membership values of the data points.



(a) Clusters (b) Membership values

Figure 48: Example for a point set where the use of effective distances results in the desired clustering. The separation between clusters is strong, illustrated by the fact that the membership vectors are nearly indicator vectors.

matrix D^{eff} can be computed recursively by a variant of Kruskal's minimum spanning tree algorithm [2]. Again, the corresponding similarity matrix S is computed by the Gaussian function (17). Fig. 48 illustrates the result of fuzzy spectral clustering based on the effective distance matrix.

4.5 Examples

The clustering results for the data provided for the 30th Fall Meeting of the working group AG–DANK at `http://www.fim.uni-passau.de/de/fim/fakultaet/lehrstuehle/ritter/ag-dank.html` are presented separately in Sect. 7.3 and Sect. 8.3.

References

[1] DEUFLHARD, P., and WEBER, M. (2005) Robust Perron cluster analysis in conformation dynamics. Linear Algebra Appl., 398, 161–184.

[2] FISCHER, B., and BUHMANN, J. M.(2002) Data Resampling for Path Based Clustering. Proceedings of the 24th DAGM Symposium on Pattern Regognition, Lecture Notes in Computer Science, 2449, Springer-Verlag, London, 206–214.

[3] FISCHER, B., ZÖLLER, T., and BUHMANN, J. M.(2001) Path Based Pairwise Data Clustering With Application to Texture Segmentation. Energy Minimization Methods in Computer Vision and Pattern Recognition, Lecture Notes in Computer Science, 2134, Springer-Verlag, Berlin, 235–250.

[4] VON LUXBURG, U. (2006) A Tutorial on Spectral Clustering. Max Planck Institute for Biological Cybernetics, Technical Report No. 149, Tübingen.

[5] SHI, J., and MALIK, J. (2000) Normalized Cuts and Image Segmentation. IEE Transactions on pattern Analysis and machine Intelligence, 22 (8), 888–905.

[6] STEWART, G. W., and JI-GUANG SUN (1990) Matrix Perturbation Theory. Computer Science and Scientific Computing, Academic Press, Boston.

[7] WEBER, M. (2006) Meshless Methods in Conformation Dynamics. Doctoral Thesis, Department of Mathematics and Computer Science, Freie Universität Berlin, Verlag Dr. Hut, München.

[8] WEBER, M., and GALLIAT, T. (2002) Characterization of Transition States in Conformational Dynamics Using Fuzzy Sets. Zuse Institute Berlin, ZIB-Report No. 02-12, Berlin.

[9] WEBER, M., RUNGSARITYOTIN, W., and SCHLIEP, A.(2006) An Indicator for the Number of Clusters Using a Linear Map to Simplex Structure. In: M. Spiliopoulou und R. Kruse und C. Borgelt und A. Nürnberger und W. Gaul (Eds.): *From Data and Information Analysis to Knowledge Engineering*, Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V., Universität Magdeburg, März 2005, Series Studies in Classification, Data Analysis, and Knowledge Organization , Springer-Verlag, Berlin, 103–110.

[10] RÖBLITZ, S. (2008) Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics. Doctoral Thesis, Freie Universität Berlin.

5 Merging Gaussian mixture components - an overview

Christian Hennig
Department of Statistical Science, UCL,
Gower St.,
London, WC1E 6BT, United Kingdom
chrish@stats.ucl.ac.uk

Abstract

The problem of merging Gaussian mixture components is discussed in a situation where a Gaussian mixture is fitted but the mixture components are not separated enough from each other to interpret them as "clusters". The problem of merging Gaussian mixtures is not statistically identifiable, therefore merging algorithms have to be based on subjective cluster concepts. Several different methods are proposed.

Keywords: model-based cluster analysis, multilayer mixture, unimodality, prediction strength, ridgeline, dip test

Introduction

The Gaussian mixture model is often used for cluster analysis (for an overview and references see Fraley and Raftery, 2002, and McLachlan and Peel, 2000). This approach is based on the assumption that $I\!R^p$-valued observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are distributed i.i.d. according to the density

$$f(\mathbf{x}) = \sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j, \Sigma_j}(\mathbf{x}), \tag{19}$$

where $\pi_j > 0 \forall j$, $\sum_{j+1}^{s} \pi_j = 1$, $\varphi_{\mathbf{a},\Sigma}$ is the density of the p-dimensional Gaussian distribution with mean vector $\mathbf{a}$ and covariance matrix Σ. Given a fixed s, the parameters can be estimated by Maximum Likelihood using the EM-algorithm. The data points can then be classified to the mixture components by maximizing the estimated a posteriori probability that $\mathbf{x}_i$ was generated by mixture component j,

$$\hat{P}(\gamma_i = j | \mathbf{x}_i = \mathbf{x}) = \frac{\hat{\pi}_j \varphi_{\hat{\mathbf{a}}_j, \hat{\Sigma}_j}(\mathbf{x}))}{\sum_{i=1}^{s} \hat{\pi}_i \varphi_{\hat{\mathbf{a}}_i, \hat{\Sigma}_i}(\mathbf{x})}, \tag{20}$$

where γ_i is defined by the two-step version of the mixture model where

$$P(\gamma_i = j) = \pi_j, \ \mathbf{x}_i | (\gamma_i = j) \sim \varphi_{\mathbf{a}_j, \Sigma_j}, \ i = 1, \ldots, n \text{ i.i.d.}. \tag{21}$$

A standard method (though not the only one) to estimate the number of components s is the Bayesian Information Criterion (BIC, Schwarz, 1978), see Fraley and Raftery (2002) for details. The outcome is called EM/BIC.

In cluster analysis usually every mixture component is interpreted as a cluster, and pointwise maximization of (20) defines the clustering. The idea is that a mixture formalizes that the underlying distribution is heterogeneous with several different populations, all of which are modelled by homogeneous Gaussian distributions. Keeping in mind that there is no unique definition of a "true cluster", and not necessarily assuming that the Gaussian mixture model assumption holds precisely, it could be said that this method employs the Gaussian distribution as the prototype shape of clusters to look for.

From a practical point of view, perhaps the most important problem with this approach is that for most applications Gaussian distributions are too restricted to formalize the cluster shapes one is interested in. For example, mixtures of two (or more) Gaussian distributions can be unimodal, and in such distributions there is no gap (and in this sense no separation) between the different Gaussian subpopulations. In many applications in which the number of clusters is not known, the EM algorithm together with the BIC yield a larger optimal number of mixture components than what seems to be a reasonable number of clusters when looking at the data.

The hierarchical principle for merging Gaussian components, which is used in the present paper, works as follows:

1. Start with all components of the initially estimated Gaussian mixture as current clusters.
2. Find the pair of current clusters most promising to merge.
3. Apply a stopping criterion to decide whether to merge them to form a new current cluster, or to use the current clustering as the final one.
4. If merged, go to 2.

Two criteria are needed, namely in step 2 a rule to find the pair of clusters best to merge and the stopping rule in step 3.

5.1 The Nature of the Problem

The merging problem looks as follows. Given a Gaussian mixture with s components as below, find $k \leq s$ and mixtures $f_1^*, \ldots, f_k^*$ of components of the original mixture so that each original Gaussian component appears in exactly one out of $f_1^*, \ldots, f_k^*$, and

$$f(\mathbf{x}) = \sum_{i=1}^{s} \pi_i \varphi_{\mathbf{a}_i, \Sigma_i}(\mathbf{x}) = \sum_{j=1}^{k} \pi_j^* f_j^*(\mathbf{x}), \tag{22}$$

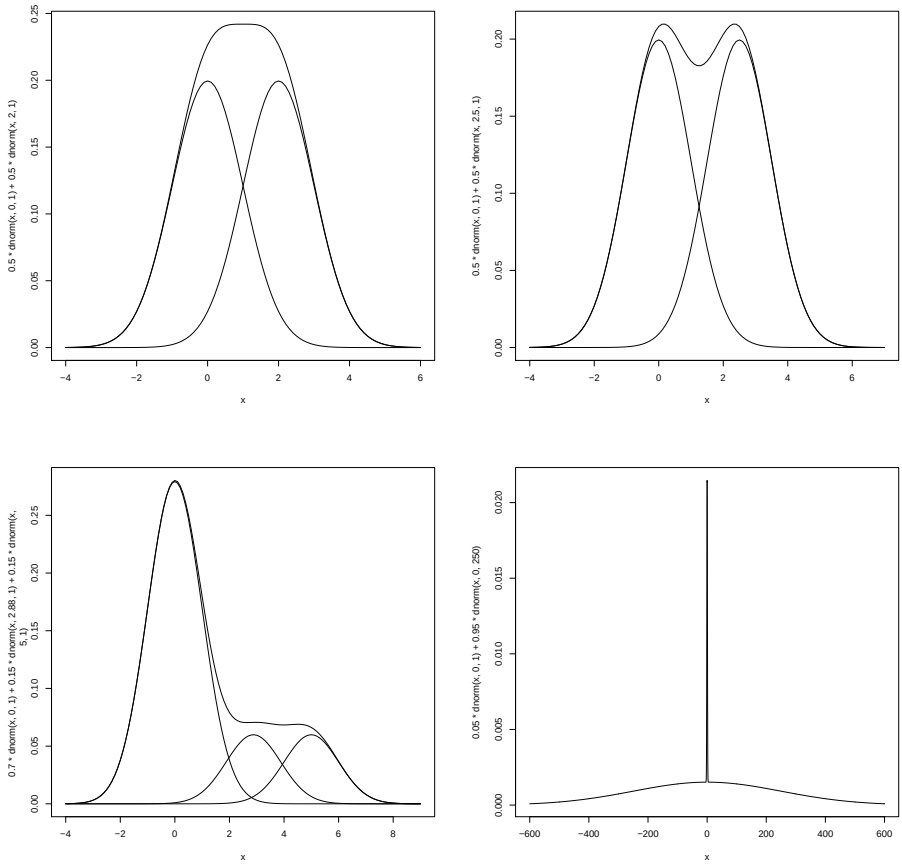


Figure 49: Four one-dimensional Gaussian mixtures.

where π_j^* is the sum of the π_i of the Gaussian components assigned to f_j^*. For datasets, clustering can be done by maximizing estimated posterior probabilities

$$\hat{P}(\gamma_i^* = j|\mathbf{x}_i = \mathbf{x}) = \frac{\hat{\pi}_j^* \hat{f}_j^*(\mathbf{x})}{\sum_{i=1}^{s} \hat{\pi}_i^* \hat{f}_i^*(\mathbf{x})}, \tag{23}$$

by analogy to (20) with $\gamma_1^*, \ldots, \gamma_n^*$ defined by analogy to (21).

From (22), however, $f_1^*, \ldots, f_k^*$ are not identifiable. Imagine $s = 3$ and $k \leq 3$. In terms of the density and therefore the likelihood, it does not make a difference whether f_1^* is a mixture of the first two Gaussian components and f_2^* equals the third one, or f_1^* mixes the first and the third Gaussian component and f_2^* equals the second one, or any other admissible combination.

It is well known that there is no objective unique definition of what a "true cluster" is, so there is necessarily a subjective component in the decision how to merge components.

The situations in Figure 49 illustrate that essentially different cluster concepts may be of interest. Particularly the role of unimodality may be controversial. Some

researchers may find it intuitive to identify a cluster with a set of points surrounding a density mode, and in most situations the unimodal mixture at the top left of Figure 49 may be regarded as generating a single cluster (except if there are strong reasons to believe that "true clusters" should at least be approximately Gaussian, be mixtures unimodal or not, i.e., not demanding any "separation" between clusters). However, in some applications the unimodal mixtures at the bottom of Figure 49 may not be regarded as a single cluster, because the modes in these examples are surrounded by dense "patterns" of the data that seem to be separated from what goes on in the tails, which is caused by other Gaussian components. But it is not clear that these mixtures in any case should not be merged into a single cluster, because there is no separating "gap" between them, which may be required to speak of "clusters".

On the other hand, multimodal mixtures may also be accepted as single clusters if the modes are not properly separated as in the upper right plot in Figure 49. Note also that ML-estimation of Gaussian mixtures applied to data generated from uniform distributions tends to come up with multimodal Gaussian mixtures.

Therefore, in order to define a suitable method for merging normals, the statistician has to decide

- whether only gaps in the density are accepted as separation between different clusters ("modality based cluster concept") or whether a dense data subset around a mode should be separated from clearly less dense data subsets even if the latter cannot be assigned to another mode ("pattern based cluster concept"),
- how strong the separation between different clusters should at least be (regardless of which of the two concepts is chosen, though the meaning of "separation" differs between them to some extent),
- what the role of the number of points in a cluster is, i.e., how strongly "cluster-shaped" small data subsets should be in order to constitute an own cluster.

The present paper offers a range of methods to deal with various decisions in these respects (for the last one, one may consider the inclusion of a "uniform noise component" in the mixture, as mentioned in the Introduction).

5.2 Methods Based on Modality

The ridgeline unimodal method

Under a strong version of the modality based cluster concept, the "strong modality merging problem" is to find a partition of the mixture components so that all resulting clusters are unimodal but any further merging would result in a cluster that is no longer unimodal. This requires an analysis of the modality of Gaussian

mixtures. The most advanced paper on this topic, to my knowledge, is Ray and Lindsay (2005). They showed that for any mixture f of s Gaussian distributions on $\mathbb{R}^p$ there is an $s-1$-dimensional manifold of $\mathbb{R}^p$ so that all extremal points of f lie on this manifold.

For $s = 2$, this manifold is defined by the so-called "ridgeline",

$$\mathbf{x}^*(\alpha) = [(1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}]^{-1}[(1-\alpha)\Sigma_1^{-1}\mathbf{a}_1 + \alpha\Sigma_2^{-1}\mathbf{a}_2], \tag{24}$$

and all density extrema (and therefore all modes which may be more than 2 in some situations) can be found for $\alpha \in [0, 1]$.

Unfortunately,for $s > 2$, Ray and Lindsay's result does not yield a straightforward method to find all nor even the number of modes. Therefore, their results can in general only be used to solve the strong modality merging problem approximately. The **unimodal ridgeline method** is defined by the hierarchical principle as follows.

1. Start with all components of the initially estimated Gaussian mixture as current clusters.
2. Using the mean vectors and covariance matrices of the current clusters (initially the Gaussian components), use the 2-component Gaussian mixture derived from these parameters on the ridgeline (24) to check whether it is unimodal for any pair of two current clusters.
3. If none of these is unimodal, use the current clustering as the final one.
4. Otherwise,
   (a) merge all of the pairs leading to unimodal mixtures.
   (b) go to step 2.

In order to apply this principle to data, means and covariance matrices are replaced by their ML-estimators for Gaussian mixture components. For mixtures of two or more Gaussians appearing as current clusters in the hierarchy, mean vectors and covariance matrices can be computed using the weights of points in the current cluster computed by summing up the weights (20) for all involved mixture components.

The ridgeline ratio method

Even if the cluster concept is modality based, in some situations the statistician may want to allow clusters that deviate from unimodality as long as the gap between the modes is not strong enough to interpret them as belonging to two separated clusters. One reason for this is that clusters may be required to be strongly separated. Another reason is that, as a result of too small sample size or particular instances of non-normality, data from unimodal underlying distributions may be approximated by EM/BIC by a multimodal Gaussian mixture.

A straightforward method to deal with this is to replace the demand of unimodality in the previous session by a cutoff value r^* for the ratio r between the minimum of the mixture density f and the second largest mode in case that there is more than one.

The dip test method

Tantrum, Murua and Stuetzle (2003) defined a hierarchical merging algorithm for the modality based cluster concept, the stopping rule of which is a sufficiently small p-value of Hartigan and Hartigan's (1985) dip test for unimodality. To use a significance test for unimodality is intuitively appealing if the statistician wants to merge components if the resulting mixture cannot be statistically distinguished from a unimodal distribution. As a modification of Tantrum, Murua and Stuetzle's method, I suggest to replace the log-likelihood difference by the ridgeline ratio r defined in Section 5.2. Here is the proposed **dip test method** in more detail:

1. Choose a tuning constant $p^* < 1$.
2. Start with all components of the initially estimated Gaussian mixture as current clusters.
3. Using the mean vectors and covariance matrices of the current clusters, use the 2-component Gaussian mixture derived from these parameters on the ridgeline (24) to compute r for any pair of current clusters.
4. Consider the data subset $\mathbf{x}^*$ of points classified to the union of the pair of current clusters maximizing r by maximizing (23).
5. Let $\mathbf{x}^{*1}$ be the projection of $\mathbf{x}^*$ onto the discriminant coordinate based on the pooled covariance matrix of the two involved current clusters, separating the two current cluster means (this is necessary because the dip test operates on one-dimensional data).
6. If the p-value of the dip test applied to $\mathbf{x}^{*1}$ is $\leq p^*$, use the current clustering as the final one.
7. Otherwise merge this pair of current clusters and go to step 3.

5.3 Methods Based on Misclassification Probabilities

The methods introduced in this Section formalize versions of the pattern based cluster concept as opposed to the modality based one. Misclassification probabilities provide an intuitive possibility to formalize separation between different clusters in a different way than density gaps. For example the two components of the scale mixture on the lower right side of Figure 49 are not separated in the sense that there are no gaps between them, but nevertheless the misclassification probability

between them is low. Obviously, the misclassification probability would be low as well in case of a strong density gap between components, so that in many clear cut situations both concepts arrive at the same clustering.

The Bhattacharyya distance method

The Bhattacharyya distance is a general distance between two distributions related to the overall Bayes misclassification probability for the 2-class problem with arbitrary class probabilities. This is bounded from above by $\exp(-d)$, where d is the Bhattacharyya distance (**reference**). For two Gaussian distributions with mean vectors and covariance matrices $\mathbf{a}_j, \Sigma_j,\ j = 1, 2,$, the Bhattacharyya distance is (Fukunaga, 1990)

$$d = \frac{(\mathbf{a_1} - \mathbf{a}_2)^t \bar{\Sigma}^{-1} (\mathbf{a_1} - \mathbf{a}_2)}{8} + \frac{1}{2} \log \left(\frac{|\bar{\Sigma}|}{\sqrt{|\Sigma_1||\Sigma_2|}} \right), \tag{25}$$

where $\bar{\Sigma} = \frac{1}{2}(\Sigma_1 + \Sigma_2)$. For data, the parameters can of course be replaced by their estimators.

The Bhattacharyya distance between two mixtures of Gaussians cannot be computed in a straightforward way. Therefore, for hierarchical merging, I again suggest to represent mixtures of Gaussians by their overall mean vector and covariance matrix and in this sense to treat them as single Gaussians. The **Bhattacharyya distance method** looks as follows:

1. Choose a tuning constant $d^* < 1$.
2. Start with all components of the initially estimated Gaussian mixture as current clusters.
3. Compute the (in case of mixtures with more than one component approximately) estimated Bhattacharyya distances d between all pairs of current clusters from their mean vectors and covariance matrices.
4. If $d < d^*$ for all pairs of current clusters, use the current clustering as the final one.
5. Otherwise, merge the pair of current clusters with maximum d and go to step 3.

Here d^* formalizes the degree of separation between clusters. By analogy to r^* above, d^* could be chosen by subject matter considerations or by simulations from borderline distributions.

Directly estimated misclassification probabilities

Instead of estimating the Bhattacharyya distance, misclassification probabilities $p_{ij} = P(\tilde{\gamma}_1^* = i|\gamma_1^* = j) = \frac{P(\tilde{\gamma}_1^*=i,\ \gamma_1^*=j)}{\pi_j^*}$ between components of a mixture distribution can also be estimated directly from the results of the EM algorithm. Here γ_1^* denotes the mixture component number that generated the first data point (or any other point according to the i.i.d. assumption, as long as only probabilities are of interest), and $\tilde{\gamma}_1^*$ is the mixture component to which the point is classified by maximizing the population version of (23), i.e., by the Bayes rule with true parameters. π_j^* can be estimated by $\hat{\pi}_j^*$. Note that

$$\hat{P}(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j) = \sum_{h=1}^{n} \hat{P}(\gamma_h^* = j|x_h)1(\hat{\gamma}_h^* = i) \tag{26}$$

is a consistent estimator of $P(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j)$, where $\hat{\gamma}_h^*$ denotes the data based classification of data point $\mathbf{x}_h$, estimating $\tilde{\gamma}_h^*$, by maximizing (23), which also defines $\hat{P}(\gamma_h^* = j|x_h)$. $1(\bullet)$ denotes the indicator function.

Therefore,

$$\hat{p}_{ij} = \frac{\hat{P}(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j)}{\hat{\pi}_j^*}$$

is a consistent estimator of p_{ij}. This works regardless of whether the mixture components are Gaussian distributions or mixtures of Gaussians. Therefore it is not needed to represent mixtures by their mean vectors and covariance matrices in order to compute $\hat{p}_{ij}$. The method of directly estimated misclassification probabilities (**DEMP method**) below therefore does not treat mixtures of Gaussians as single Gaussians in any way.

1. Choose a tuning constant $q^* < 1$.
2. Start with all components of the initially estimated Gaussian mixture as current clusters.
3. Compute $q = \max(\hat{p}_{ij}, \hat{p}_{ji})$ for all pairs of current clusters.
4. If $q < q^*$ for all pairs of current clusters, use the current clustering as the final one.
5. Otherwise, merge the pair of current clusters with maximum q and go to step 3.

5.4 The predictive strength method

Tibshirani and Walther's (2005) predictive strength approach to estimate the number of clusters is based on a different concept of misclassification. Instead of assessing

the classification of the data points to the clusters, it assesses how well it can be predicted whether pairs of points belong to the same cluster. Furthermore, instead of estimating the misclassification passively, by recomputing the clustering on subsamples, the approach does not only take into account the separation of the estimated clusters, but also the stability of the clustering solution.

For the merging problem, a special version of the predictive strength method is required. This leads to the following **predictive strength method** for the merging problem (assuming that s is the number of Gaussian mixture components estimated by EM/BIC):

1. Choose a tuning constant $c^* < 1$. For $k = 2, \ldots, s$, repeat m times:
2. Split the dataset in two halves.
3. Cluster both halves as follows:
   (a) Apply EM, fixing s.
   (b) Apply the DEMP method to the solution, stopping at k clusters.
4. Use the clustering $\mathcal{C}_1$ of the first half of the data to predict the cluster memberships (of clusters in $\mathcal{C}_1$) of the points of the second half of the data by maximizing (23) for every point of the second half with respect to the mixture components in $\mathcal{C}_1$.
5. For every cluster in the clustering $\mathcal{C}_2$ of the second half of the data, compute the proportion of correctly predicted co-memberships of pairs of points by the membership predictions of $\mathcal{C}_1$. Record the minimum over clusters $\tilde{c}$ of these proportions.
6. Repeat steps 3 and 4 exchanging the roles of the two halves.
7. Let c be the average of the $2m$ recorded values of $\tilde{c}$. Use the largest k with $c \geq c^*$ as the estimated number of clusters.

Conclusion

Several different hierarchical methods to merge Gaussian mixture components have been proposed. They correspond to different cluster concepts. The problem of merging Gaussian mixture components is not identifiable without subjective decisions about the cluster concept. Simulations, discussion, choice of the tuning constants and further details will be published elsewhere.

References

FRALEY, C. and RAFTERY, A. E. (2002) "Model-Based Clustering, Discriminant Analysis, and Density Estimation", *Journal of the American Statistical Association,* 97, pp. 611-631.

FUKUNAGA, K. (1990) *Introduction to Statistical Pattern Recognition*, 2nd edition, Academic Press, New York.

HARTIGAN, J. A. and HARTIGAN, P. M. (1985) "The dip test of unimodality", *Annals of Statistics,* 13, pp. 70-84.

MCLACHLAN, G. J. and PEEL, D. (2000), *Finite Mixture Models*, Wiley, New York.

RAY, S. and LINDSAY, B. G. (2005), "The Topography of Multivariate Normal Mixtures", *Annals of Statistics*, 33, pp. 2042-2065.

SCHWARZ, G. (1978), "Estimating the dimension of a model", *Annals of Statistics* 6, pp. 461-464.

TANTRUM, J., MURUA, A. and STUETZLE, W. (2003), "Assessment and Pruning of Hierarchical Model Based Clustering", *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C.,* pp. 197-205.

TIBSHIRANI, R. and WALTHER, G. (2005)] "Cluster Validation by Prediction Strength", *Journal of Computational and Graphical Statistics*, 14, pp. 511-528.

6 Classification of workers with different exposure levels to fumes of bitumen

Anne Spickenheuer, Monika Raulf-Heimsoth, Benjamin Kendzia, Thomas Brüning, and Beate Pesch
BGFA - Research Institute of Occupational Medicine,
German Social Accident Insurance,
Ruhr University Bochum
spickenheuer@bgfa.de raulf@bgfa.de kendzia@bgfa.de bruening@bgfa.de pesch@bgfa.de

Abstract

Bitumen is a widely used construction material. Emissions from hot bitumen applications are of long-standing health-concern. The objective of the German Human Bitumen Study is to investigate potential irritative and genotoxic effects in mastic-asphalt workers. Here discriminant analysis was used to classify mastic-asphalt and construction workers into exposure related classes using irritative biomarkers. Additionally, the workers are classified by smoking status using the same irritative biomarkers to prove whether the selected irritative variables measured in induced sputum (interleukin-8, neutrophil granulocytes, leukotriene B_4, interleukin-1β, nitrate and nitrite, and total protein) are capable of distinguishing between workers with different health effects. Linear discriminant analysis, quadratic discriminant analysis and the k-nearest neighbour method were applied. Current smokers and non-smokers could be distinguished best by 3-nearest neighbour (estimated error rate 0.14). None of the applied methods discriminated between the different exposure groups with acceptable error rates, especially highly exposed workers with a shift-exposure above 14.7 mg/m^3 fumes of bitumen were not correctly classified. Possible explanations are the small sample size of the highly exposed workers, the artificial cut-off at 14.7 mg/m^3, and the daily changing exposure levels of the workers.

Keywords: Bitumen, irritative effects, discriminant analysis

Introduction

Asphalt is a mixture of the binder bitumen with inorganic material (sand, gravel) (note: bitumen is referred to as asphalt in North America). It is used mainly for road paving. In the 1970s, bitumen replaced coal tar as a binder in Germany after tar was assessed as human carcinogen. Bitumen is the fraction of the crude oil remaining after distillation of the evaporated components and contains about 5 mg/kg benzo[a]pyrene, whereas coal tar contains approximately 5000 mg/kg benzo[a]pyrene

(IARC (1987)). The International Agency for Research on Cancer (IARC) classified extracts of steam-refined and air-refined bitumen as possible human carcinogens in Group 2B (IARC (1985, 1987)). In 2001, the German Committee for Hazardous Substances (AGS) lowered the occupational threshold limit for fumes of bitumen to 14.7 mg/m^3 (bitumen condensate standard), respectively 10 mg/m^3 (mineral oil standard) (Rühl (2001)). For mastic-asphalt workers the threshold limit was deferred, because it was assumed that a reduction of the exposure level below that threshold was not possible by technical measures. The objective of the German Human Bitumen Study was to investigate the exposure levels and potential irritative and genotoxic effects in mastic-asphalt workers. First analyses with data of this study pointed towards irritative effects of fumes of bitumen on the airways (Raulf-Heimsoth et al. (2007a), Raulf-Heimsoth et al. (2007b)).

A broad set of variables characterizes irritative effects. In order to implement this complex information in assessing health effects by exposure, discriminant analysis can be used to classify the workers into exposure-related classes regarding their health effects.

The traditional way of conducting discriminant analysis was introduced by R. A. Fisher, known as linear discriminant analysis (LDA) (Johnson and Wichern (2002)). The linearity of this method comes from the assumption of a common covariance matrix. Here, hyperplanes separate the classes. When the assumption of a common covariance matrix is not satisfied, one uses an individual covariance matrix for each group. This leads to a quadratic surface of the boundaries between discrimination regions and so this discrimination method is called quadratic discriminant analysis (QDA) (Henery (1994)). The k-nearest neighbour method classifies an observation by assigning the class label most frequently represented among the k nearest observations. Ties are broken at random. The k-nearest neighbour method is a non-parametric technique that is computationally complex. It is often successful when the decision boundary is irregular. Asymptotically the error rate of the 1-nearest neighbour method is never more than twice the Bayes rate (Duda et al. (2001)).

Here, three classification methods were applied to distinguish groups with different exposure levels of fumes of bitumen: workers with no exposure to fumes of bitumen serving as reference group, low-exposed workers with bitumen exposure lower than 14.7 mg/m^3, and highly exposed workers with a shift concentration above 14.7 mg/m^3.

6.1 Methods

The German Human Bitumen Study was conducted as a cross-sectional cross-shift study. The study group consisted of 280 bitumen-exposed men at 42 construction sites and 74 non-exposed male workers at 14 outdoor construction sites as reference group. A structured questionnaire was applied in a face-to-face interview to assess lifestyle habits, medical history, and other factors. All workers were examined after shift. Blood and urinary samples and induced sputum were collected to deter-

mine biomarkers of exposure, irritative or other health effects. Induced sputum was generated by inhalation of hypertonic saline solution and considered a non-invasive technique to collect biological material from the deeper airways. In induced sputum, cytokines were determined using commercial monoclonal "sandwich" enzyme immunoassays. During shift, personal air sampling in the workers' breathing zone was carried out to measure exposure to fumes of bitumen. The measurement was done with a German GGP sampler. Here, concentration of fumes of bitumen is given by bitumen condensate standard, which is about 1.47 times the concentration of fumes of bitumen given by mineral oil standard used in Germany until 2007 (BGIA (2008)). Further details about the study are given in Raulf-Heimsoth et al. (2007b). All study subjects provided written informed consent prior to examination. The study was approved by the Ethics Committee of the Ruhr-University Bochum and was conducted in accordance with the Helsinki Declaration.

Median and interquartile range were presented to describe the data. Linear discriminant analysis, quadratic discriminant analysis, and k-nearest neighbour were applied to classify the different groups by exposure or smoking status. The prior probabilities were set proportional to sample size. Additionally, results of LDA and QDA with equal prior probabilities were presented. Using the k-nearest neighbour method, the data was analyzed with different values of k. In the binary classification problem ties are avoided by selecting k odd. In the three group classification problem ties are broken at random. The predictive validity of the classification rules was assessed by the leave-one-out approach of Lachenbruch and Mickey (1968). The continuous variables were log-transformed because of their skewed distributions. Subjects with any missing values were excluded from the discriminant analysis. Values below limit of quantitation (LOQ) were set to 2/3 LOQ. The calculations were performed with the statistical software SAS/STAT, version 9.2 (SAS Institute Inc., Cary, NC, USA) (PROC DISCRIM). For the k-nearest neighbour method the leave-one-out method was programmed without using the CROSSVALIDATE option in SAS/STAT.

6.2 Results

Table 1: Characteristics of the study population of the Human Bitumen Study

	Reference group N = 74	Low-exposed workers $\leq$14.7 mg/m^3 N = 251	High-exposed workers $>$14.7 mg/m^3 N = 29
Age (years)	38	40	40
(median; interquartile range)	(32 - 46)	(33 - 47)	(35 - 46)
Current smokers N (%)	37 (50 %)	161 (64 %)	19 (66 %)
Fumes of bitumen (mg/m^3)	-	4.3	22.6
(median; interquartile range)		(2.2 - 7.3)	(18.1 - 34.5)

Table 1 depicts the characteristics of the study groups. 61 % of all workers were current smokers, with slightly less current smokers in the reference group (reference 50 %, low-exposed 64 %, high-exposed 66 %). Median shift concentration of fumes of bitumen in exposed workers (total) was 4.99 mg/m^3 with an interquartile range of 2.50 - 8.67 mg/m^3. 251 exposed workers had a shift concentration lower than 14.7 mg/m^3 of fumes of bitumen and 29 men had a concentration higher than 14.7 mg/m^3.

Table 2: Distribution of irritative biomarkers measured in induced sputum in German workers

	N_{miss}[a]	LOQ[b]	$N_{< LOQ}$	Reference group N = 74 Median (Q1 - Q3)[c]	Low-exposed workers N = 251 Median (Q1 - Q3)[c]	High-exposed workers N = 29 Median (Q1 - Q3)[c]
Interleukin-8	5	3 pg/ml	1	1207 (522 - 3278)	3714 (1837 - 9196)	3145 (929 - 13402)
Neutrophil[d]	13	- [$\times 10^4$]	-	5.6 (1.3 - 33.2)	4.2 (0.2 - 21.8)	15.6 (1.5 - 90.0)
Leukotriene B_4	35	11.7 pg/ml	0	1662 (1147 - 2501)	1735 (1085 - 2485)	2236 (915 - 2594)
Interleukin-1β	35	0.4 pg/ml	1	19.0 (8.7 - 34.7)	24.3 (14.3 - 50.4)	26.1 (10.9 - 36.3)
Nitrate/ nitrite	6	5 μM	38	7.5 (5.3 - 12.2)	16.2 (9.6 - 25.6)	13.5 (8.6 - 19.4)
Total protein	11	10 μg/ml	0	339 (210 - 638)	714 (430 - 1110)	411 (244 - 1165)
Interleukin-5	5	2 pg/ml	88	1.3 (1.3 - 12.4)	10.4 (5.5 - 22.2)	4.6 (1.3 - 15.0)
Interleukin-6	36	3 pg/ml	111	7.5 (2.0 - 31.0)	16.0 (2.0 - 64.0)	64.3 (2.0 - 134.6)

[a] N_{miss}: number of missing observations

[b] LOQ: limit of quantitation

[c] Q1 - Q3: interquartile range

[d] Neutrophil granulocytes

Table 2 shows the distribution of biomarkers of irritative effects in induced sputum that were candidates for the discriminant analyses. Interleukin-5 and interleukin-6 were below LOQ for 88 respectively 111 workers and, therefore, excluded from analysis. All other variables served as discriminant variables.

In order to prove whether the selected irritative parameters are capable of distinguishing between workers with different health effects, a discriminant analysis was performed for smoking status because smoking is known to induce irritative effects.

Table 3: Results of discriminant analyses with smoking status as class variable and age and irritative biomarkers[a] measured in induced sputum as discriminant variables using leave-one-out cross-validation

	Error rate	Correct classification (%)	
		Non-smoker N = 115	Current smoker N = 182
Proportional prior			
Linear discriminant analysis	0.29	62 (54%)	150 (82%)
Quadratic discriminant analysis	0.30	60 (52%)	148 (81%)
Equal prior			
Linear discriminant analysis	0.27	83 (72%)	136 (75%)
Quadratic discriminant analysis	0.31	79 (69%)	125 (69%)
3-Nearest neighbour	0.14	92 (80%)	164 (90%)

[a] interleukin-8, neutrophil granulocytes, leukotriene B_4, interleukin-1β, nitrate and nitrite, and total protein

Table 3 presents the classification results for LDA, QDA and 3-nearest neighbour to categorize non-smokers and current smokers. The 3-nearest neighbour method distinguished smokers from non-smokers better than LDA and QDA with an estimated error rate of 0.14 using leave-one-out cross-validation. Figure 50 shows the error rate as a function of the numbers of neighbours. To avoid tied votes k is selected odd. To categorize the smoking status $k = 3$ yields the smallest error rate.

In the next step, the discriminant analysis was applied to classify workers according to the different exposure levels (none, low and high exposure). The analysis was stratified by smoking status, because smoking was a strong irritative factor. The presented results in Figure 50 and Table 4 show the combined classification results for non-smokers and current smokers. Applying 2-nearest neighbour, the estimated error rate for non-smokers was 0.18 and for current smokers 0.24 using leave-one-out cross-validation. Combining the classification results of non-smokers and current smokers yields an estimated misclassification error of 0.22. Estimated error rates of k-nearest neighbour of different values of k are shown in Figure 1. The smallest estimated error rate has 2-nearest neighbour but this classification rule produces many tied votes. Five-nearest neighbour has the advantage that less ties occur. Both results together with LDA and QDA are presented in Table 4. Highly exposed workers were poorly classified by all methods except 2-nearest neighbour. Using equal priors instead of proportional priors by LDA and QDA improved the

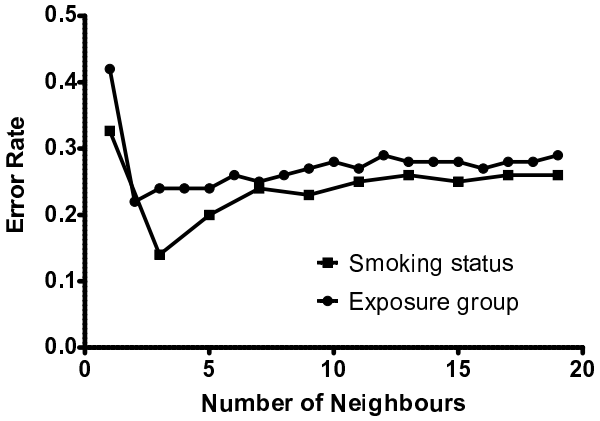


Figure 50: Estimated error rate of k-nearest neighbour with different values of k using leave-one-out cross-validation classifying smoking status and exposure group

classification of highly exposed workers and workers of the reference group, whereas the large group of low-exposed workers were classified worse.

6.3 Discussion

The 3-nearest neighbour method performed better than LDA and QDA to differentiate between non-smokers and current smokers. The two smoking groups could be well distinguished. Therefore, the selected biomarkers seem to be capable of detecting irritative effects. The classification of the three exposure groups was more difficult and none of the methods discriminated with acceptable error rates. Especially the group of the highly exposed workers could not be distinguished from the other groups. There are various possible explanations for this poor result. First, the group of the high-exposed workers was small and, therefore, likely not informative to develop a decision rule. Two-nearest neighbour was the only method that correctly classified a highly exposed worker with a higher probability than classifying at random. This method uses very few observations to classify a new subject that supports this theory. Second, the cut-off to separate highly exposed workers from low-exposed workers is artificial, and there is no clear cut point between these two groups. Third, a worker exposed during the observed shift at a high level could be exposed to lower levels in the preceding shifts and vice versa. Other analyses revealed that the observed biomarkers indicate chronic effects as well and do not clearly show short-time effects during a single working shift (data not shown). Overall, it was possible to differentiate unexposed workers from exposed workers. The differences between reference workers and exposed workers to fumes of bitumen might be caused by bitumen exposure, but also different working conditions might have contributed to the effects. In the final analysis of this study, further discriminant analysis methods shall be applied to improve the classification. One emphasis will be set on statistical methods that can deal with categorical variables as for

Table 4: Results of discriminant analyses with exposure group as class variable and age and irritative biomarkers[a] measured in induced sputum as discriminant variables stratified by smoking status using leave-one-out cross-validation

	Error rate	Correct classification (%)		
		Reference	Low-exposed workers	High-exposed workers
		N = 71	N = 204	N = 22
Proportional prior				
Linear discriminant analysis	0.25	28 (39%)	193 (95%)	1 (5%)
Quadratic discriminant analysis	0.30	25 (35%)	181 (89%)	2 (9%)
Equal prior				
Linear discriminant analysis	0.48	39 (55%)	108 (53%)	6 (27%)
Quadratic discriminant analysis	0.39	47 (66%)	132 (65%)	3 (14%)
2-Nearest neighbour	0.22	41 (58%)	179 (88%)	13 (59%)
5-Nearest neighbour	0.24	35 (49%)	189 (93%)	2 (9%)

[a] interleukin-8, neutrophil granulocytes, leukotriene B_4, interleukin-1β, nitrate and nitrite, and total protein

example CART in order to include variables with a fair amount of measurements below LOQ in the analysis. Besides classification techniques, regression models will be applied to analyze the influence of fumes of bitumen on different biomarkers as shown in the analyses presented by Raulf-Heimsoth et al. (2007b).

Acknowledgment

This study was supported by *Deutsche Gesetzliche Unfallversicherung* (DGUV).

References

BGIA - INSTITUT FÜR ARBEITSSCHUTZ DER DEUTSCHEN GESETZLICHEN UNFALLVERSICHERUNG (2008): *Messung von Gefahrstoffen - BGIA Arbeitsmappe, Expositionsermittlung bei chemischen und biologischen Einwirkungen.* Erich Schmidt Verlag, Berlin.

DUDA, R.O., HART, P.E., and STORK, D.G. (2001): *Pattern Classification.* Wiley, New York.

HENERY, R. J. (1994): Classical statistical methods. In: D. Michie, D.J. Spiegelhalter, and C.C. Taylor (Eds.); *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, New York, 17-28.

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER (1985): *IARC Monographs on the evaluation of carcinogenic risk of chemicals to humans. Polycyclic aromatic compounds.* IARC, Lyon, Volume 35, part 4, 39-81.

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER (1987): *IARC Monographs on the evaluation of carcinogenic risk of chemicals to humans. Overall evaluation of carcinogenicity: An updating of IARC monographs.* IARC, Lyon, Vol. 1-42.

JOHNSON, R.A. and WICHERN, D.W. (2002): *Applied Statistical Multivariate Analysis.* Prentice Hall, 5th edn, New Jersey.

LACHENBRUCH, P. and MICKEY, M. R. (1968): Estimation of error rates in discriminant analysis. *Technometrics, 10, 1-11.*

RAULF-HEIMSOTH, M., PESCH, B., SCHOTT, K., KAPPLER, M., PREUSS, R., MARCZYNSKI, B., ANGERER, J., RIHS, H. P., HAHN, J. U., MERGET, R., and BRÜNING, T. (2007a): Irritative effects of fumes and aerosols of bitumen on the airways: results of a cross-shift study. *Arch Toxicol, 81, 35-44.*

RAULF-HEIMSOTH, M., PESCH, B., SPICKENHEUER, A., BRAMER, R., SCHOTT, K., MARCZYNSKI, B., BREUER, D., HAHN, J. U., MERGET, R., and BRÜNING, T. (2007b): Assessment of irritative effects of fumes of bitumen on the airways using non-invasive methods - results of a cross-shift study in mastic asphalt workers. *JOEH, 4(S1), 223-227.*

RÜHL, R. (2001): Bitumen krebserzeugend? Eine Erläuterung des Gesprächskreises Bitumen zur Neubewertung durch die MAK-Kommission. *Gefahrstoffe - Reinhaltung der Luft, 61, 519-520.*

Part II

Data Analyses

As in the past years, the participants of the meeting were invited to take part in a data analysis experiment. There is a plethora of existing data analysis methods and the idea is to apply them to various data sets with different characteristics in order to learn about their strengths and weaknesses. For this purpose, three data sets were issued about two months before the event. One of them, the Tiles Data Set, is of archaeological origin made available by the working group J. Dolata/H.-G. Bartel/H.-J. Mucha. The two others are synthetic and designed by G. Ritter. Three working groups took part in the experiment, Gerhard Pöppel and Reinhard Schachtner from Infineon Regensburg, Gunter Ritter from the University of Passau, and Susanna Röblitz and Marcus Weber from the Zuse Institute Berlin. The methods employed were Projection Pursuit, Spectral Clustering, MCLUST, and the Trimmed Determinant Criterion. We now present the results ordered by data sets.

7 Roman Tiles Data Set

An introduction to the Roman tiles data set was presented in Hans–Georg Bartel's contribution, Sect. 3.

7.1 Gerhard Pöppel and Reinhard Schachtner: Analysis I by Projection Pursuit

Meanwhile it is an established tradition that for the Fall Meeting of the AG-DANK given data sets should be analyzed by the participants. This kind of contest affords an opportunity to get an impression which kind of data can be treated with which kind of method to fulfill the given task. Because different participants usually use different approaches one can learn a lot about appropriate methods.

We tried to analyze the three data sets within limited time of half a day all in all. Fifty percent of our restricted time is spent on the data set "Archaeometric data of Roman bricks and tiles from the Rhine area of Germany". The rest of the time goes in equal parts to the artificial data sets "Berlin08_synth1" and "Berlin08_synth2" which are also very interesting.

We mainly used Projection Pursuit and Mixture Models for our three analyses. We tried to analyze the data without additional background information which can be found in the corresponding papers: Dolata (2000), Mucha et al. (2005), Mucha et al. (2003).

Group	Run 10 Projection 688		SiO2	TiO2	Al2O3	Fe2O3	MnO	MgO	CaO	Na2O	K2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba		Counts
not_further_classified	0.00	Mean	68.81	1.10	16.25	**5.02**	0.05	1.63	3.56	0.58	2.80	87.64	138.19	55.76	**97.57**	131.33	165.79	33.87	258.96	25.36	491.21		75.00
A	1.00	Mean	74.14	0.66	14.82	4.23	0.03	1.14	0.83	**1.00**	3.05	68.21	82.44	35.17	63.73	143.73	112.87	36.35	**295.55**	14.24	**661.56**		107.00
B	2.00	Mean	72.86	**1.73**	**16.84**	4.53	0.03	0.90	0.62	0.22	2.16	**95.27**	**158.63**	46.60	45.94	118.72	137.63	33.88	288.87	**39.90**	422.07		100.00
C	3.00	Mean	**76.06**	0.88	15.11	3.68	0.02	0.86	0.56	0.25	2.47	69.74	75.37	32.58	34.73	130.21	110.03	24.56	221.11	25.55	456.40		62.00
D	5.00	Mean	61.42	0.63	14.57	4.89	**0.08**	**2.53**	**12.12**	0.67	2.85	84.35	99.85	**56.76**	90.58	130.76	**299.56**	26.22	135.10	14.72	416.73		253.00
E	8.00	Mean	67.48	0.65	15.03	4.77	0.06	2.33	5.47	0.83	**3.18**	81.02	99.29	48.16	83.02	**148.02**	169.33	**36.63**	268.13	15.44	544.89		63.00
																							660.00

Figure 51: Cluster characterization via the mean values of the archaeometric original variables

On one hand this is a possibly unnecessary restriction and a loss of available information, on the other hand this additional information would be not available if you did the analysis from beginning. Let us see what is the outcome if one has only the data and the task to cluster the data. In any case before we select any method we always have a view on the original variables and do some univariate statistics on that. So one gets an impression about their distributions, about univariate outliers etc. From that we decided to use Projection Pursuit (Friedman and Tukey (1974), Posse (1995)) for exploratory data analysis in order to find the "best" 2-dimensional

views to the dataset. These "best" views are defined by the highest projection index.

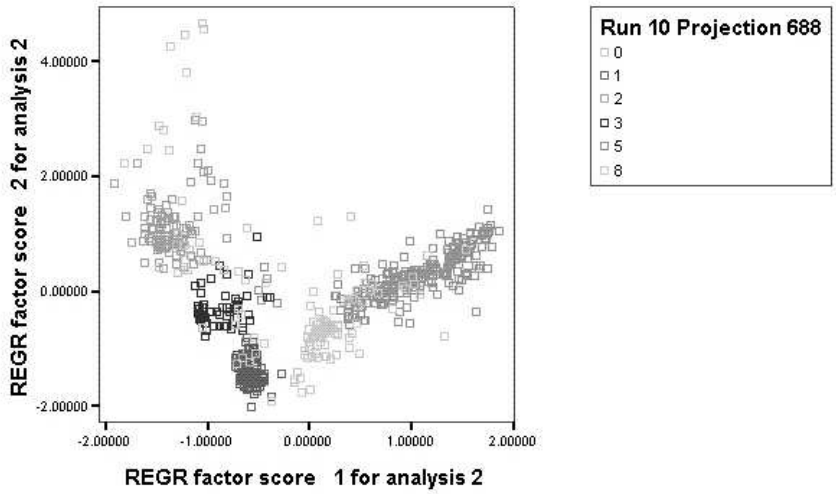


Figure 52: Scatterplot of the first two principal components of the archaeometric data

We use a Chi-Square-Index with 9 cells as described in Rohatsch et al. (2006). Instead of sphered data we use z-standardized variables with an anti-index to avoid trivial views of linear correlations. It turns out that Projection Pursuit does a good job on that dataset. One can easily mark different clusters from different views and we stopped investigating after the finding of 6 regions (5 selected clusters and one rest-cluster which includes some further substructure). To characterize the clusters, the mean values of the original variables can be considered (see Fig. 51).

Typical accumulations of oxids and trace elements characterize the clusters. In many cases also simple scatterplots after PCA of the original variables show some of the underlying structure found by Projection Pursuit, as can be seen in Fig. 52:

There are many different views of Projection Pursuit which allows the recognition of substructures in already found clusters. So it seems that Projection Pursuit can detect a lot of further details which may be of interest for the specialists.

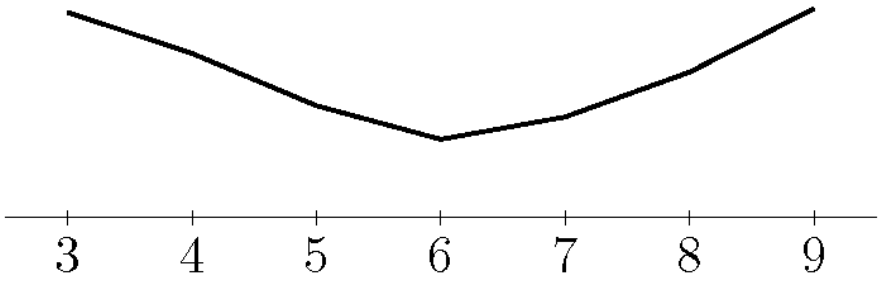


Figure 53: Tiles data: the BIC curve of the favorite solutions with three to nine clusters suggested by the posterior-density-HDBT-ratio plot.

7.2 Gunter Ritter: Analysis II by Model Based Clustering

I clustered the data set with a statistical criterion based on a heterogeneous classification model with unknown cluster sizes for normal data with full covariance matrices and allowing gross outliers, the TDC (Trimmed Determinant Criterion) according to Gallegos and Ritter (2009), see also Gallegos and Ritter (2005). Visual inspection of the 2D scatter plots suggests that the data set contains outliers. I assumed an amount of ten percent. Visual inspection also shows that some of the features are noticeably correlated. For this reason and in order to diminish sample space dimension, I deleted SiO_2, TiO_2, CaO, and K_2O from the feature list so that sample space dimension is fifteen. The number of clusters was determined with the BIC model selection criterion.

The BIC curve, see Fig. 53, clearly pleads for six clusters. However, there may be one or a few small clusters, in particular of size ≤ 15, hidden in the set of 66 discarded elements (outliers) which I did not further analyze. Cluster sizes of the favorite partition with six clusters are 145, 111, 111, 105, 61, and 61. Visual inspection of the clusters obtained suggests that there are more outliers than the assumed 66. The MnO-Y-plot of the partition is presented in Fig. 54.

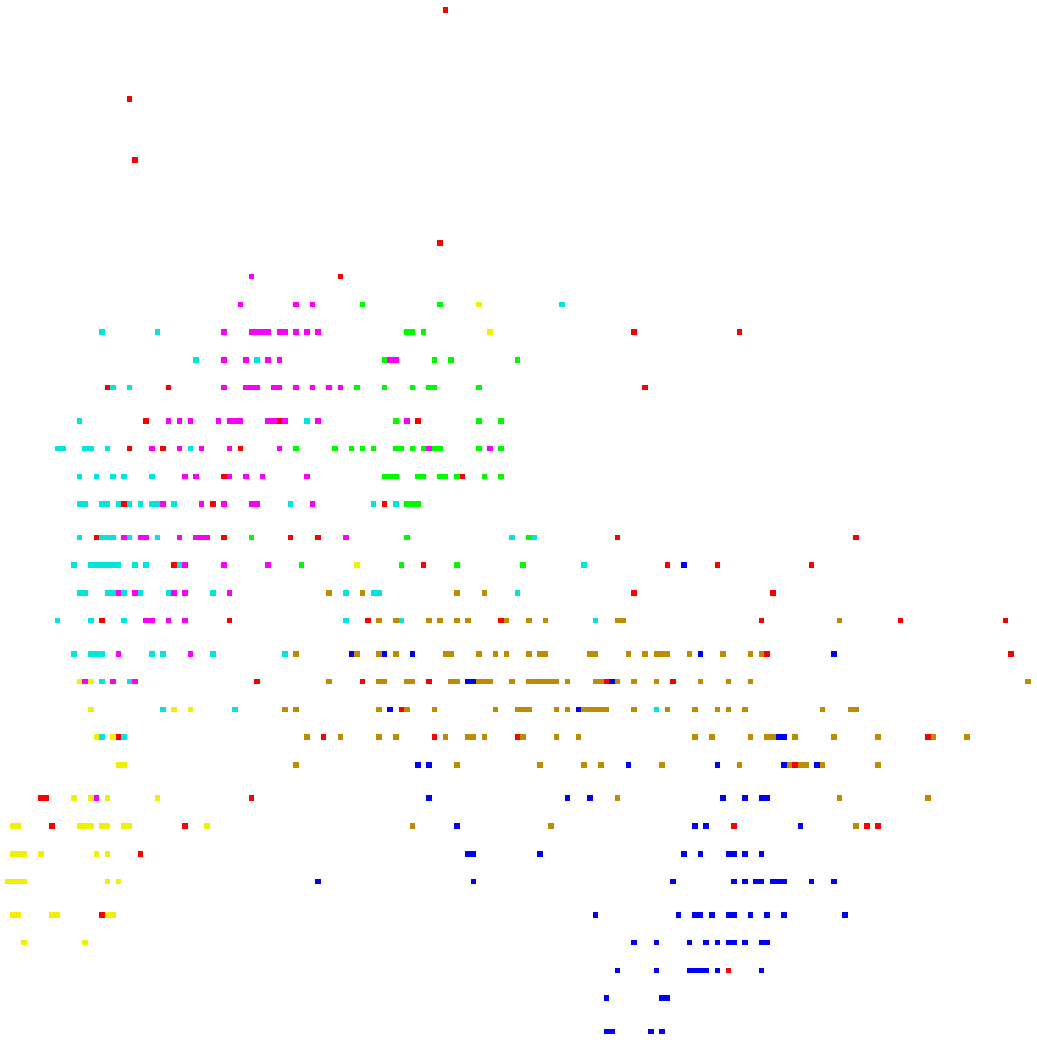


Figure 54: MnO-Y plot of the favorite partition obtained from the heteroscedastic TDC with model selection criterion BIC. Outliers are plotted in red.

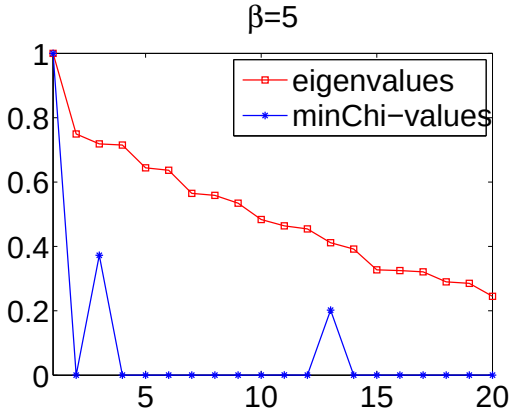


Figure 55: Tiles from Dolata: minChi-values and eigenvalues for different numbers of clusters.

7.3 Susanna Röblitz and Marcus Weber: Analysis III by Spectral Clustering

The method applied here is described by the authors at full length in Sec. 4. Because spectral clustering is scale dependent, a preprocessing step before distance computation is recommended. Since the different features have different ranges of values, we first normalized the data column-wise by dividing every column by its mean value. Afterwards, we computed the matrix D^{eff} of all pair-wise effective distances (fully connected graph). At this point, we figured out that there are two objects with the same effective distance to all other objects, namely H169 and H301. That means, the effective distances between all other objects are smaller than the effective distances between H169 or H310 and the other objects. In other words, these two objects represent outliers. The cluster algorithm will always put them into two single-object clusters. Therefore, we removed them from the data before we started the algorithm.

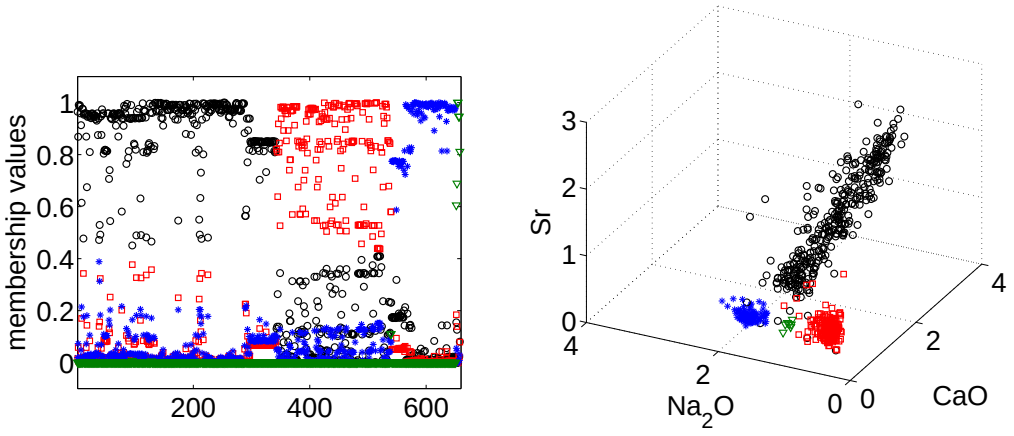
In the computation of the Gaussian similarity function (17), we set the parameter $\beta = 5$. However, the results turned out to be quite insensitive with respect to the choice of β in the interval $[0.5, 10]$. Fig. 55 illustrates the eigenvalues and the minChi-values of the initial guess for different numbers of clusters $k = 1, \ldots, 20$. Based on this information, we decided to consider the choices $k = \{4, 6, 8, 9\}$ in detail.

The resulting membership values for $k = 4$ are illustrated in Fig. 56. In fact, the separation between the four clusters is clearly visible on the (Na_2O,CaO,Sr) submanifold. Re-transformation of membership values to indicator vectors gives the assignment of objects to the clusters as listed in Tab. 5.

If we choose $k = 6$, the partition of objects into clusters is the same as for $k = 4$, accept the two objects H880 and H857, which are isolated from cluster 2 and put into single clusters.

Table 5: Tiles from Dolata: Partition into $k = 4$ clusters. The objects have been numbered according to their position in the original data file.

k	objects
1	5 6 7 8 9 10 11 12 13 14 16 47 48 49 50 51 52 53 54 55 56 57 58 60 61 63 64 65 66 67 68 69 70 71 72 73 74 80 88 90 102 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 153 154 155 156 157 158 159 160 161 162 163 164 165 167 168 169 174 181 185 187 195 199 200 201 203 204 220 226 227 228 231 232 233 235 236 237 240 241 242 249 262 264 267 270 275 276 277 280 282 283 286 287 288 290 293 294 295 296 297 299 300 301 302 303 304 305 306 307 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 346 347 350 353 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 381 382 383 384 385 386 387 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 416 417 469 472 492 493 496 506 508 509 510 511 512 513 514 515 516 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 574 575 576 577 578 579 580 581 582 583 584 585 586 587 603 604 605 606 607 613 614 615 616 617 618 619 620 621 623 624 625 626 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660
2	1 2 3 4 15 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 62 75 94 100 101 103 104 105 106 107 108 109 110 111 112 113 114 115 116 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 166 170 171 172 173 175 176 177 178 179 182 183 184 186 188 189 190 191 192 193 194 196 202 205 206 207 208 209 210 211 212 213 214 215 216 217 219 221 223 224 225 229 230 234 238 239 244 248 252 253 254 255 259 261 265 266 268 269 271 272 273 274 281 284 285 289 291 292 298 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 345 351 352 355 380 415 466 467 468 470 471 473 474 475 476 486 487 488 489 490 491 494 495 497 498 499 500 501 502 503 504 505 507 608 609 610 611 612 622 627
3	59 76 77 78 79 81 82 83 84 85 86 87 89 91 92 93 95 96 97 98 99 180 218 222 243 245 246 247 251 260 263 278 279 308 343 344 348 349 354 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 477 478 479 480 481 482 483 484 485 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602
4	197 198 250 256 257 258 573



(a) Sorted membership values.

(b) Data points in the 3-dimensional subspace (Na_2O,CaO,Sr).

Figure 56: Tiles from Dolata: Partition into $k = 4$ clusters.

For $k = 8$, we obtain three isolated objects (H888, H857, G017) and cluster 2 is split into two clusters, compare Tabs. 5 and 6. The membership values are illustrated in Fig. 57. The two views in Fig. 58 show that the clusters are indeed separated from each other.

$k = 9$ results in the same clustering as $k = 8$, accept object G018 (no. 198), which is isolated from cluster 5 and put into a single cluster.

A further increase of the number of clusters leads to a splitting of clusters and separation of outliers. Tab. 7 contains the values of the objective function $I(\mathcal{A}; X, \pi)$ from (18). The values I/k are quite similar such that it is difficult to determine the "best" number of clusters.

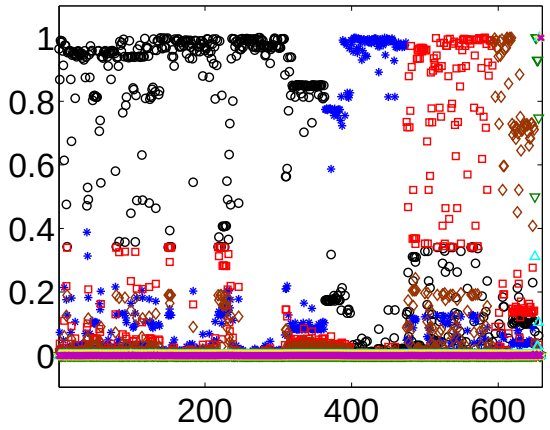
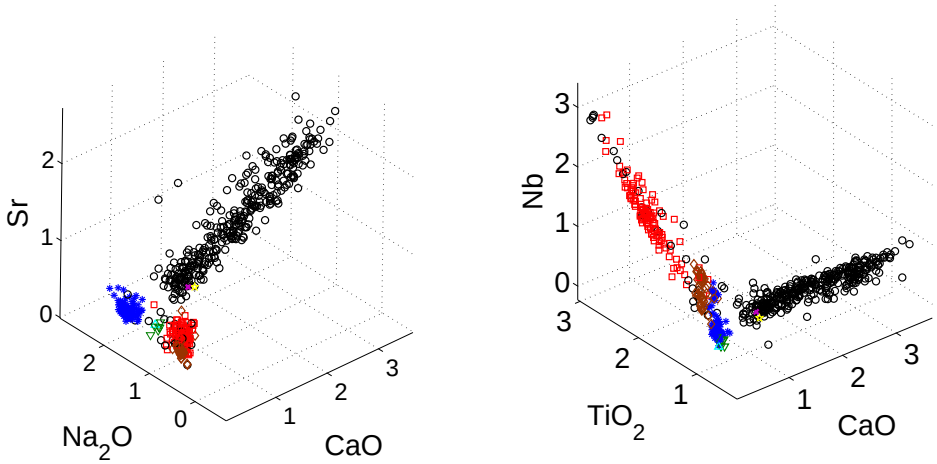


Figure 57: Tiles from Dolata: Sorted membership values for the partition into $k = 8$ clusters. Three clusters contain only one object, such that only five clusters are shown here.



(a) (b)

Figure 58: Tiles from Dolata: Data points in different sub-spaces.

Table 6: Tiles from Dolata: Partition into $k = 8$ clusters. The objects have been numbered according to their position in the original data file.

k	objects
1	5 6 7 8 9 10 11 12 13 14 16 18 47 48 49 50 51 52 53 54 55 56 57 58 60 61 63 64 65 66 67 68 69 70 71 72 73 74 80 88 90 102 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 153 154 155 156 157 158 159 160 161 162 163 164 165 167 168 169 174 179 181 185 187 189 195 199 200 201 203 204 220 223 226 227 228 231 232 233 235 236 237 240 241 242 249 253 262 264 267 270 275 276 277 280 282 283 286 287 288 290 293 294 295 296 297 299 300 301 302 303 304 305 306 307 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 328 331 334 336 346 347 350 353 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 381 382 383 384 385 386 387 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 416 417 466 467 469 470 471 472 486 487 488 489 490 491 492 493 495 496 506 508 509 510 511 512 513 514 515 516 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 574 575 576 577 578 579 580 581 582 583 584 585 586 587 603 604 605 606 607 613 614 615 616 617 618 619 620 621 623 624 625 626 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660
2	59 76 77 78 79 81 82 83 84 85 86 87 89 91 92 93 95 96 97 98 99 180 218 222 243 245 246 247 251 260 263 278 279 308 343 344 348 349 354 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 477 478 479 480 481 482 483 484 485 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602
3	1 2 3 4 15 17 19 20 39 94 103 104 108 113 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 166 170 171 172 173 175 177 178 182 183 184 186 188 190 191 192 193 194 196 202 205 206 207 208 209 210 211 212 213 214 215 216 217 219 221 224 225 229 230 234 238 244 248 252 254 255 261 265 266 268 269 284 285 289 291 292 298 326 327 329 330 332 333 335 337 338 339 340 341 342 345 351 468 473 474 475 476 497 498 499 500 501 502 503 504 505 507
4	21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 40 41 42 43 44 45 46 62 75 100 101 105 106 107 109 110 111 112 114 115 116 176 239 259 271 272 273 274 281 352 355 380 415 494 608 609 610 611 612
5	198 250 256 257 258 573
6	197
7	627
8	622

Table 7: Tiles from Dolata: Values of the objective function I in the computed (local) optimum for different cluster numbers k.

k	4	6	8	9
I	3.34	4.91	6.21	7.07
I/k	0.84	0.82	0.78	0.79

7.4 Hans-Joachim Mucha, Hans-Georg Bartel, Jens Dolata: Vergleich der Klassifikationsergebnisse zum *"Roman Tiles Data Set"*

Als Vergleichspartition für die vorhergehenden drei hier präsentierten Klassifikationsergebnisse wird die mit dem (gewöhnlichen) hierarchischen Ward-Verfahren ([10]) berechnete Partition P8DANK in acht Cluster benutzt. (P8DANK steht für die Partition in acht Cluster für den auf der DANK-Tagung verfügbar gemachten *"Roman Tiles Data Set"*.) Diese ist das Ergebnis auf Basis des quadrierten euklidischen Distanzmaßes, angewandt auf die Datenmatrix, deren Elemente (Messwerte) zuvor gemäß Wert/Mittelwert transformiert wurden (siehe Sec. 3). Die Zerlegung in acht Klassen wird durch den Ellbogentest nahe gelegt (Fig. 59). Die 8-Klassen-Partition P8DANK ist hinsichtlich der sinngebenden archäologischen Interpretation ähnlich zu den Analysen der 613 x 19-Matrix (siehe unten und Sec. 3). Der Ellbogentest zeigt, dass auch eine 5-Klassen-Partition statistisch auffällig ist.

Die Festlegung einer Vergleichspartition ist notwendig, weil es sich hier um eine reale, durch physikalische Messungen an Fundobjekten erhaltene Datenmatrix handelt, deren wahre Klassenzerlegung *a priori* unbekannt ist.

Die Partition P8DANK kann mit üblicher Standardsoftware unter Benutzung der hier erstmalig in Fig. 63 bis Fig. 74 angegeben kompletten 660 x 19-Datenmatrix erhalten werden. Hierbei liegen die ersten 613 Objekte (Proben) den Ausführungen in Sec. 3 zugrunde, während die weiteren 47 Proben mit dem Fundort Boppard am Rhein den Datensatz *"Roman Tiles"* vervollständigen. Die Werte der Variablen 1 bis 9 (Oxide) in der 660 x 19-Datenmatrix sind gerundet angegeben (siehe die entsprechenden Skalenfaktoren). Die genauen Werte findet man in [1], [6] sowie auf der Internetseite der Arbeitsgruppe AG–DANK unter `http://www.fim.uni-passau.de/fileadmin/files/lehrstuhl/ritter/agdank/TilesFromDolata.txt`.

Im Folgenden wird die Partion P8DANK jeweils den oben präsentierten Klassifikationsergebnissen P5PP+rest von Gerhard Pöppel und Reinhard Schachtner (Sec. 7.1), P6MB+outliers von Gunter Ritter (Sec. 7.2) sowie P4SC und P8SC von Susanna Röblitz und Marcus Weber (Sec. 7.3) gegenübergestellt. Die von den obigen Autoren dokumentierten Mini-Cluster sind der besseren Lesbarkeit wegen hierbei nicht berücksichtigt. Aus dem gleichen Grunde sind die Zeilen und Spalten der Kreuztabellen (Kontingenztafeln) umgeordnet worden und zwar so, dass die größten Zahlen

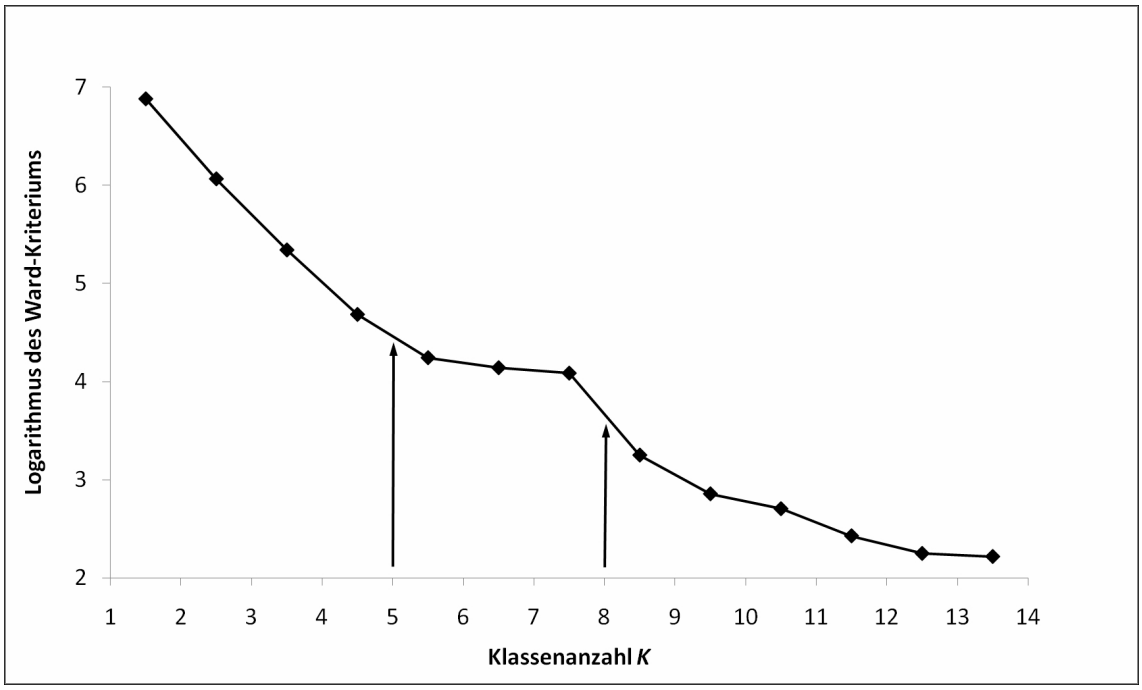


Figure 59: Auswahl einer optimalen Clusteranzahl des Ward-Verfahrens mit dem Ellbogentest

	P8DANK								
	Cluster Nr.								
P8ModWARD $\cup$ $C_{\textbf{Boppard}}$	1	2	3	4	5	6	7	8	Total
Frankfurt-Nied A	**92**								92
Frankfurt-Nied B	14	**24**							38
Unbekannt 3				7					7
Großkrotzenburg	1				**63**				64
Straßburg-Königshofen						**114**			114
Rheinzabern B, inkl. Worms	6		**88**						94
Rheinzabern A.1			5				**110**		115
Rheinzabern A.2, inkl. Unbekannt 1								**89**	89
C_{Boppard}			**45**	2					47
Total	113	24	138	9	63	114	110	89	660

Figure 60: Vergleich der Klassenzerlegungen P8ModWARD $\cup$ C_{Boppard} und P8DANK

in den 'Hauptdiagonalen' positioniert sind. Von allen hier eingesetzten Verfahren werden die Ziegel aus der Provenienz 'Straßburg-Könighofen' mit hoher Trefferquote als ein Cluster identifiziert. Dieses stabilste Cluster ist in den folgenden Kreuztabellen in fetter Schrift hervorgehoben. Weitere stabile und in hohem Grade reproduzierbare Herstellungsorte sind 'Großkrotzenburg' und 'Frankfurt-Nied A'. Das Auffinden des kleinen Clusters 'Unbekannt 3' (siehe Sec. 3 und Fig. 60) gelingt nur mit "spectral clustering" (letzte Spalte der Kreuztabellen P8DANK-P8SC und P8DANK-P4SC).

P8DANK - P5PP+rest (siehe "Analysis I"):

$$\begin{bmatrix} \mathbf{106} & 0 & 0 & 0 & 0 & (8) \\ 0 & 99 & 0 & 0 & 0 & (11) \\ 0 & 0 & 87 & 0 & 7 & (19) \\ 0 & 61 & 0 & 63 & 0 & (14) \\ 0 & 5 & 0 & 0 & 54 & (4) \\ 0 & 88 & 0 & 0 & 0 & (1) \\ 0 & 0 & 13 & 0 & 1 & (10) \\ 1 & 0 & 0 & 0 & 0 & (8) \end{bmatrix}$$

Die letzte Spalte stellt hier kein Cluster dar (die entsprechenden Zahlen sind deshalb eingeklammert), sondern laut dem Text von Pöppel und Schachtner in Sec. 7.1 einen Rest, der nicht weiter untersucht wurde. Auf die gleiche Weise ist die Menge der Ausreißer ("outliers") in der letzten Spalte der folgenden Tabelle markiert.

P8DANK - P6MB+outliers (siehe "Analysis II"):

$$\begin{bmatrix} \mathbf{111} & 0 & 0 & 0 & 0 & 0 & (3) \\ 0 & 96 & 0 & 0 & 0 & 4 & (13) \\ 0 & 0 & 76 & 24 & 0 & 0 & (10) \\ 0 & 0 & 19 & 65 & 0 & 0 & (5) \\ 0 & 0 & 50 & 16 & 61 & 0 & (11) \\ 0 & 0 & 0 & 0 & 0 & 57 & (6) \\ 0 & 15 & 0 & 0 & 0 & 0 & (9) \\ 0 & 0 & 0 & 0 & 0 & 0 & (9) \end{bmatrix}$$

P8DANK - P8SC (siehe "Analysis III"):

$$\begin{bmatrix} 137 & 0 & 0 & 0 & 0 \\ 3 & \mathbf{111} & 0 & 0 & 0 \\ 9 & 0 & 104 & 0 & 0 \\ 1 & 0 & 5 & 57 & 0 \\ 0 & 0 & 0 & 0 & 6 \\ 109 & 0 & 0 & 0 & 0 \\ 89 & 0 & 0 & 0 & 0 \\ 15 & 0 & 9 & 0 & 0 \end{bmatrix}$$

P8DANK - P4SC (siehe “Analysis III”):

$$\begin{bmatrix} 137 & 0 & 0 & 0 \\ 3 & \mathbf{111} & 0 & 0 \\ 3 & 0 & 110 & 0 \\ 0 & 0 & 24 & 0 \\ 0 & 0 & 2 & 7 \\ 1 & 0 & 62 & 0 \\ 109 & 0 & 0 & 0 \\ 89 & 0 & 0 & 0 \end{bmatrix}$$

Bislang wurde als Vergleichspartion jeweils P8DANK benutzt. Vergleicht man hingegen das Ergebnis von Gerhard Pöppel und Reinhard Schachtner (P5PP+rest) mit demjenigen von Gunter Ritter (P6MB+outliers), so erhält man vier sehr gut übereinstimmende Cluster, wie die zuhehörige Kontingenztafel verdeutlicht.

P5PP+rest - P6MB+outliers:

$$\begin{bmatrix} 134 & 0 & 0 & 0 & 4 & 103 & (12) \\ 0 & \mathbf{104} & 0 & 0 & 0 & 0 & (3) \\ 0 & 0 & 96 & 0 & 0 & 0 & (4) \\ 2 & 0 & 0 & 57 & 0 & 1 & (3) \\ 0 & 0 & 4 & 0 & 51 & 0 & (7) \\ (9) & (7) & (11) & (4) & (6) & (1) & (37) \end{bmatrix}$$

Dies scheint die beste Übereinstimmung zu geben, und sie ist fast perfekt, wenn man bedenkt, dass Pöppel und Schachtner ein Cluster weniger haben. Wie bereits weiter oben erwähnt, stellen die in Klammern angegebenen Zahlen (hier in der letzten Zeile) kein Cluster dar. Wie weiter aus der obigen Tafel erkennbar ist, besteht dieser Rest zum großen Teil aus Beobachtungen, die das Verfahren von Gunter Ritter als Ausreißer qualifiziert ($n = 37$).

Im Folgenden soll die hier benutzte Vergleichspartition P8DANK mit den in Sec. 3 skizzierten archäometrischen Ergebnissen in Beziehung gebracht werden. Es seien die in Fig. 43 der Sec. 3 charakterisierte 8-Klassen-Partition, die bei der Clusteranalyse der 613 x 19-Datenmatrix unter Verwendung des modifizierten Ward-Verfahrens erhalten wurde, mit P8ModWARD und die Menge der 47 in Sect. 3 nicht besprochenen, in Boppard gefundenen Objekte mit C_{Boppard} bezeichnet. Die in Fig. 60 wiedergegebene Gegenüberstellung der Vergleichpartition P8DANK mit P8ModWARD $\cup$ C_{Boppard} zeigt, dass beide Klassenzerlegungen im Wesentlichen übereinstimmen.

Die Objekte (Proben) der Menge C_{Boppard} werden zwei Klassen von P8DANK zugeordnet: die überwiegende Mehrheit von 45 Objekten (d.h. 95,7 %) zu der Klasse, die weitgehend mit ‘Rheinzabern B’ übereinstimmt, und der kleine Rest von nur zwei Objekten (d.h. 4,3 %) einer sonst mit ‘Unbekannt 3’ identischen Klasse. Diese Verhältnisse illustriert Fig. 61.

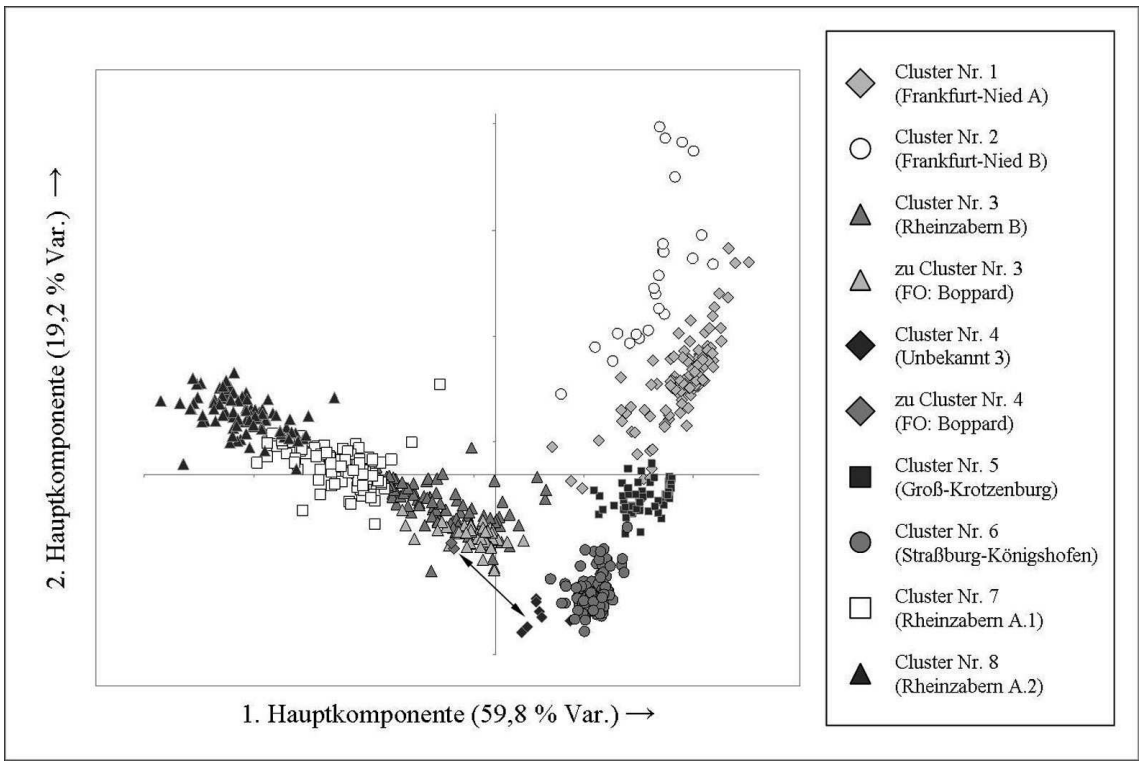


Figure 61: Auftragung der 1. und 2. Hauptkomponente mit Einfärbung der Objekte nach der Klassenzerlegung P8DANK (FO: Fundort)

P8DANK				
Cluster Nr.	Zuordnung zu Provenienz	Objektanzahl	davon	Abweichung gegenüber P8ModWARD
1	Frankfurt-Nied A	113		+ 21
2	Frankfurt-Nied B	24		– 14
3	Rheinzabern B, inkl. Worms	138	45 aus C_{Boppard}	– 6 + 5 = – 1
4	**Unbekannt 3**	9	2 aus C_{Boppard}	**± 0**
5	**Großkrotzenburg**	63		– 1
6	**Straßburg-Königshofen**	114		**± 0**
7	Rheinzabern A.1	110		– 5
8	**Rheinzabern A.2, inkl. Unbekannt 1**	89		**± 0**

Figure 62: Zur archäologischen Charakterisierung der Klassenzerlegung P8DANK

In der in Fig. 62 gegebenen Tabelle wird die Partition P8DANK hinsichtlich der Zuordnung ihrer Klassen zu den gemäß den Ausführungen in Section 3 erkannten Provenienzen von obergermanischen Heeresziegeleien charakterisiert. Das Klassifikationsresultat P8DANK rechtfertigt somit die Annahme von Produktionsstätten römischer Ziegel in Großkrotzenburg, Straßburg-Königshofen sowie in den noch nicht bestimmten Lokalitäten 'Unbekannt 1' und 'Unbekannt 3' in vollem Maße. Hier ist anzumerken, dass die eigenständige Existenz eines Ziegeleiortes 'Unbekannt 1' durch lokal-adaptive Clusteranalyse [2] sehr wahrscheinlich gemacht werden konnte, so dass 67 (d.h. 75,3 %) der 89 Objekte der Klasse 'Rheinzabern A.2' dieser Herstellungs-Provenienz zugeordnet werden können.

Auf analoge Weise ist in [6] das Bestehen von Heeresziegeleien in Worms bzw. deren Unterscheidbarkeit von solchen in Rheinzabern mathematisch-statistisch durch modellbasierte Clusteranalyse glaubhaft gemacht worden. So stammen aus der Gruppe 'Rheinzabern B' offensichtlich neben 19 weiteren die 45 erwähnten Ziegel der Menge C_{Boppard} aus Ziegeleien in Worms.

References

[1] DOLATA, J. (2000): *Römische Ziegelstempel aus Mainz und dem nördlichen Obergermanien - Archäologische und archäometrische Untersuchungen zu chronologischem und baugeschichtlichem Quellenmaterial.* Inauguraldissertation, Johann Wolfgang Goethe-Universität, Frankfurt/Main.

[2] DOLATA, J., MUCHA, H.-J. and BARTEL, H.-G. (2004): Eine Anwendung der hierarchischen Clusteranalyse: „Unbekannt 1" als Provenienz einer bekannten

Figure 63: Meßergebnisse der 660 römischen Ziegel: Die Oxide sind in der Maßeinheit % mit Skalenfaktor angegeben. *Fortsetzung folgt.*

	j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
	Scale	*10	*100	*10	*100	$*10^3$	*100	*10	*100	*100	ppm	ppm	ppm	ppm	ppm	ppm	ppm	ppm	ppm	ppm
i	Object																			
1	_2522	706	141	203	299	19	94	6	16	275	73	153	61	86	139	99	39	277	31	512
2	_2529	765	111	144	400	50	83	5	30	216	89	116	33	47	125	105	31	257	26	409
3	_2530	735	166	163	486	53	78	4	18	210	115	202	58	58	124	122	35	338	41	402
4	_328	736	158	160	454	25	103	7	18	221	93	142	38	42	132	140	41	316	35	417
5	_3661	620	62	144	511	97	251	120	55	261	81	102	55	98	127	309	27	139	12	428
6	_3662	598	65	145	472	87	268	142	66	262	76	81	70	80	118	295	26	143	17	524
7	_3663	588	65	148	495	96	283	144	69	272	80	115	56	98	133	341	30	147	14	418
8	_3664	625	69	162	515	77	254	92	69	291	98	133	65	113	150	244	28	146	14	406
9	_3665	650	68	157	479	60	244	75	66	297	88	124	63	106	147	215	31	156	15	416
10	_3666	628	62	140	477	88	251	118	69	266	95	120	53	108	133	335	29	164	13	401
11	_461	646	59	140	466	85	215	102	68	272	65	91	52	98	135	299	25	129	6	530
12	_462	673	63	152	478	60	194	62	62	275	108	107	60	111	146	253	29	135	1	543
13	_463	625	62	142	462	80	263	117	68	282	94	98	52	88	130	317	28	144	13	520
14	_464	654	64	150	436	45	245	83	69	291	95	108	58	104	149	224	27	124	13	467
15	_907	775	122	137	369	42	76	5	28	215	80	109	33	40	107	114	30	261	27	473
16	_908	673	101	206	660	21	114	2	42	263	123	136	42	69	175	95	49	285	26	381
17	_909	710	208	196	324	19	98	5	19	236	85	170	40	43	145	160	39	296	50	530
18	_910	727	177	161	511	89	88	6	16	243	74	194	71	69	113	136	39	336	40	665
19	_911	713	190	183	426	16	106	5	26	232	96	177	47	47	137	145	37	313	45	468
20	_912	764	166	147	339	24	81	8	19	199	71	152	32	32	112	131	38	335	38	442
21	F001	761	86	162	281	6	64	6	28	242	70	61	32	17	117	91	22	203	28	461
22	F002	769	86	154	296	6	57	6	26	234	69	64	32	18	119	92	24	213	27	456
23	F003	770	86	155	296	8	71	4	20	239	58	64	32	13	128	102	22	207	29	453
24	F004	764	84	159	311	7	83	3	19	239	84	62	34	21	136	103	22	219	22	471
25	F005	778	82	150	283	7	75	3	17	233	71	55	29	14	131	104	22	213	29	452
26	F006	772	85	153	300	7	80	2	19	232	85	64	32	16	132	105	21	207	30	463
27	F007	764	85	160	287	6	66	5	23	239	77	61	31	16	125	102	22	208	30	467
28	F008	750	92	169	322	7	82	4	23	252	61	57	33	19	135	95	23	219	30	447
29	F009	746	83	154	493	16	104	3	18	266	78	69	39	22	141	109	24	187	23	470
30	F010	777	81	151	269	7	58	5	23	233	65	55	32	16	120	95	20	195	29	469
31	F011	767	86	156	289	7	54	6	29	247	72	60	31	14	118	105	21	214	30	518
32	F012	766	88	158	295	7	79	3	22	235	74	64	34	15	137	107	24	215	23	474
33	F013	773	85	153	300	7	79	3	16	232	64	62	32	14	131	105	23	214	29	457
34	F014	686	89	183	700	17	148	4	21	300	89	86	46	38	162	111	27	166	21	462
35	F015	797	70	122	358	11	49	6	17	241	63	61	32	19	100	72	24	212	26	567
36	F016	805	68	118	321	11	50	6	16	240	50	47	32	17	96	71	25	211	24	572
37	F017	778	84	148	271	8	59	5	28	236	66	54	28	13	114	99	22	218	29	464
38	F018a	758	89	165	302	7	87	2	20	241	73	69	34	17	141	111	23	208	31	453
39	F018b	764	86	161	299	8	84	3	19	236	70	65	33	89	135	109	23	217	29	437
40	F019	769	87	155	303	7	79	2	17	237	84	62	32	19	134	107	24	217	31	456
41	F020	769	83	159	272	6	68	4	23	240	61	60	31	15	127	92	22	200	29	449
42	F021	768	84	160	276	7	79	2	18	232	69	65	30	15	135	107	22	213	30	448
43	F022	775	79	153	259	6	52	6	31	233	64	56	31	15	116	103	21	198	24	476
44	F023	769	86	157	288	7	78	3	22	237	67	61	30	15	133	99	22	212	30	440
45	F024	775	82	154	258	7	67	4	24	240	63	58	29	13	125	92	22	210	28	435
46	F025	774	82	153	283	7	64	4	24	234	69	59	31	17	120	89	21	209	29	428
47	F026	580	57	126	463	94	329	175	59	257	81	105	79	67	108	368	19	114	16	323
48	F027	593	57	128	464	98	299	163	61	256	70	99	75	67	127	359	19	116	16	327
49	F028	567	63	139	510	106	361	166	61	265	80	112	83	71	119	401	22	124	16	320
50	F029	570	62	138	513	105	333	166	62	264	88	111	82	72	121	381	22	126	17	317

Figure 64: Fortsetzung 1 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
51	F030	624	52	118	429	91	305	148	95	200	70	91	71	61	110	330	18	115	14	298
52	F031	619	80	138	519	100	252	126	47	258	82	87	69	56	117	323	24	191	20	322
53	F032	631	52	116	419	90	323	140	56	250	64	92	67	56	105	325	19	111	16	318
54	F033	590	60	134	483	104	258	148	57	288	82	86	77	76	116	390	19	123	15	405
55	F035	595	59	133	480	97	255	158	50	269	95	75	73	69	114	363	20	117	15	413
56	F036	650	49	109	395	84	258	139	56	237	66	87	65	54	101	304	18	103	14	314
57	F037	678	48	109	392	84	176	118	46	270	64	64	63	51	107	290	17	106	14	331
58	F038	567	57	127	463	96	268	190	53	291	69	84	73	65	113	386	21	119	15	353
59	F040	730	70	160	425	45	122	6	97	308	75	73	42	59	162	112	36	272	29	667
60	F041	638	73	170	528	66	262	65	78	315	91	108	72	104	159	226	29	135	19	417
61	F042	607	55	127	460	95	296	151	52	269	73	84	74	65	106	341	19	108	14	323
62	F043	777	84	150	275	8	73	3	25	233	65	60	31	15	127	96	23	217	28	429
63	F044	591	60	135	486	102	298	153	66	271	83	89	77	71	118	377	20	119	16	332
64	F045	629	51	117	419	89	273	146	54	249	68	82	68	59	109	352	17	110	14	377
65	F046	552	62	137	499	105	322	187	61	274	79	110	82	70	119	401	21	129	16	364
66	F047	560	65	146	538	113	337	158	69	278	81	124	83	79	126	432	22	135	17	421
67	F048	605	62	138	508	107	269	137	63	271	71	86	81	71	119	411	21	137	17	442
68	F049	631	54	119	424	91	249	143	58	255	64	84	68	66	108	374	17	130	14	484
69	F050	568	66	147	537	109	285	159	52	280	97	80	80	80	123	400	24	147	18	463
70	F051	552	60	133	486	100	298	198	54	256	80	88	76	69	112	397	20	136	16	342
71	F052	559	64	151	565	115	308	157	57	309	107	119	93	76	119	393	21	112	17	338
72	F053	593	59	131	475	98	291	157	57	260	78	87	75	67	112	392	20	130	15	347
73	F054	571	59	131	471	99	397	172	65	250	98	93	75	67	113	470	21	141	14	335
74	F055	575	60	135	503	100	246	174	51	265	98	92	78	69	118	474	21	130	17	435
75	F056	747	82	152	487	14	106	3	20	271	76	72	34	27	142	111	25	196	22	458
76	F057	749	67	141	457	40	105	7	106	289	76	75	39	52	141	104	33	295	26	564
77	F058	732	71	158	417	36	117	8	105	294	70	76	43	57	148	113	38	279	26	731
78	F059	702	73	182	458	45	144	8	93	302	83	83	47	63	167	123	41	242	26	859
79	F060	716	76	169	469	26	133	6	99	305	89	84	45	54	153	113	35	249	20	619
80	F061	733	70	151	516	20	86	10	103	272	69	78	37	48	133	146	27	281	18	913
81	F062	709	72	173	467	35	137	9	95	309	65	77	49	61	162	123	38	265	23	763
82	F063	740	67	155	397	39	117	6	97	301	84	74	41	68	158	114	36	272	25	729
83	F064	730	68	161	420	34	123	7	102	295	68	75	43	59	157	121	39	266	23	729
84	F065	735	67	159	407	69	117	7	93	292	73	75	44	55	154	115	37	255	26	797
85	F066	738	66	157	397	34	114	7	96	297	62	69	42	56	156	120	34	261	18	729
86	F067	723	70	166	445	24	129	7	99	292	70	79	49	55	155	110	37	273	18	678
87	F068	749	62	149	402	20	104	6	93	291	66	70	40	44	143	103	29	253	26	664
88	F069	730	59	130	346	17	81	45	103	345	65	58	35	34	133	175	25	264	17	561
89	F070	743	64	139	406	23	116	17	91	334	64	66	41	48	144	141	33	287	15	599
90	F072	598	57	135	526	99	223	148	77	269	67	68	55	170	112	335	19	123	16	381
91	F073	723	71	169	402	35	126	7	94	298	82	78	43	64	164	119	38	263	26	774
92	F074	736	71	155	430	20	121	6	99	300	71	76	40	81	149	109	30	278	21	578
93	F075	738	71	155	408	16	101	6	109	312	69	71	39	49	144	117	29	290	26	622
94	F076	692	135	182	497	43	152	12	36	297	96	114	50	67	158	157	39	264	41	506
95	F077	739	66	155	417	21	114	6	90	298	68	75	40	45	150	108	34	266	26	649
96	F078	762	64	137	412	22	76	8	98	274	64	64	37	47	131	149	29	272	17	556
97	F079	728	72	161	444	23	130	7	96	295	76	81	42	55	150	110	32	255	18	582
98	F080	734	70	151	465	37	126	8	103	294	72	83	46	60	149	107	36	272	24	597
99	F081	734	71	156	438	24	124	6	102	299	88	78	42	56	154	106	31	277	23	601
100	F082	775	83	139	353	19	76	8	17	243	44	56	29	31	128	140	23	223	25	442
101	F083	749	93	143	523	25	84	6	20	275	61	66	35	25	141	93	25	234	27	480
102	F084	682	61	154	447	46	228	53	70	292	88	93	65	87	151	166	22	117	16	427
103	F085	759	78	147	372	16	93	7	41	273	72	62	33	89	135	126	20	205	29	455
104	F086	743	88	170	367	12	100	4	26	249	86	71	34	94	148	118	21	177	32	443
105	F087	760	84	141	502	21	94	4	22	251	67	65	37	22	136	110	24	210	24	437
106	F088	729	96	176	404	10	113	4	25	260	85	82	37	28	160	138	23	177	32	478

Figure 65: Fortsetzung 2 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
107	F089	775	82	137	387	20	63	5	23	269	60	62	30	21	129	80	22	211	21	444
108	F090	709	82	165	546	23	117	11	68	296	86	74	40	23	146	159	23	165	22	442
109	F091	732	81	156	555	19	123	5	28	278	77	70	39	31	145	109	22	172	21	435
110	F092	732	80	163	468	18	122	8	28	257	83	69	38	56	148	194	21	162	21	439
111	F094	724	84	164	587	17	123	3	18	268	89	83	38	27	151	107	25	178	27	410
112	F095	693	87	177	699	15	148	4	25	293	102	89	51	40	161	111	29	169	26	417
113	F096	765	82	141	407	19	89	4	42	260	53	61	29	84	131	110	21	220	29	460
114	F097	755	86	159	355	11	94	4	22	253	72	64	34	22	144	116	21	179	31	438
115	F098	788	73	137	310	19	68	3	21	239	50	54	30	20	124	85	21	219	28	422
116	F099	775	73	126	423	19	77	9	32	268	53	57	31	18	121	122	24	261	25	450
117	F101	663	59	138	432	66	210	92	73	263	81	70	61	71	128	283	23	133	15	602
118	F102	642	66	157	471	61	244	82	76	294	82	88	63	95	149	248	26	126	17	447
119	F103	630	64	151	462	62	252	99	74	315	74	87	62	81	176	224	27	135	17	399
120	F104	645	66	155	480	81	255	81	79	291	93	95	64	86	144	247	25	132	17	489
121	F105	686	60	136	447	53	208	62	108	313	68	75	50	54	134	191	31	248	17	445
122	F106	681	65	144	482	73	234	56	94	292	64	85	56	63	141	190	33	264	17	503
123	F107	669	63	158	520	56	276	46	80	323	79	88	55	75	167	162	38	246	21	607
124	F108	692	62	144	471	58	233	46	94	302	74	88	53	66	152	157	34	254	23	550
125	F109	686	62	158	365	37	229	45	74	365	80	79	52	72	189	167	34	229	25	597
126	F110	659	65	165	379	43	245	62	79	359	76	84	53	68	189	189	35	224	19	596
127	F111	680	59	145	332	44	243	69	85	320	67	73	49	55	165	187	33	235	19	551
128	F112	597	60	134	486	103	289	150	60	262	74	90	75	68	112	394	21	129	14	471
129	F113	566	61	135	493	103	339	170	62	284	86	101	78	68	121	386	22	127	15	337
130	F114	606	60	141	533	117	279	132	53	250	101	85	56	66	100	357	24	129	17	552
131	F115	648	73	189	502	37	270	37	60	338	128	122	78	101	177	173	25	104	19	494
132	F116	649	53	131	397	63	213	118	71	259	66	71	53	70	124	281	52	121	13	426
133	F117	592	64	160	506	75	284	126	79	273	83	99	67	78	144	286	23	112	16	428
134	F118	619	60	149	447	64	259	115	73	304	92	94	61	77	137	264	24	119	15	465
135	F119	618	58	139	432	66	267	133	69	260	102	83	55	72	130	285	23	121	16	357
136	F120	726	187	174	454	12	83	5	22	204	107	146	45	23	122	127	37	294	47	384
137	F121	759	142	150	398	18	77	4	26	217	100	113	37	21	119	126	29	236	42	405
138	F122	721	185	175	465	13	88	5	34	206	116	147	44	25	128	150	37	288	44	408
139	F123	736	168	166	442	14	83	4	30	206	101	132	42	23	124	132	33	259	40	406
140	F124	729	162	174	408	14	91	4	20	237	119	122	37	15	130	131	30	249	39	393
141	F125	735	166	163	495	24	80	5	20	204	99	146	48	23	111	112	30	260	40	392
142	F126	739	170	161	471	21	58	7	20	193	94	140	46	19	106	95	31	259	40	424
143	F127	723	160	173	466	17	99	5	22	237	112	126	39	35	133	134	31	272	45	398
144	F128	734	168	163	515	26	76	5	16	199	90	139	47	23	109	111	30	253	43	395
145	F129	735	161	165	461	18	87	5	22	212	95	133	41	17	118	124	30	257	39	397
146	F130	731	156	168	441	17	95	5	19	236	92	124	41	20	129	130	30	262	45	398
147	F131	690	239	206	377	21	82	7	32	220	123	199	82	55	124	149	36	274	62	470
148	F132	732	172	169	446	12	80	4	24	215	103	125	40	19	122	132	31	265	47	410
149	F133	732	148	167	447	18	96	5	19	236	101	113	38	18	126	129	27	261	43	382
150	F134	695	187	195	368	20	117	10	21	287	108	129	40	67	157	156	33	245	49	439
151	F135	699	228	209	294	15	80	5	26	235	147	176	59	33	132	142	38	277	65	482
152	F136	727	129	161	436	51	112	13	31	260	82	104	47	54	128	156	33	242	34	514
153	F137	626	53	121	434	91	262	144	68	256	76	80	70	60	103	351	18	112	15	348
154	F138	614	55	127	447	96	278	147	60	260	76	93	70	71	108	380	19	113	14	307
155	F139	589	58	133	490	104	307	157	59	272	82	103	80	68	113	390	20	119	16	309
156	F140	612	54	119	445	95	275	159	57	246	81	94	73	60	101	349	19	117	14	318
157	F141	598	59	135	491	102	310	148	80	234	77	104	77	71	124	388	22	117	17	307
158	F142	579	58	130	477	97	300	169	51	281	78	84	78	77	113	459	21	127	16	413
159	F143	620	59	132	494	87	321	125	78	248	68	109	78	72	109	367	20	127	16	358
160	F144	619	55	122	445	96	272	149	54	247	70	99	71	62	104	339	19	112	15	348
161	F149	650	49	110	399	85	260	138	51	240	64	86	67	54	100	306	19	109	14	320
162	F150	598	59	130	466	98	275	156	57	270	83	95	75	70	117	406	21	135	15	431

Figure 66: Fortsetzung 3 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
163	F151	657	48	107	387	83	272	137	100	168	66	95	68	52	110	301	17	107	13	325
164	F152	587	58	134	490	100	302	157	59	273	77	110	80	75	125	377	20	115	16	312
165	F153	572	61	140	518	104	328	161	63	275	77	117	82	73	122	392	20	118	16	314
166	F154	721	175	181	341	19	101	5	25	266	128	123	40	20	146	151	32	241	49	426
167	F158	604	56	126	465	95	289	156	56	254	87	103	72	63	103	349	20	114	14	317
168	F160	608	65	119	468	96	280	160	56	239	75	101	70	58	105	342	19	128	17	338
169	F162	541	63	138	520	107	335	192	62	273	75	118	81	71	116	403	22	144	16	369
170	F164	713	214	181	393	16	91	10	29	220	105	172	53	41	122	158	32	284	57	423
171	F165	748	159	157	453	16	74	5	16	186	103	143	48	25	105	121	30	300	47	379
172	F166	763	146	143	423	26	75	6	27	201	88	90	38	22	119	134	28	219	34	371
173	F167	692	184	194	446	23	112	7	21	298	118	129	42	22	154	192	32	236	49	506
174	F172	701	52	123	362	47	229	68	130	287	58	70	47	50	114	298	27	219	13	587
175	F175	722	189	180	400	15	88	5	21	221	88	150	45	56	133	136	32	264	46	455
176	F181	762	91	147	364	15	97	5	22	281	84	64	40	30	140	113	29	211	22	452
177	F465	732	157	166	452	28	105	5	22	217	76	136	39	31	120	126	33	258	41	413
178	F466	729	164	172	385	35	107	6	17	244	80	129	35	26	136	142	28	223	40	434
179	F467	746	148	136	642	90	79	7	37	188	79	189	65	71	89	109	28	265	42	492
180	F468	712	74	167	529	34	145	5	83	298	74	101	45	63	146	105	35	243	20	722
181	G001	554	63	139	524	111	290	183	63	269	80	113	72	101	115	414	26	138	10	323
182	G002	741	173	161	456	21	82	4	22	197	87	199	55	61	114	107	35	321	37	378
183	G003	744	177	166	351	19	86	4	21	209	101	145	38	37	118	136	34	282	37	418
184	G004	713	165	178	502	21	113	5	21	235	111	154	35	36	139	148	32	241	34	374
185	G005	588	57	126	462	97	269	175	54	247	79	113	62	82	107	398	24	124	9	328
186	G006	700	202	164	739	72	78	9	25	173	119	249	89	83	82	138	34	296	49	455
187	G007	599	59	134	499	99	244	152	53	262	111	105	69	89	112	451	23	128	11	580
188	G008	746	159	157	419	25	73	11	20	184	92	160	51	53	105	148	34	322	33	392
189	G009	723	174	174	536	18	77	7	12	157	138	203	76	71	99	126	33	325	38	316
190	G010	728	150	178	393	24	94	5	30	213	89	157	51	57	130	125	35	283	29	373
191	G011	773	135	140	344	22	83	7	22	209	94	120	33	41	112	132	29	236	31	386
192	G012	721	193	177	425	16	87	6	21	215	91	171	43	48	121	151	37	280	41	497
193	G013	704	174	188	406	38	122	7	30	255	95	148	48	67	141	157	40	274	36	474
194	G014	732	174	170	443	27	86	5	13	196	96	180	55	62	118	101	35	306	35	346
195	G015	637	66	160	467	58	249	86	81	290	79	113	58	108	150	242	28	131	11	440
196	G016	736	136	145	610	57	83	11	15	217	92	118	43	40	110	156	31	298	24	383
197	G017	757	65	138	306	26	167	9	64	347	78	87	38	361	164	114	37	355	7	549
198	G018	761	62	130	307	21	169	13	62	339	60	89	37	316	148	119	37	345	7	563
199	G019	644	58	140	420	68	207	110	69	285	95	85	49	97	128	283	27	138	11	397
200	G020	628	58	134	437	79	224	124	79	296	92	82	48	88	118	362	28	150	11	389
201	G021	629	64	148	466	69	240	106	74	284	77	96	56	99	135	295	28	155	11	435
202	G022	726	199	155	614	81	86	4	13	210	91	226	82	68	95	152	33	311	44	501
203	G024	644	68	184	630	52	257	33	69	348	110	124	64	121	193	153	42	249	19	671
204	G025	655	63	148	467	74	234	83	69	283	89	109	56	107	143	221	28	138	11	405
205	G026	721	173	178	413	19	94	6	20	237	87	152	42	43	129	130	34	257	38	406
206	G027	643	168	251	411	17	118	9	25	241	98	174	76	57	134	128	34	243	37	437
207	G028	713	124	168	506	68	136	10	37	267	94	133	50	82	138	138	41	276	25	497
208	G029	709	210	179	518	24	86	5	15	218	94	193	54	57	120	130	36	306	47	426
209	G030	703	197	188	483	18	105	5	18	232	112	177	46	48	134	149	35	295	44	434
210	G031	714	185	183	457	15	94	4	14	227	101	160	43	42	123	142	34	284	40	423
211	G032	724	159	170	486	24	110	6	20	212	107	152	46	50	129	133	38	295	32	336
212	G033	693	192	190	377	24	109	8	22	363	132	151	34	42	144	97	36	275	40	354
213	G034	737	167	166	432	17	83	5	18	211	106	150	38	43	117	132	33	284	38	384
214	G035	759	150	131	578	105	66	6	40	178	69	197	65	98	86	122	30	336	35	460
215	G036	739	129	153	448	27	110	15	18	215	87	117	32	53	114	166	32	247	26	388
216	G037	693	153	195	527	42	93	7	22	218	108	156	54	56	129	142	35	288	30	392
217	G038	746	178	148	539	64	69	5	25	181	76	218	74	82	82	122	31	297	43	475
218	G039	737	71	154	438	24	125	6	97	296	82	90	37	82	146	109	31	304	11	586

Figure 67: Fortsetzung 4 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
219	G040	719	166	173	427	22	106	13	16	226	111	133	33	35	126	173	33	300	35	406
220	G041	576	60	129	492	99	271	177	63	266	66	125	66	91	110	385	25	138	12	380
221	G042	758	149	150	413	18	81	5	17	199	89	145	41	46	105	124	30	298	32	353
222	G043	728	68	160	445	58	108	10	91	292	72	91	41	73	138	157	38	280	11	854
223	G044	679	107	182	625	62	168	11	54	296	101	129	55	124	154	138	44	289	21	567
224	G045	723	146	185	393	21	79	6	23	216	94	146	51	43	108	106	27	227	35	420
225	G046	713	192	179	498	17	94	5	15	226	95	173	39	87	124	134	36	288	42	424
226	G047	613	61	143	524	95	239	123	46	308	116	120	75	97	125	346	24	121	12	358
227	G048	674	66	170	426	40	225	39	79	344	90	111	56	90	179	182	38	249	14	737
228	G049	596	57	129	499	99	250	156	46	306	104	104	66	94	118	331	21	126	12	312
229	G053	742	162	160	475	20	80	6	12	181	105	172	53	51	100	106	32	334	33	347
230	G056	698	150	186	475	67	135	8	34	274	100	136	46	82	143	144	42	290	32	516
231	G057	631	64	162	456	58	213	92	63	327	118	111	55	107	150	287	24	125	14	527
232	G059	603	58	133	483	100	265	147	53	285	85	102	64	91	116	449	23	133	13	409
233	G060	601	63	154	475	72	270	126	74	283	75	111	54	102	136	318	28	130	13	477
234	G061	729	134	187	272	18	97	5	19	255	77	143	54	66	130	117	31	274	29	455
235	G062	571	62	139	581	110	323	156	55	280	88	131	80	99	124	451	22	130	14	383
236	G063	811	99	120	323	22	60	7	17	114	63	104	66	87	78	67	47	320	16	256
237	G066	603	58	126	468	96	248	156	53	285	81	92	62	89	113	353	23	124	12	315
238	G067	711	168	182	479	18	101	5	20	239	84	144	40	40	130	140	31	271	36	387
239	G068	778	69	135	372	17	96	4	27	254	55	75	29	41	126	100	23	195	15	438
240	G069	588	60	131	502	104	302	159	55	276	87	133	70	92	112	391	22	134	12	332
241	G070	677	64	140	490	70	232	62	86	313	64	104	49	85	143	198	37	299	13	526
242	G071	660	62	152	525	57	247	63	81	308	72	107	49	97	155	190	38	270	15	561
243	G072	740	64	151	444	35	121	7	86	299	66	86	39	72	146	120	36	265	11	721
244	G073	710	199	183	470	17	94	6	21	208	94	183	54	55	122	143	37	314	45	437
245	G074	759	63	136	384	24	101	7	104	310	62	82	35	66	140	110	34	303	10	606
246	G075	760	63	138	373	23	104	6	101	311	52	79	33	60	143	107	34	311	10	602
247	G076	750	62	135	353	24	110	9	126	383	68	80	34	60	144	137	30	322	10	624
248	G077	721	176	180	346	20	110	6	22	262	103	144	35	61	143	164	35	265	39	440
249	G078	639	53	118	435	90	222	138	51	260	83	91	57	80	107	355	20	147	11	349
250	G080	757	65	131	341	26	166	12	68	342	75	96	41	277	152	103	39	342	9	545
251	G081	756	63	136	383	28	110	8	109	317	65	77	38	71	138	122	32	296	10	623
252	G083	730	229	182	355	18	68	5	18	153	91	208	79	66	85	125	34	304	56	456
253	G084	723	143	155	429	33	116	27	21	223	108	162	36	54	145	263	36	326	36	542
254	G085	749	158	156	404	25	84	6	18	212	81	169	50	50	116	111	35	314	31	387
255	G086	722	197	184	317	19	99	7	21	228	105	150	38	39	129	161	39	320	41	454
256	G087	759	64	131	336	24	168	11	62	343	77	88	40	273	153	100	38	342	7	552
257	G088	754	65	132	334	36	171	14	65	343	74	96	44	239	150	102	37	330	9	537
258	G089	755	64	131	324	32	173	13	63	346	63	97	42	253	150	104	35	326	9	531
259	G090	754	86	142	437	31	103	8	26	267	67	83	32	55	134	128	24	238	18	455
260	G091	756	63	136	380	34	95	8	110	323	61	85	37	63	139	124	33	312	8	644
261	G092	744	175	162	416	17	87	4	20	193	78	160	43	45	112	134	36	315	36	386
262	G093	686	72	124	402	67	199	92	47	240	76	116	57	71	107	265	22	151	16	376
263	G094	722	73	161	481	45	134	8	84	310	81	99	45	77	146	118	35	283	13	652
264	G095	761	84	149	292	10	74	11	21	223	56	83	31	59	118	288	25	241	15	600
265	G096	733	176	169	416	17	89	5	15	223	81	149	36	35	120	132	33	309	36	398
266	G097	749	166	159	403	19	83	4	11	208	82	136	32	34	114	133	33	306	36	383
267	G098	693	65	158	491	42	195	36	70	290	107	119	65	110	151	159	28	144	12	410
268	G099	714	202	183	448	19	90	5	16	215	125	181	52	47	120	143	35	298	46	440
269	G100	722	169	176	457	23	89	7	17	208	88	177	57	59	118	115	35	307	37	395
270	G101	590	58	126	473	100	286	168	54	262	76	116	61	88	108	394	22	129	11	454
271	G102	718	84	163	595	20	124	5	23	299	77	103	41	53	141	91	26	196	18	428
272	G103	757	82	152	405	16	104	3	22	252	53	85	32	55	137	112	24	213	18	423
273	G104	736	85	160	502	18	123	3	19	268	86	95	36	47	149	114	24	200	19	450
274	G105	753	81	155	399	15	106	4	24	253	79	85	32	42	144	117	24	205	19	422

Figure 68: Fortsetzung 5 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
275	G106	605	55	123	452	93	252	162	52	273	78	98	60	86	106	381	23	141	11	359
276	G107	641	65	160	470	49	248	83	71	288	109	119	57	110	154	224	27	129	13	416
277	G108	638	56	115	438	93	246	141	47	250	69	94	56	77	104	348	21	135	12	341
278	G109	735	71	153	444	22	125	7	94	304	76	94	37	76	147	115	32	300	11	571
279	G110	738	67	155	409	18	113	9	89	306	68	88	31	65	145	127	31	283	13	649
280	G111	583	59	132	493	102	290	165	59	281	88	121	64	90	116	383	23	121	11	308
281	G112	753	83	150	439	21	109	4	23	262	59	89	33	43	138	114	24	214	18	452
282	G113	600	69	160	476	57	295	118	71	287	109	123	58	117	134	428	28	136	14	392
283	G114	664	89	190	740	34	127	11	46	336	134	149	54	106	174	145	34	217	16	569
284	G115	720	177	170	464	17	91	11	15	222	99	155	41	46	121	169	33	271	39	414
285	G116	716	177	174	482	56	89	7	20	229	92	157	42	88	124	175	35	291	39	519
286	G117	571	62	138	518	104	258	169	54	288	85	111	67	94	120	423	25	137	14	395
287	G118	574	59	128	484	105	291	179	57	271	84	114	65	89	112	421	25	139	14	466
288	G119	625	63	153	476	69	244	108	63	287	81	95	55	107	138	239	29	129	14	394
289	G120	738	155	163	412	20	95	7	14	239	75	135	33	38	125	136	33	297	33	408
290	G121	694	70	144	435	70	186	53	68	300	78	102	48	78	147	177	36	278	13	457
291	G122	763	162	149	372	19	79	5	20	195	86	159	37	42	109	122	33	293	35	399
292	G123	715	186	180	456	15	104	6	25	211	99	161	39	42	130	152	36	304	39	415
293	G124	602	62	151	456	76	225	131	55	332	84	106	54	94	144	325	24	121	14	442
294	G125	703	61	148	471	43	191	40	68	282	96	113	58	103	144	162	26	144	13	450
295	G126	593	67	159	516	91	253	124	60	312	98	108	57	117	142	314	28	132	15	502
296	G127	608	69	166	522	58	261	101	62	310	93	116	60	121	151	256	29	135	14	537
297	G128	632	105	216	730	70	172	10	74	295	130	141	77	155	120	112	31	197	17	502
298	G129	730	166	173	378	20	105	5	17	239	114	145	35	38	135	153	35	282	35	404
299	G130	671	82	165	528	67	215	41	61	327	107	110	60	90	160	158	38	254	15	636
300	G131	722	57	133	367	57	167	44	93	296	71	90	44	89	146	163	33	252	11	566
301	G132	677	65	160	526	60	221	36	79	328	78	116	55	120	168	165	41	286	15	702
302	G133	664	66	150	518	70	261	62	93	287	96	113	50	89	146	186	38	290	15	534
303	G134	688	62	140	465	56	236	56	92	289	86	110	47	86	141	175	36	307	13	481
304	G135	609	71	171	568	82	270	88	70	320	90	122	63	123	159	276	29	132	15	454
305	G136	597	69	162	507	82	305	114	63	299	86	115	57	114	145	279	28	132	15	580
306	G139	644	49	127	438	82	180	123	69	275	93	107	65	81	108	397	21	109	11	420
307	G141	616	71	160	576	111	208	86	126	367	112	130	77	101	162	327	33	220	17	498
308	G142	752	62	140	385	26	104	8	104	330	81	75	35	70	141	127	32	300	8	608
309	G144	620	61	138	492	72	220	125	68	287	78	102	51	95	124	363	27	149	15	457
310	G145	584	66	146	505	93	262	149	82	268	82	112	54	102	127	380	29	147	16	434
311	G146	600	68	156	493	64	249	124	70	304	90	109	54	119	143	295	29	143	14	430
312	G147	596	65	142	492	113	247	139	77	279	72	97	54	102	127	410	30	165	12	622
313	G148	703	60	133	437	59	184	49	94	357	65	98	44	84	141	169	35	316	13	533
314	G149	617	66	160	545	82	260	96	62	316	83	113	59	112	156	264	28	126	15	442
315	G150	626	64	157	514	84	237	95	57	317	87	102	55	111	147	261	28	128	13	405
316	G151	589	65	157	516	91	264	131	65	302	97	108	55	133	144	338	26	125	13	490
317	G152	702	61	149	467	43	189	40	70	281	96	112	59	101	143	164	28	144	12	427
318	G153	615	68	167	531	79	254	93	56	326	96	118	62	127	158	258	29	125	14	406
319	G154	619	66	160	519	81	246	98	60	319	88	108	58	110	148	283	28	129	13	469
320	G155	599	62	155	496	83	221	130	54	301	79	109	56	98	137	307	26	124	14	467
321	G156	613	69	168	557	77	276	86	86	326	97	123	64	116	156	264	28	126	14	425
322	G157	586	66	158	529	95	288	129	60	310	95	108	57	109	146	334	28	135	16	422
323	G158	624	65	159	515	75	250	95	57	310	104	101	57	114	147	253	26	122	12	429
324	G159	602	69	167	551	85	261	101	59	328	88	111	63	118	155	276	29	126	15	426
325	G160	719	217	171	560	53	66	4	21	174	106	269	93	74	78	134	32	308	54	534
326	G161	682	205	179	763	56	119	6	25	195	111	265	93	86	106	170	35	310	48	453
327	G162	708	192	165	658	65	101	6	40	200	119	243	87	89	92	145	36	320	44	504
328	G163	747	203	150	545	54	56	4	18	146	116	255	87	80	72	126	32	327	48	459
329	G165	746	160	142	578	72	72	6	36	185	107	208	75	84	82	130	32	300	38	478
330	G166	743	162	142	589	98	73	7	38	187	100	200	75	103	87	132	33	307	39	473

Figure 69: Fortsetzung 6 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
331	G167	778	120	113	477	136	64	14	42	182	67	165	59	95	82	130	30	300	29	459
332	G168	743	145	154	371	32	117	13	22	234	88	129	33	38	121	146	30	236	31	379
333	G169	692	161	190	553	16	123	5	19	262	117	152	37	42	146	149	32	271	33	404
334	G170	718	158	153	586	101	121	10	16	281	77	168	79	82	129	174	41	293	34	507
335	G171	746	128	146	526	44	102	4	16	249	75	132	53	63	125	146	38	306	23	415
336	G172	717	173	164	554	78	113	4	22	267	97	192	72	78	123	151	42	298	38	443
337	G173	735	185	172	371	15	89	5	13	215	81	156	37	39	118	139	36	297	40	384
338	G174	744	175	162	351	18	64	8	15	240	83	138	35	36	103	137	35	299	37	395
339	G175	726	189	154	607	74	93	6	16	217	89	218	82	75	99	157	34	316	44	469
340	G176	728	191	176	381	15	90	5	19	212	113	159	41	42	122	143	35	302	41	369
341	G177	739	179	167	361	15	92	6	19	216	103	148	32	35	117	145	35	298	37	380
342	G178	698	230	203	338	18	98	5	17	244	97	192	46	55	138	167	41	307	52	476
343	G179	732	73	153	462	40	136	8	94	292	76	96	43	81	147	112	38	309	12	622
344	G180	737	66	155	411	35	121	8	93	301	74	85	37	77	148	121	38	275	11	736
345	H125	726	171	173	434	19	93	5	29	216	86	142	31	30	127	144	34	276	39	405
346	H126	570	58	132	482	98	296	178	62	271	73	117	63	84	116	400	26	126	14	295
347	H128	631	64	150	504	85	216	95	84	302	72	93	51	110	140	307	29	148	16	509
348	H129	729	69	163	410	34	127	6	102	308	78	89	34	76	162	120	41	286	15	764
349	H130	738	64	154	395	45	110	8	108	312	65	84	33	63	148	122	39	268	13	708
350	H131	622	53	121	436	95	273	147	65	250	67	108	55	73	104	420	21	115	13	308
351	H132	736	172	164	448	29	79	5	23	217	81	196	46	47	116	96	37	346	39	386
352	H133	731	86	164	484	17	122	5	27	275	67	93	32	37	144	103	29	202	18	412
353	H134	634	64	154	487	82	250	91	88	292	103	98	51	100	138	292	28	138	15	620
354	H135	728	71	154	482	42	117	8	114	299	62	92	39	70	138	120	41	313	13	737
355	H136	770	75	142	377	16	79	5	35	256	65	74	23	26	111	90	25	226	17	451
356	H137	635	64	155	477	67	229	93	80	301	74	96	49	98	141	245	29	142	15	441
357	H138	660	65	150	451	54	236	77	78	280	76	110	52	89	136	181	30	140	13	420
358	H139	652	69	167	468	48	240	64	69	311	86	115	55	104	151	178	32	138	16	439
359	H140	669	65	157	443	47	230	61	67	306	87	107	52	94	144	182	29	138	13	453
360	H141	636	67	160	504	63	254	80	80	302	93	117	53	102	148	206	30	135	15	443
361	H142	638	65	153	465	60	251	92	87	283	90	113	50	94	141	217	29	144	15	404
362	H143	669	65	155	463	55	206	64	72	287	99	95	51	97	134	192	30	143	14	681
363	H144	654	63	152	448	66	242	82	64	283	93	90	50	98	125	182	29	136	14	594
364	H145	676	65	153	449	51	244	57	68	285	93	80	51	94	130	158	30	146	16	469
365	H146	660	63	150	444	54	198	82	71	279	91	92	48	92	131	206	28	142	15	585
366	H147	649	66	163	491	51	248	66	88	308	89	114	57	103	152	190	30	134	17	489
367	H148	655	68	167	503	56	240	54	74	320	89	112	58	104	153	185	31	137	17	523
368	H149	681	59	140	434	63	255	67	71	279	85	73	47	84	121	162	27	134	14	435
369	H150	678	61	141	444	63	217	71	80	269	70	104	48	85	130	194	27	137	14	405
370	H151	687	59	140	416	56	206	66	79	286	67	94	44	84	134	186	27	143	14	411
371	H152	685	61	147	441	61	227	55	75	292	64	88	49	90	137	180	27	144	14	448
372	H153	643	65	159	474	76	233	77	64	298	105	95	55	112	137	215	30	133	17	497
373	H154	701	59	142	422	58	229	47	74	290	59	89	49	90	132	179	28	146	12	449
374	H155	689	56	133	398	61	204	75	72	281	55	83	41	79	125	190	25	146	14	423
375	H156	677	58	133	422	77	204	84	74	271	52	87	42	76	121	211	27	144	13	484
376	H157	701	60	142	423	59	272	43	70	284	80	84	49	87	126	136	26	138	13	559
377	H158	686	61	147	441	58	222	57	83	278	67	96	49	101	138	191	28	138	14	432
378	H159	661	72	171	508	64	239	44	86	305	85	122	59	114	161	166	32	146	16	436
379	H160	644	60	151	432	81	199	94	72	293	75	101	49	85	139	268	26	122	14	524
380	H161	707	82	161	571	27	136	20	23	294	76	91	32	57	145	123	28	202	20	414
381	H162	652	55	134	411	68	230	99	97	329	78	87	42	86	130	235	27	127	12	395
382	H163	689	68	178	452	38	212	17	66	329	120	124	54	126	170	137	29	125	26	475
383	H164	597	58	142	471	95	230	144	70	304	98	103	45	88	120	283	27	121	16	483
384	H165	654	67	157	478	54	240	71	78	298	76	109	49	110	144	195	31	151	14	435
385	H166	659	66	154	470	61	232	70	80	305	74	106	50	112	142	190	31	156	16	423
386	H167	614	62	144	442	75	246	130	77	263	74	89	45	84	122	271	30	147	15	467

Figure 70: Fortsetzung 7 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
387	H168	671	61	145	455	84	253	69	71	275	91	85	50	88	120	184	29	149	13	445
388	H169	643	58	142	475	72	223	103	78	267	75	94	17	97	128	247	27	127	63	395
389	H170	647	62	156	474	56	200	84	69	293	107	90	53	95	129	190	28	127	16	548
390	H171	633	68	170	499	57	224	77	77	316	117	106	55	107	152	190	29	132	16	514
391	H172	650	66	162	489	58	202	70	71	309	82	89	58	98	137	181	30	131	15	575
392	H173	635	62	141	443	73	227	113	79	258	93	92	46	92	118	257	30	158	14	564
393	H174	653	67	160	485	67	231	63	75	333	88	107	52	109	150	200	30	149	16	448
394	H175	612	71	175	550	82	299	78	91	321	77	127	60	118	159	268	30	134	17	466
395	H176	612	60	144	474	87	222	132	61	270	94	80	47	86	125	304	28	123	15	407
396	H177	616	70	170	552	103	282	82	84	300	62	113	57	98	146	277	30	134	15	473
397	H178	627	70	166	520	72	262	82	72	301	69	109	56	108	148	247	29	137	13	438
398	H179	587	64	148	517	100	252	147	68	260	64	98	49	87	129	297	28	136	17	421
399	H180	632	67	150	479	74	198	104	76	295	84	100	45	88	138	282	29	184	15	462
400	H181	586	64	140	520	114	260	153	77	256	72	87	46	84	118	331	31	158	16	411
401	H182	596	68	162	544	103	275	114	76	287	99	112	51	107	146	346	29	131	16	438
402	H183	615	70	166	531	66	277	90	77	309	98	113	56	104	154	264	31	138	16	405
403	H184	638	65	157	458	54	244	90	70	290	69	102	52	103	145	234	27	128	15	411
404	H187	632	65	149	508	99	217	102	74	268	75	94	47	87	129	242	29	153	15	438
405	H188	638	67	160	496	52	244	82	78	290	86	112	54	152	148	210	29	135	15	460
406	H189	577	66	150	509	82	275	152	69	263	72	95	47	87	124	314	30	139	14	428
407	H190	673	66	149	446	76	236	64	84	274	51	95	51	90	134	211	31	166	14	420
408	H191	647	59	136	435	64	218	106	91	273	70	100	44	75	120	275	26	165	13	414
409	H192	572	65	150	501	87	253	155	78	289	84	102	45	84	131	362	29	146	15	462
410	H193	630	70	164	490	62	272	80	76	311	89	120	55	114	152	229	31	140	16	450
411	H194	662	66	152	559	91	219	61	81	283	71	107	54	90	142	233	30	151	15	484
412	H195	569	63	140	522	112	228	172	74	270	80	74	42	83	117	357	28	146	14	402
413	H196	634	65	143	478	94	303	101	95	253	70	104	53	96	121	296	33	172	15	364
414	H197	596	68	162	515	73	276	115	69	311	94	116	53	107	150	248	31	125	16	453
415	H198	769	90	148	352	16	91	3	22	247	48	78	23	31	136	114	28	261	22	438
416	H199	596	60	128	472	101	247	160	59	260	99	81	51	82	113	445	25	159	14	346
417	H200	704	53	123	396	64	183	75	70	252	69	73	39	75	108	198	26	142	13	470
418	H201	742	66	150	415	39	120	7	103	301	73	79	35	61	149	115	41	288	13	711
419	H202	738	66	151	409	46	123	9	102	301	72	84	33	63	152	120	39	296	12	686
420	H203	739	66	152	417	43	121	7	100	303	62	80	35	64	151	113	41	293	12	651
421	H204	742	65	150	423	40	122	5	100	302	70	86	34	64	150	106	39	297	12	708
422	H205	733	65	149	418	48	124	15	103	295	57	83	33	63	147	122	39	276	10	689
423	H206	741	65	150	431	60	92	9	105	299	67	85	35	62	145	103	37	290	12	748
424	H207	737	67	151	441	46	123	7	102	302	75	81	34	68	148	117	41	290	12	757
425	H208	740	62	142	401	46	120	19	104	289	58	76	30	60	138	126	38	273	11	659
426	H209	738	69	151	447	38	120	7	105	283	68	90	34	62	140	108	41	305	11	650
427	H210	741	69	150	441	43	120	7	106	277	81	96	36	63	142	110	42	320	12	686
428	H211	737	69	151	448	40	122	7	112	283	68	92	35	63	143	112	38	312	11	678
429	H212	740	68	148	437	43	117	10	106	281	51	88	34	59	133	131	39	327	12	678
430	H213	748	61	139	390	34	109	16	102	301	59	81	30	58	136	117	35	285	12	661
431	H214	753	63	142	414	30	110	5	96	307	72	78	33	56	141	98	38	310	12	626
432	H215	736	65	151	418	37	120	8	104	335	68	84	36	72	155	113	39	297	11	700
433	H216	759	60	137	395	28	96	7	104	299	58	84	32	51	134	92	38	292	11	633
434	H217	737	67	154	418	37	99	9	104	306	65	88	33	75	144	111	35	292	11	767
435	H218	736	67	152	439	55	124	6	104	309	65	82	35	64	150	111	40	294	12	663
436	H219	729	72	158	451	38	126	9	87	295	72	95	36	62	143	140	43	324	12	712
437	H220	734	71	155	438	36	125	7	98	298	82	92	34	63	143	112	42	310	14	714
438	H221	742	63	147	417	36	117	9	85	331	67	79	33	65	146	126	38	287	12	705
439	H222	729	68	157	433	41	127	8	85	333	80	94	36	71	151	128	41	301	12	746
440	H223	744	63	146	424	42	119	6	83	333	65	80	34	66	143	110	38	293	13	722
441	H224	738	64	146	449	34	113	11	97	308	59	86	36	62	145	113	38	289	12	655
442	H225	743	64	145	453	38	120	5	91	322	73	84	33	63	145	101	35	294	12	590

Figure 71: Fortsetzung 8 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
443	H226	741	64	148	412	40	117	8	99	326	58	84	34	61	143	120	40	290	13	749
444	H227	749	62	146	393	40	116	6	98	302	62	80	30	61	147	107	38	272	13	669
445	H228	736	64	155	424	40	121	6	99	314	60	82	37	71	152	108	41	287	12	733
446	H229	734	68	157	420	44	124	7	103	299	76	90	35	66	151	113	41	300	10	726
447	H230	736	65	153	414	37	119	6	95	332	71	82	34	74	153	110	41	290	12	725
448	H231	735	67	155	435	40	123	7	103	299	82	86	35	64	152	117	39	296	13	737
449	H232	747	62	144	409	37	114	9	88	313	62	79	32	67	140	115	39	299	12	677
450	H233	737	67	148	439	37	113	7	104	332	55	82	36	64	136	114	40	314	12	739
451	H234	751	62	141	419	41	109	7	105	302	56	83	33	54	142	102	37	301	10	642
452	H235	751	60	138	401	31	106	5	95	387	58	73	30	50	133	99	37	304	12	603
453	H236	758	60	136	401	50	106	6	110	309	54	73	33	52	139	104	34	301	11	621
454	H237	745	65	146	429	33	117	6	106	300	71	83	34	58	143	99	38	315	12	648
455	H238	740	65	149	430	49	107	8	107	306	63	84	35	62	139	107	39	295	12	696
456	H239	732	65	149	415	41	123	11	85	338	69	91	36	61	149	126	39	281	12	729
457	H240	750	63	142	411	30	112	6	103	308	66	84	33	65	143	106	36	298	13	659
458	H241	757	66	136	443	27	107	6	101	276	57	88	31	58	134	92	32	321	15	526
459	H242	745	62	144	415	35	126	7	90	342	68	81	33	60	147	114	37	294	12	645
460	H243	736	65	153	418	38	121	9	98	301	72	80	33	65	150	119	39	292	11	703
461	H244	727	68	155	443	56	125	13	98	305	72	86	36	78	148	120	40	290	13	691
462	H245	738	63	145	427	45	126	11	85	341	66	86	34	68	142	125	38	286	12	672
463	H246	726	72	161	437	41	126	6	104	294	76	93	36	68	151	112	40	300	14	777
464	H247	732	67	156	433	43	126	6	88	324	73	87	34	71	151	109	41	295	13	761
465	H248	728	68	154	430	41	98	10	121	354	66	86	36	60	160	123	38	292	13	914
466	H249	765	132	130	514	55	85	6	52	190	66	164	53	67	88	93	35	363	33	354
467	H250	695	121	160	498	88	132	35	69	255	91	138	49	83	124	146	41	347	26	527
468	H251	775	108	139	359	35	80	5	44	206	82	128	36	58	113	89	31	341	23	396
469	H252	645	66	165	480	54	219	73	77	297	92	115	54	96	156	220	27	128	17	463
470	H253	745	210	153	536	50	50	3	27	152	84	249	74	57	71	113	32	325	52	429
471	H254	699	181	168	720	53	117	6	33	205	84	225	69	72	106	156	32	287	45	447
472	H255	657	59	121	410	79	200	117	58	249	66	114	53	66	108	311	25	139	15	321
473	H256	725	177	173	412	19	99	6	27	230	102	141	27	28	123	143	36	308	37	398
474	H257	691	176	196	426	30	136	6	31	293	81	144	35	52	158	169	40	277	40	461
475	H258	748	167	158	404	24	80	5	23	208	96	142	31	32	116	130	33	289	39	395
476	H259	755	160	152	390	25	78	6	23	208	92	142	33	31	112	127	34	280	37	387
477	H260	734	69	153	459	33	125	7	104	292	69	90	35	61	145	110	40	310	12	678
478	H261	747	62	146	410	35	106	8	101	299	56	85	33	59	140	101	38	290	11	731
479	H262	755	61	138	402	27	105	7	98	317	64	79	30	51	140	107	35	305	13	664
480	H263	739	68	155	389	26	115	6	106	303	75	86	30	67	147	112	38	293	13	681
481	H264	731	68	154	432	45	129	7	107	324	73	87	36	68	149	116	42	290	14	779
482	H265	735	67	157	391	33	114	9	105	297	68	87	32	66	149	110	38	282	14	716
483	H266	755	61	140	407	30	107	6	108	303	58	77	32	54	140	101	36	306	12	613
484	H267	743	63	147	400	38	119	8	106	322	55	84	31	65	143	157	40	294	11	684
485	H268	740	63	146	405	39	117	10	116	323	73	80	35	63	134	149	41	280	11	870
486	H269	645	251	202	964	122	84	3	30	152	112	323	121	81	74	112	31	338	60	356
487	H270	651	252	191	976	105	61	8	25	142	114	317	112	70	75	111	32	331	61	443
488	H271	654	251	197	914	98	67	7	23	145	107	316	116	69	69	88	33	333	61	540
489	H272	702	200	171	706	82	68	6	24	172	87	255	105	71	81	147	31	336	49	486
490	H273	645	249	199	997	135	77	5	19	148	90	320	123	69	74	123	31	336	61	429
491	H274	675	235	187	832	104	74	4	20	153	85	303	110	70	76	123	31	341	56	461
492	H275	763	93	140	354	49	84	7	89	274	81	114	39	65	118	99	43	421	23	449
493	H276	595	101	220	709	88	231	39	79	313	135	134	67	99	129	212	32	175	19	427
494	H277	766	88	147	336	21	105	7	18	253	63	77	22	42	133	121	26	255	19	449
495	H278	727	167	149	651	59	70	6	24	221	71	196	56	54	105	140	33	350	40	798
496	H279	664	61	112	422	92	193	123	71	219	79	112	58	66	91	279	29	237	13	535
497	H281	757	126	143	426	32	94	7	28	241	68	120	30	50	120	121	32	259	29	528
498	H282	709	178	180	518	22	80	6	33	213	67	157	37	41	112	109	39	325	40	430

Figure 72: Fortsetzung 9 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
499	H283	774	143	142	321	27	63	6	43	198	61	169	42	48	90	77	33	343	32	439
500	H284	763	156	148	375	28	77	5	22	200	69	127	24	29	109	127	33	297	35	411
501	H285	730	173	171	375	24	86	8	22	215	86	138	27	32	124	155	33	274	38	574
502	H286	724	173	173	473	19	92	5	23	206	75	143	34	36	117	123	37	306	40	406
503	H287	734	189	171	400	21	78	5	25	188	73	161	43	43	106	114	34	309	44	425
504	H288	719	185	180	451	16	89	5	23	197	72	152	41	37	112	111	35	330	41	422
505	H289	695	177	189	561	18	107	5	26	228	93	149	38	42	131	132	34	292	40	409
506	H290	622	67	123	456	116	281	141	89	221	69	76	37	70	92	265	34	231	14	350
507	H291	718	151	172	501	17	109	6	30	246	86	141	29	30	135	136	33	281	34	410
508	H292	579	67	166	504	66	253	102	67	607	54	114	54	99	135	291	27	119	15	405
509	H293	648	59	136	435	76	225	109	78	261	63	86	44	82	117	229	29	156	13	410
510	H294	608	67	158	516	73	247	112	71	295	95	96	51	101	142	263	28	138	15	393
511	H295	589	62	158	477	77	253	127	72	358	81	92	46	93	145	287	27	118	15	731
512	H296	605	69	161	515	64	284	106	71	311	86	103	53	111	148	261	31	133	17	395
513	H297	629	68	155	473	54	270	96	76	291	86	112	52	103	143	241	29	142	15	384
514	H298	635	72	176	529	47	263	59	79	334	115	126	62	116	169	181	29	137	18	448
515	H299	632	63	152	489	83	226	96	71	308	69	104	46	99	145	257	28	128	16	429
516	H300	558	66	158	602	102	266	146	75	312	93	116	50	95	138	377	27	125	16	539
517	H301	661	95	168	564	46	297	10	68	554	107	145	93	62	134	84	34	178	16	415
518	H302	600	65	154	517	82	250	126	65	288	84	89	50	92	130	296	30	137	15	407
519	H303	586	66	157	514	106	272	136	77	269	111	100	50	97	143	320	27	133	15	385
520	H304	612	43	90	338	119	245	193	95	290	62	63	26	42	84	472	24	169	10	354
521	H305	683	68	164	478	54	209	38	72	305	95	106	55	102	149	167	28	143	14	448
522	H306	640	66	155	495	66	225	90	66	280	79	107	52	100	142	220	27	135	14	381
523	H307	614	71	172	533	71	235	92	62	302	124	105	58	112	150	259	31	129	17	474
524	H308	628	74	178	524	52	261	69	74	302	117	132	61	113	163	213	32	132	17	411
525	H309	615	66	155	506	67	241	111	70	282	103	83	50	94	137	277	29	133	16	403
526	H310	566	64	148	542	116	256	160	80	268	97	99	46	97	127	357	28	134	15	411
527	H311	609	61	150	461	73	267	125	80	277	84	99	48	87	134	290	27	129	15	389
528	H312	619	67	165	523	73	256	92	82	297	90	111	54	99	147	226	30	132	15	419
529	H313	602	67	159	514	75	261	111	72	315	73	96	52	115	145	290	29	133	16	389
530	H314	579	64	148	502	81	257	150	74	286	67	103	46	108	134	329	28	133	15	372
531	H315	617	67	156	510	69	232	105	66	309	86	100	53	105	142	251	30	137	14	409
532	H316	633	66	165	516	56	238	75	70	343	82	105	56	108	154	210	30	123	16	467
533	H317	605	65	156	511	74	252	114	69	309	72	101	51	106	142	277	29	133	15	429
534	H318	600	66	151	491	71	290	125	79	278	82	109	48	98	136	294	30	148	16	361
535	H319	615	66	154	467	54	279	111	79	283	80	113	52	100	139	255	29	142	16	368
536	H320	592	67	154	506	75	276	131	75	278	87	108	49	98	138	306	29	140	16	382
537	H321	576	65	149	505	83	277	151	76	282	102	100	47	100	136	327	29	137	16	348
538	H322	585	66	151	511	78	265	140	77	279	83	110	49	101	136	307	28	142	15	373
539	H323	633	72	177	546	64	230	66	68	303	120	123	63	106	162	194	29	137	17	438
540	H324	637	68	168	480	43	269	72	82	309	99	123	56	120	159	208	30	139	16	414
541	H325	625	67	160	495	62	260	95	77	288	104	117	54	105	146	242	28	142	16	391
542	H326	599	70	173	525	80	264	103	69	296	126	125	57	115	164	280	27	132	17	415
543	H327	588	57	127	478	119	236	171	72	254	63	64	39	77	109	353	27	143	15	347
544	H328	543	62	149	556	126	269	179	65	293	92	102	45	90	132	374	27	124	16	406
545	H329	588	65	152	513	92	268	137	74	278	85	106	49	101	133	309	28	142	15	393
546	H330	595	57	129	473	119	259	155	68	281	72	81	41	88	124	353	26	136	13	374
547	H332	602	70	171	539	86	261	99	66	320	84	121	57	112	157	245	31	135	17	413
548	H333	591	66	154	524	104	243	131	69	301	93	90	51	107	138	312	27	141	15	408
549	H334	604	70	170	538	85	260	97	65	324	84	115	56	112	158	240	31	133	18	423
550	H335	585	59	129	498	138	223	163	70	296	72	72	39	92	122	369	29	147	13	397
551	H336	579	67	154	528	96	259	138	82	304	87	91	49	105	136	307	29	142	16	424
552	H337	607	63	144	471	88	239	130	73	301	70	85	46	108	134	282	30	145	16	382
553	H338	585	62	143	521	92	228	143	75	321	93	84	49	116	136	297	30	138	15	403
554	H339	557	64	152	534	115	225	168	66	311	101	89	46	91	135	360	28	126	16	404

Figure 73: Fortsetzung 10 von Fig. 63. *Fortsetzung folgt.*

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
555	H340	554	61	146	511	108	249	175	69	317	79	85	44	87	129	376	27	123	16	403
556	H341	603	64	152	544	104	238	113	88	340	95	88	51	104	140	235	30	140	15	454
557	H342	581	68	159	530	98	242	130	79	341	116	91	51	103	147	265	30	139	17	406
558	H343	538	61	144	533	109	276	187	65	306	72	94	42	98	130	404	26	126	15	418
559	H344	605	69	160	532	86	244	104	75	350	109	94	54	113	146	225	30	140	16	419
560	H345	544	63	149	503	110	258	184	63	308	82	87	41	90	136	395	26	121	15	407
561	H346	523	63	144	528	112	254	208	65	300	85	73	40	90	126	427	26	128	18	373
562	H347	530	62	148	518	114	276	195	62	313	89	94	41	87	134	422	25	119	15	432
563	H348	546	66	150	712	126	253	160	65	307	88	101	49	92	137	343	27	127	17	385
564	H349	567	66	148	537	95	236	157	69	331	109	80	49	92	136	297	30	136	16	355
565	H350	511	62	149	520	119	260	215	67	295	72	89	40	85	129	444	24	116	16	434
566	H351	600	66	163	518	81	242	111	66	334	120	106	53	116	148	228	28	117	17	432
567	H352	605	67	161	521	102	254	106	66	331	110	121	54	112	165	279	28	126	17	466
568	H353	600	66	162	523	88	243	111	60	334	99	103	49	102	154	270	29	120	17	432
569	H354	603	69	168	556	90	267	97	72	318	93	122	56	116	154	266	30	127	17	437
570	H355	599	67	165	537	76	252	106	67	332	91	115	52	117	156	269	29	123	17	435
571	H356	533	64	150	511	108	268	194	66	277	67	86	45	89	129	382	26	122	14	410
572	H357	546	65	156	547	101	262	172	66	298	84	90	48	96	135	391	26	127	17	399
573	H358	772	64	130	271	21	150	6	67	349	70	82	30	244	151	88	35	365	11	566
574	H359	589	68	157	522	77	269	127	74	291	96	116	52	109	143	294	29	141	16	390
575	H360	552	64	154	537	113	245	169	65	302	85	81	46	101	134	381	27	130	16	424
576	H361	569	66	150	533	98	245	155	73	297	84	80	49	101	130	341	28	135	15	361
577	H362	581	59	132	490	125	260	166	69	280	65	80	44	87	117	353	25	139	14	369
578	H363	545	64	147	539	131	243	181	64	308	89	83	47	100	131	397	27	128	14	404
579	H364	625	68	157	479	63	253	96	75	302	77	114	51	107	142	241	30	145	15	381
580	H365	619	67	158	493	68	249	103	71	289	106	115	52	108	143	249	29	143	16	395
581	H366	597	68	166	535	90	252	109	59	327	87	119	53	118	152	266	29	119	16	410
582	H367	604	66	154	483	68	287	120	74	285	88	113	51	100	136	277	32	145	17	358
583	H368	601	67	161	517	74	251	112	64	328	91	106	49	116	145	258	29	125	14	401
584	H369	562	64	161	559	116	235	140	58	350	106	91	46	92	147	355	24	117	16	514
585	H370	666	66	101	387	85	255	127	108	221	75	108	38	34	84	286	34	365	8	434
586	H371	647	68	106	401	92	269	138	114	214	75	115	41	58	81	297	33	354	9	402
587	H372	651	51	102	350	83	247	147	108	213	63	77	31	61	90	283	29	215	10	335
588	H488	744	69	142	466	29	103	12	104	269	83	86	33	63	127	91	32	327	15	521
589	H489	749	69	141	461	29	114	6	100	277	69	85	33	63	135	99	33	324	15	528
590	H490	741	69	141	472	32	117	8	98	323	73	88	32	67	122	108	34	332	15	520
591	H491	745	70	148	411	20	90	9	107	299	62	81	30	55	134	93	32	313	15	558
592	H492	754	64	137	401	30	100	11	107	304	61	81	30	66	138	101	34	315	12	601
593	H493	749	68	139	438	28	103	13	102	267	69	88	32	62	130	89	37	342	14	503
594	H494	751	67	141	443	27	112	6	98	288	65	85	31	63	135	98	35	333	15	533
595	H495	752	67	139	437	29	115	7	101	283	73	87	32	61	136	102	36	330	14	541
596	H496	750	69	140	456	31	114	6	102	280	73	87	30	57	133	100	35	329	15	503
597	H497	737	71	146	497	34	98	12	101	275	80	89	34	64	126	84	32	335	15	506
598	H498	757	66	136	433	29	108	7	101	281	58	83	31	60	133	98	30	329	13	513
599	H499	750	70	141	451	28	115	6	107	277	72	89	33	66	137	101	34	347	13	517
600	H500	764	64	132	423	29	86	9	103	271	58	80	30	55	121	78	31	318	15	516
601	H501	762	63	131	423	26	108	8	96	285	66	85	28	58	133	98	31	326	13	492
602	H502	744	68	146	428	31	94	9	103	307	62	80	31	55	138	97	34	309	14	580
603	H503	569	60	129	484	107	303	181	70	263	71	107	58	76	97	336	27	143	14	308
604	H504	540	63	137	525	108	414	188	67	263	63	101	60	90	101	334	26	137	15	301
605	H505	633	50	112	444	104	222	148	64	266	67	115	58	72	105	287	23	143	13	357
606	H506	590	59	129	476	106	279	165	62	262	84	102	55	83	104	333	27	146	14	306
607	H507	676	48	107	386	92	251	116	55	246	62	92	49	61	97	255	22	123	12	349
608	H508	745	86	169	344	20	104	4	29	250	81	83	26	32	147	116	26	207	22	468
609	H509	731	92	173	392	19	114	6	28	259	83	91	29	37	152	120	25	209	22	467
610	H510	745	85	159	398	29	109	6	35	260	76	86	3	55	144	106	24	211	21	406

Figure 74: Fortsetzung 11 von Fig. 63.

i	Name	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	V	Cr	Ni	Zn	Rb	Sr	Y	Zr	Nb	Ba
611	H511	733	87	163	447	19	129	8	32	267	66	89	28	54	150	95	27	202	21	404
612	H512	729	95	164	509	29	123	5	24	261	61	89	30	39	153	122	28	219	22	433
613	H513	589	61	132	489	107	278	159	66	277	71	111	58	82	118	343	26	148	13	336
614	H883	680	63	142	474	58	237	60	85	303	69	98	44	72	131	171	35	279	15	495
615	H838	704	61	128	431	58	192	56	98	307	60	84	41	79	120	154	35	308	12	430
616	H848	681	64	137	468	58	214	65	87	302	76	89	45	75	121	131	35	295	15	466
617	H863	666	64	139	464	62	234	76	85	317	78	100	43	73	137	186	37	288	13	491
618	H885	673	63	166	384	41	227	50	69	343	94	102	48	79	175	165	37	229	13	626
619	H882	672	66	152	508	61	246	52	86	312	89	103	45	82	150	167	40	283	15	574
620	H862	662	66	158	533	55	253	52	77	328	94	105	48	95	158	164	40	280	15	625
621	H878	666	64	158	530	56	250	50	76	324	98	104	48	83	156	158	40	264	16	618
622	H857	665	64	158	524	59	249	49	67	351	94	105	47	223	165	165	38	270	14	625
623	H858	681	65	142	493	61	234	57	90	305	77	96	45	72	138	166	37	294	14	534
624	H847	686	64	139	473	56	219	58	88	302	66	93	43	74	134	152	36	306	13	455
625	H842	702	63	131	442	59	200	55	94	297	68	100	41	66	130	160	36	314	14	468
626	H874	690	62	146	491	60	223	45	82	309	78	105	45	94	144	157	37	279	15	529
627	H880	689	66	155	420	42	178	54	64	276	103	99	49	250	127	164	34	178	11	409
628	H870	663	61	127	434	73	306	90	94	280	75	91	40	66	117	191	34	291	13	469
629	H849	696	65	136	466	59	203	53	89	301	70	90	43	72	131	148	35	305	13	489
630	H873	659	65	156	527	61	250	57	76	336	92	108	48	103	151	175	39	278	16	597
631	H851	657	64	161	384	52	247	66	73	342	94	101	45	74	179	206	37	248	15	609
632	H877	667	61	133	445	65	252	84	93	293	75	91	41	130	128	199	36	287	14	522
633	H846	678	61	133	448	62	217	72	89	309	75	90	43	71	115	170	36	290	14	440
634	H865	657	64	157	516	55	277	58	76	331	92	101	46	97	158	161	39	271	15	580
635	H855	654	65	143	498	61	265	80	89	286	71	98	43	71	137	190	39	292	16	528
636	H875	661	66	165	552	58	260	44	78	324	104	106	50	88	169	160	41	266	16	632
637	H868	671	64	141	482	56	268	66	92	290	81	104	45	78	137	184	37	289	15	590
638	H867	667	63	137	464	64	266	77	91	287	78	96	42	69	134	190	36	293	13	491
639	H850	680	64	166	384	43	198	34	57	398	105	91	48	86	186	179	37	254	14	629
640	H856	687	67	150	509	54	238	40	90	304	80	106	46	76	148	152	36	300	15	528
641	H872	666	65	157	511	51	258	50	78	334	81	107	47	106	157	175	39	275	16	624
642	H839	692	62	131	443	67	206	62	95	301	61	89	43	66	128	171	36	304	14	495
643	H852	628	59	153	378	72	224	98	69	335	71	87	44	86	165	251	40	251	15	588
644	H871	664	66	158	535	57	262	50	78	321	89	107	48	93	160	178	41	286	16	633
645	H866	671	67	146	504	57	268	60	84	292	83	107	47	83	143	177	37	306	15	523
646	H864	686	64	140	473	54	227	56	87	309	82	90	44	69	133	153	37	288	14	507
647	H840	676	65	148	499	63	220	56	87	315	80	90	46	80	144	155	40	284	15	548
648	H843	690	62	131	444	63	203	66	94	310	75	87	41	70	122	166	36	304	12	483
649	H836	663	66	162	547	62	239	46	80	327	73	106	48	90	162	162	42	270	17	636
650	H860	670	67	147	507	60	266	59	85	303	65	109	46	77	144	171	36	303	14	509
651	H844	702	62	131	443	58	193	55	92	302	78	88	43	68	127	153	37	318	13	464
652	H881	664	68	153	528	62	246	53	77	372	67	111	49	81	114	167	39	290	16	519
653	H869	674	64	143	481	67	260	62	90	297	77	99	45	81	141	176	37	288	15	520
654	H841	673	63	142	477	67	209	65	85	311	63	86	46	73	126	172	39	284	15	522
655	H859	665	65	159	532	56	280	46	76	327	90	107	48	90	163	167	40	277	16	624
656	H861	665	63	157	515	58	253	52	75	334	91	107	47	84	163	176	39	270	16	630
657	H884	671	63	159	379	49	227	57	65	361	80	94	46	81	173	191	37	261	15	604
658	H854	674	60	155	475	50	241	42	105	385	78	84	45	79	156	157	37	254	15	554
659	H845	696	61	130	432	58	199	62	94	311	65	99	44	62	126	144	34	305	12	450
660	H853	683	59	141	353	64	231	66	89	308	59	86	43	76	153	212	33	256	13	538

obergermanischen Heeresziegelei? *Archäometrie und Denkmalpflege*, 28–30.

[3] FRIEDMAN, J.H. and TUKEY, J.W. (1974): A projection pursuit algorithm for exploratory data analysis. *IEEE Transaction on Computers,* C- 23(9):881-890.

[4] GALLEGOS, M.T. and RITTER, G. (2005): A robust method for cluster analysis. *Annals of Statistics*, 33, 347–380.

[5] GALLEGOS, M.T. and RITTER, G. (2009): Trimming algorithms for clustering contaminated grouped data and their robustness. *Advances in Data Analysis and Classification*, 3. To appear

[6] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2003): Modellbasierte Clusteranalyse römischer Ziegel aus Worms und Rheinzabern. *Archäologische Informationen*, 26/2, 471–480.

[7] MUCHA, H.-J., BARTEL, H.-G. and DOLATA, J. (2005): Model-based Cluster Analysis of Roman Bricks and Tiles from Worms and Rheinzabern. In: C. Weihs and W. Gaul, W. (Eds.): *Classification - The Ubiquitous Challenge,* Springer, Berlin, 317–324.

[8] POSSE, C. (1995): Projection pursuit exploratory data analysis. *Computational Statistics & Data Analysis,* 20(6):669–687.

[9] ROHATSCH, T., PÖPPEL G. and WERNER H. (2006): Projection Pursuit for Analyzing Data From Semiconductor Environments. *IEEE Transactions on Semiconductor Manufacturings,* Vol. 19, No 1, February 2006.

[10] WARD, J.H. (1963): Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58, 236–244.

8 Synthetic Data Set 1

8.1 Gunter Ritter: Structure of the synthetic data set Berlin08_synth1

Sampling

Berlin08_synth1 is a five-dimensional numerical data set consisting of four spherical, non-normal clusters of sizes 500, 300, 300, and 200, respectively. The four cluster centers are the extreme points of a regular tetrahedron, namely,

0.8 0.8 0.8 0.8 0.0
-0.8 -0.8 0.8 0.8 0.0
0.8 -0.8 0.8 -0.8 0.0
-0.8 0.8 0.8 -0.8 0.0

Each cluster is sampled in the following way: (1) Data points are sampled from the 5D standard normal; (2) each data point is multiplied by its norm; (3) the data points are shifted by the cluster center.

Comment

Cluster centers are at the mutual distance $\sqrt{2 \cdot 1.6^2} \approx 2.26$. This means that, without the multiplication (2), the data set would consist of moderately separated, spherical normal clusters of equal (unit) variance. It could be clustered by classical k–means since cluster sizes are not too unbalanced. Of course, there would be a good number of errors since separation is not perfect. The point is the multiplication which attracts small data points to the center and drives away points of size > 1 creating a heavy tail. Nevertheless, 332 out of the 500 data points sampled for class 1 are closer to their center than to any of the other three centers. However, even a method that manages to understand the structure of the data set will probably misclassify at least 1/3 of the data (unless it is clairvoyant).

8.2 Gerhard Pöppel and Reinhard Schachtner: Analysis A by MCLUST

The Projection Pursuit approach did not reveal a clear cluster structure on the Berlin08_synth1 dataset. Therefore we utilized a "VVV-Mixture Model" approach (Fraley and Raftery (1999), Fraley and Raftery (2006), Stritt et al. (2007)). The result of the Mixture Model analysis indicates that the subgroups of the data interfuse a lot (see Fig. 75).

Group	Class		feat0	feat1	feat2	feat3	feat4		Counts
A	0.00	Mean	-0.36	-0.37	0.79	0.02	-0.08		220.00
B	1.00	Mean	0.67	0.88	1.85	0.98	-0.36		62.00
C	2.00	Mean	0.20	0.14	0.47	0.19	-0.05		362.00
D	3.00	Mean	0.13	0.19	0.85	0.29	0.14		366.00
E	4.00	Mean	0.53	0.09	0.73	0.00	-0.03		290.00
									1300.00

Figure 75: Cluster results for the Berlin08_synth1 data gained by Mixture Models

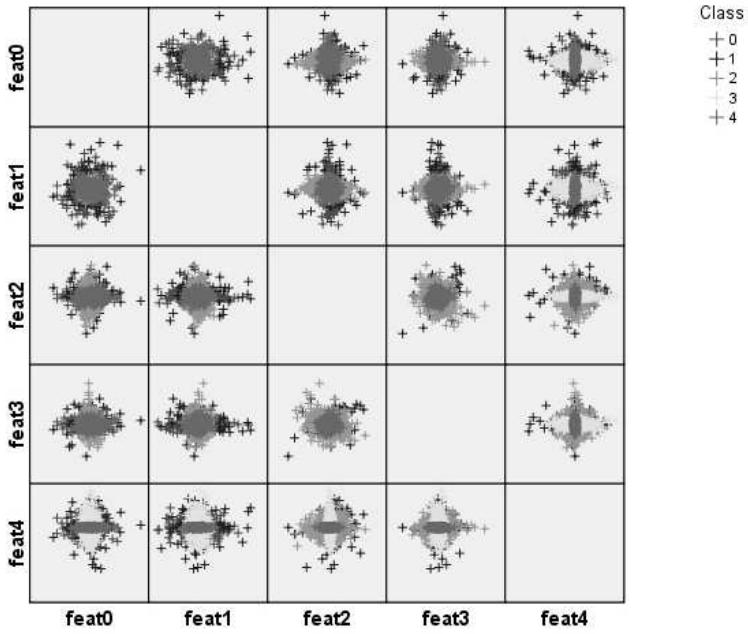


Figure 76: Scatterplots of the original variables of the Berlin08_synth1 data

Consideration of the scatterplots of the original data as shown in Fig. 76 indicates that the projection onto variable feat1 and feat2 shows an orthogonal orientation of group A (class = 0) versus group C (class = 2), whereas the projection onto variable

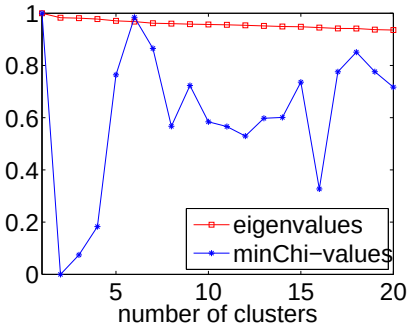


Figure 77: Synthetic data set 1: minChi-values and eigenvalues for different numbers of clusters.

feat2 and feat4 shows an orthogonal orientation of group C (class = 3) versus D (class = 4).

8.3 Susanna Röblitz and Marcus Weber: Analysis B by Spectral Clustering

This is another application of the method that is described by the authors at full length in Sec. 4. A first look at the data set did not reveal any elongated regions of data points, so we decided to use the Euclidean distance. From the distance matrix we computed a k-nearest neighbor graph with $k = 10$ and weighted the edges by

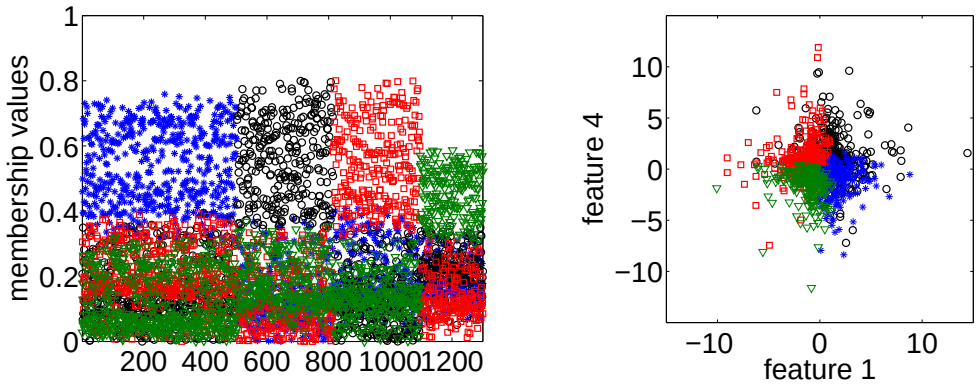
$$w_{ij_k} = \frac{1}{2^{j_k}}, \quad j_k = 1, \ldots, 10,$$

where j_k denotes the kth nearest neighbor. Afterwards, W was symmetrized by

$$w^\star_{ij} = \max(w_{ij}, w_{ji}).$$

Then the matrix $P = D^{-1}W^\star$ was used for spectral clustering. The minChi-criterion indicates that there are at most four clusters (see Fig. 77), but the resulting clusters are not well separated, see Fig. 78. The partition can be visualized best in the two-dimensional space spanned by features 1 and 4, where the clusters are more or less separated by the coordinate axes. The contingency table presented in Sec. 8.4 summarizes the assignment of objects to the clusters after re-transformation of the membership matrix χ to an indicator matrix.

In sum, we suggest that there are in fact four clusters which are not well-separated but overlap each other. Some characteristic properties are listed in Tab. 8.



(a) Sorted membership values.

(b) Data points in a 2-dimensional sub-space.

Figure 78: Synthetic data set 1: Partition into $k = 4$ clusters.

Table 8: Synthetic data set 1: Some characteristics of the identified clusters.

$k$weight	objects	χ-weighted mean	χ-weighted standard dev.
1404.50	502	(0.81, 0.79, 0.67, 0.81 − 0.10)	(1.94, 1.99, 2.00, 2.09, 1.99)
2326.57	312	(−0.82, −0.53, 0.72, 0.69, 0.08)	(1.92, 2.03, 2.04, 2.12, 1.86)
3322.00	282	(0.98, −0.46, 0.91, −0.51, 0.06)	(1.89, 2.08, 2.01, 1.99, 1.97)
4246.94	204	(−0.57, 0.50, 0.72, −0.60, −0.08)	(1.96, 2.13, 2.00, 2.00, 1.87)

8.4 Gunter Ritter and Hans-Joachim Mucha: Comparison of Results

Here, we are in a better situation for an assessment of the performance of clustering techniques than before. This is because the true classification is known and we may establish contingency tables w.r.t. the true classification.

Model–based methods of clustering depend on a statistical model. In particular, Fraley and Raftery's MCLUST is based on normal models. Although the clusters of this data set are sampled from spherical populations the "VVV-Mixture Model" is not well adapted to the data set because of its heavy tails. Therefore, MCLUST produces a good number of errors on this data set. This is seen from the contingency table below. It compares the partition obtained by Pöppel/Schachtner in "Analysis A" for the "VVV-Mixture Model" with the true partition.

$$\begin{bmatrix} 153 & 137 & 113 & 71 & 26 \\ 63 & 92 & 69 & 61 & 15 \\ 93 & 73 & 65 & 54 & 15 \\ 57 & 60 & 43 & 34 & 6 \end{bmatrix}$$

On the other hand, sphericity of the clusters offers good conditions for a distance based method such as Spectral Clustering. In fact, Röblitz/Weber approach with this method the optimal error rate predicted in Sect. 8.1. The following contin-

gency table which compares their partition obtained in "Analysis B" for Spectral Clustering with the true partition.

$$\begin{bmatrix} 366 & 52 & 57 & 25 \\ 54 & 186 & 29 & 31 \\ 45 & 49 & 183 & 23 \\ 37 & 25 & 13 & 125 \end{bmatrix}$$

Obviously, spectral clustering outperforms normal, model-based cluster analysis on this data set.

As Pöppel and Schachtner mention, *Projection Pursuit* encounters problems with this data set. The reason is the heavy overlapping. Sphericity and equal spread of all clusters is a condition for the functioning of *k–means*. However, k–means yields an unsatisfactory result on this data set due to the overlapping caused by the heavy tails.

References

[1] FRALEY, C. and RAFTERY A. (1999): MCLUST: Software for model-based cluster analysis. J. Classif. 16 (2), 297–306. Also: Technical Report No. 342, Dept. of Statistics, University of Washington.

[2] FRALEY, C. and RAFTERY A. (2006) MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering, Technical Report No. 504, Department of Statistics, University of Washington.

[3] STRITT, M., SCHMIDT-THIEME, L., and POEPPEL, G. (2007): Combining multi-distributed mixture models and bayesian networks for semi-supervised learning, Page(s): 354-362 Digital Object Identifier 10.1109/ICMLA.2007.60 .

9 Synthetic Data Set 2

9.1 Gunter Ritter: Structure of the synthetic data set Berlin08_synth2

Sampling

The construction of the numerical data set Berlin08_synth2 starts with 300 data points sampled from NV_{0,V_1}, 500 from NV_{0,V_2}, and 500 from NV_{0,V_3}, where the V_j's are the diagonal matrices with diagonals $(1, 10^{-6}, 1, 1, 1)$, $(0.25, 10^{-6}, 0.25, 1, 1)$, and $(1, 10^{-6}, 0.25, 0.25, 1)$, respectively. The whole data set is subsequently rotated by the orthogonal matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 1/2 & -1/2 \\ 0 & -1/2 & 1/2 & 1/2 & 1/2 \\ 0 & 1/2 & -1/2 & 1/2 & 1/2 \\ 0 & -1/2 & -1/2 & 1/2 & -1/2 \end{pmatrix},$$

and multiplied by 2. Finally, the third group is shifted by the vector $1.5\, e_3$, $e_3 =$ the third unit vector.

Comment

The second group lies within the first since it has the same mean and a smaller covariance matrix. It is, therefore, legitimate to speak of two clusters, one of them not normal. These two clusters are completely separated since, after rotation, the variance in the direction of the shift is small. Viewed from an appropriate direction the data set looks like two parallel needles.

9.2 Gerhard Pöppel and Reinhard Schachtner: Analysis by Projection Pursuit

Using kurtosis indices for Projection Pursuit leads to a clear structure of two disjunct clusters, which are separated along the hyperplane $0.5\text{feat}1 - 0.5\text{feat}2 + 0.5\text{feat}3 - 0.5\text{feat}4$. Since Projection Pursuit did not detect further disjunct substructures we stopped analyzing this dataset at that point. We neither did characterize in detail the clusters found.

Group	Run 1 Projection 97		feat0	feat1	feat2	feat3	feat4		Counts
A	0.00	Mean	-0.02	-0.06	1.46	0.02	0.01		500.00
B	1.00	Mean	0.02	0.04	-0.04	-0.05	0.03		800.00
									1300.00

Figure 79: Cluster Separation of the Berlin08_synth2 data found by Projection Pursuit

Note that this cluster structure can be seen in the scores of the lowest eigenvalue, consulting the scatterplots of the PCA tranformed data (see Fig. 80).

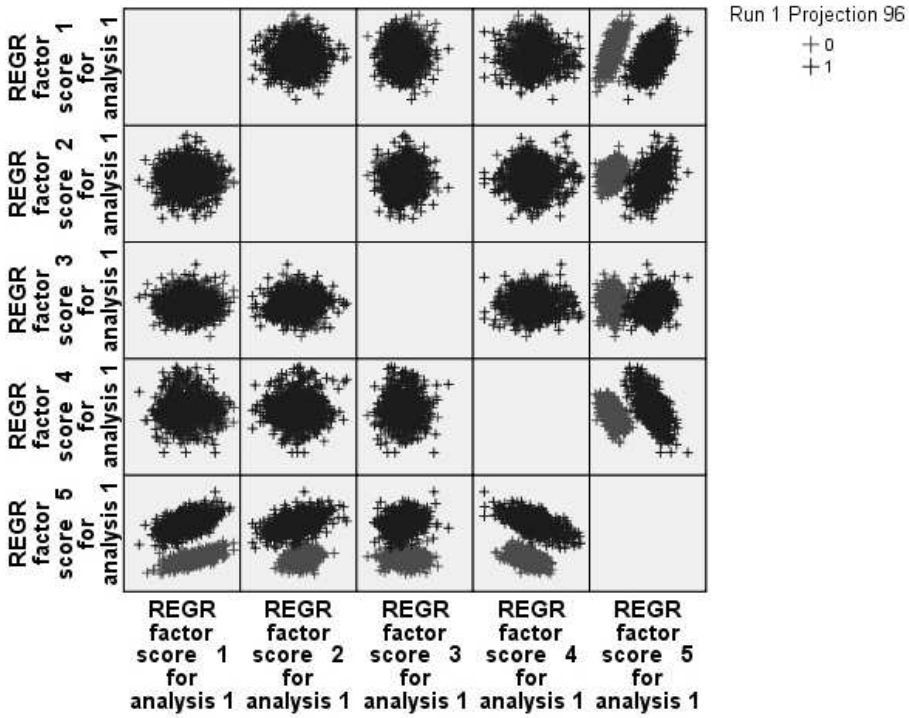


Figure 80: Scatterplots of the PCA transformed Berlin08_synth2 data

9.3 Gunter Ritter and Hans-Joachim Mucha: Comparison of Results

This data set can be interpreted as consisting of two well-separated, elliptical clusters. It is, therefore, well suited for clustering with Projection Pursuit or model–based methods. In fact, Pöppel and Schachtner present a perfect partition in two clusters obtained with Projection Pursuit. This is plain from the following contingency table which crosses their partition with the correct one.

$$\begin{bmatrix} 500 & 0 \\ 0 & 500 \\ 0 & 300 \end{bmatrix}$$

A distance-based method such as Spectral Clustering does not seem to be a method well-suited for this data set. The reason are the oblong clusters. Susanna Röblitz and Marcus Weber write: "We did not succeed in clustering this data set [by spectral clustering]. That means, from a distance-based point of view there is only one cluster. Maybe model-based clustering methods could be appropriate here."

Indeed, we applied a heteroscedastic cluster criterion based on full normal assumptions with the BIC model selection criterion. The latter was uncertain whether there were two or three clusters but both solutions that we obtained were reasonable.

Part III
List of participants

List of participants of the 30th DANK Fall Meeting, Nov. 14th to 15th 2008, Berlin

- ◯ **Bartel**, Hans-Georg, PD Dr., Humboldt-Universität zu Berlin, Institut für Chemie
- ◯ **Ey**, Rüdiger, Director CRS Custom Research Services, Berlin
- ◯ **Gaul**, Wolfgang, Prof. Dr., Universität Stuttgart
- ◯ **Hennig**, Christian, Dr., University College London (UCL), Department of Statistical Science,
- ◯ **Kurz**, Peter, TNS Infratest, München
- ◯ **Meyer**, Florian, Dipl.-Math., Universität Marburg
- ◯ **Mucha**, Hans–Joachim, Dipl.-Math., WIAS Berlin
- ◯ **Müller**, Simon, Dipl.-Math., Universität Stuttgart
- ◯ **Pöppel**, Gerhard, Dr. rer. nat., Infineon Technologies AG, Regensburg
- ◯ **Ritter**, Gunter, Prof. Dr., Universität Passau
- ◯ **Röblitz**, Susanna, Dr., Zuse Institute Berlin (ZIB)
- ◯ **Röhrl**, Norbert, Dr., Universität Stuttgart
- ◯ **Sahmer**, Karin, Dr., Institut Supérieur d'Agriculture, Lille, France
- ◯ **Spickenheuer**, Anne, Dr., Universität Stuttgart
- ◯ **Ultsch**, Alfred, Prof. Dr., Universität Marburg
- ◯ **Weber**, Marcus, Dr., Zuse Institute Berlin (ZIB)