Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Technical Report ISSN 1618 – 7776

Model-based Cluster Analysis Applied to Flow Cytometry Data of Phytoplankton

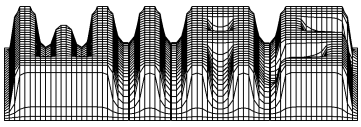
H.–J. Mucha[1], U. Simon[2] and R. Brüggemann[2]

submitted: Dec 16 2002

[1] Weierstraß-Institut für Angewandte Analysis und Stochastik
Mohrenstr. 39
10117 Berlin

[2] Leibniz-Institut für Gewässerökologie und Binnenfischerei
Müggelseedamm 310
12587 Berlin

No. 5
Berlin 2002



2000 *Mathematics Subject Classification.* 62-07, 62H30, 62H25.

Key words and phrases. Cluster analysis, K-means, data mining, principal components analysis, freshwater ecology, phytoplankton, flow cytometry.

.

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

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

Abstract

Starting from well-known model-based clustering models equivalent formulations for some special models based on pairwise distances are presented. Moreover, these models can be generalized in order to taking into account both weights of observations and weights of variables. Well-known cluster analysis techniques like the iterative partitional *K-means* or the agglomerative hierarchical *Ward* are useful for discovering partitions or hierarchies in the underlying data. Here these methods are generalised in two ways, firstly by using weighted observations and secondly by allowing different volumes of clusters. Then a more general *K-means* approach based on pair-wise distances is recommended. Simulation studies are carried out in order to compare the new clustering techniques with the well-known ones.

Afterwards a successful application in the field of freshwater ecology is presented. As an example, the cluster analysis of a snapshot from monitoring of phytoplankton (algae) is considered in more detail. Indeed, monitoring by microscope is very time- and work-consuming. Flow cytometry provides the opportunity to investigate algae communities in a semiautomatic way (Hofstraat et al. 1994). Statistical data analysis and cluster analysis can at least support the investigations. Here a combination of agglomerative hierarchical clustering and iterative clustering is recommended. In data mining, such a similar combination was proposed by Faber et al. (1994) in the field of image segmentation of Landsat data. In order to give some insight into the data under investigation several univariate, bivariate and multivariate visualizations are proposed.

Contents

List of Figures

List of Tables

1 Introduction

Cluster analysis aims at finding interesting partitions or hierarchies directly from the data without using any background knowledge. Here a partition P(I,K) is an exhaustive subdivision of the set of I objects (observations) into K non-empty clusters (subsets, groups) C_k that are pair-wise disjoint. On the other hand a hierarchy is a sequence of nested partitions. There are model-based as well as heuristic clustering techniques. At most one will set up new hypotheses about the data. At least clustering should result in practical useful partitions or hierarchies. More details and many more references can be found, for example, in the monographs of Bock (1974), Späth (1985), Jain and Dubes (1988), Kaufman and Rousseeuw (1990), Mucha (1992), and Gordon (1999). Concerning model-based clustering the papers of Banfield and Raftery (1993), Fraley (1996) as well as Fraley and Raftery (2002) are a good introduction. Successful application of model-based clustering are known, for example, from image processing (analysis of computed-tomography scans or Landsat images (see for instance, Faber et al. (1994)), and color image quantization (Murtagh, Raftery, and Starck (2001)), and from archaeometry (Mucha, Bartel, and Dolata (2002)).

Beside the most general Gaussian model for clustering two simple models will be considered here in a generalised form using weighted observations. They lead to the sum-of-squares and logarithmic sum-of-squares criterion. Both criteria can be formulated in an equivalent fashion using pair-wise distances between observations. The principle of weighting of observations is a key idea in data mining for handling cores (representatives of dense regions) and outliers. In the case of outliers one has to downweight them in order to reduce their influence. A first important attempt at downweighting of the observations goes back to Hampel (1968). Based on the theory of median absolute deviation (MAD) clear outliers in single coordinates can be downweighted and rejected. Below only a simple multivariate empirical approach of downweighting will be investigated.

As an application in the field of freshwater ecology, clustering data from a snapshot of monitoring of phytoplankton of lake Müggelsee (Berlin) is considered here. Indeed, monitoring by microscope is very time- and work-consuming. Flow cytometry provides the opportunity to investigate algae communities in a semiautomatic way followed by statistical data analysis and cluster analysis.

2 Model-based Gaussian clustering

Generally, the population of interest consists of K different subpopulations (clusters) with densities $f_k(\mathbf{x};\theta)$, $k = 1, 2, ..., K$, for some unknown vector of parameters θ. Here $\mathbf{x}$ is a J-dimensional observation. Let $\gamma = (\gamma_1, ..., \gamma_I)$ be a set of identifying labels of I observations so that $\gamma_i = k$ if $\mathbf{x}_i$ comes from the k-th cluster. In the most general classification likelihood approach, the identifying labels are chosen so as to

maximize the likelihood

$$L(\mathbf{x};\theta,\gamma) = \prod_{i=1}^{I} f_{\gamma_i}(\mathbf{x};\theta). \tag{1}$$

In the following, the focus is on the assumption that $f_k(\mathbf{x};\theta)$ is multivariate normal with the special parameters θ_k consisting of mean vector μ_k and covariance matrix Σ_k. The density $f_k(\mathbf{x};\theta)$ has the form

$$f_k(\mathbf{x};\mu_k,\Sigma_k) = (2\pi)^{-\frac{J}{2}}|\Sigma_k|^{-\frac{1}{2}}\exp\{\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\}. \tag{2}$$

Here J is the number of dimensions (variables). Banfield and Raftery (1993) developed a model-based framework for clustering by parameterizing the covariance matrix in terms of its eigenvalue decomposition. Mardia et al. (1979) described in detail the classification likelihood approach to model-based Gaussian clustering. The alternative to the classification likelihood is the mixture likelihood approach, which is recently favoured by Fraley and Raftery (2002). One reason for this is that outliers and noisy data can be handled more easily within a mixture context. In the following the focus is on two special assumptions about the covariance structure (for further reading see Fraley (1996)). Let $\mathbf{X}$ be the $(I \times J)$-data matrix under investigation consisting of I observations (objects) and J variables. When the covariance matrix is constrained to be diagonal and uniform across all groups, the well-known sum-of-squares criterion

$$V_K = \sum_{k=1}^{K} tr(\mathbf{W}_k), \tag{3}$$

has to be minimized. Herein $\mathbf{W}_k = \sum_{i\in C_k}(\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T$ is the sample cross-product matrix for the k-th cluster C_k, and $\overline{\mathbf{x}}_k$ is the usual maximum likelihood estimate of expectation values in cluster C_k.

It is well known that criterion (3) can be written in the following equivalent form without explicit specification of cluster centres (centroids) $\overline{\mathbf{x}}_k$

$$V_K = \sum_{k=1}^{K} 1/n_k \sum_{i\in C_k}\sum_{l\in C_k, l>i} d_{il}, \tag{4}$$

where n_k is the cardinality of cluster C_k, and

$$d_{il} = d(\mathbf{x}_i,\mathbf{x}_l) = (\mathbf{x}_i - \mathbf{x}_l)^T(\mathbf{x}_i - \mathbf{x}_l) = \|\mathbf{x}_i - \mathbf{x}_l\|^2$$

is the pair-wise squared Euclidean distance between two observations i and l. 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. Taking into account weights of the variables the sample cross-product matrices can be composed of two parts, and in (4), equivalently, the squared weighted Euclidean distance

$$d_Q(\mathbf{x}_i,\mathbf{x}_l) = (\mathbf{x}_i - \mathbf{x}_l)^T\mathbf{Q}(\mathbf{x}_i - \mathbf{x}_l), \tag{5}$$

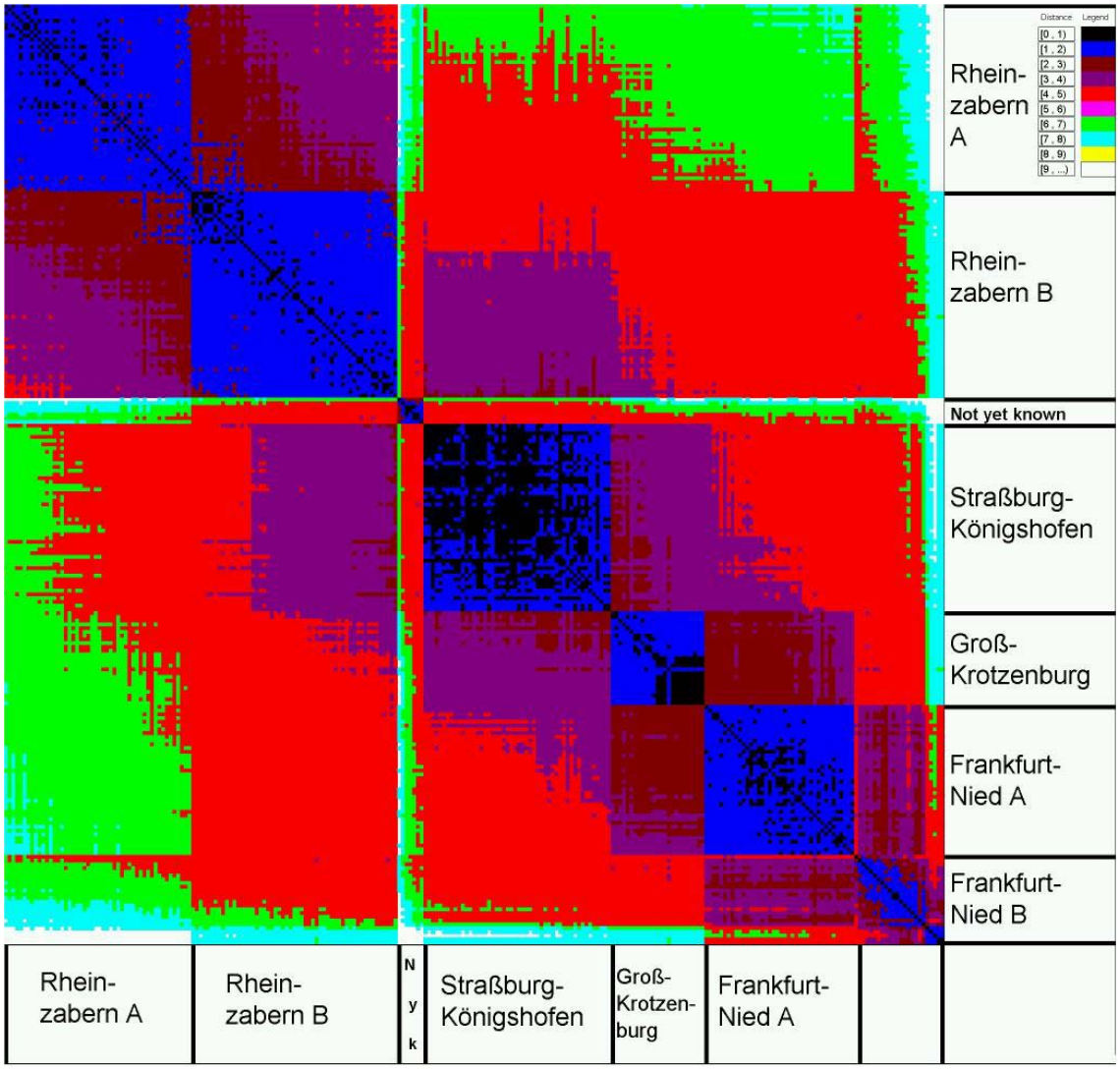


Figure 1: Fingerprint of a distance matrix (excerpt) suggesting several clusters (data: chemical compositions of Roman tiles coming from the Rhine area).

is used, where here $\mathbf{Q}$ is restricted to be diagonal. In doing so, at least scaling problems can be handled fashionable without any data preprocessing step for 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), see Mucha (1992, 1994, 1995).) Another approach of assigning weights to variables is clustering observations an subsets of variables (Friedman and Meulman (2002)). Of course, the statistical distance (5) can be generalized to cluster specific statistical distances, where instead of $\mathbf{Q}$ a matrix $\mathbf{Q_k}$ equals to the inverse within-cluster covariance matrix is used (Späth (1985)).

There are at least two well-known clustering techniques for minimizing the sum-of-squares criterion: the partitional *K-means* (MacQueen (1967), Bock (1974), Mucha (1992)) minimizes criterion (3) for a single partition P(I,K) by exchanging observations between clusters, and the hierarchical *Ward* (Ward (1963), Späth (1980), Mucha (1992)) minimizing (4) in a stepwise manner by agglomerative hierarchical clustering. For illustration purposes here only, Figure 1 shows a so-called fingerprint of a distance matrix. It expresses one adavantage of clustering based on pairwise distances, namely the visualization of arbitrary high dimensional data in only two dimensions usually (see Figure 2 below for another visualization technique). Here the color expresses the level of distance between a pair of observations (see the legend at the upper right hand corner of the picture). An intuitive impression of this figure suggests that there are several well separated clusters. The data under investigation, and thus this example of a distance matrix, accrues from archaeometry, where chemical compositions of 613 Roman tiles coming from the Rhine area are measured by X-Ray Fluorescence Analysis (RFA). Each tile is characterized by 19 variables (chemical trace elements and oxides). For further details concerning this application, see Mucha et al. (2002).

Figure 2 shows an estimated density of pairwise Euclidean distances that is a typical one in the case of several well separated clusters (top of the picture). Here the Epanechnikov kernel with the bandwidth 0.2 is applied to a randomly drawn sample from the data of the application below (data: phytoplankton data, see Section 6). Concerning nonparametric density estimation the monograph of Härdle (1990) is recommended. For reason of comparability at the bottom of Figure 2 a density estimate of pairwise Euclidean distances is given that is a typical one in the case of a randomly generated multivariate spherical Gaussian distribution without any cluster structure.

When the covariance matrix of each cluster is constrained to be diagonal, but otherwise allowed to vary between groups, the logarithmic sum-of-squares criterion

$$U_K = \sum_{k=1}^{K} n_k \log tr(\mathbf{W}_k/n_k), \tag{6}$$

has to be minimized. Once again the following equivalent formulation can be derived

$$U_K = \sum_{k=1}^{K} n_k \log(\sum_{i \in C_k} \sum_{l \in C_k, l>i} \frac{1}{n_k^2} d_{il}). \tag{7}$$

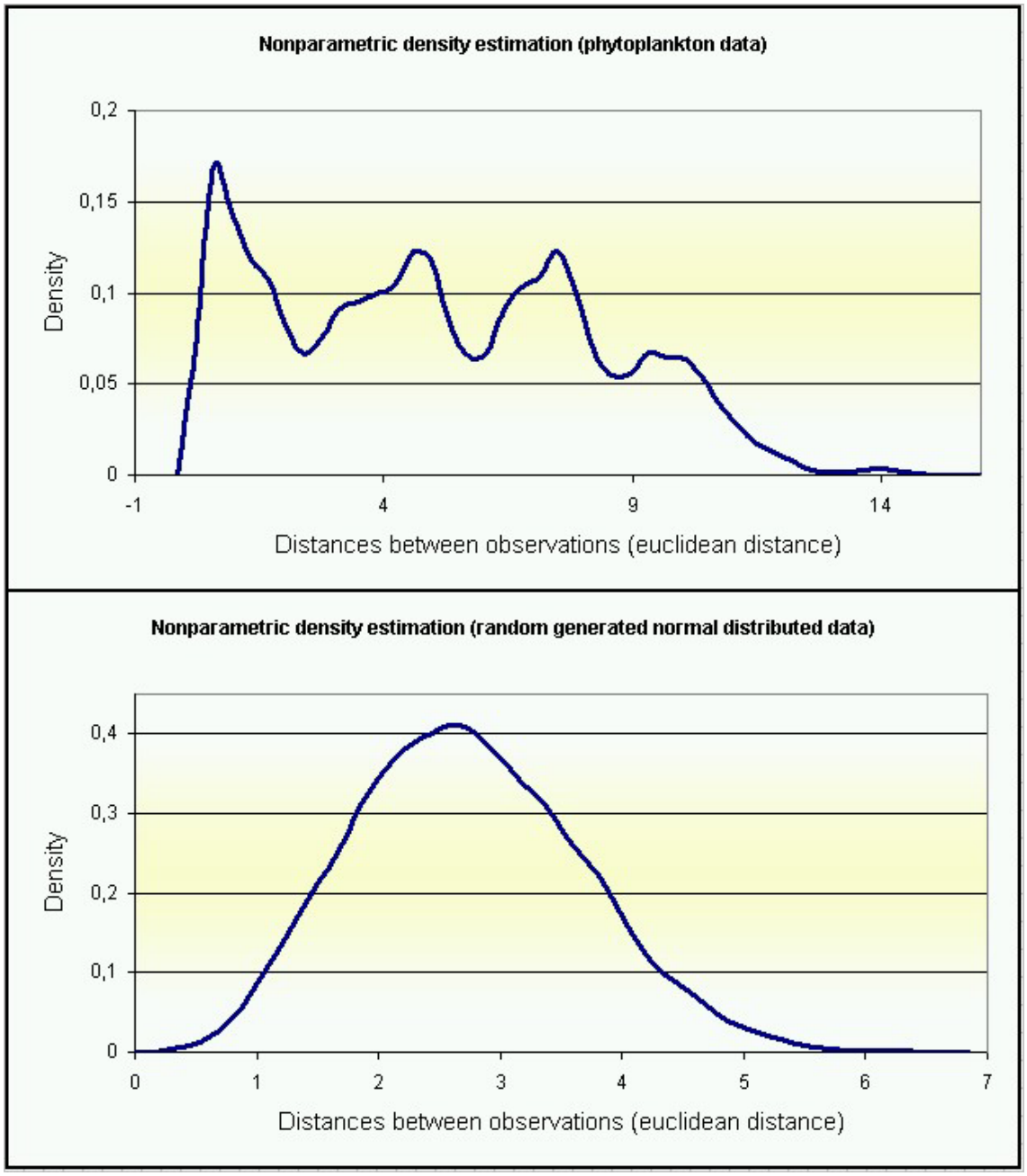


Figure 2: Distributions of pairwise distances coming from two samples of size 250 observations.

Another adavantage of clustering based on pairwise distances over clustering based directly on the $(I \times J)$-data matrix $\mathbf{X}$ is the more general meaning of distances. For example, going via distances allows cluster analysis of mixed data (quantitative and qualitative data). In that way usually at least practical useful exploratory results can be obtained.

The most general model-based Gaussian clustering is when the covariance matrix $\mathbf{W}_k$ of each cluster k is allowed to vary completely. Then the log-likelihood is maximized whenever the partition P(I,K) minimizes

$$Y_K = \sum_{k=1}^{K} n_k \log |\frac{\mathbf{W}_k}{n_k}|. \tag{8}$$

This criterion is obtained by taking advantage of the monotone log-function with regard to the densities (2) that are used in the general equation (1).

3 Model-based clustering using weighted observations

Usually, all observations have the same weight. 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, proportional to the cardinality of the cores.

The above given formulae (4) can be generalized by using positive weights of observations to

$$V_K = \sum_{k=1}^{K} v(C_k) = \sum_{k=1}^{K} \frac{1}{M_k} \sum_{i \in C_k} m_i \sum_{l \in C_k, l>i} m_l d_{il}, \tag{9}$$

where $M_k = \sum_{i \in C_k} m_i$ and m_i denote the mass of cluster C_k and the mass of observation i, respectively. Furthermore, $v(C_k)$ denotes the within-cluster variance of cluster k. Of course, such a generalisation can also be formulated based on formulae (3) by using a weighted sample cross product matrix.

Concerning the *K-means* algorithm based on exchanging observations between clusters in order to minimize (9) the following condition of exchange of an observation i from cluster k into cluster g has to be fulfilled

$$v(C_k \backslash \{i\}) + v(C_g \cup \{i\}) < v(C_k) + v(C_g),$$

where

$$v(C_k \backslash \{i\}) = \frac{1}{M_k - m_i}(\sum_{l \in C_k} \sum_{h \in C_k, h>l} m_l m_h d_{lh} - \sum_{h \in C_k} m_i m_h d_{ih})$$

and

$$v(C_g \cup \{i\}) = \frac{1}{M_g + m_i}(\sum_{l \in C_g} \sum_{h \in C_g, h>l} m_l m_h d_{lh} + \sum_{h \in C_g} m_i m_h d_{ih}).$$

Considering formulae (7) (and (6) in the case of formulation with sample cross product matrices, respectively) the generalized logarithmic sum-of-squares criterion can be derived as follows

$$U_K = \sum_{k=1}^{K} M_k \log(\sum_{i \in C_k} \sum_{l \in C_k, l>i} \frac{m_i m_l}{M_k^2} d_{il}). \tag{10}$$

According to this logarithmic sum-of-squares criterion the partitional *K-means*-like clustering algorithm is denoted here *Log-K-means* and the hierarchical *Ward*-like agglomerative method is denoted *LogWard*. Concerning the hierarchical algorithms there are special treatments of observations with low weights in use (see, for example, Mucha et al. (2002)). Because Ward's hierarchical agglomerative clustering is based on minimum incremental of sum of squares, 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 are also more general because they never require an $(I \times J)$-data matrix $\mathbf{X}$.

These more general algorithms, as the original *K-means* and *Ward*, are part of our prototype-software `ClusCorr98`® using the Excel spreadsheet environment and its data base connectivity. `ClusCorr98`® contains a set of statistical tools for data exploration with emphasis on (adaptive) clustering and multivariate graphical visualizations. The programming language is Visual Basic for Applications (VBA). Almost all numerical and graphical results presented here are made by using `ClusCorr98`®.

4 Core-based clustering

In the following simple techniques of clustering based on cores are proposed. Generally, a core is a dense region in the high-dimensional space that, for example, can be represented by its most typical observation, its centroid or, more generally, by assigning weight functions to the observations. There are at least two reasons for dealing with weighted observations or representatives of cores. First, a huge amount of data has to be clustered efficiently and a hierarchical clustering in a direct way is possible often not till then at all. Second, the problem of outliers in high-dimensional spaces has to be solved in an at least pragmatic way. Concerning these tasks there are, for example, some interesting proposals from Zhang et al. (1996) and Guha et al. (1998). The first one, called BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), performs preclustering and then uses a centroid-based hierarchical clustering algorithm. The second one, called CURE (Clustering Using REpresentatives), identifies clusters having non-spherical shapes and wide variances in size. CURE seems to be more robust against outliers than BIRCH.

Here sum-of-squares and logarithmic sum-of-squares clustering based on cores is a little bit investigated, respectively. There are many ways to deal with cores. Two of them, which are considered here, use the fast *K-means* as a preclustering step

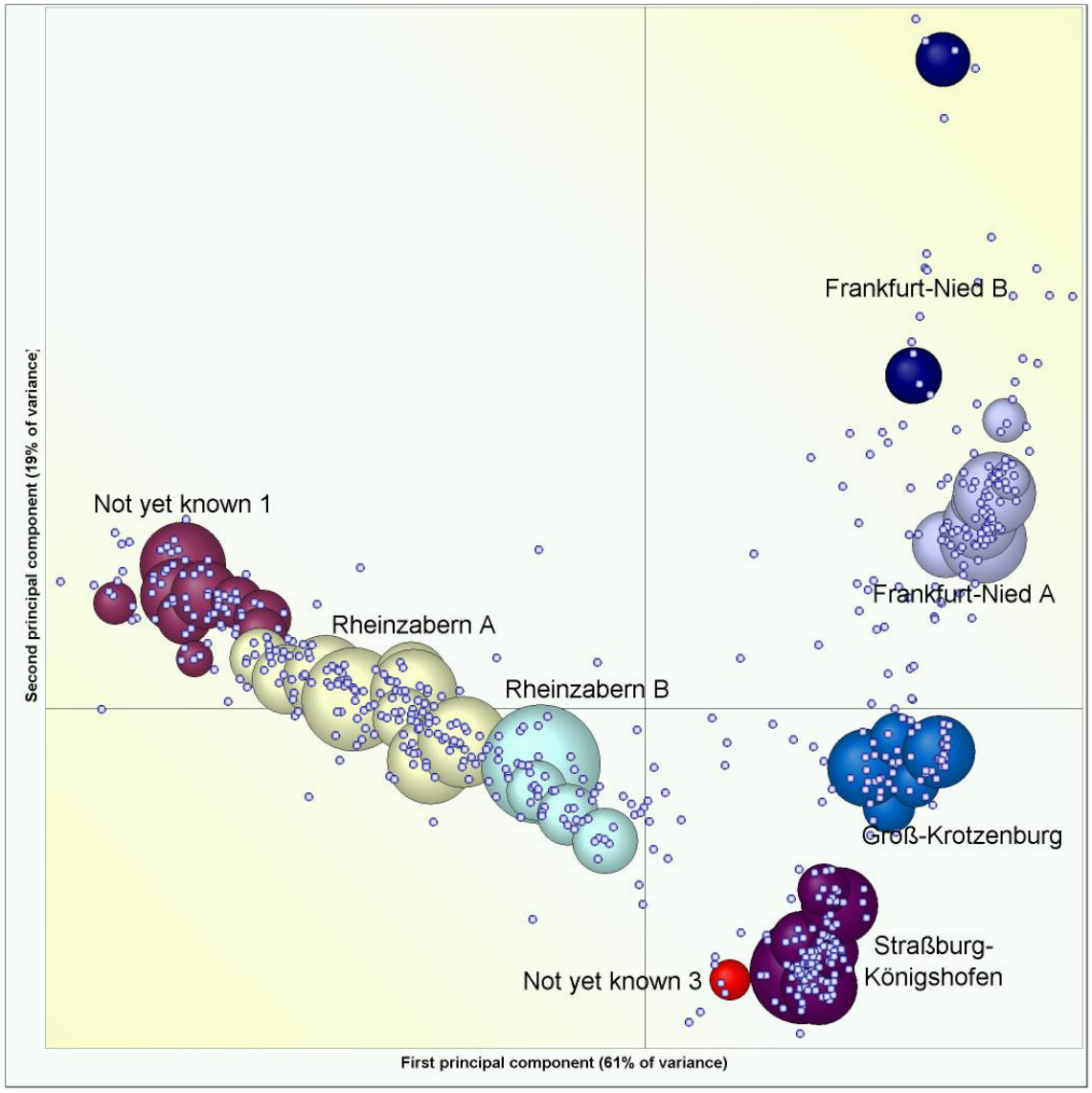
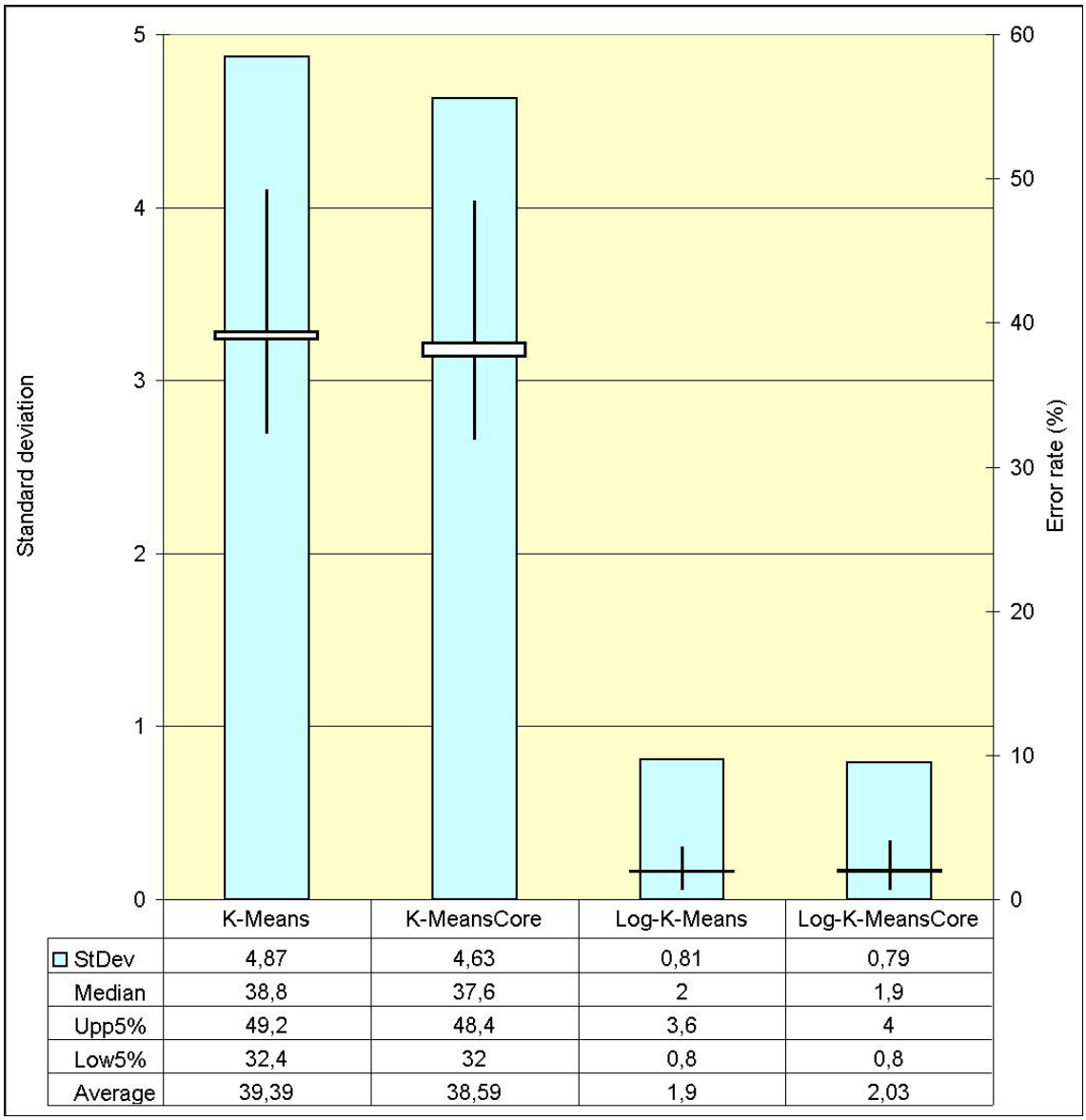


Figure 3: Principal components plot of the final 8 clusters found by core-based clustering (data: see Figure 1).



	K-Means	K-MeansCore	Log-K-Means	Log-K-MeansCore
StDev	4,87	4,63	0,81	0,79
Median	38,8	37,6	2	1,9
Upp5%	49,2	48,4	3,6	4
Low5%	32,4	32	0,8	0,8
Average	39,39	38,59	1,9	2,03

Figure 4: Summary of simulation results of clustering of the *RingNorm*-data.

in order to get so-called micro-clusters that form afterwards cores by applying an appropriate threshold value for minimum size of cores. In the first way, the performance and stability of partitional clustering based on weighted observations is investigated in a simulation study. In the other way, a successful application of hierarchical clustering based on representatives of cores is given in Figure 3. This figure is also for illustration purposes here only. It shows an example of a result of agglomerative hierarchical clustering of cores by the method *LogWard* (criterion (10)) projected into the plane of the first two principal components. Here the size of a bubble (that is, one bubble is one core) is proportional to the value of the logarithmic sum of squares within the corresponding core. Moreover the color of a bubble expresses the final clustering into 8 classes. Figure 3 suggests that there are several well separated clusters. As already mentioned above, the original data consists of 613 observations and comes from archaeometry, where 19 chemical components of Roman tiles coming from the Rhine area are measured by RFA. In Figure 3, all observations are projected into the plane of the first two principal components additionally. The same data was used for preparing Figure 1 above. PCA is a suitable multivariate visualization technique if the variance or log-variance criterion in clustering is used. A generalised PCA based on covariances can take into consideration both weights of observations and weights of variables (Mucha (1992)).

5 Simulation studies

Here the aim is to examine if core-based cluster analysis performs nearly as or better than clustering the original data set. It should be mentioned that the following small samples are drawn from a quite high-dimensional data in respect to the number of observations. Two simple examples of two class data will be investigated here. The number of variables J always equals 20. Generally for each of the examples, 200 artificially generated Gaussian samples of size $I = 300$ are drawn with equal class probabilities. They are analysed in a parallel fashion by traditional and core-based cluster analysis methods. The following simple algorithm has been applied in the case of the latter ones:

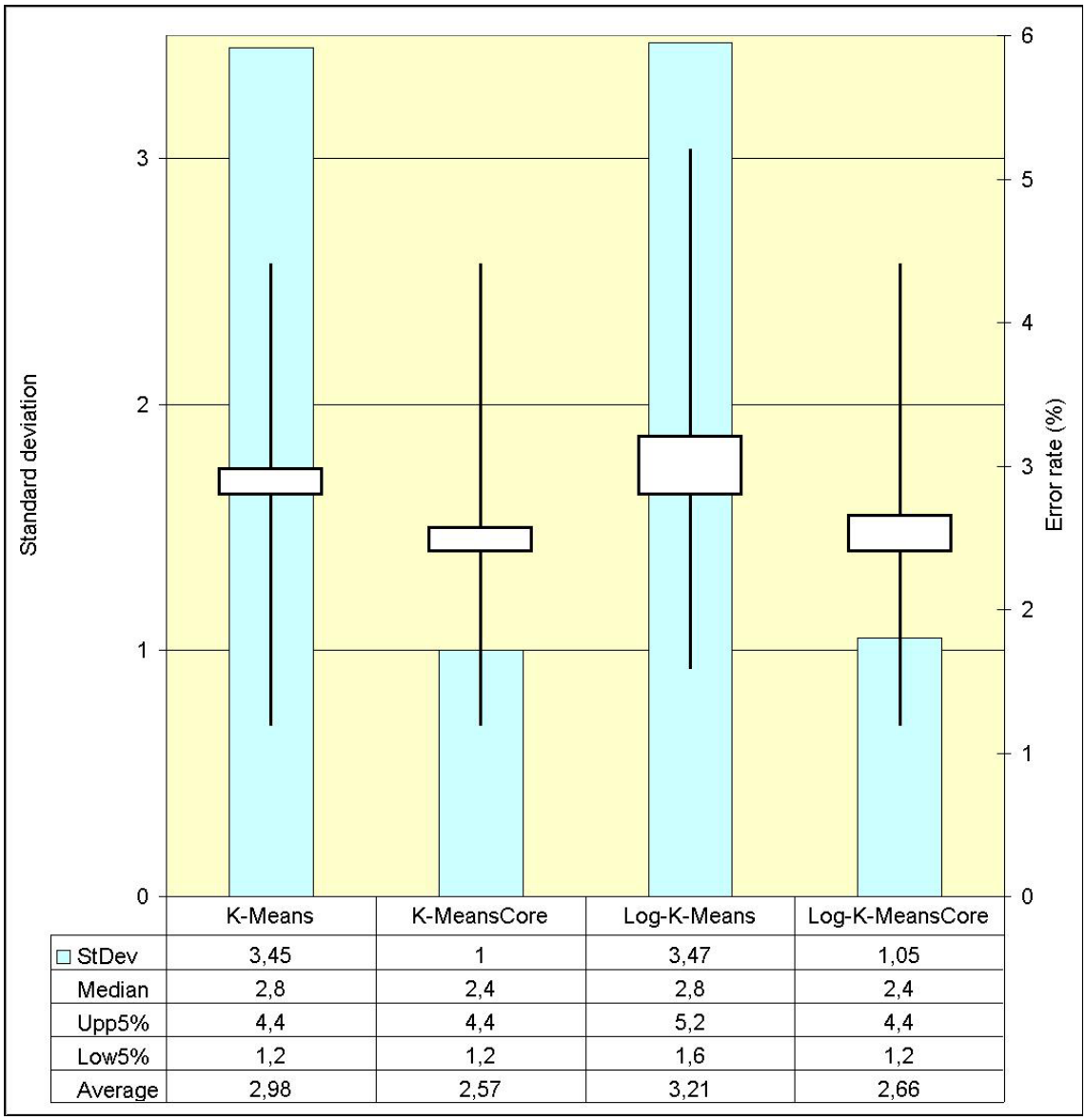
1. Preclustering of all $I = 300$ observations by *K-means* into $L = 50$ micro-clusters A_l,

2. Setting up the set of cores B_q

   $\{B_1, B_2, ..., B_Q\} = \{A_l : \#A_l \geq t, l = 1, 2, ..., L\}$,

   where t is a threshold for the minimum cardinality of a core. Here in the simulation studies below, for example, a threshold parameter $t = 2$ is used.

3. Model-based clustering (*K-means*: sum-of-squares criterion (7), and *Log-K-means*: logarithmic sum-of-squares criterion (10) with taking into account the



	K-Means	K-MeansCore	Log-K-Means	Log-K-MeansCore
StDev	3,45	1	3,47	1,05
Median	2,8	2,4	2,8	2,4
Upp5%	4,4	4,4	5,2	4,4
Low5%	1,2	1,2	1,6	1,2
Average	2,98	2,57	3,21	2,66

Figure 5: Summary of simulation results of clustering of the *TwoNorm*-data.

weights

$$m_i = \begin{cases} 1 & \text{if } \mathbf{x}_i \in B_q, q = 1, 2, ..., Q \\ \varepsilon & \text{otherwise} \end{cases}$$

Herein $\varepsilon \geq 0$ is usually nearly 0 (quasi-zero). That is, observations out of cores are downweighted (or can be even discarded by setting $\varepsilon = 0$).

Indeed, this is a hard rejection of outlying observations. Because of computational simplicity ε is chosen to be a quite small positive value (for instance, $\varepsilon = 1.E-15$) in order to assign the outlying observations to their nearest distance cluster. However, these outliers with a quasi-zero weight don't affect the clustering result. The misclassification rate measures the performance of the clustering methods. In practical applications however, where usually nothing about the supposed classes are known beforehand, other measures for performance and stability of clustering methods have to be used (Rand (1971), Hubert and Arabie (1985), and Mucha (1992, 1995)).

Example *RingNorm* (after Breiman (1996))

As already mentioned above, clustering of two class data will be investigated. Class 1 is multivariate normal with mean zero and covariance matrix 4 times the identity. Class 2 has unit covariance matrix and mean $(a, a, ..., a)$ with $a = (1/J)^{1/2}$. The appropriate clustering method for this kind of data is *Log-K-means* which minimizes criterion (6). Clustering of this kind of data is a hard problem for *K-means*, which minimizes criterion (3).

Figure 4 shows both the most important numerical results concerning the misclassification rate (in percentages) and a corresponding graphical representation of these univariate statistics. The reading of this figure is as follows. The axis at the left hand side and the bars in the graphic are assigned to the standard deviation of errors, whereas the axis at the right hand side and the box-whisker-plots are linked to all other statistics. One can see that the core-based clustering methods *K-meansCore* and *Log-K-meansCore* perform similar as the traditional ones.

Example *TwoNorm*

This two class data is also taken from Breiman (1996), but it is slightly changed. Each class is drawn from a multivariate normal distribution with unit covariance matrix. Class 1 has mean $(a, a, ..., a)$ and class 2 has mean $(-a, -a, ..., -a)$ with $a = (2/J)^{1/2}$. In order to investigate the influence of outliers modifications of the original *TwoNorm* data model are made. One out of the 150 observations of each class is randomly generated with 4 times standard deviation. As a consequence there is a high probability that the data contains at least one outlier. The most appropriate clustering method is *K-means*. For this kind of data however, the more general *Log-K-means* is also a suitable technique. It performs like *K-means* here.

Figure 5 shows both the most important numerical results concerning the misclassification rate (in percentages) and a corresponding graphical representation of these

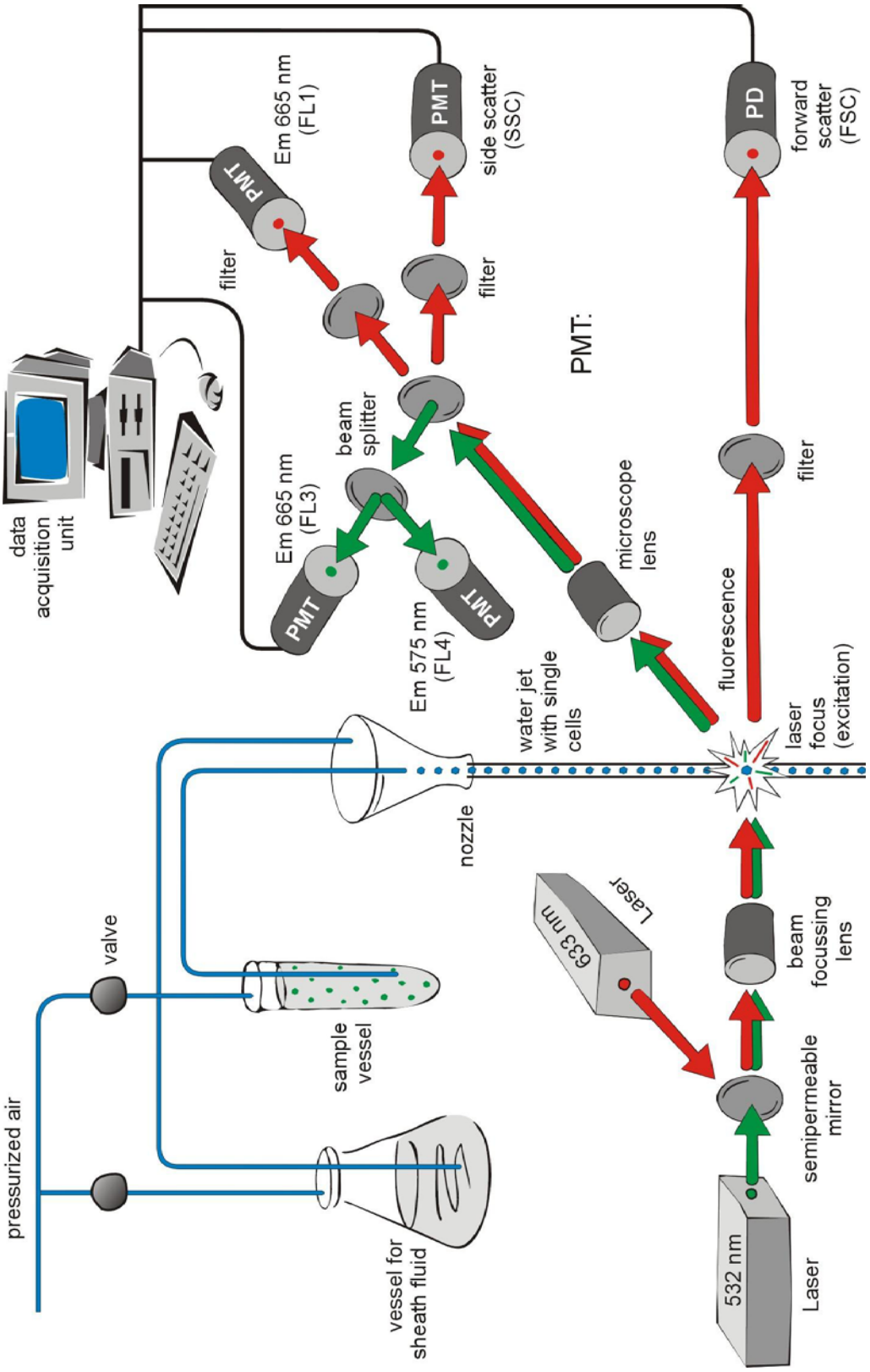


Figure 6: Schematic view of a flow cytometer

univariate statistics. The reading of this figure is like as the one of Figure 4. Obviously, the core-based clustering methods perform slightly better (0.5% in average error rate) than the traditional ones. But what is of much more importance is that the standard deviation of error rates as a measure of stability decreases to less than a third of the one of the traditional methods. The reason for this may be that the influence of the outliers is taken away by core-based clustering.

6 Application in ecology

6.1 Introduction into the problem

It is generally accepted that ecosystems are networks with a high degree of backcoupling interactions. Therefore many attempts to estimate the degree of ecosystems health, i.e. its degree of integrity are based on food web structures. For a deep discussion of integrity and health of ecosystems, see Barkmann (2001). Measures of integrity, based on food web structures need a careful taxonomic investigations. In food webs phytoplankton plays a basic role as it is mainly responsible for the utilisation of light and minerals within the food web. Therefore it may be a good strategy to start integrity measures by means of a quantitative analysis of abundance data of different types of algae. However there are severe and still not solved taxonomic problems in order to use those data routinely. In contrary to the taxonomic approach Steinberg et al. (1999) proposed an ataxonomic method, which is based on a pure analysis of size classes. Even a semiempirical theoretical frame was given in stating that there an "energy equation" and a "continuity condition" should hold (Steinberg et al. 1999). Here a more pragmatic approach is followed, which may be a basis for later theoretical investigations in the sense of integrity measures: If namely integrity measures should be used, a methodology has to be established, which also is suitable for monitoring: This demands for an (semi-) automatic measurement and an appropriate statistical evaluation.

The monitoring of phytoplankton by microscope is very time- and work-consuming. Flow cytometry provides the opportunity to investigate algae communities in a semiautomatic way (Hofstraat et al. 1994). Two different kinds of information are obtained: (1) the number of cells (here algae) per unit of sample-volume and (2) optical characteristics of each cell. These are parameters of light scatter and of fluorescence, in our study depending on the composition of photosynthetic pigments. The latter makes it possible to differentiate between different pigment-groups. In our example, the output consists of 5 parameters. Note that these pigment-groups are not identical to a taxonomic classification, but can be semiempirically related to different algae groups (Hoek 1993). To identify the different pigment-groups and to count the number of organisms belonging to each group, classes are determined manually in most cases. Anyway as 5 parameters should be taken into consideration, one has to handle a multidimensional space, where the procedure of building clusters manually might be difficult or even impossible. Thus, the technique of flow

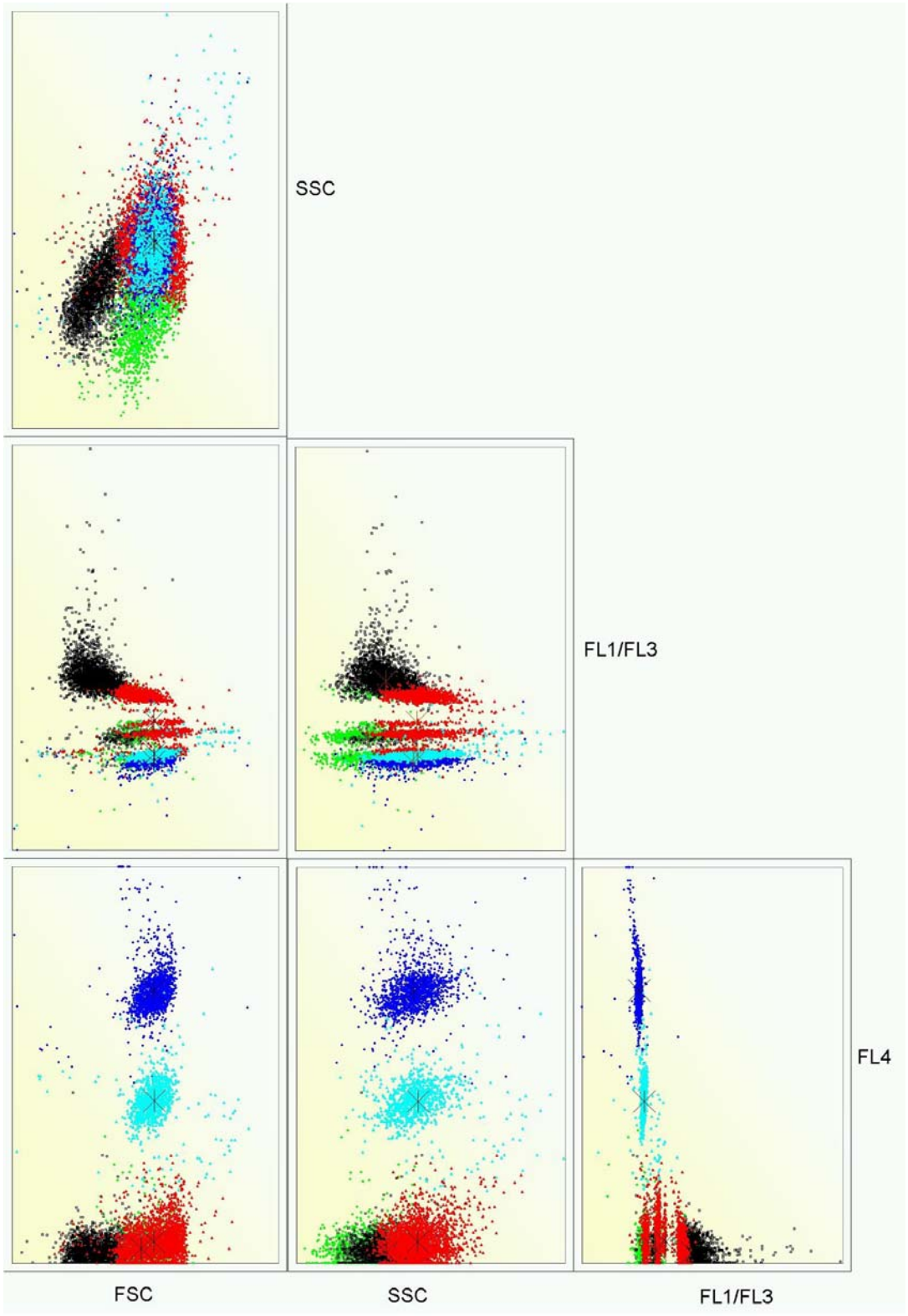


Figure 7: Scatterplot matrix of four from nine clusters that are found in a first crude clustering step by the fast K-means method. Cluster centroids are marked by big black crosses.

Parameter (short cut)	Excitation wavelength [λ]	Detected emission [λ]	Description
FSC	633 nm (red)	$>$ 610 nm	Size of the cells
SSC	633 nm (red)	$>$ 610 nm	Structure of the surface of the cells
FL1	633 nm (red)	$>$ 665 nm	Pigment chlorophyll a
FL3	532 nm (green)	$>$ 665 nm	Pigment chlorophyll a
FL4	532 nm (green)	575 nm	Pigment phycoerythrin

Table 1: Configuration of the flow cytometer

cytometry as a routine in phytoplankton monitoring and the evaluation of the data demands for an automatic data analysis by suitable mathematical tools like cluster analysis. For a comparative analysis between samples of different freshwater systems and of different seasons of the year it is also convenient, to describe each pigment-group as a Gaussian-function.

6.2 Some background of the data gathering

Flow cytometry allows single cell analysis based on information about light scatter and fluorescence. A simplified construction scheme of a flow cytometer is depicted in Figure 6. The suspended cells in the sample vessel are transported to a nozzle. In the nozzle they are surrounded by a sheath fluid and hydrodynamically focused as well as separated from each other. With a free flowing water jet cell by cell intersects two laser beams and send out pulses of scattered light and fluorescence. Light detectors, photomultiplier tubes and photodiodes, are transforming the optical pulses of each cell into electronic signals. The signals are stored as list mode files in a computer, meaning that fore every single cell up to 6 optical parameters can be recorded. As parameters of scattered light there is the forward scatter (FSC) as a rough measurement for the cell sizes and the side scatter (SSC) from which information about the structure of the surface of the cell can be deduced. In case of the fluorescence up to four different signals can be detected according to the optical attributes of the cells under investigation. In the present study a FACStarPlus (trademark of Becton Dickinson) with a red and a green laser, wavelength 633nm and 532nm respectively, has been used. The opening of the nozzle was 100μm of size, thus cells up to a diameter of about 60μm can be measured. Altogether the optical pulses of two parameter for scattered light and only three of fluorescence has been recorded (Table 1), so one fluorescence detector of the machine was not in use. The trigger parameter discriminating if a particle will be recorded or not, was FSC, thus the optical signals of all particles (algae and other particles) has been stored. For further explanation see also below.

With flow cytometry pelagic algae can be easily measured because they are already suspended in water where they live separated from each other as single cells or colonies. Because of the auto-fluorescence of their photosynthetic pigments no

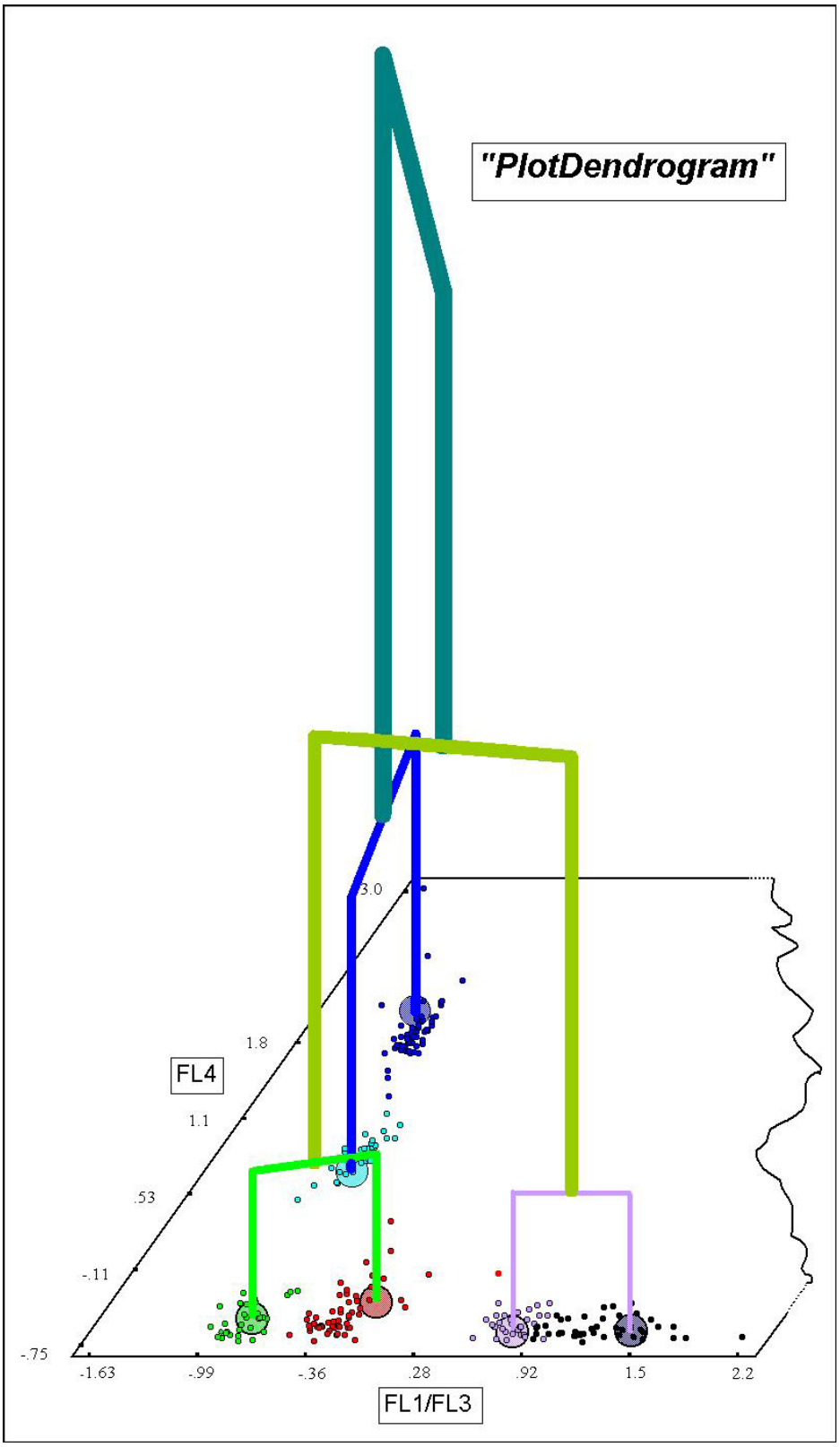


Figure 8: Graphical presentation of the result of hierarchical clustering (*Ward* method) of a randomly drawn sample of size 250 observations (data: phytoplankton data).

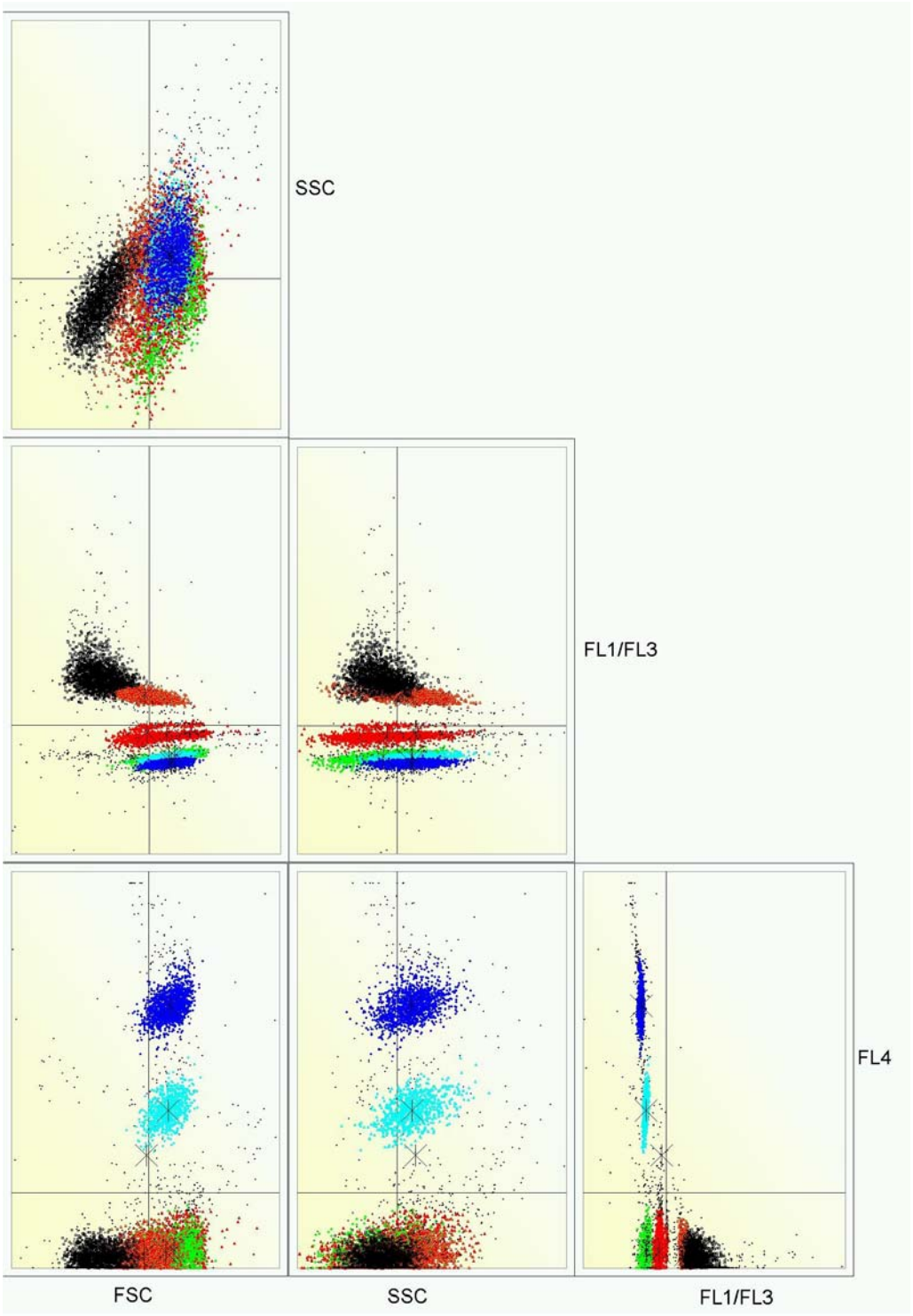


Figure 9: Result of model-based Gaussian clustering of 8786 observations.

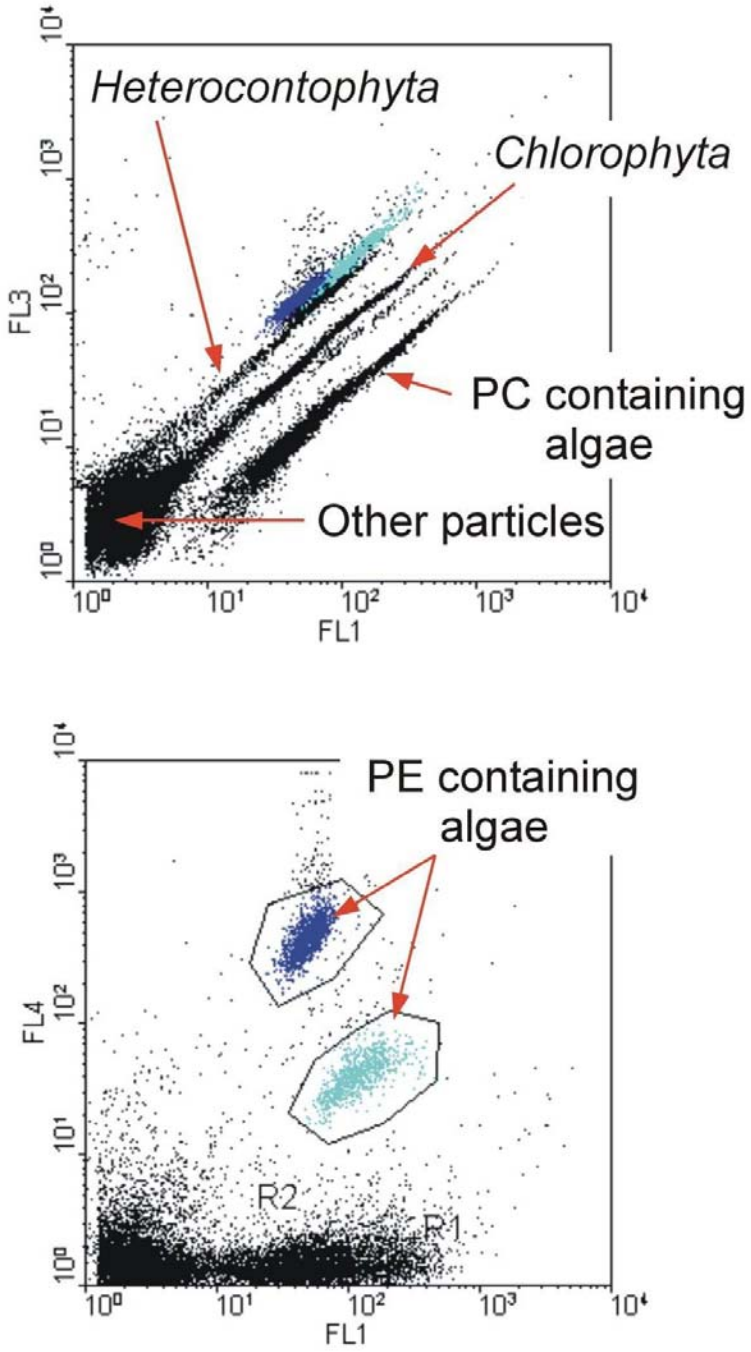


Figure 10: Flow cytometric pigment fluorescence ratio analysis

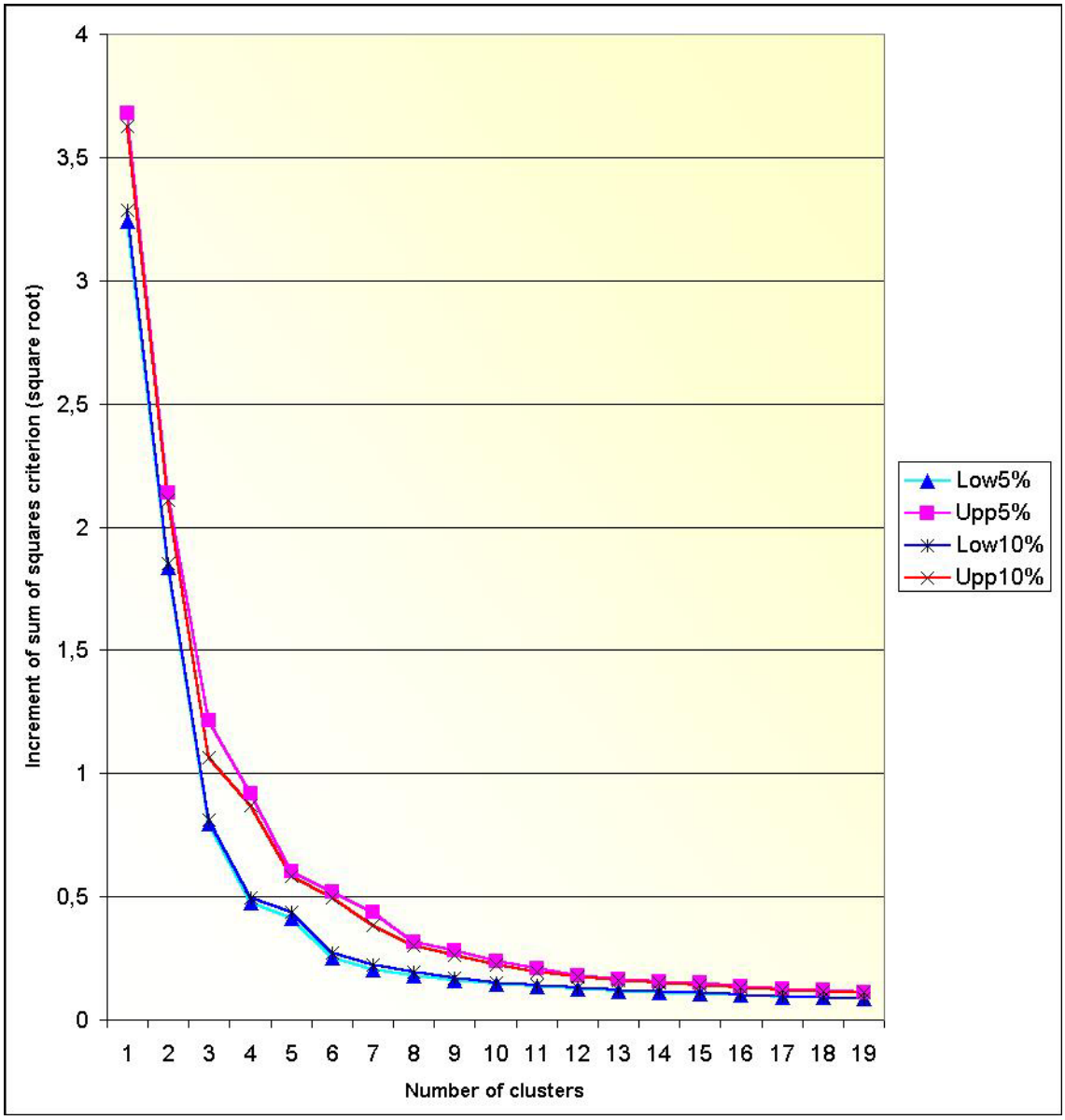


Figure 11: Hierarchical clustering: criterion values versus number of clusters

staining is necessary. From previous investigations it is known, that the fixation of algae cells can causes changes in their fluorescence characteristics (e.g. Lepesteur et al. 1993, Hall 1991, Vaulot et al. 1989). To avoid such troublesome side-effects, the probe has been measured within not more than one day and without any additional manipulation like staining or fixation. Following the approach of Steinberg et al. (1998) due to the ratio of different pigments the excitation with red and green laser light allows the differentiation of four pigment-groups. Algae containing as an accessory pigment: (1) phycocyanin (PC) concerning the taxonomic group of *cyanophyta*, (2) phycoerythrin (PE) concerning the taxonomic groups of *cyanophyta* and *cryptophyta*, (3) carotene concerning the taxonomic group of *heterocontophyta* and (4) algae with only a very low ratio between carotene and chlorophyll a (CHLa) concerning the taxonomic group of *chlorophyta* (Figure 10). Thus a semi-taxomonical differentiation between groups of algae is possible.

6.3 The data under investigation - a snapshot

In the present study a sample of lake Müggelsee is analysed. The sample was taken on the 8th of July 2002 from the upper water layer (0.5 – 3.5m). Altogether a sample volume of 1.178 ml has been analysed with flow cytometry and the five optical signals of altogether 21778 particles has been stored in the computer. The left side of Figure 10 depicts the diagram of chlorophyll a (CHLa) fluorescence $\lambda > 665$nm when excited with the two wavelengths: $\lambda = 633$nm (red) and $\lambda = 532$nm (green). Due to a transfer of energy from accessory pigments to CHLa, groups of different pigment compositions occur as stripes in the diagonal. To be able to identify the different pigment-groups, a procedure for the calibration of the cytometer has been developed, which is based on the measurement of cultured algae. The calibration makes sure, that the *chlorophyta*, containing mainly chlorophyll a, occurs as the diagonal in the centre of the dot-plot (Figure 10). Algae with high amounts of the pigment carotene (*heterokontophyta*) and with phycoerithrin (*cyanophyta*, *cryptophyta*), are located above the *chlorophyta*. Algae with pycocyanin (*cyanophyta*) are located below them. As written above in addition the accessory pigment PE can be detected separately when fluorescence emission at $\lambda = 575$nm is measured (lower dot-plot in Figure 10).

Just by optical inspection it is easy to see, that the identification of algae requires more than the three parameters used in the two dot-plots of Figure 10. Looking for example at the lower diagram in Figure 10 one can see, that there is no clear differentiation between the broad cloud of dots at the bottom of the left corner and the rising band of *chlorophyta*. From previous studies we know, that the broad cloud can be composed of algae of very small size or only little amounts of pigments, as well as detritus (dead organic material) or inorganic particles. For the purpose of phytoplankton monitoring it is of major interest, to differentiate between algae and other particles. Therefore different powerful mathematical tools has been applied, as written in the previous Sections.

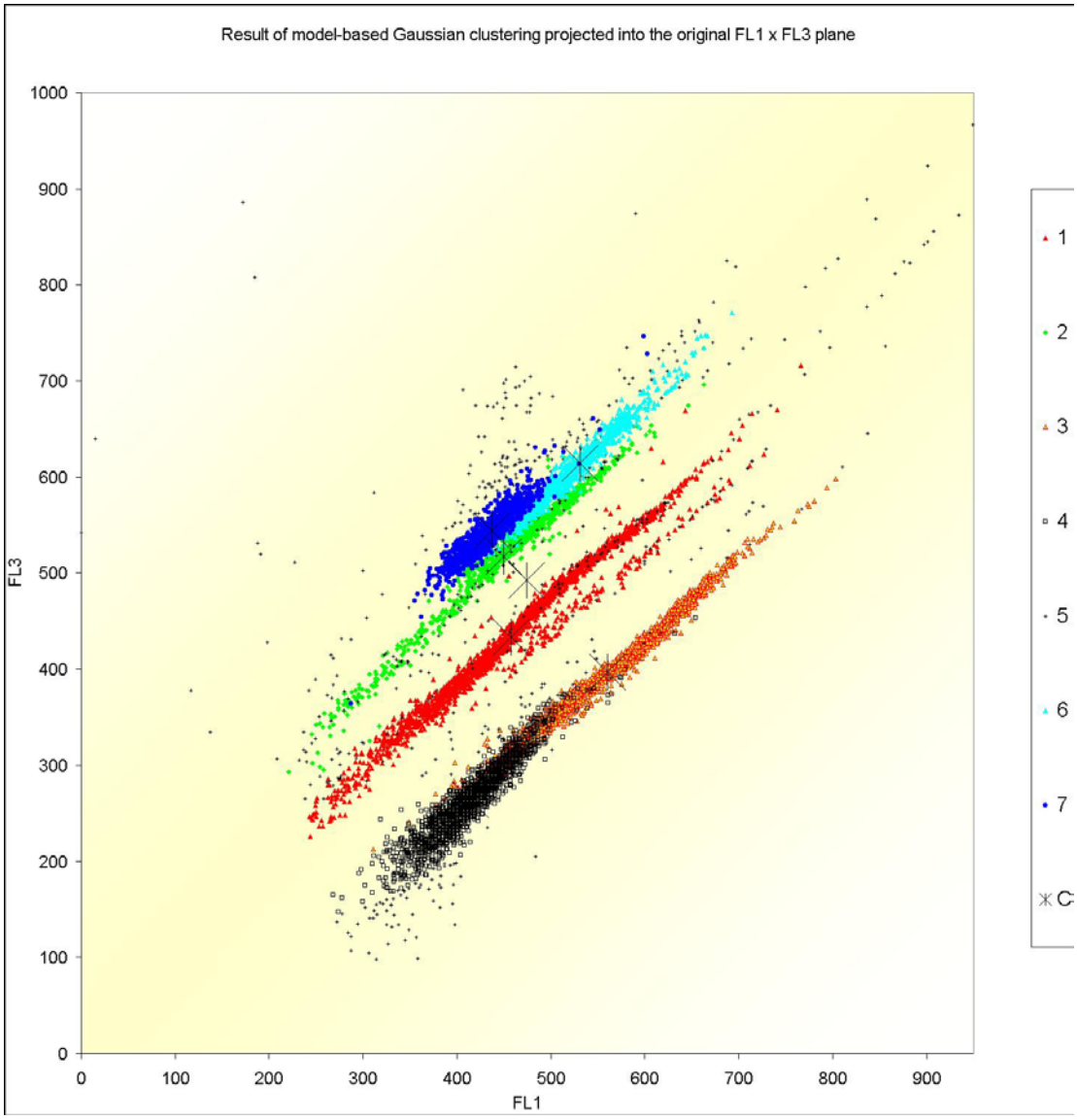


Figure 12: Final result of model-based Gaussian clustering of 8786 observations.

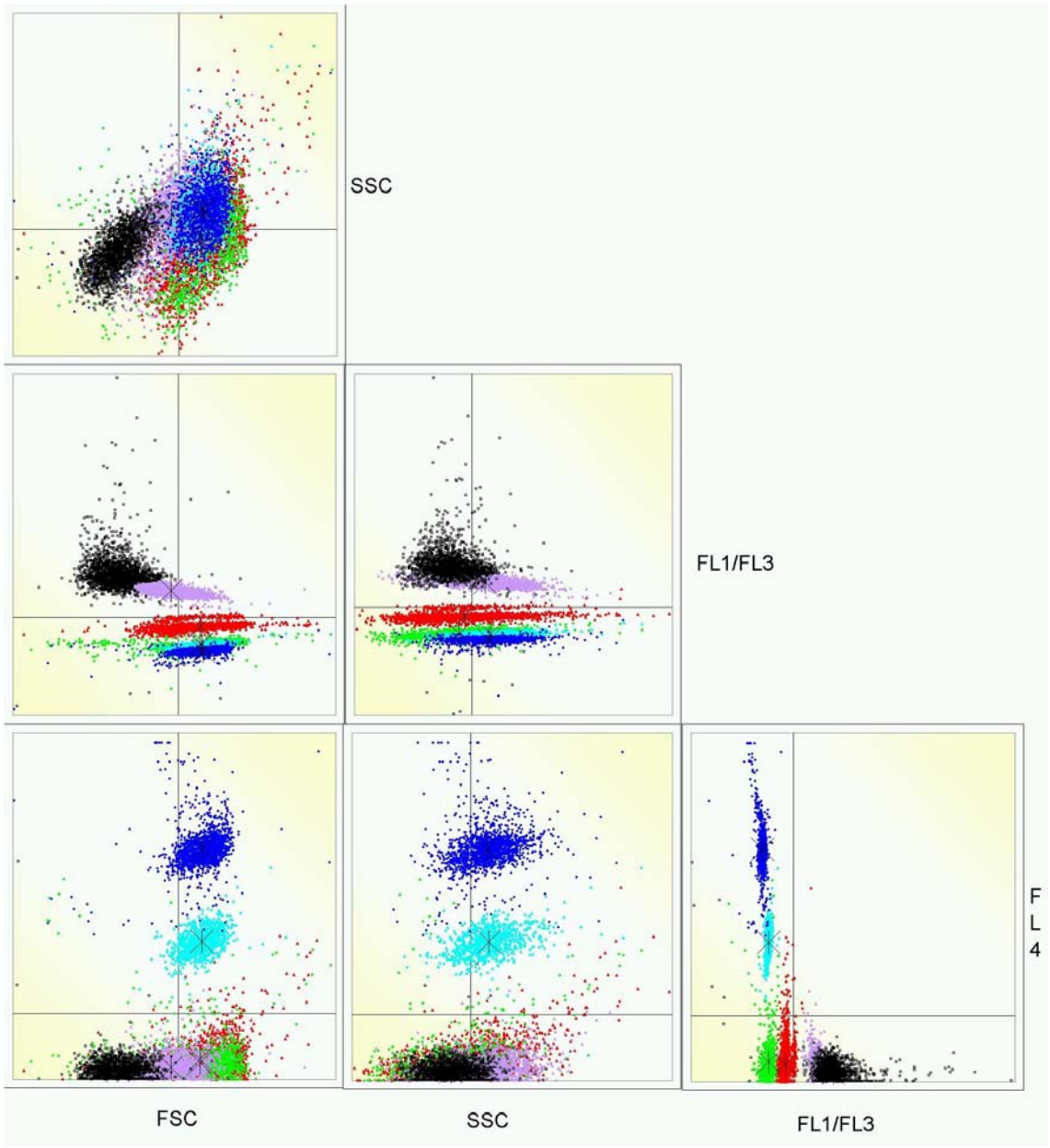


Figure 13: Scatterplot matrix of the final result of model-based Gaussian clustering

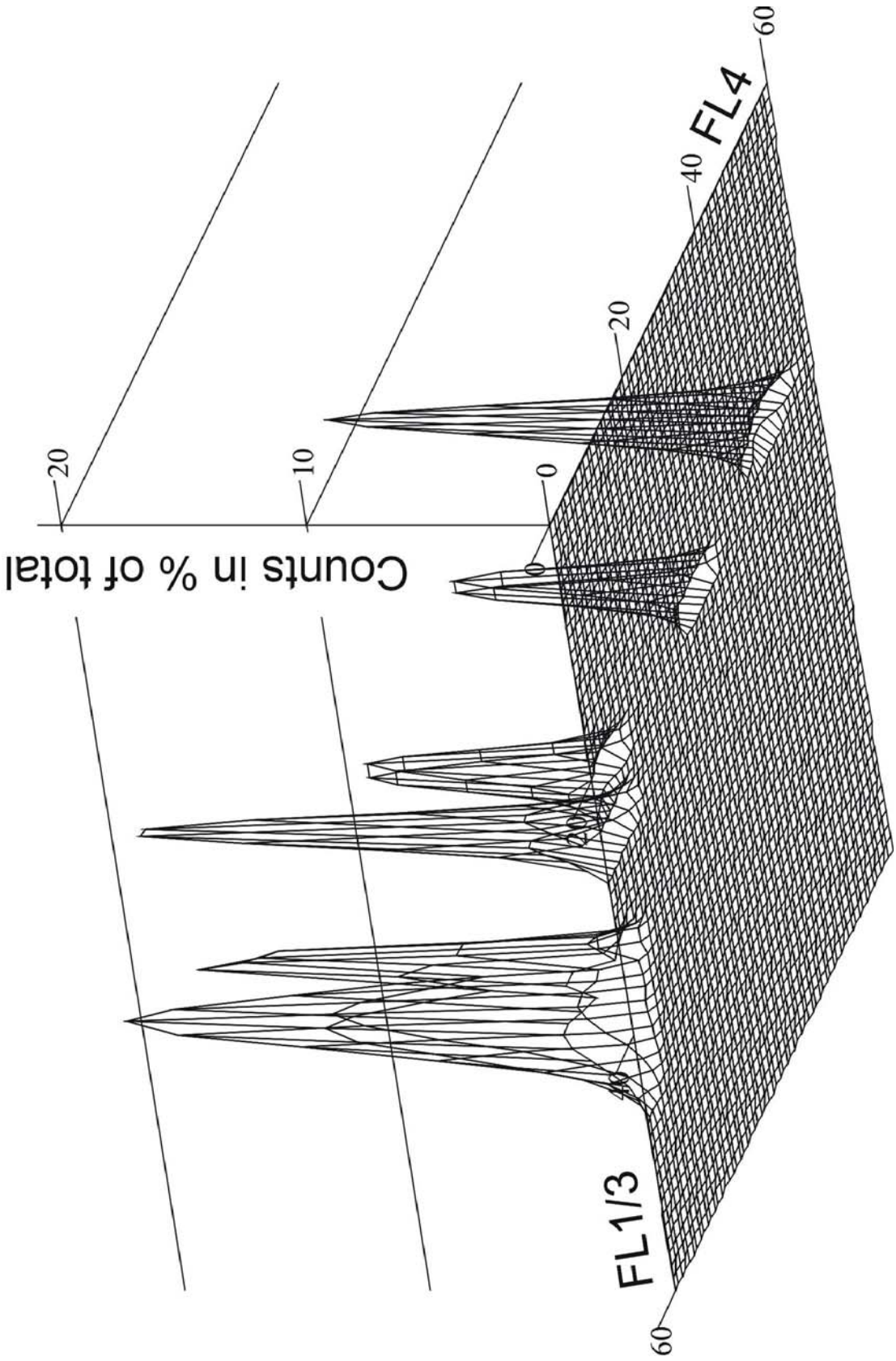


Figure 14: Pigment-groups depicted as Gaussian-functions

6.4 Results of model-based clustering

From an ecological point of view both information, the number and kind of pigment groups, and the number of organisms in every group, are of interest. To extract these information from the original data (list mode file), several mathematical tools has been used sequently. In a first crude clustering step, the original data has been clustered with the fast K-means approach. It is based on the simplest model (3) in model-based Gaussian clustering. The original variables SSC, FSC, FL1, FL3, and FL4 are used without any data preprocessing or standardisation step. Here the main focus is on dividing "interesting" particles (algae) from particles from other origin. By this clustering procedure, 8786 particles out of the 21778 events in the original data set, has been identified as cell-groups with similar optical characteristics. Figure 7 shows four interesting clusters that consist of nearly nine thousand observations. In conformity with the forthcoming results, the variables SSC, FSC, FL1/FL3 (for explanation, see below), and FL4 are used in this scatterplot matrix.

In the following these 8786 cells are declared as algae. As we are mainly interested in the pigment-bands, the ratio of FL1 and FL3 is of great importance. For that reason, the ratio of FL1 and FL3 is calculated as a new variable for further calculations. It is denoted FL1/FL3, or shortly FL1/3. From the statistical point of view this leads to a better match of the data to the Gaussian model than by using the original variables FL1 and FL3 (see the diagram at the top of Figure 10). As a consequence standardization of data is now necessary in the case of clustering based on variance criteria. This simple Gaussian model was already used in the first step.

In a second step, hierarchical clustering has been used for two main reasons. First, usually one get a good initial solution for a subsequent more complex Gaussian model. Second, one can try to find an appropriate number of clusters for further analysis. However, a better but much more expensive way for determining the number of clusters would be a simulation study based on measures for comparing clustering results (Rand (1971), Hubert and Arabie (1985), and Mucha (1992, 1995)). Here, the number of clusters is simply chosen from a simulation study based on 200 randomly drawn samples of size 250 observations from the nearly nine thousand observations. For each sample, the hierarchical Ward's method is carried out. By random sampling one get a feeling about the variation of the criterion values in respect to the number of clusters. A random sample of 250 observations seems to be sufficiently large in respect to the 4 dimensions (variables: SSC, FSC, FL1/FL3, and FL4), so that the distribution of these selected observations should reflect the distribution of observations in the entire set. Figure 8 shows the result of hierarchical clustering of such a sample, and Figure 2 gives a density estimate of their pairwise distances. Random sampling is most "natural" because the higher the density of a region the more observations are randomly drawn from the region (for a comparison, see Figure 7).

Figure 11 shows some statistics of the obtained criterion values regarding to the number of clusters. Because one is interested in low criterion values the curves, *Low5%* and *Low10%* are important for choosing the number of clusters. Obviously,

going up in the hierarchy from 6 to 5 clusters, a high increase of the criterion value can be observed. Therefore the number of six clusters has been chosen for further analysis.

In the third step, the most complex model-based Gaussian clustering is carried out. As an appropriate starting point for the model, the result of hierarchical clustering on a sample can be recommended. By doing so, the determinant criterion (8) leads to some well separated clusters, and as expected, to an additional cluster with a quite flat density collecting all observations from sparse regions (Figure 9). Banfield and Raftery (1993) recommended to assume that there is beside Gaussian models a Poisson process with an unknown intensity parameter for all these observations that do not follow the Gaussian pattern. Here another way out is proposed that is based on downweighting observations coming from sparse regions.

In the two tiny simulation studies above (Section 5), the investigations are promising with regard to improvement of performance and stability. There are many ways to find observations that come from sparce regions. Here the K-means method is applied in a similar fashion as described in Section 5 with the following modifications. The number of microclusters $L = 250$ is used and a threshold $t = 10$ is chosen, i.e. the minimum cardinality of a core is 10. Taking into account that the choice of t and L as well as the random initial partition can affect the results, the K-means clustering is carried out five times. An observation is downweighted only if it is at least two times of at most five times out of cores. Figure 12 shows the final result of weighted Gaussian clustering, depicted as a dot-plot of the two originally measured parameters of chlorophyll fluorescence FL1 and FL3. In this example, there is one group of *chlorophyta* (red), one group of *heterokontophyta* (green), two different kind of algae-groups containing phycoerithrin, PE1 (purple) and PE2 (black) and two different groups of algae with phycocyanin, PC1 (dark blue) and PC2 (light blue). Table 3 gives the parameter characteristics of the clusters in an overview. Note that the calculated parameter FL1/3 can be used as the qualitative numerical expression of the pigment ratio of the algae. Figure 13 shows the scatterplot matrix of the final result of weighted Gaussian clustering that corresponds entirely to the result that is presented in Figure 12.

A convenient way to describe the different clusters might be the expression as a Gaussian-function, as for each cluster the position with respect to the fluorescence parameters FL1/3 and FL4 can be determined exactly as well as its altitude, here in percentage of the total count. In the field of phytoplankton monitoring this approach yield some advantages, as now pigment-groups can be described in a precise mathematical way. Thus the comparison of probes of different locations and of different seasons of the year can be done easily as well as the development of structural indicators for an ecological assessment of the freshwater systems. Figure 14 shows the pigment-groups of the probe of Lake Müggelsee as Gaussian-functions.

As written above in most cases phytoplankton is counted and classified by microscope. For the probe of Lake Müggelsee both methods, microscopy and flow cytometry has been performed. So results obtained by the mathematical clustering of the

Class	Colour in graphics	Pigment group	Centroids of classes FSC	SSC	FL1	FL3	FL1/3	FL4	Total count
1	Red	Chlorophyll	483	359	464	439	1.05	51	1927
2	Green	Carotene	486	338	450	517	0.86	53	1022
3	Dark blue	PC1	405	422	552	390	1.42	35	1539
4	Light blue	PC2	271	321	409	258	1.60	26	1839
5	Purple	PE1	485	440	529	613	0.86	410	850
6	Black	PE2	490	436	439	549	0.79	687	1609
Sum									8786

Table 2: Parameter characteristics of the clusters (Phycocyanin: PC1, PC2; Phycoerithrin: PE1, PE2)

Pigment group	Cytometry [Cells * ml-1]	Microscope [Cells * ml-1]
Total number	5894	5640
Chlorophyta	1635	765
PC and PE containing algae	3392	3812
Heterokontophyta	867	887

Table 3: Comparison of data from microscope and flow cytometry

cytometry data can be compared qualitatively and quantitatively with those from microscopy. Anyway, the comparison is restricted by some technical and methodological reasons. Cells of very small size (picoplankton) in the range of 0.2 to 2μm (e.g. Sommer (1994)) and cells of a size larger than 60 μm are excluded. For that reason class 4 (PC2) in Table 2 has to be removed. The centroid of the FSC of this class is approximately half of the numerical value of all other classes. As the FSC is a parameter for the cell size, there is a high probability, that this class subsume mainly picoplanktic algae. This assumption is also confirmed by the measurement of cultured picoplanktic algae. Beyond this, the clusters of PC and PC containing algae has to be added to one group, because the taxonomic determination of the cells via microscopy, do not allows for a pigment-based differentiation. So the comparison between the microscopic count and the clustering of cytometry data is performed with only three different pigment-groups. The *chlorophyta*, phycocyanin (PC) and phycoerithrin (PE) containing algae, and *heterokontophyta* (Table 3).

For a competent comparison of the counts obtained by the two different methods of microscopy and flow cytometry, one has to take the statistical standard error of each method into account. In the case of microscopy, the standard error is correlated to the number of counted cells (Tümpling and Friedrich (1999)). Note, that the number of cells per sample volume for each group, is calculated from representative counts in some of the stripes of the sedimentation chamber according to Utermöhl (1958) using an inverted microscope. In case of the probe from Lake Müggelsee

for example, in the group of *heterokontophyta* 180 cells has been counted, in the group of PC/PE-algae it has been 308 cells and for the *chlorophyta* only 51 cells. Following the suggestion in Tümpling and Friedrich (1999), for the total count a maximal standard error of 7-10%, for the *heterokontophyta* 20%, for the PC/PE group 10% and for the *chlorophyta* even 60% is recommended as possible. For the cytometric counts of phytoplankton, up to now no statistical based estimation of the standard error is available. Anyway, first tests confirm, that a standard error of about 10% will be a conservative estimation. Figure 16 shows the comparison of the counts of both methods, when the assumptions of the maximal standard errors mentioned above are taken into account.

Figure 16 shows, that only in the group of the *chlorophyta* the counts differ out of the range of the standard errors. For the *heterokontophyta* and for the PC/PE-group the counts are in good agreement. Also the counts of the whole probes differ only within the range of the standard errors. Thus in general the cluster analysis has been successful.

7 Conclusions

The principle of weighting of observations is a key idea for handling cores and outliers. Often the stability of clustering methods like *K-means* or *Log-K-means* can be improved. However, the problem of choosing appropriate thresholds for establishing cores remains under investigation. Even for a fixed and arbitrary threshold $t = 2$, the simulation results are promising, see especially Figure 5. The threshold value $t = 1$ followed by *Ward's* method based on centroids of L cores give exactly the same clustering result as *Ward's* clustering of the original (huge) data matrix on the understanding that the latter one comes at one stage (partition) of the amalgamation process exactly over the partition of cores $P(I, L)$. Moreover, the chosen number of micro-clusters as well as the preclustering method itself affect the result of core-based clustering. The examples and simulations are figured out by the prototype-software `ClusCorr98`®. This software contains a set of statistical tools for data exploration with emphasis on model-based (adaptive) clustering and multivariate graphical visualizations. It runs under Microsoft Windows.

In general the application of a sequence of cluster analysis to the data of algae obtained by flow cytometry has been successful. Nevertheless there are some topics to discuss: First of all it has to be proved, that the first step of clustering with the fast K-means approach yields suitable results. By this procedure algae should be differentiated from other particles. The comparison with the microscopic count shows, that the differentiation in the group of *chlorophyta* might be not precise. Possible improvements of the method might be the clustering when the calculated parameter FL1/3 is added or the weighting of parameters. Furthermore the effect of choosing another number of start clusters has to be tested. Anyway as only a snapshot has been analysed here, a final decision if this method is successful has to

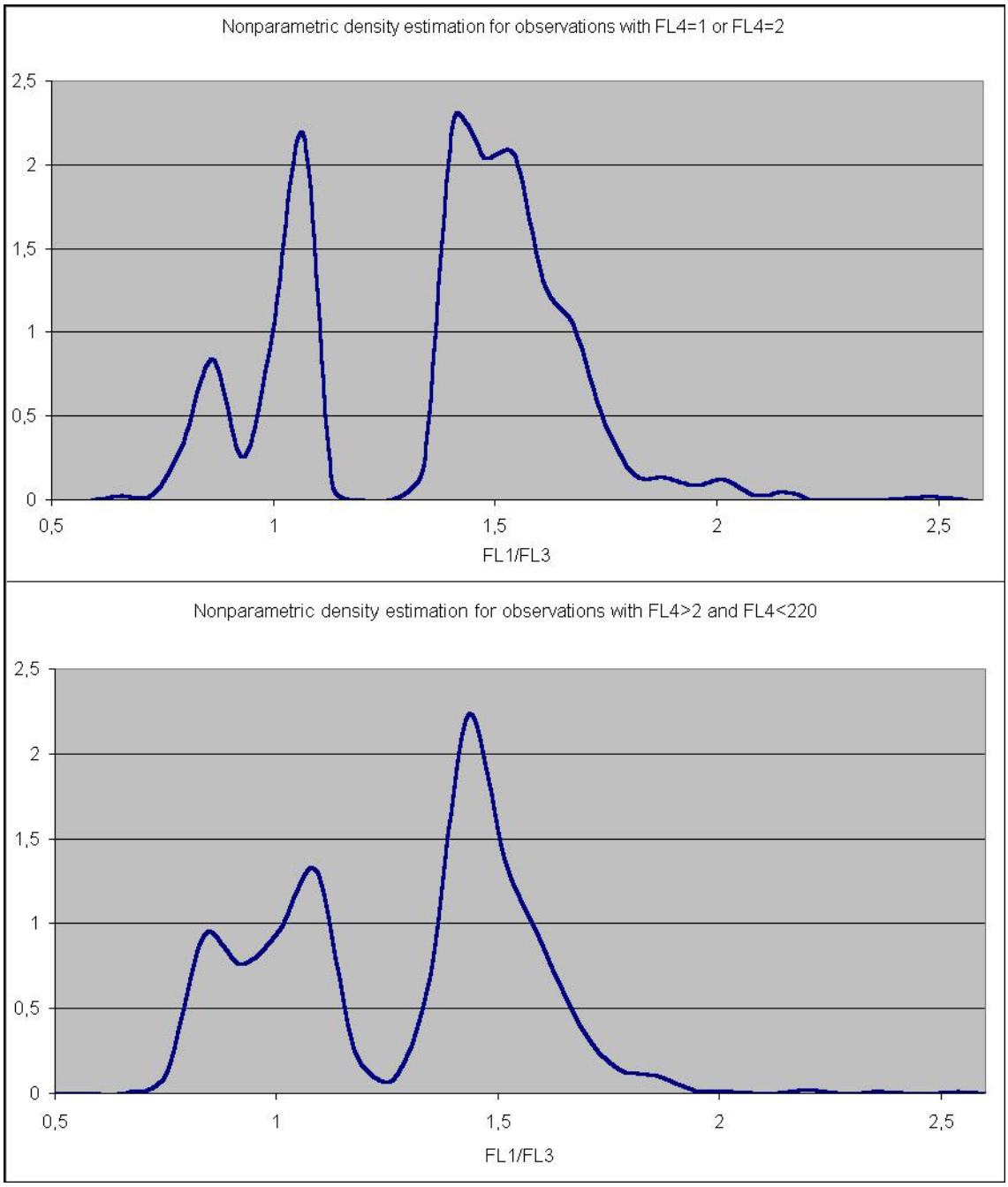


Figure 15: Nonparametric density estimations of variable FL1/FL3 under differnt restriction on FL4.

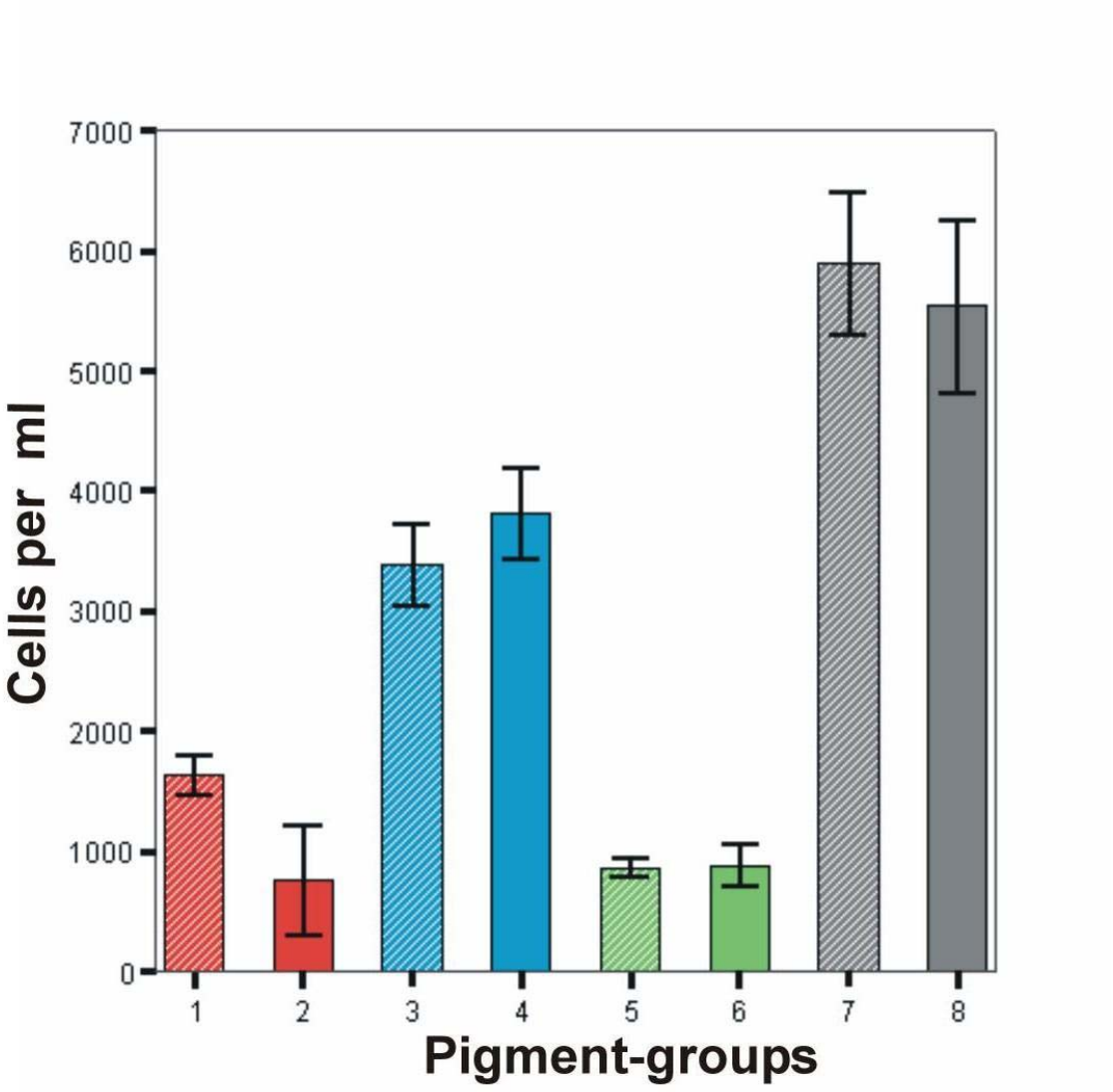


Figure 16: Count from microscopy and flow cytometry with standard error. Groups 1 and 2 = *Chlorophyta*, groups 3 and 4 = PE/PC containing algae, groups 5 and 6 = *heterokontophyta*, groups 7 and 8 = total counts. Striped boxes = cytometric counts.

be proved by analysing more probes.

In the second step, a hierarchical clustering was used to find an appropriate number of clusters for the further investigation. Here also it has to be tested, if the choice of a higher number of clusters will improve the final result. Figure 16 shows that the group of *chlorophyta* could be divided in at least two subgroups. Figure 12 confirms this impression.

Considering the variable FL4, for example, there is a very high number of observations near the detection limit (FL4=1 and FL4=2) surrounded by a comparatively sparce region (FL4>2 and FL4<15). This conspicuity disturbs the assumption of normality (see for instance, Figure 13). However, the estimates of some local densities looks similar (Figure 15). For this troublesome effect there might be two (or more) possible explanations: The optical filters in the flow cytometer do not suspend other fluorescence emissions but of wavelength of $\lambda = 575$ nm ($\pm$ 10 nm) in a proper way. Some of the algae cells of the groups mentioned above has got little amounts of the pigment phycoerithrin. These two first assumptions has to be tested in further experimental investigations. The additional consideration of the parameter FL1/3 is the bases for the fitting of a two dimensional Gaussian-function. Anyway the dimension of FL4 is only needed for a suitable visualisation of the functional responses.

As written above, the application of cluster analysis to data from flow cytometry has been successful. These first results and further investigations and simulation studies should pay much more attention to such realistic problems in ecological assessment of freshwater systems.

References

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

BARKMANN, J. (2001): *Ökologische Integrität.* Handbuch der Umweltwissenschaften 8/01, Eco-Met-Verlag, Landsberg, 3–21.

FABER, V., HOCHBERG, J.G., KELLY, P.M., THOMAS, T.R., and WHITE, J.M. (1994): Concept Extraction. A Data-mining Technique. *Los Alamos Science, 22, 123–149.*

FRALEY, C. (1996): Algorithms for model-based Gaussian Hierarchical Clustering. *Technical Report, 311.* Department of Statistics, University of Washington, Seattle.

FRALEY, C. and RAFTERY, A.E. (2002): Model-based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association, 97, No. 458, 611–631.*

FRIEDMAN, J.H. and MEULMAN, J.J. (2002): Clustering Objects on Subsets of Attributes. *report http://www-stat.stanford.edu/ jhf/ftp/cosa.pdf.* Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford.

GORDON, A. D. (1999): *Classification.* Chapman & Hall/CRC, London.

GORDON, A. D. and DE CATA, A. (1988): Stability and Influence in Sum of Squares Clustering. *Metron, 46, 347–360.*

GOWER, J.C. (1971): A General Coefficient of Similarity and some of its Properties. *Biometrics, 27, 857–874.*

GUHA, S., RASTOGI, R., and SHIM, K. (1998): CURE: An Efficient Clustering Algorithm for Large Databases. In: *Proceedings of the ACM SIGMOD Conference on Management of Data.* ACM Press, Seattle, 73–84.

HÄRDLE, W. (1990): *Applied Nonparametric Regression.* Cambridge University Press, Cambridge.

HALL, J.A. (1991): Long-term Preservation of Picophytoplankton for Counting by Fluorescence Microscopy. *Br. Phycol. J., 26, 169–174.*

HAMPLEL, F. (1968): *Contributions to the Theory of Robust Estimation.* Ph.D. thesis, University of California, Berkeley.

HOEK, C. VAN DEN, JAHNS, H.M., and MANN, D.G. (1993): *Algen.* Georg Thieme Verlag, Stuttgart.

HOFSTRAAT, J.W., ZEIJL VAN, W.J.M., VREEZE DE, M.E.J., PEETERS, J.C.H., PEPERZAK, L., COLIJN, F., and RADEMAKER, T.W.M. (1994): Phytoplankton monitoring by flow cytometry. *Journal of Plankton Research 16 (9), 1197–1224*

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

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

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

LEPESTEUR, M., MARTIN, J.M., and FLEURY, A. (1993): A Comparative Study of Different Preservation Methods for Phytoplankton Cell Analysis by Flow Cytometry. *Mar. Ecol. Prog. Ser., 93, 55–63.*

MACQUEEN, J.B. (1967): Some Methods for Classification and Analysis of Multivariate Observations. In: L. LECAM and J. NEYMAN (Eds.): *Proc. 5th Berkeley Symp. Math. Statist. Prob., Vol. 1.* Univ. California Press, Berkeley, 281–297.

MARDIA, K.V., KENT, J.T., and BIBBY, J.M. (1979): *Multivariate Analysis.* Academic Press, London.

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

MUCHA, H.-J. (1994). Adaptive Clustering with XploRe, Accompanied by Adaptive Dynamic Graphics. In: F. Faulbaum (Ed.): *SoftStat'93. Advances in Statistical Software.* Gustav Fischer Verlag, Stuttgart, 551–558.

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.

MUCHA, H.-J., BARTEL, H.-G., and DOLATA, J. (2002): Exploring Roman Brick and Tile by Cluster Analysis with Validation of Results. In: W. Gaul and G. Ritter (Eds.): *Classification, Automation, and New Media.* Springer, Heidelberg, 471–478.

MURTAGH, F., RAFTERY, A.E., and STARCK, J.-L. (2001): Bayesian Inference for Color Image Quantization via Model-Based Clustering Trees. *Technical Report, 402.* Department of Statistics, University of Washington, Seattle.

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

SOMMER, U. (1994): *Planktologie.* Springer-Verlag, Berlin.

SPÄTH, H. (1980): *Cluster Analysis Algorithms.* Ellis Horwood, Chichester.

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

STEINBERG, C.E.W. (1998): Integrity of limnic ecosystems. In: J.A. Van de Kraats (Ed.): *Let the Fish speak: The Quality of Aquatic Ecosystems as an Indicator for Sustainable Water Management.* BfG/EURAQUA, 4. technical review, Koblenz, 89–101.

STEINBERG, C.E.W., SCHÄFER, H., and BEISKER, W. (1998): Do Acid-tolerant Cyanobacteria Exist? *Acta hydrochin. Hydrobiol. 26 (1), 13–19.*

TÜMPLING, VON W. and FRIEDRICH, G. (Hrsg.) (1999): Methoden der biologischen Wasseruntersuchung. G. Fischer-Verlag, 35–51.

UTERMÖHL, H. (1958): Zur Vervollkommnung der quantitativen Phytoplanktonmethodik. *Mitteilung der internationalen Vereinigung für theoretische und angewandte Limnologie, 9, 1–38.*

WARD, J.H. (1963): Hierarchical Grouping Methods to Optimise an Objective Function. *JASA, 58, 235–244.*

ZHANG, T., RAMAKRISHNAN, R., and LIVNY, M. (1996): Birch: An efficient clustering method for very large databases. In: *Proceedings of the ACM SIGMOD Conference on Management of Data.* ACM Press, Montreal, 103–114.