Applied Geochemistry 23 (2008) 2198–2213

Contents lists available at ScienceDirect

Applied Geochemistry

journal homepage: www.elsevier.com/locate/apgeochem

Cluster analysis applied to regional geochemical data: Problems and possibilities

Matthias Templ [a,b,*], Peter Filzmoser [a], Clemens Reimann [c]

[a] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Wien, Austria
[b] Department of Register, Classification and Methodology, Statistics Austria, Guglgasse 13, A-1040 Wien, Austria
[c] Geological Survey of Norway, N-7491 Trondheim, Norway

ARTICLE INFO

Article history:
Received 2 November 2007
Accepted 11 March 2008
Available online 26 March 2008

Editorial handling by A. Danielsson

ABSTRACT

Cluster analysis can be used to group samples and to develop ideas about the multivariate geochemistry of the data set at hand. Due to the complex nature of regional geochemical data (neither normal nor log-normal, strongly skewed, often multi-modal data distributions, data closure), cluster analysis results often strongly depend on the preparation of the data (e.g. choice of the transformation) and on the clustering algorithm selected. Different variants of cluster analysis can lead to surprisingly different cluster centroids, cluster sizes and classifications even when using exactly the same input data. Cluster analysis should not be misused as a statistical "proof" of certain relationships in the data. The use of cluster analysis as an exploratory data analysis tool requires a powerful program system to test different data preparation, processing and clustering methods, including the ability to present the results in a number of easy to grasp graphics. Such a tool has been developed as a package for the R statistical software. Two example data sets from geochemistry are used to demonstrate how the results change with different data preparation and clustering methods. A data set from S-Norway with a known number of clusters and cluster membership is used to test the performance of different clustering and data preparation techniques. For a complex data set from the Kola Peninsula, cluster analysis is applied to explore regional data structures.

1. Introduction

The principal aim of cluster analysis is to partition multivariate observations into a number of meaningful multivariate homogeneous groups, i.e. to map the observations into a few centres called centroids. These centroids summarise the group information which allows getting a better overview of the data structure. A good outcome of cluster analysis will result in a number of clusters where the observations within a cluster are as similar as possible while the differences between the clusters are as large as possible. Cluster analysis must thus determine the number of classes as well as the memberships of the observations to the groups. The outcome of most cluster algorithms are memberships of 1 or 0, where 1 means that an observation has been assigned to a specific cluster and 0 otherwise. Fuzzy clustering methods allow for partial assignment, and thus the membership coefficients are in the interval [0,1]. To determine the group membership most clustering methods use a measure of similarity between the observations. Distances between the observations in the data space are generally used to express the similarity.

Cluster analysis was developed in taxonomy. The aim was originally to get away from the high degree of subjectivity when single taxonomists performed a grouping. Since the introduction of cluster analysis techniques there has been controversy about its merits (Davis, 1973 or Rock, 1988). It was soon discovered that diverse techniques can

* Corresponding author. Address: Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Wien, Austria.
E-mail addresses: matthias.templ@statistik.gv.at (M. Templ), clemens.reimann@ngu.no (C. Reimann).

0883-2927/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apgeochem.2008.03.004

yield different groupings, even when using exactly the same data. Furthermore the addition (or deletion) of just one variable in a cluster analysis can lead to completely different results. Workers may thus be tempted to experiment with different techniques and the selection of variables entered until the result of a cluster analysis fits their preconceived ideas. Readers should be very aware of the problems – cluster analysis can be applied as an "exploratory data analysis tool" to better understand the multivariate behaviour of a data set. Alternatively, principal component analysis and factor analysis (Reimann et al., 2002) use the correlation matrix for extracting "components" or "factors" from a given data set, most cluster analysis techniques use distance measures to assign observations to a number of groups. The use of correlation coefficients requires not only a normal, but even a multivariate normal distribution for all the input data (Reimann et al., 2002). This condition is almost never fulfilled when working with geochemical data (Reimann and Filzmoser, 2000). The use of distance coefficients does a priori not make any statistical assumptions about the data (except if the data are of categorical order), theoretically an ideal situation when working with geochemical data. However, geochemical data are "closed" data (compositional data expressed in units like wt.% or mg/kg, summing up to a constant) and multivariate statistical methods may thus deliver biased results (Le Maitre, 1982; Aitchison, 1986,2003). Therefore prior to performing cluster analysis appropriate data transformations have to be considered to "open" the data (Aitchison, 1986).

Distance measures will also be essential for cluster validation, i.e. measuring the quality of a clustering. In theory, it should be ideal to first use cluster analysis on a large geochemical dataset to extract more homogenous data subsets (groups) and to then perform factor analysis or discriminant analysis on these homogenous data subsets to study their multivariate data structure (Frapporti et al.,1996). Especially for data sets with many variables, it has been suggested (Everitt, 1974) to first use principal component analysis to reduce the dimensionality of the data and to then perform cluster analysis on the first few principal components. This approach has been criticised because clusters embedded in a high-dimensional space will not be properly represented by a smaller number of orthogonal components (e.g. Yeung and Ruzzo, 2001).

There are also clustering methods that are not based on distance measures, like model-based clustering (Fraley and Raftery, 1998). These techniques usually find the clusters by optimising a maximum likelihood function. The implicit assumption is that the data points forming the single clusters are multivariate normally distributed, and the algorithm tries to estimate the parameters from the normal distribution as well as the membership of each observation to each cluster.

With geochemical data, cluster analysis can be used in different ways: it can be used to cluster the variables (e.g. to detect geochemical relations between the variables) and it can be used to cluster the observations (e.g. to assign samples to certain types) to come to more homogenous data subsets for further data analysis. Furthermore, there are methods that try to group the data by simultaneously clustering objects and variables (see Ji et al., 1995,2007; Leisch, 1999; Raftery and Dean, 2004; Friedman and Meulman, 2004).

Here a variety of different methods of cluster analysis will be applied to geochemical data from a small dataset from the Oslo area, southern Norway, where 9 different plant materials (e.g., different species or leaves, wood, bark of birch and spruce) were collected at 40 sites along a 120 km long transect (Reimann et al., 2006, 2007a,b) and analysed for 25 elements. This dataset will be used as a test dataset, because for the 9 clusters corresponding to the nine materials the cluster membership is known. The second dataset is a large regional scale geochemical dataset containing 617 observations and 40 variables the KOLA dataset, (Reimann et al., 1998). Here, neither the number of clusters nor the group memberships are known. The objective of using the OSLO data set and the Kola data set for cluster analysis is to give an overview of popular cluster methods for geochemistry, to give a practical guide to apply cluster analysis in geochemistry and to investigate the following specific questions:

- Which transformations are most suitable for such compositional data and what are the effects on cluster results?
- Which distance measures are most suitable for distance based clustering methods?
- What is the influence of the actual method used and is there an ideal method for regional geochemical data?
- Is there an objective way to determine the optimum number of clusters extracted?
- Is there a graphical way to evaluate the stability or validity of clusters?
- Is an objective decision on the number and choice of elements entered into the cluster analysis possible?

Furthermore this paper will provide an overview of the implemented open source code for cluster analysis in R and provide a guideline for optimising the procedure for geochemical data.

2. Materials and methods

2.1. Oslo dataset

During the fall of 2005 a variety of different sample materials was collected at 3 km sampling intervals along a 120 km long transect crossing the city of Oslo. At 40 sample sites 9 different plant materials were collected leading to 360 samples. The plant materials were: terrestrial moss, fern, European mountain ash leaves, birch leaves, bark and wood and spruce needles, twigs and wood. Details about sampling, sample preparation, analysis and quality control can be found in (Reimann et al., 2006, 2007a,b). For 25 (out of 37) variables all analytical values were above detection for all 9 plant materials. This dataset is used here to test the performance of the different cluster methods and data preparation techniques. It is ideally suited for this purpose because different plants have very different element uptake characteristics and in addition distribute the elements differently between the wood, leaves and bark. The "logical result" of cluster analysis are thus 9 clusters, one for each plant material.

2.2. The Kola O-horizon data set

From 1992–1998 the Geological Surveys of Finland (GTK) and Norway (NGU) and Central Kola Expedition (CKE), Russia, carried out a large, international multi-media, multi-element geochemical mapping project, covering 188,000 km^2 north of the Arctic Circle. The entire area between 24° and 35.5°E up to the Barents Sea coast (Fig. 1) was sampled during the summer of 1995. Results of the "Kola Ecogeochemistry" project are documented on a web site (http://www.ngu.no/Kola) and in a geochemical atlas (Reimann et al., 1998). This atlas also provides information about sample collection, sample preparation, analysis and quality control as well as a topographical map, a geological map, a map of the vegetation zones, the location of industry and mines in the area and meteorological maps. Here only a simplified map showing the survey area and locations discussed in the text is provided in Fig. 1.

The Kola Project collected samples of terrestrial moss and 5 genetic layers of podzol profiles throughout the survey area. For testing cluster analysis on a complex regional dataset where the number of clusters is not a-priori known, as one possible example the O-horizon soil samples from the Kola Project are used here. The O-horizon was collected at 617 sites. A summary of the elements analysed in the O-horizon and used for cluster analysis is provided in Table 1.

3. Practical data set problems in the context of cluster analysis

3.1. Mixing major, minor and trace elements

In multi-element analysis of geological materials one usually deals with elements occurring in very different concentrations. In rock geochemistry, the chemical elements are divided into "major", "minor" and "trace" elements. Major elements are measured in % or tens of %, minor elements are measured in about 1% amounts, and trace elements are measured in ppm, or even ppb. This may become a problem in multivariate techniques that are based on distance coefficients because the variable with the greatest magnitude will have the greatest influence on the outcome. Therefore, one should not mix variables quoted in different units in such a multivariate analysis (Rock, 1988). Transferring all elements to just one unit (e.g. mg/kg) is not a solution to this problem, as the major elements occur in much greater amounts than the trace elements. The data matrix will thus need to be "prepared" for cluster analysis using appropriate data transformation and standardisation techniques.

3.2. Data outliers

Regional geochemical datasets practically always contain outliers. The outliers should not simply be ignored

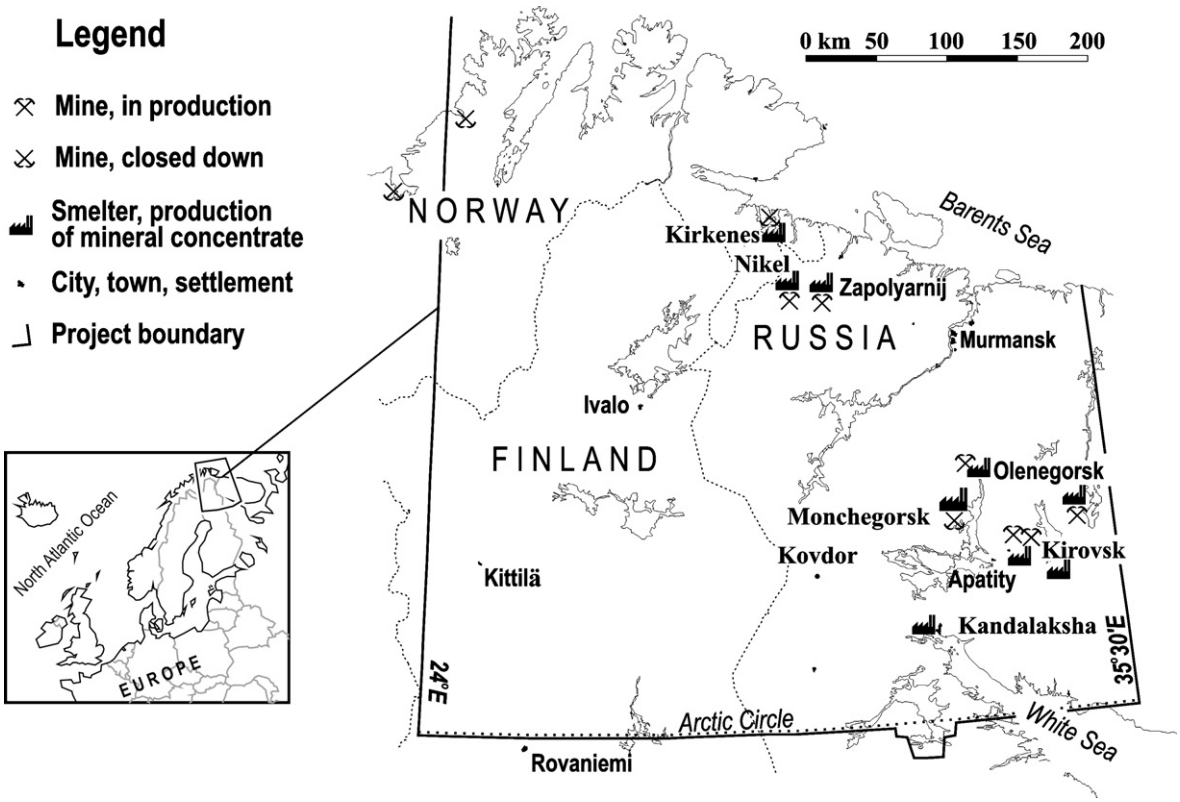

Fig. 1. General location map of the study area for the Kola Project (Reimann et al., 1998).

Table 1
Elements and summary statistics (minimum (MIN), median (MED), maximum (MAX) and spread (expressed as median absolute deviation – MAD) for the Kola O-horizon data set used here (from Reimann et al., 1998)

Element	DL	% <DL	MIN	MED	MAX	MAD
Ag	0.02	0	0.025	0.2	4.79	0.16
Al	0.2	0	372	1890	20600	1201
As	0.05	0	0.364	1.16	43.5	0.46
B	0.8	0.2	<0.8	2.15	13	0.7
Ba	0.05	0	13.9	76	290	30.3
Be	0.02	25.1	<0.02	0.04	1.87	0.04
Bi	0.02	0	0.029	0.159	1.12	0.08
C	1000	0	153000	450000	508000	3710
Ca	5	0	460	2960	25400	786
Cd	0.02	0	0.07	0.3	1.39	0.11
Co	0.03	0	0.21	1.57	96	1.11
Cr	0.4	0	0.39	2.91	109	1.75
Cu	0.01	0	2.7	9.7	4080	5.14
Fe	10	0	430	1970	44800	1245
H	1000	0	22000	61000	71000	444
Hg	0.04	0	0.094	0.227	0.974	0.05
K	200	0	300	1000	5700	297
La	0.7	4.5	<0.7	2.3	139	1.78
Mg	10	0	240	750	23800	297
Mn	1	0	11.1	126	5470	108
Mo	0.01	0	0.086	0.258	5.45	0.1
N	1000	0	5000	13000	20000	300
Na	10	3.4	<10	60	2350	29.7
Ni	0.3	0	1.5	9.2	2880	7.74
P	15	0	192	930	9280	208
Pb	0.04	0	4.1	19	1110	7.41
Rb	0.5	0	0.68	5.8	33	2.76
S	15	0	400	1530	3830	297
Sb	0.01	0	0.016	0.183	0.962	0.08
Sc	0.1	0.5	<0.1	0.5	4.1	0.3
Si	20	0	290	530	940	74.1
Sr	0.2	0	6.1	29	1430	13.6
Th	0.04	0	0.06	0.35	15.4	0.25
Tl	0.01	0	0.02	0.09	0.56	0.05
U	0.004	0	0.008	0.099	14.3	0.07
V	0.02	0	1.1	4.9	49	2.39
Y	0.1	0	0.2	0.9	69	0.59
Zn	0.4	0	12	46	198	15.1
Other parameters						
pH	0.1	0	3.2	3.85	5.6	0.22
LOI	0.1	0	33.5	89.8	98.8	6.52

In addition the detection limit (DL) and the number of samples below detection (expressed in %) are given. The element Se is excluded from analysis due to the high amount of values below detection limit.

but they have to be accommodated because they contain important information about data quality and unexpected behaviour in the region of interest. In fact, finding data outliers that may be indicative of mineralisation (in exploration geochemistry) or of contamination (in environmental geochemistry) is one of the major aims of geochemical surveys. Outliers can have a severe influence on cluster analysis, because they can disturb homogeneous clusters or fall into single clusters, depending on the clustering method used. This has to be taken into account for cluster analysis and a simple possibility would be to delete the outliers in advance or to increase the desired number of clusters. However, finding data outliers is not a trivial task, especially in high dimensions. One way of identifying such outliers is to compute robust Mahalanobis distances, i.e. Mahalanobis distances on the basis of robust estimates of location and scatter (Filzmoser et al., 2005). In the case of compositional data, outlier detection requires an appropriate transformation (Filzmoser and Hron, 2008).

3.3. Censored data

Observations below the detection limit can form a significant proportion of a variable in geochemical data sets. For statistical analysis these results are often set to a value of ½ the detection limit. However, a sizeable proportion of all data with an identical value can seriously influence any cluster analysis procedure. For the study datasets several variables had more than 25% of the data below detection limit. It is very questionable as to whether or not such elements should be included at all in a cluster analysis. Unfortunately it is often the elements of greatest interest that contain the highest number of censored data (e.g., Se) – the temptation to include these in a cluster analysis is thus high. Here all elements with more than 5% of all values below detection limit have been omitted from cluster analysis. (Helsel, 1990, 2004), (Cohen, 1991) and (Sanford et al., 1993) have studied possibilities for estimating censored data. It is, however, doubtful whether a cluster analysis should be built on estimated values.

3.4. Data distribution and data scale

Cluster analysis based on distance coefficient in general does not require that the data be normally distributed. However, it is advisable that heavily skewed data are first transformed to a more symmetric distribution. If a good cluster structure exists for a variable, a distribution can be expected which has two or more modes. A transformation to more symmetry will preserve the modes but remove large skewness and thus improve the cluster results. In many cases the log-transformation can be successfully used to approach symmetry. A more universal choice is the Box–Cox transformation (Box and Cox, 1964). Here, for each variable the optimal parameter for the Box–Cox transformation has to be determined which is a time consuming procedure with large datasets.

An additional standardisation is needed if the variables show a striking difference in the amount of variability (see discussion above, major, minor and trace elements). Different methods, all having advantages and disadvantages, exist to accommodate this requirement. The most universal method is the *z*-transformation, in which the raw data are subtracted with the mean and then divided by the standard deviation of the data. When working with geochemical data a robustified version, using median and the median absolute deviation (MAD) instead of the mean and the standard deviation respectively may be preferred.

3.5. Closed number systems

A data set is considered as "closed", when the individual variables are not independent of each other but are related, e.g. by being expressed as a percentage (or ppm, mg/kg). They sum up to a constant, e.g. 100% or 1.

The problem of undertaking statistical analyses with "closed number systems" has been much discussed in the literature (e.g., Butler, 1976; Aitchison, 1986,1992,2000,

2003; Barcelo-Vidal et al., 1996; Aitchison et al., 2000; Buccianti et al., 2006). Neglecting closure can have serious consequences for data analysis. Different data transformations like the additive logratio (alr) and the centered logratio (clr) transformation (Aitchison, 1986) or the isometric logratio (ilr) transformation (Egozcue et al., 2003) are suggested in the literature to open the data and destroy the effects of closure. For the additive logratio transformation one variable in the dataset must be chosen to open the data, this selected variable is subsequently "lost" for further data analysis. The centered logratio transformation does not depend on the results of one single other variable but uses the average of all variables. While the centered logratio transformation results in collinear data the isometric logratio transformation avoids collinearity (Egozcue et al., 2003). The computed distances between the observations of the opened data will in general be different to those coming from a simple log-transformation or from a Box–Cox transformation.

4. Distance measures

A key issue in most cluster analysis techniques is how best to measure distance between the observations (or variables). Note that "distance" in cluster analysis has nothing to do with geographical distance between two observations but is rather a measure of similarity between observations in the multivariate space defined by the entered variables. Many different distance measures exist (Bandemer and Näther, 1992).

For clustering the observations the Euclidean distance or the Manhattan distance is the most frequent choice. The latter measures the distance along the variable axes, rather than directly (Euclidean), and the cluster results are sometimes more stable (Kaufman and Rousseeuw, 1990). Usually both distance measures lead to comparable results. Other distance measures like the Gower distance (Gower, 1966), the Canberra distance , correlation based distance measures or a distance measure based on the random forest (RF™) proximity measure (Breiman, 2001) can give completely different cluster results (see, e.g., Table 2).

5. Clustering observations

One of the main problems with cluster analysis is that a multitude of different clustering methods exists. The observations need to be grouped into classes (clusters). If each observation is allocated into only one (of several possible) cluster(s) this is called "partitioning". Partitioning will result in a pre-defined (user defined) number of clusters. It is also possible to construct a hierarchy of partitions, i.e. group the n observations into several clusters. This is called hierarchical clustering.

A principally different observation allocation procedure to hard clustering, where an observation is allocated to just one cluster, is to allocate the observations to several clusters (fuzzy clustering). Fuzzy clustering allows that one observation belongs to a certain degree to several groups. In terms of applied geochemistry this procedure will often deliver the more interesting results because it reveals if an observation is influenced by several factors. The cluster solution will then show to what degree the observations are influenced by the different factors. Here the factors or processes are represented by observations that are clustered together in the data space.

5.1. Hierarchical methods

Input to most hierarchical clustering algorithms is a distance matrix (between the observations). The widely used agglomerative procedure starts with single object clusters (each observation forms its own cluster) and aggregate observations to enlarge the clusters stepwise. The computationally more intensive reverse procedure starts with one cluster containing all observations and splits the groups step by step, this procedure is called divisive clustering.

At the beginning of an agglomerative algorithm each observation forms its own class, leading to n single object clusters. The number of clusters is reduced by combining (linking) the most similar classes at each step of the algorithm. The similarity of the combined pair, a new class, can be measured to all other classes, and the next two most similar classes linked, and so on. At the end of the process there is only one single cluster left, containing all observations. A number of different methods are available for linking two clusters. Best known are complete linkage (maximum distance), single linkage (minimum distance), and average linkage (average distance). The method of Ward (Ward, 1963) merges clusters with a minimum information loss criteria based on sums of squares.

The cluster results are often displayed in a graphic called the dendrogram (see Fig. 6 for clustering variables). Horizontal lines indicate the linkage of two objects or clusters, and thus the vertical axis shows the associated height

Table 2
Results of clustering the OLSO data with 9 clusters and different clustering algorithms applied on data which are transformed with different transformations and dealing with different distance measures are summarized in the form of the number of misclassifications

	Ward					*k*-means (best results after 10 starts)					Mclust
	Eucl	Manh	Gower	Canberra	RF	Eucl	Manh	Gower	Canberra	RF	
no	30	11	7	103	26	14	47	13	96	34	35
log	1	0	3	55	22	7	16	9	54	23	8
clr	23	5	2	139	12	13	25	8	8	9	15
alr	17	22	4	87	32	22	49	35	15	30	11
ilr	14	2	2	109	1	4	7	26	10	4	0

Since the conventional *k*-means algorithm is based on random initialisation, the best solution after 10 random initialisations was taken.

or similarity as a measure of distance. Cutting the dendrogram at the height corresponding to this visible number of clusters allows assigning the objects to the clusters. Visual inspection of a dendrogram is often helpful in obtaining an initial idea of the number of clusters which is needed by a partitioning method.

5.2. Partitioning methods

In contrast to hierarchical clustering methods partitioning methods require that the number of resulting clusters be pre-determined. As noted above, when nothing is known about the observations it can be useful to first carry out a hierarchical clustering. The other possibility is to partition the data into different numbers of clusters and then evaluate the results by some method (see below). For regionalized data a more subjective but still reasonable evaluating approach is to visually inspect the location of the resulting clusters in a map. This exploratory approach can often reveal interesting data structures.

A very popular partitioning algorithm is the *k*-means algorithm. It attempts to minimise the average squared distance between the observations and their cluster centres or centroids. Starting from *k* initial cluster centroids (e.g. random initialisation by *k* observations), the algorithm assigns the observations to their closest centroids (using e.g. Euclidean distances), recomputes the cluster centres, and iteratively reallocates the data points to the closest centroid. Several algorithms exist for this purpose, those of Hartigan (1975) and MacQueen (1967) are the most popular. There are also some modifications of the *k*-means algorithm. Manhattan distances are used for *k*-medians and the centroids are the medians of each cluster. Hard competitive learning works by randomly drawing an observation from the data and moving the closest centre towards that point (e.g., Ripley, 1996). Martinetz et al. (1993) have introduced "neuralgas", this method is similar to hard competitive learning, but in addition to the closest centroid also the second closest centroid is moved at each iteration. A new high extensible toolbox for centroid clustering was recently implemented in R (Leisch, 2006). Here the user can easily try out almost any arbitrary distance measure and centroid computations for data partitioning.

Kaufman and Rousseeuw (1990) proposed several clustering methods which are implemented in a number of software packages. The partitioning method PAM (Partitioning around medoids) minimises the average distances to the cluster medians. It is thus similar to the *k*-medians method but allows the use of different distance measures. A similar method called CLARA (Clustering Large Applications) is based on random sampling. It saves computation time and is particularly appropriate for larger datasets.

The result of all these algorithms depends on the initial *k* cluster centres, which are often the *k* samples randomly selected from *n* observations. If bad initial cluster centres are selected, the iterative partitioning algorithms can lead to a local optimum that can be far away from the global optimum. This can be avoided by applying the algorithms with different random initialisations, and then selecting the best or most stable result.

Another way to approximate the global optimum is bootstrap aggregation, called bagging (Breiman, 1996). This bootstrap method generates new datasets from the available dataset of the same size by a random selection of observations with replacement from the dataset. The central idea of the bagged clustering algorithm bclust is to repeatedly apply a clustering algorithm (e.g. *k*-means) on bootstrap datasets, combine the resulting centroids to a new dataset, run a hierarchical clustering algorithm on this new dataset and cut the resulting dendrogram to get a partition into *k* clusters. The observations are then assigned to the closest centre.

5.3. Model-based methods

A method that is not based on distances between the observations but on certain models describing the shape of the clusters is called model-based clustering (Fraley and Raftery, 2002). Each cluster is described by the density of a multivariate normal distribution with a certain mean and covariance. The choice of the covariance matrix will determine the cluster shape. The Mclust algorithm selects the cluster models (e.g. elliptical cluster shape) and the number of clusters and determines the cluster memberships for all observations. The estimation of the cluster models is achieved using the Expectation–Maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm is executed on several numbers of clusters and with several sets of constraints on the covariance matrices of the clusters. Finally, the combination of model and number of groups that leads to the highest BIC (Bayesian Information Criterion) value can be chosen as the optimal model (Fraley and Raftery, 1998). The BIC measure is based on the likelihood function and penalizes the complexity of the model. It is therefore better suited to determine the "optimal" model and the "optimal number of clusters". The BIC value can also be computed for each cluster separately. As a further result, Mclust computes uncertainties for the assignment of each object to the clusters. This indicates how well the objects fall into the clusters. Objects with high uncertainty could be treated as outliers and not assigned to any of the clusters.

5.4. Fuzzy methods

In fuzzy clustering, the observations are not clearly allocated to one of the clusters, but they are "distributed" in certain degree among all clusters. Thus, for each observation a membership coefficient to all clusters is determined, providing information on how strong the observation is associated with each cluster. The membership coefficients are usually transformed to the interval [0, 1], and they can be visualised for example by using a grey scale. A popular fuzzy clustering algorithm is the fuzzy c-means (FCM) algorithm, developed by Dunn (1973) and improved by Bezdek (1981), which calculates the prototypes (most typical group characteristics) of the clusters and membership coefficients for each observation to the clusters. Another fuzzy clustering algorithm is the Gustafson–Kessel (GK) algorithm (Gustafson and Kessel, 1979). While FCM identifies clusters that tend to be rather spherical, GK is able to detect elliptically shaped clusters. The Gath–Geva (GG)

algorithm (Gath and Geva, 1989), also called the Gaussian mixture decomposition algorithm, is even more flexible. It is an extension of the GK algorithm which can also deal with different cluster sizes and densities. The GK and the GG algorithms are freely available at http://www.fuzzy-clustering.de. Just as for partitioning methods, the number of clusters resulting from fuzzy clustering needs to be chosen by the user.

6. Clustering variables

Instead of clustering the observations it is also possible to cluster the variables in order to find groups of variables that show similar behaviour. All of the methods discussed above can be used for clustering the variables. One of the best methods to display the results of clustering variables is the dendrogram (for an example see Fig. 6), calling for hierarchical clustering.

7. Evaluation of cluster validity

Because there is no universal definition of clustering, there is no universal measure with which to compare clustering results. However, evaluating cluster quality is essential since any clustering algorithm will produce several different results for every dataset. Validity measures should support the decision as to the number of natural clusters, and they should also be helpful for evaluating the quality of the individual clusters. Therefore, validity measures should provide a value for each single cluster (local validity measures), and they should also return a value for judging the quality of the overall clustering result (global validity measures).

When working with geochemical data, a rather simple method to evaluate the quality of clustering is to check the distribution of the resulting clusters on a map. The distribution of the clusters can then be evaluated against known properties of the survey area. It is also likely that clusters resulting in geographically homogeneous subgroups are more likely to have a meaning than clusters resulting in "geographical noise".

There are many different statistical cluster validity measures. Two different concepts of validity criteria – external and internal criteria – need to be considered.

External criteria compare the partitioning found with clustering with a partitioning that is known a priori. The most popular external cluster validity indices are Rand, Jaccard, Folkes and Mallows, and the Hubert indices (see e.g., Gordon, 1999, or Haldiki et al., 2002; Hubert and Arabie, 1985).

Internal criteria evaluate the clustering result of an algorithm by using only quantities and features inherent in the dataset. Most of the internal validity criteria are based on within cluster sum of squares and between cluster sum of squares. Well known indices are the Calinski–Harabasz index (Calinski and Harabasz, 1974), Hartigan's indices (Hartigan, 1975), or the Average Silhouette Width of Kaufman and Rousseeuw (1990).

From a practical point of view, an optimal value of the validity measure does not imply that the resulting clusters are meaningful. Some of these criteria evaluate only the allocation of the data to the clusters. Other criteria evaluate the form of the clusters or how well the clusters are separated. The resulting clusters only correspond to the best partition according to the validity measure selected. The measures deliver good results when a very clear cluster structure exists in the data. Unfortunately, when working with geochemical data such good clusters are rare. Thus cluster quality measures fail time and again when working with such data and the best approach to evaluating cluster quality is often to just look at the results on a map. This somewhat subjective approach can be formalised by using validity measures for each single cluster.

8. Selection of variables for cluster analysis

When using a multivariate technique, variable selection is often employed in order to reduce the dimensionality of a dataset or to learn something about the internal structure between the variables and/or observations. Often it may appear desirable to perform cluster analysis with all available observations and variables. However, the addition of only one or two irrelevant variables can have a drastic influence in identifying the clusters. The inclusion of only one irrelevant variable may be enough to hide the real clustering in the data (Gordon, 1999). The selection of the variables to enter into a cluster analysis is thus of considerable importance when working with applied earth science datasets containing a multitude of variables.

Another reason for variable selection may be a desire to focus the analysis on a specific geochemical process. Such a process is usually expressed by a combination of variables, and using these variables for clustering permits identifying those observations or areas where the process is either present or non-existent. The variables could simply be chosen based on expert knowledge. It is also possible to apply variable clustering (see above) and select variables which are in close relation (one branch of the cluster tree) to highlight a certain process. Variable clustering can of course also be used to select single key variables from each important cluster to simply reduce dimensionality for clustering observations.

9. Testing clustering with geochemical data

As described above, cluster analysis requires the analyst to make a choice on various different options like data preparation and transformation, clustering method, distance measure, cluster validity measure, etc. In the following an attempt is made to investigate the effect of some of these choices on a geochemical dataset, the OSLO data. Different selections will lead to different cluster results, and in this case the results can be easily evaluated because the membership of the observations to the nine plant materials is known.

9.1. Transformation, distance measure, clustering algorithm

The 25 variables of the OSLO dataset generally have a right-skewed distribution. This is a very common property

of geochemical data. For testing the effect of the transformation on the cluster results, the original data as well as the log-transformed data were taken. Moreover, the additive, centred and isometric logratio transformations were considered because of the compositional nature of the data. Note that for the additive logratio transformation a choice on the ratioing variable has to be made (here, Sc was chosen), and that a different choice will give a different result. The resulting data were then standardised to mean zero and variance one to ensure equal influence of the variables on the clustering method.

The effect of the distance measure used internally in the clustering algorithm is also tested. Euclidean (Eucl), Manhattan (Manh), Gower, Canberra, and Random Forest (RF) distance are used for this purpose. These distance measures are taken for the partitioning methods of Ward and for *k*-means clustering. Also partitioning with Mclust is considered which, however, does not depend on a distance measure. The different clustering methods are run with a predetermined number of nine clusters. These are expected to correspond to the nine plant materials. An evaluation of the results is then made by comparing the number of observations that have been assigned to a wrong plant material (with respect to the most dominant plant material of the observations in a cluster).

9.2. Discussion of cluster performance on the Oslo data set

Table 2 presents the results of this experiment in terms of the number of misclassified observations. (Table 2) does not favour one transformation, distance measure or clustering algorithm, but some conclusions can be drawn: No data transformation (no) or additive logratio transformation (alr) leads to poor results in almost all cases. The centred logratio transformation (clr) leads to somewhat better results. Log-transformation (log) and isometric logratio transformation (ilr) appear to be preferable options. An interesting observation is that the Ward method in combination with the Gower distance delivers relatively reliable and stable results independent of data transformation. The numbers of misclassified observations for the hierarchical procedures single, complete and average linkage were also computed (results not shown here), but the Ward procedure performs best. *k*-means clustering leads to generally high misclassification rates. *k*-means clustering works well with spherical symmetric clusters but fails for other cluster structures. It is not very likely that geochemical data form spherical clusters. Mclust looks for elliptically symmetric clusters. Here data transformation plays an important role and the ilr-transformation clearly delivers the best result. The reason is that the ilr transformation results in the best geometrical representation of compositional data. In addition to that, it can be seen that Canberra method did not perform well.

9.3. Validity measure

Fig. 2 shows a plot of validity measures resulting from clustering the scaled OSLO data, using different algorithms and various transformations. The performance evaluation of the results was made with the most simple validity measure, the average within cluster sum of squares divided by the average between cluster sum of squares. Fig. 2 shows plots of this measure against the number of clusters. Small values are preferable because this indicates homogeneous and well separated clusters. Typically, the optimal number of clusters is indicated by a knee in the plot. One should select the cluster number right before the knee occurs. For the OSLO dataset 9 clusters are expected and thus the knee should be at 9. Fig. 2 shows the validity measures for several different situations (cluster algorithms and distance measures), however, the expected "significant" change (knee effect) is usually not visible.

It is obvious that the graphs in Fig. 2 do not provide any real help in determining the optimal number of clusters. Some of the curves showing the best knee effect at the expected number of 9 clusters are methods with a very high number of misclassifications (compare Table 2). Thus neither the best cluster method nor the correct number of clusters can be selected based on this graphic. None of the other (global) internal validity measures mentioned above delivered better results. As a result the unfortunate situation is that an optimal number of clusters cannot be chosen based on these methods.

10. A tool for the exploratory use of cluster analysis

It has been shown above that the cluster results can be changed dramatically with the choice of the data transformation, the clustering method, and the distance measure. Moreover, depending on the selected validity measure, different solutions result for the optimal number of clusters. Despite the variety of cluster results, each partition could still be informative and valuable. The results can give an interesting insight into the multivariate data structure, even if the validity measure does not suggest the chosen number of clusters is optimal. Thus, it is desirable to perform cluster analysis in an exploratory context, by being able to easily change the cluster parameters and visually inspecting the results.

For this purpose, a statistical tool has been developed in R (R Development Core Team (2006), freely available at http://cran.r-project.org) as the contributed package clustTool. (Fig. 3) shows the main menu of the tool clustTool where the desired settings can be made easily. Besides the selection of data, a background map (optional), variables and coordinates, different parameters, like the distance measure, the transformation and standardisation, the clustering method, the number of clusters, and the validity measure can be selected. Depending on the selection, the clusters are presented in maps (for an example see Fig. 4) and plots of the cluster centres are provided (for an example see Fig. 5). Additionally, a summary is provided and information about the clustering is saved in an object in the R workspace.

11. Results for the example data sets and graphical presentation

The experience of the tests made in Section 9 for the OLSO data can now be used for analysing the complex Kola O-horizon dataset. The clustering tool presented in Section

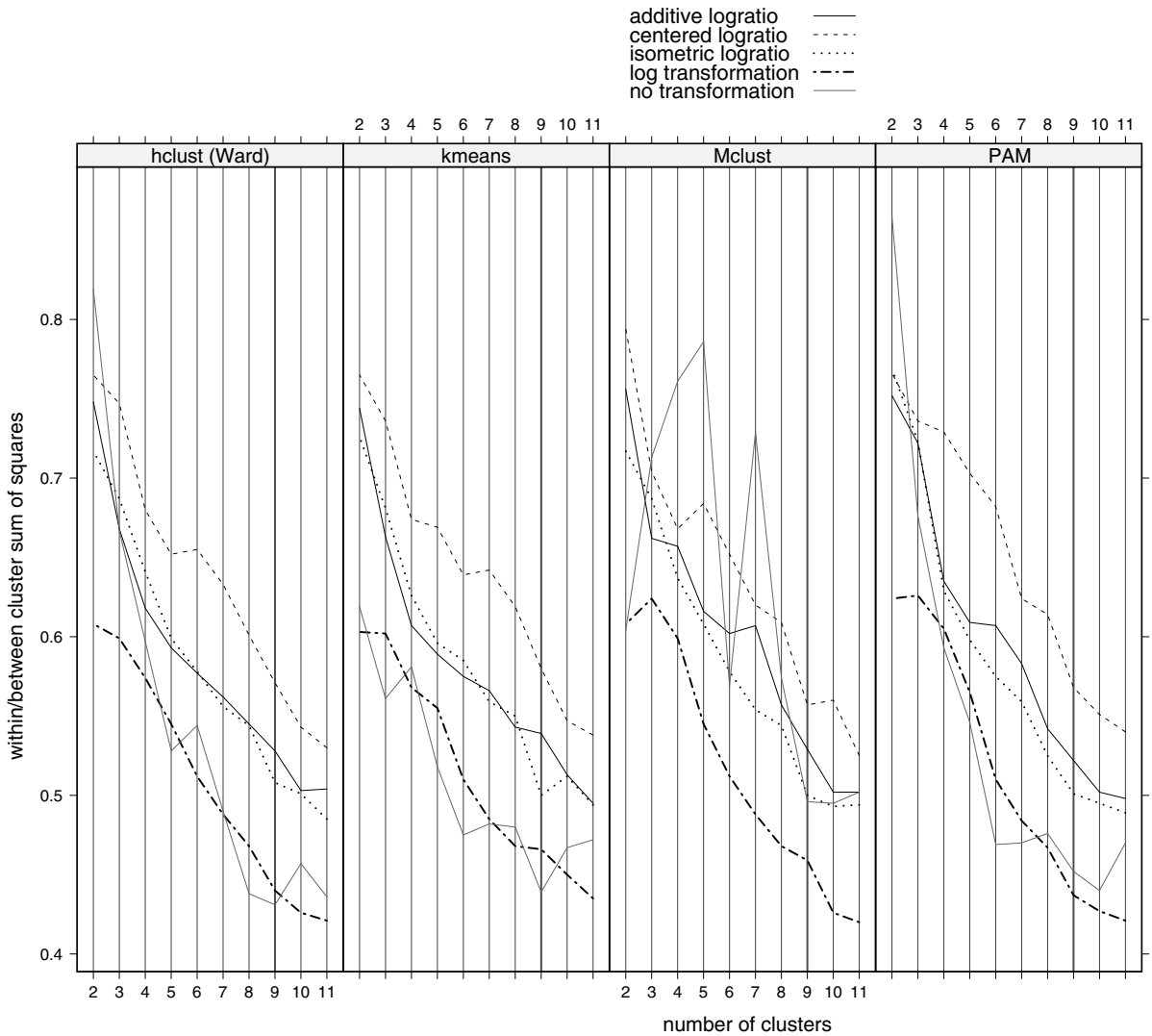

Fig. 2. Resulting validity measures of clustering 25 variables of the OSLO data (different transformations, standardised) with different methods based on the Euclidean distance measure.

10 allows application of different choices of clustering methods and a fast comparison of the results in an exploratory way.

11.1. Kola O-horizon data set clustering with Mclust

Since the algorithm Mclust worked very well for the standardised isometric log-ratio transformed OSLO data set, the same procedure was applied to the 40 elements of the Kola O-horizon data set. The optimal number of clusters is obviously difficult to choose. Because the validity measures did not provide a constructive indication about the number of clusters the new package was used to present results in maps and the desired number of clusters was increased step-wise from 3 to 9. The first result delivering informative clusters was received for 6 clusters. With higher cluster numbers, no significant new information was obtained and the clustering results could even get worse. Using other variables or applying other clustering algorithms could, however, require a different optimal number of clusters.

The validity measure BIC was used for evaluating each individual cluster. Higher values of the BIC indicate more informative clusters. Therefore, the BIC value is used for assigning grey scales to the observations in the maps. (Fig. 4) shows the resulting clusters in 6 maps from cluster 1 to 6. Cluster 5 shows low BIC values, therefore the observations are visualised as light grey points. Cluster 3 shows the input of sea spray along the Norwegian coast. Cluster 4 identifies the core areas of contamination surrounding Monchegorsk and Nikel/Zapoljarnij. Cluster 2 shows the alkaline intrusions near Apatity. Clusters 1, 5 and 6 are less informative. However, they still display clear regional patterns. Cluster 6, for example, shows the outer rim of contamination surrounding Nikel/Zapoljarnij and Monchegorsk.

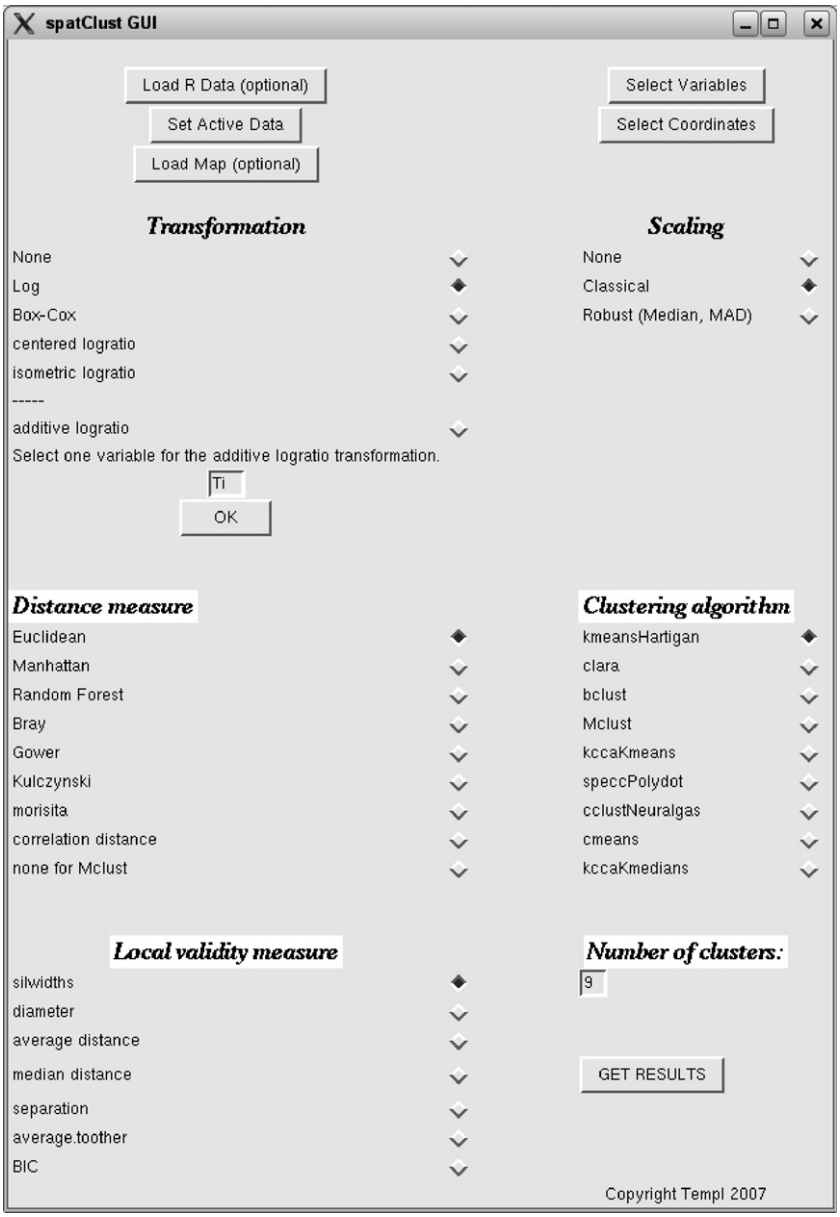

Fig. 3. Main menu of the tool clustTool for an exploratory use of cluster analysis.

In general, not only the location of the single clusters on the maps is interesting but also the geochemical composition of the clusters. For this purpose, a plot of the cluster centres is presented in Fig. 5, which aids the interpretation of the processes behind the clusters. The cluster centre is the element-wise mean of all observations of a cluster. Therefore, for each cluster all elements used for clustering are presented. In (Fig. 5) the resulting means for all 6 clusters presented in Fig. 4 are horizontally arranged. Since the variables used for clustering were standardised, each of them make the same contribution to the cluster analysis. If single elements show very high or low means for a cluster, they are highly influential for that cluster. For example, (Fig. 5) shows high means of the elements Co, Ni, As, Cu and Mo for cluster 4, identifying the Russian Ni industry.

11.2. Influence of data transformation and clustering method

The new R-package was then used to perform cluster analysis on the Kola O-horizon data using log-transformation, additive, centered and isometric logratio transformation and different clustering methods and distance measures in an exploratory way. For the additive logratio transformation the choice of the ratioing variable turned out to have a major influence on the results. The centred logratio transformation led to somewhat unstable results

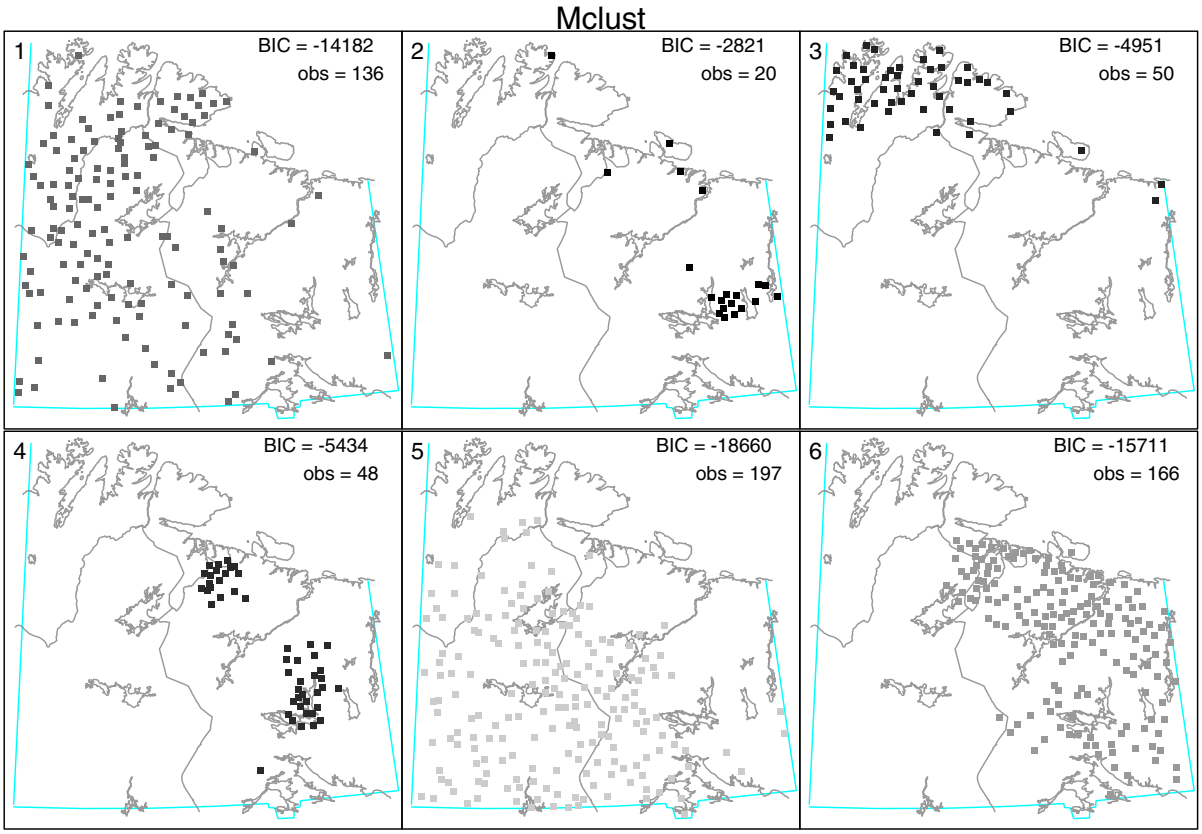

Fig. 4. Mclust for 40 elements of the Kola Project O-horizon data (isometric logratio transformed, standardised). 6 clusters were chosen and each cluster is evaluated with the BIC measure, resulting in different grey scales for the observations in the maps.

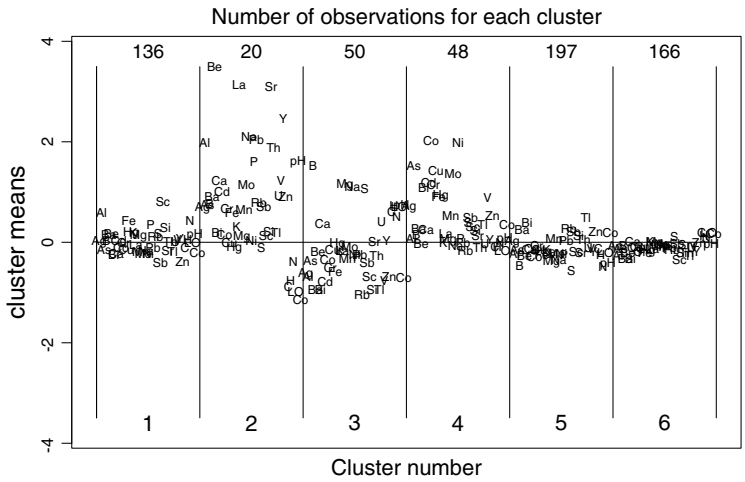

Fig. 5. Plot of the cluster centres for the cluster analysis presented in Fig. 4 (40 variables of the O-horizon data, isometric logratio transformed, standardised). High or low values of the elements suggest high influence of these elements on the observations in the corresponding cluster.

when the number of clusters was varied. Log-transformation and isometric logratio transformation leads to comparable results.

Choice of cluster method and distance measure had a major influence on the results. Mclust delivered the best interpretable clusters, showing clear geographical patterns. For geochemical data geographically defined groups are a logical result. The appearance of the clusters in maps can thus be used as a subjective quality measure. It is also possible to calculate a compactness measure of the groups

in the map which could than be compared to other validity measures.

Results with *k*-means clustering were less clear than those obtained with Mclust. Some main clusters stayed the same but did not show as clear regional patterns. In addition some other clusters were identified that could not be interpreted and did not show any regional structure.

Fuzzy clustering with all variables failed with the complex Kola dataset because the algorithm converges to equal or almost equal group memberships for all observations.

In general partitioning methods worked better than hierarchical methods. The hierarchical methods resulted in biased results with several very small clusters and some very large clusters. An exception was the method of Ward where clear and interpretable groups emerged, however, the Mclust results were still better interpretable.

With the exception of Mclust all considered methods are based on a distance measure. The Euclidean and Manhattan distance gave comparable results, the Canberra and random forest distance gave both poor results, with random forest occasionally delivering better results. The Gower distance resulted in very small clusters, which is not desirable because these will often be driven by outliers and not by larger scale geochemical processes. Several further distance measures were tested, none provided better results than Euclidean or Manhattan distance.

11.3. Variable clustering and variable selection

As an example for variable clustering the 40 elements of the Kola O-horizon data set were clustered. For variable clustering the isometric logratio transformation is not a viable choice because the direct relationship to the original variables is lost. The new artificial variables are no longer directly related to the chemical elements. Since centred logratio transformation gave relatively unstable results with hierarchical clustering the variables were simply log-transformed. Following a log-transformation and standardisation of the data, hierarchical clustering with the Ward method based on the Euclidean distances between the variables was applied resulting in the dendrogram in Fig. 6. Variables that are in close relation form a branch in the dendrogram. Thus, when focusing on certain processes, variables in a branch can be selected for the further clustering of observations. For example, the elements Cu and Ni, followed by Co are the 3 main elements emitted by the Russian smelters. Traces of Bi, Pb, Cd, As and Mo are also emitted and a sizable Bi, As, Cd and Mo-anomaly surrounds Monchegorsk (Reimann et al., 1998). The regional distribution of Na, B, Mg, Ca, Sr and the pH is dominated by the input of sea spray along the coast of the Barents Sea, while C, H, LOI, S and N are all related to the amount of organic material in the sample. The Th, U, La, Y, Al and Be branch is a geogenic dust indicator, the linked Cr, V, Fe and Sc is another dust branch but these elements may rather indicate the input of anthropogenic dust. Some of these dendrogram branches show a direct relationship to the element plot shown in Fig. 5. Thus clusters of variables are related to clusters of observations.

Some variables may not be ideally suited for cluster analysis because they do not show groups. To include such variables in cluster analysis can obscure the result. Furthermore fuzzy clustering cannot be used with too many variables. It is thus often desirable to reduce the number of variables with which cluster analysis is entered. For reducing the number of elements entered in cluster analysis one could select just one or two elements from each of the major branches of a dendrogram, or select the elements on just one branch in more detail. It is of course also possible to use geochemical arguments for the selection of variables.

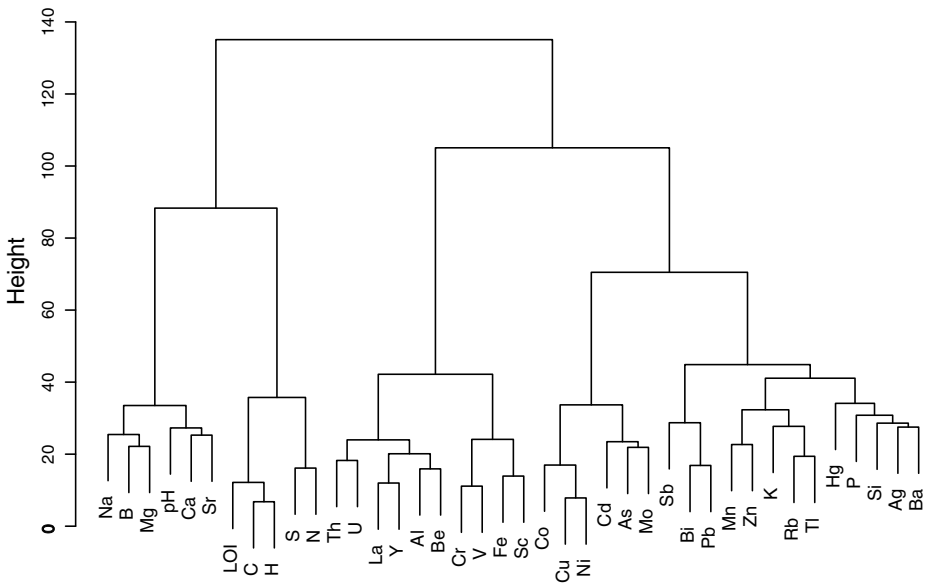

Fig. 6. Clustering the 40 variables of the O-horizon data (log-transformed, standardised) using the Ward hierarchical clustering algorithm based on the Euclidean distance between the variables. Results are presented in the form of a dendrogram.

11.4. Fuzzy clustering with selected variables

In a final example the results of fuzzy clustering on selected elements of the Kola O-horizon data are shown. The variables B, Ca, Co, Cu, Mg, Na, Ni, Sr and pH indicative of two of the main processes in the survey area (sea spray and industry – see Fig. 6) were isometric logratio-transformed and standardised. Based on the Euclidean distances, the FCM algorithm with 4 clusters is applied. With a higher number of clusters no interesting new geochemical processes could be detected. The resulting membership coefficients are shown in grey scales in Fig. 7: higher membership coefficient of an observation to a particular cluster is visualised by a darker point in the corresponding map. The plot in Fig. 8 with the cluster centres allows a better understanding of the resulting clusters: Cluster 1 is a "sea spray" cluster, and Cluster 4 is a contamination cluster. Cluster 2 appears to indicate an outer rim of contamination, while all background observations accumulate in Cluster 3.

Hence with proper variable selection FCM can provide added value in analysing geochemical datasets.

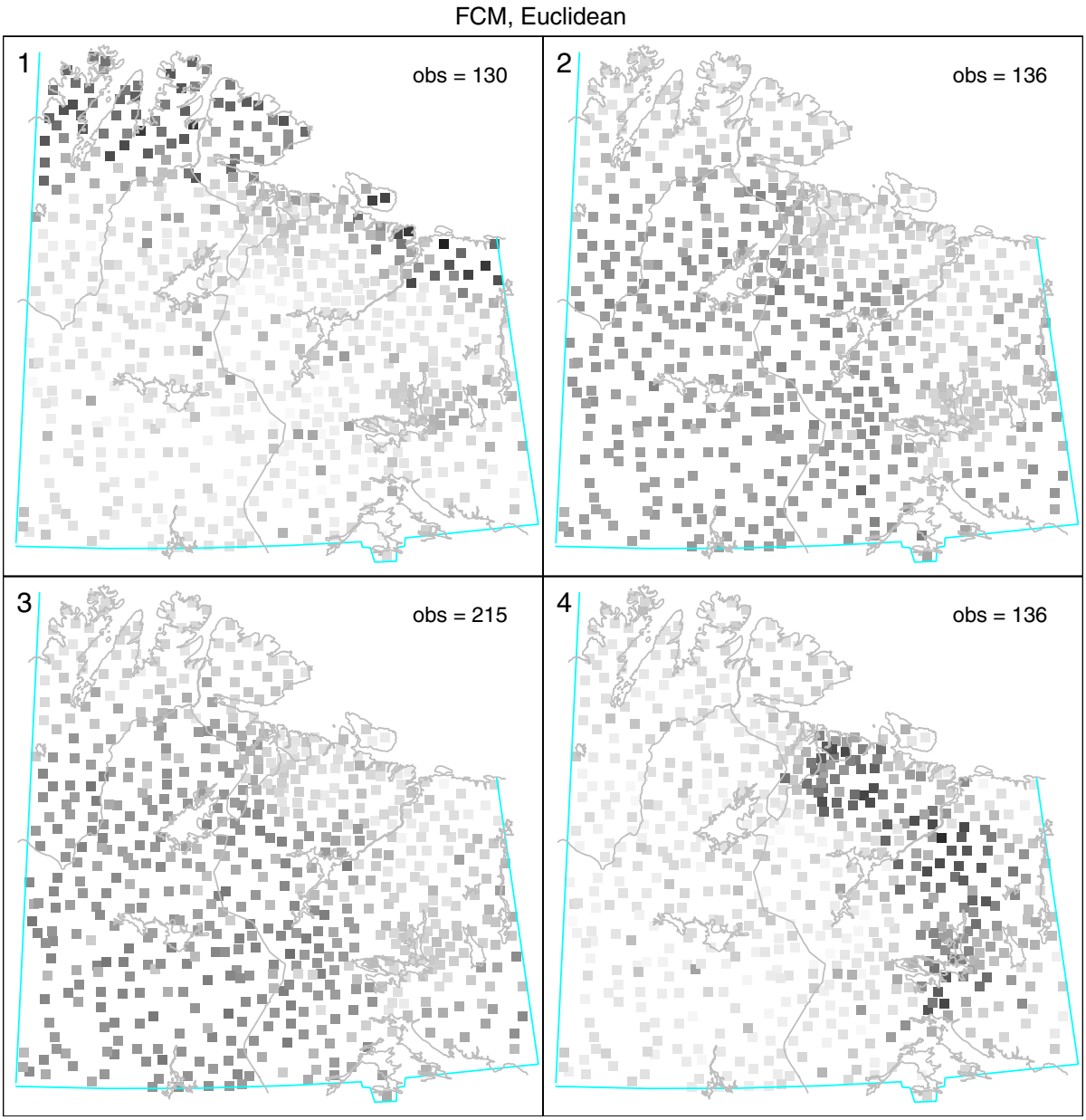

Fig. 7. Results of fuzzy clustering (FCM algorithm based on Euclidean distances) with 4 clusters of the elements B, Ca, Co, Cu, Mg, Na, Ni, Sr and pH of the Kola O-horizon data (isometric logratio transformed, standardised). The membership coefficients of the observations to the clusters are shown using a grey scale in the maps.

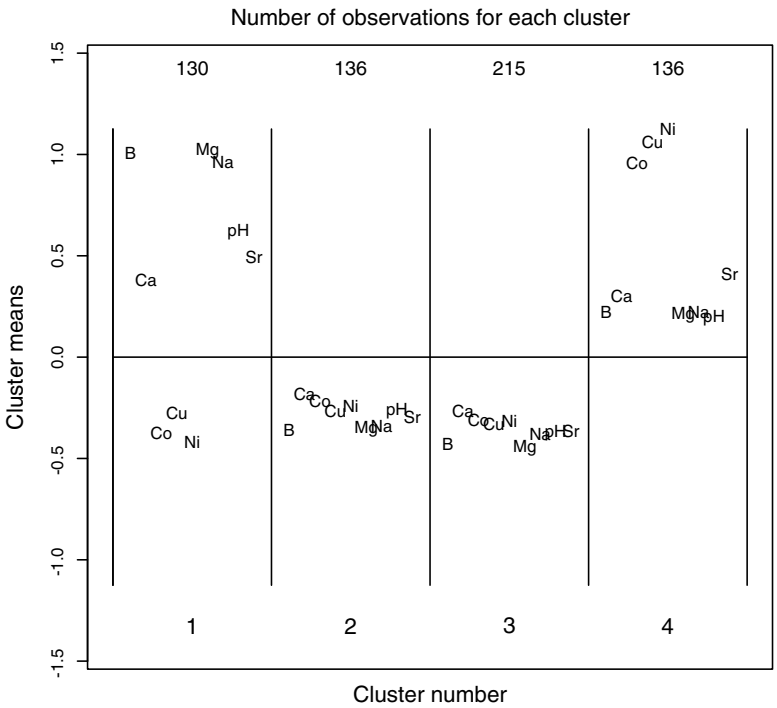

Fig. 8. Plot of the cluster centres of the fuzzy clustering results shown in Fig. 7 as a support for the interpretation of the clusters.

When working with few elements one needs to be aware that the choice of elements and adding or deleting just one element can have a drastic effect on the results of cluster analysis.

Further clustering results for the Kola Project data as well as for other geochemical data sets can be found in Templ (2003). This thesis also investigates and demonstrates the sensitivity of cluster analysis methods to data preparation and variable selection.

12. Conclusions

Like many other multivariate statistical methods, cluster analysis can be helpful in obtaining an overview of data sets with many observations and variables. It can be used to both structure the variables and to group the observations.

Of all tested data transformations simple log-transformation and isometric logratio transformation delivered the most reliable results. Because geochemical data are always compositional data the isometric logratio transformation is preferable, it can, however, not be used for variable clustering because the direct relationship to the elements is lost.

If the data show very different magnitude for the different variables (e.g., major, minor and trace elements mixed) the variables need to be standardised to mean 0 and variance 1. Following standardisation all variables will have the same influence on the results of cluster analysis. A high proportion of censored data in a variable can weaken the clustering structure. Censored variables should thus be excluded.

Euclidean and Manhattan distance performed best of the many available distance measures. Mclust does not need a distance measure.

Partitioning methods perform in general better than hierarchical methods when clustering a high number of observations. Mclust provided the most reliable and best interpretable results. Fuzzy clustering did not work with a high number of variables because of computational problems, but with a selected few variables it delivered informative results.

The global validity measures, which are aimed at assisting with the critical decision on the number of clusters, were not helpful in the actual problem. For geochemical data it turned out that mapping the location of clusters provides a good idea about the quality of clusters. The expected results with spatial data are regional clusters on the map. Too few clusters will not show clear regional patterns. When using too many clusters regional patterns tend to disappear. It is thus preferable and possible to provide a good estimate on the optimal number(s) of clusters via studying their regional distribution on a map. The rather subjective evaluation of single clusters on the map can be assisted via local validity measures.

Variable selection can be used to improve the results of cluster analysis. Variable selection can be based on expert knowledge or on the dendrogram obtained from variable clustering. Hierarchical clustering is ideal for variable clustering because the relationships between groups of variables become well visible in the dendrogram. The method of Ward is the best performing hierarchical clustering method, and even performs well for observation clustering. When performing cluster analysis with only a

few selected variables it is necessary to be aware that the addition or deletion of just one variable can change the results of cluster analysis drastically.

Overall, Mclust is an attractive method when working with geochemical data. It does not require the choice of a distance measure and computed cluster results for a whole range of numbers of clusters. The best result is then selected according to the BIC value which seems to be a sensible validity measure. Furthermore, the uncertainties of the assignment of each object to the clusters are returned and give valuable information on how well each observation fits to the clusters.

In general a powerful tool to plot the results in graphics and maps is needed in cluster analysis. A combination of maps and the plot of the cluster centres (and dendrograms for variable clustering) provide good insight into the results. The local validity measures can be displayed on the maps using different shades of grey. Different grey values on maps, representing the membership coefficient to a certain cluster, can be used to display the results of fuzzy clustering. Fuzzy clustering allows visualisation of the changing degree of membership of each observation to all clusters and is thus especially well suited to study the regional distribution of the clusters.

In general it is recommended to use cluster analysis as an exploratory method. For this purpose, the software package clustTool that runs under R has been developed. It is freely available and easy to handle via a graphical user interface. The user can choose data, coordinates, background maps, variables, different transformations, different distance measures, various cluster algorithms, determine the number of clusters, and look at the results in plots in a flexible way. The visual impression of the results, together with a pre-chosen validity measure is then helpful for deciding on the parameter selection for clustering the data. Informative results are not necessarily obtained by tuning the parameters for cluster analysis in a statistically optimal way. Expert knowledge should also be used for this purpose, e.g. for variable selection or cluster evaluation. This flexible software tool used by experts in an exploratory way can combine both strategies.

Acknowledgments

We are grateful for the valuable comments of the reviewers which resulted in an improvement of an earlier version of this manuscript.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Wiley, NY.
Aitchison, J., 1992. On criteria for measures of compositional difference. Math. Geol. 24, 365–379.
Aitchison, J., 2000. Logratio analysis and compositional distance. Math. Geol. 32, 271–275.
Aitchison, J., 2003. The Statistical Analysis of Compositional Data. Reprint. Blackburn Press, Caldwell, NJ, USA.
Aitchison, J., Barcelo-Vidal, C., Martin-Fernandez, J.A., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Math. Geol. 32, 257–271.
Bandemer, H., Näther, W., 1992. Fuzzy Data Analysis. Kluwer Academic Publication.
Barcelo-Vidal, C., Pawlowsky, V., Grunsky, E., 1996. Some aspects of transformations of compositional data and the identification of outliers. Math. Geol. 28 (4), 501–518.
Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, NY.
Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. Roy. Statist. Soc. B 26, 211–252.
Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.
Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.
Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (Eds.), 2006. Compositional data analysis in the geosciences – from theory to practice. Geological Society, London. Special Publication 264.
Butler, J.C., 1976. Principal components analysis using the hypothetical closed array. Math. Geol. 8, 25–36.
Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Statist. 3, 1–27.
Cohen, A.C., 1991. Truncated and Censored Samples. Marcel Dekker.
Davis, J.C., 1973. Statistics and Data Analysis in Geology. John Wiley and Sons, NY.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B 39, 1–38.
Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybernetics 3, 32–57.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueraz, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.
Everitt, B., 1974. Cluster Analysis. Heinemann Educational, London.
Filzmoser, P., Hron, K., 2008. Outlier detection for compositional data using robust methods. Math. Geol. 40 (3), 233–248.
Filzmoser, P., Reimann, C., Garrett, R.G., 2005. Multivariate outlier detection in exploration geochemistry. Comput. Geosci. 31, 579–587.
Fraley, C., Raftery, A., 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. 41, 578–588.
Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis and density estimation. J. Am. Statist. Assoc. 97, 611–631.
Frapporti, G., Vriend, S.P., Gaans, P.F.M., 1996. Trace element in the shallow groundwater of the Netherlands. A geochemical and statistical interpretation for the nation monitoring network data. Aquat. Geochem. 2, 51–80.
Friedman, J.H., Meulman, J., 2004. Clustering objects on subsets of attributes (with discussion). J. Roy. Statist. Soc. B 66, 815–849.
Gath, I., Geva, A., 1989. Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Anal. Intelligence 11, 773–781.
Gordon, A.D., 1999. Classification, second ed. Chapman & Hall/CRC, Boca Raton.
Gower, J.C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325–338.
Gustafson, D.E., Kessel, W., 1979. Fuzzy clustering with a fuzzy covariance matrix. Proceedings of the IEEE-CDC 2, 761–766.
Haldiki, M., Batistakis, Y., Vazirgiannis, M., 2002. Cluster validity methods. SIGMOD Record 31, 40–45.
Hartigan, J., 1975. Clustering Algorithms. John Wiley and Sons, NY.
Helsel, D.R., 1990. Less than obvious: statistical treatment of data below the detection limit. Environ. Sci. Technol. 24, 1767–1774.
Helsel, D.R., 2004. Non-detects and Data Analysis: Statistics for Censored Environmental Data. John Wiley and Sons, NY.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.
Ji, H., Zeng, D., Shi, Y., Wu, Y., Wu, X., 2007. Semi-hierarchical correspondence cluster analysis and regional geochemical pattern recognition. J. Geochem. Explor. 93, 109–119.
Ji, H., Zhu, Y., Wu, X., 1995. Correspondence cluster analysis and its application in exploration geochemistry. J. Geochem. Explor. 55, 137–144.
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data. John Wiley and Sons Inc., NY.
Leisch, F., 1999. Bagged clustering. Working Paper 51, SFB Adaptive Information Systems and Modeling in Economics and Management Science, Wirtschaftsuniversität Wien, Austria.
Leisch, F., 2006. A toolbox for *k*-centroids cluster analysis. Comput. Statist. Data Anal. 51, 526–544.
Le Maitre, R.W., 1982. Numerical Petrology. Elsevier, Amsterdam.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Springer, Berlin.
Martinetz, T., Berkovich, S., Schulten, K., 1993. Neural-gas network for vector quantization and its application to time-series prediction. IEEE Trans. Neural Networks 4, 558–569.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raftery, A.E., Dean, N., 2004. Variable selection for model-based clustering. Technical Report 452, Department of Statistics, University of Washington.

Reimann, C., Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ. Geol. 39, 1001–1014.

Reimann, C., Arnoldussen, A., Boyd, R., Finne, T.E., Koller, F., Nordgulen, O., Englmaier, P., 2007a. Element contents in leaves of four plant species (birch, mountain ash, fern and spruce) along anthropogenic and geogenic concentration gradients. Sci. Total Environ. 377, 416–433.

Reimann, C., Arnoldussen, A., Boyd, R., Finne, T.E., Nordgulen, O., Volden, T., Englmaier, P., 2006. The influence of a city on element contents of a terrestrial moss (*Hylocomium splendens*). Sci. Total Environ. 369, 419–432.

Reimann, C., Arnoldussen, A., Finne, T.E., Koller, F., Nordgulen, O., Englmaier, P., 2007b. Element contents in birch leaves, bark and wood under different anthropogenic and geogenic conditions. Appl. Geochem. 22, 1549–1566.

Reimann, C., Äyräs, M., Chekushin, V.A., Bogatyrev, I., Boyd, R., de Caritat, P., Dutter, R., Finne, T.E., Halleraker, J.H., Jæger, O., Kashulina, G., Niskavaara, H., Lehto, O., Pavlov, V., Räisänen, M.L., Strand, T., Volden, T., 1998. Environmental Geochemical Atlas of the Central Barents Region. NGU-GTK-CKE Special Publication, Norway.

Reimann, C., Filzmoser, P., Garrett, R.G., 2002. Factor analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 17, 185–206.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge.

Rock, N.M.S., 1988. Numerical Geology. Lecture Notes in Earth Sciences 18. Springer-Verlag, New York, Berlin, Heidelberg.

Sanford, R.F., Pierson, C.T., Crovelli, R.A., 1993. An objective replacement method for censored geochemical data. Math. Geol. 25, 59–80.

Templ, M., 2003. Cluster Analysis applied to Geochemical Data. Diploma Thesis, Vienna University of Technology, Vienna, Austria.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Statist. Assoc. 58, 236–244.

Yeung, K., Ruzzo, W., 2001. An empirical study on principal component analysis for clustering gene expression data. Bioinformatics 17, 763–774.