

# Automated integration of large geophysical data sets using three partitioning cluster algorithms: a comparison

H. Paasche<sup>1</sup>, D. Eberle<sup>2</sup>

1. University of Potsdam, Institute of Geosciences, Karl-Liebknecht-Str. 24, 14476 Potsdam-Golm, Germany, hendrik@geo.uni-potsdam.de
2. Council for Geoscience, Geophysics BU, Private Bag X112, Pretoria 001, South Africa, deberle@geoscience.org.za

## ABSTRACT

Since the advent of modern desktop computers, attempts have been made in various geoscientific fields towards rapid, automated and objective information extraction from suites of co-located data sets. Multivariate unsupervised classification techniques, such as cluster algorithms, have been proven valuable tools for largely automated information extraction and are for example routinely used for structural exploration and integration of multi-spectral remote sensing data sets. However, so far very few attempts have been made towards using unsupervised classification techniques for rapid, automated and objective information extraction from large geophysical data sets. In this study, we employ the crisp k-means, fuzzy c-means (FCM) and Gustafson-Kessel (GK) cluster algorithms and compare their suitability for rapid and largely automated integration of complementary geophysical data sets comprising airborne radiometric and magnetic as well as ground-based gravity data. All three data sets cover a survey area of 5000 km<sup>2</sup> located south-east of Johannesburg, South Africa. Integrated geophysical maps outlining dominant subsurface structures are obtained from each of the used cluster algorithm. Fuzzy cluster algorithms, such as the FCM and GK algorithm provide additional quantitative information about the trustworthiness of the detected subsurface units, which is considered very valuable when interpreting the finally obtained zonal maps. We will also show that the GK algorithm is most robust when it comes to the integration of data sets containing a few extreme anomalous values, e.g. as typically present in magnetic data sets, resulting in strongly skewed histograms of the data.

**Key words:** cluster analysis, airborne, Bushveld, multivariate statistics, data integration

## INTRODUCTION

Since the advent of modern desktop computers, attempts have been made in various fields of earth sciences towards rapid and automated information extraction from co-located data sets. For example, multivariate statistical analysis tools, such as cluster analyses (e.g., Kaufmann and Rousseeuw, 1990; Höppner *et al.*, 1999) have been employed for unsupervised classification of remote sensing data (e.g., Du and Lee, 1996; Ahn *et al.*, 1999), data-driven soil classification (e.g., Bragato, 2004), or the determination of facies distribution in groundwater systems (e.g., Güler and Thyne, 2004).

Cluster analysis is a generic term for different algorithms enabling the supervised or unsupervised classification of multi-parameter data sets into a number of groups or structures. A principal distinction is made between hierarchical and partitioning cluster algorithms. The former find successive clusters by division or agglomeration of previously established clusters, which often results in strongly nested clusters (e.g. Holliger *et*

*al.*, 2008), whereas the latter determine all clusters at once. Since partitioning cluster algorithms are computationally more expensive than hierarchical cluster algorithms, they have been used less frequently in the past, particularly in combination with large airborne geophysical data sets.

For example, Pirkle *et al.* (1984), Eberle (1993), Anderson-Mayes (2002), and Eberle *et al.*, (2005) used the crisp k-means cluster algorithm to analyse suites of airborne geophysical data sets aiming on the identification of statistically significant groups (clusters) on the basis of all available data sets with regard to geological or soil degradation mapping.

Innately, crisp cluster algorithms do not provide information about the statistical trustworthiness of the assignment of a sample to a certain cluster. Such trustworthiness information is by default obtained from fuzzy cluster algorithms, i.e., the fuzzy c-means (FCM) cluster algorithm, and has been found extremely valuable when analysing and interpreting a complementary geophysical data suite (Paasche *et al.*,

2006). Paasche and Eberle (in press) were probably the first employing the FCM cluster algorithm to integrate a suite of large airborne geophysical data sets for geological mapping and mineral exploration targeting. However, the Euclidian distance measure employed by the FCM cluster algorithm to detect distinct groups in the given multi-parameter data base requires the histogram distribution of the measured values of each of the considered data sets to closely match a Gaussian probability function. Processing the data sets prior to submitting those to FCM cluster analysis helps to roughly meet this condition. However, for data typically exhibiting a few extreme anomalous values, e.g., magnetic or electrical resistivity data, one may be left with more or less skewed histogram distributions. Alternatively, the GK algorithm employs a Mahalanobis norm (Mahalanobis, 1936) and thus should be able to overcome the limitations of the Euclidian-norm based crisp k-means and FCM algorithms and result in more reliable clustering results in case of data sets suffering skewed histograms.

In this study, we integrate the data base used by Paasche and Eberle (in press) using crisp k-means, FCM and GK cluster algorithms. To our knowledge, this is the first study to integrate a suite of geophysical data sets employing the GK algorithm. After a short introduction of the three cluster algorithms used in this study, we provide some information about the survey area and the available geophysical data base. Then, we will describe the preparatory processing of the data base and illustrate the results of the three clustering algorithms with regard to two different preparatory processing strategies of the airborne magnetic data set.

## CLUSTER ANALYSIS

Partitioning cluster algorithms group samples located in a multi-dimensional parameter space into a specified number of characteristic subsets or clusters by iteratively minimizing an objective function. Crisp and fuzzy cluster algorithms optimize the positions of a number of predefined cluster centres in the given parameter space. Crisp cluster algorithms, i.e. the k-means algorithm, assign each sample in the parameter space uniquely to one of the available clusters, whereas fuzzy cluster algorithms, i.e., the FCM or the GK algorithm, are devoted to the fuzzy concept of partial memberships of each sample in the parameter space to all clusters. This means, a sample may be mostly a member of a cluster, but it may also be a partial member of others. The degree of membership of the samples to the clusters, i.e., the membership values of the samples, varies between 0 and 1 and depends on its distance from the cluster centre. The membership values obtained from fuzzy cluster algorithms describe the multi-parameter data base in a fuzzy sense. However, there is need for the crisp solution emanating from the fuzzy membership information, which can be achieved by defuzzification of the fuzzy membership information (e.g., van Leeckwijk and Kerre, 1999). This goes along

with a certain information loss, which can partially be compensated by intelligent visual display.

### Crisp k-means algorithm

The crisp k-means cluster algorithm groups  $n$  samples in a  $t$ -dimensional space into a number of  $c$  clusters by iteratively minimizing the objective function

$$J = \sum_{i=1}^c \sum_{j=1}^n (d_j^{(i)} - v_i)^2 \quad (1)$$

by considering the within-cluster Euclidian distances between all samples  $d_j^{(i)}$  and the nearest cluster center  $v_i$ , where  $v_i$  is the mean of all  $d_j^{(i)}$ . The Euclidian distance measure used by the k-means algorithm privileges the finding of clusters spherically shaped in the given parameter space. An index vector  $h$  containing the cluster numbers of the  $n$  samples is obtained from

$$h_j = \arg \min (d_j - v_{1..c}). \quad (2)$$

Throughout this study, we initialize the algorithm using randomly guessed cluster centres as starting models. We repeat clustering 9 times employing different starting models each time to prove convergence. Since the optimum number of clusters suiting our data base best is not known a priori, we repeatedly perform k-means cluster analysis varying the number of clusters from 2 to 12. The solution with optimum  $c$  can then be identified by the statistical variance ratio criterion (VRC; Calinsky and Harabasz, 1974).

### Fuzzy c-means (FCM) algorithm

The FCM cluster algorithm extends the k-means algorithm by allowing for partial memberships of the samples to the  $c$  clusters. The objective function iteratively minimized is

$$J = \sum_{i=1}^c \sum_{j=1}^n (u_{ij})^f (d_j - v_i)^2, \quad (3)$$

where  $u_{ij}$  denotes the degree of membership of  $d_j$  to cluster  $i$  defined by its center  $v_i$ . Throughout this study, we set the fuzzification parameter  $f = 2$ , which is usually regarded a suitable choice (e.g., Hathaway and Bezdek, 2001). The Euclidian distance measure used by the FCM algorithm privileges the finding of clusters spherically shaped in the given parameter space. An  $n$ -element index vector  $h$  containing the number of the cluster each of the  $n$  samples has been assigned to, is obtained by membership defuzzification (e.g., van Leeckwijk and Kerre, 1999)

$$h_j = \arg \max (u_{1..c,j}). \quad (4)$$

Throughout this study, we initialize the algorithm using randomly guessed cluster centres as starting models. We repeat clustering 9 times employing different starting models each time to prove convergence. Since the optimum number of clusters suiting our data base best is not known a priori, we repeatedly perform FCM cluster analysis varying the number of clusters from 2 to 12. The solution with optimum  $c$  can then be identified by the Xie-Beni-index (XBI; Xie and Beni, 1991), which

relies upon the membership values and the distance of the two nearest cluster centres.

### Gustafson-Kessel (GK) algorithm

The GK algorithm extends the FCM algorithm by employing an adaptive distance norm (Mahalanobis norm) enabling the detection of clusters with different ellipsoidal shapes. The following objective function is iteratively minimized

$$J = \sum_{i=1}^c \sum_{j=1}^n (u_{ij})^f (d_j - v_i)^Y F_i (d_j - v_i), \quad (5)$$

The matrices  $F_i$  are used as additional optimization variables allowing each cluster to adapt its shape according to the distribution of the given data base. Since the GK algorithm is linear in  $F_i$ , we constrain  $F_i$  by

$$F_i = \frac{\det(C_i)^{\frac{1}{n}}}{C_i}, \quad (6)$$

where  $C_j$  is the fuzzy covariance matrix of the  $i$ -th cluster. A crisp solution of GK cluster analysis can be obtained by defuzzification of the obtained memberships using Equation 4.

Throughout this study, we initialize the algorithm using randomly guessed cluster centres as starting models. We repeat clustering 19 times employing different starting models each time to prove convergence. Since the optimum number of clusters suiting our data base best is not known a priori, we repeatedly perform GK cluster analysis varying the number of clusters from 2 to 12. As for the FCM algorithm, the solution with optimum  $c$  can then be identified by the Xie-Beni-index.

### SURVEY AREA AND DATA BASE

The survey area is situated in the East Rand, approximately 100 km south-east of Johannesburg, South Africa, covering more than 5000 km<sup>2</sup>. The mapped geology is fairly monotonous revealing Karoo-aged dolerite sills and sediments of the Eccla-group (Figure 1). Rock of the Bushveld Complex is concealed by the sediments of the Eccla Group in the northeast of the study area. Pyroclastic rock of the Rooiberg Group is the only sparse surface expressions of underneath Bushveld-type rock.

Geophysical data available from the study area are airborne radiometric and magnetic as well as ground-based gravity survey data (Paasche and Eberle, in press). The radiometric data were corrected for altitude, background and cosmic radiation and Compton scattering was removed from the individual channels. Data were mapped on a regular grid with 100 m spacing. The total natural gamma radiation from 0 to 3 MeV is shown in Figure 2a.

Airborne magnetic data were acquired using a Geometrics G822A caesium magnetometer. To clean the airborne magnetic data from all kind of magnetic fields coupled with ionospheric currents a base station

magnetometer was operated on the ground to record the diurnal variations of the Earth's magnetic field. Spurious magnetic fields induced by the fuselage of the aircraft were also removed and the vertical magnetic gradient was computed employing the same grid spacing as for the radiometric data (Figure 2b).

At approximately 450 stations, gravity data were collected on the ground. Bouguer anomaly values were calculated assuming a mean density of 2.67 g/cm<sup>3</sup>. The data were tied to the International Gravity Standardisation Net values and terrain corrections were applied to the data. Figure 2c shows the vertical gravity gradient map deduced from the terrain corrected data. Grid spacing is identical to those of the airborne geophysical data sets.

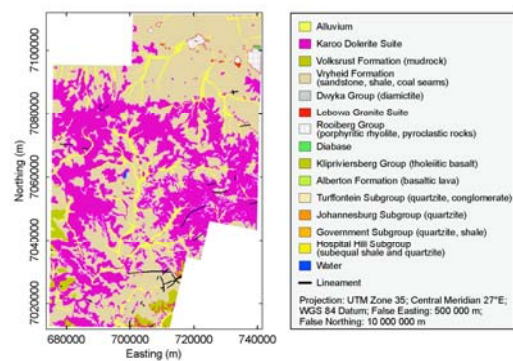


Figure 1. Geological map of the survey area (Paasche and Eberle, in press).

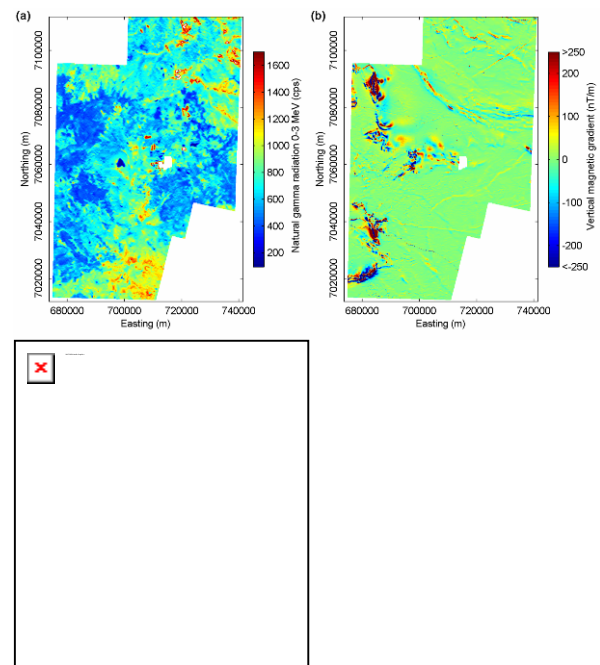
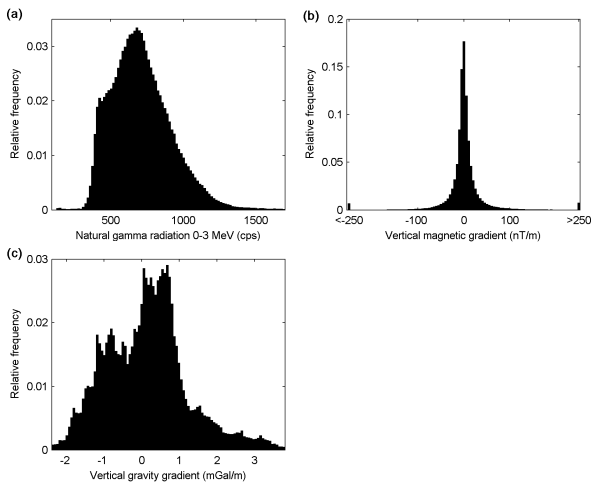


Figure 2. (a) Map displaying the total natural gamma radiation. (b) Vertical magnetic gradient map. (c) Vertical gravity gradient map. Maps (a) and (b) are compiled from airborne data, map (c) from gravity readings taken on the ground. The crosses indicate station locations. All maps are linearly equalized (modified from Paasche and Eberle, in press).

## Data processing and normalization

Prior to submitting the available data sets to clustering, we have to prepare and generate the parameter space to be clustered. Histograms of the available data sets (Figure 3) enable the identification of data sets with extremal anomalous values, which would produce long tails in the parameter space created by cross-plotting all data sets. Additionally, the dipolar nature of the magnetic data has to be overcome to ensure consistent clustering results and scaling of the data sets has to be unified to avoid perturbation of cluster analysis (e.g. Paasche *et al.*, 2006).



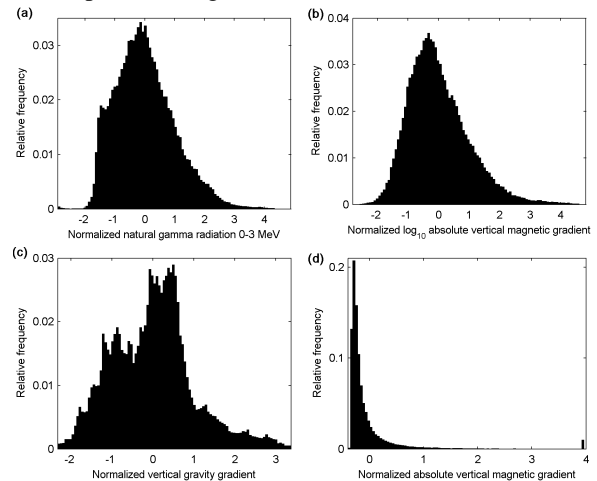
**Figure 3.** Histograms corresponding to the maps in Figure 2 (Paasche and Eberle, (in press)).

The radiometric data were submitted to a 2D median filter with a moving window of 600 x 600 m side lengths to reduce high-frequent ambient noise. Similarly, we applied a 2D median filter with 1 km x 1 km side length to the absolute values of the vertical magnetic gradient to overcome the dipolar nature of the magnetic data. The resulting histogram is strongly skewed with only a diminishing fraction of absolute data values being higher than 100 nT/m. Since k-means and FCM cluster analysis tend to produce spherically shaped clusters due to their Euclidian norm used to measure the distance from the samples to the cluster centres, in extreme cases several clusters might be required to accommodate the variance of the extreme data. To have the magnetic data set approach the requirements of the k-means and FCM algorithm, we minimize the skewness of the absolute magnetic gradient data applying Briggs logarithm. Lastly, we standardize the three processed data sets to overcome their different scaling. The normalized histograms of the processed radiometric, magnetic and the raw gravity data are shown in Figures 4a-c. These histograms are still not truly Gaussian distributed, but skewness has considerably been reduced.

To test the effect of extremely skewed data sets on the k-means, FCM and GK cluster algorithms, we prepared an alternative data base comprising the radiometric and gravimetric data sets processed and normalized as described above and the magnetic data normalized without prior taking of Briggs logarithm. The relevant

histogram is shown in Figure 4d and exhibits a significant skewness.

The normalized radiometric, magnetic, and gravimetric data sets as displayed in Figures 4a-c are now used to set up the parameter space where clustering is performed. This parameter space is referred to as “parameter space 1”. It is described by a 3-dimensional Cartesian parameter space. Each of the three axes is associated with one of the three input data types. Additionally, a second parameter space, “parameter space 2” is prepared for cluster analysis employing the normalized radiometric, magnetic and gravimetric data sets depicted in Figures 4a, 4c, and 4d.



**Figure 4.** Histograms after processing and normalization of the maps shown in Figure 2. Note, that the magnetic data have been processed using two different processing flows resulting in differently skewed histograms (cf. (b) and (d)).

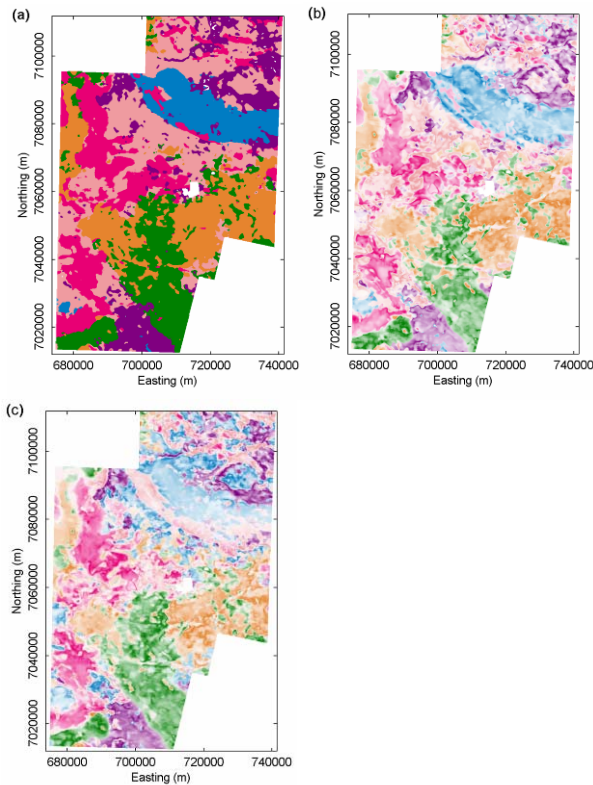
## RESULTS OF CLUSTER ANALYSES

Both parameter spaces have been clustered using the k-means, FCM and GK algorithm and integrated zonal geophysical maps have been compiled for each clustering result. We show here the 6-cluster solutions for all of the three cluster algorithms and both parameter spaces.

### Parameter space 1

The integrated zoned maps obtained by clustering the data sets in parameter space 1 using the k-means, FCM and GK algorithm are shown in Figure 5. The individual clusters are denoted by their colour. Since the crisp k-means cluster analysis does not assign each sample uniquely to the nearest cluster centre, no information is provided by the k-means algorithm about the trustworthiness of the classification (Figure 5a).

Fuzzy cluster algorithms, i.e. the FCM and the GK algorithm, also enable the generation of a crisp cluster solution by defuzzification of the obtained membership information. This membership information can be used to add extra value to the clustered zonal maps by incorporating some aspects of the membership information in the visualization of the zonal maps. Here, we follow Paasche *et al.* (2006) and set the colour

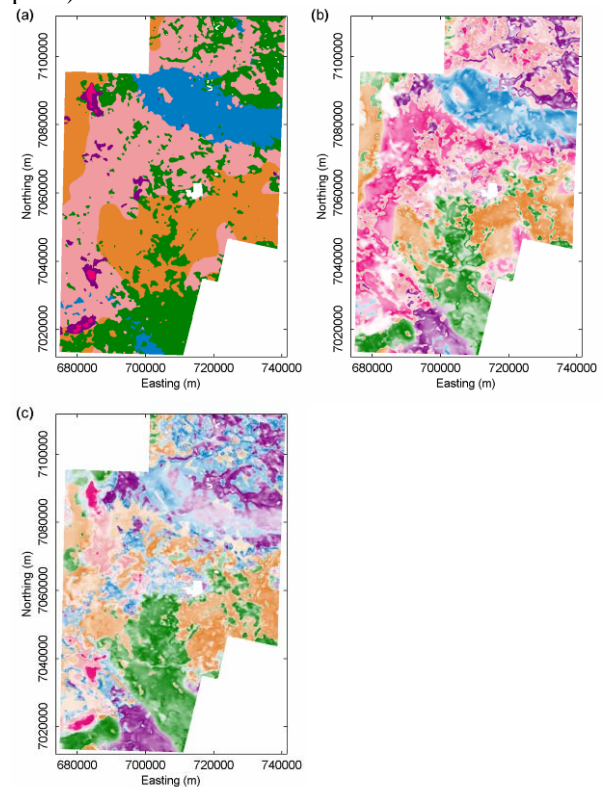


**Figure 5. Integrated zonal maps obtained from (a) k-means, (b) FCM, and (c) GK cluster analysis of the processed airborne data corresponding to the Histograms shown in Figures 4a-c. Clusters are denoted by their individual colors (cluster 1: red; 2: pink; 3: orange; 4: green; 5: purple, 6: blue). For the fuzzy clustering results, color saturation is proportional to the degree of membership a sample has been assigned to its cluster. Fully saturated samples are assigned to a cluster, whereas pale samples are merely between two or more clusters and cannot clearly be assigned to an individual cluster.**

saturation of each sample in the zonal map proportional to the membership of the sample to the cluster it has been assigned to during the defuzzification process. For example, a sample very typical for the mean properties of a cluster is plotted with high colour saturation; a sample lying rather in between two or more clusters is plotted with lower colour saturation. Thus, some information about the trustworthiness of the detected structures is added to the zonal map (see Figures 5b and 5c). However, when analysing the zonal map emanating from the FCM algorithm (Figure 5b), one can see that not only regions at transition zones between two different clusters suffer from low colour saturation. For example, at E683000/N7035000 there is a significant decrease in colour saturation in the middle of a region assigned to cluster 1 (red). Comparison to the three input data sets (Figure 2) reveals that this region corresponds to the most extreme vertical magnetic gradient anomaly. These values appear at the right edge of the histogram of the processed magnetic data (Figure 4b). Since the FCM algorithm preferentially detects spherically shaped clusters, with cluster centres at regions with high data density in the parameter space,

such few extreme values are rather far from all cluster centres hence suffering low membership values to all clusters. Statistically, this is a correct classification, but geologically, these values represent the most anomalous part of the high-magnetic gradient regions outlined by cluster 1 and should desirably not appear with lowered colour saturation in the zonal map.

The GK algorithm employs a Mahalanobis (1936) norm allowing ellipsoidal clusters, which are better suited to accommodate a few extremal values and assigning them a stronger classification to the nearest cluster centre. When comparing the zonal maps emanating from FCM and GK algorithms, it can be seen that the central areas of the regions outlined by cluster one generally exhibit higher colour saturations, which is closer to the geological interpretation of the provided data. Each of the six clusters present in the zonal map can be attributed to a geological unit (Paasche and Eberle, in press).



**Figure 6. The same as in Figure 5, but for differently processed magnetic data (Figure 4d replaces Figure 4b).**

## Parameter space 2

The integrated zoned maps obtained by clustering the data sets in parameter space 2 using the k-means, FCM and GK algorithm are shown in Figure 6. Comparison of the maps with those obtained for parameter space 1 shows that particularly the k-means and FCM cluster algorithm produced very different zonal maps. The first requires two nested clusters to accommodate the extreme anomalies of the vertical magnetic gradient whereas the latter merely ignores the few extreme values and does not use an extra cluster to accommodate them. The GK algorithm produces a zonal map that is

partly consistent with that obtained for parameter space 1, indicating an increased robustness with regard to data processing and scaling issues compared to the k-means and FCM algorithm.

## CONCLUSIONS

We have employed k-means, FCM and GK cluster algorithms to integrate three different geophysical data sets. Application of cluster algorithms requires some preparatory data processing to overcome different data scaling and to account for special requirements of the individual cluster algorithms. For data sets with almost Gaussian histograms, k-means, FCM and GK algorithm produce integrated zonal maps on the basis of all three data sets, which are structurally very similar. Fuzzy cluster algorithms enable adding extra value to the obtained zonal maps increasing the trustworthiness of the detected structures. The GK algorithm shows a significantly higher robustness regarding data scaling and skewness of data histograms.

## ACKNOWLEDGMENTS

The authors thankfully acknowledge the South African Council for Geoscience for making the geophysical data suite and geological information available for this study. This research work has been partly supported by the South African NRF (UID 69441), the International Bureau of the German BMBF (SUA 08/15), and the German DFG (PA1643/1-1).

## REFERENCES

- Ahn, C.W., Baumgardner, M.F. and Biehl, L.L., 1999, Delineation of soil variability using geostatistics and fuzzy clustering analyses of hyperspectral data: *Soils Science Society of America Journal* 63, 142-150.
- Anderson-Mayes, A.M., 2002, Strategies to improve information extraction from multi-variate geophysical data suites: *Exploration Geophysics* 33, 57-64.
- Bragato, G., 2004, Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain: *Geoderma* 118, 1-16.
- Calinski, T. and Harabasz, J., 1974, A dendrite method for cluster analysis: *Communication in Statistics* 3, 1-27.
- Du, C. and Lee, J.S., 1996, Fuzzy classification of earth terrain covers using complex polarimetric SAR data: *International Journal of Remote Sensing* 17, 809-826.
- Eberle, D., 1993, Geologic mapping based upon multivariate statistical analysis of airborne geophysical data: *International Institute for Aerospace Survey and Earth Sciences (ITC) Journal* 1993-2, 173-178.
- Eberle, D., Cole, J., Häuserer, M. and Stettler, E.H., 2005, Combined stochastic and deterministic modelling as an innovative approach to jointly interpret multi-method airborne geophysical data sets: 9<sup>th</sup> SAGA Biennial Conference, Cape Town, Expanded abstract.
- Güler, C. and Thyne, G.D., 2004, Delineation of hydrochemical facies distribution in a regional groundwater system by means of fuzzy c-means clustering: *Water Resources Research* 40, doi: 10.1029/2004WR003299.

- Hathaway, R.J. and Bezdek, J.C., 2001, Fuzzy c-means clustering of incomplete data: *IEEE Transactions on Systems, Man, and Cybernetics Part B*, 31, 735-744.
- Holliger, K., Tronicke, J., Paasche, H. and Dafflon, B., 2008, Quantitative integration of hydrogeophysical and hydrological data: Geostatistical approaches: in Darnault, C.J.G. (ed.), *Overexploitation and contamination of shared groundwater resources*: Springer, 67-82.
- Höppner, F., Klawonn, F., Kruse, R. and Runkler, T., 1999, *Fuzzy cluster analysis*: Wiley.
- Kaufmann, L. and Rousseeuw, P.J., 1990, *Finding groups in data: an introduction to cluster analysis*: Wiley.
- Mahalanobis, P.C., 1936, On the generalised distance in statistics: *Proceedings of the National Institute of Science of India* 12, 49-55.
- Paasche, H., Tronicke, J., Holliger, K., Green, A. and Maurer, H., 2006, Integration of diverse physical-property models: Subsurface zonation and parameter estimation based on fuzzy c-means cluster analyses: *Geophysics* 71, H33-H44.
- Paasche, H., Günther, T., Tronicke, J., Green, A., Maurer, H. and Holliger, K., 2007, Integrating multi-scale geophysical data for the 3D characterization of an alluvial aquifer: 13th EAGE Near Surface Geophysics Conference, Istanbul, Expanded abstract.
- Paasche, H. and Eberle, D., in press, Rapid integration of large airborne geophysical data suites using a fuzzy partitioning cluster algorithm: a tool for geological mapping and mineral exploration targeting: *Exploration Geophysics*.
- Pirkle, F.L., Howell, J.A., Wecksung, G.W., Duran, B.S. and Stablein, N.K., 1984, An example of cluster analysis applied to a large geologic data set: Aerial radiometric data from Copper Mountain, Wyoming: *Mathematical Geology* 16, 479-498.
- Van Leeckwijk, W. and Kerre, E.E., 1999, Defuzzification: criteria and classification: *Fuzzy Sets and Systems* 108, 159-178.
- Xie, X.L. and Beni, G., 1991, A validity measure for fuzzy clustering: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 841-847.