Integration and zonation of geophysical data sets with skewed histogramic data distribution using Gustafson-Kessel cluster analysis

H. Paasche (1), D. G. Eberle (2)

(1) University of Potsdam, Institute of Earth and Environmental Sciences, K.-Liebknecht-Str. 24, 14476 Potsdam-Golm, Germany, email: hendrik@geo.uni-potsdam.de
(2) Council for Geoscience, Geophysics Business Unit, Private Bag X112, Pretoria 0001, South Africa, email: deberle@geoscience.org.za

Summary
We employ the fuzzy Gustafson-Kessel (GK) cluster algorithm for rapid and objective integration of a disparate geophysical data base comprising airborne radiometric and magnetic as well as ground-based gravity data. The data base has previously been integrated using the fuzzy c-means (FCM) cluster algorithm, which required a significant amount of data pre-processing to meet the particular requirements of the FCM algorithm. Since the GK algorithm is more robust with regard to data scaling and histogramic data distribution, we repeatedly integrate the available data base while lowering the amount of preparatory data processing and scaling. The final results closely match those previously obtained by the FCM-based integration which had already been associated to known geological information in the survey area.

Introduction
Multivariate statistical analysis tools, such as partitioning cluster analyses (e.g., Höppner et al., 1999) have been proven valuable tools to quantitatively integrate complementary geophysical data sets for largely automated information extraction. Paasche and Eberle (2009) employed the fuzzy c-means (FCM) cluster algorithm to integrate a suite of airborne geophysical data sets for geological mapping and mineral exploration targeting. Since the FCM algorithm is not invariant to linear transformations, i.e. adding and/or multiplying constants to the data, the scaling/normalization of the data sets to be integrated impacts the finally detected structures. Additionally, the FCM algorithm struggles when it comes to the analysis of data sets with strongly skewed histograms, e.g., airborne magnetic data. Hence, Paasche and Eberle (2009) applied a preparatory data processing prior to data integration by FCM cluster analysis to meet the special requirements of this algorithm.

In this study, we integrate the data base used by Paasche and Eberle (2009) employing the fuzzy Gustafson-Kessel (GK) cluster algorithm. In its fundamental form, this algorithm is invariant to linear transformations and is therefore considered to be more robust with regard to the scaling of the employed data sets. Furthermore, it detects also the shape of the clusters and seems therefore more promising when it comes to the analysis of data with strongly skewed histograms. We will compare the integrated zonal map resulting from the application of the GK algorithm to those obtained by Paasche and Eberle (2009) to evaluate whether similar results can be obtained with the GK algorithm while reducing the amount of preparatory data processing and scaling.
Survey area and geophysical data base
The survey area is located 100 km southeast of Johannesburg, South Africa, covering more than 5000 km². The mapped geology is fairly monotonous, revealing Karoo-aged dolerites and sediments of the Ecca Group (Paasche and Eberle, 2009). Rocks of the Bushfeld Igneous complex are concealed by the Ecca sediments in the northeast of the study area.

Figure 1: Linearly equalized maps displaying (a) the total natural gamma radiation, (b) the vertical magnetic gradient, and (c) the vertical gravity gradient. (a) and (b) are compiled from airborne data, (c) from gravity readings taken on the ground. Crosses indicate station locations.

Figure 2: (a) - (c) Histograms corresponding to the data illustrated in Figure 1, respectively. (d) Histogram for the absolute vertical magnetic gradient resulting from the data shown in Figure 1b.

Data available from the study area comprise airborne radiometric and magnetic as well as ground-based gravity survey data (Figure 1). The radiometric data were corrected for altitude, background radiation, cosmic radiation and Compton scattering. To clean magnetic data from all kinds of magnetic fields coupled with ionospheric currents a base station magnetometer was operated on the ground. Additionally, removal of spurious magnetic fields induced by the fuselage of the aircraft was enabled and the vertical magnetic gradient has been computed.
Gravity data were collected at approximately 450 stations in the survey area. Data are Bouguer and terrain corrected. Vertical gravity gradients have been computed from the corrected data.

Figures 2a-c show the histograms corresponding to the mapped data shown in Figure 1. Since cluster analyses can not handle the dipolar nature of the vertical magnetic gradient, i.e. an anomalous subsurface structure produces a higher-than-average and a lower-than-average anomaly at the same time, we consider the absolute vertical magnetic gradient for all following analyses. The histogram of the absolute vertical magnetic gradient is significantly skewed and shown in Figure 2d.

**FCM and GK cluster analysis**

Partitioning cluster algorithms, such as FCM or GK (e.g., Höppner et al., 1999) group data points located in an $n$-dimensional space into a specified number of clusters. In this study $n=3$ since each sample in the $n$-dimensional space comprises three attribute values, which is one for each of the performed radiometric, magnetic and gravity surveys.

The unknowns in the FCM cluster analysis are the positions of the $c$ cluster centers in the $n$-dimensional space and the membership values of each sample. The fuzzy concept allows for partial memberships; that is, a sample may be mostly a member of a cluster, but it may also be partial member of others. The degree of membership of a sample to its cluster depends on its distance from the corresponding cluster center. The FCM algorithm employs a Euclidian norm to measure the distance of a sample from a cluster center. Hence, the FCM tends to detect spherical clusters of roughly equal size and is not invariant to the scaling of the data sets. For example, stretching the scale of one attribute axis of the $n$-dimensional space results in different Euclidian distance measures and perturbs the results of the FCM cluster algorithm. Furthermore, the detection of spherical clusters usually struggles when a data set exhibits a few anomalous values, which result in long tails in the $n$-dimensional space and significantly skewed histograms related to this data set (Figure 2b). When integrating the present data base using the FCM algorithm, Paasche and Eberle (2009) applied Briggs logarithm to the absolute vertical magnetic gradient data prior to cluster analysis to reduce the skewness of the data histogram.

In the GK algorithm the Euclidian distance measure of the FCM algorithm is replaced by a Mahalanobis norm (Mahalanobis, 1936). Hence, the GK algorithm determines additionally to the center positions and membership information also the shape of the clusters and thus results in rather ellipsoidal clusters. Stretching of one or several axes of the $n$-dimensional space affects the covariance of each cluster and stretches the ellipsoidal shape of the clusters. Hence, data normalization based on linear transformations, which is essentially required for a good performance of the FCM algorithm, is obsolete when employing the GK algorithm. Furthermore, ellipsoidal cluster appear more appropriate when it comes to capture the heterogeneity present in data sets with strongly skewed histograms. Since often only a few values scatter along such tail, it is beneficial to constrain the size of a cluster considered for capturing such long tails.

**Integrated zonal geophysical maps**

The geophysical data sets corresponding to the histograms in Figures 2a, c, and d serve as input for the GK cluster analysis. We employ 7 clusters and reduce a priori the size of one cluster to 50 % of the size of the others. This cluster is reserved to capture all samples related to the few extremely high absolute values of the vertical magnetic gradient (see the last class.
in the histogram depicted in Figure 2d). The zonal geophysical map emanating from the GK cluster analysis outlines dominant subsurface units on the basis of all input data sets (Figure 3a). Cluster 7 outlines the regions of extreme magnetic anomalies, which can be associated to ferruginous shale/quartzite of the Witwatersrand Supergroup. The other six clusters closely match those found by Paasche and Eberle (2009) when applying the FCM algorithm to further processed and normalized data. The detected structures have been associated to known geological information (Paasche and Eberle, 2009).

**Figure 3:** Zonal maps obtained from (a) GK cluster analysis of low-level processed and non-normalized data and (b) FCM cluster analysis of processed and normalized data (for details see Paasche and Eberle, 2009). Clusters are denoted by their colors (cluster 1: red; 2: pink; 3: orange; 4: green; 5: purple; 6: blue; 7: cyan). Color saturation is proportional to the degree of membership a sample has been assigned to its cluster. Fully saturated samples are assigned to a distinct cluster, whereas pale samples are merely between two or more clusters.

**Conclusions**

We have employed Gustafson-Kessel (GK) cluster analysis to integrate a disparate geophysical data base comprising airborne radiometric and magnetic data and ground-based gravity data. The histogram related to the magnetic data is extremely skewed and no data scaling or normalization is applied prior to GK cluster analysis. The resultant integrated zonal geophysical map outlines a number of subsurface structures, which closely match the structures previously identified and validated when employing the fuzzy c-means (FCM) cluster algorithm for the integration of the same data base after a significant amount of preparatory processing. However, employing the GK algorithm required usage of an additional cluster of reduced size to accommodate regions related to a few samples with very high absolute magnetic gradients.

**Acknowledgements**

The authors thankfully acknowledge the South African Council for Geoscience for making the data base 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**

Höppner, F., Klawonn, F., Kruse, R., Runkler, T., (1999), Fuzzy 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., Eberle, D. G., (2009), 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 40, 277-287.