A CASE STUDY ON GEOCHEMICAL ANOMALIES IDENTIFICATION THROUGH PRINCIPAL COMPONENTS ANALYSIS SUPPLEMENTARY PROJECTION

Henrique Garcia Pereira[a,*], **Sara Renca**[b], **José Saraiva**[a]

[a]*CVRM – Geosystems Center, IST, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*

[b]*Earth Sciences Dept., Univ. Coimbra, Lg. Mq. Pombal, 3000-272 Coimbra, Portugal*

*Corresponding author. Fax: +351 21 8417442; Phone: +351 21 8417247;

E-mail address: hpereira@alfa.ist.utl.pt

ABSTRACT

Based on a case study in which a single geochemical anomaly was located in the vicinity of an abandoned mine in Central Portugal, a recursive methodology for anomaly/background separation was developed. This methodology relies on the supplementary projection of each of the samples taken from a subset of 'anomaly candidates' onto the axes provided by Principal Components Analysis of the background subset. The concept of 'anomaly intensity', defined by the average of the distances from the original to the supplementary projections, is the basis for final anomaly identification.

KEYWORDS: Anomaly separation, Supplementary projection, Principal Components Analysis, Anomaly intensity

INTRODUCTION

A variety of multivariate statistical methods have been applied to the identification of metallic geochemical anomalies, namely classical discriminant analysis (Bull and Mazzucchelli, 1975), empirical ranking using an a priori index (Smith and Perdrix, 1983), and canonical variate analysis (Smith et al., 1984). These applications rely on 'external' information or on the availability of training sets (a comprehensive review of such methods is given in Howarth and Sinding-Larsen, 1983). When the 'anomaly' is considered as a set of outliers that 'emerge' from the background, the point stressed (Singh et al., 1994) is the characterisation of the background in terms of 'robust' statistics referring to each variable, one at a time. Another author (Garrett, 1989a) introduced in geochemistry the concept of robust multivariate procedures. The generalization of cumulative probability plots to the multivariate situation was also proposed (Garrett, 1989b), through the use of the chi-square plot based on the Mahalanobis distance. This approach is very promising, even though it is not completely distribution-free (a flavour of multi-gaussianity is needed for anomaly selection). Other approaches (Cheng et al., 2000), although more 'realistic' in their assumptions and 'sophisticated' in their algorithms, call for large data sets, which are not always available in practice.

In the study reported here, where a priori information is scarce and the number of samples is exiguous, a new method for anomaly separation, interpretation and quantification was devised, based on the aim of geometrically isolating two systems of relationships, rather than focusing on each variable independently.

The proposed methodology relies on the geometric properties of Principal Components Analysis of Standardized Data [cf. Vairinho et al. (1990), for the detailed description of the algorithm]. Once a group of 'anomaly candidate' samples are selected, through the inspection of their projections onto 'significant' components (from the viewpoint of the relationship between the variables that are driving the process of separation of the two populations), they are projected as 'supplementary individuals' (Greenacre, 1984) onto the axes provided by the eigenvalue decomposition of the cosine matrix (which is the analogue of the correlation matrix in the geometric formulation) of the provisional background data subset. Then, the 'anomaly intensity' is calculated by the average shift, measured along the significant components, between the original and the supplementary projection of each sample (or variable). A recursive procedure is then implemented, by choosing the subset of samples that maximises 'anomaly intensity'. The aim of this paper is to illustrate, in a comprehensive case study, a PCA-based geometrical method for anomaly separation, which does not call for large data sets or external information.

PROPOSED METHODOLOGY

Rationale

In order to cope with the difficulties in defining and quantifying globally a 'geochemical anomaly' on the grounds of a small set of empirical data of concentrations of p elements at n sampling sites, a new methodology based on Principal Components Analysis of Standardized Data (PCASD) was devised.

The rationale behind this methodology is guided by the following argument:

(i) A 'geochemical anomaly' depends on the interaction of the measured variables and cannot be accounted for by the values of those variables, taken per se.

(ii) In multi-element surveys, the separation of a 'geochemical anomaly' from the background should not be dealt with in univariate terms, but rather in the scope of a multivariate technique like PCASD, in which samples and variables are projected onto the same factorial space ['biplots', cf. Jolliffe (1986), p. 75; 'RQ-mode PCA', cf. Zhou et al. (1983)].

(iii) The criterium for identification of a sample as contributor to the anomaly relies on quantifying the 'disturbance' which that sample induces in the background system of relationships.

The proposed methodology relies essentially on the supplementary projection of samples and variables in the framework of PCASD, which is equivalent to "Simultaneous R- and Q-Mode Factor Analysis" [as named by Davis (1986), p.594]. In this framework, both individuals and properties of the original matrix **Z** can be projected onto the same space, on the grounds that the scaling process guarantees that similarity between individuals is measured by an Euclidean distance (Zhou et al., 1983).

In the framework of PCASD, the supplementary projection of a block **A** of the matrix **Z** onto another block **B** of **Z** consists simply of the calculation of **AV**, where **V** is the eigenvector matrix given by the decomposition of the cosine matrix that refers exclusively to **B** (this matrix contains the cosines of the

angles between points representing properties in the space of individuals). In geometric terms, this corresponds to finding the positions of the individuals belonging to **A**, in the space created by the individuals of **B** (responding to the objective of quantifying the 'disturbance' of **B** created by **A**). More than a mere exploratory data reduction technique, PCA is viewed here as a powerful geometrical method that unifies the traditional 'scores' of samples and 'loadings' of variables in terms of projections onto axes. This method can be used for referring the position of a set of samples on the space created by another set of samples by the 'supplementary projection' procedure, allowing also for interpretation in terms of variables in the simultaneous projection graphs.

Description

In Figure 1, the proposed methodology is described in terms of a three step flowchart.

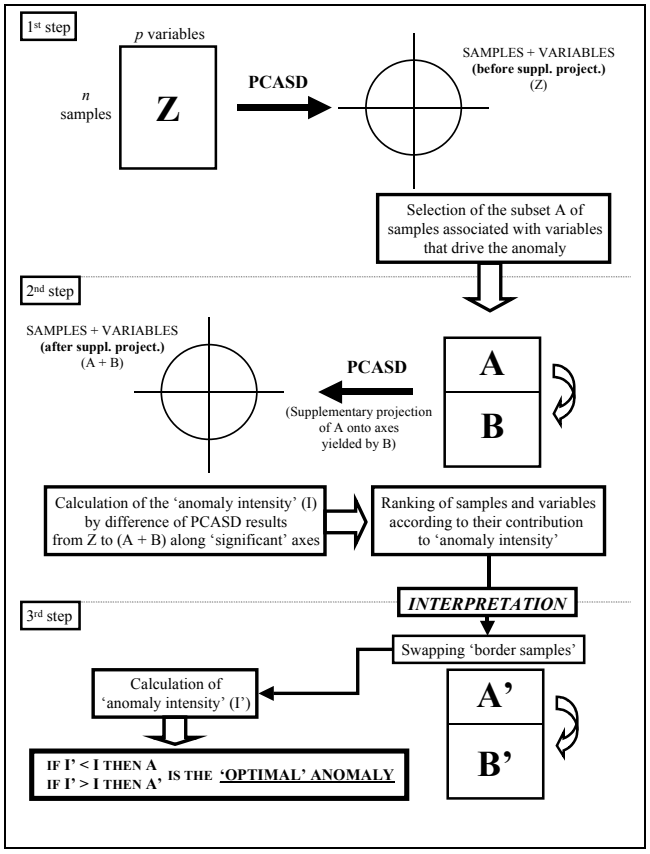

Figure 1
Flowchart of the proposed methodology

The first step of the methodology can be summarized as follows: given a matrix **Z** of *n* individuals (samples), for which *p* properties are available (the concentrations of the *p* elements measured in each sample), the problem of anomaly identification consists of selecting, in **Z**, two subsets: one referring to the background (Block **B**), and the other to the anomaly (Block **A**).

When **Z** is submitted to PCASD, the resulting axes can generally be interpreted in terms of the association of properties in a given geological and geochemical context. Also, those axes permit the separation of regions, in factor space, that are more likely to be linked to an 'anomaly' than to the

background. This is done by choosing those samples whose coordinates on 'significant' axes are higher ('significant' axes being interpreted on the grounds of the variables that are associated to the anomaly). Moreover, once the regions of factor space are identified where variables likely to cause the 'anomaly' are projected, it is easy to recognize the samples associated with those variables, due to the simultaneous projection of individuals and properties in the same space. Hence, this subset of samples is coined 'anomaly candidates', and the entire **Z** matrix is split (provisionally) into two blocks – Block **B** (background) and Block **A** (anomaly).

The second step of the methodology consists of the supplementary projection [in the framework of PCASD, cf. Lebart et al. (1984)] of the 'anomaly candidate' samples (contained in Block **A**) onto the axes provided by the eigen value decomposition of the cosine matrix referring to the background samples (contained in Block **B)**.

This second step provides the 'anomaly intensity' for each sample, given by the distance, along 'significant' axes, between the positions of the sample, before and after its supplementary projection ('significant' axes are those previously interpreted on the grounds of the variables that contribute to the anomaly, in a given geochemical context).

This distance measures the dissimilarity between a sample viewed as part of the whole set and the same sample segregated into the anomaly subset and viewed in the framework of the complementary background. The distance can be seen as the contribution of the sample to the global 'anomaly intensity', since it measures the importance of each sample in the anomalous subset (the more that the 'significant' sample properties are different from the background, the bigger is this distance).

The output of this second step is the calculation of a global intensity of the anomaly, obtained by averaging the above defined distances for the entire set of 'anomaly candidates', **A**. This 'anomaly intensity' quantifies the shift induced on the 'anomaly candidates' by positioning them in a 'provisional' background multivariate structure, given by the components of PCASD referring exclusively to **B**. Hence, the 'anomaly intensity' measures how much the background is 'disturbed' by the presence of anomalies, functioning as a criterium for judging the goodness of the provisional separation.

Subsequently, in a third step, the samples that lie in the neighbourhood of the frontiers that define the anomaly regions in factor space ('border samples') can be moved from their initial subset to the other, until the global 'anomaly intensity' is maximised. At this point, the optimal separation of the anomaly subset from the background is reached.

CASE STUDY

The above described methodology was applied to a geochemical prospecting campaign undertaken in Central Portugal, in the vicinity of the Zorro mine. This abandoned mine is located in the Central Iberian geotectonic Zone, near the contact with the Ossa Morena Zone (Fig. 2). The mineralization consists of outcropping veins of quartz (locally dolomite and ankerite), with argentiferous galena and sphalerite. The veins have a NW-SE strike and dip 60º to the East.

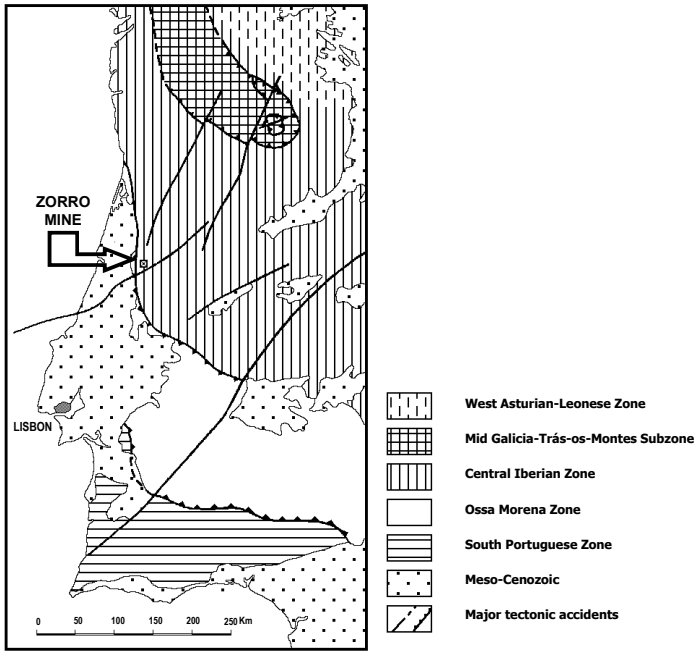

Figure 2
General geological context of the Zorro abandoned mine

The geochemical campaign provided a data set of 45 samples, taken from the B horizon of the soil, at a regular 20 meter interval along two lines, separated by 200 m and perpendicular to the strike of the veins. The first 18 samples came from the northernmost line, the other 27 from the southern line. The fraction smaller than 80 mesh (177 μm) was recovered for each sample, completely solved in a mixture of acids (HF, HNO_3 and $HClO_4$), and the concentration of the following elements was determined: Ag, Ba, Co, Cr, Cu, Fe, Li, Mn, Ni, Pb, Ti, Zn, As, Sb and Sn. The pH was also measured for each sample.

The empirical data, in the form of a nxp matrix (n=45 samples, p=16 variables), was submitted to the PCASD algorithm, giving rise to the projection of variables shown in Figure 3.

The interpretation of Figure 3 is clear – axis 1 opposes the mineralization elements (arranged in two groups – Cu and Zn, and As, Ag, Sb and Pb) to the soil components, associated to pH. Hence, the farther a sample projects from the origin, on the positive side of axis 1, the greater is its 'anomalous' character. This leads to the conclusion that the only 'significant' axis is the first, and further discussion refers only to this axis.

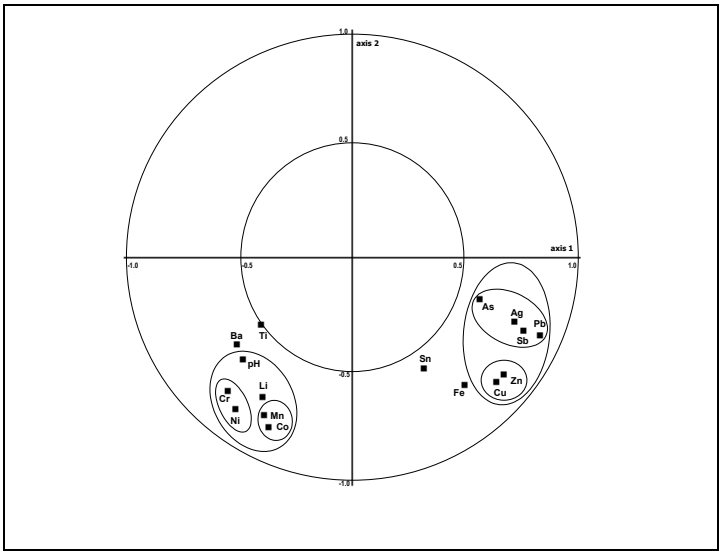

Figure 3
Projection of variables onto the first factorial plane
produced by PCASD of the entire data set
(explaining 61% of the inertia contained in the correlation matrix)

The projection of samples onto the same plane is given in Figure 4. Selection of the 'anomalous candidate samples' is straightforward, by sorting their positive co-ordinate on axis 1 in descending order: 6, 13, 7, 8, 10, 9, 11, and 12. The group of samples 39, 40, 43, 36 and 41, although projected onto the positive side of axis 1, was not selected, at this stage, as 'anomaly candidates', as these samples appear to represent only an extension of the background distribution.

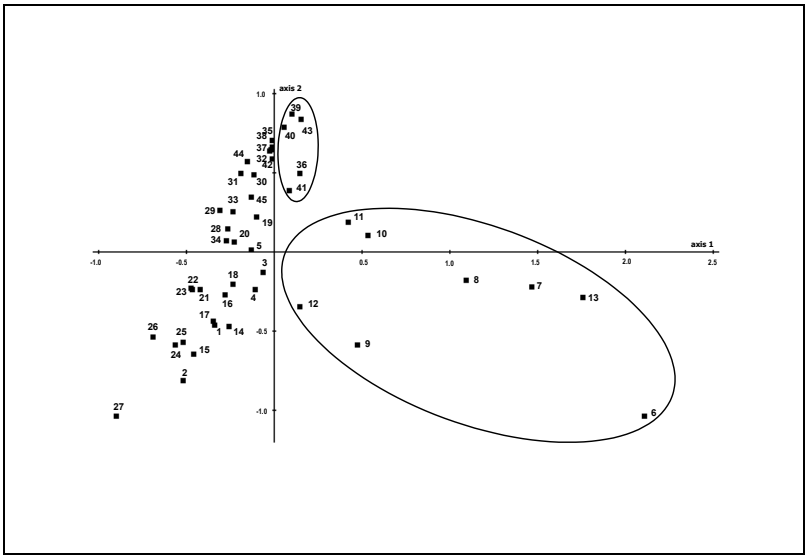

Figure 4
Projection of samples onto the first factorial plane
produced by PCASD of the entire data set
(explaining 61% of the inertia contained in the correlation matrix)

Hence, the total data set was split into a 'background matrix' of 37x16 elements, and an 'anomaly matrix' containing the selected 8 samples. The supplementary projection of the samples contained in this 'anomaly matrix' onto the axes provided by PCASD of the 'background matrix' is shown in Figure 5. As expected, the anomalous samples move away from the origin, in the direction of increasing positive co-ordinates on axis 1, and their 'importance' is reversed in some cases (now the sequence is 13, 6, 7, 8, 9, 10, 12, 11).

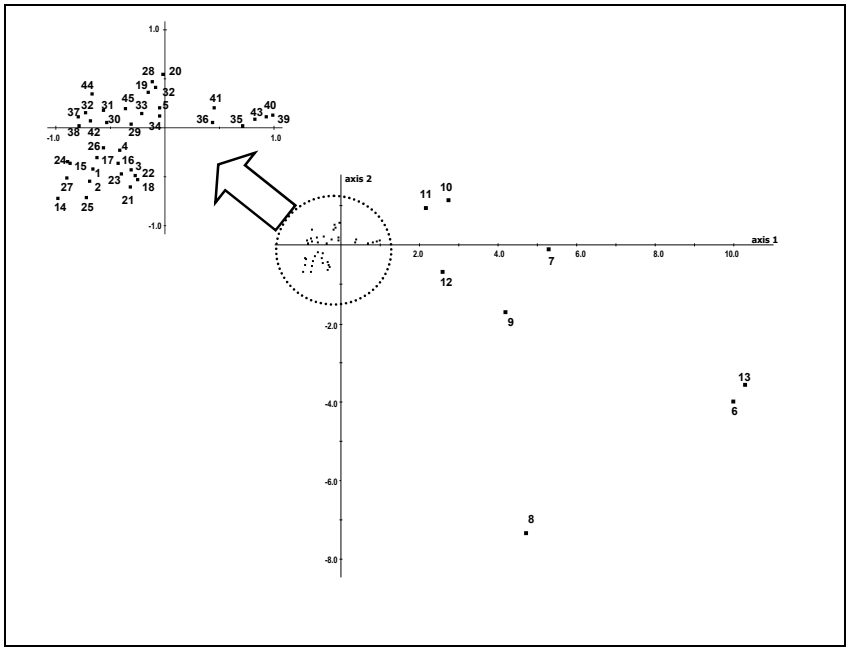

Figure 5
Projection of samples onto the first factorial plane
produced by PCASD of the background subset
(supplementary projection of anomalous samples)

With regard to variables, their position before and after the supplementary projection can be visualised in the same graph – see Figure 6. As expected, the variables that are associated to the anomaly (Cu and Zn, and As, Ag, Sb and Pb) shift towards the origin when the supplementary projection of the anomalous samples is compared to the original one (revealing a more uniform configuration, characteristic of the background).

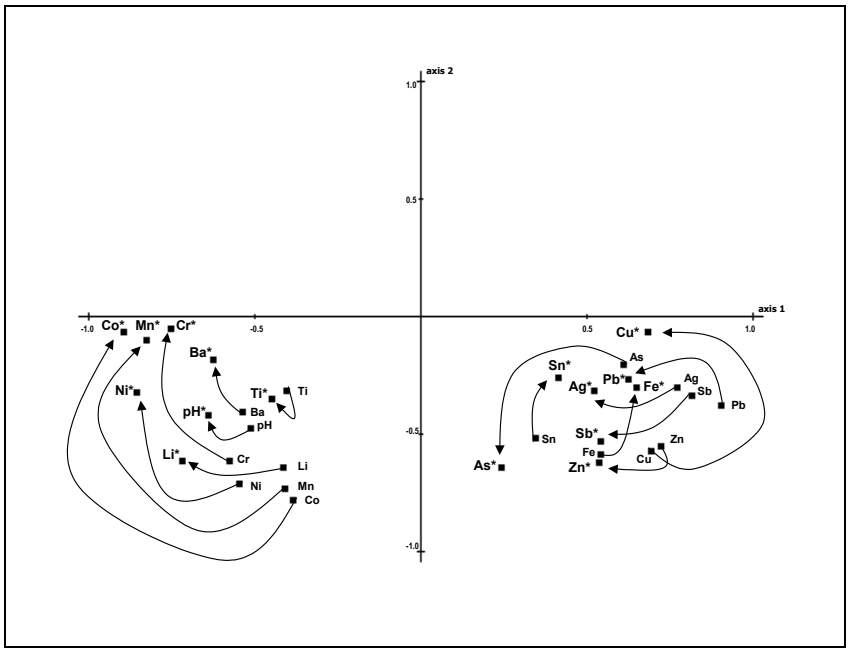

Figure 6
Plot of variables before and after
supplementary projection of the anomalous samples
(supplementary projection of each variable is denoted by *)

The importance of each variable in the anomaly formation can be quantified by the distance *dj1* (distance, along axis 1, between the positions before and after supplementary projection, cf. Fig. 6). These values are given in Table 1, showing the ranking of the 'interesting variables' for this specific anomaly.

Table 1
Distance between variable positions along axis 1, before and after supplementary projection

Variable	*dj1*
As	0.355
Pb	0.271
Sb	0.264
Ag	0.244
Zn	0.177
Cu	0.003

Regarding samples, their 'anomaly intensities' can be measured by the distance between their positions, along axis 1, before and after supplementary projection, as given in Table 2 by *di1* (note that a new sequence is obtained, now ranking samples in terms of their contribution to 'anomaly intensity').

Table 2
Distance between sample positions along axis 1, before and after supplementary projection

Sample	*di1*
13	8.523
6	7.870
7	3.789
9	3.690
8	3.598
12	2.424
10	2.178
11	1.722

The relative positions of anomalous samples in the plane of axes 1 and 2 are shown in Figure 7 by their locations before and after the supplementary projection.

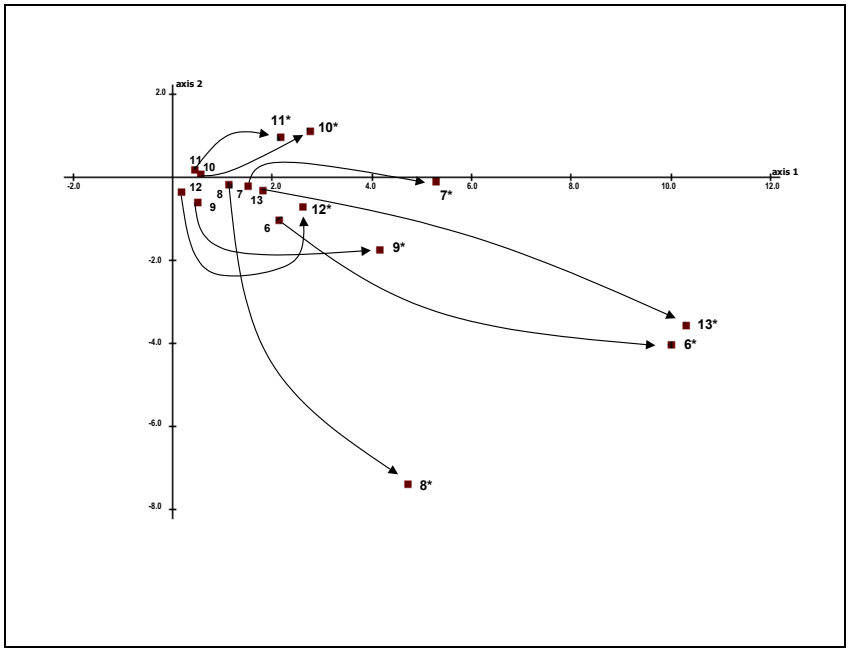

Figure 7
Plot of samples before and after
supplementary projection of the anomalous samples
(supplementary projection of each sample is denoted by *)

The global intensity of the anomaly, obtained by averaging the values of *di1* from Table 2, is 4.242. By moving samples 36, 39, 40, 41 and 43 - which have also a positive coordinate on axis 1 (see Fig. 4) -, one at a time, from the background subset to the anomaly subset, no further increase is obtained (the new global anomaly intensities are always lower than 4.242). Hence, the separation initially performed between anomaly and background cannot be improved.

CONCLUSIONS

The proposed methodology proved to be a reliable procedure for anomaly identification based on the relationships between concentrations in a set of elements determined in small sample sets. Moreover, it calls for a minimum of 'external' information and provides a good visualisation of the interdependence between variables, thus facilitating an interpretation that supports the selection of 'anomaly candidates' on factor planes. The underlying pre-requisite is only that the samples are representative of background/anomaly multivariate structure (in the sense that it allows to choose visually the 'anomaly candidates' in the PCASD graphs). The 'anomaly intensity' concept is applied to the anomaly/background optimisation procedure, which cannot rely on random selection of 'anomaly candidates'. The methodology does not provide the internal structure of each anomaly, in the sense that it cannot cope with multiple type anomalies, but it is viewed as a first step for anomaly comparison in different contexts.

ACKNOWLEDGEMENTS

The review of an early version of this paper contributed to improve its clarity. The authors are indebted to the referees, who put a serious effort into a deep analysis of the proposed methodology.

REFERENCES

Bull, A. J., Mazzucchelli, R. H., 1975. Application of discriminant analysis to the geochemical evaluation of gossans. In: Elliott, I. L., Fletcher, W. K. (Eds.), Geochemical Exploration 1974. Elsevier, Amsterdam, pp. 219-316.

Cheng, Q., Xu, Y., Grunsky, E., 2000. Integrated spatial and spectrum method for geochemical anomaly separation. Natural Resources Res. 9, 43-51.

Davis, J. C., 1986. Statistics and Data Analysis in Geology, Second Edition. Wiley, New York.

Garrett, R. G., 1989a. A robust multivariate allocation procedure with applications to geochemical data. In: Agterberg, F. P., Bonham-Carter, G. F. (Eds.), Statistical Applications in the Earth Sciences. Geological Survey of Canada, pp. 309-318.

Garrett, R. G., 1989b. The chi-square plot: a tool for multivariate outlier recognition. Jour. Geochem. Explor. 32, 319-341.

Greenacre, M., 1984. Theory and Applications of Correspondence Analysis. Academic Press, London.

Howarth, R. J., Sinding-Larsen, R., 1983. Multivariate analysis. In: Howarth, R. J. (Ed.), Statistics and Data Analysis in Geochemical Prospecting, Handbook of Exploration Geochemistry, vol. 2. Elsevier, Amsterdam, pp. 207-289 (Chapter 6).

Jolliffe, I. T., 1986. Principal Components Analysis. Springer-Verlag, Berlin.

Lebart, L., Morineau, A., Warwick. K. M., 1984. Multivariate Descriptive Statistical Analysis. Wiley, New York.

Singh, A., Singh, A. K., Flatman, G., 1994. Estimation of background levels of contaminants. Math. Geology 26, 361-388.

Smith, R. E., Perdrix, J. L., 1983. Pisolitic laterite geochemistry in the Golden Grove massive sulphide district, Western Australia. Jour. Geochem. Explor. 18, 131-164.

Smith, R. E., Campbell, N. A., Litchfield, R., 1984. Multivariate statistical techniques applied to pisolitic laterite geochemistry at Golden Grove, Western Australia. Jour. Geochem. Explor. 22, 193-216.

Vairinho, M., Fonseca, E. C., Pereira, H. G., 1990. Discrimination of gossans using principal components analysis of standardized data. Jour. Geochem. Explor. 38, 375-394.

Zhou, D., Chang, T., Davis, J. C., 1983. Dual extraction of *R*-Mode and *Q*-Mode factor solutions. Math. Geology 15, 581-606.