ELSEVIER

The Science of the Total Environment 281 (2001) 99–109

the Science of the Total Environment
An International Journal for Scientific Research into the Environment and its Relationship with Man

www.elsevier.com/locate/scitotenv

Multivariate outlier detection and remediation in geochemical databases

Gerald C. Lalor, Chaosheng Zhang*

International Centre for Environmental and Nuclear Sciences, University of the West Indies, Mona Campus, Kingston 7, Jamaica

Received 17 April 2000; accepted 2 April 2001

Abstract

In this study, outliers are classified into three types: (1) range outliers; (2) spatial outliers; and (3) relationship outliers, defined as observations that fall outside of the values expected from correlation within the dataset. The multivariate methods of principal component analysis (PCA), multiple regression analysis (MRA) and an autoassociation neural network (AutoNN) method are applied to a dataset comprising 203 samples of rare earth element (REE) concentrations in soils of Jamaica which shows the expected good correlations between the elements. PCA is shown to be effective in detection of high value range outliers, while AutoNN and MRA are effective in detection of relationship outliers. A backpropagation neural network was used to predict the 'expected values' of the outliers. Four obvious relationship outliers with unexpected low Sm concentrations were selected as an example for remediation. The predicted Sm values were confirmed on remeasurement. Neural network methods, with the advantages of being model-free and effective in solving non-linear relationship problems, appear to provide an automated and effective way for the quality control of environmental databases. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Outlier; Database; Quality control; Neural network; Principal component analysis; Multiple regression analysis; Rare earth elements

1. Introduction

The International Centre for Environmental and Nuclear Sciences (ICENS) is conducting a programme on environmental geochemistry which involves the measurement of the concentrations of very large numbers of geographically referenced samples of rocks, soils, stream sediments, air particulates, plants and animals for as many elements as possible (currently approx. 60). The analytical techniques being used are mainly atomic absorption spectrometry (AAS), neutron activation analysis (NAA) and X-ray fluorescence (XRF), and the extensive datasets are maintained in digital databases (Lalor et al., 1990). Because

* Corresponding author. Present address: Department of Geography, National University of Ireland, Galway, Ireland. Fax: +353-91-525700.
E-mail address: chaosheng.zhang@nuigalway.ie (C. Zhang).

PII: S0048-9697(01)00839-7

much of the work is semi-automated, strict quality control is necessary at every stage from sampling to final data acceptance. The use of field and analytical duplicates, analysis of standards etc., provide a measure of confidence, and various visual checks, often based on prior knowledge, on the data are routinely performed for outliers.

Outliers can be a result of mistakes and errors but often occur because of natural enrichment and anthropogenic activities. Frequently, it is the outlier groups that are of interest, however, they may distort the statistic results. Therefore, outliers should usually be identified and examined, and an effective process, especially for large datasets, would be of value and could well be of more general applicability.

The term 'outlier' can refer to an 'outlying value' of a variable or to an 'outlying sample'. When the concern is with values of a single variable, univariate statistics are applicable and the value can be detected by considering standard methods such as range (average $\pm n \times s$), histograms, box-plots, Grubb's test, *t*-test, Dixon test and Walsh test (Grubbs and Beck, 1972; Iglewicz and Hoaglin, 1993; Gibbons, 1994). When the concern is with outlying samples with more than one variable, multivariate methods are necessary (Rohlf, 1975; Hadi, 1992; Barnett and Lewis, 1994). Neural networks, which are non-linear computer models that mimic some working functions of the human brain, also have potential for outlier detection (Wong and Gedeon, 1995; Muruzabal and Munoz, 1997; Zhang et al., 1999).

For the present purposes, outliers are classified in three types: (1) range outliers, which are considered to be either too high or too low compared with the population or the majority of samples; (2) spatial outliers, usually defined as observations that are extreme with respect to their neighbouring values (Cerioli and Riani, 1999); and (3) relationship outliers, defined as observations that fall outside of the values expected from correlations within the dataset.

There are significant correlations between the concentrations of many elements in Jamaican soils (Lalor, 1995). Data values that violate such expected relationships are not always readily detectable by standard methods and may be the result of mistakes. The concentrations of rare earth elements (REEs) in soils and rocks are well known to be highly correlated and, therefore, these elements provide an excellent starting point for the relationship outlier investigation.

The purpose of this study is to assess methodologies for multivariate outlier detection, which may be widely applicable to large, high-dimensional environmental databases.

2. Methodology

The ICENS soil geochemistry database includes the results of an island wide geochemical soil survey carried out in 1988 at a sampling density of 1 per 64 km^2. It includes results on a total of 203 soil samples, including 38 site duplicate samples, which were taken several meters away from the original site. These sites to be duplicated were randomly selected, and therefore, all the 203 samples were used in this study for the purpose of outlier detection. The concentrations of 32 elements are now available from that sample set. The detection limits were used for the values that were lower than the detection limits.

2.1. Principal component analysis

Principal component analysis (PCA) is often used in data reduction to identify a small number of principal components that explain most of the variance observed in a much larger number of manifest variables and also to reveal the relationship among the variables. It has also been applied for the outlier detection (Zhang and Selinus, 1998; Zhang et al., 1998). In this study, two principal components PC1 and PC2 were defined, and the sample score plot was drawn with these as '*x*' and '*y*' co-ordinates. In the plot the outliers are located at a distance from the origin. The distance of sample scores (DSC), shown in Eq. (1), is the parameter used to assess the outliers:

$$\mathrm{DSC} = \sqrt{x_1^2 + x_2^2} \quad (1)$$

where x_1 is the sample score on PC1, x_2 is the sample score on PC2. The greater the DSC, the more likely that the sample is an outlier.

2.2. Multiple regression analysis

Multiple regression analysis (MRA) was also performed for outlier detection. In each calculation, one variable was treated as the dependent variable, and the others as independent variables. The residuals between the predicted and measured values for the dependent variable were calculated and the procedure was repeated until each had been defined as the dependent variable. The definition of outlying samples is based on the sum of absolute residuals (SAR) in Eq. (2):

$$\mathrm{SAR} = \sum_{\mathrm{i}=1}^{n} |x_{\mathrm{i}} - t_{\mathrm{i}}| \qquad (2)$$

where n is the number of variables, x_{i} is the measured value of variable i and t_{i} is the predicted value by MRA. The larger the SAR value, the more likely that the sample is an outlier.

2.3. Neural network

The most commonly used backpropagation neural network was applied in this study. A backpropagation (or back error propagation) neural network is based on fully connected, layered, feed-forward networks. It consists of one input layer, one or several hidden layers and one output layer. Each layer has one or several neurons, and the neurons of different layers are connected by weights. The errors between predicted and actual outputs were fed backwards through the network to adjust the weights on the network connections. A standard sigmoid activation function is used to control the output of a neuron to the next. The backpropagation algorithm calculates the connection weights based on initially set random weights and the differences between the output layers and the measured values are obtained. When the total error is smaller than a pre-set tolerance value, the network training stops, otherwise, the weights are adjusted by a gradient descent function and are recalculated, until the total error is smaller than the tolerance value or the epoch reaches a pre-set number. After the training, the weights are fixed and the network can be used for classification and prediction (Aleksander and Morton, 1990; Eberhart and Dobbins, 1990; Peretto, 1992).

In this study, an autoassociation backpropagation neural network (AutoNN) was used. This type of network uses exactly the input neurons (variables) as the output neurons, and aims to reproduce the trained samples (Masters, 1993; Zhang et al., 1999). All the seven available REE elements were used as both input and output variables for the AutoNN calculation. The number of neurons for the hidden layer was three and the structure of the neural network is shown in Fig. 1.

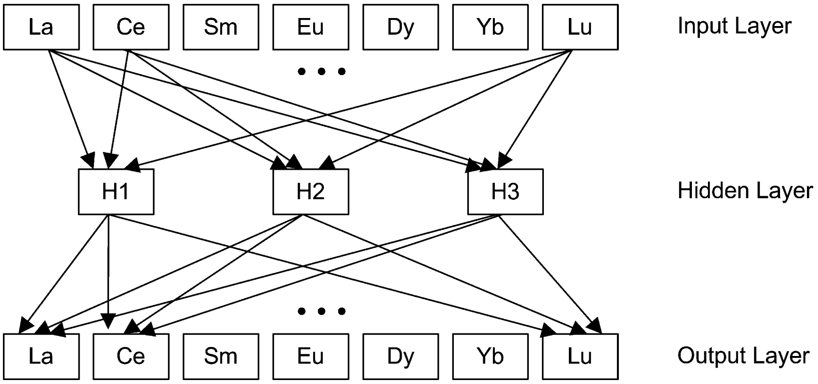

Fig. 1. Structure of the autoassociation backpropagation neural network (AutoNN) used in this study.

All the 203 samples were used as training samples, and the momentum and learning rate were 0.2 and 0.75, respectively. The neural network was trained until it reached 1000 epochs.

The similar SAR parameter in Eq. (2) with t_i now the value predicted by AutoNN was used for outlier detection.

3. Results and discussion

3.1. REEs in soils of Jamaica

The sampling sites cover soils of various lithological origins, including limestone, alluvium, shales, conglomerates, volcanic and metamorphic rocks. Jamaican soils, particularly the bauxitic soils, which overlay the white limestone, contain significant concentrations of rare earths and many other trace elements. The spatial distributions of the REEs in soils of Jamaica, for which data are available, are very similar. A typical distribution, that of lanthanum, is shown in Fig. 2.

The Central and Blue Mountains inliers, located in the central and eastern part of the island, respectively, are low in REE concentrations across a wide range of clastic sedimentary and igneous lithologies. The alluvial soils also exhibit intermediate to low REE levels. The highest concentrations are found in the bauxitic soils distributed in the middle of the island and in sedimentary soils in the southwest part.

The general information about the datasets, including skewness, kurtosis and probability of Kolmogorov–Smirnov test for normality (K–S P), are summarized in Table 1.

The concentration ranges are quite wide, and the data follow positively skewed distributions. None of the elements follows a normal distribution, because all the K–S P parameters are 0.00. As an example, the histogram of La is shown in Fig. 3.

There are clearly three peaks, which complicates outlier detection and there is a long tail with high values. Similar results are obtained for the other REEs.

The good correlation among these elements is shown in the scatter plot matrix of Fig. 4.

In each box of the matrix, the 'x' co-ordinate is the element in the row and the 'y' co-ordinate is

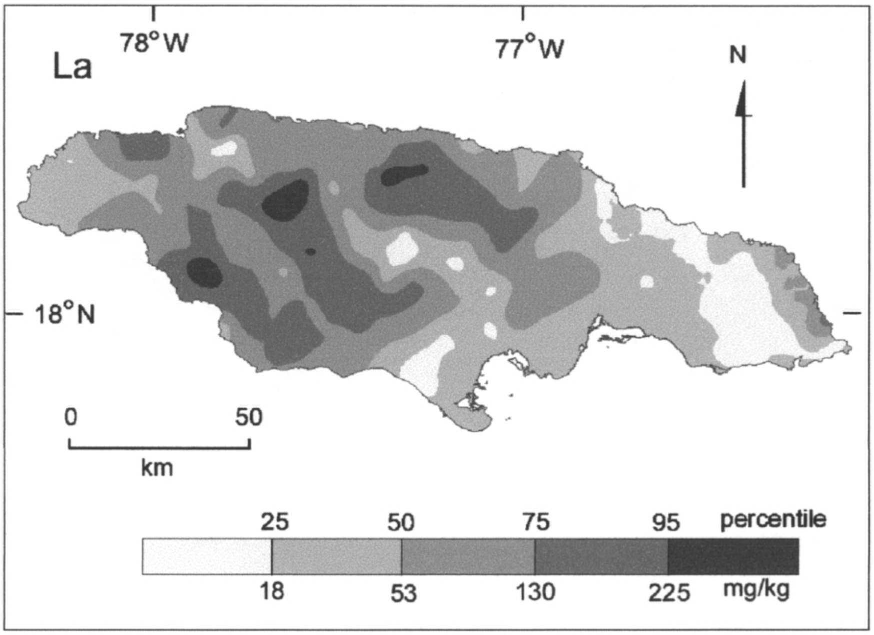

Fig. 2. Spatial distribution map of La concentration in soils of Jamaica (revised after Lalor, 1995).

Table 1
Statistical parameters of the datasets of REE concentrations in soils of Jamaica

	La	Ce	Sm	Eu	Dy	Yb	Lu
Average	84.0	112.9	12.7	3.23	12.8	7.17	1.35
S.D.	74.8	89.8	11.2	2.49	11.4	5.74	1.04
Min.	4.30	< 15.0	< 0.12	< 0.32	1.03	< 1.21	< 0.27
5%	8.06	18.0	1.37	0.60	1.79	1.61	0.36
25%	18.3	39.8	3.47	1.18	3.39	2.63	0.50
Median	53.0	79.1	8.64	2.47	8.94	4.98	0.97
75%	130.1	182.7	19.9	4.94	19.9	10.4	1.97
95%	224.9	281.3	33.8	7.53	34.4	18.6	3.44
Max.	301.7	449.7	48.8	11.6	50.7	28.5	4.82
Skewness	0.82	0.95	0.94	1.01	1.15	1.27	1.13
Kurtosis	−0.36	0.26	0.11	0.47	0.68	1.17	0.63
K-S P	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Concentration, mg/kg, $n = 203$.

the element in the column. The Pearson correlation coefficients between the REEs were positively significant at the level of 0.01. However, the correlation is not quite ‘clean’. Two samples, which are located at the upper-left corners of the scatter plot boxes in the second row and the lower-right corners of the boxes of the second column, respectively, have high Ce concentrations while those of the other elements are relatively low. Similarly, four samples show abnormally low Sm concentrations and there are similar problems with other elements. The simple scatter plot matrix provides a primary visual detection of outliers.

Because of the wide range, the concentrations of REEs were normalised to the range of 0.01–0.99 for further examination using Eq. (3):

$$x_i^* = \frac{x_i - \min(x)}{\max(x) - \min(x)} \times 0.98 + 0.01 \tag{3}$$

where x_i is the value of the i-th sample. The normalised datasets were used for both the multivariate analysis and neural network calculations.

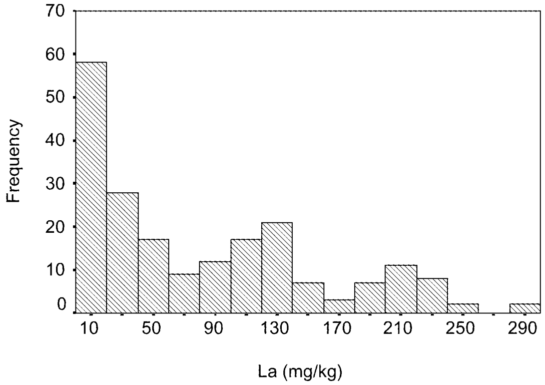

Fig. 3. Histogram of La of the raw dataset.

3.2. Outliers detected by PCA

The first two principal components (PCs) accounted for 93.3% of the total variances in PCA, and the sample scores for all the 203 samples on the first two PCs are shown in Fig. 5. Obvious outliers, A112, A112D, A177D, A078, etc., are located far away from the co-ordinate origin.

The top 20 (or 10% of the total sample number) samples based on the DSC parameter were chosen as most likely outlying samples. These are listed in Table 2. The median values are based on all 203 samples.

Compared with the medians, these 20 samples have ‘high values’ of REEs, except that the Sm concentrations for samples A112, A113, A328 and A328D are abnormally low. They have high concentrations of all the other six REEs, which is why they are detected by the PCA method.

3.3. Outliers detected by MRA

The SAR parameters for MRA are plotted in

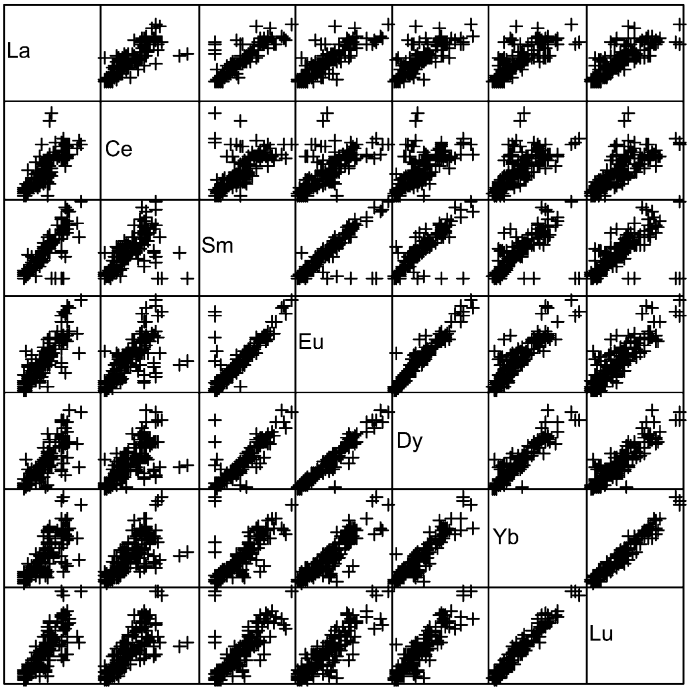

Fig. 4. Scatter plot matrix for REEs in Jamaican soils.

Fig. 6, and more details on the 20 most likely outlying samples are listed in Table 3.

The four samples with low concentrations of Sm (A328, A328D, A113 and A112) previously noted in Table 2 are top ranked as outliers by the MRA method. The highest ranked outlier, sample A044 has a very low concentration of Dy and a comparatively low concentration of Ce. Sample A112D is ranked as an outlier because of the high concentration of Ce compared with the other

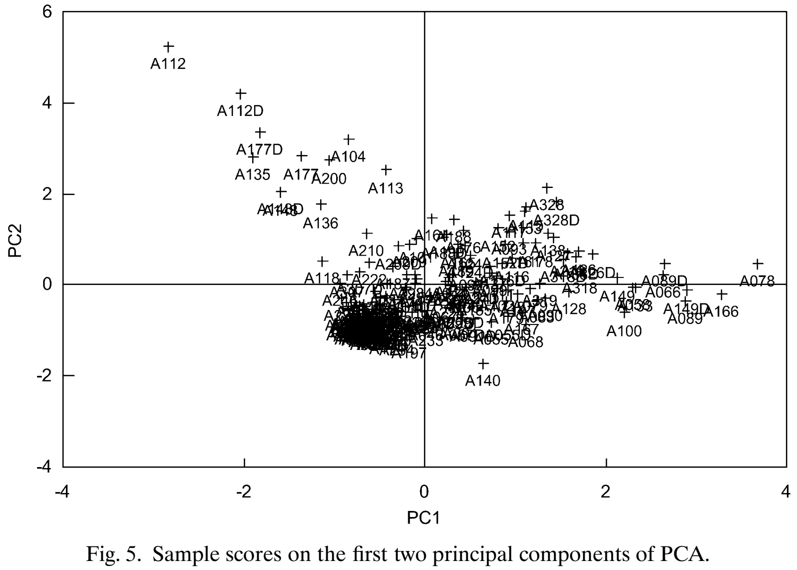

Fig. 5. Sample scores on the first two principal components of PCA.

Table 2
Top 20 possible outlying samples with normalised REE concentrations based on the index of distance of sample scores (DSC) from principal component analysis (PCA)

No.	La	Ce	Sm	Eu	Dy	Yb	Lu	DSC
(Median)	0.170	0.155	0.182	0.197	0.166	0.145	0.161	0.954
A112	0.481	0.990	< 0.011	0.339	0.308	0.369	0.415	5.984
A112D	0.444	0.898	0.357	0.297	0.273	0.321	0.378	4.686
A177D	0.723	0.624	0.286	0.206	0.178	0.222	0.290	3.835
A078	0.990	0.616	0.990	0.990	0.972	0.891	0.990	3.695
A135	0.325	0.611	0.217	0.159	0.153	0.171	0.297	3.406
A104	0.962	0.661	0.559	0.430	0.275	0.261	0.419	3.319
A166	0.803	0.509	0.871	0.917	0.990	0.635	0.691	3.282
A177	0.646	0.600	0.345	0.276	0.217	0.254	0.269	3.150
A200	0.661	0.602	0.366	0.319	0.247	0.309	0.385	2.942
A149D	0.643	0.499	0.804	0.761	0.847	0.746	0.710	2.891
A089	0.723	0.451	0.914	0.836	0.773	0.553	0.572	2.888
A089D	0.744	0.612	0.906	0.921	0.851	0.604	0.624	2.689
A066	0.770	0.477	0.759	0.766	0.712	0.746	0.777	2.633
A148D	0.304	0.470	0.177	0.099	0.127	0.174	0.301	2.607
A148	0.294	0.481	0.179	0.125	0.143	0.171	0.241	2.597
A113	0.560	0.648	< 0.012	0.599	0.588	0.472	0.499	2.572
A328	0.694	0.693	< 0.012	0.859	0.876	0.990	0.943	2.531
A328D	0.661	0.637	< 0.012	0.826	0.860	0.949	0.990	2.340
A133	0.651	0.429	0.770	0.638	0.672	0.624	0.637	2.323
A100	0.418	0.342	0.541	0.571	0.698	0.602	0.611	2.288

Table 3
Top 20 possible outlying samples with normalised REE concentrations based on the index of sum of absolute residuals (SAR) from multiple regression analysis (MRA)

Sample no.	La	Ce	Sm	Eu	Dy	Yb	Lu	SAR
(Median)	0.170	0.155	0.182	0.197	0.166	0.145	0.161	0.195
A044	0.381	0.260	0.417	0.430	0.029	0.326	0.400	1.126
A328	0.694	0.693	< 0.012	0.859	0.876	0.990	0.943	1.117
A328D	0.660	0.637	< 0.012	0.826	0.860	0.949	0.990	1.082
A112D	0.444	0.898	0.357	0.297	0.273	0.321	0.378	0.953
A147	0.071	0.308	0.320	0.270	0.243	0.221	0.184	0.857
A113	0.560	0.648	< 0.012	0.599	0.588	0.472	0.499	0.808
A112	0.481	0.990	< 0.011	0.339	0.308	0.369	0.415	0.795
A046	0.406	< 0.019	0.339	0.065	0.057	0.139	0.161	0.712
A140	0.058	< 0.016	0.052	0.432	0.384	< 0.024	< 0.027	0.695
A104	0.962	0.661	0.558	0.430	0.275	0.261	0.419	0.662
A058	0.738	0.363	0.707	0.574	0.569	0.710	0.736	0.622
A135	0.325	0.611	0.217	0.159	0.152	0.171	0.296	0.619
A117	0.762	0.599	0.647	0.605	0.599	0.534	0.419	0.589
A101	0.385	0.432	0.409	0.293	0.335	0.341	0.296	0.584
A089D	0.744	0.611	0.906	0.921	0.851	0.604	0.624	0.573
A210	0.230	0.410	0.266	0.236	0.212	0.230	0.372	0.566
A177D	0.723	0.624	0.286	0.206	0.178	0.222	0.290	0.550
A149D	0.643	0.499	0.804	0.761	0.847	0.746	0.710	0.539
A318D	0.472	0.500	0.574	0.570	0.587	0.644	0.641	0.520
A091D	0.170	0.175	0.307	0.402	0.241	0.212	0.197	0.503

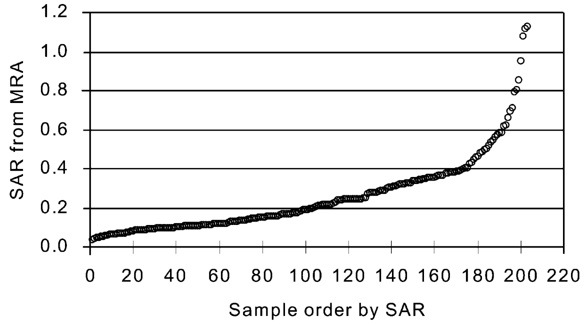

Fig. 6. Sums of absolute residuals (SAR) from multiple regression analysis (MRA).

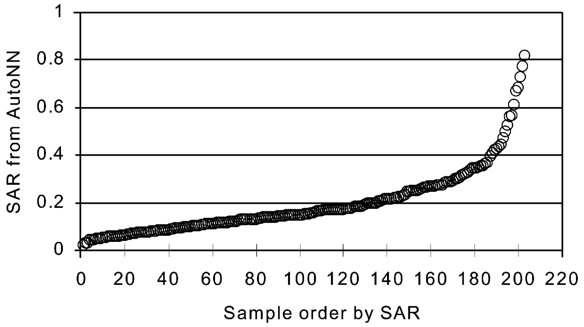

Fig. 7. Sum of absolute residuals (SAR) from autoassociation neural network (AutoNN).

REEs. The concentration of La for sample A147 appears to be too low. Most of the samples in Table 3 have shown 'relationship problems' and MRA appears to be effective in detecting such outliers in this series.

3.4. Outliers detected by AutoNN

Fig. 7 shows the results of the SAR parameter from AutoNN.

Normalized concentration data for the 20 samples with the highest SAR values are listed in Table 4. The top ranked outlier samples A140, has relatively low concentrations of La, Ce, Sm, Yb and Lu and a relatively high concentration of Eu and Dy. The third ranked outlier, sample A046, has rather high values of La and Sm, and fairly low values of Ce, Eu and Dy. The Yb and Lu concentrations for this sample are moderate. These two samples are typical 'relationship outliers' since although their concentrations are nei-

Table 4
Top 20 possible outlying samples with normalised REE concentrations based on the index of sum of absolute residuals (SAR) from the autoassociation neural network (AutoNN)

No.	La	Ce	Sm	Eu	Dy	Yb	Lu	SAR
(Median)	0.170	0.155	0.182	0.197	0.166	0.145	0.161	0.152
A140	0.058	< 0.016	0.052	0.432	0.384	< 0.024	< 0.027	0.821
A113	0.560	0.648	< 0.012	0.599	0.588	0.472	0.499	0.773
A046	0.406	< 0.019	0.339	0.065	0.057	0.139	0.161	0.728
A044	0.381	0.260	0.417	0.430	0.029	0.326	0.400	0.686
A058	0.738	0.363	0.707	0.574	0.569	0.710	0.736	0.670
A078	0.990	0.616	0.990	0.990	0.972	0.891	0.990	0.612
A112	0.481	0.990	< 0.011	0.339	0.308	0.369	0.415	0.566
A328	0.694	0.693	< 0.012	0.859	0.876	0.990	0.943	0.562
A328D	0.661	0.637	< 0.012	0.826	0.860	0.949	0.990	0.527
A112D	0.444	0.898	0.357	0.297	0.273	0.321	0.378	0.498
A089D	0.744	0.612	0.906	0.921	0.851	0.604	0.624	0.473
A147	0.071	0.308	0.320	0.270	0.243	0.221	0.185	0.445
A104	0.962	0.661	0.559	0.430	0.275	0.261	0.419	0.439
A059D	0.406	0.170	0.388	0.363	0.275	0.074	0.045	0.430
A233	0.078	0.117	0.104	0.290	0.260	0.099	0.077	0.422
A126	0.691	0.486	0.683	0.553	0.601	0.599	0.755	0.411
A166	0.803	0.509	0.871	0.917	0.990	0.635	0.691	0.395
A059	0.394	0.173	0.373	0.376	0.263	0.088	0.107	0.367
A068	0.440	0.156	0.421	0.417	0.440	0.217	0.230	0.364
A225D	0.389	0.184	0.382	0.374	0.300	0.118	0.103	0.354

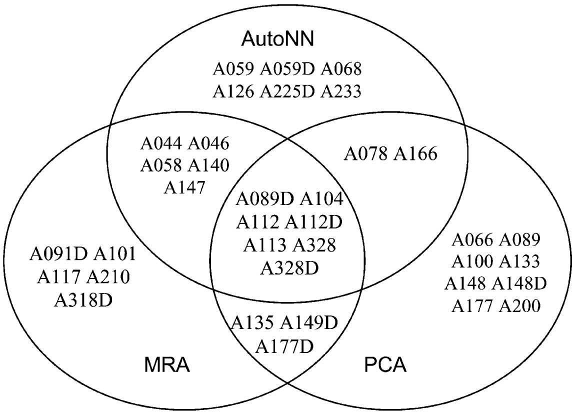

Fig. 8. Comparison of the outlying samples detected by PCA, MRA and AutoNN.

ther unusually high nor low, in some aspects they have violated the expected correlation. Samples A113, A328D, A328 and A112 that have extremely low concentrations of Sm with rather high concentrations of the other REEs are also detected by AutoNN. Sample A058 has quite high concentrations of all the REEs, except for the Ce concentration, which is relatively low. The 'relationship' problem of the 'high value outliers', samples A078, A089D, A166, etc. are also detected by the AutoNN method.

3.5. Comparison of the methods

The sets presented in Tables 2–4 are not identical and Fig. 8 compares the top 20 outliers resulting from the application of each of the three methods.

Seven samples are common to all the methods. These samples have the features of both 'relationship' and 'high value' problems. The two outlying samples detected by AutoNN and PCA and the three outlying samples detected by PCA and MRA also have both problems. The eight outliers detected only by PCA have the 'high value' problem only, and the six samples detected by only AutoNN have the 'relationship' problem. However, the 'relationship' feature of the outliers of A101, A117, A210 and A318D detected by only MRA appear rather normal (Table 3).

MRA has many statistical requirements, such as normality of the datasets and non-collinearity among the independent variables, which can hardly be satisfied in practical applications of geochemistry and, therefore, may not be fully satisfactory in this context. However, because each method has its own advantages, it is reasonable that more than one method should be applied to detect the various types of outliers.

3.6. Outlier remediation by neural network

In certain instances a reasonable way may be found to discard or replace the outliers for general statistical analysis, to reduce the negative effect of the outliers on conclusions for the majority of the population.

The prediction function of a neural network can be used to replace outliers. For the prediction, the structure of the neural network described above was revised, the variables that are to be predicted were defined as the output neurons and the other variables were the input neurons. All of the possible outlying samples were removed from the training samples, leaving only the 'good' samples to train the network.

The original Sm values for samples A112, A113, A328 and A328D were quite beyond expectation, all the other six REEs had rather high concentrations, while the Sm values were below the detec-

Table 5
Original, predicted and remeasured Sm concentrations (Sm, Sm* and Sm**, respectively) for samples A112, A113, A328, A328D, together with the other raw REEs (normalised data)

Sample no.	La	Ce	Sm	Eu	Dy	Yb	Lu	Sm*	Sm**
A112	0.481	0.990	< 0.011	0.339	0.308	0.369	0.415	0.405	0.396
A113	0.560	0.648	< 0.012	0.599	0.588	0.472	0.499	0.648	0.630
A328	0.694	0.693	< 0.012	0.859	0.876	0.990	0.943	0.796	0.939
A328D	0.660	0.637	< 0.012	0.826	0.860	0.949	0.990	0.782	0.795

tion limits. Therefore, these four samples were selected as an example to assess the methodologies used in this study. Because sample A112 also has the 'high value' problem of Ce, the other five REEs are used as input neurons and the output layer contains only Ce and Sm. The hidden layer was set to be three. The top 20 possible outliers detected by AutoNN were rejected from the training samples. The predicted Sm values (Sm*) are shown in Table 5.

The Sm values of the four samples were carefully remeasured by INAA (Table 5), and the remeasured Sm values (Sm**) are quite close to the predicted Sm values (Sm*). The four original Sm values in the database are, therefore, obviously in error, which has now been explained as a quality control mistake.

3.7. Applications of the methodologies

Because of the widely recognised good correlation among REEs in nature, it is reasonable to use multivariate methodologies to this series of elements. Generally, conventional multivariate analyses require the normality assumption of the variables and that the model should be well defined, a strong correlation among the independent variables (collinearity) may make the coefficients unstable and unreliable, finally providing misleading results. Neural network methods do not have these limitations. Problems of non-linear phenomena and violation of statistical requirements are receiving more attention and the neural network methods appear to have wide applications in quality control in environmental sciences.

Despite the fact that the neural network method is good at dealing with non-linear phenomena which may be a handicap for conventional multivariate analyses, if there is no relationship among the set of variables, all the multivariate analysis methods may be ineffective. Fortunately, this phenomenon is quite rare in environmental sciences and the methodologies used in this study, especially the neural network method, should be applicable not only to the highly correlated REE variables presented here, but also to other elements in large dimensional geochemical databases and even to wider environmental datasets. This is presently under examination.

4. Conclusion

Various types of methodologies were applied to detect outliers, which were classified into three groups: (1) range outlier; (2) spatial outlier; and (3) relationship outliers.

The good correlation among concentrations of rare earth elements in soils has been used as an example to verify the multivariate outlier detection methodologies. The sample score plot from principal component analysis is effective in detecting the high value range outliers, multiple regression analysis and an autoassociation backpropogation neural network are effective in detecting relationship outliers. As confirmed by repeat analyses the neural network methods are effective in predicting the 'expected values' of outliers.

Each method has its advantages, but since neural network methods are effective in solving nonlinear relationship problems and are model-free, they appear to be powerful automated methods for quality control in geochemistry and environmental sciences.

Acknowledgements

This early development of the geochemical database has been due to the work of numerous colleagues. Special acknowledgement is made to Dr Hilary Robotham, Dr Nedzad Limic, Mr John Preston, and Dr Mitko Vutchkov. The authors are grateful to the reviewer for the thoughtful and helpful comments.

References

Aleksander I, Morton H. An introduction to neural computing. London: Chapman and Hall, 1990.

Barnett V, Lewis T. Outliers in statistical data. 3rd edition New York: John Wiley, 1994.

Cerioli A, Riani M. The ordering of spatial data and the detection of multiple outliers. J Comput Graphical Stat 1999;8(2):239–258.

Eberhart RC, Dobbins RW. Neural network PC tools. London: Academic Press, 1990.

Gibbons RD. Statistical methods for groundwater monitoring. New York: John Wiley and Sons, 1994.

Grubbs FE, Beck G. Extension of sample sizes and percentage points for significance tests of outlying observations. Technometrics 1972;14:847–854.

Hadi AS. Identifying multiple outliers in multivariate data. J R Stat Soc Ser B 1992;54:761–771.

Iglewicz B, Hoaglin DC. How to detect and handle outliers. Milwaukee, WI: American Society for Quality Control, 1993.

Lalor GC, Rattray R, Robotham H, Thompson C. The Slowpoke 2 nuclear reactor at the University of the West Indies. Jam J Sci Technol 1990;1(1):65–77.

Lalor GC. A geochemical atlas of Jamaica. Kingston, Jamaica: University of the West Indies Press, 1995:82.

Masters T. Practical neural network recipes in C^{2+}. Academic Press: San Diego, USA, 1993:493.

Muruzabal J, Munoz A. On the visualisation of outliers via self-organising maps. J Comput Graphical Stat 1997;6(4):355–382.

Peretto P. An introduction to the modelling of neural networks. Cambridge: Cambridge University Press, 1992.

Rohlf FJ. Generalization of the gap test for the detection of multivariate outliers. Biometrics 1975;31:93–101.

Wong PM, Gedeon TD. A new method to detect and remove the outliers in noisy data using neural networks: error sign testing. Int J Sys Res Info Sci 1995;7:55–65.

Zhang CS, Selinus O, Schedin J. Statistical analysis for heavy metal contents in till and root samples in an area of southeastern Sweden. Sci Total Environ 1998;212:217–232.

Zhang CS, Selinus O. Statistics and GIS in environmental geochemistry — some problems and solutions. J Geochem Exploration 1998;64:339–354.

Zhang CS, Wong PM, Selinus O. A comparison of outlier detection methods: exemplified with an environmental geochemical dataset. The 6th International Conference on Neural Information Processing, Perth, Australia, November 16–20, 1999, vol.1, pp. 183–187.