Stochastic Environmental Research and Risk Assessment 13 (1999) 260–287 © Springer-Verlag 1999

Statistical and spatial assessment of soil heavy metal contamination in areas of poorly recorded, complex sources of pollution

Part 1: factor analysis for contamination assessment

A. Korre

Abstract The assessment of soil heavy metal contamination and the quantification of its sources and spatial extent represent a serious challenge to the environmental scientists and engineers. To date, statistical and spatial analysis tools have been used successfully to assess the amount and spatial distribution of soil contamination. However, these techniques require vast amounts of samples and a good historical record of the study area. Furthermore, they cannot be applied in cases of complex or poorly recorded contamination and provide only a qualitative assessment of the pollution sources. The author has developed a methodology that combines statistical and geostatistical analysis tools with geographic information systems for the quantitative and spatial assessment of contamination sources.

This paper focuses on the techniques that may be employed to explore the structure of a soil data set. Soil contamination data from Lavrio old mine site in Greece were used to illustrate the methodology. Through the research, it was found that principal component and factor analysis tools delineate the principal processes that drive pollution distribution. However, the spatial assessment and quantification of multiple pollution sources cannot be resolved. This aspect is explored in detail in the second paper of the series, focusing on the exploitation of principal component and factor analysis results as inputs for canonical correlation, geostatistical analysis and geographic information systems tools.

Key words soil contamination assessment, pollution sources, multivariate statistical analysis, geographic information systems

Abbreviations FA, factor analysis; GIS, geographic information systems; PCA, principal component analysis

1 Introduction

Physical degradation of the soil is one of the main environmental issues as land contamination by heavy metals or organic products is increasing not only at

A. Korre
T H Huxley School of Environment, Earth Sciences and Engineering, Imperial College of Science Technology and Medicine, London, SW7 2BP UK
Tel.: ++44 171 5947 372
Fax: ++44 171 5947 444
e-mail: a.korre@ic.ac.uk

urban sites but also in some rural areas. The importance of natural and man-made inputs of chemical elements in the surface environment in relation to the health of plants, animals and man is known and accepted. Researchers in the late 1980s and into the 1990s have been trying to address the fluxes of toxic elements in soil, water, dust and air together with their pathways to the target organism (Thornton, 1993).

Toxic elements rarely occur alone and their associations and interactions with one another and with other components of the environment are known to influence their availability to organisms and their ultimate toxicity.

In recent years, advances in computer hardware and software intensified the application of quantitative analysis techniques that would otherwise be extremely tiresome and time consuming. Research into the analysis of soil chemistry information has utilised statistical analysis tools to assist the selection of optimum soil sampling patterns, to estimate measurement uncertainty due to sampling and chemical analysis and to assist in assessing the risk of soil contamination (Ferguson, 1992; Goovaerts et al., 1997). Soil contamination research in the past has, in most cases, been quantitative in terms of levels of pollutants sampled and measured, however, the analysis of contamination has been limited to a descriptive evaluation of the sources of soil pollution (Davies and Ballinger, 1990; Li and Thornton, 1993). Therefore, the fundamental research question to be addressed in this field is the quantification of the extent of pollution, not just in levels of pollutants but also in terms of their geographic spread; the relationship between the pollutants that very often coexist; and the determination of the sources of pollutants.

One issue of major importance in soil contamination studies is to distinguish the natural background from anthropogenic anomalies. So far, this has been done by using different sample types such as different horizons of the soil profile or by comparing, statistically, rock and stream sediments, and soil analysis results (Selinus and Esbensen, 1995). However, these techniques are expensive and time consuming processes requiring vast amounts of data both from soil and baseline studies. Furthermore, these methods do not clearly associate specific sources with pollutants, and the estimation method allows only a qualitative interpretation of the results that cannot be applied in cases were baseline data is not available.

Experience in contaminated land assessment and remediation has shown that both the industry and the regulatory authorities would require a comprehensive methodology for the quantitative assessment of soil quality and its spatial extent. The author has developed a methodology which combines contemporary statistics, multivariate statistics and geostatistics with modern spatial data analysis techniques and geographic information systems (GIS) to meet these requirements, even for sites with complex background and poorly recorded history (Korre, 1997).

The overall methodology is presented in two papers, which illustrate the principles and the techniques forming the methodology. This first paper of the series highlights the differences, advantages and affiliation between principal component analysis (PCA) and factor analysis (FA) tools for the assessment of the principle processes driving soil contamination. The second paper focuses on canonical correlation statistical analysis, geostatistical analysis and geographic information systems (GIS) tools for the assessment of the sources of soil contamination.

The methodology developed is illustrated using soil chemical analysis data from Lavrio old mine site, one of the oldest examples of mining activity in

Europe. The mines are situated at the south-eastern part of the Attiki peninsula, about 60 km from Athens in Greece. The silver bearing structures around Lavrio were known and exploited for more than 25 centuries. As a result, the area around the ancient mines and processing plants is polluted with high heavy metal loads. The methodology developed enabled the author to distinguish the naturally occurring high heavy metal loads from the human induced pollution in the area.

2
Methodology for the assessment of soil contamination

The information required for all soil contamination assessment studies falls in one of two categories. Quantitative information such as metal concentrations, organic and inorganic content, pH, conductivity and the x-y co-ordinates are required for the assessment of both the extent and the spread of pollutants. However, in order to assess the nature and risk of soil contamination, additional qualitative information such as soil type, geological background, land use and many more site specific parameters are also essential. In order to integrate the two types of data, techniques such as correspondence analysis and indicator kriging have been used to analyse the qualitative information that has been coded into discrete variables (Goovaerts and Journel, 1995; Goovaerts et al., 1997). It is important to mention here that the methodology presented in this paper utilises statistical and geostatistical analysis tools for the quantitative data analysis. The qualitative information is integrated with the quantitative solution at the interpretation stage making use of the spatial referencing and transformation capabilities of GIS analysis tools.

The information recorded in any soil data set is the composite result of the parent material for the soil, the dominant processes that guide the redistribution of elements and substances as well as the pollution sources that have been active in the study area. In order to appreciate the utility of different statistical analysis tools in soil pollution assessment, one may consider each of these provisions as a separate information tier.

The simple analysis tools provide the means to clean and screen up the measured substances improving at the same time the statistical properties of the data. In addition, when combined with geostatistics and GIS, they also allow some initial evaluation of the range and extent of pollution. The PCA and FA tools examine the structure of the soil data in the next tier reducing the number of original variables to a small set of latent variables (components/factors). This tier of information reflects the principal processes that play a role in the study area and is enhanced by the synergy of geostatistical and GIS analysis. However, with these tools, it is not possible to quantify or even distinguish multiple sources of pollution that often coexist.

In some occasions, the dominant processes may relate to the sources of soil pollution since, in nature, pollutants that originate from the same source illustrate greater affinity. However, this is not always the case, especially when human activities have caused redistribution of the pollutants concerned. To illuminate this next tier of information held in the soil data set, another method of 'structural' analysis, canonical correlation is used.

One of the most important characteristics of any soil pollution survey is the spatial dimension of the data. Estimation of the measured and latent variables at unsampled locations is achieved via variogram modelling and ordinary kriging of the variables. Subsequently, the geographic database held in GIS serves as the host environment for additional spatial operations, such as vectorising the raster grids

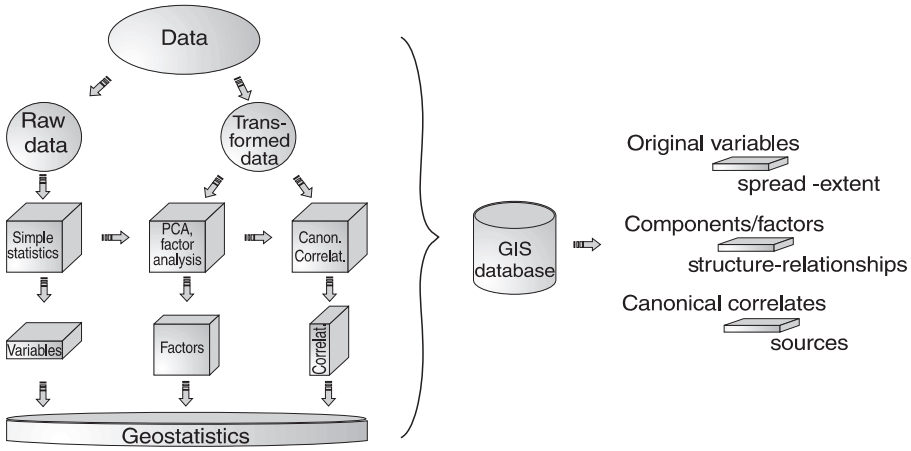

Fig. 1. Soil contamination assessment methodology developed

produced by ordinary kriging; spatial referencing between quantitative and qualitative information and graphical presentation of the results.

The overall methodology for the assessment of soil contamination and its sources is summarised in Fig. 1.

3
Principal component and factor analysis

Principal components analysis and factor analysis are statistical techniques used to investigate the structure of a data set, in an effort to identify the procedures controlling the scores of the variables in the data. The technique of PCA was first introduced by Karl Pearson (1901), however, a description of practical computation methods were established much later by Hotelling (1933). The early development of factor analysis was due to Charles Spearman (1904), who noted that most of the observed correlations could be accounted for by a simple model for the scores.

Both principal components analysis and factor analysis are statistical techniques applied to a single set of variables where the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another. The specific aims of PCA and FA for a soil data set are to summarise patterns of correlations among observed variables, reduce a large number of observed variables to a smaller number of factors, to provide an operational definition (a regression equation) for an underlying process by using observed variables, or to test a theory about the nature of underlying processes.

Mathematically, the two methods produce several linear combinations of observed variables, each linear combination being a component or factor. The factors summarise the patterns of the correlations in the observed correlation matrix and can in fact be used to reproduce the observed correlation matrix. However, since the number of factors is usually far fewer than the number of the observed variables, there is a considerable parsimony in factor analysis. Furthermore, when scores on factors are estimated for each subject, they are often more reliable than scores on individual observed variables. This is particularly useful in soil contamination studies where the factor scores on subjects, i.e. soil samples, could be further used for the spatial assessment and interpretation of the factor solution.

Steps in PCA or FA include selecting a set of variables, preparing the correlation matrix, extracting the set of factors from the correlation matrix, determining the number of factors, rotating the factors to increase interpretability and finally, interpreting the results. Although there are relevant statistical considerations to most of these steps, the final test of the analysis is interpretability. In practise, a factor is more easily interpreted when several observed variables correlate highly with it, and those variables do not correlate with other factors.

One of the problems with these techniques is that there is no criterion beyond interpretability against which to test the solution. A second problem is that, after extraction, there are infinite numbers of rotations available, all accounting for the same amount of variance in the original data, but with factors defined slightly differently. The final choice among alternatives depends on the researcher's assessment of the solution's interpretability and scientific utility. There are two general classes of rotation, orthogonal and oblique. If the rotation is orthogonal, so that all the factors are uncorrelated, the loading matrix produced portrays the results. If the rotation is oblique, there are several additional matrices: the factor correlation matrix containing the correlations among the factors and two other matrices in place of the loading matrix. These two are the structure matrix with the correlations between factors and the variables and the pattern matrix of unique relationships between each factor and each observed variable, uncontaminated by overlap among factors. The meaning of the factors after oblique rotation is ascertained from the pattern matrix. Lastly, for both types of rotation, there is a factor-score coefficient matrix used to estimate scores on factors from scores on observed variables for each individual observation.

It is worth mentioning here that in a good factor analysis the correlations in the residual correlation matrix (the differences between the observed and the reproduced correlation matrix) are small, indicating a close fit between observed and reproduced matrices. Table 1 illustrates some of the commonly encountered matrices and their descriptions.

A variety of procedures for factor extraction and rotation that are available are described in Mulaik (1972), Harman (1967) and Rummel (1970). The following paragraphs describe the most commonly used methods and give an insight into their differences and application.

3.1
Factor extraction techniques

Amongst all extraction techniques, principal components and principal factors are the most commonly used. All techniques calculate a set of orthogonal components or factors, which in combination reproduce R. The criteria used to establish the solution, such as maximising the variance or minimising the residual correlations, differ from one technique to another. Even so, the differences in solutions are small for data sets with a large sample size, numerous variables and similar communality estimates. In fact, one test of the stability of a factor analysis solution is that the same factors appear regardless of which extraction technique is employed.

One of the most important decisions is the choice between PCA and factor analysis. Mathematically, the difference occurs in the contents of the positive diagonal in the correlation matrix. In both methods the variance that is analysed is the sum of the values in the positive diagonal. This is 1.0 for all elements in PCA, which means that all the variance is distributed over the components, including the error and unique variance for each observed variable. The first principal component is the linear combination of observed variables that maxi-

Table 1. Commonly encountered matrices in principal components and factor analysis (modified after Tabachnick and Fidell, 1989)

Label	Name	Rotation	Size[a]	Description
R	Correlation matrix	Both orthogonal and oblique	p × p	Matrix of correlations between variables
Z	Variable matrix	Both orthogonal and oblique	N × p	Matrix of standardised observed variable scores
F	Factor-score matrix	Both orthogonal and oblique	N × m	Matrix of standard scores on factors or components
A	Factor loading matrix	Orthogonal	p × m	Matrix of regression-like weights used to estimate the unique contribution of each factor to the variance in a variable. If orthogonal, also correlations between variables and factors
	Pattern matrix	Oblique		
B	Factor-score coefficients matrix	Both orthogonal and oblique	p × m	Matrix of regression weights used to generate factor scores from variables
C	Structure matrix[b]	Oblique	p × m	Matrix of correlations between variables and (correlated) factors
Φ	Factor correlation matrix	Oblique	m × m	Matrix of correlations among factors
L	Eigenvalue matrix[c]	Both orthogonal and oblique	m × m	Diagonal matrix of eigenvalues, one per factor
V	Eigenvector matrix[d]	Both orthogonal and oblique	p × m	Matrix of eigenvectors, one vector per eigen value

[a] Row by column dimensions where p = number of variables, N = Number of subjects, m = number of factors or components
[b] Also called characteristic roots or latent roots
[c] Also called characteristic vectors
[d] If the matrix is of full rank, there are actually p rather than m eigenvalues and eigenvectors. Only m are of interest, however, so the remaining p-m are not displayed.#

mally separates subjects by maximising the variance of the component scores. The second component is formed from the residual correlations and is the linear combination of observed variables that extracts maximum variability uncorrelated with the first component, and the procedure continues until all variance is accounted for.

In factor analysis, on the other hand, only the variance that each observed variable shares with other observed variables is available for analysis. Exclusion of error and unique variance from factor analysis is based on the belief that such variance only obscures the recognition of underlying processes. Shared variance is calculated by communalities, which are values between 0.0 and 1.0 that are inserted in the positive diagonal of the correlation matrix. However, because unique and error variances are omitted, a linear combination of factors approximates, but does not duplicate, the observed correlation matrix and scores on observed variables.

In other words PCA analyses variance, whereas factor analysis analyses covariance (communality). The aim of PCA is to extract maximum variance from a data set with a few orthogonal components, whereas the objective of factor

analysis is to reproduce the correlation matrix with a few orthogonal factors. This implies that, whereas PCA gives a unique mathematical solution, most forms of factor analysis are not unique. The assumptions of factor analysis imply that, in general, the common factors are not linear combinations of the observed variables. In fact, even if the data contain measurements on the entire population of observations, it is not possible to compute the scores of the observations on common factors. Nevertheless, the scores can be estimated in a variety of ways.

This problem of factor score indeterminacy has led to other extraction techniques that can be considered as approximations of common factor analysis. One of these techniques is the *image factor extraction*, called so because the analysis distributes among factors the variance of an observed variable that is reflected by the other variables (the squared multiple correlation). Image factor extraction provides a unique mathematical solution and, since the components are defined as linear combinations, they are computable. However, this advantage is offset by the fact that even if the data fit the factor analysis model perfectly, the component methods do not generally recover the correct factor solution. Therefore, when the aim of the analysis is to determine the communality between the variables, this method should not be used (Dziuban and Harris, 1973; Lee and Comrey, 1979).

Another technique is the *maximum likelihood factor extraction* which was developed originally by Lawley (1963). This extraction method estimates population values for factor loadings by calculating loadings that maximise the probability of sampling the observed correlation matrix from a population. Within constraints imposed by the correlations among variables, population estimates for factor loadings are calculated in such a way that they have the greatest probability of yielding a sample with the observed correlation matrix. This method is preferred by most statisticians (Lawley and Maxwell, 1971). The advantages of this technique include that it has desirable asymptotic properties (Bickel and Doksum, 1977) and gives better results than principal factor analysis in large samples. It is also possible to test hypotheses about the number of common factors using this method. The maximum likelihood solution is equivalent to Rao's canonical factor solution which maximises the determinant of the partial correlation matrix (Morrison, 1976). Thus, as a descriptive method, it does not require a multivariate normal distribution. The validity of Bartlett's chi-squared test for the number of factors does not require approximate normality and additional regularity conditions that are usually satisfied in practise before entering multivariate analysis (Geweke et al., 1980).

Another extraction method developed by Comrey (1962) and Harman and Jones (1966), is the *unweighted least squares minimum residual factoring* (*Minres*). The goal is to minimise squared differences between the observed and reproduced correlation matrices. Only the off-diagonal differences are considered, and communalities are derived from the solution rather than being estimated as part of the solution. The procedure gives the same results as principal factors if the communalities are the same.

The method of generalised mean squares factoring also seeks to minimise off-diagonal squared differences between observed and reproduced correlation matrices, however in this case, weights are applied to the variables. Differences for variables that have substantial shared variance with other variables are weighted more heavily than the differences for variables that have substantial unique variance. In other words, differences for variables that are not as strongly related to other variables in the set are not as important to the solution.

Finally, the *alpha factor extraction*, which grew out of psychometric research, focuses interest in discovering which common factors are found consistently when repeated samples of variables are taken from a population of variables. The problem is to identify mean differences that are found consistently among samples of subjects taken from a population of subjects. However, the concern lies with the reliability of the common factors, rather than the reliability of group differences. For this reason, in alpha factoring, the communalities are estimated using iterative procedures that maximise the coefficient alpha (a measure derived in psychometrics for reliability) for the factors.

The author believes that the maximum likelihood extraction technique is the one that suits the nature of soil contamination data best. The main reasons are that the data sets generated from a potentially polluted area are most likely to include variables which are not normally distributed. In addition, the method allows to test statistically different hypotheses about the number of factors that represent the original set best. This point will be illustrated further utilising the elemental concentrations measured for soil samples from the Lavrio old mine site.

3.2 Rotation

The results of factor extraction, unaccompanied by rotation, are likely to be hard to interpret regardless of the extraction method used. Rotation is used to improve the interpretability and scientific utility of the solution. Rotation does not improve the quality of the mathematical fit between the observed and reproduced correlation matrices and, as in the case of different extraction techniques, different rotation methods tend to give similar results if the pattern of correlation in the data set is fairly clear. Different rotation techniques described in Gorsuch (1983), Harman (1967) or Mulaik (1972) fall into two main categories: the orthogonal and the oblique methods. The most widely used examples of these techniques are summarised in Table 2.

Varimax, quartimax and equamax are all orthogonal techniques, with varimax being the most commonly used of all the rotation methods available. In varimax, the objective is to maximise the variance of the loadings within factors, across

Table 2. Summary of rotational techniques

Rotational technique	Type	Objective of the analysis
Varimax	Orthogonal	Minimise complexity of factors by maximising variance of loadings on each factor
Quartimax	Orthogonal	Minimise complexity of variables by maximising variance of loadings on each variable
Equamax	Orthogonal	Simplify both variables and factors
Orthomax	Orthogonal	Simplify either factors or variables depending on the value of gamma (Γ)
Parsimax	Orthogonal	Performs an orthomax rotation for Γ = (nvar · (nfact−1)/(nvar + nfact−2), where nvar = number of variables and nfact = number of factors
Orthoblique	Oblique	Rescale factor loadings to yield orthogonal solution; nonrescaled loadings
Promax	Oblique	Orthogonal factors rotated to oblique positions
Procrustes	Oblique	Rotate to target matrix

variables. The loadings that are high after extraction become higher and those that are low become even lower. As a result, the interpretation of the factors becomes easier. Quartimax does for variables what varimax does for factors. However, the method is not nearly as popular as varimax because researchers are usually more interested in simple factors than in simple variables. Equamax is a hybrid between varimax and quartimax that tries simultaneously to simplify the factors and the variables. Mulaik (1972) reports that equamax tends to behave erratically unless the researcher can specify the number of factors with confidence.

The oblique rotation techniques are used when the researcher suspects that the processes represented by the factors are correlated. Oblique rotations offer a continuous range of correlations between factors. Of these techniques, the orthoblique rotation uses the quartimax logarithm to produce an orthogonal solution on rescaled factor loadings; therefore, the solution may be oblique with respect to the original factor loadings.

In promax rotation, an orthogonally rotated solution (usually varimax) is rotated again to allow correlations among factors. The orthogonal loadings are raised to powers (usually 2, 4, or 6) to drive small and moderate loadings to zero while larger loadings are reduced, but not to zero. As a result, even though factors correlate, simple structure is maximised by clarifying which variables do and do not correlate with each factor.

Finally in procrustes rotation, a target matrix of loading (usually 0's and 1's) is specified by the researcher, and a transformation matrix is sought to rotate extracted factors to the target if possible. If the solution can be rotated to the target, then the hypothesised factor structure is considered confirmed. Unfortunately, as Gorush (1983) reports, with procrustean rotation factors tend to be highly correlated and sometimes a correlation matrix generated by random processes is rotated to the target with apparent ease.

Factor extraction yields a solution in which observed variables are vectors that terminate at the points indicated by the co-ordinate system. The factors serve as axes for the system, the co-ordinates of each point are the entries from the loading matrix for the variable, and the length of the vector for each variable is the communality of the variable. If the factors are orthogonal, the factor axes are all at right angles to one another, and the co-ordinates of the variable points are the correlations between the common factors and the observed variables.

One of the primary objectives of principal component and factor analysis is to discover the minimum number of factor axes needed to reliably position variables. A second major goal, and the motivation behind rotation, is to discover the meaning of the factors that underlie the responses to the observed variables. Factor rotation repositions factor axes so as to make them interpretable, thus changing the co-ordinates of the variable points, while retaining the positions of the points with respect to each other.

3.3
Limitations of principal components and factor analysis

As most applications of principal components and factor analysis are exploratory in nature, both the theoretical and the practical limitations of these procedures are not very strict. The decisions upon which the number of factors and rotational schemes are selected are based on pragmatic rather than theoretical criteria.

The design of factor analysis, however, differs from other analysis methods in several important aspects (Comrey, 1973). The first task is to generate hypotheses about factors believed to underlie the domain of interest. Statistically, it is im-

portant to include enough factors so that the solution is stable. Logically, in order to reveal the process underlying the data, all relevant factors have to be included and failure to measure an important factor may distort the apparent relationships among measured factors. Next, for each hypothesised factor, some variables that are believed to be pure measures of the factor are included as marker variables. The marker variables are highly correlated with one and only one factor and, load on it regardless of extractional or rotational technique. Complexity is indicated by the number of factors with which a variable correlates; this being one factor for marker variables and up to several factors for complex variables. If variables differing in complexity are all included in an analysis, those with similar complexity levels may correlate with each other because of their complexity and not because they relate to the same factor. These variables may become trapped in factors that have little to do with the underlying processes.

An additional requirement is that the sample chosen for the analysis exhibits a spread in scores with respect to the variables and the factors measured. If all subjects achieve approximately the same score on some factor, correlations among the observed variables are low and the factor may not emerge in the analysis. Therefore, selection of subjects expected to differ on the observed variables and underlying factors is an important design consideration.

It is not advisable that the results of several samples are pooled together as they may differ with respect to some criterion or shift in time, and therefore may obscure differences rather than illuminate them. On the other hand, if different samples do produce the same factors, pooling them is desirable because of the resulting increase in sample size.

In practical terms, because both principal component analysis and factor analysis are very sensitive to the sizes of correlations, it is critical that reliable correlations are employed. Sensitivity to outlying cases, problems with missing data and degradation of correlations between poorly distributed variables of course have negative effects on the analysis and have to be corrected. Further to these considerations, a matrix that is factorable should include several sizeable correlations. The expected size depends, to some extent, on the size of the sample, with larger samples tending to produce smaller correlations. However, if no correlation exceeds 0.30, the use of factor analysis may not be appropriate.

High bivariate correlations are not, however, a definite proof that the correlation matrix contains factors. It is helpful, instead, to examine the matrices of partial correlations where pairwise correlations are adjusted for the effects of all other variables. If there are factors/components present, high bivariate correlations become very low partial correlations.

There are several more sophisticated tests of the factorability of R, like the anti-image correlation matrix and Kaiser's measure of sampling adequacy. The anti-image correlation matrix contains the negatives of partial correlations between pairs of variables with effects of other variables removed. If R is factorable there are mostly small values among the off-diagonal elements of the matrix. Kaiser's measure of sampling adequacy is a ratio of the sum of squared correlations to the sum of squared correlations plus the sum of squared partial correlations, with the value approaching 1 if partial correlations are small. Values above 0.60 are considered as being suitable for factor analysis (Tabachnick and Fidell, 1989).

After the analysis, the variables that are unrelated to others in the set are identified. These variables are usually not correlated with the first few factors although they often correlate with factors extracted later. These factors are usually unreliable both because they account for very little variance and because factors

that are defined by just one or two variables are not stable. It is important to note that a variable with a low squared multiple correlation with all other variables and low correlations with all important factors is an outlier among the variables and is usually ignored in the analysis.

Finally, cases may be unusual with respect to their scores on the components or the factors calculated with principal components or factor analysis. The deviant scores are from cases for which the factor solution is inadequate and examination of such cases for consistency is informative if it reveals the kinds of cases for which the components/factors are not appropriate.

3.4
Estimates of communalities

The difference between principal component and factor analysis is that, in the positive diagonal of *R*, communality values are used in place of 1.0's. However, there is some dispute regarding how these communality values should be estimated.

Usually, the starting estimate of communality is the SMC (squared multiple correlation) of each variable as dependent variable with the others in the sample as independent variables. As the solution develops, communality estimates are adjusted by iterative procedures to fit the reproduced to the observed correlation matrix with the smallest number of factors. The iteration finally stops when successive communality estimates are very similar.

The final estimates of communality are also SMCs, but now between each variable as dependent variable and the factors as independent variables. The final communality values represent the proportion of variance in a variable that is predictable from the factors underlying it, and they do not change with orthogonal rotation.

Image extraction and maximum likelihood extraction work differently to factor analysis, in the sense that the SMCs are used as the communality values throughout the analysis. In maximum likelihood extraction, the number of factors, rather than the communalities, is estimated and the off-diagonal correlations are forced to produce the best fit between observed and reproduced matrices.

The importance with which the estimates of communality should be regarded depends on the number of the observed variables. According to Tabachnick and Fidell (1989) if the number of variables exceeds 20, the sample SMCs will probably provide reasonable estimates of communality. Furthermore, with 20 or more variables, the elements in the positive diagonal are few compared with the total number of elements in R, and their sizes do not influence the solution very much. If the communality value for all variables in the analysis are of approximately the same magnitude, the results of principal components and factor analysis are very similar.

In the event that the estimated communalities are equal or exceed one, referred to as the Heywood and ultra-Heywood cases respectively, there is a clear indication that the analysis and the results are not valid. Very low communalities, on the other hand, indicate that there are outlying variables in the data set.

3.5
Adequacy of extraction and rotation of factor solution

Because inclusion of more factors in a solution improves the fit between the observed and reproduced correlation matrices, adequacy of extraction is tied to number of factors. The larger the number of factors extracted the better the fit

and the greater the percentage of variance explained by the factor solution. However, the larger the number of factors extracted, the less parsimonious the solution is and, if the aim is to account for all the variance (PCA) or covariance (FA) in the data set, the number of factors should match the number of variables.

The selection of the number of factors is probably more critical than the selection of extraction and rotational techniques or communality values. There are several ways to assess adequacy of extraction and the number of factors (Gorsuch, 1983).

A first quick estimate of the number of factors is obtained from the sizes of eigenvalues after a principal components extraction and, as the variance that each standardised variable contributes to a principal component's extraction is one, components with eigenvalues less than one are not as important.

A second criterion is a scree test (Cattell, 1966) of eigenvalues plotted against factors. Factors in descending order are arranged along the abscissa with eigenvalues as the ordinate. The plot is usually negatively decreasing and it is the change of slope that indicates how many factors should be selected. Under less than optimal conditions, the test is still accurate to within one or two factors.

Another test is the residual correlation matrix, where the elements of the matrix are actually partial correlations between pairs of variables with the effects of factors removed. Several moderate residuals (0.05–0.10) or a few large residuals (>0.10) suggest the presence of another factor (Tabachnick and Fidell, 1989).

For principal components extraction and maximum likelihood extraction in confirmatory factor analysis there are significance tests for number of factors. Bartlett's test evaluates all factors together and each factor separately against the hypothesis that there are no factors. However, there are disagreements regarding the use of these tests (Gorsuch, 1983).

The choice between orthogonal and oblique rotation is made after the number of reliable factors is decided. Very often the nature of the data encourages oblique rotation rather than orthogonal. In practice, the best way to decide between orthogonal and oblique rotation is to request for oblique rotation with the desired number of factors and then examine the correlations among the factors. If correlation values exceed 0.3; then there is 10% or more overlap in the variance among the factors. If this is the case and unless there are compelling reasons for orthogonal rotation, oblique rotation is in favour. Compelling reasons include the wish to compare structure in groups, need for uncorrelated factors in further analysis, or a theoretical need for orthogonal rotation.

3.6
Interpretation of factors and factor scores

The proportion of variance accounted for by a factor is the amount of variance in the original variables that has been condensed into one factor. The proportion of covariance indicates the relative importance of the factor to the total covariance accounted for by all factors. The importance of a factor is evaluated by the proportion of variance or covariance associated with the factor after rotation.

The internal consistency of the solution, in other words the certainty with which the factor axes are fixed in the variable space, is given by the squared multiple correlations of factor scores predicted from the scores on observed variables. In a good solution the SMCs range between 0 and 1, and the larger they are the more stable are the factors. Values of 0.70 or better are considered high and mean that the observed variables account for substantial variance in the factor scores (Tabachnick and Fidell, 1989).

The aim of factor interpretation is to understand the underlying dimension that unifies the group of variables loading on it. In both orthogonal and oblique rotations, loadings are obtained from the loading matrix, A, however, the meaning of loadings is different in the two rotations.

After orthogonal rotation, the values found in the loading matrix are the correlations between variables and factors. Values in excess of 0.30 for these correlations usually being chosen for interpretation. After oblique rotation, the loadings in the pattern matrix are not correlations but are a measure of the unique relationship between the factor and the variable. Because factors correlate, the correlations between the variables and the factors (structure matrix, C) are inflated by overlap between factors. It is even possible that a variable may correlate with one factor through its correlation with another factor rather than correlate directly. Comrey (1973) suggests that loadings in excess of 0.71 (50% overlapping variance) are considered excellent, 0.63 (40% overlapping variance) very good, 0.55 (30% overlapping variance) good, 0.45 (20% overlapping variance) fair, and 0.32 (10% overlapping variance) poor.

4
Data screening and preparation

4.1
Accuracy of data and missing values

There is a whole set of issues that need to be considered before the main statistical analysis of a soil data set is undertaken. A very important first step is to examine the univariate descriptive statistics for accuracy of input. Within-range variables, and plausible means and standard deviations ensure that the input data is correct. Since most multivariate procedures analyse patterns of correlation (or covariance) among variables, it is important that they are as accurate as possible. Under some rather common research conditions, correlations are overestimated, underestimated, or simply inaccurately computed. If composite variables are used and they contain in part the same items, correlations are inflated. On the other hand, a falsely small correlation between two continuous variables is obtained if the range of responses to one of the variables is restricted in the sample.

One of the most pervasive problems in data analysis is that of missing data. If only a few data points are missing in a random pattern in a large data set, the problems are usually not serious and almost any procedure for handling them yields similar results. If, however, a large number of data are missing from a small to moderate-size data set, the problems can be very serious, and in fact, it is the pattern of missing data that is more important than its volume. The decision about how to handle missing data lies among several bad alternatives:

- deleting any cases with missing values or the variables that contain them.
- estimating the missing values and using the estimates during data analysis. There are three schemes for doing so: using prior knowledge, inserting mean values, and using regression.
- using a missing data correlation matrix, where all available pairs are used to calculate each of the correlations. In such a case though, some of the correlations are more stable than others. Furthermore, they are not comparable and can go out of range in their relative sizes. The eigenvectors calculated from them can become negative; therefore, positive eigenvalues are inflated. Obviously statistics derived under these conditions are very likely distorted and this option should be used very carefully.

- treating the missing data as available data by creating a dummy variable where cases with complete data are assigned 0 and cases with missing data are assigned 1. This is very important in cases where the failure to respond is itself a very good predictor. The mean is inserted for missing values so that all cases are analysed and the dummy variable is used as another variable (Cohen and Cohen, 1975).

In soil contamination research, prior knowledge is not often available, especially in cases of unknown potentially polluted sites. What is very important is to consider why the values are missing. If for example, some samples were simply not analysed for particular soil constituents or some variables have erratic values that have been detected through the screening process or just not recorded, then the mean value is the preferred option. If, on the other hand, the value is missing due to the concentration being lower than the detection limit for a particular constituent, then the only plausible alternative to case deletion is to use the detection limit, or a value up to 10% lower than the detection limit.

4.2 Outliers

Outliers are found in both univariate and multivariate situations among all types of variables and lead to errors with no indication as to which effect they have in a particular analysis. There are four reasons for the presence of an outlier:

- incorrect data.
- failure to specify missing value codes in computer control language so that missing value indicators are read as real data.
- the outlier is not a member of the population that was intended for sampling (and should be deleted).
- the case is from the intended population but the distribution of variance has more extreme values than a normal distribution. In this case the researcher should consider changing the value of the case so that it is no longer unduly influential.

Detection of univariate outliers can be easy with the use of graphical methods or by identifying the cases with standardised scores in excess of ±3.00 as potential outliers. Multivariate outliers are cases that have an unusual pattern of scores. The statistical procedure used to screen for multivariate outliers is the computation of the Mahalanobis distance for each case. A very conservative probability estimate for a case being an outlier, say probability of less than 0.001, is appropriate (Tabachnick and Fidell, 1989). Once multivariate outliers are identified, it is important to define the variables on which the cases are deviant as it is essential to identify the kinds of cases for which results do not generalise.

Soil pollution surveys are usually carried out in areas of suspected contamination. As a result, outlying values for the variables measured are both expected and important. The options available to reduce the influence of the outliers are: to eliminate the cases; to remove the variables involved if they are highly correlated with others or they are not critical; or to transform the variables so that the scores are not so deviant. Data collection, the sampling campaign and chemical analysis of soil samples are generally time consuming and expensive. Therefore, it is crucial to conserve, if possible, the outlying cases in the data set and, to do so, transformation of the variables is the preferred solution.

4.3
Normality, linearity and homoscedasticity

The underlying principle for most multivariate procedures and most statistical tests is the assumption of multivariate normality. Multivariate normality is the assumption that each variable and all linear combinations of the variables are normally distributed. When the assumption is met, the residuals of analysis are also normally distributed and this is one way to test for normality. The other option is to examine the distributions of the variables themselves with statistical or graphical methods. For normal distributions, skewness and kurtosis are zero and their significance can be tested against the null hypothesis of zero. It is advisable to transform the variables if non-normality is found among variables or residuals. Although there is no guarantee, upon securing univariate normality, it is more likely that the multivariate normality condition will be met.

The assumption of linearity requires that there is a straight line relationship between two variables. The reasons are that the solutions are based on the general linear model, the significance tests used are based on the assumption of linearity and only the linear relationships among variables are analysed. If there are substantial non-linear relationships among variables, they are ignored unless the variables are transformed so as to capture the non-linear relationship. The assumption of homoscedasticity is that the variability in scores for one variable is roughly the same at all values of the other variable and can be evaluated through bivariate scatterplots.

Although data transformations are recommended as a remedy for outliers and failures of normality, linearity and homoscedasticity, it is also known that they may increase difficulty of interpretation. In other words, it would be far easier to interpret soil elemental concentrations rather than their log, square root or inverse transforms.

4.4
Multicollinearity and singularity

Multicollinearity and singularity are problems that occur in a correlation matrix when variables are too highly correlated. With multicollinearity the correlation is 0.90 and above and, with singularity, variables are perfectly correlated with one of the variables being a combination of one or more of the other variables. The logical problem is that, with the exception of factor analysis, redundant variables are not needed in the same analysis as they reduce the degrees of freedom for error and they actually weaken the analysis. The statistical problem is that singularity prohibits and multicollinearity renders unstable matrix inversion.

Summarising the above considerations, Table 3 presents a checklist of general guidelines for screening data. However, it is important to stress that further to the afore mentioned screening methods, each statistical analysis technique has spe-

Table 3. Data screening procedure

A.	Inspect univariate descriptive statistics for accuracy of input,
B.	Evaluate amount and distribution of missing data and deal with the problem,
C.	Identify and deal with non-normal variables,
	• Check skewness and kurtosis – probability plots
	• Transform variables
D.	Identify and deal with univariate and multivariate outliers,
E.	Check pairwise plots for nonlinearity and heteroscedasticity,
F.	Evaluate variables for multicollinearity and singularity.

cific assumptions and limitations which have to be taken into account prior to its use.

The following paragraphs provide a more detailed description of the statistical analysis techniques and the underlying theoretical background, focusing particularly on principal component and factor analysis methods for soil contamination data.

5
Principal component analysis and factor analysis for Lavrio data

5.1
Mining and geology of Lavrio area

The ancient mines of Lavrio are situated at the south-eastern part of Attiki peninsula about 60 km from Athens in Greece. The silver bearing deposits of the area have been exploited for more than 25 centuries. There are over a thousand ancient mining shafts around Lavrio which were used to explore the subsurface. Of the mixed sulphide ores of lead, zinc and iron, the ancient miners were only interested in the rich lead and silver ore. The peak of mining and processing activity was between the 6th century BC and the Roman times. At the end of the 19th century several million tons of tailings and slag, some recovered from the beaches of Lavrio, were found to be sufficiently rich that the latter day miners re-processed them for many years.

The majority of the ores of Lavrio occur within the marbles, particularly at the contacts of the marbles with the schists. The primary ore comprises of two groups of minerals: iron-manganese ore and mixed sulphides of Zn, Fe and Pb, which frequently exist together or alternate.

The adverse health effects of metals on humans due to mining activities have been well known since the ancient times and are recorded in ancient documents by various writers such as Aristotle, Dimosthen and Strabo. During the past few decades, the environmental damage has become more evident and assessable. Epidemiological studies in the area have shown a high blood lead burden in school age children which has been associated with IQ deficiencies (Lavrion Health Centre, 1989). More recent research includes soil pollution assessment studies (Korre and Durucan, 1995a, b; Demetriades et al., 1996; Korre, 1997; Durucan and Korre, 1997) and contaminated soil remediation research (Kontopoulos et al., 1996; Skoufadis et al., 1997) carried out in the same area. The soil samples utilised in the study reported in this paper series were collected over an area of more than 120 km^2. The 425 samples collected and chemically analysed for 24 elements (Table 4) were further analysed geostatistically to estimate and map the concentrations of the metals (Korre and Durucan, 1995; Korre, 1997; Durucan and Korre, 1997).

5.2
Data screening and preparation

In the particular case of the Lavrio data set, there were three variables that had missing values for some of the observations. These were four samples for Mo, 21

Table 4. Elements identified by the chemical analysis

Li	Na	K	Be	Mg	Ca	Sr	Ba
Al	La	Ti	V	Cr	Mo	Mn	Fe
Co	Ni	Cu	Ag	Zn	Cd	Pb	P

samples for Ag and eight samples for Cd. In reality, this meant that the elemental concentrations were below the detection limit of the ICP-AESpectrometer, therefore, the missing values were not replaced with the mean of the sampling population, nor they were given zero concentrations.

In order to keep the observations, rather than delete them completely and lose the information on all other variables, it was decided to assign to them the minimum values that have been detected, keeping in mind that this introduces some asymmetry at the lower end of the sampling population. This is expected to be stronger in the case of Ag where are the most missing values.

The next step was to identify possible univariate outliers in the samples and try to explain them. For this reason the data was standardised, using the mean and the standard deviation. The values that were produced were the standardised scores of the variables and those in excess of ±3.00 were the potential outliers. Such values were extracted for each variable and plotted on bubble plots in order to uncover their locations in the study area and try to explain them in conjunction with the sampling campaign notes, the geological background, the mining data and the historical data for the area.

Subsequently, it was important to minimise the influence of outliers. However, at Lavrio, extreme values are expected and represent quite a large proportion of the sampling population. As 100 different samples were identified as potential outliers, it was decided not to delete the outlying cases and to transform the variables instead.

In order to select the correct transformation for each variable, all were subjected to elementary statistical analysis with univariate and bivariate methods. From the density diagrams it was shown that most elements are far from normally distributed in the study area, with Li, K, Al, and V in near normality when the outliers are removed. All other elements (variables) were transformed using appropriate functions. Table 5 displays the transformations used for each variable.

In order to check the homoscedasticity and linearity of the data set, the variables were subjected to bivariate analysis, calculating the Pearson-product moment correlation coefficients. The transformed variables exhibited notably stronger correlations, thus enhancing the characteristics of the sampling population. This further proved the success of the transformation procedure and safeguarded the use of the transformed data set for subsequent analysis.

The test for outliers and multivariate normality after transformation revealed that there were still quite a few observations that deviated significantly from the rest of the sampling population. Since their influence could not be minimised otherwise, it was decided to remove them at this stage. Figure 2 illustrates the mahalanobis distance – chi squared plots for the original and the final data set that was subsequently used in the multivariate analysis. This final data set consisted of 415 samples out of the original 425 with no missing values for the variables and with most of them already transformed.

Table 5. Transformation methods applied to the variables

Transformation	Elements
Logarithm (log x)	Na, Mg, Ba, Cr, Mo, Mn, Fe, Co, Cu, Ag, Zn, Cd, Pb
square root ($\sqrt{x}$)	Be, La, Ti, Co, Ni, P
inverse (1/x)	Ca

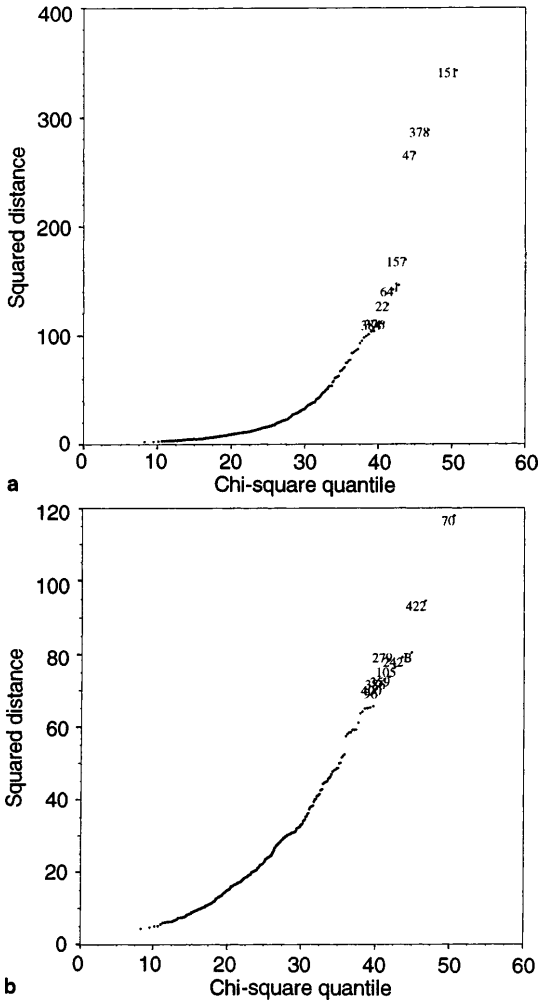

Fig. 2. Mahalanobis distance – chi squared plots for the Lavrio soil data; (a) the total sampling population after correcting for missing values and (b) the final data set used for the multivariate analysis

5.3
Principal component and factor analysis of Lavrio data

Before applying PCA and factor analysis techniques to the data, the factorability of the correlation matrix that had to be assessed. The examined correlation matrix of the data set contained numerous significant correlations. In addition, the values of partial correlation matrix were shown to be mostly low and Kaiser's measures were found to be in excess of 0.60 for the all variables except for Ca and Sr. Therefore, the correlation matrix was approved as factorable.

Principal components extraction with varimax rotation was used for the initial evaluation of the data set and to estimate the number of factors. The first 12 eigenvalues generated through the analysis are displayed graphically in the scree plot of Fig. 3.

The maximum number of components that had eigenvalues larger than 1 was five, however, since retention of so many components was considered difficult to interpret, sharp breaks in size of eigenvalues were sought using the scree test. It was found that the differences between the first five factors are large, however,

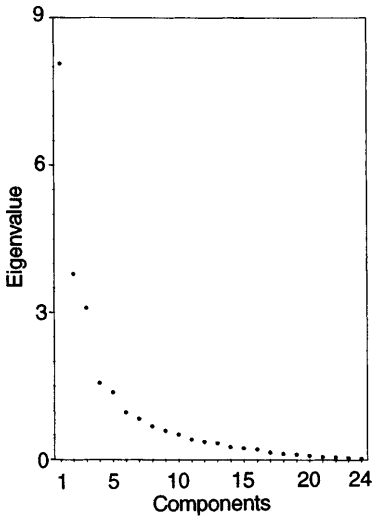

Fig. 3. Scree plot of PCA eigenvalues for the Lavrio data

Table 6. Variance explained by each component before and after varimax rotation for the Lavrio data set

Component	1	2	3	4	5
Before Rotation	8.069	3.787	3.096	1.567	1.378
After Rotation	6.359	5.378	3.216	1.696	1.476

there was little difference in variance explained by components 3, 4 and thereafter. This was taken as evidence that there were probably 3 or 4 components.

The principal components extraction model with varimax rotation was applied to the data. The results were examined for the residual correlations, the final communality estimates calculated for the 3 components, the proportion of variance explained by each component and the pattern matrix before and after rotation.

From the amount of variance explained by each component (Table 6), it was concluded that the 4th and 5th components do not add significantly to the variance explained by the first three components. On the other hand, the square off-diagonal residuals (values above 0.05) and the final communality estimates (low values) illustrated that some of the variables were not represented well by the solution. This was more evident for Ca, Sr, P and Ti. Another aspect that became apparent from the rotated component pattern plots was that the original variables in the first biplot were aligned at an angle to the two components (Fig. 4) which may indicate the presence of non-orthogonal components. These aspects were further investigated through the application of principal factor and maximum likelihood extraction techniques to the data. The following paragraphs depict the improved results that were obtained with the maximum likelihood extraction in comparison to the principal factors extraction for the Lavrio data.

For both analysis methods the communality estimates in the positive diagonal of the correlation matrix were calculated with the squared multiple correlation method (SMC). Several principal factor analysis runs were performed to find the optimum number of factors, specifying 3 to 5 factors. The trial runs with 4 and 5 factors showed that 4 factors had eigenvalues larger than one. Figure 5 illustrates

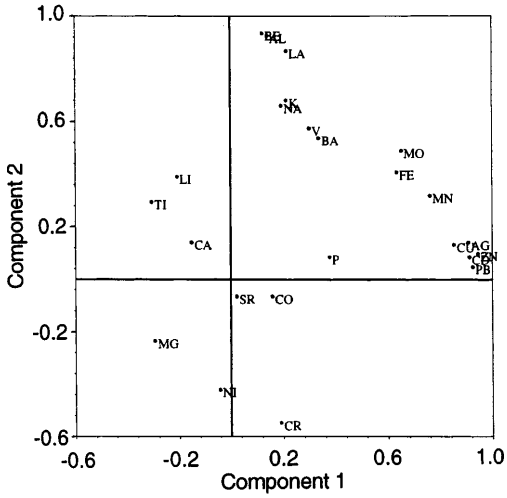

Fig. 4. Rotated component pattern plot between Component 1 and 2

the eigenvalues calculated for each factor. The 5th factor had an eigenvalue below 1.0 and no loading above 0.50, the criterion for interpretation chosen for the study. In addition, the 4th factor had eigenvalue just above unity, however, after rotation the only variable loading on the factor was Ca, which has already been identified as a possible outlier among the variables.

In the corresponding scree plot (Fig. 6) of the maximum likelihood extraction solution the first three factors showed distinctively higher eigenvalues than the subsequent factors. Therefore, only three factors were chosen for the follow-up runs of the analysis and the interpretation of the factor solution. It was notable from the final communality estimates (Table 7) that the factor extraction solution does not generalise for four of the variables: Ti, K, Ca and Mg. Therefore, these variables were not used to interpret the factors. The factors extracted were rotated using both orthogonal (varimax) and oblique methods (promax) in order to evaluate their adequacy for the data set. As shown in the scatterplots of the orthogonal factor loadings (Fig. 7) and, as further confirmed by the correlations between the three factors after oblique rotation, there is a negative correlation between FACTORs 1 and 2 ($r = -0.3448$) and negligible correlations between the other two pairs of factors.

As already mentioned, oblique rotation is adding complexities in reporting results. However, since three factors are enough to describe the data, promax rotation with varimax pre-rotation method was employed.

The results, as shown in the factor structure matrix in Table 8, strengthen the association of the factors with those variables that load significantly on them and reduce even more the values for the variables that are weakly related with the unrotated factors.

Comparing the results for the two different rotation methods proves that the variables that load strongly on the factors do not change with rotation, but those that are weakly attracted on them change positions. Another change that becomes apparent is in the relative proportion of variance explained by each factor (Table 7) which is more uniformly distributed among the factors after oblique rotation. The maximum likelihood extraction technique yielded similar results to the factor analysis solution. The three factor solution was found not to generalise for Ti, K, Mg, and Ca, and was not particularly representative for Li and Sr

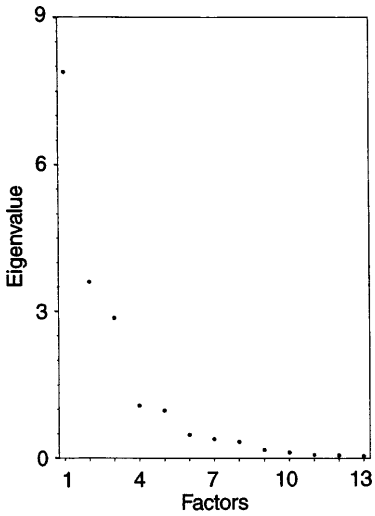

Fig. 5. Scree plot of 13 factor eigenvalues for the Lavrio data

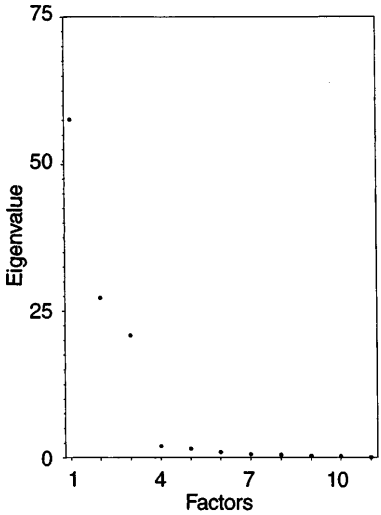

Fig. 6. Scree plot of maximum likelihood eigenvalues for the Lavrio data

Table 7. Principal factor extraction results and variance explained by each factor after varimax and promax rotation (power = 3)

Root mean square off-diagonal residuals: over-all = 0.063

Final communality estimates after varimax rotation: total = 14.366

Li	Na	K	Be	Mg	Ca	Sr	Ba
0.316	0.491	0.500	0.906	0.358	0.044	0.074	0.383
Al	La	Ti	V	Cr	Mo	Mn	Fe
0.926	0.796	0.140	0.688	0.567	0.625	0.695	0.748
Co	Ni	Cu	Ag	Zn	Cd	Pb	P
0.845	0.888	0.737	0.868	0.921	0.837	0.883	0.131

Variance explained by each factor

Rotation method	FACTOR 1	FACTOR 2	FACTOR 3
Before rotation	7.887	3.607	2.872
Varimax rotation	6.192	5.175	2.998
Promax rotation (3) eliminating other factors	5.293	4.505	2.999
Promax rotation (3) ignoring other factors	6.836	6.010	3.063

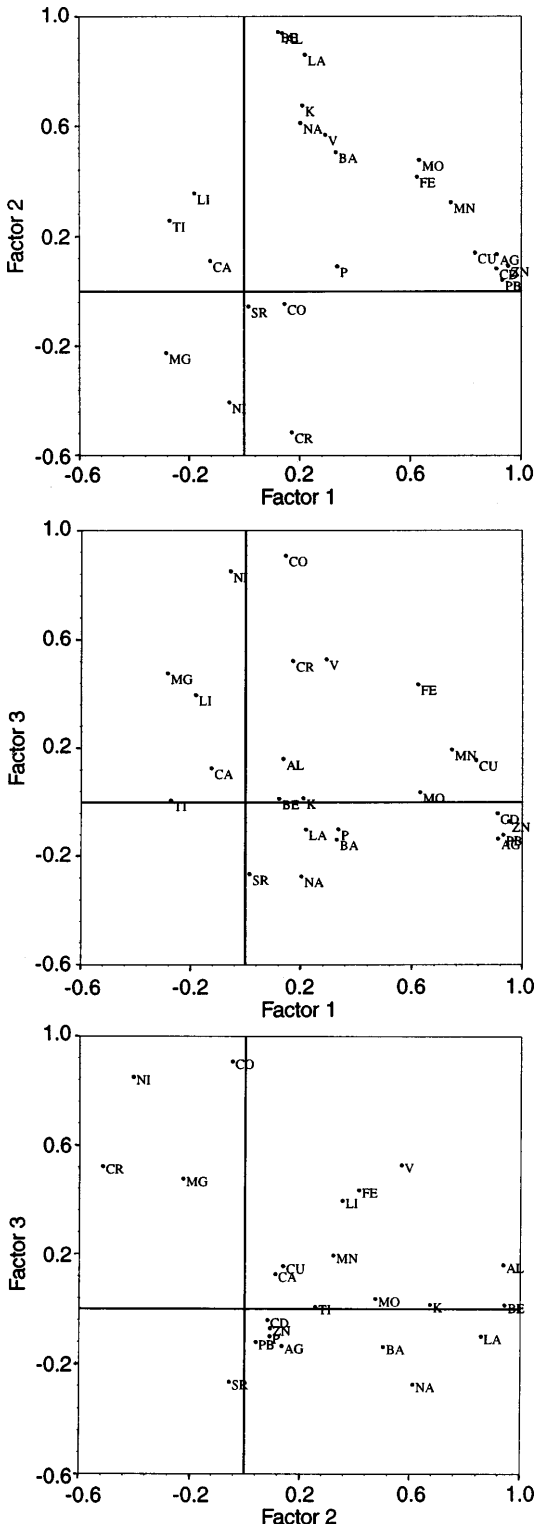

Fig. 7. Factor pattern plots after varimax rotation for the Lavrio data

Table 8. Factor pattern matrix after varimax rotation, and factor structure matrix after promax rotation (power = 3)

	Rotated factor pattern (varimax rotation)			Rotated factor structure (correlations, promax rotation)		
	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 1	FACTOR 2	FACTOR 3
Zn	0.952	0.093	−0.071	0.956	0.272	0.012
Pb	0.931	0.042	−0.121	0.924	0.216	−0.037
Ag	0.912	0.136	−0.135	0.919	0.305	−0.057
Cd	0.910	0.084	−0.041	0.911	0.255	0.038
Cu	0.832	0.141	0.156	0.847	0.900	0.226
Mn	0.744	0.323	0.195	0.789	0.462	0.250
Mo	0.630	0.477	0.037	0.698	0.589	0.076
Fe	0.622	0.415	0.435	0.688	0.533	0.474
P	0.336	0.092	−0.100	0.344	0.152	−0.073
Ti	−0.272	0.257	0.007	−0.103	0.088	0.111
Be	0.122	0.944	0.014	−0.228	0.201	−0.027
Al	0.136	0.939	0.162	0.285	0.950	0.138
La	0.219	0.859	−0.101	0.268	0.950	−0.010
K	0.210	0.675	0.015	0.349	0.883	−0.112
Na	0.202	0.612	−0.274	0.313	0.703	0.010
V	0.293	0.569	0.528	0.290	0.635	−0.277
Ba	0.330	0.505	−0.138	0.389	0.624	0.531
Co	0.145	−0.046	0.907	0.402	0.556	−0.126
Ni	−0.055	−0.405	0.850	−0.116	0.321	0.364
Cr	0.171	−0.515	0.522	0.153	−0.002	0.917
Mg	−0.284	−0.224	0.476	−0.101	−0.393	0.855
Li	−0.182	0.355	0.396	0.099	−0.464	0.554
Ca	−0.125	0.112	0.127	−0.306	−0.266	0.457
Sr	0.015	−0.055	−0.265	0.001	−0.056	−0.261

(Table 9). Furthermore, for Ca, Sr, P, Ti, Li and Mg, the weighted final communality estimates were low, extending the above list with P, whereas K was shown to recover due to a higher weight value.

After varimax rotation, the first two maximum likelihood factors extracted were found to correlate ($r = -0.3580$) slightly stronger than the corresponding factor extraction solutions. As shown in the factor reference structure plots of Fig. 8, the other pairs of variables do not correlate significantly. The resulting structure and reference structure matrix are shown in Table 10 and the relevant factor scatterplots are shown in Fig. 8, along with the reference axis correlation value and the corresponding angle values.

The simplicity of the maximum likelihood factor solution is asserted by the excellent and very good correlations between certain variables and the factors. In terms of complexity, Fe is identified as the most complex variable with fair to good loadings on all three factors.

The list of variables that correlated best with the factors after orthogonal and oblique rotation proved that the solutions were consistent. This was the case for both principal factor and maximum likelihood extraction methods.

Furthermore, comparison between the two extraction methods confirmed the adequacy of the solutions for the data set analysed since both extraction techniques gave similar results for the variables that are best correlated with the factors.

Table 9. Maximum likelihood extraction results and variance explained by each factor after varimax and promax rotation (power = 3)

Convergence criterion satisfied Significance tests based on 415 observations:
Test of H0: No common factors
vs HA At least one common factor

Chi-square = 10029.44 df = 276 Prob > chi**2 = 0.0001
Test of H0: 3 factors are sufficient
vs HA More factors are needed

Chi-square = 2133.129 df = 207 Prob > chi**2 = 0.0001
Chi-square without Bartlett's correction = 2190.447
Akaike's information criterion = 1776.447
Schwarz's Bayesian criterion = 942.594
Tucker and Lewis's reliability coefficient = 0.737
Root mean square off-diagonal residuals: Over-all = 0.064

Final communality estimates and variable weights
Total communality weighted = 105.709 Unweighted = 14.072

	Li	Na	K	Be	Mg	Ca	Sr	Ba
Communality	0.231	0.482	0.442	0.928	0.299	0.035	0.063	0.367
Weight	1.300	1.930	1.793	13.926	1.427	1.036	1.067	1.579
	Al	La	Ti	V	Cr	Mo	Mn	Fe
Communality	0.895	0.825	0.141	0.647	0.510	0.609	0.647	0.699
Weight	9.504	5.706	1.164	2.835	2.042	2.559	2.828	3.323
	Co	Ni	Cu	Ag	Zn	Cd	Pb	P
Communality	0.914	0.902	0.688	0.893	0.952	0.869	0.913	0.121
Weight	11.652	10.244	3.211	9.304	21.013	7.628	11.501	1.138

Variance explained by each factor

Rotation method	FACTOR 1	FACTOR 2	FACTOR 3
Before rotation			
Weighted	57.628	27.247	20.834
Unweighted	7.491	3.691	2.890
Varimax rotation			
Weighted	52.790	30.996	21.923
Unweighted	6.191	4.900	2.981
Promax rotation (3) eliminating other factors			
Weighted	47.160	27.966	21.913
Unweighted	5.205	4.277	2.968
Promax rotation (3) ignoring other factors			
Weighted	55.409	36.692	22.041
Unweighted	6.770	5.851	3.029

For the interpretation of the results, the extraction and rotation method selected was that of maximum likelihood with promax rotation. The reason for selecting this extraction technique is that it does not require a normal and regular multivariate population.

As proven during the screening and elementary analysis, this assumption suits the nature of the Lavrio data since the transformed variables used for the analysis are only approximately normally distributed and divergence from multivariate normality is even stronger. As for the rotation method, the presence of only three

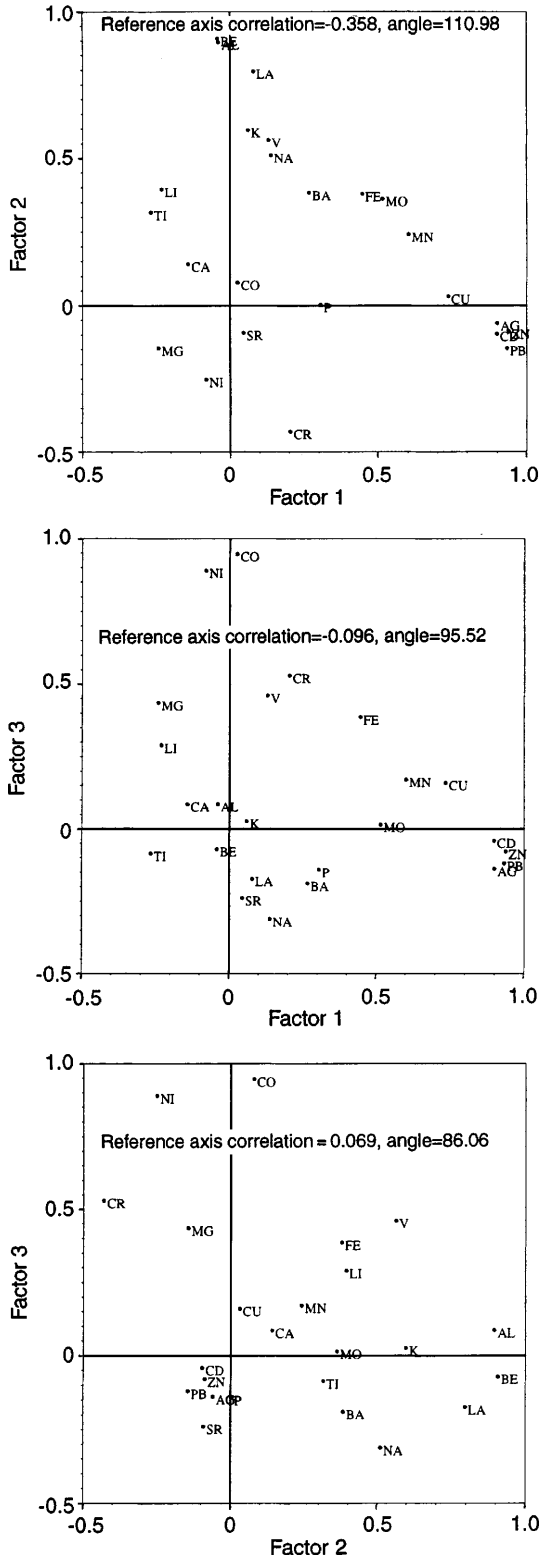

Fig. 8. Factor reference structure plots after promax rotation

Table 10. Maximum likelihood factor structure and reference structure matrices after promax rotation (power = 3)

	Rotated factor structure (correlations, promax rotation)			Reference structure (semipartial correlation, promax rotation)		
	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 1	FACTOR 2	FACTOR 3
Zn	0.969	0.265	0.002	0.939	−0.089	−0.079
Pb	0.938	0.204	−0.037	0.933	−0.145	−0.119
Ag	0.933	0.282	−0.062	0.900	−0.061	−0.138
Cd	0.927	0.240	0.036	0.900	−0.097	−0.041
Cu	0.814	0.307	0.220	0.735	0.031	0.160
Mn	0.751	0.482	0.212	0.601	0.242	0.171
Mo	0.691	0.583	0.043	0.515	0.362	0.015
Fe	0.652	0.591	0.410	0.446	0.378	0.386
P	0.318	0.124	−0.117	0.305	0.003	−0.141
Ca	−0.093	0.095	0.069	−0.143	0.142	0.086
Be	0.293	0.960	−0.110	−0.043	0.908	−0.071
Al	0.304	0.742	0.048	−0.040	0.896	0.086
La	0.373	0.890	−0.199	0.078	0.796	−0.173
K	0.292	0.661	0.008	0.060	0.597	0.026
V	0.388	0.636	0.452	0.130	0.563	0.461
Na	0.318	0.611	−0.322	0.138	0.510	−0.312
Li	−0.077	0.324	0.257	−0.231	0.394	0.290
Ba	0.417	0.519	−0.183	0.266	0.383	−0.189
Ti	−0.174	0.240	−0.121	−0.267	0.316	−0.086
Co	0.130	0.059	0.951	0.025	0.079	0.947
Ni	−0.114	−0.334	0.897	−0.082	−0.252	0.889
Cr	0.096	−0.404	0.567	0.204	−0.431	0.530
Mg	−0.281	−0.263	0.424	−0.242	−0.144	0.435
Sr	−0.006	−0.073	−0.233	0.045	−0.093	−0.239

factors introduced few complications and, since there is no absolute need for orthogonal solution, oblique rotation was preferred.

Table 11 summarises the maximum likelihood solution. The loading value selected as the interpretation criteria was 0.50. Each factor was attributed, in descending loading order, the variables that correlate best with it and that should be used for interpretation.

Table 11. Variables attributed to each factor

FACTOR 1	FACTOR 2	FACTOR 3	Not attributed to any factor
Zn	Be	Co	Fe
Pb	Al	Ni	P
Ag	La	Cr	Ca
Cd	K		Li
Cu	V		Ba
Mn	Na		Ti
Mo			Mg
			Sr

It is was then possible to identify the factors as known processes in the Lavrio area:

- FACTOR 1 is a combination of the presence of mixed sulphide ore in the study area and the human induced redistribution of the elements related to the deposits through mining,
- FACTOR 2 attracts the elements that relate with the clay content of the sampled soils and have very different geochemical properties from the elements attracted to the first factor, and
- FACTOR 3 represents, in general lines, the geological background reflecting the presence of small ultramafic bodies in the mass of the overthrust phyllite nappe.

Finally, the elements that are not attributed to any factor are either too complex in their distribution (as already mentioned for Fe), or do not comply with the processes that have been identified for the factors and the elements that agree with them. They therefore do not comply with the factor solution either.

6
Conclusions

In order to select the most appropriate technique to answer the research questions imposed in soil contamination studies, it is essential to start with the basis of all multivariate analysis techniques which is suitable data.

In this research it was found that the nature and characteristics of the available soil pollution data should drive the analysis. On the other hand the techniques employed are directed by the nature of the research questions that need to be addressed. When the research question concerns the underlying structure of the data, the appropriate method to extract this information is principal component and factor analysis.

The next tiers of the methodology are presented in the second paper of the series. The aim will be to appraise the geographic distribution of the dominant processes in soil composition and/or contamination and to enable the researchers to distinguish different sources of pollution that coexist.

References

Bikel PJ, Doksum KA (1977) Mathematical Statistics. San Francisco: Holden-Day

Cattell RB (1966) The scree test for the number of factors. Multivariate Behavioral Research 1:245–276

Cohen J, Cohen P (1975) Applied multiple regression/correlation analysis for the behavioral sciences. New York: Erlbaum

Comrey AL (1962) The minimum residual method for factor analysis. Psychological Reports 11:15–18

Comrey AL (1973) A first course in factor analysis. New York: Academic Press

Davies BE, Ballinger RC (1990) Heavy metals in soils in north Somerset, England, with special reference to contamination from base metal mining in the Mendips. Environmental Geochemistry and Health 12:291–300

Demetriades A, Stavrakis A, Vergou-Vichou K (1996) Contamination of surface soil of the Lavreotiki peninsula (Attiki, Greece) by mining and smelting activities. Mineral Wealth 98:7–15

Durucan S, Korre A (1997) Statistical and Spatial Assessment of Soil Composition and Heavy Metal Contamination around Lavrio Mine Workings, Greece. Proc. SWEMP'98 Ankara. 253–258. Balkema

Dziuban CD, Harris CW (1973) On the extraction of components and the applicability of the factor model. American Educational Research Journal 10:93–99

Ferguson CC (1992) The statistical basis for spatial sampling of contaminated land. Ground Engineering 25 (5):34–38

Goovaerts P, Journel AG (1995) Integrating soil map information in modelling the spatial variation of continuous soil properties. European Journal of Soil Science 46:397–414
Goovaerts P, Webster R, Dubois J-P (1997) Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. Environmental and Ecological Statistics 4:31–48
Gorsuch RL (1983) Factor analysis. Hillsdale, N.J.: Erlbaum
Harman HH (1967) Modern Factor Analysis. 3. Ed. University of Chicago Press
Harman HH, Jones WH (1966) Factor analysis by minimising residuals (Minres). Psychometrika. 31:351–368
Hotelling H (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24:417–441, 498–520
Kontopoulos A, Papassiopi N, Komnitsas K, Xenidis A (1996) Environmental characterisation and remediation of the tailings and soils in Lavrion. Proc. Conf. on Protection and Rehabilitation of the Environment III, Chania, Greece:484–493
Korre A (1997) A methodology for the statistical and spatial analysis of soil contamination in GIS. PhD Thesis, Imperial College, London
Korre A, Durucan S (1995) The application of geographic information systems to the analysis and mapping of heavy metal contamination around Lavrio mine workings, Greece. Proc. APCOM XXV, Brisbane:579–585
Lavrio Health Centre, Municip. of Lavrio, General State Hospital of Athens (1989) Proc. of Symp. on Health – Environment and Lead, Sounio: Greece
Lawley DN, Maxwell AE (1963) Factor analysis as a statistical method. London: Butterworth
Lawley DN, Maxwell AE (1971) Factor analysis as a statistical method. New York: Macmillan Publishing Co
Lee HB, Comrey AL (1979) Distortions in a commonly used factor analytic procedure. Multivariate Behavioural Research 14:301–321
Li X, Thornton I (1993) Multi-element contamination of soils and plants in old mining areas, U.K. Applied Geochemistry suppl. 2:51–56
Morrison DF (1976) Multivariate statistical methods. 2. Ed. New York: McGraw-Hill Book Co
Mulaik SA (1972) The foundations of factor analysis. New York: McGraw-Hill
Pearson K (1901) On lines and planes of closest fit to a system of points in space. Philosophical Magazine. 2:557–572
Rummel RJ (1970) Applied factor analysis. Evanston, III.: Northern University Press
Selinus OS, Esbensen K (1995) Separating anthropogenic from natural anomalies in environmental geochemistry. Journal of Geochemical Exploration 55:55–66
Skoufadis C, Papassiopi N, Kontopoulos A (1997) Removal of heavy metals from contaminated soils with organic acids. Proc. Engineering Geology and the Environment. 2173–2178. Balkema
Spearman C (1904) General intelligence, objectively determined and measured. American Journal of Psychology 15:201–293
Tabachnick BG, Fidel LS (1989) Using Multivariate Statistics. USA: Harper Collins Publishers
Thornton I (1993) Environmental geochemistry and health in the 1990s: a global perspective. Applied Geochemistry suppl. 2:203–210