Applied Radiation and Isotopes 67 (2009) 1721–1728

Contents lists available at ScienceDirect

Applied Radiation and Isotopes

journal homepage: www.elsevier.com/locate/apradiso

Reduction of the dimensionality and comparative analysis of multivariate radiological data

M.K. Seddeek [a], A.M. Kozae [b], T. Sharshar [c,e], H.M. Badran [d,*]

[a] Department of Physics, Faculty of Education, Suez Canal University, Al-Arish, Egypt
[b] Department of Mathematics, Faculty of Science, Tanta University, Tanta 31527, Egypt
[c] Department of Physics and Chemistry, Faculty of Education, Kafr El-Shaikh University, Kafr El-Shaikh, Egypt
[d] Department of Physics, Faculty of Science, Tanta University, Tanta 31527, Egypt
[e] Physics Department, Faculty of Science, Taif University, Taif, 888 Hawiya, Saudi Arabia

ARTICLE INFO

Article history:
Received 24 December 2008
Received in revised form
9 April 2009
Accepted 9 April 2009

Keywords:
Radium-226
Thorium-232
Potassium-40
Black sand
Statistical techniques
Classification
Multivariate analysis
Rough sets

ABSTRACT

Computational methods were used to reduce the dimensionality and to find clusters of multivariate data. The variables were the natural radioactivity contents and the texture characteristics of sand samples. The application of discriminate analysis revealed that samples with high negative values of the former score have the highest contamination with black sand. Principal component analysis (PCA) revealed that radioactivity concentrations alone are sufficient for the classification. Rough set analysis (RSA) showed that the concentration of ^{238}U, ^{226}Ra or ^{232}Th, combined with the concentration of ^{40}K, can specify the clusters and characteristics of the sand. Both PCA and RSA show that ^{238}U, ^{226}Ra and ^{232}Th behave similarly. RSA revealed that one or two of them can be omitted without degrading predictions.

1. Introduction

Many applications require projection of multidimensional data onto some lower-dimensional space. The goal of the dimension reduction is to identify the features in a multidimensional space that contribute to the classification of interest significantly and to retain most of the information of the original data. This is crucial when the original dimensionality is too high to be manageable. Multivariate statistical techniques have been used in many areas. In this work, we investigated the application of some of these techniques to the radioactivity contents and texture characteristics of sand samples. Dimension reduction was used to extract adequate information for detecting similarities, differences, and relationships among these samples.

1.1. Principal component analysis

Principal component analysis (PCA) can be used to explore data in an effort to reduce their dimension (Mardia et al., 1979; Chatfield and Collins, 1980; Krzanowski, 2000). In this technique, correlations between variables are summarized in terms of a small number of underlying factors. The method allows expressing most of the variance within a data set by means of a smaller number of factors, or principal components (PCs). Each PC is a linear combination of the parameters of the original data, whereby each successive PC explains the maximum amount of variance possible not accounted for by the previous PCs. Each PC is orthogonal to and, therefore, uncorrelated with the other PCs.

Two matrices, known as scores and loadings, give a concise and simplified description of the variance in the data set. The PC loadings define the way in which the old variables are linearly combined to form the new variables. The loadings define the orientation of the computed PC plane with respect to the original variables and indicate which variables carry the greatest weight in transforming the position of the original samples from the data matrix into their new position in the score matrix. Scores are coordinates of the samples in the established model, and they can be regarded as new variables.

The adjustment is made, first, by subtracting the mean of the variables from the value of each variable. This adjustment is made because PCA deals with the co-variances among the original variables; so, the means are irrelevant. PCs are constructed as weighted averages of the original variables. Their values for a specific row are referred to as factor scores, component scores, or

* Corresponding author. Tel.: +20 40 344 352; fax: +20 40 350 804.
E-mail address: Hussein_badran@hotmail.com (H.M. Badran).

0969-8043/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apradiso.2009.04.001

simply scores. The basic equation of PCA in the matrix notation is given by

$$\mathbf{Y} = \mathbf{WX},$$

where **X** is the data matrix, **Y** is the matrix of scores, and **W** is a matrix of coefficients. Matrix **W** is calculated as

$$\mathbf{W} = \mathbf{UL}^{-1/2},$$

where **U** is in the matrix of the eigenvectors and **L** is the diagonal of the eigenvalues of the variance-covariance matrix (**S**). The eigenvectors are the weights that relate the scaled original variables ($x_i = [X_i - mean]/\sigma$) to the factors.

The eigenvectors and the factor loadings are then calculated. The former relate the scaled original variables to the factors, while the latter represent the correlations between the variables and factors. The score coefficients are then used to form the factor scores. The factor scores are the values of the factors for a particular row of data. These score coefficients are similar to the eigenvectors, but they are scaled so that the produced scores have a variance of unity rather than a variance equal to the eigenvalue. This makes the variances of the factors identical.

Outliers are observations that are very different from the bulk of the data. To identify them, two quantities, $\mathbf{T}^2$ and Q_k, are calculated for $k = 0, 1,2,\ldots,8$. Q_k and $\mathbf{T}^2$ measures the combined variability of all the variables in a single observation. It provides a scaled distance measure of an individual observation from the overall mean and is defined as

$$\mathbf{T}^2 = [\bar{\boldsymbol{x}} - \underline{\bar{\boldsymbol{x}}}]'\mathbf{S}^{-1}[\bar{\boldsymbol{x}} - \underline{\bar{\boldsymbol{x}}}],$$

where $\bar{\boldsymbol{x}}$ represents a p-variable observation vector, $\underline{\bar{\boldsymbol{x}}}$ stands for the p-variable mean vector, and $\mathbf{S}^{-1}$ is the inverse of the covariance matrix. Q_k represents the sum of squared residuals when an observation is predicted with the first k factors.

Outliers can strongly influence the statistical analysis and compromise results (see, e.g., Baenett and Lewis, 1994). Most standard multivariate analysis techniques rely on the assumption of normality and require estimates of both the location and scale parameters of the distribution. Outliers may distort the values of these estimators arbitrarily and render the results of the application of these techniques meaningless. It is well known that the classical rule for PCA is very sensitive to outliers because the sample covariance matrix is sensitive to them (Huber, 1981). In the multivariate case, a classical way to identify outliers is to calculate Mahalanobis' distance using robust estimators of the covariance matrix and the mean vector (see Visuri et al., 2000, for references of numerous proposed robust techniques). No evidence was found of severe outliers in the present data; so, the non-robust estimation is adequate.

1.2. Discriminate analysis

Discriminate analysis (DA) can discriminate between different populations. Thus, it can simplify a description of observations by finding the patterns in perplexing data and be used to determine which variables are the best predictors. In the two-group case, DA can also be thought of as (and is analogous to) multiple regression. In general, in the two-group case (groups labeled '1' and '2'), it is possible to fit a linear equation of the type

$$group = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots\cdots + b_mx_m,$$

where a is a constant and b_1, b_2,…b_m are regression coefficients. The interpretation of the results of a two-group problem is straightforward and follows the logic of multiple regression closely: Those variables with the largest (standardized) regression coefficients are the ones that contribute most to the prediction of group membership.

When there are more than two groups, we can estimate more than one discriminate function like in the equation presented above. For example, with three groups, we can estimate: (i) a function for discriminating between Group 1 and Groups 2 and 3 combined and (ii) another function for discriminating between Groups 2 and 3. Coefficients b_i in these discriminate functions could then be interpreted as described above.

When performing a multiple group discriminate analysis, we do not have to specify how to combine groups to form various discriminate functions. Rather, some optimal combination of variables can be determined automatically so that the first function provides the most overall discrimination between groups, the second provides second most, and so on. Moreover, the functions will be independent or orthogonal, that is, their contributions to the discrimination between groups will not overlap. Computationally, a canonical correlation analysis can be performed to determine the successive functions and canonical roots (the term root refers to the eigenvalues associated with the respective canonical function). The maximum number of functions will be equal to the number of groups minus one or the number of variables in the analysis, whichever is smaller.

The influence of each of the uncorrelated variables on the discriminate analysis can be calculated. The value λ_r (removed λ) is Wilks' λ computed to test the impact of removing this variable. The value λ_a (alone λ) is Wilks' λ that would be obtained for each of the p variables if these were the only uncorrelated variables used. The Wilks' λ in the case of k groups, each with p variables, is given by

$$\lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \prod_{j=1}^{m} \frac{1}{1+\lambda_j},$$

where λ_j is the jth eigenvalue corresponding to the eigenvector, m is the minimum of $k-1$, **T** is the within-groups variance–covariance matrix, and **W** is the total variance–covariance matrix. Matrices **T** and **W** are given by

$$\mathbf{T} = \sum_{k=1}^{N}\sum_{i=1}^{N_i}(\boldsymbol{X_{ki}} - \boldsymbol{M})(\boldsymbol{X}_{ki} - \boldsymbol{M})'$$

and

$$\mathbf{W} = \sum_{k=1}^{N}\sum_{i=1}^{N_i}(\boldsymbol{X_{ki}} - \boldsymbol{M_k})(\boldsymbol{X_{ki}} - \boldsymbol{M_k})',$$

where $\boldsymbol{M}$ is the vector of means of these variables across all groups and $\boldsymbol{M_k}$ is the vector of means of observations in the kth group. The values F_v^r (removed F-value) and F_v^a (alone F-value) give the F-ratio that is used to test the significance of removing a variable and the above Wilks' λ, respectively, while F_p^r (removed F-probability) and F_p^a (alone F-probability) are the probability (significance level) of removing a variable and the above F-ratio, respectively. The result of the test is positive (the variable is important) if this value is less than the value of α (0.10). R^2 is the value that would be obtained if this variable were regressed on all other uncorrelated variables. When this R^2 value is larger than 0.99, severe multicolinearity problems exist. The variables with large R^2 should be removed (one at a time), and the analysis should be rerun every time.

1.3. Rough set analysis

Various real-life applications of rough set analysis (RSA), originated by Pawlak (1982), have shown its usefulness in many domains. This theory depends on a topological structure, called a quasi discrete topology, generated by the equivalence classes of the relation defined on the collection of data (Pawlak, 1991).