The Nucleus ***52****, No. 2 (2015) 85-87*

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)

ISSN 2306-6539 (Online)

Classification of Surface and Deep Soil Samples Using Linear Discriminant Analysis

M. Wasim[1*], M. Ali[2] and M. Daud[1]

[1]*Chemistry Division, Pakistan Institute of Nuclear Science and Technology, P.O. Nilore, Islamabad, Pakistan*

[2]*Department of Physics, Karakoram International University, Gilgit, Pakistan*

wasim1968@gmail.com; manzoor.ali@kiu.edu.pk; dm6252@gmail.com

ARTICLE INFO

Article history :
Received : 05 March, 2015
Revised : 27 May, 2015
Accepted : 16 June, 2015

Keywords :
Gilgit soil;
Naturally occurring radionuclides;
Linear discriminant analysis

ABSTRACT

A statistical analysis was made of the activity concentrations measured in surface and deep soil samples for natural and anthropogenic γ-emitting radionuclides. Soil samples were obtained from 48 different locations in Gilgit, Pakistan covering about 50 km^2 areas at an average altitude of 1550 m above sea level. From each location two samples were collected: one from the top soil (2-6 cm) and another from a depth of 6-10 cm. Four radionuclides including ^{226}Ra, ^{232}Th ^{40}K and ^{137}Cs were quantified. The data was analyzed using t-test to find out activity concentration difference between the surface and depth samples. At the surface, the median activity concentrations were 23.7, 29.1, 4.6 and 115 Bq kg^{-1} for ^{226}Ra, ^{232}Th, ^{137}Cs and ^{40}K respectively. For the same radionuclides, the activity concentrations were respectively 25.5, 26.2, 2.9 and 191 Bq kg^{-1} for the depth samples. Principal component analysis (PCA) was applied to explore patterns within the data. A positive significant correlation was observed between the radionuclides ^{226}Ra and ^{232}Th. The data from PCA was further utilized in linear discriminant analysis (LDA) for the classification of surface and depth samples. LDA classified surface and depth samples with good predictability.

1. Introduction

Radiation dose to all living organisms on earth originates mostly from natural radionuclides [1]. These include ^{40}K and radionuclides from ^{238}U and ^{232}Th series. The contribution of anthropogenic radionuclides to total dose is almost negligible. Uranium, thorium and potassium are present in earth crust at an average concentration of 2.7 µg g^{-1}, 9-10 µg g^{-1}and 1.4 % respectively [2-4]. Igneous rocks contain relatively higher amount of uranium (0.03-4.7 µg g^{-1}) as compared to sedimentary rocks (1.5-2.2 %) [5]. The same is true for thorium and potassium. Classification of rock types is usually performed by a large number of factors, such as its physical properties and chemical constituents. Using instruments, it is now possible to generate a huge dataset containing a large number of variables. These include wavelengths, energies, elements, radionuclides etc. It is difficult to explore patterns inside a large dataset employing single variables with statistical data analysis tools. Multivariate methods have evolved with time, which help explore such datasets. Chemometrics provides powerful data analysis tools to interpret large data sets. Elemental contents along with chemometric methods have been used for the differentiation of healthy and diseased subjects such as hepatitis-C [6] and cancer patients [7]. Elemental composition has also been extensively employed in provenance studies [8]. There are several techniques from pattern recognition realm which facilitate this classification. Principal component analysis (PCA) is the most common method to explore data for this purpose [9]. Other more sophisticated techniques include discriminant analysis, cluster analysis, neural networks, support vector machine etc. [10, 11].

A number of analytical techniques have been used to collect soil data. In recent years, several pattern recognition methods have been employed in conjunction with different analytical techniques to classify soils using landscape [12], infrared spectrometry [13], radionuclides [14, 15] etc. Environmental monitoring studies also generate a large amount of data. The variations in environmental data include in homogeneity in the matrix and differences due to different source profiles. Dose from soil radioactivity depends upon the distribution of individual radionuclides in soil and their migration behaviour. The current paper explores the distribution of natural (^{226}Ra, ^{232}Th and ^{40}K) and man-made (^{137}Cs) radionuclides in surface and sub soil of Gilgit. Multivariate methods have been applied to find out any differences in the distribution of radionuclides on the surface and below. This characterization will help to understand soil properties for ecological, environmental and agricultural studies.

2. Experimental

In this study, data for 48 soil samples collected from the surface and at a depth of 6 – 10 cm from Gilgit

* Corresponding author

covering 50 km^2 areas has been used. Gilgit is located at 35^{o}55$'$ N and 74^{o}17$'$ E, at an average altitude of 1,500 m above sea level with 0.1 million population. The area has alluvial and lacustrine deposits due to blockage of river by glacial or landslide debris [16]. From each location, two samples were collected, mixed, dried in air, pulverized, sieved, homogenized and stored in plastic bottles. The sampling procedure and experimental conditions, in detail, can be found in our previous work [17].

3. Results

The activity concentrations (Bq kg^{-1}) of ^{226}Ra, ^{232}Th, ^{137}Cs and ^{40}K in the surface soil [17] have already been reported. The average activity for ^{226}Ra, ^{232}Th, ^{137}Cs and ^{40}K was 25.4, 29.1, 4.6 and 115 Bq kg^{-1} respectively for surface samples and was respectively 25.8, 35.1, 2.9 and 191 for depth samples. Activity concentrations of all radionuclides followed log-normal distribution in all samples except for ^{40}K, which was normally distributed in depth samples. There were significant positive correlations between the same radionuclides found in the surface and depth samples.

4. Discussion

The hypothesis that there is no difference between the means of paired samples is tested using one sided paired difference *t*-test [10]. The difference of activity between the surface and depth samples has been plotted in Fig. 1. It shows that most of the activity concentrations for ^{226}Ra, ^{232}Th and ^{137}Cs are higher in the surface samples. It was, however, the other way around for ^{40}K. The application of paired difference *t*-test confirms this prediction at 95% confidence level except for ^{226}Ra. There are two samples (Nos. 8 and 18) having relatively higher activity concentrations of ^{226}Ra. If these two samples are removed from the data and the *t*-test is applied again, it shows that activity concentration of ^{226}Ra is also significantly higher in the surface than in the depth samples.

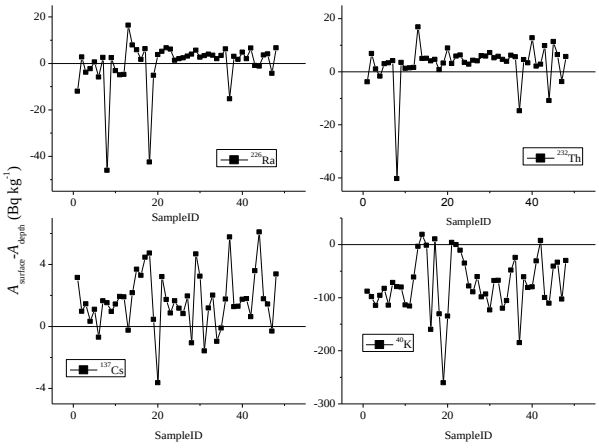

Fig. 1: Activity difference ($S_{surface}$ - A_{dept}) plotted for different radionuclides

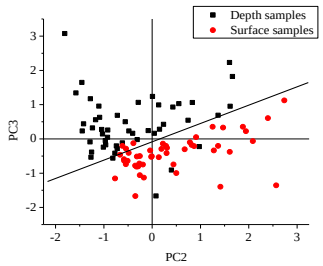

Fig. 2: Scores plot showing a distribution of two classes of samples across the line

4.1 PCA

PCA decomposes multivariate data into a set of abstract eigenvectors and an associated set of abstract eigen values [11, 18, 19]. Each eigen value presents a portion of the total variation in the data and each eigenvector is a linear combination of the original variables. PCA has been applied in this study to explore patterns within objects and within variables [20]. PCA was applied on the standardized data because activity concentrations of different radionuclides were on different scales. Data was standardized using Z-score [21]. The data matrix used for PCA was of size (96×4), *i.e.* 96 samples (48 surface sample + 48 depth samples) and 4 radionuclides. The first three principal components explained 97% of the total variance. Fig. 2 illustrates scores plot of second principal component (PC2) against the third principal component (PC3). The loadings plot, for the same PCs, is presented in

Fig. 3 reveals ^{232}Th and ^{226}Ra in positive correlation. In Fig. 2, the depth samples represented by black solid

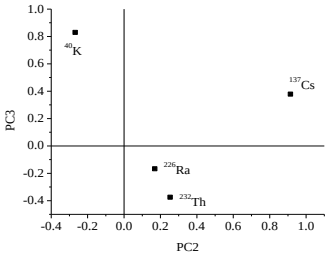

Fig. 3: Loadings plot shows positive correlation between ^{226}Ra and ^{232}Th

squares are clearly distinct from the surface samples represented by red solid circles. In fact, it seems possible to draw a line between the two groups. If above the line, a sample will belong to the depth otherwise to the surface.

4.2 Linear Discriminant Analysis (LDA)

A better way to find distribution of samples between two or more classes is the use of the LDA [21]. It is a linear method with discriminating character. The method maximizes the variance between the classes and minimizes the variance within the classes. It can be represented by projecting the objects on to a line at right angles to the discriminating line. It is possible to

determine the centre of each class along the projection and if the distance to the centre of class A is greater than that to class B, the object is placed in class A, and vice versa. In supervised pattern recognition, the distance of an object is defined from the centre of a class. Using this distance the linear discriminant function is calculated as;

$$f_i = (\bar{x}_A - \bar{x}_B) C_{AB}^{-1} x_i' \quad (1)$$

where C_{AB} is the pooled variance – covariance matrix,

$$C_{AB} = \frac{(N_A - 1)C_A + (N_B - 1)C_B}{(N_A + N_B - 2)} \quad (2)$$

N_A represents the number of objects in class A, and C_A is the variance – covariance matrix for this group with $\bar{x}_A$ the corresponding centroid. Mahalanobis distance from the centre of class A is measured as:

$$d_{iA} = \sqrt{(x_i - \bar{x}_A) C_A^{-1} (x_i - \bar{x}_A)'} \quad (3)$$

The predicted class for each object is the one whose centroid it is closest to the object.

Discriminant analysis was performed on the scores of PC2 and PC3. The results yielded 95.8 % samples from the depth as correctly classified and 79.2 % from surface. This study shows that a combination of PCA and LDA may be applied to classify or differentiate samples on the basis of their activity concentrations if any differences exist in their activity levels.

5. Conclusions

In this study, distribution behaviour of four radionuclides (^{226}Ra, ^{232}Th, ^{137}Cs and ^{40}K) was determined in the surface and deep soil samples. The samples were collected from the surface and at a depth of 6 - 10 cm from Gilgit. Principal component analysis revealed a positive significant correlation between ^{232}Th and ^{226}Ra. The *t*-test revealed that the levels of activity concentrations were significantly higher in the surface than in depth for ^{226}Ra, ^{232}Th and ^{137}Cs, whereas, the activity of ^{40}K was higher in depth than in the surface samples. These differences in activity concentrations at the surface and at depth were exploited to discriminate samples as surface or sub surface samples using principal component analysis and linear discriminant analysis. Modeling using linear discriminant analysis showed good predictability for this distribution.

References

[1] R. L. Kathren, "NORM sources and their origins". Appl. Radiat. Isot., vol. 49, p. 149-168, 1998.

[2] UNSCEAR, Effects of Atomic Radiation, United Nations, New York, 1982.

[3] NCRP, Ionizing radiation exposure of the population of the United States, National Council on Radiation Protection and Measurements, Bethesda, 1988.

[4] H. J. M. Bowen, Environmental chemistry of the elements, Academic Press, London, 1979.

[5] NCRP, Report No. 50, Environmental Radiation Measurements, National Council on Radiation Protection and Measurement, Washington DC, 1992.

[6] G. R. Lloyd, S. Ahmad, M. Wasim and R. G. Brereton, "Pattern recognition of inductively coupled plasma atomic emission spectroscopy of human scalp hair for discriminating between healthy and hepatitis C patients", Anal. Chim. Acta, vol. 649, p. 33-42, 2009.

[7] Z. Zhang, H. Zhuo, S. Liu and P. de B Harrington, "Classification of cancer patients based on elemental contents of serums using bidirectional associative memory networks", Anal. Chim. Acta, vol. 436, p. 281-291, 2001.

[8] C. Peltz and M. Bichler, "Classification of archaeologically stratified pumice by INAA", J. Radioanal. Nucl. Chem., vol. 248, p. 81-87, 2001.

[9] I. Jolliffe, "Principal component analysis", Wiley Online Library, 2015.

[10] Massart, B. Vandeginste, L. Buydens, S. De Jong, P. Lewi and J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.

[11] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. D. Jong, P. J. Lewi and J. Smeyers-Verbeke, "Handbook of chemometrics and qualimetrics, Part B", Elsevier, Amsterdam, 1998.

[12] A. Zhu, "Mapping soil landscape as spatial continua: the neural network approach". Water Resources Research, vol. 36, p. 663-677, 2000.

[13] P. H. Fidêncio, I. Ruisánchez and R. J. Poppi, "Application of artificial neural networks to the classification of soils from Sao Paulo state using near-infrared spectroscopy", Analyst, vol. 126, p. 2194-2200, 2001.

[14] S. Dragović and A. Onjia, "Classification of soil samples according to their geographic origin using gamma-ray spectrometry and principal component analysis". J. Environ. Radioact., vol. 89, p. 150-158, 2006.

[15] S. Dragovic and A. Onjia, "Classification of soil samples according to geographic origin using gamma-ray spectrometry and pattern recognition methods". Appl. Radiat. Isot., vol. 65, p. 218-224, 2007.

[16] J. F. Ivanac, D. M. Traves and D. King, "Records of the Geological Survey of Pakistan, vol. VIII Part 2", Geological Survey of Pakistan, 1956.

[17] M. Ali, M. Wasim, M. Arif, J. H. Zaidi, Y. Anwar and F. Saif, "Determination of the natural and anthropogenic radioactivity in the soil of Gilgit—a town in the foothills of Hindukush range", Health Phys., vol. 98 (Supplement 2), p. S69-S75, 2010.

[18] S. Wold, K. Esbensen and P. Geladi, "Principal component analysis". Chemom. Intell. Lab. Syst., vol. 2, p. 37-52, 1987.

[19] M. Daud, M. Wasim, N. Khalid, J. H. Zaidi and J. Iqbal, "Assessment of elemental pollution in soil of Islamabad city using instrumental neutron activation analysis and atomic absorption spectrometry techniques", Radiochim. Acta, vol. 97, p. 117-122, 2009.

[20] M. Wasim, M. S. Hassan and R. G. Brereton, "Evaluation of chemometric methods for determining the number and position of components in high-performance liquid chromatography detected by diode array detector and on-flow ^{1}H nuclear magnetic resonance spectroscopy", Analyst, vol. 128, p. 1082-1090, 2003.

[21] R. G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, Wiley, Chichester, 2003.