PHYSICAL CHEMISTRY 2006

Proceedings

of the 8th International Conference on Fundamental and Applied Aspects of Physical Chemistry

September 26-29,
Belgrade, Serbia

ISBN	86-82139-26-X
Title:	Physical Chemistry 2006. (Proceedings)
Editors	Prof. dr A. Antić-Jovanović
Published by:	The Society of Physical Chemists of Serbia, Studentski trg 12-16, P.O.Box 137, 11001 Belgrade, Serbia
Publisher:	Society of Physical Chemists of Serbia
For publisher:	Prof. dr S. Anić, president of the Society of Physical Chemists of Serbia
Printed by:	"Jovan" Printing and Published Comp; 250 Copies; Number of Pages: x + 442; Format B5; Printing finished in September 2006.
Text and Layout:	Aleksandar Nikolić

250 – copy printing

PATTERN RECOGNITION METHODS FOR CLASSIFICATION OF SOILS BASED ON THEIR RADIONUCLIDE CONTENT

S. Dragović[1] and A. Onjia[2]

[1] *Institute for the Application of Nuclear Energy – INEP, Banatska 31b, Belgrade, Serbia*
[2] *Vinča Institute of Nuclear Sciences, P.O. Box 522, Belgrade, Serbia*

Abstract

Multivariate data analysis methods were used to recognize and classify soils of unknown geographic origin. A total of 103 soil samples were differentiated into classes, according to regions in Serbia and Montenegro from which they were collected. Their radionuclide (^{226}Ra, ^{238}U, ^{235}U, ^{40}K, ^{134}Cs, ^{137}Cs, ^{232}Th and ^{7}Be) activities detected by gamma-ray spectrometry were then used as the inputs in different pattern recognition methods. The prediction ability of linear discriminant analysis (LDA), *k*-nearest neighbours (KNN), soft independent modelling of class analogy (SIMCA) and artificial neural network (ANN) were 82.8%, 88.6%, 60.0% and 92.1%, respectively.

Introduction

There are only few studies on employing the chemometric approach to describe the correlation between locations and radioactive contamination, in which artificial neural networks was used [1,2]. This approach is used predominantly for identifying of radioactive isotopes [3], quantitative gamma-ray spectrometry analysis [4] and for optimization of gamma-ray spectrometric measurements [5,6].

In this work the feasibility of using pattern recognition techniques to predict soil origin based on its radionuclide content was examined. Four commom pattern recognition techniques, LDA, KNN, SIMCA and ANN, were employed to classify soil samples based on their radionuclide contents determined by gamma-ray spectrometry. Since a predominant part of the radioactivity in soil is derived from the decay of the primordial radionuclides, ^{238}U, ^{232}Th and ^{40}K and the significant amounts of ^{137}Cs are also present in soil as a result of nuclear weapon tests and nuclear accidents, these radionuclides and their descendants are used for prediction. Soil samples (n=103) were collected from fifteen geographic regions of Serbia and Montenegro and their activities determined using an HPGe gamma-ray spectrometer ORTEC-AMETEK model GEM 25 (34% relative efficiency and 1.65 keV FWHM for ^{60}Co at 1.33 MeV, 8192 channels).

Details on employed pattern recognition methods can be found elsewhere [7]. It is only important to indicate the differences in thgeir characteristics and in the way in which each of them define the classification rules. LDA is a parametric method which searches for optimal boundaries between classes while it assumes that all classes have the same multi-normal distribution and that they are linearly separable. KNN is a non-parametric method which is very simple from a mathematical point of view and free from statistical assumptions; however, it is very sensitive to gross inequalities in the number of objects in each class. ANN does not impose any condition on the data structure. SIMCA is based on principal components for each category and critical distances with probabilistic signification; hence this implies that a spatial and probabilistic structure is present in the data.

Results and Discussion

Table 1 shows the activities (Bq/kg) of radionuclides in the analyzed soil samples. In general, radionuclide activities in these soils varied by a factor of up to 3-5 for ^{226}Ra, ^{238}U, ^{235}U, ^{40}K and ^{232}Th, 7 for ^{7}Be, 13 for ^{134}Cs, and 21 for ^{137}Cs. The distribution parameters (skewness and kurtosis) indicated normal (Gaussian) distribution.

Table 1. Descriptive statistics of radionuclide activities (Bq/kg) in soil samples

Variable	Mean	SD	Min.	Max.	Median	Skewness	Kurtosis
^{226}Ra	30.8	9.14	13.6	54.9	29.9	0.42	-0.45
^{238}U	29.7	9.42	14.6	53.4	30.2	0.44	-0.33
^{235}U	1.37	0.46	0.51	2.61	1.38	0.53	0.22
^{40}K	567	164	271	919	593	-0.17	-0.73
^{232}Th	40.7	13.5	18.3	83.4	39.7	1.20	2.18
^{134}Cs	0.09	0.07	0.02	0.26	0.05	0.91	-0.56
^{137}Cs	48.3	26.2	5.25	112	42.9	0.44	-0.39
^{7}Be	1.79	1.05	0.54	3.94	1.50	0.66	-0.91

The reliability of the classification models achieved was studied in terms of recognition ability (percentage of the members of the training set correctly classified) and prediction ability (percentage of the members of the test set correctly classified by using the rules developed in the training step).

Table 2 shows recognition and prediction abilities afforded with each multivariate technique. The recognition ability of these methods ranged from 86.4 to 100% and the prediction ability from 60.0 to 92.1%.

LDA produced a recognition rate of 95.1%; however, prediction ability was less satisfactory (82.8%). The highest rate of recognition achieved by KNN classifier was 86.4% and that of prediction was 88.6%. The number of neighbours (*k*) assayed were 1,3,5,7 and 9. The results obtained by SIMCA were very poor, giving only 60.0% of correct assignation for samples from test set containing autoscaled data. Meancentering of the data resulted in prediction ability of only 42.8%. In this application, ANN method overperforming the classification capabilities of other techniques. All soil samples presented to the network were correctly classified, i.e. the recognition ability was 100% and the prediction ability was 92.1%.

Table 2. Recognition and prediction ability of LDA, KNN, SIMCA, and ANN to classify soil samples

Method	Recognition ability (%)	Prediction ability (%)
LDA	95.1	82.8
KNN (K=1; inverse squared Euclidean distance)	86.4	88.6
SIMCA (normal range; α=0.05)	86.4	60.0
ANN (8×7×15; η=0.1; μ=0.1; sigmoidal transfer function)	100.0	92.1

Neural network are able to reproduce the training data in much more detail than other statistical techniques, which is particularly important if the specific problem under consideration requires a more detailed solution which the usual statistical approach based on means and variances may be unable to provide. The main advantage of neural network-based decision models lies in their adaptive behaviour.

In gamma-ray spectrometry ANNs have shown many advantages because of their full-parallel computing, simple implementation and nonlinearity. Automated pattern recognition systems such as neural networks are particularly useful in hazardous environments since an environment can be monitored without risk to human operators.

References

[1] M. Kanevski, R. Arutyunyan, L. Bolshov, V. Demyanov, M. Maignan, Geoinform., 1996, **7**, 5-11.

[2] M. Kanevski, V. Demyanov, M. Maignan, Conference ICANN 1997, 1997, Lausanne.

[3] P. Olmos, J.C. Diaz, J.M Perez, P. Gomez, V. Rodellar, P. Aguayo, A. Bru, G. Garcia Belmonte, J.L. de Pablos, IEEE Trans. Nucl. Sci., 1991, **38**, 971-975.

[4] V. Pilato, F. Tola, J.M. Martinez, M. Huver, Nucl. Instr. Meth. A, 1999, **422**, 423-427.

[5] S. Dragović, A. Onjia, S. Stanković, I. Aničin, G. Bačić, Nucl. Instr. Meth. Phys. Res. A, 2005, **540**, 455-463.

[6] S. Dragović, A. Onjia, Appl. Radiat. Isot. 2005, **63**, 363-366.

[7] D. Massart, B. Vandeginste, S. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, 1998, New York, USA.