Multivariate Statistical Tools for Classification of Soils and Ground Waters in Apulian Agricultural Sites

P. Ielpo[1,2], L. Trizio[3], D. Cassano[1], G. Pappagallo[1], A. Lopez[1], V. F. Uricchio[1]
[1] Water Research Institute-National Research Council, Bari, 70132, Italy
[2] Institute of Atmospheric Sciences and Climate, National Research Council, 73100, Lecce, Italy
[3] Arpa Puglia, Direzione scientifica U.O. Aria, corso Trieste 27, Bari

Abstract

In the paper results obtained from multivariate statistical techniques such as PCA (Principal Component Analysis), CA (Cluster Analysis), DA (Discriminant Function Analysis) and RBF-NN (Radial Basis Function Neural Networks) applied to wide data sets of ground waters and soils are shown. The obtained information can support for the water and soil resource management.

Introduction

Multivariate statistical techniques are widely used in environmental data

In the present paper multivariate statistical techniques such as PCA, CA, DA and RBF-NN have been applied to soils data set (306 samples and 15 parameters). The results obtained have been compared with those obtained by the same statistical methods to the ground waters data set (1009 samples and 15 parameters). Both data sets were obtained from monitoring activity performed during the years 2004-2007, in the frame of the project "Improvement of the Regional Agro-meteorological Monitoring Network". The project funded by the Agriculture and Food Authority of Apulia Region concerned all the Apulian provinces.

Materials & Methods

The chemical and physical parameters analyzed on soil samples were skeleton, sand, silt, clay, pH, electrical conductivity, water content, total carbonate, organic carbon, total nitrogen, extractable K_2O, extractable Na, extractable Ca, extractable Mg, assimilated P_2O_5. The chemical and physical analyses of soil samples were carried out according to official guidelines proposed by the National Agriculture Authority in a specific law [1]. The parameters analyzed on groundwater samples were pH, Electrical Conductivity (EC), Total Dissolved Solids (TDS), Dissolved Oxygen (DO), Chemical Oxygen Demand (COD), the major ions (ie. Na^+, Ca^{2+}, Mg^{2+}, K^+, Cl^-, NO_3^-, SO_4^{2-} and HCO_3^-), vital organism at 22 °C and 36 °C. The chemical and physical analyses of the water samples were carried out according to official guidelines proposed by the National Agriculture Authority in a specific law [2].

Multivariate statistical analyses of the soil and groundwater data sets included principal component analysis (PCA), cluster analysis (CA), discriminant function analysis (DFA), and radial basis function neural network (RBF-NN) [3].

Results

Table 1 shows the performance of the DFA model in terms of soil classification. The model was able to discriminate among the five provinces in a good manner (67.2%). The lowest discrimination was associated with Brindisi province data set. In terms of forecasting DFA model shows an overall performance of 65.4%.

Table 1: DFA Classification confusion matrix for soil data set.

from \ to	Bari	Brindisi	Foggia	Lecce	Taranto	Total	% correct
Bari	43	0	12	3	2	60	71.67%
Brindisi	3	13	3	6	7	32	40.63%
Foggia	16	0	30	2	2	50	60.00%
Lecce	1	4	1	46	9	61	75.41%
Taranto	7	2	0	4	40	53	75.47%
Total	70	19	46	61	60	256	67.19%

Considering the training set results of RBF-NN (not shown here) in comparing with the classification results of DFA, it is possible to note that the RBF-NN shows a very high accuracy for all the provinces.

Table 2: DFA Forecasting confusion matrix for soil data set.

From \ To	Bari	Brindisi	Foggia	Lecce	Taranto	Total	% correct
Bari	50	0	14	5	6	75	66.67%
Brindisi	4	13	3	8	8	36	36.11%
Foggia	18	0	38	2	2	60	63.33%
Lecce	3	6	1	52	11	73	71.23%
Taranto	9	2	0	4	47	62	75.81%
Total	84	21	56	71	74	306	65.36%

Among the input variables those with higher importance for soils data set are total carbonate and Electrical Conductivity.

Table 3: Relative importance of input variables obtained fron RBF-NN.

Variable	**Importance (%)**		
Total carbonate	100	Organic Carbon	19
Elect. Cond.	98	PH (H_2O)	15
extractable K_2O	84	extractable Na	12
Silt	75	extractable Mg	11
Skeleton	64	extractable Ca	10
Clay	44	Assimilated P_2O_5	9
Total Nitrogen	21	water content	8

Conclusions

The results from statistical methods applied to ground waters and soils data sets pointed out that the groundwater pollution sources pressuring the sites were similar among the provinces [3] and among these, the agricultural practices and marine water intrusion are those relevant; the application to soil data set has better pointed out the state of soil's fertility among Apulia provinces. Generally the RBF-NN gave better performance than DFA in classification and it suggested that among parameters investigated the electrical conductivity and Mg^{2+} are the variables with greater relative importance for the ground waters, while total carbonate and electrical conductivity are those with greater relative importance for soils.

References

1) National Agriculture Authority, Official methods of soil chemical analyses, Decreto Ministeriale, 13/09/1999

2) National Agriculture Authority, Official methods for analysing of agricultural usage and livestock waters, Decreto Ministeriale, 23/03/2000

3) Ielpo, P., Cassano, D. Uricchio, V. F. Lopez A, Pappagallo, G., Trizio, L., de Gennaro G., Identification of pollution sources and classification of Apulia region groundwaters by multivariate statistical methods and neural networks. *T ASABE* 56 (6) (2013), 1377-1386

P. Ielpo et al., Multivariate Statistical Tools for Classification of Soils and Ground Waters in Apulian Agricultural Sites