IDENTIFICATION OF POLLUTION SOURCES AND CLASSIFICATION OF APULIA REGION GROUNDWATERS BY MULTIVARIATE STATISTICAL METHODS AND NEURAL NETWORKS

P. Ielpo, D. Cassano, V. F. Uricchio, A. Lopez, G. Pappagallo, L. Trizio, G. de Gennaro

ABSTRACT. Multivariate statistical techniques, including discriminant function analysis (DFA), cluster analysis (CA), principal component analysis (PCA), absolute principal component score (APCS), and radial basis function neural network (RBF-NN), were applied to a data set formed by 905 samples and 15 parameters, including pH, electrical conductivity (EC), total dissolved solids (TDS), dissolved oxygen (O₂), chemical oxygen demand (COD), Na⁺, Ca²⁺, Mg²⁺, K⁺, Cl⁻, NO₃⁻, SO₄²⁻, HCO₃⁻, and vital organism at 22°C and 36°C, of groundwater samples collected in the Apulia region (southern Italy). Among all collected samples, only samples showing values for all parameters were used to compose the data set on which the multivariate statistical techniques were applied. PCA and APCS allowed us to identify, for each province as well as the sites diverging from the main cluster, the pollution sources pressuring the sampling sites investigated: they were identified as fertilizer applications, the use of unpurified irrigation water, marine water intrusion, and calcareous characteristics of the soil. We found that the groundwater pollution sources pressuring the sites were similar among the five Apulian provinces (Foggia, Bari, Brindisi, Lecce, and Taranto). Moreover, for each province, marine water intrusion showed the highest contribution. The application of DFA to the data set allowed us to obtain good results in discriminating among four provinces, with the exception of Taranto. The model also gave good performance results in forecasting. However, RBF-NN provided more accurate results than DFA and confirmed that EC had the greatest relative importance. This is probably due to the different salinity among the sites (Na⁺ also showed good discriminant importance). In fact, with PCA and APCS, it was possible to observe that EC, together with Na⁺, Cl⁻, and TDS, was the parameter that most often showed high loading values, and the scattered samples with these loading values were collected at sites in which marine water intrusion had been hypothesized. The results obtained by multivariate statistical methods can be useful both in guiding stakeholders and in providing a valid tool to authorities for assessing and managing groundwater resources.

Keywords. Classification, Cluster analysis, Discriminant function analysis, Forecasting, Groundwaters, Neural networks, Principal component analysis, Source apportionment.