Global NEST Journal, Vol 14, No 1, pp 3-9, 2012
Copyright© 2012 Global NEST
Printed in Greece. All rights reserved

SOIL CONTAMINATION INTERPRETATION USING SELF-ORGANIZING MAPS

Ts. VOYSLAVOV
S. TSAKOVSKI
V. SIMEONOV*

Chair of Analytical Chemistry
Faculty of Chemistry
University of Sofia "St. Kl. Okhridski"
1164 Sofia, J. Bourchier Blv. 1, Bulgaria

Received:17/08/11
Accepted: 03/10/11

*to whom all correspondence should be addressed:
e-mail: VSimeonov@chem.uni-sofia.bg

ABSTRACT
The present study deals with the problem of soil contamination risk assessment for a region being actively impact by the routine production of Kremikovtzi metallurgical plant near Sofia, Bulgaria. Although the production is now cancelled, the soil pollution is present and needs careful assessment. The application of self-organizing maps classification strategy of Kohonen makes it possible to identify: a/ pollution sources in the region of interest; b/ spatial patterns of similarity of polluted sites and the reason for the specific pollution.

KEYWORDS: heavy metals, soil pollution, chemometrics, risk assessment.

1. INTRODUCTION

The anthropogenic activity is one of the main reasons for soil pollution. Especially high loads are due to contamination by heavy metals whose sources are mainly metallurgical plants, urban heating processes, agricultural treatment of land etc. The polluting emissions are mainly fixed in the upper soil zone and even after a partial remobilization of the heavy metals it could bring to groundwater contamination, increased input to agricultural plants and, thus, harmful effects in the human food chain.

The traditional assessment of soil pollution is based on the regular routine of comparison of allowable threshold values with the results of monitoring. This approach is even a required action in environmental agencies, agricultural administration and managing organization. Very often solving a particular problem concerning the soil pollution or respective decision making is based solely on single results and not on a more generalized model about the state of the soil pollution in a certain region. The application of multivariate statistical approaches to the problem allows a better classification, modeling and interpretation of the soil monitoring data. This environmetric strategy makes it possible to detect relationships between the chemical pollutants and specific soil parameters, between sampling sites and, therefore, to achieve a stratification of the pollution. Further, it becomes possible to identify possible pollution sources and to construct apportioning models allowing the determination of the contribution of each identified source to the formation of the total pollutant mass (Stanimirova *et al*., 2006; Einax and Soldt, 1995; Singh *et al.,* 2008; Andrade *et al*., 2007; Buszewski and Kowalkowski, 2006; Kemper and Sommer, 2002; Lee *et al.,* 2004; Stanimirova *et al*., 2009; Terrado *et al.,* 2007; Perez Pavon *et al*., 2008). The careful exploratory data analysis of polluted soil samples proves to be an important tool to assess the soil quality in endangered regions. Various approaches are usually applied in order to collect sufficient information from the monitoring results in order to interpret and model the data sets. This approach is, indeed, an advanced assessment procedure, which is very effective for seriously polluted region. In such cases the simple interpretation of the monitoring results involving only standard comparison of registered heavy metals concentrations with accepted as hazardous threshold values, does not allow a real data interpretation. The pollution process is, in principle, a multivariate one and only multivariate data treatment is appropriate for assessment activities.

In the last decades of the 20th century the annual production in the Kremikovtzi steelwork near City of Sofia, capital of in Bulgaria was about 1 million tons of steel. In addition, alloys and inorganic compounds, like lime, barite, and dolomite were produced. In recent investigations only the vertical distribution of some elements in different soil types was in focus (Schulin *et al.*, 2007). Recently, a geostatistical data interpretation of the heavy metals data set was published trying to localize the soil pollution status quo (Schaefer *et al.,* 2010). It is the aim of the present study to perform a multivariate statistical treatment of another data set of heavy metals collected on the polluted Kremikovtzi soils in order to detect additional details of the data set structure by the use of self-organizing maps of Kohonen (SOM) technique. This study improves the pollution assessment around the factory using 65 samples from the region inside and outside of the metallurgical plant.

2. METHODS

2.1 Investigation area, soil sampling, preparation, and analysis

The steelwork Kremikovtzi is located about 20 km in the north east of the Bulgarian capital Sofia. For this investigation 62 samples were taken in an area of about 64 km^2. 16 sampling points are located directly in the boundaries of the steelwork (Figure 1).

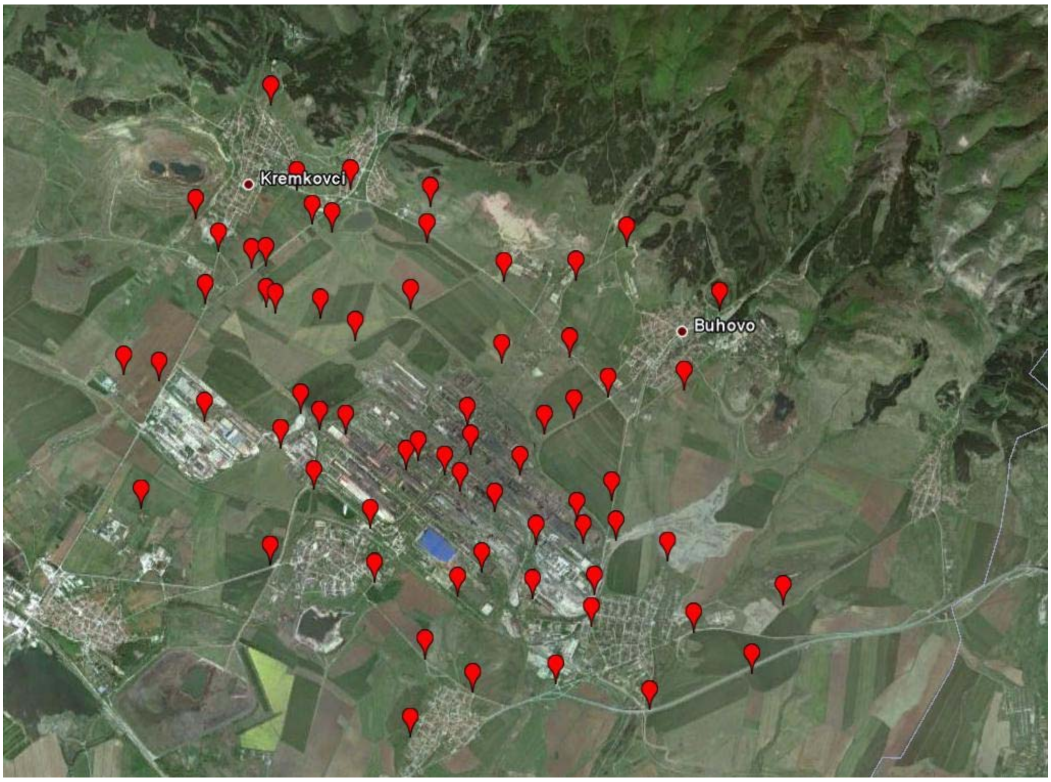

Figure 1. Sampling point location

Soil samples from the upper 0- to 20-cm layer were taken. The soil samples were dried, homogenized, and passed through a 2-mm-sieve. A microwave (power: 1200 W) aqua regia digestion with a mixture of 21 mL HCl (c = 12 mol L^{-1}) and 7 mL HNO_3 (c = 15.8 mol L^{-1}) according to the German standard (DIN EN 13346, 2001) was performed with 0.5 g of the soil twice for each sampling site. After cooling, the solutions were completed to 100 mL with diluted HNO_3 (c = 0.5 mol L^{-1}). Afterwards, the concentration of Cd, Cu, Fe, Mg, Mn, Na, Pb, and Zn was determined by means of different atomic spectroscopic techniques. All examined metals could be detected in concentrations higher than the detection limits. The trueness of the analytical methods was verified by analyzing certified reference material IAEA/Soil-7.

In Table 1 the basic statistics of the analytical results is presented (mean values, maximal and minimal value, standard deviation).

Table 1. Basic statistics of the monitoring results

Analyte	Cd	Cu	Fe	Mg	Mn	Na	Pb	Zn
Mean	1.2	86.3	58.7	5.0	4.3	315.4	318.3	237.7
Max	0.2	34.9	22.0	1.4	0.6	112.9	39.3	64.5
Min	6.1	291.6	301.2	13.1	15.4	1420.2	1624.8	1672.4
St. Dev.	1.0	56.4	47.7	2.8	3.9	214.4	348.0	255.8

2.2 Chemometric Approach

The SOM is an algorithm (Kohonen, 2001) used to visualize and interpret large high-dimensional data sets; it is an unsupervised pattern cognition method similar to cluster analysis. The main advantage of SOM is the simultaneous classification of variables and objects (sampling locations). Typical applications are visualizations of process states or financial results by representing the central dependencies within the data on the map. The map consists of a regular grid of processing units called neurons.

A model of some multidimensional observations, possibly a vector consisting of features (variables), is associated with each unit. The map attempts to represent all available observations with optimal accuracy using a restricted set of models. At the same time the models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other. Fitting of the model vectors is usually carried out by a sequential regression process, where *t* = 1,2,... is the step index. For each sample *x(t)*, the winner index *c* (best matching unit - BMU) is first identified by the condition:

$$\forall i, \|x(t) - m_c(t)\| \le \|x(t) - m_i(t)\| \tag{1}$$

where *i* is the node indication and *m* is node vector.

When the BMU has been found, the weight vectors of the SOM are updated so that the BMU is moved closer to the input vector in the input space.

Then, all the model vectors or a subset of them belonging to the nodes centered around node *c* = *c*(**x**) are updated as:

$$m_i(t+1) = m_i(t) + h_{c(x),i}(x(t) - m_i(t)) \,. \tag{2}$$

Here, $h_{c(x),i}$ is the “neighborhood function”, a decreasing function of the distance between the *i*-th and *c*-th nodes on the map grid. This regression is usually reiterated over the available objects.

The trained map can be graphically presented by 2D planes for each variable, with the variable distribution values being indicated by different colors on the different regions of the map. Additionally, the node “coordinates” (vectors) can be clustered by the non-hierarchical K-means classification algorithm. Throughout the calculations a software package for MATLAB was used.

3. RESULTS AND DISCUSSION

In Figure 2 SOMs for all parameters of interest and all sampling sites are presented. The color scale at the right site of the map indicates the concentration levels for the heavy metals analyzed at each sampling location (Cd and Cu are in μg g^{-1}, the rest in mg g^{-1}). It is readily seen that three specific patterns of heavy metals distribution are formed for the sampling area of interest. The first pattern resembles the similarity of the SOMs for Cd, Cu, Fe, Pb and Zn. This is an important indication about the role of a specific source of these metals in the polluted soils, obviously the major metallurgical activity – production of iron and steel in the time of highest production activity of the plant.

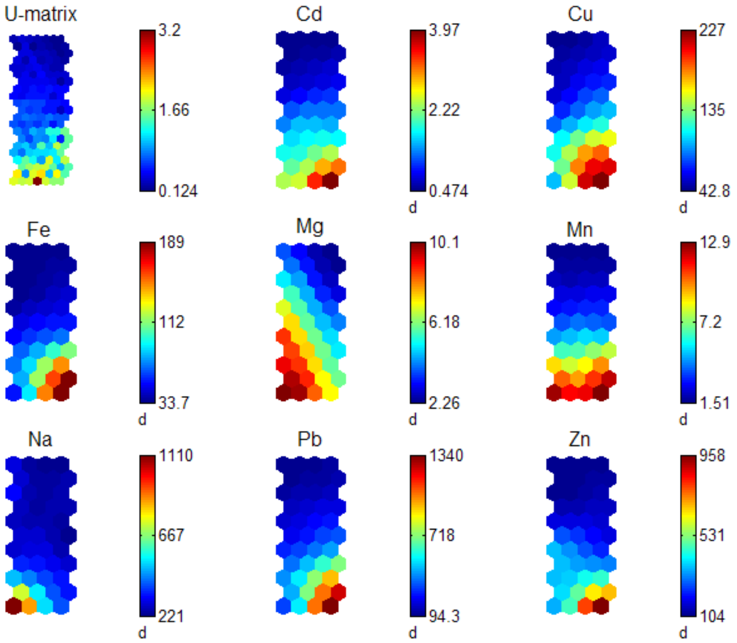

Figure 2. SOMs for all pollutants and all sampling points

In this case the highest metal concentrations are found in the right bottom corner of the maps. It describes the specific locations polluted with the metals mentioned and by the source discussed.

The second pattern consists of the maps for sodium and magnesium, which indicate enhanced metal concentrations on the left bottom corner of the maps. Obviously, the specificity in this case is related to the effect of another source of pollution (dolomite production) causing higher metal impact in another location around the metallurgical plant.

Except for these two basic soil pollution sources, another pattern of pollution could be detected if the third SOM pattern is interpreted. It includes only one map (that of manganese) characterized by increased concentrations of the metal both on left and right botton corner of the SOM. Thie particular area, therefore, indicates a mixed type of soil pollution caused by the simultaneous impact of the “metallurgical” and “dolomite” source.

Figure 3 represents the typical component plane plot where the similar patterns of maps are grouped in clusters of similarity. This figure proves completely the interpretation performed by the separate maps.

Table 2. Average values and standard deviations for the chemical parameters for each cluster

Variables		Average values				Standard deviations			
Element	Unit	1 (n=48)	2 (n=9)	3 (n=2)	4 (n=3)	1 (n=48)	2 (n=9)	3 (n=2)	4 (n=3)
Cd	µg g^{-1}	0.79	1.92	2.52	4.27	0.49	0.46	0.59	1.95
Cu	µg g^{-1}	62.56	165.74	79.26	232.03	21.31	46.60	17.58	51.69
Fe	mg g-1	40.62	105.81	40.35	218.72	9.57	33.90	5.94	73.58
Mg	mg g^{-1}	4.16	7.53	10.30	6.87	2.24	2.51	1.47	2.46
Mn	mg g^{-1}	2.50	8.82	13.46	14.24	1.35	2.21	1.64	1.42
Na	µg g^{-1}	257.80	380.97	1328.29	365.73	92.62	107.40	129.98	106.95
Pb	µg g^{-1}	181.56	662.25	241.67	1524.47	98.83	251.98	40.83	91.17
Zn	µg g^{-1}	159.09	369.48	327.25	1040.53	116.99	90.00	193.79	646.44

In the next step of data interpretation an effort was made to separate the sampling locations into groups of similarity using SOMs. In Figure 4 the sample sites grouping into four clusters and the number of hits into each node of the map are presented.

The first cluster is most populated (48 sampling points), in the second one there are 9 objects, in the third – 2 and in the fourth – 3 samples.

It is of substantial importance to determine the discriminator parameters for each one of the clusters in order to better understand the data structure and the pollution patterns for the region of interest. In Table 2 the averages of the variables and their standard deviations for each one of the clusters identified are shown.

In Table 3 each cluster is characterized by the group median values of each of the pollutants. Additionally, Kruskal – Wallis one-way analysis of variance test and Mann-Whitney test is given to illustrate the similarities and dissimilarities between clusters formed. It is seen that the different patterns are reliably described by several specific features as follows:

Cluster 1: It is the biggest group of objects. The averages for the pollutants in this case are actually gross averages of the other three groups. This is an indication that most of the region of interest has a homogeneous type of pollution characterized by relatively high levels but still below the threshold of serious pollution. The homogeneity is indicated also by relatively low levels of the relative standard deviations for all averages of the pollutants as compared to the deviations in the other groups.

Cluster 2: The lowest levels of cadmium and manganese are found for the rest of the clusters (2, 3 and 4). Since the average value for sodium is very close to the lowest calculated (for cluster 4), it could be assumed that the cluster presents a group of sampling points located in an area not affected by the dolomite production (low levels of sodium and manganese) with moderate impact from the metallurgical factor (iron and steel production).

Cluster 3: This cluster represents a tiny area in the region of the plant indicated by only two, which is strongly affected by the dolomite production. It shows highest concentrations for magnesium and sodium and lowest concentrations of all other heavy metals whose pollution effect is mostly related to the steel and iron production.

Cluster 4: This is also a small group of sampling points which reveal a completely different pattern as compared to that of cluster 3. The limited area in this case is seriously polluted by the activity of the steel and iron production (highest averages for the concentrations of Cd, Cu, Fe, Mn, Pb, and Zn), lowest concentrations for Na and Mg (dolomite production tracers).

The classification procedure gives the opportunity to identify typical soil pollution sources in the region of interest and to indicate the spatial variety of the pollution. These patterns are presented in Fig. 5.

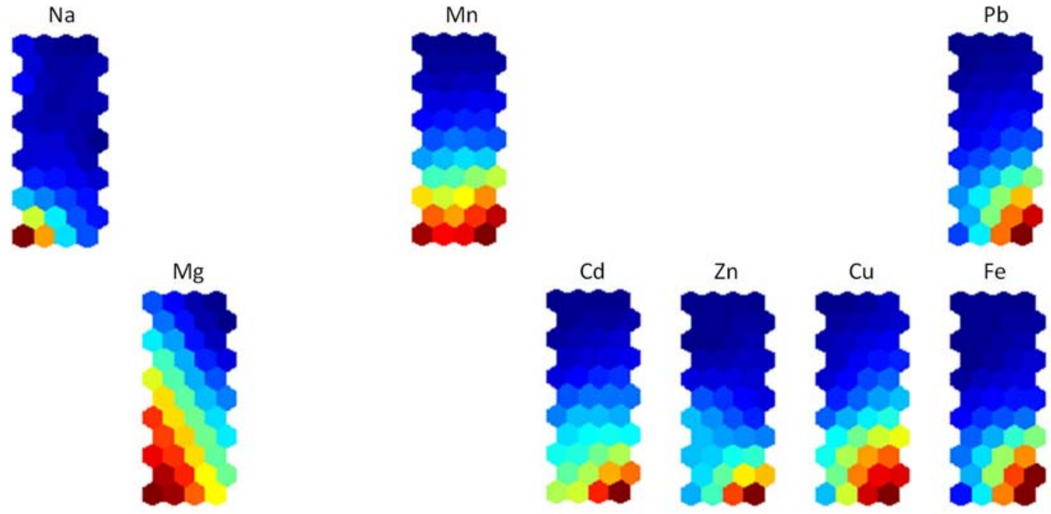

Figure 3. Components plane configuration

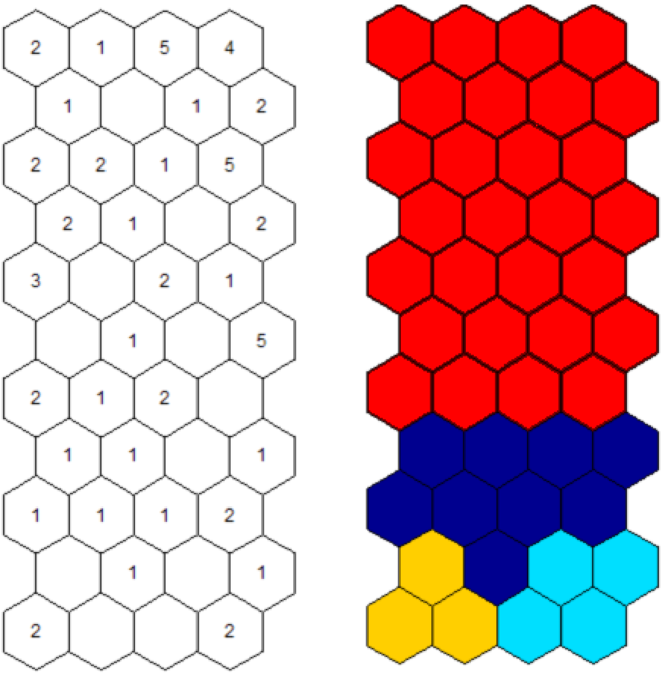

Figure 4. Clustering of the sampling sites and the number of hits for each node

Table 3. Statistical assessment of differences between cluster median values of variables (statistical level of significance lower than p<0.001 was marked as ‘‘++’’, while lower than p<0.05 as ‘‘+’’)

Variables		Median values				K-W test	Mann-Whitney U test					
Element	Unit	1 (n=48)	2 (n=9)	3 (n=2)	4 (n=3)		1-2	1-3	1-4	2-3	2-4	3-4
Cd	µg/g	0.67	1.77	2.52	4.55	++	++	+	++			
Cu	µg/g	58.88	168.98	76.26	204.93	++	++		++			
Fe	mg/g	39.94	108.88	36.15	195.14	++	++		++	+	+	
Mg	mg/g	3.70	7.68	9.26	6.37	++	++	+	+			
Mn	mg/g	2.09	8.37	12.29	14.72	++	++	+	++		+	
Na	µg/g	241.12	339.92	1236.38	405.19	++	++	+		+		
Pb	µg/g	152.83	752.65	241.67	1501.96	++	++		++	+	+	
Zn	µg/g	126.55	387.54	327.25	1068.78	++	++	+	++			

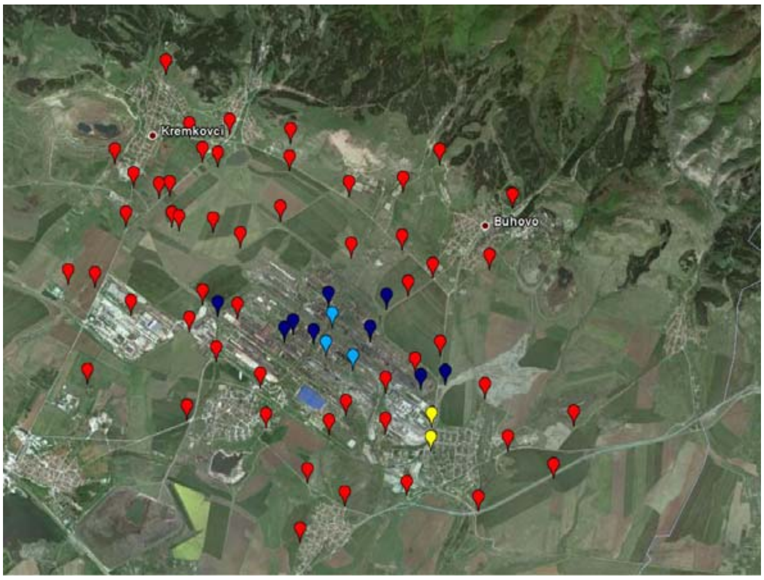

Figure 5. Soil pollution patterns in the region of interest

4. CONCLUSION

The use of self-organizing maps classification made it possible to project the multivariate soil monitoring data set and to determine various ways of grouping of sampling points depending on the chemical pollutant concentrations. Thus, a complete scheme of the polluted plant area could be constructed indicating both spatial and chemical pollution patterns. This approach works with relatively limited number of pollution indicators and sampling points allowing, however, a complete description of the hazards caused by the anthropogenic activity.

REFERENCES

Andrade J.M., Kubista M., Carlosena A. and Prada D., (2007), 3-Way characterization of soils by Procrustes rotation, matrix-augmented principal components analysis and parallel factor analysis, *Anal. Chim. Acta*, **603**, 20-29.

Buszewski B. and Kowalkowski T., (2006), A new model of heavy metal transport in the soil using nonlinear artificial neural networks, *Environ. Eng. Sci.,* **23**, 589–595.

Einax J.W. and Soldt U., (1995), Geostatistical investigations of polluted soils, *Fres. J. Anal. Chem.,* **351**, 48 – 53.

Kemper T. and Sommer S., (2002), Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy, *Environ. Sci. Technol.,* **36,** 2742 – 2747.

Kohonen T., (2001), Self-organizing maps of massive databases, *Int. J. Eng. Intell. Sys.Elec. Eng.Commun.,* **9**, 179-185.

Lee C.K., Ko E.J., Kim K.W. and Kim Y.J., (2004), Partial least square regression method for the detection of polycyclic aromatic hydrocarbons in the soil environment using laser-induced fluorescence spectroscopy, *Wat. Air Soil Pollut.*, **158**, 261 – 275.

Perez Pavon J.L., Garcia Pinto C., Guerrero Pena A. and Moreno Cordero B., (2008), Headspace mass spectrometry methodology: application to oil spill identification in soils, *Anal. Bioanal. Chem.,* **391**, 599 – 607.

Schaefer K., Einax J. W., Simeonov V. and Tsakovski S., (2010), Geostatistical and multivariate statistical analysis of heavily and manifoldly contaminated soil samples, *Anal. Bioanal. Chem.,* **396**, 2675-2683.

Schulin R., Curchod F., Mondeshka M., Daskalova A. and Keller A., (2007), Heavy metal contamination along soil transect in the vicinity of the iron smelter of Kremikovtzi (Bulgaria), *Geoderma*, **140**, 52-61.

Singh K.P., Malik A. Basant A. and Ojha P., (2008), Vertical characterization of soil contamination using multi-way modeling – A case study, *Environ. Monit. Assess*., **146,** 19 - 32.

Stanimirova I., Zehl K., Massart D.L., Vander Heyden Y. and Einax J.W., (2006), Chemometric analysis of soil pollution data using the Tucker N-way method, *Anal. Bioanal. Chem.,* **385**, 771-779.

Stanimirova I., Kita, A., Malkowski E., John E. and Walczak B., (2009), N-way exploration of environmental data obtained from sequential extraction procedure, *Chemom. Intell. Lab. Sys.,* **96**, 203 – 209.

Terrado M., Kuster M., Raldua D., Lopez De Alda M., Barcelo D. and Tauler R., (2007), Use of chemometric and geostatistical methods to evaluate pesticide pollution in the irrigation and drainage channels of the Ebro river delta during the rice-growing season, *Anal. Bioanal. Chem.*, **387**, 1479 – 1488.