S E N S ' 2 0 0 6
Second Scientific Conference with International Participation
SPACE, ECOLOGY, NANOTECHNOLOGY, SAFETY
14 – 16 June 2006, Varna, Bulgaria

AIR QUALITY ASSESSMENT BY MULTIVARIATE STATISTICS

Vasil Simeonov[1], Stefan Tsakovski[1], Pavlina Simeonova[2]

[1]Faculty of Chemistry, University of Sofia "St. Kl. Okhridski", 1164 Sofia, 1, J. Bourchier Blvd.,Bulgaria, vsimeonov@chem.uni-sofia.bg
[2]Institute of Solid State Physics, Bulgarian Academy of Sciences, 1784 Sofia, 72, Tzarigradsko Chaussee Blvd, Bulgaria, poly-sim@issp.bas.bg

Key words: aerosols, environmetrics, apportioning, pollution sources

Abstract. *The present communication deals with the application of several chemometrical methods (cluster and principal components analysis, source apportioning on absolute principal components scores) to an aerosol data collection from Arnoldstein, Austria. It is convincingly shown that six latent factors explaining almost 80 % of the total variance are responsible for the data structure and are conditionally identified as "fertilizer", secondary emission", "lead smelter", "traffic", "salt" and "soil dust". Further more, the contribution of each identified source to the formation of the particle total mass and chemical compounds total concentration is calculated. Thus, a reliable assessment of the air quality in the region of observation is achieved. The apportioning models obtained are checked for adequateness and validated. It is explained why for sodium and magnesium non-adequate models are obtained. The latent factor contribution models can be further used for risk assessment and respective decision making. Additionally, it is commented why chemometrics could be successfully applied as sustainability metrics in various aspects of interpretation of the state of "sustainable development"*

Introduction

In the recent years the concept of sustainable development has been successfully introduced and exploited not only as legislative and political formula but also as a background of various research projects. The question asked is not if a certain activity is sustainable but what recommendation is needed in order an activity to be considered as sustainable. In sense of the environmental problems this seems quite formal and undefined [1]. For instance, the traditional acceptance of the sustainability idea is to find a sound compromise between the constantly increasing needs of the mankind for energy and raw materials on one hand, and, on the other, the social requirement for clean environment and better chances for the coming generations.

It is obvious that the rapid technological and social development requires some kind of sustainability metrics in order to control and implant the system of sustainable development. The challenge to be environmentally relevant has led to the development of two important concepts for sustainability indicators:

- The P – S – R indicator concept (the *pressure* of the socio-economic activities into natural systems leads to observable changes in the *state* of the environmental systems, which causes respective *response* or socio-economic measures to reduce the hazardous effects);

- The D – P – S – R indicator concept (the socio – economic *drivers* cause the *pressure*, which changes the *state* and calls for *response*).

The support of these basic concepts reflected in the creation of various eco – efficiency indicators, which are tracking and reporting energy, waste and water parameters required for the definition of sustainability [2]. Such indicators have been introduced in the economics and there are efforts to find appropriate eco – efficiency indicators for the social life. However, this metrics leads to univariate estimates of the real environmental problems.

The natural environment is, in deed, multivariate complex system and its quality assessment related to sustainability requires multivariate approaches and metrics. The capability of the chemometrics and the environmetrics to handle multivariate systems and objects has helped for many environmental studies [3-6] to reach correct data classification, modeling and interpretation. Thus, chemometrics turns to be a very effective tool for problem solving and decision-making. It is the aim of the present study to illustrate some of the multivariate solutions offered by chemometrics when applied to environmental studies.

Experimental

Sampling site and sampling procedure

The sampling site Arnoldstein is located in the Austrian province Carynthia at height 564 m a.s.l. (altitude 46^{o} 33'31'' and longitude 13^{o} 42'12''). The site is located near to the border with Slovenia. At a distance (southwest direction) of nearly 1.5 km an industrial region is active with factories for wastes recycling (oil and fat emulsions, aluminium slags, rails, masts etc.), for polymer production and a lead smelter is still active

The aerosol data collection was gathered in the period between March 1999 and February 2000. The sampling was performed by the use of a high – volume sampler (Digitel DHA-80), which is a completely automated device described in details elsewhere [7]. The aerosol particles of the class PM_{10} are collected on a daily basis on quartz fiber filters (QAT-UP, Pallflex, USA) allowing in this way determination of the carbon content. The complete description of the sampling devise and the pre-sampling preparation of the filters could be found in [7].

Analytical procedures

The particle total mass was determined by weighing of the sampling filters before and after sampling according to CEN standard [8].

The determination of the water-soluble ions (cations: sodium, ammonium, potassium, magnesium and calcium; anions: chloride, nitrate, sulfate) was performed by the use of two ion-chromatographic systems after extraction of the filters by deionised water in ultrasonic bath for 20 min.

The concentration of the heavy metals was determined by the use of atomic absorption spectrometry. One quarter of the filter was cut by a ceramic scissor and the sample was weighted and extracted with 10 mL 10 % HNO_3. The detailed analytical procedures are described elsewhere [9,10].

The analytical procedure for determination of carbon (total carbon, TC, black carbon, BC and organic carbon, OC) used the developments of the well-established approaches of Puxbaum [9] for sample burning in oxygen atmosphere (TC), optical determination (BC) and the difference between TC and BC for OC determination.

Chemometrical methods

In the data treatment approaches of the environmetrics both unsupervised and supervised techniques are used.

Cluster analysis is a well-known and widely used classification approach for environmetrical purposes with its hierarchical and non-hierarchical algorithms [3,5].

In order to cluster objects characterized by a set of variables (e.g. sampling sites by chemical concentrations or pollutants), one has to determine their similarity. To avoid influence of the data size, a preliminary step of data scaling is necessary (e.g. autoscaling or z – transform, range scaling, logarithmic transformation) where normalized dimensionless numbers replaces the real data values. Thus, even serious differences in absolute (concentration) values are reduced to close numbers. Then, the similarity (or more strictly, the distance) between the objects in the variable space can be determined. Very often the Euclidean distance (ordinary, weighted, standardized) is used for clustering purposes. Thus, from the input matrix (raw data) a similarity matrix is calculated. There is a wide variability of hierarchical algorithms but the typical ones include the single linkage, the complete linkage and the average linkage methods. The representation of the results of the cluster analysis is performed either by a tree-like scheme called dendrogram comprising a hierarchical structure (large groups are divided into small ones) or by tables containing different possible clusterings.

Principal components analysis (PCA) is a typical display method, which allows estimating the internal relations in the data set. There are different variants of PCA but basically, their common feature is that they produce linear combination of the original columns in the data matrix (data set) responsible for the description of the variables characterizing the objects of observation. These linear combinations represent a type of abstract measurements (factors, principal components) being better descriptors of the data structure (data pattern) than the original (chemical or physical) measurements. Usually, the new abstract variables are called latent factors and they differ from the original ones named manifest variables. It is a common finding that just a few of the latent variables account for a large part of the data set variation. Thus, the data structure in a reduced space can be observed and studied [5].

According to the theory of PCA the scores on the PCs (the new co-ordinates of the data space) are a weighted sum of the original variables (e.g. chemical concentrations):

Score (value of object *I* along a PC *p*) = $\gamma_{1p} Y_1 + \gamma_{2p} Y_2 + \ldots + \gamma_{kp} Y_k$

where Y is indication of the variable value (e.g. concentration) and γ are the weights (called loadings). The information hidden in the loadings can also be displayed in loadings plots. It is important to note that PCA requires very often scaling of the input raw data to eliminate dependence on the scale of the original values.

Multiple regression on principal components (apportioning models) is a very important environmetric approach [11]. It makes it possible to apportion the contribution of each identified by PCA latent factor (emission source) to the total mass (concentration) of a certain chemical variable. The first step is performance of PCA, identification of latent factors, then determination of the absolute principal components scores (APCS) and multiple regression of the total mass (dependent variable) on the APCSs (independent variables).

Results and Discussions

The data was treated by the use of the STATISTICA 6.0 package.

The data clustering with respect to determine possible relationships between the variables (chemical components) gives following significant clusters (data matrix of 113 objects or sampling days and 20 variables):

C1: **K^+, Cu, Zn, As, BC, OC**
C2: **NO_3^-, NH_4^+, SO_4^{2-}**
C3: **Cd, Pb, Ni, V**
C4: **Cr, Fe, Mn**
C5: **Mg^{2+}, Cl^-, Na^+**
C6: **Ca^{2+}**

The cluster analysis followed the Ward's method of linkage and squared Euclidean distance as similarity measure. The cluster significance was determined by separation at distances 1/3 D_{max} and 2/3 D_{max}.

The linkage of the chemical variables into 6 clusters is an indication about the complex character of the pollution emitters in the region.

In order to obtain information about the data structure and identify latent factors responsible for it the data collection was treated by principal components analysis (Varimax rotation, scree plot validation and Malinowski's test for significance of the factor loadings). In Table 1 the factor loadings for 6 principal components, which explain nearly 80 % of the total variance of the system are presented.

Table 1. Factor loadings (*Marked loadings are statistically significant)*

	PC1	PC2	PC3	PC4	PC5	PC6
Cl^-	**0.669**	-0.028	0.072	0.043	0.613	-0.224
NO_3^-	0.241	**0.824**	0.013	0.032	0.119	-0.152
SO_4^{2-}	0.097	**0.606**	0.178	0.044	-0.069	0.648
Na^+	0.216	0.087	0.305	0.045	**0.852**	0.058
NH_4^+	0.166	**0.901**	0.075	-0.009	0.036	0.282
K^+	**0.886**	0.051	0.072	-0.153	0.102	0.050
Ca^{2+}	-0.064	0.045	0.020	0.148	-0.027	**0.901**
Mg^{2+}	0.019	0.139	-0.030	0.421	**0.765**	-0.077
As	**0.509**	0.454	0.426	0.308	0.168	0.056
Cd	0.128	-0.122	**0.769**	0.014	-0.105	0.024
Cr	0.060	0.022	0.004	**0.884**	-0.006	-0.028
Cu	**0.734**	-0.098	0.238	0.253	0.065	0.342
Fe	0.083	-0.020	0.016	**0.744**	0.232	0.137
Mn	0.416	0.190	0.284	**0.657**	0.216	0.185
Ni	0.200	0.465	**0.561**	0.248	0.317	-0.292
Pb	0.097	0.164	**0.769**	0.034	0.266	0.048
V	-0.035	0.547	**0.658**	0.013	0.220	0.277
Zn	**0.720**	0.344	0.254	0.226	-0.032	-0.134
BC	**0.696**	0.351	0.039	0.190	0.311	-0.412
OC	**0.830**	0.378	-0.074	0.215	0.065	-0.052
Expl.Var.	20.6%	15.4%	12.5%	11.8%	10.9%	9.6%

It is seen that six latent factors determine the data structure. These factors are related to the existing emission sources in the region and are conditionally named “combustion” (explaining nearly 21 % of the variance) “secondary emission (about 15 % explanation), “lead wastes” (with 12.5% of total variance), “vehicles” (with almost 12 %), “salt” (with 10.9 %), and the last “mineral dust” factor.

The first latent factor reveals the high loadings for potassium, copper, zinc, black and organic carbon and could be identified as a factor originating from the combustion products of the waste recycling plant. The second factor indicates the contribution of secondary emissions (high loadings for ammonium sulfate and nitrate) to the local aerosol formation. The high factor loadings of cadmium and lead in the third latent factor demonstrate the impact of lead smelter treating lead batteries. The regional traffic load is marked by latent factor. The contribution of the earth crust and the soil is shown by the high loadings of the mineral components sodium, magnesium, calcium and partially chloride in the last two latent factors being typical for the Alpine region.

In the next stage of the chemometric study a source apportioning procedure was applied [11], which allows determining the contribution (in %) of each identified source in the formation of the particle total mass or measured chemical concentrations. In Table 2 the regression models (regression using the absolute principal components scores) for each chemical parameter and for the total mass are presented. The intercept (intcpt) indicates the unexplained mass or concentration. The determination coefficient R^2 is a measure for the model validity.

Table 2. Source apportioning for the particle total mass (TM) and chemical concentrations for each identified latent factor (PC) in %. The regression analysis does not give significant results for contribution of PC5

	Intcpt	PC1	PC2	PC3	PC4	PC5	PC6	R^2
Cl^-		100.0						0.45
NO_3^-		20.2	79.8					0.73
SO_4^{2-}	17.4		32.2	9.2			41.2	0.81
Na^+	Non- adequate model							
NH_4^+	12.4	13.8	53.7				20.1	0.92
K^+		100.0						0.79
Ca^{2+}					25.2		74.5	0.83
Mg^{2+}	Non- adequate model							
As		33.7	22.2	20.1	24.0			0.74
Cd	21.3			78.7				0.59
Cr	31.2				68,8			0.77
Cu		49.9		10.4	21.1		18.7	0.77
Fe					100.0			0.55
Mn		20.9	6.3	9.9	55.3		7.4	0.76
Ni			33.5	40.1	26.4			0.58
Pb	49.5			50.5				0.64
V	15.4		30.6	35.6			18.5	0.81
Zn		51.9	17.1	11.9	19.1			0.74
BC		59.4	5.5		35.1			0.52
OC		63.0	5.2		31.8			0.71
TM		32.0	21.9	4.6	28.8		12.7	0.96

The highest part of the particle total mass is explained by the contribution of the combustion, traffic and secondary emission sources. The model shows a good validity (R^2 = 0.96).

Similarly, the contribution of the emission sources to the formation of the total concentrations of the chemical parameters can be found and estimated. No adequate models could be obtained for the apportioning of magnesium and sodium.

Conclusions

The multivariate statistical assessment of the air quality in the region of the site Arnoldstein, seems to be a good example how the multivariate approach to sustainability works. Monitoring and chemometrics are that combination, which considers the environmental system in its whole complexity. It is our deep conviction that chemometrics could make a lot about the concept of sustainability in all of its aspects – ecological, technological, economic and even social.

References

[1] Siemann W (2003) Umweltgeschichte, Verlag C.H. Beck, München

[2]. Eco-efficiency indicators workbook, (2003) http://www.nrtee.ca/publications/eco-efficiency_workbook/

[3] Einax JW, Zwanziger HW, Geiβ S (1997) Chemometrics in environmental analysis, VCH, Weinheim

[4] Simeonov V (2002) In: Encyclopedia of Environmetrics, Wiley, New York

[5] Massart DL, Vandeginste B. G. M, Buydens L M C, De Jong S, Lewi PJ, Smeyers-Verbeke J (1998) Handbook of chemometrics and qualimetrics; data handling in science and technology, parts A and B, Elsevier, Amsterdam

[6]. Hopke PK (1991) Receptor modeling for air quality management, Elsevier, NY

[7] Berner A (1978), *Chem Ing Techn* 50: 399

[8] CEN Norm – (1998) pr EN 1234

[9] Puxbaum H, Rendl J (1983), *Mikrochim Acta* I: 263.

[10] Hansen AD, Rosen H, Novakov T (1984), *The Sci Total Envir* 36: 191.

[11] Thurston GD, Spengler JD (1985), *Atmos Environ* 19: 9.