ECOLOGICAL CHEMISTRY AND ENGINEERING

Vol. 14, No. 7 2007

Vasil SIMEONOV*

RECEPTOR MODELING OF AIR CONTAMINANTS

MODELOWANIE RECEPTORÓW POLUTANTÓW POWIETRZA

Summary: The aim of the present paper is to discuss the options for studying the environmental contaminants by the use of chemometric strategies. The problem with source apportioning of environmental data is considered through the receptor modeling approach being very effective statistical tool in chemometrics. Data sets from two major regions (region of Krakow, Poland and several Austrian cities) are classified, apportioned and modeled by the use of cluster analysis, principal components analysis, self-organizing maps approach and chemical mass balancing modeling. It is shown that in many cases the balancing could be achieved by already classical methods like cluster and factor analysis but in other cases additional strategies contribute significantly to the adequate modeling and interpretation process. It is our deep conviction that a reliable source apportionment of environmental contaminants and correct estimation of the contribution of the pollution sources to the total mass can be expected if various strategies are involved to one and the same object of study. The methods mentioned are extremely important as information sources for risk assessment observations and for decision making procedures.

Keywords: chemometrics, environmental chemistry, data mining, source apportioning

Introduction

The careful monitoring of natural systems turns to be one of the most important tool not only for assessment of the real state of the human environment but also for taking political decisions, often with global impact. However, environmental data possess high variability due to various influences like geographical location of the monitoring sites, dynamic conditions in the atmosphere and hydrosphere, geological limitations, and many anthropogenic pollution sources. Usually, the output of the monitoring process is a large sheet of numbers indicating concentration levels at the sampling sites involved. Very often it is still generally accepted that satisfactory information could be extracted if the monitoring results are simply compared with allowable threshold values officially introduced by decision-making institutions. No need to say that this is an outdated approach and point of view. The monitoring data obtained should be considered in their integrity as a set of variables characterizing different environmental objects (natural

* Chair of Analytical Chemistry, Faculty of Chemistry, University of Sofia “St. Kl. Okhridski”, 1164 Sofia, J. Bourchier Blvd. 1, BG, email: VSimeonov@chem.uni-sofia.bg

water, air, and soil samples and sampling sites). As manual consideration of a large data set is practically impossible, the only reasonable way of data classification, projection, modeling, and reliable interpretation is the application of chemometric methods, of intelligent data and exploratory analysis [1–10].

The intelligent data analysis performed by chemometrics includes, in principle, two major approaches. In one of them the similarity between data parameters (*eg* between sampling regions or between chemical parameters) is object of consideration, and, in the other case, projection of multidimensional data on a simple plane is sought in order to visualize links and relations within the set of monitoring results. But much more important chemometric task seems to be the detection an internal balance of the environmental data, which reveals relationships between the monitoring results obtained and the really existing pollution sources and contaminants in a certain urban or rural environment. Actually, this is what the politicians and the decision makers really want – a simple and understandable presentation of the environmental balance: the level of pollution for a region and the contribution of the separate pollution sources to the formation of total pollution. Then is much easier to offer solution of the problem. The decision makers usually neglect the sophisticated theoretical considerations and it makes difficult to find acceptable problem solving.

The management of the water and air quality is not an easy problem. In general it involves the identification of the sources of materials emitted into the water or air, the quantitative estimation of the emission rates of the pollutants, the understanding of the transport of substances from the sources to certain locations (*eg* coastal regions for water phase or to downwind locations for the air), and the knowledge of physical and chemical transformation processes that can occur during that transport. All of those elements have to be put together into a chemometric model that can be used to estimate the changes in observable airborne or water concentrations that might be expected to happen if various actions are taken.

The aim of the present communication is to offer some simple schemes for risk assessment and pollutant balance of data from environmental monitoring of aerosols.

Theoretical considerations

The classical efforts over the past thirty years in the mathematical modeling of dispersion of pollutants in the environment have been significantly improved but there are still many instances when the models are unsatisfactory to permit the full development of effective and efficient environment quality management strategies. Obviously, other approaches are necessary to assist in the identification of pollutants and the apportionment of the observed pollutant concentrations in order to state that we are able to balance the processes of contamination, to look for responsibility or to prevent “hot spot” episodes. Such methods are often called receptor-oriented or receptor models since they are focused on the behavior of the ambient environment at the point of impact as opposed to the source-oriented dispersion models that focus on transport, dilution, and transformation that occur beginning at the source and following the pollutants to the sampling (receptor) site.

The modern variety of receptor modeling methods is dedicated mostly to aerosol particles analysis and there are convincing examples for their adequate use. All methods are divided into two major categories: with previously known sources and without known sources of pollution. For the first case the principle of mass balance is valid. It states that the mass conservation is at hand and a mass balance analysis can be used to identify and apportion sources of airborne particulate matted in the atmosphere. In order to carry out calculations with this mode of balancing, one need a preliminary constructed and measured set of really existing emission sources of pollution in the neighborhood. For the other mode of action no preliminary sources of pollution have to known in advance.

Balancing and apportionment without known source profiles

Very often it is difficult to obtain source profiles and it is inappropriate to use composition data from other locations. Thus, it becomes necessary to extract information on the sources from the ambient monitoring data. There are a number of factor analysis methods that can be used to identify and apportion pollutant sources from the environment. The basic factor analysis approach is the already classical chemometric method principal components analysis (PCA). However, PCA does not provide a direct balancing and apportionment. Alternative approaches able to do this include absolute principle components analysis (APCS), target transformation factor analysis (TTFA), positive matrix factorization (PMF), UNMIX [11–15]. Although all of these factor methods suffer from the problem of rotational degrees of freedom, each of them has its specificity. Recent studies indicate that variations of artificial neural networks can be used to provide information on the likely location and elemental source profiles for a given receptor site.

Despite of the different names given to several of the variety of forms of eigenvector analysis (factor analysis, PCA, principal components factor analysis, empirical orthogonal function analysis, Karhunen – Loeve transform, etc.), all the methods have the same basic objective – the compression of data into fewer dimensions and identification of the structure of interrelationships that exists between the variables measured or the cases in consideration. Thus, a new set of variables is introduced as linear combinations of the initial (real) variables so that the observed variation in the system can be reproduced by a smaller number of new (latent) variables. Since PCA can only be performed on a set of samples in which the various sources contribute different amounts of pollutant species to each sample, the balancing has to be expanded to a matrix equation of the type $\mathbf{Z} = \mathbf{A} \cdot \mathbf{F}$, where $\mathbf{Z}$ is the matrix of sample vectors, $\mathbf{A}$ is the matrix of loading vectors related to the source composition, and $\mathbf{F}$ is the matrix of scores that are related to the contribution of that source type to the variance of the measured variable.

After the pollution sources identification by the application of PCA, the next calculational step in modeling and balancing of pollution impacts is the apportioning itself. It is performed mostly by absolute principal components analysis (APCA). The procedure introduced by Thurston and Spengler [11] is well developed and often applied for apportionment purposes. The method estimates source profiles and contributions but a serious disadvantage is error propagation in centering and un-centering of data. This balancing approach accepts that all sources have been identified by the principal components analysis and all of them participate in the source contribution procedure. As we shall see in some of the case studies, the source identification by PCA is not always an easy and correctly solved problem.

Several interesting approaches for balancing environmental data have been developed by Paatero [13, 14]. One of them is called positive matrix factorization (PMF). Initially the problem was solved iteratively using alternating least squares. In this case, one of the matrices (known from PCA or factor analysis), A or F, is taken as known and the chi-squared is minimized with respect to the other matrix as a weighted linear least square problem. Then the roles of A and F are reversed so that the matrix that has just been calculated is fixed and the other is calculated by minimizing Q, the process then continues until convergence.

There are many evidences that artificial neural networks (ANN) are also used to look at the receptor modeling problem when the source profiles are not known. The self-organizing ANN method of Kohonen [16] has been presented for local scale problems with one sampling site and for multiple sampling sites. This method can analyse a three dimensional data block as a whole and yield both source profiles and geographical information on the identified emission sources. In the Results and Discussion section an application of the method of self-organizing maps (SOM) will illustrate the opportunities for pollution balancing interpretation for a large-scale case study.

Self-organizing maps (SOMs) are a data visualization technique invented by Professor Teuvo Kohonen, which reduce the dimensions of data through the use of self-organizing neural networks. The problem that data visualization attempts to solve is that humans simply cannot visualize high dimensional data as is so techniques are created to help us understand this high dimensional data. The way SOMs go about reducing dimensions is by producing a map of usually 1 or 2 dimensions, which plots the similarities of the data by grouping similar data items together. So SOMs accomplish two things, they reduce dimensions and display similarities.

The first part of a SOM is the data. The idea of the self-organizing maps is to project the n-dimensional data into something that can be better understood visually (it would be a 2 dimensional image map). The second component to SOMs are the weight vectors. Each weight vector has two components to them. The first part of a weight vector is its data. This is of the same dimensions as the sample vectors and the second part of a weight vector is its natural location.

The way that SOMs go about organizing themselves is by competeting for representation of the samples. Neurons are also allowed to change themselves by learning to become more like samples in hopes of winning the next competition. It is this selection and learning process that makes the weights organize themselves into a map representing similarities.

Balancing and apportionment with known source profiles

In principle, a mass balance equation can be written to account for all m chemical species in the n samples as contributions from p independent sources:

$$\mathbf{x_{ij}} = \Sigma\ \mathbf{c_{ik}} \cdot \mathbf{s_{kj}}\ \text{(for } k = 1 \text{ to } p\text{)}$$

where x_{ij} is the i-th elemental concentration measured in the j-th sample, c_{ik} is the gravimetric concentration of the i-th element in material form from the k-th source, and s_{ik} is the airborne mass concentration of material from k-th source contributing to the j-th sample.

There exist a set of natural physical constraints on the system that must be considered in developing any model for identifying and apportioning the sources of airborne particulate mass. The fundamental, natural physical constraints that must be obeyed are:

1. The original data must be reproduced by the model; the model must explain the observations.
2. The predicted source compositions must be non-negative; a source cannot have a negative percentage of an element.
3. The predicted source contributions to the aerosol must all be non-negative; a source cannot emit negative mass.
4. The sum of the predicted elemental mass contributions for each source must be less than or equal to total measured mass for each element; the whole is greater than or equal to the sum of its parts.

When modeling with known source profiles is used, the most common approach is, undoubted, the chemical mass balance (CMB) method [17].

The following set of linear equations expresses the essence of the CMB (it resembles entirely the idea of a mass balance model mentioned previously):

$$\mathbf{c_{ik}} = \Sigma\mathbf{a_{ij}}\ \mathbf{s_{jk}}\ \text{(for } j = 1 \text{ to } m\text{),}$$

where c_{ik}, the concentration of chemical species j in the particulate sample at receptor site k, equals the sum over m source types of the product of a_{ij}, the relative concentration of chemical constituent i in the fine particle emission from source j, multiplied by s_{jk}, the increment to total fine particulate mass concentration at receptor site k originating from source j. The system of equations states that the ambient concentration of each mass balance species must result only from the m sources included in the model and that no selective loss or gain of species i occurs in transport from the source to the receptor site. Therefore, the selection of mass balance compounds must be limited to:

1. Species for which all major sources are included in the model.
2. Species that do not undergo selective removal by chemical reaction or other mechanisms over the time scale for transport between the source and the receptor site.
3. Species, which are not significantly formed by chemical reactions in the atmosphere.

Source profiles are the mass abundances (fraction of total mass) of a chemical species in source emissions. They are intended to represent a category of source rather than individual emitters. The number and meaning of these categories is limited by the degree of similarity between the profiles. Mathematically, this similarity is termed "collinearity", which means that two or more of the CMB equations are redundant and the set of equations cannot be solved. Owing to measurement error, however, CMB equations are never completely collinear in a mathematical sense. When two or more source profiles are "collinear" in a CMB solution, standard errors on source contributions are often very high. Determining the degree of collinearity is one of the main objectives of CMB validation.

The generalized categories of the source profiles resemble the emission inventories in a certain location like "coal-burning" category, "vegetative burning and cooking" category, "diesel exhaust" category etc. For each of the chosen source profiles a series of measurements must be performed in order to collect the source profiles data set. The organization of the measurements, the chemical analysis of the species, and the error estimations is a quite complex experimental task.

Another very important part of the data preparation for the CMB modeling is the performance of the receptor measurements. They could be considered as a subset of the source profile measurements and must include at least those species in the source profiles that allow sources to be separated ("tracer" species).

After data collection one has to take into account the fundamental assumptions, potential deviations, and the validation options of the CMB procedure, namely:

1. The compositions of the source emissions have to be constant over the period of ambient and source sampling.
2. The chemical species do not react with each other, *ie* they act linearly.
3. All sources with a potential for significant contributing to the receptor have been identified and have had their emissions characterized.
4. The source compositions are linearly independent of each other.
5. The number of sources or source categories is less than or equal to the number of chemical species.
6. Measurement uncertainties are random, uncorrelated, and normally distributed.

All these assumptions are fairly restrictive and will never be totally complied with in actual practice. Fortunately, the CMB model can tolerate deviations from these assumptions, though these deviations increase the stated uncertainties of the source contribution estimates. Besides, there are a lot of optimization studies, which minimize the effect from the possible deviations.

The CMB estimates are finally tested to see how sensitive they are to the various input data. It has to be mentioned that the calculation procedure is offered as a software package by EPA (CMB 8.2 as last version).

Experimental

Data collection for the case studies

The apportioning and balancing problem considered above is illustrated by two major case studies – in Krakow area, Poland and in Vienna, Austria.

Integrated emissions, air quality and health impact – case study Krakow

In 2004 and 2005 a big research project concerning the air quality in the area of Krakow, Poland was launched sponsored by the European Commission. The organization of the project was coordinated by the Joint Research Center (JRC), Institute for Environment and Sustainability (IES), Transport and Air Quality Unit. Many Polish academic and governmental institutions as well as groups of researchers from many European countries and from US EPA participated to various extent in the integrated project with full title: "Krakow Integrated Project: Particulate Matter: From Emissions to Health Effects". Finally, a workshop in Krakow took place to present the outcome of the project.

In brief, the main goal of this project was in line with European Environment and Health and the Urban Environment thematic strategies to develop an integrated approach to assess the effects of toxic emissions, the resulting air quality and their impacts on human health. In a stepwise approach a methodology is being developed in a confined and well-characterized location. The city of Krakow and its surroundings comprise an area with typical emission sources suitable for a large-scale study. Coal in this region is still widely used in residential heating appliances.

Six sampling sites were chosen for data collection (PM10 particulate matter), conditionally named AGRI (rural site), INDU (industrial site), TRAF (traffic site), POLI (urban site), ZAKO (site in Zakopane, considered to be non-polluted by industrial sources), and HOUSE (in-door pollution measurements). For apportioning purposes altogether 85 cases were involved for different time periods. The chemical species determined in the particulate matter were quite a lot – soluble major ions, silica, aluminum, heavy metals, soot, organic carbon (mainly polyaromatic hydrocarbons or PAHs). A variety of analytical methods were applied to determine the species concentrations. The techniques of sampling, sample preparation, and chemical analysis are not the aims of the presentation and are only mentioned.

Case study Austrian Urban Air Quality

Since many years the group of atmospheric chemistry, Institute of Chemical Technologies and Analytics, Technical University of Vienna, headed by Prof. Hans Puxbaum is actively engaged in large-scale projects dealing with the Vienna air quality as well as the air quality in other large Austrian cities. These projects are financed and supported by the Vienna municipality and all local authorities. In all of the projects a chemometric data classification, modeling and interpretation is necessary done in almost all cases by our research group of chemometrics, Faculty of Chemistry, University of Sofia.

In the results section some cases will be presented from the project "Quellenanalyse PM10 Belastung". They illustrate the inevitable role of chemometrics in pollutants balancing.

In brief, particulate matter PM10 was collected from 10 urban sites in and around the city of Vienna or in Graz and Linz. Major (soluble) ions, heavy metals and carbonaceous content were analytically determined for a two-year period.

Results and discussion

Case study Krakow

The data available were treated initially by cluster analysis and PCA (in addition, the identified latent factors were included in the Thurston – Spengler apportioning procedure with APCS).

In Fig. 1 the hierarchical dendrogram for clustering (z-transformed input data, Ward's method of linkage, squared Euclidean distance as similarity measure) of the chemical species (variables) for all samples (sampling sites, the whole sampling period, sum of all PAH chemical compounds) is shown. It can be readily seen that two major clusters are formed: the one of them collects those species, which are known tracers for coal combustion like soot, organic carbon, bromine or for secondary emissions like ammonium, sulfate, nitrate; the second one includes typical soil and road dust markers like aluminium, titanium, silica, iron and calcium.

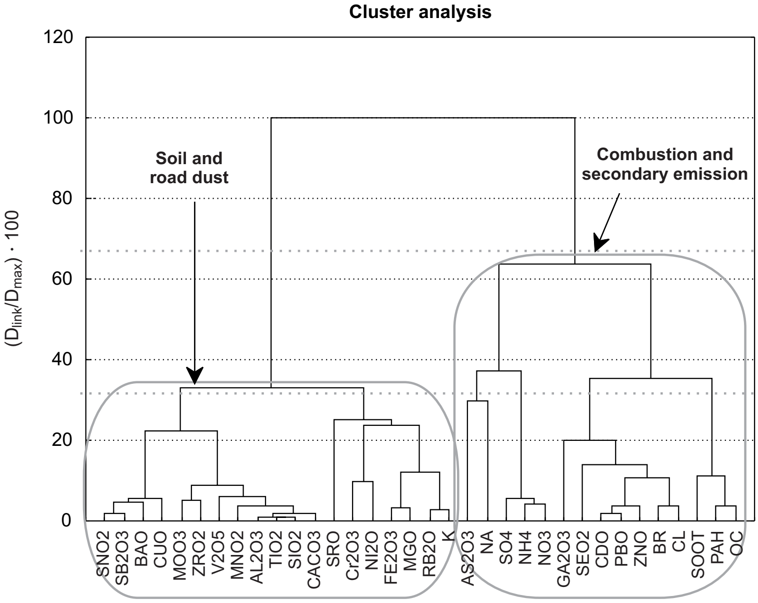


Fig. 1. Hierarchical dendrogram for all Krakow data (chemical variables linkage)

Using cluster classification we get the initial information about the possible major emitters in the region named by us as soil and road dust source and combustion and secondary emission source.

In the further identification and apportioning effort, PCA of the same data set was performed. Next three figures (Figs. 2–4) illustrate the physical meaning of the identified latent factors by indicating the significant factor loadings for each latent factor. It was found that three latent factors explain more than 85 % of the total variance of the system. As in the case with cluster analysis, we were able to give conditional names to the latent factors as follows: PC 1 – "soil and road dust"; PC 2 – "secondary emission" and PC 3 – "combustion by stationary and mobile sources".

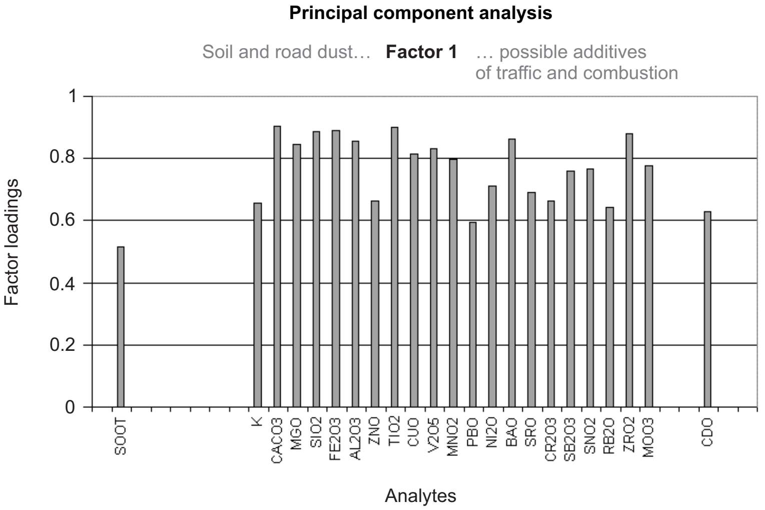


Fig. 2. Factor loadings plot for PC1 for all Krakow data

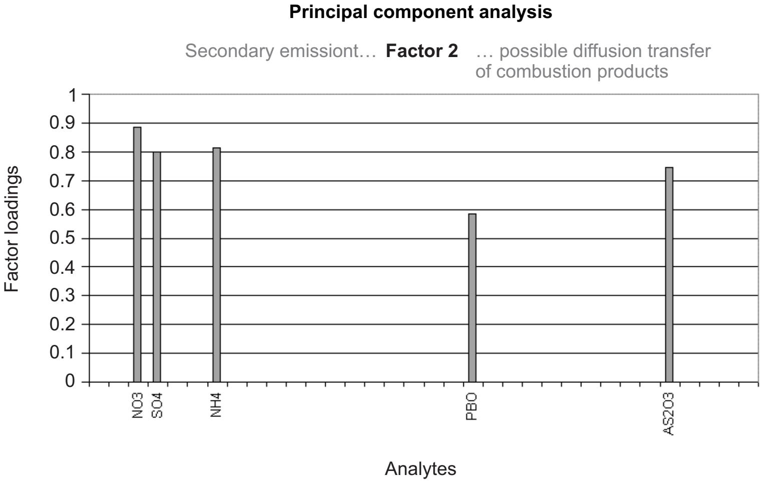


Fig. 3. Factor loadings plot for PC2 for all Krakow data

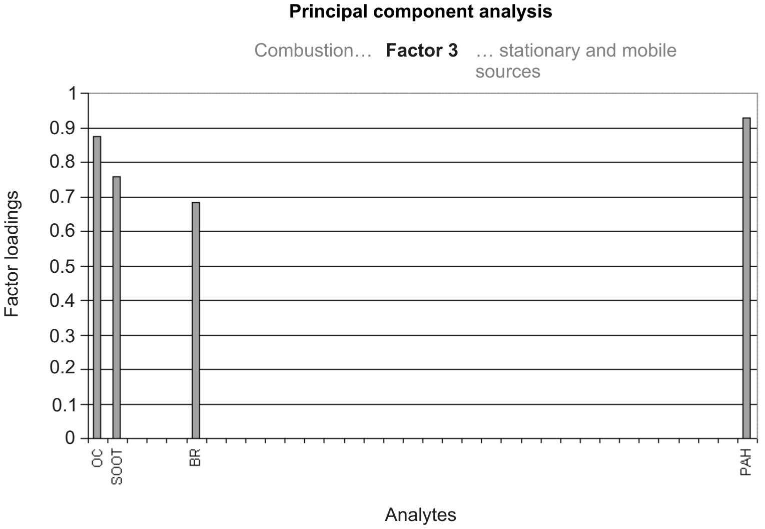


Fig. 4. Factor loadings plot for PC3 for all Krakow data

In this chemometric analysis we achieved a slightly better resolution of the impact of the possible local pollutants – three instead of two factors influence the balancing mechanism. However, the analysis of the latent factor structure indicates again, that PC1 is by no means only "soil and road dust" source but a more complex one with possible supplement of combustion or traffic sources to the dust one (tracers for combustion are obviously lead, cadmium, vanadium). Further, PC2 reveals not only secondary emissions impact but possible combustion products transport. Finally, PC3 shows traffic addition (part of PAHs are vehicle combustion tracers; the same holds true for soot) to the coal combustion factor, which is a dominant local pollutant.

The performance of the source apportionment by the Thurston – Spengler regression method using absolute principle components scores for the total mass (TM) of the PM10 aerosols gave the numbers shown in Fig. 5.

The balancing procedure indicates that the highest contribution to the aerosol total mass has the secondary emission source (61%), followed by the combustion source

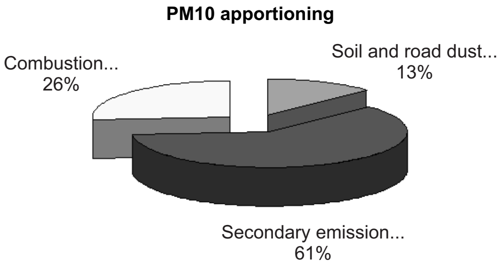


Fig. 5. Source contributions to the total mass according to Thurston – Spengler method for all Krakow data

(26%) and the lowest contribution goes to the soil and road dust factor (13%). It has to be stressed that this is a logical balancing keeping in mind that, for instance, coal combustion emitters are involved in secondary emission and in the soil dust latent factor. Thus, the coal combustion proves to be the main pollutant in the whole Krakow area.

In order to get specific information from the different sampling sites a self--organizing map (SOM) data classification was performed. It has to be reminded that the general idea of the SOM method is to map the data from a higher dimensional space onto a lower, usually, two-dimensional (2D) space. The latter consists of i nodes arranged in the 2D plane as neighboring hexagons or squares. The mapping preserves the topology of the original data space. In this way the 2D plane (called U matrix) resembles the space reduction and the cluster abilities of SOM. The topological neighborhood concept (taking into account distances between the nodes) as well as the mapping of the original data onto a grid of nodes allows better visualization and further interpretation.

Classification of sites by SOM

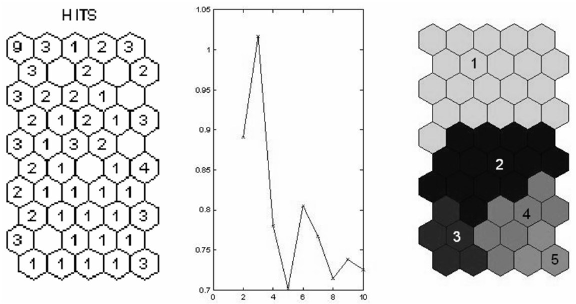


Fig. 6. SOM classification (unit hits and Davies-Boulduin index) for all Krakow data

The map unit hits (distribution of cases in each sampling site) in each map node are presented in Fig. 7. The map unit hits and the component planes visualize clearly the relation between objects and variables, as it will be commented later. The unified distance matrix (U matrix) represents also the distances between the units. The high values indicate a cluster border; uniform areas or low values indicate the clusters themselves. The U matrix could be also used for classification of the sampling sites. This approach lacks the drawback of the classical clustering, namely once the object is linked to a certain cluster, it remains linked for the rest of the clustering process. A training algorithm constructs the nodes in SOM in order to represent the whole data set and their weights are optimized at each iteration step. Thus, the optimal topology is guaranteed. In our study the non-hierarchical K-means classification algorithm was applied. The different values of k (predefined number of clusters) were tried and the sum of squares for each run was calculated. Finally, the best classification with the lowest Davies-Bouldin index (also shown graphically in Fig. 6) is chosen. It is seen that five clusters configuration has the lowest index.

The SOM classification obtained was then compared with the real data from the starting data set (all sites, all measurements). The results are marked in Figs. 7–11 for each sampling site. For better interpretation the real concentration of tracer species are given along with the classification results. Thus, a unexpectedly reasonable interpretation of some of the pollution events and sources is achieved.

Altogether 85 observations for 5 sampling sites were available (site AGRI – 15, site INDU – 15, site POLI – 15, site TRAF – 15, site ZAKO – 13, sites HOUSE for indoor pollutants – 12). Their SOM classification has indicated as follows:

– *Group 1* (*cluster 1*): it consists of totally 44 cases (the number for each site is presented in Fig. 7). Almost all indoor cases (11 out of 12) belong to this cluster. A very high similarity between in- and outdoor measurements is observed (33 outdoor cases out of all 73 are included in Group 1). The concentration levels of

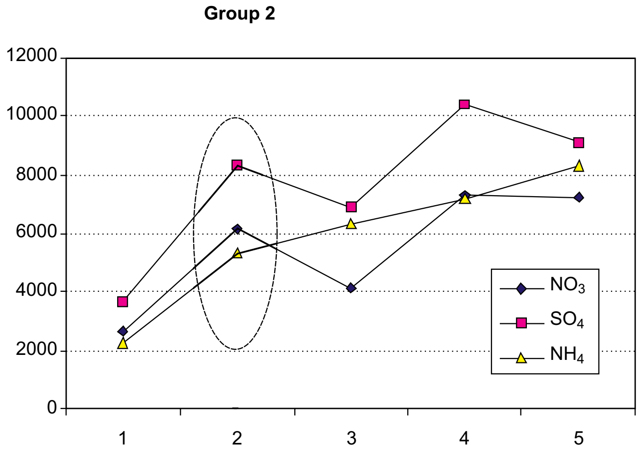
Group 1

AGRI	9
INDU	7
POLI	7
TRAF	5
ZAKOPANE	5
HOUSE	11
TOTAL	44

Almost all indoor cases (11 out of 12).
High level of similarity between indoor and outdoor cases for significant numbers of observation (33 out of all outdoor 74 measurements).
The lowest concentrations for all species are observed.
Probably this is the **"Background group"** for the sampling period and region.

Fig. 7. Interpretation of SOM Group 1 by the input data values

AGRI	3
INDU	4
POLI	4
TRAF	5
ZAKOPANE	3
HOUSE	1
TOTAL	**20**



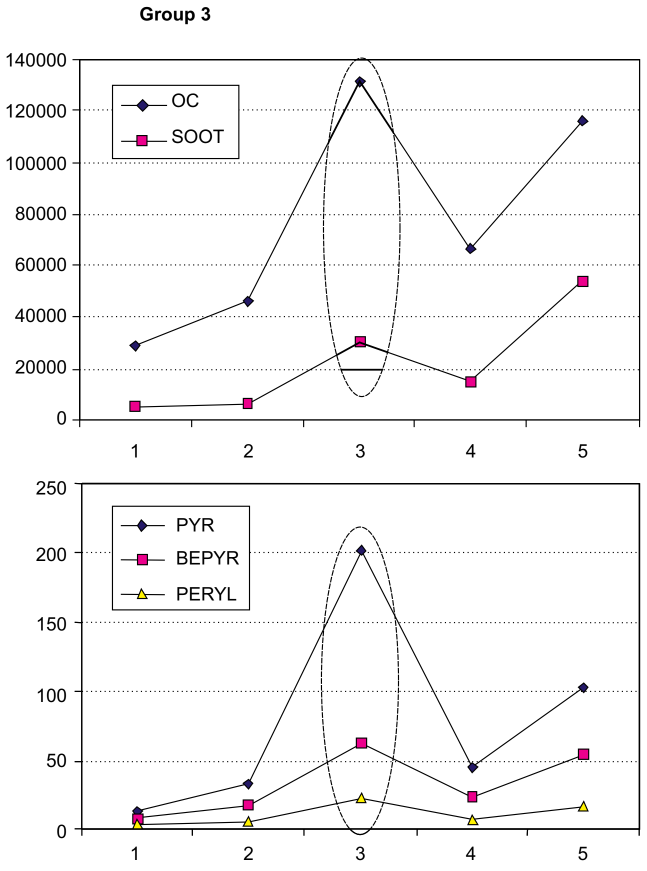
The dominating time period is 29.01–04.02.06 (15 out 20).
The possible reason for its formation – the North Atlantic air transport.
Low concentrations for PAH and dust but increased concentrations for secondary aerosol mass.

Fig. 8. Interpretation of SOM Group 2 by the input data values

polluting species for all cases, which belong to this group, are the lowest in the set. Probably, this group could be conditionally named "background cluster" as it a subject of non-polluting impacts;

– *Group 2* (*cluster 2*): totally 20 cases are included in the cluster of similarity (again the numbers of cases from each site is indicated in Fig. 8). It is characterized by low concentrations of PAHs but increased concentrations of secondary aerosols, especially for the time period between 29.01. and 4.02. 2006 (15 cases out of all 20 fall into this period). Therefore, this is a typical case for "hot spot" event with dominating role of the "secondary emission" source of pollution due probably to specific meteorological reasons (low ambient temperature and north-west wind direction);
– *Group 3* (*cluster*): it consists of a small number of cases, only 7 in total (Fig. 9), five of them from the ZAKO (Zakopane) site. Obviously, the situation resembles a strictly local event characterized with enhanced concentrations of carbonaceous species (elemental and organic carbon). If the sampling period for the cases into

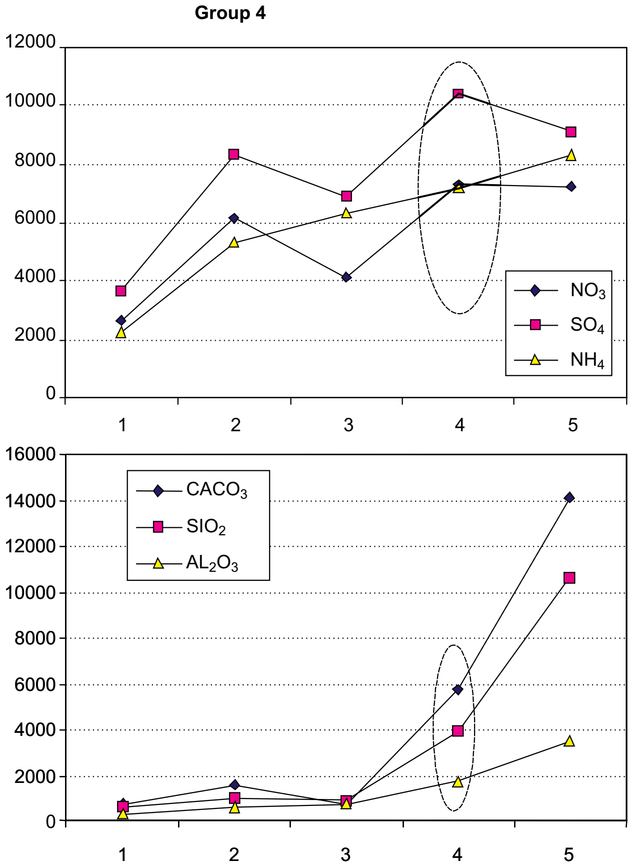
AGRI	0
INDU	1
POLI	1
TRAF	0
ZAKOPANE	5
HOUSE	0
TOTAL	**7**



This is the small group dominantly by Zakopane cases.
Probably these are regional events.
Highest concentration of carbonaceous particles.
Probably due to increased residential combustion

Fig. 9. Interpretation of SOM Group 3 by the input data values

AGRI	2
INDU	2
POLI	1
TRAF	4
ZAKOPANE	0
HOUSE	0
TOTAL	**9**



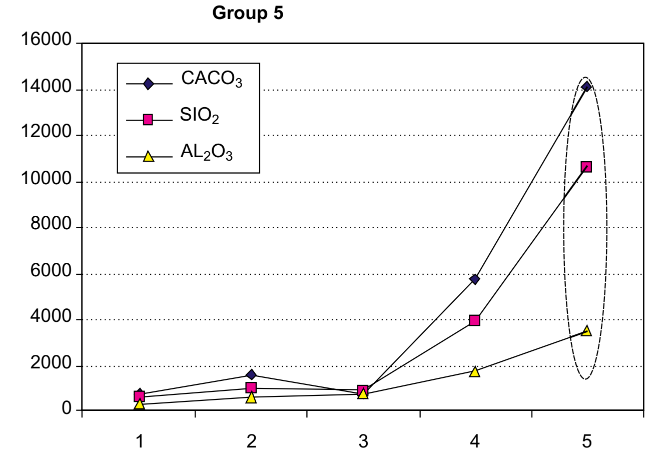
Little group with regional impact due mainly to secondary emission and local traffic and dust.

Fig. 10. Interpretation of SOM Group 4 by the input data values

consideration (15–17.01. and 5–6.02.2006) is compared with the ambient temperature, it could be found that these are the days with very low temperatures. The “coal combustion” source is probably responsible for the events due to increased appliance of coal for domestic heating in Zakopane;

- *Group 4* (*cluster 4*): nine cases are involved (exact numbers for each case are given in Fig. 10) and this is a typical “urban” or “anthropogenic” group related to the combined polluting impact of secondary aerosol emissions, local traffic, and mineral dust (indications for this balancing are the increased concentrations of secondary emission species (ammonium, sulfate, nitrate) and dust tracers like calcium, silica and aluminum;
- *Group 5* (*cluster 5*): this is a second “urban” cluster, which comprises only urban sites (Fig. 11). Highest concentrations of dust components are observed within a small sampling period (16–17.01.2006). It is probably a strictly local event due to air re-circulation, which is partially confirmed by meteorological data and back trajectory analysis.

AGRI	1
INDU	1
POLI	2
TRAF	1
ZAKOPANE	0
HOUSE	0
TOTAL	**5**



Small group of cases (16–17.01.06) with regional impact due to possible air re-circulation.
Highest concentration of dust sources

Fig. 11. Interpretation of SOM Group 5 by the input data values

Finally, chemical mass balancing with previously know emission source profiles was performed on the Krakow air-monitoring data. In the first step a list of source profiles was considered. The possible major sources in the Krakow are over 20 (among them iron ore sinter plant, blast furnace, cement kiln – coal fired, coal combustion power plant, coke gas, coal combustion, commercial boilers with coal combustion, residential coal combustion, residential wood combustion etc.). In addition to own measurements of the chemical composition of the source profiles, literature sources were adapted (secondary ammonium, sulfate, nitrate; road salt, road pavements, tire debris, rock and crust material, and vehicle/tires/brakes composite). A multitude of combinations of source profiles were subjected to CMB calculations including the profiles that represent the emissions most likely to influence receptor concentrations. Profiles of similar chemical composition were often found to be collinear and automatically rejected by the CMB 8.2 software package. Five sources remained robustly significant in all the simulations and were finally retained for the source apportionment:

1. Coal fired small residential stoves or boilers;
2. Secondary emissions (ammonium, sulfate, nitrate);
3. Vehicle/traffic source;
4. Re-suspension of road dust (rock and/or pavement combined with road salt);
5. Coal fired small boilers.

It could be readily seen that the source profiles chosen for CMB calculations correspond in general with the sources identified by cluster analysis, PCA, and SOM. It is, however expected that CMB will give a better resolution of the apportionment both by a more reasonable source collection (and identification) and by the introduction of more tracers.

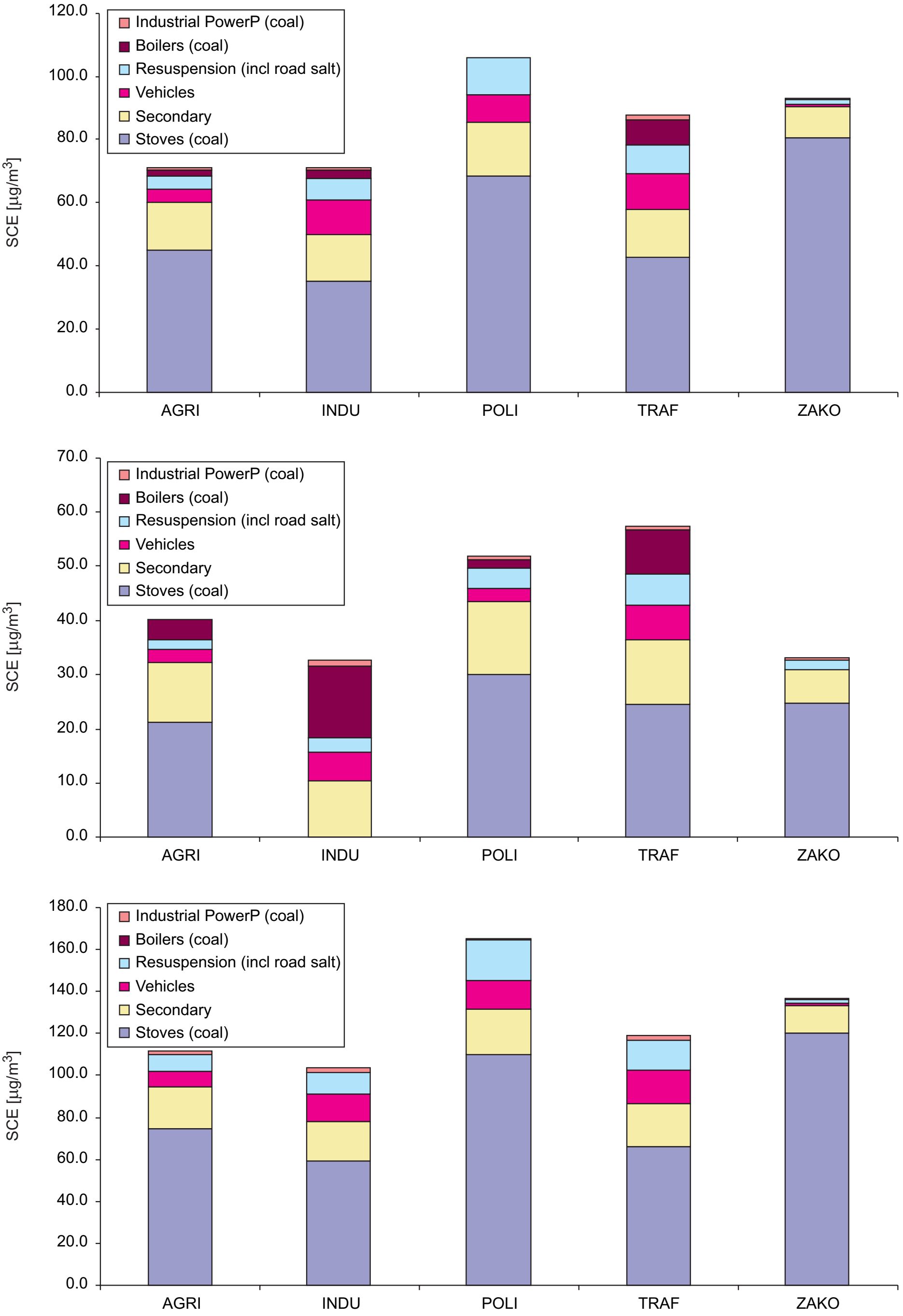


Fig. 12. Source contribution estimates by CMB to the total aerosol mass at sites AGRI, INDU, POLI, TRAF, and ZAKO

In Fig. 12 the source contribution estimates by CMB to the total aerosol mass at all five sampling sites from the Krakow region except for the in-door samples (AGRI, INDU, POLI, TRAF, and ZAKO) are given.

Various monitoring events are presented in the figure-averages for all days for one of the sampling session; average for days considered as “clean” (TM less than 70 μg/m^3); average for episode days being subject to polluting events (TM higher than 70 μg/m^3). In all situations the dominant polluting source is the residential coal combustion in stoves and boilers. Secondary emissions contribute also significantly to the total mass for all events. The same conclusion holds true for daily source contribution estimates. The traffic and re-suspension contributions complete the balancing since the contribution of industrial pollution is quite low.

Case study Austria

The problem and its solution does not seem quite different for the second case study, which involves more than 10 sampling sites in the three largest Austrian cities – Vienna, Graz and Linz. The principles of sampling and aerosol analysis do not differ substantially from that for Krakow. The sampling sites in the three cities are mostly urban and no organic components are included in the chemical analysis procedure. Thus, except major soluble ions and heavy metals, the input data sets offered soot, organic carbon, and carbonate.

Monthly averages were used for the apportioning procedures. Most of the sites are from Vienna, one site is from Graz, and one – from Linz. The sampling allowed separation of the aerosols in two categories PM 10 (coarse fraction) and PM 2.5 (fine fraction). The data set were subject to cluster analysis, principal components analysis and chemical mass balance modeling. Since the volume of the work done is too big to be presented in full scale, only several chosen examples will illustrate the main goal of the case study – to offer a good apportioning strategy as required by the European commission.

In the previous case study the main features of the chemometric approaches informing on polluting sources in a certain region and their contributions to the formation of the aerosol total mass were discussed. In this second case study we would like to stress on some possible problems in the apportioning procedure. As illustrative examples the PM10 balancing in Vienna (site AKH, typical urban site, downtown) and in Linz (seriously polluted industrial region) were chosen.

In Fig. 13 the hierarchical dendrogram (Ward’s method of linkage, z-transformed input data, squared Euclidean distance as similarity measure) as graphical projection of the cluster analysis for data from AKH site (PM10) is presented (linkage of chemical variables).

Four clusters are formed but it is not simple to interpret the relations between the chemical species. For instance, sulfate is clustered along with chromium and oxalate in a well-defined group. The common origin as a prerequisite for the linkage is doubtful. The same holds true for the other clusters where some of the links are logical from

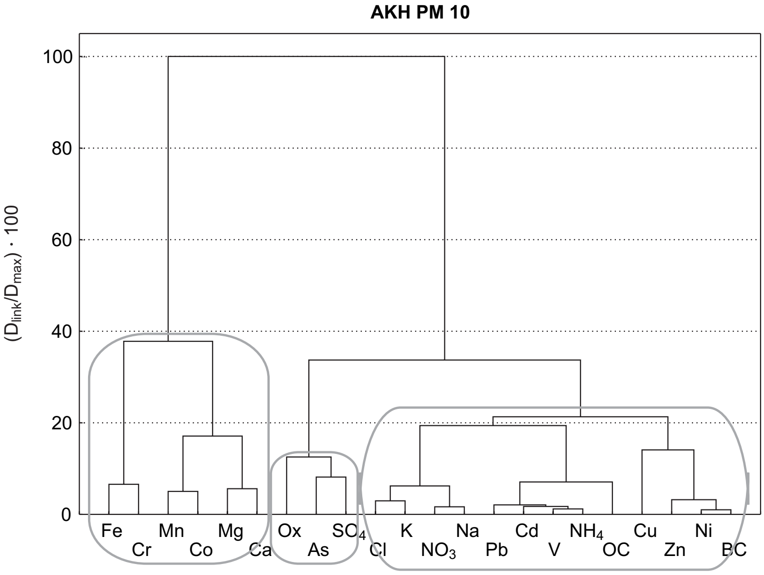


Fig. 13. Hierarchical dendrogram for all AKH site data (chemical variables linkage)

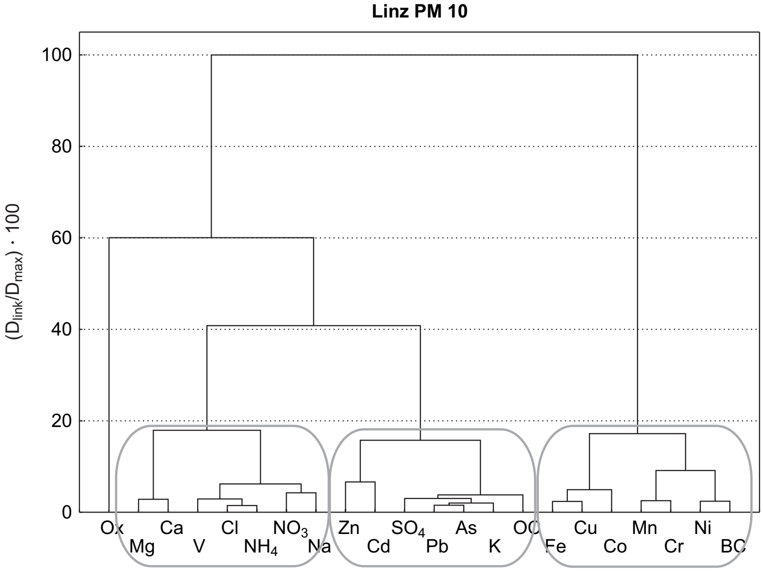


Fig. 14. Hierarchical dendrogram for Linz site data (chemical variables linkage)

environmetric point of view (ammonium, nitrate), others – not (potassium, carbonaceous species).

Next figure (Fig. 14) shows the dendrogram (same method of clustering) of the Linz site.

The three clusters formed contain chemical species whose common origin (natural or anthropogenic source) is quite unbalanced (organic carbon linkage with some heavy metals for instance).

The next step in searching for balance and apportionment was the performance of PCA. In Tables 1 and 2 the factor loadings for data sets from AKH and Linz are given. The statistically significant loadings are marked. The Varimax rotation mode on scaled data was applied in the chemometric analysis. The number of latent factors was traditionally chosen by the scree-plot view, by the values of the total variance explained by the latent factors and by the condition for eigenvalue for the significant factors to be higher than 1.

Again, we get very complex latent factors and their correct interpretation is difficult. The possible identification of the latent factors is presented under the tables and it is readily seen that they are of mixed origin. If one accepts that PC1 for site AKH is combination of traffic and coal combustion polluting sources, then the real apportionment will be handicapped since no division between traffic impact and coal combustion impact will be possible. The same holds true for the other identified latent factors.

Table 1

AKH PM10 Factor loadings

Species	PC1	PC2	PC3	PC4
BC	**0.906**	0.196	0.192	0.186
OC	**0.654**	0.369	0.342	–0.254
NA	**0.698**	0.133	0.569	–0.062
NH_4	0.620	**0.720**	0.054	–0.252
K	0.476	**0.544**	0.499	–0.293
CA	0.103	–0.226	**0.633**	0.606
MG	0.1782	0.0640	**0.883**	0.335
CL	**0.687**	0.140	0.565	–0.130
NO_3	**0.774**	0.224	0.483	–0.139
SO_4	0.065	**0.964**	–0.135	–0.002
AS	0.275	**0.734**	0.397	0.110
CD	**0.608**	0.677	0.018	–0.171
CO	0.251	–0.140	0.089	**0.901**
CR	–0.260	0.032	**–0.876**	0.324
CU	0.361	0.388	–0.141	**0.604**
FE	–0.277	–0.369	–0.436	**0.628**
MN	–0.282	0.016	0.022	**0.897**
NI	**0.878**	0.107	0.272	0.151
PB	**0.707**	0.645	–0.133	–0.141
V	**0.7291**	0.650	0.172	0.001
ZN	**0.865**	0.365	0.095	0.197
Expl. Var. %	32.8	20.7	17.9	16.1

The four latent factors could be identified as follows:

- PC1: (32.8%): BC, OC, Na, Cl, NO_3, Ni, Pb, V, Zn, Cd – Combination of traffic and coal combustion sources??
- PC2: (20.7%): NH_4, SO_4, As, K – Secondary emission source??
- PC3: (17.9%): Mg, Ca, Cr – Road dust source ??
- PC4: (16.1%): Mn, Co, Fe , Cu – Mineral dust source??

Absolutely same conclusions could be offered with the site from Linz. The case is even worse and the identification of the polluting sources impossible. This is obviously due to the strongly polluted atmosphere in the region of Linz.

Table 2

LINZ PM10 Factor loadings

Species	PC1	PC2	PC3
BC	**0.613**	0.606	0.439
OC	**0.818**	0.412	0.2941
NA	0.504	0.135	**0.776**
NH_4	**0.707**	0.181	0.654
K	**0.919**	0.084	0.283
CA	0.147	0.346	**0.881**
MG	0.3686	0.522	**0.716**
CL	**0.706**	0.308	0.601
NO_3	0.464	0.338	**0.761**
SO_4	**0.877**	0.043	0.4297
AS	**0.862**	0.251	0.399
CD	**0.839**	0.106	0.316
CO	–0.056	**0.922**	0.276
CR	0.143	**0.897**	0.156
CU	0.310	**0.866**	0.090
FE	0.153	**0.955**	0.123
MN	0.294	**0.823**	0.403
NI	0.500	**0.778**	0.355
PB	**0.878**	0.330	0.265
V	**0.729**	0.431	0.472
ZN	**0.868**	0.176	–0.074
Expl. Var. %	39.4	29.5	22.7

Three latent factors are found:

- PC1: (39.4%): BC, OC, NH_4, K, Cl, SO_4, As, Cd, Pb, V, Zn – Traffic, combustion, secondary emissions??
- PC2: (29.5%): Co, Cr, Cu, Fe, Mn, Ni – Undefined??
- PC3: (22.7%): Na, Ca, Mg, NO_3 – Road dust ??

The only possible decision in this case is the application of CMB with preliminary known source profiles. In this case study it was not possible to carry out experiments for determination of the local pollution source profiles and they were chosen from literature source, mainly from USA. A list of source profiles offered by US EPA (Speciate 3.2 data bank) was thouroughly tested and compared to local conditions. Finally, 8 source profiles were chosen (steel production, crustal dust, Diesel combustion, paved road dust, petrol combustion, brake debries, coke combustion, wood combustion) for CMB modeling. The profiles were checked for collinearity and only non-correlated sources were included in the calculation. The calculational results are summarized and averaged for quarters.

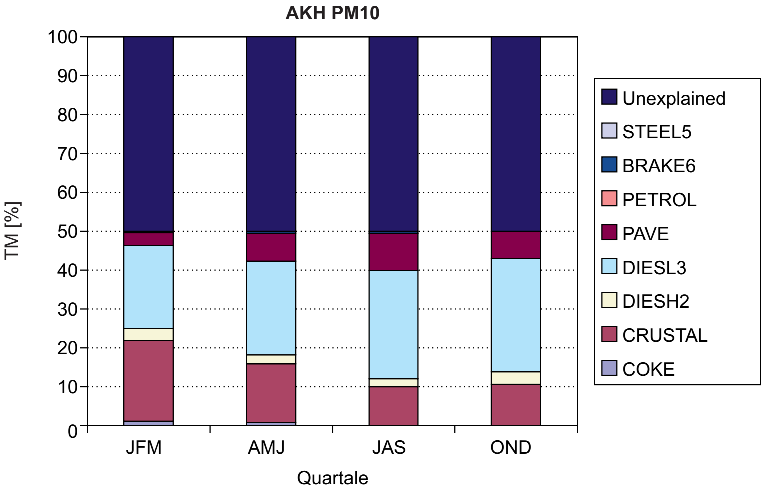


Fig. 15. Source contribution estimates by CMB to the total aerosol mass (by quarters) for AKH site

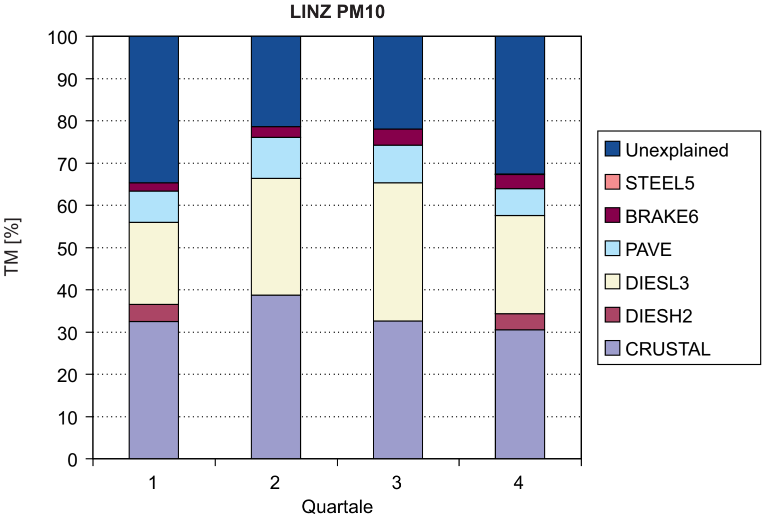


Fig. 16. Source contribution estimates by CMB to the total aerosol mass (by quarters) for Linz site

Next two figures (Figs. 15 and 16) show the percent contribution of each source to the total aerosol mass at the two sites.

For the Vienna site (Fig. 15) paved road and crustal dust are the main contributors to the total mass both in summer and winter periods. The traffic source (diesel combustion vehicles) is also important polluting source. A substantial part of the total mass remains unexplained since no measurements of silica, aluminum and other major components are done. The models miss also the role of secondary emissions but they can be easily attributed to the unexplained part.

For the Linz site we detect the same major contributors but, additionally, a steel production source influences the total mass apportioning. The unexplained part of the total mass includes the unavoidable secondary emissions impact.

Thus, despite the fact that the construction of the source profiles was to some extent artificial, the CMB modeling made the source apportioning very realistic and corresponding to the local environmental conditions.

Conclusion

The chemometric approaches and strategies as applied to monitoring data sets deliver very abundant and specific information about details in the environment, which usually remain hidden behind the figures of real concentrations and allowable thresholds. The present study discloses some better options for problem solving and decision making in source apportionment and balancing of environmental pollutants. The combination of already classical chemometric approaches like cluster and factor analysis with more advanced strategies like SOM or receptor modeling with previously known pollution source profiles by CMB gives a complete idea of the contribution of pollutions to the "hot spots" events in a region and the impact of each separate source to the total pollution.

Acknowledgement

The author would like to express his sincere gratitude to Dr. Stefan Tsakovski, University of Sofia, Prof. Dr. Hans Puxbaum, Vienna Technical University, and Dr. Bo Larsen, JRC-IES, Italy for the help in preparation of this work.

References

[1] Simeonova P., Bogoeva L., Kashukeeva K. and Dimitrova D.: Ann. Univ. Sofia Fac. Chem., 2006, **98/99**, 223–231.

[2] Simeonova P. and Lovchinov V.: J. Optoelec. Adv. Mater., 2005, **7**, 419–423.

[3] Simeonova P.: Ecol. Chem. Eng., 2006, **13**, 1021–1032.

[4] Simeonova P., Simeonov V. and Andreev G.: Centr. Europ. J. Chem., 2003, **2**, 121–136.

[5] Samara C., Kouimtzis T., Tsitoridou R., Kanias G. and Simeonov V.: Atmos. Environ., 2003, **37**, 41–54.

[6] Simeonov V., Tsakovski S., Lavric T., Simeonova P. and Puxbaum H.: Microchim. Acta, 2004, **148**, 293–298.

[7] Stanimirova I. and Simeonov V.: Chemom. Intell. Lab Syst., 2005, **77**, 115–121.

[8] Simeonova P.: Ann. Univ. Sofia Fac. Chem., 2007, **100**, (in press).

[9] Simeonova P., Sarbu C., Spanos Th., Simeonov V. and Tsakovski S.: Centr. Europ. J. Chem., 2006, **4**, 68–80.

[10] Simeonov V.: *Environmetric Strategies to Classify, Interpret and Model Risk Assessment and Quality of Environmental Systems*, [in:] Technological Choices for Sustainability, Sikdar S., Glavic P. and Jain R. (eds.), Springer, Berlin–Heidelberg 2004, pp. 147–164.

[11] Thurston G. and Spengler J.: Atmos. Environ. 1985, **19**, 9–26.

[12] Hopke P.: *The Mixture Resolution Problem Applied to Airborne Particle Source Apportionment*, [in:] Chemometrics in Environmental Chemistry, Einax J. (ed.), Springer Verlag, Heidelberg 1995, pp. 47–86.

[13] Paatero P. and Tapper U.: Chemom. Intell. Lab. Syst., 1993, **18**, 183–194.

[14] Anttila P., Paatero P., Tapper U. and Jarvinen O.: Atmos. Environ. 1995, **29**, 1705–1709.

[15] Hopke P. (ed.): Receptor Modeling for Air Quality Management, Elsevier Science, Amsterdam 1991.

[16] Kohonen T.: Self-Organizing and Associative Memory, 3rd Edition, Springer Verlag, New York 1989.

[17] Watson J., Chow J. and Pace T.: *Chemical Mass Balance*, [in:] Receptor Modeling for Air Quality Management, Hopke P. (ed.), Elsevier Science Publishers, Amsterdam 1991, pp. 83–116.

MODELOWANIE RECEPTORÓW POLUTANTÓW POWIETRZA

Streszczenie

Przedyskutowano możliwości zastosowania metod chemometrycznych do badań zanieczyszczenia środowiska z wykorzystaniem strategii chemometrycznych. Problem określenia źródła danych środowiskowych jest rozważony za pomocą modelowania receptora, będącego bardzo efektywnym narzędziem statystycznym w chemometrii. Dane pomiarowe dotyczące dwóch regionów (miasta Krakowa, Polska i kilku miast austriackich) zostały opracowane za pomocą analizy skupień, analizy głównych składowych, map samoorganizujących i bilansu mas chemicznych. Wykazano, że w wielu przypadkach bilansowanie mogłoby być dokonane klasycznymi metodami, jak skupienia i analiza czynnikowa, ale w innych przypadkach dodatkowe metody przyczyniają się znacząco do odpowiedniego modelowania i interpretacji wyników. Zastosowanie różnych metod do analizy wyników badań pozwali na określenie środowiskowych źródeł zanieczyszczeń i poprawne obliczenie wkładu tych źródeł w całkowitą masę zanieczyszczeń. Wspomniane metody są ważne jako źródło informacji przy ocenie ryzyka i przy opracowywaniu procedur decyzyjnych.

Słowa kluczowe: chemomometria, chemia środowiska, eksploracja danych, określanie źródeł