W19 State of the art in methods and software for the identification, resolution and apportionment of contamination sources

Organized by Romà Tauler[1], Pentti Paatero[2] and Philip Hopke[3] with the assistance of Ronald C. Henry[4], Cliff Spiegelman[5], Eun Sug Park[6], and Richard L. Poirot[7]

[1].Department of Environmental Chemistry, IIQAB-CSIC, Jordi Girona 18-26, Barcelona 08034, Spain, e-mail rtaqam@iiqab.csic.es

[2] Department of Physical Sciences, University of Helsinki, Box 64, FIN-00014. e-mail: Pentti.Paatero@helsinki.fi

[3] Center for Air Resources Engineering and Science, Clarkson University, Box 5708, Potsdam, NY 13699-5708, e-mail hopkepk@clarkson.edu

[4] Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA 90089-0231, e-mail rhenry@usc.edu

[5] Department of Statistics, Texas A&M University, College Station, TX 77843-3143, e-mail cliff@stat.tamu.edu

[6] System Planning, Policy, and Environmental Research Group, CE/TTI Building,3135 TAMU, College Station, TX 77843-3135, e-mail e-park@tamu.edu

[7] Department of Environmental Conservation, Vermont Agency of Natural Resources, Waterbury, VT 05671, e-mail: rich.poirot@dec.anr.state.vt.us

Abstract: Current approaches and recent developments and software related with multivariate factor analysis and related methods in the analysis of environmental data for the identification, resolution and apportionment of contamination sources are discussed and compared

Keywords*: Environmental Modeling, Factor Analysis related methods, Identification and Resolution of Contamination Sources, Source Apportionment*,

1. INTRODUCTION

The main purpose of an introduction is to enable the paper to be understood without undue reference to other sources. It should therefore have sufficient background material for this purpose. Generally, highly specialized papers will not need an extensive introduction as interested readers may be expected to be familiar with current literature on the subject. On the other hand, when a paper is likely to interest people working in fields outside the immediate area of the paper, the introduction should contain background material which could otherwise be scattered throughout the literature.

Environmental systems (air, water, soils, biota...) are very complex systems and it is necessary to obtain simplified descriptions of them in order to produce mathematical models capable of being calculated on current computer technologies. Thus, although significant improvements have been made over the recent years in the mathematical modeling of transport, dilution, transformation, diffusion and dispersion of contaminants in the environment, there are still many cases where these models (usually based on the solution of large differential equation systems) are insufficient to allow full development of effective and efficient environmental quality management strategies. Moreover, operating these models in an appropriate way requires a detailed knowledge and control of a large number of parameters, which makes this approach unrealistic in many practical situations. Thus, it is necessary to have other approaches available to assist in the identification of contamination sources, in the determination of

their distribution (geographical, temporal, among environmental compartments,...) and in their apportionment at a particular sampling point.

Environmental monitoring studies often produce huge amounts of measured physical parameters and chemical concentrations evaluated at distant geographical sites and during different time periods. Moreover, these parameters and chemical concentrations are also estimated at different environmental compartments (i.e. air, water, sediments, biota...). All these data sets are difficult to handle and evaluate in a simple and fast way using simple univariate statistical and modeling tools, especially due to their large size and to their multicomponent and multivariate nature. In order to discover relevant patterns and sources of variation in these large environmental data sets, the application of modern chemometric methods based in statistical multivariate data analysis and in factor analysis is proposed. The basic assumption of these methods is that each of the parameters or chemical concentrations measured in a particular sample are mostly affected by different contributions coming from independent sources. By using these methods, specific point sources and diffuse area sources of contaminants in the environment and their origin (natural, anthropogenic, industrial, agricultural...) can be identified and their relative distribution among samples (geographical, temporal, among different environmental compartments distributions) can be evaluated. At each sampling site, relative source quantitative apportionment is estimated allowing an assessment of their environmental impact, distribution and time evolution.

2. DATA SETS

Environmental data sets are usually organized in data tables or data matrices, corresponding to one sampling time period or environmental compartment of the monitoring campaign. Rows of these data matrices identify the investigated samples (e.g. different sampling sites) and columns identify the measured variables (physical parameters, concentrations of chemical contaminants or other environmental parameters). Variables having very few values above the measurement detection limit should be removed before multivariate data analysis is applied. When a particular chemical compound is not detected, its concentration value may be set equal to half its detection limit (Fharnham et al., 2002). For missing values, imputation methods have been proposed (Walczak et al., 2001) and whenever they are a small fraction of the measured values, they may be estimated without loosing the data structure needed for application of multivariate data analysis tools. Last, but not least important is the data weighting problem. A critical aspect to consider is data uncertainties. It has been shown (Paatero, 1997) that traditional scaling and autoscaling weighting schemes based in the variables data variance are in many cases problematic, because they may overestimate the influence of variables with low signal to noise ratios. A statistically sounder and more rigorous approach is based in the use of data uncertainties and in their inclusion in the definition of the objective function to minimize. A more controversial topic is the discussion of what to do when these uncertainties are not available (Paatero, 2003).

3. MODELS AND METHODS

The fundamental equation describing the general bilinear model used to solve the factor analysis problem is stated as follows:

$$x_{ij} = \sum_{n=1}^{N} g_{in} f_{nj} + e_{ij} \qquad \text{Equation 1}$$

In Equation 1, x_{ij} refers to measured variable j (physical parameter or chemical concentration) in sample i; f_{nj} refers to the contribution of variable j to source profile n; g_{in} refers to the contribution of this source n to sample i;, and. e_{ij} gives the unmodeled part of x_{ij} considering a total number of N environmental sources, hopefully equals only to experimental and instrumental noise if all sources of physical-chemical variation are captured by the model. Therefore Equation 1 assumes that the measured parameters or concentrations x_{ij}, (apart from noise) are a weighed (scores, g_{in}) sum of a reduced number (N<< I or J number of samples or number of variables) of contributions from distinct environmental sources. Written in matrix form the same bilinear Equation (Equation 2) is written

$$\mathbf{X} = \mathbf{G}\mathbf{F}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equation 2}$$

where now **X** is the matrix of all measurements (j=1,...,J variables in i=1,...,I samples) **G** is the matrix of score profiles (distribution of the N contamination sources among samples), $\mathbf{F}^{\mathbf{T}}$ is the matrix of loading profiles (composition of the N composition sources) and **E** is the noise or error matrix containing the variance not explained by the model defined by the N environmental sources described in **G** and **F.** Since only **X** is known initially, the matrix decomposition described by Equations 1 and 2 is ambiguous (not unique) unless constraints are applied.

3.1 Principal Component Analysis and Factor Analysis derived methods

One first approach to solve the bilinear model of Equations 2 is Principal Component Analysis (Jolliffe, 2002). In this approach, matrix factorization or decomposition of Equation 2 is performed under orthogonal constraints for both **G** and $\mathbf{F}^T$. Moreover, loadings (rows of $\mathbf{F}^T$ matrix) are also normalized (i.e. this matrix becomes orthonormal) and forced to be in the direction of explaining maximum variance. Under such constraints, PCA provides unique solutions and interpretation of variance is straightforward since scores and loadings are orthogonal (not overlapped). Using a small number of principal components a considerable amount of data variance is usually captured since many of the analyzed variables are correlated. Therefore, interpretation and visualization of main features and trends of the data set under study, i.e. of main contamination sources, are readily available from score and loading plots. However and due to precisely to PCA mathematical constraints, solutions may be useful for variance interpretation but they do not have a direct physical interpretation. PCA decomposition does not estimate the 'true' underlying (latent) sources of data variance but a linear combination of them fulfilling orthogonal constraints. This means that although these solutions have good mathematical properties, they do not have a physical meaning. For instance, both **G** and $\mathbf{F}^T$ will have negative values and uncorrelated profiles, whereas expected profiles for 'true' environmental sources defined by **G** and $\mathbf{F}^T$ should not have these profiles negative and they may be also strongly correlated. Moreover source apportionment (quantitative assessments of source contributions at each sample) cannot be performed due to the applied constraints also.

The problem related with the extraction of non-negative profiles, improving interpretation and allowing source apportionment has been addressed in different ways. For instance rotation of PCA factor matrices to simplify interpretation like in varimax orthogonal rotation, scores uncentering (to make them positive) and regression to total sample mass has been proposed in the alternative approach called Absolute Principal Component Analysis (APCA, Thurston et al., 1985). However, when source impacts are low, negative values in scores are difficult to handle and produce undesirable results. Alternatively, several methods derived from some kind of Target Factor Analysis have been proposed, like Confirmatory Factor Analysis (Christensen et al, 2002), which tries to use efficiently previous knowledge available about the nature of the investigated source profiles. However, use of these approaches is in general limited because of the limited number of known point source profiles (for instance in atmosphere contamination some profiles like crustal, combustion, vehicle-traffic, soil,... profiles). The problem is even more difficult when diffusion contamination sources are also involved as it is the general case in environmental studies.

3.2 Alternatives to PCA based methods

New approaches have been proposed in the recent years to solve the factor analysis problem previously stated in Equations 2 and 3. As described below, these methods place restrictions on the possible source profiles defined in **G** and $\mathbf{F}^T$, to require them to met certain physical constraints (e-g. non-negative source impacts and composition) instead of purely based mathematical constraints like orthogonality or variance independency. In this presentation several of these methods will be discussed and compared.

Unmix

The Unmix model has been developed for the US EPA (Henry et al., 1990, 2003) and has several unique features. Unmix has an advanced computationally intensive algorithm to estimate the number of sources than can be seen above the noise level in the data. Given this estimated number of sources, Unmix uses PCA to reduce the dimensionality of the data space. Geometrical concepts of self-modeling curve resolution are used to ensure the results obey (to within error) non-negativity constraints on source compositions and contributions. This is, however, not sufficient to uniquely determine the source compositions and contributions (see also below multivariate curve resolution method). Additional constraints determined from the data are needed. These are estimated by looking for the edges in the data determined by points where one source is small compared to other sources. Other features of Unmix are its ability to handle missing data (Henry et al. 1999), so often encountered in environmental monitoring studies, and the ability to gather estimates of uncertainties in the source compositions. Version 4, the latest version of Unmix is available from Dr. Gary Norris, norris.gary@epa.gov). This version includes identification of influential data points and variables that can be excluded from the analysis, and automatic selection of the best models. Running time has been dramatically reduced by giving the model a "memory" of previous solutions based on a method that uses the duality between sources and source contributions demonstrated in Henry (2005).

Positive Matrix Factorization (PMF) and Multilinear Engine (ME)

Whereas PCA based methods and Unmix are essentially based on eigenvector analysis, which in fact can be also considered as a least-squares analysis using a particular set of constraints and minimizing the sum of squared residuals for the model described by Equations 1 and 2, Positive Matrix Factorization (PMF, Paatero, 1997) takes a very different approach to the same factor analysis problem. PCA and related methods usually scale or normalize data and this scaling will lead to distortions in the analysis. In fact the optimal scaling of the data would be to scale each data point individually so as to have the more precise data having more influence on the solution than points that have higher uncertainties. PMF takes the approach of an explicit least squares approach in which the method minimizes the objective function Q:

$$Q = \sum_{i=1}^{I} \sum_{j=1}^{J} \left| \frac{x_{ij} - \sum_{n=1}^{N} g_{in} f_{nj}}{s_{ij}} \right|^2 \qquad \text{Equation 3}$$

where s_{ij} are estimates of the uncertainties in the jth variable measured in the ith sample. The factor analysis problem is to minimize Q(E) with respect to **G** and $\mathbf{F}^{\mathbf{T}}$ with the constraint that each of the elements of these two matrices are to be non-negative.

Over the last past years different algorithms have been developed and applied to solve the PMF problem (Paatero, 1997, Polissar et al. 1998; Ramadan et al. 2003), and more recently and alternative approach has been proposed that provides a more flexible modeling system, the multilinear engine (ME, Paatero, 1999), with several expansions to handle different type of problems. One of these extensions takes into account modeling source contributions using multifactor physical and meteorological effects (such as wind direction and speed, day/week/season variations, precipitation, and so on, Paatero, 2002). Also ME can be easily handle even more complex models related with multiset and multiway data set arrangements, like the trilinear model for three-way data analysis (Hopke, 1998, Yakovleva, 1999).

Recently, ME has been used to in exposure assessments to examine the sources of particles that are joint among different kinds of samples. For example, Hopke (2003) examine data from multiple environments (outdoor, indoor, apartments, and people) around a residential facility for elderly inhabitants using a model that includes factors that contribute to all four types of samples (external factors) and factors that only contribute to the indoor, apartment and personal samples (internal factors). Similar models have been applied to an exposure panel study in the Raleigh-Chapel Hill, NC area (Zhao, 2006a).

It has also been used to develop a complex spatial model that examined the distribution of particle mass across the eastern United States (Paatero, 2003). The factor analytic model was enhanced by modeling the dependence of $PM_{2.5}$ concentrations on temperature, humidity, pressure, ozone concentrations, and wind velocity vectors. The model comprises 12 general factors across the spatial domain, augmented by 5 urban-only factors intended to represent excess concentration present in urban locations only. The computed factor components or concentration fields are displayed as concentration maps, one for each factor, showing how much each factor contributes to the average concentration at each location. The factors are also displayed as flux maps that illustrate the spatial movement of $PM_{2.5}$ aerosol, thus enabling one to pinpoint potential source areas of $PM_{2.5}$.

Multivariate Curve Resolution Alternating Least Squares (MCR-ALS*)*

Another possible complementary and/or alternative method to perform PCA bilinear matrix decomposition given in Equations 1 and 2 is Multivariate Curve Resolution (MCR, Tauler et al., 1995a; Tauler, 1995b). This method was initially developed to investigate evolving processes of multicomponent systems by means of spectroscopic methods. However, it may be easily extended to investigate environmental sources in the analysis of large monitoring data tables (Salou et al., 1997, Tauler et al. 2000, Tauler et al. 2004) and also to resolve component profiles in mixture analysis problems in general (de Juan et al. 2003) In MCR methods, loadings and scores are not constrained to be orthogonal like in PCA, but to fulfill a particular set of physical constraints like non-negativity, normalization, unimodality (single peak shaped profiles), closure (mass-balance), selectivity, local rank, shape (Gaussian, Lorentzian...) and hard-modeling (equilibrium, kinetic or any other physical or chemical law). All these constraints may be introduced in alternating least squares (ALS) algorithms (Tauler et al., 1995a; Tauler 1995b; de Juan et al. 2003; Jaumot et al., 2005) in an optional and flexible way. The goal of MCR-ALS when applied to environmental data tables is to investigate how contamination sources really are in physical terms (loadings) and how they are distributed among samples (scores). However, since only matrix **X** is known and only soft constraints like non-negativity, profile normalization and/or mass-balance (receptor

models), are in general applied, unique solutions are not guaranteed in MCR-ALS in general and rotational and intensity ambiguities may persist (Tauler et al., 1995a). A method to evaluate these effects after MCR-ALS resolution and how to calculate maximum and minimum band boundaries of the set of feasible solutions (Tauler, 2001) and resampling (Jaumot et al., 2004) error intervals have been proposed. A new approach taking into account uncertainities in measured data and using Total Least Squares (Van Huffel et al., 1991) is under development.

Other Statistical Approaches

Spiegelman and Dattner [1993a, 1993b] developed an algorithm for selecting species to use in a receptor models as well as a linear programming approach to fitting the model. Recently, there has been a series of work by statisticians often jointly with environmental engineers to provide the estimates having good statistical properties in multivariate receptor modeling [Park et al. 2001; Park et al. 2002 a&b; Christensen and Sain 2002; Park et al. 2004; Gajewski and Spiegelman 2004]. In Park et al. (2001), a time series extension of multivariate receptor modeling was developed to account for temporal dependence in air pollution data into estimation of source compositions and uncertainty estimation. A different approach for dealing with temporal dependence was suggested by Christensen and Sain (2002). Park et al. (2002 a) proposed new sets of realistic identifiability conditions for the parameters in Equations 1 and 2 and developed the Constrained Nonlinear Least Squares (CNLS) estimators for the parameters. A Bayesian approach that can handle the unknown number of pollution sources and unknown identifiability conditions simultaneously with estimation of model parameters has also been developed (Park et al. 2002b and Park et al. 2004). The method computes the marginal likelihoods and/or the posterior probabilities using a computational technique known as the Markov chain Monte Carlo (MCMC) for a range of plausible models (rather than a single model) selected by varying the number of sources and identifiability conditions. Gajewski and Spiegelman (2004) developed estimators that are robust to outliers.

Multiway data Analysis

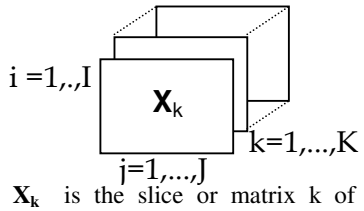
The factor analysis bilinear model shown in Equations 1 and 2 can be extended to the simultaneous analysis of multiple data sets using data matrix augmentation. Thus, bilinear methods like PCA, Unimix, PMF, ME and MCR-ALS can be easily adapted to multiset and multiway data sets by matrix augmentation or cube unfolding (matricizing). More involved trilinear and multilinear models have been also proposed for three-way and multi-way data arrangements in the investigation of environmental contamination sources. In particular trilinear models for three-way data are described by equations:

$$x_{ijk} = \sum_{n=1}^{N} g_{in} f_{jn} z_{kn} + e_{ijk} \qquad \text{Equation 4}$$

in element-wise way, or in matrix way for each individual matrix or cube slice,

$$X_k = G Z_k F^T + E_k \qquad \text{Equation 5}$$

In Equation 4, x_{ijk} is the measured physical parameter or concentration of component j at sample i under condition k. There are three ways, directions, orders or modes of measurement. These three modes indicate that component j was analyzed at sample i at a particular situation or condition k, usually time or environmental compartment (water, sediment or biota). The whole data set can be organized in a data 'cube' or parallelepiped as shown in the Figure



where $\mathbf{X_k}$ is the slice or matrix k of the data parallelepiped, which is modelled by Equation 5, and Z_k is a diagonal matrix. This trilinear model described by Equations 4 and 5 is also called the PARAFAC model (Bro, 1997; Smilde et al. 2004). In the trilinear model, all slices in the three-way data set are decomposed using the same **G** (scores) and $\mathbf{F^T}$ (loadings), differing only in their relative amounts expressed in the different $\mathbf{Z_k}$ diagonal matrices. Trilinear models, and by extension multilinear models, provide unique decompositions and they are the natural extension of bilinear models. They are useful for data exploration and interpretation. However, they assume that there is no system variation since they impose equal scores and loading profiles (same shape) for all data matrices simultaneously analyzed, and therefore, they are in many circumstances, too rigid, and do not allow the resolution of the 'true' underlying sources of data variation, simply because the data do not behave like in the postulated trilinear models. More flexible models including Tucker models have been proposed and used in this context (Smilde et al., 2004) A compromise between 'softer' bilinear models and 'harder' trilinear and multilinear models should be

considered in practice according to the data structure encountered for a particular data set. There are also multiway data sets coming from systems that provide time and size resolved constituents (e.g. particles) that require a different model to resolve them. There are approaches that can be used to resolve such data as well.

3.3 Other Related Techniques

Factor analysis models cannot provide full resolution of the specific contributions of sources with similar composition. For example, it is common to see a source profile dominated by one component (e.g. sulfate) that is ascribed to particular emission sources (e.g. coal-fired power plants). However, in order to identify the likely locations of such sources, methods that include the transport need to be included. It is possible to examine the influence of local sources using the wind directions measured during the sampling periods. Several methods, non-parametric regression and conditional probability function analysis, are available to identify specific local sources. When the transport is from longer distances, the flow can be characterized by air parcel back trajectories that estimate where the fluxes (e.g. air) were located prior to its arrival at the sampling site. The information from the trajectories can be incorporated directly into the factor analysis or the factor analysis results can be used as input to a second set of models that use them and the back trajectories to infer likely origins of the pollutants. Several models available to use air parcel back trajectories include potential source contribution function (PSCF) and residence time analysis (RTA).

Nonparametric Regression Methods

Local sources of airborne pollutants have been identified by nonparametric regression of hourly concentrations of primary pollutants versus wind direction and speed (Henry et al., 2002; Yu et al. 2004). Also known as kernel smoothing, nonparametric regression does not make any assumptions as to the functional form of the relationship between the predictor and predicted variables. Even fundamental assumptions such as mass conservation are not required. Nonparametric regression can determine the direction of a local source from the monitoring site with unprecedented accuracy. Using the wind speed, the approximate distance to the source can be estimated as well.

Conditional Probability Function

The conditional probability function (CPF) (Ashbaugh, 1985) analyzes point source impacts from varying wind directions using the source contribution estimates from PMF coupled with the wind direction values measured on site (Kim, 2003). The CPF estimates the probability that a given source contribution from a given wind direction will exceed a predetermined threshold criterion. The same daily contribution was assigned to each hour of a given day to match to the hourly wind data. The CPF is defined as

$$CPF_{\Delta\theta} = \frac{m_{\Delta\theta}}{n_{\Delta\theta}} \qquad \text{Equation 6}$$

where $m_{\Delta\theta}$ is the number of occurrence from wind sector $\Delta\theta$ that exceeded the threshold criterion, and $n_{\Delta\theta}$ is the total number of data from the same wind sector. In this study, 24 sectors were used ($\Delta\theta$ = 15 degrees). Calm wind (< 1 m/sec) periods were excluded from this analysis due to the isotropic behavior of wind vane under calm winds. From tests with several different percentile of the fractional contribution from each source, a threshold criterion of the upper 25 percentile was chosen to define the directionality of the sources. The sources are likely to be located to the direction that have high conditional probability values.

Potential Source Contribution Function

The Potential Source Contribution Function (PSCF) receptor model was originally developed by Ashbaugh (1985). It has been applied in a series of studies over a variety of geographical scales. In a PSCF analysis, both chemical and meteorological data for each filter sample are needed. Air parcel back trajectories ending at a receptor site are calculated from the meteorological data with a trajectory model. Trajectories are represented by segment endpoints. Each endpoint has two coordinates (e.g., latitude, longitude) representing the central location of an air parcel at a particular time. To calculate the PSCF, the whole geographic region covered by the trajectories is divided into an array of grid cells whose size is dependent on the geographical scale of the problem so that the PSCF will be a function of locations as defined by the cell indices i and j.

Let N be the total number of trajectory segment endpoints during the whole study period, T. If n segment trajectory endpoints fall into the ij-th cell (represented by n_{ij}), the probability of this event, A_{ij}, is given by

$$P[A_{ij}] = \frac{n_{ij}}{N} \qquad \text{Equation 7}$$

where $P[A_{ij}]$ is a measure of the residence time of a randomly selected air parcel in the ij-th cell relative to the time period T.

Suppose in the same ij-th cell there is a subset of m_{ij} segment endpoints for which the corresponding trajectories arrive at a receptor site at the time when the measured concentrations are higher than a pre-specified criterion value. In this study, the criteria values were the calculated mean values for each species at each site. The probability of this high concentration event, B_{ij}, is given by $P[B_{ij}]$,

$$P[B_{ij}] = \frac{m_{ij}}{N} \qquad \text{Equation 8}$$

Like $P[A_{ij}]$ this subset probability is related to the residence time of air parcel in the ij-th cell but the probability B is for contaminated air parcels.

The potential source contribution function (PSCF) is defined as

$$PSCF_{ij} = \frac{P[B_{ij}]}{P[A_{ij}]} = \frac{m_{ij}}{n_{ij}} \qquad \text{Equation 9}$$

$PSCF_{ij}$ is the conditional probability that an air parcel which passed through the ij-th cell had a high concentration upon arrival at the trajectory endpoint. There are several problems with the PSCF analysis approach. Near the edge of the spatial domain of the back trajectories, there are relatively few trajectories in any given grid cell. In many of the studies (e.g., Zeng *et al.*, 1989; Cheng *et al.*, 1993a&b), an arbitrary weight function is used to reduce the values in cells with few endpoints.

Residence Time Analysis

An initial effort was made by Ashbaugh (1983) to make use of air parcel back trajectories to identify likely source locations for particulate sulfur observed at the Grand Canyon. A gridded array is created around the sampling location. Trajectories are a sequence of segments, each of which represents a fixed amount of time. Thus, each endpoint can be considered to be an indication that the air parcel has spent a given time within that grid cell. The total "residence time" that air spends in the given cell would be the total number of endpoints that fall into that cell. These values can be plotted over a map. The residence time values associated with high or low concentration can be plotted to examine likely directions from which contaminated or clean air is transported to the sampling site.

The problem with this method is that all of the trajectories begin at the receptor site and thus, the residence time is maximum in the cells surrounding the sampling location. Ashbaugh (1985) suggest one solution to this problem that will be described below. An alternative method which has come to be called Residence Time Analysis was developed by Poirot and Wishinski (1986). In their method, they first interpolate along each trajectory segment to estimate the fraction of time spent in each grid cell and then summing the residence time for that cell. They propose a method to adjust the resulting grid cell values for the geometrical problem of high values in the region immediately adjacent to the receptor site.

In the RTA approach, a variety of different metrics can be applied to the resultant counts of hours in the equal-area grid squares. One set of RTA metrics, referred to as "concentration-based sorting" begins with the conversion of the gridded trajectory hours to "probability fields" in which, for a given scenario of dates, the "upwind probability" of trajectory location in a given grid square is defined as the fraction of hours in that square compared to the total hours in all of the cell. An "everyday probability field" is calculated for a scenario of all sample days at the receptor, and provides an indication of areas most likely to be upwind of the receptor on a long-term or climatological basis. A "high day probability field" can be calculated for various definitions of "high" contributions at the receptor, for example upper 50^{th}, 75^{th}, or 90^{th} percentile days, etc. The "incremental probability" for a given high day scenario is defined as the difference between the high day and everyday probability fields.

A second series of RTA metrics, referred to as "location-based sorting" calculates a summary statistic (mean, median, percentile, etc.) from concentrations (or in this case source contributions) at the receptor for all days with trajectories residing over a each grid square. The summary statistic is weighted by the hours over square of the individual trajectories. As with the PSCF metric, the results from location-based sorting are sensitive to the sparse trajectory coverage of distant grid squares, and a censoring function is applied to exclude calculations in squares with sparse coverage.

4 CONCLUSIONS

Comparison of results obtained using models and methods previously described will be discussed and summarized during the workshop. Extension of these models and methods to problems in other scientific areas and communities in environmental modeling and global change studies will be attempted and their participation is encouraged during the workshop presentations and discussions.

5. REFERENCES

Ashbaugh, L.L., 1983. A Statistical Trajectory Technique for Determining Air Pollution Source Regions, *J. Air. Pollut. Contr. Assoc*. 33, 1096-1098.

Ashbaugh, L.L., Malm, W.C., Sadeh, W.Z., 1985. A residence time probability analysis of sulfur concentrations at Grand Canyon national park. *Atmos. Environ.* 19(8), 1263-1270.

Bro R., PARAFAC. Tutorial and applications,1997. *Chemom. Intell. Lab.Syst.*, 38, 149-171.

Cheng, M.D., Hopke, P.K., Zeng, Y., 1993a. A Receptor-Oriented Methodology for Determining Source Regions of Particle Sulfate Composition Observed at Dorset, Ontario, *J. Geophys. Res.* 98, 16839-16849.

Cheng, M.-D., Hopke, P.K., Barrie, L.., Rippe, A., Olson, M., Landsberger, S., 1993b. Qualitative Determination of Source Regions of Long-Range Transported Aerosol Using Data Collected at Canadian High Arctic, *Environ. Sci. Technol.* 27, 2063-2071.

Christensen W.F. and Sain S.R. 2002. Accounting for dependence in a flexible multicariate receptor model. *Technometrics*, 16, 222-60.

de Juan A. and Tauler R, 2003. Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution. *Anal. Chim.Acta*, 500,195-210.

Fharnham I.M., Singh A.K., Stetzenbach K.J.and Johannesson K.H, 2002. Treatment of non-detects in multivariate analysis of groundwater geochemistry data, *Chemom. Intell. Lab.Syst.*, 60, 265-281

Gajewski, B.J. and Spiegelman, C.H., 2004. Correspondence Estimation of the Source Profiles in Receptor Modeling, *Environmetrics,* 15, 613-634.

Henry R. C. 2003. Multivariate Receptor Modeling by N-dimensional Edge Detection, *Chemom. Intell. Lab. Syst.,* **65**,179 - 189.

Henry R. C. 2005. Duality in multivariate receptor models, *Chemometrics and Intelligent Lab. Sys*., **77** , 59-63.

Henry R.C., Lewis C.W., Hopke P.K. Williamson H.J., 1984. Review of receptor model fundamentals, *Atmos. Environ*., 18, 1507-17.

Henry R.C. and Kim B.M., 1990. Extension of self-modelling curve resolution to mixtures of more than three components. Part 1: finding the basic feasible region. Chemom. Intell. Lab.Syst., 8, 205-16.

Henry, R.C., E.S. Park, C.H. Spiegelman 1999. Comparing a new algorithm with the classic methods for estimating the number of factors, *Chemom. Intell. Lab. Syst.,* 48, 91 - 97.

Henry, R. C., Y-S Chang, C. H. Spiegelman 2002. Locating Nearby Sources of Air Pollution by Nonparametric Regression of Atmospheric Concentrations on Wind Direction. *Atmos. Environ.,* **36**, 2237-2244.

Hopke P.K., Paatero P., Jia H., Ross R.T., and Harshman R.A.Three-way (PARAFAC) factor analysis examination and comparison of alternative computational methods as applied to ill-conditioned data. three-way data analysis, 1998. *Chemom. Intell. Lab. Syst.*, 43, 25-42.

Hopke, P.K., Ramadan, Z., Paatero, P., Norris, G., Landis, M., Williams, R., Lewis, C.W., 2003. Receptor Modeling of Ambient and Personal Exposure Samples: 1998 Baltimore Particulate Matter Epidemiology-Exposure Study, *Atmos. Environ.,* **37**, 3289-3302.

Jaumot J., Gargallo R., de Juan A. and Tauler R., 2005. An user friendly interface for MCR-ALS : a new tool for Multivariate Curve Resolution in MATLAB. *Chemom. Intell. Lab. Syst.*, 76, 101-110

Jaumot, J., Gargallo, R., Tauler, R., 2004. Estimation of error propagation and prediction intervals in MCR-ALS using resampling methods, *J. of Chemomet***,** 18, 327–340

Jolliffe I.T., 2002. *Principal Component Analysis*, Springer, 2nd Ed., New York.

Kim, E., Hopke, P.K., Edgerton, E.S., 2003. Source identification of Atlanta aerosol by Positive Matrix Factorization. *Journal of Air and Waste Management Association* 53, 731-739.

Paatero, P., 1997. A Weighted Non-Negative Least Squares Algorithm for Three-Way 'PARAFAC' Factor Analysis, *Chemom. Intell. Lab. Syst.* **38**, 223-242

Paatero P., 1999. The multilinear engine -a table driven least squares program for solving multilinear problems including the n-way parallel factor analysis model, *J. of Computational and Graphical Statistics*, **8**, 854-888

Paatero P. and Hopke P.K., 2002. Utilizing wind direction and wind speed as independent variables in multilinear receptor modelling studies. *Chemom. Intell. Lab. Syst.*, **60**, 25-41.

Paatero P. and Hopke P.K., 2003. Discarding or downweighting high-noise variables in factor analytic models. *Anal. Chim. Acta*., **490**,227-289.

Paatero, P., Hopke, P.K., Hoppenstock, J., Eberly, S., 2003. Advanced Factor Analysis of

Spatial Distributions of PM2.5 in the Eastern U.S., *Environ. Sci. Technol.,* **37**, 2460-2476.

Park E.S., Guttorp P., Henry R.C., 2001. Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC, *Journal of the American Statistical Association*, **96**, 1171-1183.

Park, E.S., Spiegelman, C.H., and Henry, R.C., 2002a. Bilinear Estimation of Pollution Source Profiles and Amounts by Using Multivariate Receptor Models, (with discussions), *Environmetrics,* **13**, 775-798.

Park, E.S., Oh, M.S., and Guttorp, P., 2002b. Multivariate Receptor Models and Model Uncertainty, *Chemom. Intell. Lab. Syst.,* **60**, 49-67.

Park, E.S., Guttorp, P., and Kim, H., 2004. Locating Major PM_{10} Source Areas In Seoul Using Multivariate Receptor Modeling, *Environmental and Ecological Statistics*, **11**, 9-19.

Poirot, R.L., Wishinski, P.R., 1986. Visibility, Sulfate, and Air Mass History Associated with the Summertime Aerosol in Northern Vermont, *Atmos. Environ.* 20, 1457-1469.

Polissar. A.V., Hopke, P.K., Malm, W.C. and Sisler, J.F., 1998. Atmospheric aerosol over Alaska: 2. Elemental composition and sources *J. Geophys. Res.*, **103**, 19045-19057.

Smilde, A., Bro, R. and Geladi, P, 2004. Multi-way Analysis Applications in the Chemical Sciences, Wiley, NewYork.

Ramadan, Z., Eickhout, B., Song, X., Buydens, L.M.C. and Hopke, P.K., 2003. Comparison of Positive Matrix Factorization and Multilinear Engine for the source apportionment of particulate pollutants, *Chemom. Intell. Lab.Syst.*, 66, 15-28

Salou J., Tauler R., Bayona J. and Tolosa I., 1997. Input Characterization of Sedimentary Organic Chemical Markers in the Northwestern Mediterranean Sea by Exploratory Data Analysis. Environ. Sci. and Technol., 1997, 31, 3482-3490

Spiegelman, C.H. and Datttner, S., 1993. Applying and Developing Receptor Models to the 1990 El Paso Air Data: A Look at Receptor Modeling with Uncharacterized Sources and Graphical Diagnostics, *Anal. Chim. Acta*, 277, 347-356.

Spiegelman, C.H. and Datttner, S., 1993. Multivariate Chemometrics, A Case Study: Applying and Developing Receptor Models for the 1990 El Paso Winter PM_{10} Receptor Modeling Scoping Study. In G.P Patil and C.R. Rao (Eds), *Multivariate Environmental Statistics* (pp.509-524). Amsterdam: North-Holland.

Tauler, R., Smilde, A. and Kowalski, B.R. Selectivity, 1995. Local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. of Chemom.*, 9, 31-58

Tauler, R. 1995. Multivariate curve resolution applied to second order data. *Chemom. Intell. Lab. Syst.*, 30, 133-146

Tauler R., Barcelo D. and Thurman E.M., 200. Multivariate correlations between concentration of selected herbicides and derivatives in outflows from selected US midwestern reservoirs. *Environ. Sci. Technol.*, 34, 3307-3314

Tauler R., 2001. Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. *J. of Chemom.*, 15, 627-646

Tauler R., Lacorte S., Guillamón M., Cespedes R., Viana P. and Barceló D., 2004. Resolution of main environmental contamination sources of semivolatile organic compounds in surface waters of Portugal using chemometric compounds. *Environ. Toxicol. Chem.,* 23, 565-575

Thurston G.D. and Spengler D., 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston, *Atmos. Environ.*, 19, 9-25

Van Huffel, S., Vandevalle, J., 1991. The Total Least Squares Problem. Computational Aspects and analysis, *SIAM*, Philadelphia.

Walczak, B. and Massart, D.L., 2001. Dealing with missing data. Part I. *Chemom. Intell. Lab.Syst.* 58, 15-27

Yakovleva, E., Hopke, P.K., Wallace, L. 1999. Receptor Modeling Assessment of PTEAM Data, *Environ. Sci. Technol.* **33**, 3645-3652.

Yu, K.N.; Y.P. Cheung; T. Cheung; and R. C. Henry 2004. Identifying the impact of large urban airports on local air quality by nonparametric regression. *Atmos. Environ.* **38**,4501-4507.

Zeng, Y., Hopke, P.K., 1989. A Study on the Sources of Acid Precipitation in Ontario, Canada, *Atmos. Environ.* 23, 1499-1509.

Zhao, W., Hopke, P.K., Norris, G., Williams, R., Paatero, P. 2006. Source Apportionment and Analysis on Ambient and Personal Exposure Samples with a Combined Receptor Model and an Adaptive Blank Estimation Strategy, *Atmos. Environ.* (in press).