JOURNAL OF CHEMOMETRICS
J. Chemometrics 2003; **17**: 255–265
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.796

Recent developments in receptor modeling

Philip K. Hopke*
Department of Chemical Engineering, Clarkson University, Box 5705, Potsdam, New York 13699-5705, USA

Received 15 September 2002; Revised 15 December 2002; Accepted 27 January 2003

Receptor modeling is the application of data analysis methods to elicit information on the sources of air pollutants. Typically, it employs methods of solving the mixture resolution problem using chemical composition data for airborne particulate matter samples. In such cases, the outcome is the identification of the pollution source types and estimates of the contribution of each source type to the observed concentrations. It can also involve efforts to identify the locations of the sources through the use of ensembles of air parcel back trajectories. In recent years, there have been improvements in the factor analysis methods that are applied in receptor modeling as well as easier application of trajectory methods. These developments are reviewed. Copyright © 2003 John Wiley & Sons, Ltd.

KEYWORDS: receptor modeling; air pollution

1. INTRODUCTION

The management of air quality is a difficult but an important problem. In general it involves the identification of the sources of materials emitted into the air, the quantitative estimation of the emission rates of the pollutants, the understanding of the transport of the substances from the sources to downwind locations and the knowledge of the physical and chemical transformation processes that can occur during that transport. All of those elements can then be put together into a mathematical model that can be used to estimate the changes in observable airborne concentrations that might be expected to occur if various actions are taken. Such actions could include the initiation of new sources as new industries are built and begin to function and the imposition of emission controls on existing facilities in order to reduce the pollutant concentrations.

However, the atmosphere is a very complex system and it is necessary to simplify greatly the descriptions of reality in order to produce a mathematical model capable of being calculated on even the largest and fastest computers. Significant improvements have been made over the past 20 years in the mathematical modeling of dispersion of pollutants in the atmosphere. However, there are still many instances when the models are insufficient to permit the full development of effective and efficient air quality management strategies, particularly for airborne particulate matter. Hence it is necessary to have other methods available to assist in the identification of sources and the apportionment of the observed pollutant concentrations to those sources. Such methods are 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 the transport, dilution and transformations that occur beginning at the source and following the pollutants to the sampling or receptor site. These methods have been applied primarily to airborne particulate matter. In the United States, there are two size ranges of particles that are regulated by the US Environmental Protection Agency. Particulate matter in the air with aerodynamic diameters less that 10 μm is called PM_{10} where as the mass concentration of particles less than 2.5 μm is termed $PM_{2.5}$. A comprehensive view of the field can be found in Hopke [1]. A review through 1996 is provided as part of a larger report by Seigneur *et al.* [2]. Thus, in this paper, work presented since 1997 will be highlighted.

*Correspondence to: P. K. Hopke, Department of Chemical Engineering, Clarkson University, Box 5705, Potsdam, New York 13699-5705, USA.
E-mail: hopkepk@clarkson.edu

2. BACKGROUND

The fundamental principle of receptor modeling is that mass conservation can be assumed and a mass balance analysis can be used to identify and apportion sources of airborne particulate matter in the atmosphere. This methodology has generally been referred to within the air pollution research community as *receptor modeling* [1,3]. The approach to obtaining a data set for receptor modeling is to determine a large number of chemical constituents such as elemental concentrations in a number of samples. Alternatively, automated electron microscopy can be used to characterize the composition and shape of particles in a series of particle samples. In either case, a mass balance equation can be written to account for all m chemical species in the n samples

as contributions from p independent sources:

$$x_{ij} = \sum_{p=1}^{p} g_{ip} f_{pj} + e_{ij} \quad (1)$$

where x_{ij} is the measured concentration of the jth species in the ith sample, f_{pj} is the concentration of the jth species in material emitted by source p, g_{ip} is the contribution of the pth source to the ith sample and e_{ij} is the portion of the measurement that cannot be fitted by the model.

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 particle mass [4]. The fundamental, natural physical constraints that must be obeyed are as follows:

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 elemental concentration.
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 the total measured mass for each element; the whole is greater than or equal to the sum of its parts.

While developing and applying these models, it is necessary to keep these constraints in mind in order to be certain of obtaining physically realistic solutions.

Thus, receptor modeling is a variation on the 'spectro-chemical mixture resolution' problem in chemometrics. However, there are some additional complicating aspects since source profiles do not remain constant in contrast to molecular spectra and the environmental data tend to have much higher noise in the measurements. This comparison is discussed in more detail by Hopke [5].

3. RECEPTOR MODELS

3.1. Sources known

There are a variety of ways to solve Equation (1) depending on what information is available. If the number and nature of the sources in the region are known (i.e. p and f_{ik}s), then the only unknown is the mass contribution of each source to each sample, g_{kj}. This approach was first independently suggested by Winchester and Nifong [6] and by Miller *et al.* [7]. The problem is typically solved using an effective-variance least-squares approach [8] that is now generally referred to as the chemical mass balance (CMB) model. Software [9] is available from the US Environmental Protection Agency at www.epa.gov/ttn/SCRAM. Solution methods using multivariate calibration methods have also been proposed, and are summarized in the earlier review [2]. There have not been any new method developments in this area nor have there been many new source profiles developed that are readily available. The US Environmental Protection Agency's library is contained in SPECIATE, which is available from www.epa.gov/ttn/CHIEF. There are some profiles, particularly for spark-ignition and diesel vehicles, that have been measured and have not yet been added to the database. This work and recent CMB studies have been reviewed by Chow and Watson [10]. It is anticipated that an update to the SPECIATE library will be done in 2003 and, thus, will become more readily available.

3.2. Sources unknown

The area of active method development has been in the methods to be used when the source profiles are not known. These are forms of factor analysis, but completely different from traditional principal components analysis and related techniques. In factor analysis, the problem is expanded to the solution of the source profiles and contributions over a set of samples. Thus, the basic equation in matrix form is

$$\mathbf{X} = \mathbf{GF}' \quad (2)$$

The two new approaches are UNMIX [4,11–13] and positive matrix factorization (PMF) [14–16].

3.2.1. UNMIX

UNMIX is based on an eigenvalue analysis. The UNMIX model is a new type of multivariate receptor model based on principal component analysis (PCA). The model uses a new transformation method based on the self-modeling curve resolution (SMCR) techniques. Since a unique solution is not possible [17], the SMCR technique restricts the feasible region of the real solution into a small region with explicit physical constraints, such as source compositions must be greater than or equal to zero. Explicit physical constraints form linear inequality constraints in the space spanned by the eigenvectors, and these constraints form the feasible region in eigenvectors' space. UNMIX is designed to resolve the most important sources contributing to the measured mass concentrations.

The model was applied to 1986 PM_{10} data from Los Angeles [18]. Source compositions and their source apportionments of the major source categories, such as roadway, marine, crustal and secondary sources, were estimated at six sites in the South Coast Air Basin. To reduce further the size of the feasible region in this model and to estimate source compositions better, carbon monoxide and ozone gas data were used along with the PM_{10} data. These two gases were used as additional physical constraints for the model application because carbon monoxide and ozone are unique tracers for the motor vehicle and secondary sources, respectively. In addition, a new concept of additional physical constraint, a stoichiometric constraint, was also used.

Major source categories identified are motor vehicle, crustal, secondary and marine sources. The factor analysis (FA) or PCA model cannot distinguish spatially and temporally correlated sources. From the FA or PCA standpoint, spatially and temporally correlated sources are perceived as a single source because they almost always impact the receptor site at the same time. Although these are not a single source, they can be assumed as a pseudo-single composite source. The motor vehicle and road dust sources are spatially and temporally correlated as the road dust is resuspended in the air when the motor vehicle passes over the road.

Recently, the model has also been applied to data from Phoenix, AZ [19]. The analysis generated source profiles and

overall average percentage source contribution estimates for five source categories: gasoline engines ($33 \pm 4\%$), diesel engines ($16 \pm 2\%$), secondary sulfate ($19 \pm 2\%$), crustal/soil ($22 \pm 2\%$) and vegetative burning ($10 \pm 2\%$). In this study, the authors were able to separate motor vehicle contributions into diesel and spark-ignition sources. Diesel emissions were identified by high elemental carbon relative to the organic carbon whereas spark ignition vehicles had a profile with more organic than elemental carbon. They found a substantial difference in the contribution of diesel emissions between weekend and weekday samples.

The US Environmental Protection Agency has developed this model as a stand-alone program which has been in beta tests during 2002 and is expected to be available from www.epa.gov/ttn/SCRAM in the near future.

3.2.2. *PMF*

Positive matrix factorization takes a very different approach to the factor analysis problem. All of the other methods use an eigenvector analysis based on a singular value decomposition (SVD). The **X** matrix can also be defined:

$$\mathbf{X} = \mathbf{USV}' = \overline{\mathbf{U}}\,\overline{\mathbf{S}}\,\overline{\mathbf{V}}' + \mathbf{E} \tag{3}$$

where $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ are the first p columns of the **U** and **V** matrices. The **U** and **V** matrices are calculated from eigenvalue–eigenvector analyses of the $\mathbf{XX}'$ and $\mathbf{X}'\mathbf{X}$ matrices, respectively. It can be shown [20,21] that the second term on the right-hand side of Equation (3) estimates **X** in the least-squares sense that it gives the lowest possible value for

$$\sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij}^{\,2} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \sum_{p=1}^{P} g_{ip}f_{pj}\right)^2 \tag{4}$$

Thus, an eigenvector analysis is an *implicit* least-squares analysis in that it is minimizing the sum of squared residuals for the model. Paatero and Tapper [22] showed that in effect in PCA, there is scaling of the data by column or by row and that this scaling will lead to distortions in the analysis. They further showed that optimum 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. However, they showed that point-by-point scaling results in a scaled data matrix that cannot be reproduced by a conventional factor analysis based on the singular value decomposition. Thus, PMF takes the approach of an *explicit* least-squares approach in which the method minimizes the object function:

$$Q = \sum_{j=1}^{n}\sum_{i=1}^{m}\left(\frac{x_{ij} - \sum_{p=1}^{P} g_{ip}f_{pj}}{s_{ij}}\right)^2 \tag{5}$$

where s_{ij} is an estimate of the 'uncertainty' in the jth variable measured in the ith sample. The factor analysis problem is then to minimize $Q(\mathbf{E})$ with respect to **G** and **F** with the constraint that each of the elements of **G** and **F** is to be non-negative.

Over the past few years, several approaches to solving the PMF problem have been developed. Initially, a program called PMF2 utilized a unique algorithm [14] for solving the factor analytic task. For small- and medium-sized problems, this algorithm was found to be more efficient than ALS methods [23]. Subsequently, an alternative approach that provides a flexible modeling system was developed for solving the various PMF factor analysis least-squares problems [15]. This approach, called the multilinear engine (ME), has been applied to an environmental problem [24–26], but has not yet been widely used.

PMF2 was initially applied to data sets of major ion compositions of daily precipitation samples collected over a number of sites in Finland [27] and samples of bulk precipitation [28] in which they are able to obtain considerable information on the sources of these ions. Polissar *et al.* [29] applied the PMF2 program to Arctic data from seven National Park Service sites in Alaska as a method to resolve the major source contributions more quantitatively.

Recently, there has been a series of applications of PMF to various source/receptor modeling problems. Polissar *et al.* [30] re-analyzed an augmented set of Alaskan NPS data and resolved up to eight sources. Xie *et al.* [24,31] made several analyses of data from an 11-year series of particulate matter samples taken at Alert, NWT. Polissar *et al.* [32] examined the semicontinuous aerosol data collected by NOAA at their atmospheric observatory at Barrow, Alaska. Lee *et al.* [33] applied PMF to urban aerosol compositions in Hong Kong. They were able to identify up to nine sources that provided a good apportionment of the airborne particulate matter.

Paterson *et al.* [34] applied PMF to air quality and temperature data collected at a series of sites around the southern end of Lake Michigan in 1997 and used three factors to reproduce 75% of the variation in the data. Huang *et al.* [35] analyzed elemental composition data for particulate matter samples collected at Narragansett, RI, using both PMF and conventional factor analysis. They were able to resolve more components with more physically realistic compositions with PMF. Hence the approach is attracting interest because it does have some inherent advantages, particularly through its ability to weight each data point individually. PMF is more complex and harder to use, but it appears to provide improved resolution of sources and better quantification of impacts of those sources than PCA [35].

Yakovleva *et al.* [36] applied PMF to particle composition data taken from personal samplers and also indoor and outdoor samplers around the home in which the person with the sampler lived. These data were analyzed as both two-way and three-way problems. In the three-way analysis, it is possible to ascertain the extent of indoor particle concentration that is of ambient origins.

Chueinta *et al.* [37] analyzed PM_{10} composition data. They introduced a source contribution rose analogous to a wind rose to help provide information on the direction of the source relative to the receptor site. Ramadan *et al.* [38] applied PMF to a set of daily data from Phoenix, AZ. In this analysis, separate profiles for diesel and spark-ignition vehicles. Lewis *et al.* [19] analyzed the same data and found relatively similar results for those sources that contribute the largest amounts to the ambient mass concentrations.

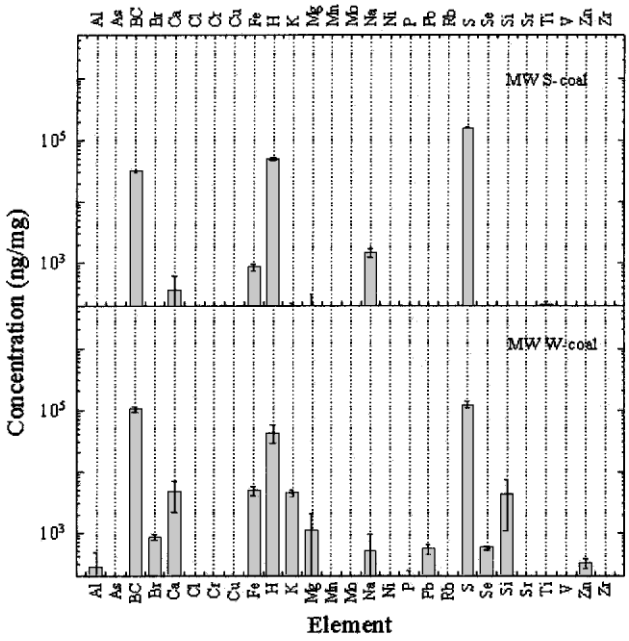

Figure 1. PMF-derived source profiles for the sulfur factors from Underhill, VT taken from Polissar *et al.* [39]. Top, summer coal-fired power plants; bottom, winter coal-fired power plants.

Aerosol chemical composition data for $PM_{2.5}$ samples collected during the period from 1988 to 1995 at Underhill, VT, were analyzed by Polissar *et al.* [39]. An 11-factor solution was obtained. Sources representing wood burning, coal and oil combustion, the coal combustion emissions plus photochemical sulfate production, metal production plus municipal waste incineration and the emissions from motor vehicles were identified. Emissions from smelting of non-ferrous metal ores, arsenic smelting, and soil particles and particles with high concentrations of Na were also identified by PMF.

Polissar *et al.*'s results show an interesting feature of factor analysis solutions. Two sulfur-dominated factors were identified. The first S factor also has the highest loadings of Se (Figure 1, top) suggesting emissions from coal-fired power plants. The factor has an annual cycle with maxima during the winter/spring season and minima in the summer (Figure 2, top). It is thus denoted W-Coal. The second S factor has the opposite annual cycle with the summer maxima (Figure 2, bottom). It has a much higher S/Se ratio (Figure 1, bottom). It is suggested that the second S factor represents the photochemically enhanced sulfate production from SO_2 in the summer (S-Coal). Sources related to the second S factor provide the highest fine particulate matter mass concentrations. It is likely that the two S factors represent the same coal-fired power plants with the differences between the factors representing the extremes of photochemical production of sulfate from the emitted SO_2. Figure 3 shows the Se/S scatter plot for Underhill. Se is generally an important tracer for coal combustion. The slope for the linear regression for the winter subset of the data points is higher than that for the summer data while the corresponding correlation coefficient is higher for the summer. Thus, although there is a single source type, two factors are required to reproduce the variability in the source profile arising from the differences in seasonal atmospheric processes. It is this type of variability that makes receptor modeling different from the standard mixture resolution problem in chemometrics.

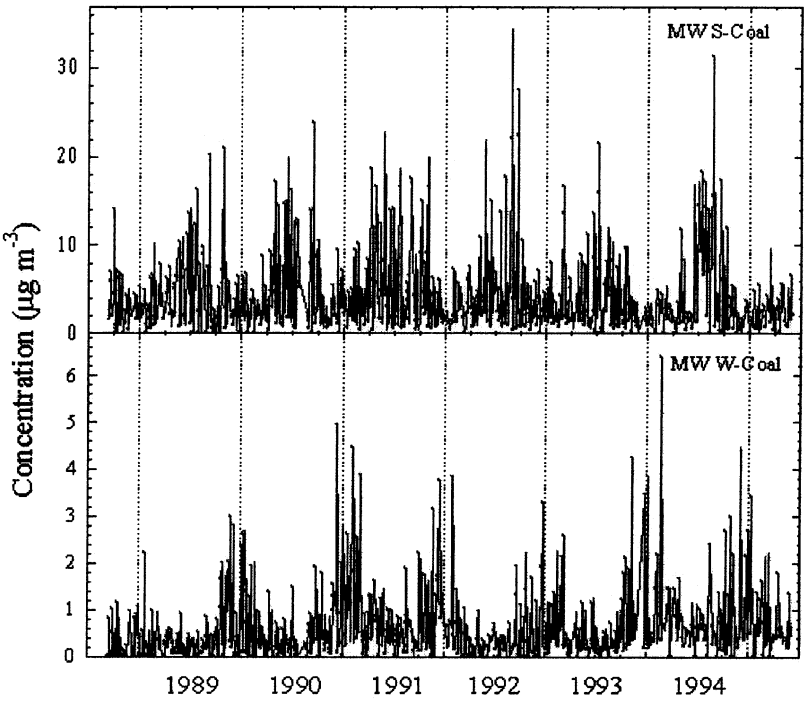

Figure 2. PMF-derived source contributions for the sulfur factors from Underhill, VT taken from Polissar *et al.* [39]. Top, summer coal-fired power plants; bottom, winter coal-fired power plants.

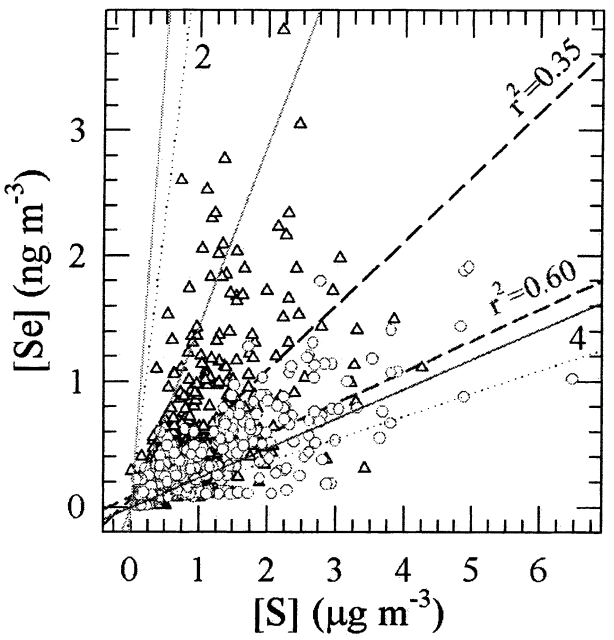

Figure 3. Concentrations of Se versus S. Dashed lines represent linear regression fits for the 'winter' and 'summer' groups of points. The two solid lines represent the Se/S ratios for the two resolve 'coal' factors.

Song *et al.* [40] analyzed similar data from Brigantine, NJ and Washington, DC and compared these results with those from Underhill. In these data, similar two sulfur factors (winter and summer) were observed.

3.2.3. *Comparison of UNMIX and PMF*

Both UNMIX and PMF have been applied to the Underhill, VT data [41]. The UNMIX model (run with 838 observations of 12 input variables) identified seven sources. The PMF model (run with 853 observations of 27 input variables) identified 11 sources. Both models reproduce the daily mass reasonably well. UNMIX, which included mass as an input variable, reproduced the average and daily mass concentrations somewhat more closely (essentially attributing all the mass among the seven identified sources). PMF, which apportioned mass by regression from the daily mass and source scores, leaves some of the mass unexplained by the 11 sources it identified from the 27 non-mass input variables.

In most cases the source profiles are similar, the daily source contributions are well correlated ($R^2 > 0.75$), and the slopes of the daily mass comparisons are generally within 25% of 1:1. Notable exceptions occur for the MW Winter Coal and the Canadian Mn sources, which have similar composition profiles and highly correlated daily contributions, but show substantially higher mass contributions from UNMIX than for PMF–hence the sources are 'similar,' but their mass attributions differ. The Soil sources have similar profiles for elements included in both models, but show the poorest daily correlation ($R^2 = 0.73$), and a mass contribution from the PMF Soil source which is about 50% higher than from the UNMIX Soil source. Despite these discrepancies, the authors felt it reasonable to conclude that for the seven sources identified by the UNMIX model, there were 'similar' counterpart sources also identified by PMF.

3.3. Advanced Factor Models

Airborne concentrations due to specific sources may display a sharp directional pattern with respect to wind directions. In these cases, concentrations are high when the air arrives from certain direction(s) while concentrations associated with other directions are low or nil. Such a non-linear dependence cannot be directly modeled so that wind information would be included in a factor analytical model as one or a few special variables, used in parallel with the ordinary variables, the concentrations. There may be other similar kinds of effects such as weekend/weekday activity patterns, time of day, time during the year, etc., that significantly affect the observed elemental concentrations. The non-linear variables can be included in the model as *independent* or *free* variables [42].

In the present expanded PMF analysis, the bilinear Equation (1) is augmented by another more complicated set of equations that contain modeling information. In its most basic form, the contribution r_{ijp} of source p is represented by the following expression:

$$r_{ijp} = m_{ip} f_{jp} = \mathbf{D}(\delta_i, p)\mathbf{V}(v_i, p) f_{jp} \tag{6}$$

The known values δ_i and v_i indicate wind direction and wind speed on day i. The symbols **D** and **V** represent matrices, consisting of unknown values to be estimated during the fitting process. Their columns numbered p correspond to source number p. For typographic reasons, their indices are shown in parentheses, not as subscripts. The index value δ_i for day i is typically obtained by dividing the average wind direction of day i (in degrees) by 10 and rounding to nearest integer. As an example, if source 2 comes strongly from the wind direction at 90°, then the element **D**(9,2) is likely to become large. The values v_i are obtained from a chosen classification of wind speeds. The following classification was used in this work: 0–1.5–2.5–3.5–5.8–∞ m s^{-1}. Hence $v_i = 2$ for such days when the average wind speed is between 1.5 and 2.5 m s^{-1}.

In component form, the equations of the model are

$$x_{ij} = \sum_{p=1}^{P} g_{ip} f_{jp} + e_{ij}$$

$$x_{ij} = \sum_{p=1}^{P} m_{ip} f_{jp} + e'_{ij} = \sum_{p=1}^{P} \mathbf{D}(\delta_i, p)\mathbf{V}(v_i, p) f_{jp} + e'_{ij} \tag{7}$$

The notation m_{ip} does not indicate a factor element to be determined, such as g_{ip}, but the expression defined by the physical model in question. In different physical models, m_{ip} will correspond to different expressions. Because the variability of m_{ip} is restricted by the model, the second set of Equations (7) will produce a significantly poorer fit to the data than the first set of Equations (7). The physical model, m_{ip}, is one of the multiple possible models depending on the understanding of the system under study while the mass balance in the first set of equations should be much more applicable. Thus, the error estimates connected with the second set of equations must be (much) larger than the error estimates connected with the first set of equations.

The task of solving this expanded PMF model means that values of the unknown factor matrices, **G**, **F**, **D** and **V**, are to

be determined so that the model fits the data as well as possible. In other words, the sum-of-squares value Q, defined by

$$Q = \sum_{i=1}^{I}\sum_{j=1}^{J}(e_{ij}/\sigma_{ij})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(e'_{ij}/\sigma'_{ij})^2 \quad (8)$$

is minimized with respect to the matrices **G**, **F**, **D** and **V**, while the residuals e_{ij} and e'_{ij} are determined by Equations (7). The error estimates σ'_{ij} must be specified (much) larger than the corresponding error estimates σ_{ij}.

Since there are other sources of variation such as weekend/weekday source activity patterns or seasonal differences in emission rates or in atmospheric chemistry, additional factors are included in the model. In this case, wind direction, wind speed, time of year and weekend/weekday will be used. In this case, 24 1-h average values are available for wind speed and direction. Time of year will be aggregated into six 2-month periods or *seasons*, indicated for each day i by the index variables σ_i. (The Greek letter σ is used for two purposes: σ_{ij} indicates the error estimates of data values, while σ_i indicates the season number for day i.). For the values $i = 1$ to $i = 60$, $\sigma_i = 1$, meaning that January and February belong to the first season. For the values $i = 61$ to $i = 121$, $\sigma_i = 2$, and so on.

The non-linear dependences are now defined by the following multilinear expression:

$$m_{jp} = \mathbf{W}(\omega_i, p)\mathbf{S}(\sigma_i, p)\sum_{h=1}^{24}\mathbf{D}(\delta_{ih}, p)\mathbf{V}(v_{ih}, p) \quad (9)$$

where $\mathbf{D}(\delta_{ih}, p)$ is the element of **D** with the index for the wind direction during hour h of day i for the pth source, $\mathbf{V}(v_{ih}, p)$ is the element of **V** with the index for the wind speed during hour h of day i for the pth source, $\mathbf{W}(\omega_i, p)$ is the element of **W** with the index corresponding to day i for the weekday/weekend factor for the pth source, and $\mathbf{S}(\sigma_i, p)$ is the element of **S** with the index corresponding to the time-of-year classification of day i for the pth source. Each of these matrices, **D**, **V**, **W** and **S**, contain unknown values to be estimated in the analysis. The specific factor elements used to fit a particular data point are selected based on the hourly (**D**, **V**) or daily (**W**, **S**) values of the corresponding variables. Hence these auxiliary variables are not fitted, but serve as indicators to the values to be fitted.

The expanded model to be fitted thus consists of the basic bilinear equations plus a set of multilinear equations specifying the physical model:

$$x_{ij} = \sum_{p=1}^{P} g_{ip}f_{jp} + e_{ij} \quad (10)$$

$$x_{ij} = \sum_{p=1}^{P} m_{ip}f_{jp} + e'_{ij} = \sum_{p=1}^{P}\sum_{h=1}^{24}\mathbf{D}(\delta_{ih}, p)\mathbf{V}(v_{ih}, p)\mathbf{W}(\omega_i, p)\mathbf{S}(\sigma_i, p)f_{jp} + e'_{ij} \quad (11)$$

The multilinear engine (ME) was used to solve this problem with non-negativity required for all of the elements of the matrices being estimated [15].

This model was applied to a set of simulated data created by the EPA [43]. Sixteen distinct source profiles were used in Palookaville simulation: nine point sources, four industrial complexes, one area source and two highways. Hourly meteorological data including wind speed and direction were used in the ISC3 model to estimate the concentrations at the receptor site. The area profile was a mixture of dust and road profiles. All source profiles with the exception of the petroleum refinery were fixed. The latter profile had some built-in variability (coefficient of variation (CV) of approximately 25%). Temporal modulation of the source strengths (50% CV for most) was found to be essential in being able to resolve the sources by PMF or UNMIX. A total of 366 24-h samples were generated at the receptor site.

Comparisons with known true data indicate that the analysis is successful. More factors could be determined than by the state-of-the-art bilinear technique PMF. Fifteen of the 16 sources could be resolved. Close inspection of the results reveals that minor rotational problems still remain. They are mainly visible so that the strongest elements of the strongest factors tend to appear in the weaker factors. The directional information derived from the wind direction factors pointed to each of the point sources. The mass apportionment was much closer to the true values than could be obtained with the simple bilinear modeling. This analysis was based on 24-h concentrations and 1-h weather data. The success of the analysis demonstrates that high-resolution weather data may significantly enhance the usefulness of 24-h concentration data.

4. METHODS INCORPORATING BACK TRAJECTORIES

The dispersion models described elsewhere in this paper describe the transport of the particles from a source to the sampling location. However, using an analogous model of atmospheric transport, a model calculates the position of the air being sampled backward in time from the receptor site from various starting times throughout the sampling interval. The trajectories are then used in residence time analysis (RTA), areas of influence analysis (AIA), quantitative bias trajectory analysis (QTBA), potential source contribution function (PSCF) and residence time weighted concentrations (RTWC). AIA, QTBA and RTWC have only been used in a single publication for each method and those results are reviewed by Seigneur *et al.* [2]. Only PSCF and RTA have been used in recent published studies.

4.1. Residence time analysis

In RTA, 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, hence the residence time is maximum in the cells surrounding the sampling location. Ashbaugh *et al.* [44] suggested one solution to this problem that will be described below as potential source contribution function analysis. An alternative method which has come to be called residence time analysis was developed by Poirot and Wishinski [45]. In their method, they first interpolated along each trajectory segment to estimate the fraction of time spent in each grid cell and then summed the residence time for that cell. They proposed a method to adjust the resulting grid cell values for the geometric problem of high values in the region immediately adjacent to the receptor site.

Poirot *et al.* [41] performed an RTA analysis of the source contributions to particles collected at Underhill, VT that we estimated with both UNMIX and PMF. Figure 4 compares the RTA upwind incremental probability fields for the highest 10% of daily source contributions for the seven similar sources identified independently by the UNMIX and PMF models. The strong similarities in the incremental probability plots for the 'similar sources' identified by PMF and UNMIX are not surprising, given the strong correlations between the modeled source contributions

4.2. Potential source contribution function

The potential source contribution function (PSCF) receptor model was originally developed by Ashbaugh *et al.* [44] and Malm *et al.* [46]. It has been applied in a series of studies over a variety of geographical scales [47–51]. Air parcel back trajectories ending at a receptor site 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 ijth cell (represented by n_{ij}), the probability of this event, A_{ij}, is given by

$$P[A_{ij}] = \frac{n_{ij}}{N} \tag{12}$$

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

Suppose in the same ijth 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 $P[B_{ij}]$:

$$P[B_{ij}] = \frac{m_{ij}}{N} \tag{13}$$

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

The potential source contribution function (PSCF) is defined as

$$P_{ij} = \frac{P[B_{ij}]}{P[A_{ij}]} = \frac{m_{ij}}{n_{ij}} \tag{14}$$

where P_{ij} is the conditional probability that an air parcel which passed through the ijth cell had a high concentration upon arrival at the trajectory endpoint. Although the trajectory segment endpoints are subject to uncertainty, a sufficient number of endpoints should provide accurate estimates of the source locations if the location errors are random and not systematic. Cells containing emission sources would be identified with conditional probabilities close to one if trajectories that have crossed the cells effectively transport the emitted contaminant to the receptor site. The PSCF model thus provides a means to map the source potentials of geographical areas. It does not apportion the contribution of the identified source area to the measured receptor data.

Xie *et al.* [52] used PSCF to examine the locations of the sources identified by the PMF analysis of the data from Alert. The results of these analyses were in agreement with earlier efforts that examined the PSCF maps for the individual chemical constituents in the particle samples.

Poissant [53] used PSCF to examine the likely source locations for total gaseous mercury observed in the St Lawrence River valley. During the winter, fall and spring period the distribution of potential sources reasonably reproduces the North American Hg emission inventory. However, because a single fixed criterion was over the entire year and transport from many of the strong source areas was weak during the summer months, few source areas were observed during the summer data where the concentrations were the lowest.

Polissar *et al.* [54] examined the particle data (black carbon, light scattering and condensation nuclei counts) collected at Point Barrow, Alaska. They found that they could distinguish between biogenic sources of the small particles seen only with the condensation nuclei counter from anthropogenic larger particles that scatter and absorb light. The biogenic particles came primarily from the open areas of the North Pacific Ocean whereas most of the anthropogenic particles came from known industrialized areas of Russia.

Polissar *et al.* [39] applied PSCF to the PMF results for the data from Underhill, VT. The results helped to clarify the nature of the sources. For example, the high degree of overlap in the source regions for the winter and summer 'coal-fired power plant' source type suggested that the observed sulfate was coming from the same emission sources. The primary difference then between the winter and summer periods was the degree of photochemical activity leading to the kind of differences in Se/S ratios that can be seen in Figure 3 and was discussed earlier in this paper.

4.3. Comparison of RTA and PSCF results

The only direct comparison of these two trajectory ensemble methods to date is that of Poirot *et al.* [41]. Figure 5 compares the areas identified by the two methods for two of the sources, east coast oil-fired power plants and Canadian non-

ferrous metal smelters. At this choice of criterion value for PSCF, there is a trailing effect beyond the region in which the smelters are situated. A higher criterion value would reduce, but not eliminate this trailing effect. Other comparisons between these two methods provided very similar results, indicating that they can both provide approximately equivalent information on the likely source locations that contribute material to a specific receptor site.

5. FUTURE DIRECTIONS

The US Environmental Protection Agency has deployed an extensive network of samplers to collect particle samples in urban areas for particle composition measurements. This Chemical Speciation Network will have 54 sites for which samples are collected every third day and more than 200 other sites where data are collected either every third day or every sixth day. Thus, in a few years there will be an extensive database available for receptor modeling of urban areas.

There have been new developments in semicontinuous sampling/analysis systems for major components of the ambient aerosol. Commercial instruments are now available for sulfate, nitrate and organic and elemental carbon. These systems typically provide hourly data. Research instruments have been developed to measure ions or trace elements on a

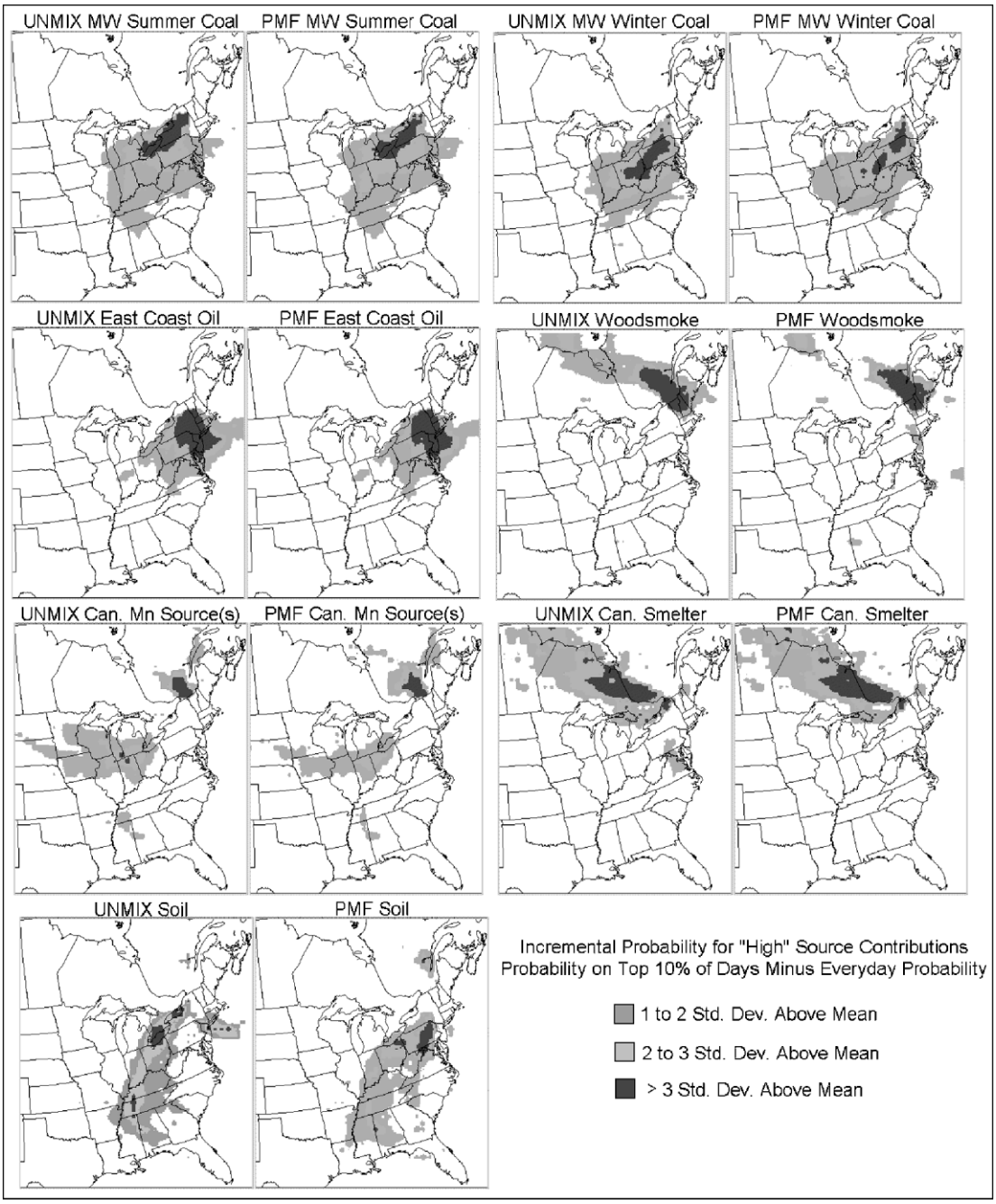

Figure 4. RTA incremental probability fields for top 10% of daily UNMIX and PMF source contributions at Underhill, VT [41].

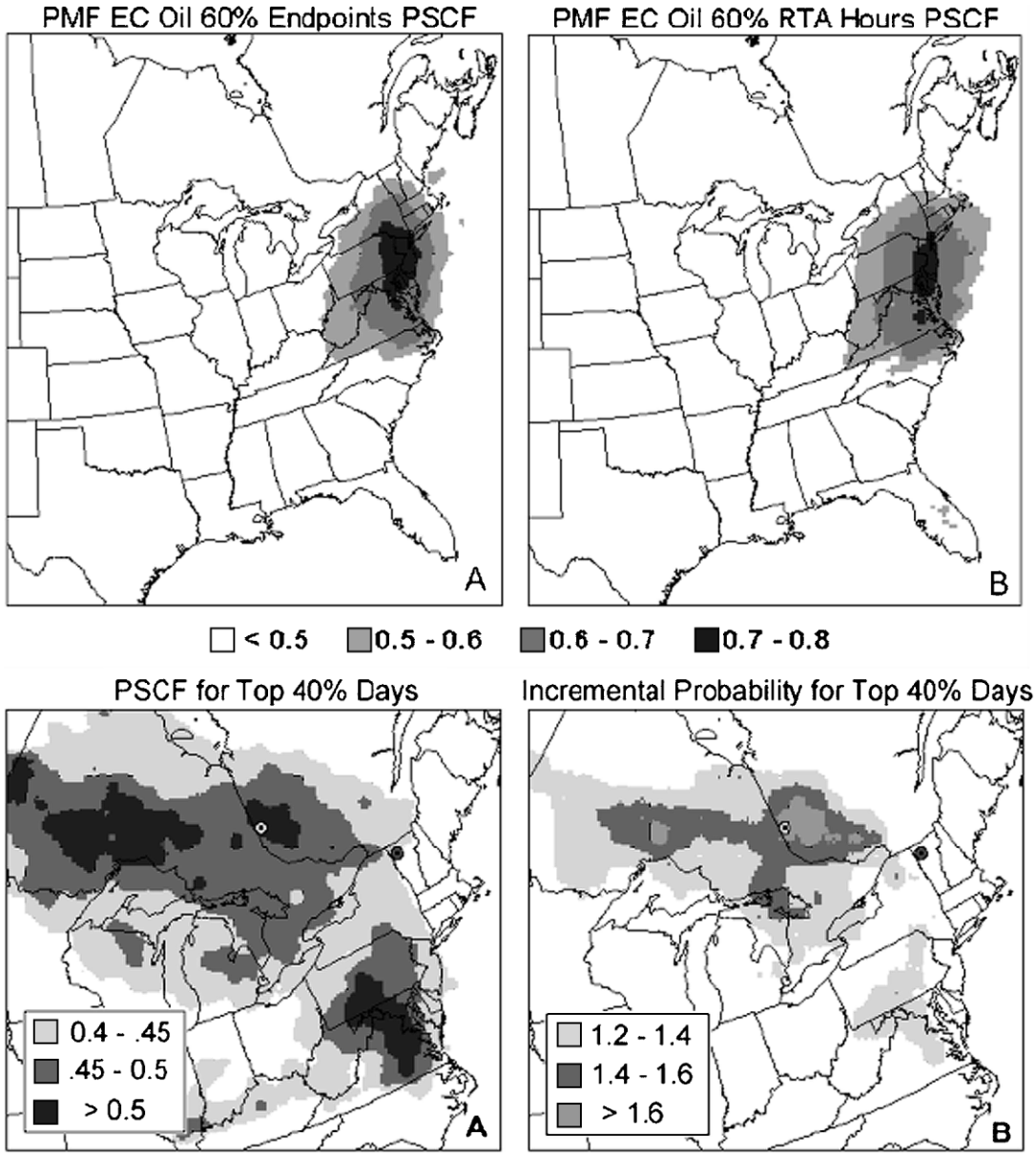

Figure 5. Comparison of PSCF (A) and RTA (B) trajectory ensemble method results based on the PMF analysis of Polissar *et al.* [39], illustrated by ratio of 60th percentile to every day for PMF 'East Coast Oil' source (top) and 'Canadian Smelter' (bottom).

time scale of every 30 min. These instruments permit the observation of rapidly changing concentrations that occur when the sampling site is affected by the plume of particular point sources. Such data will make factor analysis models much more effective because the wind data can be used to determine that certain downwind source contributions can be forced to zero. With greater variability in the data, more sources will be more reliably extracted from the data.

There is growing interest in the use of resolved source contributions in epidemiological studies of the relationships between airborne particle and adverse human health effects. It is thought unlikely that all particles have equal toxicity, hence the problem then exists of how to organize data characterizing particle samples to enter appropriate statistical models. There are two many chemical components typically measured and there is often high correlation among them because they do come from a limited number of common sources. Hence it is anticipated that there will be an increased demand for easy-to-use software that will permit even complex receptor models to be applied to a wider variety of available data.

Finally, as the EPA begins to declare areas of the United States to be in non-attainment of the $PM_{2.5}$ ambient air quality standard, there will need to be application of these new receptor model methods to data such as those from the Speciation Network to provide information for state and local air quality management strategy development. A similar problem will arise in Europe as the new European PM_{10} standards start to be enforced and areas are identified that have problems that require identification and quantitative apportionment of particle sources. Thus, receptor models continue to be developed and improved and there

appears to be a substantial need for the application in the near future.

REFERENCES

1. Hopke PK (ed) *Receptor Modeling for Air Quality Management*, Elsevier: Amsterdam, 1991.
2. Seigneur C, Pai P, Louis JF and Hopke PK and Grosjean D. *Review of Air Quality Models for Particulate Matter*, Report 4669. American Petroleum Institute: Washington, DC, 1997.
3. Hopke PK. *Receptor Modeling in Environmental Chemistry*. Wiley: New York, 1985.
4. Henry RC. Multivariate receptor models. In *Receptor Modeling for Air Quality Management*, Hopke PK (ed). Elsevier: Amsterdam, 1991; 117–147.
5. Hopke PK. The mixture resolution problem applied to airborne particle source apportionment. In *Chemometrics in Environmental Chemistry*, Springer: Heidelberg, 1995; 47–86.
6. Winchester JW and Nifong GD. Water pollution in Lake Michigan by trace elements from pollution aerosol fallout. *Water Air Soil Pollut* 1971; **1**: 50–64.
7. Miller MS and Friedlander SK and Hidy GM. A chemical element balance for the pasadena aerosol. *J. Colloid Interface Sci.* 1972; **39**: 65–176.
8. Cooper JA and Watson JG and Huntzicker JJ. The effective variance weighting for least squares calculations applied to the mass balance receptor model. *Atmos. Environ.* 1984; **18**: 1347–1355.
9. Watson JG, Robinson NF, Chow JC, Henry RC, Kim BM, Pace TG and Meyer EL and Nguyen Q. The USEPA/DRI chemical mass balance receptor model, CMB 7.0. *Environ. Software* 1990; **5**: 38–49.
10. Chow J and Watson J. Review of $PM_{2.5}$ and PM_{10} apportionment for fossil fuel combustion and other sources by the chemical mass balance receptor model. *Energy Fuels* 2002; **16**: 222–260.
11. Henry RC and Kim BM. Extension of self-modeling curve resolution to mixtures of more than three components. Part 1. Finding the basic feasible region. *Chemom. Intell. Lab. Syst.* 1989; **8**: 205–216.
12. Kim BM and Henry RC. Extension of self-modeling curve resolution to mixtures of more than three components. Part 2. Finding the complete solution. *Chemom. Intell. Lab. Syst.* 1999; **49**: 67–77.
13. Kim BM and Henry RC. Extension of self-modeling curve resolution to mixtures of more than three components. Part 3. Atmospheric aerosol data simulation studies. *Chemom. Intell. Lab. Syst.* 2000; **52**: 145–154.
14. Paatero P. Least squares formulation of robust, non-negative factor analysis. *Chemom. Intell. Lab. Syst.* 1997; **37**: 23–35.
15. Paatero P. The multilinear engine—a table-driven least squares program for solving multilinear problems, including the *n*-way parallel factor analysis model. *J. Comput. Graph. Stat.* 1999; **8**: 854–888.
16. Paatero P, Hopke PK and Song XH and Ramadan Z. Understanding and controlling rotations in factor analytic models. *Chemom. Intell. Lab. Syst.* 2002; **60**: 253–264.
17. Henry RC. Current factor analysis models are ill-posed. *Atmos. Environ.* 1987; **21**: 1815–1820.
18. Kim BM and Henry RC. Application of SAFER model to the Los Angeles PM_{10} data. *Atmos. Environ.* 2000; **34**: 1747–1759.
19. Lewis CW, Norris GA and Henry RC and Conner TL. Source apportionment of Phoenix $PM_{2.5}$ aerosol with the UNMIX receptor model. *J. Air Waste Manage. Assoc.* 2003; **53**: 325–338.
20. Lawson CL and Hanson RJ. *Solving Least-Squares Problems*. Prentice-Hall: Englewood Cliffs, NJ, 1974.
21. Malinowski ER. *Factor Analysis in Chemistry* (2nd edn) Wiley: New York, 1991.
22. Paatero P and Tapper U. Analysis of different modes of factor analysis as least squares fit problems. *Chemom. Intell. Lab. Syst.* 1993; **18**: 183–194.
23. Hopke PK, Paatero P, Jia H and Ross RT and Harshman RA. Three-way (PARAFAC) factor analysis: examination and comparison of alternative computational methods as applied to ill-conditioned data. *Chemom. Intell. Lab. Syst.* 1998; **43**: 25–42.
24. Xie Y-L, Hopke PK, Paatero P and Barrie LA and Li S-M. Identification of source nature and seasonal variations of Arctic aerosol by the multilinear engine. *Atmos. Environ.* 1999; **33**: 2549–2562.
25. Ramadan Z, Eickhout B, Song X-H and Buydens LMC and Hopke PK. Comparison of positive matrix factorization (PMF) and multilinear engine (ME-2) for the source apportionment of particulate pollutants. *Chemom. Intell. Lab. Syst.* 2003; in press.
26. Chueinta W and Hopke PK and Paatero P. A multilinear model for spatial pattern analysis of the measurement of haze and visual effects (MOHAVE) project. *Environ. Sci. Technol.* 2003; submitted.
27. Juntto S and Paatero P. Analysis of daily precipitation data by positive matrix factorization. *Environmetrics* 1994; **5**: 127–144.
28. Anttila P, Paatero P and Tapper U and Järvinen O. Application of positive matrix factorization to source apportionment: results of a study of bulk deposition chemistry in Finland. *Atmos. Environ.* 1995; **29**: 1705–1718.
29. Polissar AV, Hopke PK and Malm WC and Sisler JF. The ratio of aerosol optical absorption coefficients to sulfur concentrations, as an indicator of smoke from forest fires when sampling in polar regions. *Atmos. Environ.* 1996; **30**: 1147–1157.
30. Polissar AV, Hopke PK and Malm WC and Sisler JF. Atmospheric aerosol over Alaska: 2. Elemental composition and sources. *J. Geophys. Res.* 1998; **103**: 19045–19057.
31. Xie YL, Hopke P, Paatero P and Barrie LA and Li SM. Identification of source nature and seasonal variations of Arctic aerosol by positive matrix factorization. *J. Atmos. Sci.* 1999; **56**: 249–260.
32. Polissar AV, Hopke PK, Paatero P, Kaufman YJ, Hall DK, Bodhaine BA and Dutton EG and Harris JM. The aerosol at Barrow, Alaska: long-term trends and source locations. *Atmos. Environ.* 1999; **33**: 2441–2458.
33. Lee E and Chan CK and Paatero P. Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong. *Atmos. Environ.* 1999; **33**: 3201–3212.
34. Paterson KG, Sagady JL, Hooper DL, Bertman SB and Carroll MA and Shepson PB. Analysis of air quality data using positive matrix factorization. *Environ. Sci. Technol.* 1999; **33**: 635–641.
35. Huang S and Rahn KA and Arimoto R. Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island. *Atmos. Environ.* 1999; **33**: 2169–2185.
36. Yakovleva E and Hopke PK and Wallace L. Receptor modeling assessment of PTEAM data. *Environ. Sci. Technol.* 1999; **33**: 3645–3652 (1999).
37. Chueinta W and Hopke PK and Paatero P. Investigation of sources of atmospheric aerosol in urban and suburban residential areas in Thailand by positive matrix factorization. *Atmos. Environ.* 2000; **34**: 3319–3329.
38. Ramadan Z, Song X-H and Hopke PK. Identification of sources of Phoenix aerosol by positive matrix factorization. *J. Air Waste Manage. Assoc.* 2000; **50**: 1308–1320.
39. Polissar AV and Hopke PK and Poirot RL. Atmospheric

aerosol over Vermont: chemical composition and sources. *Environ. Sci. Technol.* 2001; **35**: 4604–4621.
40. Song X-H and Polissar AV and Hopke PK. Sources of fine particle composition in the Northeastern US. *Atmos. Environ.* 2001; **35**: 5277–5286.
41. Poirot RL, Wishinski PR and Hopke PK and Polissar AV. Comparative application of multiple receptor methods to identify aerosol sources in Northern Vermont. *Environ. Sci. Technol.* 2001; **35**: 4622–4636.
42. Paatero P and Hopke PK. Utilizing wind direction and wind speed as independent variables in multilinear receptor modeling studies. *Chemom. Intell. Lab. Syst.* 2002; **60**: 25–41.
43. Willis RD. *Workshop on UNMIX and PMF as applied to $PM_{2.5}$, 14–16 February 2000*. US EPA, RTP, NC. Report No. EPA/600/A-00/048. US Environmental Protection Agency: Research Triangle Park, NC, 2000.
44. Ashbaugh LL, Malm WC and Sadeh WZ. A residence time probability analysis of sulfur concentrations at ground canyon national park. *Atmos. Environ.* 1985; **19**: 1263–1270.
45. Poirot RL and Wishinski PR. Visibility, sulfate, and air mass history associated with the summertime aerosol in Northern Vermont. *Atmos. Environ.* 1986; **20**: 1457–1469.
46. Malm WC and Johnson CE and Bresch JF. Application of principal component analysis for purposes of identifying source–receptor relationships. In: *Receptor Methods for Source Apportionment*, Pace TG (ed). Air Pollution Control Association: Pittsburgh, PA, 1986; 127–148.
47. Gao N and Cheng M-D and Hopke PK. Potential source contribution function analysis and source apportionment of sulfur species measured at Rubidoux, CA during the Southern California Air Quality Study, 1987. *Anal. Chim. Acta* 1993; **277**: 369–380.
48. Gao N and Cheng MD and Hopke PK. Receptor modeling for airborne ionic species collected in SCAQS, 1987. *Atmos. Environ.* 1994; **28**: 1447–1470.
49. Gao N and Hopke PK and Reid NW. Possible sources of some trace elements found in airborne particles and precipitation in Dorset, Ontario. *J. Air Waste Manage. Assoc.* 1996; **46**: 1035–1047.
50. Cheng MD and Hopke PK and Zeng Y. A receptor methodology for determining source regions of particle sulfate composition observed at Dorset, Ontario. *J. Geophys. Res.* 1993; **98**: 16839–16849.
51. Cheng MD and Gao N and Hopke PK. Source apportionment study of nitrogen species measured in Southern California in 1987. *J. Environ. Eng.* 1996; **122**: 183–190.
52. Xie Y-L, Hopke PK, Paatero P and Barrie LA and Li S-M. Locations and preferred pathways of possible sources of Arctic aerosol. *Atmos. Environ.* 1999; **33**: 2229–2239.
53. Poissant L. Potential sources of atmospheric total gaseous mercury in the St. Lawrence River valley. *Atmos. Environ.* 1999; **33**: 2537–2547.
54. Polissar AV and Hopke PK and Harris JM. Source regions for atmospheric aerosol measured at Barrow, Alaska. *Environ. Sci. Technol.* 2001b; **35**: 4214–4226.