ELSEVIER

Chemometrics and Intelligent Laboratory Systems 60 (2002) 25–41

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Utilizing wind direction and wind speed as independent variables in multilinear receptor modeling studies

Pentti Paatero [a], Philip K. Hopke [b,*]

[a] *Department of Physics, University of Helsinki, Helsinki FIN-00014, Finland*
[b] *Departments of Chemical Engineering and Chemistry, Clarkson University, Potsdam, NY 13699-5705, USA*

Abstract

The problem of identifying sources of airborne pollutants and providing quantitative estimates of the contributions of each of those sources is important for airborne particulate matter. Various forms of factor analysis have been applied to this problem. However, in factor analysis, there is the fundamental problem of rotational ambiguity that makes the problem ill-posed. Thus, the incorporation of additional information can be useful in improving the solutions. Especially for identifying local sources, wind data (direction and speed) could be valuable additional information in such receptor modeling. However, wind data cannot be used directly as *dependent* variables in factor analytic modeling because the dependence of observed concentrations on wind variables is far from linear. An expanded multilinear model has been developed in which the wind direction, speed and other variables are included as *independent* variables. For each source, the analysis computes a directional profile that indicates how much of the concentrations are explained by the factors depending on wind direction, speed, and other values. This model has been tested using simulated data developed by the U.S. Environmental Protection Agency as part of a workshop to test advanced factor analysis methods. For most of the local sources, well-defined directional profiles were obtained. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Receptor modeling; Positive matrix factorization; Source apportionment; Factor analysis; Meteorological data; Particle composition data

1. Introduction

With the promulgation of new National Ambient Air Quality Standards (NAAQS) for airborne particulate matter, there is a renewed interest in improved methods to identify and quantitatively apportion sources of particle mass. Such methods will be needed in the near future to analyze the data that will be obtained through the national chemical speciation network that will begin to be deployed in 2000 and will eventually produce large quantities of elemental composition data that will need to be analyzed. The results of such source identification and apportionment efforts will then be utilized as part of the State Implementation Planning (SIP) process to develop efficient and effective strategies to bring the particulate matter concentrations into compliance with the NAAQS values. The identification and

[*] Corresponding author. Department of Chemical Engineering, Clarkson University, P.O. Box 5705, Potsdam, NY 13699-5705, USA. Tel.: +1-315-268-3861; fax: +1-315-268-6654.
E-mail address: hopkepk@clarkson.edu (P.K. Hopke).

PII: S0169-7439(01)00183-6

apportionment of pollutants to their sources is called receptor modeling.

Currently, there are two basic approaches to the receptor model problem depending on the extent of *a priori* knowledge that is available about the number and nature of the sources. If the pollution sources are known and the compositions of the emissions have been measured, then the Chemical Mass Balance (CMB) model [1,2] can be applied. This model provides a quantitative estimate of the contribution of each identified source and the corresponding uncertainty in that estimate. However, in many cases, the sources have not been identified or their emissions characterized.

If the source information is not known, multivariate receptor models [3,4] can be applied. These models estimate the number and nature of the sources from only the ambient data. However, there are limitations to the ability of factor analysis to produce unambiguous results that make the factor analysis problem ill-posed [5]. The imposition of constraints such as non-negativity of the source profile and source contribution values can reduce this rotational ambiguity. Thus, new factor analysis methods have been developed that incorporate such constraints. UNMIX developed by Henry [3,6] applies constraints externally to the eigenvector analysis used to identify the number of underlying source profiles. The model has been applied to the data from Los Angeles, CA [7].

An alternative method, Positive Matrix Factorization (PMF), uses a least squares approach to solve the factor analysis problem and can integrate the non-negativity constraints into the optimization process [8]. This approach as implemented in the programs called PMF2 and PMF3 have been applied to a number of data sets including precipitation [9,10] and urban [11,12] and remote [13–16] site particulate matter compositions. In the study of data from Alert, N.W.T., it was found that the data were best reproduced by a more complicated model [17]. In order to fit this model, Paatero [18] developed a more flexible fitting algorithm called the multilinear engine (ME) that can be applied to fit any model that can be expressed as a sum of products. The availability of this tool that can efficiently fit complex data models has enabled the construction of the present model. 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 non-linear dependency cannot be directly modeled so that wind information would be included in a factor analytic 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. The nature of the resulting expanded factor analysis model is described in the next section. The application of this model will be demonstrated using the simulated data prepared by the Environmental Protection Agency for a workshop held in February 2000 to compare the performance of UNMIX and PMF [19].

2. Data analysis

In this study, a generalization of the ordinary bilinear (factor analytic) model has been used for modeling source–receptor data. The ordinary factor analysis model can be written as

$$\mathbf{X} = \mathbf{G}\mathbf{F}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where **X** is the matrix of ambient elemental concentrations, **F** is the matrix of source profiles, **G** is the matrix of source contributions, and **E** is the matrix of residuals that are not fit by the model. In this paper, composition profiles run along columns of **F**, not along rows. This creates a consistent notation where all factor matrices are organized similarly. For this reason, **F** is transposed in Eq. (1).

To present the expanded factor analysis approach, the model is described from the viewpoint of one source, denoted by p. In reality, there are several sources and the observed concentrations are sums of contributions due to all sources, $p=1,\ldots,P$. In the customary bilinear analysis, the contribution r_{ijp} of source p on day i to concentration of chemical species j is represented by the product $g_{ip}f_{jp}$, where g_{ip} corresponds to the strength of source p on day i,

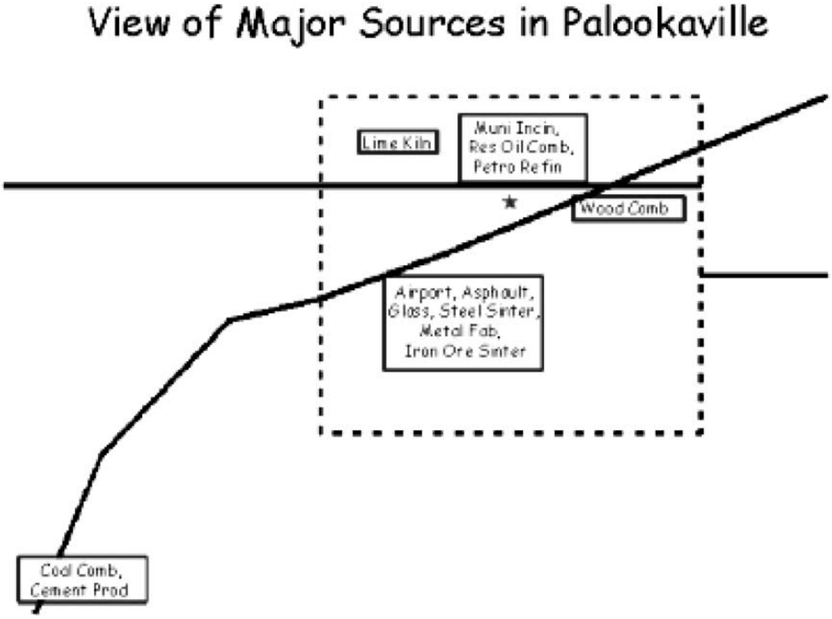

Fig. 1. Location of main sources relative to the receptor location.

and f_{jp} corresponds to the concentration of compound j in the emission signature of source p.

In the present expanded PMF analysis, the bilinear Eq. (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:

$$r_{ijp} = m_{ip} f_{jp} = \mathbf{D}(\delta_i, p)\mathbf{V}(v_i, p) f_{jp} \tag{2}$$

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. Because of 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 the 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 $\boldsymbol{v}_i$ are obtained from a chosen classification of wind speeds. The following classification was used in this work: $0-1.5-2.5-3.5-5.8-\infty$ m/s. Thus, $v_i = 2$ for such days when the average wind speed is between 1.5 and 2.5 m/s.

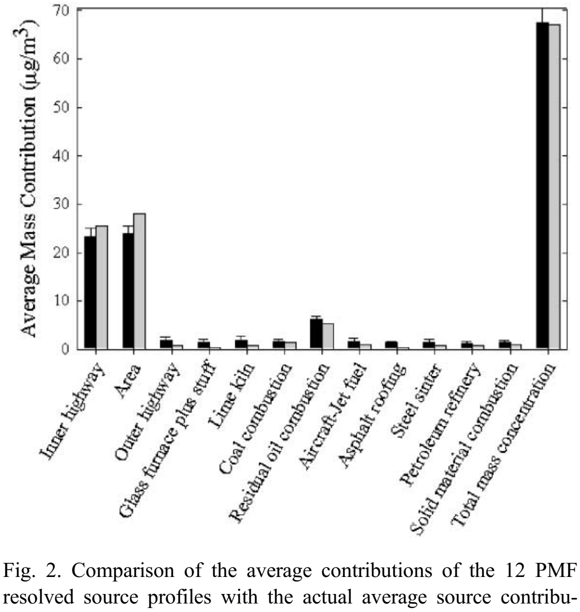

Fig. 2. Comparison of the average contributions of the 12 PMF resolved source profiles with the actual average source contributions. The results of the conventional PMF analysis are represented by the gray bars while the black bars present the known true values.

In component form, the equations of the model are:

$$
\begin{aligned}
x_{ij} &= \sum_{p=1}^{P} g_{ip} f_{jp} + e_{ij} \\
x_{ij} &= \sum_{p=1}^{P} m_{ip} f_{jp} + e'_{ij} \\
&= \sum_{p=1}^{P} \mathbf{D}(\delta_i, p)\mathbf{V}(v_i, p) f_{jp} + e'_{ij} \qquad (3)
\end{aligned}
$$

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 Eq. (3) will produce a significantly poorer fit to the data than the first set of Eq. (3). The physical model, m_{ip}, is one of 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 models

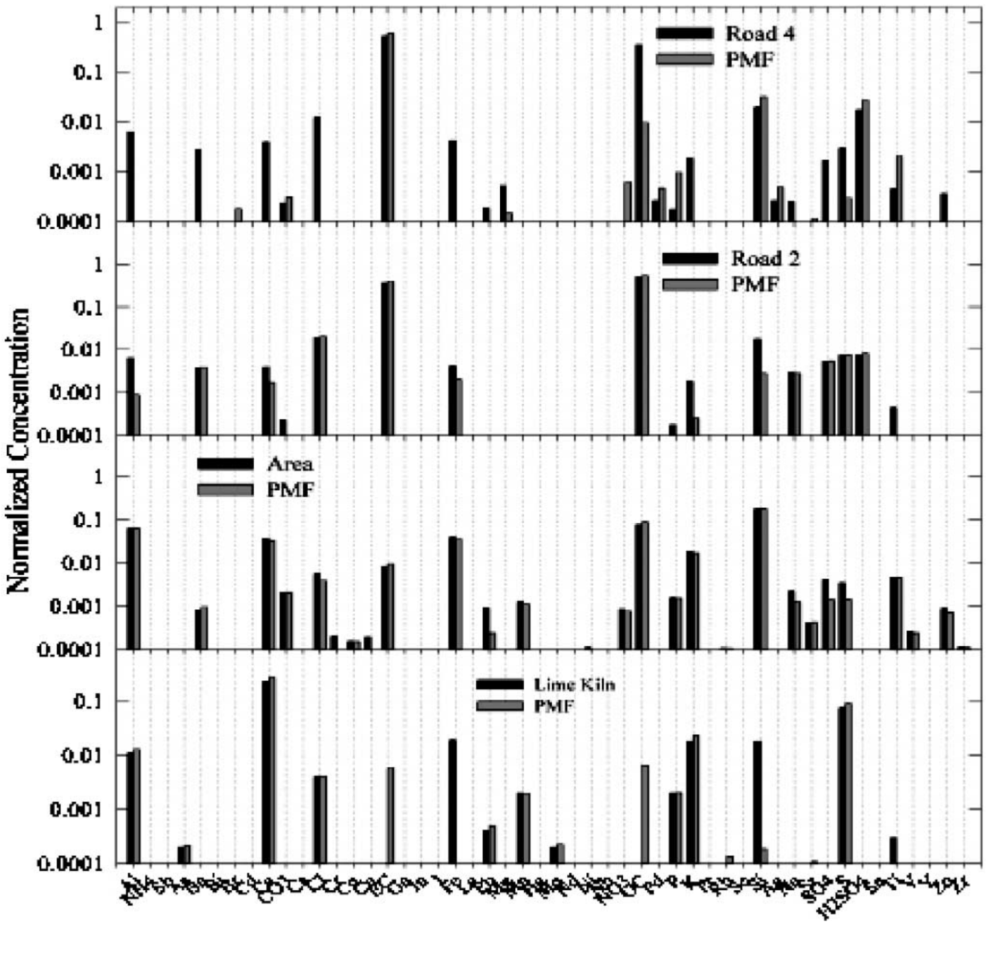

Fig. 3. Source profiles for road 4, road 2, area, and lime kiln compared to the true profiles.

fits the data as well as possible. In other words, the sum-of-squares value Q, defined by

$$Q = \sum_{i=1}^{I} \sum_{j=1}^{J} (e_{ij}/\sigma_{ij})^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (e'_{ij};/\sigma'_{ij})^2 \qquad (4)$$

is minimized with respect to the matrices **G**, **F**, **D**, and **V**, while the residuals $\boldsymbol{e}_{ij}$ and e_{ij}' are determined by Eq. (3). 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, twenty-four 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 variable σ_i (the Greek letter σ is used for two purposes: σ_{ij} indicates the error

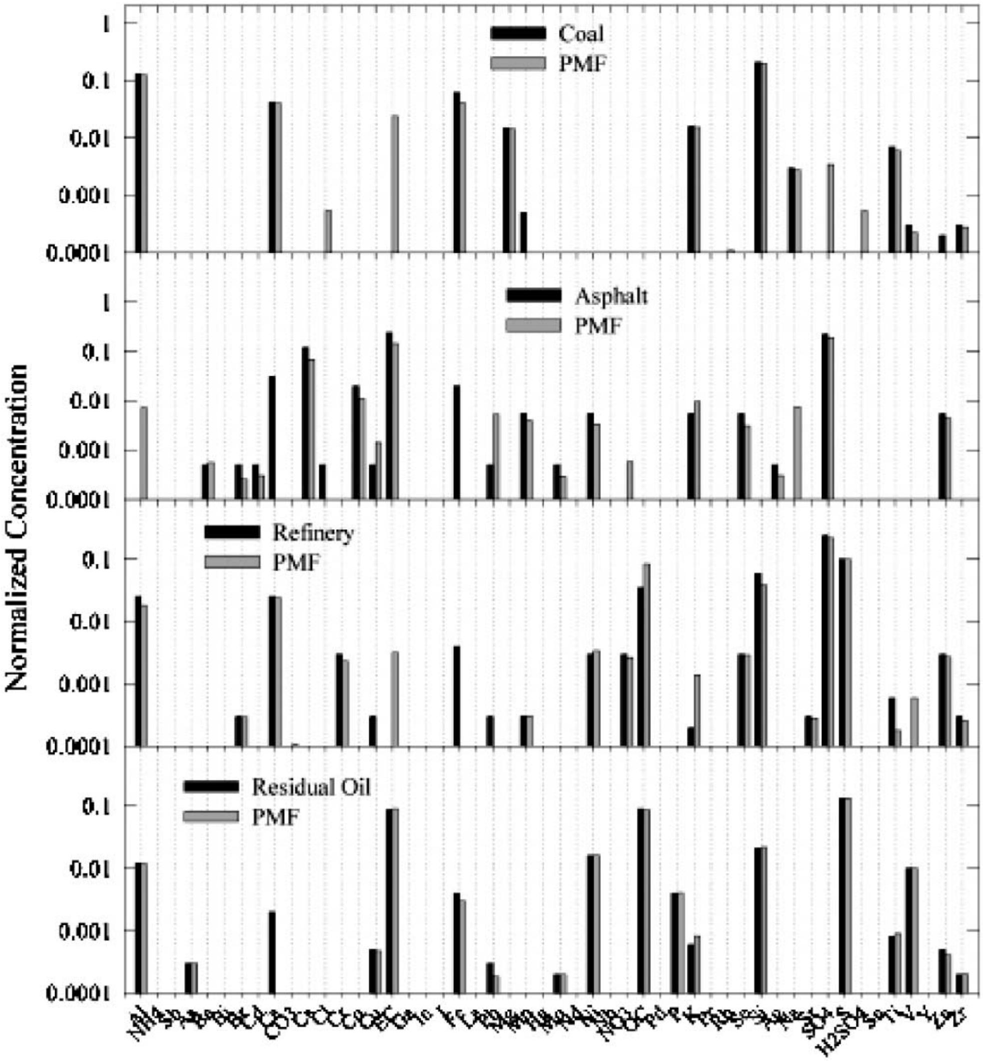

Fig. 4. Source profiles for coal, asphalt, refinery, and residual oil combustion compared to the true profiles.

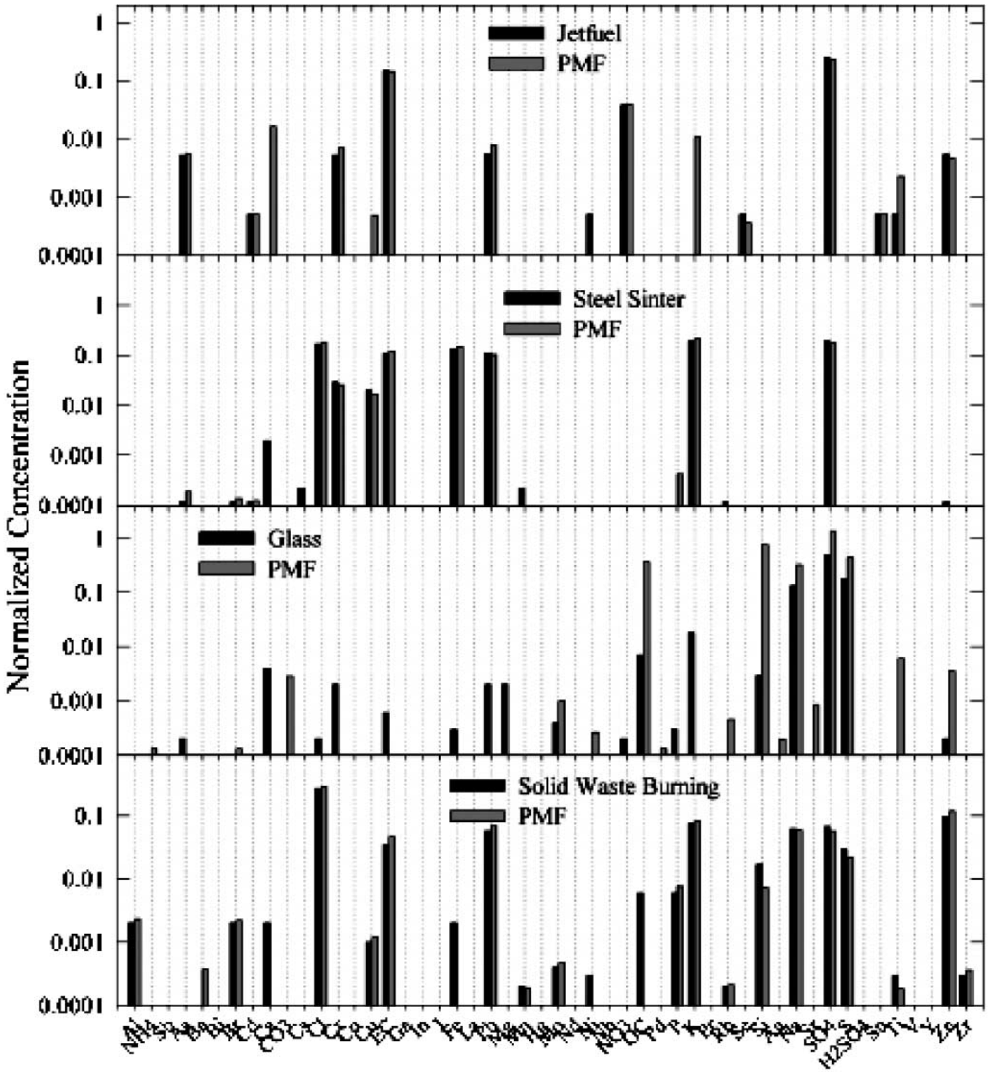

Fig. 5. Source profiles for jetfuel, steel sinter, glass, and incinerator compared to the true profiles.

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.

Instead of the basic Eq. (2), the non-linear dependencies are now defined by the following multilinear expression:

$$m_{ip} = \sum_{h=1}^{24} \mathbf{D}(\delta_{ih},p)\mathbf{V}(v_{ih},p)\mathbf{W}(\omega_i,p)\mathbf{S}(\sigma_i,p) \tag{5}$$

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. Thus, these auxiliary variables are not fitted, but serve as indicators to the values to be fitted.

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

$$x_{ij} = \sum_{p=1}^{P} g_{ip} f_{jp} + e_{ij} \tag{6}$$

$$x_{ij} = \sum_{p=1}^{P} m_{ip} f_{jp} + e'_{ij} = \sum_{p=1}^{P} \sum_{h=1}^{24} \mathbf{D}(\delta_{ih},p) \times \mathbf{V}(v_{ih},p)\mathbf{W}(\omega_i,p)\mathbf{S}(\sigma_i,p) f_{jp} + e'_{ij} \tag{7}$$

The multilinear engine (ME) was used to solve this problem with non-negativity required for all of the elements of the matrices being estimated [18]. The following values were used as input to the program: x_{ij}, the corresponding error estimates σ_{ij} and σ_{ij}' , and the index variables δ_{ih}, v_{ih}, ω_i, and σ_i.

3. Data description

Sixteen distinct source profiles were used in Palookaville simulation—nine point sources, four industrial complexes, one area source, and two highways. The layout of the sources is shown in Fig. 1. 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 variability of approximately

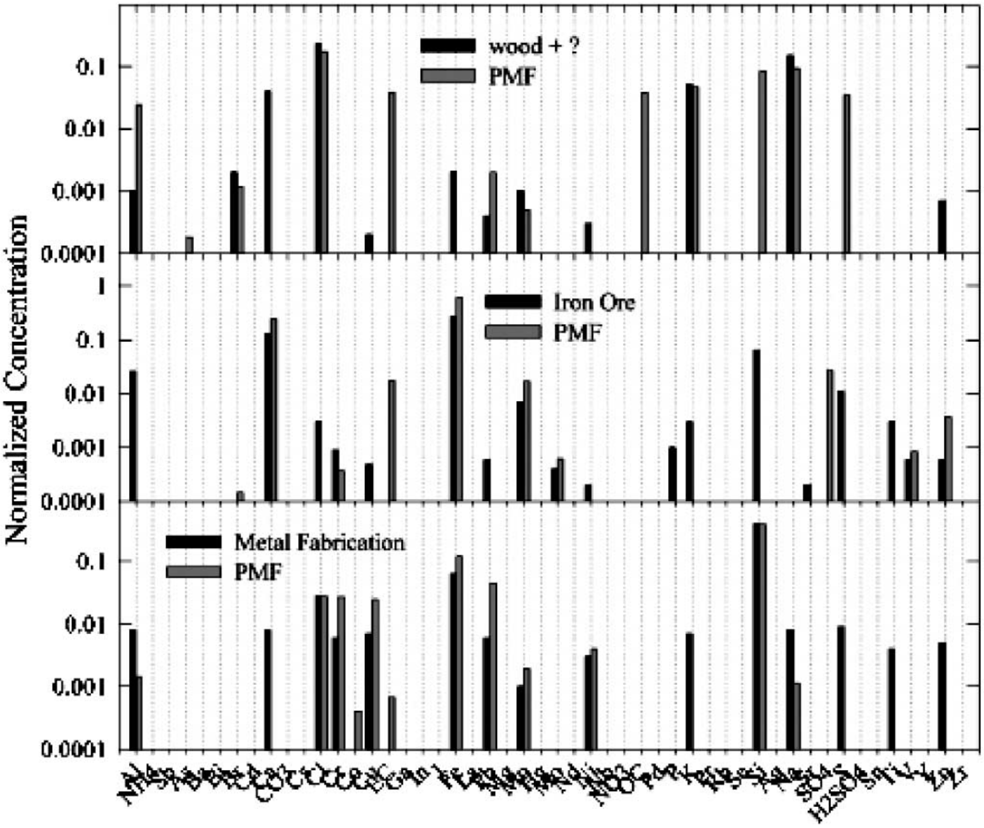

Fig. 6. Source profiles for wood combustion, iron ore, and metal fabrication compared to the true profiles.

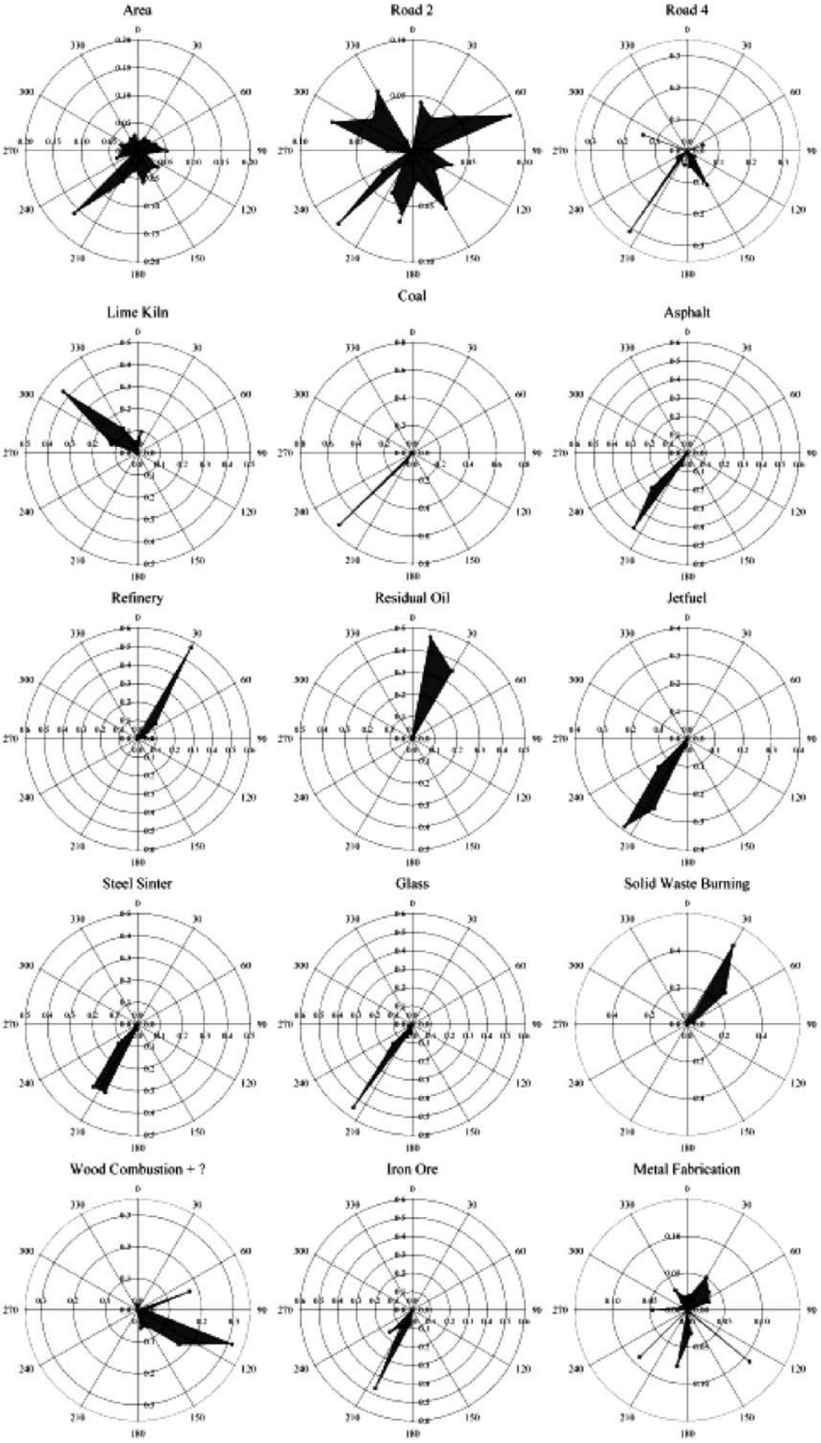

Fig. 7. Wind direction factors for each of the sources. The columns of matrix **D** are plotted in polar coordinates.

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.

The data matrix was augmented by one artificial variable: all identified concentration were subtracted from the measured mass concentration. The resulting variable might be called *unidentified mass*. The presence of such variable is useful for limiting the range of possible rotations: such rotations are prevented that would cause the concentration of unidentified mass to go negative in some factor(s).

4. Practical details

4.1. Weekday/weekend factors

The weekday/weekend factors has been simplified as follows. In principle, the matrix **W** has dimension 2 by *p*. The first row specifies the coefficients for weekdays, and the second row for weekend days. In this work, the weekday coefficients have been fixed at unity so that they may be omitted from the actual equations. Then only the second row remains in effect. Its elements specify the average strength of each factor on weekend days, relative to the strength in weekdays.

4.2. Modification of the equations to accommodate modeling errors in strong factors

In this data set, there are three strong non-directional sources that are not very well described by the physical model in Eq. (5). It turned out that it was impossible to determine the weakest (15th) factor if Eq. (7) was used in its original form. When the number of factors was increased from 14 to 15, one of the strongest factors split in two because of the poor fit. For this reason, Eq. (7) was modified to be

$$
\begin{aligned}
x_{ij} &= \sum_{p=1}^{P} (\lambda_p g_{ip} + (1-\lambda_p) m_{ip}) f_{jp} + e'_{ij} \\
&= \sum_{p=1}^{P} \left(\lambda_p g_{ip} + (1-\lambda_p) \sum_{h=1}^{24} \mathbf{D}(\delta_{ih}, p) \right. \\
&\qquad \left. \times \mathbf{V}(v_{ih}, p) \mathbf{W}(\omega_i, p) \mathbf{S}(\sigma_i, p) \right) f_{jp} + e'_{ij} \qquad (8)
\end{aligned}
$$

For all but the strongest sources, the coefficients λ_p were set to zero. For the strongest factors, $\lambda_p = 0.8$ was used. In this way, the less-than-perfect model of the strong factors did not mask the 15th factor. This problem will exist for any system in which some sources have strong wind directional dependence and some do not. The separation of the object function through the use of the λ_p reduces the effect of the wind directionality in the model on the non-directional sources.

4.3. Computation in stages

It is common to start a multilinear analysis at a pseudorandom starting point. In this work, it was necessary to run in stages. First, an initial analysis was computed with a smaller number of factors, typically 13, by using a pseudorandom start. Then the strongest factors were identified and non-zero values were assigned to the corresponding coefficients λ_p. The number of factors was increased by one, and the new factor was initiated by using pseudorandom values. The old factors were started from the values that resulted from the previous computation. After computing the results, the number of factors was again increased by one. This was continued until the emerging factor was not meaningful.

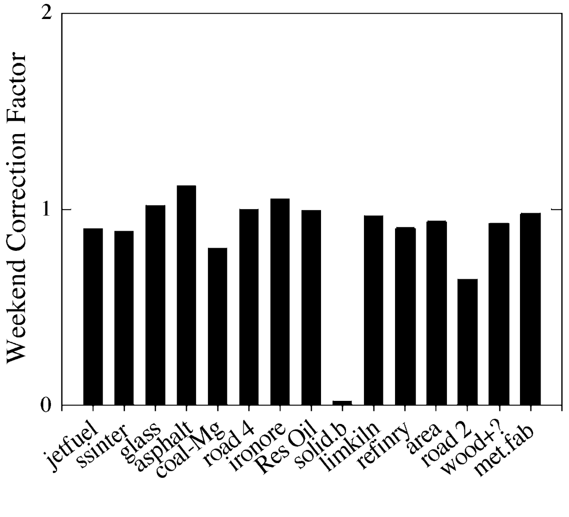

Fig. 8. Weekend/weekday correction factor (values on the second row of matrix **W**).

4.4. The scale for wind directions

When the use of wind information was discussed above, it was suggested that wind direction is classified into intervals of uniform width, typically 10° or 20°. It is, however, quite possible to use intervals of varying widths, in analogy with classifying wind speeds. In this data, the point sources with directions between 200° and 240° appear with very sharp directional definition. For this reason, 20 directional intervals were chosen, most of them having width of 20°. Between 200° and 240°, four intervals of width 10° were specified.

4.5. Choosing the error estimates

For the bilinear Eq. (6), the error estimates σ_{ij} were set according to the noise that was introduced in the simulation. The relative level of this noise (expressed in percent of the true values) was specified in simulation description. In the description, detection limits were also specified for all elements. In this work, it is assumed that error estimates of low concentrations must not be smaller than one-third of the specified detection limit. With real data, this would be normal practice. However, in this simulated data set, the relative noise level remains the same also for lowest

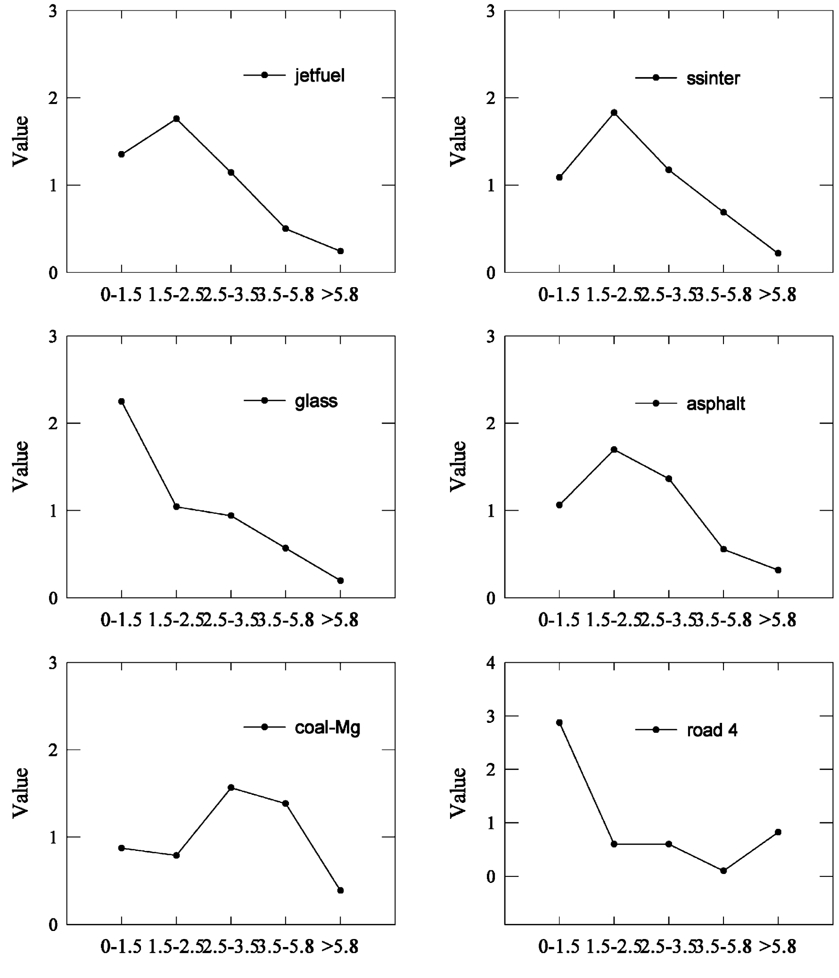

Fig. 9. Wind speed factors (columns of matrix **V**) for jetfuel, steel sinter, glass, asphalt, coal, and road 4.

concentrations (this fact was revealed after the workshop). Once this fact is known, one could extract more information from the data by decreasing the error estimates of the lowest concentrations. Such modification was not attempted in this study because it would never work with any real data. The lessons from the simulation study will be most valuable if one avoids such techniques that never succeed with real measurements.

For the multilinear Eqs. (7) or (8), the error estimates σ_{ij}' were specified as a fixed multiple of the corresponding bilinear values σ_{ij}: $\sigma_{ij}' = 8\sigma_{ij}$. The multiplier (= 8) was chosen so that the contribution to **Q** from the bilinear equations was three times the contribution from the expanded model equations. However, the value of the multiplier is not critical. Practically the same results would be obtained with values ranging from 7 to 10, say.

4.6. Regularization

The data suffer from the fact that concentrations are integrated over 24 h. It is not possible to attribute the collected concentrations to different hours of the day with certainty. Solving the model is an *ill-posed problem.* It was noticed during this study that different factor values might produce practically the same fit to data. In order to avoid spurious results, it is necessary to *regularize* the model. The following regularization was applied in the present work.

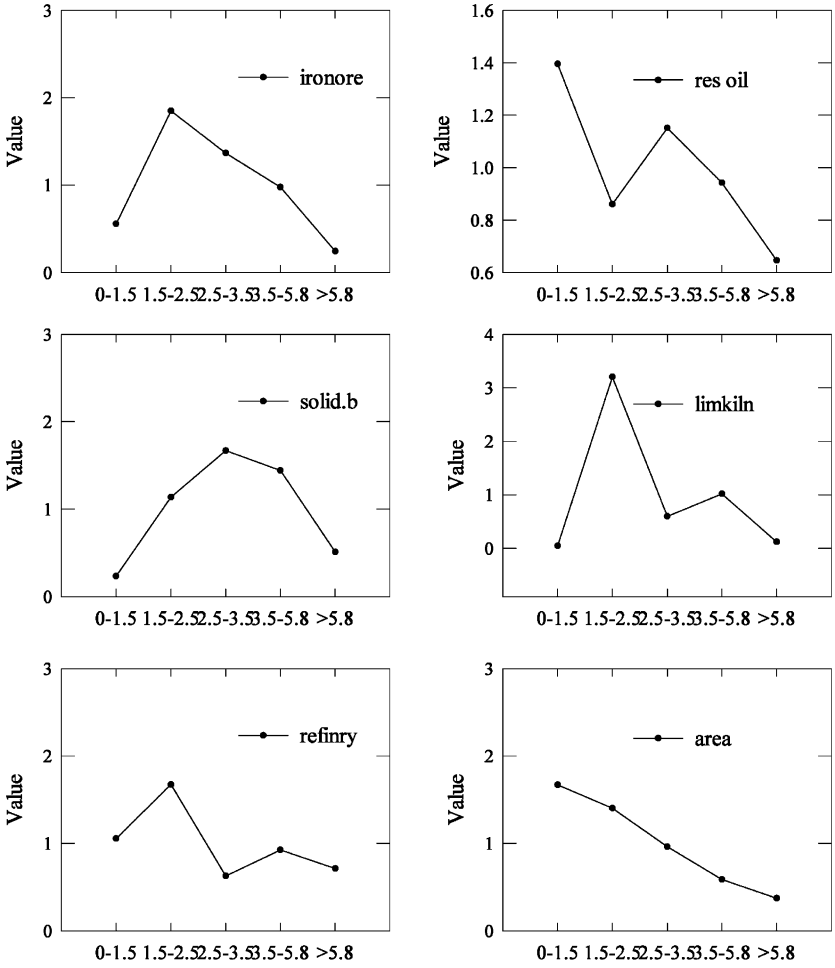

Fig. 10. Wind speed factors (columns of matrix **V**) for iron ore, residual oil combustion, incinerator, lime kiln, refinery, and area.

The seasonal factors $\mathbf{S}(k,p)$ and the weekend coefficients $\mathbf{W}(2,p)$ were pulled towards unity. In other words, auxiliary equations were added to the model, such as

$$\mathbf{S}(k,p) = 1 \quad (k = 1,\ldots,6, \quad p = 1,\ldots,P) \qquad (9)$$

The error estimates connected to these equations were specified so that the contributions to **Q** from these equations were a few percent of the contributions from main equations. In other words, introduction of these equations was not allowed to worsen the fit noticeably. Nevertheless, the computed factors changed clearly. Such variation disappeared from the seasonal and weekend factors that was not essential for achieving a good fit.

4.7. The multilinear engine script

The details described above were implemented as commands in the script that guides the operations of the multilinear engine program. The script should be understandable for anybody with programming background in BASIC, Fortran, or C. The script is available from the authors. The program was run on a PC computer equipped with a Pentium II processor and 96 MB of memory. Typically, between 1000 and 2000 iteration steps were needed for convergence. One step took approximately 1 s.

5. Results

The conventional PMF analysis of the Palookaville data produced a 12-factor solution that clearly identified the major sources [19]. For the main sources, the source profiles computed by PMF produced an excellent match to the true profiles used in the simulation. However, several profiles of the minor sources were less well reproduced and some of the mass was apportioned from the most significant sources to some of the minor sources. The comparison of the conventional factor analysis to the known average source strengths is shown in Fig. 2.

Using the expanded wind-based model, it was possible to extract 15 of the 16 source profiles employed in the simulation. Only the cement plant could not be separated. These resulting source profiles

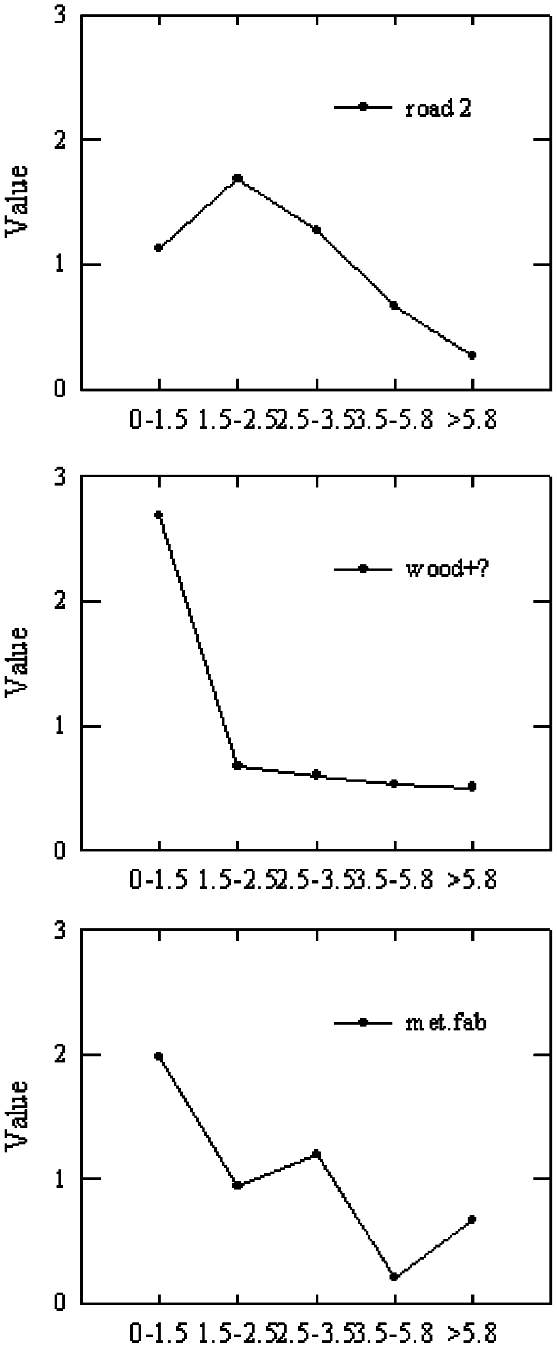

Fig. 11. Wind speed factors (columns of matrix **V**) for road 2, wood combustion, and metal fabrication.

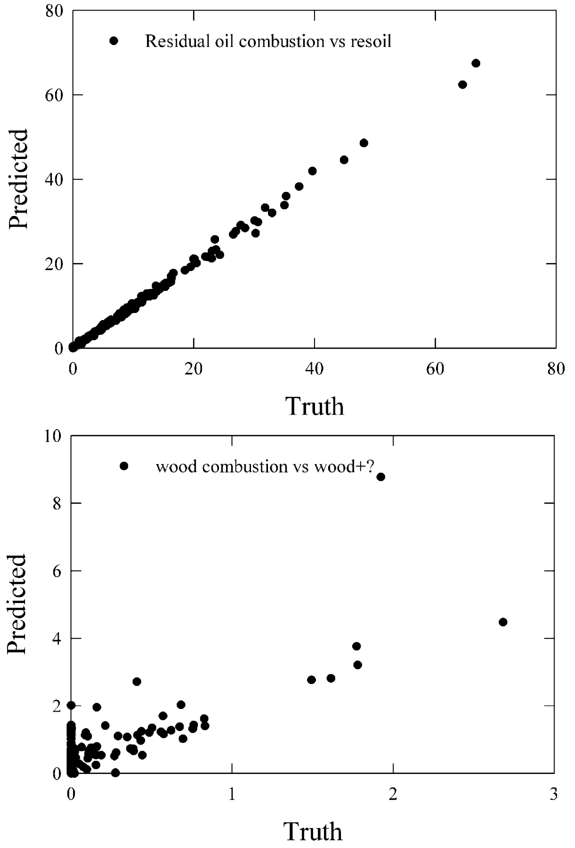

Fig. 12. Predicted versus true mass contributions for residual oil and wood combustion.

can be compared to the known profiles used to prepare the data. These comparisons are shown in Figs. 3–6. Thus, this analysis has resolved an additional three of the weak sources. In general, there is a good agreement between the extracted and the true profiles. There are cases of a number of specific elements in particular profiles that are over- or underpredicted. This result is particularly important given that no specific rotations were imposed on the results. It is the expanded modeling that reduces the rotational ambiguity in the problem.

The wind direction factors, the columns of matrix **D**, are displayed in Fig. 7. The point sources (lime kiln, coal, asphalt roofing, petroleum refinery, residual oil combustion, jet fuel combustion, steel sinter, glass furnace, municipal incinerator, wood combustion, iron ore dust, and metal fabrication) generally show strong directional behavior that agrees well with the distribution of the sources. These results are extremely encouraging in terms of being able to identify the direction from which the source materials arrives at the receptor site. For the distributed sources (area, road 2, and road 4), there is still some defined directionality. The major road source (road 2) is mainly along the roads that cross just to the south of the receptor site. Thus, the major wings in the road 2 directional pattern point along these four directions. For the area source that is uniformly distributed around the receptor site, there are also preferred directions in the wind direction. The directionality shown in the area source figure may reflect these features in the data formulation process. However, one cannot exclude the possibility that the directionality of the area source is an artefact caused by too little regularization being applied.

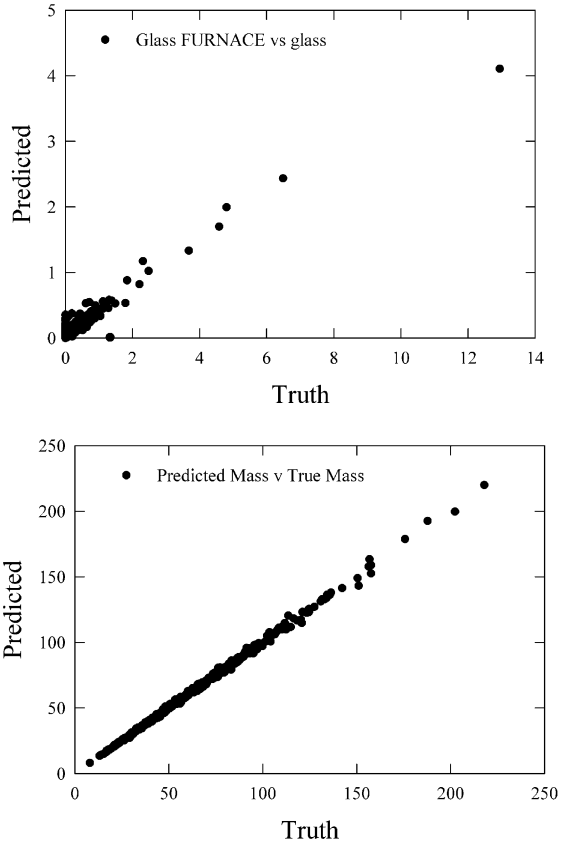

Fig. 13. Predicted versus true mass contributions for glass and the total mass values.

The weekend/weekday correction factors, corresponding to the relative strength of each source on weekend days, are shown in Fig. 8. The incinerator was set to be off on weekends and it can be seen that the correction factor for this source is almost zero. There was also reduced traffic on the major road (road 2) on weekend days, in agreement with what was built into the data creation process.

Figs. 9–11 show the profiles for the wind speed factor, columns of matrix **V**. Many of these factors have very low values for the lowest wind speed range, indicating weak transport from the point sources to the receptor site. The general trend is that these factors decrease with increasing wind speed. This trend is explained by a dilution effect: with increasing wind speed, the same emitted mass is distributed to a larger volume of air so that the concentration decreases. Without having more specific details on the preparation of these data, it is not possible to compare these results to the true relationships used to create the data. However, in general, the results seem sensible. For the sources "wood" and "metal fabrication", the computed wind speed dependency appears unrealistic. This is not surprising because

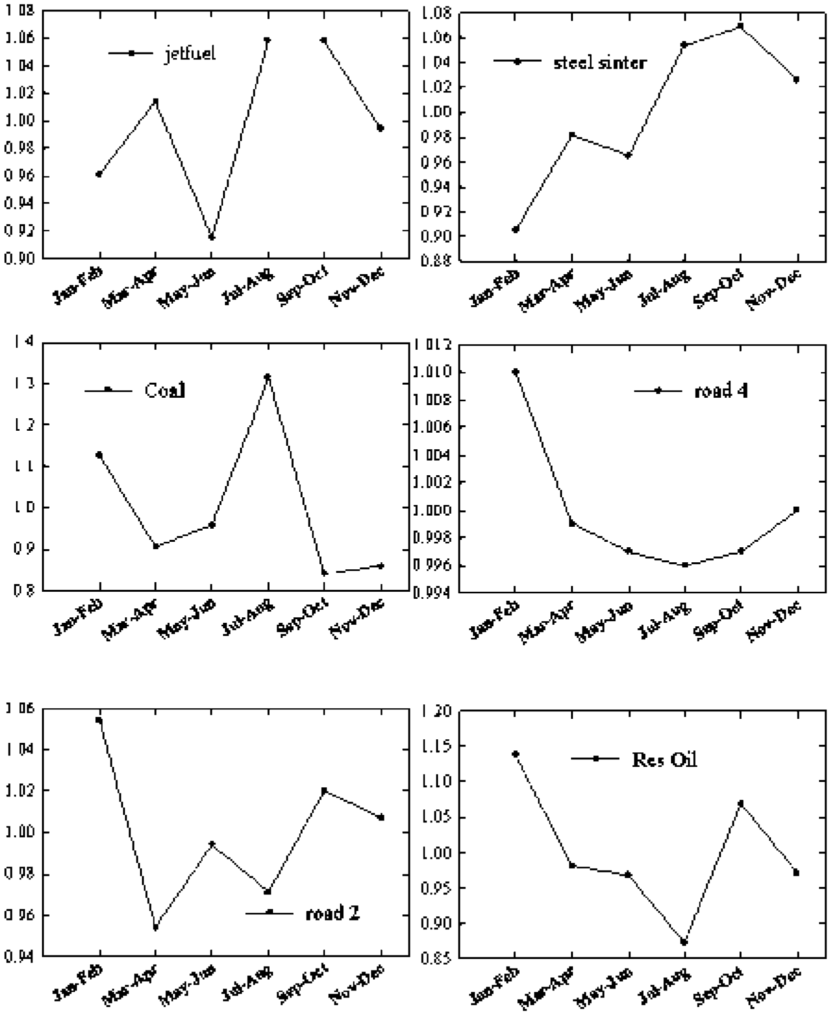

Fig. 14. Seasonal factors for jetfuel, steel sinter, coal, road 4, road 2, and residual oil.

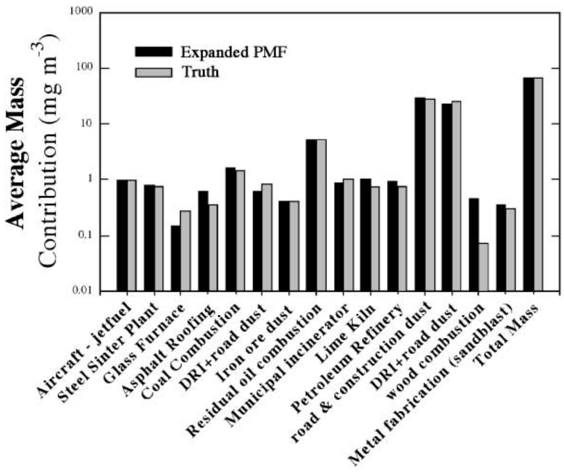

Fig. 15. Average mass apportionment by the present analysis compared with the true average mass contributions.

these factors are so weak that their identification is just barely possible.

Figs. 12 and 13 show several of the seasonal or time-of-year patterns (= columns of matrix **S**). Again, it is not possible to directly compare these results with the data creation process.

Figs. 14 and 15 show the contribution values for the individual samples for several of the sources. The parameters describing the quality of the reproduction of the true contributions values are provided in Table 1. It can be seen that, generally, the relationship of the estimated contributions for individual samples with the true values has a very high correlation with many of the r^2 values approaching 1. For the sources making significant contributions, the slopes are generally close to 1. The ability to accurately resolve the weaker sources depends on how distinctive the source profiles are relative to the other sources and the level of uncertainty in the data.

The apportionment of the average aerosol mass to the sources is shown in Fig. 15. It can be seen that for all of the higher contribution sources, there is excellent agreement between the estimated and the true contributions.

5.1. Interpretation of results

The independent variables do not necessarily occur randomly in arbitrary combinations. During different times of year, wind may come predominantly from different directions or average wind speeds may be different. Thus, some part of variation in source strengths might be explained alternatively by the seasonal factors or by the wind-related factors. Similarly, wind speed and wind direction may be correlated. Because of such correlations, caution is needed when interpreting the results.

Table 1
Parameters describing the predicted versus the true source contributions

Source	Mass (μg^{-3})	Intercept	Error	Slope	Error	r	r^2
Aircraft	1.01	0.04	0.01	0.999	0.003	0.998	0.996
Asphalt	0.62	0.04	0.01	1.693	0.012	0.992	0.983
Coal	1.62	0.06	0.02	1.066	0.004	0.997	0.994
Road 4	0.62	0.18	0.01	0.528	0.009	0.946	0.895
Road 2	23.35	0.03	0.13	0.914	0.004	0.996	0.993
Glass	0.15	0.06	0.00	0.340	0.004	0.970	0.941
Iron ore	0.41	0.20	0.01	0.504	0.015	0.864	0.746
Lime kiln	1.02	0.42	0.03	0.781	0.013	0.950	0.903
Metal fabrication	0.34	0.18	0.02	0.548	0.025	0.756	0.572
Incinerator	0.89	0.04	0.01	0.839	0.002	0.999	0.997
Residual oil	5.40	0.06	0.03	0.996	0.002	0.999	0.998
Refinery	0.95	0.10	0.02	1.086	0.008	0.991	0.982
Area	29.67	1.2	0.2	1.010	0.005	0.996	0.992
Steel sinter	0.79	0.12	0.01	0.888	0.009	0.982	0.964
Wood combustion	0.44	0.29	0.02	2.03	0.08	0.811	0.658
Total mass	67.28	0.12	0.18	1.000	0.002	0.999	0.998

Wind speed is correlated with mixing properties of the atmosphere. Thus, some effects attributed to wind speed might in fact be caused by variations in mixing height or by inversion situations.

Variations observed at receptor site may be due to variations in source strength or variations in transport path. This ambiguity should be taken into account when interpreting the seasonal (time-of-year) factors. The variation shown by a seasonal factor may be caused by variations in the activity of the source. For example, a coal-fired power plant may be off during low demands of electricity. On the other hand, weather conditions may be different during different parts of the year, thus influencing the efficiency of the transport path from the source to the receptor site. Such differences of transport efficiency can be represented by the seasonal factors. In summary, any variation seen in the seasonal factors may indicate seasonal variation of source strength or of transport conditions or of both.

6. Conclusions

The customary bilinear factor analytic model is enhanced so that a structural expanded factor model is fitted simultaneously with the original bilinear model. The structural model reduces the rotational ambiguity of the solution. In addition, the structural factors, such as wind direction dependence, aid in identifying the sources that correspond to the computed factors.

Two meteorological variables (wind speed and wind direction) plus two calendar variables (the season of each observation, and the weekday/weekend status of each observation) are used as independent variables in the structural model. Each observation can be envisioned as being mapped into this four-dimensional space; weather data and calendar information determine the mapping. For each source, its dependence on these four variables is determined when the model is fitted. A fully unique solution is not expected because of two reasons: (1) the four independent variables of the structural model are correlated; (2) concentrations are integrated over 24 h, thus losing much detail connected with the meteorological variables. By applying regularization, a unique ‘‘smooth’’ solution was obtained, at the expense of possibly losing some detail.

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. 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.

This analysis was based on 24-h concentrations and 1-h weather data. The success of the analysis demonstrates that the high-resolution weather data may significantly enhance the usefulness of 24-h concentration data. Recording high-resolution weather data costs much less than gathering high-resolution concentrations. It is suggested that a cost-optimal measuring strategy should record frequent and comprehensive weather information even if the concentrations are integrated over longer times.

Acknowledgements

We would like to thank Charles Lewis, Shelly Eberly and Gary Norris of the U.S. Environmental Protection Agency for providing the data and their very useful comments on the results of these analyses. We would also like to thank Dr. Ron Henry of the University of Southern California for useful discussions of these results during the EPA-sponsored workshop on factor analysis methods held in February 2000. This work was supported by the U.S. Environmental Protection Agency under contracts 9D-1808-NTEX and 0D-6202-NAEX.

References

[1] J.A. Cooper, J.G. Watson, J.J. Huntzicker, The effective variance weighting for least squares calculations applied to the mass balance receptor model, Atmos. Environ. 18 (1984) 1347–1355.

[2] J.G. Watson, J.C. Chow, T.G. Pace, Chemical mass balance, in: P.K. Hopke (Ed.), Receptor Modeling for Air Quality Management, Elsevier, Amsterdam, 1991, pp. 83–116.

[3] R.C. Henry, Multivariate receptor models, in: P.K. Hopke (Ed.), Receptor Modeling for Air Quality Management, Elsevier, Amsterdam, 1991, pp. 117–147.

[4] P.K. Hopke, Receptor modeling for air quality management, in: R.E. Hester, R.M. Harrison (Eds.), Issues in Environmental Science, Issue 8, Royal Society of Chemistry, Cambridge, UK, 1997, pp. 95–117.

[5] R.C. Henry, Current factor analysis models are ill-posed, Atmos. Environ. 21 (1987) 1815–1820.

[6] B.-M. Kim, R.C. Henry, Extension of self-modeling curve resolution to mixtures of more than three components: Part 2. Finding the complete solution, Chemom. Intell. Lab. Syst. 49 (1999) 61–77.

[7] B.M. Kim, R.C. Henry, Application of SAFER model to the Los Angeles PM_{10} data, Atmos. Environ. 34 (2000) 1747–1759.

[8] P. Paatero, Least squares formulation of robust, non-negative factor analysis, Chemom. Intell. Lab. Syst. 37 (1997) 23–35.

[9] S. Juntto, P. Paatero, Analysis of daily precipitation data by positive matrix factorization, Environmetrics 5 (1994) 127–144.

[10] P. Anttila, P. Paatero, U. Tapper, O. Järvinen, Application of positive matrix factorization to source apportionment: results of a study of bulk deposition chemistry in Finland, Atmos. Environ. 29 (1995) 1705–1718.

[11] W. Chueinta, P.K. Hopke, P. Paatero, Investigation of sources of atmospheric aerosol at urban and suburban residential areas in Thailand by positive matrix factorization, Atmos. Environ. 34 (2000) 3319–3329.

[12] E. Lee, C.K. Chan, P. Paatero, Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong, Atmos. Environ. 33 (1999) 3201–3212.

[13] K.G. Paterson, J.L. Sagady, D.L. Hooper, S.B. Bertman, M.A. Carroll, P.B. Shepson, Analysis of air quality data using positive matrix factorization, Environ. Sci. Technol. 33 (1999) 635–641.

[14] A.V. Polissar, P.K. Hopke, W.C. Malm, J.F. Sisler, Atmospheric aerosol over Alaska: 2. Elemental composition and sources, J. Geophys. Res. 103 (1998) 19045–19057.

[15] A.V. Polissar, P.K. Hopke, P. Paatero, Y.J. Kaufman, D.K. Hall, B.A. Bodhaine, E.G. Dutton, J.M. Harris, The aerosol at Barrow, Alaska: long-term trends and source locations, Atmos. Environ. 33 (1999) 2441–2458.

[16] Y.L. Xie, P. Hopke, P. Paatero, L.A. Barrie, S.M. Li, Identification of source nature and seasonal variations of arctic aerosol by positive matrix factorization, J. Atmos. Sci. 56 (1999) 249–260.

[17] Y.-L. Xie, P.K. Hopke, P. Paatero, L.A. Barrie, S.-M. Li, Identification of source nature and seasonal variations of arctic Aerosol by the multilinear engine, Atmos. Environ. 33 (1999) 2549–2562.

[18] P. Paatero, The multilinear engine—a table-driven least squares program for solving multilinear problems, including the *n*-way parallel factor analysis model, J. Comput. Graphical Stat. 8 (1999) 854–888.

[19] R.D. Willis, Workshop on UNMIX and PMF as applied to $PM_{2.5}$, 14–16 February 2000, U.S. EPA, RTP, NC, U.S. Environmental Protection Agency, Report No. EPA/600/A-00/048, Research Triangle Park, NC, 2000, 26 pp.