Analysis of PM2.5 large-area measurements by using extended (multilinear) factor analytic models

Pentti Paatero, University of Helsinki, Finland,
email: Pentti.Paatero@Helsinki.fi
Shelly Eberly, US EPA,
Philip K. Hopke and Janjira Hoppenstock, Clarkson University, Potsdam N.Y., USA.

Nov. 30th, 2001.

Introduction

Factor analytic models have been used for receptor modeling for a long time, first based on the Principal Component Analysis (PCA) approach. Recently models based on the non-negatively constrained Positive Matrix Factorization (PMF) approach have gained popularity. In the analysis of temporal-spatial data, factor analytic models have been called Empirical Orthogonal Functions (EOF). Such models consider the data as organized in a two-way matrix. One dimension (row numbers) corresponds to time while the second dimension (column numbers) corresponds to the two spatial dimensions x and y. Data on one row of the matrix correspond to observations made in a single day at all locations of the measuring network. Similarly, data on one column represent the time series measured at one location. In the EOF approach, the matrix is analyzed similarly as in PCA, although the name of the method is different, probably because of historical reasons.

The data matrix is denoted by **X** and its elements by x_{ij}, i=1,...,I, j=1,...,J. The equation for EOF can be written in matrix form as

$$\mathbf{X} = \mathbf{G}\mathbf{F}^T + \mathbf{E} \qquad (1.1)$$

In component form, the equation is

$$x_{ij} = \sum_{p=1}^{P} g_{ip} f_{jp} + e_{ij} \qquad (1.2)$$

The columns of matrix **F** represent spatial coefficients, each column of **G** contains a time series, and **E** is the matrix of residuals that are not fitted by the model. The task of the fitting is to determine the elements of **G** and **F** so that the norm of **E** is minimized. Each pair of corresponding columns of **G** and **F** represents a ”concentration field”, an entity that has a certain time behavior and a certain spatial pattern. In the framework of EOF, the columns of **G** are constrained to be orthogonal to each other, and similarly the columns of **F**. The fitting is usually achieved by Singular Value Decomposition (SVD).

The equations (1.1) and (1.2) are also fundamental in the approach called Positive Matrix Factorization (PMF) (Paatero 1997). The additional constraints are different, however. In PMF, the elements of **G** and **F** are usually required to be non-negative. Also, the weight of each data value x_{ij} is individually adjustable in the definition of the norm that is to be minimized. Application of PMF to PM2.5 spatial-temporal data has been successful (Eberly and Cox, to be published).

The present work describes an enhanced model that utilizes auxiliary or *parametric variables* that have been measured concurrent with the PM2.5 data values. Technically, the enhanced model is an example of *quasi-multilinear* fitting. In practice, the fitting of the model is based on the program Multilinear Engine (ME-2) (Paatero 1999). Description of the mathematical model is written in the form of a *script*. The program ME-2 reads the script and fits the model by using the information in the script.

Terminology

Time and location are *independent variables* in this modeling. The concentration of PM2.5 is a *dependent variable*, meaning that the values of PM2.5 are not chosen at will but depend on the values chosen for location and time.

The present work builds on the methodology that was first published by Paatero and Hopke (2001). In that paper, a number of auxiliary variables, such as wind speed and wind direction, were used for improving the factor analytic model of source apportionment. In that paper those variables were called independent. Such terminology may be confusing because the values of auxiliary variables do depend on place and time. In the present paper, a different term is adopted, based on the idea that on one hand, the auxiliary variables do depend on place and time, while on the other hand they do influence the dependent variables. Hence they are called *parametric variables*. Similarly, the term *parametric factors* is reserved for such factor elements that represent the dependence of the independent variable PM2.5 on values of parametric variables.

The experimental data to be analyzed

PM2.5 measurements made at every third day of year 2000 were analyzed. The domain of measurements consists of the area between 32 and 43 degrees North and 72 and 96 degrees West. Within this area, 304 locations were included in the data set. Some of these locations consist of a single station, while others represent averages of several stations situated within a single grid cell. There are no missing values in PM2.5 data. Of the 304 locations, 82 were classified as rural, the rest as urban.

In this work, the following parametric variables were used: 24h-average temperature T, 24h-average specific humidity Q , 24h-average pressure P (= deviation from average pressure at the location), ozone daily maximum 8-hour average concentration Z, and 6AM-9AM average wind velocity vector (V_X,V_Y). Ozone data is provided for May-October, inclusive, only. The meteorological variables were not monitored at the geographical locations of the PM2.5 stations. Instead, the nearest available met station was connected to each PM2.5 location. A very small number of missing values occur for some met stations. The corresponding PM2.5 data have been omitted from the analysis as if they are missing values.

The quasi-multilinear model for PM2.5 distributions

The mathematical model is defined by the following general equation:

$$x_{ij} = \sum_{p=1}^{P} m_{ijp} g_{ip} f_{jp} + e_{ij} = \sum_{p=1}^{P} r_{ijp} + e_{ij} \qquad (2.1)$$

The matrices **G** and **F**, consisting of unknown factor elements, have a similar meaning as in the 2-way factor analytic model. In contrast, the values m_{ijp} are not unknown adjustable values but functionals that depend on the values of parametric variables at site j on day i. The functionals can in principle be defined in different ways. In this work, the dependence is defined as a product of effects related to different parametric variables:

$$m_{ijp} = \mathbf{T}_p(T_{ij})\, \mathbf{Q}_p(Q_{ij})\, \mathbf{P}_p(P_{ij})\, \mathbf{Z}_p(Z_{ij})\, \mathbf{V}_p(V_{ijX}, V_{ijY}) \qquad (2.2)$$

The definitions of the first four functionals are structurally similar. For each factor, each functional has a unique definition. In contrast, each definition is the same for all sites and for all times. Intuitively this can be understood so that the shapes of functionals attempt to mirror laws of physics and chemistry. Those laws are the same for all times and all places. During the iteration, the shapes of the functionals are determined numerically so that a best possible fit (smallest possible values of the residuals) are achieved. In practice, the functionals are implemented as tables that define the value of any functional, e.g. $\mathbf{T}_1$, for all values of temperature T, creating a dependence $\mathbf{T}_1(T)$. The values contained in the tables that determine the

functionals are called *parametric factors*. In the fitting algorithm, they are treated similarly as the usual factor elements g_{ip} and f_{jp}.

The fifth functional **V** is slightly different: it defines each one of the p values $\mathbf{V}_1$ to $\mathbf{V}_p$ as a two-way table, depending on the two wind velocity components (V_X, V_Y).

The functionals $\mathbf{Z}_p(Z)$, describing the dependence of concentration of PM2.5 on ozone concentration, are omitted from equation (2.2) for those days (in winter) when ozone data is not available.

The fitting of the model means that the sum-of-squares expression Q be minimized. Assuming that the individual errors e_{ij} are statistically independent, the expression is defined as

$$\begin{aligned} Q = Q_m + Q_a &= \sum_{i=1}^{I}\sum_{j=1}^{J}(e_{ij}/\sigma_{ij})^2 \quad + Q_a \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\left(x_{ij} - \sum_{p=1}^{P} m_{ijp} g_{ip} f_{jp}\right)/\sigma_{ij}\right)^2 \quad + Q_a \end{aligned} \tag{2.3}$$

The values σ_{ij} are uncertainties associated with each original data value x_{ij}, i=1,...,I, j=1,...,J. In the present work, each σ_{ij} was specified by the expression $\sigma_{ij} = 1\mu\mathrm{g/m^3} + 0.08x_{ij}$. The symbol Q_a denotes the auxiliary sum of squares that is created by the auxiliary equations, used for regularization and normalization, to be discussed later. The numerical values obtained for Q_m and Q_a were typically 28000 and 7800, respectively.

Multiple solutions

In contrast to simpler factor analytic models, many multilinear models have many local minima of the sum-of-squares expression Q. When the iteration is started from different random initial values, different solutions will be obtained. It is necessary to perform multiple computations. From the obtained results, one inspects those that have reached the lowest values of Q and chooses such solutions that offer a meaningful interpretation.

The general factors and urban factors

It is of considerable interest to know how the PM2.5 concentrations in urban locations differ from the concentrations in surrounding rural areas. Conceptually we may understand that there is an overall component of PM2.5 that has been distributed throughout the troposphere. This component will be present equally in rural and urban locations. In addition there will be recently released PM2.5 that occurs mostly near its origins in the urban areas.

In the present model, a number of factors were reserved for only representing such components of PM2.5 that occur at urban locations. Technically this was achieved so that for the last five factors (p = 13,..., 17), the spatial factor elements f_{jp} were forced to be zero corresponding to all rural locations j. These factors are called *urban factors*. The other factors (p = 1,..., 12) are called *general factors*.

The term *urban excess* denotes the increase of PM2.5 in urban areas with respect to rural background stations. One should note that there may be more of urban excess than what is contained in the urban factors. The explanation is that the contributions r_{ijp} for the general factors (p = 1,..., 12) may well have larger values at urban locations than in rural ones. The urban factors only explain such part of the urban excess that has different behavior than the general factors with respect to time or with respect to parametric variables. The dilution effect is one example of such a different behavior: the concentration of locally emitted PM2.5 will be lower if the wind is stronger, while the spread-out PM2.5 concentration will not be significantly influenced by the wind speed.

The separation of the urban factors has not been without problems in the present computations. The reason is that in some areas, particularly near some edges of the domain, there are large numbers of urban locations without close-by rural stations. In such areas, the urban-only factors are able to explain all locations. This distorts the separation so that in these areas, the urban-only factors tend to explain a significant fraction of the general PM2.5, too. A paradoxical result emerges: in order to understand what is going on in the urban areas, more rural stations are urgently needed!

Presentation of results

Contributions of individual factors

In 2-way factor analysis, there is no question about how to present the results: the columns of factor matrices **G** and **F** contain the dependence of each factor on time and space. The columns can be plotted as such. The only consideration is about normalization; one can normalize either **G** or **F** columns.

Experience has shown that with the multilinear model (2.1), displaying columns of **G** or **F** can be quite misleading. This is caused by the fact that the parametric variables may be partially collinear with time and also with space coordinates. As an example, the average humidity depends on time of year. Thus the columns of **G** do not correspond well to the behavior of the factor with respect to time of year because indirect or hidden time dependence may be present as the dependence of the factor on humidity. A better picture is obtained by using the contributions r_{ijp}, defined in equation (2.1). Each value $r_{ijp}= m_{ijp}g_{ip}f_{jp}$ indicates how much the p^{th} factor contributes to the data value x_{ij}. The contributions have the dimension of the original data values, i.e. micrograms per cubic meter.

Time averaged contributions of each factor p are obtained so that the values r_{ijp} are averaged over time, i.e. over the first index i. Plotted in a map, the time-averaged contributions represent the spatial patterns of factors. Such maps are shown on the attached figure pages. Time averaged flux density vectors (see below) can be plotted together with the time-averaged contributions. The flux pointers indicate the average direction of movement of the PM2.5 concentration that the factor explains at each site.

Similarly, **location-averaged contributions** represent the average time dependence of each factor p. These time series give a better picture of the time behavior of the factors that the columns of matrix **G**.

The values m_{ijp} in equation (2.1) are dimensionless. Similarly, the five functionals in equation (2.2) are also dimensionless. The values of a functional specify how much the contribution to PM2.5, by the factor in question, increases/decreases with different values of the argument = the parametric variable. Example: the contributions of factor 9 increase with increasing humidity. Thus the values of the functional $\boldsymbol{Q}_9(Q)$ are above 1 for high values of humidity, (Q>10g/Kg, say) and below unity for lower values of humidity.

For each functional, its average value has been normalized to unity for each factor. Typically, the values of functionals are defined for 12 values of the argument, a parametric variable (humidity, temperature, etc.). For in-between argument values, the nearest defined value is used. If there is no dependence, (e.g. factors 3 and 4 do not seem to depend on ozone concentration) then all values of the functional are ≈1 for the factor in question, $\boldsymbol{Z}_3(Z) \approx 1$ and $\boldsymbol{Z}_4(Z) \approx 1$. The functionals are plotted versus their arguments. Thus the argument axis shows the dimension of the parametric variable, g/Kg, ^{o}F, mb, ppb, or m/s. The wind velocity functionals $\boldsymbol{V}_p$ depend on two variables, wind velocity components V_x and V_y. It is tricky to display this dependence. In the accompanying plots, the dependence is shown by variable dot size, so that the area of the black dot indicates the value of $\boldsymbol{V}_p$ for the coordinates V_x and V_y of the dot.

Flux of PM2.5 explained by individual factors

It is useful to know how each factor explains the movement of PM2.5 aerosol. In that way, one may learn about relationships between sources and affected regions. Inspection of the parametric wind velocity factors $\mathbf{V}_1(V_X,V_Y)$ to $\mathbf{V}_p(V_X,V_Y)$ may sometimes be misleading: a high value in the wind factor table may be of little significance if the wind vector corresponding to the large value only occurs infrequently. A

better picture is obtained by considering the **flux density** of PM2.5. The flux density vector φ is defined as the product of concentration and velocity. The flux density contribution vector φ_{ijp} of factor p to data point x_{ij} is $\varphi_{ijp}=(r_{ijp}V_{ijX}, r_{ijp}V_{ijY})$. By averaging the components of φ_{ijp} with respect to time and location one obtains the average (overall) flux density vector of factor p. By only averaging with respect to time, one obtains the **flux density map** of each factor p. The flux density vectors in the p^{th} map show the spatial details of the movement of that fraction of PM2.5 that corresponds to factor p. The dimension of the flux density vector is $\mu g/m^3 \cdot m/s$. The dimension may also be written as $\mu g\ m^{-2}\ s^{-1}$. This dimension suggests that average flux density indicates the average mass that traverses one square meter in one second, when the surface is vertical and perpendicular to the average flux direction. In this definition, averaging is over the period of measurements.

The **flux** of PM2.5 from a specific region is a measure of the net amount of PM2.5 that is transported from the region (net amount is the difference of amounts that are transported out and in). In principle, the flux is obtained by integrating the flux density through a closed surface that encloses the region. The flux density is well known near the surface because both PM2.5 concentration and wind velocity are known. If wind velocity is known throughout the mixing layer, then flux transported by the mixing layer can be calculated because it may be assumed that PM2.5 concentration is constant throughout the layer. A quantitative estimate is thus obtained for the PM2.5 emission from the enclosed region.

Instability of the multilinear model

Experience has shown that various multilinear models exhibit a tendency to instability in the following sense. With error-free data, the model works beautifully. If more and more error is introduced in the data, the factors tend to assume unphysical shapes, such as random-looking oscillations between large and small values. This phenomenon appears to be similar to such instability that is observed when performing regression or least squares fitting when the basis vectors are almost collinear. The well-known solution is called *regularization* or *ridge regression*. Additional terms are introduced in the least squares expression so that these terms tend to damp the meaningless oscillations while introducing as little bias as possible.

Instability was observed with the PM2.5 multilinear model. As an example, the shapes of temperature or humidity factors could contain two strong maxima and two deep minima. Smoothing equations were hence introduced in the model. These equations specify that the first differences and/or the second differences of all parametric factors should be equal to zero. The sigma values for these equations were adjusted so that the smoothing equations for one parametric factor matrix typically contributed 100 units to the value of Q. In this way, almost all parametric factors assumed plausible shapes. At the same time, the main Q value Q_m of the fit increased but this increase was not alarmingly high, on the order of 1000 units. This increase is less than the increase of Q if one factor is omitted from the model. – Introduction of smoothing equations brings a certain aspect of arbitrariness in the model. This same arbitrariness has been the subject of much debate when considering ridge regression. However, it is well known that useful results are obtained if this arbitrariness is accepted.

Analyzing the uncertainty of the computed results

Importance of individual factors

It is natural to ask about the importance of different factors for achieving a good explanation of the measured data. High concentrations alone do not necessarily mean that a factor is important: it could happen that the model contains two almost similar high-concentration “twin” factors. Then neither one is important because the other one may be able to explain the contents of both factors. However, if one of the twin factors is removed, then in the reduced model the remaining one would be important if the other factors would be too different so that they could not approximate the original load of the two twin factors. In general one would expect a factor to be important if it is different than the others, it is not extremely weak, and there are no free rotations that might influence the factor.

The importance of individual factors has been explored by observing how well the other factors are able to compensate when one factor is artificially made weaker. The increase of Q (dQ) was used as the measure of the uniqueness or importance of each factor in the following way. One factor p in turn was decreased by multiplying its factor elements g_{ip} by a chosen factor $\alpha<1$. All factor elements belonging to factor p were "frozen" so that they were regarded as constants, not variables, in the fitting process. The remaining factors were fitted in order to achieve the best possible fit, given the distortion in the chosen factor p. In this fitting, all aspects of all factors (except for factor p) were adjustable. Different values of α seem to give consistent results. The value $\alpha = 0.8$ may be chosen as representative. The two most important factors (with $\alpha = 0.8$) were F9 (dQ = 460) and F10 (dQ = 327). For the least important factors, the value of dQ was dramatically smaller. For all the factors F11, F12, F13, F14, and F15, dQ was between 30 and 48. For the remaining factors, dQ was between 104 and 232.

Confidence intervals

A script has been written for the Multilinear Engine for the purpose of determining confidence intervals of functionals of factor elements. In the present case, the functionals were chosen to be contributions by individual factors to three consecutive observations at one chosen site at a time. The script forces a functional to deviate from its best-fit value and determines how far the functional may go before the Q value of the fit grows past a chosen limit. In simulation studies (Paatero, unpublished), it was found that the increase of Q by 4 units gives a reasonable confidence (95%, say) for the intervals. The increase of 4 was chosen for this work. However, no confidence should be quoted because the statistical properties of errors are not known in this real-world study.

Mathematically, the confidence intervals (l,u) are computed in the following way. Denote all factor elements (elements of **G** and **F** and the parametric factor elements) collectively as the vector **f**. Let C(**f**) be the functional whose lowest and highest possible values l and u are to be determined. (Often, C(**f**) simply consists of a single element of **f**). Denote the best-fit value of Q by Q_{opt}. Then the limits are obtained as

$$\begin{aligned} l &= \min_{Q(\mathbf{f})<Q_{opt}+4} C(\mathbf{f}) \\ u &= \max_{Q(\mathbf{f})<Q_{opt}+4} C(\mathbf{f}) \end{aligned} \qquad (4.1)$$

If a higher confidence is desired, then a larger allowed increment value can be used instead of the value 4 chosen in this work.

The sites for confidence studies were chosen such that one or two factors were strong while other factors only made small contributions. The contributions of the two strongest factors were estimated. From each of the following three areas, three locations were selected: in North Carolina (factors F3, F5), near Philadelphia, PA (F8, F11), and near Chicago (F9, F12). The contributions over one summer period (days 65, 66, and 67) and one winter period (days 116, 117, and 118) were estimated.

When the full data set was used, the following results were obtained: for the stronger factors, the half-width of the confidence interval was typically 20% to 25% of the best-fit value, symmetrically above and below the best-fit value. For weaker factors, the width could be 30% or more. When the sum of the contributions of the two main factors was estimated, the interval was not wider than for one of the individual factors. When the smoothing was decreased for the time series factors, the best-fit results changed, typically within their confidence bands. Simultaneously, the confidence bands became slightly narrower.

When 45% of data points were rejected (see below), the results changed as follows. The best-fit values changed, often going outside their full-data confidence intervals. The confidence intervals increased markedly. The width of the lower half of the interval was typically 60% of the best-fit value. In 25% of all cases, the lower limit extended to zero. The width of the upper half of confidence intervals was typically 70% to 120% of the best-fit value. — Inspection of data values and random omissions reveals that

typically two of the three target days had been omitted for the investigated sites. This offers partial explanation for the significant loss of accuracy. In fact, one may be surprised that even this level of performance was possible.

Modifying the data matrix by omitting selected elements from the matrix

Collecting all the information in the PM2.5 matrix is expensive. It is reasonable to ask if essentially the same results could be obtained by measuring fewer data points (either at fewer locations, or in fewer days, or both). This question was addressed so that various configurations of data values were omitted from the data matrix of year 2000 (122 days by 304 locations).

Eberly and Cox (to be published, 2002) studied the omission problem with the 2-way PMF model. They generated "bootstrapped" copies of the matrix **X** by randomly picking 122 rows from **X** so that each instance of picking a single row is independent from the other choices. Then, on the average, 37% of all rows do not get picked at all, while some rows enter the new matrix twice or three times (Poisson distribution with parameter μ=1). They found out that the bootstrapped copies of **X** could be successfully analyzed into 7 or 10 factors so that the shapes of spatial factors did not change in an essential way.

In the present work, in different cases, typically 40 to 50 % of all data values were omitted, as if these values never existed. The following sequence of less and less severe omissions was tried.

- Omit all data of every second day (50% omission)
- Omit randomly chosen 68% of all urban data points (the number of omitted points is 50% of all data points)
- Omit randomly chosen urban points (number = 45% of all points) with a biased probability: the probability of omission depends on the number of neighbors that there are within a prespecified limit distance of 0.72 degrees. If the closest neighbor is farther away, the location is unconditionally accepted.
- Omit randomly chosen urban points (number = 45% of all points) with a doubly biased probability: the probability of omission increases for crowded points, as above, but decreases for points that represent the median of several sites.

In the first case, the multilinear analysis into 12+5 factors clearly failed: the parametric factors oscillated wildly between large and small values, the spatial factors lost much of their shape. Several of the spatial factors lost their identity entirely. The same amount of smoothing was applied as in the original full-data runs.

In the last case, the analysis can well be called successful. Small changes could be seen in the spatial factors. However, it is not *a priori* clear whether these are changes to the better or to the worse. The reason is that the original data, containing crowded or *clustered* regions, does not represent a uniform or balanced geographical sampling. Omitting crowded points has a useful de-clustering effect that improves on the original distribution of points. -- Analysis of residuals has revealed that locations (grid cells) representing averages or medians of several sites have smaller residuals than locations that only represent a single site. The smaller residuals are (partly) caused by averaging out random error in the data from a multi-site grid cell. Hence less information is lost when omitting a location that corresponds to a single site in comparison to its close neighbor that is the average of several sites.

The second case is a partial failure. The parametric factors are still unrealistic. However, the definition of some of the spatial factors is clearly better than in the first case. The difference between the first two cases demonstrates that optimization of the network is not a trivial task. It is not just the number of observations that counts, it is also how they are placed in the space-time configuration. Schemes that discard all values of certain days seem to be the worst ones. On the other hand, fully random omissions are not necessarily the best ones. It will also be necessary to consider schemes where some locations ("A stations") operate

more often while other locations ("B stations") only perform measurements on every second or every third time the A stations do.

The third case comes close to the last one. It represents a result that could perhaps be used in practice. The significant difference between the second and third cases underlines the fact that concentrations at close-lying locations are strongly correlated. Example: Assume that the measurement c(S,D) at site S on day D has already been included in the data set. If site S' is close to S, then the concentration c(S',D) at S' on day D is strongly correlated with c(S,D). The value c(S',D) does not contain much of useful information because it may be predicted based on c(S,D). Hence more information is gained by including c(S",D), measured at a more remote location S", than by including c(S',D).

Discussion

The uncertainty principle

In quantum mechanics, the Heisenberg uncertainty principle is well known. The principle says that if information about one aspect of the state of a particle is sharpened then less information may be gained about other aspect(s) of the state. The same principle can be discerned in the current statistical results. Certain factors have a sharp spatial definition, e.g. the "New York" factor. In contrast, the other properties of the New York factor are undefined: there is no preference of any wind directions, etc. Some other factors have clear distinction between summer/winter days but unsharp spatial definition, and so on.

In contrast to quantum physics, the amount of statistical information can be increased. If more information is available, then the results can become more sharp for more than one aspect of the PM2.5 distribution. Having more information will allow that more factors are used in the analysis and each factor will describe a more sharply defined subclass of all PM2.5. With any given amount of information, there will be a maximal number of factors that can be meaningfully determined. Attempts to use more factors will produce meaningless unphysical or noisy shapes for the different aspects of the factors.

Omitting data points

The present analysis is based on one year of information, on 122 days of data. (Although there is data from year 1999, that data was not included in the present study because of the large number of missing values.) It was seen that a successful analysis into 12+5 factors is possible if all days are available. However, deleting half of the days did not allow the multilinear analysis into 12+5 factors. A basic 2-way analysis into 7 or 10 factors was still possible if 37% of all days were omitted.

It is concluded that the situation will improve as soon as two years of data are available. From the full two-year data set of size 244x304, more factors can be determined, allowing a more detailed analysis of the PM2.5 distribution. The extent of this improvement cannot be predicted, however. Alternatively, the present performance level can be maintained even with higher omission percentages. Thus it is not meaningful to try to quantify how much information could be gained from the present one-year data set by different omission strategies, because of two reasons: (1) the situation that one year only is available is soon over and then it does not make sense to use one year only, (2) it is not clear how representative the year 2000 happens to be in the long run. Quantitative assessments are not reliable if based on a single year.

Fitting especially high values

Factor analytic techniques are basically geared for analyzing the typical or average behavior. Thus it is necessary to check the fitting of highest concentrations. This is also necessary because the US legislation focuses on the highest concentrations, not on the average ones.

The following table illustrates how PM2.5 values above 50, 60, and 70 $\mu g\ m^{-3}$ are fitted (recall that the allowed limit is 65 $\mu g\ m^{-3}$). The rows marked "Alarm OK" indicate the numbers of successfully identified high-concentration cases where both the data value and the corresponding fitted value exceed the limit L

specified on the top line. The other rows indicate numbers of failed cases. “Alarm failed” means cases where the data value exceeded the limit while the fitted value did not, i.e. no alarm was sounded. “False alarm” indicates cases where the fitted value exceeded the limit without reason, i.e. when the data value did not exceed the limit. In a successful fit, the numbers of failures should be small in comparison to the numbers of successes.

Method of analysis	Outcome	Limit L=50	Limit L=60	Limit L=70
		Number of cases		
Use parametric factors, use all data, 12+5 factors	Alarm OK	23	5	2
	Alarm failed	29	9	6
	False alarm	6	0	0
Use parametric factors, use 55% of data, 12+5 factors	Alarm OK	18	1	1
	Alarm failed	34	13	7
	False alarm	42	17	5
The basic 2-way PMF, use all data, 7 factors	Alarm OK	4	2	1
	Alarm failed	63	15	8
	False alarm	0	0	0

This table shows that the highest values are not especially well fitted. There are 5+9=14 data values above 60. When all data are used, then 5 of the 14 get a fit above 60. With 55% of data fitted, only one of the 14 get a fit >60. What is worse, with 55% of data there are 17 false alarms above 60. Similar results are seen for the other levels 50 and 70.

The third method, the basic 2-way PMF model, was computed with a more complete data set, hence the numbers of cases are slightly larger. It is seen that practically speaking, this model does not predict any of the large values. Of the 67 data values above 50, the model fits 4 so that the fitted value is above 50. This is especially significant as 9 of those 67 data values even exceed 70.

Including additional or supporting information

Limiting the spatial extent of individual factors.

The spatial shape of most factors consists of a clearly limited high-concentration region or “blob” plus low (non-zero) concentration values scattered all around the domain. The wide-spread noise-like stray concentrations have nothing to do with the physical or meteorological reasons that have created the blob. It is believed that random fluctuations in the data are the cause of noise-like spatial coefficients. Thus the result will be more informative if the noise coefficients are eliminated. For this reason, the final results were computed in the following way. The spatial factor maps were inspected and the clear blobs were determined. Latitude-longitude limits in the form of rectangular “boxes” were specified around the blobs, allowing ample margins between the box and the blob. No spatial coefficients were allowed to be non-zero outside the boxes. For each factor, the box was specified individually. There are a few factors that have a wide spatial distribution without any clearly defined blob. For those factors, no spatial constraints or especially mild constraints were specified.

The spatial constraints caused that the spatial stray components also decreased inside the box, so that the distinction of the blob and the surrounding domain became even more clear. This supports the understanding that the stray components were originally caused by a time-space rotation which was

"undone" by specifying the limiting boxes. The increase of Q, caused by the spatial limits, is small (less than 2000 units). Also, no unnatural details are seen. On the other hand, a parsimonious solution (containing as few fitted values as possible) is generally preferable. Thus it is suggested that the latitude-longitude-limited solution is preferable to the original solution that was computed without spatial limits.

The question of spatial constraining is connected with the question of analyzing as large domains as possible. In order to avoid edge effects, it would be useful to analyze large domains within a single model. A large domain means that a large number of factors is needed in order to represent the different conditions in different parts of the domain. If each factor is allowed to carry stray components all around the domain, then the proportion of noise-like terms will increase in comparison to the structured part of the model, eventually weakening the structure of the model and preventing the analysis of large domains. It is believed that with increasing domain size, the spatial constraining will become a must so that no successful detailed analysis is possible without spatial constraining of stray components.

Smoothing the G *(time) factors*

As already discussed, there is collinearity between time and the parametric variables. For this reason, some variation of PM2.5 concentration may be explained either by the parametric factors or by variation of **G** values. It is more useful to have the parametric factors explain the variation of PM2.5, because then it may be possible to understand the underlying chemical or physical processes. For this reason, strong smoothing was applied to the columns of **G**. This smoothing attempts to enforce a day-to-day smooth behavior of the columns of **G**. Although this smoothing appears to have some useful effect, its effect was not as strong as anticipated. The time factors, and also the corresponding location-averaged contributions, still show strong variation whose origins are not clear. The usefulness of the smoothing is questionable. Confidence interval studies indicate that more narrow intervals are obtained if there is less smoothing of columns of **G**.

What should be improved?

When the sum-of-squares expression (2.3) was defined, it was assumed that the individual errors are independent. However, it is well known that this assumption is not true. It would be good to improve the model so that correlation of data errors is taken into account. Unfortunately such a change is in conflict with the present structure of the program ME-2: the program is based on the assumption that Q consists of a sum of squares of quantities, without cross terms. It is not clear if the program can modified so that cross terms, i.e. interactions of individual error values, could be included in the model.

Inspection of flux maps reveals that at certain locations, the flux pointers point in conflicting directions. This may be caused by non-representative wind data that is influenced by local conditions. Conflicting wind vectors mean that the model cannot use wind factors as well as should be possible. Some form of preprocessing of wind data will be needed.

References

Pentti Paatero, Least squares formulation of robust non-negative factor analysis, *Chemometrics and Intelligent Laboratory Systems* **37** (1997) 23-35.

Pentti Paatero, The Multilinear Engine - a Table-driven Least Squares Program for Solving Multilinear Problems, Including the n-way Parallel Factor Analysis Model. *Journal of Computational and Graphical Statistics* (1999), Vol **8**, Number 4, 854-888.

Paatero, P. and Hopke, P.K., Utilizing Wind Direction and Wind Speed as Independent Variables in Multilinear Receptor Modeling Studies. Accepted to *Chemometrics and Intelligent Laboratory systems*, (2001).