Atmospheric Environment 43 (2009) 3989–3997

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Comparison of the results obtained by four receptor modelling methods in aerosol source apportionment studies

R. Tauler [a,*], M. Viana [a], X. Querol [a], A. Alastuey [a], R.M. Flight [b], P.D. Wentzell [b], P.K. Hopke [c]

[a] Institute of Environmental Assessment and Water Studies, IDAEA-CSIC, C/Jordi Girona 18-26, 08034 Barcelona, Spain
[b] Department of Chemistry, Dalhousie University, Halifax, NS B3H 4J3, Canada
[c] Center for Air Resources Engineering and Science, Clarkson University, Box 5708, Potsdam, NY 13699-5708, USA

ARTICLE INFO

Article history:
Received 1 February 2009
Received in revised form
6 May 2009
Accepted 12 May 2009

Keywords:
Atmospheric pollution
Particulate matter
Source apportionment
Receptor modeling
PMF
MCR-ALS
MCR-WALS

ABSTRACT

In this work the performance and theoretical background behind two of the most commonly used receptor modelling methods in aerosol science, principal components analysis (PCA) and positive matrix factorization (PMF), as well as multivariate curve resolution by alternating least squares (MCR-ALS) and weighted alternating least squares (MCR-WALS), are examined. The performance of the four methods was initially evaluated under standard operational conditions, and modifications regarding data pre-treatment were then included. The methods were applied using raw and scaled data, with and without uncertainty estimations. Strong similarities were found among the sources identified by PMF and MCR-WALS (weighted models), whereas discrepancies were obtained with MCR-ALS (unweighted model). Weighting of input data by means of uncertainty estimates was found to be essential to obtain robust and accurate factor identification. The use of scaled (as opposed to raw) data highlighted the contribution of trace elements to the compositional profiles, which was key to the correct interpretation of the nature of the sources. Our results validate the performance of MCR-WALS for aerosol pollution studies.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

In recent years, there has been an increased interest in the application of chemometrics (Massart et al., 1997) to different environmental research fields, ranging from water to air pollution (Einax et al., 1997; Hopke, 1985). One aspect of the application of chemometrics to environmental pollution research is often referred to as source apportionment, receptor modelling and/or mixture analysis discipline. Recent examples of such work can be found in Europe (Jeanneau et al., 2008; Viana et al., 2008), the US (Ke et al., 2007; Shrivastava et al., 2007) and Asia (Bi et al., 2007; Srivastava and Jain, 2007). In the fields of pollution sciences (air or water), source apportionment models aim to re-construct the emissions from different sources of pollutants based on ambient data registered at monitoring sites (Lee et al., 2006).

In atmospheric sciences the most commonly used multivariate data analysis methods are principal component analysis, PCA (Jolliffe, 2002), singular value decomposition, SVD (Golub and Van Loan, 1989), the UNMIX (Henry and Kim, 1990; Henry, 2003) and positive matrix factorization, PMF (Paatero and Tapper, 1994), which have already proven their reliability in a vast number of receptor modelling and source apportionment works (Salvador et al., 2004; Ogulei et al., 2006; Song et al., 2006). Both the UNMIX and PMF methods have been adopted by US EPA as robust methods (i.e. accurate for different operating conditions) for air quality management and can be freely downloaded from its web page (http://www.epa.gov) In the present study we introduce the application of the multivariate curve resolution alternating least squares method, MCR-ALS (Tauler et al., 1995; Tauler, 1995; de Juan and Tauler, 2003) to atmospheric pollution data. This method has been already successfully used to solve different types of problems (Tauler et al., 2004; Jaumot et al., 2006; Felipe-Sotelo et al., 2006; Peré-Trepat et al., 2007). In addition, the recently proposed extension of the basic MCR-ALS algorithm that includes uncertainty information, multivariate curve resolution with weighted alternating least squares, MCR-WALS (Wentzell et al., 2006) is also considered. None of them, MCR-ALS nor MCR-WALS, have been used previously for air pollution source apportionment studies.

Current research directions point towards the widespread use of source apportionment data as input in health effects studies and mitigation plans (Viana et al., 2008), and thus the data from different studies must be comparable. Our study adds to recent intercomparison exercises (Song et al., 2006; Rizzo and Scheff, 2007) as it presents the evaluation of the performances of four receptor modelling methods, two of which (MCR-ALS, MCR-WALS) have been

* Corresponding author.
E-mail address: Roma.Tauler@idaea.csic.es (R. Tauler).

1352-2310/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2009.05.018

applied to aerosol science for the first time. This evaluation allowed us not only to test the sensitivity of the different methods, but also to analyse the theoretical causes for the differences obtained. Our results aim to facilitate the interpretation of receptor modelling results, and to encourage the understanding of the different outputs provided by different receptor modelling methods, in order for researchers to be able to apply the optimal method for each case study.

2. Methods

Four receptor modelling methods (PCA, PMF, MCR-ALS and MCR-WALS) were applied to a single dataset containing PM_{10} (atmospheric particulates with aerodynamic diameter <10 μm) speciation from an urban background site under industrial influence in Northern Spain.

The following different situations were examined: (a) using raw data; (b) using scaled data; (c) including or excluding estimated uncertainties (PCA and MCR-ALS vs PMF and MCR-WALS); and (d) including and excluding an offset in the multilinear regression analysis. Consequently, the objectives of this paper are two-fold: (a) to compare the overall performance of the different receptor modelling methods under their usual conditions using raw data; and (b) to compare the effects of modifying the operational conditions of the methods.

2.1. Study area, analytical determinations and uncertainty calculations

Details on the study area, previous source apportionment analyses and specific details on PM_{10} sampling may be found in Viana et al. (2006a). Analyses of chemical components were performed according to the methodology described by Querol et al. (2001) for a total of 87 valid PM_{10} samples. Concentration values below the detection limit were substituted by half of the detection limit as recommended in the literature (see Farnham et al., 2002).

In this work sample-specific or feature-specific uncertainties were not experimentally available and they were estimated to be proportional to estimated concentrations and also with a constant term related to their limit of detection. When the value of a variable in a sample is above its detection limit, the uncertainty of this value is considered to be 10% of its value plus the detection limit value divided by three (i.e. when $x_{ij} > \mathrm{LOD}$, $s_{ij} = 0.1 \times x_{ij} + \mathrm{LOD}/3$). On the other hand, when the value of a variable is below or equal to its detection limit, its uncertainty is estimated to be 20% its value plus its detection limit divided by three (i.e. when $x_{ij} \leq \mathrm{LOD}$, $s_{ij} = 0.2 \times x_{ij} + \mathrm{LOD}/3$). The uncertainties calculated using these formulas give values, which are mostly determined by a high proportional value and a low constant value. This produces a heterocedastic noise structure, which can be handled properly by weighted schemes of PMF and MCR-WALS procedures (see below). One of the reasons for the choice of this calculation was the fact that the experimental data were obtained using four different analytical techniques (ICP-AES, ICP-MS, FIA and CHNS analysis), which resulted in different analytical uncertainties. By using estimated uncertainties instead of sample-specific uncertainties, this bias was reduced. The use of estimated uncertainties instead of sample-specific uncertainties has been documented by Kim and Hopke (2007).

With this noise structure, the average signal-to-noise ratio was estimated to be approximately 8 (mean $(S/N) = 7.7$). This average is calculated dividing each data value by its associated uncertainty, and then averaging for all the values. Proportional noise is known to produce a heterocedastic structure, with each estimation of s_{ij} for each x_{ij} value being different and not constant, but of independent nature and without correlation. Maximum likelihood estimations of bilinear model parameters under uncorrelated heterocedastic noise (Wentzell et al., 2006, 1997) should consider uncertainties s_{ij}, especially for those cases where uncertainties can be high. The effect of introducing noise uncertainties during the data analysis stage is discussed below.

2.2. Data analysis

The four methods used in this work (PCA, MCR-ALS, MCR-WALS and PMF) are based on the same bilinear model, which is described by:

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

$$\mathbf{X} = \mathbf{G}\mathbf{F}^{\mathbf{T}} + \mathbf{E} \tag{2}$$

In Eq. (1), x_{ij} refers to a particular experimental measurement of concentration for species j (one of the analytes) in one particular sample i (1 sampling day). Individual experimental measurements are decomposed into the sum of N contributions or sources, each one of which is described by the product of two factors, one (f_{nj}) defining the relative amount of the considered variable j in the source composition (loading of this variable on the source) and another (g_{in}) defining the relative contribution of this source in this sample, i (score of the source on this sample). The sum is extended to $n = 1, \ldots, N$ sources, leaving the measurement unexplained residual stored in e_{ij}. Ideally e_{ij} should contain only experimental error, but in practice it may also contain small unknown and unmodelled contributions. Eq. (2) describes the same model in a more compact way using matrix algebra notation. The data matrix of measurements $\mathbf{X}$ is decomposed into the product of two factor matrices, the loadings matrix ($\mathbf{F}^{\mathbf{T}}$) defining the chemical composition of the sources, and the scores matrix ($\mathbf{G}$) related to the contributions or distribution of these sources into the different samples. The noise matrix $\mathbf{E}$ contains the experimental error as well as unmodelled variance sources not included in the N considered components. These variance sources can have large contributions in real data sets and they are not expected to follow a normal distribution.

The first decision to consider in this type of analysis is the number of components to include in the model, i.e. the number of sources to consider in $\mathbf{G}$ and $\mathbf{F}^{\mathbf{T}}$ factor matrices. Our approach is more practical than theoretical and it is based on two goals. First, the selected number of sources should explain the major systematic or deterministic variance (not the random or stochastic one) and, second, that these sources should have interpretable composition (loadings) and contribution (scores) profiles. Keeping this in mind, tools like principal component analysis or singular value decomposition already provide an estimation of the amount of variance explained by a particular number of components and also what amount of new variance is added when a new component is considered. On the other hand, the detailed observation of the loading and scores profiles, together with the knowledge of the problem, allow the proposal of a particular model.

PCA performs the decomposition described by Eqs. (1) and (2) under the very specific constraints of factor orthogonality ($\mathbf{G}$ and $\mathbf{F}^{\mathbf{T}}$), normalization (only $\mathbf{F}^{\mathbf{T}}$ in PCA) and maximum explained variance for successive extracted components. Under these constraints, PCA has the favorable mathematical property of unique solutions, i.e. there is only one possible solution. PCA is especially useful to investigate the variance structure of a particular data set, i.e. how much data variance can be explained by a particular number of components. However there are two problems with the application of PCA in environmental studies: (a) the interpretability of the sources; and (b) noise propagation effects on PCA results. PCA solutions are in fact an abstract mathematical linear combination of the true variance

sources and may bear little resemblance to the true variance sources. For instance, whereas PCA solutions have positive and negative score and loading elements to achieve the orthogonality properties, composition and distribution profiles of the true pollution sources must be non-negative and are generally not orthogonal. Factor rotations like varimax orthogonal rotation (Jolliffe, 2002) has been proposed to improve factor interpretability, but it is in general difficult to achieve this interpretability completely from only abstract mathematical factor orthogonal rotations. It is therefore more reasonable to look for solutions in agreement with more natural constraints, such as non-negativity instead of orthogonality, but the mathematical property of uniqueness is then lost. MCR and PMF methods described below are based on the application of more natural constraints like non-negativity.

Another problem, mentioned previously, is that the solutions obtained by ordinary PCA or MCR-ALS are only optimal in case of independent and identically distributed (iid) errors (Wentzell et al., 1997). This assumption cannot be made in general and is not satisfied by the example in this work, where relatively large uncertainties in the measurements are assumed to be proportional to concentrations. In these cases, alternative algorithmic least squares approaches to calculate optimal solutions fulfilling maximum likelihood (ML) assumptions (Wentzell et al., 1997, 2006) have been proposed. In this work, this type of approach has been considered for the MCR-WALS and PMF methods.

An additional aspect to consider in this discussion is the selection of a more convenient data pre-treatment strategy. Results of the application of PCA will be presented for raw data, for scaled variables (each variable value is divided by its standard deviation over all samples or matrix rows) and for autoscaled variables (variables are mean centered and scaled). In the PCA literature (Jolliffe, 2002), the more common data pre-treatment is autoscaling, which allows the assignment of equal weight to all variables and removes their constant offsets. However, due to non-negativity constraints and also to PM_{10} source apportionment (see below), autoscaling is not usually applied in MCR and PMF methods. For MCR and PMF, results will be shown for raw and scaled data only. As it was already stated in Section 2.1, data scaling will allow focusing on the composition of the sources in lower concentration elements.

MCR-ALS and PMF both decompose the experimental data (Eqs. (1) and (2)) using non-negativity constraints instead of orthogonality constraints. However, they differ in the algorithms used to decompose the experimental data matrix and also (of minor importance), in the normalization of loading and scores profiles. It is not the purpose of this paper to make a detailed comparison of these two algorithms, but to compare their results for the same experimental data set. This comparison has not been performed previously and we consider it of interest.

PMF was developed to cope with uncertainties and error propagation problems and to achieve more statistically sound maximum likelihood solutions (Paatero and Tapper, 1994) in the analysis of noisy environmental (air pollution) data. From its initial development, error estimates of the experimental measurements were necessarily included in PMF. The residual error square sum to be minimized is weighted accordingly. In contrast, MCR-ALS was initially developed for spectroscopic data analysis of evolving mixtures of chemical components (de Juan and Tauler, 2003). In most of the cases, spectroscopic measurements provide relatively precise measurements, with low (typically 1–3% of the measured signal or lower) and fairly uniform uncertainties. In these cases applications of MCR-ALS without consideration experimental uncertainties have produced good results (de Juan and Tauler, 2003). However, in the case of noisier experimental data, as for the environmental data set analyzed in the present work, this situation is more complex and should be reconsidered. This work compares the results obtained by the conventional MCR-ALS method with those obtained by its counterpart, a new MCR-WALS method, where uncertainty estimations are incorporated in a weighted alternating least squares algorithm. In this new approach a maximum likelihood total least squares solution, TLS (Van Huffel and Vandewalle, 1991) is incorporated in place of the standard least squares solution (see for more details Schuemans et al., 2005; Wentzell et al., 2006). In both MCR-WALS and PMF, the noise structure of the experimental data was considered to be heterocedastic, where the different measured variables (concentrations of the different chemical elements and compounds in the samples) were considered to have different uncertainties (see uncertainties calculation in previous Section 2.1).

Another aspect to be considered in this work is the principle of mass conservation, which can be assumed to be fulfilled for the investigated data set. To achieve the source apportionment of the total mass, the independent gravimetric measurement of total sample masses in the form of PM_{10} (see experimental section) was used. In PCA of autoscaled (mean centered) data, source apportionment requires first eliminating the effect of mean-centering in the scores. As a result of this method, the so-called absolute principal component scores, APCS (Thurston and Spengler, 1985; Viana et al., 2006b), are obtained, and a multilinear regression model between them and the total daily PM_{10} mass is performed. For MCR-ALS, MCR-WALS and PMF, such a source apportionment can be performed directly using the resolved scores (either from scaled or unscaled data, since this theoretically should not modify them significantly). The mass balance equation to perform this multilinear regression can be written as in Eq. (3)

$$\mathbf{m} = \mathbf{G}\,\mathbf{b} \tag{3}$$

where $\mathbf{m}$ $(I, 1)$ is the column vector of PM_{10} mass measurements for each sample (I samples in total), $\mathbf{G}$ (I, N) is the scores matrix obtained in the resolution of the model (Eq. (2)) considering N sources and using the different approaches (in the case of PCA, instead of $\mathbf{G}$, APCS are considered here) and $\mathbf{b}$ $(N, 1)$ is the regression vector giving optimal fit of PM_{10} for a given $\mathbf{G}$ considering the N variance sources. Eq. (3) may include an offset term (as an additional element in regression vector $\mathbf{b}$ and a column of ones in the scores matrix $\mathbf{G}$) to account for constant apportions not provided by the resolved components in Eqs. (1) and (2).

The average mass contribution percentage of different sources to the particulate matter can be estimated by the equation:

$$\%\mathbf{aver} = \left(\mathbf{1}./\mathbf{m}^{\mathbf{T}}\right)\mathbf{G}\,diag(\mathbf{b})(100/\mathrm{I}) \tag{4}$$

In Eq. (4), $\%\mathbf{aver}$ $(1, N)$ is a vector that gives the average percentage contribution of each source to the particulate matter, $\mathbf{1}$ $(1, I)$ is a row vector of ones, $\mathbf{m}$ $(I, 1)$ is a column vector of particulate mass concentrations, $\mathbf{G}$ (I, N) is the matrix of source contributions, $\mathbf{b}$ $(N, 1)$ is the regression vector from Eq. (3), and I is the number of samples. The Hadamard (element-wise) quotient is indicated by "./" and "diag" is used to convert $\mathbf{b}$ into a diagonal matrix (N, N).

To evaluate the performance of the different methods and to allow their comparison, the following magnitudes were defined:

1) Explained variances, either for individual components or for the sum of all the considered components in a particular model:

$$R^2 = 100\left(1 - \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} e_{ij}^2}{\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2}\right), e_{i,j} = x_{i,j} - \widehat{x}_{i,j} \tag{5}$$

where x_{ij} is the experimental value in the data matrix and $\widehat{x}_{ij}$ is the corresponding calculated value using a bilinear model based method, like PCA, PMF or MCR-ALS.

Table 1
Percentage of explained variances R^2 by individual components (1–6), their sum (SUM) and when they are considered together in the model (ALL) (see Eq. (5)). In PCA and VARIMAX (V_{max}) methods, loadings 1–6 are ordered according to the amount of explained variances. In MCR-ALS, MCR-WALS and PMF, loadings 1–6 are ordered according to pairwise correlation coefficients found when score profiles are compared for unscaled and scaled data.

Method	1	2	3	4	5	6	ALL	SUM
PCA raw	84.7	11.0	1.3	1.0	0.8	0.5	99.4	99.4
V_{max} raw	46.3	43.8	2.3	3.7	2.5	0.7	99.4	99.4
PCA scaled	72.8	6.2	3.7	2.6	2.2	1.8	89.2	89.2
V_{max} scaled	23.4	21.0	19.2	11.9	8.5	4.7	89.2	89.2
PCA auto	29.7	16.5	9.9	6.3	5.9	4.8	73.2	73.2
V_{max} auto	19.5	15.6	11.2	11.2	8.7	7.0	73.2	73.2
Method	1 (steel)	2 (crustal)	3 (marine)	4 (valley)	5 (traffic)	6 (pigment)	ALL	SUM
ALS	12.6	6.7	13.4	45.2	28.3	35.2	99.3	141.4
WALS	14.7	25.4	10.9	50.0	13.4	8.7	87.6	123.2
PMF	14.2	28.2	12.4	43.9	13.8	10.1	85.6	122.6
ALSS	20.2	26.3	19.7	29.0	7.1	23.9	88.9	126.3
WALSS	21.7	28.5	14.0	20.6	12.5	15.3	82.1	112.7
PMFS	20.2	30.0	13.7	19.8	14.7	16.4	82.8	114.9

2) Pairwise correlation coefficients, r^2, for the comparison of loading and score profiles obtained by the different receptor modelling methods:

$$r^2 = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{6}$$

where in this case **x** and **y** are simply two profiles (loadings or scores) from different methods to be compared. The correlation coefficients serve two purposes. First, for MCR-ALS, MCR-WALS and PMF, the components are not necessarily extracted in the same order for each method, even when the same constraints and number of components are used. Therefore, it is necessary to match the components by examining the correlations of the loadings or scores between the two methods. When comparing scaled vs unscaled results, this comparison is generally done with scores, which should be invariant with respect to column scaling. The second purpose of the correlation coefficients is to evaluate the similarities of the profiles extracted by two methods in the validation and interpretation of results.

Applications of the different methods compared here have been described in previous publications, for PCA (Jolliffe, 2002; Thurston and Spengler, 1985); for PMF (Paatero and Tapper, 1994; Lee et al., 2006); for MCR-ALS (Tauler et al., 2004; Jaumot et al., 2006; Felipe-Sotelo et al., 2006; Peré-Trepat et al., 2007), and for MCR-WALS (Wentzell et al., 2006).

3. Results and discussion

3.1. Data description

In the original experiments, 63 variables (corresponding to different species concentrations in μg m^{-3} and ng m^{-3}) were measured for each of the samples, out of which, 34 were used in the source apportionment analyses. The remaining variables were excluded either because of their low signal-to-noise ratios ($S/N < 2$) or because of the small number of measurements available (% >DL = 65% of the samples or lower). A total number of 87 samples were finally analyzed. Details about experimental data are given separately in the Supporting Information section (see Supporting Information, Table S1 and Figs. S1 and S2). Some variables have more influence than the rest: C_{total} (total carbon), Al_2O_3, Ca, K, Na, Mg, Fe, SO_4^{2-}, NO_3^-, NH_4^+. In order to give equal weight to all the investigated variables and to allow for a better source identification, raw data were scaled to equal variance. In this way, all variables showed a comparable range of variation and contribute with the same weight to data analysis. Uncertainties for scaled data were taken from previously evaluated uncertainties for unscaled data and divided by the same standard deviations previously obtained to scale the raw experimental data.

3.2. Model analysis

3.2.1. Model selection and performance

The selection of the optimal number of components (see Supporting Information and Table S2) was carried out by singular value decomposition, SVD, analysis (Golub and Van Loan, 1989). In Table 1, the explained variances (Eq. (5)) for each component (1, 2, 3, 4, 5 and 6) together with the explained variances for the full model including all the components (ALL) and the sum of the explained variances for the individual components (SUM) are given. Since PCA and varimax components are orthogonal, the two last columns, ALL and SUM, for these two methods are equal. For PCA of data without any pre-treatment, most of the variance (>95%) is concentrated in the first two components. Adding more components up to six, the explained variance reaches 99.4%. Varimax orthogonal rotation produces a more even distribution of the variance in the two first components, with values of 46.3 and 43.8%. For scaled data also, a large part of the variance is explained by the first two components (79%), and increasing up to 89.2% for six components. Varimax rotation changes considerably the amounts of variance distributed among the different components, as expected. Finally for autoscaled data, the amount of variance explained by six components is reduced to 73.2, and the amount of explained variance among components changed in a less pronounced way. Note that the first principal component of autoscaled data explains only 29.7% of variance compared to 84.7 and 72.8% for raw and scaled data. These large differences should be related to the effect of mean-centering, which eliminates offset contributions for all the variables. In contrast, these offset

Table 2
Comparison of loading profiles obtained by different methods using pairwise correlation coefficients, r^2 (see Eq. (6)). For the loadings comparisons using raw data, WALS loadings are taken as a reference. For the loadings comparisons using scaled data, WALSS loadings are taken as a reference. Loadings 1–6 are ordered according to pairwise correlation coefficients found when scores profiles are compared for unscaled and scaled data.

Method	1 (steel)	2 (crustal)	3 (marine)	4 (valley)	5 (traffic)	6 (pigment)
ALS	0.6019	0.6595	0.9663	0.9905	0.8862	0.9667
WALS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
PMF	0.9998	0.9994	0.9669	0.9976	0.9622	0.9966
ALSS	0.9394	0.9190	0.9515	0.9070	0.1949	0.9259
WALSS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
PMFS	0.9986	0.9991	0.9956	0.9406	0.9486	0.9976

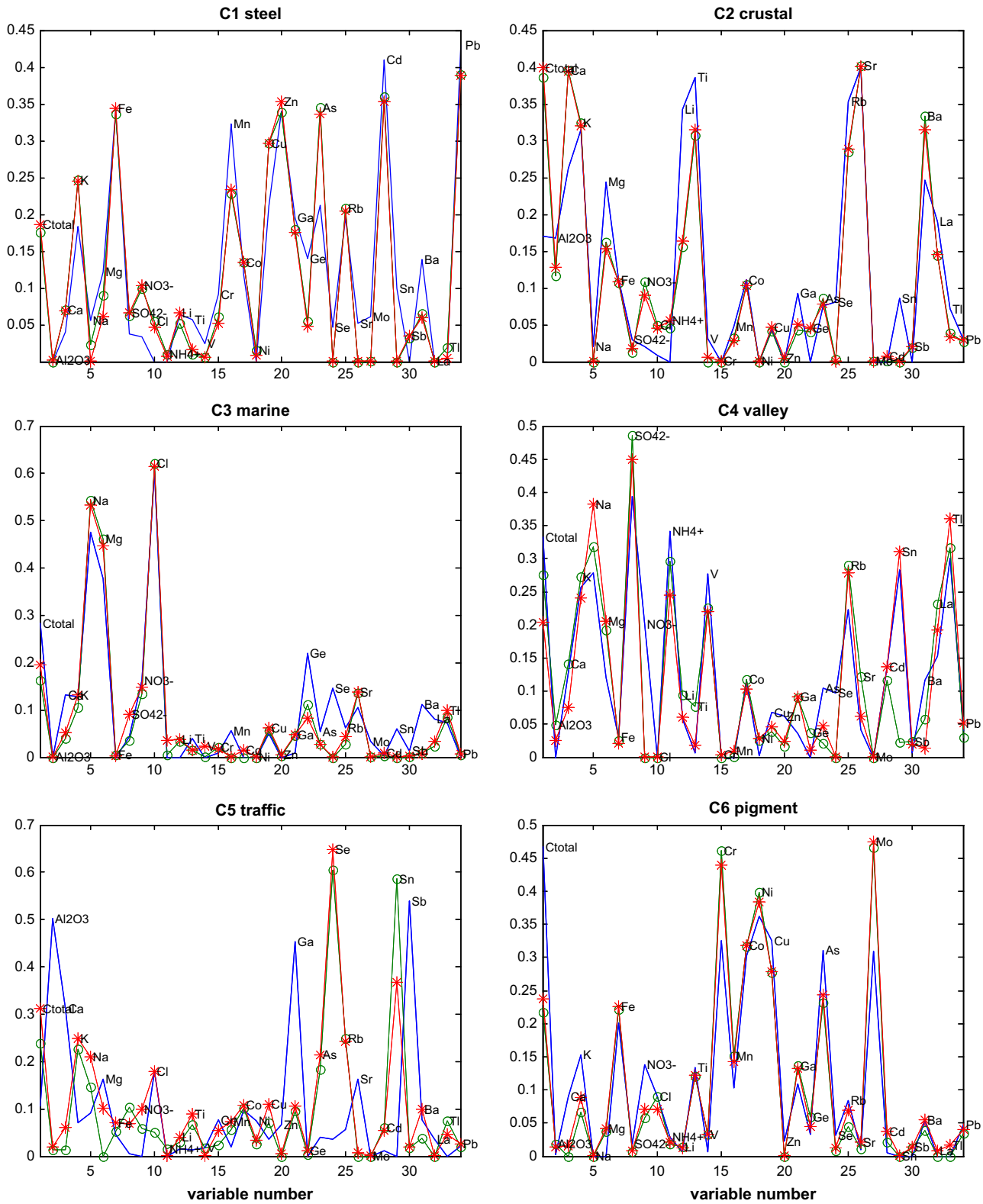

Fig. 1. Comparison of loading profiles in the analysis of scaled data obtained by MCR-ALS (blue line, ALS), PMF (green line with 'o' symbol) and MCR-WALS (red line with '*'). Identification of components C1–C6 is given in the title of each subplot. *X* axes indicates variable number (each species is also identified in the plot by the corresponding chemical symbol). *Y* axes are normalized to one (loadings were normalized to unit length). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

contributions are mostly present in the first principal component of (non-mean centered) raw and scaled data. Although in most PCA studies it is recommend that experimental data are mean centered to eliminate data offset constant contributions, for the purposes of source profile resolution, identification (tracers) and total mass (PM_{10}) source apportionment, it is easier to perform the study directly with raw or scaled data (avoiding negative values produced by mean-centering), and extend the analysis to minor components carrying little variance. In terms of source interpretation of major components, similar results can be achieved with both approaches (either mean-centering or not), but in terms of source apportionment and resolution of the 'true' source profiles for minor components, it is generally better to perform the analysis without mean-centering. For PCA and varimax methods, the components were ordered according the amount of explained variances (orthogonal components).

Results for MCR-ALS, MCR-WALS and PMF for raw and scaled data are also shown in Table 1 (ALSS, WALSS and PMFS for scaled data). To maintain consistency with the rest of the tables and figures of this work, the ordering, correspondence and labelling of the different components was performed according to the identification of their loading and scores profiles (see below). Individual explained variances (Eq. (5)) between MCR-WALS and PMF for the different components shows a very good agreement. In contrast, for MCR-ALS and weighted MCR-WALS, this agreement is only found for some of the resolved components (1st, 3th and 4th), while for others (2nd, 5rd and 6th) the explained variances were significantly different. Another relevant aspect to comment on here is the difference in the explained variances finally achieved for these two approaches. Whereas MCR-ALS with six components explains 99.3% of the data variance (similar to PCA for raw data), MCR-WALS explains only 87.6% of the data variance. It is important to recognize that this difference of approximately 12% between the two approaches is not indicative of an inferior model for MCR-WALS, which would be the conventional interpretation in the presence of iid normal errors. The objective of MCR-WALS is not to explain the largest amount of variance, but rather the largest amount of meaningful variance. The difference shows clearly the tendency to over-fit experimental data in unweighted ALS and PCA methods. This tendency was also confirmed in simulation studies performed at the same time as this work, but these results will not be discussed in detail here. Depending on the noise structure, but especially in the proportional error case, noise enters into the model parameters easily if it is not taken into account, as in ordinary PCA and MCR-ALS. In these methods, eigenvectors are preferentially used to account for random (high noise) variations in large signals rather than important systematic (low noise) variations in small signals, so important information can be lost. Another interesting aspect is the comparison between the last two columns of Table 1, where explained variances for the full model of six components (ALL) for the different applied methods are compared with the sum of the individual explained variances (SUM). The difference between these two quantities gives a measure of the amount of overlap (i.e. the extent of orthogonality) among the components for a particular model. Whereas for PCA and varimax models there is no variance overlap because the profiles of the different components are orthogonal (ALL and SUM give the same number for PCA and varimax results), in the case of ALS, the same difference grows up to approximately 40%. The same happens for PMF and WALS results which, for scaled data, have a variance overlap of approximately 30%. Resolved component profiles (see below) by MCR-ALS, MCR-WALS and PMF are expected to overlap strongly. Orthogonality (zero overlap) among source profiles is not normally encountered in real world applications, which simply means that different sources typically have some common patterns.

Table 3
Comparison of scores profiles obtained by different methods using pairwise correlation coefficients, r^2 (see Eq. (6)). WALSS scores have been taken as a reference for comparison in all cases (scaled and unscaled data).

Method	1 (steel)	2 (crustal)	3 (marine)	4 (valley)	5 (traffic)	6 (pigment)
ALS	0.4546	0.6369	0.9679	0.9259	0.6165	0.6078
WALS	1.0000	0.9999	0.9998	0.9998	0.9999	1.0000
PMF	0.9985	0.9878	0.9817	0.9779	0.9851	0.9989
ALSS	0.9548	0.9145	0.9581	0.9520	0.1851	0.9216
WALSS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
PMFS	0.9986	0.9875	0.9819	0.9794	0.9860	0.9990

3.2.2. Including uncertainties in the analysis of raw and scaled data. MCR-ALS vs MCR-WALS

This section examines the differences in the results obtained by MCR-ALS when uncertainties are included (WALS) or not (ALS) in the analysis of raw and scaled data. This analysis was carried out for MCR, but the results obtained may be extrapolated to the other methods (PMF and PCA).

In Table 2, a quantitative evaluation of the similarity among loading profiles is performed from their pairwise correlation coefficients (Eq. (6)). In this evaluation, due to scaling, this comparison is performed separately for unscaled data (ALS vs WALS) and scaled data (ALSS vs WALSS). It is clear that for non-scaled data, the similarities between loading profiles of the last four components obtained either by ALS or WALS are high ($r^2 > 0.89$), whereas larger differences are obtained for the first ($r^2 = 0.60$) and second ($r^2 = 0.66$) components.

In Table 2, loadings 1–6 were ordered and identified according to pairwise correlation coefficients found when score profiles were compared for unscaled and scaled data (see below). In Fig. 1 and Table 2, these profiles were identified as produced by different possible sources based on their elemental composition. The shape of the loading profiles for scaled data (Fig. 1) changes substantially compared to the same profiles for raw data (Fig. S3 in Supplementary Information section) because of the larger contribution of the chemical elements at lower concentrations (see Supplementary Information section, elements 12–34 in Table S1 and Fig. S3).

For scaled data, all the components (with the exception of number 5, traffic, with a very low correlation $r^2 = 0.18$), the agreement in the tracers between the two approaches is rather good (correlation coefficients r^2 in Table 2 are always larger than 0.90). By using scaled data, the presence of tracers with lower mass contribution is much more evident and thus more informative, resulting in the more precise identification of the emission sources. Component 1 is identified as an industrial source (steel metallurgy), with Fe, Mn, Zn and Cd as the main tracers. Component 2 is a mineral source of crustal origin with relatively high contributions of Ca, Al_2O_3, Mg, K, Ti, Rb, Sr and Ba. Component 3 is a marine source with high contributions of Cl^- and Na, but also of NO_3^- and Mg. Component 4 is a regional source originating from a nearby valley (Viana et al., 2006a), with ammonium sulphate ($SO_4^=$ and NH_4^+) and nitrate aerosols (NO_3^-) as well as marine tracers (Na), metals (V, Co) and crustal elements (Rb, Sr). This regional source refers to the transport of pollutants on the regional-scale from the coast towards the study area, by means of breeze circulations and channelled through the Nervión river valley (see description of this regional source in Viana et al., 2006a). Component 5 is somewhat ambiguous and there is disagreement for some of the tracers. Its composition changes depending on the selected method and also when data were scaled or not. It has been finally assigned to a traffic source due to the high loadings of C_{total}, Sn and Sb (well-known traffic tracers, see Viana et al., 2008) confirmed by PMF and WALS, see below. Component 6 was assigned to a pigment manufacture industry since it gave high contributions for metals typically used in

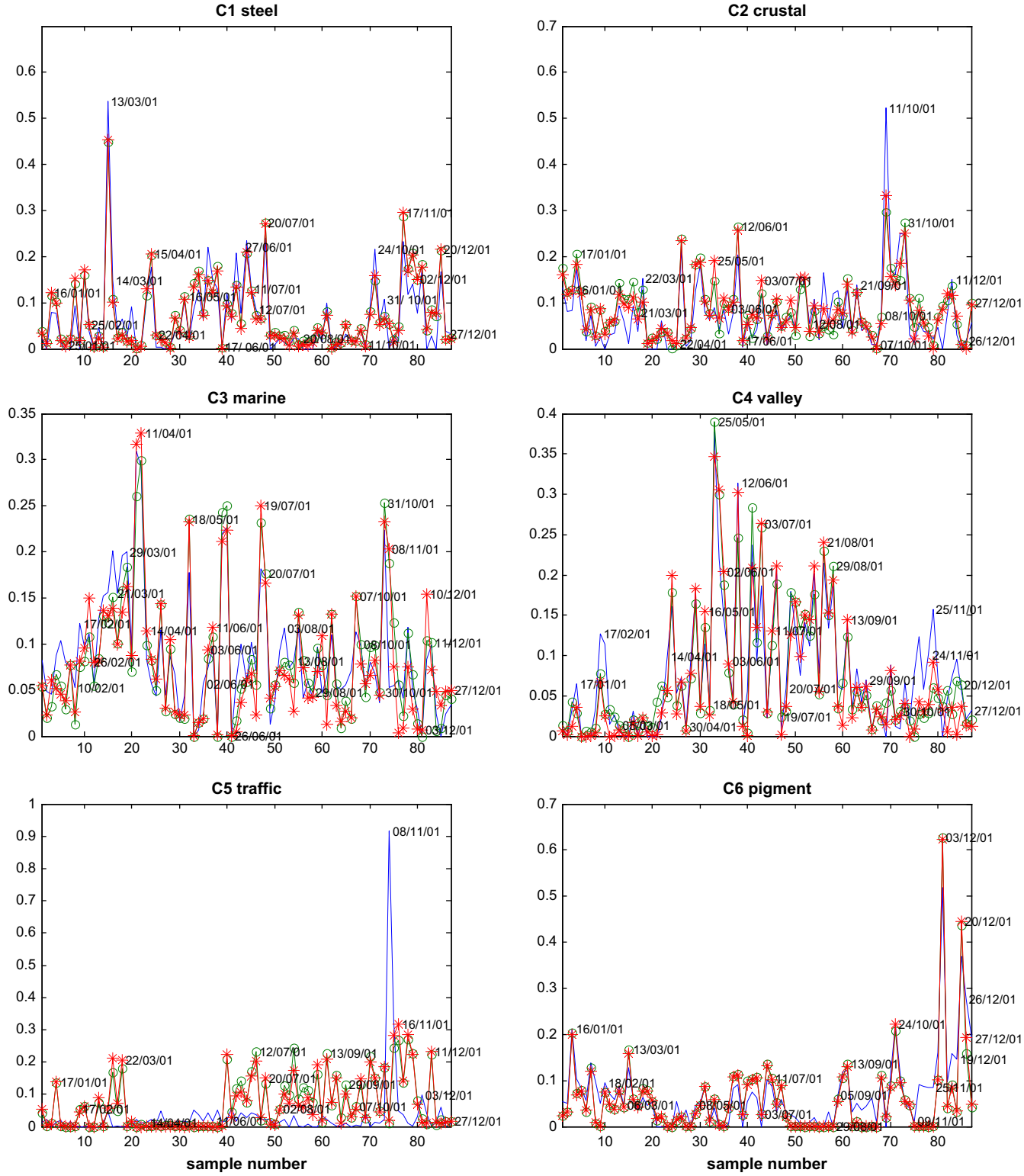

Fig. 2. Comparison of score profiles in the analysis of scaled data obtained by MCR-ALS (blue line, ALS), PMF (green line with 'o' symbol) and MCR-WALS (red line with '*'). Identification of components C1–C6 is given in the title of each subplot. The *X* axes indicates sample number. Some of the dates when these samples were obtained are given in the plots. *Y* axes are normalized to one (MCR-ALS, MCR-WALS and PMF scores were normalized to unit length for their comparison). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a pigment plant located in the area under study (Cr, Co, Mo, Cu). Total carbon is always mixed in the different component loading profiles (especially for unscaled data), which is a consequence of its large presence in all samples (on average it accounted for around a 23% of the total particulate matter mass). For non-scaled data, only the first 11 major elements have large contributions to the loading profiles (see Fig. S3, in the Supplementary Information section).

There is another important distinction that should be made between MCR-ALS and MCR-WALS in this comparison. For MCR-WALS, the results can be considered to be scale invariant. In other words, the results of MCR-WALS for the unscaled and scaled data are essentially identical except for the scaling that was applied. This is not generally true for the ALS algorithm. This invariance for MCR-WALS is expected, since the uncertainties scale with the data and the objective function remains unchanged. This is an important advantage for the WALS algorithm, since it means that the scaling does not fundamentally affect the results, only their presentation. A similar advantage is expected for PMF.

3.2.3. Comparison of MCR-WALS with PMF results for raw and scaled data

In Fig. 1 loading profiles obtained by WALS and PMF are also compared for scaled data. In Fig. S3 of the Supplementary Information section, the same comparison is also given for unscaled data. In both cases, the agreement between the loading profiles obtained by the two methods was very good (for all components $r^2>0.94$). The main conclusion is that WALS and PMF produced essentially the same results, with very little differences among them. The only noticeable discrepancy is in the significant contribution of Sn in the fourth component (valley) resolved by PMF, whereas this contribution was more important in the fifth component (traffic) resolved by WALS (Fig. 1). From these results, it is also clear that scaling the data allowed for a better identification of the tracers for each possible source, and therefore allowed a better source identification. The fact that the two methods, MCR-WALS and PMF, identified and resolved the same sources corroborated the previous source identification and also validated the performance of these two chemometric methods based on completely different algorithm approaches.

In Table 3 the comparison among the scores obtained by the different methods is shown. Correlation coefficients for PCA and varimax score profiles are not given in this table since they were obtained under completely different constraints and their shapes are obviously totally different from those obtained by the other three methods. For their visual comparison in Fig. S3, all score profiles were renormalized. Similar to what was observed for the comparison of loading profiles obtained by the different methods, for both raw and scaled data, a better agreement was obtained between score profiles obtained by WALS and PMF, and a worse agreement was observed between score profiles obtained by (unweighted) ALS compared to those obtained by WALS or PMF (especially for component 5 "traffic" for scaled data). In general, the agreement was better for scaled data than for unscaled data. Score profiles obtained by ALS, WALS and PMF for scaled data are plotted in Fig. 2. The agreement between WALS and PMF scores is very good. Our main conclusion therefore is that the best approach was working with scaled data and residual weighting using either WALS or PMF. As mentioned previously, the correspondence among score profiles measured using pairwise correlation coefficients was used to order and identify correctly the components obtained using the different methods, and especially to deduce what is the correspondence among the components resolved using scaled and non-scaled data. This correspondence cannot be performed directly using the loading profiles since scaling changes their shapes significantly (compare loadings of Fig. 1 for scaled data with loadings of Fig. S3 in the Supplementary section for unscaled data). Another way to check for the correct correspondence among components for scaled and non-scaled data is to perform the inverse scaling operation for the scaled loadings and compare them with the loadings obtained for raw unscaled data. When this is done (results not shown), the component correspondence given in tables and figures of this work is corroborated. It was also confirmed that uncertainty weighting in WALS or PMF significantly improved the recoveries of the unscaled profiles for reasons noted earlier.

In the Supporting information section, comparison of results of source apportionment for the total measured PM_{10} mass obtained by linear regression is given. From this comparison it is concluded that PMF and MCR-WALS give results very similar to each other also for source apportionment. We concluded therefore that the most reliable results were obtained when uncertainties weighting (MCR-WALS and PMF) was considered. The good agreement between the results obtained by these two methods confirmed and validated the identification of the components and the assignation of the corresponding contamination sources.

Acknowledgements

This work was partially funded by the Spanish Ministry of Education and Science (Secretaría de Estado de Universidades e Investigación), and by project CTQ2006-15052-C02-01. Funding was also provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.atmosenv.2009.05.018.

References

Bi, X., Feng, Y., Wu, J., Wang, Y., Zhu, T., 2007. Source apportionment of PM_{10} in six cities of northern China. Atmospheric Environment 41, 903–912.
de Juan, A., Tauler, R., 2003. Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution. Analytica Chimica Acta 500, 195–210.
Einax, J.W., Zwanziger, H.W., Greiss, S., 1997. Chemometrics in Environmental Analysis. VCH, Wiley, New York.
Farnham, I.M., Singh, A.K., Stetzenbach, K.J., Johannesson, K.H., 2002. Treatment of non-detects in multivariate analysis of groundwater geochemistry data. Chemometrics and Intelligent Laboratory Systems 60, 265–281.
Felipe-Sotelo, M., Gustems, Ll., Hernàndez, I., Terrado, M., Tauler, R., 2006. Geographical and temporal distribution of tropospheric ozone in Catalonia (North-East Spain) during the period 2000–2004. Atmospheric Environment 40, 7421–7436.
Golub, G.H., Van Loan, Ch.F., 1989. Matrix Computations, second ed. John Hopkins Univ. Press, London.
Henry, R.C., 2003. Multivariate receptor modeling by N-dimensional edge detection. Chemometrics and Intelligent Laboratory Systems 65, 179–189.
Henry, R.C., Kim, B.M., 1990. Extension of self-modelling curve resolution to mixtures of more than three components. Part 1, finding the basic feasible region. Chemometrics and Intelligent Laboratory Systems 8, 205–216.
Hopke, P.K., 1985. Receptor Modelling in Environmental Chemistry. Wiley, New York.
Jaumot, J., Eritja, R., Tauler, R., Gargallo, R., 2006. Resolution of a structural competition involving dimeric G-quadruplex and its C-rich complementary strand. Nucleic Acids Research 34, 206–216.
Jeanneau, L., Faure, P., Montarges-Pelletier, E., 2008. Evolution of the source apportionment of the lipidic fraction from sediments along the Fensch River, France: a multimolecular approach. Science of the Total Environment 398, 96–106.
Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Springer, New York.
Ke, L., Ding, X., Tanner, R.L., Schauer, J.J., Zheng, M., 2007. Source contributions to carbonaceous aerosols in the Tennessee Valley Region. Atmospheric Environment 41 (39), 8898–8923.
Kim, E., Hopke, P.K., 2007. Comparison between sample-species specific uncertainties and estimated uncertainties for the source apportionment of the speciation trends network data. Atmospheric Environment 41, 567–575.
Lee, J.H., Hopke, P.K., Turner, J.R., 2006. Source identification of airborne $PM_{2.5}$ at the St. Louis-Midwest Supersite. Journal of Geophysical Research 111, D10S10.
Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., de Jong, S., Lewi, P.J., Smeyers-Verbeke, J., 1997. Data Handling in Science and Technology, Handbook of Chemometrics and Qualimetrics, vols. 20A and 20B. Elsevier, Amsterdam.
Ogulei, D., Hopke, P.K., Zhou, L., Pancras, J.P., Nair, N., Ondov, J.M., 2006. Source apportionment of Baltimore aerosol from combined size distribution and chemical composition data. Atmospheric Environment 40, S396–S410.
Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126.
Peré-Trepat, E., Lacorte, S., Tauler, R., 2007. Alternative calibration approaches for LC–MS quantitative determination of coeluted compounds in complex

environmental mixtures using multivariate curve resolution. Analytica Chimica Acta 595, 228–237.

Querol, X., Alastuey, A., Rodríguez, S., Plana, F., Ruiz, C.R., Cots, N., Massagué, G., Puig, O., 2001. PM_{10} and $PM_{2.5}$ source apportionment in the Barcelona Metropolitan area, Catalonia, Spain. Atmospheric Environment 35, 6407–6419.

Rizzo, M.J., Scheff, P.A., 2007. Fine particulate source apportionment using data from the USEPA Speciation Trends Network in Chicago, Illinois: comparison of two source apportionment models. Atmospheric Environment 41, 6276–6288.

Salvador, P., Artíñano, B., Alonso, D.G., Querol, X., Alastuey, A., 2004. Identification and characterisation of sources of PM_{10} in Madrid (Spain) by statistical methods. Atmospheric Environment 38, 435–447.

Schuemans, M., Markovsky, I., Wentzell, P.D., Van Huffel, S., 2005. On the equivalence between total least squares and maximum likelihood PCA. Analytica Chimica Acta 544, 254–267.

Shrivastava, M.K., Subramanian, R., Rogge, W.F., Robinson, A.L., 2007. Sources of organic aerosol: positive matrix factorization of molecular marker data and comparison of results from different source apportionment models. Atmospheric Environment 41, 9353–9369.

Song, Y., Xie, S., Zhang, Y., Zeng, L., Salmon, L.G., Zheng, M., 2006. Source apportionment of $PM_{2.5}$ in Beijing using principal component analysis/absolute principal component scores and UNMIX. Science of the Total Environment 372, 278–286.

Srivastava, A., Jain, V.K., 2007. Size distribution and source identification of total suspended particulate matter and associated heavy metals in the urban atmosphere of Delhi. Chemosphere 68, 579–589.

Tauler, R., 1995. Multivariate curve resolution applied to second order data. Chemometrics and Intelligent Laboratory Systems 30, 133–146.

Tauler, R., Smilde, A.K., Kowalski, B.R., 1995. Selectivity, local rank, 3-way data-analysis and ambiguity in multivariate curve resolution. J. Chemometr. 9, 31–58.

Tauler, R., Lacorte, S., Guillamón, M., Cespedes, R., Viana, P., Barceló, D., 2004. Resolution of main environmental contamination sources of semivolatile organic compounds in surface waters of Portugal using chemometric compounds. Environmental Toxicology and Chemistry 23, 565–575.

Thurston, G.D., Spengler, J.D., 1985. A quantitative assessment of source contribution to inhalable particulate matter pollution in Metropolitan Boston. Atmospheric Environment 19, 9–25.

Van Huffel, S., Vandewalle, J., 1991. The Total Least Squares Problem: Computational Aspects and Analysis. SIAM, Philadelphia.

Viana, M.M., Querol, X., Alastuey, A., Ibarguchi, J.I., Menendez, M., 2006a. Identification of PM sources by principal component analysis (PCA) coupled with wind direction data. Chemosphere 65 (12), 2411–2418.

Viana, M., Zabalza, J., Querol, X., Alastuey, A., Santamaría, J.M., Gil, J.I., Menéndez, M., Hopke, P.K., 2006b. Comparative Analysis of PMF and PCA-MLRA Results for PM_{10} at an Industrial Site in Northern Spain. In: Proceedings of the Summit on Environmental Modelling and Software (IEMSs 2006). University of Vermont, Vermont, USA., ISBN 1-4243-0852-6 978-1-4243-0852-1.

Viana, M., Kuhlbusch, T.A.J., Querol, X., Alastuey, A., Harrison, R.M., Hopke, P.K., Winiwarter, W., Vallius, M., Szidat, S., Prévôt, A.S.H., Hueglin, C., Bloemen, H., Wåhlin, P., Vecchi, R., Miranda, A.I., Kasper-Giebl, A., Maenhaut, W., Hitzenberger, R., 2008. Source apportionment of PM in Europe: a meta-analysis of methods and results. Journal of Aerosol Science 39, 827–849.

Wentzell, P.D., Andrews, D.T., Hamilton, D.C., Faber, K., Kowalski, B.R., 1997. Maximum likelihood principal component analysis. Journal of Chemometrics 11, 339–366.

Wentzell, P.D., Karakach, T.K., Roy, S., Martinez, M.J., Allen, C.P., Werner-Washburne, M., 2006. Multivariate curve resolution of time course microarray data. BMC Bioinformatics 7, 343.