Atmospheric Environment 197 (2019) 159–165

Contents lists available at ScienceDirect

Atmospheric Environment

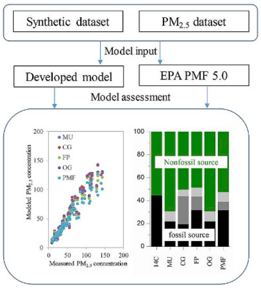
journal homepage: www.elsevier.com/locate/atmosenv

Development and assessment of a receptor source apportionment model based on four nonnegative matrix factorization algorithms

Haitao Liu[a,c], Chongguo Tian[b,*], Zheng Zong[b,d], Xiaoping Wang[e], Jun Li[d], Gan Zhang[d]

[a] School of Economics and Management, Harbin Engineering University, Harbin, 150007, China
[b] Key Laboratory of Coastal Zone Environmental Processes and Ecological Remediation, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, 264003, China
[c] Graduate School, Heilongjiang University of Science & Technology, Harbin, 150022, China
[d] Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Kehua Street No. 511, Guangzhou, 510640, China
[e] Ludong University, Yantai, 264025, China

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Source apportionment
Receptor model
Non-negative matrix factorization
Model development
Model assessment
Radiocarbon measurement

ABSTRACT

This study developed a receptor model, comprising four non-negative matrix factorization algorithms: the multiplicative update method; the optimal gradient method; the highly efficient, monotonic, fixed-point method; and the conjugate gradient method. The feasibility and performance of the developed model for emission source apportionment were assessed, using both a synthetic dataset, and an ambient $PM_{2.5}$ dataset. The results from the US EPA's positive matrix factorization (PMF) 5.0 model were used for the assessment. Modeled results for the synthetic data showed that the range of factor contributions to most matrix elements solved by the four algorithms covered actual values. Modeled results, using the ambient dataset as the input, showed that the four algorithms in the developed model, and the PMF model, identified the same eight emission sources, and apportioned similar source contributions to $PM_{2.5}$. Comparisons between the modeled organic carbon, and the elemental carbon source apportionments and radiocarbon measurements, suggested that combined application of multiple algorithms could satisfactorily apportion emission source contributions for one, or a few, specified samples among a receptor dataset, thus confirming the excellent source apportionment ability of the proposed model.

[*] Corresponding author. Yantai Institute of Coastal Zone Research, CAS, China.
E-mail address: cgtian@yic.ac.cn (C. Tian).

https://doi.org/10.1016/j.atmosenv.2018.10.037
Received 6 June 2018; Received in revised form 18 October 2018; Accepted 21 October 2018
Available online 22 October 2018

1. Introduction

Receptor models have been widely used to quantitatively apportion sources, and their respective contributions to pollutants in the environment. This has been particularly true for the atmospheric environment, in recent years (Belis et al., 2013; Hopke, 2016). Understanding source apportionment is important for designing effective strategies for reduction of contaminant levels in the environment. These models are based on mathematical analysis of pollutant concentrations, measured at a sampling site (receptor site), to infer the source types, and estimate their contributions to monitored site pollutant concentrations (Hopke, 2016). Widely used receptor-based source apportionment models include Principal Component Analysis/Multiple Linear Regression (PCA/MLR), UNMIX, Positive Matrix Factorization (PMF), and Chemical Mass Balance (CMB) (Belis et al., 2013; Hopke, 2016). Provision of source profiles to apportion mass is a prerequisite for the use of the CMB model (Norris and Duvall, 2014) and such requirements are highly restrictive, since the identification of all sources influencing the data at a receptor site is often hard (Baek et al., 1997). The PMF model is thus preferred, because it does not require prepared source profiles to derive a source apportionment (Norris and Duvall, 2014; Wang et al., 2015).

The PMF model decomposes a matrix of speciated sample data into two matrices, to quantify source contributions to the samples. The results of the source apportionments are obtained using the constraint that no sample can have significantly negative source contributions. The source types need to be interpreted and identified by users based on their composition or fingerprints (Norris and Duvall, 2014; Paatero and Tapper, 1994). The PMF model, which is principally an algorithm of alternating non-negative least squares, was proposed and developed, and has been updated, by inclusion of the conjugate gradient method, to enhance its capacity and efficiency (Paatero et al., 2014; Paatero and Tapper, 1994). The method has been applied to the newest version of the PMF model (EPA PMF 5.0) released by the U.S. Environmental Protection Agency (EPA) (Norris and Duvall, 2014). Both the algorithm of alternating nonnegative least squares, and the conjugate gradient method are used widely for solving non-negative matrix factorization (NMF). In fact, PMF is a subset of NMF (Wang and Zhang, 2013). The first NMF algorithm was a multiplicative update method (Lee and Seung, 1999). It has been extensively applied to signal processing, computer vision, machine learning, and data mining (Alexandrov and Vesselinov, 2014; Fu et al., 2016; Karoui et al., 2012). In order to meet the needs of the application in question, several types of algorithms, including gradient descent, quasi-Newton, and hierarchical NMF have been developed so far, and have been further modified and extended for various special purposes (Berry et al., 2007; Cichocki et al., 2008; Laudadio et al., 2016). The existing NMF algorithms can be divided into four categories: basic NMF, constrained NMF, structured NMF, and generalized NMF (Wang and Zhang, 2013). However, these algorithms cannot produce realistic solutions when they are directly applied for source apportionment of environmental contaminants, mainly because, unlike PMF, these algorithms cannot comprehensively consider the uncertainty of sample species (Norris and Duvall, 2014; Paatero et al., 2014).

Multiple receptor modelling algorithms have often been used to apportion source, for sets of environmental data, and more reasonable source contributions could be identified by comparison of modeled results (Nguyen et al., 2013; Shi et al., 2009b; Tauler et al., 2009). Given the limitation of the existing NMF algorithms, and the advantage provided by the aforementioned combined application, the aim of this study was to build an improved source apportionment model. The new model will be based on four different NMF algorithms, and our study will examine its effectiveness by undertaking comparisons with the results derived from the developed model and EPA PMF 5.0, using a synthetic dataset and an ambient $PM_{2.5}$ dataset. Radiocarbon measurement was introduced for the assessment of model effectiveness.

2. Methods and materials

2.1. Model development

The NMF and PMF model algorithms have the same features. They decompose a data matrix (V) into two matrices (W and H) as shown in Eq. (1)

$$V_{ij} = \sum_{r=1}^{p} W_{ir} H_{rj} + e_{ij} \tag{1}$$

where p is the number of factors and e_{ij} is the residual for a component in the ith row and jth column of matrix V. However, in general, they decompose a matrix by minimizing a different objective function. The objective function (Q_{NMF}) of the NMF algorithm is expressed using equation (2).

$$\min Q_{NMF} = f(W, H) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(V_{ij} - \sum_{r=1}^{p} W_{ir} H_{rj} \right)^2 = \|V - WH\|_F^2 \tag{2}$$

where the matrix elements of W and H are subject to constraints greater than or equal to zero. The objective function for the PMF model (Q_{PMF}) is expressed using equation (3).

$$\min Q_{PMF} = f(W, H) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\left(V_{ij} - \sum_{r=1}^{p} W_{ir} H_{rj} \right) / u_{ij} \right]^2 = \| \frac{V - WH}{u} \|_F^2 \tag{3}$$

where the elements of matrices W and H are also constrained to values greater than or equal to zero, while u is the uncertainty matrix, and $||\bullet||_F$ is the Frobenius norm. The difference between the two objective functions is that the PMF model requires user-provided uncertainty data (u) associated with an original matrix (V) to weight each element in V. This design provides users with a convenient manner to consider the confidence in the measurement data. For example, while data below the detection limit can be retained, for use in the model after adjusting for the associated uncertainty, they have less influence on the solution, compared with measurements above the detection limit (Norris and Duvall, 2014; Paatero et al., 2014).

Four NMF algorithms, which were originally used to decompose a data matrix based on Eq. (2), were modified to decompose a data matrix based on Eq. (3). The four NMF algorithms comprise a multiplicative update method (MU) (Lee and Seung, 1999), an optimal gradient method (OG) (Guan et al., 2012), a highly efficient monotonic fixed-point algorithm (FP) (Li and Zhang, 2009) and a conjugate gradient algorithm (CG) (Abd El Aziz and Khidr, 2015). These methods belong to the gradient descent algorithm mentioned above. They iterate to find local minima (Eq. (2)) because the solution domain of a decomposed matrix is not a convex set due to the limitation of the non-negative constraints. The methods start iteration by randomly generating initial matrices of non-negative W and H, and terminate the iteration by setting fixed iteration times, or a maximum allowable tolerance.

After considering the uncertainty matrix, the MU algorithm can be rewritten as shown in equation (4).

$$\begin{aligned} W_{ir}^{k+1} &= W_{ir}^{k} [(V_{ij}/u_{ij}^2)(H_{rj}^k)^T]_{ir} / [(W_{ir}^k H_{rj}^k)/u_{ij}^2 (H_{rj}^k)^T]_{ir} \\ H_{rj}^{k+1} &= H_{rj}^{k} [(W^{k+1})^T (V_{ij}/u_{ij}^2)]_{rj} / [(W_{ir}^k H_{rj}^k)/u_{ij}^2 (H_{rj}^k)^T]_{rj} \end{aligned} \tag{4}$$

where the superscript T indicates partial derivatives to the elements of W and H, and the superscripts k and $k+1$ are iteration counters. The OG method associated with the uncertainty matrix can be expressed by equation (5).

$$
\begin{aligned}
&L_j = \|(W_{ir}/u_{ij})^T(W_{ir}/u_{ij})\|_2\\
&H_{rj}^{k+1} = \max\left\{\left[{Y_{rj}}^k - \frac{{W_{ir}}^T}{L_j}\left(\frac{W_{ir}{H_{rj}}^k - V_{ij}}{u_{ij}^2}\right)\right], 0\right\}\\
&\alpha_{k+1} = (1 + \sqrt{4\alpha_k^2 + 1})/2\\
&Y_{rj}^{k+1} = H_{rj}^k + \frac{\alpha_k - 1}{\alpha_{k+1}}(H_{rj}^k - H_{rj}^{k-1})
\end{aligned}
\tag{5}
$$

where the meanings of superscripts *T*, *k*, and *k + 1* are the same as those in equation (4), and $||\bullet||_2$ is the Euclidean norm. The method alternatively updates *H*, *α*, and *Y*, until convergence to obtain an optimal solution. The FP method associated with the uncertainty matrix can be expressed as shown in equations (6) and (7).

$$
W_{il}^{k+1} = \begin{cases} \max\left[\left(\left(\frac{V_{ij}}{u_{ij}^2}H_{lj}^{T^k}\right) - \sum_{r=1, r\neq l}^{p}\frac{W_{ir}^k H_{rj}^k}{u_{ij}^2}H_{rj}^{T^k}\right)/(HH^T)_{ll}^k, 0\right] & \|H_{l\cdot}\| > 0\\ rand & \|H_{l\cdot}\| = 0 \end{cases}
\tag{6}
$$

$$
H_{lj}^{k+1} = \begin{cases} \max\left[\left(\left(W_{il}^{T^k}\frac{V_{ij}}{u_{ij}^2}\right) - \sum_{r=1, r\neq l}^{p} W_{ir}^{T^k}\frac{W_{ir}^{k+1}H_{rj}^k}{u_{ij}^2}\right)/(W^TW)_{ll}^k, 0\right] & \|W_{\cdot l}\| > 0\\ 0 & \|W_{\cdot l}\| = 0 \end{cases}
\tag{7}
$$

where the meanings of the superscripts *T*, *k*, and *k + 1* are the same as those in equations (4) and (5). The *rand* in equation (6) indicates that W_{il} at the *k* + 1st iteration is set as a random number, when $||W_{\cdot l}||$ is equal to zero. The CG algorithm associated with the uncertainty matrix can be written as shown in equation (8).

$$
\begin{aligned}
&A = \|(W_{ir}/u_{ij})^T(W_{ir}/u_{ij})\|_2\\
&d^0 = g^0 = -{W_{ir}}^T\left(\frac{W_{ir}{H_{rj}}^k - V_{ij}}{u_{ij}^2}\right)\\
&H_{rj}^{k+1} = \max\left({H_{rj}}^k - \frac{g^k g^k}{d^{Tk}Ad^k}d^k, 0\right)\\
&r^{k+1} = r^k - \alpha^k A d^k\\
&d^{k+1} = g^k + \frac{g^{k+1}g^{k+1}}{d^k g^k}d^k
\end{aligned}
\tag{8}
$$

where the meanings of the superscripts *T*, *k*, and *k + 1* are the same as those in equations (4)–(6), and $||\bullet||_2$ is the Euclidean norm. The method iteratively updates *d*, *H*, and *r*, to score the optimal solution.

2.2. Experimental setup

The four NMF algorithms were compiled using Matlab R2016b software, and two experiments were conducted to assess model performance. To identify the most optimal factor contributions and profiles, two model experiments were run 100 times, with each experiment commencing from a different starting point. The same iteration times and starting point design were used for the final solution of the PMF modelling (Norris and Duvall, 2014). The original matrices, as the starting points for each run, were randomly generated and systematically modified using their respective approaches to chart the optimal path to the best-fit solution (global minimum). The best solution was typically identified by the lowest Q value along the path (equation (3)). A simulation run starting from random matrices cannot guarantee that the solution is the best solution, so therefore, to maximize the chance of reaching the global minimum, it may instead find a local minimum from the 100 runs mentioned above.

The first experiment used a synthetic matrix. To build a synthetic matrix as the model input (matrix *V*, described above), a matrix with 23 rows and eight columns identified by the PMF model in a previous study (Zong et al., 2016) was adopted for matrix *W* and a random matrix with eight rows and 100 columns was generated for matrix *H*, as described above. The matrices *W* and *H* were termed W_O and H_O, respectively, in order to distinguish them from the later-modeled matrices, *W* and *H*. Matrix *V* was produced by multiplying matrix W_O with matrix H_O. An uncertainty matrix *u* was generated by multiplying matrix *V* by 0.1. Both matrices *V* and *u* were used as input data, to derive both the developed model, and the PMF model (Norris and Duvall, 2014).

The other experiment used a matrix of ambient $PM_{2.5}$ components as the model input for *V*. The $PM_{2.5}$ samples were collected at the sampling platform of a national station for background atmospheric monitoring in North China from December 2011 to January 2013. The national station is located on Tuoji Island, a small island with an area of 7.1 km^2 located at the demarcation line between the Bohai Sea and Yellow Sea as shown in Fig. S1 of the Supporting Information (SI). The concentration data for organic carbon (OC), elemental carbon (EC), water-soluble ions (i.e., sodium [Na^+], ammonium [$NH_4{}^+$], potassium [K^+], magnesium [Mg^{2+}], calcium [Ca^{2+}], chloride [Cl^-], nitrate [$NO_3{}^-$], and sulfate [$SO_4{}^{2-}$]), and metals (i.e., vanadium [V], manganese [Mn], iron [Fe], chromium [Cr], nickel [Ni], copper [Cu], zinc [Zn], arsenic [As], cadmium [Cd], and lead [Pb]), in $PM_{2.5}$, formed the input matrix (*V*), and its associated uncertainty matrix (*u*). The matrices *V* and *u* were those used to derive the PMF model for model assessment in a previous study (Wang et al., 2017). Sample information, and details of the chemical analyses, are presented in the text of SI, and our previous study (Wang et al., 2014, 2017). The $PM_{2.5}$ concentrations and chemical components are summarized in Table S1 of SI. The same *V* and *u* matrices were applied as input data to derive the developed model and the PMF model.

2.3. Assessment of model performance

For the model experiment with a synthetic matrix as input, changes of Q values with factor number increase were used to assess the capacity for determining source number. Decreasing Q values, with increasing iterations, was used to assess calculative efficiency. The *W* matrices decomposed by the developed model and PMF model were compared with the matrix W_O, to examine calculation precision. Average absolute error (AAE) of the overall source contributions to total mass, in the synthetic dataset, was used as an aggregative indicator, to examine the calculating precision, which can be calculated as shown in equation (9) (Shi et al., 2009b).

$$
AAE = \frac{1}{n}\cdot\sum_{i=1}^{n}\left|\frac{a_i - b_i}{b_i}\right|\cdot 100\%
\tag{9}
$$

where n is the number of the sources, a_i is the estimated contribution of the ith source, and b_i is the true contribution of the ith source. A low value for AAE indicates that the estimated contributions are close to the true values. Collinearity was examined by correlation analysis among factors in matrix *W*.

For the model experiment with ambient $PM_{2.5}$ data as input, the overall source contributions to the total mass, solved by the four algorithms and the PMF model, were compared. Model performance was assessed in more detail by comparing the source apportionment results of carbonaceous components (OC and EC) in $PM_{2.5}$, with radiocarbon (^{14}C) measurements. Recent studies have shown that ^{14}C measurements can unambiguously discriminate between fossil and non-fossil sources of carbonaceous particles, as ^{14}C is completely depleted in fossil fuel emissions due to its half-life of 5730 years, whereas non-fossil carbon sources (e.g., from biomass burning or biogenic emissions) show ^{14}C levels similar to those of atmospheric CO_2 (Liu et al., 2013, 2014; Zhang et al., 2015). Thus, ^{14}C measurements of OC and EC fractions can directly quantify their fossil and non-fossil source contributions. Four seasonally merged samples, and three outlier samples, of ^{14}C measurements, were used for the assessment. The four samples were winter of 2011, and spring, summer, and autumn of 2012, termed as winter, spring, summer, and autumn, respectively, for later analysis. The three outlier samples were those with the highest OC, EC, and $PM_{2.5}$ concentrations for the entire sampling period, and were termed OC_{max}, EC_{max}, and PM_{max}, respectively, for later analysis. For the comparison, the modeled source contributions were classified into two carbon source groups (fossil and non-fossil). The contribution fractions of fossil or non-fossil carbon sources to OC and EC were subsequently compared to the ^{14}C results of the seven samples. The contribution fractions (R) of non-fossil or fossil sources to OC or EC, classified from the developed model and PMF results, were determined by using the formula shown in equation (10).

$$R_{ij} = \sum_{k=1}^{n} W_{ik}H_{kj} \Big/ \sum_{k=1}^{p} W_{ik}H_{kj} \tag{10}$$

where W and H are the factor contributions and factor profiles, respectively, i represents the OC or EC species, j is a specified sample, n is the number of fossil or non-fossil carbon sources, and p is the total number of sources (Zong et al., 2016).

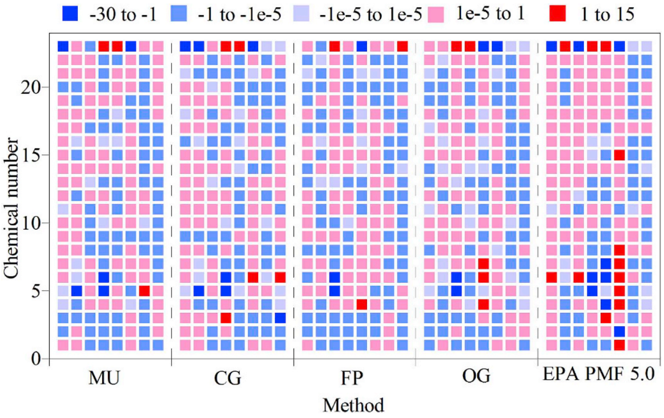


Fig. 1. Differences between matrix W decomposed by the four developed algorithms and EPA PMF 5.0 for matrix W_0.

3. Results and discussion

3.1. Model assessment using synthetic data

Figure S2 of SI shows variation of Q values calculated by the four algorithms, with modelling factor numbers increased from six to ten. The Q values calculated by all four algorithms showed the most significant declines in the model experiments with eight factors, indicating that the fitting results simulated by the model experiments achieved more significant more significant improvement. The factor number equaled the source number to generate the synthetic matrix used in this study, as mentioned above, suggesting that these algorithms could capture the number of main sources quite well (Wang et al., 2017).

Figure S3 of SI displays the changes to the Q values brought about through iterations modeled by the four developed NMF algorithms, with eight factors. The four algorithms were subjected to iterations less than or equal to 20000, or as long as the modeled Q value was less than or equal to 2.0×10^{-5}. The MU method showed the smoothest downtrend in Q values among the four algorithms, and this value reduced to 98×10^{-5}, until the iterations reached the specified upper limit. Both the OG and FP algorithms generated a moderate downward gradient for their Q values. The OG and FP algorithms reached the critical limitation of 2.0×10^{-5}, after 2534 and 1479 iterations, respectively. The CG algorithm had the most rapid convergence, and its Q value achieved the critical level of 2.0×10^{-5} after just 65 iterations. This highly efficient performance supported replacing the alternating non-negative least square algorithm with the conjugate gradient method, in the PMF model (Norris and Duvall, 2014; Paatero et al., 2014). The iterations and Q value assessed by the PMF model, using the synthetic matrix as input data are also shown in Fig. S3 of SI. After 290 iterations, the PMF model obtained its optimal solution, when the corresponding Q value was equal to 3.0×10^{-5}. The number of iterations by the PMF model was three times that of the CG method used in present study, although this number was clearly lower than those for the other three algorithms in the developed model. The large number of iterations for the PMF model might be due to additional considerations in this model, because the CG algorithm has been termed as Multilinear Engine-2, in the PMF model (Norris and Duvall, 2014; Paatero et al., 2014). The Q value of the PMF model was larger than the Q values determined by the OG, FP, and CG algorithms, and this was considered as likely to be due to additional considerations in the PMF model. The low Q values shown in Fig. S3 of SI indicated that, in general, the four algorithms and the PMF model could decompose the synthetic matrix satisfactorily. The downtrends of Q values indicated that CG was the most efficient matrix decomposition method, followed by FP, OG, and MU, among the four algorithms in the developed model.

To understand model performance in more detail, the modeled errors were analyzed. Input data for the original matrix V was generated, by multiplying W_0 and H_0 (as described in the method section), indicating that once matrix W or H was determined by these decomposition methods, the other matrix was produced deterministically. Thus, only W matrices were used to assess the performance of the algorithms, by comparing them with matrix W_0. Matrices of W, obtained by the four algorithms in the developed model and PMF model, are displayed in Tables S2–S6 of SI. The large differences between matrices W and W_0 require more attention, because these differences could possibly distract model users from the truth, and affect source diagnoses in real-life emission source apportionment. The differences between matrices W and W_0, generated by the four algorithms and the PMF model, were merged into a data column and the column was sorted into descending order. In this study, the first 2.5% and last 2.5% of the data column were defined as data showing a significant difference (≥ 1.0 or ≤ -1.0) from the overall level. These significant differences were distributed again into the four algorithms and the PMF model, as displayed in Fig. 1. The MU, CG, FP, OG algorithms and the PMF model had 8, 12, 6, 9, and 22 such differences, respectively. The number difference was attributed to them calculated by different algorithms and random startup. A comparison of these differences showed both positive and negative deviations for a specified element. For instance, the difference in the first row and first column was less than zero, for the algorithms MU and CG, and the PMF model, and positive for FP and OG. This is an important property for source apportionment of contaminants if the algorithms are combined, because the actual values fall within the variation ranges of the differences. This property is similar to the finding that the combined application of different receptor models can effectively improve source resolution results (Shi et al., 2009a, 2009b; Tue et al., 2013). Correlation analysis showed insignificant correlation among source factors in matrices W solved by the four algorithms (see Table S7 of SI), indicating these algorithms are well to overcome the near collinearity problem (Shi et al., 2009a).

The average source contributions, assessed by the four algorithms and PMF model, to the total mass of all samples in the synthetic dataset, and their AAE, are listed in Table 1. According to the AAEs, the results from the MU algorithm were the closest to true values, followed by the algorithms CG, OG, and FP, and the PMF model. The assessments of model performance were not inconsistent with the results assessed by the comparison between matrices W and W_0. For instance, the FP method had the least number of significant differences between matrices W and W_0, while exhibiting largest AAE value. This inconsistency makes it difficult to determine which among the four algorithms is best, but fortunately, the contribution values from all the sources were within the scopes of the source contributions estimated by the four algorithms, as shown in Table 1. The averaged AAEs of the four algorithms were also less than those of each algorithm, and combination of

Table 1
Average source contributions to the synthetic matrix.

Source	Synthetic contributions	Estimated contributions						
		MU	OG	FP	CG	PMF	Ave(4)	Ave(5)
Source1	11.45	10.72	11.90	11.98	10.54	7.58	11.28 ± 0.66	10.54 ± 1.60
Source2	15.07	16.20	18.18	17.54	14.58	10.20	16.63 ± 1.38	15.34 ± 2.85
Source3	11.35	11.32	10.30	10.01	15.50	7.71	11.78 ± 2.20	10.97 ± 2.56
Source4	13.66	14.63	14.65	14.87	13.60	8.89	14.44 ± 0.49	13.33 ± 2.26
Source5	12.84	11.86	8.91	11.01	12.85	14.54	11.11 ± 1.39	11.79 ± 1.85
Source6	12.05	10.50	12.06	9.91	9.96	13.47	10.61 ± 0.87	11.18 ± 1.39
Source7	11.86	11.75	11.91	12.50	11.63	23.43	11.95 ± 0.33	14.24 ± 4.60
Source8	11.71	13.02	12.09	12.18	11.03	14.19	12.08 ± 0.71	12.50 ± 1.05
AAE		6.73	9.43	10.37	9.15	34.63	6.28	7.19

Note: Ave(4) and AVE(5) are averaged results of the four algorithms, and the four algorithms and PMF model, respectively.

these algorithms and PMF model, suggesting that the combined application of the four algorithms could achieve a narrower range of source contributions, covering true source contribution. Therefore, the four algorithms could be seen as single entity to apportion sources of ambient $PM_{2.5}$ data.

3.2. Model assessment using ambient $PM_{2.5}$ data

Using the findings from our previous study (Wang et al., 2017), model experiments of eight factors were performed, using the developed model, and the results were compared with those simulated by applying the PMF model (Wang et al., 2017). The modeled concentrations of each factor for the four algorithms were correlated with those provided by the PMF model and the correlation coefficients are listed in Table S8 of SI. In general, seven factors had significantly high correlation coefficients, with seven sources identified by the PMF model, indicating their good correspondence. The seven factors were traffic dust, industrial processes, biomass burning, vehicle emissions, mineral dust, shipping emissions, and sea salt. The last factor generated by the four algorithms did not correspond very highly to the factor of coal combustion solved by the PMF model, but correlated well with the factor of industrial processes (see Table S8 of SI). The ratios of OC to EC calculated by the algorithms MU (3.71), CG (3.89), FP (0.79), and OG (2.93) were higher than that derived by the PMF model (0.68). The high ratios indicated that the last factor solved by the four algorithms contained more combustion signals from domestic coal (Cao et al., 2007; Wang et al., 2014). These analyses suggested that the last factor could also be considered to be sourced from coal combustion.

Fig. S4 of SI displays a scatter plot of measured $PM_{2.5}$ concentrations versus modeled $PM_{2.5}$ concentrations by the four algorithms and the PMF model. The modeled and measured $PM_{2.5}$ concentrations were regarded as dependent variables and independent variables, respectively, and a simple linear regression analysis was performed, with the regression equations shown in Fig. S4 of SI. The slopes are close to values of one, and these high correlation coefficients of the regression equations indicate that these algorithms, and the PMF model, captured the temporal variation in $PM_{2.5}$ concentrations well. The algorithms MU, CG, and OG have slopes closer to one, and smaller intercepts, than those provided by the FP and PMF models, suggesting the stronger apportionment capability of these three algorithms.

Fig. 2 displays the averaged contributions of the eight sources to the $PM_{2.5}$ mass concentration apportioned by the four algorithms and the PMF model. In general, these methods make similar source contributions to the $PM_{2.5}$ mass concentrations, as indicated by the significant correlation coefficients among them in Table S9 of SI. Statistical source contributions to $PM_{2.5}$ mass concentrations apportioned by these four algorithms showed that biomass burning, coal combustion, shipping emissions, mineral dust, sea salt, industrial processes, vehicle emissions, and traffic dust contributed 25.5% ± 5.64%, 15.6% ± 3.99%, 13.3% ± 3.04%, 13.2% ± 2.18%, 9.54% ± 3.06%, 9.41% ± 1.43%, 9.02% ± 1.67%, and 4.38% ± 3.28%, to the $PM_{2.5}$ mass concentrations, respectively. Biomass burning was identified as the largest $PM_{2.5}$ mass concentrations contributor, although the value was lower than that apportioned (27.5%) by the PMF model. A review study compared the source apportionments of $PM_{2.5}$ solved by the PMF and CMB models, and found that the former's solutions showed a strong possibility that the contribution of biomass burning to $PM_{2.5}$ was overestimated (Zhang et al., 2017). The relatively low contribution of biomass burning could suggest that the combined application of the four algorithms overcame the overestimation of the contribution of biomass burning to $PM_{2.5}$, apportioned by the PMF model. In addition, such contribution fractions (mean ± standard deviation) were likely to be more accurate than an exact value of source apportionment, identified by a model, because the actual values fall within the contribution range of synthetic matrix V, solved by the four algorithms, as illustrated in Section 3.1.

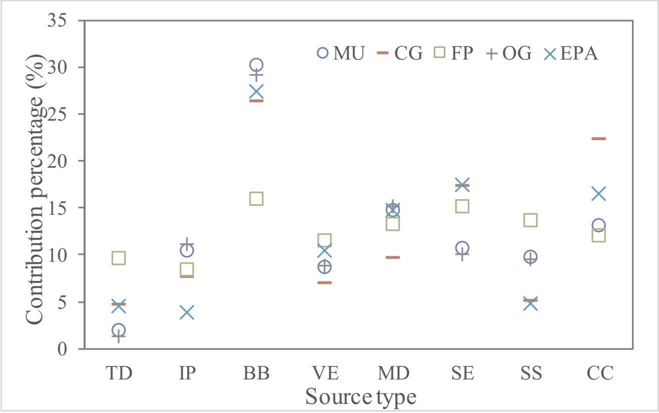


Fig. 2. Averaged contributions of eight sources to $PM_{2.5}$ mass concentrations solved by the MU, OG, FP, and CG algorithms, and the PMF model. The eight sources are traffic dust (TD), industrial processes (IP), biomass burning (BB), vehicle emissions (VE), mineral dust (MD), shipping emissions (SE), sea salt (SS), and coal combustion (CC).

3.3. Model assessment based on ^{14}C measurements

According to the source types described in Section 3.2, coal combustion, shipping emissions, vehicle emissions, and industrial processes were classified as fossil sources, while biomass burning and sea salt were merged into the non-fossil source (Wang et al., 2017; Zong et al., 2016). Mineral dust and traffic dust were not considered in this classification, as they could not be apportioned quantitatively into either fossil or non-fossil sources. The contributions of the two unsorted sources make us assess only irrelevant source apportionments, by the algorithms and the PMF model, according to the overestimation of the source contributions classified from the simulated results, compared

Table 2
Comparison of ^{14}C measurement and source contributions of fossil and non-fossil sources to OC and EC apportioned by the four algorithms and PMF model.

Sample	Method	OC		EC	
		Fossil	Non-fossil	Fossil	Non-fossil
Winter	^{14}C measurement	44.6	55.4	58.2	41.8
	Four algorithms	21.9–43.9	56.1–78.1	44.0–63.6	36.4–56.0
	EPA PMF	39.1	60.9	55.4	44.6
Spring	^{14}C measurement	40.4	59.6	58.2	41.8
	Four algorithms	15.0–38.3	61.7–85.0	41.8–62.7	37.3–58.2
	EPA PMF	29.8	70.2	45.1	54.9
Summer	^{14}C measurement	32.8	67.2	48.5	51.5
	Four algorithms	16.2–38.7	61.3–83.8	34.4–55.8	44.2–65.6
	EPA PMF	35.1	64.9	54.7	45.3
Autumn	^{14}C measurement	41.8	58.2	55.1	44.9
	Four algorithms	9.64–29.8	70.3–90.4	28.3–50.0	50.0–71.7
	EPA PMF	24.7	75.3	40.6	59.4
EC_{Max}	^{14}C measurement	45.4	54.6	54.1	45.9
	Four algorithms	14.8–48.7	51.3–85.2	35.7–67.4	32.6–64.3
	EPA PMF	62.5	37.5	77.4	22.6
OC_{Max}	^{14}C measurement	21.2	78.8	43.3	56.7
	Four algorithms	8.82–33.1	66.9–91.2	27.2–54.6	45.4–72.8
	EPA PMF	23.2	76.8	36.8	63.2
PM_{Max}	^{14}C measurement	40.1	59.9	59.9	40.1
	Four algorithms	7.57–29.0	71.0–92.4	42.4–73.2	26.8–57.6
	EPA PMF	14.3	85.7	23.9	76.1

with ^{14}C measurements (Zong et al., 2016). Such a comparison assessment was described in the SI, as text, Table S10, and Fig. S5.

Given that traffic dust is thought to contain a higher fraction of fossil carbon than mineral dust, we classified traffic dust as an additional fossil source, and mineral dust as an additional non-fossil source to estimate the performance of the developed model. The method was confirmed as a suitable tool to assess the modelling capacity of the PMF model in our previous study (Wang et al., 2017). After this addition, the comparisons between ^{14}C measurements and the contributions from fossil and non-fossil sources to OC and EC apportioned by the four algorithms, and the PMF model, have been listed in Table 2. Most ^{14}C measurements were in the range of the source contributions apportioned by the four algorithms, indicating that the developed model could give more reasonable source apportionments than those identified by the PMF model, particularly with respect to the three outlier samples. A previous study demonstrated that the PMF model could capture the source apportionments of the four seasonal samples better than those of the three outlier samples (Wang et al., 2017). The developed model could capture the source apportionments for one or a few specified samples better, demonstrating its excellent capacity for diagnosing source contributions within a short period, such as a haze event.

4. Conclusions

In this study, a receptor model including four NMF algorithms, namely MU, OG, FP, and CG, was developed for emission source apportionment. To examine the feasibility and performance of the model, synthetic and ambient $PM_{2.5}$ datasets were decomposed and analyzed by the developed model. In addition, the results of the PMF model were used for the assessment. In the scenario with the synthetic dataset, the CG algorithm showed the least convergence steps, followed by the FP and OG algorithms under the same error tolerance, while the MU algorithm had the slowest convergence speed. The range of factor contributions to most matrix elements solved by the four algorithms covered the actual values. The model scenario using the ambient dataset as input showed that both the developed model and the PMF model could satisfactorily capture temporal variations to $PM_{2.5}$ concentrations. The four algorithms in the developed model and the PMF model identified the same eight emission sources, and apportioned similar source contributions of $PM_{2.5}$. Comparison between the modeled OC and EC source apportionments, and ^{14}C measurements suggested that the combined application of multiple algorithms could adequately apportion the source contributions for some relatively complicated receptor samples.

Notes

The authors declare no competing financial interest.

Acknowledgment

This work was financially supported by Key Laboratory of Coastal Environmental Processes and Ecological Remediation of the Chinese Academy of Sciences (Grant No: 2016KFJJ01), and the Natural Scientific Foundation of China (NSFC; Grant Nos: 41471413 and 41473101).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2018.10.037.

References

Abd El Aziz, M., Khidr, W., 2015. Nonnegative matrix factorization based on projected hybrid conjugate gradient algorithm. Signal Image Video P 9, 1825–1831.

Alexandrov, B.S., Vesselinov, V.V., 2014. Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. Water Resour. Res. 50, 7332–7347.

Baek, S.-O., Choi, J.-S., Hwang, S.-M., 1997. A quantitative estimation of source contributions to the concentrations of atmospheric suspended particulate matter in urban, suburban, and industrial areas of Korea. Environ. Int. 23, 205–213.

Belis, C.A., Karagulian, F., Larsen, B.R., Hopke, P.K., 2013. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmos. Environ. 69, 94–108.

Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J., 2007. Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. 52, 155–173.

Cao, J.J., Lee, S.C., Chow, J.C., Watson, J.G., Ho, K.F., Zhang, R.J., Jin, Z.D., Shen, Z.X., Chen, G.C., Kang, Y.M., Zou, S.C., Zhang, L.Z., Qi, S.H., Dai, M.H., Cheng, Y., Hu, K., 2007. Spatial and seasonal distributions of carbonaceous aerosols over China. J. Geophys. Res. 112, D22S11.

Cichocki, A., Zdunek, R., Amari, S.i., 2008. Nonnegative matrix and tensor factorization [Lecture Notes]. IEEE Signal Process. Mag. 25, 142–145.

Fu, X., Sidiropoulos, N.D., Ma, W.K., 2016. Power spectra separation via structured matrix factorization. IEEE Trans. Signal Process. 64, 4592–4605.

Guan, N., Tao, D., Luo, Z., Yuan, B., 2012. NeNMF: an optimal gradient method for nonnegative matrix factorization. IEEE Trans. Signal Process. 60, 2882–2898.

Hopke, P.K., 2016. A review of receptor modeling methods for source apportionment. J. Air Waste Manag. Assoc. 66, 237–259.

Karoui, M.S., Deville, Y., Hosseini, S., Ouamri, A., 2012. Blind spatial unmixing of multispectral images: new methods combining sparse component analysis, clustering and non-negativity constraints. Pattern Recogn. 45, 4263–4278.

Laudadio, T., Croitor Sava, A.R., Sima, D.M., Wright, A.J., Heerschap, A., Mastronardi, N., Van Huffel, S., 2016. Hierarchical non-negative matrix factorization applied to three-dimensional 3 T MRSI data for automatic tissue characterization of the prostate. NMR Biomed. 29, 751–758.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.

Li, L., Zhang, Y.-J., 2009. FastNMF: highly efficient monotonic fixed-point nonnegative matrix factorization algorithm with good applicability. J. Electron. Imag. 18, 033004.

Liu, D., Li, J., Zhang, Y., Xu, Y., Liu, X., Ding, P., Shen, C., Chen, Y., Tian, C., Zhang, G., 2013. The use of levoglucosan and radiocarbon for source apportionment of $PM_{2.5}$ carbonaceous aerosols at a background site in East China. Environ. Sci. Technol. 47, 10454–10461.

Liu, J., Li, J., Zhang, Y., Liu, D., Ding, P., Shen, C., Shen, K., He, Q., Ding, X., Wang, X., Chen, D., Szidat, S., Zhang, G., 2014. Source apportionment using radiocarbon and organic tracers for $PM_{2.5}$ carbonaceous aerosols in Guangzhou, South China: contrasting local- and regional-scale haze events. Environ. Sci. Technol. 48, 12002–12011.

Nguyen, Q.T., Skov, H., Sørensen, L.L., Jensen, B.J., Grube, A.G., Massling, A., Glasius, M., Nøjgaard, J.K., 2013. Source apportionment of particles at station nord, North east Greenland during 2008–2010 using COPREM and PMF analysis. Atmos. Chem. Phys. 13, 35–49.

Norris, G., Duvall, R., 2014. EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide. U.S. Environmental Protection Agency National Exposure Research Laboratory, Washington, pp. 1–124.

Paatero, P., Eberly, S., Brown, S.G., Norris, G.A., 2014. Methods for estimating uncertainty in factor analytic solutions. Atmos. Meas. Tech. 7, 781–797.

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.

Shi, G.-L., Feng, Y.-C., Zeng, F., Li, X., Zhang, Y.-F., Wang, Y.-Q., Zhu, T., 2009a. Use of a nonnegative constrained principal component regression chemical mass balance model to study the contributions of nearly collinear sources. Environ. Sci. Technol. 43, 8867–8873.

Shi, G.-L., Li, X., Feng, Y.-C., Wang, Y.-Q., Wu, J.-H., Li, J., Zhu, T., 2009b. Combined source apportionment, using positive matrix factorization–chemical mass balance and principal component analysis/multiple linear regression–chemical mass balance models. Atmos. Environ. 43, 2929–2937.

Tauler, R., Viana, M., Querol, X., Alastuey, A., Flight, R.M., Wentzell, P.D., Hopke, P.K., 2009. Comparison of the results obtained by four receptor modelling methods in aerosol source apportionment studies. Atmos. Environ. 43, 3989–3997.

Tue, N.M., Takahashi, S., Subramanian, A., Sakai, S., Tanabe, S., 2013. Environmental contamination and human exposure to dioxin-related compounds in e-waste recycling sites of developing countries. Environ. Sci. Proc. Impacts 15, 1326–1331.

Wang, F., Lin, T., Feng, J., Fu, H., Guo, Z., 2015. Source apportionment of polycyclic aromatic hydrocarbons in PM2.5 using positive matrix factorization modeling in Shanghai, China. Environ. Sci. Proc. Impacts 17, 197–205.

Wang, X., Chen, Y., Tian, C., Huang, G., Fang, Y., Zhang, F., Zong, Z., Li, J., Zhang, G., 2014. Impact of agricultural waste burning in the Shandong Peninsula on carbonaceous aerosols in the Bohai Rim, China. Sci. Total Environ. 481, 311–316.

Wang, X., Zong, Z., Tian, C., Chen, Y., Luo, C., Li, J., Zhang, G., Luo, Y., 2017. Combining positive matrix factorization and radiocarbon measurements for source apportionment of $PM_{2.5}$ from a national background site in North China. Sci. Rep. 7, 10648.

Wang, Y., Zhang, Y., 2013. Nonnegative matrix factorization: a comprehensive review. IEEE Trans. Knowl. Data Eng. 25, 1336–1353.

Zhang, Y.-L., Schnelle-Kreis, J., Abbaszade, G., Zimmermann, R., Zotter, P., Shen, R.-r., Schäfer, K., Shao, L., Prévôt, A.S.H., Szidat, S., 2015. Source apportionment of elemental carbon in Beijing, China: insights from radiocarbon and organic marker measurements. Environ. Sci. Technol. 49, 8408–8415.

Zhang, Y., Cai, J., Wang, S., He, K., Zheng, M., 2017. Review of receptor-based source apportionment research of fine particulate matter and its challenges in China. Sci. Total Environ. 586, 917–929.

Zong, Z., Wang, X., Tian, C., Chen, Y., Qu, L., Ji, L., Zhi, G., Li, J., Zhang, G., 2016. Source apportionment of $PM_{2.5}$ at a regional background site in North China using PMF linked with radiocarbon analysis: insight into the contribution of biomass burning. Atmos. Chem. Phys. 16, 11249–11265.