Atmospheric Environment 43 (2009) 4951–4958

Contents lists available at ScienceDirect

Atmospheric Environment

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

Sensitivity of a molecular marker based positive matrix factorization model to the number of receptor observations

YuanXun Zhang [a], Rebecca J. Sheesley [a], Min-Suk Bae [a], James J. Schauer [a,b,*]

[a] *Environmental Chemistry and Technology Program, University of Wisconsin–Madison, 660 North Park Street, Madison, WI 53706, USA*
[b] *Wisconsin State Laboratory of Hygiene, University of Wisconsin–Madison, 2601 Agriculture Drive, Madison, WI 53718, USA*

ARTICLE INFO

Article history:
Received 7 February 2009
Received in revised form
3 June 2009
Accepted 7 July 2009

Keywords:
Source attribution
Factor analysis
Organic aerosol

ABSTRACT

To investigate the impact of the number of observations on molecular marker-based positive matrix factorization (MM-PMF) source apportionment models, daily $PM_{2.5}$ samples were collected in East St. Louis, IL, from April 2002 through May 2003. The samples were analyzed for daily 24-h average concentrations of elemental and organic carbon, trace elements, and speciated particle-phase organic compounds. A total of 273 sets of observations were used in the model and consisted of all valid sets of observations from the year long data set minus one sixth of the measurements, which were collected every 6th day and were analyzed by different chemical analysis techniques. In addition to the base case of 273 samples, systematic subsets of the data set were analyzed by PMF. These subsets of data included 50% of the observations (135–138 days), 33% of the observations (90–92 days) and 20% of the observations (52–56 days). In addition, model runs were also examined that used 48-h, 72-h, 6-day, and weekly average concentrations as model inputs. All MM-PMF model runs were processed following the same procedures to explore the stability of the source attribution results. Consistent with previous MM-PMF results for East St. Louis, the main sources of organic aerosol were found to be mobile sources, secondary organic aerosols (SOAs), resuspended soil and biomass combustions, as well as an *n*-alkane dominated point source and other combustion sources. The MM-PMF model was reasonably stable when the number of observations in the input was reduced to ninety, or approximately 33% of observations present in the base case. In these cases, the key factors including resuspended soil, mobile and secondary factors, which accounted for more than 70% of the measured OC concentrations, were stable as defined by a relative standard deviation (RSD) of less than 30%. Similar results were obtained from the smaller data subsets, but resulted in larger uncertainties, with several of these factors yielding RSD of greater than 30%. The three factors with the largest OC contributions were more stable than the other minor factors, even when the number of observations was nominally 50 days. Secondary organic aerosol (SOA) was the most stable factor observed in the model runs. Since it is unclear if these results can be broadly applied to all MM-PMF models, additional studies of this nature are needed to assess the broader applicability of these conclusions. Until such studies are implemented, this paper provides a foundation to design future studies in sampling strategies for source apportionment using MM-PMF.

1. Introduction

An important capability needed within the atmospheric aerosol research community and for air pollution control development is source apportionment, which assigns the pollutant concentrations or their associated environmental impacts to sources. During the past several decades, different source apportionment models have been developed, which provide quantitative information about source contributions to support control strategy design (Cass, 1998; Chow and Watson, 2002). One such model, positive matrix factorization (PMF), has been widely used in many source apportionment studies in a variety of locations (e.g., Hopke et al., 2006; Jaeckels et al., 2007; Kim et al., 2003a; Paatero et al., 2005; Shrivastava et al., 2007; Zhang et al., 2008a).

The PMF model is a multivariate factor analysis tool based on the mass balance hypothesis, which states that the ambient pollutants are linearly summed from a number of time-variant source profiles. Therefore, PMF is an approach to find a new, reduced dimensional

---

* Corresponding author at: Environmental Chemistry and Technology Program, University of Wisconsin–Madison, 660 North Park Street, Madison, WI 53706, USA. Tel.: +1 608 262 4495; fax: +1 608 262 0454.
E-mail address: jjschauer@wisc.edu (J.J. Schauer).

1352-2310/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2009.07.009

set of basis vectors that can reproduce the original data. In other words, it determines unknown source profiles and source contributions from a time series of observations (Polissar et al., 2001; USEPA, 2008).

Typically, PMF uses trace elements, and organic and elemental carbon (OC/EC), as well as secondary ions, as the input data matrix to explore the "co-variances" between species and their associated sources (e.g., Kim and Hopke, 2004; Kim et al., 2003a). In these traditional PMF models, which use trace elements, EC, OC and ions, large data sets of observations have been used to obtain more robust quantitative source contribution assessments. Thus, many PMF studies apply one or more years of observations as the input data matrix (e.g., trace elements PMF studies such as Kim and Hopke, 2004; Kim et al., 2003a; Paatero and Hopke, 2003).

The use of tracers with high source specificity in PMF models can enhance the interpretation of the factors derived from the model. Consequently, the use of organic molecular markers in PMF has resulted in considerable progress in source attribution, as shown in the recent molecular marker PMF (MM-PMF) studies by Jaeckels et al. (2007) and Shrivastava et al. (2007).

Given the complexity and costs associated with the measurement of molecular markers, there are usually fewer observations available for MM-PMF studies than traditional PMF models. The typical number of observations used in an MM-PMF model is around 100 daily observations according to the previous studies by Jaeckels et al. (2007), Shrivastava et al. (2007) and Bullock et al. (2008), which have 120, 99 and 96 receptor observations, respectively. Notably, all the aforementioned studies have observed relative reliable and robust MM-PMF results, which have been evaluated in the context of existing knowledge of sources in these study areas and other receptor models such as Chemical Mass Balance (CMB) models. The consistencies and biases of these different models have been well documented. A critical question in the application and interpretation of MM-PMF studies is whether or not larger numbers of observations can significantly enhance the stabilities of the MM-PMF, and reduce the bias between these different receptor models. Likewise, the relatively high expense and effort involved in the molecular marker measurements provides motivation to determine if acceptable MM-PMF results can be derived from fewer receptor observations in order to save time and expense. To this end, there is a need to understand the sensitivities of MM-PMF stabilities to the receptor observation numbers to better design future studies that seek to use MM-PMF.

The present study focuses on the sensitivity of PMF source apportionment of $PM_{2.5}$ organic aerosols in relation to the number of receptor observations in East St. Louis, IL. More than three hundred daily 24-h observations, including $PM_{2.5}$ metals, ions, organic and element carbon (EC), and speciated organic matters, were obtained for this purpose at the St. Louis-Midwest Supersite during 2002–2003. Additionally, subsets of this large data set were used to investigate the sensitivities of PMF solutions to the number of receptor observations. This study has important implications for future PMF source apportionment studies and the strategies used for data collection.

2. Methods and procedures

2.1. Sampling and chemical analysis

Daily $PM_{2.5}$ samples were collected at the St. Louis-Midwest Supersite, and analyzed for 24-h average OC, EC, metals, trace elements and speciated organic compounds from May 1st, 2001 through April 30th, 2003 (Bae et al., 2004). OC and EC were determined using a thermal-optical ECOC analyzer (Sunset Laboratory Inc., Forest Grove, OR) (NIOSH, 1996). The Harvard/EPA Annular Denuder System (HEADS) was used to collect ambient particulate matter NO_3^-, SO_4^{2-}, and NH_4^+, as well as gaseous SO_2, HNO_3, HNO_2, and NH_3. Ionic compounds were analyzed by ion chromatography (Babich et al., 2000). Samples collected by $PM_{2.5}$ Harvard Impactors were analyzed for elements by X-Ray Fluorescence (XRF) Spectroscopy. Desert Research Institute (DRI) performed the ion and elemental analyses. Speciated organic matters, including *n*-alkanes, fatty acids, levoglucosan, polycyclic aromatic hydrocarbons (PAHs), hopanes and steranes, etc., were measured using solvent extraction and gas chromatography mass spectrometry (GCMS) methods. One set of the samples, collected on a 1-in-6 day sampling schedule (EPA 2001–2003 1-in-6 day schedule), along with all of the samples from April 2002 through May 2003 were analyzed by solvent extraction GCMS (Bae et al., in press; Schauer et al., 2002; Sheesley et al., 2004), and the remaining samples from May 2001 through April 2002 were analyzed using thermal desorption-GCMS (TD-GCMS) (Sheesley et al., 2007). The samples used in this study, 5-in-6 day samples from April 2002 through May 2003, were analyzed by a modified version of the method used by Bae et al. (in press). The 1-in-6 day samples analyzed by Bae et al. (in press) were extracted using a mixture of methanol and dichloromethane, while the samples used in this study were extracted using only dichloromethane. Although the biases of key molecular markers introduced by the above solvent difference or by method difference (GCMS and TD-GCMS) can statistically be ignored in source apportionment studies (Sheesley et al., 2007), to completely eliminate the possible statistical bias introduced by the measurement difference, only the 5-in-6 day samples from April 2002 through May 2003 were selected as the input data set.

All the quantified species were blank corrected. The uncertainties for molecular marker analysis was determined by

$$\sqrt{(20\% \times Conc.)^2+(STDEV_{blk})^2},$$

where *Conc.* is the measured concentration of the species and $STDEV_{blk}$ is the standard derivation of the measured concentrations of the field blanks including all sample analysis batches.

2.2. Preparation of PMF input data set

Daily average concentrations of the 48 selected species (listed in Table S1 in Supplemental materials) were measured as described above on a 5-in-6 day sampling schedule from April 2002 through May 2003. Candidate PMF species were screened based on the following considerations. First, only species that are known to have some source specificity were included. In this study, some *n*-alkanes and all of the fatty acids were excluded from the input data set because they are ubiquitous in almost all sources of atmospheric aerosols (Schauer et al., 1999a,b, 2001; Zhang et al., 2008b). Second, aluminum and silicon have been used in the past in molecular marker source apportionment models as tracers of resuspended soils (Sardar et al., 2005). In the current data set, a poor linear relationship between aluminum and silicon concentrations was present due to a low signal to noise (S/N) ratio of aluminum concentrations, and aluminum measurements were removed from the data set. Finally, all of the daily average observations were reviewed, and days with missing molecular markers, EC, OC or trace metals were removed for completeness. Additionally, extreme events of very high concentrations were removed as a recommended step for data processing for PMF model inputs (Paatero, 1999; Paatero and Hopke, 2003; USEPA, 2005), and as recommended by Jaeckels et al. (2007). Table S2 in the Supplemental materials lists all the days that were removed from the model input, including: 1) days with missing and invalid data and 2) 11 days with extremely high OC (5 observations with