

Assessing positive matrix factorization model fit: a new method to estimate uncertainty and bias in factor contributions at the measurement time scale

J. G. Hemann¹, G. L. Brinkman², S. J. Dutton², M. P. Hannigan², J. B. Milford², and S. L. Miller²

¹Department of Applied Mathematics, University of Colorado, Boulder, USA

²Department of Mechanical Engineering, University of Colorado, Boulder, USA

Received: 3 January 2008 – Published in Atmos. Chem. Phys. Discuss.: 14 February 2008

Revised: 4 December 2008 – Accepted: 5 December 2008 – Published: 22 January 2009

Abstract. A Positive Matrix Factorization receptor model for aerosol pollution source apportionment was fit to a synthetic dataset simulating one year of daily measurements of ambient PM_{2.5} concentrations, comprised of 39 chemical species from nine pollutant sources. A novel method was developed to estimate model fit uncertainty and bias at the daily time scale, as related to factor contributions. A circular block bootstrap is used to create replicate datasets, with the same receptor model then fit to the data. Neural networks are trained to classify factors based upon chemical profiles, as opposed to correlating contribution time series, and this classification is used to align factor orderings across the model results associated with the replicate datasets. Factor contribution uncertainty is assessed from the distribution of results associated with each factor. Comparing modeled factors with input factors used to create the synthetic data assesses bias. The results indicate that variability in factor contribution estimates does not necessarily encompass model error: contribution estimates can have small associated variability across results yet also be very biased. These findings are likely dependent on characteristics of the data.

1 Introduction

Air pollution comprised of particulate matter smaller than 2.5 μm in aerodynamic diameter (PM_{2.5}) has been associated with a significant increased risk of morbidity and mortality (Dockery et al., 1993; Pope et al., 2002; Peel et al., 2005). Existing regulations have focused on average and peak PM_{2.5} concentrations (μg m⁻³). To help policy makers design more

targeted and cost-effective approaches to protecting public health and welfare, an understanding of the association between PM_{2.5} sources and morbidity and/or mortality needs to be developed.

The Denver Aerosol Sources & Health study (DASH) has been undertaken to understand the sources of PM_{2.5} that are detrimental to human health. PM_{2.5} filter samples are collected daily from a centrally located site in Denver, CO. Speciated PM_{2.5} is quantified including sulfate, nitrate, bulk elemental and organic carbon, trace metals, and trace organic compounds. These speciated PM_{2.5} data are used as input to a receptor model, Positive Matrix Factorization (PMF), for pollution source apportionment. The PMF model fit yields characterizations of pollution sources, known as *factors*, with respect to their contributions to total measured PM_{2.5}, as well as their chemical profiles. Ultimately, an association will be explored between the individual factor contributions and short-term, adverse health effects, including daily mortality, daily hospitalizations for cardiovascular and respiratory conditions, and measures of poor asthma. For example, historical records of daily hospitalizations due to respiratory problems might be regressed against the daily concentrations of PM_{2.5} pollution from diesel fuel combustion (as estimated by PMF) over the same time span. Having measures of uncertainty associated with the contribution of diesel fuel combustion to PM_{2.5}, at the daily time scale, may lead to more reliable characterization of the role diesel fuel combustion has in daily health effects data.

PMF is a factor analytic method developed by Paatero and Tapper in 1994 (Paatero and Tapper, 1994) that has been widely used for pollution source apportionment modeling (Anderson et al., 2001; Kim and Hopke, 2007; Larsen and Baker, 2003; Lee et al., 1999; Polissar et al., 1998; Ramadan et al., 2000). The objective of this paper is to present a novel



Correspondence to: J. G. Hemann
(josh.hemann@colorado.edu)

method that has been developed to quantify uncertainty and bias in a PMF source apportionment model as it is applied to speciated PM_{2.5} data. Uncertainty in a PMF solution exists at a number of levels and is important to quantify, especially if the solutions will inform environmental and health policy decisions.

Uncertainty can stem from the data and from the PMF model itself. With respect to the data, uncertainty in the solution is imparted through measurement error as well as random sampling error. For the PMF model, there is generally “rotational ambiguity” in the solutions (i.e. solutions are not unique); further, solutions based upon the same data can vary depending upon how the model parameters are set. Past studies have considered these aspects, primarily by using the statistical method of the bootstrap to analyze model fit results. For example, Heidam (1987) considered the uncertainty in factor profiles due to receptor model uncertainty by varying the model parameters in models fit to bootstrapped datasets.

The Environmental Protection Agency’s Office of Research and Development distributes two software products, EPA PMF 1.1 (Eberly, 2005) and EPA Unmix 6.0 (Norris et al., 2007), which incorporate the bootstrap to analyze receptor model fit results. The software can be used to assess uncertainty in factor profile estimates and has been used by studies such as Chen et al. (2007) and Olson et al. (2007) to characterize sources of PM_{2.5}. Few studies, however, have addressed uncertainty in factor contribution estimates. Two examples are Nitta et al. (1994) and Lewis et al. (2003), though the estimates come from different source apportionment models and pertain to average contribution variability.

The method presented in this paper estimates, at the measurement time scale, bias and variability due to random sampling error in factor contribution estimates. Replicate datasets are created using a circular block bootstrap, and the subsequent application of two novel techniques make such estimation possible. First, neural networks are used for matching factors across PMF results on that data. Second the measurements resampled across the replicate datasets are tracked within the PMF solutions. This discussion describes the method in the context of application to a synthetic PM_{2.5} dataset, which was designed to simulate DASH data, fit by the PMF model. Using synthetic data allows assessment of model fit as well as a way to validate the method itself.

2 Methodology

Presented here is a method of assessing uncertainty in source apportionment model results using two different measures: bias and variability due to random sampling error. The method goes beyond computing these measures in terms of “average values” and gives estimates at the measurement time scale.

A synthetic time series of daily PM_{2.5} measurements is used in which the concentrations of chemical species are de-

rived from published source profiles and source contributions consistent with the Denver area. The solution from applying PMF can be compared with “known” profiles and contributions, allowing estimates of bias to be computed.

A circular block bootstrap generates additional data by re-sampling, with replacement, from the original synthetic measurement series. Each new dataset, or replicate, is again fit by the PMF model to apportion the PM_{2.5} mass to factors.

The first novel aspect pertains to how factors are sorted between solutions. For each solution the factors should correspond to the same real-world pollution sources. The factors need to be aligned such that “factor *k*” in each solution always refers to the same factor. To accomplish this factor alignment, or matching, the standard approach has been to use scalar metrics like linear correlation to match a factor from one solution to the “closest” factor in another solution. This is the approach taken by the EPA PMF 1.1 software, where it is specifically the time series of factor contributions that are matched between solutions. In contrast, the present work takes the novel approach of using Multilayer Feed Forward Neural Networks (NN), trained to perform pattern recognition, to align factors between PMF solutions. Further, using the intuitive notion that pollution sources are characterized best by the chemical species they emit, the matching is based on factors’ profiles rather than their contribution time series. The NN approach is a robust factor matching technique: it avoids the sensitivity to outliers that is problematic when using measures such as linear correlation and replaces it with a method that is capable of capturing linear as well as non-linear relationships.

The second novel aspect in the method presented here is the tracking of the measurement days resampled in each bootstrapped dataset. Through this bookkeeping it is possible to arrive at a collection of PMF results for each factor’s contribution on each day. Accordingly, descriptive statistics can be computed for each factor contribution on each day.

2.1 Positive Matrix Factorization

PM_{2.5} pollution is typically comprised of dozens of chemical species emitted from multiple sources. The concentration of each species may be treated as a random variable observed over time. The statistical technique of *factor analysis* can be used to explain the variability in these observations as linear combinations of some unknown subset of the sources, called factors. In traditional factor analysis approaches, including Principal Components Analysis, the variance-covariance matrix of the observations is used in an eigen-analysis to find the factors that explain most of the variability observed. The uncertainty in the observations, for all variables, is assumed to be independent and normally distributed. These assumptions are often not valid in the context of air pollution measurement data. In contrast, PMF – a receptor-based source apportionment model – offers an alternative technique that is based upon a least squares method, and measurement uncertainties

can be specific to each observation, correlated, and non-normal in distribution. Further, the factors resultant from PMF need not be orthogonal, which is an important quality when trying to associate modeled factors to real-world pollution sources that can be highly temporally correlated but are nonetheless important to characterize separately (e.g. diesel versus gasoline fuel combustion).

Given a matrix of observed PM_{2.5} concentrations, **X**, PMF attempts to solve

$$\mathbf{X} = \mathbf{G}\mathbf{F} + \mathbf{E} \quad (1)$$

by finding the matrices **G** and **F** that recover **X** most closely, with all elements of **G** and **F** strictly non-negative. **G** is the matrix of factor contributions (or “scores” in traditional factor analysis terminology), where G_{ik} is the concentration factor k contributed to the total PM_{2.5} observed in sample i . **F** is the matrix of factor profiles (or “loadings”), where F_{kj} is the fraction at which species j makes up factor k . Finally, **E** is the matrix of residuals defined by

$$E_{ij} = X_{ij} - \sum_{k=1}^p G_{ik}F_{kj} \quad (2)$$

G and **F** are found through an alternating least squares algorithm that minimizes the sum of the normalized, squared residuals, Q

$$Q = \sum_{i=1}^n \sum_{j=1}^m \left(\frac{E_{ij}}{S_{ij}} \right)^2 \quad (3)$$

where E_{ij} is weighted by S_{ij} , the uncertainty associated with the measurement of the j th pollutant species in the i sample. The ability to weight specific observations with specific uncertainties allows PMF to handle data that include heterogeneous measurement uncertainty, outliers, values below measurement detection limits, and missing values. As such, PMF can often yield better results than traditional factor analysis methods (Huang et al., 1999).

An algorithm for implementing PMF is available as a commercial software library, PMF2 (Paatero, 1997). The work presented here uses PMF2 version 4.2, and specifically, the pmf2wopt executable file (Paatero, 2007). PMF2 has numerous optimization parameters that can be set by the user, and methods of choosing these values have been published elsewhere (Paatero, 2000; Paatero et al., 2002, 2005). Since the focus of this paper is on a method of assessing uncertainty and bias in PMF solutions, the discussion of fine-tuning the numerous algorithm parameters is kept to a minimum. Two PMF2 parameters are especially important to the PMF model fit and deserve mention. First, the number of factors in the model, p , must be set by the user. In the present work, eight and nine factor solutions are considered, with the primary focus on the results for the nine factor solutions. The other important parameter is FPEAK, which controls the rotational freedom of the possible solutions. It is advised that FPEAK

values range between -1 and 1 , with positive values causing extremes in the **F** matrix (values near 0 or 1) and negative values causing extremes in the **G** matrix. In the present work, FPEAK is zero for all PMF2 solutions, which corresponds to the default setting.

2.2 Synthetic data

Given that the results of pollution source apportionment models may ultimately be used as critical components of environmental policy and regulatory decisions, it is especially important to assess their quality. One approach for evaluating receptor models is the use of synthetic data, which is defined as simulated PM_{2.5} measurements rather than actual observations (Willis, 2000). Predefined sources are used, along with their respective contributions and profiles, to create the **G** and **F** matrices in Eq. (2). With **G** and **F** defined **X** can be calculated directly and given as input (along with uncertainty estimates) to the PMF2 software, where the resultant **G** and **F** matrices can then be compared with the actual values to assess model fit.

The method of creating synthetic datasets followed in this paper is described in detail in Brinkman et al. (2006) and Vedal et al. (2007). Briefly, nine pollutant sources were used (Table 1), which contributed concentrations of 39 chemical species (Table 3), over 365 synthetic sampling days. The synthetic measurements were assumed to come from a single receptor site. Table 1 also lists the references used to generate the annual contributions, chemical profile, temporal patterns and variability for each source. Table 2 shows the lag zero cross correlations between the source contributions. With respect to PMF modeling, the relatively high cross-correlations between some of the input source contribution time series has the implication that some of these sources may be harder to cleanly separate from others.

Distinct time series for the contributions from each source were generated by starting with average contribution estimates from preliminary DASH studies and the Northern Front Range Air Quality Study (Watson et al., 1998), then adding day-to-day variations reflecting both random variability and hypothesized weekly or seasonal patterns, as appropriate. Daily totals for the nine source contributions were normalized to match actual daily PM_{2.5} levels observed in Denver in 2003. It should be noted that the presence of additional sources, such as secondary organic aerosols, could complicate application of PMF to observed data. The matrix of data uncertainties, **S** from Eq. (3), is computed as follows. Measurement detection limits, detection limit uncertainty, and measurement uncertainty associated with typical analytical techniques used to speciate PM_{2.5} filter samples (Ion Chromatography, Thermal Optical Transmission, and Gas Chromatography/Mass Spectrometry), were incorporated into the PMF input via

$$S_{ij} = \sqrt{(\alpha_j X_{ij})^2 + (\beta_j D_j)^2} \quad (4)$$

Table 1. Synthetic PM_{2.5} sources.

Source	References
Secondary Ammonium Sulfate	Lough (2004)
Secondary Ammonium Nitrate	Lough (2004)
Gasoline Vehicles	Watson et al. (1998); Chinkin et al. (2003); Cadle et al. (1999); Hildeman et al. (1991); Rogge et al. (1993a)
Diesel Vehicles	Watson et al. (1998); Chinkin et al. (2003); Hildeman et al. (1991); Rogge et al. (1993a); Schauer (1998)
Paved Road Dust	Watson et al. (1998); Chinkin et al. (2003); Hildeman et al. (1991); Rogge et al. (1993b)
Wood Combustion	Watson et al. (1998); Fine et al. (2004)
Meat Cooking	Watson et al. (1998); Schauer et al. (1999)
Natural Gas Combustion	Hildeman et al. (1991); Hannigan (1997); Rogge et al. (1993d)
Vegetative Detritus	Hildeman et al. (1991); Hannigan (1997); Rogge et al. (1993c)

Table 2. Source contribution cross-correlations, Lag=0.

	Amm Sulfate	Amm Nitrate	Gasoline	Diesel	Road Dust	Wood	Meat	Natural Gas	Veg
Ammonium Sulfate	1	0.55	0.89	0.57	0.73	0.24	0.81	0.68	0.49
Ammonium Nitrate	.	1	0.39	0.26	0.32	0.82	0.35	0.8	-0.15
Gasoline Vehicles	.	.	1	0.62	0.74	0.13	0.81	0.64	0.54
Diesel Vehicles	.	.	.	1	0.61	0.08	0.28	0.33	0.31
Paved Road Dust	1	0.13	0.57	0.5	0.37
Wood Combustion	1	0.14	0.77	-0.34
Meat Cooking	1	0.64	0.53
Natural Gas	1	0.08
Vegetative Detritus	1

where for species j , α_j is the measurement uncertainty, β_j is the detection limit uncertainty, and D_j is the detection limit. Table 3 contains the α , β and D associated with each species. The S_{ij} uncertainties were incorporated into the final data matrix \mathbf{X}' with the following formula

$$X'_{ij} = X_{ij} + S_{ij}Z_{ij} \quad (5)$$

where Z_{ij} is a random number drawn from a standard normal distribution. If X'_{ij} was less than the detection limit associated with measuring species j , then a value of one-half the detection limit was substituted in the final data matrix.

2.3 The bootstrap

The bootstrap is a computationally intensive method for estimating the distribution of a statistic, the statistic itself being an estimator of some parameter of interest (Efron, 1979). The essence of the method is to create replicate data by re-sampling, with replacement, from the original observations of a random variable. For each replicate dataset the statistic of interest is computed, and the distribution of these values serves as an estimate for the random sampling distribution

of the statistic. The properties of this distribution are then used to make inferences about the parameter of interest. In the present context, each pollutant species' time series represents realizations of a random variable. The \mathbf{F} and \mathbf{G} matrices resulting from PMF's fitting of these data are functions of these random variables, thus, each element of those matrices may be considered a statistic. Previous studies using PMF have focused on analyzing the \mathbf{F} matrix, the matrix of factor profiles. This discussion takes a different tack, with the statistic of interest being each element of the \mathbf{G} matrix, the matrix of factor contributions over time.

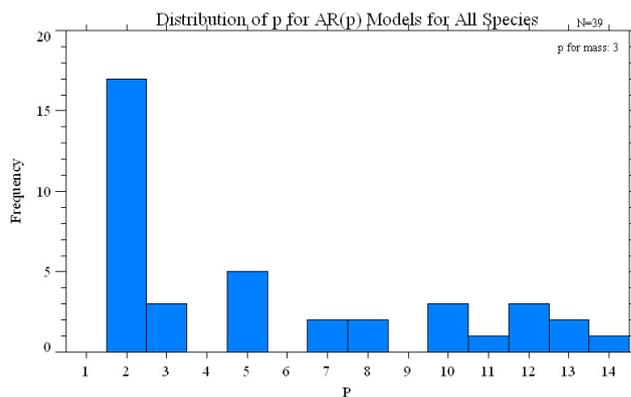
2.3.1 Dependent data considerations

Much of bootstrap theory is based upon the assumption that the data are comprised of observations of independent and identically distributed (iid) random variables. Time series data, however, are typically serially correlated. Singh (1981) showed that the bootstrap can be inconsistent in estimating the distribution of statistics based upon dependent data. Since then, numerous modifications of the original iid bootstrap have been formulated to better handle dependent data

Table 3. Synthetic PM_{2.5} species, measurement detection limits (D), measurement errors (α), and detection limit uncertainties (β).

Species #	Species Name	D	α	β
		(ng/m ³)	(%)	(%)
1	Elemental Carbon	13	10	197
2	Organic carbon	3.6	14	100
3	Nitrate	0.094	5	296
4	Sulfate	0.20	3	614
5	Ammonium	0.22	6	1057
6	n-Tricosane	0.97	8	125
7	n-Tetracosane	1.4	9	186
8	n-Pentacosane	1.1	8	209
9	n-Hexacosane	1.1	7	214
10	n-Heptacosane	1.1	6	245
11	n-Octacosane	1.2	4	207
12	n-Nonacosane	0.96	4	234
13	n-Triacontane	0.92	6	185
14	n-Hentriacontane	0.92	8	145
15	n-Dotriacontane	0.23	9	108
16	n-Tritriacontane	0.16	9	96
17	n-Tetracontane	0.093	8	93
18	Oleic acid	6.6	14	39
19	n-Pentadecanoic acid	1.4	15	235
20	n-Hexadecanoic acid	48	16	116
21	n-Octadecanoic acid	26	14	137
22	Acetovanillone	0.064	14	212
23	Coniferyl aldehyde	1.9	14	86
24	Syringaldehyde	0.62	12	139
25	Acetosyringone	0.97	11	468
26	Retene	0.079	12	316
27	Alkyl Cyclohexanes	0.085	8	123
28	Benzo[k]fluoranthene	0.0068	10	649
29	Benzo[b]fluoranthene	0.0068	10	649
30	Benzo[e]pyrene	0.0056	15	272
31	Indeno[1,2,3-cd]pyrene	0.017	13	1052
32	Indeno[1,2,3-cd]fluoranthene	0.021	13	1052
33	Benzo[ghi]perylene	0.023	10	169
34	Coronene	0.021	13	23
35	Cholestanes	0.21	8	65
36	Hopane	0.045	17	202
37	Norhopane	0.050	19	113
38	Homohopanes	0.013	18	342
39	Oxygen	41	3	106

(Carlstein, 1986; Kunsch, 1989; Liu and Singh, 1992). One approach often used for time series data is to resample blocks of successive observations. If the blocks are of sufficient length, l , and the series is only weakly dependent, then the observations within each block may be considered independent of the observations within the other blocks. Further, if the series is stationary then all blocks will share the same l -dimensional joint distribution. These two conditions allow the blocks themselves to be treated as independent and identically distributed observations to which the iid bootstrap can be applied. This approach is currently used by the EPA PMF 1.1 software tool, which uses a 3-day, Moving Block Bootstrap (MBB).

**Fig. 1.** Distribution of the lag dependence parameter, p , for the 39 pollutant species.

In the EPA's bootstrap implementation, as well as this study, measurement days are resampled. In the present case, realizations of a composite random variable comprised of 39 pollutant species are resampled, with replacement, from the original synthetic data. To investigate an appropriate bootstrap block size the serial correlation of each pollutant species in the synthetic dataset, as well as total PM_{2.5} mass, was examined. An Auto Regressive (AR) time series model was fit to each species' series and the optimal lag parameter, p , was found, with p constrained between 1 and 14. Figure 1 shows the distribution of these lag values.

While concentrations for most species were serially correlated with only the previous two day's concentrations, some species had longer lag-dependence. The aggregate mass of all species had lag-three dependence. It has been shown that the consistency of approximations yielded by the block bootstrap is sensitive to block size, with optimal block size critically dependent upon the size of the data as well as the statistic for which the distribution is being estimated (Hall et al., 1995; Lahiri, 2001). Given these findings, and that practical methods for choosing block size are based on examining the lag-dependence in the data (Politis and White, 2004), it seems difficult to choose a single block size for resampling measurement days of speciated PM data.

An additional complication in using the block bootstrap with such data is that while all block bootstrap schemes are designed to handle serial correlation they also assume the data are from a stationary stochastic process. To the authors' knowledge, there are no published results in which a block bootstrap was used on speciated PM time series data that had tested for, or transformed to, stationarity prior to resampling. This assumption is ignored in the present work, but it deserves consideration in future applications of the block bootstrap.

The present work uses 1000 replicate datasets, each generated by the circular block bootstrap (CBB) of Politis and Romano (1992) with a block size, b , of four. *Circular* refers to "wrapping" the data end-to-beginning such that for a length

N series, $X_{N+k}=X_k$. A block length of four was chosen since the lag-dependence for mass was three days, and the majority of species had lag-dependence of five or less days. As an example, to create a bootstrap replicate data matrix \mathbf{X}^* , let \mathbf{Y} be a copy of the original matrix \mathbf{X} , and append the first $b-1$ rows of \mathbf{X} to the bottom of \mathbf{Y} . Divide the rows of \mathbf{Y} into N consecutive size b blocks. If X_i and Y_i denote the i th row of \mathbf{X} and \mathbf{Y} , respectively, then the blocks would be

block 1 : $Y_1, Y_2, Y_3, Y_4=X_1, X_2, X_3, X_4$

block 2 : $Y_2, Y_3, Y_4, Y_5=X_2, X_3, X_4, X_5$

...

block N : $Y_N, Y_{N+1}, Y_{N+2}, Y_{N+3}=X_N, X_1, X_2, X_3$

Construct \mathbf{X}^* by selecting N/b blocks randomly, with replacement, from \mathbf{Y} . Construct an associated replicate matrix of uncertainties, \mathbf{S}^* , by selecting these same blocks from \mathbf{S} . Note that if N/b is not an integer, round up and truncate as needed. For example, in the present case with $N=365$ and $b=4$, 92 blocks would be chosen to create \mathbf{X}^* . The last three elements of the 92nd block would not be used, yielding a total of 365 rows (measurement days) resampled from \mathbf{X} . Importantly, under CBB resampling each row of \mathbf{X} has the same probability of being resampled, which is not the case under MBB resampling (Lahiri, 2003).

2.3.2 Record keeping

In the following results an important component in implementing the bootstrap is the tracking of days resampled in each replicate dataset. Since the essence of bootstrapping is resampling with replacement it is possible that for any given replicate dataset some synthetic measurement days are included multiple times and other days are not included at all. By keeping track of which days are resampled in which replicate dataset it is possible to assess factor contribution uncertainty and bias at the daily time scale.

2.4 Factor matching

The use of the bootstrap yields a collection of factor contribution matrices, G^k , $k=1, \dots, B$, where B is the number of bootstrap replicate datasets. The collection of matrices may be considered as a single, three-dimensional matrix \mathbf{G}' , with elements $G'_{ij}{}^k$, $i=0, \dots, N-1$; $j=0, \dots, P-1$; $k=0, \dots, B$, where N is the number of samples and P is the number of factors (note that the \mathbf{G} matrix associated with the original data and “base case” solution is also included in \mathbf{G}'). While the nature of the factors that the PMF2 algorithm finds should be stable across the “bootstrap” solutions, the ordering of the factors within those solutions may be different. Before computing statistics on the elements of \mathbf{G}' , the dimension indexed by j must be sorted, such that across the $B+1$ matrices factor j is always associated with the same real-world pollution source. The typical approach to matching and sorting factors has relied on comparisons between the factor contribution time series,

using linear correlation to match a given factor from a bootstrap solution (or a factor from another analysis method) to the “closest” factor in a base case solution.

There are several concerns with this approach. First, “closeness” is measured with a scalar metric that is highly sensitive to outliers. Second, the bootstrap replicate data sets will not preserve the temporal patterns seen in the original data when viewed over the course of the entire sampling period (although there are block bootstrap methods that seek to address this). Correlation ceases to be a useful measure once there is no temporal consistency between the contribution time series being compared. Third, there are no clear rules for what constitutes sufficient correlation, especially in cases where two factors in one bootstrap solution are highly (or even poorly but equally) correlated with the same factor in the base case solution. The use of this metric invites “data dredging”, where the practitioner must make ad hoc choices to separate and match factors.

The present discussion employs an approach that the authors believe to be novel and robust when applied to aerosol pollution data: neural networks are used to match factors between bootstrap solutions and the base case solution based upon their profiles. Matching on profiles addresses the second issue noted above, while the use of neural networks rather than correlation address the first and third issues. The need to classify a measured spectrum (profile) with a known reference spectrum is a problem found in multiple scientific settings, most notably in the analysis of stellar spectra and data from hyperspectral remote sensing. Findings in these fields may be useful in the modeling of aerosol pollution data and are considered briefly. Work by van der Meer (2006) found that a spectral similarity measure based on correlation was more sensitive to noisy data than other traditional measures based on Euclidean distance or spectral angle. Further, Shafri et al. (2007) reported that neural networks were accurate at classifying spectra from remote sensing of tropical forests, especially when compared to measures based on spectral angle. Tong and Cheng (1999) found that using neural networks was superior to using maximum correlation when classifying gas chromatography mass spectrometry data. Based on these findings, as well as the findings presented herein, the authors are confident that using neural networks to match factor profiles allows the bootstrap technique to be better leveraged. The factor matching process can be easily automated, adapted to complex patterns and new, possibly noisy, data, and can avoid subjective “closeness” thresholds required when using less robust measures like correlation.

2.5 Neural networks

Artificial Neural Networks are statistical modeling methods capable of characterizing highly non-linear functions, doing so by approximating the behavior of the brain. The term “artificial” is used to distinguish this numerical approximation of biological, adaptive, cognition from those biological systems. In general, this is understood in statistical modeling, and Artificial Neural Networks are simply referred to as Neural Networks (NN). Excellent introductions to the subject can be found in Haykin (1998) and Munakata (1998).

The specific type of NN used in this study is called a *Multi-layer Feed Forward Network*. This type of network relies on *supervised learning*, in which the network is given inputs and learns how to transform it into desired outputs. The learning is encoded in numerical weights defining the strength of connection between elements in the network. Weights are found through quasi-Newton optimization incorporated with the backpropagation method, where “backpropagation” refers to the ground-breaking algorithm developed in the 1970s and 1980s (Rumelhart et al., 1986; Werbos, 1974), allowing neural networks to classify linearly inseparable patterns. A trained network, characterized by its structure and its weights, can then be given novel input and transform it to the correct output. In the present work that transformation is classification: given a new factor profile, the trained neural networks will classify it as a known type, or possibly classify it as unknown.

2.5.1 Neural network configuration

In the present work, NN software from Visual Numerics’ IMSL[®] C Numerical Library, version 6.0, is used. The structure of the network is three fully connected layers, with 39 nodes in the input layer, five nodes in the hidden layer, and two nodes in the output layer. All nodes in the hidden and output layers use a logistic activation function. The values of the two output nodes range between 0 and 1. An output of [1,0] indicates a perfect match between an input factor profile and the profile that particular network was trained to classify as a “Yes”. Likewise, an output of [0,1] indicates a perfect non-match between an input factor profile and the learned profile. A “Yes” match is only possible if the first output node has a value of at least 0.95, with the second node having a value no larger than 0.05.

The performance of the network depends heavily upon the data used for training. It is well established that neural networks can be unstable when data used for training varies greatly in scale; therefore, transformation and normalization of data are typical preprocessing steps. In this study, factor profiles are normalized before being learned by the networks. In “raw” form, the rows of the \mathbf{F} matrix correspond to factor profiles and each row sums to 1. Viewing factor profiles this way can sometimes result in factors that are difficult to distinguish, since some species will be present in large amounts

in many factors (e.g. organic carbon and oxygen). To make plots of factor profiles more visibly distinguishable the following normalization is done,

$$F'_{kj} = \frac{F_{kj}}{\sum_{k=1}^P F_{kj}} \quad (6)$$

where F'_{kj} is the relative weighting species j has in factor k 's profile when considering all other factors. When viewing factor profiles under this normalization, species common to many factors are damped and marker species become more pronounced, as compared to viewing the raw profiles.

2.5.2 Training data

Five datasets were used to train the neural networks. One dataset was the original synthetic data, with the remaining four being bootstrap replicates. PMF was fit to each dataset, with the solution associated with the original dataset being the base case. The base case factor profiles were normalized (as described in Sect. 2.5.1), plotted, and visually compared to the normalized factor profile plots associated with the bootstrap solutions. For each bootstrap solution the factors were reordered to match the base case ordering. Figure 2 shows the end result of this process, with the five factor profile plots used to classify each factor for the neural network training, as well as the actual profile used to create the original synthetic dataset.

2.6 Method steps

Having discussed the major components of the method for analyzing factor contribution uncertainty and bias, it is helpful to summarize their relationship in the following steps:

Step 1: Using the synthetic PM_{2.5} data and measurement uncertainties, compute a base case PMF model fit that has P factors. In the present work, $P=9$.

Step 2: Create T bootstrap replicate data matrices, with corresponding uncertainty matrices, and fit each set with PMF. These results, in addition to the base case result, will serve as the neural network training data. In the present work, $T=4$.

Step 3: For each training replicate dataset, visually compare the normalized bootstrap factor profiles versus the normalized base case profiles, and define the factor matching between the results. Reorder the factors to be consistent with the base case factor ordering.

Step 4: For each factor, train a neural network to learn its normalized profile, as well as what is not its profile (thus, there will be P networks). For each factor there will be $T+1$ profiles to be learned as “Yes” patterns. The remaining $P-1$ profiles associated with the base case results are learned as “No” patterns.

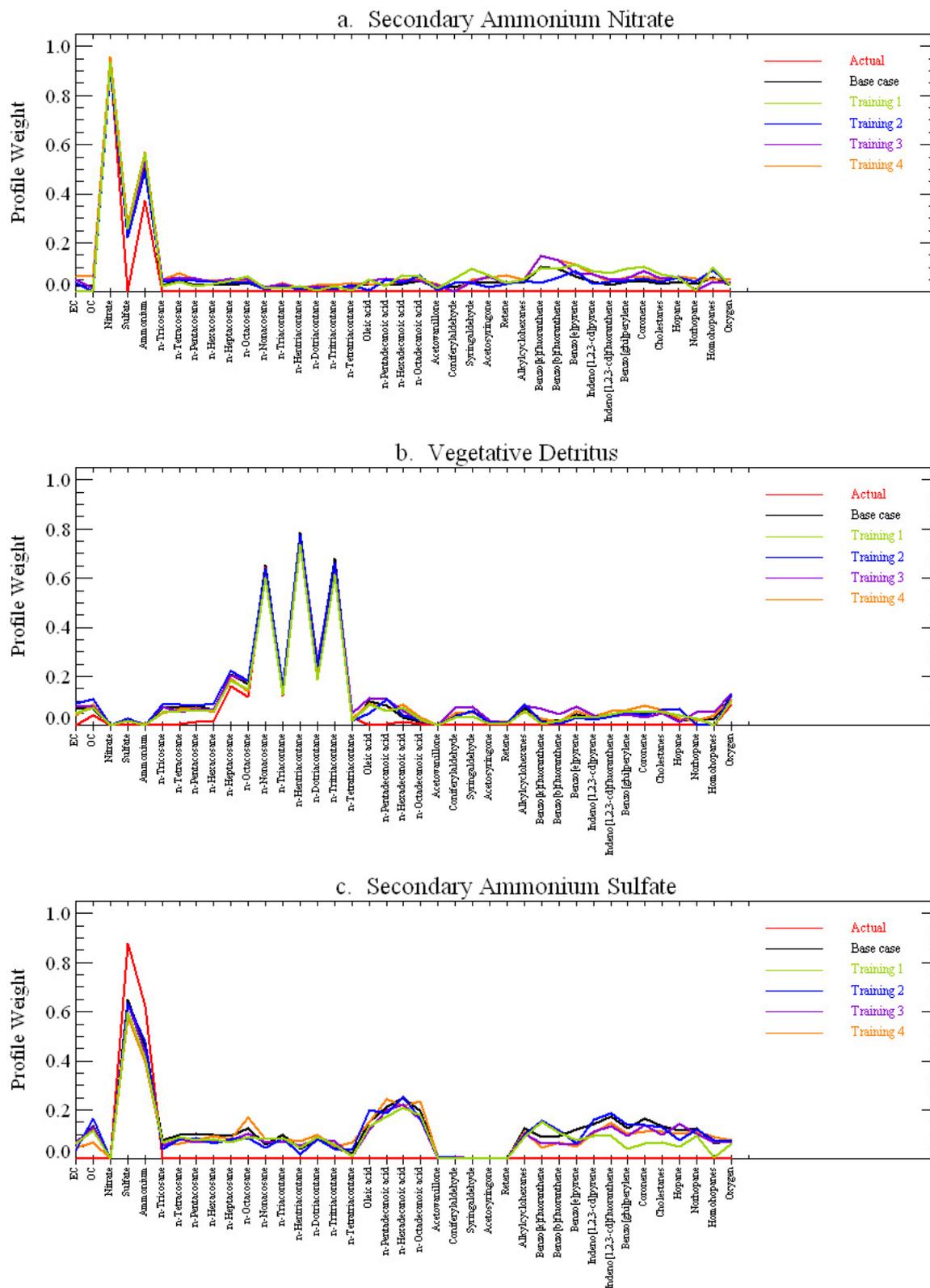


Fig. 2. Plots of five normalized profiles for each factor learned by the neural networks. The thicker, black line represents the profile associated with the “base case” solution, while the thicker red line indicates the actual profile used to create the synthetic data. The remaining four colors correspond to factor profiles for “bootstrap” solutions based on resampled data. The factor ordering is now relative to the ordering in the base case solution.

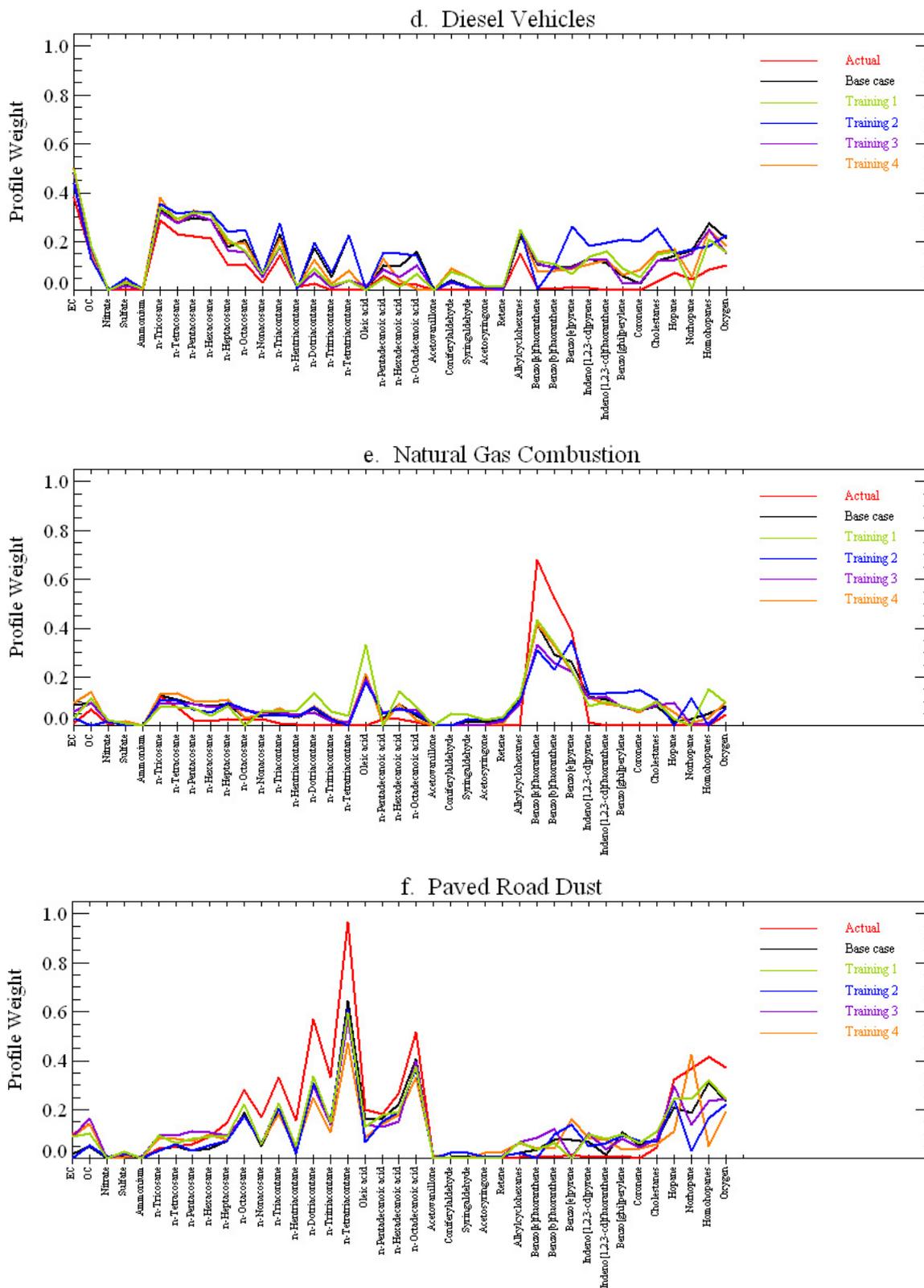


Fig. 2. Continued.

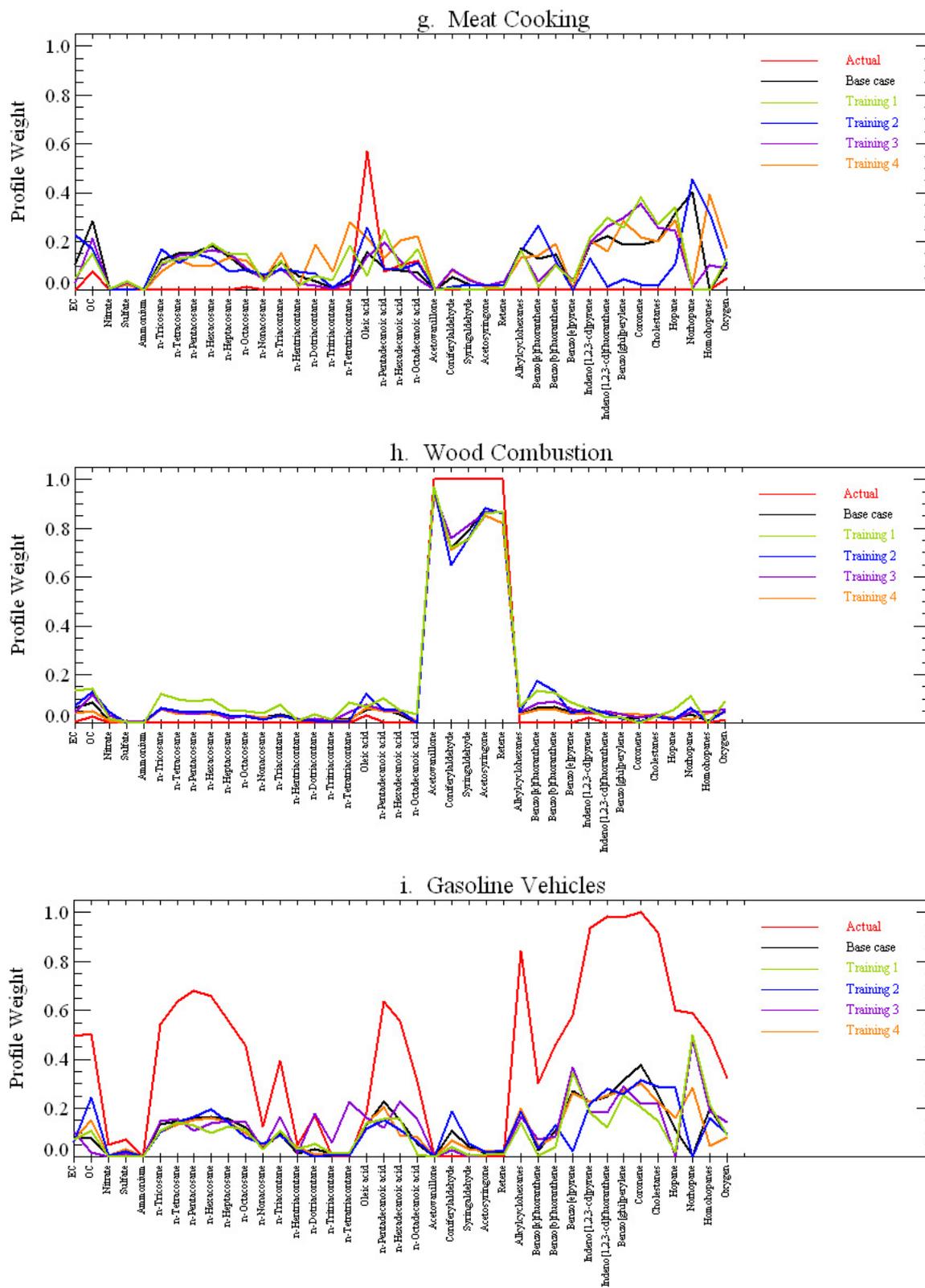


Fig. 2. Continued.

Step 5: Create B new bootstrap replicate datasets and fit each one with PMF. These results will be used to assess PMF model fit uncertainty. In the present work, $B=1000$

Step 6: For each bootstrap solution, allow each of the P neural networks to examine each of the P normalized factor profiles. Each network should identify a unique factor profile as a “Yes”, with all others classifying the profile as “No”. Reorder the factors in the bootstrap solution accordingly.

Step 7: Parse the factor contribution data by day-factor combination. For example, consider examining the bias and variability in the PMF solutions for the 3rd factor on day 126. All $B+1$ datasets would be searched for where the original day 126 was resampled. This collection of indices would be used to index into the 3rd column of the corresponding solutions’ \mathbf{G} matrix to get factor 3’s contribution on day 126. The distribution of values is then compared with the actual contribution used to create the original synthetic dataset.

Note that Step 3 is what establishes the supervisor for the *supervised learning* algorithm used to train the neural network. The role of the neural network is to learn the classification defined by an expert human observer, such that when new factor profiles are analyzed, they are classified as would the expert. In Step 6, it is possible that a bootstrap factor can be matched with more than one base case factor, or, perhaps, it cannot be matched to any base case factor. In either case, that particular solution is excluded from the collection of other solutions. In this way, after the last replicate dataset has been fit by PMF, the collection of solutions correspond to the case where bootstrap factors were matched uniquely to base case factors.

3 Results

Eight and nine factor PMF models fit to the original synthetic data were comparable, in terms of sums of the normalized, squared residuals, Q , the residuals associated with specific species, and the physical interpretability of the factors. Seven and ten factor solutions were judged less optimal with respect to these same measures. Accordingly, descriptive statistics are presented for both eight and nine factor simulations in Table 4. (The *kurtosis* statistic relates to the peakedness of the distribution; a value near 0 is associated with *mesokurtic* distributions, of which the Normal (μ, σ^2) distribution is the most common example.) To facilitate comparison of model fitting results with the contributions of the nine sources used to create the synthetic data, all other results pertain to the simulation using a nine factor solution. Figure 3 presents plots of factor contribution time series for all nine factors. Each plot shows the base case series, the actual series, and two bands defined by empirical quantiles of the simulation results: the interquartile range and the 5th–95th percentile range. The plots show the data and quantiles smoothed by a 5-day boxcar average, in the hopes of focusing attention on the gross features of the series and not the daily fluctuations

Table 4. Simulation statistics for 8 and 9 factor solutions.

	8 Factors	9 Factors
Number of bootstrap replicate datasets:	1000	1000
Number of datasets for which PMF2 failed to converge to a solution:	2	1
Number of datasets for which factors could not be uniquely matched:	45	263
<i>Q</i> -value Statistics		
Sample Size:	953	736
Mean:	3823.43	3299.16
Standard Deviation:	101.14	73.11
Skewness:	0.10	0.12
Kurtosis:	0.28	0.02

in contribution. Figure 4 is a histogram showing the distribution of contributions associated with a specific factor on a specific day, as an example of how the method allows assessment of contribution uncertainty at the daily time scale.

4 Discussion

4.1 Factor contribution plots

The results of applying the method to the synthetic PM_{2.5} data demonstrate several types of PMF solutions. The first is exemplified by the contribution time series plots for factors 1 and 3 (Fig. 3a and c). Here, PMF’s solutions, over hundreds of resampled datasets, show low variability and moderate bias when compared to the actual contribution time series. Factors 2, 4, 6, and 8 represent solutions in which the temporal pattern matches closely with the actual respective contribution time series, but the bias is large. Finally, factors 5, 7, and 9 have solutions that match poorly with respect to bias, variability, and temporal pattern, against the known contributions. It is worth recalling from Table 2 that these three factors had moderately strong correlations with each other in the contribution time series used to create the synthetic data.

The preceding results should not be generalized with respect to how well PMF models real-world pollution sources, as the results are based on synthetic data. However, if the synthetic data is assumed to be a close approximation of data likely to be actually observed, then application of the method to synthetic data representing a specific situation could help identify sources for which PMF contribution estimates should be carefully scrutinized.

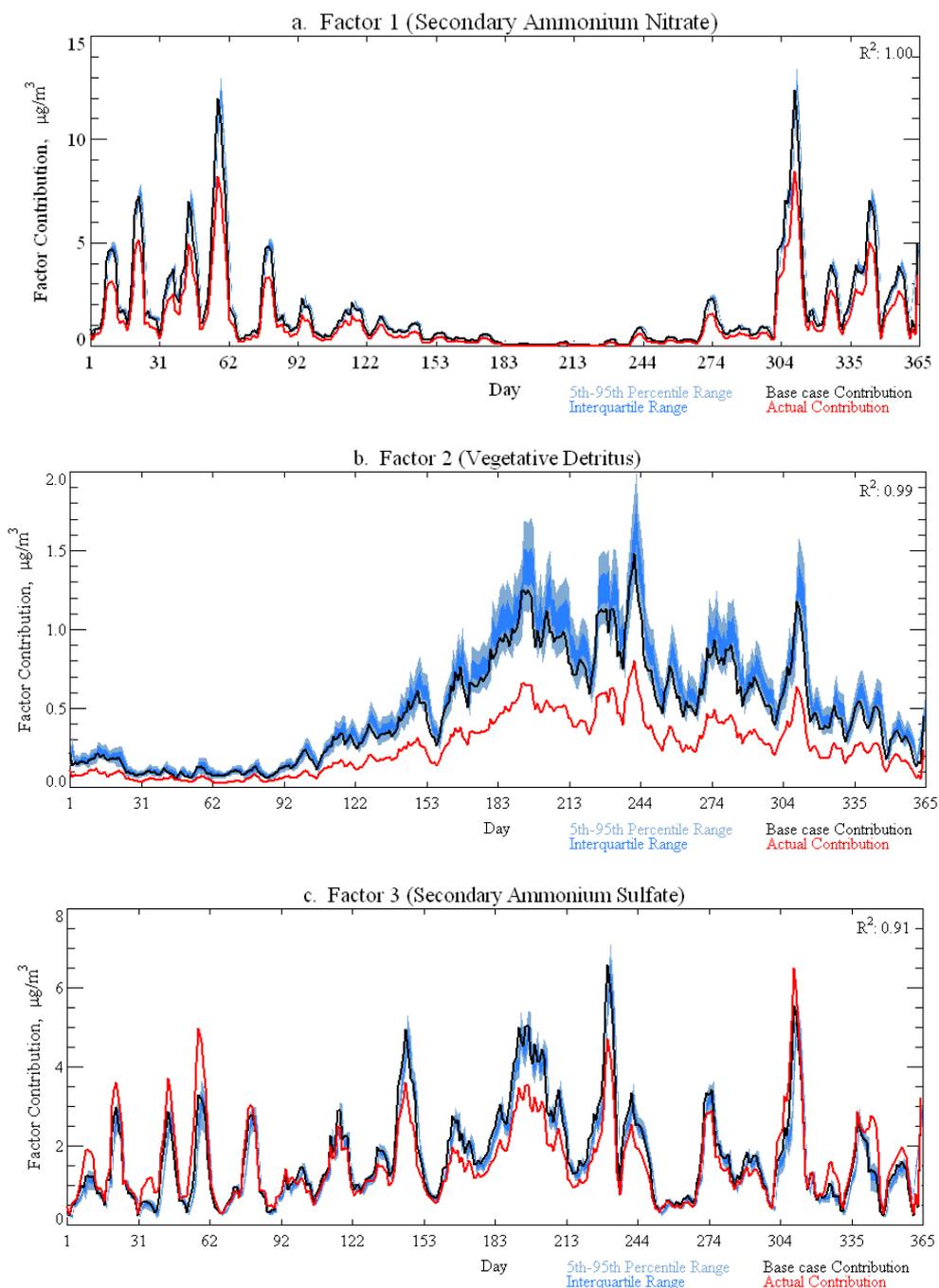


Fig. 3. Comparison of PMF bootstrap solutions for factor contributions versus actual factor contributions. Each plot corresponds to a different factor, showing the actual contribution time series, the time series corresponding to the base case PMF solution, and two bands based on the empirical quantiles of the bootstrap solutions. The listed coefficient of correlation is with respect to the base case and actual contribution time series. The factor ordering is relative to the base case solution.

4.2 Uncertainty, variability, and bias

Application of the method to synthetic, daily, measurements of $\text{PM}_{2.5}$ yields estimates of variability and bias in daily factor contributions, which can be used in an uncertainty analysis of the PMF model fit. However, the uncertainty in re-

sults brought out by fitting PMF to resampled data is likely different compared to the uncertainty in the results due to model assumptions. For example, how would the solutions change if seven or eight factors were instead considered; if certain pollutant species were added or removed; if different assumptions were made about measurement errors; if a

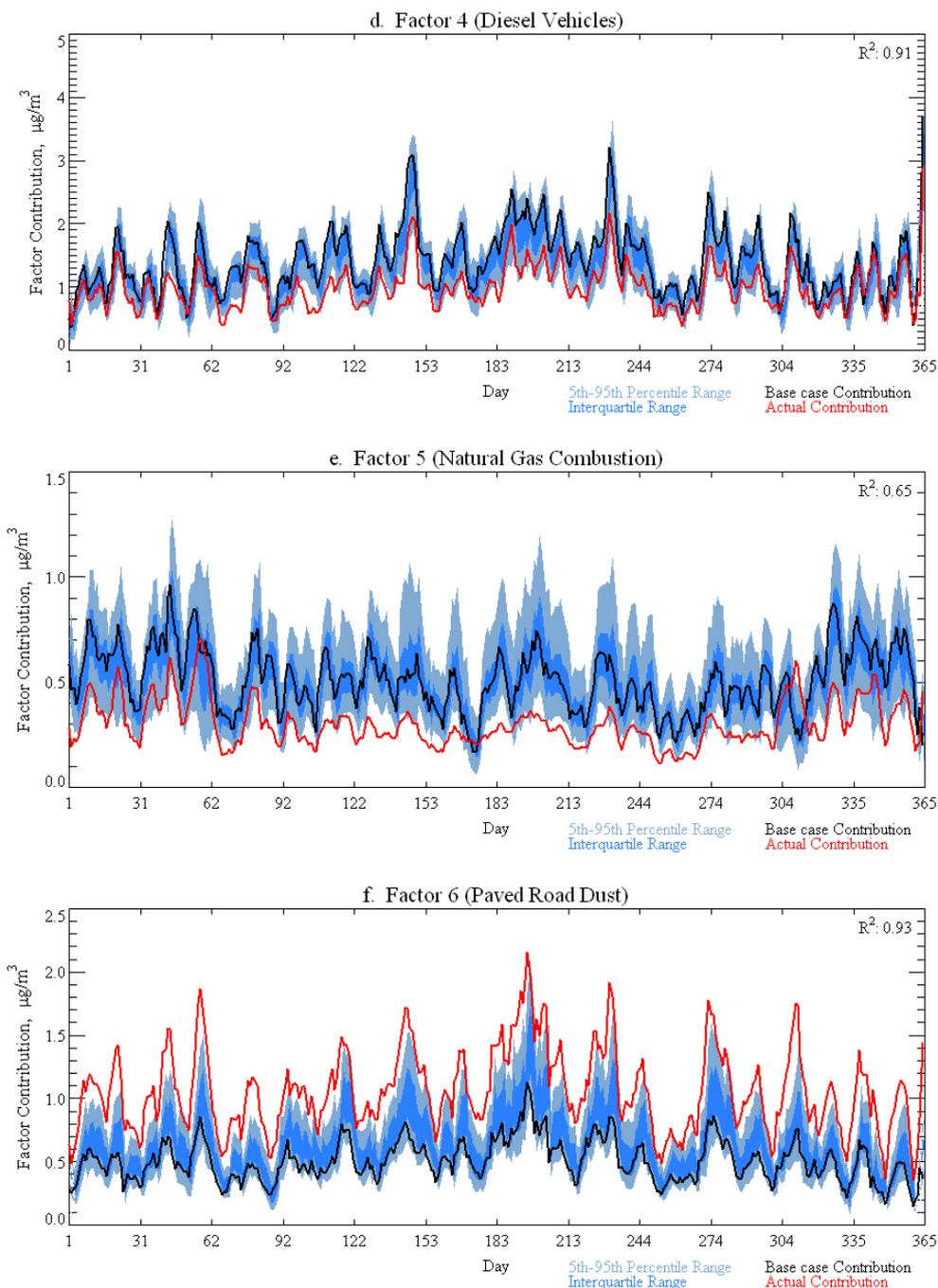


Fig. 3. Continued.

different source apportionment model was used altogether? As Chatfield (1995) notes, “*It is indeed strange that we often admit model uncertainty by searching for a best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true.*” In the present work, as much as possible, model assumptions have been avoided: input data was not filtered after seeing preliminary output, and PMF2 parameters were

set to avoid assumptions about the distribution or “quality” of the data. Still, the use of PMF as the receptor model, the chemical species included in the analysis, and the number of factors to be characterized, were all choices and are clearly subjective. The present work seeks to offer a method for assessing uncertainty in model fit when it is assumed that the model is valid, and this distinction should be kept in mind.

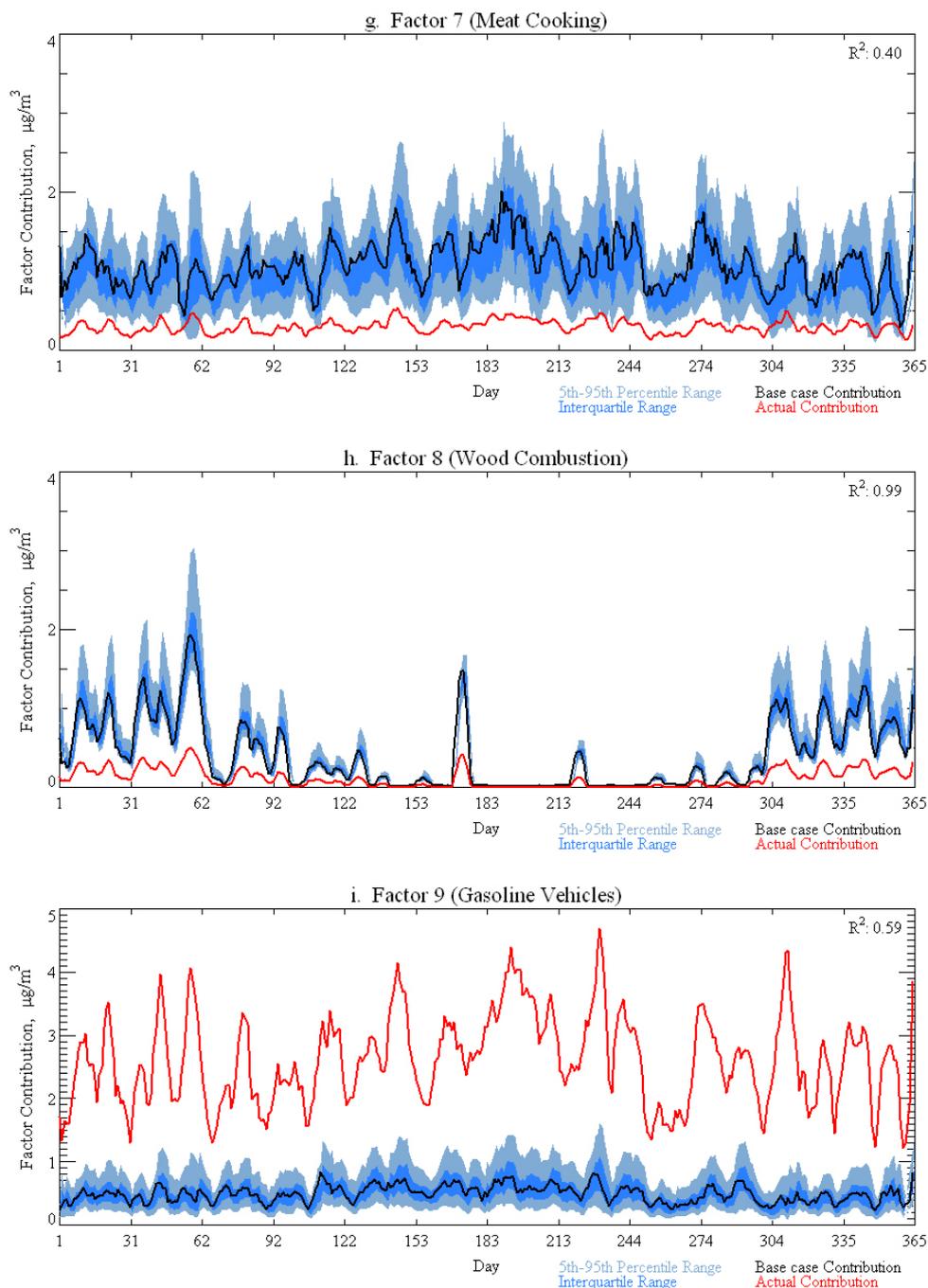


Fig. 3. Continued.

It is clear from the factor contribution plots in Fig. 3 that the PMF estimates for some factors are badly biased from the known contributions. Such bias is assessable because the PMF results are based upon synthetic data in which the true factor contributions and profiles are known. The results of using the presented method indicate that PMF is able to fit some synthetic pollution sources' contributions reasonably well, while other synthetic sources have approximations that

have large variability, bias, and generally are in error. Does PMF perform similarly for observed data, and if so, what characteristics of the data might result in some sources fit well and others not? Further investigation is needed and the authors are hopeful that practitioners will use methods, such as the one presented here, to further assess their pollution source apportionment results.

4.3 Eight versus nine factor solutions

In simulations using nine factor solutions, the rate at which the neural network factor matching method failed to uniquely match bootstrapped factors to base case factors was generally close to 25%. Typically this was just one bootstrap factor matching with two base case factors for a given bootstrap solution, with the remaining bootstrap factors having a unique match. Specifying only eight factors allowed PMF to collapse two correlated factors – meat cooking and natural gas – leading to generally more distinguishable results, thus unique factor matching failed at only a 5% rate. The nine factor solutions are presented in Figs. 2 and 3 to allow easy comparison with the actual factor information. It is important to note that if using the traditional factor matching method, based on the maximum linear correlation between the contribution time series, the nine factor bootstrap solutions could be difficult to sort. Consider Fig. 3d and f for the contribution time series plots for factors 4 (diesel fuel combustion) and 6 (paved road dust), respectively. If linear correlation alone was used as the metric for matching, it is easy to imagine how often factor 4 might be labeled 6 sometimes and vice versa. The time series plots would accordingly show larger interquartile ranges, which would be purely an artifact of the factor matching technique “lumping apples with oranges”, misleading the practitioner into inferring that PMF’s fit was more uncertain than it in fact was. In contrast, the neural network method uses pattern recognition to classify factor profiles that may differ in some species between datasets, differences that may be enough to throw off measures like linear correlation, but not an expert observer.

4.4 The nonparametric versus parametric bootstrap

The method used here makes use of a nonparametric bootstrap for creating replicate datasets. The term “nonparametric” refers to the fact that the bootstrap resamples the data itself, as opposed to data from a generating process for which parameters would have to be set. The parametric approach assumes some data generating process is an accurate approximation for the data actually in hand. In the present setting, however, there are often dozens of chemical species comprising PM_{2.5} data, each likely characterized by a different probability density function and cross correlation with other species. Accordingly, the parametric bootstrap does not appear to be a feasible tool. Another version of a parametric bootstrap resamples residuals from a model fit. In the present context this approach could be outlined as: Given a data matrix \mathbf{X} , use PMF to find a factorization; bootstrap rows from the resulting residuals matrix, \mathbf{E} , and add them to rows of \mathbf{X} to create a new data matrix, \mathbf{X}^* ; use PMF to find a factorization for \mathbf{X}^* ; repeat the previous steps as desired. A fundamental assumption of this approach is that the model is the true model, and, given this assumption, residuals should be independent and identically distributed. For

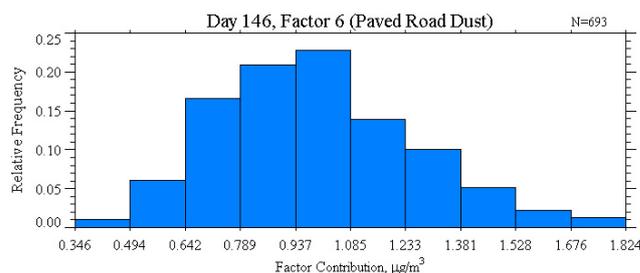


Fig. 4. Histogram of results associated with PMF solutions for factor 6’s contribution on day 146. This example represents a “vertical slice” from the contribution time series in Fig. 3f and can be calculated for any factor-day combination.

the simulation presented here, however, the base case solution had associated residuals for eight of the 39 species that failed some basic test of independence (for example, runs up, runs below the mean, and length of runs). While it may be possible to fine tune the PMF2 algorithm settings to improve the results, the residuals bootstrap rests upon the assumption that the model from which the residuals come is “true”. The authors believe that this is inappropriate given the level of model uncertainty present. In contrast, the nonparametric bootstrap employed here gives focus to the PM_{2.5} data itself, avoiding assumptions about the validity of the model fit to that data. The surrounding method for assessing that model’s quality is equally applicable to PMF results as it is to another source apportionment model, and is applicable to assessing estimates of source profiles as well as estimates of source contributions.

4.5 Method improvements

The method presented here can likely be made even more robust, and the authors propose two options to explore. The first is to consider the neural networks, with respect to how they are trained and their structure. The networks could train on more information than just the scaled factor profiles. For example, additional input-layer nodes could encode information about factor contributions or important tracer/marker species. Additionally, different network architectures could be explored, for example, adding hidden layers, hidden layer nodes, changing the node activation functions, or the initialization of weights.

The second option pertains to assessing replicate datasets before fitting the PMF model, and fundamentally, this might include examining the choice of the bootstrap method itself. In the present discussion replicate datasets generated by bootstrapping were not examined in any way for being “realistic” prior to the PMF model fit. Heidam (1987) presented a bootstrap method in which the replicate datasets were first screened by looking at their associated covariance matrices. If a given covariance matrix was not representative of the covariance structure assumed to be truly in the data, then that

bootstrapped dataset was not fit by the source apportionment model. There are numerous accept-reject criteria that could be employed such that non-representative replicate datasets would not be fit by PMF. For example, if certain marker species or “rare event” sampling days were deemed critical to the model fit, replicate datasets could be tested for sufficient representation of those data before use in subsequent analyses. This approach was avoided in the present discussion in order to focus on the method’s performance with as few practitioner-defined assumptions as possible. In certain settings, however, such assumptions may be warranted.

With respect to the underlying choice of bootstrap method, the effect of block length choice for speciated PM data should be explored. It is known that the *stationary block bootstrap* (SBB) of Politis and Romano (1994), which uses random block lengths, is less sensitive to block size misspecification when compared to the CBB employed in the present work, or the moving block bootstrap (MBB) used in EPA PMF 1.1 (Politis and Romano, 1994). Thus, using the SBB could provide a simple way of mitigating block size misspecification. More sophisticated (and harder to implement) methods of addressing block size choice also exist. For example, Christensen and Sain (2002) provide a bootstrap variant and goodness-of-fit test for choosing a block size for resampling multivariate data, and Rajagopalan (1999) developed a *k-nearest-neighbor* bootstrap method for resampling from a multivariate state space. Future investigation into, and application of, bootstrapping schemes that best reproduce the correlation structure in multivariate data is needed.

Acknowledgements. This work is supported under the National Institute of Health grant 972343/7 R01 ES012197-02. The authors wish to thank Balaji Rajagopalan, Department of Civil Engineering, University of Colorado, for assistance with understanding multivariate, nonparametric time series simulation methods; Shelly Eberly, formerly with the US Environmental Protection Agency, for assistance with understanding the methods used by EPA PMF 1.1 software; Sverre Vedal, Department of Environmental and Occupational Health Sciences, University of Washington, for assistance with understanding the role of PMF model results in health effects studies. Finally, the authors wish to thank the editors and referees for their many helpful comments on how to improve the manuscript.

Edited by: S. Pandis

References

- Anderson, M. J., Daly, E. P., Miller, S. L., and Milford, J. B.: Source apportionment of exposures to volatile organic compounds: II. Application of receptor models to TEAM study data, *Atmos. Environ.*, 36, 3643–3658, 2002.
- Brinkman, G., Vance, G., Hannigan, M. P., and Milford, J. B.: Use of synthetic data to evaluate positive matrix factorization as a source apportionment tool for PM_{2.5} exposure data, *Environ. Sci. Technol.*, 40, 1892–1901, 2006.
- Cadle, S. H., Mulawa, P., Hunsanger, E. C., Nelson, K., Ragazzi, R. A., Barrett, R., Gallagher, G. L., Lawson, D. R., Knapp, K. T., and Snow, R.: Light-duty motor vehicle exhaust particulate matter measurement in the Denver, Colorado, area, *J. Air Waste Manage.*, 49, 164–174, 1999.
- Carlstein, E.: The use of subseries values for estimating the variance of a general statistic from a stationary sequence, *Ann. Stat.*, 14, 1171–1179, 1986.
- Chatfield, C.: Model uncertainty, data mining and statistical inference, *J. Roy. Stat. Soc. A Sta.*, 158, 419–466, 1995.
- Chen, L. W. A., Watson, J. G., Chow, J. C., and Magliano, K. L.: Quantifying PM_{2.5} source contributions for the San Joaquin valley with multivariate receptor models, *Environ. Sci. Technol.*, 41, 2818–2826, 2007.
- Chinkin, L. R., Coe, D. L., Funk, T. H., Hafner, H. R., Roberts, P. T., Ryan, P. A., and Lawson, D. R.: Weekday versus weekend activity patterns for ozone precursor emissions in California’s south coast air basin, *J. Air Waste Manage.*, 53, 829–843, 2003.
- Christensen, W. F. and Sain, S. R.: Accounting for dependence in a flexible multivariate receptor model, *Technometrics*, 44, 328–337, 2002.
- Dockery, D. W., Pope, C. A., Xu, X. P., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G., and Speizer, F. E.: An association between air-pollution and mortality in 6 United-States cities, *New Engl. J. Med.*, 329, 1753–1759, 1993.
- Dominici, F., McDermott, A., Daniels, M., Zeger, S. L., and Samet, J. M.: Revised analyses of the national morbidity, mortality, and air pollution study: Mortality among residents of 90 cities, *J. Toxicol. Env. Heal. A*, 68, 1071–1092, 2005.
- Eberly, S. I.: EPA PMF 1.1 User’s Guide, U.S. Environmental Protection Agency, Research Triangle Park, NC, 2005.
- Efron, B.: 1977 Rietz lecture – bootstrap methods – another look at the jackknife, *Ann. Stat.*, 7, 1–26, 1979.
- Fine, P. M., Cass, G. R., and Simoneit, B. R. T.: Chemical characterization of fine particle emissions from the fireplace combustion of wood types grown in the midwestern and western United States, *Environ. Eng. Sci.*, 21, 387–409, 2004.
- Hall, P., Horowitz, J. L., and Jing, B. Y.: On blocking rules for the bootstrap with dependent data, *Biometrika*, 82, 561–574, 1995.
- Hannigan, M. P.: Mutagenic particulate matter in air pollutant source emissions and in ambient air, Doctorate of Philosophy, California Institute of Technology, Pasadena, CA, 221 pp., 1997.
- Haykin, S.: *Neural networks: A comprehensive foundation*, Prentice Hall, 2nd Ed., Upper Saddle River, 1998.
- Heidam, N. Z.: Bootstrap estimates of factor model variability, *Atmos. Environ.*, 21, 1203–1217, 1987.
- Huang, S. L., Rahn, K. A., and Arimoto, R.: Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island, *Atmos. Environ.*, 33, 2169–2185, 1999.
- Kiefer, N. M. and Vogelsang, T. J.: A new asymptotic theory for heteroskedasticity-autocorrelation robust tests, *Economet. Theor.*, 21, 1130–1164, 2005.
- Kim, E. and Hopke, P. K.: Source identifications of airborne fine particles using positive matrix factorization and US Environmental Protection Agency positive matrix factorization, *J. Air Waste Manage.*, 57, 811–819, 2007.
- Kunsch, H. R.: The jackknife and the bootstrap for general stationary observations, *Ann. Stat.*, 17, 1217–1241, 1989.
- Lahiri, S. N.: Effects of block lengths on the validity of block resampling methods, *Probab. Theory Rel.*, 121, 73–97, 2001.
- Lahiri, S. N.: *Resampling methods for dependent data*, Springer

- series in statistics, Springer-Verlag, New York, 2003.
- Larsen, R. K. and Baker, J. E.: Source apportionment of polycyclic aromatic hydrocarbons in the urban atmosphere: A comparison of three methods, *Environ. Sci. Technol.*, 37, 1873–1881, 2003.
- Lee, E., Chan, C. K., and Paatero, P.: Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong, *Atmos. Environ.*, 33, 3201–3212, 1999.
- Lewis, C. W., Norris, G. A., Conner, T. L., and Henry, R. C.: Source apportionment of Phoenix PM_{2.5} aerosol with the Unmix receptor model, *J. Air Waste Manage.*, 53, 325–338, 2003.
- Lough, G. C.: Sources of metals and NMHCs from motor vehicle roadways, Doctor of Philosophy University of Wisconsin, Madison, WI, 2004.
- Munakata, T.: Fundamentals of the new artificial intelligence: Beyond traditional paradigms, in: Graduate texts in computer science, edited by: Grles, D. and Schneider, F. B., Springer-Verlag, New York, 1998.
- Nitta, H., Ichikawa, M., Sato, M., Konishi, S., and Ono, M.: A new approach based on a covariance structure model to source apportionment of indoor fine particles in Tokyo, *Atmos. Environ.*, 28, 631–636, 1994.
- Norris, G. A., Vedantham, R., and Duvall, R.: EPA Unmix 6.0 Fundamentals & User Guide, US Environmental Protection Agency, Research Triangle Park, NC, 2007.
- Olson, D. A., Norris, G. A., Seila, R. L., Landis, M. S., and Vette, A. F.: Chemical characterization of volatile organic compounds near the World Trade Center: Ambient concentrations and source apportionment, *Atmos. Environ.*, 41, 5673–5683, 2007.
- Paatero, P. and Tapper, U.: Positive matrix factorization – a nonnegative factor model with optimal utilization of error-estimates of data values, *Environmetrics*, 5, 111–126, 1994.
- Paatero, P.: Least squares formulation of robust non-negative factor analysis, *Chemometr. Intell. Lab.*, 37, 23–35, 1997.
- Paatero, P., Hopke, P. K., Song, X. H., and Ramadan, Z.: Understanding and controlling rotations in factor analytic models, *Chemometr. Intell. Lab.*, 60, 253–264, 2002.
- Paatero, P., Hopke, P. K., Begum, B. A., and Biswas, S. K.: A graphical diagnostic method for assessing the rotation in factor analytical models of atmospheric pollution, *Atmos. Environ.*, 39, 193–201, 2005.
- Paatero, P.: User's guide for positive matrix factorization programs PMF2 and PMF3, part 1: tutorial, University of Helsinki, Finland, 2007.
- Peel, J. L., Tolbert, P. E., Klein, M., Metzger, K. B., Flanders, W. D., Todd, K., Mulholland, J. A., Ryan, P. B., and Frumkin, H.: Ambient air pollution and respiratory emergency department visits, *Epidemiology*, 16, 164–174, 2005.
- Polissar, A. V., Hopke, P. K., and Paatero, P.: Atmospheric aerosol over Alaska – 2. Elemental composition and sources, *J. Geophys. Res.-Atmos.*, 103, 19045–19057, 1998.
- Politis, D. N. and Romano, J. P.: A Circular Block-Resampling Procedure for Stationary Data, in: Exploring the Limits of Bootstrap, edited by: LePage, R. and Billard, L., John Wiley, New York, 1992.
- Politis, D. N. and Romano, J. P.: The stationary bootstrap, *J. Am. Stat. Assoc.*, 89, 1303–1313, 1994.
- Politis, D. N. and White, H.: Automatic block-length selection for the dependent bootstrap, *Economet. Rev.*, 23, 53–70, 2004.
- Pope, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D.: Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution, *J. Amer. Med. Assoc.*, 287, 1132–1141, 2002.
- Rajagopalan, B. and Lall, U.: A k-nearest-neighbor simulator for daily precipitation and other weather variables, *Water Resour. Res.*, 35, 3089–3101, 1999.
- Ramadan, Z., Song, X. H. and Hopke, P. K.: Identification of sources of Phoenix aerosol by positive matrix factorization, *J. Air Waste Manage.*, 50, 1308–1320, 2000.
- Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of fine organic aerosol. 2. Noncatalyst and catalyst-equipped automobiles and heavy-duty diesel trucks, *Environ. Sci. Technol.*, 27, 636–651, 1993.
- Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of fine organic aerosol. 3. Road dust, tire debris, and organometallic brake lining dust – roads as sources and sinks, *Environ. Sci. Technol.*, 27, 1892–1904, 1993.
- Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of fine organic aerosol. 4. Particulate abrasion products from leaf surfaces of urban plants, *Environ. Sci. Technol.*, 27, 2700–2711, 1993.
- Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of fine organic aerosol. 5. Natural-gas home appliances, *Environ. Sci. Technol.*, 27, 2736–2744, 1993.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, *Nature*, 323, 533–536, 1986.
- Schauer, J. J.: Source contributions to atmospheric organic compound concentrations: Emissions measurements and model predictions, Doctor of Philosophy, California Institute of Technology, Pasadena, CA, 399 pp., 1998.
- Schauer, J. J., Kleeman, M. J., Cass, G. R., and Simoneit, B. R. T.: Measurement of emissions from air pollution sources. 1. C-1 through C-29 organic compounds from meat charbroiling, *Environ. Sci. Technol.*, 33, 1566–1577, 1999.
- Shafri, H. Z. M., Suhaili, A., and Mansor, S.: The performance of maximum likelihood, spectral angle mapper, neural network and decision tree classifiers in hyperspectral image analysis, *Journal of Computer Science*, 3, 419–423, 2007.
- Singh, K.: On the asymptotic accuracy of Efron's bootstrap, *Ann. Stat.*, 9, 1187–1195, 1981.
- Tong, C. S. and Cheng, K. C.: Mass spectral search method using the neural network approach, *Chemometr. Intell. Lab.*, 49, 135–150, 1999.
- van der Meer, F.: The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery, *Int. J. Appl. Earth Obs.*, 8, 3–17, 2006.
- Vedal, S., Dutton, S. J., Hannigan, M. P., Milford, J. B., Miller, S. L., Rabinovitch, N., and Sheppard, L.: The Denver Aerosol Sources and Health (DASH) Study: 1 Overview, *Atmos. Environ.*, in press, 2009.
- Watson, J. G., Fujita, E., Chow, J., Zielinska, B., Richards, L. W., Neff, W., and Dietrich, D.: Northern Front Range Air Quality Study final report, Desert Research Institute, Fort Collins, CO, 1998.
- Werbos, P. J.: Beyond regression: New tools for prediction and analysis in the behavioral sciences, Doctorate of Philosophy, Harvard University, Cambridge, 1974.