Atmos. Chem. Phys. Discuss., 8, S4881–S4891, 2008
www.atmos-chem-phys-discuss.net/8/S4881/2008/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Assessing positive matrix factorization model fit: a new method toestimate uncertainty and bias in factor contributions at the daily timescale" *by* J. G. Hemann et al.

J. G. Hemann et al.

Received and published: 15 July 2008

The authors would like to thank the Referee for his/her thoughtful and detailed comments. Regarding the issues raised,

1)
The authors agree with the Referee that other pollution sources are likely present in other contexts, and inclusion of such sources may complicate PMF modeling. With respect to the Denver metro area, the authors feel that the choice of sources used in the synthetic data are well-founded by the Northern Front Range Air Quality Study

(NFRAQS, Watson et al., 1998). The authors will add a qualifying sentence to section 2.2 (Synthetic Data): *"It should be noted that the presence of additional sources, such as secondary organic aerosols, could complicate application of PMF to observed data."*

2)

The authors agree that the discussion at the end of section 4.1 should be clarified. The creation of, and testing with, synthetic data is not truly a component of the new method. Synthetic data is used in the present work simply as a means of validating the method. The authors stand by the assertion however that the new method could help identify pollution sources that require greater effort to accurately model. If realistic synthetic data exists for a specific context, and if PMF results based on that data consistently show some factors poorly modeled, then PMF results for those same factors identified in observed data from that context should be carefully scrutinized. Accordingly, the authors will modify the last sentence of section 4.1 to read *"Thus, the application of the method to synthetic data representing a specific situation could help identify sources for which contribution estimates should be carefully scrutinized."*

3)

The authors are unclear about the Referee's reference to "true uncertainty." The authors assume that the Referee is concerned about the sometimes large biases between factor contribution estimates and the true contributions used to create the synthetic data, as well as the sometimes large variability in contribution estimates. It should be stressed that application of the new method yields results that have heretofore not been presented: estimates of factor contribution variability at the measurement time scale, where the variability estimated is the variability in PMF's results due to random sampling error in the data. This variability is in contrast to the uncertainty reported by the PMF2 software (via the G_std-dev, F_std-dev and rotmat matrices), which is borne out of both noise in the data as well as the fact that the ***X* =**

ACPD

8, S4881–S4891, 2008

Interactive Comment

GF factorization is generally not unique (Paatero, 2007, part 1).

The authors agree with the Referee that the PMF results are troubling, at least for some of the factors. However, it is not necessarily true that PMF should perform better when fitting synthetic data versus observed data. The Referee's concern about application of PMF results to subsequent epidemiological studies is critically important but outside the scope of the present work. Given that PMF results may be used not just for characterization of air pollution, but also for public health policy decisions, the presented method is hopefully part of what seems to be an increased desire to estimate pollution source apportionment uncertainties. For example, Christensen (2008) and Park et al. (2000) have presented Bayesian approaches to pollution source apportionment that naturally yield estimates of uncertainty for every element of the profile and contribution matrices.

To enhance the discussion in section 4.2, the authors propose adding the following:

"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 source 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

ACPD

8, S4881–S4891, 2008

the one presented here, to further assess their pollution source apportionment results."

4)
The authors appreciate the Referee's concern. A Conclusion section was not included because the authors felt that it would be needlessly repetitive, especially since the main contributions of the work concern a new method for analyzing PMF results, rather than the results themselves.

5)
The authors appreciate the Referee's desire to see more information about how the synthetic data was created. The authors feel that the methods used to create the synthetic data are extensively referenced and in order to keep the paper streamlined the authors will refrain from adding discussion. The Referee is asked to see the authors' response to a similar question by Anonymous Referee 1 (item 1). The authors propose adding the factor contribution correlations table from this response to the manuscript in section 2.2 (Synthetic Data).

6)
The data are assumed to come from a single receptor site and this will be noted in section 2.2 (Synthetic Data).

7)
The authors agree that this information about the synthetic data should be included to help the reader. Table 2 will be updated to include the measurement detection limits and errors associated with each species.

8)

ACPD

8, S4881–S4891, 2008

Quantifying the benefit of using the presented neural network-based approach versus the traditional, linear correlation-based approach would be very interesting yet very difficult to actually compute. It is important to note that the traditional method is typically employed via the *EPA PMF 1.1* software. (The authors say "typically" based on their informal surveying of published literature in which uncertainty analysis via the bootstrap is applied to PMF results.) The following discussion is within that context.

At the outset of the work presented, a variation of the traditional factor matching method (based on linear correlation) was employed. The variation was to identify tracer/marker species for each factor (that is, species which were predominantly present due to the activity of a single factor), and use that information along with linear correlation between factor contribution time series. The method is outlined as follows:

Let

N = Number of sampling days

M = Number of species measured per sampling day

G = The *NxP* matrix of factor contributions, where ***G****(i,j)* corresponds to the j^{th} factor's contribution on the i^{th} day.

F = The *PxM* matrix of factor profiles, where ***F****(i,j)* corresponds to how much the i^{th} factor accounts for the j^{th} pollutant species.

For a given bootstrap solution we want to match factors (columns of ***G***) with their closest basecase counterpart, where *bootstrap solution* refers to the ***F*** and ***G*** matrices from PMF fitting a bootstrap replicate dataset, and *basecase* refers to the ***F*** and ***G*** matrices resultant the PMF fit on the original data. The goal of factor matching is to collect the ***F***

ACPD

8, S4881–S4891, 2008

Interactive Comment

and ***G*** matrices output from PMF fitting of bootstrapped datasets and reorder their rows and columns, respectively, such that, for example, column 2 always refers to "factor 2", regardless of which solution is under consideration. Then, bootstrap factor *j* is matched best with basecase factor *i* if

a) Bootstrap factor *j* accounts for most of species *k* if species *k* is a "marker" for basecase factor *i*

b) Bootstrap ***G****(*,j)* has the highest linear correlation with basecase ***G****(*,i)*, with the correlation also being greater than some threshold

This method could be massaged into working well for a given dataset. For example, by applying the method over and over again to bootstrap replicate datasets, a good correlation threshold could be found, as well as good thresholds defining what made a species a marker species for each factor. However, the method was generally not robust and was unacceptable because

a) The thresholds would typically need to be painstakingly adjusted, ad hoc, for even the slightest changes in the data.

b) The thresholds for defining marker species are based on the basecase solution, but due to random sampling error in the data, marker species would be not always be present for all factors in the bootstrap replicate datasets.

c) Bootstrap replicate datasets do not preserve the temporal patterns seen in the original data when viewed over the course of the entire measurement period. Accordingly, using a non-robust statistic like linear correlation to match bootstrap contribution time series with basecase contribution time series is fatally flawed.

ACPD

8, S4881–S4891, 2008

Interactive Comment

Practically addressing issues a) and b) force the practitioner into data dredging and ad hoc choices of goodness-of-fit. Issue c) is addressed in the *EPA PMF 1.1* software by allowing many-to-one factor matching (multiple bootstrap factors may match to the same base case factor within a given solution) and lumping non-matchable factors across bootstrap solutions into a single factor (Eberly, 2005). The problem with this behavior is that factors that are temporally correlated (e.g. road dust and diesel combustion) may easily have their labeling incorrectly interchanged across solutions. The variability in the PMF solutions across bootstrapped data would then be due to more than just the effect of random sampling error in the data. The variability could also be an artifact of the factor matching process inconsistently classifying factors.

The use of a neural network-based factor matching approach is a central component of the new method. Factor contribution uncertainty cannot be estimated unless there is a reliable way to order solutions consistently across bootstrap data sets. By training neural networks to learn factor profiles, as opposed to using linear correlation between factor contribution time series, the authors believe that issues a, b and c are effectively dealt with. In order to make quantitative statements about how the new method performs compared to the traditional method the practitioner would need to assess, among other things, the factor matching misclassification rate. Thus, the practitioner would have to examine the factor matching between the basecase solution and each bootstrap solution associated with hundreds of bootstrap replicate datasets and find all of the mismatches, as defined subjectively by the practitioner. This proportion could then be compared between the two methods. However, it is presently impossible to calculate a misclassification proportion in results from *EPA PMF 1.1* because the software does not output graphical or numerical results for every single bootstrap solution (only summary information is reported).

9)

ACPD

8, S4881–S4891, 2008

The authors respect the Referee's opinion but are satisfied with the current organization: bootstrapping is discussed initially to provide background on what methods of assessing uncertainty in pollution source apportionment have been tried. Section 2 is Methodology, and because the work is primarily focused on presenting a new method, this section has increased content and discussion.

10)
The authors thank the Referee for his/her attention to detail. A reference to PMF2 will be added to section 2.1.

11)
The authors agree that the eight versus nine factor results could be clarified. The eight and nine factor solutions are briefly considered in section 3 (Results) because those choices would be the ones likely made by practitioners using common methods for choosing the number of factors. Given that the known solution has nine factors, the main focus of the work is on the simulations generating nine factor solutions. Accordingly, the offending sentence at the end of section 2.1 will be changed to *"In the present work, eight and nine factor solutions are considered, with the primary focus on the results for the nine factor solutions."*

12)
The authors agree that FPEAK is an important parameter to the PMF algorithm. However, the choice of FPEAK (or any of the myriad PMF tuning parameters) is not directly relevant to the method presented for analyzing PMF (or another pollution source apportionment model) results. The Referee is asked to see the authors' response to Anonymous Referee 1 (item 1) regarding FPEAK.

ACPD
8, S4881–S4891, 2008

13)
The authors agree that the mention of including oxygen for mass closure is misleading: in fact, oxygen is included merely for accounting of the species typically measured in observed data. The sentence in section 2.2 regarding oxygen and 30% will be removed.

14)
Regarding the inclusion of ammonium sulfate and ammonium nitrate, the authors agree with the Referee that PMF is likely not able to model the sources of these secondary species. These species were carried along in the analysis as significant constituents of Denver aerosol, according to the NFRAQS study.

15)
As with the inclusion of oxygen, ammonium is included primarily for "accounting." Importantly, ammonium is known to be in abundance in the Denver metro area airshed and its inclusion helps the reader understand the nature of Denver's inorganic aerosols.

16)
The authors agree with the Referee and will update the legends accordingly.

17)
The factor profiles used in the neural network training were created via the same method used for generating all factor profiles, PMF model fitting of bootstrap replicate datasets. Generally speaking, it can be said that the variability in the factor profile estimates is due to the effect of random sampling error in the data on PMF. However, a more detailed discussion about why profile estimates for some factors deviate greatly from the known profiles is beyond the scope of the present work.

ACPD
8, S4881–S4891, 2008

18)
The authors appreciate the Referee's suggestion, but the authors felt that the seven and 10 factors were unlikely to be chosen if practitioners used the commonly used methods for choosing the number of factors (e.g. assessing ***Q*** values, the residuals associated with specific species, and the physical interpretability of the factors). Accordingly, the simulation statistics for only the eight and nine factor solutions are presented.

References

Christensen, W.F.: Integrating diverse sources of airshed information in pollution source apportionment, presented at the 19th annual conference of The International Environmetrics Society, Kelowna, Canada, 2008. http://people.ok.ubc.ca/zhrdlick/ties08/invited.htm (accessed May 30, 2008)

Eberly, S. I.: EPA PMF 1.1 User's Guide, U.S. Environmental Protection Agency, Research Triangle Park, NC, 2005.

Paatero, P., Hopke, P. K., Song, X. H., and Ramadan, Z.: Understanding and controlling rotations in factor analytic models, Chemometrics and Intelligent Laboratory Systems, 60, 253-264, 2002.

Paatero, P.: User's guide for positive matrix factorization programs PMF2 and PMF3, part 1: tutorial, University of Helsinki, Finland, 2007.

Paatero, P.: User's guide for positive matrix factorization programs PMF2 and PMF3,

ACPD

8, S4881–S4891, 2008

part 2: reference, University of Helsinki, Finland, 2007.

Park, E.S., Oh, M.S., Guttorp, P.: Multivariate receptor models and model uncertainty, Technical Report Series 060, National Research Center for Statistics and the Environment, 2000.

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.

Interactive comment on Atmos. Chem. Phys. Discuss., 8, 2977, 2008.