Supplement of Atmos. Chem. Phys. Discuss., 15, 10123–10162, 2015
http://www.atmos-chem-phys-discuss.net/15/10123/2015/
doi:10.5194/acpd-15-10123-2015-supplement

Atmospheric Chemistry and Physics Discussions

Supplement of

Receptor modelling of both particle composition and size distribution from a background site in London, UK

D. C. S. Beddows et al.

Correspondence to: R. M. Harrison (r.m.harrison@bham.ac.uk)

1. PMF Methodology

PMF solves the general receptor modelling problem using constrained, weighted, least-squares (see Reff et al. (2007) for a review of PMF methods). The general model assumes there are *p* sources, source types or source regions (termed factors) impacting a receptor (in this case the North Kensington monitoring site), and linear combinations of the impacts from the *p* factors to the observed concentrations of the various species or in this case size bins for the SMPS+APS spectra plus auxiliary measurements (including: gas concentrations, meteorological data and traffic measurements) (see equation 1). Mathematically, observation x_{ij} , or in this case the particle size distribution plus auxiliary measurements, at the receptor is given by the matrix representation whose elements are,

$$x_{ij} = \sum_{h=1}^{p} g_{ij} \cdot f_{hj} + e_{ij} \tag{1}$$

where the j^{th} size bin (or auxiliary measurement) on the i^{th} hour. The term g_{ik} is the contribution of the k^{th} factor to the receptor on the i^{th} hour, f_{kj} is the fraction of the k^{th} factor that contributes to measurement *j*, and e_{ij} is the residual for the j^{th} measurement on the i^{th} hour. In PMF, only x_{ij} are known and the goal is to estimate the contributions (g_{ik}) and the fractions (f_{ij}). It is assumed that the contributions and number fractions are all non-negative, hence the "constrained" part of the least-squares. Furthermore, PMF uses uncertainties measured for each of the x_{ij} size-bin. Hours with high uncertainty are not allowed to influence the estimation of the contribution and fractions as much as those with small uncertainty, hence the "weighted" part of the least squares.

Given the above, it is task of PMF to minimise the sum of the squares Q, see equation 2.

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{e_{ij}}{s_{ij}} \right)^2 \tag{2}$$

where s_{ij} is the uncertainty in the j^{th} measurement for hour *i*. PMF also operates a robust mode, meaning that "outliers" are also not allowed to overly influence the fitting of the contributions and profiles.

The elements of the matrix S, is derived from the uncertainties entered by the user. These can either be entered directly as a matrix using the X_std-dev file or using one of various ad-hoc computations available to PMF. In general, the X_std-dev entries should predict the average size of the residual of the data value in question; this is the 'golden rule' in assigning these values (private communication with Pennti Paatero, 2010). The method chosen for calculating these values is based on the method used by Ogulei et al. (2006a, 2006b). In this, S is calculated using equation 3,

$$s_{ij} = t_{ij} + v_{ij} \max(|x_{ij}||y_{ij}|) \tag{3}$$

Where x_{ij} are the actual data values and y_{ij} are the equivalent data values fitted by PMF. Matrices t_{ij} and v_{ij} are given by

$$t_{ij} = T(x_{ij} + \overline{x_j}) \tag{4}$$

$$v_{ij} = V \tag{5}$$

Ogulei et al. (2006) used constant T values between 0.01 and 0.05 in their work which was decided through trial and error. Similarly, the constant values were chosen such that their calculated Q value was the closest to the theoretical value it could be. In this work, the optimum values were derived my methodically scanning though all the possible combinations (see Table S2 optimum results). The choice of the number of factors is a compromise according to Lee et al. (1999). Two measures *IM* and *IS* were derived from the elements (r_{ij}) of the scaled residual matrix **R** (equations 6 and 7) to help further decide on the optimum settings**.** Using these the compromise between using too few factors which combine sources of different nature and using too many factors dissociate into two or more non-existing factors was made.

$$\boldsymbol{R} = \left(\boldsymbol{X} - \boldsymbol{GF}\right) / \boldsymbol{S} \tag{6}$$

$$IM = \max\left(\frac{1}{n} \cdot \sum_{i=1}^{n} r_{ij}\right) \tag{7}$$

$$IS = \max\left(\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(r_{ij} - \bar{r}_j\right)^2}\right) \tag{8}$$

Each column of R represents the quality of the fitting of each species (in this case particle size bin count dN/dLog(Dp) or the auxiliary metric) to the product GF and from this the values of IM and IS are calculated. IM is the maximum individual column mean and IS is the maximum individual column standard deviation. Both serve as indicators to identify the species having the least fit and the most imprecise fit, respectively. Using the critical number of factors, IS and IM drop in value to a plateau and this drop indicates the minimum number of factors that should be used in the model.

The maximum number of factors is selected using a third value taken as the maximum element outputted from the rotation matrix (rotmat). This matrix is used for detecting the degree of rotational freedom of the factors. It is only qualitative in nature but can be used to reveal if factors have excessive rotational freedom. In this case, a small number of factors should be used. Choosing the largest element in the rotational matrix can show the worst case in the rotational freedom and on increasing the number of factors a critical point will be reached where the value of the largest element will increase from a plateau. This critical point indicates the upper range of factors recommended in the model.

Having selected the optimum settings (Table S1), the model was optimised by increasing the uncertainties of the outliers by varying degrees to reduce their influence on the results. For all elements where the Scaled Residual (SR) was between 4 and 7, the uncertainties were increased by a factor of 10 and for all elements where SR was greater than 7, the uncertainties were multiplied by 15. To a lesser degree, all elements with an SR between 3 and 4 were multiplied by 3 and those with an SR between 2 and 3 were multiplied by 1.5. Similarly, in order to get a reasonable OC/EC ratio our PM_{10} a FKEY = 5 for OC had to be applied to our Traffic Factor.

The output from PMF2 analysis was then scaled to the measured concentration using a scaling constant, z_k, was taken as the measured total PM_{10} mass (or NSD concentration) within each factor as defined the in parametric form,

$$x_{ij} = \sum_{k=1}^{p} (z_k g_{ik}) \left(\frac{f_{kj}}{z_k} \right) \quad (9)$$

Such scaling results in unit less factors F which describe the characteristics of the sources and time series G with units of $\mu g/m^3$.

Also included in the output is a value of the Effective Variation EV.

$$\mathrm{EV_{kj}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{m}} \left| \mathrm{g_{ik}} \mathrm{f}_{kj} \right| / \mathrm{s_{ij}}}{\sum_{\mathrm{j=1}}^{\mathrm{m}} \left(\sum_{\mathrm{h=1}}^{\mathrm{p}} \left| \mathrm{g_{ih}} \mathrm{f_{hj}} \right| + \left| \mathrm{e_{ij}} \right| \right) / \mathrm{s_{ij}}} \quad \text{(for } k = 1 \ldots \mathrm{p}) \quad (10)$$

For a matrix X[n,m] of *n* observations with *m* constituents of interest, this is a dimensionless quantity with a value from 0.0 (no variance explained) to 1.0 (100% variance explained). Defined by equation (10), it summarizes how important each factor element is in explaining one column of the observed matrix X, i.e. equation (10) defines how the j^{th} element if the k^{th} of the factor F explains the j_{th} column of the input data matrix X. EV is by and large most useful in discerning which constituents of a factor are the most important, since a large EV indicates that this particular factor explains a major proportion of that species variability. For a given *p* factor solution there a (p+1) 'factor' is outputted from PMF2 which represents the residual. Referring to it to as the Not Explained Variation, NEV_{kj}, the following rule should is applied in that its values should not exceed 0.25 and that we should consider that the variable question practically as "not explained". Given that $\left(\sum_{k=1}^{p} EV_{kj}\right) + NEV_{kj} = 1$ we can also say that the Total EV or TEV = $\left(\sum_{k=1}^{p} EV_{kj}\right) > 0.75$ to be an explained variable. For the PM_{10} work the TEV values varied from 0.79 to 0.91 and for the NSD work the TEV varied from 0.79 to 0.97. Similarly for the PM_{10} + NSD solution the TEV varied between 0.75 and 0.92. In other words, the TEV values gave further confidence in the chosen PMF solutions. As expected, as we slackened off the uncertainties in of either the PM_{10} or NSD uncertainties in the PM_{10} + NSD analysis, the TEV values of the corresponding constituents become "not explained". When slackening the PM_{10} uncertainties, the TEV values of the PM_{10} values fell below 0.75 and ranged 0.683 and 0.806. Most noticeable were TNA, TCL, WNH4, WNO3 and TMG which had even low TEV values, all between 0.568 and 0.663. But the NSD constituents driving the PMF solution with tight uncertainties ranged between a reassuring value of 0.883 and 0.951. Likewise for a *slack* NSD model, the TEV values of the PM_{10} varied from 0.7820 and 0.9110 and the TEV values for the NSD ranged from 0.712 and 0.838 with the lowest size bin (at 17.78 nm) registering a TEV = 0.556.

References

Fuller, G.W., Tremper, A.H., Baker, T.D., Yttri, K.E., Butterfield, D.: Contribution of wood burning to PM10 in London, Atmos. Environ., 87, 87-94, 2014.

Lee, E., Chan, C.K., Paatero, P.: Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong, Atmos. Environ. 33 3201-3212, 1999.

Ogulei, D., Hopke, P.K., Wallace, L.A.: Analysis of indoor particle size distributions in an occupied townhouse using positive matrix factorization, Indoor Air, 16, 204-215, 2006a.

Ogulei, D., Hopke, P.K., Zhou, L., Pancras, J.P., Nair, N., Ondov, J.M.: Source apportionment of Baltimore aerosol from combined size distribution and chemical composition data, Atmos. Environ., 40, S396-S410, 2006b.

Paatero, P.: Private communication, 2010.

Reff, A., Eberly, S.I. and Bhave, P.V.: Receptor modeling of ambient particulate matter data using positive matrix factorization: Review of existing methods, JAWMA, 57, 146-154, 2007.

Sandradewi, J., Prévôt, A. S. H., Weingartner, E., Schmidhauser, R., Gysel, M., Baltensperger, U.: A study of wood burning and traffic aerosols in an Alpine valley using a multi-wavelength Aethalometer, Atmos.Environ., 42, 101-112, 2008.

Table S1. Parameters optimising the modelled metrics.

Parameters						*Optimisation metrics*						
Input Data	Num of Factors	FKEY	T	V	FPEAK	IM	IS	ROT	Q_{Theory}	Q	Num of Neg. Outliers	Num of Pos Outliers
PM_{10}	6	5	0.026	0.1	0	0.20	1.31	0.24	17,544	17,575	0	0
NSD	4	-	0.054	0.02	0.1	0.24	1.44	0.004	681,474	678,305	0	1
PM_{10} and NSD	5	-	0.052	0.07	-0.25	0.30	1.45	0.019	30,090	30,114	0	0

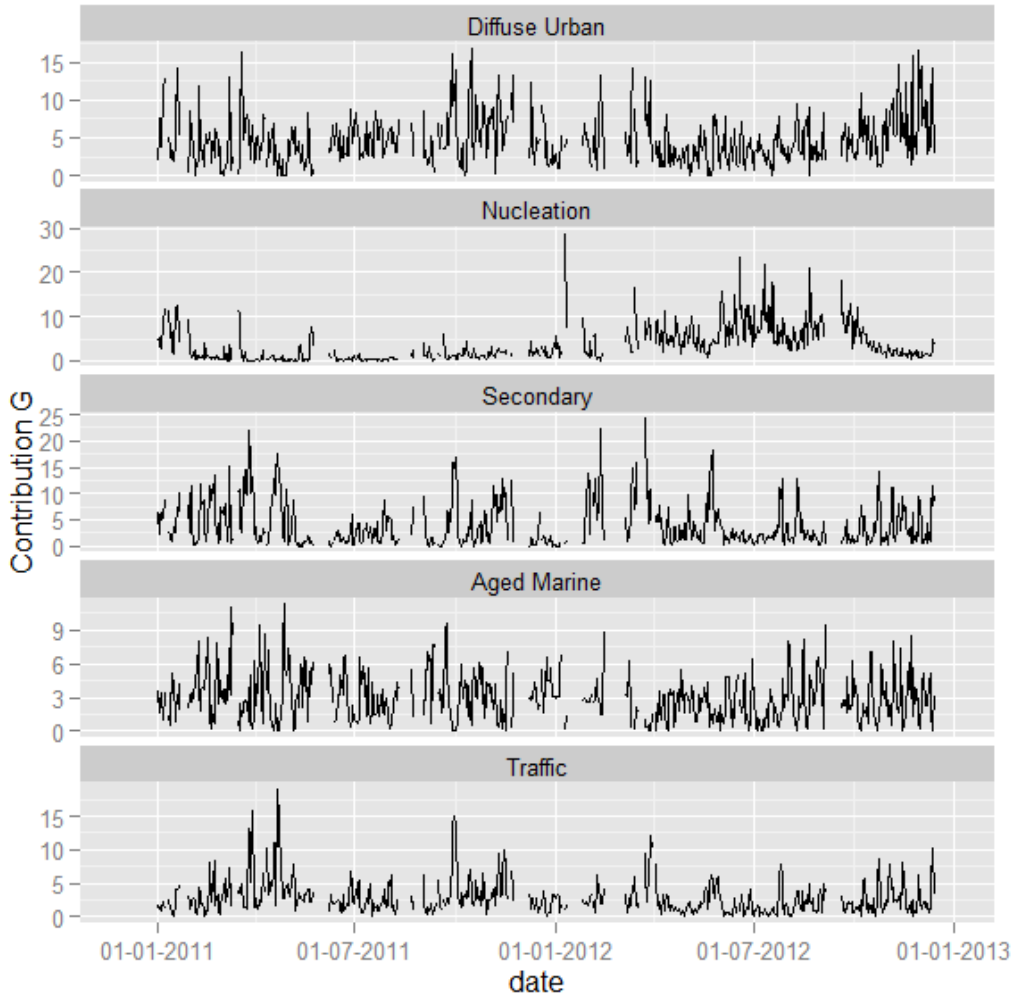

Figure S1. Five factor solution : plotting the matrix G

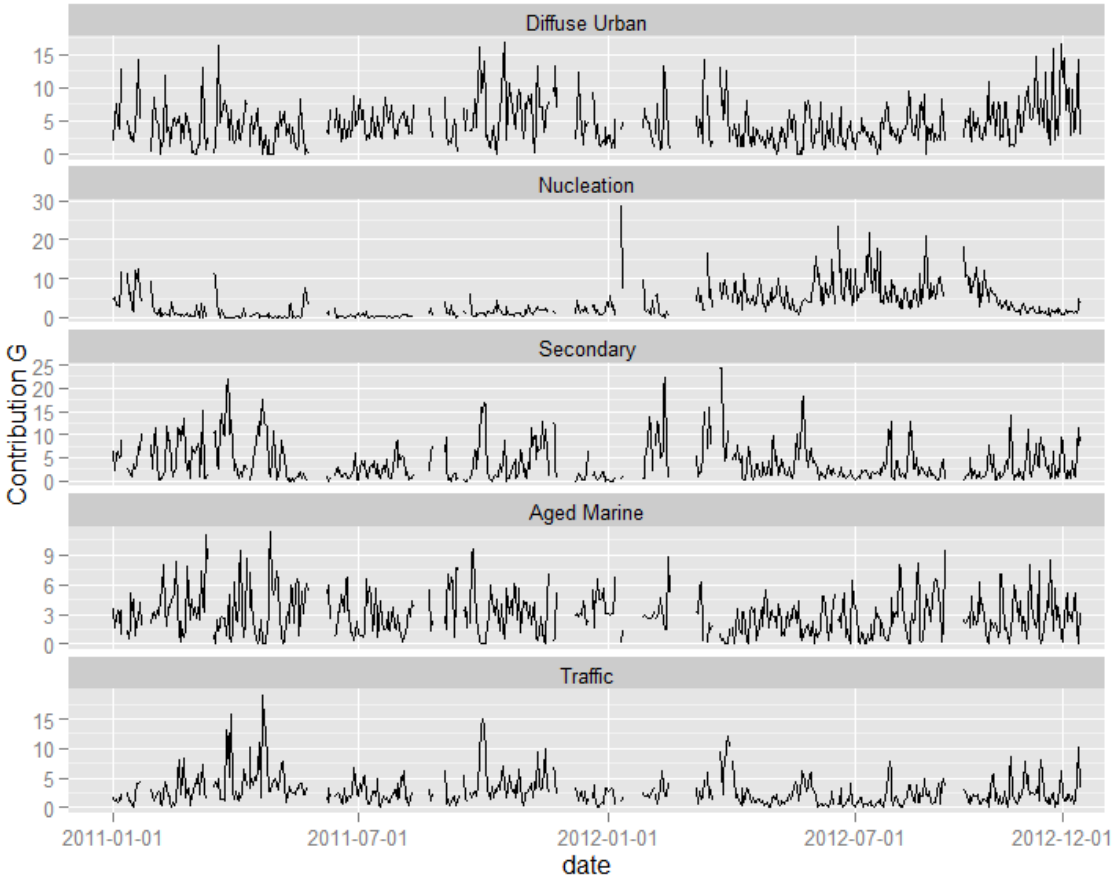

Figure S1. Five factor solution : plotting the matrix G.