

Supplement of Atmos. Chem. Phys., 15, 10107–10125, 2015  
<http://www.atmos-chem-phys.net/15/10107/2015/>  
doi:10.5194/acp-15-10107-2015-supplement  
© Author(s) 2015. CC Attribution 3.0 License.



*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](mailto:r.m.harrison@bham.ac.uk))

The copyright of individual parts of the supplement might differ from the CC-BY 3.0 licence.

## 1. PMF Methodology

PMF solves the general receptor modelling problem using constrained, weighted, least-squares applied to the input data  $x$  which represent a matrix of concentrations, albeit particle or PM, measured at specific intervals during the study (Reff et al. (2007)). The general model assumes there are  $p$  factors  $F$  which are interpreted as fixed emission source profiles and impact the receptor site by various amounts - represented by the scores  $G$  - during the measurement. . . , PMF determines the profiles of these factors and calculates their contribution  $G$  such that the sum of linear combinations  $G \times F$  of closely matches the measured concentration. . . Mathematically, the observation  $x_{ij}$ , at the receptor is represented in the matrix equation  $X = G \times F + E$  whose elements are,

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

The measurements (PM<sub>10</sub> mass or NSD size concentration) are indexed by the integer  $j$  for the  $j^{\text{th}}$  time step (hour or day)our. The term  $g_{ik}$  is the contribution of the  $k^{\text{th}}$  factor to the receptor site on the  $j^{\text{th}}$  hour/day,  $f_{kj}$  is the fraction of the  $k^{\text{th}}$  factor (PM<sub>10</sub> mass or NSD size concentration) that contributes to measurement  $j$ . Matrix  $E$ , comprises of elements  $e_{ij}$  which are the residual values for the  $j^{\text{th}}$  measurement on the  $i^{\text{th}}$  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. measurements with high uncertainty are not allowed to influence the estimation of the contribution and fractions as much as those with small uncertainty, thus giving the “weighted” part of the least squares.

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

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

where  $s_{ij}$  is the uncertainty in the  $j^{\text{th}}$  measurement for hour/day  $i$  and PMF can be operated in a robust mode, meaning that “outliers” are also not allowed to influence the fitting of the contributions and profiles. The elements of the matrix  $S$ , are derived from the uncertainties entered by the user and these can either be entered directly as a matrix using the X\_std-dev file or by 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). In our work, 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,

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

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

$$t_{ij} = T(x_{ij} + \bar{x}_j) \quad (4)$$

$$v_{ij} = V \quad (5)$$

Ogulei et al. (2006) chose the values of T and V by trial and error until their calculated Q value was the closest to the theoretical value it could be. In our work, a computer was used to loop through all the possible combinations of T and V for the optimum number of factors, calculating Q and various other optimisation metrics for each loop (Table S1).

To decide on the optimum number of factors three metrics were used. The first two, *IM* and *IS*, were calculated using the elements ( $r_{ij}$ ) of the scaled residual matrix **R** (equations 6 and 7) (Lee et al, 1999). By plotting *IM* and *IS* against factor number, the number of factors could be found, above which, no further increases in factor number would improved the fit

$$\mathbf{R} = (\mathbf{X} - \mathbf{GF}) / S \quad (6)$$

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

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

In essence, each column of R (from which the values, IM and IS are calculated) represents the quality of the fitting of each species (mass or number concentration) to the product GF.. IM is the maximum individual column mean and IS is the maximum individual column standard deviation. Hence for an increasing number of factors, the minimum number is identified when both IS and IM drop in value to a plateau. The maximum number of factors is selected using the maximum valued 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. 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. Hence, by superimposing plots of both IM, IS and ROT vs factor number on top of each other, a range of factor numbers can be selected to investigate. In addition, bivariate plots of the time series of each factor were used to identify the maximum number of factor that could be used before factor splitting occurred. This was identified by unusually high correlations which showed that two factors were from the same source.

Given the the optimum settings for the different data sets (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. In addition, once the optimum solution had been selected and the factors identified, the models were considered for further tuning using FKEY. FKEY is used in the model to constrain elements in the F matrix to modify an output that

does not agree with a priori information. In our PM<sub>10</sub> solution, the OC/EC ratio for traffic was greater than the accepted value of 0.8. To get a reasonable OC/EC ratio, a FKEY of 5 was applied to the OC column of our Traffic Factor

The chosen solution was scaled to the measured concentration using a scaling constant,  $z_k$ , using equation (9),

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

Such scaling results in a unit less factors  $F$  which describe the characteristics of the sources and time series  $G$  with units of  $\mu\text{g}/\text{m}^3$  or  $\text{cm}^{-3}$  for mass and number concentration respectively. This scaling was not carried out for the combined NSD and PM<sub>10</sub> solution due the difficulty in assigning a meaningful unit to the altered  $G$  values.

The interpretation of the factors was assisted with the Effective Variation  $EV$  which is defined by equation 10.

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

For a matrix  $X[n,m]$  of  $n$  observations with  $m$  constituents of interest,  $EV$  is a dimensionless quantity with a range of 0.0 (no variance explained) to 1.0 (100% variance explained). It shows which factor is the most important in explaining one column of the observed values in the 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 is a  $(p+1)$  factor outputted from PMF2 which represents the residual. Referring to it to as the 'Not Explained Variation',  $NEV_{kj}$ , the following rule should be applied in that its values should not exceed 0.25 and that we should consider that the variable in question is practically "not explained". Given that  $(\sum_{k=1}^p EV_{kj}) + NEV_{kj} = 1$  we can also say that the Total EV or TEV =  $(\sum_{k=1}^p EV_{kj}) > 0.75$  to be an explained variable. For the PM<sub>10</sub> 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<sub>10</sub> + 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 slacked off the uncertainties in of either the PM<sub>10</sub> or NSD uncertainties in the PM<sub>10</sub> + NSD analysis, the TEV values of the corresponding constituents become "not explained". When increasing the PM<sub>10</sub> uncertainties, the TEV values of the PM<sub>10</sub> values fell below 0.75 and ranged 0.683 and 0.806. Most noticeable were TNA, TCL, WNH4, WNO3 and TMG which had even lower 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<sub>10</sub> 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.

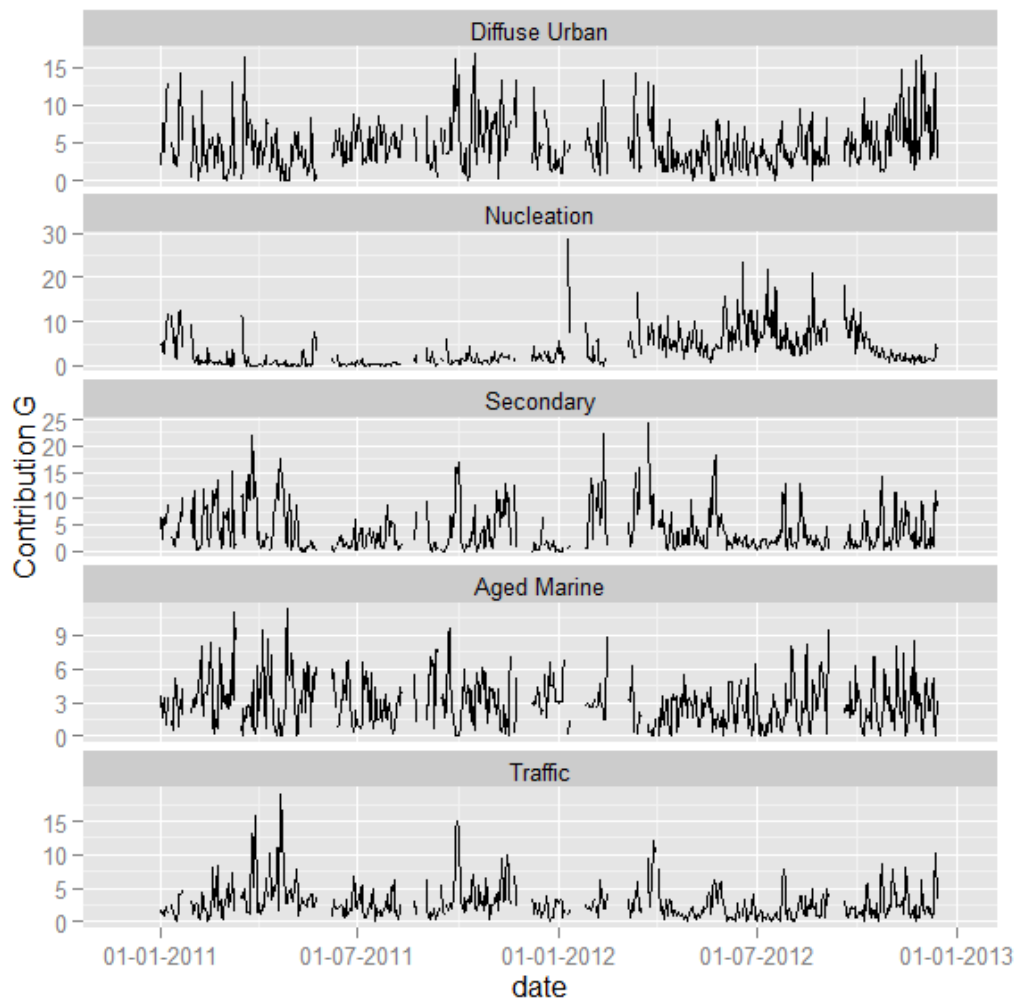
Sandradewi, J., Prévôt, A. S. H., Weingartner, E., Schmidhauser, R., Gysel, M., Baltensperger, U.: A study of wood burning and traffic aerosols in an Alpine valley using a multi-wavelength Aethalometer, *Atmos. Environ.*, 42, 101-112, 2008.

**Table S1.** Parameters optimising the modelled metrics.

<i>Parameters</i>						<i>Optimisation metrics</i>						
Input Data	Num of Factors	FKEY	T	V	FPEAK	IM	IS	ROT	Q <sup>Theory</sup>	Q	Num of Neg. Outliers	Num of Pos Outliers
PM <sub>10</sub>	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 <sub>10</sub> and NSD	5	-	0.052	0.07	-0.25	0.30	1.45	0.019	30,090	30,114	0	0

**Table S2.** Change in the percentage values shown in Pie chart in Figure 3.

Cluster	Diffuse Urban (25.9%)	Nucleation (19.9%)	Secondary (21.6%)	Aged Marine (16.7%)	Traffic (15.9%)
6	3.7	4.6	-5.9	0.7	-3.1
2,4,7	1.9	1.9	-8.9	7.6	-2.4
3	-2.6	3.8	-0.7	1.3	-1.7
1	-5.9	-8.1	13.0	-3.4	4.3
5	-1.8	-9.1	13.3	-9.3	6.9



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