Contributions from transport, solid fuel burning and cooking to primary organic aerosols in UK cities: Supplementary material – Selection of numbers of factors

J. D. Allan[1], P. I. Williams[1], W. T. Morgan[2], C. L. Martin[2], M. J. Flynn[2], J. Lee[3], E. Nemitz[4], G. J. Phillips[4], M. W. Gallagher[2] and H. Coe[2].

[1]{National Centre for Atmospheric Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK}

[2]{School of Earth, Atmospheric and Environmental Sciences, The University of Manchester, Oxford Road, Manchester M13 9PL, UK}

[3]{National Centre for Atmospheric Science, The University of York, Heslington, York YO10 5DD, UK}

[4]{Centre for Ecology and Hydrology, Bush Estate, Penicuik, Midlothian EH26 0QB, UK}

Correspondence to: J. D. Allan (james.allan@manchester.ac.uk)

Scope

The purpose of this supplementary material is to present the data used to determine the most appropriate number of factors to be used in each of the datasets presented in the paper when subjected to PMF analysis. All analysis was performed using the tools presented by Ulbrich et al. (2009). Note that variations according to rotationality are not covered here; these are discussed in the main article. The solutions from each analysis run were critically appraised according to two criteria:

I. The uniqueness of the factors derived

II. The numerical robustness of the solutions

The main method of determining uniqueness was through inspection of the temporal and mass spectral profiles of the solutions. Factors that bore a strong similarity were taken to be indicative of 'splitting', i.e. separate factors being derived that represent variations within a single factor. This is to be avoided as split factors may represent variance within multiple physical factors and therefore cause mass to be attributed incorrectly, known as 'mixing'. Numerical stability was quantitatively evaluated by performing bootstrapping analysis, randomly resampling in the time dimension. The numerical uniqueness of individual solutions was also verified by repeating the analysis and varying the initialisation seed. All of these methods are discussed by Ulbrich et al. (2009) and references therein.

Only once these tests were passed were solution sets used for further validation and analysis on a chemical basis as described in the main articles. Ones that fail to meet these criteria were rejected. It was noted that if a solution set with a given number factors was deemed unreliable, those with greater than this number also failed. As noted in the main article, this approach tended to result in residuals larger than what would be considered optimal according to the error model, which indicates that the number of factors used is insufficient to capture all of the chemical variability within the dataset. However, this analysis also shows that it is not possible to sufficiently constrain a greater number of factors with the data available, so the derived data should be seen as the 'best estimate' solution for the given number of factors with the caveat that additional, weaker factors may also be present that are not resolvable with this factor model. The fractional contributions from unknown factors are considered minor compared to the uncertainty introduced by rotational ambiguity (see main article).

The solution sets are presented according to campaign. The basis for accepting or rejection of the solution sets are shown. The solutions for greater numbers of factors than those of the rejected sets are not shown but in all cases, the diagnostics used as a basis for rejection consistently deteriorated further as the number of factors increased.

All solutions shown are for fpeak = 0. All graphs are shown with the factors unsorted, with factors stacked in ascending order as output by the PMF toolkit. The bootstrapping analysis was performed with 20 iterations and the results grouped according to the uncentred *r* (normalised dot product) between mass spectral profiles.

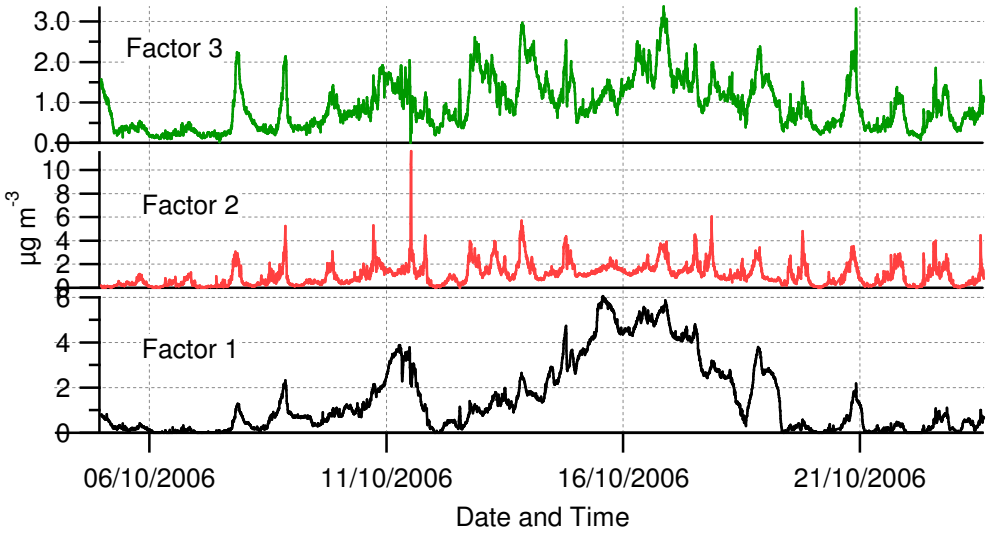
The profiles in terms of mass spectral profiles and time series were averaged and standard deviations derived on a point-by-point basis.

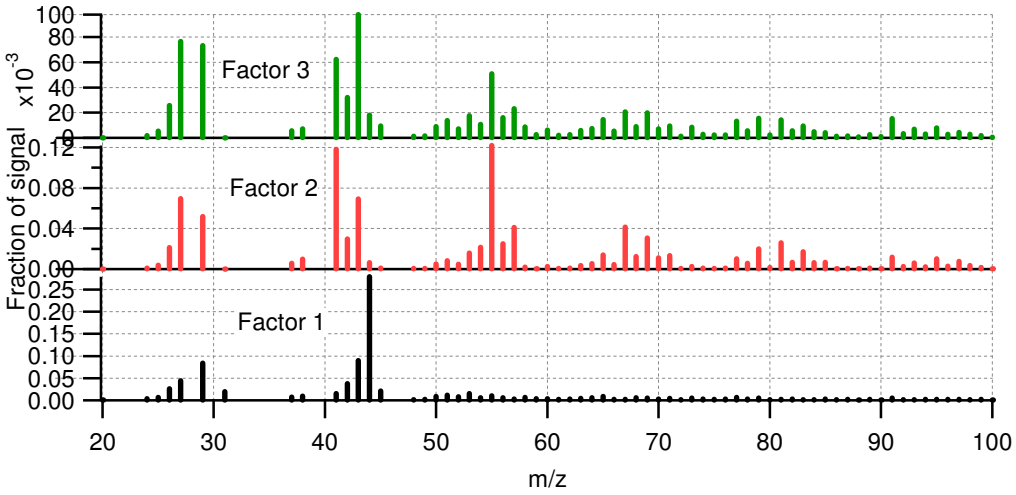
In order to quantitatively evaluate and compare and numerical stability within the outputs of the bootstrapping analysis, suitable metrics must be derived for the variance of solutions within both the time and mass spectral dimensions. The main time series diagnostic reported here is the mean of the standard deviations for each factor, reported as a percentage of the overall mean mass concentration. A mean is not deemed suitable to describe the stability of the mass spectral profiles because while upwards of 150 peaks are used for the analysis, the chemical assignment of factors is typically based on less than 10 peaks. Instead, the averaged mass spectral profile derived from the bootstrapping analysis is inspected and the greatest standard deviation within the spectrum is reported. Because the mass spectral profiles are already normalised, scaling is not needed.

1. REPARTEE 1

1.1.3 factor solution

Unlike the other datasets presented, a 3 factor solution was chosen for REPARTEE 1. This successfully produced reasonably unique time series and mass spectral profiles, as shown:



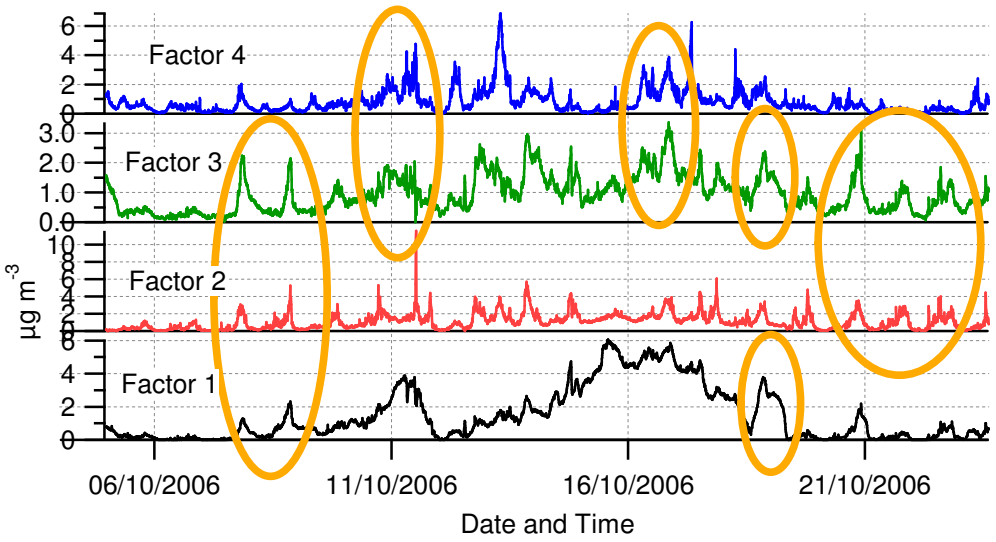


All the time series contain features unique to each factor. Factor 1 can be defined as unique according to its mass spectral profile as it is the only factor with a large peak at *m*/*z* 44. Factors 2 and 3 can be differentiated by the ordering of the peak magnitudes at *m*/*z* 41, 43 and 55.

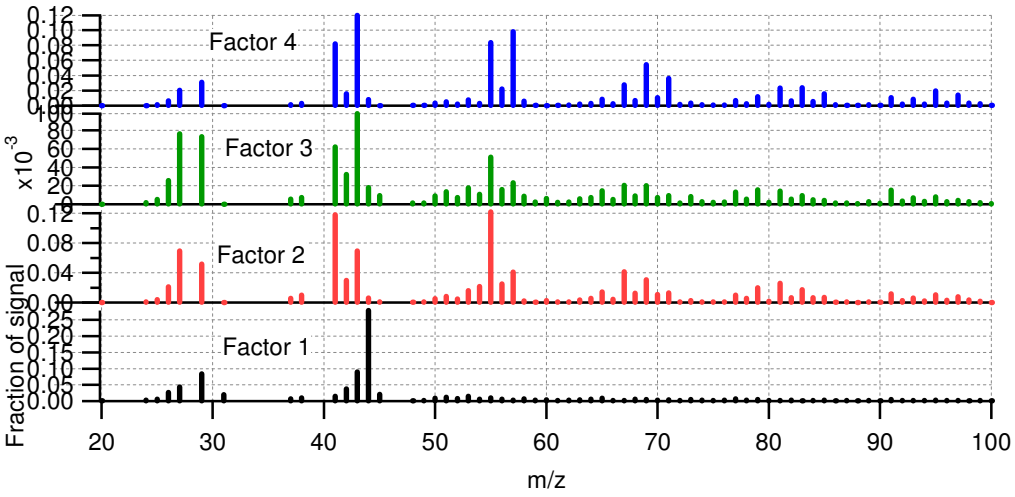
The solutions were found to vary little when the initialisation seed was changed. These factors were also found to be highly robust when subjected to bootstrapping analysis; the mean standard deviations were found to be 1.3, 2.9 and 2.0 % of the mean mass concentration according to the time series and the maxima of the standard deviations associated with the peaks of the three respective factor profiles were 0.25, 0.11 and 0.18 % of the total signal. This solution set was deemed acceptable for further analysis.

1.2.4 factor solution

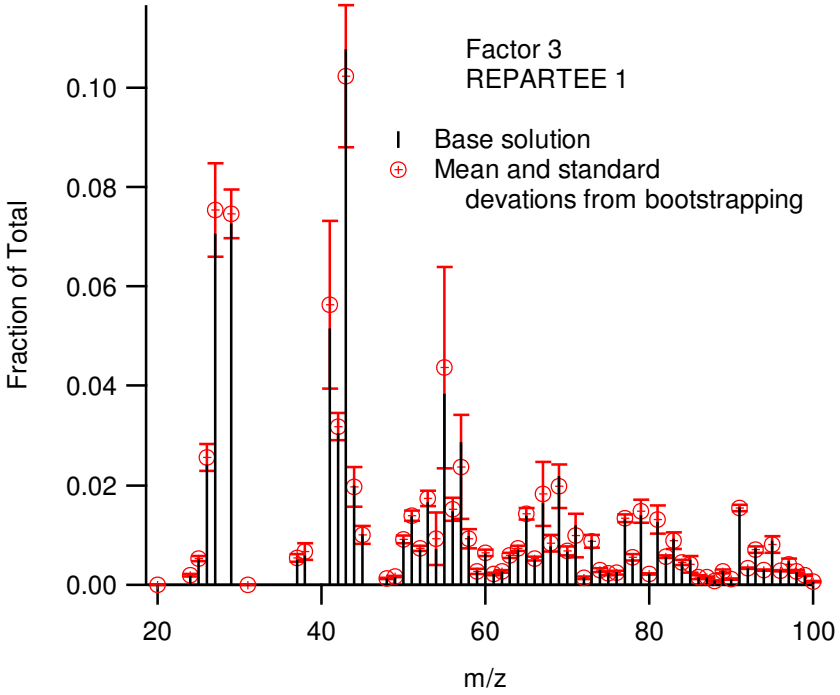
When moving to a 4 factor solution for REPARTEE 1, the Q/Q_{exp} decreased from 10.5 to 8, but similarities between factors begin to become evident. The time series of factor 3 in particular shows few unique features, with all events echoed in other time series (with the exception of the period 12-14 October 2006). Commonalities with other time series are circled for clarity.



In addition, the mass spectral profile of factor 3 shows few distinct features, with the relative sizes of *m/z* 41, 43 and 55 resembling factor 4 and the pattern of peaks above *m/z* showing a strong resemblance to that of factor 2.



This would suggest that there is a significant amount of splitting and merging between factors caused by the introduction of the new factor. The 4 factor solution was found to be not as invariant of initialisation seed; with a seed of 2, a second solution was found with a slightly modified factor 3. The fractional signal at m/z 57 was absent and the time series was reduced by 21 % on average. Bootstrapping analysis also showed a decrease in solution stability; the mean standard deviations within the time series were greatly increased at 6.1, 13, 16 and 15 % of the mean loadings and the peak standard deviations within the mass spectral profiles increasing to 1.6, 0.65, 2.0 and 0.86 % of the total signals. Factor 3 in particular shows a large variation on key peaks, such as *m/z* 55, as can be seen below:

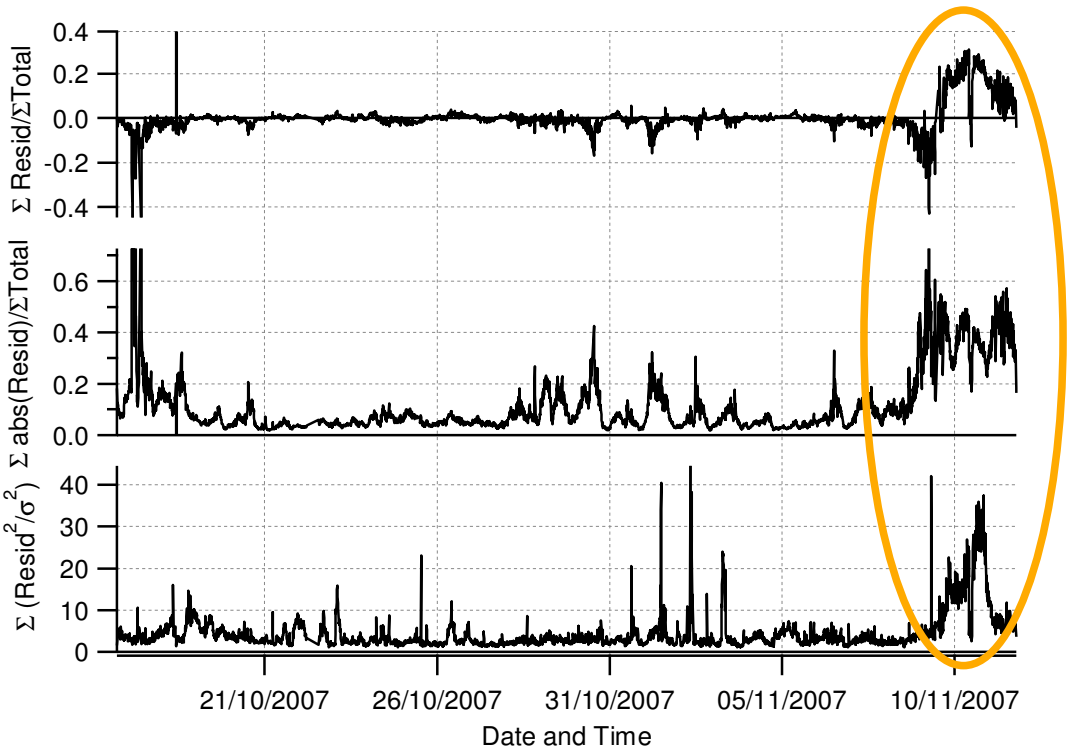


As a result of these tests, this solution set is not deemed to be acceptable, either on the grounds of uniqueness of factors or numerical stability. While the addition of the fourth factor is undoubtedly capturing additional chemical variability not accounted for by the 3 factor solution, the derived data are considered unreliable for the purposes of interpretation of atmospheric composition.

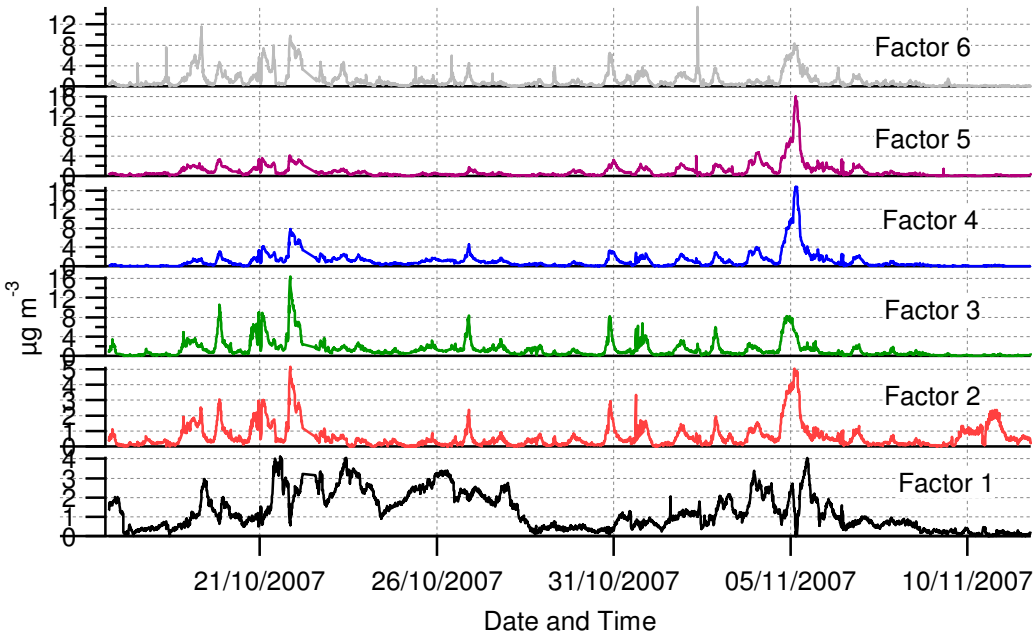
2. REPARTEE 2

2.1. Exclusion of atypical data

During the REPARTEE 2 experiment, an organic event took place starting on 9 November 2007 that caused unusual behaviour during PMF analysis. When a 4 factor solution was generated, the three principal scaled residual diagnostics (summed residual scaled to total loading, summed absolute residual scaled to total loading and summed square of the residual scaled to total modelled variance) all showed this event to be poorly fitted in comparison to the rest of the experiment:



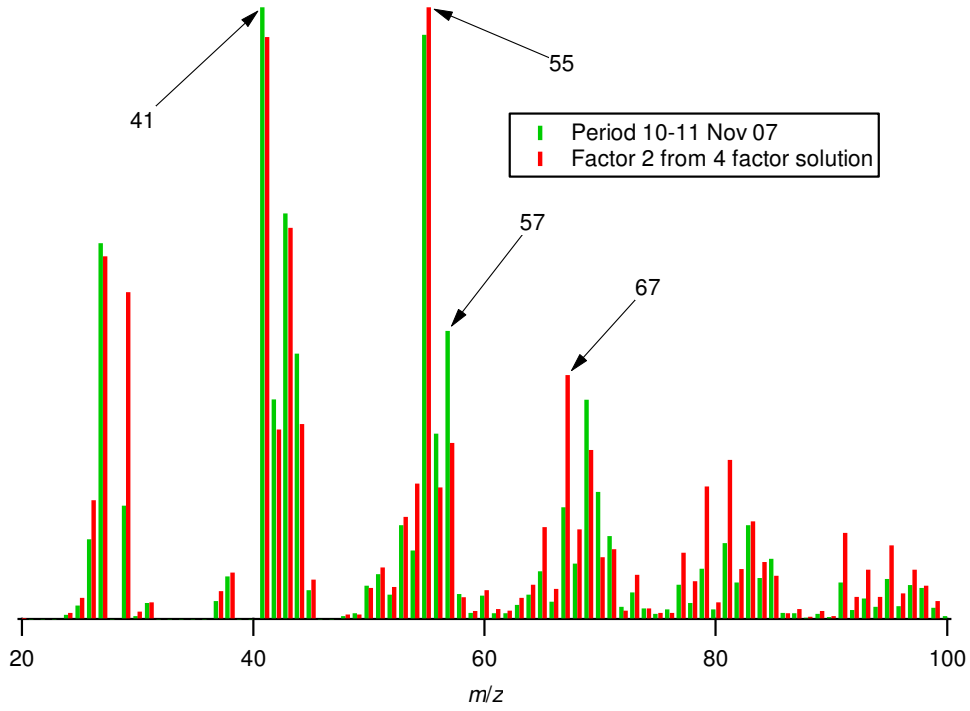
It was found that this anomaly in the residuals would exist for any solution set of 5 factors or less. When a 6 factor solution was generated, this event was represented by a single factor (factor 2 in this set):



Note that outside of the event, factor 2 is almost identical to factor 3. This indicates that the organic activity during the 9-10 November period is too different chemically to allow the entire measurement period to be described by less than 6 factors. Because solution sets of greater than 4 factors are considered unreliable for this experiment

(see below), it was decided that this event should be removed from the dataset. This does not necessarily mean it is 'bad' data as such, but that the organic activity within this period is not representative of the majority of the measurement period.

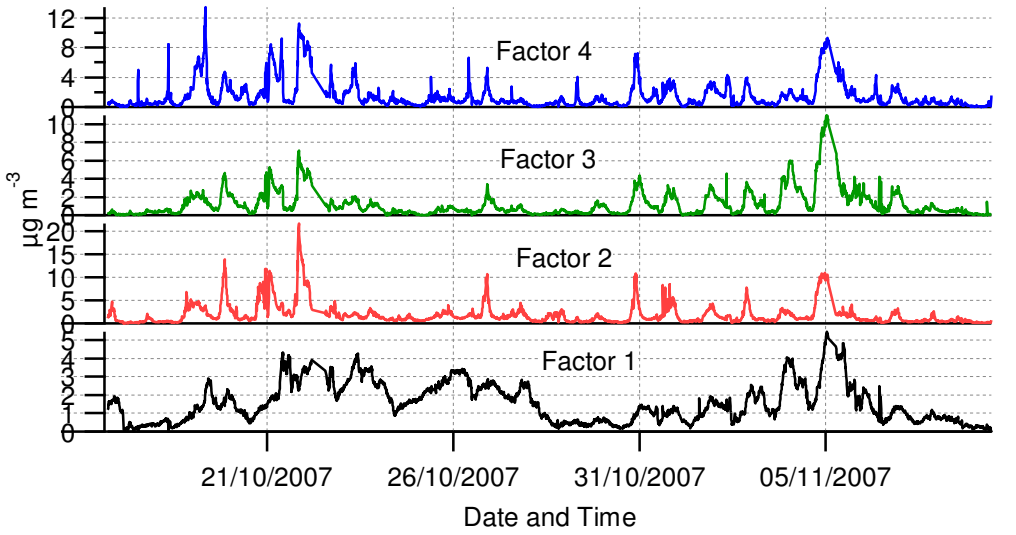
An inspection of the problematic period shows why its inclusion causes problems. Below is a comparison between the averaged mass spectrum during this period and factor 2 from the 4 factor solution:

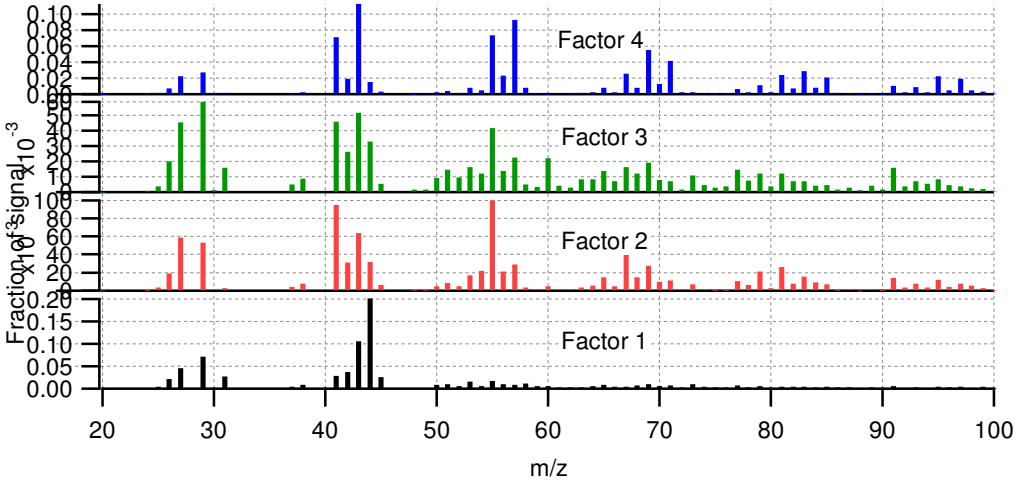


The period's mass spectrum bears a strong similarity with the factor at *m/z* 41, 43 and 55, however certain key peaks differ, such as *m/z* 57, 67, 79 and 81. These features are not adequately described by any of the factors produced from solution sets of 5 factors or less. The exact chemical nature of the organic aerosol sampled during this period is not known and will be the subject of further investigation.

2.2.4 factor solution

The 4 factor solution for REPARTEE 2 produced factors that, at first glance, showed a certain degree of similarity between time series, particularly between factors 2, 3 and 4. However, a much better separation was found between the mass spectral profiles:



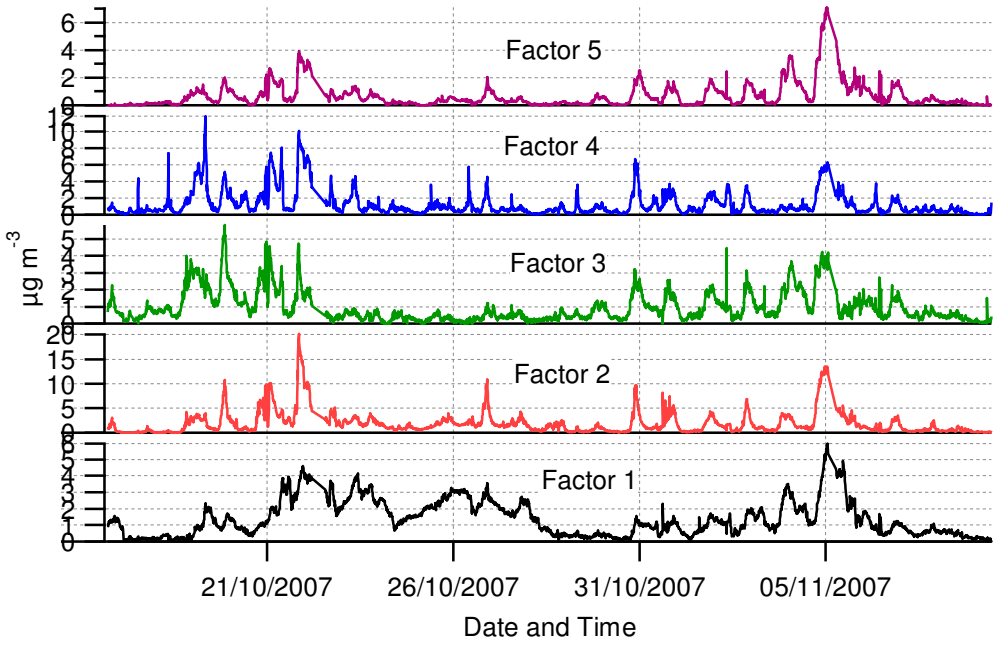


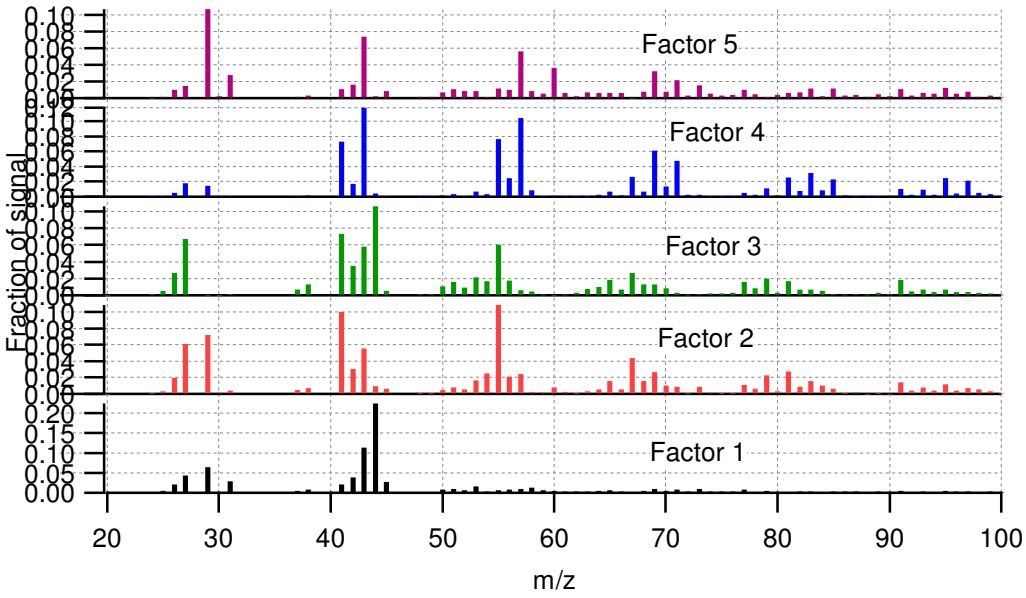
Unlike the 4 factor solution derived for REPARTEE 1, good separation was found between the mass spectral profiles. Factor 1 could be identified by its peak at m/z 44, while factors 2 and 4 are distinct by the ordering of the peaks at m/z 41, 43, 55 and 57. Factor 3 is unique in that it has a very distinct peak at m/z 60, which is not present in any of the other factor profiles. This solution set also fared well when subjected to bootstrapping analysis. The standard deviations of the time series for the four factors were 2.3, 2.7, 2.6 and 4.0 % and the maximum peak standard deviation of the mass spectral profiles 0.27, 0.40, 0.38 and 0.24 %, indicating that the solutions were numerically stable. Crucially, the defining m/z 60 peak of factor 3 was found to be very stable at 2.2 % of the total signal, with the standard deviation of the bootstrapping being only 0.066 % and the mean differing from the base solution by only 0.003 %.

For these reasons, the 4 factor solution set was deemed suitable for further analysis, in spite of the similarities in the some of the features of the time series. In some cases, real meteorological phenomena may contribute to these apparent similarities (such as variations in the mixing rate for primary emissions) so this alone was not deemed sufficient grounds for rejection.

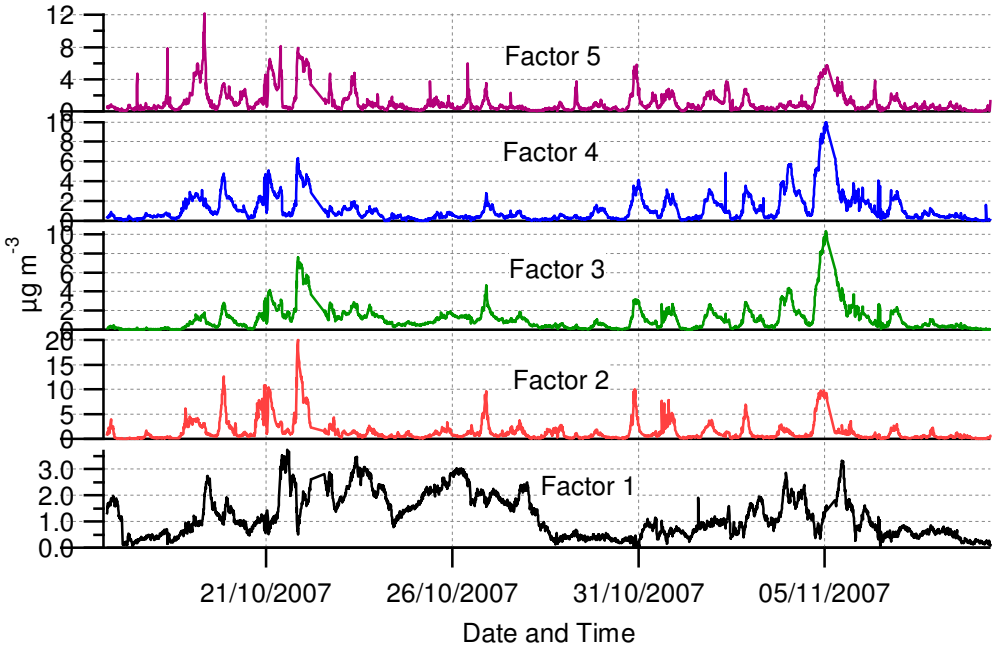
2.3.5 factor solution

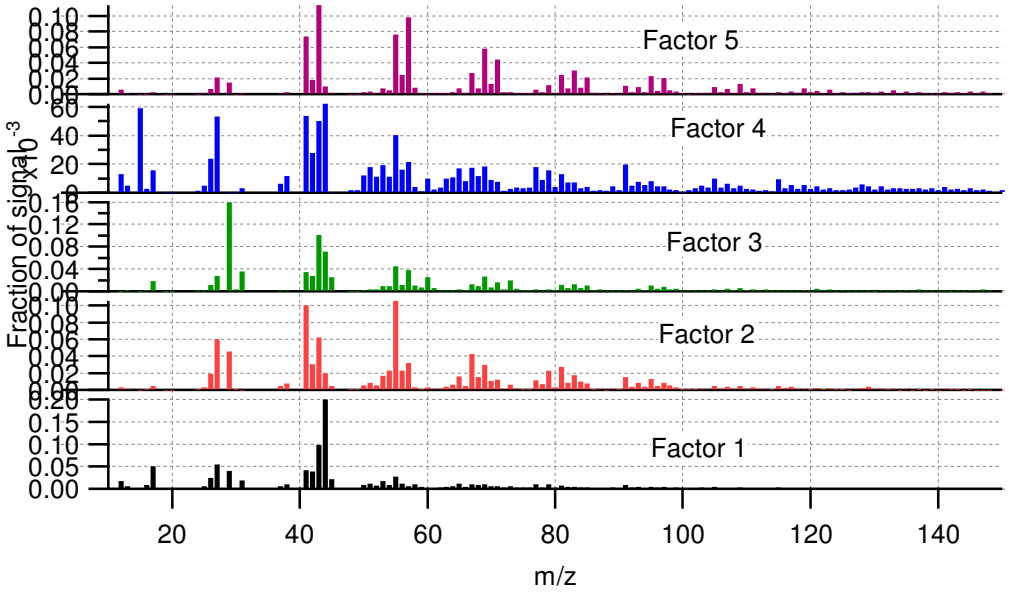
Unlike previous solutions presented, the 5 factor solution for REPARTEE 2 showed considerable dependency on the initialisation seed used, producing two general types of solution depending on the exact seed used. The first solution (using a seed of 0) produced the following profile:





This solution caused the distinct m/z 60 signal to be present in a much larger fraction of factor 5 with a very distinctive mass spectral pattern. Factors 1, 2 and 4 remained distinguishable, however factor 3 began to resemble a hybrid of factors 1 and 2. The second solution (achieved with a seed of 1) produced the following profile:





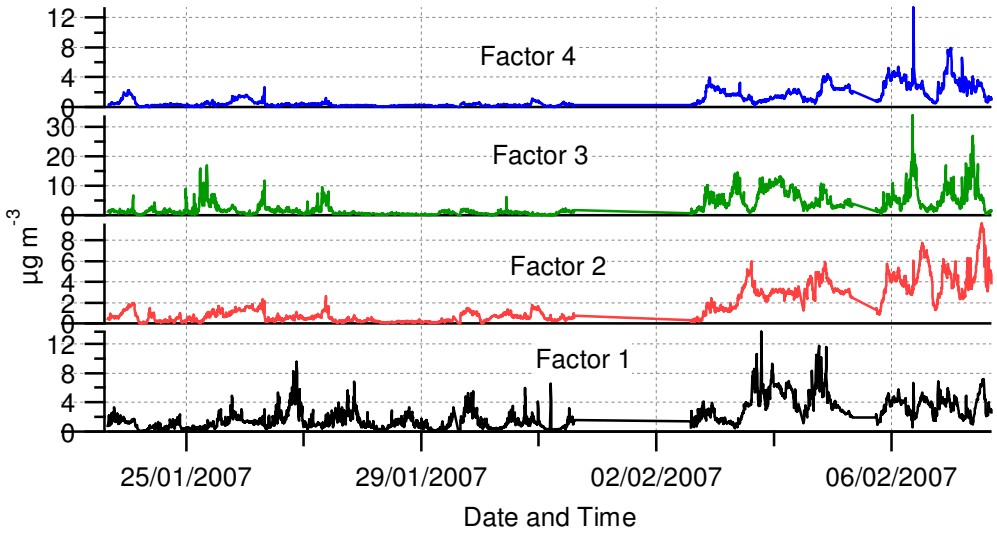
Factors 1, 2 and 5 now resemble factors 1, 2 and 4 from the 4 factor solution. The distinctive m/z 60 signal is now split between factors 3 and 4 and neither species now possess any truly unique mass spectral species. This fact and the close similarity between the time series of these factors (r = 0.90) would imply that factor 3 from the 4 factor solution had split into what was now factors 3 and 4 in the 5 factor solution. Bootstrapping analysis was found to be problematic with this solution set, with the algorithm consistently failing to converge (according to the default criteria) with the resampled datasets. This would indicate a lack of robustness in the solutions.

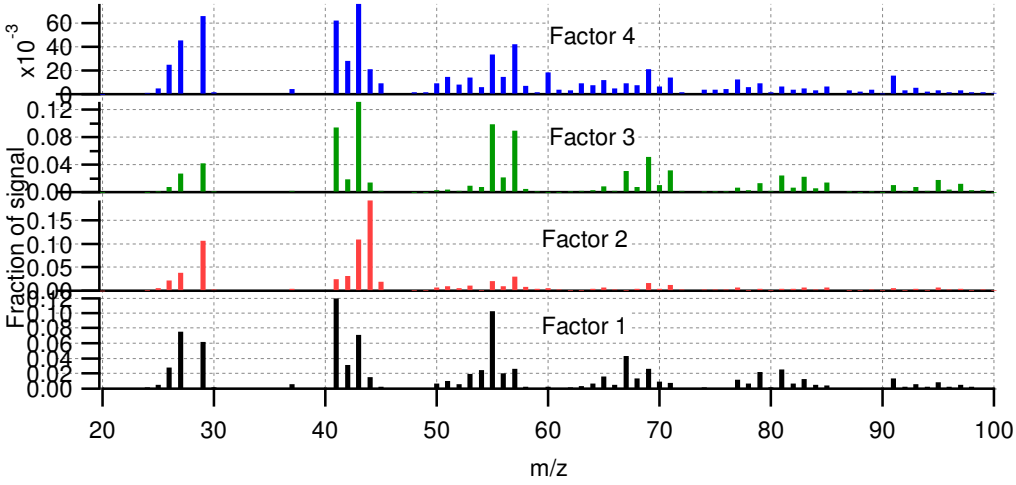
Based on the nature of the splitting, the addition of a fifth factor is clearly reflecting chemical variability within factor 3 as identified in the 4 factor solution. While this is real variation, the lack of numerical stability shows that this is not being reliably captured with the 5 factor solution set. For this reason, this solution set was not deemed usable for further analysis.

3. Manchester

3.1.4 factor solution

The 4 factor solution from the Manchester dataset produced factors distinct in their time series and mass spectral profiles:

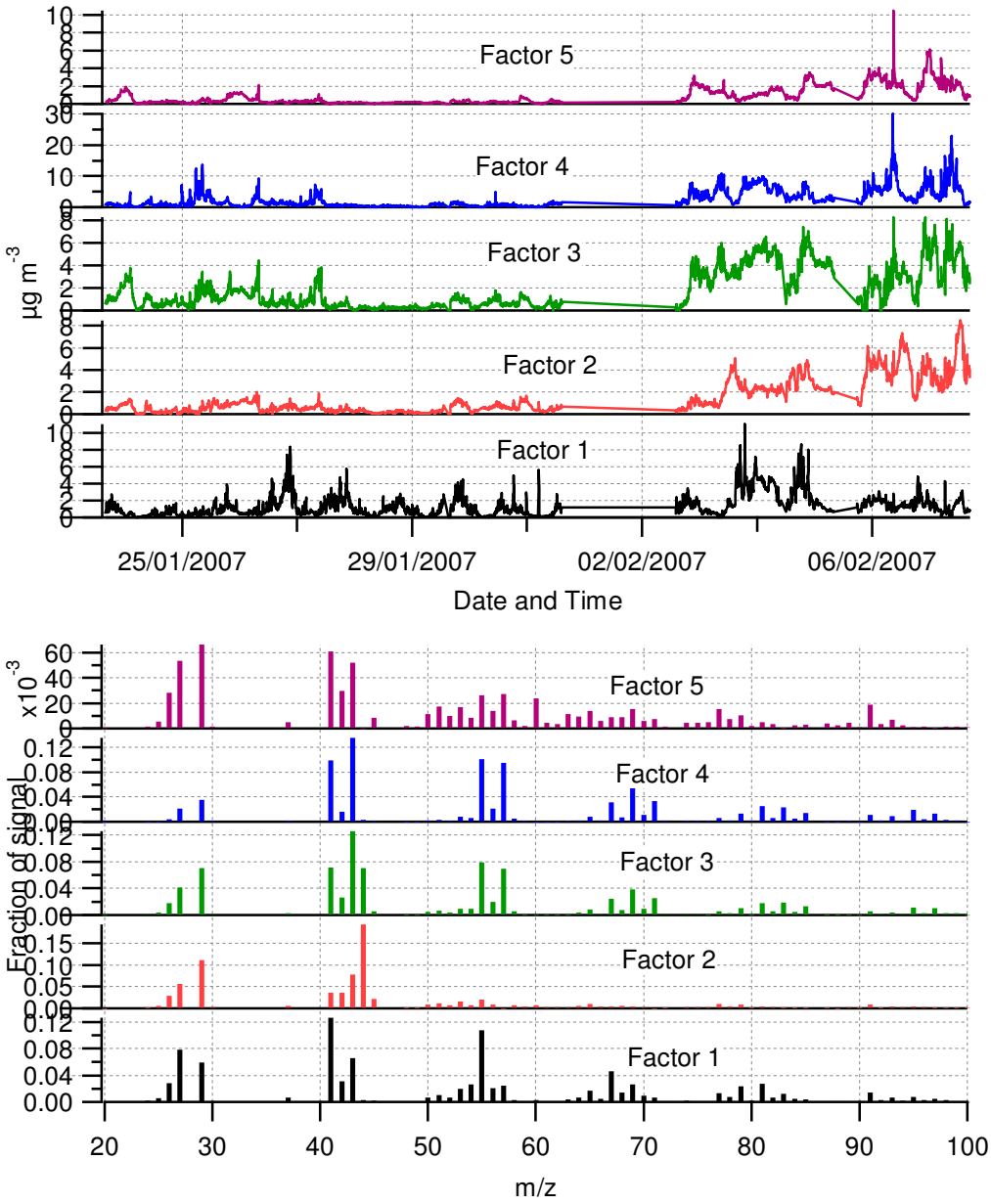




Much like the 4 factor solution from REPARTEE 2, the factors could be distinguished according to unique features at m/z 41, 43, 44, 55, 57 and 60. When bootstrapping was performed, the mean time series standard deviations derived were 5.4, 3.6, 2.6 and 5.6 % and the maximum peak standard deviations of the mass spectra 0.43, 0.66, 0.053 and 0.92 %

3.2.5 factor solution

Much like REPARTEE 2, the 5 factor Manchester solution set begins to show evidence of similarities between both the time series and mass spectra:



In particular, the mass spectral profiles of factors 3 and 4 are almost identical, with the exception of the contribution from m/z 44, and their time series show many similarities in their features, which would indicate a degree of factor splitting.

The solutions were also found not to be numerically stable. A strong dependence on the initialisation seed was noted, with a large amount of signal redistribution between factors 1 to 4 (factor 5 remained largely stable). In some solutions, factors 3 and 4 became almost identical, with the m/z 44 signal becoming diminished. For instance, a seed of 1 yielded an uncentred *r* value between the two factors of 0.99 (compared to 0.91 for seed = 0 above). When bootstrapping analysis was performed, 11 out of the 20 solutions could not be accurately classified according to the 5 factors of the base

case. Of the ones that could be classified, the mean relative standard deviations of the 5 time series were increased at 6.8, 3.3, 8.8, 11 and 8.1 %. The maximum mass spectral profile standard deviations were slightly elevated at 0.43, 0.47, 0.30, 0.47 and 1.2 %. Given the evident splitting and strong dependency on initialisation seed, the 5 factor solution set was deemed not suitable for further analysis.

4. Summary

The solution sets for the three experiments can be summarised with the following diagnostics:

Solution set	Q/Q_{exp}	Seed dependence?	SD_{TS}/TS (%)	Max(SD_{MS}) (%)
REPARTEE 1 (3 factors)	10.50	No	1.3, 2.9, 2.0	0.25, 0.11, 0.18
REPARTEE 1 (4 factors)	8.02	Yes	6.1, 13, 16, 15	1.6, 0.65, 2.0, 0.86
REPARTEE 2 (4 factors)	3.90	No	2.3, 2.7, 2.6, 4.0	0.27, 0.40, 0.38, 0.24
REPARTEE 2 (5 factors)	3.18	Yes	Did not converge	Did not converge
Manchester (4 factors)	16.70	No	5.4, 3.6, 2.6, 5.6	0.43, 0.66, 0.053, 0.92
Manchester (5 factors)	14.88	Yes	6.8, 3.3, 8.8, 11, 8.1	0.43, 0.47, 0.30, 0.47, 1.2

The solution sets were chosen as follows: 3 factors for REPARTEE 1; 4 factors for REPARTEE 2; and 4 factors for Manchester. In each case, the transition to a greater number of factors was accompanied by the following:

I. An increase in qualitative similarity between the time series of some of the factors

II. A reduction in uniqueness of mass spectral profiles in terms of key marker peaks

III. A loss of invariance with respect to initialisation seed

IV. An increase in the mean standard deviation of the bootstrapped time series, with at least one factor reporting a value greater than 10 % with respect to the overall mean

V. An increase in the maximum standard deviation of the bootstrapped mass spectral profile, with at least one factor reporting a value greater than 1 % of the total signal.

While the first two of these observations may be considered subjective, the latter three provide an objective and consistent means of selecting the appropriate number of factors for each experiment. When higher order solution sets were examined, it was found that these five observations continued in their trends. Given that the solution sets chosen can be considered defendable, lower order solutions sets have been excluded from further analysis, as these will result in less accurate assessments of the aerosol composition.

Reference:

Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., and Jimenez, J. L.: Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data, Atmos. Chem. Phys., 9, 2891-2918, 2009.