ANSTO/E-756

Australian Government

Nuclear-based science benefiting all Australians

Receptor Modelling with PMF2 and ME2 using aerosol data from Hong Kong

by

J.Crawford

D. Cohen

L. Dyer

W. Zahorowski

Prepared within the Institute for Nuclear Geophysiology
Australian Nuclear Science and Technology Organisation

May 2005

AUSTRALIAN NUCLEAR SCIENCE AND TECHNOLOGY ORGANISATION

LUCAS HEIGHTS SCIENCE AND TECHNOLOGY CENTRE

Receptor Modelling with PMF2 and ME2 using aerosol data from Hong Kong

by

J. Crawford
D.D. Cohen
L. Dyer
W. Zahorowski

ABSTRACT

A number of techniques, such as principal component analysis and factor analysis, have been used in receptor modelling where measured aerosol composition at the sampling site are analysed in order to determine the likely source contributions. In this study factor analysis with non-negative factor elements has been carried out using two techniques as implemented in the PMF2 and ME2 computer codes. The various analysis techniques provided by the two programs are illustrated using measured data at Hong Kong as a case study, which covers a period of three years (2001 to 2003). Both analysis techniques resulted in similar results which are presented in this report. Data bootstrapping was also carried out as an additional check on the quality of the results.

ISSN 1030-7745
ISBN 1 920791 043

Table of Contents

List of Figures

List of Tables

1. Introduction

Receptor modelling [1], as applied to air quality studies, apportions measured aerosols of chemical concentrations at the sampling site to their possible sources [2]. If the composition profiles of all sources contributing to the measured concentrations at a sampling site are known, the mass balance model becomes a multilinear regression problem [3]. However, if a series of samples has been analysed without any substantial information being available on the sources, multivariate data analysis methods can be utilised. These methods are based on the analysis of the correlation between measured concentrations of chemical species, assuming that highly correlated compounds are emitted from the same sources.

Factor analysis techniques are multivariate data analysis methods that are commonly used in environmental studies to estimate the number and composition of the sources, and their contributions to the samples taken at receptors. The most common form of factor analysis is Principal Component Analysis (PCA) [4]. PCA extracts the principal components explaining the majority of the variance in the data matrix that is then qualitatively interpreted as possible sources. PCA suffers from a number of drawbacks, e.g. the factors are not always physically explainable and it cannot handle missing and below-detection-limit data.

To overcome some of the limitations of PCA, new approaches to the solution of factor analysis problems have been developed. In particular, two implementations of Positive Matrix Factorisation have been used in this study, PMF2 and ME2 (Multilinear Engine), to assess and compare their usefulness in the identification of origins of atmospheric pollution.

2. The Factor Analysis Model

A 2-way, bilinear factor analysis problem can be represented as:

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} \qquad (1)$$

or

$$x_{i,j} = \sum_{n=1}^{p} g_{i,n} f_{n,j} + e_{i,j} \qquad (2)$$

where the matrix **X** contains the measured quantities at the receptor, i.e. $x_{i,j}$ represents the concentration of chemical species *j* in the *i*th sample. Matrices **G** and **F** are factor matrices to be determined and **E** is the matrix of residuals. If **n** observations are available, each containing **m** chemical species and if a **p**-factor model is being considered, **G** is a **n** x **p** matrix of source contributions to the samples, describing the temporal variation of the sources. The matrix **F** is a **p** by **m** matrix of source chemical compositions, or source profiles.

Equation (1) represents a 2-way problem in that the matrix **X** is two dimensional, i.e temporal samples form rows of the matrix with the concentration of each chemical species stored in the columns. For situations where the measured data can be classified into separate groups, e.g. seasonal variation of measurements, the matrix **X** could be three-dimensional, referred to as a 3-way problem. In addition, equation (1)

is a bilinear problem in that the values are expressed as a sum of products of the two elements of **G** and **F**. On the other hand, the well known PARAFEC [6] model is a 3-way trilinear model:

$$x_{i,j,k} = \sum_{p=1}^{P} g_{i,p} b_{j,p} f_{k,p} + e_{i,j} \qquad (3)$$

where the 3-way measured data is explained by the product of 3 factor matrices, **G**, **B** and **F**.

2.1 *Positive Matrix Factorisation*

The objective of Positive Matrix Factorisation (PMF) analysis, [4], is to minimise the objective or penalty function Q, defined as:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{e_{i,j}^2}{s_{i,j}^2} \qquad (4)$$

where $e_{i,j}$ are the error terms specified in equation (2) and $s_{i,j}$ is a specified standard deviation of each of the data values. The minimisation of the sum of the square of the errors between the measured data and the model is carried out under the constraints that the factor elements remain non-negative.

2.2 *PMF2*

PMF2 [7, 8] is a program, which implements a weighted least squares approach to perform positive matrix factorisation of measured data. PMF2 solves the 2-way bilinear model, while a second program, PMF3, has also been developed for the solution of 3-way, trilinear models. The programs provide a number of options to control the solutions process which are specific to factor analysis.

2.3 *ME2*

The drawback of the PMF2 and PMF3 suite of programs is that the form of the model is predetermined, i.e. bilinear models are solved by PMF2 and trilinear by PMF3. ME2 [9] is a more recent technique for fitting multilinear and quasi-multilinear mathematical expressions or models to two-, three- and many-dimensional data. The multilinear equations and the required information are presented to the program in a user generated input file from which the program builds and solves the model. This gives the user a greater freedom in defining the form of the model and the imposition of constraints on the solution.

3. Case Study

The Hong Kong study site, Figure 1, is located on Cape D'Aguilar at the south-eastern end of Hong Kong Island at 22.22^{o}N, 114.25^{o}E atop a 60m cliff facing the South China Sea. The population density of Cape D'Aguilar is relatively low and the nearest major urban/industrial town of Chai Wan is 10km away. A detailed description of the site and sampling techniques is given in [10]. In summary, 24-hour averaged $PM_{2.5}$ samples are taken twice a week, on Wednesday and Sunday. In this study data from 2001 to 2003 was used. The total aerosol mass sampled is presented in Figure 2, where the vertical axis represents mass concentrations in ng/m^3 with sampling time on the horizontal axis.

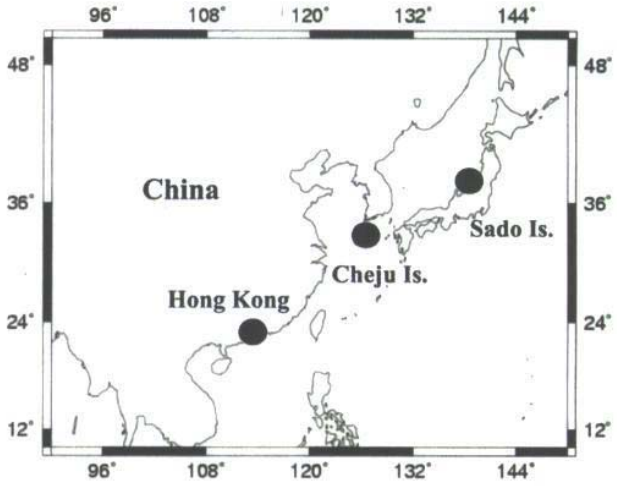

Figure 1: Hong Kong Study Site

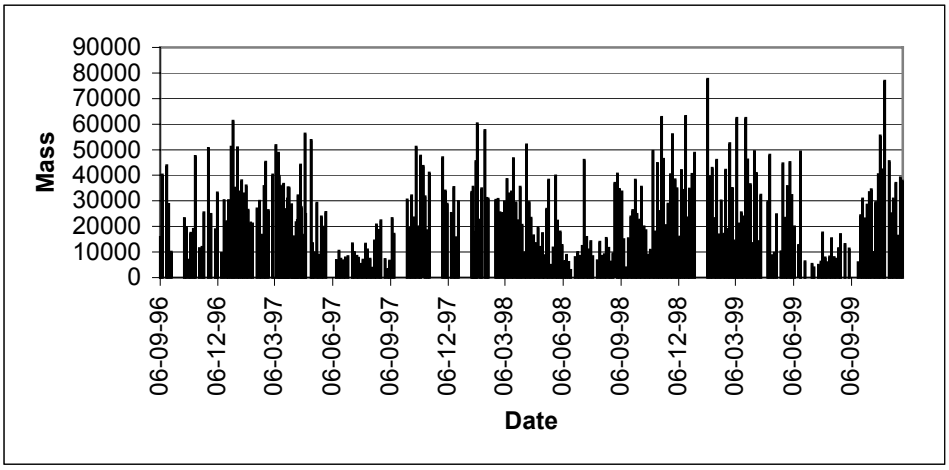

Figure 2: Daily averaged total mass fractions.

Figure 2 clearly shows that there were selected periods of the year when the mass concentrations were particularly high. These periods have been identified in previous studies also and were related to significant international events such as dust emissions from desert regions in China as well as favourable meteorological conditions bringing pollution from eastern China into Hong Kong [28]. Aside from this, a seasonal variation is observed. Chemical compositions of atmospheric aerosols in Hong Kong has been studied elsewhere, one such study is presented in [18] - most of the chemical species show seasonal variations, which reflect the weather conditions: i.e. low concentrations in the rainy season of summer and high concentrations for the rest of the year.

The samples have been analysed using accelerator based ion beam analysis (IBA) techniques, resulting in the identification of 21 elements; H, Na, Al, Si, P, S, Cl, K, Ca, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Br, Pb, Black Carbon. The analysis techniques

have been presented in [27,28]. Principal Component Analysis has previously been carried out on a subset of this data and possible source factors identified [10]. In the current study positive matrix factorisation is implemented using PMF2 and ME2.

4. Data Analysis with PMF2

The user controls the working of PMF2 via a fixed format input file, which offers a number of options including such control as to the number of iterations to be carried out, convergence criteria, etc. Some of these options are discussed in the subsequent sections, but for a full description see the user guide [8].

In undertaking a factor analysis study, uncertainties arise from a number of sources:

- Choice of the number of factors, see section 4.5 for a number of techniques that can be employed to determine the number of factors.
- Having determined the number of factors, multiple solutions may exist:
  - Where the optimisation problem to be solved may contain a number of local minima. Section 4.6 describes the use of different initial guesses to search the solution space for the best solution.
  - A number of different solutions may provide the same Q value in the objective function. This is referred to as rotational freedom and is described in section 4.7.

The following sections use the Hong Kong data to illustrate the use of PMF2 in carrying out factor analysis.

4.1 *PMF2 Model Formulation*

PMF2 solves the problem specified in equation (1) under non-negativity constraints, i.e. only zero or positive values of the factor matrices are considered. This implies that mathematically the problem is redefined to include the constraints. One might better understand the internal workings of PMF2 by looking at the simplified recasting of the model [11]:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{e_{i,j}^2}{s_{i,j}^2} - \alpha \sum_{i=1}^{m} \sum_{j=1}^{p} \log g_{i,h} - \beta \sum_{i=1}^{m} \sum_{j=1}^{p} \log f_{i,h} + \gamma \sum_{i=1}^{m} \sum_{j=1}^{p} g_{i,h}^2 + \delta \sum_{i=1}^{m} \sum_{j=1}^{p} f_{i,h}^2 \quad (5)$$

The coefficients α and β control the strength of two penalty terms which prevent the factors **G** and **F** from becoming negative. Whenever an element in **G** or **F** is about to become negative, the corresponding term in the objective function (i.e. $-\log(f_{i,j})$) approaches plus infinity and causes an increase of Q. By controlling the strength of the penalty terms, the user may allow arbitrarily close approaches to zero but the exact value of zero is never reached. The penalty can be controlled by the *Fkey* input matrix to PMF2, where elements of the factors are selectively chosen to be pulled down to zero. The user also has control over the coefficients through the *lims* parameter in the input file [8], which controls how close to zero the constrained components of the solution may get. Smaller values result in values closer to zero. Similarly, γ and δ control two regularisation terms which remove the rotational indeterminacy and control the scaling of the left and right factors.

Experimenting with the regularisation parameters is detailed in section 4.7.

4.2 *Standard deviation of observations*

The values, $s_{i,j}$, for the standard deviations of the observations are specified by the user. Generally one should attempt to specify the standard deviation values so that they indicate the agreement which is expected between the model and the data array. A number of alternative forms for specifying the standard deviations are available in PMF2, see [7]. In environmental work, different columns of the array **X** typically contain concentrations of different elements, having different error characteristics. In this study two forms of the standard deviation were considered. The first form is that when the error mode flag, EM is set to -14, which results in a standard deviation of the form:

$$s_{i,j} = c1_{i,j} + c2_{i,j}\sqrt{\max\left(\left|x_{i,j}\right|,\left|y_{i,j}\right|\right)} + c3_{i,j}\max\left(\left|x_{i,j}\right|,\left|y_{i,j}\right|\right) \qquad (6a)$$

where, the values of $c1_{i,j}$, $c2_{i,j}$ and $c3_{i,j}$ are specified by the user and the matrix **Y** is the model fit, i.e. **Y** = **GF**. Thus the weight will depend both on the measured value and on the fitted value.

The second form is when the error mode flag, EM is set to –12, which results in a standard deviation of the form:

$$s_{i,j} = c1_{i,j} + c2_{i,j}\sqrt{\left|x_{i,j}\right|} + c3_{i,j}\left|x_{i,j}\right| \qquad (6b)$$

in which case the weight depends only on the measured values. In our case study, some differences were observed between the results obtained when using the two forms of error modes. The results are presented in section 4.10.

4.3 *Outliers and the robust mode*

Unless the errors in the data are approximately normally distributed and there are no outliers, the robust mode needs to be used. In this mode, if the residual exceeds a specified amount, α - the outlier threshold distance, times the specified standard deviation of the corresponding observation, the data point is processed as an outlier, i.e. if

$$\frac{\left|x_{i,j} - y_{i,j}\right|}{s_{i,j}} > \alpha \qquad (7)$$

the influence on the estimation and the value of Q would be the same as if:

$$\frac{\left|x_{i,j} - y_{i,j}\right|}{s_{i,j}} = \alpha \qquad (8)$$

In this study α=4 is used, which is recommended by the author of the code based on his experience [8].

4.4 *Solution and Convergence of the algorithm*

The problem of determining the best solution for **F** and **G** in equation (1), becomes that of finding the values of **F** and **G** that minimises Q in equation (5). An iterative algorithm is implemented which progresses from an initial guess for the elements of the factors **G** and **F**, updating them systematically (using the Gauss-Newton method), until convergence is achieved. In the PMF2 implementation random numbers can be

used to initialise the matrices or those from a previous calculation can be input into the current calculation.

Convergence criterions are specified by the user, through the *lims* parameters, which determine that convergence will be reached when the relative change in Q is below the specified value for a specified number of iterations in succession. A maximum number of iterations is also specified, at which time the program will terminate if the convergence criterion has not been met.

Based on the adequacy of the selected model to explain the observations, the expected value of Q at convergence is approximately equal to the number of data values minus the number of essential free parameters fitted to the data [12]. This value is often called the degrees of freedom:

$$E(Q) = nm - p(n+m) \qquad (9)$$

In the presence of outliers it may be difficult to know if the observed value of Q is as expected or too large. It may be more helpful to investigate the distribution of the scaled residuals ($e_{i,j}/s_{i,j}$). If the vast majority of these values are <2.0, then an increase in the Q value is probably due to outliers.

In the current study (**n**=)266 data measurements are available for (**m**=)21 chemical species. When 8 factors are used, in the case of the accepted solution, the expected value of Q according to equation (9) is 3290. On convergence a value of 3393 for Q was obtained, which is close to the expected value. The scaled residuals are randomly distributed and only a small number has an absolute value greater than 2.0 but all scaled residuals have an absolute value less than 2.2.

4.5 *Estimating the number of sources/factors*

Various methods have been proposed for the estimation of the number of sources that contributed to the measurement at the detector site [1, 5], e.g. the number of eigenvalues of the correlation matrix, and other statistical methods, such as the NUMFACT algorithm [26]. However, it appears that no mathematical criteria are able to predict the "correct" number of factors [8]. Rather, the question is which number of factors gives the most useful solution.

In PMF, the choice of the number of factors is a compromise [5]. Using too few factors will combine sources of different nature together. Using too many factors will make a real factor further dissociated into two or more non-existing sources. In this study we have carried out factor analysis with varying the number of factors followed by inspection of the error of the fit and the form of the resulting factors. A detailed analysis of the residual matrix is presented in [5] where techniques that can be used to give a bound on the number of factors is presented. This analysis has been carried out for this case study and the results are presented in section 4.5.1. However, in this case this analysis did not produce conclusive results.

In this case study the Q value results in a drop when the number of factors is increased (see Table 1 where the Q value at convergence is given for 7, 8 and 9 factors). For 9 factors the expected value of Q according to equation (9) is 3003, whereas the fit resulted in a smaller than expected Q value, indicating that a good fit is being obtained as a result of the freedom due to the larger number of factors used.

The expected Q value for 7 factors is 3577, whereas the actual value obtained in our fit is 4525, indicating a poor choice for the number of factors. The best fit to the expected and actual value of Q is when 8 factors are used, i.e. 3393 achieved as compared to 3290 as expected and thus the ratio is closest to 1.0.

Table 1 : PMF2: Q values for different number of factors.

Factor	Q	Expected Q	Ratio
7	4525	3577	1.26
8	3393	3290	1.03
9	2613	3003	0.87

An additional analysis of matrix *rotmat* is presented in [5] (and detailed in the following section) in accessing the number of factors. By plotting the largest value in *rotmat* against the number of factors, one can see a significant increase in the largest value of *rotmat* as the number of factors is increased above the optimum.

4.5.1 Determining the number of factors by analysis of matrix *rotmat* and residual matrix

A matrix *rotmat* is generated by PMF2 [12, 8], which is the basis for estimating the uniqueness of the computed factor values. *rotmat* is a **p** x **p** matrix of standard deviations of rotational coefficients, i.e.

$$rotmat_{i,j} = stddev\left(t_{i,j}\right) \qquad (11)$$

Roughly, each value of *rotmat* indicates how large a rotation can be carried out without increasing the penalty function (Q), by more than one unit. It is 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 chosen. Choosing the largest element in *rotmat* can show the worst case in rotational freedom. In Figure 3 we see some increase when 11 factors are used and a significant increase with 12 factors, thus 11 factors or less should be used.

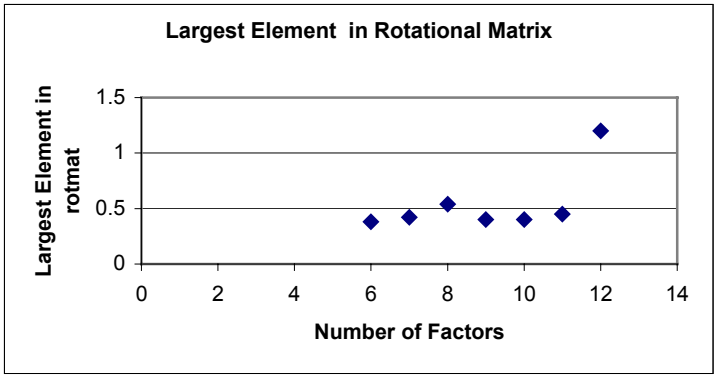

Figure 3: Largest element in the rotational matrix.

In addition to the rotational matrix (*rotmat*) an analysis of the information from the scaled residual matrix (**R**) [5] is used to define the number of factors and reduce the ambiguity due to manual judgment. Each element in matrix **R** is

$$r_{i,j} = \frac{e_{i,j}}{s_{i,j}}$$

Each column in **R** represents the quality of the fitting of each chemical species to the product of **GF**. Generally speaking the more the factors, the better the fit. For each specific number of factors, two parameters are obtained from **R**: IM, the maximum individual column mean, and IS, the maximum individual column standard deviation, where

$$\text{IM} = \max\left(\frac{1}{n}\sum_{i=1}^{n} r_{i,j}\right) \text{ for j = 1, … ,m}$$

and

$$\text{IS} = \max\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(r_{i,j} - \bar{r}_j\right)^2}\right) \text{ for j = 1, … ,m}$$

IM and IS serve as indicators to identify the chemical species having the least fit and the most imprecise fit, respectively. They can also be used to identify the number of factors. When the number of factors increases to a critical value, IM and IS will experience a drastic drop.

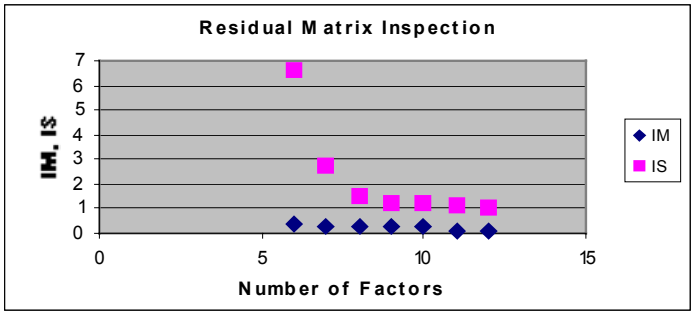

Figure 4: Maximum individual column mean (IM) and standard deviation (IS).

Inspecting Figure 4, shows that from 8 to 10 factors should be used, where IM and IS have similar values. An analysis of 9 factors was carried out, however when examining the **G** matrix, a correlation between the columns was found to exist (see Figure 5), where two columns were correlated. A similar analysis was carried out with 8 factors in which case no correlations were found.

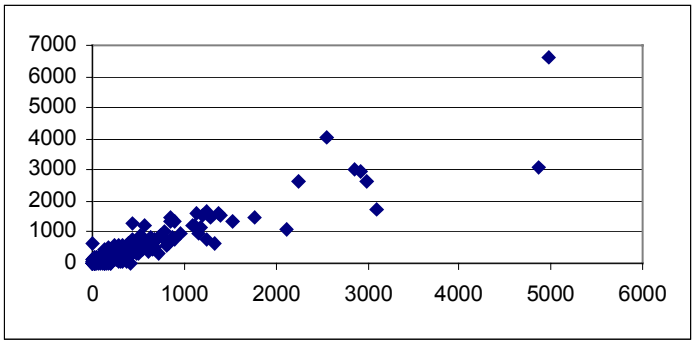

Figure 5: Correlation of two columns of matrix G.

4.6 *Multiple Solutions*

The objective function Q, may posses several local minima, leading to multiple solutions of the factorisation problem. The presence of multiple solutions may be investigated so that one re-runs the algorithm using several different sets of pseudorandom values as the initial values for the factor matrices **G** and **F** (controlled through the *seed* value in the input file to PMF2). The advice given by the author of PMF2 [11], is to regard all the solutions more or less independently from each other. Inspect all of them and either accept the most sensible if a meaningful choice is possible or else regard the solution as ambiguous, perhaps reporting two different explanations for the same data.

For this study, seed values 1 to 20 were tested resulting in almost equal Q values (see Table 2) and on examination almost identical factors were obtained.

Table 2: PMF2: Q values for different seeds.

Seed	Q	Seed	Q
1	3393.2671	11	3393.1147
2	3393.1899	12	3393.1899
3	3393.1897	13	3393.1921
4	3393.1709	14	3393.1970
5	3393.1521	15	3393.1782
6	3393.1687	16	3393.1583
7	3393.2148	17	3393.1155
8	3393.2102	18	3393.1528
9	3393.1873	19	3393.1797
10	3393.2144	20	3393.1960

4.7 *Rotational Freedom*

In the bilinear model an additional form of non-uniqueness results due to rotational ambiguity, or different transformations of **F** and **G** that produce as good a fit to the measured data. This is demonstrated by the identity:

$$GF = GTT^{-1}F \qquad (10)$$

where T is any non-singular square matrix. The expression **GT** and $\mathbf{T^{-1}F}$ represent a pair of factors which are "equally good" as the original pair, **G** and **F**. Understanding and controlling rotations is discussed in [12]. There are two alternatives for handling rotational freedom in PMF2 [5]:

- positive values of the user-specified parameter *Fpeak* change the object function in such a way that concentration values in the matrix **F** tend to assume extreme values in both directions, i.e. either close to zero or close to the upper bound, depending on the form of normalisation chosen. Negative values of *Fpeak* result in the same effect on matrix **G**.
- individual elements of matrix **F** may be pulled towards zero by introducing special penalty terms in the objective function. The penalty elements are controlled by the corresponding element in the *Fkey* matrix. This imposes a restriction on the accepted solutions, thus reducing the rotational freedom.

The form of the penalty depends on what value each element is set to:

- key = 0: element is constrained to non-negative values only
- key = 1: element is free to take any values
- key > 1: element is bound to zero, the strength of the bond depends on the chosen value, e.g. if key = 3 it is weak, else if key = 10, it is very strong

Note that these manipulations should be carried out on an already obtained solution, by using the restart feature. Also, forcing the fitting process by *Fpeak* and *Fkey* restricts the solutions and thus Q might be larger. One must take care that this increase is not too large.

Values of *FPeak* from -1 to 1 were tested. This produced a range of Q values (Table 3) where local minima's were found to cluster around *FPeak* = 0 (Figure 6). As can be seen from the figure a number of solutions result in a similar Q value, any of which could be a solution. The solution with *Fpeak* of 0 was analysed and the resulting factors could be physically explained based on a number of previous studies, see section 4.10.

Table 3: PMF2: Variance of Q, showing rotational freedom.

FPEAK	Q
-1	3755.5447
-0.75	3783.2278
-0.5	3441.0317
-0.25	3404.6904
-0.2	3401.3337
-0.1	3394.7385
-0.01	3393.2004
-0.001	3393.2515
0	3393.1147
0.001	3393.1931
0.01	3393.2073
0.1	3395.4932
0.2	3408.8235
0.25	3412.5498
0.5	3682.4041
0.75	3755.5862
1	3870.0200

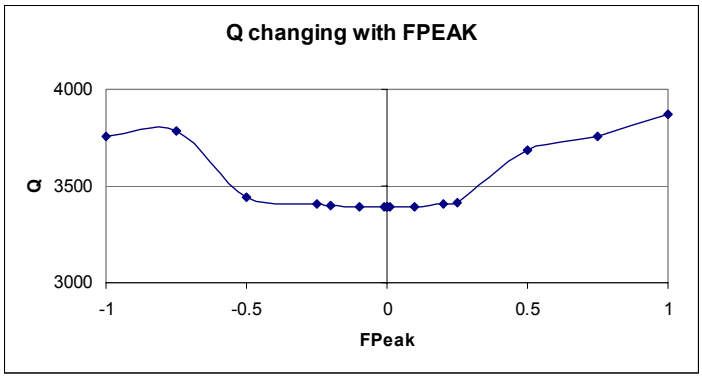

Figure 6 : Q results for different FPeak values.

A third method, which allows detailed control of rotations is available which uses the matrix *rotcom*, although the author of the code discourages the use of this.

4.8 *Standard Deviation of the Factor Matrices*

PMF2 produces two matrices, G_std-dev and F_std-dev, which give the uncertainty of each element of the factor matrix. The standard deviation for the factor matrix **G** are generated under the assumption that **F** is kept fixed and similarly for the **F** matrix, under the assumption that **G** is kept fixed.

An example of the computed standard deviation can be seen in Figure 7, where the results for the soil factor are shown.

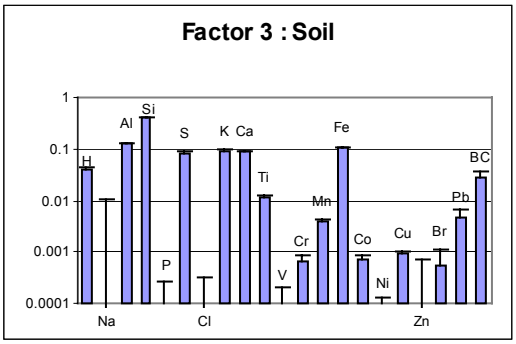

Figure 7: Plot for the soil factor showing standard deviation as calculated by PMF2.

4.9 *Explained Variation*

A quantity referred to as the explained variation (EV) [8], summarises how important each factor element is in explaining one row or column of the observed matrix. The equations used to evaluate EV are documented in [8]:

$$EV(G)_{i,k} = \frac{\sum_{j=1}^{m} \frac{\left|g_{i,k} f_{k,j}\right|}{s_{i,j}}}{\sum_{j=1}^{m}\left(\sum_{h=1}^{p} \frac{\left|g_{i,h} f_{h,j}\right| + \left|e_{i,j}\right|}{s_{i,j}}\right)} \qquad \text{for k=1,…,p}$$

$$EV(G)_{i,k} = \frac{\sum_{j=1}^{m} \frac{\left|e_{i,j}\right|}{s_{i,j}}}{\sum_{j=1}^{m}\left(\sum_{h=1}^{p} \frac{\left|g_{i,h} f_{h,j}\right| + \left|e_{i,j}\right|}{s_{i,j}}\right)} \qquad \text{for k=p+1}$$

In source appropriation studies, the element number j of the (p+1st) row of the EV for **F** indicates how much of the variable number j remains unexplained. In PMF2 an

option is provided to print the factors in order of importance as calculated by EV. Analysis of the results using this technique have not been carried out in this study, this section has been included for completeness.

4.10 *Discussion of PMF results*

The eight factors generated by PMF2 are presented in Figure 8. The results with 8 factors were found to be stable, i.e. the characteristics of the 8 factors did not change significantly with different starting seeds.

Analysis of the elements within the 8 factors revealed 8 possible sources. Factor 1 is an ocean (seaspray) factor, identified here as *Sea1,* with a strong representation from Na and Cl which is in line with the "fresh marine" source from Qin *et al.* [32] who had high contributions of Na, Cl and Mg. K and Ca were also present in our ocean factor, as they are minor constituents of sea salt [22]. The PMF technique allows the Black Carbon element that appears in this factor to be pulled down to zero (see section 6) as it is not expected from marine sources. This has also been noted in Qin *et al*. [32] who have adjusted their carbon element in the marine factor to better fit emission patterns from sources. They believe it is particularly significant to do this as the carbon fractions of their total mass was large, and could obscure finer signals. Factor 2 is a chlorine depleted ocean factor that we have identified as *Sea2*. Qin *et al*. [32] found a similar factor, in their PMF2 analysis using 9 factors, with high contributions of Na, Mg and S (rather than Cl). Chlorine depleted marine aerosols have been reported by Lee *et al.* [5] due to inhomogeneous loss of Cl to reaction with nitric acid. This factor also has K and Ca, which are minor components of sea salt [22].

The five elements Al, Si, Ti, Ca, Fe are commonly the major elements that indicate a soil factor [27,28] and thus Factor 3 is identified as *Soil*. Ho *et al.* [24] analysed the composition of paved road dust and soil in Hong Kong and found that the constituent with greatest concentration was Si. Urban soil also contained Si as its highest abundance and Al with the second highest. It was also found that paved road dust samples are enriched in K, Ca and Fe. All of these have high contributions in Factor 3.

Factor 4 has high H content and identified as *Org/Coal*. The elemental analysis of this data is described in [10], where it is stated that ammonium sulfate and organic matter were the only two measurable major sources of hydrogen.

Factor 5, has been identified as a biomass smoke source, labelled *Smoke*, as H from organic material and K from vegetation have been found by Cohen *et al.* [27,28] to have the highest factor loadings and thus a strong correlation in their PCA analysis. Also, Yli-Tuomi *et al.* [31] have associated a biomass burning factor with high concentrations of Black Carbon and K. These are all major elements within our factor.

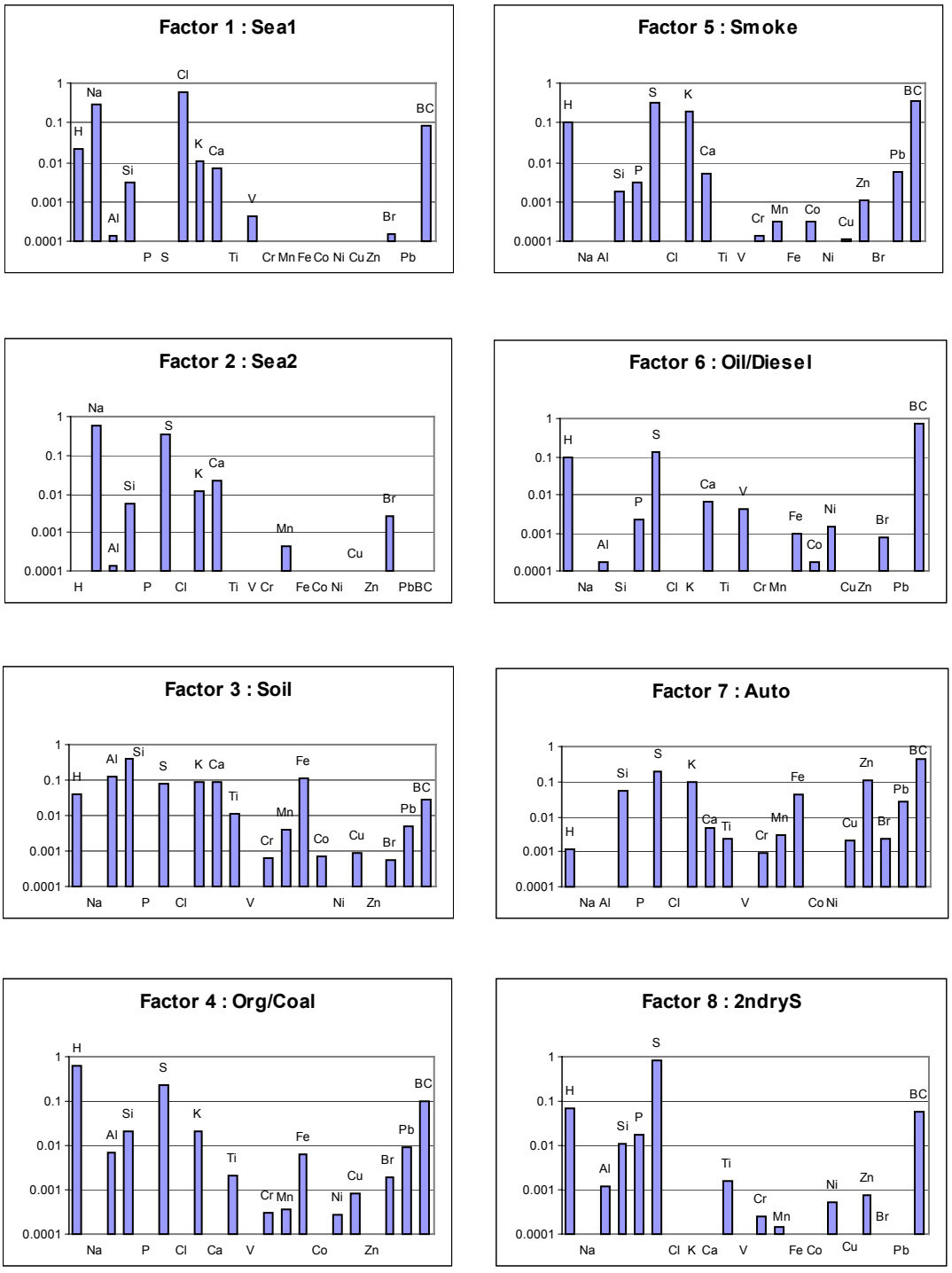

Figure 8: PMF2 results for 8 Factors, with EM=-14.

Factor 6 has been identified as *Oil/Diesel* as a number of authors have matched high V and Ni to burning of oil in power-plants [Qin *et al*. 32]. Also, Fung and Wong [25] have identified S and V as constituents of oil combustion sources and Cohen *et al.* [28] found that their *Diesel* factor was high in Ni, V, Black Carbon and H from organic matter. These elements that have been identified are major contributors in our factor and probably associated with oil and fossil fuel combustion associated with power production and diesel motor vehicles.

Factor 7 was identified as *Auto*, as Zn and Pb (prominent in this factor) have been reported by various authors [25; 28] to be markers for motor vehicle emission. Lee *et al.* [5] also reported the presence of Ca and Fe (which are both present in this factor) in their motor vehicle emission factor, which they concluded was due to road dust.

Factor 8, named *2ndrySulf* is a secondary factor (i.e. due to reactions within the atmosphere) of ammonium sulphate. A secondary sulfate factor has also been described in Ramadan [3].

4.10.1 PMF2 results with EM=-12

The results presented so far have used error mode: EM=-14. When EM=-12 was used, similar results were obtained (see Figure 9) with only some minor differences. These differences include: Na present in Factor 4 (*Org/Coal*) (when EM=-12), which wasn't present when EM=-14 was used. Also, higher H concentration was observed in Factor 7 (*Auto*) and lower P concentration in Factor 6 (*Oil/Diesel*), when EM=-12 was used as opposed to when EM=-14. However, these differences do not significantly affect the classification of the factors as the contribution of these elements is only small.

In terms of the convergence statistics, the main difference is that when EM=-14 was used only 4 points exceeded the outlier limit of 4 as specified in section 4.3, whereas when EM=-12 was used, 9 points exceeded the outlier limit of 4. A Q value of 3393 was reached for EM=-14 and 4287 when EM=-12. Inspection of the scaled residuals showed more values above 2.0, when EM=-12 and a number of values were around 3.6, thus the higher Q value. Calculating $s_{i,j}$ by equation 6a (when EM=-14) takes into consideration both the maximum measured and maximum calculated value, whereas equation 6b (when EM=-12) takes into consideration only the maximum measured value. This could well explain the smaller Q value and scaled residuals observed when EM=-14, which follow the established analysis techniques, *i.e.* that of Q approaching an expected value, as specified in section 4.4, and the scaled residuals generally within the range of –2.0 and +2.0. When examining Q and the residuals one needs to keep in mind that a larger Q value and some larger residuals will be observed when EM=-12 is used.

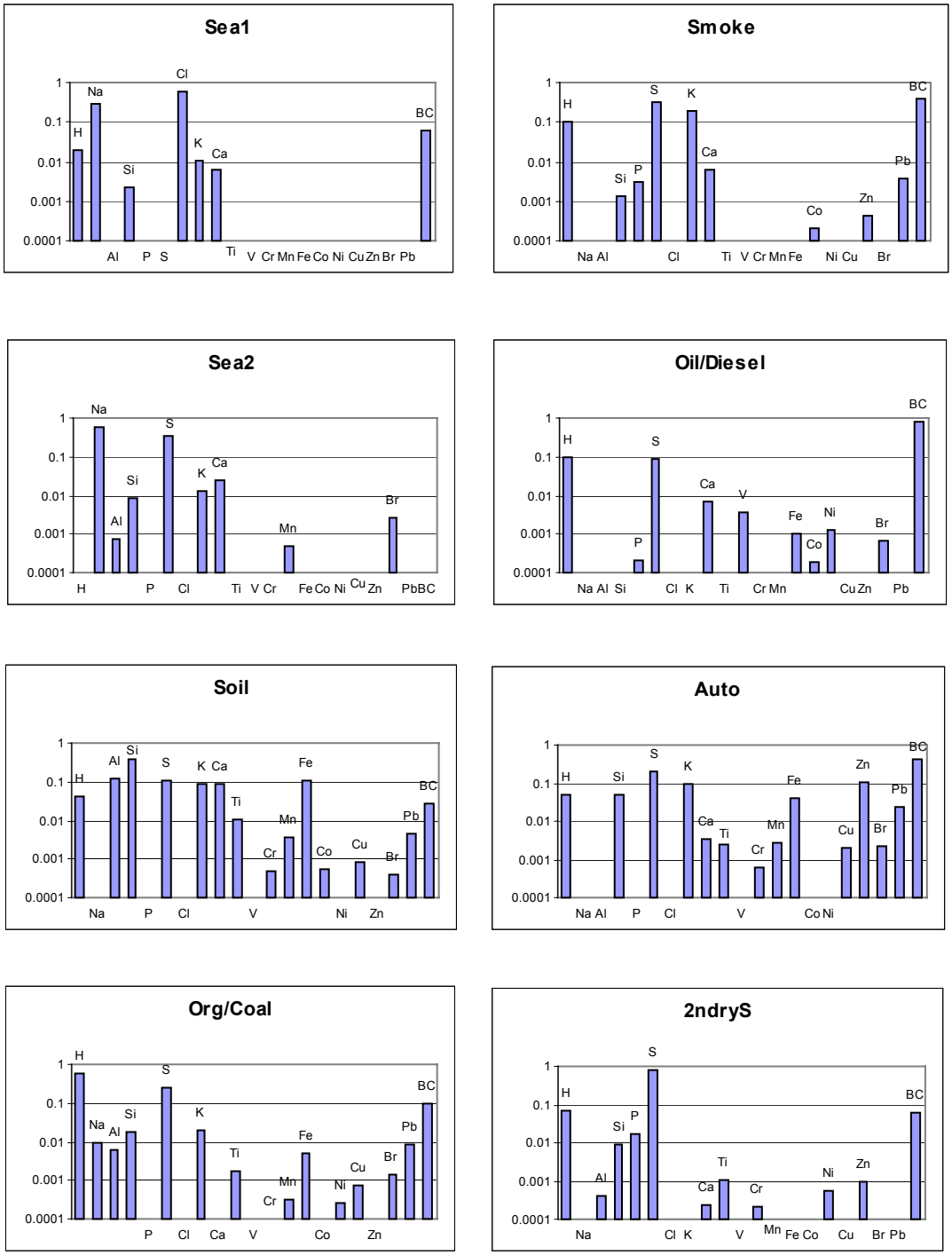

Figure 9: PMF2 results for 8 Factors, EM=-12.

5. Data Analysis with ME2

The Multilinear Engine (ME) [13] solves general multilinear models. The multilinear least squares model to be fitted is specified by enumerating all equations that together make up the model.

In PMF2 a fixed model is fitted to the measured data with a number of predetermined constraints on the form of the solution. The user has some control of the fit through a set of defined parameters and what subset of the predefined output is to be printed to which file and under what format. ME2 provides an engine to solve models, which are specified by the user by using a scripting language. The user is also provided with a library, ME2libr, which implement the appropriate code to perform some common analysis, some of which functionality was available under PMF2.

For example, if the 2-way bilinear model given in equation (1) is to be solved, subject to the constraints that the first column of the factor matrix is to be smooth, then the objective function used for optimisation (i.e. equation 4) will need to be augmented as follows:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{e_{i,j}^2}{s_{i,j}^2} + \frac{1}{w} \sum_{j=1}^{p-1} \left(f_{j+1,1} - f_{j,1} \right)^2 \qquad (12)$$

Where, the second term introduces a penalty on the difference between successive elements of column 1 of the factor matrix **F**, and w is a chosen weight. While in PMF2 the form of the constraints is fixed, in ME2, setting up of constraints is under complete control of the user. A facility for specifying non-negativity constraints is provided through the concept of *child arrays*. *Child arrays* are associated with each array and contain additional information about the data in the arrays. For factor matrices, a *child array* known as *fkey*, can be used to specify if the corresponding element is constrained, and an additional two *child arrays*, known as *flow* and *fhigh* can be used to specify the lower and upper limits. In addition, factor elements can be given a value and marked as *locked* at that value or can be forced to zero through a *masked* option.

Given that the user has freedom to specify the model equations, associated constraints and regularisation (known as auxiliary equations), a library is provided which generates some of the more commonly used auxiliary equations. The library also provides a number of service routines, e.g. array manipulation for pre- and post-processing of arrays and data input and output. For example, a subroutine *Expvariat*, provides code to compute the Explained Variation for the basic 2-way model.

Technical details of the ME2 implementation are given in [14], and the scripting language, used to define the model, is detailed in [15].

5.1 *Comparison of PMF2 and ME2*

5.1.1 ME2 Model Formulation

The user has the freedom to specify the form of the model to be fitted to their data. Information on possible models and how this is specified to ME2 is given in [13, 14 and 15].

5.1.2 Standard deviation of observations

ME2 offers the same options for error modes as available in PMF2. The error mode specified by EM=-14 is used in this analysis, for comparison with the PMF2 results.

5.1.3 Outliers and the robust mode

ME2 has a similar option to PMF2 available to specify the robust mode and outlier distances. However, in ME2 the absolute value in equation (7) is replaced by:

$$-\alpha_1 < \frac{x_{i,j} - y_{i,j}}{s_{i,j}} < \alpha_2$$

That is, the lower and upper limits can be specified by setting the two values, α_1 and α_2. For the ME2 study presented in this report the outlier threshold distance, i.e. α_1 and α_2, have both been set to 4, as for PMF2.

5.1.4 Solution and Convergence of the algorithm

In ME2 convergence criteria are specified in the same way as for PMF2, i.e. through the *lims* parameters. The scaled residuals of the fit are available in the *child array, sres*.

5.1.5 Estimating the number of sources/factors

Analysis of the number of factors for ME2 would be similar to PMF2, however for this study, given that the analysis on the number of factors was carried out under PMF2, 8 factors were used under ME2.

5.1.6 Multiple Solutions

ME2 faces the same issues as PMF2 in finding the global minimum and once again the ability to initialise the factor elements with different seeds is provided to give the user some freedom in exploring the solution space.

5.1.7 Rotational Freedom

When solving 2-way bilinear models with ME2, the same issues arise as in PMF2. To some extent some of the rotational freedom is removed in solving multi-way, multilinear problems. As with PMF2, the user can reduce the rotational freedom by using normalisation of the factor elements or pulling factor elements to a specified value. However, in ME2 there is no equivalent to the PMF2 *Fkey* matrix, or the *Fpeak* option. Rather, the user has to specify the equation to achieve this, e.g. a subroutine called *Pullto*, supplied in the library, generates the required equations to pull columns of a factor matrix towards a given value. Also the *child array fkey* in ME2 provides a facility to specify constrains on the factor elements.

In order to perform rotations in ME2 [12] the following additional terms can be included in the objective function:

$$Q^n = \frac{\sum_{h=1}^{p}\left(1 - \sum_{j=1}^{m} f_{h,j}^2\right)^2}{\alpha^2} \qquad (13)$$

$$Q^p = \beta^2\left(\sum_{h=1}^{p}\sum_{i=1}^{n} g_{i,h}\right)^2 \qquad (14)$$

The term Q^n defined in equation (13) attempts to normalise the rows of **F** to unit norm. The parameter α should be chosen small enough so that in the computed solution, the norms of rows of **F** deviate from unity at most by a small value (e.g. 0.01). The term Q^p defined by equation (14) attempts to pull the sum of all elements of **G** towards zero. A tedious analysis shows that this attempt leads to rotations in the direction of negative *Fpeak*, as in PMF2, i.e. the parameter β^2, corresponds to negative values of *Fpeak*. Positive values of *Fpeak* can be simulated by equation (13) and (14) where **G** and **F** have been interchanged.

The role of the Jacobian matrix in analysing uniqueness is described in [14]. If the Jacobian matrix in not of full rank (a certain combination of the columns is a zero vector), the solution is not locally identifiable. The gradient at the computed solution can be accessed through the *child matrix* for each factor matrix, called *fgrad.*

5.1.8 Standard Deviation of Factor Matrices

An automated way of generating the standard deviations of the factor matrices has not been provided. A technique that could be investigated is that of obtaining a solution, followed by restarting with the fixed solution for one of the factor matrices while solving for the second matrix, perhaps with different seeds.

5.1.9 Explained Variation

A sample subroutine has been provided in the library to allow the calculation of the Explained Variation. A similar formula to that used in PMF2 is implemented.

5.2 ***Bilinear Model under ME2***

As a first step the Hong Kong data was analysed under ME2 using a bilinear model, with the normalisation of columns of the matrix **F**, i.e. the following two equations are solved:

$$x_{i,j} = \sum_{p=1}^{P} g_{i,p} f_{p,j} + e_{i,j} \qquad (15)$$

$$\sum_{p=1}^{P} f_{p,j} = 1 \qquad (16)$$

subject to non-negativity constrains, as defined by the built-in option of setting *F.key* to low limits and *F.flow* to a lower limit of zero, as detailed in [13].

5.2.1 ME2 Results and Discussions

The data was analysed under ME2, where 10 runs were carried out, and 8 factors were used, as determined through the PMF analysis. The seed value of the random number generator used for initialisation of the factor matrices, was initialised to 1 and for each successive run was incremented by 100. Analysing the results of the 10 runs showed that the major elemental contributions were the same in the results of most runs, with some of the minor elements moving between the factors.

The 9th run (Seed=801) obtained the smallest Q (Table 4) which was chosen for further analysis. Inspection of the scaled residuals showed that only a very small

number had an absolute value greater than 2, with the largest absolute value of 2.3. An experiment was undertaken, where columns of the factor matrices were pulled down to certain values, *i.e.* adding an auxiliary equation in which the sum of nominated columns of the factor matrices is constrained to a given value. The minimum Q value obtained from this procedure was 4047, implying that some rotational freedom exists in the solution. This was not perused further as pulling individual elements to zero was considered more appropriate for this work (see section 6).

The ME2 generated factors with the smallest Q value (Figure 10), thus indicating the best fit to the data, corresponded very closely to the PMF results (Figure 8), with the only noticeable difference being the amount of Na appearing in Org/Coal factor.

Table 4 : ME2: Q values for different seeds.

Run	Seed	Q
1	1	4253
2	101	3977
3	201	4640
4	301	3489
5	401	3536
6	501	3408
7	601	3448
8	701	3974
9	801	3401
10	901	4512

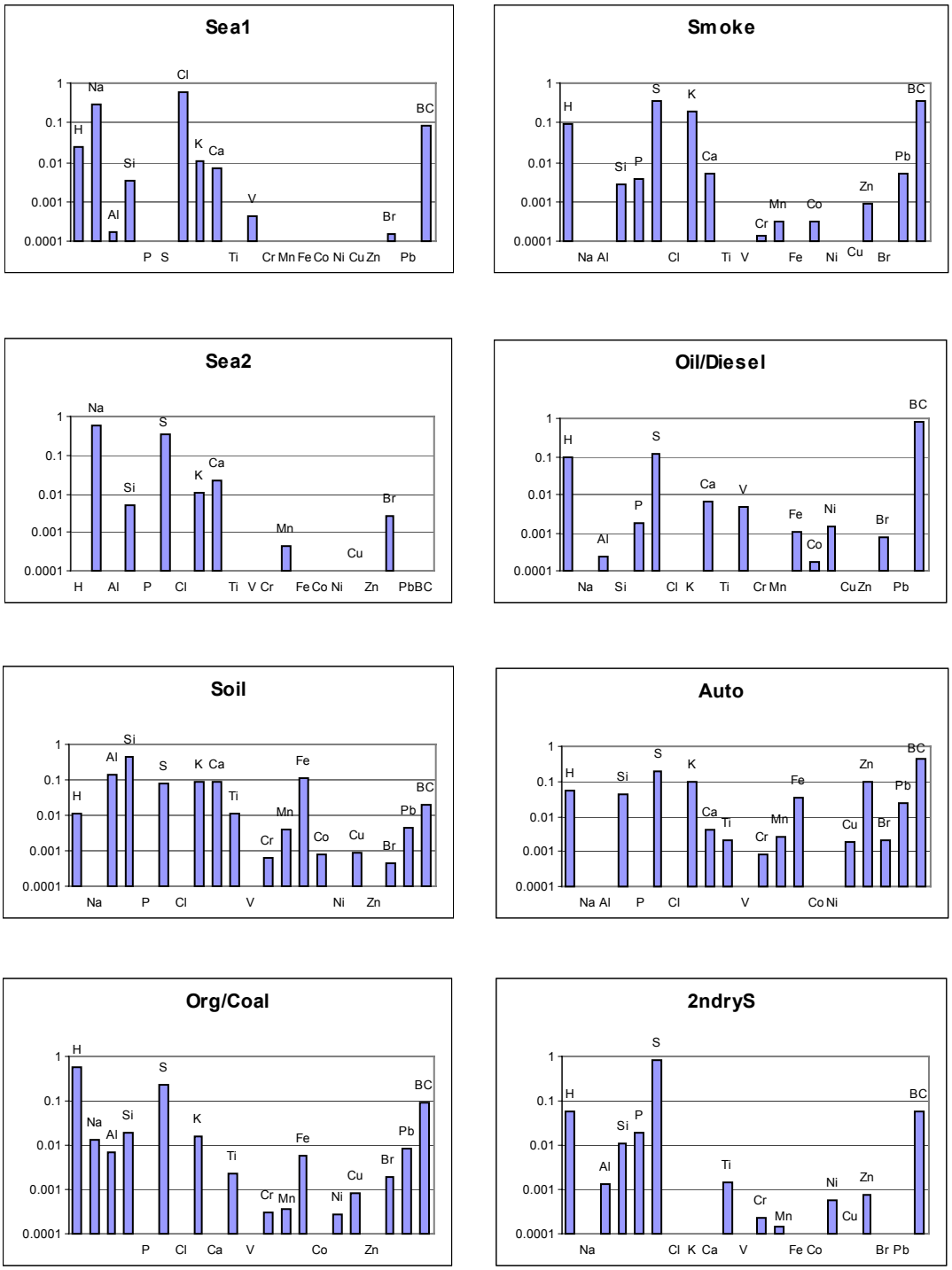

Figure 10 : ME2 results for 8 Factors.

6. Pulling Down Factor Elements

On examination of the *Sea Spray*, *Soil* and *Auto* factors a decision was made to pull down the Black Carbon element from these factors. This was achieved very simply by setting the corresponding element of the *fkey* child matrix to *masked*. The new solution obtained resulted in a slightly higher Q value, that of 3472. The new factors are shown in Figure 11. Aside from the Black Carbon being pulled down some of the other elements were affected; H was removed from *Soil* and so was Na from the

Org/Coal factor, with the Black Carbon component of the *Org/Coal* factor increasing. Also K from *Auto* was removed and an increase in *Org/Coal* was observed. Pulling down the same elements in the corresponding factors under PMF2 resulted in similar results, see Figure 12.

A number of additional analyses of the data were carried out following a review of these results [33]. Given that the S concentration is high in the measurements, an analysis was carried out where the error for S was increased, following which it was found that some of the Black Carbon was reduced from the *Auto* and *2ndryS* factors and an increase was observed in the *Org/Coal* factor. The presence of Br and Pb in the *Auto* and *Ord/Coal* factors was examined by applying a weak pulling down of these two elements from the two factors, in which case it was found that there was a stronger association of Pb with the elements in the *Auto* factor than in the *Org/Coal* factor.

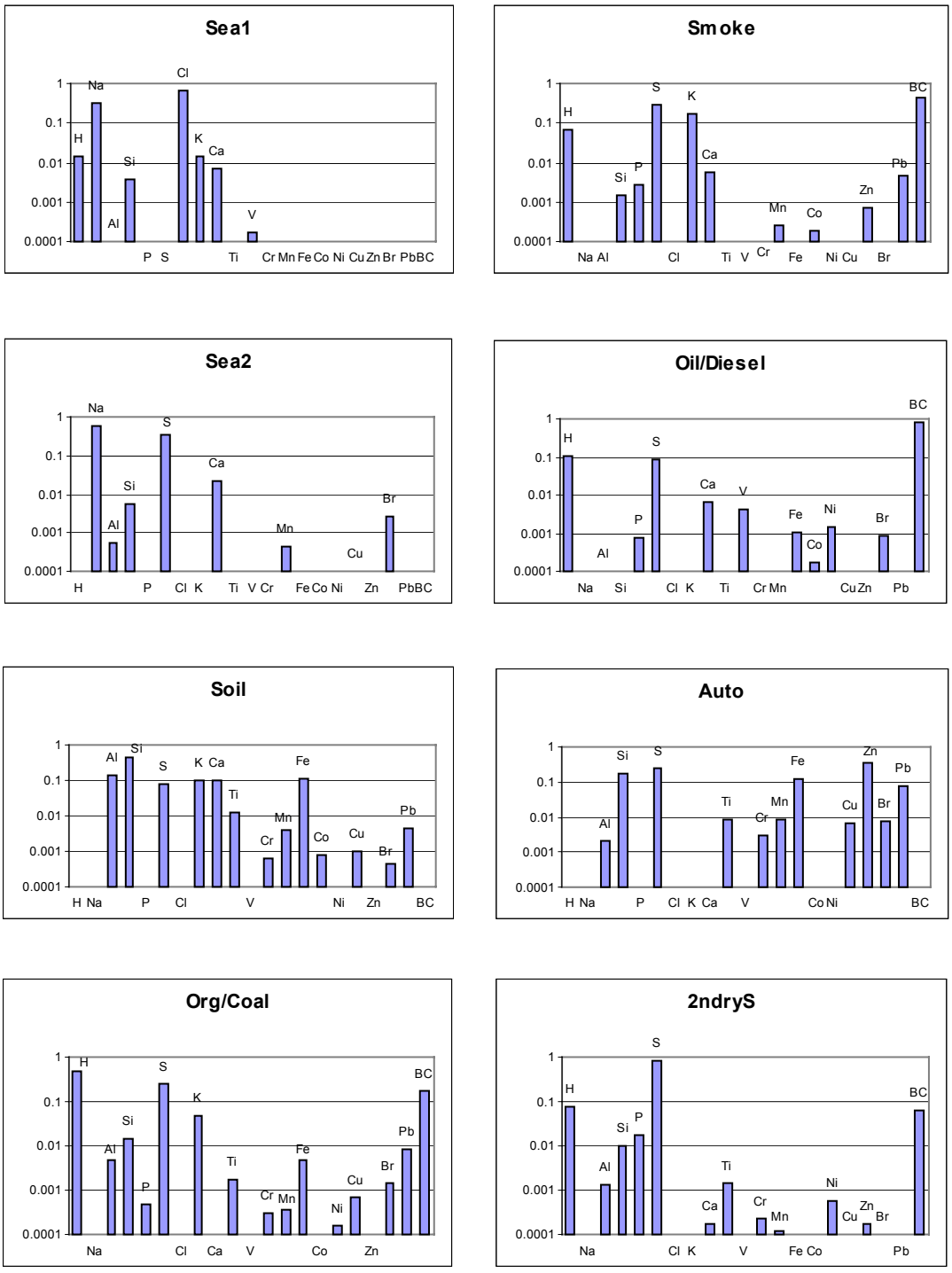

Figure 11: ME2 Results after pulling down selected elements.

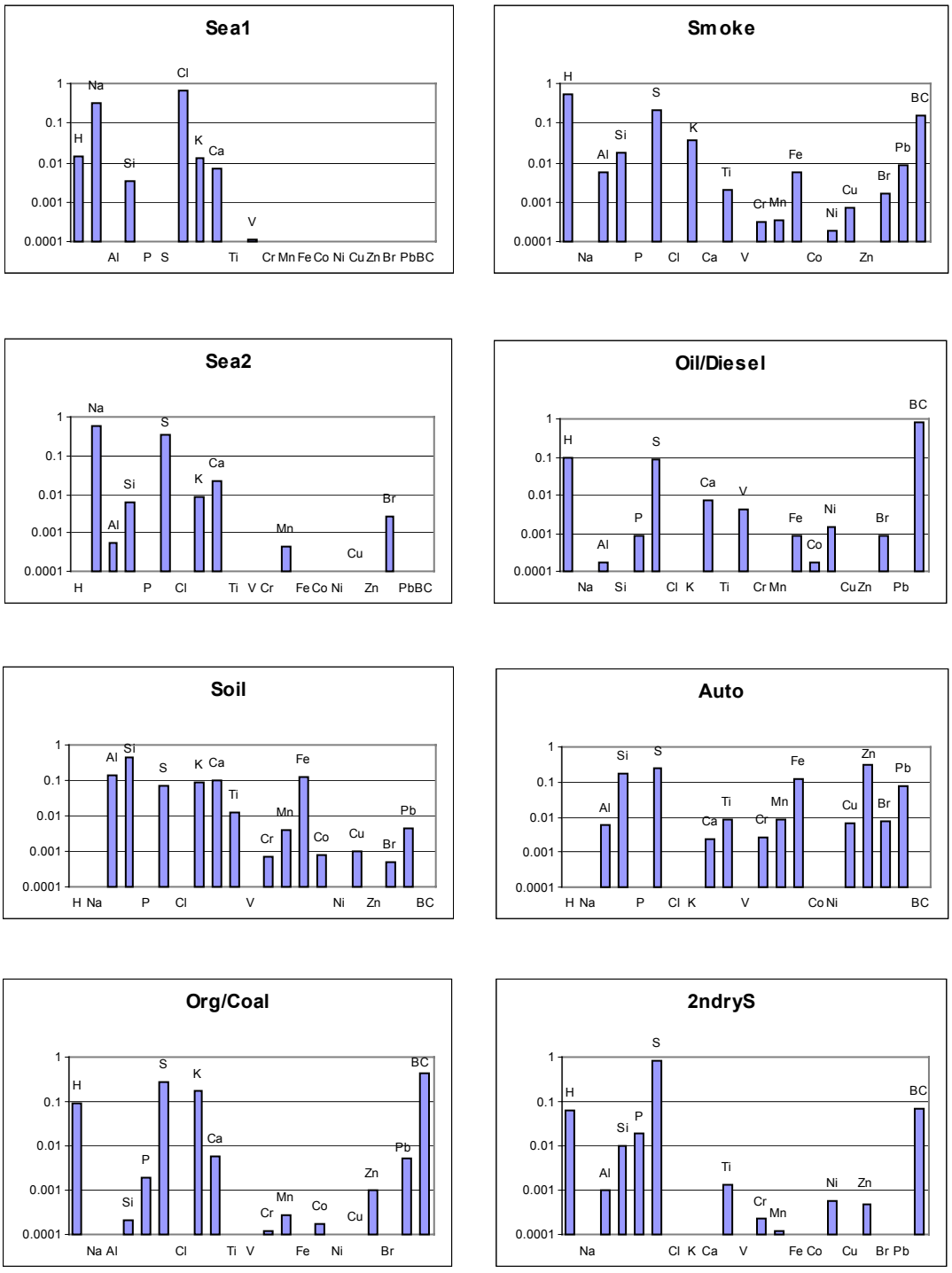

Figure 12: PMF2 Results after pulling down selected elements.

7. Multivariate Linear Regression

Once the factor analysis has been carried out, further analysis might be to perform a multivariate linear regression fit to the total mass, **C**. That is, while the factor analysis is carried out on data matrix, **X**, which contains the mass of each element, a regression fit of the following form:

$$\mathbf{C} = \mathbf{GZ} + \mathbf{E}$$

can be carried out on the total mass data, *i.e*. **C**, which is a vector of **n** measurements, **G** is an **n** by **p** matrix, as obtained by the factor analysis, **Z** and **E** are vectors of **p** entries being the result of the fit and the residuals, respectively . Thus **Z** contains the contribution to the total mass from each of the factors. Performing a linear least squares fit resulted in **Z**=(2.0, 2.9, 3.0, 4.5, 3.6, 3.2,2.0, 3.5). From these coefficient an approximate contribution from each of the factors to the total measured mass can be derived. The results of this analysis are shown in Table 5. In Table 5 the contribution from each factor is given as a percentage of the total measured mass. The highest contributors to the total mass are the *Org/Coal* factor with 24% and *2ndryS* with 23% followed by *Smoke* of 19% and *Oil/Diesel* of 12%. This is not surprising as these factors have high S, Black Carbon or H content, which were the three elements identified with the highest contribution to the total mass.

Table 5 : Contribution to total measured mass from each factor.

Factor	% contribution to mass
Sea1	2
Sea2	10
Soil	8
Org/Coal	24
Smoke	19
Oil/Diesel	12
Auto	2
2ndryS	23

Similar analysis can be carried out by multiplying elements of the **G** and **F** matrices, as obtained by the factor analysis, to obtain mass contribution from each element within the identified source factors. The results are shown in Table 6. This analysis results in a similar contribution to the total mass from each factor, with the main difference being in the reduced fraction of *Org/Coal*, and a slight increase in *Auto*, being due to the solution of a different representation of the problem.

Table 6 : Breakdown of contributions to total mass.

	Soil	Oil/Diesel	Sea2	Smoke	2ndryS	Org/Coal	Sea1	Auto
H	0	29656	0	26466	35639	190601	863	0
Na	0	0	164644	0	0	0	19544	0
Al	25153	29	142	0	643	1899	5	166
Si	80888	0	1533	561	4992	5753	225	12965
P	0	211	0	1100	8756	183	0	0
S	15289	23199	91818	117080	401655	93200	0	20246
Cl	0	0.2866	0	0	0	0	39464	0
K	18581	0	0	67722	0	18549	885	0
Ca	17411	1920	6127	2340	85	0	415	0
Ti	2186	0	0	0	685	692	4	626
V	0	1256	0	0	0	0	10	0
Cr	122	0	0	39	107	112	0	218
Mn	760	0	113	104	58	136	0	672
Fe	21653	289	0	0	0	1790	0	9947
Ca	148	47	0	74	0	0	5	0
Ni	0	423	0	0	290	63	0	0
Cu	174	0	25	24	43	263	1	526
Zn	0	0	0	282	87	0	0	26538
Br	80	255	715	0	43	570	3	553
Pb	850	0	0	1804	0	3177	0	5837
BC	0	229266	0	175476	31992	69256	0	0
% of Total	8.6	13.4	12.4	18.4	22.7	18	2.9	3.6

8. Using Bootstrapping to examine the results

It was pointed out in [26] that measurement errors affect the analysis when determining the number of factors. A similar approach was used here to examine how sensitive the obtained solution is to measurement error. The original data was used as a pool to randomly generate samples to be analysed individually, in a bootstrapping fashion [29]. Twenty samples of 266 observations were generated from the original data using random sampling with replacement. The method employed here was to generate 10 solutions for each of the 20 samples, following which the optimum solution is selected. The 8 fingerprints generated by the optimum solutions from each bootstrap were then pooled and classified into 8 source fingerprints using cluster analysis. Thirteen out of the twenty solutions resulted in 8 unique fingerprints, however the other seven solutions all had a variation of the same factor, i.e. the factor identified as *Sea2*, in some cases resulted with high contribution of Black Carbon and H. As a result the classification software incorrectly classified it as the *Auto* factor. The seven solutions were analysed by hand and the factor was appropriately classified. Figure 13 shows the average of each element in each factor as generated by the 20 runs. While some variations are present, the dominant elements exhibit the same behaviour in all fingerprints with some difference seen in the factor labelled

Sea2. The error bars show the standard deviation of the 20 solutions, showing that little difference between the solutions existed and we can conclude that the obtained solution represents the situation well.

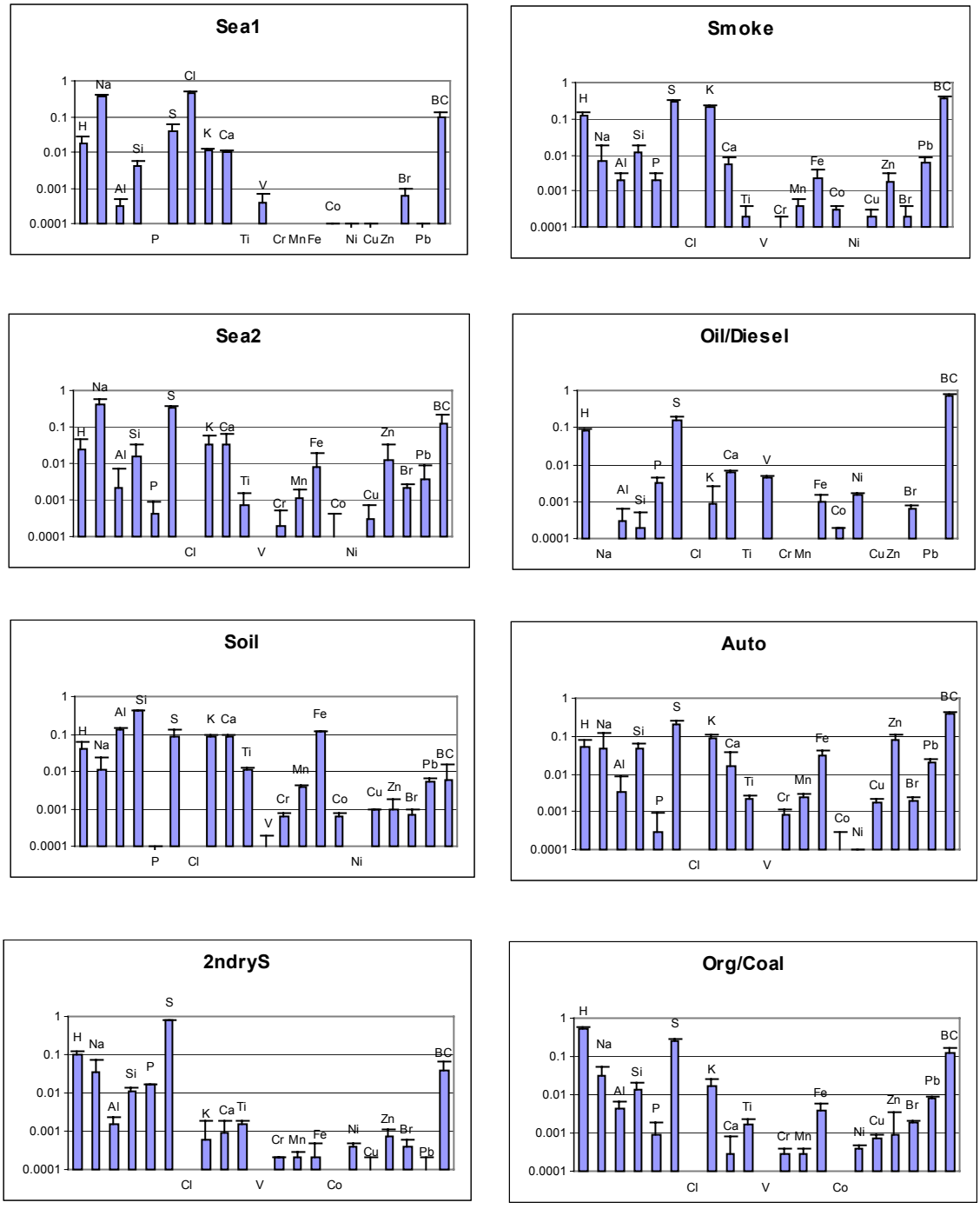

Figure 13: Average of 20 runs under ME2.

9. Conclusion

A 2-way factor analysis has been carried out on aerosol measurements at Hong Kong using two separate programs; PMF2 and ME2. The results from PMF2 were successfully reproduced using ME2 even though the two programs implement different solution techniques.

The results from the two programs were found to be very similar, giving us confidence in the two techniques. When EM=-14 was used in both PMF2 and ME2, the results were almost identical with only minor differences in the smaller contributing elements, thus not affecting the classification of the factors. These results are also compatible with previous studies of a subset of the data using Principal Component Analysis.

Pulling down of factor elements wasn't fully exploited by the authors, due to the desire of not applying personal judgement to the analysis. However following a discussion of these results [33], the benefit of applying a weak pulling down of elements was investigated and fully appreciated. When unexplained elements appear to have an association with a certain factor, the degree of this association can be examined by applying a weak pulling down of the element.

The strength of ME2 lays in the ability to perform multi-linear multi-way analysis. The ability to introduce seasonal variations as well as information on air parcel trajectories will be explored in a subsequent publication.

10. Acknowledgments

The authors would like to acknowledge the helpful discussion of this work with Prof. Philip K. Hopke during his visit to ANSTO.

11. References

1. Henry R.C., Multivariate receptor models – current practice and future trends, Chemometrics and Intelligent Laboratory Systems 60 (2002) 43 – 48.

2. Hopke P.K, Receptor Modeling for Air Quality Management, Elsevier, Amsterdam, 1991.

3. Ramadan Z., Eickhout B., Song X., Buydens L.M.C and Hopke P.K., Comparison of Positive Matrix Factorisation and Multilinear Engine for source appropriation of particulate pollutants, Chemometrics and Intelligent Laboratory Systems 66 (2003) 15-28.

4. Jollife I.T. , Principal component Analysis, Spring_Verlag, New York, 1986

5. Lee E., Chan C.K. and Paatero P. Application of positive matrix factorisation in source appropriation of particulate pollutants in Hong Kong, Atmospheric Environment 33 (1999) 3201-3212.

6. Harshman R.A., Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multi-mode Factor Analysis, UCLA Working Papers in Phonetics, 16, 1-84.

7. Paatero P. and Tapper U, Positive Matrix Factorisation: A non-negative factor model with optimal utilisation of error estimates of data values, Environmetrics, Vol 5, 111-126 (1994).

8. Paatero P, User's Guide for Positive Matrix Factorization programs PMF2 and PMF3, Parts 1 and 2. Feb 19, 2004.

9. Paatero P, The Multilinear Engine – a Table-driven Least Squares Program for Solving Multilinear Problems, Including the n-way Parallel Factor Analysis Model. Journal of Computational and Graphical Statistics, (1999), Vol 8, Number 4, pp 854-888.

10. Cohen D.D., Garton D., Stelcer E. and Hawas O. Multi-elemental Analysis and Characterisation of Fine Aerosols at Several Key ACE-Asia Sites, J of Geophysical Research, Vol 109, D19S12, 2004

11. Paatero P. Least squares formulation of robust non-negative factor analysis, Chemometrics and Intelligent Laboratory Systems 37 (1997) 23-35.

12. Paatero P, Hopke P.K, Song X, Ramadan Z, Understanding and controlling rotations in factor analysis models, Chemometrics and Intelligent Laboratory Systems 60 (2002) 253-264.

13. Paatero P., User's guide for the Multilinear Engine program "ME2" for fitting multilinear and quasi-multilinear models, February 10, 2000.

14. Paatero P., The Multilinear Engine – a Table-driven Least Squares Program for Solving Multilinear Problems, Including the n-way Parallel Factor Analysis Model. Journal of Computational and Graphical Statistics (1999)_, Vol 8, Number 4, pp 854-888.

15. Paatero P. The Multilinear Engine (ME-2) script language (v 1.095), May 2004.

16. Draxler R.R and Rolph G.D., Hybrid single-particle Lagrangian integrated trajectory (HYSPLIT), Model, http://www.arl.noaa.gov/ready/hysplit4.html

17. Biegalski S.R., Hopke P.K, Total Potential Source Contribution Function Analysis of Trace Elements Determination in Aerosol Samples Collected near Lake Huron, Environ. Sci. Technol, 1004 38, 4276-4284.

18. Qin Y, Chan C.K and Chan L.Y, Characteristics of chemical compositions of atmospheric aerosols in Hong Kong: spatial and seasonal distributions, The Science of the Total Environment 206 (1993) 25-37.

19. Zahorowski, W, Chambers, S. and Henderson-Sellers, A. “Ground based radon-222 observations and their application to atmospheric studies.” Journal of Environmental Radioactivity, in press, 2004.

20. Kim E., Hopke P.K, Paatero P and Edgerton E.S, Incorporation of parametric factors into multilinear receptor model studies of Atlanta aerosol, Atmospheric Environment 37 (2003) 5009-5021.

21. Paatero P and Hopke P.K. Utilizing wind direction and wind speed as independent variables in multilinear receptor modelling studies. Chemometrics and Intelligent Laboratory Systems, v60 25.

22. Rae F, Bridgman H.A., Paatero P and Cohen D, Assessing the relationship between fine particle chemistry, source and transport using ME2, unpublished report.

23. Ho K.F., Lee S.C., Chan C.K., Yu J.C., Chow J.C. and Yao X.H, Characterization of chemical species in $PM_{2.5}$ and PM_{10} aerosols in Hong Kong, Atmospheric Environment 37 (2003) 31-39.

24. Ho K.F., Lee S.C, Chow J.C and Watson J.G, Characterization of PM_{10} and $PM_{2.5}$ source profiles for fugitive dust in Hong Kong, Atmospheric Environment 37 (2003) 1023-1032.

25. Fung Y.S and Wong L.W.Y, Apportionment of Air Pollution Sources by Receptor Models in Hong Kong, Atmospheric Environment Vol 29, No 16, pp2041-2048, 1995.

26. Park E.S, Henry R.C. and Spiegelman C.H., Determining the Number of Major Pollution Sources in Multivariate Air Quality Receptor Models, NRCSE, Technical Report Series, NRCSE-TRS No. 034.

27. Cohen D.D., Stelcer E., Hawa O. and Garton D., IBA methods for characterising of fine particulate atmospheric pollution: a local, regional and global research problem, Nuclear Instruments and Methods in Physics Research B 219-220 (2004) pp145-152.

28. Cohen D.D., Garton D., Stelcer E. and Hawas O., Accelerator based studies of atmospheric pollution processes, Radiation Physics and Chemistry 71 (2004) pp759-767.

29. Efron B. and Tibshirani R, An Introduction to the Bootstrap, Chapman & Hall, London, 1993.

30. Qin Y., Oduyemi K., and Chan L.Y., Comparative testing of PMF and CFA testing, Chemometrics and Intelligent Laboratory Systems 61 (2002) 75-87.

31. Yli-Tuomi T., Kopke P.K., Paatero P., Shamsuzzoha Basunia M., Landsberger S., Viisanen Y., and Paatero J., Atmospheric aerosol over Finnish Arctic: source analysis by the multilinear engine and the potential source contribution function, Atmospheric Environment 37 (2003) 4381-4392.

32. Qin Y., Chan C.K., and Chan L.Y., Characteristics of chemical compositions of atmospheric aerosols in Hong Kong: spatial and seasonal distributions, The Science of the Total Environment 206 (1997) 25-37.

33. Hopke K. P., 2005, Clarkson University, Potsdam, NY. Private communication.