REGIONAL CO-OPERATIVE AGREEMENT

INTERNATIONAL ATOMIC ENERGY AGENCY

NAHRES-52
Vienna, 2000

WORKBOOK ON

DATA ANALYSIS

Prepared by
Philip K. Hopke

FOR PARTICIPANTS IN THE UNDP/RCA/IAEA
SUB-PROJECT ON AIR POLLUTION AND ITS TRENDS
Project No. RAS/97/030/A/01/18

EDITORIAL NOTE

This document is not a formal publication of the International Atomic Energy Agency (IAEA), and the views expressed in it do not necessarily reflect those of the IAEA or of governments of the Member States or organizations under whose auspices the work described herein was carried out. The use in this document of particular designations of countries or territories does not imply any judgment by the IAEA as to the legal status of such countries or territories, of their authorities and institutions or of the delimitation of their boundaries.

A Workbook on Data Analysis

Philip K. Hopke
Department of Chemistry
Clarkson University
Potsdam, NY 13699-5810, USA

Table of Contents

INTRODUCTION

As a consequence of various IAEA programmes to sample airborne particulate matter and determine its elemental composition, the participating research groups are accumulating data on the composition of the atmospheric aerosol. It is necessary to consider ways in which these data can be utilized in order to be certain that the data obtained are correct and that the information then being transmitted to others who may make decisions based on such information is as representative and correct as possible. In order to both examine the validity of those data and extract appropriate information from them, it is necessary to utilize a variety of data analysis methods. The objective of this workbook is to provide a guide with examples of utilizing data analysis on airborne particle composition data using a spreadsheet program (EXCEL[1]) and a personal computer based statistical package (StatGraphics[2]).

SAMPLE ANALYSIS METHODS

Other workbooks in this series have been prepared covering sampling of airborne particulate matter and the analysis of those samples by nuclear-based analytical methods. The three principal methods are X-Ray Fluorescence (XRF), Ion-Beam Analysis (IBA), and Instrumental Neutron Activation Analysis (INAA). Each of these nuclear methods have advantages and disadvantages and these are summarized in Tables I to 3. Typically the choice of methods is determined largely by the availability of a method. However, if there are choices that can be made, these tables in conjunction with the workbooks can help provide useful information on each method. The different methods have different elemental coverage and those differences can be important in source identification and apportionment. For example, INAA does not provide analyses of S and Pb which are important species in fine particle samples. However, it provides excellent results for Na and Al that can be very useful for coarse particle mass apportionment. the choice of analytical method depends on the problem being studied as well as the nature of the systems available. The data analysis process must take into account the potential limitations of the analytical approaches used to obtain the data.

[1] ©Microsoft Corporation
[2] ©Statistical Graphics Corporation

Table 1. Comparison of the advantages and disadvantages of XRF for airborne particle analysis

Advantages	Disadvantages
1. Nondestructive, sample available for analysis by other techniques	1. No chemical information
2. Minimal sample preparation	2. Filter medium selective - requires homogeneous, thin deposit on non-fibrous filter
3. Analysis time shorter than NAA; varies with spectrometer characteristics,	3. Analysis time longer than PIXE
4. Multielemental (Na to U)[1]	4. Particle size effects for low Z elements[2]
5. Sensitivity and applicability over a wide concentration range (g/g to 100% for many elements)	5. Interelement interference[2]
7. Good accuracy and precision	6. Blind spots due to limitations of excitation mode or overlapping X-ray lines[2]
8. Possibility of semi-quantitative estimation of concentrations	
9. Possibility of automation for unattended analysis of samples	
10. Relatively cheap, does not require a major irradiation facility	
11. IAEA software support	

[1] Analysis in vacuum with low exciting energy can improve detection of low Z elements.

[2] Approaches for correction available.

Table 2. Comparison of the advantages and disadvantages of Ion Beam Analysis methods for airborne particle analysis.

Advantages	Disadvantages
1. Non destructive (low beam currents)	1. No chemical information
2. Minimal sample preparation	2. Requires homogeneous solid samples
3. Quantitative (±3-5% if careful)	3. Cannot easily do liquid and gaseous samples.
4. Multi-elemental (H to U)	4. Some particle size effects for low atomic number materials.
5. Fast, less than 15 minute irradiation times	5. Not cost effective for a few samples
6. Sensitive (µg/g to 100% for many elements)	6. Requires access to an ion beam accelerator (3MV)
7. Can handle small sample sizes (<1mg)	
8. Cost effective for hundreds of samples	

Table 3. Comparison of the advantages and disadvantages of Instrumental Neutron Activation Analysis methods for airborne particle analysis.

Advantages	Disadvantages
1. Non destructive	1. No chemical information
2. Minimal sample preparation	2. Long time to results for many elements
3. Quantitative	3. Samples have residual radioactivity
4. Multi-elemental (Na to U)	4. Not cost effective for a few samples
5. Fast, less than 15 minute irradiation times	5. Requires access to nuclear reactor.
6. Sensitive (µg/g to 100% for many elements)	6. Does not include all elements
7. Can handle small sample sizes (<1mg)	7. Sensitivity varies from element to element

DATA VALIDATION

The process of examining the data obtained in sampling and analysis campaigns such as those performed under the auspices of the IAEA programmes on airborne particulate matter is technically termed data validation. Data validation refers to those activities performed after the fact, that is, after the data have been collected. Data validation and quality control techniques are somewhat different. Quality control methods are employed to minimize the amount of bad data being collected, while data validation seeks to prevent bad data from getting through the data collection and storage process without being recognized as problematic. The idea is to prevent such data from being influential in decision making. Data validation should be performed as soon as possible after data collection so that questionable data can be checked by recalling information on unusual events and on specific meteorological conditions that may have affected the environment being sampled. A detailed document on data validation is available from the U.S. Environmental Protection Agency at http://es.epa.gov/ncerqa/qa/qa_docs.html. This site contains a number of useful documents to describe a quality management system that would be helpful for any environmental monitoring and analysis program.

Also, such validation efforts can help to identify problems that then result in prompt application of corrective actions. Such an iterative process can lead to a higher rate of capture of acceptable quality data. Data validation is the process by which data are filtered and either accepted or flagged for further investigation following a predetermined set of criteria based on the study objectives and the type of data being examined.

Data validation procedures can be conveniently divided into 4 categories:

□ Routine check and review procedures which should be used to some extent in every data validation process,

□ Tests for internal consistency of the data,

□ Tests for consistency with previous data, (historical or temporal consistency), and

□ Tests for consistency with other data sets, collected at the same time or under similar conditions (consistency of parallel data sets).

It is possible to learn a considerable amount about the quality of a data set by examining the data itself. Data validation methods can be used to identify potential problem points that require further examination. These points can arise because of unusual events that produce very high or very low values. It may be that these samples should be excluded from further analysis because they do not represent the normally observed situation. They also may be present because of errors in the sampling and analysis process and examination may permit identification of problems. If the data permit, new values can be calculated after the process has been fixed. Alternatively it may only be possible to insure that future data do not include such errors. The key point is that the data should be examined for their quality before starting to use them for other purposes such as receptor modeling so that time is not spent in analyzing problematic data.

ILLUSTRATIVE EXAMPLE

Through this workbook, examples of how to apply the methods will be presented using a data set that has been reported by Alpert and Hopke (1981). These data are the result of XRF analysis of fine particle samples taken using a dichotomous sampler located on the campus of Washington University in St. Louis, MO. The samples were collected during July and August 1976. The samples were collected in twelve hour increments (midnight to noon and noon to midnight). Only the fine particle (<2.5 μm) samples will be discussed. At this time in the United State, gasoline still contained tetraethyl lead and there were limited controls on point sources of atmospheric particulate emissions. Thus, this set provides an opportunity to illustrate the various methods described in the subsequent sections. These data can be obtained at ftp://ftp.clarkson.edu/pub/hopkepk/IAEA.

PLOTTING

As an initial approach to the review of any data, it is useful to plot the data in a variety of ways. The objective is to begin to identify any "ususal" data points as well as explore the potential relationships among the variables within the data set. Two types of data plots are recommended to be routinely produced; time series plots of each measured.

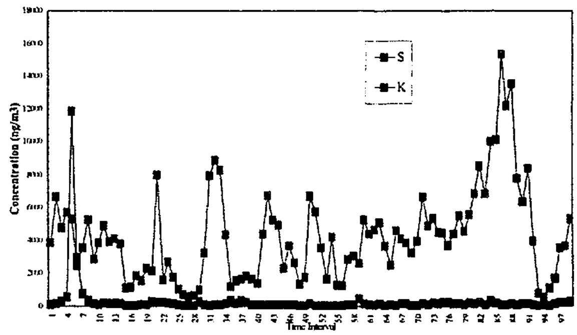

Figure 1. Plot of the sulfur and potassium concentrations measured in St. Louis, MO over the period of July and August 1976.

To illustrate time series plots, Figure 1 provides plots of K and S as a function. It can be seen that in the K data, there are two high values while the rest of the values are very much lower. This behavior is in contrast with the S series which has regular variations over the whole of the time series. When those high points are examined, it is found that they occur for the noon to midnight sample of July 4 and the midnight to noon sample on July 5. These high values are the result of fireworks that were displayed near the sampling site to celebrate the 200^{th} anniversary of the signing of the United States Declaration of Independence. Thus, these two points can be clearly identified as real data. There is no error in the chemical analysis, but they are the result of an unusual and non-recurring source. Thus, although it is good that these high values can be easily identified, it is of little assistance in understanding the normally occurring sources.

A second example of time series plots is shown in Figure 2 where the Pb and Br values are presented. It can be seen that there is a strong similarity in the time patterns for the two elements. The congruence of these two series suggests that these elements are likely to be strongly related to one another and thus, potentially coming from the same emission source.

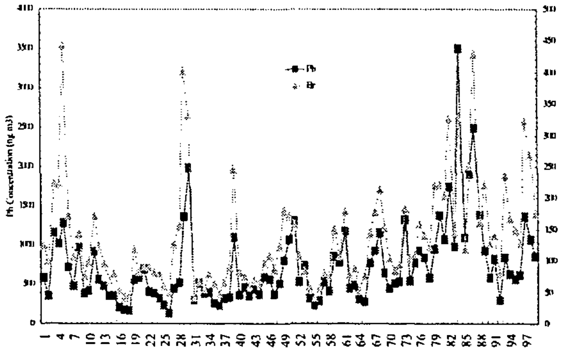

Figure 2. Plot of lead and bromine concentrations measured in St. Louis, MO during July and August 1976.

The relationship between lead and bromine can be further explored by plotting the concentration values in a biplot. In this figure it can be seen that there is generally increasing Br values as Pb increases. However, there is some spread in the relationship suggesting the possibility of more than one source of one or the other of these elements. Such a diagram can help to identify elemental relationships and those will be discussed in the next section.

ELEMENTAL RELATIONSHIPS

There are a variety of relationships that can be used to check if there are significant problems with the data. These problems can be systematic or random. For example, if motor vehicles are burning lead gasoline, there should be a well defined relationship between lead and bromine as seen in Figure 3. The typical Br to Pb ratio for the fuel is around 0.48 which is shown as the solid line in the figure. However, Br is lost from the lead halide particles by reaction with gaseous acids. For example,

$$PbBr_2 + HNO_3 \rightarrow PbBrNO_3 + HBr \qquad (1)$$

Similar reactions can be written for a reaction with sulfuric acid. Thus, typically aged urban aerosol containing leaded gasoline combustion particles show a ratio of the order of 0.30 as can be seen in Figure 3 as the dot-dash line.

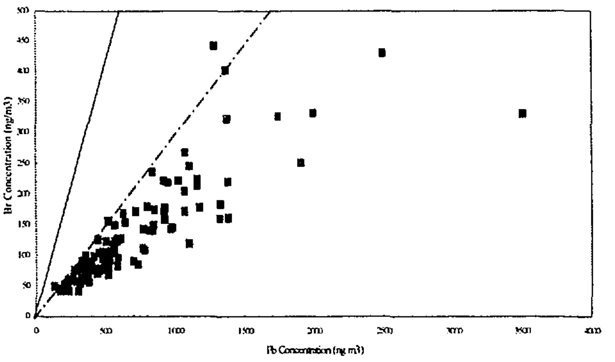

Figure 3. Plot of Br concentration against the Pb concentration for the data from St. Louis, MO in July and August 1976.

A number of similar relationships should typically be observed. For example, aluminum and silicon should be closely related in both coarse and fine sized particles that would reflect the ratio of these elements in local soils. A number of other sources will also produce clearly observed relationships between pairs of elements.

MATERIAL BALANCE ANALYSIS

Material balance is a methodological approach that can combine the elements into composite variables that represent significant sources of airborne particles. It is then possible to examine whether the sum of these composite variables, the reconstructed mass, approximates the measured mass.

For example, sodium and chlorine can be combined to produce the source that we can call "sea salt." Sea salt is produced by wave and sea spray producing droplets that then can evaporate to form solid sea salt particles. The relative elemental composition should then be the same as that of seawater. It has been found to be very useful in several large monitoring programs such as IMPROVE in the United States and ASP in Australia to examine whether or not the measured elemental concentrations are close to the measured mass after applying some reasonable assumptions regarding those elements that are not measured. For example, sea salt is estimated by

$$'Salt' = 2.5 \bullet Na$$

There is an implicit assumption that all of the sodium comes from sea salt. There is some sodium in soil, but it is typically small enough (see Appendix A) that the increment of sodium from soils can be neglected.

There can be displacement of chlorine from the sea salt particles by reaction with gaseous acids like HNO_3 and H_2SO_4 produce through photochemical reactions in the air.

$$NaCl + H_2SO_4 = Na_2SO_4 + 2HCl$$

$$NaCl + HNO_3 = NaNO_3 + HCl$$

Thus, the use of the formula would produce an erroneous estimate of sea salt if these reactions are taking place to a significant extent.

Similarly all of the sulfur that would be detected in samples can generally be assumed to be in the form of ammonium sulfate [$(NH_4)_2SO_4$]. When the particles are collected, the sulfur is generally present as a mixture of ammonium bisulfate and ammonium sulfate. However, given the typical time between sampling and analysis and the access the samples have to ammonia being emitted by laboratory personnel, the samples will be fully neutralized by the time the analyses are completed. Thus, $[(NH_4)_2SO_4] = 4.125 \cdot S$.

The crustal elements can be expected to be present in the airborne particles as oxides. Thus, given measurements for a series of major crustal elements (Al, Si, Ca, Ti, and Fe), it is possible to estimate the concentration of soils based on their typical oxidized forms (Al_2O_3, SiO_2, CaO, TiO_2, and Fe_2O_3). In this approach, the presence of MgO, Na_2O, and K_2O as well as some associated water are included by assuming that the oxides of the other elements account for 86% of the mass (Malm *et al.*, 1994). Using these assumptions leads to an equation of the form:

$$[Soil] = 2.20 \bullet [Al] + 2.49 \bullet [Si] + 1.63 \bullet [Ca] + 2.42 \bullet [Fe] + 1.94 \bullet [Ti]$$

These values work well for the United States, but because iron can be present as both FeO and Fe_2O_3. Also the iron coefficient is inflated to include the contribution of K to fine soil. Thus, there could be some variations in these coefficients.

Cohen has found that biomass burning (bush fires) are a sufficiently important source that it is useful to define a "smoke" variable which is based on the measured potassium value less the crustal potassium.

$$[Smoke] = (K - 0.6 \bullet Fe)$$

Biomass burning also contributes elemental and organic carbon to the aerosol mass. Thus, the equation will necessarily underestimate the contribution of smoke to the total aerosol mass. In addition, this parameter may produce negative values in urban areas where there may be sources emitting extra iron into the air. However, in areas where there is active burning, it can prove useful to estimate the contribution of biomass burning to the aerosol mass.

In ion beam analysis, it is possible to estimate the amount of organic carbon through the PESA determination of the H concentration in the sample. Malm *et al.* (1994) have suggested that organic matter mass can be estimated using the following expression:

$$[OMH] = \frac{1}{f_{OH}}([H] - 0.25[S])$$

where [H] is the concentration of hydrogen in the sample, [S] is the sulfur concentration, and a value of $1/f_{OH} = 11$ was found by forcing [OMH] to equal an alternatively estimated value of organic matter mass concentration. The sulfur is accounted for in the equation because $(NH_4)_2SO_4$ represents a significant contribution to the hydrogen concentration in the sample. It is assumed that there will be relatively low concentrations of NH_4NO_3.

Combining these various pseudoelements permits an estimate of reconstructed fine mass [RCFM].

$$[RCFM] = [sulfate] + [EC] + [OMH] + [SOIL]$$

where [EC] is the elemental carbon as measured by the reflectometer. One can then plot the reconstructed mass against the actual mass values. If there is no IBA available for [H] measurements, then the reconstructed mass will be reduced, but it can still be useful to examine the reconstructed mass.

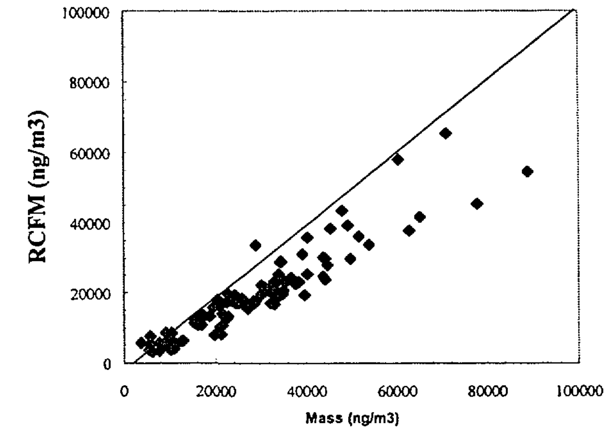

Figure 4. Reconstructed mass against the measured mass for the St. Louis data. The line is drawn to shown the 1:1 relationship.

To illustrate this approach, Figure 4 shows the reconstructed mass against the measured fine mass concentration. These samples wre only analyzed via XRF so there are no EC or OMH data available. Thus, the reconstructed mass substantially underestimates the measured mass values. In this case, we know that motor vehicles contribute a significant amount of airborne particle mass particularly for the fine particle samples. Thus, we can improve the reconstruction of the mass for these samples by estimating the contribution from motor vehicles to the aerosol mass. If source profiles for that period are examined, it was found that lead would typically be 20% of the emitted particle mass. Thus, another pseudoelement, cars, can be defined as

$$[\mathit{Cars}] = 5*[\mathit{Pb}]$$

If this additional summary variable is added to the reconstructed mass, the results shown in Figure 5 are obtained. It can be seen that the addition of this source type has produced relatively good reproduction of the measured mass. Thus, an understanding of the nature of the sources and utilizing the information that is available can still provide an opportunity to examine the mass closure in the analysis.

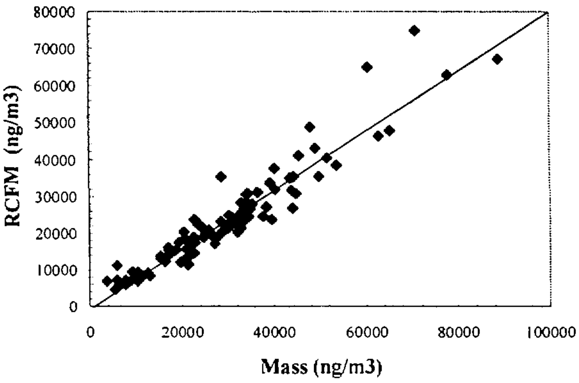

Figure 5. Reconstructed mass plotted as a function of the measured mass for the St. Louis data after the addition of [Cars]. The line is drawn to show the 1:1 relationship.

It is important that consideration be made of the specific types of important sources that may be operating in the area where the samples were collected. This section has provided some general ideas about a set of pseudovariables that have proven useful in other regions. They may be useful in other air sheds, but some thought needs to be included to develop a logical framework for the calculations.

FACTOR ANALYSIS

In studies of the environment, many variables are measured to characterize the system. However, not all of the variables are independent of one another. Thus, it is essential to have mathematical techniques that permit the study of the simultaneous variation of multiple variables. One such analyses is based on examining the relationships between pares of variables. This correlation analysis, however, does not provide a clear view of the multiple interactions in the data. Thus, various forms of eigenvector analysis are used to convert the correlation data into multivariate information. Factor analysis is the name given to one of the variety of forms of eigenvector analysis. It was originally developed and used in psychology to provide mathematical models of psychological theories of human ability and behavior [Harman, 1976]. However, eigenvector analysis has found wide application throughout the physical and life sciences. Unfortunately, a great deal of confusion exists in the literature in regard to the terminology of eigenvector analysis. Various changes in the way the method is applied has resulted in it being called factor analysis, principal components analysis, principal components factor analysis, empirical orthogonal function analysis, Karhunen-Loeve transform, etc., depending on the way the data are scaled before analysis or how the resulting vectors are treated after the eigenvector analysis is completed.

All of the eigenvector methods have the same basic objective; the compression of data into fewer dimensions and the identification of the structure of interrelationships that exist between the variables measured or the samples being studied. In many chemical studies, the measured properties of the system can be considered to be the linear sum of the term representing the fundamental effects in that system times appropriate weighing factors. For example, the absorbance at a particular wavelength of a mixture of compounds for a fixed path length, z, is considered to be a sum of the absorbencies of the individual components

$$\frac{a(\lambda)}{z} = \epsilon_1 C_1 + \epsilon_2 C_2 + \ldots + \epsilon_p C_p \qquad (10)$$

where ε_i is the molar extinction coefficient for the ith compound at wavelength λ and C_i is the corresponding concentration. Thus, if the absorbencies of a mixture of several absorbing species are measured at m various wavelengths, a series of equations can be obtained.

$$\frac{a(\lambda_j)}{z} = \sum_{k=1}^{p} \epsilon(\lambda_j)_k C_k \qquad (11)$$

If we know what components are present and what the molar extinction coefficients are for each compound at each wavelength, the concentrations of each compound can be determined using a multiple linear regression fit to these data. However, in many cases neither the number of compounds nor their absorbance spectra may be known. For example, several compounds may elute from an HPLC column at about the same retention time so that a broad elution peak or several poorly resolved peaks containing these compounds may be observed. Thus, at any point in the elution curve, there would be a mixture of the same components but in differing proportions. If the absorbance spectrum of each of these different mixtures could be measured such as by using a diode array system, then the resulting data set would consist of a number of absorption spectra for a series of n different mixtures of the same compounds.

$$\frac{a_i(\lambda_j)}{z} = \sum_{k=1}^{p} \epsilon(\lambda_j)_k C_{ki} \qquad \begin{matrix} j = 1,m \\ i = 1,n \end{matrix} \qquad (12)$$

For such a data set, factor analysis can be employed to identify the number of components in the mixture, the absorption spectra of each component, and the concentration of each compound for each of the mixtures. Similar problems are found throughout analytical and environmental chemistry where there are mixtures of unknown numbers of components and the properties of each pure component are not known *a priori.*

Another use for such methods is in physical chemistry where the measured property can also be related to a linearly additive sum of independent causative processes. For example, the effects of solvents on the proton NMR shifts for non-polar solutes can be expressed in a form of

$$d_{ia} = \sum_{j=1}^{p} U_{ij} V_{ja} \tag{13}$$

where d_{ia} is the chemical shift of solute i in solvent a, U_{ij} refers to the jth solute factor of the ith solvent, and V_{ja} refers to the jth solvent factor of the ath solvent with the summation over all of the physical factors that might give rise to the measured chemical shifts. Similar examples have been found for a variety of chemical problems and are described by Malinowski [1991].

Finally, similar problems arise in the resolution of environmental mixtures to their source contributions. For example, a sample of airborne particulate matter collected at a specific site is made up of particles of soil, motor vehicle exhaust, secondary sulfate particles, primary emissions from industrial point sources, etc. It may be of interest to determine how much of the total collected mass of particles comes from each source. It is then assumed that the measured ambient concentration of some species, x_i, where i=1, m measured elements, is a linear sum of contributions from p independent sources of particles. These species are normally elemental concentrations such as lead or silicon and are given in ng of element per cubic meter of air. Each kth source emits particles that have a profile of elemental concentrations, a_{ik}, and the mass contribution per unit volume of the kth source is f_k. When the compositions are measured for a number of samples, an equation of the following form is obtained.

$$z_{ij} = \sum_{k=1}^{p} a_{ik} f_{kj} \tag{14}$$

The use of factor analysis for this type of study has been reviewed by Hopke [1985].

The factor analysis problem can be visualized with the following example. Suppose a series of samples are taken in the vicinity of a highway where motor vehicles are using leaded gasoline and a steel mill making specialty steels. For these samples, measurements of Pb, Br, and Cr are made. This set of data can then be plotted in a three dimensional space as in Figure 6. A cloud of points can be observed.

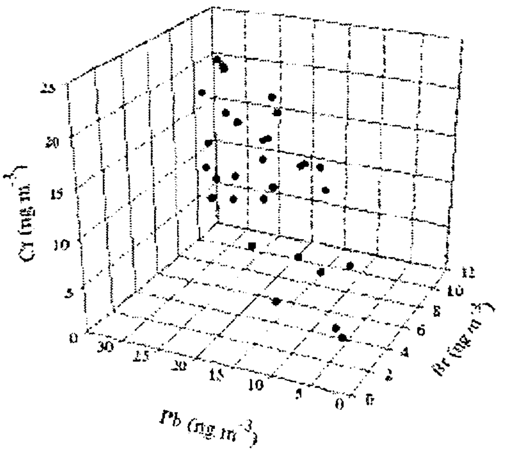

Figure 6Three dimensional plot of simulated data.

However, it is known that there are only two particle sources. The problem is then to determine the true dimensionality of the data and the relationships among the measured variables. That is goal of a factor analysis. In the case of this example, the relationships can be observed with a simple rotation of the axes so that we look down onto the figure so that the Cr axis sticks out of the page. This view is seen in Figure 7. Now it can be seen that the data really cluster around a line that represents the Pb-Br relationship in the particles emitted by the motor vehicles. The Cr values are distributed vertically and are independent of the other two elements. Factor analysis of this problem would find two sources and provide the relationship between the lead and bromine.

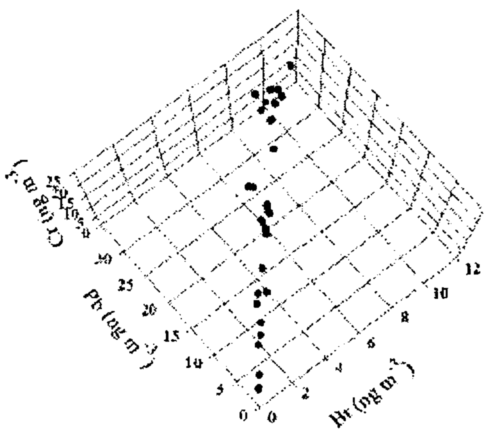

Figure 7Plot of the simulated data as viewed from above relative to the view in Figure 1.

Principal Component Analysis

The most common form of factor analysis is Principal Components Analysis (PCA). This method is generally available in most computer packages for statistical analysis. Since PCA can only be performed on a set of samples in which the various sources contribute different amounts of particles to each sample, the mass balance needs to be expanded to a matrix equation.

$$Z = A \cdot F \tag{15}$$

where Z is a matrix of sample vectors, A is the matrix of loading vectors related to the source compositions, and F is the matrix of scores that are related to the contribution of that source type to the variance of that particular measured variable. The mathematical details of PCA are given in Appendix A.

The general result of a PCA analysis is illustrated in Figure 8. In a PCA analysis, the data are normalized by subtracting a mean value, $\bar{x}$ and divided by the standard deviation, σ.

$$z_{ij} = \frac{x_{ij} - \bar{x}}{\sigma}$$

In this case each standardized variable has a mean value of zero and a standard deviation of 1. The variance of a variable is given by σ^2. The line representing the maximum variance for these two variables is shown in Figure 8 as well as a line that is orthogonal to the first that will reproduce the remaining variance in the variables.

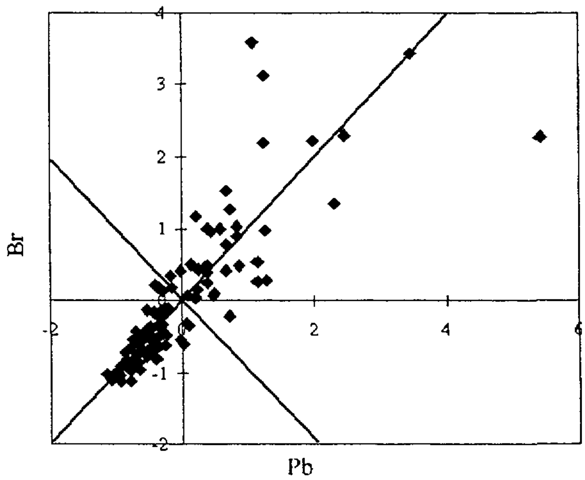

Figure 8. Plot of the standardized values of the same values of bromine and lead as shown in Figure 3. The thin lines depict the original axes. The heavy lines depict the directions of maximum variance.

PCA can produce components that have a strong relationship with only one variable. Since it is quite reasonable for many environmental systems to show factors that produce such single variable behavior, it is advisable to use a principal components analysis and extend the number of factors to those necessary to reproduce the original data within the error limits inherent in the data set.

By design, the eigenvector analysis compresses the information content of the data set into as few eigenvectors as possible. Thus, in considering the number of factors to be used to describe the system, it is necessary to carefully examine the problems of reconstructing both the variability within the data and reconstructing the actual data itself. Following the diagonalization of the correlation or covariance matrix, it is necessary to make the difficult choice of the number of factors, p, to use in the subsequent analysis. This problem occurs in any application of an eigenvector analysis of data containing noise. In the absence of error, the eigenvalues beyond the true number of sources become zero except for calculational error. The choice becomes more difficult depending on the error in the data. Several approaches have been suggested [Duewer *et al.*, 1976; Hopke *et al.*, 1980].

A large relative decrease in the magnitude of the eigenvalues is one indicator of the correct number of factors. If the data were without error, then there would be a clear distinction between factors with non-zero eigenvalues and those whose values were zero. It can often be useful to plot the eigenvalues as a function of factor number and look for sharp breaks in the slope of the line [Cattell, 1966]. If the eigenvalue is a measure of the information content of the corresponding eigenvector, then only sufficiently "large" eigenvalues need to be retained in order to reproduce the variation initially present in the data. One of the most commonly used and abused criteria for selecting the number of factors to retain is retaining only those eigenvalues greater than 1 [Guttman, 1954]. The argument is made that the normalized variables each carry one unit of variance. Thus, if an eigenvalue is less than one, then it carries less information than one of the initial variables and is therefore not needed. However, Kaiser and Hunka [1973] make a strong argument that although eigenvalue greater than one does set a *lower* limit on the number of factors to be retained, it does not set a simultaneous upper bound. Thus, there must be at least as many factors as there are eigenvalues greater than one, but there can be more than that number that are important to the understanding of the system's behavior.

Hopke [1982] has suggested a useful empirical criterion for choosing the number of retained eigenvectors. In a number of cases of airborne particulate matter composition source identification problems, Hopke found that choosing the number of factors containing variance greater than one *after* an orthogonal rotation provided a stable solution. Since the eigenvector analysis artificially compresses variance into the first few factors, reapportioning the variance using the rotations described in the next section will result in more factors with total variance greater than one than there are eigenvalues greater than one. In many cases, this number of factors will stay the same even after rotating more factors.

Because the matrices produced by the diagonalization process have been calculated in a way to maximize the amount of variance contained in each factor, they may not reflect the pattern of variables representative of a particle source. Thus, the factors are rotated generally to achieve what is termed "simple structure" [Hopke *et al.*, 1976]. A varimax rotation is most commonly used in such analyses. More details of factor axis rotations are given in Appendix B.

Application of PCA to Illustrative Data

In the St. Louis data set, a total of 124 samples were obtained in each the fine and coarse fractions. A summary of the data is presented in Table 4. Data were missing for 24 pairs of samples leaving a total of 100 pairs of coarse and fine fraction samples. In order to produce the best possible source resolutions, it is vital to have accurate measurements of the particulate mass as well as determinations for as many elements as possible. Of the twenty-seven elements determined for each of the 100 remaining samples, over 50% of the determinations of ten elements had values that were missing or below the detection limits. Since a complete and accurate data set is required to perform a factor analysis, these ten elements were eliminated from the analysis. No element for which more than about 33% of the values are below detection limits should be included in the analysis. A value is required for each value including those values that are below the detection limits. A good approach to providing these values are to use a uniform random number generator to provide a number between zero and the detection limit. In EXCEL©, the function, =RAND(), can be used to provide this value. Thus, multiplying this function by the detection limit will provide a useful value. However, it is important to convert these results into values since each time the spreadsheet is opened will result in a different value. Thus, after all of the below detection limit values have been replaced, the block of data is selected. Under the "Edit" menu, select "Copy". Then again in "Edit", select "Paste Special" and check "Values". This function will replace all of the formulae in the data block with the current values. If this worksheet is then saved, it will be the same each time it is opened.

The exclusion of elements leads to a loss of information. In this example, arsenic was excluded because almost all of the values were below the detection limits. Arsenic determinations by x-ray fluorescence are often unreliable because of an interference between the arsenic K x-ray and the lead L x-ray. A neutron activation analysis of these samples would produce better arsenic determinations. Reliable data for arsenic may be important to the differentiation of coal flyash and crustal material; these two materials have very similar source profiles. However, a low percentage of measured values for species can lead to distortions in the resulting analysis.

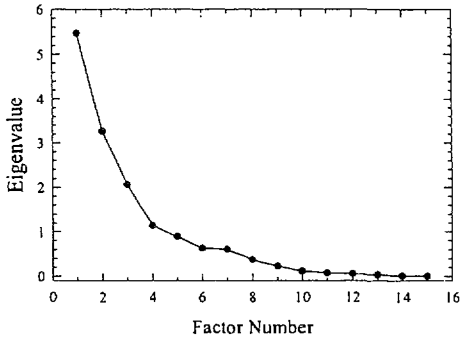

Figure 9. Eigenvalues for the St. Louis data set as a function of factor number.

The eigenvalues for this data set are given in Table 4. There are 4 factors with eigenvalue greater than 1.0. However, there are values of 0.8. 0.6. and 0.6. A plot of these eigenvectors as a function of factor number is given in Figure 9. It can be seen that there is a break in the curve at 7 factors and then a smooth decrease from factor 8 up to factor 15. There are at least 4 factors, but it appears likely that there will be 7 factors with total valiance greater than 1 after rotation.

The varimax rotated loadings are presented in Table 5. The loadings can be considered to be a correlation coefficient between the original variables with the new factor variables that represent a linear combination. However, it is not appropriate to use tables of statistical significance for correlation coefficients because those significance values are based on an assumption that the variables being correlated are each normally distributed. Environmental data show positively skewed, heavily tailed distributions and thus, do not fit the underlying assumption. A useful rule of thumb is that loading values over 0.50 are generally considered significant although smaller values may still provide insights into the nature of the sources. The total variance in each factor is calculated as the sum of the squared loadings, a_{ij}^2, for the given factor. Each of these seven factors have variance greater than 1 after rotation. It is now necessary to interpret the pattern of elements for each factor.

Table 4. Eigenvalues for PCA of July-August 1976 Fine Fraction Data from RAPS Site 112

Factor Number	Eigenvalue
1	5.472
2	3.265
3	2.064
4	1.143
5	0.892
6	0.633
7	0.607
8	0.377
9	0.228
10	0.116
11	0.080
12	0.059
13	0.041
14	0.014
15	0.009

The first factor is unusual in that it has high values for Al, K, Ca, and Sr. It is not obvious what might be the source of such an unusual combination of elements. In this case, the examination of the corresponding scores, f_{kj}, can prove helpful. A time series plot of the scores looks identical to the plot of the K values in Figure 1. The score values are low with the exception of 2 samples. These scores are in standardized space. Thus, a score of 0 means that the is an average contribution of the given source to the given sample. The large score suggests a specific source that was only operating during these two sampling period. These samples were taken during the period of noon to midnight on July 4 and midnight to noon on July 5 and again is bicentennial fireworks. We can compare the average values for the measure variables in the complete data sets and in a subset in which the samples for July 4 and 5 have been deleted. This comparison is provided in Table 6.

Table 5. Rotated Factor Loading Matrix for the Site 112 Fine Fraction Data

	Factor Number						
Element	1	2	3	4	5	6	7
Al	**0.619**	**0.766**	-0.040	0.049	-0.085	-0.054	-0.019
Si	0.067	**0.981**	0.000	-0.034	-0.060	-0.097	-0.052
S	-0.040	-0.041	0.183	0.118	**0.957**	0.104	0.049
Cl	0.248	-0.016	0.187	0.473	**0.735**	0.265	0.137
K	**0.982**	-0.002	0.034	0.075	0.055	0.078	-0.021
Ca	**0.827**	0.379	0.062	0.226	0.001	0.042	0.178
Ti	0.062	0.115	0.118	0.169	0.096	-0.003	**0.955**
Cr	0.143	0.012	**0.778**	0.271	0.177	0.220	0.201
Mn	-0.022	0.137	**0.885**	0.143	0.144	0.240	0.004
Fe	-0.048	**0.830**	0.243	0.124	0.087	0.125	0.419
Cu	0.173	0.016	0.159	0.239	0.130	**0.859**	-0.070
Zn	-0.053	-0.086	0.439	0.087	0.159	**0.762**	0.112
Br	0.295	0.061	0.179	**0.872**	0.154	0.182	0.186
Sr	**0.984**	-0.018	0.023	0.068	0.055	0.051	-0.004
Pb	0.010	0.044	0.446	**0.747**	0.367	0.238	0.097
Variance	3.214	2.429	1.986	1.823	1.742	1.634	1.246

Table 6. Comparison of Data With and Without Samples from July 4th and 5th (ng/m^3) RAPS Station 112, July and August 1976, Fine Fraction.

Element	With July 4th & 5th Mean	Without July 4th & 5th Mean[a]
Al	220 ± 30	200 ± 30
Si	440 ± 60	450 ± 60
S	4370 ± 310	4360 ± 320
Cl	90 ± 10	80 ± 9
K	320 ± 130	150 ± 9
Ca	110 ± 10	110 ± 10
Ti	63 ± 13	64 ± 13
Mn	17 ± 3	17 ± 3
Fe	220 ± 20	220 ± 20
Ni	2.3 ± 0.2	2.3 ± 0.2
Cu	16 ± 3	15 ± 3
Zn	78 ± 8	75 ± 8
Se	2.7 ± 0.2	2.7 ± 0.2
Br	140 ± 9	130 ± 8
Sr	5 ± 4	1.1 ± 0.1
Ba	19 ± 5	15 ± 4
Pb	730 ± 50	720 ± 50

Table 7. Eigenvalues for PCA of July-August 1976 Fine Fraction Data from RAPS Site 112 without the Fireworks Samples.

Factor Number	Eigenvalue
1	5.495
2	3.746
3	1.446
4	0.950
5	0.865
6	0.747
7	0.581
8	0.326
9	0.296
10	0.238
11	0.125
12	0.080
13	0.057
14	0.036
15	0.012

Table 8. Rotated Factor Loading Matrix for the Site 112 Fine Fraction Data without Fireworks Samples

Element	1	2	3	4	5	6	7
Al	**0.970**	-0.060	-0.110	-0.024	0.118	-0.053	-0.040
Si	**0.976**	0.004	-0.030	-0.075	0.095	-0.055	-0.080
S	--0.072	0.184	**0.956**	0.097	0.008	0.065	0.046
Cl	-0.034	0.198	**0.786**	0.462	0.121	0.210	0.135
K	0.284	0.169	0.081	0.079	**0.889**	0.141	-0.090
Ca	**0.706**	0.037	-0.065	0.360	0.239	-0.013	0.295
Ti	0.141	0.126	0.111	0.171	0.019	-0.022	**0.948**
Cr	0.035	**0.801**	0.183	0.250	-0.004	0.210	0.213
Mn	-0.094	**0.887**	0.156	0.149	0.200	0.120	-0.000
Fe	**-0.800**	0.275	0.131	0.105	0.161	0.104	0.375
Cu	-0.006	0.228	0.147	0.192	0.082	**0.918**	-0.030
Zn	--0.208	0.467	0.179	0.131	**0.524**	**0.526**	0.082
Br	0.082	0.195	0.188	**0.913**	0.091	0.145	0.172
Sr	**0.554**	-0.007	-0.042	0.046	**0.708**	-0.041	0.234
Pb	0.032	0.474	0.404	**0.716**	0.038	0.197	0.072
Variance	3.507	2.159	1.897	1.892	1.744	1.325	1.305

It is clear that the presence of these 2 samples within a set of 99 has created a single factor. Such a potent source could significantly distort the whole analysis. Since this source is unique to this particular occasion and it is the average properties of the system that are of interest, another analysis with these samples deleted is performed. The eigenvalue results for this analysis is presented in Table 7 and the loadings are presented in Table 8. Now there are only 3 eigenvalues greater than 1, but there still appear to be up to 7 factors. To examine the best model, each of the possible solutions including 5, 6, and 7 factors need to be examined to see which makes the most physical and chemical sense and which provides a reasonable overall fit to the data. An 8 factor solution is clearly one with too many factors since the 8th factor does not have any loading with a significant value. After examining these solutions, the 7 factor solution is presented in Table 8. The key is to be able to interpret them in terms of likely emission sources.

The first factor is crustal elements with the exception of iron. We cannot distinguish between soil or flyash based on the elements that are available so it could be either or a combination of both source types. The Fe loading has a high negative value meaning that when there is significant soil/flyash input to the samples, there is a significantly decreased iron concentration. There is a large iron/steel complex to the northeast of the site in Granite City as discussed earlier which would provide high Fe values as well as several iron/steel works on the Missouri side of the Mississippi River. Major sources of flyash or soil generally lie in other directions and thus, the samples for which there is high Fe from the ferrous metal works are those which there is low values of the other crustal species and vice-versa.

The second factor has high loadings for Cr and Mn. These elements would be expected to be related to iron/steel works. It would have been more logical if Fe had also had a high positive loading on this factor, but it appears that the negative correlation with the other crustal elements dominates over any correlation between Fe and these other ferrous metals. The third factor is regional sulfate, but again there is a curious additional loading from Cl. It would be expected that S is poorly correlated with the primary particulate emissions since local emissions of S would be in the form of SO_2. It takes time to convert the SO_2 into $SO_4^=$ and the dry deposition rate of SO_2 is much more rapid than that of particulate sulfate.

The fourth factor is clearly motor vehicles. These samples were taken in 1976 when leaded gasoline was still in widespread use in the United States. The lead variance is also partially seen in the second factor indicating that the steel complex being observed is probably the one in Granite City where a secondary lead smelter was in operation. In the 5th factor, Sr, Zn, and K are featured. Although there is no correlation with Pb, this factor is probably an indication of refuse incinerators. So much of the lead variance is associated with motor vehicles and the Granite City source area, that it does not appear here. Potassium and strontium are strongly related to the combustion of paper particularly when it would be expected that there would be no wood combustion in the area.

The 6th factor is related to Cu and Zn. There are two significant sources of non-ferrous metals in the Sauget, IL area to the east of site 105. It has been found [Rheingrover and Gordon, 1988] that the Cu plant is a source of fine particles while the Zn smelter is a source of larger particles. Thus, we assign this factor to being the Cu products plant. The final factor has a high loading only for Ti. There was a large paint pigment plant that produces TiO_2 and is a major source of fine particle titanium in the area.

Quantitative Factor Analysis Methods

Thus, the factor analysis can identify major source types and examine some of the relationships among the measured compositions. However, it does not provide a quantitative apportionment in the same way that a CMB analysis can. The problem is that the data are centered and scaled before they are analyzed. It is possible to unscale them and them uncenter them to produce an Absolute Principal Components Analysis (APCA) using the procedure outlined by Thurston and Spengler [1985]. This approach has been successfully applied in a number of cases [Maenhaut and Cafmeyer, 1987; Andrade *et al.*, 1993]. However, it does involve centering of the data and uncentering of factor scores which propogates the uncertainties in the mean values twice. Thus, other methods have been developed that avoid centering the data and thus, does not have to transpose the results back to the true zero in the system.

One approach that avoids centering and then uncentering is Target Transformation Factor Analysis that has been discussed by Hopke [1989; 1995]. Although it has been applied to a number of data sets, it has been replaced with better factor analysis methods. Another approach to providing quantitative source resolutions is Source Apportionment by Factors with Explicit Constraints (SAFER). In this approach, the space in which the factor solution lies is truncated by imposing one or more of the natural physical constraints discussed above. SAFER has been described by Henry and Kim [1989] and by Henry [1991]. In addition to the natural constraints, it is also possible to incorporate additional constraints based on additional knowledge of the system. If the range in which the concentration of an element must lie can be estimated, then these limits can be added to the analysis. SAFER has been applied to interpretation of airborne organic compound data from Atlanta, Georgia [Henry *et al.*, 1994]. In this study, they extracted the vehicle-related hydrocarbon source compositions from the ambient sample analytical results with quite reasonable results.

Another new approach to factor analysis is Positive Matrix Factorization (PMF). In order to properly scale the data, it is necessary to look explicitly at the problem as a least-squares problem. To begin this analysis, the elements of the "Residual Matrix," E are defined as

$$e_{ij} = x_{ij} - \hat{x} = x_{ij} - \sum_{k=1}^{p} f_{ik} \cdot g_{kj} \qquad (19)$$

where i =1,..., m elements

j =1,..., n samples

An "object function," Q, which is to be minimized as a function of G and F is given by

$$Q(E) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\frac{e_{ij}}{s_{ij}} \right]^2 \qquad (20)$$

where s_{ij} is an estimate of the “uncertainty” in the ith variable measured in the jth sample. The factor analysis problem is then to minimize Q(E) with respect to G and F with the constraint that each of the elements of G and F is to be non-negative.

PMF was initially applied to data sets of major ion compositions of daily precipitation samples collected over a number of sites in Finland [Juntto and Paatero, 1994] and samples of bulk precipitation [Anttila *et al.*, 1995] in which they are able to obtain considerable information on the sources of these ions. Polissar *et al.* [1996] applied the PMF2 program to Arctic data from 7 National Park Service sites in Alaska as a method to more quantitatively resolve the major source contributions.

Recently there has been a series of applications of PMF to various source/receptor modeling problems. Polissar *et al.* [1998] have reanalyzed an augmented set of Alaskan NPS data and resolved up to 8 sources. Xie *et al.* [1999a&b] have made several analyses of data from an 11 year series of particulate matter samples taken at Alert, N.W.T. Polissar *et al.* [1999] have examined the semicontinuous aerosol data collected by NOAA at their atmospheric observatory at Barrow, Alaska.. Lee *et al.* [1999] have applied PMF to urban aerosol compositions in Hong Kong. They were able to identify up to 9 sources that provided a good apportionment of the airborne particulate matter. Paterson *et al.* [1999] applied PMF to air quality and temperature data collected at a series of sites around the southern end of Lake Michigan in 1997 and used three factors to reproduce 75% of the variation in the data. Huang *et al.* [1999] analyzed elemental composition data for particulate matter samples collected at Narragansett, R.I. using both PMF and conventional factor analysis. They were able to resolve more components with more physically realistic compositions with PMF. Thus, the approach is gaining interest because it does have some inherent advantages particularly through its ability to individually weight each data point. PMF is someone more complex and harder to use, but it appears to provide improved resolution of sources and better quantification of impacts of those sources.

CHEMICAL MASS BALANCE ANALYSIS

It is also possible to calculate the source contributions for individual samples if the number and the nature of those samples are known. Using source profiles derived by Alpert and Hopke (1981), the methodology can be demonstrated. This analysis can be performed using the regression capability in EXCEL®. It is essential that the Analysis Toolbox has been installed. It is not part of the default installation and thus, must explicitly be installed.

It is essential to have good estimates of the source profiles. There are source profiles available as a self-extracting zip file, SPECIATE.EXE from the U.S. Environmental Protection Agency at http://www.epa.gov/ttn/chief/software.html. Because there are errors in the measurement of the source profiles, the US Environmental Protection Agency prefers the use of a model that incorporates these errors in the analysis using effective variance least squares (EVLS) [Watson *et al.*, 1990; Watson *et al.*

1991]. Cheng *et al.* (1988) have suggested that in many cases, the value of including these errors in the analysis has relatively little effect on the results. However, if the EVLS analysis is desired, the DOS program CMB7 that implements the algorithm can be downloaded from http://www.epa.gov/ttn/scram/t22.html. A new Windows-compatible program, CMB8, is supposed to be released in the spring of 2000.

A CMB analysis is performed on a vector representing the composition of an individual sample, **x**, using a matrix of source compositions, **A**. The model can be written as

$$\mathbf{x} = \mathbf{Af} \tag{20}$$

Multiplying both sides by the diagonal weight matrix, **W**, gives

$$\mathbf{Wx} = \mathbf{WAf} \tag{21}$$

where

$$\mathbf{W} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\sigma_3^2} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \frac{1}{\sigma_m^2} \end{bmatrix} \tag{22}$$

Multiplying both sides by the transpose of the source profile matrix, $\mathbf{A}^t$, yields

$$\mathbf{A^tWx} = \mathbf{A^tWAf} \tag{23}$$

Solving for the source contribution vector, **f**, is possible by multiplying both sides by the inverse of $\mathbf{A^tWA}$

$$(\mathbf{A^tWA})^{-1}\mathbf{A^tWx} = \mathbf{f} \tag{24}$$

which is an ordinary weighted least squares solution. In EXCEL, there is no direct provision for weighted regression. However, it can be easily accomplished by multiplying each value along a row of the source profile matrix and the corresponding value in the **x** vector by the weight, $1/\sigma^2$. To illustrate this approach, a single sample and the appropriate set of source profiles is given in Table 9 while the weighted source profile matrix is given in Table 10. These source profiles were taken from Albert and Hopke (1981).

Table 9. Source Profiles and Data for the EXCEL CMB Analysis

Element	Motor Vehicles	Sulfate	Flyash/soil	Paint	Refuse	Limestone	112184-12	Error	Weight
Al	5000	1100	53000	0	0	30000	108.2	47	0.000
Si	0	1900	130000	0	7000	150000	206.1	49	0.000
S	20	240000	19000	6000	0	0	3859.6	339	0.000
Cl	2400	1100	0	4600	22000	16000	29.9	5	0.040
K	1400	1600	15000	5700	48000	15000	121.1	13	0.006
Ca	11000	0	16000	34000	1200	188000	60.8	11	0.008
Ti	0	700	2500	110000	0	1500	60.9	7	0.020
Mn	0	0	700	4800	8600	1600	7.5	2	0.250
Fe	0	1100	36000	90000	36000	34000	91	5	0.040
Ni	80	40	42	11	700	200	0.8	2	0.250
Cu	600	10	0	0	8700	600	8.3	1	1.000
Zn	800	0	0	3700	65000	4200	54.9	3	0.111
Se	100	100	1	200	200	20	2.9	1	1.000
Br	30000	30	2500	0	50	3100	124.7	6	0.028
Sr	90	10	150	100	5	300	0.6	1	1.000
Ba	700	50	70	28000	500	700	12.9	16	0.004
Pb	107000	6500	5000	0	46000	11000	588.5	30	0.001

The solution is them obtained using the data analysis tool, REGRESSION. In this procedure, the weighted measured concentration column is the dependent variable block. The weighted source profiles are the independent block. The intercept should be set to be 0. Part of the resulting output tables are presented in Tables 11 and 12. It can be seen that there is a high quality fit, but there is a negative coefficient for the limestone source. There is also low t-values for both the limestone source and flyash/soil. Thus, one repeats the analysis with that source omitted.

The results for the five source fit are given in Tables 13 and 14. All of the coefficients are positive, but the value for flyash/soil is statistically insignificant since the standard error is twice the source contribution value. Hence the analysis is performed one more time without

Table10. Weighted Source Profiles and Fine Particle Sample Composition

Element	Motor Vehicles	Sulfate	Flyash/Soil	Paint Pigment	Refuse Incineration	Limestone	112184F
Al	1.100e+07	2.430e+06	1.171e+08	0.000e+00	0.000e+00	6.627e+07	2.390e+05
Si	0.000e+00	4.562e+06	3.121e+08	0.000e+00	1.700e+07	3.602e+08	4.948e+05
S	2.298e+06	2.758e+10	2.184e+09	6.895e+08	0.000e+00	0.000e+00	4.435e+08
Cl	6.000e+04	2.750e+04	0.000e+00	1.150e+05	5.500e+05	4.000e+05	7.475e+02
K	2.366e+05	2.704e+05	2.535e+06	9.633e+05	8.112e+06	2.535e+06	2.047e+04
Ca	1.331e+06	0.000e+00	1.936e+06	4.114e+06	1.452e+05	2.275e+07	7.357e+03
Ti	0.000e+00	3.430e+04	1.225e+05	5.390e+06	0.000e+00	7.350e+04	2.984e+03
Mn	0.000e+00	0.000e+00	2.800e+03	1.920e+04	3.440e+04	6.400e+03	3.000e+01
Fe	0.000e+00	2.750e+04	9.000e+05	2.250e+06	9.000e+05	8.500e+05	2.275e+03
Ni	3.200e+02	1.600e+02	1.680e+02	4.400e+01	2.800e+03	8.000e+02	3.200e+00
Cu	6.000e+02	1.000e+01	0.000e+00	0.000e+00	8.700e+03	6.000e+02	8.300e+00
Zn	7.200e+03	0.000e+00	0.000e+00	3.330e+04	5.850e+05	3.780e+04	4.941e+02
Se	1.000e+02	1.000e+02	1.000e+00	2.000e+02	2.000e+02	2.000e+01	2.900e+00
Br	1.080e+06	1.080e+03	9.000e+04	0.000e+00	1.800e+03	1.116e+05	4.489e+03
Sr	9.000e+01	1.000e+01	1.500e+02	1.000e+02	5.000e+00	3.000e+02	6.000e-01
Ba	1.792e+05	1.280e+04	1.792e+04	7.168e+06	1.280e+05	1.792e+05	3.302e+03
Pb	9.630e+07	5.850e+06	4.500e+06	0.000e+00	4.140e+07	9.900e+06	5.297e+05

Table 11. Summary statistics from EXCEL for the 6 source solution.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9937
R Square	0.9875
Adjusted R Square	0.8909
Standard Error	0.3185
Observations	17

Table 12. Results of the CMB analysis with all 6 sources

	Coefficients	Standard Error	t Stat	P-value
Intercept	0			
Motor Vehicle	0.0039	0.00036	10.89	3E-07
Sulfate	0.0198	0.00297	6.66	4E-05
Flyash/Soil	0.0004	0.00046	0.85	0.4118
Paint Pigment	0.0004	0.00014	2.88	0.015
Refuse Incineration	0.0007	3.5E-05	20.57	4E-10
Limestone	-2E-04	0.00021	-0.93	0.3733

Table 13. Summary statistics from EXCEL for the 5 source solution.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9932
R Square	0.9865
Adjusted R Square	0.8987
Standard Error	0.3167
Observations	17

Table 12. Results of the CMB analysis with 5 sources

	Coefficients	Standard Error	t Stat
Intercept	0		
Motor Vehicles	0.003853	0.000353	10.926240
Sulfate	0.019490	0.002937	6.636182
Flyash/Soil	0.000210	0.000413	0.508394
Paint Pigment	0.000410	0.000142	2.895766
Refuse Incineration	0.000703	0.000032	21.811297

A similar analysis can be performed even more easily using StatGraphics since it is easier to define the individual sources to be included in the model. With EXCEL, the sources being included in the analysis must be in a rectangular block. Thus, since the flyash source appears in the middle of the block, it would have to be deleted or moved to the edge of the block in order to perform the 4 source analysis.

REFERENCES

Alpert, D.J. and P.K. Hopke (1981) A Determination of the Sources of Airborne Particles Collected During the Regional Air Pollution Study, *Atmospheric Environ.* 15:675-687.

Andrade, F., C. Orsini, and W. Maenhaut (1993) Receptor Modeling for Inhalable Particles in Sao Paulo, Brazil, *Nucl. Instr. Methods* B75:308-311.

Anttila, P., P. Paatero, U. Tapper, and O. Järvinen (1995) Application of Positive Matrix Factorization to Source Apportionment: Results of a Study of Bulk Deposition Chemistry in Finland, *Atmospheric Environ.* 29:1705-1718.

Cattell, R.B. (1966) *Handbook of Multivariate Experimental Psychology*, Rand McNally, Chicago, pp. 174-243.

Cheng, M.D., P.K. Hopke, and D. Jennings (1988) The Effects of Measurement Errors, Collinearity and Their Interactions on Aerosol Source Apportionment Models, *Chemom. Intell. Lab. Syst.* 4:239-250.

Chueinta, W., P.K. Hopke and P. Paatero (2000) Investigation of Sources of Atmospheric Aerosol Urban and Suburban Residential Areas in Thailand by Positive Matrix Factorization, *Atmospheric Environ.* In press.

Cohen, D.D. (1997) Ion Beam Analysis Methods in Aerosol Analysis, Quality Assurance and Inter-Technique Comparisons, in *Harmonization of health Related Environmental Measurements using Nuclear and Isotopic Techniques*, International Atomic Energy Agency, IAEA-SM-344/23, 457-472.

Duewer, D.L., B.R. Kowalski, and J.L. Fasching (1976) Improving the Reliability of Factor Analysis of Chemical Data by Utilizing the Measured Analytical Uncertainty, *Anal. Chem.* 48:2002-2010.

Exner, O., Additive Physical Properties, *Collection of Czech. Chem. Commun.*, 31:3222-3253 (1966).

Goldberg, E.D. (1963) Chemistry - the Oceans as a Chemical System in *The Sea*, H.M. Hill, ed., Wiley-Interscience, New York, pp 3-25.

Guttman, L. (1954) Some Necessary Conditions for Common Factor Analysis, *Psychometrika* 19:149-161.

Henry, R.C. (1991) Multivariate Receptor Models, In: *Receptor Modeling for Air Quality Management*, P.K. Hopke, ed., Elsevier Science Publishers, Amsterdam, 117-147.

Henry, R.C. (1997a) Receptor Model Applied to Patterns in Space (RMAPS), Part 1 - Model Description, *J. Air Waste Manag. Assoc.* 47:216-219.

Henry, R.C. (1997b) Receptor Model Applied to Patterns in Space (RMAPS), Part II -Apportionment of Airborne Particulate Sulfur from Project MOHAVE, *J. Air Waste Manag. Assoc.* 47:220-225.

Henry, R.C. (1997c) Receptor Model Applied to Patterns in Space (RMAPS), Part III -Apportionment of Airborne Particulate Sulfur in Western Washington State, *J. Air Waste Manag. Assoc.* 47:226-230.

Henry, R.C. and B.M. Kim (1989) Extension of Self-Modeling Curve Resolution to Mixtures of More Than Three Components. Part 1. Finding the Basic Feasible Region, *Chemom. Intell. Lab. Syst.* 8:205-216.

Henry, R.C., C.W. Lewis, and J.F. Collins (1994) Vehicle-Related Hydrocarbon Source Compositions from Ambient Data: The GRACE/SAFER Methods, *Environ. Sci. Technol.* 28:823-832.

Hopke, P.K. (1982) Comments on 'Trace Element Concentrations in Summer Aerosols at Rural Sites in New York State and Their Possible Sources' by P. Parekh and L. Husain and 'Seasonal Variations in the Composition of Ambient Sulfur-Containing Aerosols' by R. Tanner and B. Leaderer, *Atmospheric Environ.* 16:1279-1280.

Hopke, P.K. (1985) *Receptor Modeling in Environmental Chemistry*, John Wiley & Sons, Inc., New York.

Hopke, P.K. (1989) Target Transformation Factor Analysis, P.K. Hopke, *Chemometrics and Intelligent Laboratory Systems* 6:7-19

Hopke, P.K. (1995) The Mixture Resolution Problem Applied to Airborne Particle Source Apportionment, *Chemometrics in Environmental Chemistry,* W. Einax, ed., Springer-Verlag, Heidelberg. 47-86.

Hopke, P.K., E.S. Gladney, G.E. Gordon, W.H. Zoller, and A.G. Jones, The Use of Multivariate Analysis to Identify Sources of Selected Elements in the Boston Urban Aerosol, *Atmospheric Environ.*, 10:1015-1025 (1976).

Hopke, P.K., R.E. Lamb and D.F.S. Natusch (1980) Multielemental Characterization of Urban Roadway Dust, *Environ. Sci. Technol.* 14:164-172.

Juntto, S. and P. Paatero (1994) Analysis of daily precipitation data by positive matrix factorization, *Environmetrics* 5:127-144.

Kaiser, H.F. and S. Hunka (1973) Some Empirical Results with Guttman's Stronger Lower Bound for the Number of Common Factors, *Education and Psych. Measurement* 33:99-102.

Lee, E., C.K. Chan, and P. Paatero (1999) Application of Positive Matrix Factorization in Source Apportionment of Particulate Pollutants in Hong Kong, *Atmospheric Environ.* 33:3201-3212.

Maenhaut, W. and J. Cafmeyer (1987) Particle-Induced X-Ray Emission and Multivariate Techniques: An Application to the Study of the Sources of Respirable Atmospheric Particles in Gent, Belgium, *J. Trace and Microprobe Techniques* 5:135-158.

Malinowski, E.R. (1991) *Factor Analysis in Chemistry*, Wiley, New York, 2nd Ed..

Malm, W.C., Sisler, J.F., Huffman, D., Eldred, R.A., and Cahill, T.A. (1994) Spatial and Seasonal Trends in Particle Concentration and Optical Extinction in the United States, *J. Geophys. Res.* 99:1347-1370.

Mason, B. (1966) *Principles of Geochemistry*, Wiley, New York.

Paatero, P. (1997) Least Squares Formulation of Robust, Non-Negative Factor Analysis, *Chemom. Intell. Lab. Syst.* 37:23-35.

Paatero, P. (1999)The Multilinear Engine --- a Table-driven Least Squares Program for Solving Multilinear Problems, Including the n-way Parallel Factor Analysis Model, J. Computational and Graphical Stat. 8: 1-35.

Paatero, P. and U. Tapper (1993) Analysis of Different Modes of Factor Analysis as Least Squares Fit Problems, *Chemom. Intell. Lab. Syst.* 18:183-194.

Paatero, P. and U. Tapper (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics* 5:111-126.

Paterson, K.G. , J.L. Sagady, D.L. Hooper , S.B. Bertman , M.A. Carroll , and P.B. Shepson (1999) Analysis of Air Quality Data Using Positive Matrix Factorization, *Environ. Sci. Technol.* 33: 635-641

Polissar, A.V., P.K. Hopke, W.C. Malm, J.F. Sisler (1996) The Ratio of Aerosol Optical Absorption Coefficients to Sulfur Concentrations, as an Indicator of Smoke from Forest Fires when Sampling in Polar Regions, *Atmospheric Environ.* 30:1147-1157.

Polissar, A.V., P.K. Hopke, W.C. Malm, J.F. Sisler (1998) Atmospheric Aerosol over Alaska: 2. Elemental Composition and Sources, *J. Geophys. Res.* 103: 19,045-19,057.

Polissar, A.V., P.K. Hopke, P. Paatero, Y. J. Kaufman, D. K. Hall, B. A. Bodhaine, E. G. Dutton and J. M. Harris (1999) The Aerosol at Barrow, Alaska: Long-Term Trends and Source Locations, *Atmospheric Environ.* 33: 2441-2458.

Watson, J.G., N.F. Robinson, J.C. Chow, R.C. Henry, B.M. Kim, T.G. Pace, E.L. Meyer, and Q. Nguyen (1990) The USEPA/DRI chemical mass balance receptor model, CMB 7.0, *Environ. Software* 5:38-49.

Watson, J.G., J.C. Chow, and T.G. Pace (1991) Chemical Mass Balance, in *Receptor Modeling for Air Quality Management*, Hopke, P.K., ed., Elsevier Science Publishers, Amsterdam, 83-116.

Xie, Y.-L., P. K. Hopke, P. Paatero, L. A. Barrie and S.-M. Li (1999a) Identification of Source Nature and Seasonal Variations of Arctic Aerosol by the Multilinear Engine, *Atmospheric Environ.* 33: 2549-2562.

Xie, Y. L., Hopke, P., Paatero, P., Barrie, L. A., and Li, S. M. (1999b). Identification of source nature and seasonal variations of Arctic aerosol by positive matrix factorization. *J. Atmos. Sci.* 56:249-260.

Yakovleva, E., P.K. Hopke and L. Wallace (1999) Receptor Modeling Assessment of PTEAM Data, *Environ. Sci. Technol.* 33: 3645-3652 (1999).

APPENDIX A Reference Concentrations

Composition of Seawater as compiled by Goldberg (1963).

Element	Concentration	X/Na	Element	Concentration	X/Na
	(ppm)			(ppm)	
Li	0.027	2.571E-06	As	0.003	2.857E-07
Be	0.0000006	5.714E-11	Se	0.004	3.81E-07
F	1.3	0.00012	Br	65	0.0061905
Na	10500	1	Rb	0.12	1.143E-05
Mg	1350	0.1285714	Sr	8	0.00076
Al	0.01	9.524E-07	Y	0.0003	2.857E-08
Si	3	0.0002857	Nb	0.00001	9.524E-10
S	885	0.0842857	Mo	0.01	9.524E-07
Cl	19000	1.8095238	Ag	0.0003	2.857E-08
K	380	0.0361905	Cd	0.00011	1.048E-08
Ca	400	0.0380952	Sn	0.003	2.857E-07
Sc	0.00004	3.81E-09	Sb	0.0005	4.762E-08
Ti	0.001	9.524E-08	I	0.06	5.714E-06
V	0.002	1.905E-07	Cs	0.0005	4.762E-08
Cr	0.00005	4.762E-09	Ba	0.03	2.857E-06
Mn	0.002	1.905E-07	La	0.0003	2.857E-08
Fe	0.01	9.524E-07	Ce	0.0004	3.81E-08
Co	0.0005	4.762E-08	W	0.0001	9.524E-09
Ni	0.002	1.905E-07	Au	0.000004	3.81E-10
Cu	0.003	2.857E-07	Hg	0.00003	2.857E-09
Zn	0.001	9.524E-08	Pb	0.00003	2.857E-09
Ga	0.00003	2.857E-09	Th	0.00005	4.762E-09
Ge	0.00007	6.667E-09			

Crustal Composition as Compiled by Mason (1966)

	Crustal			Crustal	
Element	Abundance	X/Al	Element	Abundance	X/Al
H	1400	0.01722	Rh	0.005	0.00000
Li	20	0.00025	Pd	0.01	0.00000
Be	2.8	0.00003	Ag	0.07	0.00000
B	10	0.00012	Cd	0.2	0.00000
C	200	0.00246	In	0.1	0.00000
N	20	0.00025	Sn	2	0.00002
O	466000	5.73186	Sb	0.2	0.00000
F	625	0.00769	Te	0.01	0.00000
Na	28300	0.34809	I	0.5	0.00001
Mg	20900	0.25707	Cs	3	0.00004
Al	81300	1.00000	Ba	425	0.00523
Si	277200	3.40959	La	30	0.00037
P	1050	0.01292	Ce	60	0.00074
S	260	0.00320	Pr	8.2	0.00010
Cl	130	0.00160	Nd	28	0.00034
K	25900	0.31857	Sm	6	0.00007
Ca	36300	0.44649	Eu	1.2	0.00001
Sc	22	0.00027	Gd	5.4	0.00007
Ti	4400	0.05412	Tb	0.9	0.00001
V	135	0.00166	Dy	3	0.00004
Cr	100	0.00123	Ho	1.2	0.00001
Mn	950	0.01169	Er	2.8	0.00003
Fe	50000	0.61501	Tm	0.5	0.00001
Co	25	0.00031	Yb	3.4	0.00004
Ni	75	0.00092	Lu	0.5	0.00001
Cu	55	0.00068	Hf	3	0.00004
Zn	70	0.00086	Ta	2	0.00002
Ga	15	0.00018	W	1.5	0.00002
Ge	1.5	0.00002	Re	0.001	0.00000
As	1.8	0.00002	Os	0.005	0.00000
Se	0.05	0.00000	Ir	0.001	0.00000
Br	2.5	0.00003	Pt	0.01	0.00000
Rb	90	0.00111	Au	0.004	0.00000
Sr	375	0.00461	Hg	0.08	0.00000
Y	33	0.00041	Tl	0.5	0.00001
Zr	165	0.00203	Pb	13	0.00016
Nb	20	0.00025	Bi	0.2	0.00000
Mo	1.5	0.00002	Th	7.2	0.00009
Ru	0.01	0.00000	U	1.8	0.00002

APPENDIX B

EIGENVECTOR ANALYSIS

Dispersion Matrix

The initial step in the analysis of the data requires the calculation of a function that can indicate the degree of interrelationships that exists within the data. Functions exist that can provide this measure between the two variables when calculated over all of the samples or between the samples when calculated over all of the variables. The most well-known of these functions is the product-moment correlation coefficient. To be more precise, this function should be referred to as the correlation about the mean. The "correlation coefficient" between two variables, x_i and x_k, over all n samples is given by

$$\frac{\sum_{j=1}^{n} (x_{ij} - \overline{x_i})(x_{kj} - }{(x_{ij} - \overline{x_i})^2 \Big)^{1/2} \Big(\sum_{j=1}^{n} (x_{kj}}$$

The original variables can be transformed by subtracting the mean value and dividing by the standard deviation,

$$z_{ij} = \frac{x_{ij} - \overline{x_i}}{\sigma_i}$$

Using the standardized variables, equation 6 can be simplified to

$$r_{ik} = \frac{1}{n} \sum_{j=1}^{n} z_{ij} z_{kj}$$

The standardized variables have several other benefits to their use. Each standardized variable has a mean value of zero and a standard deviation of 1. Thus, each variable carries 1 unit of system **variance** and the total variance for a set of measurements of m variables would be m.

There are several other measures of interrelationship that can also be utilized. These measures include covariance about the mean as defined by

$$c_{ik} = \sum_{j=1}^{n} d_{ij} d_{kj}$$

where

$$d_{ij} = x_{ij} - \overline{x_i}$$

are called the deviations, and $\bar{x}_i$ is the average value of the ith variable; The covariance about the origin is defined by

$$c_{ik}{}^{o} = \sum_{j=1}^{n} x_{ij} x_{kj}$$

and the correlation about the origin

$$_{ik}{}^{o} = \frac{\sum_{j=1}^{n} x_{ij} x_{kj}}{(\sum_{j=1}^{n} x_{ij}{}^{2} \sum_{j=1}^{n} x_{kj}{}^{2})^{1/}}$$

The matrix of either the correlations or covariances, called the dispersion matrix, can be obtained from the original or transformed data matrices. The data matrix contains that data for the m variables measured over the n samples. The correlation about the mean is given by

$$R_{m} = ZZ'$$

where Z′ is the transpose of the standardized data matrix Z. The correlation about the origin

$$= Z^{o} Z^{o'} = (XV)(XV$$

where

$$z_{ij}{}^{o} = \frac{x_{ij}}{(\sum_{j=1}^{n} x_{ij}{}^{2})^{1/2}}$$

which is a normalized variable still referenced to the original variable origin and V is a diagonal matrix whose elements are defined by

$$v_{ik} = \delta_{ik} \, (\sum_{j=1}^{n} x_{ij}^{2})^{1/2}$$

This normalized variable also carries a variance of 1, but the mean value is not zero. The covariance about the mean is given as

$$C_m = DD'$$

where D is the matrix of deviations from the mean whose elements are calculated using equation 10 and D′ is its transpose. The covariance about the origin is

$$C_o = XX'$$

the simple product of the data matrix by its transpose. As written, these product matrices would be of dimension m by m and would represent the pairwise interrelationships between variables. If the order of the multiplication is reversed, the resulting n by n dispersion matrices contain the interrelationships between samples.

The relative merits of these functions to reflect the total information content contained in the data have been discussed in the literature [Rozett and Petersen, 1975; Duewer *et al.*, 1976]. Rozett and Petersen [1975] argue that since many types of physical and chemical variables have a real zero, the information regarding the location of the true origin is lost by using the correlation and covariance about the mean that include only differences from the variable mean. The normalization made in calculating the correlations from the covariances causes each variable to have an identical weight in the subsequent analysis. In mass spectrometry where the variables consist of the ion intensities at the various m/e values observed for the fragments of a molecule, the normalization represents a loss of information because the variable metric is the same for all of the m/e values. In environmental studies where measured species concentrations range from the trace level (sub part per million) to major constituents at the percent level, the use of covariance may weight the major constituents too heavily in the subsequently analyses. The choice of dispersion function depends heavily on the nature of the variables being measured.

Another use of the correlation coefficient is that it can be interpreted in a statistical sense to test the null hypothesis as to whether a linear relationship exists between the pair of variables being tested. It is important to note that the existence of a correlation coefficient that is different from zero does not prove that a cause and effect relationship exists. Also, it important to note that the use of probabilities to determine if correlation coefficients are "significant" is very questionable for environmental data. In the development of those probability relationships, explicit assumptions are made that the underlying distributions of the variables in the correlation analysis are normal. For most environmental variables, normal distributions are uncommon. Generally, the distributions are positively skewed and heavy tailed. Thus, great care should be taken in making probability arguments regarding the significance of pairwise correlation coefficients between variables measured in environmental samples.

Another problem with interpreting correlation coefficients is that environmental systems are often truly multivariate systems. Thus, there make be more than two variables that covary because of the underlying nature of the processes being studied. Although there can be very strong correlations between two variables, the correlation may arise through a causal factor in the system that cannot be detected

For each of the equations previously given in this section, the resulting dispersion matrix provides a measure of the interrelationship between the measured variables. Thus, in the use of a matrix of correlations between the pairs of variables, each variable is given equal weight in the subsequent eigenvector analysis. This form of factor analysis is commonly referred to as an R-mode analysis. Alternatively, the order of multiplication could be reversed to yield covariances or correlations between the samples obtained in the system. The eigenvector analysis of these matrices would be referred to as a Q-analysis. The differences between these two approaches will be discussed further after the approach to eigenvector analysis has been introduced.

Eigenvector Calculation

The primary goal of eigenvector analysis is to represent a data matrix as a product of two other matrices containing specific information regarding the sources of the variation observed in the data. It can be shown [Horst, 1963] that any matrix can be expressed as a product of two matrices

$$X_{nm} = A_{np} F_{pm}$$

where the subscripts denote the dimensions of the respective matrices. There will be an infinite number of different A and F matrices that satisfy this equation.

To divide the matrix into two cofactor matrices as in equation 19, a question is raised about the minimum value p can have and still yield a solution. This value is the "rank" of matrix X [Horst, 1963, p. 335]. The rank clearly cannot be greater than a matrix's smaller dimension and the rank of a product moment matrix cannot be greater than the smaller of the number of columns or the number of rows. The rank of the product moment matrix must be the same as that of the matrix from which it was formed.

Associated with the idea of matrix rank is the concept of linear independence of variables. We can look at the interrelationships between columns (or rows) of a matrix and determine if columns (or rows) are linearly independent of one another. To understand linear independence let us examine the relationship between the two vectors in Figure 11.

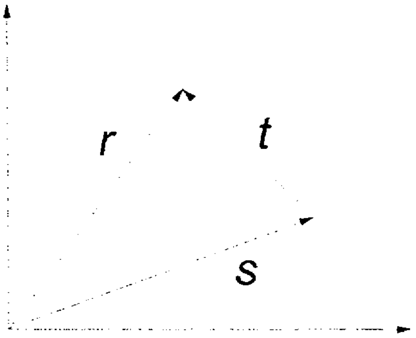

Figure 11. Illustration of the interrelationship between two vectors.

We can find the vector **t** such that

$$\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{t}$$

The vector **t** can then be generalized to be the resultant of the sum of **r** and **s** with coefficients a_r and a_s:

$$\boldsymbol{t} = a_r \boldsymbol{r} + a_s \boldsymbol{s}$$

If **t** = 0, then **r** and **s** are said to be collinear or linearly dependent vectors. Thus, a vector **y** is linearly dependent on a set of vectors, $\mathbf{v}_1$, $\mathbf{v}_2$,...,$\mathbf{v}_m$ if

$$a_1 v_1 + a_2 v_2 + \ldots + a$$

and at least one of the coefficients, a_i, is non-zero. If all of the a_i values in equation 22 are zero, then **y** is linearly independent of the set of vectors, $\mathbf{v}_i$. The number of linearly independent column vectors in a matrix defines the minimum number of dimensions needed to contain all of the vectors. The idea of the rank or true dimensionality of a data matrix is an important concept in receptor modeling as it defines the number of separately identifiable, independent sources contributing to the system under study. Thus finding the rank of a data matrix will be an important task. In addition, the ability to resolve sources of material with similar properties or the resolution of various receptor models needs to be carefully examined. To examine this question, several additional mathematical concepts need to be discussed.

A given data matrix can be reproduced by one of an infinite number of sets of independent column vectors or basis vectors that will describe the axes of the reduced dimensionality space. The rank of the matrix can be determined and a set of linearly independent basis vectors can be developed by the use of an eigenvalue analysis. In this discussion, only the analysis of real, symmetric matrices such as those obtained as the minor or major product of a data matrix will be discussed. Suppose there exists a real, symmetric matrix R that is to be analyzed for its rank, remembering that the rank of a product moment matrix is the same as the data matrix from which it is formed.

An eigenvector of R is a vector **u** such that

$$Ru = u\lambda$$

where λ is an unknown scalar. The problem then is to find a vector so that the vector Ru is proportional to **u**. This equation can be rewritten as

$$Ru - u\lambda = 0$$

or

$$(R - \lambda I)u = 0$$

implying that **u** is a vector that is orthogonal to all of the row vectors of (R - λI). This vector equation can be considered as a set of p equations where p is the order of R:

$$+ \ u_2r_{12} + u_3r_{13} + \ldots$$

$$1 - \lambda) + u_3r_{23} + \ldots$$

$$\cdot \quad \cdot \quad \ldots$$

$$_{p2} + u_3r_{p3} + \ldots + u_p($$

Unless **u** is null vector, equation 25 can only hold if

$$R - \lambda I = 0$$

There is a solution to this set of equations only if the determinant of the left side of the equation is zero:

$$|R - \lambda I| = 0$$

This equation yields a polynomial in λ of degree p. It is then necessary to obtain the p roots of this equation, λ_i, i = 1, p. For each λ_i there is an associated vector $\mathbf{u}_i$ such that

$$R\boldsymbol{u}_1 - \boldsymbol{u}_1\lambda_1 = 0$$

If these λ_i values are placed as the elements of a diagonal matrix Λ, and the eigenvectors can be collected as columns of the matrix U, then we can express equation 25 as

$$RU = U\Lambda$$

The matrix U is a square orthonormal so that

$$U'U = UU' = I$$

Postmultiplying equation 30 by U′ yields

$$R = UU'\Lambda$$

Thus, any symmetric matrix R may be represented in terms of its eigenvalues and eigenvectors:

$$\iota_1\boldsymbol{u}_1' + \lambda_2\boldsymbol{u}_2\boldsymbol{u}_2' + \ldots +$$

so that R is weighted sum of matrices $\mathbf{u}_i\mathbf{u}_i'$, of order p by p and of rank 1. Each term is orthogonal to all other terms so that for i $\neq$j

$$\boldsymbol{u}_i'\boldsymbol{u}_j = 0$$

and

$$\boldsymbol{u}_i\boldsymbol{u}_i'\boldsymbol{u}_j\boldsymbol{u}_j' = 0$$

Premultiplying equation 30 by U′ yields

$$U'RU = \Lambda$$

so that U is a matrix that reduces R to a diagonal form. The eigenvalues have a number of useful properties [Joreskog *et al.*, 1976].

1. Trace Λ = trace R; the sum of the eigenvalues equals the sum of the elements in the principal diagonal of the matrix.

2. $\Pi^{p}_{i=1}\ \lambda_i = |R|$; the product of the eigenvalues equals the determinant of the matrix. If one or more of the eigenvalues is zero, then the determinant is zero and the matrix R is called a singular matrix. A singular matrix cannot be inverted.

3. The number of non-zero eigenvalues equals the rank of R.

Therefore, if for a matrix R of order p, there are m zero eigenvalues, the rank of R is (p - m). The (p - m) eigenvectors corresponding to those non-zero linearly independent vectors of R.

Another approach can be taken to examine the basic structure of a matrix. This method is called a singular value decomposition of an arbitrary rectangular matrix. A very detailed discussion of this process is given by Lawson and Hanson [1974]. According to the singular value decomposition theorem, any matrix can be uniquely written as

$$R = UDV'$$

where R is an n by m data matrix, U is an n by n orthogonal matrix, V, is an m by m orthogonal matrix, and D is an n by m diagonal matrix. All of the diagonal elements are non-negative and exactly k of them are strictly positive. These elements are called the singular values of R. The column values of the U matrix are the eigenvectors of XX$'$. The column values of the V matrix are the eigenvectors of X$'$X. Zhou *et al.* [1983] show that the R- and Q-mode factor solutions are interrelated as follows

$$X = \rlap{\overbrace{U}^{A_Q}\,\overbrace{D\ V'}^{F_Q}}\underbrace{\,}_{F_R}\ \underbrace{}_{A_R}$$

Although there has been discussion of the relative merits of R- and Q-mode analyses in the literature [Rozett and Petersen, 1975; Hwang *et al.*, 1984], the direction of multiplication is not the factor that alters the solutions obtained. Different solutions are obtained depending on the direction in which the scaling is performed [Heidam and Kronborg, 1985]. Thus, different vectors are derived depending on whether the data are scaled by row, by column, or both. Zhou *et al.* [1983] discuss this problem in more detail.

By making appropriate choices of A and F in equation 38, the singular value decomposition is one method to partition any matrix. The singular value decomposition is also a key diagnostic tool in examining collinearity problems in regression analysis [Belsley *et al.*, 1980]. The application of the singular value decomposition to regression diagnostics is beyond the scope of this chapter.

In the discussion of the dispersion matrix, it becomes necessary to discuss some of the semantical problems that arise in "factor" analysis. If one consults the social science literature on factor analysis, a major distinction is made between factor analysis and principal components analysis. Because there are substantial problems in developing quantitative models analogous to those given in the introduction to this chapter in equations 1 to 5, the social sciences want to obtain "factors" that have significant values for two or more of the measured variables. Thus, they are interested in factors that are common to several variables. The model then being applied to the data is of the form:

$$z_{ij} = \sum_{k=1}^{p} a_{ik} f_{kj} + d_i U_{ij}$$

where the standardized variables, z_{ij}, are related to the product of the common factor loadings, a_{ik}, by the common factor scores, f_{kj}, plus the unique loading and score. The system variance is therefore partitioned into the common factor variance, the specific variance unique to the particular variable, and the measurement error.

$$\text{System Variance} = ($$

In order to make this separation, an estimation is made of the partitioning of the variance between the common factors and the specific factors. A common approach to this estimation is to replace the 1's on the diagonal of the correlation matrix with a estimate of the "communality" defined by

$$h^2 = \sum_{k=1}^{p} a_{ik}^2$$

The multiple correlation coefficients for each variable against all of the remaining variables are often used as initial estimates of the communalities. Alternatively, the eigenvector analysis is made, the communalities for the initial solution are then substituted into the diagonal elements of the correlation matrix to produce a communality matrix. This matrix is then analyzed and the process repeated until stable communality values are obtained.

The principal components analysis simply decomposes the correlation matrix and leads to the model outlined in equation 39 without the d_iU_{ij} term. It can produce components that have a strong relationship with only one variable. This single variable component could also be considered to be the unique factor. Thus, both principal components analysis and classical factor analysis really lead to similar solutions although reaching these solutions by different routes. Since it is quite reasonable for many environmental systems to show factors that produce such single variable behavior, it is advisable to use a principal components analysis and extend the number of factors to those necessary to reproduce the original data within the error limits inherent in the data set.

Typically, this approach to making the eigenvector analysis compresses the information content of the data set into as few eigenvectors as possible. Thus, in considering the number of factors to be used to describe the system, it is necessary to carefully examine the problems of reconstructing both the variability within the data and reconstructing the actual data itself.

Number of Retained Factors

Following the diagonalization of the correlation or covariance matrix, it is necessary to make the difficult choice of the number of factors, p, to use in the subsequent analysis. This problem occurs in any application of an eigenvector analysis of data containing noise. In the absence of error, the eigenvalues beyond the true number of sources become zero except for calculational error. The choice becomes more difficult depending on the error in the data. Several approaches have been suggested [Duewer *et al.*, 1976; Hopke *et al.*, 1980].

A large relative decrease in the magnitude of the eigenvalues is one indicator of the correct number of factors. It can often be useful to plot the eigenvalues as a function of factor number and look for sharp breaks in the slope of the line [Cattell, 1966]. If the eigenvalue is a measure of the information content of the corresponding eigenvector, then only sufficiently "large" eigenvalues need to be retained in order to reproduce the variation initially present in the data. One of the most commonly used and abused criteria for selecting the number of factors to retain is retaining only those eigenvalues greater than 1 [Guttman, 1954]. The argument is made that the normalized variables each carry one unit of variance. Thus, if an eigenvalue is less than one, then it carries less information than one of the initial variables and is therefore not needed. However, Kaiser and Hunka [1973] make a strong argument that although eigenvalue greater than one does set a *lower* limit on the number of factors to be retained, it does not set a simultaneous upper bound. Thus, there must be at least as many factors as there are eigenvalues greater than one, but there can be more than that number that are important to the understanding of the system's behavior.

Hopke [1982] has suggested a useful empirical criterion for choosing the number of retained eigenvectors. In a number of cases of airborne particulate matter composition source identification problems, Hopke found that choosing the number of factors containing variance greater than one *after* an orthogonal rotation provided a stable solution. Since the eigenvector analysis artificially compresses variance into the first few factors, reapportioning the variance using the rotations described in the next section will result in more factors with total variance greater than one than there are eigenvalues greater than one. In many cases, this number of factors will stay the same even after rotating more factors.

For a different type of test, the original data are reproduced using only the first factor and compared point-by-point with the original data. Several measures of the quality of fit are calculated including chi-squared

$$\chi^2 = \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{(x_{ij} - \bar{x}_i)^2}{\sigma_{ij}^2}$$

where x_{ij} is the reconstructed data point using p factors and σ_{ij} is the uncertainty in the value of x_{ij}. The Exner function [Exner, 1966] is a similar measure and is calculated by

$$p = \left[\sum_{j=1}^{n} \sum_{i=1}^{m} \frac{(x_{ij} - \bar{x}_i)^2}{(x^o - \bar{x}_i)^2} \right]$$

where x^o is a grand ensemble average value. The empirical indicator function suggested by Malinowski [1977] can be used for this purpose and is calculated as follows:

$$\left[\sum_{j=p+1}^{n} \frac{\lambda_j}{n(m-p)} \right] .$$

$$IND = \frac{RSD}{(m-p)^2}$$

$$\left[\sum_{j=p+1}^{n} \frac{\lambda_j}{m(n-p)} \right]$$

$$IND = \frac{RSD}{(n-p)^2}$$

where λ_j are the eigenvalues from the diagonalization. This function has proven very successful with spectroscopy results [Malinowski, 1977]. However, it has not proven to be as useful with other types of environmental data [Hopke, 1989]. Finally, the root-mean-square error and the arithmetic average of the absolute values of the point-by-point errors are also calculated. The data are next reproduced with both the first and second factors and again compared point-by-point with the original data. The procedure is repeated, each time with one additional factor, until the data are reproduced with the desired precision. If p is the minimum number of factors needed to adequately reproduce the data, then the remaining n-p factors can be eliminated from the analysis. These tests do not provide unequivocal indicators of the number of factors that should be retained. Judgement becomes necessary in evaluating all of the test results and deciding upon a value of p. In this manner the dimension of the A and F matrices is reduced from n to p.

The compression of variance into the first factors will improve the ease with which the number of factors can be determined. However, their nature has now been mixed by the calculational method. Thus, once the number of factors has been determined, it is often useful to rotate the axes in order to provide a more interpretable structure.

Rotation of Factor Axes

The axis rotations can retain the orthogonality of the eigenvectors or they can be oblique. Depending on the initial data treatment, the axes rotations may be in the scaled and/or centered space or in the original variable scale space. The latter approach has proved quite useful in a number of chemical applications described by Malinowski [1991] and in environmental systems as described by Hopke [1985]. To begin the discussion of factor rotation, it is useful to describe how one set of coordinate system axes can be transformed into a new set. Support that it is necessary to change from the coordinate system x_1, x_2 to the system of y_1, y_2 by rotating through angle θ as shown in Figure 12. For this two-dimensional system, it is easy to see that

$$= \cos\theta\ x_1 + \sin\theta\ x$$
$$= -\sin\theta\ x_1 + \cos\theta$$

In matrix form, this equation could be written as

$$Y' = X'T$$

where T is the transformation matrix.

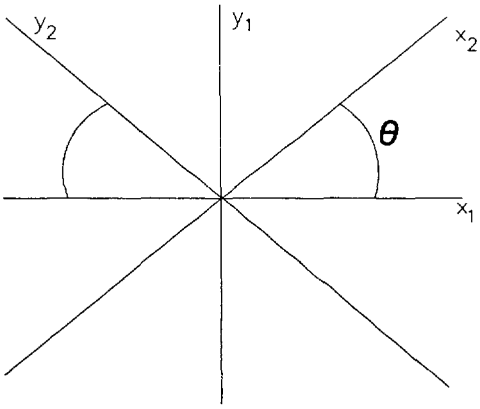

Figure 12. Illustration of the rotation of a coordinate system (x_1, x_2) to a new system (y_1, y_2) by an angle Θ.

In order to obtain a more general case, it is useful to define the angles of rotation in the manner shown in Figure 13. Now the angle θ_{ij} is the angle between the ith original reference axis and the jth new axis. Assuming that this is a rotation that maintains orthogonal axes (rigid rotation), then, for two dimensions,

$$\theta_{12} = \theta_{11} + 90^o$$
$$\theta_{21} = \theta_{11} + 90^o$$
$$\theta_{22} = \theta_{11}$$

There are also trigonometric relationships that exist,

$$\theta_{21} + 90^{\circ}) = \cos \theta_{21}$$
$$\theta_{12} + 90^{\circ}) = \sin (90'$$

so that equation 48 can be rewritten as

$$= \cos \theta_{11}\, x_1 + \cos \theta_{2}$$
$$= \cos \theta_{12}\, x_1 + \cos \theta_{2}$$

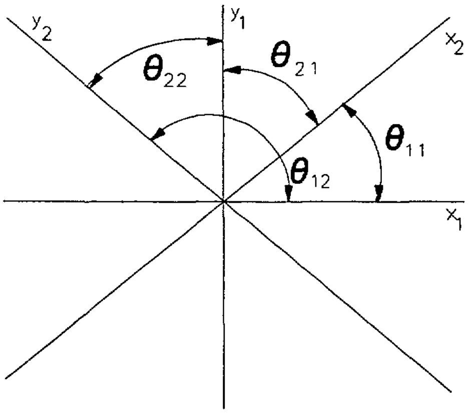

Figure 13. Illustration of the rotation of a coordinate system (x_1, x_2) to a new coordinate system (y_1, y_2) showing the definitions of the full set of angles used to described the rotation.

This set of equations can then be easily expanded to n orthogonal axes yielding

$$x_1 + \cos \theta_{21}\, x_2 + \cdots +$$
$$x_1 + \cos \theta_{22}\, x_2 + \cdots +$$
$$\cdot \quad \cdot \quad \cdot \quad \cdots \quad .$$
$$x_1 + \cos \theta_{2n}\, x_2 + \cdots +$$

A transformation matrix T can then be defined such that its elements are

$$t_{ij} = \cos \theta_{ij}$$

Then, for a collection of N row vectors in a matrix X with n columns,

$$Y = XT$$

and Y has the coordinates for all N row vectors in terms of the n rotated axes. For the rotation to be rigid, T must be an orthogonal matrix. Note that the column vectors in Y can be thought of as new variables made up by linear combinations of the variables in X with the elements of T being the coefficients of those combinations. Also a row vector of X gives the properties of a sample in terms of the original variables while a row vector of Y gives the properties of a sample in terms of the transformed variables.

Single copies of this report are available cost-free on request from:

Section of Nutritional and Health-Related Environmental Studies
Division of Human Health
Department of Nuclear Sciences and Applications
International Atomic Energy Agency
P.O. Box 100
A-1400 Vienna
Austria