US EPA ARCHIVE DOCUMENT

US EPA ARCHIVE DOCUMENT

United States Environmental Protection Agency

EPA Unmix 6.0 Fundamentals & User Guide

RESEARCH AND DEVELOPMENT

US EPA ARCHIVE DOCUMENT

EPA/600/R-07/089
June 2007
www.epa.gov

EPA Unmix 6.0 Fundamentals & User Guide

Gary Norris, Ram Vedantham, Rachelle Duvall
U.S. Environmental Protection Agency
National Exposure Research Laboratory
Research Triangle Park, NC 27711

Ronald C. Henry
24017 Ingomar Street
West Hills, CA 91304

U.S. Environmental Protection Agency
Office of Research and Development
Washington, DC 20460

Notice: Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy. Mention of trade names and commercial products does not constitute endorsement or recommendation for use.

US EPA ARCHIVE DOCUMENT

Disclaimer

EPA through its Office of Research and Development funded and managed the research and development described here. The User Guide has been subjected to Agency review and is cleared for official distribution by the EPA. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

US EPA ARCHIVE DOCUMENT

TABLE OF CONTENTS

US EPA ARCHIVE DOCUMENT

LIST OF FIGURES

US EPA ARCHIVE DOCUMENT

LIST OF TABLES

US EPA ARCHIVE DOCUMENT

SECTION 1. INTRODUCTION

The underlying philosophy of Unmix is to let the data speak for itself. Unmix seeks to solve the general mixture problem where the data are assumed to be a linear combination of an unknown number of sources of unknown composition, which contribute an unknown amount to each sample. Unmix also assumes that the compositions and contributions of the sources are all positive. Unmix assumes that for each source there are some samples that contain little or no contribution from that source. Using concentration data for a given selection of species, Unmix estimates the number of sources, source compositions, and source contributions to each sample.

It is well known that the general mixture problem and the special case of multivariate receptor modeling are ill posed problems. There are simply more unknowns than equations and thus there may be many wildly different solutions that are all equally good in a least-squares sense. Statisticians say that these problems are not identifiable. One approach to ill-posed problems is to impose conditions that add additional equations, which then define a unique solution. The most likely candidates for these additional conditions, or constraints, are the non-negativity conditions imposed by the physical nature of the problem. Source compositions and contributions must be non-negative. Unfortunately, it has been shown that non-negativity conditions alone are not sufficient to give a unique solution and more constraints are needed (Henry, 1987). Under certain rather mild conditions, the data themselves can provide the needed constraints (Henry, 1997). This is how Unmix works. However, sometimes the data do not support a solution. In this case Unmix will not find one. While some might judge this a disadvantage, it is actually a positive benefit to the user. Few modeling approaches let the user know clearly when a reliable solution is not possible.

If the data consists of many observations of M species, then the data can be plotted in an M-dimensional data space where the coordinates of a data point are the observed concentrations of the species during a sampling period. If there are N sources, the data space can be reduced to an N-1-dimensional space. It is assumed that for each source there are some data points where the contribution of the source is not present or small compared to the other sources. These are called edge points and Unmix works by finding these points and fitting a hyperplane through them; this hyperplane is called an edge (if N = 3, the hyperplane is a line). By definition, each edge defines the points where a single source is not contributing. If there are N sources, then the intersection of N-1 of these hyperplanes defines a point that has only one source contributing. Thus, this point gives the source composition. In this way the composition of the N sources are found, and from this the source contributions are calculated so as to give a best fit to the data. The Unmix modeling process is explained by simple graphical examples in Henry (1997), and a list of manuscripts that provide details on the algorithms used by Unmix can be found in Section 8.

US EPA ARCHIVE DOCUMENT

The term "Source" should be considered short for “Source type.” This more general term accounts for the potential that there could be a cluster of sources within short distances of each other and/or there could be multiple sources along the wind flow pattern reaching the receptor thereby creating source types.

SECTION 2. INSTALLING Unmix

EPA Unmix 6.0 runs in a Microsoft Windows environment (2000, XP, and Vista). Special instructions for running Unmix under the Windows Vista Operating System are listed in Appendix A. When correctly installed, double clicking on the icon (shown below) on the Desktop will start the program.

2.1 Hardware and Software Requirements

The speed of execution depends on the memory, processor speed etc. of the user's computer. A minimum of 80 Mb of disk space is required to install the program.

2.2 Installation Directions

1) Download the installation executable from the File Transfer Protocol (FTP) site.
2) Double-click on the EPA Unmix 6.0 Standalone Installation.exe file.
3) Follow the on-screen instructions to complete the installation.

If Unmix 6.0 has already been installed on your computer and you are updating the program with a new version, please follow the instructions in Appendix B. To uninstall Unmix, use the “Add or Remove Programs” option (from Start => Control Panel), select the EPA Unmix 6.0 on the window that opens and press “Remove.” After that, also remove Matlab Component Runtime (MCR) program from the same list. Please be aware that this process does not completely remove all files related to EPA Unmix 6.0. Files and folders that may have been created on your computer cannot be removed by this process. Those will have to be removed individually.

2.3 Testing the Installation

Double click the EPA Unmix 6.0 icon. The first time the program is run, MATLAB creates a machine copy of the code in the MCR folder. This slightly increases the amount of time to start the program and subsequent use of the model will not require re-creation of the code.

US EPA ARCHIVE DOCUMENT

An EPA disclaimer will first be displayed. Select OK and the program's Main Window Graphic User Interface (GUI) will be displayed.

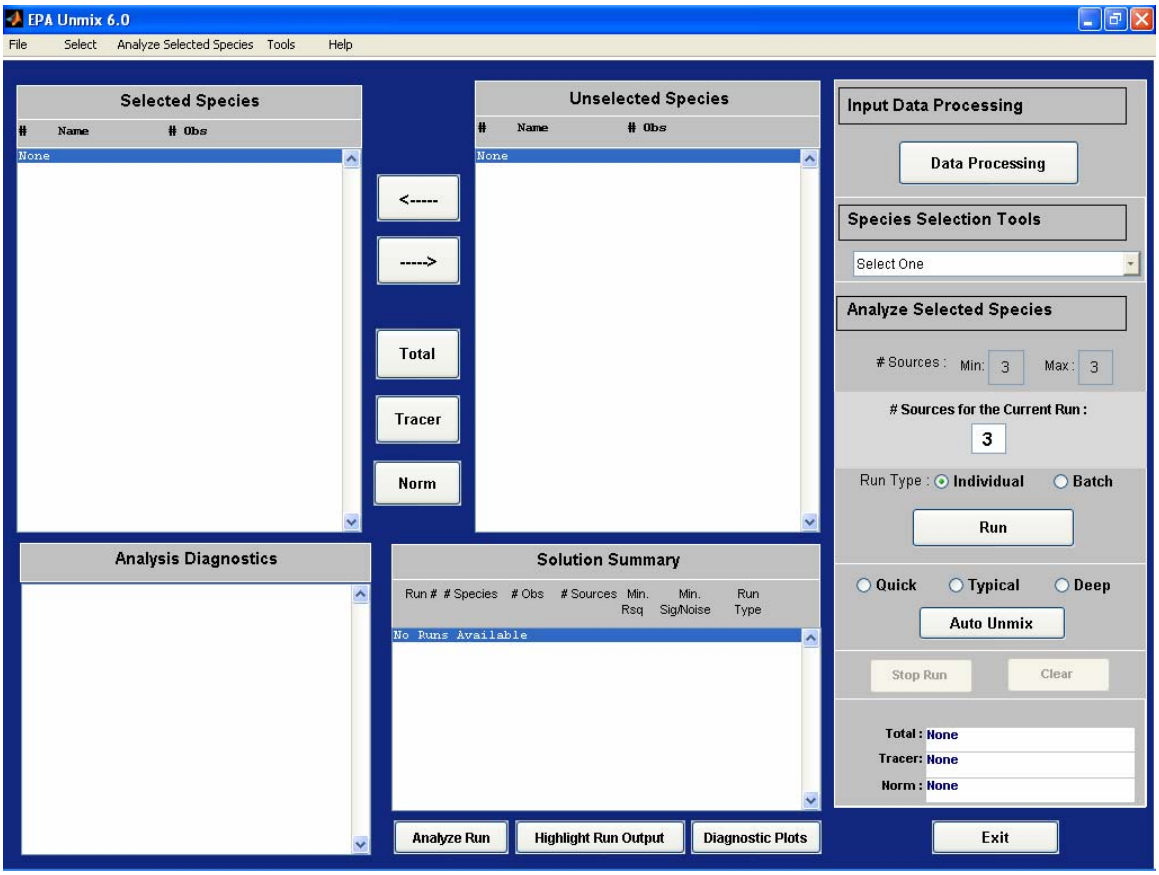


Figure 1: Main window

It should be noted that a prolonged session of use of EPA Unmix can lead to slower and slower response. This is due to the intensity of memory manipulation required by a scientific program compared to a word processing program. Memory allocations tend to become fragmented and memory management uses more and more of the available computer resources. As such, the Windows Operating System (OS) is not ideally suited for intense scientific computations. When the program response speed appears to slow down perceptibly, the user should save the current profile, shut down the program, and re-start the program.

SECTION 3. BASIC OPERATIONS

This section walks through the basic sequence of operations to produce a receptor model of the data consisting of the source compositions and source contributions that reproduce the data. Unmix procedure diagrams that cover both the basic and advanced operations can be found in Appendix D.

US EPA ARCHIVE DOCUMENT

The following discussions assume that the program has been installed in the default directory of C:\Program Files\EPA Unmix 6.0. Please make appropriate changes if the program has been installed in a directory other than the one mentioned above.

3.1 Input Data

Unmix accepts delimited (*.txt, *.dat, *.csv) and Microsoft Excel data files. Files with dates and dates and time can also be used. The species names must be placed in the first row of the data file and if dates (mm/dd/yyyy) or dates and times (hh:mm) are provided they must be placed in the first and first and second columns, respectively. Missing values can either be characters (e.g. XX) or specific numerical values (e.g.-99).

The recommended units for input files are ppb for gases such as NO_2 and SO_2, ppm for CO, and $\mu g/m^3$ for PM mass and species. Replace negative or zero data with half the method detection limit for the species**.** If zero values are in the data set, a warning message will appear and the number of zero values for each species will be displayed in the Data Processing window. Table 1 provides a summary of the input data parameters.

Table 1: Input Data Parameters

Data File Types	Delimited *.txt, *.dat, *.csv, or *.xls
Data Location:	
Species Names	First row of data file
Date Only	First column of data file
Date & Time	First and second columns of data file
Data Format:	
Date	MM/DD/YYYY (e.g. January 1, 2007 = 01/01/2007)
Time	HH:MM in 24 hour cycle (e.g. 1:45 PM = 13:45)
Missing Values	Characters (e.g. XX) or Numerical Values (e.g. -99)
Recommended Units:	
NO_2 or SO_2	ppb
CO	ppm
PM Mass & Species	$\mu g/m^3$
Negative or Zero Data	Replace with 1/2 the method detection limit for the species

Five example data files have been provided that are located in the C:\Program Files\EPA Unmix 6.0\Data folder. A data file is read by selecting the File ▶ New

US EPA ARCHIVE DOCUMENT

Input Data command in the upper left corner of the main window. The input data file window will open as shown in Figure 2. Input data file errors are typically caused by not correctly identifying the file's date and time information on the Input Data File window.

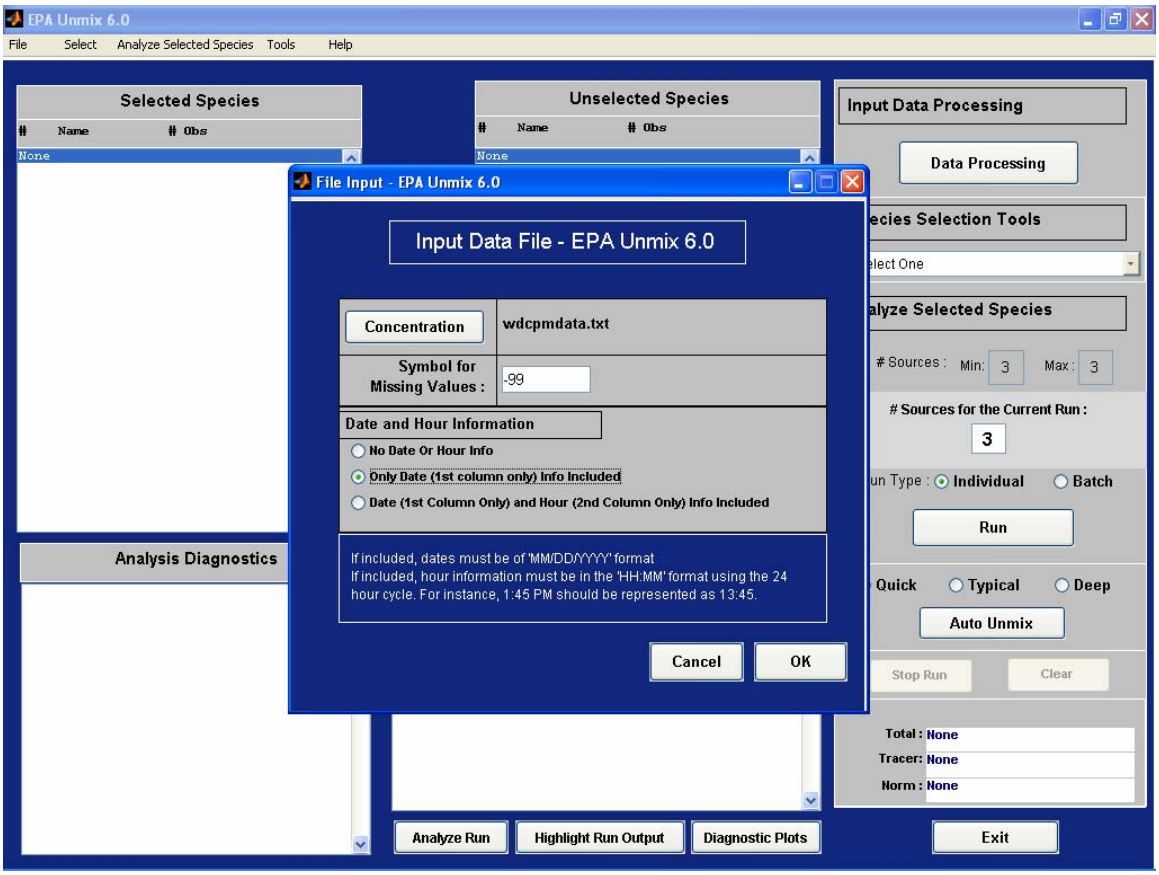


Figure 2: Input Data File

Input the file wdcpmdata.txt from the C:\Program Files\EPA Unmix 6.0\Data folder. These data are from the Washington, D.C. Interagency Monitoring of Protected Visual Environments (IMPROVE) PM monitoring site. Date information is included in the file and the missing value symbol or code is –99. The Unmix Data Processing window will open as shown in Figure 3 after inputting the data.

US EPA ARCHIVE DOCUMENT

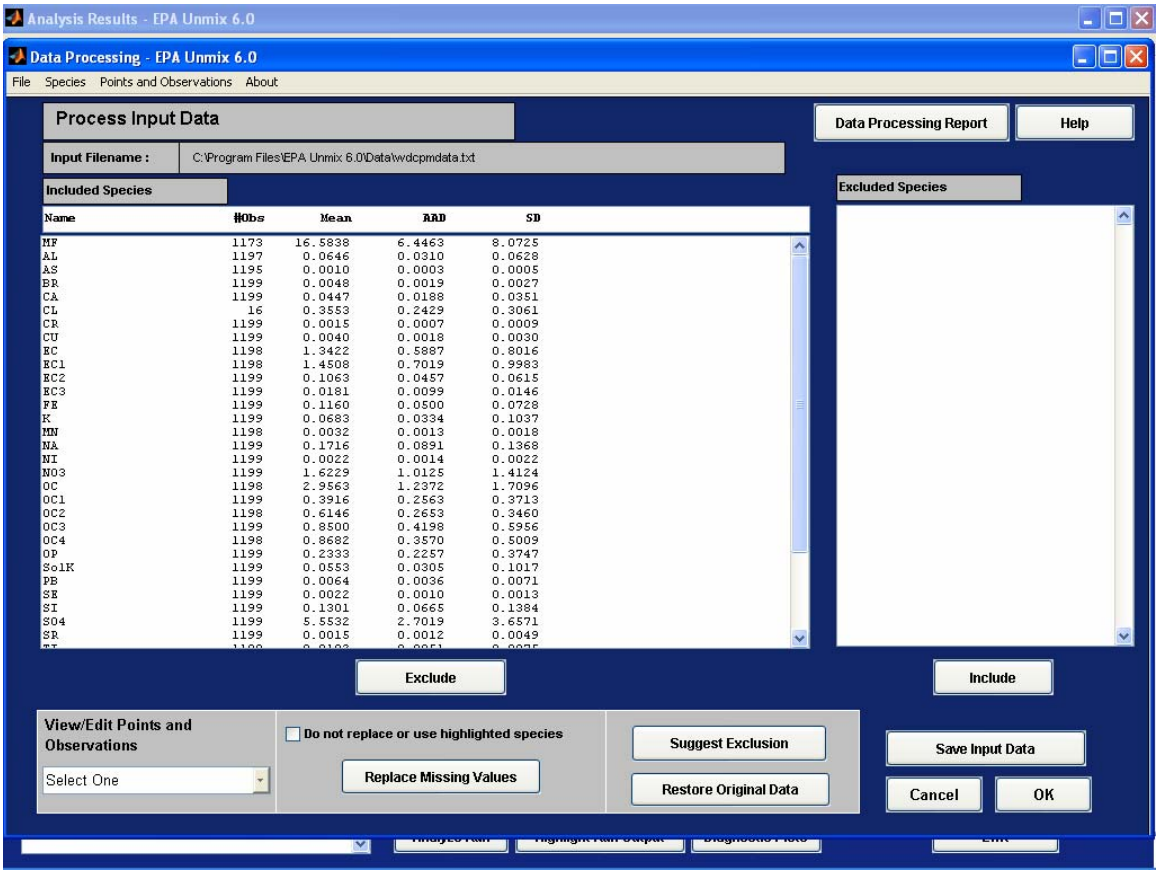


Figure 3: Data Processing window

A number of statistical parameters are displayed for each species: number of observations (# obs), mean, average absolute deviation (AAD), and the standard deviation (SD). The AAD is defined as the $mean(|Y|-\overline{Y})$, where Y is the species concentration, $\overline{Y}$ is the mean of the species concentrations, and $|Y|$ is the absolute value of Y. This measure does not square the distance from the mean, so it is less affected by extreme observations.

3.2 Suggest Exclusion

The results of all multivariate receptor models are degraded by including species in the model that are dominated by noise. Inevitably, errors in the noisy species will spread to all the sources determined by the model. Thus, it is best to remove species that are known to have a high level of noise from possible inclusion in the model. A general caveat concerning selection of species is in order. Sometimes more is less, that is, adding more species to the model can, under certain conditions, be detrimental to the model. Usually, the greater the number of species, the model produces more accurate and stable results. The additional information or 'signal' contributed by adding a species is larger than the error or 'noise'. However, if the species has a lot of error, it will to some degree mix into

US EPA ARCHIVE DOCUMENT

the whole model and corrupt the results. Thus, the model may have a better signal-to-noise ratio without the species in question. Generally, it is best to first add species that have the smallest measurement error.

Upon request, Unmix will provide a recommendation of species to exclude from the analysis. To utilize this feature, select the Suggest Exclusion button located in the data processing window and consider all species. You can use this tool on a small set of species or all of the species. For a small set of species, highlight the species you wish to include before pressing the "Suggest Exclusion" button.

Factor analysis is used to estimate the fraction of the variance of each species associated with factors it has in common with other species and the fractional amount that is associated uniquely with each species (technically known as the uniqueness). It is recommended that species with more than 50 percent of the variance due to error, or specific variance (SV) be considered for exclusion from further Unmix modeling. However, if there is a species that is known to be important and it has only a little more than 50 percent error, then it should not be excluded from the modeling process. In addition, species that are suggested for exclusion can be further evaluated using the View/Edit Points and Observations (in particular the View Time Series Plots, View/Edit Observations and Points, and View/Edit Influential Points commands) by selecting the appropriate option and choosing the currently highlighted species option in the ensuing options window. To unselect a species, hold the Ctrl key and click the left mouse key on the species you wish to unselect.

It is best to first exclude all species that do not pass even the minimum quality requirements. These excludable species may be those with a low number of observations or species with a low number of concentrations above the detection limit. For the wdcpmdata data, first select the species that you want to exclude from Unmix analysis. Select (highlight) OMC (organic carbon x constant for converting measured carbon to organic mass), ammSO4, ammNO3, SO2 (gas), and OP (pyrolized carbon). Then, select the Exclude button to move the species to the Excluded Species box.

The remaining species may be subject to more rigorous mathematical testing to determine if they are compatible with the rest of the data set. Select the Suggest Exclusion button and choose the "All Included" in the ensuing question box. After a few moments, a message will be displayed stating that some of the species have over 75% of their values missing. The species with the low number of values are Cl and Zr which only have 16 and 192 values (MF or mass has 1178 values), respectively. Select OK and the species that are recommended for exclusion will appear highlighted. Species such as NA are recommended for exclusion because they have a SV greater than 0.50 or 50 percent. Select the Exclude button to move the selected species from the Included to Excluded Species box. The highlighted species are shown in Figure 4 and the lists of included and excluded species are shown in Figure 5.

US EPA ARCHIVE DOCUMENT

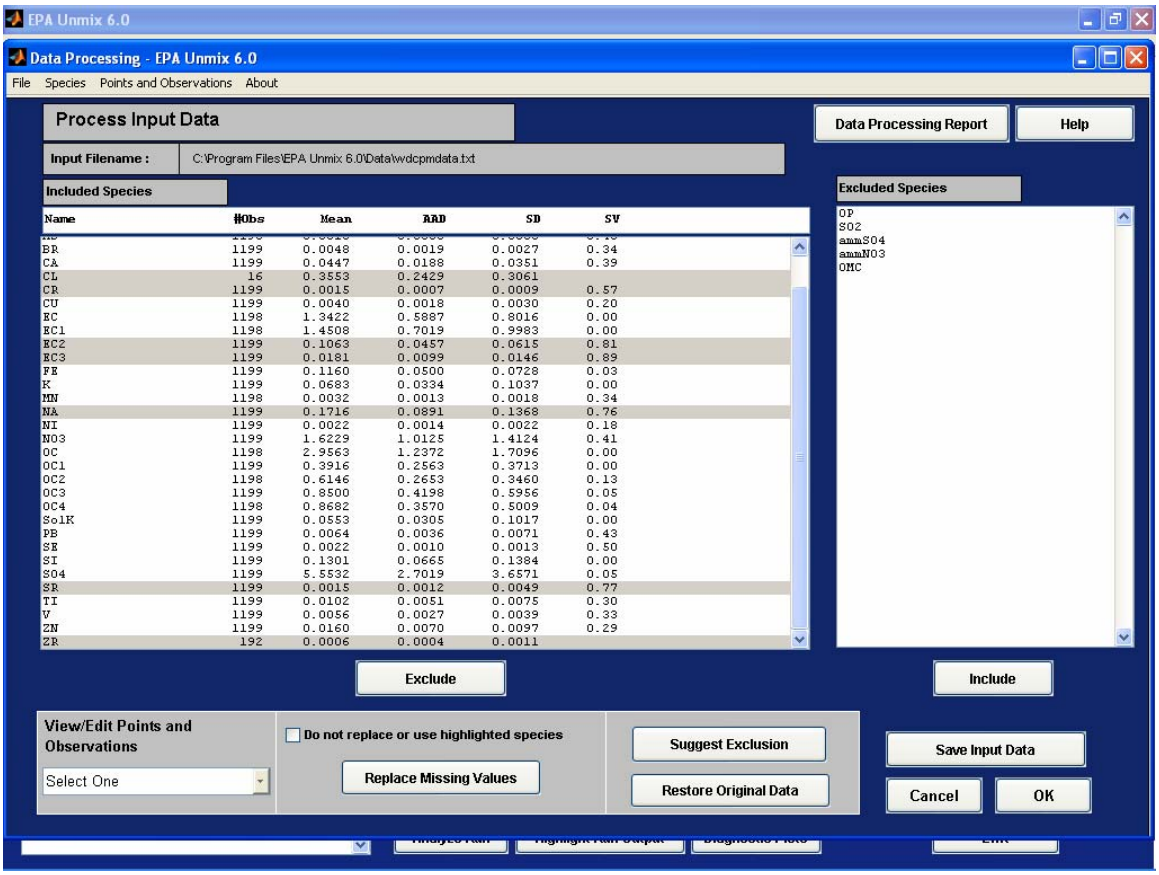


Figure 4: Suggest Exclusion

US EPA ARCHIVE DOCUMENT

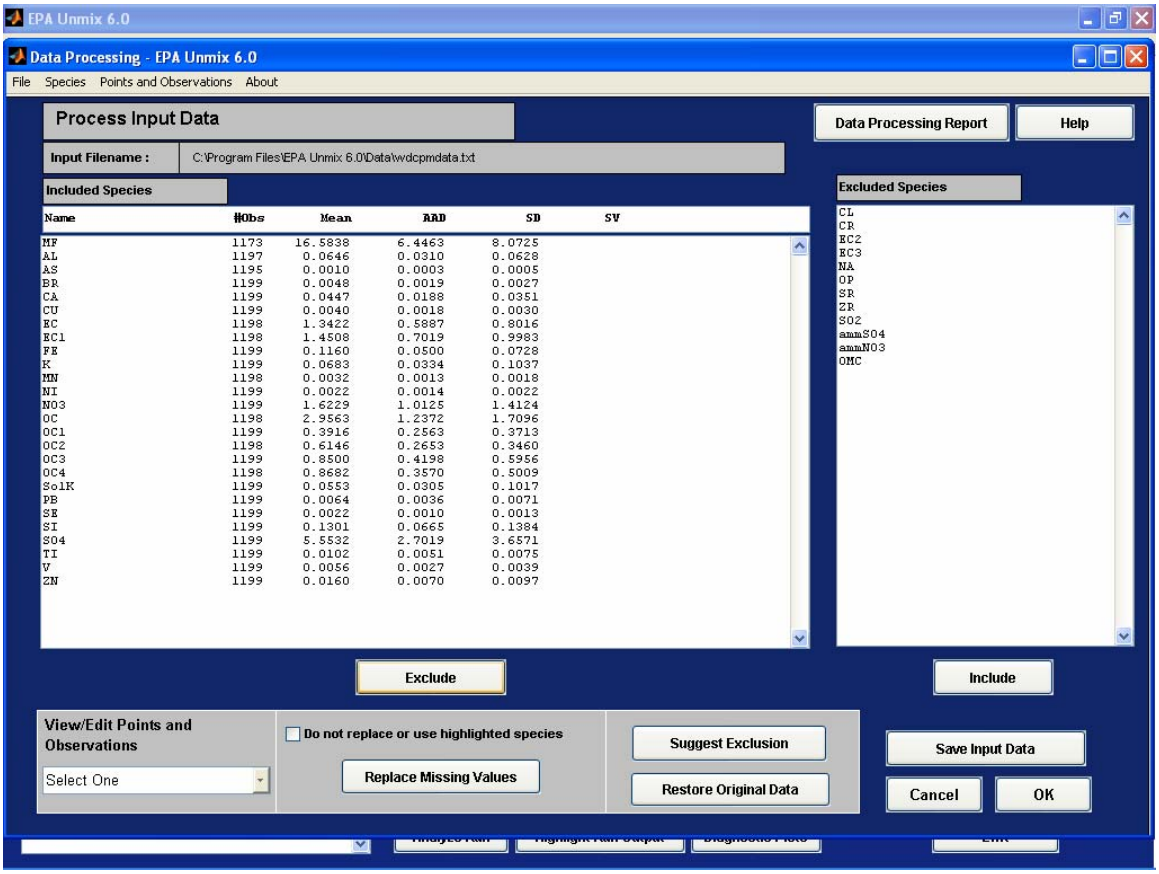


Figure 5: Included and Excluded Species

The included species from this window will be displayed on the main Unmix window after selecting the OK button. In the main Unmix window, species listed in the Selected Species box are used in the Unmix analysis. Species can be moved from the Unselected Species box to the Selected Species box by highlighting the specie(s) in the Unselected Species box and selecting the ← button. Species can also be moved to the Unselected Species box by highlighting the specie(s) in Selected Species box and selecting the → button. It is important to differentiate between excluded species and unselected species. The excluded species are simply not included in the analysis. Unselected species are available. They are essentially on the sidelines and can be brought in anytime for analysis.

The three buttons below the two arrow buttons are used to identify the species highlighted in the left-hand side selection box as the total species (TS), a tracer species (only emitted by one source), or the variable used to normalize the source compositions. Usually the normalization and TS are the same, as this gives a source composition as a mass fraction. However, in some applications, one may wish to normalize to some standard species to be consistent with other reported normalized compositions. In this case the normalization species will not be the same as the TS. If a TS is set, Unmix tests the source compositions to

US EPA ARCHIVE DOCUMENT

ensure that the sum of the species in a source is not greater than the TS. Thus, if the user wants Unmix to apply this constraint, Unmix must be informed as to which species is the total species. The total, tracer and the norm species buttons are toggle buttons. That is, after highlighting a species in the Selected Species window (left side box), repeated pressing of these buttons will alternatively select and unselect the chosen species as the requested type of species. Once a species is selected as a total, tracer, or normalization species, it can be deselected by highlighting the species in the left-hand box and selecting the same button again. Thus, if a species is set as a tracer and no tracer is desired, then it can be deselected by highlighting it and selecting the Tracer button. Specific examples of using the total, tracer, and normalization options for VOCs and PM can be found in Mukerjee et al. (2004), and Lewis et al. (2003).

3.3 Initial Species

Unmix requires the selection of species from the input file for the model. Selected species can be determined by the user or by using the Unmix Species Selection Tools. The Select Initial Species command uses the species with the largest loadings in the varimax factor analysis of the data to find a selection of species that gives a 4 or 5 source model that has very good signal-to-noise properties. If a 4 or 5 source solution is not found, a 3 source solution is attempted.

Before using the Unmix Species Selection Tools, move the species with high mean mass concentration over to the Species box. For example, select the species with a mean mass concentration greater than 1 $\mu g/m^3$. Select the Data Processing button in the upper right corner of the main window to go back to the Data Processing window in order to review the data. The species with mass concentrations greater than 1 $\mu g/m^3$ in the wdcpmdata are MF, EC, EC1, NO3, OC, and SO4 (species means are shown in the Data Processing window). Other species such as SI could be added, since it is a marker for soil or crustal material and it is typically present in quantities above the analytical method detection limit (XRF). Select the OK button to return to the main window. Note that you can track when species are designated as Total, Tracer, or Norm in the bottom right corner of the main window. In the current example, Figure 6 shows that the species MF is designated as Total and Norm. In order to view the results as mass fractions, highlight MF, and select the Total and Norm buttons. Select the first Species Selection Tool which is the Select Initial Species command shown in Figure 6.

US EPA ARCHIVE DOCUMENT

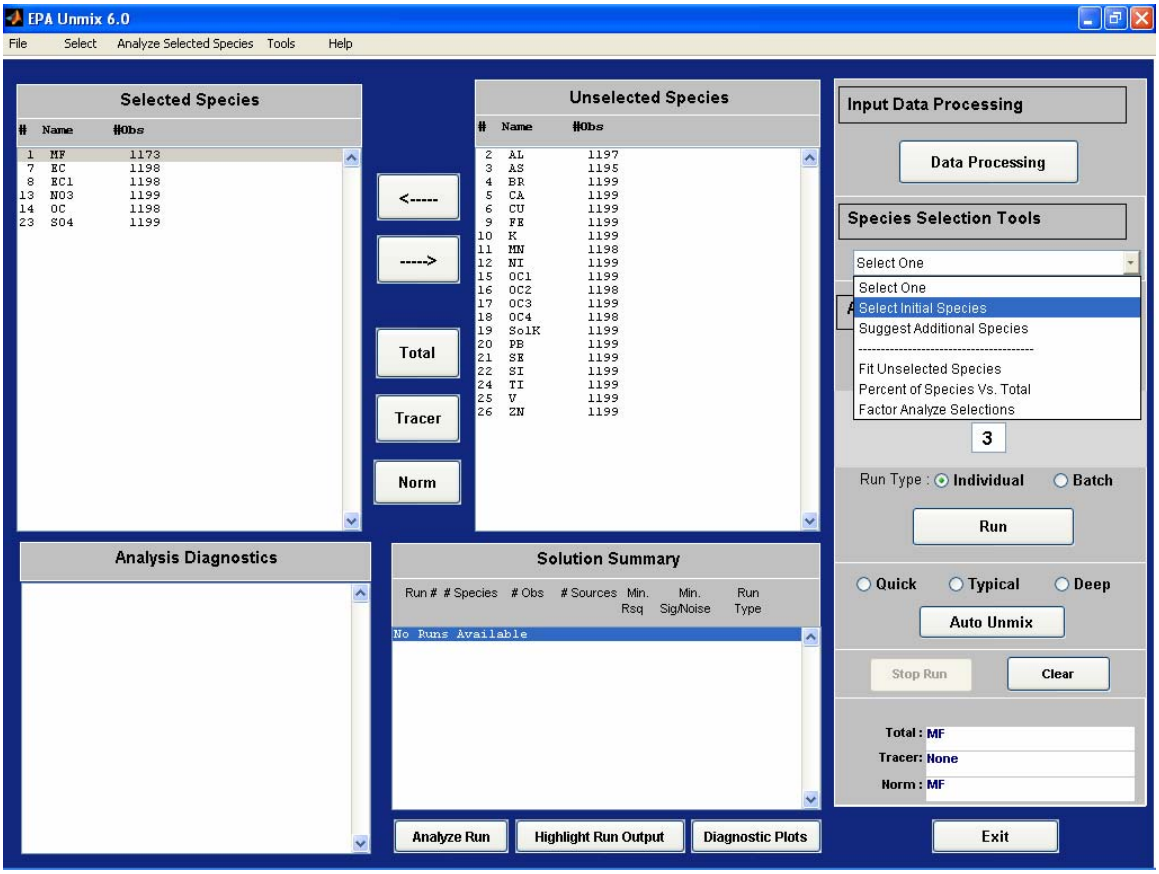


Figure 6: Select Initial Species

The Analysis Results window (Figure 7) shows the Unmix solution with the added species. A 5 source solution was found by adding Al, OC3, OC4, and SI to the initial list of selected species. In addition, the species are automatically moved from the Unselected to Selected Species box in the main window as shown in Figure 8. If an error message is displayed stating “Factor Rotation did not converge”, the user should replace zero’s or negative values in the input data set with a missing value code. The Replace Missing Data command in the Data Processing window can be used to replace the values before running Select Initial Species.

The preamble to the Unmix results is pretty much self-explanatory. The line that gives the Min Rsq, etc. does require some explanation. Min Rsq is the smallest r-squared value for any species in the model (r-squared for any species is greater than this value). The Min Sig/Noise is the smallest estimated signal-to-noise ratio of any of the factors included in the model.

US EPA ARCHIVE DOCUMENT

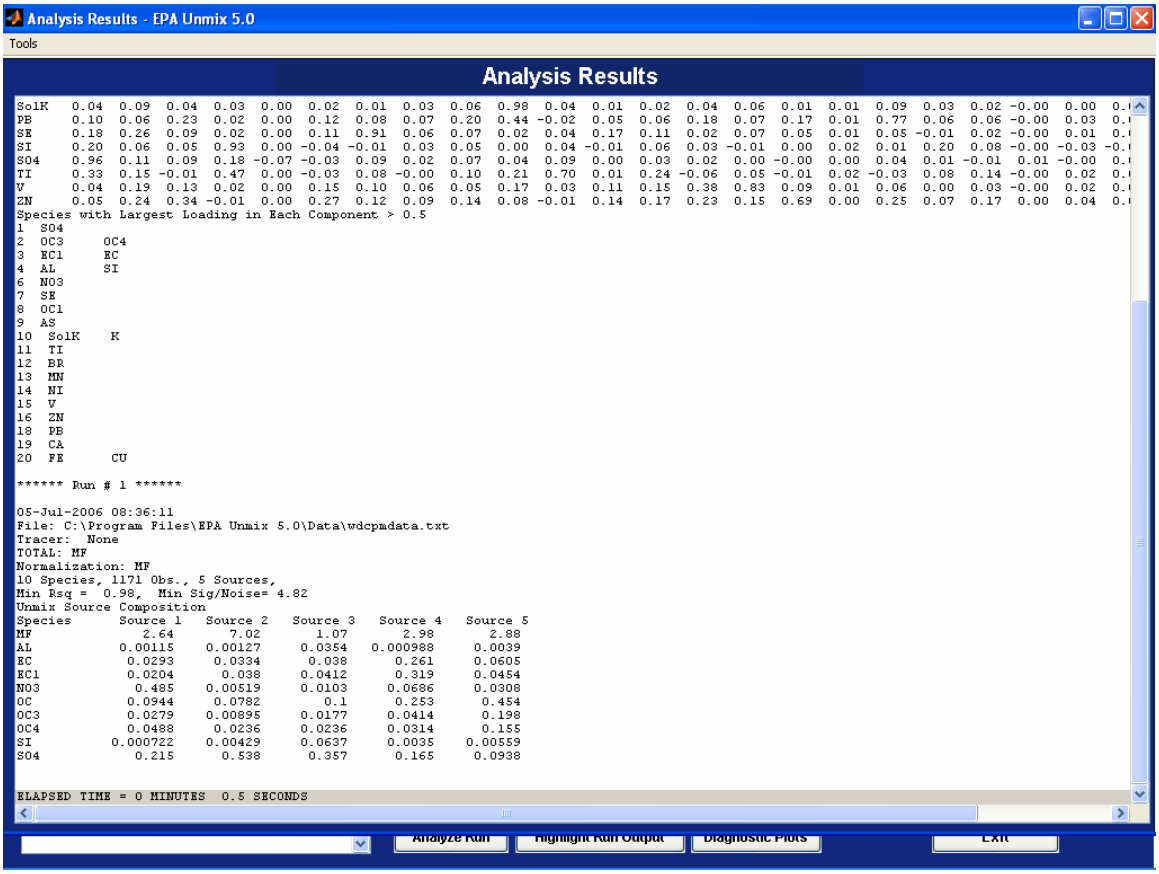


Figure 7: Initial Species Source Profiles

Select the main Unmix window to view a summary of the results. The Solution Summary box displays the summary information for the run. The highlighted line in the Solution Summary box tells us that there are 10 species and that there are 1171 observations. In addition, the r-squared values and the signal-to-noise ratios are listed for the run. In the example above, the data can be explained with a five source model with a minimum r-squared value of 0.98. This means that at least 98% of the variance of each species can be explained by five sources. Thus, the number in the Unmix display is the minimum r-squared value over all the species, not the overall r-squared of the fit. The run type is "I" for individual runs or "B" for batch mode (see Section 7). The signal to noise ratio is calculated by a procedure known as NUMFACT, which is described with several examples in Henry *et al.* (1999). The minimum number of sources is 3 and the recommended number of sources in the # Sources for the Current Run box is determined by the NUMFACT algorithm. The maximum number of potential sources is the number of sources with a signal-to-noise ratio greater than 1.5 (Analysis Diagnostics box). The user may see slightly different values for r-squared and signal-to-noise than those shown in Figure 8 because of the Monte Carlo nature of the underlying calculation. The user may wish to override the automatic selection and enter a new number of sources between 3 and maximum number of sources in the # Sources for the Current Run entry box.

US EPA ARCHIVE DOCUMENT

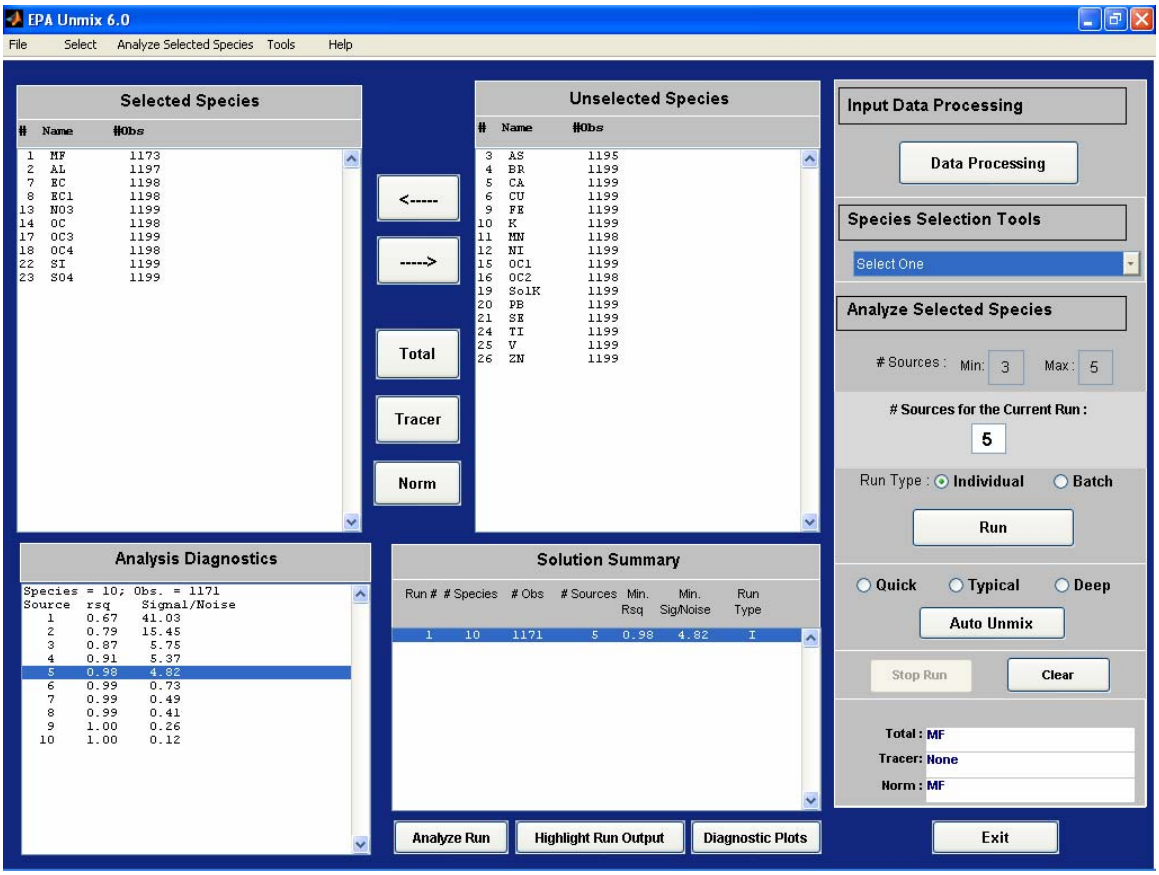


Figure 8: Initial Selected Species

3.4 Suggest Additional Species

The Suggest Additional Species command is used to create a list of species that can be added to an Unmix solution. Select the second Species Selection Tool, Suggest Additional Species command as shown in Figure 9. Select the All option to run both the SAFER and Influential Points (IP) Algorithm. Select the All button to run both the SAFER and Influential Points (IP) algorithms (see Section 5.2) and use the default spread parameters for the IP. The window shown in Figure 10 will appear after the SAFER and IP progress bars are displayed. Please note that selecting both the SAFER and IP can take a while to calculate depending on the number of unselected species and data observations.

US EPA ARCHIVE DOCUMENT

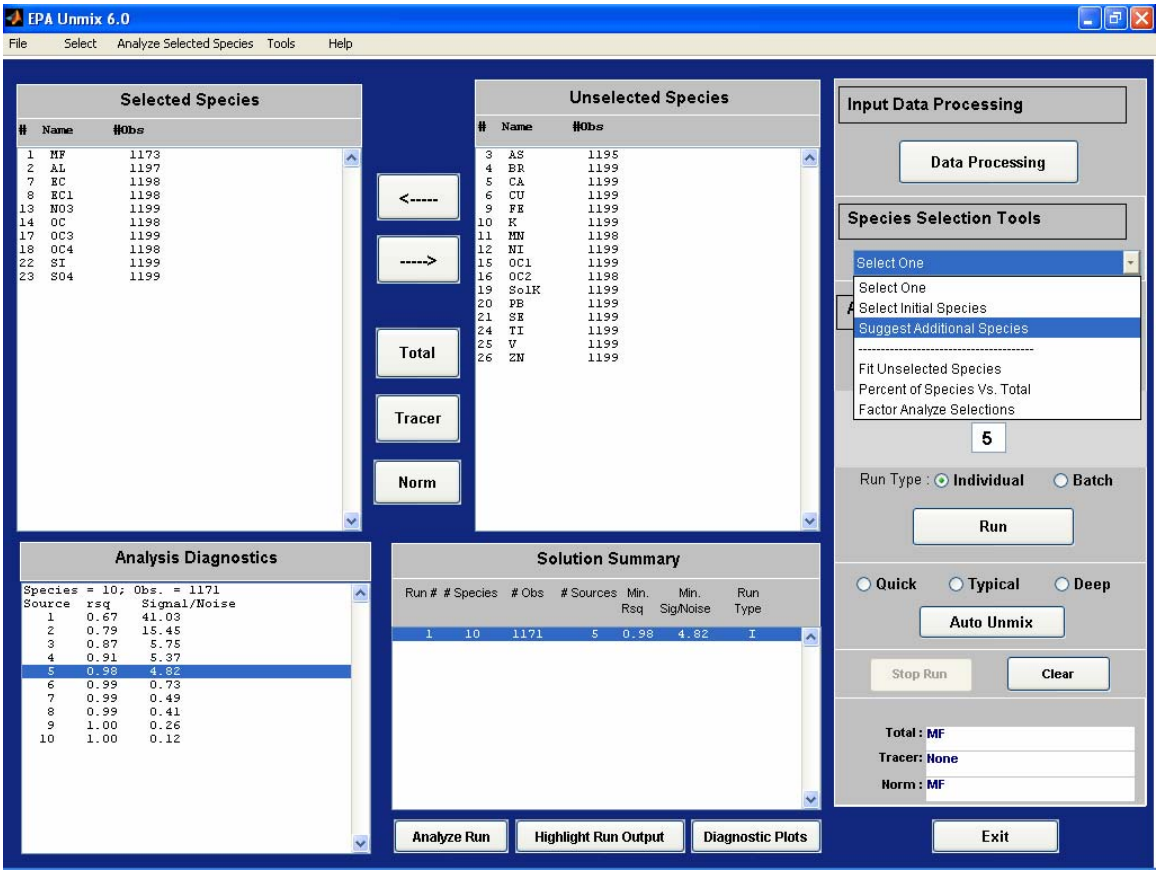


Figure 9: Suggest Additional Species

US EPA ARCHIVE DOCUMENT

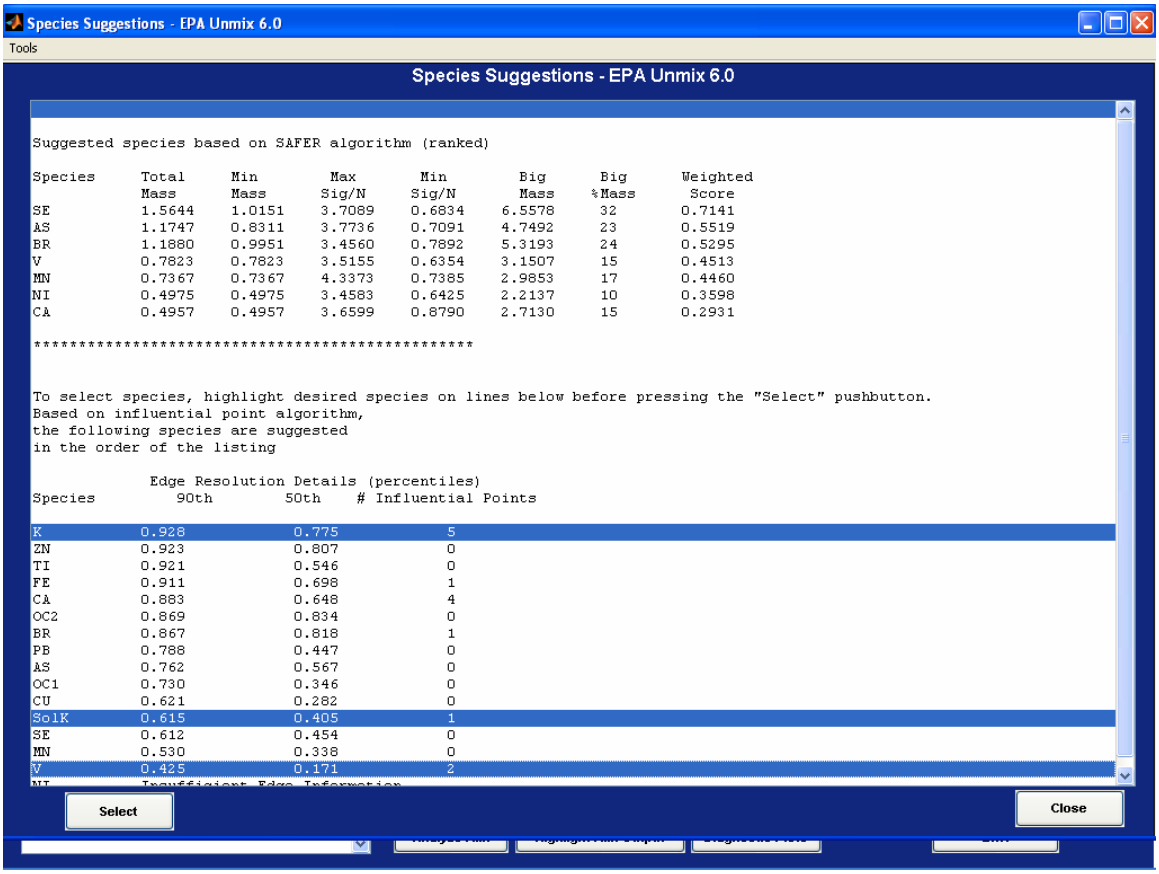


Figure 10: Suggestion Species

If the SAFER algorithm cannot suggest any additional species to add, only the influential species information will be displayed. Choose additional species by selecting the species in the list below the "****" line. Select species from the SAFER Algorithm that have a Max Signal to Noise (Max Sig/N) greater than 2 and Minimum Signal to Noise (Min Sig/N) less than 1. If a TS was selected, the SAFER Algorithm output also includes columns labeled Big Mass and Big Mass Percent. The idea is to distinguish between species that contribute to sources that explain only a small amount of TS. Assume two species explain 4 percent of the TS on average. However, the source contributions for one of the sources are sometimes very big, in fact sometimes this "small" source explains over 40 percent of the mass. A solution with such a source is to be preferred to a solution with a source that is always there but at a low level. In addition, it is recommended to select species from the Influential Point Algorithm that have a high edge resolution (90^{th} percentile greater than 0.80) or a low number of influential points.

The Suggest Additional Species command can be run multiple times by evaluating the species with the highest Weighted Score in the Unmix solution. If a feasible solution is found after adding the species to the Selected Species, use the Analyze Run, Fit Diagnostics command to evaluate the solution. For this

US EPA ARCHIVE DOCUMENT

example the following species are recommended: Se, Ca, and Si. If the number of significant and strong species abruptly decreases after adding a species, remove the species from the selected species box, and either try the next recommended species or finish using the Suggest Additional Species command. This process has been automated and is available by selecting the Auto Unmix (AU) command.

Another option is to select multiple species based on the SAFER Algorithm, Influential Points Algorithm, and species that are useful for identifying sources such as Se for coal combustion. However, selecting multiple species may result in a non-feasible solution. Select species that are components of potential PM sources in Washington DC: K (crustal, wood smoke), SolK (wood smoke), and V (residual oil) (Figure 10). The species will be highlighted in the Unselected Species box. Select the ← button to move the species to the Selected Species box and select the Run button. A new 7 source solution is displayed in the Analysis Results window (Figure 11). Two additional Unmix options are available for selecting additional species: Auto Unmix (AU) and Batch Mode. These options are discussed in sections 4 and 7.

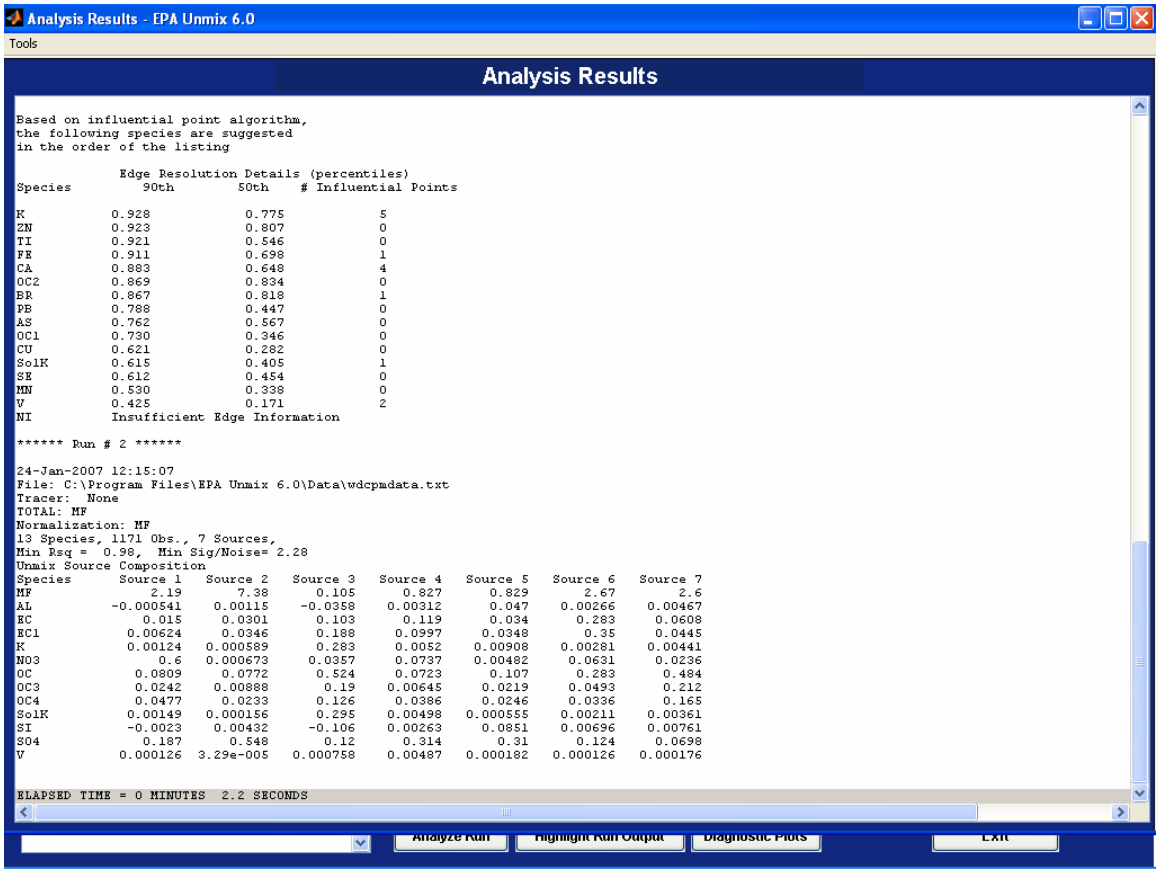


Figure 11: Suggest Additional Species Source Profiles

US EPA ARCHIVE DOCUMENT

3.5 Plot Distribution

Select the MF line of Run # 2 from the Analysis Results window and select the Tools, Plot Distribution command to create a pie chart showing the average source contributions (Figure 12). The secondary sulfate source (source 2) contributes 44% of the MF or $PM_{2.5}$. The color bar and legend can be added/customized and the figure can be saved.

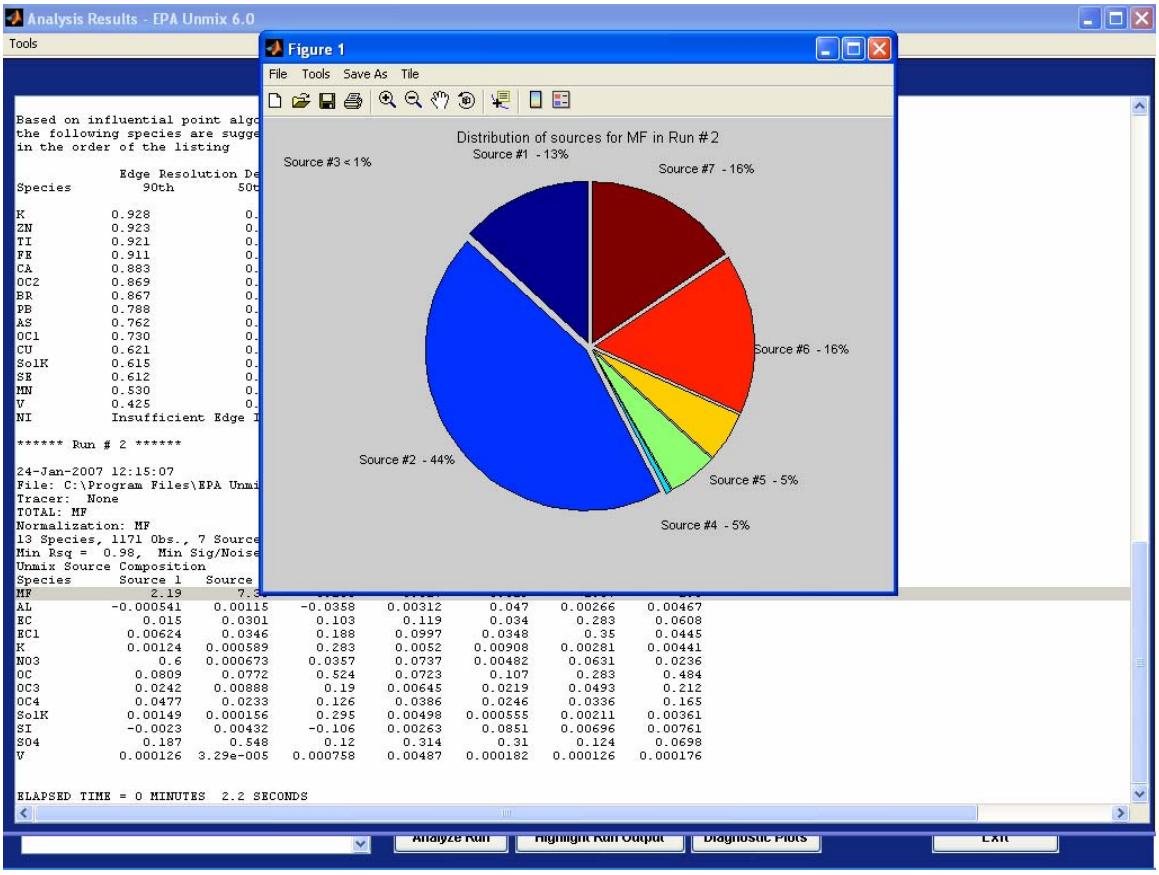


Figure 12: Analysis Results - Plot Distribution

3.6 Evaluating Results

An individual run can be evaluated by selecting one of the buttons below the Solution Summary box in the main Unmix window: Analyze Run, Highlight Run Output, or Diagnostic Plots. The Highlight Run Output option highlights the solution in the Analysis Results window. The Analyze Run window, shown in Figure 13, lists outputs that can be exported to the Analysis Results Window, text file, or Excel file. The predicted results from Unmix can be calculated by adding the Selected Species Data and Species Residual Files in Excel.

US EPA ARCHIVE DOCUMENT

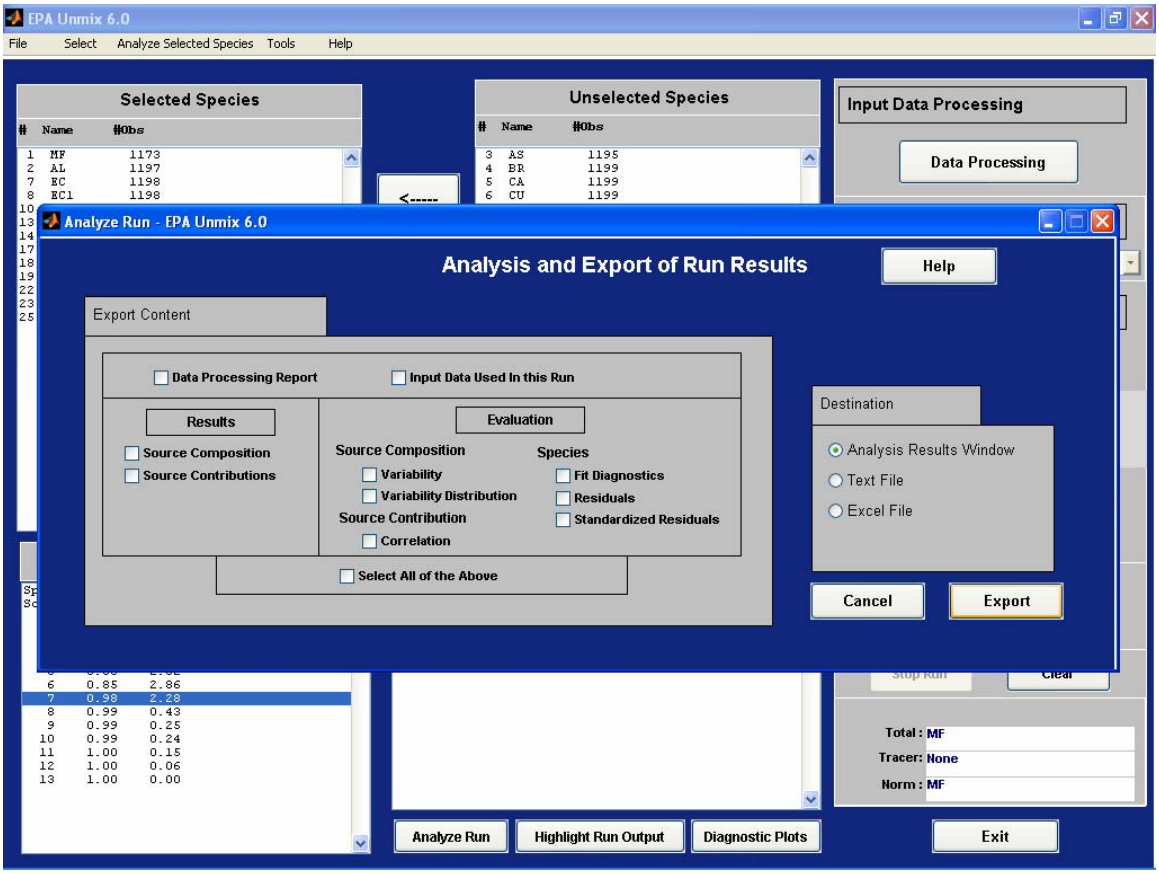


Figure 13: Analyze Output window

Fit Diagnostics example output is shown in Figure 14. The diagnostics include the regression statistics between the predicted and measured species concentrations, whether any species have a significant negative bias, strong/significant species in a source, and details on the variability distribution. If a normalization species was selected, the source contribution output will be for the species. For example, if PM was selected as the normalization species, then the source contributions output will contain PM source contributions for each sample.

US EPA ARCHIVE DOCUMENT

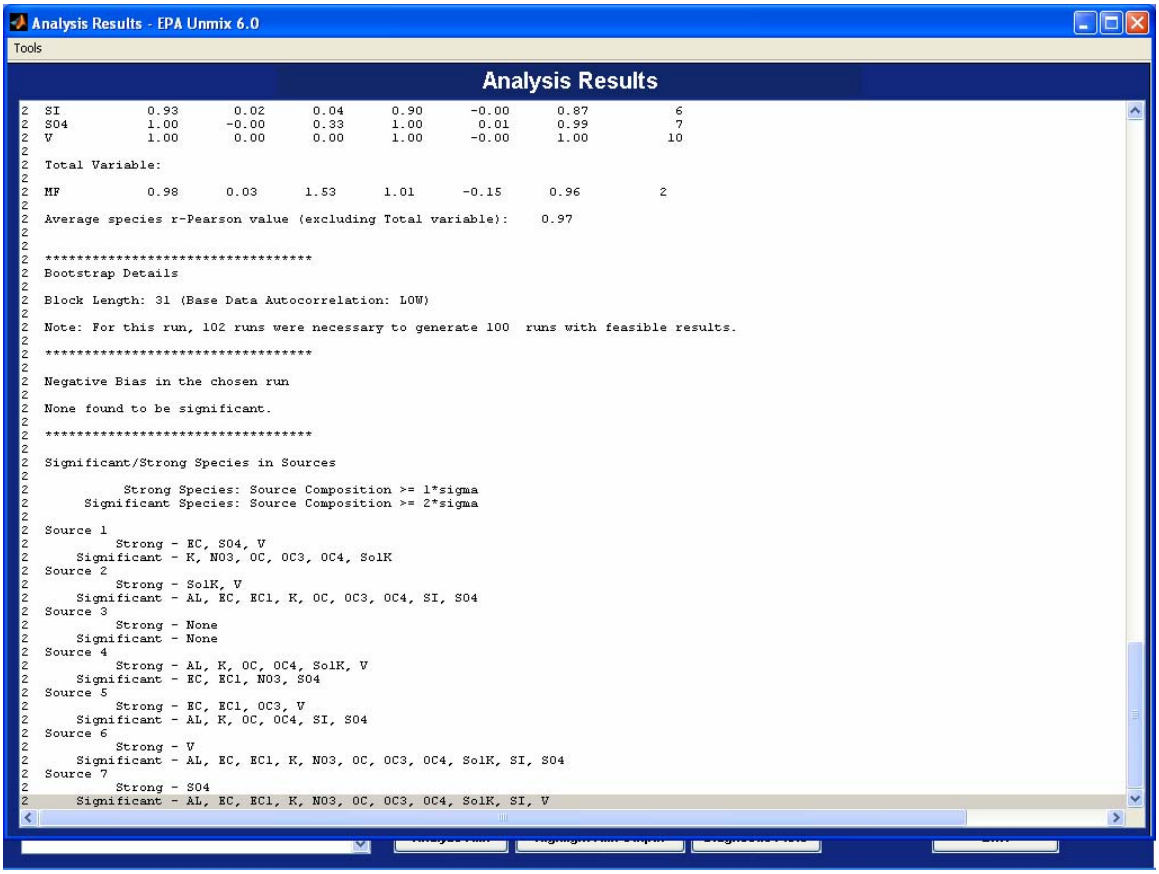


Figure 14: Analysis Results – Fit Diagnostics Example

It should be noted that Unmix can generate results with negative concentrations for species and the TS. The occasional small negative value is due to the effects of errors. Allowing for small negatives is necessary to reduce bias in estimating source compositions for species that are zero or very small. If a species is significantly negative, it is recommended that the species be removed from the Selected Species box. Reducing the number of sources in the model generally results in a solution without a negative TS. The Fit Diagnostics and guidance on interpreting the output is shown in Table 2.

US EPA ARCHIVE DOCUMENT

Table 2: Fit Diagnostics Guidance

<table>
<tr><td>Fit remaining species</td><td>Try adding species with r^2 greater than 0.80 in descending order, one at time, to the Selected Species box and run Unmix. Select species that will add meaning to the existing profile.</td></tr>
<tr><td>Correlation, differences, and regression coefficients</td><td>Evaluate species with low r^2 values (< 0.30) or with greater than 1 outlier for 100 points using the View/Edit Points and Observations tools. If species still have r^2 values less than 0.30 and do not significantly aid in the interpretation of the profiles, remove the species from the Selected Species box.</td></tr>
<tr><td>Bootstrap Details</td><td>After evaluating the bootstrap variability percentile summary in the Fit Diagnostics and the Diagnostic Plots, Variability Distribution option, use the following guidelines for the number of attempts to obtain 100 feasible solutions.
<table>
<tr><th>Data Set Size</th><th># of attempts to obtain 100 feasible solutions (rough estimates only)</th></tr>
<tr><td>> 600 samples</td><td>Up to 140</td></tr>
<tr><td>400 – 600 samples</td><td>Up to 150</td></tr>
<tr><td>< 400 samples</td><td>Up to 160</td></tr>
</table></td></tr>
<tr><td>Negative Bias</td><td>Species with significant negative bias should be evaluated in more detail using the View/Edit Points and Observations tools. Remove any species from the Selected Species box with significant negative bias.</td></tr>
<tr><td>Significant/Strong Species</td><td>Strong Species have a contribution to a source that is greater than or equal to 1 times the standard deviation of the bootstrap estimated variability (sigma). Significant species have a contribution to a source that is greater than or equal to 2 times sigma. Most sources should have both strong and significant species, with the significant species having large signals-to-noise ratio. Look for reasonableness of the profiles and how species group. Only one or two sources should have neither significant nor strong species.</td></tr>
<tr><td>Species Report</td><td>
<table>
<tr><th>Values</th><th>Interpretation</th></tr>
<tr><td>0</td><td>Base run source profile value not contained in the IQR.</td></tr>
<tr><td>1</td><td>Base run source profile contained in the IQR but not centered.</td></tr>
<tr><td>2</td><td>Base run source profile contained in the IQR and is centered.</td></tr>
<tr><td>+</td><td>2.5th percentile value of source profiles from the bootstrap runs > 0</td></tr>
</table>
IQR – interquartile range, between the 25th and 75th percentile. Interpretation of species in source variability percentiles should focus on species with little influence of outliers (2+). Species not strongly influenced by outliers are a 1+ and should be interpreted. Species that are impacted more by outliers are 0, and generally have low contributions to a source and should be interpreted with caution. Total Mass species should be category 1+ or 2+, however, a source with a total mass summary value of 1 is acceptable because the source may explain species variability without having a significant contribution to the total mass. Each source should have multiple 1+ species.</td></tr>
</table>

US EPA ARCHIVE DOCUMENT

In the main Unmix window, select the Diagnostic Plots button. The Diagnostic Plots window contains many plotting options and an example from the Fit Diagnostics option is displayed in Figure 15. The predicted vs. measured concentrations are shown in the top plot and the regression results are shown in the bottom plot. A figure is generated for each of the selected species. Use the << Previous (Prev) and Next >> buttons to view each of the figures.

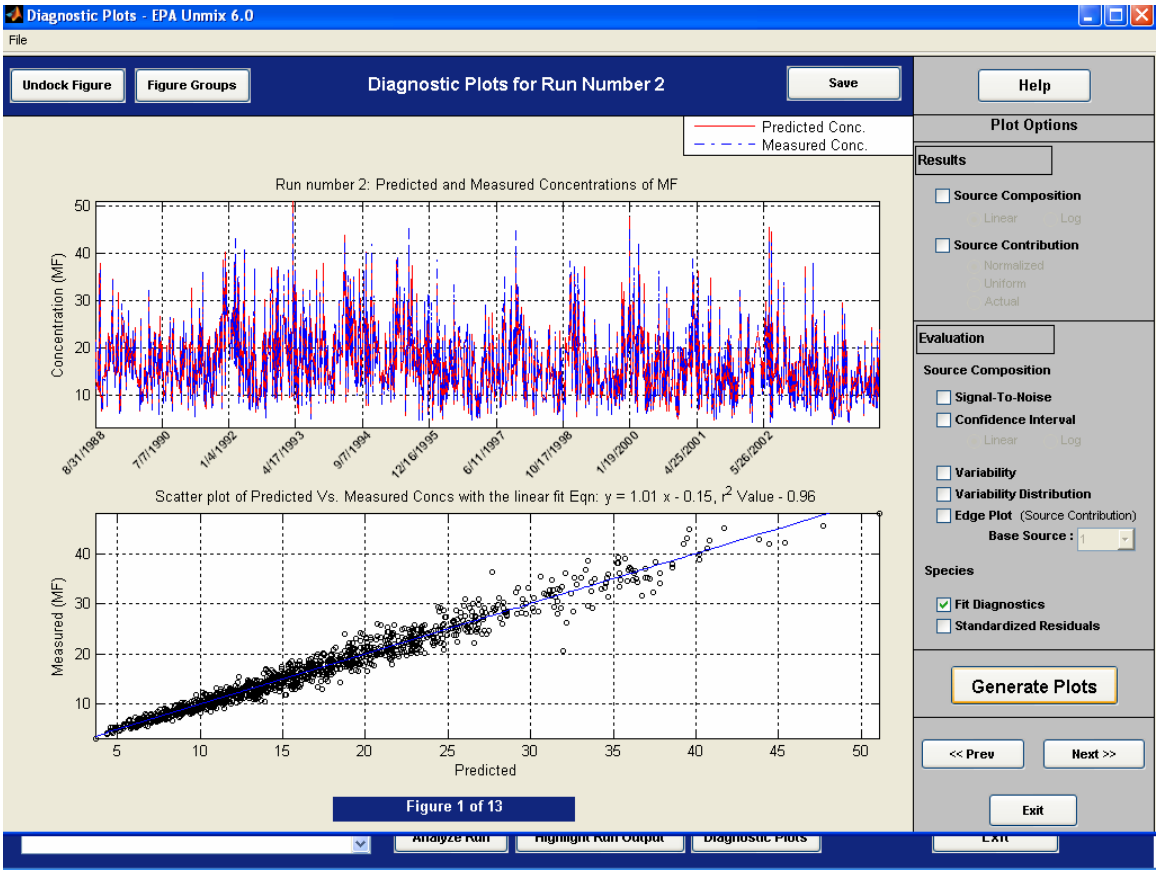


Figure 15: Diagnostic Plots – Fit Diagnostics Example

The majority of the options in the Analyze Run and Diagnostic Plots window are self-explanatory and a Help button is available for details. Some explanation is required for the Source Contribution plots: Normalized scale plots have source contributions that are scaled to range between 0 and 100, Uniform scale plots have source contributions that are scaled to range between 0 and 1, Actual scale plots are not scaled and these plots will display the true source contribution values.

In addition to evaluating the Fit Diagnostics plots, the standardized residuals should be evaluated by selecting the Diagnostic Plots, Standardized Residuals option. The standardized residuals from residuals (difference between predicted

US EPA ARCHIVE DOCUMENT

and measured concentrations) are calculated by first mean centering the residuals and dividing the result by the standard deviation of the residuals.

The standardized residuals distribution should appear similar to the standard normal distribution (zero mean and standard deviation of 1) with the majority of standardized residuals should be between -3 and +3. Species with significant number of standardized residuals outside this range should be evaluated in more detail using the Data Processing window under View/Edit Points and Observations options.

The figures can be saved using the Save button. If many pages of figures are saved using the Save All button, the figures are automatically named with the figure type followed by the figure number (i.e. Source Composition_fig_1). These figures can be grouped together in one file using Adobe Acrobat's Create PDF "From Multiple Files" option.

3.7 No Feasible Solution

Unmix looks for edges in a multidimensional plot of the data. If N sources are sought, Unmix needs to find N edges. From each edge, the normalized source contributions of a source can be estimated. If fewer than N edges are found, earlier versions of Unmix simply reported "No Feasible Solution", even though it may have found N-1 edges. Unmix also reports "No Feasible Solution" if N or more edges are found but these lead to source compositions that have negative source contributions that are too large or too numerous. Again, earlier versions only reported "No Feasible Solution" even though information was gathered about the sources. Unmix 6.0 reports estimates of partial solutions and other information in order to give more guidance to the user to produce a better solution.

The nature of the partial solutions reported by Unmix depends on the number of edges found and whether or not a total species has been set by the user. In the following, 'edge' and 'possible source' are used interchangeably. When no feasible solution is found, there are three possible types of partial solutions.

- If a total species is not set, Unmix reports the species that have a correlation ≥ 0.8 with the contributions associated with the edge. If there are no such species, 'None' is reported.
- If a total species is set, the behavior of the solution depends on the number of edges found. If the number of edges is ≤ the number of sources, Unmix does a complex calculation to estimate the percentage of the total associated with each possible source (edge). Unmix reports the estimated total percentage and the species that have 50% or more of their average concentration explained by the edge. If there are no such species, 'None' is reported.
- If the number of edges is > the number of sources, Unmix reports the

US EPA ARCHIVE DOCUMENT

correlation of the total species with the possible source and the species that have a correlation of ≥ 0.8 with the possible source. If there are no such species, 'None' is reported.

A partial solution is generated when Cu, and Zn are added to the selected species shown in Figure 8. Seven sources are listed with major species (Figure 16). The largest possible source is secondary sulfate with MF and SO4 (44%). Other possible sources are nitrate (NO3), crustal (Al, SI), motor vehicles (EC, EC1), Zn, and Cu sources. One source was also identified that has no major species.

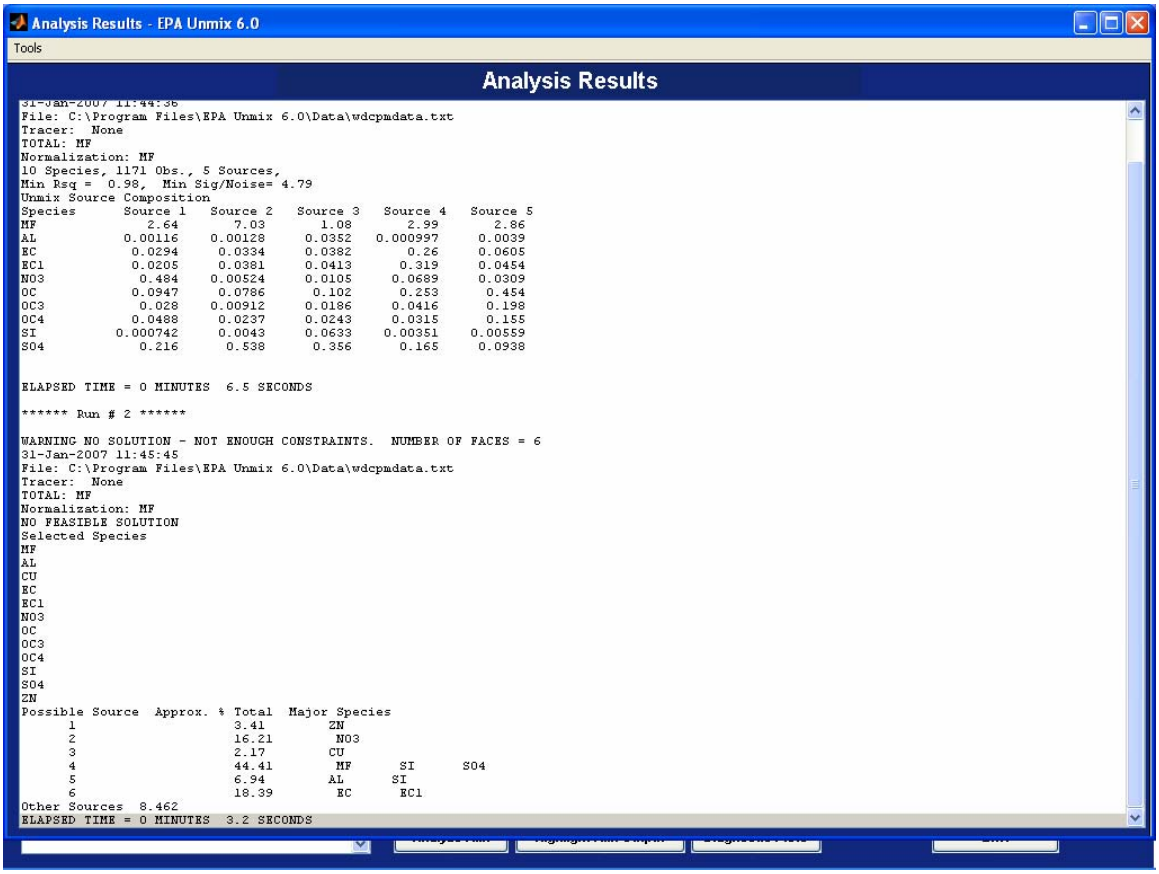


Figure 16: Partial Solution

3.8 Estimate Source Profile Uncertainties

The variability of the base run, commonly referred as uncertainties in source profile, are estimated using a block bootstrap method. A note on the use of the terminology used is necessary. Source Profile Uncertainties is a phrase generally used by researchers to evaluate the robustness or the stability of a chosen source profile. The source profile is considered robust or stable if small changes to the input data produces proportionally small change in the results. In essence,

US EPA ARCHIVE DOCUMENT

the bootstrapping technique helps measure the variability in the source profile produced by the variability in the input concentration data. However, it is important to note that variability and uncertainty are not technically equivalent. Variability is a point estimate of the uncertainty just like the mean being a point estimate of the distribution. Uncertainty associated with the source profile can be constructed only by running multiple block bootstrap runs on the same same source profile. This is similar to the process of collecting samples (eg. Number of hurricanes to hit Florida in September) to be able to construct the underlying unknown distribution. In other words, uncertainty is the distribution associated with the sample space of variability.

Bootstrap data sets are constructed by sampling, with replacement, from the original data set. This randomly re-sampled data may not retain the positive serial correlation of the original, which could lead to errors canceling out when they should not. The solution is to break the original data into blocks of data that are long enough to retain the serial correlation. The bootstrap samples are obtained by re-sampling the blocks of data with replacement. Blocks of data are chosen until the new data created has the same number of observations as the original data. This data set is then used as the input to the Unmix. Since, there is no guarantee of a feasible solution for every bootstrap data set, bootstrap data sets are created and run until one hundred feasible solutions are obtained. They are used to calculate the standard deviation (or sigma) and percentile distribution of the source compositions. The variability computation algorithms are described in detail in Appendix C.

The singular value decomposition of the bootstrap sample is calculated and the duality principle translates the known (normalized) source contributions into edges in the bootstrap sample's principal component space, called the bootstrap V-space. These edges are used as initial guesses to find edges in the bootstrap V-space. After finding these new edges, everything is the same as in basic Unmix with one exception. Occasionally if there is a lot of error or outliers in the data, the initial guesses (as good as they are) do not converge on a new edge. If this occurs, the Variability Algorithm will generate a different initial guess to look for the edge.

Variability or uncertainties in the source compositions are estimated for feasible solutions by selecting the Analyze Run button, Variability option. The Analysis Results Window will display three types of variability distribution diagnostics for each source. The first one is the sigma, and the composition divided by 2 times sigma. The composition divided by 2 times sigma represents the signal to noise ratio and is greater than 1 for species that contribute significantly to a source. The next is the set of percentile values on the range of bootstrap run values. The 2.5^{th} to 97.5^{th} range in percentiles is an estimate of the 95% confidence interval that accounts for the non-negativity constraint in Unmix. Finally, a new method is used to provide a 90% and 95% confidence intervals. This method provides the range of bootstrap run values as a percent of the base run value and centered

US EPA ARCHIVE DOCUMENT

about the base run values. We will refer to this as the Discrete Difference Percentile (DDP) method and its algorithm is described below.

Before describing the DDP method, we will provide the justification for introducing a new metric. While the first two methods (Sigma and Percentile) provide a statistical overview of the nature of the bootstrap runs, neither method provides a reportable statistic that connects the source profile that is being evaluated to the summary information provided by those methods . The sigma method provides the variation about the mean whereas the percentile method provides an insight in to the spread about the median values of the bootstrap source profiles. But, there isn't a clear method to compare the mean nor the median of the bootstrap source profiles to the chosen source profile under investigation. The DDP method is designed to address these shortcomings.

The DDP method contains the following steps. The procedure is applied to each entry in the source profile matrix.

1. For the chosen entry in the source profile matrix, collect the corresponding entries in the bootstrap matrices.
2. Construct the absolute difference values by taking the absolute value of the difference between the chosen source profile entry and the corresponding entries in the bootstrap source profile matrices.
3. Compute the 90^{th} and 95^{th} percentile value of the collected absolute difference values.
4. Repeat this process for each entry in the chosen source profile matrix.

For example, the second source shown in Figure 11 (secondary sulfate) has a 95 % confidence level value of 4% for SO4 with the mass fraction value of 0.55 (MF specified as normalization species). The 95% confidence interval using the ranked value metric states that 95% of the bootstrap source composition species mass fraction lies between 0.53 (0.55 – 4%*0.55 = 0.55 – 0.02) and 0.57 (0.55 + 4%*0.55 = 0.55 + 0.02). If a normalization species is not used, the confidence interval can be calculated using the example shown above. Relative source contribution values are typically reported for PM studies and if a normalization species is specified the ranked value output also provides the percent and confidence intervals. For source 2, the MF percent is 44 and the relative 95% confidence interval is 8%. The contribution of the secondary sulfate to MF or $PM_{2.5}$ is 44 ± 8%.

The advantages of the DDP method over the other two methods are two-fold. Firstly, the method gives an exact quantity that describes the variation about the base run source profile which is exactly the quantity that is being evaluated and not the variation about the mean or the median of the bootstrap source profiles. Second, the use of percentile to predict the variation about the base run source profile is a statistical value and not a value tied a specific set of bootstrap runs. The use of sigma to describe the variation is one such case where the

US EPA ARCHIVE DOCUMENT

descriptive value (sigma) is tied to the specific bootstrap run and may vary from one bootstrap run to another. Nevertheless, the user should use all the available inferential information to evaluate the base run source profile. If, in particular, the conclusions to be drawn from the three methods vary significantly, the user should attempt to explain the possible implications of the differences while reporting the results.

It is recommended that the DDP method output, at the minimum, be included in any reports or publications to summarize the source profile variability associated with a base run source composition. Also, the user is advised to choose the 90% confidence level for small and/or noisy data set and 95% for medium and large data sets (greater than 250 observations).

The variability information can also be viewed graphically by selecting the Diagnostic Plots button, Confidence Interval option. The plot displays the 2.5th to 97.5th percentile range and the source profile composition is shown by an asterisk (*) for each source. If a 2.5th percentile value is negative, it is displayed as 10^{-4} in the plot (see Section 3.6 for a discussion of small negative values). An Variability Estimate Plot for the source profiles is shown in Figure 17.

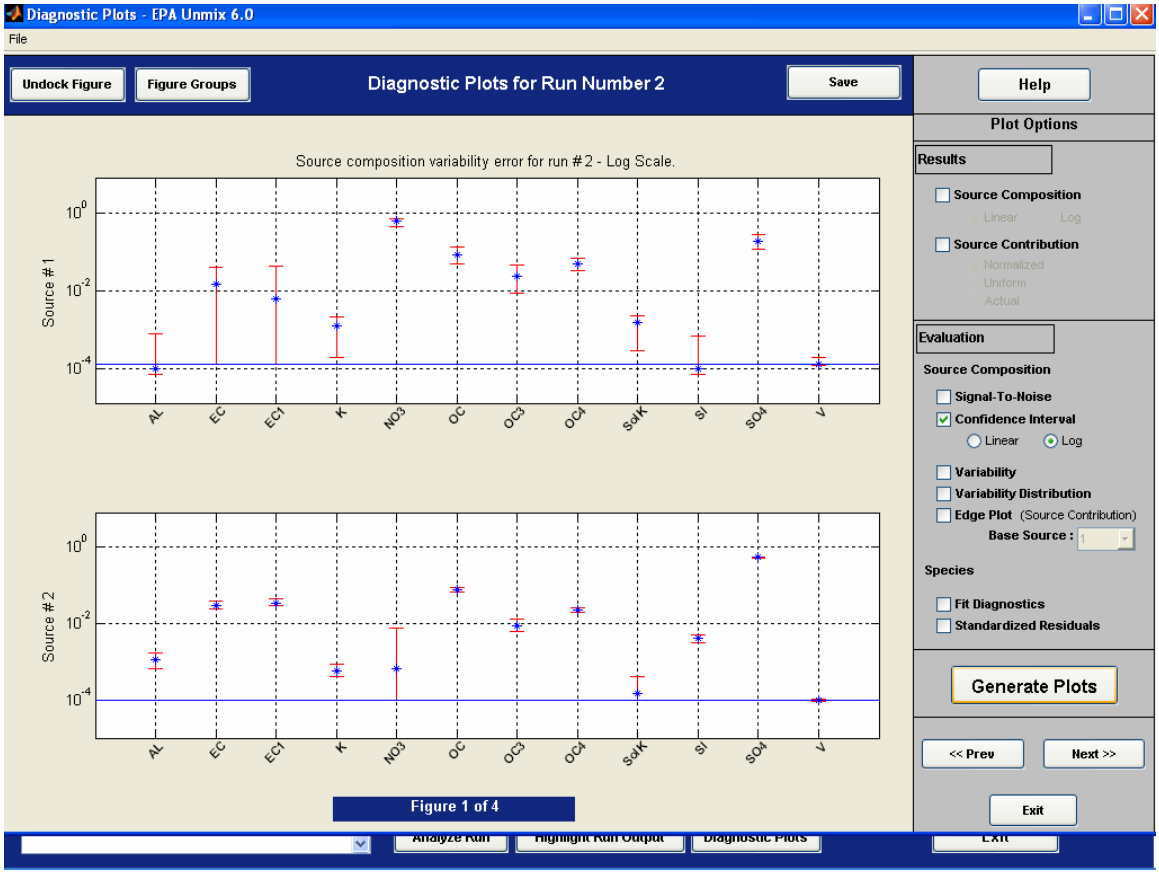


Figure 17: Variability Estimate Plot

US EPA ARCHIVE DOCUMENT

The source profile variability plots should be evaluated before interpreting the source profiles. Select the Diagnostics Plots, Variability option and two subplots will be generated for each source as shown in the following figure. Use the Next button to view the other 6 sources.

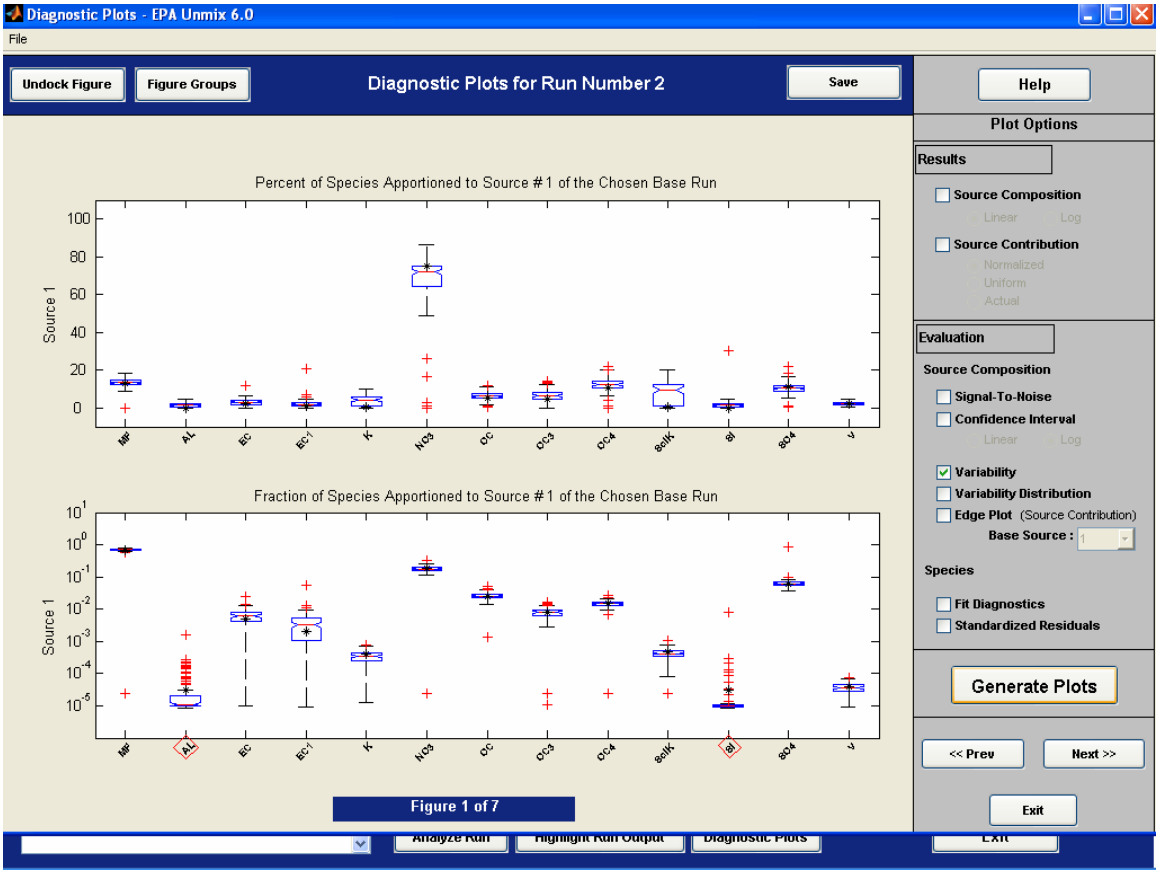


Figure 18: Source Profile Variability Plot

The first subplot titled “Percent of Species Apportioned to Source #1 of the Chosen Base Run” displays the variability in the percent of the species apportioned to the currently viewed source. The figure is generated as follows.

1. If the chosen base profile contains M species and N sources, then the bootstrap matrix will of the dimension M x N x 100, where 100 equals the number of feasible bootstrap runs that are used to plot the above figure. Each box plot in the figure above uses 100 data points for the each species and source. There will M box plots per subplot, one for each species and N sets of subplots, one for each source.
2. For each bootstrap run, the bootstrap matrix is normalized using the row sum.
3. The base run profiles are also normalized by their row sums.

US EPA ARCHIVE DOCUMENT

4. The values for each species and source over all the bootstrap runs are then used to construct the box plot figure. The normalized values from the base run are marked by "*" in the subplot.
5. The red pluses represent the outliers in the plotted data set.

The first subplot highlights species that contributed to the source. In other words, the higher valued species in this subplot are candidates that highly influenced this source in its composition. It should be noted that the pie chart percentages in Figure 12 may not match the percent of species attributed to a source since they are source contribution (profile x source contribution) and profile based calculations, respectively.

The second subplot represents the same bootstrap matrix using a slightly different metric. The procedure used to plot the bottom subplot is as follows.

1. Similar to the top plot, each box plot in the figure uses 100 data points for the each species and source, M box plots per subplot, one for each species and N sets of subplots, one for each source.
2. For each bootstrap run, the bootstrap matrix is normalized using the column sum. The base run source profile is normalized the same way.
3. The normalized values are used to plot the box plot. Due to this normalization, the y-axis limits will always range between 0 and 1. Therefore, the axis is marked in logarithmic scale to highlight the species with smaller values and ranges. The normalized profiles are marked by "*" as in the top subplot.
4. The species titles for which the normalized profiles do not fall within the inter-quartile range are shown in a red outlined box (see SI in the second subplot in Figure 18).

The first subplot in Figure 18 highlights NO3 as being the largest significant contributors to Source 1. The subplot matches well with the output from the sigma-based output shown in Figure 14 which listed the significant species contributing to Source 1 were K, NO3, OC, OC3, OC4, and SolK.

The second subplot shows the variation in the species apportionment. This subplot should be mostly used to confirm inferences from the first subplot. For instance, an influential species from the first subplot may be confirmed if their variations are not too large in the second subplot. The vice versa may hold true in some cases. An influential candidate may be rejected if their variations appear to be larger than expected in the second plot

Nevertheless, in both cases, the user should use all available information before arriving at a conclusion. This includes the information of known local sources, data set anomalies, and other inferential data provided by the model. As is true with any chosen method, certain data are favored more than the others. The user should be aware of the limitations and adjust their conclusion accordingly.

US EPA ARCHIVE DOCUMENT

The second subplot in Figure 18 shows that the SI contributing to Source 1 should be interpreted with caution since the variability is significantly higher in the second subplot compared to the first and the presence of many outliers. This species is shown in a red outlined box because the base run "*" is not within the IQR. In these cases, the user should use other available information such as local emission inventory and/or measured source profile information and ensure that the output from the model is not over-interpreted. Source Profile Variability is an invaluable tool that helps prevent the model from identifying incorrect or in some cases non-existent sources.

The second subplot results in Figure 18 are summarized in tabular form in the Analyze Run, Fit Diagnostics option (see Figure 19). The ranking goes from 0 to 2 with 0 being highly suspect to 2 being the ideal. Along with that, the presence of the "+" sign indicates the non-negative nature of the 2.5th percentile value of the source composition value from the bootstrap runs.

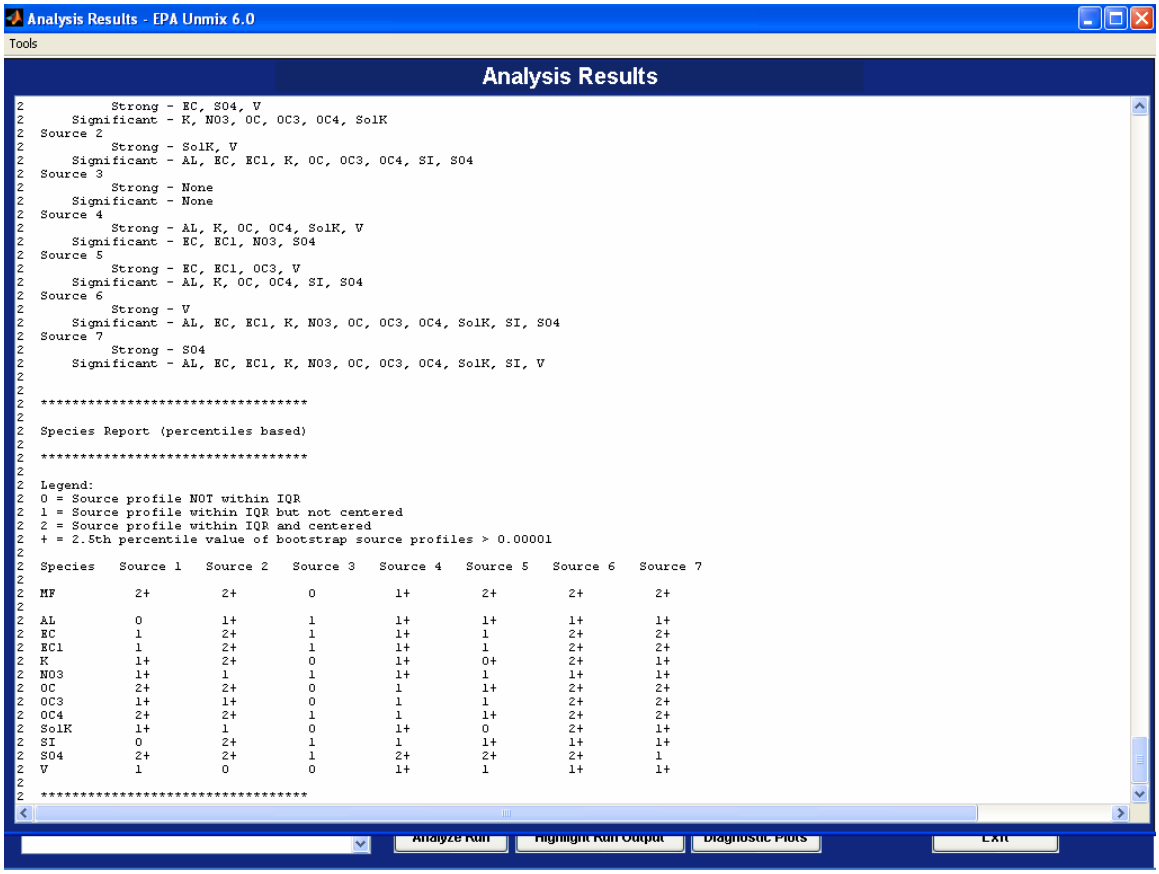


Figure 19: Species Report (percentile based)

The distribution of the variability estimates should be evaluated for small or medium size data sets (less than 600 observations) or data sets that contain infrequent high species measurements. Select the Analyze Run button, and the

US EPA ARCHIVE DOCUMENT

Variability Distribution option. A table is generated that shows the profile and variability, the results of 4 independent runs of the variability estimate, and the coefficient of variation (CV) of the species uncertainties. The results can also be plotted by selecting the Diagnostic Plots button, Variability Distribution option. The ratios of the variability estimate to the profile value are plotted for each species and source.

The variance of repeated variability estimates should be evaluated before selecting a final solution. This evaluation is especially important for small data sets (< 250 observations) and for data sets that have infrequent impacts by sources. The coefficient of variation (CV %) of uncertainties from the selected run and four additional runs is determined by selecting the Analyze Run button, Variability Distribution option. A CV less than 25% is preferred for each of the species in a source; however, some species that are at concentrations near their analytical method detection limit may have higher CV values. The results can also be plotted as a ratio of the variability distribution to the species concentration by selecting the Diagnostic Plots, Variability Distribution option.
These plots display the spread of sigma values obtained from 5 variability estimation runs. The individual variability values are normalized to the median variability so that the median value of the transformed values of the sigma values from the variability runs is always 1. The plots show the distribution of the normalized values with the red lines ranging from the normalized minimum to the normalized maximum values. The blue star denotes the normalized median values and is always present at 1. The data related to any species that shows a large spread should be analyzed thoroughly.

3.9 Run Profiles

Run profiles or selections can be saved using the Save Profile command as shown in Figure 18. The profiles are saved as .umx files that can be opened and edited in Microsoft Excel. The default folder for these files is C:\Program Files\EPA Unmix 6.0\Unmix Profiles. Using the data and selected species shown in Figure 11, select the File and Save Profile commands and save the file as wdcpmdata profile.umx.

US EPA ARCHIVE DOCUMENT

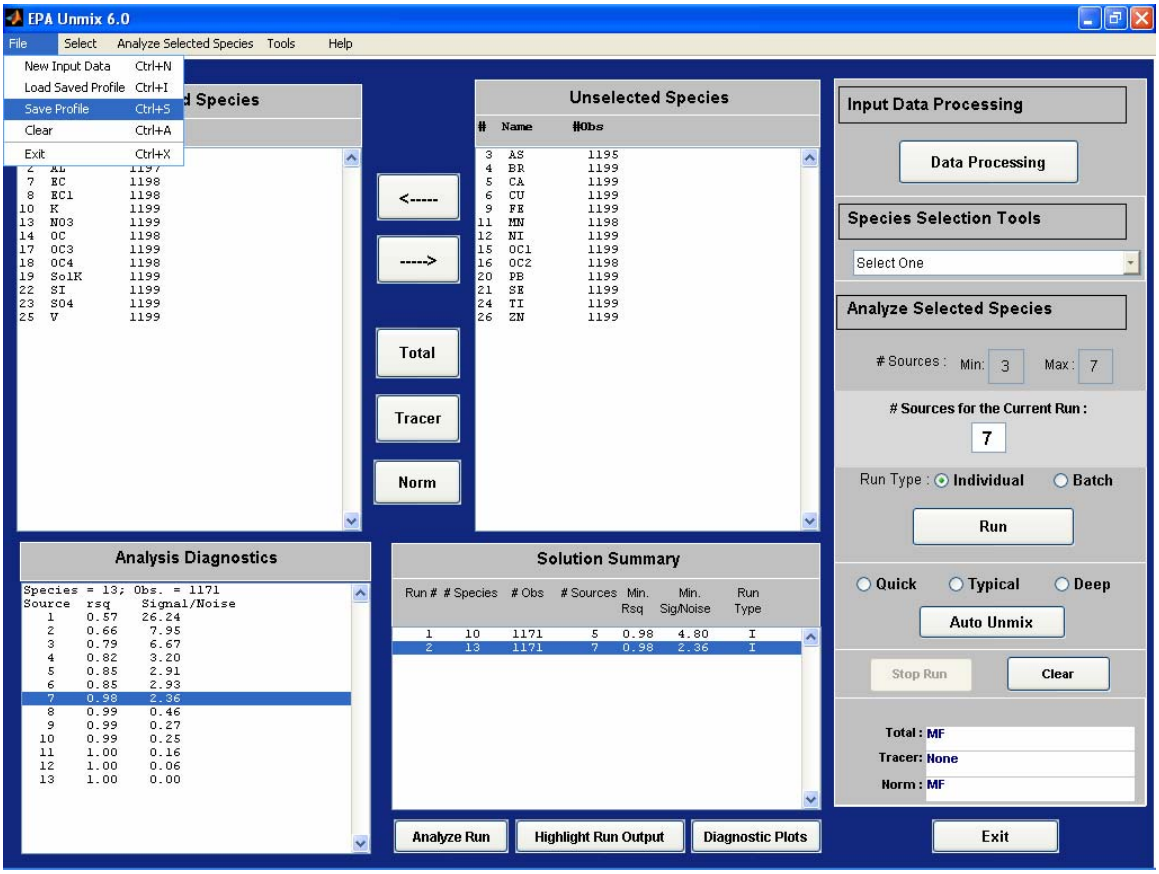
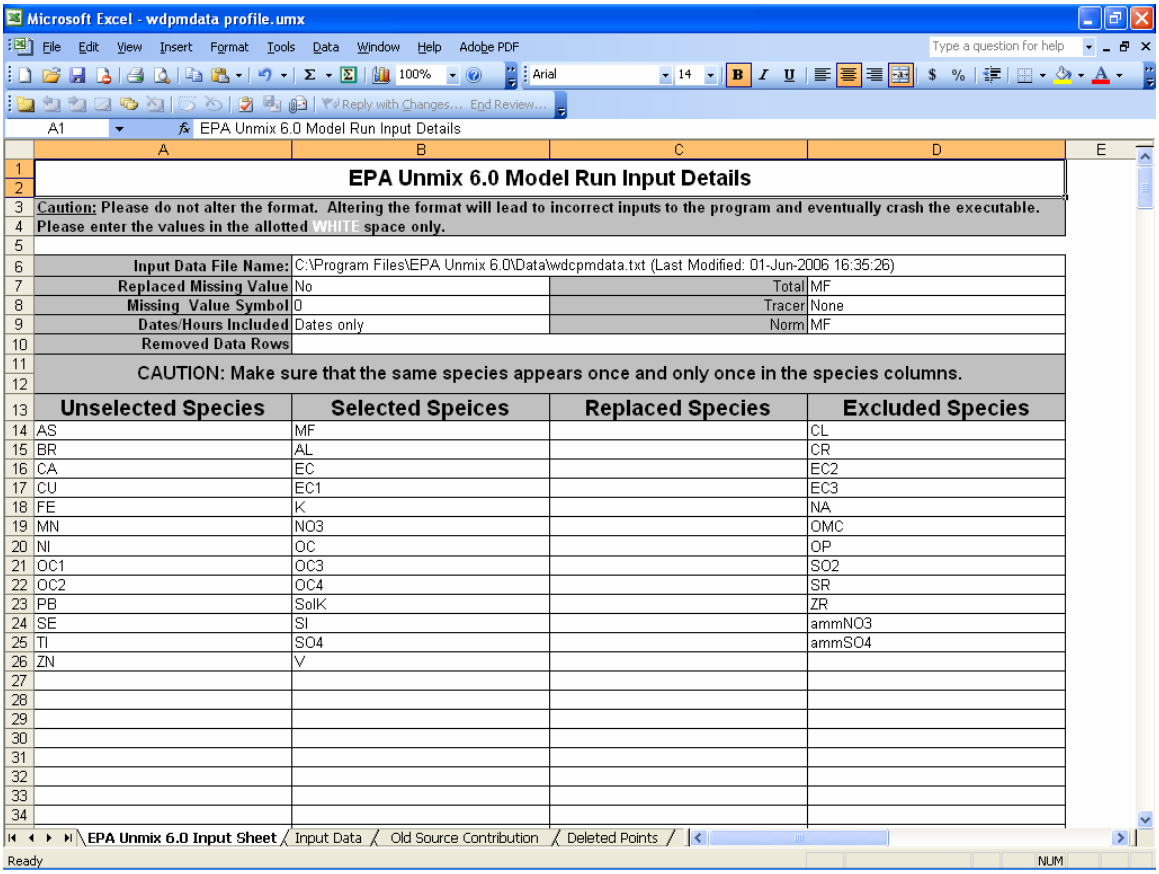


Figure 20: Save Current Run Choices

The umx file can be opened by starting Microsoft Excel, selecting the File and Open commands, changing the directory to the C:\Program Files\EPA Unmix 6.0\Unmix Profiles folder, and changing the file name extension to *.umx. After selecting the wdcpmdata profile.umx file, the following spreadsheet will open (Figure 19). To reload the run profile information into Unmix, select the File and Load Saved Profile commands from the main window in Unmix.

US EPA ARCHIVE DOCUMENT



Microsoft Excel - wdpmdata profile.umx

File Edit View Insert Format Tools Data Window Help Adobe PDF

A1 EPA Unmix 6.0 Model Run Input Details

EPA Unmix 6.0 Model Run Input Details

Caution: Please do not alter the format. Altering the format will lead to incorrect inputs to the program and eventually crash the executable.
Please enter the values in the allotted WHITE space only.

Input Data File Name:	C:\Program Files\EPA Unmix 6.0\Data\wdcpmdata.txt (Last Modified: 01-Jun-2006 16:35:26)		
Replaced Missing Value	No	Total	MF
Missing Value Symbol	0	Tracer	None
Dates/Hours Included	Dates only	Norm	MF
Removed Data Rows			

CAUTION: Make sure that the same species appears once and only once in the species columns.

Unselected Species	Selected Speices	Replaced Species	Excluded Species
AS	MF		CL
BR	AL		CR
CA	EC		EC2
CU	EC1		EC3
FE	K		NA
MN	NO3		OMC
NI	OC		OP
OC1	OC3		SO2
OC2	OC4		SR
PB	SoilK		ZR
SE	SI		ammNO3
TI	SO4		ammSO4
ZN	V		

EPA Unmix 6.0 Input Sheet | Input Data | Old Source Contribution | Deleted Points

Ready NUM

Figure 21: Umx File

SECTION 4. AUTO Unmix

Auto Unmix (AU) works by finding an initial model with the species selected using a varimax-rotated factor analysis of the data. The remaining species are added one at a time looking for new models with one more source than the initial or base model. Each new model found then becomes a base model, and the process of adding the remaining species one at a time is repeated. For each group of models with N sources, a model Figure of Merit (FOM) between 0 and 1 is calculated, with a FOM of 1 representing the best possible model. Eventually, the process ends when adding new species does not give new models with more sources. At this point, the uncertainties in the source compositions are calculated and models with too much variability (as defined below) are eliminated. Finally, a new modified FOM is calculated, and the models are stored in order of decreasing FOM.

The FOM calculation depends on whether a TS such as $PM_{2.5}$ has been selected. It is strongly recommended that TS be selected in the data since it makes ranking the models more reliable. The amount of TS explained by each source is as an important part of the FOM calculation, and if no TS is selected,

US EPA ARCHIVE DOCUMENT

then AU uses a similar method to calculate the FOM that leaves out the parts that require TS.

Assume that one or more base models with N-1 sources have been found. Further assume that adding remaining species one at a time to the base models results in K new models with N sources. The FOM is calculated as a weighted sum of five parameters:
r_1 = signal-to-noise ratio of the Nth principal component,
r_2 = difference between the signal-to-noise ratio of principal components N and N+1,
r_3 = percent of average TS associated with the new source,
r_4 = minimum percent of average TS of any source in the model, and
r_5 = percent TS associated with the large new source contributions.

The signal-to-noise ratio of each component in the singular value decomposition of the data is calculated by the NUMFACT algorithm (Henry, 1999). All else being equal, the larger the value of r_1, the signal-to-noise ratio of the Nth principal component, the better the model. Also, it is better to have a large difference between the signal-to-noise ratio of components N and N+1 since this implies that the information in the data (the signal) is concentrated in the first N principal components. So larger values of r_1 and r_2 are better, and the FOM should reflect this.

The next three parts of the FOM are only defined if the data contains TS. For r_3, the percent of average TS associated with the new source, it is obviously better that the new source added should explain as much of the total as possible. It is also better if r_4, the minimum percent of average TS of any source in the model, be as large as possible since this tends to preclude having one or more sources that explain very little mass. Finally, it is better when the newly added source contribution is large (greater than 3 sigma) and that the percentage of TS explained by the source is also large.

The FOM is calculated so that models that maximize the values of r_1 to r_5 have the largest FOM. Thus, if there are K models with N sources and r_{ik} is the value of r_i for the kth model, the evaluation number for the kth source FOM_k is calculated as

$$EN_k = \frac{1}{\sum_{j=1}^{5} w_j} \sum_{i=1}^{5} \frac{w_i r_{ik}}{\max_k (r_{ik})},$$

where the weights w_i, i = 1,…,5 are as follows:
w = (1 1 2 2 2) if TS explained by the new source is less than 5 percent and
w = (1 1 2 1 1) if the TS explained by the new sources is greater than 10 percent.
In between, the weights are linear. The terms in the denominator in the equation for FOM are normalization factors that ensure that FOM is between 0 and 1. Models are sorted by the FOM and the bigger the FOM, the better the model. If

US EPA ARCHIVE DOCUMENT

there is no TS in the data, then the FOM is calculated just as above, but the weights of parts that depend on TS are set to 0 (*w* = (1 1 0 0 0) for all models).

The process of adding species one at a time to obtain models with ever-greater number of sources continues until no new models are found. Uncertainties for the source compositions of the final set of models are then calculated and models with too much error are eliminated. If the data has TS, then models are eliminated that have at least one source with TS less than 2 times the estimated error. If there is no TS, models are eliminated if for at least one source the source compositions are all less than twice the estimated error. Finally, a new FOM is calculated for the remaining models. In this final calculation, the same formula is used as given above, but r_3 is replaced by the min r^2 for the model and the weights are (1 1 2 1 2).

The parameters MaxParents, TSmin, SNMin, and MaxNumSolutions can be changed. If a TS was selected, then TSmin is the minimum average percentage of TS due to any source in a solution. The purpose of this is to prevent solutions with sources that have very small contributions to the TS. As AU searches for solutions, it uses each existing solution as a base from which to look for solutions with one additional source. The number of existing solutions that will be followed at each level is limited to MaxParents. If it is set to 5, for example, AU will only follow the best 5 solutions at each level. The larger the value of MaxParents, the more solutions Auto Unmix will find, but the run time will be increased. MaxNumSolutions is the maximum number of solutions that Unmix will report. SNMin is a highly technical parameter that controls the minimum signal-to-noise ratio allowed in the solution. AU has three levels: Quick, Typical, and Deep. The specific settings for each of the options are shown below.

Quick: MaxParents = 2, TSmin = 2, SNMin = 1.5, MaxNumSolutions = 2, Unselected species r^2 min = 0.5, feasible variability runs = 50, maximum number of variability runs = 100;

Typical: MaxParents = 5, TSMin = 2, SNMin = 1.5, MaxNumSolutions = 5, Unselected species r^2 min = 0.3, feasible variability runs = 100, maximum number of variability runs = 200;

Deep: MaxParents = 10, TSMin = 1, SNMin = 1.5, MaxNumSolutions = 10, Unselected species r^2 min = 0.3, feasible variability runs = 100, maximum number of variability runs = 200.

This procedure can take from a few minutes to over an hour depending on the AU settings and the number of unexcluded species, and automates the typical steps that are used to develop the best possible Unmix solution based on the data. An example of the AU algorithm is shown below using the wdcpmdata and the Typical AU settings. Use the initial selected wdcmpdata species shown in Figure 6 and select MF as the Total and Normalization (Norm) variable. On the

US EPA ARCHIVE DOCUMENT

right side of the Main window, select the Typical option for AU and select the AU button as shown in Figure 20. Select “consider all species” to evaluate all of the species in the Unselected Species box.

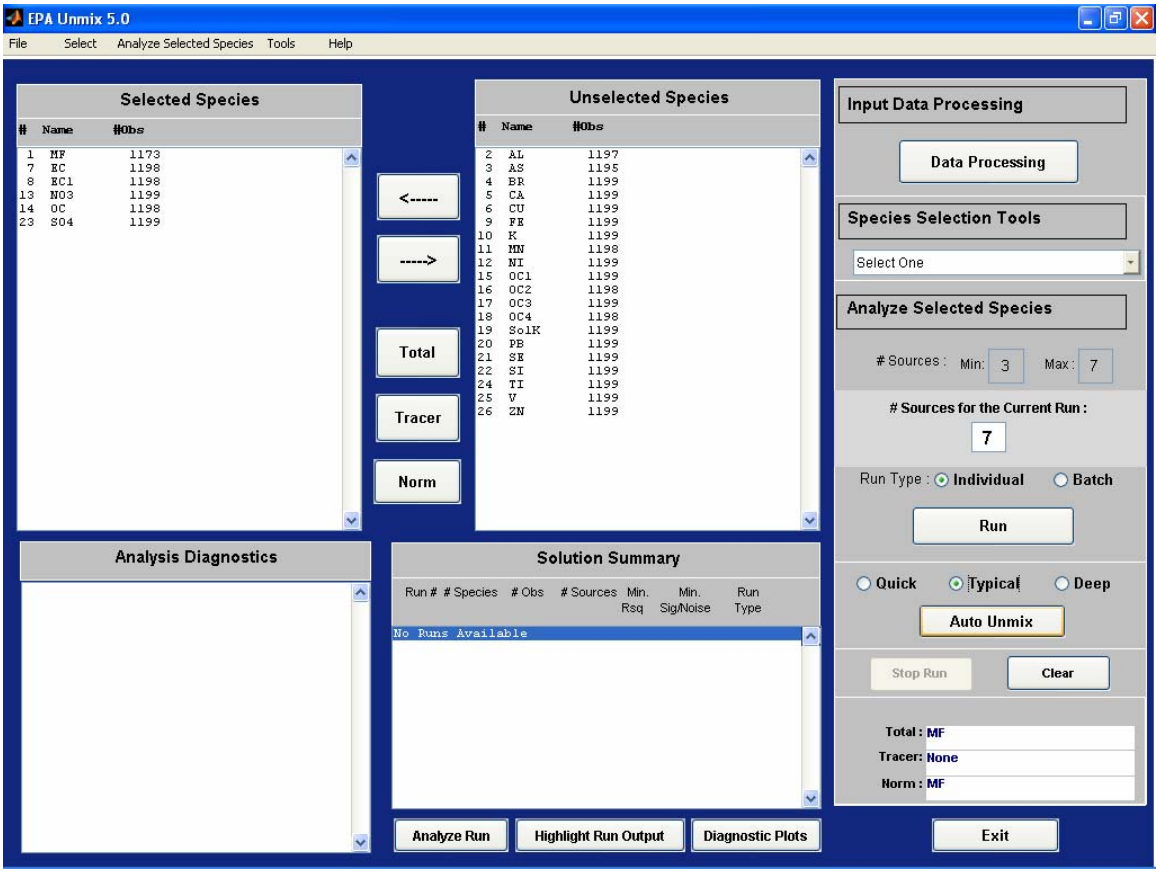


Figure 22: Auto Unmix command

A window will open that shows the AU progress and the species that are being evaluated (# shown next to species in Selected and Unselected Species boxes). The analysis results from AU are displayed in Figure 21 and plotted in Figure 22 by selecting the Diagnostic Plots button, Variability Estimate (log) option, Figure Groups button, and Number of Plots per page set to All. Two ten source solutions were found with FOM values of 0.94 and 0.96. The FOM value is a relative ranking and may not match these values due to the random number generator used in the model. In addition, the number of solutions may not be the same due to the randomized sequence of species selection. However, the best solution is always displayed.

US EPA ARCHIVE DOCUMENT

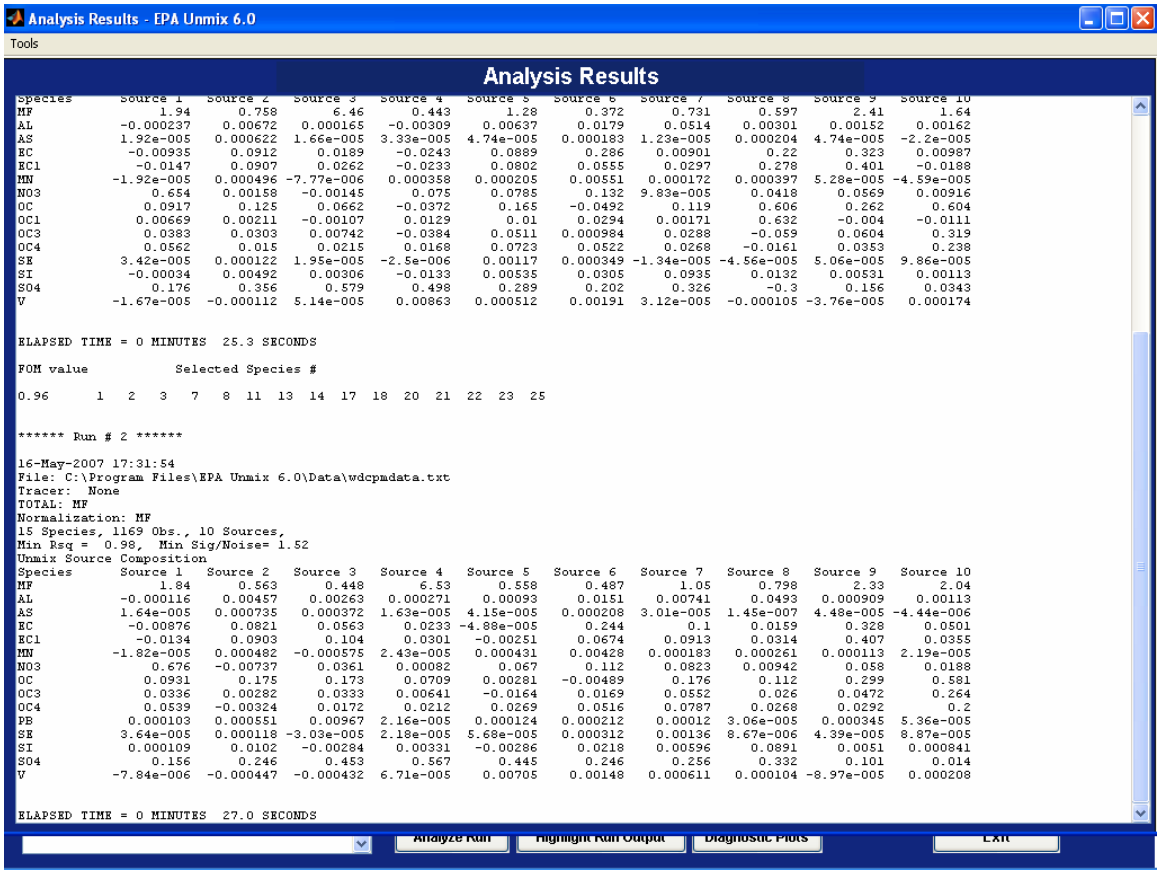


Figure 23: AU result for wdcpmdata

US EPA ARCHIVE DOCUMENT

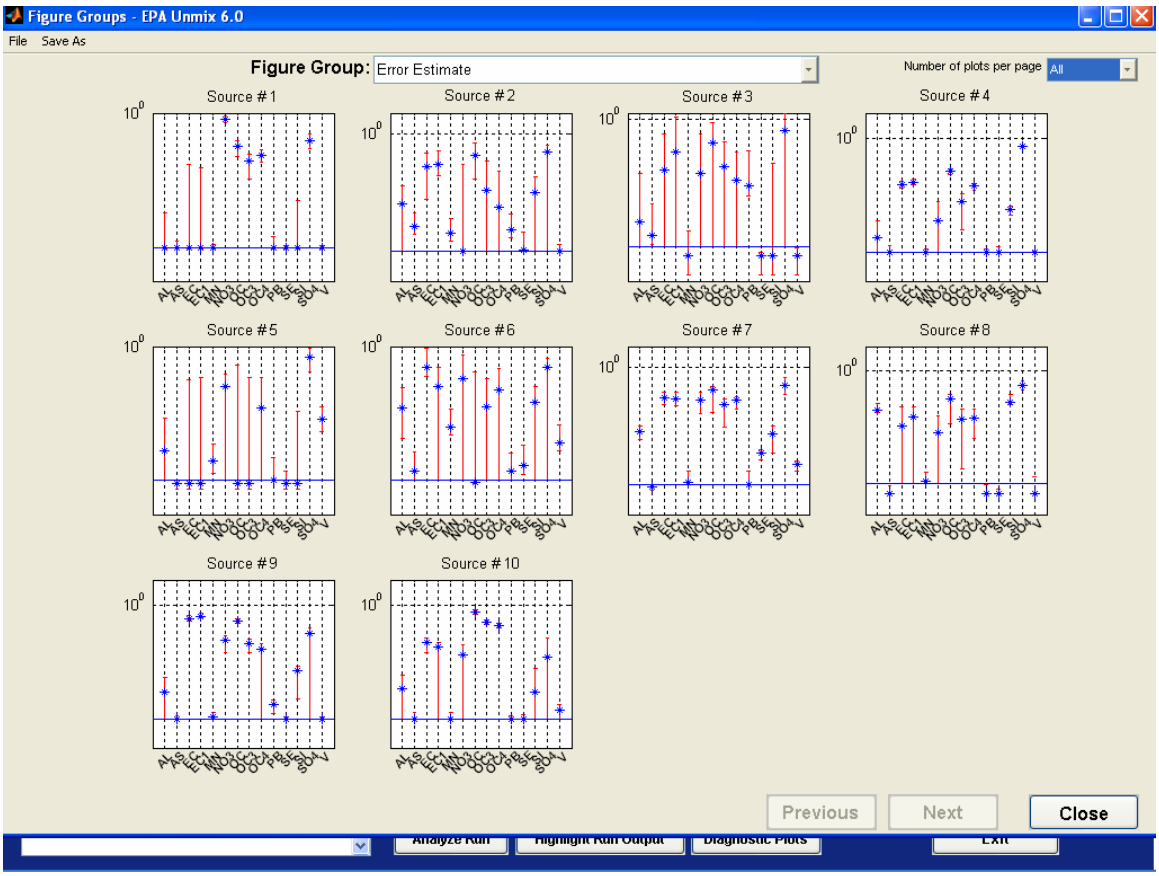


Figure 24: AU Source Profiles

The species from the highest FOM are loaded into the Selected Species box. Evaluate the AU solution in more detail by selecting the Analyze Run and Diagnostic Plots buttons. The Fit Additional Species command can be used to find species (i.e. species with R^2 greater than 0.80) that can be added to the AU solution. Evaluate any final solutions with the Diagnostic Plots ▶ Fit Diagnostics and Analyze Run ▶ Variability Distribution commands.

A warning message that "No additional species found" will be displayed if AU cannot find any additional solutions. Try removing some of the species in the Selected Species box or do not use a TS. Another option is to try running AU with the Deep setting.

SECTION 5. ADVANCED OPERATIONS

The basic strength of Unmix, as with all receptor models, is that it relies on the data; the basic weakness of Unmix is that it relies on the data. A number of problems may afflict a species and make it unsuitable for selection to be part of the model. A common problem is that the species may have missing

US EPA ARCHIVE DOCUMENT

concentration data, however, the Replace Missing Values (see Section 5.5) command in Unmix can be used to replace missing data. Another common problem is the existence of outliers in the values of the species. These can often be detected using scatterplots as described in the section on View/Edit Observations & Points, or by using the Influential Points command (see Section 5.2). Sometimes a species may have a lot of noise associated with it. Measurement error is one source of noise, especially when the species is just above the minimum detectable limit. Finally, a species may not be suitable because it violates the assumption inherent in all receptor models that the source compositions are approximately constant. If the mass fraction of a species varies enough, it will destroy the constraints in the data that Unmix uses to obtain a solution. Lewis *et al.* (1998) discusses some possible data problems and how to identify these.

5.1 Influential Observations

The tools for evaluating influential observations and points are listed in the lower left corner of the Data Processing window under View/Edit Points and Observations. Select the View/Edit Points and Observations list on the Data Processing window, after removing the wdcpmdata species that were recommended for exclusion in Figure 5. Select the View/Edit Observations & Points command and consider all species as seen in Figure 23. Select MF as the base species.

US EPA ARCHIVE DOCUMENT

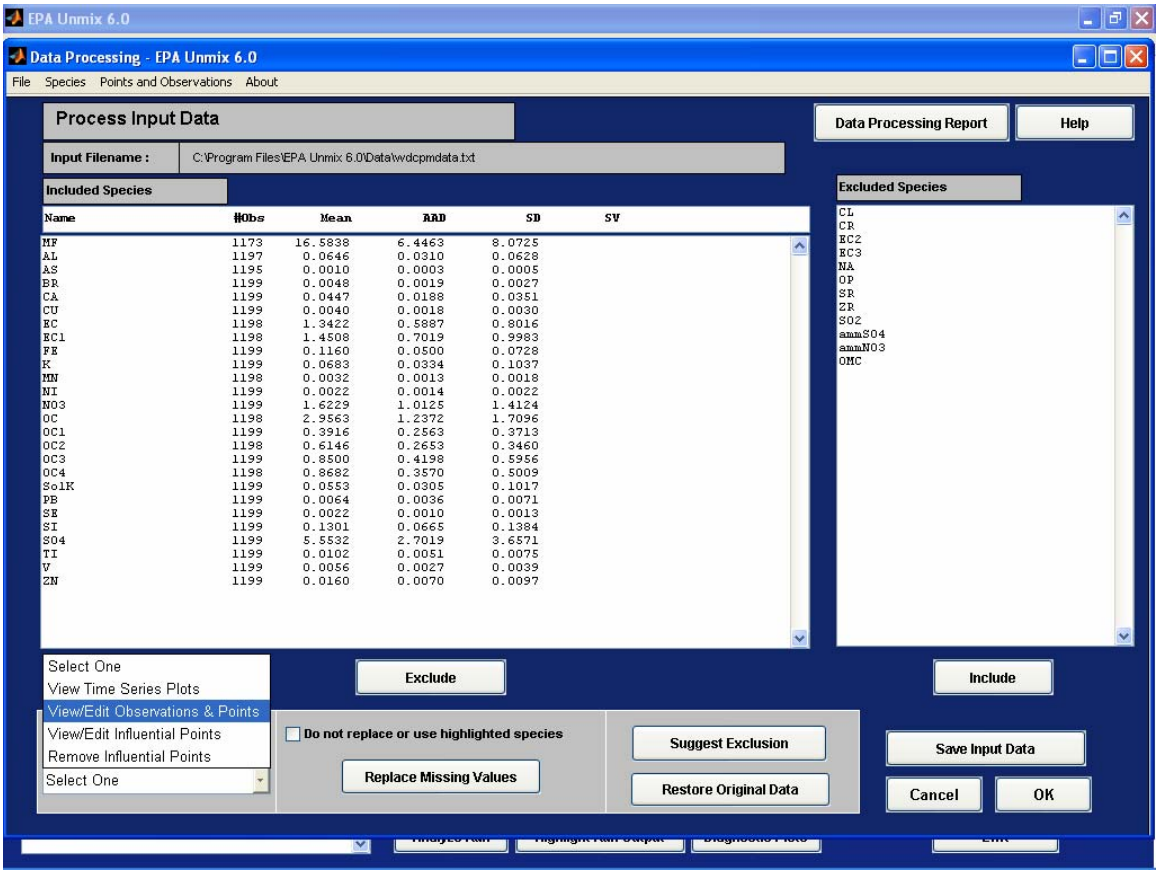


Figure 25: View/Edit Observations & Points

All of the species concentrations are plotted against the selected base species and the edges are displayed with dashed red lines (Figure 24). The size of the figures can be increased by reducing the number of species or by selecting a figure and the Undock button. Observations or data points contributing to the poor upper edges can be deleted in the Species vs. Base Figure by selecting the point in the figure with the left mouse button. For example, go to the OC1 figure (4th column, 3rd row) and choose the high OC1 point by placing the tip of the arrow near the point, holding down the left mouse key and increasing the size of the selection box until it contains the point. Release the left mouse key and the point will be identified with a red square and the other species for that sample will be circled in the other plots. Multiple points can also be selected in a plot by choosing a location near one of the points in the plot (not outside of the plot), holding down the left mouse key, increasing the size of the box until the points are within the box, and releasing the mouse key.

US EPA ARCHIVE DOCUMENT

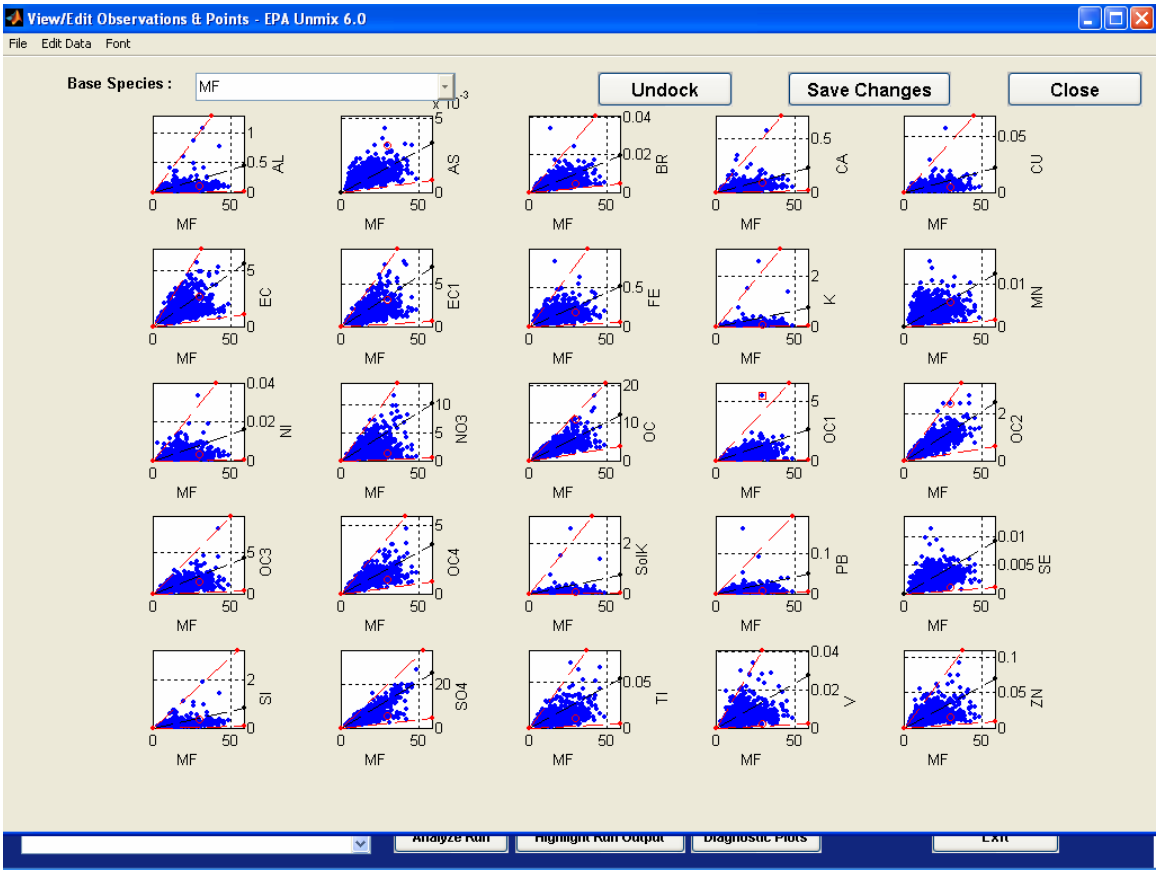


Figure 26: View/Edit Observations & Points plot high OC1 point

Select the Edit Data ▶ Delete Selected Observation(s) command from the top left of the View/Edit Observations & Points window. After all of the species related to the deleted observation are removed, new edges are drawn (as displayed in Figure 25) and the observation number of the deleted point is recorded in the Data Processing Report. If a figure is too small to select an individual point, reduce the number of highlighted species before selecting the View/Edit Observations & Points command. Another option if the figure is small is to select a figure and then select the Undock button. This will create a window with only the selected figure. Undocked figures cannot be used for selecting observations. The data points can be restored in the Edit Data command using the Restore Most Recently Added or Restore All options.

US EPA ARCHIVE DOCUMENT

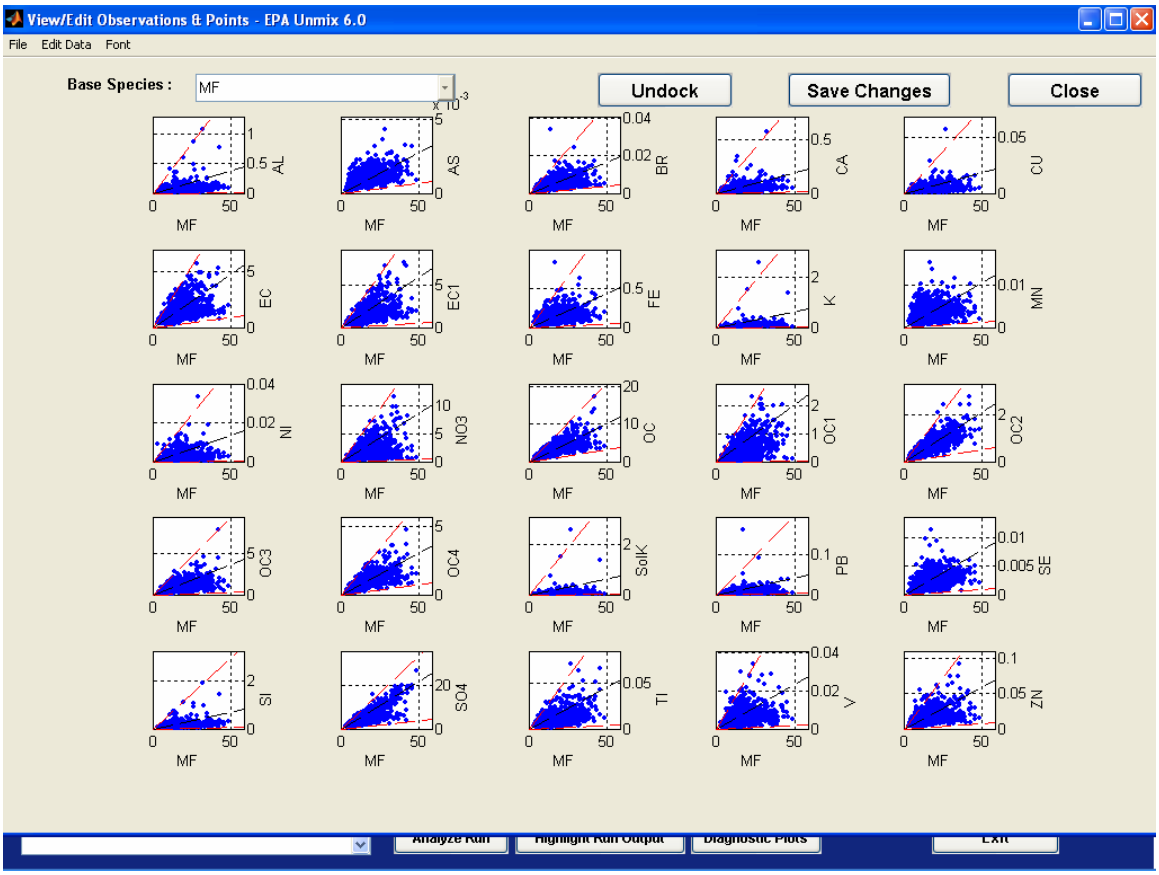


Figure 27: New edges in View/Edit Observations & Points plot

Select the Save Changes and Close buttons on the View/Edit Observations & Points window. Select the Data Processing Report button in the upper right corner of the Data Processing window to view the report, or export the report from the Analyze and Export Run Results window, Data Processing Report option. An example report is shown in the following Figure 26. The sample collected on 04/08/1992 was deleted and the OC1 value was 5.46.

US EPA ARCHIVE DOCUMENT

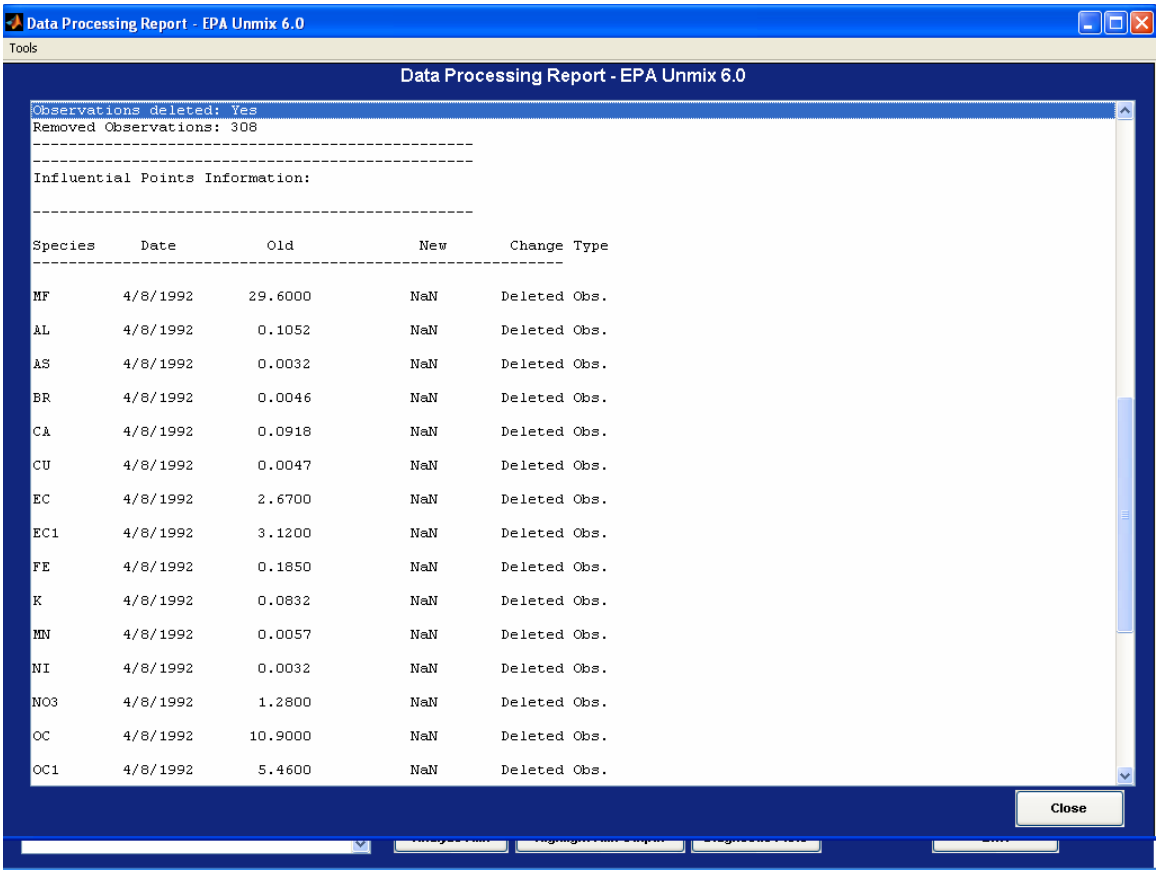


Figure 28: Data Processing Report

A point can be identified in the View/Edit Observations & Points plot by selecting the Edit Data ▶ Datacursor Mode command. For example, select the high K value in the K vs. MF plot. Figure 27 shows that the high K value was on 07/05/2000 and was most likely impacted by 4th of July fireworks. This individual point can be removed using the Edit Data ▶ Delete Selected Point(s) command. The selected non-base species point in the observation is replaced with a missing value symbol and the other species concentrations in the observation remain unchanged. The Replace Missing Value command can be used to replace the deleted point with a value that is consistent with the other species concentrations in the observation.

US EPA ARCHIVE DOCUMENT

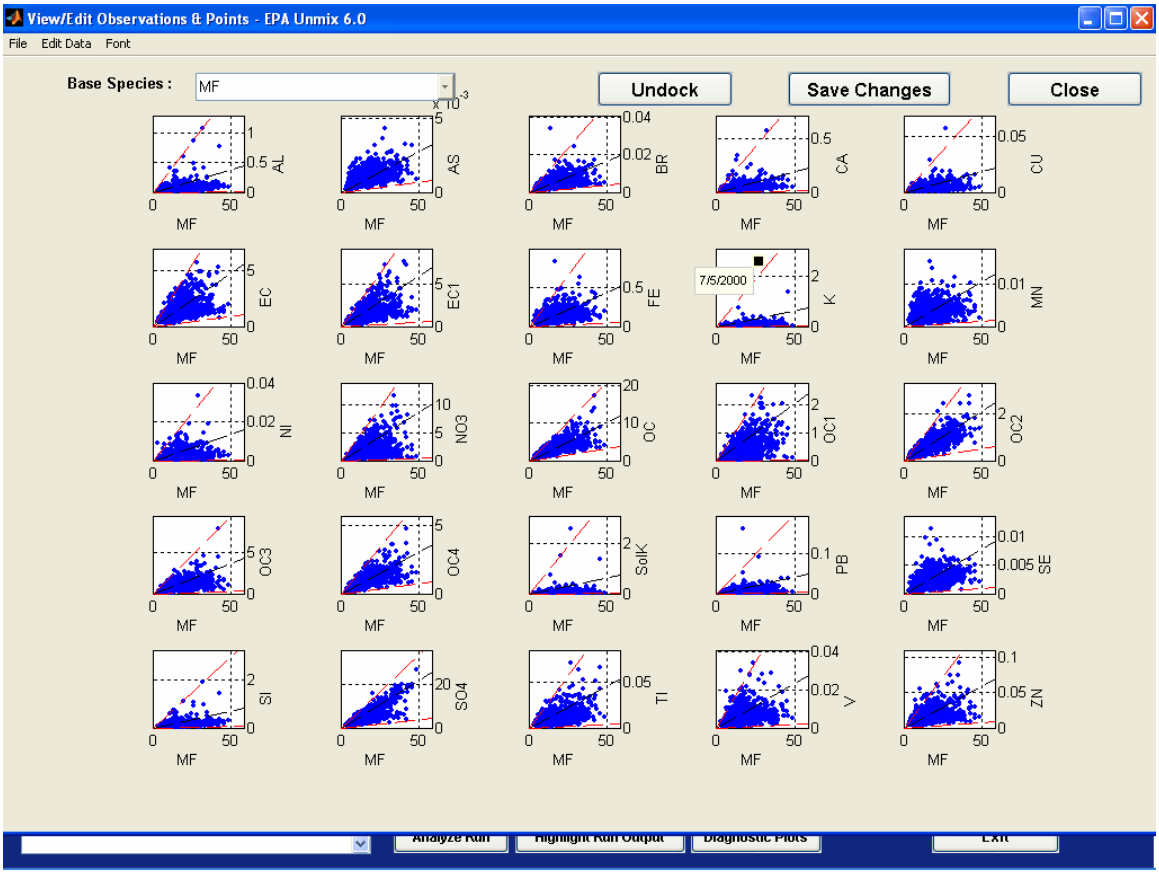


Figure 29: Datacursor Mode

The Datacursor Mode command is turned off by re-selecting the Edit Data ▶ Datacursor Mode command. After the Datacursor mode is turned off, the figures are re-drawn resulting in a delay in the window commands being available. Save the modified data set by selecting Save Changes. It should be noted that Unmix saves the deleted point information in the Run Profile (see Section 3.9).

5.2 Influential Points

The View/Edit Influential Points command identifies those points that influence the definition of an edge significantly. A point is considered to be highly influential, if after removal of the point, the edge moves significantly. Influential points can significantly affect the nature of the solution produced by the SAFER algorithm for a chosen group of species. That is, the SAFER algorithm that failed to produce a feasible solution for a chosen set of species can produce a feasible solution for the same set of species after deleting just one observation deemed highly influential from one of the chosen species. The View/Edit Influential Points command can be used to remove a single point which then can be replaced using the Replace Missing Value command.

US EPA ARCHIVE DOCUMENT

Influential points are identified using three parameters: P, *alpha,* and K. P is the fraction of the total number of points near the edge that are used in the statistical modeling of the edge. The value for P is 0.25, and if there are N points, the algorithm only looks at the 25 percent closest to the edge; there are n1 = P*N of these. An adjustable parameter *alpha* is related to determining influential points. Before discussing *alpha*, the working of the algorithm must be explained. The following discusses in detail how the influential points algorithm works.

The axes of the data are rotated to align with the edge. Thus, each data point's perpendicular distance from the edge becomes its y-coordinate (dy1) and the x-coordinate is distance along the edge (dx1). The algorithm looks for points with unusually large negative values of dy1 (points below the edge) and points with unusually large distances along the edge as given by dx1.

Since most air quality data are approximately lognormally distributed, dx1 is assumed to be lognormally distributed. Next, the values of dx1 are converted to a standard normal distribution by taking logs and subtracting the mean and dividing by the standard deviation. The parameter *alpha* takes values from 0 to 1: *alpha* = 1 corresponds to allowing 1 outlier in the sample of n1 points near the edge, and *alpha* = 0.5 corresponds to allowing 0.5 outliers (on average). The smaller the *alpha*, the fewer points are flagged as possible influential points. Thus, the alpha or longitudinal spread ranges from 0 to 0.35 with a default value of 0.05, with the larger values flaging more points as influential.

Most points are flagged as influential because they lie too far below the edge (dy1 values are large and negative). Because of random noise in the data, some of the points will be below the edge. Assume a uniform distribution of points near the edge with simple Gaussian noise, it can be shown that the squared distance of the point from the edge divided by the sum of the squared distances from the edge should be less than a multiple of 3/n1. This multiple is defined as K or the transverse spread which ranges from 2 to 15, with a default of 10. In this case, the smaller values of K would flag more points as influential.

In addition to identifying influential points these plots also display the "Max. Common Source Contribution." One can get some additional information from the lower edge in a plot of a species (V) versus a species (T). Assume V has more than one source and that the source with the largest fraction of V is source S. The inverse of the slope of the edge is an estimate of the fraction of V in source S. Furthermore, the slope times the mean of V divided by the mean of T is an estimate of the maximum fraction of average T contributed by source S. This information can be useful in helping to decide if the influential points that have been identified by the algorithm should be eliminated or not. Also, this kind of information can be helpful in deciding if a small source associated with a single species found by Unmix is physically reasonable.

US EPA ARCHIVE DOCUMENT

The edge resolution number is useful in characterizing the edge definition. The value 0 suggests an unresolved edge and 1 a perfectly resolved edge. The intermediate values accurately indicates how "good" the edge is. The edge resolution number displayed pertains to the edge defined by the red dashed line (prior to removing the influential points). This is likely to improve when the suggested influential points are removed.

Load the STN_PM.xls data set from the C:\Program Files\EPA Unmix 6.0\Data folder. This is an example particulate matter data set from a Speciation Trends Network (STN) site. The data file contains dates in the first column and the missing value symbol is -99 (Figure 28).

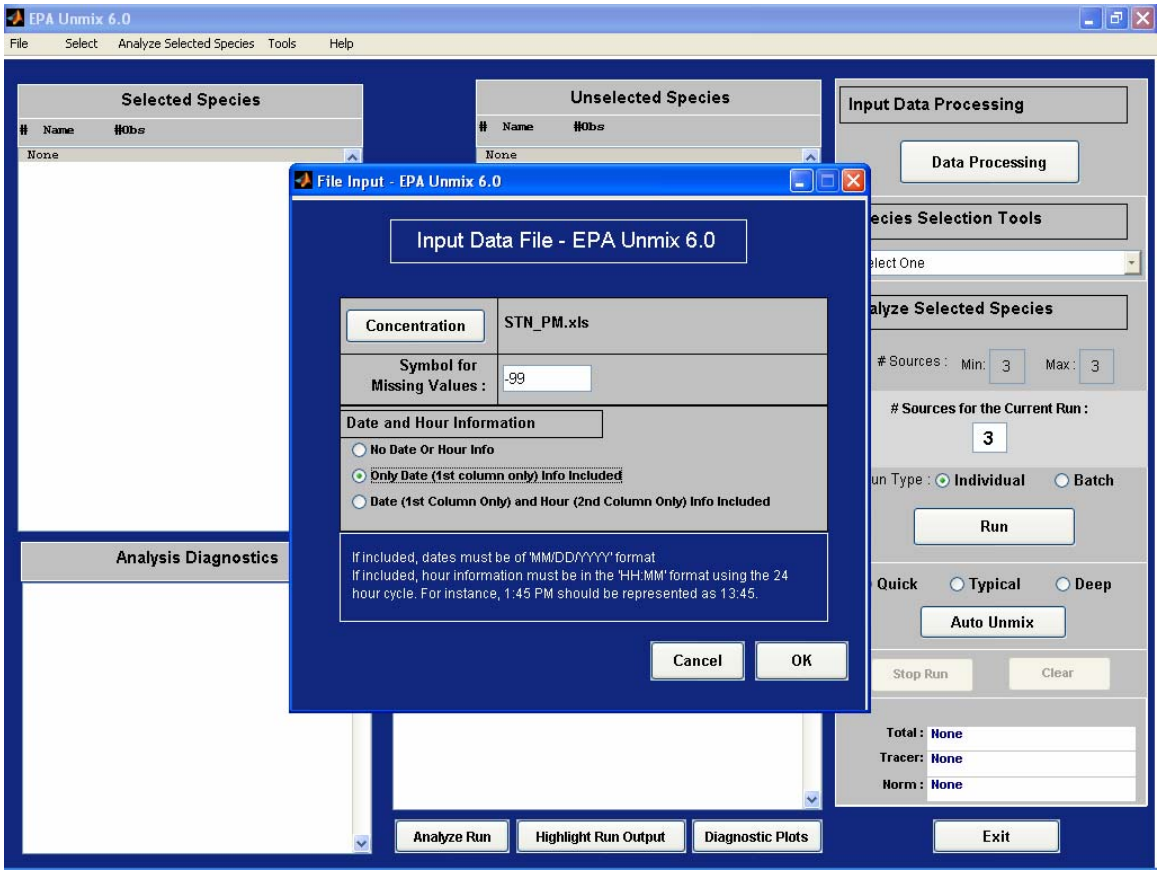


Figure 30: Example STN PM data set

Select the Suggest Exclusion button ▶ and then exclude the suggested species by selecting the Exclude button. A large number of the trace elemental species are recommended for exclusion (Figure 29).

US EPA ARCHIVE DOCUMENT

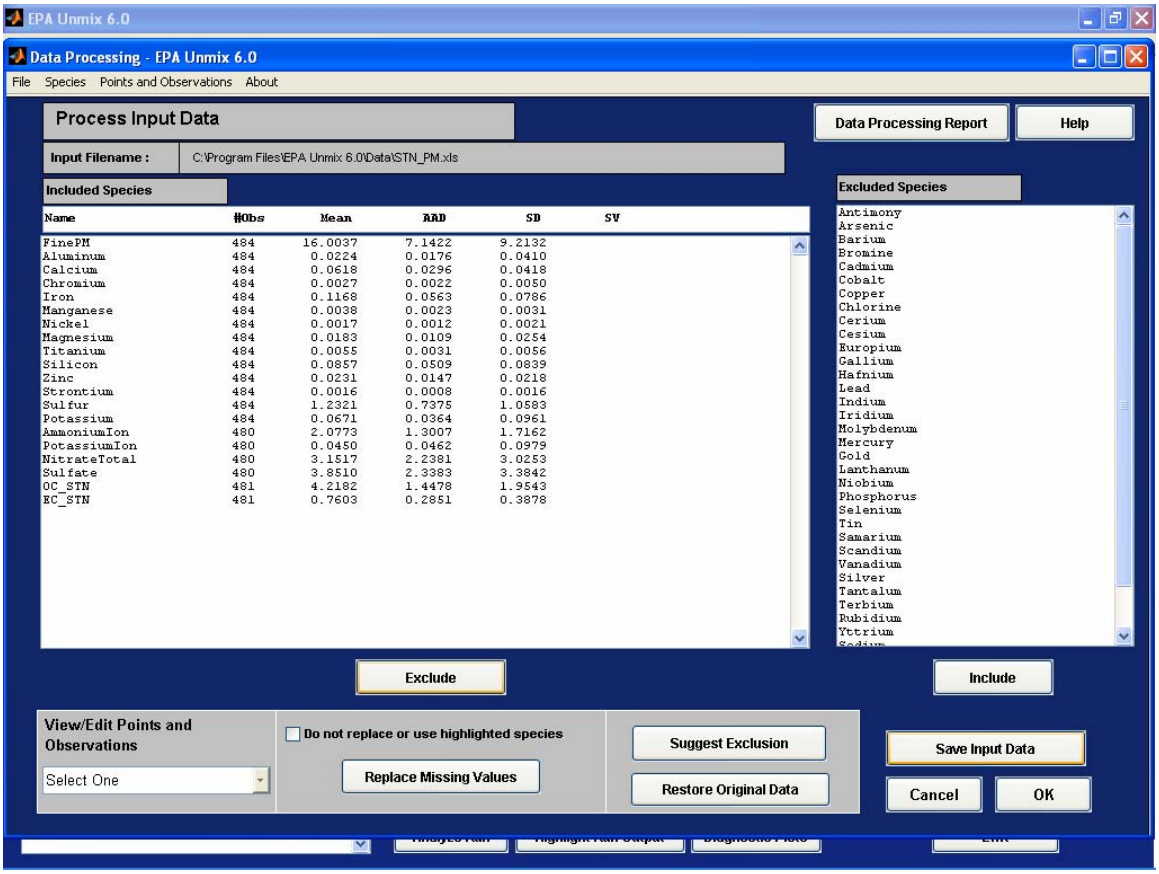


Figure 31: Example STN PM excluded species

Select the View/Edit Points and Observations list at the bottom left corner of the window and select the View/Edit Influential Points command. Use the default longitudinal and transverse spread parameters (Figure 30). All of the influential points can also be deleted without viewing the plots by selecting the Remove Influential Points option in the View/Edit Points and Observations list.

US EPA ARCHIVE DOCUMENT

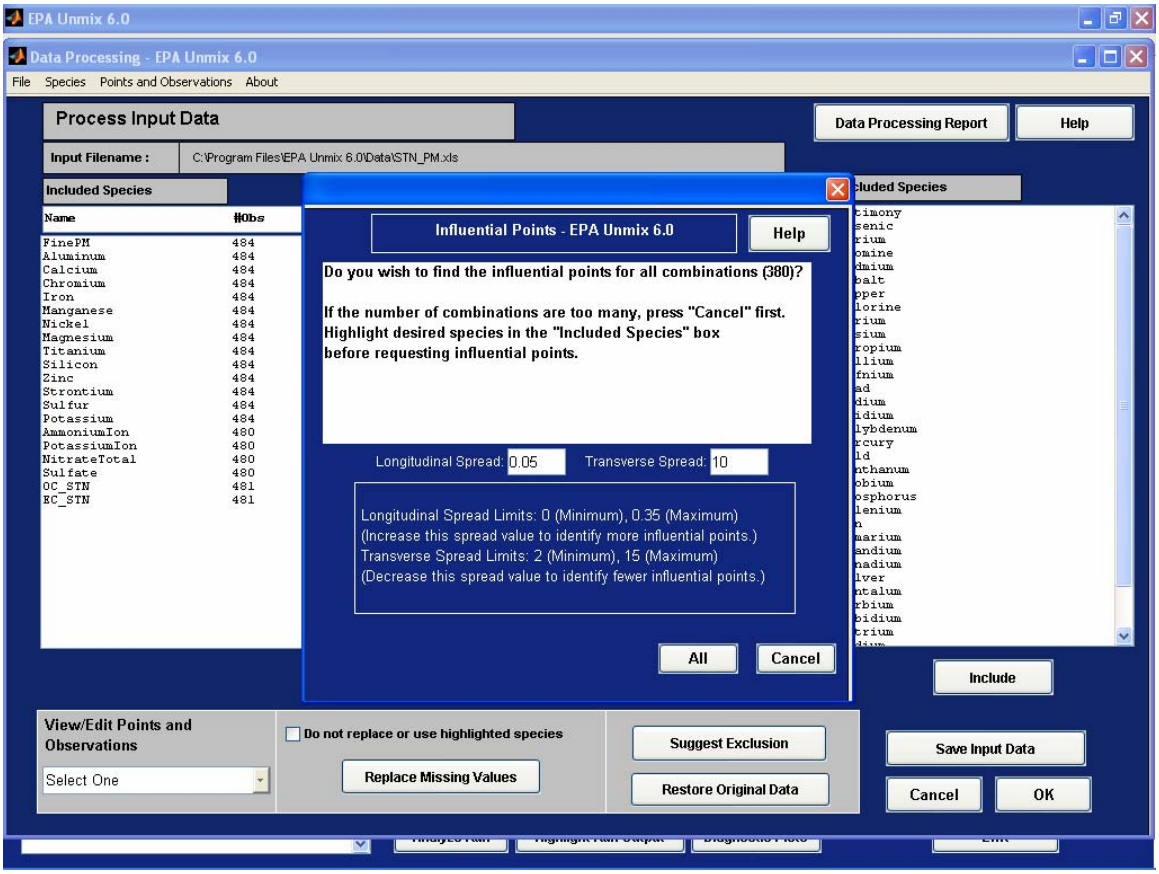


Figure 32: Influential Points command

Select the All button to find the influential points for all combinations of the selected species. In the scatter plot (Figure 31), the red circled points are the influential points. The red dashed line is the edge defined by the data values when none of the influential points are removed. The black solid line is the mean edge after removing all of the influential points and the black dashed lines below and above the black solid lines are the lower and upper estimates of the edges after the influential points have been removed. Use the Next and Previous buttons to view the species combinations with influential points. The influential point information is also listed in the Data Processing Report. Use the Next button to select the 3rd figure, the Potassium vs. Zinc plot.

US EPA ARCHIVE DOCUMENT

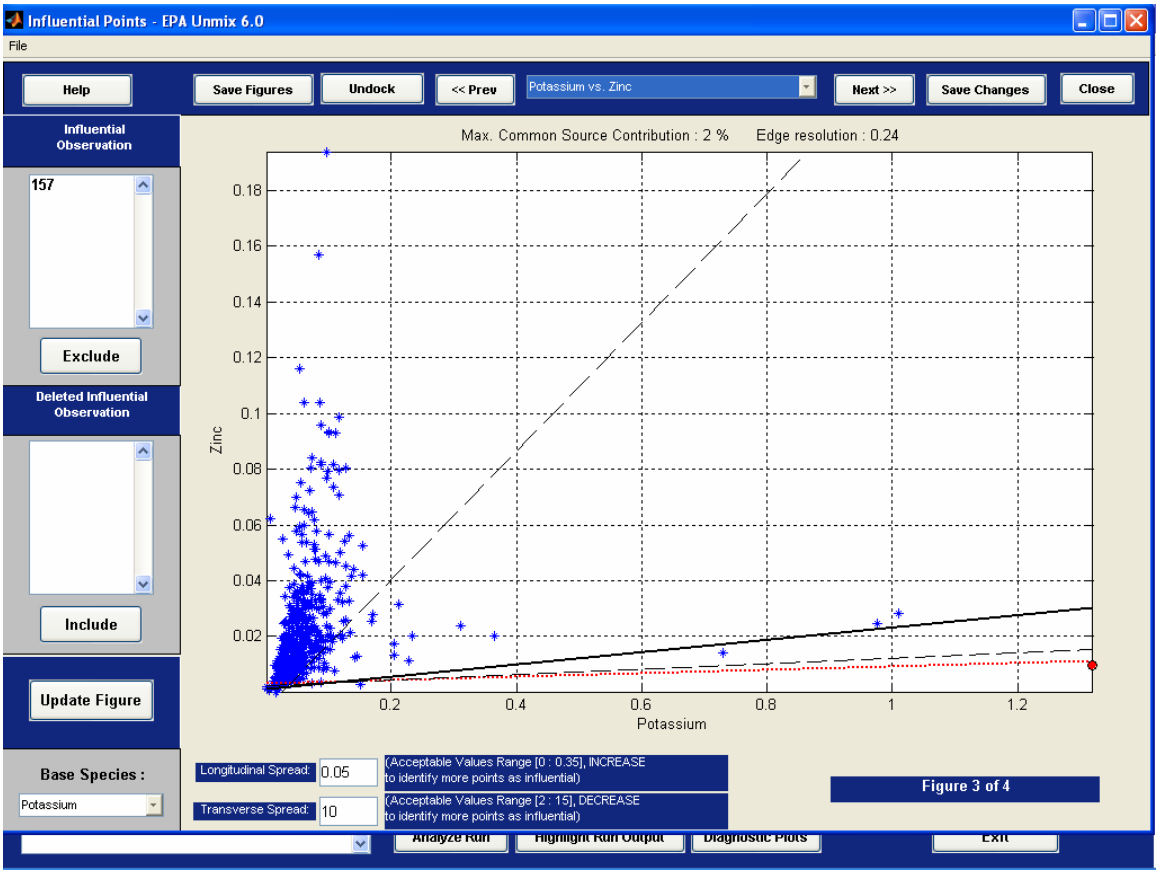


Figure 33: Influential potassium point

Highlight observation 157 in the Influential Observation list and the date of the sample and longitudinal distance will be displayed on the plot (Figure 32).

US EPA ARCHIVE DOCUMENT

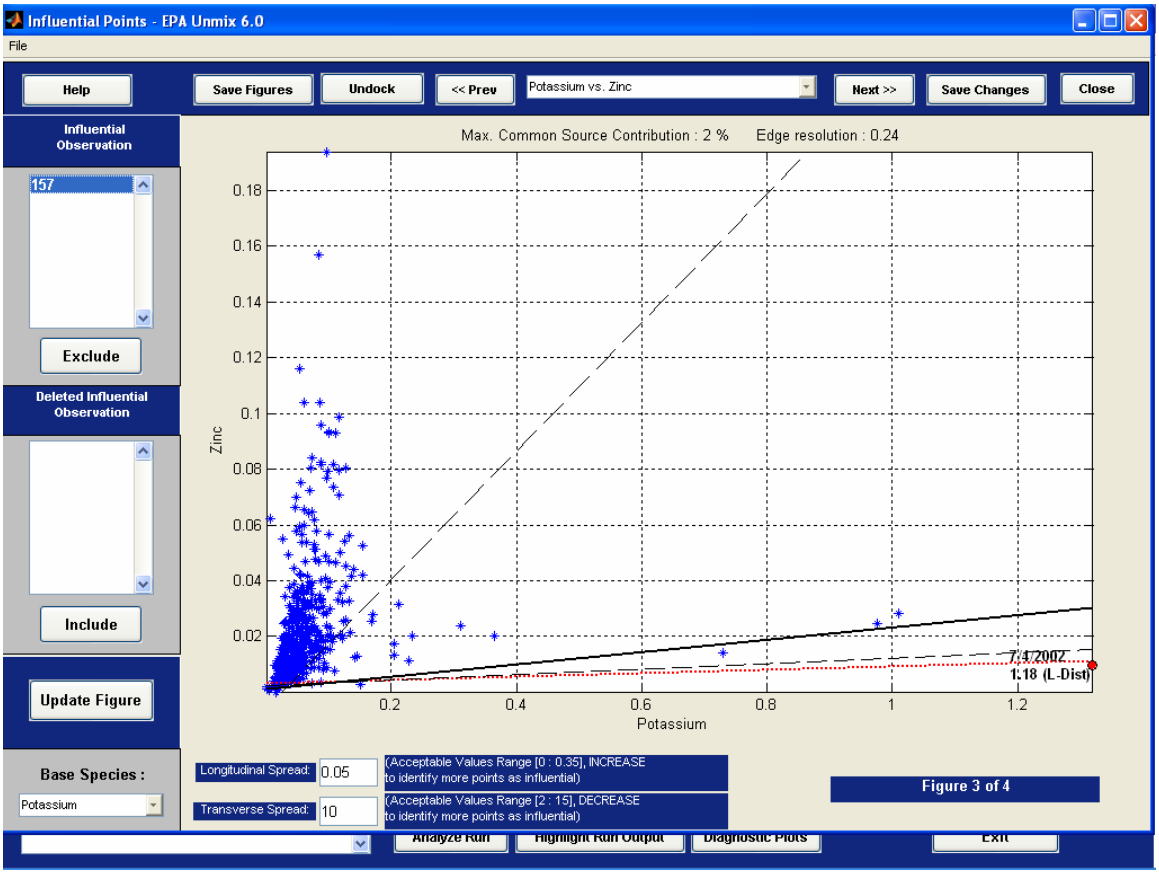


Figure 34: Influential points

The high potassium concentration is from a sample collected on the 4th of July (fireworks). Select the Exclude button to move the observation number to the Deleted Influential Observation list. Select the Update Figure button to view the figure without the influential point. The edge resolution improves after removing the point by increasing from 0.24 to 0.84.

The spread parameters in the Influential Points window can be changed by changing the longitudinal spread or transverse spread parameters and selecting the Update Figure button. Increasing the longitudinal spread to 0.08 captured three additional high Potassium points. The high Potassium points can be deleted using the updated list of influential points or View/Edit Observations & Points command, Edit Data ▶ Delete Selected Point(s) command. Use the Replace Missing Values command to replace the deleted values.

5.3 Apportionment of Species Not in the Model

Once Unmix has found a solution, the user may be interested in how well the rest of the species in the data are fit by the sources in the solution. The unselected species are evaluated by selecting the Species Selection Tools, Fit Unselected

US EPA ARCHIVE DOCUMENT

Species command (Figure 33). Figure 34 shows the results for the wdcpmdata species (Refer to Figure 11).

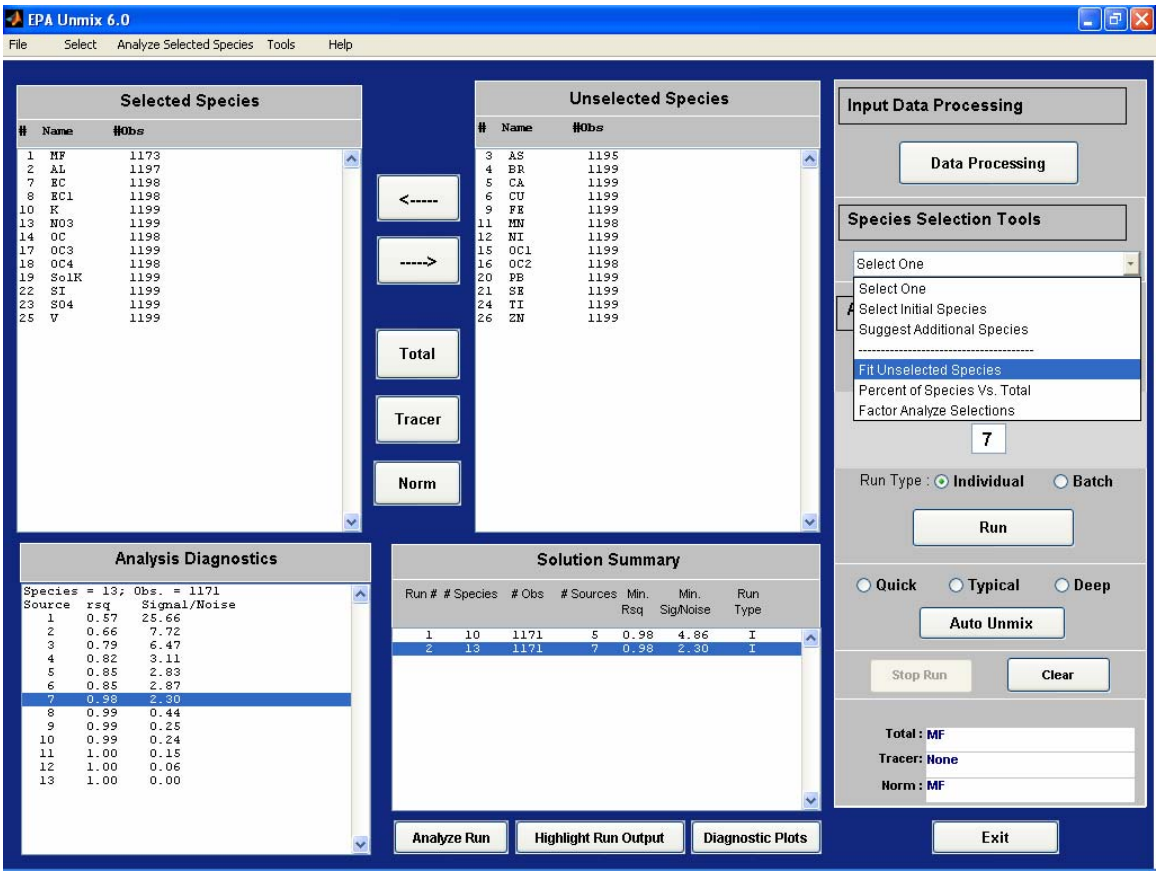


Figure 35: Fit Unselected Species command

US EPA ARCHIVE DOCUMENT

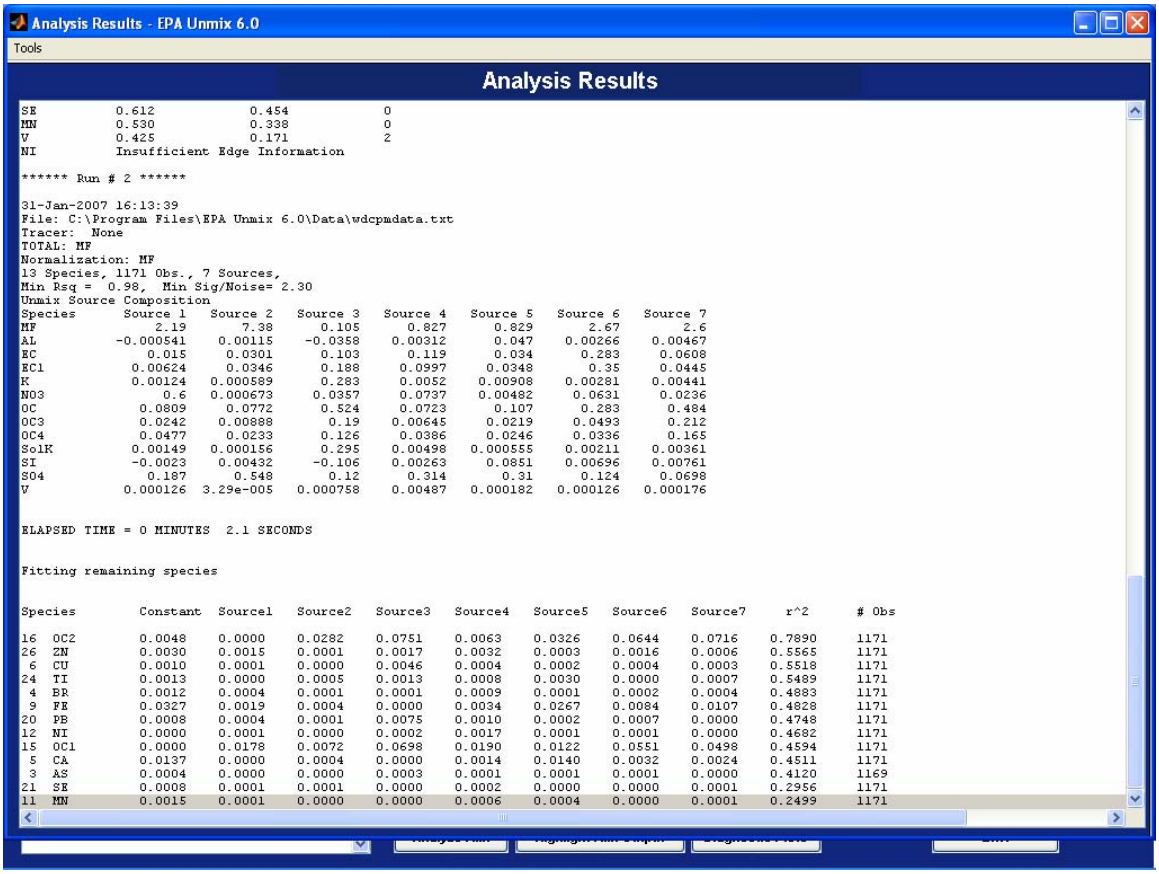


Figure 36: Fit Unselected Species Results

Each row is the result of a non-negative least squares regression of the source contributions (normalized to a mean of 1) and the species. The amount of the species in each of the sources, the amount not explained by the sources or constant, and coefficient of determination are displayed. Results are also sorted by the species R^2.

These results can also help guide the selection of additional species that can be added to the Unmix solution. For example, OC2 has a high R^2 value and after adding it to the Selected Species Window and selecting the Run button, the solution in Figure 35 is produced. This command can be used to create a consistent set of species for comparing EPA Unmix and EPA PMF profiles.

US EPA ARCHIVE DOCUMENT

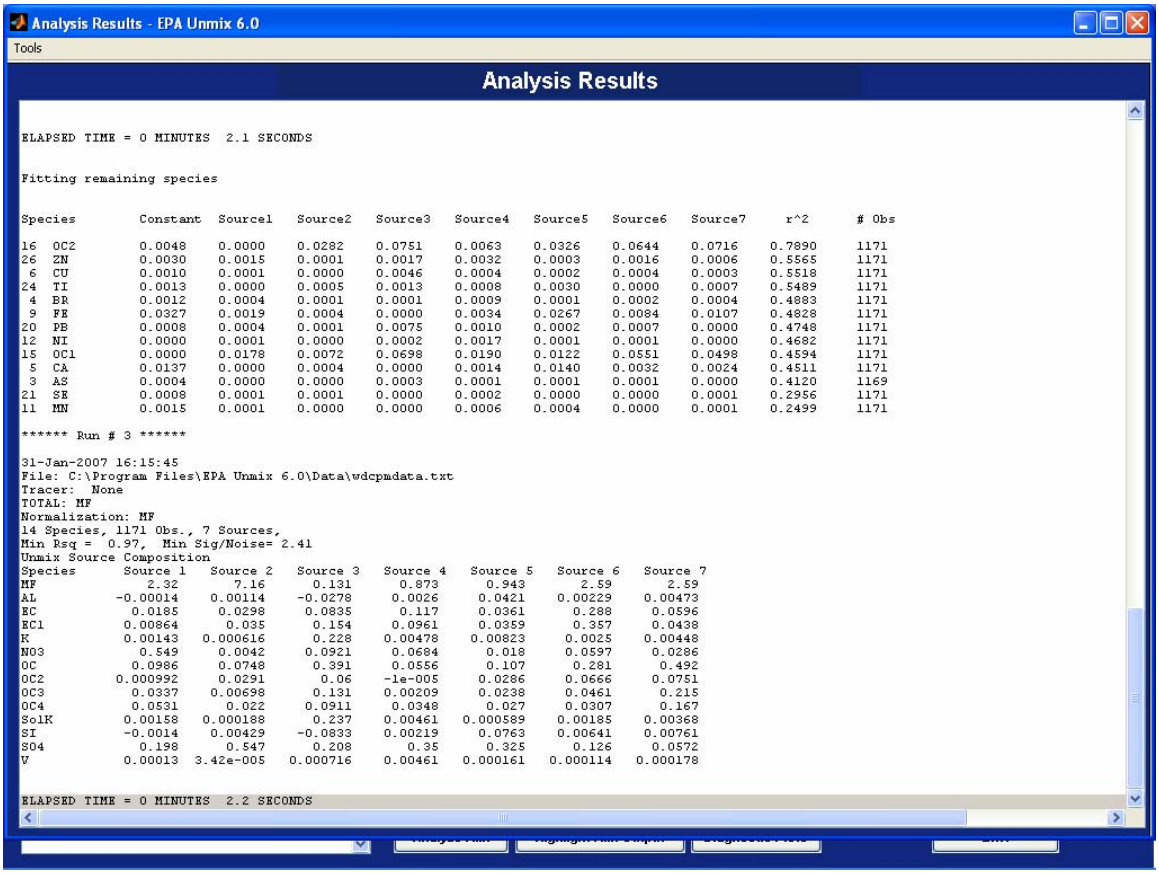


Figure 37: Adding species from Fit Unselected Species

5.4 Factor Analysis

The Unmix results can be compared to the typical factor analysis approach (varimax rotated factor analysis) by selecting the Species Selection Tools, Factor Analyze Selections command. Use the Factor Analyze Selections command to evaluate the Unmix results shown in Figure 36.

US EPA ARCHIVE DOCUMENT

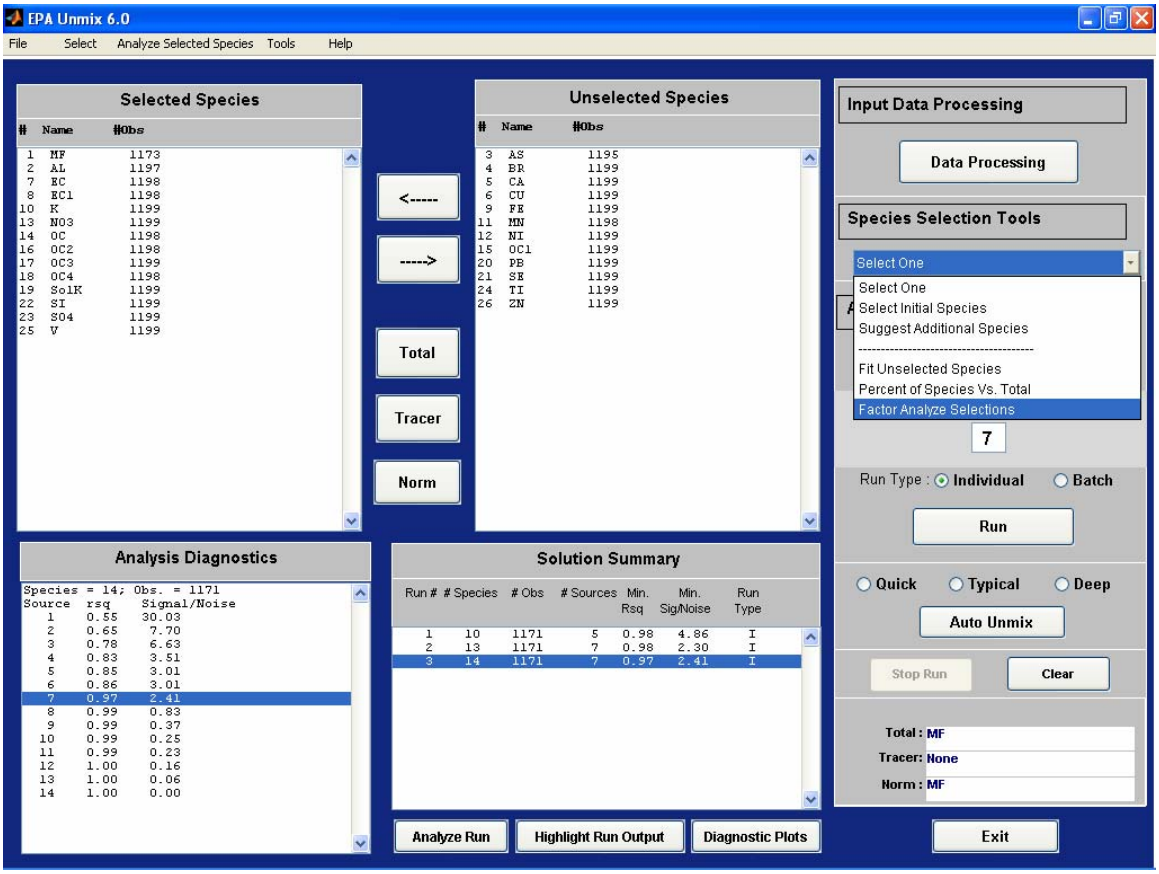


Figure 38: Factor Analyze Selections command

The factor, matrix scores, covariance matrix scores, explained covariance, and Scree plot are displayed. Select the Display Loadings button to show the loadings. The number of factors can be changed by selecting a value in the Number of Factors list. Selecting the Display Loadings button again will show the loadings associated with the new number of factors (Figure 37).

US EPA ARCHIVE DOCUMENT

Figure 39: Factor Analysis Results

Four factors have Eigenvalues greater than 1, which has been used as a cutoff for determining the number of factors in air pollution data. Compared to the factor analysis, Unmix is extracting three additional factors from the data. The factor analysis results can also be compared to the strong and significant species listed in the Fit Diagnostics output. Factor 1 represents motor vehicles with high species loadings on EC, EC1, NO3, OC, OC3, and OC4. Factor 2 represents crustal material with high factor loadings on Al and Si. Factor 3 represents wood burning with high loadings on K and SolK. Factor 4 represents secondary sulfate and additional combustion aerosol with high loadings on EC, EC1, OC2, and SO4.

5.5 Replace Missing Data

One common question when applying Unmix is what to do with species that have many missing values. Unmix does not use data from a sample if even one of the selected species has a missing value. For most species with many missing values, the solution is simply not to include these species in the model. However, sometimes the species are important because they could be indicators of a source. Selenium, nickel, and vanadium are examples of elements that often have many values below minimum detectable limits or infrequent point

US EPA ARCHIVE DOCUMENT

source impacts, but when these have high values, it may signify the impact of important sources such as coal combustion and residual oil combustion.

In general, missing values result from two causes. First is mechanical failure of the sampler, loss of the sample or other irretrievable event. Nothing can be done in these cases. Second and more often, missing data are below the minimum quantifiable (or detectable) limit. Data may be too low to quantify for two reasons: the amount in the sample is very low, or the amount in the sample is not small, but the species detection limit is raised by the presence of a large, near-by, interfering species.

Unmix allows the option of replacing missing values. Table 3 illustrates how a missing value is estimated. The fundamental idea is to find a value that is consistent with the ratios of the species in the data. In this example, 200 of the 400 values of VOC3 less than 100 in the umtest.txt data set were assumed to be missing. The method of finding a number to fill in a missing value is illustrated using row 15 of the umtestR.txt data. These data are reproduced as the first column in the table. The next column gives the smallest ratio of VOC3, when it is not missing, divided by each remaining VOC. The maximum ratio is in the next column.

To understand the algorithm, first consider VOC1. It has a concentration in the sample of 1.86, and when this is multiplied by the min and max ratios we get the concentrations 1.99 and 145.74 in the last two columns. The first is the smallest concentration of VOC3 that would be consistent with the observed ratios with VOC1 in the data. If VOC3 were smaller than this value, the ratio with VOC1 would be smaller than anything seen in the rest of the data. Similarly, if VOC3 is greater than 145.74, this would be outside the range of ratios with VOC1 seen in the data. Now VOC3 could have a value anywhere between 1.99 and 145.74 and be consistent with the rest of the VOC1 data.

The same calculation of limits is made for each VOC and Total. Notice that the smallest value of VOC3 that is consistent with all the data is the maximum of the minimum values for each species, which is 5.99. By the same reasoning, the largest value of VOC3 that is consistent with the data is the minimum of the maximum values given in the Table, or 15.67. This puts a fairly tight range of 5.99 to 15.67 on the possible values for the missing VOC3 concentration. The simple arithmetic mean might be chosen as the best estimate of the missing value. This is true if the distribution of the values is symmetric around the mean. However, air quality data is usually not so distributed, most often following a skewed distribution such as a lognormal or weibull distribution. In this case, the geometric mean is a better estimate of the most likely value. Thus, the final estimate of the missing VOC3 value is the geometric mean of 5.99 and 15.67, or 9.69. This compares well to the value of 11.83 in the original data. Finally, it is possible for the estimated minimum value to be greater than the estimated maximum value; if this happens the missing value is not replaced.

US EPA ARCHIVE DOCUMENT

Table 3: Missing Value Estimation

	Concentration	Min. Ratio to VOC 3	Max. Ratio to VOC3	Min. Conc. from ratios	Max. Conc. from ratios
VOC1	1.86	1.07	78.36	1.99	145.74
VOC2	8.84	0.68	1.91	5.99	16.88
VOC3	*	*	*	*	*
VOC4	3.40	0.77	8.29	2.61	28.19
VOC5	1.13	2.49	15.65	2.81	17.69
VOC6	3.61	0.92	5.53	3.33	19.97
VOC7	3.60	0.98	6.99	3.53	25.17
VOC8	3.47	0.84	8.71	2.92	30.22
VOC9	4.03	1.18	5.29	4.77	21.30
Total	61.92	0.09	0.25	5.64	15.67
Max of the minimums				5.99	
Min of the maximums				15.67	
Geometric mean		9.69	Value used to replace missing value		
Original concentration		11.83			

This method of filling in missing values has some obvious limitations. For the missing values of a species to be properly estimated, the species cannot be missing all or almost all of the time. If the species has almost all of its values missing, then the ratios of the non-missing data to the other species will not be representative and the estimates will be unreliable. Also, the greater number of species in the data, the better the method will work. Of course, one must always be cautious in replacing missing data. A basic assumption is that the conditions and contributing sources during the periods of missing data are the same as for the rest of the data. If for some reason this is not the case, then filling in missing data based on the existing data may not be advisable. Experience has shown that this method of dealing with missing data does not degrade Unmix solutions. In some cases, species cannot be replaced when the minimum replacement value is at least 1% greater than the maximum replacement value.

As a broad guideline, a species should have no more than about two-thirds missing data, but this depends on how many data points there are in the whole data set. Fewer points would require less missing data. Be sure to select all the species that have mostly non-missing data. It is these species that will be used by the algorithm to estimate the species with many missing values. It is not recommended at this time to mix data of different types (particulate data and gas data). Missing total data should not be replaced. Data files with and without missing data (umtest.txt) and with missing data (umtestR.txt) are included with EPA Unmix in the data folder. The umtestR data are the same as the umtest, except that VOC species 2 and 6 have some missing data. Load the umtestR data. The data do not have any date or time information and the missing value symbol is -99. In the Data Processing window, highlight the TOTAL variable in the Included Species box and select the Do not replace or use highlighted species checkbox. Next, select the Replace Missing Values button (Figure 38).

US EPA ARCHIVE DOCUMENT

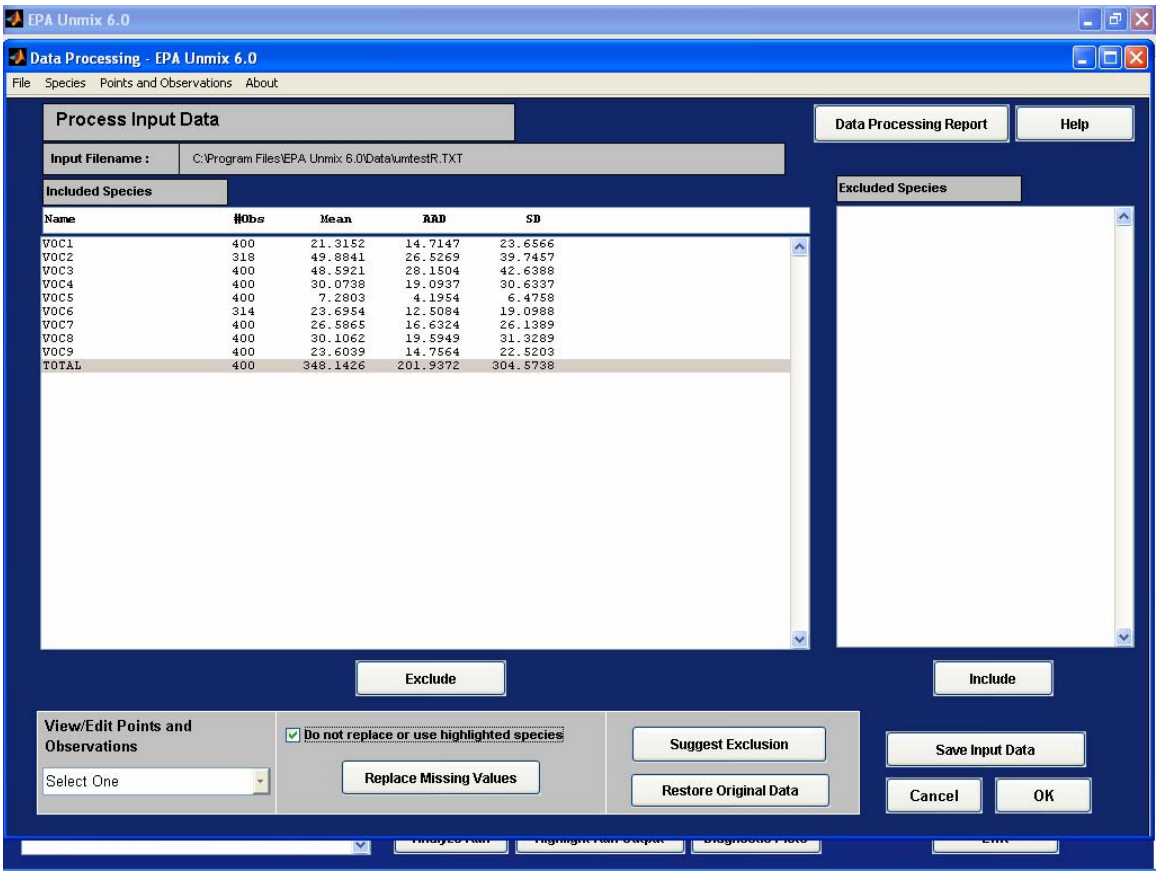


Figure 40: Replace Missing Values command

The number of replaced values and the replaced mean (RM) are displayed. The modified input data file can be saved for use by other programs using the Save Input Data button. The replaced values are reported in the Data Processing Report (Figure 39).

US EPA ARCHIVE DOCUMENT

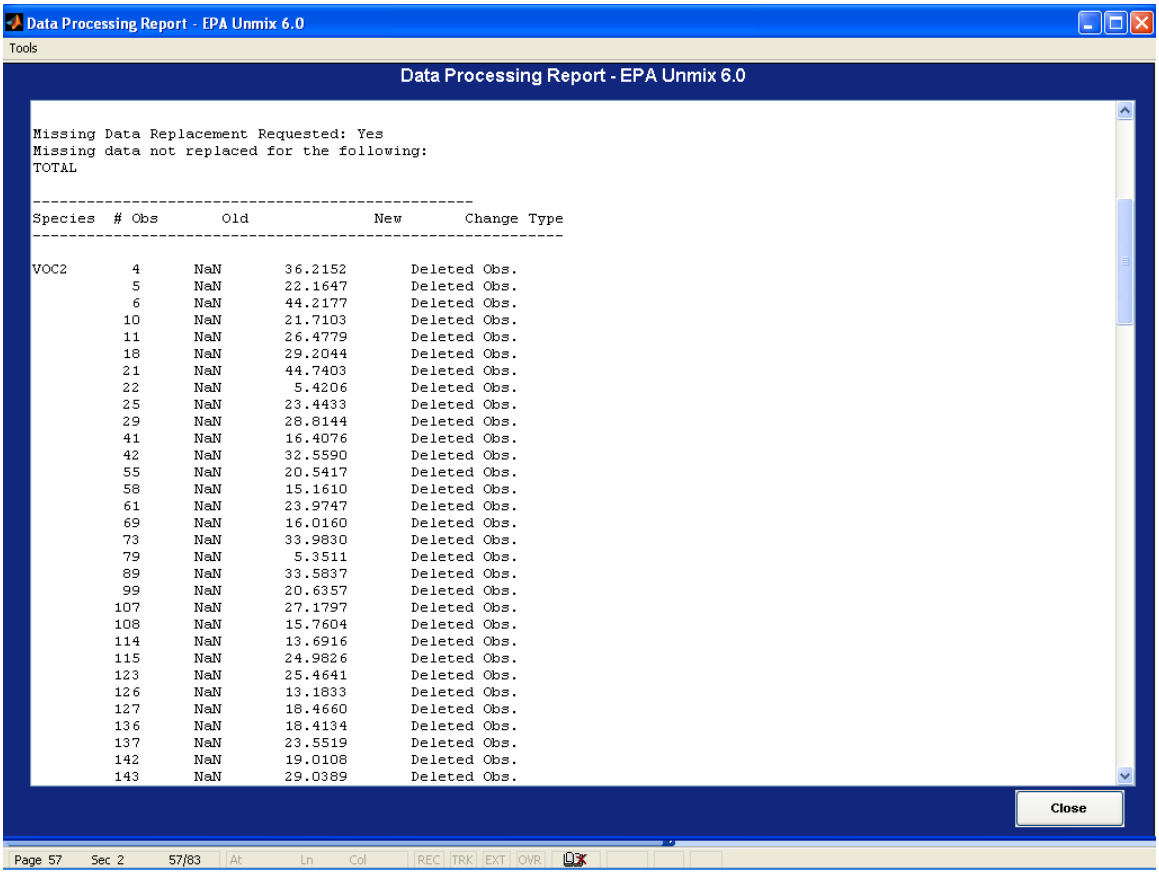


Figure 41: Replaced missing values

Close the Data Processing Report and select the OK button in the Data Processing window. Move all of the species to the Selected Species box. Set VOC1 as a Tracer, Total as the Total and Norm variable, and select the Run Type I option and Run button. Unmix results from using the original umtest data and after replacing the data in umtestR are shown in Figure 40.

US EPA ARCHIVE DOCUMENT

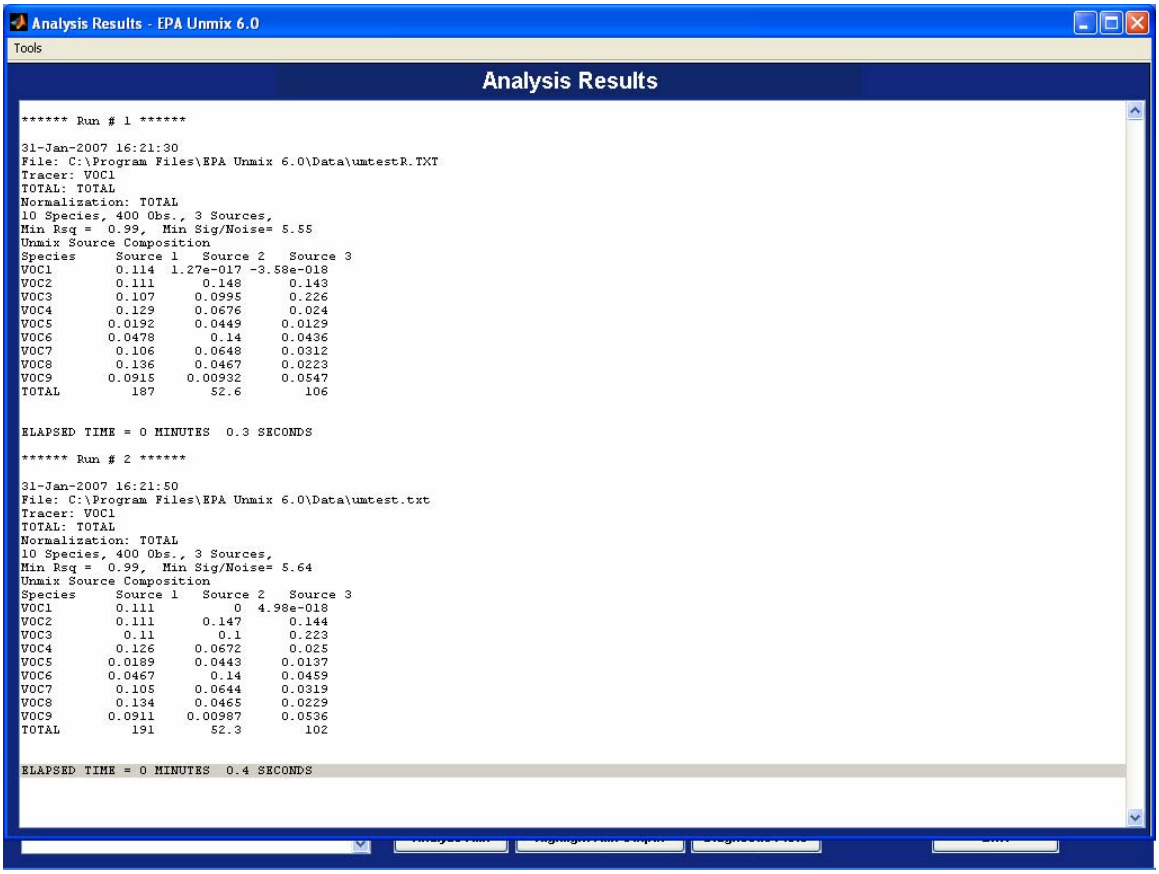


Figure 42: Comparison of umtestR and umtest results

SECTION 6. ADVANCE PLOTTING OPTIONS

EPA Unmix can be used to create custom figures that can be printed and exported for use in other programs. All plots can be saved by selecting File and the Save command (Figure 41).

US EPA ARCHIVE DOCUMENT

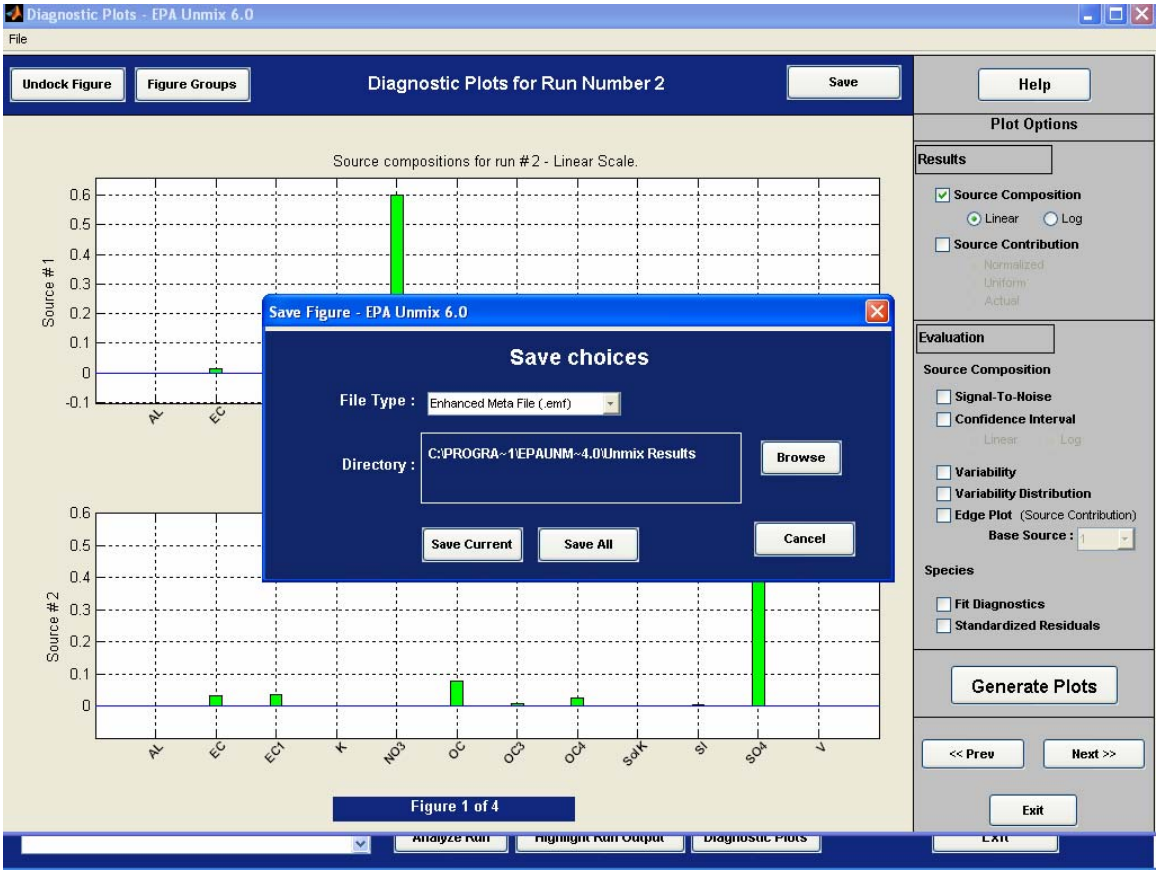


Figure 43: Saving Diagnostic Plots

6.1 Figure Groups

The following figures can be created using the results from the analysis of the wdcpmdata species from Run # 2 in Figure 11. Select the Diagnostics Plots button from the main window and select Source Composition (Log), Source Contribution (Actual), and Fit Diagnostic options. The source profiles (in mass fraction) will be displayed with two profiles on each chart (Figure 42).

US EPA ARCHIVE DOCUMENT

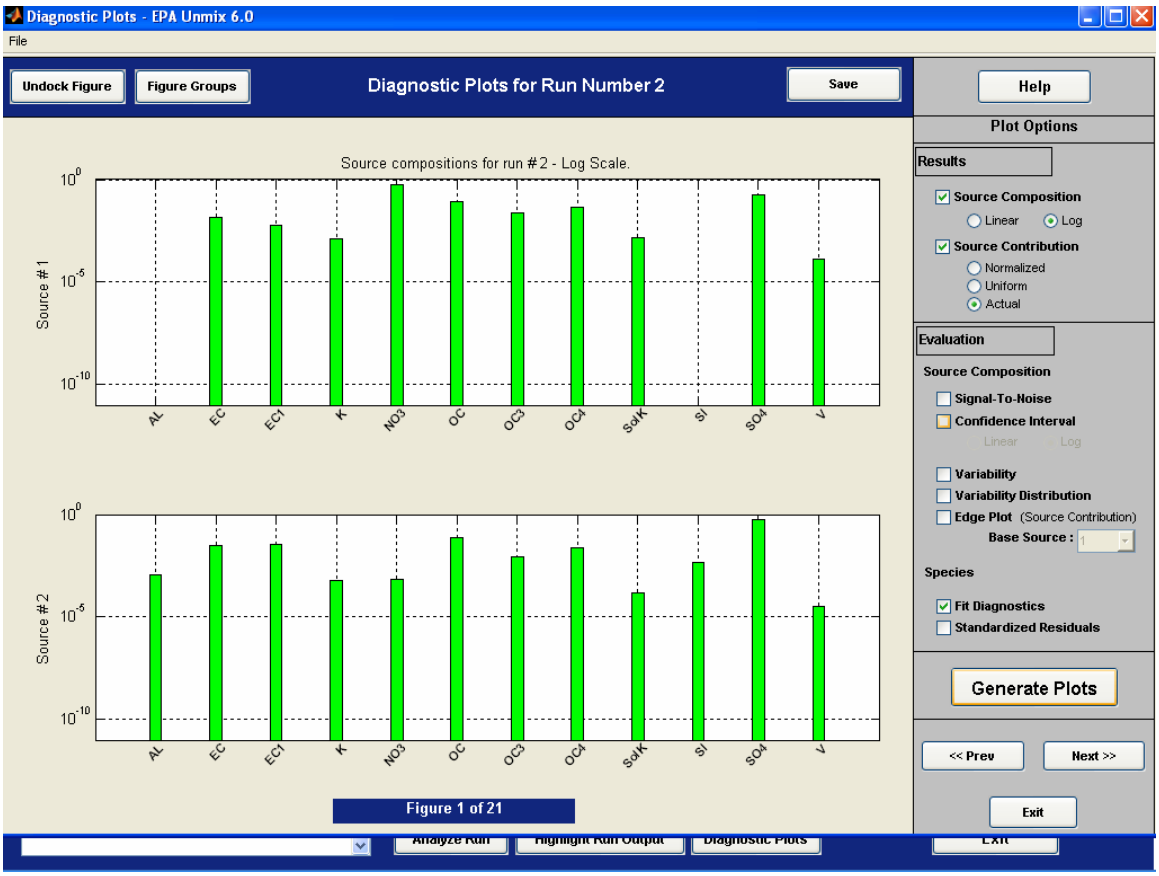


Figure 44: Source profile plots

Select the Figure Groups button and set the number of plots per page to All. A new figure will be created that contains all of the source profiles (Figure 45). Select the Figure Groups list to view the groups: Source Composition, Source Contribution, Fit Diagnostics – Scatter Plot, and Fit Diagnostics – Time Series.

US EPA ARCHIVE DOCUMENT

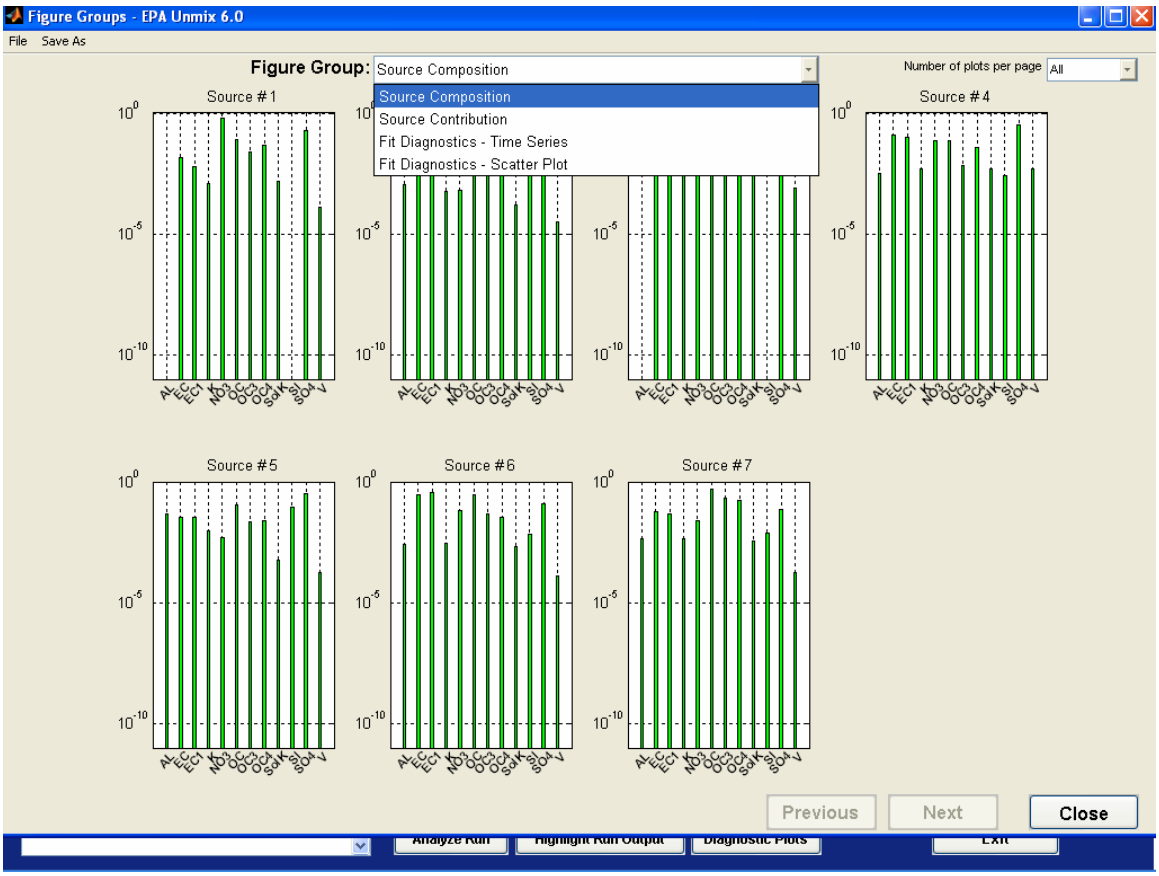


Figure 45: Figure groups

6.2 Edge Plots

A basic explanation of how Unmix uses edges in the data is found in Henry, 1997. An example of the edge plots is shown below using the umpmdata.txt file located in the C:\Program Files\EPA Unmix 6.0\Data directory. Load the umpmdata data (no date or hour information, missing value symbol is -99). Exclude the gases (NO2 and CO) and exclude the suggested species that are recommended for exclusion (As, Sr) and select the OK button.

Move MASS, Al, Si, S, K, Fe, OC, EC, and Sol_K (soil corrected K) from the Unselected to Selected Species window. Set MASS as the Total and Norm species and select the Run button (Figure 44). These selected species give a four source model.

US EPA ARCHIVE DOCUMENT

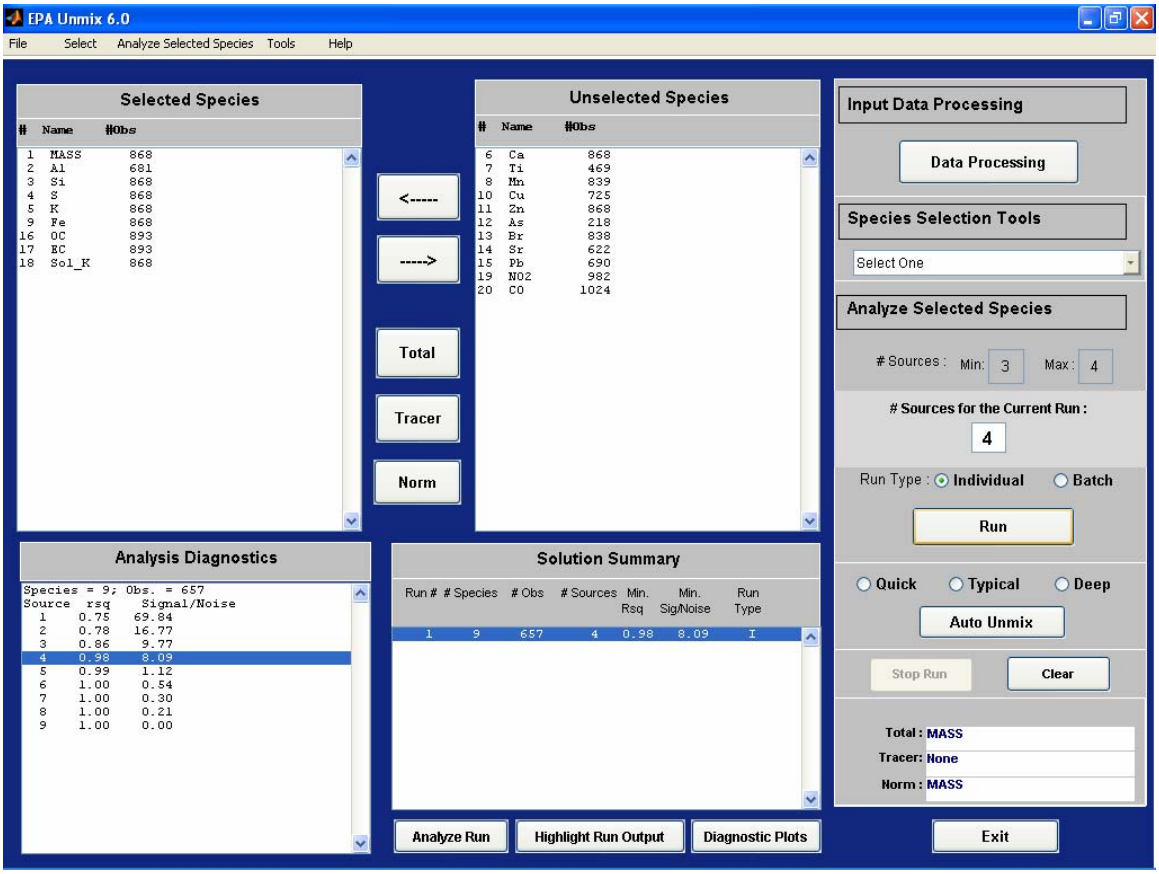


Figure 46: Umpdata edge plot example

To generate a plot of the source profiles, select the Diagnostic Plots button, Source Composition (linear) option, and Generate Plots button. Select the Figure Groups button to display the profiles on one plot. The four sources are vegetative or wood smoke (source 1), secondary sulfate (source 2), motor vehicle (source 3), and crustal (source 4) as seen in Figure 45.

US EPA ARCHIVE DOCUMENT

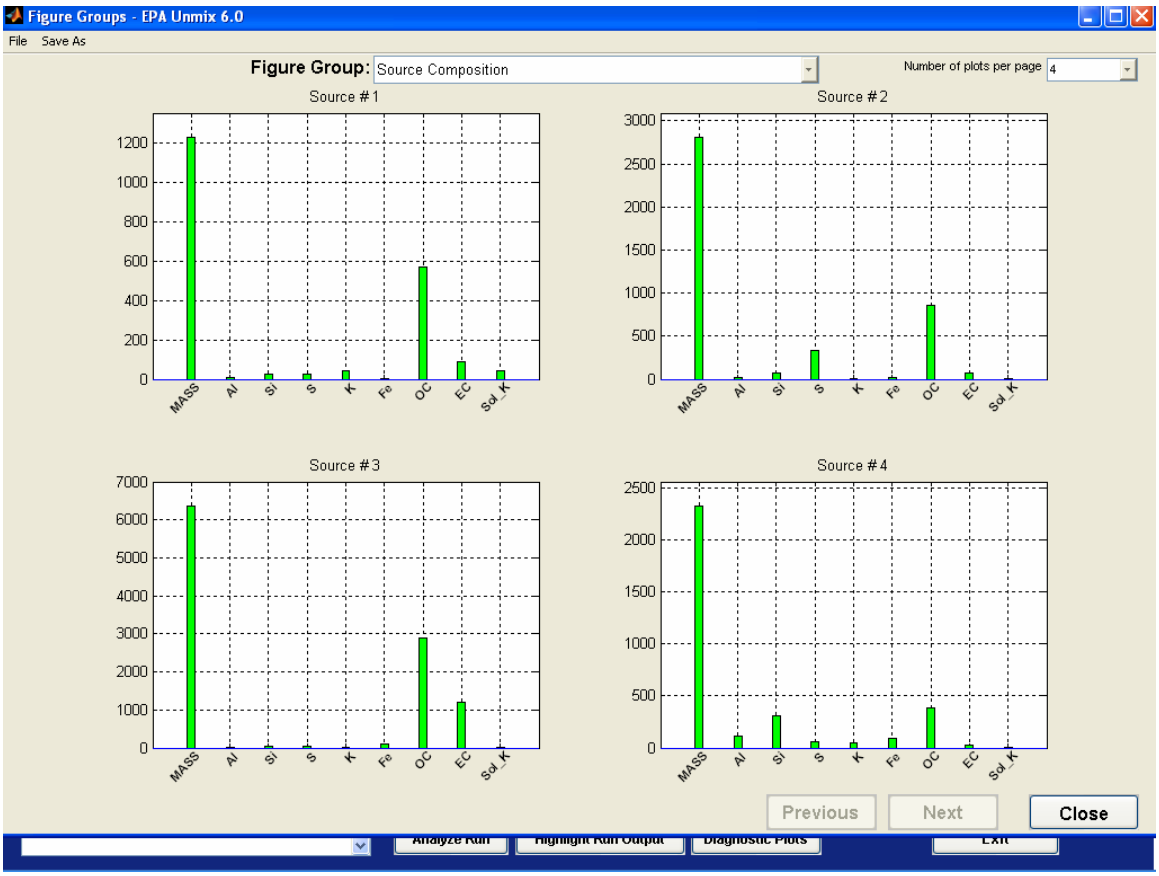


Figure 47: Umpdata source profiles

Unselect the Source Composition option and select the Edge plot option. Select source 2 (secondary sulfate) as the base factor. The edge plots are shown in Figure 46.

US EPA ARCHIVE DOCUMENT

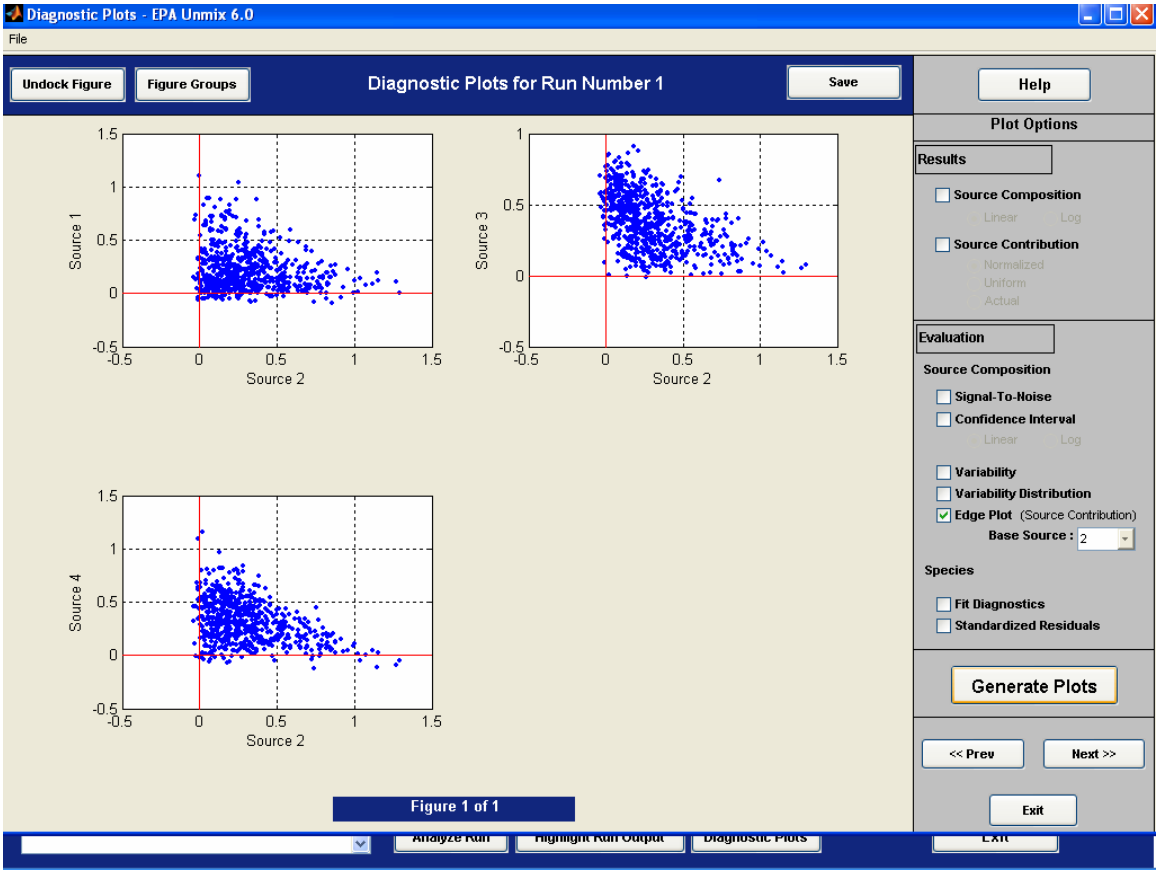


Figure 48: Edge Plots

The numbers plotted are the fraction of the sample that is from each source, thus the numbers lie between 0 and 1, except for the effects of error. The x and y axes are the edges. In both plots, the y-axis is the edge associated with source 2, the secondary sulfate. Points that are near the y-axis have very small contributions from vegetative or wood smoke source (source 1). All the edges in these plots are typical of "good" edges. Points can be selected by holding the left mouse button down and drawing a rectangle around them. Select the points with low motor vehicle contributions (Source 3). The selected points are contained within red squares while red circles are drawn around the same samples in the other figures (Figure 47).

US EPA ARCHIVE DOCUMENT

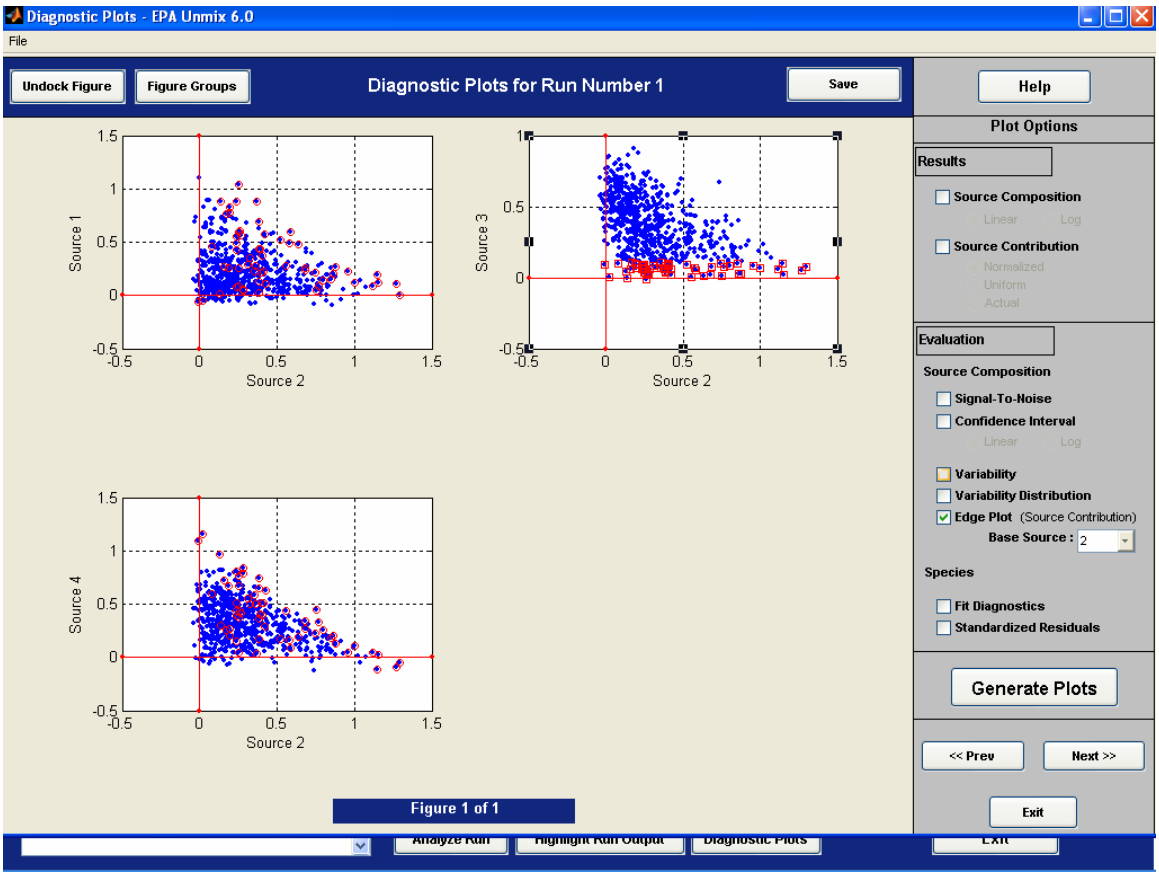


Figure 49: Selected Points in Edge Plots

The observation numbers are also displayed in the Analysis Results window. An example of poor edges can be seen by adding Zn to the species and running Unmix again. Evaluate the new five source solution by re-plotting the edge plots against the secondary sulfate source (source 3). Select the points near the x-axis in the source 4 plot (Figure 48).

US EPA ARCHIVE DOCUMENT

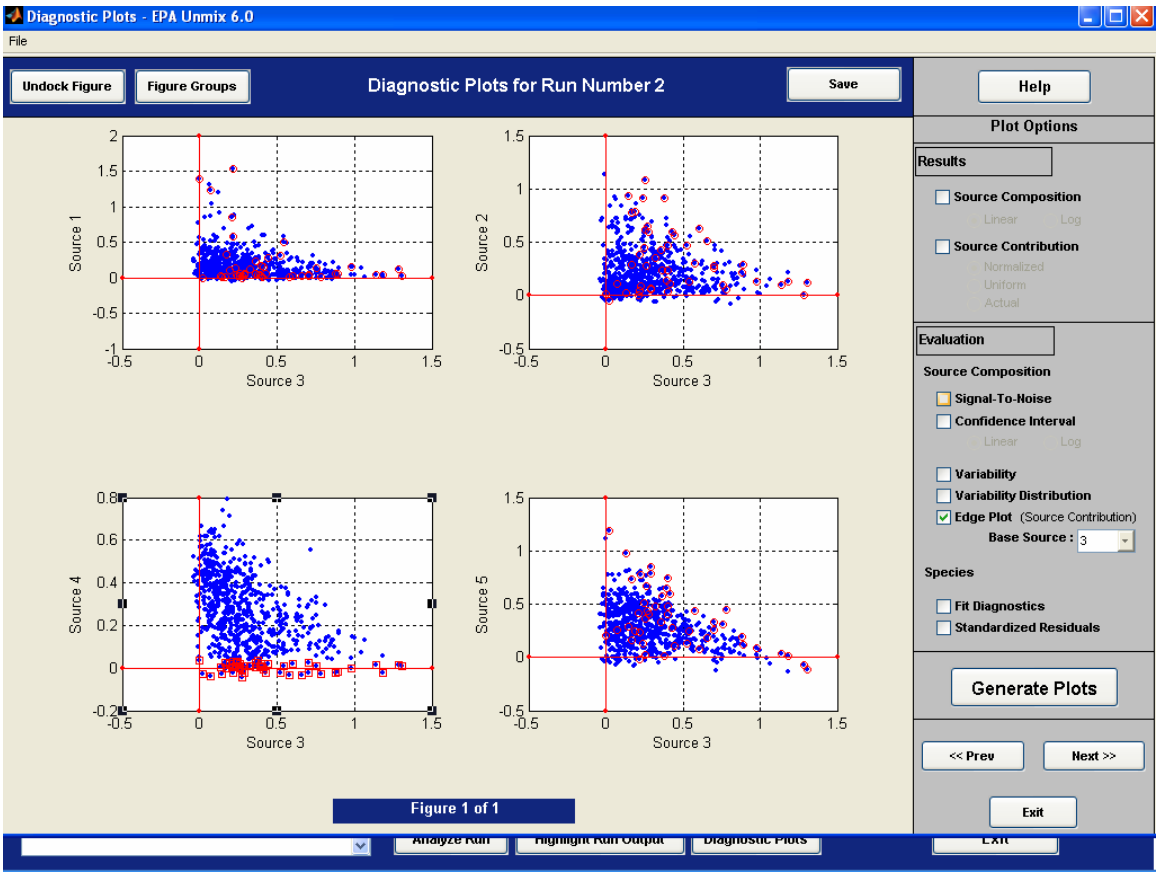


Figure 50: Example of poorly defined edges

The source 4 plot has just a few points near the x-axis that define the edge. In general, edges that are dependent on just a few points will be more greatly affected by errors than edges that are defined by many points. Since Unmix uses edges to find the source compositions, poor edges will lead to increased variability in the source compositions.

SECTION 7. BATCH MODE

The Batch Mode command can be used to evaluate the addition of species recommended by the Suggest Additional Species command. Batch mode will add all combinations of the highlighted species in the Unselected Species window to the Selected Species and evaluate each set of species. To run the Batch Mode command, select the Batch mode radio button and select the Run button. The following example uses the wdcpmdata data file and the selected wdcpmdata species from Run # 2 in Figure 11. Run the Influential Point Algorithm to obtain information on the species that have not been highlighted in the Unselected Species box (Figure 49).

US EPA ARCHIVE DOCUMENT

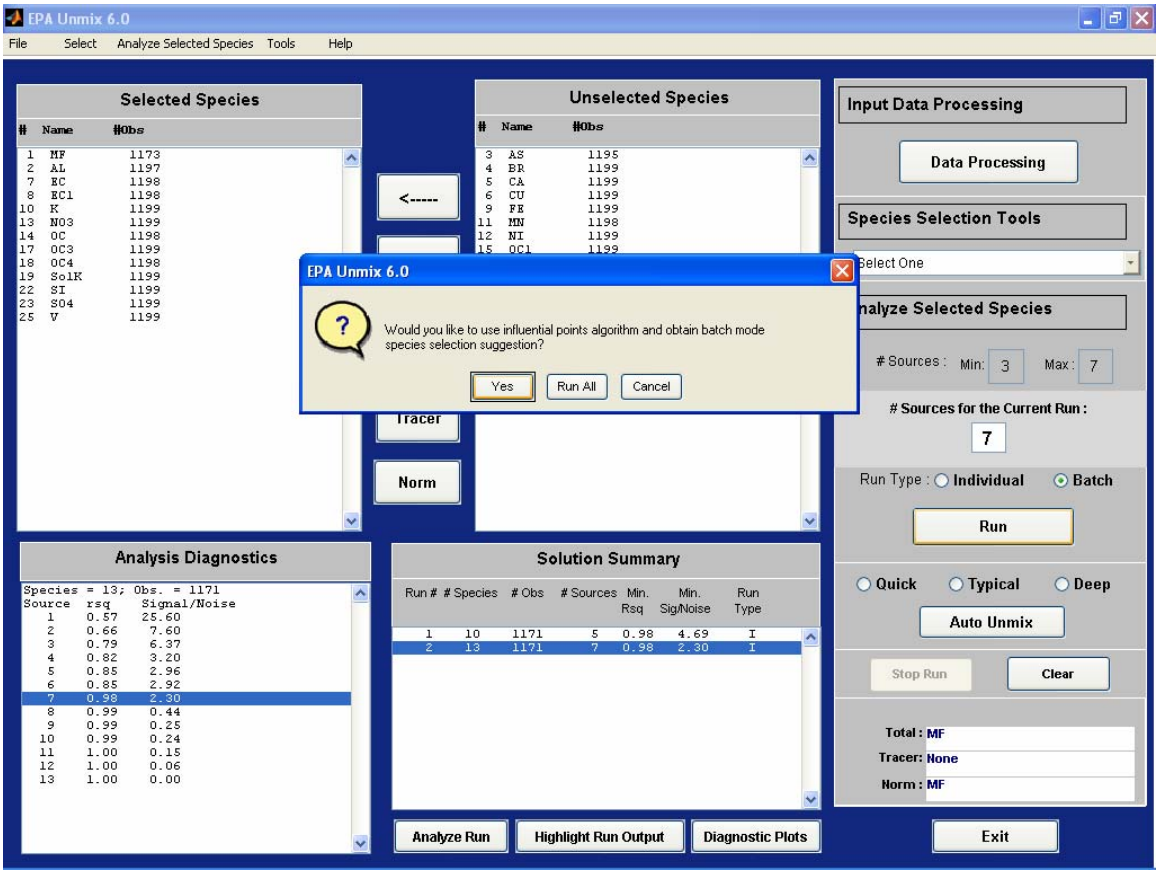


Figure 51: Batch mode influential point option

Select species with good edge resolution and low number of influential points from the Species Suggestions Window. For this example, select Zn, Ti, Fe, Ca, OC2, Pb, and Se. Choose the Select button and the Batch Mode Preferences window will open (Figure 50). Save the Batch mode results in a data file and select the OK button.

US EPA ARCHIVE DOCUMENT

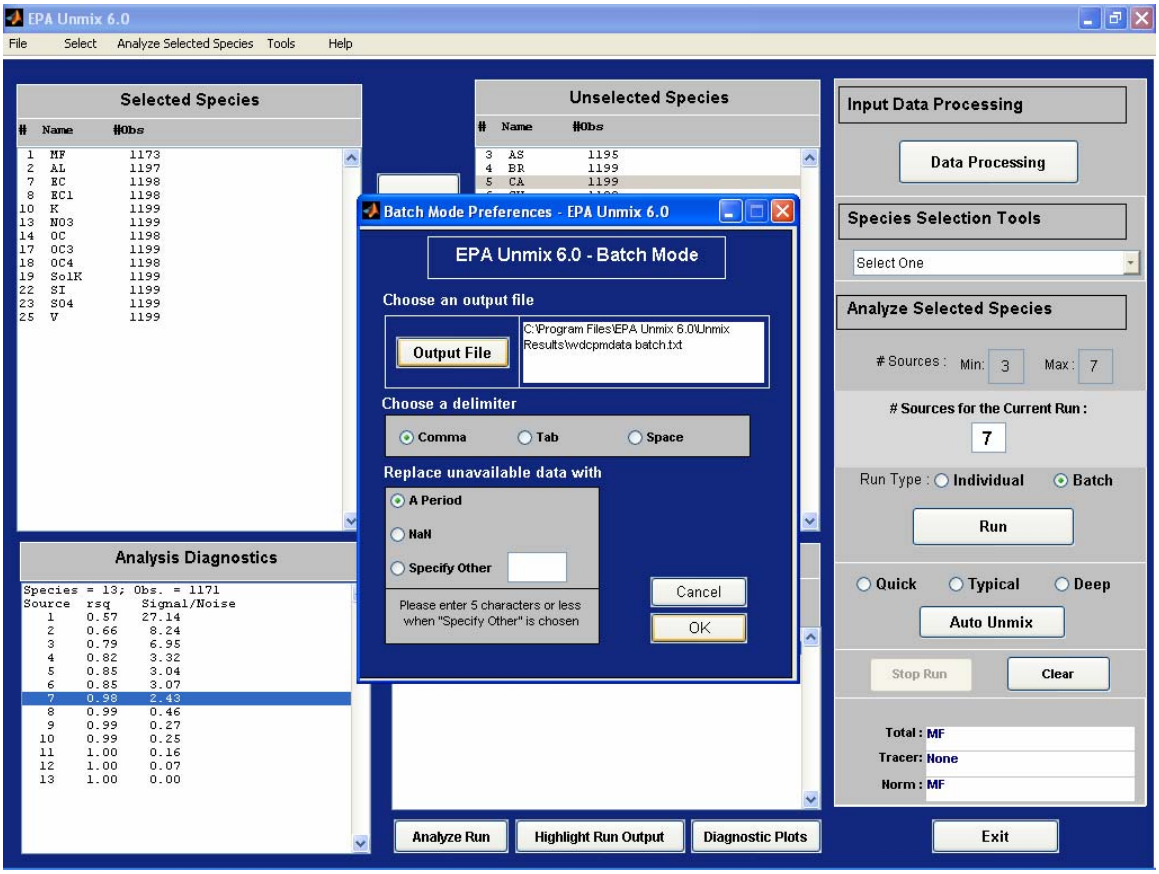


Figure 52: Batch Mode Preferences

Progress bars will show the Batch mode status. The Batch mode can be stopped by selecting the Stop Run button and the results will be saved in the output file. The Solution Summary box will list the Batch Mode results (Run Type "B") as shown in Figure 51. The Analyze Run and Diagnostic Plots buttons are not available for Batch Mode model results.

US EPA ARCHIVE DOCUMENT

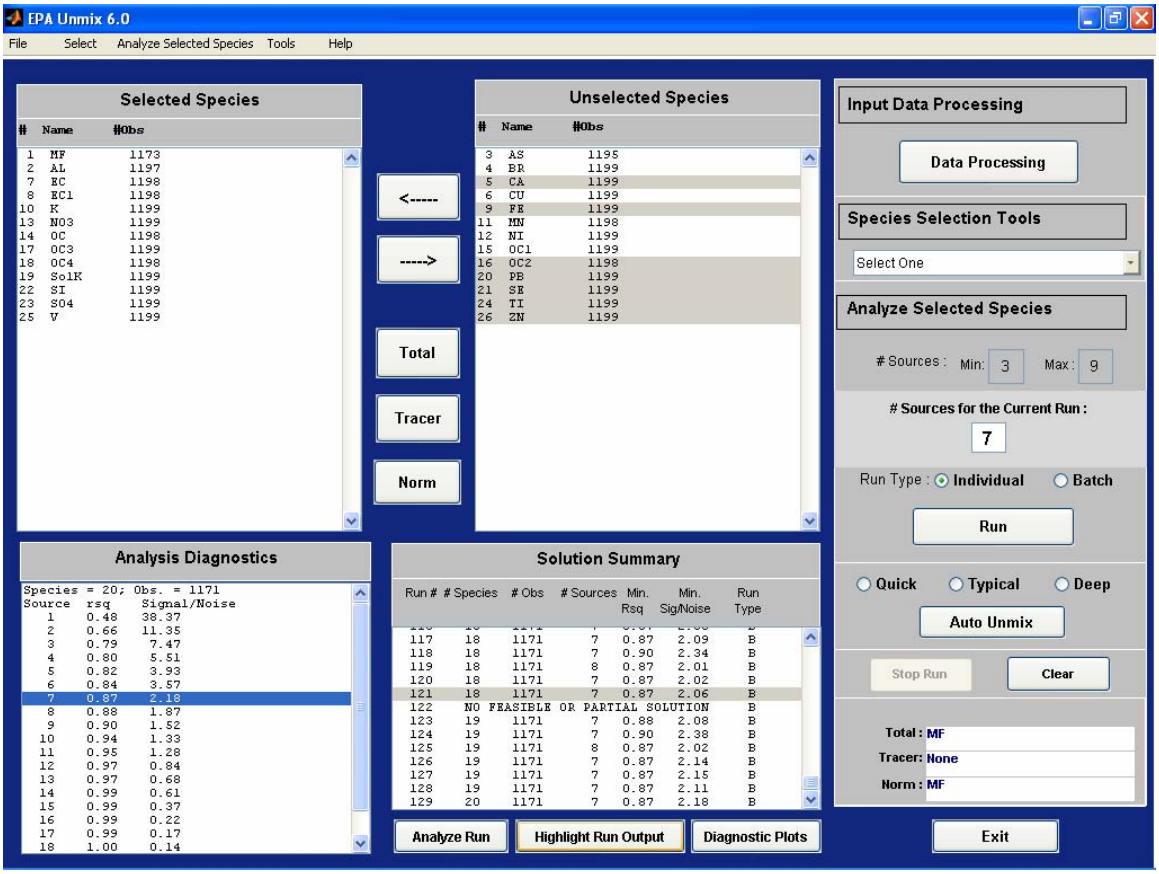


Figure 53: Batch Mode Solution Summary

Choose Run # 121 from the Solution Summary and select the Highlight Run Output button to view the results (Figure 51). The run number may be different due to the number of individual runs prior to the batch mode run but will have 18 species and be before the “NO FEASIBLE OR PARTIAL SOLUTION” and before the # Species increases to 19. Five species were added to the selected species: OC2, PB, SE, TI, and ZN (Figure 52).

US EPA ARCHIVE DOCUMENT

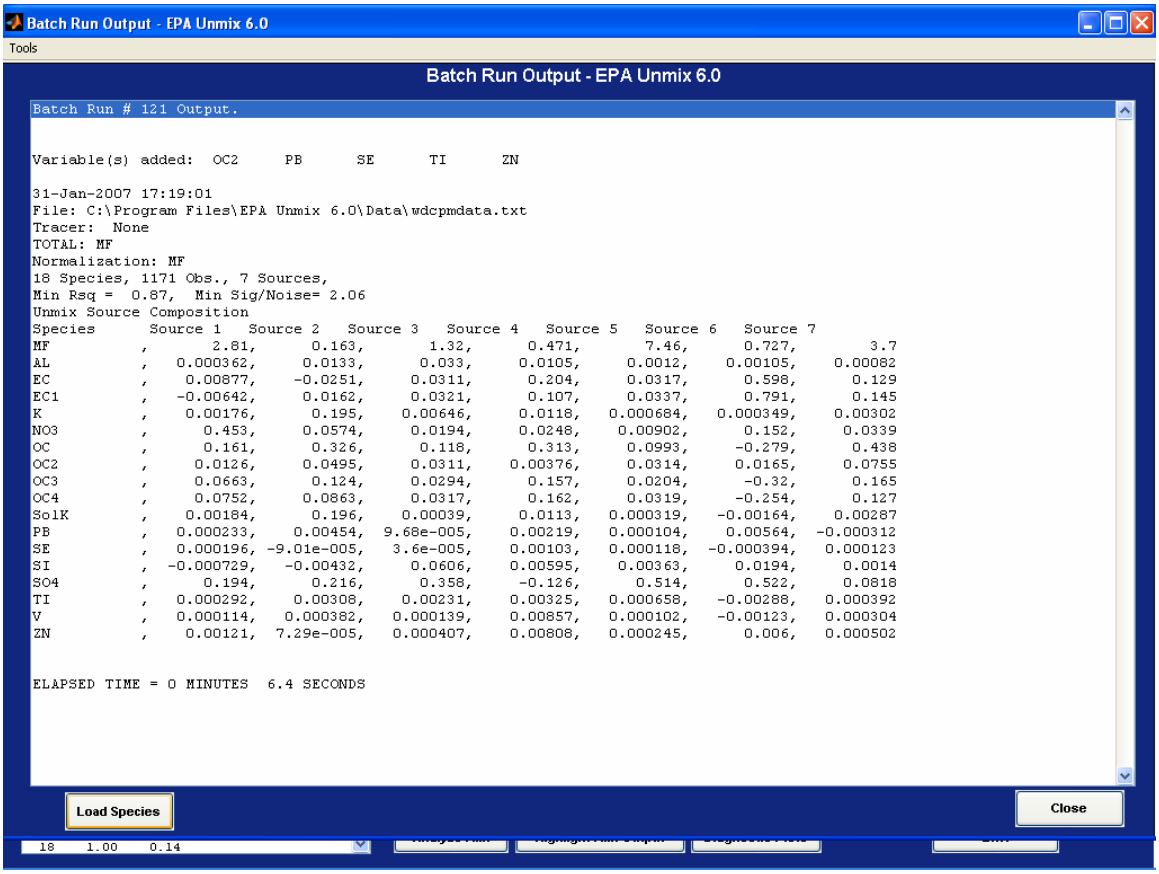


Figure 54: Batch Mode Analysis Results

Select the Load Species button to add these species to the Selected Species box and select the Run button. The new solution can now be evaluated with the Analyze Run and Diagnostic Plots options.

US EPA ARCHIVE DOCUMENT

SECTION 8. Unmix PUBLICATIONS

Henry R.C. (2005) Duality in multivariate receptor models. Chemometrics and intelligent laboratory systems. 77: 59 – 63.

Henry R.C. (2003) Multivariate receptor modeling by N-dimensional edge detection. Chemometrics and intelligent laboratory systems. 65: 179 – 189.

Henry R.C. (2002) Multivariate receptor models- current practices and future trends. Chemometrics and intelligent laboratory systems 60: 43- 48.

Eun Sug. Park, C. Spiegelman, and R.C. Henry (2000) Estimating the number of factors to include in a high-dimensional multivariate bilinear model. *Communications in Statistics, Simulation & Computation* 29: 723-746.

Henry, R.C. and B.- M. Kim (1989) A Factor Analysis Model with Explicit Physical Constraints, *Transactions Air Pollut. Control Assoc.* 14:214–225.

Henry, R.C. and B.-M. Kim (1990) 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 Composition from Ambient Data: The GRACE/SAFER Method, *Environ. Sci. Technol.* 28:823–832.

Henry, R.C. (1997) History and Fundamentals of Multivariate Air Quality Receptor Models, *Chemom. Intell. Lab. Syst.* 37:525–530.

Henry, R.C., and C. Spiegelman (1997) Reported Emissions of Volatile Organic Compounds are not Consistent with Observations, *Proc. Nat. Acad. Sci.* 94:6596–6599.

Henry, R.C., E.S. Park, and C.H. Spiegelman (1999) Comparing a New Algorithm with the Classic Methods for Estimating the Number of Factors, *Chemom. Intell. Lab. Syst*. 48:91–97.

Hopke. P.K., et al. (2006) PM source apportionment and health effects: 1. Intercomparison of source apportionment results, Journal of Exposure Science and Environmental Epi. 16:275-286.

Kim, B.-M., and R.C. Henry (1999) Extension of Self-Modeling Curve Resolution to Mixtures of More Than Three Components. Part 2: Finding the Complete Solution, *Chemom. Intell. Lab. Syst*. 49:67–77.

US EPA ARCHIVE DOCUMENT

Kim, B.-M., and R.C. Henry (2000) Application of the SAFER Model to Los Angeles PM10 Data, *Atmos. Environ*. 34:1747–1759.

Kim B-M and R. C. Henry. Extension of self-modeling curve resolution to mixtures of more than three components. Part 3. Atmospheric aerosol data simulation studies (2000) *Chemometrics and Intelligent Laboratory Systems* 52: 145-154.

Lewis, C.W.; Henry, R.C.; Shreffler, J.H. (1998) An exploratory look at hydrocarbon data from the Photochemical Assessment Monitoring Network, *J. Air & Waste Manage. Assoc..*48: 71 – 76.

Lewis, C.W.; Norris, G.; Henry, R. (2003) Source Apportionment of Phoenix PM2.5 Aerosol with the Unmix Receptor Model, *J. Air & Waste Manage. Assoc..* 53: 325-338.

Mukerjee, S.; Norris, G.A.; Smith, L.A.; Noble, C.A; Neas, L.M.; Ozkaynak, A.H., Gonzales, M. (2004) Receptor Model Comparisons and Wind Direction Analyses of Volatile Organic Compounds and Submicrometer Particles in an Arid, Binational, Urban Air Shed *Environ. Sci. Technol.* 38: 2317 - 2327.

Kim B-M and R. C. Henry. Application of the SAFER model to Los Angeles PM10 data (2000). *Atmos. Environ.* 34:747-1759.

Park E.S., Spiegelman C, and Henry R.C. Estimating the number of factors to include in a high-dimensional multivariate bilinear model (2000) *Communications in Statistics, Simulation & Computation* 29:723 – 746.

Poirot, R.L., P.R. Wishinski, P.K. Hopke, and A.V. Polissar (2001) Comparative Application of Multiple Receptor Methods to Identify Aerosol Sources in Northern Vermont, *Environ. Sci. Technol.*, 35:4622-4636.

Willis, R.D. (2000) Workshop on Unmix and PMF as Applied to $PM_{2.5}$, U.S. Environmental Protection Agency Report No. EPA/600/A-00/048, Research Triangle Park, NC, June 2000.

Willis, R.D. W.D. Ellenson, T.L. Conner (2001) Monitoring and Source Apportionment of Particulate Matter Near a Large Phosphorous Production Facility, *J. Air & Waste Manage. Assoc.* 51: 1142-1166.

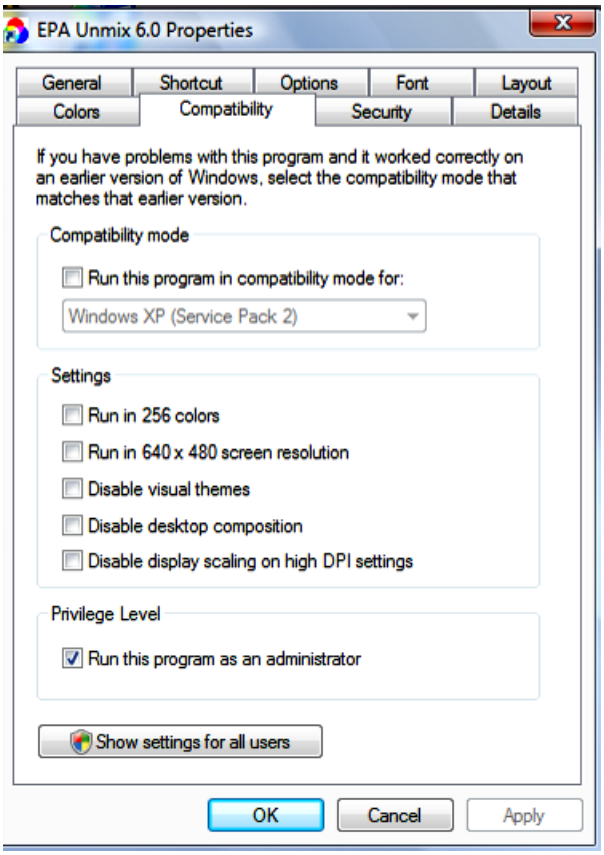
Xin-Hua Song, Alexandr V. Polisssar, Phillip K. Hopke, Sources of fine particle composition in northeastern US (2001) *Atmos. Environ.* 35:5277-5286.

US EPA ARCHIVE DOCUMENT

APPENDIX A: INSTRUCTIONS FOR RUNNING UNDER WINDOWS VISTA

Installation of EPA Unmix on a machine with Windows Vista operating system (OS) does not differ other operating systems. The installation still requires system administrator privilege. However, Windows Vista users have an additional constraint due to the inner workings of their OS. While users on Windows XP machines can run EPA Unmix as regular users, users on Windows Vista machines are required to run the machines as system administrators to be able to run EPA Unmix. EPA Unmix will not run on Windows Vista without this change.

To implement this, **right mouse** click on the EPA Unmix shortcut (found on the desktop) and select the "Properties" menu option from the pop menu. A window resembling the image shown here appears. Select the "Compatability" tab. In the bottom of the tab sheet, check the "Run the program as an administrator"



checkbox (as showns above) in the "Privilage Level" section and press the "OK" pushbutton. This will ensure that the user can run EPA Unmix program on their machine.

The user should contact their system administrator if they are unable to execute the above mentioned steps to avail themselves the ability to run EPA Unmix.

US EPA ARCHIVE DOCUMENT

APPENDIX B: INSTALLING A NEW VERSION OF EPA Unmix

If anyone wishes the install a new version of EPA Unmix when they already have a version on their machine, use the following steps.

1) Double-click on the installation package and keep pressing "Next" until the installation is complete.
2) Find the unmix6R_mcr sub-folder in the Program Files\EPA Unmix 6.0 folder and delete the folder. Since this folder is created on the user's machine, deleting does not have deleterious effect. It only ensures that this folder is correctly generated on the user's machine.
3) Double-click on the icon titled EPA Unmix 6.0 on the desktop.

US EPA ARCHIVE DOCUMENT

APPENDIX C: VARIABILITY CALCULATION ALGORITHM

Unmix processes input data to produce feasible solutions by placing equal weights on all chosen observations. An important aspect of the Unmix approach is that the solution finding method is constructive in nature. That is, the model first defines a solution space and follows it through with finding a "point" in space that is called a feasible solution. Thus, in constructive models, it is imperative that the construction technique that produced a feasible solution be confirmed by other means. One of the more popular methods is the bootstrapping method. (The complimentary process to construction is finding solution through a process of elimination where a solution is singled out as the most feasible solution from among a multitude of possible solutions.)

In constructive solution approach with equal weights criterion, the magnitude of a few observations can have disproportional effect on the solution. The solution may be a "point" at the edge of the feasible solution space. The absence of some observations can either produce a solution point far away from the original solution point (qualitatively different solution) or, in the worst case, place the point outside the feasible solution space (non-feasible solution). Therefore, it is important to investigate the certainty (or uncertainty) of the solution of interest before proceeding to interpreting. We will, henceforth, refer the data set used to obtain the feasible solution as the base data set. Variability estimates associated with a feasible solution are computed using the bootstrap method. Data sets, termed as bootstrap variates, are created by sampling the base data set. Each variate is analyzed using the same model to obtain a range of values from the bootstrap runs and thus the Variability associated with the feasible solution obtained from the base data set.

Bootstrapping Method

Let $\mathbf{X} = \{X_1, X_2, X_3 \ldots X_N\}$ be an N-observation vector where X_i is an M-vector of concentration measurements of M species. The vector **X** is assumed to have temporal association in the following manner: The observations are the result of analysis done on the specimen samples collected from a filter placed at a site of interest. The samples can range from 30-minute averages to daily averages. In some rare cases, data from multiple sites may be present. In those cases, it is hard to classify the time interval of the input data set.

Based on the data set, a subset of the species and observations are chosen for further analysis to determine the source of the species. Source apportionment models such as EPA Unmix may be used to analyze the data to suggest composition and contribution matrices for the chosen set of species and observations. Using the composition matrix, sources can be identified by known signatures of well known sources. In Unmix, the presence of a composition and its associated contribution matrices is called a feasible solution for the chosen set of species and observations. This pair of composition and contribution matrices

US EPA ARCHIVE DOCUMENT

will be called the base feasible solution. This solution can be analyzed for its robustness using the bootstrap runs.

An important detail in creating the bootstrap variate is the sampling method. If the observations were samples from an independent, identically distributed random distribution, simple sampling techniques should suffice. But, the data sets used in source apportionment models are time series data that can be categorized as dependent data. Simple sampling techniques can create bootstrap variates that vary intrinsically from the base data set and lead to incorrect Variability estimates. Serial correlation, defined as the statistical dependence of a datum on its predecessors, is the most important intrinsic quality of time series data. Block sampling of the base data produces bootstrap variates that preserve serial correlation of time series data. The next important and obvious question is to ask what optimal block length produces an acceptable bootstrap variate. The block length that produces the bootstrap variates that is qualitatively and quantitatively comparable to the base data set is called the optimal block length. The quantitative requirement is easily defined. The size of the bootstrap variates must be the same as the base data set. Qualitative comparisons between the data sets are harder to define. We will address that later in this document. It is worth noting that that optimal block length alone will not be enough to produce qualitatively comparable data sets. However, it is a step in the right direction.

Block Length Calculation Schemes

Various schemes have been suggested to derive the optimal block length for a given data set. Many of them involve using a statistical quantity of a bootstrap variate with the same statistical quantity of the base data set. Frequently mentioned statistics are bias or variance, one-sided probability, two-sided probability, and auto-correlation.

Among the schemes to arrive at a reasonable block length to build bootstrap variates, we will compare two schemes that help decide on the optimal block length. One is based on the trial and error estimation using one of the three parameters associated with a data set: variance, one-sided and two-sided distribution as described in the paper titled “On blocking rules for the bootstrap with dependent data” by Hall, Horowitz and Jing (1995). This method will be referred as variance estimation method. The other method is based on the spectral estimation via a flat-top lag-window as described in the paper titled “Automatic Block-Length Selection for the Dependent Bootstrap” by Politis and White (2004) and will be referred as the spectral estimation method.

Using a fixed block length for all data sets or even a step function approach to fixing the block length can have deleterious effects. For instance, using a block length of 3 can produce bootstrap variates that are much more similar to the base data set. In the example data set, it was found that the maximum deviation of the bootstrap variates generated using block length of 3 were consistently

US EPA ARCHIVE DOCUMENT

lower than the maximum deviation of bootstrap variates generated using block length suggested by one of the two methods. In other words, bootstrap variates using the block length of 3 produced variates more similar to the base data set more often. While a block length of 3 may be appropriate for some data sets, caution should be exercised when determining the block length of each base data set.

Block Length Calculation - Variance Estimation Method

Hall, Horowitz and Jing (1995) have shown that the choice of the optimal block length for dependent data depends significantly on the yardstick chosen to measure the quality of a bootstrap variate. They have shown that the optimal block length is of $o(N^{1/3})$ when the yardstick used to compare is either the bias or the variance, $o(N^{1/4})$ when one-sided distribution is used and $o(N^{1/5})$ when two-sided distribution is used as the yardstick. The symbol o() implies a proportional relationship. That is, when bias or variance is used as the yardstick, the optimal (*) block length $L^* = k\, N^{1/3}$ for some positive constant *k*.

The following recursive algorithm is suggested to obtain the optimal block length value:

The algorithm suggests using a “seed” block length first to obtain the estimates of the chosen statistical quantity. Then, choose subsets of the base data set and try all other block lengths smaller than the seed block length. Isolate the size of the specific subset and its block length that has the minimum value for the chosen statistical quantity, and then create the next iterate for the block length. Use this iterate as the new seed and follow the procedure detailed above. When the seed and the computed block length obtained from the seed are sufficiently close, the iteration is stopped and the current value of the seed to considered the optimal block length.

Block Length Calculation - Spectral Estimation Method

Politis and White (2004) have detailed a method based on spectral estimation to suggest an optimal block length. A plug-in method is suggested to derive at an optimal block length for a given data set. This method appears to factor the serial correlation aspect of the given data set to derive the optimal block length. The yardstick used to measure the quality of a bootstrap variate is the variance. As in the Hall, Horowitz and Jing (1995) paper, Politis and White (2004) show that the optimal block length is of $o(N^{1/3})$. The plug-in method suggests a formula to compute the constant of proportionality. The formula suggests optimal block length in the L^* is given by

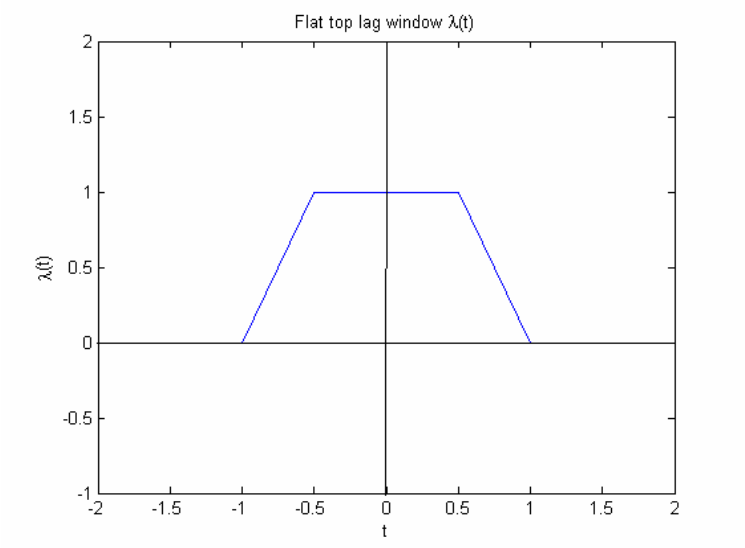
US EPA ARCHIVE DOCUMENT

$$L^* = \left[\left(\frac{2\hat{G}}{\hat{D}} \right)^{\frac{1}{3}} \right] N^{\frac{1}{3}},$$

where

$$\hat{G} = \sum_{k=-M}^{M} \lambda\left(\frac{k}{m}\right) | k | \hat{R}(k),$$

the flat-top lag window is following function given by



Also,

$$\hat{R}(k) = \frac{1}{N} \sum_{1=1}^{N-|k|} \left(X_i - \overline{X}_N\right)\left(X_{i+|k|} - \overline{X}_N\right),$$

$$\hat{D} = 4g^2(0) + \frac{2}{\Pi} \int_{-\Pi}^{\Pi} \left(1+\cos(w)\right)\, g^2(w)\, dw,$$

and

$$g(w) = \sum_{k=-M}^{M} \lambda\left(\frac{k}{m}\right) \hat{R}(k) \cos(wk).$$

In addition, $M = 2\hat{m}$, where $\hat{m}$ is the smallest value of m such that $| \hat{\rho}(\hat{m}+k) | < 2\sqrt{\frac{\log N}{N}}$ for k = 1…5 where

$$\hat{\rho}(k) = \frac{\hat{R}(k)}{\hat{R}(0)}.$$

US EPA ARCHIVE DOCUMENT

Input Data for Block Length Calculation

There are two schools of thought on the type of input data that should be used in calculating the optimal block length. One approach is to use the input concentration data (Measured) and its serial correlation structure to obtain the optimal block length. The other approach is to use the matrix obtained from multiplying composition matrix by the contribution matrix. This matrix is referred as the predicted concentration. The difference between the predicted and measured concentration data are the so-called residuals.

In EPA Unmix, the use of measured concentration is preferred over predicted concentrations for the following reasons. First, EPA Unmix differs conceptually from other models that use input data uncertainty information. The spirit of EPA Unmix is to use the data as provided by the user without any modification. Since, predicted data can be viewed as reconstructed data, use of predicted data in the bootstrapping algorithm violates this spirit. Secondly, the choice of measured concentration versus predicted concentration is similar to the choice of evaluating the algorithm's sensitivity versus the solution's sensitivity. The use of the predicted concentrations can underestimate uncertainties since the input data (predicted concentrations) to the bootstrap procedure tends to be less noisy when compared to the measured concentration. Bootstrap variates constructed from predicted concentrations can have qualitatively different underlying noise structure compared to the bootstrap variate constructed from measured concentrations.

Block Length Calculation Evaluation

In the experiments conducted using different sized data sets and criteria, the spectral estimation appeared to be stable. The variance estimation method failed with noisier data sets. The block lengths suggested by the variance estimate method were sometimes larger than the number of observations in the data set. Real data challenges algorithms since the noise found in real data is colored and not white in nature and hence does not cancel out intrinsically. The spectral estimation method fared better with noisy data sets compared to the variance estimation method.

This conclusion was reached using known information about the particular data sets such as the data collection interval, etc. In addition, using the block length suggested by the spectral estimation method reduced the number of total attempts to get 100 feasible solutions. That is, more quality variates were generated using the block length suggested by the spectral estimation method. The source apportionment model tends to reject variates by stating that no feasible solution can be found when it fails to define the space in which a possible solution can exist.

US EPA ARCHIVE DOCUMENT

Nevertheless, this experiment cannot guarantee success of one method over the other for all types of data sets. In this, the evaluated algorithm shares the following critical characteristic of any scientific approach: There is a possibility of the existence of a real data set that can stump the model algorithm and the model remains relevant only as long as the algorithm can be tweaked to meet the new challenges.

Input Bootstrap Variates Analysis

The basic aim of the bootstrap method is to run the model a number of times using altered versions of the base data set called the bootstrapped variates. Bootstrapped variates are created by sampling the base data set with replacements. By feeding the model with bootstrapped variates and by using the resulting variations in the source compositions, the stability of the base data sources can be confirmed. An issue of vital importance is the nature of bootstrapped variates. Currently, all sets generated using the abovementioned sampling scheme are considered valid. However, it is likely that the bootstrapped variate might not exhibit the underlying features of the base data set. An extreme case can result when a small percentage of the observations from the base data set are used to generate the bootstrapped variate. Other cases might include the more frequent occurrence of known outliers. Although, outliers are part and parcel of any realistic data, excessive occurrence can substantially change the nature of the data set resulting in sources unrelated to the base data sources and therefore not a valid variation of the base data set. Using this and other data sets differing qualitatively from the base data set (a.k.a. rogue data sets) is akin to introducing a new data set in the middle of a bootstrap run.

The other extreme is when the bootstrap variates appears to be only a slight variation of the base data set. This leads to a different problem of not adequately testing the solution. These variates have to be ignored as well.

Thus, it is imperative that bootstrap variates be analyzed and confirmed to have the similar but measurably distinct underlying features as found in the base data set *prior* to feeding the bootstrap variate to the model for evaluation.

A possible argument against such analyzing of bootstrap variates is that this process tends to underestimate uncertainties. This argument is vacuous on two counts. First, one needs to know the unknown (Variability) to suggest that the computed result is an underestimation (of the unknown). It is impossible to guess even the range of the uncertainties due to the mathematics used in the model and the sampling process to suggest possible sources. Secondly, this process ensures that the uncertainties are more believable since only quality variates are fed into the model, and the resulting solution is evaluated impartially. Thus, analyzing the input bootstrap variate can only help produce a more justifiable picture of the Variability associated with a chosen solution.

US EPA ARCHIVE DOCUMENT

In the experiments that have been conducted, the Variability range (2.5th percentile and 97.5th percentile) obtained from the analyzing input variates have both decreased and increased when compared to the results obtained without any analyzing process. Thus, the argument that the analyzing process tends to underestimate Variability is again found to be without merit.

An alternative to this process is the post-screening of the resulting solutions. The bootstrap variates are used as generated by random sampling of the base data observations with replacements, but the resulting bootstrap source compositions are screened for its proximity to base source compositions. Such selective use of bootstrap run results may violate the principles of the bootstrap approach of analyzing base solutions and therefore lead to a less believable Variability picture.

Analysis Method of Bootstrap Variates

Using a random number generator, a number ranging between 1 and the total number of observations in the base data set is generated. The block of observations of the size suggested by spectral estimation (as the optimal block length) starting from that number is chosen for the bootstrap variate. Continue this process of generating a random number and stringing together the block of observation from the base data set until the size of the bootstrap variate equals the size of the base data set. The block length guidance is ignored if there are not enough observations available from the current starting point. That is, if N is the number of observations and K is the block length, then any random number R greater than N-K will have only N-R+1 observations available.

Use this bootstrap variate to be compared to the base data set using the Generalized Singular Value Decomposition (GSVD) method. Let B be the base data set and B_n be the nth bootstrap variate. Then, by using the GSVD method, B and B_n can written as

$$B = U*C*X'$$
$$B_n = V*S*X'.$$

In the above expression, U and V are unitary matrices, X, a (usually) square matrix, and nonnegative diagonal matrices C and S such that

$$C'*C + S'*S = I,$$

where I is the identity matrix. The matrices C and S can be viewed as the cosine and sine decomposition matrices. Hence, the angular distance between the data sets is

US EPA ARCHIVE DOCUMENT

$$\theta_m = \operatorname{arc\,cot}\left(\frac{c_m}{s_m}\right) - \frac{\pi}{4},$$

where c_m and s_m are the m^{th} entries in the C and S diagonal matrices and θ_m indicates the relative closeness of the m^{th} common "factor" between the two data sets.

The closer the angles are to zero indicates positive alignment or similar nature of the underlying structure of the base data set and bootstrap variate while angles that tend toward $\frac{\pi}{4}$ shows almost no alignment of the underlying structure of the data sets. In this model, bootstrap variates whose ***maximum*** angular deviation from base data set structure does not exceed $\frac{\pi}{50}$ are considered too similar to the base data set and are rejected. Also, those bootstrap variates whose ***minimum*** angular deviation from the base data set exceeds $\frac{\pi}{8}$ are considered too dissimilar and are also rejected. This requirement allows a generous amount of variation on the base data set with 5% to 50% variation on the underlying structure of the base data set. Thus, $\left[\frac{\pi}{40}\ \frac{\pi}{8}\right]$ is the appropriate angle interval so that noise is a factor and ensures that the bootstrap variates are more than being just minor variations of the base data set while retaining the underlying structural features of the base data set.

Base Data Set Classifications

The underlying nature of the base data set dictates the nature of the bootstrap variates in terms of their proximity to the base data set. For instance, if a base data set has localized high values, good bootstrap variates, as described in the earlier section, may be harder to produce. In such cases, the base data is adjudged to be relatively rigid in its structure. Therefore, the base data set can be classified into three categories based on their structural rigidity of "High", "Medium" and "Low". If the base data set values are not localized, then a good bootstrap variate may be produced with high probability. Such data sets are then termed to have "Low" rigidity associated with its structure. We quantify the classification using the following method.

All angular data associated with a bootstrap variate of all rejected bootstrap variates are collected. If the median value falls with the interval $\left[\frac{\pi}{8}\ \frac{\pi}{6}\right]$, then the base data set is classified as "Low". If the median falls within the interval $\left[\frac{\pi}{6}\ \frac{\pi}{5}\right]$,

US EPA ARCHIVE DOCUMENT

then the base data is classified as “Medium”. Otherwise, the base data is classified as “High.”

All data sets have high and low values. The angle intervals are used to quantify the possible significance of a small set of observations while generating bootstrap variates. If a small set of values have disproportionate effect on the generation of bootstrap variates, then the base data set is deemed to be either in the “Medium” or “High” category in terms of its rigidity in its underlying structure.

Suspension of Bootstrap Variates Analysis

Also, in the interest of speeding up the bootstrap process, the model suspends analyzing the bootstrap variates under the following circumstance: If the number of rejected bootstrap variates does not exceed five(5) after the first twenty five(25) attempts, then any further analysis of bootstrap variates are suspended. The assumption behind this decision is that the initial low probability of obtaining a poor bootstrap variate implies the base data and bootstrap creation procedure are both robust and will produce high quality end results even without the analysis of the bootstrap variates.

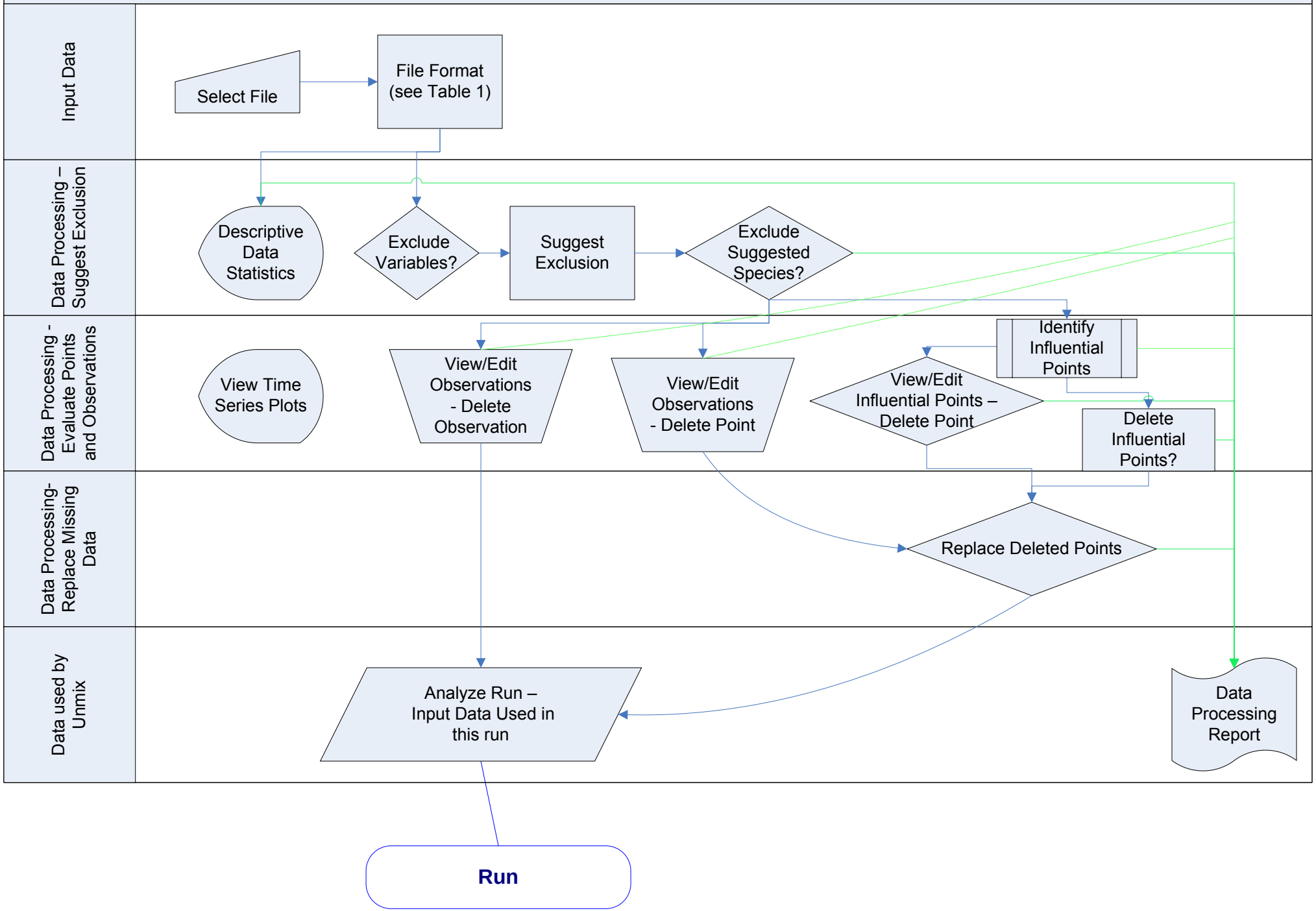
US EPA ARCHIVE DOCUMENT

APPENDIX D: PROCEDURE DIAGRAMS

US EPA ARCHIVE DOCUMENT

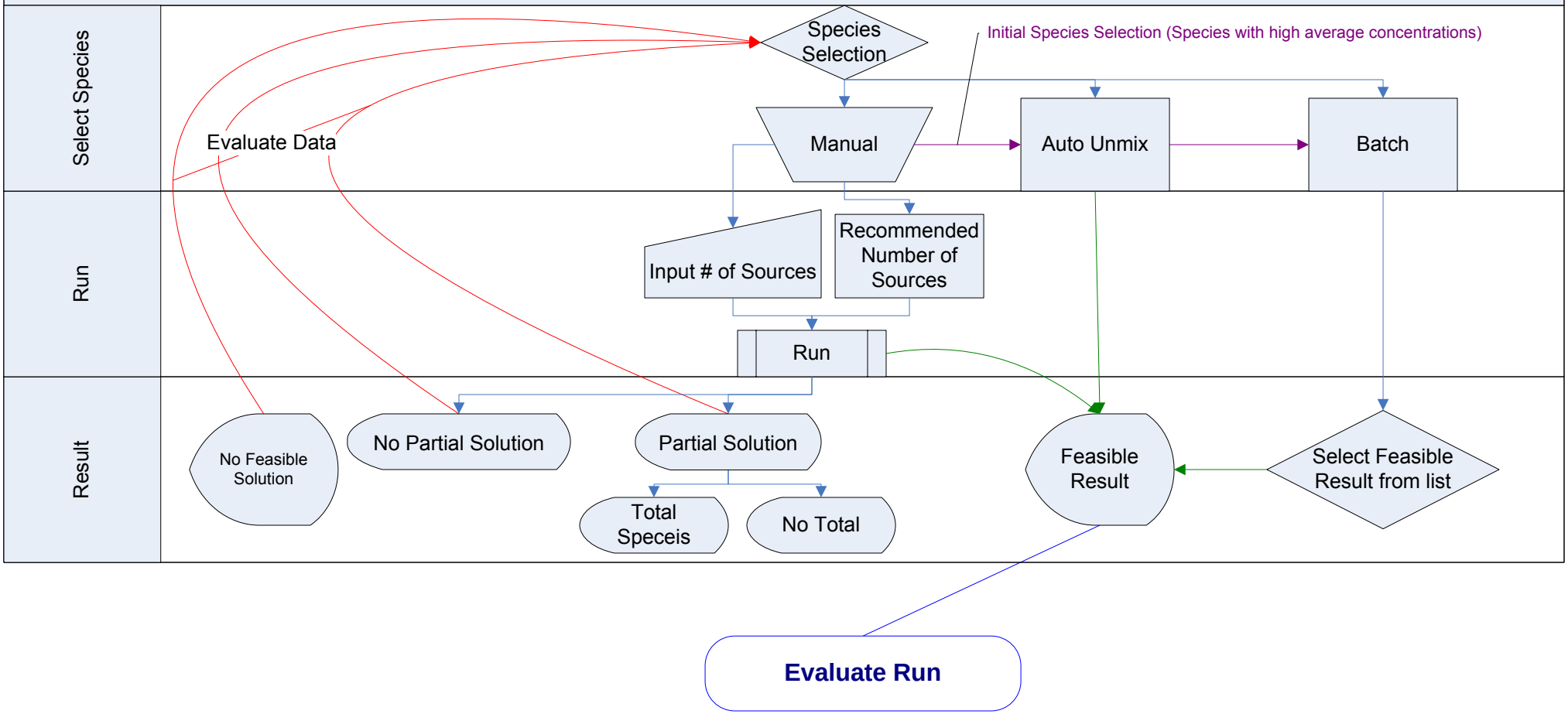
Appendix C – Unmix Procedure Diagrams

Data Processing

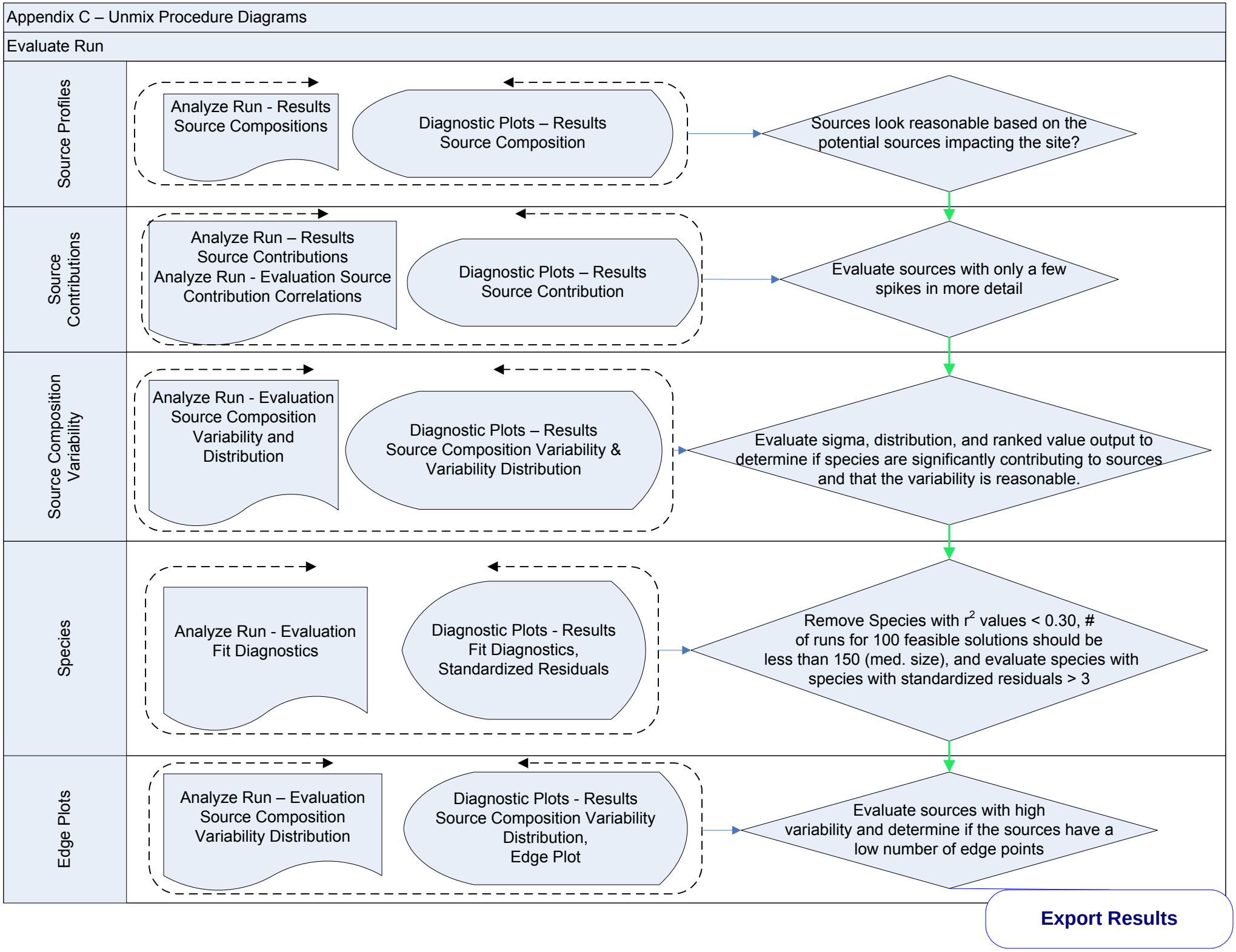


Appendix C – Unmix Procedure Diagrams

Run



Evaluate Run


Appendix C – Unmix Procedure Diagrams
Evaluate Run
Source Profiles
Analyze Run - Results Source Compositions
Diagnostic Plots – Results Source Composition
Sources look reasonable based on the potential sources impacting the site?
Source Contributions
Analyze Run – Results Source Contributions
Analyze Run - Evaluation Source Contribution Correlations
Diagnostic Plots – Results Source Contribution
Evaluate sources with only a few spikes in more detail
Source Composition Variability
Analyze Run - Evaluation Source Composition Variability and Distribution
Diagnostic Plots – Results Source Composition Variability & Variability Distribution
Evaluate sigma, distribution, and ranked value output to determine if species are significantly contributing to sources and that the variability is reasonable.
Species
Analyze Run - Evaluation Fit Diagnostics
Diagnostic Plots - Results Fit Diagnostics, Standardized Residuals
Remove Species with r2 values < 0.30, # of runs for 100 feasible solutions should be less than 150 (med. size), and evaluate species with species with standardized residuals > 3
Edge Plots
Analyze Run – Evaluation Source Composition Variability Distribution
Diagnostic Plots - Results Source Composition Variability Distribution, Edge Plot
Evaluate sources with high variability and determine if the sources have a low number of edge points
Export Results

US EPA ARCHIVE DOCUMENT

United States
Environmental Protection
Agency
Office of Research
and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
$300

EPA/600/R-07/089
June 2007
www.epa.gov

Please make all necessary changes on the below label, detach or copy, and return to the address in the upper left-hand corner.

If you do not wish to receive these reports CHECK HERE

☐; detach, or copy this cover, and return to the address in the upper left-hand corner.

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT No. G-35

US EPA ARCHIVE DOCUMENT

Recycled/Recyclable
Printed with vegetable-based ink on paper that contains a minimum of 50% post-consumer fiber content processed chlorine free