UNMIX Version 2 Manual

Prepared for the

United States Environmental Protection Agency

March, 2000

By

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

INTRODUCTION **3**

INSTALLATION **3**

HARDWARE AND SOFTWARE REQUIREMENTS 3
INSTALLING UNMIX 3
STARTING UNMIX AND TESTING THE INSTALLATION 4

BASIC OPERATIONS **5**

INPUT DATA 5
SELECT SPECIES FOR THE MODEL 6
ANALYZE THE SELECTION 7
RUN UNMIX 9
ESTIMATE ERRORS IN THE SOURCE COMPOSITIONS 10
OUTPUT 11
TROUBLESHOOTING 11

ADVANCED OPERATIONS **13**

DIAGNOSTIC PLOTS 13
UTILITIES AND OPTIONS 13

SELECTING SPECIES TO INCLUDE IN THE MODEL AND PARTICULATE DATA EXAMPLE **16**

INTRODUCTION 16
HINTS FOR SELECTING SPECIES 16
EXAMPLES OF EDGE PLOTS 20
MISSING DATA 23

REFERENCES **23**

ACKNOWLEDGEMENTS **23**

Introduction

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 assumes that the data and the compositions and contributions of the sources are all strictly positive (zero data values are not allowed). UNMIX further assumes that for each source there are some samples that contain little or no contribution from that source. For a given selection of species, UNMIX estimates the number of sources, the source compositions, and source contributions to each sample.

A word about the underlying philosophy of UNMIX. The goal is to let the data speak for itself. Thus, some have called this approach self-modeling. 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, more constraints are needed (Henry, 1987). Under certain rather mild conditions, the data itself 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. No solution is better, if not more satisfying, than an unreliable one.

This is the user's manual for a beta test version of UNMIX, multivariate receptor modeling and analysis software. The software and manual are provided 'as is' for evaluation purposes only, no warrantee is expressed or implied. All the distributed files and this manual are copyright 1998, 1999 by Ronald C. Henry. None may be altered, copied, or distributed without written permission

Installation

Hardware and Software Requirements

UNMIX runs under Matlab, a high level language for numerical and graphical analysis of data. Matlab is a registered trademark of the Mathworks, Inc. For all but one feature of UNMIX, only the basic Matlab package is needed. However, for one operation described below, the Optimization Toolbox is required, and must be purchased separately form the Mathworks, Inc. More information on Matlab and obtaining it may be found at http://www.mathworks.com/. UNMIX has been tested on a PC with Matlab version 5.3 and should work with all higher versions whether running on a PC, or UNIX machine. UNMIX will not run on Macs.

UNMIX is numerically intensive. Computers with processors that do not have good numerical processing units, e.g. Intel's Celeron processor, are not recommended. A clock speed of 400 MHz and 128 Meg of RAM are recommended but not required. UNMIX assumes a screen display of at least SVGA (800 x 600) resolution.

Installing UNMIX

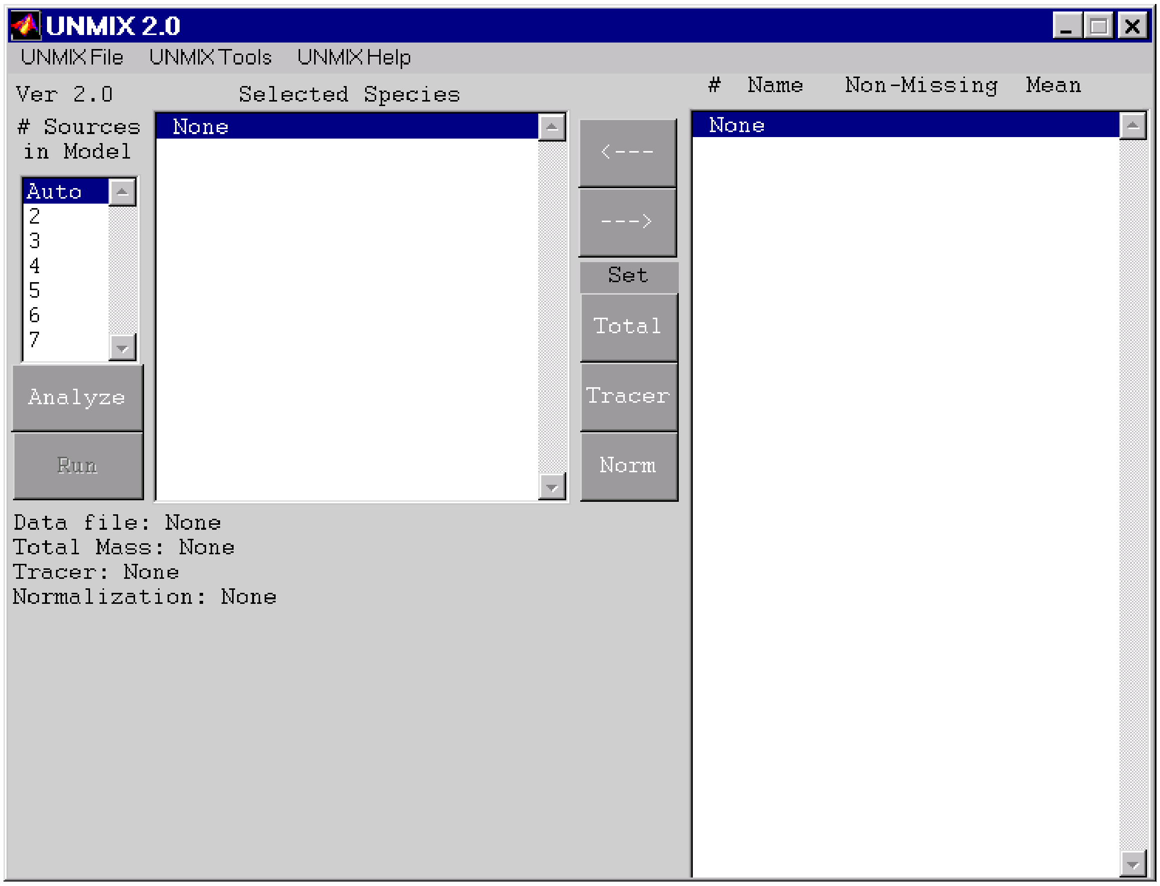
The UNMIX 2 package consists of the following files:
unmix2.p – the main UNMIX Matlab program,
UMgui.p – the Matlab program that draws the Graphical User Interface (GUI),
UMgui.mat - The Matlab workspace that contains the data needed to draw the GUI,
umgoalfom.p - a Matlab program needed by unmix2.m,
umplotbd.p – a Matlab program needed by unmix2.m,

umfomcalc.p– a Matlab program needed by unmix2.m,
UMtotalmin.p - a Matlab program needed by unmix2.m,
umtest.txt – a set of simulated VOC composition data,
umpmdata.txt – a set of real particulate composition data, and
UManual.pdf (this file).

These files must be copied to a directory listed in the Matlab path, or the directory that contains them must be added to the Matlab path. To find the Matlab path, type path at the command prompt in the Matlab command window. It is recommended, but not at all necessary, that the UNMIX files be copied to its own directory, e.g., 'C:\unmix' on PC systems. Use the addpath command to add this directory to the path, e.g., addpath('C:\unmix'). On Macintosh and PC systems Matlab's pathtool may be use to add the UNMIX directory to the path.

Starting UNMIX and Testing the Installation

Start Matlab 5.2 or higher; UNMIX will not run properly on lower versions. In the Matlab Command window type 'unmix2', this opens the Graphical User Interface. It should look similar to Figure 1 below. Most UNMIX output, except graphs, appears in the Matlab command window. Run through the examples in the next section to ensure UNMIX is working properly.



Selected Species: None

# Name Non-Missing Mean: None

```
Data file: None
Total Mass: None
Tracer: None
Normalization: None
```

Figure 1. UNMIX main window at startup.

Basic Operations

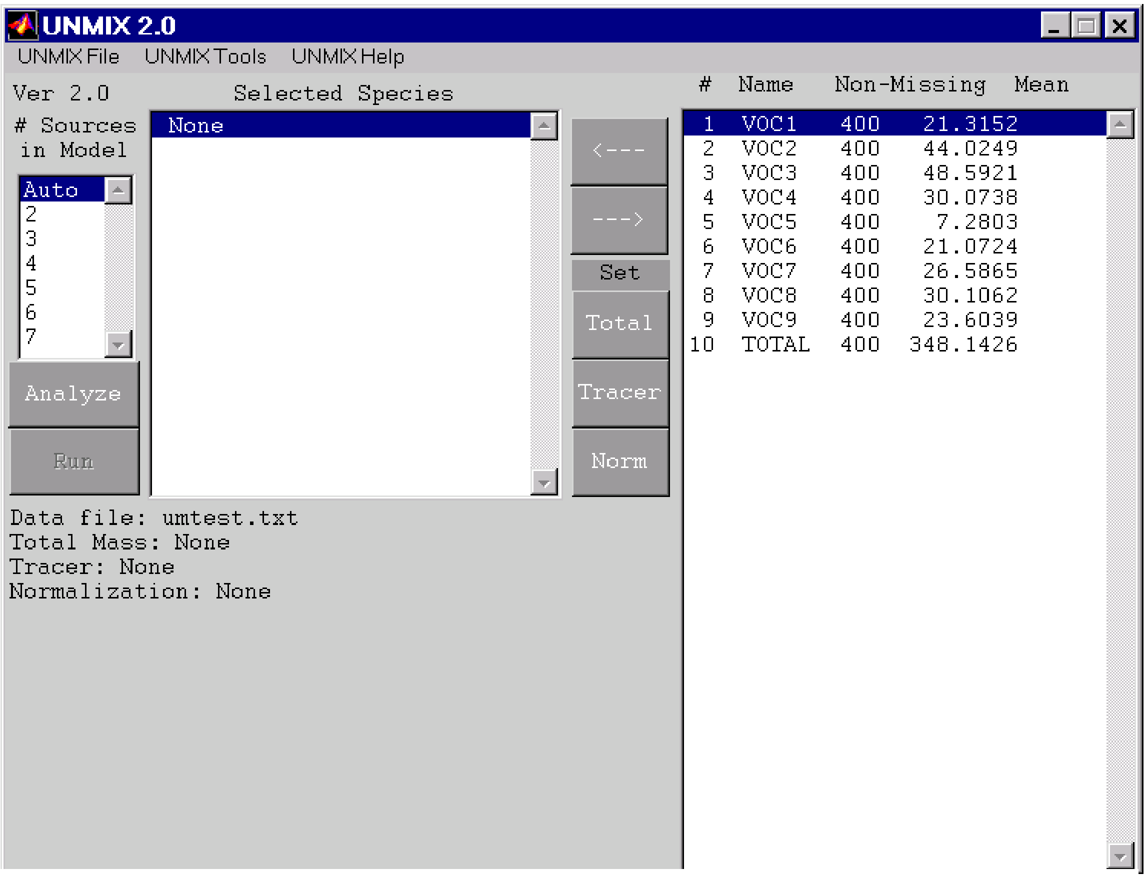
This section walks through the sequence of operations that will usually be followed to produce a receptor model of the data consisting of the source compositions and source contributions that reproduce the data. The sequence of operations is:

1. read the data using the **Input Data** command from the **UNMIX File** menu ;
2. select the species to be included in the model by highlighting species in the right-hand side box, and clicking the left arrow button;
3. set total mass, tracer, and normalization species;
4. click the **Analyze** button to determine the number of sources that may be included in the model;
5. select the number of sources to included in the model from the box above the **Analyze** button;
6. run UNMIX by clicking the **Run** button; and
7. estimate the errors in the source compositions .

The section concludes with how to output the solutions to the command window, file or plot, and some basic troubleshooting.

Input Data

UNMIX requires that data files be simple text files in the format described below. A data file is read by selecting **Input Data** from the **UNMIX File** menu. Do this now and input the file umtest.txt, which contains simulated data from three sources with 15% error (1 sigma). You will be asked to give the missing value code, in this case use the default value of –99. The UNMIX command window should look like Figure 2 after reading the data.



Selected Species: None

```
#  Name   Non-Missing  Mean
1  VOC1   400    21.3152
2  VOC2   400    44.0249
3  VOC3   400    48.5921
4  VOC4   400    30.0738
5  VOC5   400     7.2803
6  VOC6   400    21.0724
7  VOC7   400    26.5865
8  VOC8   400    30.1062
9  VOC9   400    23.6039
10 TOTAL  400   348.1426
```

```
Data file: umtest.txt
Total Mass: None
Tracer: None
Normalization: None
```

Figure 2. UNMIX main window after loading the data in umtest.txt.

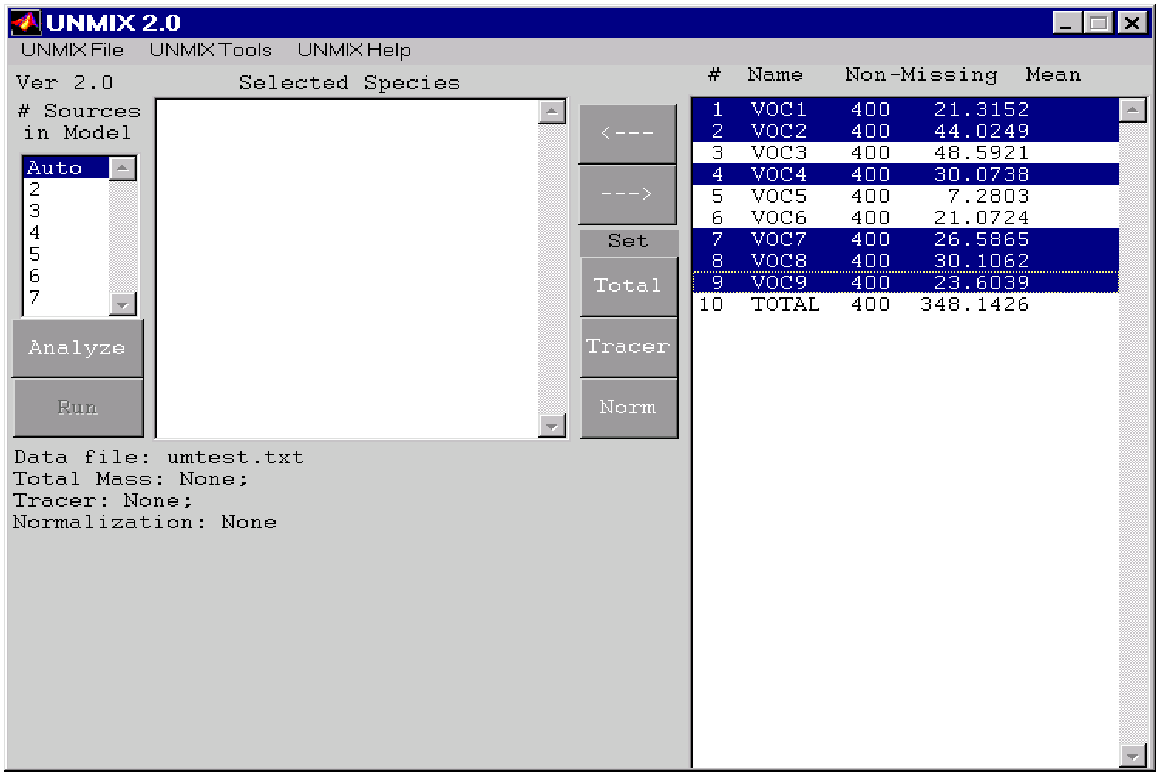
The right-hand box contains the species number, name, the number of non-missing values, and the average value of all non-missing values. Also notice that the name of the data file appears below the left-hand box.

Data File Format

UNMIX data files are flat ASCII files with spaces or tabs separating the numbers. The first row must contain the column labels. These labels may be any lengths, however imbedded spaces are not allowed, e.g. the label SPECIES 1 is not allowed, but SPECIES_1 is acceptable. After the first row, all the columns must contain only numbers and each row must contain the same number of columns. Missing data must be given a numerical value such as -999 or 0. Only one missing value code is allowed, if the data contains several different codes, or the codes are non-numeric all must be converted to one numerical missing value code. A spreadsheet is a convenient tool for this. From the spreadsheet, you should save the data file as a tab-separated text file. The files umtest.txt and unpmdata.txt are examples of UNMIX data files. Note that no special provision for using columns with dates or times is made in UNMIX, however, these are allowed as long as they have a valid label and no non-numeric characters such as Feb or AM. Columns containing month, day, and year will be read in like any other variable and these can be used in like any other variable in scatter plots that are made by UNMIX. Currently UNMIX makes no provision to treat certain columns of the data file as time variables for use in time series plots. However, as long as the date and/or time columns do not contain non-numeric data, there is no harm, and possibly some advantage, in leaving these columns in the data file.

Select Species for the Model

The most important choice that the UNMIX user makes is the selection of the species to be included in the model. This is done by highlighting the species in the right-hand side box and then moving these species to the left-hand side box by clicking on the button with the arrow pointing to the left. Species are removed from the model by highlighting in the right-hand side box and clicking on the button with the arrow pointing to the right. The following Figures 3 and 4 illustrate this process.

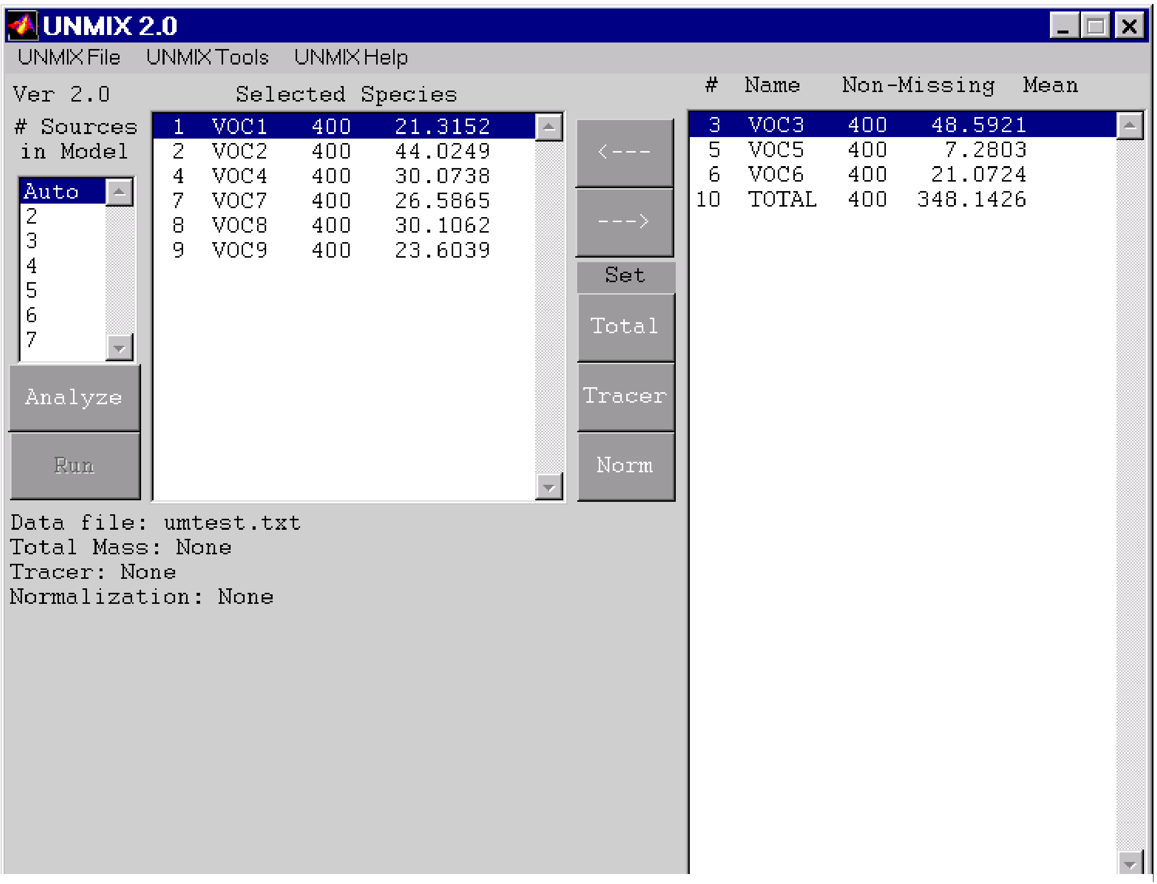


#	Name	Non-Missing	Mean
1	VOC1	400	21.3152
2	VOC2	400	44.0249
3	VOC3	400	48.5921
4	VOC4	400	30.0738
5	VOC5	400	7.2803
6	VOC6	400	21.0724
7	VOC7	400	26.5865
8	VOC8	400	30.1062
9	VOC9	400	23.6039
10	TOTAL	400	348.1426

Data file: umtest.txt
Total Mass: None;
Tracer: None;
Normalization: None

Figure 3. UNMIX main window showing selected species before pressing the ← button.

Clicking the← button gives the result in Figure 4.



```
Ver 2.0              Selected Species
# Sources   1  VOC1    400    21.3152
 in Model   2  VOC2    400    44.0249
            4  VOC4    400    30.0738
Auto        7  VOC7    400    26.5865
2           8  VOC8    400    30.1062
3           9  VOC9    400    23.6039
4
5
6
7

Data file: umtest.txt
Total Mass: None
Tracer: None
Normalization: None
```

```
#   Name    Non-Missing   Mean
3   VOC3    400      48.5921
5   VOC5    400       7.2803
6   VOC6    400      21.0724
10  TOTAL   400     348.1426
```

Figure 4. UNMIX main window as in Figure 3, but after pressing the ← button.

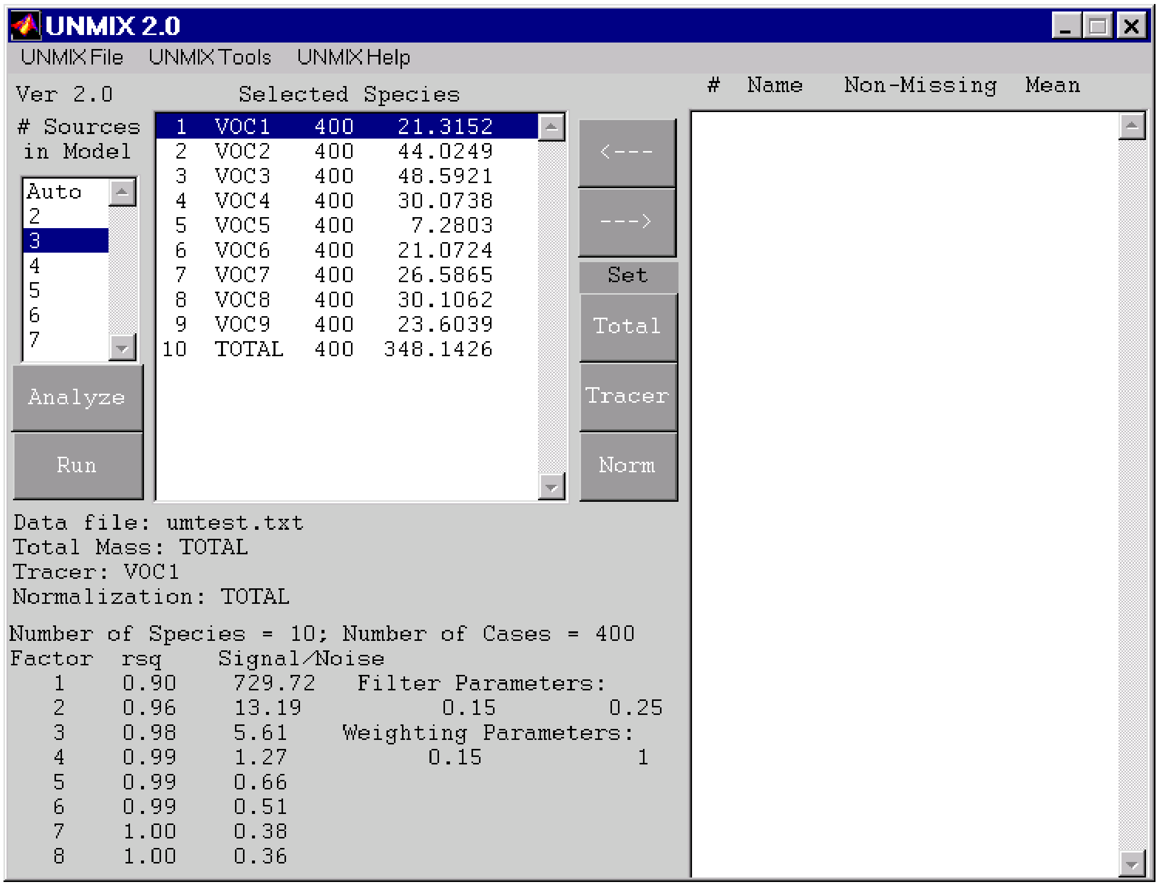
We want all the species in the model, so highlight and select the remaining species. The following results are for all species in the model.

Analyze the Selection

The next step is to analyze the selection of species that appears in the left-hand side selection box. But first, one can identify some species as special. 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 mass variable, a tracer species, i.e., one that has only one source, or the variable used to normalize the source compositions. Usually the normalization and total mass species are the same, as this gives a source composition as a mass fraction. Once a species is selected as a total mass, tracer, or normalization species, it can be deselected by highlighting the species in the left-hand box and pressing the same button again. Thus, if VOC1 is set as a tracer and no tracer is desired, then it can be deselected by highlighting it and pressing tracer.

In this case we choose VOC1 as a tracer and TOTAL as the total mass and normalization variable. Note the box in the upper right labeled # of Sources. If Auto is selected when the **Analyze** button is clicked, then UNMIX will determine the possible number of sources in the data and highlight these in the # of Sources box. Having thus chosen, we now click the **Analyze** button and the result is shown in Figure 5.

We see that a lot of information has appeared. Under the data file name, the names of the total mass, tracer, and normalization variables are given. Note that UNMIX does not require any of these to be identified, but it often helps if they are.



```
Selected Species
 1  VOC1    400    21.3152
 2  VOC2    400    44.0249
 3  VOC3    400    48.5921
 4  VOC4    400    30.0738
 5  VOC5    400     7.2803
 6  VOC6    400    21.0724
 7  VOC7    400    26.5865
 8  VOC8    400    30.1062
 9  VOC9    400    23.6039
10  TOTAL   400   348.1426
```

```
Data file: umtest.txt
Total Mass: TOTAL
Tracer: VOC1
Normalization: TOTAL

Number of Species = 10; Number of Cases = 400
Factor   rsq     Signal/Noise
   1     0.90     729.72   Filter Parameters:
   2     0.96     13.19          0.15         0.25
   3     0.98     5.61     Weighting Parameters:
   4     0.99     1.27          0.15             1
   5     0.99     0.66
   6     0.99     0.51
   7     1.00     0.38
   8     1.00     0.36
```

Figure 5. UNMIX main window showing the result of analyzing the selected species.

Finding the number of sources

The bottom of the window tells us that there are 10 species selected and that there are 400 cases with all-non-missing data (in this case there is no missing data). Under this is seen the r-squared value and the signal-to-noise ratio for each of the factors in the data. In the example above, the r-squared value for a model with three sources is 0.98. This means that at least 98% of the variance of each species can be explained by three sources. 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 number of sources (or factors) that will be used in the model is the number of factors with signal-to-noise ratios greater than 2.0 and r-squared values greater than 0.8.

If **Auto** is selected in the # of sources box when **Analyze** is pushed, then all the factors which meet the criteria above will be highlighted in the # of sources box. UNMIX will not automatically select two sources since this case is seldom of interest. For the current example, there is only one possibility, three sources, since only with three factors do we have r-squared greater than 0.8 and the signal-to-noise greater than 2.0. If 4, 5, 6, or 7 factor models met these criteria, all these would be highlighted in the box. Generally, if **Auto** selects several possible numbers of sources, one would run UNMIX only for the largest number of sources that meet the criteria. If this does not give a feasible solution than one should try the next largest number of sources. Alternatively, if **Auto** gives several possible number of sources, one can just run UNMIX and UNMIX will attempt to find a model with all the numbers of sources highlighted in the box. Of course, the user can override the automatic selection of the number of sources and highlight whatever selection of number of sources they wish.

A note on the meaning of a one factor model. In the example above, the one source model has an r-squared of 0.9 and a signal to noise ratio of over 700. This means that a reasonable model could consist of a single source whose composition is given by the average value of the species divided by the average total mass. Since it is an average value, the single source model will almost always have a high signal to noise ratio. However, usually, as in the case above, one is interested in finding more than one source.

The number of sources is limited to a maximum of seven. It is highly unlikely that more than seven sources can be reliably estimated from a real data set. For if there are 8 sources, then the largest contribution that can be made by the smallest source is 100/8 or 12.8 percent of the total mass. In most cases where two or three sources alone contribute 50 percent of the total mass, the largest minimum contribution is only 50/(8-2) to 50/(8-3) or 8.3 to 10 percent, and very probably less than this. Errors in measurements, and perhaps more importantly, even rather small variations in the source compositions make it almost impossible to reliably determine a source that contributes less than 10 percent of the total mass. There are exceptions, of course, such as the case where a small source has a tracer that has little measurement error and with a stable composition. However, in the interest of realistic source apportionment, UNMIX allows no more than seven sources.

The Filter and Weighting Parameters given are for advanced users and are described later, the beginner should use the default values given. The final steps are to run UNMIX and estimate the errors in the source composition.

Run UNMIX

The model is run by clicking the **Run** button. Notice that the **Run** is not active until the **Analyze** button has been clicked. If the **Run** button is ever gray-out and not active, just click on **Analyze** to activate it. After clicking on **Run,** the model output appears in the Matlab command window. This is:

```
Number of Sources =3
28-Feb-2000 12:29:02
File: umtest.txt
Tracer: VOC1
Total Mass:
Normalization: TOTAL
Filter Parameters: 0.15        0.25
Weighting Parameters: 0.15           1
10 Variables, 400 Cases, 3 Factors,
Min Rsq =  0.98;  Min Sig/Noise= 5.80; Strength = 4.53
UNMIX2 Source Composition Solution #1
        Source 1    Source 2    Source 3
VOC1     0.10523     0.00000    -0.00000
VOC2     0.11301     0.14690     0.14430
VOC3     0.11425     0.10226     0.22017
VOC4     0.12166     0.06648     0.02601
VOC5     0.01907     0.04383     0.01439
VOC6     0.04813     0.13879     0.04816
VOC7     0.10119     0.06381     0.03273
VOC8     0.12798     0.04610     0.02344
VOC9     0.08836     0.01060     0.05255
TOTAL  201.96129    47.56278    96.29586
```

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, i.e., the r-squared for any species is greater than or equal to this value. The Min Rsq is recommended to be 0.8 or greater. The Min Sig/Noise is the smallest estimated signal-to-noise ratio of any of the factors included in the model. Min Sig/Noise is recommended to be 2.0 or greater. Finally, the strength number represents the overall confidence in the model. Its minimum value is 1, and it should be as large as possible. Values greater than 3 are ok, but values greater than 10 are better.

Since TOTAL is the normalization variable, the source compositions are given as a mass fractions, and the values under TOTAL are the average amount of the TOTAL apportioned to each source. **If a normalization species is not set**, then the source compositions are normalized to give the average amount of each species associated with each source, as shown below. In this case the source contributions are normalized to have a mean value of 1.

```
File: umtest.txt
Tracer: VOC1
Total Mass:
Normalization:
Filter Parameters: 0.15        0.25
Weighting Parameters: 0.15          1
10 Variables, 400 Cases, 3 Factors,
Min Rsq =  0.98;  Min Sig/Noise= 5.54; Strength = 4.47
UNMIX2 Source Composition Solution #1
        Source 1    Source 2    Source 3
VOC1    21.25229     0.00000    -0.00000
VOC2    22.82377     6.98693    13.89582
VOC3    23.07375     4.86401    21.20194
VOC4    24.57025     3.16203     2.50441
VOC5     3.85109     2.08450     1.38574
VOC6     9.71940     6.60105     4.63764
VOC7    20.43699     3.03502     3.15143
VOC8    25.84681     2.19275     2.25746
VOC9    17.84552     0.50394     5.06062
TOTAL  201.96129    47.56278    96.29586
```

For the record, the "true" source compositions (mass fraction) and apportionment are given by:

	Source1	Source2	Source3
VOC1	0.1067	0.0000	0.0000
VOC2	0.1135	0.1348	0.1500
VOC3	0.1148	0.0752	0.2160
VOC4	0.1234	0.0770	0.0172
VOC5	0.0187	0.0476	0.0124
VOC6	0.0500	0.1471	0.0409
VOC7	0.1026	0.0736	0.0237
VOC8	0.1294	0.0397	0.0208
VOC9	0.0938	0.0135	0.0421
TOTAL	200.00	48.340	99.990

Estimate Errors in the Source Compositions

Since there was 8 % (1 sigma) error added to the data the results are not expected to be exactly the same as the true values. We need to estimate the errors in the UNMIX source composition to find out if it is indeed acceptably close to the true values.

One sigma errors in the source compositions are estimated by going to the **Estimate Errors** selection under the **UNMIX Tools** menu. The results are:

```
UNMIX2 Solution # 1 Source Composition Errors
VOC1    0.0064     0.0000     0.0000
VOC2    0.0038     0.0162     0.0072
VOC3    0.0087     0.0217     0.0129
VOC4    0.0071     0.0128     0.0066
VOC5    0.0009     0.0047     0.0014
VOC6    0.0028     0.0145     0.0050
VOC7    0.0048     0.0078     0.0049
VOC8    0.0070     0.0093     0.0046
VOC9    0.0044     0.0079     0.0074
TOTAL  15.9014     8.7244    11.4173
```

Comparing these errors with the difference of the true and UNMIX estimate shows that all UNMIX estimates are within 2 sigma of the true values.

Output

This section describes how to display and save the source compositions, error estimates, and the source contributions found by UNMIX.

Command Window and File Output

The current source compositions and the errors in these can be redisplayed in the Matlab command window by selecting **Output Data** from the **UNMIX File** menu. Both can also be saved to a file. When saving to a file, be aware of the following features. If you opt to save to an existing file, UNMIX will append the new results to the contents of the existing file. The file will not be overwritten. This allows one to keep a running record of results that are worthy of saving. This feature cannot be turned off. If you want the file to contain only the data currently saved, then you must give a new name to the save file. Finally, the information saved to the file is echoed to the command window. This cannot be disabled at this time.

Plot Output

Source contributions are usually too numerous to be simply sent to the Matlab command window. Instead, **Output Data** allows one to choose to plot the source contributions or to write them to a file. If the normalization variable is set to total mass, then the source contributions are in the same units as total mass. If the normalization variable is set to ‘none’, then the source contributions are normalized to a mean of one.

Troubleshooting

Reading Files

If the input file is not of the specified format, UNMIX will not be able to read it. If there are problems reading the data, Check to see that

1. the label names in the first row do not contain spaces,
2. the missing value code is numeric,
3. the column delimiter is a space or a tab,
4. all data values are numeric, and
5. all rows have a full complement of species.

Spreadsheets are very useful for identifying these and other types of problems with data files.

Matlab Command Window Errors

Sometimes UNMIX may respond to an operation with a warning or an error message displayed in the Matlab command window. A common warning is

```
Warning: Matrix is close to singular or badly scaled.
```

Often these warning are benign and can be ignored. If actual error messages occur, not just warnings, the best solution is to click the **Analyze** button, set the number of sources, and click the **Run** button. Sometimes one has to go through this cycle twice to resolve the problem. If this does not resolve the problem and if one has deleted data from the diagnostic plots, it is a good idea to restore the original data using the command from the menu **UNMIX Tools → Utilities → Restore Original Data**. Then go through the **Analyze,** set number of sources, and **Run** cycle again. This should resolve most problems.

Advanced Operations

The next sections describe tools, utilities, and diagnostics provided to assist in developing the best possible UNMIX model of the data. Examples of the use of all these commands are found in the Particulate Data Example section at the end of this manual.

Diagnostic Plots

UNMIX provides two types of diagnostic plots. The first produces scatterplots of the raw data. The second is a visualization of the edges in the data that form the additional constraints that are at the heart of UNMIX.

Raw Data Plots

Menu selection **UNMIX Tools → Diagnostic Plots → Selected Species vs Tracer** plots all the species in the selected species box versus the species set as a tracer. One purpose in this is to identify possible outliers and other problem data points. Points in the scatter plots are selected by using the mouse to draw a rectangle around them while holding down the left mouse button. Selected are identified by a magenta circle about the point (To deselect points, make a rectangle with no points in it.) Data points so selected can be deleted using the **UNMIX Tools → Delete Points** menu selection on the figure menu. The row numbers of the selected points are displayed in the Matlab command window. The most recently deleted points can be restored by **UNMIX Tools → Restore Points** menu selection in the figure window. The lines in the plots showing edges in the data are redrawn after points are deleted or restored. Be patient, if there are many points, this may take sometime.

The final selection on the **UNMIX Tools** menu is **UNMIX Tools → Change Font**, which allows the user to change the plot font to make it more readable. Once the plot window has been closed the points are permanently deleted until the data are reloaded or the **Restore Original Data** command is given. UNMIX does not make changes to the user's data file. The user is responsible for permanent removal of rows from the data file.

An example of how to use the scatterplots is found in the Particulate Data Example section. The relationship of scatterplots to receptor modeling is discussed in Henry *et al.* (1994).

Edge Plots

Menu selection **UNMIX Tools → Diagnostic Plots → Edge Plots** allows the advanced user to examine the edges in the data that define the additional constraints used by UNMIX. An example is given in the Particulate Data Example section. Edges in the data are the source of the additional constraints used by UNMIX to find a unique multivariate receptor model that fits the data (Henry, 1997). If there are three sources, the edges are indeed the edges of a triangle that contains the data in a special plot, thus the name edge plots. For more sources, i.e., in higher dimensions, the edges are properly called hyperplanes. Without getting too technical, the edge plots are a way of visualizing the "edges" in the data, even in higher dimensions. A source is associated with the edge for which that source's contributions are small or zero. The existence of such an edge is one of the primary assumptions of UNMIX.

The edge plot command asks first for a base source. One should pick a large source that is well determined. Next, UNMIX plots of the distance of all data points from the base source versus the distance of the data points from the edge defined by each source. The edges are the x and y axes of the plots. One is looking for edges that are ragged, or that are the result of just a few points. Such edges are indicative of outliers in the data or other problems. This command is most useful for those expert in the theory of UNMIX and experienced in it application.

Utilities and Options

These are special purpose UNMIX tools.

Restoring Original Data

Menu selection **UNMIX Tools → Utilities → Restore Original Data** does just that, if one has used the diagnostic scatterplots to delete some data points.

Species Selection Tools

The following commands in the **UNMIX Tools → Utilities** menu may help identify species to add to the model. These tools are not foolproof and are not guaranteed to find the best possible selection of sources for the model.

Suggest species that may not change the number of sources

Menu selection **UNMIX Tools → Utilities → Add Species with Same Number of Sources** looks at all the species that have not been selected and finds those that can be added to an existing the model and will not likely increase the number of sources. The method is to regress each unselected species against the source contributions and take all those with r-squared values of greater than 0.85.

Suggest species to add that increase the number of sources

Menu selection **UNMIX Tools → Utilities → Suggest Species for More Sources** identifies those species that when added to the selection will require more sources to explain the data. The species in the left-hand box are required to be in the model, the user then highlights species in the right-hand box that are considered optional. The method is to add the optional species two at a time to the required species and run the number of factors algorithm repeatedly and report the species that most often give additional sources with signal to noise ratios greater than 3.

Removing Negative Values from the Source Composition

Menu selection **UNMIX Tools → Utilities → Remove Negatives from the Source Composition** does just that. Just specify the source composition when prompted and UNMIX will vary produce a model with all positive source compositions that is as close to the original as possible. This command requires an advanced function found in the Matlab Optimization Toolbox, and will not work if this is not installed. This command may be useful if one cannot find a feasible solution for a selection of species. By turning off all the source composition filtering, as descried below, UNMIX will output all possible source compositions, even those with unrealistic negative values. This command can then be applied to one or more of these to produce a feasible source composition.

UNMIX Overnight

Perhaps the easiest way to find the best selection of variables is simply to try a large number of possible combinations. The **UNMIX Tools → Utilities → UNMIX Overnight** allows one to specify two sets of species, ones that are required to be in the model and ones that are optional. UNMIX will then run through all possible combinations of the optional species plus the required species and automatically calculate all feasible models and put the output in a user-specified file. If there are N optional species, then there will be $2^N - 1$ combinations. If N = 7, there are 127 combinations and if each one takes an average of 5 minutes, 10 and a half hours will be needed, thus the name of the command. A feasible model is one that has more than two sources, r-squared greater than 0.8, and the signal to noise ratio for the number of sources is greater than 2.

Options

Menu selection **UNMIX Tools → Options** has three choices. The first is a simple way to disable and enable the audible signals that UNMIX . The two remaining choices are for the advanced user and are described below. Beginners should stick with the defaults.

Filtering UNMIX Output

Menu selection **UNMIX Tools → Options → Filter Source Composition Output** sets two parameters used to reject a source composition with negative values. Call the first parameter P1. A source composition is rejected if the sum of the negative values of any species is greater than the fraction P1 of the average value of the species. The default value is 0.15, i.e., if the sum of the negative values for any species in a source composition is greater than 15 percent of its mean, the source composition is rejected and not displayed by UNMIX. Thus, if P1 is set to 1, nothing will be rejected.

The second parameter P2 is active only when a total mass variable is set. In this case, P2 is the maximum fraction of the total mass of the species that is due the negative values. The default value is 0.2, which means that up to 20 percent of the mass of a source may be negative values. Since P2 is only active if a total mass species is set, it has no effect if a total mass is not set.

When the a total mass species is set, UNMIX automatically rejects any species with negative total mass. To turn off this feature, do not set a total mass species.

Hints for filtering

1. For no filtering, that is, to see all solutions that UNMIX has found, no matter how many or how big the negative values, do not set a total mass species and set both filter parameters to 1.
2. To reject only solutions with negative total mass, but not other filtering, set the total mass species and set the filter parameters to 1.

These settings are good when taking a first, preliminary look at the data. If these settings lead to too many solutions, go back to the default values of the filter parameters.

Data Weighting Factors

Menu selection **UNMIX Tools → Options → Data Weighting Factors** allows the advanced user to set two parameters that define how very small and very large values in the data are treated. The beginner should use the default values. The first parameter P3 defines the lower fraction of the data that is to be partially discounted. The default value is 0.15, this means that the lower 15 percent of the data are weighted to have less influence on the UNMIX calculations. The second parameter P4 defines the upper bound. The default value is 1.0. If P4 were 0.9, say, then the upper 10 percent of the data would be weighted to have less influence on the UNMIX calculations.

Selecting Species to Include in the Model and Particulate Data Example

Introduction

The basic strength of UNMIX is that it relies on the data; the basic weakness of UNMIX is that it relies on the data. There are a number of problems that 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 many missing data points. This is dealt with in a later section. 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 Diagnostic Plots. 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. Problems caused by outliers and very small, noisy values may be overcome by data weighting. This approach is described in the Advanced Operations section above. 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. There may be no remedy for this difficulty other than not using the species in the model. Lewis *et al.* (1998) discusses some possible data problems and how to identify these.

Hints for Selecting Species

There is not one fixed method that will always lead to the best possible model of the data. Some guidelines to two approaches that may help are given in this section. This discussion will include examples using the data set umpmdata.txt supplied with the UNMIX package. This data set is a subset of actual particulate data obtained from the U. S. Environmental Protection Agency. It is used here for illustrative purposes only. To ensure this, the location where the samples were taken is not revealed and the units of data are not given.

One approach is the bottom-up method. First, by trial and error find a selection of 5 or 6 major species that give a 3-source model (or maybe a 4-source model) with a large minimum signal-to-noise ratio. It is good to start with species that have large average values as these will tend to have the fewest missing values and the least measurement noise. For example, load the umpmdata.txt file in to UNMIX and select the six species with the largest mean values: Mass, Si, S, Fe, OC, and EC, where OC is organic carbon and EC is elemental carbon. This gives a good 3-source model :

```
File: umpmdata.txt
Tracer:
Total Mass:
Normalization: MASS
Filter Parameters: 0.15        0.25
Weighting Parameters: 0.15           1
6 Variables, 839 Cases, 3 Factors,
Min Rsq =  0.94;  Min Sig/Noise= 13.72; Strength = 6.93
UNMIX2 Source Composition Solution #1
        Source 1    Source 2    Source 3
MASS   3214.30980  2818.38829  6062.71772
Si        0.01674     0.11166     0.00308
S         0.10871     0.01953     0.00748
Fe        0.00382     0.03918     0.01081
OC        0.35034     0.20326     0.47468
EC        0.03238     0.06962     0.16626
```

At this point, two tools are available to help the UNMIX modeler in selecting species to add to the model. The first identifies species that might be added to the model without increasing the number of sources. Use the menu selection **UNMIX Tools → Utilities → Add Species with Same Number of Sources**. In the example, the results are Ca, Al, and CO (Carbon Monoxide). We add all these to the model and find that there indeed are still only 3 sources. One might not want to include Al as it has more missing values than CO and Ca, and adds no new information over Si. The new UNMIX results are:

```
File: umpmdata.txt
Tracer:
Total Mass:
Normalization: MASS
Filter Parameters: 0.15        0.25
Weighting Parameters: 0.15           1
9 Variables, 602 Cases, 3 Factors,
Min Rsq =  0.93;  Min Sig/Noise= 19.48; Strength = 10.94
UNMIX2 Source Composition Solution #1
        Source 1    Source 2    Source 3
MASS   3632.88029  3231.93171  6077.83527
Al        0.00850     0.03899     0.00257
Si        0.02400     0.10296     0.00824
S         0.09242     0.01678     0.00873
Ca        0.00898     0.03759     0.00454
Fe        0.00736     0.03499     0.01333
OC        0.33924     0.23837     0.45876
EC        0.04185     0.07044     0.16947
CO        0.06958     0.05928     0.19347
```

Source 3 is high in CO (carbon monoxide) and thus associated with vehicles. Source 2 is high in Si and Al so is associated with soil dust. Finally, Source 1 is high in S and can be considered to be secondary particulate plus other sources.

Now, the problem is to add species that will increase the number of sources in the model. Use the menu selection **UNMIX Tools → Utilities → Suggest Species for More Sources**. In this case, UNMIX needs a list of species to consider. In the right-hand box we highlight those species with more than 544 non-missing values. Choose K, Mn, Zn, Br, Sr, Pb, and Sol_K (Soluble potassium) as possible species and run **Suggest Species for More Sources.** The suggestions are Sol_K, Br, Zn, Mn, and K. Start by adding the species with the fewest missing values first, which are Sol_K and K. The result is a good 4-source solution. After adding Zn, we get the following 5-source solution:

```
File: umpmdata.txt
Tracer:
Total Mass:
Normalization: MASS
Filter Parameters: 0.15        0.25
Weighting Parameters: 0.15           1
12 Variables, 602 Cases, 5 Factors,
Min Rsq =  0.95;  Min Sig/Noise= 5.24; Strength = 6.24
UNMIX2 Source Composition Solution #1
        Source 1    Source 2    Source 3    Source 4    Source 5
MASS   633.59661  1239.94800  2793.99242  5939.64566  2349.21882
Al      -0.00168     0.00679     0.01016     0.00294     0.05083
Si       0.00830     0.02185     0.02731     0.00818     0.13329
S        0.02800     0.02008     0.11215     0.00462     0.02518
K        0.00444     0.03650     0.00434     0.00431     0.01891
Ca       0.00826     0.01005     0.00953     0.00422     0.04786
Fe       0.02553     0.00267     0.00758     0.01424     0.04053
Zn       0.02244     0.00054     0.00001     0.00079     0.00037
OC       0.29877     0.45223     0.31212     0.46801     0.16943
EC       0.31720     0.09552     0.01200     0.16869     0.02293
Sol_K    0.00331     0.03352     0.00061     0.00319     0.00071
CO       0.10228     0.03228     0.07017     0.21715     0.01242
```

Most of the Zn is found in Source 1, however, this source explains only a small amount of the total mass. On this basis, Zn is removed and Br and Mn tried. This leads to the following source compositions:

```
File:umpmdata.txt
Tracer:
Total Mass:
Normalization: MASS
Filter Parameters: 0.15        0.25
Weighting Parameters: 0.15          1
13 Variables, 580 Cases, 6 Factors,
Min Rsq =  0.95;  Min Sig/Noise= 3.22; Strength = 3.74
UNMIX2 Source Composition Solution #1
        Source 1    Source 2    Source 3    Source 4    Source 5    Source 6
MASS   751.32693  1797.77746  2130.38570  1948.22260  4823.40666  1747.28887
Al       0.02663     0.02058     0.00852     0.00715    -0.00384     0.06025
Si       0.07384     0.06468     0.02163     0.02509    -0.01113     0.15261
S        0.02389    -0.00770     0.13148     0.05402     0.00273     0.02791
K        0.06213     0.01157     0.00333     0.00774     0.00144     0.02068
Ca       0.02972     0.02823     0.00733     0.00968    -0.00322     0.05334
Mn       0.00019     0.00272     0.00011     0.00020     0.00003     0.00086
Fe       0.00893     0.04186     0.00473     0.01020     0.00807     0.04189
Br       0.00003    -0.00007     0.00023     0.00152     0.00026    -0.00010
OC       0.30556     0.23203     0.28914     0.43094     0.52888     0.13619
EC       0.05047     0.25073     0.00758     0.05399     0.17497    -0.01524
Sol_K    0.05204     0.00274     0.00038     0.00431     0.00296    -0.00016
CO      -0.13513     0.14002     0.05016     0.12853     0.24409    -0.02127
```

Another approach to selecting species is to examine the scatterplots of all the major species against total mass. For this example use the **UNMIX Tools → Diagnostic Plots → Selected Species vs Tracer** command to produce the plot in Figure 6., where Mass has been set to the Tracer. Don't forget to unset it before running UNMIX**.** In these plots we look for species that have good upper edges. A good upper edge means that the species is effective in explaining total mass. These species are Al, Si, S, K, Ca, Mn, Fe, Br, OC, EC, Sol_K and CO. Zn, Sr and Pb do not have good upper edges, so are not chosen. The result is the same set of species that was determined above. What constitutes a "good" edge is somewhat subjective.

For the record, the estimated errors in the 6-source receptor model given above are:

```
UNMIX2 Solution # 1 Source Composition Errors
MASS   272.3885   353.1870   240.2877   370.9282   425.9801   295.1618
Al       0.0189     0.0038     0.0014     0.0040     0.0022     0.0088
Si       0.0510     0.0117     0.0036     0.0102     0.0056     0.0222
S        0.0310     0.0160     0.0076     0.0066     0.0037     0.0065
K        0.0465     0.0019     0.0008     0.0015     0.0009     0.0026
Ca       0.0199     0.0053     0.0013     0.0041     0.0021     0.0077
Mn       0.0004     0.0006     0.0001     0.0001     0.0001     0.0001
Fe       0.0071     0.0072     0.0015     0.0025     0.0014     0.0053
Br       0.0003     0.0002     0.0000     0.0002     0.0000     0.0001
OC       0.1190     0.0364     0.0193     0.0386     0.0173     0.0441
EC       0.0782     0.0502     0.0113     0.0184     0.0101     0.0201
Sol_K    0.0403     0.0010     0.0006     0.0008     0.0004     0.0008
CO       0.1990     0.0216     0.0107     0.0240     0.0138     0.0279
```

Notice the large relative error in Source1, which is high in Sol_K and associated with wood smoke or other vegetative burning. Considering the small contribution of this source to the total mass, it may be wise to eliminate K and Sol_K from the model and try adding another species.

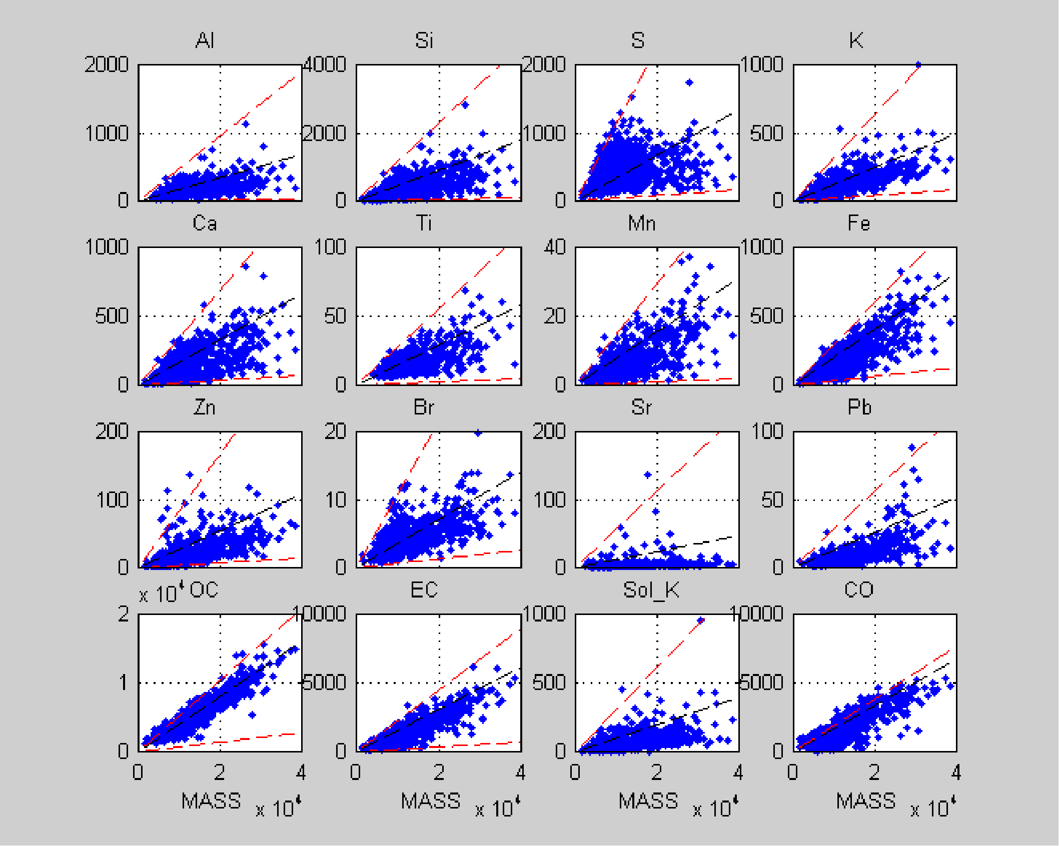


Figure 6. Scatterplots of selected species versus total mass.

Examples of Edge Plots

This section is intended as a brief introduction to the use of the edge plot command found under the menu selection **UNMIX Tools → Diagnostic Plots → Edge Plots.** Load the particle composition data file `umpmdata.txt` and select the species Mass, Al, Si, S, Ca, Fe, OC, EC and CO, as at the beginning of the section above. These species give a three-source model that was given in the example in the previous section. The three sources are sulfur, soil, and vehicle exhaust, in that order. Select source 1, the sulfur source as the base source. The edge plots are shown in Figure 7.

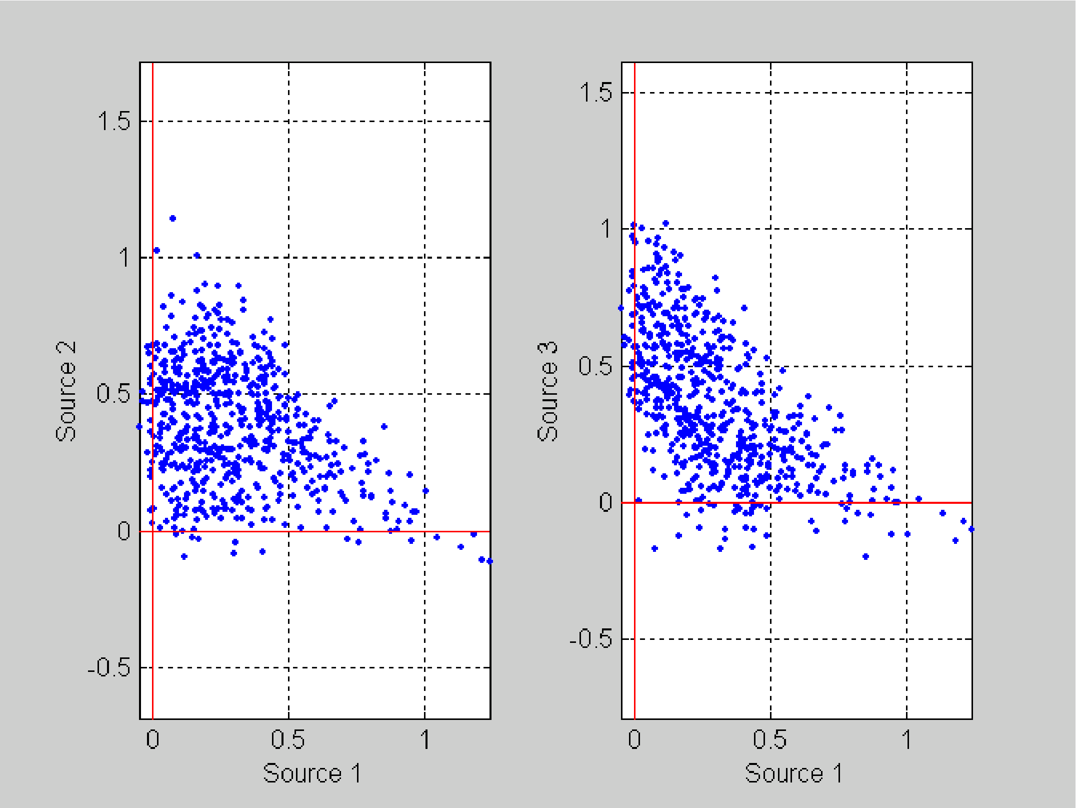


Figure 7. Edge plots for the three source model in the previous section.

The x and y axes are the edges. In both plots, the y-axis is the edge associated with source 1, the sulfur source. Points that are near the y-axis have very small contributions from the sulfur source. In the first plot, points near the x-axis are associated with small contributions of source 2, soil. Finally, the x-axis in the second plot is associated with small values of source three, the vehicle exhaust source. All the edges in these plots are typical of "good" edges.

One can select points by holding the left mouse button down and drawing a rectangle around them, as in Figure 8.

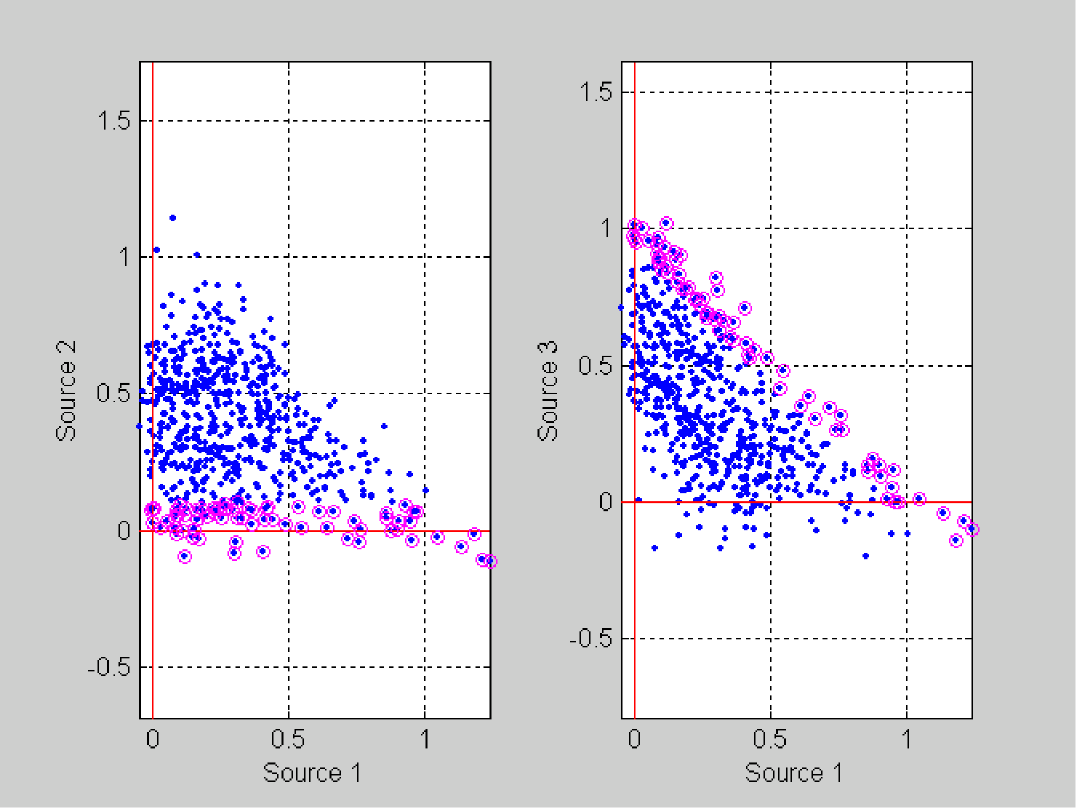


Figure 8. Same as Figure 7 showing the points with low contribution from source 2.

The points with magenta circles in Figure 8 are those close to the x-axis in the first plot, therefore these data points have little or no contribution from source 2, the soil source. (The row numbers of the selected points are displayed in the Matlab command window.) The same points are circled in the second plot. Notice that since there are only three sources these points form an edge in the second plot as well. This will NOT happen in general for plots with more than three sources.

An example of poor edges can be seen by adding Zn to the species and running UNMIX again. In this case one gets the model:

```
10 Variables, 602 Cases, 4 Factors,
Min Rsq =  0.94;  Min Sig/Noise= 9.03; Strength = 6.79
UNMIX2 Source Composition Solution #1
        Source 1    Source 2    Source 3    Source 4
MASS   553.01927  2994.05222  6645.39803  2757.44650
Al      -0.00068     0.00873     0.00285     0.04639
Si       0.01084     0.02431     0.00838     0.12174
S        0.03459     0.10544     0.00767     0.02069
Ca       0.00889     0.00881     0.00439     0.04396
Fe       0.03029     0.00531     0.01247     0.03816
Zn       0.02364     0.00024     0.00080     0.00049
OC       0.26669     0.33053     0.46787     0.19946
EC       0.32773     0.01792     0.16035     0.03928
CO       0.13577     0.05383     0.19576     0.03081
```

The edge plots for this model look like Figure 9 below.

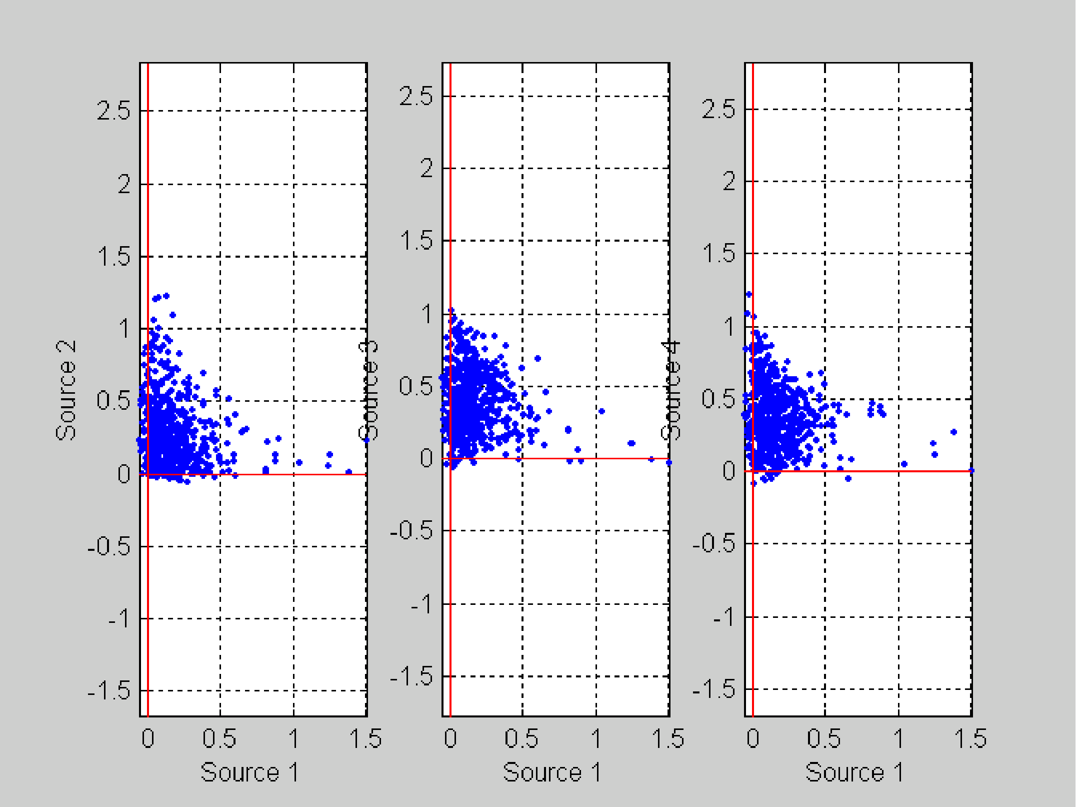


Figure 9. Edge plots for 4-source model showing examples of poor edges.

In the second plot it is seen that just a few points near the x-axis define the edge for source 3. The edge for source 4 in the third plot is also rather poor. In general, edges that are dependent on just a few points will be more greatly affected by errors that edges that are defined by many points, as seen in the first plot for the edges for source 1 and 2. Since UNMIX uses edges to find the source compositions, poor edges will lead to large uncertainties in the source compositions.

Missing Data

One of the most vexing questions that arises when applying UNMIX is what to do with species that have lots of missing values. UNMIX cannot 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 some significant, suspected source. Arsenic, selenium, nickel, and vanadium are elements that often have many values below minimum detectable limits, but when these have high values it may signify the impact of important sources.

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. In our sample particulate data set Al is an example of this last type of problem and As is an example of a species with concentrations that are just too low.

The remedy for species such as As is simple, replace the values that are below the detection limit with some small value, one half the detection limit is usually chosen. In the case of species with missing values not because the values are low but because the detection limit is high, a different approach is needed. In this case, one can often regress the species on one or more species that have few missing values. If the regression has a high r-squared, over 0.9, then the regression equation can be used to estimate the value of the species when it is missing. In the example particulate data set one can regress Al on Si to get the regression equation:

Al = 0.369Si , r-squared = 0.95. (There is no constant, because it was not significantly different from zero.)

This equation can be used to estimate Al very accurately when it is missing but Si is not.

References

Henry R. C., 1987. Current Factor Analysis Receptor Models are Ill-Posed. *Atmospheric Environment*, **21**:1815-1820.

Henry, R. C. History and Fundamentals of Multivariate Air Quality Receptor Models, 1997. *Chemometrics and Intelligent Laboratory Systems.* **37**:525-530.

Henry, R. C.; C. W. Lewis and John F. Collins, 1994. Vehicle-Related Hydrocarbon Source Composition from Ambient Data: The GRACE/SAFER Method. *Environ. Science & Technology,* **28**:823-832.

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, *Chemometrics and Intelligent Laboratory Systems*, **48**: 91-97.

Lewis, C. W.; R. C. Henry and J. H. Shreffler, 1998. An Exploratory Look at Hydrocarbon Data from the Photochemical Assessment Monitoring Stations Network*, J. Air & Waste Management Assoc*., **48**: 71 – 76.

Acknowledgements

A number of people made suggestions that led to improvements I UNMIX. These include Charles W. Lewis, Gary Norris, and Shelly Eberly of the U. S. EPA. Special thanks is given to Rich Poirot of the State of Vermont and Cliff Spiegelman of Texas A&M University.