EPA/#
June 2000

Workshop on UNMIX and PMF As Applied to $PM_{2.5}$

14-16 February 2000
U.S. EPA, RTP, NC

Final Report

by

Robert D. Willis
ManTech Environmental Technology, Inc.
Research Triangle Park, NC 27709

Contract No. 68-D5-0049

Project Officer
Portia Britt

Work Assignment Manager
Charles W. Lewis
National Exposure Research Laboratory
Human Exposure and Atmospheric Sciences Division

National Exposure Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

Notice

The U.S. Environmental Protection Agency through its Office of Research and Development funded and managed the research described here under Contract 68-D5-0049 to ManTech Environmental Technology, Inc. It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document.

NERL-RTP-HEASD-00-161 TECHNICAL REPORT DATA

1. REPORT NO. EPA/600/A-00/048	2.	3.RECIPIENT'S ACCESSION NO.
4. TITLE AND SUBTITLE Workshop on UNMIX and PMF as Applied to PM2.5		5.REPORT DATE
		6.PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Robert D. Willis		8.PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS ManTech Environmental Technology, Inc. Research Triangle Park, NC 27711		10.PROGRAM ELEMENT NO.
		11. CONTRACT/GRANT NO. 68-D5-0049
12. SPONSORING AGENCY NAME AND ADDRESS National Exposure Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711		13.TYPE OF REPORT AND PERIOD COVERED EPA Report; February 14-16, 2000
		14. SPONSORING AGENCY CODE EPA/600/09

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This report is the proceedings of a workshop convened to evaluate two new air quality receptor models, Positive Matrix Factorization (PMF) and UNMIX. The workshop was held in Research Triangle Park, NC, during February 14 - 16, 2000 and was sponsored jointly by EPA's Office of Research and Development (ORD) and Office of Air Quality Planning and Standards (OAQPS). The workshop evaluation of PMF and UNMIX was accomplished by examining the results of applying both models to two ambient $PM_{2.5}$ data sets, one real and one synthetically generated. Both data sets were supplied in advance to a proponent of each model (UNMIX: Dr. Ron Henry, University of Southern California; PMF: Dr. Phil Hopke, Clarkson University). Each brought to the workshop the results of independently applying their model to both data sets. The report briefly summarizes the technical exchange and major conclusions reached during the workshop.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b.IDENTIFIERS/ OPEN ENDED TERMS	c.COSATI
18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)* UNCLASSIFIED	21.NO. OF PAGES
	20. SECURITY CLASS *(This Page)* UNCLASSIFIED	22. PRICE

GENERAL DISCLAIMER

This document may have problems that one or more of the following disclaimer statements refer to:

- This document has been reproduced from the best copy furnished by the sponsoring agency. It is being released in the interest of making available as much information as possible.
- This document may contain data which exceeds the sheet parameters. It was furnished in this condition by the sponsoring agency and is the best copy available.
- This document may contain tone-on-tone or color graphs, charts and/or pictures which have been reproduced in black and white.
- The document is paginated as submitted by the original source.
- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Contents

Volume I: Workshop Proceedings

Agenda 1
Introduction 3

Session 1: 14 February, a.m. 4
- Opening Remarks 4
- Session 1A: UNMIX Methodology 4
  - UNMIX Results on Synthetic Data Set 5
  - UNMIX Results on the Phoenix Data Set 6
- Session 1B: PMF Methodology 6
  - PMF Results on Synthetic Data Set 8
  - PMF Results on Phoenix Data Set 8
- Session 1C: Overview of Synthetic Data Set Results 8

Session 2: 14 February, p.m. 10
- Session 2A: Description of the Synthetic Data Generation Process 10
- Session 2B: Processing of Synthetic Data and Resulting Solutions for PMF 10
- Session 2C: Processing of Synthetic Data and Resulting Solutions for UNMIX 10
- Session 2D: Description of Metric of the Goodness of Fits of the Solutions and the Results of Applying the Metric 11

Session 3: 15 February, a.m. 13
- Session 3A: Phoenix Source Apportionment Studies 13
- Session 3B: Phoenix NERL Platform Studies—Data Quality Issues and Supplementary Analyses 14
- Session 3C: PMF Analysis of Phoenix Data 14
- Session 3D: UNMIX Analysis of Phoenix Data 15

Session 4: 15 February, p.m. 16
- Session 4A: Reexamination of the Synthetic Data Results 16
- Session 4B: Demonstration of UNMIX Program 16
- Session 4C: Demonstration of the PMF Program 17
- Session 4D: Potential Effects of Data Artifacts on Receptor Modeling Results 17
- Session 4E: Open Discussion 19

Session 5: 16 February, a.m. 20
- Session 5A: Application of PMF in the Northern Great Lakes: A Tale of Two Studies 20
- Session 5B: Discussion of FPEAK, Open Discussions, and Workshop Conclusion 20

References 22
Attendees 24

Volume II: Appendices

Appendix 1A:	UNMIX User's Manual and Presentation Materials for UNMIX Theory and Applications	1A-1
Appendix 1B:	A Guide to Positive Matrix Factorization	1B-1
Appendix 1C:	Presentation Materials for Overview of Synthetic Data Set Results	1C-1
Appendix 2A:	Presentation Materials for Description of Synthetic Data Generation Process	2A-1
Appendix 2B:	Presentation Materials for Processing of Synthetic Data and Resulting Solutions for PMF	2B-1
Appendix 2C:	Presentation Materials for Processing of Synthetic Data and Resulting Solutions for UNMIX	2C-1
Appendix 2D:	Presentation Materials for Description of Metric of the Goodness of Fits of the Solutions and the Results of Applying the Metric	2D-1
Appendix 3A	Presentation Materials for Phoenix Source Apportionment Studies	3A-1
Appendix 3B	Presentation Materials for Phoenix NERL Platform Studies	3B-1
Appendix 3C	Presentation Materials for PMF Analysis of Phoenix Data	3C-1
Appendix 3D	Presentation Materials for UNMIX Analysis of Phoenix Data	3D-1
Appendix 4D	Presentation Materials for Potential Effects of Data Artifacts on Receptor Modeling Results	4D-1
Appendix 5A	Presentation Materials for Application of PMF in the Northern Great Lakes	5A-1
Appendix 5B	Presentation Materials for Discussion of FPEAK	5B-1
Appendix 6	Journal Article Excerpt Concerning EPA's PAMS Guidance on Reporting of Low Concentration Data	6-1

Volume I
Workshop Proceedings

Agenda for Workshop on UNMIX and PMF as Applied to $PM_{2.5}$
Dates: 2/14/2000–2/16/2000
Location: EPA Administrative Building Auditorium, RTP, NC

February 14, 8:30 a.m. – 5:00 p.m.

Morning Session: (Session 1)
General presentations on the methodology behind the tools and a brief presentation of the solutions found for both the Phoenix and the synthetic data set. This session is geared toward a general audience with the purpose of giving an overview of the tools and the results from their applications. The following 4 sessions will go into the details and will be at an advanced technical level, thus not for a general audience.

8:30–8:45	Welcome and Introductions (Chuck Lewis, ORD, and John Bachmann, OAQPS)
8:45–10:00	Presentation on UNMIX methodology and results for Phoenix and synthetic data set (Dr. Ron Henry)
10:00–10:15	Break
10:15–11:30	Presentation on PMF methodology and results for Phoenix and synthetic data set (Dr. Phil Hopke)
11:30–12:00	Overview describing the synthetic data set and a pictorial presentation of how close the tools reproduce the "known" profiles (OAQPS)
12:00–1:00	Lunch

Afternoon Session: (Session 2)
Thorough discussions of the results from the *synthetic data set* analysis. Includes description of the data generation, the metric used by EPA to determine how well the tools reproduced the "known" profiles, data preprocessing (e.g., outlier identification). selection criteria for which species to use in the models and the number of sources to try to fit, and a description of the solutions (identification of the fitted sources and the uncertainties with these solutions).

1:00–1:15	Description of the data generation process (OAQPS)
1:15–2:00	Presentation of processing of synthetic data and resulting solutions for PMF (Dr. Phil Hopke)
2:00–2:45	Presentation of processing of synthetic data and resulting solutions for UNMIX (Dr. Ron Henry)
2:45–3:00	Break
3:00–4:00	Description of metric of the goodness of fits of the solutions and the results of applying the metric (OAQPS)
4:00–5:00	General discussion topics such as what it means to say that one solution is better than another, how to use "known" profiles to compare with derived solutions for source identification, and whether it is realistic to have an automated source identification process (General discussion)

February 15, 8:00 a.m. – 5:00 p.m.

Morning Session: (Session 3)
Thorough discussions of the results from the *Phoenix* analysis. Includes steps used to preprocess the data to identify potential outliers, selection of species and number of sources used in the model, estimates of confidence (error bars) in the source compositions and contributions, and degree of fit obtained.

8:00–8:45	Results from other recent source apportionment studies in Phoenix (Mark Hubble, Arizona Department of Environmental Quality)

8:45–9:00 Data quality issues associated with Phoenix measurements used in current analyses, and supplementary analyses (SEM and trajectory analyses) performed to confirm sources (ORD)

9:00–12:00 (Break when needed.) Presentations by Hopke and Henry on their respective Phoenix analyses, addressing the issues listed above.

12:00–1:30 Lunch

Afternoon Session: (Session 4)

Thorough discussions on how the tools really work. In trying to use the tools over the past few months, EPA has had some questions about operating the tools and interpreting the output. This session will be a "question and answer" session, where many of the questions will have examples to illustrate them.

1:30–1:45 Reexamination of the synthetic data results (OAQPS)

1:45–2:15 Demonstration of UNMIX Program (Dr. Ron Henry)

2:15–2:45 Demonstration of PMF Program (Dr. Phil Hopke)

2:45–3:15 Potential effects of MDL on modeling results (Rich Poirot, Vermont Department of Environmental Conservation)

3:15–5:00 Open discussions on how the tools really work. Questions of interest include:

(1) Can the tools identify a source that has a discrete profile change? How different do the before and after profiles have to be for the tools to find two unique sources? (OAQPS has constructed an example.)

(2) Should the measured total mass or the reconstructed mass ($PM_{2.5}$) be included as a fitting species or not?

(3) How to identify and handle outliers?

(4) UNMIX specific questions: What are the equations behind R^2 and strength/noise? What do they measure? How are "edges" fit, especially in light of errors? Do the interior (non-edge) points have any influence on the solution? Why is it that UNMIX uses at most ~15 species and finds at most ~6 sources? Why does UNMIX often find no feasible solution? How does a user wisely use the new feature in UNMIX2 that allows for source compositions with very negative entries? Implications of not using MDLs and uncertainties (which is a continuation of the discussion started in (3))?

(5) PMF specific questions: What is rotmat and how can it be used to understand better how much rotation freedom there is in the solution? What is the appropriate FPEAK to use? Should multiple passes be made using various FPEAKS: one pass to improve source identification at the expense of the contribution component, and the second pass to accurately reflect the contribution component at the expense of source identification? How are FPEAK, FKEY, and GKEY implemented? Are they part of the regularization component of Q? (OAQPS has constructed an example that shows slightly negative FPEAKs are preferable.)

February 16, 8:30 a.m. – 12:00 p.m.

Morning Session: (Session 5)

Discussion of general problems and potential solutions regarding issues such as treatment of secondary sources, regional vs local source identification, and recommendations for further research and testing of methods. Discuss why factor analysis is "ill-posed" (i.e., produces infinitely many solutions) and begin a discussion about how to use multiple receptors with these tools.

8:30–9:15 Results from applying PMF to data from the Lake Michigan area (Dr. Kurt Paterson, Michigan Technological University)

9:15–12:00 Work on issues listed above.

12:00 End of workshop

Introduction

This report provides a summary of the Workshop on UNMIX and Positive Matrix Factorization (PMF) as Applied to $PM_{2.5}$. This 2½-day workshop was held at the EPA administrative building auditorium in Research Triangle Park, NC, during 14–16 February 2000. Sponsored jointly by EPA's Office of Research and Development (ORD) and Office of Air Quality Planning and Standards (OAQPS), the workshop was intended to facilitate an exchange of technical information on the use of two source apportionment tools as applied to particulate matter (PM). PMF and UNMIX represent the current state of the art in multivariate receptor modeling. Both methodologies attempt to generate source contribution estimates as well as source compositions using only the ambient data.

The workshop evaluation of PMF and UNMIX was accomplished by examining the results of applying both models to two ambient $PM_{2.5}$ data sets, one real and one synthetically generated. Both data sets were supplied in advance to a proponent of each model (UNMIX: Dr. Ron Henry, University of Southern California: PMF: Dr. Phil Hopke, Clarkson University). Each brought to the workshop the results of independently applying their model to both data sets. The source contributions underlying the synthetic data set were of course known to the EPA personnel who generated the data set, but this information was not made available prior to the workshop.

Approximately 40 attendees representing primarily EPA, universities, and state environmental agencies attended the workshop. A list of attendees is provided at the end of this volume.

The purpose of this report is to briefly summarize the technical exchange and major conclusions reached during the workshop. The organization of the report follows the workshop agenda. The text of the report is intentionally brief to spare the reader from overwhelming detail. Interested readers who seek more detailed information are referred to the appendices (Volume II) for hard copies of individual presentations and supporting materials.

The references given at the end of this report are intended to provide a complete list of all known publications relating to the theory and application of PMF and UNMIX.

In addition to this report, the workshop was recorded on videotape and the tapes are available for loan on request from Dr. Charles Lewis, EPA (tel: 919-541-3154; e-mail: lewis .charlesw@epa.gov).

Session 1
14 February, a.m.

Opening Remarks

Chuck Lewis (ORD), John Bachmann (OAQPS), and Shelly Eberly (OAQPS)

Chuck Lewis opened the workshop by acknowledging the efforts of Shelly Eberly who was the primary organizer of the workshop and who alternated with Chuck Lewis as session moderator. Lewis stressed that the workshop was not intended as a "shoot-out" between two competing receptor modeling approaches in order to declare a winner. Rather, the intent was to provide researchers with a better understanding of the methods in order to assess the potential of these tools for regulatory and research applications.

Lewis provided the following definition of receptor models:

> *Receptor models are mathematical procedures for identifying and quantifying the sources of ambient air pollutants and their effects at a site (receptor)*
>
> - *primarily on the basis of concentration measurements at the receptor, and*
> - *generally without need of emissions inventories and meteorological data.*

The two multivariate receptor models that are the subject of the workshop are much more complicated to understand and use than those presently in common usage. The potential reward for the complexity is that these models "do it all." That is, they generate both source contributions and source profiles, all from ambient data.

John Bachmann, Associate Director of OAQPS, stressed the importance of receptor modeling from the regulatory perspective. Receptor models can provide important scientific support for current (or future) PM standards. In addition, receptor models can be an important tool in understanding the associations among PM, visibility, and health effects, and in developing regulatory control strategies. State-of-the-art tools such as UNMIX and PMF, as well as experienced users of these tools, will be needed to interpret the large quantity of data expected from the $PM_{2.5}$ Speciation Monitoring Network.

Shelly Eberly had members of the audience introduce themselves and briefly describe their experience in receptor modeling.

The remainder of Session 1 consisted of overviews of the UNMIX and PMF models and results by their principal proponents, Drs. Ron Henry and Phil Hopke, respectively, and an overview of the synthetic data set. Session 1 was intended as a less technical summary of the methods and results for the benefit of managers and others who were unavailable for the entire workshop.

Session 1A: UNMIX Methodology

Dr. Ron Henry, University of Southern California
(Full presentation is in Appendix 1A.)

Dr. Henry presented the theory of the UNMIX model from a geometric perspective. The fundamental problem for receptor models is posed as follows: Given an ambient data set, find—with as few assumptions as possible—the number of sources, the composition and contributions of the sources, and the uncertainties. However, the problem as presented in the conventional mass balance formulation is statistically ill-defined, i.e., there exist an infinite number of solutions that have the same root mean squared error and that satisfy the non-negativity requirement for source compositions and contributions. The keys to finding a unique solution are therefore (1) to determine the number of sources in the data that are above the noise level, and (2) to find additional constraints that limit the number of solutions.

The UNMIX model takes a geometric approach to these two key problems that exploits the covariance of the ambient data. Simple two-element scatterplots of the ambient data provide a basis for understanding the UNMIX model. For example, a straight line and high correlation for Al versus Si can indicate a single source for both species (soil), while the slope of the line gives information on the composition of the soil source. In the same data set, iron does not plot on a straight line against Si, indicating other sources of Fe in addition to soil. More importantly, the Fe-Si scatterplot reveals a lower edge. The points defining this edge represent ambient samples collected on days when the only significant source of Fe was soil. Success of the UNMIX model hinges on the ability to find these "edges" in the ambient data from which the number of sources and the source compositions are extracted. UNMIX uses principal component analysis to find edges in m-dimensional space, where m is the number of ambient species. The problem of finding edges is more properly described as finding hyperplanes that define a simplex. The vertices at which the hyperplanes intersect represent pure sources from which source compositions can be determined. However, there is measurement error in the ambient data that "fuzzes" the edges making them challenging to find. UNMIX employs an "edge-finding" algorithm to find the best edges in the presence of error. Once the edges are found, the major issue remains of estimating the number of sources. UNMIX finds the number of sources using a resampling technique (NUMFACT algorithm) in which random subsets of samples are successively fit with UNMIX. Results for major sources change little during the resampling, while minor sources show considerable variability. NUMFACT calculates a signal-to-noise (S/N) ratio for each factor, and results with real data sets indicate that a S/N ratio >2 is an effective rule of thumb in estimating the number of quantifiable sources.

Using only ambient data, UNMIX outputs the following information:

- Number of sources
- Composition of each source
- Source contributions to each sample
- Uncertainties in the source compositions
- Apportionment of the average total mass, if total mass is included in the model

The major assumptions employed in UNMIX are as follows:

- Source compositions remain approximately constant.
- There are at least N*(N-1) points that have low or no impact from each of the N sources, i.e., need some points with one source missing or low.

Advantages of the UNMIX tool were given as the following:

- No assumptions about the number or compositions of sources are needed.
- No assumptions or knowledge of errors in the data are needed.
- UNMIX automatically corrects source compositions for effects of chemical reactions.

A major difference between UNMIX and PMF is that UNMIX does not make explicit use of errors or uncertainties in the ambient concentrations. This is not to imply that the UNMIX approach regards data uncertainty as unimportant, but rather that the UNMIX model results implicitly incorporate error in the ambient data.

UNMIX Results on Synthetic Data Set

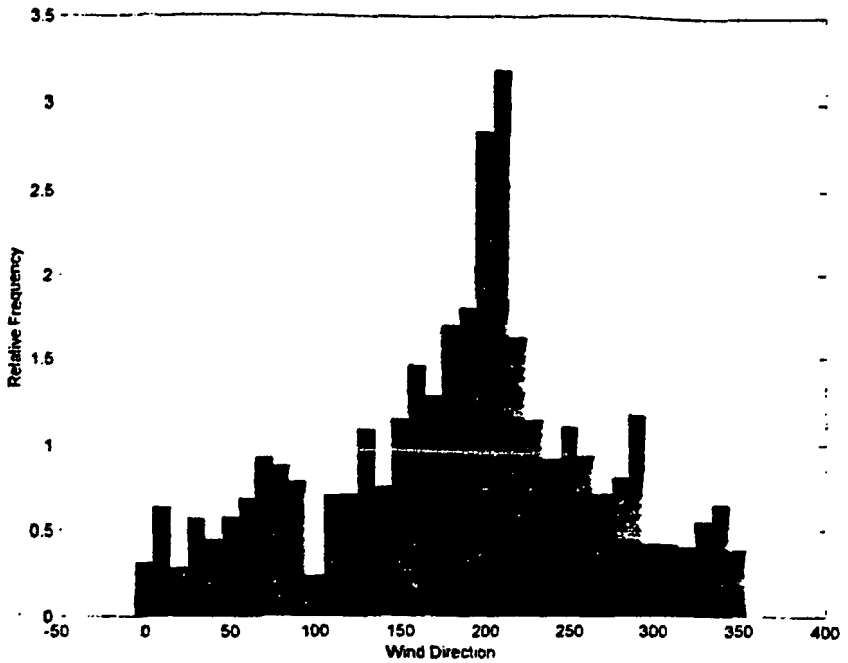
Henry summarized his seven-source UNMIX model for the synthetic data set. UNMIX source apportionment results are summarized in the following table:

Source	Mean Source Contribution ($\mu g/m^3$)
Soil	28
Vehicles	25
Steel sinter	6
Residual oil	5
Combustion	4
Palladium source	3
Asphalt roofing	2

The soil, vehicles, residual oil, and combustion sources had S/N ratios significantly above 2. The remaining sources are not statistically quantifiable but are identifiable in terms of characteristic species. The remaining sources included asphalt roofing (defined by Cs and Co), steel sinter (Cu, Cr), aircraft jet fuel (As, NO_3), as well as sources associated with Mg, Pd, and Se.

Using wind-directional analysis, Henry showed that one can extract information on source locations even for sources that are well into the noise. As an example, Henry showed a wind-directional plot of the steel sinter source (next page). Only the highest 10% of the data (samples showing the highest estimated contribution from the steel sinter source) are plotted. The hourly

Steel Sinter



wind directions for these samples are normalized to the hourly wind-direction data for all samples and the relative frequency is then plotted for each 10-degree wind sector. The plot shows that on days when the steel sinter source has a high expected source contribution, the winds are three times more likely to be from 200 to 220 degrees than the average frequency over all samples.

UNMIX Results on the Phoenix Data Set

Dr. Henry presented a six-source UNMIX solution for the Phoenix $PM_{2.5}$ data set as summarized in the following table:

Source	Mean Source Contribution ($\mu g/m^3$)
Vehicles	4.7
Secondaries	2.6
Soil	1.8
Diesels	1.2
Vegetative burning	0.7
Unexplained	1.6

Secondaries include sulfates and organic carbon. Source compositions are shown in Appendix 1A. It should be noted that the "unexplained" source represents a real source (or mixture of real sources) that was extracted by UNMIX but could not be specifically identified.

The identification of the "diesel" source hinged on the high Mn concentration and the high OC and EC concentrations, as well as the fact that this source contributed only one-fourth as much on the weekends as on weekdays. Henry speculated that the Mn is a fuel additive used (probably illegally) by diesel truck operators to prevent engine fouling. Time-series plots for the different sources are consistent with their identification, e.g., vehicle source peaks during the winter months, while the secondary source peaks during the summer.

Session 1B: PMF Methodology

Dr. Philip Hopke, Clarkson University
(Full presentation is in Appendix 1B.)

PMF is a recently developed least squares formulation of factor analysis with built-in non-negativity constraints. PMF was developed by Dr. Pentti Paatero in Finland in the mid-1990s. The tool is currently being refined jointly by Paatero and Hopke. The following is excerpted from Hopke and Song, Appendix 2B:

"Suppose X is a n by m data matrix consisting of the measurements of n chemical species in m samples. The objective of multivariate receptor modeling is to determine the number of aerosol sources, p, the chemical composition profile of each source, and the amount that each of the p sources contributes to each sample.

The factor analysis model can be written as:

$$X = GF + E \tag{1}$$

where G is a n by p matrix of source chemical compositions (source profiles) and F is a p by m matrix of source contributions (also called factor scores) to the samples. Each sample is an observation along the time axis. so F describes the temporal variation of the sources. E represents the part of the data variance unmodeled by the p-factor model.

In PMF, sources are constrained to have nonnegative species concentration, and no sample can have a negative source contribution. The error estimates for each observed data point were used as point-by-point weights. The essence of PMF can thus be presented as:

$$\min_{G,F} Q(X,\sigma,G,F) \tag{2}$$

where

$$Q = \left\| \frac{(X - GF)}{\sigma} \right\|_{F,G}^2 = \sum_i \sum_j \left(\frac{e_{ij}}{\sigma_{ij}} \right)^2 \tag{3}$$

$$e_{ij} = x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj} \tag{4}$$

with $g_{ik} \geq 0$ and $f_{kj} \geq 0$ for $k = 1,\ldots,p$, and σ is the known matrix of error estimates of X. Thus, this is a least squares problem with the values of G and F to be determined. That is, G and F are determined so that the Frobenius norm of E divided by σ (point-wise) is minimized. As shown by Paatero and Tapper [1], it is impossible to perform factorization by using singular value decomposition (SVD) on such a point-by-point weighted matrix. PMF uses a unique algorithm in which both G and F matrices are varied simultaneously in each least squares step. The algorithm was described by Paatero [2].

Application of PMF requires that error estimates for the data be chosen judiciously so that the estimates reflect the quality and reliability of each of the data points. This feature provides one of the most important advantages of PMF, the ability to handle missing and below-detection-limit data by adjusting the corresponding error estimates. In the simulated data, there were some below-detection-limit values for different chemical species. As the input to the PMF program, the concentration data and the associated error estimates were constructed as follows: For the measured data (above detection limit), the concentration values were used directly, and the error estimates were built as the analytical uncertainty plus a quarter of detection limit. For the below-detection-limit data, half of the detection limit was used as the concentration value, and as the error estimate as well. This strategy [3] appeared to work well in the present study."

Excerpt from Appendix 1B:

"Another important aspect of weighting of data points is the handling of extreme values. Environmental data typically shows a positively skewed distribution and often with a heavy tail. Thus, there can be extreme values in the distribution as well as true "outliers." In either case, such high values would have significant influence on the solution (commonly referred to as leverage). This influence will generally distort the solution and thus an approach to reduce their influence can be a useful tool. Thus, PMF offers a "robust" mode. The robust factorization based on the Huber influence function [Huber, 1981] is a technique of iterative reweighing of the individual data values."

A critical step in PMF analysis is the determination of the number of sources. Plots of the scaled residuals for all species can help determine the number of factors. It is desirable to have symmetric distributions and to have all the residuals within ±3 standard deviations. If there is asymmetry or a larger spread in the residuals, then the number of factors should be reexamined.

Note: The definition of F and G are interchanged throughout this report. In some places F represents the source compositions and G represents the source contributions and in other places F represents the source contributions and G represents the source compositions. From a mathematical perspective, this is permissible, although it may lead to confusion for the reader. Most of the current literature refers to F as the source composition matrix and G as the source contribution matrix.

PMF Results on Synthetic Data Set

Hopke presented a nine-factor solution for the simulated data set as summarized in the following table:

Source	Mean Source Contribution ($\mu g/m^3$)
Area source	26
Inner highway	24
Residual oil combustion	6
Steel sinter	1.5
Asphalt roofing	2
Municipal incinerator	1
Petroleum refinery	1
Lime kiln	5
Extra area source	2

The major sources were the area source and the inner highway source. All factors showed a reasonable relationship to the true source profiles provided to the modelers: For many factors, concentrations of most species were within 1-sigma uncertainty of the synthetic concentrations. Plots of residuals for selected species were generally symmetric and were contained within ±2 sigma. Residual plots are a useful aid in deciding how many factors are optimal. In the case of the synthetic data set, residual peaks for some species were relatively broad and asymmetric when fewer than nine factors were used. A scatterplot of the modeled mass versus the synthetic mass showed excellent agreement.

PMF Results on Phoenix Data Set

PMF yielded a six-source model for the Phoenix $PM_{2.5}$ data set as summarized in the following table:

Source	Mean Source Contribution ($\mu g/m^3$)
Biomass burning	4.4
Motor vehicles	3.5
Coal-fired power plant	2.1
Soil	1.9
Smelter	0.5
Sea salt	0.1

Motor vehicle emissions and biomass burning were the major sources. It is noteworthy that PMF was able to extract the sea-salt factor even though concentrations for the key determining species (Na and Cl) were mostly below their respective detection limits. This source was not found with the UNMIX model because the Na and Cl were not good-fitting species. Time-series plots for the six factors showed that most source contributions generally peaked during the winter; however, the sea-salt source showed aperiodic episodes. Modeled mass and observed mass were generally in good agreement. PMF was also applied to the PM_{coarse} data and a five-factor model gave best results. The five sources were identified as (1) soil, (2) construction, (3) road dust, (4) sea salt, and (5) coal-fired power plant. Soil and construction were the major sources.

In summary, Hopke cited the following advantages of PMF:

- PMF allows optimal weighting of individual data points. This in turn makes it possible to include less robust species (those with many missing values or values below the detection limit) that may nevertheless define real sources.
- PMF provides for natural inclusion of non-negativity and other constraints.
- The PMF approach will allow future inclusion of better algorithms for finding the optimal number of factors.

Session 1C: Overview of Synthetic Data Set Results

Shelly Eberly, OAQPS

(Full presentation is in Appendix 1C.)

Ms. Eberly provided a brief overview of the synthetic data and a comparison of the PMF and UNMIX results to the synthetic data. Eberly's remarks addressed the following topics:

- A description of how the synthetic data set was generated.
- Discussion of the 16 distinct sources that were input into the model. (Temporal modulation of the synthetic sources was critical in being able to resolve individual sources.)
- The geographic layout of "Palookaville."
- A summary of the average source contributions used to generate Palookaville's ambient data.
- Summary of the materials provided to the analysts (Hopke and Henry).
- Summary of the materials received from the analysts.
- Side-by-side comparison of the sources identified by UNMIX and PMF and the source contribution estimates.

Comparison of UNMIX and PMF results to the known results. This comparison is shown below:

Comparison to Known Profiles
(Amended)*

Sources identified by both tools
(known / UNMIX / PMF)

- Area / Soil / Area	28 / 28 / 26
- Inner Hwy / Vehicle / Inner Hwy	26 / 25 / 24
- Residual Oil Combustion	5 / 5 / 6
- Muni. Incin	1 / 4 / 1
- Steel Smlter	0.8 / 6 / 1.5
- Asphalt Roofing	0.4 / 2 / 2

Source identified by UNMIX only
- Palladium source (~3)

Sources identified by PMF only
(known / PMF)

- Petro Refin / Petro Refin	0.8 / 1
- Lime kiln / Lime kiln	0.8 / 5
- Coal Comb / Extra Area	1.5 / 2

*Note: Originally the above chart did not have the "municipal incinerator" source in the category of "Sources identified by both tools." UNMIX had identified the source, but under the label "Combustion source located to NE of site."

- Comparison of UNMIX and PMF residual oil combustion source profiles to the synthetic source profile.

- Scatterplots of UNMIX source strength versus true source strength and PMF source strength versus true source strength for the residual oil combustion source.

Eberly offered the following conclusions:

- The largest three known sources were correctly identified by both tools and the modeled mass was close to the simulated mass for all three sources.

- The fourth largest source (coal combustion, presence of source withheld from analysts) was not identified by either tool. PMF found a source similar to the coal combustion source but identified it as an extra area source. UNMIX did not find the source.

- Three to four smaller known point sources were identified but the estimated source contributions were larger than the true source strengths.

Following Eberly's presentation, the session was opened for questions to any of the previous presenters. Eberly was asked how the synthetic uncertainties and minimum detection limits (MDLs) were determined. Response: Each of the 50 species had a single MDL and a single uncertainty, which were fixed across the entire data set. For each species a number was randomly chosen between 5% and 10%. These numbers were used as the coefficients of variation (CVs) for log-normal distributions of the measurement errors of the species. Daily random measurement error drawn from this distribution was applied after the "true" species concentration at the receptor was computed.

An MDL for each species was provided. These MDLs were computed as a function of the average concentration and the species' measurement error CV. Specifically, the MDL for each species was computed as the maximum of 1.5 × CV × (mean concentration) and 0.001 $\mu g/m^3$. The data below the MDL were not modified in any way.

As a consequence of not modifying the data below MDL, Henry pointed out that scatterplots of certain species revealed an unrealistic structure of sub-MDL data in the synthetic data set. For example, although all values of iodine were below the MDL, scatterplots of iodine values versus other selected species showed high r^2 values, indicating that the synthesized iodine data were not truly noise.

Session 2
14 February, p.m.

Drs. Hopke and Henry described in more detail their PMF and UNMIX solutions for the synthetic data set. Dr. Basil Coutant discussed goodness of fit (GOF) metrics for evaluating receptor model solutions and the results of applying GOF metrics to the PMF and UNMIX solutions.

Session 2A: Description of the Synthetic Data Generation Process

Dr. Basil Coutant, OAQPS
(Full presentation is in Appendix 2A.)

Dr. Coutant provided a more detailed description of how the synthetic data set was generated. Sixteen distinct source profiles were used in Palookaville—nine point sources, four industrial complexes, one area source, and two highways. The area profile was a mixture of dust and road profiles. All source profiles with the exception of the petroleum refinery were fixed. The latter profile had some built-in variability (coefficient of variation of approximately 25%). Temporal modulation of the source strengths (50% CV for most) was found to be essential in being able to resolve the sources by PMF or UNMIX. A total of 366, 24-h samples were generated at the receptor site.

There was further discussion regarding MDLs. Data below the MDL should be noise with no structure. What does it mean to quote a value below the MDL? Some laboratories report values and uncertainties only for data above the MDL, while other labs (and the IMPROVE network on occasion) report values below MDL. Lewis presented EPA documentation reflecting the EPA view that it is perfectly allowable to report sub-MDL values (at least in the AIRS database for VOCs). See Appendix 6, quote from *JAW'MA* 48, 71 (1998).

Session 2B: Processing of Synthetic Data and Resulting Solutions for PMF

Dr. Phil Hopke
(Full presentation is in Appendix 2B.)

Dr. Hopke described how the synthetic data set was analyzed. Initial trials with PMF yielded low Q values indicative of incorrect weighting of the data. Alternative data weights were evaluated until the Q values became more reasonable (approximately equal to the sample size). At this point, plots of residuals are very helpful in determining the optimum number of factors. Generally, residual peaks that are broad for a whole suite of elements imply the need for more factors; residual peaks that are positively skewed imply the need for another factor(s); residual peaks that are negatively skewed imply the need for fewer factors. PMF with nine factors seemed to yield the best results. Trials with eight factors left some residual peaks with positive tails, while PMF with 10 factors failed to extract a physically interpretable 10th factor. Scatterplots of predicted mass versus the actual mass reveal whether PMF results consistently underpredict or overpredict the known mass and may provide additional guidance on whether the optimal number of factors has been used. The PMF model was run multiple times starting with totally random source profiles to ensure there was a robust solution.

Session 2C: Processing of Synthetic Data and Resulting Solutions for UNMIX

Dr. Ron Henry
(Full presentation is in Appendix 2C.)

Dr. Henry typically begins an UNMIX analysis with graphical analysis of the data. UNMIX provides the ability to

view scatterplots of the data. Scatterplots of all species versus mass are very useful in choosing those species that influence the mass and should be included in the analysis. Henry looks for straight lines between species, which can suggest a common source. He also tries to select species whose scatterplots yield well-defined edges. Scatterplots can also be used to identify outliers in the data, which can be removed if desired.

Henry typically runs UNMIX multiple times, varying the fitting species and/or the number of factors. UNMIX will consistently extract the major sources, but the minor sources come and go during successive runs. Wind-frequency plots can be helpful in locating and identifying sources, even weak sources that cannot be quantified. Based on these plots, Henry located his Palookaville sources as follows: residual oil combustion (10–30 degrees); incineration combustion (broad, 30–50 and 60–80); Se source (broad, 20–40); steel sinter (200–220); aircraft jet fuel (200–220); asphalt roofing (210–230); Pd source (260–280); Mg source (215–235). Interestingly, the location for the airport (aircraft jet fuel source) determined by Henry disagreed with the airport location as shown on the Palookaville map (see Appendix 1C), which placed the airport north of the receptor. Subsequent examination of the synthetic data set simulation by OAQPS revealed that the airport, asphalt roofing manufacture, and steel sinter sources were, in fact, inadvertently located in the same place—about 200 degrees from north, just as found by Henry and in subsequent wind-direction analyses by Hopke.

Session 2D: Description of Metric of the Goodness of Fits of the Solutions and the Results of Applying the Metric

Dr. Basil Coutant, OAQPS.
(Full presentation is in Appendix 2D.)

Dr. Coutant discussed goodness of fit (GOF) metrics developed by EPA to determine how well the tools reproduced the "known" profiles and/or contributions. Ideally, one would like a single GOF number that can indicate how closely the model results approximate the profile matrix or the contribution matrix. Two GOF metrics were described—a mean based and a median based, both of which measure the relative error in the apportioned species mass from a source. Both metrics sum these relative errors for the largest three sources only.

Both metrics were applied to the PMF and UNMIX synthetic data set solutions. The mean- and median-based GOFs yielded substantially different results. In particular, the mean-based metric is very sensitive to the largest relative errors. In these metrics developed by Coutant, all species are treated equally (no weights). There was some discussion as to the merits of (1) unequal weighting of species and (2) making the metrics independent of the number of fitting species. In addition to GOF metrics for the source profiles, Coutant described GOF metrics for (1) the source contribution matrix and (2) the raw data, and discussed the results of applying these metrics to the PMF and UNMIX solutions. Coutant presented an algorithm intended to automatically identify source profiles generated by UNMIX or PMF. For a given source profile, the algorithm finds the best match from a list of candidate profiles. (These might come from the SPECIATE source profile library, for example). The automated profile identification algorithm was applied to the PMF source profiles with promising results. The algorithm works better as more species are included. A minimum of 30 species is recommended. Some of the audience expressed concern about making such a tool available to inexperienced receptor modelers, while others felt that such a tool could assist even experienced receptor modelers in coming up with a short list of potential source identifications. There followed some discussion of the quality and reliability of SPECIATE source profiles. SPECIATE profiles certainly have error associated with them; are these errors considered in the spectral matching algorithm? In some cases, automated source identification using the SPECIATE library might be a step backward compared to reliance on knowledge of local sources. Coutant concluded that "the profile GOF metrics have worked well: they let one objectively identify sources, [and] they provide a systematic way of measuring the overall quality of the fit."

Session 2 concluded with a general discussion and questions for the presenters. Henry responded to a question about physical constraints in UNMIX. UNMIX presently does not allow the user to impose constraints on the source profiles (e.g., the user may know from experience that a certain species is absent from a source), but this could be implemented in future versions. PMF presently has only non-negativity constraints built-in, but it is possible through the regularization functions to force specific source contributions or profile components toward zero. Henry expressed his concern that the errors in both tools are not being properly estimated. As a next step in model validation, Henry proposed development of a synthetic data set with variable source profiles and more realistic error structure. UNMIX and PMF should be run on 1000 different data sets and the errors estimated by the models should be compared with the standard error of the synthetic data set to see if the model error estimates are realistic.

There was some discussion regarding how well the models deal with secondary aerosols. Basically, secondaries are a challenge for the models. In the case of regional transport, one might be able to combine UNMIX or PMF with back-trajectory methods or regional transport models. Stratifying the ambient data set by season and/or wind direction may improve the apportionment of secondaries; however, one must be careful not to make the data sets too small in the process.

Asked how Henry and Hopke view each other's model, Henry reiterated his philosophy that it is best to do as little as possible to the data and let the data speak for itself. He

expressed his concern that by weighting the data as PMF does, one runs the risk of putting additional distance between the statistical model and the physical reality. Hopke argues that the ability to weight individual data points allows the modeler to extract the most information from the data.

Session 3
15 February, a.m.

Session 3 began with a description of the Phoenix area and results from three earlier source apportionment studies. This was followed by results from an independent analysis of the same Phoenix data set to which the UNMIX and PMF models were applied. The session concluded with thorough discussions of the UNMIX and PMF results from Phoenix, including steps used to preprocess the data to identify potential outliers, selection of species and number of sources used in the model, estimates of confidence (error bars) in the source compositions and contributions, and degree of fit obtained.

Session 3A: Phoenix Source Apportionment Studies

Mark Hubble, Arizona Department of Environmental Quality
(Full presentation is in Appendix 3A.)

Mark Hubble described the Phoenix geography, meteorology, and major emissions sources. Hubble also presented results from three source apportionment studies carried out in the Phoenix area:

1. 1989–1990 Urban Haze Study (principal investigators: John Watson and Judith Chow, Desert Research Institute)

2. 1994–1995 Maricopa Association of Governments/DRI Brown Cloud Analysis (principal investigators: Tom Moore et al., Arizona Department of Environmental Quality, and Eric Fujita, Desert Research Institute)

3. 1994–1996 ADEQ/ENSR Analysis (principal investigators: Tom Moore et al., Arizona Department of Environmental Quality, and Steven Heisler, ENSR)

The first two studies were conducted during the fall and winter, while the last study was conducted during all seasons. The Urban Haze Study used conventional chemical mass balance (CMB7) to apportion fine mass ($PM_{2.5}$) and light extinction to source categories. Local motor vehicle and geological source profiles were generated. The Brown Cloud Study used conventional and extended CMB to apportion fine mass only. The extended CMB included selected semivolatile organic compounds and polycyclic aromatic hydrocarbons to separately apportion gasoline and diesel combustion. The ADEQ/ENR Study used conventional CMB to apportion fine mass and light extinction.

Results from the first two studies were in general agreement and showed that motor vehicles contributed the bulk of $PM_{2.5}$ (in the range of 44–75%) and that geological sources were typically the second most abundant source of ($PM_{2.5}$), accounting for approximately 10–20% of $PM_{2.5}$. Ammonium nitrate and ammonium sulfate were smaller but significant contributors to $PM_{2.5}$. The third study differed from the first two studies in that it was conducted year-round and it attempted to apportion vegetative burning using soluble potassium. The apportionment results showed a significant increase in vegetative burning (11–17% of $PM_{2.5}$) and geological sources (26–33%) at the expense of motor vehicles (typically <40%of $PM_{2.5}$). However, the vegetative burning source is probably overestimated since the model indicates that it contributes 15–20% of $PM_{2.5}$ during the summer months, when there should be little vegetative burning.

In conclusion:

- All studies show that most fine mass comes from combustion.

- All show similar proportions between geological and combustion source categories.

- All show rather low contributions from secondary nitrate and sulfate.

Session 3B: Phoenix NERL Platform Studies—Data Quality Issues and Supplementary Analyses

Dr. Gary Norris, NERL

(Full presentation is in Appendix 3B.)

Dr. Norris discussed the following topics in regard to the Phoenix NERL platform data:

- NERL Platform data (measurements, sampling equipment)
- Receptor modeling results
- Scanning electron microscopy results
- Health effects studies

The NERL monitoring platform in Phoenix provided data that was submitted to Drs. Henry and Hopke for the UNMIX and PMF analyses. The data consisted of collocated measurements from a dual fine-particle sequential sampler (DFPSS), a dichotomous sampler, TEOMs, and a 10-m meteorological tower. The DFPSS data were the subject of analysis unless otherwise noted. The data were collected between 1 February 1995 and 30 June 1998. Norris et al. carried out their own chemical mass balance receptor modeling study, which has recently been submitted for publication. This study attributes 42.2% of $PM_{2.5}$ to motor vehicles, 24.5% to road dust, 17% to secondary organics, 9.5% to ammonium bisulfate, 5.4% to wood smoke, and 1.4% to marine aerosol. Norris suggested that secondary organics may represent a positive artifact on the quartz filter, which may account for some of Hopke's "biomass burning" source and Henry's "secondary" source.

Scanning electron microscopy was used to validate the receptor model results and to provide evidence for additional weak sources. For example, back-trajectories pointing toward the Pacific combined with SEM images of salt aerosols provided confirmation of the marine source. SEM also identified particles suggestive of smelting operations and an unrelated source(s) of Pb particles.

Health effects associated with the Phoenix aerosol were analyzed in a recent study by Mar et al. (Associations between Air Pollution and Mortality in Phoenix, 1995–1997). Cardiovascular mortality was significantly associated with $PM_{2.5}$, coarse PM, and elemental carbon. Factor analysis revealed that combustion-related pollutants (motor vehicle exhaust and vegetative burning) and secondary aerosols (sulfates) were associated with cardiovascular mortality.

Session 3C: PMF Analysis of Phoenix Data

Dr. Phil Hopke

(Full presentation is in Appendix 3C.)

Dr. Hopke discussed his PMF analysis of the Phoenix data. Hopke found a six-source model for Phoenix. In order of descending mass contribution, these sources were biomass burning, motor vehicles, coal-fired power plant, soil, Cu smelter, and sea salt. Time-series plots of the six sources showed reasonable seasonal trends. Sea salt and soil were episodic in nature; motor vehicles, biomass burning, and perhaps the Cu smelter source appear to peak in winter. Wind-directional analysis of the copper smelter source might clarify whether this is being transported across the Mexican/U.S. border. Because PMF allows the user to fill in missing data or replace sub-MDL data, Hopke was able to use Na, Cl, and Cu species to advantage in extracting the sea-salt and copper smelter sources, in contrast to the UNMIX solution.

Determining the number of factors to include in the model is a multistep process. After obtaining a trial PMF solution, the total mass ($PM_{2.5}$) is regressed on the source contributions to apportion the mass to each of the sources. If any of the coefficients in this regression are negative, then there likely are too many factors in the model. Another technique for evaluating the number of factors is to examine the standardized residuals by species. If these residuals are not symmetric or if there are a number of residuals more than three standard deviations from the mean, this may indicate there are too many or too few factors (although it may also indicate that the uncertainties provided to PMF by the user are not appropriate).

Once the number of factors has been determined, then the correct rotation for the solution needs to be determined. One easy way to rotate the solution is through the parameter FPEAK. Graphing Q against different values of FPEAK is a useful diagnostic for selecting the appropriate rotation. As a general rule of thumb, one should increase FPEAK until Q starts to rise.

Although the selection of the number of factors and the appropriate rotation are presented here as independent steps, they, in fact, interact. For example, after selecting FPEAK, one should reexamine the residuals to be sure they are still small and symmetric and reexamine the regression coefficients to be sure they are still non-negative.

As an aside, Hopke separately applied PMF to data collected with the DFPSS and to data collected with the collocated dichot sampler. The results lent support to the modeling results since the resulting source profiles for the two samplers looked very similar with the exception of sea salt and soil. These typically represent coarse-fraction intrusion and were affected by the different inlet efficiencies for the two sampling systems.

Note: An eight-source model, whose results differ considerably from the six-source model presented at the workshop and are much more similar to the UNMIX results, has been submitted for publication (Ramadan et al., *JAWMA*, in press).

Session 3D: UNMIX Analysis of Phoenix Data

Dr. Ron Henry

(Full presentation is in Appendix 3D.)

Dr. Henry discussed his six-source solution for Phoenix using UNMIX. He excluded Na, Cl, and Cu from the list of fitting species because scatterplots versus mass indicated that little mass was associated with these species. Also, there were a large number of measured values below the detection limit. Henry's six sources in order of decreasing mass contribution were non-diesel vehicles (37%), secondaries (20%), soil (15%), diesels (10%), vegetative burning (5%), and unexplained (12%). In contrast to Hopke, Henry used soil-corrected potassium as a fitting species. The correction was made by using the lower edge in the potassium versus silicon scatterplot as an estimate of the soil potassium. Non-soil potassium proved to be very important in being able to extract the weak vegetative burning source. The secondaries source was high in S and organic carbon. The unexplained source, distinguished by Br and OC, is probably a mixture of sources according to Henry. (Phoenix has a surprising number of local OC sources according to Henry, although regional transport of OC is another possibility.) Several factors supported Henry's labeling of the diesel source. First was the high EC component. Second, Henry compared the diesel contributions on weekdays versus weekends and found nearly a factor of 4 decrease on the weekends, consistent with commercial truckers' reluctance to work on weekends. (The other sources, if anything, may have shown a tendency toward higher contributions on the weekends.) Third, some research on the Internet indicated that it is common practice among truckers (though possibly illegal) to add MMT (an octane-enhancing fuel additive) to their fuel to minimize engine fouling. This could then explain the large Mn component in the diesel source profile. Unfortunately, no traffic count data were available in the Phoenix area showing the number of diesel vehicles on weekends versus weekdays. Henry also presented the 1-sigma source composition errors that can be generated by UNMIX. By dividing each contribution in the source profile matrix by its associated error, one calculates the normalized signal-to-noise values for the source profiles. With the exception of vegetative burning (the weakest source), the great majority of these values are greater than 2.

Time-series plots of the six sources showed reasonable seasonal cycles. Vegetative burning and non-diesel vehicle sources peaked in winter, while the secondaries peaked in September–October. "Unexplained" had no discernible pattern. In contrast to the synthetic data set, wind-directional plots showed little directionality to the sources, and any directional trends that did show up were probably driven by seasonal changes in wind direction. (Winds are more likely to come from the north during the winter and the top 10% samples for the vehicle source are most likely to occur in the winter, so the wind-direction plot for the vehicle source will be skewed toward the north.)

As an aside, Henry included PM_{10} and $PM_{2.5}$ masses from collocated TEOM samplers in the UNMIX model and generated a seven-source solution. Six of the sources reproduced the previous six-source solution very well. In addition, the DFPSS fine mass and the TEOM fine mass apportioned to each of the six sources were in remarkable agreement. The additional seventh source appeared to be associated with PM_{10}.

Session 3 concluded with a brief comparison of the UNMIX and PMF solutions to the Phoenix data as summarized in the following table:

PMF		UNMIX	
Biomass burning	35%	Non-diesel	37%
Motor vehicles	28%	Secondary	20%
Coal-fired power plant	17%	Soil	15%
Soil	15%	Diesel	10%
Smelter	4%	Vegetative burning	5%
Sea salt	1%	Unexplained	12%

There were some major differences in the two solutions. The largest source in the PMF solution was biomass burning, accounting for nearly 35% of the mass. By comparison, UNMIX's vegetative burning accounted for only 5% of the mass. It is worth noting that Henry used non-soil K to extract his vegetative burning source, while Hopke did not. Hopke speculates that his biomass burning source may be a combination of Henry's diesel and unexplained sources, which account for about 22% of the mass. Motor vehicles account for about 28% of the mass in PMF versus 47% in UNMIX (combining diesels plus non-diesel). Based on profile similarities, Hopke's coal-fired power plant source, accounting for about 17% of the mass, appears to be equivalent to Henry's secondaries source, representing 20% of the mass. Hopke's soil source accounts for about 15% of the mass, the same as Henry's soil source estimate.

Session 4
15 February, p.m.

Session 4 included a reconsideration of the synthetic data results, discussions of how the tools really work, and live demonstrations of PMF and UNMIX by Hopke and Henry.

Session 4A: Reexamination of the Synthetic Data Results

Shelly Eberly

Eberly reviewed the PMF and UNMIX results for the synthetic data set and made some corrections. Specifically, UNMIX identified four sources larger than noise, including the municipal incinerator source (identified by Henry as solid material combustion) with an estimated strength of 4 $\mu g/m^3$. Both tools tended to overestimate the contributions from the minor sources. Henry explained that this is simply a consequence of the fact that both tools attempt to explain all of the observed mass with only seven or nine sources rather than the 16 sources that were used to generate the synthetic data. Therefore, some of the source contributions will necessarily be overestimated. Henry emphasized the need to put error bars on estimated source contributions when comparing results from different tools.

Other issues pertaining to the synthetic data results included the actual location of the airport in Palookaville. With regard to putting labels on sources, Henry encouraged modelers to provide a one-sentence justification for each source label so that readers will understand how the sources were identified.

Session 4B: Demonstration of UNMIX Program

Dr. Ron Henry

Dr. Henry presented a live demonstration of the UNMIX program. UNMIX is copyrighted to Henry. The current version (UNMIX2.1) is available at no charge from Dr. Henry, who requests that users not distribute the program to others. E-mail Dr. Henry at rhenry@usc.edu to request a copy. In addition to the program, users will receive a user's manual (PDF format) and some test input files. Users must have MatLab 5.3 in order to run UNMIX.

Ambient data is input to UNMIX as a flat ASCII text file with column headings. UNMIX has a user-friendly Windows interface. UNMIX provides some statistical measures to guide the user toward the best solution. These include minimum r-square (r^2) and minimum signal-to-noise (S/N). Recommended values are $r^2 > 0.8$ and S/N > 2. UNMIX allows the user to set one species as a "tracer" if desired. This forces all measured mass for that species into one source. UNMIX has an option for displaying scatterplots of any species against any other species. This is very useful in selecting fitting species. In the same plots one can identify outliers and remove them (temporarily) from the data set. One can also display "edge" plots. Henry recommends this as a good way to find out which species are important. In selecting fitting species, Henry had the following suggestions: (1) Major species must be included or the model won't be able to find a solution. (2) Select "robust" species—i.e., those with few missing or sub-MDL values. (3) Use as few species as possible, since each additional species adds error to the analysis and usually degrades the S/N. UNMIX outputs include the source composition matrix and the source contributions. Additionally, UNMIX can estimate errors in the UNMIX source compositions using a bootstrap approach in which the model is applied to 100 random subsets of the data. "UNMIX overnight" is another useful feature that allows the user to try all possible subsets of a selected set of fitting species in order to find the optimal solution. This can be a lengthy process and the user will probably want to limit the number of candidate species to seven or less.

Future improvements that Henry would like to see include (1) a stand-alone version that would not require MatLab and could potentially run much faster, (2) the ability to input constraints on source compositions, and (3) the ability to save

"fitting sessions" with all pertinent information so users can remember where they've been or reproduce earlier analyses. Asked whether the quoted uncertainties in the ambient data could be used to some advantage, Henry reiterated his philosophy that it is best to assume that you know nothing about the data and that, in his experience, uncertainties are often meaningless. Nevertheless, Henry did not entirely rule out the possibility that future versions of UNMIX may try to use the information present in the quoted uncertainties.

Session 4C: Demonstration of the PMF Program

Dr. Phil Hopke

The PMF programs are available from Dr. Pentti Paatero via the ftp site rock.helsinki.fi/pub/misc/pmf. First-time users can get PMF for a 6-month free trial period after which there is a license fee. PMF is still primarily a research tool and does not have a nice graphical interface. Researchers interested in learning to use PMF are invited to spend a week with Hopke at Clarkson University.

PMF can be run through a programmer's file editor (PFE), which is free shareware downloadable from the Internet. Every PMF job begins by setting up an *.INI file, which contains all the parameters needed for the analysis, including the file names and paths for input data files.

The output of PMF includes three matrices: the G matrix of source contributions. the F matrix of source compositions, and the matrix of residuals. PMF also outputs a text file containing a log of the current analysis session. The G matrix can be input to a statistics program in order to carry out the regression versus mass to get the scaled source contributions. The PMF program has no built-in diagnostic tools (e.g., for displaying residual plots).

Looking to the future. the PMF program may not be refined. Instead, programming efforts may be directed entirely into the Multilinear Engine (ME) program, which Hopke sees as replacing PMF (Paatero, 1999). ME is considered more flexible in its ability to handle the imposition of physical constraints. A wish list for future versions of ME includes a much more user-friendly graphical interface, the ability to input fixed source profiles or ratio constraints (e.g., Al:Si ratio), and a stand-alone version with built-in diagnostics (e.g., residual plots), which will obviate the need to export results to other software packages. Hopke speculated that it might be possible to automate to some extent the search for the optimal FPEAK by, for example, increasing FPEAK until there is a substantial rise in Q.

Further discussion of MDLs revealed a general consensus that there is considerable lack of agreement on the meaning of MDLs and how they are reported by various labs. Lewis provided the following definitions of the limit of detection (equivalent to the MDL) and limit of quantitation:

From Lloyd Currie, pg. 289, in "X-Ray Fluorescence Analysis of Environmental Samples," T.G. Dzubay, ed., Ann Arbor Science (1977):

Limit of Detection $= 3.29\ \sigma_0$
(false positive risk = 5%,
false negative risk = 5%)

Limit of Quantitation $= 10\ f\ \sigma_0$
(RSD of measured
concentration = 10%)

where $\sigma_0 = (1.0 - 1.4)$ x standard deviation of blank
and $f = 1$

It was noted that the above definitions define *method* limits, as distinguished from *sample* limits. The latter vary from sample to sample and are more realistic limits because they include the effects of spectral interferences due to other analytes present in the particular sample. Some labs report the fixed-method MDLs, and some report variable-sample MDLs. Also, some labs report values below the MDL, while others do not. Some statisticians argue for reporting only raw values plus uncertainties and dispense with the concept of MDLs. Hopke is currently investigating the use of a statistical method known as "multiple imputation" as a way to use existing data to impute missing data, but this research is in a preliminary stage. The discussion did not lead to any resolution of the difficult issue of how best to handle and report nondetected values.

Session 4D: Potential Effects of Data Artifacts on Receptor Modeling Results

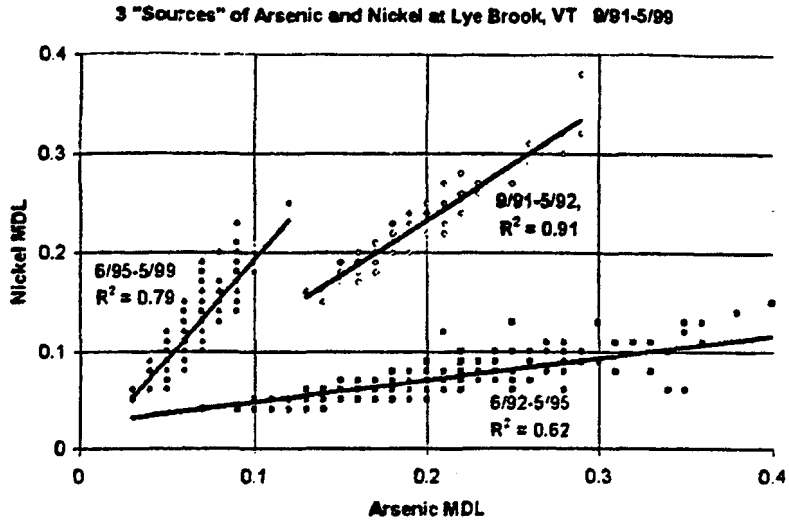
Rich Poirot, Vermont Department of Environmental Conservation

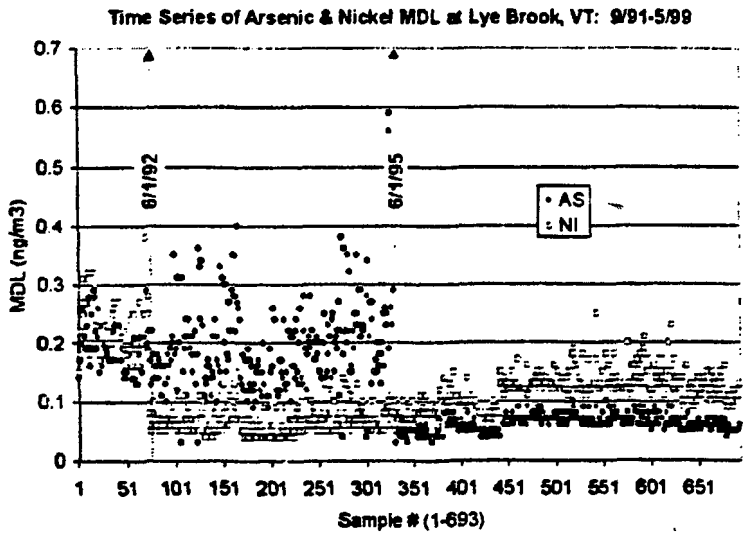
(Full presentation is in Appendix 4D.)

Data artifacts, which can include measurement errors, uncertainties, and various hole-filling replacements for nondetects, can interfere with the identification of real sources. Poirot discussed his experience with UNMIX and dealing with nondetect data. There are two choices for dealing with nondetects: one can censor the input data to screen out all nondetects, or one can use some hole-filling techniques to replace nondetects. The former approach can create a small and biased subset of the original data. Poirot discussed the results of using various hole-filling techniques to modify the input data for UNMIX calculations. In the end, Poirot

felt that simple replacement of nondetect values with zeros (or some small constant) yielded the most consistent and interpretable UNMIX results.

Poirot showed a series of slides lending support to those who mistrust reported uncertainties and MDLs. For example, Ni and As measurements at Lye Brook, VT, are totally uncorrelated and yet the reported uncertainties exhibit a significant positive correlation (top figure below). Also, he said, "although concentrations of Ni and As are uncorrelated, their MDLs are highly correlated, both as a function of three methods changes in different time periods, and also within each of three different reporting periods" (bottom figure below).





Poirot went on to say that "[this is] possibly due to common interferences or instrumental drift, but not due to changing ambient concentrations. Generally, in most long-term measurement programs both ambient concentrations and detection limits are likely to decrease over time, creating the possibility of false positive correlations between source activity for some elements and lab activity for other elements." This latter point was elegantly demonstrated by a plot of same-day, above-MDL As concentrations at Acadia and Mt. Rainier IMPROVE sites. The measured concentrations exhibit no correlation (as expected given the continental distance between sites). However, same-day As MDLs for these sites are correlated, generally due to "progress" (improving detection limits over time) in the 10+ year IMPROVE network. Poirot also provided evidence for "misquantified" MDLs for Al in IMPROVE data. He presented some encouraging results, which showed that despite wide differences in data preprocessing and model input, both UNMIX and PMF identified three common sources in an IMPROVE-like data set. However, artifacts associated with changes in Se MDLs due to a change from PIXE to XRF analysis during the sampling period clearly influenced the UNMIX and PMF results in different ways.

Poirot concluded by saying. "Data Artifacts, including MDLs and uncertainties as reported by labs and/or as processed by data analysts, can and do influence receptor model results."

Session 4E: Open Discussion

There was further discussion of the MDLs. It was not known whether the EPA $PM_{2.5}$ Speciation Monitoring Network will report the single method-based MDLs or the daily-varying sample MDLs. Henry reconsidered his distrust of reported MDLs and uncertainties and found it to be justified. In situations where one cannot afford to lose data due to nondetects, Henry recommends just replacing the nondetects with zero or a small constant.

Important MDL-related questions include the following: How are MDL and uncertainty values determined by analytical laboratories? Do these reported values have the same meaning at different labs or in different measurement programs? How have analytical methods and the resultant data changed over the course of a measurement program? And finally, what are the best ways of processing this information as input to different receptor models?

In response to the question of whether or not to use mass as a fitting species, Henry and Hopke expressed different philosophies. Henry likes to include the mass so that the total mass is apportioned just like the species mass. Hopke has traditionally kept the mass separate and likes to use the results of the mass regression analysis as an added check on the validity of the model results.

Henry expressed his concern that the errors reported in both UNMIX and PMF have not been given adequate scrutiny. Hopke believes that the error estimates in PMF are almost certainly overestimates.

Several members of the audience commented on the dreaded UNMIX message informing the user that there was "no feasible solution" to a problem. Henry responded that rather than viewing this as a bug or deficiency in UNMIX, it should instead be viewed as a valuable feature in that a bad solution is worse than no solution.

There was some discussion about dealing with outliers. Henry relies heavily on UNMIX scatterplots to identify outliers. He urged caution in eliminating suspected outliers because, if real, they can provide very important information about source compositions. Hopke typically does a principal components analysis of the data and plots factor scores to identify outliers. The "robust mode" option in PMF automatically downweights outliers (but does not eliminate them) so that they do not exert too much influence. If the user knows that a certain sample is an outlier (e.g., fireworks on the Fourth of July), then it is best to remove that data point before performing UNMIX or PMF analysis.

The interpretation of source profiles remains one of the biggest challenges in using these tools. Receptor modeling should not be done in a vacuum. Ideally, the modeler will have intimate knowledge of the modeled airshed, or will work closely with someone who does. Rich Poirot suggested creating an informal, unofficial bulletin board or site where modelers could share source profiles (accompanied by some descriptive information) generated by UNMIX or PMF. Lewis would like modelers to show their profiles in publications. With emphasis on $PM_{2.5}$, there is likely to be increased mass being apportioned to regional sources, which are typically dominated by secondary species. It would be useful to compile a library of regional "fingerprints." Such a library could be helpful in proper source identification. There are some good tools such as residence time analysis, back-trajectory analysis, and partial source contribution function (PSCF) analysis for identifying and quantifying regional impacts. Hopke showed how PSCF was able to trace a Ni-V factor in Vermont back to residual oil combustion in the Eastern urban corridor.

Session 5
16 February, a.m.

Session 5A: Application of PMF in the Northern Great Lakes: A Tale of Two Studies

Dr. Kurt Paterson, Michigan Technological University
(Full presentation is in Appendix 5A.)

Dr. Paterson presented an overview of two studies conducted in the Northern Great Lakes in which PMF was applied. The first study involved source apportionment of a mixture of trace gases and particulate matter in order to identify the sources that influence air quality in the northern Great Lakes. PMF extracted three sources identified by Paterson as biogenic (defined by isoprene), local, and regional transport. Paterson combined PMF with residence time analysis, met data analysis, and time-series analysis to confirm the identification of the sources. In the second study PMF was used on particle size distribution data, not to apportion sources, but to extract distinct factors that could reveal the dynamics of different particle modes. The original data comprised 100 size ranges from 5 nm to 7.5 μm and 1046 half-hour samples. PMF collapsed this data into six factors, which fell out into distinct particle size ranges and which exhibited different dynamic properties. Two factors, for example, showed strong diurnal cycles. Two factors were most influenced by long-range transport. And $PM_{2.5}$ mass was most influenced by particles in the size range 220–800 nm. The chemical composition data for these samples are now available and Paterson will repeat these analyses, adding in the composition data and using both PMF and UNMIX.

Session 5B: Discussion of FPEAK, Open Discussions, and Workshop Conclusion

Depending on the input data set, PMF may generate multiple solutions that are all equally valid within the rotational ambiguity of the PMF model. Somehow the user must decide which rotation is the best. FPEAK is one parameter available in PMF that allows the user to try various rotations. Positive FPEAK values force the source composition matrix toward more extremes (zeros for some species and large percentages for other species) and the source contribution matrix toward less extremes, while negative FPEAK values produce the opposite effect. Eberly presented a simple example (seven samples, three species, two sources) to show the effect of FPEAK (see Appendix 5B). PMF was executed with FPEAK values of -0.5, 0.0, and 0.5, and the resultant source composition and contributions were presented. All three of these possible solutions are consistent with the measurements recorded at the receptor, that is, the masses balance. Examination of the solutions shows that (1) for the negative FPEAK value, the source contributions are the most extreme, including some days when one source is not contributing, and (2) for the positive FPEAK value, the source compositions are the most extreme, including a species whose proportions are 0.01 and 0.85.

UNMIX was also run on the simple example and the results were presented. UNMIX produces only one solution and this solution had compositions and contributions similar to those from PMF where the FPEAK value was -0.5. The reason for this is that the UNMIX algorithm assumes there are days when each source is not contributing to the receptor. That is, UNMIX seeks sources for which there are some contributions near zero, and this is similar to what PMF does with negative FPEAK values.

As mentioned, a requirement of UNMIX is that there must be sampling days when each source disappears or is insignificant. How does UNMIX handle a source like motor vehicles in Washington, DC, which never turns off? Henry responded that this was the reason for putting the "tracer" option in UNMIX. This option allows the user to select one species as a tracer. This constrains the UNMIX solution by forcing all of the tracer species mass into one source. For motor vehicles, Henry recommended using CO as a tracer (not perfect, but usually good enough). Without a tracer in this case, UNMIX may not find a feasible solution.

Is there a rule of thumb for the number of samples needed by UNMIX or PMF? It is really a signal-to-noise problem. PMF has been applied to as few as 40 samples, but typically there is not enough variability present in so few data points to be able to pull out distinct factors. Recent work by John Ondov (PM2000

Charleston Conference) has shown that by sampling with high time resolution (half-hour) one can dramatically improve the signal/noise for sources with temporal variability. Henry offered the following rule of thumb for UNMIX: 200–300 samples may get you five sources; 2000–3000 samples may be needed to extract 9–10 sources.

How can receptor modelers take advantage of the EPA Speciation Monitoring Network now coming online? What tools are available to interpret these data? Instead of modeling multiple species at a single site, one can model a single species across multiple sites. In this way, one can extract spatial concentration gradients, which, combined with wind-direction analysis, can identify source locations. Alternatively, one can model multiple species at multiple sites using three-way factor analysis (Hopke et al., 1998).

A member of the audience pointed out the discrepancies in the UNMIX and PMF solutions for the Phoenix data, most notably the mass apportioned to vegetative burning in the two models. Are such discrepancies the result of applying different models, or the result of different people interpreting the same information? Hopke responded by saying that the modeler needs to tap into the local expertise to help identify important sources and to screen out unreasonable solutions. It is always good to come at a problem with as many tools as possible. If one can get similar solutions using both PMF and UNMIX, this adds confidence to the results.

Henry proposed a strategy that combines UNMIX and PMF and should yield defensible solutions. In this combined approach, the modeler might start with UNMIX to estimate the number of factors and to get good starting source profiles. UNMIX profiles could be used as starting profiles for PMF, since PMF is particularly good at finding smaller sources and including additional species. (This will shorten the PMF analysis since the model does not have to start with random profiles.) Applying other information such as wind-direction plots, one can probably come up with 10 or more sources. The ability to look at residuals in PMF can be very helpful as a quality check at the end of the modeling process.

Kurt Paterson suggested that it would be very useful to have thorough training tutorials for both PMF and UNMIX showing detailed applications of the tools in actual case studies.

There was a broad discussion regarding the roles of regional planning bodies and state regulatory agencies in dealing with compliance issues. Within a few years, regulatory agencies will need to address reductions at both the regional and local levels, with the regional planning bodies probably taking the lead. Will the state and regional agencies have the resources and the expertise to utilize the latest modeling tools? How can PMF and UNMIX be used to separate the regional from the local sources? Can the IMPROVE and Speciation networks be combined in some way to help separate regional sources from local sources? Henry responded that there presently exist a handful of good tools for dealing with regional sources. The challenge is for someone to put these tools together and make people aware that they exist. Perhaps the EPA regional offices can play a role in disseminating information about these tools or providing training to state agencies.

References

PMF Publications

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

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

Garrido Frenich, A., M. Martínez Galera, J.L. Martinez Vidal, D.L. Massart, J.R. Torres-Lapasió, K. De Braekeleer, J.H. Wang, and P.K. Hopke (2000) Resolution of Multicomponent Peaks by OPA, PMF and ALS, *Anal. Chim. Acta* 411:145–155.

Hopke, P.K.,P. Paatero, H. Jia, R.T. Ross, and R.A. Harshman (1998) Three-Way (PARAFAC) Factor Analysis: Examination and Comparison of Alternative Computational Methods as Applied to Ill-Conditioned Data, *Chemom. Intell. Lab. Syst.* 43:25–42.

Hopke, P.K., Y. Xie, and P. Paatero (1999) Mixed Multiway Analysis of Airborne Particle Composition Data, *J. Chemom.* 13:343–352.

Hopke, P.K. (2000) A Guide to Positive Matrix Factorization, report prepared for the Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, under Contract No. 9D-1808-NTEX, January 2000.

Hopke, P.K. (2000) Application of Source Apportionment Methods to the State Implementation Planning Process, report prepared for the Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, under Contract No. 9D-1808-NTEX, January 2000.

Huang, S., K.A. Rahn, and R. Arimoto (1999) Testing and Optimizing Two Factor-Analysis Techniques on Aerosol at Narragansett, Rhode Island, *Atmos. Environ.* 33:2169–2185.

Juntto, S., and P. Paatero (1994) Analysis of Daily Precipitation Data by Positive Matrix Factorization, *Environmetrics* 5:127–144.

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

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

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

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

Paatero, P., and U. Tapper (1994) Positive Matrix Factorization: A Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values, *Environmetrics* 5:111–126.

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

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

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

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

Ramadan, Z., X.-H. Song, P.K. Hopke (2000) Identificaton of Sources of Phoenix Aerosol by Positive Matrix Factorization. *JAWMA*. In press.

Xie, Y.-L., P.K. Hopke, and P. Paatero (1998) Positive Matrix Factorization Applied to Curve Resolution Problem, *J. Chemom.* 12:357–364.

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

Xie, Y.-L., P.K. Hopke, and P. Paatero (1999c) Calibration Transfer as a Data Reconstruction Problem, *Anal. Chim. Acta* 384:193–205.

Xie, Y. L., P. Hopke, P. Paatero, L.A. Barrie, and S.M. Li (1999b) Identification of Source Nature and Seasonal Variations of Arctic Aerosol by Positive Matrix Factorization, *J. Atmos. Sci.* 56:249–260.

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

UNMIX and SAFER (the Parent Model of UNMIX) Publications

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, 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.

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.

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 (2000) Extension of Self-Modeling Curve Resolution to Mixtures of More Than Three Components. Part 3, *Chemom. Intell. Lab. Syst.* Submitted.

Park, E.S., C. Spiegelman, and R.C. Henry (2000) Bilinear Estimation of Pollution Source Profiles and Amounts by Using Receptor Models, *Communications in Statistics, Simulation & Computation*, 29(3). In press.

UNMIX/PMF Receptor Modeling Workshop Attendees
14–16 February 2000
U.S. EPA, RTP, NC

Name	Affiliation	Phone	E-mail	Mailing Address
Charles Lewis	U.S. EPA	919-541-3154	lewis.charlesw@epa.gov	U.S. EPA NERL (MD-47) RTP, NC 27711
Philip K. Hopke	Clarkson University	315-268-3861	hopkepk@clarkson.edu	Clarkson University Box 5810 Potsdam, NY 13699-5810
Ron Henry	University of Southern California, Los Angeles	213-740-0596	rhenry@usc.edu	
John Bachmann	U.S. EPA	919-541-5359	jbachmann@epa.gov	U.S. EPA NERL (MD-10) RTP, NC 27711
Scott Kegler	U.S. EPA	919-541-4906	kegler.scott@epa.gov	U.S. EPA NERL (MD-52) RTP, NC 27711
Mark Hubble	Arizona DEQ	602-207-4481	mhl@ev.state.az.us	Arizona DEQ 3033 N. Central Ave. Phoenix, AZ 85012
Lara Autry	U.S. EPA	919-541-5544	autry.lara@epa.gov	U.S. EPA NERL (MD-14) RTP, NC 27711
Donna Kenski	U.S. EPA	312-886-7894	kenski.donna@epa.gov	EPA Region 5 77 W. Jackson Blvd. Chicago, IL 60604
Tom Pace	U.S. EPA	919-541-5634	pace.tom@epa.gov	U.S. EPA NERL (MD-14) RTP, NC 27711
Larry Cox	U.S. EPA	919-541-2648	cox.larry@epamail.epa.gov	U.S. EPA NERL (MD-75) RTP, NC 27711
N. Dean Smith	U.S. EPA	919-541-2708	smith.dean@epa.gov	U.S. EPA NERL (MD-61) RTP, NC 27711
Basil Coutant	U.S. EPA	919-541-5028	coutant.basil@epa.gov	U.S. EPA NERL (MD-14) RTP, NC 27711
Alan Vette	U.S. EPA	919-541-1378	vette.alan@epa.gov	U.S. EPA NERL (MD-56) RTP, NC 27711
Anne Rea	U.S. EPA	919-541-0053	rea.anne@epa.gov	U.S. EPA NERL (MD-56) RTP, NC 27711
Teri Conner	U.S. EPA	919-541-3157	conner.teri@epa.gov	U.S. EPA NERL (MD-46) RTP, NC 27711

Name	Affiliation	Phone	E-mail	Mailing Address
Tom Braverman	U.S. EPA	919-541-5383	braverman.tom@epa.gov	U.S. EPA NERL (MD-14) RTP, NC 27711
Allan Marcus	U.S. EPA	919-541-0636	marcus.allan@epa.gov	U.S. EPA NERL (MD-52) RTP, NC 27711
Shao-Hang Chu	U.S. EPA	919-541-5382	chu.shao-hang@epa.gov	U.S. EPA NERL (MD-15) RTP, NC 27711
Ned Meyer	U.S. EPA	919-541-5594	meyer.ned@epa.gov	U.S. EPA OAQPS (MD-14) RTP, NC 27711
Peter Frechtel	U.S. EPA	919-541-1173	frechtel.peter@epa.gov	U.S. EPA OAQPS (MD-14) RTP, NC 27711
Eugene Kim	University of Washington	206-526-2909	ugene@u.washington.edu	
Tom Rosendahl	U.S. EPA	919-541-5314	rosendahl.tom@epa.gov	U.S. EPA OAQPS (MD-15) RTP, NC 27711
Cynthia Howard Reed	U.S. EPA	703-648-5222	howard.cynthia@epa.gov	U.S. EPA 12201 Sunrise Valley Dr. 555 National Center Reston, VA 20192
Barbara Parzygnat	U.S. EPA	919-541-5474	parzygnat.barbara@epa.gov	U.S. EPA OAQPS (MD-14) RTP, NC 27711
Bob Willis	ManTech Environmental	919-541-2809	willis.robert@epa.gov	ManTech Environmental P.O. Box 12313 RTP, NC 27709
Kaz Ito	New York University	914-731-3540	kaz@env.med.nyu.edu	New York University, Env. Med. 57 Old Forge Rd. Tuxedo, NY 10987
Shaibal Mukerjee	U.S. EPA	919-541-1865	mukerjee.shaibal@epa.gov	U.S. EPA NERL (MD-47) RTP, NC 27711
John Langstaff		919-967-6649	jlangstaff@mindspring.com	1105 Valley Park Dr. Chapel Hill, NC 27514
Melissa Gonzales	U.S. EPA	919-966-7549	gonzales.melissa@epa.gov	U.S. EPA MD-58A RTP, NC 27711
Tom Coulter	U.S. EPA	919-541-0832	coulter.tom@epa.gov	U.S. EPA MD-47 RTP, NC 27711
Rich Poirot		802-211-3807	richpo@dec.anr.state.vt.us	103 South Main St. Waterbury, VT 05671
Steve Fudge		919-933-9501	fudge.steve@ecrweb.com	1129 Weaver Dairy Rd. Chapel Hill, NC 27514

Name	Affiliation	Phone	E-mail	Mailing Address
Shelly Eberly	U.S. EPA	919-541-4128	eberly.shelly@epa.gov	U.S. EPA NERL (MD-14) RTP, NC 27711
Jong Hoon Lee	Rutgers University	732-932-0306	jhlee@aesop.rutgers.edu	Rutgers University 14 College Farm Rd. New Brunswick, NJ 08901
Barbara Turpin	Rutgers University	732-932-9540	turpin@aesop.rutgers.edu	Rutgers University 14 College Farm Rd. New Brunswick, NJ 08901
Kurt Paterson	Michigan Tech University	906-487-3495	Paterson@mtu.edu	Michigan Tech University Dept. Civil & Env. Engineering 1400 Townsend Drive Houghton, MI 49931