

May 16, 2003

Erik Gribbin
Data Analysis Team Leader
TCEQ (MC-164)
P. O. Box 13087
Austin, Texas 78711-3087

STI Ref. No. 900700

Re: Exploratory Source Apportionment of Auto-GC Data

Dear Erik,

As part of Work Order No. 31985-20, Exploratory Source Apportionment of Auto-GC Data, enclosed is the final report for Task 3. The objective of Task 3 was to complete source apportionment analyses and prepare a report discussing the analyses.

In previous analyses of the auto-GC data collected at the Clinton Drive site in the Houston Ship Channel, Sonoma Technology, Inc. (STI) described general characteristics of the data and proceeded to more detailed analyses of VOC characteristics (e.g., composition, concentration ranges, diurnal profiles) during ozone episodes. For this work order, we performed exploratory source apportionment of the 1998-2001 Clinton auto-GC data using receptor-based factor analysis models. These models, such as Positive Matrix Factorization (PMF), require a relatively large data set for which the auto-GC data are ideally suited. PMF extracts "factors" which are essentially profiles (or fingerprints) of source emissions as they appear at the receptor (sampling site). The analyst then infers the source type from the factor composition, diurnal or seasonal variation, and wind-direction dependence.

In our research of the literature, there were no published applications of PMF to auto-GC data and little guidance on applying PMF to such a large (more than 21,000 records) data set. There are many decisions to be made in applying PMF to the Clinton auto-GC data set including the selection of model settings, treatment of missing data, treatment of data below detection, and selection of data subsets (e.g., all data, morning-only data, data by wind quadrant). We have used existing guidance to perform a technically sound application of PMF, but the work should be considered exploratory because additional model-sensitivity tests should be conducted when more funds are available.

STI developed a database of auto-GC data from the Clinton Drive site during 1998-2001 for use in PMF; samples and species were carefully screened, resulting in over 21,000 data

points used in source apportionment. A corresponding uncertainty file was also generated, with an uncertainty value scaled to each data point. This utilizes one of the strengths of PMF, which is its ability to weight each individual data point. Before this robust source apportionment tool was applied, simple factor and cluster analyses were conducted. This gave a range of factors and further demonstrated the heavy influence of wind direction on species' concentrations. PMF was then applied; the large number of data points and extreme outlying concentrations of nearly all species severely complicated this exploratory work. Fifteen sources were identified by their composition fingerprints, temporal characteristics (i.e., time of day, day of week, season) and wind-direction dependencies. Further temporal analyses were conducted by time of day, day of week and season, as well as detailed wind-direction analysis, including use of a Conditional Probability Function (CPF). Source profiles were scaled by reactivity, and their reactivity potentials compared. Sources were investigated for differences on mornings of ozone episodes during the summer; six sources were found to have significantly higher weight percents on mornings of ozone episodes. Outlying model residuals of reactive species were investigated, and no relationship with high ozone was found. Another receptor model, UNMIX, was applied, which found five factors that were composites of PMF factors.

As part of this work order, we delivered an interim report in March 2003 entitled "Preliminary Analyses and Assembly of Houston Auto-GC 1998-2001 Data for Exploratory Source Apportionment". The information in the interim report was incorporated into the final report. Key findings include the following:

- *This exploratory application of PMF to an hourly PAMS VOC data set was successful in identifying and quantifying VOC sources in the Houston Ship Channel area.* This hourly data set was particularly useful when apportioning sources by wind direction, time of day, day of week, and season. Combining individual species reactivity with PMF source profiles provided additional insights into the relative importance of source contributions to ozone formation.
- *Fresh emissions occur all day.* Only small differences between morning and afternoon factor analyses were found, suggesting that fresh emissions occur all day (as observed by Brown and Main, 2002) and that depletion of reactive species by atmospheric reactions should not overly interfere with the source apportionment.
- *The auto-GC data provide a rich and useful database for receptor-based source apportionment.* Diurnal and wind direction patterns in factor strength were useful in aiding in the identification of factors; these patterns can only be inspected fully with hourly data.
- *The mix of VOC sources at the Clinton Drive site is complex.* Significant differences were found in both the number and composition of factors (sources) by wind direction, illustrating the complex mixture of emissions that impact the Clinton Drive site. The best PMF solution identified 15 factors, which included the following:
 - A motor vehicle factor (on average 4% of total mass) and a diesel factor (on average 2% of total mass). A mixed aromatic factor was also found (on average 12% of total mass) that likely has some motor vehicle influence.

- An industrial flare factor comprised of acetylene, ethane, ethane, and n-butane was identified. While there may be some mobile source influence in this factor, factor strength does not decrease on weekends, indicating it is mainly from stationary sources.
- Factors with high concentrations of reactive olefins and aromatic hydrocarbons to the south and east.
- Source profiles were scaled by reactivity coefficients, and *no factor, compound, or compound class (such as olefins) dominated the overall reactivity potential*; this finding is consistent with earlier results. Dramatic shifts in importance are found when the factor compositions are weighted by reactivity, including a decrease in importance of the background+fresh and evaporative factors (5, 10, 11) and an increase in importance of the C2-C5 olefins (Factors 4, 7, 15) and aromatic hydrocarbons (Factor 8).
- *Only six factors contributed more to the VOC mix on mornings of ozone episodes (by median weight percent) at a 95% confidence level than on other mornings.* These included industrial flare (Factor 1), heavy aromatic hydrocarbons (2), motor vehicles (3), solvents (10), light paraffins (11), and industrial/mobile aromatic hydrocarbons (12). This analysis highlights that *aromatic hydrocarbons may be more important* than previously thought in ozone formation. While the high concentrations of the more reactive compounds (e.g., light olefins) appear to support a high “background” of ozone, high concentrations of the aromatic hydrocarbons may provide additional potential to form ozone and push ozone concentrations above 125 ppb. Also, the inclusion of the industrial flare factor is significant, as earlier analyses were not able to determine that industrial flare concentrations were higher on episode days.

We have enjoyed working with TCEQ on this project. Please call either Hilary or me if you have any questions regarding the final report.

Sincerely,

Steven G. Brown
Air Quality Analyst

Hilary R. Hafner
Sr. Manager, Air Quality Data Services

Enclosure

cc. Neil Wheeler (STI)
Paul Roberts (STI)



Sonoma Technology, Inc.

1360 Redwood Way, Suite C
Petaluma, CA 94954-1169
707/665-9900
FAX 707/665-9800
www.sonomatech.com

**EXPLORATORY SOURCE APPORTIONMENT
OF HOUSTON'S CLINTON DRIVE AUTO-GC
1998-2001 DATA**

**FINAL REPORT
STI-900700-2317-FR**

**By:
Steven G. Brown
Hilary R. Hafner
Sonoma Technology, Inc.
1360 Redwood Way, Suite C
Petaluma, CA 94954-1169**

**Prepared for:
Erik Gribbin
Texas Commission on Environmental Quality
MC 164
P.O. Box 13087
Austin, TX 78711-3087**

May 15, 2003

This page is intentionally blank.

ACKNOWLEDGMENTS

This work is sponsored by the Texas Commission on Environmental Quality (TCEQ) as part of the Modeling Assistance Project II, Work Order Number 31985-03-20. Erik Gribbin is the TCEQ Work Order Manager.

This page is intentionally blank.

TABLE OF CONTENTS

<u>Section</u>	<u>Page</u>
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	vii
LIST OF TABLES	xiii
1. INTRODUCTION.....	1-1
1.1 Background.....	1-1
1.2 Purpose.....	1-1
1.3 Source Apportionment	1-1
1.3.1 Overview.....	1-1
1.3.2 Approach to Source Apportionment.....	1-3
1.3.3 Source Apportionment Tools.....	1-4
1.4 Key Findings.....	1-4
1.5 Overview of Report.....	1-7
2. SOURCE APPORTIONMENT	2-1
2.1 Mathematical Framework.....	2-1
2.2 Estimation of Weights.....	2-3
2.3 Estimation of the Number of Factors.....	2-4
2.4 Factor Rotations	2-5
2.5 Other Parameters.....	2-6
2.6 Mass Apportionment.....	2-6
2.7 Conditional Probability Function.....	2-7
3. DATABASE PREPARATION	3-1
3.1 Data Assembly and Reduction.....	3-1
3.2 Treatment of Missing Data, Data Below Detection, and Development of Uncertainty.....	3-3
3.3 Selection of Species	3-3
3.4 Summary of Data Set for Receptor Modeling	3-5
4. STATISTICAL ANALYSIS	4-1
4.1 Factor Analysis	4-1
4.1.1 Overall Results.....	4-1
4.1.2 Factor Analysis by Year	4-3
4.1.3 Factor Analysis in Summer	4-6
4.1.4 Factor Analysis by Time of Day.....	4-7
4.1.5 Factor Analysis by Wind Direction	4-9
4.2 Cluster Analyses	4-11
4.2.1 Overall Cluster Analysis.....	4-12
4.2.2 Cluster Analysis by Time of Day	4-13

TABLE OF CONTENTS (Concluded)

<u>Section</u>	<u>Page</u>
5. PMF SOURCE APPORTIONMENT	5-1
5.1 Finding the Optimal Solution	5-1
5.1.1 Gaining Convergence	5-1
5.1.2 Determining the Number of Factors and Rotation	5-2
5.2 Final Solution.....	5-4
5.3 Temporal Analyses	5-23
5.3.1 Seasonal Variations	5-24
5.3.2 Day of Week Variations	5-28
5.3.3 Time of Day Variations	5-30
5.4 Wind Direction Analysis.....	5-34
5.5 Conditional Probability Function.....	5-44
5.6 Scaling Source Profiles by Reactivity.....	5-52
5.6.1 Ozone Production Potential: Reactivity Scales	5-52
5.6.2 PMF Sources Scaled by MIR Reactivity.....	5-54
5.7 Source Strength on Ozone Episode Days	5-56
5.8 Outlying Residuals and High Ozone.....	5-58
5.9 UNMIX Solutions.....	5-60
6. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS	6-1
6.1 Database Preparation	6-1
6.2 Factor and Cluster Analysis Results	6-2
6.3 PMF Analysis.....	6-2
6.4 Future Work	6-5
7. REFERENCES	7-1
APPENDIX A: PLOTS OF RESIDUALS FOR EACH SPECIES FROM THE 15-FACTOR PMF SOLUTION	A-1
APPENDIX B: EMISSION INVENTORY MAPS OF STATIONARY SOURCES IN THE HOUSTON AREA	B-1

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1-1. Map of Houston area with auto-GC sites in green.....	1-2
1-2. Average distribution of the 15-factor PMF solution for VOC data collected at the Clinton Drive site, 1998-2001.....	1-6
1-3. Average distribution of the reactivity-weighted factors for VOC data collected at the Clinton Drive site, 1998-2001.....	1-7
3-1. Scatter plot of 2-methylpentane versus 3-methylpentane over all samples used for source apportionment; $r^2 = 0.99$	3-5
4-1. Scatter plot matrix of ethane ethene, propane, and propene in the source apportionment data set for Clinton Drive.	4-6
4-2. Cluster analysis of species at Clinton Drive in 1998-2001.....	4-12
4-3. Cluster analysis of species at Clinton Drive during the morning in 1998-2001.....	4-13
4-4. Cluster analysis of species at Clinton Drive during the afternoon/evening in 1998-2001	4-14
5-1. Q values as a function of FPeak for a 15-factor solution.....	5-3
5-2. Reconstructed PMF mass versus measured mass in ppbC for the 15-factor solution with $F_{peak} = 0.2$	5-3
5-3. Average contribution of each factor to the total mass	5-8
5-4. Percent of each species and the percent of mass from each species in Factor 1.....	5-9
5-5. Percent of each species and the percent of mass from each species in Factor 2.....	5-10
5-6. Percent of each species and the percent of mass from each species in Factor 3.....	5-11
5-7. Percent of each species and the percent of mass from each species in Factor 4.....	5-12
5-8. Percent of each species and the percent of mass from each species in Factor 5.....	5-13
5-9. Percent of each species and the percent of mass from each species in Factor 6.....	5-14
5-10. Percent of each species and the percent of mass from each species in Factor 7.....	5-15
5-11. Percent of each species and the percent of mass from each species in Factor 8.....	5-16
5-12. Percent of each species and the percent of mass from each species in Factor 9.....	5-17

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
5-13. Percent of each species and the percent of mass from each species in Factor 10.....	5-18
5-14. Percent of each species and the percent of mass from each species in Factor 11.....	5-19
5-15. Percent of each species and the percent of mass from each species in Factor 12.....	5-20
5-16. Percent of each species and the percent of mass from each species in Factor 13.....	5-21
5-17. Percent of each species and the percent of mass from each species in Factor 14.....	5-22
5-18. Percent of each species and the percent of mass from each species in Factor 15.....	5-23
5-19. Illustration of a box whisker plot and a notched box whisker plot as defined by SYSTAT statistical software.....	5-24
5-20. Notched box whisker plots of Factors 1-4 weight percent by month	5-26
5-21. Notched box whisker plots of Factors 5-8 weight percent by month	5-26
5-22. Notched box whisker plots of Factors 9-12 weight percent by month	5-27
5-23. Notched box whisker plots of Factors 13-15 weight percent by month	5-27
5-24. Notched box whisker plots of Factors 1-4 weight percent by day of week	5-28
5-25. Notched box whisker plots of Factors 5-8 weight percent by day of week	5-29
5-26. Notched box whisker plots of Factors 9-12 weight percent by day of week	5-29
5-27. Notched box whisker plots of Factors 13-15 weight percent by day of week	5-30
5-28. Notched box whisker plots of hourly weight percents of Factors 1 through 4	5-32
5-29. Notched box whisker plots of hourly weight percents of Factors 5 through 8	5-32
5-30. Notched box whisker plots of hourly weight percents of Factors 9 through 12	5-33
5-31. Notched box whisker plots of hourly weight percents of Factors 13 through 15	5-33
5-32. Notched box whisker plots of concentrations of Factors 5 and 6 by hour	5-34
5-33. Concentration of Factor 15 by hour and by wind octant	5-34
5-34. Median concentration and weight percent of Factor 1 by wind direction	5-36

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
5-35. Median concentration and weight percent of Factor 2 by wind direction	5-37
5-36. Median concentration and weight percent of Factor 3 by wind direction	5-37
5-37. Median concentration and weight percent of Factor 4 by wind direction	5-38
5-38. Median concentration and weight percent of Factor 5 by wind direction	5-38
5-39. Median concentration and weight percent of Factor 6 by wind direction	5-39
5-40. Median concentration and weight percent of Factor 7 by wind direction	5-39
5-41. Median concentration and weight percent of Factor 8 by wind direction	5-40
5-42. Median concentration and weight percent of Factor 9 by wind direction	5-40
5-43. Median concentration and weight percent of Factor 10 by wind direction	5-41
5-44. Median concentration and weight percent of Factor 11 by wind direction	5-41
5-45. Median concentration and weight percent of Factor 12 by wind direction	5-42
5-46. Median concentration and weight percent of Factor 13 by wind direction	5-42
5-47. Median concentration and weight percent of Factor 14 by wind direction	5-43
5-48. Median concentration and weight percent of Factor 15 by wind direction	5-43
5-49. Factor 14 concentrations by wind direction by day and night	5-44
5-50. CPF of Factor 1 by concentration and weight percent.....	5-45
5-51. CPF of Factor 2 by concentration and weight percent.....	5-45
5-52. CPF of Factor 3 by concentration and weight percent.....	5-46
5-53. CPF of Factor 4 by concentration and weight percent.....	5-46
5-54. CPF of Factor 5 by concentration and weight percent.....	5-47
5-55. CPF of Factor 6 by concentration and weight percent.....	5-47
5-56. CPF of Factor 7 by concentration and weight percent.....	5-48

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
5-57. CPF of Factor 8 by concentration and weight percent.....	5-48
5-58. CPF of Factor 9 by concentration and weight percent.....	5-49
5-59. CPF of Factor 10 by concentration and weight percent.....	5-49
5-60. CPF of Factor 11 by concentration and weight percent.....	5-50
5-61. CPF of Factor 12 by concentration and weight percent.....	5-50
5-62. CPF of Factor 13 by concentration and weight percent.....	5-51
5-63. CPF of Factor 14 by concentration and weight percent.....	5-51
5-64. CPF of Factor 15 by concentration and weight percent.....	5-52
5-65. Total reactivity by factor.....	5-55
5-66. Median source strength on mornings of ozone episodes and non-episodes June-September 1998-2001	5-57
5-67. Notched box whisker plots of Factors 1 and 2 weight percent on non-episode and episode mornings during June-September 1998-2001	5-58
5-68. Notched box whisker plots of Factors 3 and 10 weight percent on non-episode and episode mornings during June-September 1998-2001	5-58
5-69. Notched box whisker plots of Factors 11 and 12 weight percent on non-episode and episode mornings during June-September 1998-2001	5-58
5-70. Residuals of ethene and propene versus ozone concentration during May-October 1998-2001.....	5-59
5-71. Residuals of 1,3-butadiene and trans-2-butene versus ozone concentration during May-October 1998-2001	5-59
5-72. Residuals of toluene and m/p-xylenes versus ozone concentration during May-October 1998-2001.....	5-60
5-73. Percent of each species in UNMIX Factor 1	5-62
5-74. Percent of each species in UNMIX Factor 2	5-62
5-75. Percent of each species in UNMIX Factor 3	5-63

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
5-76. Percent of each species in UNMIX Factor 4	5-63
5-77. Percent of each species in UNMIX Factor 5	5-64
A-1. Scaled residuals of ethane (ethan), ethene (ethyl), propane (propa) and propene (prpyl).....	A-3
A-2. Scaled residuals of isobutane (isbta), n-butane (nbuta), acetylene (acety) and t-2-butene (t2bte).....	A-3
A-3. Scaled residuals of isopentane (ispna), n-pentane (npnta), t-2-pentene (t2pne) and 2,2-dimethylbutane (v22dmb).....	A-4
A-4. Scaled residuals of 2-methylpentane (v2mpna), isoprene (ispre), n-hexane (nhexa) and methylcyclopentane (mcpna)	A-4
A-5. Scaled residuals of benzene (benz), cyclohexane (cyhxa), 2-methylhexane (v2mhxa) and 2,2,4-trimethylpentane (v224tmp)	A-5
A-6. Scaled residuals of n-heptane (nhept), methylcyclohexane (mcyhx), toluene (tolu) and 2-methylheptane (v2mhhep)	A-5
A-7. Scaled residuals of n-octane (noct), ethylbenzene (ebenz), m/p-xylenes (m_pxy) and n-nonane (nnon)	A-6
A-8. Scaled residuals of n-propylbenzene (npbz), m-ethyltoluene (metol), 1,3,5-trimethylbenzene (v135tmb) and o-ethyltoluene (oetol).....	A-6
A-9. Scaled residuals of 1,2,4-trimethylbenzene (v124tmb), n-decane (ndec), 1,2,3-trimethylbenzene (v123tmb) and p-diethylbenzene (pdeben).....	A-7
A-10. Scaled residuals of n-undecane (nundc), 1,3-butadiene (v13buta), and unidentified (uidvoc).....	A-7
B-1. Map of designated emission sections in the Houston area (1-8, plus 9a and 9b)	B-3
B-2. Emission inventory map of stationary sources of ethene in the Houston area	B-4
B-3. Emission inventory map of stationary sources of propene in the Houston area.....	B-5
B-4. Emission inventory map of stationary sources of butenes in the Houston area.....	B-6
B-5. Emission inventory map of stationary sources of 1,3-butadiene in the Houston area.....	B-7
B-6. Emission inventory map of stationary sources of pentenes in the Houston area.....	B-8

LIST OF FIGURES (Concluded)

<u>Figure</u>	<u>Page</u>
B-7. Emission inventory map of stationary sources of toluene in the Houston area	B-9
B-8. Emission inventory map of stationary sources of xylenes in the Houston area.....	B-10
B-9. Emission inventory map of stationary sources of ethyltoluene in the Houston area	B-11
B-10. Emission inventory map of stationary sources of trimethylbenzenes in the Houston area	B-12

LIST OF TABLES

<u>Table</u>	<u>Page</u>
3-1. Number of data records by year in which all species were missing at Clinton Drive in 1998-2001.	3-1
3-2. Number of data missing or reported as 0 concentrations for selected species, significant periods when these data occurred, and the action taken.	3-2
3-3. Species not included in factor analysis and source apportionment, the reasoning behind using a surrogate species and that surrogate species' correlation with the excluded species.....	3-4
3-4. AIRS code, abbreviation, hydrocarbon name, and species group for the species used in receptor modeling tasks.	3-5
3-5. Number of records by season in the data set.....	3-7
3-6. Number of records by season in the data set.....	3-7
3-7. Number of records by day of week in the data set.....	3-7
3-8. Number of records by hour in the data set. Note that hours between 0000 and 0300 are times when the auto-GC is in calibration mode, and is not sampling.....	3-7
4-1. Factors, percent of variance the factor accounts for, key species in the factor, and likely sources at Clinton Drive 1998-2001.	4-2
4-2. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in 1998.	4-4
4-3. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in 1999.	4-4
4-4. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in 2000.	4-5
4-5. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in 2001.	4-5
4-6. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in June-September 1998-2001.	4-7
4-7. Factors, percent of variance which are account for by the factor, key species in the factor, and likely sources at Clinton Drive in 1998-2001 during the morning.....	4-8

LIST OF TABLES (Concluded)

<u>Table</u>	<u>Page</u>
4-8. Factors, percent of variance which are account for by the factor, key species in the factor, and likely sources at Clinton Drive in 1998-2001 during the afternoon/evening	4-9
4-9. Number of factors, percent variance accounted for by the factor, and what species that were not included in a factor by wind octant at Clinton Drive in 1998-2001	4-10
4-10. Factors, their likely source, and in what wind octant they were identified at Clinton Drive in 1998-2001	4-10
5-1. Defaults, general range, and final settings used in this work for outlier influence, iteration control, chi2 variation, and maximum number of steps	5-2
5-2. R2 value of the reconstructed mass versus expected mass for solution by number of factors used	5-2
5-3. Important species, average % of the total mass, and likely source of each factor identified by PMF	5-7
5-4. Reactivity values for selected hydrocarbons.....	5-54
5-5. Results of two-sample t-tests for each factor in the June-September 1998-2001, 0500-0900 CST: whether episode or non-episode median weight percents are higher and whether these differences are different at a 95% confidence level.....	5-57
5-6. Species used in UNMIX and their abbreviations.....	5-61
6-1. Summary of the 15-factor solution for PMF using 1998-2001 auto-GC data collected at Clinton Drive	6-4

1. INTRODUCTION

1.1 BACKGROUND

As a part of monitoring efforts to better understand ozone precursor concentrations and composition in the Houston area, the Texas Commission on Environmental Quality (TCEQ) collected hourly speciated volatile organic hydrocarbon (VOC) data at several sites for several years using automatic gas chromatographs (auto-GCs, **Figure 1-1**). The VOC data are collected to assess the characteristics (e.g., composition, ozone formation potential) of VOCs, investigate spatial and temporal variability in VOCs, and assess the capability of models to simulate the conditions that lead to episodes of high ozone concentrations. The auto-GCs record hourly concentrations of nearly 60 hydrocarbons. Other air quality measurements (such as ozone and NO_x) and meteorological data are collocated at these sites. In previous work assignments, Sonoma Technology, Inc. (STI) acquired, validated, and characterized the 1998-2001 auto-GC data (Brown and Main, 2002).

1.2 PURPOSE

In this work assignment, exploratory source apportionment of the Clinton Drive auto-GC data was performed to explore possible emission source types of VOCs in the Houston Ship Channel (HSC). Factor and cluster analyses and two enhanced factor analysis tools, Positive Matrix Factorization (PMF) and UNMIX, were used to identify likely sources. The factor analysis tools provide factors which can be related to emission source types and which estimate the quantitative contribution of each factor in every sample. Thus, the variation of source strength by time of day, day of week, and wind direction can be explored.

This report details the assembly of the data for source apportionment, provides the preliminary results of the factor and cluster analyses, and summarizes the final results from PMF and UNMIX, including all details from the interim report (Brown and Hafner, 2003). The results are exploratory in nature because we did not exhaust all options in the factor analyses; suggestions and ideas for future analyses are included in Section 6.4, Future Work.

1.3 SOURCE APPORTIONMENT

1.3.1 Overview

Receptor modeling is a mathematical procedure for identifying and quantifying the sources of ambient air contaminants at a receptor, primarily on the basis of ambient concentration measurements at that receptor (also called source apportionment). Multivariate receptor models require the input of data from multiple samples and extract the source apportionment information from all of the sample data simultaneously. The reward for the extra complexity of these models is that they estimate not only the source contributions but also the source compositions (profiles). Two such models are PMF and UNMIX which are based on



Figure 1-1. Map of Houston area with auto-GC sites in green.

factor analysis. In recent years, the development of PMF and UNMIX and subsequent applications to hydrocarbon data have been supported by the U.S. Environmental Protection Agency (EPA).

Source apportionment techniques are based on similar assumptions and needs:

- The composition of source emissions is assumed not to change during travel from the point of emission (where the source profile is defined) to the point of receptor site measurements. While less of an issue with receptor-based models, the analyst needs to understand the potential changes to the emission source composition that occur while in transit.
- Measurement uncertainties are assumed to be random, uncorrelated, and normally distributed; the effects of deviations from this assumption are unknown.
- The models require complete samples (i.e., a concentration for every VOC in the sample) from which to work.
- The models require that species concentrations vary. The practical implication of this assumption is that the analyst should not include highly collinear species in the data set.
- The models require reasonable estimates of the uncertainty associated with the ambient concentration measurements.

1.3.2 Approach to Source Apportionment

An example approach to source apportionment is outlined in the PAMS Data Analysis Workbook (Main and Roberts, 2000) as follows:

1. Understand the airshed geography and topography using maps, photographs, site visits, etc.
2. Investigate the size, composition, and location of emission sources.
3. Understand the typical meteorology of the site, including diurnal and seasonal variations.
4. Investigate the spatial and temporal characteristics of the data, including meteorological dependence.
5. Investigate the relationships among species using scatter plot matrices, correlation matrices, and other statistical tools.
6. Apply cluster and factor analysis techniques using standard statistical packages to get an overall understanding of species relationships and groupings by time of day, day of week, season, episode, etc.
7. Apply UNMIX to investigate the possible number of factors and source profiles.
8. Apply PMF using the number of factors determined by UNMIX and/or factor analysis to obtain source profiles with more species, detailed mass apportionment, and the temporal variation in source strengths.

9. Apply the chemical mass balance (CMB) model using standard source profiles and using source profiles from PMF.
10. Evaluate and compare results between the three source apportionment methods.

The first five suggested steps have been performed in previous work assignments by STI (Brown and Main, 2002; Main et al., 2001; Main and Brown, 2002a), TCEQ, and other contractors. This work assignment focuses on cluster, factor, and PMF analysis applications.

1.3.3 Source Apportionment Tools

Factor analysis, cluster analysis, UNMIX, and PMF are all useful tools in examining data. Of these, only UNMIX and PMF are able to develop detailed source profiles from the ambient data. However, PMF generally allows for more data points and species to be used than UNMIX, by utilizing samples with missing or below-detection data, which would be discarded by UNMIX. One of PMF's greatest strengths is its ability to consider each individual data point individually by using uncertainties tailored to each species' concentration in every sample, something UNMIX is not able to do. Therefore, PMF was chosen to be the focus of this exploratory source apportionment work. Factor analysis, cluster analysis, and UNMIX are used to supplement the PMF efforts and potentially give alternative views on the data.

1.4 KEY FINDINGS

Several key findings resulted from this work:

- *This exploratory application of PMF to an hourly PAMS VOC data set was successful in identifying and quantifying VOC sources in the Houston Ship Channel area.* This hourly data set was particularly useful when apportioning sources by wind direction, time of day, day of week, and season. Combining individual species reactivity with PMF source profiles provided additional insights into the relative importance of source contributions to ozone formation.
- *Fresh emissions occur all day.* Only small differences between morning and afternoon factor analyses were found, suggesting that fresh emissions occur all day (as observed by Brown and Main, 2002) and that depletion of reactive species by atmospheric reactions should not overly interfere with the source apportionment.
- *The auto-GC data provide a rich and useful database for receptor-based source apportionment.* Diurnal and wind direction patterns in factor strength were useful in aiding in the identification of factors; these patterns can only be inspected fully with hourly data.
- *The mix of VOC sources at the Clinton Drive site is complex.* Significant differences were found in both the number and composition of factors (sources) by wind direction, illustrating the complex mixture of emissions that impact the Clinton Drive site. The best PMF solution identified 15 factors (**Figure 1-2**), which included the following:

- A motor vehicle factor (on average 4% of total mass) and a diesel factor (on average 2% of total mass). A mixed aromatic factor was also found (on average 12% of total mass) that likely has some motor vehicle influence.
- An industrial flare factor comprised of acetylene, ethane, ethane, and n-butane was identified. While there may be some mobile source influence in this factor, factor strength does not decrease on weekends, indicating it is mainly from stationary sources.
- Factors with high concentrations of reactive olefins and aromatic hydrocarbons to the south and east.
- Source profiles were scaled by reactivity coefficients, and *no factor, compound, or compound class (such as olefins) dominated the overall reactivity potential*; this finding (**Figure 1-3**) is consistent with earlier results. Dramatic shifts in importance are found when the factor compositions are weighted by reactivity, including a decrease in importance of the background+fresh and evaporative factors (5, 10, 11) and an increase in importance of the C2-C5 olefins (Factors 4, 7, 15) and aromatic hydrocarbons (Factor 8).
- *Only six factors contributed more to the VOC mix on mornings of ozone episodes (by median weight percent) at a 95% confidence level than on other mornings.* These included industrial flare (Factor 1), heavy aromatic hydrocarbons (2), motor vehicles (3), solvents (10), light paraffins (11), and industrial/mobile aromatic hydrocarbons (12). This analysis highlights that *aromatic hydrocarbons may be more important* than previously thought in ozone formation. While the high concentrations of the more reactive compounds (e.g., light olefins) appear to support a high “background” of ozone, high concentrations of the aromatic hydrocarbons may provide additional potential to form ozone and push ozone concentrations above 125 ppb. Also, the inclusion of the industrial flare factor is significant, as earlier analyses were not able to determine that industrial flare concentrations were higher on episode days.

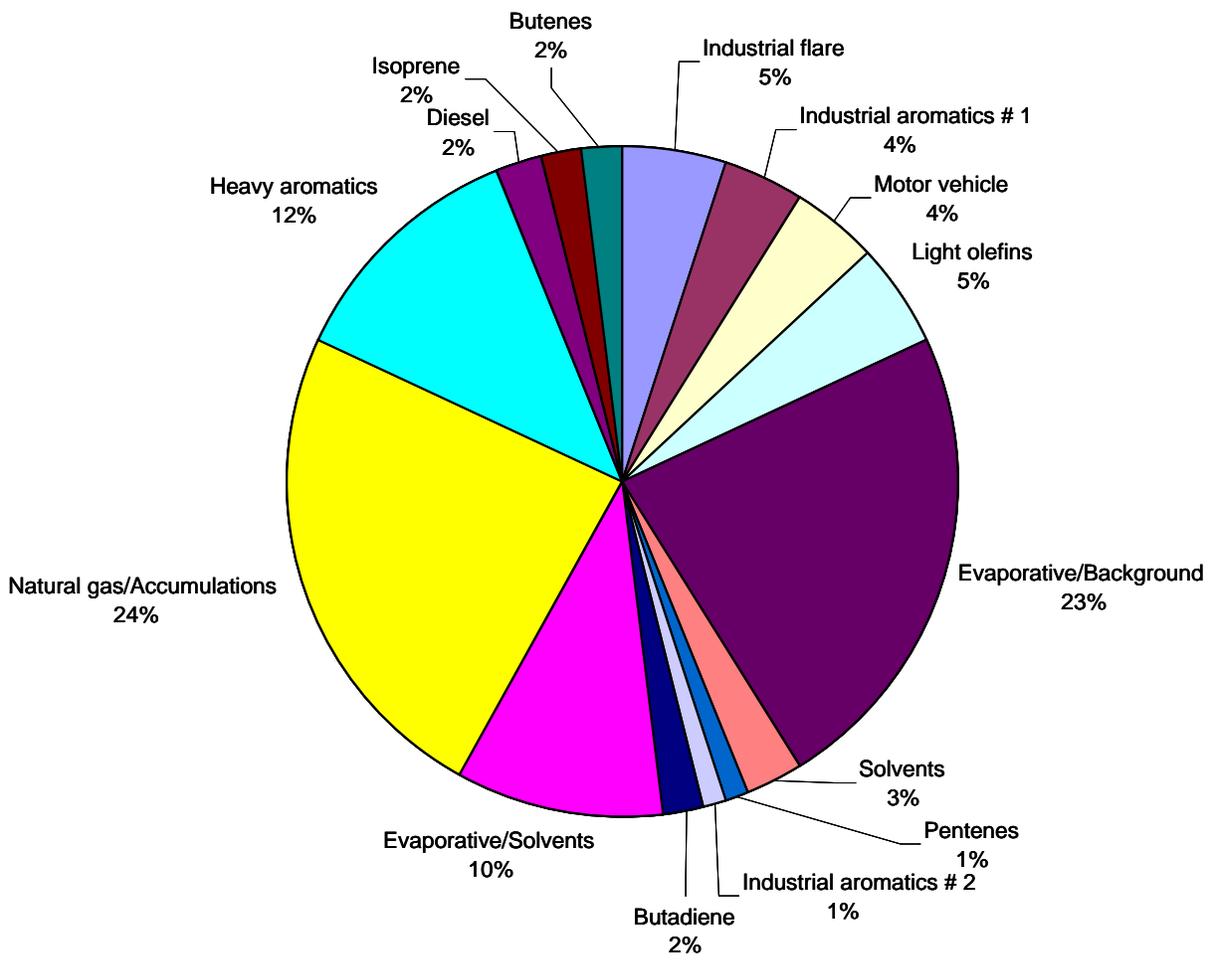


Figure 1-2. Average distribution of the 15-factor PMF solution for VOC data collected at the Clinton Drive site, 1998-2001.

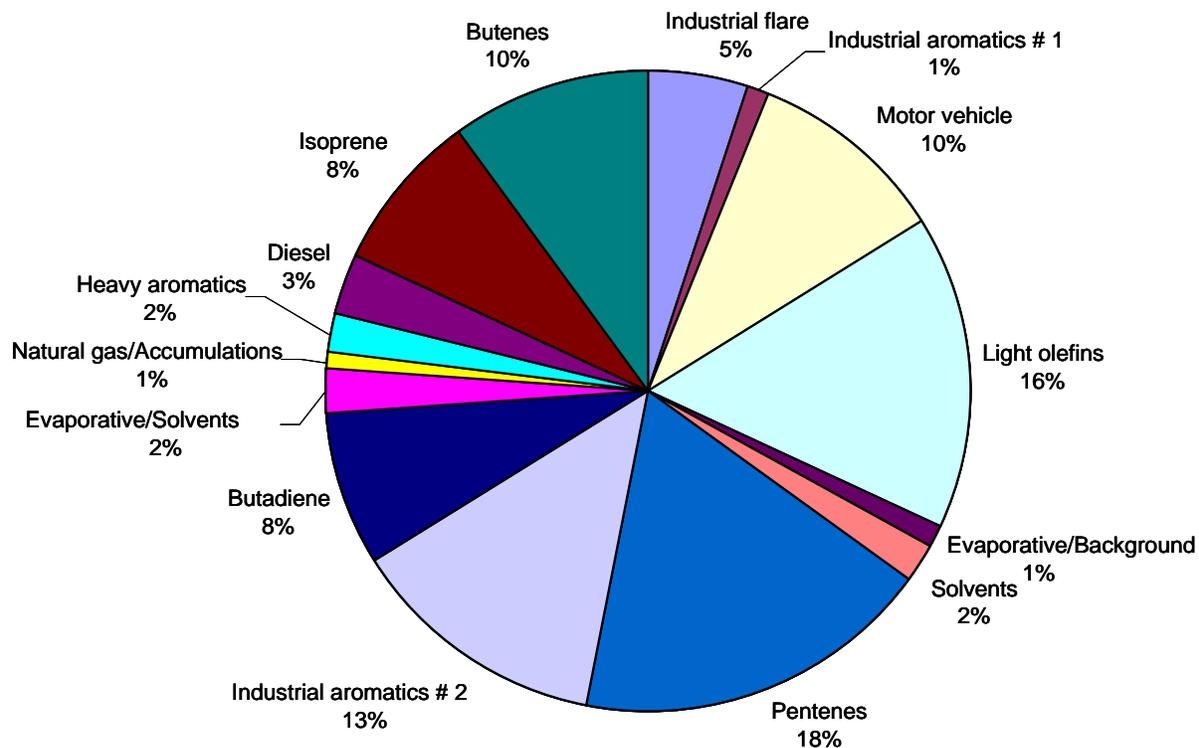


Figure 1-3. Average distribution of the reactivity-weighted factors for VOC data collected at the Clinton Drive site, 1998-2001.

1.5 OVERVIEW OF REPORT

This report presents a discussion of the workings and details of PMF (Section 2); the development of the source apportionment data set and its implications (Section 3); the detailed statistical analyses using factor and cluster analysis (Section 4); the results of the PMF solution and their implications (Section 5); and the summary, conclusions, and recommendations (Section 6). References are provided in Section 7. Appendices contain plots of residuals for each species used in PMF (Appendix A) and plots of stationary-source emissions for reactive species (Appendix B).

2. SOURCE APPORTIONMENT

The receptor-based source apportionment tool, PMF, was selected for this project. PMF has been applied to a number of data sets including PM_{2.5} data from Alaska (Polissar et al., 1998); Vermont (Polissar et al., 2001; Ramadan et al., 2000); Phoenix (Ramadan et al., 2000); the northeastern United States (Song, et al., 2001) and Hong Kong (Lee, et al., 1999). STI has applied PMF to a number of data sets, including VOCs, size-segregated PM_{2.5} species, and a combination of PM_{2.5}, semi-volatiles and VOCs at a single site.

To run PMF, an input file of concentrations by time is needed, though without headings or time index; only the species that will be used can be included. A corresponding uncertainty file is also used, which is an exact match to the input file, but with the uncertainty for each data point instead of the concentration. The added complexity of PMF also means there are a number of model parameters that need to be considered, including the number of factors to use and the treatment of outliers; these are detailed in this section.

The model outputs two files, one of which is a set of source profiles (also called fingerprints) that gives the relative distribution of each species in each factor. The analyst must then infer what source (or mix of sources) each factor represents. To do this, the analyst uses knowledge of the sources in the area, key species or unique tracers, relationships among species, and transport and transformation of pollutants. The other file is the relative strength of each factor by sample, retaining the original time index. This file can also be matched to the total mass for each sample and, through a multi-linear regression, be used to properly scale the factor strengths to the original concentration units.

This section describes the general workings and framework of PMF, including the ability of PMF to consider each individual data point separately, the optimization of the final solution by using different number of factors and rotations, the investigation of model parameters and how they influence the final solution, the technique for apportioning mass, and the various ways to examine the results.

2.1 MATHEMATICAL FRAMEWORK

The mathematical framework of PMF is described in full detail in previous work (Paatero, 1997), and is summarized here. A data set of ambient data can be viewed as a data matrix \mathbf{X} of n by m dimensions, in which n is the number of samples where m chemical species were measured. The goal of multivariate receptor modeling is to identify p number of sources, the chemical profile of each source, and the amount of mass contributed by each source in an individual sample. The model can therefore be written as

$$X = F \cdot G + E$$

or

$$x_{ij} = \sum_{h=1}^p g_{ih} f_{hj} + e_{ij}$$

where

F is a p by m matrix of source profiles for all species j

G is a p by n matrix of source contributions to each sample i (retaining its original time index)

E represents the residuals, i.e., the part of the data variance that does not fit the model

p is the number of factors

n is the number of samples

m is the number of chemical species

Results are constrained by a penalty function so that no sample can have a negative source contribution (in **G**), and that no species (in **F**) can have a negative concentration in any profile.

The goal of PMF is to minimize the Q value (the sum of squares):

$$\min_{G,F} Q(X, \mathbf{s}, G, F)$$

where

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

with

$$e_{ij} = x_{ij} - \sum_{h=1}^p g_{ih} f_{hj}$$

where g_{ik} and f_{kj} are forced to be ≥ 0 for $i = 1, \dots, n, j = 1, \dots, m, k = 1, \dots, p$, and \mathbf{s}_{ij} is the known matrix of error estimates of **X**. Simply put, this is a least squares problem in which **G** and **F** are determined in such a way that Q (and therefore e_{ij}) is minimized. PMF utilizes a unique algorithm in which both **G** and **F** are varied simultaneously in each least squares step. Paatero and Tapper (1994) and Paatero (1997) further detail this iteration sequence and a global optimization scheme in which the joint solution is directly determined. PMF can run in “robust” mode, in which the influence of extreme outliers is diminished. This is extremely useful in environmental data, in which “true” outlying concentrations often occur in a lognormal distribution. The Q function is modified in the robust mode as follows:

$$Q = \left\| \frac{(X - GF)}{\mathbf{s}} \right\|_{G,F}^2 = \sum_i \sum_j \left(\frac{e_{ij}}{h_{ij} \mathbf{s}_{ij}} \right)^2$$

where

$$h_{ij} = \begin{cases} 1 & \text{if } |e_{ij}/s_{ij}| \leq \mathbf{a}, \text{ and} \\ \text{otherwise} & |e_{ij}/s_{ij}|/\mathbf{a} \end{cases}$$

α is the outlier distance parameter (i.e., the upper limit before data points are treated as outliers). Typically values of 2.0, 4.0, and 8.0 are chosen for α , 4.0 being the default value.

2.2 ESTIMATION OF WEIGHTS

Factor analyses require complete samples (i.e., a concentration for each compound in every sample). Typically, species with a significant number of samples below the detection limit or missing are eliminated from the analysis. One of the strengths of PMF compared to other source apportionment tools such as UNMIX or principal component analysis (PCA) is that PMF can individually weigh (consider) each data point. This feature allows the analyst to adjust the influence of each data point depending on the confidence in the measurement and retain data that would otherwise be screened out. Data below detection can be retained for use in the model, with the associated uncertainty adjusted so these data points are less important to the model solution (i.e., these data have less influence on the solution than measurements above the detection limit). By individually weighing data, PMF also allows missing data to be retained; the analyst can substitute the overall mean concentration for missing data and adjust the uncertainty accordingly, so that these data also have only a small impact on the final solution. Thus, careful assembly of the data is required to prepare the most complete data set with reasonable estimates of uncertainty.

Missing data and data below the detection limit are different and need to be treated differently in PMF:

- Missing data are instances in which concentrations are not determined; thus, the concentrations are completely unknown.
- Data below detection are instances in which concentrations are below the analytical measurement detector's limit of detection; these data are often reported as zero.

Methods for replacing and developing uncertainty values for missing and below-detection-limit data are documented in previous work with PMF (e.g., Polissar, et al., 2003; Lee, et al., 2002; Poirot, et al., 2001; Polissar, et al., 1998; Polissar, et al., 2001; Ramadan, et al., 2000; Song, et al., 2001). Following these earlier works, missing values and values below the detection limit in this project were substituted according to the following:

$$\begin{array}{ll} x_{ij} = v_{ij} & \text{for concentrations above the detection limit} \\ x_{ij} = \text{MDL}_j/2 & \text{for data below the minimum detection limit (MDL)} \\ x_{ij} = \bar{v}_j & \text{for missing values} \end{array}$$

where:

x_{ij} = estimated concentration

v_{ij} = measured concentration

\bar{v}_j = mean of the measured concentration of a species over all data

i = sample

j = species

Since the solution found by PMF relies on both the concentration data and on the error estimates, these error estimates must be chosen judiciously so that they reflect the quality and reliability of each data point. As discussed above, the three types of data that are typically found are observed concentrations, data known to be below the MDL, and missing data. Error estimates that are most commonly used (e.g., Hopke, et al., 2003; Lee, et al., 2002; Poirot, et al., 2001; Polissar, et al., 1998; Polissar, et al., 2001; Ramadan, et al., 2000; Song, et al., 2001) are

$$\sigma_{ij} = 10\% * v_{ij} \quad \text{for determined values}$$

$$\sigma_{ij} = \text{MDL}_{ij}/2 + \text{MDL}_{ij}/3 \quad \text{for data below MDL}$$

$$\sigma_{ij} = 4 \bar{v}_{ij} \quad \text{for missing values}$$

where:

σ_{ij} = uncertainty

v_{ij} = measured concentration

MDL = minimum detection limit

\bar{v}_{ij} = mean of the measured concentration of a compound over all data

I = sample

J = species

The error for data below the MDL is 166%, and error for missing data is 400%. Thus, the missing and below-detection-limit data have much less weight in comparison to actual measured values, so these data are less important to the solution.

2.3 ESTIMATION OF THE NUMBER OF FACTORS

Typically, a simple factor analysis can give an estimate of the appropriate number of factors. However, this is simply a starting place. Because PMF is more robust and weights each individual data point depending on its error estimate, it is often able to discern more factors than other factor analysis tools. Four methods were used to establish a reasonable number of factors: optimizing the Q-value, examining the residuals, repeating the analysis using different starting points, and comparing the reconstructed (modeled) mass versus the measured mass.

A useful indicator of the optimum number of factors is the Q-value, which should theoretically be equal to the number of degrees of freedom, i.e., the number of data points in the array minus the total number of elements in the resultant factor matrices. Each resultant fitted data point, assuming the errors are properly estimated, should add approximately 1 to the

Q-value and be able to approach the theoretical value of Q. Therefore, it can be tempting to examine the estimated Q-value as a function of the number of factors to determine the appropriate number of factors. However, the answers still must make physical sense. The presence of missing and below-detection data, even with appropriate error estimates such as those presented earlier, still cause the calculated Q-value to deviate from its theoretical value. It has been seen in previous work (e.g., Yakovleva, et al., 1999) that once an appropriate number of factors are included in the fit, additional factors do not significantly improve the Q-value. Therefore, while examining changes in the Q-value with additional factors is useful, it should not be relied upon solely to determine the number of factors.

A large spread in the residuals, e_{ij} , generally beyond ± 3 standard deviations, indicates that the number of factors is likely incorrect. Also, groups of mostly positive or mostly negative residuals should ideally not occur. In data sets with “true” outlying values, even with adjusting the outlier parameter h_{ij} , residual values beyond ± 3 standard deviations may still occur and should be investigated further and understood.

Multiple solutions may be found depending on the starting point of the model. This is an inherent disadvantage in a least-squares approach in which, depending on the starting point, a local minimum may be found that is not necessarily the global minimum. This can be avoided in PMF by initiating random values for the **F** and **G** matrices by a different seed number. By repeating the analysis a number of times (five is generally sufficient), it can be determined if there are multiple solutions; multiple solutions indicate that the number of factors should be re-examined. One or zero alternative solutions are a good indication that the number of factors is correct.

Lastly, a multi-linear regression of the reconstructed mass from the PMF factors versus the measured mass is another good indicator of whether the number of factors is correct. This is further discussed in Section 2.1.6, but in general, if a negative coefficient for a factor arises from this analysis, it indicates that the number of factors is incorrect. While PMF is constrained so that no source emits negative mass, forcing the model to an incorrect number of factors can result in a negative mass scaling coefficient.

2.4 FACTOR ROTATIONS

Rotational ambiguity is inherent in factor analyses such as PMF, UNMIX, PCA and simple factor analysis. A unique solution may not be found even with the global minimum from the least-squares process, and no statistical criteria can be utilized to choose among the equivalent solutions. The non-negativity constraints on the system in PMF can reduce the rotational freedom, though this is often not enough to find a unique solution, and rotational ambiguity remains.

A key feature of PMF is that rotations are part of the fitting process and not applied after the extraction of factors, which is done in eigenvector-based methods such as PCA and simple factor analysis (e.g., with Varimax rotation). PCA and factor analysis have difficulty interpreting non-negativity constraints, since the rotation is done after the least-squares fit. By performing rotations during the least-squares fit, which really represent additions and subtractions (Paatero and Tapper, 1994), the combination of zero values and non-negativity

constraints reduce rotational ambiguity. This result is generally not simply a rotated image of the original unrotated result, as in PCA, but is slightly changed since the rotation is occurring simultaneously with the least-squares fitting.

One way rotations are managed in PMF is by the parameter FPeak. The default is zero, and by using a non-zero value PMF is forced to (1) subtract columns of **G** from each other and adding them to corresponding rows of **F** (“negative” direction) or (2) subtract rows of **F** from each other and add to columns of **G** (“positive” direction). This complicated procedure is described further in (Paatero, 1997). With each change in FPeak, the Q-value also changes, since the rotation is integrated into the least-squares minimization. Hopke (2003) noted that often slightly positive FPeak values give more easily interpretable results, and often the highest FPeak value before the substantial rise in Q often yields the best result. However there is no theoretical basis for choosing a particular FPeak value, or of interpreting the change in Q with FPeak, it simply is useful to inspect the range of possibilities.

2.5 OTHER PARAMETERS

In addition to FPeak, robust/non-robust mode, number of factors, and outlier distance there are other parameters that are important. One is the “seed” parameter, which designates a pseudorandom starting point for the least-squares minimization process. By changing this starting point, repeat analyses can be performed, so that it can be confirmed that a unique solution was reached. While there is only one global minimum to the solution, there are often local minima, so changing the initial value ensures that PMF searches the range of possible minima. Typically, if multiple solutions are obtained by just altering the “seed” parameter, then the selection criteria of other parameters discussed earlier are not optimal.

Another set of parameters which can be altered is the iteration control table, whose values control the rate and the final solution of each of the three stages of the model. The first set is “lims” values, which are weight coefficients for the logarithmic penalty function acting on the (non-negatively constrained) Factors **G** and **F**. These “lims” values also help ensure that extreme and unreal values do not occur. There are three lines in this control table: the first two lines control the first two stages, thereby influencing the rate and path of convergence for the final stage, and the last line controls the final result. Each stage ends when there have been a certain number of consecutive steps (i.e., “steps” parameter set to 4) where the absolute change in Q was less than a prescribed value (“chi²”). This final “chi²” value can often be low, such as 0.01, though for larger data sets (such as with multi-year hourly auto-GC data), this value needs to be much larger.

2.6 MASS APPORTIONMENT

While the results of PMF reproduce the data and are constrained so that mass contributions are non-negative, they are not properly scaled against the total measured mass. Therefore the results need to be scaled using a multiplicative scaling factor. By introducing a “1” into the equation (s_k/s_k), the results can be scaled to appropriate units:

$$x_{ij} = \sum_{h=1}^p g_{ih} f_{hj} = \sum_{k=1}^p f_{kj} \frac{s_k}{s_k} g_{ik}$$

where the sum of the source contributions g_{ik} should be equal to the total measured mass. By using a multi-linear regression of the measured mass against the source contributions, with a constant of zero, the scaling constants s_k can be determined for each source.

$$m_j = \sum_{k=1}^p s_k g_{kj}$$

where, as mentioned earlier, these s_k values must be non-negative. If the regression yields a negative value, it suggests the wrong number of factors was used. By scaling each of the g_{ik} factors by their appropriate scaling factor, the original mass units are regained. The source profiles f_{kj} must therefore be divided by s_k .

Both resultant matrices after scaling can be examined in a variety of insightful ways. The source contributions can be analyzed temporally by (1) relative mass contribution (i.e., the percent of the total mass attributed to a factor), (2) mass contribution (e.g., $\mu\text{g}/\text{m}^3$), and (3) relative source strength, i.e., mass contribution normalized so that the average of the source over all data points is unity. Source profiles can be analyzed (1) by concentration, (2) by composition (i.e. the percent of mass in a factor from each species), and (3) by species distribution (i.e. the percent of each species in each factor).

2.7 CONDITIONAL PROBABILITY FUNCTION

The conditional probability function (CPF) (Ashbaugh, et al., 1985; Kim, et al., 2002) can be used to identify in what direction high concentrations of individual sources identified by PMF are likely to originate. Sources are likely to be located in the direction of high conditional probability, since this function identifies where high concentrations originate. CPF is defined as

$$CPF = \frac{m_q}{n_q}$$

where m_q is the number of data points in the wind sector θ that are higher than the 25th percentile over all data (this can be any percentile), and n_q is the total number of data points over all data from the same wind sector. Samples of calm winds (i.e., < 1 m/s) should be excluded. In this work, 16 wind sectors of 22.5 degrees each were used.

3. DATABASE PREPARATION

Database preparation for use in receptor modeling is comprised of several steps: data assembly, data reduction (i.e., eliminating records), treatment of missing and below-detection-limit data, and selection of the hydrocarbons to be modeled. In addition to the concentration database, an accompanying uncertainty file needs to be constructed. The focus of this demonstration-level analysis is on the Clinton Drive site from which near-continuous hourly data were collected from 1998 through 2001. These data were previously validated by both TCEQ and STI (Main et al., 2001; Main and Brown, 2002c) and analyzed to assess the role of VOCs on ozone exceedances in the Houston area (Brown and Main, 2002).

3.1 DATA ASSEMBLY AND REDUCTION

The Clinton Drive data set for 1998 through 2001 should contain roughly 35,000 samples based on 1-hr average samples collected 24 hours a day, every day. There were a number of instances in which all species (i.e., the entire record or sample) were missing. These samples were removed (see **Table 3-1**). Data were missing for a variety of reasons, including calibration checks or operational downtimes.

Table 3-1. Number of data records by year in which all species were missing at Clinton Drive in 1998-2001.

Year	No. of Records in Which All Species Were Missing
1998	1627
1999	2419
2000	1484
2001	4125

Next, samples were found and documented in which most of the hydrocarbons were reported but in which species of interest were missing (listed in **Table 3-2**). TCEQ analyses have highlighted ethene, propene, 1,3-butadiene, xylenes, and toluene because of their high ozone formation potential. In previous PAMS data analyses (Main, 2001a; Main and Brown, 2002d), these species have been found to be commonly abundant in concentrations well above the detection limit in nearly all urban samples. The following adjustments were made to the data set:

- Deletion of samples in which concentrations of ethene and propene were either missing or reported as 0 (below detection).
- Deletion of samples during time periods when xylenes, benzene, and toluene were all missing.

- Deletion of samples in which the total nonmethane organic compound (TNMOC) values were not reported. TNMOC values are required in our analyses for two reasons: (1) the unidentified fraction (i.e., the difference between the sum of identified species and the total sample mass) cannot be computed unless TNMOC is available—and the unidentified contribution is one of the key variables used in the source apportionment; and (2) TNMOC is necessary as a quality control (QC) check of the source apportionment results in which the mass predicted by the model is reconstructed and compared to the measured mass.
- Deletion of samples flagged as “invalid” or “suspect” during data validation efforts.

Overall, these reductions in data resulted in 21,105 hourly records for source apportionment during the 1998-2001 period.

Table 3-2. Number of data missing or reported as 0 concentration for selected species, significant periods when these data occurred, and the action taken.

Species	No. of Samples Missing or 0	Significant Periods	Comment
Ethane	4	None	Records excluded
Ethene	763	9/1/00 – 9/30/00 (510 records, missing) 12/20/01 1300 – 12/31/01 2300 (242 records, 0)	Exclude all records missing or = 0
Propane	6	None	Records excluded
Propene	285	2/1/00 – 2/6/00 (125 records, missing) 3/9/01 – 3/10/01 (43 records, missing)	Exclude all records missing or = 0
1,3-butadiene	579 (missing only)	7/27/00 – 8/24/00 1200 (557 records)	Records excluded
m/p-xylenes	316	9/15/99 – 9/18/99 (80 records) 10/17/99 1200 – 10/21/99 1600 (69 records)	Exclude these 2 periods because benzene, xylenes and toluene were all missing
Toluene	187	9/15/99 – 9/18/99 (80 records) 10/17/99 1200 – 10/21/99 1600 (69 records)	See xylenes
Benzene	196	9/15/99 – 9/18/99 (80 records) 10/17/99 1200 – 10/21/99 1600 (69 records)	See xylenes

Table 3-2. Number of data missing or reported as 0 concentrations for selected species, significant periods when these data occurred, and the action taken.

Species	No. of Samples Missing or 0	Significant Periods	Comment
n-decane	1982 (missing only)	2/1/99 – 2/17/99 (352 records) 10/17/00 – 10/31/00 (571 records) 12/1/00 – 12/31/00 (613 records)	Not excluded
n-undecane	509 (missing only)	7/1/99 – 7/15/99 0600 (301 records) 10/17/99 1200 – 10/21/99 1600 (69 records)	Not excluded
Unidentified (missing due to no TNMOC)	2774 (no 0 concentrations)	4/1/98 – 4/16/98 (209 records) 5/1/98 – 5/30/98 (600 records) 6/2/98 – 6/6/98 (107 records) 6/10/98 – 7/3/98 (408 records) 7/15/98 – 8/2/98 (466 records) 11/1/00 – 11/30/00 (627 records)	Exclude all records in which TNMOC and unidentified were missing

3.2 TREATMENT OF MISSING DATA, DATA BELOW DETECTION, AND DEVELOPMENT OF UNCERTAINTY

As noted in Section 2.2, one of the strengths of PMF is the ability to handle missing and below-detection-limit data by adjusting the corresponding error estimates of these data points. The only exception to the method described in Section 2.2 was made for isoprene. Isoprene is predominantly from biogenic sources and is the only tracer of biogenic activity in the auto-GC's target species list. Previous work in Houston (Brown and Main, 2002) and elsewhere (Main et al., 1999a; Main et al., 1999b; Main and O'Brien, 2001; Main and Brown, 2002b) shows that isoprene exhibits a clear diurnal pattern that is different from other species. Biogenic isoprene emissions are a function of sunlight and temperature. Thus, isoprene concentrations vary monthly with biogenic activity (i.e., less biogenic emissions activity results in lower isoprene concentrations in winter). Due to these natural concentration variations, the substitution of missing isoprene data with simply the *overall* annual mean of isoprene would distort this diurnal and seasonal pattern, perhaps biasing the data to the point that the biogenic factor would be obscured in the model. Therefore, the mean concentration of isoprene by month and by hour was computed and substituted for missing isoprene data (1274 values total, or about 6% of the isoprene data) to ensure that a typical monthly diurnal pattern was left intact. Note that these data were also assigned correspondingly higher uncertainty than measured data, as described in Section 2.2.

3.3 SELECTION OF SPECIES

Not all hydrocarbons reported by the auto-GC were used in the model. Some species, such as styrene, have been shown to be unreliable due to analytical limitations (Main et al., 1999a; Main, 2001b; Main and Brown, 2002c). Also, using species that are highly collinear (i.e., that always vary together), such as 2-methylpentane and 3-methylpentane, can artificially

influence what factors are identified by the model. Therefore, only one of a pair of highly collinear species should be included in source apportionment. **Table 3-3** lists the species that were excluded from both the factor and cluster analyses and the PMF analysis. An example scatter plot of 2-methylpentane and 3-methylpentane demonstrating the extreme collinearity (r^2 of 0.99 over all 21,000 samples) of these species is shown in **Figure 3-1**.

Table 3-3. Species not included in factor analysis and source apportionment, the reasoning behind using a surrogate species and that surrogate species' correlation (r^2) with the excluded species.

Species not included	Reasonable surrogate	Reason for not including	r^2
1-butene, cis-2-butene	Trans-2-butene	Only one butene isomer is needed because these isomers are highly collinear	0.97
Cyclopentene, Cyclopentane	–	Low variance	–
1-pentene, cis-2-pentene	Trans-2-pentene	Only one pentene isomer is needed because these isomers are highly collinear	0.99
2,2-dimethylbutane	2,3-dimethylbutane	Highly collinear	0.94
3-methylpentane	2-methylpentane	Highly collinear	0.99
3-methylhexane	2-methylhexane	Highly collinear	0.99
2,3-dimethylpentane	2,4-dimethylpentane	Highly collinear	0.95
2,3,4-trimethylpentane	2,2,4-trimethylpentane	Highly collinear	0.99
3-methylheptane	2-methylheptane	Highly collinear	0.94
o-xylene	m/p-xylenes	Highly collinear	0.94
Isopropyl benzene	n-propylbenzene	Highly collinear, can coelute	0.93
p-ethyltoluene	–	Significant number of missing or below-detection data	–
Styrene	–	Significant number of missing or below-detection data, high analytical uncertainty	–

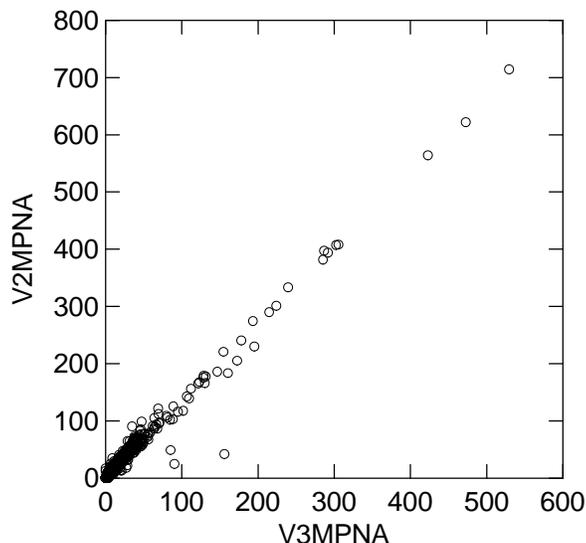


Figure 3-1. Scatter plot of 2-methylpentane (v2mpna) versus 3-methylpentane (v3mpna) over all samples used for source apportionment (21,105 records); $r^2 = 0.99$.

3.4 SUMMARY OF DATA SET FOR RECEPTOR MODELING

The resulting data set of 21,105 records contains hydrocarbon concentration data collected from 1998 through 2001 at the Clinton Drive site and accompanying uncertainty data. The species in the model include those listed in **Table 3-4**. The number of records by year, season, day of week, and hour in the data set is given in **Tables 3-5 through 3-8**; overall, the data are well-distributed across all years, seasons, days, and hours.

Table 3-4. AIRS code, abbreviation, hydrocarbon name, and species group (O=olefin, P=paraffin, A=aromatic) for the species used in receptor modeling tasks.

Page 1 of 2

AIRS code	Abbreviation	Hydrocarbon	Species Group
43206	acety	Acetylene	O
43203	ethyl	Ethylene	O
43202	ethan	Ethane	P
43205	prpyl	Propylene	O
43204	propa	Propane	P
43214	isbta	Isobutane	P
43212	nbuta	n-Butane	P
43216	t2bte	trans-2-Butene	O
43221	ispna	Isopentane	P
43220	npnta	n-Pentane	P

Table 3-4. AIRS code, abbreviation, hydrocarbon name, and species group (O=olefin, P=paraffin, A=aromatic) for the species used in receptor modeling tasks.

AIRS code	Abbreviation	Hydrocarbon	Species Group
43243	ispre	Isoprene	O
43226	t2pne	trans-2-Pentene	O
43284	22dmb	2,2-Dimethylbutane	P
43285	2mpna	2-Methylpentane	P
43231	nhexa	n-Hexane	P
43247	24dmp	2,4-Dimethylpentane	P
45201	benz	Benzene	A
43248	cyhxa	Cyclohexane	P
43263	2mhxa	2-Methylhexane	P
43250	224tmp	2,2,4-Trimethylpentane	P
43232	nhept	n-Heptane	P
43261	mcyhx	Methylcyclohexane	P
45202	tolu	Toluene	A
43960	2mhhep	2-Methylheptane	P
43233	noct	n-Octane	P
45203	ebenz	Ethylbenzene	A
45109	m/pxy	m/p-Xylene	A
43235	nnon	n-Nonane	P
45209	npbz	n-Propylbenzene	A
45207	135tmb	1,3,5-Trimethylbenzene	A
45208	124tmb	1,2,4-Trimethylbenzene	A
45211	oetol	o-Ethyltoluene	A
45212	metol	m-Ethyltoluene	A
45218	mdeben	m-diethylbenzene	A
45219	pdeben	p-diethylbenzene	A
45225	123tmb	1,2,3-trimethylbenzene	A
43238	ndec	n-Decane	P
43954	nundc	n-Undecane	P
43218	13buta	1,3-butadiene	O
	Uidvoc	Unidentified (TNMOC-Sum of PAMS)	

Table 3-5. Number of records by year in the data set.

Year	N Records
1998	4955
1999	5862
2000	6102
2001	4186

Table 3-6. Number of records by season in the data set.

Season	Months	N Records
Spring	Mar – May	4542
Summer	Jun – Aug	5246
Fall	Sep – Nov	5342
Winter	Dec - Feb	5975

Table 3-7. Number of records by day of week (DOW) in the data set.

DOW	N Records
Monday	2941
Tuesday	3025
Wednesday	3008
Thursday	3086
Friday	3065
Saturday	3060
Sunday	2920

Table 3-8. Number of records by hour in the data set. Note that hours between 0000 and 0300 are times when the auto-GC is in calibration mode, and is not sampling.

Hour (local time)	N Records	Hour (local time)	N Records
0000	663	1200	913
0100	625	1300	913
0200	580	1400	915
0300	622	1500	927
0400	929	1600	947
0500	929	1700	949
0600	931	1800	959
0700	925	1900	956
0800	936	2000	959
0900	906	2100	949
1000	901	2200	946
1100	901	2300	924

4. STATISTICAL ANALYSIS

Once the working data set was established, statistical analyses were completed to determine what groupings exist in the data and the estimated number of factors that make up the hydrocarbon composition. In addition to performing these analyses using the entire data set, data were analyzed by year, time of day (morning and afternoon), and wind direction to assess the spatial and temporal variation of the factors.

4.1 FACTOR ANALYSIS

There are two main goals with factor analysis: (1) determine the relationships among the measured parameters, and (2) find the number of factors/sources that explain most of the variance in the data. Factor analysis was completed using SYSTAT software with a Varimax rotation. Rotation enables the program to further interpret species loadings in individual factors, with the goal of reducing the number of factors with which each species is associated. Varimax rotation is an orthogonal rotation method that minimizes the number of variables that have high loadings on each factor, facilitating interpretation of the factor. In addition to determining potential sources, the number of factors found by factor analysis were used as a starting point for more robust source apportionment in PMF.

4.1.1 Overall Results

Factor analysis identified nine factors over the entire data set, accounting for 79.3% of the variance. While this is somewhat lower than that often achieved with PAMS data (usually 83%-89%, e.g., Brown and Hafner, 2003), the Clinton Drive site is subject to extremely high concentrations of all compounds at all times of the day, month, and year, which complicates the ability of simple factor analysis to resolve sources. Additionally, changes in industrial emissions or practices, as well as changes in gasoline content and vehicle traffic from 1998 to 2001, may further complicate results. Initially, only 7 factors were identified and benzene, n-pentane, and the unidentified mass were not included in any of these factors. This result was surprising because these species are typically strongly identified with ubiquitous sources such as automobile exhaust or industry (in the case of benzene) and evaporative emissions or solvent use (in the case of n-pentane). Expansion of the factor analysis to include 9 factors resulted in the inclusion of benzene and n-pentane and an increase in the variance accounted for; however, the unidentified mass was not included in any factor even when the model was forced to find 10 or 11 factors.

The nine-factor solution is detailed in **Table 4-1**. One factor contains heavy aromatic hydrocarbons, likely from a motor vehicle/industrial emissions combination, which accounts for 24% of the overall variance. Because major freeways exist in the heavily industrialized HSC, the sources are likely difficult for the simple factor analysis to separate. Also, as noted in earlier reports (Brown and Main, 2002), a bias likely exists in the analytical technique that makes factor analysis group compounds that elute close together in the GC. A second factor contains some species associated with motor vehicle emissions (accounting for 17.2% of the variance), such as

toluene, n-hexane, and methylpentanes, but the absence of acetylene and xylenes in the factor suggests that this factor may include an industrial source.

Other factors include

- A grouping of C4 and C5 alkanes and alkenes, which may be due to an analytical bias or indicative of an industrial olefin/paraffin¹ source.
- A group of C2 and C3 olefins and paraffins, also likely indicative of an industrial olefin source.
- Heavy (C9-C11) alkanes, possibly diesel emissions or an industrial sources.
- Biogenic (isoprene) accounting for 3% of the overall variance, demonstrating that the bulk of VOCs at Clinton Drive are anthropogenic. This high reactivity of isoprene may also cause this low number.
- Separate 1,3-butadiene only and benzene only factors. 1,3-butadiene is predominately from industrial sources and has a much faster reaction rate than other analyzed VOCs, which support its isolation in a factor. Benzene is emitted from a variety of sources, and often in different proportions to other aromatic hydrocarbons such as toluene and xylenes, which may be the cause of its placement in a factor by itself. Another possibility is that a significant benzene source exists near Clinton Drive that overwhelms any benzene signature from other sources; using wind analysis with PMF results may address this possibility.
- A grouping of butanes and 2,2,4-trimethylpentane, likely from evaporative emissions.

Table 4-1. Factors, percent of variance the factor accounts for, key species in the factor, and likely sources at Clinton Drive, 1998-2001.

Factor #	% Variance	Key Species	Likely Source
1	24.0	Acetylene, xylenes, ethane, trimethylbenzenes, ethyltoluenes, n-decane	Motor vehicles + industrial aromatic hydrocarbons
2	17.2	Toluene, n-hexane, methylpentanes	Industrial?
3	9.2	i- and n-pentane, pentenes, butenes, n-butane	Olefin/paraffin source – industrial
4	8.2	Propane, ethane, propene, ethene	Light olefin/ paraffin – industrial
5	5.8	C9, C10, C11 alkanes	Heavy alkane source – industry, diesel
6	3.2	Isoprene	Biogenic
7	3.3	1,3-butadiene	Industrial
8	5.5	i- and n-butane, 2,2,4-trimethylpentane	Evaporative
9	2.9	Benzene	Industrial

¹ Note that olefin/paraffin is synonymous with alkene/alkane. Alkanes are saturated hydrocarbons and alkenes are unsaturated hydrocarbons with one double bond.

4.1.2 Factor Analysis by Year

Factor analysis was performed on the data from the Clinton Drive site separated by year to investigate whether the results (i.e., the factors) varied significantly by year. Differences in factors could be the result of changes in the chemical characterization of emissions from year to year. Ten factors accounting for 85.4% and 83.1% of the variance in 2000 and in 1999, respectively, were found while 9 factors accounting for 84.7% of the variance in 2001 and 8 factors accounting for 81.3% of the variance in 1998 were found. The factors that were found each year were similar in composition and in variance explained. The range in number of factors is small (8 to 10) and seems acceptable. The Clinton Drive site is subject to sundry sources that vary in emission strength throughout the year as observed in the concentration data. This analysis also indicates there are multiple small but significant sources of VOCs.

Details on factors for each year are shown in **Tables 4-2 through 4-5**. A number of common factors were found in each year; often a change in the number of factors between years can be attributed to slight changes in the groupings of species that result in more or fewer factors. One common factor that accounted for the most variance in each year was a heavy aromatic hydrocarbon factor (including trimethylbenzenes and ethyltoluenes) associated with motor vehicle species such as acetylene, toluene, and/or xylenes. The combination of the heavy VOCs with species mostly associated with motor vehicles suggests that emissions from both industrial and mobile sources are likely emitted from the same direction.

Another common factor is one of light olefins (ethene, propene) and paraffins (ethane, propane). These hydrocarbons are often emitted in high concentrations from industrial sources in the HSC and tend to be present in high concentrations in the same air parcels. However, as shown in **Figure 4-1**, there are a number of large outliers in which only one or two of the species are found in high concentrations together, which distorts the general correlation. Other combinations include butanes and pentanes, which are from multiple sources but often correlate, and heavy alkanes, n-decane and n-undecane, which are significant in diesel exhaust and industrial activities. Common among all years were factors that included only isoprene (a likely biogenic signature), 1,3-butadiene (an industrial signature—1,3-butadiene has a much lower residence time in the atmosphere than the other PAMS target hydrocarbons), and benzene (which is emitted from both mobile and industrial sources).

The similarity among the factor results from year to year shows that combining the data from all years seems to be a reasonable approach.

Table 4-2. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in 1998.

Factor #	% Variance	Key Species	Likely Source
1	27.0	Ethyltoluenes, acetylene, trimethylbenzenes, xylenes, C10 and C11 alkanes, unidentified, 2,2,4-trimethylpentane	Motor vehicles + industrial aromatic hydrocarbons
2	20.6	Toluene, unidentified, methylpentane, C5-C7 alkanes	Industrial solvent use?
3	9.4	i- and n-butane, butenes, pentenes, i-pentane	Evaporative emissions; olefin/paraffin industrial source
4	8.7	Propane, ethane, propene, ethene	Light olefin/paraffin – industrial
5	5.8	C9, C10, C11 alkanes	Heavy alkane source – industry, diesel
6	3.4	Isoprene	Biogenic
7	3.3	1,3-butadiene	Industrial
8	3.1	Benzene	Benzene source

Table 4-3. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in 1999.

Factor #	% Variance	Key Species	Likely Source
1	23.0	Ethyltoluenes, acetylene, trimethylbenzenes, C10 and C11 alkanes, unidentified, 2,2,4-trimethylpentane, ethane	Motor vehicles + industrial aromatic hydrocarbons
2	18.2	C4-C7 alkanes, pentene, methylpentanes, methylhexane, 2,2,4-trimethylpentane,	Evaporative emissions; industrial solvent use?
3	9.3	Propane, ethane, propene, ethene	Light olefin/ paraffin – industrial
4	3.5	1,3-butadiene	Industrial source
5	6.6	C9, C10, C11 alkanes	Heavy alkane source – industry, diesel
6	3.2	Isoprene	Biogenic
7	7.6	Butanes, pentanes	Evaporative emissions
8	3.2	Xylenes	Xylenes source
9	5.4	Toluene, methylheptane	Solvent use?
10	3.1	Benzene	Benzene source

Table 4-4. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in 2000.

Factor #	% Variance	Key Species	Likely Source
1	25.6	Ethyltoluenes, acetylene, trimethylbenzenes, toluene, ethane, xylenes	Motor vehicles + industrial aromatic hydrocarbons
2	14.5	C4-C7 alkanes, methylpentanes, methylhexane, butene, pentene, C9 alkanes	Evaporative emissions; industrial solvent use?
3	7.9	Propane, ethane, propene, ethene	Light olefin/ paraffin – industrial
4	7.6	butanes, 2,2,4-trimethylpentane	Evaporative emissions
5	3.8	1,3-butadiene	Industrial source
6	6.4	C10-C11 alkanes	Heavy alkane source – industry, diesel
7	9.8	Pentanes, hexane	Evaporative emissions?
8	3.3	Isoprene	Biogenic
9	3.7	Isobutane	?
10	3.0	Benzene	Benzene source

Table 4-5. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in 2001.

Factor #	% Variance	Key Species	Likely Source
1	25.9	Ethyltoluenes, acetylene, trimethylbenzenes, ethane, xylenes, C9, C10 alkanes	Motor vehicles + industrial aromatic hydrocarbons
2	14.1	C4-C5 alkanes, butene, pentene, unidentified	Evaporative emissions; industrial solvent use?
3	9.8	Propane, ethane, propene, ethene	Light olefin/ paraffin – industrial
4	8.2	n-butane, toluene, 2,2,4-trimethylpentane	Evaporative emissions?
5	3.4	Isoprene	Biogenic
6	10.7	C6-C7 alkanes, methylpentane, methylhexane	Solvent use?
7	3.4	1,3-butadiene	Industrial source
8	5.8	C10-C11 alkanes	Heavy alkane source – industry, diesel
9	3.4	Benzene	Benzene source

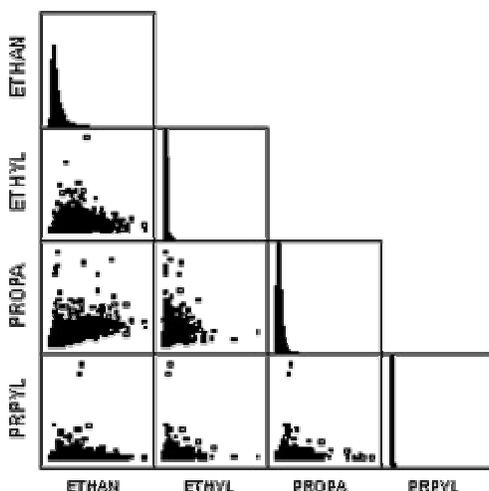


Figure 4-1. Scatter plot matrix of ethane (ethan), ethene (ethyl), propane (propa), and propene (prpyl) in the source apportionment data set for Clinton Drive.

4.1.3 Factor Analysis in Summer

Factor analysis was also performed using only data collected during the summer (June through September) of all years in order to focus on the periods when significant ozone levels occur. A total of 11 factors were identified, accounting for 81.8% of the variance. The unidentified fraction was not included in any factor, even when expanding to more factors. Results are provided in **Table 4-6**.

Many factors are similar to those identified in overall and annual factor analyses, including heavy aromatic hydrocarbon/motor vehicle, C6-C7 alkane, light olefin/alkane, 1,3-butadiene, C9-C11 alkane, isoprene, and benzene factors. In the summer, the pentenes and butenes were grouped in their own factor, while butanes and pentanes made up two other factors. In the annual analyses, these compounds were combined in a single factor. It is likely the faster removal of the olefins relative to the alkanes under the enhanced photochemical conditions of the summer lead to more differences in these species, hence, their apportionment to separate factors. Additionally, the butanes were associated with toluene and 2,2,4-trimethylpentane and may be due to more evaporative emissions in the summer. A xylenes factor was also identified, again likely due to the higher reaction rate of these aromatic hydrocarbons compared to other aromatic species.

Table 4-6. Factors, percent of variance accounted for by the factor, key species in the factor, and likely sources at Clinton Drive in June-September 1998-2001.

Factor #	% Variance	Key Species	Likely Source
1	20.0	Ethyltoluenes, acetylene, trimethylbenzenes, C10 alkanes	Motor vehicle, heavy industrial aromatic hydrocarbons
2	15.1	C6-C7 alkanes, methylpentanes, methylhexanes	Evaporative emissions or solvent use
3	8.1	Propane, ethane, propene, ethene	Light olefin/ paraffin – industrial
4	8.4	butanes, toluene, 2,2,4-trimethylpentane	Motor vehicle, evaporative emissions
5	3.3	1,3-butadiene	Industrial source
6	6.2	Pentanes	Evaporative emissions
7	6.1	C10-C11 alkanes	Heavy alkane source – industry, diesel
8	3.2	Isoprene	Biogenic
9	3.0	Xylenes	Industrial source and higher reaction rate than other aromatic hydrocarbons
10	3.0	Benzene	Benzene source
11	5.4	Butenes, pentenes	Industrial source, higher reaction rates than C4-C5 alkanes

4.1.4 Factor Analysis by Time of Day

VOC concentrations vary significantly by time of day because of changes in meteorology and emissions source activities. Generally, VOC concentrations are higher at night when mixing heights are low and photochemistry is not occurring. As the day progresses, solar radiation increases, resulting in the loss of highly reactive species relative to less reactive species. The breakup of the morning boundary layer dilutes the emissions that have been trapped overnight. Changes in wind direction influence what source emissions directly impact a site. The activities of some emission sources, such as motor vehicles, have distinct diurnal patterns. Higher temperatures in the midday can lead to increased evaporative emissions. Some biogenic emissions, such as isoprene, increase with sunlight and temperature. All these competing patterns may affect the ability of both factor analysis and PMF to correctly resolve distinct sources; thus, factor analysis was performed on the data segregated by time of day to investigate whether significant changes in factors occur. If significant differences exist between morning

and afternoon/evening due to degradation of primary emissions, data may have to be initially separated by time of day before running PMF so that sources are not obscured.

Details of factors found using data in the morning (0000 to 1100 CST) and afternoon/evening (1200 to 2000 CST) are provided in **Tables 4-7 and 4-8**. Nine factors in the morning accounted for 80.4% of the variance. N-nonane, toluene, and 2,2,4-trimethylpentane were not included in any factor, and expansion to more factors did not include the latter two species or significantly increase the total variance accounted for. Eight factors in the afternoon/evening were identified, accounting for 81.0% of the overall variance; benzene was not included in any factor, and expansion to more factors did not include benzene.

Generally, the factors were similar between the morning and afternoon. Benzene formed a small factor in the morning but was not incorporated into any factor in the afternoon. A pentane factor was found in the morning, while a butane/2,2,4-trimethylpentane factor was found in the afternoon. These compounds are emitted from both combustion and evaporative sources (e.g., motor vehicle or industrial emissions); the diurnal differences in factors may represent differences in wind patterns and boundary layer height which affect accumulation of these relatively less reactive species. The overall similarities between the morning and afternoon/evening is encouraging and indicates that *a priori* separation of data before PMF is not necessary and that atmospheric reactions will likely not complicate or obscure factors.

Table 4-7. Factors, percent of variance which are account for by the factor, key species in the factor, and likely sources at Clinton Drive in 1998-2001 during the morning (0000–1100 CST).

Factor #	% Variance	Key Species	Likely Source
1	24.2	Ethyltoluenes, acetylene, trimethylbenzenes, ethane, xylenes, C10 alkane	Motor vehicle
2	15.6	C6-C7 alkanes, methylpentanes, methylhexanes	Evaporative emissions or solvent use
3	9.4	Butenes, pentenes, butanes	Evaporative emissions?
4	8.9	Propane, ethane, propene, ethene	Light olefin/ paraffin – industrial
5	3.3	1,3-butadiene	Industrial source
6	7.1	C10-C11 alkanes, unidentified	Heavy alkane source – industry, diesel
7	3.2	Isoprene	Biogenic
8	5.8	Pentanes	Evaporative emission?
9	2.9	Benzene	Benzene source

Table 4-8. Factors, percent of variance which are account for by the factor, key species in the factor, and likely sources at Clinton Drive in 1998-2001 during the afternoon/evening (1200–2000 CST).

Factor #	% Variance	Key Species	Likely Source
1	23.6	Ethyltoluenes, acetylene, trimethylbenzenes, xylenes	Motor vehicle
2	19.8	C6-C7 alkanes, methylpentanes, methylhexanes, toluene, unidentified	Industrial/solvent
3	9.0	Butenes, pentenes, pentanes	Evaporative emissions?
4	9.0	Propane, ethane, propene, ethene	Light olefin/ paraffin – industrial
5	6.8	C9-C11 alkanes	Heavy alkane source – industry, diesel
6	3.3	Isoprene	Biogenic
7	3.3	1,3-butadiene	Industrial source
8	6.2	Butanes, 2,2,4-trimethylpentane	Evaporative emissions

4.1.5 Factor Analysis by Wind Direction

Previous analyses of auto-GC data in Houston (Brown and Main, 2002) showed significant differences in concentration levels and overall composition by wind direction. Factor analysis by wind direction may identify small but sometimes significant sources that would otherwise be obscured in an analysis of all data. In addition, similar source types that one expects to see from several wind quadrants (e.g., motor vehicle emissions) should have similar composition and diurnal profiles in each quadrant—a way of ground-truthing the solutions. A large number of small but different factors in different wind directions would suggest that more than the estimated 9 to 12 factors should be investigated in PMF. Results of the number and composition of factors found by wind direction are summarized in **Tables 4-9 and 4-10**. Each wind octant is 45°, with wind octant 1 corresponding to winds from the north (337.5°–22.5°), wind octant 2 corresponding to winds from the northeast (22.5°–67.5°), etc.

There were significant differences in both the number of factors found and the types of factors found from different wind octants, even though the same species were used, and all species were accounted for in each wind octant. Only 3 factors were found in wind octant 8, while 11 were found in wind octant 4; total variance accounted for, however, was similar at 82% and 86%, respectively. The wide variation in the number of factors by wind octant further suggests that some sources can have a significant impact on a sampling site depending on the wind direction, and that some of these sources may be obscured in the overall data set. This will further complicate PMF analysis because a large number of potential, but often small, sources seem to impact the Clinton Drive site.

Only a few common factors were found among wind octants, and even these factors differed as to which species were included. A heavy aromatic hydrocarbon/motor vehicle signature was found in each wind direction, and a light olefin/alkane source, a heavy alkane source, a C4-C5 olefin/alkane source, a C6-C8 source, 1,3-butadiene, and isoprene factors were found in many wind directions. A likely industrial isoprene source (i.e., isoprene with other compounds) was found in wind octants 6 (southwest) and 7 (west), and a 1,3-butadiene source associated with other compounds was found in wind octants 1, 6, 7, and 8. Other factors containing single VOCs, such as propene and xylenes, were found in some wind directions, suggesting a single significant source in the wind quadrant that may get obscured when using all data.

Table 4-9. Number of factors, percent variance accounted for by the factor, and what species (if any) that were not included in a factor by wind octant at Clinton Drive in 1998-2001.

Wind Octant	No. of Factors	% Variance	Species not included in any factor
1 (N)	6	86.3%	None
2 (NE)	6	83.2%	None
3 (E)	7	82.6%	None
4 (SE)	11	85.8%	None
5 (S)	9	79.7%	None
6 (SW)	6	75.4 %	None
7 (W)	5	82.0 %	None
8 (NW)	3	81.9 %	None

Table 4-10. Factors, their likely source, and in what wind octant they were identified at Clinton Drive in 1998-2001.

Page 1 of 2

Key Species	Likely Source	Found in Wind Octant
Ethyltoluenes, acetylene, trimethylbenzenes + others	Motor vehicle/ industrial	1-8
C6-C7 alkanes, methylpentanes, methylhexanes + others	Evaporative	3, 4, 5, 6
Butenes, pentenes, pentanes, butanes	Evaporative, industrial	1-3, 5, 7
Propane, ethane, propene, ethene	Light olefin/ paraffin – industrial	1-6
C10-C11 alkanes	Industrial/diesel	1-6
Isoprene	Biogenic	1-5, 8

Table 4-10. Factors, their likely source, and in what wind octant they were identified at Clinton Drive in 1998-2001.

Key Species	Likely Source	Found in Wind Octant
1,3-butadiene	Industrial	2, 3, 4, 5
1,3-butadiene with light olefins and paraffins	Industrial	1
1,3-butadiene, butene, i-butane	Industrial	6
Propene	Significant industrial source	1
Butanes	Significant industrial source	4
Pentanes	Significant industrial source	4
Butenes, pentenes, trimethylpentane	Industrial	4
Xylenes	Significant industrial source	4
Benzene	Significant industrial source	4,7
Isobutane, propane	Evaporative	5
Trimethylpentane, n-butane	Evaporative	5
Benzene, isoprene	Industrial isoprene	6
Isoprene, 1,3-butadiene, butene	Industrial isoprene	7
C2-C7 alkanes, ethene, pentene	Olefin/paraffin source	7
C2-C7 alkanes, propene, butene, 1,3-butadiene, ethene, benzene, pentene	Combination of other factors	8

4.2 CLUSTER ANALYSES

Cluster analysis is a multivariate procedure for detecting natural groupings in data. This analysis provides a graphic depiction of the relationships among data groupings, such as individual hydrocarbon species, samples collected at different sites or times of day, etc. Depending on the complexity of the hydrocarbon mix at a site, one to several clusters or factors may be needed to account for a majority of the variability in the data. SYSTAT statistical software was used to prepare cluster analyses. The clustering was computed using Pearson product-moment correlation coefficients for each pair of objects, similar to that used in factor analysis.

4.2.1 Overall Cluster Analysis

Cluster analysis was completed on the entire data set for another method to establish relationships between species that factor analysis may not have found. Cluster analysis can be helpful in determining if source apportionment results are consistent with the data set. Results for the whole data set are shown in **Figure 4-2**. To interpret the figure, consider a vertical line drawn at the arbitrary distance of 2. At this distance, many of the heavy aromatic hydrocarbons and alkanes are in one cluster, suggesting these species vary together, which is consistent with earlier factor analysis. Surprisingly, isoprene was included in the core cluster; factor analysis and knowledge of general emissions patterns in the HSC indicate that this species is from a unique source and should not be associated with other species. It may be that industrial sources of isoprene are significant enough to affect the cluster analysis. The factor analyses by wind quadrant show a likely industrial, rather than biogenic, source of isoprene.

Following are other observations from the cluster analysis results:

- The light olefins and paraffins are clustered separately from most species, consistent with factor analysis.
- The unidentified fraction is clustered separately from all species, also consistent with factor analysis.
- The butanes were not included in any cluster and may be indicative of one or several significant butane sources in the area.

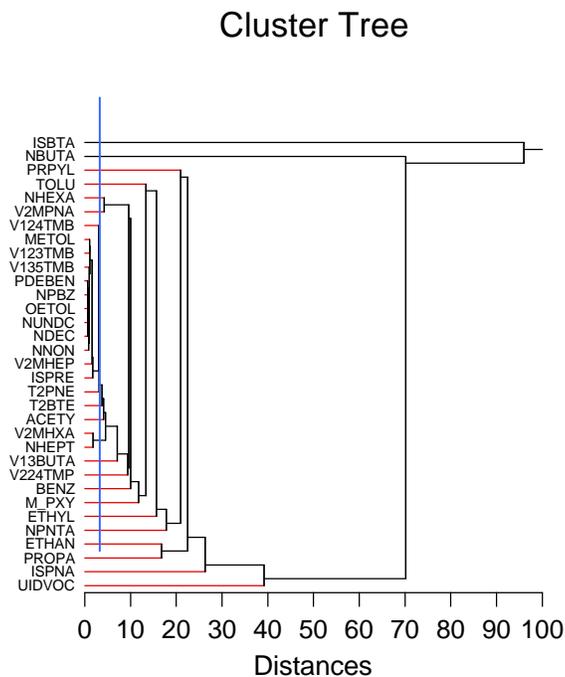


Figure 4-2. Cluster analysis of species at Clinton Drive in 1998-2001.

4.2.2 Cluster Analysis by Time of Day

Similar to factor analysis, cluster analyses were completed on the data by time of day. Cluster analyses on morning (0000-1100 CST) and afternoon/evening (1200-2000 CST) data are shown in **Figures 4-3 and 4-4**, respectively. To interpret Figure 4-3, consider a vertical line drawn at an arbitrary distance of 2. Similar to the overall results, the heavy aromatic hydrocarbons and alkanes are in this main cluster. Again, similar to the overall results, the light olefins and paraffins are somewhat removed from the main cluster, isoprene is within the main cluster, and the unidentified fraction and butanes are far removed from most species.

In Figure 4-4, consider a vertical line drawn at an arbitrary distance of 5. Heavy aromatic hydrocarbons and alkanes are in this main cluster. Isoprene, the light olefins and paraffins, unidentified hydrocarbons, and butanes all show a similar pattern as in previous analyses. Overall, the changes in the cluster distribution between morning and afternoon are minimal, which suggests that source apportionment using all data should be effective and that, due to the wealth of fresh emissions, atmospheric degradation of species should not overly affect the results.

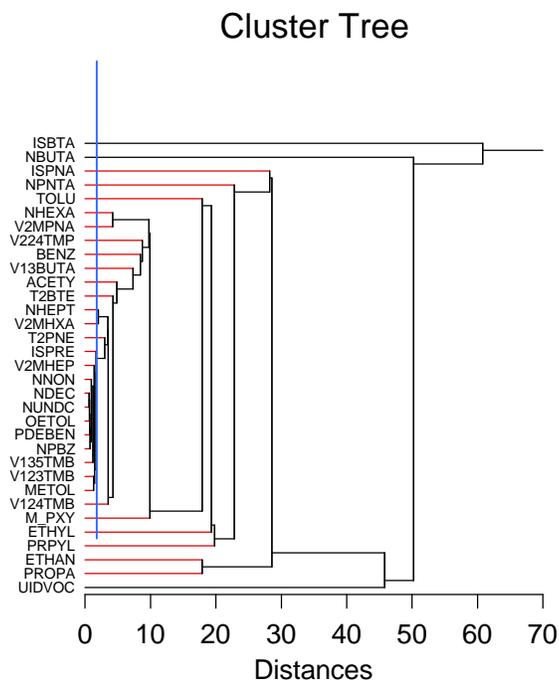


Figure 4-3. Cluster analysis of species at Clinton Drive during the morning (0000-1100 CST) in 1998-2001.

Cluster Tree

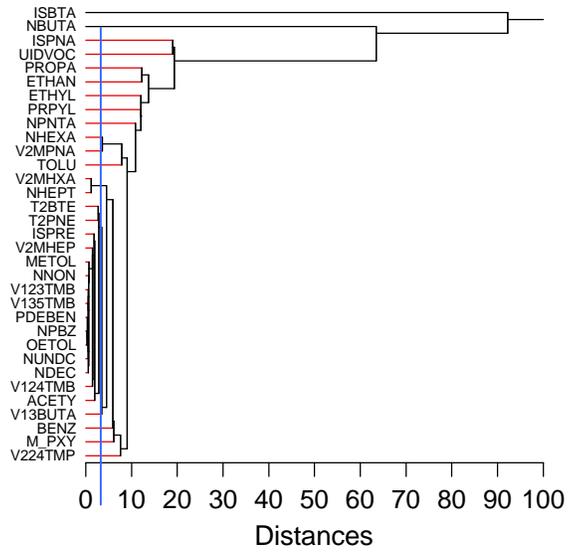


Figure 4-4. Cluster analysis of species at Clinton Drive during the afternoon/evening (1200-2000 CST) in 1998-2001

5. PMF SOURCE APPORTIONMENT

Based on the initial results found by factor and cluster analysis, the next step was to run PMF on the data set. This, however, is an iterative task, as many parameters of the model need to be adjusted and multiple scenarios run. One of the strengths of PMF is that it has a number of parameters which affect the outcome, although more time is needed to run more scenarios. This section details the results of changing various parameters, the final results, and detailed analyses of these results.

5.1 FINDING THE OPTIMAL SOLUTION

As discussed earlier and detailed in Section 1.3, PMF has a number of parameters that can be adjusted in the process of finding the optimal solution. While factor analysis and data analysis are useful in helping determine the basic structure of the data, source apportionment is still an iterative process. This section details the specific changes to parameters for the Clinton Drive data set. This data set is extremely large, more than an order of magnitude larger than what is typically used with PMF, so much of this work was exploratory.

5.1.1 Gaining Convergence

Finding convergence was a significant task; default settings often work on smaller data sets but did not work on this extremely large auto-GC data set. Gaining convergence to a model solution requires a combination of the number of factors selected, the influence of outliers, the iteration control (influencing the rates of convergence), and the prescribed change of Q (χ^2) for the incremental and final solutions. Factor analysis (Section 4) showed that factor number and source strength varies widely by wind direction, so the “true” number of factors is likely a combination of all wind directions, i.e., more than the highest number identified by factor analysis. Therefore, a range of factor numbers was explored, from 11 to 18 factors. Convergence, examination of residuals, and mass apportionment by multi-linear regression all determined whether a given number of factors was correct.

The influence of outliers on the model plays an important role in source apportionment, and especially in PMF, in which the user is able to determine the overall influence of outliers on the least-squares minimization calculations by altering the “outlier” function. Often this is set to 4.0, but it was found that the higher value of 8.0 yielded better (i.e., more easily interpretable, with a higher amount of the mass well apportioned and correlation with the expected mass) results for this data set. This is most likely due to the high number of extreme outliers of nearly all species observed in the data (Brown and Main, 2002). The default rates of convergence were too tight, and more steps were generally needed to gain convergence than is often needed for smaller data sets. Additionally, the prescribed change in Q (χ^2) for the incremental and final solutions was too tight on the default setting, as would be expected for large data sets (Paatero, 2000). These parameters were increased to various values in an iterative process to find the tightest available constraints on Q (and in theory therefore the best solution). The default and the final settings used to obtain the final solution are detailed in **Table 5-1**.

Table 5-1. Defaults, general range, and final settings used in this work for outlier influence, iteration control, χ^2 variation, and maximum number of steps. FPeak parameter was left at 0 until later in the process and is detailed in Section 4.1.2.

Parameter	Default	Range Considered	This work
Outlier	4.0	2.0, 4.0, 8.0	8.0
Chi ² (iterations 1, 2, 3)	0.5, 0.5, 0.3	0.3-10	5, 5, 0.5
Maximum cumulative number of steps (iterations 1, 2, 3)	100, 150, 200	Beyond 500 would most likely not improve solution	100, 150, 300

5.1.2 Determining the Number of Factors and Rotation

While a number of different factors were explored, most gave solutions in which convergence was not achieved without a significant increase in the χ^2 value, or in which the multi-linear regression mass apportionment coefficients were negative. Only sets of 12, 13, and 15 factors gave satisfactory results. These sets were further explored by rotation change (FPeak) and by examination of the resultant source profiles and time series.

Q values were examined for each of the three sets as a function of FPeak value. These results for the 15-factor solution are shown in **Figure 5-1**. While change in Q does not indicate which combination of factor number and FPeak is optimal, it is useful in determining what combinations result in a significant increase in Q and therefore would not be the “best” solution. By this analysis it appears that only small rotations (i.e., FPeak between -0.3 and 0.3) are useful. Additionally, the r^2 of the reconstructed mass versus the measured (expected) mass can also be of use. If the model does a poor job reconstructing the mass, it indicates the incorrect number of factors were used. Correlation coefficients of reconstructed versus actual (expected) mass for solutions of 12, 13, and 15 factors are shown in **Table 5-2**. The set of 15 factors best re-apportions the mass (shown in **Figure 5-2**), and may indicate that this is the best solution. Further analysis of source profiles and their variations in time (i.e., by hour, day of week, season) are needed to determine which set makes sense based on our understanding of emissions and atmospheric reactions in the area.

Table 5-2. R^2 value of the reconstructed mass versus expected mass for solution by number of factors used.

N factors	r^2 reconstructed mass versus expected mass
12	0.87
13	0.86
15	0.91

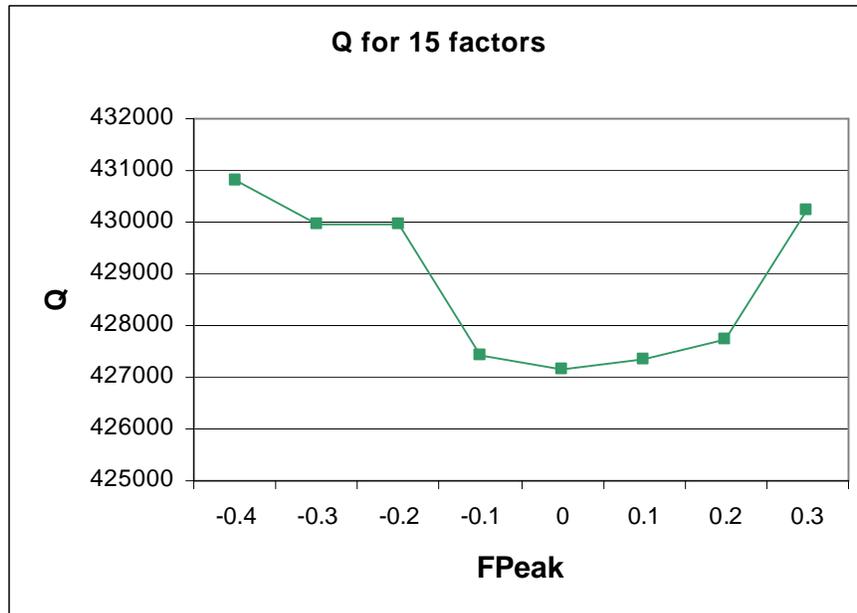


Figure 5-1. Q values as a function of FPeak for a 15-factor solution.

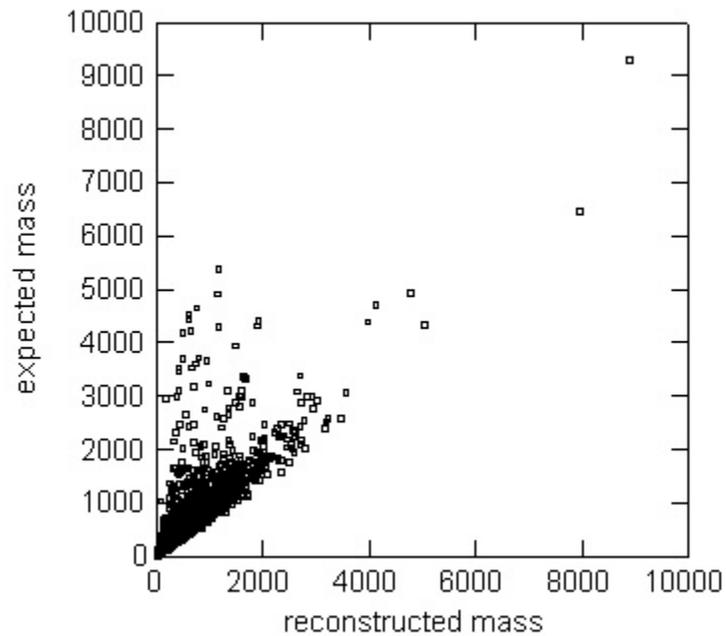


Figure 5-2. Reconstructed PMF mass versus measured (expected) mass in ppbC for the 15-factor solution with $F_{peak} = 0.2$ ($r^2=0.91$).

5.2 FINAL SOLUTION

After close analysis of the source profiles and temporal variations, the 15-factor solution with an $F_{Peak} = 0.2$ was selected. The solution had the best reconstruction of mass and one that made the most physical sense. To check this result, the residuals were inspected, and additional runs with different starting locations were completed.

Analysis of the residuals of the data matrix for each set of solutions is important; these should generally have a normal distribution, with few points beyond ± 3 standard deviations. A large percentage of data points exceeding this range indicates that an inappropriate number of factors were used, or that the uncertainty estimates were incorrect. These are shown in Appendix A for each species for the 15-factor solution with $F_{Peak} = 0.2$. One note is that with the large number of extreme outliers in the data set, it is likely that there may be some residuals beyond ± 3 standard deviations. However, the number of these outlying residuals was always less than 0.1% of the total number of data points, so overall the residuals supported using the 15-factor solution, since other solutions had a similar or worse distribution of residuals.

In addition to examining the residuals, it is also important to ensure that multiple solutions do not exist, since there is a global minimum, but there may also be local minima that PMF may get attracted to. Therefore, it is important to repeat runs with a different starting location. Only one multiple solution was found with 15 factors and an $F_{Peak} = 0.2$, though it was discarded since it had negative mass coefficients. **Table 5-3** lists the dominant species, the average percentage of the total mass and the average mass in each factor. The average percentage of the total mass from each factor is shown in **Figure 5-3**. Source profiles by composition (i.e., the percent of mass in a factor from each species) and by species distribution (i.e., the percent of each species in each factor) are shown in **Figures 5-4 through 5-18**.

Factor 1 was characterized by ethane, ethene, n-butane, and acetylene, with some pentanes, and on average accounted for 5% of the total mass. This is a somewhat surprising combination and may be due in part from industrial flares; earlier work indicated acetylene and ethene were evident in flare emissions (Brown and Main, 2002). Additionally, motor vehicle exhaust may contribute to this factor, based on a fairly high amount (50%) of the total acetylene that is included. Examining median concentration and weight percent by wind direction supports this hypothesis, because this factor is high with wind from the northwest (the direction of the freeway) and from the east (industrial). However, this factor does not decrease on the weekend, which suggests that there is low influence from mobile sources (see Section 5.3.2). Therefore, it appears that this factor may be mostly an industrial flare signature.

Factor 2 was characterized by p-diethylbenzene and n-propylbenzene, with over half the mass due to the unidentified fraction (though there is only 15% of the total unidentified in this factor). This factor is likely from an industrial aromatic source (#1), though it is possible that these species have a similar analytical bias. Wind direction analysis (detailed in later sections), however, suggests that there are sources of this factor to the south and southwest, which would indicate that this factor is real, and its potential sources need to be further researched.

Factor 3 is likely a motor vehicle source, with an abundance of benzene, toluene, xylenes, 2,2,4-trimethylpentane, and acetylene; these species are all typical of motor vehicle exhaust.

This factor also shows a decrease on Sundays compared to other days, again typical of motor vehicle emissions. This factor, on average, comprises only 4% of the total VOC concentration. While industrial emissions are thought to have a large impact on the Clinton Drive site, it is unlikely that motor vehicles play such a small role. Part of Factor 1 is also likely from mobile sources, and some of the later factors with influences from other paraffins and aromatic hydrocarbons are also likely to have some motor vehicle influence that could not be separated out by PMF.

Factor 4 contains approximately 75% of the light olefins, ethene and propene, and on average accounts for 5% of the total concentration. This light olefin factor is from industrial processes in the HSC to the east and south, as shown in later wind direction analysis. These compounds are among the most reactive VOCs, and the reactivity and ozone formation potential of this factor is further discussed in Sections 5.6 and 5.7.

Factor 5, on average, accounts for 25% of the VOCs at Clinton Drive, and is made up predominantly of butanes, which are among the most abundant species. This is a combination of evaporative emissions and general background from both mobile and point sources in the HSC. These butanes alone generally make up 16% of the total VOC mass at Clinton Drive, and the addition of some pentanes made this factor even higher in total mass. While these compounds are not very reactive, their high concentrations indicate they can be important in ozone formation (Brown and Main, 2002; Brown et al., 2002).

Factor 6 identified consisted of mid-weight paraffins, from C6 to C9 (though without trimethylpentanes). The overall mass was low (on average 3% of the total) and is likely to originate from solvent use to the south-southeast. As mentioned earlier, a small fraction of these compounds may also be from mobile sources.

Factor 7 had all of the pentenes, plus more than half the isopentane and half the n-pentane but accounted, on average, for only 1% of the total concentrations. This factor is likely from an industrial pentene source and was predominantly from the south and east-southeast, indicating the source or sources may be in that direction. Pentenes are one of the groups of compounds identified by TCEQ as important in ozone formation in the area, and this factor is expected to be significant when scaled by reactivity (Section 5.6).

Factor 8 consisted of 1,2,3-trimethylbenzene, the unidentified fraction, and small fractions of 1,2,4-trimethylbenzene, propene, and acetylene. This factor did not have as clear a dependence on wind direction as the other factors though it was higher with winds from the west and southwest. This is likely another industrial aromatic source (#2), though the lack of wind direction dependence may also suggest that the dominant species in this factor, 1,2,3-trimethylbenzene, has an analytical bias that forced this factor.

Factor 9 was dominated by 1,3-butadiene, an extremely reactive species with high ozone formation potential, and isobutene, and is from industrial 1,3-butadiene sources. These sources are located throughout the HSC, and while this factor, on average, was only 2% of the total mass, its high reactivity potential likely makes it important in ozone formation.

Factor 10 was a mix of mid-weight paraffins from C5 to C7 and likely due to evaporative emissions and solvent use from both industrial facilities, as well as oil and gas leaks and spills from vehicles on the road. On average, it comprised about 10% of the overall VOC concentrations at Clinton Drive, though the species involved are not very reactive and probably not very important in ozone formation.

Factor 11 consisted of the light paraffins ethane and propane and, on average, was 24% of the total VOC concentration. Both of these species are abundant, and because of their low reactivity, they tend to accumulate in the urban atmosphere. They are also prominent in emissions from natural gas usage. Due to their abundance, they can play a small role in ozone formation. This is likely an accumulation and natural gas factor, with contributions from both mobile and industrial sources. Concentrations of this factor are generally higher with winds from the east, though on a weight percent basis, this factor is higher with winds from the north. This difference is likely due to the higher amounts of VOC emissions in the HSC to the east and south, which dilute this factor's prominence in these directions. With winds from the north, where emissions are less, this background factor is a larger amount of the total mass.

Factor 12 was a mixed heavy aromatic factor with most of the m-ethyltoluene and 1,3,5-trimethylbenzene, as well as some o-ethyltoluene, 1,2,4-trimethylbenzene, and xylenes. Overall this heavy aromatic factor was, on average, 12% of the mass and is mostly due to industrial emissions, though there may be some motor vehicle influence as well. This factor was predominantly from the south, southwest, and west, confirming a probable combination of industrial sources. It may be that these species have a particular analytical bias, so PMF was not able to sort out between the multiple sources.

Factor 13 consisted of most of the C10 and C11 alkanes, plus 50% of the C9 alkanes and 30% of the xylenes. These heavy alkanes are often used as diesel markers, and the predominance of this factor from many wind directions is consistent with diesel sources in many directions, from both trucks on the freeways to the west and north and tracks and trains to the south and east. Also, it decreased in concentration and weight percent on weekends, another likely indication of mobile sources. On average, it was only 2% of the total VOC mass.

Factor 14 had all of the isoprene, the only biogenic marker among the PAMS species analyzed by the auto-GC. This factor is predominantly biogenic in origin and, on average, was only 2% of the VOC loading. High concentrations of this factor occurred in the winter and during the night, which are times of minimal biological activity and therefore should be times of very low biogenic isoprene. Previous analyses (Brown and Main, 2002) have shown that industrial isoprene emissions were evident during the night and winter, and it appears that these emissions were included in this factor. This is not surprising because these industrial emissions are both infrequent (and therefore do not enough have variation to appear as a separate factor) and high in concentration (and therefore treated as an outlier and weighted less). However, while this factor includes both biogenic and industrial isoprene, instances when high concentrations occur during the night or winter can be attributed to industrial sources, while other periods are mostly biogenic.

Factor 15 consisted of all the butene and was, on average, 2% of the total VOCs. This industrial butene factor occurred almost exclusively with winds from the south, the direction of

multiple butene point sources. These industrial facilities are not the only butene sources in the area (see Appendix B), and the dominance of this factor only with winds from the south may indicate that butenes emitted from these other facilities to the east react away too fast to impact Clinton Drive. Their importance should not be discounted, however, as the butenes are reacted away to form ozone and can still impact Clinton Drive with their secondary byproduct.

Table 5-3. Important species, average % of the total mass, and likely source of each factor identified by PMF.

Factor	Important Species	Average % of total mass	Specific Wind Direction	Likely Source
1	Ethane, ethene, acetylene, n-butane	5%	NW, E	Industrial flare
2	p-diethylbenzene, n-propylbenzene	4%	S, SW	Industrial aromatic #1
3	Acetylene, benzene, 2,2,4-trimethylpentane, toluene, xylenes	4%	SW, W, NW, SE	Motor vehicle
4	Ethene, propene	5%	E, S	Light olefin
5	Butanes	23%	E, S	Evaporative emissions + background
6	C6-C9 alkanes, unidentified	3%	S, SSE	Solvents
7	Pentenes, pentanes	1%	S, ESE	Pentene source
8	1,2,3-trimethylbenzene, unidentified, 1,2,4-trimethylbenzene	1%	SW, W, NE	Industrial aromatic #2
9	1,3-butadiene, isobutane	2%	E, S	Butadiene source
10	C5-C7 alkanes	10%	E, SE, S	Evaporative + solvents
11	Ethane, propane	24%	N, E	Accumulation + natural gas
12	Ethyltoluenes, 1,3,5-trimethylbenzene, xylenes	12%	S, SW, W	Heavy aromatics
13	C9-C11 alkanes, unidentified, xylenes	2%	E, SE, S, SW, W, NW	Diesel
14	Isoprene	2%	W, E, S	Biogenic, also possibly from industrial source
15	Butene	2%	S	Butene source

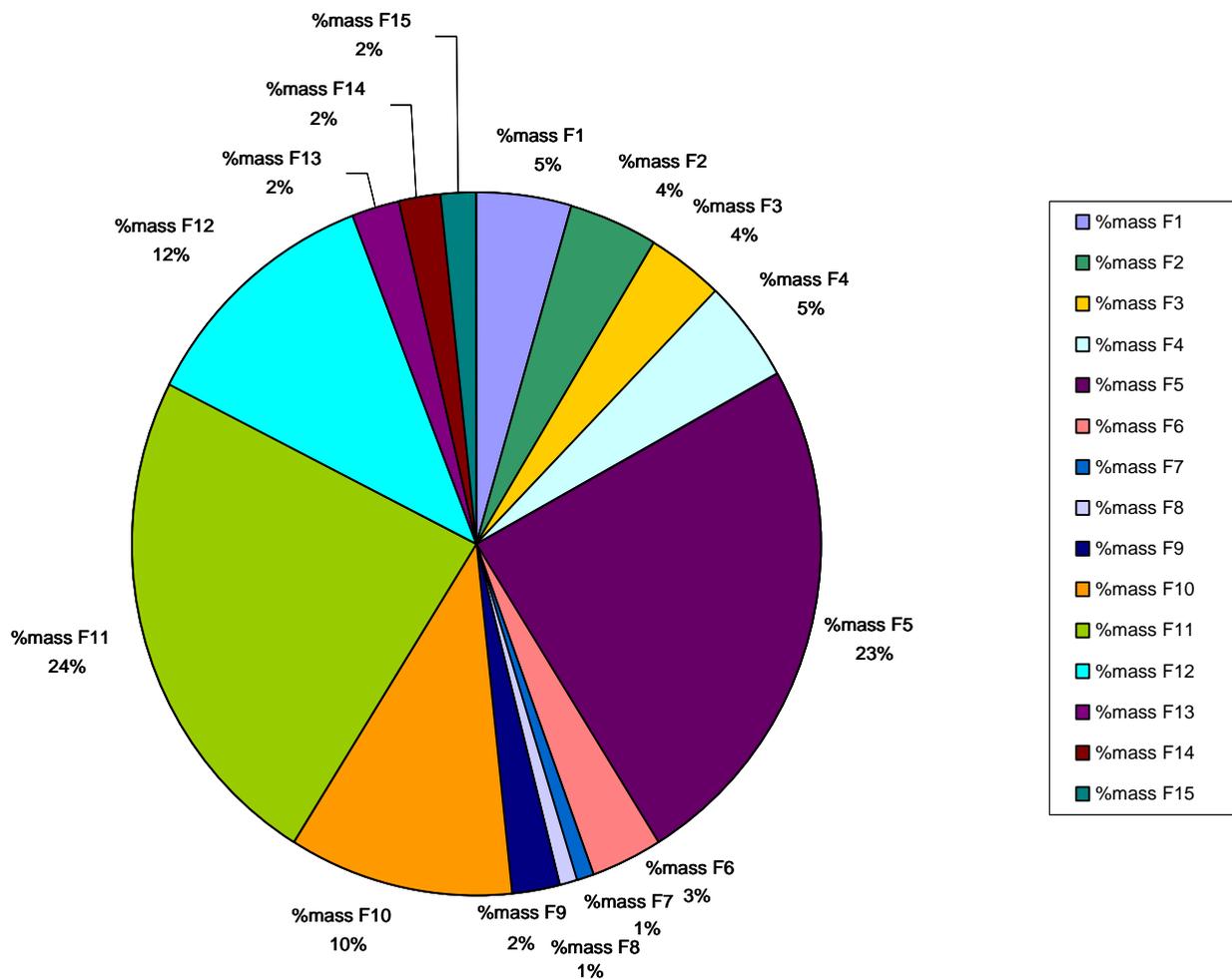


Figure 5-3. Average contribution of each factor (F) to the total mass (ppbC).

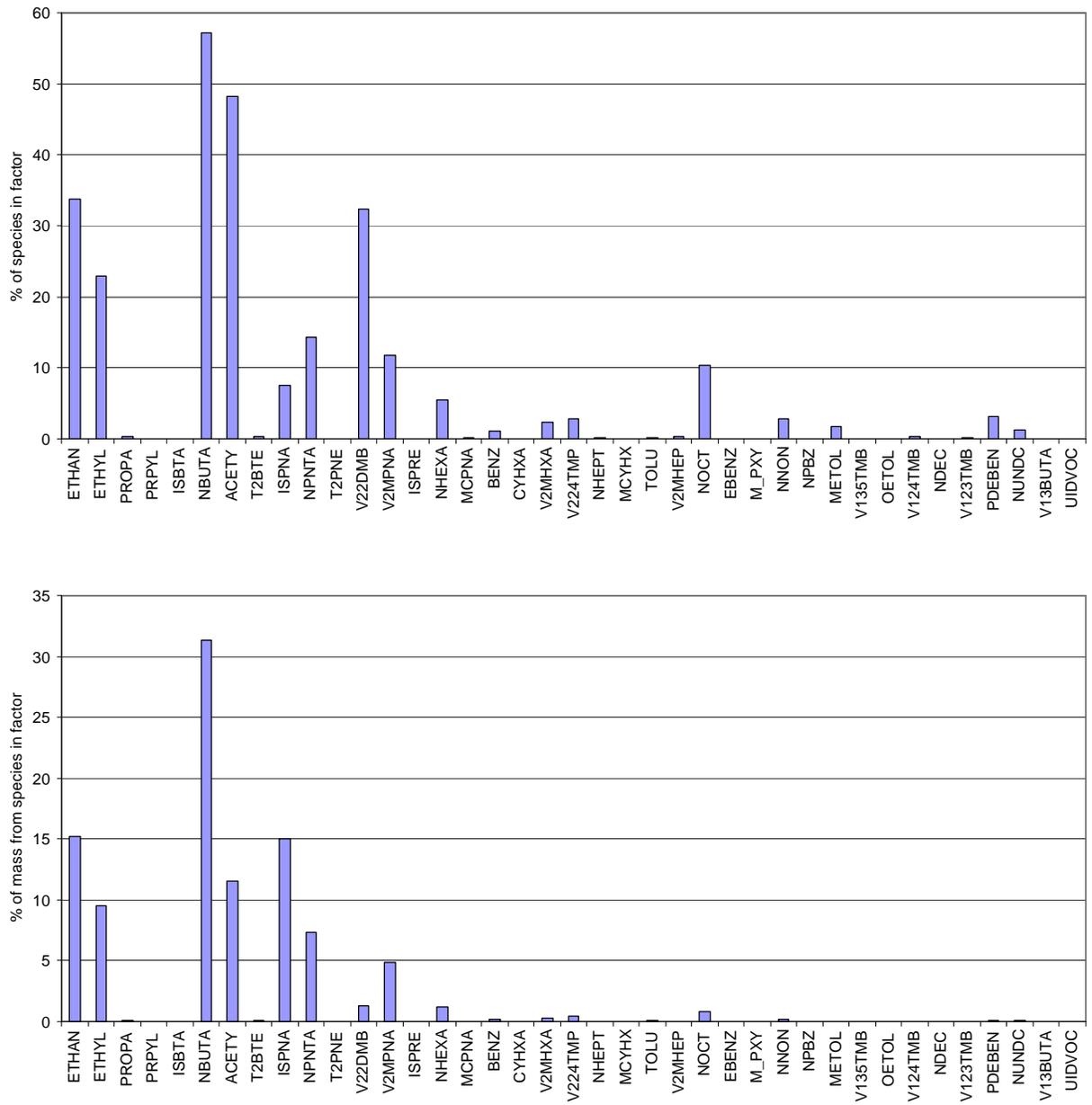


Figure 5-4. Percent of each species and the percent of mass from each species in Factor 1.

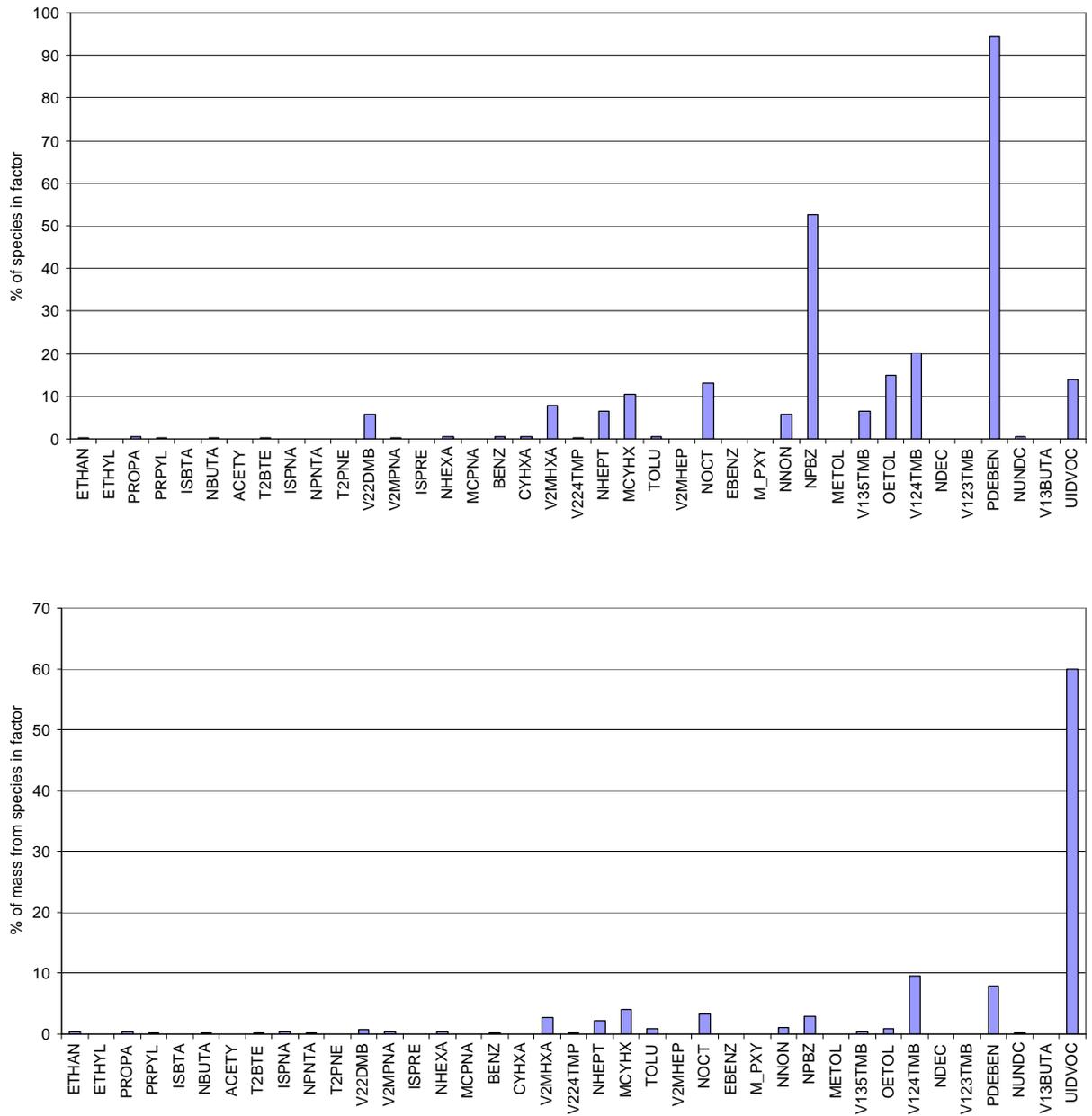


Figure 5-5. Percent of each species and the percent of mass from each species in Factor 2.

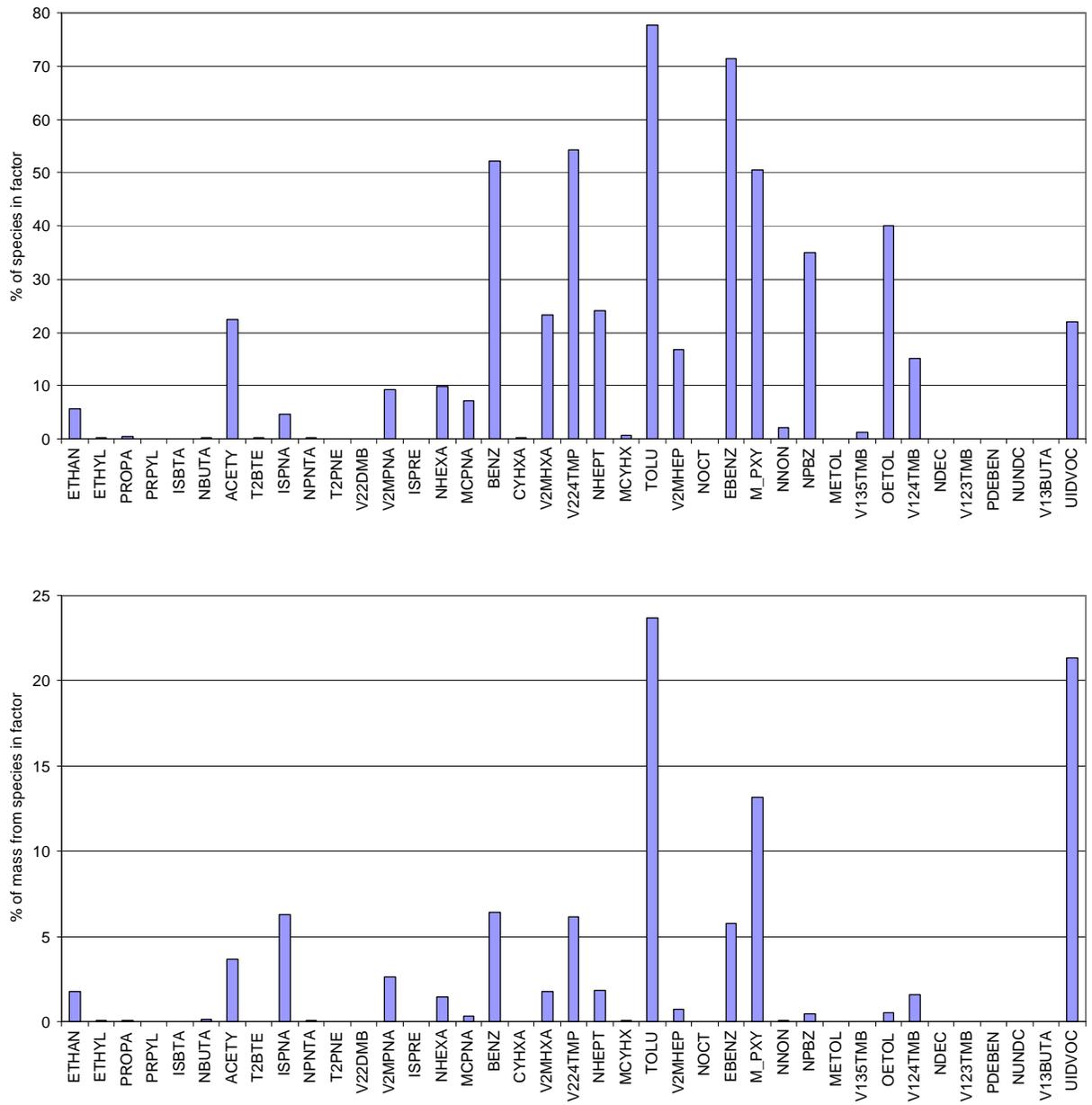


Figure 5-6. Percent of each species and the percent of mass from each species in Factor 3.

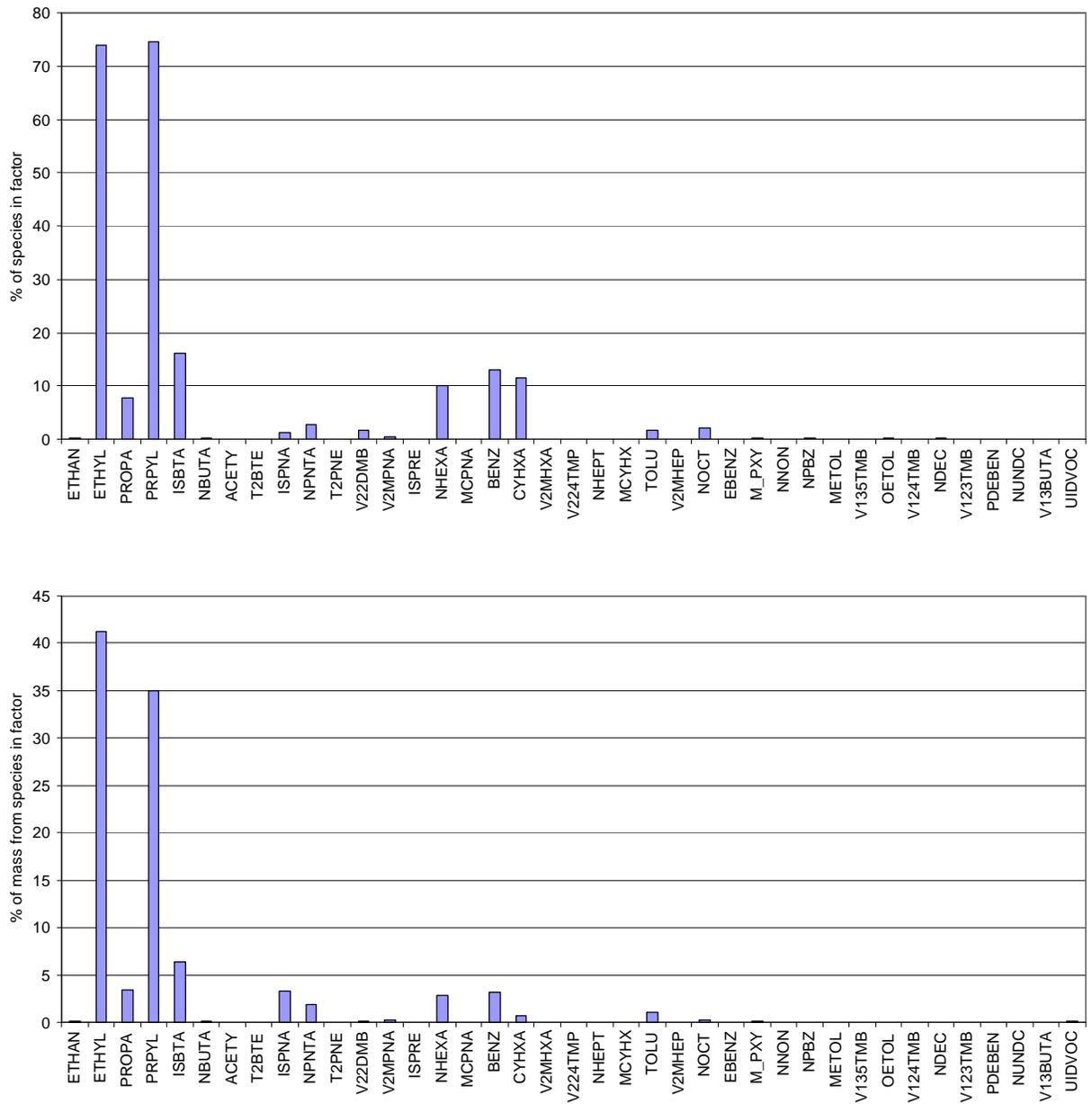


Figure 5-7. Percent of each species and the percent of mass from each species in Factor 4.

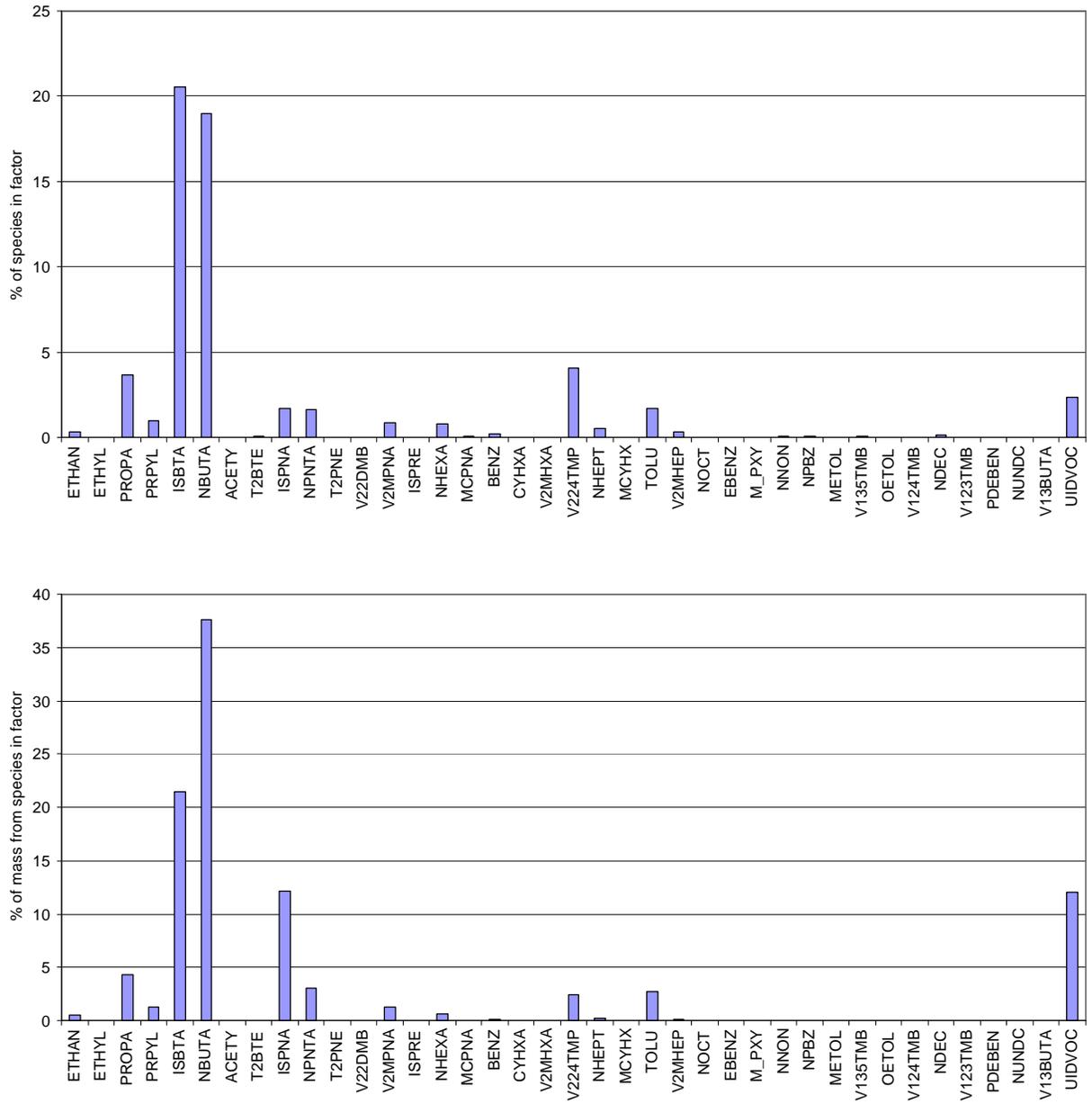


Figure 5-8. Percent of each species and the percent of mass from each species in Factor 5.

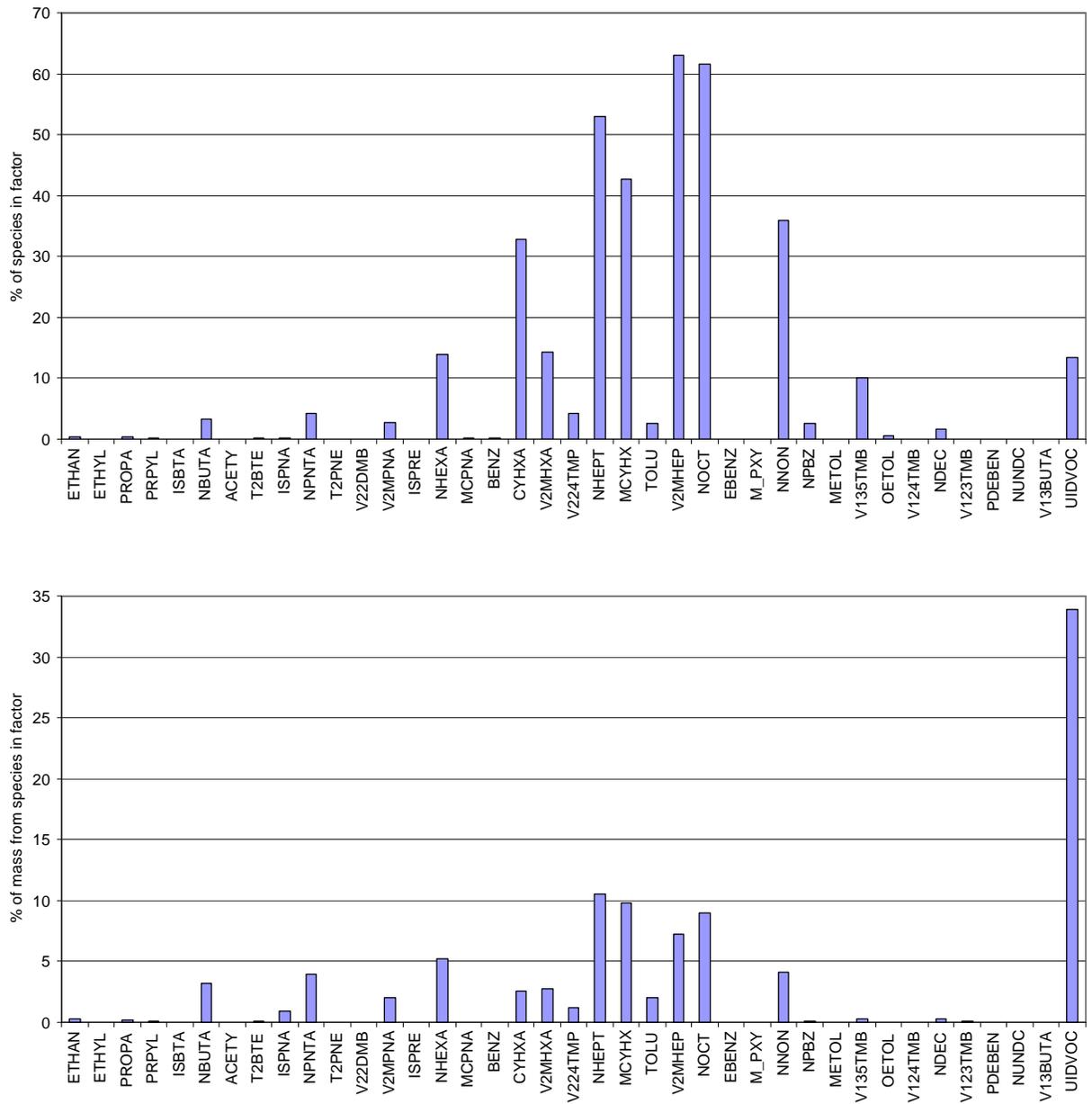


Figure 5-9. Percent of each species and the percent of mass from each species in Factor 6.

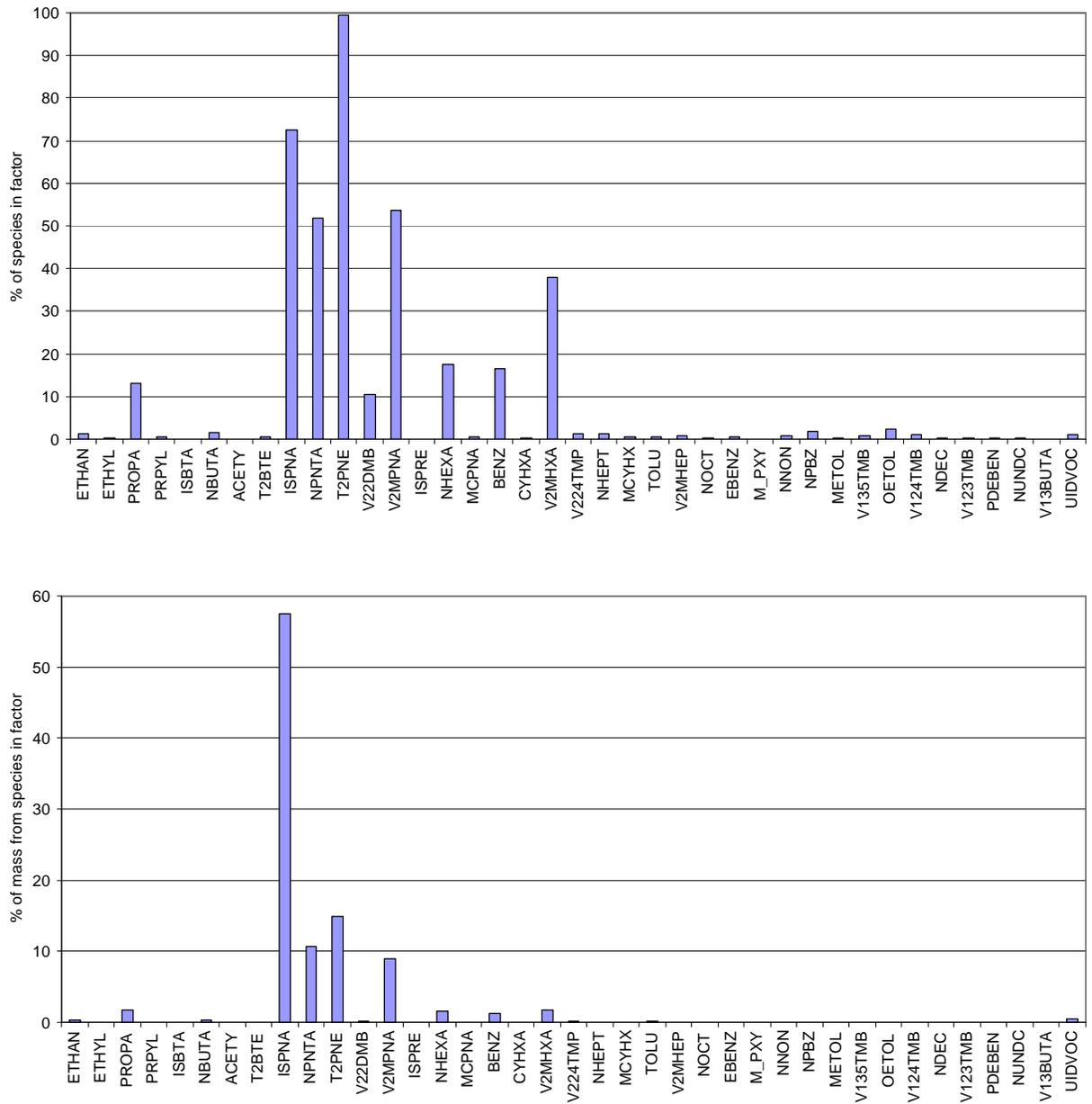


Figure 5-10. Percent of each species and the percent of mass from each species in Factor 7.

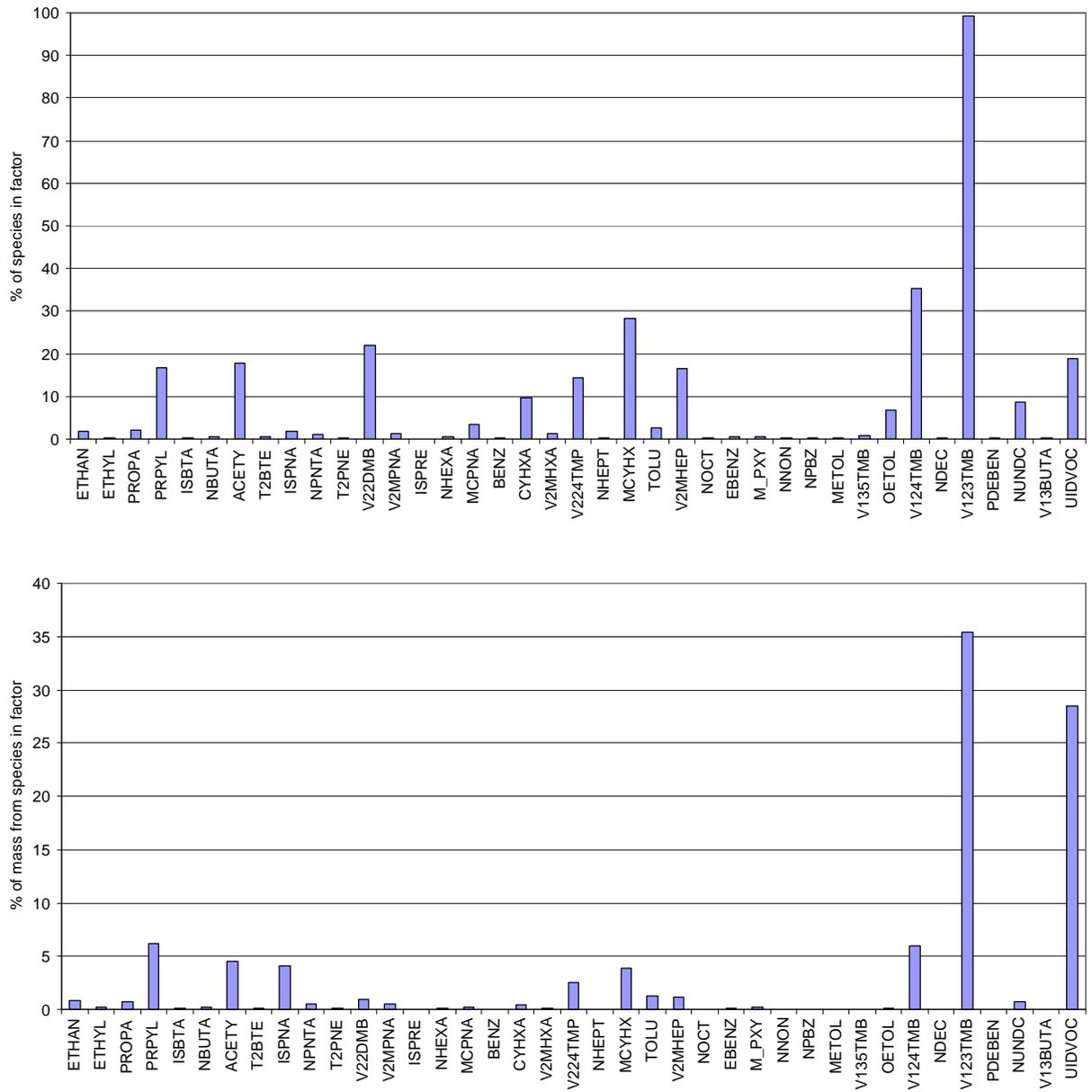


Figure 5-11. Percent of each species and the percent of mass from each species in Factor 8.

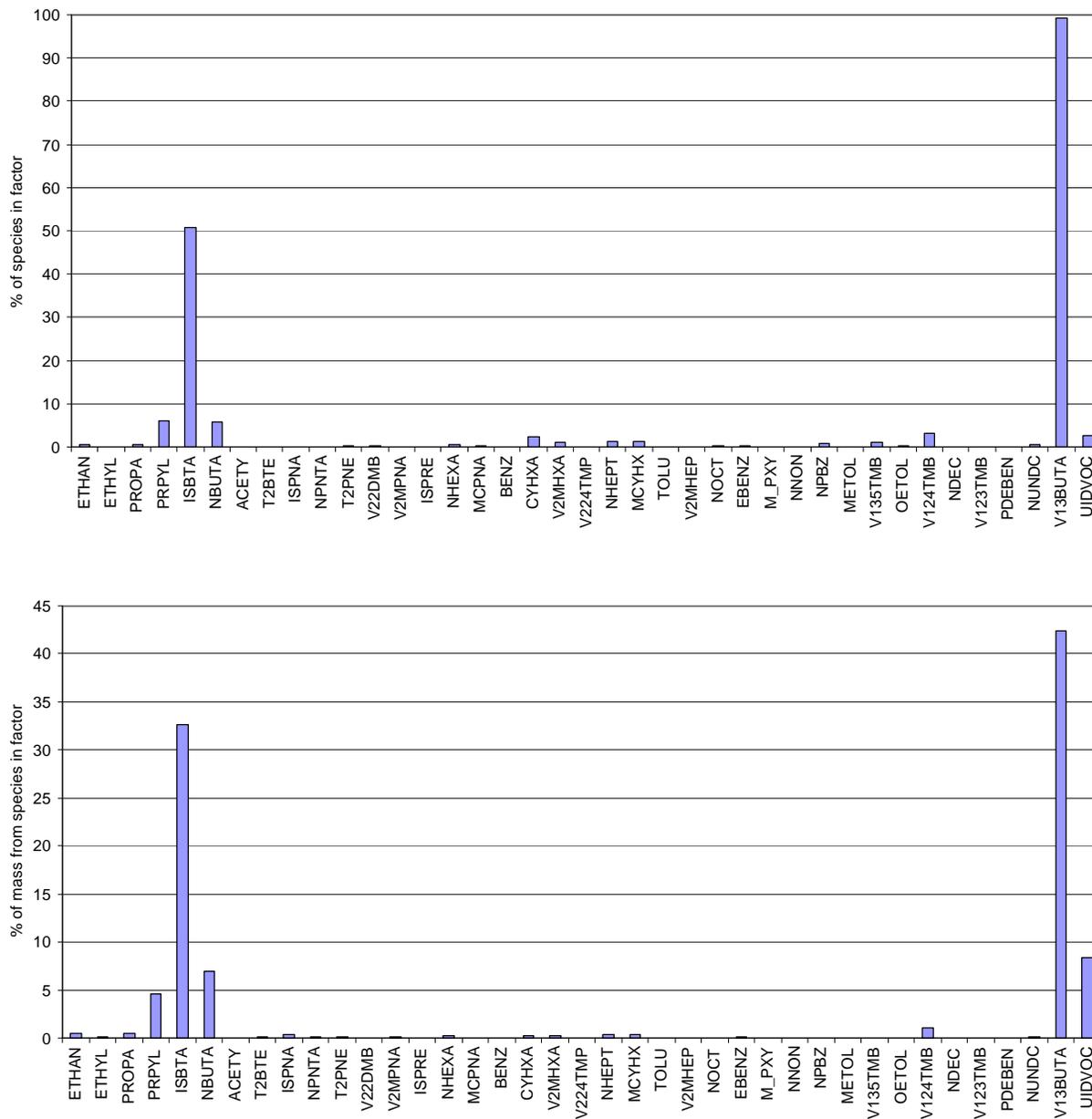


Figure 5-12. Percent of each species and the percent of mass from each species in Factor 9.

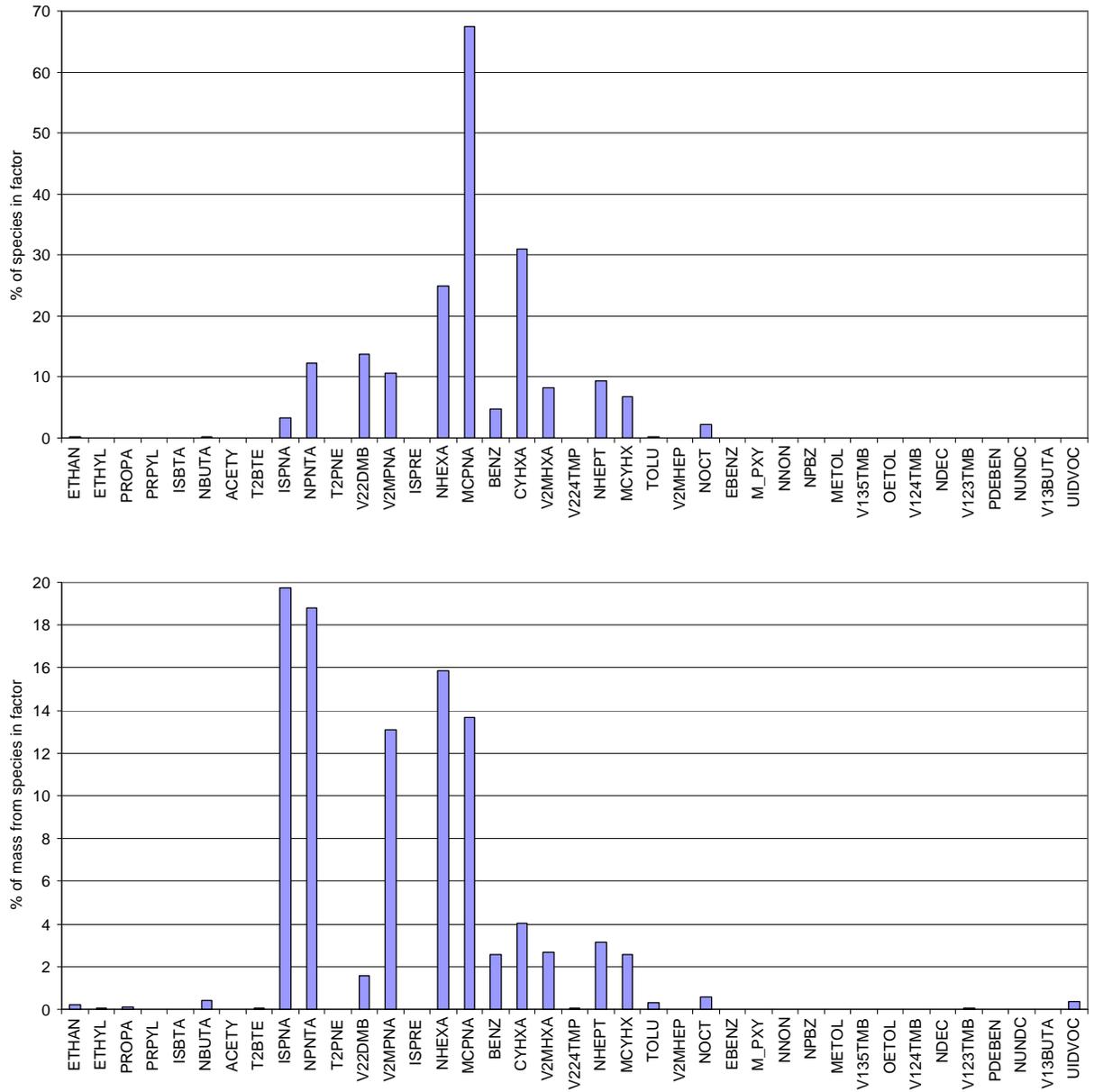


Figure 5-13. Percent of each species and the percent of mass from each species in Factor 10.

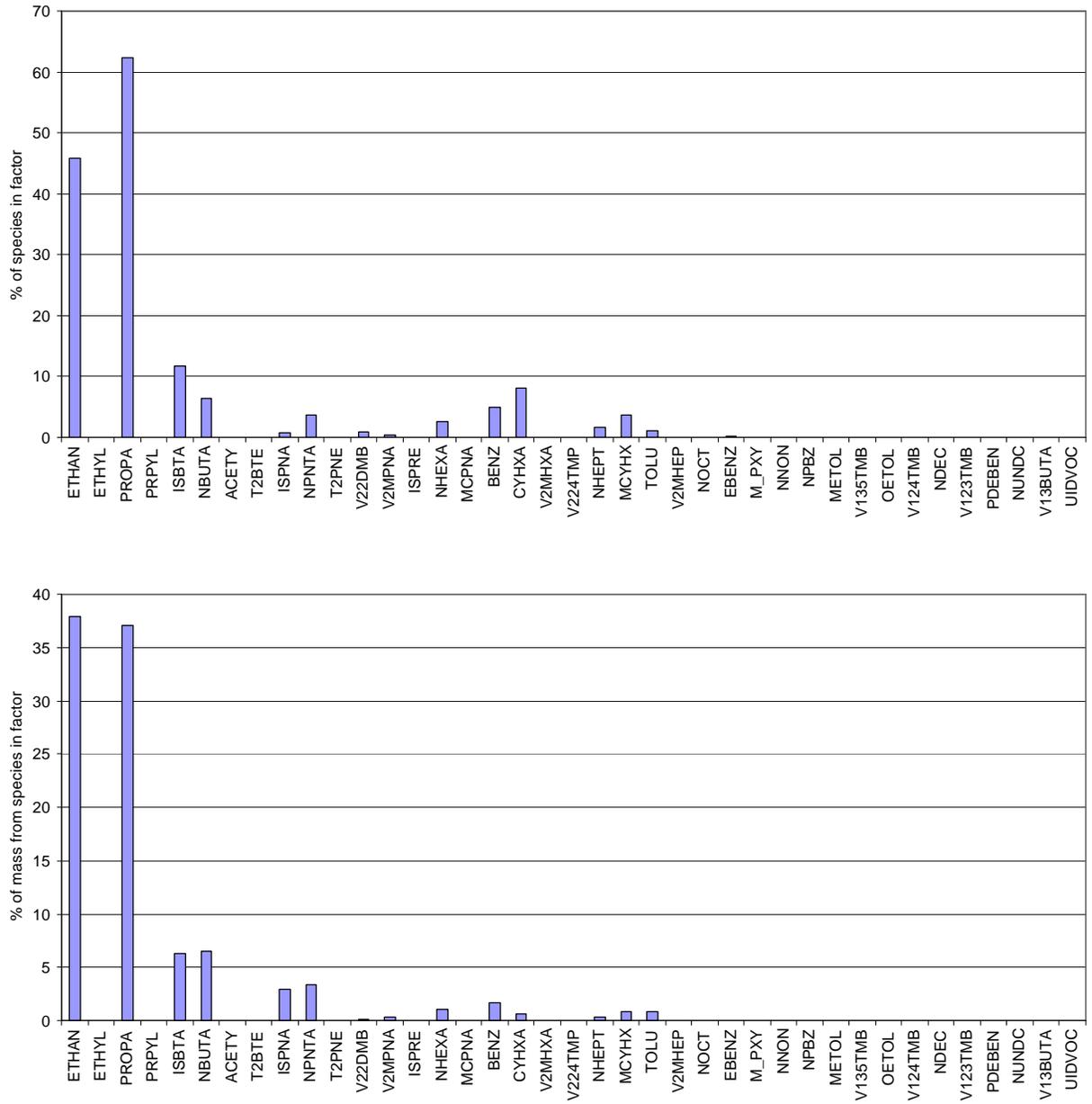


Figure 5-14. Percent of each species and the percent of mass from each species in Factor 11.

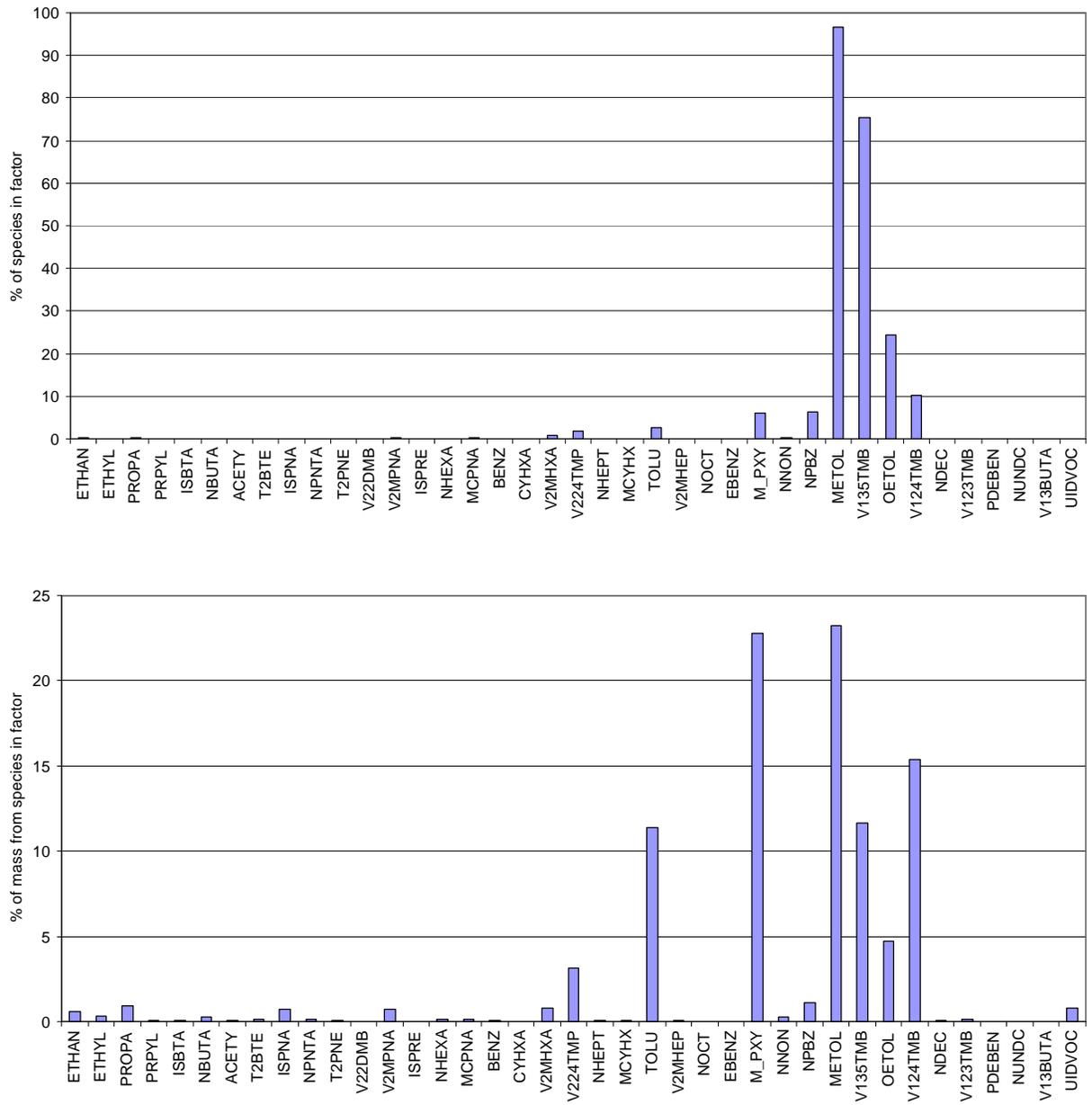


Figure 5-15. Percent of each species and the percent of mass from each species in Factor 12.

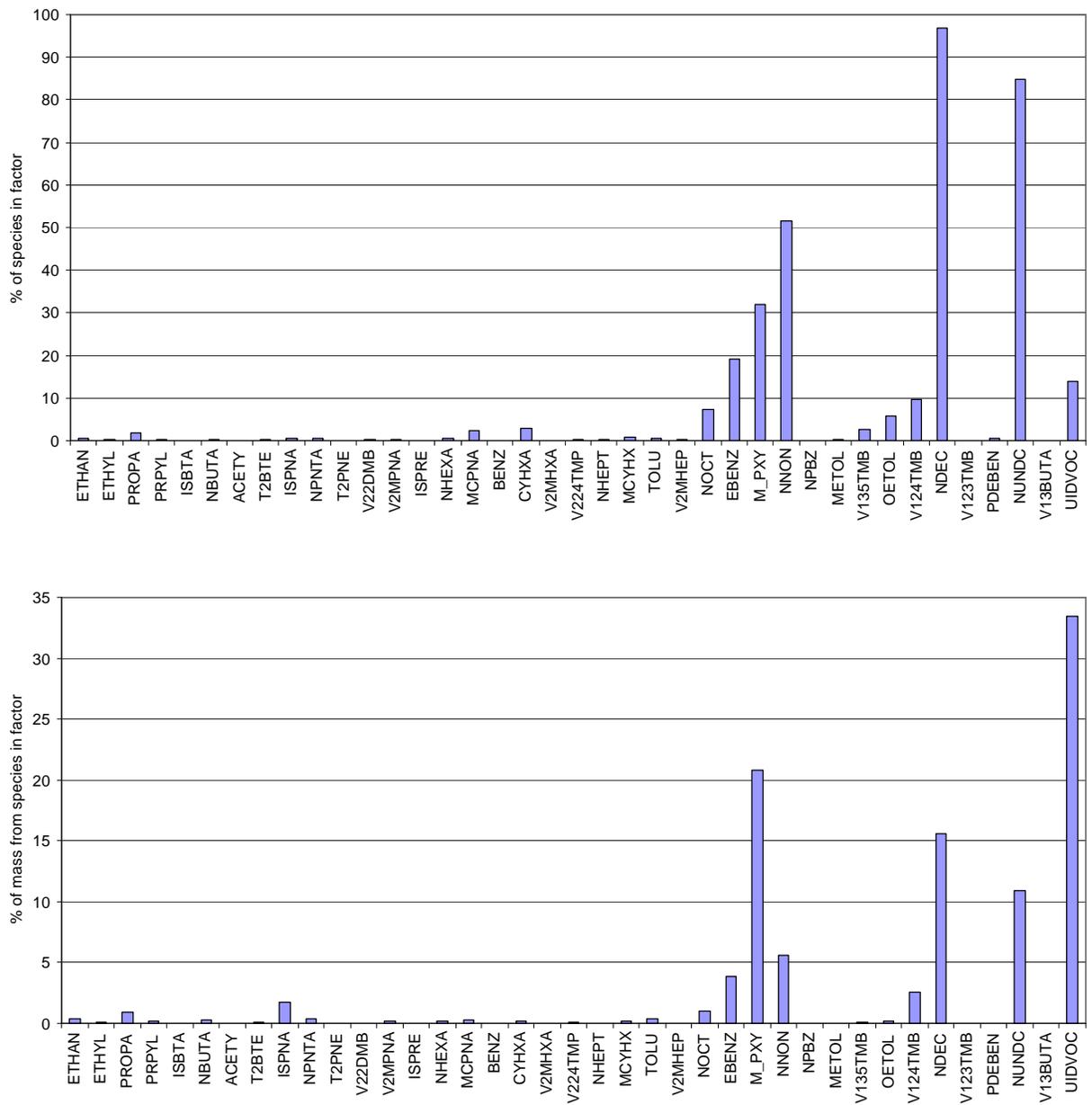


Figure 5-16. Percent of each species and the percent of mass from each species in Factor 13.

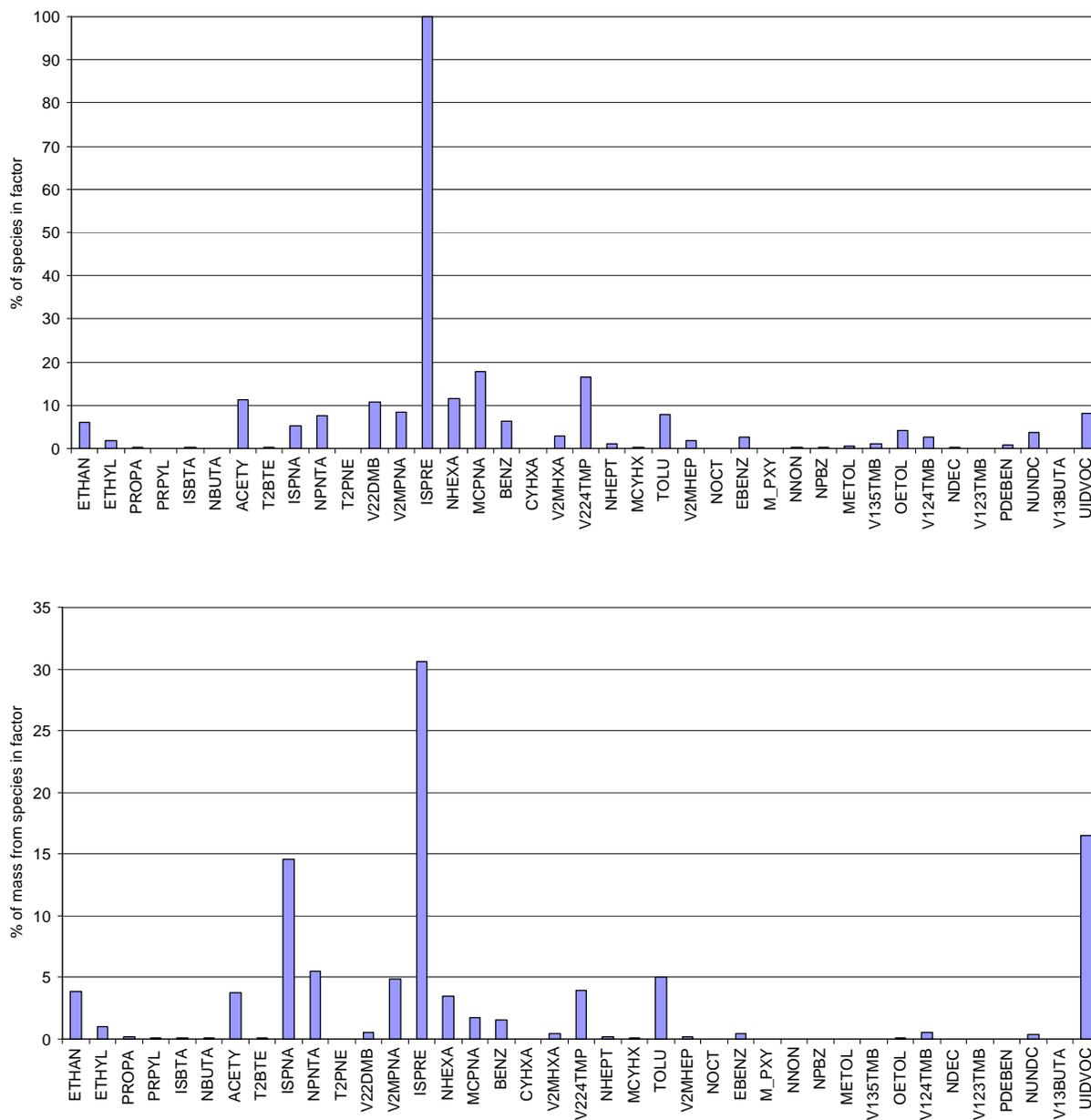


Figure 5-17. Percent of each species and the percent of mass from each species in Factor 14.

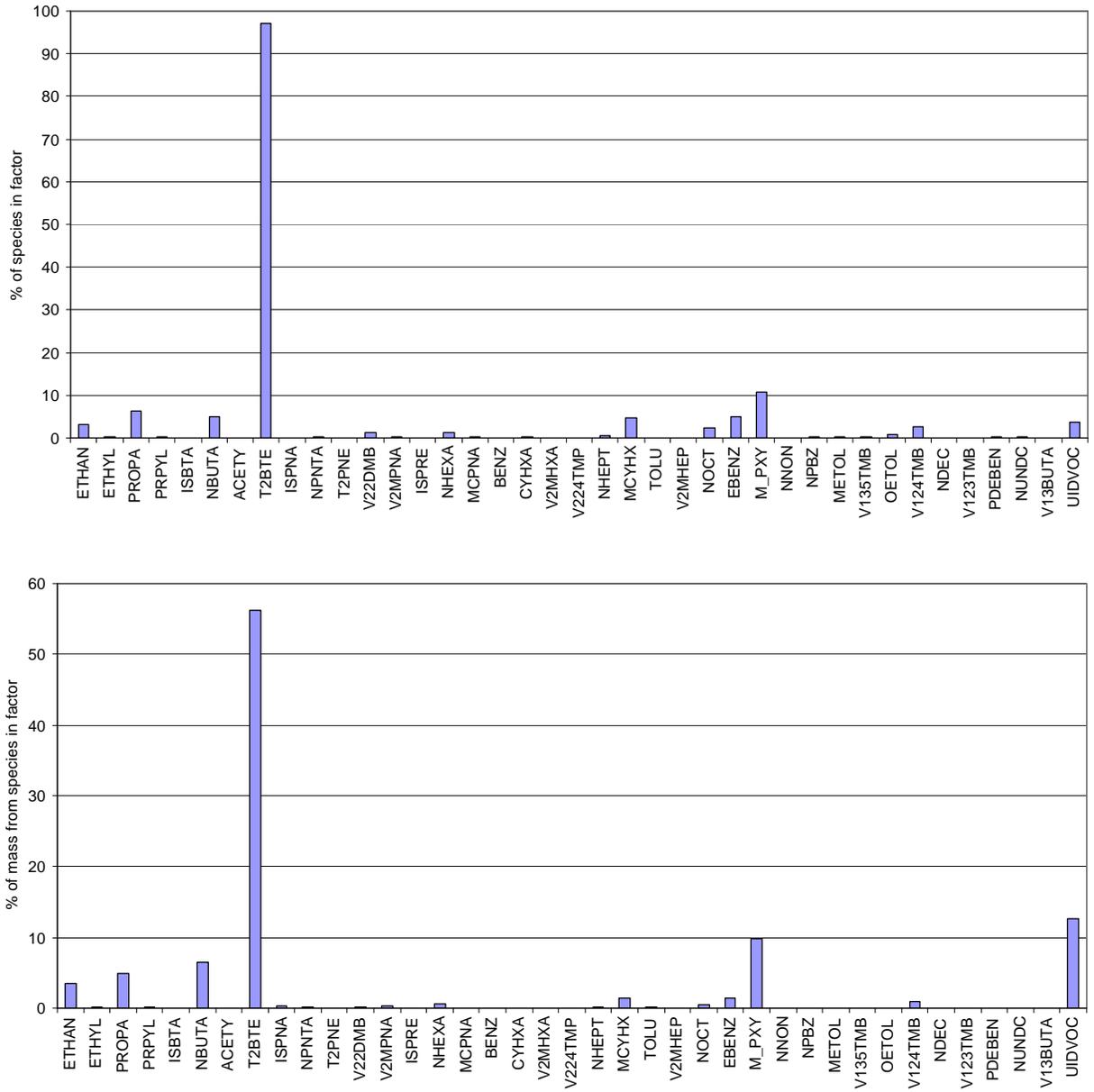


Figure 5-18. Percent of each species and the percent of mass from each species in Factor 15.

5.3 TEMPORAL ANALYSES

Temporal analyses, such as by season, day of week, or time of day, can lend more insight into the behavior of the identified factors and confirm the proper identification of the factors. For example, the biogenic factor is expected to be very low in the winter and at night; if this is not so, we need to understand why. Also, the influence of mobile sources should be evident by a

decrease in concentration on Sundays and have peaks in the morning and evening corresponding to commute hours.

5.3.1 Seasonal Variations

The mass contribution (ppbC) as well as the weight percent contribution of each factor by month was examined. Box whisker plots are commonly used to display a large amount of data and are particularly useful in assessing differences among data. Box whisker plots are drawn in different ways by different software programs. However, most box whisker plots show an interquartile range (i.e., 25th to 75th percentile) and some way to illustrate data outside this range. **Figure 5-19** shows an illustrated box whisker and notched box whisker plot. The box shows the 25th, 50th (median), and 75th percentiles. The whiskers always end on a data point; when the plots show no data beyond the end of a whisker, the whisker shows the value of the highest or lowest data point. The whiskers have a maximum length equal to 1.5 times the length of the box (the interquartile range). If there are data outside this range, the points are shown on the plot and the whisker ends on the highest or lowest data point within the range of the whisker. The “outliers” are also further identified with asterisks representing the points that fall within 3 times the interquartile range from the end of the box and circles representing points beyond this.

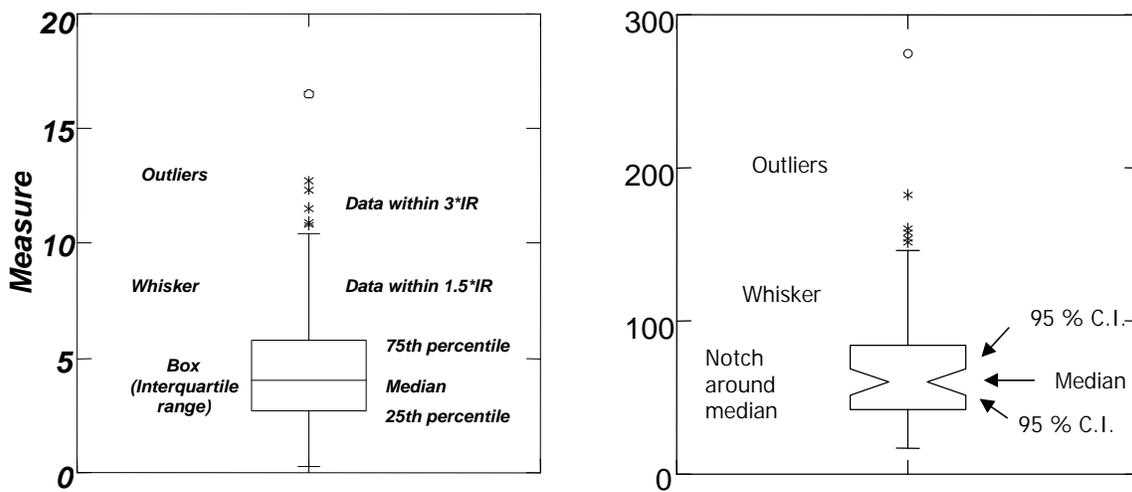


Figure 5-19. Illustration of a box whisker plot and a notched box whisker plot as defined by SYSTAT statistical software.

Since sample size is also an important consideration when one begins to stratify data, notched box whisker plots (see Figure 5-19) have been used to analyze data in this study. These plots include notches that mark confidence intervals. The boxes are notched (narrowed) at the median and return to full width at the lower and upper confidence interval values². We selected

² SYSTAT literature uses methodology documented by McGill, Tukey, and Larsen (1978) to show simultaneous confidence intervals on the median of several groups in a box whisker plot. If the intervals around two medians do not overlap, one can be confident at about the 95% level that the two population medians are different.

95% confidence intervals. If the 95% confidence interval is beyond the 25th or 75th percentile, then the notches extend beyond the box (hence the “folded” appearance).

Notched box whisker plots for each factor’s weight percent by month are shown in **Figures 5-20 through 5-23**. A number of trends are evident:

- Many factors have a peak in summer, including 2 (industrial aromatic), 3 (motor vehicle), 6 (solvents), 7 (pentenes), 8 (trimethylbenzenes), 9 (butadiene), 12 (heavy aromatics), 13 (diesel) and 14 (biogenic).
- The solvent and mobile source factors are expected to be higher in the summer due to increased temperatures which volatilize higher concentrations of species.
- Factor 12 shows a similar seasonal trend as the well characterized mobile source factors, further evidence that Factor 12 has a strong influence from mobile sources.
- The biogenic factor is expected to rise significantly in the summer, correlating with the large increase in biogenic activity. High concentrations due to industrial emissions (i.e., more than 2 ppbC) of this factor occur even in the winter and indicate that some part of this factor is industrial in origin even in the summer, but the two signatures could not be separated by PMF.
- The butadiene and pentenes factors’ weight percents peak in May, which has been seen in previous work (Brown and Main, 2002) to be a month of extremely high butadiene concentrations, though the cause and source are unknown.
- Factor 2 is generally high all spring and summer before dropping off significantly in the winter, while Factor 8 has a distinct peak in July and August, further illustrating that these are indeed separate sources, though again the exact cause of these trends are not well-characterized.

Three factors exhibited highest weight percents in the winter: 1 (industrial flares), 5 (evaporative/backgrounds), and 11 (accumulation/natural gas). The pairing of the first factor with the two accumulation and general background factors is interesting. This may indicate that this industrial flare factor is a general background of flares from many sources and, therefore, may not be able to be isolated. It also further suggests there is minimal motor vehicle influence, because the distinct mobile source factors (3 and 13, plus some of 12) peak in the summer. These accumulation/background factors probably have a higher weight percent in the winter because a number of other factors, as shown earlier, decrease in the winter, which makes the accumulation/background factors’ weight percents higher.

Three factors showed no distinct seasonal trend: 4 (light olefins), 10 (evaporative solvents), and 15 (evaporative/backgrounds). Factor 15 showed a small decrease in the summer, likely due to accelerated destruction of the reactive butenes by more intense solar radiation. Some sort of trend, either similar to the butenes or the pentenes, would be expected for the light olefin factor, but its weight percent remains fairly constant throughout the year. However, concentrations are lower in the summer, which is consistent with increased depletion by

photochemistry, similar to the butenes. The evaporative emissions factor shows a strange and sudden decrease in April, but otherwise remains fairly constant throughout the year in both concentration and weight percent.

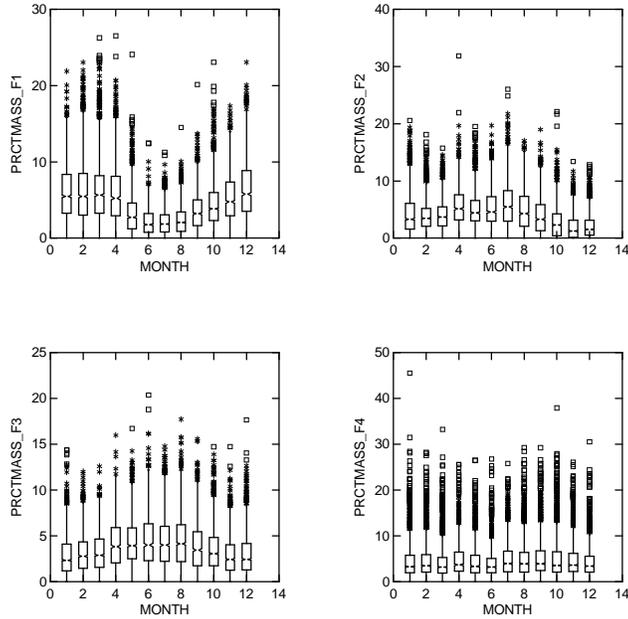


Figure 5-20. Notched box whisker plots of Factors 1-4 weight percent by month.

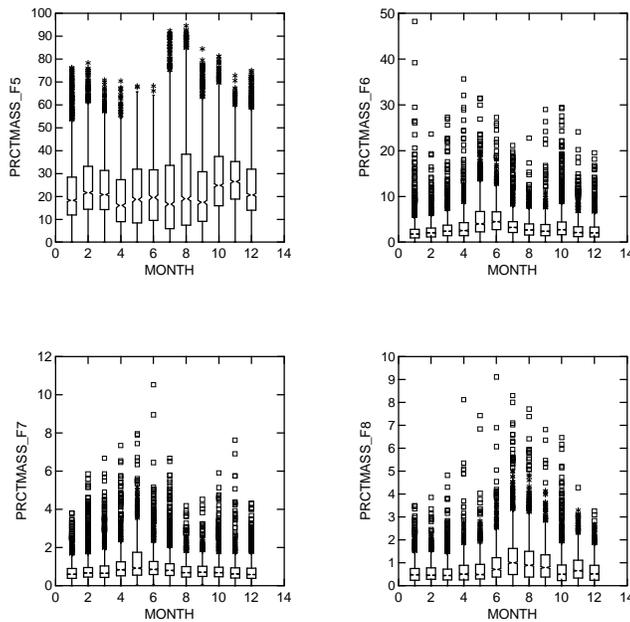


Figure 5-21. Notched box whisker plots of Factors 5-8

weight percent by month.

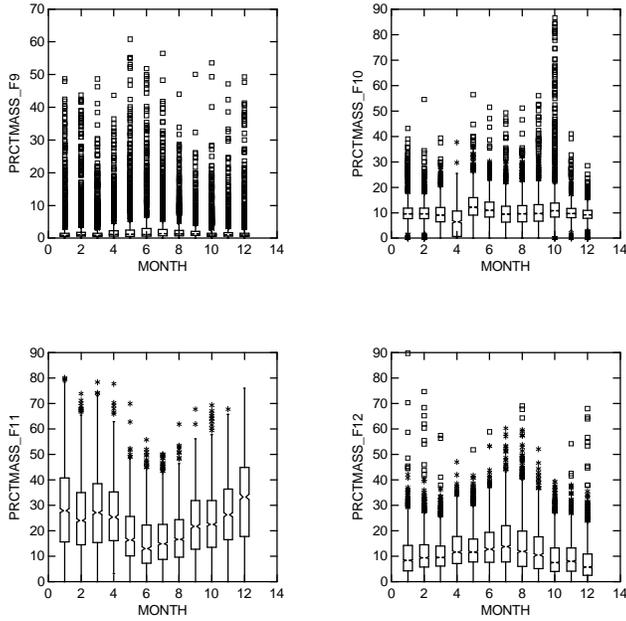


Figure 5-22. Notched box whisker plots of Factors 9-12 weight percent by month.

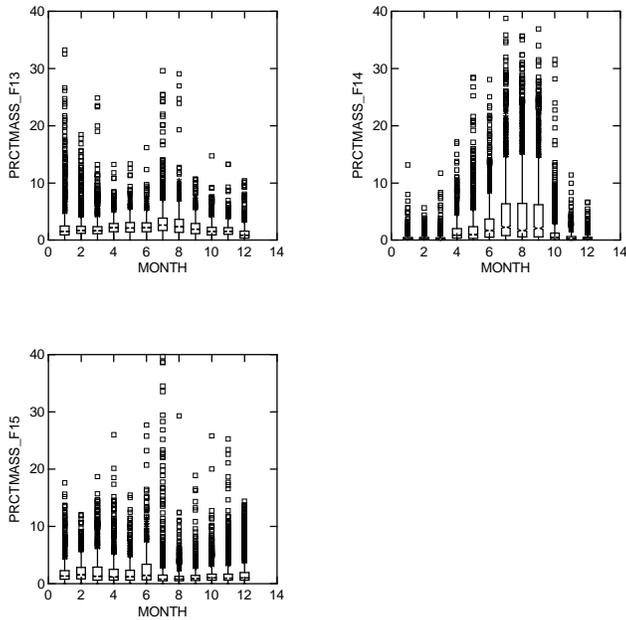


Figure 5-23. Notched box whisker plots of Factors 13-15 weight percent by month.

5.3.2 Day of Week Variations

Species from mobile sources generally show a decrease between the weekdays and weekends and are lowest on Sundays. If the factors attributed to mobile sources are well apportioned and properly identified, they should decrease on the weekends. Other factors, such as industrial or background/accumulation factors, should have little day-of-week difference.

Figures 5-24 through 5-27 show notched box whisker plots of each of the fifteen factors by day of week. Factors 3 (motor vehicle), 12 (heavy aromatics), and 13 (diesel) show a modest decrease on Sundays. Factor 1 (industrial flares), which may have some mobile source influence, did not show a decrease on the weekend, which may suggest that this factor is mostly industrial in origin. Previous work (Brown and Main, 2002) demonstrated that industrial activity occurs independent of the day of week. Factors attributed to stationary sources showed little day-of-week variation, consistent with the identification of these factors.

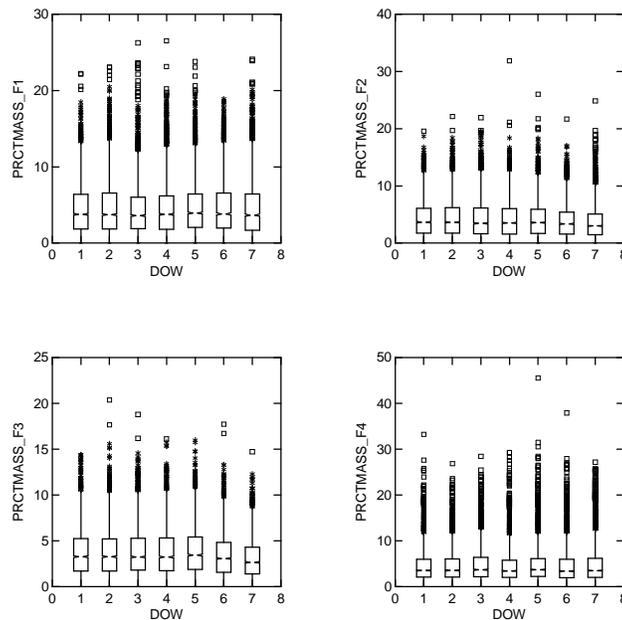


Figure 5-24. Notched box whisker plots of Factors 1-4 weight percent by day of week (1=Monday).

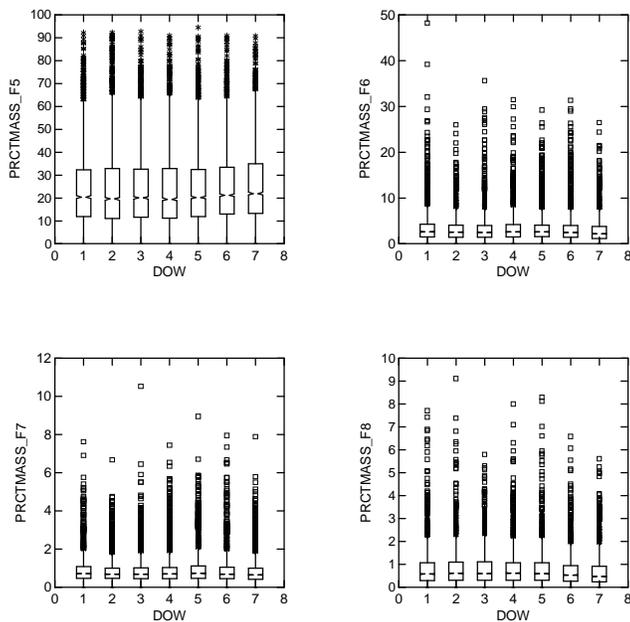


Figure 5-25. Notched box whisker plots of Factors 5-8 weight percent by day of week (1=Monday).

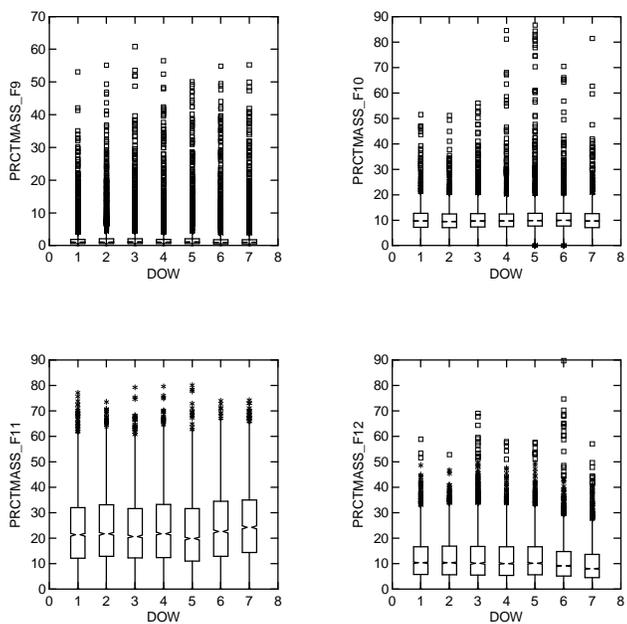


Figure 5-26. Notched box whisker plots of Factors 9-12 weight percent by day of week (1=Monday).

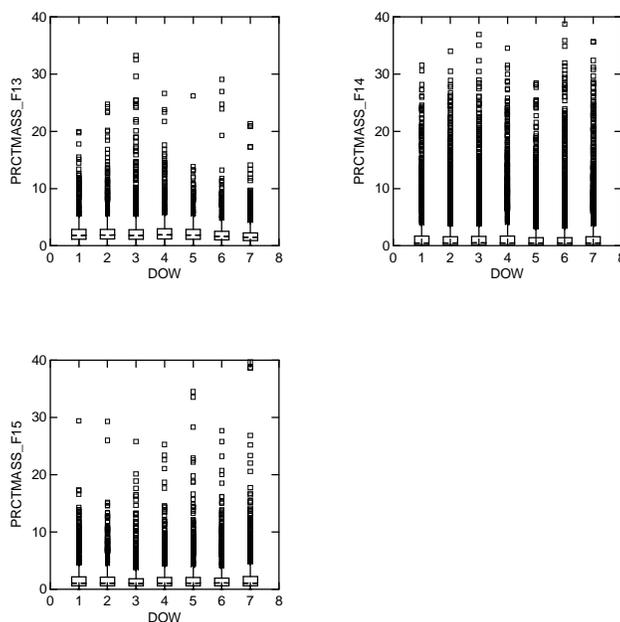


Figure 5-27. Notched box whisker plots of Factors 13-15 weight percent by day of week (1=Monday).

5.3.3 Time of Day Variations

Concentrations and composition often change over the course of a day as different air masses, mixing heights, winds, and emissions influence a particular site. Often emissions are highest in the morning when low mixing heights, minimal winds, and lack of solar radiation encourage accumulation of emissions. VOCs from mobile sources often show a morning and afternoon/evening peak associated with the rush hour, while industrial emissions, though they often accumulate in the morning, do not have such a pattern. Box plots of each of the fifteen factors' weight percent by hour are shown in **Figures 5-28 through 5-31**.

Factors 3 (motor vehicle) and 12 (heavy aromatics) had distinct and relatively large peaks in the morning (0600-0700 CST) and evening (1700-1800 CST) that are consistent with a large mobile source influence. This further supports the idea that these factors as predominantly mobile source in origin.

Factors 1 (industrial flare), 2 (industrial aromatic hydrocarbons), and 7 (pentenes) each showed small peaks in the median weight percent in the morning and afternoon/evening. These rises in weight percent in the morning and afternoon/evening are consistent with mobile source emissions, indicating that it is likely that the acetylene in Factor 1, some of the aromatic hydrocarbons in Factor 2, and some of the pentenes/pentanes in Factor 7 are from motor vehicles. However, the fact that these are only small rises, especially compared to the mobile source Factors 3 and 12, also suggests that stationary sources are more important in these factors. It may also be that these hourly variations are due to meteorology only, with accumulation in the morning, and older, secondary air masses influencing the site in the afternoon as the Bay breeze

brings morning emissions that were advected away from Houston back over the site. Previous work (Brown et al., 2002) indicated this is a frequent occurrence, and that these air masses can have high ozone and be subjected to an injection of “fresh” emissions of reactive species as they pass over the HSC.

Six factors had weight percent peaks in the morning only, with factors 8 (1,2,3 trimethylbenzene), 4 (light olefins), 9 (butadiene), and 10 (C5-C7 paraffins) at 0700 CST, factor 11 (light paraffins, accumulation) at 0300 CST, and factor 13 (diesel) at 1000 CST. The first four factors’ peaks suggest that their species accumulate during the nighttime, consistent with previous analyses showing near-continuous nighttime emissions of these compounds accumulating under the low mixing heights, etc., at night. The early morning peak at 0300 CST of Factor 11 is mainly due to an increase in concentration of other factors later in the morning; concentrations of this factor remain relatively steady from 0100 to 0800 CST (thus its weight percent is decreased as other VOC concentrations increase). The mid-morning weight percent peak of Factor 13 is interesting and may be due to diesel truck traffic in conjunction with trains idling while being unloaded/loaded in the morning.

Factors 5 (butanes), 6 (C6-C9 paraffins) and 15 (butenes) saw their weight percents peak in the afternoon or evening, often with a minima in the morning. The first two factors follow a somewhat similar pattern, with median weight percent maxima at 1400 CST and 2000 CST, respectively, and weight percent lows at 0600 CST and 0300 CST, respectively. Their concentration profiles by hour are nearly identical (see **Figure 5-32**), with median concentration maxima at 2000 CST and minima at 1300 CST. The lower weight percents in the morning are likely due to increases in other species’ concentrations, while these factors’ concentrations remain fairly constant. Their rise in the afternoon and evening may be due to the breakdown of reactive precursors of these species due to photochemistry. The last factor, composed of reactive butenes, has a sharp decrease in the early morning, most likely due to sunrise and the beginning of photochemistry and breakup of the boundary layer. The apparent rise in the afternoon and evening is likely due to continued industrial emissions of butenes in the HSC being advected over the Clinton Drive site by the afternoon Bay breeze. This is further illustrated in **Figure 5-33**, which shows an increase in both the number of data points from the south (the direction of highest influence for this factor), and the concentration of the factor in the late afternoon after a decrease during the middle of the day. This further demonstrates that concentrations can be heavily dependent on wind direction, and that emissions of these reactive compounds occur throughout the day.

Factor 14 (isoprene, mostly biogenic) had a peak in the early afternoon, typical of biogenic emissions. The presence of outlying concentrations during the nighttime again demonstrates that industrial isoprene emissions are also included in the factor.

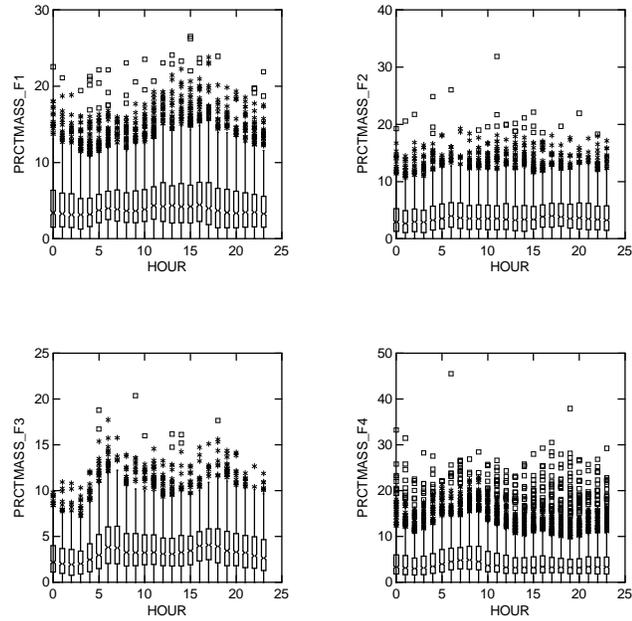


Figure 5-28. Notched box whisker plots of hourly weight percents of Factors 1 through 4.

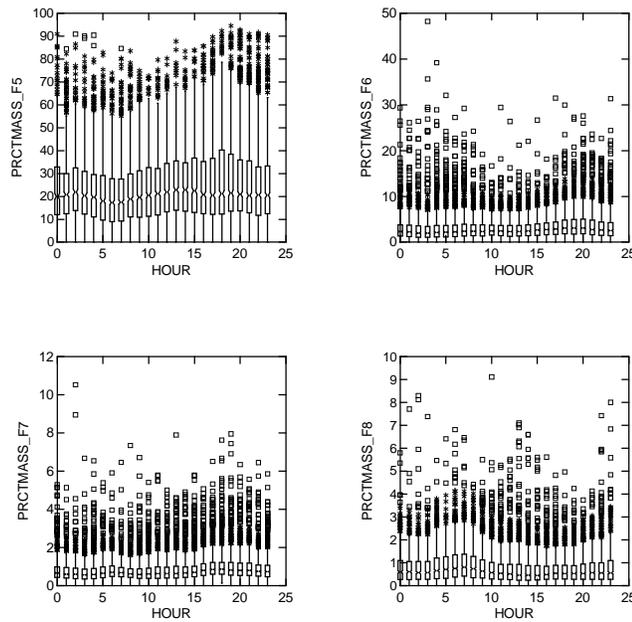


Figure 5-29. Notched box whisker plots of hourly weight percents of Factors 5 through 8.

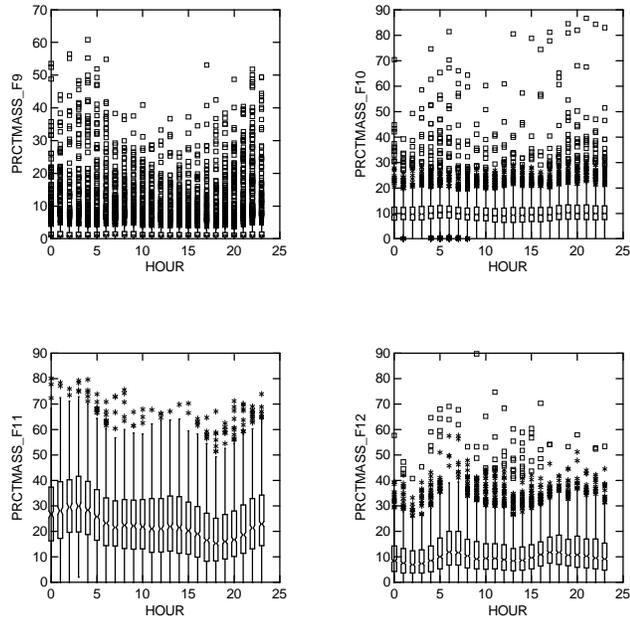


Figure 5-30. Notched box whisker plots of hourly weight percents of Factors 9 through 12.

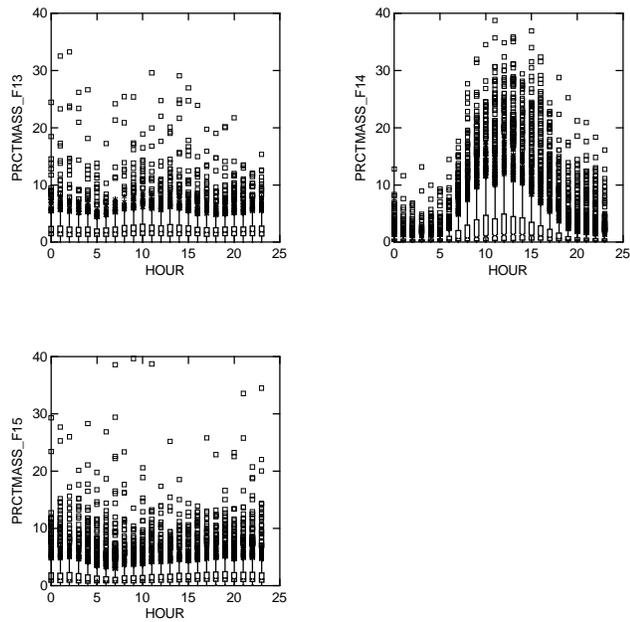


Figure 5-31. Notched box whisker plots of hourly weight percents of Factors 13 through 15.

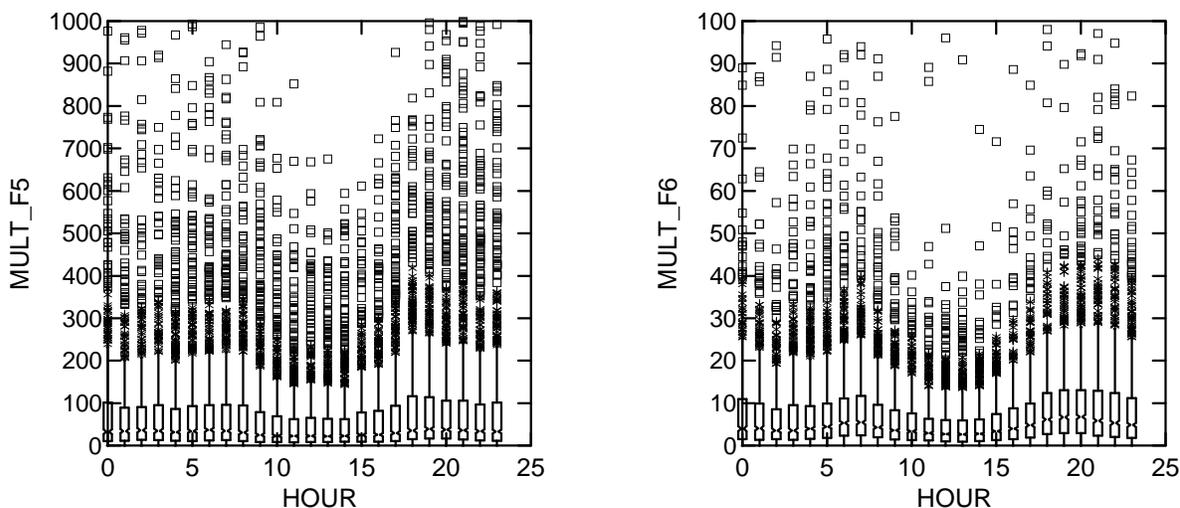


Figure 5-32. Notched box whisker plots of concentrations (ppbC) of Factors 5 and 6 by hour.

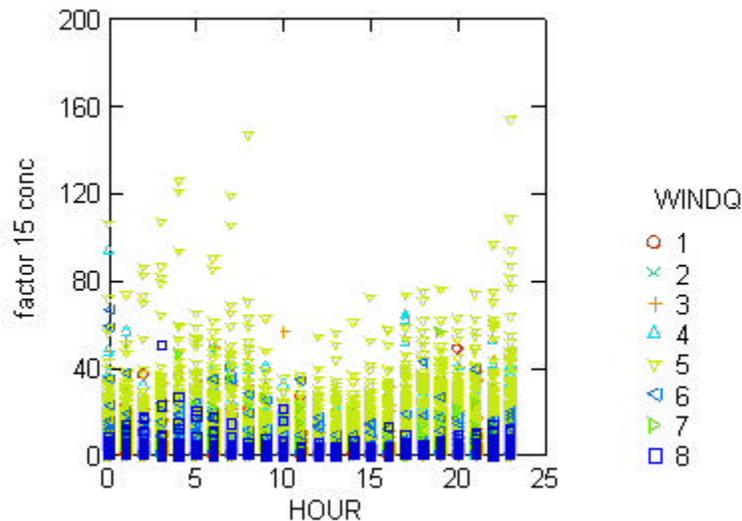


Figure 5-33. Concentration (ppbC) of Factor 15 by hour and by wind octant (1 = north, 5 = south).

5.4 WIND DIRECTION ANALYSIS

Meteorology and wind direction play an important role in both the level of VOC concentrations impacting a site and in ozone formation in the HSC. The median concentration and weight percent of each factor by wind direction was found in order to assist in identifying source areas of each factor. Results are shown in **Figures 5-34 through 5-48**. This analysis is further supplemented by the use of the conditional probability function (CPF) in Section 5.5.

Factor 1 (industrial flares) was found to be higher on both a concentration and weight percent basis with winds from the northwest, and by concentration with winds from the east. It is likely that the weight percent is lower from the east because this is the direction of high concentrations of other factors as well, so Factor 1 makes up less of the total loading. These directions suggest that this factor may have some motor vehicle influence from the freeways to the northwest in addition to industrial activity in the HSC to the east and northwest.

Factors 2 (propyl- and ethyl-benzene), 8 (industrial aromatic hydrocarbons), and 12 (heavy aromatics) exhibited concentration and weight percent spikes with winds from the south and southwest (2), southwest, west, and northeast (8), and south, southwest, and west (12). South, southwest, and west appear to be the directions of high aromatic emissions, though other source areas to the northeast (Factor 8) are also evident. The separation of these factors by PMF suggests that there may be distinct facilities or groups of sources for each factor, that emissions on a temporal scale are different (i.e., different release/upset times), and/or that the variation of the species is impacted by the species-specific depletion rates by photochemistry.

Factor 3 (motor vehicle) did not show a large difference among wind directions, though concentrations and weight percents were higher with winds from the northwest, the direction of the freeway, which is consistent with the identification of this factor. The lack of a definitive wind direction for this factor is also consistent with mobile emissions, since mobile emissions are emitted throughout the Houston area in every direction around the Clinton Drive site. A similar pattern is seen with Factor 13, identified as diesel emissions. This factor has multiple peaks from every direction except the north and northeast, indicative of a general background of diesel emissions likely from both trucks (freeway directions) and trains (tracks to the south of Clinton Drive).

Factor 4 (light olefins) was one of many factors that are likely industrial in origin and that had their highest concentrations (and weight percents) with winds from the east. This factor also had high concentrations from due south, the direction of another heavily industrialized area in the HSC. This trend was similar for factor 5 (butanes), with high concentrations and weight percents from the east and south.

Factors 6 (C6-C9 alkanes, solvents), 7 (pentenes), 9 (butadiene), 10 (C5-C7 alkanes, evaporative), and 15 (butenes) all exhibited their highest concentrations and weight percents with winds from the south, and their second highest concentrations with winds from the east and/or southeast. These factors may be related since they come from a similar source region, though with such a high density of sources along the HSC it is difficult to ascertain. Also, while this analysis suggests that there may be dominant sources of reactive butenes and pentenes, emission inventory maps (see Appendix B) show a number of significant sources of C4-C5 olefins in addition to those to the south of Clinton Drive. It may be that the sources to the south are closer to the Clinton Drive site, so emissions from this direction are fresher and have not been reacted away. Other source regions may impact the Clinton Drive site, but with their butenes and pentenes depleted by photochemistry.

Unlike previous factors, Factor 8 (industrial aromatic hydrocarbons) showed higher concentrations and weight percents with winds from the southwest, west, and northeast. This difference in source regions suggests that these trimethylbenzenes may have a number of

significant sources outside the immediate HSC to the south and east. This difference also suggests that this factor may be real and not caused by an analytical bias.

Factor 11 (accumulation/natural gas) exhibited a somewhat different distribution than other factors, with high concentrations from the north and east and high weight percent from the north. This factor is likely part general background, with the peaks in these directions due to accumulation and transport of these light paraffins from the north with the morning land breeze (Brown et al., 2002). The easterly spike may indicate the “fresher” emissions from the HSC area that are associated with many other factors of more reactive VOCs.

Factor 14 (biogenic + industrial isoprene) had peaks in concentration from the west, south, and east and weight percent from the west and southwest. These may be the directions of increased plant and tree coverage or industrial point sources of isoprene. A further investigation of Factor 14 concentrations by wind direction at day and night is demonstrated in **Figure 5-49**. This graph shows a number of high concentrations of Factor 14 occurring during nighttime hours (8 p.m. to 6 a.m.), which are most likely from industrial sources because biogenic activity is minimal during the nighttime. A number of these outliers, however, occur from the north (in addition to the west), which is not (by median) a direction of consistently high concentrations and may suggest yet another source of industrial isoprene in that direction.

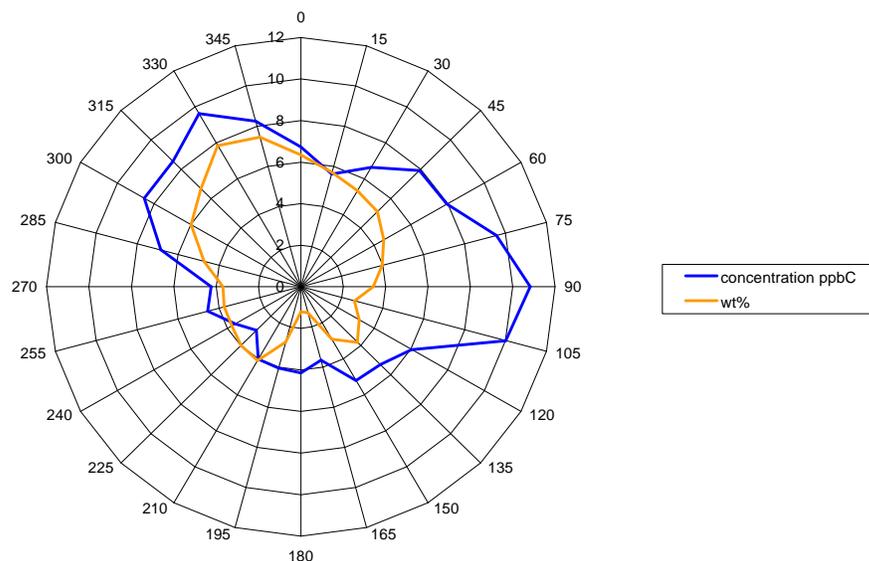


Figure 5-34. Median concentration and weight percent of Factor 1 by wind direction.

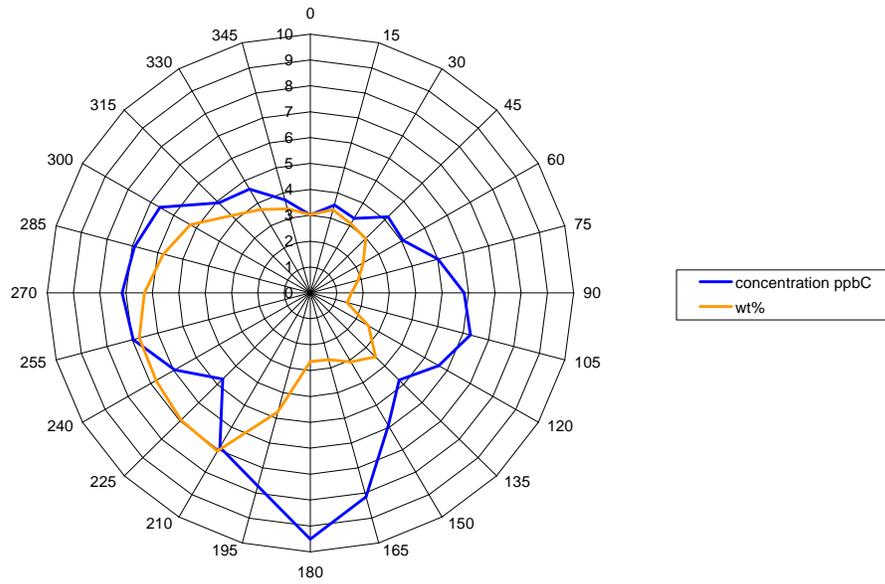


Figure 5-35. Median concentration and weight percent of Factor 2 by wind direction.

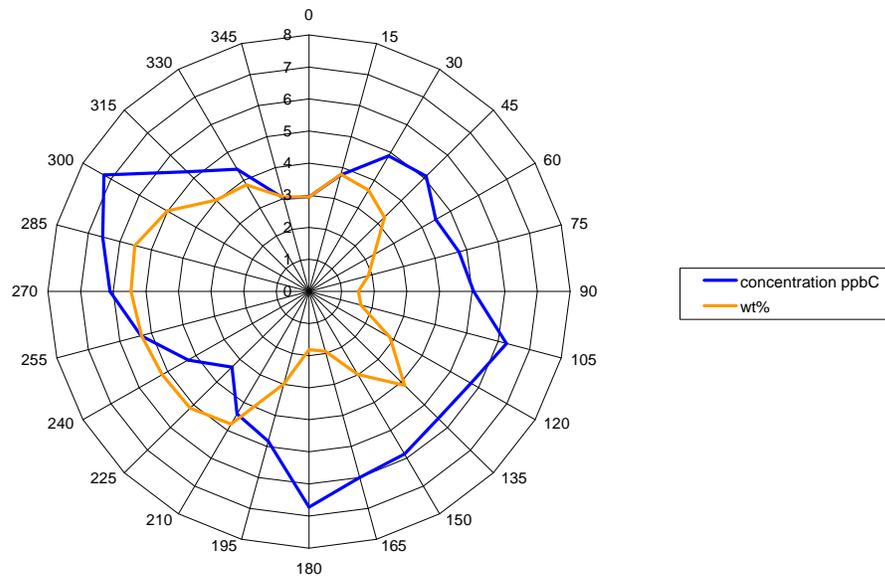


Figure 5-36. Median concentration and weight percent of Factor 3 by wind direction.

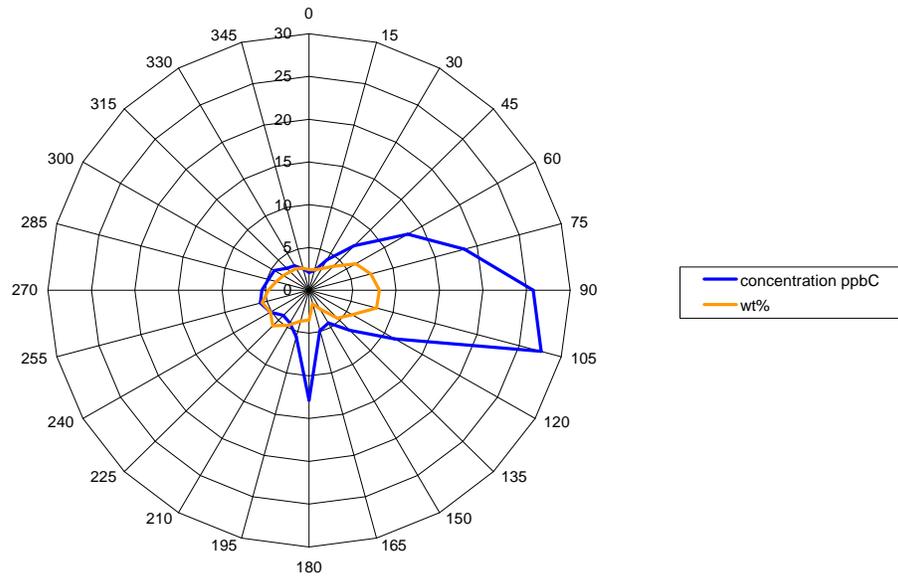


Figure 5-37. Median concentration and weight percent of Factor 4 by wind direction.

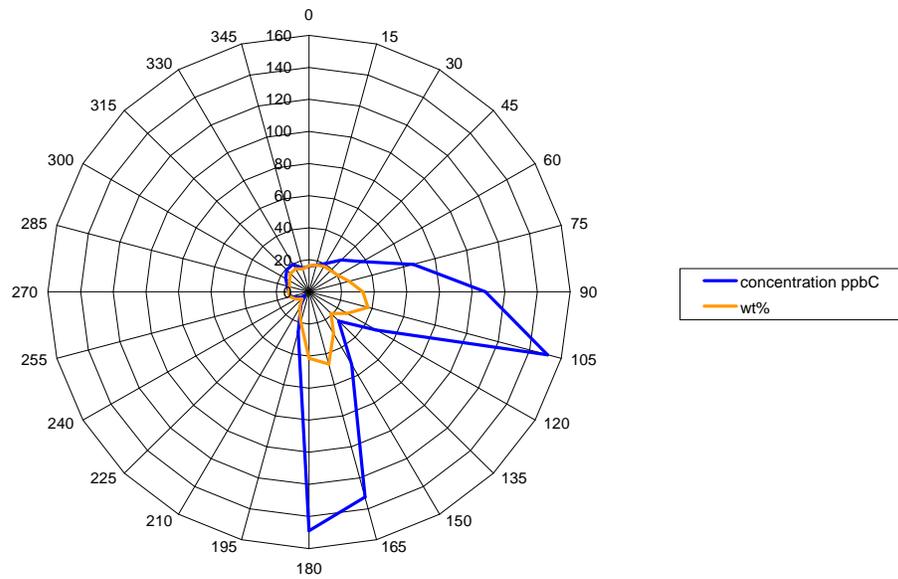


Figure 5-38. Median concentration and weight percent of Factor 5 by wind direction.

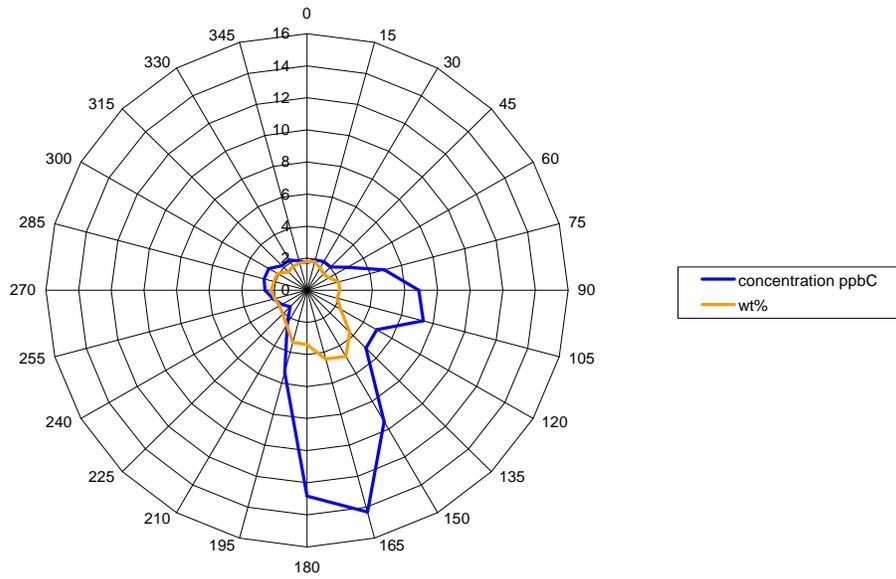


Figure 5-39. Median concentration and weight percent of Factor 6 by wind direction.

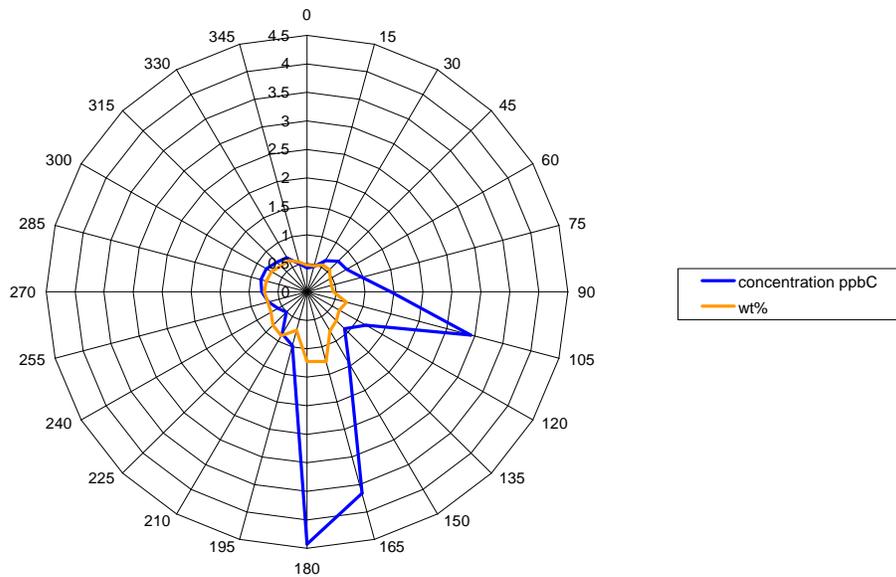


Figure 5-40. Median concentration and weight percent of Factor 7 by wind direction.

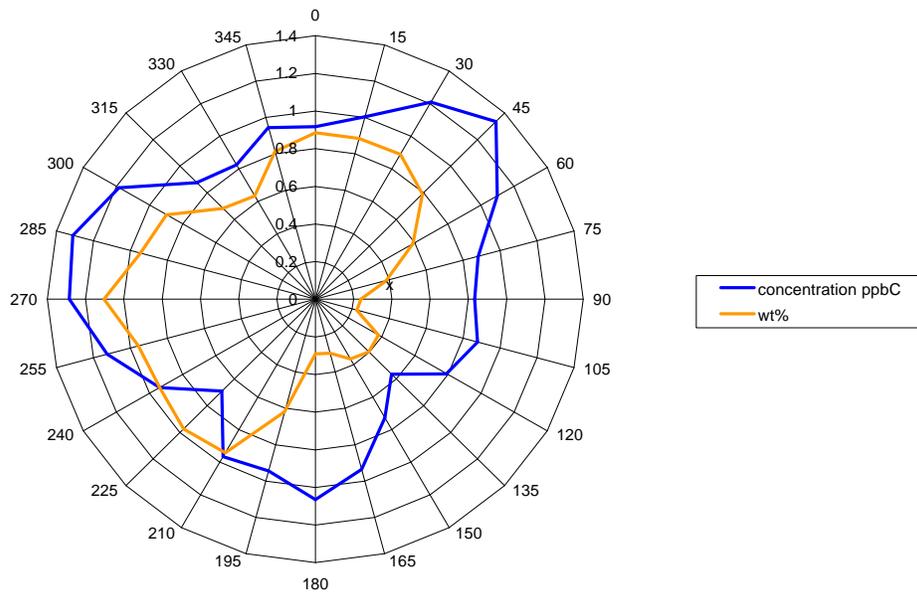


Figure 5-41. Median concentration and weight percent of Factor 8 by wind direction.

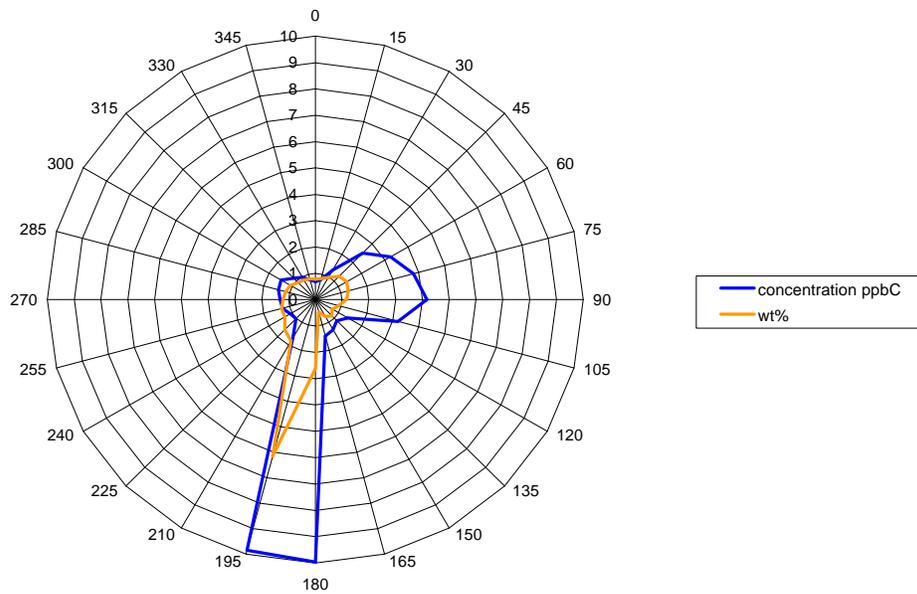


Figure 5-42. Median concentration and weight percent of Factor 9 by wind direction.

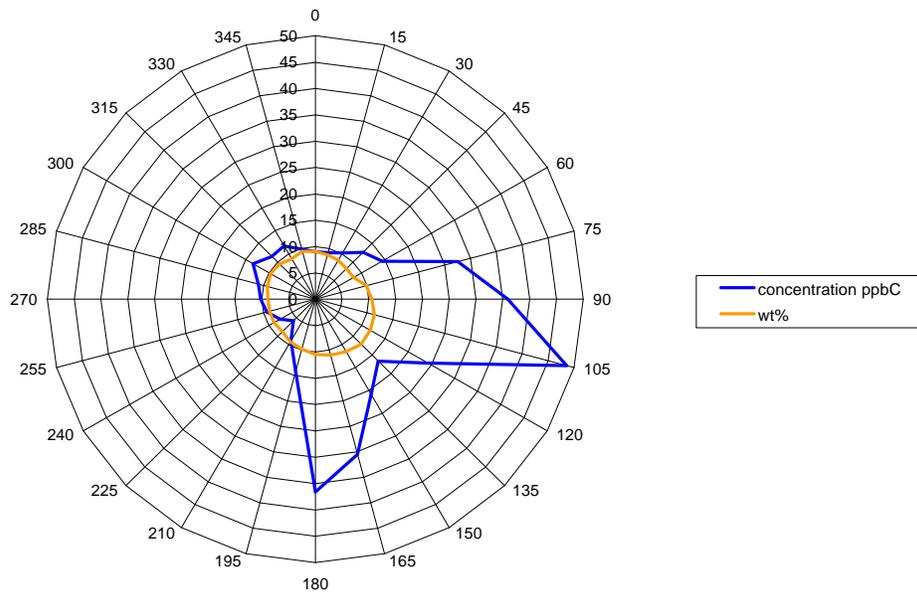


Figure 5-43. Median concentration and weight percent of Factor 10 by wind direction.

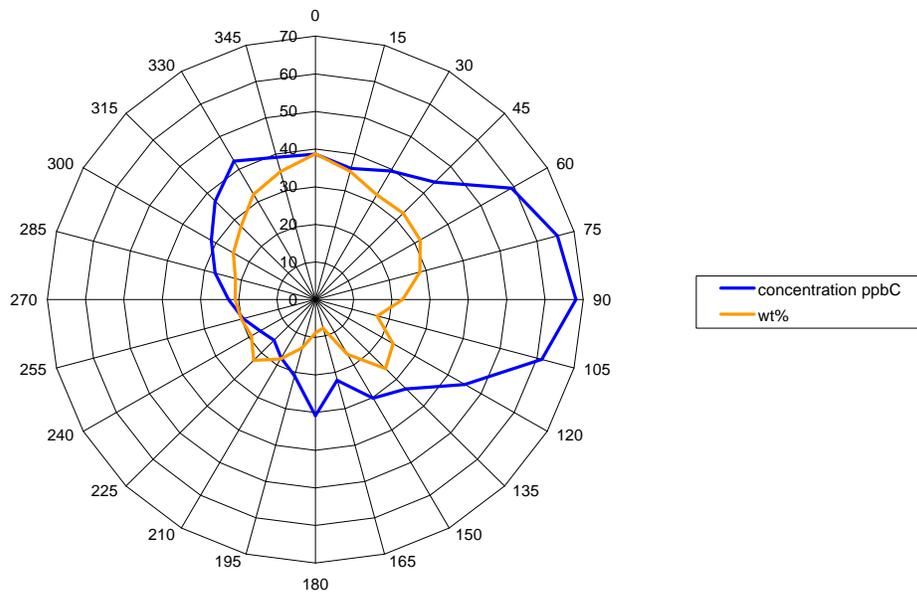


Figure 5-44. Median concentration and weight percent of Factor 11 by wind direction.

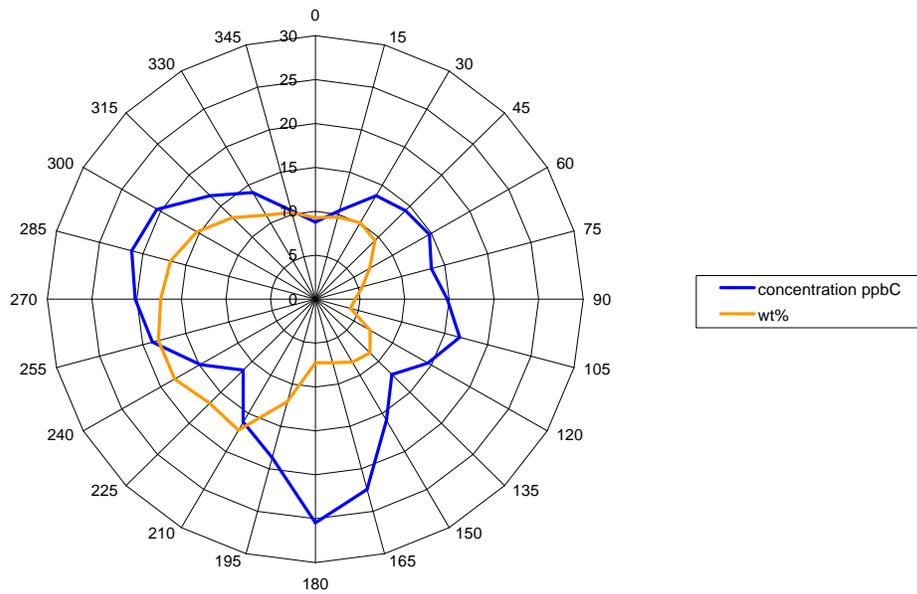


Figure 5-45. Median concentration and weight percent of Factor 12 by wind direction.

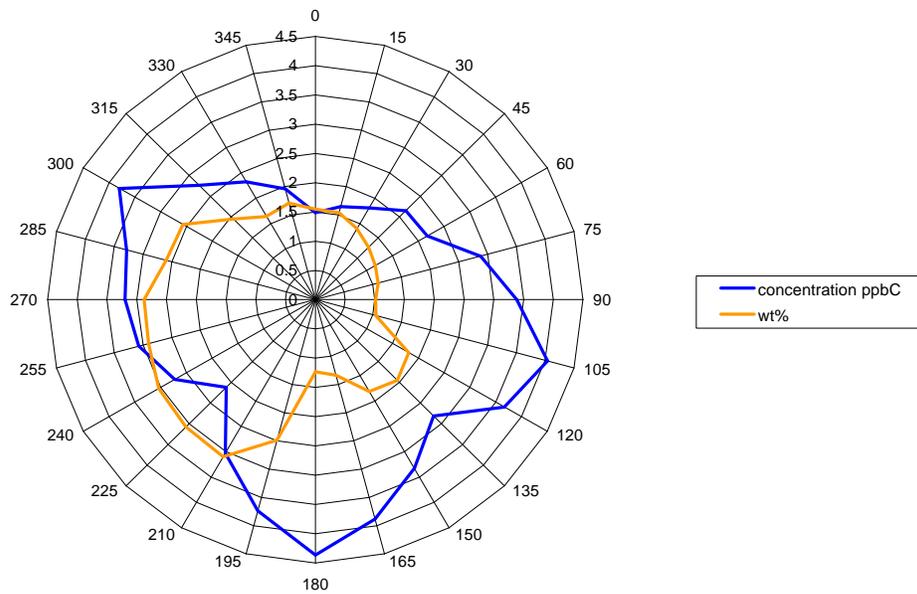


Figure 5-46. Median concentration and weight percent of Factor 13 by wind direction.

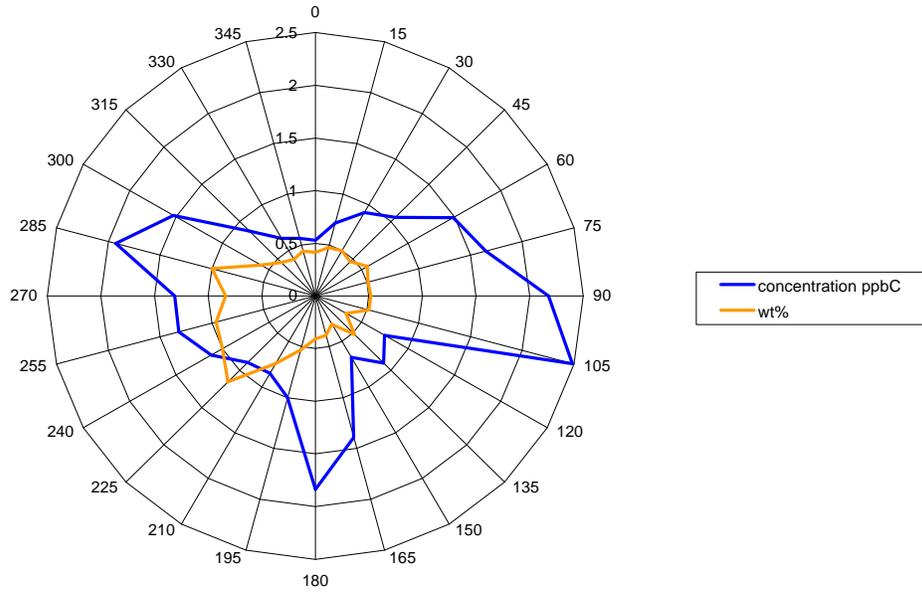


Figure 5-47. Median concentration and weight percent of Factor 14 by wind direction.

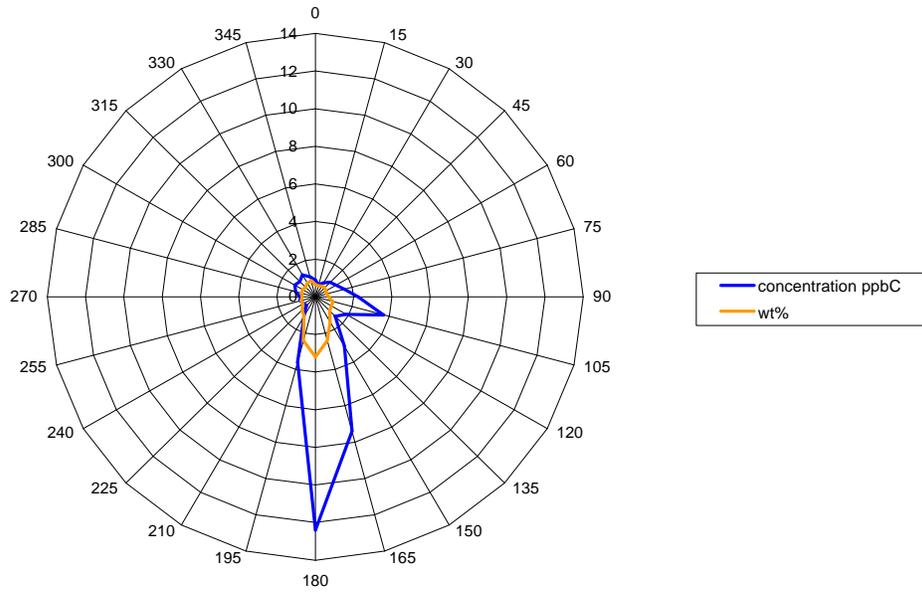


Figure 5-48. Median concentration and weight percent of Factor 15 by wind direction.

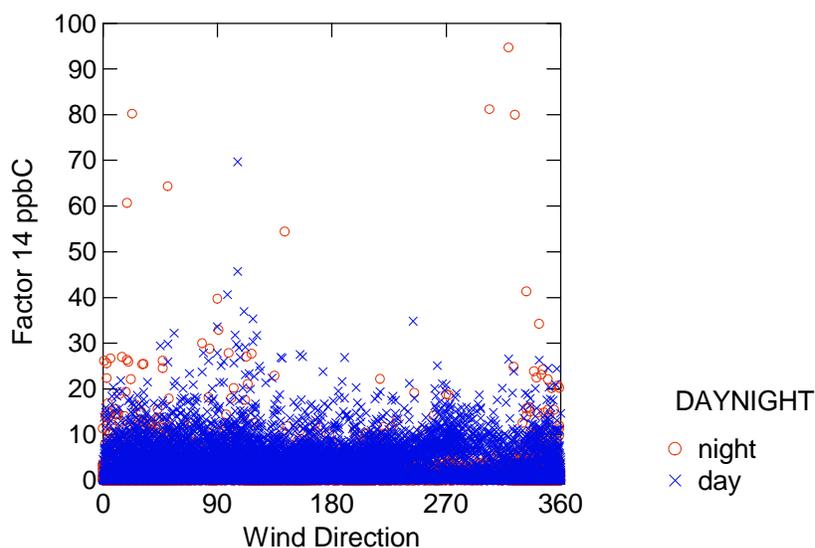


Figure 5-49. Factor 14 (isoprene) concentrations (ppbC) by wind direction by day (6 a.m.-8 p.m.) and night (8 p.m.-6 a.m.).

5.5 CONDITIONAL PROBABILITY FUNCTION

The conditional probability function (CPF), as previously described in Section 2.7, can be used to determine the areas from which factors are most likely to have high concentrations and, therefore, give a better estimation of source direction. The CPF was calculated for the top 25th percentile for each factor by 24 wind sectors of 15 degrees each for both concentration and weight percent. Results for each factor are shown in **Figures 5-50 through 5-64**.

Generally, results were similar to earlier analyses in Section 5.4 using median values. The high number of data points is likely the cause, and the consistence between the two methods further validates the association of factors with specific source regions. Some exceptions were evident, however, and may give a better idea of the location of significant sources. Factor 4 (light olefins) has a rise in weight percent probability to the southwest that was not seen when looking only at median values, suggesting that there are sources of these compounds in this direction in addition to the east and south. The eighth factor (industrial aromatic hydrocarbons) saw a spike in both concentration and weight percent in the CPF to the northeast that was not shown when using median values, yielding another potential source area besides the southwest and west.

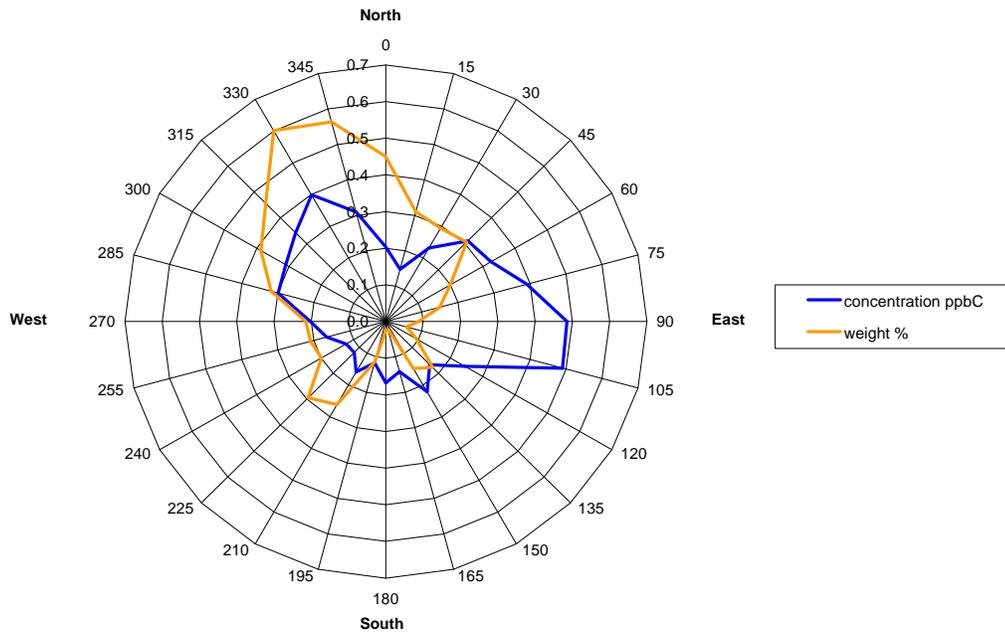


Figure 5-50. CPF of Factor 1 by concentration (ppbC) and weight percent.

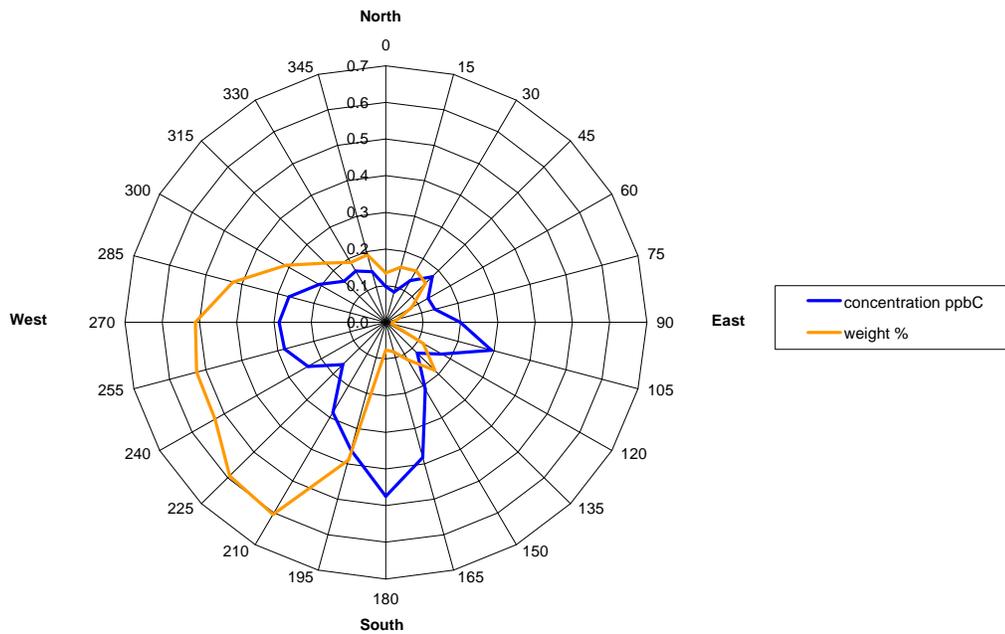


Figure 5-51. CPF of Factor 2 by concentration (ppbC) and weight percent.

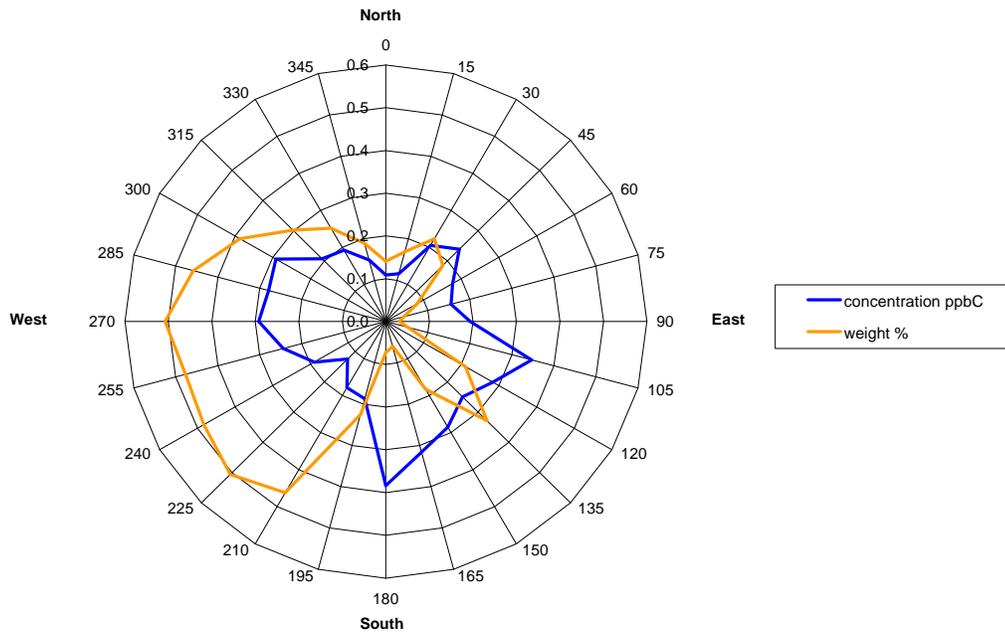


Figure 5-52. CPF of Factor 3 by concentration (ppbC) and weight percent.

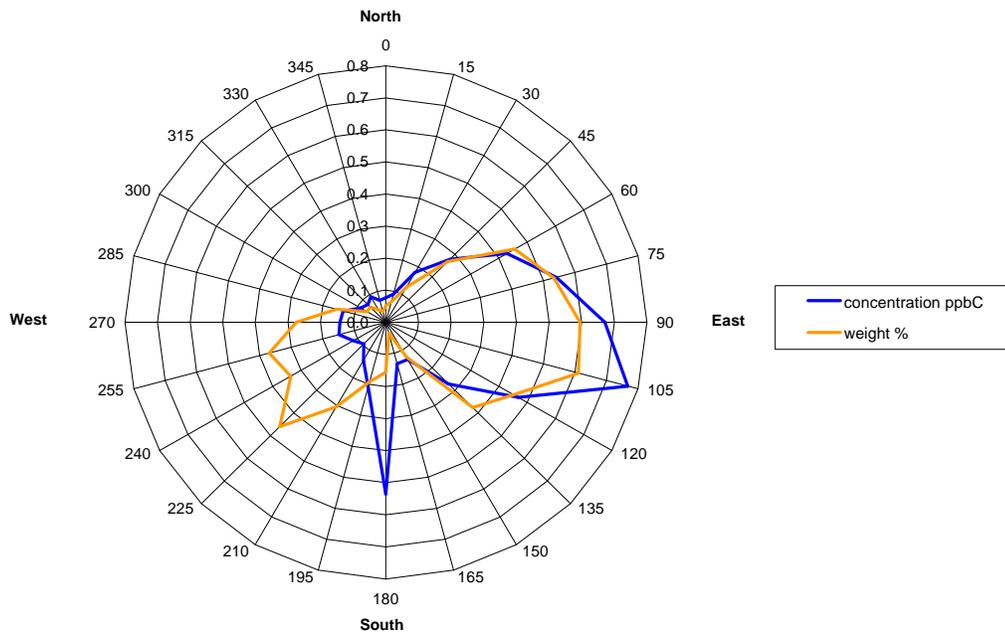


Figure 5-53. CPF of Factor 4 by concentration (ppbC) and weight percent.

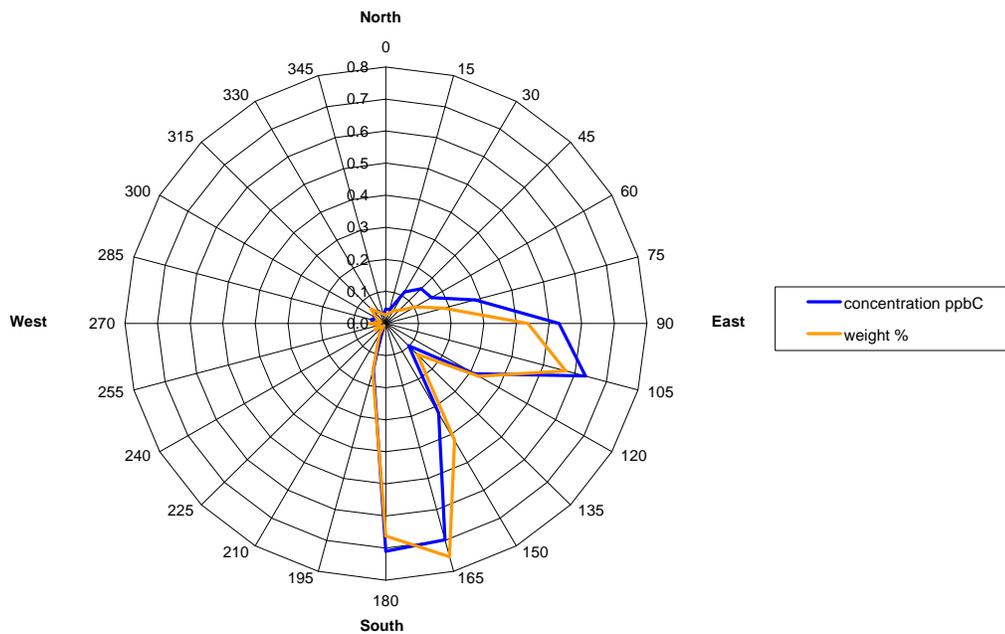


Figure 5-54. CPF of Factor 5 by concentration (ppbC) and weight percent.

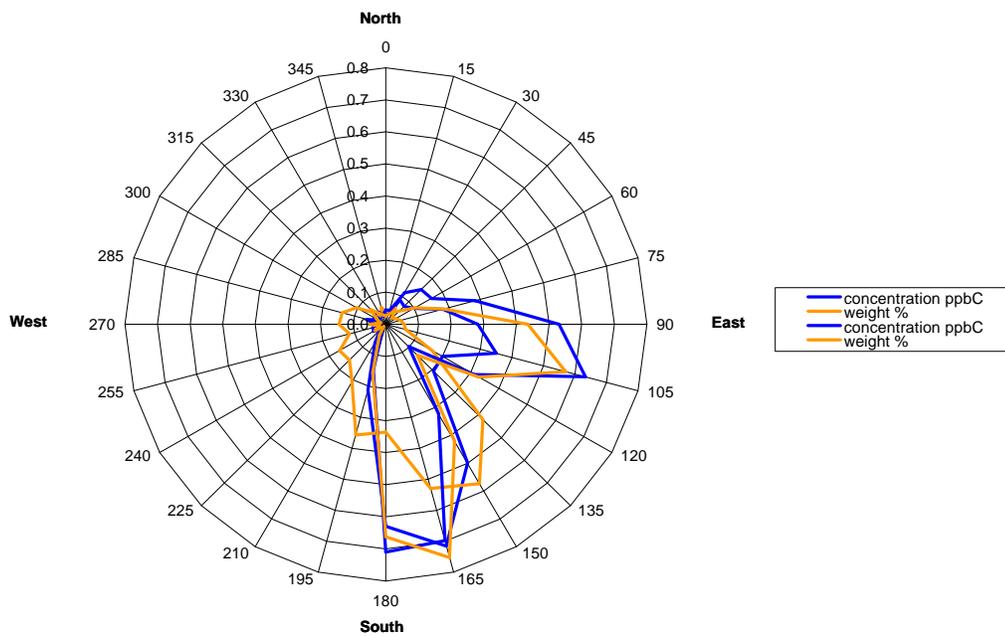


Figure 5-55. CPF of Factor 6 by concentration (ppbC) and weight percent.

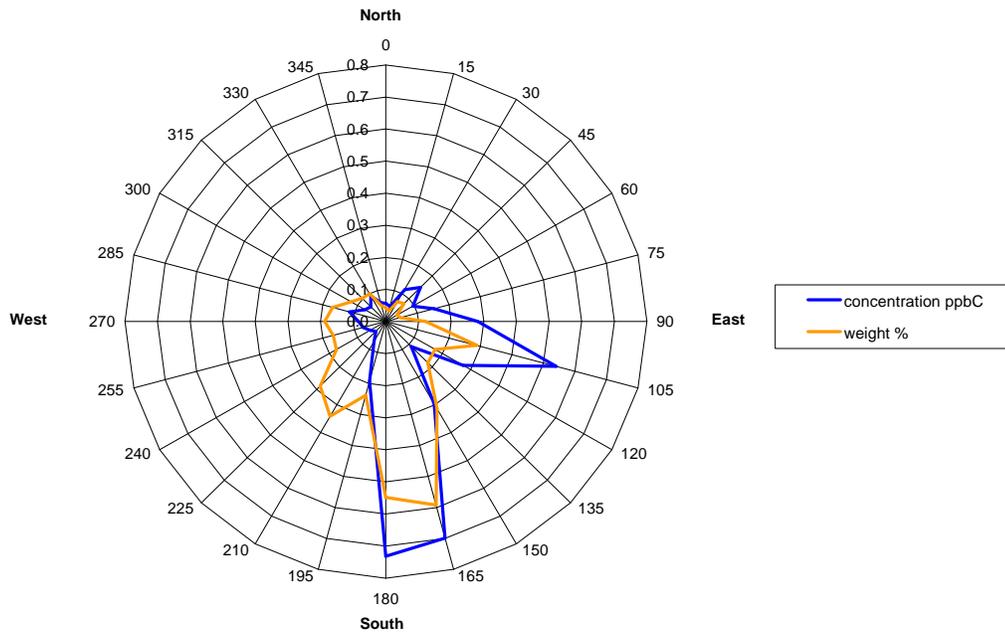


Figure 5-56. CPF of Factor 7 by concentration (ppbC) and weight percent.

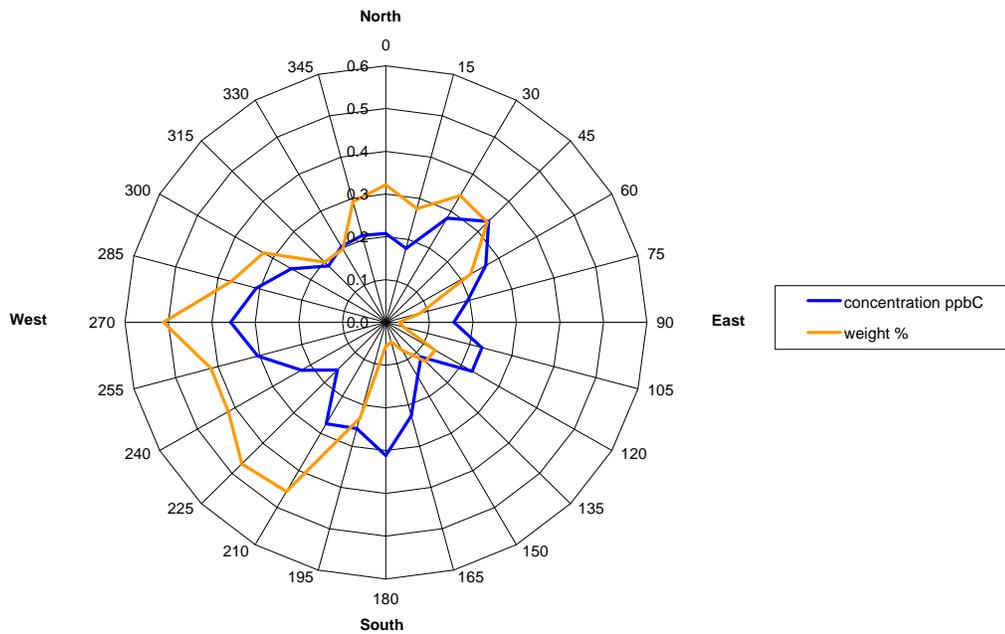


Figure 5-57. CPF of Factor 8 by concentration (ppbC) and weight percent.

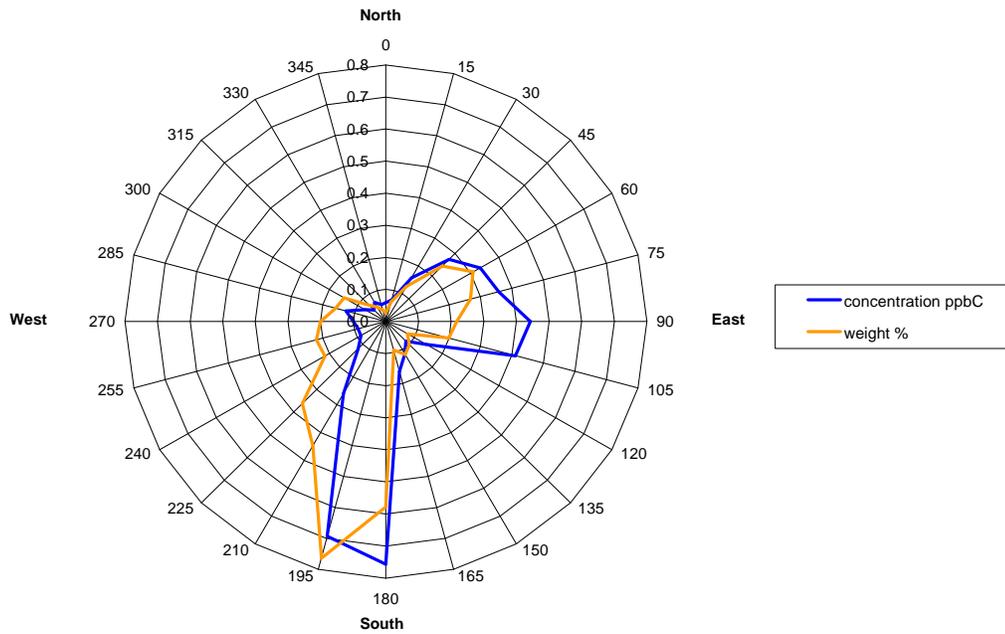


Figure 5-58. CPF of Factor 9 by concentration (ppbC) and weight percent.

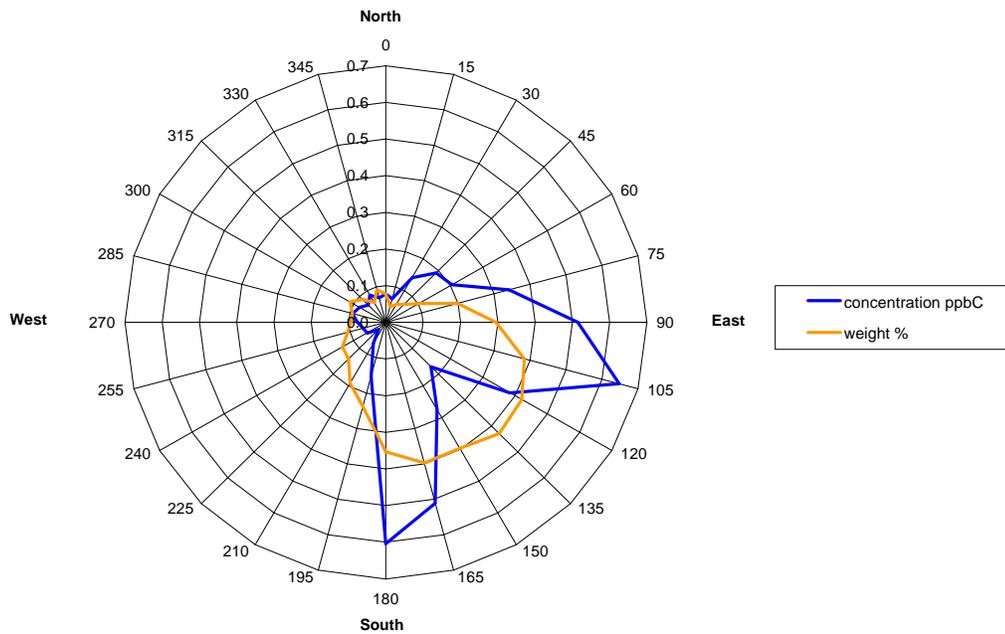


Figure 5-59. CPF of Factor 10 by concentration (ppbC) and weight percent.

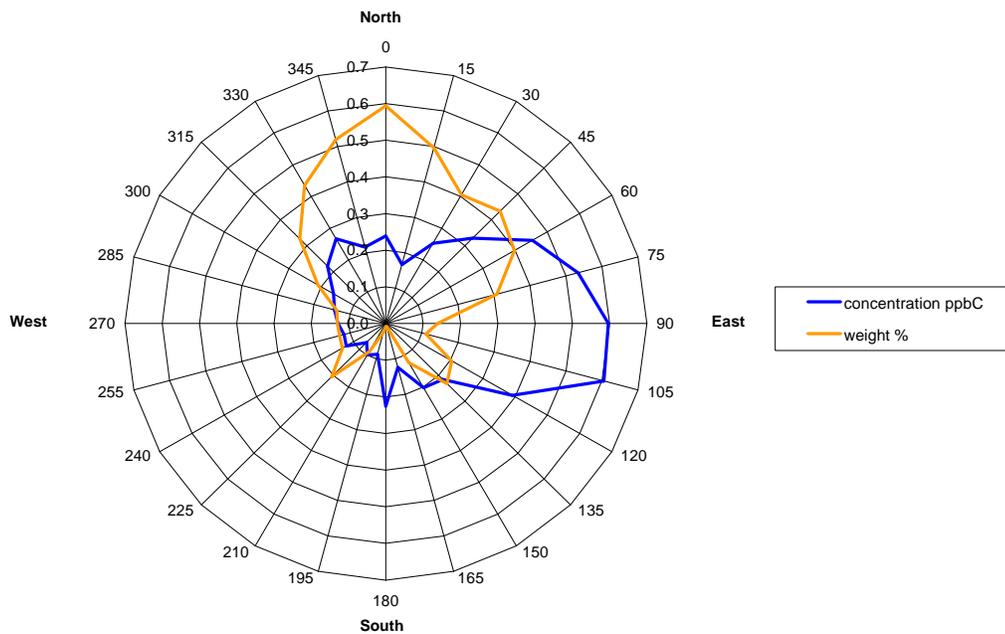


Figure 5-60. CPF of Factor 11 by concentration (ppbC) and weight percent.

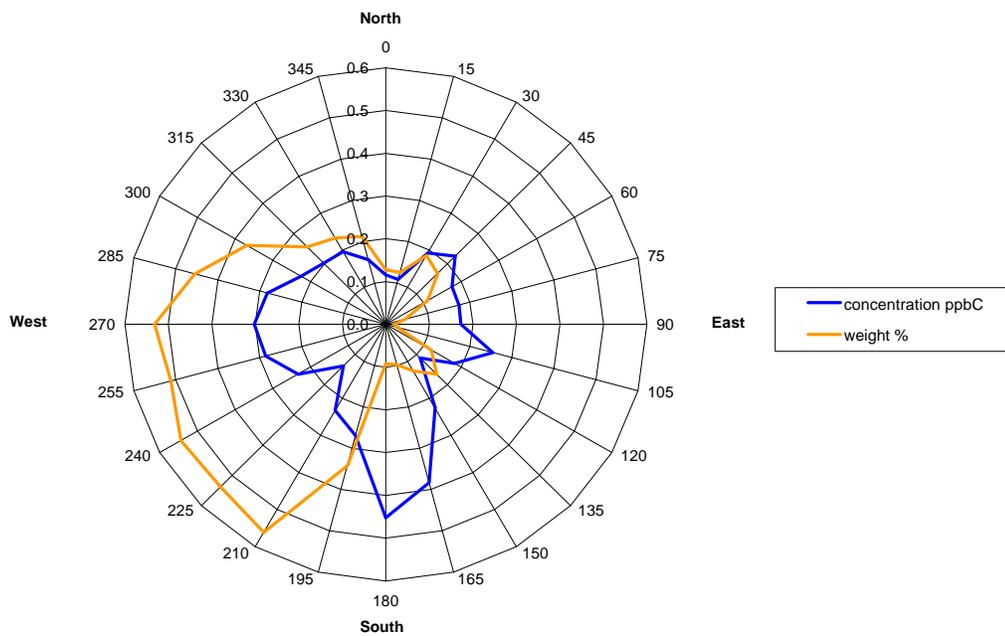


Figure 5-61. CPF of Factor 12 by concentration (ppbC) and weight percent.

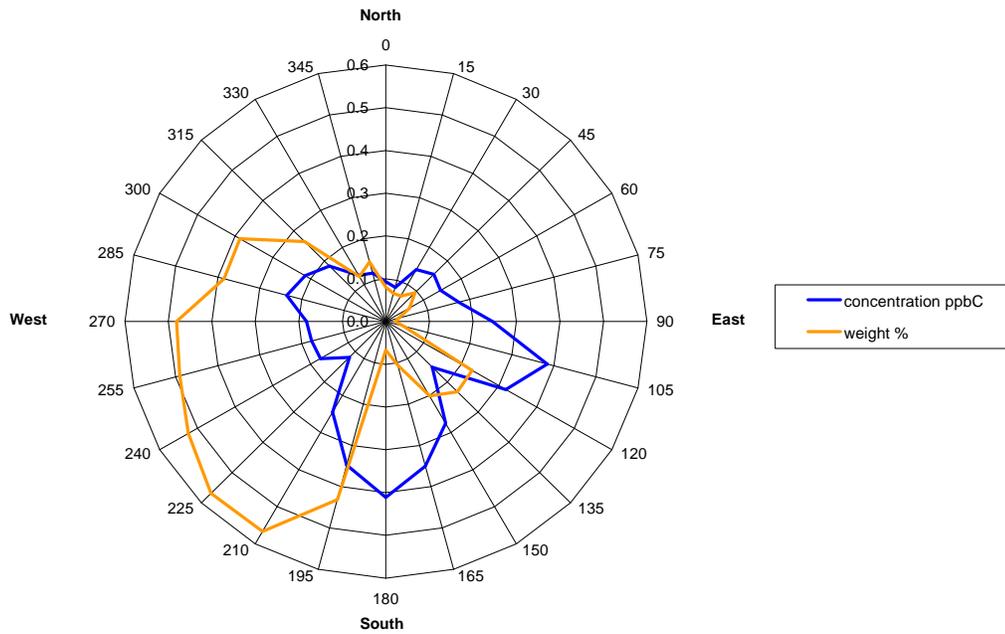


Figure 5-62. CPF of Factor 13 by concentration (ppbC) and weight percent.

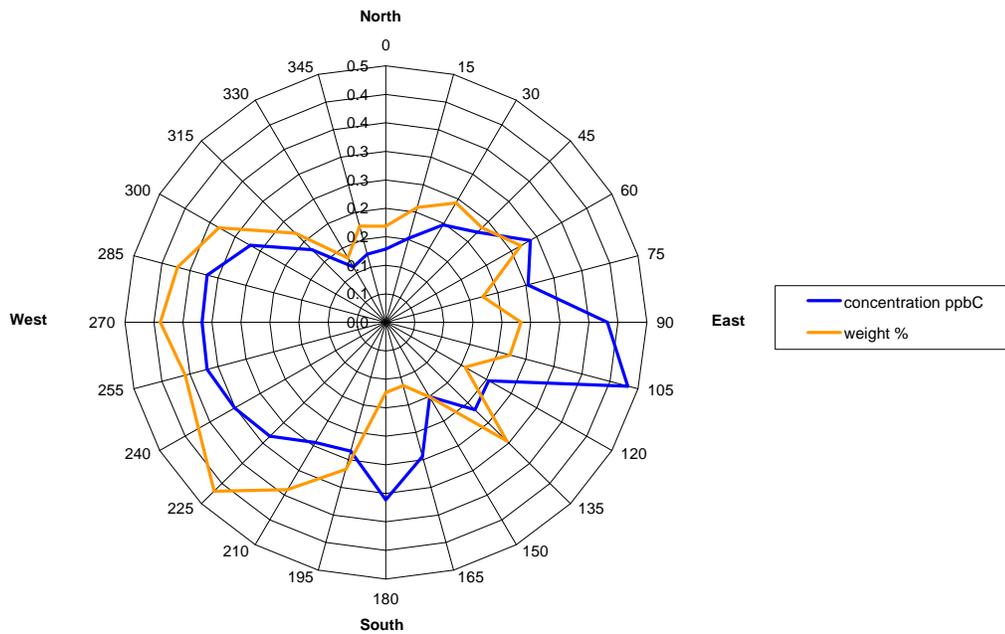


Figure 5-63. CPF of Factor 14 by concentration (ppbC) and weight percent.

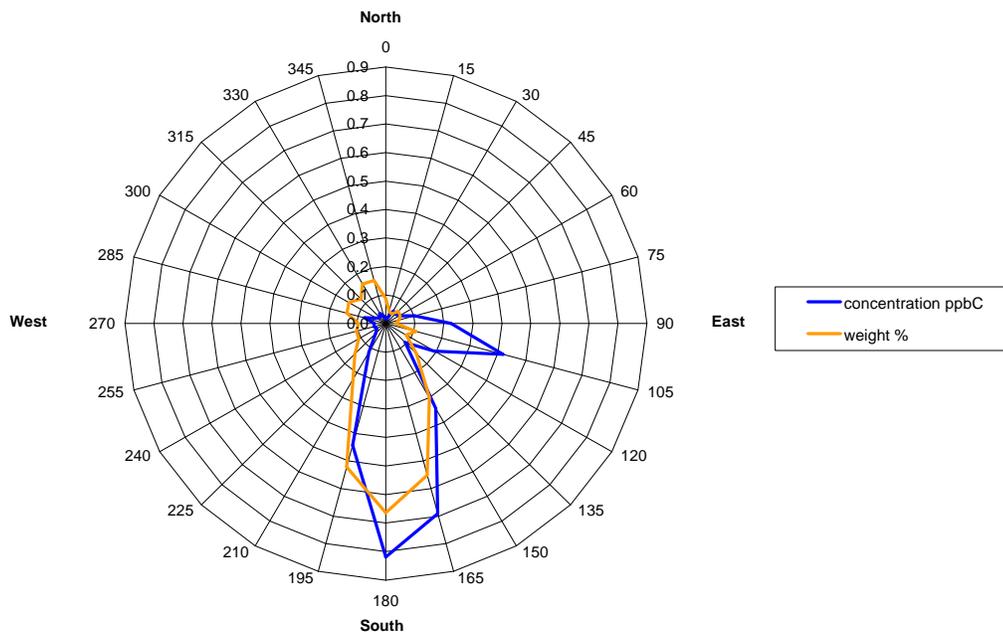


Figure 5-64. CPF of Factor 15 by concentration (ppbC) and weight percent.

5.6 SCALING SOURCE PROFILES BY REACTIVITY

Source profiles can also be scaled by ozone formation potential to examine the reactivity potential of each source. This can potentially give insight into what factors and source areas are the most important in ozone formation. Other variations on this analysis would be to run PMF from an already scaled data set, to scale weight percents and concentrations of each factor for every sample by an average reactivity (this could also lead to more detailed reactivity analysis between ozone episode and non-episode days), or to complete the CPF using a reactivity-scaled data set.

5.6.1 Ozone Production Potential: Reactivity Scales

The degradation of VOCs by photochemistry and the resulting conversion of NO to NO₂ and formation of ozone do not occur at the same rate for all VOCs. The ozone formation potential of a specific hydrocarbon depends on its concentration, structure, and removal pathways. If a reactive compound is low in concentration, it will generally not have a high ozone formation potential while a somewhat unreactive compound with a high concentration may have a larger ozone formation potential. One scale on which to gauge VOC ozone formation potential is the hydroxyl reactivity scale (OH) (Atkinson, 1989, 1994), which utilizes the reaction coefficient of an individual hydrocarbon with hydroxyl radical. This is strictly the rate at which the hydrocarbon is oxidized by hydroxyl radical only and does not consider competing removal mechanisms for either the VOC or hydroxyl radical or the influence from the overall composition of VOCs in an air mass.

Incremental reactivity (Carter, 1994, 2001) is the change in ozone caused by adding a small amount of test VOC to the emissions in an episode, divided by the amount of test VOC added: g ozone/g C or moles ozone/mole C. Incremental reactivity may be used to assess the effect of changing emissions of a given VOC on ozone formation, to compare the ambient VOC mix among sites or episodes, or to investigate VOCs important to ozone formation. This scale considers NO_x sinks as well as the generation and loss of hydroxyl radicals, all of which affect the rate of reaction for VOCs. The maximum incremental reactivity (MIR) scale was developed by W.P.L. Carter (1994) and used in "low emission vehicles and clean fuels" regulations in California. The MIR list was recently expanded to include more VOCs, and MIR values were updated (Carter, 2001).

In assessing VOC data, analysts have found that the MIR scale is most useful in a relative (i.e., whether an ambient sample more reactive than another) rather than absolute (i.e., how much ozone can be generated with this air parcel) manner. Furthermore, the uncertainty associated with MIR scale values and the notion that total reactivity equals the sum of incremental reactivities from individual species is unverified. The analyst needs a low unidentified fraction of total non-methane organic carbon (TNMOC) to best assess the potential reactivity of a hydrocarbon mixture. If high unidentified fractions exist, this analysis is less useful. When comparing samples, the weight percent of each hydrocarbon multiplied by its reactivity is often used. Scaling by a sample's TNMOC allows for differences of the entire sample to be assessed on a relative basis (see Equation 5-1), instead of on a per species basis (via concentration (reactivity as in Equation 5-2).

$$\text{By concentration: } [\text{ppbCHC}] * \frac{\text{molO}_3}{\text{molair}} = \frac{\text{molC}}{\text{molair}} * \frac{\text{molO}_3}{\text{molC}} = \text{ppbO}_3 \quad (5-1)$$

$$\text{By weight percent: } \frac{[\text{ppbCHC}]}{\text{ppbCTNMOC}} * \frac{\text{molO}_3}{\text{molC}} = \frac{\text{ppbO}_3}{\text{ppbCTNMOC}} \quad (5-2)$$

where:

HC = a particular hydrocarbon
 TNMOC = total non-methane organic carbon

There are a number of differences between the two reactivity scales. One is that carbonyl compounds are much more reactive on the MIR scale than on the OH scale. Another is that isoprene is much more reactive on the OH scale, so that even small amounts become significantly amplified. Propene is much more reactive than ethene on the OH scale, but less so on the MIR scale. Lastly, styrene is much more reactive on the OH scale than on the MIR scale, so that low amounts appear more significant on the OH scale than on the MIR scale. Values for a number of species on the OH and MIR reactivity scales are given in **Table 5-4**.

It is often useful to find the relative contribution of each hydrocarbon or species family to the total reactivity on both scales. This is done by dividing the individual compound's concentration or weight percent on the reactivity scale by the sum of all species' concentration or weight percent on the reactivity scale. There is no difference whether this reactivity composition is calculated by concentration or weight percent because both the numerator (hydrocarbon

weight percent x reactivity) and denominator (sum of all hydrocarbons x their reactivities) are scaled by the total identified fraction when using weight percent numbers. These values cancel out and yield the same result as if pure concentration values were used. This is shown in Equations 5-3 and 5-4.

$$\text{Contribution of HC by concentration: } \frac{[HC] \times R_{HC}}{\sum_k [HC_k] \times R_{HC_k}} * 100\% = \% \text{ Reactivity from HC} \quad (5-3)$$

$$\text{Contribution of HC by weight percent: } \frac{\frac{HC}{TNMOC} \times R_{HC}}{\sum_k \frac{HC}{TNMOC} \times R_{HC_k}} * 100\% = \% \text{ Reactivity from HC} \quad (5-4)$$

where:

- HC = a particular hydrocarbon
- R = reactivity coefficient
- TNMOC = total non-methane organic carbon

Table 5-4. Reactivity values (MIR and OH) for selected hydrocarbons.

Compound	MIR Reactivity (mol O ₃ /mol C)	OH Reactivity (rate constant with OH (10 ¹²) (cm ³ molecule ⁻¹ s ⁻¹))
Ethene	2.65	8.5
Propene	3.38	26.3
n-butane	0.4	2.4
Trans-2-butene	4.07	64
Isopentane	0.51	3.7
Cis-2-pentene	2.99	67
m/p-xylene	2.06	23.6
Toluene	1.09	5.95
1,3,5-trimethylbenzene	3.12	57.5
Isoprene	3.03	101

5.6.2 PMF Sources Scaled by MIR Reactivity

The species in the 15 sources identified by PMF were multiplied by their respective MIR reactivity value. This allows for the total reactivity of each factor to be calculated, by summing the scaled values for each species in each factor. The reactivities of each factor are presented in **Figure 5-65** as a pie chart similar to the mass apportionment shown in Figure 5-4.

Factor 7 (pentenes) was found to be the source with the highest total reactivity, followed by Factors 4 (light olefins), 8 (industrial aromatic hydrocarbons), 3 (motor vehicle), and 15 (butenes). Other factors with a high total reactivity include 9 (butadiene), 14 (isoprene), and 1 (industrial flares). These results are generally what would be expected, because these are the

factors with the most reactive compounds. One note is that Factors 2 (industrial aromatic hydrocarbons) and 12 (mixed aromatic) were low in total reactivity, despite having reactive aromatic hydrocarbons; this was mostly due to a high amount of mass in these factors from the unidentified fraction, whose reactivity potential is unknown. If this fraction is high in carbonyl compounds such as formaldehyde and acetaldehyde, then these factors would have a significantly larger reactivity. Other factors with low reactivity potential include Factors 5 (butanes), 6 (C6-C9 paraffins), 10 (C5-C7 paraffins), 11 (light paraffins), and 13 (diesel). All of these factors have their largest mass contributions from paraffins, which have a relatively low reactivity.

Overall, these results are consistent with earlier reactivity analyses of auto-GC data (Brown and Main, 2002) that showed that no single compound, or even compound class (i.e., olefins), dominated the total reactivity. The contribution from the pentene, butene, butadiene, and isoprene factors is higher than results using only the pentenes, butenes, 1,3-butadiene, and isoprene species, respectively, and may indicate that these species are more important than has been previously indicated. The high reactivity associated with Factors 4 (light olefins), 8 (industrial aromatic hydrocarbons), and 3 (motor vehicles) is consistent with other results with the light olefins, heavy aromatic hydrocarbons, and toluene and xylenes. Factor 1's reactivity is consistent with a mix of contributions from ethene, n-butane, and acetylene. The low reactivity calculated for Factor 5 (butanes) is lower than previous analyses have suggested, in which the C4-C5 alkanes were 12%-18% of the average reactivity at various auto-GC sites in Houston (Brown and Main, 2002). This result from PMF seems to make more physical sense because these compounds are not very reactive. In the PMF analysis, the butanes factor appears to have less mass than the total butanes since some butanes are apportioned into other factors (such as Factors 1, 4, 9, 11, and 15), thus reducing the impact the butane factor has on the total reactivity.

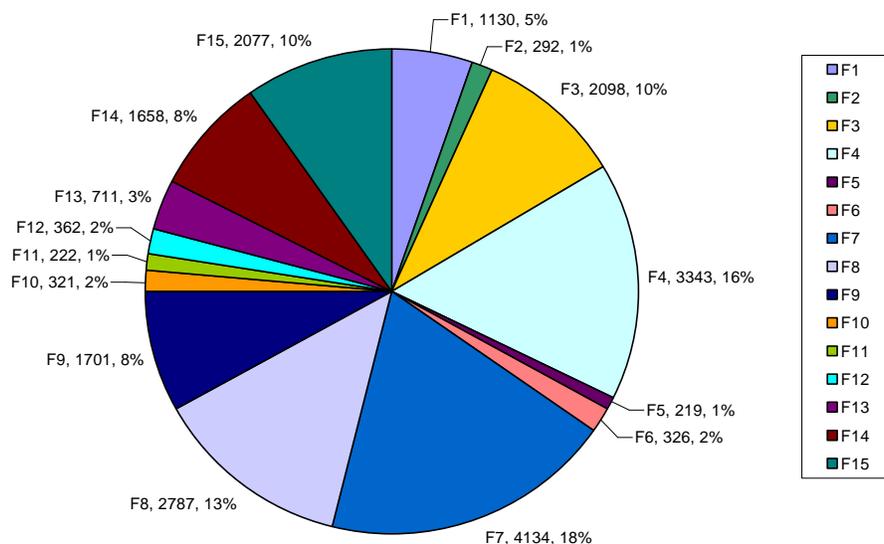


Figure 5-65. Total reactivity (concentration * MIR) by factor.

5.7 SOURCE STRENGTH ON OZONE EPISODE DAYS

TCEQ's definition of an ozone "episode" as a day on which a 1-hr average ozone concentration exceeded 125 ppb at an ozone monitor in the Houston area was used. The list of episodes was provided by TCEQ and confirmed by STI in order to be consistent with other investigations. All samples at all auto-GC sites on these days were then flagged as an episode day.

The strength of each source during mornings (0500-0900 CST) of ozone episodes and non-episodes during the summer (June-September 1998-2001) were investigated in order to determine what factors, if any, are higher on a weight percent basis on episode days and may therefore be linked to high ozone events. The median weight percent of each factor on episode and non-episode mornings during the summer is shown in **Figure 5-66**. Results of two-sample t-tests investigating whether differences between episode and non-episode mornings were significant are detailed in **Table 5-5**. These analyses utilize 1005 hourly data points of non-episode morning data and 202 hourly data points of morning episode data.

Nine of the 15 factors' weight percents were higher (by median) on episode mornings than on non-episode mornings; of these, 6 had statistically significant differences. These include Factors 1 (industrial flares), 2 (industrial aromatic hydrocarbons), 3 (motor vehicles), 10 (mid-range paraffins, solvents), 11 (light paraffins), and 12 (heavy aromatic hydrocarbons). Notched box whisker plots of these factors by episode and non-episode are shown in **Figures 5-67 through 5-69**. The significant rise of the paraffins (Factors 10, 11, and some 1) on ozone episode days may actually be due to the breakdown of more reactive compounds that resulted in both the high ozone and paraffins as secondary products. The heavy aromatic hydrocarbons in Factors 2 and 12 and partly in 3 may indicate that these compounds are more important to ozone formation than previously thought. While it is doubtful that these aromatic hydrocarbons were mainly responsible for the high levels of ozone on these episode days, it may be that they were sufficient to increase the ozone level just enough to trigger an ozone episode, since levels of other reactive compounds (light olefins, butenes, pentenes) can be high at any time of the day, week, and year and may force a high "background" level of ozone.

The motor vehicle signature (Factor 3) was also higher on ozone episode mornings, and while still a small amount of the total VOCs (about 6%), this is consistent with earlier analyses showing toluene to be significant at the Clinton Drive site on episode mornings in 1998 and 1999. Higher industrial flares contribution (Factor 1) on episode mornings is also significant, because previous analyses were not able to identify or apportion this source using individual compounds. Overall, the fact that multiple factors are significantly higher on ozone episode mornings suggests that this source apportionment analysis may be more effective in gauging what sources impact ozone formation than analysis of individual compounds, which gave mixed results (Brown and Main, 2002).

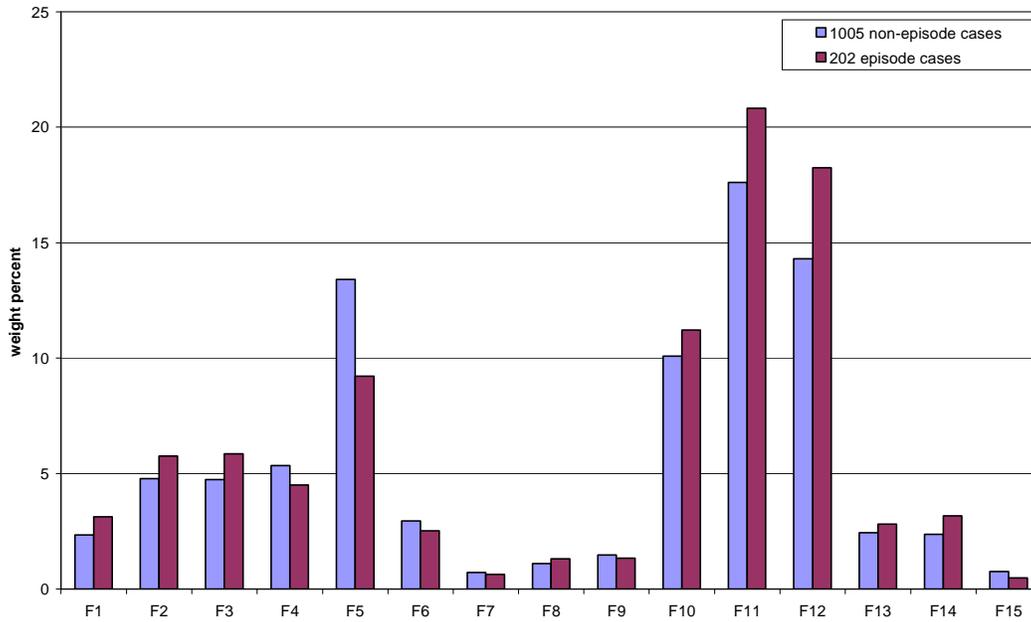


Figure 5-66. Median source strength (weight percent) on mornings (0500-0900 CST) of ozone episodes and non-episodes June-September 1998-2001.

Table 5-5. Results of two-sample t-tests for each factor in the June-September 1998-2001, 0500-0900 CST: whether episode or non-episode median weight percents are higher and whether these differences are different at a 95% confidence level.

Factor	Which median weight percent is higher?	Significant at a 95% confidence level?
1	Episode	Yes
2	Episode	Yes
3	Episode	Yes
4	Non-episode	Yes
5	Non-episode	Yes
6	Non-episode	Yes
7	Non-episode	No
8	Episode	No
9	Non-episode	Yes
10	Episode	Yes
11	Episode	Yes
12	Episode	Yes
13	Episode	No
14	Episode	No
15	Non-episode	Yes

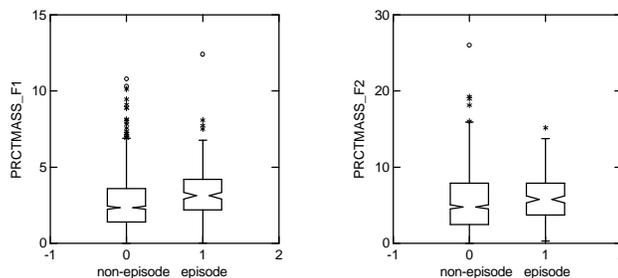


Figure 5-67. Notched box whisker plots of Factors 1 and 2 weight percent on non-episode and episode mornings (0500-0900 CST) during June-September 1998-2001.

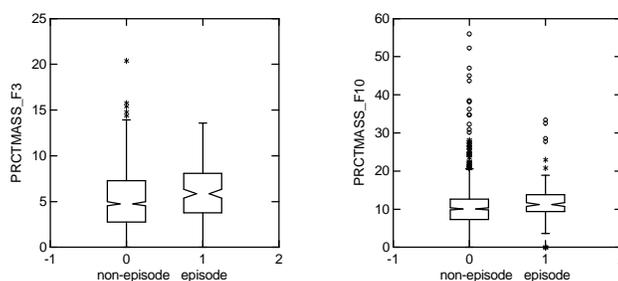


Figure 5-68. Notched box whisker plots of Factors 3 and 10 weight percent on non-episode and episode mornings (0500-0900 CST) during June-September 1998-2001.

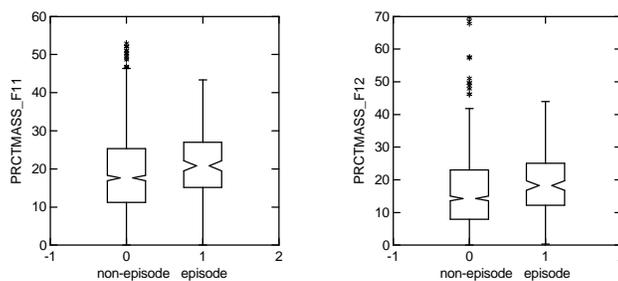


Figure 5-69. Notched box whisker plots of Factors 11 and 12 weight percent on non-episode and episode mornings (0500-0900 CST) during June-September 1998-2001.

5.8 OUTLYING RESIDUALS AND HIGH OZONE

With the combination of a large data set, a high number of outliers of most species, the fact that these outliers are all likely true concentrations, and that the uncertainties are not completely characterized, it is inevitable that some residuals will exceed ± 3 standard deviations. While the number of outlier residuals is small (i.e., less than 0.1% of the data for each species), it is interesting to investigate whether these outlying residuals are linked to high

ozone. If so, this could help further identify what conditions are needed for accelerated ozone formation.

The residuals for a number of reactive species were examined versus ozone concentrations. Results for ethene, propene, 1,3-butadiene, t-2-butene, toluene, and xylenes are shown in **Figures 5-70 through 5-72**. Generally, most high ozone concentrations occur where residuals are small and not beyond the ± 1 range. This indicates that the poorly modeled points are not linked to high ozone, so analyses of PMF results and ozone episodes appear to be valid.

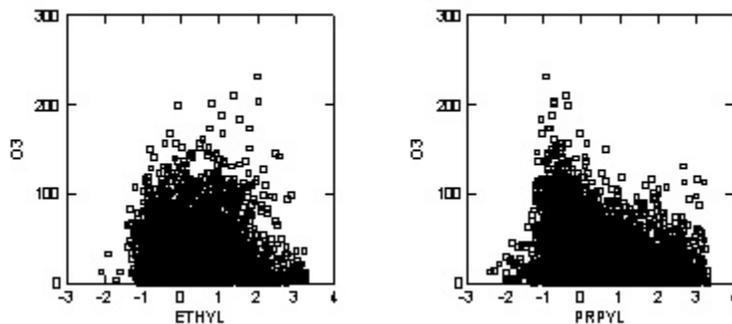


Figure 5-70. Residuals of ethene (ethyl) and propene (prpyl) versus ozone concentration during May-October 1998-2001.

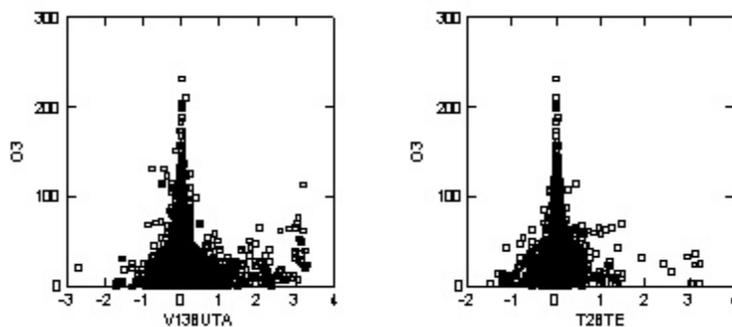


Figure 5-71. Residuals of 1,3-butadiene (v13buta) and trans-2-butene (t2bte) versus ozone concentration during May-October 1998-2001.

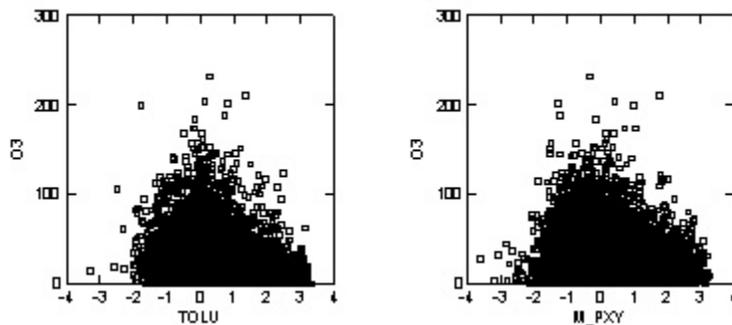


Figure 5-72. Residuals of toluene (tolu) and m/p-xylenes (m_pxy) versus ozone concentration during May-October 1998-2001.

5.9 UNMIX SOLUTIONS

UNMIX was also utilized to gain a different perspective on the source apportionment results from PMF. The same data set was used, except that missing data were treated as missing and not substituted by their average concentration as done in the PMF analysis. UNMIX does not have the corresponding error matrix that PMF has to downweight these points. Additionally, not all species that were used in PMF were used in UNMIX, because at some point the addition of additional species forces a decrease in the number of solutions found with UNMIX, or leads to no feasible solution being found. Utilizing twenty important species, listed in **Table 5-6**, only five factors were found using UNMIX; these are detailed in **Figures 5-73 through 5-77**. An additional note is that some species had negative contributions to a factor, which is physically impossible. For the UNMIX solutions, zero was substituted for these negative values in the graphs. In contrast, one strength of PMF is that it uses a non-negativity constraint.

The first UNMIX factor is dominated by isobutane, with some n-butane as well. This is analogous to the PMF butane factor (5). The second UNMIX factor had all of the 1,3-butadiene, as well as some butene and C2-C3 paraffins and ethene. This is similar to PMF Factors 4, 11, and 15. The third factor had most of the propene, as well as contributions from the light paraffins, ethene, n-butane, and pentanes. Factor 4 had butenes, pentenes, pentanes, and unidentified fraction. The fifth factor had toluene, benzene, acetylene, xylenes, and C10-C11 paraffins, indicative of mobile source influence, as well as accumulation species such as ethane and propane. Overall, PMF appears to yield a more detailed and understandable solution than UNMIX; further work with this model is needed, since both factor analysis and PMF suggest that more than the 5 factors identified by UNMIX exist.

Table 5-6. Species used in UNMIX and their abbreviations.

Abbreviation	Species
ETHAN	Ethane
ETHYL	Ethene
PROPA	Propane
PRPYL	Propene
ISBTA	Isobutane
NBUTA	n-butane
ACETY	Acetylene
T2BTE	t-2-butene
ISPNA	Isopentane
NPNTA	n-pentane
T2PNE	t-2-pentene
BENZ	Benzene
TOLU	Toluene
EBENZ	Ethylbenzene
M_PXY	m/p-xylenes
V124TMB	1,2,4-trimethylbenzene
NDEC	n-decane
NUNDC	n-undecane
V13BUTA	1,3-butadiene
UIDVOC	unidentified

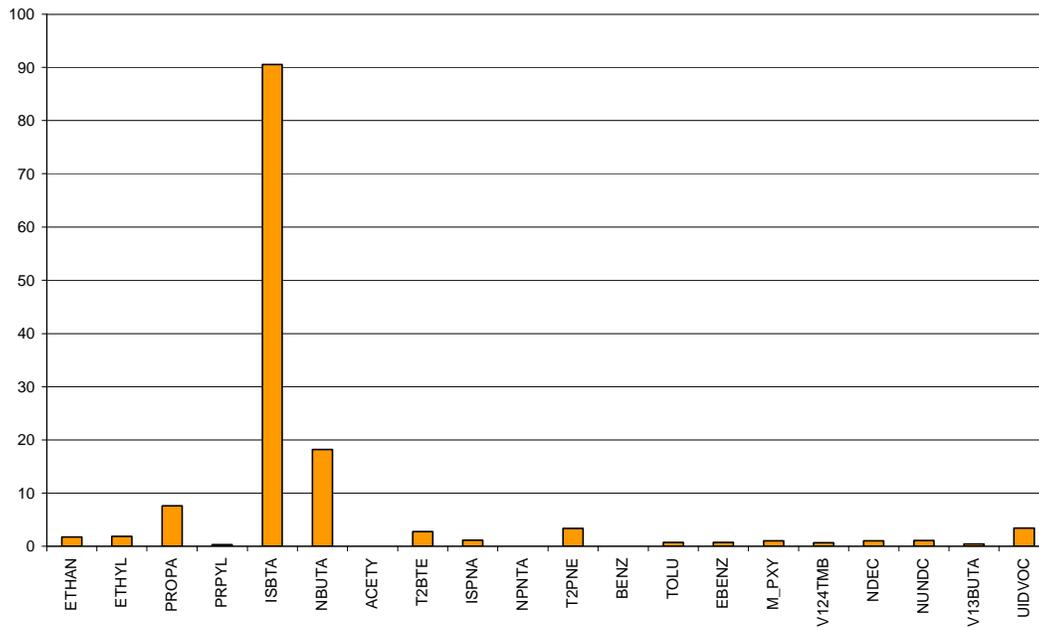


Figure 5-73. Percent of each species in UNMIX Factor 1.

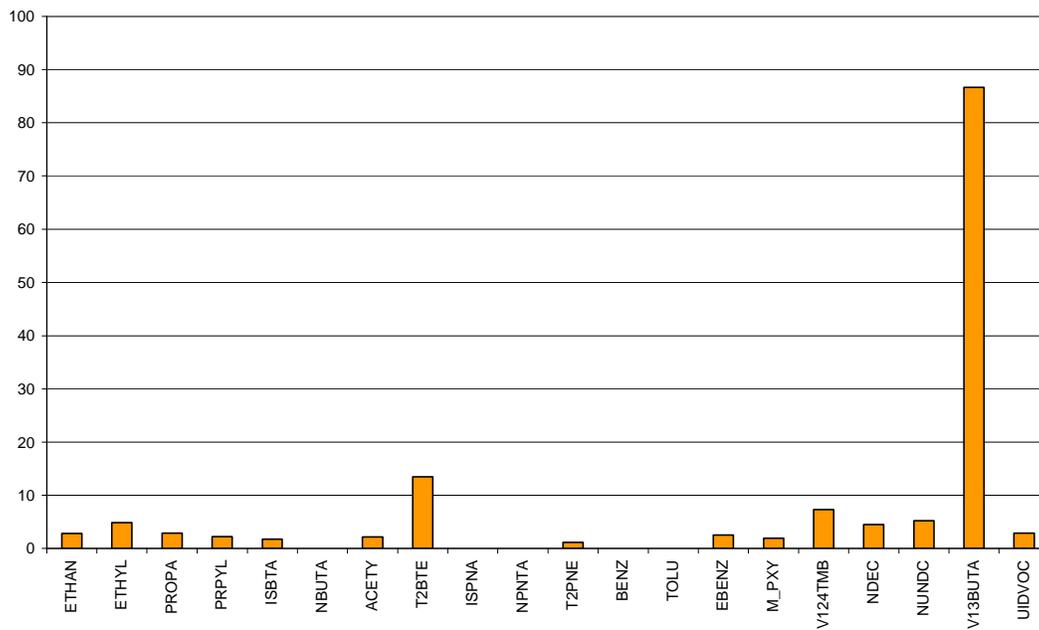


Figure 5-74. Percent of each species in UNMIX Factor 2.

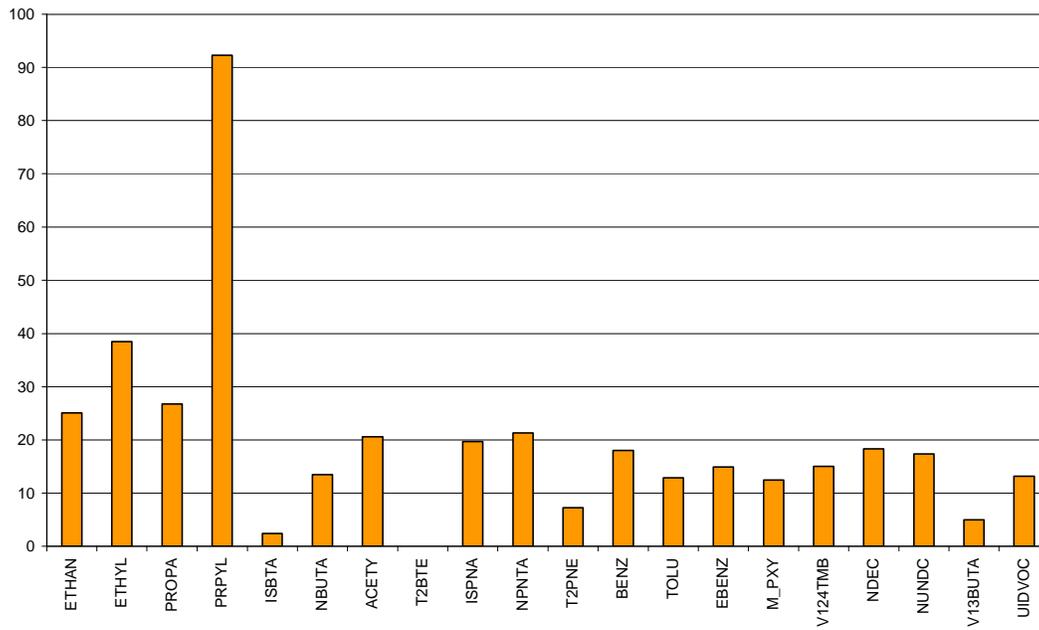


Figure 5-75. Percent of each species in UNMIX Factor 3.

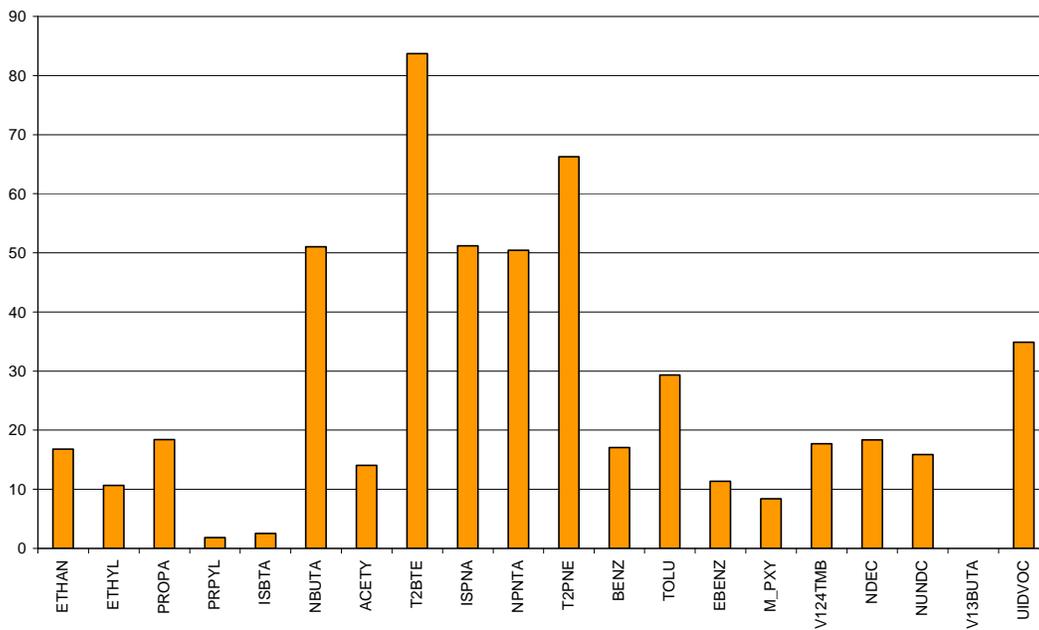


Figure 5-76. Percent of each species in UNMIX Factor 4.

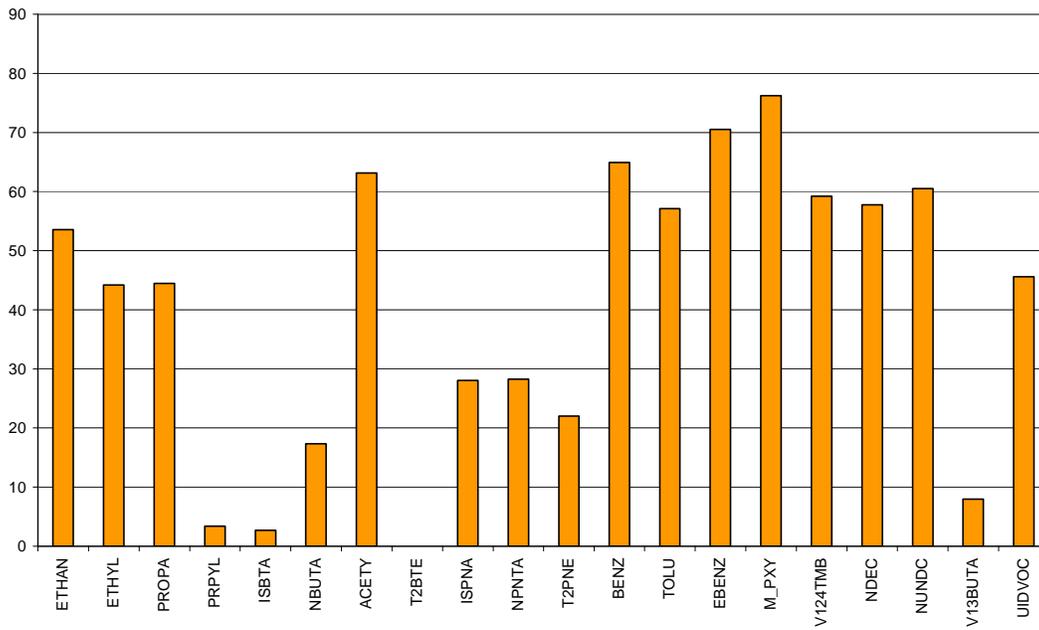


Figure 5-77. Percent of each species in UNMIX Factor 5.

6. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This report details the data assembly, data reduction (i.e., eliminating records), treatment of missing and below-detection data, preparation of uncertainties for all data values, and selection of the hydrocarbons to be modeled. The report also shows results from investigations of the data using cluster and factor analyses, which assist us in setting the number of factors used by PMF and in understanding potential differences in the results related to year, season, time of day, and wind direction. Results from PMF analysis are analyzed, with resultant factors identified as specific sources, their source regions characterized, and their temporal variations characterized. Results from preliminary UNMIX analysis are also discussed.

6.1 DATABASE PREPARATION

Factor analyses require complete samples (i.e., a concentration for each compound in every sample) and typically samples with missing species or species below the detection limit are excluded from analysis. One of the strengths of PMF compared to other source apportionment tools is that PMF can individually consider each data point. This feature allows the analyst to adjust the influence of each data point depending on the confidence in the measurement, and retain data that would otherwise be screened out. In preparing the database, the following decisions were made:

- The following samples were excluded from the data base: samples in which all data were missing; samples in which ethene, propene, or TNMOC/unidentified values were missing; samples collected during two periods when benzene, xylenes, and toluene were all missing; and samples flagged as suspect or invalid during data validation.
- Published methods for replacing, and assigning uncertainty to, missing and below-detection-limit data were employed. The uncertainty was adjusted so that these data were less important to the model solution (i.e., these data had less influence than measurements well above the detection limit).
- For most hydrocarbons, annual mean concentrations were assigned to missing data, following published methods. However, missing isoprene data were treated differently to account for changes in biogenic emissions by month and time of day; monthly mean values by time of day were assigned.
- In order to reduce collinearity problems, not all species were included in the source apportionment. Styrene data were not used because of questions regarding the capability of the auto-GCs to accurately measure this VOC.

The resulting, carefully constructed, database contains over 21,000 samples collected over four years; data were well-distributed by year, month, day of week, and hour. The accompanying uncertainty estimates are critical to PMF model performance.

6.2 FACTOR AND CLUSTER ANALYSIS RESULTS

Factor and cluster analyses using SYSTAT statistical software were performed as a preliminary investigation of the data set. We have found in other projects that SYSTAT's factor analysis provides us with a good idea of the number of factors which we may expect to be identifiable using PMF. We also used this "simple" factor analysis to investigate how the number and composition of the factors varied with time of day and wind direction. In summary,

- Nine factors were found using the entire data set, indicating that at least this many factors should be identifiable using PMF.
- Only small differences between morning and afternoon factor analyses were found, suggesting that fresh emissions occur all day (as observed by Brown and Main, 2002) and that atmospheric reactions should not overly interfere with the source apportionment.
- Significant differences were found in both the number and composition of factors by wind octant, illustrating the complex mixture of emissions that impact the Clinton Drive site. Using these findings, initial applications of PMF were made to test solutions varying from 9 to 18 factors.

6.3 PMF ANALYSIS

The large (21,000 records) and highly variable data set provided challenges to PMF. Outliers in the data, which occur for nearly all species, at all times of day and year, were especially challenging. However, after many sensitivity runs of the model, a 15-factor solution was selected for PMF application based on model performance parameters and the uniqueness of the factors (e.g., in composition, day of week, or diurnal variation). The reconstructed mass showed a very good ($r^2=0.91$) correlation with the measured (expected) mass, indicating the solution adequately represented the data. The identified factors, average weight percent, and temporal and wind direction variations are summarized in **Table 6-1**. There were a number of surprises and key findings:

- The likely motor vehicle factor contributed much less (4%) to the overall TNMOC than expected. We suspect that some of the motor vehicle emissions are included in other factors—PMF could not cleanly split this source from the industrial sources with similar species emissions using auto-GC data alone.
- A diesel component was identified, but the contribution to total VOC mass was small. It was encouraging to be able to identify a diesel component separately from other mobile source emissions. The small contribution to TNMOC was not surprising because the only PAMS target VOCs linked to diesel emissions are C10 and C11 alkanes (even higher carbon numbers are generally in the particle phase); these compounds comprise a very small portion of TNMOC. While not necessarily important to ozone formation, this finding is useful for other source apportionment investigations.
- The separation of two aromatic sources of little mass (ethyl- and propyl-benzene in Factor 2 and 1,2,3-trimethylbenzene in Factor 8) was unusual; we would have expected these low concentration compounds to be more strongly associated with other aromatic hydrocarbons.

- An industrial flare factor (tentatively), comprised of acetylene, ethane, ethane, and n-butane was identified. While there may be some mobile source influence in this factor, factor strength does not decrease on weekends, indicating it is mainly from stationary sources.
- Major source regions were identified to the east and south of the Clinton Drive site, the areas of densest industrial activity, but some factors were associated with other directions such as northwest and southwest indicating other sources impacting the site.

In addition to working with the concentration data, species in each source profile were scaled by MIR reactivity coefficients, to gain an understanding of what factors (i.e., sources) are potentially the most important for ozone formation. In terms of total reactivity, the pentenes factor was found to be the highest (18%), followed by the light olefins (16%), trimethylbenzenes/unidentified (13%), motor vehicle (10%), butenes (10%), butadiene (8%), and isoprene (8%). These results were consistent with earlier analyses of auto-GC data which showed that no single compound or compound class (such as olefins) dominated the total reactivity. Olefins, however, may have more influence than is apparent since it is likely that some portion of them are already reacted away before impacting the sampling site. This variation is difficult to characterize, though carbonyl compound sampling may give insight into this. Of these factors, only the motor vehicle, trimethylbenzenes/unidentified and isoprene had higher median weight percents on mornings of ozone episodes. Note that the unidentified mass contribution to reactivity is unknown.

Source strength on mornings of ozone episode days and non-episode days in the summer over all years was also investigated; sources with a higher weight percent on mornings of episode days may be closely linked to the high ozone. Six factors were significantly higher on ozone episode mornings: industrial flare (1), heavy aromatic hydrocarbons (2), motor vehicles (3), solvents (10), light paraffins (11), and industrial/mobile aromatic hydrocarbons (12). This analysis indicates that aromatic hydrocarbons may be more important to ozone formation than observed in the earlier analyses. While the high levels of the more reactive compounds (e.g., light olefins) appear to support a high “background” of ozone, high concentrations of the aromatic hydrocarbons may provide additional formation of ozone to increase levels above 125 ppb.

We also explored the application of UNMIX to the data set. However, UNMIX allowed for fewer species, allowed fewer samples (since substitutions and separate weighting schemes are not accommodated), and resulted in only five factors. These factors appeared to be combinations of the separate factors identified by PMF.

Table 6-1. Summary of the 15-factor solution for PMF using 1998-2001 auto-GC data collected at Clinton Drive.

Factor	Average % of TNMOC	Average % of total reactivity	Estimated Source Type	Key Species	Weekday-Weekend Variation?	When is daily peak?	Prominent Wind Direction
1	5%	5%	Industrial flares	Ethane, ethene, n-butane and acetylene	None	Small morning + evening	E, NW
2	4%	1%	Industrial aromatic hydrocarbons #1	Unidentified fraction, diethylbenzene and propylbenzene	None	Small morning + evening	S, SW
3	4%	10%	Motor vehicle	Benzene, toluene, xylenes, acetylene and 2,2,4-trimethylpentane	Lower on weekend	Morning + evening	SW, W, NW, SE
4	5%	16%	Industrial light olefins	Ethene and propene	None	Morning only	E, S
5	16%	1%	Evaporative emissions/background	Butanes	None	Afternoon/evening	E, S
6	3%	2%	Solvent use	C6-C9 paraffins	None	Afternoon/evening	SSE
7	1%	18%	Industrial pentene source	Pentenes, some pentanes	None	Small morning + evening	S, ESE
8	1%	13%	Industrial aromatic hydrocarbons #2	Unidentified fraction and trimethylbenzenes	None	Morning only	N, E
9	2%	8%	Butadiene sources	1,3-Butadiene	None	Morning only	S
10	10%	2%	Evaporative emissions/solvents	C5-C7 paraffins	None	Morning only	E, SE, S
11	24%	1%	Accumulated emissions and natural gas	Ethane and propane	None	Morning only	E, N
12	12%	2%	Heavy aromatic sources	Ethyltoluene, trimethylbenzenes and xylenes	Lower on weekend	Morning + evening	E, N
13	2%	3%	Diesel	C10-C11 alkanes and xylenes	Lower on weekend	Morning only	W, N (likely trucks); S, E (likely trains, shipping)
14	2%	8%	Biogenic with outliers from industry	Isoprene	None	Noon	W, E, S
15	2%	10%	Industrial butene source	Butenes	None	Afternoon/evening	S

6.4 FUTURE WORK

As an exploratory exercise, this application of PMF provided useful results. As the project progressed, we identified many additional analyses that would be potentially useful to explore:

Additional applications with the Clinton Drive data set

- Work with emission inventory staff and emission inventory maps to try to more precisely define the factors.
- Prepare monthly mean temporal distribution of factors (e.g., stacked bar with each factor strength by hour in September) for comparison to 1993 CMB work (Fujita et al., 1995).
- Evaluate the emission inventory by comparing factors to the emission inventory on a species-specific basis.
- Perform trajectory analyses of selected factors for days with highest impact. Compare to trajectory analyses of days with high ozone.
- Use factors (source profiles) developed from Clinton Drive using the 1998-2001 data and apply CMB to the 2002 Clinton Drive data.
- Apply PMF to the 2002 Clinton Drive data.
- Convert the concentration data set to a reactivity-weighted data set and rerun PMF. Currently, we weighted the resulting average composition of the factors after performing source apportionment on the concentration data. Starting with reactivity-weighted data may lead to different source identification.
- Scale source strengths by an average/median reactivity (determined by source profiles) and further investigate total reactivity temporally (i.e., by time of day, ozone episodes).
- Complete CPF using a reactivity-scaled data set as listed above.

Additional PMF sensitivity runs

- Utilize additional VOC and criteria species such as CO, NO_x and O₃, to investigate the changes in source profiles and strengths.
- Compare runs with and without the missing isoprene data substitution method used in this report.
- Run PMF with data separately by year to investigate whether source profiles change. Annual differences may indicate changes in emission types and frequency of emissions.
- Run PMF with only data from the summer, the period of high ozone concentrations and exceedances to see if factor number, composition, or strength is different.
- Exclude samples of extremely high total mass and rerun the model. While these data were well-modeled, other samples in the highest 10th percentile of the mass were not, which may be due to the influence of the few extreme outlying concentrations.

Applications to other HSC sites

- Prepare data sets for the Deer Park, HRM 3, Channelview, Baytown, Aldine, and Bayland Park sites and apply PMF.
- Investigate the spatial, diurnal, and wind direction dependence of the factors based on results from multiple sites to further validate and confirm identified sources and their impacts.
- Run PMF on 24-hr toxics data from the Houston area and compare the results with the auto-GC findings.

Synthesis of results

- Compare the results from each site for consistency: were the compositions of factors such as motor vehicle emissions consistent between sites? Do similar factors point to the same wind sectors?
- Compare PMF results to 1993 COAST CMB results.
- Compare current work to PMF results from PM_{2.5} data and 24-hr toxics data.

Recommendations for monitoring and analyses

- Install an aethalometer at Clinton Drive; once several months of hourly data are collected, perform PMF using auto-GC and aethalometer BC data. This may better resolve the diesel component, which is of interest for toxics monitoring.
- Sample/analyze for carbonyl compounds for an extended period (e.g., an entire summer) and include in PMF runs. These compounds are reactive and abundant, and their inclusion may better resolve and apportion the unidentified fraction, as well as give further insight into sources affecting ozone formation.

7. REFERENCES

- Ashbaugh L.L., Malm W.C., and Sader W.Z. (1985) A residence time probability analysis of sulfur concentrations at Grand Canyon National Park. *Atmospheric Environment* **19** (8), pp. 1263-1270.
- Atkinson R. (1989) Kinetics and mechanisms of the gas-phase reactions of the hydroxyl radical with organic compounds, Monograph 1. *Journal of Physical and Chemical Reference Data*.
- Atkinson R. (1994) Gas-phase tropospheric chemistry of organic compounds, Monograph 2. *Journal of Physical and Chemical Reference Data*.
- Brown S.G. and Main H.H. (2002) Acquisition, review, and analysis of auto-GC VOC data in Houston area, 1998-2001. STI-900670-2224-FR, July.
- Brown S.G., Roberts P.T., Buhr M.P., and Wheeler N.J.M. (2002) Characterization of 2001 event-triggered VOC and carbonyl compound samples. STI-900680-2188-FR, August.
- Brown S.G. and Hafner H.R. (2003) Prescribed data analyses for Phoenix, Arizona, hydrocarbon data collected in 2001. Final report, STI-902263-2285-FR, January.
- Carter W.P.L. (1994) Development of ozone reactivity scales for volatile organic compounds. *Journal of the Air & Waste Management Association* **44**, pp. 881-899.
- Carter W.P.L. (2001) The SAPRC-99 chemical mechanism and updated reactivity scales. Final report, available at <<http://pah.cert.ucr.edu/~carter/>>.
- Fujita E.M., Lu Z., Sagebiel J.C., Robinson N.F., and Watson J.G. (1995) VOC source apportionment for the Coastal Oxidant Assessment for southeast Texas. Report.
- Hopke P.K. (2003) A Guide to Positive Matrix Factorization. early 2003.
- Kim E., Hopke P.K., Larson T.V., and Covert D.S. (2002) Analysis of ambient particle size distributions using UNMIX and positive matrix factorization. *Environmental Science & Technology* (submitted).
- Lee E., Chan C.K., and Paatero P. (1999) Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong. *Atmospheric Environment* **33**, pp. 3201-3212.
- Lee J.H., Yoshida Y., Turpin B.J., Hopke P.K., Poirot R.L., Liou P.J., and Oxley J.C. (2002) Identification of sources contributing to mid-Atlantic regional aerosol. *Journal of Air and Waste Management Association* **52**, pp. 1186-1205.
- Main H.H., Alcorn S.H., and Roberts P.T. (1999a) Characteristics of volatile organic compounds in the Mid-Atlantic region. Report, STI-998482-1869-FR2, November.

- Main H.H., Hurwitt S.B., and Roberts P.T. (1999b) Spatial and temporal characteristics of California PAMS and long-term trend site VOC data (1990-1997). Report, STI-998241-1883-FR, May.
- Main H.H. and Roberts P.T. (2000) PAMS Data Analysis Workbook: Illustrating the use of PAMS data to support ozone control programs. STI-900243-1987-FWB, September.
- Main H.H. (2001a) PAMS analysis for Southern California: characteristics of VOC data 1994-1997. Presented at the *SCOS97-NARSTO Data Analysis Conference, Los Angeles, CA, February 13-15* (STI-2054).
- Main H.H. (2001b) Toxics data analysis, Phoenix 1994-2000. Presentation made to *Arizona Department of Environmental Quality, Phoenix, AZ, November 28, 2001* (STI-901484-2178).
- Main H.H. and O'Brien T. (2001) PAMS data validation for the Northeast and Mid-Atlantic states, 1994-1999. Report, STI-900860-2040-FR, June.
- Main H.H., O'Brien T., Hardy C., Wharton S., and Sullivan D. (2001) Characterization of Auto-GC data in Houston. Report, STI-900610-2112-EO, August.
- Main H.H. and Brown S.G. (2002a) Preliminary analyses of Houston auto-GC 1998-2001 data: episode/non-episode differences. Report, STI-900670-2165-IR, March.
- Main H.H. and Brown S.G. (2002b) Validation and analysis of Atlanta, Georgia 1999 and 2000 PAMS data. Technical memorandum, STI-999434-2177-TM, March.
- Main H.H. and Brown S.G. (2002c) Acquisition, review and analysis of auto-GC data in the Houston area. Presented to *Texas Natural Conservation Commission, Austin, TX, August 28* (STI-900670-2260).
- Main H.H. and Brown S.G. (2002d) PAMS data validation for the northeast states 2000-2001. Report, STI-902071-2205-FR, September.
- Paatero P. and Tapper U. (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, pp. 111-126.
- Paatero P. (1997) Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems* **18**, pp. 183-194.
- Paatero P. (2000) User's guide for positive matrix factorization programs PMF2 and PMF3, part 1: tutorial. February.
- Poirot R.L., Wishinski P.R., Hopke P.K., and Polissar A.V. (2001) Comparative application of multiple receptor methods to identify aerosol sources in northern Vermont. *Environmental Science & Technology* **35** (23), pp. 4622-4636.

- Polissar A.V., Hopke P.K., Paatero P., Malm W.C., and Sisler J.F. (1998) Atmospheric aerosol over Alaska 2. Elemental composition and sources. *Journal of Geophysical Research* **103** (15), pp. 19045-19057.
- Polissar A.V., Hopke P.K., and Poirot R.L. (2001) Atmospheric aerosol over Vermont: chemical composition and sources. *Environmental Science & Technology* **35** (23), pp. 4604-4621.
- Ramadan Z., Song X.-H., and Hopke P.K. (2000) Identification of sources of Phoenix aerosol by positive matrix factorization. *Journal of Air and Waste Management Association* **50**, pp. 1308-1320.
- Song X.-H., Polissar A.V., and Hopke P.K. (2001) Sources of fine particle composition in the northeastern U.S. *Atmospheric Environment* **35**, pp. 5277-5286.
- Yakovleva E., Hopke P.K., and Wallace L. (1999) Receptor modeling assessment of particle total exposure assessment methodology data. *Environmental Science & Technology* (published) **33** (20), pp. 3645-3652, American Chemical Society (10.1021/es981122i). Available on the Internet at http://pubs3.acs.org/acs/journals/doi/lookup?in_doi=10.1021/es981122i; last accessed on 5/1/03.

APPENDIX A

PLOTS OF RESIDUALS FOR EACH SPECIES FROM THE 15-FACTOR PMF SOLUTION

This page is intentionally blank.

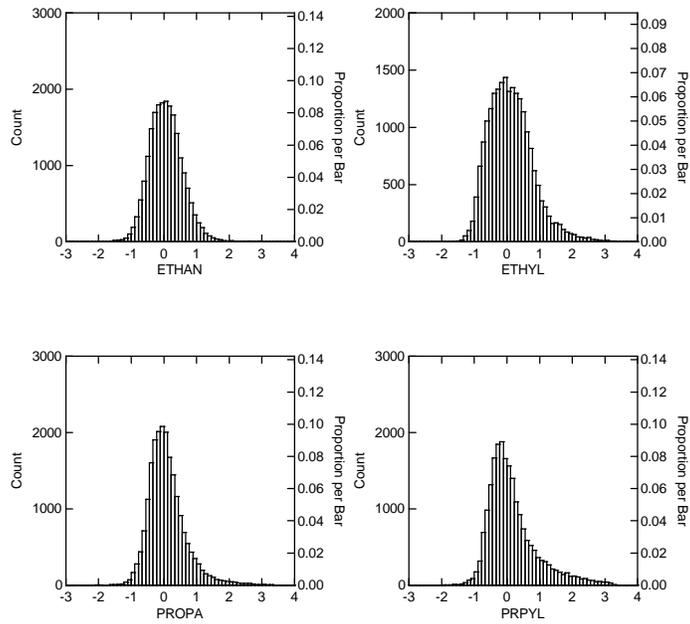


Figure A-1. Scaled residuals of ethane (ethan), ethene (ethyl), propane (propa) and propene (prpyl).

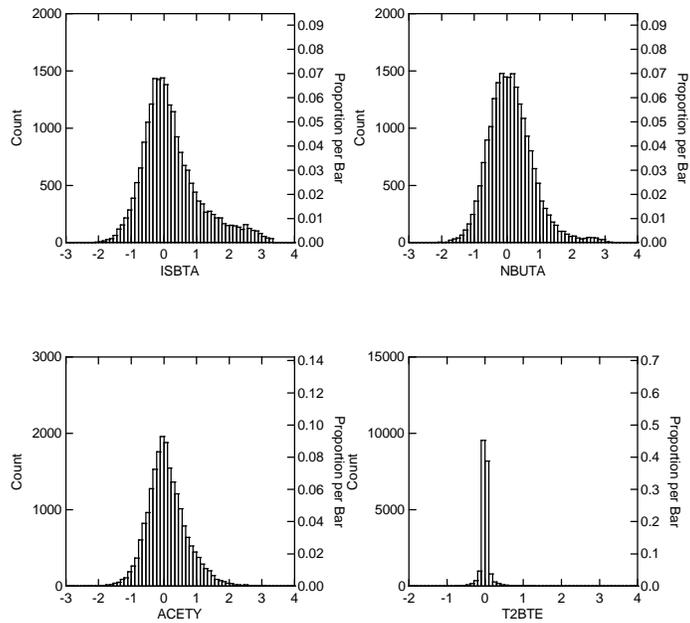


Figure A-2. Scaled residuals of isobutane (isbta), n-butane (nbuta), acetylene (acety) and t-2-butene (t2bte).

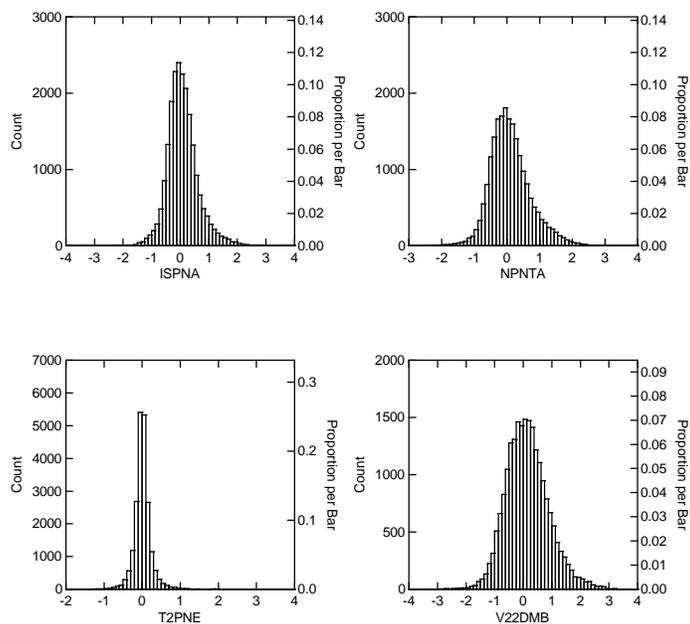


Figure A-3. Scaled residuals of isopentane (ispna), n-pentane (npnta), t-2-pentene (t2pne) and 2,2-dimethylbutane (v22dmb).

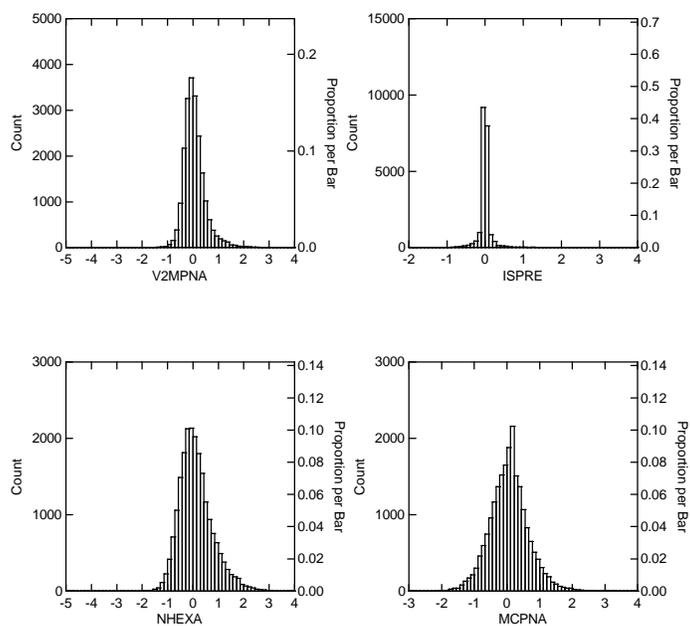


Figure A-4. Scaled residuals of 2-methylpentane (v2mpna), isoprene (ispre), n-hexane (nhexa) and methylcyclopentane (mcpna).

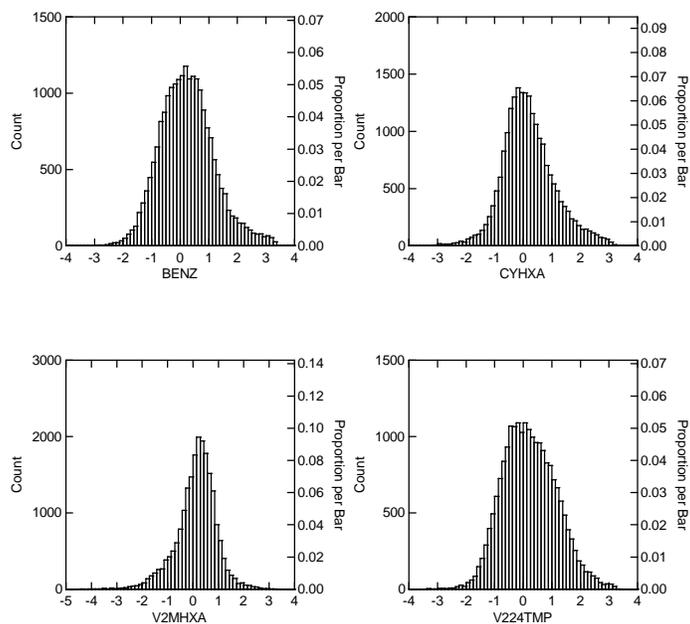


Figure A-5. Scaled residuals of benzene (benz), cyclohexane (cyhxa), 2-methylhexane (v2mhxa) and 2,2,4-trimethylpentane (v224tmp).

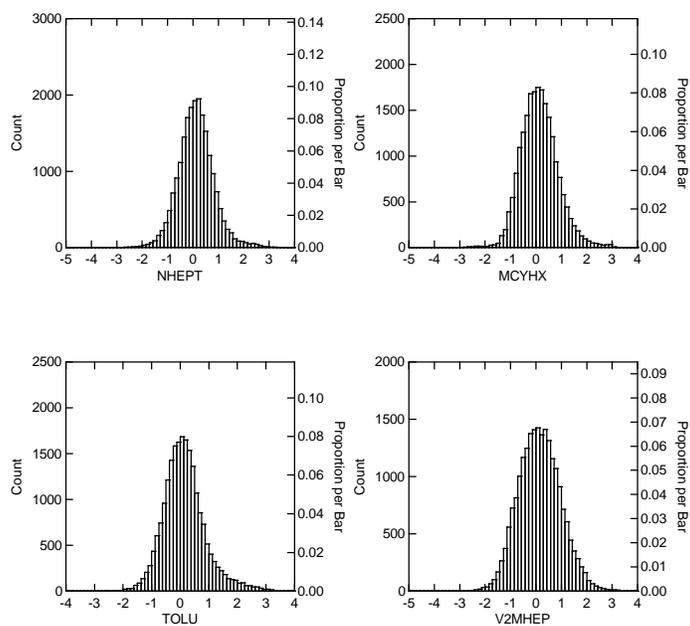


Figure A-6. Scaled residuals of n-heptane (nhept), methylcyclohexane (mcyhx), toluene (tolu) and 2-methylheptane (v2mhhep).

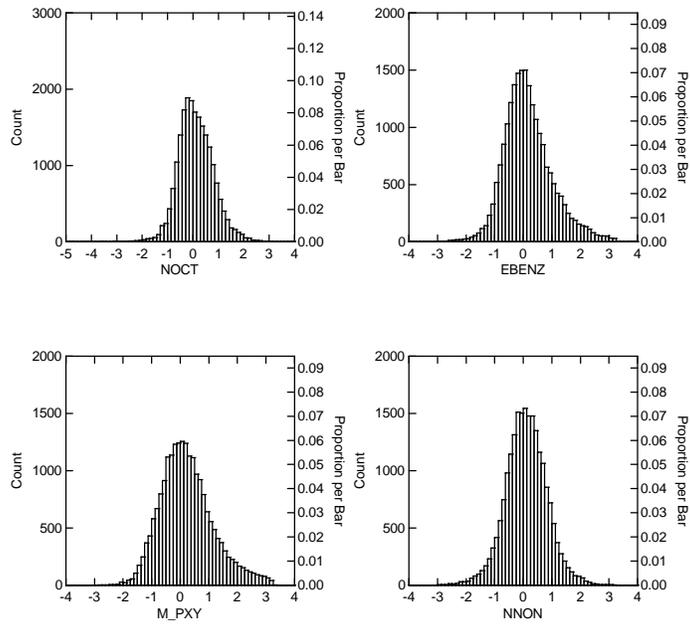


Figure A-7. Scaled residuals of n-octane (noct), ethylbenzene (ebenz), m/p-xylenes (m_pxy) and n-nonane (nnon).

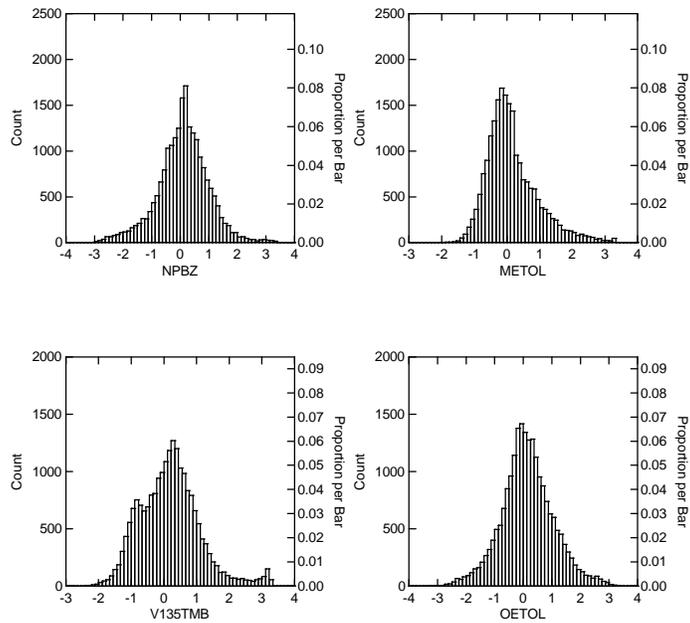


Figure A-8. Scaled residuals of n-propylbenzene (npbz), m-ethyltoluene (metol), 1,3,5-trimethylbenzene (v135tmb) and o-ethyltoluene (oetol).

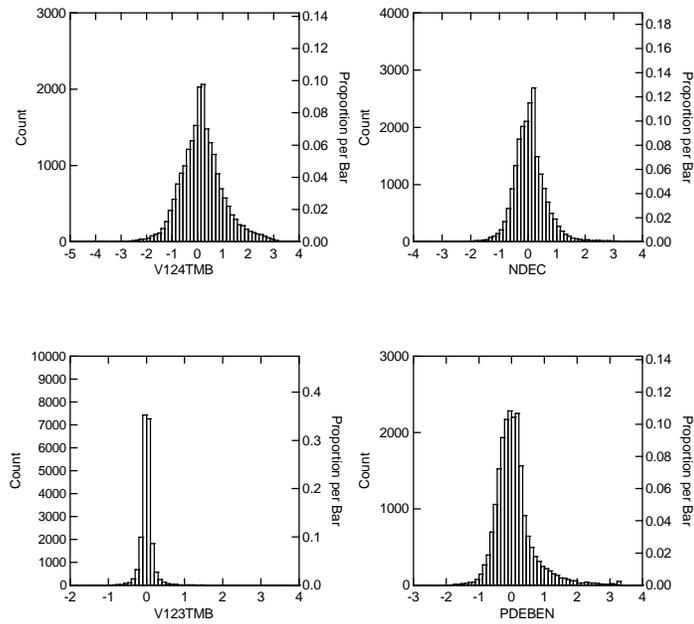


Figure A-9. Scaled residuals of 1,2,4-trimethylbenzene (v124tmb), n-decane (ndec), 1,2,3-trimethylbenzene (v123tmb) and p-diethylbenzene (pdeben).

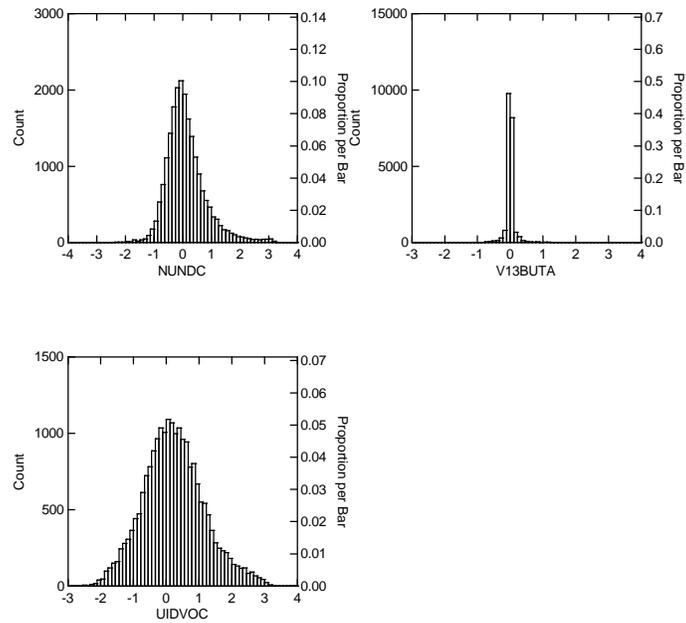


Figure A-10. Scaled residuals of n-undecane (nundc), 1,3-butadiene (v13buta), and unidentified (uidvoc).

APPENDIX B

EMISSION INVENTORY MAPS OF STATIONARY SOURCES IN THE HOUSTON AREA

This page is intentionally blank.

Emission Sections

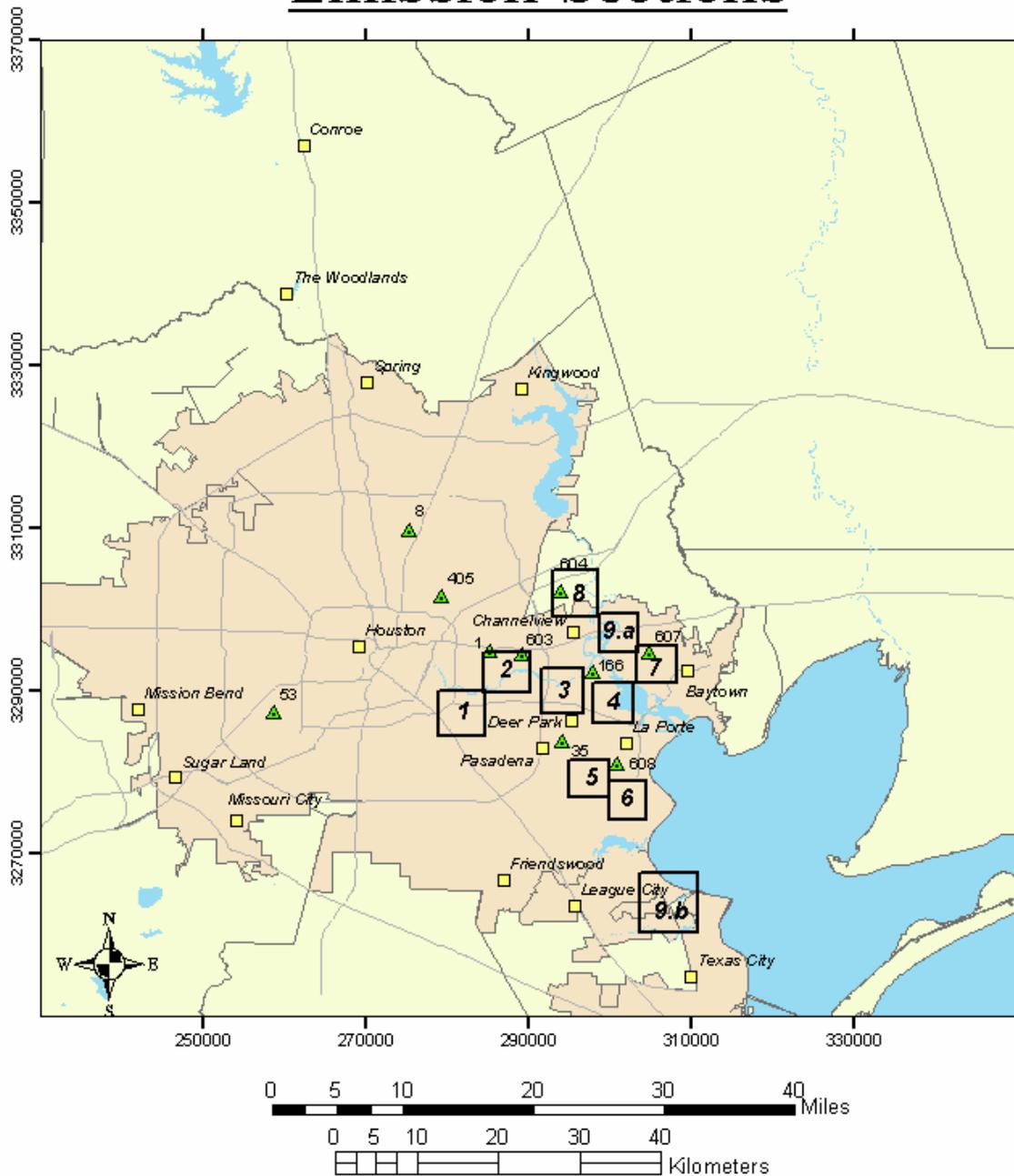


Figure B-1. Map of designated emission sections in the Houston area (1-8, plus 9a and 9b).

Ethene Emission Density

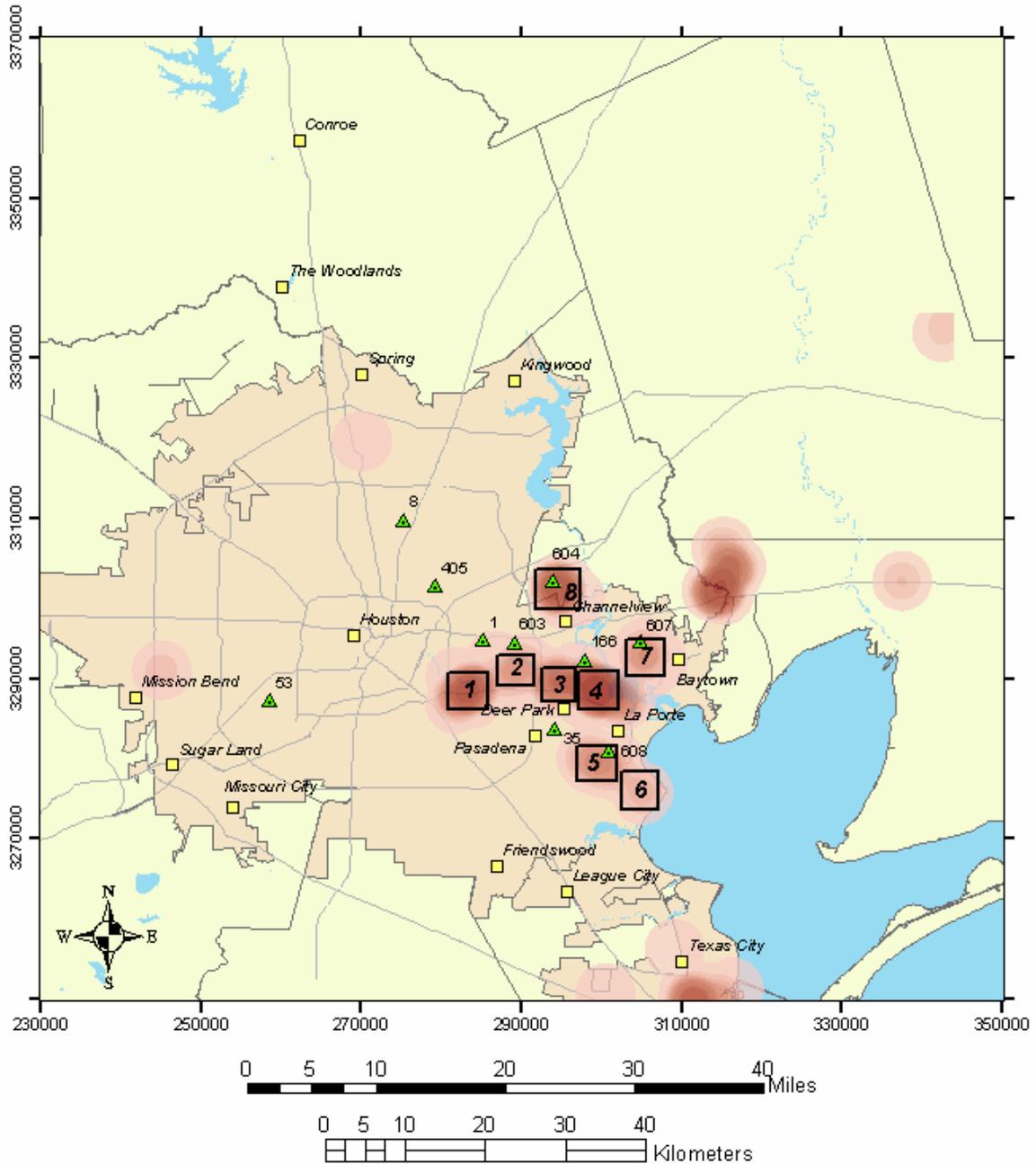


Figure B-2. Emission inventory map of stationary sources of ethene in the Houston area.

Propene Emission Density

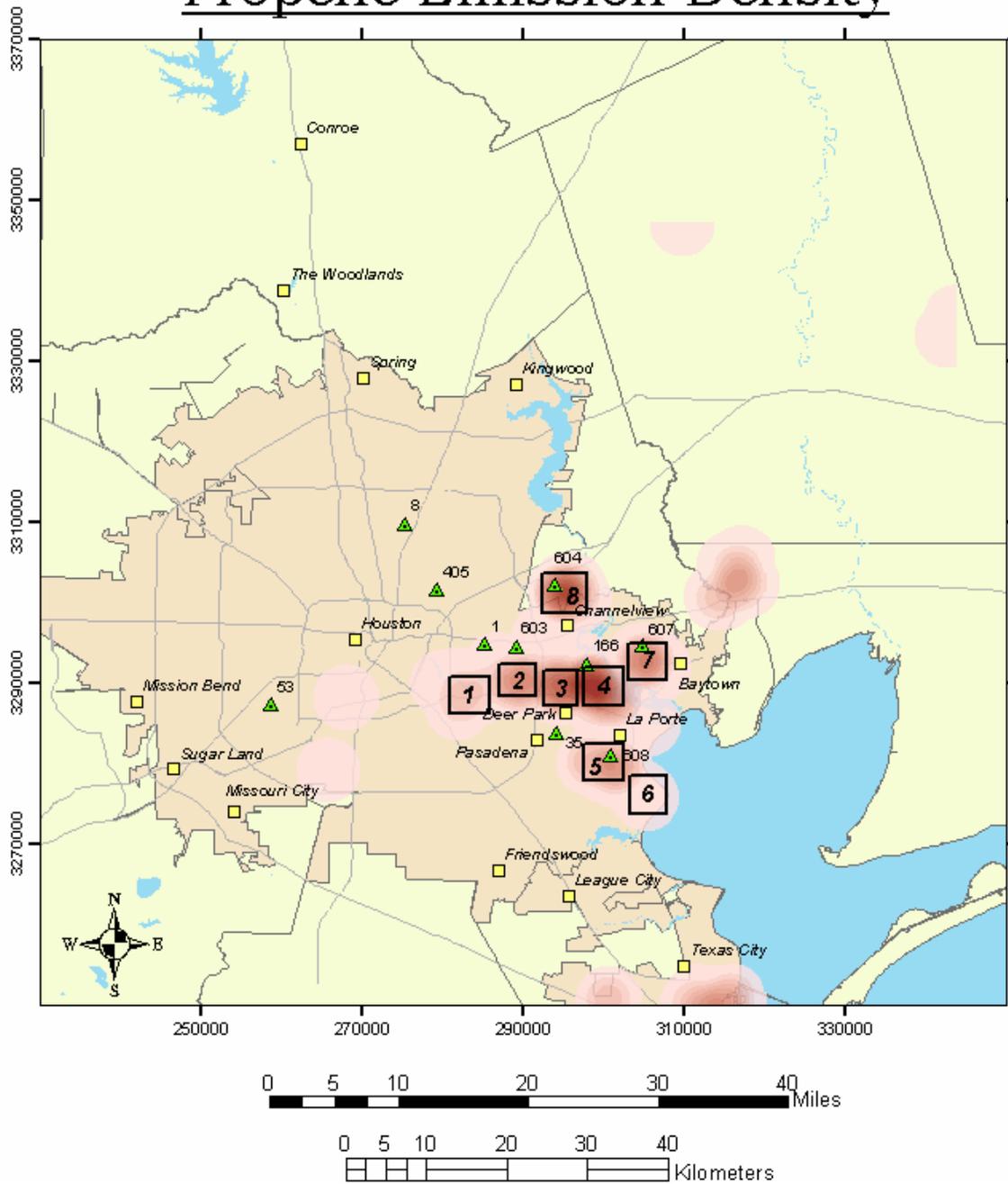


Figure B-3. Emission inventory map of stationary sources of propene in the Houston area.

Butenes Emissions Density

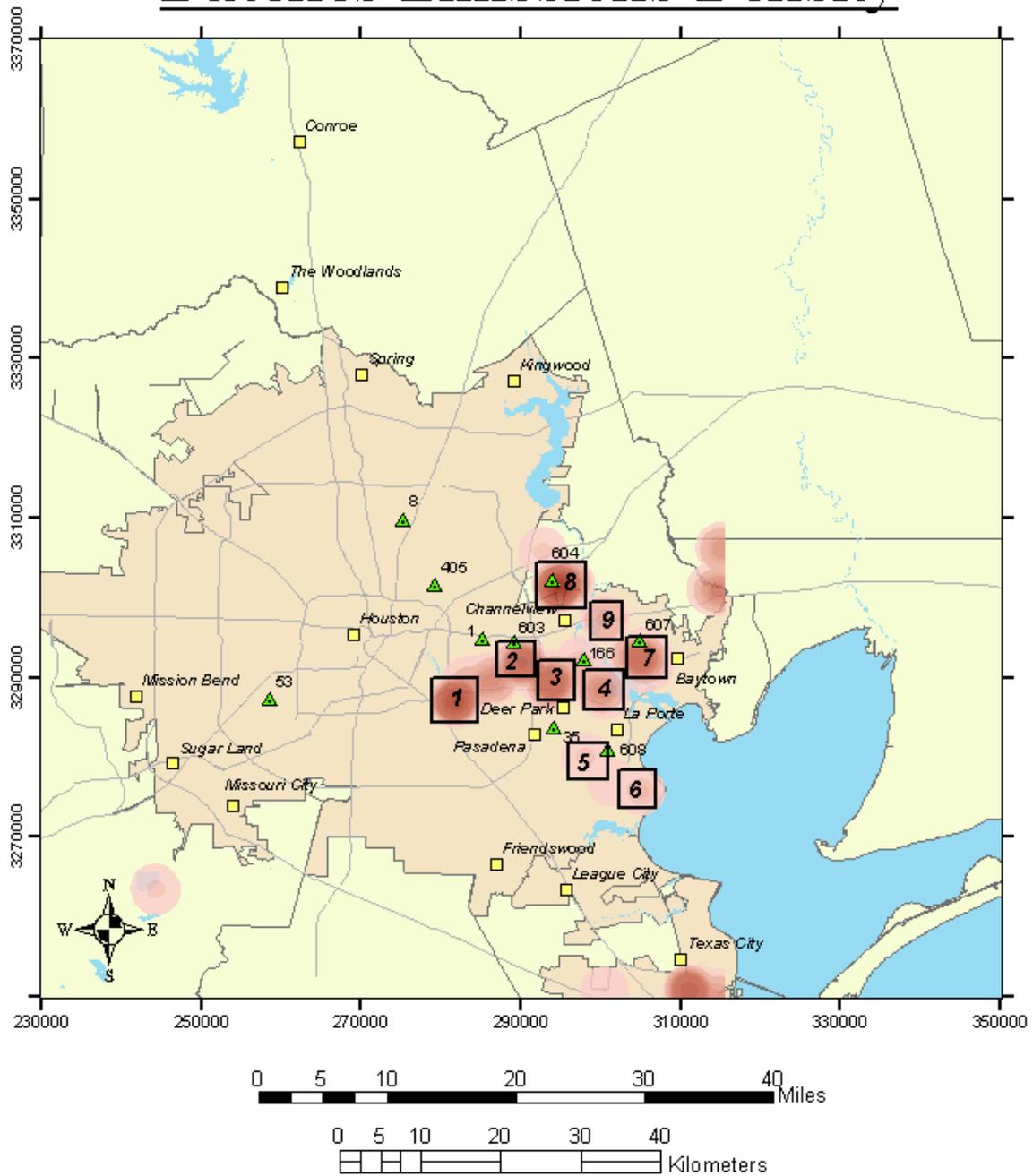


Figure B-4. Emission inventory map of stationary sources of butenes in the Houston area.

Pentenes Emission Density

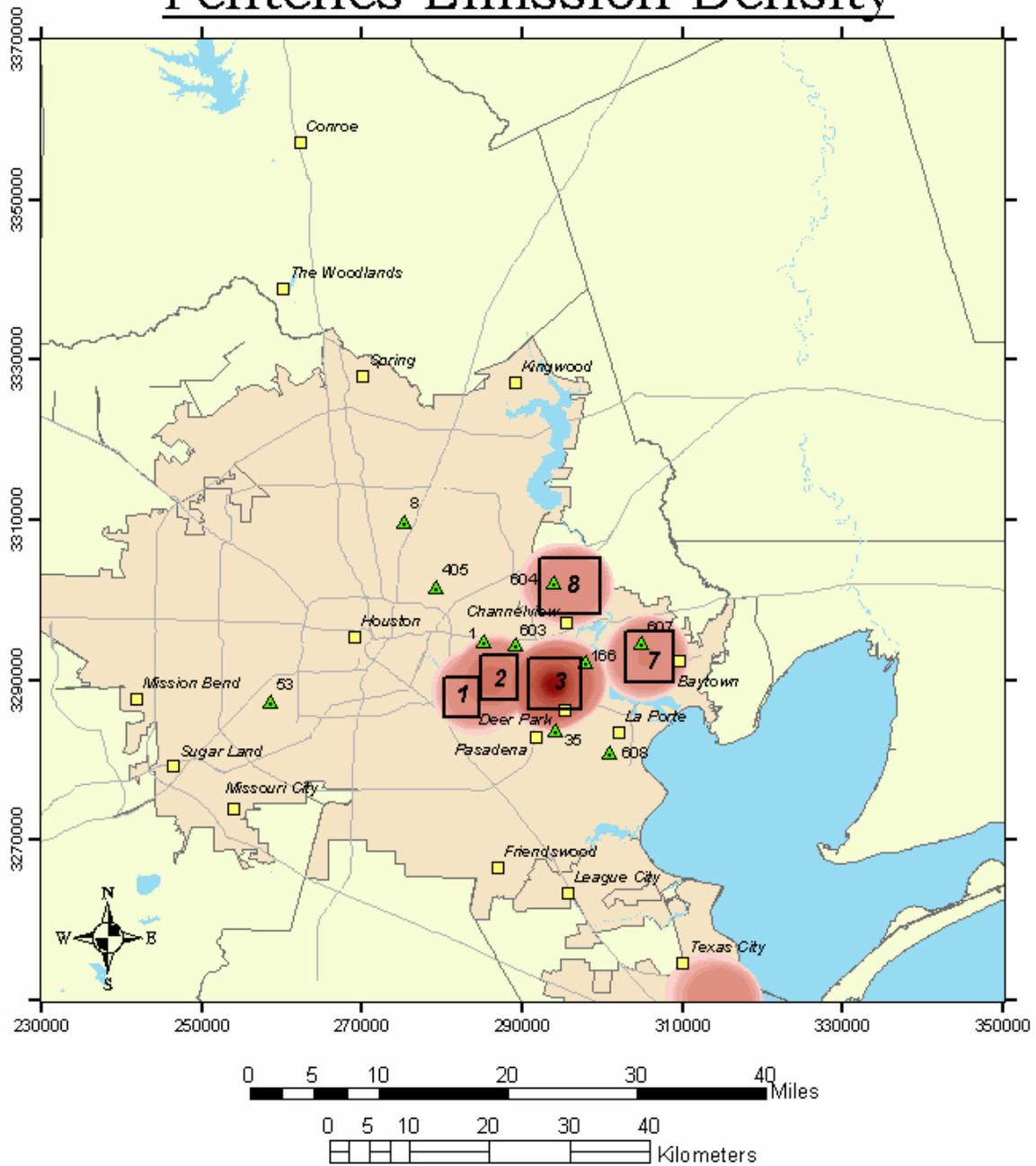


Figure B-6. Emission inventory map of stationary sources of pentenes in the Houston area.

Toluene Emission Density

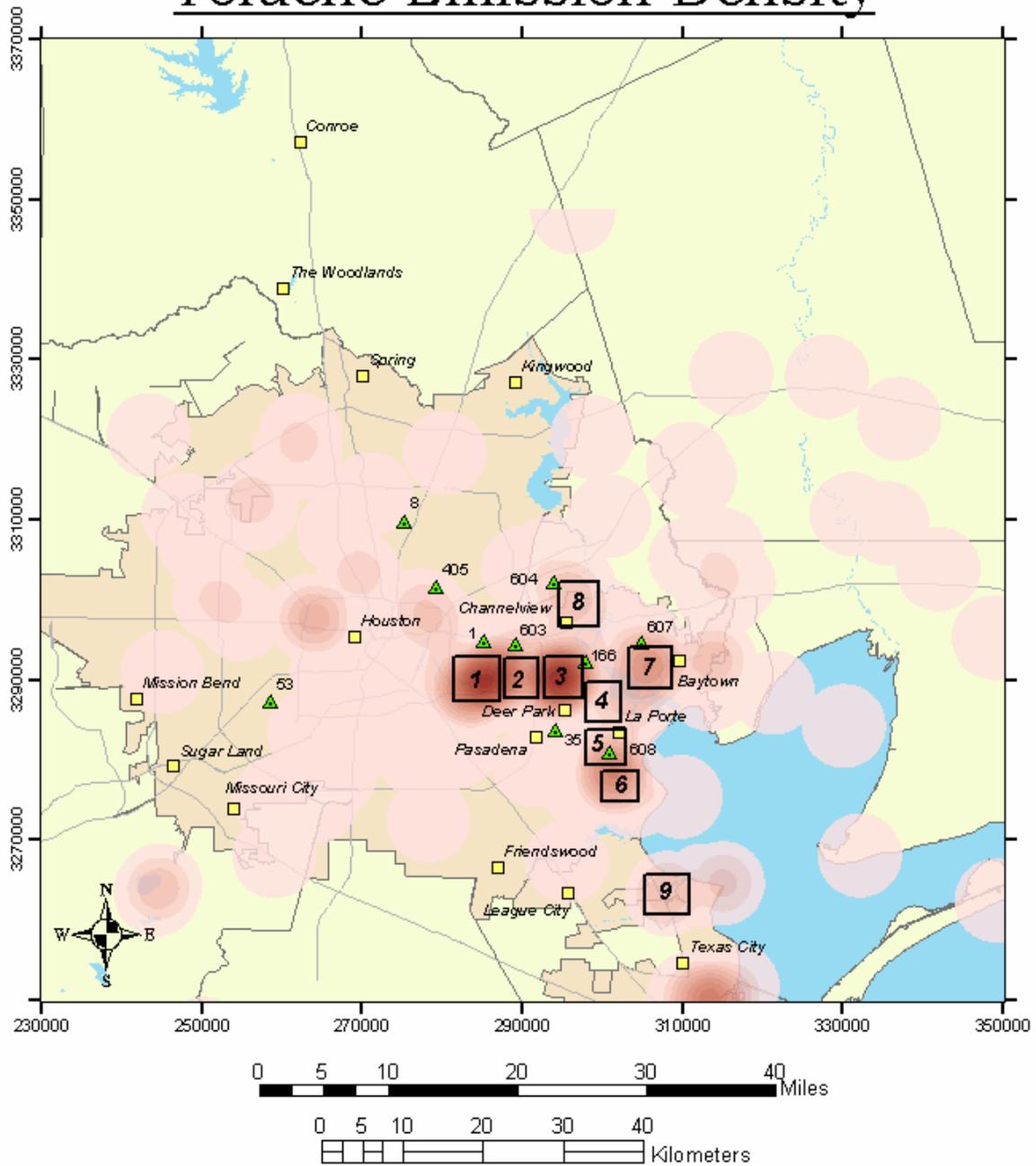


Figure B-7. Emission inventory map of stationary sources of toluene in the Houston area.

Xylenes Emission Density

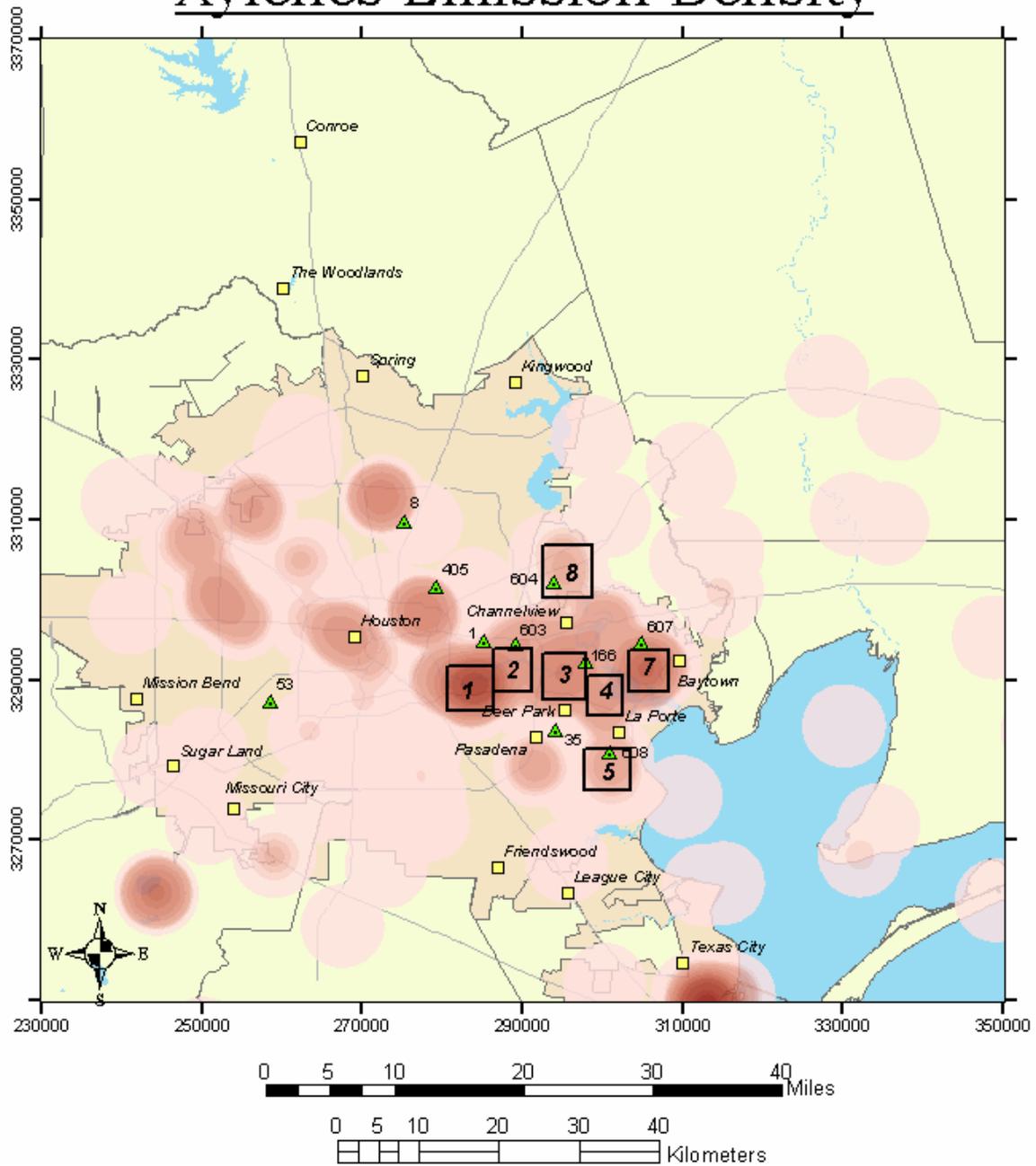


Figure B-8. Emission inventory map of stationary sources of xylenes in the Houston area.

Ethyltoluene Emission Density

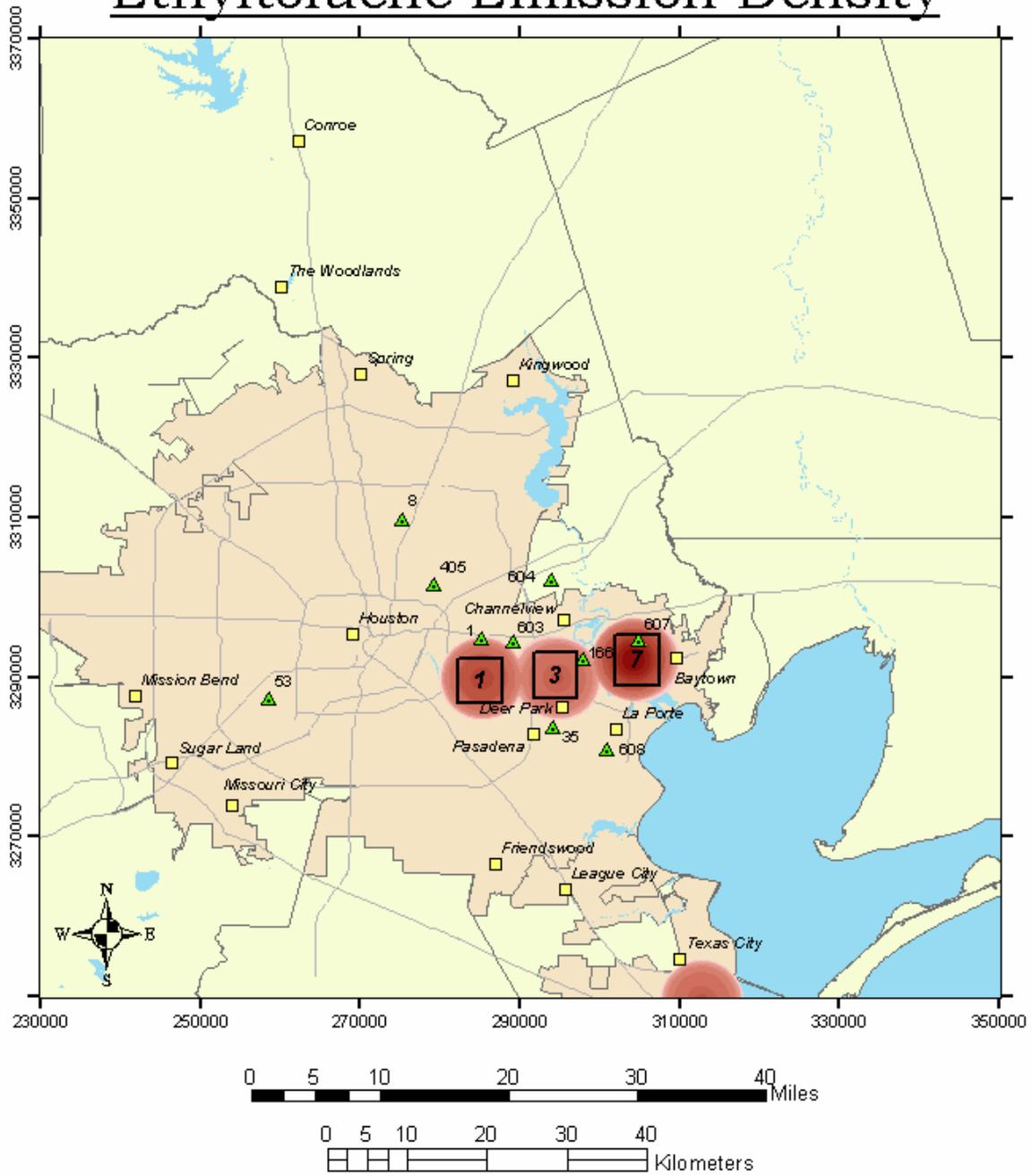


Figure B-9. Emission inventory map of stationary sources of ethyltoluene in the Houston area.

Trimethylbenzene Emission Density

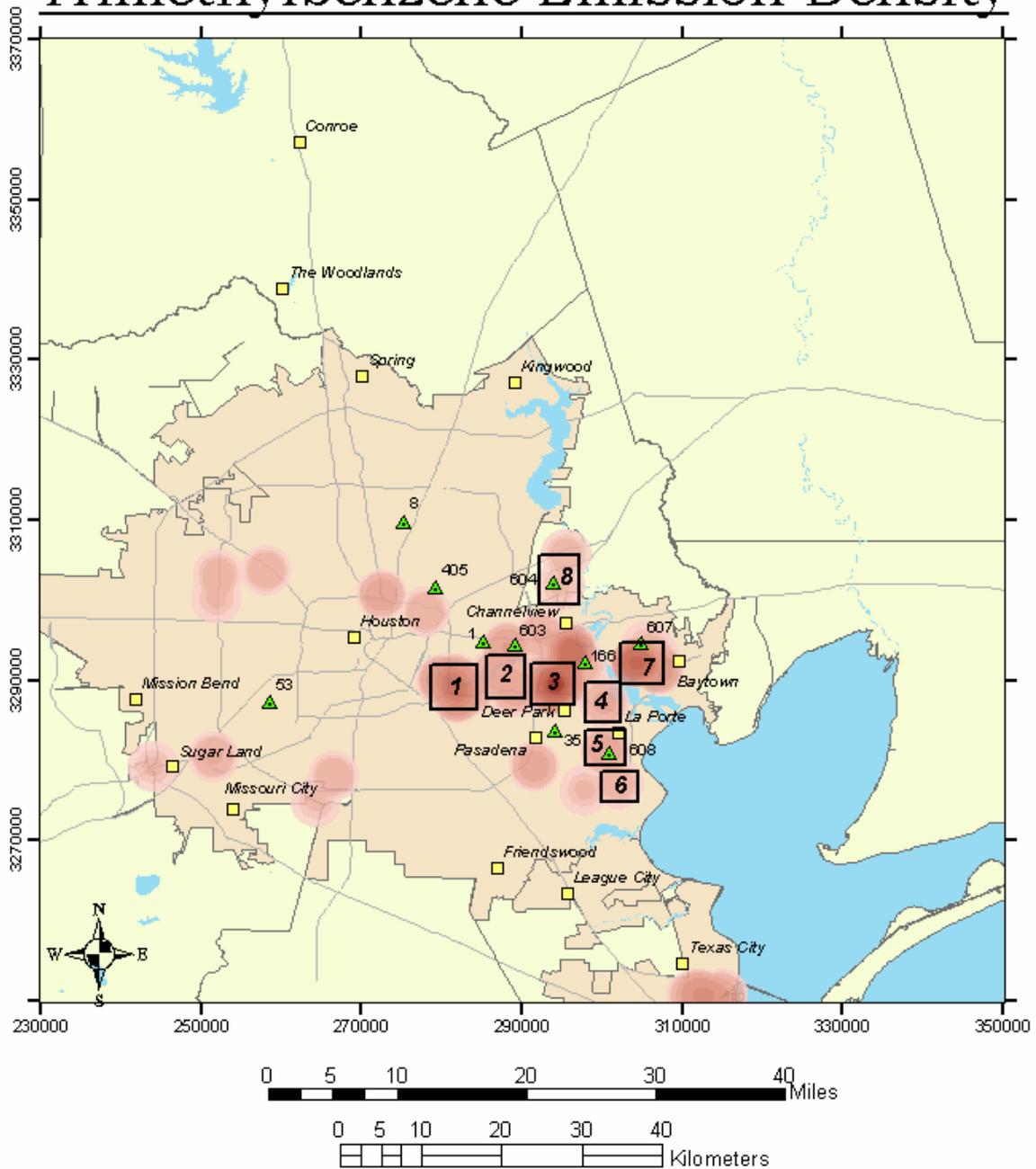


Figure B-10. Emission inventory map of stationary sources of trimethylbenzenes in the Houston area.