A WORKBOOK FOR
EXPLORATORY ANALYSIS OF PAMS DATA

June, 1995

Bill Cox

Emissions, Monitoring and Analysis Division
Office of Air Quality Planning and Standards
U.S Environmental Protection Agency

Research Triangle Park, NC.

Contents

[Note: The figures referenced in this document were not available electronically; If the reader wishes to receive a hard copy of this document including allfigures, contact Mark Schmidt @ 919-541-2416]

INTRODUCTION

State and local air pollution control agencies have begun implementation of an ambitious new program (Federal Register, 1993) to monitor ozone precursors in ozone nonattainment areas designated as serious, severe or extreme. In addition, the new network of Photochemical Assessment Monitoring Stations (PAMS) will supplement existing monitoring networks with new data on air toxics and meteorological measurements needed to interpret pollutant transport and accumulation. Data from the PAMS network will provide air quality planners with vital new information needed to understand and control ozone precursors and toxic organics more effectively.

PAMS data will ultimately be used to meet a variety of specific data objectives such as corroboration of emission inventories, refinement of model inputs, and empirical evaluation of trends and ozone precursor relationships. Preliminary data analysis plans have been proposed (Stoeckenius T. E., et. al, 1994) for beginning such analysis, with emphasis on first steps to explore and evaluate the consistency of VOC species.

The purpose of this workbook is to describe exploratory data methods useful in preliminary investigation of data collected from PAMS. In essence, this workbook illustrates a condensed sampling of "first look" analyses which will become part of a

much broader set of analytical approaches to achieve stated PAMS objectives. Since many of the specific analytic methods are presently evolving, supplementary guidance will be provided as these data analytic methods mature. While the methods in this workbook are not new (Hoaglin, et.al, 1983 and 1985), they have been tested (Stoeckenius T. E., et. al., 1995) and should prove useful in evaluating data consistency and potential data "outliers".

The analysis and graphics were produced using SAS (SAS Institute, 1993) or S-PLUS (Mathsoft, 1993) running on a 486-based PC under Microsoft Windows 3.1. SAS was used because it is supported by EPA while S-PLUS offers a less-expensive alternative with a wide range of robust statistical methods useful in exploratory analysis.

Since the emphasis is on exploratory analysis, the graphics shown in the workbook are not "presentation" quality. Most of the graphs are typical of those that might be produced through interactive analysis at a terminal followed by production of a working "hardcopy" for subsequent review by other analysts.

The reader is assumed to have some knowledge of the PAMS program and the type of data produced. The data used in this analysis were collected at a "type 2" PAMS site in Baltimore Maryland during the summer of 1993. The raw hourly data were retrieved from AIRS and manipulated for input into SAS and S-PLUS using standard data reduction utilities.

Because of the size of the collected data base (e.g. over 60 organic species) only selected species or species combinations were included in these examples. For convenience, a scheme proposed for analysis of data measured in Atlanta (Cohen, 1992) was used: acetylene, ethylene, olefins--(the sum of butenes and pentenes), isoprene, toluene, xylene--(the sum of three related species), benzene and total non-methane organic compounds. In addition, hourly concentrations of ozone, NO, NOx, NO2 and CO were included along with available hourly meteorological parameters.

The workbook begins with examples of graphical methods for summarizing PAMS data completeness. Next, simple procedures for illustrating diurnal patterns are pursued, with the view that typical diurnal patterns define a frame of reference for detecting some types of data anomalies. This is followed by examples of methods for comparing organic concentrations among data categories including weekend vs weekday differences. Methods for factoring out meteorology are illustrated, including use of species ratios and simple statistical models that relate species concentrations to selected meteorological parameters. Next, multivariate cluster methods are illustrated for interpreting the interrelationships among organics and for detecting potential data outliers. Finally, a section is included on methods to investigate the relationships between ozone and meteorological parameters and organic species.

While most of the examples appear to confirm what we already know about basic species relationships, exploratory methods offer an approach for developing new hypotheses where data do not appear to support our current understanding. Because 1993 was a "start-up" year, discovery of data anomalies should come as no surprise. As the PAMS monitoring program matures, we also would expect for such anomalies to become less frequent. It is important that the results from exploratory analysis be made available to monitoring operators and quality assurance specialists to ensure that steps are taken to upgrade the overall quality of PAMS data.

DATA COMPLETENESS

We begin the analysis by examining simple graphical methods to display data completeness. The intent is to quantify overall success in reporting data for the entire monitoring period but also to focus on patterns of missing data that may affect subsequent data interpretation. This task is somewhat challenging because of the large number of hourly parameters evaluated (approximately 20 for this analysis) and the relatively long span over which data are collected (24 hours per day for 90+ summer days).

Figure 1 consists of side-by-side box plots of the number of hours reported per day for each component or species at the

Maryland PAMS site. The median number of hours reporting per day is indicated by the dark line at the center of each box plot. The wide shaded areas span the distance between the 25th and 75th percentiles (interquartile range). The narrow shaded areas (whiskers) extend from the quartiles up to a distance equal to 1.5 times the interquartile range. Values outside this latter range are indicated as isolated dots.

For the organic species, the median number of hours reported across the three months (92 days) is approximately 17 per day. Xylene is the exception having a median reporting rate of only 9 hours per day. For the other continuous pollutant measurements (Ozone, NO, NO2, NOx and CO), most days are relatively complete (median approximately 23 per day). The meteorological parameters are also relatively complete with the exception of relative humidity for which few measurements have been reported. While these boxplots provide a good overview of data completeness, they do not show which times of the day (e.g. morning vs afternoon) are most problematic.

Figure 2 is an array that illustrates missing ethylene data by hour of the day (horizontal-axis) for each July day (vertical-axis) at the Maryland PAMS site. From this array, it is easy to identify periods of the day which are more apt to have missing data as well as any trends in data completeness among days of the month. For July, reported ethylene values after July 4 appear

relatively complete, except for a tendency towards missing values during the morning hours.

The ability to detect relationships among species may be seriously affected when the quantity of missing data is large or when missing data among species occur at different times. Although we do not illustrate the process here, it should be relatively easy to identify concurrent patterns of missing data that may present problems in analysis by simply overlaying these arrays for two or more parameters of interest.

For this particular data set, most hours (refer also to figure 1) reported either all of the organic species (except xylene) or none. Thus, analysis to examine the interrelationship among organic species is not limited by missing values for any single species but by the general availability of organic data for each candidate hour. Furthermore, since the non-organic components are relatively complete, data interpretations involving organic and non-organic species are limited by the availability of the organic species.

DIURNAL PATTERNS

Most of the pollutant species measured through the PAMS program are known to have well defined diurnal cycles that are related to both source activity (e.g. traffic) and familiar diurnal meteorological patterns. Figure 3 shows an example of

those diurnal patterns using the PAMS data taken at the Baltimore site in 1993. Each panel contains 24 box plots corresponding to each hour of the day. The panels are grouped by parameter type, i.e., organic species (figure 3a), in-organic species (figure 3b) and meteorological parameters (figure 3c).

Box plots for organic species (figure 3a) appear to indicate clearly defined diurnal trends in spite of being quite noisy. Median values for each species, except isoprene, show a tendency for higher morning and evening concentrations. These patterns appear to coincide with typical diurnal meteorological conditions, i.e., higher measured concentrations during peak traffic hours when wind speeds and vertical mixing are relatively low and lower measured concentrations during mid-day when traffic is light and vertical mixing is greatest. Relatively extreme concentrations occur sporadically among all hours for acetylene, olefins, toluene and xylene. For ethylene and isoprene the most notable extremes appear to be confined to the mid-morning hours.

Box plots for non-organic species (figure 3b) are also quite noisy with the exception of ozone. Diurnal patterns for ozone are relatively smooth with characteristically low values in the early morning hours followed by highest values in early to mid-afternoon hours. Mean diurnal patterns for NO and NOX indicate low values during mid-day hours and highest values during early morning and later evening hours. For CO, the more typical

concentrations have a pattern similar to NO, although peak levels from mid-day on appear somewhat unusual.

Box plots for meteorological parameters (figure 3c) are well behaved and exhibit familiar diurnal patterns. In Baltimore, wind speeds rise steadily during morning hours and peak in early afternoon. As expected, surface temperature and solar radiation closely track the daily solar cycle. Relative humidity (though the quantity of data is limited) is typically highest in early morning and late evening and lowest during the mid-day hours reflecting the general inverse association with temperature.

With these typical patterns in mind, we examine diurnal patterns for individual days to evaluate daily consistency and to expose potential problems with the quality of the individual values. Using SAS/INSIGHT, a graphic panel similar to that shown in figure 3a, was generated for each day. By "paging" through the results for each of the 92 summer days, it is relatively easy to spot patterns that stand radically apart from the more typical patterns shown in figure 3.

For example, figure 4 shows diurnal patterns for August 14 in which one particular hour (hour 10) stands apart from rest of the day. For isoprene, the concentration reported for that hour (252 ppb) is more than 15 times larger than any other hour reported on that day. Moreover, each of the other five organic species reports the highest value for hour 10, usually exceeding values for each adjacent hour by over a factor of 10.

Figure 5 shows a similar plot for CO on four consecutive days (July 9-12) for which the pattern of CO values appear quite unusual. On each day, reported CO levels ramp upward periodically with uncharacteristic discontinuities in early morning and evening hours. Levels of CO on these days are considerably higher than any other day for which CO data is reported.

Examination of daily diurnal patterns in each parameter (including meteorological), should be routinely performed as a preliminary quality control check on the data. If such data are confirmed to be incorrect, clearly they should be deleted before the data base is used in any subsequent analysis.

COMPARISONS AMONG ORGANICS

The relationship among the various species is largely governed by hour-to hour and day-to-day variations in source activity levels, atmospheric dispersion and, during daylight hours, photochemically driven transformations. Because the process is so complex, we will focus initially on the organic species for a time period during the day when concentrations are relatively high. Hour 6 (5-6 AM average) was chosen for this purpose, since data for this particular hour are relatively complete and not affected by photochemical processes occurring later in the day.

Figure 6 shows frequency distributions of acetylene and the logarithm of acetylene using the hour 6 concentration values. The distribution based on logarithms removes the apparent skewness and thus mitigates the impression that the largest values are an "outliers" in some sense. Furthermore, the q-q plot for the log values appears straight and provides support for the notion that acetylene (and perhaps other organic species) are approximately log-normally distributed. Approximate log-normality of organic species was also established in other analysis (Stoeckenius T. E., et. al., 1995) using data from the Houston area. The log-normality assumption will become an important consideration later on as we explore and test hypotheses regarding differences between weekday and weekend concentrations.

Figure 7 shows side-by-side boxplots for six of the organic species using the hour 6 Baltimore PAMS data. In this case, the data are all plotted using the same concentration scale, making it easy to compare the relative magnitude and distribution of values among species. Median concentrations of xylene and toluene (approximately 12 ppb) are largest followed by ethylene and acetylene (approximately 4 to 5 ppb), olefins (approximately 2 ppb) and isoprene (1 ppb). The distributions are slightly skewed as indicated by the tendency toward a few isolated large values. Again, a log-transformation of these data would probably remove this apparent skewness.

Side-by-side boxplots are also useful for comparing distributions for different data subsets. For example, figure 8 compares the distribution of acetylene on weekends (Saturday and Sunday) vs weekdays. Since acetylene is strongly associated with vehicle emissions (Scheff P. A., et. al, 1989), it is not surprising to find that acetylene values are lower on weekends when morning traffic would be expected to be relatively light compared to weekdays. For xylene, median values on weekdays and weekends are the roughly the same although there appears to be greater scatter on weekend days. Differences between acetylene and xylene weekday to weekend patterns are not surprising since sources other than traffic (e.g. solvent usage) can contribute to the total observed xylene concentration.

Because meteorology has such an important affect on pollutant concentrations, variations in meteorological conditions should be considered before drawing any conclusion about weekday vs weekend differences. A major reason for including meteorological variates in such analysis is to reduce any bias caused by the coincidental occurrence of favorable meteorology with the effect being examined. For example, if weekends were unusually windy, lower concentrations due to dilution might erroneously be ascribed to lower traffic emissions on Saturdays and Sundays. Another reason for including meteorology is to lower the residual variance used in judging the significance of the weekday-weekend effect. In the next section, we will discuss

how meteorological effects might be statistically modeled so that inferences may be drawn about differences in concentration levels among data categories (i.e., weekend vs weekdays).

METEOROLOGICAL INFLUENCES ON ORGANICS

Fluctuations in meteorological conditions are known to play an important role in affecting measured concentrations. Since meteorology affects different pollutants in different and sometimes complex ways, it may be difficult to confirm suspected relationships with limited data. Nevertheless, simple exploratory methods may be useful, especially for less reactive pollutants or during periods when photochemical activities are not dominant.

Figure 9 shows pairwise scatter plots between acetylene and xylene and three of the meteorological parameters (wind speed, wind direction and temperature), again using the data for hour 6. Both acetylene and xylene appear to have a strong inverse relationship with early morning wind speed and a very weak relationship with temperature and wind direction. Of the other two meteorological variables, relative humidity is only available for 1-2 percent of the hours while solar intensity is not a factor for this time of day.

Based on the appearance of these graphs, a simple log-linear model was used to describe morning acetylene as a function of

wind speed along with a categorical variable (ie. weekend and weekday) to account for differences between weekdays and weekends (refer back to figure 8). The results, summarized in figure 10, show that wind speed and weekday-weekend differences explain approximately 67 percent (R-Square=0.67) of the variation. Because the log of acetylene is modeled, the parameter estimates may be interpreted as the fractional change in acetylene per unit change in the independent variable (weekday or wind speed). For example, the parameter estimate for wind speed is -0.38 which means that for every 1 meter/sec increase in wind speed, acetylene decreases by approximately 38 percent. Of greater interest, is the fact that weekday values of acetylene, adjusted for wind speed differences between weekend and weekdays, are typically 60 percent higher than values on the weekends. Although the sample size here is relatively small (52 values), the difference in concentrations between weekdays and weekends is statistically significant. The model fits the data reasonably well as indicated by the close linear fit between the observed and predicted values and the linearity of the log-normal q-q residual plot. Although refinements are possible (e.g., a plot of the residuals vs wind speed would be an appropriate diagnostic check), a simple model of this type might be an adequate starting point for building more complex models to explain interspecies differences (and similarities) within an area.

SPECIES RATIOS

Another way to seek normalization of meteorological (and other) effects, is through use of ratios of selected species. For example, the ratio of individual VOC species components to TNMOC and the ratio of TNMOC to NOx have been suggested as meaningful. The presumption is that differences in the ratio among contrasting data subsets (e.g. weekends vs weekdays) are dominated by differences among source and emission characteristics, since meteorological effects (e.g. wind speed) common to each ratio component are factored out in the calculation of the ratio.

As an example, Figure 11 shows two graphs using the ratio of TNMOC to NOx. Overall, the median ratio among the 50 data values is approximately 6. Although not illustrated on these plots, two "outliers", early in the sampling period (June 1 and 3) have been removed. The graph on the left indicates that the ratio is more or less independent of wind speed, at least for the morning values. This suggests that the ratio has served the purpose of factoring out the effects of wind speed on each component.

The graph on the right, shows the ratios in the form of box plots for each weekday, beginning with Sunday (Wkday=1) and continuing through Saturday (Wkday=7). Although the dataset is very small, (approximately 7 values per day), there is a hint that ratios on Sundays are slightly higher than other days of the

week. Again, the statistical significance of the day-to-day differences could be tested using procedures similar to those described earlier in testing for weekend vs weekday effects for acetylene.

In a sense, use of ratios represent a simplification of more sophisticated methods involving source apportionment that lie outside the scope of this document. The following discussion will focus on a more generalized method (multivariate) for examining the interrelationships among species. The view is toward trying to explain how day-to-day covariations among certain species may be used to infer common underlying factors or causes.

MULTIVARIATE METHODS

Because many of the organic pollutant species originate from the same source category, we would expect to see a statistical association among those species as source activity and atmospheric dispersion varies over time. Figure 12 is a matrix scatter plot using the data for hour 6 at the Baltimore site. Isoprene was omitted from this plot since values are generally very low during this time period.

Acetylene and ethylene appear highly correlated with each other but less so with other species. Presumably, this strong association is in part due to gasoline combustion and resulting

vehicle emissions. Likewise, xylene, toluene and benzene also appear highly correlated with each other and less so with other species, presumably because they result from several common sources including both combustion, complete evaporation of raw gasoline, and other evaporative sources. Olefins, which also originate from vehicle related emissions, appear to be positively correlated with acetylene and ethylene and to a lesser degree with xylene, toluene and benzene.

Cluster analysis is one of several multivariate technique well suited for examining interpollutant relationships and helpful in identifying potential data outliers. Many of the popular clustering techniques begin with a single cluster that is essentially a linear combination of the variables used in the analysis. The second step breaks the initial single cluster into two separate clusters where each cluster is composed of one or more of the original variables. At this stage, all of the variables have been assigned to one of two clusters (groups) that hopefully "explain" a large portion of the combined variation of all of the original variables. The process continues by breaking each subsequent cluster into smaller clusters of variables until an objective stopping criteria is met. Variables (species) that are common to a given cluster are generally highly correlated with one another and have less (but still possibly large) correlation with variables in other clusters.

Figure 13 illustrates the outcome from application of a hierarchical clustering technique using the organic data. The vertical axis (reading from the top) indicates the proportion of the total variance explained by 1 cluster (86 percent), 2 clusters (93 percent) and finally 3 clusters (98 percent). In this case, the large fraction (86 percent) of the total variation explained by only 1 cluster reflects the large positive correlations that exists among the 6 species. These large positive correlations in turn reflect common source (e.g. vehicular emissions) and meteorological influences that affect all species simultaneously. The second step results in two clusters--one cluster, containing ethylene, acetylene, and olefins, and the second cluster, containing xylene, benzene and toluene. These two clusters represent subtle but distinct "signals" that are probably related to the impact of two (or more) source categories. For example, cluster 1 could represent roadway emissions since the three components are strongly related to gasoline combustion from vehicles. Cluster 2 is perhaps more difficult to interpret but probably reflects a combination of events related to evaporative losses and combustion.

In the third step, olefins break apart from the acetylene-ethylene cluster to form a single species cluster. It is not clear what signal (if any) is being sent at this point. Clearly, had more of the original species been used (and more days of data

been available), three or more interpretable clusters could have easily emerged.

Since clusters are essentially weighted averages of the variables within each cluster, a cluster score can be computed using the scoring coefficients and the individual (normalized) concentrations. In effect, the scores represent the presence or strength of that cluster for that particular day. Figure 14 shows the scatter plots of the three cluster scores along with several rotated three-dimensional plots. From these plots, no values appear radically apart from the overall body of data. In the event that one or more cluster scores appear to be "outliers", the contributing species values for that day should be investigated further to determine if the underlying data is potentially erroneous or whether the extremes might simply be related to unusual or unexpected source activity associated with that day (e.g. gasoline spill, etc).

This multivariate approach closely resembles many of the techniques used in receptor modeling (e.g., Henry R. C., et. al., 1994) to distinguish and apportion contributions from various emissions sources to observed data. We anticipate that application of more refined approaches, using a more complete suite of VOC species, will result in more informative and quantitative assessments of source contributions.

OZONE RELATIONSHIPS

Hourly ozone levels typically peak sometime between early to late afternoon at most monitoring sites in the U.S. From previous analysis of national weather and state reported ozone data over the past decade, the relationship between peak daily ozone levels and a variety of meteorological conditions has been well established (Cox and Chu, 1993). Since the PAMS program provides for similar measurements, it seems reasonable to explore the relationship between daily ozone and meteorology at PAMS sites and to compare results with historical results where appropriate.

Figure 15 is a matrix scatter plot using daily maximum 1-hour ozone and several daily meteorological parameters derived from the Baltimore PAMS data. The ozone plots (top row), suggest that the log of daily ozone has a strong positive association with daily maximum temperature, a moderately strong inverse association with morning average wind speed and a weak positive association with mid-day average solar radiation. Note also that solar radiation and temperature appear to have a weak positive association.

Using standard linear regression methods, a log-linear model was used to fit daily maximum ozone as a function of daily maximum temperature, wind speed and solar radiation. The results suggested that only temperature and wind speed were important

predictors of ozone. Solar radiation was insignificant, probably due to the overwhelming dominance of the ozone and temperature association. Figure 16 summarizes the relationship along with partial regression plots of the two significant predictors. Since a log-linear model was used, we may interpret the regression coefficients as a fractional change in ozone per unit change in the independent variable. In this case, daily ozone increases approximately 2.5 percent for each 1 degree (F) increase in temperature and decreases by approximately 8 percent for each 1 m/s increase in wind speed.

These results compare reasonably well with historical analysis (e.g., Cox and Chu, 1993) using ozone monitoring in the Baltimore area over the period from 1981-1991. The coefficients for both temperature (0.025 vs 0.032) and morning average wind speed (~ -0.08 vs -0.06) are quite close to one another. As data from additional sites and years become available, the significance of such comparisons may take on greater meaning.

Finally, the regression model was modified to include selected organic species (hour 6) to determine if additional variation in ozone could be explained. For these limited data, no organic species has a statistically significant relationship with ozone for reasons that we may only speculate on at this point.

SUMMARY

Data produced from the PAMS monitoring program will provide valuable new information needed by air quality planners to more effectively control ozone precursors and toxic pollutants. The purpose for this workbook has been to provide a quick overview of exploratory methods useful in preliminary investigation of PAMS data prior to more extensive analysis to support specific data objectives. Graphical methods are emphasized as being particularly useful for examining the shape of data distributions and for detection of potential outliers. Likewise, graphical displays of diurnal patterns define a useful frame of reference for detecting certain types of data anomalies. Side-by-side boxplots provide for simple visual comparisons among pollutant species and help in revealing differences in concentration levels that may be attributed to differences in categories of source activity (e.g., weekend vs weekday). Methods for incorporating meteorological data into the analysis of ozone and species relationships may prove useful for removing potential bias and for increasing the power associated with specific hypotheses of interest.

REFERENCES

Cohen J. and Stoeckenius T. E. (1992) Analysis of Sources of Variability in the Atlanta 1990 Ozone and Ozone Precursor Study Data. SYSAPP-92/094, Systems Application International, San Rafael, California 94903.

Cox W. M. and Chu S (1993) Meteorologically Adjusted Ozone Trends in Urban Areas: A probabilistic approach. Atmospheric Environment 4.

Federal Register, (58 FR 8542), Ambient Air Quality Surveillance-Final Rule, February 12, 1993.

Henry R. C., Lewis C. W. and Collins J. F. (1994) Vehicle-Related Hydrocarbon Source Compositions from Ambient Data: The GRACE/SAFER Method. Environmental Science & Technology 28

Hoaglin D. C., Mosteller F., and Tukey J. W. (1983) Understanding Robust and Exploratory Data Analysis. John Wiley, New York.

Hoaglin D. C., Mosteller F., and Tukey J. W. (1985) Exploring Data Tables, Trends, and Shapes. John Wiley, New York.

Mathsoft (1993) S-PLUS Users Manual, Version 3.2. Mathsoft, Inc, Seattle, WA.

SAS Institute Inc. (1993) SAS/INSIGHT User's Guide, Version 6, Second Edition. SAS Institute Inc., Cary, NC.

Scheff P. A., Wadden R. A., Bates B. A., and Aronian P. F. (1989) Source Fingerprints for Receptor Modeling of Volatile Organics. Journal Air Pollution Control Association 39.

Stoeckenius T. E., Ligocki, M. P., Cohen J. P., Rosenbaum A. S., and Douglas, S. G. (1994) Recommendations for Analysis of PAMS Data. SYSAPP94-94/011r1, Systems Application International, San Rafael, California 94903.

Stoeckenius T. E., Ligocki M. P., Shepard, S. B. and Iwamiya R. K. (1995) Analysis of PAMS data: Example Application to Summer 1993 Houston and Baton Rouge Data. SYSAPP-94/115d, Systems Application International, San Rafael, California 94903.