Data Analysis and Validation Support for PM2.5 Chemical Speciation Networks- #82

Max R. Peterson and Edward E. Rickman
Research Triangle Institute
3040 Cornwallis Road
P.O. Box 12194
Research Triangle Park, NC 27709
James B. Homolya
C339-02
U.S. EPA Mailroom
Research Triangle Park, NC 27711

ABSTRACT

The U.S. Environmental Protection Agency (EPA) has developed software and workshop training materials to assist State, Local, and Tribal monitoring agencies in their review and validation of PM2.5 chemical speciation data reported monthly by EPA's contract laboratory (RTI). Nine hands-on training workshops have been given (three each in Research Triangle Park NC, Chicago IL, and Denver CO) in 2002 with a total attendance of 109 people. Workshop materials included: the Speciation Data Validation Analysis Tool (SDVAT), which is a Microsoft Access software application; an SDVAT Users Guide; overviews of data sources, data validation, and the SDVAT; and hands-on exercises. Higher-level data analysis features of the SDVAT include data completeness, time series analysis, mass concentration reconstruction analysis, and major component distribution analysis.

This paper presents data analysis features of the current version of the SDVAT, its usefulness in identifying potential problem data, and its flexibility in meeting the needs of individual users. Some possible upgrades to the SDVAT will also be presented, including data analyses that combine air concentration data for a pollutant with meteorological data to determine compass directions toward the principal sources for that pollutant at a given sampling site.

INTRODUCTION

In May 2001, OAQPS/EPA-RTP issued a work assignment through a quality assurance contract with Research Triangle Institute (RTI) to develop a Microsoft Excel-based software tool and hands-on training materials for use by State, Local, and Tribal monitoring agencies in validating PM2.5 chemical speciation data. The Speciation Data Validation Analysis Tool (SDVAT) was developed during the summer and previewed by a local group of data reviewers and statisticians from the State of North Carolina in September 2001. The SDVAT was modified to accommodate as many of the suggestions from reviewers (from the State of North Carolina, RTI, and EPA) as possible within the usual budget and time limitations.

A users guide[1] for the SDVAT was finalized and training materials[2] for a one-day hands-on workshop were developed during the fall of 2001. Three consecutive hands-on workshops were presented January 28-31, 2002, at EPA in the Research Triangle Park, NC, with a total of 39 people participating. Revisions to the SDVAT and training materials were made based on

lessons learned in the first round of workshops, and the revised tool and materials were used in a second series of three consecutive workshops at EPA Region 5 headquarters in Chicago, IL, April 30-May 2, 2002, with a total of 40 people attending. One or two minor changes were made to the SDVAT for the third series of three consecutive workshops at EPA Region 8 headquarters in Denver, CO, July 16-18, 2002, with a total of 30 people attending. All participants in the workshops received copies of the SDVAT, the users guide, and the workshop materials.

Currently, user feedback is being solicited from the 109 people who attended one of the workshops, and ways to use meteorological data in the SDVAT are being developed and evaluated. If there is sufficient interest, the SDVAT will be upgraded and made available to monitoring agencies. Training may also be offered if there is sufficient demand.

OVERVIEW OF PM2.5 CHEMICAL SPECIATION DATA

Rolling Schedule for PM2.5 Speciation Data

RTI makes partially validated data reports (for $PM_{2.5}$ samples received for analysis 30-59 days earlier) available to monitoring agencies by the 15th of each month. Data are made available in four formats (three Microsoft Excel spreadsheets and one rich text file) on a secure web site (user name and password required) to a data contact specified by each monitoring agency. Monitoring agencies have 45 days to validate the data and notify RTI of any changes. RTI has 15 days to make changes and upload validated data to AIRS. This schedule places data in AIRS within 90-119 days of receipt of samples by RTI.

RTI Monthly Analytical Reports

Table 1 summarizes the data included in RTI monthly analytical reports. Mass and concentration data are provided for either 58 or 59 analytes (depending on sampler type[3]), and the remainder of the data items are taken from the Field Data and Custody Forms that are completed by field sampling personnel. The number of field data items included varies by sampler type, but the total number of data items for each sample averages about 100.

Table 1. RTI Monthly Analytical Reports.

• Field sampling data – Data from *Field Sampling Data and Custody Forms*: Number of data items varies by sampler type but averages about 40.
• Analysis data for 59 analytes – Gravimetric Mass – Ions by Ion Chromatography[3]: SO_4^{2-}, NO_3^-, NH_4^+, Na^+, K^+ – Carbon by Thermal-Optical Transmittance Analysis: OC, EC, CO_3^{2-}, TC, OCX2 – Elements by X-Ray Fluorescence: Al, Sb, As, Ba, Br, Cd, Ca, Ce, Cs, Cl, Cr, Co, Cu, Eu, Ga, Au, Hf, In, Ir, Fe, La, Pb, Mg, Mn, Hg, Mo, Ni, Nb, P, K, Rb, Sm, Sc, Se, Si, Ag, Na, Sr, S, Ta, Tb, Sn, Ti, V, W, Y, Zn, Zr

Data Validation Checks

Data validation is performed at four levels numbered 0 through 3. Level 0 includes a basic review of data with respect to their provenance or origin. Level 1 includes the process of evaluating the correctness and acceptability of individual items or groups of items within the data set using statistical methods and other screening techniques. Levels 2 and 3 include higher-level validation activities such as correlations between sites, time-series analysis, and other analyses and correlations as well as modeling.

RTI partially validates the PM2.5 speciation data provided to monitoring agencies[4]; and the monitoring agencies then confirm RTI's validation checks and perform additional, higher-level checks.[5] Level 0 and Level 1 data validation checks are performed by personnel in RTI's sample handling facility, analytical laboratories, and data entry and management areas. Examples of RTI's validation checks and some of the validation checks typically performed by monitoring agencies are shown in Table 2.

Table 2. Validation Activities.

• RTI: Level 0 – Sample Identification – Operator Observations – Sampler Flags – Shipping & Disassembly – Laboratory Checks (per SOPs) – Range Checking - flow rate - exposure duration - elapsed time before retrieval - holding times	• RTI: Level 1 – Mass balance - sum of species concs./mass conc.) – Charge balance - cation/anion ratio – Analyte correlations - sulfate/sulfur ratio $\simeq$ 3
• Monitoring Agency: Levels 0 and 1 – Confirm RTI validation checks	• Monitoring Agency: Levels 2 and 3 – Data Completeness – Time Series Analysis – Mass Reconstruction Analysis – Species Distribution Analysis – Others - correlations between sites - collocated bias and precision - modeling

RTI partial validation checks compare individual sample concentrations to expected concentration ranges developed from results for a large number of samples; monitoring agencies confirm those checks and add temporal and spacial checks.

OVERVIEW OF THE SDVAT

The Speciation Data Validation Analysis Tool (SDVAT) was developed to assist monitoring agencies in reviewing and validating data from RTI's monthly analytical reports. The SDVAT can be run on a personal computer with a Pentium 300 MHz (or similar) processor, 256 MB RAM, adequate hard disk space, Windows 95/98/ME/NT/2000, and Microsoft Excel 2000 (Microsoft Office 2000 Professional suite preferred). The SDVAT is a Microsoft Access application in a single file that is installed by simply copying it to the hard drive of a personal computer. The application can be renamed, and multiple databases can reside in the same directory or in different directories on the same computer.

The SDVAT is useful for verifying RTI Level 0 and Level 1 validation checks, combining data from multiple reports into a single database, creating charts and graphs for Level 2 and Level 3 validation checks, looking at temporal and spacial data trends, and for flagging data with user-defined flags. The user selects data based on sampler and site, chemical analysis, analytes, date range, sample types, and data validity types. A "User Omit" feature allows the user to exclude data from an analysis by the SDVAT. Output is exported to a stand-alone Microsoft Excel Workbook with pre-formatted chart sheets linked to data worksheets. Formatting of the charts and data in the workbook can be changed by the user (in Excel).

Four types of analyses are currently available in the SDVAT: data completeness, time series analysis, mass reconstruction analysis, and species distribution analysis. These analyses and the output produced by the SDVAT are described below.

Data Completeness

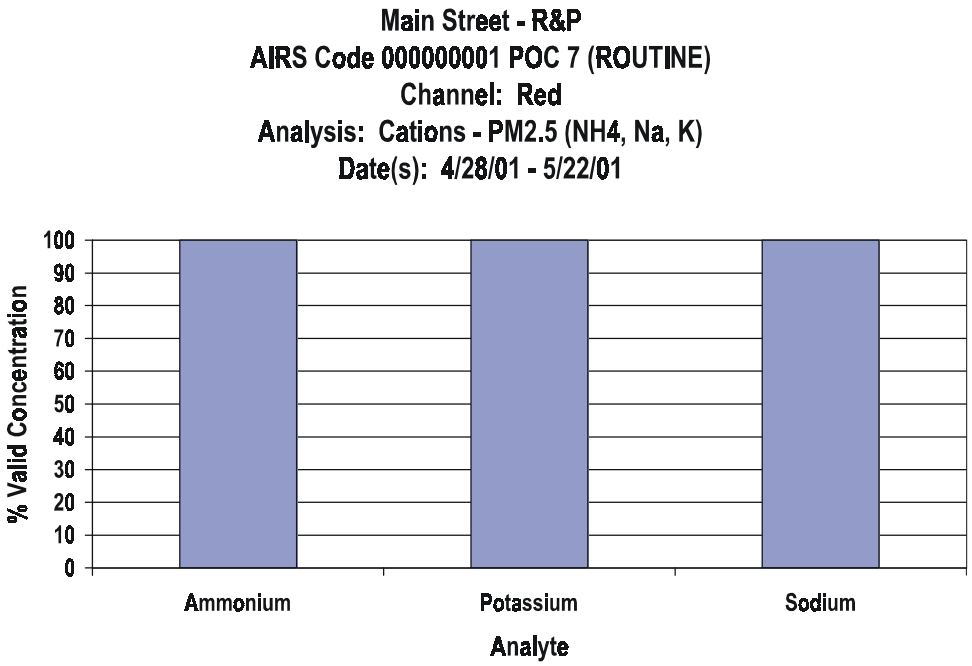
Data completeness is an indicator of data capture and is calculated separately for each analyte or parameter. The equation used to calculate data completeness in the SDVAT is shown in Equation 1.

Equation 1. Data completeness is computed as percent of results that are valid.

$$\%\ \text{Valid} = \frac{\text{Number of Valid Results}}{\text{Total Number of Results}}\ (100\%)$$

Figure 1 shows a typical column chart for data completeness. The selection criteria chosen by the user are in the title of the Excel chart. The data selected by the user was for samples collected on a sampler named the "Main Street - R&P." The AIRS Code for the site (fictitious in this case) and the POC (pollutant occurrence code) for the sampler are given in the second line of the header. Channels on R&P samplers are designated by colors, and "Red" is the code for the channel containing the nylon filter from which ions are extracted and measured. The analysis chosen was "Cations - PM2.5 (NH4, Na, K)," and the date range chosen for samples was "4/28/01 - 5/22/01." The chart shows 100% completion for the three analytes in samples collected for the Main Street - R&P over the chosen time period.

Figure 1. Data Completeness Column Chart



Time Series Analysis

A time series analysis is a simple plot of an analyte or some field-reported value versus time and requires no calculations. Time series plots can be generated as xy-line plots (Figure 2) and as stacked column plots (Figure 3).

Figure 2. Time Series Analysis XY-Plots

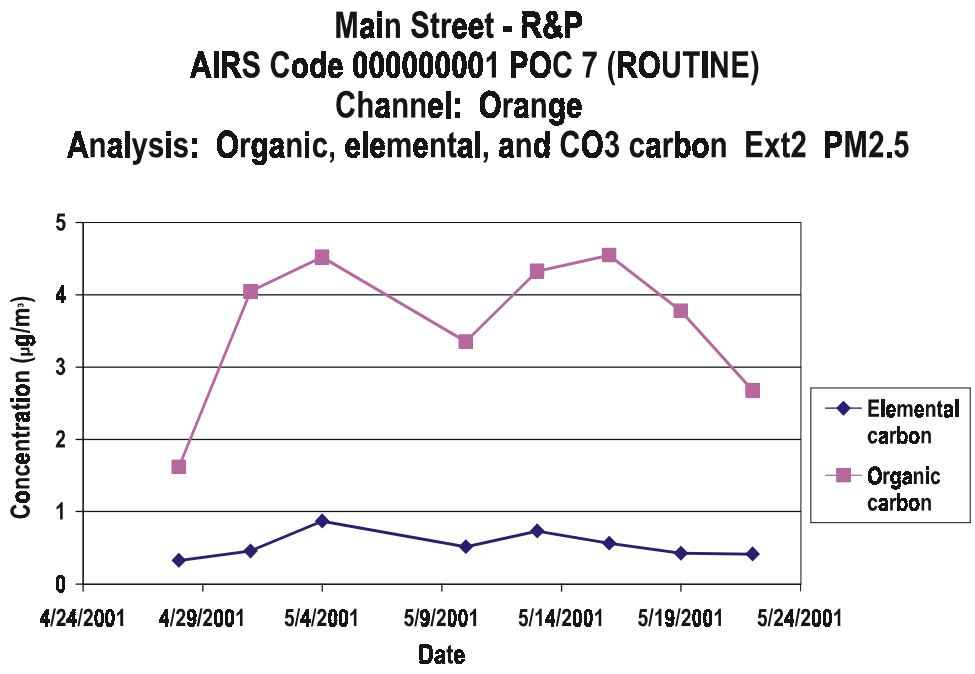
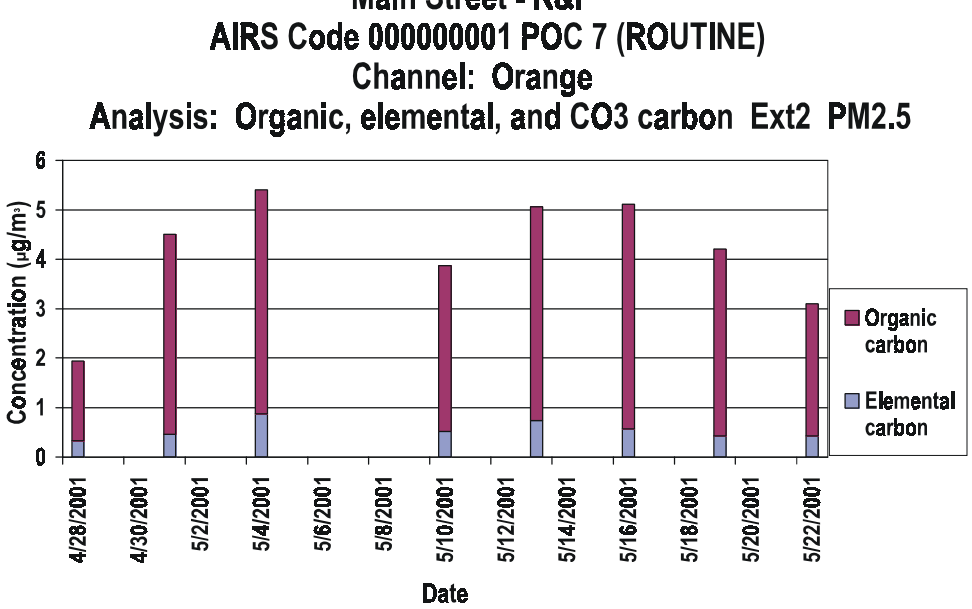


Figure 3. Time Series Stacked Column Chart



Mass Concentration Reconstruction Analysis

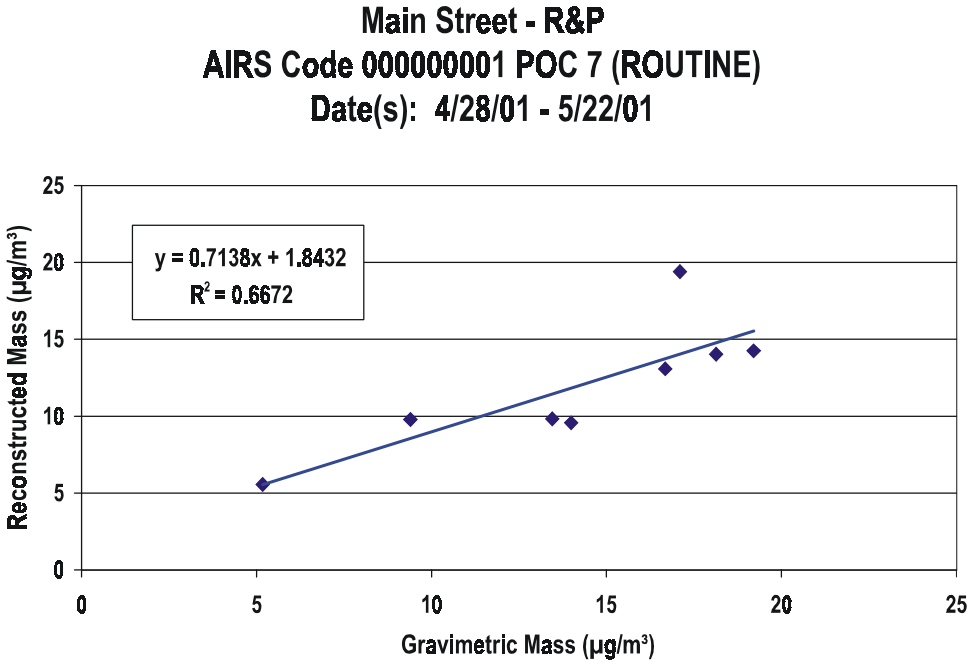
Mass concentration reconstruction analysis is an xy-plot of of gravimetric mass concentration vs. reconstructed mass concentration with a linear trend line added to the graph. Reconstructed mass concentration is calculated according to Equation 2.

Equation 2. Reconstructed mass concentration is the sum of the concentrations of the anions, cations, carbon, and trace elements (excluding S, Na, and K, which are measured as ions) for a sample with no attempt to correct for species not measured (oxygen, chlorine, other metals, etc.).

$$\begin{bmatrix}\text{Reconstructed}\\ \text{Mass}\\ \text{Concentration}\end{bmatrix} = \begin{bmatrix}\sum\text{Anions(IC)} + \sum\text{Cations(IC)} + \text{Total Carbon}\\ +\sum\text{Trace Elements(XRF), excluding S, Na, K}\end{bmatrix}$$

Figure 4 shows a typical mass concentration reconstruction plot. The data point with a reconstructed mass concentration of almost 20 µg/m³ is quite far from the trend line compared to the other data points, and the PM2.5 sample for which that point was calculated should probably be investigated as part of the monitoring agency's data validation process.

Figure 4. Mass Concentration Reconstruction XY-Line Plot



Species Distribution Analysis

Species distribution analysis is a graphic representation of the composition of PM2.5 by seven major components: nitrate (total[3]), sulfate, ammonium, organic carbon, elemental carbon, a crustal component, and other. The contribution of the crustal component is currently calculated by the SDVAT according to Equation 3.

Equation 3. The crustal component is calculated as the sum of the concentrations of five common elements multiplied by coefficients used to approximate additional mass contributions from for oxygen attached to those elements in crustal materials.

$$\left[\begin{array}{c}\text{Crustal}\\ \text{Component}\end{array}\right] = 2.2[\text{Al}] + 2.49[\text{Si}] + 1.63[\text{Ca}] + 2.42[\text{Fe}] + 1.94[\text{Ti}]$$

An improvement under consideration for the SDVAT would allow changes in the coefficient in front of the five major elements to adjust for differences in composition of crustal material in different parts of the country. The current coefficients are those used in the IMPROVE program.

The "other" component is calculated according to Equation 4. If "other" is less than zero, then "other" is set equal to zero.

Equation 4. "Other" represents the remainder of species not included in the major components listed above.

$$\text{Other} = \left[\begin{array}{c}\text{Gravimetric}\\ \text{Mass}\\ \text{Concentration}\end{array}\right] - \left[NO_3^- + SO_4^{2-} + NH_4^+ + \text{OC} + \text{EC} + \left[\begin{array}{c}\text{Crustal}\\ \text{Component}\end{array}\right]\right]$$

Species distribution output choices include xy-line plots (Figure 5), stacked-column charts (Figure 6), and a pie chart (Figure 7) of the averages for the seven major components.

Figure 5. Species Distribution Analysis XY-Line Plots

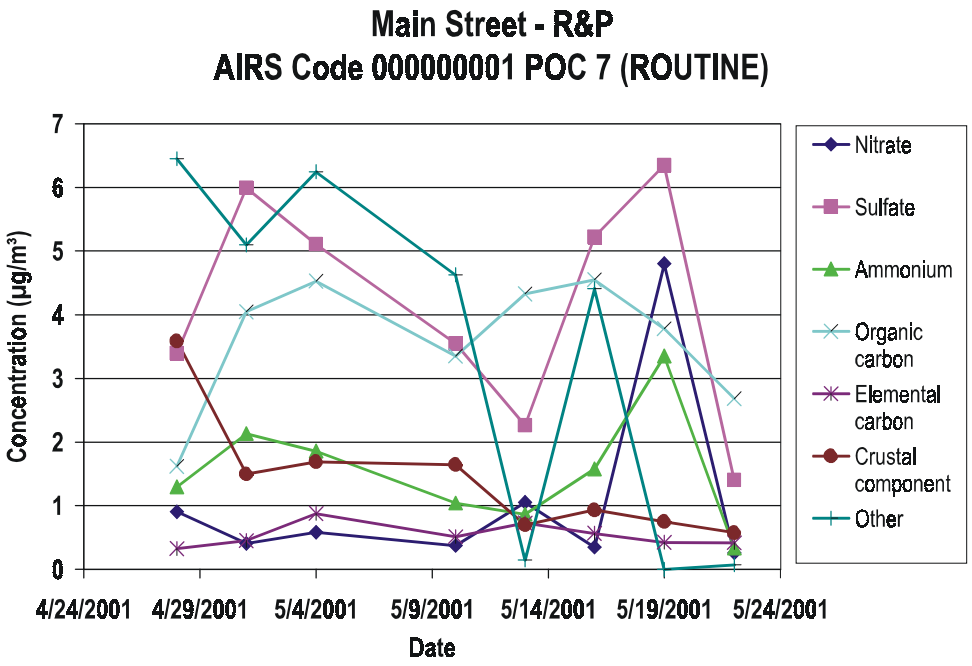


Figure 6. Species Distribution Analysis Stacked Column Plots

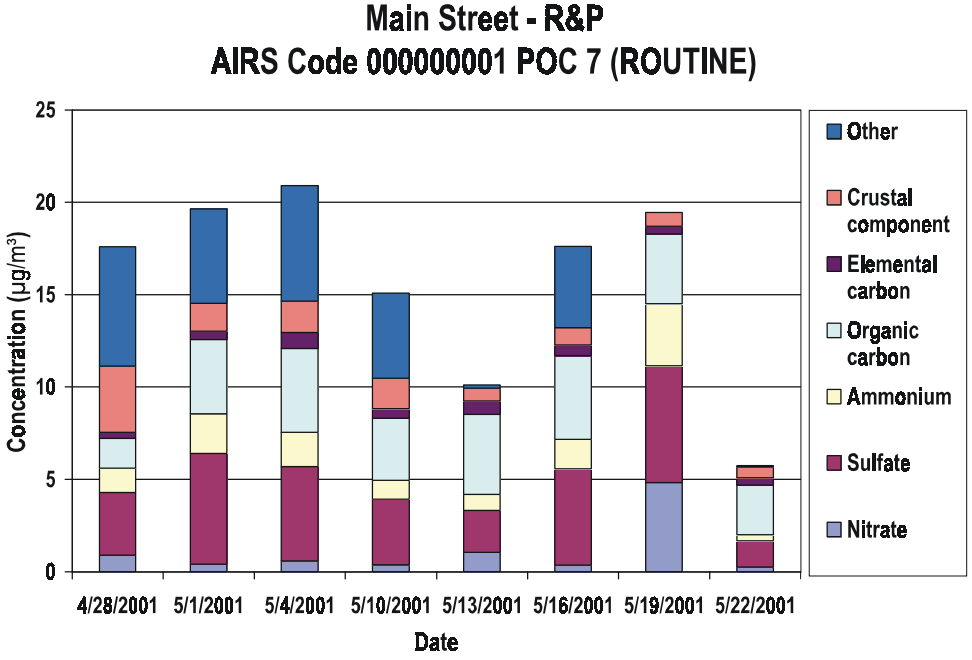
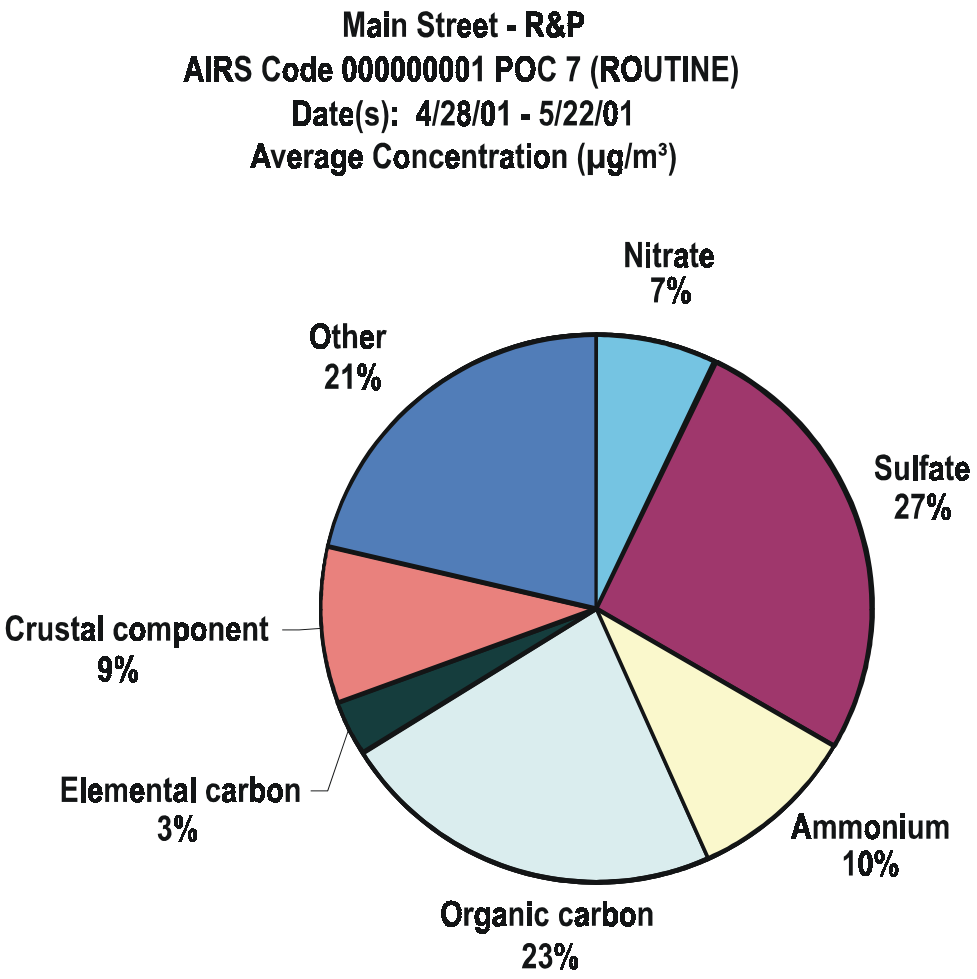


Figure 7. Species Distribution Analysis Pie Chart (Averages)



Other Features of the SDVAT

The SDVAT allows import of multiple data reports into a single database; deletion or export of selected data; use of user-defined flags; and selection of data based on user-assigned flags. The quantity of data that can be stored and manipulated by the SDVAT is limited primarily by the storage capacity, speed, and memory of the computer system used.

USE OF METEOROLOGICAL DATA

Background

The two big questions for air monitoring agencies are "What's in the air?" and "Where did it come from?" For the PM2.5 Chemical Speciation program, the monthly analytical reports provide answers to the first question; the second question requires additional information.

The very first work assignment EPA issued regarding data validation software listed input of meteorological data as a feature for the proposed software tool; however, time and funding resources were not sufficient to develop a procedure for getting meteorological data into the SDVAT or to determine useful and appropriate ways to use the data once it was there. Because of the importance of meteorological data in source identification and the power and ease-of-use of the SDVAT, RTI funded an internal research and development project in the summer of 2002 to study the use of meteorological data in source identification as it might be incorporated into an upgraded version of the SDVAT. The use of meteorological data for source identification described in the following section was developed under the RTI-funded project and is currently being evaluated.

One Possible Way to Use Meteorological Data in the SDVAT

The National Weather Service (NWS) provides, for a fee, detailed weather data for sites all across the U.S. NWS monthly data reports provide daily summaries and hourly readings for, among many other parameters, wind direction, which is a key parameter for transporting a pollutant from a source to a sampler at a monitoring site. The daily summary value for wind direction is a resultant value determined by taking the vector sum of wind speed and wind direction readings for that day. Because the resultant wind direction takes into account wind speed and because the vector sum is a mathematical computation, it is possible that the wind did not blow at all from the direction calculated as the resultant wind direction for that day. Excel xy-plots of pollutant concentration versus resultant wind direction (similar to plots that would be created from an upgraded version of the SDVAT) give inconclusive results at best, especially for a small data set (<30 samples). In addition, multiple sources for a particular pollutant in different directions from the sampling site complicate the trends in such plots.

A more accurate way to approach wind direction is to use the hourly data. Obviously, the hourly wind direction readings provide more detailed information, but PM2.5 chemical speciation measurements are daily values. The procedure in Table 3 shows one possible way to use hourly wind direction readings to determine directions toward pollutant sources from a sampling site.

Table 3. One Possible Way to Use Hourly Wind Direction Readings for Source Direction Identification.*

- **For each sample:**
  - Get hourly wind direction readings for nearby NWS station taken during sampling period.
  - Count, by 10s of degrees, the number of hours the wind blew from each of the 36 directions during sampling.
- **For an analyte of interest over many samples:**
  - For each wind direction, determine slope (by linear regression analysis) of a plot of air concentration vs. count (or fraction) of hours wind blew from that direction during sampling.
  - Do radar plot of slopes by wind direction.

*Developed as part of an RTI-funded internal research and development project.

Figure 8 shows half of the slope plots (18 of 36 wind directions) calculated for gravimetric mass concentration using actual speciation data from a monitoring site (designated Site A) in a medium-sized city with relatively clean air and hourly wind direction data taken during each sampling period at a nearby NWS weather station. The trend lines added to the plots in Figure 8 indicate that some wind directions give positive slopes, which indicate relatively dirty directions or directions towards sources, and some give negative slopes, which indicate relatively clean directions or directions with lower emissions.

Figure 8. Site A: Gravimetric Mass Concentration versus Percent of Hourly Wind Direction Readings

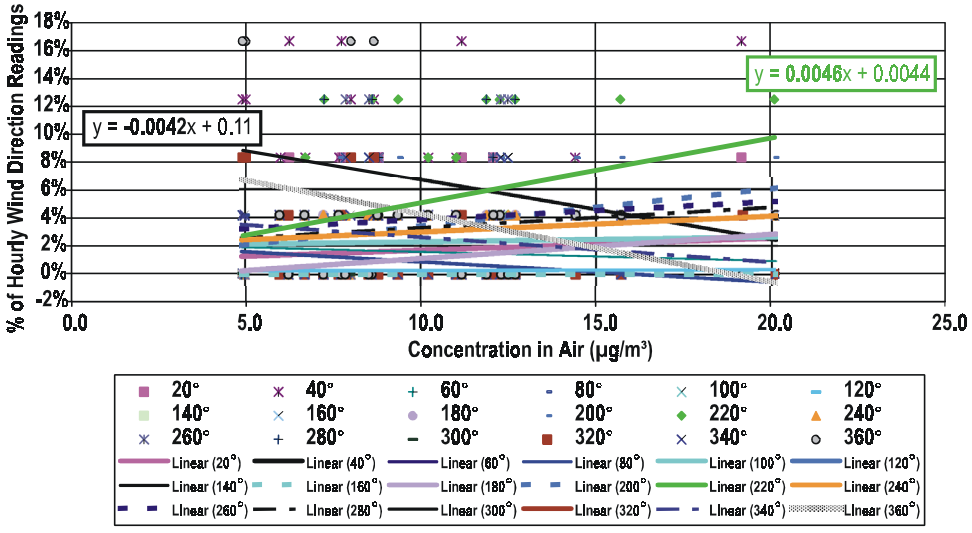
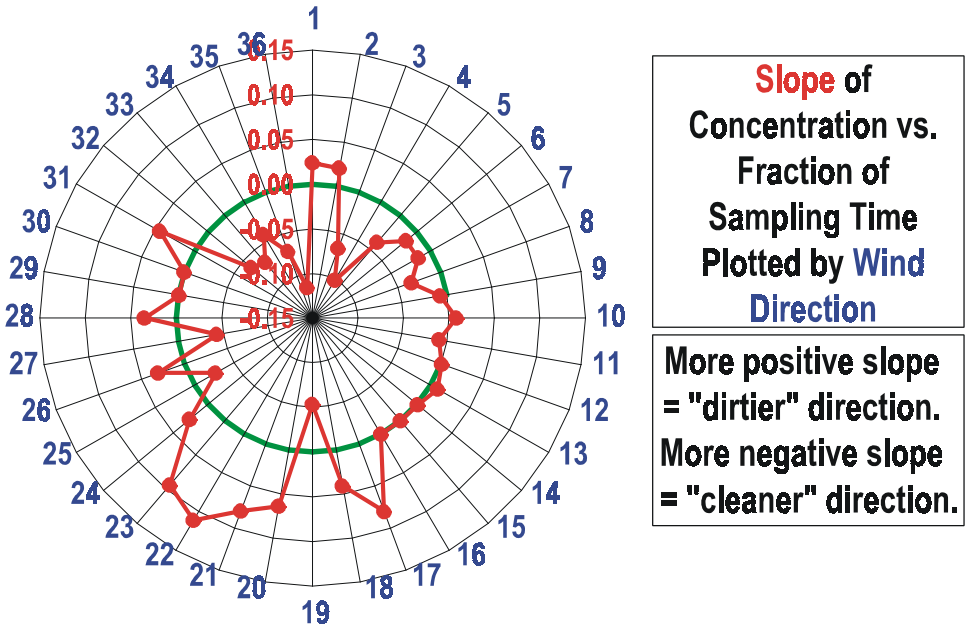


Figure 9 is an Excel radar chart with 36 categories (corresponding to the 36 possible wind directions) with the slope of the trend line (from a plot of percentage of hourly wind direction readings for that direction versus measured analyte concentration in air for each sample in the selected data set) plotted radially from the center along the corresponding radius. The green circle was added to the chart to indicate the location of zero on the slope axis. Slopes outside the green circle indicate "dirtier" directions; and slopes inside the green circle indicate "cleaner" directions. This chart could be created for any one of the 58 or 59 analytes[3] measured in the PM2.5 chemical speciation program for any selected time span.

Figure 9. Site A: Slope of Gravimetric Mass Concentration versus Fraction of Hourly Wind Direction Plots In a Radar Plot



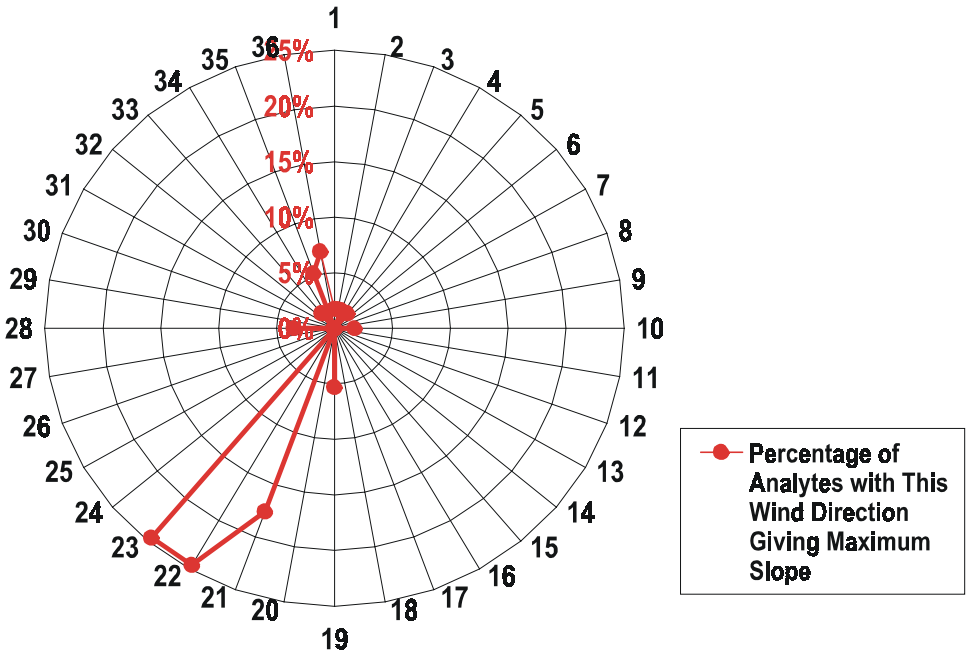
The large number of analytes (Table 1) measured in the PM2.5 chemical speciation program provide an additional dimension to the data set. (This approach must be used with caution because equal weight is assigned to all analytes, including those present at concentrations below their detection limits.[6]) Table 4 gives the percentage of the speciation analytes for which a given wind direction gives the maximum slope, and Figure 10 shows the results of plotting the data in Table 4 in an Excel radar chart. Note the similarities in potential source directions in Figures 9 and 10.

Table 4. Site A: Percentage of Analytes With Maximum Slope by Wind Direction.*

Wind Direction	Percentage of Analytes for Which This Wind Direction Gives Maximum Slope
10°	1.75%
20°	1.75%
30°	1.75%
50°	1.75%
100°	1.75%
190°	5.26%
210°	17.54%
220°	24.56%
230°	24.56%
280°	3.51%
330°	1.75%
350°	5.26%
360°	7.02%

*Wind directions for which no analytes (0%) have a maximum slope have been removed from the table. The number of analytes varies by sampler type (see Table 1).

Figure 10. Site A: Percentage of Analytes with Maximum Slope by Wind Direction Radar Plot



A map of the city in which the samples used in this example were collected indicates a medium-sized airport[7] in a southwest direction (220°-230°) from the sampling site, which agrees well with the radar plot in Figure 10. The directions toward less significant sources have not been investigated.

A second sampling site in a medium-sized city with more pollution sources has also been investigated with similar results. Figure 11 shows the radar plot of the slopes of plots of gravimetric mass concentration versus fraction of hourly wind direction readings by direction. The sampling site is located in the northeastern quadrant of the city. Figure 12 is a radar plot of percentage of the speciation analytes for which a given wind direction gives the maximum slope. A medium-sized airport,[7] which is located some distance from the sampling site, is a candidate for sources contributing from the 290° direction. Other potential sources around the second site are being investigated.

Figure 11. Site B: Slope of Gravimetric Mass Concentration versus Fraction of Hourly Wind Direction Plots In a Radar Plot

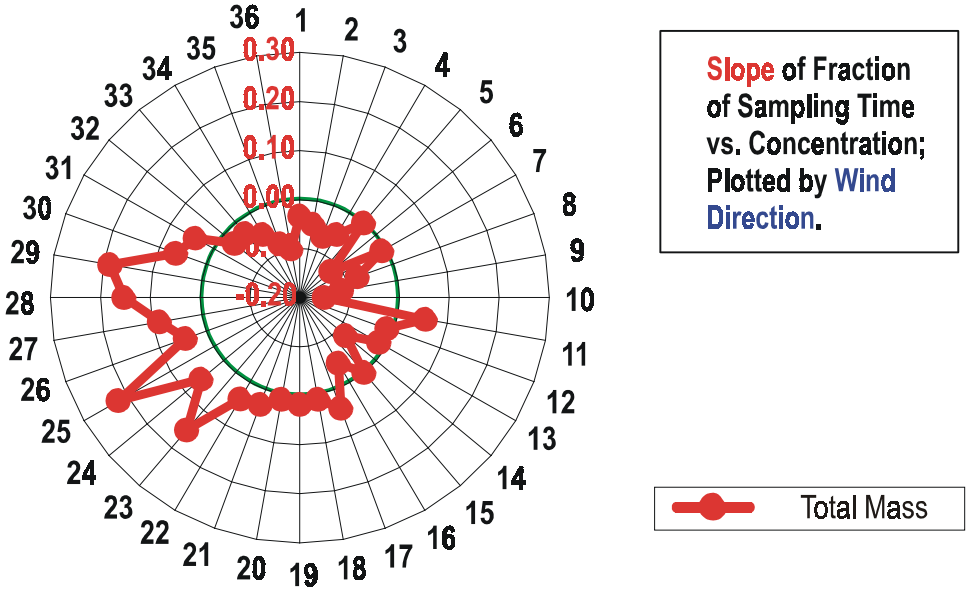
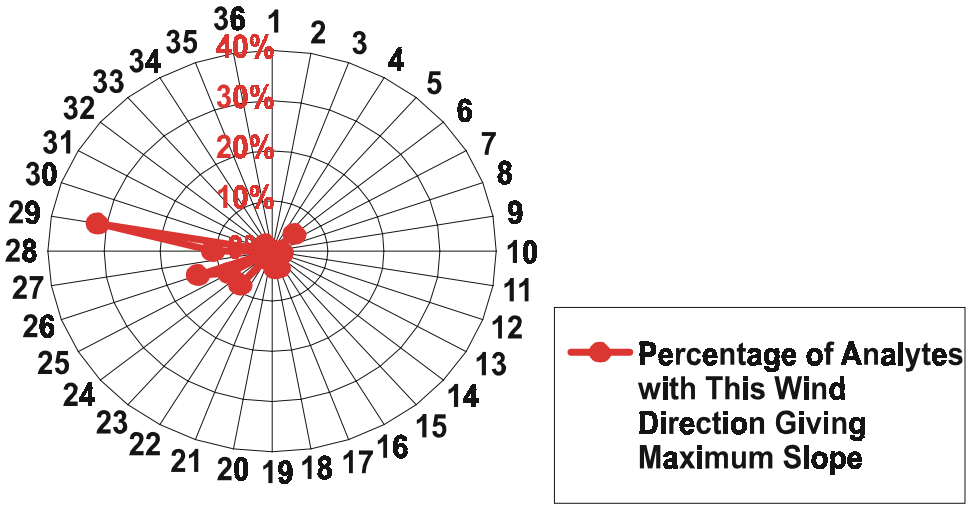


Figure 12. Site B: Percentage of Analytes with Maximum Slope by Wind Direction Radar Plot



CONCLUSIONS

The SDVAT is currently a powerful tool for performing data completeness, time series analysis, mass reconstruction analysis, and species distribution analysis. Addition of data analyses including meteorological data could extend use of the SDVAT to source identification. At least one possible way to use hourly meteorological data has been identified and should be evaluated. Ways to get appropriate meteorological data linked to PM2.5 samples and into a file format that could be imported into the SDVAT are being considered.

Comments and suggestions are being solicited from current users of the SDVAT. Software upgrades and future training workshops depend upon the usefulness and value of the SDVAT to State, Local, and Tribal monitoring agencies.

ACKNOWLEDGMENTS

Development of the SDVAT and all training materials used in the workshops and the workshop presentations themselves were funded through EPA Contract 68-D-98-032, Work Assignments 3-08, 4-02, and 4-03.

The study of possible ways to use meteorological data in PM2.5 source identification was funded by RTI as an internal research and development project.

REFERENCES AND NOTES

1. Rickman, E.; Lloyd, J.; Simone, E.; Andrews, L. *Speciation Data Validation Analysis Tool (SDVAT) User's Guide*; prepared for the U.S. Environmental Protection Agency/Office of Air Quality, Planning, and Standards under U.S. EPA Contract 68-D-98-032, Work Assignments 3-08, July 8, 2002.

2. Peterson, M.R. *Data Validation Analysis Support for the PM2.5 Speciation Networks*; prepared for the U.S. Environmental Protection Agency/Office of Air Quality, Planning, and Standards under U.S. EPA Contract 68-D-98-032, Work Assignments 3-08, 4-02, and 4-03, April 2002.

3. For samplers with separate channels for each of the three filters (nylon, teflon, and quartz), nitrate is reported as total nitrate (extracted from the nylon filter). For samplers in which the nylon filter is placed downstream from and in the same channel as the teflon filter, nitrate is reported as nonvolatile nitrate (extracted from the teflon filter) and volatile nitrate (extracted from the downstream nylon filter); and total nitrate can be calculated as the sum of nonvolatile nitrate and volatile nitrate.

4. A complete description of RTI's Level 0 and Level 1 data validation procedures can be found in *Draft Data Validation Process for the PM2.5 Chemical Speciation Network*, available at http://www.epa.gov/ttn/amtic/pmspec.html. July 5, 2000.

5. A discussion of monitoring agency validation of PM2.5 chemical speciation data can be found in *Speciation Trends Network Quality Assurance Project Plan*. Available at http://www.epa.gov/ttn/amtic/pmspec.html. February 28, 2001.

6. If the Percentage of Analytes with Maximum Slope by Wind Direction Radar Plot is included in a future version of the SDVAT, the software could allow users to select which analytes to include in the calculations. This would allow users to omit from the calculations analytes present only at concentrations below their respective minimum detection limits.

7. Airports, which certainly may not be a major source of $PM_{2.5}$ overall, were singled out as potential sources in this study solely because they appear on public maps while other potential stationary sources typically do not.