Paper 136-27

Using SAS® Graphics Capabilities to Display Air Quality Data
Debbie Miller, National Park Service, Denver, CO

ABSTRACT
SAS software is used to create detailed charts showing air quality data. These charts will be used by state air quality specialists to develop plans to comply with new air quality regulations. SAS/GRAPH® procedures are used to generate the charts.

This paper highlights the use of the Annotate facility to add annotations unique to each chart that greatly enhance the amount of information that can be conveyed. This facility allows you to automatically place annotations on each graph with respect to other chart elements such as bar height and x-axis location. Instructions for creating the annotations are stored in a data set and referenced later in a SAS/GRAPH procedure such as the GCHART procedure.

This paper is intended for intermediate and advanced SAS users who are familiar with SAS/GRAPH. There is no limitation with respect to operating system.

INTRODUCTION
In the 1977 amendments to the Clean Air Act, Congress established as a national goal the prevention of any future and the remedying of any existing impairment to visibility resulting from manmade air pollution in 156 national parks and wilderness areas designated as federal 'Class I' areas. Visibility impairment is caused by gases and particles that affect the passage of light through the atmosphere. This impairment is often in the form of a uniform haze that washes out color and makes distant landmarks more difficult or impossible to see. Particles are typically the major contributor to visibility impairment and although some particles have natural sources, many have manmade origins. Particles that cause visibility impairment are very small, generally less than 10 μm in diameter. A recent ruling by the Environmental Protection Agency (EPA) known as the Regional Haze Rule (RHR) directs states to restore visibility on the haziest days in Class I areas to 'natural background' conditions.

Particles that cause haze are generally composed of six chemical constituents: 'fine' nitrate, sulfate, soil, organic, and light-absorbing carbon particles with diameters less than 2.5 μm, and 'coarse' mass particles with diameters between 2.5 and 10 μm. Such chemically speciated particle concentration data have been collected at many of these Class I areas beginning in 1988. As states begin to develop control strategies to meet the goal of the RHR, air quality planners need to be able to look at historical data and determine whether seasonal patterns exist in the distribution of the clearest and haziest days, which chemical compounds in the particles contribute most to visibility impairment, and whether year to year differences in visibility conditions exist. They also want to be able to identify outliers caused by events such as fires or dust storms. A graphical display provides a rapid method for examining multiple years of data and answering these types of questions.

A SPECIALIZED BAR CHART TO DISPLAY AIR QUALITY DATA
Figure 1 shows an example of a stacked bar chart created using SAS software that shows the contributions of each of the six categories of particulate mass to each day's extinction. The term 'extinction' refers to the amount of light removed from a scene by scattering and absorption, and is measured in units of inverse megameters (Mm^{-1}). For each year, days are ranked by their total extinction, and the top 20% are identified. Each of these 20% haziest days is displayed on a bar chart. The height of each bar indicates the total light extinction due to particles. The relative contribution of each chemical component is indicated by a different colored portion of each bar. The contribution of sulfates to each day's extinction is indicated by the yellow portion of each bar, nitrates by the red portion, fine soil by the brown portion, coarse mass by the gray portion, organics by the green portion, and light-absorbing carbon by the black portion. Data samples have historically been collected each Wednesday and Saturday; all possible sample days in a given year are indicated on the horizontal axis, allowing seasonal patterns in the distribution of haziest days to be easily identified. For various reasons, it is possible on any given sample date to be missing one or more components, or to be missing the entire sample. An empty bar indicates that all components are missing on a sample date. The absence of any bar on a particular date indicates that a sample was collected, but that day was not determined to be one of the 20% haziest days. Vertical lines on the charts mark the boundaries between calendar quarters. The mean extinction value of the year's 20% haziest days is displayed in the title above each chart, allowing for year to year changes in mean haziness to be identified.

Where at least one component of particle mass is available, substitution is performed for any missing components. For each mass type, mass concentration data are grouped by month across all years of data, and the 10th percentile is determined for each mass type. The 10th percentile is then substituted for each missing component in order to achieve a complete sample without artificially inflating a day's total extinction. Each occurrence of a component substitution is indicated in the figures by a small letter above the day. On the legend on the bottom of each graph, a letter is bolded in the name of each component indicating the letter used to represent a substitution. For example, a red 'N' indicates that nitrate was substituted, a green 'O' indicates that a substitution was made for organics, and a black 'L' indicates that elemental (or light-absorbing) carbon was substituted. Where more than one substitution has been performed, the letters are stacked above one another.

The first example in Figure 1 shows three years of data from Shenandoah National Park in Virginia. It is apparent

from the grouping of the stacked bars that the majority of each year's haziest days at Shenandoah National Park occur during the summertime. The use of contrasting colors on the bars makes it easy to get a qualitative feel for the relative contributions of the different mass components to total light extinction. The abundance of yellow indicates that sulfates are by far the largest contributors to extinction. By including three successive years on one graph, year to year changes in mean extinction can be seen. It is also apparent from the many N's that appear above the bars that a problem existed in 1998 with nitrate data collection, resulting in many lost samples.

The second example in Figure 2 illustrates data from Badlands National Park in South Dakota. The data from this park appear quite differently from Shenandoah. Here the haziest days are distributed throughout the year. Sulfates and organics are the most important contributors to haze during the warmer months, but nitrates become highly significant contributors during the winter. Unlike Shenandoah, wintertime extinction values are often higher than in summer.

USING THE ANNOTATE FACILITY
The bar charts are generated using the GCHART procedure. The Annotate facility is used in three places to enhance the basic charts, providing crucial information about the data. To use the Annotate facility, a data set is created that contains instructions for creating graphical annotations. This data set is then referenced within a graphics procedure statement to generate the graphics elements.

CREATING THE CHART LEGEND
Below is a sample of the code used to create the Annotate data set LEGEND. The sample shown creates the gray bar and text that show the color assigned to coarse mass extinction. Similar code is used for each bar in the legend.

The variables X and Y provide the coordinates for the annotations. These are fixed for the first bar drawn; subsequent bars are referenced from the previous bar. The internal variables XLAST and YLAST retain the last X and Y coordinates used in generating non-text graphics for later use by other non-text graphics functions. Similarly, the internal variables XLASTT and YLASTT retain the last X and Y coordinates used in generating text graphics. Specifying the function MOVE updates XLAST and YLAST to the new X and Y locations given in that data line without generating graphics. The value of the FUNCTION variable specifies the annotation to be created. The value BAR generates a bar drawn from XLAST, YLAST to the current X, Y values, and the value LABEL places text at the specified X, Y coordinates. Using the FUNCTION value CNTL2TXT copies the XLAST and YLAST locations to XLASTT and YLASTT without generating graphics, and thus the location for the text output 'C' can be given relative to the end of the gray bar.

The values of the variables XSYS and YSYS specify the type of coordinate system used by the X and Y values of the annotations. The value '1' specifies an absolute coordinate system where the X and Y values are percentages of the data area, and the value '7' specifies a

relative coordinate system where units are percentages of the data area. This allows the X-axis parameters of the gray bar to fixed, and the X-axis location of the text description to be be placed relative to the end of the bar.

```
data legend;
length function color style $ 8 text
$ 12;
retain ysys hsys '4' line 0;
/*------------------------------*/
/* draw coarse mass bar and text */
/*------------------------------*/
function='move'; x=10; y=0.7;
xsys='1';
output;
function='bar'; x=14; y=1.4;
color='gray80'; style='solid';
xsys='1';
output;
function='cntl2txt';
output;
function='label'; text='C';
color='gray80'; style='swissb';
size=0.61; position='0'; y=0.7;
xsys='7'; x=+0.8;
output;
function='label'; text='oarse Mass';
color='black'; style='swissl';
size=0.61; x=.; y=0.7; output;
```

The data set LEGEND is referenced in a GSLIDE procedure statement to create the legend. The bar charts are later replayed three at a time in a GREPLAY procedure statement after all individual charts are created along with the output of PROC GSLIDE. Letters are highlighted in the names of the mass types in the legend. These indicate the letter used to flag a data substitution. For example, the 'i' in the word 'Soil' is displayed in bold brown; a brown 'I' above a bar indicates a substitution for soil was performed on that particular day.

DRAWING LINES TO DEFINE THE SEASONS
The Annotate facility is also used to create the vertical lines indicating the boundaries between the calendar quarters. These lines make it easier to visually detect seasonal patterns in the distribution of the data. An Annotate data set is created for each year of data from the chart data set that contains the extinction values (TEMPEXT2). The data sets are named LINEYYYY, where YYYY is the year (e.g., 1999). The value of the variable XSYS specifies that the x-axis coordinate system units are data values. The internal variable XLAST is updated by using a MOVE function with a MIDPOINT statement to retrieve the x-axis location of the current bar. A small amount is then added to the variable X to move the x-axis location just to the right of the bar.

The value of the YSYS variable indicates that the coordinate system units for the y-axis are percentages of the data area. When the last day of March, June, or September is encountered, a data line is added to draw a line from the bottom to the top of the display area by specifying a starting y-value of 0 and an ending y-value of 100. The function DRAW is used to generate the line, and the variable LINE indicates the type of line (dashed). This procedure allows the lines to be generated dynamically for each year of data to be plotted. The code below illustrates the creation of the Annotate data sets.

```
data line&j;
format midpoint mmddyy10.;
length color $ 8;
retain when 'b' line 14 size 0.2 ysys
'1' hsys '2';
set tempext2;
by site year month date;
lastm=last.month; lastd=last.date;
color='white';
midpoint=date; subgroup=type;
function='move'; xsys='2'; y=0;
output;
function='move'; xsys='A'; x=+0.4;
output;
if month in (3 6 9) then do;
  if lastm=1 and lastd=1 then do;
  function='draw'; color='gray40'; x=+0;
  y=100; output; end;
end; run;
```

LABELING DATA SUBSTITUTIONS

The Annotate data set BARLABEL is used to identify days where data substitution was performed. During initial processing of the extinction data set, a flag is created for each mass type and set to 1 on each day that substitution for that mass type is performed. BARLABEL is created from the chart data set (TEMPEXT2) containing the extinction values. When a flag indicating a substitution is encountered, variables are added to the data set with instructions for placing the corresponding letter above the bar. The variables X and Y provide the coordinates for the annotations; the values of the variables XSYS and YSYS specify that the coordinate system units are data values. The annotations can thus be automatically referenced to the locations of the stacked bars. The x-axis location is specified by the location of each bar, in this case the date, using a MIDPOINT statement. The y-axis location of the first letter to be placed above each bar is created in an earlier data step, and set to the value of the height of the bar +1; successive letters (if necessary) are positioned above the previous letter by adding an incremental value to the y-coordinate (y=y+add). The variable FUNCTION specifies that the annotation to be created is a text label, and the variable TEXT specifies the text for the label. The data set BARLABEL is later referenced from the PROC CHART statement to produce the annotations. The code below shows how the Annotate data set is created.

```
data barlabel;
format midpoint mmddyy10.;
length color style $ 8;
retain  when 'a' style 'duplex' xsys
ysys '2' position '2' size 0.8 hsys '4';
set tempext2; by site date;
first=first.date;
if first=1 then add=0;
midpoint=date;
subgroup=type; text=' '; color='black';
function='label';
if type='ECM' and cmmiss=1 then do;
  text='C'; y=y+add; add+0.05*high;
  color='gray40'; end;
if type='ENO3' and no3miss=1 then do;
  text='N'; y=y+add; add+0.05*high;
  color='red'; end;
if type='EOC' and ocmiss=1 then do;
  text='O'; y=y+add; add+0.05*high;
  color='green'; end;
```

```
if type='ESOOT' and lacmiss=1 then do;
  text='L'; y=y+add; add+0.05*high; end;
if type='ESO4' and so4miss=1 then do;
  text='S'; y=y+add; add+0.05*high;
  color='gold'; end;
if type='ESOIL' and soilmiss=1 then do;
  text='I'; y=y+add; add+0.05*high;
  color='lirbr'; end;
run;
```

CONCLUSION
The flexibility of the Annotate facility makes it possible to create many types of custom annotations that greatly enhance basic charts. By creating a program to produce these charts, the process can be automated, allowing for the rapid mass production of hundreds of detailed charts. The charts created by this program will greatly assist state air quality planners nationwide in their efforts to reduce atmospheric haze.

REFERENCES
Malm W. C. (1999), *Introduction to Visibility*, Cooperative Institute for Research in the Atmosphere (CIRA), Colorado State University, Fort Collins, Colorado.

SAS Institute, Inc. (1999), *SAS/GRAPH Software: Reference, Version 8,* Cary, NC: SAS Institute Inc.

Sisler J. F., Malm W. C. and Gebhart K. A. (1996), *Spatial and Seasonal Patterns and Long Term Variability of the Composition of the Haze in the United States: An Analysis of Data from the IMPROVE Network*, Report by Cooperative Institute for Research in the Atmosphere (CIRA), Colorado State University, Fort Collins, Colorado.

U.S. Environmental Protection Agency Office of Air Quality Planning and Standards (2001), *Guidance for Tracking Progress Under the Regional Haze Program*, Draft.

ACKNOWLEDGMENTS

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author at:
    Debbie Miller
    National Park Service Air Resources Division
    P.O. Box 25287
    Denver, CO 80225-0287
    Work Phone: (303) 987-6947
    Fax: (303) 969-2822
    Email: debbie_c_miller@nps.gov

Figure 1. Reconstructed Extinction of 20% Worst Days--Missing Values Replaced With 10th Percentile
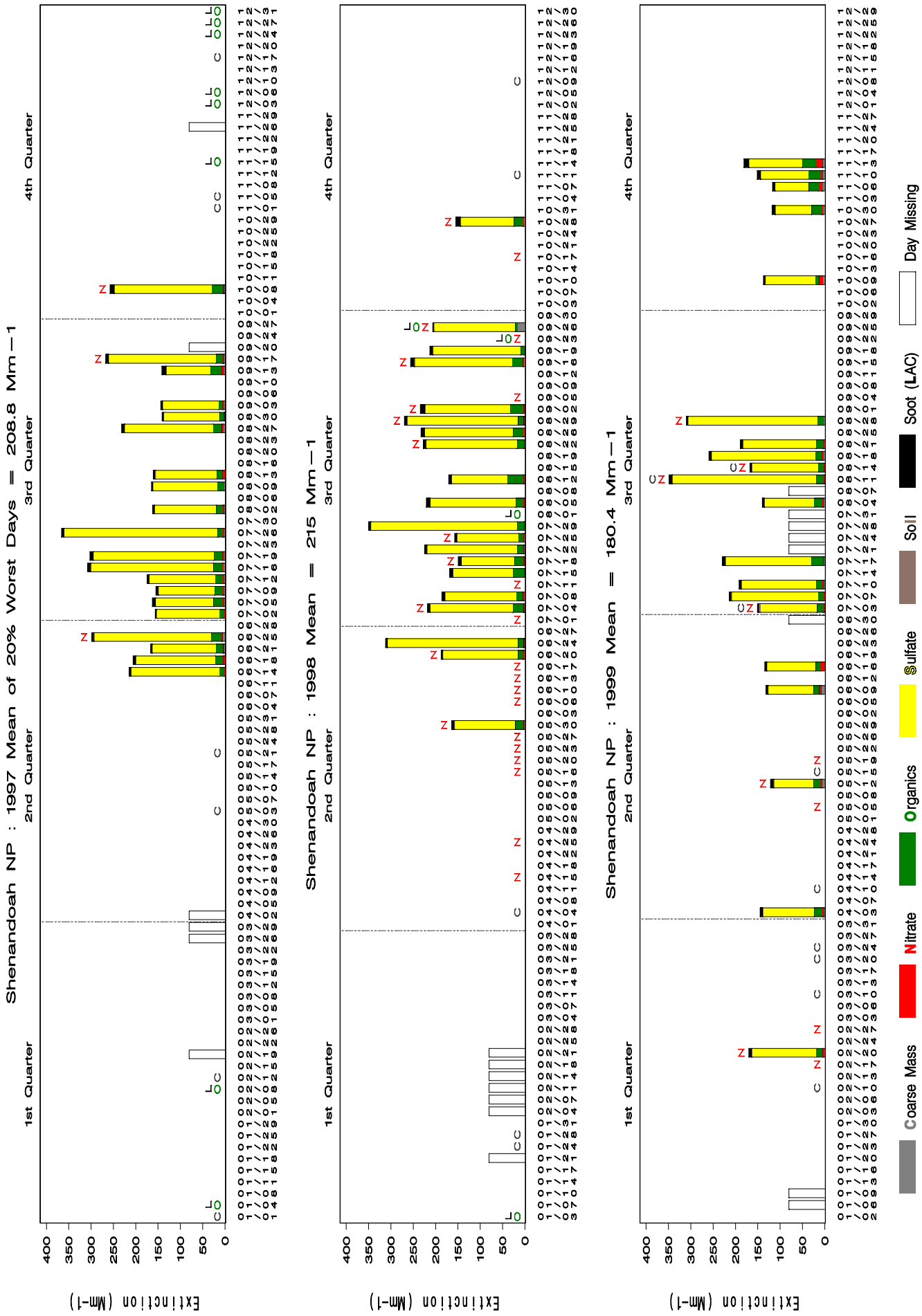
Figure 2. Reconstructed Extinction of 20% Worst Days--Missing Values Replaced With 10th Percentile