

SAS/GIS* and SAS/SPECTRAVIEW* for Visualization-Driven Analysis: An Application Supporting Ambient Ozone Control Policy

Thomas W. Sager¹, University of Texas at Austin, Austin, Texas

M. W. Hemphill, Texas Natural Resource Conservation Commission, Austin, Texas

ABSTRACT

The national standard for ground-level ozone pollution is currently topical. The current EPA standard for ozone is set in terms of the maximum of hourly levels throughout a day. A revision of the standard to a moving average of hourly levels is under consideration. Neither standard nor revision specifically considers population exposures or the effect of weather and other mitigating/aggravating factors. SAS/GIS and SAS/SPECTRAVIEW permit the display of relevant spatial-temporal maps of ground-level ozone concentrations that can inform the judgment of policy-makers, with the support of SAS analytical tools available for deeper probing. The integration of visualization tools with powerful statistical analysis makes possible a new approach to analysis of spatially/temporally distributed variables that we have elsewhere termed "visualization-driven analysis." We illustrate the application of this approach to modeling ambient ozone pollution. Maps of the geographic distribution of ozone concentrations can be displayed over time, based on either the current or revised standard, and adjusted for weather and other factors. The ozone measurements data can be combined with U.S. Census Bureau data to produce map displays of population exposure over time. The application of these tools provides an effective, scientifically-based approach for the development of control policy.

1. INTRODUCTION

As an article of faith, popular culture now accords to computers a role of chief facilitator of change. Among the pervasive influences of computers that are just now reaching general public awareness is the role of computers in the burgeoning area of geographic analysis.

"Geographic information system (GIS) computer programs that manipulate and analyze spatial data represent one of the hottest growth areas in the world of computing. Depending on what is included in the category 'GIS software,' the market is now worth between \$550 million and \$1 billion, and is growing at 15-24% every year."¹

This quote, from a feature article in a respected scholarly journal, indicates a growth rate that would double the GIS software market in 3-4 years! Clearly, the demand from both scientific and commercial users of GIS software is exploding. With the growth in demand, software publishers are rushing to compete by offering new products or by adding products to flesh out existing lines. With the release in late 1995 of SAS/GIS and SAS/SPECTRAVIEW, SAS Institute has offered tools for geographic analysis that compete on traditional terms with long-established players such as ESRI, with its market leading ARC/INFO* software, but that also offer the potential for raising the GIS software bar by integrating SAS/GIS and SAS/SPECTRAVIEW with the rest of the SAS* system. Through SAS/GIS, one can invoke the data analytic and data management capabilities, in which SAS has traditionally excelled, to support geographic analysis.

At the recent SUGI 21 Conference, we presented our views of how these new computer tools should be applied.² We

¹ Reichardt, Tony (1996). "Environmental GIS: The World in a Computer", in *Environmental Science & Technology*, 30:8, 340A - 343A.

² Sager, Thomas W. and Hemphill, M. W. (1996). "Temporal and Geographical Statistical Modeling: A New Paradigm", in *Proceedings of the Twentyfirst*

proposed a new paradigm we called *visualization-driven analysis* (VDA), characterized by a high level of interaction between geographical displays and analyses of those displays. SAS/GIS should be an appropriate tool for implementation of our vision. In this paper, we examine some of the possibilities of this use of SAS/GIS and its integration with the rest of the SAS system through an illustrative demonstration. The demonstration addresses an important and timely environmental issue: control of ground-level ozone pollution in urban areas. Although the questions that drive these analyses are of fundamental national interest and importance, we are not here focused upon providing answers. Rather, our intent is to illustrate software tools, so we avoid more than cursory interpretation of the products of our demonstration. Moreover, the inherently static constraints of print media preclude conveyance of a feel for the dynamic possibilities of a computer demonstration. So, for the most part, we include only samples of final products.

Despite differences, GIS software has a common feature - the ability to integrate and manipulate spatial data. When linked to displays over time, GIS computer programs provide the necessary tools for spatial-temporal data analysis. Map data can be divided into two categories: spatial data and attribute data. Spatial data consist of the coordinates of geographic features, such as monitoring stations, roads, and county boundaries, which are examples of zero, one, and two-dimensional spatial features, respectively. Spatial maps in SAS/GIS are built up in layers, for which a set of transparencies is an apt metaphor. One transparency may show county boundaries; another roads; and a third locations of monitoring stations. By overlaying the transparencies, one can build up a map, layer by layer, turning on the features that one wishes to view by including appropriate layers, and turning off other layers (removing transparencies). Attribute data are geographically varying data, such as ozone concentrations, associated with geographic features. When spatial and attribute data are

linked, one can produce choropleth or isopleth maps that show by color or contour where pollution is more severe. These capabilities are standard in GIS packages. What is now new is the ability to perform sophisticated statistical modeling and analysis of maps through linkages to tools such as SAS/STAT. Most GIS packages have at most limited statistical analysis capabilities, laboriously difficult to implement.

2. ILLUSTRATIVE DEMONSTRATION

Our illustrative demonstration concerns the use of SAS/GIS to support policy making and evaluation for control of ground-level ozone pollution. The discussion of the demonstration proceeds in three parts. First, we set forth some brief background that will facilitate appreciation for the crucial role of integrated statistical modeling with spatial-temporal data analysis in furthering this objective. Then we describe the implementation of relevant spatial and attribute data in SAS/GIS. Finally, we describe sample products that can be produced in this framework.

Air pollution from ground-level ozone adversely affects more than 7 million Texans - principally in major metropolitan areas. Further, ozone is a problem across our nation. The U.S. Environmental Protection Agency (EPA) has recently required Photochemical Assessment Monitoring Stations (PAMS) in areas where ground-level ozone is a problem, such as Houston, Texas. This monitoring provides concurrent measurements of:

- ozone,
- precursors of ozone (nitrogen oxides and volatile organic chemicals), and
- meteorology.

The overall design of PAMS should enable characterization of:

- precursor emission sources within the area,
- transport of ozone and its precursors into and out of the area, and

- photochemical processes related to high ozone measurements (episodes).

Specific EPA objectives for ozone that must be addressed include assessing time trends and determining spatial and diurnal variability.

The current regulatory standard for ozone is that no part of an urban area should exceed a concentration level of 120 parts per billion for any hour of a day more than three days in any three years. The EPA has announced it is considering revising this National Ambient Air Quality Standard (NAAQS) for ozone. One alternative being considered would use a moving average of hourly levels. Limitations of both data and technology have not allowed the expression of the NAAQS for ozone in terms of the number of people affected, weather, or other mitigating/aggravating factors. This could change with federal mandates for area-wide sampling, and subsequent analysis of the air quality measurement data considering critical weather data.

We believe GIS programs are necessary to respond to the EPA requirements. Both spatial and temporal analyses must be integrated to respond fully to EPA requirements. Thus, the analysis and interpretation of the PAMS data and the reconsideration of the ozone standard provide opportunities to evaluate our visualization-driven analysis paradigm through SAS software.

Within the context of the two issues outlined above, the questions which we address in this paper include:

- To what extent does SAS/GIS provide the spatial display capability needed for the analysis required by EPA?
- How do SAS graphical displays of data interact with SAS/GIS?
- What tools are available now within SAS/GIS to manipulate data spatially - and what additional tools are needed or wanted?

We now turn to describing the relevant spatial and attribute data.

2. MAP DEVELOPMENT

Houston, Texas occupies most of Harris County. Houston is appropriate for this demonstration, as it has an out-of-control ozone problem,³ a large long-standing network of ozone monitoring stations, and high quality spatial maps. In addition, we have a number of years of experience in analyzing Houston ozone data.

Spatial Data

SAS/GIS offers the ability to import spatial data from other formats such as ARC/INFO and SAS/GRAPH. Initially GIS map layers were imported from ARC/INFO coverages. These coverages were provided as ARC/INFO export files courtesy of the GIS Section of the Texas Natural Resource Conservation Commission TNRCC.⁴ ARC/INFO export files require no special processing prior to import into SAS/GIS, which automatically generates the necessary processing code (stored in the LOG window).⁵ Once imported, the converted ARC/INFO coverages are ready for display and analysis. Figure 1 displays three active SAS/GIS layers, illustrating zero, one, and two-dimensional geographic features that were imported from ARC/INFO:

- 1) AIRMON: points representing several monitoring sites in Harris County where samples of air have been collected periodically;
- 2) HWYS: lines which spatially display major highways around Houston; and
- 3) CNTYBND: the boundary of Harris County.

Additional spatial data layers were available for Harris County, including displays of cities,

³ Houston is conceded to have the second most significant ozone pollution problem in the United States. Los Angeles is first.

⁴ The TNRCC is the State agency charged with administering pollution control programs.

⁵ For the demonstration, SAS for Windows, version 6.11, was used, running on a Pentium processor under Windows NT version 3.51. All programs and figures were run and produced in this environment. References in the text to software features are to this environment.

streams, lakes, and streets. The latter have all been turned off for Figure 1 by making their layers inactive (by clicking their buttons in the map view window).

Attribute Data

The Statistical Sciences Team of the monitoring Operations Division of TNRCC has developed a SAS database called MOTHER (Monitoring Operations Tiered Hierarchical Environmental Records) that archives historical ambient air pollutant and meteorology data collected throughout Texas on a generally hourly basis from 1970 through the present. For the current paper, hourly ozone averages for nine monitoring sites in Harris County were extracted as SAS data sets from MOTHER. In addition, TNRCC supplied five-minute ozone measurements from other sources for selected high ozone episodes in Microsoft Excel® format. The Excel files were converted to SAS data sets for our analyses.

Episode Analysis

Regulatory authorities have identified certain time periods as episodes of high ozone. These periods are subject to intense analysis for determination of factors that can be modified to limit future episodes. Measurements as frequent as every five minutes are archived for episodes. One such episode occurred in Harris County during August 18 through 21, 1993. The highest ozone measurement in this episode, 231 ppb, occurred at 4:00 p.m. on August 19, 1993 at a monitoring site on Aldine Drive in north Houston. Several of our GIS-related analyses focus on this occurrence.

Trend Analysis

Historical data from MOTHER were used to analyze ozone trends. Ozone data for up to 9 monitoring sites operating between 1980 and 1995 were used for display and analysis of long-term trends. Ozone and certain weather variables, such as temperature, wind speed and wind direction, are archived at hourly intervals for each monitoring station. In order to link such attribute data with the spatial data, there must be a variable in common between the spatial data and the attribute data that is

unique to each spatial feature with which one wishes to link an attribute value. For example, we added an identifier variable SITECODE to the AIRMON layer in Figure 1. Each AIRMON point is a monitoring station that can be identified by a unique code in a national coding system such as the AIRS or SAROAD codes, or simply by a home-spun code like "HAL" for the "Houston, Aldine Drive" site. In the SAS dataset containing the attribute data, we also created a SITECODE variable which had the value HAL for the observation having the data for the Aldine site, and different codes for the other sites. By this means, ozone values at a given time, representing a geographic section of the spatial-temporal data could be associated with the spatial map. Although the imported ARC/INFO map data are expressed in coordinates of latitude and longitude, it is not necessary that the attribute data include latitude and longitude, or any other coordinate system. What is necessary is that there be a linking variable like SITECODE above in both the spatial and attribute data.

Each layer of a map will be of one and only one type: point, line, or area. Each type can display linked attribute data, but only in a manner appropriate to its type. For example, in the manner described in the preceding paragraph, we linked ozone measurements to the monitoring stations that produced them. We envision making the size or color of the points proportional to their associated ozone levels. Because these ozone data are linked to points, the data cannot directly be used for a choropleth map, with different colors and/or shading representing different ozone concentration levels. The latter application requires linking the ozone data to an area layer. This requires defining a new layer, with geographic coordinates describing the boundaries of the regions to be colored/shaded differently. One solution with SAS software is to interpolate the monitoring station data to a fine rectangular latitude-longitude grid over the Houston region by PROC GCONTOUR, use a DATA step to turn the grid into an input map dataset in format suitable for PROC GMAP, but instead import the map dataset into SAS/GIS as a new layer, group the grid cells by ozone levels and link the resulting SAS dataset to the new layer by associating ozone level with

unique grid cell identifier. The result will be a choropleth map in SAS/GIS. However, the area boundaries, which are the rectangular perimeters of all of the grid cells are visible and there is no way to turn the area boundary display off in SAS version 6.11 (a coming enhancement for version 6.12). Consequently, we have not included an example of this sort of display. No doubt more satisfactory solutions are possible within SAS, at the cost of greater complexity.

The main linkage between SAS/GIS and the rest of the SAS system is provided through the capability of performing *actions* for selected map features.⁶ An action is defined by selecting View → Actions → Define from the menu bar in the Map window. For example, with the ozone levels linked to the AIRMON layer, as above, we can define a new action by selecting Browse under the Type field in the GIS Actions Definition window. We select the SITECODE variable in both the Spatial Vars and Data Set Vars. Then whenever air monitoring sites are selected on the map, an FSBROWSE window will open to display the ozone and other values in the associated attribute SAS dataset. For our purposes, the Program action proves most useful. It allows the execution of complete SAS programs against selected map features. We now discuss a few such analyses that could be executed in this manner. Although interpretations of the data are not a central issue in this paper, we use our understandings of the general features of the data to judge our satisfaction with the applications of the tools we develop.

3. SELECTED PRODUCTS OF THE ILLUSTRATIVE VDA DEMONSTRATION

Ground-level urban ozone typically exhibits an annual cycle. A typical example is shown in Figure 2, for the monitoring station located on Aldine Drive in Houston, Texas. Hourly data are available for this station from 1980-95. To

reduce plot clutter, we condense the 24 hours/day * 16 years = 384 potential measurements for a given day of the year into a single point (excluding Feb 29) by first finding the daily maximum ozone value and then averaging the 16 daily maxima. The fitted curve comes from a regression involving 4th-degree polynomials and trigonometric functions produced by PROC GLM, whereas the plot itself is PROC GPLOT output. We observe that higher ozone values appear in July and August, whereas lower values generally occur in colder months. Trend analyses mandated by EPA should account for this annual ozone cycle in the statistical modeling of this site.

Figure 3 resulted from a combination of PROC GLM to deseasonalize the historical ozone data at the Aldine site (discussed in Figure 2) and PROC AUTOREG to fit a time series model to the residuals from PROC GLM. Figure 3 attempts to portray the potential difference between ground-level ozone with controls and without controls. The solid line in Figure 3 represents the observed trend of the annual mean of the daily maximum ozone concentrations at Aldine as actually measured. The dashed line represents an estimate of what the trend might have been without controls on ozone. The dashed line, with higher ozone values, is obtained from a time series regression of seasonally-adjusted ozone on important weather predictor variables, including the maximum daily temperature, mean wind speed, proportion of time the wind is from various directions, and on a time trend variable. The coefficient of the trend variable is negative, indicating that after controlling for the weather, the trend in ozone concentrations is down at Aldine. Since weather is presumably the most important set of natural factors leading to ozone, what remains is the effect of humans (and possibly some weather factors of lesser importance).

PROC GCONTOUR yielded Figure 4. The hourly ozone measurements from nine Houston monitoring sites were obtained for 4:00 - 5:00 p.m. on August 19, 1993. Default interpolations of the nine values to a 200-by-200 grid for the Houston region were provided by PROC G3GRID as input to PROC GCONTOUR. The ozone gradients increase

⁶ For details, see chapter 4 of *SAS/GIS Software: Usage and Reference*, version 6, First Edition, (1995), SAS Institute Inc., Cary, N.C.

from lower right (southeast) to upper left (northwest). Lows below 40 ppb are estimated within the lowest contour. The highest measured value (231 ppb) occurred at the Aldine Drive site within the top contour. These contours can be spatially related to our GIS map and shown for a portion of Harris County (including Houston). The development of area analyses is too new to comment of whether this display is typical of ozone episodes. The spatial representation of ozone data is designed to assist in characterizing precursor emission sources as required by EPA.

Ozone is not emitted directly from sources. Instead, ozone is formed indirectly by chemical reactions among air pollutants in the presence of sunlight. Precursor pollutants that lead to the formation of ozone include nitrogen oxides (NO_x) and volatile organic compounds (VOCs). Figure 5 displays a time trend for ozone, NO_x, and temperature (a proxy for the intensity of sunlight) for one monitoring site during an episode of extremely high ozone on August 19, 1993. The level of time detail shown is not usually available for most historical data, as five-minute data are not ordinarily archived for more than a few months. Instead, hourly data are preserved, which are averages of all the data collected within each hour. The superposition of the ozone step function on the five-minute ozone data reveals the sort of detail that is lost in condensing the five-minute data for archiving. The shapes of the NO_x and ozone traces are typical for the region and season of the year. We observe that the peak value for NO_x occurs early in the morning, whereas ozone peaks in the late afternoon. This display would be useful in developing insights for modeling (1) to support forecasting of episode days, (2) to assist in estimating missing values, and (3) to incorporate meteorological and VOC data as sources of variation. This PROC GPLOT output overlays four different data sources that have been merged in time order, with the PLOT2 feature to provide the right-hand vertical scale for NO_x and temperature.

4. FUTURE APPLICATIONS OF SPATIAL-TEMPORAL ANALYSIS

Early standards for ground-level air pollution were based on point monitoring. More recently EPA has expressed interest in considering area-wide population exposure for interpretations of air quality measurements. It may be reasonable to extend our statistical models and use area-wide (spatial) analyses to decide whether an area is in attainment of a mandated standard. The expression of our standard would then be in terms of a model – adaptable to different geographic areas and able to incorporate new information, whether the information deals with control technology or effects. The development of a model-based standard begins with creation of new technological approaches for spatial-temporal analysis of data. The SAS System for Information Delivery, with SAS/GIS, SAS/SPECTRAVIEW, and other components, offers a strong base for what we refer to as *visualization-driven analysis*, or VDA.

REFERENCES

- Sager, Thomas W. and Hemphill, M. W. (1996). "Temporal and Geographical Statistical Modeling: A New Paradigm", in *Proceedings of the Twentyfirst Annual SAS Users Group International Conference*, Cary, NC: SAS Institute, Inc., 1327-35.
- Reichardt, Tony (1996). "Environmental GIS: The World in a Computer", in *Environmental Science & Technology*, 30:8, 340A - 343A.
- SAS Institute Inc.(1995). *SAS/GIS Software: Usage and Reference*, Version 6, First Edition, Cary, NC: SAS Institute Inc., 273 pp.

ACKNOWLEDGMENTS

The authors express their appreciation to the following for their help and support: Tom Winn, SAS Technical Services (Cary, N.C.), SAS Austin Regional Office, Monitoring Operations Division and GIS Section of TNRCC, Center for Statistical Sciences at University of Texas, Environmental Monitoring and Analysis Program of EPA. Responsibility for the content of this article is solely that of the authors.

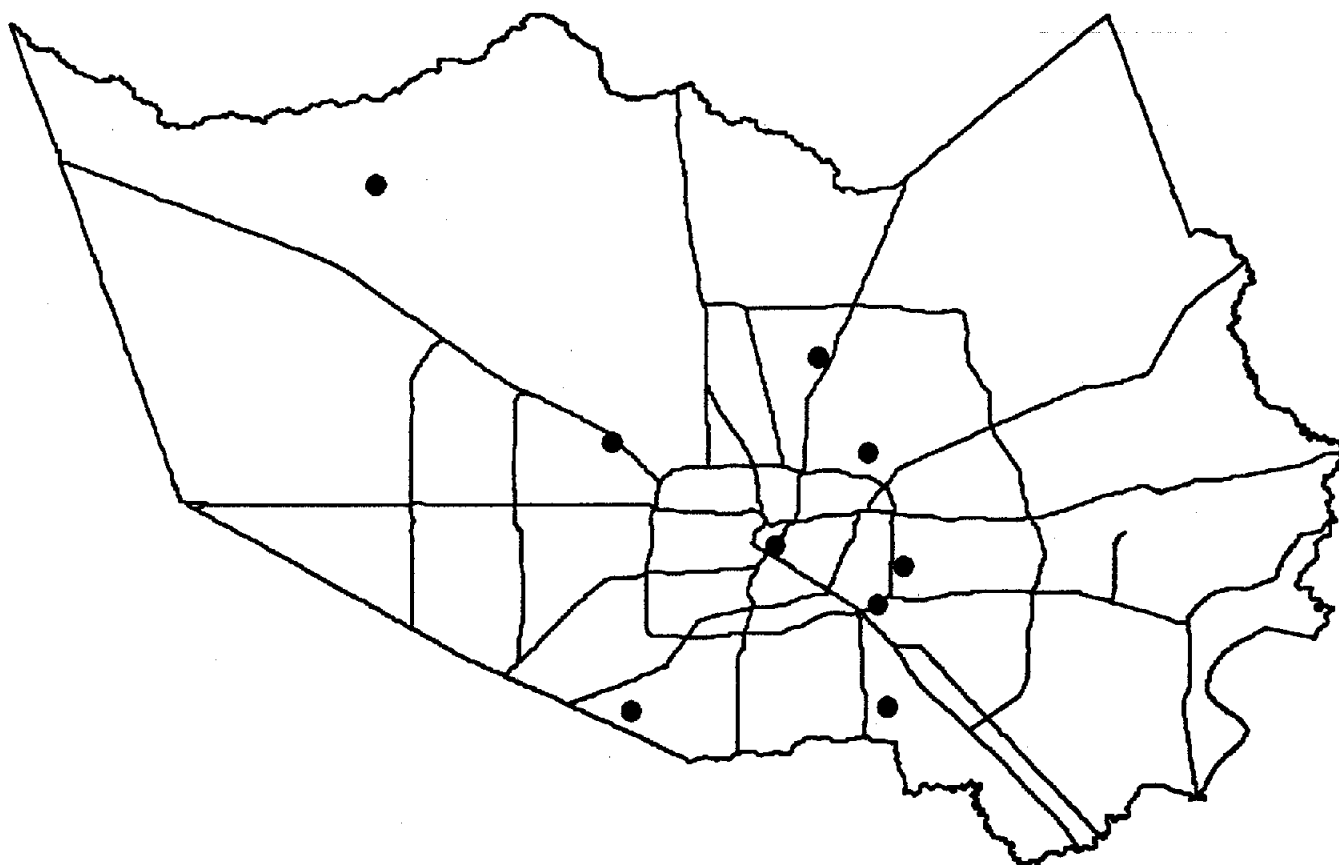
SAS, SAS/GIS, SAS/SPECTRAVIEW, and SAS/STAT software are registered trademarks or trademarks of SAS Institute, Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Contact author:

Dr. Thomas W. Sager
Department of MSIS, CBA 5.202
University of Texas at Austin
Austin, TX 78712
(512) 471-3322
tomsager@mail.utexas.edu

¹ Research related to this article supported by EPA under cooperative agreement no. 821801011. The article has not been subject to the review of the EPA and thus does not necessarily reflect the view of the agency and no official endorsement should be inferred.



Harris County – 9 Ozone CAMS – 8/19/93

● AIRMON

— HWYS

■ CNTYBND

Figure 1

Annual OZONE Cycle

Quartic-Trig Trend of Mean Daily LOG(Max OZONE)
Site HAL, 1980-95

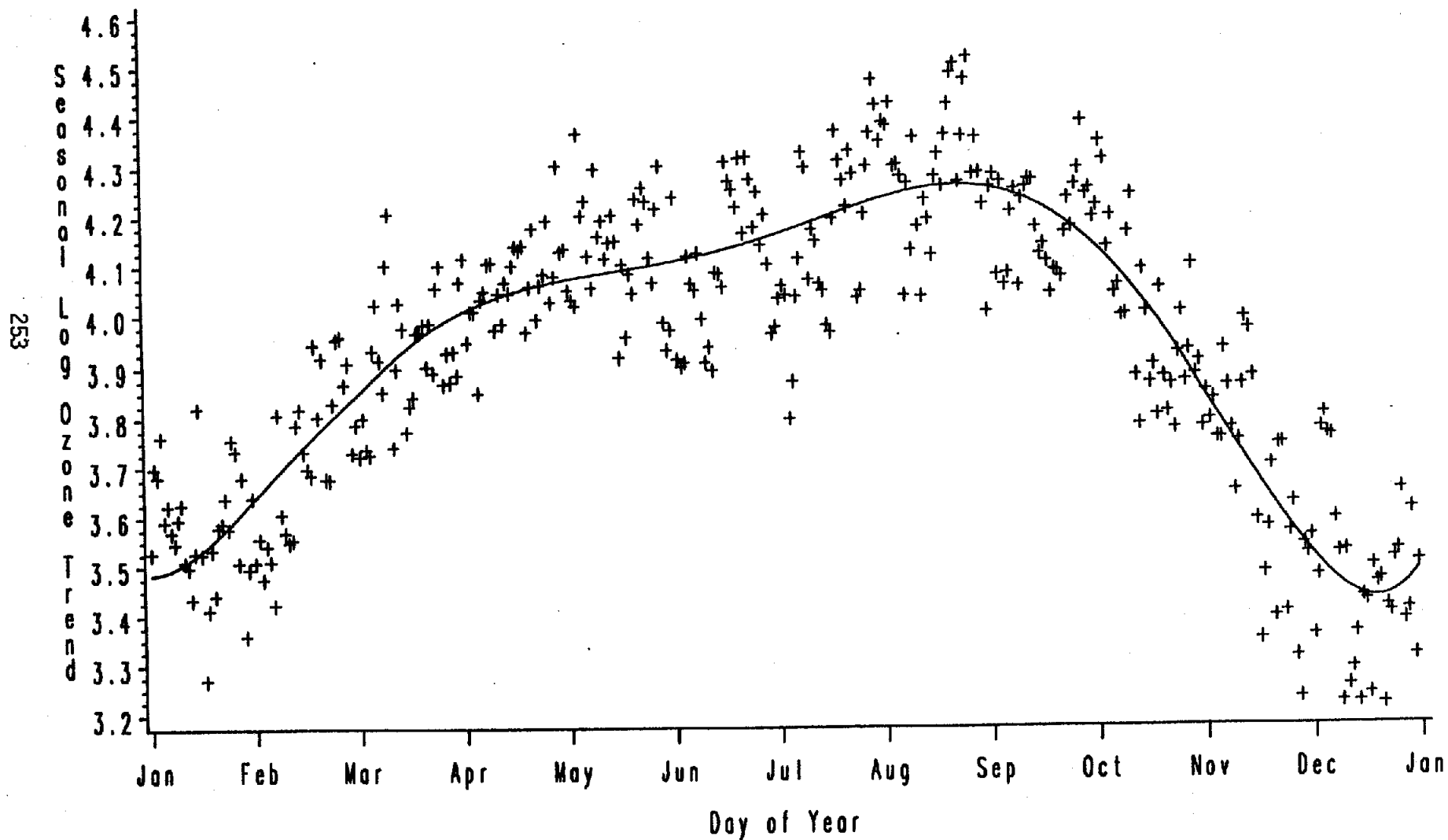


Figure 2

Potential Historical Impact of Ozone Control Strategies

Houston Aldine Site

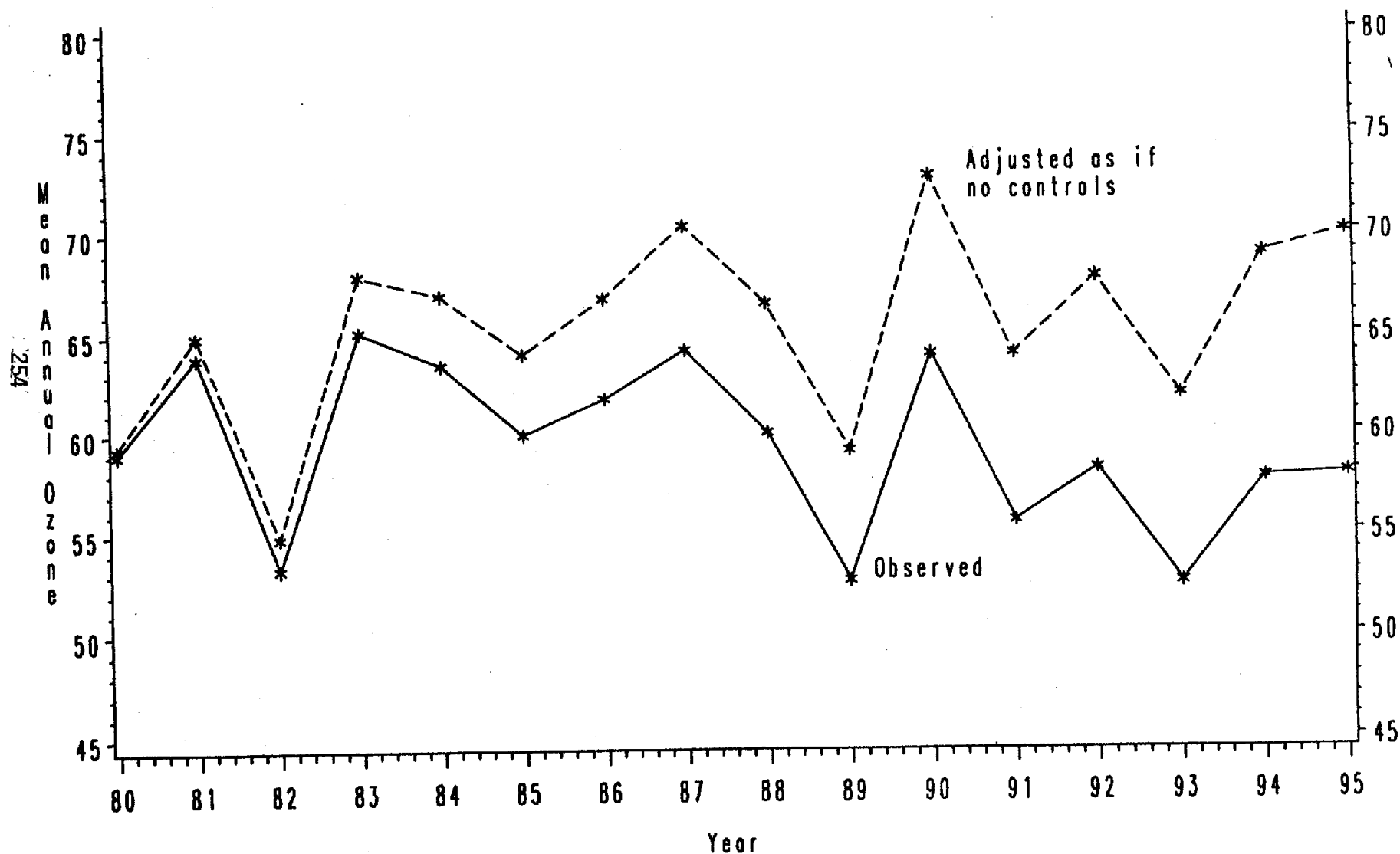


Figure 3

Houston Ozone 4:00pm August 19, 1993

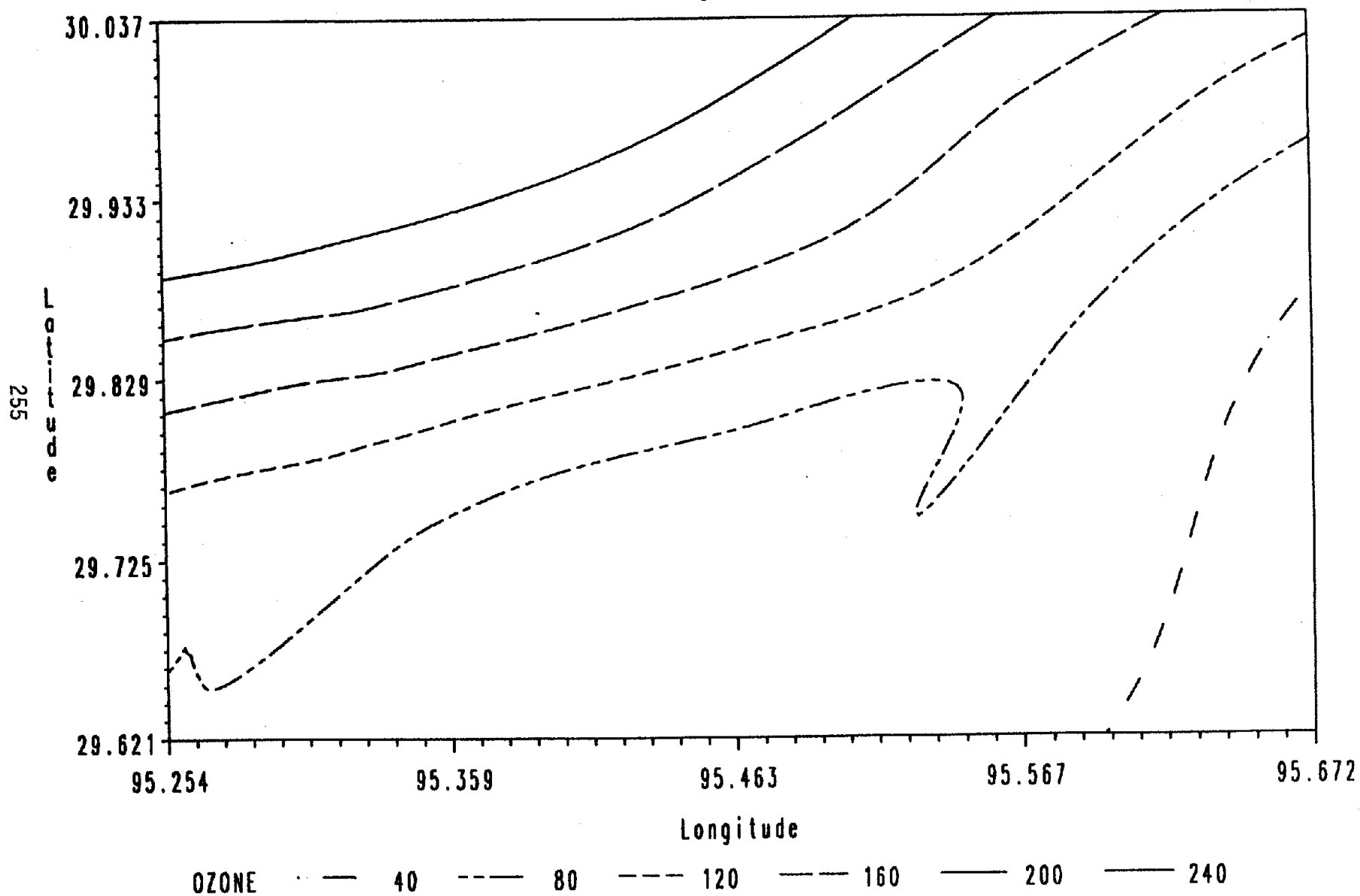


Figure 4

Five-Minute Ozone, Temperature, and NO_x at Houston Aldine

August 19, 1993
With Superimposed Hourly Ozone (Step Function)

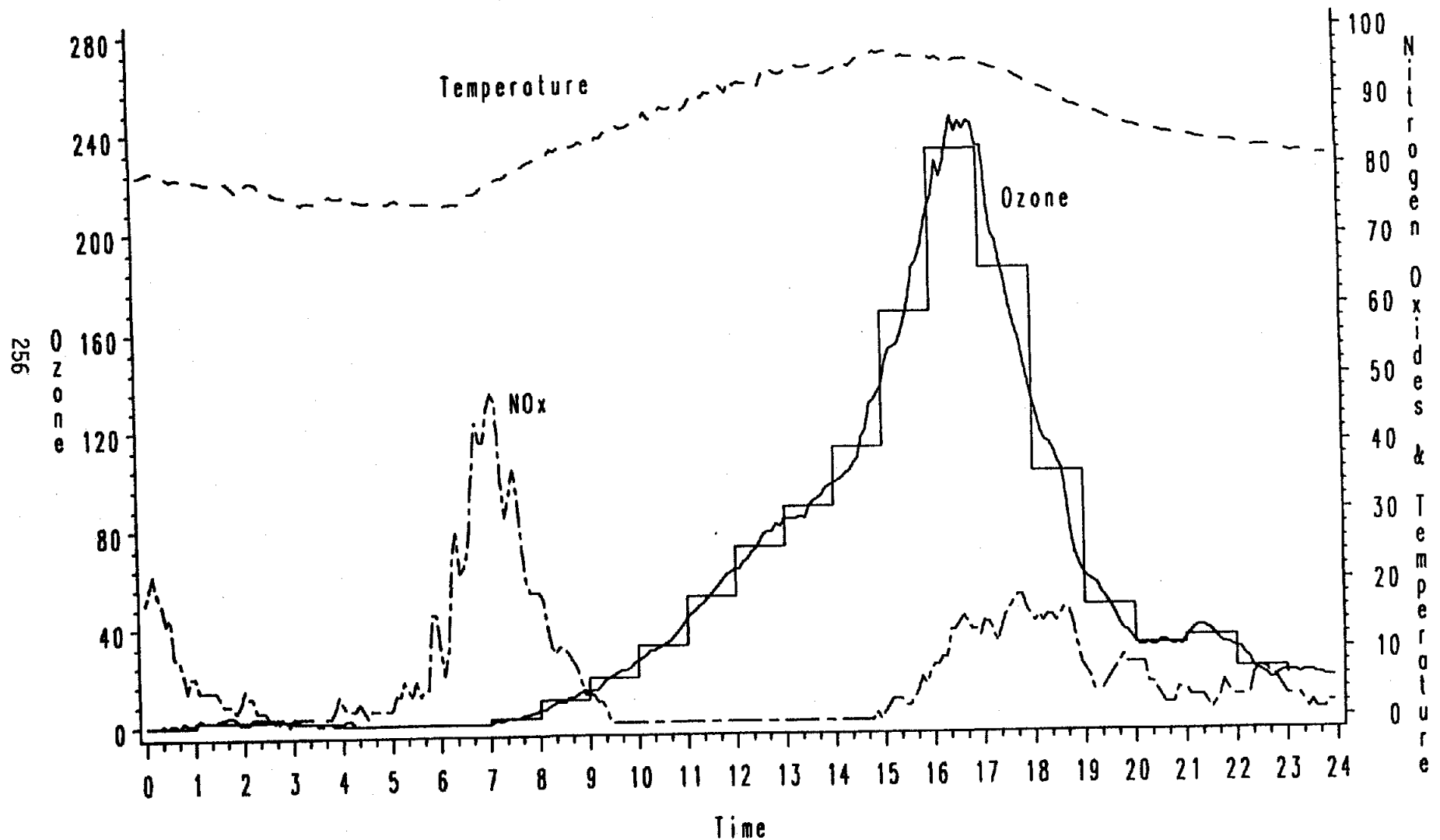


Figure 5