NISS

Case Study: Using the AIRS Database — Report to EPA/CEIS

Peter Bloomfield and Yuntae Kim

Technical Report Number 93
March, 1999

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Case Study: Using the AIRS Database

Report to EPA/CEIS

Peter Bloomfield and Yuntae Kim
National Institute of Statistical Science
and
Department of Statistics, North Carolina State University

September, 1998

Executive Summary

The Center for Environmental Information and Statistics (CEIS) of the U.S. Environmental Protection Agency (EPA) has prepared a draft "Protocol for Statistical Assessment of EPA Program Office Data Systems for Secondary Purposes", which in part states that

> The statistical assessment of Agency databases for secondary use purposes will be in two parts:
>
> 1) a quantitative description of the data in the data bases; and
>
> 2) a case study using a hypothetical secondary use (in most cases we select the secondary use of "Environmental Status Quality and Trends for a Local Community").

The National Institute of Statistical Science (NISS) has entered into a Cooperative Agreement with CEIS to assist with the assessment. NISS's first effort under the Agreement has been to carry out just such a case study, specifically of ozone status and trends for the Research Triangle area of North Carolina. The results of the

study are presented in the body of this report. In particular, Figure 7 on page 21 shows the trend for the area, adjusted for concomitant changes in meteorology.

In the course of the study we needed to identify sources of air quality and meteorological data, and obtain the data required for the analysis. This process, documented in some detail below, revealed both strengths and weaknesses in the resources that are currently readily accessible. Our entry point was AIRSWeb, a Worldwide Web site established by the Office of Air Quality Planning and Standards, the purpose of which is declared to be:

> AIRSWeb gives you access to air pollution data for the entire United States.

While this site does indeed provide various summaries of air quality data, it does not at present provide access to the original measurements. Through a chain of contacts begun at AIRSWeb and ending with local (non-EPA) experts, we were able to identify and obtain the ozone data that we needed.

Since ozone concentrations are strongly affected by prevailing meteorological conditions, we needed observations of various meteorological variables at relevant times and locations in order to assess the status and especially the trends in ozone. We were unable to identify a source for the required data using AIRSWeb, and turned instead to colleagues, who led us to a resource of the State of North Carolina where the data were acquired.

With ozone and meteorological data in hand, we were able to follow the pattern established in an earlier study to extract an estimate of the trend in ozone concentration, adjusted for concomitant changes in meteorology. Even with such a pattern to follow, informed judgment is required in making various choices, including:

- treatment of missing values;
- precise specification of model for dependence on meteorology;
- specification of the form of the trend.

The association of one of the authors with the earlier study provided the background to make reasonable choices, but this would have been difficult for someone approaching such a study with less background.

Our principal finding is that *it is not possible at this time to identify and retrieve the data required for a study of this kind without assistance from experienced personnel.* In particular, we needed guidance in:

- identifying monitoring locations that were relevant to a specific Metropolitan Statistical Area (MSA);
- retrieving the identified time series data;
- obtaining the meteorological data needed to interpret and adjust the observed trends in the ozone measurements.

We also needed to draw on our own experience in analyzing air quality data to make the choices listed above.

The difficulties that we encountered highlight some of the areas where access to information could be improved. Some deficiencies seem to be readily rectified: adding a geographical search engine to the data retrieval process, incorporating original measurements, and so on. Others may be more difficult: providing automated expert assistance, or eliciting the purpose of a study in order to provide links to relevant parts of the literature.

Lessons Learned : Comments on CEIS Review Questions

CEIS has formulated questions that a typical secondary user might want to be answered. In carrying out the NISS case study, we have needed to address questions similar to many of these. The questions are listed below, with our comments on how easy or difficult we believe it is to find answers.

The comments on all questions except 1 were made on the assumption that the hourly ozone data of Research triangle area would ultimately be available on the AIRSWeb site.

1. *How comprehensive is the database?*

   While the AIRSWeb site provides considerable information and various summaries about the data collected by EPA, it doesn't currently provide access to the raw data, the hourly ozone data of Research Triangle area. It will be more useful when the data can be obtained directly from the web site.

2. *Can the database be used for spatial analysis?*

Yes, since the AIRSWeb site provides the location information (longitude and latitude) of the ozone monitoring stations and currently permits listing of stations by county or by MSA.

However, in our case, discussions with a local (non-EPA) expert identified a nearby station where the ozone concentration measurements are relevant to the Research Triangle area, and several stations within the MSA that are not judged relevant. An additional query based on geographical radius would have allowed us to identify the nearby site. It is not clear how to incorporate the expert knowledge as to which locations provide the most relevant observations.

3. *Can the database be used for temporal analysis?*

   Yes, since the hourly ozone data were time-stamped, although the data format was not especially convenient.

   The varying amounts of data available in different years (ozone seasons) and the incompleteness of the data within years complicate the statistical analyses. These issues have been explored elsewhere (Bloomfield et al. [1993, 1996]), and we shall take advantage of their explorations in managing the problem in the context of the current exercise. It would be desirable to make other users of AIRSWeb aware of such information.

   Trends, a temporal analysis result, can be represented in many ways, including the simple linear trend used by Cox and Chu [1993], and the natural cubic spline function used in this study. Choosing among these must be based on the observed nature of the data and the use to which the trend is to be put. Since the same issue would arise in other trend analyses, it would similarly be desirable to make other users of AIRSWeb aware of the issue.

4. *How consistent are the variables over space and time?*

   This study was based on data covering several years but at locations in only a small geographical region. No inconsistencies were found.

5. *Can data be linked with information from other databases?*

   In our study, linkage to meteorological data was essential, but it is not provided on AIRSWeb site at present. The present exercise has been delayed by the need for meteorological observations to accompany the ozone concentration observations. While AIRSWeb provides access to some such data,

they were not available for the area of our study. More complete meteorological data, or links to other databases, are essential for the data that *are* in AIRS to be of most use.

The study that established the pattern being followed here was carried out at a time when National Ambient Air Quality Standard (NAAQS) was based on daily maximum 1-hour concentrations, and consequently the analysis was all based on such maxima. The current NAAQS is based on the maximum 8-hour average concentration, which was the daily summary used in the present study. The form of the NAAQS and the process that resulted in them are described in several places, and in particular are easily obtained over the web. There does not however seem to be a link from AIRSWeb to any of those locations.

6. *How accurate are the data?*

   The measurement of ozone was in integer values of PPM, which was accurate enough for our case study to analyze the relationship between ozone and meteorology factors.

7. *What are the limitations?*

   We needed guidance in:

   - identifying monitoring locations that were relevant to a specific Metropolitan Statistical Area (MSA);
   - retrieving the identified time series data;
   - obtaining the meteorological data needed to interpret and adjust the observed trends in the ozone measurements.

8. *How can I get information?*

   We were able to contact people with detailed knowledge of the data we sought with relatively few telephone calls and e-mail messages. In the chain of contacts for ozone data, the starting point was the email address of contact point on the AIRSWeb. For meteorological data we found no help on AIRSWeb, but colleagues led us to the State Climate Office.

9. *Is there documentation?*

   When we acquired the hourly ozone data for the Research Triangle area, we could get some explanation files, including variable description and station

information, along with the compressed raw data, through email from the local expert.

Contents

1 Introduction **8**

2 EPA Data **8**

3 Meteorological Data **15**

4 Trend Analysis **15**

5 Out-of-sample Validation **20**

6 Comparison with the model of Cox and Chu **23**

List of Tables

1 Summary information on eight ozone monitoring stations in the Research Triangle, NC, area. 13
2 Parameter estimates. 19
3 Validation results 22
4 Comparison with the model of Cox and Chu 24

List of Figures

1 AIRSWeb list of ozone monitoring stations in the Raleigh-Durham-Chapel Hill MSA. 10
2 AIRSWeb list continued. 11
3 Locations of ozone monitoring stations in the Research Triangle, NC, area. 12
4 Sample records from a file of ozone measurements. 13
5 Mean diurnal profiles of ozone concentrations at 8 monitoring stations in the Research Triangle, NC, area. 14
6 Availability of daily summaries of ozone concentrations. 16
7 Network average maximum 8-hour average ozone concentration and met-adjusted trend. 21
8 Seasonal component of fitted ozone model. 22

1 Introduction

CEIS has prepared a draft "Protocol for Statistical Assessment of EPA Program Office Data Systems for Secondary Purposes", which in part states that

> The statistical assessment of Agency databases for secondary use purposes will be in two parts:
>
> 1) a quantitative description of the data in the data bases; and
>
> 2) a case study using a hypothetical secondary use (in most cases we select the secondary use of "Environmental Status Quality and Trends for a Local Community").

In December, 1997, NISS initiated its own case study of one particular Agency database, namely the Aerometric Information Retrieval System (AIRS) database of air quality measurements.

A Research Assistant with no previous experience working with air quality data was assigned to carry out a follow-up study of an earlier NISS project (Bloomfield et al. [1993, 1996]) in which ozone measurements from the Chicago area were analyzed (referred to below as the "Chicago study"). The goal of the Chicago study was to model the relationship between surface ozone concentrations and meteorology, with a view to

- understanding the impact of meteorology on observed trends in ozone, and
- adjusting trend estimates for the effects of meteorology.

The goals of the current follow-up study are

- to obtain appropriate air quality and meteorological data for the Research Triangle area of North Carolina, and
- to carry out an analogous study of the relationship between surface ozone concentrations and meteorology for these data.

2 EPA Data

Initial contact with the EPA was made through the World Wide Web site

`http://www.epa.gov/airsweb/`

This site provides information about data made available by the EPA, including monitor locations. Following links successively to "Monitors", "Queries", and "Site" led to a form-based screen which allowed us to request ozone monitoring stations in the Metropolitan Statistical Area (MSA) of Raleigh-Durham-Chapel Hill (MSA 6640). The result of the query is shown in Figures 1 and 2.

AIRSWeb does not at present offer direct access to the data, but does provide an electronic mail contact for data access. An e-mail exchange with the contact led to the office of Air Quality in the Department of Environmental Health and Natural Resources of the State of North Carolina, and ultimately to Wayne Cornelius (a graduate of the Department of Statistics, NCSU). Discussions with Dr. Cornelius identified seven of these stations for which reasonably extensive ozone data are available and are thought to be relevant to exposures in the Research Triangle area, and an additional nearby station that is not on the list obtained through AIRSWeb. Locations of all stations are shown in Figure 3.

Dr. Cornelius sent by e-mail a number of compressed and encoded data files containing the available hourly ozone concentration measurements for the eight identified monitoring stations. The first five records from one file are shown in Figure 4. The interpretation of certain fields is as follows:

2–3: state code—"37" denotes North Carolina;

4–6: county code—"063" denotes Durham County;

7–10: site ID, unique within county—site "0013" in Durham County is at 2700 North Duke Street in Durham, NC;

11–15: parameter code—"44201" denotes ozone;

18–20: units—"007" denotes "parts per million by volume" (ppm);

24–29: date in "yymmdd" format;

30–31: hour of first observation in record—"00", "08", or "16";

32: decimal point locator—"3" means that 3 digits follow the decimal point;

33–36: first observation;

37: validity code for first observation—blank signifies valid data;

. . .

AIRSWeb
EPA Office of Air Quality Planning And Standards

Sites Report

13 Matches

For State of North Carolina And Parameter Code of 44201 And MSA Code of 6640
And Ordered by Columns site.state_code,site.county_code,site.site_id

Click a Column Heading for Description

Monitor	Subordinate Tables	State Code	County Code	Site ID	Street Address	Latitude	Longitude	City Code
Monitor	Subordinate Tables	37	037	0004	RT4 BOX62 PITTSBORO NC27312	35.758889	-79.165278	00000
Monitor	Subordinate Tables	37	037	0098	MONCURE PLANT - SOUTH SITE	35.615833	-79.045833	00000
Monitor	Subordinate Tables	37	063	0013	2700 NORTH DUKE STREET	36.035556	-78.904722	19000
Monitor	Subordinate Tables	37	063	1001	4340 E GREER ST.	36.061111	-78.775	00000
Monitor	Subordinate Tables	37	063	8001	ALEXANDER DR., N. OF HIGHWAY 54	35.9025	-78.87	19000
Monitor	Subordinate Tables	37	069	0001	431 S HILLSBOROUGH ST FRANKLINTON NC	36.0975	-78.463611	00000
Monitor	Subordinate Tables	37	101	0002	3411 JACK ROAD CLAYTON NC 27520	35.5	-78.4375	12860
Monitor	Subordinate Tables	37	101	0099	HIGHWAY 301 & SR 2141	35.569722	-78.185833	00000
Monitor	Subordinate Tables	37	183	0014	E MILLBROOK JR HI 3801 SPRING FOREST RD	35.856111	-78.575556	55000
Monitor	Subordinate Tables	37	183	0015	808 NORTH STATE STREET	35.788333	-78.622222	55000

Clicking on the word Monitor will return all the **Monitors** for this site.
Clicking on the word Subordinate Table will return all the **Tangent Street** data for this site.
Clicking on the County_code value will return the **County** data for this site.

From row #0 to row #10
Next 10
Create ASCII file of this query.

Sites Report
Go to EPA Home Page AIRSWeb Home Page

August 25, 1998 12:15 AM
Comments?

Figure 1: AIRSWeb list of ozone monitoring stations in the Raleigh-Durham-Chapel Hill MSA.

AIRSWeb
EPA Office of Air Quality Planning And Standards

Sites Report

13 Matches

For State of North Carolina And Parameter Code of 44201 And MSA Code of 6640
And Ordered by Columns site.state_code,site.county_code,site.site_id

Click a Column Heading for Description

Monitor	Subordinate Tables	State Code	County Code	Site ID	Street Address	Latitude	Longitude	City Code
Monitor	Subordinate Tables	37	183	0016	201 NORTH BROAD STREET	35.585	-78.794722	25300
Monitor	Subordinate Tables	37	183	0017	5033 TV TOWER RD GARNER NC 27529	35.683333	-78.55	25480
Monitor	Subordinate Tables	37	183	2001	HWY 98 WAKE FORREST WATER TREATMENT PLAN	35.970833	-78.490833	70540

Clicking on the word Monitor will return all the **Monitors** for this site.
Clicking on the word Subordinate Table will return all the **Tangent Street** data for this site.
Clicking on the County_code value will return the **County** data for this site.

Create ASCII file of this query.

Figure 2: AIRSWeb list of ozone monitoring stations in the Raleigh-Durham-Chapel Hill MSA (continued).

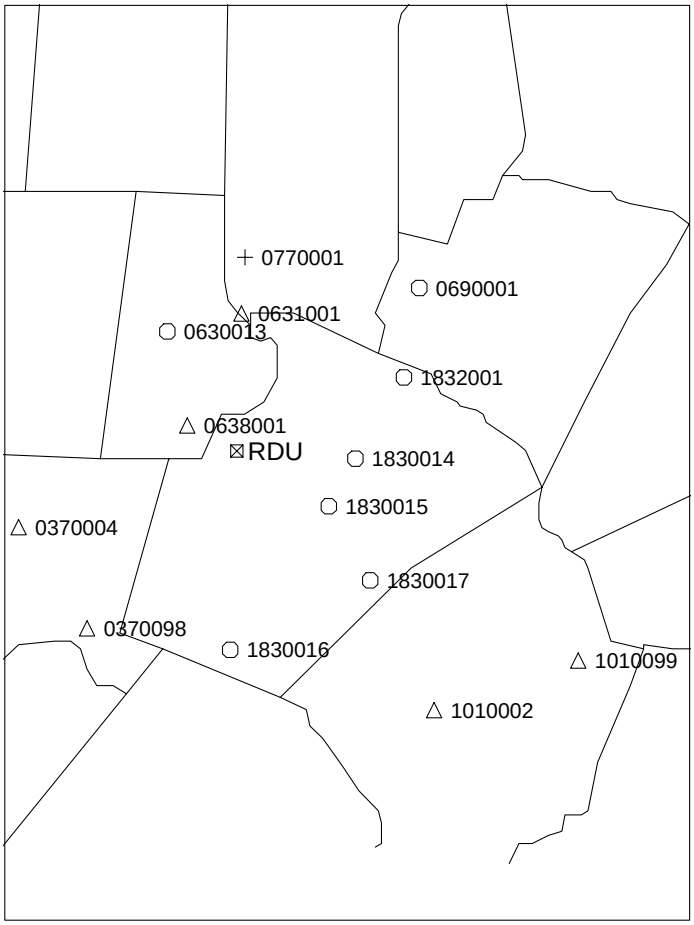

Figure 3: Locations of ozone monitoring stations in the Research Triangle, NC, area. “RDU” shows the location of RDU International Airport, the site of the National Weather Service station. Key: circles indicate stations identified through AIRSWeb for which adequate ozone data were available and deemed relevant; triangles indicate stations identified through AIRSWeb for which adequate ozone data were not available or were available and deemed not relevant; and plus indicates the one additional station.

```
1370630013442011100701993050400300130017 0019      0013 0010 0012 0011       I
1370630013442011100701993050408300140021 0025 0031 0035 0038 0037 0035       I
1370630013442011100701993050416300320031 0019 0018 0006 0006 0010 0015       I
1370630013442011100701993050500300170018 0018      0013 0005 0004 0003       I
1370630013442011100701993050508300060011 0016 0019 0016 0025 0030 0034       I
```

Figure 4: Sample records from a file of ozone measurements.

Table 1: Summary information on eight ozone monitoring stations in the Research Triangle, NC, area.

Station ID	Urban/Suburban/ Rural	Dates of First data	Dates of Last data	% missing
370630013	S	05-04-93	10-31-97	16.7
370770001	S	04-01-90	10-31-97	16.5
370690001	U	07-07-93	10-31-97	14.6
371832001	R	04-01-90	10-31-93	15.9
371830014	S	04-01-90	10-31-97	16.5
371830015	U	08-01-91	10-31-97	54.3
371830016	U	04-27-94	10-31-97	13.3
371830017	S	07-23-93	10-31-97	52.9

68–71: eighth observation;

72: validity code for eighth observation;

Fixed-format data of this kind are easily incorporated into computer programs, whether hand-coded programs in languages such as FORTRAN or data analysis systems such as SAS and S-PLUS. The last was used for the analyses shown below.

Some summary information for the selected stations is presented in Table 1. Mean daily profiles are shown in Figure 5. These show afternoon maxima between 50 and 60 ppb and overnight minima between 10 and 20 ppb at each site, and are otherwise typical of suburban locations.

For much of the subsequent analysis, the data were reduced to a single summary quantity for each station for each day. In the light of the current National Ambient Air Quality Standard (NAAQS), this was taken to the maximum running 8-hour average concentration. The maximum was taken over all 8-hour windows

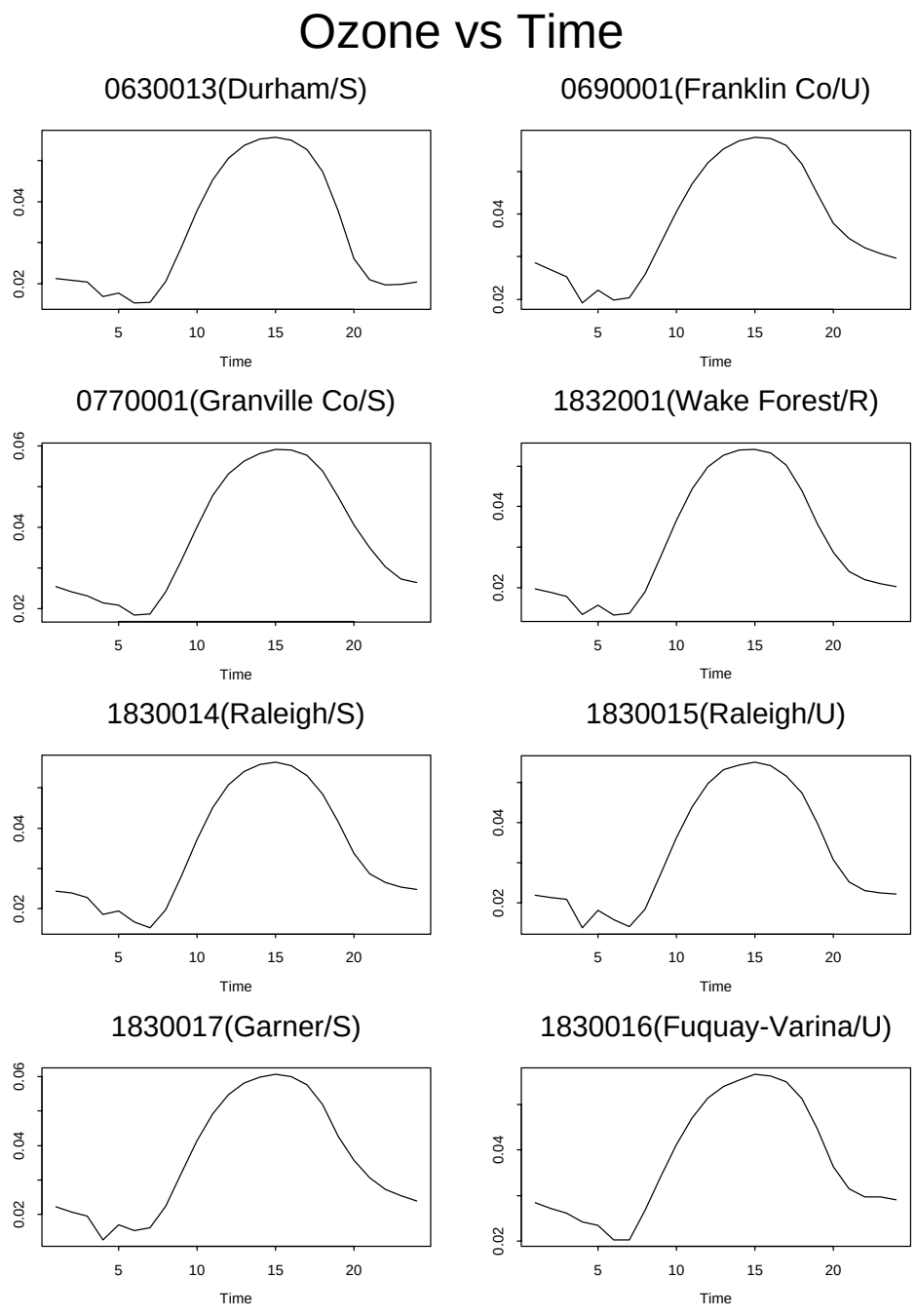

Figure 5: Mean diurnal profiles of ozone concentrations at 8 monitoring stations in the Research Triangle, NC, area.

within the day for which none of the 8 observations was missing. For all stations but one, there were days on which some observations were recorded, but no complete 8-hour windows, and hence no such maximum. The highest number of affected days was 9, and the total across stations was 32 days. Figure 6 shows the availability of the summary by station. No observations were made during the winter months (November to March).

3 Meteorological Data

A cursory query of AIRSWeb revealed only one site in North Carolina with meteorological data, located in Mecklenburg County. This location is somewhat remote from the Research Triangle, making the data only marginally useful. The State Climate Office of the State of North Carolina, operated by the Department of Marine, Earth, and Atmospheric Sciences of North Carolina State University, can make meteorological data available on request, and has provided data for Raleigh-Durham International Airport first for the calendar years 1993–1997 and more recently for the additional earlier years 1990–1992. The location makes the observations appropriate for the entire Research Triangle area.

We identified the State Climate Office as a resource after discussions with colleagues, and not from any EPA source. We view this as a weakness of the present arrangements.

4 Trend Analysis

In the Chicago study, the following model was fitted to a network average daily maximum one-hour ozone concentration:

$$\begin{aligned}\texttt{o3.1hr} = &\left(\mu_0 + \frac{t_0 + t_1\texttt{maxt} + t_2\texttt{maxt}^2 + t_3\texttt{maxt}^3 + t_{l1}\texttt{tlag1} + t_{l2}\texttt{tlag2}}{1 + \texttt{wspd}/v + \texttt{wspd700}/v_{700} + \texttt{wlag}/v_l}\right)\\ &\times (1 + r\texttt{rh} + r_l\texttt{rhlag})(1 + o\texttt{opcov})(1 + \omega\texttt{vis})\\ &\times (1 + m_u\texttt{mean.u} + m_v\texttt{mean.v})\\ &\times (1 + \tau\texttt{year})\\ &+ a_1\texttt{c1} + b_1\texttt{s1} + a_2\texttt{c2} + b_2\texttt{s2}\end{aligned}$$

Here italicized quantities are parameters estimated by nonlinear least squares, and the quantities in the `typewriter` type-face are variables:

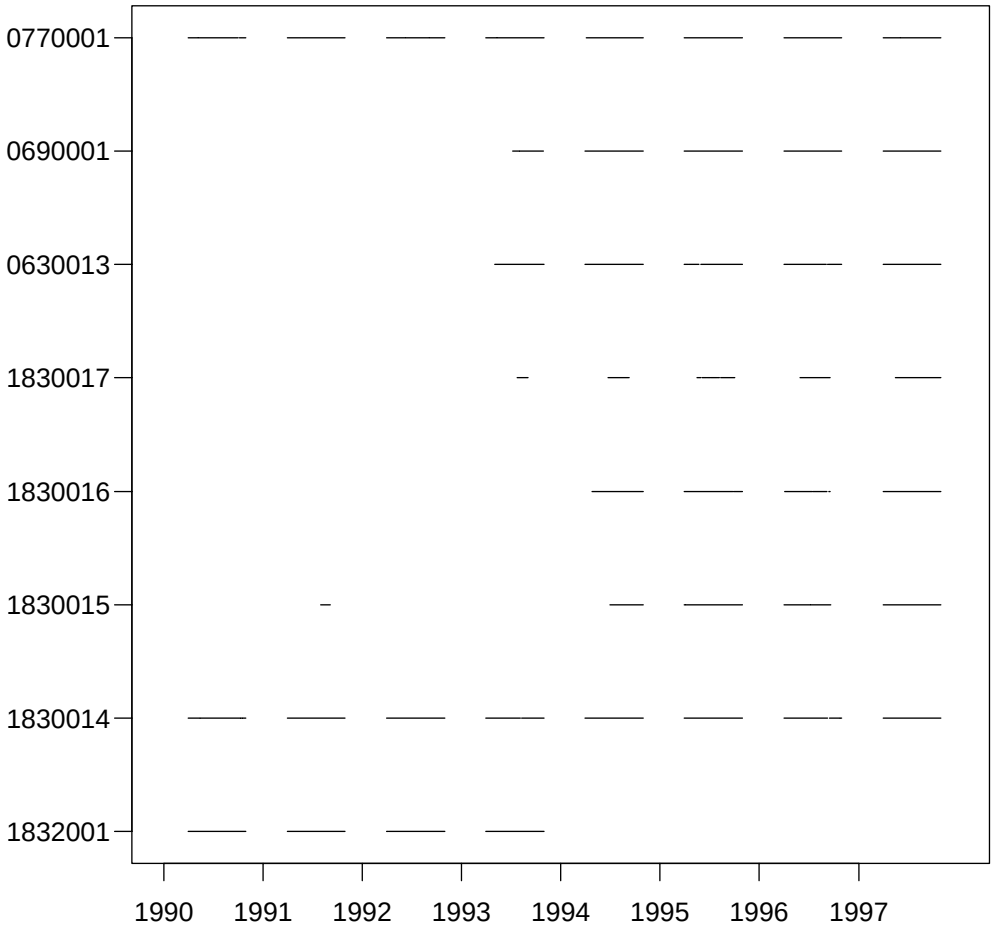

Figure 6: Availability of daily summaries of ozone concentrations.

o3.1hr: network average daily maximum one-hour ozone concentration;

maxt: maximum temperature, 9:00 a.m.–6:00 p.m.;

tlag1: 24 h average temperature, previous day;

tlag2: 24 h average temperature, two days earlier;

wspd: noon wind speed;

wspd700: 700 millibar wind speed at 1200 UTC (6:00 a.m. local time);

wlag: 24 h average wind speed, previous day;

rh: noon relative humidity;

rhlag: 24 h average relative humidity, previous day;

opcov: noon opaque cloud cover;

vis: noon visibility;

mean.u, mean.v: 24 h average of the west-to-east and south-to-north components of wind, respectively;

year: time in years relative to January 1, 1985;

c1, s1, c2, s2: cosine and sine functions with annual and semiannual frequencies, respectively.

All explanatory variables were centered at convenient values, except for the cosine and sine functions, which were centered at their means.

Since the trend term $(1 + \tau \texttt{year})$ was fitted as an integral part of a model that incorporates the association of ozone concentrations with the meteorological variables, it is interpreted as a *met-adjusted* trend estimate. That is, if there are inter-annual variations in meteorological variables that lead to corresponding changes in ozone concentrations, these are expressed in the meteorological part of the model, and have no impact on the value of τ. The trend term thus represents only those changes in ozone concentrations that cannot be attributed to meteorology.

For the Research Triangle study, a slightly different model was used.

- The response variable was a network average maximum 8-hour average ozone concentration. This was calculated from the individual station maximum 8-hour average ozone concentrations in a way similar to that used in the Chicago study. The change from maximum one-hour concentration to maximum 8-hour average concentration was made to reflect the new form of the NAAQS for ozone.

- Since upper air observations played only a minor role in the Chicago study, they were omitted from the present analysis.

- Similarly, `opcov` and `vis` were not available and were omitted.

- Lagged temperature, wind speed, and relative humidity were found to be not statistically significant, and were omitted.

- The cubic term in `maxt` was also found to be not statistically significant, and was similarly omitted. The resulting quadratic was found to vanish at approximately 80°F; with `maxt` centered at this temperature, no constant term t_0 was needed.

- To obtain a more flexible representation of trend, the linear function `year` was replaced by a natural cubic spline representation with five degrees of freedom (five "knots").

- A constant term added to the seasonal cosine and sine functions was found to be statistically significant, and was included. In the Chicago study, inclusion of such a term was explored but rejected.

Thus, the following model was fitted to a network average daily maximum 8-hour ozone concentration of Research Triangle area.

$$\begin{aligned}\texttt{o3.8hr} = &\left(\mu_0 + \frac{t_1\texttt{maxt} + t_2\texttt{maxt}^2}{1 + \texttt{wspd}/v}\right) \\ &\times (1 + r\texttt{rh}) \\ &\times (1 + m_u\texttt{mean.u} + m_v\texttt{mean.v}) \\ &\times (1 + \texttt{spline trend}) \\ &+ \mu_1 + a_1\texttt{c1} + b_1\texttt{s1} + a_2\texttt{c2} + b_2\texttt{s2}\end{aligned}$$

Table 2: Parameter estimates.

Parameter	Estimate	Standard error	t-statistic
μ_0	34.1754	5.7164	5.97
t_1	1.4943	0.1625	9.19
t_2	0.0177	0.0045	3.86
v	29.7383	12.8698	2.31
r	-0.0137	0.0027	-4.96
m_u	0.0078	0.0031	2.48
m_v	0.0090	0.0032	2.78
a_1	-11.2250	2.7251	-4.11
b_1	2.8102	0.9695	2.89
a_2	-4.6865	1.3276	-3.52
b_2	-1.5847	0.9626	-1.64
μ_1	23.1169	5.4509	4.24
ν_1	-0.1060	0.0537	-1.97
ν_2	-0.1042	0.0823	-1.26
ν_3	-0.0793	0.0568	-1.39
ν_4	-0.2954	0.1320	-2.23
ν_5	0.1177	0.0396	2.97

As before, italicized quantities are parameters estimated by nonlinear least squares, and the quantities in the `typewriter` type-face are variables. The only new variable is

`o3.8hr`: network average daily maximum 8-hour average ozone concentration (ppb);

Parameter estimates obtained by nonlinear least squares are shown in Table 2. It should be noted that the tabulated standard errors are calculated in a way that makes them valid only under standard assumptions, which are unlikely to hold for data such as these. In the Chicago study, standard errors were obtained using alternative techniques that allow for certain departures from these assumptions,

and were found to be up to three times the magnitude of those calculated under the standard assumptions.

The last five parameters are coefficients of basis cubic spline functions, and define the trend; the individual values have no particularly intuitive interpretation. The F-statistic for testing that the trend function is constant takes the value 4.01, with 5 and 831 degrees of freedom, resulting in a P-value of 0.0013. Thus the trend is apparently highly significant; the calculation is however subject to the same caveat as the standard errors.

Figure 7 shows the network average series and the fitted spline trend function which may be interpreted as the predicted ozone concentration on each day, assuming that all meteorological variables are at their centering values, and adjusted for the seasonal effects. The trend function shows little overall change from start to finish, but the initial drop and subsequent rise are noteworthy.

Figure 8 shows the estimated form of the additive seasonal component $\mu_1 + a_1\texttt{c1} + b_1\texttt{s1} + a_2\texttt{c2} + b_2\texttt{s2}$. It is similar in form to the corresponding component found in the Chicago study, except that it elevated by the constant μ_1 that was omitted in that case.

5 Out-of-sample Validation

The model described above was developed for data from 1993-1997, and the parameter estimates reported in Table2 were fitted to the same data. The model was then validated by predicting ozone concentrations for 1990-1992 based on the corresponding meteorology, without refitting the parameters.

For the validation not only the meteorological data for the out-of-sample period, but also the ozone concentration data for the same duration are needed. However only two of the seven ozone monitoring stations whose data were used in the model's construction are available for this duration. The data from these two stations were not considered adequate to cover the spatial domain covered by the 1993-1997 ozone concentration data of the seven stations. A "typical" value was constructed by analysis of the whole record, and used as the response variable in the validation.

Since the spline function used for the trend in our model can not be reliably extended beyond the range of sample, we used alternative forms of trend in this validation.

Results of the validation are shown in Table 3, with relevant statistics from the data used for fitting.

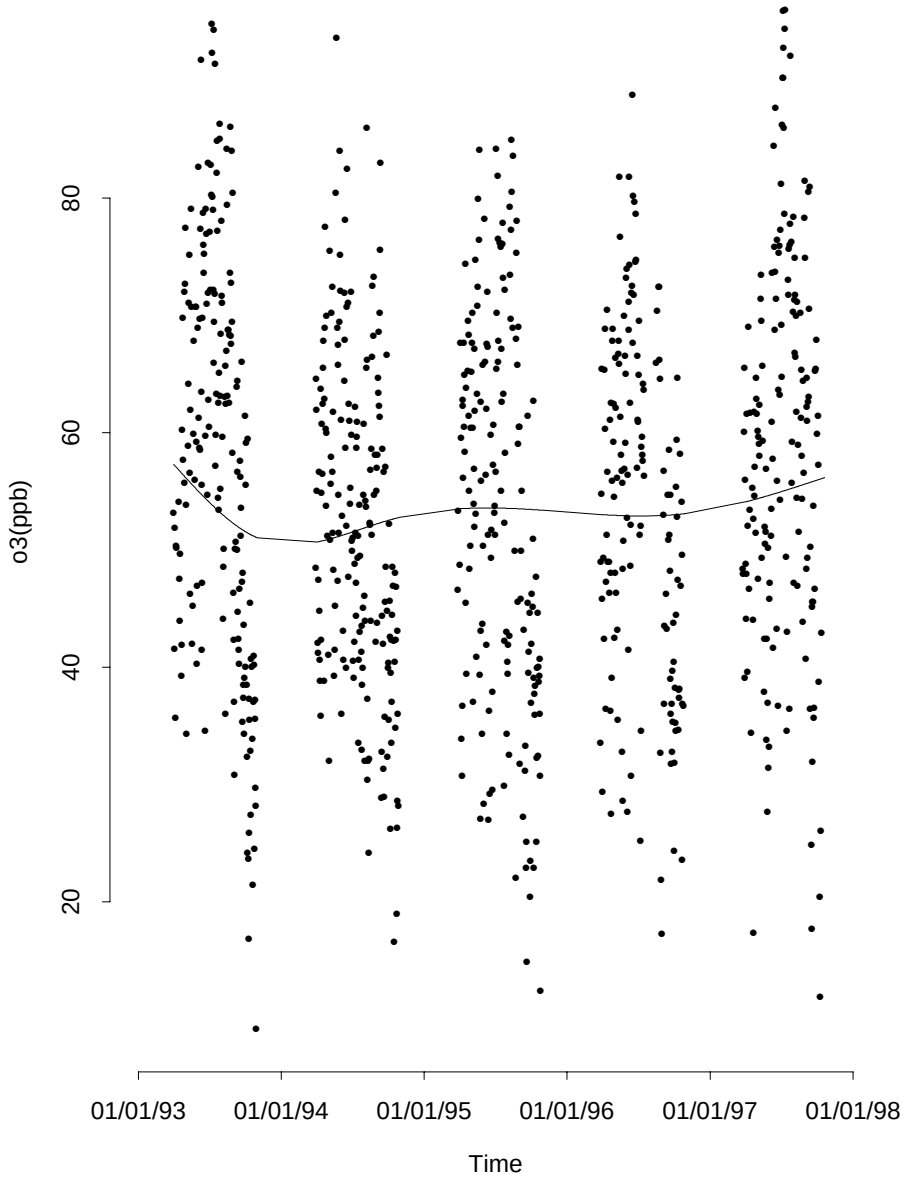

Figure 7: Network average maximum 8-hour average ozone concentration and met-adjusted trend.

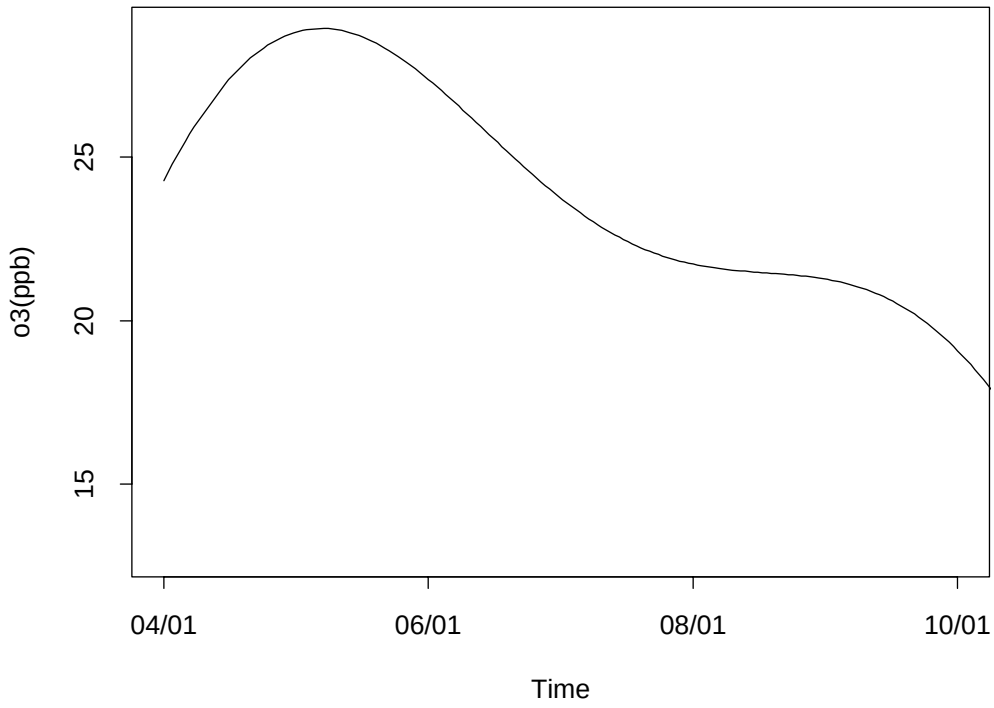

Figure 8: Seasonal component of fitted ozone model.

Table 3: Validation results (ppb)

Trend model	Validation data 1990–1992		Fitting data 1993–1997
	Mean	R.M.S.	R.M.S.
No trend	-2.840	9.619	10.12
Linear trend	-0.912	9.262	10.09

Note that for both trend assumptions, the predictive performance out of sample is actually better than the fitting performance within sample, suggesting that the model is not overspecified.

6 Comparison with the model of Cox and Chu

Cox and Chu [1993] also constructed a model for obtaining adjusted ozone trends. They fitted a model for the one-hour network maximum value rather than the maximum 8 hour average value of ozone concentration, as here. They assumed a linear relationship between the explanatory factors (meteorological and trend factors) and the response variable (logarithm of network maximum ozone concentration) and no seasonal factors, while in our model, a nonlinear relationship including seasonal factors is fitted to the raw concentrations, not the logarithms.

Specifically, Cox and Chu [1993] assumed a probability model based on the Weibull distribution as follows:

$$Prob(Y_i > \mathsf{y}) = \exp\left\{-(\mathsf{y}/\sigma_i)^{\lambda}\right\},$$

Here Y_i = daily network maximum ozone concentration for day i, σ_i = scale parameter for day i, λ = shape parameter. The scale parameter for any given day is allowed to vary as a function of the meteorological conditions in the following manner:

$$\sigma_i = \exp\left\{\sum \beta_j * M_{ij} + \tau * T\right\},$$

where M_{ij} = meteorological parameter j on day i, T = year ($T = 1, 2, ...$). The meteorological parameters were maximum surface temperature and average values of wind speed (7-10 a.m.), temp$\times$wind speed (a.m.), wind speed (1-4 p.m.), relative humidity (10 a.m.-4 p.m.), mixing height (a.m.), opaque cloud cover. They obtained maximum likelihood estimates of the coefficients based on this model.

In our fitting of their model for the data of Research Triangle area for 1993-1997, all the meteorological factors they used were included except mixing height and opaque cloud cover, which are not available to us.

The fitting results of their model are distinctly worse than the presented model, as shown in Table 4. On the basis of the simplest measure, mean squared error of prediction, the Cox and Chu [1993] model (13.23 ppb) was worse than ours

Table 4: Comparison with the model of Cox and Chu (ppb)

Model	Validation data 1990–1992		Fitting data 1993–1997	
	Mean	R.M.S.	Mean	R.M.S.
Presented	-0.912	9.262	0	10.09
Cox and Chu	0.801	12.473	0.736	13.23

(10.09 ppb) assuming linear trend. Often, a favorable mean squared error of prediction can result from overparameterization. This typically causes bad fitting out-of-sample, so the possibility can be checked by out-of-sample validation of the two models. In the case of the Cox and Chu [1993], the mean squared error of prediction of out-of-sample validation using the 1990-1992 data is 12.47 ppb which is still somewhat higher than 9.26 ppb of the presented model assuming linear trend.

References

Peter Bloomfield, J. Andrew Royle, Laura J. Steinberg, and Qing Yang. Accounting for meteorological effects in measuring urban ozone levels and trends. *Atmos. Env.*, 30:3067–3077, 1996.

Peter Bloomfield, J. Andrew Royle, and Qing Yang. Accounting for meteorological effects in measuring urban ozone levels and trends. Technical report, NISS, 1993.

William M. Cox and Shao-Hang Chu. Meteorologically adjusted ozone trends in urban areas: a probabilistic approach. *Atmos. Env.*, 27B:425–434, 1993.