NISS

Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends

Peter Bloomfield, Andy Royle and Qing Yang

Technical Report Number 1
June, 1993

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

Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends*

Peter Bloomfield Andy Royle
Qing Yang

National Institute of Statistical Sciences

and

Department of Statistics
North Carolina State University

Preliminary Report: 9 June, 1993

Abstract

Surface ozone levels are determined by the strengths of sources and precursor emissions, and by the meteorological conditions. Observed ozone concentrations are valuable indicators of possible health and environmental impacts. However, they are also used to monitor changes and trends in the sources of ozone and of its precursors, and for this purpose the influence of meteorological variables is a confounding factor. This report describes a study of ozone concentrations and meteorology in the Chicago area. The data are described using a variety of exploratory methods, including median polish and principal components analysis. The key relationships observed in these analyses are then used to construct a model relating ozone to meteorology. The model can be used to estimate that part of the trend in ozone levels that cannot be accounted for by trends in meteorology, and to "adjust" observed ozone concentrations for anomalous weather conditions. The model

*Research supported by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0.

parameters are estimated by nonlinear least squares. Its goodness of fit is assessed by comparison with nonparametric regression results (lowess).

Keywords: Ozone concentration, meteorological adjustment, nonlinear regression, nonparametric regression.

Contents

1 Introduction **4**

2 Description of Data **5**
- 2.1 Ozone data 5
- 2.2 Meteorological data 9

3 Preliminary Analyses **11**
- 3.1 Diurnal cycles 11
- 3.2 Imputation of missing values 13

4 Summarizing the Network **15**
- 4.1 Principal components analysis 15
- 4.2 Network average 23

5 Modeling Ozone Concentrations **23**
- 5.1 Surface meteorology 23
- 5.2 Upper air meteorology 29
- 5.3 Lagged meteorological variables 30
- 5.4 Seasonal structure and trend 30
- 5.5 Fitted coefficients and standard errors 33

6 Discussion of the Model **35**
- 6.1 Quality of the fitted model 35
- 6.2 Predicted and adjusted percentiles 38
- 6.3 Interpretation of model components 41
- 6.4 Cross validation 44
- 6.5 Jackknifing 46
- 6.6 Network maximum 49

7 Conclusions **54**

List of Tables

1 The ozone monitoring stations. 6
2 Meteorological variables and station locations. 9
3 Results of principal components analysis for 16 stations. 17
4 Station loadings for the first five principal components. 18
5 Coefficients in the fitted model. 34
6 Root mean square residual and one day lagged correlation coefficient, by month 37
7 Predicted residual mean and root mean square by year. 44
8 Jackknifed estimates and standard errors. 48
9 Coefficients in the fitted model for the network maximum. 50

List of Figures

1 Locations of the ozone monitoring stations. 8
2 Locations of the surface and upper air weather stations. 10
3 Daily typical values for station 170317002. 12
4 Diurnal cycle for station 170317002. 12
5 Imputation of 2 p.m. observations for station 170317002. 13
6 Quantile-quantile plot of imputed 2 p.m. observations for station 170317002 against valid 2 p.m. observations. 14
7 Locations of the 16 stations used in principal components analysis. 16
8 Station loadings for component 1. 19
9 Station loadings for component 2. 20
10 Station loadings for component 3. 21
11 Station loadings for component 4. 22
12 Scatter plots of ozone against temperature, wind speed, relative humidity, and cloud cover. 25
13 Scatter plots of ozone against wind direction and barometric pressure. 26
14 Nonparametric regression surface for ozone against temperature (maximum from 09:00 to 18:00) and noon relative humidity. . . . 26
15 Nonparametric regression surface for ozone against temperature (maximum from 09:00 to 18:00) and noon wind speed. 27
16 Residuals from meteorological model against day of year. 31

17 Nonparametric regression surface for ozone against season and the fit from the nonseasonal model. 32
18 Quantile-quantile plot of the residuals from the fitted model against the Gaussian distribution. 36
19 Nonparametric regression surface for magnitude of residuals against season and fitted value. 37
20 Actual and predicted 95th percentiles of ozone by year. 39
21 Actual and adjusted 95th percentiles of ozone by year. 40
22 Fitted polynomial effect of temperature at zero wind speed. 41
23 Fitted seasonal effect. 42
24 Quantile-quantile plot of predicted root mean square residual by year, against a bootstrap reference distribution. 46
25 Quantile-quantile plot of ratio of jackknifed standard errors to adjusted standard errors. 49
26 Fitted polynomial effect of temperature at zero wind speed for network maximum 51
27 Fitted seasonal effect for network maximum. 52
28 Actual and predicted 95th percentiles of network maximum ozone by year. 53
29 Actual and adjusted 95th percentiles of network maximum ozone by year. 53

1 Introduction

Current knowledge about trends in surface ozone concentrations has been reviewed recently by the National Research Council (1991, Chapter 2), who divided efforts to quantify the impact of meteorology on ozone into those based on classificatory methods, and those based on regression methods. The former have the advantage of nonparametric flexibility, but are essentially uninterpretable. The latter give interpretable models, but are heavily tied to linear models. Cox and Chu (1992) used a generalized linear model approach, introducing interaction between two key meteorological variables (temperature and wind speed) by including their product.

The modeling effort described here was based on ozone and meteorological data from the Chicago area (Cox, 1992). Graphical displays and nonparametric modeling, described in detail in Section 5, were used to confirm the identities of the key meteorological variables and their interactions, and to choose functional forms for modeling the dependence of ozone. Nonlinear least squares methods were then used to fit a model incorporating these variables and interactions.

When the model is extended to include a dependence on year, the associated coefficient may be interpreted as an estimate of proportional trend in ozone, allowing for changes in the meteorological variables. The model may also be used to construct "adjusted" ozone levels, that is, estimates of what the concentration of ozone would have been on a given day or in a given year, had the meteorological conditions been different.

The development of the model is driven by empirical study of the association of ozone levels with the various meteorological variables. This approach complements that used by Comrie and Yarnal (1992), who constructed a synoptic climatology of atmospheric circulation and ozone concentrations in Pittsburgh, Pennsylvania. Kelly, Ferman and Wolff (1986) similarly studied the meteorological conditions associated with high and low ozone days in Detroit, Michigan. The conclusions of such studies provide considerable guidance in the construction of models such as that developed here.

Feister and Balzer (1991) have also studied the dependence of ozone concentrations on meteorology, but with a view towards short-term forecasting. Their model includes the previous day's ozone concentration, and they find that this is the most important variable. Previous values of ozone have not been used in the present study, as their incorporation in a model makes its use for adjustment problematic.

2 Description of Data

2.1 Ozone data

The ozone data consisted of hourly averages at the 45 stations described in Table 1. The locations of the ozone monitoring stations are shown in Figure 1.

Most of the 45 stations recorded data only during the summer months, although some were operated essentially year-round. In all the analyses described subsequently, data were limited to the ozone "season" of 1 April to 31 October.

Table 1: The ozone monitoring stations. "AIRS" is the EPA air quality data base. "MSA" is the Metropolitan Statistical Area identifier for the station location. Dates of first and last observations are given in "yymmdd" form. Codes are: R-residential, I-industrial, C-commercial, A-agricultural, M-mobile; S-suburban, U-urban, R-rural.

AIRS Site ID	Latitude	Longitude	MSA	State	First and Last Dates		Code
170310001	41.669	87.733	1600	IL	880612	911030	RS
170310003	42.061	88.003	1600	IL	810206	821216	–
170310009	41.747	87.732	1600	IL	810331	820526	RU
170310032	41.757	87.545	1600	IL	870331	911031	IU
170310037	41.979	87.669	1600	IL	850723	911231	RU
170310038	41.705	87.632	1600	IL	810331	810912	RS
170310044	41.790	87.583	1600	IL	811001	861112	RS
170310045	41.923	87.634	1600	IL	810101	860105	CU
170310050	41.708	87.568	1600	IL	820101	911231	IS
170310053	41.739	87.726	1600	IL	821130	870914	RS
170310062	41.917	87.573	1600	IL	890516	890727	AS
170310063	41.877	87.634	1600	IL	910729	911231	MU
170310064	41.790	87.602	1600	IL	890701	911031	RS
170311002	41.616	87.558	1600	IL	810101	901031	RS
170311003	41.984	87.792	1600	IL	810331	911030	MU
170311601	41.668	87.990	1600	IL	830407	911030	IS
170312301	41.874	87.849	1600	IL	810101	820831	CS
170313005	42.060	87.753	1600	IL	810101	820831	CS
170314002	41.855	87.753	1600	IL	830217	911031	RS

continued on next page

Table 1 (continued)

AIRS Site ID	Latitude	Longitude	MSA	State	First and Last Dates		Code
170314003	42.039	87.898	1600	IL	850321	911031	RS
170315001	41.514	87.645	1600	IL	810101	820201	–
170316002	NA	NA	1600	IL	830701	841016	CS
170317002	42.062	87.676	1600	IL	810101	911231	RS
170318003	41.631	87.568	1600	IL	910401	911031	RS
170431002	41.856	88.077	1600	IL	810226	831231	RS
170436001	41.813	88.073	1600	IL	840208	911231	AS
170890005	42.050	88.273	620	IL	810101	911231	RS
170970001	42.177	87.865	3965	IL	810101	911231	RS
170971002	42.391	87.847	3965	IL	810101	911231	RS
170971003	42.446	88.103	3965	IL	900606	911031	AR
170973001	42.290	87.982	3965	IL	810101	911231	RS
171110001	42.221	88.241	1600	IL	810101	911231	RS
171971007	41.278	88.220	3690	IL	810227	891108	AR
171971008	41.592	88.049	3690	IL	810101	911231	RS
180890011	41.627	87.481	2960	IN	810131	820818	CU
180891016	41.600	87.335	2960	IN	820901	910930	RU
180892001	41.600	87.383	2960	IN	810101	820815	IU
180892002	41.631	87.521	2960	IN	811002	820831	RU
180892008	41.639	87.501	2960	IN	810101	910930	CS
181270020	41.631	87.087	2960	IN	830930	911031	IR
181270021	41.563	87.076	2960	IN	810512	851031	AR
181270024	41.617	87.199	2960	IN	831115	910930	RS
181270903	NA	NA	2960	IN	810101	810131	IS
181271004	41.463	87.044	2960	IN	830430	851018	RS
181271005	41.467	87.061	2960	IN	810430	821031	CU

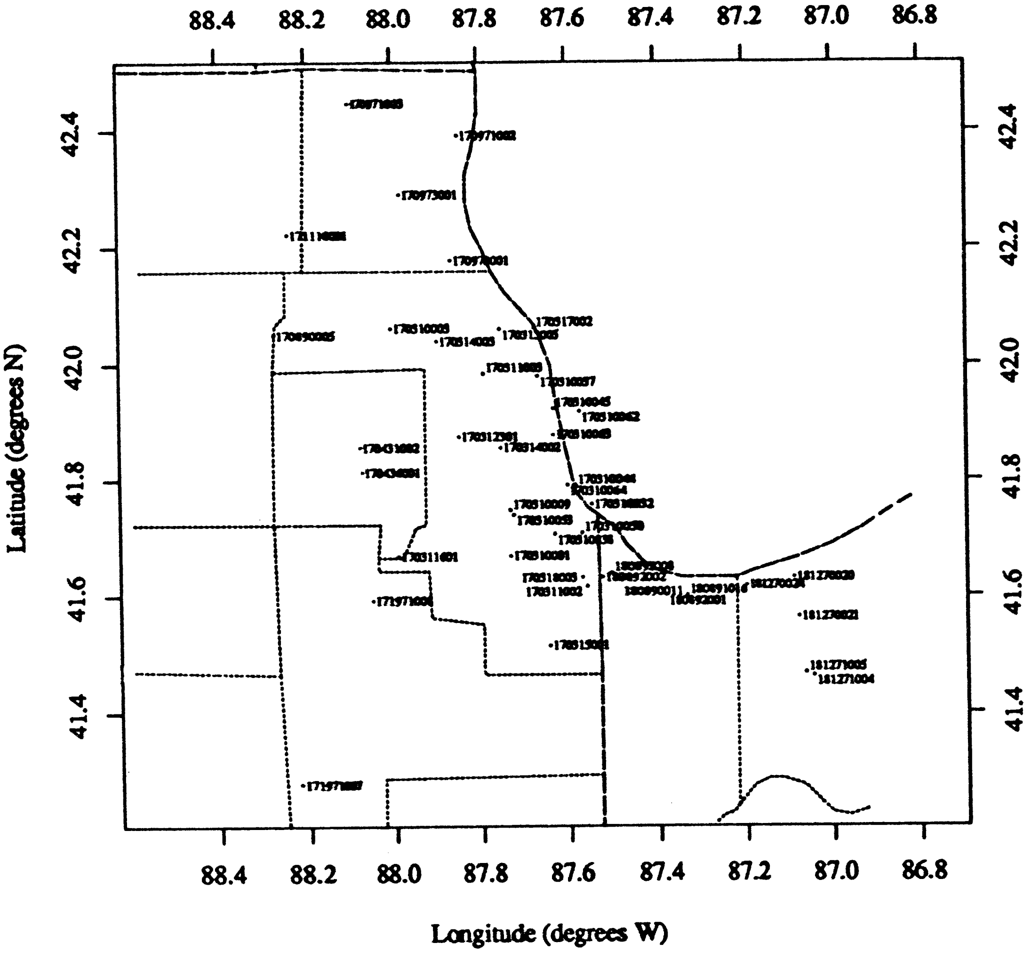


Figure 1: Locations of the ozone monitoring stations.

2.2 Meteorological data

Surface and upper air meteorological variables were collected at the two stations shown in Figure 2. The meteorological variables are described in Table 2. The surface observations were made each hour, while the upper air soundings were made largely at 00Z and 12Z (00:00 and 12:00 UTC, respectively; occasional soundings at 23Z and 11Z were taken as being at 00Z and 12Z, respectively; other soundings were ignored). The upper air measurements were made at many levels; these always included 950mb, 850mb, 700mb, and 500mb, which were the only levels at which the data were used.

Table 2: Meteorological variables and station locations.

Variable	Units	Surface	Upper	Symbol
Total cloud cover	%	Yes		`totcov`
Opaque cloud cover	%	Yes		`opcov`
Ceiling height	m	Yes		`cht`
Barometric pressure	mb	Yes		`pr`
Temperature	°F	Yes	Yes	`t`
Dewpoint temperature	°F	Yes	Yes	`td`
Relative humidity	%	Yes	Yes	`rh`
Specific humidity	g/kg	Yes	Yes	`q`
Wind direction	° from N	Yes	Yes	`wdir`
Wind speed	m/s	Yes	Yes	`wspd`
Visibility	km	Yes		`vis`
Height of pressure layer	m		Yes	`ht`
Latitude		41.98°	40.67°	
Longitude		87.90°	89.68°	

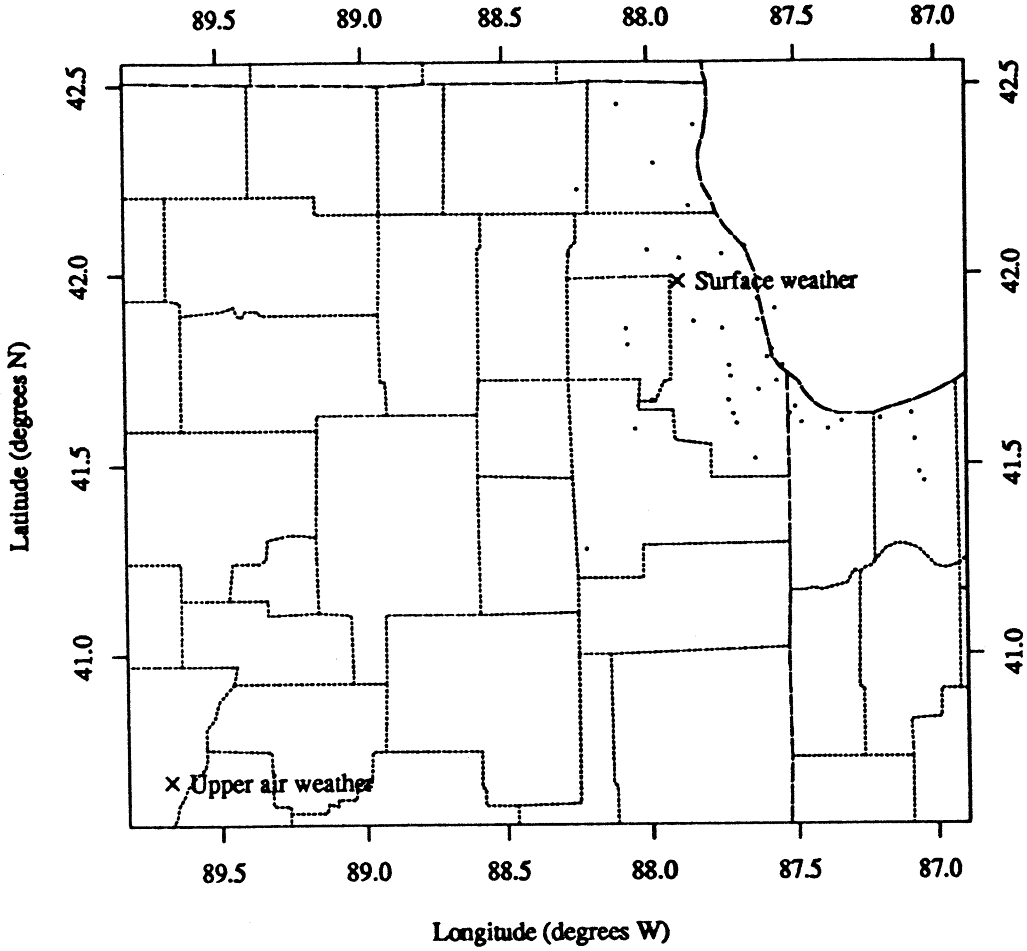


Figure 2: Locations of the surface and upper air weather stations. Dots indicate ozone monitoring stations. Note change of scale.

3 Preliminary Analyses

3.1 Diurnal cycles

The recorded ozone concentrations for a given station may be written as a two-way array:

$$y_{d,h} = \text{concentration on day } d \text{ at hour } h.$$

The diurnal cycle for the station and a "typical" value for each day were obtained by decomposing the logarithms of the data as

$$\log y_{d,h} = \mu + \alpha_d + \beta_h + \epsilon_{d,h}. \tag{1}$$

The decomposition was made using median polish (Tukey, 1977), as implemented in S (Becker, Chambers and Wilks, 1988, see `twoway`). The advantage of median polish, which is closely related to least absolute deviations fitting of equation (1) (Bloomfield and Steiger, 1983), is that it is influenced more by the body of a distribution and less by its tails than methods based on averages. Thus the "typical" values for an hour or for a day are less affected by deviations from normal levels that last for less than half a day. Median polish also works well in the presence of missing data (Tukey, 1977, Chapter 10).

The decomposition was made on the logarithmic scale, to correspond to a multiplicative decomposition of the actual ozone concentrations. This is appropriate when effects are expected to be proportional; for instance, when the typical diurnal profile is expected to be scaled by the daily effect, rather than offset by it. The *diagnostic plot* (Tukey, 1977, Section 10F) indicated that the decomposition was more satisfactory on the logarithmic scale.

Figure 3 shows the fitted daily typical values, $\exp(\mu + \alpha_d)$, for a station with one of the longer records. The seasonal cycle is visible, as are the unusually high values in 1988. Figure 4 shows the corresponding diurnal cycle, $\exp(\mu + \beta_h)$. This shows the characteristic behavior of urban ozone: a mid-afternoon maximum, followed by a decay to an early morning minimum, which is sharpened by NO scavenging during the morning rush hour, and is followed by a rise through the late morning and the middle part of the day. This rise is caused initially by mixing of the surface layer with the relatively ozone-rich residual layer above it, and later by photochemical production of ozone. For this station, the root mean square residual in the median polish decomposition (on the logarithmic scale) was 0.62, indicating quite large proportional variations of the observed data around the fitted values.

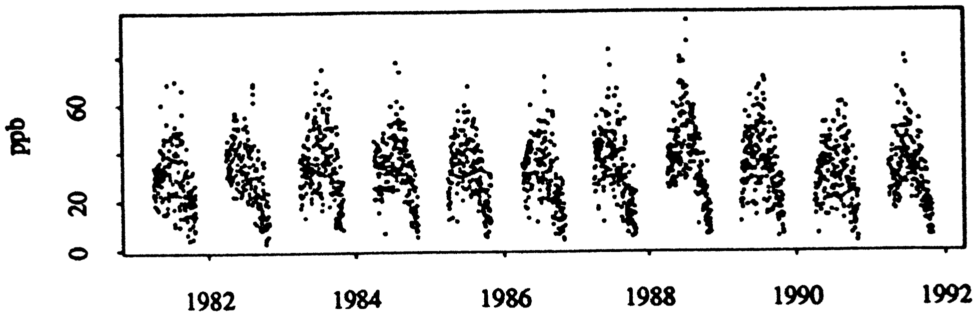


Figure 3: Daily typical values for station 170317002.

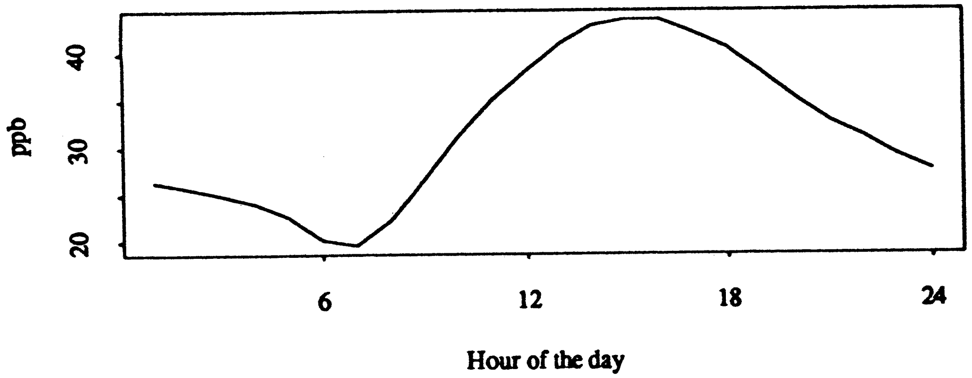
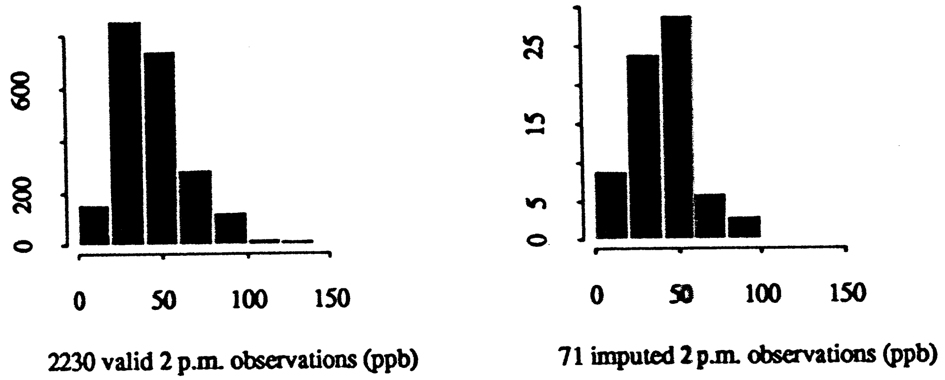


Figure 4: Diurnal cycle for station 170317002.

3.2 Imputation of missing values

The decomposition (1) was used to impute values for missing hourly ozone concentrations. If $y_{d,h}$ is missing, but the first three terms on the right hand side of (1) are available, then their sum provides a "fitted value" for the logarithm of the ozone concentration at that hour. These three terms are available if there are any data for the relevant station on the same day but at different hours, and for that hour but on different days. At every station there was enough data to construct a reliable estimate of the diurnal cycle terms $\beta_1, \beta_2, \ldots, \beta_{24}$, so this procedure gave imputed values for all missing data other than where entire days were missing. However, it was found that the imputed values were unreliable when there were fewer than 2 valid hourly averages between 9:00 a.m. and 6:00 p.m., and they were not used in this situation.

The root mean square residual in equation (1) of 0.62 means that the imputed value typically differed from the missed value by a factor of 0.5 to 2. However, their *distribution* was quite similar to the valid data. Figure 5 shows the histograms of the valid and imputed 2 p.m. observations for the station shown in Figures 3 and 4. Figure 6 is the corresponding quantile-quantile plot (Becker et al., 1988, see `qqplot`). Here and later, the order statistics of the smaller sample are graphed against estimated quantiles from the larger sample, obtained by interpolation and



2230 valid 2 p.m. observations (ppb)

71 imputed 2 p.m. observations (ppb)

Figure 5: Imputation of 2 p.m. observations for station 170317002.

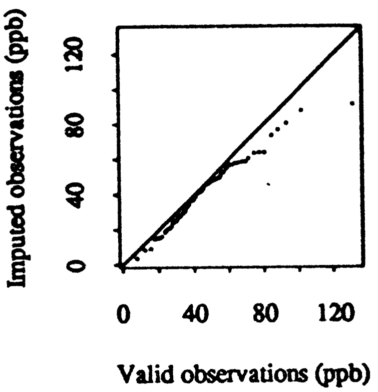


Figure 6: Quantile-quantile plot of imputed 2 p.m. observations for station 170317002 against valid 2 p.m. observations.

by assuming that the ith order statistic in a sample of size n estimates the $(i - \frac{1}{2})/n$ quantile (Becker et al., 1988, see `quantile`). The distribution of the imputed data agrees closely with that of the valid data up to about 60 ppb, but the largest imputed values appear to be too small. However, this affects only a handful of days, and should lead to negligible bias in further analyses. For instance, 26 of the 2230 valid 2 p.m. values exceed 120 ppb, suggesting that 1 or so of the missed values might have been in this range, while the largest imputed value was below 90 ppb. However, the percentages of values above 120 ppb are 1.166% in the valid data and 1.130% in the data set including the imputed values.

A further level of imputation was used for the analyses described in following sections, including principal components analysis (Section 4.1) and nonlinear modeling (Section 5). Where an entire day of data was missing, but at least some data were measured on both the preceding and succeeding days, the logarithmic-scale daily effect α_d was imputed as the average of the two neighboring values. This is equivalent to using the geometric mean on the original scale of measurement. Days with fewer than 2 valid observations between 9:00 a.m. and 6:00 p.m. were treated in the same way: although α_d was available, it was flagged as missing, and replaced by the average of the neighboring values, if available. For most stations there were 10 or fewer days meeting these conditions, so this had relatively little effect.

4 Summarizing the Network

4.1 Principal components analysis

The joint behavior of ozone concentrations at the various stations was explored using principal components analysis of the daily maximum concentration. This was carried out computationally through the singular value decomposition of a data matrix whose rows corresponded to days of data, and whose columns corresponded to the stations. Since no missing values were allowable, an appropriate subset of stations and days had to be found. To minimize missing data, both levels of imputation described in Section 3.1 were used. The matrix was constructed sequentially, beginning with the station with the most complete data, and successively adding stations in descending order of amount of data. The first 16 stations yielded 662 days of data, while adding the 17th reduced this to 274. The locations of the 16 stations chosen for the analysis are shown in Figure 7. The singular values and some related quantities are shown in Table 3. The station loadings for the first 5 components are shown in Table 4. Figures 8 to 11 show the results of spatial interpolation (Becker et al., 1988, see `interp`) of the station loadings for the first four components.

The first component, accounting for 79% of the variance of the 16 station sub-network, is a weighted average of the station values, with weights ranging from 0.20 to 0.32. Stations close to Lake Michigan carry the higher weights, while those further from the lake have lower weights. The second and third components account for a further 5% and 4% of variance, respectively. They represent gradients across the network, from East to West and from North to South, respectively. The remaining singular values are not well separated, and most are presumably noise. However, the fourth component, while not well distinguished from the rest, nevertheless has some spatial interpretation as a contrast between the interior and boundary stations. Components 2 to 4 may reflect meteorological conditions; this will be explored elsewhere.

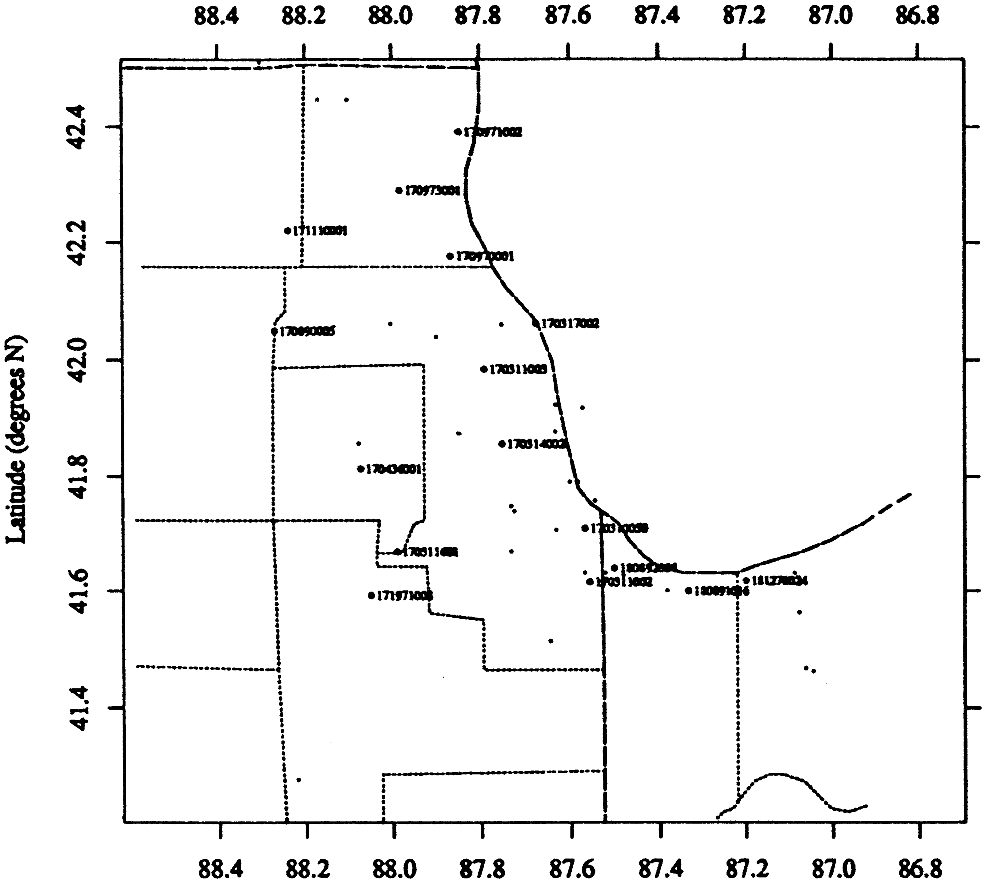


Figure 7: Locations of the 16 stations used in principal components analysis. Small dots indicate stations not used.

Table 3: Results of principal components analysis for 16 stations.

Singular value	Percent of variance	Cumulative percent
1828.81	78.8523	78.85
451.36	4.8031	83.66
415.50	4.0703	87.73
284.81	1.9124	89.64
268.73	1.7026	91.34
240.70	1.3659	92.71
230.02	1.2474	93.95
221.47	1.1564	95.11
210.90	1.0486	96.16
197.13	0.9162	97.08
169.28	0.6756	97.75
161.87	0.6178	98.37
156.91	0.5805	98.95
144.68	0.4935	99.44
118.49	0.3310	99.77
97.98	0.2263	100.00

Table 4: Station loadings for the first five principal components.

Station	Component				
	1	2	3	4	5
181270024	0.2786	0.3956	-0.3512	-0.1620	0.5381
170436001	0.2065	-0.3170	-0.1700	-0.0380	0.0692
170311601	0.2359	-0.3976	-0.2663	0.0826	-0.0573
170311003	0.2525	-0.0622	0.1928	-0.5339	0.1447
170314002	0.2512	-0.0159	0.0759	-0.4037	-0.4021
180891016	0.2399	0.3188	-0.2732	0.2106	-0.1318
170311002	0.2334	0.1039	-0.1675	-0.3765	-0.0154
170310050	0.2384	0.1691	-0.0237	-0.0391	-0.2313
180892008	0.2574	0.2667	-0.2379	0.2017	-0.5233
171971008	0.2306	-0.3262	-0.2942	0.1316	0.0619
170970001	0.2593	-0.0300	0.3581	-0.1165	0.0287
170973001	0.2492	-0.1252	0.3090	0.0005	-0.0682
170971002	0.2959	0.0772	0.4390	0.3147	-0.0893
171110001	0.2043	-0.2771	0.0052	0.1937	0.0834
170890005	0.2185	-0.3230	-0.0716	0.1242	0.1124
170317002	0.3202	0.2423	0.2567	0.3261	0.3715

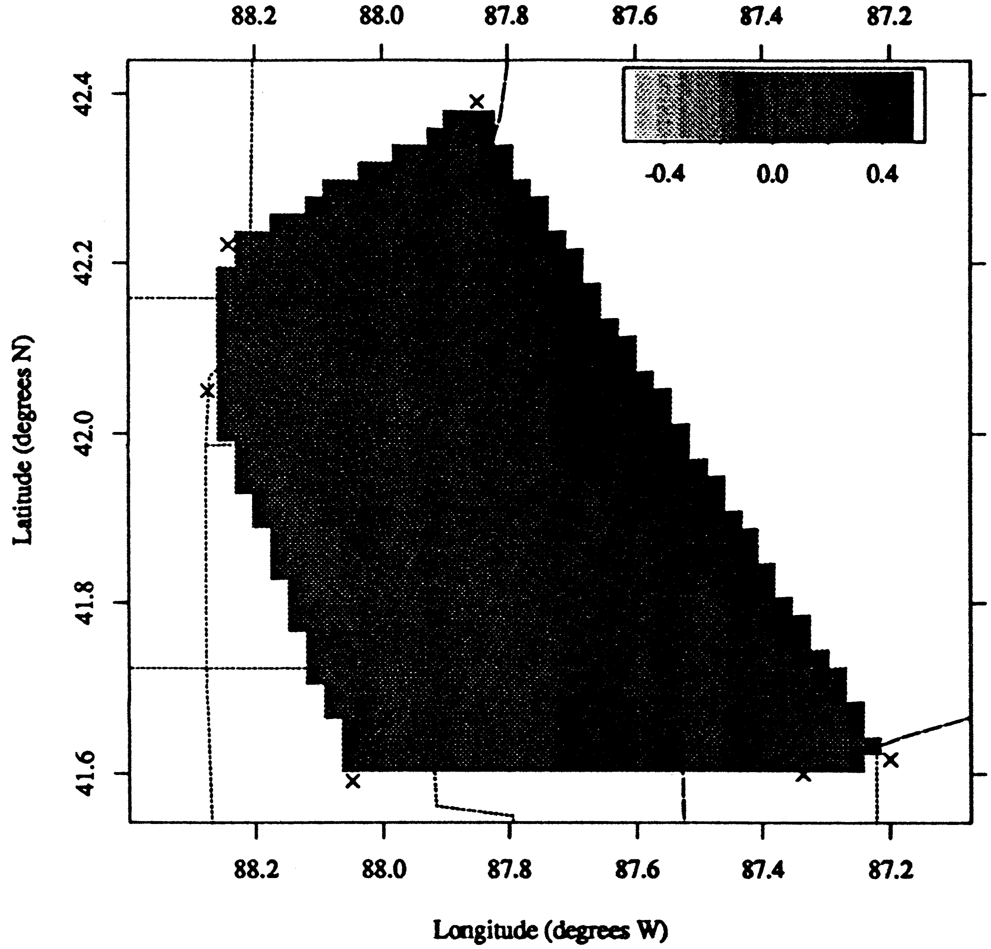


Figure 8: Station loadings for component 1.

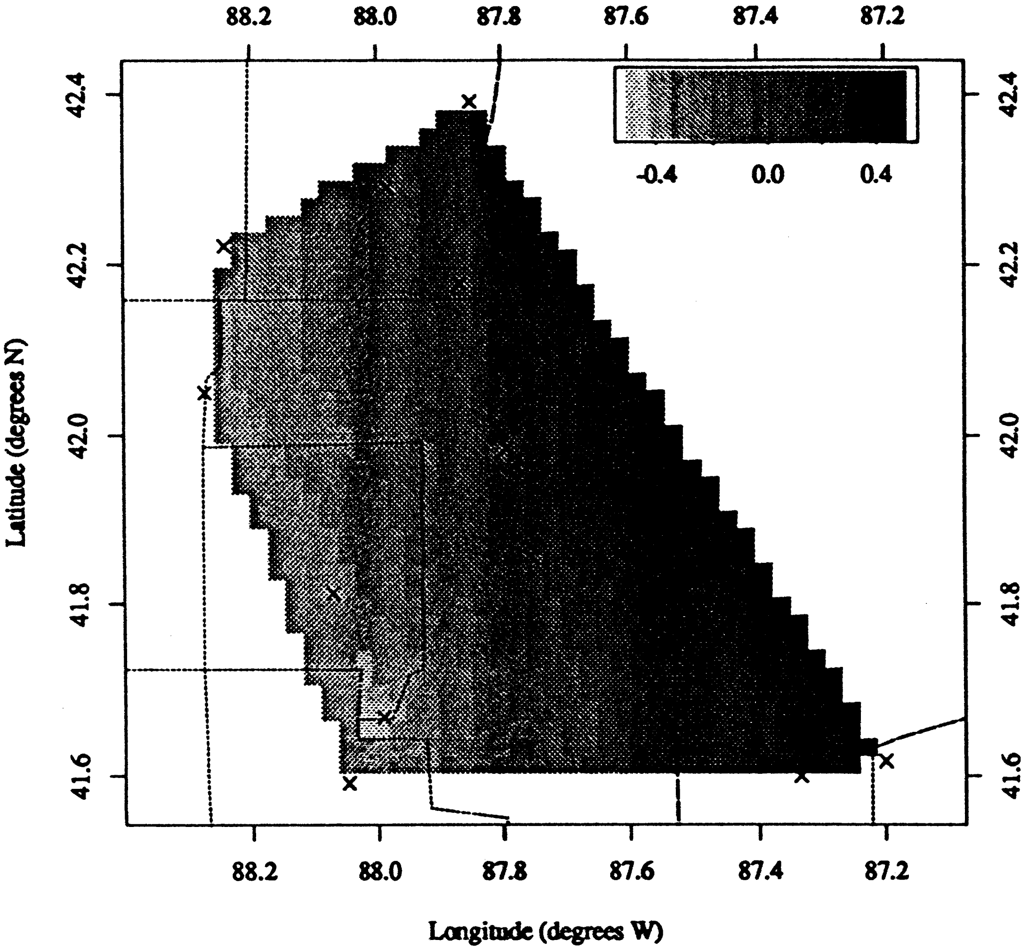


Figure 9: Station loadings for component 2.

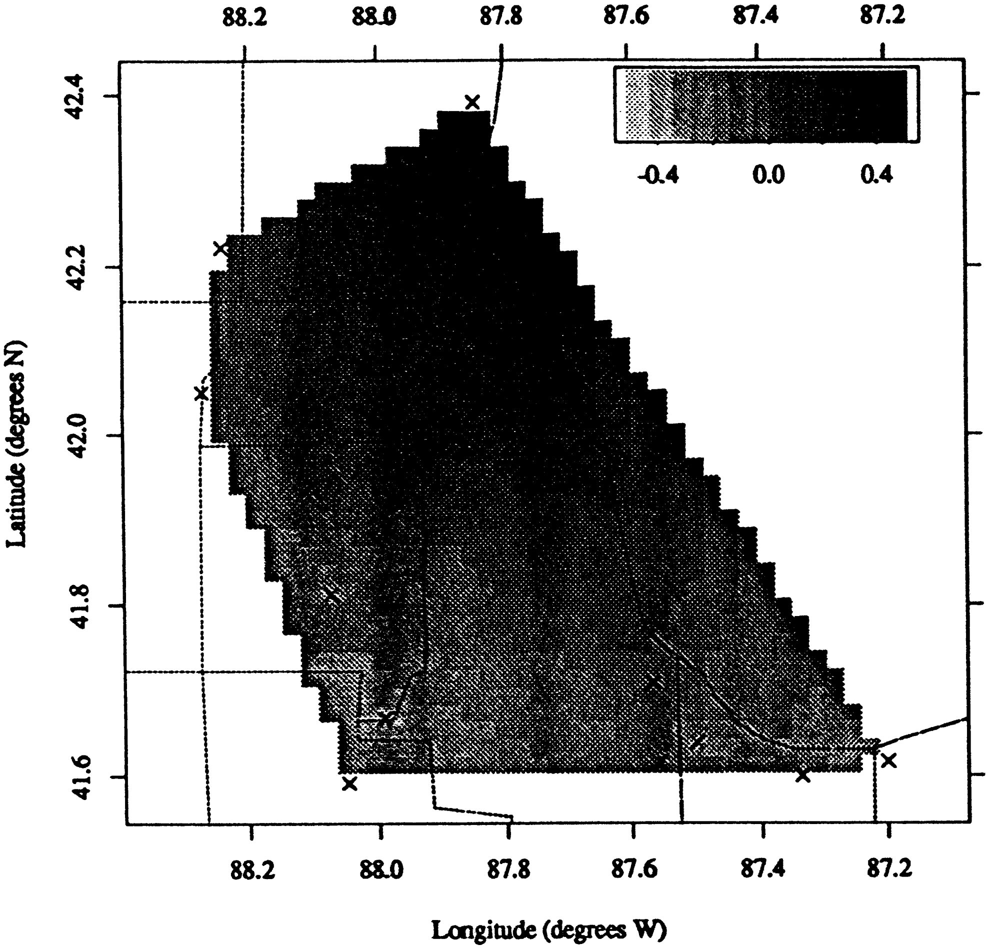


Figure 10: Station loadings for component 3.

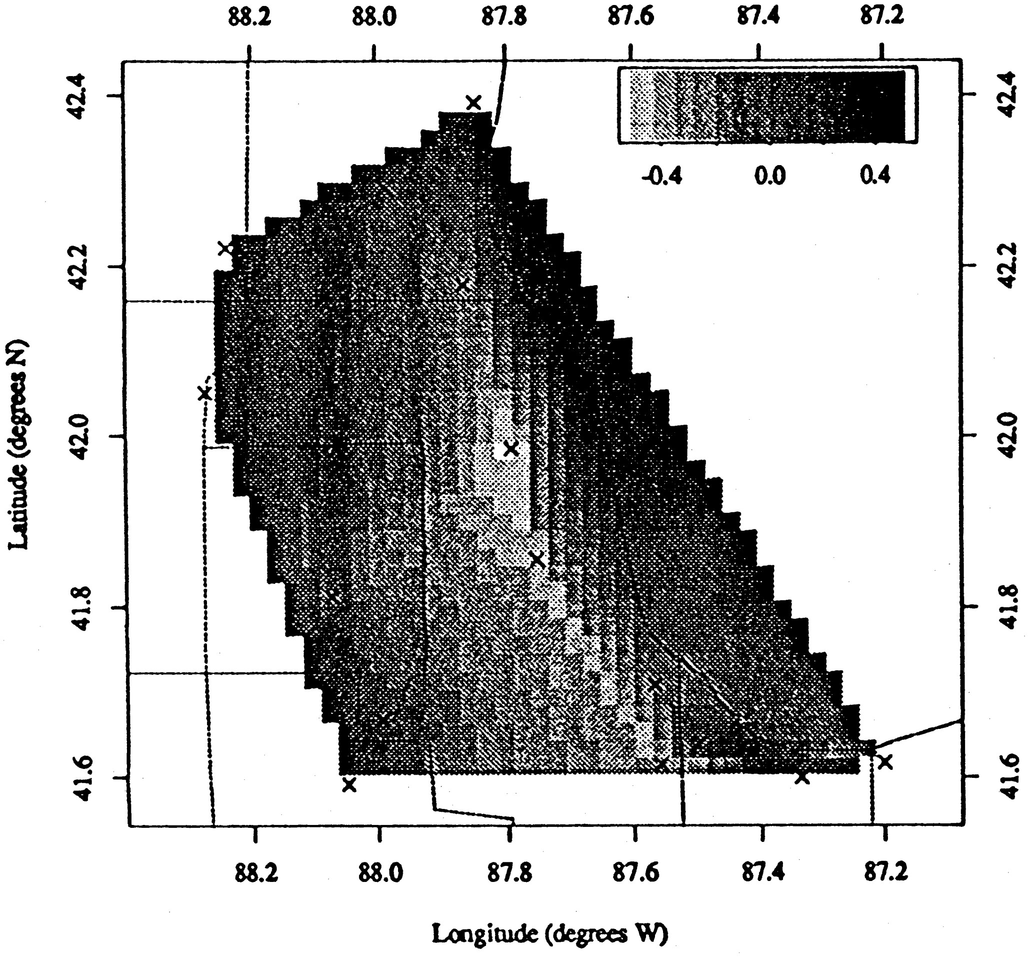


Figure 11: Station loadings for component 4.

4.2 Network average

The principal components analysis of Section 4.1 showed that most of the variance was associated with a single component. It was based on a subset of 662 days of data, because of the data requirements of the method, and consequently the principal component time series could be calculated only for the same subset of days. However, it was essentially a weighted average of the 16 stations, which suggested that a similar network average, calculated for a more extensive set of days, would provide a better basis for the subsequent modeling.

Median polish was used a second time to provide an outlier-resistant summary. Both levels of imputation (Section 3.1) were used to construct daily maximum ozone concentrations by station. The daily maxima were written as a two-way array:

$$y_{d,s} = \text{maximum concentration on day } d \text{ at station } s,$$

and this was decomposed as

$$y_{d,s} = \mu' + \alpha'_d + \beta'_s + \epsilon'_{d,s}.$$

Note that this decomposition is on the original scale of the observations, rather than the logarithmic scale used in Section 3.1. The corresponding diagnostic plot indicated that this was a satisfactory scale. The results were used to construct a network average of the station daily maxima, $\mu' + \alpha'_d$. They were also used to impute values where the array had missing data, in a way exactly parallel to that used in section 3.1, and a network maximum was computed from the completed array.

When the resulting network average was restricted to the set of days on which the principal components analysis was based, its correlation with the dominant component was 0.9896. Thus the new series may be regarded as an extension of the dominant component to the whole 2,354 days of the record.

5 Modeling Ozone Concentrations

5.1 Surface meteorology

The National Research Council (1991, Chapter 2) reviewed previous efforts to relate ozone concentration data to meteorological variables, finding that temperature, wind speed, relative humidity, and cloud cover were important variables.

Other variables mentioned were wind direction, dew point temperature, sea level pressure, and precipitation. All of these except precipitation were included in the suite of meteorological data available for the present study (surface barometric pressure was used in place of sea level pressure).

Figure 12 shows scatter plots of the network average daily maximum ozone concentration, described in Section 4.2, against four surface meteorological variables. Temperature is measured by the maximum from 9:00 a.m. to 6:00 p.m., while the other variables are for noon. The curve overlayed on each graph is a cubic least squares smoothing spline with 6 degrees of freedom. Figure 13 shows corresponding plots for surface wind direction and barometric pressure; dew point temperature was not considered, as it is a function, albeit nonlinear, of temperature and relative humidity, and highly correlated with temperature.

It is evident that temperature has the strongest effect, that the effect would be well approximated by a low order polynomial, and that relative humidity and wind speed appear to be the next most important variables. To explore the dependence of ozone level on these variables two at a time, the lowess method of nonparametric regression (Cleveland, 1979; Chambers and Hastie, 1992, see `loess`) was used to find regression surfaces. Figure 14 shows the surface for ozone against temperature and relative humidity. The surface indicates that the polynomial effect of temperature is maintained at all levels of relative humidity, and that the effect of relative humidity is reasonably linear at all levels of temperature, but with a larger slope where the temperature effect is larger. This suggests that a model of the form

$$\text{ozone} = (\text{polynomial in temperature}) \times (\text{linear function of relative humidity}) \qquad (2)$$

might express the joint effect of these two variables. Polynomial regression of ozone against temperature suggests that the polynomial needs to be of order at least 3; raising the order from 3 to 4 increases R^2 from 0.6052 only to 0.6065, suggesting that order 3 is adequate. Fitting equation (2) by nonlinear least squares leads to $R^2 = 0.6509$. The dimensionless function of relative humidity is estimated to be $1 - 0.00498\%^{-1} \times (\text{relative humidity} - 50\%)$, which decreases from 1.249 to 0.751 as relative humidity increases from 0% to 100%. The R^2 for this fit may be compared with that for the lowess fit of Figure 14, which was 0.66; the equivalent number of parameters was 10.2. This indicates that the parametric model, with only 5 parameters, performs very nearly as well as the nonparametric one.

The corresponding regression surface for ozone against temperature and wind

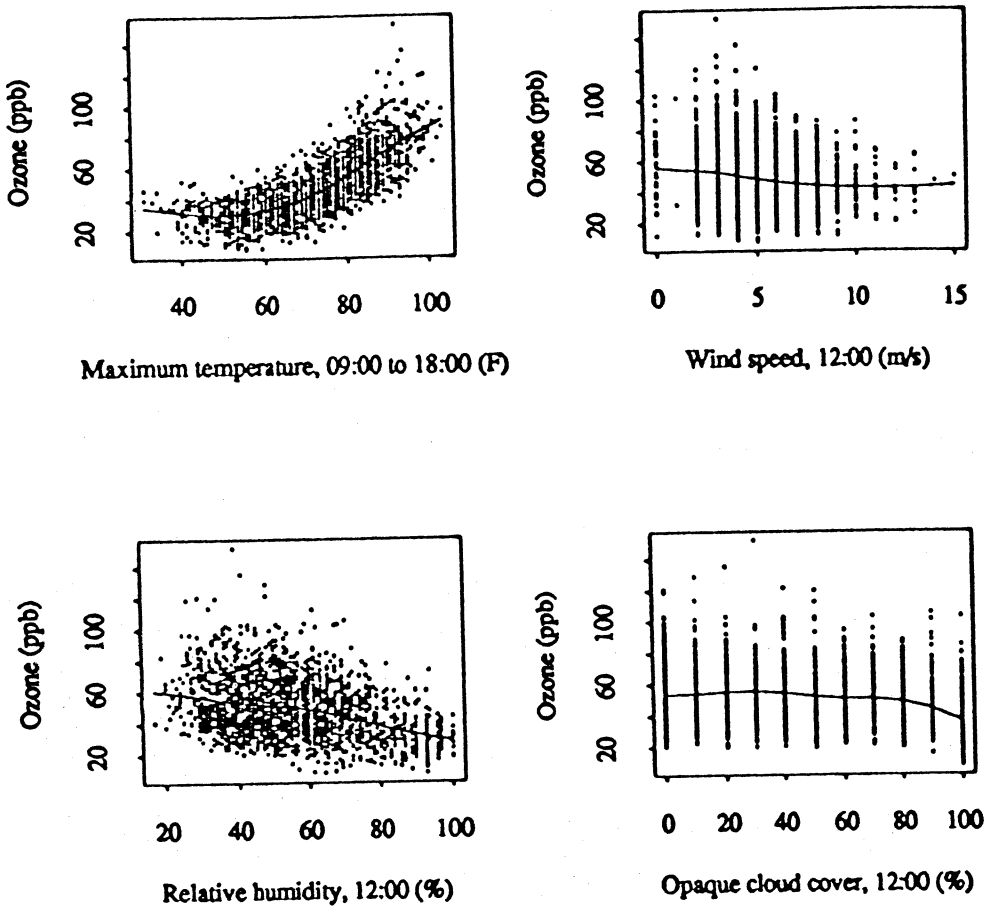


Figure 12: Scatter plots of ozone against temperature, wind speed, relative humidity, and cloud cover.

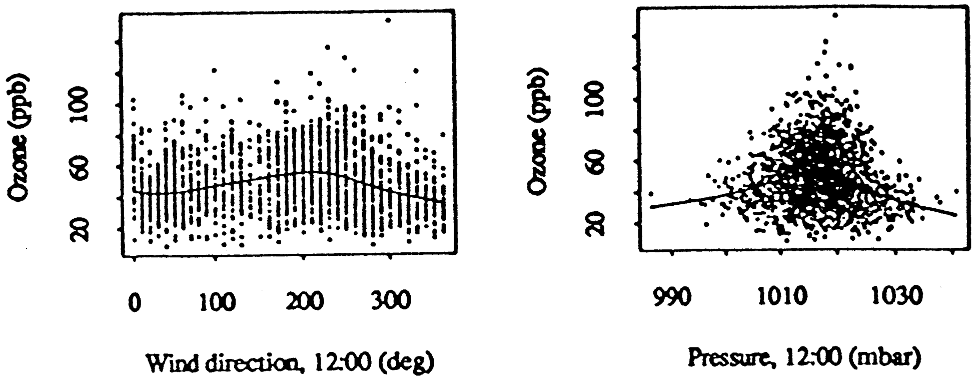


Figure 13: Scatter plots of ozone against wind direction and barometric pressure.

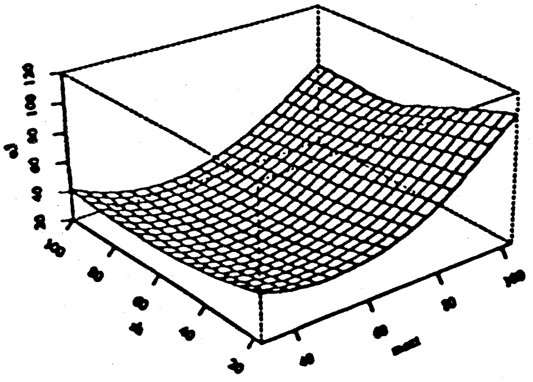


Figure 14: Nonparametric regression surface for ozone against temperature (maximum from 09:00 to 18:00) and noon relative humidity.

speed is shown in Figure 15. This surface shows similarly that the polynomial effect of temperature is maintained at all levels of wind speed, but that the effect of wind speed is neither linear nor the same (nor even proportional) at different wind speeds. Rather, increasing wind speed is associated with a *drop* in ozone at high temperatures, leveling off above around 6 m/s, but with a slight *rise* in ozone at low temperatures. This behavior could be captured in a model of the form

$$\text{ozone} = \text{constant} + (\text{polynomial in temperature}) \times (\text{function of wind speed}) \quad (3)$$

and the surface suggests that the function of wind speed might be of the form

$$\frac{1}{1+\frac{\text{wind speed}}{v}},$$

where v is a critical speed at which the effect of this dimensionless factor drops from 1 to 0.5. The nonlinear least squares fitted value of v is 7.72 m/s, and $R^2 = 0.6280$. That the effect of wind speed differs from that of relative humidity is shown in the fitted value of the constant term in equation (3), 40.8 ppb. This term would have to be set to 0 ppb for the form of the equation to reduce to the simpler multiplicative form. Forcing this change reduces R^2 to 0.6143 (and

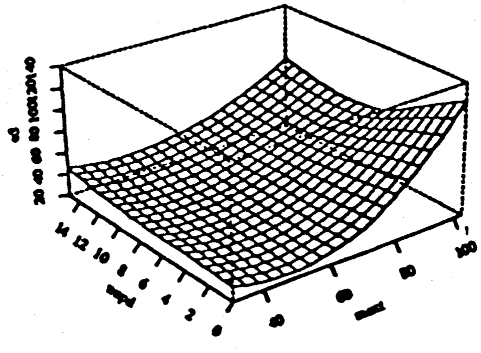


Figure 15: Nonparametric regression surface for ozone against temperature (maximum from 09:00 to 18:00) and noon wind speed.

increases v to 51 m/s). By contrast, if such a constant is included in equation (2), it is estimated to be -1.2 ppb, and R^2 is unchanged to 5 decimal places.

Again, the R^2 for the model (3), 0.6280, may be compared with that for the lowess fit, which was 0.63 with the equivalent of 10.2 parameters. The parametric model performs well by comparison.

Equations (2) and (3) may be combined as

$$\begin{aligned}\text{ozone} = {} & \{\text{constant} \\ & + (\text{polynomial in temperature}) \times (\text{function of wind speed})\} \\ & \times (\text{linear function of relative humidity}), \qquad (4)\end{aligned}$$

with wind speed entering in the same form as before. The R^2 for the combined model is 0.6752, and the coefficients change only slightly: the multiplier of (relative humidity – 50%) becomes $-0.00505\%^{-1}$, and v becomes 6.85 m/s. The corresponding lowess fit has $R^2 = 0.68$ for the equivalent of 16.7 parameters, again indicating good performance of the parametric model, which here has 7 parameters.

Study of the residuals from combined models such as equation (4) showed that ozone also depends on visibility and opaque cloud cover. On the other hand, these residuals showed essentially no dependence on barometric pressure, despite the association shown in Figure 13. High temperatures tend to be associated with pressures close to 1015 mbar, and when the effect of temperature is removed from the ozone concentrations, the association with pressure is removed as well. Much but not all of the association of ozone concentration with wind direction is similarly removed by taking out the effect of temperature.

Noon values of visibility and opaque cloud cover were used, and these were included in the model by multiplying the right hand side of equation (4) by

$$(\text{linear function of visibility}) \times (\text{linear function of opaque cloud cover}).$$

Including these terms in order raised R^2 to 0.7016 and 0.7035, respectively. The dependence of ozone levels on wind direction suggested including a similar term, of the form

$$\begin{aligned}1 + {} & \text{linear combination of} \cos\left(\frac{2\pi \times \text{wind direction}}{360}\right) \\ & \text{and} \sin\left(\frac{2\pi \times \text{wind direction}}{360}\right). \qquad (5)\end{aligned}$$

Including this term raised R^2 to 0.7064. Multiplying the cosine and sine by wind speed gave a better fit and has greater physical meaning, since the term can then be interpreted as the inner product of the wind vector with a vector of coefficients. This change increased R^2 to 0.7074. Finally, it was found that averaging the wind vector over the hours 06:00 to 18:00 gave a still better fit, with $R^2 = 0.7108$.

Equation 5 was adequate for these data because the (remaining) dependence on wind direction was simple. In many cases the dependence of ozone concentration on wind direction is multimodal, in which case a longer Fourier series would be needed. This requires including the sines and cosines of small multiples of wind direction.

5.2 Upper air meteorology

The residuals from the model described in the preceding section were graphed against and correlated with upper air meteorological variables from the 12Z sounding (06:00 CST). The only substantial association was with wind speed, and this was stronger at 700 mbar than at 950 mbar, 850 mbar or 500 mbar.

First the effect of including upper air wind speeds in the wind speed interaction with temperature was explored. The factor

$$\frac{1}{1 + \frac{\text{wind speed}}{v}},$$

was extended by adding the four upper air wind speeds, with arbitrary weights, in the denominator. The weights were then estimated by nonlinear least squares. The coefficient of 950 mbar wind speed was very small and negative. The coefficients of 850 mbar and 500 mbar wind speed were both smaller than that of 700 mbar wind speed, consistently with the ordering of the correlation coefficients. The 700 mbar term was therefore retained, and the model was refitted with the wind factor extended to

$$\frac{1}{1 + \frac{\text{surface wind speed}}{v_s} + \frac{\text{700 mbar wind speed}}{v_{700}}}.$$

This gave an R^2 of 0.7204.

Next the effect of including the 700 mbar wind vector in the wind vector part of the model was studied. The term

$$1 \; + \; \text{linear combination of (wind speed)} \times \cos\left(\frac{2\pi \times \text{wind direction}}{360}\right)$$

$$\text{and (wind speed)} \times \sin\left(\frac{2\pi \times \text{wind direction}}{360}\right).$$

was extended by adding a corresponding linear combination of the components of the 700 mbar wind vector. However, this increased R^2 by only 0.0003, indicating that there is negligible further information in the upper air wind vector beyond that in the surface wind vector. A similar experiment with the 950 mbar wind vector also gave negative results. This extension was not used further.

5.3 Lagged meteorological variables

The residuals from the preceding analysis were examined for association with lagged surface meteorological variables by using multiple linear regression methods. The lagged variables were all 24 hour averages. This examination showed that there were moderately strong effects of temperature at lags 1 and 2, and relative humidity and wind speed at lag 1. The model was therefore extended by including these lagged temperature variables linearly in the temperature factor, and by adding lagged relative humidity and wind speed into the corresponding factors.

These additions raised R^2 to 0.7499. The coefficients of lagged temperatures were both negative, with that of the lag 2 temperature three times the size of that of lag 1 temperature, though still only one thirtieth of the coefficient of the same day's temperature. The coefficient of lagged relative humidity was the same sign and size as that of the same day's relative humidity. Similarly, the critical wind speed for lagged wind speed was similar to that for the same day's wind speed. The multiple regression had indicated that neither variable needed to be lagged more than one day.

The residuals were also fitted linearly to lagged upper air variables, in the form of the average of the 00Z sounding for the given day (which is made at 6:00 p.m. on the previous day) and the 12Z sounding for the previous day (made at 6:00 a.m. on the previous day). The t-statistics for these variables, while larger in magnitude than their nominal 95% points, were all smaller than those for the lagged surface variables. The lagged upper air variables were not included in the model.

5.4 Seasonal structure and trend

The residuals from all models were also found to have seasonal dependence. For instance, the residuals from the model described above are plotted against day of year in Figure 16. The graph shows an essentially linear decline over the ozone season, and to model this effect a seasonal term was included in the model. To give a reasonable fit both in the present context and in a model with no meteorological factors, the term was taken to be a short Fourier series, with the annual and semiannual frequencies represented, and with the mean removed. When this term was included as another multiplicative factor, it raised R^2 considerably, to 0.7982. The alternative of an additive seasonal term gave a somewhat higher R^2 of 0.8037.

To explore further the choice between an additive seasonal term and a multiplicative one, lowess was again used to give a nonparametric view. Here ozone was fitted as a function of season and of the fit from the previous (nonseasonal) model. Figure 17 shows the lowess surface, which is not easy to interpret. The seasonal change is larger at fitted values of 90–100 ppb than at low fitted values such as 20–30 ppb, which is consistent with a multiplicative combination of effects. However, there are relatively few data points with high fitted values near the ends of the season. Up to 60–80 ppb, the surface shows a more nearly constant seasonal change, suggestive of an additive combination. The R^2 for the lowess fit was 0.80, essentially the same as that obtained with both the multiplicative seasonal term and the additive version. The residual root mean square for lowess, 8.225 ppb, falls between those for the additive fit (8.203 ppb) and the multiplica-

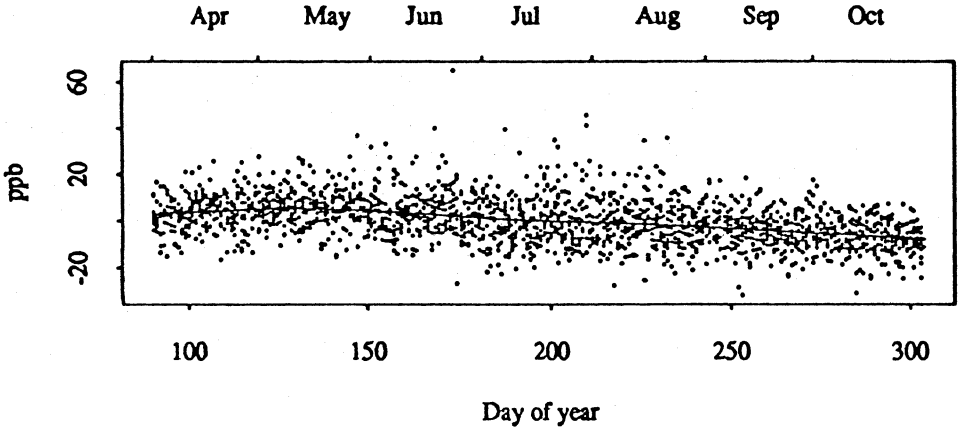


Figure 16: Residuals from meteorological model against day of year.

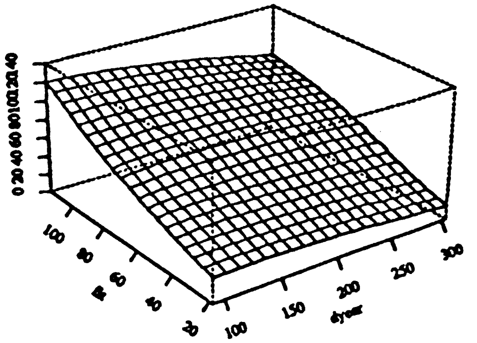


Figure 17: Nonparametric regression surface for ozone against season and the fit from the nonseasonal model.

tive fit (8.317 ppb). Although the choice is not clear cut, the additive form was chosen for the later analyses.

The model is easily extended to estimate a trend in ozone concentrations, of any chosen form. The simplest extension is the incorporation of a factor that is linear in time into the model, as an extra multiplicative factor in the main part of the model. Since possible trends in ozone concentration that are traceable to trends in causative factors such as temperature are accounted for in the model, the fitted coefficient in the trend term represents an estimate of that part of the trend that is not explained by meteorology, or in other words an *adjusted* trend. An unadjusted trend may also be calculated by omitting all meteorological variables. The adjusted trend was found to be -2.7%/decade, while the unadjusted trend was $+5.3\%$/decade. Adding the trend term to the model with meteorological variables resulted in a relatively small rise in R^2 to 0.8042. The R^2 for the model with no meteorology, only trend and seasonality, was 0.3084, showing that meteorological effects as formulated in the model account for around 50% of the variance in ozone concentrations.

The 8%/decade difference between the adjusted and unadjusted trend estimates is large relative to all of the trend standard errors calculated below. It is caused by bias in the misspecified model that omits meteorology, the removal of which is one of the motives for constructing these models.

5.5 Fitted coefficients and standard errors

The trend coefficients discussed in the previous section are of quantitative interest, as are others of the fitted coefficients in the model. It is therefore desirable to associate a standard error with each of them. This is possible in nonlinear least squares fits (Gallant, 1987, for instance), but typically requires the usual assumptions of constancy of variance and lack of correlation in the residuals. These assumptions are easily seen to be false for the residuals from the present model (see Section 6.1). If these requirements are ignored, the standard errors reported by nonlinear least squares programs (Chambers and Hastie, 1992, see `nls`, for instance) are invalid. However, Gallant (1987, Sections 2.1, 2.2) describes methods for correcting the variance estimates for heteroscedasticity and serial correlation in the errors. When combined in a way that allows for serial correlation with a 2–3 day span, these corrections gave the standard errors shown in Table 5. The final form of the model was

$$\begin{aligned}
\text{ozone} \sim\ & \{\text{constant} \\
&\quad + (\text{polynomial in temperature}) \times (\text{function of wind speed})\} \\
& \times (\text{linear function of same day and lag 1 relative humidity}) \\
& \times (\text{linear function of visibility}) \\
& \times (\text{linear function of opaque cloud cover}) \\
& \times (\text{linear function of mean surface wind vector} \\
& \times (\text{linear function of time in years}) \\
& + (\text{seasonal model})
\end{aligned}$$

where the polynomial in temperature is cubic in same-day temperature plus linear in lags 1 and 2 temperature, the function of wind speed is

$$\frac{1}{1+\frac{\text{surface wind speed}}{v_s}+\frac{\text{700 mbar wind speed}}{v_{700}}+\frac{\text{lag 1 wind speed}}{v_l}},$$

and the seasonal model is a linear combination of the cosines and sines of the annual and semiannual frequencies. It may be written specifically as

$$\texttt{o3} \sim (\texttt{mu0} + (\texttt{t0} + \texttt{t1} * (\texttt{maxt} - 60)$$

Table 5: Coefficients in the fitted model, with standard errors computed conventionally and adjusted for heteroscedasticity and serial correlation.

Coefft.	Fitted Value	Conventional		Adjusted	
		Standard error	*t* Value	Standard error	*t* Value
mu0	3.977e+01	1.6149160	24.6262	1.5570528	25.5414
t0	-1.011e+01	6.8563271	-1.4747	6.6209256	-1.5271
t1	3.015e+00	0.4637083	6.5029	0.5347362	5.6391
t2	9.480e-02	0.0141952	6.6780	0.0165989	5.7109
t3	-1.541e-03	0.0002857	-5.3931	0.0003189	-4.8325
t11	-2.531e-02	0.1758384	-0.1439	0.1729924	-0.1463
t12	-8.198e-01	0.1851373	-4.4280	0.2053805	-3.9916
vh	6.055e+00	1.2663745	4.7815	1.5009998	4.0341
vh700	1.358e+01	2.8454335	4.7728	3.7519675	3.6196
vhl	4.686e+00	1.1571911	4.0498	1.4042068	3.3374
r	-3.384e-03	0.0003513	-9.6329	0.0003664	-9.2342
rl	-2.044e-03	0.0003430	-5.9603	0.0003769	-5.4244
op	-6.976e-04	0.0001289	-5.4120	0.0001417	-4.9215
v	-7.658e-03	0.0006391	-11.9820	0.0008258	-9.2733
m.u	5.245e-03	0.0014633	3.5841	0.0015789	3.3218
m.v	9.877e-04	0.0015238	0.6482	0.0016596	0.5952
y	-2.685e-03	0.0011339	-2.3680	0.0014404	-1.8640
a1	-8.028e+00	1.4207495	-5.6504	1.3814916	-5.8110
b1	4.085e+00	0.4924658	8.2946	0.4682407	8.7237
a2	-2.678e+00	0.6501565	-4.1195	0.7043474	-3.8026
b2	-1.124e+00	0.4750715	-2.3656	0.5213316	-2.1557

$$\begin{aligned}
&\qquad\quad +\mathtt{t2} * (\mathtt{maxt} - 60)^2 + \mathtt{t3} * (\mathtt{maxt} - 60)^3 \\
&\qquad\quad +\mathtt{tl1} * (\mathtt{tlag1} - 60) + \mathtt{tl2} * (\mathtt{tlag2}) - 60) \\
&\quad *1/(1 + \mathtt{wspd/vh} + \mathtt{wspd700/vh700} + \mathtt{wlag/vhl})) \\
&*(1 + \mathtt{r} * (\mathtt{rh} - 50) + \mathtt{rl} * (\mathtt{rhlag} - 50)) \\
&*(1 + \mathtt{op} * (\mathtt{opcov} - 50)) \\
&*(1 + \mathtt{v} * (\mathtt{vis} - 12)) \\
&*(1 + \mathtt{m.u} * \mathtt{mean.u} + \mathtt{m.v} * \mathtt{mean.v}) \\
&*(1 + \mathtt{y} * (\mathtt{year} - 1985)) \\
&+ \mathtt{a1} * \mathtt{cos}(2 * \mathtt{pi} * \mathtt{year}) + \mathtt{b1} * \mathtt{sin}(2 * \mathtt{pi} * \mathtt{year}) \\
&\quad + \mathtt{a2} * \mathtt{cos}(4 * \mathtt{pi} * \mathtt{year}) + \mathtt{b2} * \mathtt{sin}(4 * \mathtt{pi} * \mathtt{year}). \quad (6)
\end{aligned}$$

The seasonal cosine and sine terms were in fact deviations from their respective means.

Most of the t-ratios for variables in the model are at least 2 in absolute value. Aside from the trend coefficient y, the only parameters that have t-ratios lower than 3 are associated with other parameters with higher t-ratios. Thus all meaningful groupings of parameters have a high level of statistical significance.

6 Discussion of the Model

6.1 Quality of the fitted model

The root mean square residual from the fitted model is 8.195 ppb. The lower and upper 2.5% points of the residuals are -14.6 ppb and 17.6 ppb, respectively, while the quartiles are -5.0 ppb and 4.3 ppb. Thus model predictions differ from the actual values by up to ± 5 ppb about half the time, and by up to ± 16 ppb about 95% or the time. Figure 18 is a quantile-quantile plot of the residuals against the Gaussian distribution. The points would fall on a straight line if the distribution of the residuals were exactly Gaussian in shape. In the figure, the behavior is roughly linear up to a Gaussian quantile of around 2, or in other words for the lower 97.5% of the distribution. Above this level, the points fall above the extrapolation of the straight line, meaning that the quantiles of the residuals are progressively higher than the corresponding quantiles of the Gaussian distribution that would fit the lower part of their distribution. The highest residual, 65 ppb, is at around the 8σ level.

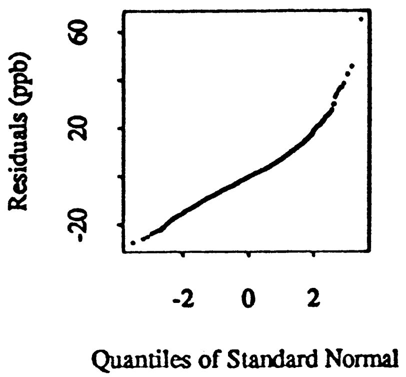


Figure 18: Quantile-quantile plot of the residuals from the fitted model against the Gaussian distribution.

Table 6 shows the root mean square residual and the first lagged correlation coefficient, by month. The standard error of a month's root mean square is around 6% of the observed value, and therefore ranges from 0.4 to 0.6. The seasonal variation in the root mean square is therefore highly significant. The standard error of each correlation coefficient is around 0.06; thus the correlations in June and July are not significantly different from zero, while the higher correlations are clearly different from these mid-summer values.

The seasonal rise in root mean square residual parallels the rise in mean levels associated with summer temperatures. If the extra variability is solely caused by the higher mean value, it might be eliminated by reexpressing ozone concentrations on a "variance-stabilizing" scale. However, it may be caused partly by differing meteorological variability in the different seasons, and not simply by the change in mean levels. This issue may be addressed by exploring the dependence of the magnitude of the residuals on the fitted values and the season. Figure 19 shows the regression surface that results from using lowess to fit such a model nonparametrically. It appears that there is a strong relationship between the magnitude of the residuals and the fitted values early in the season, but that it becomes weaker as the season progresses. The decrease in residual magnitude at high fitted values late in the season is presumably spurious, there being relatively

Table 6: Root mean square residual and one day lagged correlation coefficient, by month

Month	Rms	Number in rms	Correlation	Number in correlation
4	6.492	317	0.28475	295
5	7.203	336	0.25152	320
6	9.205	318	0.06903	298
7	10.404	331	0.06727	312
8	8.975	331	0.22361	315
9	7.467	318	0.42167	298
10	6.515	332	0.28397	313

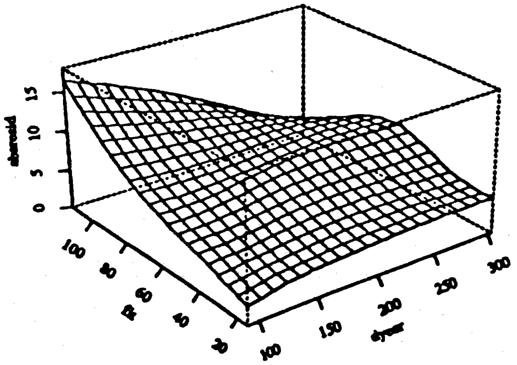


Figure 19: Nonparametric regression surface for magnitude of residuals against season and fitted value.

few high fitted values at these dates.

It should be noted that the adjusted standard errors tabulated in Section 5.5 (page 34) are valid in the presence of heteroscedasticity and serial correlation of the form displayed in this section. In particular, they remain valid when the serial correlations are not stationary, provided they are short term in nature. Standard errors that are valid for longer term correlation are discussed in Section 6.5 (page 46). However, ordinary least squares parameter estimates may be inefficient in these cases. Efficiency can be partially restored by reexpressing the ozone concentrations on a variance-stabilizing scale. In the present case, Figure 19 suggests that the magnitude of the residuals is roughly proportional to the fitted values, meaning that the logarithms of ozone concentrations would have more nearly constant variance.

It must be recognized, however, that the model is only an empirical approximation to the actual physical/chemical mechanism whereby meteorological variables influence ozone concentrations. When a fitting procedure such as least squares is used to fit the model, it produces estimates of the values of the parameters that make the empirical model as close as possible to the actual mechanism, in the least squares sense. If the model were fitted differently, for instance by least squares on the logarithmic scale, the fitted parameters would be estimates of different parameter values, namely those that make the model best approximate the actual mechanism on the new scale. In the case of a logarithmic reexpression, the effect is to estimate a model that fits the data better at low concentrations, but worse at high concentrations. In other words, rexpression may produce more efficient parameter estimates, but they may be estimates of less appropriate parameter values. It is for this reason that all the fitting described in this report has been carried out on the original scale, with standard errors computed in a way that makes them valid in the face of heteroscedasticity and serial correlation, rather than on reexpressed data.

6.2 Predicted and adjusted percentiles

One aspect of model performance is how well it predicts the highest levels of ozone. To address this question the model predictions of 95% points by season were calculated. The model prediction for a given day consists of a probability distribution, whose mean is the predicted value for that day. The distribution was taken to be the empirical distribution of the residuals from the model (6), centered at the predicted value. These prediction distributions were averaged within years (actually within ozone seasons, April–October of each year). Figure 20 shows the actual 95th percentiles and those of the yearly averaged prediction distributions. The model percentiles track the actual percentiles well, with the largest deviation occurring in 1987. Cox and Chu (1992) carried out a similar exercise for the network maximum values rather the network typical value, as here. Their Figure 3 shows similar model ability to track the observed 95th percentile, but with somewhat poorer agreement, presumably reflecting the noisier character of the network maximum *versus* the network typical value.

To explore the effect of the residual distribution on the prediction of percentiles, the calculation of predicted percentiles was repeated using a Gaussian shape for the prediction distribution with a standard deviation of 8.195 ppb. This gave essentially the same predicted quantiles, presumably because of the close agreement between the percent points of the distribution of the residuals and those of a Gaussian distribution, up to the 97.5% level.

One of the major uses of a model such as that discussed above is to allow for the effect of year-to-year variations in the explanatory variables. Adjustment of overall trends in ozone levels was described in Section 5.4. An individual day's ozone value may also be adjusted, by adding the predicted value for an adjusted set of meteorological variables to the residual for the actual day. Figure 21

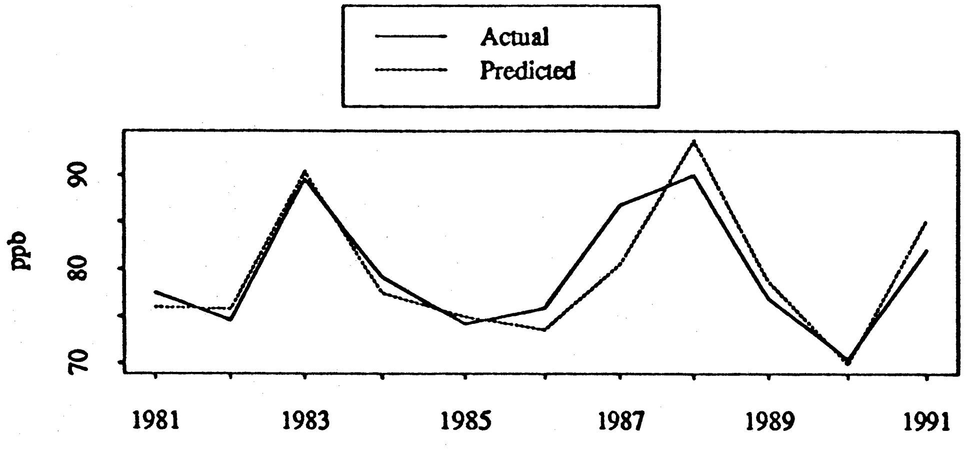


Figure 20: Actual and predicted 95th percentiles of ozone by year.

shows the 95th percentiles by year of such adjusted ozone values, as well as the actual percentiles. As in Cox and Chu (1992), the meteorological variables in the model (6) were adjusted linearly season by season, so that the mean and variance from June to September matched those for all 11 years combined.

The adjusted percentiles show somewhat less year-to-year variation than the actual percentiles, indicating that much of the variation was associated with meteorological variability. Again, one year is exceptional: in 1987 the adjusted percentile is very close to the actual value. However, since the 95th percentile is determined by the values of the largest 10 or so residuals, the exceptional behavior in 1987 may be due simply to sampling variability.

Cox and Chu (1992) also constructed adjusted ozone percentiles, for the network maximum value. Their Figure 6 shows much less variability around a downward linear trend than does our Figure 21, despite the fact that Cox and Chu's calculation is for the 99th percentile of the network maximum, rather than the 95th percentile of the network typical value. However, their construction does not involve the observed residuals, as was done here, but is closer to the calculation of *predicted* percentiles for the adjusted meteorology.

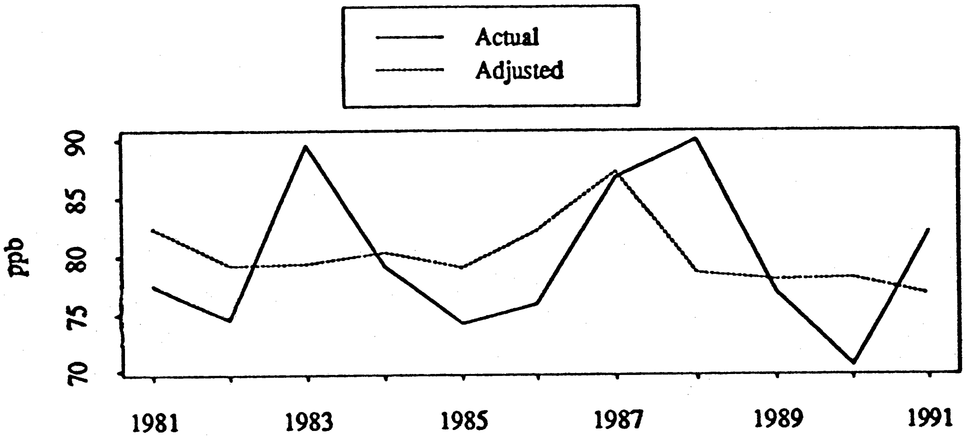


Figure 21: Actual and adjusted 95th percentiles of ozone by year.

6.3 Interpretation of model components

The parameters of the model (6) have various interpretations. As always, these interpretations are only suggestive of actual physical or chemical causes of ozone variations.

`mu0` : Predicted ozone level when temperature × wind speed interaction term is zero (that is, high wind speed or temperature = zero of polynomial, around 60°F), and all other variables are at their centering values. 40 ppb.

`t0, t1, t2, t3, t11, t12` : Coefficients of a cubic polynomial in maximum temperature and lagged temperature, which added to `mu0` gives the predicted ozone level for a given temperature at zero wind speed, and all other variables at their centering values. The polynomial plus `mu0` is shown in Figure 22 (horizontal line indicates `mu0`). In constructing the figure, lagged temperatures were taken as equal to same day temperature.

`vh, vh700, vhl` : Critical speeds for surface wind, 700 mb wind, and lagged surface wind, respectively, at which wind speed factor drops to one half, with the other wind speeds at zero. 6 m/s, 14 m/s, 5 m/s.

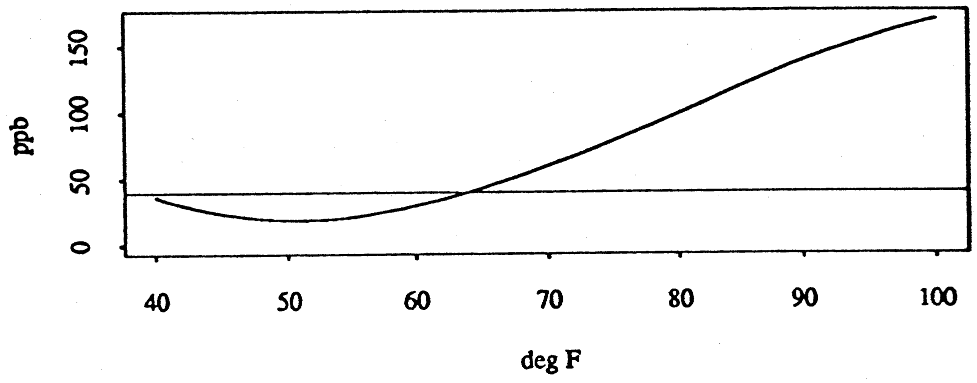


Figure 22: Fitted polynomial effect of temperature at zero wind speed.

`r, rl` : Effects of relative humidity and lagged relative humidity. $-0.003\%^{-1}$, $-0.002\%^{-1}$. The factor drops from 1.27 to 0.73 as both humidities rise from 0% to 100%.

`op` : Effect of opaque cloud cover. $-.0007\%^{-1}$. The factor drops from 1.03 to 0.97 as opaque cloud cover rises from 0% to 100%.

`v` : Effect of visibility. $-.008\text{km}^{-1}$. The factor drops from 1.09 to 0.91 as visibility rises from 0km to 24km.

`m.u, m.v` : The relative effect of the 24hr mean wind vector is its vector (inner) product with the vector $(\texttt{m.u}, \texttt{m.v}) = (0.0052, 0.0010)(\text{m/s})^{-1}$. Maximum when wind is from the south. The effect of a 5 m/s wind varies from 1.03 to 0.97 as the wind direction changes from south to north.

`y` : Trend parameter, $-0.0027\text{yr}^{-1} = -2.7\%/\text{decade}$.

`a1, b1, a2, b2` : Coefficients of the annual and semiannual cosines and sines. The seasonal term is shown in Figure 23. The location of the maximum at May 1 rather than at April 1 may reflect the small number of terms in the Fourier series rather than a true increase from April 1 to

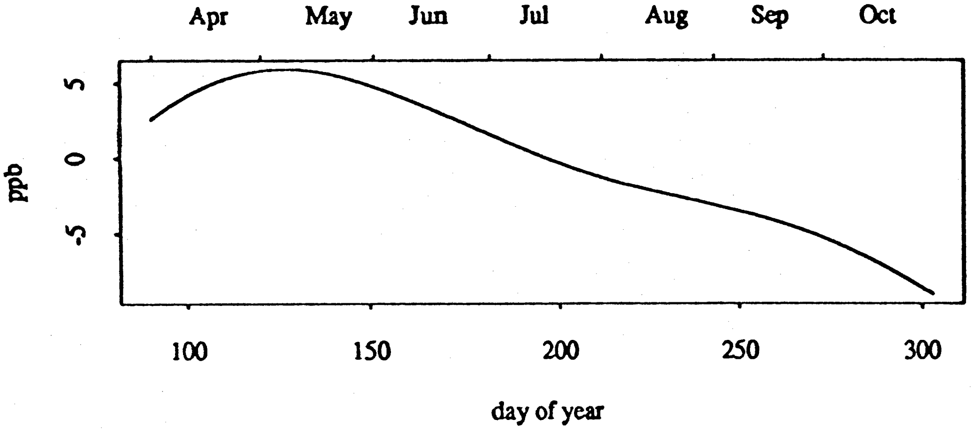


Figure 23: Fitted seasonal effect.

April 30. The fitted values at the start of the ozone season are influenced by the observed values at the end of the season, because of the periodicity of this component. However, a similar rise through the month of April is perceptible in Figure 16 (page 31).

The need for a seasonal term in the model indicates that there are other determinants of surface ozone concentrations than those incorporated in the model, with systematic seasonal variations. The observation that the seasonal term fits better in an additive form rather than as a further multiplicative factor suggests that it reflects variations in a source of ozone rather than in precursors of ozone. However, the difference in quality of fit was not clear cut (R^2 of 0.8037 rather than 0.7982), and does not amount to more than a suggestion.

Altshuller (1987) reviewed information on background tropospheric ozone concentrations, and stated that during the warmer months it appears to be in the range 10–20 ppb. Although Altshuller did not quantify the seasonal variability in background tropospheric ozone, springtime maxima in oxidants and ozone in particular have been observed by Wakamatsu, Uno, Ueda and Uehara (1989), for example. Thus the effect shown in Figure 23 is qualitatively consistent with seasonal variations in background tropospheric ozone concentrations.

The model (6) contains no term that may be interpreted as an estimate of the average background ozone level. The coefficient `mu0` is independent of temperature and wind speed, but still multiplies the factors associated with the remaining meteorological variables. An intercept may be added to the model by extending the "seasonal" part to

$$\begin{aligned}&\texttt{mu1} + \texttt{a1} * \texttt{cos(2} * \texttt{pi} * \texttt{year)} + \texttt{b1} * \texttt{sin(2} * \texttt{pi} * \texttt{year)}\\ &+ \texttt{a2} * \texttt{cos(4} * \texttt{pi} * \texttt{year)} + \texttt{b2} * \texttt{sin(4} * \texttt{pi} * \texttt{year)}\end{aligned}$$

(recall that in (6) the cosine and sine terms had means of zero, by construction). The fitted value of `mu1` is 10.5 ppb, and the seasonal coefficients are essentially unchanged. Thus the background level of ozone is estimated as the curve in Figure 23 (page 42), offset by +10.5 ppb, in reasonable agreement with Altshuller's statement. However, the standard error of this estimate is 5.6 ppb, and it is therefore poorly quantified. The estimates of the coefficients `mu0` and `mu1` are highly (and negatively) correlated; their sum is well estimated, but neither coefficient is individually. In particular the estimated value of `mu1` is barely significantly different from zero, and therefore this extended seasonal model has not been used elsewhere.

6.4 Cross validation

As with any model containing many parameters, the possibility exists that (6) is over-parameterized. This may be explored by refitting the model to subsets of the data, and studying

- the quality of predictions from the refit model to the remainder of the data (*cross validation*), and
- the variability of the parameter estimates in the refit models (*jackknifing*).

In the present case, where the focus is on the possible effects of interannual variations in conditions, it is appropriate to construct the subsets based on years (or "ozone seasons"). This was done by leaving out one year at a time.

Table 7 gives some results. The line for a given year gives the number of days of data used in the model for that year, and the mean and root mean square error when the model refit to the remainder of the data is used to predict that year. This statistic is a grouped version of the PRESS statistic discussed by Cook

Table 7: Predicted residual mean and root mean square by year.

Year	Number of days	Mean	Root mean square
1981	204	-0.1860	8.878
1982	208	-1.2221	8.371
1983	210	-0.6555	9.776
1984	208	-0.1020	8.490
1985	209	0.7356	7.629
1986	206	-0.4128	8.371
1987	209	1.4206	8.628
1988	212	-0.7896	7.162
1989	208	2.3823	8.240
1990	204	0.9323	7.372
1991	205	-2.9085	8.088

and Weisberg (1982, Section 2.2.3). The overall root mean square of the cross-validated prediction errors is 8.302 ppb, which compares very favorably with the root mean square residual of the fit to all the data, 8.195 ppb.

There is some variability from year to year in Table 7. However, it is not clear how it relates to the characteristics of either ozone levels or meteorology. For instance, both 1983 and 1988 experienced high ozone levels (see Figure 20, page 39, for example), and yet 1983 shows the highest root mean square error and 1988 the lowest. Note also that 1987, while showing the poorest fit in the predicted 95th percentile (see the same figure), has a mean squared error that is exceeded in two other years. The possibility that these variations were caused solely by sampling error was explored by comparing the values in the second column of Table 7 with quantiles of a reference distribution. Some care has to be taken in constructing the reference distribution, because of the following features of the context.

- There is variability in sample size from year to year.
- The residuals from the fitted model are noticeably nonGaussian (see Figure 18, page 36).
- Each line in Table 7 summarizes an entire season, and the residuals do not have the same distribution across the season (see Figure 19, page 37).

To allow for these features, the reference distribution was constructed by bootstrap methods (Efron, 1979; Wu, 1986), in the following steps.

1. Select a season length randomly from the first column of Table 7.
2. Construct a random season as a sample of the dates from April 1 to October 31, without replacement, with size equal to the season length selected in step 1.
3. For each date in the random season, choose a residual at random from the population of residuals labeled by that date. Note that there are at most 11 such residuals.
4. Calculate the standard deviation of the resulting random season of residuals.

This procedure was carried out 250 times to produce simulation estimates of the quantiles of the reference distribution. It should be noted that the reference distribution was constructed from the residuals from the model when fitted to all of the data, rather than from "leave one out" predicted residuals. However, these were very similar in overall magnitude, and the substitution should not cause a noticeable bias.

Figure 24 shows the ordered values from the second column of Table 7, graphed against the bootstrap quantiles. The graph suggests that at most the highest observed value, that for 1983, is more extreme than could be explained by sampling variability, and by only around 0.5 ppb.

6.5 Jackknifing

The "leave out one year" parameter estimates discussed in Section 6.4 may be used to construct "pseudo-values" (Tukey, 1958) from

$$\text{pseudo-value} = Y(\text{estimate from all data}) - (Y - 1)(\text{leave one out estimate}),$$

where Y is the number of years, 11. The pseudo-values may then be treated as a sample of values estimating the given parameter. Their mean is used as a less biased "jackknifed" estimate of the parameter, and the standard error of

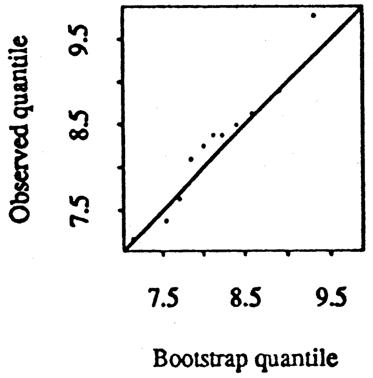


Figure 24: Quantile-quantile plot of predicted root mean square residual by year, against a bootstrap reference distribution.

the mean provides a standard error for either the original parameter estimate or the jackknifed estimate. This standard error is valid in the presence of arbitrary variances and covariances of the model residuals, provided only that they are uncorrelated across years. Miller (1974) discussed the use of the jackknife for linear regression models, and proposed its use also for nonlinear models. Hinkley (1977) showed that in general a weighted jackknife performs better for unbalanced situations such as regression models, and other modifications were suggested by Wu (1986) for both linear and nonlinear models. However, in the present case each year should be reasonably similar to the others, which suggests that the extra complication of the weighted jackknife outweighs any performance advantage.

The jacknifed estimates and standard errors are shown in Table 8. The jackknifed parameter estimates differ very little from the overall values shown in Table 5 (page 34). The jackknifed standard errors are also generally similar to those in Table 5, the most notable difference being the increase in the standard error of the trend coefficient from 1.4 %/decade to 3.4 %/decade. The associated change in the $|t|$-statistic from nearly 2 (allowing for heteroscedasticity and short-term correlation) to less than 1 suggests that the information about trend in the data is overstated in the standard errors in Table 5. However, the standard errors in Table 8 are based on only 10 degrees of freedom, and the effects of sampling variations must not be neglected.

Figure 25 shows the ratios of the two sets of standard errors, ordered and graphed against the quantiles of the χ distribution with 10 degrees of freedom. If there were no long-term or interannual effects, *and* if the dispersion matrix of the parameter estimates were diagonal, the points would be expected to lie close to the solid diagonal line. Correlations among the parameter estimates would tend to reduce the slope of the points, but this effect is expected to be small. The highest point in the figure, corresponding to the trend coefficient, falls far enough above the line to preclude the possibility that it is the result of sampling variability. The jackknife standard error is therefore preferred for inferences about the trend coefficient.

The dotted line in Figure 25 has a slope of 1.2, and provides a good compromise match to the remaining points. This suggests that the jackknife standard errors are estimating values up to 20% larger than the adjusted standard errors of Table 5. The adjusted standard errors multiplied by 1.2 should therefore give rise to conservative inferences about all parameters other than the trend, at the same time being less sensitive to sampling variability than the jackknife standard errors.

Table 8: Jackknifed estimates and standard errors.

Coefft.	Fitted Value	Standard error	*t* Value
mu0	3.980e+01	1.6932387	23.5035
t0	-1.042e+01	8.0915148	-1.2881
t1	3.050e+00	0.7471219	4.0830
t2	9.606e-02	0.0267512	3.5908
t3	-1.557e-03	0.0003707	-4.2002
t11	-2.382e-02	0.2129552	-0.1119
t12	-8.322e-01	0.3544919	-2.3477
vh	6.018e+00	2.0013063	3.0071
vh700	1.353e+01	4.0252847	3.3612
vh1	4.679e+00	1.7690150	2.6447
r	-3.385e-03	0.0002265	-14.9437
r1	-2.046e-03	0.0003206	-6.3810
op	-6.951e-04	0.0001320	-5.2658
v	-7.640e-03	0.0009879	-7.7332
m.u	5.284e-03	0.0011824	4.4688
m.v	9.919e-04	0.0015540	0.6383
y	-2.646e-03	0.0034115	-0.7757
a1	-8.024e+00	1.2199726	-6.5774
b1	4.084e+00	0.5528235	7.3876
a2	-2.684e+00	0.4108871	-6.5331
b2	-1.129e+00	0.3694495	-3.0560

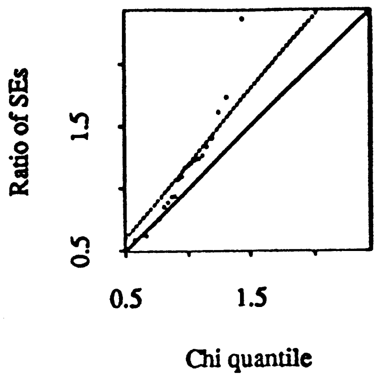


Figure 25: Quantile-quantile plot of ratio of jackknifed standard errors to adjusted standard errors, against the χ distribution with 10 degrees of freedom, with lines of slope 1 and 1.2.

6.6 Network maximum

The modeling effort described above dealt with only the daily network *typical* value. The median polish analysis that gave rise to these network typical values was also used to construct a network *maximum*. Fitting the same form of model to the network maximum gave $R^2 = 0.6731$ and a root mean square residual of 15.611 ppb. The fitted coefficients are shown in Table 9. The major changes from the model for the network typical value are as follows.

`mu0` : Increased from 40 ppb to 59 ppb.

`t0, t1, t2, t3, t11, t12` : The same day temperature coefficients are generally larger. This is to be expected, as they are absolute effects. However, the lagged temperature effects are smaller, and both are now nonsignificant. The polynomial plus `mu0` is shown in Figure 26 (horizontal line indicates `mu0`). As in constructing Figure 22, lagged temperatures were taken as equal to same day temperature.

`vh, vh700, vhl` : All critical wind speeds are lower, meaning that the effects of high temperature are moderated more rapidly by increasing wind speed

Table 9: Coefficients in the fitted model for the network maximum, with standard errors computed conventionally and adjusted for heteroscedasticity and serial correlation.

Coefft.	Fitted Value	Conventional		Adjusted	
		Standard error	*t* Value	Standard error	*t* Value
mu0	5.851e+01	2.704e+00	21.6400	2.817e+00	20.7682
t0	-8.222e+00	1.526e+01	-0.5389	1.510e+01	-0.5444
t1	5.540e+00	1.370e+00	4.0449	1.499e+00	3.6964
t2	1.626e-01	3.900e-02	4.1699	4.318e-02	3.7664
t3	-1.615e-03	6.316e-04	-2.5578	6.969e-04	-2.3179
tl1	-1.019e-01	4.806e-01	-0.2120	4.854e-01	-0.2099
tl2	-7.087e-01	4.170e-01	-1.6994	4.568e-01	-1.5514
vh	2.908e+00	8.012e-01	3.6292	9.122e-01	3.1873
vh700	1.167e+01	3.791e+00	3.0774	4.343e+00	2.6863
vhl	2.965e+00	1.013e+00	2.9261	1.249e+00	2.3746
r	-1.796e-03	4.579e-04	-3.9214	5.258e-04	-3.4150
rl	-2.048e-03	4.456e-04	-4.5967	5.315e-04	-3.8537
op	-6.176e-04	1.674e-04	-3.6893	1.783e-04	-3.4647
v	-7.002e-03	8.283e-04	-8.4542	9.134e-04	-7.6662
m.u	2.574e-03	1.863e-03	1.3814	1.915e-03	1.3438
m.v	-1.328e-02	1.916e-03	-6.9274	2.089e-03	-6.3554
y	-9.492e-03	1.441e-03	-6.5882	1.828e-03	-5.1939
a1	-4.235e+00	2.688e+00	-1.5758	2.912e+00	-1.4543
b1	6.380e+00	9.344e-01	6.8283	1.091e+00	5.8466
a2	-1.441e+00	1.240e+00	-1.1623	1.401e+00	-1.0289
b2	2.531e-01	9.072e-01	0.2790	1.075e+00	0.2354

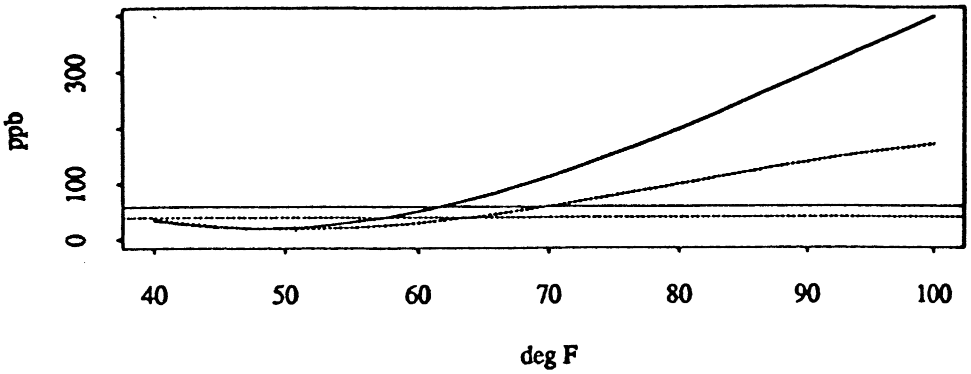


Figure 26: Fitted polynomial effect of temperature at zero wind speed for network maximum. Dotted line shows values for network typical value, shown in Figure 22.

for the network maximum than for the network typical value.

`r, rl, op, v` : The proportional effects of relative humidity, opaque cloud cover, and visibility are all lower for the network maximum than for the network typical value.

`m.u, m.v` : The proportional impact of the wind vector is now greatest when winds are from the north, and is nearly three times as large as for the network typical value.

`y` : The estimated trend parameter for the network maximum is -0.0095yr^{-1} = -9.5%/decade, notably larger than for the network typical value. It is apparently statistically significant, with respect to the standard errors adjusted for heteroscedasticity and serial correlation. Jackknife standard errors have not been computed, but if the same ratio were found as for the trend in the network typical value, the jackknife standard error for the trend would be 4.8%/decade, and the trend would be barely significant.

`a1, b1, a2, b2` : The seasonal term is shown in Figure 27, together with the corresponding curve for the network typical value. The curves are essentially

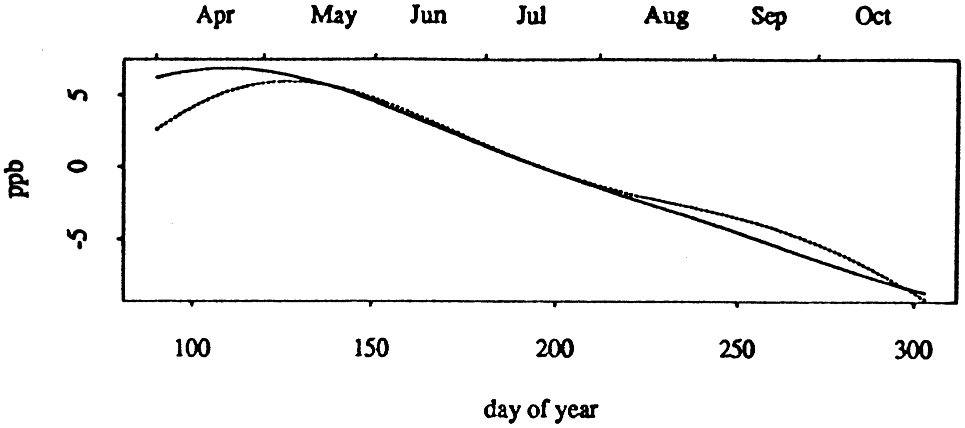


Figure 27: Fitted seasonal effect for network maximum. Dotted line shows values for network typical value, shown in Figure 23.

identical, the largest difference being the somewhat earlier maximum. The close agreement between the seasonal curves contrasts sharply with the difference between the temperature curves (Figure 26), and tends to support the interpretation of this part of the model as the effect of seasonal variations in background ozone concentrations.

Actual, predicted, and adjusted 95th percentiles for the network maximum, corresponding to those in Figures 20 and 21 (pages 39 and 40), are shown in Figures 28 and 29. These figures show similar performance to those for the network typical value. The relatively good performance in 1987 suggests that the problems in the prediction and adjustment of the 95th percentile for the network typical value were due to chance effects.

The results of this section may be compared with those of Cox and Chu (1992), who fitted a Weibull probability model to the same network maximum time series. In the model the logarithm of the Weibull scale parameter was represented as a linear function of selected meteorological variables and linear trend. Parameters were fitted by maximum likelihood. The trend was estimated as $-0.0054/\text{yr} = -5.4\%/\text{decade}$ with a standard error (computed by bootstrapping 3 day sequences) of 3.7%/decade. The difference between Cox and Chu's estimate

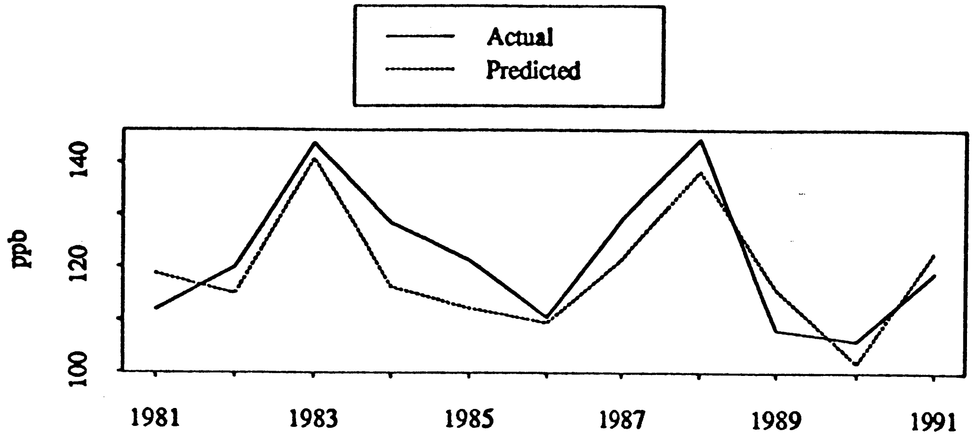


Figure 28: Actual and predicted 95th percentiles of network maximum ozone by year.

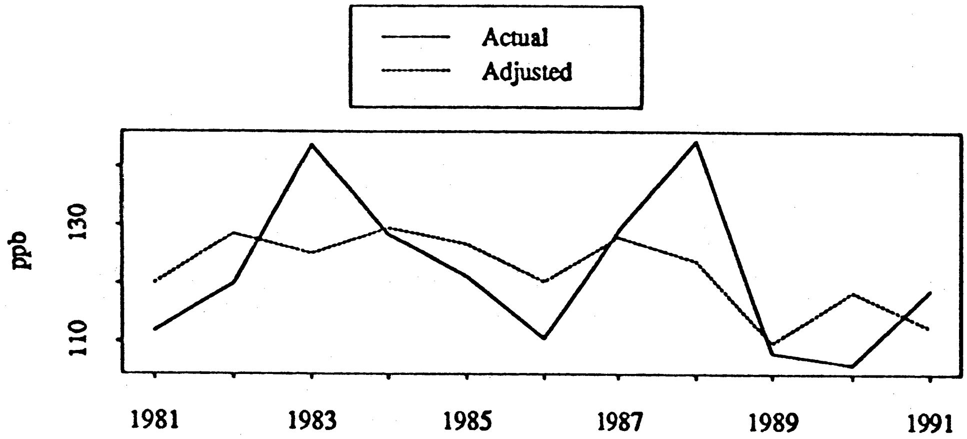


Figure 29: Actual and adjusted 95th percentiles of network maximum ozone by year.

and the one obtained from the present model is around one standard error. They are therefore reasonably consistent with each other, given the differences between the models and the fitting criteria.

7 Conclusions

The daily maximum one-hour average surface ozone concentrations from the 45 stations were highly correlated. A 16-station subnetwork with essentially the same spatial coverage as the whole network showed a dominant principal component that accounted for 78% of the variance, and the corresponding time series was nearly perfectly correlated ($\hat{\rho} = 0.99$) with a simple "typical" value for the network, obtained by median polish.

The network typical value time series shows nonlinear and nonadditive dependence on various meteorological quantities, including individual measurements and constructs (averages). There is also strong seasonal dependence, even after allowing for the effects of the meteorological variables. The dependence can be approximated by a nonlinear parametric model, and when the parameters are fitted by least squares, the model accounts for 80% of the variance of the ozone concentration data. The root mean square residual is 8.2 ppb. The model may be extended to include a trend parameter, which is estimated to be $-2.7\%/\text{decade}$, with a (jackknife) standard error of $3.7\%/\text{decade}$. This represents an estimate of trend *adjusted* for meteorological variability. If the meteorological variables are omitted, the model contains just seasonality and trend, and the trend estimate is found to be $+5.3\%/\text{decade}$, representing the *unadjusted* trend in the surface ozone concentrations.

References

Altshuller, A. P. (1987). 'Estimation of the natural background of ozone present at sruface rural locations', *Journal of the Air Pollution Control Association* 37, 1409–1417.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The New S Language*, Advanced Books and Software. Pacific Grove, California: Wadsworth.

Bloomfield, P. and Steiger, W. L. (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms*, Progress in Probability and Statistics. Boston: Birkhäuser.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Pacific Grove, CA: Wadsworth.

Cleveland, W. S. (1979). 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* 74(368), 829–836.

Comrie, A. C. and Yarnal, B. (1992). 'Relationships between synoptic-scale atmospheric circulation and ozone concentrations in metropolitan pittsburgh, pennsylvania', *Atmospheric Environment* 26B, 301–312.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Monographs on Statistics and Applied Probability. New York: Chapman and Hall.

Cox, W. M. (1992). Personal communication.

Cox, W. M. and Chu, S.-H. (1992). Meteorologically adjusted ozone trends in urban areas: A probability approach, U.S. Environmental Protection Agency, Technical Support Division MD-14, Research Triangle Park, NC 27711.

Efron, B. (1979). 'Bootstrap methods: Another look at the jackknife', *Annals of Statistics* 7, 1–26.

Feister, U. and Balzer, K. (1991). 'Surface ozone and meteorological predictors on a subregional scale', *Atmospheric Environment* 25A, 1781–1790.

Gallant, A. R. (1987). *Nonlinear Statistical Models.* New York: Wiley.

Hinkley, D. V. (1977). 'Jackknifing in unbalanced situations', *Technometrics* 19, 285–292.

Kelly, N. A., Ferman, M. A. and Wolff, G. T. (1986). 'The chemical and meteorological conditions associated with high and low ozone concentrations in southeastern michigan and nearby areas of ontario', *Journal of the Air Pollution Control Association* 36, 150–158.

Miller, R. G. (1974). 'The jackknife—a review', *Biometrika* **61**, 1–15.

National Research Council (1991). *Rethinking the Ozone Problem in Urban and Regional Air Pollution.* Washington, DC: National Academy Press.

Tukey, J. W. (1958). 'Bias and confidence in not quite large samples (abstract)', *Annals of Mathematical Statistics* **29**, 614.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, Massachussetts: Addison-Wesley.

Wakamatsu, S., Uno, I., Ueda, H. and Uehara, K. (1989). 'Observational study of stratospheric ozone intrusions into the lower troposphere', *Atmospheric Environment* **23**, 1815–1826.

Wu, C. F. J. (1986). 'Jackknife, bootstrap and other resampling methods in regression analysis', *Annals of Statistics* **14**, 1261–1295.