NISS

Rural Ozone and Meteorology:Analysis and Comparison with Urban Ozone

Peter Bloomfield, J. Andrew Royle and Qing Yang

Technical Report Number 5
December, 1993

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

Rural Ozone and Meteorology: Analysis and Comparison with Urban Ozone*

Peter Bloomfield J. Andrew Royle
Qing Yang

National Institute of Statistical Sciences

and

Department of Statistics
North Carolina State University

Preliminary Report: 1 December, 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 rural ozone concentrations and meteorology in the rural area surrounding Chicago, using methods similar to an earlier study of urban Chicago ozone concentrations. The results are broadly similar to those of the earlier study; key departures are noted.

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

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

Contents

List of Tables

List of Figures

1 Introduction

In an earlier report, Bloomfield, Royle and Yang (1993) studied the effects of meteorology on surface ozone concentrations in urban Chicago. Data from a network of rural stations surrounding Chicago are analyzed here to determine the relationship of rural ozone levels to meteorology, and to compare that with the corresponding relationship between meterology and urban ozone levels. The earlier report should be consulted for details of the statistical methods.

2 Description of Data

2.1 Ozone data

The ozone data consisted of hourly averages at the 12 stations described in Table 1. The locations of the rural ozone monitoring stations are shown in Figure 1.

Most of the 12 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.

2.2 Meteorological data

Surface and upper air meteorological variables were collected at the two stations shown in Figure 1. The meteorological variables are described in Table 2. The surface observations were made each hour, while the upper air soundings were

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.

AIRS Site ID	Latitude	Longitude	MSA	State	First and Last Dates	
170190004	40.124	88.230	1400	IL	810101	911231
171192007	38.793	90.040	7040	IL	810311	911231
171431001	40.746	89.586	6120	IL	810101	911226
172010009	42.228	89.077	6880	IL	810101	911231
180970042	39.647	86.249	3480	IN	810101	910930
181630013	38.114	87.536	2440	IN	810202	910930
191632011	41.648	90.431	1960	IA	810219	911231
260370001	42.798	84.394	4040	MI	810101	911031
260812001	43.042	85.413	3000	MI	810101	911031
261611001	42.156	83.778	440	MI	820401	911031
550890005	43.321	87.941	5080	WI	810101	910604
551171002	43.669	87.740	7620	WI	811023	910228

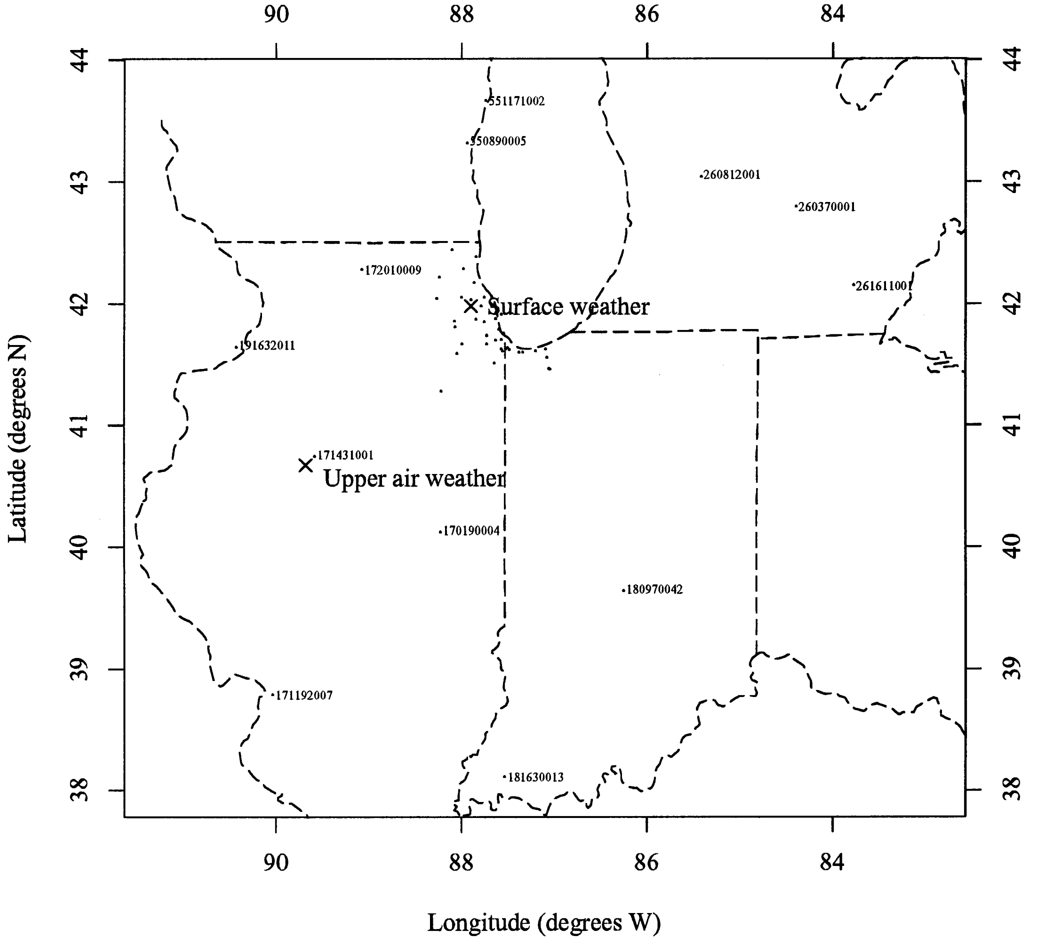

Figure 1: Locations of the ozone monitoring stations with weather stations. Unlabeled dots represent the urban ozone monitoring stations.

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°	

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.

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 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 2 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 3 shows the corresponding diurnal cycle, exp$(\mu + \beta_h)$, which is similar in profile to the curve shown by Bloomfield et al. (1993), but with a slightly higher range. For this station, the root mean square residual in the median polish decomposition (on the logarithmic scale) was 0.47, indicating quite large proportional variations of the observed data around the fitted values.

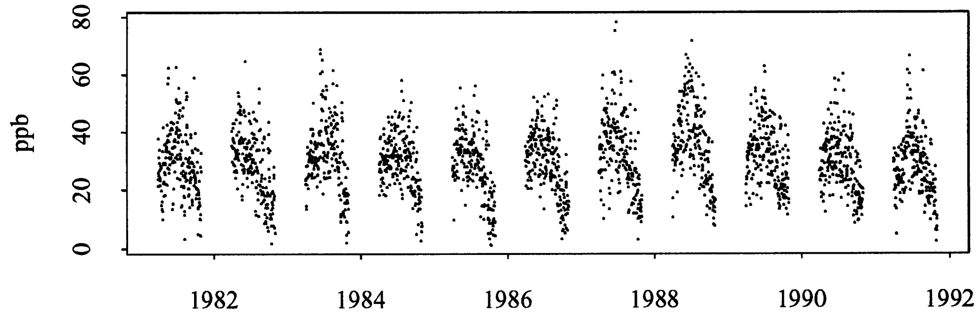

Figure 2: Daily typical values for station 171431001.

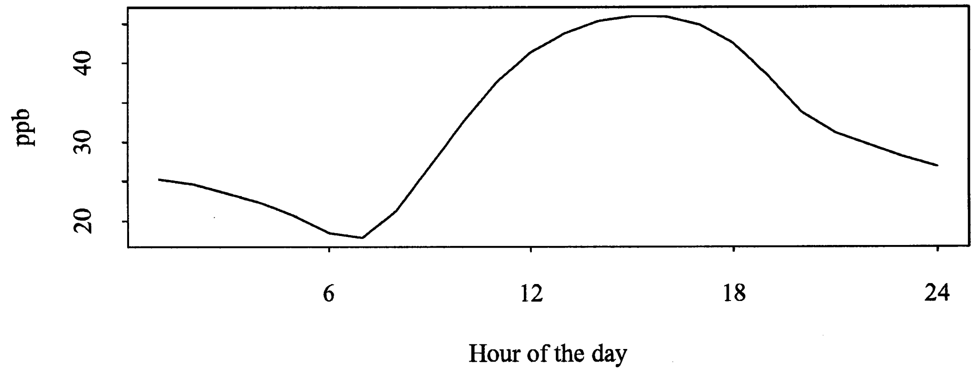

Figure 3: Diurnal cycle for station 171431001.

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 0.47 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 4 shows the histograms of the valid and imputed 2 p.m. observations for the station shown in Figures 2 and 3. Figure 5 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 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 55 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.

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.

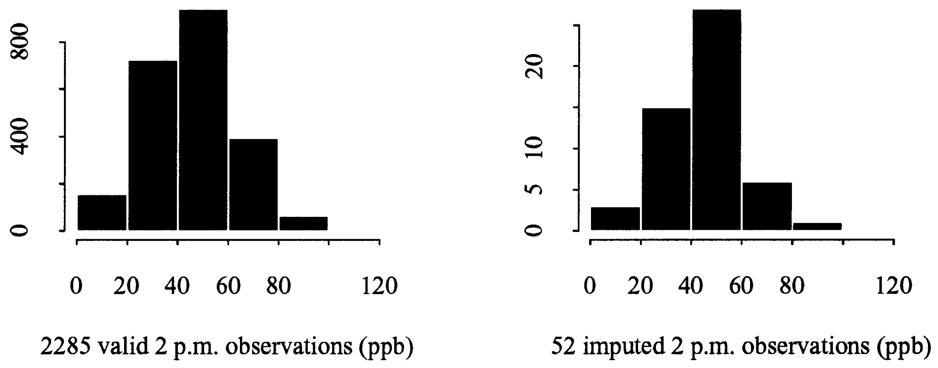

Figure 4: Imputation of 2 p.m. observations for station 171431001.

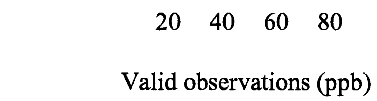

Figure 5: Quantile-quantile plot of imputed 2 p.m. observations for station 171431001 against valid 2 p.m. observations.

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. There were 1371 coincidental days among the 12 rural stations, thus all of the stations were used in the principal components analysis. 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 6 to 9 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 61% of the variance of the 12 station network, is a weighted average of the station values, with weights ranging from 0.25 to 0.32. Stations to the west of Chicago carry the higher weights, while those in the remainder of the network have lower weights. The second and third components account for a further 7.5% and 5.34% of variance, respectively. They may appear to represent gradients across the network, from North to South and from East to West, respectively. However, the loadings of the 2nd component are large on Wisconsin, and those for the 3rd component are large for the Michigan stations, indicating that the components may be representing groups of similar stations. The remaining singular values are not well separated, and account for between 1.9% and 4.4% of the variance of the network each.

The dominant component explains less of the variance than the corresponding component in the urban study (61% *versus* 79%). The ambiguity in the 2nd and 3rd components did not arise in that study, in which the interpretation as gradients seemed clear.

4.2 Network average

The principal components analysis of Section 4.1 showed that most of the variance was associated with a single component. However, it was essentially a weighted average of the 12 stations, which suggested that a similar network average, would

Table 3: Results of principal components analysis for 12 stations.

Singular value	Percent of variance	Cumulative percent
7363.85	60.8565	60.86
913.48	7.5492	68.41
636.18	5.2576	73.66
532.72	4.4025	78.07
436.63	3.6084	81.67
409.12	3.3810	85.06
374.70	3.0966	88.15
331.51	2.7397	90.89
310.30	2.5644	93.46
294.36	2.4326	95.89
266.54	2.2028	98.09
230.95	1.9086	100.00

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

Station	Component				
	1	2	3	4	5
170190004	-0.2789	-0.1971	-0.0295	-0.0832	-0.1989
171192007	-0.3168	-0.3993	-0.3075	0.1393	0.1426
171431001	-0.2772	-0.0956	-0.0418	-0.3767	0.0405
172010009	-0.2578	0.0154	-0.0190	-0.3704	0.0061
180970042	-0.3180	-0.2839	0.0837	0.2103	-0.2184
181630013	-0.3174	-0.3210	-0.1760	0.4326	0.2234
191632011	-0.2845	-0.0794	0.0088	-0.6253	0.1629
260370001	-0.2465	0.1793	0.6013	0.1169	0.4965
260812001	-0.2959	0.2052	0.3123	0.1838	0.1660
261611001	-0.2742	0.0499	0.3963	0.0608	-0.7294
550890005	-0.2983	0.4776	-0.2603	0.0424	0.0137
551171002	-0.2889	0.5454	-0.4249	0.1391	-0.0900

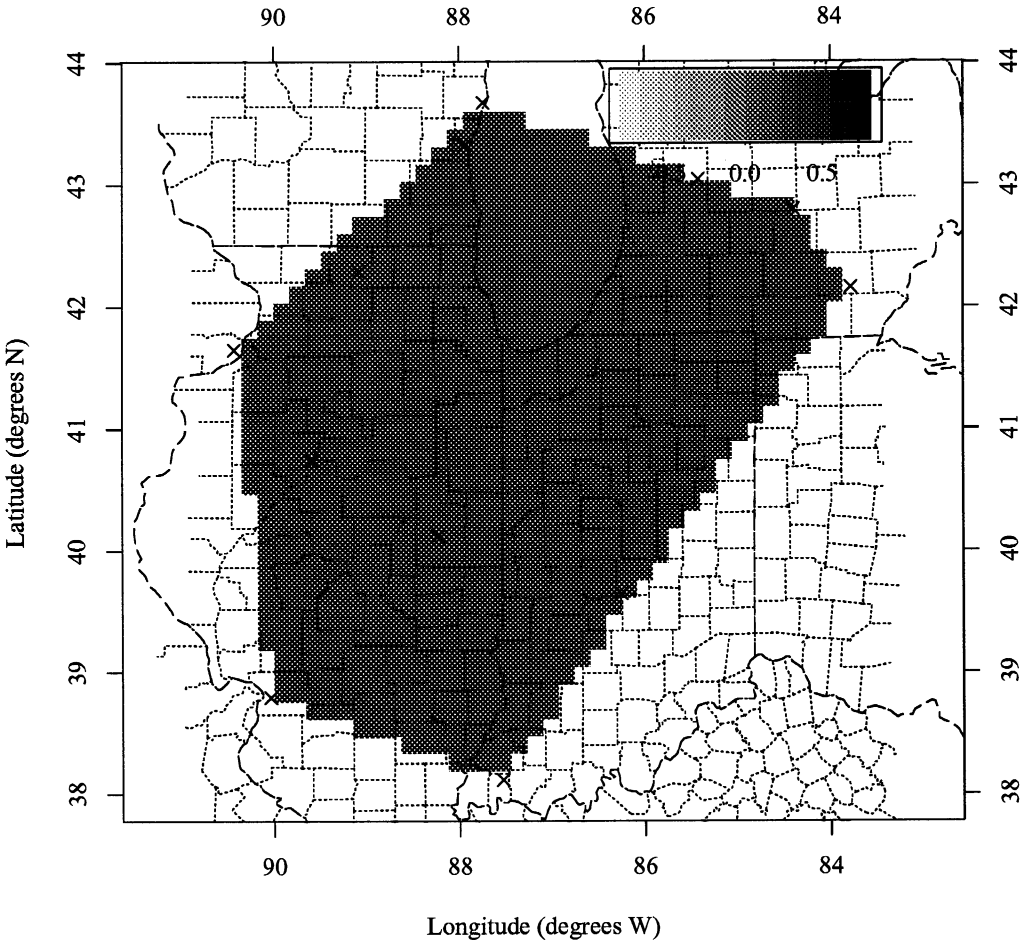

Figure 6: Station loadings for component 1.

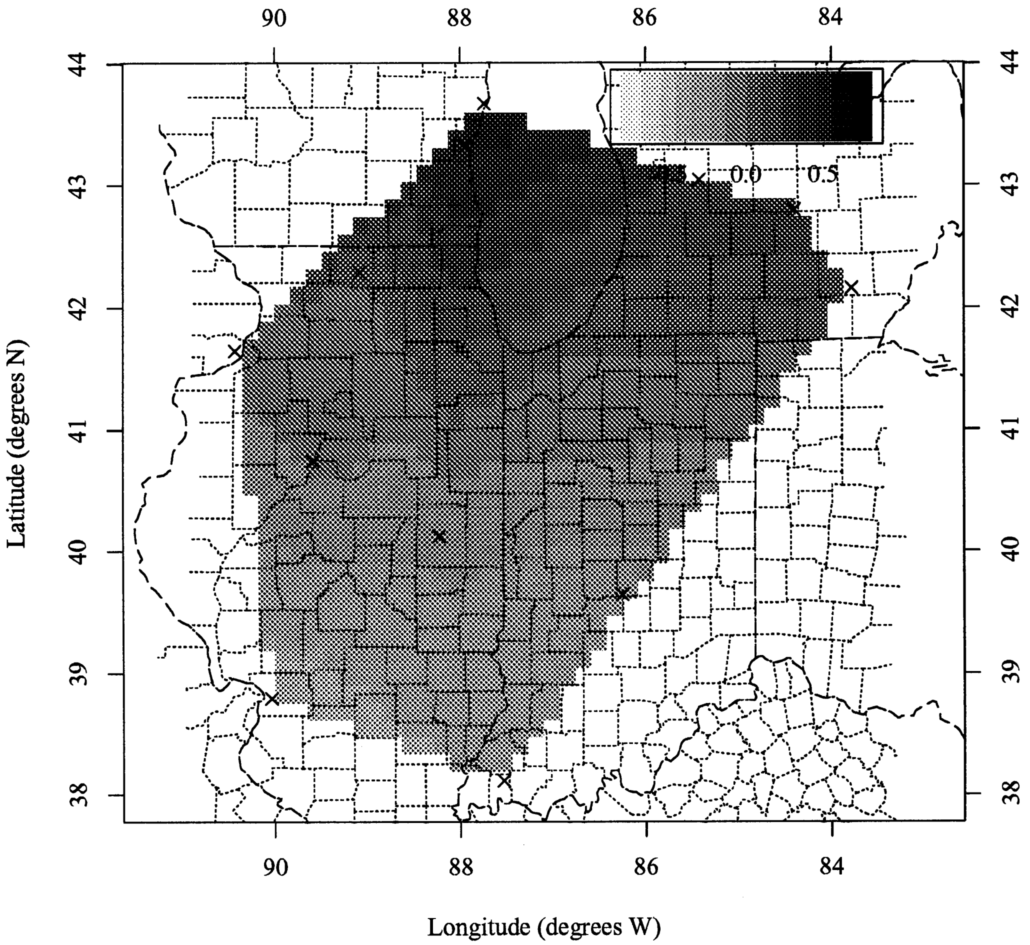

Figure 7: Station loadings for component 2.

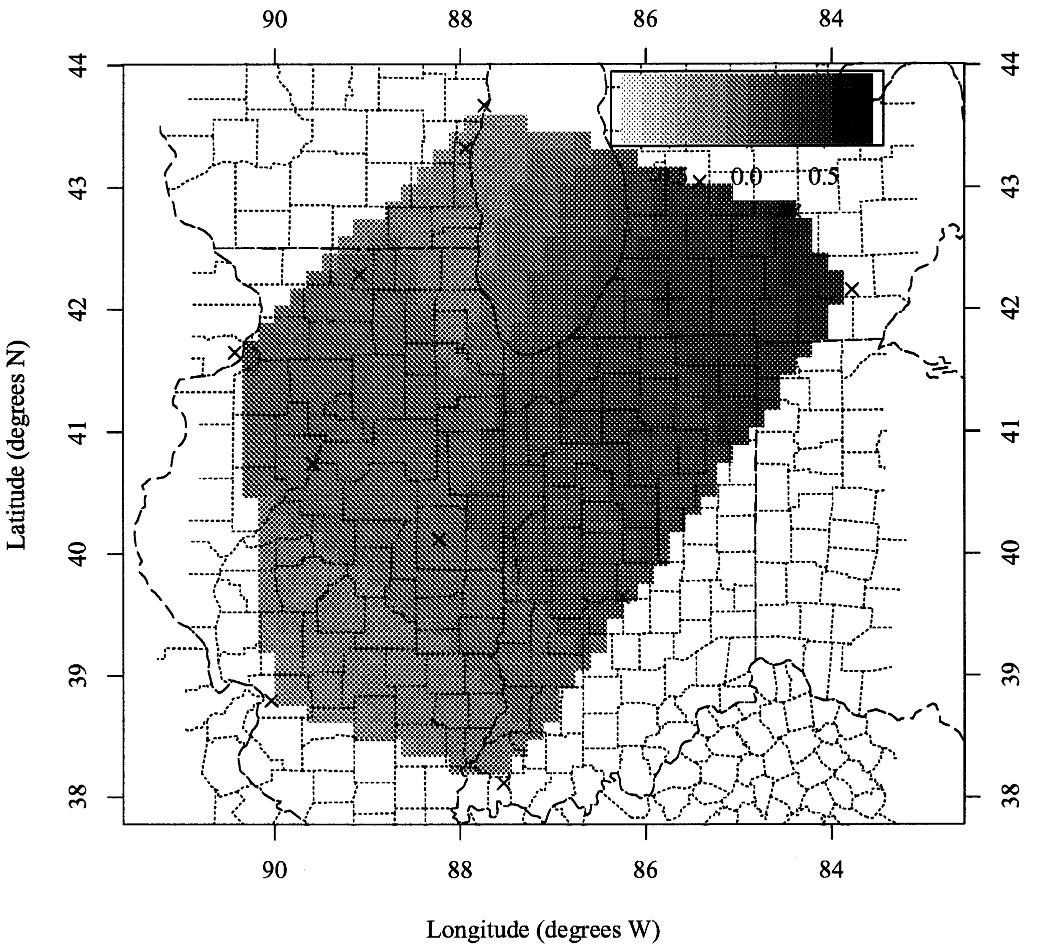

Figure 8: Station loadings for component 3.

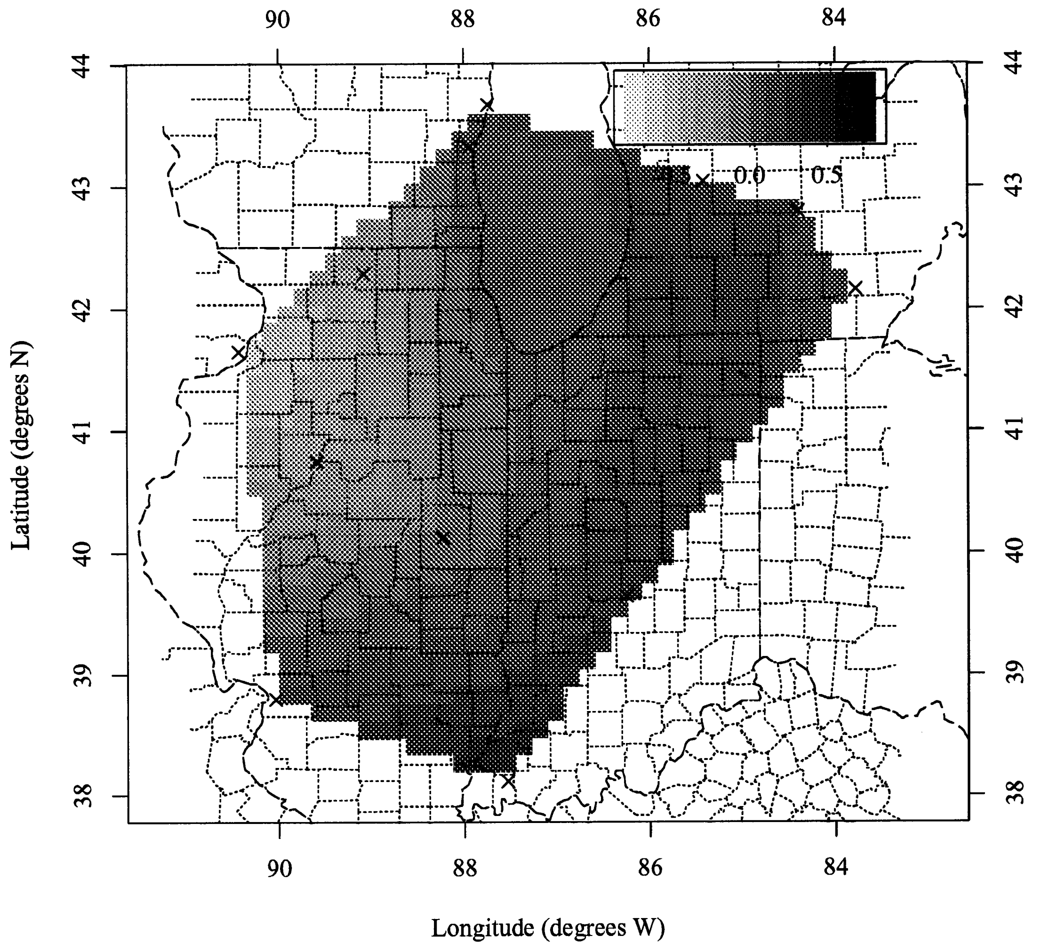

Figure 9: Station loadings for component 4.

provide a 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 0.9857. 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 10 shows a plot of the rural typical value against the urban typical value used in Bloomfield et al. (1993). This plot shows that while the average ozone tends to be slightly higher for the rural network, all of the very large values occur in the urban network. It is evident that the typical values from both networks are highly correlated. The estimated correlation coefficient was 0.9027. Thus it makes some sense to fit the same model to the rural data as was fit to the urban data. Doing this has the advantage of allowing direct comparisons to be made between the two networks for a given meteorological variable or variables.

Figure 11 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 12 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 13 shows the surface for ozone against temperature and relative humidity. The surface which is very similar to that of for the urban data, 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

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

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 does not increase the R^2 in the fourth decimal place ($R^2 = 0.5541$), suggesting that order 3 is adequate. Fitting equation (2) by nonlinear least squares leads to $R^2 = 0.602$. The dimensionless function of relative humidity is estimated to be $1 - 0.00405\%^{-1} \times (\text{relative humidity} - 50\%)$, which decreases from 1.203 to 0.798 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 13, which was 0.61; the equivalent number of parameters was 10.2. This indicates that the parametric model, with only 5 parameters, performs very nearly as well

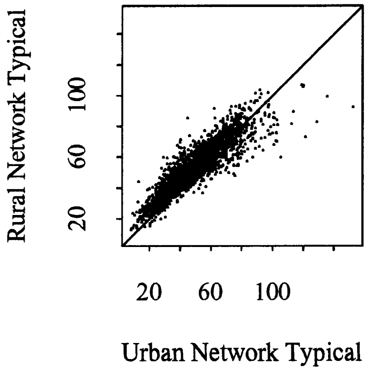

Figure 10: Plot of urban typical value against rural typical value with a line of slope 1.

as the nonparametric one.

The corresponding regression surface for ozone against temperature and wind speed is shown in Figure 14. 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 slight *rise* in ozone at low temperatures. At high temperatures, ozone levels are shown as slightly *decreasing* and then rising again as wind speed increases above 8 m/s, but this portion of the surface is determined by very few data points.

The behavior shown in Figure 14 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 20.36 m/s, and

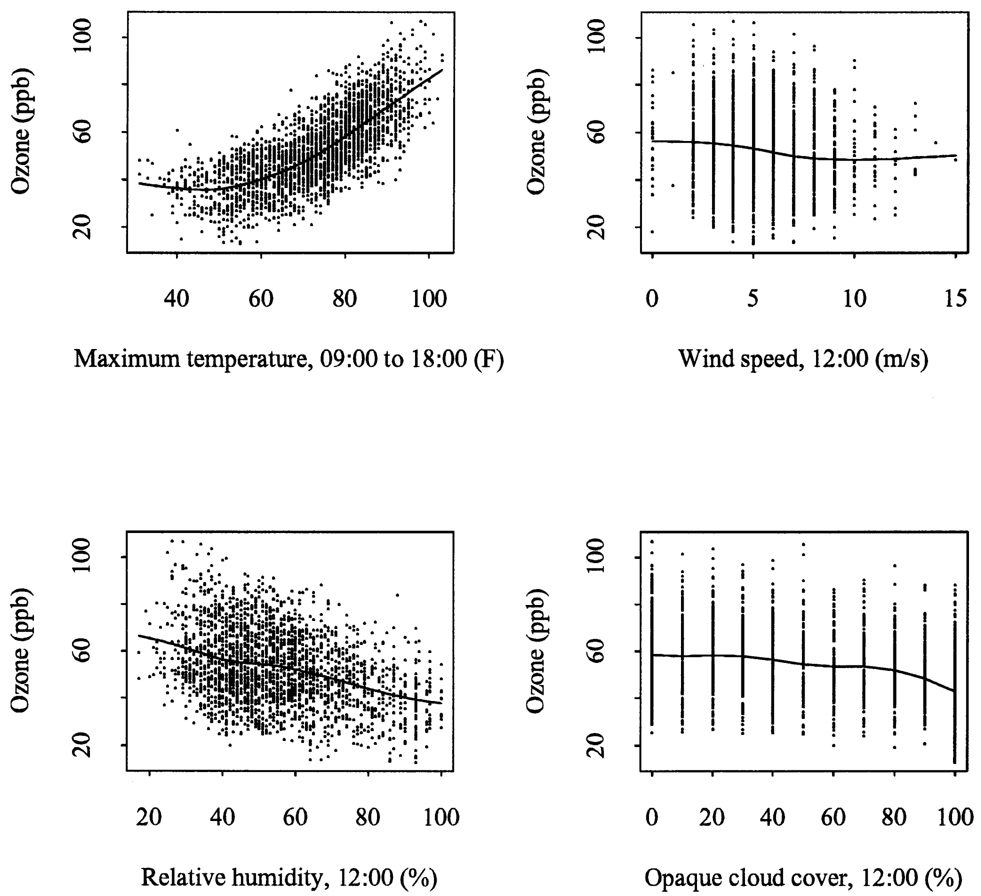

Figure 11: Scatter plots of ozone against temperature, wind speed, relative humidity, and cloud cover.

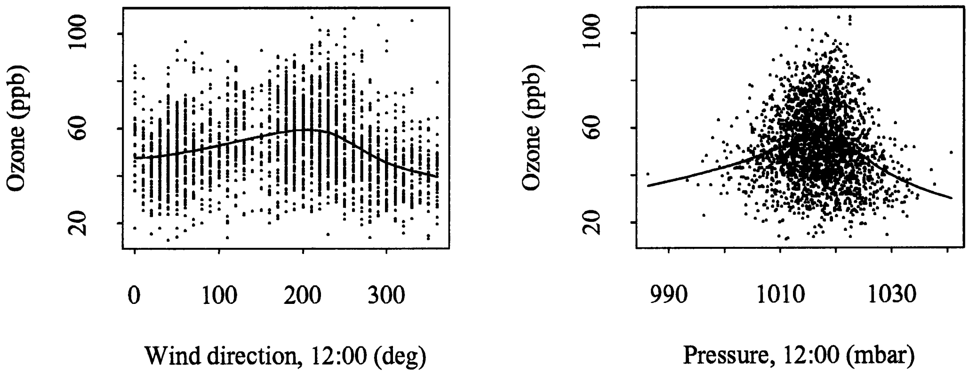

Figure 12: Scatter plots of ozone against wind direction and barometric pressure.

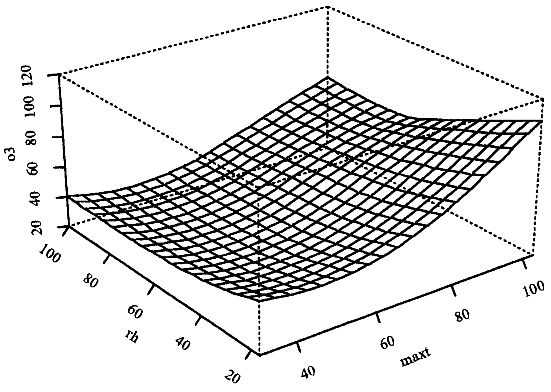

Figure 13: Nonparametric regression surface for ozone against temperature (maximum from 09:00 to 18:00) and noon relative humidity.

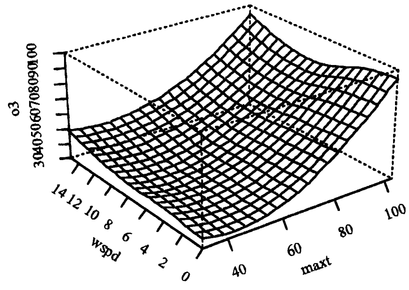

Figure 14: Nonparametric regression surface for ozone against temperature (maximum from 09:00 to 18:00) and noon wind speed.

$R^2 = 0.5583$. 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), 48.2 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 0.5549 (and increases v to 22.85 m/s). By contrast, if such a constant is included in equation (2), it is estimated to be 23.80 ppb, and R^2 increases from 0.6020 to 0.6036.

Again, the R^2 for the model (3), 0.5583, may be compared with that for the lowess fit, which was 0.56 with the equivalent of 10.2 parameters. The parametric model performs well by comparison.

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

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

with wind speed entering in the same form as before. The R^2 for the combined model is 0.6070, and the coefficients change only slightly: the multiplier of (relative humidity $-$ 50%) becomes $-0.00408\%^{-1}$, and v becomes 17.62 m/s. The corresponding lowess fit has $R^2 = 0.62$ for the equivalent of 16.7 parameters, again indicating good performance of the parametric model, which here has 7 parameters.

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 0.6314 and 0.6361, respectively. The dependence of ozone levels on wind direction suggested including a similar term, of the form

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

Including this term raised R^2 to 0.6498. 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 0.6502. Finally, it was found that averaging the wind vector over the hours 06:00 to 18:00 gave a still better fit, with $R^2 = 0.6563$.

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

In the analysis of the urban ozone, it was found that the only substantial association between ozone and upper air meteorological variables was with 700 mbar wind speed.

The effect of including upper air wind speed in the wind speed interaction with temperature was explored by extending the factor

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

by adding the upper air wind speed at 700 mbar, with arbitrary weight, in the denominator. The weight was then estimated by nonlinear least squares. 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 0.6702, an increase in R^2 of 0.0140. This increase was larger than in the urban analysis where the increase in R^2 by adding the 700 mbar wind speed was only 0.0096.

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

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

was extended by adding a corresponding linear combination of the components of the 700 mbar wind vector. However, this increased R^2 by only 0.0001, indicating that there is negligible further information in the upper air wind vector beyond that in the surface wind vector.

5.3 Lagged meteorological variables

The lagged surface meteorological variables that were found to be important in the urban analysis were studied here by using multiple linear regression methods. The lagged variables were all 24 hour averages. The urban analysis indicated 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 0.7084. The coefficient of lag 1 day temperature was positive while the coefficient of lag 2 day temperature was negative and approximately five times the magnitude of the lag 1 day temperature coefficient. Both coefficients were considerably smaller in magnitude than the coefficient of the same day temperature. The coefficient of lagged relative humidity was the same sign and slightly larger in magnitude than 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.

Lagged upper air variables were found to add little to the model in terms of increasing the R^2 and had marginally significant t-statistics. Thus, 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 15. The graph, which is very similar to that for the urban data, 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 semi-annual frequencies represented, and with the mean removed. When this term was included as another multiplicative factor, it raised R^2 considerably, to 0.7867. The alternative of an additive seasonal term gave a somewhat higher R^2 of 0.7964.

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 16 shows the lowess surface, which, as for the urban data 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 0.79, essentially the same as that obtained with both the multiplicative seasonal term and the additive version. The residual root mean square for lowess, 7.330 ppb, falls between those for the additive fit (7.288 ppb) and the multiplicative fit (7.459 ppb). Although the choice is not clear cut, the additive form was chosen for the later analyses to maintain consistency with the urban analysis.

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

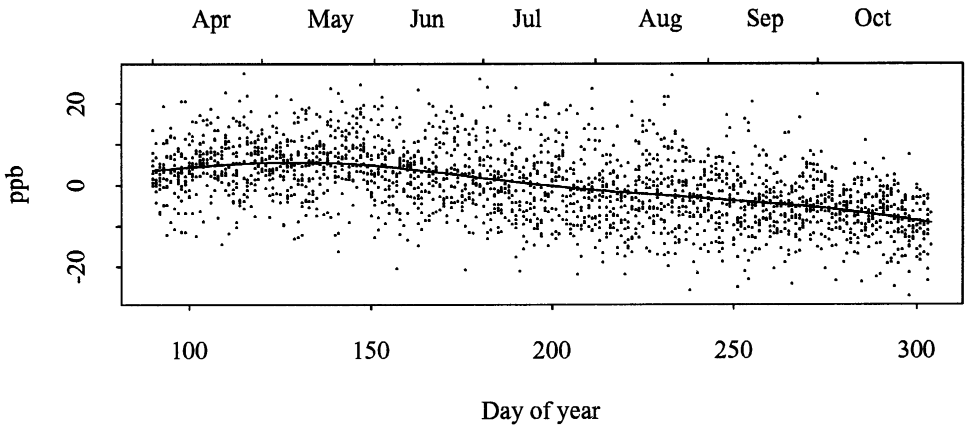

Figure 15: Residuals from meteorological model against day of year.

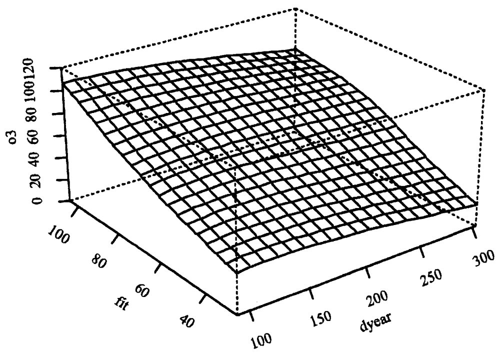

Figure 16: Nonparametric regression surface for ozone against season and the fit from the nonseasonal model.

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 $-1.1\%/\text{decade}$, while the unadjusted trend was $+4.4\%/\text{decade}$. Adding the trend term to the model with meteorological variables resulted in a relatively small rise in R^2 to 0.7965. The R^2 for the model with no meteorology, only trend and seasonality, was 0.3247, showing that meteorological effects as formulated in the model account for around 47% of the variance in ozone concentrations, similar to that found in the urban ozone analysis as might be expected. The *adjusted* and *unadjusted* trends for the urban analysis were $-2.7\%/\text{decade}$, and $+5.3\%/\text{decade}$ respectively.

The $5.4\%/\text{decade}$ 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

$$\begin{aligned}\text{ozone} \sim\ & \{\text{constant} \\ & \qquad + (\text{polynomial in temperature}) \times (\text{function of wind speed})\} \\ & \times (\text{linear function of same day and lag 1 relative humidity})\end{aligned}$$

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	4.165e+01	1.8033263	23.098	1.7141305	24.3000
t0	1.035e+01	4.0753794	2.540	4.0212753	2.5737
t1	1.638e+00	0.2158631	7.587	0.2357416	6.9476
t2	4.230e-02	0.0059948	7.056	0.0065990	6.4103
t3	-8.049e-04	0.0001451	-5.548	0.0001727	-4.6603
tl1	2.931e-01	0.1176979	2.490	0.1178735	2.4867
tl2	-4.543e-01	0.1054086	-4.310	0.1099967	-4.1305
vh	2.472e+01	8.1121112	3.047	9.5918592	2.5767
vh700	1.725e+01	3.3052389	5.218	3.6711659	4.6980
vhl	5.784e+00	1.2494826	4.629	1.4332493	4.0354
r	-1.638e-03	0.0002892	-5.663	0.0003408	-4.8067
rl	-2.898e-03	0.0002767	-10.473	0.0003237	-8.9534
op	-1.062e-03	0.0001080	-9.832	0.0001250	-8.4961
v	-5.035e-03	0.0005513	-9.133	0.0006673	-7.5451
m.u	6.698e-03	0.0012141	5.517	0.0013408	4.9954
m.v	7.355e-03	0.0012694	5.794	0.0013732	5.3560
y	-1.072e-03	0.0009509	-1.127	0.0014068	-0.7619
a1	-8.599e+00	1.2705609	-6.768	1.3582624	-6.3308
b1	4.639e+00	0.4395628	10.553	0.4472495	10.3716
a2	-2.866e+00	0.5764051	-4.972	0.7148538	-4.0089
b2	-1.271e+00	0.4207220	-3.021	0.5180679	-2.4530

$\times$ (linear function of visibility)
$\times$ (linear function of opaque cloud cover)
$\times$ (linear function of mean surface wind vector
$\times$ (linear function of time in years)
$+$ (seasonal model)

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

$$\begin{aligned}
\texttt{o3} \sim\; & (\texttt{mu0} + (\texttt{t0} + \texttt{t1} * (\texttt{maxt} - 60) \\
& \qquad\qquad + \texttt{t2} * (\texttt{maxt} - 60)^2 + \texttt{t3} * (\texttt{maxt} - 60)^3 \\
& \qquad\qquad + \texttt{tl1} * (\texttt{tlag1} - 60) + \texttt{tl2} * (\texttt{tlag2}) - 60) \\
& \qquad * 1/(1 + \texttt{wspd/vh} + \texttt{wspd700/vh700} + \texttt{wlag/vhl})) \\
& * (1 + \texttt{r} * (\texttt{rh} - 50) + \texttt{rl} * (\texttt{rhlag} - 50)) \\
& * (1 + \texttt{op} * (\texttt{opcov} - 50)) \\
& * (1 + \texttt{v} * (\texttt{vis} - 12)) \\
& * (1 + \texttt{m.u} * \texttt{mean.u} + \texttt{m.v} * \texttt{mean.v}) \\
& * (1 + \texttt{y} * (\texttt{year} - 1985)) \\
& + \texttt{a1} * \texttt{cos}(2 * \texttt{pi} * \texttt{year}) + \texttt{b1} * \texttt{sin}(2 * \texttt{pi} * \texttt{year}) \\
& \quad + \texttt{a2} * \texttt{cos}(4 * \texttt{pi} * \texttt{year}) + \texttt{b2} * \texttt{sin}(4 * \texttt{pi} * \texttt{year}). \quad (6)
\end{aligned}$$

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 $\texttt{y}$, 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 7.288 ppb. The lower and upper 2.5% points of the residuals are -13.99 ppb and 15.23 ppb, respectively, while the quartiles are -4.54 ppb and 4.17 ppb. Thus model predictions differ from the actual values by up to ±4.5 ppb about half the time, and by up to ±15 ppb about 95% or the time. Figure 17 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 for all Gaussian quantiles, thus this would seem to indicate that the residuals are approximately Gaussian. This differs from the result for the urban data, where the upper 2.5% of the distribution was markedly stretched out relative to 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 0.5. The seasonal variation in the root mean square is therefore highly significant. The standard error of each correlation coefficient is around 0.06; thus all correlations are significantly differnt from zero, and appear to be roughly equal. This contrasts with the behaviour in the urban data, where the June and July correlations were essentially zero, and were accompanied by some what higher root mean squares.

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 18 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 late in the season, and little relationship early in the season. Due to the fact that there are few high fitted values late in the season, this effect, which is almost exactly opposite to that observed in the urban data, may be spurious.

It should be noted that the adjusted standard errors tabulated in Section 5.5

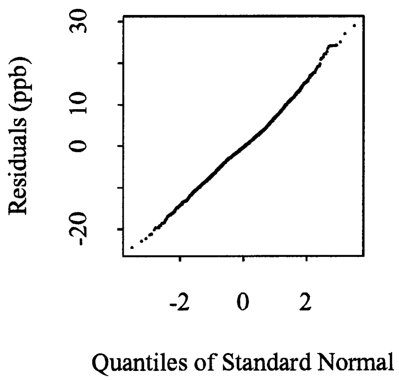

Figure 17: Quantile-quantile plot of the residuals from the fitted model against the Gaussian distribution.

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

Month	Rms	Number in rms	Correlation	Number in correlation
4	6.222	317	0.3057	295
5	6.732	336	0.3555	320
6	7.643	318	0.2550	298
7	8.293	331	0.3694	312
8	8.199	331	0.3488	315
9	7.347	318	0.3205	298
10	6.067	332	0.4566	313

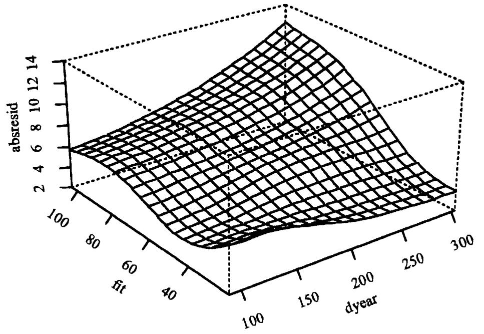

Figure 18: Nonparametric regression surface for magnitude of residuals against season and fitted value.

(page 30) 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 41). 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 18 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, reexpression 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 19 shows the actual 95th percentiles and those of the yearly averaged prediction distributions. The model percentiles track the actual percentiles well, with the largest deviations occurring in 1981, 1987 and 1988. The performance is similar to that in the urban study. 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 7.288 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 20 shows the 95th percentiles by year of such adjusted ozone values, as well as the

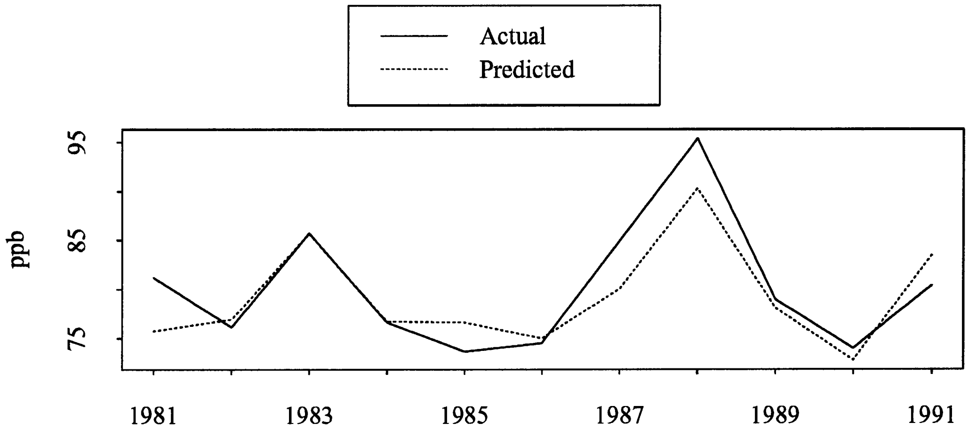

Figure 19: Actual and predicted 95th percentiles of ozone by year.

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. The highest values, in 1988, are only partially adjusted, while the lower values in 1987, still high by comparison with other years are essentially unadjusted.

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 20, 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.

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

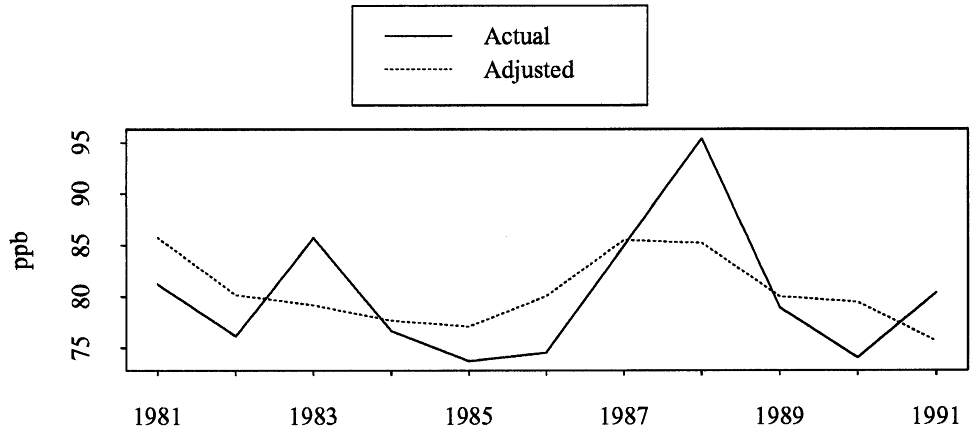

Figure 20: Actual and adjusted 95th percentiles of ozone by year.

variations. Except where noted, the values are consistent with those for the urban data.

`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. 42 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 21 (horizontal line indicates `mu0`). In constructing the figure, lagged temperatures were taken as equal to same day temperature. The curve is similar to that for the urban data, but with a somewhat lower maximum at high temperatures.

`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. 24.7 m/s, 17.2 m/s, 5.8 m/s. All are higher (indicating *weaker* dependence) than for the urban data, especially for unlagged surface wind speed.

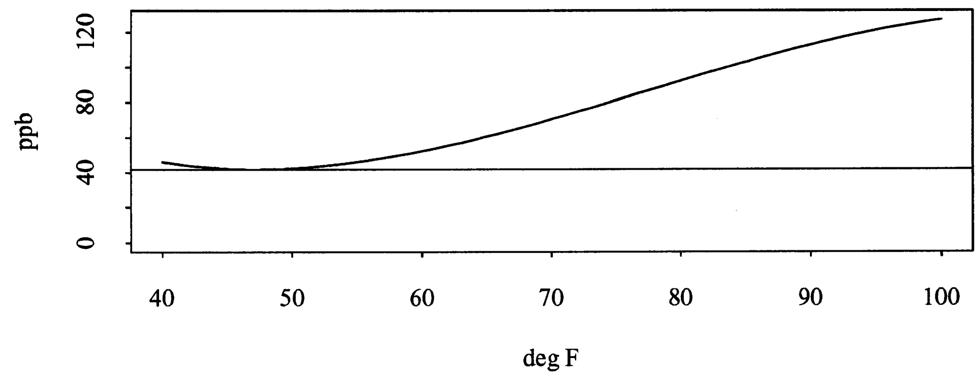

Figure 21: Fitted polynomial effect of temperature at zero wind speed.

`r, rl` : Effects of relative humidity and lagged relative humidity. $-0.0016\%^{-1}$, $-0.0029\%^{-1}$. The factor drops from 1.23 to 0.78 as both humidities rise from 0% to 100%.

`op` : Effect of opaque cloud cover. $-.00106\%^{-1}$. The factor drops from 1.05 to 0.94 as opaque cloud cover rises from 0% to 100%.

`v` : Effect of visibility. $-.005\text{km}^{-1}$. The factor drops from 1.06 to 0.94 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 $(\texttt{m.u}, \texttt{m.v}) = (0.0067, 0.0074)(\text{m/s})^{-1}$. Maximum when wind is from the southwest. The effect of a 5 m/s wind varies from 1.05 to 0.95 as the wind direction changes from southwest to northeast.

`y` : Trend parameter, $-0.00107\text{yr}^{-1} = -1.07\%/\text{decade}$.

`a1, b1, a2, b2` : Coefficients of the annual and semiannual cosines and sines. The seasonal term is shown in Figure 22. The location of the maximum in early May 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

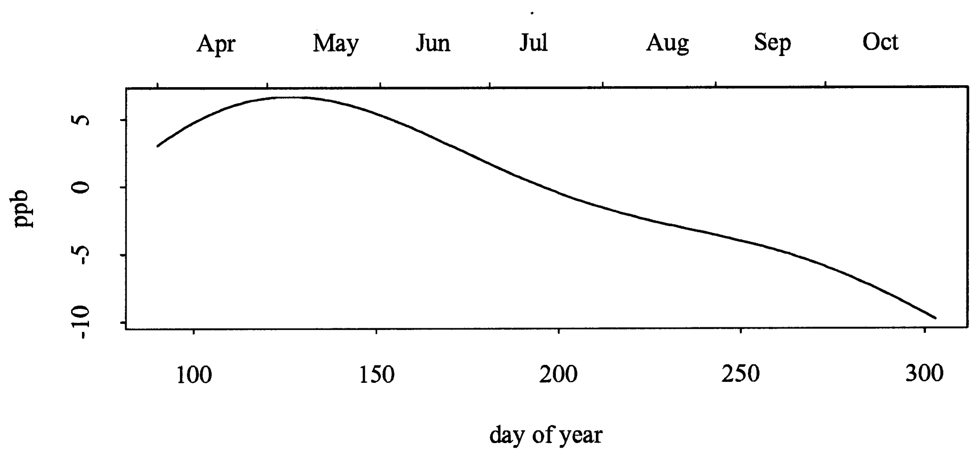

Figure 22: Fitted seasonal effect.

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 15 (page 28).

6.4 Cross validation

As for the urban data, the model was studied by cross validation and jackknifing by whole seasons. 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 and Weisberg (1982, Section 2.2.3). The overall root mean square of the cross-validated prediction errors is 7.517 ppb, which compares very favorably with the root mean square residual of the fit to all the data, 7.288 ppb.

There is some variability from year to year in Table 7, which was evaluated by combarison with a "bootstrap" population. Figure 23 shows the ordered values from the third column of Table 7, graphed against the bootstrap quantiles. The graph suggests that at most the two highest observed values, those for 1981 and 1991, are more extreme than could be explained by sampling variability, and by

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

Year	Number of days	Mean	Root mean square
1981	204	0.5408	8.712
1982	208	-1.4923	7.626
1983	210	0.1565	7.294
1984	208	-0.9784	6.732
1985	209	-1.3757	6.734
1986	206	-0.3026	7.077
1987	209	2.1880	7.116
1988	212	2.8545	7.739
1989	208	2.6512	7.132
1990	204	-0.1076	7.243
1991	205	-4.8142	8.963

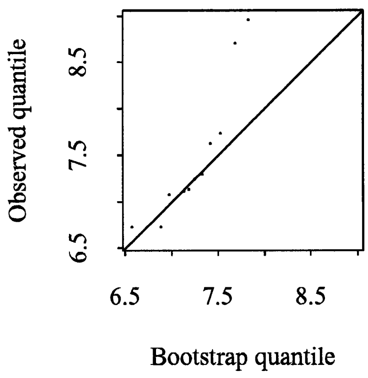

Figure 23: Quantile-quantile plot of predicted root mean square residual by year, against a bootstrap reference distribution.

only around 0.5 ppb.

6.5 Jackknifing

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 30). 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 4.8 %/decade. This had no effect on the significance of the associated $|t|$-statistic which declined from 0.73 to 0.20 (allowing for heteroscedasticity and short-term correlation). 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 24 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 24 has a slope of 1.24, and provides a good compromise match to the remaining points. This suggests that the jackknife standard errors are estimating values up to 24% larger than the adjusted standard errors of Table 5. The adjusted standard errors multiplied by 1.24 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.

7 A Revised Model for Rural Ozone

Although fitting the same model to the rural ozone as was fit to the urban ozone is desireable for comparative purposes, and to a certain extent makes sense due to the high correlation between the typical values of the two networks, it is likely

Table 8: Jackknifed estimates and standard errors.

Coefft.	Fitted Value	Standard error	t Value
mu0	4.166e+01	1.370e+00	30.414
t0	1.025e+01	4.646e+00	2.207
t1	1.648e+00	3.126e-01	5.273
t2	4.263e-02	9.475e-03	4.500
t3	-8.170e-04	3.403e-04	-2.401
tl1	2.961e-01	1.247e-01	2.375
tl2	-4.576e-01	1.588e-01	-2.881
vh	2.503e+01	1.282e+01	1.953
vh700	1.720e+01	4.044e+00	4.254
vhl	5.774e+00	1.281e+00	4.506
r	-1.636e-03	3.091e-04	-5.291
rl	-2.887e-03	3.526e-04	-8.190
op	-1.059e-03	1.265e-04	-8.376
v	-4.999e-03	8.723e-04	-5.730
m.u	6.755e-03	1.857e-03	3.638
m.v	7.404e-03	1.734e-03	4.270
y	-9.793e-04	4.776e-03	-0.205
a1	-8.588e+00	1.482e+00	-5.797
b1	4.643e+00	6.800e-01	6.828
a2	-2.868e+00	6.652e-01	-4.312
b2	-1.275e+00	3.097e-01	-4.116

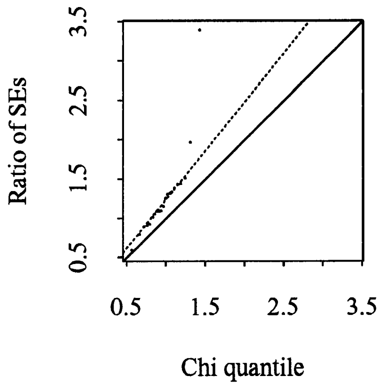

Figure 24: Quantile-quantile plot of ratio of jackknifed standard errors to adjusted standard errors, against the χ distribution with 10 degrees of freedom, with lines of slope 1 and 1.24.

that differences do exist in the dependence of ozone on meteorological variables. These differences are explored here.

The surface meteorological variables ceiling height and barometric pressure were added to the model of Section 5.1 as $(1 + \texttt{c} * \texttt{cht})$ and $(1 + \texttt{p} * \texttt{pr})$ respectively and were found to increase the R^2 by less than 0.001. Total cover was added to the model as $(1 + \texttt{op} * \texttt{opcov} + \texttt{to} * \texttt{totcov})$, and increased the R^2 by only 0.002. All three of these model fits were done without the trend, seasonal, upper air, and lagged surface components of the model. Due to the small increases in R^2 as a result of adding these variables, they were not considered further, and left out of the final model.

The only significant upper air meteorological variable found to be important in the urban analysis was 700 mbar wind speed. The residuals of the surface met model from Section 5.1 were explored for dependence on upper air met variables using linear regression. It was found that wind speed, relative humidity, and height of pressure layer were all important in explaining significant amounts of the residual variation.

Overall, the suite of upper air met variables at the 950 mbar height explained more of the residual variation than the upper air met variables at other heights.

However, the variables at the 500 mbar height explained nearly as much residual variation, and the variables at the 850 mbar height explained the least residual variation. The single most important upper air met variable was relative humidity at 950 mbar. Adding this variable to the non-linear model in the form $(1 + \texttt{r} * \texttt{rh} + \texttt{ru} * \texttt{rh950})$ increased the R^2 by more than 0.02. As was the case in the urban model, 700 mbar wind speed was quite important. The marginal increase in R^2 by adding this term to the non-linear model in the same form as in Section 5.2 was approximately 0.014. Aside from these two variables, little else added significantly to the R^2 of the rural model. Adding these two upper air variables to the model with trend and seasonal components increased the R^2 from 0.7537 to 0.7863.

The residuals from Section 5.1 were also used to search for lagged surface meteorological variables that have an important effect on ozone. Using linear regression, it was found that relative humidity, wind speed, temperature and pressure at lags 1 and 2 days, were important in explaining some of the variation of these residuals.

Lagged 1 day relative humidity was the most important of these variables when it was added to the non-linear model. By adding this variable to the model in the form $(1+\texttt{r}*\texttt{rh}+\texttt{ru}*\texttt{rh950}+\texttt{lrh}*\texttt{rhlag1})$, the R^2 was increased by nearly 0.024. Adding lagged 1 day wind speed as in Section 5.3 added nearly 0.006 to the R^2 of the model. Both relative humidity and temperature lagged 2 days, added approximately 0.014 to the R^2 of the fit, however relative humidity lagged 2 days in the presence of relative humidity lagged 1 day added only marginally to the R^2. Adding the lagged temperature variables to the model in the presence of the other lagged variables, upper air variables and seasonal component, increased the R^2 by only 0.0015, thus they were not used in the final model. Pressure added very little in terms of R^2, and was not used further. Adding relative humidity and wind speed lagged 1 day to the model with the 2 upper air variables of the preceeding section and the trend and seasonal components, increased the R^2 from 0.7863 to 0.7996.

The addition of 950 mbar relative humidity, and the removal of the two lagged temperature terms from the urban model increased the R^2 of the final model by approximately 0.003, from 0.7967 to 0.7996. This required 1 less parameter. The root mean square residual from this revised fitted model was 7.229 ppb, 0.05 less than that of the urban model fit to the rural data. The parameter estimates from this revised fit, and the conventional and adjusted (for serial correlation) standard errors are given in Table 9.

Table 9: Coefficients in the revised 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	3.988e+01	1.874e+00	21.279	1.756e+00	22.705
t0	1.133e+01	3.591e+00	3.157	3.390e+00	3.344
t1	1.466e+00	1.614e-01	9.085	1.769e-01	8.287
t2	3.983e-02	4.967e-03	8.020	5.439e-03	7.324
t3	-7.580e-04	1.235e-04	-6.138	1.442e-04	-5.257
vh	3.663e+01	1.364e+01	2.686	1.634e+01	2.242
vh700	2.080e+01	3.709e+00	5.610	4.102e+00	5.072
vhl	6.952e+00	1.408e+00	4.936	1.606e+00	4.329
r	-7.701e-04	2.909e-04	-2.648	3.316e-04	-2.322
rh95	-1.721e-03	2.040e-04	-8.437	2.371e-04	-7.259
rhl	-2.540e-03	2.778e-04	-9.141	3.241e-04	-7.836
op	-9.692e-04	1.074e-04	-9.027	1.223e-04	-7.923
v	-5.509e-03	5.444e-04	-10.120	6.479e-04	-8.502
m.u	6.977e-03	1.208e-03	5.775	1.290e-03	5.409
m.v	7.524e-03	1.175e-03	6.406	1.281e-03	5.874
y	-1.448e-03	9.429e-04	-1.535	1.376e-03	-1.052
a1	-7.729e+00	1.237e+00	-6.247	1.326e+00	-5.831
a2	-3.046e+00	5.729e-01	-5.317	7.071e-01	-4.307
b1	4.931e+00	4.224e-01	11.674	4.350e-01	11.336
b2	-1.454e+00	4.183e-01	-3.475	5.077e-01	-2.863

There were few noticeable changes in the parameter estimates of this revised model. In particular, the estimate of trend changed from $-1.07\%/\text{decade}$ to about $-1.45\%/\text{decade}$, although it was still statistically insignificant. The estimated coefficient for 950 mbar relative humidity was negative and approximately twice the magnitude as surface relative humidity.

8 Conclusions

The daily maximum one-hour average surface ozone concentrations from the 12 stations were highly correlated. A principal component analysis of these 12 stations showed a dominant principal component that accounted for 61% of the variance, and the corresponding time series was nearly perfectly correlated ($\hat{\rho} = 0.98$) 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 about 80% of the variance of the ozone concentration data. The root mean square residual is 7.279 ppb. The model may be extended to include a trend parameter, which is estimated to be $-1.0\%/\text{decade}$, with a (jackknife) standard error of $4.8\%/\text{decade}$. 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 $+4.4\%/\text{decade}$, representing the *unadjusted* trend in the surface ozone concentrations.

A slightly revised model that did not include lagged temperature and included relative humidity at the 950 mbar height was fit to the rural data and was found to increase the R^2 and change the parameter estimates slightly.

The fact that the fits of this model to both the urban and rural data are very similar in terms of R^2, and in both cases quite good, suggests that the physical and chemical processes underlying ozone formation in both urban and rural settings are approximated nicely by this statistical model. Furthermore, the fact that the fits are similar is encouraging since the major processes of ozone formation are indeed similar in both environments.

References

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., Royle, J. A. and Yang, Q. (1993), Accounting for meteorological effects in measuring urban ozone levels and trends, Technical Report 1, National Institute of Statistical Sciences.

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.

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. 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.

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

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

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