Ozone Exposure and Population Density in Harris County, Texas

R. J. CARROLL, R. CHEN, E. I. GEORGE, T. H. LI, H. J. NEWTON, H. SCHMIEDICHE, and N. WANG

We address the following question: What is the pattern of human exposure to ozone in Harris County (Houston) since 1980? While there has been considerable research on characterizing ozone measured at fixed monitoring stations, little is known about ozone away from the monitoring stations, and whether areas of higher ozone correspond to areas of high population density. To address this question, we build a spatial–temporal model for hourly ozone levels that predicts ozone at any location in Harris County at any time between 1980 and 1993. Along with building the model, we develop a fast model-fitting method that can cope with the massive amounts of available data and takes into account the substantial number of missing observations. Having built the model, we combine it with census tract information, focusing on young children. We conclude that the highest ozone levels occur at locations with relatively small populations of young children. Using various measures of exposure, we estimate that exposure of young children to ozone decreased by approximately 20% from 1980 to 1993. An examination of the distribution of population exposure has several policy implications. In particular, we conclude that the current siting of monitors is not ideal if one is concerned with population exposure assessment. Monitors appear to be well sited in the downtown Houston and close-in southeast portions of the county. However, the area of peak population is southwest of the urban center, coincident with a rapidly growing residential area. Currently, only one monitor measures air quality in this area. The far north-central and northwest parts of the county are also experiencing rapid population growth, and our model predicts relatively high levels of population exposure in these areas. Again, only one monitor is sited to assess exposure over this large area. The model we developed for the ozone prediction consists of first using a square root transformation and then decomposing the transformed data into a trend part and an irregular part, the latter modeled as a Gaussian random field with both time and space correlations. Due to the large number of observations and high-dimensional optimization problem, we developed a fast method to estimate the parameters of the model. The model and estimation method are general and can be used in many problems with space–time observations.

KEY WORDS: Data transformation; Gaussian random field; Missing data; Spatial statistics; Spatial–temporal modeling; Time series analysis.

1. INTRODUCTION

Ambient ozone pollution in urban areas represents one of the nation's most pervasive environmental problems. While the decreasing stratospheric ozone layer may lead to increased instances of skin cancer, high ambient ozone intensity has been shown to cause damage to the human respiratory system as well as to agricultural crops and trees (Lefohn and Runeckles 1987; Lippmann 1989). Four metropolitan areas in Texas currently are not in compliance with the National Ambient Air Quality Standard (NAAQS): Houston, Beaumont/Port Arthur, El Paso, and Dallas/Fort Worth. Among these, the Houston area was rated **severe,** second only to the Los Angeles area in the entire nation. The current project concentrates on the ozone pollution problems in the Houston area. The data analyzed are hourly ozone measurements between 1980 and 1993, along with concurrent meteorological variables and demographic variables. The project has three major goals:

- Provide information (and/or tools to obtain such information) about the amount and pattern of missing data, as well as the quality of the ozone and the meteorological measurements.
- Build a model of ozone intensity to predict the ozone concentration at any given location within Harris County at any given time between 1980 and 1993.
- Apply this model to estimate exposure indices that account for either a long-term exposure or a short-term high-concentration exposure, and also relate census information to different exposure indices to achieve population exposure indices.

Ground-level ozone has been studied extensively in the literature. For example, Cox and Chu (1992), Horowitz (1980), and Smith and Huang (1993) studied daily maximum ozone concentration using extreme value theory. Bloomfield, Royle, and Yang (1993a,b) constructed a nonlinear regression model for hourly average ozone data in the Chicago area. Guttorp, Meiring, and Sampson (1994) used a space–time model to analyze ground-level ozone data. Niu (1996) used a nonlinear additive model for ozone series. Cressie (1993, p. 274) listed a number of other references to pollution data analyses and general space-time modeling methods.

Most analyses use aggregation to cope with large datasets or else incorporate only space or time correlation into models or else incorporate them separately. In this research we attempt to reconstruct the ozone surface, particularly at locations other than monitoring stations. We note that we have not used meteorological adjustment, which tends to

R. J. Carroll is Professor of Statistics, Nutrition and Toxicology, H. J. Newton is Professor and Head, Rong Chen is Associate Professor, Ta-Hsin Li and Naisyin Wang are Assistant Professors, and Henrik Schmiediche is Lecturer, Department of Statistics, Texas A&M University, College Station, TX 77843. Edward George is Professor, Department of Management Science and Information Systems, University of Texas, Austin, TX 78712. The list of authors is in alphabetical order. This research was supported in part by a contract from the Texas Natural Resource and Conservation Commission (TNRCC). Carroll and Wang were supported by National Cancer Institute grant CA-57030. Chen and George were supported by National Science Foundation grants DMS-9301193 and DMS-9404408. The authors thank David Sullivan of the TNRCC for many helpful discussions, and the editor and three referees for detailed comments.

Journal of the American Statistical Association
June 1997, Vol. 92, No. 438, Applications and Case Studies

estimate the ozone surface under idealized meteorology. (References on meteorological adjustment can be found in Bruntz, Cleveland, Graedel, Kleiner, and Warner 1974; Lamb, Guenther, Gay, and Westberg 1987; National Research Council 1991; and Pagnotti 1990). Our main concern is the risk assessment of health to overall ground-level ozone. Further research may include health risk assessment of ozone due to emission, where meteorological adjustment will be the main focus.

In Section 2 we describe the ozone and meteorological data and discuss the quality and missingness of the data. Despite a missingness rate of approximately 20%, we conclude that the quality and quantity of observed data warrant building a model. This model, discussed in Section 3, consists of first using a square root transformation and then decomposing the transformed data into a trend part and an irregular part, the latter modeled as a Gaussian random field. Inspection of sample correlations of the detrended transformed data strongly suggests an exponential form for the space–time correlation function of the random field, which depends on seven parameters. In Section 4 we discuss various methods for estimating these parameters, including a new fast cross-validation–type method that handles both the massive amounts of data involved and the considerable amount of missing data. We give the results of our methods for modeling and predicting ozone in Section 5.

In Section 6 we discuss combining the ozone predictions obtained from our model with measures of population density obtained from census data to obtain population ozone exposure indices. Such indices and their standard errors can be calculated for any time and any location and can be summarized if desired by averaging individual time–location values over time and/or locations. An interesting implication of our results for Harris County is that there has been a large decrease in the ozone exposure of children from 1980–1993. Finally, we present some conclusions in Section 7.

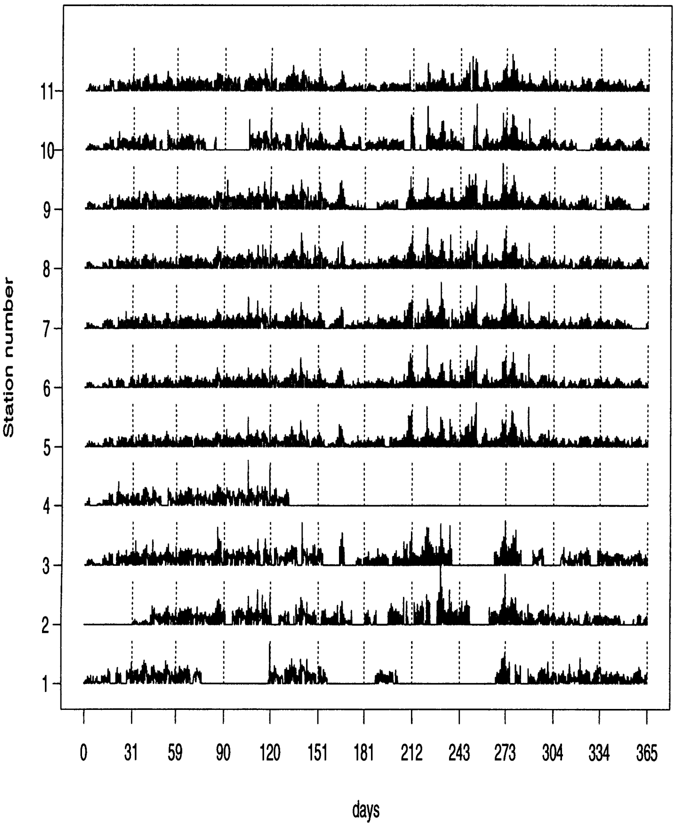

Figure 1. Ozone Measurements for 1993. Observations at zero are missing. The dotted lines indicate the end of each month.

2. THE OZONE AND WEATHER DATA

In this section we describe the data provided to us by the Texas Natural Resources Conservation Commission (TNRCC). Hourly measurements of the level of ambient ozone (in parts per billion [ppb]) in Harris County from 1980–1993 were recorded by 9 to 12 monitoring stations, the number varying by year. Figure 6 shows the locations of the 11 monitoring stations operating in Harris County in 1993. Each station is marked with a station number. The dotted lines are the major highways within the county. Besides the ozone level, each station also recorded three meteorological variables: temperature, wind speed, and wind direction.

2.1 Data Quality

With such a large dataset, we began our analysis by evaluating the quality of the data. This evaluation entailed two aspects: (a) identifying the proportion and location of the missing data, and (b) validating the accuracy of observed data.

We first focus on the ozone measurements. These are relatively high-quality data, because the monitors were calibrated weekly and review procedures were used to eliminate measurements suffering from temporal inconsistency. There is no known instrumental change for the period in our study. A major issue faced in this (and indeed, any) analysis of ozone is the pattern and extent of missing data. Figure 1 displays the hourly ozone measurements for 1993 at each of the 11 stations; observations at zero are missing. We see immediately that station 4 was shut down before mid-year, whereas station 1 had significant periods of shutdown. Overall, excluding station 4, about 20% of the ozone measurements are missing.

Examination of Figure 1 clearly shows that the data are not missing completely at random, but instead systematic patterns of missingness occur. Stations go off-line for fixed (sometimes long) periods, and the measurements are calibrated according to a fixed schedule. However, the figure indicates that the missing data are missing at random in the sense of Little and Rubin (1987); that is, observations are missing because of factors other than the ozone measurements themselves, such as scheduled calibrations, device failures, and so on. There is no evidence from the plots (or discussions with the TNRCC) that missing data tend to be of a particular size (e.g., high or low). The assumption that ozone measurements are missing at random is a crucial part of our analysis, because, as we discuss in Section 4, the missing-at-random assumption allows us to do a form of imputation.

Temperature data also suffer from problems of missing values. These data appear to be of somewhat lower quality than the ozone data, principally because temporal inconsistencies were not addressed prior to construction of the data file. We calculated the differences in temperature between consecutive times at a single station and the differences at the same time among neighboring stations. When we found gross discrepancies, we modified the suspicious observations to be missing.

Fortunately, both the ozone data and our modified temperature data appear to be fairly high-quality data, with reasonably low spatial and short-term temporal variability. This enabled us to use missing-value imputation techniques with these variables for our spatial and temporal modeling of ozone concentration throughout Harris County. However, other meteorological variables, such as wind speed, appear to have substantial variability as well as significant numbers of missing values. These drawbacks complicate the use of these variables in modeling ozone throughout Harris County. We discuss wind speed and direction further in the next section.

In what follows, it is important to keep in mind the vast extent of the data base with which we are working. Even ignoring missing data, 11 stations providing hourly measurements for 14 years provided more than 1,300,000 ozone measurements. Fitting a spatial–temporal model with so many observations while at the same time taking missingness into account is not feasible unless one develops quick computational techniques.

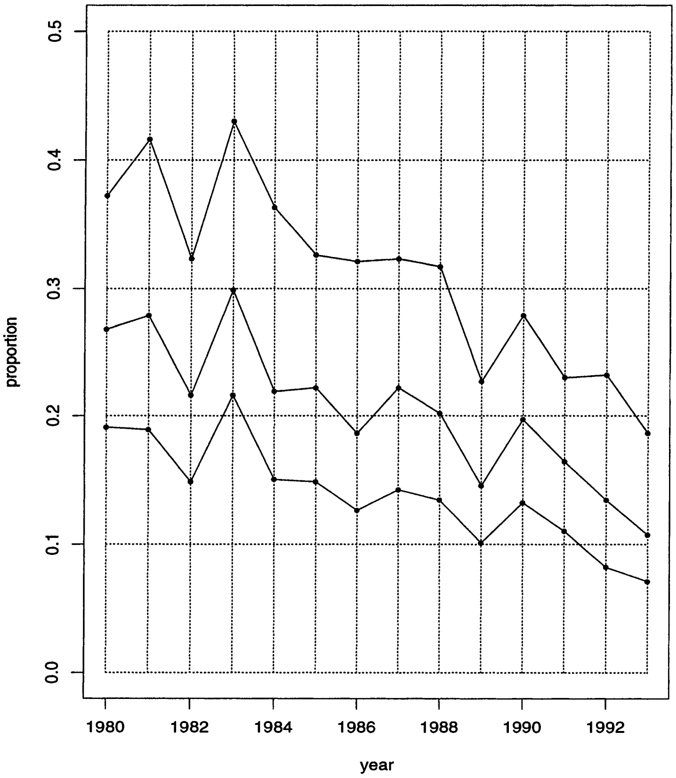

Figure 2. Annual Proportion of Days Above 80 ppb, 100 ppb, and 120 ppb in Harris County, 1980–1993.

2.2 Proportional of Days Above Thresholds

To get a preliminary sense of ozone trends throughout Harris County from 1980–1993, we tabulated the proportion of days that any station reports any hour for which ozone concentration is above a particular threshold. Missing values were simply ignored. Figure 2 displays for each year the proportion of days above each of the three thresholds 80, 100, and 120 ppb. The steady decline of these proportions is clear from the plot. For example, the proportion of days above 80 ppb declined from around .4 to below .2 during the time period. Although it is not shown here, there is also a strong monthly effect in the number of days above thresholds. The declines in high ozone days are apparent in all summer months, and are particularly pronounced in July.

3. BUILDING A SPACE–TIME MODEL FOR OZONE

Although the raw trends portrayed in Figure 2 provide evidence of reduced ozone levels in Harris County, both the missing-value structure and the spatial dimension have been ignored. In this section we describe a comprehensive model for the ozone level throughout Harris County. Using the available data, this spatial–temporal model can be used to obtain a map of the ozone intensity throughout Harris County at any time from 1980–1993.

3.1 A Square Root Transformation

Let $\mathrm{oz}(x,t)$ denote the ozone level at spatial location x and time t, where x is the vector of longitude and latitude. Extensive exploratory analyses (not shown here) of the ozone data showed phenomena very similar to that found by Haslett and Raftery (1989) in their analysis of long-term records of wind speeds at 12 recording stations in Ireland—namely, skewed distributions and standard deviations correlated with means. We thus decided to follow their lead and use a square root transformation of the ozone data. We thus define

$$Y(x,t) = \sqrt{\mathrm{oz}(x,t)}.$$

Because ozone is highly related to sunlight and temperature, and these variables are quite variable during a 24-hour period, we decided not to do any time aggregation of the data.

The model we use is of the form

$$Y(x,t) = g(t) + \varepsilon(x,t), \qquad (1)$$

consisting of two components: $g(t)$, a deterministic function of month, hour, temperature, and possibly other meteorological data; and $\varepsilon(x,t)$, a random process which captures spatial and temporal variation due to other factors. The rest of this section deals with the details of these components.

3.2 The Deterministic Trend

Because we need a model that is predictive not just at the monitoring stations, but throughout Harris County, any variables placed in the deterministic part of model (1) must be observable or at least predictable throughout Harris County. The fact that temperature is relatively constant throughout the county at any given time makes it easy to include it in the deterministic part of the model. In fact, we include linear and quadratic terms in the median temperature of the monitoring stations as predictor variables in the model.

On the other hand, including wind speed and/or wind direction in the deterministic trend is much more problematic. As mentioned previously, wind speed appears to be spatially very variable. For example, the average coefficient of variation (the ratio of standard deviation to the mean) of the 11 monitoring sites over the 8,760 hours in 1993 is .8914. It is difficult to construct any function of wind speed that would allow for reasonable extrapolation away from the monitoring sites.

Another difficulty with incorporating wind speed has to do with goodness of fit. It is well known in time series analysis that noisy exogenous variables may help explain a response variable, but typically are not very helpful in prediction. Wind speed appears to be just such a variable. We added wind speed and the interaction of wind speed and temperature to our model and used ordinary regression techniques to predict the ozone concentration at monitoring stations where wind speed was observed. We found that including wind speed in the model actually increased the prediction error for the 1993 data by 5%, compared to the model without wind speed. So wind speed actually decreased the quality of the fit using this technique. Keeping in mind that the main goal is to estimate population exposure in Harris County, it seems reasonable to treat wind speed and direction as part of the random process component in (1).

Our model for $g(t)$ is thus

$$g(t) = \alpha_{\text{hour}} + \beta_{\text{month}} + \gamma_1 \text{temp}(t) + \gamma_2 \text{temp}^2(t), \quad (2)$$

where temp(t) is the median temperature over Harris County at time t. In this model α_{hour} accounts for the overall hourly level of ozone, and β_{month} accounts for the overall monthly level.

Note that the trend is constant over space, which we feel is justified both by the data and because Harris County is a relatively small and flat area, without strong geographical variations. Including linear and quadratic terms of longitude and latitude in the deterministic trend does not reduce the variance of the random field significantly.

3.3 The Random Process

Spatial and temporal variation is dealt with by letting $\varepsilon(x,t)$ be a real-valued stationary *Gaussian random field* (GRF) with mean 0. GRF's are standard models for nondeterministic spatial–temporal variation (see, e.g., Handcock and Wallis 1994).

The commonly used kriging method (Cressie 1989; Isaaks and Srivastava 1990; Journel and Huijbregets 1979) and smoothing spline method (Laslett 1994; Wahba 1983) are designed for spatial statistics without time direction correlation. The work of Haslett and Raftery (1989) and Handcock and Wallis (1994) tries to eliminate the time–direction correlation before modeling the spatial structure.

The key feature of our GRF is that it deals with both space and time correlation. To get a feel for the nature of the correlation of the GRF in both the time and space domains, we began by fitting the trend part of the model by ordinary least squares and examining the resulting residual process. We calculated the autocorrelations and cross-correlations for all pairs of monitoring stations for time lags 0, 1, 2, and so on, and for each time lag we plotted the correlations as a function of the Euclidean distance in degrees between stations. We noticed that each plot was exponential in nature, and so we plotted the logarithm of the correlations versus distance. In calculating these correlations, we used only pairs of data at a fixed time lag when both values were observed. The resulting plots for 1993 for time lags 0, 2, 4, and 6 are displayed in Figure 3, along with the least squares line fit to the points on each plot. The plots for all time lags are on the same scale so as to show a variety of features. First, the scatter is strongly linear for each time lag, and the lines steadily move down for increasing lags. Second, the degree of scatter increases only slightly with lag. For lag 0, the line appears to have an intercept close to 0 but probably below 0. This arises later in terms of a "nugget effect" in our final correlation model.

Thus we concluded that the sample log correlations appeared to be well modeled by linear functions. We then wondered whether the slopes and intercepts of these linear functions might follow some simple function of time lag, so we plotted them for each year from 1980–1993 and found that the plots followed closely quadratic functions of time lag. To illustrate this, Figure 4 shows a plot of the slopes and intercepts versus time lags 0–8 for 1991, 1992,

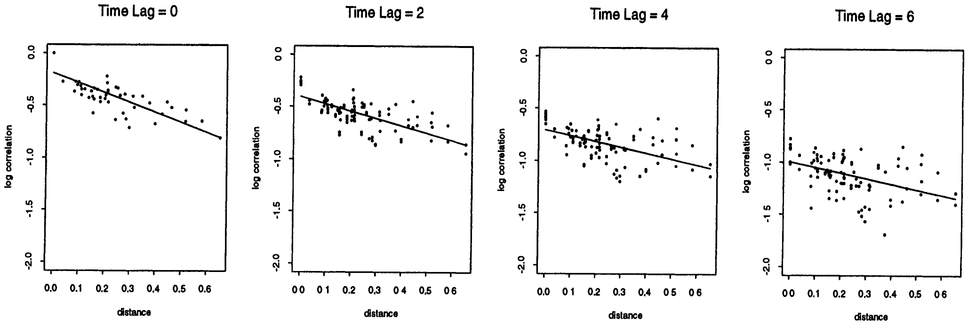

Figure 3. Logarithm of Sample Cross-Correlations Versus Distance Between all Pairs of Stations for Detrended Transformed Ozone Data for Time Lags 0, 2, 4, and 6 for 1993.

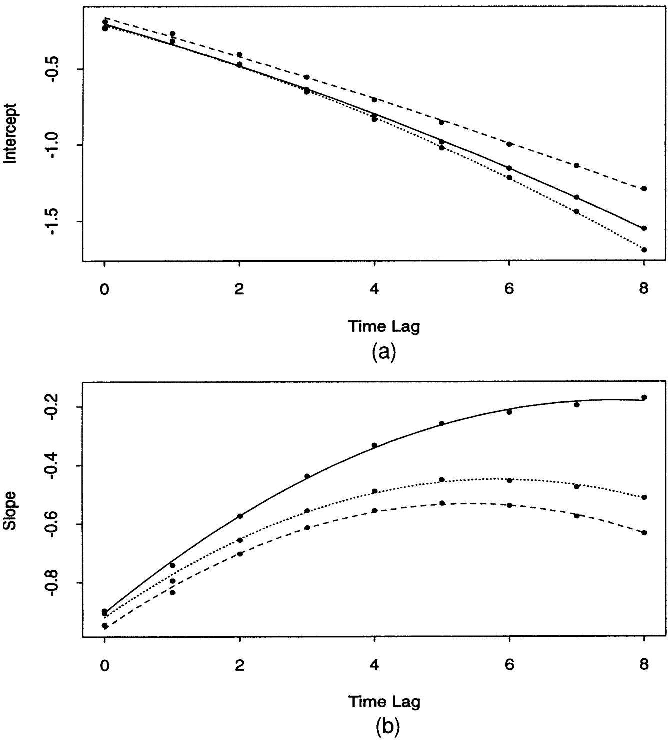

Figure 4. Intercepts (a) and Slopes (b) Versus Time Lag of the Least Squares Lines for Fitting Log Sample Correlation Versus Distances: ———, 1991; · · · ·, 1992; - - - -, 1993.

and 1993. We have superimposed on each plot a quadratic function fit to the points. The fit is remarkably good.

One final question that we had was whether the correlation function of the GRF is isotropic; that is, whether the spatial correlation for two stations is invariant to the relative position of the two stations. In particular, we considered the fact that the Houston area has a southeast–northwest prevailing wind pattern.

Figure 5 is a plot of log correlation versus distance for various time lags using two symbols, an "x" for two stations that lie along the prevailing wind region and an "o" for stations in a region perpendicular to the prevailing wind. Least squares lines for each of the two types are also plotted. Because there do not appear to be large differences in the lines, we concluded that isotropy is a reasonable assumption.

Our space–time covariance function is thus of the form

$$\operatorname{cov}(\varepsilon(x_1,t_1),\varepsilon(x_2,t_2)) = \sigma^2\rho(d,v),$$

where d is the Euclidean distance between locations x_1 and x_2, $v = |t_2 - t_1|$ is the time lag between two times, and the correlation function ρ is given by

$$\rho(d,v) = \begin{cases} 1 & \text{if } d = v = 0 \\ \phi_v^d \psi_v & \text{otherwise,} \end{cases}$$

where

$$\log(\psi_v) = a_0 + a_1 v + a_2 v^2$$

and

$$\log(\phi_v) = b_0 + b_1 v + b_2 v^2.$$

If a_0 is not 0, then we have the so-called nugget effect; that is, spatial correlations less than 1 at very small distances. This can be due to measurement error and other causes (Cressie 1993; Journel and Huijbregts 1979). The nugget effect cannot happen with continuous process. However, if the observed process is the sum of a continuous process and a *white* process (measurement error, say), then the nugget effect is inevitable.

The correlation of the random field is the product of two factors. The first factor, ϕ_v^d, depends on both the time difference and the distance between the two locations. The second factor, ψ_v, depends only on the time difference. Finally, note that counting σ^2, the model has seven parameters.

Although we have not been able to show analytically that ρ is a positive definite function, we have inverted correlation matrices based on it for a wide variety of time lags, distances, and parameter values.

4. ESTIMATION AND PREDICTION

In this section we discuss possible estimation and prediction methods for models such as ours and describe in detail the methods that we decided to use for the ozone data.

4.1 Trend Estimation

In principle, it would be possible to obtain efficient estimates of the parameters of the model component $g(t)$ in (2), using the correlation structure and generalized least squares. To do this for a single year would require the inversion of an n by n covariance matrix where n, equal to 87,600 for 1993 (8,760 hours for each of 10 stations after eliminating station 4, the majority of whose values are missing), is the total number of observations. This inversion can be simplified by treating the data as a vector time series that is covariance

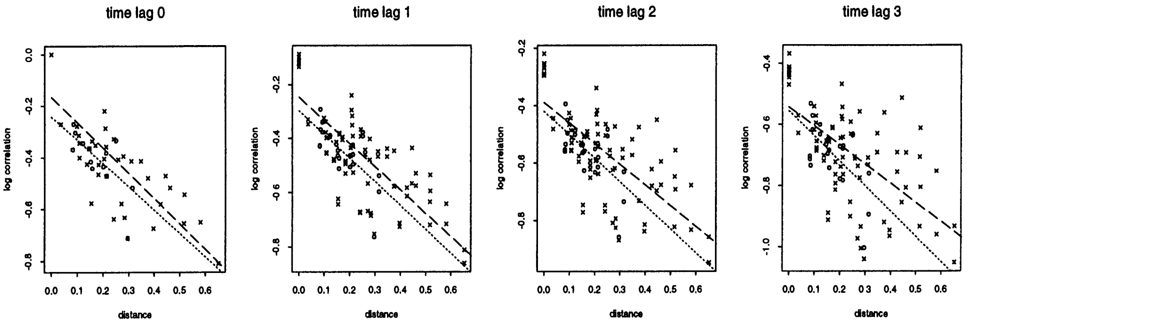

Figure 5. Estimated Correlation as Function of Distance for 1993, With × Indicating Northwestern–Southeastern Pairs of Stations and o Indicating Northeastern–Southwestern Pairs of Stations.

stationary because of the form of the correlation function of the GRF. We let N be the number of hours and p be the number of stations for a year, let $\mathbf{e} = (e_1^T, \ldots, e_N^T)^T$, where e_t contains the p values of the GRF at time t at the monitoring stations; and let $\mathbf{\Sigma} = \text{var}(\mathbf{e})$. Note that $\mathbf{\Sigma}$, which is the covariance matrix in the generalized least squares problem, is a block Toeplitz matrix, and we could use the extensive results on inverting such matrices (see e.g., Dietrich 1993 and references therein). This is still a large computational task.

Not only is the inversion a large task, but we feel that it is unnecessary. Under the missing at random assumption, ordinary least squares using only the observed data provides a consistent estimate of the deterministic trend model parameters. Although least squares may offer reduced efficiency due to lack of independence, it nonetheless provides consistent estimates of the parameters.

To further reduce computational overhead, in fitting the random process we treated the deterministic trend as known and equal to the least squares estimates. We feel that this assumption is justified by noting that because of the large number of observations, the least squares estimation error was negligible compared to the observed ozone variation. For example, the ratio of the average standard error of the estimated trend to the standard deviation of the random field was .017 for 1993.

4.2 Parameter Estimation for the Gaussian Random Field

By treating the trend parameters as known, the data can be adjusted at each location by subtracting out the estimated trend.

4.2.1 Naive Estimators. Our first estimates of the a's and b's in ρ are ordinary least squares estimates in fitting the regression model

$$\log(\hat{\rho}(d,v)) = a_0 + a_1 v + a_2 v^2 + (b_0 + b_1 v + b_2 v^2)d + \varepsilon,$$

where $\hat{\rho}(d,v)$ is the sample correlation coefficient of two stations at distance d and time lag v.

4.2.2 Maximum Likelihood Estimators. At this point, it would be attractive to use maximum likelihood to estimate the parameters of the GRF. However, finding the maximum likelihood estimators (MLE's) requires repeated inversion of the same huge covariance matrix discussed in Section 4.1. Again, because of the stationary vector time series structure of the problem, it may be possible to phrase our correlation structure into a state–space framework for the vector time series $\mathbf{e}$, and then apply Kalman filter methods to evaluate the likelihood. This would be attractive also because such a method would handle our extensive missing data (see, e.g., Shumway and Stoffer 1982).

We decided not to pursue MLE's because the computations involved are still very large. Further, because the primary use of our model is prediction, it is not obvious that fully efficient MLE's of the a's and b's are needed or are most appropriate. (See Tiao and Xu 1993 for a discussion of the relative merits of using MLE's in formulas for predictors versus using parameters optimized for prediction purposes, particularly when there is no absolute guarantee of model correctness.)

4.2.3 A Fast Cross-Validation–Type Method. Given values θ of the parameters of the GRF, let $\mathbf{z}_i(1;\theta), \ldots, \mathbf{z}_i(N;\theta)$ be the univariate time series obtained as the errors in predicting the GRF at the ith monitoring station using all the other stations; that is, the time series $\mathbf{z}_1, \ldots, \mathbf{z}_p$ are the leave-one-station-out prediction errors for the p stations. We discuss calculating such predictors in the next section.

Our proposed estimation procedure is to minimize

$$S(\theta) = \sum_{i=1}^{p} \sum_{t=1}^{N} z_i^2(t;\theta),$$

where the sum is only over times and locations where ozone is not missing.

To handle the missing observations, we use an EM-type algorithm (Dempster, Laird, and Rubin 1977; Little and Rubin 1987; Shumway and Stoffer 1982) to evaluate $S(\theta)$ for a specified θ. The first step of this algorithm is to fill in a missing residual at time t with the average of the observed residuals at time t. If there are no residuals at time t, then we fill in a 0. This gives us a full N by p matrix $\mathbf{W}_0$. We then perform a sequence of "updating cycles" giving matrices $\mathbf{W}_1$, $\mathbf{W}_2$, and so on, until there is negligible change in the matrices. Each cycle consists of a step for each monitoring station. At the jth step (for $j = 2, \ldots, p-1$), the missing data at the jth station are predicted using the observed and filled-in data for stations $1, \ldots, j-1$ from the current cycle and the observed and filled-in data for stations $j+1, \ldots, p$ from the previous cycle. The first station is predicted from the last $p-1$ of the previous cycle, whereas the last station is predicted from the first $p-1$ stations of the current cycle.

Note that ordinary leave-one-observation-out cross-validation (Allen 1974; Stone 1974; Wahba and Wold 1975) would not be suitable here, because our ultimate purpose is to predict ozone levels at locations other than monitor sites.

Our estimation procedure accomplishes two important objectives. First, it gives us a sensitivity analysis on the naive estimators; second, it provides a full dataset that we can use in prediction.

4.3 Predicting the GRF

Predicting a value of the GRF ideally consists of finding the conditional expectation of $\varepsilon(x,t)$ given all Np data points, another massive computational task. To simplify things, we consider finding the predictor at location x and time t given only the data within K time units of time t; that is,

$$\hat{\varepsilon}_K(x,t) = E[\varepsilon(x,t)|\mathbf{S}_K(\mathbf{t})],$$

where $\mathbf{S}_K(\mathbf{t}) = (\mathbf{e}_{t-K}^T, \ldots, \mathbf{e}_{t+K}^T)^T$ is a $(2K+1)p$-dimensional vector. Thus

$$\hat{\varepsilon}_K(x,t) = \gamma_K^T \mathbf{\Sigma}_K^{-1} \mathbf{S}_K(\mathbf{t}),$$

Table 1. Trend Parameter Estimates for 1993

Hour	1	2	3	4	5	6
α	5.574	5.587	5.551	5.513	5.330	5.161
Hour	7	8	9	10	11	12
α	4.844	5.133	5.854	6.497	6.934	7.244
Hour	13	14	15	16	17	18
α	7.441	7.515	7.504	7.364	7.124	6.726
Hour	19	20	21	22	23	24
α	6.179	5.764	5.557	5.449	5.432	5.486
Month	1	2	3	4	5	6
β	−.442	.888	1.067	1.567	.630	−1.644
Month	7	8	9	10	11	12
β	−2.259	−1.854	−.782	.012	−.089	0
	1	2				
γ	−.147	.00162				

where the elements of $\gamma_K = \text{cov}[\mathbf{S}_K(\mathbf{t}), \varepsilon(x,t)]$ and $\mathbf{\Sigma}_K = \text{var}[\mathbf{S}_K(\mathbf{t})]$ are easily found from the covariance function of the GRF.

To find prediction intervals, we can use the fact that

$$\begin{aligned} V_K(x,t) &\equiv E\{[\varepsilon(x,t) - \hat{\varepsilon}_K(x,t)]^2 | \mathbf{S}_K(\mathbf{t})\} \\ &= \sigma^2 - \gamma_K^T \mathbf{\Sigma}_K^{-1} \gamma_K. \end{aligned}$$

We discuss the choice of K and the effect of not using all of the data in prediction in our discussion of our results in Section 6.

4.4 Predicting Ozone

As we did with the GRF, we predict ozone level at a location x and time t given not all of the data but rather all ozone data within K time units of time t. We define the vector $\mathbf{Q}_K(\mathbf{t})$ in the same way we defined $\mathbf{S}_K(\mathbf{t})$ except for ozone rather than ε's. Then, treating the estimated trend as the true trend, we have

$$\begin{aligned} &\widehat{\text{oz}}_K(x,t) \\ &\quad = E[\text{oz}(x,t)|\mathbf{Q}_K(\mathbf{t})] = E\{[g(t) + \varepsilon(x,t)]^2 | \mathbf{S}_K(\mathbf{t})\} \\ &\quad = g^2(t) + 2g(t)E[\varepsilon(x,t)|\mathbf{S}_K(\mathbf{t})] + E[\varepsilon^2(x,t)|\mathbf{S}_K(\mathbf{t})] \\ &\quad = [g(t) + \hat{\varepsilon}_K(x,t)]^2 + V_K(x,t). \end{aligned}$$

If we let $w_K(x,t) = E[\varepsilon^2(x,t)|\mathbf{S}_K(\mathbf{t})] = \hat{\varepsilon}_K^2(x,t) + V_K(x,t)$, then we have

$$\begin{aligned} &\text{var}[\text{oz}(x,t)|\mathbf{Q}_K(\mathbf{t})] \\ &\quad = E(\{\text{oz}(x,t) - E[\text{oz}(x,t)|\mathbf{Q}_K(\mathbf{t})]\}^2 | \mathbf{Q}_K(\mathbf{t})) \\ &\quad = E\{[\text{oz}(x,t) - \widehat{\text{oz}}_K(x,t)]^2 | \mathbf{Q}_K(\mathbf{t})\} \\ &\quad = E(\{2g(t)[\varepsilon(x,t) - \hat{\varepsilon}_K(x,t)] + \varepsilon^2(x,t) \\ &\qquad\quad - w_K(x,t)\}^2 | \mathbf{S}_K(\mathbf{t})) \\ &\quad = 4g^2(t)V_K(x,t) + 4g(t)\text{cov}[\varepsilon(x,t), \varepsilon^2(x,t)|\mathbf{S}_K(\mathbf{t})] \\ &\qquad + \text{cov}[\varepsilon^2(x,t), \varepsilon^2(x,t)|\mathbf{S}_K(\mathbf{t})]. \end{aligned}$$

Because the ε's are jointly normally distributed, the ε's given $\mathbf{S}_K(\mathbf{t})$ are also normally distributed and

$$\varepsilon(x,t)|\mathbf{S}_K(\mathbf{t}) \sim N(\hat{\varepsilon}_K(x,t), V_K(x,t)).$$

Hence

$$\text{cov}[\varepsilon(x,t), \varepsilon^2(x,t)|\mathbf{S}_K(\mathbf{t})] = 2\hat{\varepsilon}_K(x,t)V_K(x,t)$$

and

$$\begin{aligned} &\text{cov}[\varepsilon^2(x,t), \varepsilon^2(x,t)|\mathbf{S}_K(\mathbf{t})] \\ &\qquad\qquad = 4\hat{\varepsilon}_K^2(x,t)V_K(x,t) + 2V_K^2(x,t), \end{aligned}$$

and thus

$$\begin{aligned} &\text{var}[\text{oz}(x,t)|\mathbf{Q}_K(\mathbf{t})] \\ &\quad = 4g^2(t)V_K(x,t) + 8g(t)\hat{\varepsilon}_K(x,t)V_K(x,t) \\ &\qquad + 4\hat{\varepsilon}_K^2(x,t)V_K(x,t) + 2V_K^2(x,t) \\ &\quad = 4[g(t) + \hat{\varepsilon}_K(x,t)]^2 V_K(x,t) + 2V_K^2(x,t). \end{aligned}$$

5. RESULTS

In this section we describe the results of applying the methods of the previous section to the ozone data. Because of the amount of data involved and the possibility of changes in the model parameters over time, we have analyzed each of the 14 years of data separately.

5.1 Trend Estimates

In Table 1 we give parameter estimates for the trend part of the model for 1993. The results for other years are similar.

Table 2. GRF Parameter Estimates for 1980–1993

Year	a_0	a_1	a_2	b_0	b_1	b_2	σ^2
80	−.1757	−.1608	−.0051	−1.8354	.2942	−.0205	4.672
81	−.2359	−.1310	−.0109	−1.4770	.1654	.0004	4.588
82	−.3147	−.1085	−.0104	−1.0178	.1493	−.0038	4.487
83	−.2815	−.1248	−.0097	−1.1111	.1721	−.0090	4.318
84	−.2453	−.1190	−.0115	−1.0405	.1807	−.0113	3.863
85	−.2426	−.1188	−.0068	−.9835	.1737	−.0094	3.776
86	−.3506	−.1328	−.0091	−.9229	.1983	−.0155	3.469
87	−.3244	−.1384	−.0065	−1.0486	.1985	−.0090	3.763
88	−.2646	−.1303	−.0045	−.7383	.1768	−.0118	3.909
89	−.2493	−.1091	−.0102	−1.0800	.1704	−.0072	3.905
90	−.2781	−.1125	−.0076	−.8153	.1938	−.0160	4.035
91	−.2096	−.1216	−.0063	−.9088	.1962	−.0135	3.496
92	−.2159	−.1190	−.0083	−.9209	.1612	−.0137	2.723
93	−.1708	−.1131	−.0044	−.9582	.1563	−.0142	2.431

Table 3. Estimates of log ψ_V in GRF Correlation Function for 1980–1993

	Time lag						
Year	*0*	*1*	*2*	*3*	*4*	*5*	*6*
1980	−.176	−.342	−.518	−.704	−.900	−1.107	−1.324
1981	−.236	−.378	−.542	−.727	−.935	−1.164	−1.416
1982	−.315	−.434	−.573	−.734	−.915	−1.117	−1.339
1983	−.282	−.416	−.570	−.743	−.936	−1.148	−1.379
1984	−.245	−.376	−.530	−.706	−.906	−1.129	−1.375
1985	−.243	−.368	−.507	−.660	−.827	−1.006	−1.200
1986	−.351	−.492	−.653	−.831	−1.027	−1.242	−1.475
1987	−.324	−.469	−.627	−.798	−.982	−1.180	−1.390
1988	−.265	−.399	−.543	−.696	−.857	−1.028	−1.207
1989	−.249	−.369	−.508	−.669	−.849	−1.051	−1.272
1990	−.278	−.398	−.533	−.684	−.849	−1.030	−1.226
1991	−.210	−.338	−.478	−.631	−.797	−.976	−1.167
1992	−.216	−.343	−.487	−.647	−.824	−1.017	−1.227
1993	−.171	−.288	−.415	−.550	−.694	−.847	−1.009

Note that hour 1 corresponds to the hour between 12:00 A.M. and 1:00 A.M. Therefore, the low coefficients corresponding to hours 6 to 8 indicate that the early morning between 5 and 7 A.M. has the lowest hourly average. Similarly, the early afternoon hours between 1 and 4 P.M. have the highest averages.

The estimated β's, the last of which is constrained to be 0, represent the monthly adjustments to the average ozone measurements. The low coefficients in the summer months are due to the fact that temperature variables are included in the model.

5.2 The GRF Covariance Function

In Table 2 we give the estimates of the parameters of the covariance function of the GRF for each year 1980–1993 as well as the sample variances of the GRF. These estimates are based on using $K = 6$ in the prediction algorithm. We chose $K = 6$ by using values 1, 2, and so on until the estimates changed very little. The estimators appear to be quite consistent across years, and the estimates of a_0 do not appear to be estimating 0. Because of the excellent fit of the quadratic functions to the slopes and intercepts, our final estimates of the a's and b's are actually very close to the naive estimates described earlier.

Tables 3 and 4 contain the logarithms of the ψ's and ϕ's corresponding to the a's and b's. For 1993, the leave-one-station-out prediction mean squared error for the detrended random field data is .9768, compared to the variance of the detrended random field itself, which is 2.4306, giving an R^2-type measure of $1 - (.9768/2.4306) = .598$. For the other years, this R^2-type measure ranges from 42% to 60%.

6. ESTIMATING POPULATION OZONE EXPOSURE

A major payoff of the space–time model for ozone intensity is that it can be used to construct meaningful indices of population exposure to ozone. Such indices are obtained by combining the ozone intensity maps with the actual population density. The motivation for doing this is clear: High levels of ozone matter in particular if they occur where population density is high. We have focused our efforts on children age 5 and younger. We have done so for two reasons: (1) We want to consider a "sensitive group" within the general population; and physiological development of young children may be impaired by exposure to ozone; and (2) the lack of mobility of young children makes a population exposure index weighted by their population densities at a given area more meaningful than one for a more mobile subpopulation.

Table 4. Estimates of log ϕ_V in GRF Correlation Function for 1980–1993

	Time lag						
Year	*0*	*1*	*2*	*3*	*4*	*5*	*6*
1980	−1.835	−1.562	−1.329	−1.138	−.987	−.878	−.810
1981	−1.477	−1.311	−1.145	−.977	−.809	−.640	−.470
1982	−1.018	−.872	−.735	−.604	−.482	−.367	−.260
1983	−1.111	−.948	−.803	−.676	−.567	−.476	−.404
1984	−1.041	−.871	−.724	−.600	−.498	−.418	−.362
1985	−.984	−.819	−.674	−.547	−.439	−.350	−.280
1986	−.923	−.740	−.588	−.467	−.377	−.318	−.290
1987	−1.049	−.859	−.687	−.534	−.398	−.280	−.181
1988	−.738	−.573	−.432	−.314	−.220	−.149	−.102
1989	−1.080	−.917	−.768	−.633	−.513	−.407	−.316
1990	−.815	−.637	−.492	−.378	−.296	−.246	−.228
1991	−.909	−.726	−.570	−.441	−.339	−.264	−.216
1992	−.921	−.773	−.653	−.560	−.495	−.457	−.446
1993	−.958	−.816	−.703	−.618	−.561	−.533	−.533

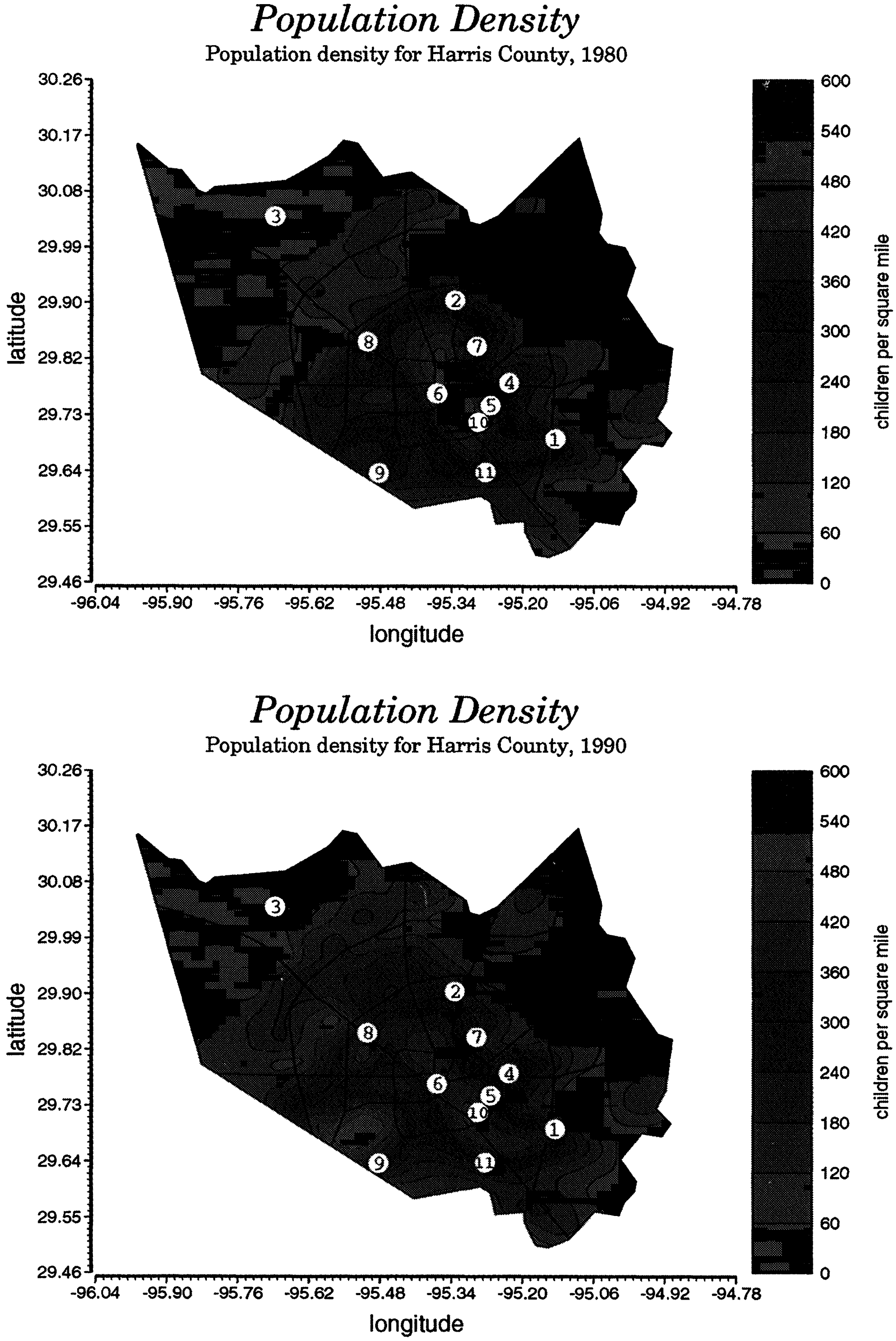

Figure 6. Population Densities for Children Age 5 and Younger for 1980 and 1990. Ozone monitoring stations are marked with the station number. Dotted lines are the major highways in Harris County. Longitude is negative to reflect the fact that it is degrees west. A degree of latitude is approximately 69.2 miles, whereas at 30 degrees latitude, a degree of longitude is approximately 64.4 miles. The thickness of the contour lines serve as a visual cue to the rate at which the values change in the region. (The wider the contour lines, the more gradual the change.)

The section is outlined as follows. We first display maps of population densities, we then discuss the concept of exposure indices. Finally, we pick a particular exposure index and display its trend since 1980.

6.1 Population Density

To determine the density of young children in Harris County at the beginning and near the end of our time period, we obtained the number of children age 5 and younger at each of the 515 census tracts for 1980 and the 582 tracts for 1990. (Some tracts split or combined between 1980 and 1990.) In the 1990 data, the census includes a latitude and longitude for each tract, while we were able to reconstruct what latitude and longitude were in 1980 using a census tracts compatibility table. Given the latitudes and longitudes and the numbers of children, we built the population density maps displayed in Figure 6 in a manner analogous to kernel density estimation with a uniform kernel. We superimposed a 25 by 25 rectangular grid on Harris County. We calculated the density for a cell within the grid by summing up the population sizes of all tracts within one bandwidth from the center of the cell and then dividing by the number of cells within the counting range. We chose the bandwidth to be one cell size to keep the features of the data and to obtain a smooth contour. From the 25 by 25 grid, we drew the maps using the same methods we used in drawing the ozone maps in Figure 8.

To obtain population densities between 1980 and 1990, we used linear interpolation between the 1980 and 1990 measurements, assuming that the population size at each location was either linearly increasing or linearly decreasing. We also used the same linear functions to extrapolate the population densities for the three years beyond 1990.

Within each year, we treat the population density as constant, and we denote the value at a particular location x by $p(x)$. Figure 6 displays the population densities of children age 5 and younger for 1980 and 1990. Note that both graphs are on the same scale, and that the volumes under the two surfaces are equal to the total populations of children for the two years. (The populations of people of all ages in Harris County were 2,409,547 and 2,818,199 in 1980 and 1990, whereas the populations of children age 5 and under were 238,417 and 290,545.) The growth in total population from 1980–1990 is clear, particularly in the decrease in the amount of low population areas from 1980–1990. Also, the growth of population away from the south-central Houston area toward all other areas is obvious.

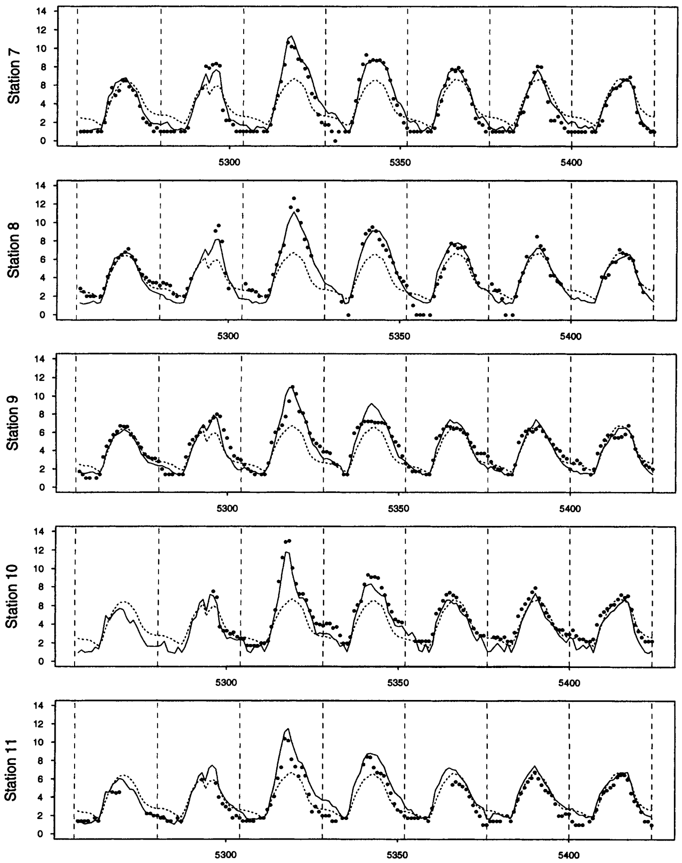

Figure 7. Prediction of Transformed Ozone for August 8–14, 1993. The dots stand for observed data, the dashed lines for trend-only prediction, and the solid lines for trend-plus-random-field prediction.

One of the more striking aspects of Figure 6 concerns the placement of monitoring sites. Although the sites in 1980 were located largely near areas with high population density (the locations were largely consistent with those in 1993), by 1993 an important area of high population density (the area southwest of the center of Houston) was served by only one monitoring site.

6.2 Exposure Indices

We define the exposure index $e(A, T, \mathbf{z})$ in area A over some time period T by

$$e(A, T, \mathbf{z}) = \int_A \int_T w(x, t, z_1) i(x, t, z_2)\, dx\, dt, \qquad (3)$$

where $w(x, t, z_1)$ is a weight function tailored to the purpose at hand, $i(x, t, z_2)$ is the ozone intensity, and $\mathbf{z} = (z_1, z_2)$ is a vector of covariate variables. If $w(x, t, z_1) = p(x, t, z_1)$ in (3), where $p(x, t, z_1)$ is the population density identified by demographic variables z_1, then we call $e(A, T, z)$ a population exposure index in area A over time T for covariate z_1. It is important to make the distinction between the average amount of ozone, obtained using $w(x, t, z_1) \equiv 1$, and the per-person exposure to ozone. The former is an interesting measure, but ultimately it is the latter that is of public health interest.

Various exposure indices, based on average ozone or average per-person exposure, have been discussed in the literature. A critical defining factor is the time period of interest, which generally can be classified into two categories: long-term periods and short-term high-concentration periods. For instance, average ozone contour maps over a given time period T can be formed by plotting $e(x, T, \mathbf{z})$ for a given T with $w(x, t, z_1)$ equal to 1.0; that is, without taking demographics into account. Population exposure contour maps use $w(x, t, z_1)$ equal to the population density. Lefohn and Runeckles (1987) have found that different ozone exposure indices (long-term or short-term) should be used to explain crop losses for different crop species. Similar results were also found in human and animal studies (National Research Council 1991, pp. 31–37). As mentioned at the beginning of this section, the two major categories considered are long-term average exposure and short-term high-concentration exposure. Although we have performed many other analyses, we focus on the time period March–October, 10:00 A.M.–6:00 P.M.,—the period with the highest ozone levels in 1993.

Average exposure and per-person population exposure indices based on (3) can be easily calculated as follows. Let A_G be a grid of points (x_i, t_j), where location x_i is in area A and time t_j is in period T. Given the ozone and population exposure maps for a time period T over A_G, we can calculate

$$E_{x_i,T} = \frac{\sum_{j=1}^{N_T} \mathrm{oz}(x_i, t_j)}{N_T}, \qquad (4)$$

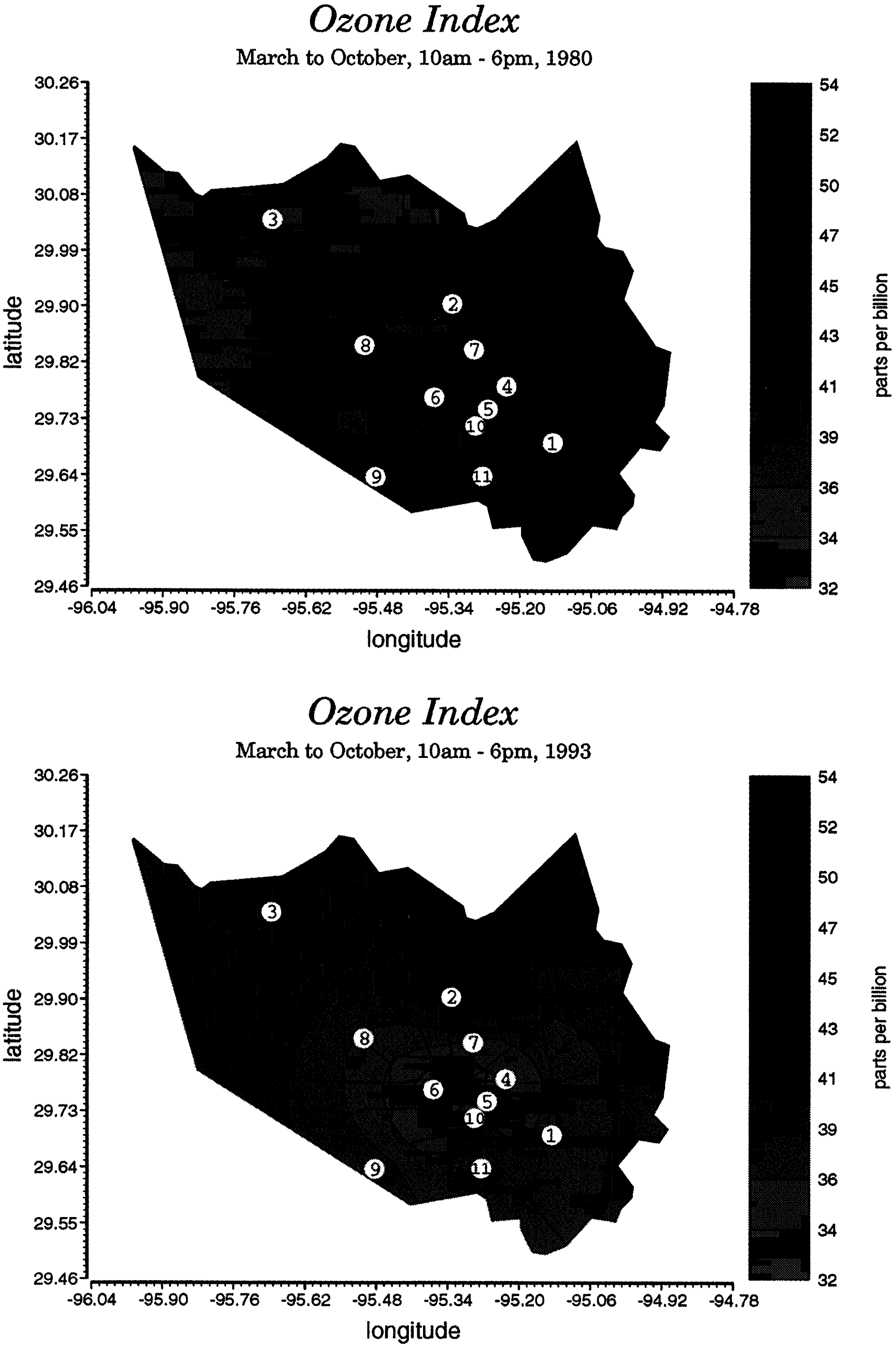

Figure 8. Ozone Maps for 1980 and 1993.

$$E_{A_G,T} = \frac{\sum_{i=1}^{N_G} E_{x_i,T}}{N_G}, \qquad (5)$$

and

$$PE_{A_G,T} = \frac{\sum_{i=1}^{N_G} p(x_i) E_{x_i,T}}{\sum_{i=1}^{N_G} p(x_i)}, \qquad (6)$$

where N_T is the number of time observations in T and N_G is the number of locations in G. Note that (5) and (6) are natural estimates of (3) when the $w(x, t, z_1)$ equals 1.0 (average ozone) and $p(x)$ (population exposure to ozone). Finding standard errors of these indices is then a straightforward but time-consuming calculation involving the covariances of $\widehat{oz}$.

6.3 Trends in Population Exposure

Given parameter estimates of the spatial–temporal model, we can predict ozone at any time and place within Harris County. In Figure 7 we display the results for one summer week in 1993 at five monitoring stations. The time series plots (dots) of the (transformed) ozone measurements for one summer week are given, with the corresponding deterministic trends (dashed lines) and trend-plus-random-field predictions (solid lines) overlaid. The important feature of these graphs is that our model tracks peak ozone levels much better than just using the trend.

To further illustrate prediction using our model, consider the ozone maps for 1980 and 1993 in Figure 8, which we constructed as follows. First, we placed a 25 by 25 grid within the county, and at each time between 10:00 A.M. and 6:00 P.M. during March–October 1993 (the times when ozone levels are high in Houston), we predicted the value of ozone at each point in the grid. Then, to summarize ozone for the entire period, we averaged across time the predictors at each grid point. Finally, we drew the map itself using the numerical interpolation routine in the MISHA software developed by Hardin and Schmiediche (1992) and Schmiediche and Hardin (1993). We chose a 25 by 25 grid for two reasons. First, it seemed to be the coarsest grid that we could use that preserved the features of ozone and also led to a smooth map. Second, the size of the grid determines how long it takes to calculate standard errors of further indices based on the map, and it is desirable to make the grid as coarse as possible. Figure 8 shows the dramatic decrease in ozone from 1980 to 1993, particularly in the center of Houston.

With the ability to predict ozone level at any time and location, the annual population exposure indices described in Section 6.2 can be calculated. These are the quantities of most interest.

From an ozone map, we can construct a single annual population exposure index for the 10:00 A.M. to 6:00 P.M., March–October time period, by combining the mapped values with the population density at the 25 by 25 grid. Figure 9a displays these indices for each year from 1980–1993. The display is striking, because it shows a steady decrease in ozone exposure for young children, typically on the order of 20% from the early 1980s to the present .The standard deviation of the averages of the predicted ozone values can be calculated, because it is only a function of the prediction covariances. The index for a single year is an average over our 25 by 25 grid and over all the hours from 10:00 A.M. to 6:00 P.M. in March–October. The covariance of two ozone predictors is a function of the covariance matrix Σ. With $K = 6$, it took approximately 20 cpu hours on a Sparc 20 to calculate one standard deviation. For 1993, the standard deviation of the index is .15; those for other years range from .15 to .22.

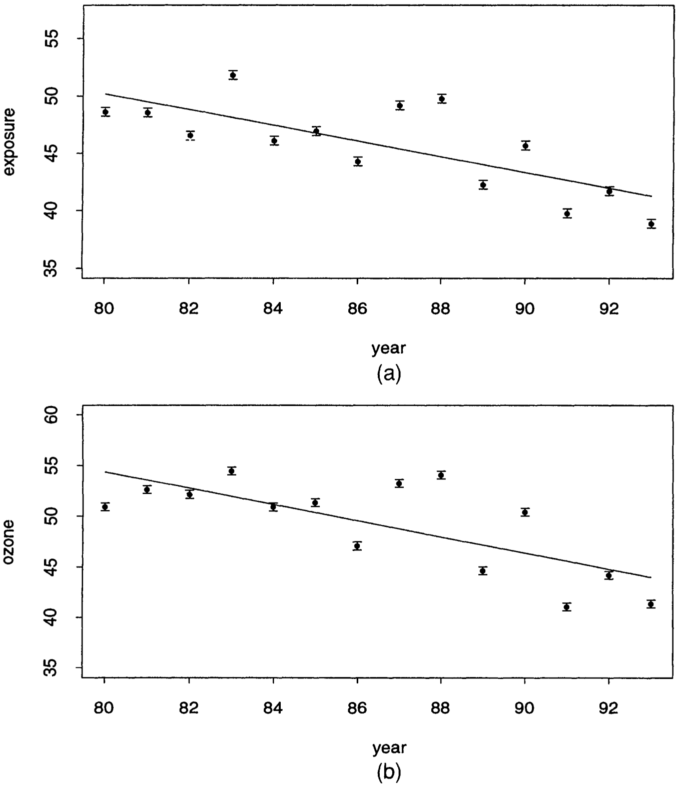

Figure 9. Predicted Average Population Exposure (a) and Predicted Average Ozone (b) Throughout Harris County for March–October, 10:00 A.M.–6:00 P.M., 1980–1993.

Figure 9b shows the average ozone level (an exposure index, if the population is uniformly distributed) from 1980–1993. Again, we see a substantial decrease. For 1993, the standard deviation is .19; those for other years are similar.

7. CONCLUSIONS

Ozone levels have clearly decreased in Harris County, Texas in terms of the average ozone measurements throughout the county, the average ozone in different parts of the county, and in the per-person ozone computed taking population trends for young children into account. A simple station-by-station analysis also shows decreases in ozone. The distinguishing feature of our problem is that by focusing on per-person exposure, we have been forced to build a model for any point in Harris County at any time from 1980–1993, and not just as fixed monitoring stations. The need for such a model has led us to consideration of two types of predictor variables: those that are measured with reasonable quality and show low spatial variability and hence are predictable throughout the county, and those of lower quality or with large spatial variability, which are thus not predictable away from the monitoring stations. The spatial–temporal model that we built has a deterministic component based on time and temperature, with a GRF component that accounts for other factors as well as random variation.

The deterministic component of the model is simply estimated by ordinary least squares. The more interesting statistical problem is the estimation of parameters in the GRF. Here we took account of three important features: (1) The dataset is itself huge, making methods such as maximum likelihood infeasible; (2) there are substantial amounts of missing data; and (3) the goal is prediction throughout the county. We have developed a simple, generally applicable estimation procedure that optimizes prediction variance and handles missing data.

We carried out an initial application of our procedure. As indicated by Figure 9a, it appears that exposure to ozone among young children is decreasing substantially over time. We have performed many other analyses of these data, all of which point to the same conclusion.

[Received January 1995. Revised September 1996.]

REFERENCES

Allen, D. M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method of Prediction," *Technometrics*, 16, 125–127.

Bloomfield, P., Royle, A., and Yang, Q. (1993a), "Accounting for Meteorological Effects in Measuring Urban Ozone Level and Trends," Technical Report 1, National Institute of Statistical Sciences, Research Triangle Park.

——— (1993b), "Rural Ozone and Meteorology: Analysis and Comparison With Urban Ozone," Technical Report, 5, National Institute of Statistical Sciences, Research Triangle Park.

Bruntz, S. M., Cleveland, W. S., Graedel, T. E., Kleiner, B., and Warner, J. L. (1974), "Ozone Concentrations in New Jersey and New York: Statistical Association With Related Variables, *Science*, 186.

Cox, W. M., and Chu, S-H. (1992), "Meteorology-Adjusted Ozone Trends in Urban Areas: A Probability Approach," *Atmospheric Environment*, 27B, 425–434.

Cressie, N. A. (1989), "Geostatistics," *The American Statistician*, 43, 197–202.

——— (1993), *Statistics for Spatial Data*, New York: Wiley-Interscience.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Dietrich, C. R. (1993), "Computationally Efficient Cholesky Factorization of a Covariance Matrix With Block Toeplitz Structure," *Journal of Statistical Computation and Simulation*, 45, 203–218.

Guttorp, P., Meiring, W., and Sampson, P. D. (1994), "A Space–Time Analysis of Ground-Level Ozone Data," *EnvironMetrics*, 5, 241–254.

Handcock, M. S., and Wallis, J. R. (1994), "An Approach to Statistical Spatial–Temporal Modeling of Meteorological Fields," *Journal of the American Statistical Association*, 89, 368–390.

Hardin, J. W., and Schmiediche, H. (1992), "MISHA: A Computational and Graphical Tool," *Computing Science and Statistics, Proceedings of the 24th Symposium on the Interface*, pp. 293–297.

Haslett, J., and Raftery, A. E. (1989), "Space–Time Modeling With Long-Memory Dependence: Assessing Ireland's Wind Power Resource," *Applied Statistics*, 38, 1–21.

Horowitz, J. (1980), "Extreme Values for Nonstationary Stochastic Processes: An Application to Air Quality Analysis," *Technometrics*, 22, 469–478.

Isaaks, E. H., and Srivastava, R. M. (1990), *An Introduction to Applied Geostatistics*, New York: Oxford University Press.

Journel, A. G., and Huijbregts, C. J. (1979), *Mining Geostatistics*, New York: Academic Press.

Lamb, B., Guenther, A., Gay, D., and Westberg, H. (1987), "A National Inventory of Biogenic Hydrocarbon Emissions," *Atmospheric Environment*, 21, 1695–1705.

Laslett, G. M. (1994), "Kriging and Splines: An Empirical Comparison of Their Predictive Performance in Some Applications," *Journal of the American Statistical Association*, 89, 391–400.

Lefohn, A. S., and Runeckles, V. C. (1987), "Establishing Standards to Protect Vegetation-Ozone Exposure/Dose Considerations," *Atmospheric Environment*, 21, 561–568.

Lippmann, M. (1989), "Health Effects of Ozone: A Critical Review," *Journal of the Air Waste Management Association*, 39, 672–695.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

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

Niu, X-F. (1996), "Nonlinear Additive Models for Environmental Time Series, With Applications to Ground-Level Ozone Data Analysis," *Journal of the American Statistical Association*, 91, 1310–1321.

Pagnotti, V. (1990), "Statistical Ozone Levels and Control by Seasonal Meteorology," *Journal of the Air Waste Management Association*, 40, 206–210.

Schmiediche, H., and Hardin, J. W. (1993), "Graphics Keys: A Resource Database Approach to Extensible Graphics," in *Proceedings of Statistical Graphics Section, American Statistical Association*, pp. 1–10.

Shumway, R. H., and Stoffer, D. S. (1982), "An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm," *Journal of Time Series Analysis*, 3, 253–264.

Smith, R. L., and Huang, L-S. (1993), "Modeling High Threshold Exceedances of Urban Ozone," Technical Report 6, National Institute of Statistical Sciences, Research Triangle Park.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Tiao, G. C., and Xu, D. (1993), "Robustness of Maximum Likelihood Estimates for Multi-Step Predictions; The Exponential Smoothing Case," *Biometrika*, 80, 623–641.

Wahba, G. (1983), "Bayesian Confidence Intervals for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society*, Ser. B, 45, 133–150.

Wahba, G., and Wold, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics*, Ser. A, 4, 1–17.