Statistical Models for Monitoring and Regulating Ground-level Ozone

Eric Gilleland[1] and Douglas Nychka[2]

Abstract

The U.S. Environmental Protection Agency's (EPA) National Ambient Air Quality Standard (NAAQS) for ground-level ozone is now based on the fourth-highest daily maximum 8-hour average ozone level (FHDA). Standard geostatistical models may not be appropriate for interpolating such a statistic off of a network of monitoring sites. In this paper we compare the performance of different statistical models in predicting this standard at locations where monitors are not located. We give special attention to two models: a daily model that uses an autoregression to account for spatial and temporal dependence, and a seasonal model that assumes the FHDA field is Gaussian and employs spatial statistical techniques. Based on five seasons of ozone data collected in and around North Carolina, we find that the daily model is superior enough to the seasonal model to warrant its added complexity.

1 Introduction

The application of statistical techniques to environmental problems often involves a tradeoff between simple methods that are easily implemented and interpreted and more complicated methods that may have smaller errors. In this paper we compare simple and complicated statistical models for interpolating the U.S. Environmental Protection Agency (EPA) National Ambient Air Quality Standard (NAAQS) for ground-level ozone off of a network of monitoring sites. A recent change in the NAAQS for ground-level ozone is based on the fourth-highest value from the daily sequence of maximum 8-hour average ozone measured over an ozone season. This order statistic will be referred to here as FHDA. The standard requires that the average FHDA for three consecutive seasons (years) be below 84 parts per billion (ppb) in order for a location to be in attainment. Understanding the spatial distribution for the FHDA for a given ozone season presents a new statistical problem for infering regions of attainment or nonattainment because it is not clear that the FHDA field (a field of order statistics) is Gaussian–a fundamental assumption of most standard spatial statistical techniques. This problem is addressed in this work.

Although the use of spatial statistics for interpolating air quality measurements would not be disputed by a statistical audience, surprisingly the use of monitoring data in a regulatory context is often limited to point locations. For example, a region that does not have a monitoring station would not have any information about the amount of ozone exposure on the population or environment

[1] *Corresponding author address*: Eric Gilleland, National Center for Atmospheric Research, Research Applications Program, Boulder, CO 80307-3000; email: ericg@ucar.edu

[2] National Center for Atmospheric Research, Geophysical Research Project

of this region from the EPA; even if a contiguous region with monitoring has violated air quality standards. Accordingly, Holland *et al.* [13] argue for the introduction of modern statistical methods to understand the spatial and temporal extent of pollution fields based on monitoring data. Given the range of statistical backgrounds associated with the regulatory community, it is appropriate to propose statistical methods that are simple and understandable to a broad group when such methods provide an accurate and defensible analysis. In particular, for interpolating the FHDA standard, it is useful to ascertain the feasibility of approximate statistical methods that treat the FHDA statistics directly. From this perspective we compare two statistical models. The first, a fairly complex model, uses a spatial AR(1) model for daily ozone measurements and samples the FHDA field conditional on the daily data for the entire season. This approach will be referred to as the *daily model*. The second model, referred to as the *seasonal* model, is a geostatistical model that predicts the FHDA field from the network values using standard best linear unbiased prediction, or kriging (see Cressie [1] or Stein [17] for more details on kriging). This seasonal model is similar to the model proposed by Fuentes [6], except that the region of interest here is much smaller and so can be assumed to be spatially and temporally stationary. A third approach that will be used as a benchmark estimates the FHDA field by way of a thin plate spline (see Green and Silverman [11] or Hastie and Tibshirani [12] for details on thin plate splines). This last method is generic and uses the least amount of information concerning the actual air quality context.

The paper is organized as follows. Section 2 describes the ozone monitoring data. Section 3 describes how the ozone data is standardized and presents the daily and seasonal models. Section 4 presents a comparison of these models along with thin plate spline interpolation as a benchmark. One interesting aspect is that the daily model implies a covariance function for the FHDA field, and this choice is included in the results for the seasonal approach. Our results suggest that the daily model is worth the extra effort and complexity. Section 5 discusses the results and suggests some future research.

2 Ozone Monitoring Data

Data used for these analyses consist of maximum daily 8-hour average ozone levels measured in parts per billion (ppb) for the 72 monitoring stations in a study region centered on North Carolina (Fig. 1). The dashed rectangle in Fig. 1 shows a region around the Research Triangle Park (RTP) in North Carolina in which the models will be interpolated onto a 15×15 grid. These data are a subset of the 513 stations covering the eastern United States used by Fuentes [6] and can be obtained from the web through `http://www.cgd.ucar.edu/stats/Data`.

Partly because of the high cost of operation, ozone monitoring occurs only during the hotter

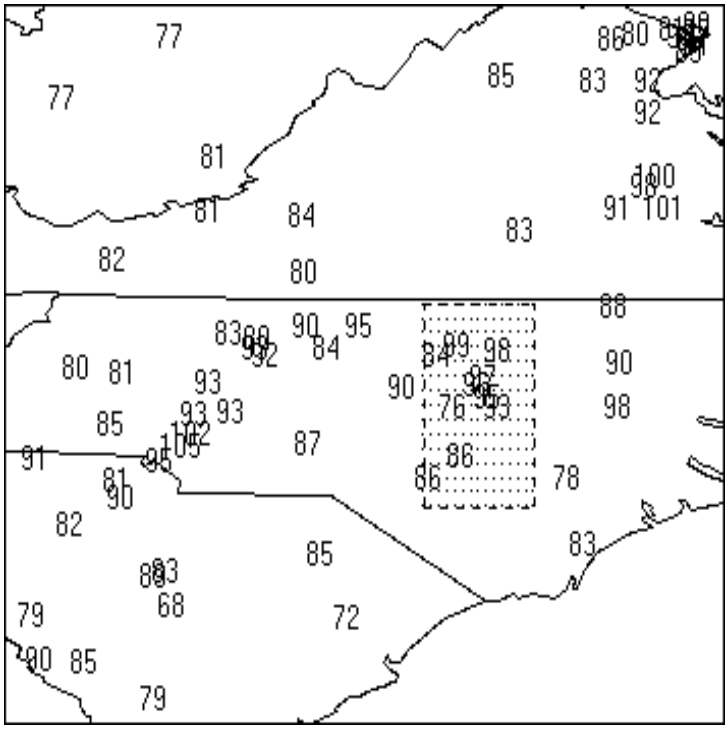

Figure 1: *Network of ozone monitoring locations with a rectangular region (with* 15×15 *grid) in the Research Triangle Park (RTP) in North Carolina. The numbers represent the fourth-highest value for 1997.*

months when weather conditions are most conducive to forming ozone; this "ozone season" essentially spans the months from April through October. For these analyses, the data cover five seasons (1995-1999), with each season consisting of 184 days.

From a total of $66,240$ possible daily data measurements, there are fewer than $5,000$ missing observations. Although the data have missing values, it is not serious enough in estimating the autoregressive models for the individual stations; and, in our judgement, each day has a reasonable representation of reporting stations. In effect, we assume that the missing data are missing at random so that they do not affect parameter inferences.

3 Daily model for ozone and a seasonal model for FHDA

Ideally one would like to know the complete multivariate distribution of the FHDA field. This would enable us to not only interpolate to values off of the monitoring network, but to better understand the error associated with such an interpolation. Because the FHDA statistic is an order statistic based on a serially correlated sample, it is difficult to derive a form for its distribution. As an

alternative, we suggest a model for daily ozone measurements and then infer the distribution of the FHDA through aggregating the daily model over the season. The daily model uses a spatial AR(1) model to model the daily maximum 8-hour averages and then uses Monte Carlo sampling to approximate the conditional distribution of the FHDA field given the observed network data. One important practical issue is whether this constructive and rigorous approach has any benefits over a simple spatial interpolation of the FHDA sample statistics. We divide the presentation of the model into a preliminary standardization, an autoregressive time series for individual stations and a spatial model for the innovations. Section 3.3 describes the algorithm to simulate a process from the daily model.

3.1 Standardizing the Data

Ozone has a seasonal dependence even during the relatively short ozone season described in section 2. It is useful to account for this seasonality as a fixed effect before modeling space-time structure. Let $O(\mathbf{x}, t)$ denote the maximum 8-hour average ozone at location $\mathbf{x}$ and day t. The following standardization is used for the daily maximum 8-hour ozone measurement

$$O(\mathbf{x}, t) = \mu(\mathbf{x}, t) + \sigma(\mathbf{x})u(\mathbf{x}, t), \tag{1}$$

and we assume that $u(\mathbf{x}, t)$ for any given location and time has mean zero and variance 1; and $\sigma(\mathbf{x})$ is estimated from the residuals, and is allowed to vary over space. Note that μ is a function of both seasonality and space, and the variation of $\mu(\mathbf{x}, t)$ over space has the potential to account for large-scale spatial trend.

The seasonal means are smoothed over space using a singular value decomposition approach. The m individual station time series are regressed on an intercept and three sine and cosine pairs with periods 365, 365/2 and 365/3; and let $\mathbf{B}$ denote the $m \times 7$ matrix of regression coefficients across all the stations. Next, $\mathbf{B}$ can be decomposed as $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{D}$ is a diagonal matrix of the singular values of $\mathbf{B}$. By setting some of the singular values of $\mathbf{D}$ to zero (call the resulting matrix $\mathbf{D}^*$), the multiplication $\mathbf{B}^* = \mathbf{U}\mathbf{D}^*\mathbf{V}^T$ yields a matrix of the original regression parameters, but having reduced the parameter variability across stations. For the analyses here, the first three principal components explained about 96% of the variance, and thus were retained; and, in this case, results in smoothing the estimated parameters over space. Finally, the estimates of μ and σ based on station locations are extrapolated to unobserved locations using thin plate spline interpolation.

3.2 Daily Model

Given the standardized process, $u(\mathbf{x}, \mathbf{t})$ (from (1)), we consider a spatial AR(1) model.

$$u(\mathbf{x}, t) = \rho(\mathbf{x})u(\mathbf{x}, t-1) + \varepsilon(\mathbf{x}, t) \tag{2}$$

The shocks, $\varepsilon(\mathbf{x}, t)$, are assumed to be independent over time and to be a mean zero Gaussian process over space with spatial covariance

$$\sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x}')}\psi(d(\mathbf{x}, \mathbf{x}')) \tag{3}$$

Here, $\psi(d(\mathbf{x}, \mathbf{x}'))$ is considered to be an isotropic and stationary correlation function where $d(\mathbf{x}, \mathbf{x}')$ is a metric to measure separation between locations. For this application, the great circle distance, transformed to ensure positive definiteness, is used for $d(\mathbf{x}, \mathbf{x}')$. That is, $d(\mathbf{x}, \mathbf{x}') = 2\sin(\frac{h}{2})$, where $h \in [0, \pi]$ denotes the angular great circle distance between $\mathbf{x}$ and $\mathbf{x}'$ (see Gneiting [9], Weber and Talkner [18], Freeden et al. [5] sec. 5.8.1, and Gaspari and Cohn [7]). Model (2) with covariance (3) implies a space-time covariance function

$$\mathrm{Cov}(u(\mathbf{x}, t), u(\mathbf{x}', t-\tau)) = \frac{(\rho(\mathbf{x}))^\tau \sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x}')}\psi(d(\mathbf{x}, \mathbf{x}'))}{1-\rho(\mathbf{x})\rho(\mathbf{x}')}, \tau = 0, 1, 2, \ldots \tag{4}$$

Thus, if the AR(1) parameters are not constant over space, then (i) the spatial process $u(\mathbf{x}, t)$ is not stationary and (ii) covariance model (4) is not space-time separable. Covariance model (4) is not fully symmetric because $C(u(\mathbf{x}, t), u(\mathbf{x}', t-\tau))$ and $C(u(\mathbf{x}', t), u(\mathbf{x}, t-\tau))$ generally differ from each other (see Gneiting [8] for a discussion of full symmetry). The violation of full symmetry is physically justifiable here because ozone is often transported by wind in only one direction.

3.3 Sampling the distribution of FHDA conditioned by the monitoring data

Under the assumption that all the components of the data model are known, there is a straightforward algorithm for sampling the FHDA field conditional on the observed data. This algorithm is quite efficient and uses the autoregressive structure over time to recursively generate the daily process. Let $\mathbf{x}_0$ be a location where ozone is unobserved. A spatial prediction for the FHDA at this location involves two steps. One first obtains a sample of the time series of daily ozone measurements at this location conditional on the observed data (for all locations and all times). Next one calculates the FHDA for this series. By elementary probability, the resulting FHDA statistics will be a sample of the FHDA field at $\mathbf{x}_0$ conditional on the data. Repeating these two steps, one can generate a random sample that approximates the FHDA conditional distribution; and, of course, the sample mean is a point estimate for the conditional expectation of the FHDA at $\mathbf{x}_0$. The conditional variance can be used as a measure of uncertainty.

Sampling from the conditional distribution of the ozone field is simplified by the autoregressive structure over time and the restriction of spatial dependence to the shocks in the AR(1) innovation. In this section we will assume that all parameters $(\mu(\mathbf{x},t), \sigma(\mathbf{x}), \rho(\mathbf{x}), \psi)$ are fixed quantities and known, but more will be said about this assumption in section 5. Also let $\{\mathbf{x}_k$, for $1 \leq k \leq m\}$ be the station locations. Based on these assumptions it is sufficient to find the conditional distribution of $\{u(\mathbf{x}_0,t), 1 \leq t \leq T\}$ given $\{u(\mathbf{x}_k,t),\ 1 \leq t \leq T,$ and $1 \leq k \leq m\}$ because the standardized random variables can always be transformed back to the raw scale of the measurements. Moreover, knowledge of $\{u(\mathbf{x},t),\ 1 \leq t \leq T\}$, for any $\mathbf{x}$ is equivalent, through the autoregressive relationship, to knowledge of $\{u(\mathbf{x},1), \varepsilon(\mathbf{x},t),\ 2 \leq t \leq T\}$. Recall that the AR shocks are temporally independent so that the conditional distribution for the ozone fields at $\mathbf{x}_0$ can be found based on the much simpler conditional distribution of $\varepsilon(\mathbf{x_0},t)$ given $\{\varepsilon(\mathbf{x_k},t),\ 1 \leq k \leq m\}$. Thus we can easily generate a conditional ozone field by considering the conditional field of the AR(1) shocks and then transform these results to the original scale of measurements. Specifically, we assume the spatial shocks to be multivariate Gaussian so that

$$[\varepsilon(\mathbf{x_0},t)|\{\varepsilon(\mathbf{x_k},t),\ 1 \leq k \leq m\}] \sim \text{Gau}(M, \Sigma_{\varepsilon_0|\varepsilon}), \tag{5}$$

where, if $\underline{c}_0 = \text{cov}(\varepsilon(\mathbf{x}_0,t), \varepsilon(\mathbf{x}_k,t),\ 1 \leq k \leq m)$, $\mathbf{K}$ is the covariance matrix $[\text{cov}(\varepsilon(\mathbf{x}_i,t), \varepsilon(\mathbf{x}_j,t))]$ ($\mathbf{x}_i$, $1 \leq i \leq m$ and $\mathbf{x}_j$, $1 \leq j \leq m$ station locations), and $\underline{\varepsilon}$ the vector of observed spatial shocks; then $M = \underline{c}_0^T \mathbf{K}^{-1} \underline{\varepsilon}$ and $\Sigma_{\varepsilon_0|\varepsilon} = \text{cov}(\varepsilon(\mathbf{x}_0,t), \varepsilon(\mathbf{x}_0,t)) - \underline{c}_0^T \mathbf{K}^{-1} \underline{c}_0$. Note that for predicting to a single location (as described here), $\text{cov}(\varepsilon(\mathbf{x}_0,t), \varepsilon(\mathbf{x}_0,t))$ is a scalar quantity.

Because the number of locations and grid points here is small, it is sufficient to use the Cholesky decomposition method, but for larger domains other faster approximate methods can be implemented (see, for example, Davis [2], Gotway [10], and Fang and Tacher [3] for details on this and other methods).

The algorithm for conditional sampling of the FHDA field is now summarized below.

1. Initialize the time series by interpolating $u(\mathbf{x}_0,1)$ from $u(\mathbf{x}_k,1)$ $(1 \leq k \leq m)$ using a thin plate spline.

2. For $t = 2, \ldots, T$ sample the spatial shocks from (5).

3. Accumulate the sampled shocks and initial value using the autoregressive relationship (2) to obtain a conditional realization of the standardized process $u(\mathbf{x}_0,t)$.

4. Unstandardize the simulated data, and compute the FHDA at $\mathbf{x}_0$.

The shocks at a station location are based on the actual daily observations and so the sample is tied explicitly to the data. If in fact $\mathbf{x_0}$ is a station location and the spatial process has a zero nugget

variance, then the resulting conditional sample will just be the observed data. Thus the "conditional realization of the FHDA field" will be the FHDA statistic for that station's measurements. It should be noted that this algorithm works because we assume complete observations at the station locations. It would be more complicated if observations were sparse over time. For these data there are no instances where there are missing observations at a given time point at every station. Therefore, when shocks are sampled from the conditional distribution, locations that have missing values are simply not used in the calculation for that time point. Although it is possible to sample from $[u(\mathbf{x}_0, 1)|u(\mathbf{x}_k, 1), 1 \leq k \leq m]$ in Step 1 exactly, we have found that sampling from a geostatistical model fit to the standardized fields is adequate.

In this algorithm it is straightforward to replace the conditional sampling of a single location with a vector, or grid of locations. Thus one obtains a conditional field with spatial and temporal dependence among the grid points consistent with the space-time model. In addition, this algorithm can be modified simply to simulate a space-time process that follows this model. In this case, it is necessary to substitute the unconditional sample (7) for the conditional sample (5) of the shocks at Step 2. This unconditional Monte Carlo result is used in the next section to identify an approximate Gaussian model for the FHDA field.

3.4 Seasonal Model

For the seasonal model we posit that the FHDA field is approximately Gaussian distributed, so the main modeling issue is to derive a suitable covariance function. Although extreme value theory suggests that, in the limit, extreme order statistics will not follow a Gaussian distribution; it is not clear that for the ozone application, the fourth highest out of a sample of 184, will achieve the limiting case. In fact, the bivariate FHDA distribution derived from simulations of correlated bivariate normal random variables with unit variances is well approximated by a bivariate normal. This perhaps surprising result is a partial justification of the Gaussian assumption.

In the next section we present results based on a standard geostatistical technique for estimating a stationary covariance from the FHDA values. However, with access to a daily model, one can also compute a covariance model for the FHDA field by Monte Carlo simulation. One generates many realizations of the space-time ozone process and accumulates independent realizations of the FHDA field. The sample covariances are then used to fit a covariance function for the FHDA field. That is, let $\underline{Y}_i = (Y_{i1}, \ldots, Y_{im})'$ be the i^{th} $(i = 1, \ldots, N)$ sampled FHDA values obtained from the algorithm in section 3.3 using the unconditional distribution (7) in Step 2. Then, for each pair of locations (Y_{ij}, Y_{ik}), compute $\hat{C}_Y = \frac{1}{N}\sum_i^N (Y_{ij} - \bar{Y}_i)(Y_{ik} - \bar{Y}_i)$, and finally fit a covariance (ψ_m) to the $\hat{C}_Y$'s as a function of distance. We note that this is a purely computational exercise; and by increasing the Monte Carlo sample size, one can obtain arbitrarily accurate estimates of the second moments

for the FHDA spatial process implied by the daily model.

4 Results

4.1 Daily Model Results

Individual autoregressive models were fit by maximum likelihood to the standardized station data to estimate the parameter of the AR(1) model. An AR(2) model was also tried, but did not show significant improvement to warrant its added complexity. The AR(1) parameter estimate varies across stations, which is to be expected even under stationarity. To assess stationarity of the AR shocks, a local correlogram is fit for each station location using an exponential covariance function. The estimated nugget variance and estimated range parameters do not vary significantly across the domain; the estimated nugget ranges from 0.83 to 1.03 ppb (0.04 ppb) and the estimated range parameter from 164 to 328 miles (33 miles), where standard errors are given in parentheses, and units have been converted into miles for more intuitive interpretations. This suggests that the spatial shocks field can be approximated by a stationary process. The general shape of the empirical correlations suggested fitting a mixture of exponentials function to these correlations.

$$\psi(d(\mathbf{x},\mathbf{x}')) = \alpha \exp(-d(\mathbf{x},\mathbf{x}')/\theta_1) + (1-\alpha)\exp(-d(\mathbf{x},\mathbf{x}')/\theta_2) \tag{6}$$

where θ_1 represents a short range correlation and θ_2 a long range correlation. Correlation model (6) allows the spatial field to be interpreted as the sum of two independent spatial processes with possibly different correlation scales. The smoothness of ψ will still be linear at zero, but the shape will be modified for short distances. The reader should note that unlike a geostatistical analysis for a single field, the correlations associated with the shocks are statistics based on a large ($n > 500$) sample size, which enables enough accuracy to facilitate modeling detailed features such as the mixture component. Fig. 2 (a) shows the fitted correlogram for the spatial shocks as a function of distance of separation along with box plots of the binned sample correlations.

The fitted function is shown as a dashed line in Fig. 2. Within a distance of about 50 miles the shocks show a correlation of around 0.8 or higher, and remain above 0.6 as far out as 100 miles. The fitted parameter estimates are $\hat{\alpha} \approx 0.130$ (0.02), $\hat{\theta}_1 \approx 11$ miles (3.37 miles) and $\hat{\theta}_2 \approx 272$ miles (16.89 miles) with parametric bootstrap standard errors in parentheses.

The fitted model was used to generate conditional fields for the FHDA for each year and the 15×15 rectangular grid in the RTP subregion (Fig. 1). One thousand Monte Carlo realizations were used to approximate the distribution. Results for 1997 (Fig. 3 (a) and (b)) give predicted FHDA for the RTP in 1997 that are mostly above the 84 ppb standard. Most areas in the north and west are found to be more than one model standard error of prediction (MPSE) above 84 ppb, and most

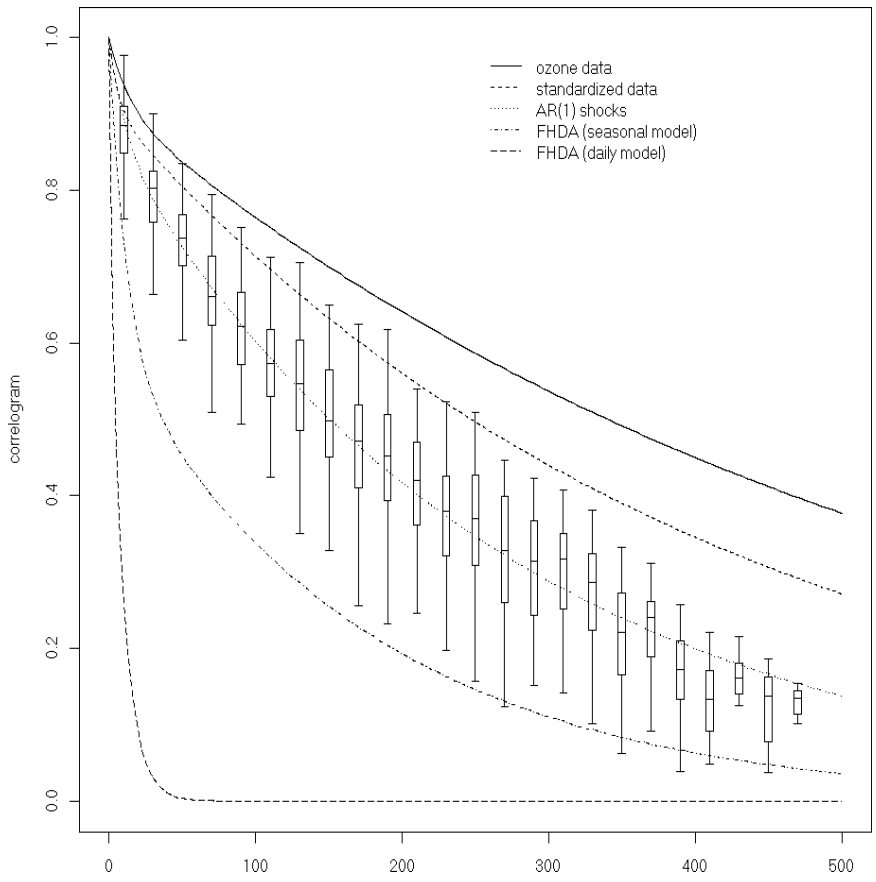

Figure 2: *Fitted empirical correlation functions for original daily maximum 8-hour average ozone measurements, the standardized daily values, the spatial AR(1) shocks, unconditional (seasonal model) and conditional (daily model) simulations of the FHDA field. The box plots are of correlations binned by distance (miles) for the empirical spatial shocks.*

locations found to be in attainment are within one MPSE of being at or above this threshold. It should be noted that a down-side of this approach is that the daily model is not able to account for occassional large ozone values that appear in the data based on a Gaussian assumption at a daily time scale so that it does not accurately capture the variability of the FHDA field.

4.2 Seasonal Model Results

The seasonal approach applies a spatial model directly to the FHDA values, so a key step is to estimate a covariance function for this field. It should be noted that no standardization of the FHDA values is performed, but that a linear spatial trend is included in the model. Empirical variograms for each of the 5 seasons indicate that almost all of the spatial dependence in the FHDA field appears to be limited to a very short range less than 100 miles. A mixture of exponentials

variogram

$$\gamma(d(\mathbf{x},\mathbf{x}')) = \sigma^2(1-\alpha\exp(-d(\mathbf{x},\mathbf{x}')/\theta_1) - (1-\alpha)\exp(-d(\mathbf{x},\mathbf{x}')/\theta_2))$$

was fit using all five years of data with parameter estimates: $\hat{\sigma} \approx 7.37$ ppb, $\hat{\alpha} \approx 0.38$, $\hat{\theta}_1 \approx 0.62$ miles and $\hat{\theta}_2 \approx 48.61$ miles and subsequently converted to a covariance function, ψ_v.

For comparison to estimating the covariance from the FHDA variogram, a covariance function was estimated from unconditional simulations of the daily model. That is, in step 2 of the daily model algorithm, we sample shocks from the *unconditional* distribution

$$\varepsilon(\mathbf{x}_k, t) \sim \text{Gau}(\mathbf{0}, \mathbf{K}),\ 1 \le k \le m \tag{7}$$

($\mathbf{K}$ the same as in (5)) instead of sampling from the conditional distribution (5).

Based on a Monte Carlo sample of 600 FHDA simulated fields, a mixture of exponentials (6) was fit to the empirical correlations, call it ψ_m. The estimated parameters are $\hat{\alpha} \approx 0.51$, $\hat{\theta}_1 \approx 8.66$ and $\hat{\theta}_2 \approx 128.76$. For each choice of covariance model, the seasonal model is fit using the function `Krig` from the R [15] package `fields` (Nychka *et al.* [14]).

The thin plate spline model was fit (using `Tps` from the R [15] package `fields` (Nychka *et al.* [14])) with a linear drift and the smoothing parameter chosen by generalized cross-validation.

The seasonal model, ψ_m, for 1997 shows similar predictive results (Fig. 3 (c) and (d)) to those of the daily model (Fig. 3 (a) and (b)). It appears that there are fewer areas predicted to be out of attainment with this model than with that of the daily model, and the seasonal model generally favors attainment more than the daily model. However, some of the areas found to be in (or out of) attainment are within one MPSE of being out of (or in) attainment. Generally, the MPSE's for both models are very similar, but the seasonal model MPSE is as high as 1 ppb higher than that of daily model in areas, such as the southeast corner of the RTP region, where station locations are not situated nearby.

4.3 Model Comparison

Model-based standard errors can either be reliable or misleading depending on the adequacy of the spatial model. It is, therefore, of interest to use cross-validation to evaluate the average prediction error of these methods. Specifically, let $\hat{Y}_{-i}$ represent the predicted FHDA value at station i, having removed this station's data from the analysis, and Y_i be the observed FHDA for station i. The cross-validation residuals (CV) are given by

$$\text{CV} = \hat{Y}_{-i} - Y_i,\ i = 1, \ldots, m. \tag{8}$$

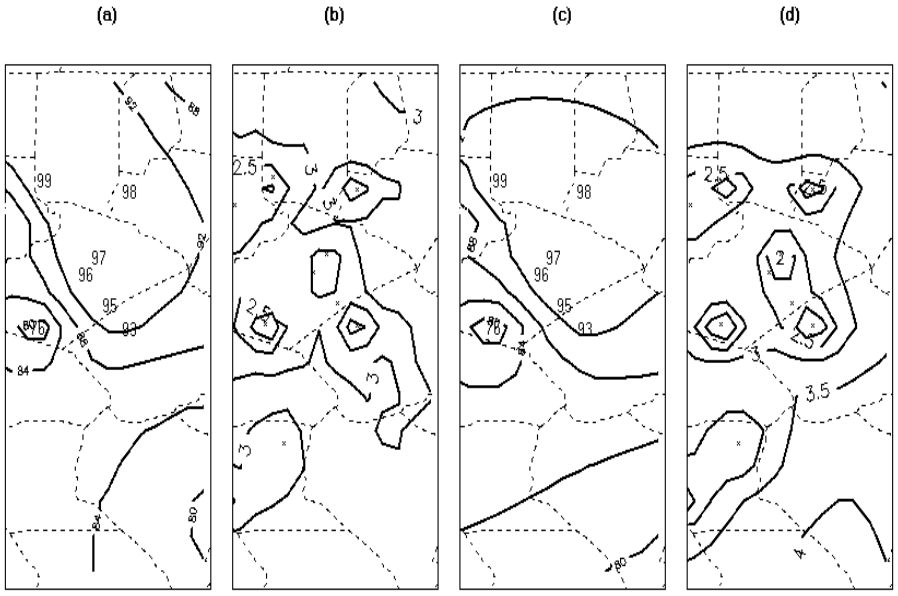

Figure 3: *Daily ((a) and (b)) and seasonal model (ψ_m) ((c) and (d)) prediction results for RTP region in 1997, where (a) and (c) are predicted FHDA; and (b) and (d) are model prediction standard errors (MPSE). Observed FHDA for 1997 are overlayed (larger numbers) on plots (a) and (c) at station locations. Dashed lines are county boundaries.*

The standard leave-one-out procedure was applied to each monitoring location for each method. Table 1 reports the root mean squared error (RMSE) of Eq. (8) for each year. The seasonal model (both ψ_m and ψ_v) CV RMSE and thin plate spline are very similar for each year. The daily model CV RMSE is consistently lower than the other models, but only slightly. Of course, it is desired to have a model that is correct (low CV) and has low MPSE. In addition to slightly better CV values, the daily model has comparable MPSE (Table 2) with the thin plate spline, which are lower than those of the seasonal models. All of this suggests that the daily model out performs the other models compared here; at least enough to warrant its added complexity.

5 Discussion

Although care is needed in generalizing results from a specific data set to other cases, this work has shown a preference to analyze the FHDA standard using a daily model for ozone and then aggregating over the season to infer the FHDA field. The results for the North Carolina study region show that the seasonal model is reasonable, but the daily model is generally more accurate,

Table 1: *Leave-one-out cross-validation (CV) root mean squared error (RMSE) (ppb) for predicting FHDA.*

	Thin Plate Spline	Seasonal Model (ψ_v)	Seasonal Model (ψ_m)	Daily Model
1995	5.34	5.19	5.33	4.73
1996	5.61	5.51	5.68	4.84
1997	6.27	6.03	6.05	4.59
1998	5.00	4.98	4.93	3.25
1999	6.25	6.47	6.30	4.91

Table 2: *Averages of model prediction standard errors (MPSE) (ppb).*

	Thin Plate Spline	Seasonal Model (ψ_v)	Seasonal Model (ψ_m)	Daily Model
1995	2.23	5.68	5.27	2.67
1996	2.49	5.96	5.90	2.85
1997	2.91	6.41	6.02	3.01
1998	2.75	5.35	4.85	2.93
1999	4.34	6.76	6.22	2.94

based on the CV measures of RMSE. Of course, the MPSEs for the daily model are not as accurate as the other approaches because the daily model is not able to account for occassional large ozone values that appear in the data based on a Gaussian assumption at a daily time scale.

Conceptually, the daily model has advantages in using fairly simple statistical components on a daily scale that can produce relatively complicated seasonal statistics. For example, as long as a standard correlation function is reasonable for $\psi(d(\mathbf{x}, \mathbf{x}'))$ from (3), the entire daily model can be fit using standard geostatistical and regression methods even if the observed FHDA field is nonstationary. We believe that part of the success of the daily model is that much of the spatial correlation and the nonstationarity of the raw measurements can be accounted for by standardizing the process and building in a temporal evolution. Additionally, this standardization probably accounts for the spatial trend better than the seasonal approach, and may explain the better performance of the daily model. While the seasonal model is much simpler and easier to employ in general, it can actually be more complicated if the FHDA field is not stationary. The lack of long-range correlation structure in the FHDA field simulated by the daily model approach (conditional on the data) and reaffirmed by empirical variograms of the observed FHDA field suggest that standard spatial techniques may not be very effective at predicting the FHDA at locations relatively far from any monitoring station.

Fig. 2 contrasts the different correlation scales among different transformations of the ozone field, and we note the marked difference between daily fields and the seasonal FHDA. Specifically, at about 100 miles the daily model FHDA fitted correlogram function shows nearly zero correlation, whereas that of the seasonal model shows a correlation above 0.3. This is further justified by the greater MPSE found by both the seasonal model (Fig. 3 (d)) and the thin plate spline (not shown) at locations away from the monitoring network. Additionally, the apparent correlation structure in the FHDA field, found from using the daily model approach without conditioning on the data, may be an artifact of the model. One source of model bias may be the lack of a heavy tail distribution for the AR shocks.

The data used here are from a relatively small, homogeneous subset of a much larger network of monitoring stations, and the more practical application of our work is to extend the analysis to the Eastern United States. Stationarity for this field cannot be assumed (see, for example, Fuentes [6]) and, indeed, some heterogeneity might be modeled using additional covariates. Generally, meteorological data, such as temperature, is difficult to use for the daily model approach because at any time point the temperatures at two locations can vary greatly, and meteorological measurements are generally gathered only on a coarse spatial scale. Therefore, incorporation of such covariates becomes problematic. Two possible covariates, however, that would be easy to incorporate into the daily model are elevation and aspect, and preliminary results have suggested these are useful.

In this work we have considered some parameter uncertainty in parts of the models, but have not propagated the uncertainty into the FHDA fields. A fully Bayesian model could perhaps synthesize covariates, model parameters, and any uncertainty associated with them in an efficient manner. Note that by varying the model parameters in the algorithm, one can include uncertainty into the daily model analysis resulting from uncertainty in the parameters. Although a fully Bayesian approach may be the most elegant solution, bootstrapping is a good compromise in terms of less demands for new software and computing resources. For example, one could use a parameteric bootstrap to generate a sample of parameters that reflect the uncertainty (in a frequentist sense!) with respect to the MLEs. These values would then be used to generate the conditional FHDA fields.

We understand that the use of an extreme order statistic (4^{th} largest) suggests a standard sensitive to the tail of the ozone distribution. For this reason, a productive extension of our models is to incorporate methods from extreme value theory to explicitly model the frequency of large, but rare, ozone events. Part of the benefit of this approach is a statistical description for the entire tail of the distribution, rather than just a particular order statistic or quantile. The biggest challenge, however, would be to incorporate spatial correlations into such a model; an active area of research. This might be accomplished by fitting Generalized Pareto (GP) distributions at each location with values from other stations used as covariates in the scale parameter. Another idea would be to fit

GP distributions without any spatial covariates and introduce correlations through a penalty based on the scale parameter. Some recent work on characterizing extremal dependencies can be found in Schlather and Tawn [16], Ferro and Segers [4] and the references therein.

In closing, although this problem suggests ample areas of new research, we also believe the daily model provides a substantial improvement in interpolating monitoring data with respect to the regulatory standard. Moreover, our methods are easily implemented with supporting packages in the R [15] environment, and so can be used by a broad group of scientists beyond statisticians.

6 Acknowledgements

This research was supported by National Science Foundation grant DMS 9815344. The authors thank William Cox for making the ozone data sets available, and the reviewers for detailed and helpful comments.

References

[1] Cressie, Noel A.C. *Statistics for Spatial Data (Revised Edition).* Wiley Interscience, New York, 1993.

[2] Davis, Michael W. Production of conditional simulations via the LU triangular decomposition of the covariance matrix. *Mathematical Geology*, 19 (2): 91–98, 1987.

[3] Fang, Jiannong and Tacher, Laurent An efficient and accurate algorithm for generating spatially-correlated random fields. *Communications in Numerical Methods in Engineering*, 19: 801–808, 2003.

[4] Ferro, C.A.T. and Segers, J. Inference for clusters of extreme values. *J.R. Statist. Soc. B*, 65 (2): 545–556, 2003.

[5] Freeden, W., Gervens, T. and Schreiner, M. Constructive approximation on the sphere with applications to geomathematics. *Clarendon Press*, 1998.

[6] Fuentes, Montserrat Statistical assessment of geographic areas of compliance with air quality. *Journal of Geographic Research–Atmosphere*, 108(D24), 2003.

[7] Gaspari, G. and Cohn, S.E. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125:723–757, 1999.

[8] Gneiting, Tilmann Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97 (458):590–600, 2002.

[9] Gneiting, Tilmann Correlation functions for atmospheric data analysis. *Quarterly Journal of the Royal Meteorological Society*, 125:2449–2464, 1999.

[10] Gotway, Carol A. The use of conditional simulation in nuclear-waste-site performance assessment. *Technometrics*, 36 (2): 129–141, 1994.

[11] Green, P.J. and Silverman, B.W. *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, 2-6 Boundary Row, London SE1 8HN, UK, 1994.

[12] Hastie, T.J. and Tibshirani, R.J. *Generalized Additive Models.* Chapman and Hall, London; New York, 1990.

[13] Holland, David M., Cox, William M., Scheffe, Rich, Cimorelli, Alan J., Nychka, Douglas and Hopke, Philip K. Spatial prediction of air quality data. *Environmental Manager*, Aug. 2003.

[14] Nychka, D., Meiring, W., Royle, J.A., Fuentes, M. and Gilleland, E. Fields: R Tools for Spatial Data. `http://www.cgd.ucar.edu/stats/software`, 2002.

[15] R Development Core Team R: A language and environment for statistical computing. *R Foundation for Statistical Computing* Vienna, Austria, ISBN 3-900051-00-3. `http://www.R-project.org`, 2003.

[16] Schlather, Martin and Tawn, Jonathan A. A dependence measure for multivariate and spatial extreme values: properties and inference. *Biometrika*, 90: 139–156, 2003.

[17] Stein, Michael L. *Interpolation of spatial data: some theory for kriging.* Springer-Verlag, 175 Fifth Ave., New York, N.Y. 10010, 1999.

[18] Weber, R.O. and Talkner, P. Some remarks on spatial correlation function models. *Mon. Weather Rev.* 121:2611–2617, 1993.