Paper: Statistical models for monitoring and regulating ground-level ozone

Eric Gilleland and Douglas Nychka, Environmetrics, 2005

1 Background

As we know, ground-level ozone triggers a variety of health problems even at very low levels and may cause permanent lung damage after long-term exposure. It is one of six common pollutants for which the Clean Air Act requires EPA to set National Ambient Air Quality Standards (NAAQS) for ozone. EPA has been revising the standard, and in 2004, EPA revised NAAQS to be based on the fourth-highest daily maximum 8-h average ozone level (FHDA). To attain the EPA primary standard, the 3-year average of FHDA concentrations measured at each monitor within an area must not exceed 0.08 ppm (parts per billion).

There are 513 monitoring stations covering eastern Unites States, but still region without monitor does not have any information about the amount of ozone exposure on the population or environment of this region, because the use of monitoring data in a regulatory context is often limited to point locations. And that brings out the necessity of constructing some statistical models to interpolate from a network of monitoring data. Standard geostatistical model is a widely used spatial model, and it is referred to as the *Seasonal Model* in this paper. However, understanding the spatial distribution for the FHDA for a given ozone season presents a new statistical problem for inferring regions of attainment or non-attainment 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. From this perspective, the authors 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, a geostatistical model (*Seasonal Model*) that predicts the FHDA field from the network values using standard best linear unbiased prediction, or kriging. A third approach that will be used as a benchmark estimates the FHDA field by way of a thin plate splines.

2 Ozone Monitoring Data

Data used for these analysis consist of maximum daily 8-h average ozone levels measured in parts per billion (ppb) for the 72 monitoring stations in a study region centered on North Carolina (Figure 1). These data are a subset of the 513 stations covering the eastern United States used by Fuentes (2003) and can be obtained from the web through $http://www.cgd.ucar.edu/stats/Data$.

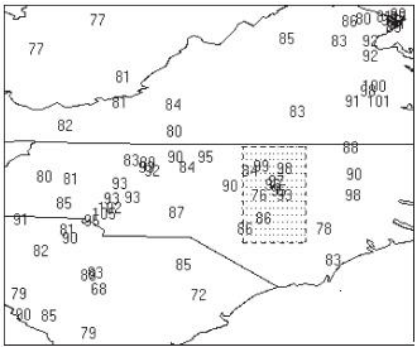

Figure 1: Network of ozone monitoring locations with a rectangular region (with $15 * 15$ grid) in the Research Triangle Park (RTP) in North Carolina. The numbers represent the fourth-highest value for 1997.

3 Daily Model for Ozone

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, the authors suggest a model for daily ozone measurements and then infer the distribution of the FHDA through aggregating the daily model over the season.

3.1 Standardizing the data

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

$$\boldsymbol{O}(\mathbf{x}, t) = \mu(\mathbf{x}, t) + \sigma(\mathbf{x})u(\mathbf{x}, t) \tag{3.1}$$

Assume $u(\mathbf{x}, t)$ for any given location and time has mean zero and variance 1; $\sigma(\mathbf{x})$ is estimated from the residuals, and is allowed to vary over space. Note $\mu(\mathbf{x}, t)$ is a function of both seasonality and space.

3.2 Daily model

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

$$u(\mathbf{x}, t) = \rho(\mathbf{x})u(\mathbf{x}, t-1) + \varepsilon(\mathbf{x}, t) \tag{3.2}$$

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

$$\sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x'})}\psi(d(\mathbf{x}, \mathbf{x'})) \tag{3.3}$$

Here, $d(\mathbf{x}, \mathbf{x'})$ is a metric to measure separation between locations, and $\psi(d(\mathbf{x}, \mathbf{x'}))$ is considered to be an isotropic and stationary correlation function. Equation (3.2) and (3.3) implies a space-time covariance fuction

$$Cov(u(\mathbf{x}, t), u(\mathbf{x'}, t-\tau)) = \frac{\rho(\mathbf{x})^\tau\sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x'})}\psi(d(\mathbf{x}, \mathbf{x'}))}{1-\rho(\mathbf{x}\rho(\mathbf{x'}))}, \tau = 0, 1, 2, ... \tag{3.4}$$

3.3 Sampling the distribution of FHDA conditional by the monitoring data

Let $\mathbf{x}_0$ be a location where ozone is unobserved, and $\{\mathbf{x}_k, 1 \le k \le m\}$ be the station locations. Under the assumption that all the components of the data model are known, we can assume the spatial shocks to be multivariate Gaussian.

$$[\varepsilon(\mathbf{x}_0, t)|\{\varepsilon(\mathbf{x}_k, t), 1 \le k \le m\}] \sim Gau(M, \Sigma_{\varepsilon_0|\varepsilon}) \tag{3.5}$$

Because the number of locations and grid points here is small, it is sufficient to use the Cholesky decomposition method. And so $\underline{c}_0$=$Cov(\varepsilon(\mathbf{x}_0, t), \varepsilon(\mathbf{x}_k, t), 1 \le k \le m)$, $K = Cov(\varepsilon(\mathbf{x}_i, t), \varepsilon(\mathbf{x}_j, t))$ for $1 \le i \le m, 1 \le j \le m$ station locations, and $\underline{\varepsilon}$ the vector of observed spatial shocks; then $M = \underline{c}_0^T K^{-1}\underline{\varepsilon}$ and $\Sigma_{\varepsilon_0|\varepsilon} = Cov(\varepsilon(\mathbf{x}_0, t), \varepsilon(\mathbf{x}_0, t)) - \underline{c}_0^T K^{-1}\underline{c}_0$. Note that for predicting to a single location (as described here), $Cov(\varepsilon(\mathbf{x}_0, t), \varepsilon(\mathbf{x}_0, t))$ is a scaler quantity.

The algorithm for conditional sampling of the FHDA field is summarized:

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

(*ii*) For $t = 2, ..., T$, sample the spatial shocks from (3.5).

(*iii*) Accumulate the sampled shocks and initial value using the autoregressive relationship (3.2) to obtain a conditional realization of the standardized process $u(\mathbf{x}_0, t)$.

(*iv*) Unstandardize the simulated data, and compute the FHDA at $\mathbf{x}_0$.

Repeating the steps, one can generate a random sample that approximated the FHDA conditional distribution.

4 Model Comparison

The seasonal model for 1997 shows similar predictive results (Figures (c),(d)) to those of the Daily Model (Figure (a),(b)).

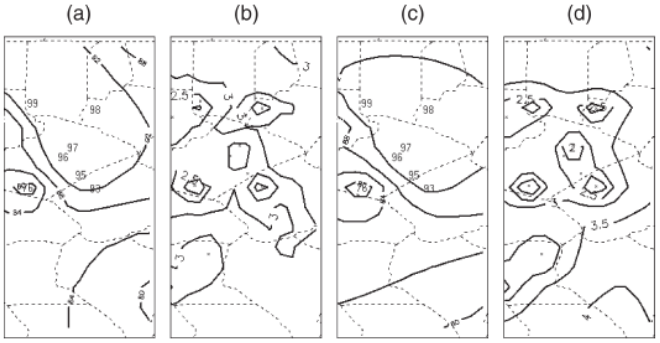

Figure 2: (a) and (c) are predicted FHDA; (b) and (d) are model prediction standard errors(MPSE)

The standard leave-one-out cross-validation (CV) procedure was applied to each monitoring location for each method. Table 1 reports the root mean squared error(RMSE) of the CV for each year. The seasonal model CV RMSE and thin plate spline are very similar, while the daily model CV RMSE is consistently lower. In addition to slightly better CV values, the daily model has comparable MPSEs, in Table 2, with the thin plate spline, which are lower than those of the seasonal model.

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

All of this suggests that the daily model outperforms the other models compared here.