Spatio-Temporal Modeling of $PM_{2.5}$ Data with Missing Values

Richard L. Smith, Stanislav Kolenikov and Lawrence H. Cox

University of North Carolina and

National Center for Health Statistics *

March 19, 2003

Abstract

We propose a method of analyzing spatio-temporal data by decomposition into deterministic nonparametric functions of time and space, linear functions of other covariates, and a random component that is spatially though not temporally correlated. The resulting model is used for spatial interpolation, and especially for estimation of a spatially-dependent temporal average. The results are applied to part of the $PM_{2.5}$ network established by the United States Environmental Protection Agency (USEPA), covering three southeastern U.S. states. A novel feature of

*Richard L. Smith (rls@email.unc.edu) is professor and Stanislav Kolenikov (skolenik@email.unc.edu) is a graduate student at the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. Lawrence H. Cox is Associate Director, National Center for Health Statistics, Centers for Disease Control and Prevention, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782. He was formerly Senior Mathematical Statistician, U.S. Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, NC 27711. This work was supported by EPA Cooperative Agreement CR-827737-01-0. Richard L. Smith is also supported by NSF grants DMS-9971980 and DMS-0084375. Part of the paper was written while R.L.S. was visiting the Geophysical Statistics Project at the National Center for Atmospheric Research; this Project is supported by NSF grant DMS-9815344. The authors thank Dr. David Holland of the EPA for providing the data used in the study.

the analysis is a variant of the expectation-maximization (EM) algorithm to account for missing data. The results show, amongst other things, that a substantial part of the region is in violation of the proposed long-term average standard for $PM_{2.5}$.

1 Introduction

The classical theory of spatial statistics (or geostatistics), represented by numerous books such as those of Cressie (1993), Chilès and Defliner (1999) and Stein (1999), is concerned primarily with the case of a single realization from a spatially correlated stochastic process. The more recent theory of spatio-temporal processes, represented by the present volume, extends this to processes correlated in both time and space. In between these extremes lie space-time processes whose random component is spatially but not temporally correlated, so that there are independent replications of a spatial process. We believe that processes of this type are appropriate for many forms of geophysical data, but they also pose methodological problems in their own right. One of these, that is one of the principal foci of the present paper, is the treatment of missing data.

The application that has motivated this paper is the analysis of $PM_{2.5}$ data (particulate matter of aerodynamic diameter 2.5 μm or less) from a network set up by the United States Environmental Protection Agency (USEPA). Under the Clean Air Act and Amendments, USEPA is responsible for maintaining standards on air pollutants that are "requisite to protect human health" with "an adequate margin of safety". Among the currently regulated pollutants is PM_{10} (particulate matter of aerodynamic diameter 10 μm or less). Under a new standard, proposed in 1997 but not yet implemented, this is extended to include $PM_{2.5}$. Specifically, the proposed standard requires that (a) the three-year average of the 98th percentile of $PM_{2.5}$ should not exceed 50 $\mu g/m^3$, (b) the arithmetic mean (over all monitors within a given region) of the three-year average of daily $PM_{2.5}$

levels should not exceed 15 $\mu g/m^3$. Further information about the background of these standards and the epidemiological basis behind them is given by Cox (2000).

In 1999, USEPA established a network of some 800 $PM_{2.5}$ monitors, to supplement the much longer-established PM_{10} monitoring network. This new network has been used to gather information about the spatial distribution of $PM_{2.5}$, which is needed to help design a long-term $PM_{2.5}$ network and can also be used to answer other research questions, such as whether spatial interpolation could help in determining individual exposure to air pollution (National Academy of Sciences, 1998). Satisfactory answers to this and related questions, however, require methodological research on spatio-temporal analysis of $PM_{2.5}$ data.

As a first step to answering these questions, this paper presents a spatio-temporal analysis of part of the 1999 data, from the three states of North Carolina, South Carolina and Georgia, within which there were 74 monitors from which we calculated weekly $PM_{2.5}$ means. Preliminary statistical analysis suggested that the $PM_{2.5}$ field could be represented as the sum of nonparametric spatial and temporal trends, together with a random component that is spatially but not temporally correlated (in other words, the one-week time aggregation interval is long enough for successive obsevations to be uncorrelated). By using geostatistical methods to interpolate the random component, we are able to estimate the weekly average $PM_{2.5}$ at any point in the region, and hence to estimate derived quantities such as the long-term average at any site. As with any statistical interpolation procedure, a major question of interest is the uncertainty of the estimation procedure.

A particular methodological issue raised by our analysis is how to deal with the rather high proportion of missing data (around 28%) in the context of a maximum-likelihood fitting of a spatial or spatio-temporal model. Two methods are outlined to deal with this. For a pure spatial model without any temporal dependence, it is possible to calculate an exact likelihood function

by computing and inverting the spatial covariance matrix for each week's data — typically, the spatial covariance matrix is different for each week because the available monitoring network is different for each week. This is feasible but computationally inefficient, and even conceptually may not work for a truly spatio-temporal model that includes temporal as well as spatial correlations. The alternative method, which is in principle applicable to any spatio-temporal model for the data, uses the expectation-maximization (EM) algorithm to account for the conditional distributions of missing observations. We give particular attention to this second method, and variants known as the GEM and ECM algorithm, and show how it may be used to calculate approximate maximum likelihood estimators for the spatial model under consideration.

The structure of the paper is as follows. Section 2 describes the data and poses the research questions. Section 3 presents the semiparametric model that accounts for trends in space and time, as well as for the residual spatial covariance. Section 4 describes the principle of the EM algorithm, and shows how it can be applied in our setting. Section 5 presents the estimation results, and Section 6 concludes.

2 Data Used in This Study

The data used in this research is a part of the EPA data set for $PM_{2.5}$ collected for 49 weeks during 1999. The observation frequencies generally vary from site to site: most sites have observations recorded once in three days, but some have daily records, and others have much sparser records. Information about the monitors includes geographic position (latitude and longitude); urbanization classified as rural, urban or suburban; land use classified as agricultural, industrial, commercial, residential or forest; altitude of the monitor; the measurement method; and some other technical

information.

We only used a fraction of this rich data set related to North Carolina, South Carolina, and Georgia. There were 74 monitors across those states (23 in Georgia, 35 in North Carolina, 16 in South Carolina), mapped in Fig. 1. No data are available for Georgia in the fourth quarter of the year. The data were further aggregated into weekly averages: for each week and each monitoring station, a suitably weighted average is computed based on all readings available during that week. There is a possibility of some bias by this method, because $PM_{2.5}$ values are typically lower at weekends than weekdays, but we ignore that aspect here.

We ended up with 2613 observations. The proportion of missing data is rather high: comparing the above figure with $74 \times 49 = 3626$ observations that should be in the complete data set, almost 28% of the data are missing.

3 Building a Spatio-temporal Model

This section presents initial analysis of the data described in Section 2, leading up to the detailed specification of the model (5). Then, Section 4 describes the detailed approach to fitting that model.

3.1 Transforming the raw data

Initial inspection of the data shows that both the mean and variance of the $PM_{2.5}$ data tend to be higher in Georgia than the other two states. It would be desirable to find a variance-stabilizing transformation, i.e one that makes the variance approximately constant across all stations. Two possibilities are (a) a square root transformation, (b) a logarithmic transformation. It would be possible to consider more general families of transformations, such as the Box-Cox transformation

$y \to (y^\lambda - 1)/\lambda$, but we shall confine ourselves here to the square root and logarithmic transforms.

Fig. 2 plots the variance for each station against the mean for each station, using (a) the original $PM_{2.5}$ data, (b) square root of $PM_{2.5}$, (c) natural logarithm of $PM_{2.5}$. It is obvious that plot (a) shows an increase of variance with the mean. The other two plots both show approximate constancy of variances with the exception of two stations which have much larger variances than the remainder. These two stations are more prominent outliers in plot (c) than plot (b), and this gives some reason to prefer plot (b), i.e. the square root transform.

As noted originally by Box and Cox (1964), if different data transformations are to be compared in terms of standard statistical criteria such as residual sums of squares, it is necessary first to rescale the data. In the case of the square root and logarithmic transformations, this means replacing y_i with either $C_1\sqrt{y_i}$ or $C_2 \log y_i$, where $C_1 = 2\sqrt{\dot{y}}$ or $C_2 = \dot{y}$. Here $\dot{y} = \left(\prod_{i=1}^n y_i\right)^{1/n}$ is the geometric mean of the observations $y_1, ..., y_n$. These transformations will be used in subsequent comparisons.

3.2 The time trend

Since we have only one year of observations, it is not possible to do a detailed decomposition into seasonal and long-term trends; all the variation must be assumed seasonal in origin. It might be possible to "explain" some of the seasonal variation by including meteorological information at individual sites, as has been done for example in a similar context by Holland *et al.* (2000). In another paper looking at PM_{10} in the Pittsburgh area, Daniels *et al.* (2001) argued that if meteorological effects are properly accounted for, there is then no need to model any spatial dependence, but it would be very surprising if such a conclusion were to hold in general for air pollution data sets. In the present analysis, no attempt has been made to incorporate meteorological effects, but we consider two simple nonparametric approaches:

1. Model each weekly mean as a separate "week effect" as in standard analysis of variance.

2. Use a smooth function to represent the weekly trend over the whole year.

The latter approach is implemented using B-splines (Green and Silverman 1994), in which an unknown smooth function is approximated as the weighted sum of the basis functions:

$$B(x) = \begin{cases} \frac{3|x|^3 - 6x^2 + 4}{6}, & -1 \le x \le 1, \\ \frac{(2-|x|)^3}{6}, & 1 < |x| \le 2, \\ 0, & 2 < |x|. \end{cases} \tag{1}$$

$$\hat{f}(t) = \alpha_0 + \sum_{k=1}^{K} \alpha_k \delta_k(t), \quad t \in [0, T], \quad \delta_k(t) = B\left[\frac{K}{T}\left(t - \frac{Tk}{K}\right)\right], \tag{2}$$

where t is the time index and T the total observation time. Coefficients $\alpha_0, \ldots, \alpha_K$ can be estimated by ordinary least squares (OLS) or generalized least squares (GLS) regression, the latter being appropriate when the data points are correlated. The number of basis functions K controls the smoothness of the fitted function: the smaller K, the smoother the function, but at the cost of less precise agreement between the data and the fitted function.

Fig. 3 shows the time trend fitted by simple weekly effects and by B-splines with $K = 20$. Other values of K were tried with similar results. Plot (a) shows the result of a single time trend fitted to all the data points, with the actual data points (all stations) overlaid on the plots. Plots (b)–(i) show the same estimated trends but with different subsets of the data, corresponding to individual states (plots (b)–(d)) or different land uses (plots (e)–(i)). The results tend to confirm that there is a similar shape of trend across all subsets of the data, but with upward or downward shifts, e.g. the values for Georgia seem clearly above the underlying trend while those for forest areas are clearly below. Incidentally, plot (d) for Georgia confirms that the overall higher means in Georgia

are not merely the result of Georgia data being unavailable during the later part of the year, but the Georgia values are generally higher than those elsewhere.

Based on these plots, we provisionally conclude that a common time trend (represented by week effects or by B-splines) may be applied to all the stations, but is shifted up or down at each station by a constant that depends on the location and land use of the station.

3.3 The space trend

The trend in space was also estimated non-parametrically via the bivariate version of splines, known as *thin-plate splines* (Green and Silverman 1994). The basis function for this spline evaluated at the point (x, y) is given by

$$\Psi(x, y) = r^2 \log r \tag{3}$$

where $r = \sqrt{x^2 + y^2}$ is the distance from the origin (i.e., the knot of the spline). The overall spatial trend is represented as

$$\psi_{x,y} = \beta_0 + \beta_1 x + \beta_2 y + \sum_{j=1}^{J} \beta_{j+2} \Psi(x - x^{(j)}, y - y^{(j)}), \tag{4}$$

$(x^{(j)}, y^{(j)})$ denoting the coordinates of the jth knot. Green and Silverman (1994) take J equal to the total number of sites (i.e., one knot per monitor) but with an additional "bending energy" penalty term to force some smoothness into the fitted function. A simpler alternative, used here, is to use J directly as a smoothing parameter, i.e. we force $\psi_{x,y}$ to be smoother by restricting the number of knots. For a given value of J, we select the knots by K-means clustering (Hartigan and Wong (1979)): group the 74 monitoring locations into J clusters and then take the cluster centers as the knots of the spline. In our comparisons we have taken $J = 10,\ 20,\ 30,\ 40$ and 50.

Another part of the model that can be thought of as a component of the spatial trend are

the additive terms that account for differences in the landscape surrounding the observation site. Those are coded as the set of four dummy variables for the five possible land uses. The analysis of Section 3.2 suggests that we could also include a dummy variable to denote the state in which the site lies, but this was not used in the final analysis because it would imply, counterintuitively, that there should be a jump shift in the $PM_{2.5}$ level at a state boundary. Instead, we rely on the thin-plate spline representation to include any spatial variation on a statewide scale.

3.4 Comparing regression models

The previous subsections suggest a variety of different regression analyses, in which we may take any of $PM_{2.5}$, $\sqrt{PM_{2.5}}$ or $\log PM_{2.5}$ as the independent variable, time trends modeled by week effects or by B-splines with various numbers of knots K, the spatial trends similarly modeled by thin-plate splines with knots, and other effects such as land use. The simplest way to compare these different models is to fit an ordinary least squares (OLS) regression, ignoring the possible effects of spatial and/or temporal correlation. To compare different models, we have used the AIC and BIC criteria, defined by $AIC = n \log s^2 + 2p$, $BIC = n \log s^2 + p \log n$, where n is the number of data points, p the number of regressors and s^2 the mean square of the residuals. Although these are only two out of many possible variable selection criteria, they are particularly useful in the present context given their simplicity of computation when comparing large numbers of models.

We do not give detailed tables of regression analyses but instead summarize our broad conclusions as follows:

1. Of the three transformations of $PM_{2.5}$ (none, square root or logarithmic), the square root transformation was best in all cases where they were directly compared.

2. The weekly trend is best modeled by a simple week effect, which gives a better fit to the data than any of the B-spline regression models.

3. Comparing spatial trends modeled by thin-plate splines with $K = 10$, 20, 30, 40, 50 knots, AIC chooses $K = 40$ while BIC (which always favors a model with a smaller number of parameters) selects $K = 10$. For our subsequent analyses we have compromised between these two conclusions and chosen $K = 20$.

4. Modeling land use (A, C, F, I, R) as a dummy variable always gave a significant result.

5. There was no evidence of any interaction between land use and week.

Based on these conclusions, the subsequent analysis uses the model

$$y_{st} = \omega_t + \psi_s + \theta_s + \eta_{st}, \tag{5}$$

where y_{st} is the square root of the mean $PM_{2.5}$ at location s in week t; ω_t denotes the week effect; ψ_s is the spatial trend at location s (writing $s = (x, y)$, this is given by (4), a thin-plate spline with 20 knots); θ_s is the effect due to land-use at location s; and η_{st} is a random error.

The next step is to look for time and spatial correlations among the η_{st} variables, represented by the residuals when the model (5) is fitted by OLS.

3.5 Temporal autocorrelations

For each of the 74 stations, the residuals η_{st} were viewed as a time series in t, and the first five autocorrelations computed. A standard time series technique for the statistical significance of sample autocorrelations is to compare them with $2/\sqrt{T}$, where T is the length of the time series. This corresponds approximately to a 5% test of significance. In the present case, the value of T is

different for the different stations, but we have taken the average value in order to make a direct comparison across all the stations. Fig. 4 shows all 74 autocorrelation plots superimposed, with the critical $\pm 2/\sqrt{T}$ values shown as horizontal lines. Very few of the individual autocorrelations exceed the critical values. From this we conclude that there is no temporal autocorrelation among the residuals, and it is safe to proceed with a purely spatial analysis.

3.6 Variograms

For spatial data sets, a common method of looking for spatial correlation is through the variogram. Under assumptions of stationarity and isotropy, the variogram is defined by

$$2\gamma(h) = \mathrm{E}\{(\eta_{s_1 t} - \eta_{s_2 t})^2\}$$

where h is the distance between two spatial locations s_1 and s_2. It is typically calculated by grouping the possible values of h into bins, and computing one value by taking the sample average of all $(\eta_{s_1 t} - \eta_{s_2 t})^2$ values for which $|s_1 - s_2|$ lies within a given bin. Although there are many variants on this basic algorithm (Cressie (1993)) we stick to this procedure here.

In measuring the distance between two monitoring stations with latitudes θ_1 and θ_2 and longitudes ϕ_1 and ϕ_2 (converted to radians), we use the formula

$$\text{Distance} = 12732.40 \arcsin(B) \quad \text{(km)}$$

where

$$4B^2 = (\cos\theta_1 \cos\phi_1 - \cos\theta_2 \cos\phi_2)^2 + (\cos\theta_1 \sin\phi_1 - \cos\theta_2 \sin\phi_2)^2 + (\sin\theta_1 - \sin\theta_2)^2.$$

This is the geodesic distance between two locations, treating the earth's surface as that of a sphere. For plotting variograms, distances were grouped into bins of width 25 km.

Fig. 5(a) shows eight variograms computed from regression residuals η_{st}: an overall variogram in which all the data are combined, and separate variograms for each of the three states and each of four "seasons" defined by weeks 0–11, 12–23, 24–35 and 36–49. Some features apparent from visual inspection of these plots include:

- there appear to be significant differences between states and between seasons, with the variogram for Georgia in particular standing out as sitting above the other variograms (i.e. inter-site variances are larger in Georgia than the other two states);
- the variograms do not appear to be of the traditional "nugget–range–sill" form (Cressie 1993) — there is indeed evidence of a nugget effect (i.e. a non-zero limit in the variogram as the distance between two stations tends to 0), but there is no evidence that the variogram levels off to a finite "sill" at any particular range;
- another option is to standardize the data prior to calculating the variogram, by normalizing the residuals at each site so that the sample standard deviation is 1. This is done for the variograms in Fig. 5(b). Although there is some indication that this helps (e.g. the variogram for Georgia no longer stands out as different from all the others), the general characteristics remain the same, i.e. there still appear to be significant differences among the eight variograms plotted, and they do not show a clear-cut sill and range.

For the present analysis, although there is evidence that the data exhibit non-constant variances and possibly other forms of nonstationarity, we have chosen to ignore these features, largely for reasons of computational simplicity. However, for future work, and especially when trying to extend the present three-states analysis to cover the whole U.S., we believe that it will be important to take account of these features.

A separate issue is the functional form of variogram that we should fit. Although many parametric variogram functions are known (see e.g. Cressie (1993) or Chilès and Delfiner (1999)), most of them have a finite sill and correspond to second-order stationary spatial processes. A weaker form of stationarity is *intrinsic stationarity*, which allows the possibility of infinite sill. For example, one form of intrinsically stationary spatial process that does not reduce to a second-order stationary process is the model

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0, \\ \alpha(\theta_1 + h^{\theta_2}) & \text{if } h > 0, \end{cases} \tag{6}$$

where α is an overall scaling constant, $\theta_1 > 0$ implies a nugget effect, and $0 < \theta_2 < 2$. Although other parametric forms of variogram were tried for comparison, (6) is used as our preferred model for the remainder of the present paper.

4 Maximum likelihood estimation and the EM algorithm

4.1 Principles of the EM and related algorithms

If we make the additional assumption that the data are multivariate normal, then the most natural method to fit the spatial model is the method of maximum likelihood. Because of missing data, the number of observations available in each one-week time period is different. We may write down the likelihood function for each week as

$$l(\beta, \theta | y_t) = (2\pi)^{-n_t/2} |\Sigma_t(\theta)|^{-1/2} \exp\big[-\frac{1}{2}(y_t - x_t\beta)^T \Sigma_t(\theta)^{-1}(y_t - x_t\beta)\big] \tag{7}$$

where the subindex t denotes time (week), n_t is the number of observations available in week t, y_t is the vector of measured $PM_{2.5}$ concentrations in week t, x_t is the matrix of explanatory variables

in week t and β_t is the vector of trend coefficients in week t. In principle, the method is to calculate $l(\beta, \theta | y_t)$ for each week t, multiply over all t to obtain an overall likelihood function, and maximize with respect to β and θ. There are, however, two difficulties with this plan.

The first is that for the model (6), there does not exist a stationary covariance matrix Σ_t. The solution is to use *generalized covariances*, see e.g. Cressie (1993), Section 5.4 or Stein (1999), Section 2.9, who both cite Matheron (1973) as the originator of the concept. The semivariogram (6) defines an *intrinsic random function of order 0* (IRF0) and for this we can write

$$Cov\left\{\sum_s \nu_s \eta_{s,t}, \sum_{s'} \kappa_{s'} \eta_{s',t}\right\} = \sum_s \sum_{s'} \nu_s \kappa_{s'} G(||s - s'||), \tag{8}$$

provided $\sum_s \nu_s = \sum_{s'} \kappa_{s'} = 0$. Here G is known as the generalized covariance function: however in the case of an IRF0, it suffices to take $G = -\gamma$. There is a more general concept of an intrinsic random function of order k (IRFk for $k \geq 0$), but the case $k = 0$ suffices for our present application.

To apply this in the context of (7), let $\bar{y}_{\cdot t}$ denote the sample mean of y_{st} over all observed sites s in week t, and let y_t^* denote the space-centered vector in which each y_{st} is replaced by $y_{st}^* = y_{st} - \bar{y}_{\cdot t}$. Similarly, let x_t^* denote the space-centered matrix of covariates and $\eta_{st}^* = \eta_{st} - \bar{\eta}_{\cdot t}$ the space-centered vector of errors. Elementary calculations based on (8) (with $G = -\gamma$) show that

$$Cov\{\eta_{t,s}^*, \eta_{t,s'}^*\} = \frac{1}{n_t}\sum_{s_1} \gamma(||s - s_1||) + \frac{1}{n_t}\sum_{s_1} \gamma(||s' - s_1||) - \gamma(||s - s'||) - \frac{1}{n_t^2}\sum_{s_1}\sum_{s_2} \gamma(||s_1 - s_2||), \tag{9}$$

the sums with respect to s_1 and s_2 being taken with respect to all stations available at time t.

In (7), we may therefore replace y_t by y_t^*, x_t by x_t^*, Σ_t by the covariance matrix Σ_t^* with entries (9), and the maximum likelihood method works. This is essentially the algorithm of Kitanidis (1983).

The second issue about equation (7) is that as things stand, the method is time consuming, because of the need to compute the inverse and determinant of a different Σ_t^* matrix for each

week of the data. The problem would be much easier if there were no missing data, because then Σ_t^* would be the same for every week t, and the determinant and inverse would only have to be calculated once on each evaluation of the likelihood function. The expectation-maximization or EM algorithm (Dempster, Laird and Rubin (1977), Little and Rubin (1987), McLachlan and Krishnan (1997)) provides a way out of this difficulty.

The EM algorithm is an iterative algorithm to obtain maximum likelihood estimates for a parametric model with missing data, for the case when the data are missing at random, i.e. the probability that a given variable is not observed in a given instance is independent of the true value of that variable. This assumption would not be appropriate if, for example, the machinery was liable to break down at very high or very low levels of $PM_{2.5}$, but there is no reason to think this is the reason for missing values in the current data set.

The algorithm alternates expectation (E) and maximization (M) steps. At iteration k of the expectation step, the conditional expected value of the log likelihood given the observed data y_{obs} and the current value of the parameter vector $\theta^{(k)}$ by using the underlying parametric model (notation follows Little and Rubin (1987)):

$$Q(\theta|\theta^{(k)}, y_{obs}) = \int l(\theta^{(k)}|y) f(y_{miss}|y_{obs}, \theta^{(k)})\, dy_{miss}. \tag{10}$$

One way to compute this is to impute the expected values of the missing data, and any functions of the missing data, conditional on y_{obs} and $\theta^{(k)}$. If there is a sufficient statistic for the model, then it is enough to compute the expected value of this statistic conditional on the observed values of the variables involved, and on the current parameter values.

At the maximization step, the full likelihood is maximized with respect to the parameters by

using the imputed missing values or the expected values of the sufficient statistic:

$$\theta^{(k+1)} = \arg\max Q(\theta|\theta^{(k)}, y_{obs}) \tag{11}$$

The procedure is iterated to convergence, or in practice until successive values of the parameter vector $\theta^{(k)}$ differ by no more than some specified small tolerance. As shown in the cited references, under mild regularity conditions the EM algorithm converges to the maximum likelihood estimator.

Two variants of the EM algorithm may speed up convergence. The first is that in the M step, the conditional likelihood is not maximized but only made to be larger than the previous iteration:

$$Q(\theta^{(k+1)}|\theta^{(k)}, y_{obs}) > Q(\theta^{(k)}|\theta^{(k)}, y_{obs}). \tag{12}$$

This is known as the generalized EM or GEM algorithm.

The second variant is known as the expectation-conditional maximization or ECM algorithm (McLachlan and Krishnan (1997)). In this, the parameter vector is split into components, and the conditional likelihood maximized with respect to each component at each iteration, conditional on the current values of the other components. In the case of (7), the natural components are β and θ. Thus in the ECM algorithm, at each iteration, β is estimated by generalized least squares (GLS) conditional on the current value of θ, and then the likelihood is maximized with respect to θ conditional on the current estimate of β. In contrast, the EM algorithm maximizes with respect to β and θ simultaneously at each iteration. ECM requires less computation but typically no more iterations to convergence than EM. The results reported in this paper are based on the EM algorithm, but preliminary studies suggest that either GEM or ECM, or a combination of both, leads to similar estimates and significantly faster convergence.

One point that needs to be made about all the EM algorithms is that they do not produce standard errors in the way Newton-Raphson full likelihood procedures do.

4.2 Implementation

Only the response variable (measurement of $PM_{2.5}$ concentrations, in $\mu g/m^3$) is affected by missing data. All the design variables are observed perfectly. The methods have been programmed in both Fortran and in Stata software (StataCorp. 2001).

The expectation step calculates the fitted values for the GLS regression, using these as predicted values for the missing data:

$$\hat{y}_{st} = \begin{cases} y_{st}, & y_{st} \text{ is non-missing,} \\ x_{st}^T \hat{\beta}^{(k)}, & y_{st} \text{ is missing,} \end{cases} \tag{13}$$

where x_{st} is the vector of covariates associated with observation y_{st}. Note that this is not a strict E step, because that would require a full kriging prediction of the missing y_{st}, which is as computationally demanding as a direct evaluation of the exact likelihood. We have found that, in practice, (13) produces results very similar to the maximum likelihood estimator, though its theoretical properties are a question for future research.

The M step treats $\hat{y}_{st}$ as if they were all observed data, and maximizes the likelihood with respect to parameter vectors β and θ. We return to the E step to update the values of $\hat{y}_{st}$, and so on.

The starting values of the parameters for the algorithm are the OLS regression results for the regression part of the parameter vector, and some reasonable guesses for the covariance part.

4.3 Kriging

The form of spatial interpolation we use in this paper is known as universal kriging because it combines the regression function $x_t\beta$ with prediction of the spatial random field η_t. The basic mathematics of universal kriging have been given by numerous authors, e.g. Cressie (1993), Section

3.4 or Stein (1999), Section 1.5, but we give an independent derivation here to make clear how the method is applied to construct a time-averaged $PM_{2.5}$ field.

Suppose in week t we have observation vector y_t and we want to predict the value $y_{s_0,t}$ at some unmonitored site s_0. The non-standard feature of this problem is that we want to make such predictions simultaneously for several values of t, together with averages or possibly other functions over time, but still based on a single common estimator of β. The usual derivation of universal kriging is based on only a single replication of the random field.

Suppose the vector of covariates at time t at site s_0 is denoted $x_{s_0,t}$ (assumed known), and write

$$y_{s_0,t} = x_{s_0,t}^T \beta + \eta_{s_0,t}. \tag{14}$$

Write $\begin{pmatrix} \Sigma_t & \tau_t \\ \tau_t^T & \sigma_t^2 \end{pmatrix}$ for the joint covariance matrix of $\begin{pmatrix} \eta_t \\ \eta_{s_0,t} \end{pmatrix}$.

Based on η_t alone, the natural predictor of $\eta_{s_0,t}$ would be $\hat{\eta}_{s_0,t} = \tau_t^T \Sigma_t^{-1} \eta_t$ with prediction error variance $\sigma_t^2 - \tau_t^T \Sigma_t^{-1} \tau_t$; this follows from standard theory of conditional means and variances in the multivariate normal distribution. For the case where β is unknown but the covariances Σ_t, τ_t and σ_t^2 are known (the standard set-up of universal kriging), the appropriate procedure is to estimate β by the GLS estimator $\hat{\beta}$, and the point predictor of $y_{s_0,t}$ is then given by

$$\hat{y}_{s_0,t} = x_{s_0,t}^T \hat{\beta} + \tau_t^T \Sigma_t^{-1} (y_t - x_t \hat{\beta}). \tag{15}$$

Combining (14) and (15), we see that

$$\hat{y}_{s_0,t} - y_{s_0,t} = \left(x_{s_0,t}^T - \tau_t^T \Sigma_t^{-1} x_t\right) (\hat{\beta} - \beta) + \left(\tau_t^T \Sigma_t^{-1} \eta_t - \eta_{s_0,t}\right). \tag{16}$$

The key point in the calculation of prediction error variance is that the two terms in (16) are independent: this follows because $\tau_t^T \Sigma_t^{-1} \eta_t - \eta_{s_0,t}$ is independent of η_t and hence of $\hat{\beta} - \beta = (x_t^T \Sigma_t^{-1} x_t)^{-1} x_t^T \Sigma_t^{-1} \eta_t$.

In the case where all the observations are taken at a single time point t, the covariance matrix of $\hat{\beta}$ is $(x_t^T \Sigma_t^{-1} x_t)^{-1}$, the variance of $\tau_t^T \Sigma_t^{-1} \eta_t - \eta_{s_0,t}$ is $\sigma_t^2 - \tau_t^T \Sigma_t^{-1} \tau_t$, and (16) quickly leads to

$$E\left\{\hat{y}_{s_0,t} - y_{s_0,t}\right\}^2 = \left(x_{s_0,t}^T - \tau_t^T \Sigma_t^{-1} x_t\right)\left(x_t^T \Sigma_t^{-1} x_t\right)^{-1}\left(x_{s_0,t} - x_t^T \Sigma_t^{-1} \tau_t\right) + \sigma_t^2 - \tau_t^T \Sigma_t^{-1} \tau_t,$$

which is equivalent to the usual characterization of the mean squared prediction error in universal kriging, e.g. Stein (1999), equation (11), page 8.

For predictions over multiple time points, a typical calculation is the following. Suppose we are interested in $\sum_{t=t_1}^{t_2} \hat{y}_{s_0,t}$ as a predictor of $\sum_{t=t_1}^{t_2} y_{s_0,t}$. (In practice we usually divide by $t_2 - t_1 + 1$, to make it an average over the weeks from t_1 to t_2.) By (16),

$$\sum_{t=t_1}^{t_2} \hat{y}_{s_0,t} - \sum_{t=t_1}^{t_2} y_{s_0,t} = \sum_{t=t_1}^{t_2} \left(x_{s_0,t}^T - \tau_t^T \Sigma_t^{-1} x_t\right)(\hat{\beta} - \beta) + \sum_{t=t_1}^{t_2} \left(\tau_t^T \Sigma_t^{-1} \eta_t - \eta_{s_0,t}\right) \tag{17}$$

and the two terms in (17) are independent: therefore

$$\begin{aligned} E\left\{\left(\sum_{t=t_1}^{t_2} \hat{y}_{s_0,t} - \sum_{t=t_1}^{t_2} y_{s_0,t}\right)^2\right\} &= \left\{\sum_{t=t_1}^{t_2} \left(x_{s_0,t}^T - \tau_t^T \Sigma_t^{-1} x_t\right)\right\} Cov\{\hat{\beta}\} \left\{\sum_{t=t_1}^{t_2} \left(x_{s_0,t} - x_t^T \Sigma_t^{-1} \tau_t\right)\right\} \\ &\quad + \sum_{t=t_1}^{t_2} \left(\sigma_t^2 - \tau_t^T \Sigma_t^{-1} \tau_t\right), \end{aligned} \tag{18}$$

where $Cov\{\hat{\beta}\}$ denotes the covariance matrix of $\hat{\beta}$, which is calculated at the same time as $\hat{\beta}$ itself is calculated.

Note that even in the case $t_1 = t_2$, the formula (18) does not reduce to (17), because the $\hat{\beta}$ is still based on the entire data set. The two formulae are equivalent only if the entire analysis is based on a single time point.

For the case of an intrinsically stationary process, as defined by (6), these calculations cannot be directly applied because the covariance matrices (Σ_t, etc.) are not defined. However, the preceding calculations are applied with y_t, x_t and Σ_t replaced by y_t^*, x_t^* and Σ_t^* respectively, as already explained in connection with likelihood maximization.

One further point should be mentioned about the implementation of these formulae. The variograms we have fitted include a nugget effect. This is most usually interpreted as a measurement error: the observed data are viewed as a smooth random field plus measurement error. The kriging formulae predict the smooth random field ignoring the measurement error. For this reason, even if s_0 is very close to a monitor, the predicted value will not be the same as the monitor. (In different language, this version of kriging is a smoothing procedure, not an interpolation procedure.) For that reason, the predicted random fields look smoother than the raw data. Another aspect of this is the interpretation of σ_t^2 in (18). This can be calculated two ways: with or without the nugget. In the following calculation, it has been calculated without the nugget, so that the quoted prediction error variance reflects the estimated variance compared with the smooth random field without measurement error. If the predicted random field were to be compared with future monitor values, it would be necessary to increase σ_t^2 to take account of measurement error.

As with most kriging calculations, the mean squared prediction errors given by (18) allow for the estimation of β but do not allow for the estimation of θ_1 and θ_2; it is possible to do this, approximately via Taylor expansions (e.g. Zimmerman and Cressie (1992)) or by Bayesian methods (Handcock and Stein (1993)), but both methods involve additional computation and we do not consider them further here.

5 Results

Several runs of the algorithm were tried with different covariates for the regression part and different variogram models, but for the present discussion we focus on just one model and one variogram. The variogram is (6), and the regression model includes the following covariates: four independent

dummy variables corresponding to land-use effects, linear terms in latitude and longitude, and 40 terms corresponding to spatial thin-plate spline basis functions (3). The eventual decision to include 40 spline basis functions was taken after repeating the model fit with 10, 20, 30, 40 and 50 basis functions, using likelihood ratio tests to compare the fits of different models. Since the model is based on space-centered variables $y^*_{st} = y_{st} - \bar{y}_{\cdot t}$, there is no need for a separate "week effect"; however, the kriging step gives us predictions for $y_{s_0 t} - \bar{y}_{\cdot t}$ at each unmonitored site s_0, and we must add $\bar{y}_t$ to this, to obtain a prediction for $y_{s_0 t}$. Thus, the week effects are still reflected in the final predictions, but they are not estimated as part of the regression model.

Since the E-step in (13) is only an approximate E-step, ignoring the correlations among ys, questions remain open about how close the resulting EM-based estimates are to the true MLE, and whether the standard errors obtained from standard likelihood asymptotics can be trusted in this setting. We have not attempted to resolve these questions theoretically, but we have made numerical comparisons between the EM approach and exact maximum likelihood. In these comparisons, the parameters θ_1 and θ_2 in (6) are estimated as those values that maximize the (EM or exact) likelihood function, and the standard errors are estimated in the usual way from the observed information matrix. The parameter α is estimated as $\hat{\alpha} = G^2/(n-p)$ with standard error $\hat{\alpha}\sqrt{2/(n-p)}$ where $G^2 = \sum_t (y^*_t - x^*_t \hat{\beta}) \Sigma^*_t(\theta)^{-1} (y^*_t - x^{*T}_t \hat{\beta})$ is the generalized residual sum of squares calculated using the GLS estimator $\hat{\beta}$; here n is the total number of non-missing observations and p is the number of parameters estimated in the regression model. The parameter estimates and standard errors are reported in Table 1. The estimates of θ_1 and θ_2 are very comparable in the two approaches: that of α is somewhat smaller in the EM approach, as are all three standard errors.

The most likely explanation for $\hat{\alpha}$ and the standard errors being smaller in the EM approach is that the E step of the EM algorithm treats all the $\hat{y}_{st}$ values as known, not allowing for the fact that

Table 1: Comparison of EM estimates with exact maximum likelihood (standard error in parentheses)

Method	θ_1	θ_2	α
MLE	2.06	0.92	0.061
	(0.35)	(0.097)	(0.0017)
EM	2.13	0.92	0.049
	(0.29)	(0.083)	(0.0012)

some are estimated. A first-order correction to this would be to divide $\hat{\alpha}$ by 0.72, and the standard errors by $\sqrt{0.72}$, where 0.72 is the proportion of non-missing data. With this correction, the EM estimate of α becomes .0675, and the three standard errors become 0.35, 0.098, .0019, much closer to the maximum likelihood values.

Following the parameter estimation, we perform a kriging operation to construct a predicted $PM_{2.5}$ surface, for the whole region, for each week of the study, and also averaged over all weeks, using the procedure described in Section 4.3. For this, the maximum likelihood estimates from Table 1 were used. At the end of this procedure, $\bar{y}_{\cdot t}$ is added to the prediction $\hat{y}^*_{s_0,t}$ to obtain a prediction $\hat{y}_{s_0,t}$; the actual predicted $PM_{2.5}$ is $\hat{y}^2_{s_0,t}$. Approximate mean squared prediction errors are calculated in this last step using a Taylor approximation.

Because we have identified five land use types, there is a question to decide how to handle that variable in the predictions, since the current data base available for the present analysis does not include land uses at sites away from the monitors. This problem has been handled by treating all the prediction sites as residential sites. This is because the main focus of interest, for human health

assessment, is the level of pollution at residential sites: so we focus on these. Of course, it would be desirable in a future analysis to take account of the different land use types more thoroughly.

The final results presented here are the predicted $PM_{2.5}$ surface for week 33, which according to the estimated week effects was the week with highest $PM_{2.5}$ values in the entire series — this therefore reflects the $PM_{2.5}$ field under the most adverse conditions observed in the present data set — and the estimated average $PM_{2.5}$ field over the 49-week time period. The latter is particularly interested in assessing which parts of the field are in violation of the proposed 15 $\mu g/m^3$ standard for long-term mean $PM_{2.5}$. The results are shown in Fig. 6(a) for week 33 and Fig. 6(c) for the overall average, while the corresponding estimates of root mean squared prediction error are shown in Fig. 6(b) and Fig. 6(d) respectively. The latter plots both show an apparent peak of mean squared prediction error in a small region on the NC/SC border, just south of Charlotte — we have no explanation of this, but the same phenomenon was observed in different plots using different versions of the spatial model.

Finally in Fig. 7, we show a spatial map of the estimated probability that any given location is in violation of the 15 $\mu g/m^3$ standard. This was calculated using the predicted value of the 49-week mean at each site, together with the mean squared prediction error, using a normal probability approximation to calculate the probability that the true mean exceeds 15 $\mu g/m^3$ at each site. As can be seen, much of the region appears to be in violation of the standard. For comparison, the cities of Atlanta (A), Charlotte (C) and Raleigh (R) are marked on the plot: the first two seem to be clearly in violation of the standard and the third may well be as well.

6 Conclusions

This paper proposes a model for spatial-temporal $PM_{2.5}$ data in which the $PM_{2.5}$ field is represented as a sum of three fixed components and a random component. The fixed components consist of a weekly time-trend effect, common to all the stations, a smooth spatial effect represented by thin-plate splines, and a land-use component. The random component is spatially but not temporally correlated. The estimation procedure emphasizes simultaneous estimation of the fixed and random components, and we propose a kriging methodology that also takes account of the fixed as well as random components of the model.

A particular feature of the data is the presence of missing values, which complicates evaluation of the likelihood function. As an alternative to exact evaluation, we propose an approximate method based on the EM algorithm. The comparisons in the present paper suggest that the approximate method produces results comparable to the true MLE.

The results are applied to predict the overall field of $PM_{2.5}$, together with estimated mean squared prediction errors, both for one week when $PM_{2.5}$ was especially high, and for the overall means. The results imply that substantial portions of the three states, and Georgia in particular, appear to violate the federal standard on $PM_{2.5}$.

References

Box, G.E.P. and Cox, D.R. (1964), An analysis of transformations (with discussion). *J.R. Statist. Soc. B* **26**, 211–246.

Cox, L.H. (2000), Statistical issues in the study of air pollution involving airborne particulate matter. *Environmetrics* **11**, 611–626.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty.* John Wiley, New York.

Cressie, N. (1993), *Statistics for Spatial Data.* Second edition, John Wiley, New York.

Daniels, M.J., Lee, Y-D. and Kaiser, M. (2001), Assessing sources of variability in measurement of ambient particulate matter. *Environmetrics* **12**, 547–558.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J.R. Statist. Soc. B* **39**, 1–38.

Green, P. J., and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Monographs on Statistics and Applied Probability, Vol. 58. Chapman & Hall.

Gould, W., and Sribney, W. (1999). *Maximum Likelihood Estimation with Stata.* Stata Press, College Station, TX.

Handcock, M.S. and Stein, M. (1993), A Bayesian analysis of kriging. *Technometrics*, **35**, 403-410.

Hartigan, J.A. and Wong, M.A. (1979), A K-means clustering algorithm. *Applied Statistics* **28**, 101–108.

Holland, D.M., De Oliveira, V., Cox, L.H. and Smith, R.L. (2000), Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics* **11**, 373–393.

Kitanidis, P.K. (1983), Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research* **19**, 909-921.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* Wiley Series in Probability and Statistics, John Wiley, New York.

Matheron, G. (1973), The intrinsic random functions and their applications. *Adv. Appl. Prob.* **5**, 439–468.

McLachlan, G.J., and Krishnan, T. (1997) *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics, John Wiley, New York.

Meng, X.-L. and van Dyk, D. (1997), The EM algorithm — an old folk song sung to a fast new tune (with discussion). *J.R. Statist. Soc. B* **59**, 511–567.

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging.* Springer Verlag, New York.

StataCorp. (2001). *Stata Statistical Software: Release 7.* College Station, TX: Stata Corporation.

Zimmerman, D.L. and Cressie, N. (1992), Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.* **44**, 27-43.

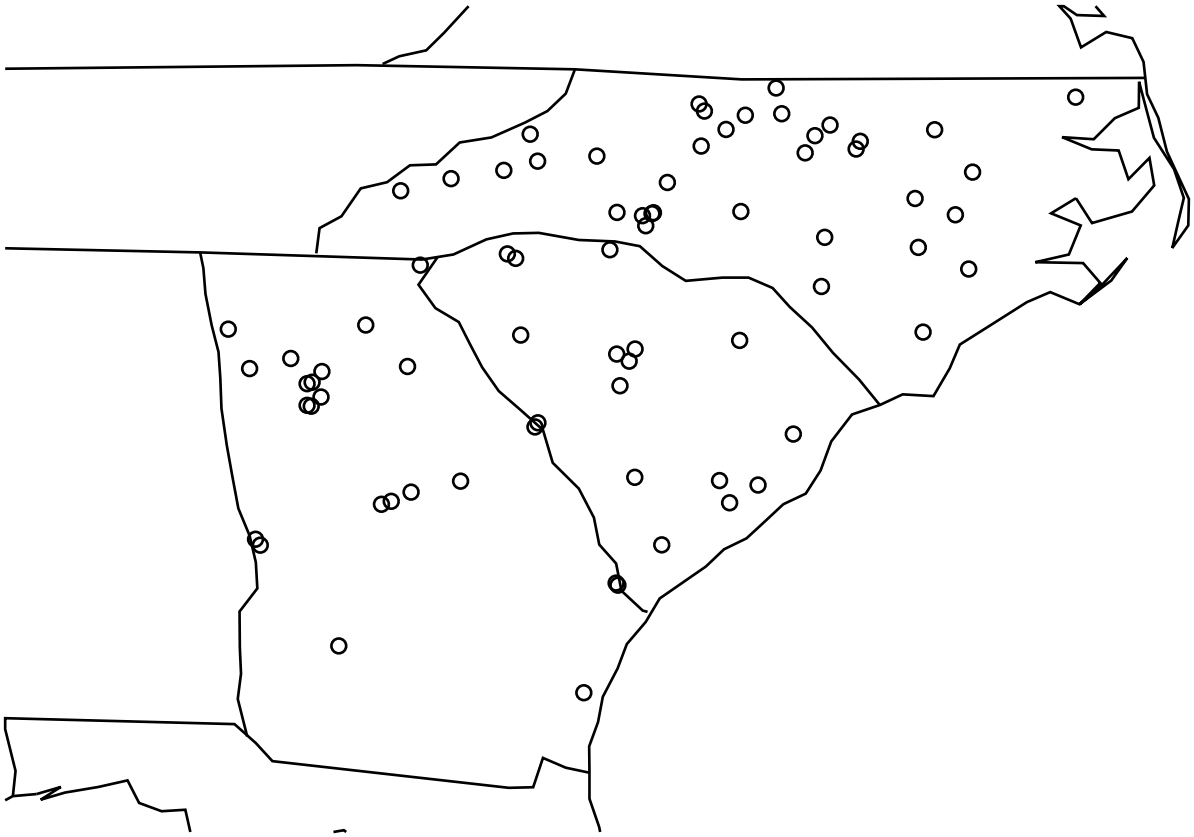

Figure 1: Map of 74 monitors

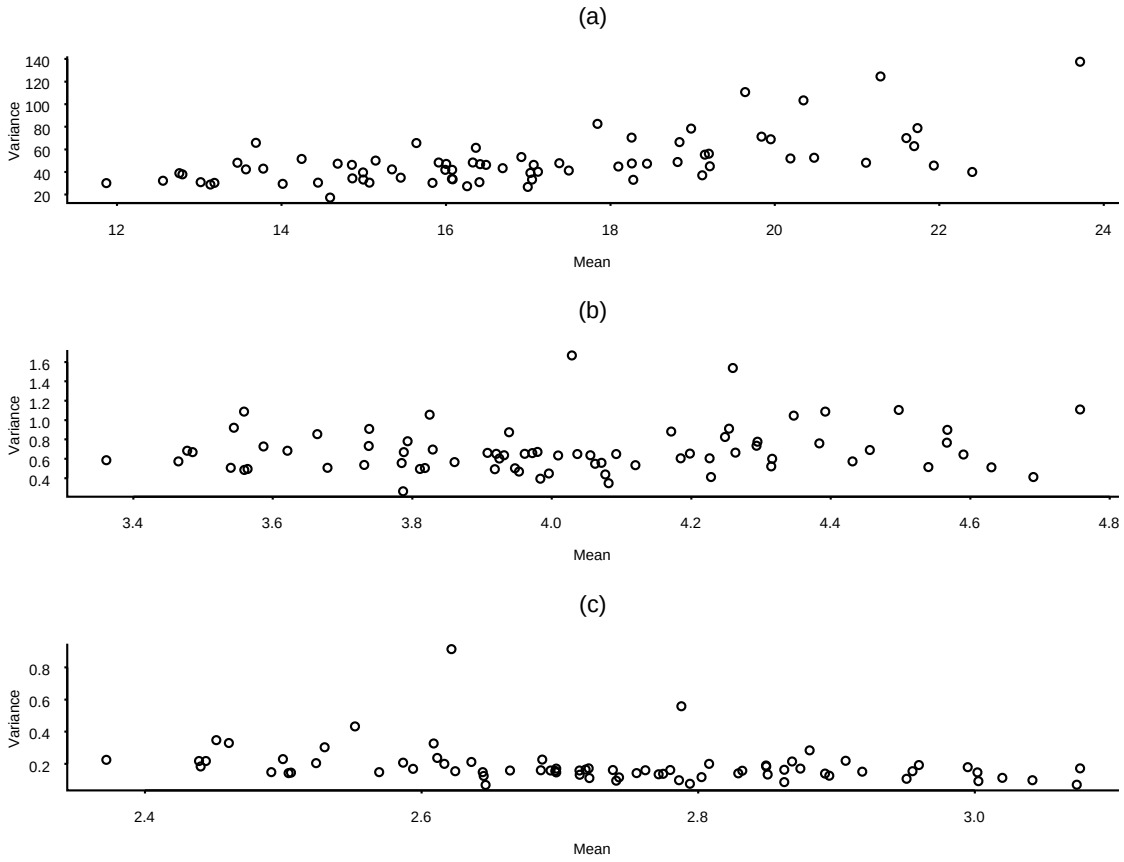

Figure 2: Variance *vs.* mean plot for $PM_{2.5}$ values at each of the 74 stations. (a) Original data, untransformed. (b) Square root transform. (c) Logarithmic transform.

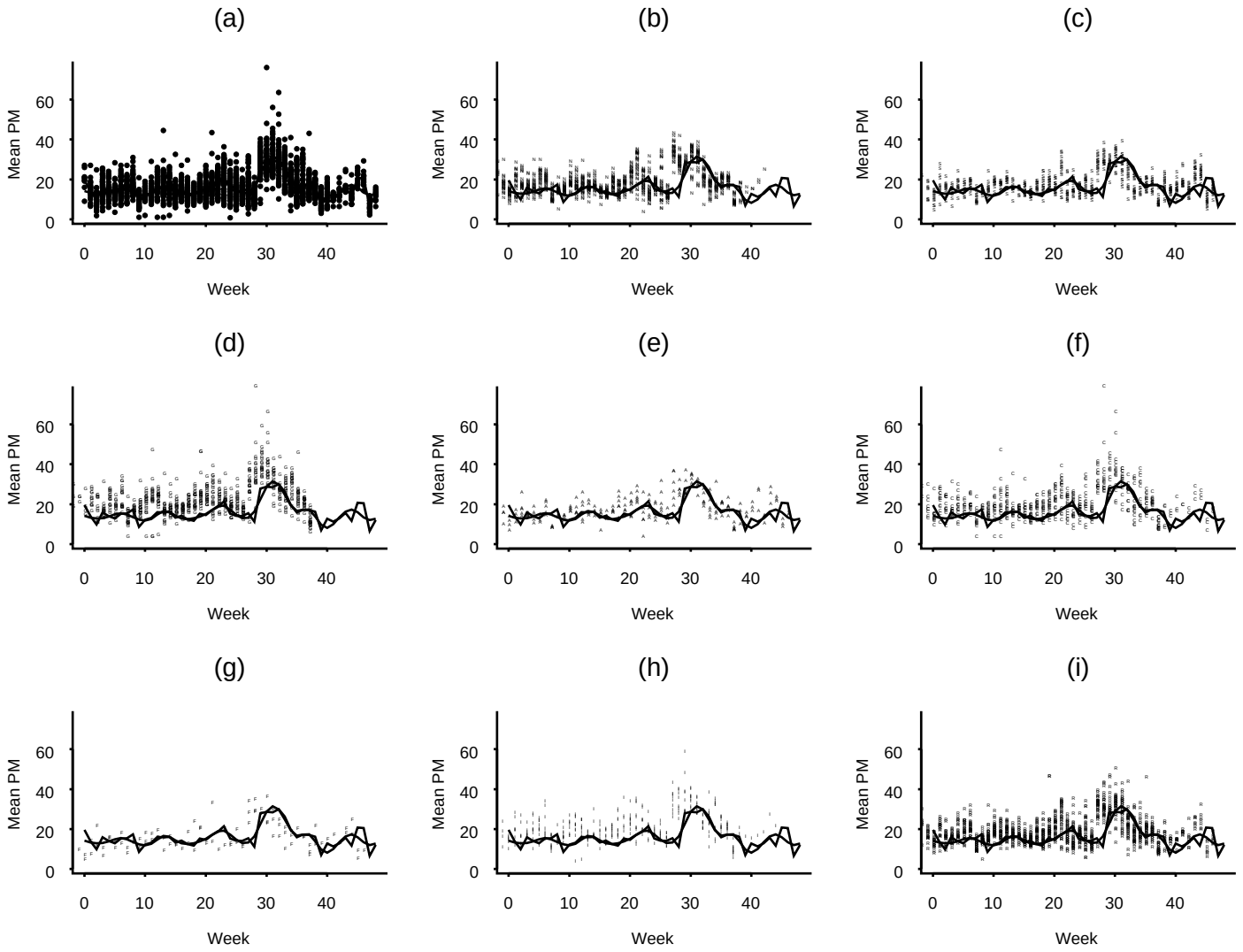

Figure 3: The comparison of the fitted trend and the raw data for subpopulations. (a) All data combined; (b) North Carolina; (c) South Carolina; (d) Agricultural sites; (e) Commercial sites; (f) Forest sites; (g) Industrial sites; (h) Residential sites. Plotted curves are the *overall* fits of the weekly effect and the result of a 20-DF smoothing spline.

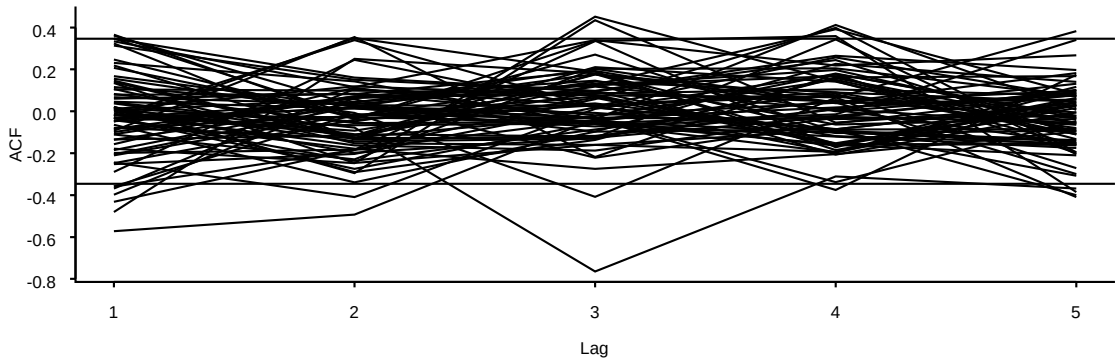

Figure 4: Time-autocorrelation plots for the 74 stations with approximate 95% confidence bands

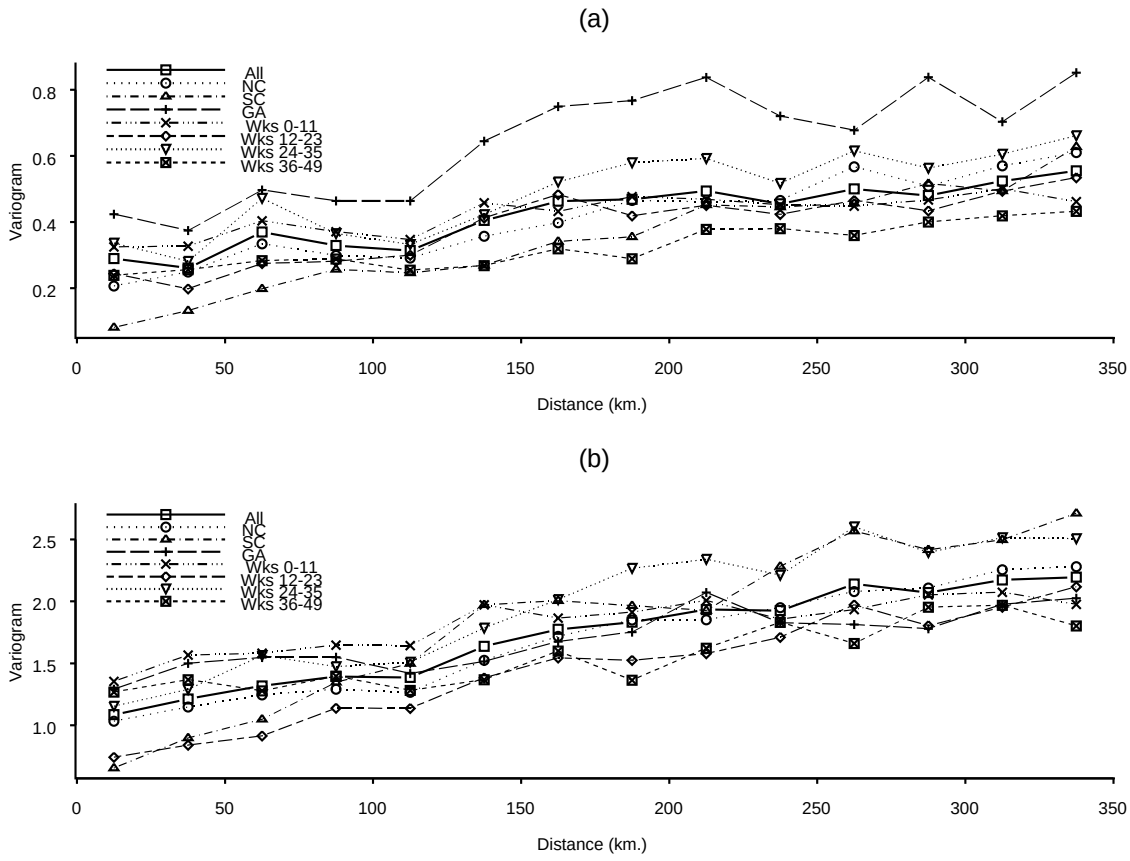

Figure 5: Variogram plots for residuals after fitting time trend, spatial trend and type effects; all data combined, and separate plots by state and by season. (a) Without standardizing variances. (b) After standardizing the sample variance of residuals at each station to be 1.

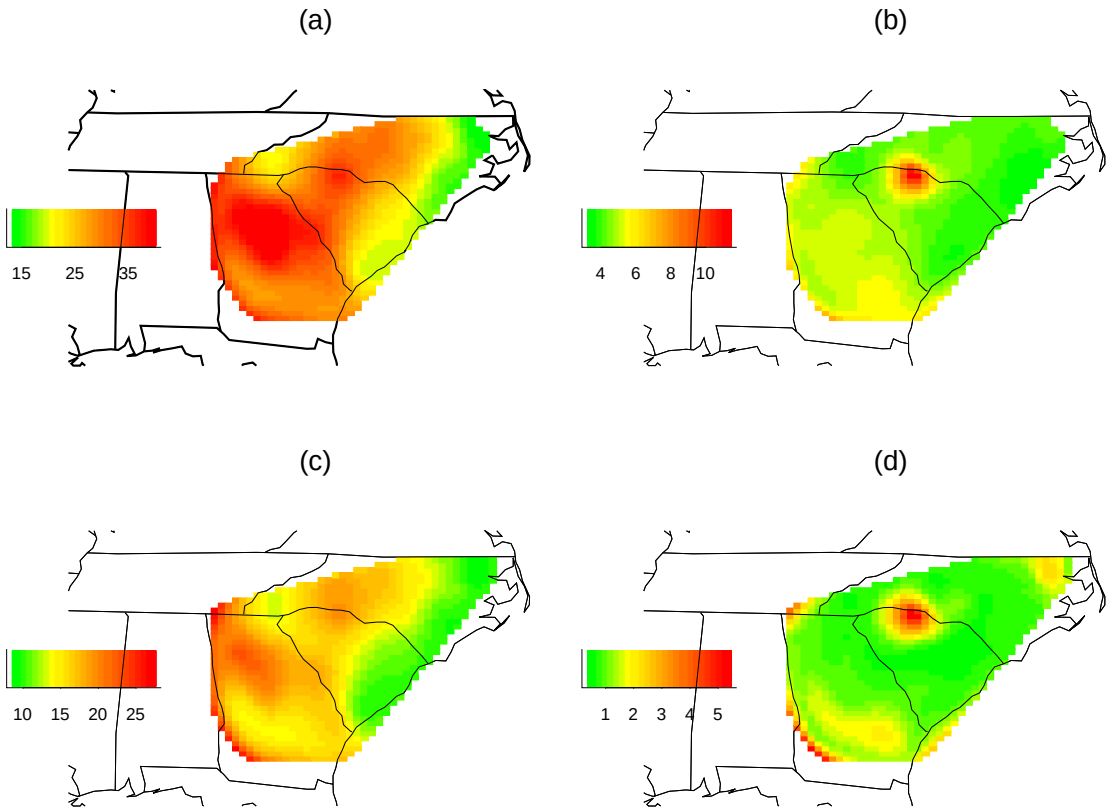

Figure 6: Plots of the predicted surface for $PM_{2.5}$. (a) Predicted surface for week 33. (b) Estimated prediction standard error for week 33. (c) Predicted surface for average of weeks 1–49. (d) Estimated prediction standard error for average of weeks 1–49.

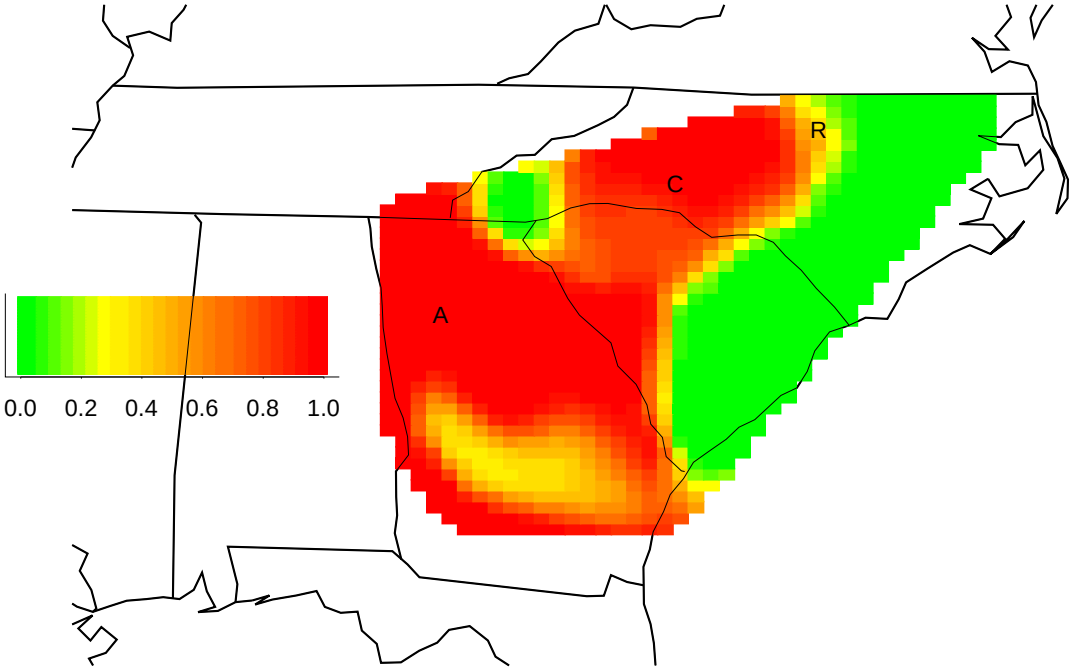

Figure 7: Plot of estimated probability that any given location is in violation of the proposed standard for long-term mean $PM_{2.5}$. The symbols A, C and R mark the cities of Atlanta, Charlotte and Raleigh.