ESTIMATING THE MEAN LEVEL OF FINE PARTICULATE MATTER: AN APPLICATION OF SPATIAL STATISTICS

Richard L. Smith
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, N.C., U.S.A.

Boston University
April 16 2004

I. CONSTRUCTING A MAP OF ANNUAL AVERAGES OF FINE PARTICULATE MATTER

Based on Smith, Kolenikov and Cox, *Journal of Geophysical Research*, 2003.

II. SOME THEORETICAL ASPECTS OF BAYESIAN SPATIAL PREDICTION

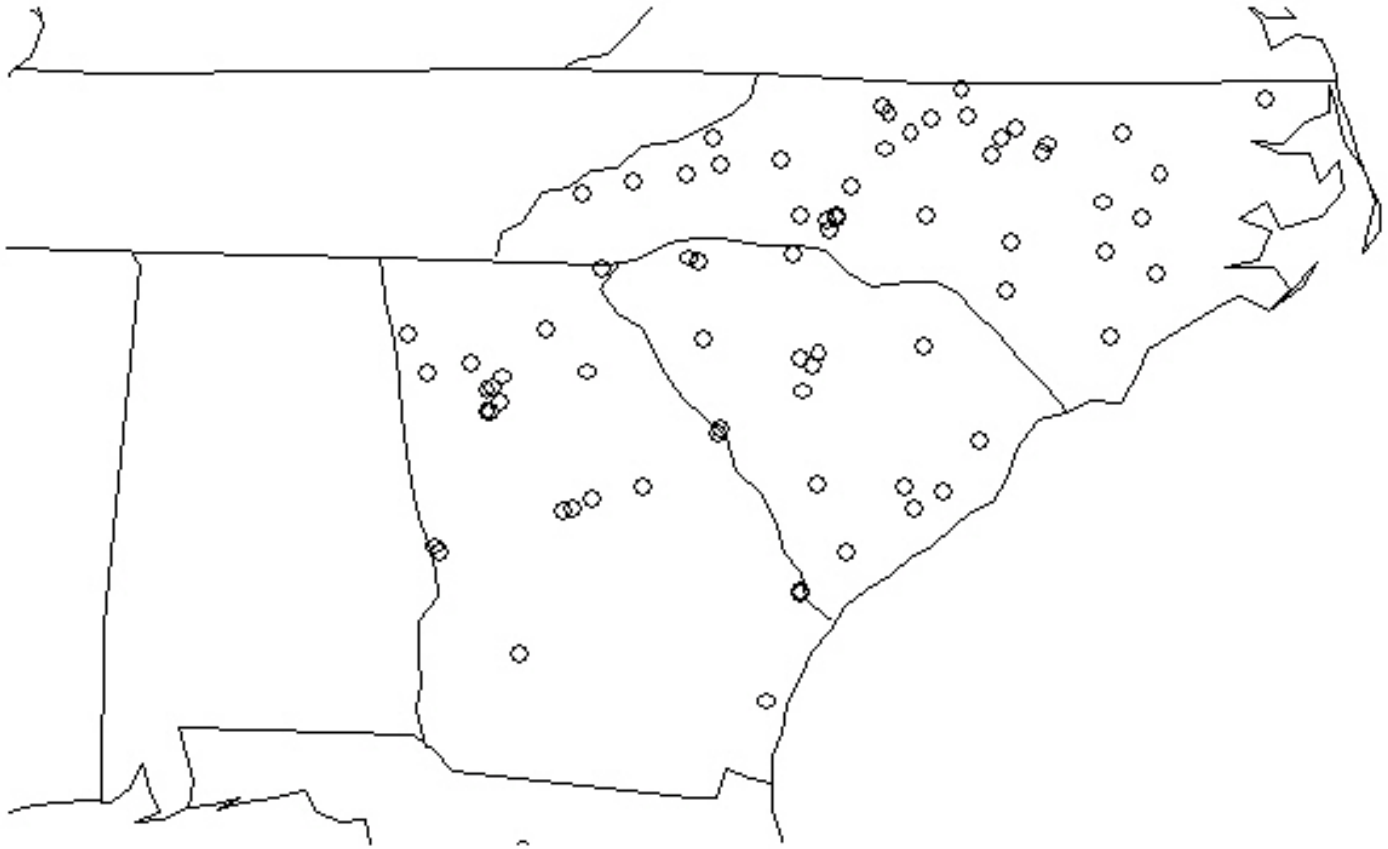
Work currently in progress, with Z. Zhu, J. Ibrahim and E. Shameldin (UNC).

Background to Part I

A new set of air pollution standards, first proposed in 1997, is finally being implemented by the U.S. Environmental Protection Agency (EPA). One of the requirements is that the mean level of fine particulate matter ($PM_{2.5}$) at any location should be no more than 15 $\mu g/m^3$. A network of several hundred monitors has been set up to assess this.

The present study is based on 1999 data for a small portion of this network, 74 monitors in North Carolina, South Carolina and Georgia. We converted the raw values to weekly averages, but even so more than $\frac{1}{4}$ of the data are missing. The EPA also recorded a "land-use" variable, classified as one of five types of land-use: agricultural (A), commercial (C), forest (F), industrial (I) and residential (R).

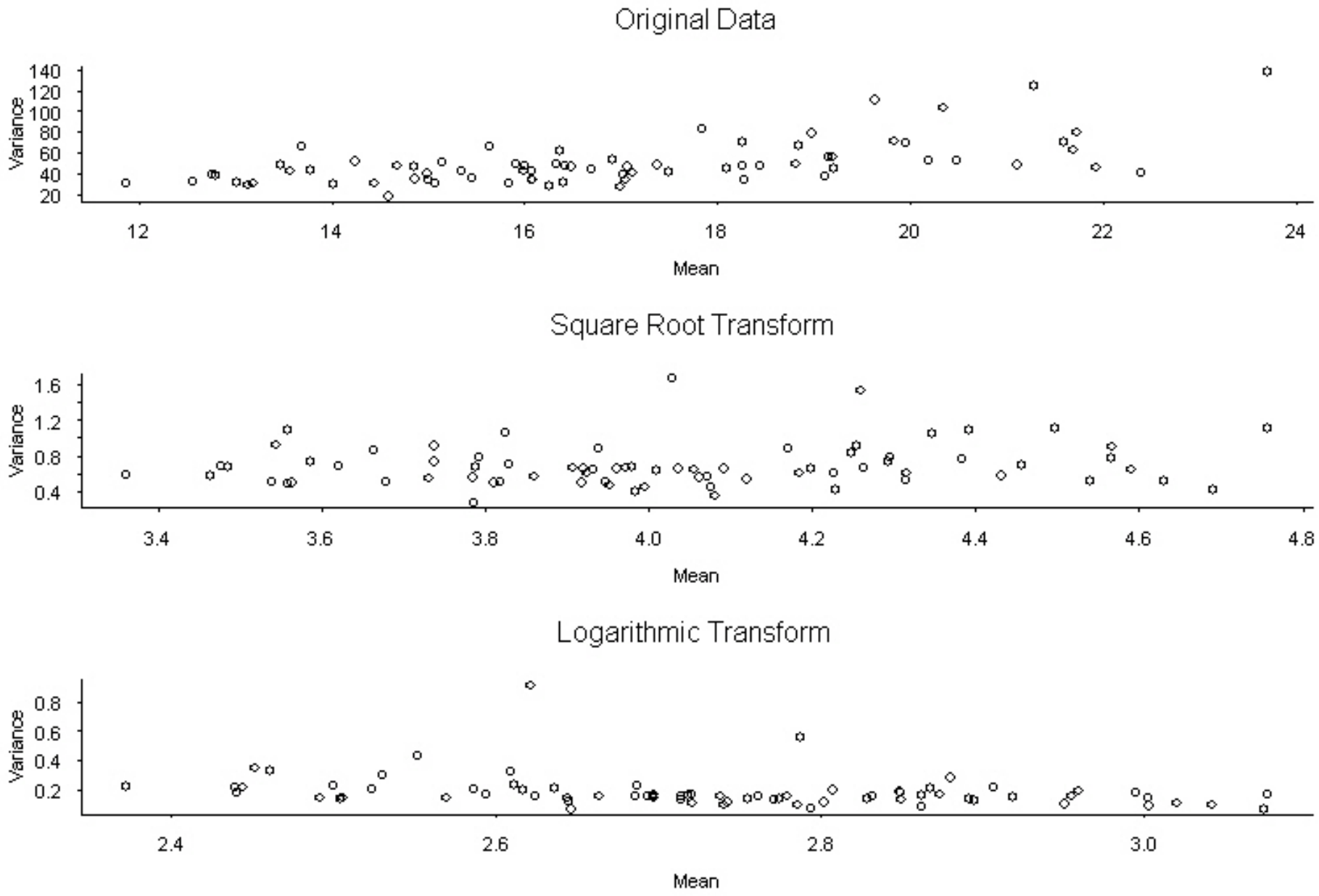
Map of 74 Stations



Exploratory data analysis

The first issue considered is whether to make any transformation, such as square roots or logarithms, of the raw $PM_{2.5}$ values. We show a plot of sample variance against sample mean, across all 74 stations, for each of three transformations, (a) no transformation, (b) square root transformation, (c) logarithmic transformation. On the basis that (b) is the closest fit to a constant-variance model, the rest of the analysis is based on the square root of $PM_{2.5}$ as a variance-stabilizing transformation.

Mean-Variance Plots

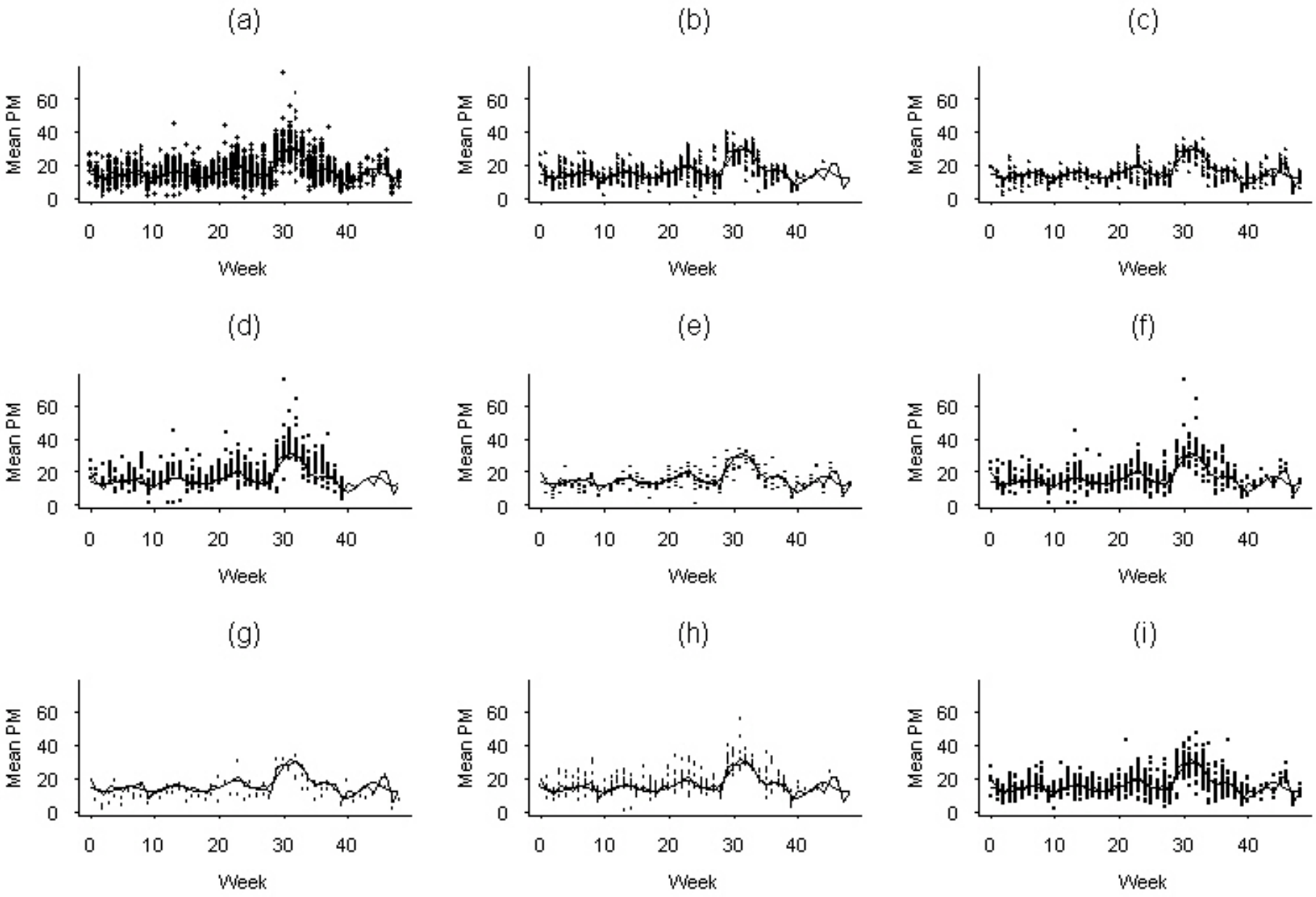


The time trend

The time trend was estimated both as a B-spline smooth curve and (more simply) by using a weekly indicator variable to represent the overall mean level for that week.

Fig. (a) shows both versions of the fitted time trend, with all data points superimposed. Also shown are the same fitted time trend curves, but with different portions of the data superimposed, (b)–(d) corresponding to each of the three states, (e)–(i) corresponding to each of the five land-use variables. The results show a significant discrepancy between states, with Georgia values generally higher than the overall mean, while the land-use variables show significant variations in the directions one would expect.

Non-linear Trend with Selected Subsets of Data



These comparisons suggest the model

$$y_{xt} \;=\; w_t + \psi_x + \theta_x + \eta_{xt} \tag{1}$$

in which y_{xt} is the square root of $\mathrm{PM}_{2.5}$ in location x in week t, w_t is a week effect, ψ_x is the spatial mean at location x (in practice, estimated through a thin-plate spline representation), θ_x is a land-use effect corresponding to the land-use as site x, and η_{xt} is a random error.

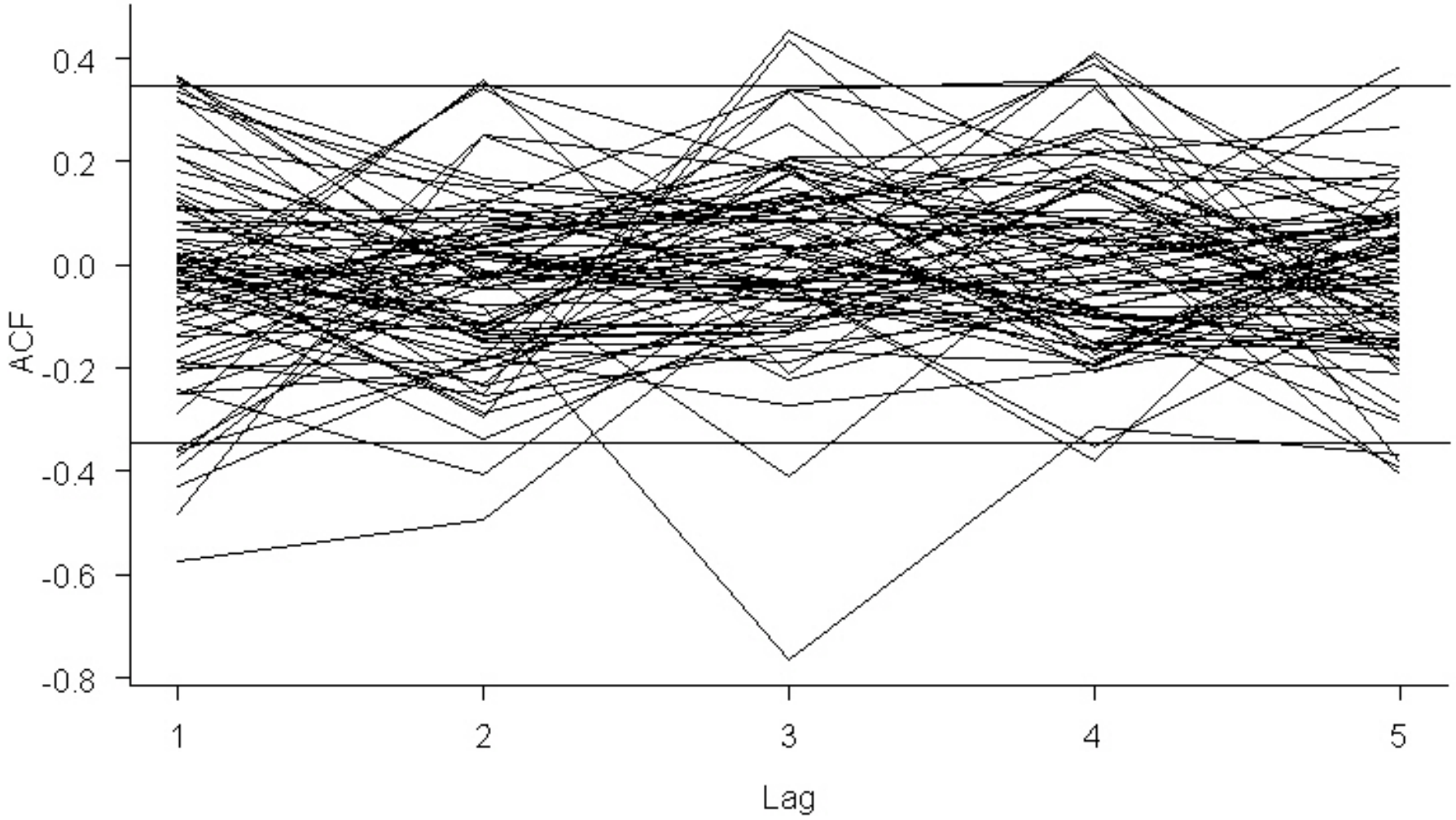
So far we have ignored temporal and spatial correlations among the η_{xt}, but we consider these next.

Spatial and temporal dependence

Take residuals from preceding linear regression.

Plots of autocorrelations suggest series are uncorrelated in time but not in space.

Autocorrelation Plots for 74 Stations

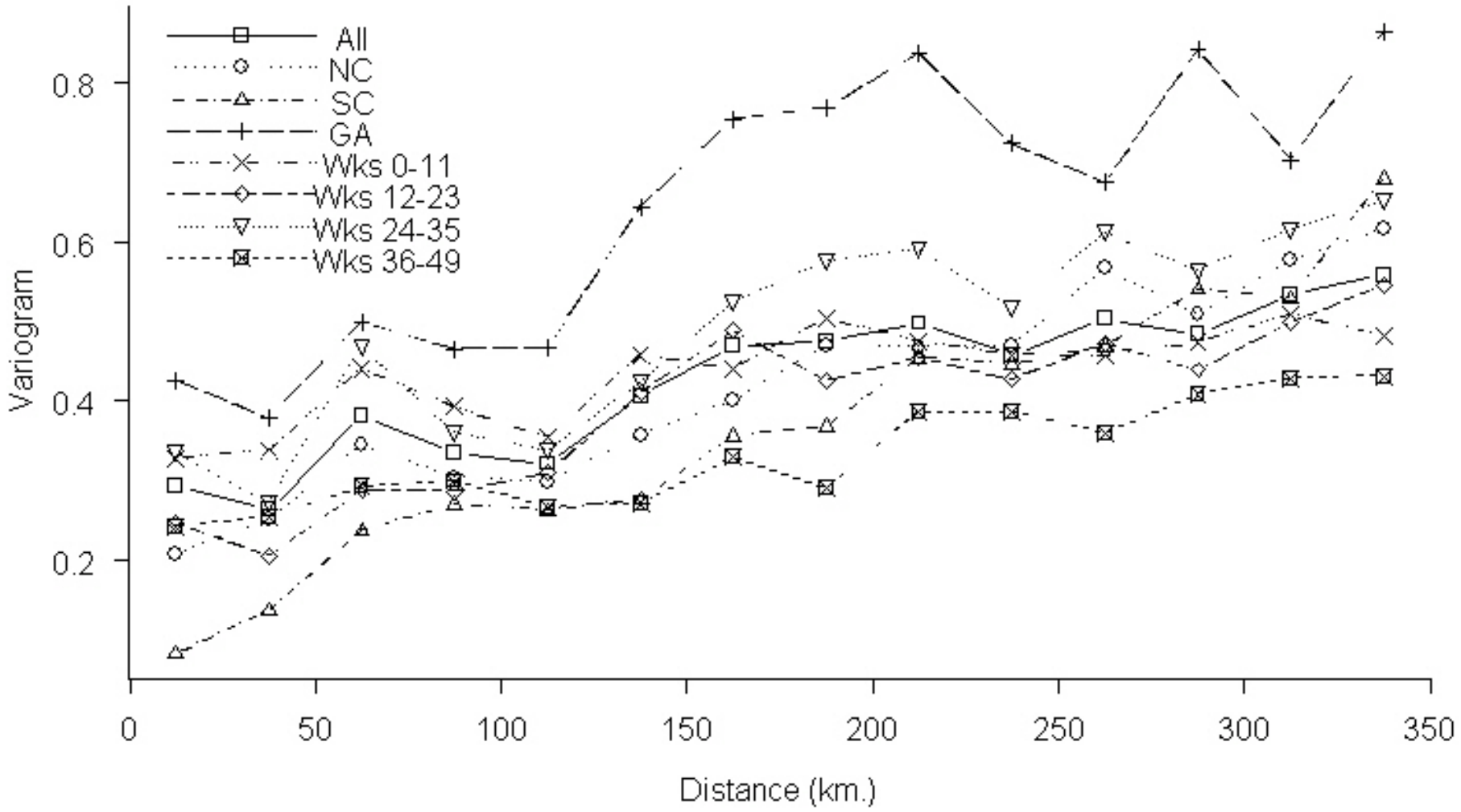


We show variograms of residuals from simple linear regression, where a number of subsets of the data (classified by state and also by season) have been identified to look for comparability of the estimated variogram among different subsets of data. Key points are

- Substantial inhomogeneity among subgroups despite initial variance stabilization

- Does not seem to follow standard nugget-range-sill shape

Variogram Plots for Selected Subsets of Data



We fit the power law variogram

$$\gamma(h) \;=\; \begin{cases} 0 & \text{if } h = 0, \\ \theta_0 + \theta_1 h^{\lambda} & \text{if } h > 0, \end{cases} \tag{2}$$

where $\theta_0 > 0,\ \theta_1 > 0,\ 0 \leq \lambda < 2$.

To fit this model by maximum likelihood, we need the concept of *generalized covariances*, introduced by Matheron (1973). For modern references see Cressie (1993), Chilès and Delfiner (1999) or Stein (1999). In the present context the key formula is the following: for an intrinsically stationary process defined by a semivariogram γ,

$$\begin{aligned} & Cov\left\{\sum_x \nu_x \eta_{x,t}, \sum_{x'} \kappa_{x'} \eta_{x',t}\right\} \\ = & \sum_x \sum_{x'} \nu_x \kappa_{x'} G(||x - x'||), \end{aligned}$$

provided $\sum_x \nu_x = \sum_{x'} \kappa_{x'} = 0$. Here G is known as the generalized covariance function: however for an intrinsically stationary process, it suffices to take $G = -\gamma$.

Practical implementation:

In (1), replace each y_{xt} by $y^*_{xt} = y_{xt} - \frac{1}{n_t}\sum_{x'} y_{x't}$ where the second sum is over all x' values available in week t; n_t is the number of such x' values in a given week. With some further simplifications we replace (1) by

$$y^*_{xt} \;=\; \psi^*_x + \theta^*_x + \eta^*_{xt} \tag{3}$$

where

$$\begin{aligned} Cov\{\eta^*_{x,t}, \eta^*_{x',t}\} &= \frac{1}{n_t}\sum_{x_1} \gamma(||x - x_1||) \\ &+ \frac{1}{n_t}\sum_{x_1} \gamma(||x' - x_1||) - \gamma(||x - x'||) \\ &- \frac{1}{n_t^2}\sum_{x_1}\sum_{x_2} \gamma(||x_1 - x_2||). \end{aligned} \tag{4}$$

The model defined by (2)—(4) may now be fitted by maximum likelihood.

There are additional complications because of the missing values, which mean that n_t and the fitted covariance matrix are different from week to week. The present data set is relatively small and we were still able to compute exact maximum likelihood, but some variants of the EM algorithm (Little and Rubin 1987, McLachlan and Krishnan 1997) were also used, and remain the focus of further research.

Results

The model (3) was fitted to the data values from which each weekly mean had been subtracted. The residuals η^*_{xt} were assumed independent at different time points but with spatial covariances given by (4) with (2). As an example of the results, the maximum likelihood of the parameter θ_2 was 0.92 with standard error 0.097. Since a linear variogram corresponds to $\theta_2 = 1$, this shows that the spatial dependence is not significantly different from a linear variogram.

The fitted model was then used to construct a predicted surface, with estimated root mean squared prediction error (RMSPE), for each week of the year and also for the average over all weeks. The latter is of greatest interest in the context of EPA standards setting.

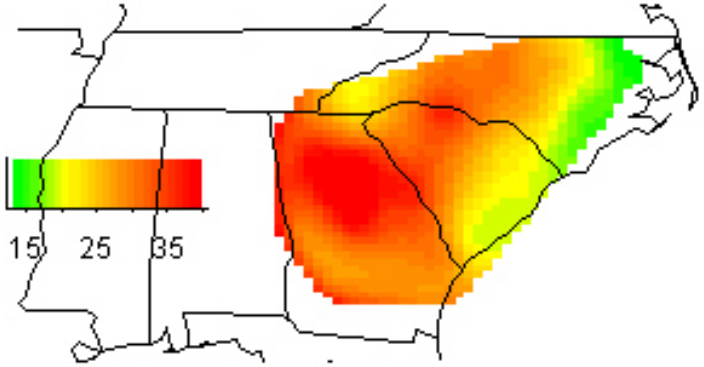
We show the predicted surface and RMSPE for week 33 (the week with highest average $PM_{2.5}$) and overall for the annual mean. WE also show shows the estimated probability that any particular location exceeds the 15 $\mu g/m^3$ annual mean standard. These maps are based on kriging the residuals η^*_{xt} in (2) and then combining them with the estimated fixed effects for ψ^*_x and θ^*_x, transforming back to the original scale of the data for the actual plots. The RMSPE values used here take into account the averaging of kriged values, but do not take account of the additional uncertainty in estimating the parameters θ_1 and θ_2. Fig. 6 is based on the assumption that (on a square root scale) the difference between the predicted and true values, scaled by the RMSPE, has a standard normal distribution.

It can be seen that substantial parts of the region, including the western portions of North and South Carolina and virtually the whole of the state of Georgia, appear to be in violation of the standard. Of the three major cities marked on the last figure, Atlanta and Charlotte are clearly in the "violation" zone; Raleigh is on the boundary of it.

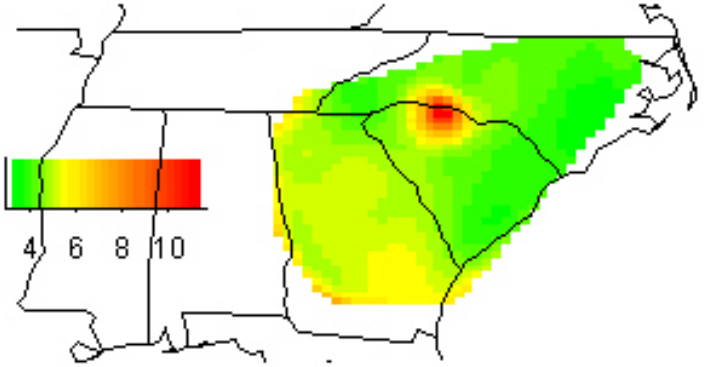
In future work, we hope to extend this analysis to other parts of the country (this will certainly involve consideration of non-stationary spatial models), to analyze more recent data, and to consider the associated "network design" questions.

Predicted PM2.5 Surfaces and RMSPEs

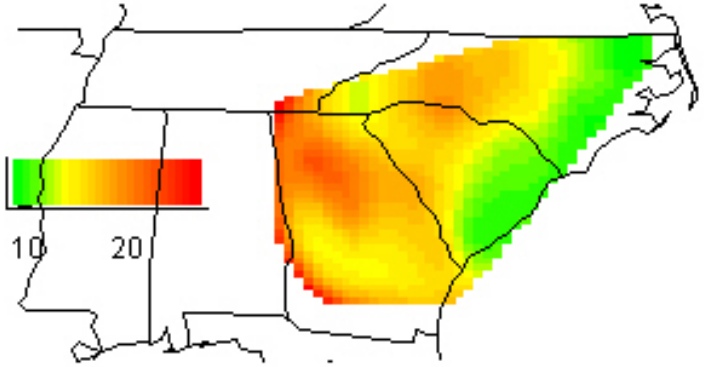
Predicted Surface for Week 33



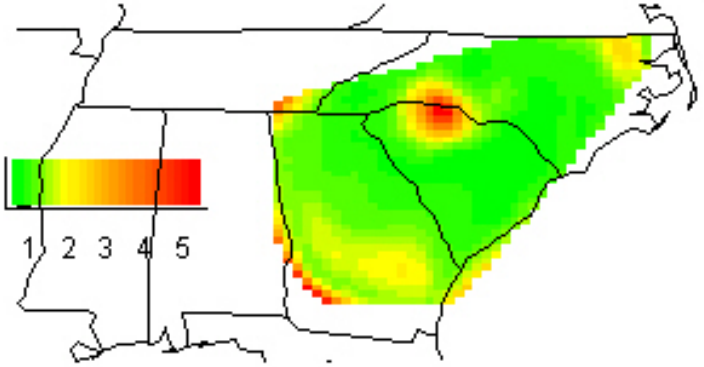
RMSPE for Week 33



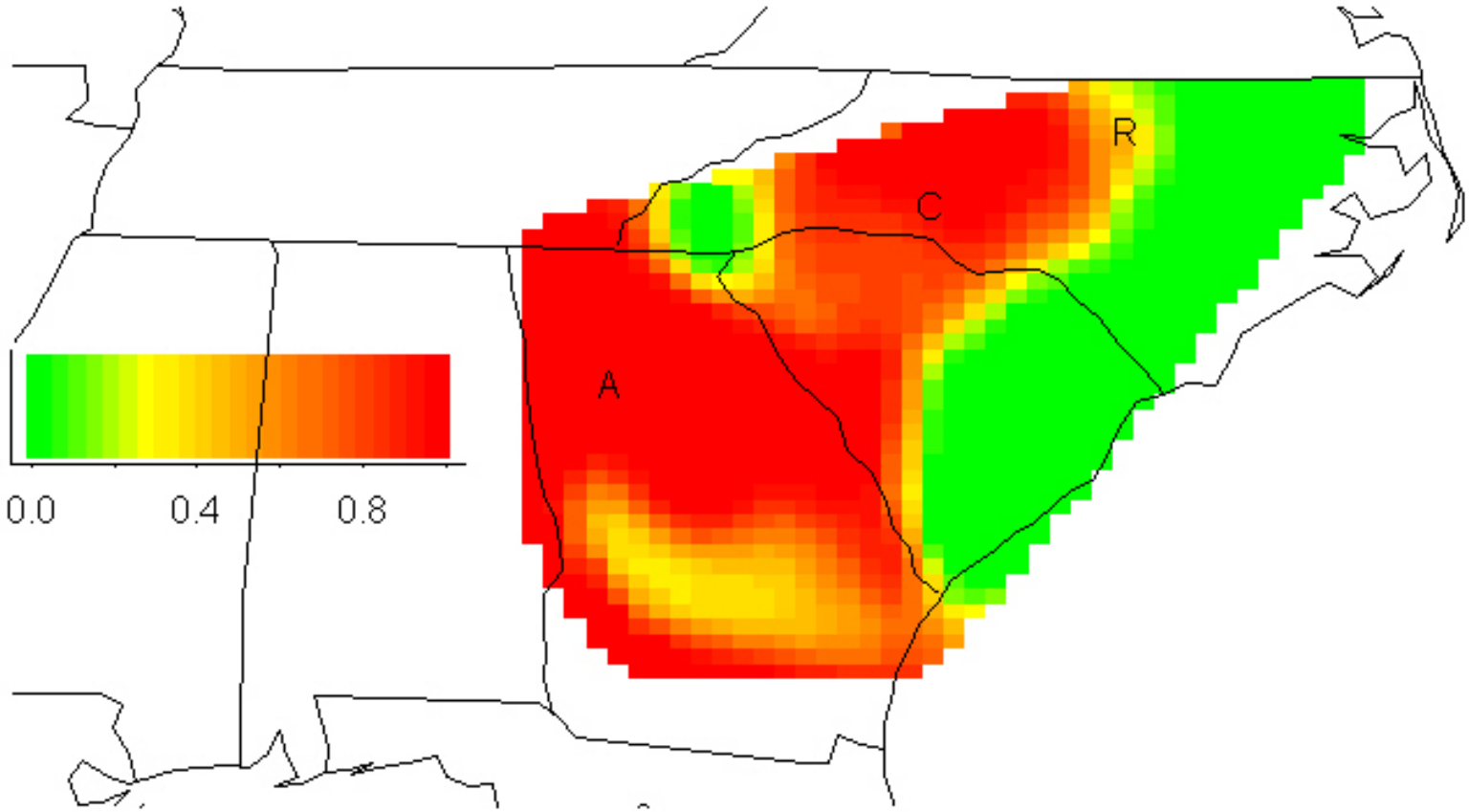
Predicted Surface for Annual Average



RMSPE for Annual Average



Probability of Exceeding Standard



II. SOME THEORETICAL ASPECTS OF SPATIAL PREDICTION

We assume data follow a *Gaussian random field* with mean and covariance functions represented as functions of finite-dimensional parameters.

Define the prediction problem as

$$\begin{pmatrix} Y \\ Y_0 \end{pmatrix} \sim N\left[\begin{pmatrix} X\beta \\ x_0^T\beta \end{pmatrix}, \begin{pmatrix} V & w^T \\ w & v_0 \end{pmatrix}\right] \tag{5}$$

where Y is an n-dimensional vector of observations, Y_0 is some unobserved quantity we want to predict, X and x_0 are known regressors, and β is a p-dimensional vectors of unknown regression coefficients. For the moment, we assume V, w and v_0 are known.

Where notationally convienient, we also define $Y^* = \begin{pmatrix} Y \\ Y_0 \end{pmatrix}$ and write (5) as

$$Y^* \sim N[X^*\beta, V^*]. \tag{6}$$

Specifying the Covariances

The most common and widely used spatial models (stationary and isotropic) assume the covariance between components Y_i and Y_j is a function of the (scalar) distance between them, $C(d_{ij})$. For example,

$$C_\theta(d) = \sigma \exp\left(-\frac{d}{\rho}\right), \tag{7}$$

where $\theta = (\sigma, \rho)$ (exponential),

$$C_\theta(d) = \sigma \exp\left\{-\left(\frac{d}{\rho}\right)^2\right\}, \tag{8}$$

where $\theta = (\sigma, \rho)$ (Gaussian),

$$C_\theta(d) = \frac{\sigma}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\nu^{1/2}d}{\rho}\right)^\nu \mathcal{K}_\nu\left(\frac{2\nu^{1/2}d}{\rho}\right), \tag{9}$$

where $\mathcal{K}_\nu$ is a modified Bessel function and we have $\theta = (\nu, \sigma, \rho)$ (Matérn).

Estimation

Model of form

$$Y \sim N[X\beta,\ V(\theta)]$$

where the unknown parameters are (β, θ) and $V(\theta)$ is a known function of finite-dimensional parameters θ.

Methods of estimation:

1. Curve fitting to the variogram, based on residuals from OLS regression.

2. Maximum likelihood (MLE)

3. Restricted maximum likelihood (REML)

The Main Prediction Problem

Assume model (5) where the covariances $V,\ w,\ v_0$ are known but β is unknown. The classical formulation of *universal kriging* asks for a predictor $\widehat{Y}_0 = \lambda^T Y$ that minimizes $\sigma_0^2 = E\left\{(Y_0 - \widehat{Y}_0)^2\right\}$ subject to the unbiasedness condition $E\left\{Y_0 - \widehat{Y}_0\right\} = 0$.

The classical solution:

$$\begin{aligned}
\widehat{Y}_0 &= w^T V^{-1} Y + (x_0 - X^T V^{-1} w)^T (X^T V^{-1} X)^{-1} X^T V^{-1} Y, \\
\sigma_0^2 &= v_0 - w^T V^{-1} w + (x_0 - X^T V^{-1} w)^T (X^T V^{-1} X)^{-1} (x_0 - X^T V^{-1} w).
\end{aligned}$$

In traditional geostatistics, the covariances are estimated in a separate estimation step assuming a parametric model such as one of (7)–(9). However, the estimation is then ignored in applying universal kriging. This is potentially a problem, because we would expect the prediction variance to be larger if we took into account the uncertainty in estimating θ.

Bayesian methods provide a potential way round this difficulty, because in a Bayesian analysis we integrate out the predictive density with respect to the posterior density of all the unknown parameters. This is straightforward to implement via MCMC and is starting to be implemented in some widely available packages (GeoR, GeoBugs). However, this raises the question of what are the sampling properties of such procedures. The aim of the present research is to investigate such questions via asymptotic theory.

Bayesian Reformulation of Universal Kriging

Assume the model (5) or equivalently (6). Suppose β (the only unknown parameter, for the moment) has a prior density which is assumed uniform across $\mathcal{R}^p$. The Bayesian predictive density of Y_0 given Y is then

$$p(Y_0 \mid Y) = \frac{\int f(Y^* \mid \beta) d\beta}{\int f(Y \mid \beta) d\beta}. \tag{10}$$

This may be rewritten in the form

$$p(Y_0 \mid Y) = (2\pi)^{-1/2} \frac{|V^*|^{-1/2} |X^{*T} V^{*-1} X^*|^{-1/2} e^{-G^{*2}/2}}{|V|^{-1/2} |X^T V^{-1} X|^{-1/2} e^{-G^2/2}}. \tag{11}$$

where $G^2 = Y^T \{V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1}\} Y$ is the generalized residual sum of squares for the vector Y and G^{*2} is the same thing for the vector Y^*.

However, with some algebraic manipulation we can show,

$$\frac{|V^*||X^{*T}V^{*-1}X^*|}{|V||X^TV^{-1}X|} = \sigma_0^2, \tag{12}$$

$$G^{*2} = G^2 + \frac{(Y_0 - \widehat{Y}_0)^2}{\sigma_0^2}. \tag{13}$$

The Bayesian predictive density then becomes

$$p(Y_0 \mid Y) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2}\left(\frac{Y_0 - \widehat{Y}_0}{\sigma_0}\right)^2\right\} \tag{14}$$

Thus, in the case where β is the only unknown, we have rederived universal kriging as a Bayesian predictor.

However, because of the usual (frequentist) derivation of universal kriging, it follows that in this case, Bayesian procedures have exact frequentist properties, e.g. a Bayesian 95% prediction interval for Y_0 will indeed cover the true Y_0 in 95% of repeated samples.

Now consider the case where θ is also unknown. We assume θ has a prior density $\pi(\theta)$, independent of β.

The Bayesian predictive density of Y_0 given Y is now

$$\begin{aligned} p(Y_0 \mid Y) &= \frac{\int\int f(Y^* \mid \beta, \theta)\pi(\theta)d\beta d\theta}{\int\int f(Y \mid \beta, \theta)\pi(\theta)d\beta d\theta} \\ &= (2\pi)^{-1/2}\frac{\int |V^*|^{-1/2}|X^{*T}V^{*-1}X^*|^{-1/2}e^{-G^{*2}/2}\pi(\theta)d\theta}{\int |V|^{-1/2}|X^TV^{-1}X|^{-1/2}e^{-G^2/2}\pi(\theta)d\theta}. \end{aligned}$$

Using (12) and (13), $p(Y_0 \mid Y)$ may be rewritten

$$\frac{\int |V|^{-1/2}|X^TV^{-1}X|^{-1/2}e^{-G^2/2}(2\pi\sigma_0^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{Y_0-\hat{Y}_0}{\sigma_0}\right)^2\right\}\pi(\theta)d\theta}{\int |V|^{-1/2}|X^TV^{-1}X|^{-1/2}e^{-G^2/2}\cdot\pi(\theta)d\theta}. \tag{15}$$

(15) is of the form

$$p(Y_0 \mid Y) = \tilde{\psi} = \frac{\int e^{\ell_n(\theta)} \psi(\theta) \pi(\theta) d\theta}{\int e^{\ell_n(\theta)} \pi(\theta) d\theta} \tag{16}$$

where $e^{\ell_n(\theta)}$ is the *restricted likelihood* of θ given the data Y and $\psi(\theta) = (2\pi\sigma_0^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{Y_0 - \hat{Y}_0}{\sigma_0}\right)^2\right\}$ is the predictive density we are trying to evaluate, written as a function of θ.

The function $e^{\ell_n(\theta)}$ may be alternatively derived from purely frequentist considerations as the likelihood of a set of orthogonal contrasts in the original X space. However, it has been known since Harville (1974) that this is equivalent to the Bayesian derivation as an integrated likelihood with respect to β. The best regarded (frequentist) estimator of θ is the so-called restricted maximum likelihood or REML estimator $\widehat{\theta}$ which maximizes $\ell_n(\theta)$.

Solution of (16): Use *Laplace approximation*.

First, some notation. Let

$$\begin{aligned} U_i &= \frac{\partial \ell_n(\theta)}{\partial \theta^i}, \\ U_{ij} &= \frac{\partial^2 \ell_n(\theta)}{\partial \theta^i \partial \theta^j}, \\ U_{ijk} &= \frac{\partial^3 \ell_n(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k}, \end{aligned}$$

where θ^i, θ^j... denote components of the vector θ.

Suppose inverse of $\{U_{ij}\}$ matrix has entries $\{U^{ij}\}$.

We shall introduce other quantities such as $Q(\theta)$ and $\psi(\theta)$ that are functions of θ, and where needed, we use suffixes to denote partial differentiation, for example $Q_i = \partial Q/\partial\theta^i$, $\psi_{ij} = \partial^2\psi/\partial\theta^i\partial\theta^j$. All these quantities are evaluated at the true θ unless denoted otherwise. The maximum likelihood estimator (MLE) is denoted $\widehat{\theta}$ with components $\widehat{\theta}^i$. The MLE of ψ is $\widehat{\psi} = \psi(\widehat{\theta})$. Any expression with a hat on it, such as $\widehat{U}_{ijk}$, means that it is to be evaluated at the MLE $\widehat{\theta}$ rather than the true value θ.

Using *summation convention*, define

$$\mathcal{D} = \frac{1}{2} U_{ijk} U^{ik} U^{j\ell} \psi_\ell - \frac{1}{2} (\psi_{ij} + 2\psi_i Q_j) U^{ij} \tag{17}$$

and let $\widehat{\mathcal{D}}$ denote the same expression where all terms have hats. With these conventions, an application of Laplace's integral formula leads to

$$\tilde{\psi} = \hat{\psi} + \widehat{\mathcal{D}}, \tag{18}$$

accurate to $O_p(n^{-1})$.

Apply to predictive inference: recast as predictive distribution function (rather than density) so

$$\psi(y; Y, \theta) = \Phi\left(\frac{y - \lambda^T Y}{\sigma_0}\right)$$

where $\widehat{Y}_0 = \lambda^T Y$ and σ_0^2 are the point prediction and MSPE under universal kriging.

$$
\begin{aligned}
\psi'(y;Y,\theta) &= \frac{1}{\sigma_0}\Phi'\left[\frac{y-\lambda^T Y}{\sigma_0}\right], \\
\psi''(y;Y,\theta) &= \frac{1}{\sigma_0^2}\Phi''\left[\frac{y-\lambda^T Y}{\sigma_0}\right], \\
\psi_i(y;Y,\theta) &= \frac{\partial}{\partial\theta^i}\left\{\frac{y-\lambda^T Y}{\sigma_0}\right\}\Phi'\left[\frac{y-\lambda^T Y}{\sigma_0}\right], \\
\psi_{ij}(y;Y,\theta) &= \frac{\partial^2}{\partial\theta^i\partial\theta^j}\left\{\frac{y-\lambda^T Y}{\sigma_0}\right\}\Phi'\left[\frac{y-\lambda^T Y}{\sigma_0}\right] \\
&\quad + \frac{\partial}{\partial\theta^i}\left\{\frac{y-\lambda^T Y}{\sigma_0}\right\}\frac{\partial}{\partial\theta^j}\left\{\frac{y-\lambda^T Y}{\sigma_0}\right\}\cdot \\
&\quad \cdot\ \Phi''\left[\frac{y-\lambda^T Y}{\sigma_0}\right].
\end{aligned}
$$

Define a maximum likelihood or "plug-in" formula by $\widehat{\psi}(y;Y) = \psi(y;Y,\widehat{\theta})$. The Bayesian predictor, $\tilde{\psi}(y;Y)$, is defined by (16) in which $\psi(\theta)$ is replaced by $\psi(y;Y,\theta)$. Equations (17) and (18) so far give an approximation to $\tilde{\psi}(y;Y)$, accurate to $O_p(n^{-1})$.

Now let us consider the corresponding quantile problem. Suppose we are interested in determining the value y_P for which the event $Y_0 \leq y$ has conditional probability P, given Y. Here P is a fixed constant between 0 and 1. We can define two natural estimators by inverting the maximum likelihood and Bayesian predictive distribution functions. Specifically, $\widehat{y}_P$ satisfies $\widehat{\psi}(\widehat{y}_P;Y) = P$ and $\tilde{y}_P$ satisfies $\tilde{\psi}(\tilde{y}_P;Y) = P$. In the case $P_1 = \alpha/2,\ P_2 = 1-\alpha/2$, the intervals $(\widehat{y}_{P_1}, \widehat{y}_{P_2})$ and $(\tilde{y}_{P_1}, \tilde{y}_{P_2})$ are natural candidates for a $100(1-\alpha)\%$ *prediction interval* for y_0. One of the questions of interest is what is the true coverage probability of either of these intervals in repeated sampling — it would be ideal if the coverage probability was exactly $1-\alpha$.

For $\widehat{y}_P$, we simply substitute maximum likelihood estimator for unknown parameters throughout, and obtain the exact solution

$$\widehat{y}_P = \widehat{\lambda}^T Y + \widehat{\sigma}_0 \Phi^{-1}(P) \tag{19}$$

where $\Phi^{-1}(\cdot)$ is the inverse standard normal distribution function.

For $\widetilde{y}_P$, a Taylor expansion based on (18) suggests the approximation

$$\widetilde{y}_P = \widehat{y}_P - \frac{\widehat{\mathcal{D}}(\widehat{y}_P)}{\widehat{\psi}'(\widehat{y}_P; Y)}. \tag{20}$$

Here $\widehat{\mathcal{D}}(\widehat{y}_P)$ is defined by (17) where we evaluate all functions of θ at $\theta = \widehat{\theta}$ and, in addition, evaluate the function $\psi(y; Y, \theta)$ at $y = \widehat{y}_P$; we also define $\widehat{\psi}'(y; Y) = \psi'(y; Y, \widehat{\theta})$. Since $\widehat{\mathcal{D}} = O_p(n^{-1})$, it follows that (20) is also accurate to $O_p(n^{-1})$.

Future Work

1. Investigate the computational properties of this procedure as an alternative to MCMC.

2. Bias in coverage probability — how far do the true coverage probabilities of either the likelihood or Bayesian prediction intervals differ from their nominal levels?

3. Design of a network: a reasonable criterion for design of a monitoring network might be to do it to minimize the expected length of a Bayesian prediction interval (of some quantity of particular interest, such as the statewide average of particular matter). The present approach allows for a combination of *estimative* and *predictive* criteria which is in line with recent research in this field (Zhu, 2002, U. Chicago thesis).