SPATIAL INTERPOLATION OF AIRBORNE PARTICULATES AND ITS APPLICATION TO EPIDEMIOLOGICAL STUDIES

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

S4 Colloquium: Spatial Structures in the Social Sciences
Brown University
October 8 2004

R.L. Smith, S. Kolenikov and L.H. Cox (2003), Spatio-temporal modeling of PM2.5 data with missing values. *J. Geophys. Res., 108* (D24), 9004, doi:10.1029/2002JD002914, 2003.

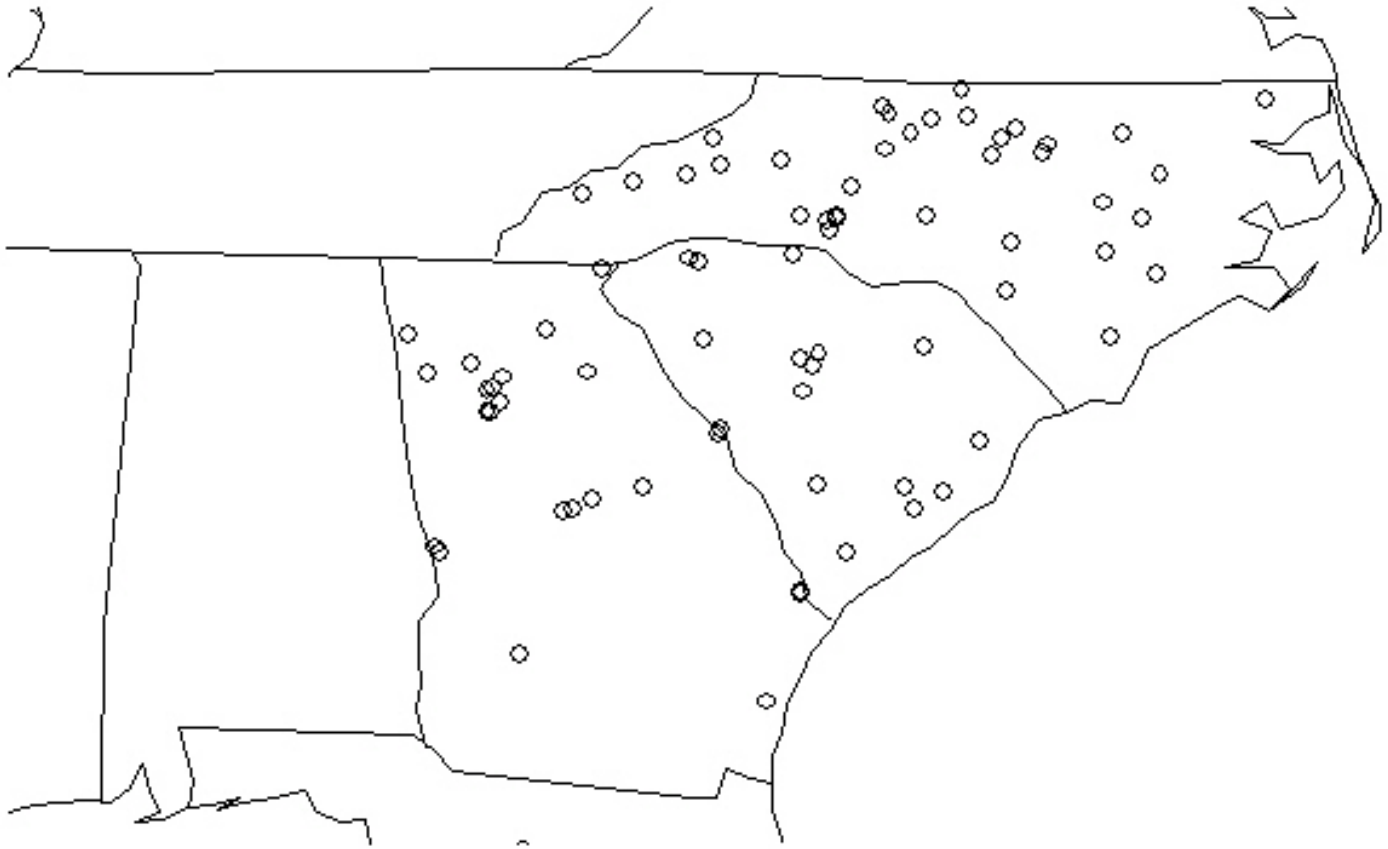
http://www.unc.edu/depts/statistics/postscript/rs/Smith-JGR-2003.pdf

Background

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).

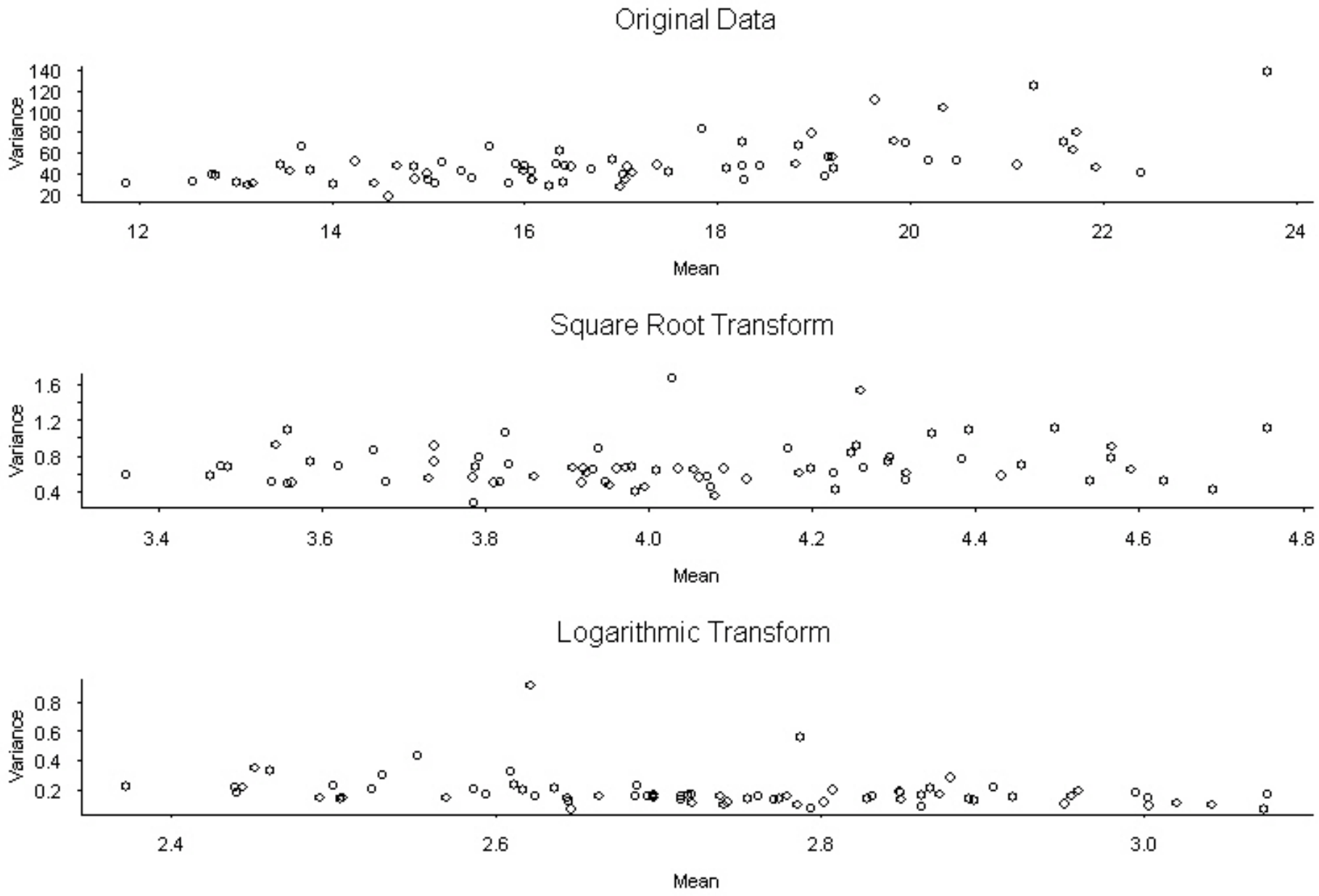
Map of 74 Stations



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

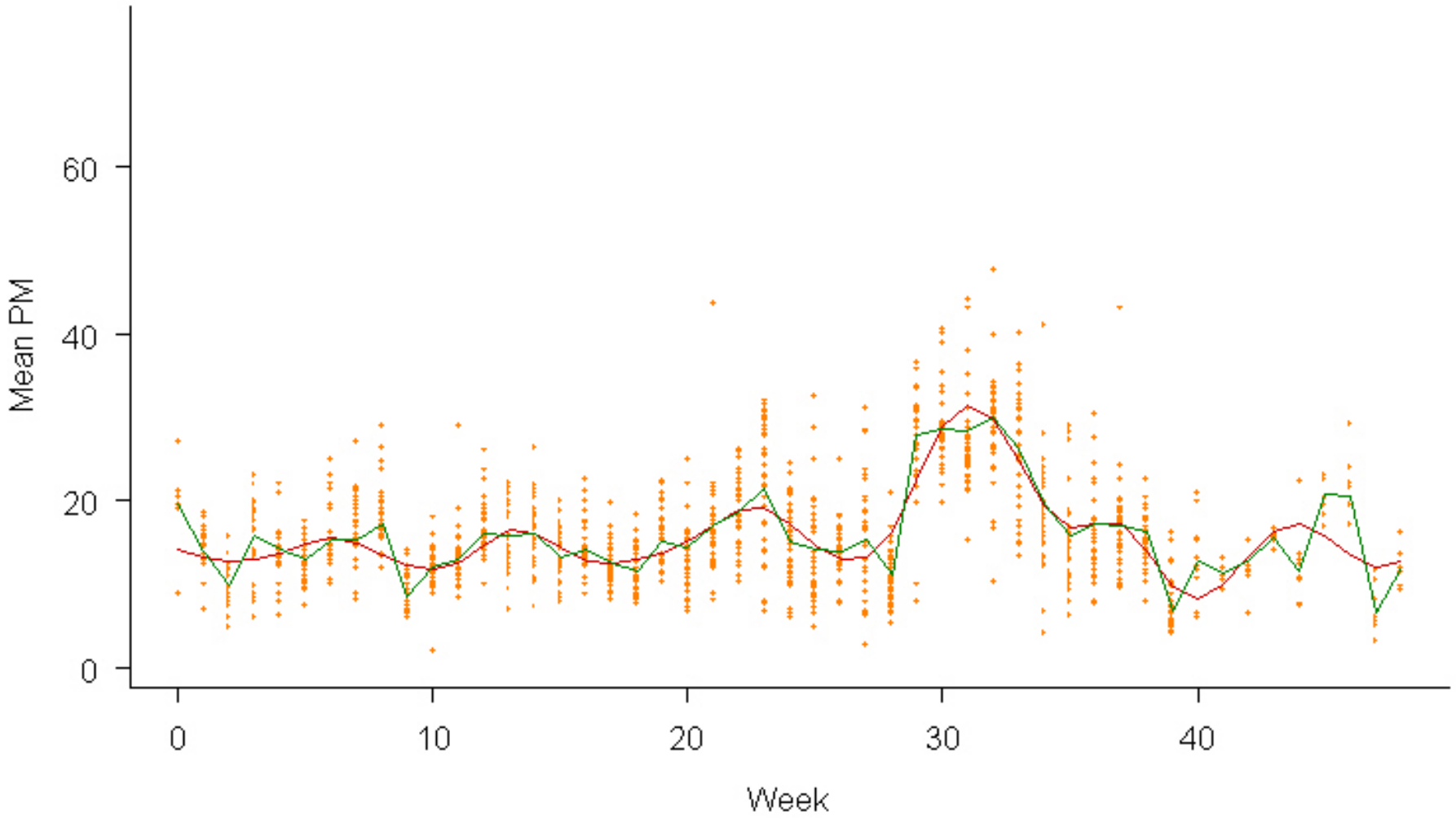


The time trend

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.

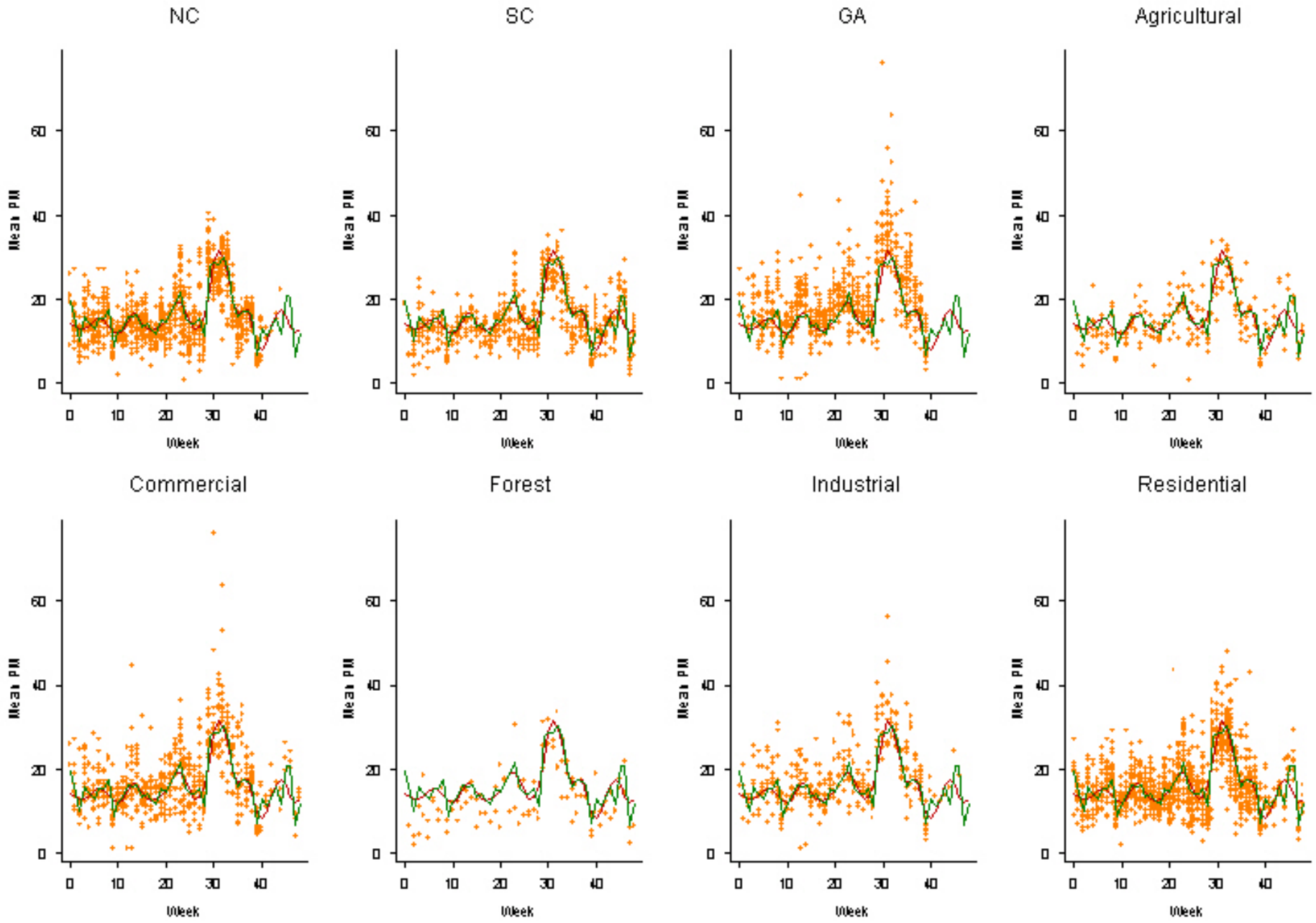
Plotting these two curves with the original data superimposed supports the notion that the entire data set is following roughly the same time trend.

Time Trend Fits to Entire Data Set



If we plot various subsets of the data (against the same weekly trends calculated from the whole data), they support the contention that the same overall time trend applies to all the data, though it's clear that some subsets are systematically higher or lower than others.

Overall Time Trends 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.

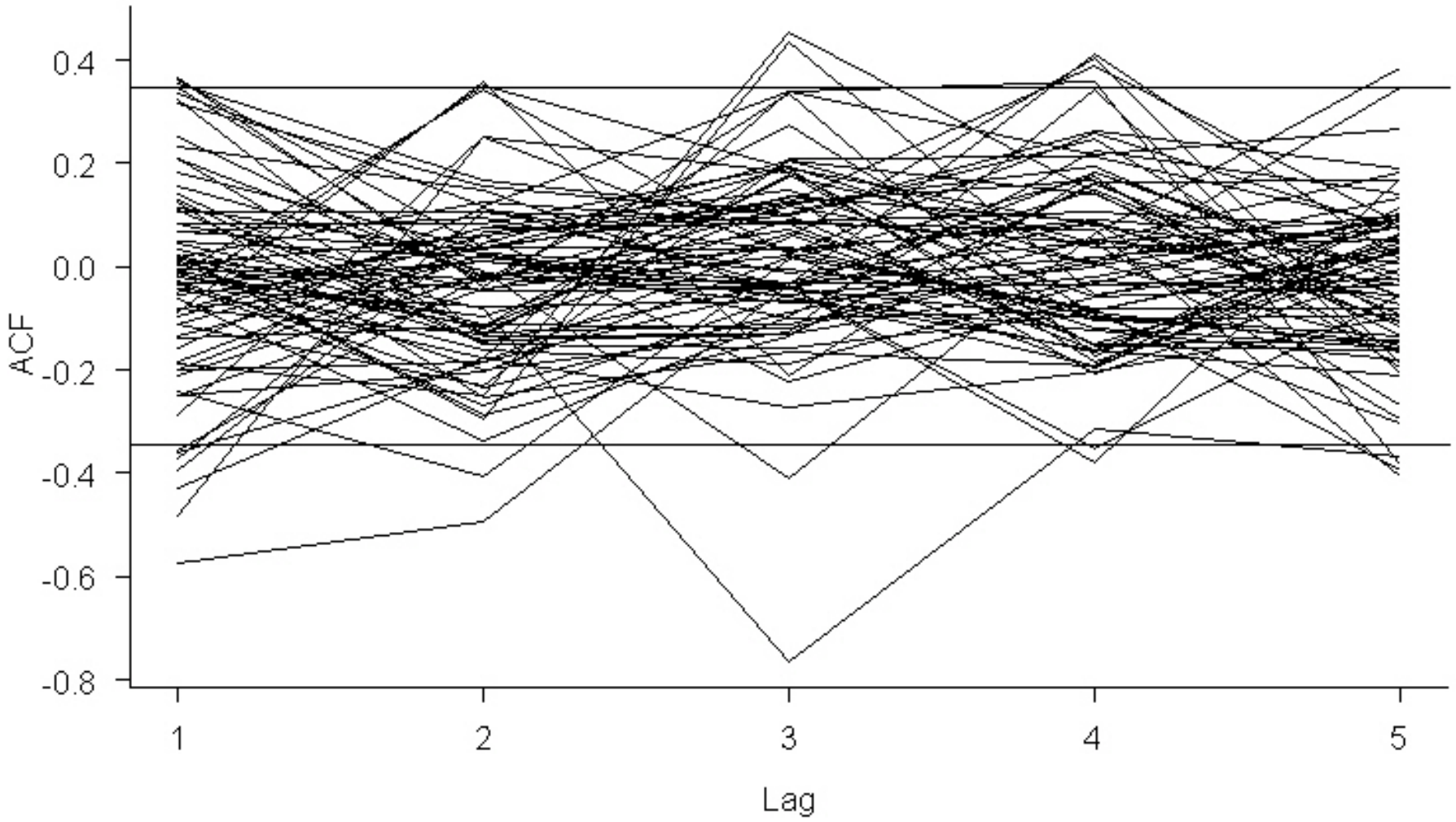
So far we have ignored temporal and spatial correlations among the η_{xt}, but we consider these next.

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

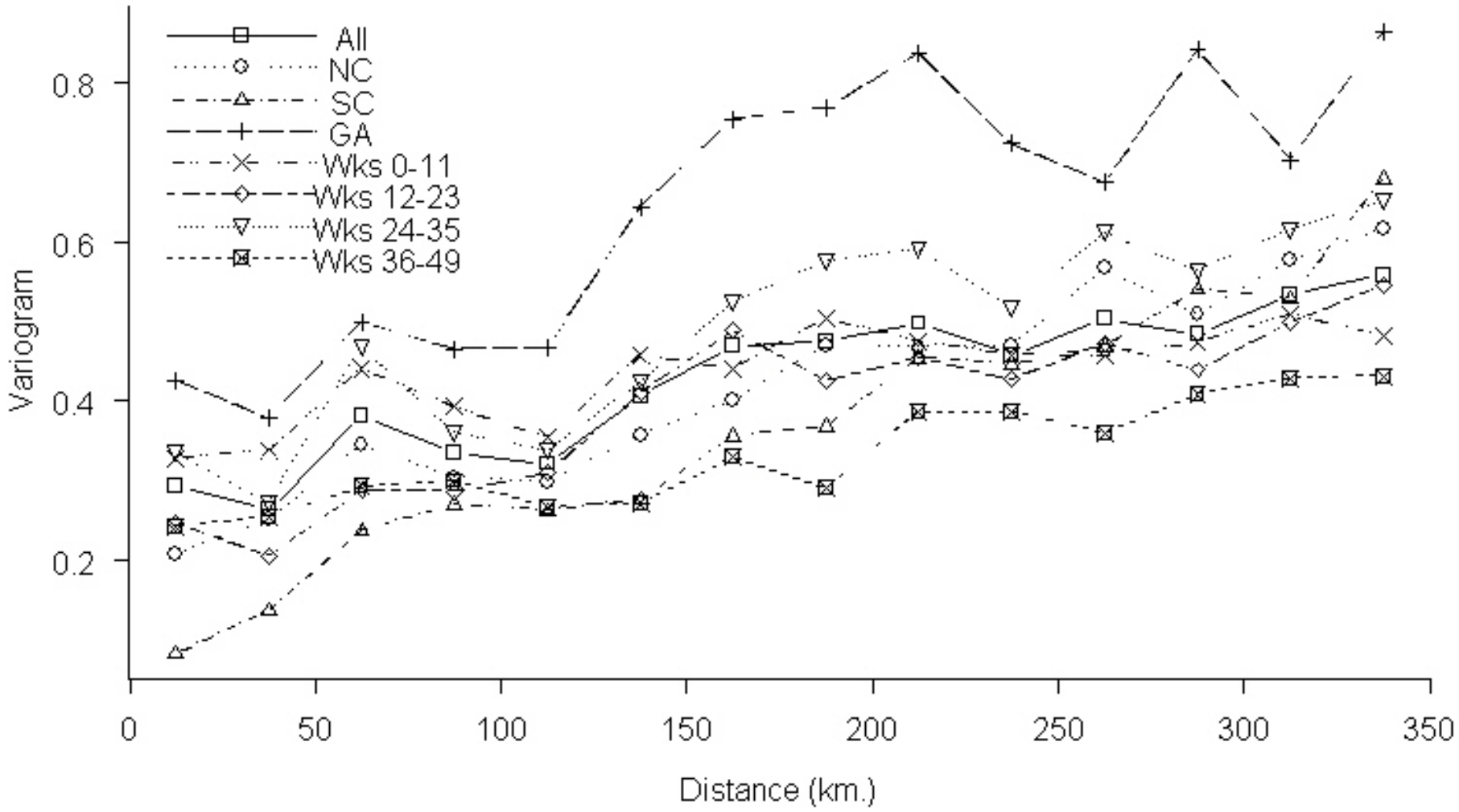


We show variograms of residuals from simple linear regression, where a number of subsets of the data (classified by state and also by season) have been identified to look for comparability of the estimated variogram among different subsets of data. Key points are

- 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

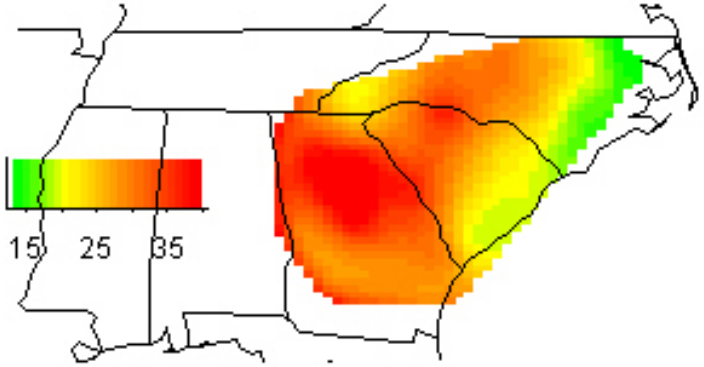
The model (3) was fitted to the data values from which each weekly mean had been subtracted. The residuals η^*_{xt} were assumed independent at different time points but with spatial covariances given by (4) with (2). As an example of the results, the maximum likelihood estimate of the parameter θ_2 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.

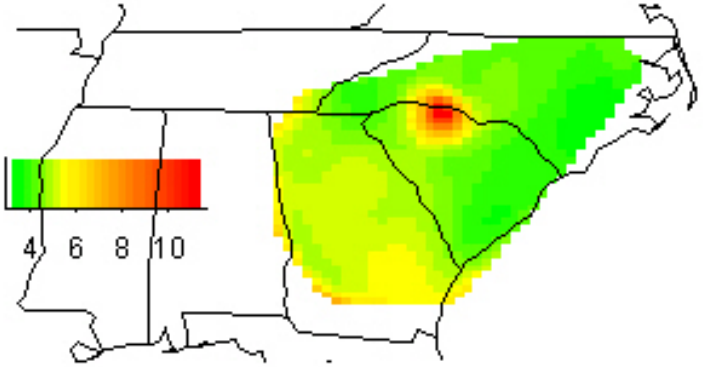
We show the predicted surface and RMSPE for week 33 (the week with highest average $PM_{2.5}$) and overall for the annual mean. We also show the estimated probability that any particular location exceeds the 15 $\mu g/m^3$ annual mean standard. These maps are based on kriging the residuals η^*_{xt} in (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 calculation of exceedance probabilities assumed that (on a square root scale) the difference between the predicted and true values has a normal distribution with standard deviation given by the RMSPE.

Predicted PM2.5 Surfaces and RMSPEs

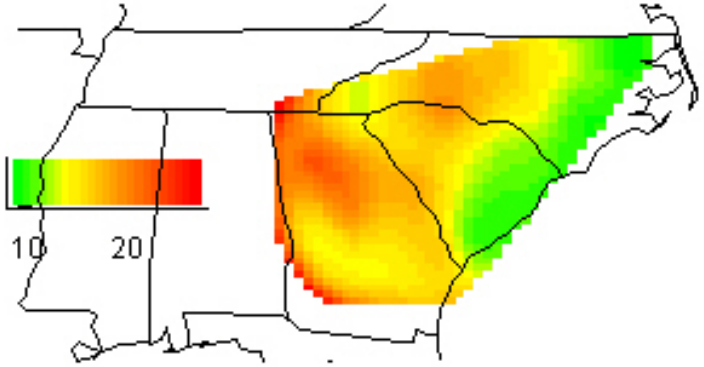
Predicted Surface for Week 33



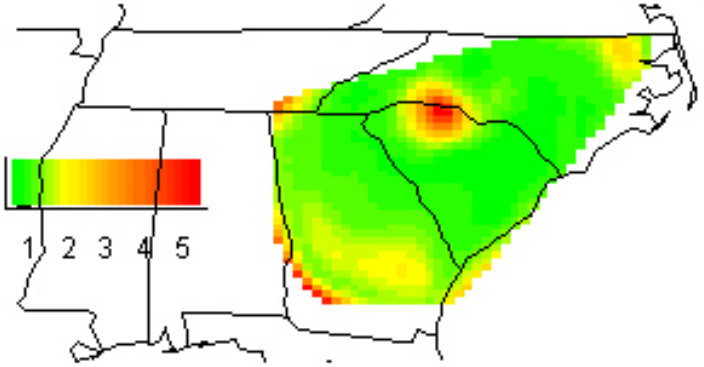
RMSPE for Week 33



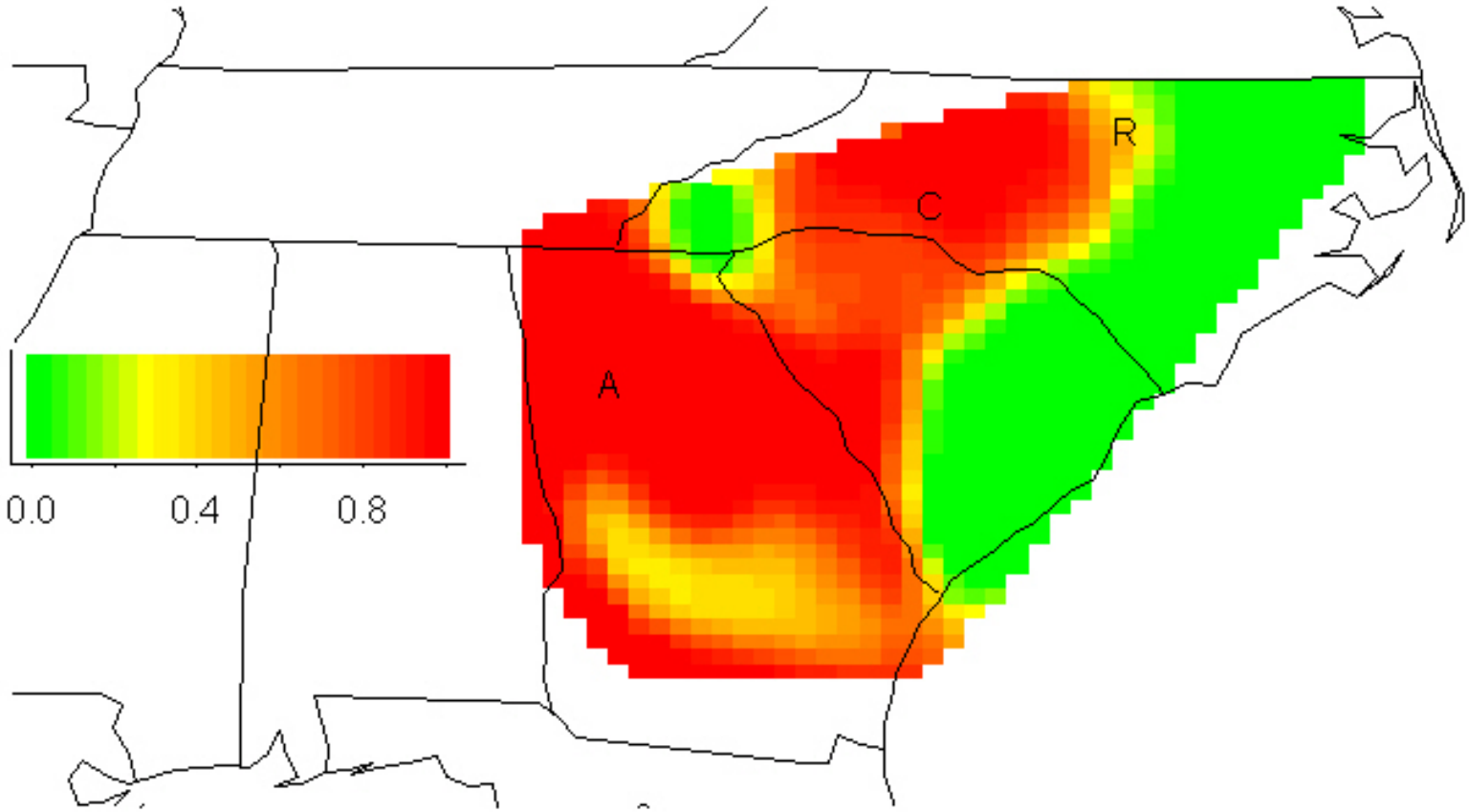
Predicted Surface for Annual Average



RMSPE for Annual Average



Probability of Exceeding Standard



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.

Ongoing and Proposed Future Work

Part of a project *The Environmental Epidemiology of Arrhythmogenesis in WHI* headed by Eric Whitsel (UNC Department of Epidemiology)

The Background: There have been many studies showing an association between increased levels of atmospheric particles (PM_{10} and $PM_{2.5}$) and adverse health effects such as mortality in the elderly population. A specific causal mechanism has not been identified, but there is a particularly strong association with deaths due to cardio-vascular failure. This suggests the need for more studies focussing specifically on measures of cardio-vascular health.

The Women's Health Initiative is a prospective study of 68,133 post-menopausal women. Among other data, the experimenters have recorded ECGs and used them to calculate various measures of heart-rate variability, such as SDNN and RMSSD.

After a log transformation, these have been regressed on various covariates, including age of the subject, sin-cosine terms representing seasonality and $PM_{2.5}$. The $PM_{2.5}$ estimates are based on kriging calculations based on data from available monitors, to estimate the $PM_{2.5}$ level at the subject's home address.

We have become particularly interested in the effect of interpolation error. This will be combined with other aspects of measurement error, such as errors in geocoding addresses, or discrepancies between personal and ambient exposure.

A Few Preliminary Results

In a preliminary study based on 757 participants in North Carolina, the decrease in SDNN associated with a 10 μg/m^3 rise in $PM_{2.5}$ was 8.4%.

In 12 perturbations of the same analysis, where the spatially interpolated $PM_{2.5}$ was subjected to a further perturbation to represent the estimated effect of interpolation error, this was reduced to 7.7%.

Similarly, the decrease in RMSSD associated with a 10 μg/m^3 rise in $PM_{2.5}$ was 10.2%. The average estimate in 12 perturbations was 9.3%.

In neither case was the change in parameter estimate statistically significant when compared with the standard errors of those estimates, but if the same regression coefficients appeared in the full study, they would be significant differences.

This implies the need for a more systematic approach to take into account the error of spatial interpolation.

Summary and Conclusions

The main focus of our work has been on a new methods for spatial interpolation of $PM_{2.5}$, taking both temporal and spatial trends into account to reduce the variability in a single day's measurements.

By allowing us to assess the variability of the interpolation either for a single time point or for a time average, we can measure the overall quality of the interpolation and also provide insights into the design of a monitoring network.

Preliminary results suggest that taking interpolation errors into account is important for the assessment of epidemiological effects, and we hope to develop this much further in future work.