

Spatio-Temporal Analysis of the PM_{2.5} field

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

Background

In 1997, the EPA proposed a new particulate matter standard based on $PM_{2.5}$, to take effect alongside the earlier standard for PM_{10} .

- the three-year average of the 98th percentile of $PM_{2.5}$ should not exceed $50 \mu\text{g}/\text{m}^3$,
- the arithmetic mean (over all monitors within a given region) of the three-year average of daily $PM_{2.5}$ levels should not exceed $15 \mu\text{g}/\text{m}^3$.

However, at the time the standard was promulgated, no nationwide network of $PM_{2.5}$ monitors existed.

The present study is based on a subset of 74 monitors set up during 1999, in the three states of North Carolina, South Carolina and Georgia (Fig. 1.)

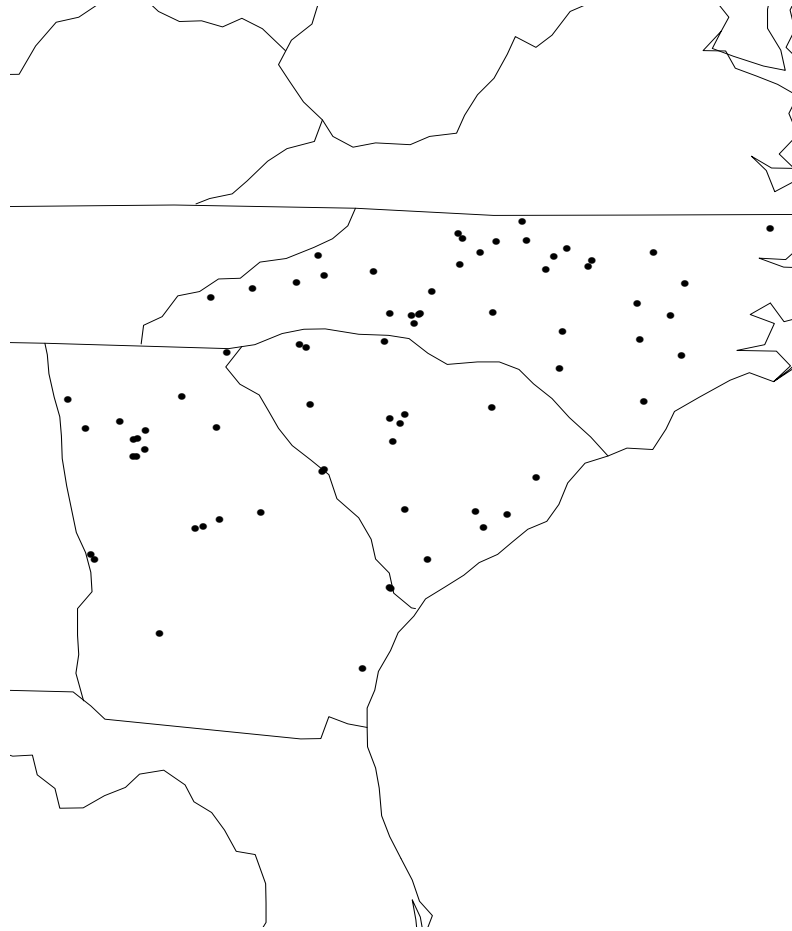


Fig. 1: Map of 74 monitors

We analyzed the data as a spatio-temporal field, with the following features:

- Square root transformation stabilizes the variance (Fig. 2)
- Deterministic effects due to time (“week” variable or B-spline trend), space (thin-plate splines) and other effects such as land use (agricultural, commercial, forest, industrial, residential) (Fig. 3)
- Residuals from simple regression showed little evidence of time autocorrelation (Fig. 4)
- However, variogram plots showed evidence of long-range spatial dependence (Fig. 5)

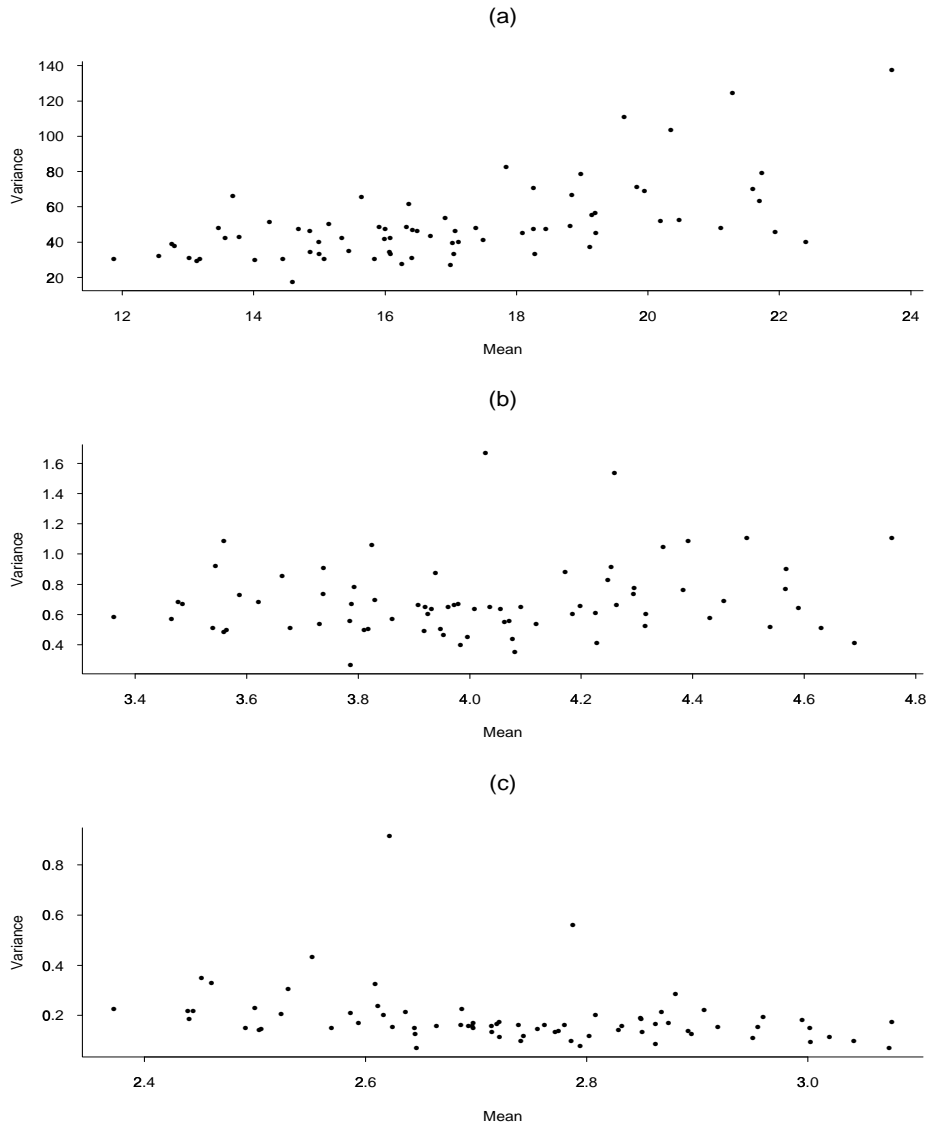


Fig. 2: Variance vs. mean plot for PM_{2.5} values at each of the 74 stations. (a) Original data, untransformed. (b) Square root transform. (c) Logarithmic transform.

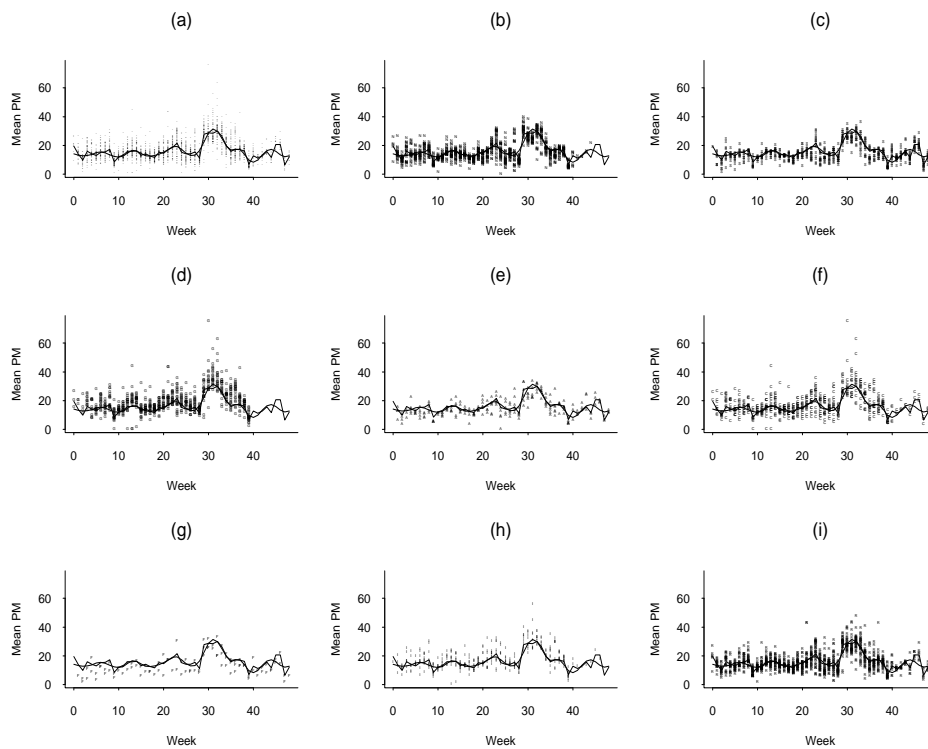


Fig. 3: The comparison of the fitted trend and the raw data for subpopulations. (a) All data combined; (b) North Carolina; (c) South Carolina; (d) Agricultural sites; (e) Commercial sites; (f) Forest sites; (g) Industrial sites; (h) Residential sites. Plotted curves are the *overall* fits of the weekly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

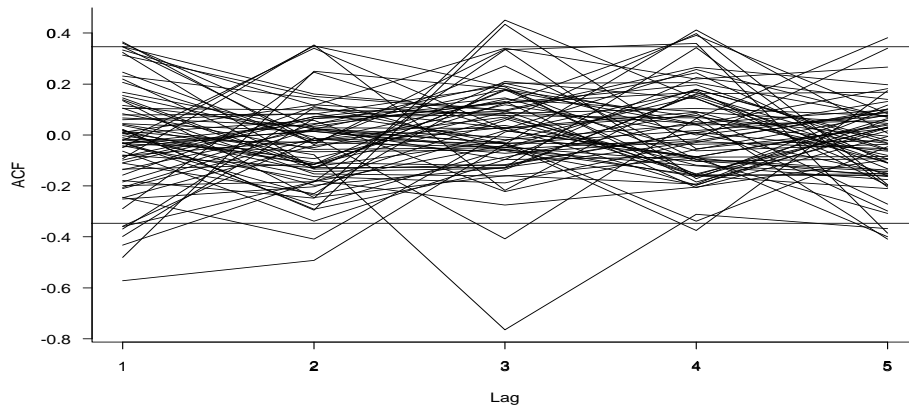


Fig. 4: Time-autocorrelation plots for the 74 stations with approximate 95% confidence bands

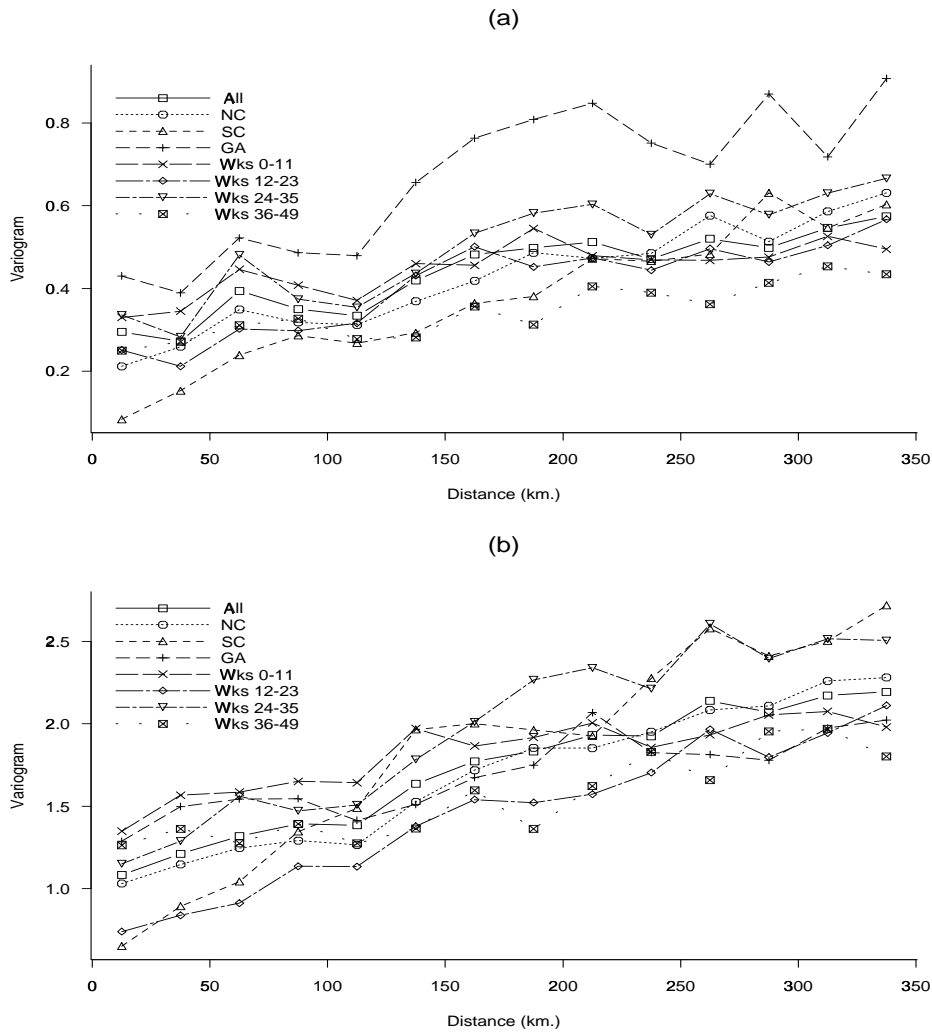


Fig. 5: Variogram plots for residuals after fitting time trend, spatial trend and type effects; all data combined, and separate plots by state and by season. (a) Without standardizing variances. (b) After standardizing the sample variance of residuals at each station to be 1.

Statistical Model

$$y_{st} = \psi_s + \phi_{\ell_s} + \theta_t + \eta_{st} \quad (1)$$

where

y_{st} is $\sqrt{\text{PM}_{2.5}}$ for location s and week t

ψ_s is spatial mean at location s (represented by thin-plate splines)

ϕ_{ℓ_s} is discrete effect for landuse ℓ_s at location s

θ_t is week effect (represented as fixed effect for each week but could also use B-splines etc.)

η_{st} is a spatial random field in s , independent for each t with $\text{Var}\{\eta_{st} - \eta_{s't}\} = \alpha(\theta_1 + h_2^\theta)$ for $h > 0$, where $h = \|s - s'\|$, $\theta_1 > 0$, $0 < \theta_2 < 2$.

Complications:

- Fitting method: uses REML
- Missing data: about 25% data missing. Exact REML estimation involves inverting a different covariance matrix for each week. Alternatives considered based on interpolating missing data, e.g. through a spatial version of the EM algorithm. This is ongoing work.

End result is to reconstruct the annual mean $PM_{2.5}$ field across the whole region, with associated standard errors. Also shown was a direct comparison with a simple linear interpolation of individual monitor means (Fig. 6).

Conclusion: Much of the region, in particular nearly all the state of Georgia, is currently in violation of the new standard.

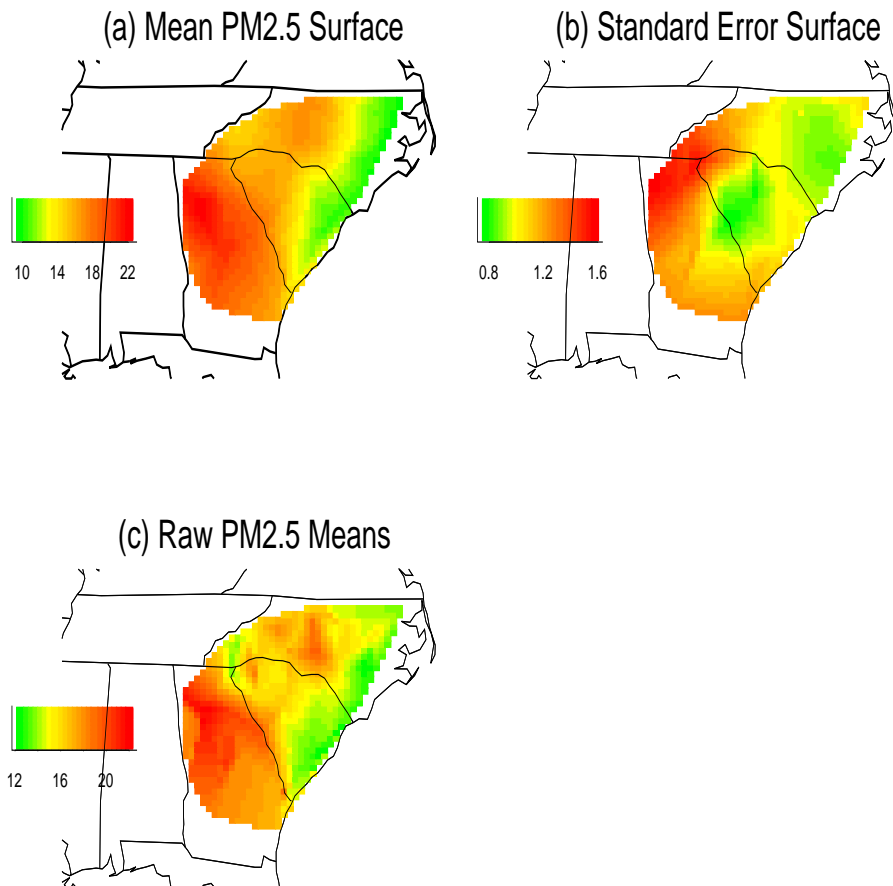


Fig. 6: Plots of the reconstructed surface for overall annual mean PM_{2.5} at each location. (a) Mean surface reconstructed by kriging. (b) Standard error of the surface in (a). (c) Plot of raw data with linear interpolation on a triangulation (S-PLUS “interp” routine).

Issues for SAMSI

1. Statistical methodology questions
2. Applications
3. Interactions with numerical models

Statistical methodology questions

1. Nonstationary spatial models (in two senses)
2. Fully spatial-temporal models (separability?)
3. Methods of estimation — curve fits to the variogram, MLE, REML, BME, fully Bayesian methods
4. Efficient computation, e.g. “good” approximations to the MLE (P. Caragea thesis)
5. Non-Gaussian models

Applications

1. Assessing compliance with air pollution standards
2. Choosing the monitor locations (network design problem)
3. Exposure measures for human health studies (critical question: does kriging provide an adequate basis for assessing measurement error?)
4. Other pollutants, e.g. would the same approach work for monitoring diesel fuel emissions?

Interactions with numerical models

If the objective is to interpolate the $PM_{2.5}$ field, would we be better off using a numerical model? How can we efficiently incorporate data into a numerical model?

1. Use the same techniques but on gridded model output data
2. Use of monitor data to estimate parameters and assess uncertainties in numerical models
3. Combining monitor data and numerical model data into a single prediction approach (data assimilation problem)