

Spatio-Temporal Analysis of the Longitudinal PM_{2.5} Data with Missing Values

Stanislav Kolenikov Richard L. Smith

University of North Carolina, Chapel Hill

Lawrence H. Cox

National Center for Health Statistics

1

Focus on the particular matter

Health statistics evidence: links between the airborne particulate matter and human mortality, especially in elderly ages. Smaller particles penetrate deeper into the lungs \Rightarrow more detrimental effects.

1997 US EPA standard on the particular matter: PM_{2.5} (particulate matter of aerodynamic diameter of 2.5 μm or less) regulations to operate alongside the earlier standard for PM₁₀.

- (a) the 3-year average of the 98th percentile $\leq 50 \mu\text{g}/\text{m}^3$
- (b) the arithmetic mean of the 3-year average of daily PM_{2.5} levels $\leq 15 \mu\text{g}/\text{m}^3$

2

Data

- US EPA 1999 data from some 780 continental US monitors;
- variables: $PM_{2.5}$ concentration; latitude and longitude; the area type; altitude; etc.
- frequency varies from daily to \approx weekly;
- lots of missing data

Restricted / collapsed data set: 74 monitors (NC, SC, GA), 52 weeks — need for spatio-temporal modelling?

3

The trend

Generalized additive model for the trend:

$$y_{it} = \phi_{spatial}(i) + \phi_{temp}(t) + \phi_{area}(i) + \varepsilon_{it}, \quad (1)$$

Each of $\phi(\cdot)$ has its own semiparametric model.

The simplest is ϕ_{area} which is 12 possible shifts according to the possible combinations of { rural, urban, suburban } \times { agricultural, industrial, commercial, residential, forest }

4

Temporal trend

The trend in time is modelled through B-splines

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

$$\phi_{temp}(t) = \alpha_0 + \sum_{k=1}^K \alpha_k \delta_k(t), \quad t \in [0, T], \quad \delta_k(t) = B\left(\frac{K}{T}\left(t - \frac{Tk}{K}\right)\right) \quad (3)$$

Smoothness of the spline: 52 weeks — $K = 20$ knots.

5

Spatial trend

Spatial trend is approximated by the thin plate spline expansion:

$$\phi_{spatial}(\mathbf{z}) = \beta_x x + \beta_y y + \sum_{j=1}^J \beta_j \psi(z_1 - x^{(j)}, z_2 - y^{(j)})$$
$$\psi(x, y) = \frac{r \log r}{16\pi}; \quad r = \sqrt{x^2 + y^2} \quad (4)$$

of nodes J :

- full sample?
- a subsample of sites?
- clustering: a few nodes somewhere in between

6

Spatial covariance

The error process is assumed to be uncorrelated in time and have some spatial covariance:

$$\Omega = \text{Cov}(\epsilon_{ti}, \epsilon_{sj}) = 2\alpha(1 - (1 - \delta_{ij})\kappa) \exp\left(-\left[\frac{d(i, j)}{R}\right]^p\right) \delta_{st}, \quad (5)$$

α : variance; > 0 ;

$d(\cdot, \cdot)$ distance between the sites;

R : the range parameter;

p : the power (shape) parameter (the exponential model for $p = 1$, and Gaussian model for $p = 2$);

κ : the nugget effect

7

EM algorithm: E step

1. calculate the predicted values/residuals:

$$\hat{y}_{it}^{(h)} = \begin{cases} y_{it}, & y_{it} \text{ is observed;} \\ x'_{it}\beta^{(h-1)}, & \text{otherwise} \end{cases} \quad (6)$$

$$e_{it}^{(h)} = y_{it}^{(h)} - x'_{it}\beta^{(h-1)} \quad (7)$$

2. calculate $\Omega^{(h)} = E[ee'| \text{iteration } h]$:

$$E[e_{it}^{(h)} e_{jt}^{(h)} | \text{iteration } h] = \begin{cases} e_{it}^{(h)} e_{jt}^{(h)}, & \text{both } y_{it} \text{ and } y_{jt} \text{ are observed;} \\ \hat{\omega}_{ij}(\theta_{var}), & \text{otherwise} \end{cases} \quad (8)$$

8

EM algorithm: M step

1. Maximizing over the variance parameter space

$$l(\theta_{var}|X, \beta^{(h-1)}, Y^{(h)}, \Omega^{(h)}) \rightarrow \max_{\theta_{var}} \quad (9)$$

2. Maximizing over the trend space

$$\beta^{(h)} = (X'\Omega^{(h)-1}X)^{-1}X'\Omega^{(h)-1}Y^{(h)} \quad (10)$$

Initial values: $\beta^{(0)} =$ OLS estimates based on the available cases.

Results

Estimates of the semivariogram parameters

| Parameter | Point estimate | Reported s.e. | “Corrected” s.e. |
|-----------|----------------|---------------|------------------|
| α | 2.917 | 0.049 | 0.058 |
| R | 2.305 | 0.210 | 0.247 |
| p | 1.444 | 0.139 | 0.164 |
| κ | 0.450 | 0.027 | 0.032 |

The correction is for the overall amount of missing information

| Area type | Rural | Suburban | Urban |
|--------------|--------|----------|--------|
| Agricultural | -1.716 | -1.585 | n/a |
| # sites | 5 | 1 | 0 |
| Commercial | 1.162 | 0.935 | -0.267 |
| # sites | 1 | 10 | 8 |
| Forest | -1.667 | n/a | n/a |
| # sites | 3 | 0 | 0 |
| Industrial | -0.943 | -0.692 | -0.805 |
| # sites | 1 | 7 | 2 |
| Residential | -2.647 | base | 0.021 |
| # sites | 1 | 18 | 17 |

Results

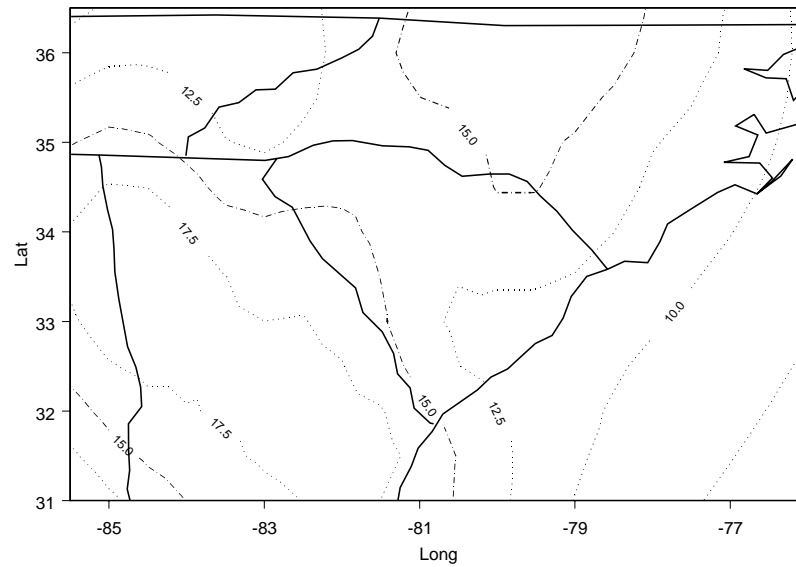


Figure 1: Week 1 of observations

Results

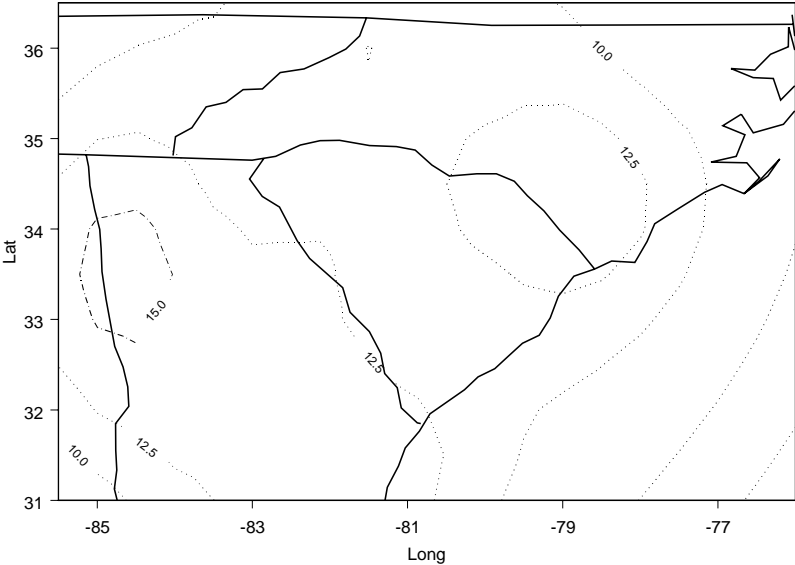


Figure 2: Week 10 of observations

Results

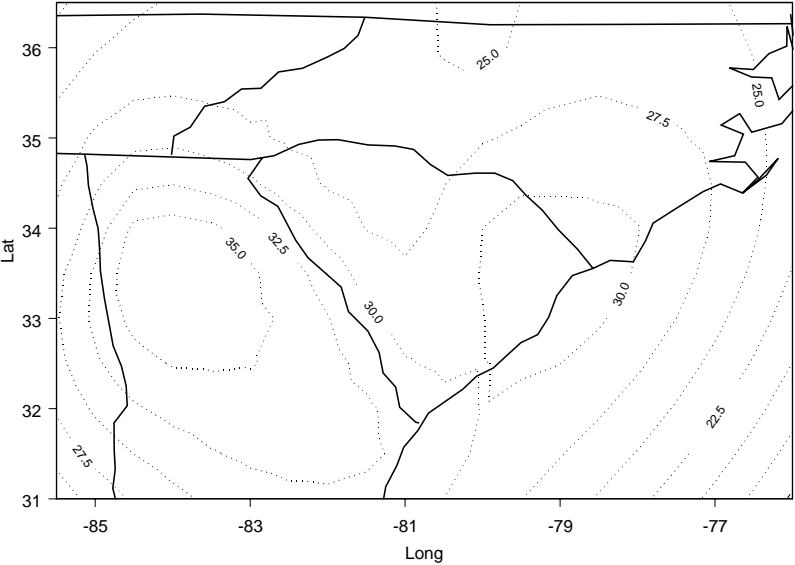


Figure 3: Week 30 of observations

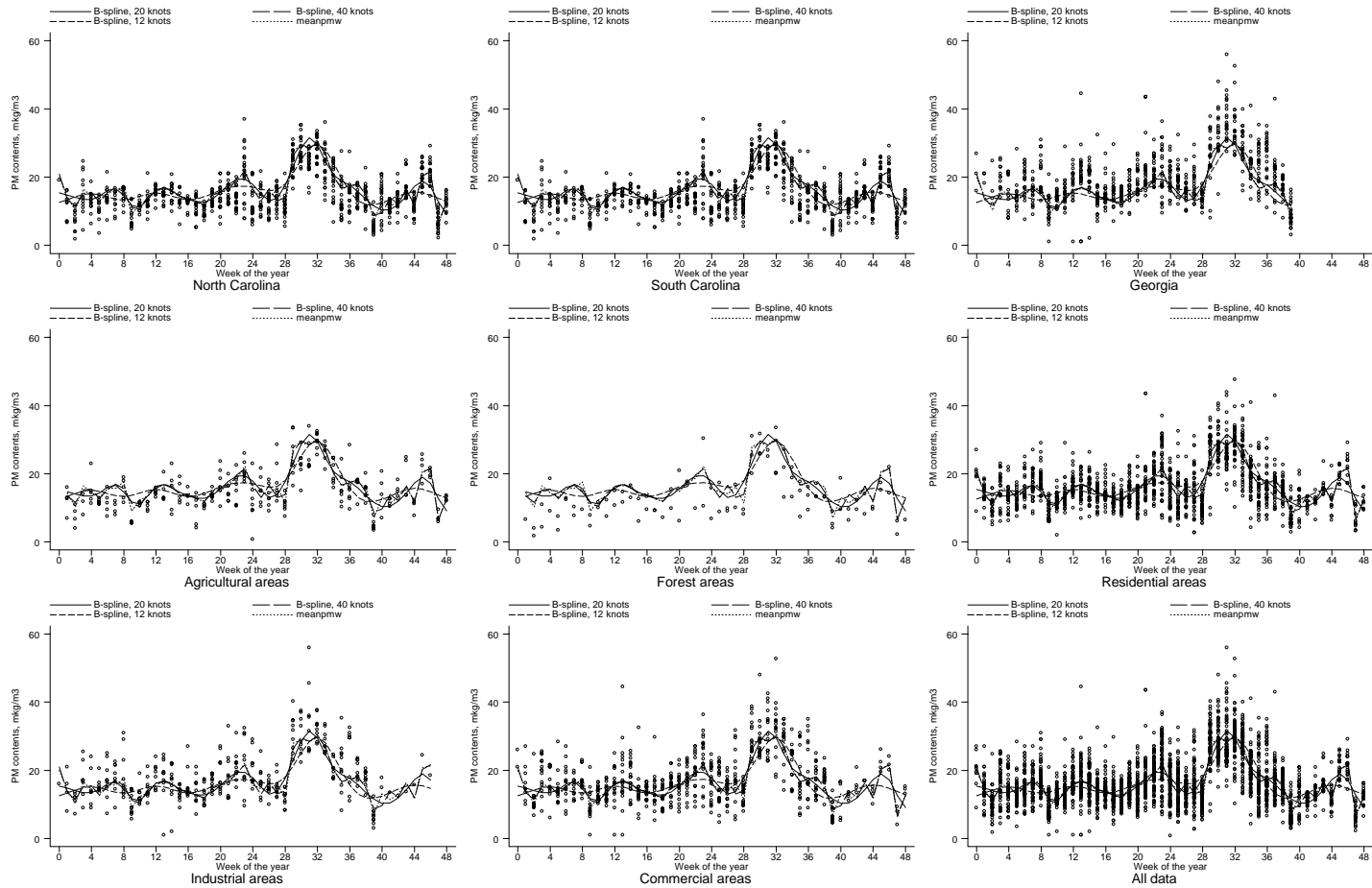


Figure 4: The comparison of the fitted trend and the raw data for subpopulations.