Challenges in Integrating Remote Sensing and Ground Monitoring Data to Estimate $PM_{2.5}$ Concentrations

Chris Paciorek
Department of Biostatistics
Harvard School of Public Health
www.biostat.harvard.edu/˜paciorek

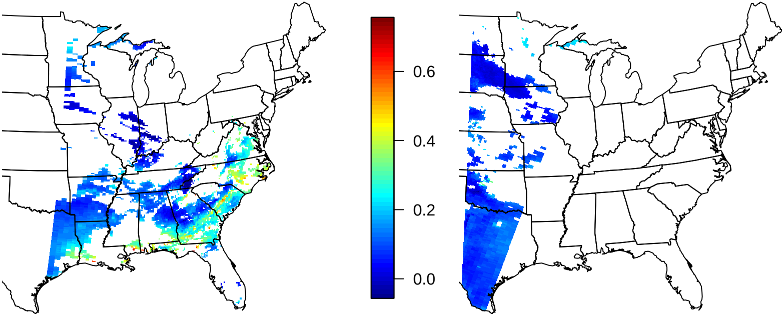
with Yang Liu, HSPH Environmental Health

Research supported by HEI 4746-RFA05-2/06-7

Setting

- To study chronic health effects of PM, estimating spatial heterogeneity in exposure is critical.
- Satellite retrievals of aerosol (AOD) may help, particularly in suburban and rural areas far from monitors.
- Bayesian statistical modeling holds promise for integrating ground measurements of $PM_{2.5}$, satellite-retrieved AOD, GIS and weather information for prediction.
- Output of broader project is intended as a data product for use in various studies of chronic health effects:
  - eastern U.S. (east of 100 W longitude)
  - 4 km grid resolution
  - monthly, 2000-2006

MODIS, July 14, 2004



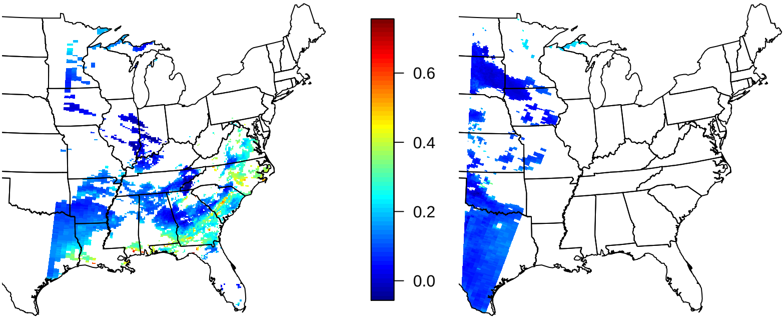
Statistical Challenges

- AOD (aerosol optical depth) measurements estimate total column aerosol.
  - AOD is a noisy and biased proxy for $PM_{2.5}$ with low correlation with $PM_{2.5}$ at high temporal and spatial resolution.
  - Potential spatial correlation in bias of AOD as a proxy for $PM_{2.5}$ poses identifiability issues.
- AOD retrievals are frequently missing.
- Various sources of information are mis-aligned in space and time.
- Full space-time modeling with large remote-sensing datasets is challenging.

Key Questions

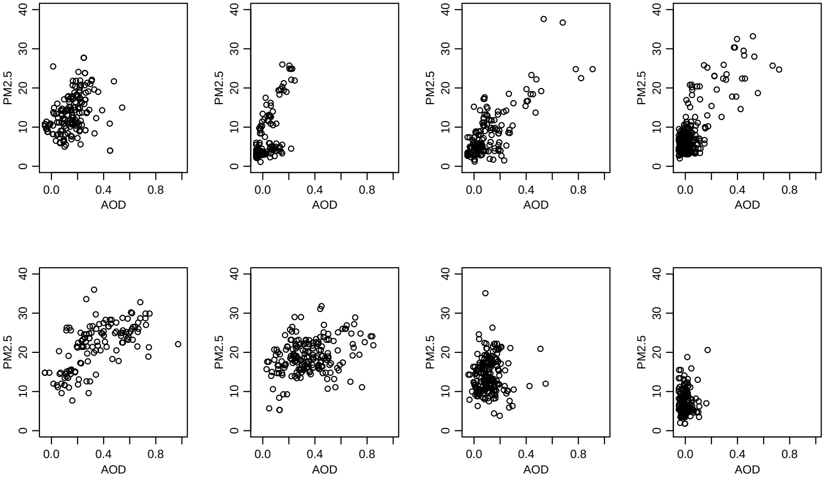
- Should we model spatially-correlated bias in AOD as a proxy for $PM_{2.5}$?
  - What are the implications for identifiability?
- Does including AOD in the model truly improve predictions of $PM_{2.5}$ concentrations conditional on other information?
  - GIS-based covariates
  - Spatial smoothing
  - Meteorological covariates

MODIS, July 14, 2004



Associations of PM and AOD

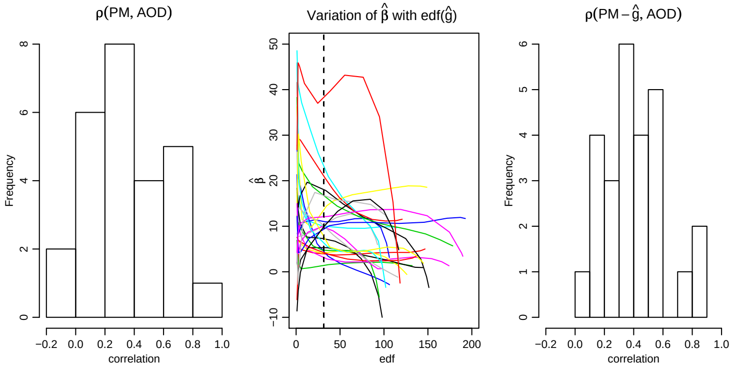
Raw Associations of Spatial Pairs



Associations of PM and AOD

Adjusting pairs for large-scale spatial patterns

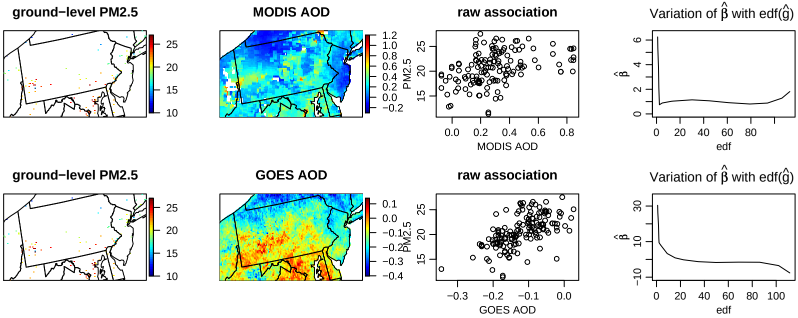
$$PM_{it} = g(s_i) + \beta AOD_{it} + \epsilon_{it}$$



Drawbacks of Daily Data

- Few days have AOD retrievals: 39% for GOES, 12% for MODIS, 3% for MISR
- Even on days with large number of retrievals, strength of association with PM is weak.
- Question: Does averaging in time for chronic exposure estimation help get around this?

Monthly Case Study: July 2004



Modeling Approaches

- Use AOD as data
  - Two likelihoods
  - Issue of relative influence of the two data sources on the latent process
    - No inherent gold standard
  - Model structure for bias of AOD is critical
  - Naturally deals with missing AOD
- Use AOD as a covariate
  - $PM_{2.5}$ treated as gold standard
  - Inherent calibration of AOD and $PM_{2.5}$
  - Requires latent AOD process to avoid having missing covariate values

Model: AOD as data

Likelihood for monthly average data:

$$\text{PM}_i = y_i \sim \mathcal{N}(\mu + P(s(i)) + \sum_k f_k(z_{k,i}), \sigma^2_{y,i})$$

$$\text{AOD}_m = a_m \sim \mathcal{N}(\beta_0 + \phi(s_m) + \beta_1(\mu + P(s_m)), \sigma^2_{a,m})$$

- $f_k(\cdot)$, $k = 1, \ldots, K_f$ are nonparametric regression functions of within-grid cell covariates.
- $\phi(s)$ is spatially-correlated additive bias.

Latent $\text{PM}_{2.5}$ process, $P(s)$, on 4 km grid:

$$P(s_m) = \sum_k h_k(w_k(s_m)) + g(s_m)$$

- $h_k(\cdot)$, $k = 1, \ldots, K_h$ are nonparametric regression functions of grid cell-scale covariates.
- $g(s)$ is Gaussian spatial process.

Smooth term structure

Thin plate spline-based smooth terms, evaluated on the grid:

$$
\begin{aligned}
g &= Zb_g \\
\phi &= Zb_\phi \\
b_g &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_g^2) \\
b_\phi &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\phi^2)
\end{aligned}
$$

- Z is a thin plate spline basis matrix, following Ruppert, Wand, and Carroll (2003), *Semiparametric Regression*.
- $b_{(\cdot)}$ are basis coefficients for the given smooth term.
- Variance components, $\sigma^2_{(\cdot)}$, penalize complexity.
- Regression smooths, $f_k(\cdot)$ and $h_k(\cdot)$, are represented in a similar fashion.

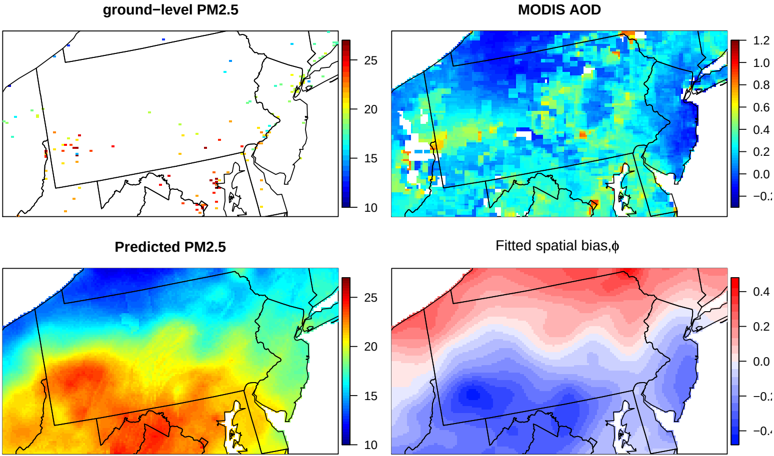
MCMC Implementation

- Because of conditional conjugacy, $\psi = \{b_g, b_\phi, b_f, b_h, \beta_0, \mu\}$ can be sampled from its exact conditional.
- Importance: g, ϕ, f_k, and h_k are all competing to explain the spatial patterns in the data; joint sampling accounts for this dependence.
- Also, there is high dependence between the spline coefficients and their associated variance component (e.g., between b_g and σ_g^2).
  - Therefore, jointly sample: $\{\sigma_g^2, \psi\}$, $\{\sigma_\phi^2, \psi\}$, $\{\{\sigma_{f_k}^2\}_{k=1,\ldots,K_f}, \psi\}$, $\{\{\sigma_{h_k}^2\}_{k=1,\ldots,K_h}, \psi\}$.
  - Joint sampling is done with a Metropolis proposal for the variance component and then sampling ψ from its conditional normal, with a single acceptance decision and a Hastings adjustment needed because we are not sampling from the joint conditional.

Possible models for spatial structure

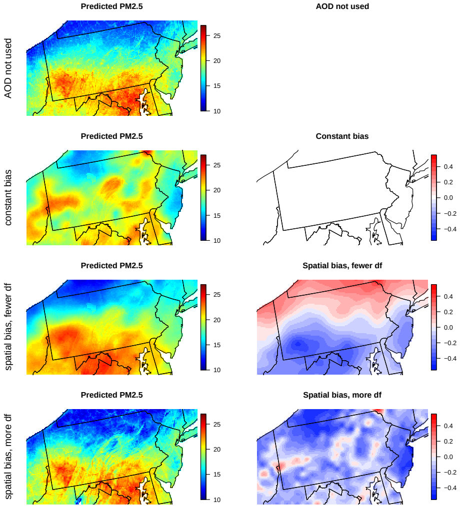
- Knot-based thin plate penalized splines:
  - Efficient for smooth processes (i.e., few knots).
  - For rough processes (many knots), joint Gibbs sampling of coefficients is slow.
- An alternative is the GMRF representation of the thin-plate spline (see Speckman/Sun/Yue, Rue and Held)
  - Sparse precision matrices make this efficient.
  - Again, joint sampling of process values and hyperparameter is critical.
  - Harder to set up joint sampling of GMRF process with other processes in model.

Results



Model discounts AOD hotspots, attributing them to the bias term.

Sensitivity to Assumptions about Bias



Model: AOD as covariate

Likelihood for monthly average $PM_{2.5}$:

$$PM_i = y_i \sim \mathcal{N}(\mu + P(s(i)) + \sum_k f_k(z_{k,i}), \sigma^2_{y,i})$$

Latent $PM_{2.5}$ process, $P(s)$, on 4 km grid:

$$P(s_m) = \beta_1(s_m)A(s_m) + \sum_k h_k(w_k(s_m)) + g(s_m)$$

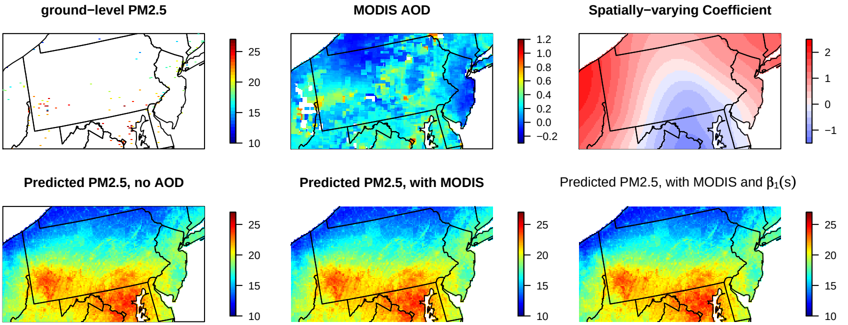
Model: Covariate imputation

This model requires $A(s)$ observed on the full grid, so we need to separately model the AOD process, which we do using a thin-plate spline-based GMRF model:

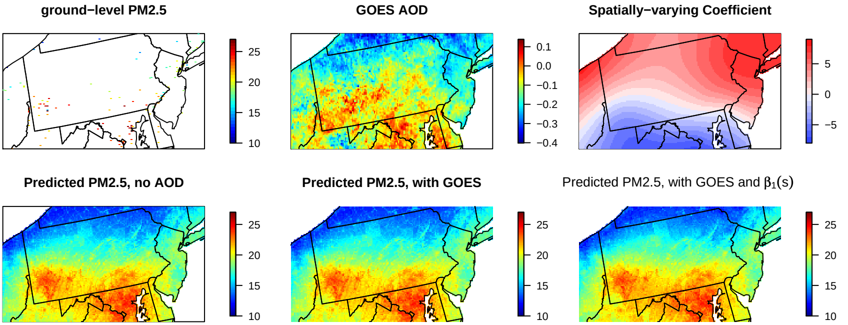
$$a_m \sim \mathcal{N}(\gamma_0 + A(s_m), \sigma^2_{a,m})$$
$$A(s) \sim \text{GMRF}(\tau^2)$$

$\{A(s), \tau^2\}$ sampled jointly (and efficiently) following Rue and Held (2005) (GMRFLib C functions)

Using MODIS as a Covariate



Using GOES as a Covariate



Key Questions

- Should we model spatially-correlated bias in AOD as a proxy for $PM_{2.5}$?
  - Two-likelihood model fit and substantive assessment suggest spatial bias term is critical.
- Does including AOD in the model truly improve predictions of $PM_{2.5}$ concentrations conditional on other information?
  - Raw correlations are weak and do not indicate strong fine-scale association of AOD and PM2.5.
  - Use of AOD as a covariate suggests inclusion provides only limited additional information but current analysis is only initial step.

Statistical Summary

- Spatial modeling allows investigation of key questions in the use of remote sensing data in this arena.
    - Bayesian models allow for a variety of specifications of AOD-PM relationship.
- Data contain an endless array of complications; a major challenge is choosing the key aspects to focus on in the modeling.
- Knot-based spline and carefully-chosen GMRF models provide necessary computational efficiency to handle remote sensing data for single time.
    - Extension to space-time will introduce additional computational complexity.

Next Steps

- Model comparison, including cross-validation, to fully assess usefulness of AOD.
- Full space-time modelling over multiple months.
- Assessment of and accounting for non-ignorable missingness.