Challenges in Integrating Remote Sensing and Ground Monitoring Data When Estimating PM$_{2.5}$ Concentrations for Chronic Health Studies

Chris Paciorek and Yang Liu
Departments of Biostatistics and Environmental Health
Harvard School of Public Health

October 19, 2007

This research was supported by HEI 4746-RFA05-2/06-7.

Setting

- For studies of the chronic health effects of PM, estimating spatial heterogeneity in exposure is critical.

- Satellite retrievals of aerosol (AOD) may help estimate concentrations at locations far from monitors, particularly in suburban and rural areas.

- Bayesian statistical modeling holds promise for integrating ground measurements of PM$_{2.5}$, satellite-retrieved AOD, GIS and weather information to predict monthly PM$_{2.5}$ concentrations at fine spatial resolution.

- Output 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

Statistical Challenges

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

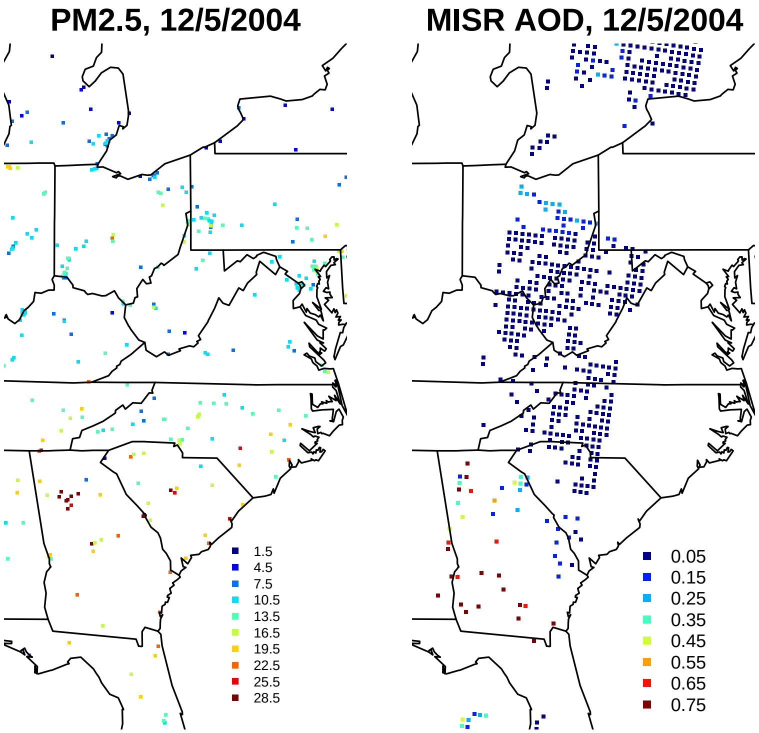
Key Questions

- 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 beyond a model that uses ground data, GIS-based covariates and meteorology?

Data sources

Comparison of PM$_{2.5}$ and AOD: one day



PM2.5, 12/5/2004

MISR AOD, 12/5/2004

PM$_{2.5}$ and covariate information

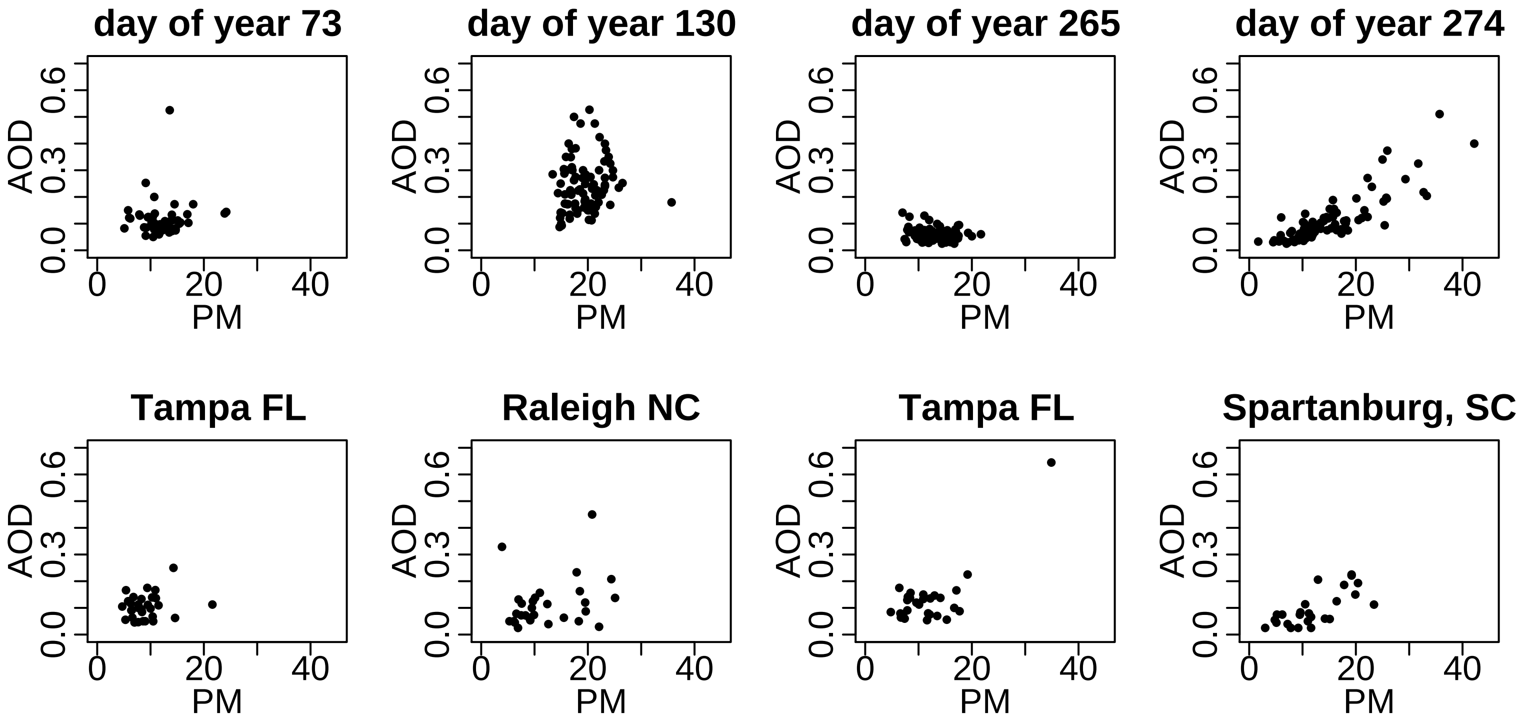
- PM$_{2.5}$ measurements from AQS and IMPROVE: daily average, every 1, 3, or 6 days
- Weather data at 32 km, 3 hour resolution from North American Regional Reanalysis
- GIS-derived information: distance to roads by road class, population density, land use; aggregated to 4 km grid resolution

AOD information

- MISR AOD: once per day; nominal 17.6 km resolution, a given location is measured every 4-7 days, 10:30 am
- MODIS AOD: once per day; nominal 10 km resolution, a given location is measured every 1-2 days, 10:30 am
- GOES AOD: every half hour during daylight; nominal 4 km resolution

Correspondence of AOD and PM$_{2.5}$

Cross-sectional (top row) and longitudinal (bottom row) associations for days and locations with many co-occurring observations:



day of year 73

day of year 130

day of year 265

day of year 274

Tampa FL

Raleigh NC

Tampa FL

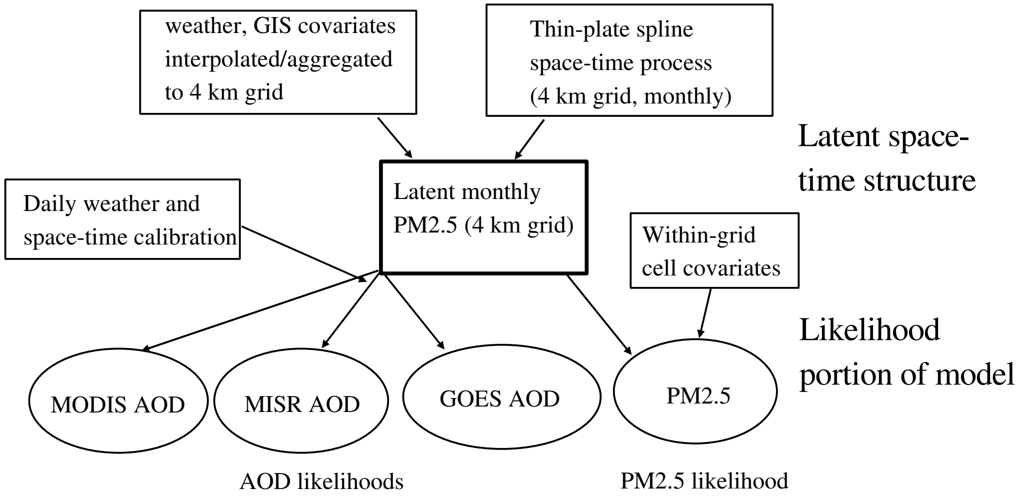
Spartanburg, SC

Plots indicate noisiness of the daily AOD - PM$_{2.5}$ relationship. Calibration with a regression model to adjust for daily meteorology and spatial and temporal variations helps, as does averaging to the month (not shown), but AOD remains a noisy proxy.

Modelling Overview

- Goal is a space-time model that accounts for covariates and incorporates both PM$_{2.5}$ and AOD data. Here we specify separate likelihood terms for each source of data and use a knot-based thin plate spline based representation for computational efficiency.

- AOD is treated as data and not a covariate because of high degree of missingness; for MODIS and MISR, data are aligned to the 4 km grid.


weather, GIS covariates interpolated/aggregated to 4 km grid
Thin-plate spline space-time process (4 km grid, monthly)
Latent space-time structure
Daily weather and space-time calibration
Latent monthly PM2.5 (4 km grid)
Within-grid cell covariates
Likelihood portion of model
MODIS AOD
MISR AOD
GOES AOD
PM2.5
AOD likelihoods
PM2.5 likelihood

Statistical Model for mid-Atlantic Region, July 2004

As a case study to understand the issues involved in jointly modeling PM$_{2.5}$ and AOD, we fit a model for one month for a subregion of the eastern United States using only MODIS AOD.

Likelihood for monthly average PM$_{2.5}$ (y_i) and monthly average AOD (a_m):

$$\begin{aligned} \text{PM}_i = y_i &\sim \mathcal{N}(\mu + P(s(i)) + \sum_k f_k(z_{k,i}), \sigma^2_{y,i}) \\ \text{AOD}_m = a_m &\sim \mathcal{N}(\beta_0 + \beta_1(\mu + P(s_m) + \phi(s_m)), \sigma^2_{a,m}) \end{aligned}$$

$f_k(\cdot),\ k = 1, \ldots, K_f$ are nonparametric regression functions of within-grid cell covariates; β_0 and β_1 are additive and multiplicative bias terms; $\phi(s)$ is spatially-correlated additive bias. Error variances are heteroscedastic to account for differing numbers of daily observations.

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

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

$h_k(\cdot),\ k = 1, \ldots, K_h$ are nonparametric regression functions of grid cell-scale covariates, and $g(s)$ is spatial process defined on the 4 km grid.

Smooth term structure

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

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

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, $\sigma_{(\cdot)}^2$, 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.

This is particularly important because g, ϕ, f_k, and h_k are all competing to explain the spatial patterns in the data; joint sampling accounts for this dependence.

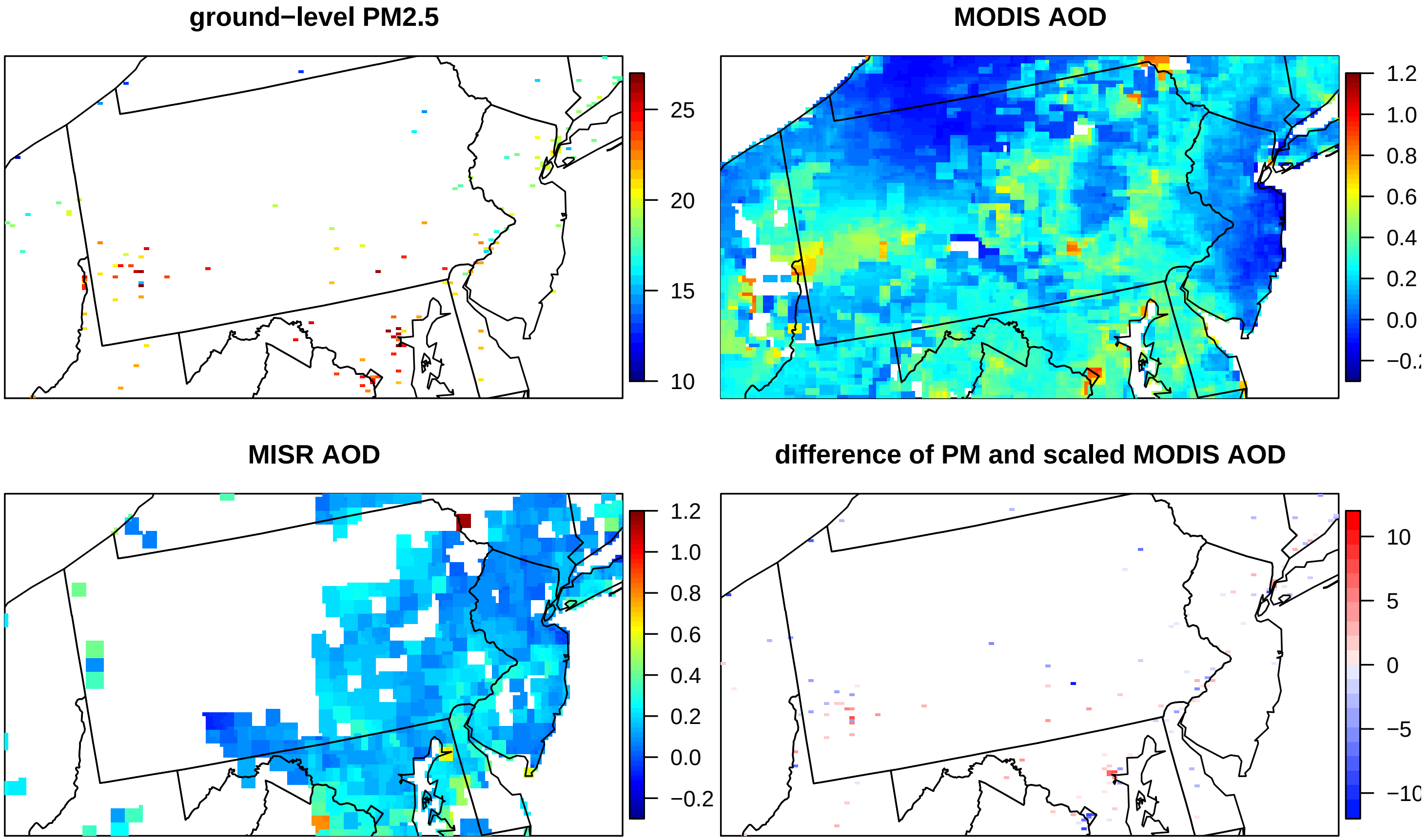
Also, there is high dependence between the spline coefficients and their associated variance component (e.g., between b_g and σ_g^2).

Therefore, good mixing is achieved by jointly sampling each of the following groupings: $\{\sigma_g^2, \psi\}$, $\{\sigma_\phi^2, \psi\}$, $\{\{\sigma_{f_k}^2\}_{k=1,\ldots,K_f}, \psi\}$, $\{\{\sigma_{h_k}^2\}_{k=1,\ldots,K_h}, \psi\}$.

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.

Also jointly sample $\{\beta_1, \psi\}$.

Raw Data for Regional Case Study for July 2004


ground−level PM2.5

MODIS AOD

MISR AOD

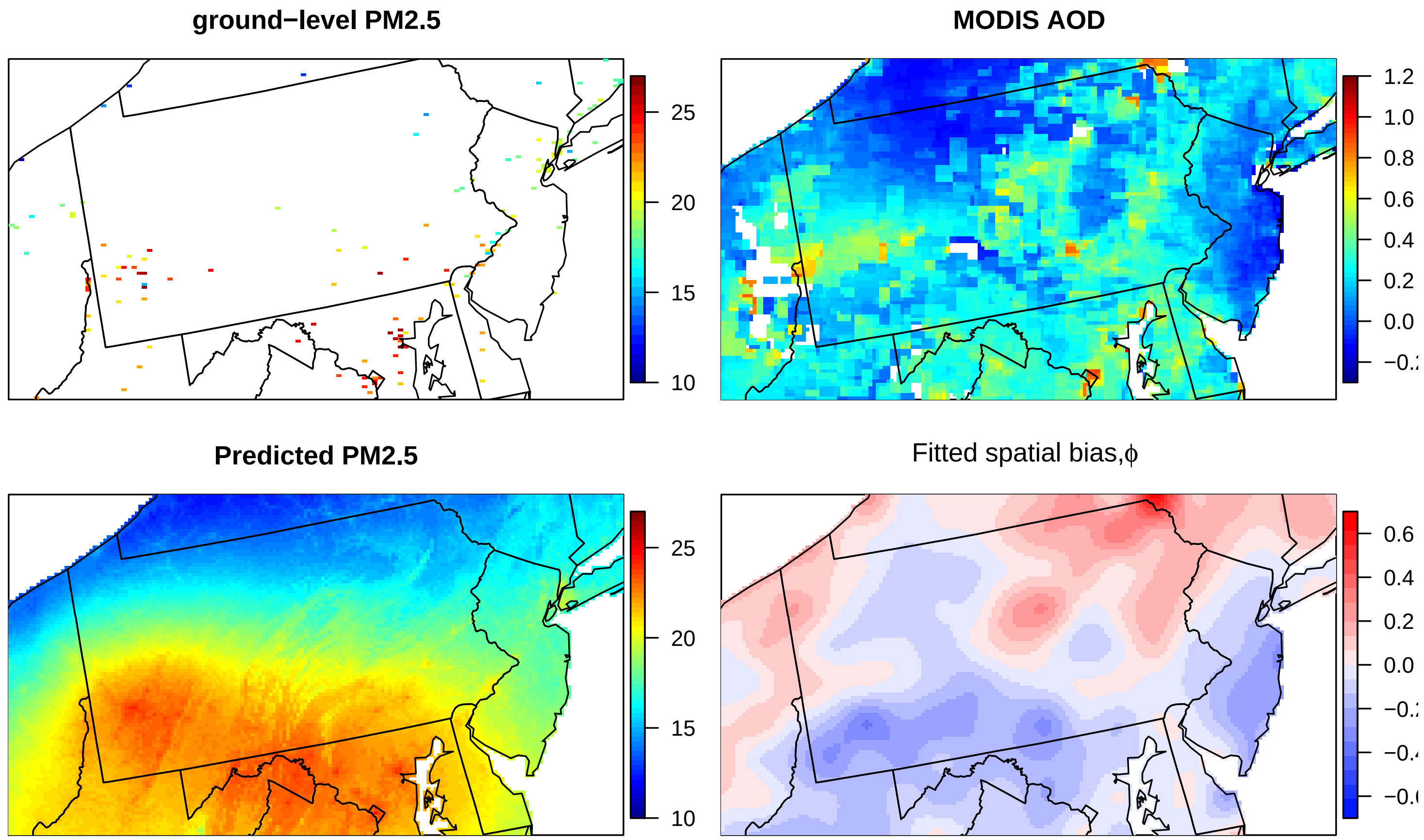
difference of PM and scaled MODIS AOD

Data are monthly averages.

AOD as a proxy for $PM_{2.5}$ appears to be not just noisy but also spatially-structured, with local areas in which AOD seems to either over- or under-predict PM.

Missingness for MISR is a big problem, so MISR is excluded for now.

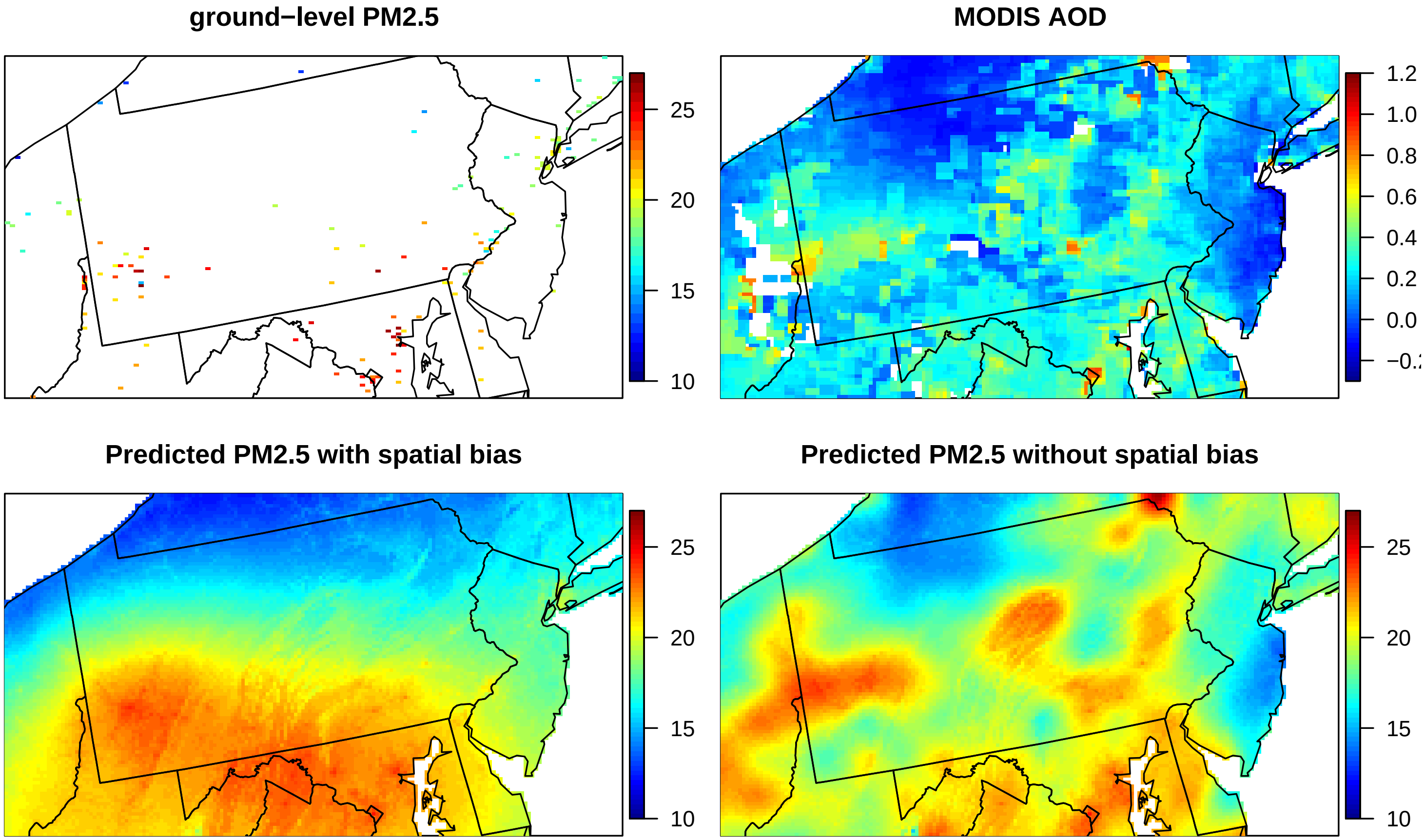
Results for the Model with Spatially Correlated Bias



Model discounts hot spots of AOD by attributing them to ϕ.

Fine-scale heterogeneity introduced through the smooth regression functions, $h_k(\cdot)$.

Key Question: Correlated or Independent AOD Noise?



ground−level PM2.5

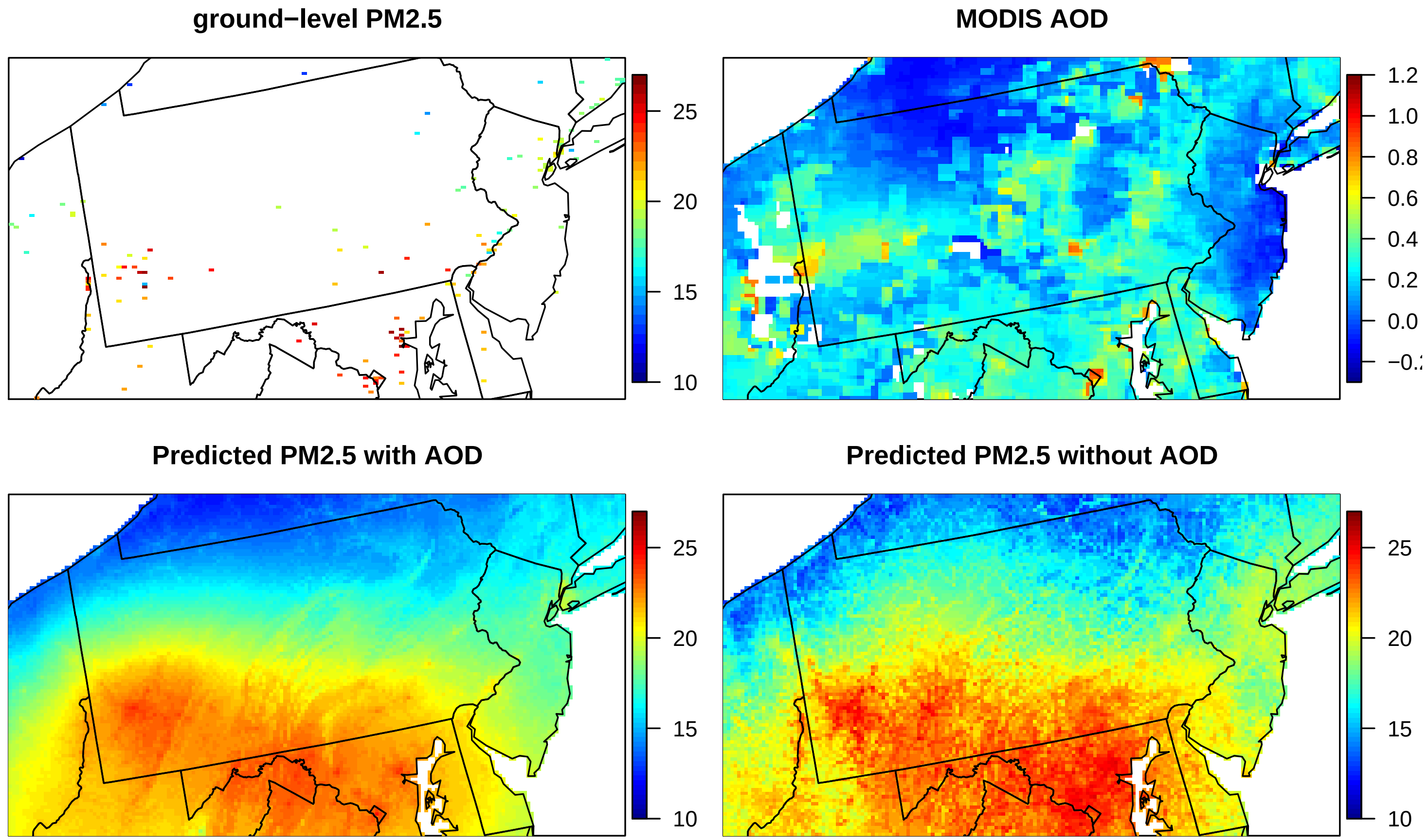
MODIS AOD

Predicted PM2.5 with spatial bias

Predicted PM2.5 without spatial bias

Assumptions about structure of bias drastically change predictions. More model comparison is needed, but patterns of MODIS AOD and scientific understanding suggest that spatially-structured bias is more reasonable.

Key Question: Does Inclusion of AOD Improve Exposure Predictions?



ground−level PM2.5

MODIS AOD

Predicted PM2.5 with AOD

Predicted PM2.5 without AOD

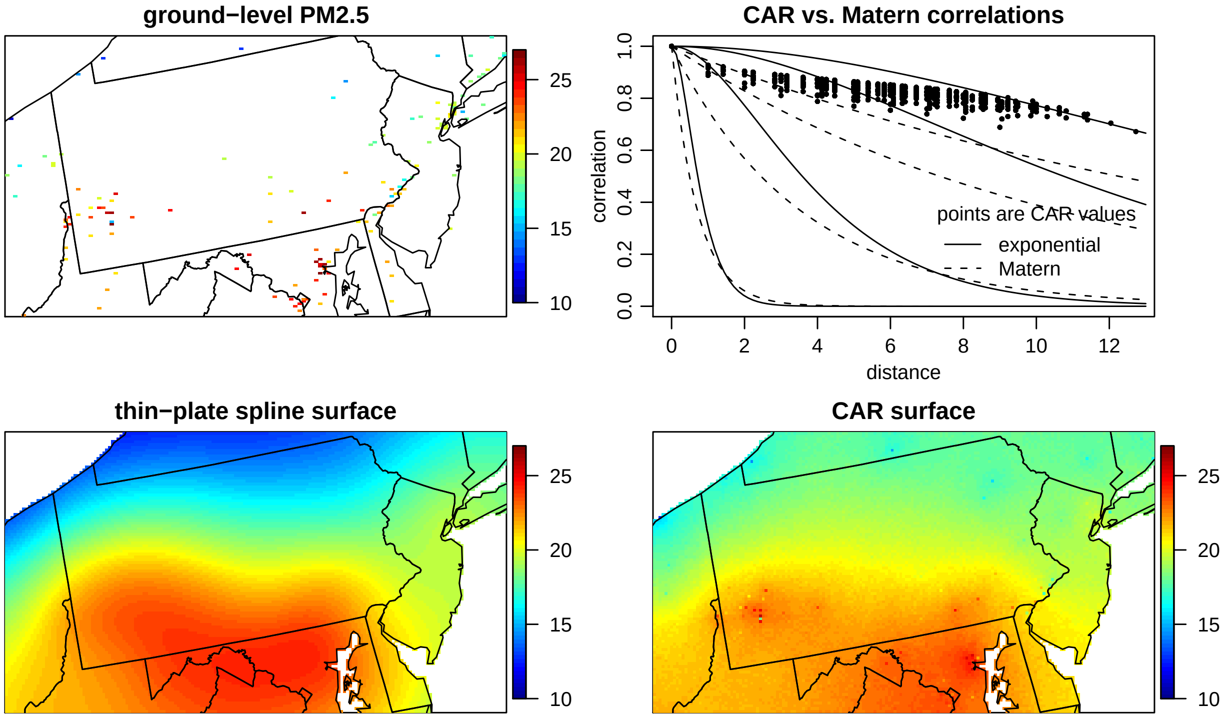
Patterns are somewhat similar and usefulness of AOD is not a foregone conclusion because of spatial bias term; conclusions await further model comparison.

Next Steps

- Model comparison, including cross-validation, to assess usefulness of AOD and need for spatial bias term.
- Better understanding of spatial bias and identifiability.
- Additional covariates, including information on pollution emissions.
- Inclusion of MISR and GOES AOD data.
- Additional subregions
- Assessment of and accounting for non-ignorable missingness.
- Full space-time modelling over multiple months.

Sidebar: CAR models on grids with sparse observations

Question: when working with latent processes on a grid, are CAR models a good choice? Here we work with a very simple model based only on the ground PM observations and simple spatial smoothing.



- Results: CAR correlation falls off quickly but levels off and drops slowly, causing both hot spots and global smoothing. This contrasts with exponential and Matern correlations with various spatial range values, which show larger local correlations that then drop more quickly at longer distances.

- Conclusion: CAR models are a poor choice for latent spatial processes with sparse observations because of their implied spatial correlation function.
- Also, distance-based weight decay functions do not materially alter the CAR spatial correlation function in a helpful way (not shown).

Introduction

Modeling

Case Study

Final Thoughts