ASSESSING SYSTEMATIC DISCREPANCIES BETWEEN POLLUTION OBSERVATIONS AND PROXY (SATELLITE OR PHYSICAL MODEL) VARIABLES USING STATISTICAL MODELING

Christopher J. Paciorek[1]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA; www.biostat.harvard.edu/~paciorek/research/presentations/presentations.html

INTRODUCTION

- Increasingly researchers seek to use proxy variables such as remote sensing retrievals and deterministic model output to improve spatial characterization of pollution concentrations
- Statisticians have developed methods for 'data fusion' that seek to improve predictions of environmental processes by combining sparse gold standard data with the proxy information.
- Current data fusion models assume that the error or discrepancy in the proxy relative to the true underlying environmental process of interest is a combination of white noise and very smoothly-varying spatial discrepancy.
- Here I propose a flexible model for discrepancy between the proxy and the truth.
- The model is able to discount the proxy at scales at which there is little correspondence between proxy and gold standard.
- In addition, the modeling approach holds promise for improving understanding of how the association of proxy and gold standard varies by scale.

DATA SOURCES

Remote Sensing Observations

- MODIS AOD: 16 day orbit repeat, observations every 1-2 days at 10:30 am for a given location, 10 km nominal resolution; averaged to the month after calibration to meteorology; 2001-2007 available – we use 2004.
- CMAQ $PM_{2.5}$: 36 km resolution; half-hour estimates; averaged to month; 2001

$PM_{2.5}$ and Covariate Information

- $PM_{2.5}$ measurements from AQS and IMPROVE: daily average, every 1, 3, or 6 days; averaged to the month
- Weather data at 32 km, 3 hour resolution from North American Regional Reanalysis
- GIS-derived information: distance to roads (and road density) by road class, population density, land use
- NEI point source and county-level area emissions

SUMMARY OF RESULTS

1). Carefully-specified Markov random field (MRF) spatial models can capture a variety of types of spatial structure in the discrepancy term.

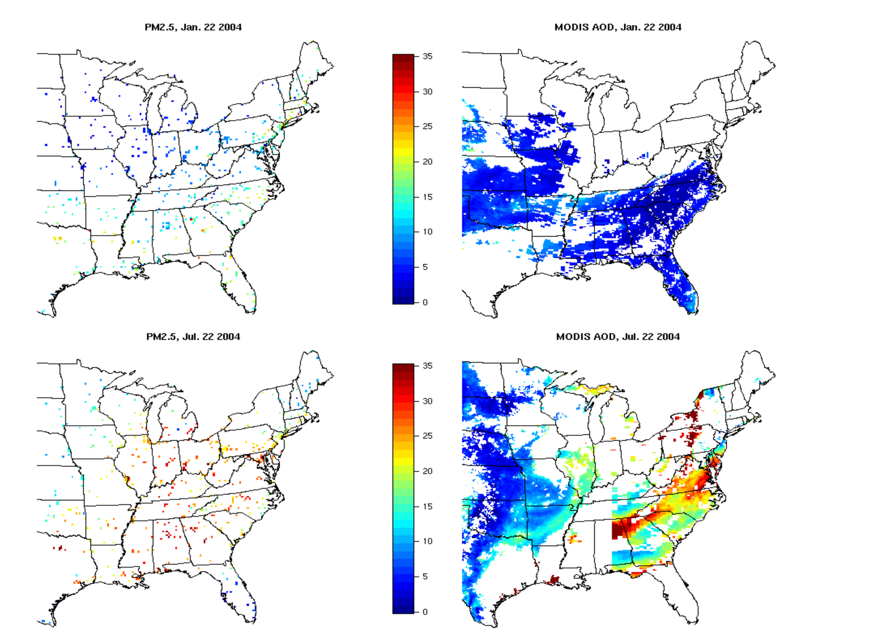
2.) The sparse matrix representations of MRFs allow for efficient computations that other statistical representations do not.

3.) In the AOD and CMAQ examples here, the model estimates that the discrepancy dominates the model for the proxy, heavily downweighting the contribution of the proxy to the final predictions. This suggests that AOD and CMAQ are not helpful in prediction of $PM_{2.5}$ in the contexts examined here.

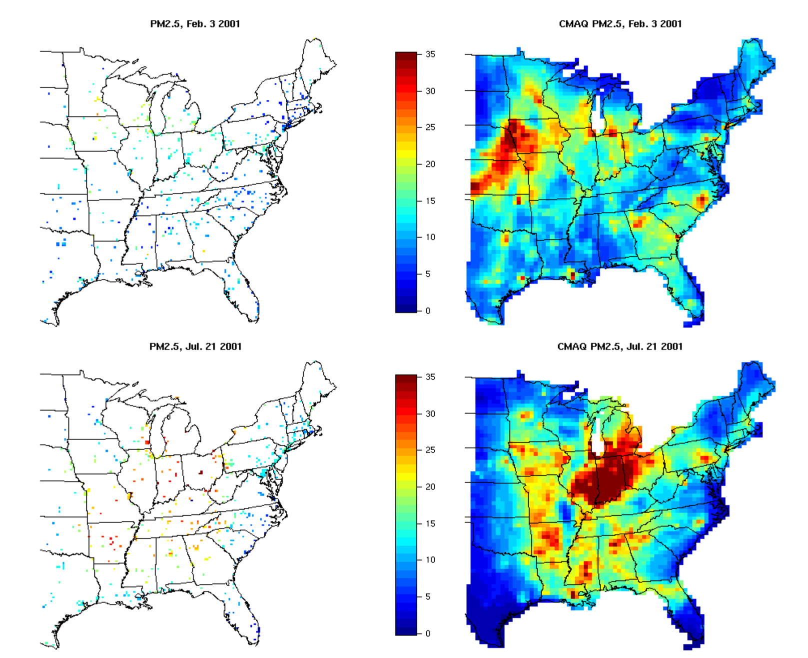
EXPLORATORY ANALYSIS IN EASTERN U.S.

Key Question: At what scales are spatial patterns in proxies reflective of patterns in ground-level $PM_{2.5}$?

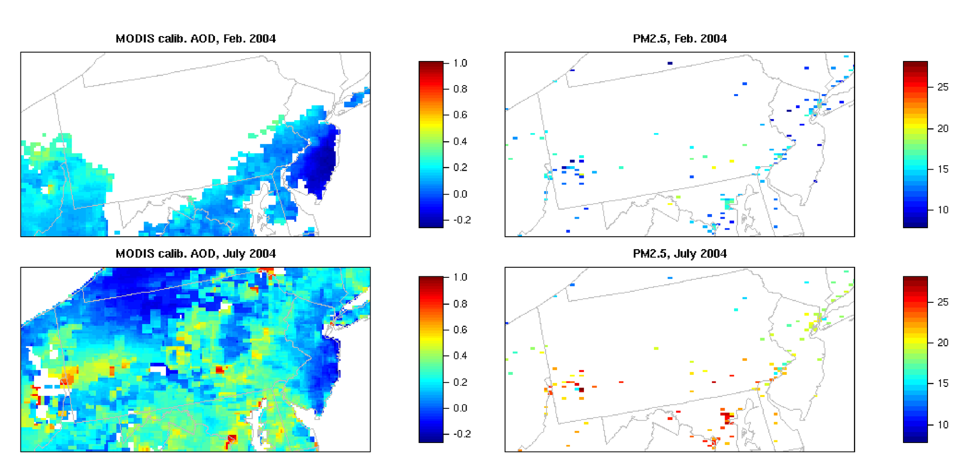
Daily MODIS AOD-PM comparison examples



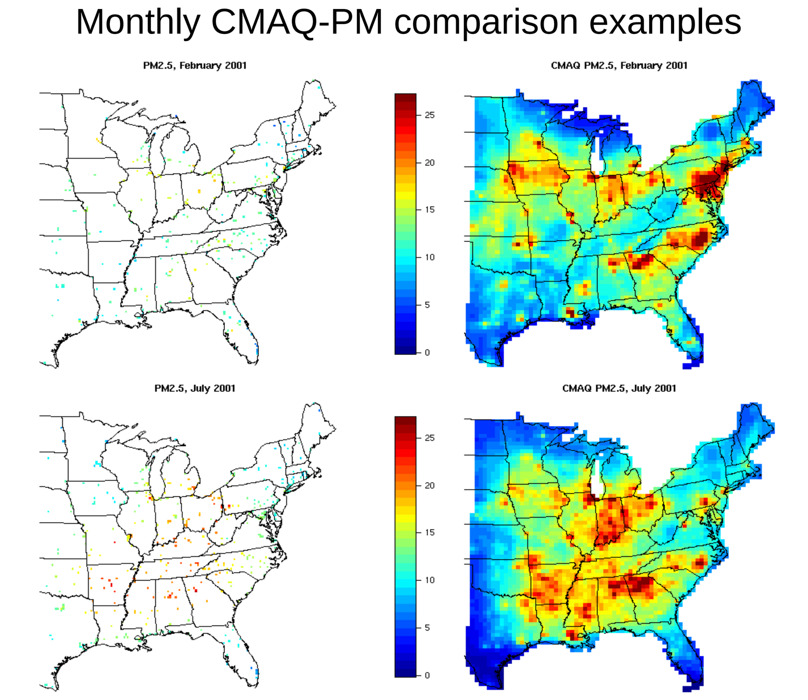
Daily CMAQ-PM comparison examples



Monthly MODIS AOD-PM comparison examples



Monthly CMAQ-PM comparison examples



Associations of MODIS AOD retrievals and $PM_{2.5}$

	Raw MODIS AOD	Calibrated MODIS AOD
2004 Daily values; eastern U.S.		
Overall correlation (longitudinal plus cross-sectional)	0.60	0.64
Average of daily (cross-sectional) correlations	0.35	0.45
Average of daily, April-October only	0.42	0.50
2004 Yearly averages, eastern U.S.		
Overall correlation	0.14	0.36***
2004 Yearly averages; Pennsylvania Focal Region		
Overall correlation	0.09	0.49***
April-October only	-0.11	0.41***

*** Caution; much of this correlation is driven by spatial calibration and does not seem to represent predictive ability of calibrated AOD.

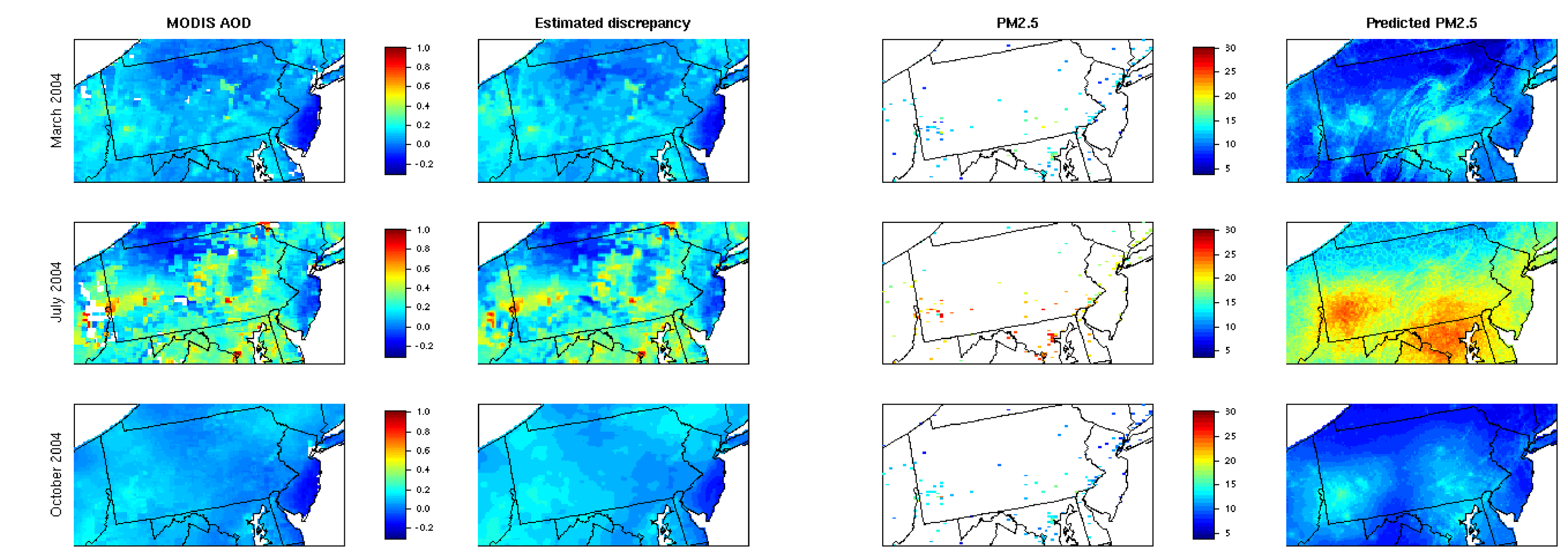
Associations of CMAQ-estimated $PM_{2.5}$ and monitored $PM_{2.5}$

	CMAQ PM2.5, Layer 1***
Overall correlation (longitudinal plus cross-sectional) of daily values	0.56
Average of daily (cross-sectional) correlations	0.50
Correlation of yearly averages	0.51

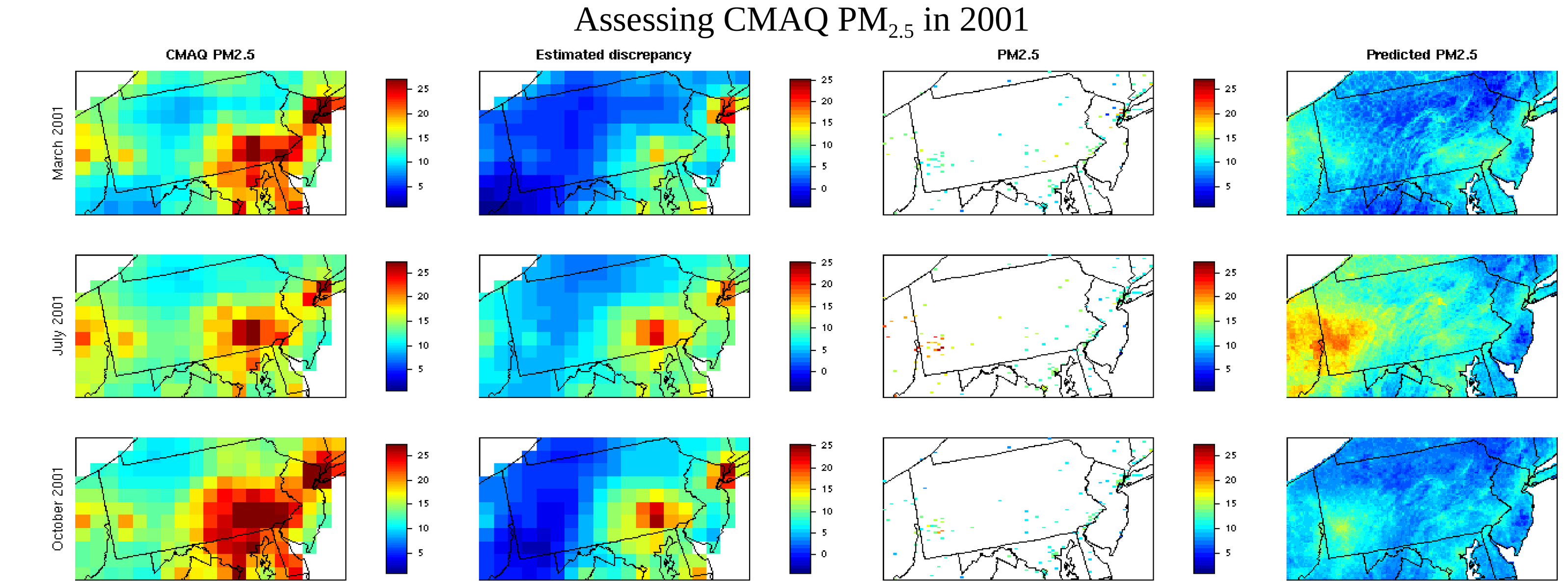
***Averaging first three layers results in very high correlation with first layer alone.

MODEL RESULTS, MID-ATLANTIC REGION

Assessing calibrated MODIS AOD in 2004



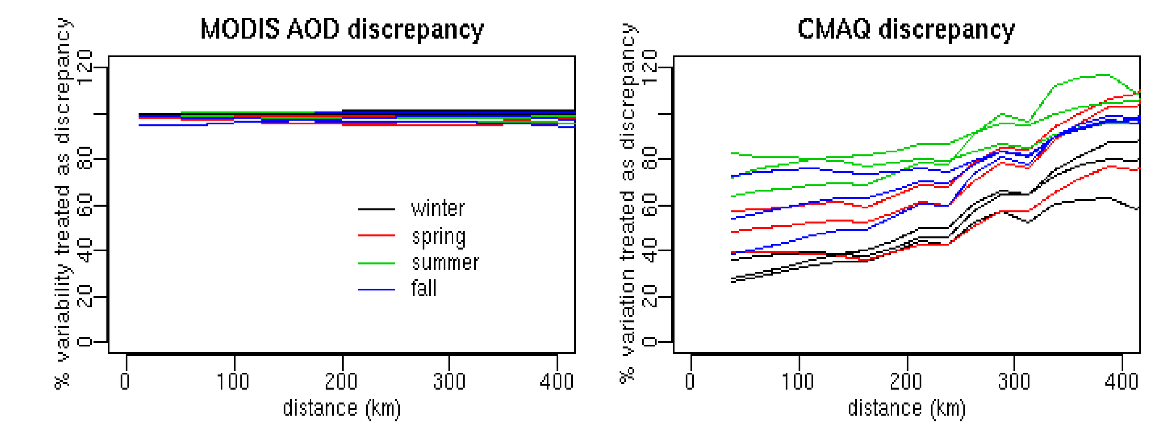
Assessing CMAQ $PM_{2.5}$ in 2001



Proportion of variation in proxy explained by the discrepancy as a function of spatial scale

The proposed variogram ratio is: $R(d) = \frac{\text{Variog}(\phi)}{\text{Variog}(\beta_1 P) + \text{Variog}(\phi + \beta_1 P)}$

This makes use of an idea introduced by Jun and Stein (2004)



All variability in MODIS AOD is being accounted for in the discrepancy term, while for CMAQ $PM_{2.5}$, some of the variability at smaller scales is accounted for in the latent $PM_{2.5}$ process.

Despite this, for CMAQ, as for MODIS AOD, the proxy contributes little to predictive ability, as seen below.

Predictive ability of various model specifications

	Monthly R^2 (correlation)	Yearly R^2 (correlation)
Model with calibrated MODIS AOD, 2004		
Core model	0.80 (0.89)	0.65 (0.81)
No AOD	0.80 (0.90)	0.63 (0.80)
No discrepancy term	<0 (0.18)	<0 (<0)
Discrepancy forced very smooth	0.71 (0.84)	0.50 (0.71)
Model with CMAQ PM2.5, 2001		
Core model	0.74 (0.87)	0.51 (0.79)
No CMAQ	0.77 (0.88)	0.61 (0.79)
No discrepancy term	0.46 (0.74)	<0 (0.40)
Discrepancy term forced very smooth	0.60 (0.81)	<0 (0.73)
Core model without covariates	0.72 (0.85)	0.31 (0.56)
Core model, no covariates or CMAQ	0.71 (0.85)	0.29 (0.55)

STATISTICAL MODEL OF SYSTEMATIC DISCREPANCY USING MARKOV RANDOM FIELDS

Likelihoods for monthly average data:

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

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

- $f_k(\cdot)$, $k = 1, \ldots, K_f$ are nonparametric regression functions of within-grid cell covariates. These help to account for fine-scale variation in the gold standard.
- $\phi(s)$ is the key spatially-correlated discrepancy term.

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, specified as a thin plate spline.

- $\phi(s)$ is specified as a Gaussian Markov random field (MRF) with a neighborhood structure that gives an approximation to a thin plate spline (TPS)

$$\phi \sim \mathcal{N}(0, \kappa Q^{-1})$$

- κ controls the amount of spatial smoothing
- The rows of Q give the neighbor weights for a given element of ϕ.

Standard 0–1 neighbor weights

$$\begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array}$$

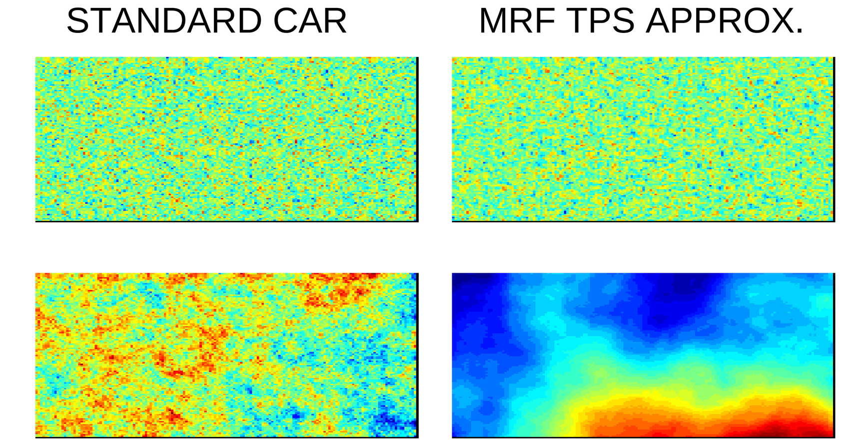
Weights for thin plate spline approx.

$$\begin{array}{ccccc} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{array}$$

- Q is very sparse, so the matrix calculations for Bayesian estimation are very fast. We work with 17,500 elements in ϕ for the AOD model.
- Other statistical representations cannot handle this dimensionality and still be able to represent both smooth and wiggly processes.

Standard CAR models cannot represent smooth processes (left) while MRF TPS approximation can represent both noisy and smooth processes (right).

STANDARD CAR MRF TPS APPROX.



The main idea is that if the proxy well-represents the truth at large but not small scale, the discrepancy term acts to account for spatial autocorrelation.
If the proxy well-represents the truth at small but not large scale, the discrepancy corrects for this mismatch, provided sufficient gold-standard data.

ONGOING WORK

- Development of a full spatio-temporal model to avoid assumption of independence between months.
- Simulation-based assessment of the modeling approach.

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

Thanks to AER for providing CMAQ output. Thanks also to Yang Liu for processing and discussions.