INTEGRATING SATELLITE AND MONITORING DATA TO RETROSPECTIVELY ESTIMATE MONTHLY $PM_{2.5}$ CONCENTRATIONS IN THE EASTERN U.S.

Christopher J. Paciorek[1] and Yang Liu[2]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA; [2]Department of Environmental Health, Harvard School of Public Health, Boston, MA

www.biostat.harvard.edu/~paciorek/research/presentations/presentations.html

INTRODUCTION

- Remote sensing observations of aerosol hold promise for adding information about $PM_{2.5}$ concentrations beyond that from monitors, particularly in suburban and rural areas with limited monitoring.
- AOD (aerosol optical depth) observations are frequently missing, and noisy and biased relative to $PM_{2.5}$.
- Bayesian statistical modeling holds promise for integrating AOD, PM2.5, and GIS and weather information to predict monthly $PM_{2.5}$ concentrations on a fine grid (4 km).
- Key challenges include:
- 1.) formulation of a statistical model to relate observations to a latent space-time process representing true $PM_{2.5}$ in a way that accounts for spatial and temporal mismatch and nature of error and bias.
- 2.) representation of the latent process that provides appropriate spatial and temporal correlation while allowing for computationally-efficient statistical estimation

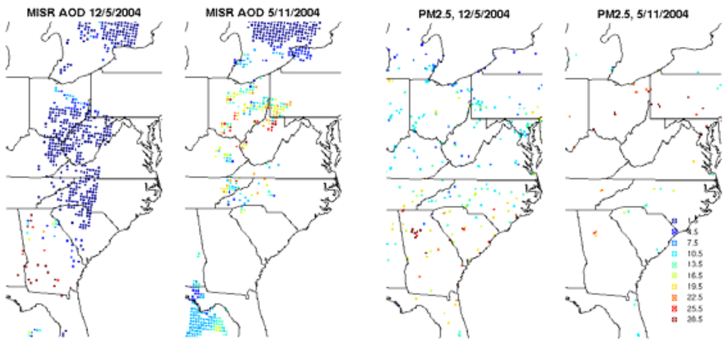
SUMMARY OF INTERIM RESULTS

- Daily MISR AOD shows little association with ground monitors of $PM_{2.5}$ across time and at individual stations.
- Calibration of MISR AOD to $PM_{2.5}$ measurements, modified by weather variables and spatial and temporal bias terms, improves correlations between AOD and $PM_{2.5}$, particularly when averaging over time.
- There is limited evidence that missing MISR AOD observations are associated with the level of $PM_{2.5}$.
- Satellite AOD holds some promise for enhancing predictions of $PM_{2.5}$, but is likely most useful at monthly or yearly temporal scale.
- Ability of satellite AOD to improve predictions relative to models based on $PM_{2.5}$ data, weather and GIS variables is a key question.
- Conditional autogregressive (CAR) space-time models hold promise for computationally efficient latent process estimation in a Bayesian statistical framework.
- CAR models can account for spatial and temporal correlation induced by underlying physical reality, areally-integrated satellite observations, and time averaging of incomplete satellite observations.

ONGOING AND NEAR-TERM WORK

- Calibration of GOES and MODIS AOD observations with $PM_{2.5}$, modified by weather variables and spatial and temporal bias terms.
- Comparison of strength of association of AOD with $PM_{2.5}$ for the different satellite instruments.
- Assessment of spatial and temporal scales at which satellite AOD is useful for estimating $PM_{2.5}$.
- Ongoing data processing and matching of satellite observations and GIS variables to base 4 km grid.
- Full development of daily- and monthly-scale Bayesian statistical models for $PM_{2.5}$ prediction based on CAR framework.
- Initial model fitting for small region and several month time period to assess computational feasibility and compare daily/monthly approaches.

DATA SOURCES



Remote Sensing Observations

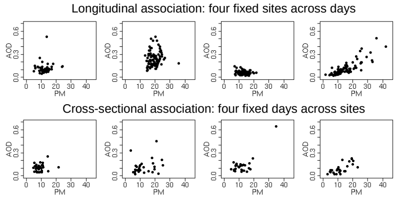
- MISR AOD: 16 day orbit repeat, observations every 4-7 days at 10:30 am for a given location, 17.6 km resolution
- MODIS AOD: 16 day orbit repeat, observations every 1-2 days for a given location, 10 km resolution
- GOES AOD: observations every half hour, 4 km resolution

$PM_{2.5}$ 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

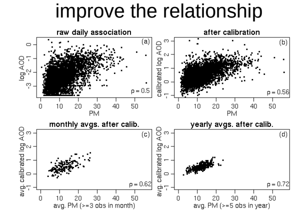
ASSESSMENT AND CALIBRATION OF MISR AOD

AOD not strongly related to daily PM



Scatterplots of AOD against PM across site for four individual days (top row) and for AOD against PM across time for four individual sites (bottom row) suggest that at the daily scale and without calibration, the association is weak and variable.

Calibration and temporal averaging improve the relationship

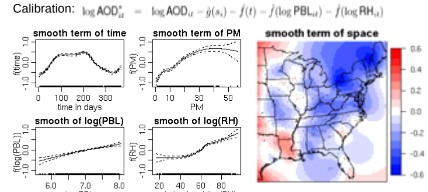


Log AOD vs. PM before and after calibration with RH, PBL, and spatial and temporal bias terms (top row). Average calibrated log AOD against average PM over a month and over a year (bottom row).

Statistical calibration of AOD to PM

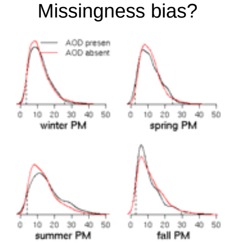
GAM model: $\log \mathrm{AOD}_{it} = g(s_i) + f(t) + f(\log \mathrm{PBL}_{it}) + f(\log \mathrm{RH}_{it}) + f(\mathrm{PM}_{it}) + \epsilon_{it}$

Calibration: $\log \mathrm{AOD}^*_{it} = \log \mathrm{AOD}_{it} - \hat{g}(s_i) - \hat{f}(t) - \hat{f}(\log \mathrm{PBL}_{it}) - \hat{f}(\log \mathrm{RH}_{it})$



Relationships of log(AOD) with PM as modified by time, space, log(PBL), and RH. Smooth terms indicate how each factor affects the bias in log(AOD) as a proxy for PM. For example during the summer (days 150-240), log(AOD) is more positively offset (biased) with respect to PM than in the winter. GAM provides calibration of log(AOD) at daily scale that allows averaging to longer time periods.

Missingness bias?



After adjustment for space, time, and PBL, there is some evidence that missing AOD indicates lower (~2 ug/m3) PM in summer and higher (0.67 ug/m3) PM in fall, with little difference in winter and spring.

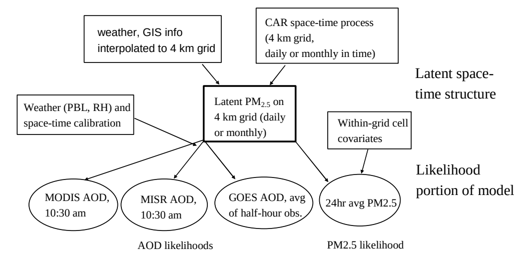
STATISTICAL MODELLING

Challenges:

- Large data sources and desire for fine-scale prediction
- AOD is a biased and noisy reflection of $PM_{2.5}$
- Need for spatial and temporal correlation in modelling $PM_{2.5}$
- Spatial correlation of AOD errors
- Irregular sampling of both AOD and $PM_{2.5}$ in space and time
- Missingness of AOD may be related to $PM_{2.5}$ levels
- Spatial mismatch of data sources (point data plus varying areal units)

Basic solutions:

- Calibrate AOD to $PM_{2.5}$ (partly as preprocessing, partly in model)
- Relate all quantities to latent PM2.5 variable on base 4km grid
- Treat AOD at natural resolution, as weighted averages of $PM_{2.5}$ on base grid, with calibration
- Use conditional autoregressive (CAR) space-time statistical models to build space-time correlation in computationally feasible manner (use weights decaying with distance to ensure adequate spatial correlation)
- Use weather and GIS information to help estimate $PM_{2.5}$



Likelihood Terms

PM2.5 likelihood with $f(z)$ smooth terms of distance to major road, K_Y mapping observations to grid cells, and X the latent space-time PM process on the 4 km grid at daily or monthly resolution:

$$Y \sim \mathcal{N}(f(z) + K_Y X, \sigma^2 I)$$

AOD likelihood (separate terms for MISR, MODIS, or GOES) with A^* pre-calibrated based on weather data and spatial and temporal bias terms, b_0 and b_1 bias terms, K_A a weight matrix mapping pixels to grid cells, and P a sparse spatial CAR precision matrix representing spatially-correlated AOD error

$$\log A^* \sim \mathcal{N}(b_0 1 + b_1 K_A X, P^{-1})$$

AOD missingness likelihood with W representing augmented data in the probit regression formulation and M a vector missingness indicators for potential AOD observations

$$\begin{aligned} W &\sim \mathcal{N}(m_0 1 + m_1 K_A X, I) \\ M &= 1(W > 0) \end{aligned}$$

Latent Process Representation and Fitting

Latent process formulation with $\sum f_k(z_k)$ smooth terms of weather and GIS variables and g a latent space-time process with CAR space-time structure on the 4 km grid at daily or monthly resolution and sparse space-time precision matrix, Q.

$$\begin{aligned} X &= \sum f_k(z_k) + g \\ g &\sim \mathcal{N}(0, Q^{-1}) \end{aligned}$$

The CAR representation allows for Bayesian MCMC estimation of g the core latent space-time process by sampling g as a Gibbs step

$$\begin{aligned} g|A^*, W, Y &\sim \mathcal{N}\big(V^{-1}\big(b_1 P^T K_A^T(\log A^* - b_0 1) + \\ &\quad m_1 K_A^T(W - m_0 1) + \sigma^{-2} b_1 K_Y^T(Y - f(z))\big), V\big) \\ V^{-1} &= Q + b_1^2 K_A^T P K_A + m_1^2 K_A^T K_A + \sigma^{-2} K_Y^T K_Y \end{aligned}$$

The conditional precision matrix, V^{-1}, is sparse because all components and products are sparse, which allows very efficient Gibbs sampling

Other quantities, namely variance components and regression terms and W, are estimated within the MCMC procedure in (hopefully) computationally efficient steps.

Build Model at Daily or Monthly Level?

Daily model

- More naturally treats daily observations
- Satellite pixels represented as weighted averages of 4 km grid cells
- $PM_{2.5}$ data relatively sparse
- Much more computationally intensive
- Monthly latent $PM_{2.5}$ estimated as average of latent daily estimates on grid

Monthly model

- Aggregate data to the month after daily satellite calibration; more computationally feasible
- Need to assign AOD measurements to multiple 4 km cells and then average within cells
- AOD and $PM_{2.5}$ monthly averages do not have constant error variance (varying number of days)
- Unusual induced correlations of time-averaged AOD.

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