Enhancing EPHT with Satellite-Driven PM$_{2.5}$ Exposure Modeling and Epidemiology – Year 2 Report

Yang Liu, Ph.D.
Rollins School of Public Health
Emory University

NASA Public Health Review Meeting
Santa Fe, NM
September 14-16, 2011
Project Team

- Emory/RSPH: Yang Liu (PI), Jeremy Sarnat, Mitch Klein, Xuefei Hu, Heather Strosnider, and Erika Rees
- MSFC/USRA: Dale Quattrochi, Bill Crosson, Mohammad Al-Hamdan, Maury Estes, Sue Estes, Sarah Hemmings, and Gina Wade
- CDC/NCEH: Judy Qualters, Paul Garbe, Helen Flowers, and Ambarish Vaidyanathan

Sponsored by NASA Applied Science Program grant NNX09AT52G under J. Haynes and Sue Estes
Research Objectives

- Extend the spatial coverage of the PM\(_{2.5}\) indicators in Tracking Network with satellite data
- Provide timely estimates of county average PM\(_{2.5}\) health indicators
- Evaluate satellite PM\(_{2.5}\) estimates as an alternative exposure data source in environmental epidemiologic studies and using independent ground sampling
Technical Approach

- Meteorological data: NARR
- Satellite data: MISR/MODIS/GOES/OMI
- Land use: NCLD or MODIS
- Ground truth: EPA & IMPROVE

\[ [PM_{2.5}] = F(\text{satellite aerosol retrievals, meteorological calibration, source adjustment}) \]

- Estimated PM$_{2.5}$ concentration surface
  - Prospective sampling
  - Comparison with HBM outputs
  - Epidemiological modeling
Year 2 Progress Summary

- Proposed Tasks:
  1. Spatial model development and comparison of NARR and NLDAS (Emory, manuscript submitted)
  2. Initiation of prospective sampling (Emory)
  3. AOD calibration with AERONET (Emory, MSFC)
  4. Nearest neighbor approach development (MSFC)

- Ahead of Schedule:
  5. Initiation of epidemiological analysis (Emory)

- Need More Work
  6. Comparison with HBM
Study Domain

- Number of monitoring sites: 119
- Exposure modeling domain: 700 x 700 km²
- SEARCH sites: 2 independent validation sites
Geographically Weighted Regression Model

GWR allows model parameters to vary in space to better capture spatially varying AOD-PM relationship – major advantage over global regression models.

Model Structure

\[
[PM_{2.5}]_ {(x,y)} \sim \beta_0(x,y) + \beta_1(x,y) \times AOD + \beta_2(x,y) \times PBL + \beta_3(x,y) \times RH \\
+ \beta_4(x,y) \times Temp + \beta_5(x,y) \times Wind\_Speed + \beta_6(x,y) \times Forest\_Cover
\]

Datasets (2003):
- \(PM_{2.5}\) – EPA / IMPROVE daily measurements
- AOD – MODIS collection 5 (10 km) or GASP (4 km)
- Meteorology – NLDAS-2 (14 km) or NARR (32 km)
- Land use: NLCD 2001

Model is fitted at daily level
Model Fitting Results

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Obs. Per Day</td>
<td>101</td>
</tr>
<tr>
<td>Model Days</td>
<td>137 (37.5%)</td>
</tr>
<tr>
<td>Total Obs.</td>
<td>4,477</td>
</tr>
</tbody>
</table>

No residual spatial autocorrelation was found in 78% of the daily GWR models.

Local $R^2$ values vary in time – daily model is necessary.

$\geq$10 matched data records is needed to stabilize the model.
Model Performance Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model $R^2$</td>
<td>0.86</td>
<td>0.56</td>
<td>0.92</td>
</tr>
<tr>
<td>CV $R^2$</td>
<td>0.70</td>
<td>0.22</td>
<td>0.85</td>
</tr>
</tbody>
</table>

SEARCH site predictions

<table>
<thead>
<tr>
<th>Site</th>
<th>N</th>
<th>Annual PM$_{2.5}$ ($\mu$g/m$^3$)</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHM</td>
<td>85</td>
<td>19.1</td>
<td>0.90</td>
</tr>
<tr>
<td>JST</td>
<td>87</td>
<td>15.3</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Putting all the data points together, we see unbiased estimates

Slope = 0.98

Slope = 0.97
Spatial Pattern of Model Bias

Model Fitting

Cross Validation

Negative and positive model / CV residuals are randomly distributed.
Model Predicted Mean PM$_{2.5}$ Surface

Note: annual mean calculated with 137 days
Comparison with Other Models

**Pros:**

- Better performance than global regression models
- Better reflection of temporal variability than LUR models
- Stronger physical base than kriging models
- Simpler and faster than air quality models

**Cons:**

- Integration with air quality models?
- Statistical data filling is under study
- Higher resolution data will become available soon

Manuscript submitted to *Environmental Research*
Strategy

- Identify a “hot” and a “cool” pixel based on ratios of GWR daily PM$_{2.5}$ concentrations over regional mean.
- 3 sampling locations > 3 km apart in each 12 km pixel
- ~20 24-hr samples in the next 6-9 months

So far, 3 sites located, portable samplers tested, made 2 sampling trips.
A Closer Look

2. Field Sampling
Rational and Approach

For satellite data to be considered a reliable source of exposure estimates in health studies, both the spatial pattern and absolute levels of predicted PM$_{2.5}$ concentrations are important.

General calibration model structure (fitted annually)

\[
\text{AERONET AOD} = \alpha + \beta_1 \times \text{satellite AOD} + \text{season} \\
+ \beta_2 \times \text{satellite AOD} \times \text{season}
\]

Caveat: without calibration, MODIS can’t be used for seasonal trend analysis, GOES can’t be used for either seasonal or interannual trend analysis
Rational and Approach

Problem: cloud cover causes a lot of data missingness. Without any treatment, best possible coverage is ~ 50%.

Hypothesis: missing AOD values due to small clouds can be filled with its nearest neighbors without significantly disturbing the predicted PM$_{2.5}$ surface.

Method: maximum distance over which nearby observations may be used to fill in missing grid cell values = 20 km
Preliminary Results

4. Nearest Neighbor Filling

<table>
<thead>
<tr>
<th>Coverage (%)</th>
<th>N_days</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>365</td>
<td>46.04</td>
</tr>
<tr>
<td>NN</td>
<td>365</td>
<td>65.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coverage (%)</th>
<th>RMSE (μg/m³)</th>
<th>Relative Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw_NARR</td>
<td>5.61</td>
<td>60.4</td>
</tr>
<tr>
<td>NN_NARR</td>
<td>4.82</td>
<td>66.8</td>
</tr>
</tbody>
</table>

NN filling: (1) improve coverage (2) improve model performance
Plan of epidemiological analysis

1. Communicate with epidemiologists on data format, structure, and modeling needs
2. Generate daily PM$_{2.5}$ estimates using calibrated, nearest neighbor-filled MODIS AOD for 2000 – 2007
3. Spatially join with zip code level patient addresses
4. Work with epidemiologists to develop space-time model
5. Evaluate resulted exposure-response functions
Year 3 Tasks

- **Emory**
  - MODIS/GOES data fusion
  - Final GWR PM$_{2.5}$ modeling
  - Development of new model structure
  - Field sampling and sample analysis
  - Health effects modeling and evaluation

- **MSFC**
  - Further study of gap filling techniques
  - Finalization of gridded aerosol data

- **CDC**
  - Comparison between HBM and satellite
  - Project benefit assessment