Bayesian Nonparametric Ensemble (BNE) algorithm for predictions of high spatiotemporal PM$_{2.5}$ concentrations

Vijay Kumar$^1$, Jaime Benavides$^1$, Carlos Carrillo-Gallegos$^2$, Arlene Fiore$^3$, John Paisley$^3,4$ and Marianthi-Anna Kioumourtzoglou$^1,4$

$^1$Environmental Health Sciences, Columbia University, New York, NY, $^2$Earth and Planetary Sciences, MIT, Cambridge, MA,
$^3$Electrical Engineering, Columbia University, New York, NY, $^4$Data Science Institute, Columbia University, New York, NY

Background and Methods

Motivating Problem:
- Air pollution associated with several adverse health outcomes [1]
- Scarcity of ground level monitors means we must model exposure
- Satellite data essential in filling gaps, but requires translation from column to surface measurements
- Current health studies use predictions from a single dataset
- We present a novel ensembling method to incorporate several different exposure datasets (with various methodologies) and provide uncertainty

Bayesian Nonparametric Ensemble (BNE) algorithm
- Integrate information across existing spatiotemporal prediction models
- Weights each model by its spatiotemporal predictive accuracy
  - i.e., spatiotemporal weights
- Provide spatiotemporal uncertainty of predictions
  - Including due to existing model disagreement

Flow-chart

Preparation

Training data AOD, CO, PM$_{2.5}$

Training set 80%

Testing set 20%

PM$_{2.5}$ for input models

Data preparation spatiotemporal

Bayesian algorithm

Normalized emissions

BNE predictions

Table: Daily PM$_{2.5}$ base models

Results

Input Exposure Models:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Spatial Resolution</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNE et al. 2019</td>
<td>1 x 1 km$^2$</td>
<td>AOD from MODIS, gap filled with random forest approach, incorporating AOD data, land use regression variables, and meteorology from KEAR</td>
</tr>
<tr>
<td>BNE et al. 2021</td>
<td>1 x 1 km$^2$</td>
<td>Generalized additive model of three ML methods: Incorporating AOD data, AOD from MODIS, SDOSS CHEM, CMAQ (v4.3), and land use variables</td>
</tr>
<tr>
<td>ESIMAP CMAQ (v4.3)</td>
<td>12 x 12 km$^2$</td>
<td>USEPA Air Quality Model (CMAQ) (v4.3.2), driven by WRF (v4.3), meteorology from the National Emissions Inventory (NEI) for 2017</td>
</tr>
<tr>
<td>FASID</td>
<td>12 x 12 km$^2$</td>
<td>Fusion of CMAQ (v4.3) chemical/transport model (VOC and NOx) [4.3.2] and the National Emissions Inventory (NEI) for 2017</td>
</tr>
<tr>
<td>WRF-CMAQ [R. Kumar et. al 2021]</td>
<td>12 x 12 km$^2$</td>
<td>Biweekly assimilating AOD retrievals from MODIS and CO retrievals from MOPITT into CMAQ, with WRF providing meteorological input</td>
</tr>
</tbody>
</table>

Table: Daily PM$_{2.5}$ base models

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Spatial Resolution</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNE et al. 2019</td>
<td>1 x 1 km$^2$</td>
<td>AOD from MODIS, gap filled with random forest approach, incorporating AOD data, land use regression variables, and meteorology from KEAR</td>
</tr>
<tr>
<td>BNE et al. 2021</td>
<td>1 x 1 km$^2$</td>
<td>Generalized additive model of three ML methods: Incorporating AOD data, AOD from MODIS, SDOSS CHEM, CMAQ (v4.3), and land use variables</td>
</tr>
<tr>
<td>ESIMAP CMAQ (v4.3)</td>
<td>12 x 12 km$^2$</td>
<td>USEPA Air Quality Model (CMAQ) (v4.3.2), driven by WRF (v4.3), meteorology from the National Emissions Inventory (NEI) for 2017</td>
</tr>
<tr>
<td>FASID</td>
<td>12 x 12 km$^2$</td>
<td>Fusion of CMAQ (v4.3) chemical/transport model (VOC and NOx) [4.3.2] and the National Emissions Inventory (NEI) for 2017</td>
</tr>
<tr>
<td>WRF-CMAQ [R. Kumar et. al 2021]</td>
<td>12 x 12 km$^2$</td>
<td>Biweekly assimilating AOD retrievals from MODIS and CO retrievals from MOPITT into CMAQ, with WRF providing meteorological input</td>
</tr>
</tbody>
</table>

Mean Concentration: 6.84 ± 0.7 ppb

Mean Concentration: 6.53 ± 0.8 ppb

Mean Concentration: 5.03 ± 0.7 ppb

References


Completed 2010-2016; more years, pollutants coming!