Satellite data for air quality exposure assessment and epidemiology

Yuming Guo, MD, PhD
Professor of Global Environmental Health and Biostatistics
Monash University School of Public Health and Preventive Medicine
Outline

- Why perform air pollution exposure assessment
- Measurement error
- Challenges
- Machine learning
- The role of satellite data
- Wildfire smoke exposure assessment
- Opportunities
Why perform air pollution exposure assessment?
The role of air pollution data in environmental risk assessment

- Exposure level
  - Air pollution concentration

- Effect estimates
  - Relative risk

- Population Attributable Fraction

- Population exposure

- Attributable Disease Burden
Why perform air pollution exposure assessment?

- Limited exposure data, for both space scale and time scale.
What variables are involved in exposure assessment?

- Satellite data.
- Land use information (greenness, road density, urban cover...).
- Weather conditions (temperature, humidity, rainfall, wind speed...).
- Spatiotemporal correlations/trends.
- Others (wildfire, population density...).
Measurement error
Measurement error in air pollution exposure assessment

Estimating spatiotemporal distribution of PM1 concentrations in China with satellite remote sensing, meteorology, and land use information

Gongbo Chen a, Luke D. Knibbs b, Wenyi Zhang c, Shanshan Li a, Wei Cao d, Jianping Guo e, Hongyan Ren d, Boguang Wang f, Hao Wang g, Gail Williams h, N.A.S. Hamm h, Yuming Guo a

\[ y = 1.01x + 0.17 \]
R-squared=59%
RMSE=22.5

\[ y = 0.96x + 0.24 \]
R-squared=71%
RMSE=13.0

\[ y = 1.02x - 2.14 \]
R-squared=77%
RMSE=11.4

Figure. Scatterplots of 10-fold cross-validation for daily, monthly and seasonal estimation of PM1 concentrations (µg/m3)
Measurement error in environmental risk assessment

- Random measurement error lead to underestimated effect estimates.
- Does this mean random measurement error lead to underestimated burden of disease attributable to the exposure?

The relationship between simulated exposure (X) and response outcome (Y)
The more accurate exposure assessment, the more accurate effect estimates and the more accurate estimation of disease burden!
Challenges
Challenges for air pollution exposure assessment

- Missing values in satellite data.

**Figure.** Aerosol optical depth (550 nm) from Moderate Resolution Imaging Spectroradiometer (MODIS)-Aqua over Asia.

The aerosol optical depth (AOD) data are mapped to show: TAU_ORG: original data considering varying pixel size; TAU_GRD: high resolution gridded data with no filling of empty grids; and TAU_HRG: high resolution gridded data with spatial filling at the edge of the swath.

Gupta et al. High-Resolution Gridded Level 3 Aerosol Optical Depth Data from MODIS. Remote Sensing.
Challenges for air pollution exposure assessment

- Filling missing values of satellite data.

### Table 2. Comparison of geostatistical and machine learning (ML) interpolation for MAIAC AOD product.

<table>
<thead>
<tr>
<th>Interpolation Method</th>
<th>Coverage (%)</th>
<th>Computation Time</th>
<th>CV RMSE</th>
<th>CV R^2</th>
<th>CV MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Before Interpolation</strong></td>
<td>15.46</td>
<td>\</td>
<td>\</td>
<td>\</td>
<td>\</td>
</tr>
<tr>
<td><strong>Geostatistical Algorithms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>87.31</td>
<td>55:38:56:56</td>
<td>0.17</td>
<td>0.75</td>
<td>20.56</td>
</tr>
<tr>
<td>ST Kriging</td>
<td>67.73</td>
<td>128:56:43:56</td>
<td>0.17</td>
<td>0.78</td>
<td>20.36</td>
</tr>
<tr>
<td>IDW</td>
<td>45.22</td>
<td>45:35:28:39</td>
<td>0.18</td>
<td>0.65</td>
<td>21.35</td>
</tr>
<tr>
<td>Kriging</td>
<td>42.37</td>
<td>88:46:37:24</td>
<td>0.17</td>
<td>0.66</td>
<td>20.89</td>
</tr>
<tr>
<td>NN</td>
<td>21.43</td>
<td>15:39:27:93</td>
<td>0.19</td>
<td>0.49</td>
<td>25.38</td>
</tr>
<tr>
<td><strong>ML Algorithms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>98.64</td>
<td>120:55:28:65</td>
<td>0.15</td>
<td>0.89</td>
<td>18.00</td>
</tr>
<tr>
<td>XG</td>
<td>98.64</td>
<td>18:00:38:20</td>
<td>0.15</td>
<td>0.85</td>
<td>19.06</td>
</tr>
<tr>
<td>SVM</td>
<td>98.64</td>
<td>19:04:47:64</td>
<td>0.17</td>
<td>0.72</td>
<td>19.41</td>
</tr>
<tr>
<td>BRNN</td>
<td>98.64</td>
<td>18:45:36:22</td>
<td>0.17</td>
<td>0.70</td>
<td>22.39</td>
</tr>
<tr>
<td>GBM</td>
<td>98.64</td>
<td>06:35:47:65</td>
<td>0.18</td>
<td>0.69</td>
<td>25.17</td>
</tr>
<tr>
<td>GAM</td>
<td>98.64</td>
<td>01:05:38:20</td>
<td>0.17</td>
<td>0.62</td>
<td>21.88</td>
</tr>
<tr>
<td>LASSO</td>
<td>98.64</td>
<td>01:18:30:23</td>
<td>0.19</td>
<td>0.49</td>
<td>28.03</td>
</tr>
</tbody>
</table>

TS (two-step interpolation), IDW (inverse distance weighting), ST Kriging (spatio-temporal kriging), NN (nearest neighbors with 30 km buffer), RF (Conditional Inference Random Forest), XG (extreme gradient boosting), XGBoost, SVM (support vector machine), BRNN (Bayesian regularized neural network), GBM (gradient boost model), GAM (generalized additive model), and LASSO (least absolute shrinkage and selection operator); a tested by Windows 10 system in 3.4 GHz computer with 16 GB of RAM.

Machine learning
Which models are usually used for exposure assessment?

- Linear regression.
- Generalized linear regression or generalized additive regression.
- Mixed effect model.
- Bayesian spatiotemporal model.
- Geospatial model, e.g., Kriging, inverse distance weighting.
- Chemical transport model.
- Machine learning and deep learning
Machine learning for air pollution exposure assessment

- How to choose machine learning models? Which one is best?
- How to apply machine learning in environmental exposure assessment easily?
- Do ensemble machine learning models perform better than single machine learning model?
### Table 2. PM$_{2.5}$ prediction performances of DEML model and five benchmark models from 2015 to 2019 in Italy.

<table>
<thead>
<tr>
<th>Year</th>
<th>Measurement</th>
<th>GBM</th>
<th>SVM</th>
<th>RF</th>
<th>XGBoost</th>
<th>SL$^a$</th>
<th>DEML$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>$R^2$</td>
<td>0.69</td>
<td>0.79</td>
<td>0.85</td>
<td>0.81</td>
<td>0.85</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>RMSE (μg/m$^3$)</td>
<td>9.25</td>
<td>6.42</td>
<td>6.49</td>
<td>7.23</td>
<td>6.47</td>
<td>5.54</td>
</tr>
<tr>
<td>2016</td>
<td>$R^2$</td>
<td>0.72</td>
<td>0.80</td>
<td>0.84</td>
<td>0.81</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>RMSE (μg/m$^3$)</td>
<td>7.74</td>
<td>6.51</td>
<td>5.84</td>
<td>6.33</td>
<td>5.82</td>
<td>5.18</td>
</tr>
<tr>
<td>2017</td>
<td>$R^2$</td>
<td>0.74</td>
<td>0.81</td>
<td>0.85</td>
<td>0.81</td>
<td>0.85</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>RMSE (μg/m$^3$)</td>
<td>8.20</td>
<td>7.19</td>
<td>6.41</td>
<td>7.09</td>
<td>6.38</td>
<td>5.37</td>
</tr>
<tr>
<td>2018</td>
<td>$R^2$</td>
<td>0.70</td>
<td>0.78</td>
<td>0.86</td>
<td>0.82</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>RMSE (μg/m$^3$)</td>
<td>7.44</td>
<td>6.22</td>
<td>5.18</td>
<td>5.69</td>
<td>5.13</td>
<td>4.43</td>
</tr>
<tr>
<td>2019</td>
<td>$R^2$</td>
<td>0.68</td>
<td>0.76</td>
<td>0.84</td>
<td>0.79</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>RMSE (μg/m$^3$)</td>
<td>7.34</td>
<td>6.42</td>
<td>5.13</td>
<td>5.78</td>
<td>5.12</td>
<td>4.55</td>
</tr>
<tr>
<td>Total</td>
<td>$R^2$</td>
<td>0.51</td>
<td>0.76</td>
<td>0.83</td>
<td>0.70</td>
<td>0.83</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>RMSE (μg/m$^3$)</td>
<td>10.4</td>
<td>7.42</td>
<td>6.23</td>
<td>8.20</td>
<td>6.23</td>
<td>5.38</td>
</tr>
</tbody>
</table>
Deep Ensemble Machine Learning Model

The advantages:
1. Have high prediction performance
2. Avoid over-fitting through cross-validation analysis
3. Set the optimal non-negative weight for each base-model/meta-model
4. Minimizes the extent to the empirical experience in select models
5. Assessed and compared models’ results directly

The disadvantage:
1. Be cautious to select features
2. Be sensitive to missing values
3. Need more time to run big data
The role of satellite data
The role of satellite data in air pollution exposure assessment

Figure 4. The PM$_{2.5}$ prediction performance of the DEML models with and without AOD as a predictor from 2015–2019 in Italy. The x-axis indicates the observed daily PM$_{2.5}$ in the monitor stations; the y-axis indicates the estimated PM$_{2.5}$ by the DEML model. The points represent the corresponding PM$_{2.5}$ for both observed and predicted values. The solid line represents a regression line for the observed and predicted PM$_{2.5}$ by using the simple linear regression. $R^2$ is the coefficient of determination for the unseen independent data. (A) The DEML prediction model including AOD. (B) The DEML prediction model without AOD. Note: DEML, the three-stage stacked deep ensemble machine learning method; PM$_{2.5}$, particulate matter with aerodynamic diameter <2.5 μm; RMSE, the root mean square error (micrograms per cubic meter).
The role of satellite data in air pollution exposure assessment

- In following scenario, we might not need satellite data:
  1. Have enough observed air pollution data and predictors (correlated with satellite data) in a specific region; and
  2. Don’t predict air pollution in the locations far away from the training region; and
  3. Don’t predict air pollution in the period outside the training period.

- Correspondingly, in following scenarios, we need satellite data:
  1. Have limited observed air pollution data and predictors in a specific region; or
  2. Predict air pollution in the locations far away from the training region; or
  3. Predict air pollution in the period outside the training period.
Wildfire smoke exposure assessment
Wildfire smoke exposure assessment
Wildfire smoke exposure assessment

- Ground-based PM$_{2.5}$
- Satellite remote sensing products
- Meteorology (temperature, humidity, wind speed, rainfall, air pressure, sunshine duration)
- Land use information (land cover, vegetation, road types)
- Others (elevation, fire, population density)

Mixed ensemble machine learning
- Pollution emission inventory
- Chemical transport model

Predicted daily concentrations of bushfire PM$_{2.5}$
Wildfire smoke exposure assessment

Annual average PM$_{2.5}$ (μg/m$^3$)

- < 15
- 15 to < 30
- 30 to < 45
- 45 to < 60
- 60 to < 75
- ≥ 75
Estimating daily wildfire-related PM$_{2.5}$ (µg/m3) 2000-2016
Chen et al. Mortality risk attributable to wildfire-related PM2.5 pollution: a global time series study in 749 locations. The Lancet Planetary Health.
Challenges in estimating wildfire smoke concentrations

1. Many studies just simply use total concentrations of air pollutants during fire period as wildfire-related concentrations.

2. Low spatial resolution.

3. It is hard to directly validate wildfire smoke concentration with the measured air pollution data.
Opportunities
Available big data (observed data including those from low cost sensors, remote sensing, weather data), makes it possible to perform accurate prediction.

Machine learning /deep learning technologies provides better predication performance than traditional models.

High performance computer / and cloud analysis are available to perform big data analysis.
Thank you!

Improving the health of populations in a changing and inequitable world