IMPROVING POLLUTION SOURCE RESOLUTION FOR REAL TIME LOW COST SENSORS USING WIDELY AVAILABLE DATA RESOURCES
A PROOF OF CONCEPT

L. Drew Hill (presenting), Ramboll, Environment & Health, San Francisco, USA
Ajay Pillarisetti, Division of Environmental Health Sciences, University of California, Berkeley, USA
Kirk R Smith, Division of Environmental Health Sciences, University of California, Berkeley, USA
Shari Libicki, Ramboll, Environment & Health, San Francisco, USA
RAMBOLL IN BRIEF

- Independent engineering and design consultancy and provider of management consultancy
- Founded 1945 in Denmark
- 14,000 experts
- Close to 300 offices in 35 countries
- Particularly strong presence in the Nordics, the UK, North America, Continental Europe, Middle East and Asia Pacific
- Owned by Rambøll Fonden

Services across the markets:
- Buildings
- Transport
- Planning & Urban Design
- Water
- Environment & Health
- Energy
- Management Consulting
WHAT AFFECTS THE RELATIONSHIP BETWEEN SENSOR READINGS AND ACTUAL CONCENTRATIONS? (PM2.5, OPTICAL)

Inter-device hardware inconsistencies

Environmental factors, cross-sensitivity
- Temperature, relative humidity

Aerosol properties
- Distributions of size and shape
- Aerosol refractive index
- Particle density

adapted from Litton et al 2004
Machine Learning (ML)

- Very good at uncovering, assessing hidden and complex relationships
- Until very recently, the domain of mathematicians and computer scientists
- Computing advances, open source programming have made ML and Ensemble methods accessible to (more of) the general public
- One of the most important aspects of ML: picking the right variables
- ML is now the domain of subject matter experts (like us!).
Specific makeup of local point, area sources

Traffic
- Time of day: Fraction of total ambient aerosols coming from mobile vs. point sources
- Ratio of diesel to non-diesel
- Ratio of clunkers to ... not clunkers

Environmental phenomena, like wild fires
- Intermittent source
- Produce aerosols of size, shape, refractive index different from those of common urban sources

Meteorology
- Wind direction, speed
  - Regional and local transport
  - Determines upstream sources, dilution
- Precipitation, fog
- Air pressure
WHAT INFLUENCES THESE FACTORS?

Specific makeup of local point, area sources

Traffic
- Time of day: Fraction of total ambient aerosols coming from mobile vs. point sources
- Ratio of diesel vs. non-diesel
- Ratio of clunkers vs. non-clunkers

Environmental phenomenon, e.g., forest fires
- Intermittent source
- Produce aerosols of size, shape, and refractive index different from those of traffic, industrial sources

Meteorology (regional and local transport)
- Wind direction, speed
  - Determines upstream sources, dilution
- Precipitation, fog
USE PUBLIC DATA SOURCES, ADVANCED STATISTICS TO ASSESS AND EXPLOIT CHANGES IN THESE FACTORS RELEVANT TO SENSOR RESPONSE

raw output + some initial calibration (optional) → assess sensor-relevant changes in aerosol properties → data that can be applied to improve sensor utility and the actionability of sensor output
PROOF OF CONCEPT – METHODS

- Plantower sensor data (5 min.) from 5 Clarity Node devices throughout N. California, provided by Clarity
  - Concentration estimates of PM10, PM2.5, PM1.0; temperature; relative humidity
- Collocated with regulatory-grade monitors February – August 2018
PROOF OF CONCEPT – METHODS

Mean\textsubscript{Ref}: 9.0 \, \text{ug/m3}

\sigma_{\text{Ref}}: 8.1 \, \text{ug/m3}

Mean\textsubscript{CN\_raw}: 7.6 \, \text{ug/m3}

\sigma_{\text{CN\_raw}}: 13.0 \, \text{ug/m3}

Reference = 5.0 + 0.52 (‘Raw’ Sensor Estimate)
(uncalibrated) Clarity Output: Reference, by unit

- Variation within units over time
- Variation between units

Overall, the ratio observed is not steady over the assessment period ($\sigma: 1.5$)

![Graph showing ratio by clarity node (also location)]
Concurrent data collected from publicly accessible sources:

- Meteorology (3 closest NOAA ISD-listed stations to each location)
Concurrent data collected from publicly accessible sources:

- Meteorology (3 closest NOAA ISD-listed stations to each location)
- Hourly average PM2.5 concentrations from BAAQMD, SJVAPCD sites (excluding those used in colocation)

https://www.arb.ca.gov/aqmis2/aqdselect.php
Concurrent data collected from publicly accessible sources:

- Meteorology (3 closest NOAA ISD-listed stations to each location)
- Hourly average PM2.5 concentrations from BAAQMD, SJVAPCD sites (excluding those used in colocation)
- Daily indicator of nearby wildfires (> mid-March)
  - ABBA, geosphere package (75 km radius)
PROOF OF CONCEPT – METHODS

Machine Learning (ML), Ensemble Methods

1. Deep Neural Net
   - Multi-layer, feed-forward perceptron
   - 18710 data points, 126 covariates (~ 2.4 million cells)
   - 90%/10% cross validation

2. A ensemble of
   - Random Forests
   - Support Vector Machines
   - GLM, GLM net
   - Ultimate sample size: 5586 data points, 66 covariates (~ 370,00 cells)
   - 10-fold cross validation

\[ \varphi = \frac{\text{Raw Clarity PM2.5 Estimate (ug/m3)}}{\text{Reference PM2.5 Value (ug/m3)}} \]
PROOF OF CONCEPT – RESULTS

- Deep Neural Network:
  - Moderate predictive power, well-fit, moderate error
  - Variable importance: nearby NOAA and regulatory monitor data show high importance

<table>
<thead>
<tr>
<th></th>
<th>Mean φ observed</th>
<th>Mean φ predicted</th>
<th>r² Obs. Vs. Pred</th>
<th>β² Obs. Vs. Pred</th>
<th>RMSE validation</th>
<th>RMSE train</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.67 (σ: 1.1)</td>
<td>0.76</td>
<td>~ 0.35</td>
<td>1.17</td>
<td>0.88</td>
<td>1.04</td>
</tr>
</tbody>
</table>

17% underestimation
PROOF OF CONCEPT – RESULTS

• Ensemble (RF, SVM, GLM, GLM net):
  • Low bias, moderate error
    • Strongly predicted ratio as it changed
    • Thus, likely a strong predictor of changes in aerosol properties and potentially nearby source characteristics

<table>
<thead>
<tr>
<th>Mean φ observed</th>
<th>Mean φ predicted</th>
<th>Ensemble Avg. RMSE</th>
<th>$\beta^2$ Obs. Vs. Pred (obs &lt; 7)</th>
<th>adj-$r^2$ Obs. Vs. Pred (obs &lt; 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.71 ($\sigma$: 1.4)</td>
<td>0.66 ($\sigma$: 0.51)</td>
<td>1.60</td>
<td>1.04 (SE: 0.01)</td>
<td>0.48</td>
</tr>
</tbody>
</table>

$\text{Mean } \phi_{\text{observed}} = 0.71$, $\text{Mean } \phi_{\text{predicted}} = 0.66$, $\text{Ensemble Avg. RMSE} = 1.60$, $\beta^2 = 1.04$, adj-$r^2 = 0.48$. 4% underestimation.
PROOF OF CONCEPT – RESULTS

Ensemble (RF, GLM, GLM net, SVM):

- Ratios can be used to reliably produce estimates of true hourly average local PM2.5 mass concentrations
- Low bias across nodes, low/moderate error
- Ratio & Clarity output allowed reliable reconstruction of reference values
  - Better in some nodes than other
INSIGHTS, NEXT STEPS

Using publicly available data, a machine learning-enhanced statistical model can be trained to:

- strongly predict hourly changes in the relationship between sensor output and PM2.5 concentrations
  - Identify key changes in local pollution source contributions, important events
  - account for location-based and inter-unit differences with good accuracy

Such a model leverages and highly relies upon local, sophisticated low-cost sensor output

- Clarity Node provides estimates of PM1 and PM10, allows model to consider changes in size distribution

Such a model can reliably produce estimates of true hourly average local PM2.5 concentrations

Future work should explore the ability of such a model to predict low-cost sensor calibration factors in near real-time (~ hourly)

Future models should explore local traffic data
REFERENCES

THANK YOU!

For providing the Node/FEM colocation datasets.

Collaborator Shari Libicki for good feedback on early drafts, and the organization for allowing me to utilize our resources to pursue this area of work.

Collaborators Ajay Pillarisetti and Kirk Smith, who’ve tolerated years of brainstorming and provided good comments.
PM2.5 – CURRENT MEASUREMENT TECHNIQUES

- Environment-based (not exposure)
- Provide daily or hourly average estimates
- Reliable monitors are few and far between
  - In US: about 1 per 350,000 people (USEPA 2016 via Hill 2017)
  - Air pollution concentrations are variable at a hyper-local level
  - Potential for exposure misclassification
  - Suboptimal actionability of resulting information
- Monitors are expensive (thousands to tens of thousands of $), bulky, noisy, delicate, require expertise to operate

Source: EPA.gov
TYPICAL PM2.5 SENSOR CALIBRATION

Lower cost for a reason

- Low-cost hardware
  - Signal can be noisy
  - Getting better every month!

- Measure proxy of metric of interest: mass concentration (μg/m³)
  - Light scattering
    - High correlation with mass concentration for a given aerosol, device, and software

Pillarisetti et al 2017

Gravimetrically-Adjusted DustTrak
TYPICAL PM2.5 SENSOR CALIBRATION

Lower cost for a reason

• Low-cost hardware
  • Signal *can* be noisy
  • Getting better every month!

• Measure proxy of metric of interest: mass concentration (ug/m3)
  • Light scattering
    • High correlation with mass concentration for a given *device*, software setup, and aerosol.

Pillarisetti et al 2017
TYPICAL PM2.5 SENSOR CALIBRATION

Lower cost for a reason

- Low-cost hardware
  - Signal can be noisy
  - Getting better every month!

- Measure proxy of metric of interest: mass concentration (ug/m3)
  - Light scattering
    - High correlation with mass concentration for a given device, software setup, and aerosol.

Pillarisetti et al 2017
TYPICAL PM2.5 SENSOR CALIBRATION

Lower cost for a reason

- Low-cost hardware
  - Signal can be noisy
  - Getting better every month!

- Measure proxy of metric of interest: mass concentration (ug/m3)
  - Light scattering
    - High correlation with mass concentration for a given device, software setup, and aerosol.
    - Highly sensitive to changes in aerosol properties (e.g. size distribution, refractive index, shape)

- Require extensive calibration
  - Best: real time colocation with FEM/FRM
  - Better, uncommon: real time dynamic calibration via inter and intra-device comparison
  - Common: lab co-location vs. specific aerosol mixture