Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles

Brian Reich
Department of Statistics, NCSU

Department of Biological and Agricultural Engineering, NCSU, 2020
Collaborators

Yawen Guan, Margaret Johnson, Matthias Katzfuss, Elizabeth Mannshardt, Kyle Messier and Joon Jin Song
Classic air pollution monitoring scheme

- Since the clean air act of 1970, the US EPA has been monitoring and regulating air pollution

- They rely on a small number of stationary monitors collecting daily data

- These data are used to
  - Uphold national air quality standards
  - Follow trends over space and time
  - Study health effects of air pollution
Stationary N0₂ monitors in NC¹

¹https://deq.nc.gov/about/divisions/air-quality/air-quality-monitoring
US average NO$_2$ by year$^2$

NO$_2$ Air Quality, 1980 - 2018
(Annual 98th Percentile of Daily Max 1-Hour Average)
National Trend based on 22 Sites

1980 to 2018 : 61% decrease in National Average

$^2$https://www.epa.gov/air-trends/nitrogen-dioxide-trends
Nonattainment counties

Counts Designated "Nonattainment"
for Clean Air Act’s National Ambient Air Quality Standards (NAAQS) *

Legend **
- County Designated Nonattainment for 6 NAAQS Pollutants
- County Designated Nonattainment for 5 NAAQS Pollutants
- County Designated Nonattainment for 4 NAAQS Pollutants
- County Designated Nonattainment for 3 NAAQS Pollutants
- County Designated Nonattainment for 2 NAAQS Pollutants
- County Designated Nonattainment for 1 NAAQS Pollutant

3https://www3.epa.gov/airquality/greenbook/mapnpoll.html
Epidemiological studies

- A common approach is to regress daily health outcomes onto city average daily air pollution\(^4\)

- A more expensive approach is to estimate individual exposure for a small number of subjects\(^5\)

- Health effects are also studied using controlled experiments\(^6\)

---

\(^4\) e.g., Dominici et al, JAMA, 2006
\(^5\) e.g., Larkin and Hystad, CEHR, 2017
\(^6\) https://www.nap.edu/catalog/24618/controlled-human-inhalation-exposure-studies-at-epa
Epidemiological studies

These studies have established links between air pollution and several adverse health outcomes including:

- Cardiovascular diseases
- Respiratory diseases
- Cancer
- Pre-term birth

An estimated 4.2 million premature deaths globally are linked to ambient air pollution\(^7\)

\(^7\)https://www.who.int/airpollution/ambient/health-impacts/en/
New sources of monitoring data

- Air pollution epidemiology relies on a few stationary monitors per city.
- The field is undergoing a paradigm shift due to fine-resolution mobile monitors.
- We analyze data collected from a car driving around the city and continuously measuring air pollution.
- Some cities now have thousands of low-cost stationary sensors.
- Phone apps are under development.
Data from Google StreetView cars

Photo: Apte (2017)
Mobile data: Oakland NO$_2$

- Two cars were deployed from June 2015 to May 2016

- Starts on weekdays at $\approx 9$am, drove $\approx 6$-8 hours each day

- Measurements are taken at roughly every second.

- Large missing data (car maintenance, sensor failure, etc.)
Example daily observations of $\log(\text{NO}_2)$

- Car A and B drove from 8am - 2pm
- 12,389 observations, covered less than a third of Oakland
Objectives

Develop a statistical model for real-time, high resolution forecasting of air pollution

- Can we develop accurate maps for epi studies?
- Can we make forecasts that help people avoid exposure?
- How far ahead can we reasonably forecast air pollution?
- How many cars should be deployed for future studies?
- Are cars more efficient than stationary monitors?
Statistical challenges

- Data are large \((n \approx 900,000 \text{ observations})\)
- Data are streaming
- Data are extremely sparse in space and time
- Data are noisy and subject to outliers
- Process is likely dynamic and nonstationary
We took temporal block medians to dampen effects of extremes
Example landuse covariates

Road Type

City of Oakland Zoning

Highway • Major • Residential

Commercial • Industrial • Residential • NA
Principal components of landuse variables

PC 1

PC 3

PC 4

PC 5
Non-spatial landuse regression

Let $Y_t(s)$ be the log(NO$_2$) at time $t$ and location $s$

$$Y_t(s) = X_t(s)^T \beta + \epsilon_t(s), \quad \epsilon_t(s) \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$$

where $X_t(s)$ contains

- The first seven PCs
- Four trig functions for hourly diurnal cycle
- Interactions between the PCs and trig functions

$R^2 \approx 0.16$ and residuals are correlated
Results from landuse regression

Results from landuse regression
We add a spatiotemporal process to capture dependence

\[ Y_t(s) = X_t(s)\beta + \eta_t(s) + \epsilon_t(s), \quad \epsilon_t(s) \sim \text{iid} \mathcal{N}(0, \tau^2) \]

The Gaussian process \( \eta \) has covariance

\[ \text{Cov} \{\eta_t(s), \eta_{t'}(s')\} = \sigma^2 \exp \left\{ -\sqrt{||s - s'||^2 / \rho + |t - t'|^2 / \phi} \right\} \]

The range parameters \( \rho \) and \( \phi \) determine the extent of spatial and temporal dependence

The covariance parameters to be estimated are \( \theta = \{\sigma^2, \tau^2, \rho, \phi\} \)
Maximum likelihood analysis is impossible.

The likelihood depends on function of the spatial covariance matrix, which is huge \((n \times n)\).

Overcoming this computational bottleneck is one of the main challenges in spatial statistics.

There are now many approaches\(^9\).

\(^9\)e.g., Heaton et al, 2019, JABES
Computation

- We use the Veccia approximation to estimate the covariance parameters

- Training the model based on the joint distribution of all the observations is too slow

- Instead we regress the current observation $Y_i$ onto the recent past $\mathcal{N}_i$

  - If $\mathcal{N}_i = \{Y_1, \ldots, Y_{n-1}\}$ this is exact and slow

  - If $\mathcal{N}_i \subset \{Y_1, \ldots, Y_{n-1}\}$ this is approximate and fast
Computation

- The neighboring sets are observations in the recent past
  \[ N_j = \{ \text{obs between } l \text{ and } l + u \text{ minutes prior to obs } i \} \]
- The results are insensitive to \( u \) so we set \( u = 60 \)
- Taking \( l = 0 \) gives the best approximation to the likelihood
- Often it is better to include some distant neighbors\(^\text{10}\)

\(^{10}\)Gramacy and Apley, 2015, Technometrics
Cross validation design

- Train the model using both cars’ data prior to time $t$

- Predict the value for both cars at time $t + h$ for $h \in \{5, 15, 60\}$ minutes

- We compare the non-spatial (“X”), spatial (“S”) and spatiotemporal (“ST”) methods

- We also compare fitting the model using one car’s data and predicting the other car (“Car AB”)

- Methods are compared using the correlation between observed and predicted

- Repeat using raw data (1 sec) and 1-minute block medians
Prediction correlation using 1 sec block medians

<table>
<thead>
<tr>
<th>Model</th>
<th>I</th>
<th>5 mins</th>
<th>15 mins</th>
<th>60 mins</th>
<th>Car AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
<td>0.08</td>
</tr>
<tr>
<td>S</td>
<td>-</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.09</td>
</tr>
<tr>
<td>ST</td>
<td>0</td>
<td>0.45</td>
<td>0.25</td>
<td>0.18</td>
<td>0.09</td>
</tr>
<tr>
<td>ST</td>
<td>5</td>
<td>0.58</td>
<td>0.36</td>
<td>0.28</td>
<td>0.10</td>
</tr>
<tr>
<td>ST</td>
<td>15</td>
<td>0.57</td>
<td>0.36</td>
<td>0.31</td>
<td>0.10</td>
</tr>
<tr>
<td>ST</td>
<td>60</td>
<td>0.55</td>
<td>0.38</td>
<td>0.28</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Spatiotemporal methods for mobile monitoring data
## Prediction correlation using 1 min block medians

<table>
<thead>
<tr>
<th>Model</th>
<th>l</th>
<th>5 mins</th>
<th>15 mins</th>
<th>60 mins</th>
<th>Car AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
<td>0.19</td>
</tr>
<tr>
<td>S</td>
<td>-</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.21</td>
</tr>
<tr>
<td>ST</td>
<td>0</td>
<td>0.59</td>
<td>0.44</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>ST</td>
<td>5</td>
<td>0.64</td>
<td>0.56</td>
<td>0.46</td>
<td>0.26</td>
</tr>
<tr>
<td>ST</td>
<td>15</td>
<td>0.64</td>
<td>0.56</td>
<td>0.45</td>
<td>0.26</td>
</tr>
<tr>
<td>ST</td>
<td>60</td>
<td>0.63</td>
<td>0.55</td>
<td>0.45</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Estimated diurnal trends

9:00

12:00

15:00

18:00
## Correlation parameter estimates

<table>
<thead>
<tr>
<th>Block size</th>
<th>$l$</th>
<th>Spatial $R^2$ ratio</th>
<th>Spatial range (km)</th>
<th>Temporal range (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 sec</td>
<td>0</td>
<td>1.00</td>
<td>0.95</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.63</td>
<td>4.82</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.54</td>
<td>3.95</td>
<td>9.53</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.77</td>
<td>1.38</td>
<td>2.32</td>
</tr>
<tr>
<td>1 min</td>
<td>0</td>
<td>0.92</td>
<td>3.52</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.64</td>
<td>5.21</td>
<td>9.24</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.57</td>
<td>5.43</td>
<td>28.72</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.60</td>
<td>3.62</td>
<td>4.19</td>
</tr>
</tbody>
</table>
15 minutes ahead forecasts of NO$_2$

- Forecast for 15:00 using the data from 13:45 to 14:45 on May 5, 2016

- As expected, standard errors are lowest where data has been obtained most recently from the two cars.
Dynamic model

- We envision the model being refitted periodically to adapt to evolving environmental, traffic and emissions patterns.

- We refit the model in a sliding window of training data to study changes in parameter estimates and performance.

- For each week from 12/07/2015 to 05/13/2016, we use the data from the previous \( w \) weeks to train the model.

- We compute 15-minute ahead prediction mean squared error for that week.

- We use \( l = 60 \) and 15-second block median data.
Dynamic model – Spatial range estimates ($w = 21$)

Blue dots are estimates, black dots are bootstrap samples
Dynamic model – Temporal range estimates ($w = 21$)

Blue dots are estimates, black dots are bootstrap samples
Dynamic model – prediction MSE

![Graph showing Dynamic model prediction MSE from Dec 2015 to May 2016 with different line types for 2 Week, 6 Week, 12 Week, 21 Week, and Static models. The graph shows the MSPE (Mean Squared Prediction Error) over time, with peaks and valleys indicating variability in prediction accuracy.]
Network design

How many cars should be deployed? How many fixed-location sensors would provide the same quality of prediction?

Deploy mobile and stationary sensors and simulate data
Network design

![Graph showing the relationship between MSPE and the number of monitors for different methods: Station Forecast, Station Interpolation, Mobile Forecast, Mobile Interpolation. The graph illustrates a decrease in MSPE as the number of monitors increases for all methods.](image-url)
Summary and future projects

▶ Our work\textsuperscript{11} shows that short-term forecasting of air pollution at a high spatial resolution is possible

▶ Future work:
  ▶ Fuse mobile and stationary sensors
  ▶ Model extremes
  ▶ Multi-city analysis
  ▶ Design efficient sampling routes

▶ Works supported by NIH and NSF

▶ THANKS!