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

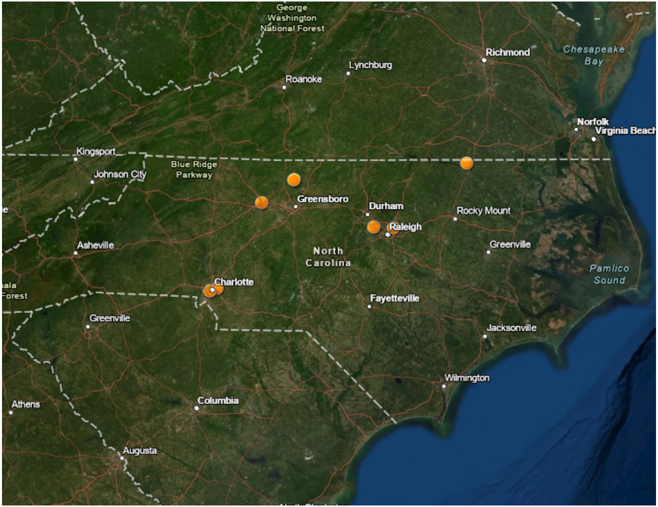
Collaborators

Yawen Guan, Margaret Johnson, Matthias Katzfuss, Elizabeth Mannshardt, Kyle Messier and Joon Jin Song

Classic air pollution monitoring scheme

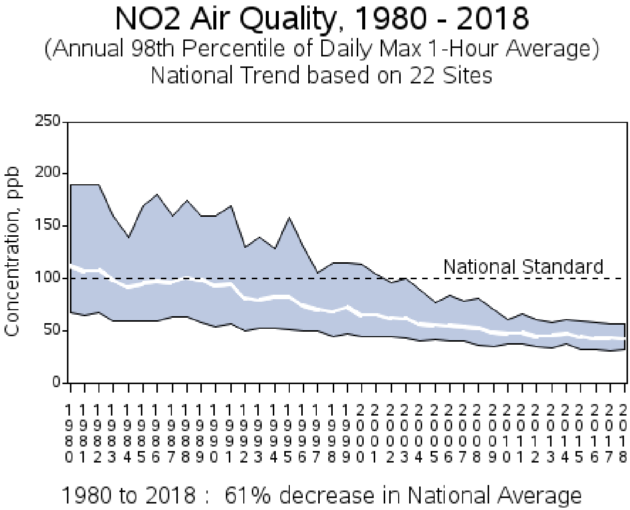
- Since the clean air act of 1970, the US EPA has been monitoring and regulating air pollution
- They rely on a small number of stationary monitors collecting daily data
- These data are used to
  - Uphold national air quality standards
  - Follow trends over space and time
  - Study health effects of air pollution

Stationary $N0_2$ monitors in NC[1]



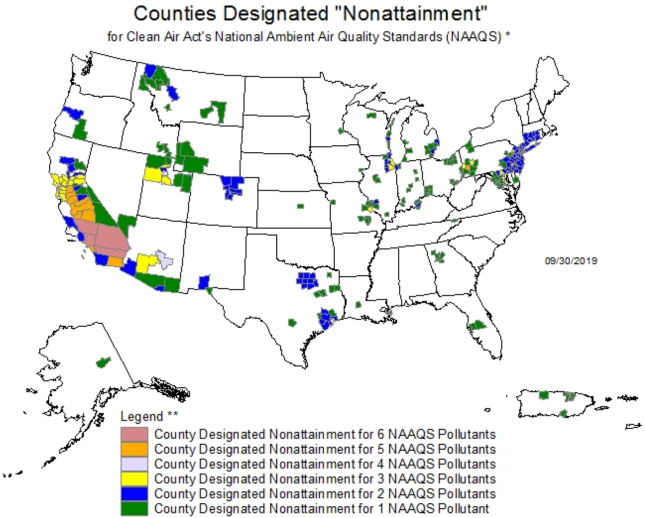
[1] https://deq.nc.gov/about/divisions/air-quality/air-quality-monitoring

US average NO_2 by year[2]



[2] https://www.epa.gov/air-trends/nitrogen-dioxide-trends

Nonattainment counties[3]



[3] https://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

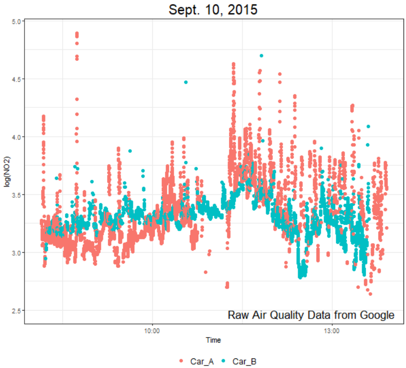
Data from Google StreetView cars[8]

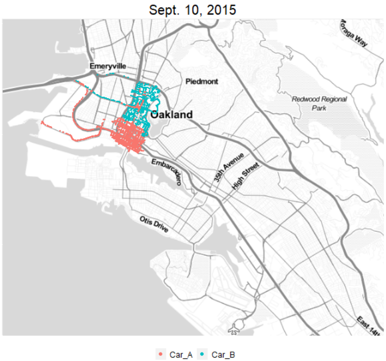
[8] Photo: Apte (2017)

Mobile data: Oakland NO_2

- Two cars were deployed from June 2015 to May 2016
- Starts on weekdays at $\approx$ 9am, 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(NO_2)





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

Objectives

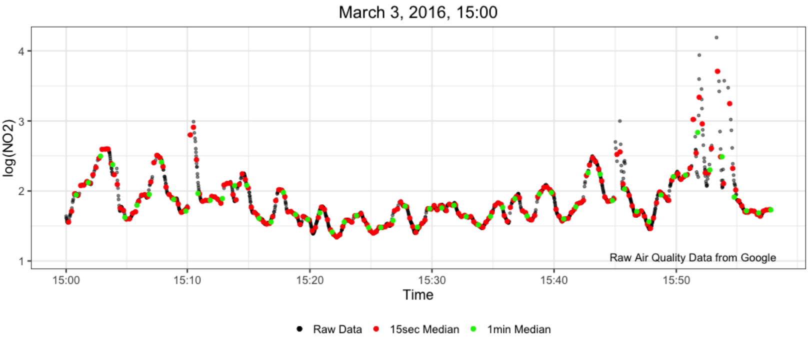
Develop a statistical model for real-time, high resolution forecasting of air pollution

- 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 future studies?
- Are cars more efficient than stationary monitors?

Statistical challenges

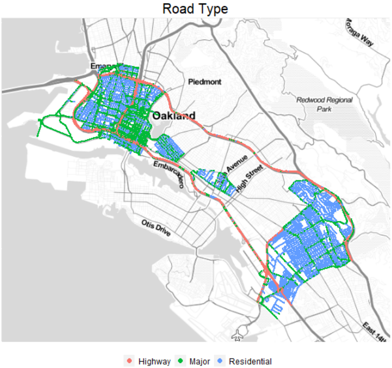
- Data are large ($n \approx$900,000 observations)
- Data are streaming
- Data are extremely sparse in space and time
- Data are noisy and subject to outliers
- Process is likely dynamic and nonstationary

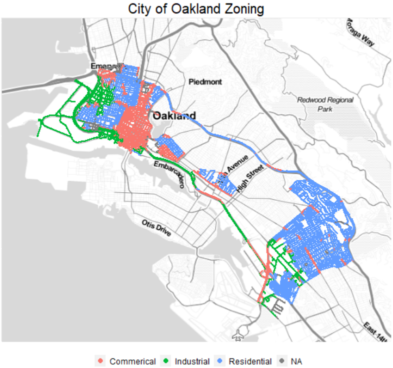
Temporal aggregation



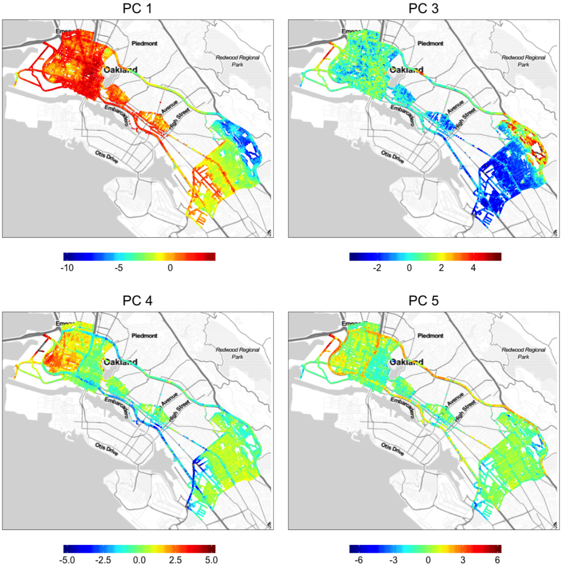
We took temporal block medians to dampen effects of extremes

Example landuse covariates





Principal components of landuse variables



Non-spatial landuse regression

Let $Y_t(\mathbf{s})$ be the log(NO_2) at time t and location $\mathbf{s}$

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

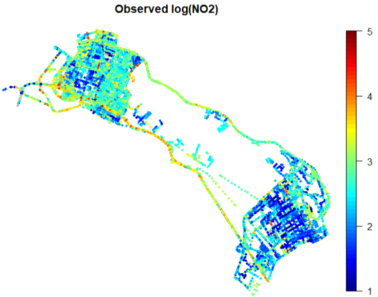
where $\mathbf{X}_t(\mathbf{s})$ contains

- 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

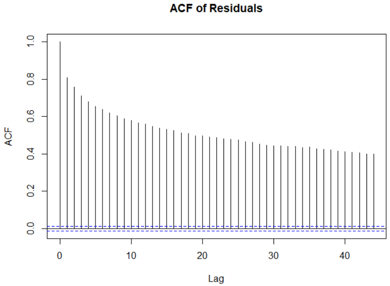
Results from landuse regression

Observed vs. Predicted, Oct. 29, 2015 – Dec. 18, 2015



Results from landuse regression

Observed vs. Predicted, Oct. 29, 2015 – Dec. 18, 2015



Spatiotemporal landuse regression model

- We add a spatiotemporal process to capture dependence

$$Y_t(\mathbf{s}) = \mathbf{X}_t(\mathbf{s})\boldsymbol{\beta} + \eta_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad \epsilon_t(\mathbf{s}) \stackrel{iid}{\sim} N(0, \tau^2)$$

- The Gaussian process η has covariance

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

- The range parameters ρ and ϕ determine the extent of spatial and temporal dependence

- The covariance parameters to be estimated are $\boldsymbol{\theta} = \{\sigma^2, \tau^2, \rho, \phi\}$

Computation

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

Computation

- The neighboring sets are observations in the recent past

$$\mathcal{N}_i = \{\text{obs between } l \text{ and } l+u \text{ minutes prior to obs } i\}$$

- The results are insensitive to u so we set $u = 60$
- Taking $l = 0$ gives the best approximation to the likelihood
- Often it ts better to include some distant neighbors[10]
- We pick l by cross-validation

[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

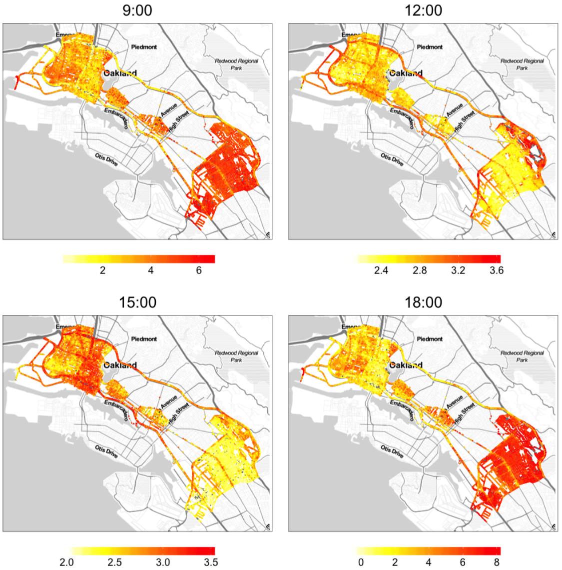
Prediction correlation using 1 sec block medians

		Prediction lag			
Model	l	5 mins	15 mins	60 mins	Car AB
X	-	0.18	0.18	0.18	0.08
S	-	0.27	0.27	0.27	0.09
ST	0	0.45	0.25	0.18	0.09
ST	5	0.58	0.36	0.28	0.10
ST	15	0.57	0.36	0.31	0.10
ST	60	0.55	0.38	0.28	0.09

Prediction correlation using 1 min block medians

		Prediction lag			
Model	l	5 mins	15 mins	60 mins	Car AB
X	-	0.28	0.28	0.28	0.19
S	-	0.34	0.34	0.34	0.21
ST	0	0.59	0.44	0.29	0.26
ST	5	0.64	0.56	0.46	0.26
ST	15	0.64	0.56	0.45	0.26
ST	60	0.63	0.55	0.45	0.26

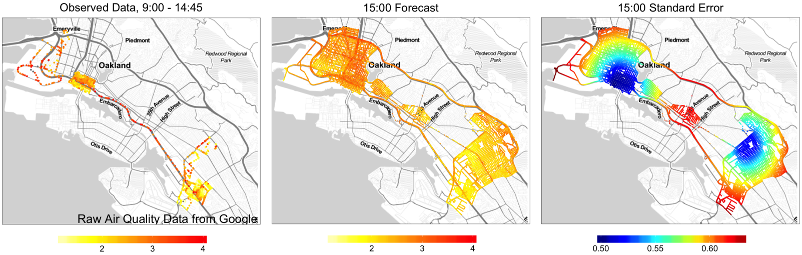
Estimated diurnal trends



Correlation parameter estimates

Block size	l	Spatial R^2 ratio	Spatial range (km)	Temporal range (hr)
1 sec	0	1.00	0.95	0.19
	5	0.63	4.82	3.83
	15	0.54	3.95	9.53
	60	0.77	1.38	2.32
1 min	0	0.92	3.52	0.23
	5	0.64	5.21	9.24
	15	0.57	5.43	28.72
	60	0.60	3.62	4.19

15 minutes ahead forecasts of NO_2

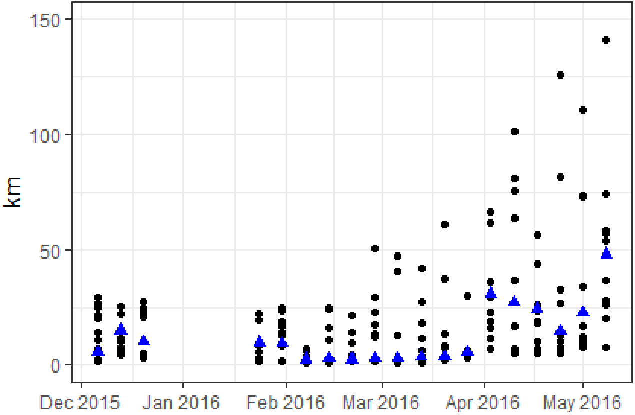


- Forecast for 15:00 using the data from 13:45 to 14:45 on May 5, 2016
- As expected, standard errors are lowest where data has been obtained most recently from the two cars.

Dynamic model

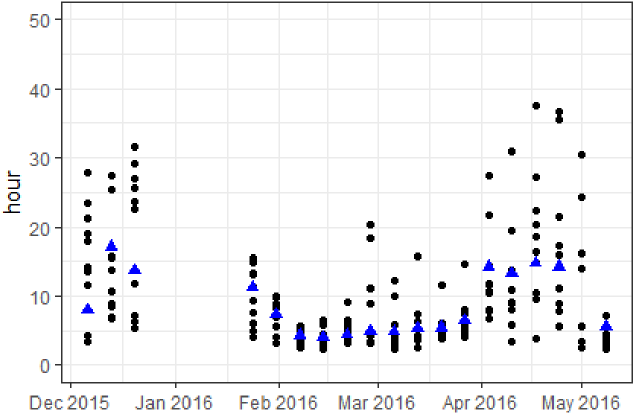
- We envision the model being refitted periodically to adapt to evolving environmental, traffic and emissions patterns
- We refit the model in a sliding window of training data to study changes in parameter estimates and performance
- For each week from 12/07/2015 to 05/13/2016, we use the data from 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 ($w = 21$)



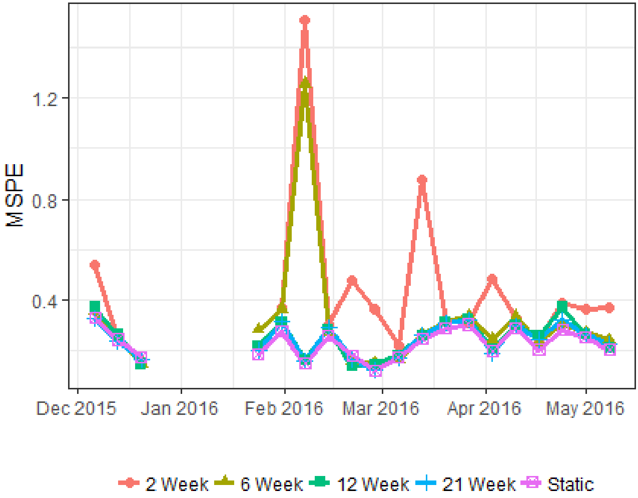
Blue dots are estimates, black dots are bootstrap samples

Dynamic model – Temporal range estimates ($w = 21$)



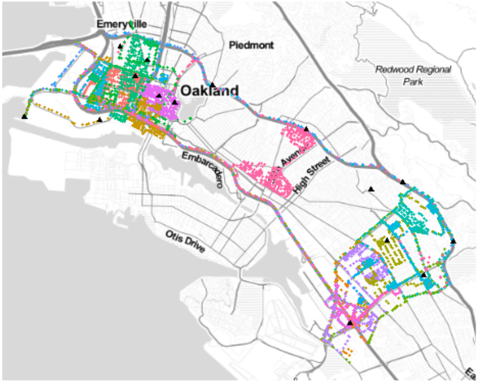
Blue dots are estimates, black dots are bootstrap samples

Dynamic model – prediction MSE



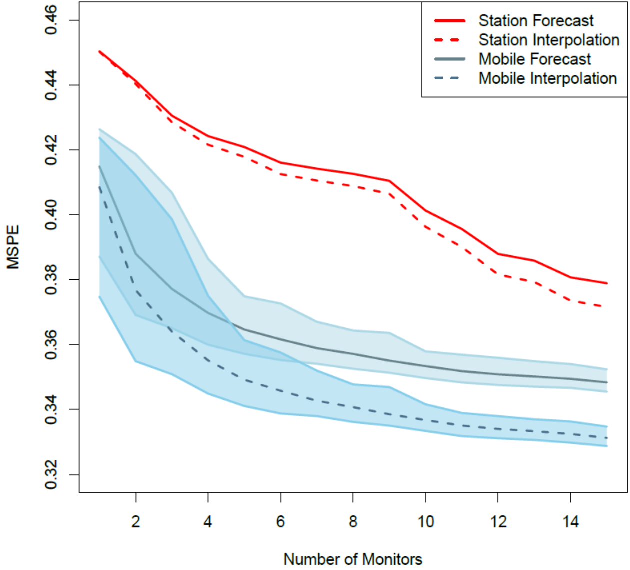
Network design

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



Summary and future projects

- Our work[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!**

[11] Guan et al (2020). Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicle. In press, *Journal of the American Statistical Association*.