Advances in spatial causal inference

Brian Reich

North Carolina State University

University of Minnesota, 2020

Collaborators

Alex Larsen

Yawen Guan

Others: Maggie Johnson, Ana Rappold, Shu Yang and others

Overview

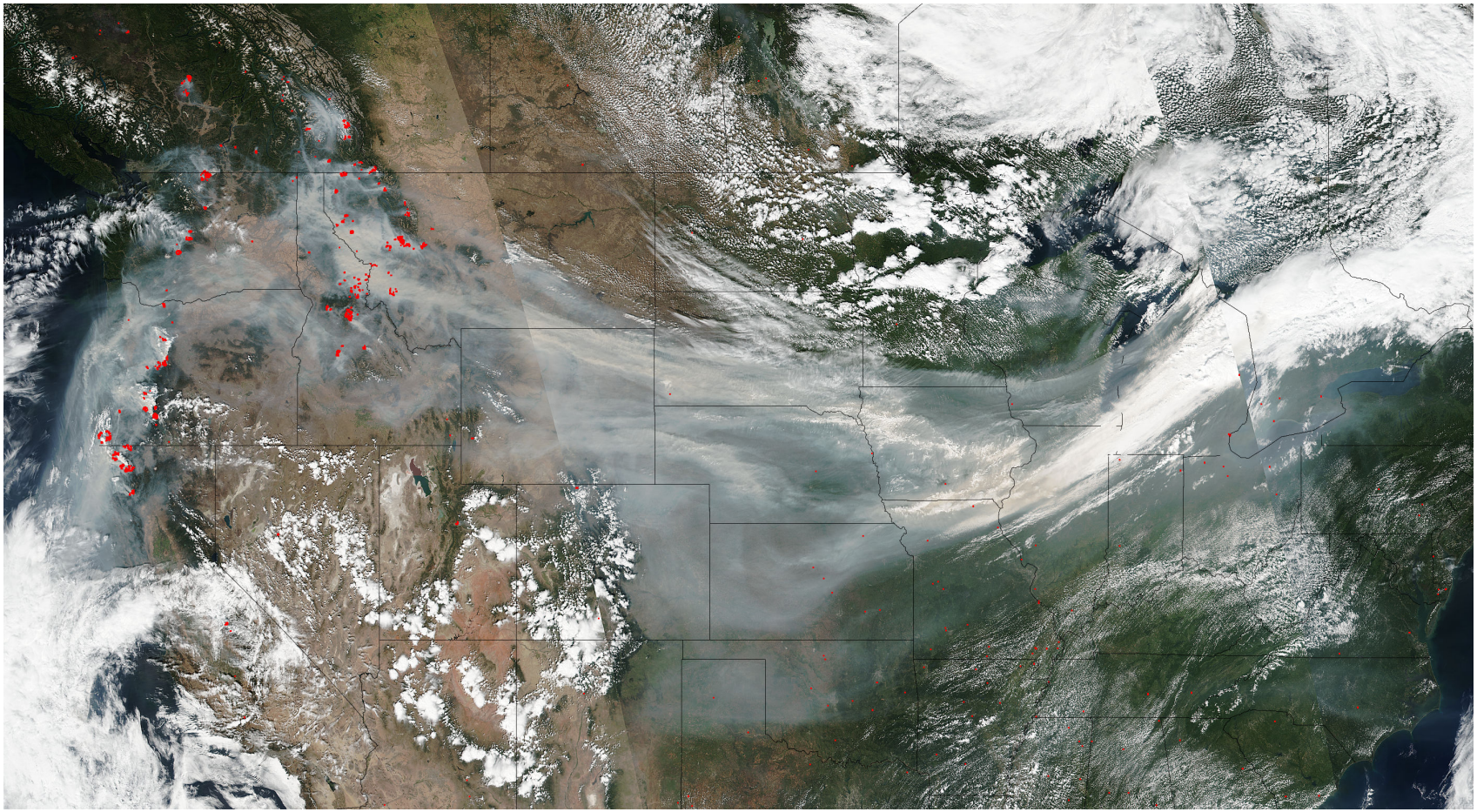
- Ambient air pollution has well-established adverse health effects
- The EPA has been regulating air pollution for many years and it has substantially reduced in the US
- The classic approach relies on a sparse network of stationary monitors
- We'll discuss methods based on two new data streams:
  - Project 1: Causal inference using numerical models to adjust for confounders (Larsen et al, in revision)
  - Project 2: Hyper-local spatiotemporal modeling using mobile monitors (Guan et al, 2020, *JASA*)

Increase in wildfires in the US

- Since the 70's, the rate of large wildfires (1000+ acres) has doubled
- In that time, the rate of very large wildfires (10,000+ acres) has increased fivefold[1]
- This poses an increasing health threat:
  - NPR (9/11/17) *Is All That Wildfire Smoke Damaging My Lungs?*
  - NYT (9/17/17): *As Wildfires Burn West, Ash Rides Wind High Across the U.S.; Large Erratic Blazes are Posing a Bigger Threat to People*

[1] climatecentral.org, 2012

Smoke plumes carry pollution across the continent



Wildfires' contribution to air pollution

- Fine particular matter ($PM_{2.5}$) is a criteria pollutant monitored by the EPA to protect human health
- In the US, emissions of $PM_{2.5}$ from most sources are steadily declining
- Forest fire smoke remains a major contributor, and may increase with climate change, population, and land-use change
- Many studies are now investigating the health effects of fire smoke

Causal analysis using numerical models

- How much $PM_{2.5}$ is causally-attributed to wildfires?
- How much health burden is causally-attributed to wildfires?
- These questions are difficult to answer because only total $PM_{2.5}$ (background + fire) can be measured
- Observational data, long-range transport and confounding environmental factors are also challenges
- We combine numerical models and observational data
- Causal analysis lays bare the assumptions needed to interpret the results causally

Related work on spatial causal inference

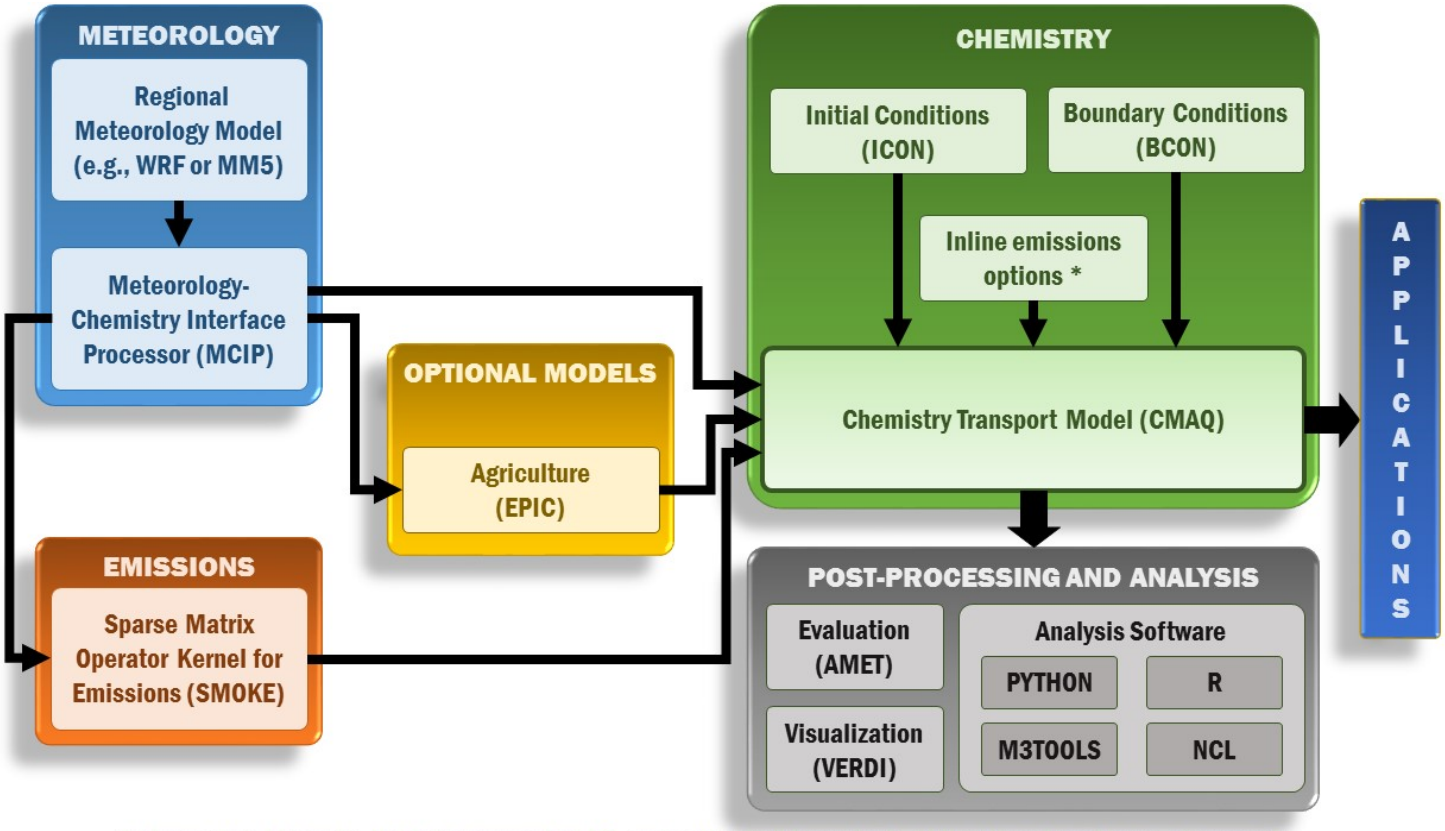
- Downscaler methods (e.g., Berrocal et al[2]) correct for spatially-varying bias in model output – no treatment effects
- Zigler et al have a series of papers[3] on causal effects of air pollution regulation – no numerical models or counterfactuals
- Detection and attribution climate studies[4] run models in different regimes – only observe data in one regime

[2] e.g., Berrocal, Gelfand, and Holland, AOAS, 2010

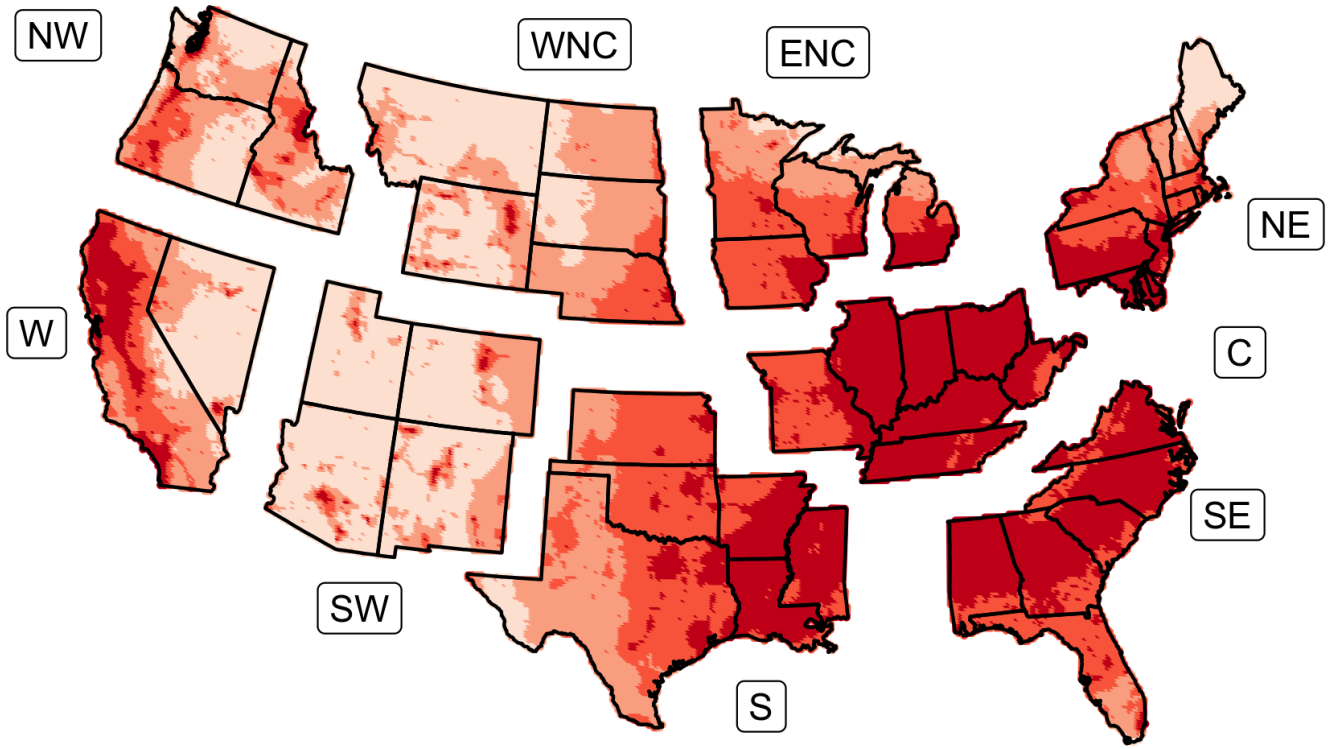
[3] e.g., Zigler, Dominici, Wang, Biostatistics, 2012; Zigler and Dominici, AJE, 2014; Zigler et al, HEI, 2016

[4] e.g., Katzfuss et al, JGR, 2017

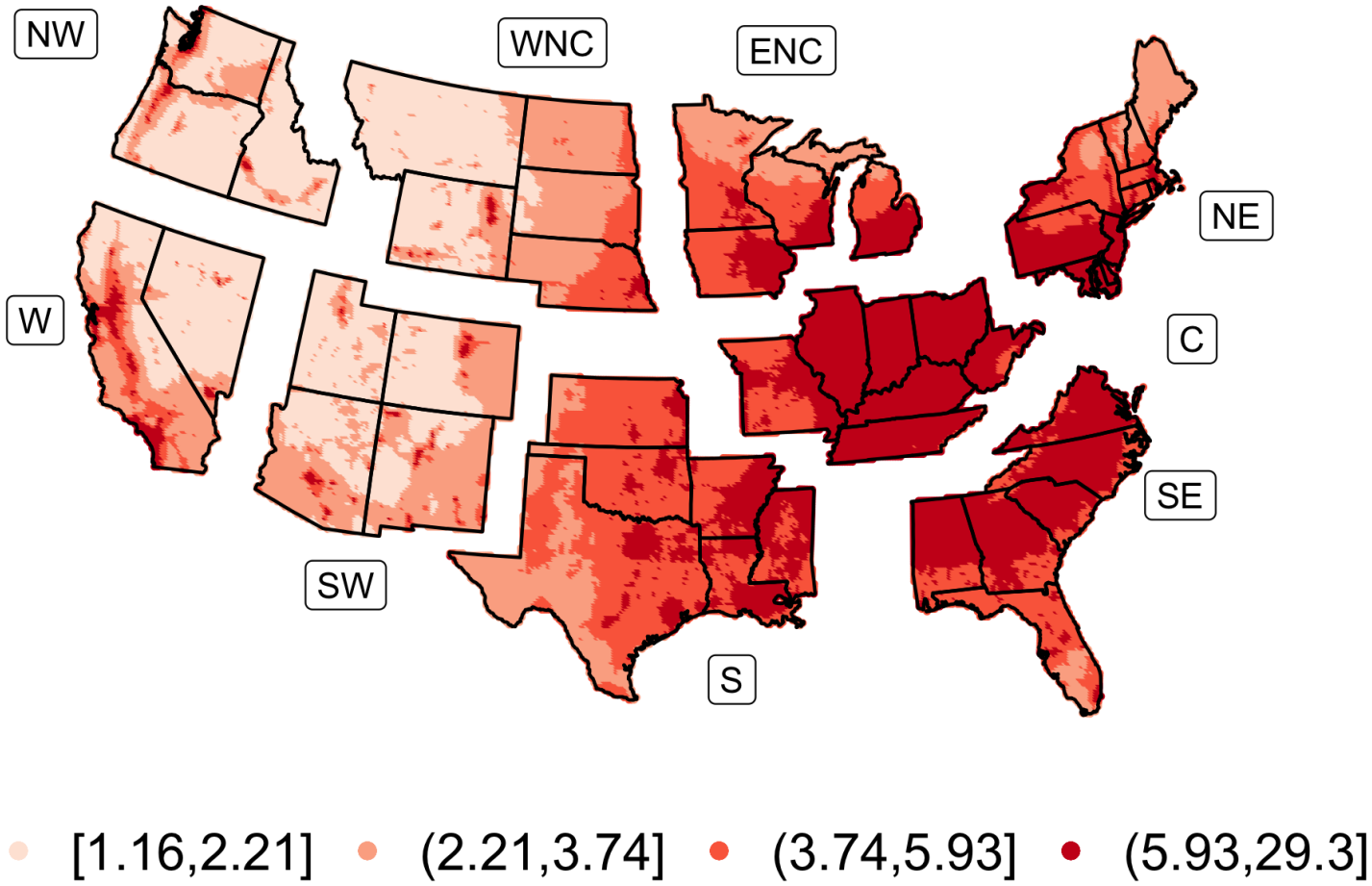
CMAQ: The Community Multiscale Air Quality Modeling System



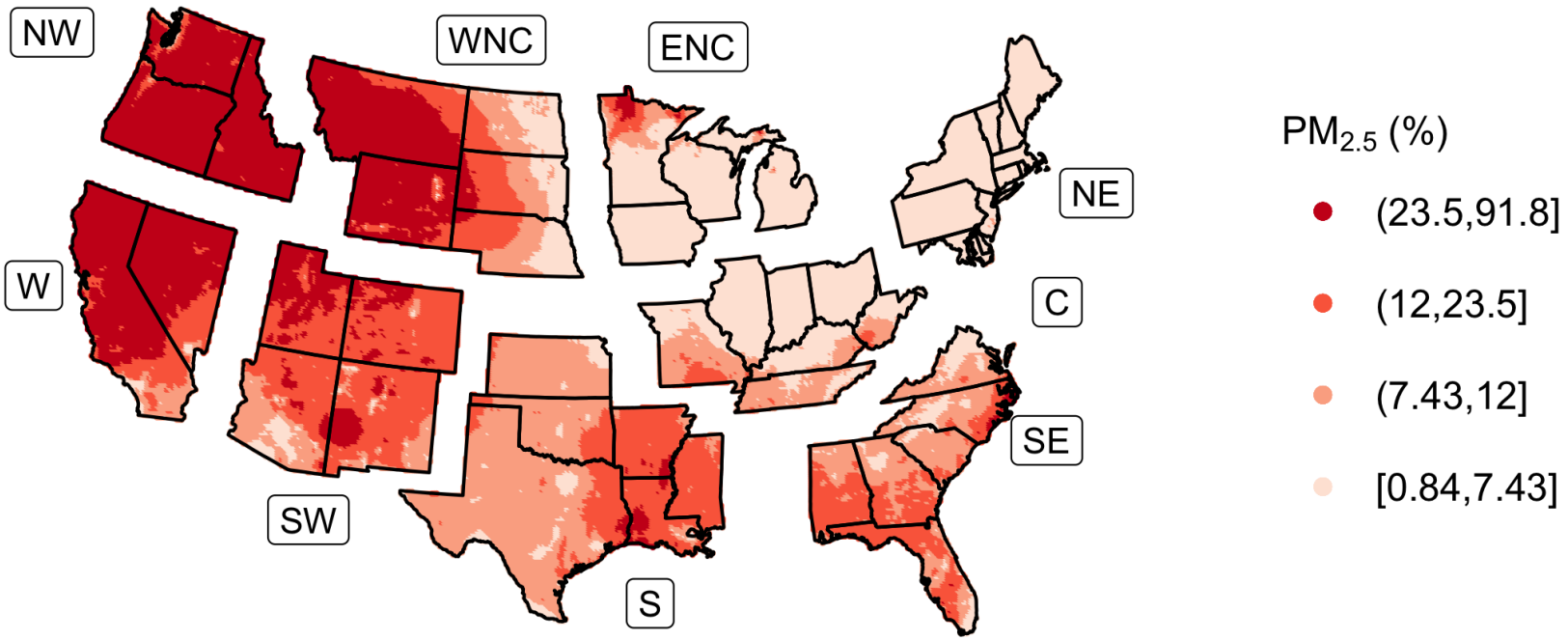
Annual average CMAQ (12km $\times$ 12km)



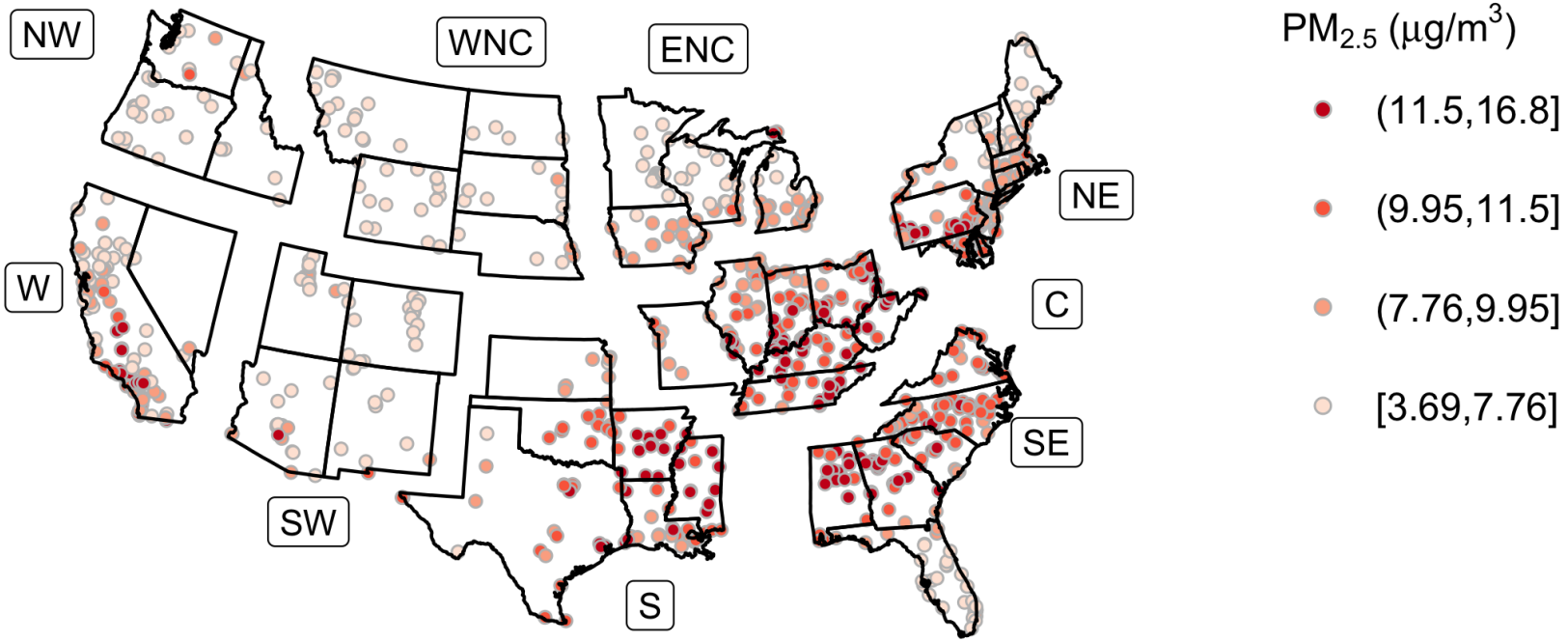
CMAQ run without fires



Difference between the runs (as % of total)

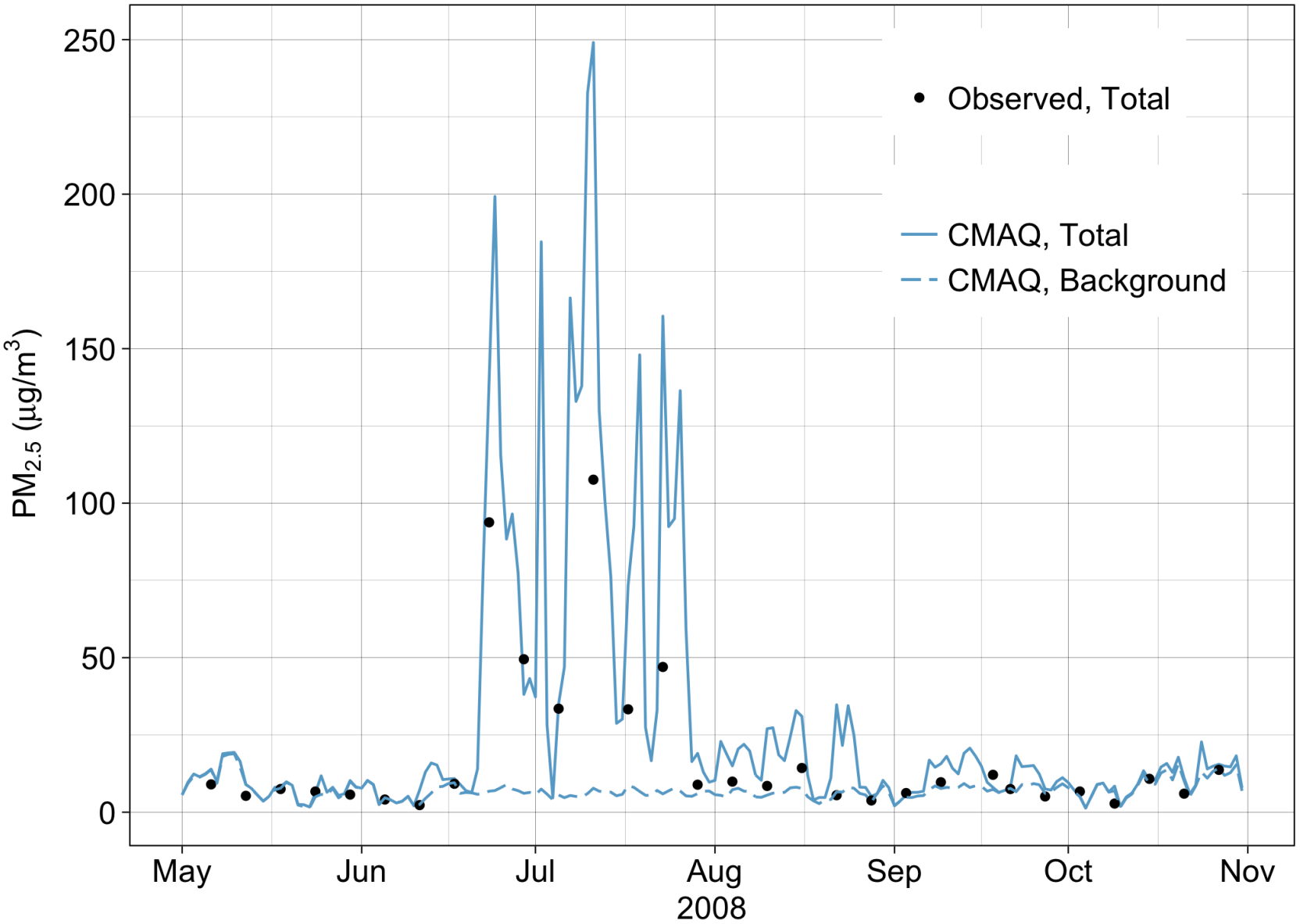


EPA Monitoring Stations (background + fire)



$PM_{2.5}$ is measured every 3-6 days; this is the 2008-2012 average

Time series plot for one site in Northern CA



Data sources and notation

- Monitor data at location s and day t: $Y_t(\mathrm{s})$
- CMAQ non-fire run: $\hat{\theta}_t(\mathrm{s})$
- CMAQ fire run minus non-fire run: $\hat{\delta}_t(\mathrm{s})$
- Other confounders (emissions, wind, land type): $\mathrm{X}_t(\mathrm{s})$
- Binary indicator of a fire: $A_t(\mathrm{s})$
- The collection a process across space is bold, e.g.,

$$\mathbf{A}_t = \{A_t(\mathrm{s}); \mathrm{s} \in \mathcal{D}\}$$

Potential outcomes framework

- The $PM_{2.5}$ at s depends on the fire status at all sites, A
- This is called *interference* or *spill-over*
- When envisioning counterfactual outcomes we must consider all sites simultaneously

Potential outcomes framework

- The "treatment" is the regime
  - $R = 0$: world without forest fires
  - $R = 1$: current world with forest fires

- $Y_t(\mathrm{s}, 0)$ and $Y_t(\mathrm{s}, 1)$ are the potential $\mathrm{PM}_{2.5}$ outcomes

- Model: $Y_t(\mathrm{s}, 0) = \theta_t(\mathrm{s})$ and $Y_t(\mathrm{s}, 1) = \theta_t(\mathrm{s}) + \delta_t(\mathrm{s})$

- θ_t and δ_t are stochastic processes

Potential outcomes framework

- The causal effect is

$$\Delta(s) = E[Y_t(s,1) - Y_t(s,0)] = E[\delta_t(s)]$$

  where the average is over the distribution of covariates X and fires A over the entire spatial domain

- **Challenge**: we never observe data under $R = 0$

- To address this we use
  - CMAQ output
  - Causal assumptions

Assumptions

We assume there exist

- $C_t(\mathrm{s}) \in \{0, 1\}$ where s is affected by smoke iff $C_t(\mathrm{s}) = 1$
- Bias-correction functions B_0 and B_1

so that the following assumptions hold:

(A1) Consistency: $Y_t(\mathrm{s}) = Y_t[\mathrm{s}, C_t(\mathrm{s})]$

(A2) No unmeasured confounders given model output:

$$\theta_t(\mathrm{s}) = B_0[\hat{\theta}_t(\mathrm{s})] + e_{1t}(\mathrm{s}) \text{ and } \delta_t(\mathrm{s}) = B_1[\hat{\delta}_t(\mathrm{s})] + e_{2t}(\mathrm{s}),$$

where $\mathrm{e}_t(\mathrm{s}) = [e_{1t}, e_{2t}]$ is independent of X, A and **C**

Assumptions

(A1) Consistency: $Y_t(\mathrm{s}) = Y_t[\mathrm{s}, C_t(\mathrm{s})]$

(A2) No unmeasured confounders given model output:

$$\theta_t(\mathrm{s}) = B_0[\hat{\theta}_t(\mathrm{s})] + e_{1t}(\mathrm{s}) \text{ and } \delta_t(\mathrm{s}) = B_1[\hat{\delta}_t(\mathrm{s})] + e_{2t}(\mathrm{s}),$$

where $\mathrm{e}_t(\mathrm{s}) = [e_{1t}, e_{2t}]$ is independent of X, A and **C**

Are these assumptions reasonable?

- (A1) assumes that we have some observations we are sure are not affected by fire smoke (including spillover)...probably OK?
- (A2) assumes that the CMAQ modelers have included the important drivers of fine particulate matter....maybe OK? Have we accounted for all feedbacks?

Causal interpretation

Theorem 1:*Under (A1) and (A2) and assuming that* $\mathbf{C}_t$ *is not degenerate, then the parameters in the potential outcome model are identifiable via the distribution of observed data,* $Y_t \mid \hat{\theta}_t, \hat{\delta}_t, \mathbf{C}_t$.

- By Theorem 1, causal parameter estimation only requires inspecting the implied model for $Y_t(\mathrm{s})$ and confirming parameter identification.
- We specify parametric models for the bias correction functions B_0 and B_1 and the spatial process $\mathrm{e}_t(\mathrm{s})$.
- We then argue that all parameters, including the correlation between counterfactuals, are identifiable.
- This serves as a basis for using a Bayesian approach to estimating $\Delta(\mathrm{s})$.

Spillover/interference

- Defining the intervention as the fire regime instead of individual fires is key for two reasons:
  1. mimics the numerical model simulation
  2. limits the number of potential outcomes
- An alternative is to define potential outcomes for each potential fire-presence vector, $\mathbf{A}_t = [A_t(\mathbf{s}_1), ..., A_t(\mathbf{s}_n)]$
- This would require 2^n potential outcomes, and very complicated modeling and marginalization
- Limitation: the proposed framework cannot estimate effects of individual fires or specific features of fires

Bayesian hierarchical model

The model is:

$$\begin{aligned} Y_t(\mathrm{s}) &= \theta_t(\mathrm{s}) + C_t(\mathrm{s})\delta_t(\mathrm{s}) + \varepsilon_t(\mathrm{s}) \\ \theta_t(\mathrm{s}) &= B_0[\hat{\theta}_t(\mathrm{s})] + e_{1t}(\mathrm{s}) \\ \delta_t(\mathrm{s}) &= B_1[\hat{\delta}_t(\mathrm{s})] + e_{2t}(\mathrm{s}) \end{aligned}$$

- Measurement error (iid) is ε
- Background $PM_{2.5}$ bias adjustment:

$$B_0[\hat{\theta}_t(\mathrm{s})] = a_0(\mathrm{s}) + b_0(\mathrm{s})\hat{\theta}_t(\mathrm{s})$$

- Fire-contributed $PM_{2.5}$ bias adjustment:

$$B_1[\hat{\theta}_t(\mathrm{s})] = a_1(\mathrm{s}) + b_1(\mathrm{s})\hat{\delta}_t(\mathrm{s})$$

Bayesian hierarchical model

The model is:

$$\begin{aligned} Y_t(\mathrm{s}) &= \theta_t(\mathrm{s}) + C_t(\mathrm{s})\delta_t(\mathrm{s}) + \varepsilon_t(\mathrm{s}) \\ \theta_t(\mathrm{s}) &= B_0[\hat{\theta}_t(\mathrm{s})] + e_{1t}(\mathrm{s}) \\ \delta_t(\mathrm{s}) &= B_1[\hat{\delta}_t(\mathrm{s})] + e_{2t}(\mathrm{s}) \end{aligned}$$

- Fire indicator: $C_t(\mathrm{s}) = 1$ if $\hat{\delta}_t(\mathrm{s}) > \tau$
- Spatial process $\mathrm{e}_t(\mathrm{s})$ follows a bivariate Matern process

$$\mathrm{Cov}[e_{jt}(\mathrm{s}), e_{kt}(\mathrm{s}')] = M(||\mathrm{s} - \mathrm{s}'||; \rho_{jk})$$

where M is the Matern covariance function and ρ_{jk} are covariance parameters

Identification

- The mean is $E[Y_t(s)]$ is

$$\alpha_0(s) + \beta_0(s)\hat{\theta}_t(s) + \alpha_1(s)C_t(s) + \beta_1(s)[C_t(s)\hat{\delta}_t(s)]$$

- Identification follows assuming $\{1, \hat{\theta}_t(s), C_t(s), C_t(s)\hat{\delta}_t(s)\}$ are not perfectly collinear

- The covariance $\text{Cov}[Y_t(s), Y_t(s')]$ is

$$\begin{cases} M(||s - s'||; \rho_{11}) & \text{if } C_t(s) = C_t(s') = 0 \\ M(||s - s'||; \rho_{12}) & \text{if } C_t(s) \neq C_t(s') \\ M(||s - s'||; \rho_{22}) & \text{if } C_t(s) = C_t(s') = 1 \end{cases}$$

- By examining the correlation between pairs separately by $C_t(s)$ the correlation parameters are identifiable

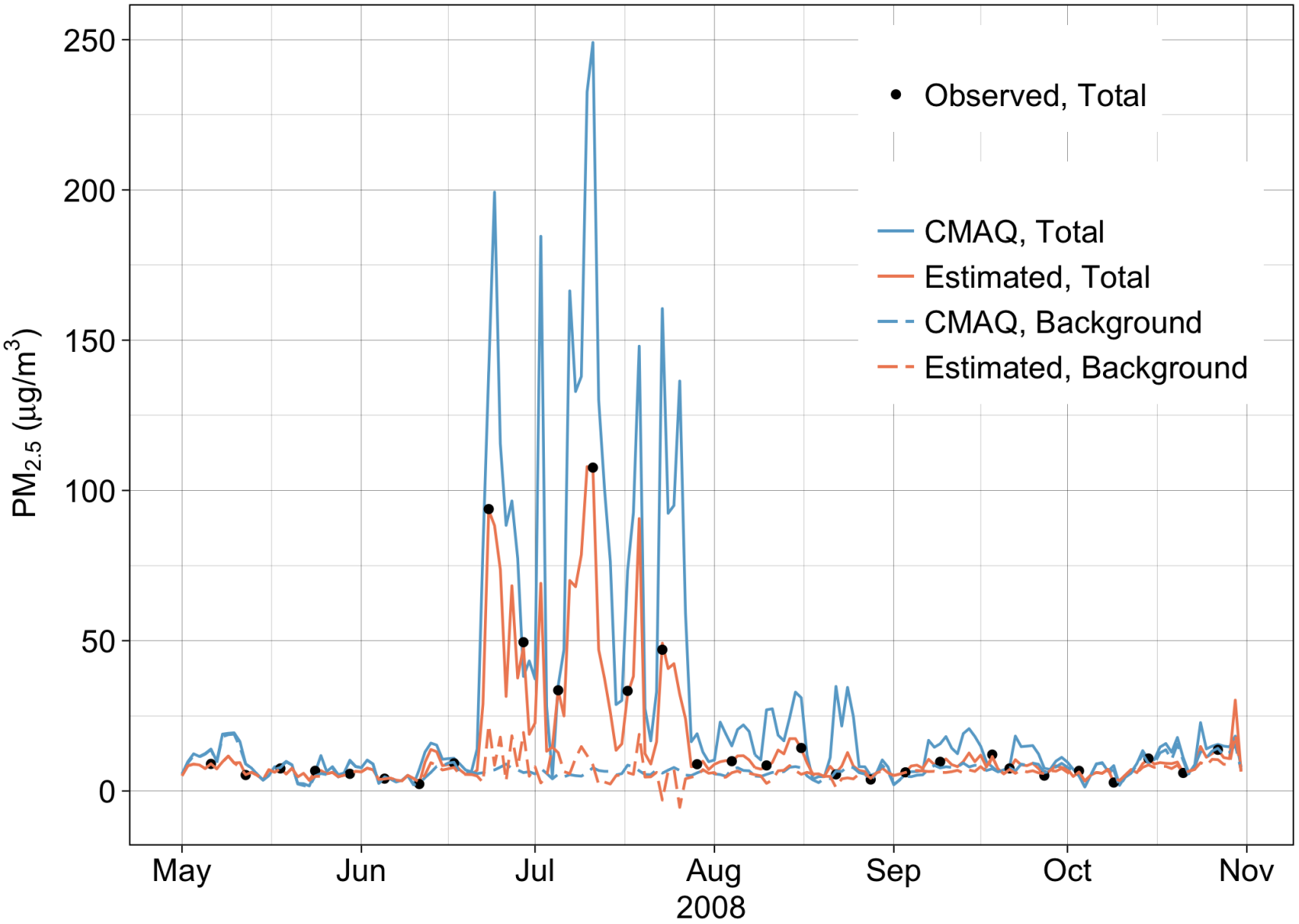
Posterior inference

- We approximate the causal effect $\mathrm{E}[C_t(\mathrm{s})\delta_t(\mathrm{s})]$ by

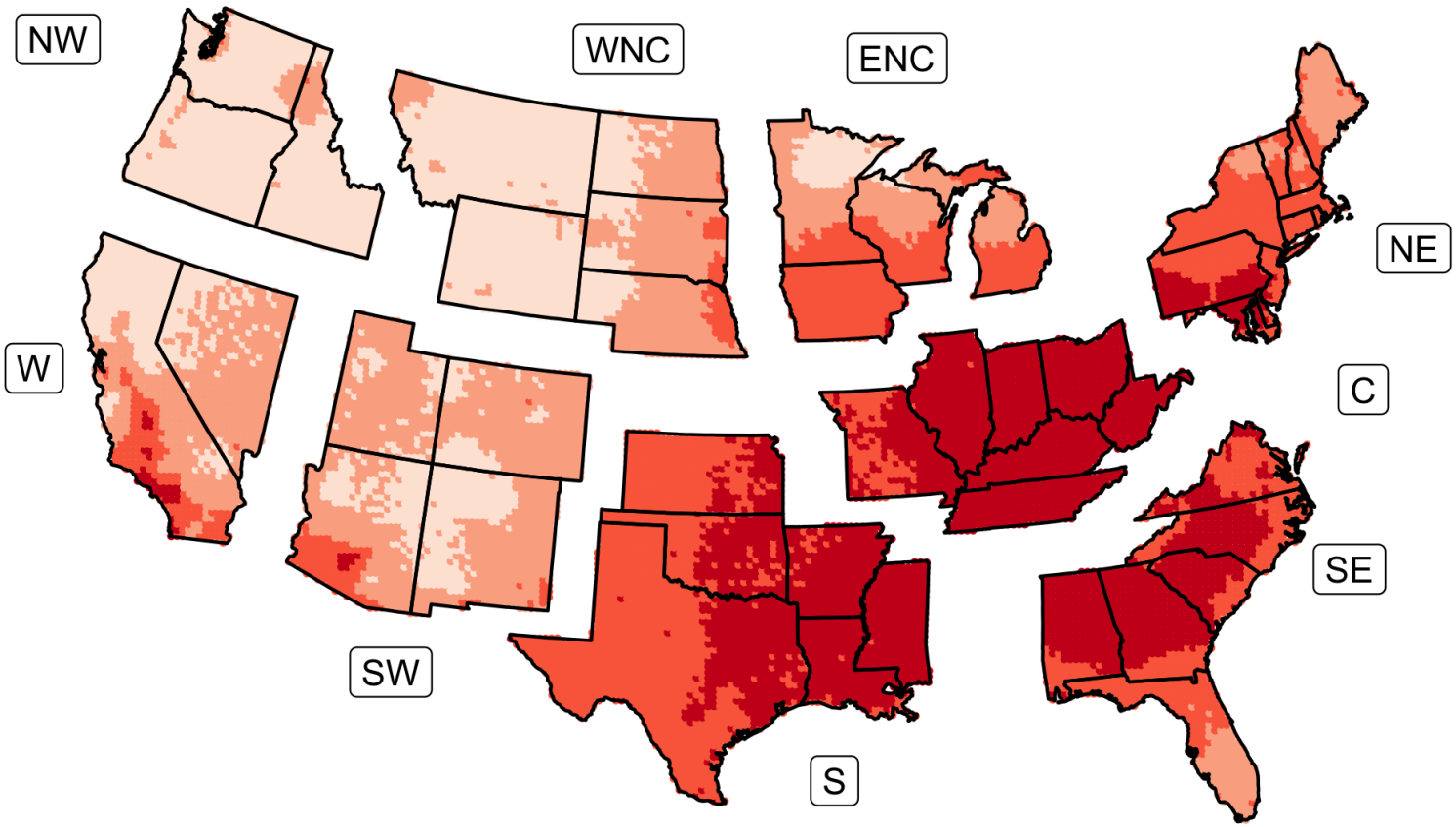
$$\Delta(\mathrm{s}) = \sum_{t=1}^{T} C_t(\mathrm{s})\delta_t(\mathrm{s})$$

- The Bayesian framework gives the full posterior for Δ
- Under (A1) and (A2) this has a causal interpretation
- The model is fit separately in 9 subregions
- We assume independence across days

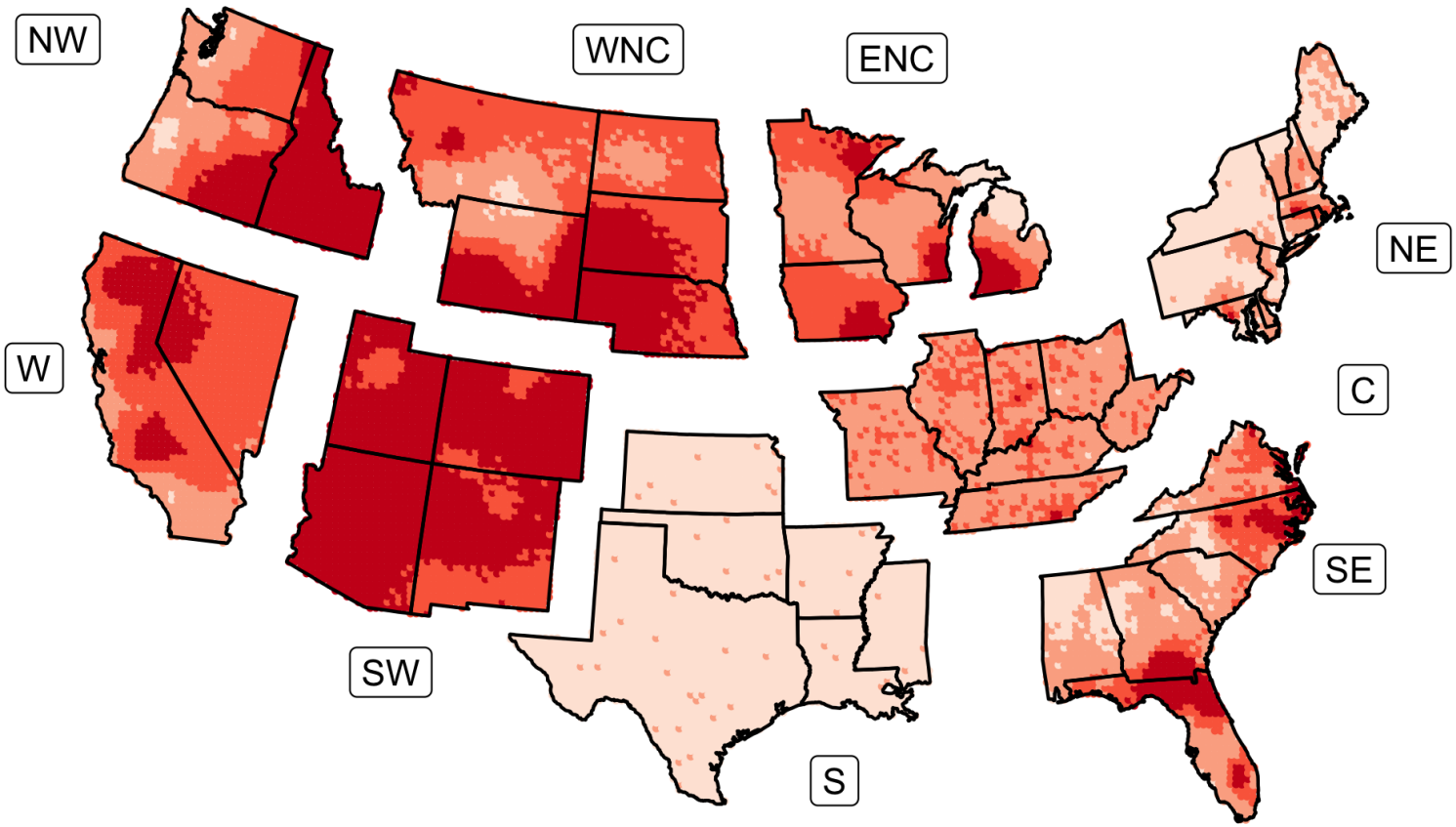
Data and estimates for one site in CA



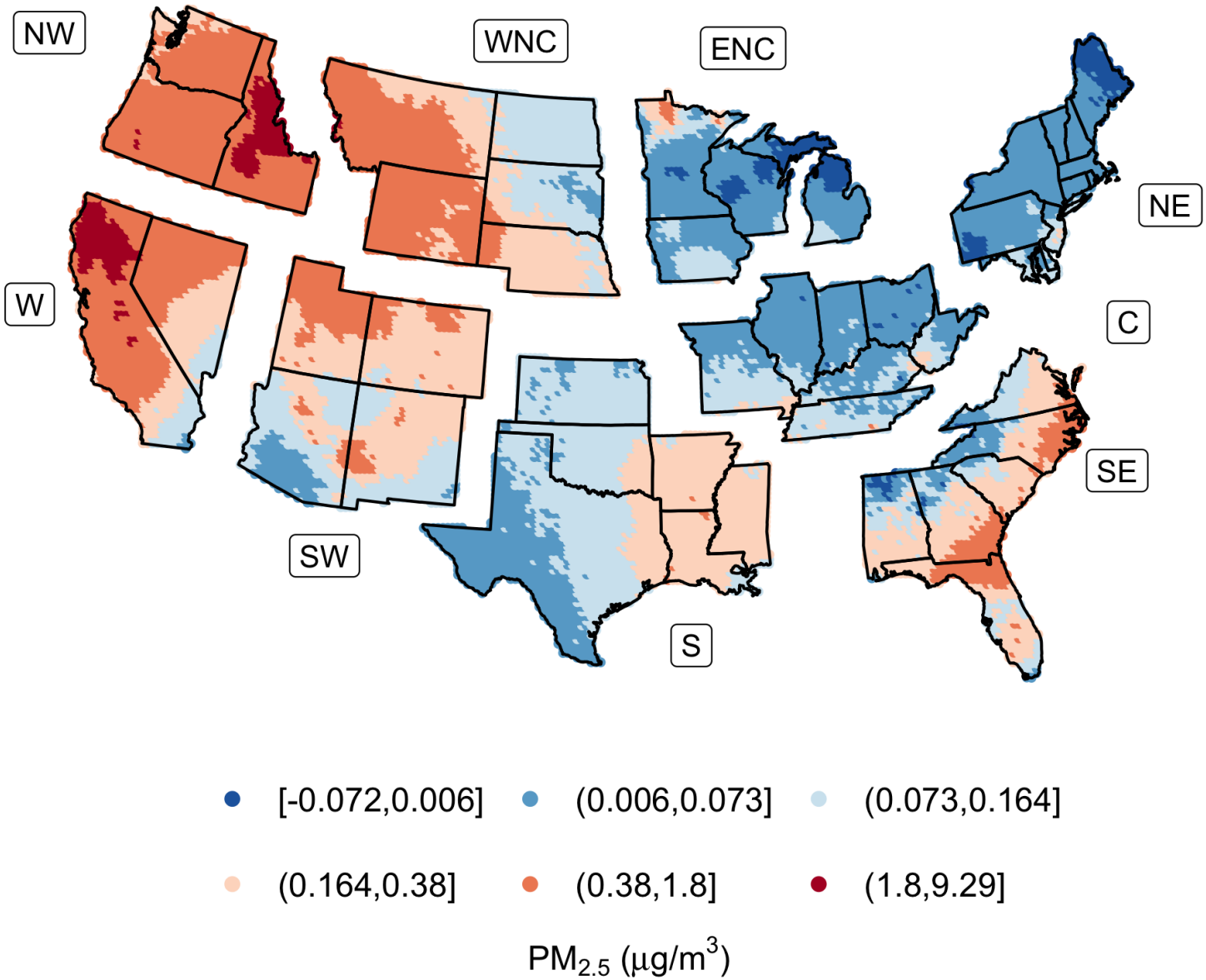
Average $\theta_t(s)$ over the 2008-2012 wildfire season



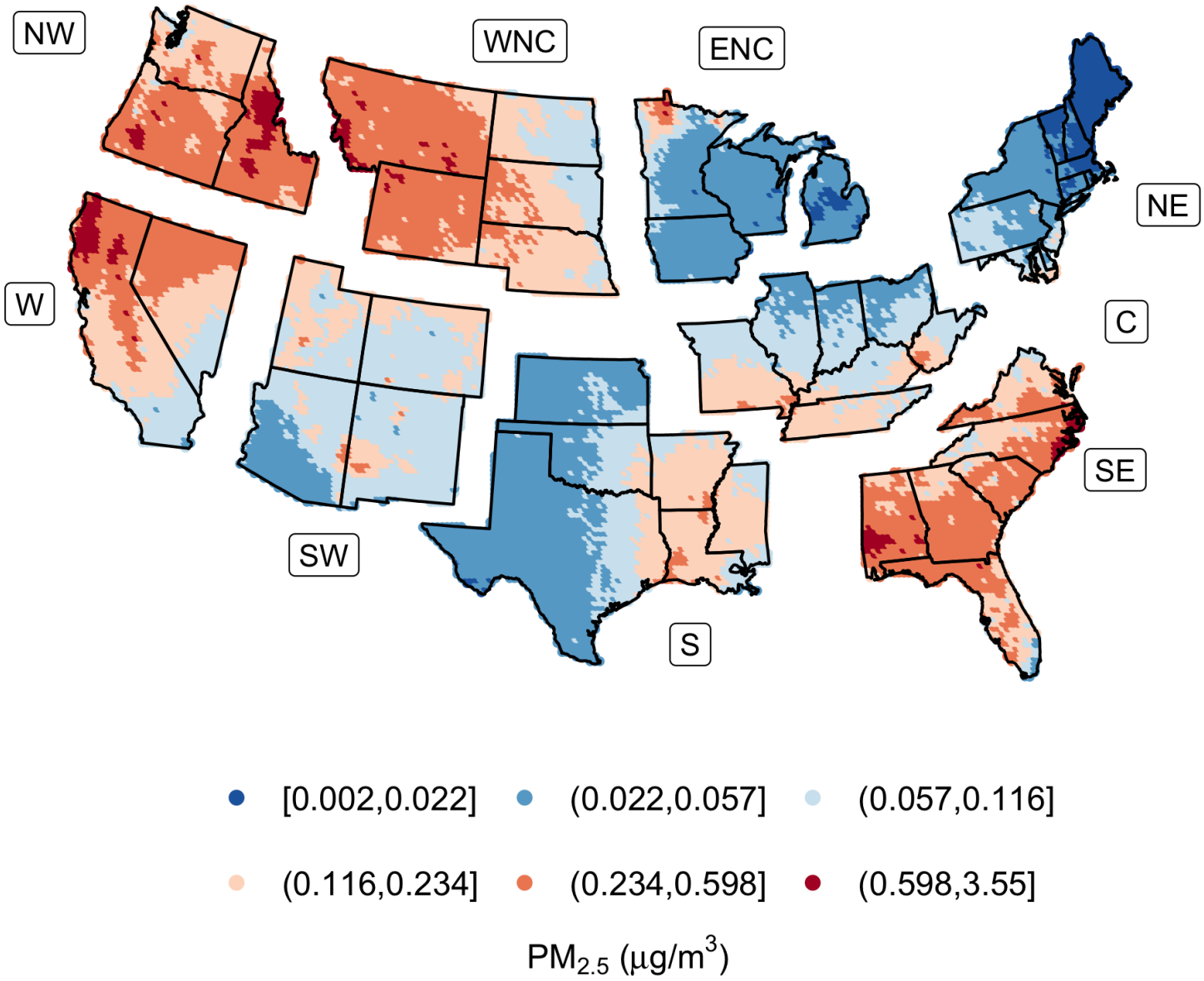
Posterior mean slope, $b_1(s)$



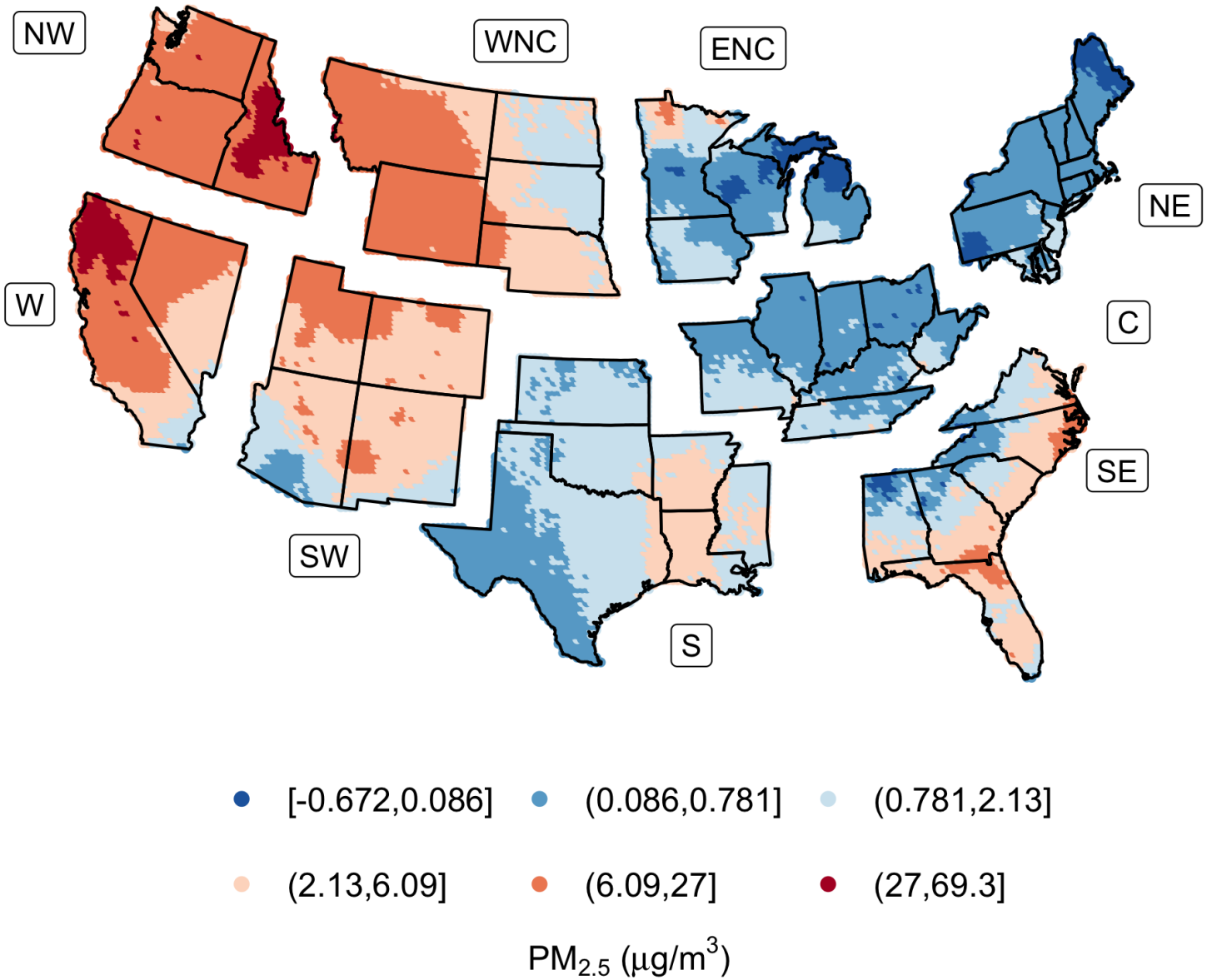
Causal estimate, $\Delta(s)$, posterior mean



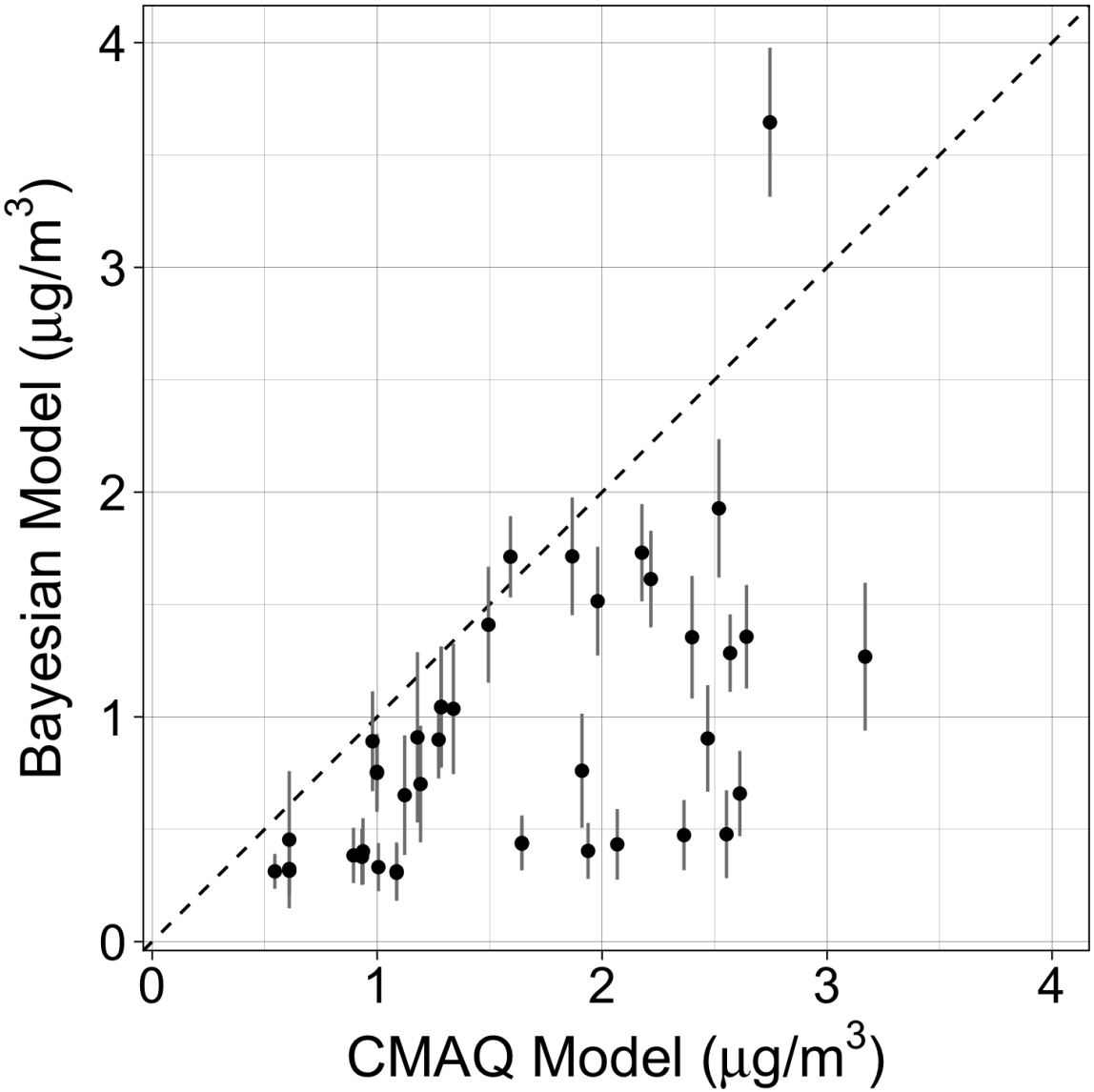
Causal estimate, $\Delta(s)$, posterior SD



Fire-Contributed PM as % of Total



Causal Estimate vs. CMAQ at monitors in the NW



Estimated[5] number of attributable hospitalizations

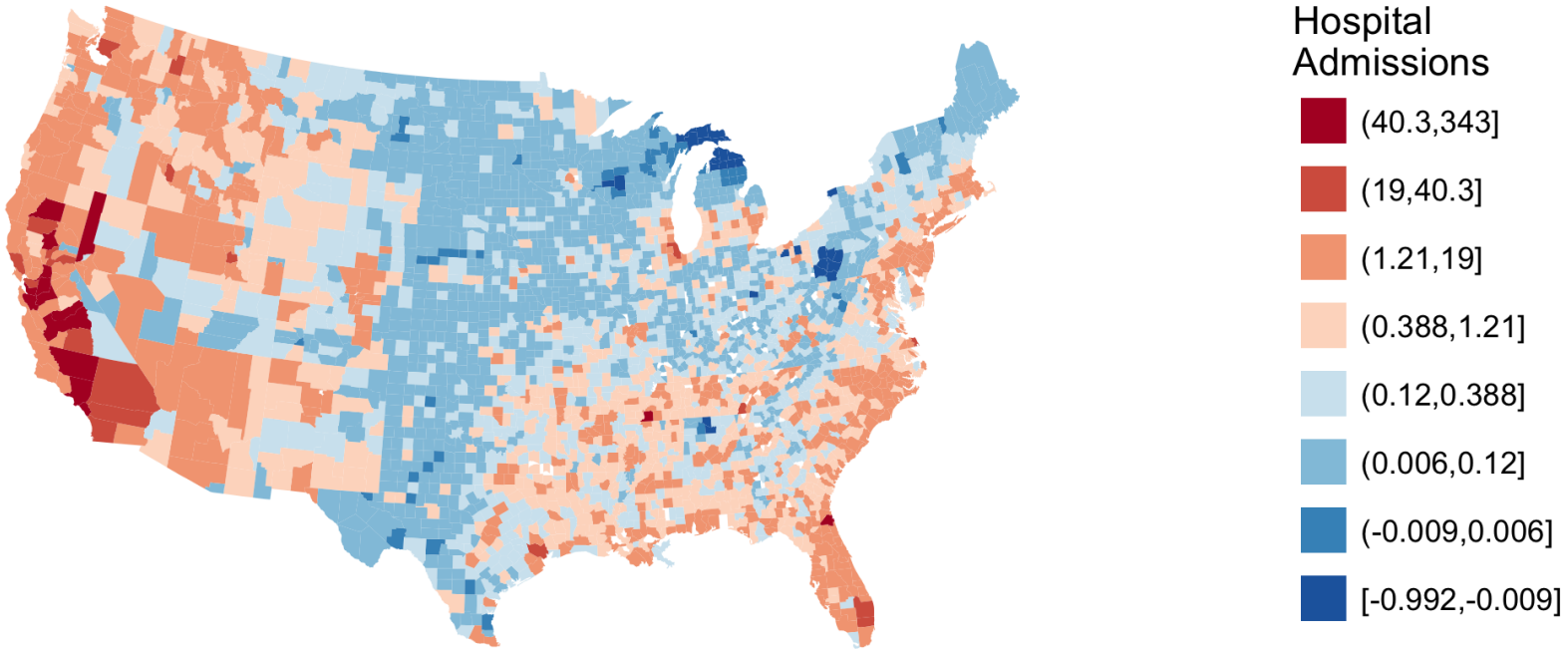
Region	Method	Age Group 0-1	65-99	0-99
Central	Bayesian	91	170	396
	CMAQ	603	1134	2626
ENC	Bayesian	18	40	88
	CMAQ	92	205	436
South	Bayesian	111	202	456
	CMAQ	601	1092	2462
Southeast	Bayesian	196	395	901
	CMAQ	625	1260	2881
Southwest	Bayesian	100	75	248
	CMAQ	182	156	484

[5] Relative risk of wildfire-contributed $PM_{2.5}$ taken from Delfino et al (2009)

Estimated number of attributable hospitalizations

Region	Method	Age Group 0-1	65-99	0-99
Northeast	Bayesian	49	109	234
	CMAQ	178	427	895
Northwest	Bayesian	89	194	423
	CMAQ	177	392	848
West	Bayesian	427	714	1631
	CMAQ	912	1548	3507
WNC	Bayesian	47	28	130
	CMAQ	93	57	257

Cumulative health burden by county



Summary and future projects

- We derived the assumptions needed to make causal claims using numerical model output
- We found large contributions of fires to PM, especially in the west, even after adjusting for model bias
- Future work:
  - Continuous treatments
  - Causal quantile effects
  - Analyze individual fires

Changing scales!

- 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 from sensors strapped to Google StreetView cars
- Some cities now have thousands of low-cost sensors
- Phone apps are under development

Photo: Apte (2017)

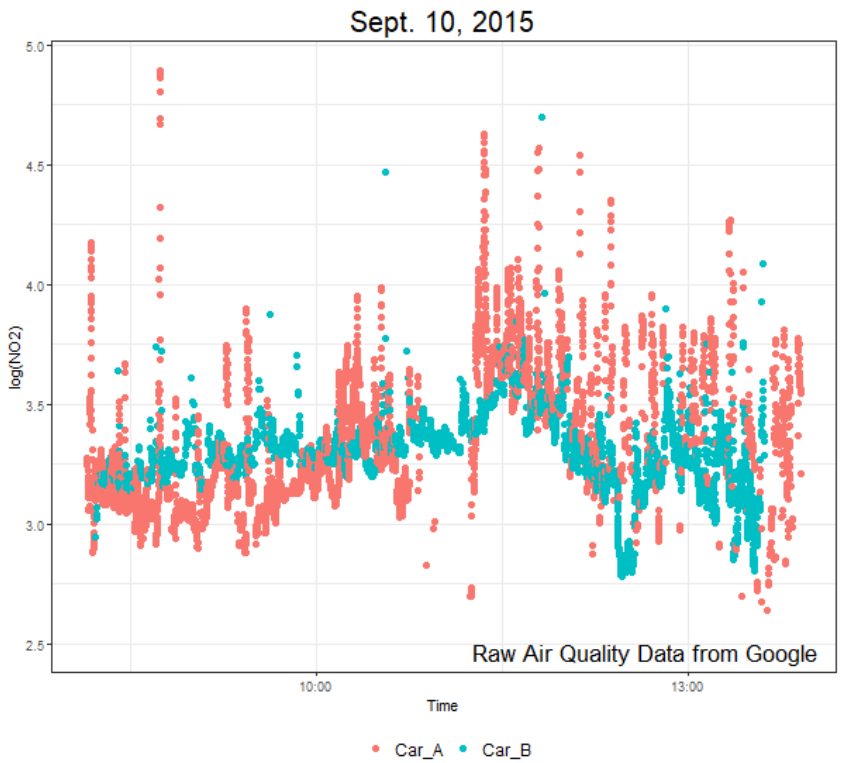
Mobile data: Oakland NO_2

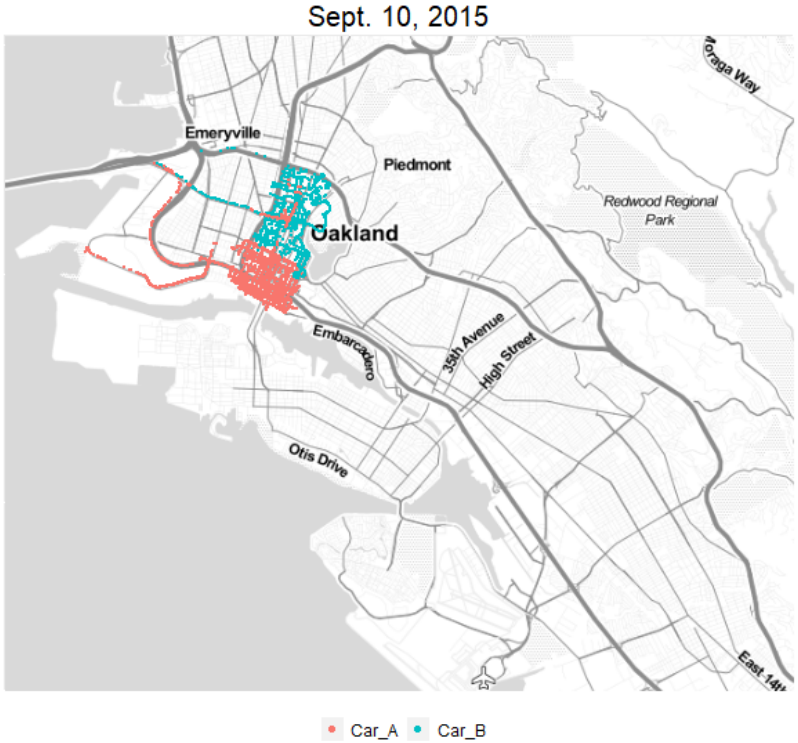
- Two cars were deployed from June 2015 to May 2016
- Starts on weekdays at $\approx$ 9am, drove $\approx$ 6-8 hours each day
- Measurements are taken at roughly every second.
- Large missing data (car maintenance, sensor failure, etc.)

Statistical Challenges:

- Data are large ($\approx$900,000 observations in current dataset)
  - Computationally prohibitive to fit full spatiotemporal models
- Data are extremely sparse in space and time
  - Maximum of two observations at a time point.
  - Only a small region of Oakland is sampled on each day.
- Data are noisy and subject to outliers

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

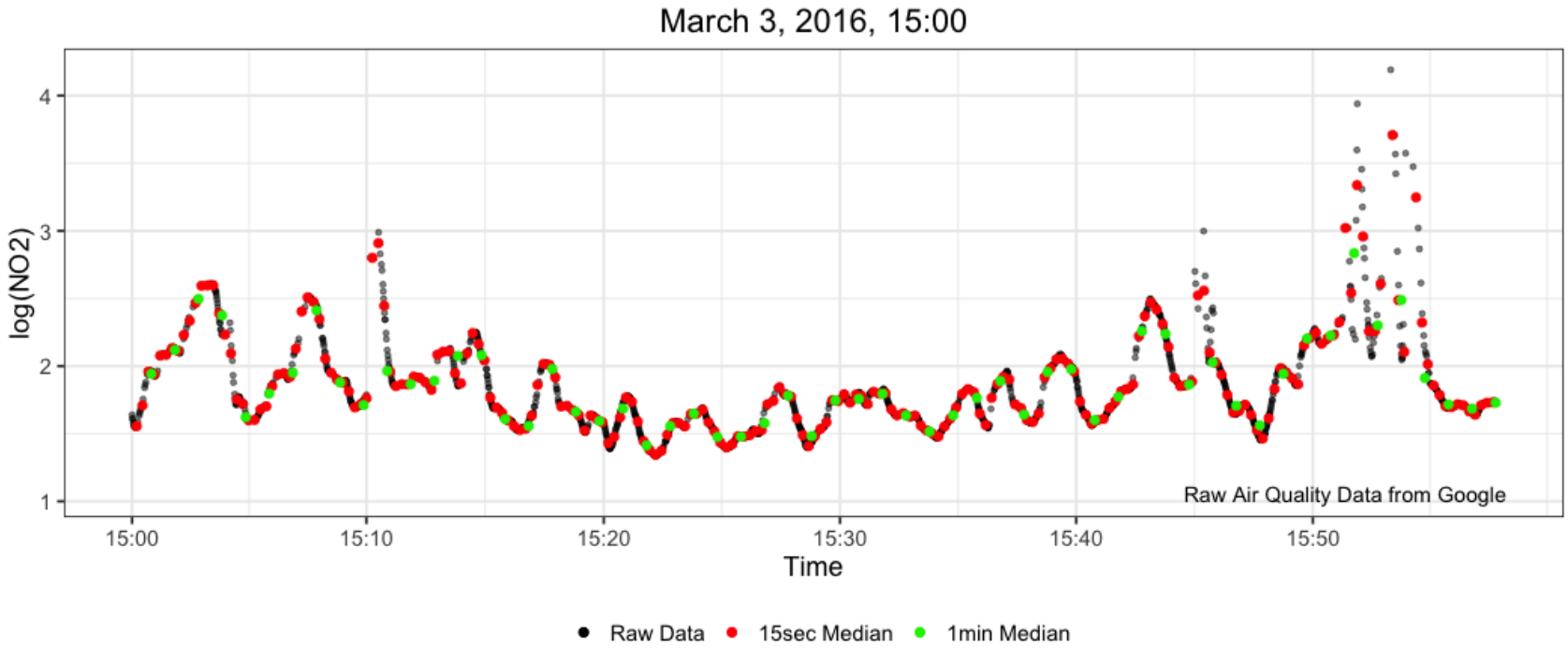
Develop a statistical model incorporating landuse covariates and spatiotemporal dependence for real-time, high resolution (30m) forecasting of air pollution.

1. How well and how far ahead can we reasonably forecast air pollution levels?

2. If we were to design a new Google car study, how many cars should be deployed to improve prediction?

3. Is deploying sensors on cars more efficient than a fixed-location sensor network?

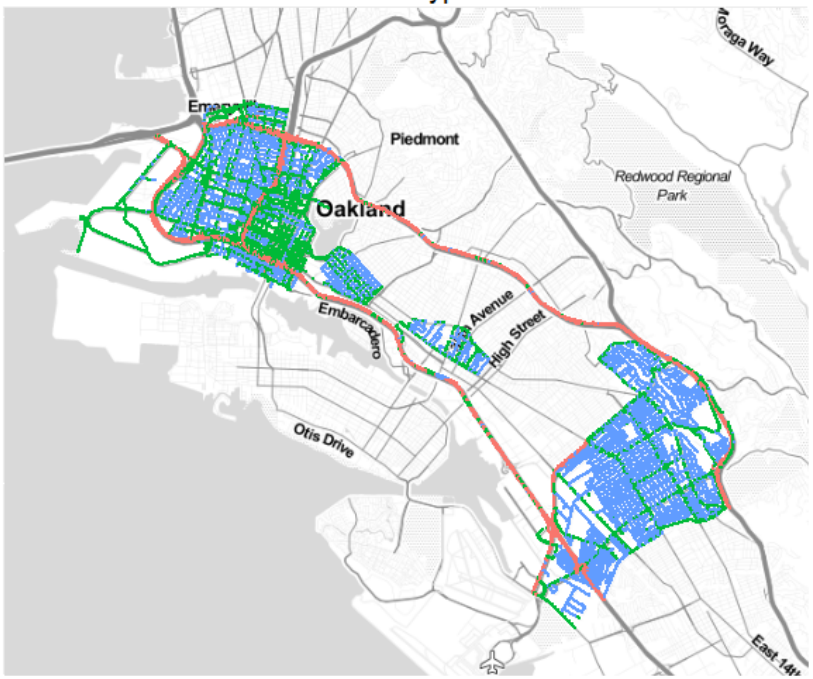
Temporal aggregation



We took temporal block medians to dampen effects of extremes

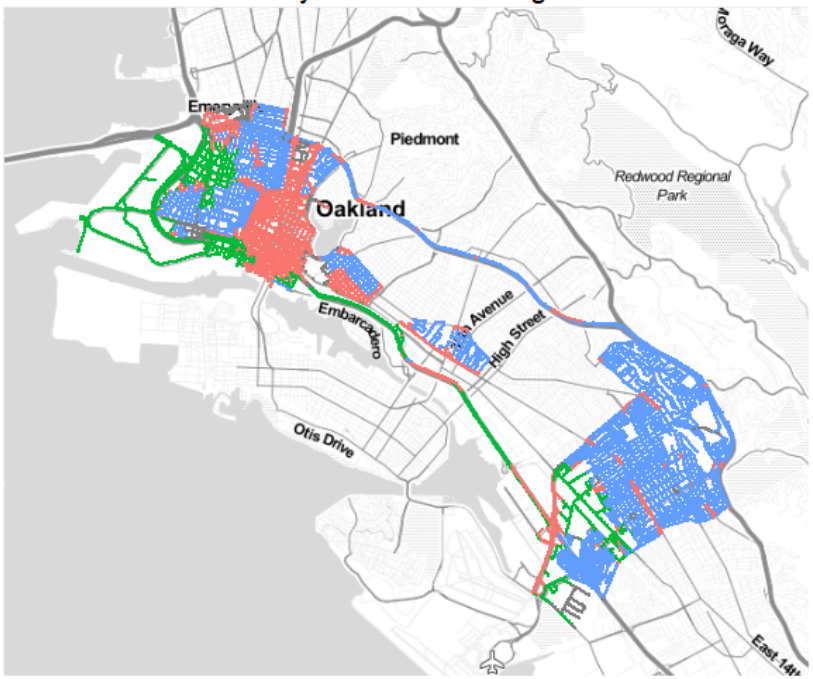
Example landuse covariates

Road Type



Highway Major Residential

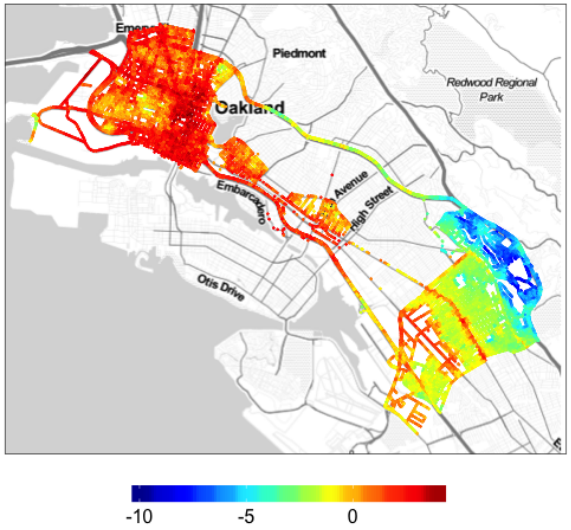
City of Oakland Zoning



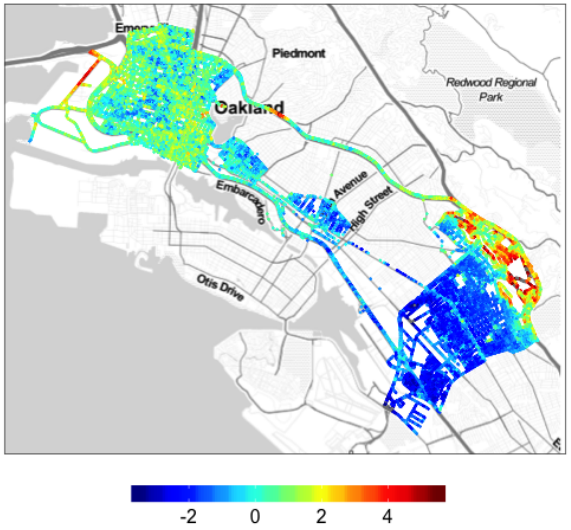
Commerical Industrial Residential NA

PCA of landuse variables

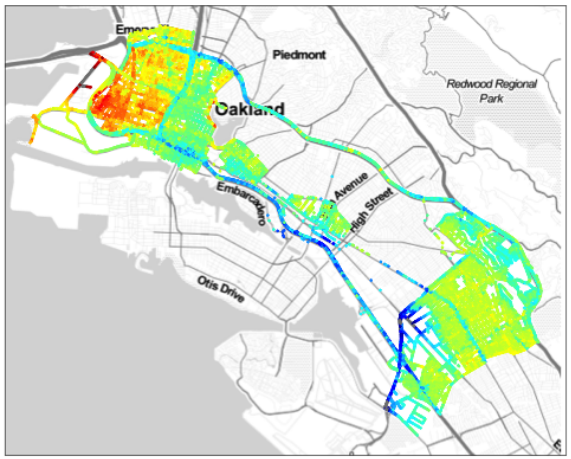
PC 1



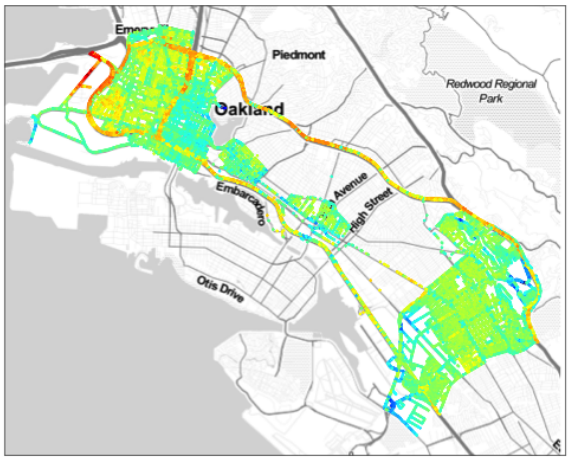
PC 3



PC 4



PC 5



Non-spatial landuse regression

Let $Y_t(s)$ be the log(NO_2) at time t and location s

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

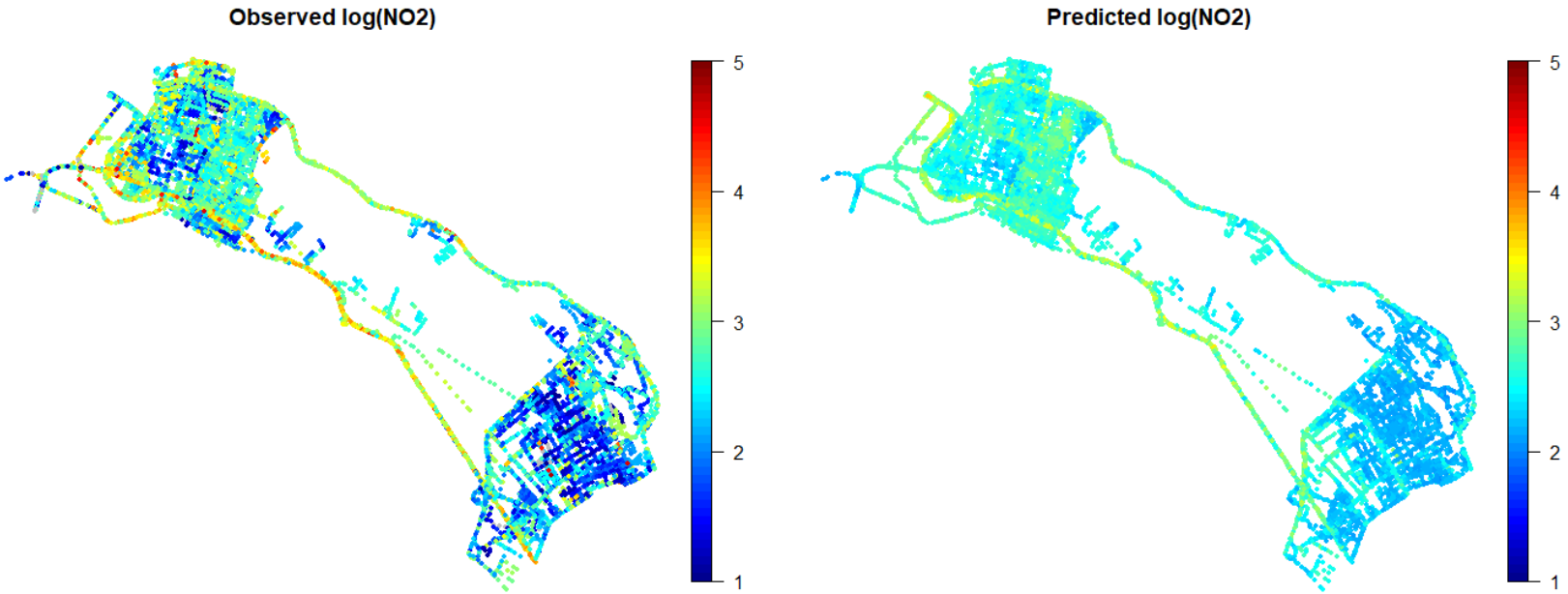
where $X_t(s)$ contains

- 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

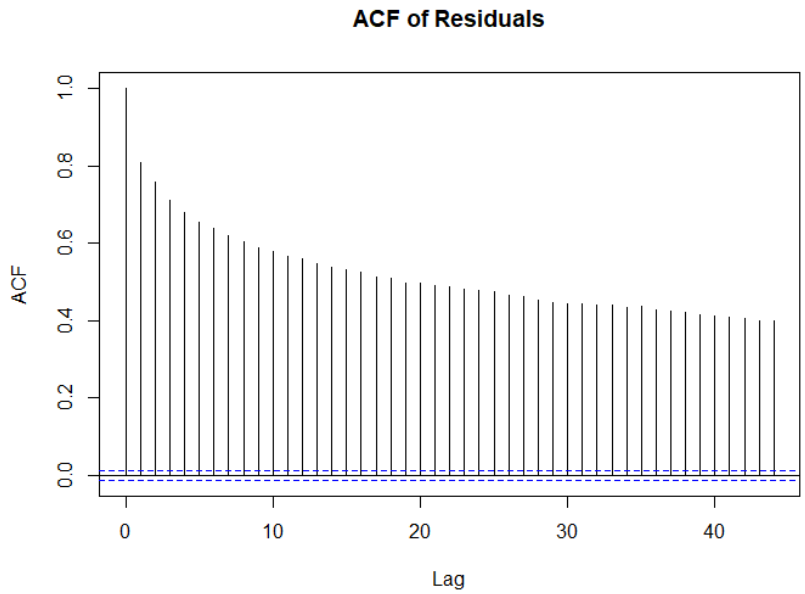
Results from landuse regression

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

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

- The covariance is

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

- ρ and ϕ determine the spatial and temporal dependence.

Computation

- MLE for the full dataset is impossible
- We first estimate β using OLS and compute residuals e_i
- We use the Veccia approximation to estimate the covariance parameters
- The joint distribution is approximated by the product of conditional distributions

$$f(e_1, ..., e_n) \approx \prod_{i=1}^{n} f\left(e_i | e_j \in \mathcal{N}_i\right)$$

where N_i is the set of "neighbors" for observation i

- This is fast if the neighboring sets are small

Computation

- The neighboring sets are observations in the recent past

$$\mathcal{N}_i = \{\text{obs between } l \text{ and } l + 60 \text{ minutes prior to obs } i\}$$

- Taking $l = 0$ gives the best approximation to the likelihood

- Gramacy and Apley (2015) show its better to include some distant neighbors

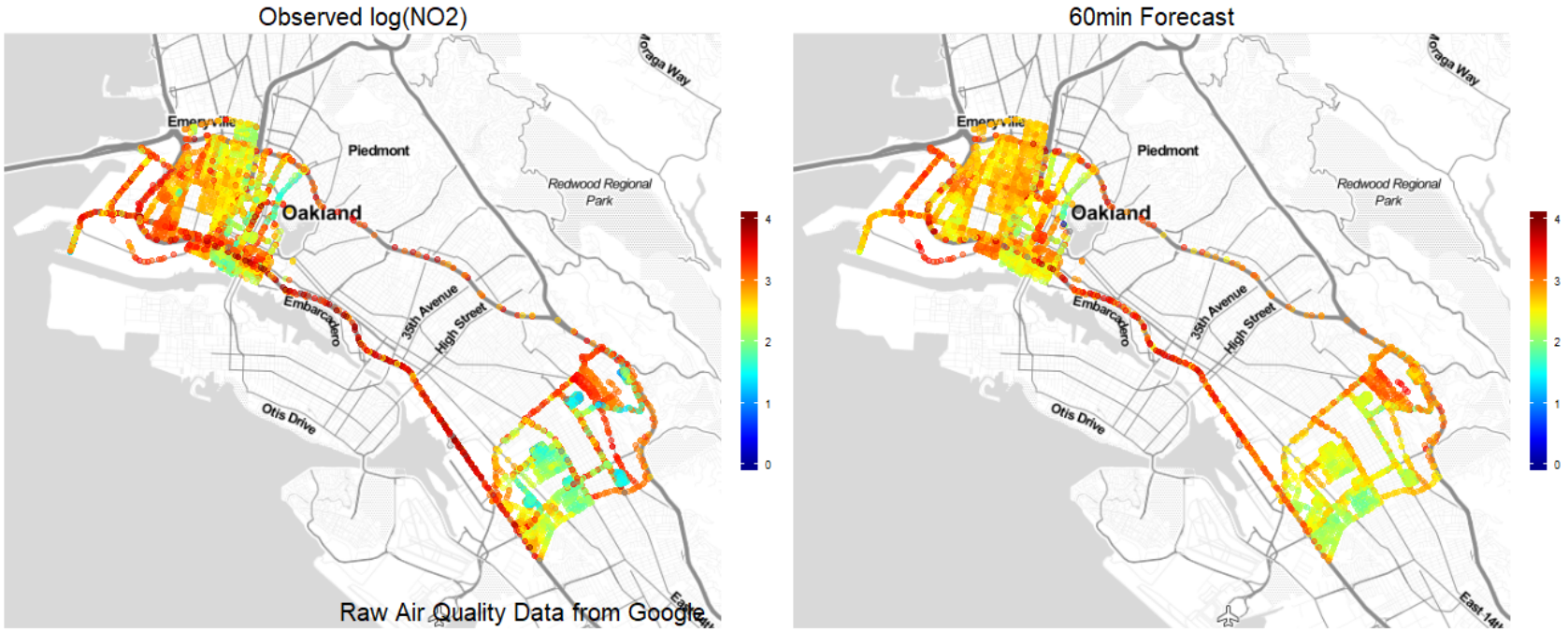
Prediction correlation using 1 sec block medians

Model	l	Prediction lag 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

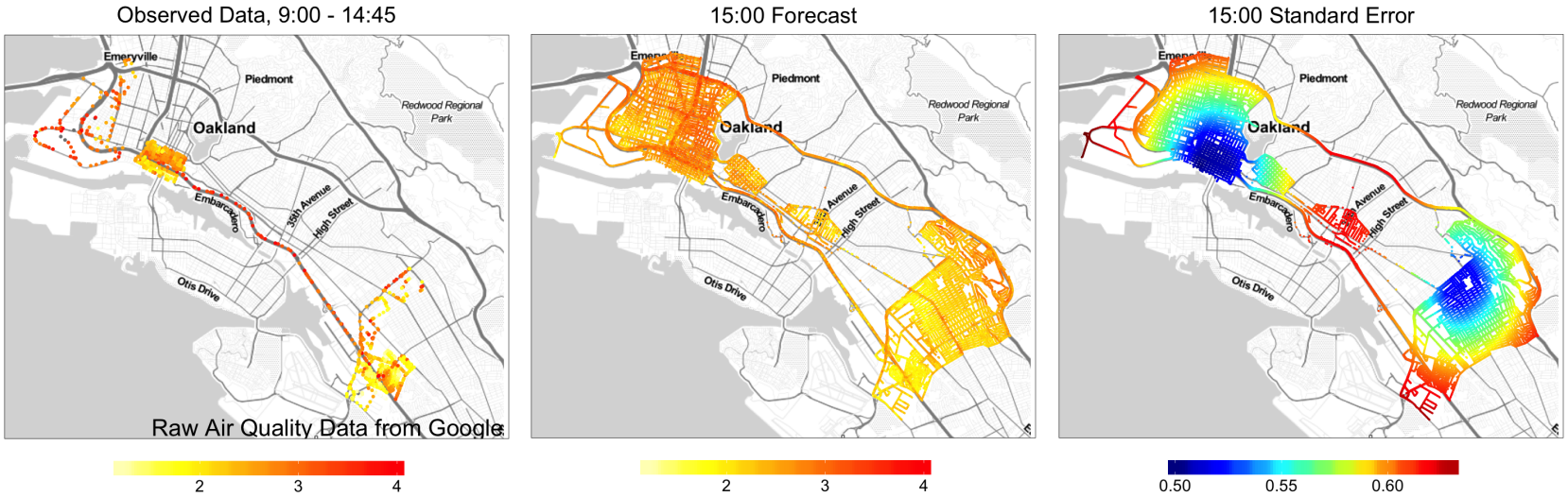
Prediction correlation using 1 min block medians

Model	l	Prediction lag 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

Forecast for Dec 2015 - Feb 2016



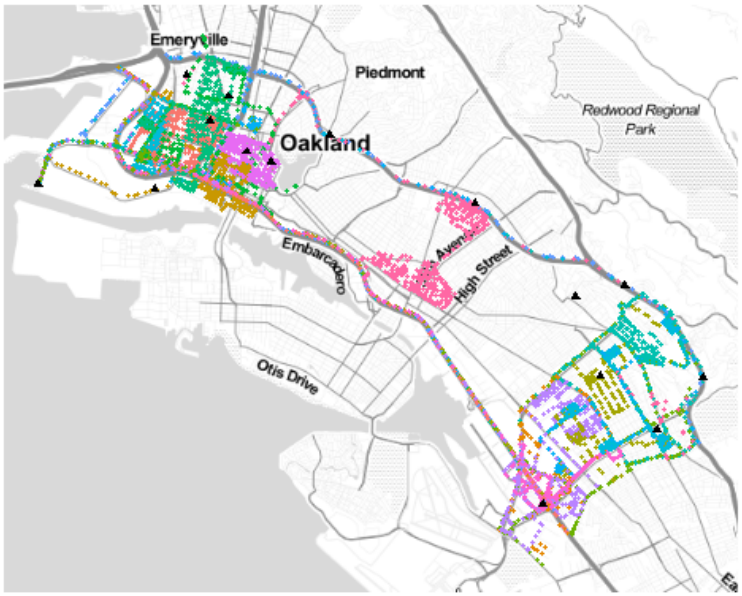
15 minutes ahead forecasts of NO_2



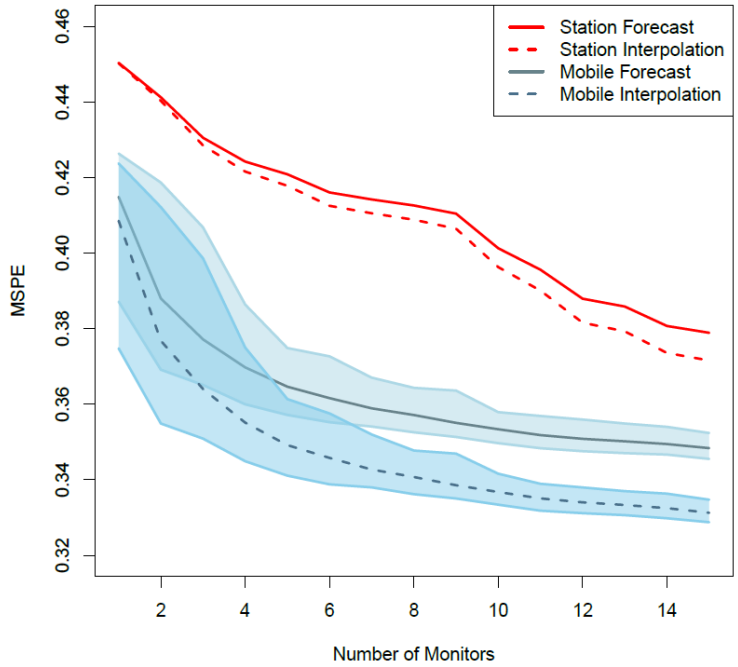
- Forecast for 15:00 using the data from 13:45 to 14:45 on May 5, 2016
- As expected, standard errors are lowest where data has been obtained most recently from the two cars.

Network design

Objective: To improve NO_2 prediction how many Google cars should be deployed? How many fixed-location sensors would provide the same quality of prediction?



Deploy different number of mobile and fixed-location sensors.



Prediction performance comparison.

Summary and future projects

- Our work shows that short-term forecasts of air pollution at a high spatial resolution are possible
- Future work:
  - Fuse mobile and stationary sensors
  - Model extremes
  - Multicity analysis
  - Design efficient routes
- Works supported by NIH and NSF
- **THANKS!**