Making Sense of Sensors: Predictive Modeling Approach

Nikolay Bliznyuk, Associate Professor
Departments of ABE, Biostatistics & Statistics
University of Florida

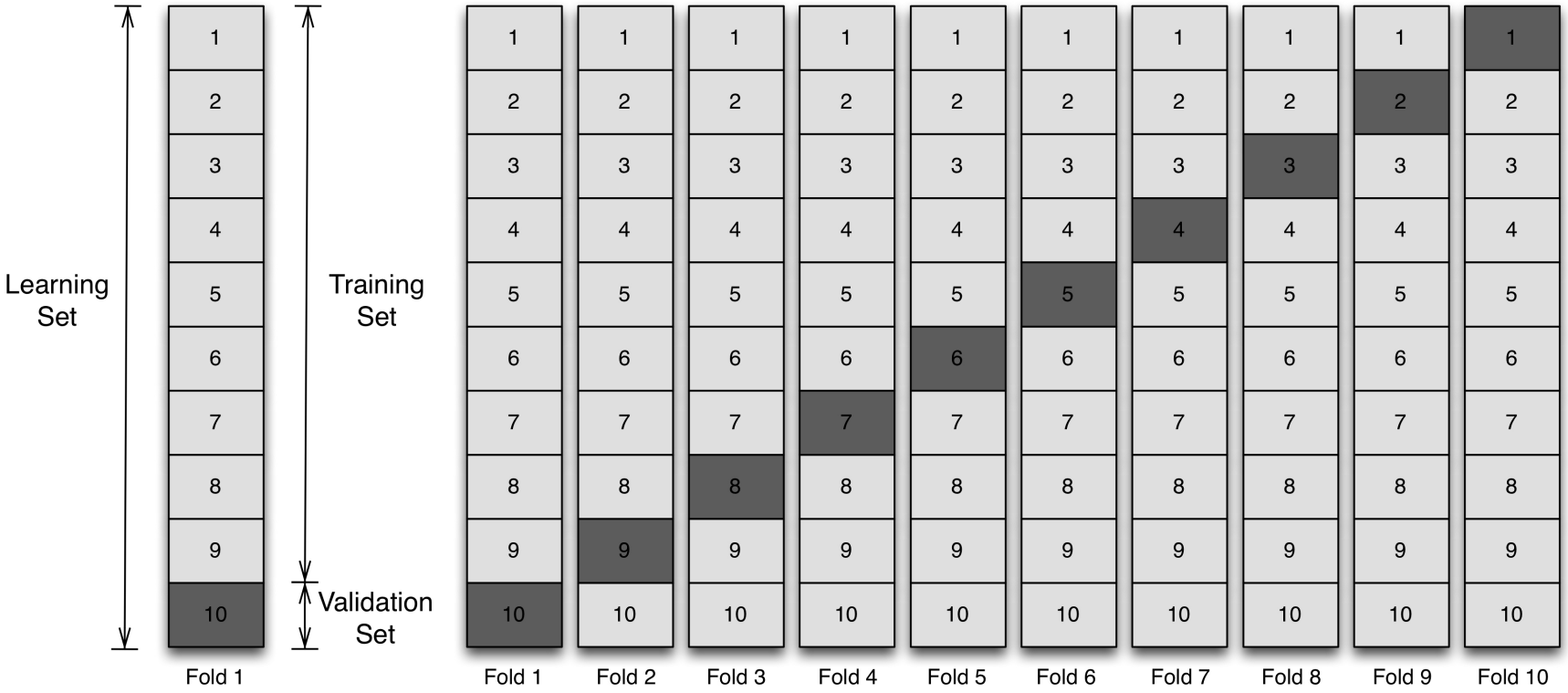
14 November, 2018

Predictive Modeling - Goals and Approach

Setup - (Supervised) Statistical/Machine Learning (SML)

- Y is the response (quantitative or categorical),
- X is the set of covariates/predictors; often high-dimensional ("large p")
- Relate $Y \sim X$ as
  - $Y = f(X) + \epsilon$ (regression), or
  - $Pr(Y = k) = g(X)$ (classification), or
  - $Y \approx$ blackBox(X) ("black box" machine learning)
- Goal: good predictive performance of Ys on a left-out set of data (aka "test" data). Care less about inference.
- Fitting/training: have models use a subset of data ("training set") to learn the relationship $Y \sim X$ and calibrate tuning parameters, then predict outside of training sample.
- Approach: K-fold cross-validation to tune parameters and estimate predictive performance (test RMSE or misclassification rate).

K-fold Cross-Validation - Illustration



Rose (2010, 2016)

(Almost) Everything is a Sensor (or Acts Like It)

Case Studies:

1. Predicting type I diabetes (T1D) in young children using high-dimensional “multiomics" data (gene expression, blood metabolites, gut microbiome, etc).

   Omics platforms outputs act as sensors. E.g., gene expression data - measure activity of $\approx$ 21,000 genes.

   Y - incidence of T1D (binary). X - omics profiles.

2. Predicting household-level water use in an urban area.

   Individual households act as sensors.

   Y - monthly water use data; X - demographic, economic & climate variables.

3. Predicting particulate matter air pollution in an urban setting.

   A network of “rotating" monitors recording black carbon (BC) concentration acts as a sensor.

   Y - BC readings; X - space, time, pop. and climate variables.

Case Studies: Complicating "Stylized Features"/Properties

Three Case Studies:

1. Predicting type I diabetes (T1D) in young children

   "Large p, small n"; missing data; standardization and normalization

2. Predicting household-level water use in an urban area.

   Temporal & spatial dependence/heterogeneity; very large n

3. Predicting particulate matter air pollution in an urban setting.

   Multiple related datasets of different space-time coverage at different temporal resolutions. A lot of missing data.

   A highlight: pooling data of different types using hierarchical Bayesian modeling

Case Study: Predicting T1D in Children using TEDDY Study Data

Objectives

- Create predictive models using “omics" datasets collected in a large scale study of T1D
- Test models for each omics dataset on left out data
- Combine best performing omics models into a single predictive model
- Test final models on held out data

TEDDY Study

- Study was run by The Environmental Determinants of Diabetes in the Young (TEDDY)
- Study was conducted at 6 clinical centers (Seattle, WA, Denver, CO, Augusta, GA, Munich, Germany, Turku, Finland, and Malmo, Sweden)
- Several thousand children with a genetic predisposition to type one diabetes (T1D) were chosen
- Subjects were followed from birth and visited clinical centers roughly every 3 months until 4 years old
- Case-control study: Subjects who developed T1D were matched with healthy children of the same or similar age, gender, genotype, and location

Data Structure

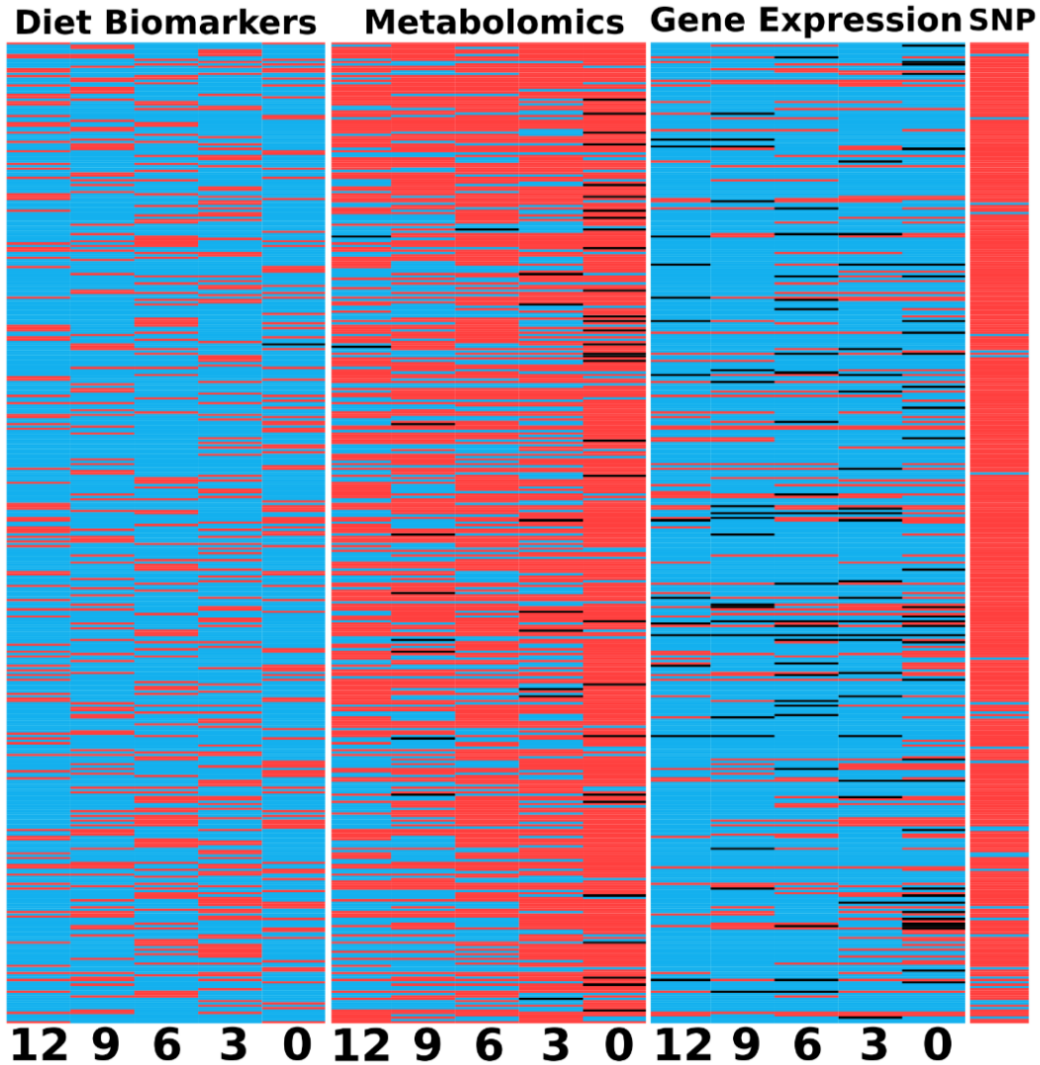
- For our study the year prior to seroconversion (typical precursor of T1D) was used
- Data was divided into 3 month timepoints to track with testing schedule
- Timepoints are defined as months until seroconversion
- Helped to cope with high level of missing data
- Demographic data was included in each omics data set for each subject

TEDDY Study

Our datasets range from 400 to 1,400 subjects. The omics datasets include:

- Gene expression, $p \approx 21000$ (dimension reduction is critical)
- Metabolomics, $p \approx 1300$
- Dietary Biomarkers
- SNPs
- Microbiome

Pattern of Data Missingness

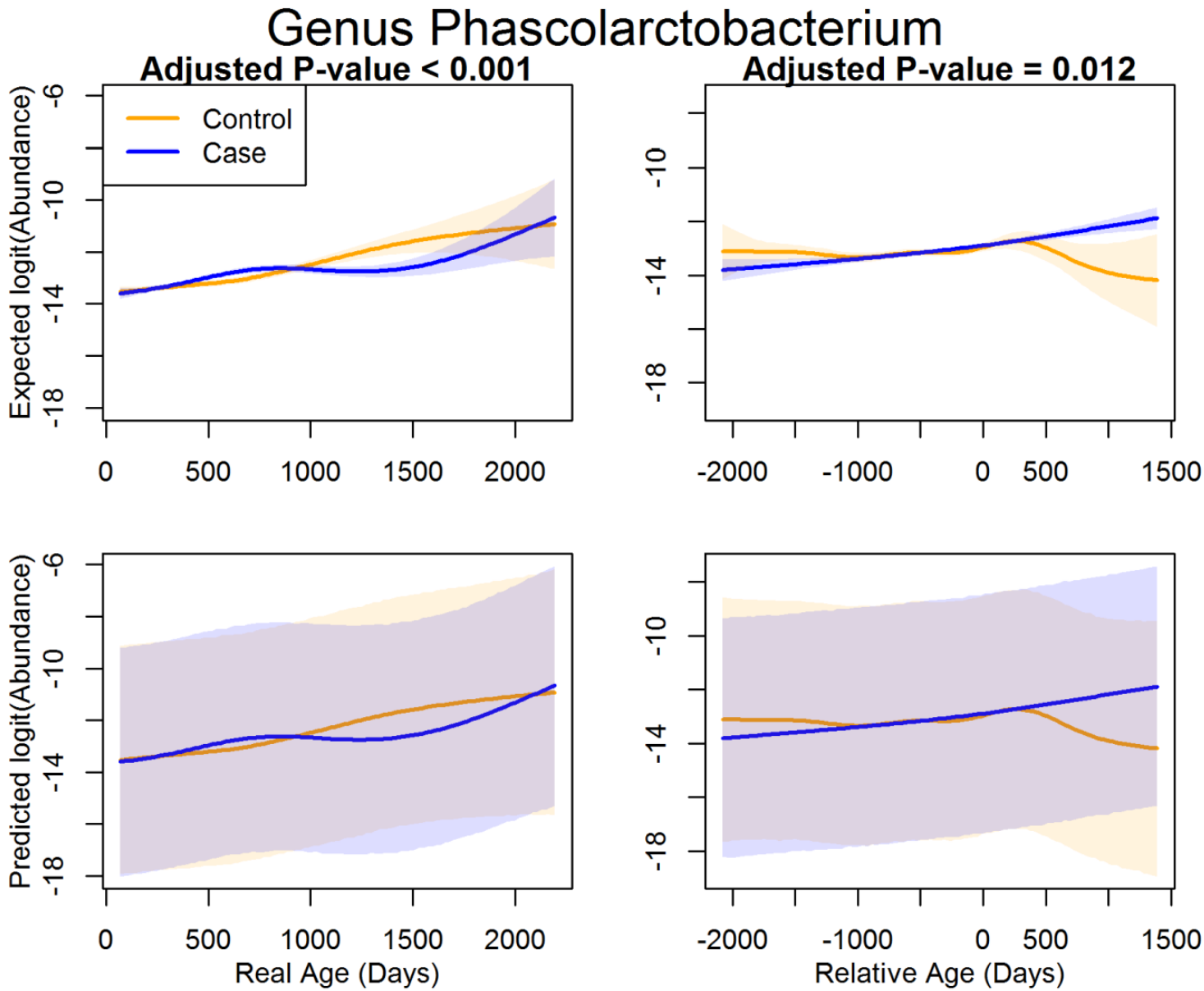


Multiomic Data Availability

Available data by omics type

Data	TP0	TP3	TP6	TP9	TP12
Gene, meta, SNP	157	110	105	95	94
Gene, meta, SNP, microbio	103	84	77	65	58
Gene, meta, SNP, biomark	46	40	26	39	26
All	31	35	26	29	18

Inference vs Prediction



Best-Performing Multiomic Models across All Timepoints

Best performing multiomic models across all timepoints

Gene	Meta	SNP	Model	Misclassification
x	x	x	RF	41.8
	x	x	RF	42.4
x	x		SVM	43.2
x		x	RF	44.7

Best-Performing Multiomic Models

Time	Model	Misclassification
0	GBM	41.0
3	RF	40.7
6	SVM	42.1
9	SVM	36.7
12	SVM	30.9

Case Study: Towards Predictive Modeling of "Big" Utilities Data - Statistical and Machine Learning Methods for Predicting Household Water Demand

Data

- Our dataset consists of water usage data from Tampa Bay Water
- One million unique customer records of monthly water usage
- Data stretches from approximately 1998 to 2010 (varies by parcel)
- 55 additional covariates, including parcel-specific and weather variables
- Weather data applied over 2km by 2km pixel grid

Training and Testing Sets

- The test and training sets were selected by year and month
- The test set is made up of the year stretching from Feb, 2009 to Jan, 2010
- The model was tested for each month in the test set separately
- One month ahead predictions were made for each month using all of the previous time points
- As a result the training set increases as we move further into the test set
- Using the one month ahead method simulates a real world situation where data is continuously added to the model
- k months ahead predictions were also done

Spatial and Temporal Structure I

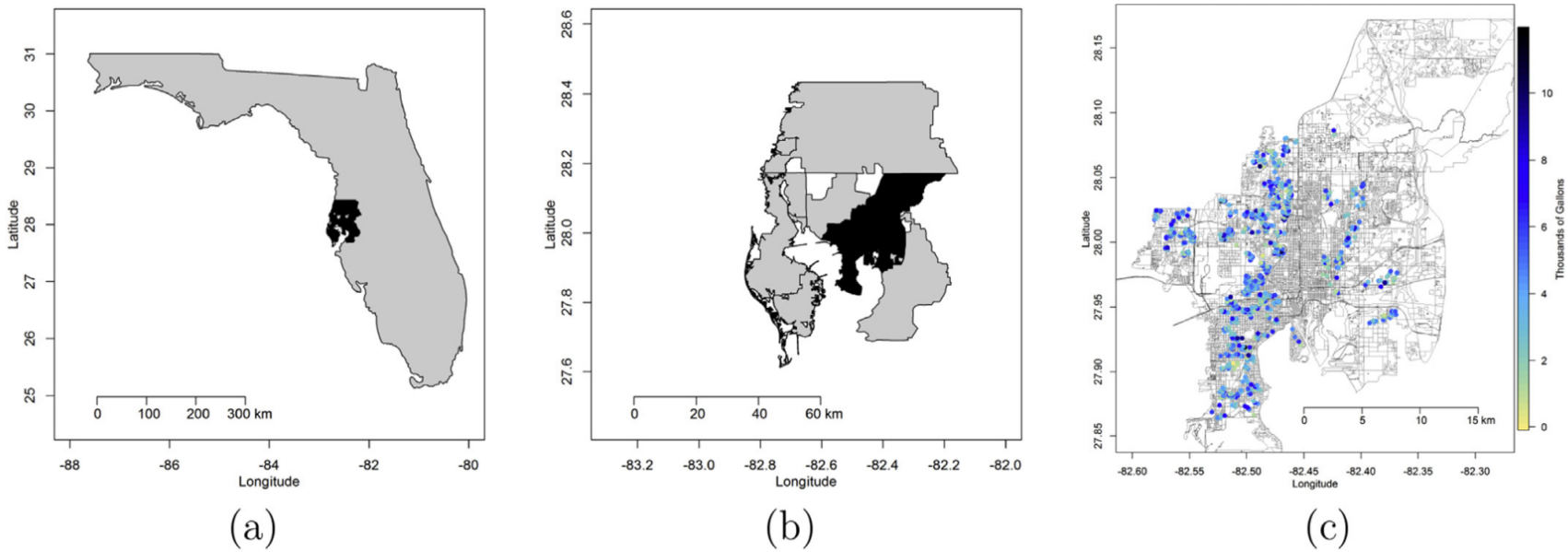
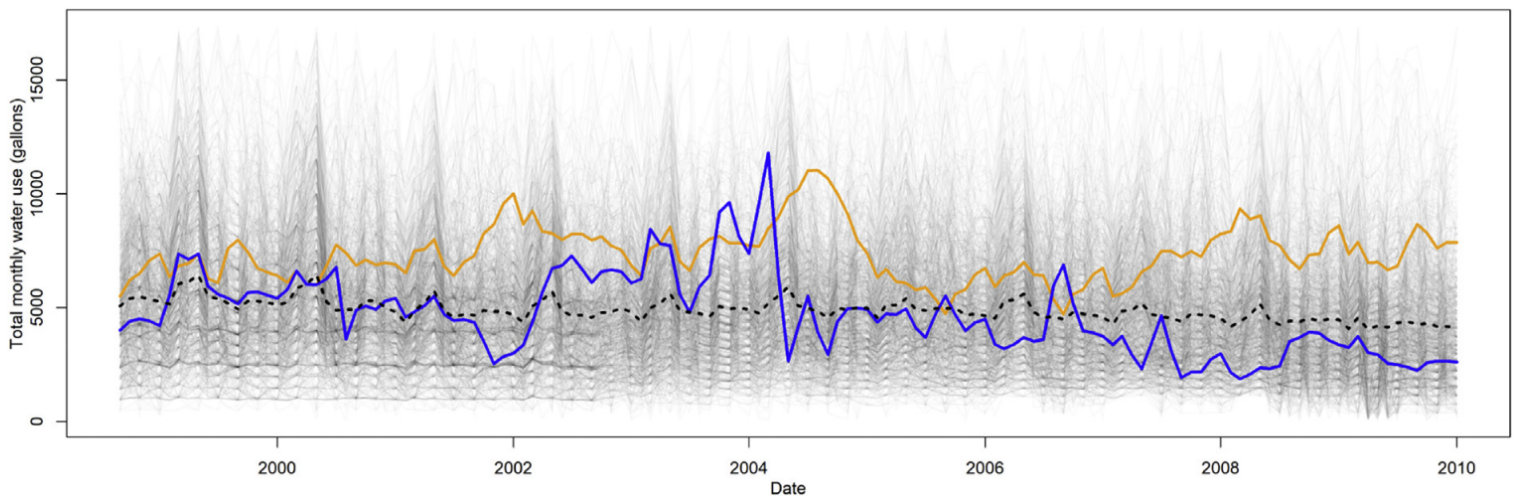


Fig. 1. (a) The region in Florida to which Tampa Bay Water supplies water. (b) The city of Tampa within the greater bay area. (c) Locations of households included in this study and average total monthly water use.



Spatial and Temporal Structure II

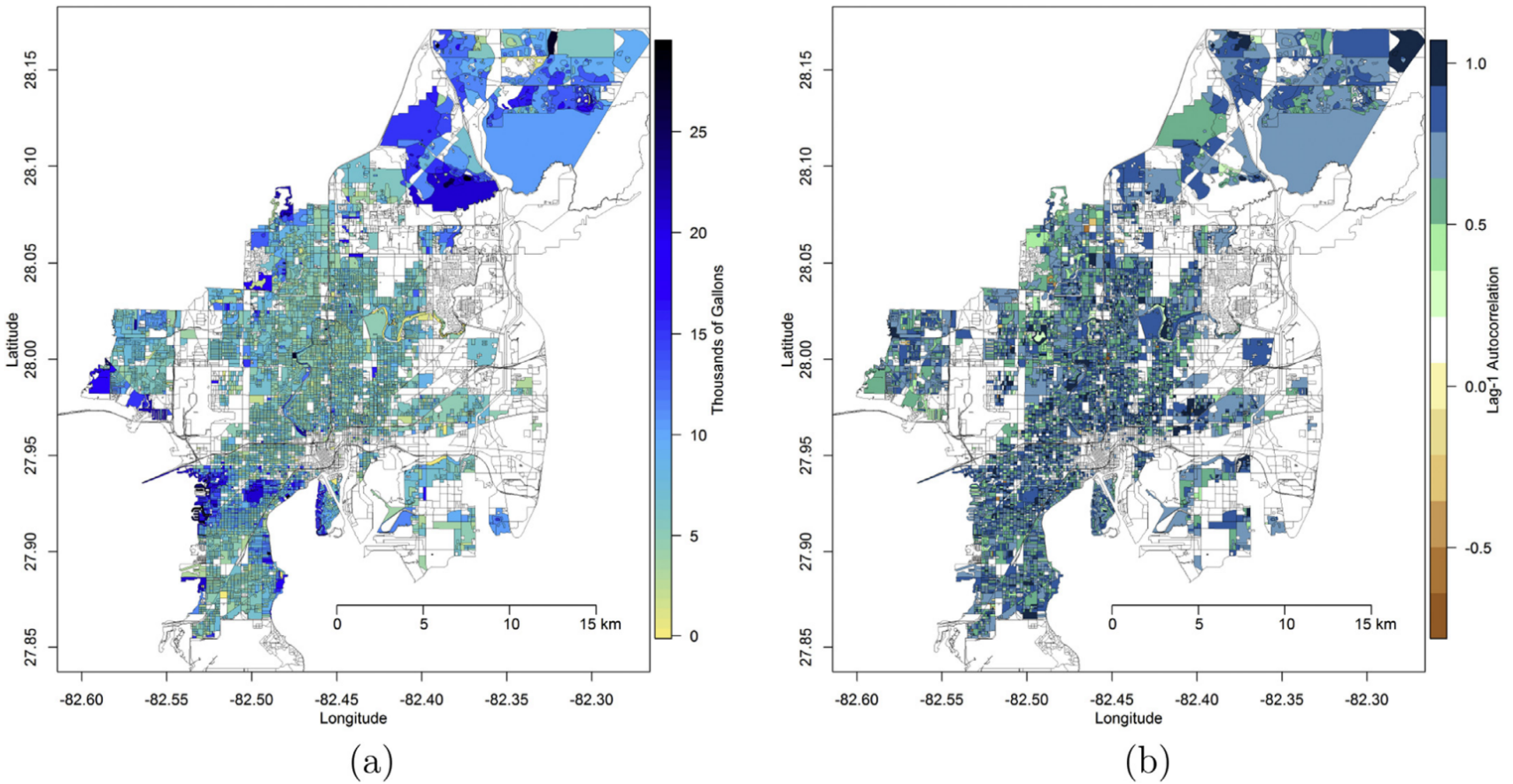


Fig. 3. (a) The average total water use per month for each census block from 1999 to 2009. (b) The estimated lag-1 autocorrelation coefficient from an exploratory AR (1) model fit to census-level data.

Prediction Quality Metrics for Comparison

Table 2

Definition of each metric used for comparing forecast quality across models. Each metric is computed using the same training and test set for each model. $\widehat{y}_i$ is the forecast for the i^{th} observation of total water use y_i, and (i) and $(\widehat{i})$ are the i^{th} indices of the observations ranked by the sizes of the observed (y) and predicted $(\widehat{y})$ values, respectively. $\widehat{y}_i^{(l)}$ and $\widehat{y}_i^{(u)}$ are the lower and upper bounds of the $(1-\alpha)100\%$ prediction intervals, respectively. $\mathbb{I}(\cdot)$ is the indicator function that takes value 1 if the condition in the argument is true, and 0 otherwise; e.g., if $f(x) = \mathbb{I}(1 \leq x \leq 3)$, then $f(0) = 0$ and $f(2) = 1$.

$$\text{RMSE} = \{\textstyle\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2/N\}^{1/2}$$

$$\text{GINI} = \textstyle\sum_{i=1}^{N}[\sum_{j=1}^{i} y_{\widehat{(j)}}/\sum_{j=1}^{N} y_j - (\widehat{i})/N]/\sum_{i=1}^{N}[\sum_{j=1}^{i} y_{(j)}/\sum_{j=1}^{N} y_j - (i)/N]$$

$$\text{AWPI} = \textstyle\sum_{i=1}^{N}(\widehat{y}_i^{(u)} - \widehat{y}_i^{(l)})/N$$

$$\text{ECPI} = \textstyle\sum_{i=1}^{N}\mathbb{I}(\widehat{y}_i^{(l)} \leq y_i \leq \widehat{y}_i^{(u)})/N$$

$$\text{NOIS} = \text{AWPI} + \tfrac{2}{\alpha}\textstyle\sum_{i=1}^{N}[(\widehat{y}_i^{(l)} - y_i)\mathbb{I}(\widehat{y}_i^{(l)} > y_i) + (y_i - \widehat{y}_i^{(u)})\mathbb{I}(y_i > \widehat{y}_i^{(u)})]/N$$

Prediction Quality Metrics for Comparison

Point measures:

- RMSE, the root mean squared error of predicted and observed values
- Gini, a score of how well the model correctly predicts the ordering of the results

Interval measures:

- AWPI, the average width of prediction intervals
- ECPI, the empirical coverage of prediction intervals
- NOIS, the negatively oriented interval score:

$$\text{NOIS}_i = y_i^{(upper)} + \frac{1}{\alpha}\left(y_i^{(obs)} - y_i^{(upper)}\right)\mathbb{I}_{\left[y_i^{(obs)} > y_i^{(upper)}\right]}$$

Performance: one month ahead prediction

1-month forecast assessments for each model, averaged over 12 months.

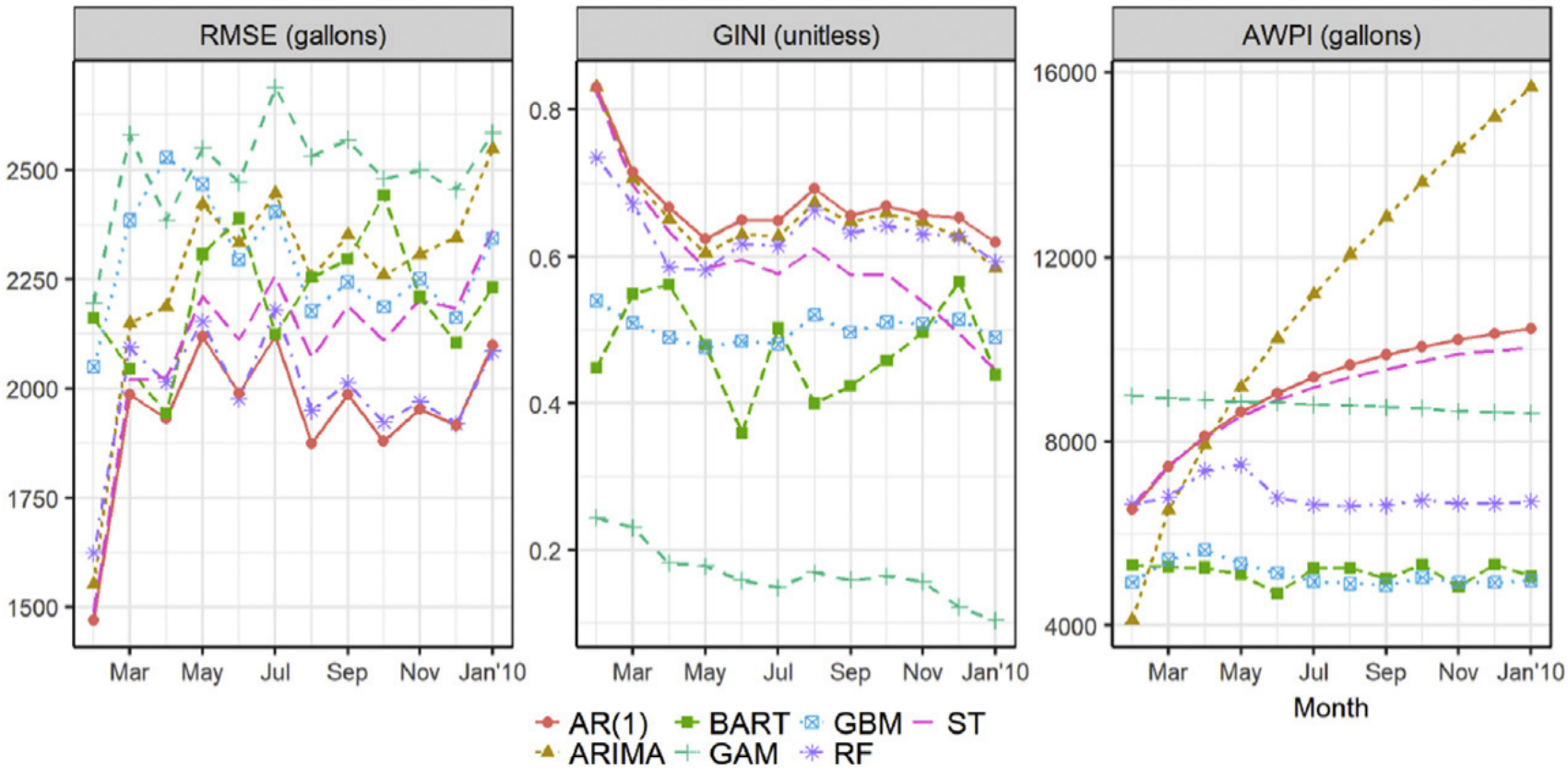
	RMSE	GINI	AWPI	ECPI	NOIS
Shared monthly means	2604.26	0.009	9156.01	0.954	10955.50
Individual monthly means	2234.36	0.590	8071.93	0.962	9827.94
Random effects	2199.75	0.602	7845.96	0.954	9758.60
Linear regression	2496.05	0.161	4572.85	0.625	20816.21
AR (1)	1246.21	0.879	6417.36	0.944	8073.36
ARIMA	1276.78	0.889	4170.48	0.922	7326.39
ST	1265.88	0.877	6467.76	0.949	8055.11
RF	1908.88	0.690	6892.60	0.910	9369.87
BART	2125.00	0.540	5037.23	0.710	15326.53
GBM	2275.28	0.513	4993.50	0.709	16257.64
GAM	2465.99	0.206	8562.03	0.936	10972.02

Performance: k months ahead prediction I

k-months forecast assessments for each model, averaged over 12 months.

	RMSE	GINI	AWPI	ECPI	NOIS
Shared monthly means	2617.49	0.004	9175.17	0.954	10961.47
Individual monthly means	2315.70	0.556	8057.30	0.955	9971.84
Random effects	2281.43	0.567	7829.43	0.950	9903.72
Linear regression	2500.03	0.157	4606.16	0.630	20613.25
AR (1)	1950.74	0.674	9149.45	0.958	10668.27
ARIMA	2275.85	0.658	11063.49	0.935	10911.00
ST	2113.48	0.597	8948.82	0.962	10312.72
RF	1996.65	0.634	6806.43	0.887	9810.98
BART	2213.64	0.474	5146.76	0.723	15214.26
GBM	2295.72	0.503	5096.52	0.677	19747.03
GAM	2502.06	0.168	8791.54	0.939	11083.02

Performance: k months ahead prediction II



Case Study: Predictive Modeling of Air Pollution by Black Carbon in the Greater Boston Area Using Data from Multiple Sources

Problem and Goals

Goal

Build a model for spatio-temporal prediction of the particulate matter air pollution (black carbon, BC) process.

Motivation

- When controlled for other risk factors, exposure to particulate matter in the air in urban areas has been shown to be positively associated with elevated mortality and morbidity.
- For locally-generated pollutants, spatial predictions using only a central-site monitor can incur significant measurement error.
- Models based on a single study often fail to provide adequate coverage over a given spatial region and time period.

Approach

Use Bayesian hierarchical modeling to combine information from multiple related datasets.

What is BC?

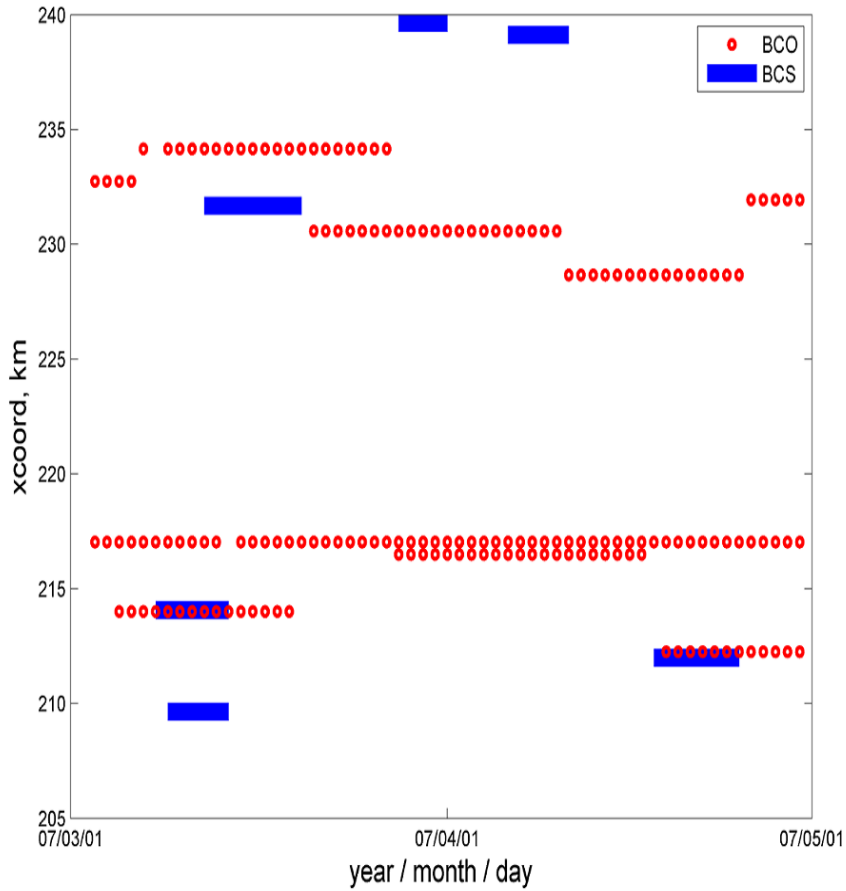
BC particles are formed through incomplete combustion of fuels

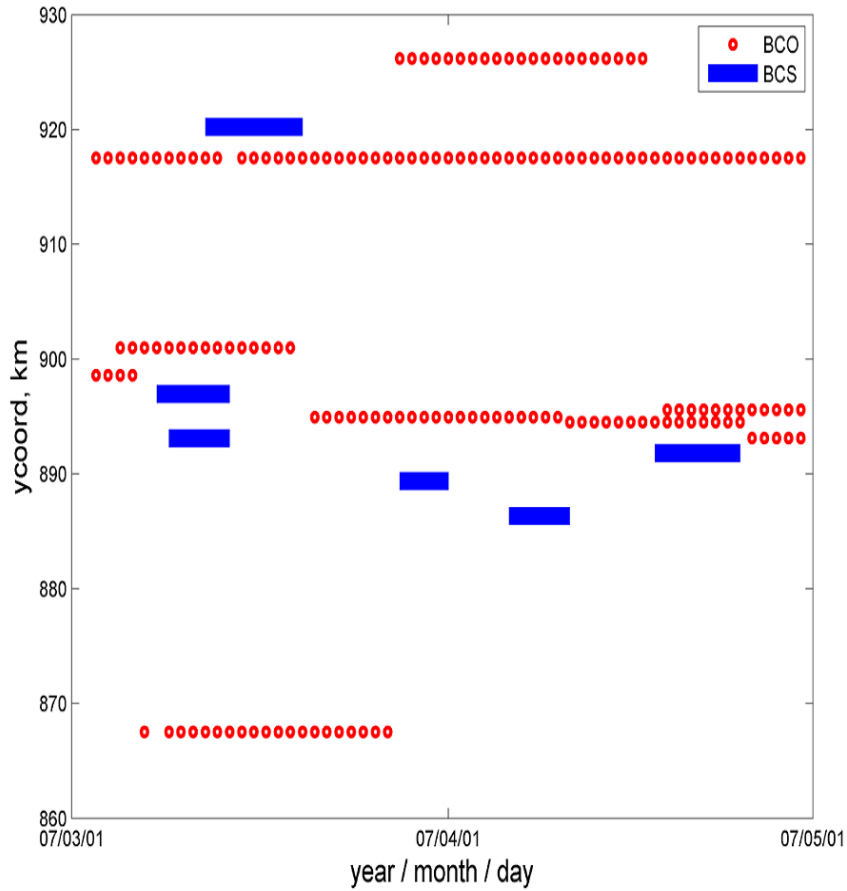
Data

Monitors are scattered irregularly in space and operate irregularly in time.

- BCO - Outdoor concentrations of BC from 3 studies.
  - over 6,500 daily averages from a total of about 80 sites
  - bulk of the data comes from ≤ 15 sites
- BCI - Indoor concentrations of BC
  - ≈ 300 daily averages from a total of 45 distinct households
  - monitoring sites overlap spatially with 30 BCO sites
- BCS - average multi-day concentrations of indoor BC. Equivalently, Sums of daily indoor BC concentrations.
  - 1 reading per household from about 100 households
  - data aggregated over multiple days; individual daily average concentrations are not available
  - average monitoring time per household - 7 days

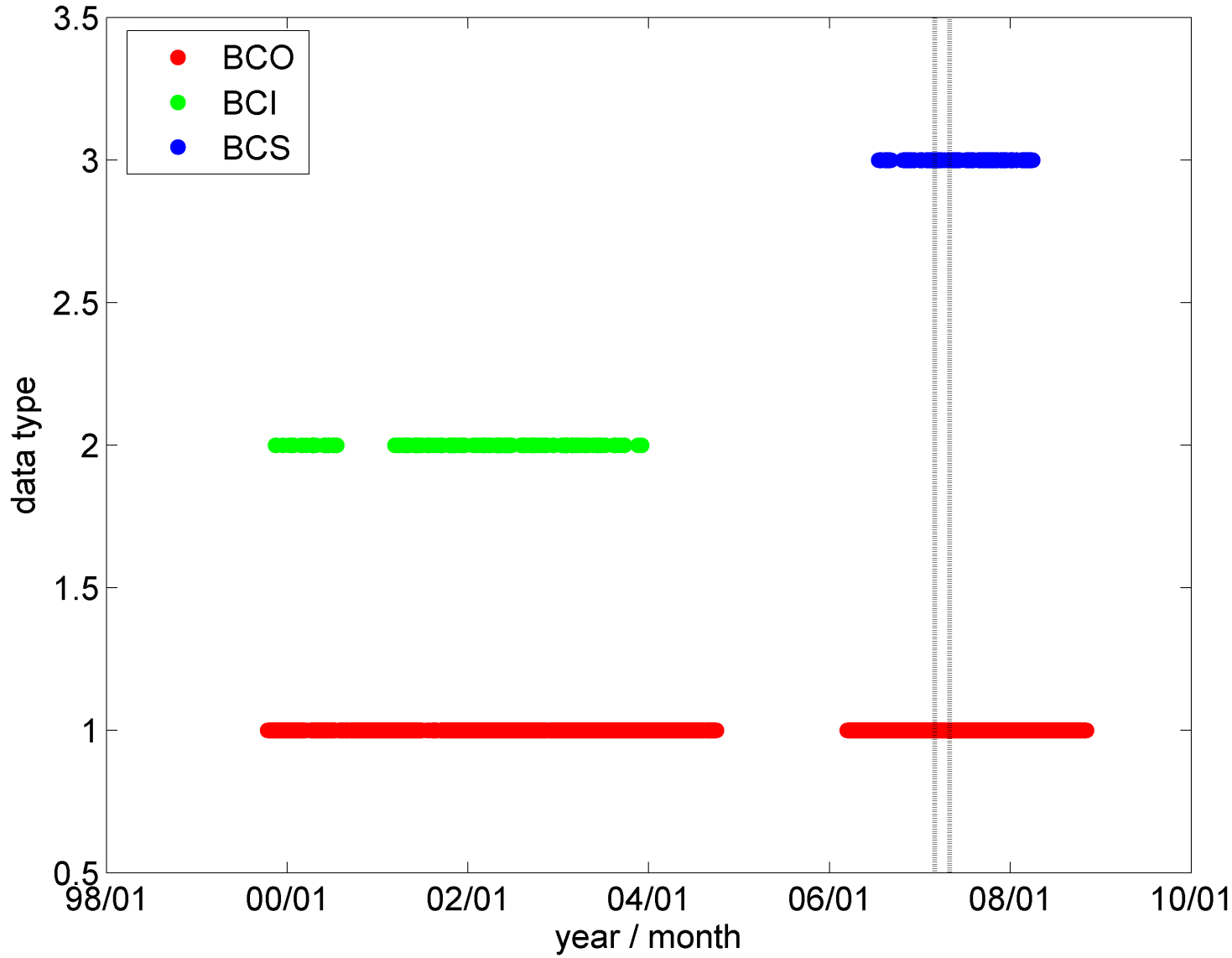
Illustration: s-t coverage by BCO and BCS monitors



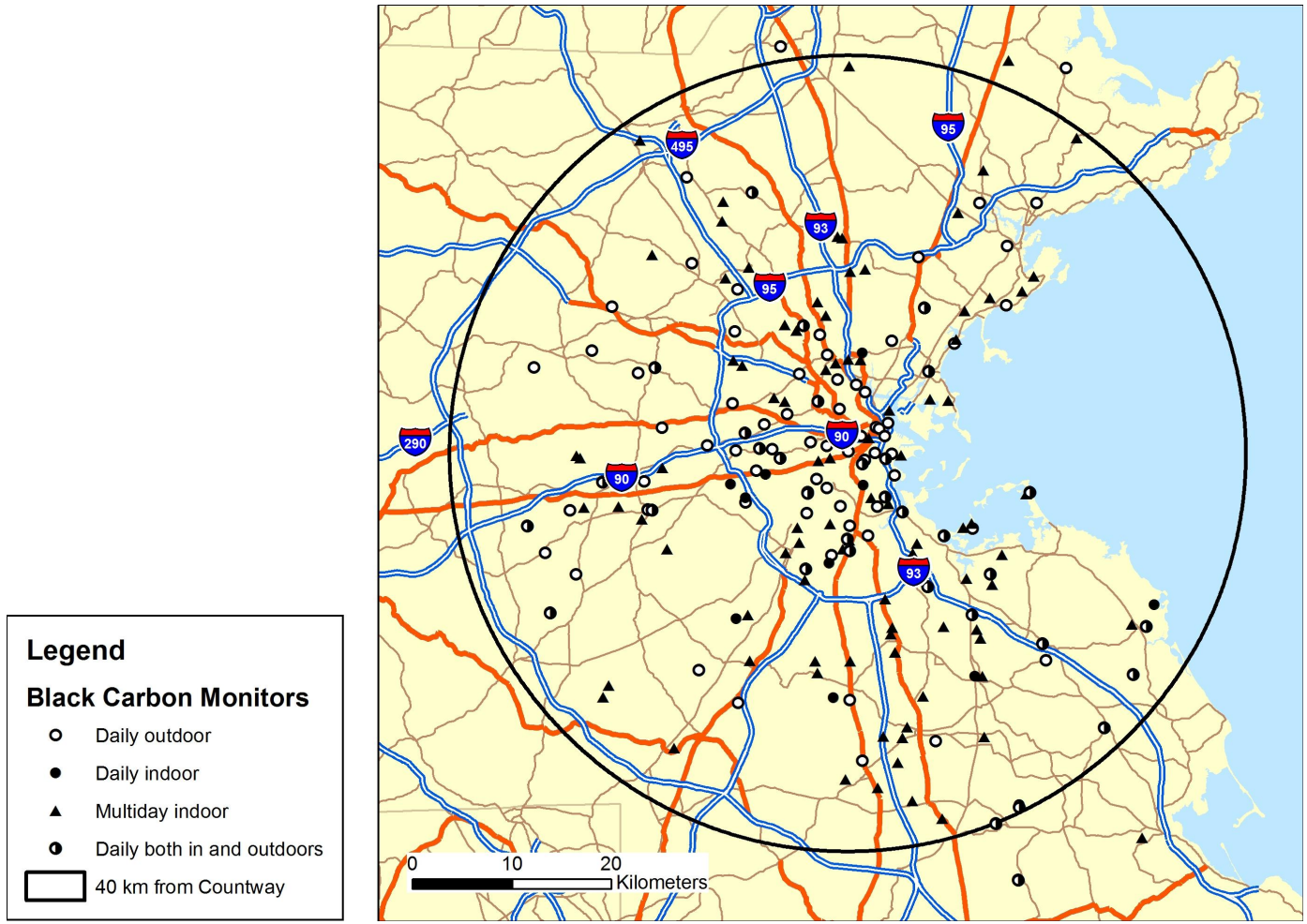


$\Rightarrow$ Need to solve the change of support problem (e.g., Gelfand, Zhu, Carlin, 2001, and references therein).

Temporal Coverage by Monitor Type



Spatial Distribution of Monitors



$\Rightarrow$ Use all data to reduce the prediction error of the unknown exposure process (in spatio-temporal misalignment problem).

A Null Model - Gryparis et al., (2007)

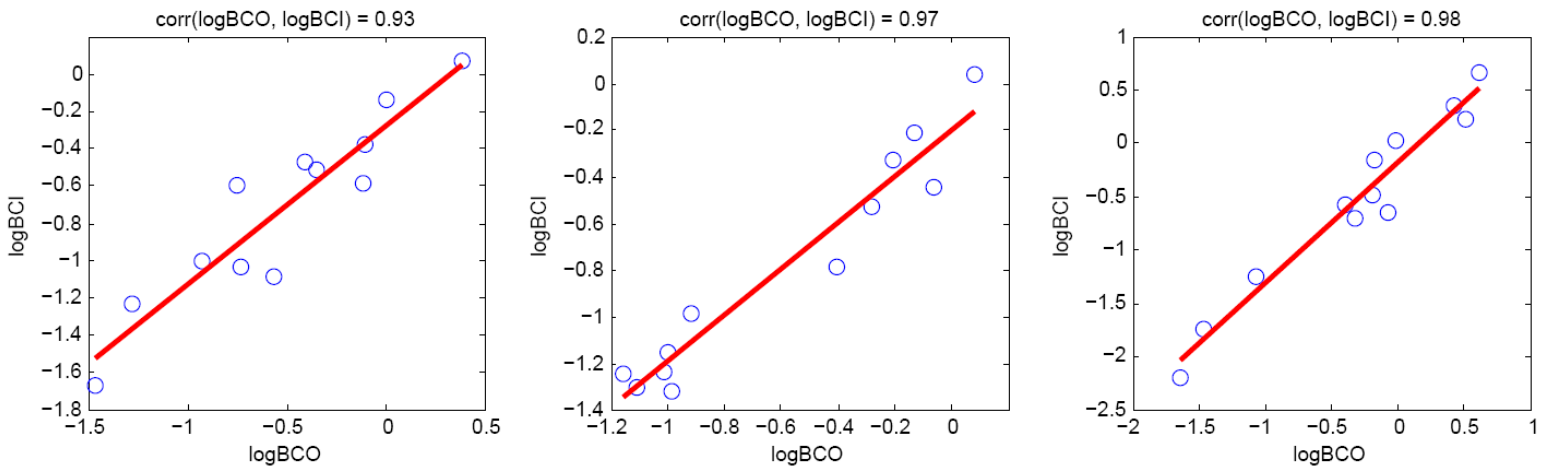
$$\eta_{ij} = c_{ij}^{\mathsf{T}} w + \epsilon_{ij}^{\eta} \qquad \text{(latent process)}$$

$$Y_{ij}^{O} = \eta_{ij} + \epsilon_{ij}^{O} \qquad \text{(log outdoor BC)}$$

$$Y_{ij}^{I} = \alpha_{0i} + \alpha_1 \eta_{ij} + \epsilon_{ij}^{I} \qquad \text{(log daily indoor BC)}$$

- Indexing: i for spatial site, j for day, e.g., $\epsilon_{ij}^{\eta} = \epsilon^{\eta}(\texttt{xy}_i, \texttt{time}_j)$
- Gryparis et al. (2007), set $\epsilon_{ij}^{\eta} = 0$.

Validating the daily indoor BC model



A Nonlinear Statistical Model

$$\eta_{ij} = c_{ij}^{\mathsf{T}} w + \epsilon_{ij}^{\eta} \qquad \text{(latent process)}$$

$$Y_{ij}^{O} = \eta_{ij} + \epsilon_{ij}^{O} \qquad \text{(log outdoor BC)}$$

$$Y_{ij}^{I} = \alpha_{0i} + \alpha_1 \eta_{ij} + \epsilon_{ij}^{I} \qquad \text{(log daily indoor BC)}$$

$$Y_{i}^{S} = \alpha_{0i} + g_i(\eta_i; \alpha_1) + \epsilon_{i}^{S} \qquad \text{(log multiday indoor BC)},$$

where $g_i(\eta_i; \alpha_1) = \log \sum_j \exp(\alpha_1 \eta_{ij})$

Motivation

Suppose $\epsilon_{ij}^{I} = 0. \Rightarrow Y_{ij}^{I} = \alpha_{0i} + \alpha_1 \eta_{ij}$
$\Rightarrow Y_{i}^{S} = \alpha_{0i} + \log \sum_j \exp(\alpha_1 \eta_{ij}).$

Here, $\epsilon^{\eta}(\cdot, \cdot)$ is a continuous Gaussian spatio-temporal process.

Structure of the Latent Process

The linear model for the latent process is
$\eta(\mathtt{xy}_i, \mathtt{time}_j) = c(\mathtt{xy}_i, \mathtt{time}_j)^{\mathsf{T}} w + \epsilon^{\eta}(\mathtt{xy}_i, \mathtt{time}_j)$.

Consider the semiparametric model $c(\mathtt{xy}_i, \mathtt{time}_j)^{\mathsf{T}} w = obs(\mathtt{xy}_i, \mathtt{time}_j)^{\mathsf{T}} \beta + f_S(\mathtt{xy}_i) + f_T(\mathtt{time}_j) + f_{ST}(\mathtt{xy}_i, \mathtt{time}_j)$, where

- $obs(\mathtt{xy}_i, \mathtt{time}_j)$ is a vector of observable predictors; e.g., meteorology, population, distance to roadway.
- f_S, f_T and f_{ST} are smooth spatial, temporal and spatio-temporal trends/surfaces.
- Represent the smooth trends using basis functions, ϕ and ψ.
  - $f_S(\mathtt{xy}) = \boldsymbol{\phi}(\mathtt{xy})^{\mathsf{T}} w_S$,
  - $f_T(\mathtt{time}) = \boldsymbol{\psi}(\mathtt{time})^{\mathsf{T}} w_T$ (annual cyclic trend),
  - $f_{ST}(\mathtt{xy}, \mathtt{time}) = \{\boldsymbol{\phi}(\mathtt{xy})^{\mathsf{T}} \otimes \boldsymbol{\psi}(\mathtt{time})^{\mathsf{T}}\} w_{ST}$.
- Thus, $w = (\beta^{\mathsf{T}}, w_S^{\mathsf{T}}, w_T^{\mathsf{T}}, w_{ST}^{\mathsf{T}})^{\mathsf{T}}$.

Bayesian Paradigm

$\boldsymbol{\theta}$: parameters of interests; $\mathbf{D}$: data

Data model (likelihood): $\mathbf{D} \,|\, \boldsymbol{\theta} \sim p(\mathbf{D} \,|\, \boldsymbol{\theta})$

- Prior knowledge about $\boldsymbol{\theta}$:
  - Bayesian: a random variable $\sim \pi(\boldsymbol{\theta})$ (prior distribution).
  - Frequentist: a fixed value $\boldsymbol{\theta}_0$.
- Bayes theorem: prior distribution + likelihood $\Longrightarrow$ posterior distribution with a pdf/pmf

$$p(\boldsymbol{\theta} \,|\, \mathbf{D}) = \frac{p(\mathbf{D} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_\Theta p(\mathbf{D} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

- Main advantage: conceptually easy uncertainty quantification through the posterior distribution and predictive distribution.

Bayesian Inference

- $p(\boldsymbol{\theta} \mid \mathbf{D}) \sim$ a simple and common distribution. The posterior mean, confidence interval, and other key quantities are available in explicit expressions.
- If they are not explicitly available, use their sample versions.
  - Sometimes it is possible to draw iid samples from $p(\boldsymbol{\theta} \mid \mathbf{D})$.
  - If not, Markov Chain Monte Carlo (MCMC) algorithms are used to draw dependent and approximated samples.
    - Gibbs sampling: $\theta_1^{(g+1)} \sim p(\theta_1 | \boldsymbol{\theta}_{-1}^{(g)}, \mathbf{D}), \ldots, \theta_p^{(g+1)} \sim p(\theta_p | \boldsymbol{\theta}_{-p}, \mathbf{D})$.
    - Metropolis-Hasting algorithm: Given the gth sample $\boldsymbol{\theta}^{(g)}$, propose a new state $\tilde{\boldsymbol{\theta}}$ from some proposal distribution; accept it as the $(g+1)$th sample with probability α, which depends on the proposal and the posterior distribution; otherwise $\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)}$.
    - Metropolis-Hasting within Gibbs sampling

Summary and Conclusions

- “Big Data" is a hot topic of research . . .
- . . . but “big" or “small" is relative to the approach/algorithm.
- Another somewhat vague dimension to consider - “messiness" (e.g., due to data “missingness").
- “Big" and “messy" data are often intractable.
- If “messy", “small" data can become “big".
- Other complicating features: dependence (temporal, spatial), modeler’s philosophy/beliefs (“Bayesianity").
  Only simple models are tractable with “big" data.
- A fancy ML method is not necessarily the best.
- Is a single point-level test error summary (RMSE or misclassification rate) sufficient?

$\Rightarrow$ Working with “messy" data is hard. Each study presented was very labor intensive and required over a year of full-time work (typically more).

Thank you!