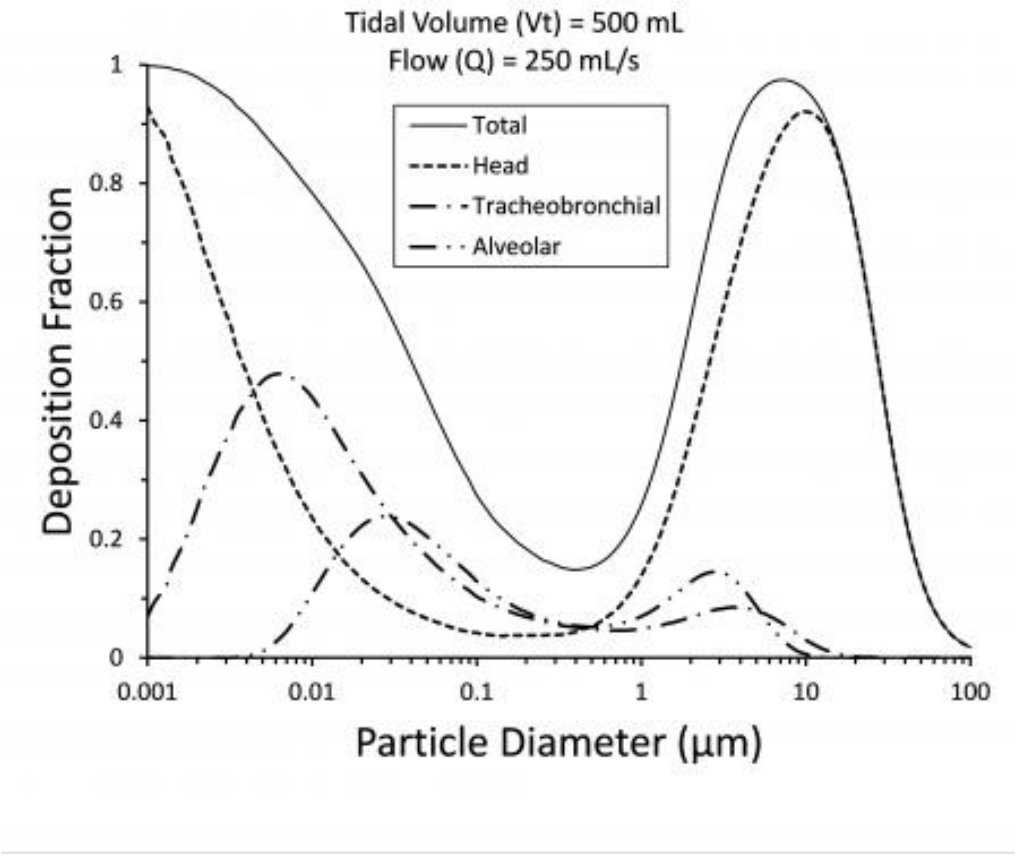
A spatio-temporal prediction model based on support vector machine regression:
Ambient Black Carbon in three New England States.

Yara Abu Awad

What is Black Carbon?



- **Is an air pollutant**
- **Generated by combustion (more on this later)**
- **Small particles that enter the alveoli of the lung**

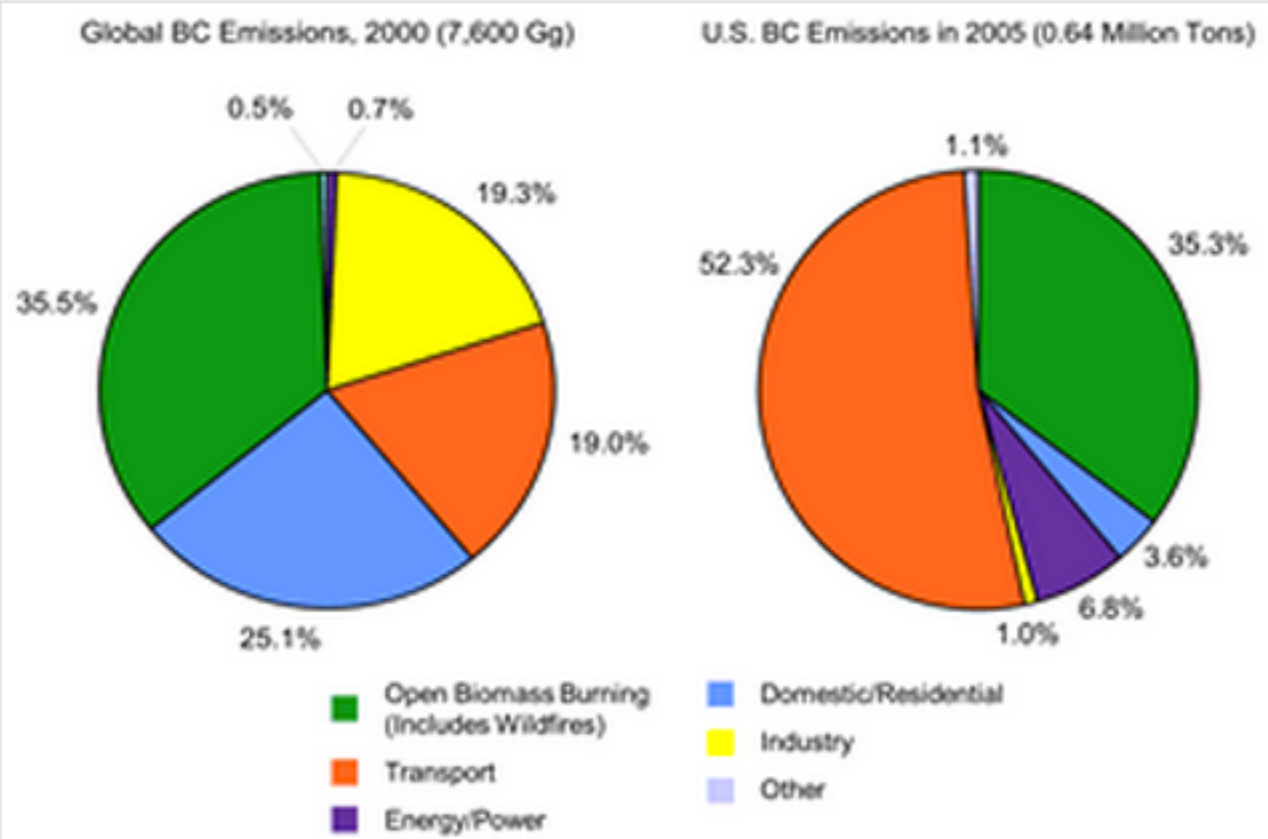
Image source: www.epa.gov

Why do we care about Black Carbon?

- **Acute exposure effects**: respiratory[1], cardiovascular (CVD) and total mortality[2]
- **Chronic exposure effects**: blood pressure[3], lung function[4,] cognitive function[5], all-cause, cardiovascular and lung cancer mortality[6]

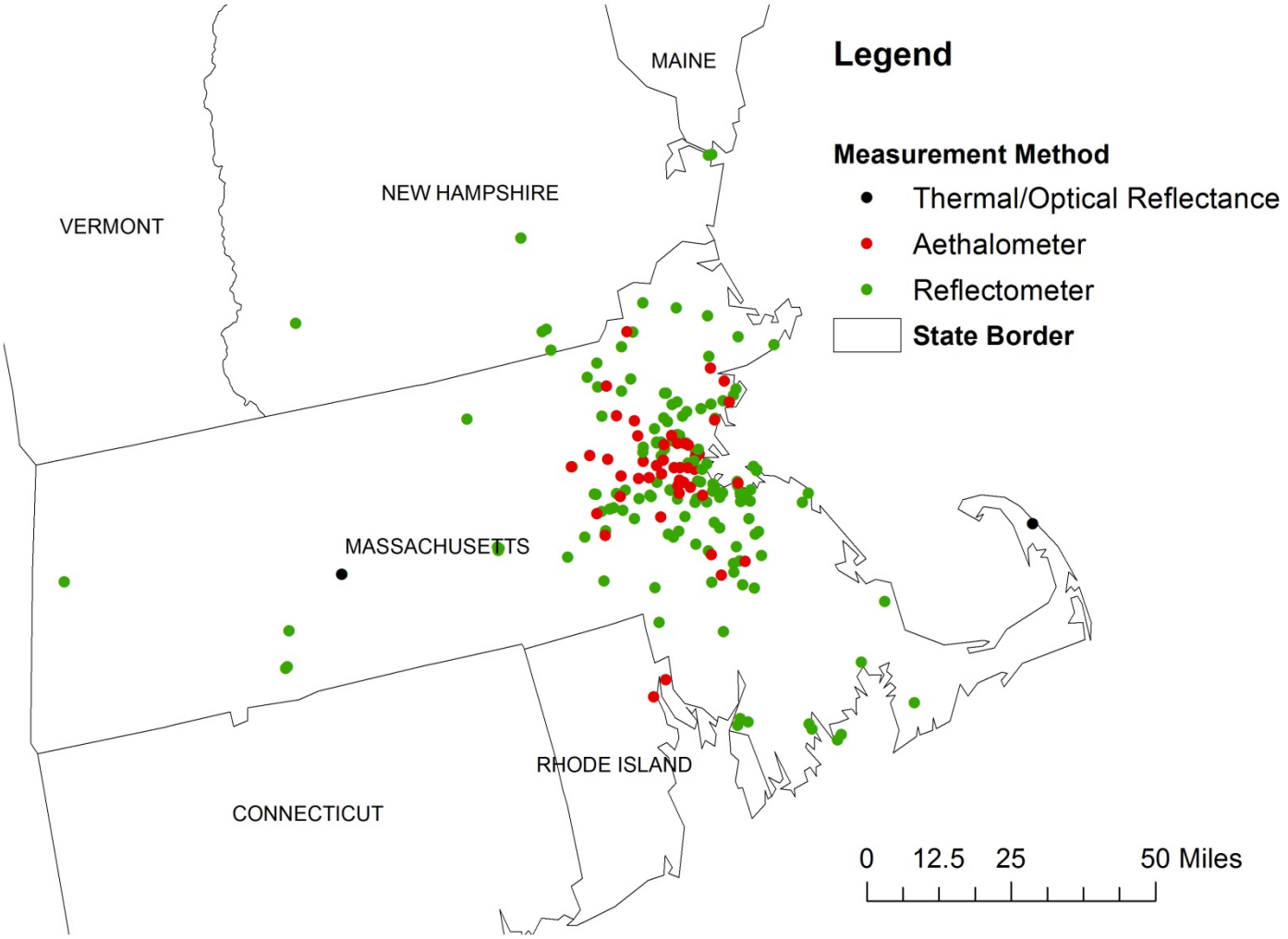
1. Bremner et al. (1999), 2. Maynard, Coull, Gryparis, & Schwartz (2007, 3. Schwartz et al. (2012), 4. Lepeule et al. (2014), 5. Power et al. (2010), 6. Beelen et al. (2008); Filleul et al. (2005); Smith et al.,(2009)

Black Carbon

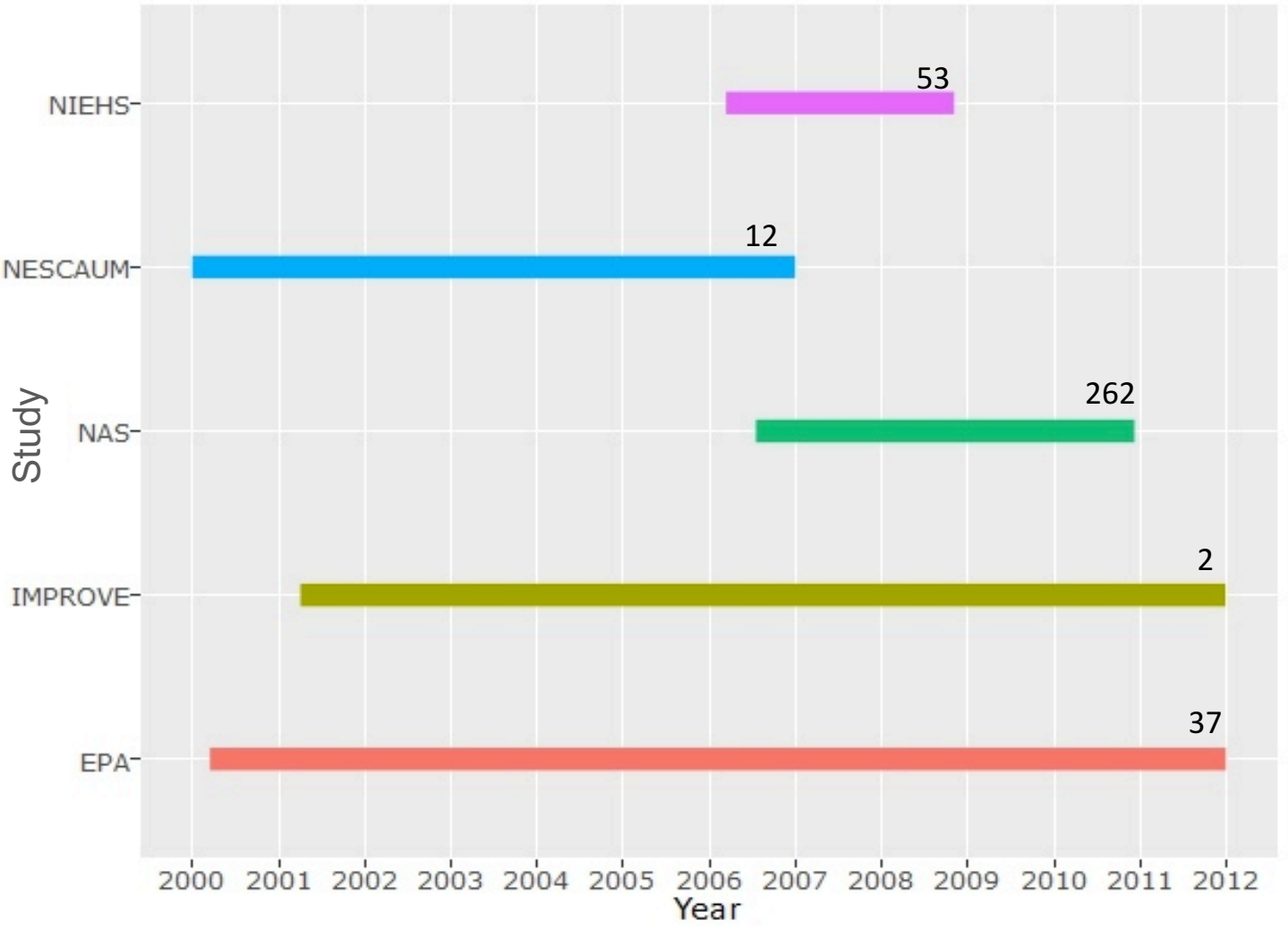


BC Emissions by Major Source Category
(Source: Lamarque et al., 2010 and US EPA)

Monitoring data



Data Sources



24,301 measurement days

Predictors

Temporal

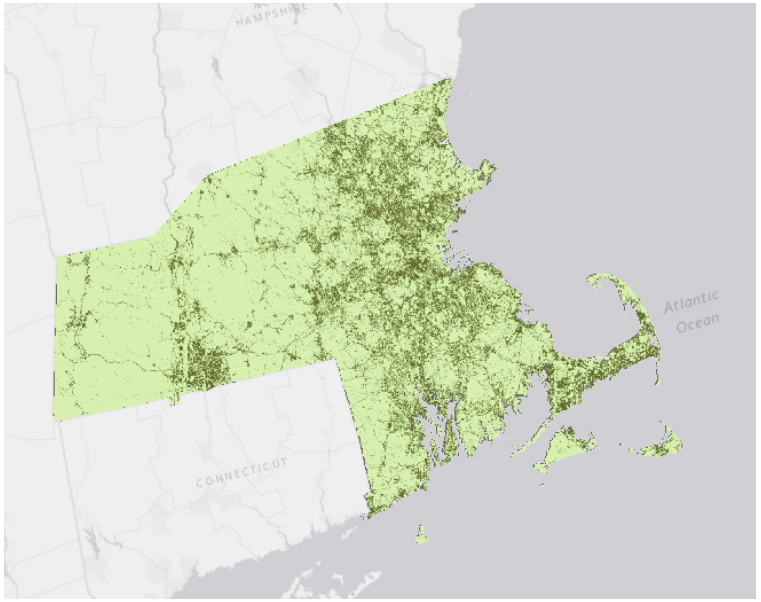
- Temperature, wind speed, visibility, dew point, sea-level pressure, and relative humidity
- height of the planetary boundary Layer
- average daily BC and $PM_{2.5}$ concentrations at one location
- Weekday / weekend
- Cold (Nov - Apr) and warm (May - Oct)

Spatial

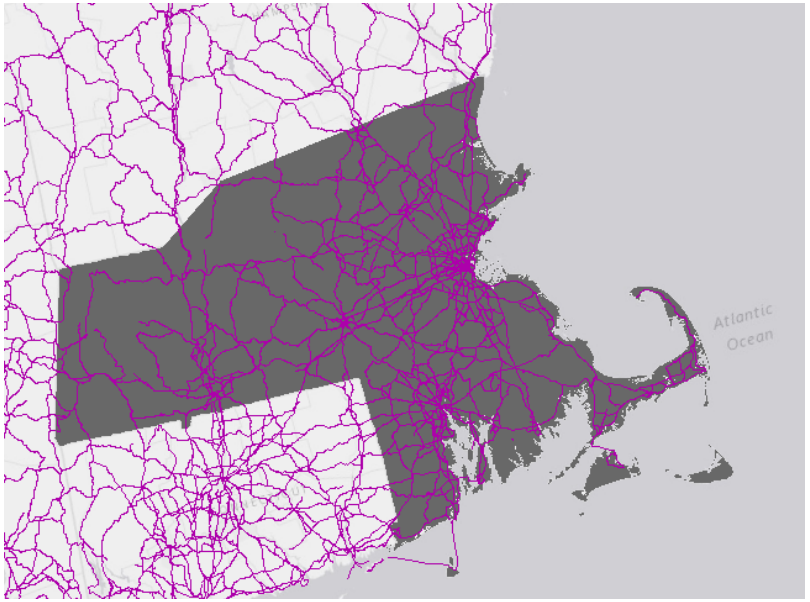
- Proximity to transportation
- Topographical characteristics
- Neighborhood characteristics

Spatial Predictors

% Low Development



Truck Routes



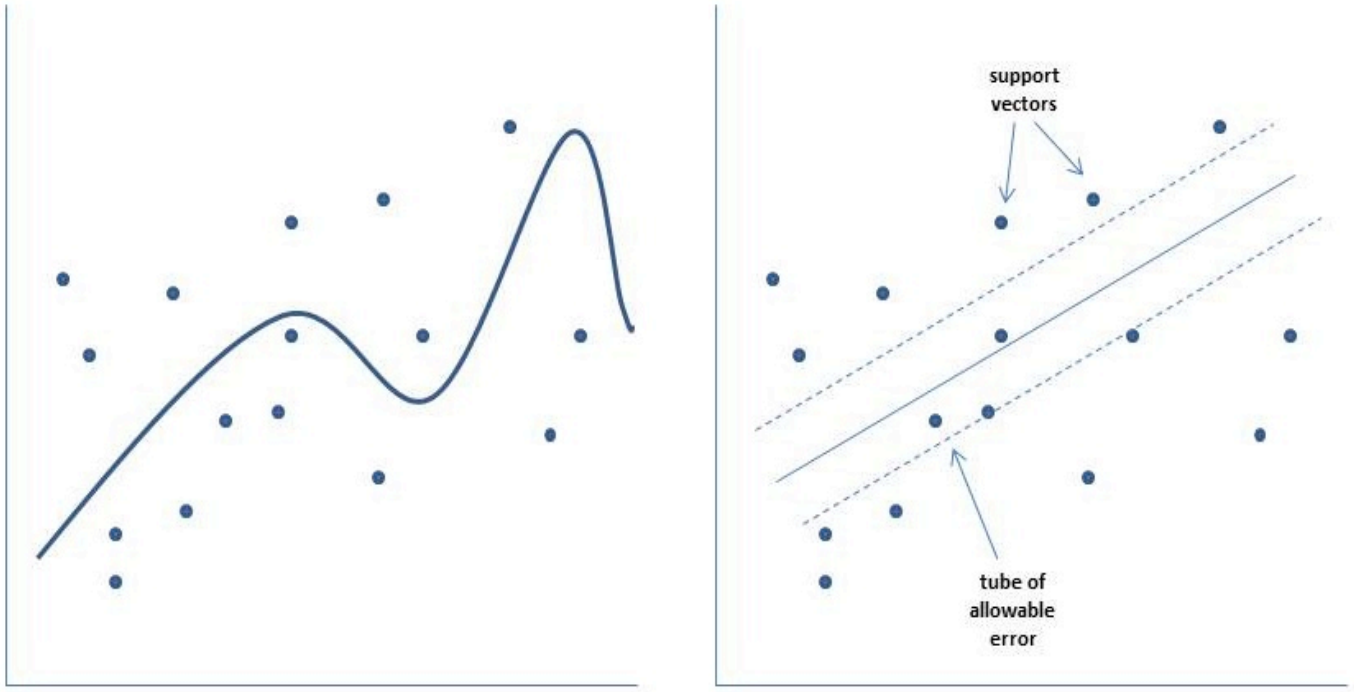
nu-Support Vector regression

- Mapping data into a multidimensional feature space
- Specifying allowable error: % of observations estimated with error
- Ridge penalty

Gaussian Kernel

$K(x,x') = \exp(-\gamma|x - x'|^2)$,

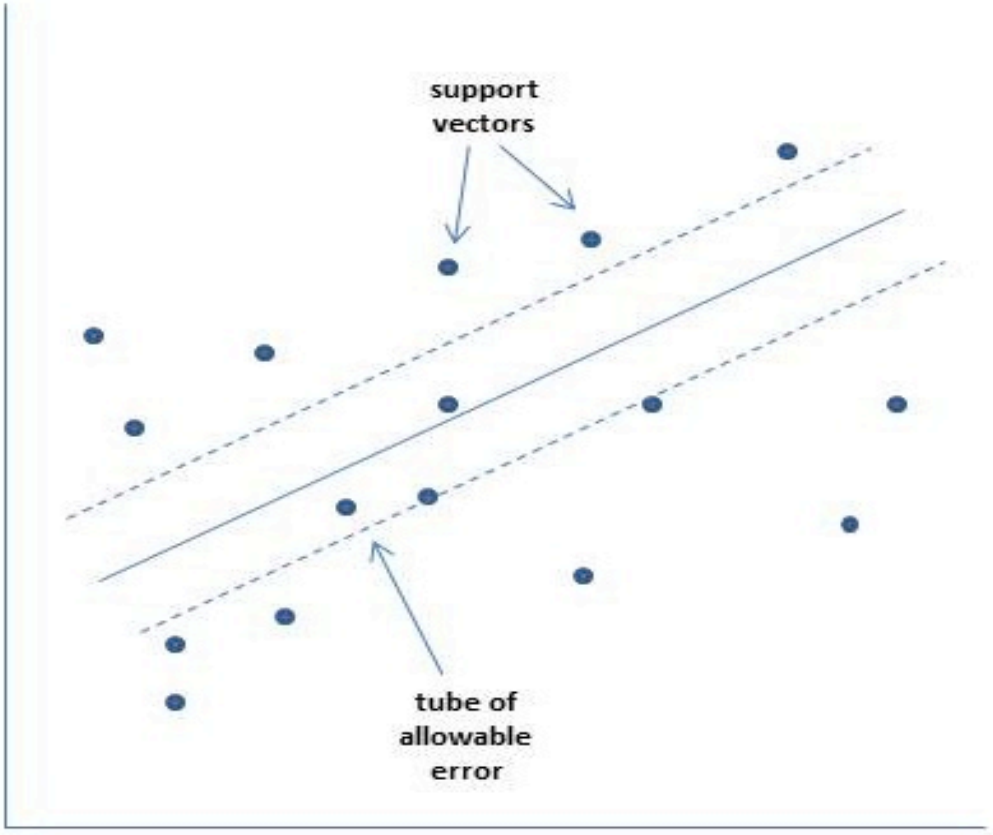
where $\gamma = \frac{1}{2\sigma^2}$



a) untransformed space

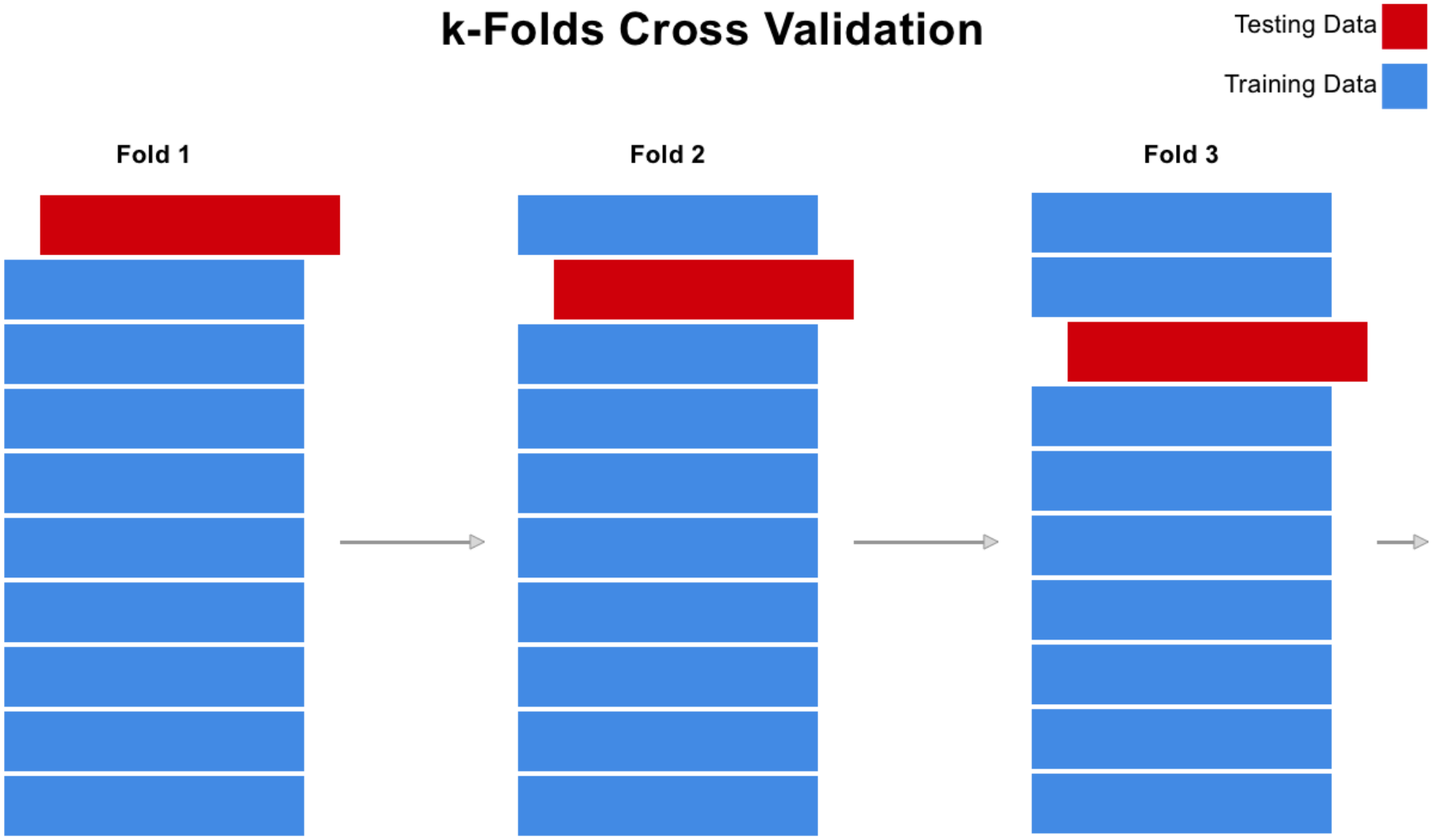
b) transformed space

Minimizing Error

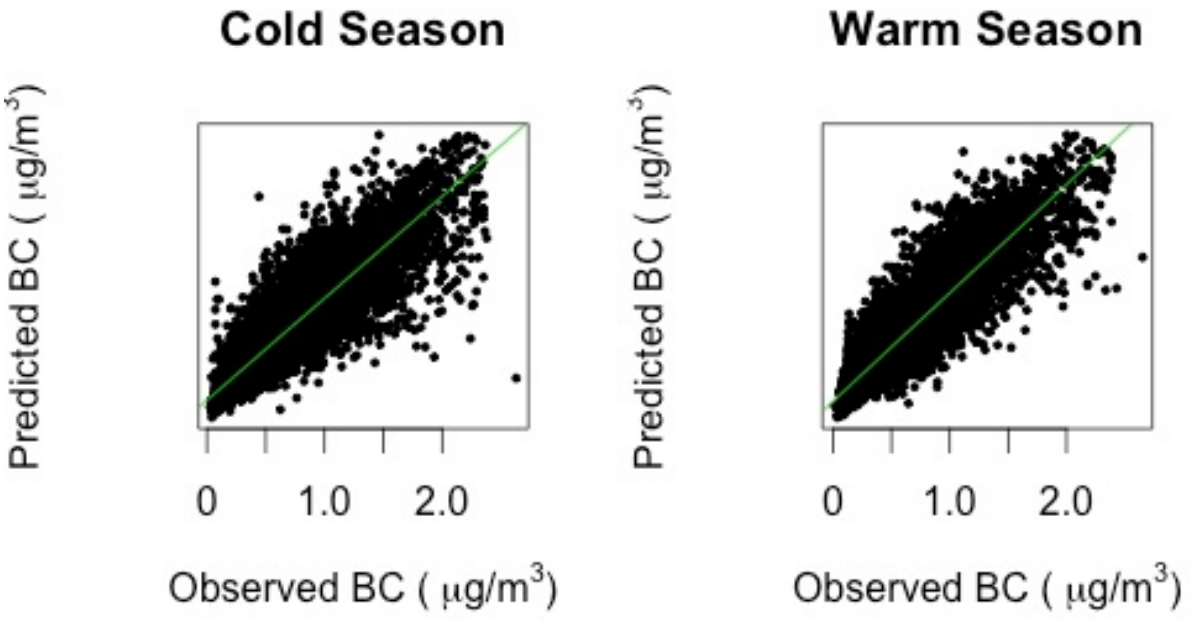


b) transformed space

Assessing Model Performance



Model Performance: Predicted vs Observed



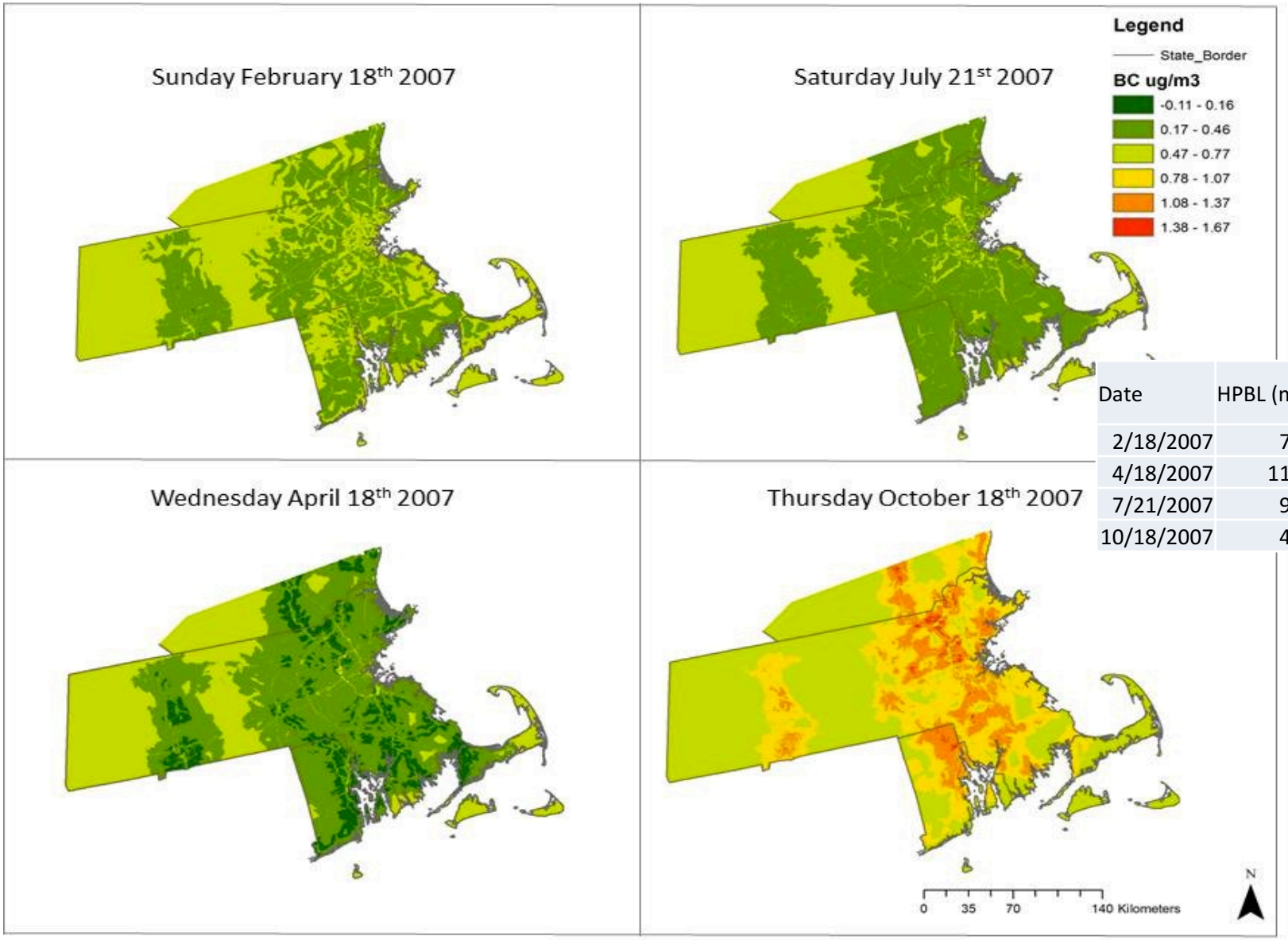
10-fold CV R^2 = 0.73 and 0.75

Sensitivity to Gamma and Nu

nu	gamma	R^2
0.65	0.03	0.72978
0.65	0.06	0.71303
0.65	0.015	0.72489
0.33	0.03	0.72511
0.95	0.03	0.72957

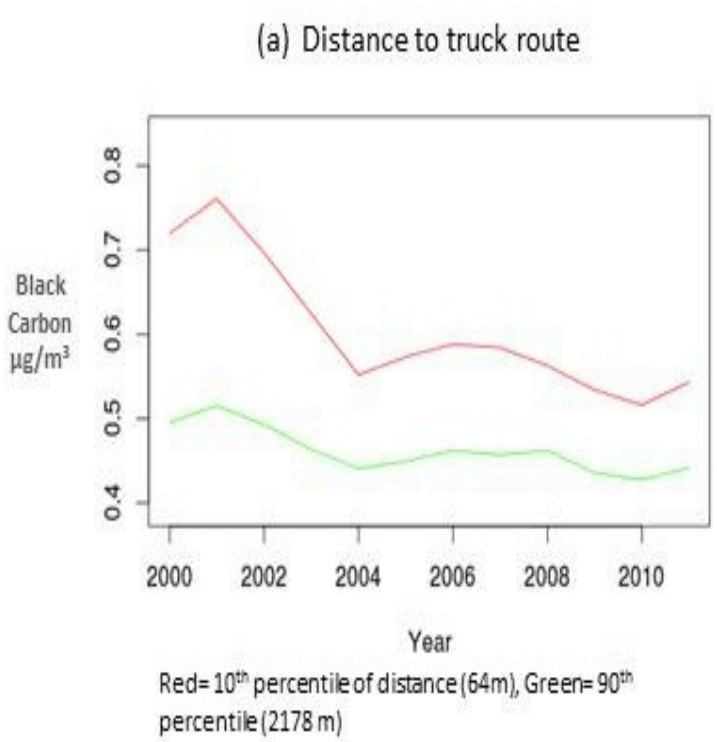
Gamma controls smoothness and how quickly coefficients move towards zero
Nu is the upper bound on the fraction of errors and the lower bound on the fraction of support vectors

Predicted Black Carbon

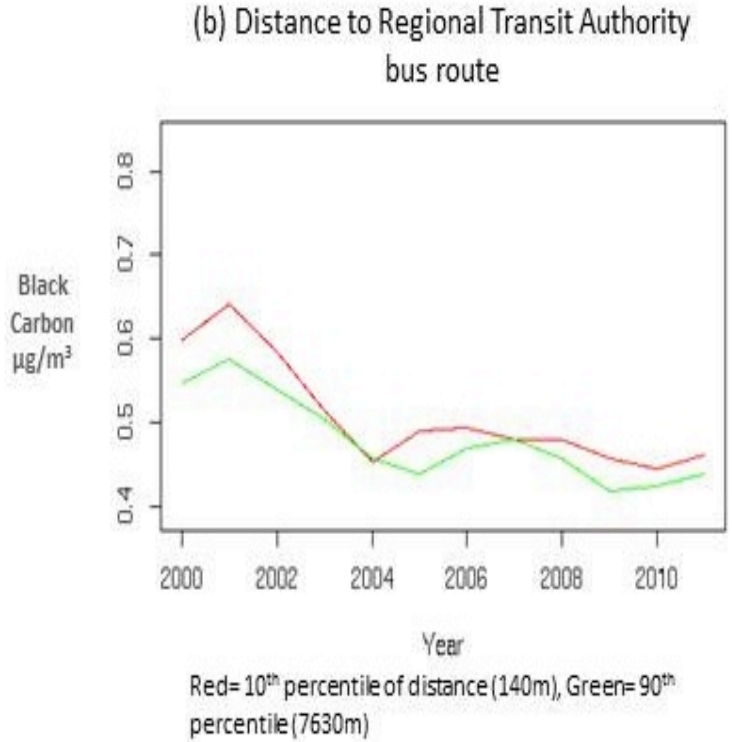


Date	HPBL (m)
2/18/2007	777
4/18/2007	1153
7/21/2007	924
10/18/2007	455

Capturing spatial variance



Red= 10th percentile of distance (64m), Green= 90th percentile (2178 m)



Red= 10th percentile of distance (140m), Green= 90th percentile (7630m)

Update: Ensemble Modeling

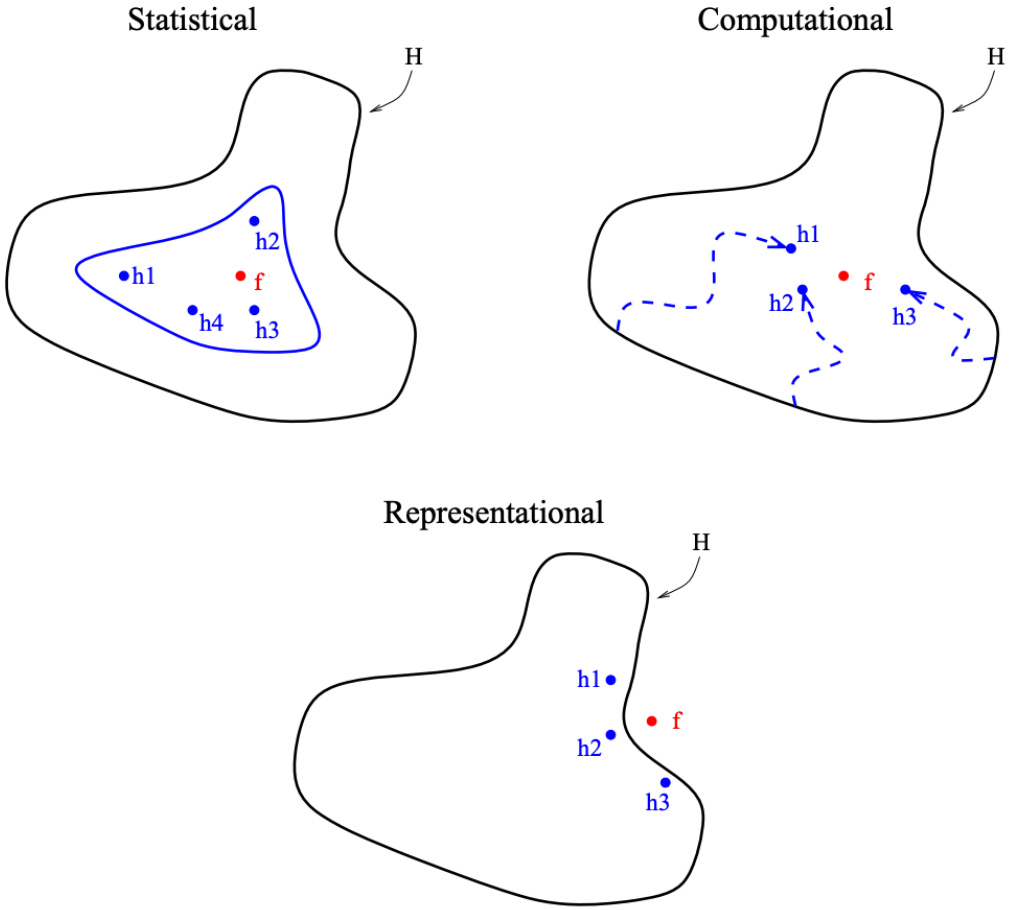


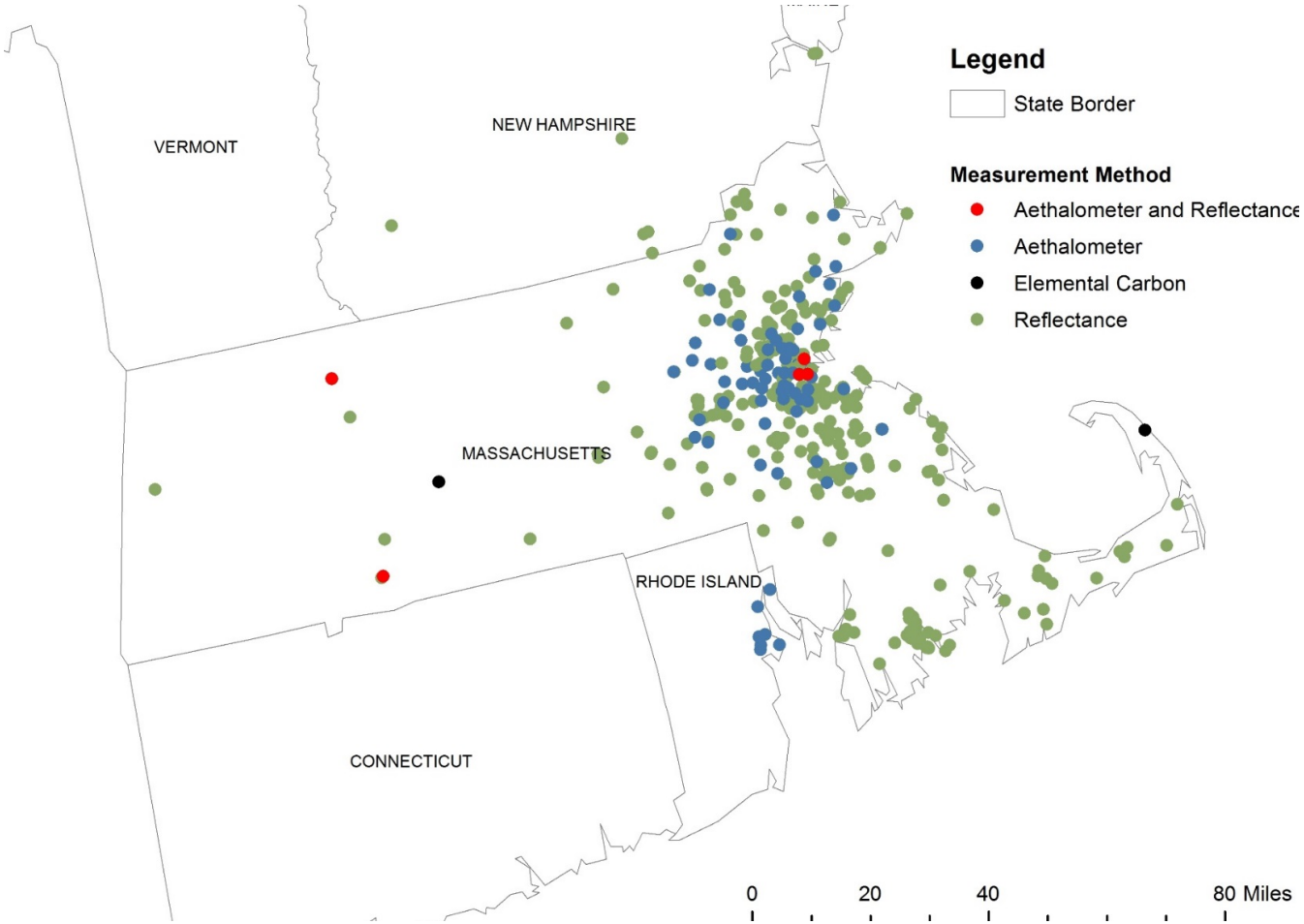
Fig. 2. Three fundamental reasons why an ensemble may work better than a single classifier

Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, 2000.

Updated Model

- Extended model to 2015
- 25,000 observations added for a total of ~47,000 observations
- An ensemble ML method which leverages the capabilities of multiple learners

Updated Monitoring Data



Algorithms Selected

- nuSVR
- Random Forest
- Gradient Boosting
- Deep Learning
- Generalized Additive Model (GAM)

Algorithms Selected

- nuSVR -> R Package e1071
- Random Forest
- Gradient Boosting
- Deep Learning

R package h2o

- Generalized Additive Model (GAM) -> R Package mgcv

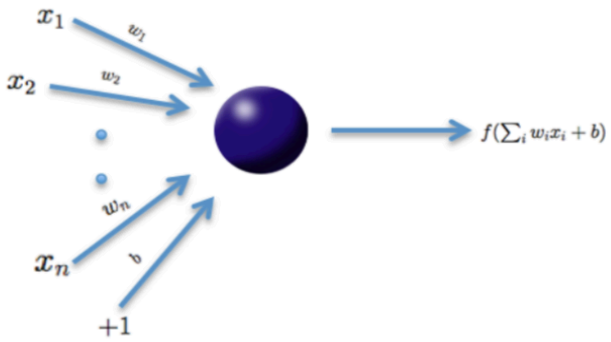
Random Forest

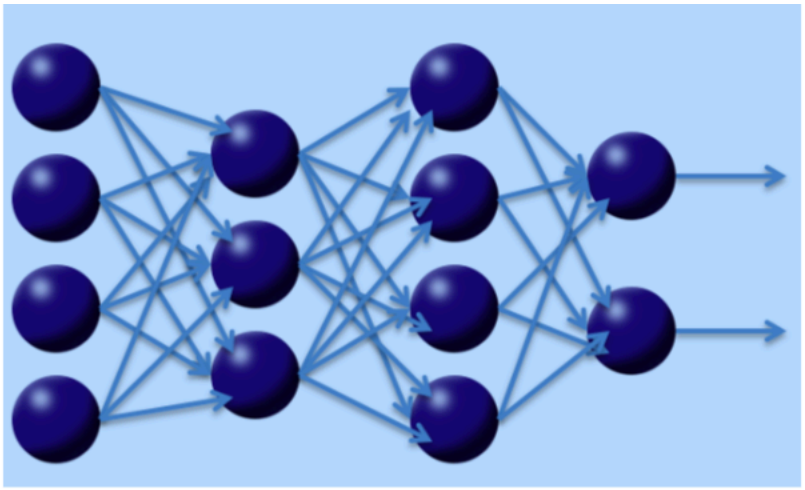
- Fits multiple prediction trees by randomly sampling and then resampling the data and then takes the mean.
- Also randomly samples prediction variables and accounts for interaction terms.

Gradient Boosting

- Combines multiple weak models to form one prediction model
- This final model performs better than the individual models.

Deep Learning





- Fits a neural network to the data.
- ‘H2O follows the model of multi-layer, feed-forward neural networks for predictive modeling.’

Selection of Algorithm Parameters

1. Specify a range of possible values
2. Run 10 fold cross validation over all possible combinations
3. Select model with best held out R^2

GAM Ensemble model

BCi =β_0 + $s(svr\ predictions$)*EC*Ref+ $s(rf\ prediction)$*EC*Ref + $s($ $gb\ prediction)$*EC*Ref + $s(dl\ prediction)$*EC*Ref + ei

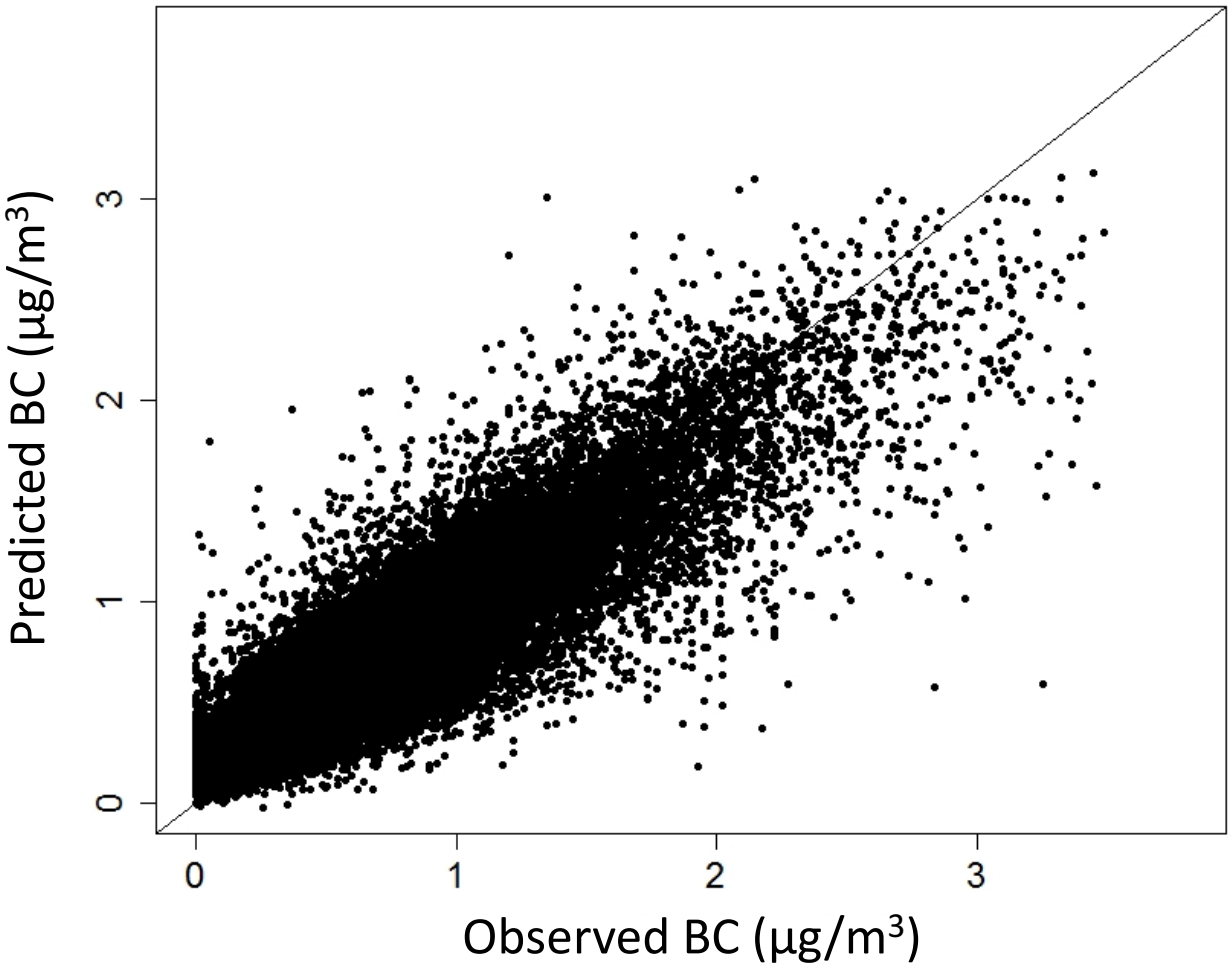
where

- *BCi* is the BC concentration on the *ith* day
- Ref is an indicator variable if the method of measurement was reflectance
- EC is an indicator variable for Elemental Carbon observations

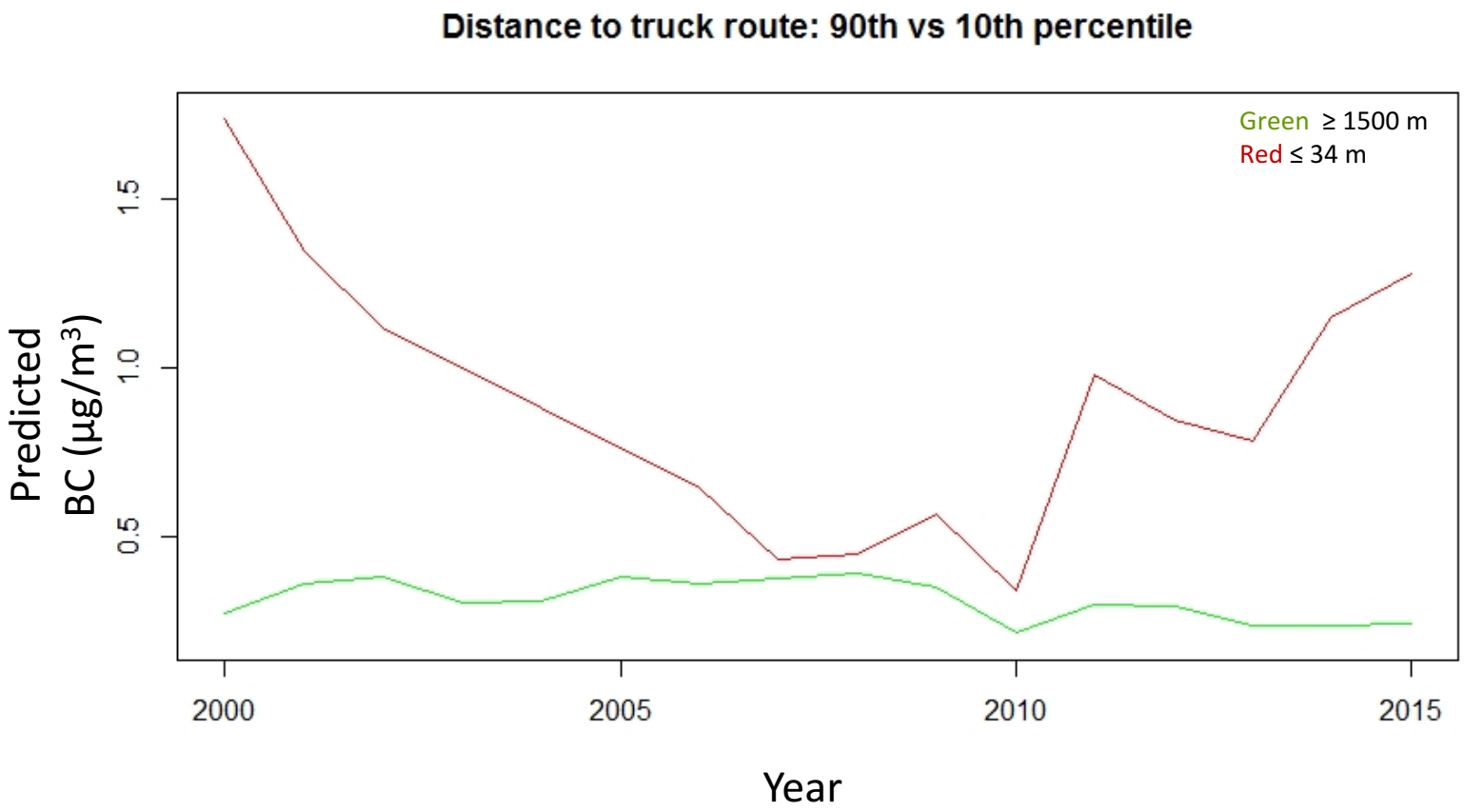
New Model Performance

Calculated in held-out test data in 10-fold Cross Validation		
	Mean R^2	Slope
Nu-SVR	0.73	1.10
Random Forest	0.81	1.09
Gradient Boosting	0.83	1.09
Deep Learning	0.77	1.01
Ensemble model	0.81	1.01

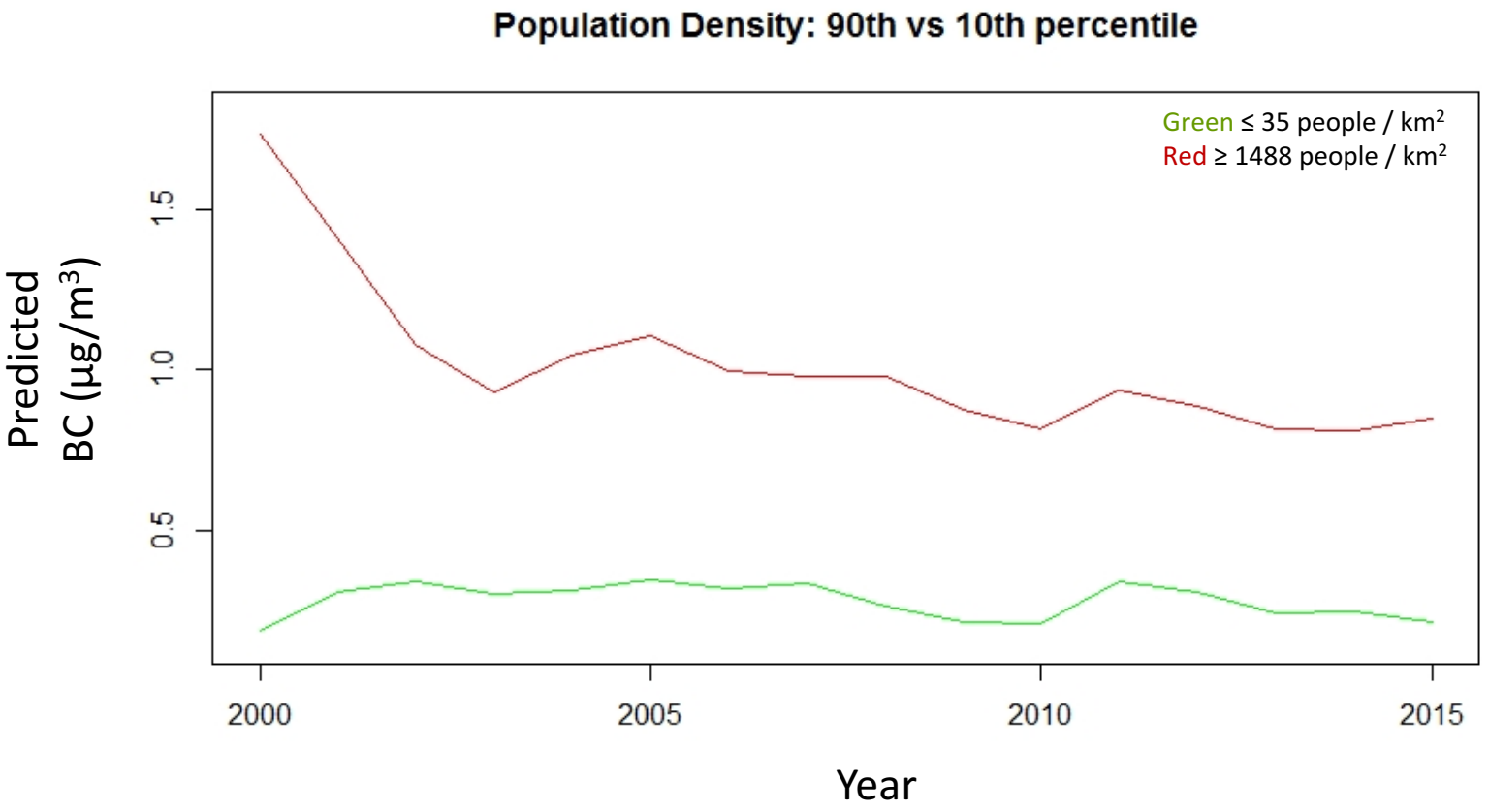
Predicted vs observed in held out data



Spatial predictors over time



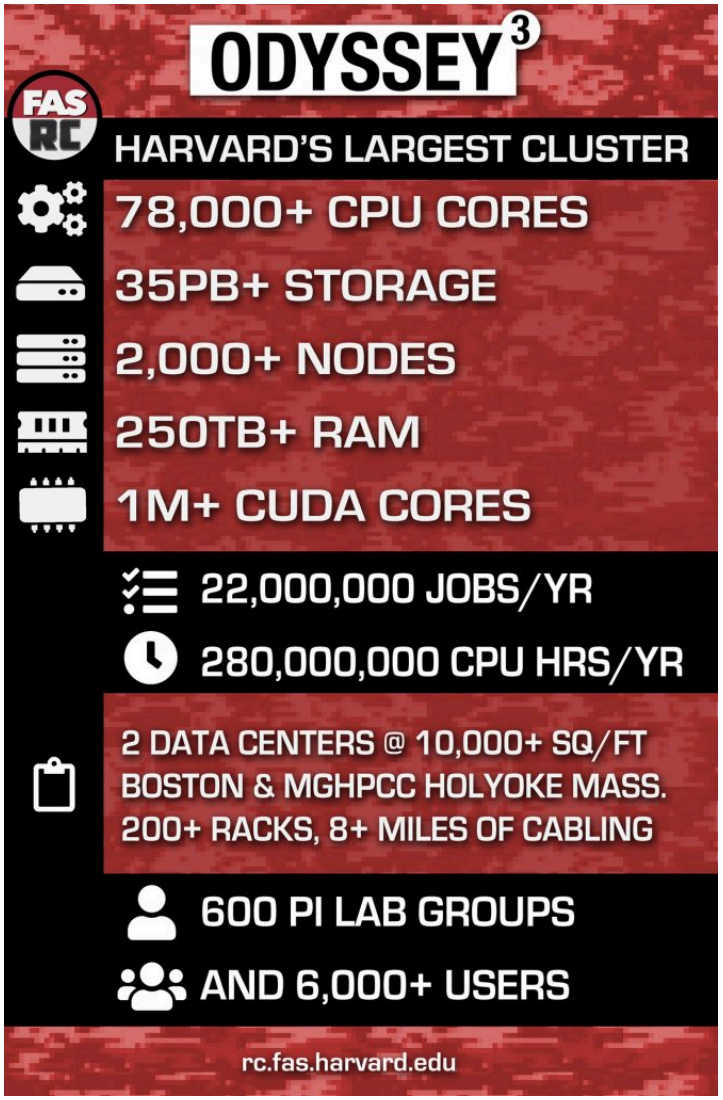
Spatial predictors over time



Predictions MA Parcel Data

2 million addresses!!

Cluster computing



1. Split data into 2000 pieces
2. Run on 2000 nodes

Thank you!