Application of Machine Learning Algorithms for Estimating Daily $PM_{2.5}$ Concentrations

By

Runing Huo

B.S., Beijing Forestry University, 2019

M.S., Michigan Technological University, 2021

Thesis Committee Chair: Howard Chang, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics

2023

Abstract

Application of Machine Learning Algorithms for Estimating Daily $PM_{2.5}$ Concentrations

By Runing Huo

Background: The detrimental impact of $PM_{2.5}$ air pollution is widespread, as it has been linked to premature mortality and a diverse range of health concerns such as cardiovascular and respiratory illnesses. Machine learning approaches offer several advantages for predicting $PM_{2.5}$ levels at locations without monitoring data. These include the ability to handle complex and large datasets, detect nonlinear associations, and provide accurate and adaptable solutions.

Objectives: Compare the prediction ability of four machine learning algorithms with three types of cross-validation experiments using data from 2018 in California.

Methods: Four machine learning algorithms were applied in this analysis: random forest, Bayesian additive regression trees (BART), gradient boosting and soft Bayesian additive regression trees (SoftBART). We performed 3 types of 10-fold cross-validations (ordinary, spatial, and temporal) using, R-squared, mean absolute error (MAE), and root-mean square error (RMSE). We also obtained average predictions of $PM_{2.5}$ concentrations at 1km spatial resolution for January, April, July, Octobe in 2018.

Results: In the cross-validation analysis, we found the random forest performed the best with highest R-squared and smallest RMSE and MAE values. Random forest model also the least computationally intensive approach. Gradients boosting and BART model with larger number of trees are the second-best model. When using small number of trees, SoftBART model behaved similarly with the BART model.

Conclusions: In this study, we demonstrated the superior predictive performance of random forest, which is a commonly used method for predicting daily $PM_{2.5}$ concentrations.

Application of Machine Learning Algorithms for Estimating Daily $PM_{2.5}$ Concentrations

By

Runing Huo

B.S., Beijing Forestry University, 2019

M.S., Michigan Technological University, 2021

Thesis Committee Chair: Howard Chang, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics
2023

Acknowledgement

I would like to express my gratitude to the Rollins School of Public Health at Emory University for their unwavering support throughout my academic journey. My two-year tenure at RSPH equipped me with a strong foundation in Biostatistics and helped me discover my passion for bioinformatics, which has now become the focal point of my career aspirations.

I am especially grateful to my thesis advisor, Howard Chang, for his constant encouragement and unwavering support. He not only provided me with an opportunity to delve into machine learning research but also shared his invaluable expertise and provided me with sincere and constructive feedback throughout my thesis work. His guidance was instrumental in shaping the outcome of my research.

I would also like to extend my appreciation to Danlu Zhang, a team member who generously shared the entire dataset with me, which served as the primary source material for my study. Her willingness to assist me and her friendly demeanor made my research experience a positive one.

Last but not least, I am indebted to my family, who supported my decision to study abroad and offered me comfort during times of distress. Despite being miles apart, their unwavering love made me feel connected and never alone.

Table of Contents

1. Introduction

$PM_{2.5}$ refers to fine inhalable particles less than or equal to 2.5 micrometers in aerodynamic diameter, which can get deep into the respiratory system through breathing. $PM_{2.5}$ air pollution is a global problem because it is known to cause premature deaths and a variety of health issues, including the development and exacerbation of cardiovascular and respiratory diseases [1]. Continued monitoring and understanding of $PM_{2.5}$ health effects are crucial for protecting public health.

To measure $PM_{2.5}$ in the ambient environment, various air monitoring techniques are used, including techniques can provide (1) real-time data on the concentration (e.g., tapered element oscillating microbalance (TEOM), beta attenuation, and laser particle counters), and (2) integrated mass samplers that utilize filters. Over the past decades, these $PM_{2.5}$ measurements have allowed us to conduct studies to understand impact of ambient $PM_{2.5}$ on human health and the environment. Particularly, the establishment of air quality monitoring networks and the development of air quality index (AQI) enable us to regulate and communicate levels of $PM_{2.5}$ and other air pollutants to the public [2].

Monitoring measurements of $PM_{2.5}$ are typically only available sparsely in space. This motivated the development of many statistical and machine learning approaches to predict the distributions of $PM_{2.5}$ particles with various predictors, including meteorological factors, land cover characteristics and satellite imagery. Machine learning models, such as neural network, random forest and gradient boosting are among the most commonly employed approaches [3-5]. Some advantages of machine learning models include the ability to handle complex and large predictors and identify non-linear relationships, in a flexible and accuracy manner.

In this analysis, we compared several different machine learning models to predict of daily average concentration of $PM_{2.5}$ in California during the year 2018, a year with significant wildfire activities. Specifically, the four models are random forest, gradient boosting, Bayesian additive regression trees (BART) and soft Bayesian additive regression trees model.

Random forest, gradient boosting and BART models have been used in an air pollution previously [6-9]. They all have been demonstrated to provide accurate prediction. But SoftBART is a new model that haven't been applied on air pollution data analysis yet.

2. Materials and Methods

2.1 Motivating Datasets

We obtained daily average $PM_{2.5}$ concentration measurements for California and Oregon in 2018 from the US Environmental Protection Agency's Air Quality System and the PurpleAir monitoring network. A series of quality control and bias-correction steps were applied to measurements from PurpleAir [10]. The year 2018 was chosen due to it being the most severe wild fire season in this region.

Satellite-derived aerosol optical depth (AOD) was obtained using the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm at a 1 km spatial resolution. Missing AOD values were gap-filled based on modelled AOD data and meteorology following a previously developed approach [11] . We also obtained simulations from the Community Multiscale Air Quality (CMAQ) models at 12km resolution. CMAQ is a numerical transport model that incorporates information from emission, atmospheric chemistry and transport [12-13] .

Several meteorological variables were obtained the North American Land Data Assimilation System phase 2, include including air temperature, specific humidity, surface pressure, surface downward longwave and shortwave radiation, U and V wind component, total precipitation, and potential evaporation, with a spatial resolution of 0.125° × 0.125°. Land cover variables was obtained from the National Land Cover Database at 30-meter resolution and include the percentage of each land cover type (water, developed area, barren, forest, shrubland, herbaceous area, cultivated area, and wetland). Additional land use variables included annual population counts, total length of different road by type (highways, primary road, secondary road, tertiary road, and local road) and elevations.

To create the analytic dataset, each monitor locations of $PM_{2.5}$ levels were assigned to an AOD 1km grid cell. Inverse-distance weighting was used to assign predictor variables at coarse resolutions (e.g., CMAQ simulation and meteorology) to the AOD grid cell. The sinusoidal projection was used calculate Euclidean distances.

2.2 Statistical Analysis

2.2.1 Machine Learning Models

We first examined prediction performance of three traditional machine learning models, which are Random Forest (RF), Gradient Boosting (GB) and Bayesian Additive Regression Tree (BART) models.

Random forest is a popular machine learning algorithm based on constructing a large number of decision trees, and predictions are these trees are averaged to give a final output. RF is a powerful and versatile machine learning algorithm that can be used for various tasks, including both regression and classification. It is widely used in various applications, such as image classification, natural language processing, and recommendation systems [14-15]. In RF, a subset of the training data which is randomly sampled with replacement is used to train a decision tree, and process the repeated to generate multiple trees.

One of the key advantages of RF is that it can handle a large number of predictors and identify their relative importance p. It also reduces overfitting, which is a common problem in decision trees, by combining the predictions of multiple trees.

We used the R package “randomForest” to perform the training of the model. The number of trees is 100 and the other settings of the parameters are default, the exact parameters are listed in the Table 3.1.

Gradient Boosting is another algorithm of ensemble methods for regression and classification problems. There are 4 common variations of GB, including Gradient Boosting Machine (GBM), Extreme Gradient Boosting Machine (XGBM), LightGBM and CatBoost. These methods are widely used on the data analysis of healthcare, finance [16-17]. Typically, GB builds an ensemble of weak decision trees and the final prediction corresponds to the sum of prediction from each tree. The idea behind the GB is to update each base-learner (i.e., decision tree) by using the negative gradient of the previous model's loss function, which corresponds to minimize the loss function by the gradient descent method, which we describe below.

First, we initialize the model and its associated loss function $L(y_i, \rho)$ with a constant value f_1,

$$f_1 = argmax_\rho \sum_{i=1}^{N} L(y_i, \rho)$$

In our analysis, we assume outcome is Gaussian with mean and the loss function corresponds to squared error loss. Then for each m =1 to 1000 trees, we calculate the negative gradient, where $f_m(x_i)$ is the m^th^ tree:

$$\tilde{y}_i = -\frac{\partial L\big(y_i, f_m(x_i)\big)}{\partial f_m(x_i)}.$$

Using the base learner $h_m(x)$, line search is used to calculate step size ρ_m and minimize loss function. Then the updated model is:

$$f_m(x_i) = f_{m-1}(x) + \rho_m h_m(x_i, w_m).$$

If shrinkage is used, the update is:

$$f_m(x) = f_{m-1}(x) + \nu \rho_m h_m(x, w_m).$$

We used GBM with R package "gbm" to implement the GB, assuming the Gaussian distribution outcome model. The number of trees was set as 1000, and the maximum depth of each tree was 4 and shrinkage parameter was 0.01 which is also called learning rate.

Bayesian Additive Regression Trees (BART) model is also based on decision trees. It is similar to GB in that the final prediction represents the sum of predications from all trees. Bayesian backing fitting via Markov Chain Monte Carlo (MCMC) is employed for Bayesian inference to obtain the posterior distribution of each prediction.

The BART model is given below:

$$Y_i = \sum_{k=1}^{K} T_k(M_k; x_i) + \epsilon$$

$$\epsilon \sim N_n(0, \sigma^2 I_n)$$

where $T_k(M_k; x_i)$ represents the tree structure, and M_k represents the parameters of terminal nodes dependent on predictor vector x_i.We used the R package "BART" to perform the model training and the default setting of the function "wbart". The number of decision trees are 10 or 200, the number of MCMC iterations for burn in is 100, the total number of iterations are 1100.

Soft Bayesian Additive Regression Trees (SoftBART) is a variate machine learning algorithm of BART model, it is first generated by Linero and Yang in 2018 [18]. The main difference between SoftBART model and BART model is that SoftBART uses a Bayesian hierarchical model to calibrate the predicted values of the model to better match the actual values in the data. This is done by modeling the residuals of the original BART model as a Gaussian process and adjusting the predicted values accordingly.

SoftBART model can take less time to make more accurate predictions. It also has some similarities with BART model. Both of them are tree-based algorithms and use Bayesian framework. They both use MCMC to estimate the posterior distribution of the model parameters. They take long time for calculation when applied on large dataset.

2.2.3 Prediction Performance Comparison

Cross-validation (CV) is a method to evaluate out-of-sample prediction performance of algorithms by separating the dataset into two mutually exclusion sets: the training set and the testing set. The training set is used to perform learning and testing set is used to validate a model's predictions. In a k-fold CV study, the dataset is divided into *k* segments evenly. The training and testing process is conducted in *k* rounds. In each round, we use each segment as the testing set, and the other *k-1* folds as training set. In the end, an out-of-sample prediction is obtained for each data point in the entire dataset. In all CV analysis, RF, GB, BART and SoftBART were applied to the same training and testing datasets in each fold.

We used three different types of 10-fold cross validation: traditional 10-fold cross validation, spatial 10-fold cross validation, and temporal 10-fold cross validation. The difference among them is how the dataset is partitioned into training sets and testing sets.

The traditional 10-fold cross validation separates the entire dataset randomly into 10 folds. To evaluate prediction performance, we calculated coefficient of determination(R^2), root mean squared error (RMSE) and mean absolute error (MAE). A larger R^2, smaller RMSE and smaller MAE are preferred. Specifically, R^{2}takes value from 0 to 1 and is a measure that describes the proportion of the variance in the dependent variable explained by the model. of the equation for R^2 is:

$$R^2 = 1 - \frac{Sum\ Squared\ Regression\ (SSR)}{Total\ Sum\ of\ Squares\ (SST)} = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where, y_i is the observational data, $\widehat{y_i}$ is the predicted values, $\bar{y}$ is the mean of observational data.

RMSE represents the average distance from predicted values to the out-of-sample data. RMSE is given by (N is the number of observations):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{N}}$$

Mean absolute error (MAE) is another useful metric. The difference between RMSE and MAE is that MAE measure the absolute distance from predicted values to the observations. It is less sensitive to the outliers than RMSE, because it doesn’t square the difference between predicted and observational values. MAE is given by

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \widehat{y_i}|}{N}$$

In addition to the traditional CV analysis, we also considered spatial 10-fold CV. where data from 10% of all monitors were designated as the testing set, and the other ninety percent of monitors were used as the training set. The spatial CV is designed to evaluate prediction performance associated with spatial interpolation. Similarly, in a temporal 10-fold CV analysis, data from 10% of unique dates were designated as the testing set in order to evaluate model’s performance to predict days without any observations.

The parameters used in the three types of cross-validations for different methods are given below:

Table 3.1 Parameters Used in the Cross-Validation Analysis

Random Forest	ntree=100, mtry=sqrt(p), nodesize=5
BART	ntree=10 and 200, nskip=100
Gradient Boosting	ntree=1000, shrinkage=0.01, depth=4
Soft BART	ntree=5, num_burn=100

3. Results

3.1 Results of Cross-Validation Experiments

In the dataset used to perform cross-validation, there are total 56, 274 records of PM$_{2.5}$ measurements, consists datapoints from 365 days and 206 different air quality monitoring locations (158 from California and 48 from Oregon). There are 38 predictors, include meteorological parameters, land-cover variables, and geographical information of the air monitoring devices.

The histogram and boxplot showed below indicate that our outcome variable PM$_{2.5}$ has a highly right-skewed distribution, and there are many outliers. Most of the values are from 1 to 50, but outliers are distributed between 50 to 420. The correlation plot shows complicated correlations among predictors, most of the variables are weakly correlated, but there are several highly correlated variables (e.g., elevation and surface pressure).

Figure 3.1 Histogram, Boxplot of PM$_{2.5}$ in 2018

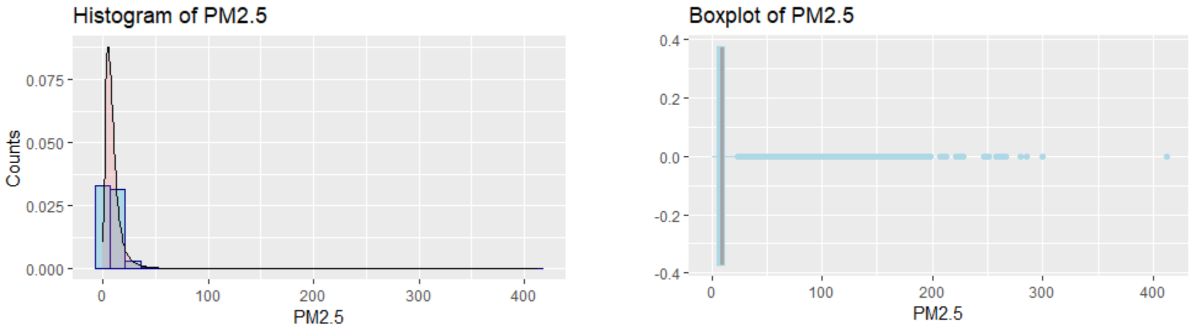
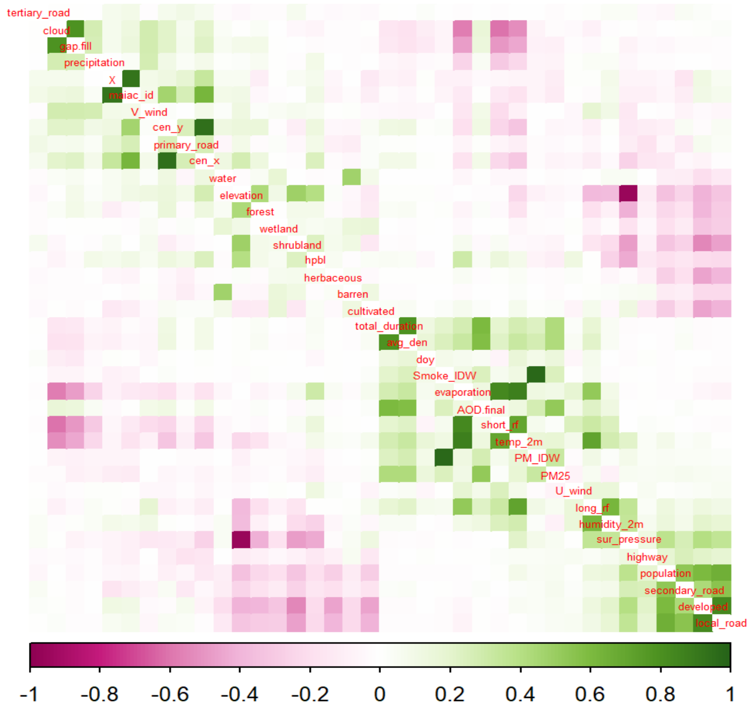


Figure 3.2 Correlation Plot of Predictors



3.1.1 Traditional 10-fold Cross-Validation

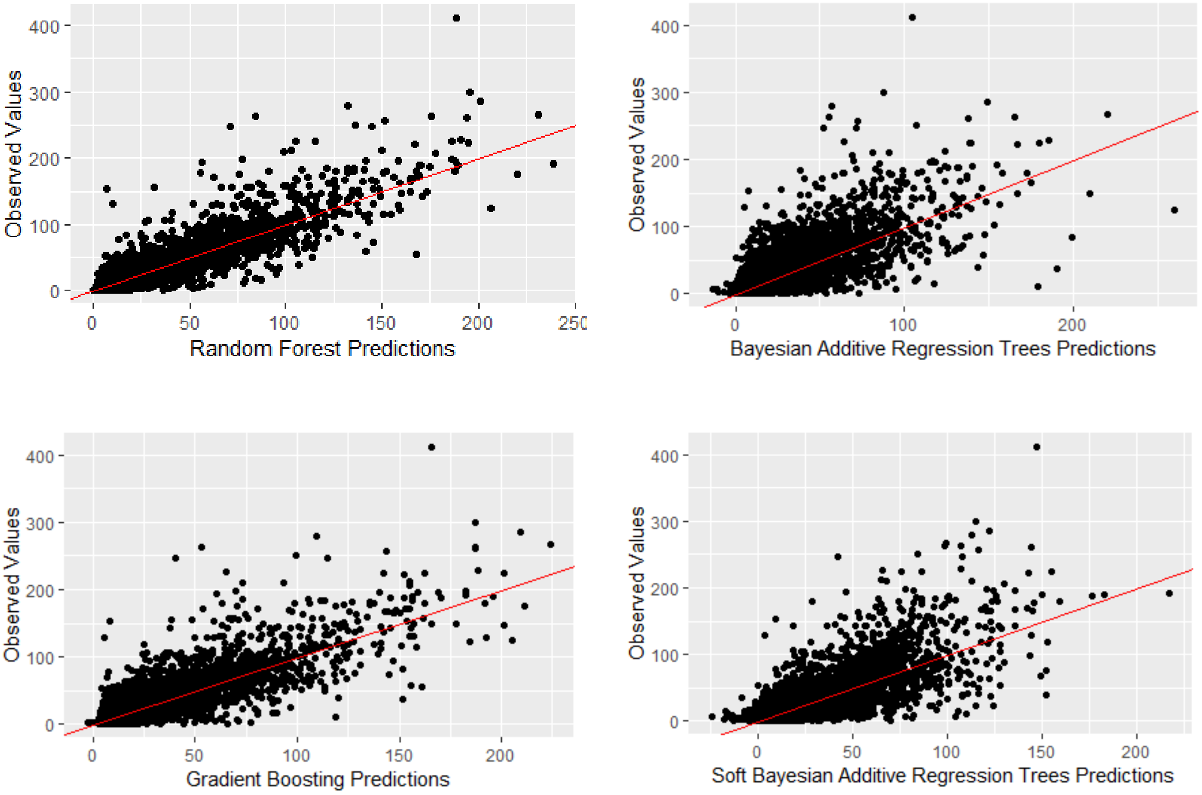
. The results in Table 3.1.1 showed that the RF model behaved the best and have the highest R-squared, followed by the BART model with 200 trees. Soft BART model behaved better then BART model when both have smaller number of trees.

Table 3.1.1 Prediction performance with traditional 10-fold Cross-Validation

Type of Cross-Validation	ML Methods	R^2	RMSE	MAE
	Random Forest (100)	0.833	5.661	2.705
Traditional 10-fold Cross-Validation	BART (10)	0.572	8.989	4.772
	BART (200)	0.742	6.99	3.590
	Gradient Boosting (1000)	0.719	7.298	3.940
	Soft BART (5)	0.623	8.447	4.439

From the predicted versus observed plots, we can see that the datapoints from RF are more compact andshow stronger linearity. The plot of Gradient Boosting model looks similar with Random Forest one, but with more dispersed datapoints. The plots of BART model (with 10 trees) and SoftBART model have weaker linearity and the datapoints distributed more spread out from the diagonal. The plots showed the same conclusions with the R-squared and RMSE.

Figure 3.1.1 Predicted and Observational Data in Traditional 10-fold Cross-Validation Experiments



3.1.2 Spatial 10-fold Cross-Validation

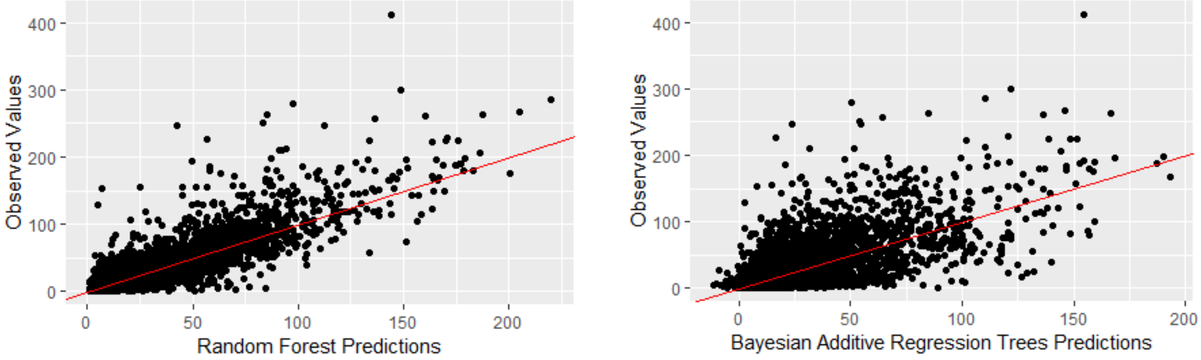
The spatial cross-validation results given in Table 3.1.2 have similar results in the traditional cross-validation experiment. The Random Forest model is still the best one, but compared with the results from traditional cross-validation, the R-squared is smaller and RMSE, MAE are larger. The Gradient Boosting is the second best one, and the Soft BART model is better than the BART model even though it only has 5 trees.

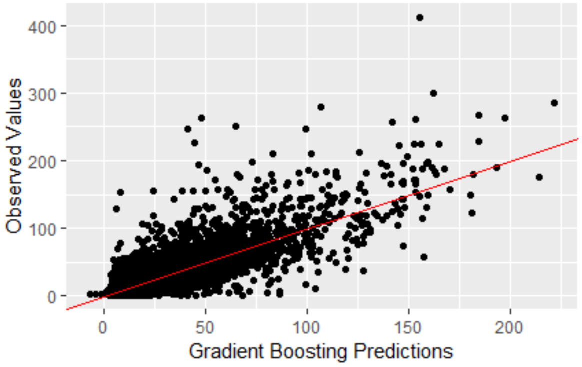
Table 3.1.2 Prediction performance in Spatial 10-fold Cross-Validation Experiments

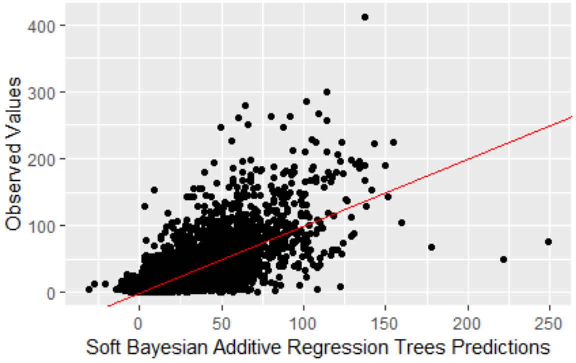
Type of Cross-Validation	ML Methods	R^2	RMSE	MAE
Spatial 10-fold Cross-Validation	Random Forest (100)	0.785	6.454	3.263
	BART (10)	0.497	9.734	5.228
	BART (200)	0.550	9.565	5.734
	Gradient Boosting (1000)	0.695	7.623	4.076
	Soft BART (5)	0.575	8.947	4.642

From the predicted versus observed plots generated with the spatial 10-fold cross-validation, we find that the Random Forest still looks the best compared with the other models. The datapoints from SoftBART model stay closer than BART model.

Figure 3.1.2 Predicted and Observational Data in Spatial10-fold Cross-Validation Experiments







3.1.3 Temporal 10-fold Cross-Validation

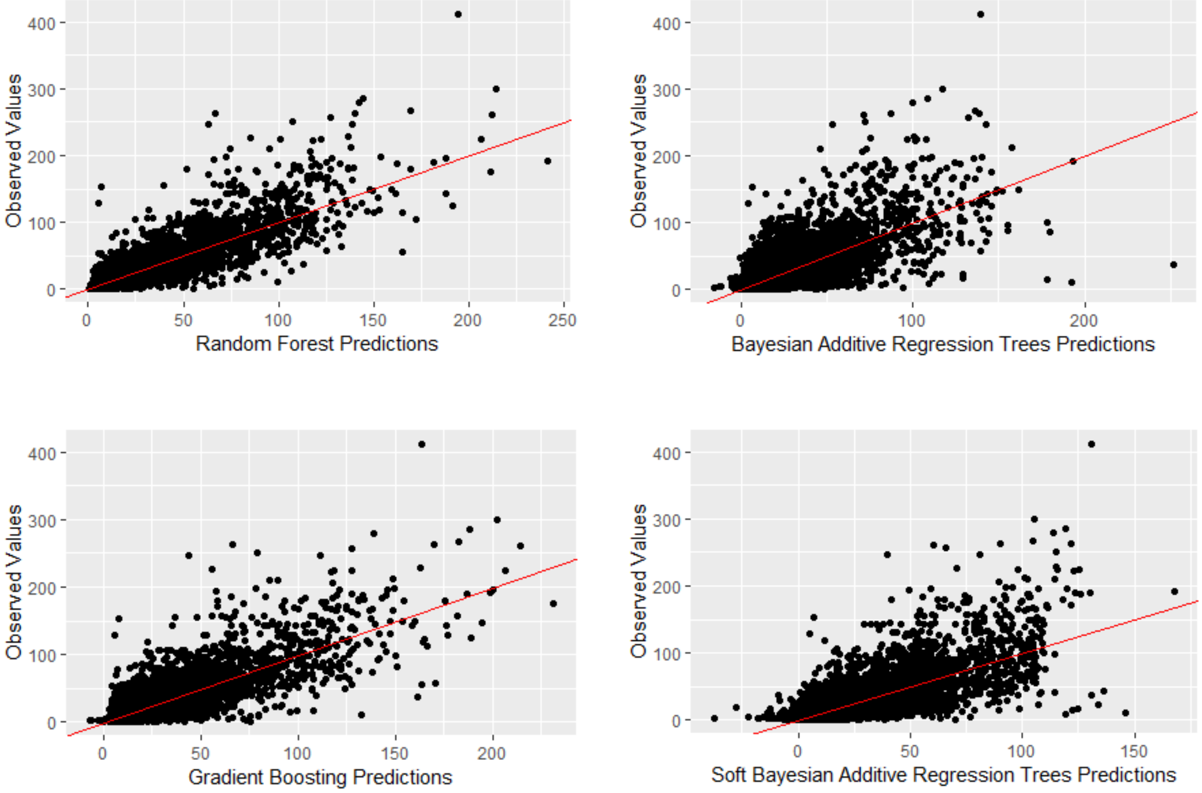
Results from temporal cross-validation are similar to the traditional cross-validation, the Random Forest model being the best one, followed by Gradient Boosting model.

Table 3.1.3 Prediction performance in Temporal 10-fold Cross-Validation Experiments

Type of Cross-Validation	ML Methods	R^2	RMSE	MAE
Temporal 10-fold Cross-Validation	Random Forest (100)	0.776	6.548	3.154
	BART (10)	0.539	9.319	4.888
	BART (200)	0.681	7.792	3.883
	Gradient Boosting (1000)	0.691	7.639	4.045
	Soft BART (5)	0.576	8.940	4.698

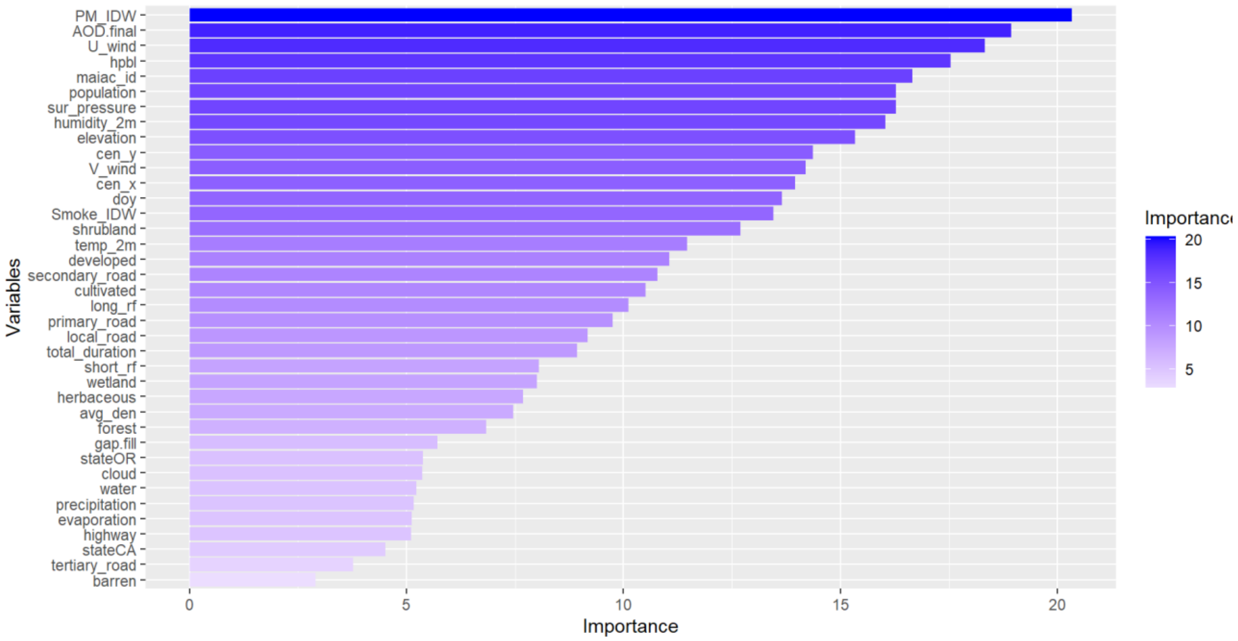
From the predicted versus observed plots generated by the temporal 10-fold cross validation, we observe that Soft BART model behaved better than the BART model when generated small number of trees, though it still has worse prediction than Random Forest did.

Figure 3.1.3 Predicted and Observational Data in Temporal 10-fold Cross-Validation Experiments



To quantify the contribution of individual predictors, we also trained a RF model with the full dataset, and examine variable importance (Figure 3.1.4). The top 9 predictors are PM simulated from CMAQ, aerosol optical depth, U-direction wind speed, boundary layer height, grid cell, population density, surface pressure, relative humidity, and elevation.

Figure 3.1.4 Importance of Variables in Random Forest Model

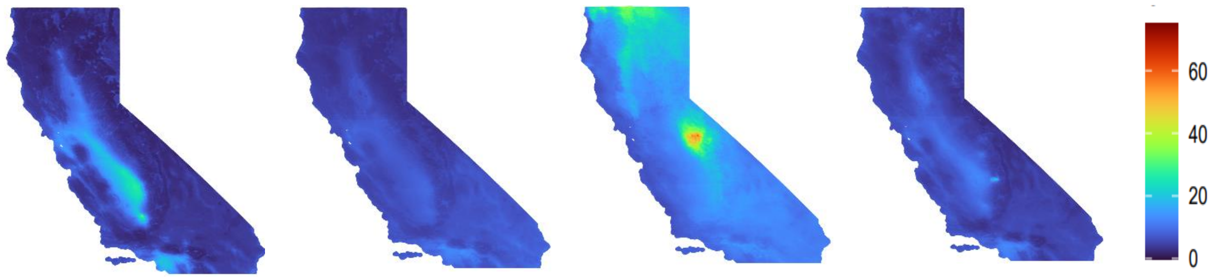


3.2 Results of Predictions

Figures 3.2.1 to 3.2.3 show the monthly average PM2.5 concentrations predicted from RF, GB and SoftBART for California in January, April, July and October. We see that the distributions of $PM_{2.5}$ levels were different between the four months, with the highest concentration observed in the middle part of California and in July.

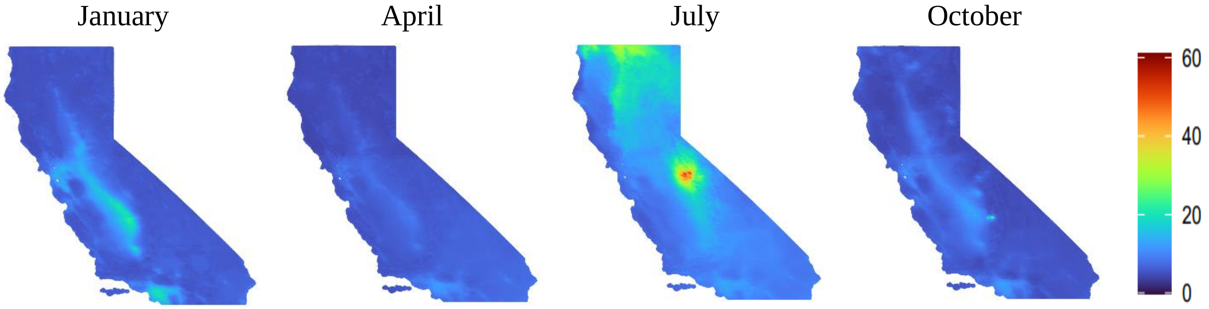
Figure 3.2.1 Predicted Concentrations of $PM_{2.5}$ in California from Random Forest Model

January April July October



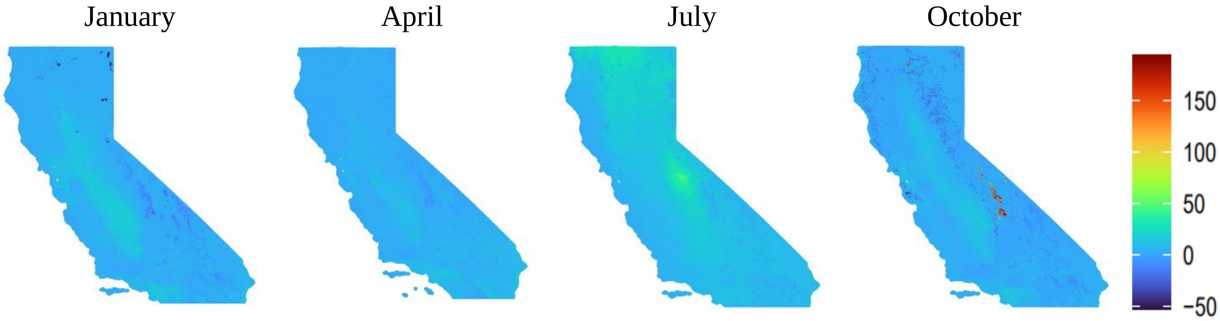
The plots we got from GB model are similar with the RF model, but the range of the predictions are a little bit high. July is still the month with highest concentrations of $PM_{2.5.}$

Figure 3.2.2 Predicted Concentrations of $PM_{2.5}$ in California from Gradient Boosting Model


January April July October

In the predictions of SoftBART, we can see the range of the predictions became much larger, and there are some extreme values appeared in October. But from the whole state view, July is still the month has higher concentrations of $PM_{2.5.}$

Figure 3.2.3 Predicted Concentrations of $PM_{2.5}$ in California from SoftBART Model


January April July October

4. Discussion

In the analysis, we evaluated the prediction performance of four machine learning models: Random Forest, BART, Gradient Boosting and SoftBART. From three types of cross-validations (traditional, spatial and temporal), RF consistently performed the best, as determined by its highest R-squared and lowest RMSE and MAE. RF has the additional advantage of having the less computation time. In BART model, we observed that increasing the number of trees can improve prediction, despite additional memory and run-time requirements. SoftBART model behaved better than BART model when both used small number of trees. Gradient Boosting model is often the second-best model.

Random Forest model is an efficient and accurate model to use It can prevent overfitting by creating multiple decision trees, each trained on a different subset of the data and a random subset of the predictors. It can also provide the importance of each feature in the model, which helps better interpretation of the results. This information may be particularly useful as a tool to select variables to as inputs to methods that cannot handle large predictor set, such as Treed Gaussian Process [19].

There are several follow-up analyses that can be built upon the current work. Particularly, the choice of parameters setting in BART and SoftBART should be examined more comprehensively by changing the number of trees, depth of trees, and prior's distributions' parameter or MCMC iterations. Secondly, we can continue to work on prediction plots of BART modes and see the differences.

References:

[1] Callen MS, Lopez JM, Mastral AM (2012) Apportionment of the airborne PM_{10} in Spain. Episodes of potential negative impact for human health. J Environ Monit : JEM 14:1211–1220

[2] Li QF, Wang-Li L, Liu Z, Heber AJ. Field evaluation of particulate matter measurements using tapered element oscillating microbalance in a layer house. J Air Waste Manag Assoc. 2012 Mar;62(3):322-35. doi: 10.1080/10473289.2011.650316. PMID: 22482290.

[3] Meng, X.; Hand, J.L.; Schichtel, B.A.; Liu, Y. Space-timeœ trends of $PM_{2.5}$ constituents in the conterminous United States estimated by a machine learning approach, 2005–2015. Environ. Int. 2018, 121, 1137–1147.

[4] Geng, G.; Meng, X.; He, K.; Liu, Y. Random forest models for $PM_{2.5}$ speciation concentrations using MISR fractional AODs. Environ. Res. Lett. 2020, 15, 034056.

[5] Bekkar, A., Hssina, B., Douzi, S. *et al.* Air-pollution prediction in smart city, deep learning approach. *J Big Data* 8, 161 (2021). https://doi.org/10.1186/s40537-021-00548-1

[6] Yu R, Yang Y, Yang L, Han G, Move OA. RAQ-A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors (Basel). 2016 Jan 9;16(1):86. doi: 10.3390/s16010086. PMID: 26761008; PMCID: PMC4732119.

[7] Li Z, Chen Y, Tao Y, Zhao X, Wang D, Wei T, Hou Y, Xu X. Mapping the personal $PM_{2.5}$ exposure of China's population using random forest. Sci Total Environ. 2023 May 1;871:162090. doi: 10.1016/j.scitotenv.2023.162090. Epub 2023 Feb 9. PMID: 36764537.

[8] Zhang T, Geng G, Liu Y, Chang HH. Application of Bayesian Additive Regression Trees for Estimating Daily Concentrations of $PM_{2.5}$ Components. Atmosphere (Basel). 2020 Nov;11(11):1233. doi: 10.3390/atmos11111233. Epub 2020 Nov 16. PMID: 34322279; PMCID: PMC8315111.

[9] Zalakeviciute R, Rybarczyk Y, Alexandrino K, Bonilla-Bedoya S, Mejia D, Bastidas M, Diaz V. Gradient Boosting Machine to Assess the Public Protest Impact on Urban Air Quality. *Applied Sciences*. 2021; 11(24):12083. https://doi.org/10.3390/app112412083

[10] Bi J, Wildani A, Chang HH, Liu Y. Incorporating low-cost sensor measurements into high-resolution PM2. 5 modeling at a large spatial scale. Environmental Science & Technology. 2020 Jan 13;54(4):2152-62.

[11] J. Bi, J. H. Belle, Y. Wang, A. I. Lyapustin, A. Wildani, Y. Liu, Impacts of snow and cloud covers on satellite-derived $PM_{2.5}$ levels. *Remote sensing of environment* 221, 665-674 (2019).

[12] K. R. Baker, M. C. Woody, G. S. Tonnesen, W. Hutzell, H. O. T. Pye, M. R. Beaver, G. Pouliot, T. Pierce, Contribution of regional-scale fire events to ozone and $PM_{2.5}$ air quality estimated by photochemical modeling approaches. *Atmospheric Environment* 140, 539-554 (2016).

[13] P. D. Koman, M. Billmire, K. R. Baker, R. de Majo, F. J. Anderson, S. Hoshiko, B. J. Thelen, N. H. F. French, Mapping Modeled Exposure of Wildland Fire Smoke for Human Health Studies in California. *Atmosphere (Basel)* 10, (2019).

[14] Balyan R, Crossley SA, Brown W 3rd, Karter AJ, McNamara DS, Liu JY, Lyles CR, Schillinger D. Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study. PLoS One. 2019 Feb 22;14(2):e0212488. doi: 10.1371/journal.pone.0212488. PMID: 30794616; PMCID: PMC6386302.

[15] A. Ajesh, J. Nair and P. S. Jijin, "A random forest approach for rating-based recommender system," *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, India, 2016, pp. 1293-1297, doi: 10.1109/ICACCI.2016.7732225.

[16] Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O; written on behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med. 2019 Apr;7(7):152. doi: 10.21037/atm.2019.03.29. PMID: 31157273; PMCID: PMC6511546.

[17] Davis, J., Devos, L., Reyners, S., & Schoutens, W. (2021). Gradient boosting for quantitative finance. In *Journal Of Computational Finance* (Vol. 24, Issue 4). Risk Waters Group. https://doi.org/10.21314/JCF.2020.403

[18] Antonio R. Linero & Yun Yang, 2018. "Bayesian regression tree ensembles that adapt to smoothness and sparsity," *Journal of the Royal Statistical Society Series B, Royal Statistical Society,* vol. 80(5), pages 1087-1110, November.Handle: *RePEc:bla:jorssb:v:80:y:2018:i:5:p:1087-1110* DOI: 10.1111/rssb.12293

[19] Gramacy, R. B., & Lee, H. K. H. (2008). Bayesian Treed Gaussian Process Models with an Application to Computer Modeling. *Journal of the American Statistical Association*, *103*(483), 1119–1130. http://www.jstor.org/stable/27640148