Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____	_____
Tianyu Zhang	Date

An Application of Bayesian Additive Regression Trees (BART) to Estimate Daily Concentrations of PM2.5 Components in California

By

Tianyu Zhang
Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Howard H. Chang, PhD
Thesis Advisor

_____

Yang Liu, PhD
Reader

An Application of Bayesian Additive Regression Trees (BART) to Estimate Daily Concentrations of PM2.5 Components in California

By

Tianyu Zhang

B.S., Shanghai Jiaotong University, 2018

Thesis Advisor: Howard H. 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 Science in Public Health
in Biostatistics
2020

Abstract

An Application of Bayesian Additive Regression Trees (BART) to Estimate Daily Concentrations of PM2.5 Components in California

By Tianyu Zhang

Background: Fine particulate matter ($PM_{2.5}$), defined as particles that have an aerodynamic diameter of less than 2.5 micrometers, represents a complex mixture of solids and liquids that are small enough to pass through the upper respiratory system and penetrate deep into the lungs. Studies have found associations between adverse health outcomes and specific $PM_{2.5}$ species, such as sulfate, nitrate and carbon-containing species. It's important to accurately measure the concentration of $PM_{2.5}$ and its component to support additional epidemiological studies and perform health impact analyses.

Methods: In this work, we examine the use of Bayesian Additive Regression Tree (BART) for predicting concentrations of 4 major components of $PM_{2.5}$: elemental carbon (EC), organic carbon (OC), nitrate (NO_3), and sulfate (SO_4). BART employs a sum-of-trees model and the prediction is based on the average of a set of decision trees. Meteorological variables, population size, land use variables, numerical model simulations (CMAQ), and satellite-derived fractional aerosol optical depth (AOD) in California during the period 2005 to 2014 were used as predictors for $PM_{2.5}$ species concentrations. We evaluated the importance of $PM_{2.5}$, numerical model simulations and AODs by leaving or keeping them in the model.

Results: After tuning parameters in the model to achieve a prediction coverage probability of about 95%, our model consistently results in a R^2 between 0.64 and 0.83 in 5-fold ordinary and spatial leave-on-monitor-out cross-validation (CV) experiments for four species of interest when $PM_{2.5}$ itself is a predictor. When $PM_{2.5}$ is not a predictor, the models achieved a smaller R^2 from 0.52 to 0.72. In spatial CV experiments, including AOD parameters or CMAQ simulations can improve R^2, especially when total $PM_{2.5}$ mass is not included as a predictor. The relative importance of different AOD parameters varies across $PM_{2.5}$ components. AOD_3 and AOD_2 are most important for NO_3 and OC respectively. For SO_4, many AOD parameters show moderate importance.

Conclusions: Collocated $PM_{2.5}$, fractional AOD and CMAQ simulations are important predictors for daily concentrations of $PM_{2.5}$ component EC, OC, NO_3 and SO_4. The ensemble learning method BART provides good prediction accuracy, as well as uncertainty measures that can be utilized in subsequent analyses.

An Application of Bayesian Additive Regression Trees (BART) to Estimate Daily Concentrations of PM2.5 Components in California

By

Tianyu Zhang

B.S., Shanghai Jiaotong University, 2018

Thesis Advisor: Howard H. 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 Science in Public Health
in Biostatistics
2020

Acknowledgements

I would like to thank Rollins School of Public Health and the Department of Biostatistics and Bioinformatics for providing me with the space, equipment and abundant education and research resources in the past two years. I've gained the knowledge necessary for this research as well as many future applications here.

I especially would like to thank my thesis advisor Howard H. Chang, Ph.D. for guiding me through the process of this study from selecting the topic, conducting analysis to writing of the thesis. He is very nice to students and I was always inspired by him and the weekly group meeting held by him. I've benefited so much from the process of working with him.

I also would like to thank my reader Yang Liu, Ph.D., who provided me with the data and precious advice on the analysis process and the writing of the thesis.

Finally, I would like to thank my family and friends for being my emotional support during the past two years. They make me believe in myself and everything I do.

Table of Contents

1. Introduction **1**

2. Methods **3**

2.1 Data *3*

2.2 Modeling *4*

3. Results **5**

4. Discussion **11**

5. Bibliography **15**

Appendix A. Data and Data Processing **17**

Bibliography *19*

Appendix B. Supplementary Tables and Figures **21**

1. Introduction

Understanding the spatial and temporal distribution of ambient air pollution is an active research area due to its harmful health effects. Exposures to air pollution have been associated with the development and exacerbation of various chronic heart and lung diseases, as well as premature deaths (Fann et al. 2019; Thurston et al. 2016; Janssen et al. 2013). Fine particulate matter ($PM_{2.5}$), defined as particles that have an aerodynamic diameter of less than 2.5 micrometers, is regulated worldwide by government agencies. $PM_{2.5}$ represents a complex mixture of solids and liquids that are small enough to pass through the upper respiratory system and penetrate deep into the lungs. Various studies have found associations between adverse health outcomes and specific $PM_{2.5}$ species, particularly sulfate, nitrate, and carbon-containing species (Schlesinger 2007; Rohr and Wyzga 2012). Hence, it's important to accurately measure the concentration of $PM_{2.5}$ and its component to support additional epidemiological studies and perform health impact analyses (Grahame 2014). However, due to resource constraints, the availability of $PM_{2.5}$ components measurements is more limited than to other air pollutants. Therefore, it is important to develop methods to estimate concentrations of $PM_{2.5}$ components at locations and at time points without monitoring data.

Various models have been proposed to tackle this problem. On one hand, there are traditional geostatistical models, including national spatial models for annual average concentrations of $PM_{2.5}$ species (Bergen et al. 2013), generalized additive models using principal components of predictors (Li et al. 2017), and models based on chemical transport modeling that utilizes information on anthropogenic emissions of primary $PM_{2.5}$

and $PM_{2.5}$ precursors (Kelly, Reff, and Gantt 2017). On the other hand, applications of machine learning methods also have demonstrated excellent prediction accuracy. For example, random forest was used by Meng et al. to estimate the $PM_{2.5}$ specific concentration in the United States (Meng et al. 2018). While machine learning methods have been widely applied to estimate total $PM_{2.5}$ mass (Hu et al. 2017; Lary, Lary, and Sattler 2015; Niu et al. 2017), their applications to $PM_{2.5}$ species have been more limited.

In this work, we examine the use of Bayesian Additive Regression Tree (BART) (Chipman, George, and McCulloch 2012) for predicting concentrations of 4 major components of $PM_{2.5:}$ elemental carbon (EC), organic carbon (OC), nitrate (NO_3), and sulfate (SO_4) BART employs a sum-of-trees model, meaning that the prediction is based on the average of a set of trees where each decision tree contributions a small proportion of the prediction. This form of ensemble learning has been shown to improve prediction accuracy in many applications (Weyuker, Ostrand, and Bell 2010; Linero 2017; Hern et al. 2015). More importantly, BART is a probabilistic model-based learning method that provides straight-forward uncertainty quantification for predictions (e.g. via prediction standard error and prediction intervals).

We applied BART to $PM_{2.5}$ species data from California during the period 2005 to 2014, as used in a previous study by Franklin et al. paper (Franklin, Kalashnikova, and Garay 2017). We also utilized meteorological variables, population size, land use variables, numerical model simulations, and satellite-derived fractional aerosol optical depth (AOD). We demonstrated that multiple BART parameters can be tuned to achieve a

balance between coverage probability of prediction intervals and prediction accuracy. Understanding how different predictors influence and contribute to $PM_{2.5}$ prediction is a well-known challenge for machine learning methods (Huang et al. 2018). Hence, we also examined the variable importance (measured by the times a variable is used in the BART model across trees) to investigate the usefulness of AODs, numerical model simulations and total $PM_{2.5}$ mass in predicting the concentration of the components. Our results confirmed that satellite-derived fractional AOD and numerical model simulations play a heavy role in predicting the components, especially when collocated $PM_{2.5}$ level is unavailable.

2. Methods

2.1 Data

Data in the state of California and the surrounding area from the year 2005 to 2014 were used. The daily concentration of $PM_{2.5}$ species: elemental carbon (EC), organic carbon (OC), nitrate (NO_3), and sulfate (SO_4) are our objectives. The predictors include Community Multiscale Air Quality (CMAQ) simulation, satellite-derived fractional aerosol optical depth (AOD), meteorological and geographical data. We used the same dataset compiled in Geng et al. (2020) and detailed information about data and data processing steps can be found in the Appendix.

2.2 Modeling

The BART methodology is based on a sum-of-trees regression model and regularization priors on the parameters. Let Y_i be the i^{th} observation and $x_i = (x_{i1}, \dots, x_{ip})$ be the corresponding vector of p-many predictors. A BART model with m-many trees can be described as:

$$Y_i = \sum_{j=1}^{m} g(x_i; T_j, M_j) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

where T_j represents the jth tree that consists of a set of internal nodes of decision rules and a set of b-many terminal nodes $M_j = \{\mu_{1j}, \mu_{2j}, \dots, \mu_{bj}\}$. The component ϵ_i represents independent mean-zero Normal error with variance σ^2. The decision rules for each tree are all binary and in the form $x_i \leq c$ or $x_i > c$ for a continuous x_i. Each unique combination of values in x_i is associated with a terminal node according to the sequence of decision rules of that tree. This process of assigning each x_i value at one terminal node and assigning it the value μ_{ij} is represented by function $g(x; T_j, M_j)$. Thus Y represents the sum of m such binary trees. Furthermore, prior distributions over all the parameters, $(T_1, M_1), \dots, (T_m, M_m), \sigma$ is imposed on the model to allow for parameter regularization. We assume the priors of (T_j, M_j) are independent of each other and of that of σ. The priors of M_j given T_j are independent of each other. A commonly used variable importance measure in BART is represented by how often a variable is used in all trees. The BART model can be fitted in the package BART in R.

We used this model to predict $PM_{2.5}$ concentrations of elemental carbon (EC), organic carbon (OC), sulfate (SO_4) and nitrate (NO_3). Tuning parameters that control for the

number of trees and the depth of the trees are adjusted to achieve the correct coverage probability of the 95% posterior prediction intervals for in-sample data. 5-fold cross-validation and spatial cross-validation are conducted. We are interested in the importance of three types of predictors: total $PM_{2.5}$ mass, CMAQ simulated $PM_{2.5}$ components, and fractional AODs. Models with different set of the above three types of predictors were constructed to test the importance of each of them. The analysis of prediction performance with and without $PM_{2.5}$ data is of particular interest because of the potential to leverage the larger $PM_{2.5}$ monitoring network for estimating $PM_{2.5}$ species.

3. Results

Table 1 presents the prediction performance of BART in 5-fold ordinary and spatial leave-one-monitor-out CV experiments. For all $PM_{2.5}$ species, prediction performance is poorer for spatial CV compared to 5-fold CV. Based on R^2 and RMSE, $PM_{2.5}$ total mass is an important variable for predicting $PM_{2.5}$ components. By including $PM_{2.5}$, in both 5-fold and spatial CV experiments, we see the largest improvement in prediction associated with OC and the smallest improvement with SO_4. When $PM_{2.5}$ is included as a predictor, 5-fold CV R^2 decreases in the order of OC, SO_4, EC, and NO_3; this is likely because both OC and SO_4 are major constituents of $PM_{2.5}$ mass. In spatial CV, R^2 decreases in the order of SO_4, OC, NO_3, and EC, which can be explained by the higher spatial heterogeneity associated with NO_3 and EC concentrations.

		With $PM_{2.5}$			Without $PM_{2.5}$		
		R^2	RMSE[1]	95% Cvg[2]	R^2	RMSE[1]	95% Cvg[2]
5fold CV	EC	0.76	0.36	0.95	0.69	0.41	0.95
	OC	0.83	1.22	0.95	0.59	1.92	0.96
	SO_4	0.79	0.49	0.96	0.72	0.57	0.95
	NO_3	0.75	1.29	0.95	0.62	1.60	0.95
Spatial CV	EC	0.64	0.45	0.94	0.52	0.52	0.93
	OC	0.71	1.62	0.92	0.46	2.20	0.93
	SO_4	0.75	0.53	0.95	0.70	0.59	0.95
	NO_3	0.72	1.36	0.94	0.55	1.76	0.95

Table 1. 5-fold ordinary and leave-one-monitor-out spatial cross-validation (CV) results for evaluating BART prediction performance using tuned parameters, with and without $PM_{2.5}$ total mass in the predictor set.

[1]root mean-square prediction error

[2]empirical coverage probability of the 95% prediction interval

RMSE and R^2 for using default BART settings are given in Supplementary Table S1. We found when using the default setting with prior distributions and 200 trees, the models showed evidence of overfitting as the 95% prediction intervals have lower coverage probability than desired. This under-coverage is likely due to an under-estimation of the true residual variability in the model. However, when we reduce the number of trees and decrease the depth of trees, we can achieve a better 95% coverage probability, sacrificing little R^2. In some cases, R^2 improves further with tuning, especially in predicting at locations without monitors (e.g. spatial CV for NO_3).

Figure 1 shows the usefulness of including AOD parameters or CMAQ simulations in the set of predictors. In spatial CV, including AOD parameters or CMAQ simulations can improve R^2, especially when $PM_{2.5}$ is not included as a predictor. However, including

$PM_{2.5}$ as a predictor results in greater R^2 improvement compared to including AOD and/or CMAQ. AOD parameters are most useful for predicting NO_3 and OC when $PM_{2.5}$ is not included as a predictor. Prediction performance for SO_4 and EC depend less on the inclusion of $PM_{2.5}$, AOD, and CMAQ compared to other species. Similar observations are found for RMSE (Supplementary Figure S2), 5-fold CV experiments (Supplementary Figure S3), and BART fitted with default settings. We tuned BART to have the desired 95% interval coverage for 5-fold CV. The resulting spatial CV predictions all have coverage above 90% regardless of the set of predictors used (CMAQ or AOD). However, the default BART predictions result in poorer coverage sometimes under 80% (Supplementary Figure S4).

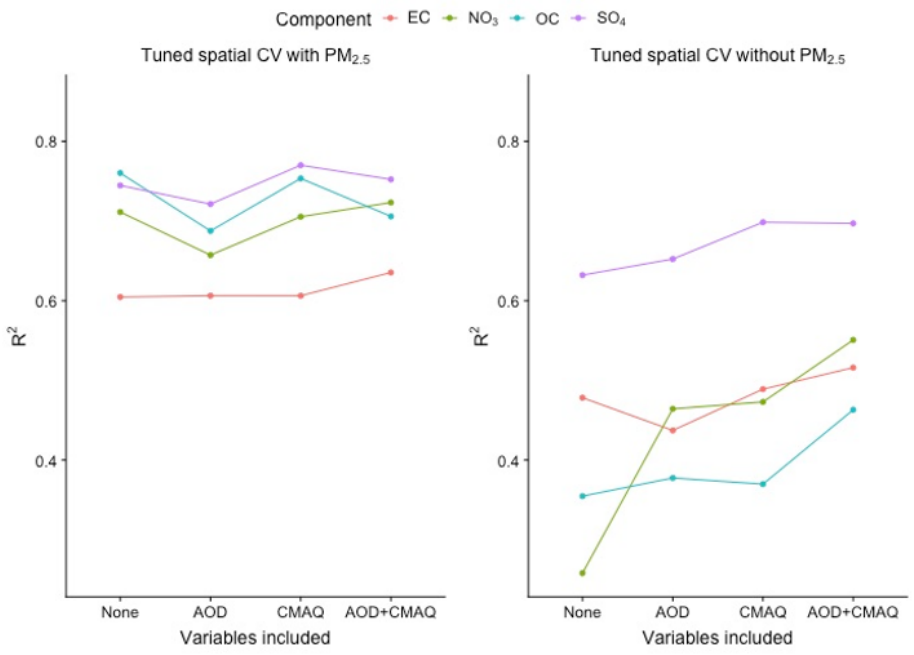

Figure 1. R^2 of leave-one-monitor-out spatial cross-validation (CV) results comparing the inclusion of AOD-only, CMAQ-only, AOD and CMAQ in addition to other meteorological and land use variables.

Figure 2 describes the importance for AOD parameters in different BART models with and without the presence of other predictors. Here variable importance is measured by calculating the number of times a variable is used for splitting nodes across MCMC iterations. The pattern of variable importance is robust in models with only AOD (red), with AOD and CMAQ (green), or with AOD and $PM_{2.5}$ (blue). However, the relative importance of different AOD parameters varies across $PM_{2.5}$ components. For predicting NO_3, AOD_3 is highly important, followed by MISR total AOD, spherical, and non-spherical AOD. But for OC, AOD_2 is the most important. For SO_4, many AOD parameters show moderate importance. For EC, none of the AOD parameter shows high importance. We note that our AOD predictors are not independent. For example, AOD_3 is part of total AOD and part of non-spherical AOD. BART is able to handle highly correlated predictors because of the sum-of-tree approach. Specifically, the estimation of each individual decision tree is based on the model residuals accounting all other trees. Hence highly correlated predictors may be less likely to appear across trees.

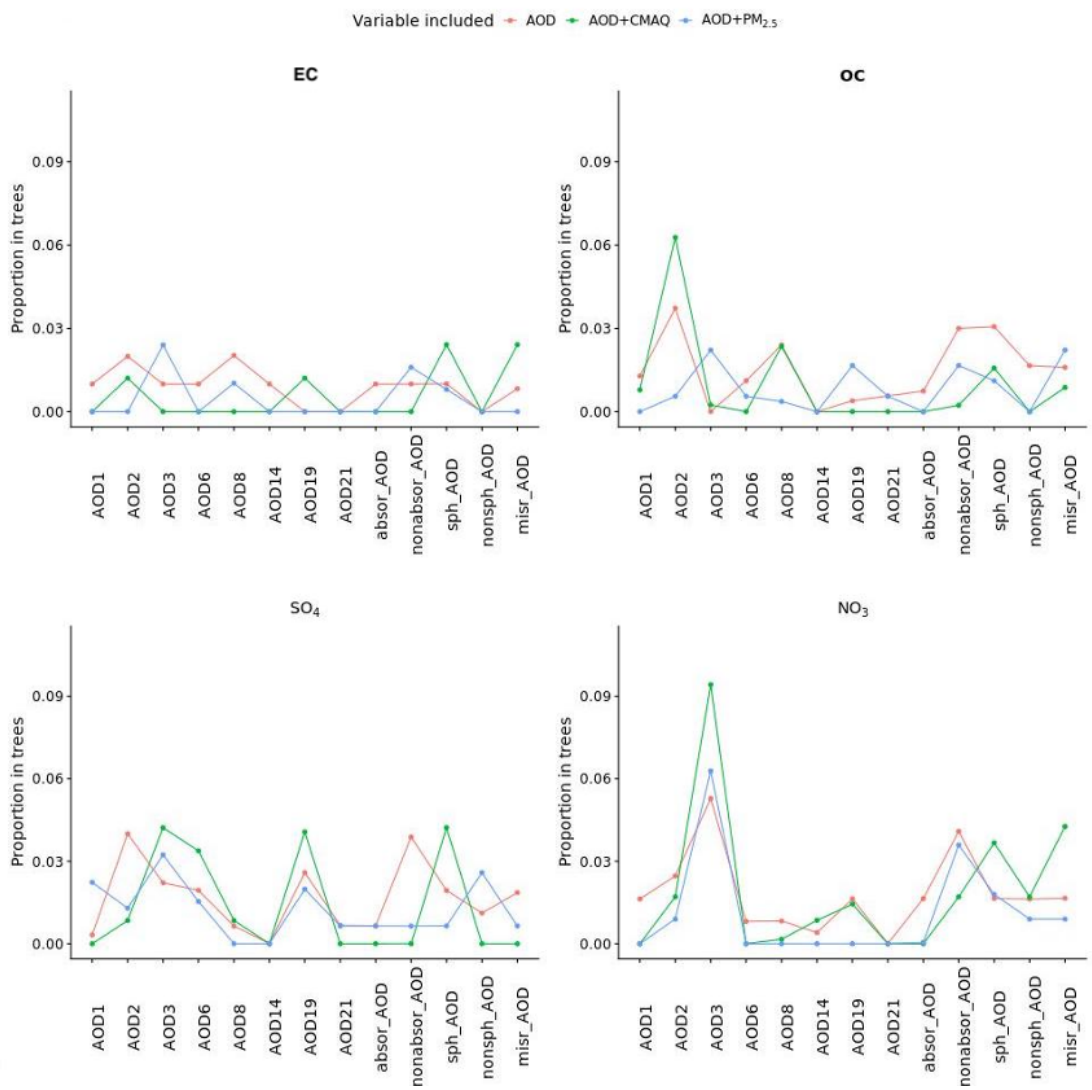

Figure 2. Variable importance of individual AOD fractional components for BART with tuned parameters under different predictor sets (with AOD, with AOD and CMAQ, with AOD and $PM_{2.5}$ total mass). All models include meteorology and land use predictors.

Figure 3 describes the importance for CMAQ simulations in different BART models with and without the presence of other predictors. Similar to AOD, the pattern of variable importance for CMAQ is robust across models. All $PM_{2.5}$ components depend on CMAQ heavily, specifically on the corresponding pollutant (i.e. CMAQ simulation for EC has

highest importance for predicting EC concentration). $CMAQ_{NH4}$ is also predictive of SO_4 and NO_3 because ammonium nitrate and ammonium sulfate are major components of $PM_{2.5}$. Generally, including $PM_{2.5}$ reduces the importance of CMAQ simulations.

When BART is not tuned, variance importance is less distinct across different AOD parameters (Supplementary Figure S5) and different CMAQ simulations (Supplementary Figure S6), demonstrating that turning BART also results in more interpretable models. Finally, supplementary Figure S7 shows the variable importance of all variables in a BART model without $PM_{2.5}$. We observe different meteorological and land use variable importance for different $PM_{2.5}$ constituents. For example, specific humidity is an important predictor for SO_4 and NO_3, while impervious surface is an important predictor for EC and OC.

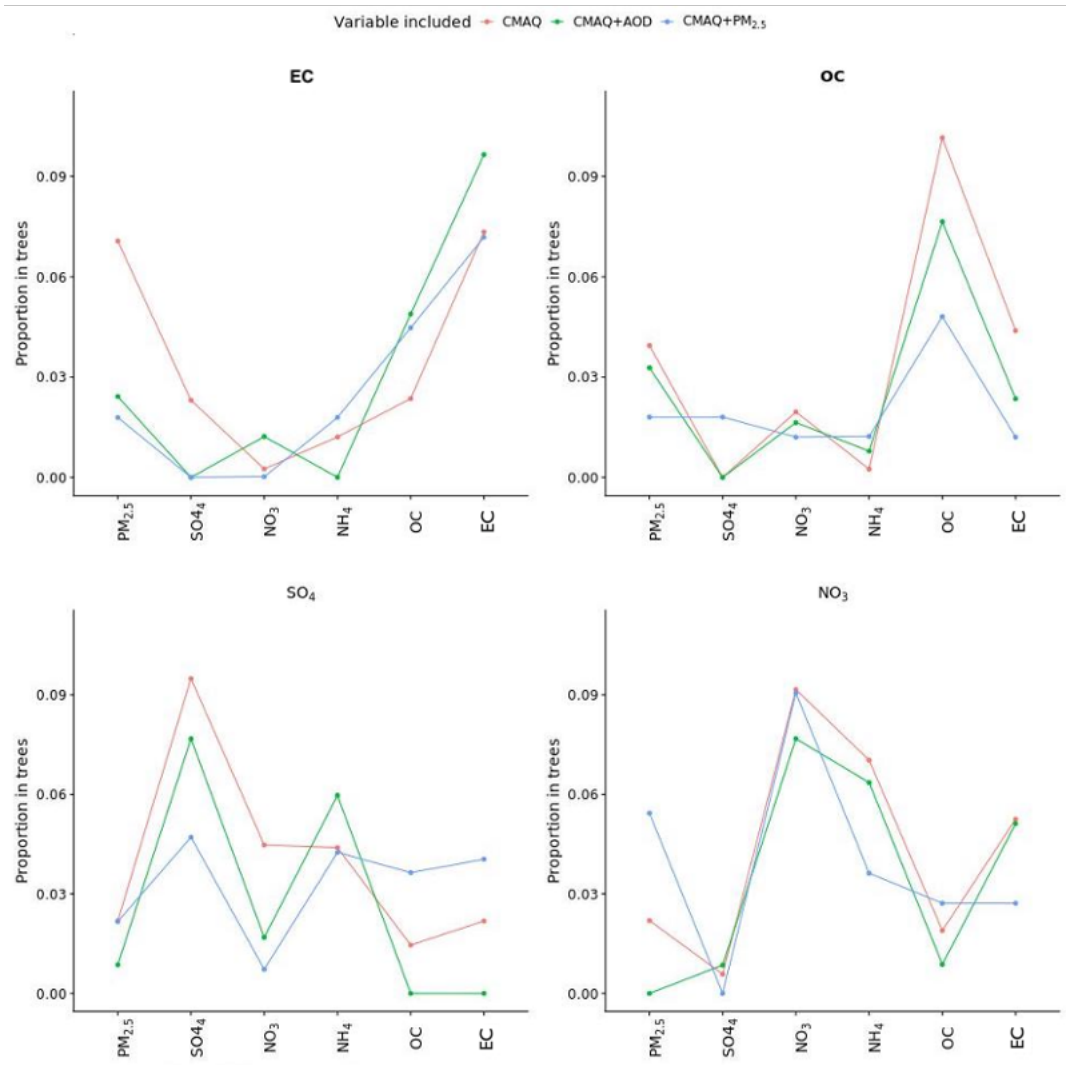

Figure 3. Variable importance of individual CMAQ variables for BART with tuned parameters under different predictor sets (with CMAQ, with AOD and CMAQ, with CMAQ and $PM_{2.5}$ total mass). All models include meteorology and land use predictors.

4. Discussion

In this paper, we demonstrate the usefulness of BART for predicting $PM_{2.5}$ components. A national spatial exposure model with partial least squares and universal kriging for predicting annual average concentrations of $PM_{2.5}$ elemental carbon, organic carbon,

silicon and sulfur achieved R^2 ranging from 0.62 to 0.95 (Bergen et al. 2013). Though comparison between national and regional models is challenging. This is because a national analysis usually has greater variability in the predictor values, resulting in better model. A Random forest model reached a spatial R^2 0.62, 0.62, 0.54 and 0.58 for $PM_{2.5}$ species sulfate, nitrate, OC and elemental carbon (EC) concentrations using the same predictors in spatial CV (Geng et al. 2020). A hybrid prediction model using a chemical transport model as well as land use regression at 1 km × 1 km grid cell showed a ten-fold CV and leave-one-day-out CV prediction R^2 results around 0.70–0.80 for $PM_{2.5}$ components sulfate, nitrate, organic carbon, elemental carbon, ammonium, sea salt and dust (Di, Koutrakis, and Schwartz 2016). However, the above methods do not provide uncertainty measures that can be subsequently used in health impact and health effect analyses. Furthermore, unlike Di et al. (2016), we did not include observed $PM_{2.5}$ component as spatial or lagged predictors. While this can result in improved R^2 when conducting cross-validation analysis, the actual prediction performance will likely depend on the spatial location and availability of monitoring data.

One disadvantage of BART is that it is not designed specifically for spatially correlated data such as ambient air pollution levels. Although geographic information such as latitude and longitude were used as predictors in our model, BART may not capture small-scale spatial dependence in the outcome. Hence, it would be interesting to incorporate spatially correlated residual into BART and examine the potential improvement in its performance. Thus, our future work includes adding spatial extension to BART as follows. We allow the residual of model to have two components, a spatially

correlated residual that depends solely on the spatial coordinates and an independent residual. Let s denote the spatial location of the outcome, our extended BART is shown below:

$$Y_{i,s} = \sum_{j=1}^{m} g\left(x_i; T_j, M_j\right) + \epsilon_s + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

where $\sum_{j=1}^{m} g\left(x_i; T_j, M_j\right)$ is the same mean structure as in ordinary BART, ϵ_i is the independent mean-zero Gaussian residual error that independently and identically follows $N(0, \sigma^2)$, and the spatial residual part ϵ_s is a mean-zero Gaussian process with stationary and isotropic covariance function Σ. Element Σ_{ij} is determined by a parametric covariance function $C(d; \theta)$ where d is the Euclidean distance between the locations of observations i and j. The covariance function $C(d; \theta)$ could be the commonly used squared exponential covariance function, the exponential covariance function, or the more general Matérn covariance function which has two parameters that include the other two as special cases. Estimation can be achieved by modifying the existing MCMC implement of BART in C++. Specifically, in each step of the MCMC iteration, we first subtract the observed y_i by the spatial residual estimated at that location and then input that into the BART model. Model develop and application are currently in progress.

In conclusion, this study showcases the application of the statistical learning method, BART, in modeling ambient air pollution. BART has received increasing attention in machine learning because it combines elements of ensemble learning and statistical inference. Specifically, BART can provide model-based uncertainties in predictions and be flexibly extended to incorporate external information with the use of probabilistic distributions on model parameters. Our results also highlight the relative importance of

$PM_{2.5}$, AODs and CMAQ simulations for predicting daily concentrations of $PM_{2.5}$ components.

5. Bibliography

Bergen, Silas, Lianne Sheppard, Paul D Sampson, Sun-young Kim, Mark Richards, Sverre Vedal, Joel D Kaufman, and Adam a Szpiro. 2013. "A National Prediction Model for PM2.5 Component Exposures and Measurement Error – Corrected Health Effect Inference." *Environmental Health Perspectives* 121 (9): 1017–25.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2012. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 6 (1): 266–98. https://doi.org/10.1214/09-AOAS285.

Di, Qian, Petros Koutrakis, and Joel Schwartz. 2016. "A Hybrid Prediction Model for PM2.5 Mass and Components Using a Chemical Transport Model and Land Use Regression." *Atmospheric Environment* 131: 390–99. https://doi.org/https://doi.org/10.1016/j.atmosenv.2016.02.002.

Fann, Neal, Evan Coffman, Anjum Hajat, and Sun Young Kim. 2019. "Change in Fine Particle-Related Premature Deaths among US Population Subgroups between 1980 and 2010." *Air Quality, Atmosphere and Health* 12 (6): 673–82. https://doi.org/10.1007/s11869-019-00686-9.

Franklin, Meredith, Olga v. Kalashnikova, and Michael J. Garay. 2017. "Size-Resolved Particulate Matter Concentrations Derived from 4.4 Km-Resolution Size-Fractionated Multi-Angle Imaging SpectroRadiometer (MISR) Aerosol Optical Depth over Southern California." *Remote Sensing of Environment* 196: 312–23. https://doi.org/10.1016/j.rse.2017.05.002.

Geng, Guannan, Xia Meng, Kebin He, and Yang Liu. 2020. "Random Forest Models for PM2.5 Speciation Concentrations Using MISR Fractional AODs." *Environmental Research Letters*. https://doi.org/10.1088/1748-9326/ab76df.

Grahame, Thomas J. 2014. "PM2.5 Species: Importance of Accurate Measurement." *Epidemiology* 25 (4): 615. https://doi.org/10.1097/EDE.0000000000000112.

Hern, Belinda, Adrian E Raftery, R Stephen, and C O Jul. 2015. "Bayesian Additive Regression Trees Using Bayesian Model Averaging," 1–32.

Hu, Xuefei, Jessica H. Belle, Xia Meng, Avani Wildani, Lance A. Waller, Matthew J. Strickland, and Yang Liu. 2017. "Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach." *Environmental Science and Technology* 51 (12): 6936–44. https://doi.org/10.1021/acs.est.7b01210.

Huang, Keyong, Qingyang Xiao, Xia Meng, Guannan Geng, Yujie Wang, Alexei Lyapustin, Dongfeng Gu, and Yang Liu. 2018. "Predicting Monthly High-Resolution PM2.5 Concentrations with Random Forest Model in the North China Plain." *Environmental Pollution* 242: 675–83. https://doi.org/10.1016/j.envpol.2018.07.016.

Janssen, N. A.H., P. Fischer, M. Marra, C. Ameling, and F. R. Cassee. 2013. "Short-Term Effects of PM2.5, PM10 and PM2.5-10 on Daily Mortality in the Netherlands." *Science of the Total Environment* 463–464: 20–26. https://doi.org/10.1016/j.scitotenv.2013.05.062.

Kelly, James T., Adam Reff, and Brett Gantt. 2017. "A Method to Predict PM2.5 Resulting from Compliance with National Ambient Air Quality Standards."

Atmospheric Environment 162: 1–10. https://doi.org/10.1016/j.atmosenv.2017.05.009.

Lary, D. J., T. Lary, and B. Sattler. 2015. "Using Machine Learning to Estimate Global PM2.5 for Environmental Health Studies." *Environmental Health Insights* 9 (S1): 41–52. https://doi.org/10.4137/EHI.S15664.

Li, Shuang, Liang Zhai, Bin Zou, Huiyong Sang, and Xin Fang. 2017. "A Generalized Additive Model Combining Principal Component Analysis for PM2.5 Concentration Estimation." *ISPRS International Journal of Geo-Information* 6 (8). https://doi.org/10.3390/ijgi6080248.

Linero, Antonio R. 2017. "A Review of Tree-Based Bayesian Methods" 24 (6): 543–59.

Meng, Xia, Jenny L. Hand, Bret A. Schichtel, and Yang Liu. 2018. "Space-Time Trends of PM 2.5 Constituents in the Conterminous United States Estimated by a Machine Learning Approach, 2005–2015." *Environment International* 121 (August): 1137–47. https://doi.org/10.1016/j.envint.2018.10.029.

Niu, Mingfei, Kai Gan, Shaolong Sun, and Fengying Li. 2017. "Application of Decomposition-Ensemble Learning Paradigm with Phase Space Reconstruction for Day-Ahead PM2.5 Concentration Forecasting." *Journal of Environmental Management* 196: 110–18. https://doi.org/10.1016/j.jenvman.2017.02.071.

Rohr, Annette C, and Ronald E Wyzga. 2012. "Attributing Health Effects to Individual Particulate Matter Constituents." *Atmospheric Environment* 62: 130–52. https://doi.org/https://doi.org/10.1016/j.atmosenv.2012.07.036.

Schlesinger, Richard B. 2007. "The Health Impact of Common Inorganic Components of Fine Particulate Matter (PM2.5) in Ambient Air: A Critical Review." *Inhalation Toxicology* 19 (10): 811–32. https://doi.org/10.1080/08958370701402382.

Thurston, George D., Richard T. Burnett, Michelle C. Turner, Yuanli Shi, Daniel Krewski, Ramona Lall, Kazuhiko Ito, et al. 2016. "Ischemic Heart Disease Mortality and Long-Term Exposure to Source-Related Components of U.S. Fine Particle Air Pollution." *Environmental Health Perspectives* 124 (6): 785–94. https://doi.org/10.1289/ehp.1509777.

Weyuker, Elaine J., Thomas J. Ostrand, and Robert M. Bell. 2010. "Comparing the Effectiveness of Several Modeling Methods for Fault Prediction." *Empirical Software Engineering* 15 (3): 277–95. https://doi.org/10.1007/s10664-009-9111-2.

Appendix A. Data and Data Processing

Our study area includes the state of California, as well as an additional 80km buffer as shown in Figure S1 (Geng et al. 2020). A 1km x 1km grid was constructed for this region for defining various predictors.

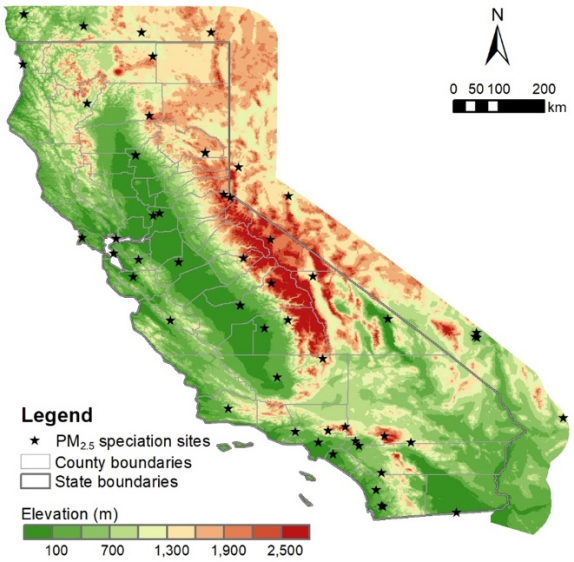

Figure S1. Locations of $PM_{2.5}$ monitors. Elevations are shown in the background. This figure is produced from Geng et al. (2020).

Daily concentrations of $PM_{2.5}$ sulfate, nitrate, organic carbon (OC), and EC in the study area were obtained from the CSN network (http://aaqsdr1.epa.gov) and the IMPROVE network (http://views.cira.colostate.edu/fed) between 2005 and 2014. Overall, there are 55 $PM_{2.5}$ monitors in our research region and their locations are shown in Figure 1. Following similar data processing steps in Meng et al. (2018), OC and EC measurements from CSN were converted to the IMPROVE standards. Each monitor was assigned to a 1km grid cell.

We obtained satellite-derived AOD from Multi-angle Imaging SpectroRadiometer (MISR). MISR simultaneously retrieves data from nine different angles, which provides data to distinguish the aerosol particles. We downloaded Aerosol Data V23 level 2 for the years 2005-2014 from the NASA Earthdata portal (https://search.earthdata.nasa.gov/) which contains 74 different aerosol components. In addition, eight tractional AOD components (i.e., component 1, 2, 3, 6, 8, 14, 19 and 21) were developed to represent the different particle shapes, scattering properties and effective radius for a log-normal distribution. We used the following equation to convert any MISR aerosol observation to the fractional AOD components:

$$\text{AOD } i = \frac{\sum_{j=1}^{74} \alpha AOD_{\text{mixture j}} \times Fraction_{\text{component } i \text{ in mixture } j}}{Number\ of\ successful\ mixtures}$$

where $AOD_{\text{mixture j}}$ is the AOD mixture j; $Fraction_{\text{component } i \text{ in mixture } j}$ is the contribution of component i to the AOD for mixture j; if mixture j is retrieved successfully, then α=1, otherwise α=0. We also considered different sums of the 8 AOD components for absorbing, non-absorbing, spherical and non-spherical particles. The numerical model simulations used in this study were based on the Community Multiscale Air Quality (CMAQ) model version 2.0.0 which used meteorological conditions from the Meteorological Research and Forecast (WRF) model version v3.4. Details of the model configuration for WRF and CMAQ can be found in Zhang et al. (2019). The National Land Cover Database (NLCD) was used as the input to the WRF model, and the Meteorology-Chemistry Interface Processor version 4.1.3 was used to generate the input to the atmospheric parameters in CMAQ model. The chemical boundary conditions for the CMAQ model were derived from the annual-specific simulation of the global GEOS-

Chem model. Anthropogenic emissions inputs were based on data from 2005, 2008 and 2011 National Emissions Inventories.

Daily temperature, wind speed and humidity data for the spatial resolution of approximately 13 km were obtained from the North America Land Data Assimilation Systems phase 2 (NLDAS-2, http://ldas.gsfc.nasa.gov/nldas/). Meteorological data were all averaged between 9:00 am to 12:00 pm and interpolated to the 1km grid cells by inverse-distance weighting. Elevation data were based on the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Map (GDEM) (https://asterweb.jpl.nasa.gov/gdem.asp) version 2.

Data on road networks were obtained from ESRI StreetMap USA (Environmental Systems Research Institute, Inc., Redlands, CA). Impervious surface, forest cover, shrub cover and cultivated land cover information at 30 m spatial resolution were taken from NLCD (https://www.mrlc.gov) for the year 2006 and 2011. Population data at 1 km spatial resolution were extracted from the LandScan Global Population Database (https://landscan.ornl.gov/). Elevation, impervious surface, forest cover, shrub cover and cultivated land cover were averaged while population and road lengths were summed within each 1-km grid cell.

Bibliography

Geng, Guannan, Xia Meng, Kebin He, and Yang Liu. 2020. “Random Forest Models for PM2.5 Speciation Concentrations Using MISR Fractional AODs.” *Environmental Research Letters*. https://doi.org/10.1088/1748-9326/ab76df.

Meng, Xia, Jenny L. Hand, Bret A. Schichtel, and Yang Liu. 2018. “Space-Time Trends of PM 2.5 Constituents in the Conterminous United States Estimated by a Machine Learning Approach, 2005–2015.” *Environment International* 121 (August): 1137–47. https://doi.org/10.1016/j.envint.2018.10.029.

Zhang, Yuqiang, Kristen M. Foley, Donna B. Schwede, Jesse O. Bash, Joseph P. Pinto, and Robin L. Dennis. 2019. “A Measurement-Model Fusion Approach for Improved Wet Deposition Maps and Trends.” *Journal of Geophysical Research: Atmospheres* 124 (7): 4237–51. https://doi.org/10.1029/2018JD029051.

Appendix B. Supplementary Tables and Figures

		With $PM_{2.5}$			Without $PM_{2.5}$		
		R^2	RMSE[1]	95% Cvg[2]	R^2	RMSE[1]	95% Cvg[2]
5fold CV	EC	0.76	0.36	0.95	0.69	0.41	0.95
	OC	0.83	1.22	0.95	0.59	1.92	0.96
	SO_4	0.79	0.49	0.96	0.72	0.57	0.95
	NO_3	0.75	1.29	0.95	0.62	1.60	0.95
Spatial CV	EC	0.64	0.45	0.94	0.52	0.52	0.93
	OC	0.71	1.62	0.92	0.46	2.20	0.93
	SO_4	0.75	0.53	0.95	0.70	0.59	0.95
	NO_3	0.72	1.36	0.94	0.55	1.76	0.95

Table S1. BART performance in default setting

[1]root mean-square prediction error

[2]empirical coverage probability of the 95% prediction interval

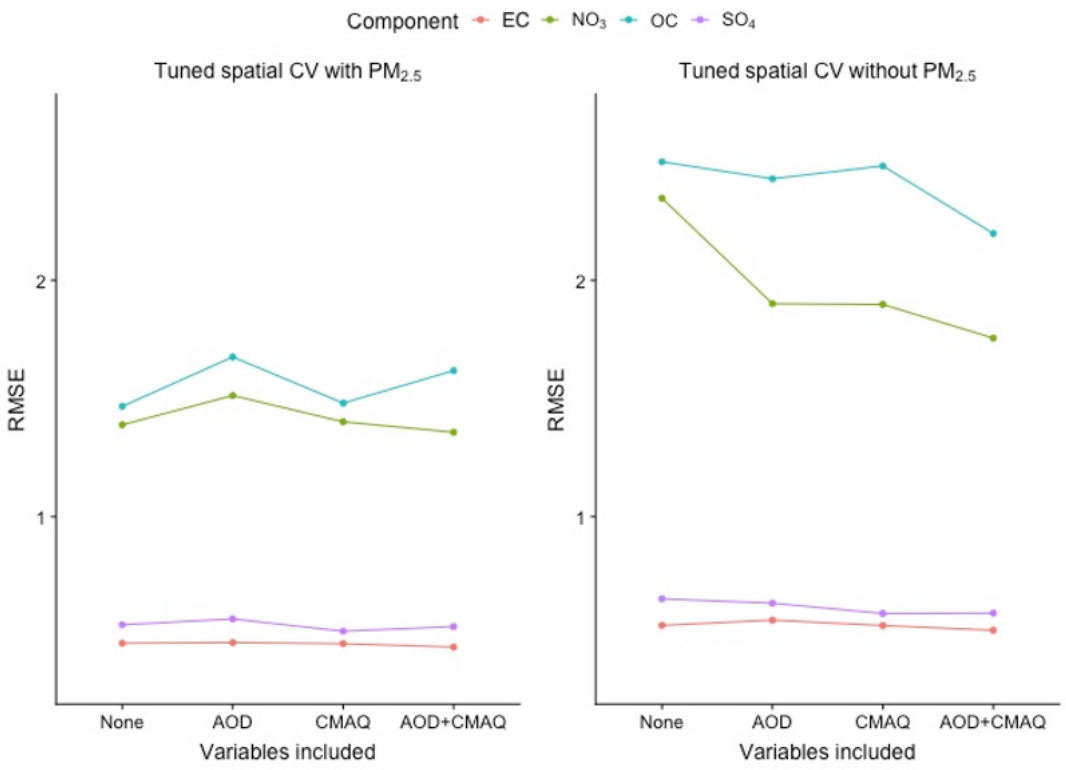

Figure S2. RMSE in spatial CV when parameters are tuned

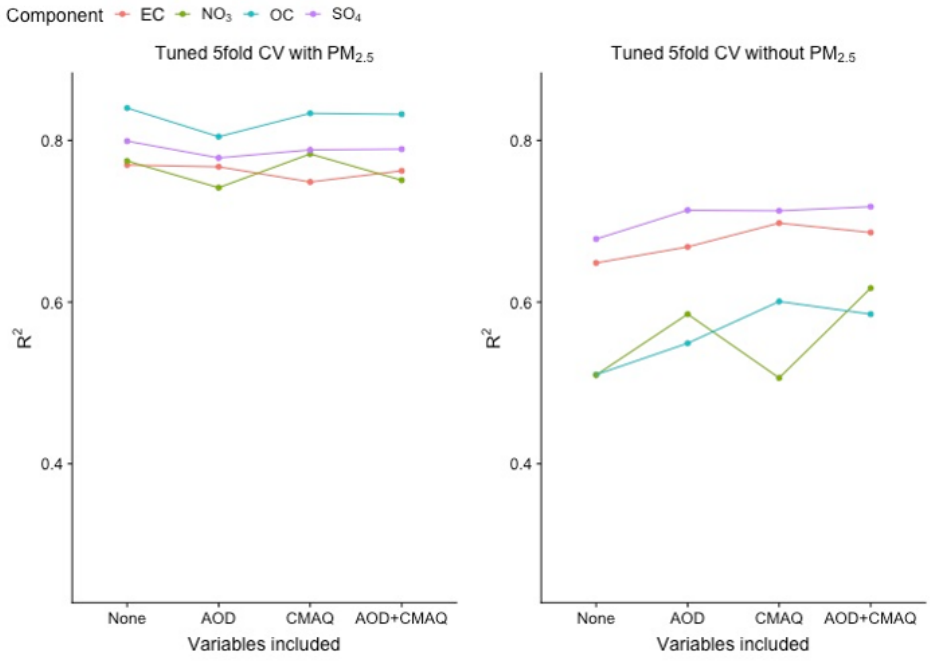

Figure S3. R^2 in 5fold CV when parameters are tuned

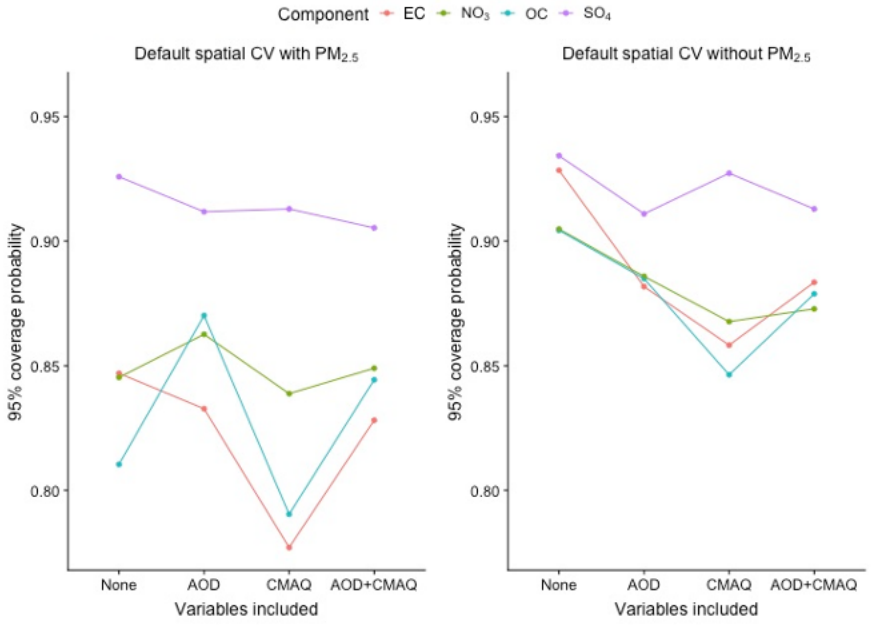

Figure S4. 95% coverage probability in spatial CV in default setting

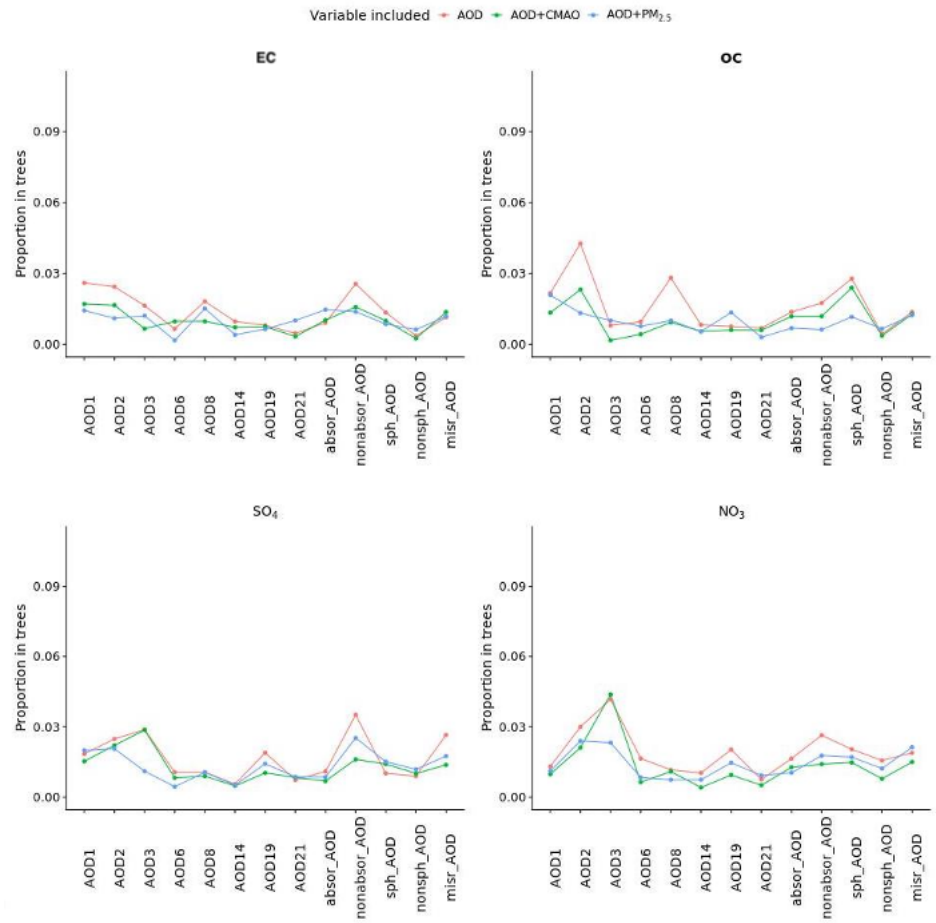

Figure S5. AOD component importance in default setting

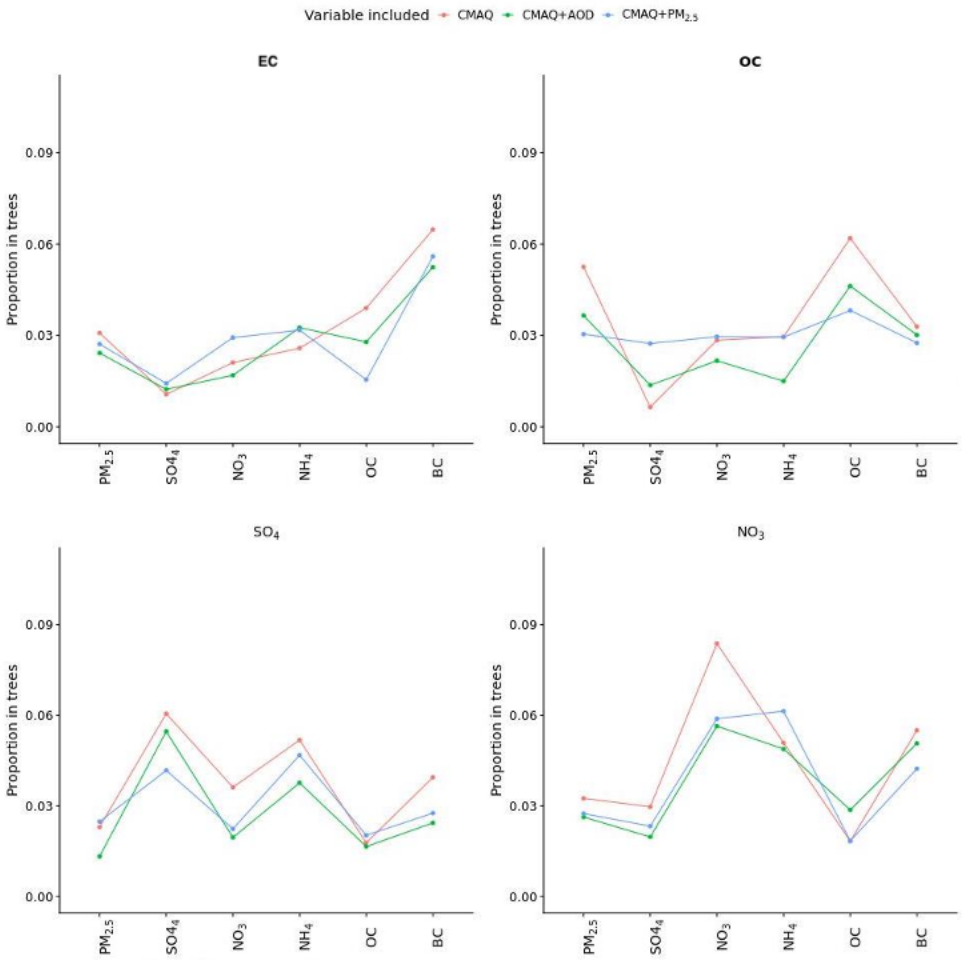

Figure S6. CMAQ component importance in default setting

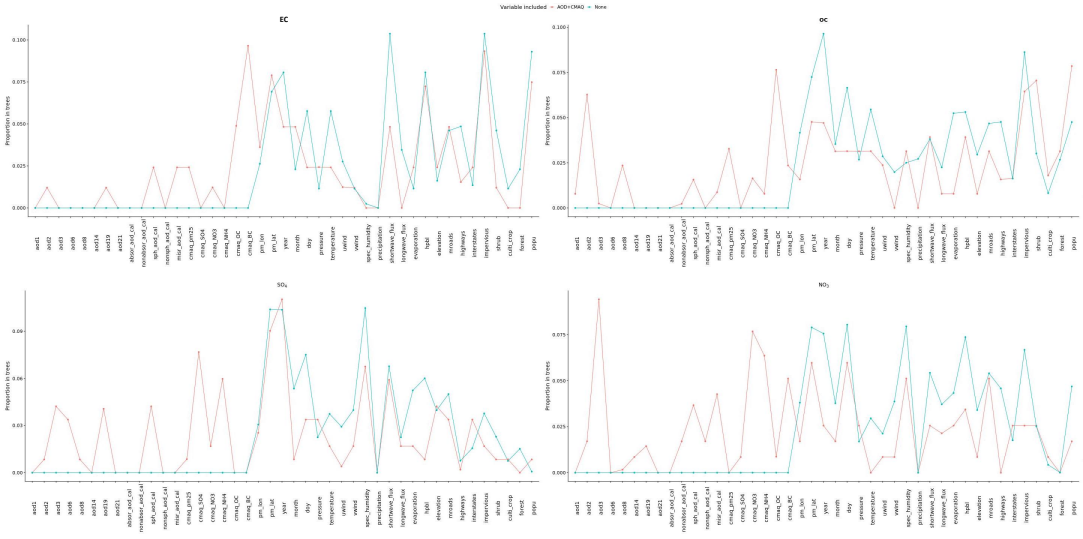

Figure S7. Variable importance when parameters are tuned and $PM_{2.5}$ is not a predictor