University of Cincinnati

Date: 3/23/2016

I, Richard C Brokamp, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Biostatistics (Environmental Health).

It is entitled:

Land Use Random Forests for Estimation of Exposure to Elemental Components of Particulate Matter

Student's name: **Richard C Brokamp**

This work and its defense approved by:

Committee chair: Patrick Ryan, Ph.D.

Committee member: Roman A. Jandarov, PH.D.

Committee member: Marepalli Rao, Ph.D.

UNIVERSITY OF Cincinnati

19684

Land Use Random Forests for Estimation of Exposure to Elemental Components of Particulate Matter

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Division of Biostatistics and Bioinformatics

of the Department of Environmental Health

of the College of Medicine

by

Richard Coleman Brokamp

March 2016

B.S. University of Cincinnati

Committee Chair: Patrick H. Ryan, Ph.D.

Abstract

Particulate matter (PM) has long been known to have a negative effect on public health. Epidemiological studies associating air pollution and other sources of PM often rely on land use modeling for exposure assessment. This approach relies on the association of characteristics of the surrounding land with PM concentrations. Land use regression (LUR) is the most commonly implemented land use model and has several drawbacks, including model instability due to correlated predictors and an inability to capture non-linear relationships and complex interactions. Here, I utilize the machine learning random forest model within a land use framework to generate a novel land use random forest (LURF) model. Using ambient air sampling data from the Cincinnati Childhood Allergy and Air Pollution (CCAAPS) study, I developed LURF and LUR models for eleven elemental components of particulate matter, including Al, Cu, Fe, K, Mn, Ni, Pb, S, Si, V, Zn. We show that LURF models utilized a higher number and more diverse selection of land use predictors than the LUR models. Furthermore, the LURF models were more accurate and precise predictors of all elemental PM concentrations, except for Fe, Mn, and Ni.

To extend the usability of the LURF models, I utilized the recent application of the infinitesimal jackknife (IJ) to the random forest model in order to estimate the prediction variance. The IJ theorems were originally verified under the assumptions of traditional random forest framework, namely using CART trees and bootstrap resampling. Alternatives to the traditional random forest, such as subsampling instead of bootstrap resampling and conditional inference trees instead of CART trees have been shown to increase the accuracy

of the random forest algorithm and eliminate its variable selection bias. Here, I conduct simulation experiments to show that the IJ performs well when using these random forest variations. Specifically, using the conditional inference tree instead of the CART tree and subsampling instead of bootstrap resampling results in increasing the accuracy and precision of the IJ estimator of random forest prediction variance.

To associate the exposure of elemental components of PM with respiratory health, I applied the novel LURF model to the CCAAPS cohort. The exposures of children in this ongoing, prospective birth cohort located in Cincinnati, Ohio were calculated using their residential address history. Comparison of estimated elemental exposures with total PM2.5 estimated exposure showed that they were not correlated. Lung function and asthma testing was conducted on all children at age seven. We found that Al, Fe, Pb, Si, Zn, and total PM2.5 were associated with decreased lung function on their own, but after unconfounding the effect of exposure to PM with neighborhood level effects, the associations generally disappeared.

Finally, I discuss the current limitation of two-stage models that associate spatial pollutants with health effects, namely omitting the uncertainty from the exposure assessment stage. When incorporating this uncertainty in the future, the increased accuracy and precision of LURF models compared to LUR models should allow for more precise estimation of the health effects of these pollutants.

Acknowledgments

I would like to thank Dr. Patrick Ryan for guiding me throughout the PhD process and serving as a personal and career mentor. He has shown me by example what a successful and responsible scientific researcher is and I am very thankful for the opportunities that he has created for me. I also want to thank Dr. MB Rao for serving as my academic advisor and inspiring me to forever maintain an enthusiastic pursuit of learning. Thanks also to Dr. Roman Jandarov for serving on my dissertation committee, Dr. Erin Haynes and Dr. Jeff Welge for serving on my qualifying examination committee, and the faculty and staff in the UC Department of Environmental Health and the CCHMC Department of Biostatistics and Epidemiology for their time and support.

Dedication

This work is dedicated to my parents who were my first teachers and mentors. I will always be grateful for the sacrifices you have made to help me succeed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Exposure assessment for air pollution is a complex problem due to the high spatial and temporal variation of airborne pollutants. The current regression-based approaches have several drawbacks, including model instability due to correlated predictors and the inability to capture non-linear relationships and complex interactions. The random forest is a regression alternative whose strengths address the disadvantages of regression in a land use model setting. Here, we develop novel land use random forest models and show that they are more accurate and precise than land use regression models.

Although random forests have been proven to be more accurate than common parametric techniques like regression, they still remain underused because most researchers utilize them only for prediction and not for interpretation. A recent application of the infinitesimal jackknife to the resampling distribution has allowed for estimation of prediction variances of traditional random forests. Variations to the traditional random forest algorithm such as different resampling methods and different base learner types have become more widely used because they increase prediction accuracy and reduce variable selection bias. However, because the infinitesimal jackknife formulas are proven under the assumption of traditional random forests, they are hard to verify when using these common variations. Here, we conduct simulation experiments to test the application of the infinitesimal jackknife estimator on the common random forest variations.

Although particulate matter has long been known to have a negative effect on public health, recently the individual elemental components of air pollution have been shown to be important as well. Here, we use our novel land use models to estimate the exposure of a cohort of children from the Cincinnati Childhood Allergy and Air Pollution Study to elemental components of particulate matter. We show that total particulate matter exposure does not fully represent the individual elemental exposure signature. Furthermore, we associate exposure to individual elemental particulate matter components with respiratory health.

Chapter 2

Land Use Models for Elemental Components of Particulate Matter

Land Use Models for Elemental Components of Particulate Matter in an Urban Environment: A Comparison of Regression and Random Forest Models

Cole Brokamp[1], Roman Jandarov[1], MB Rao[1], Grace LeMasters[1,2], Patrick Ryan[1,3]

[1]Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA

[2]Division of Asthma Research, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[3]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Corresponding Author:

Cole Brokamp

University of Cincinnati

Kettering Lab Building, 160 Panzeca Way, Cincinnati, OH 45267-0056

E-mail: brokamrc@mail.uc.edu

Acknowledgments:

The authors thank the study and clinic staff for their efforts in study coordination, subject recruitment, data management, and data collection. They also thank the participating CCAAPS families for their time and effort. This work was supported by grants from the National Institute of Environmental Health Sciences (5R01ES011170 and R01ES019890).

The authors declare no competing financial interests.

Abstract

Background: Exposure assessment for elemental components of particulate matter (PM) using land use modeling is a complex problem due to the high spatial and temporal variations in pollutant concentrations at the local scale. Land use regression (LUR) models often fail to capture complex interactions and non-linear relationships. The increasing availability of big spatial data and machine learning methods present an opportunity for improvement in PM exposure assessment models.

Objectives: Our objective was to develop a novel land use random forest (LURF) model and compare its accuracy and precision to a LUR model for elemental components of PM in the urban city of Cincinnati, Ohio.

Methods: PM2.5 and eleven elemental components were measured at 24 sampling stations from the Cincinnati Childhood Allergy and Air Pollution Study (CCAAPS). Over 50 different predictors associated with transportation, physical features, community socioeconomic characteristics, greenspace, land cover, and emission point sources were used to construct LUR and LURF models. Cross validation was used to quantify and compare model performance.

Results: LURF and LUR models were created for Al, Cu, Fe, K, Mn, Ni, Pb, S, Si, V, Zn, and total PM2.5 in the CCAAPS study area. LURF utilized a more diverse and greater number of predictors than LUR in its final models. Cross validation revealed that LURF models had a lower predictive error than LUR models for all elements except Fe, Mn, and Ni. Furthermore, LUR models showed a differential exposure assessment bias and had a higher prediction error variance. LURF models predictions were approximately ten fold more precise compared to LUR model predictions. Random forest and other machine learning methods should be incorporated into future land use models for more accurate and precise exposure assessment.

2.1 Introduction

2.1.1 Land Use Regression Models

Land use modeling assumes that the spatial distribution of air pollutant concentrations are directly related to the use of the surrounding land. Physical features like elevation as well as the location and intensity of known pollutant sources including industrial emitters and traffic have been found to correlate well with pollutant concentrations [1, 2]. Specifically, land use regression (LUR) uses predictors within a regression framework and has been the main focus of almost all land use models, becoming a popular tool for exposure assessment in air pollution research [3, 4, 5, 6]. However, it is well known that land use modeling is a complex problem due to the high spatial and temporal variations in pollutant concentrations on the local scale [7, 8]. Although LUR models are valuable tools for air pollution exposure studies, the methodology has not included current predictive modeling techniques. Therefore, there is an opportunity to improve the accuracy and precision of land use models, resulting in better exposure assessment for air pollution related epidemiological studies.

2.1.2 Using Random Forest in Land Use Models

Land use models inherently use a high number of features that are highly correlated, for example, the length of highways within 100, 200, 300, and 400 meters. Selection of which features to use in the final model is the outstanding challenge in land use model building and several approaches have been implemented (see [9] for a review), all of which revolve around stepwise variable selection in a regression framework. Inclusion of correlated predictors generate problems for regression, often leading to unstable model estimates and variance inflation [10]. Another challenge rising from regression-based land use models is the difficulty in capturing non-linear relationships and complex interactions. Because of the usually small sample size ($n = 20$ to 40) and very large number of possible predictors ($p = 50$ to over 500), it is often not feasible to evaluate all possible regression models.

Random forests are resistant to all of these problems. A key advantage of random forest is its ability to capture complex and non-linear relationships between predictors and the outcome with small sizes of training data. Random forests may be more accurate predictors of pollutant concentrations if they can indeed capture more patterns based on land use data. Indeed, a random forest has been empirically shown to estimate concentrations of nitrogen dioxide based on land use data in the urban area of Geneva with a lower error when compared to regression [11]. This was a non-peer reviewed preliminary analysis included in a book and the authors did not compare the model's cross validated performance with a traditional land use regression model. We hypothesize that implementing land use random forest (LURF) models as an alternative to LUR models could result in more accurate and precise estimates of PM2.5 elemental component concentrations. Furthermore, since most land use models are concerned with accurate predictions rather than interpretability, these type of learners seem ideal for land use based predictions of elemental concentrations.

2.1.3 Random Forests

Random Forests (covered in more detail in Chapter 3) are often implemented in prediction analyses because of their increased accuracy and resistance to multi-collinearity and complex interaction problems as compared to linear regression [10]. In a recent study, random forest was found to be the most accurate classification algorithm among 179 classifiers, based on 121 different data sets [12]. Often times, such as in the case of land use models, accurate prediction is desired over interpretability and so random forests are an ideal candidate to improve the current implementation of land use models which all rely on regression approaches. The technique itself is an ensemble learning method that builds on bagging – specifically the bootstrapped aggregation of several regression trees – to predict an outcome. Bagging is most often used to reduce the variance of an estimated prediction function and is most useful for models which are unbiased but have a high variance, like regression trees [10]. Random Forests, first proposed by Breiman [13], modify the bagging technique by ensuring

that the individual trees are de-correlated by using a bootstrap sample for each tree and also randomly selecting a subset of predictors for testing at each split point in each tree. The random forest comes with the advantages of tree-based methods, namely the ability to capture complex interactions and maintain low bias, while at the same time alleviating the problem of high variance of predictions usually associated with tree-based methods by growing the individual trees to a very deep level (usually one observation per terminal node) and averaging their predictions.

2.1.4 Land Use Models for Elemental PM2.5 Components

Particulate matter (PM) is a complex mixture of chemical and elemental constituents and epidemiological studies have shown that these components and their sources are associated with adverse cardiovascular and respiratory health outcomes in adults [14, 15, 16]. Further studies suggest that certain components of PM2.5 are responsible for adverse health effects and characterizing these health effects of PM components has been identified as a research priority by the National Research Council for the National Academies [17]. Recently, successful LUR models have been developed for PM components in twenty areas in Europe as a part of the ESCAPE study [18] and for an urban area in Canada [19]. These land use models have allowed for assessment of exposure to individual components of PM and the study of their association with health outcomes [20, 21, 22]. Although some models have been developed, limited information on PM components has impeded progress in identifying their health effects [23]. A land use model for elemental components of PM in the United States has not been created.

2.1.5 Innovation and Relevance to Environmental Health

All previously used elemental exposure assessment models have been based on LUR. This chapter specifically lays out a framework for a novel LURF used to predict elemental concentrations of air pollution and compares the model accuracy and precision of LUR to LURF.

The air sampling data was originally collected as a part of the Cincinnati Childhood Allergy and Air Pollution Study (CCAAPS), which aims to assess the health impact of air pollution on an ongoing prospective birth cohort of high-risk atopic children in the Cincinnati, Ohio area [24, 25]. The models were built in order to assess exposure of the CCAAPS cohort to components of air pollution to determine their effect on respiratory health.

2.2 Methods

2.2.1 Elemental PM2.5 Measurements

Measurements were collected at 24 sites across Cincinnati, Ohio as a part of CCAAPS, with full details available elsewhere [3]. Briefly, sites were selected based on the location of the CCAAPS cohort as well as wind direction, and proximity to pollution sources. Figure 2.1 shows the location of the CCAAPS sampling sites and the birth addresses for the CCAAPS cohort. Between 2001 and 2004, PM2.5 samples were collected on 37-mm Teflon membrane filters and 37-mm quartz filters with Harvard-type Impactors. The increase in weight of the Teflon filters after sampling was used to determine the total PM2.5 mass [26] and X-ray fluorescence was used with the quartz filters to determine elemental concentrations for a total of 38 elements. Traffic related air pollution (TRAP) was calculated as the fraction of elemental carbon that was attributable to traffic by using a multivariate receptor model [27, 28], UNMIX, to identify source signatures. One of the signatures was identified as TRAP because it was similar to comparison measurements conducted for cluster sources of trucks and buses [26] in Cincinnati, Ohio. Mean elemental concentrations for each site were considered missing if at least 75% of their measurements were classified as below the threshold of measurement certainty. For implementation of the land use models, in addition to total PM2.5 and TRAP, we restricted our building of models to the following eleven elements, which were selected for their previous association with health effects and a high percentage ($\geq$ 75%) of detected samples: Al, Cu, Fe, K, Mn, Ni, Pb, S, Si, V, and Zn. All

elements had complete information for all sites ($n = 24$) except for V, which had one site with missing concentration information. All concentrations were log transformed prior to building models and back transformed to their natural scale after predictions.

2.2.2 Land Use Predictors

The predictors made available for inclusion in the final models for each element were based on previously validated LUR models for elements in PM2.5 [18, 19] and on a previously validated LUR model for TRAP built using the same ambient sampling data in Cincinnati, Ohio [3]. Where applicable, geographic predictors were extracted based on the area within circles centered on the sampling site locations with varying buffer radii. In brief, land use predictors included measurements related to road location, traffic intensity, elevation, community-level socioeconomic status, greenspace, land cover, and emission point sources. See Table 2.2 for a full list of predictors, their units of measurement, and buffer radii. Section 2.4.4 contains figures that illustrate some of the land use predictors for the Cincinnati area and the location of the CCAAPS cohort (Section 2.3.5).

Transportation. Roadway data (both distance and total length) were extracted from 2002 TIGER/Line shapefiles [29] for each of the 5 classes of roads and railroads. Class 1 roads are primary highways with limited access; class 2 roads are primary roads without limited access; class 3 roads are secondary and connecting roads; class 4 roads are local, neighborhood, and rural roads; and class 5 roads are vehicular trails [30]. Annual average daily truck count on both interstates and highways were obtained from the Ohio Department of Transportation based on data from 2002 - 2005. Bus routes were obtained from local transit authorities and intersections were identified as locations where class 3 or 4 roads intersected one another. Figure 2.10 shows the Class 1 and 2 roads and Figure 2.11 shows the average daily truck count on interstates and highways.

Physical Features. An elevation raster supplied by the US Geological Survey was used to identify the elevation and the mean elevation within a buffer radius. The standard deviation of elevation as well as the fraction of elevation points more than 20 meters uphill (or downhill) were separately calculated as a measure of the elevation gradient.

Community Characteristics. Population totals for each census block were retrieved from the 2000 US Census [31]. For varying buffer radii, the population count was defined as the sum of the total population of all census blocks for which the census block centroid was contained within the buffer radius. The population density was the population count divided by the total area of the census blocks which were included in the population count. Eight census tract level variables (fraction that graduated high school, fraction of households in poverty, median household income, fraction of population receiving public assisted income, fraction of houses that are vacant, median home value, white fraction of population, and black fraction of population) from the Census 2010 5-year American Community Survey for all census tracts in the counties where CCAAPS subjects resided were used to create a deprivation index [32]. Because of the high correlation among the individual variables, principal components analysis was used to extract representative measures of socioeconomic status for each census tract. The first principal component explained 60.6% of the total variance and was called the "deprivation index" because of its high loadings on the fraction that graduated high school, fraction of households in poverty, median household income, fraction of population receiving public assisted income, fraction of houses that are vacant, and median home value, with a higher value representing a census tract with increased deprivation. The index was normalized to a range of [0,1] by subtracting the minimum and dividing by difference of the resulting range. Figure 2.15 shows the deprivation index of each census tract for the study area.

Greenspace. Greenspace was estimated using satellite-derived normalized difference vegetation index (NDVI) images. A raster image of the Cincinnati area was obtained from

the United States Forest Service and the average NDVI within varying buffer radii of each sampling site was extracted. NDVI ranges from -1 to 1 and a higher value represents more surrounding greenspace. Figure 2.13 shows the greenspace of the study area. Briefly, a cloud-free composite image with a resolution of roughly 100 by 100 feet for all of the Cincinnati area was created based on individual images collected in June of 2000 that differed by no more than 15 calendar days. Imagery digital numbers were converted to top of atmosphere reflectance (ToAR) using the standard Landsat calibration process. ToAR was then converted to surface reflectance by using the 6S atmospheric correction procedure as described previously [33].

Land Cover. The 2001 National Landcover Database from the United States Geological Survey was used to extract the percentage of each land class within varying buffer radii from each location. The raster file classifies 30 by 30 meters grids of land into 15 different land use classes (see Table 2.2 for the full list of land classes). The classes and an overview of their distribution in the study area are shown in Figure 2.12.

NEI Point Sources. The 2011 National Emissions Inventory (NEI) is a national compilation of emissions sources collected from state and local agencies as well as information from the Environmental Protection Agency (EPA) emissions programs including the Toxics Release Inventory (TRI). Point source sites and total emissions were obtained from the NEI for PM2.5, PM10, and the available modeled elements (Ni, Pb, and Mn). Land use models extracted from the NEI data included the distance to the nearest point source, total number of point sources, total point source emissions, average point source emissions, and point source emissions weighted by inverse distance to the source. The location of the PM10 point sources from the NEI are shown in Figure 2.14.

2.2.3 Land Use Regression (LUR) Models

The approach for building the LUR models were based on other successful implementations of land use regression models for elemental components of PM2.5 in urban areas of Europe [18] and the urban city of Calgary, Alberta [19] as well as for TRAP in Cincinnati [3]. In general, the approach is a forward selection based algorithm which begins by ranking every variable based on their association with the elemental concentration. Each predictor in Table 2.2 was initially ranked based on their *model* R^2 value from a univariate regression with the elemental concentrations at all 24 sites. Because of the inherent correlation between variables of the same category (i.e. length of class 1 roads within 100 m and length of class 1 roads within 200 m), only one variable from each category was considered for inclusion into the model. The initial regression model was fit using only the highest associated predictor and the remaining predictors were tried for addition to the model in order of decreasing association with the elemental concentration. At each step, the predictor was retained only if it increased the *adjusted model* R^2 by at least 1% and the p value for the predictor and all other predictors already retained in the model were less than 0.1.

2.2.4 Land Use Random Forest (LURF) Models

The approach for building the LURF models were based on implementations taken previously in the literature for microarray data [34, 35, 36, 37, 38, 39, 40]. Like land use modeling of elemental concentrations, these type of studies utilize predictors (genes) that outnumber the sample size, are littered with noise, and are often highly correlated with one another. Because these types of problems play to the strengths of random forests, these models outperformed other gene mining techniques like regression and linear discriminant analysis. As in the LUR models, the best buffer radii for each variable category was determined based on the *model* R^2 value from a univariate regression with the elemental concentrations. In order to rank the importance of all the land use predictors, an initial random forest was trained using the best predictor from each category. The random forest variable importance measure (see

Section 3.1.1) was used to rank the land use predictors in order of decreasing importance and several random forests were built, each one by removing the least important predictor one at a time. The variable importance was used from the initial random forest and not recalculated at each step, as it has been shown that this can lead to severe overfitting [41]. The random forest picked as the final model was the random forest with the largest *pseudo* R^2, which was then optimized on m_{try} based on the *pseudo* R^2. See Section 3.1.1 for details on the random forest algorithm, m_{try}, and *pseudo* R^2.

2.2.5 Model Predictions

The final models were used to predict concentrations at the sampling sites using the land use predictor data. Since models were built using training data on the log scale, predictions were back transformed to their natural scale. Furthermore, some of the LUR models predicted concentrations much higher than the observed range, so the predictions for each LUR model were censored at a maximum limit of three times the highest observed concentration. For the 762 predictions in the CCAAPS cohort, this resulted in censoring for two predictions for Al and K; three predictions for Si, Zn, and PM2.5; four predictions for Mn and S; and 13 predictions for Ni. For the cross validation predictions, only one prediction from the LUR model for TRAP was censored. None of the CCAAPS or cross validation predictions for any of the elements from the LURF models exceeded three times the maximum observed value so these were not censored.

2.2.6 Cross Validated Model Accuracy

Cross validation of the accuracy of land use models is an important step because it quantifies the accuracy of the model when it is used to make predictions based on new observations. In the specific case of land use models, this will estimate the accuracy for when the model is used to predict elemental concentrations at new locations not included in the original sampling sites. Leave one out cross validation (LOOCV) was used with the predictor selection step

included as a part of the cross validation, so each fold of the LOOCV resulted in a different final model. Heatmaps were used to show the selection frequency of each land use variable for each element. The root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.1}$$

of all cross validation folds were calculated for each element, where y_i and $\hat{y}_i$ are the actual and predicted concentrations at each of the $n = 24$ total sites. Furthermore, the mean absolute prediction error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2.2}$$

was calculated for each cross validation fold as the absolute difference between the predicted and actual value divided by the actual value. The MAPE along with its 95% confidence interval was plotted for each element and model type.

2.2.7 Study Cohort

The Cincinnati Childhood Allergy and Air Pollution Study (CCAAPS) is an ongoing prospective birth cohort of high-risk atopic children [24, 25]. Children born between October 2001 and July 2003 in the Greater Cincinnati and Northern Kentucky region were screened by birth record and enrolled if they lived less than 400 meters or more than 1,500 meters from the nearest major road [25]. Furthermore, each study participant must have had at least one parent with symptoms of rhinitis, asthma, or eczema and allergic sensitization by a positive skin prick test result to one of 17 aeroallergens. Informed consent was obtained and the study was approved by the University of Cincinnati Institutional Review Board. Both types of land use models, LUR and LURF, were used to generate average exposure estimates for all elemental components of PM, TRAP, and total PM2.5 using each child's annual residential address history. CCAAPS exposure predictions for TRAP were also compared to the

original LUR model [3].

2.2.8 Computing

All statistical and geospatial computing was done in `R`, version 3.1.2 [42], using the `rgdal` [43], `rgeos` [44], and `sp` [45] packages. The code used to calculate the land use predictors and generate exposure estimates for each location has been made into an R package and is available online at `https://github.com/cole-brokamp/aiRpollution`. The code used to build and crossvalidate the LUR and LURF models is available on request.

2.3 Results

2.3.1 PM2.5 Elemental Measurements

Measurements from the 24 sites were collected and described as annual daily averages in ng/m^3. Figure 2.2 shows the concentrations as a boxplot on the log scale. In general, measurements had similar ranges to elemental PM2.5 components from previous studies [18, 20, 21, 22, 19]. More specific descriptive numbers along with the variance are listed in Table 2.1. Figure 2.3 illustrates the Spearman correlation matrix of all of the measured concentrations. All elements, including TRAP and PM2.5, were highly correlated with one another. However, K, Ni and S were less correlated compared to the rest of the elements.

2.3.2 Land Use Models

Both LUR and LURF models were built for total PM2.5 mass, TRAP, and eleven elemental components of PM2.5: Al, Cu, Fe, K, Mn, Ni, Pb, S, Si, V, and Zn. Land use variables as well as the varying buffer radii made available for selection by each model are listed in full in Table 2.2.

LUR The final LUR models for most pollutants (Table 2.3) resulted in a high fraction of explained variance, with Al, Cu, Fe, Mn, Pb, Si, Zn, and TRAP all having a *model* R^2 of at least 0.9. All other pollutants had a *model* R^2 of at least 0.75, except for Ni (*model* R^2 = 0.49). The most commonly selected land use predictor was the fraction of highly developed land from the National Landcover Database, utilized in the models for all pollutants except for S, Ni, and PM2.5. Transportation related variables also dominated the models, with truck traffic volume and length of roads or bus routes being the most common. Other than the use of the deprivation index for the Pb model, the use of elevation in the V model, and the use of PM2.5 point sources in the models for Cu, Mn, and PM2.5, only transportation and land use variables were selected. Of note, intersections, railroads, population, greenspace variables, and element specific point sources were not selected for any of the LUR pollutant models.

LURF The final LURF models (Table 2.4) also showed a generally high fraction of explained variance with Cu, Fe, Mn, Pb, Zn, and TRAP having a *model pseudo* R^2 of at least 0.8. All other models had a *model pseudo* R^2 of at least 0.5 except for K, Ni, and V. Although the *model* R^2 from the LUR model is not directly comparable to the *model pseudo* R^2 from the LURF model, it is interesting to note that Ni and K were two of the three worst performing pollutant models in both model types. In general, the LURF models utilized a higher number and more diverse selection of land use predictors for each pollutant than the LUR models. The fraction of highly developed land was still common, appearing in all models except for K, Ni, and V. However, other variables not included in the LUR models were commonly utilized in the LURF models. Examples include greenspace (used in all models except S, V, K, and TRAP), the deprivation index (used in models for Fe, Zn, Ni, Si, Pb, Al, and PM2.5), intersections (used in models for Zn and Pb), and railroads (used in models for Cu and V). Furthermore, the PM2.5 point sources were utilized more often (in models for Cu, Fe, Ni, K, and Mn) and the elemental specific point source variables were

utilized for Mn and Ni, but not for Pb or PM2.5. The optimization of m_{try} resulted in low values relative to the total number of variables in each final model, suggesting that this use of auxiliary noise in the random forest was useful in increasing the model accuracies.

2.3.3 Land Use Predictor Selection

Although Tables 2.3 and 2.4 describe the land use predictors selected for the final models, variable selection was a part of the model building process and thus was included in the cross validation assessment of accuracy. For each fold of the cross validation, different land use predictors were selected as the models were created and Figures 2.4 and 2.5 describe the frequency of selection of the land use variables across these folds for the LUR and LURF models, respectively. In the figures, a darker cell corresponds to a higher frequency of selection for a land use predictor in a model. Overall, it is clear that final group of land use predictors selected for each model varied depending on which fraction of training data was used, which means land use predictor selection is dependent on the sampling sites chosen to train the models.

With a few exceptions, the final LUR models (described in detail in Section 2.3.2) depended exclusively on land use and transportation related predictors. This is the case in the cross validation also, with highly developed landcover, length of bus routes, and length of roadways being selected very frequently (Figure 2.4). However, some of the models built during cross validation included predictors related to greenspace, deprivation, population, and elevation. Although these appeared less frequently, it is still important to note that they could have appeared in the final models if the elemental sampling stations were only slightly different. More specifically, the elements with the most variability with respect to the group of selected land use predictors were Al, S, Si, V, TRAP, and PM2.5. These were also the worst performing models, as assessed by cross validation (Table 2.5 and Figure 2.6).

The final LURF models (described in detail in Section 2.3.2) utilized a higher number and more diverse selection of land use predictors. Although more predictors were selected

compared to the LUR models, highly developed landcover, length of bus routes, and length of roadways again dominated the LURF models. However, in contrast to the LUR models greenspace was utilized heavily in the LURF models. Here, the models with the highest variability in the final selection of land use predictors (Cu, K, Ni, S, Si, PM2.5) were again associated with the highest MAPE.

2.3.4 Cross Validated Model Accuracy

Leave one out cross validation (LOOCV) was used to quantitatively compare the accuracy between the LUR and LURF models. Each site was left out once and the remaining 23 sites were used to create both a LUR and LURF model to predict the elemental PM2.5 concentrations. This was repeated for all 24 sites. The cross validated RMSE (Equation 2.1) and MAPE (Equation 2.2) for each element and model type are presented in Table 2.5. MAPE is equivalent to the difference between the actual and predicted concentration as a fraction of the actual concentration. This allows for comparison between models using different elements and also for a better sense of the magnitude of the error in terms of the inaccuracy rather than in terms of $(ng/m^3)^2$. The MAPE along with its 95% confidence interval is also plotted in Figure 2.6 for each model.

The LURF models for all elements except Cu, Fe, Ni, and Pb had a RMSE lower than that for the corresponding LUR models. When using MAPE rather than RMSE to evaluate the model accuracies, using the LURF model instead of the LUR model decreased the MAPE for all elements except for Fe, Mn, Ni. The difference in the MAPE of the model types for both the Fe and Mn models were less than 0.01 and the Ni model increased from 1.00 to 1.17. The biggest reduction in the MAPE when using the LURF model instead of the LUR model was seen for K, Si, V, Zn, TRAP, and PM2.5; all of which decreased by at least 0.1. The APE for these elements were also much more variable when using the LUR models as compared to the LURF models, seen in the confidence intervals in Figure 2.6. Figure 2.7 shows the individual predictions according the observed concentrations for each

fold of the cross validation according to model type. Here, it can be seen that the LUR models often make predictions that are extremely high compared to the actual value and the LURF predictions. It is these extreme errors that are likely driving the large variation in the mean of the APE. Although when averaging the APEs together, the decrease when using LURF compared to LUR is not as noticeable, the main advantage of the LURF is its lack of extremely high and erroneous predictions. Using RMSE to evaluate the model accuracy further masks this trend because the errors are in squared concentration units, whereas the MAPE gives a better description of the relative magnitude of the error.

2.3.5 CCAAPS Exposures

The final land use models (both LUR and LURF) were used to estimate elemental exposures for the CCAAPS cohort by using their primary residential birth record address. Figure 2.8 plots the exposure estimate for each model with the LUR predictions on the x-axis and the LURF predictions on the y-axis. For lower concentration predictions the LUR and LURF models generally agree. However, when predicting higher concentrations the LUR models are likely to predict much higher exposures than the LURF models. In this case, when the LUR predictions are much higher than the LURF over predictions, the LUR predictions are likely overestimating exposure, as seen in the model cross validation (Section 2.3.4).

Predictions from the older LUR model originally used to assess TRAP exposure [3] were also compared to the TRAP exposure predictions for CCAAPS using the newer LUR and LURF models (Figure 2.9). Here, it can be seen that the older LUR model predictions are more similar to the newer LURF model predictions than the newer LUR model predictions. Specifically, high exposures predicted by the older LUR model were not predicted as high using either of the new land use models; however, this disagreement seems to be much more severe with the newer LUR model. If the CCAAPS exposure assessments follow the same pattern as seen in the cross validation predictions, this suggests that the older LUR model did not suffer from same differential misclassification bias that the newer LUR model showed.

This is likely due to overfitting of the linear regression model because of the high number of available land use predictors.

2.4 Discussion

Here, we have successfully created LUR and LURF models for elemental components of PM. To our knowledge, this is the first study to do so for an American location. We have also shown that our novel land use models based on random forests are more accurate than LUR models for most of the elements. Furthermore, we identified a differential bias in exposure assessment using the LUR models which was not present using the LURF models. As assessed by LOOCV, the best performing models ($MAPE < 0.3$) were for Cu, Fe, K, Mn, Pb, Si, Zn, TRAP, and total PM2.5. Models for Al, S, and V performed moderately well ($MAPE < 0.5$), but the model for Ni performed the worst by far ($MAPE = 1.0$).

2.4.1 Previous Work

Two other LUR models have been developed for elemental components of PM [18, 19], both of which used regression based approaches. Specifically, the model created for Calgary, Alberta [19] used models specific to summer and winter seasons to predict elemental components of PM10. Their measured elemental concentrations were similar to ours and followed the same correlation patterns, with all elements except for S being highly correlated with one another. They found that industrial point sources explained the most variance in their models, followed by developed land use. Although our elemental LUR models did not incorporate any pollutant point source information, highly developed land use did explain the largest amount of variation in almost all of the models. They did include other potential predictors, like traffic volume, road density, housing, and population density, but, unlike our LUR models, these did not explain much variance in their final models. The authors found that 11 out of their 30 elemental models had a model R^2 of at least 0.7 for both seasons,

whereas ten of our eleven elemental models had a model R^2 greater than 0.7. The ESCAPE study [18] developed LUR models for elemental components of both PM10 and PM2.5 for twenty different areas of Europe. Again, the correlation patterns and concentrations of measured elements were similar to our results. The model R^2 for each element varied greatly across the locations, but on average, they found a model R^2 greater than 0.7 for two of eight total modeled elements. Similar to our results, they found the elements with the highest model R^2 to be Cu and Fe, and the element with the lowest model R^2 to be Ni. Neither of these two studies utilized a method other than regression, and neither studies reported the precision of their estimates, leaving uncertainty in the validity of the exposure assessment that might be used for epidemiological studies.

2.4.2 Model Accuracy

Overall, we found that models with the most variation in the set of final selected land use predictors had the worst performance. This is expected because a higher dependence of the final model structure on the training data means that the model is doing a poor job of capturing the variability and will perform with poor accuracy during cross validation.

The identification of relatively lower accuracy when predicting relatively high concentrations in the LUR models means that this misclassification is differential and could result in biased associations with health outcomes. This problem was not found in the LURF models and highlights the advantage in our novel model, which is not only an increased accuracy, but a decreased variance of the amount of prediction error. This problem was specifically seen in the TRAP model, where some sites were over predicted by at least ten fold by the LUR model, whereas the LURF models never over predicted concentrations by even one fold.

2.4.3 Model Precision

Another advantage of the LURF model over the LUR model is its increased precision; this is expanded upon by example in Section 3.1 but warrants further study because it will likely have large implications in the use of exposure assessment predictions for epidemiological studies. Indeed, most studies that utilize land use models ignore precision in exposure assessment [21, 46, 22] which likely leads to biased estimates when it comes to association with health effects. Incorporating uncertainty in exposure assessment during association with health effects would help to mitigate this problem and is a promising avenue for future research (see Chapter 5 for more details).

2.4.4 Conclusion

LURF will be a useful exposure assessment tool for epidemiological studies associating elemental components of PM with health effects. More generally, random forest and other machine learning methods should be incorporated into future land use models for more accurate and precise exposure assessment.

Tables

Table 2.1: Summary of measured elemental concentrations, TRAP, and total PM2.5 used to train the land use models. All elements contained complete measurements for all 24 sites, except for V, which had one site with a missing measurement. Concentration units are ng/m^3.

Element	Minimum	25th Percentile	Median	Mean	75th Percentile	Maximum	SD
Al	16.7	28.3	35.6	42.1	46.9	157.6	28.2
Cu	1.0	1.6	2.1	3.0	3.9	8.0	1.9
Fe	50.9	62.5	83.0	112.0	139.7	342.3	73.7
K	48.6	56.8	65.9	67.1	75.4	111.8	14.8
Mn	1.4	1.9	2.4	3.3	4.4	8.9	2.1
Ni	0.2	0.4	0.6	1.0	1.0	6.2	1.3
Pb	2.0	2.6	2.9	4.7	3.9	19.7	4.6
S	819	1,267	1,653	1,648	1,861	3,151	557
Si	55	83	102	122	124	484	86
V	0.2	0.3	0.4	0.4	0.5	1.4	0.3
Zn	8.5	11.4	14.6	28.7	21.5	156.9	38.6
TRAP	200	335	370	485	608	1,020	248
Total PM2.5	12,595	13,558	17,396	17,582	19,665	28,623	4,110

Table 2.2: Land use predictors considered for inclusion in final models.

Predictor	Units	Buffer radius in meters (intervals)
Transportation		
Distance to nearest Class 1 road	meters	n/a
Distance to nearest Class 2 road	meters	n/a
Distance to nearest Class 3 road	meters	n/a
Distance to nearest Class 4 road	meters	n/a
Distance to nearest Class 5 road	meters	n/a
Length of roads: Class 1	meters	100-1000 (50)
Length of roads: Class 2	meters	100-1000 (50)
Length of roads: Class 3	meters	100-1000 (50)
Length of roads: Class 4	meters	100-1000 (50)
Length of roads: Class 5	meters	100-1000 (50)
Average daily truck count on interstates	count	100-1000 (50)
Average daily truck count on highways	count	100-1000 (50)
Number of major intersections	count	50-1000 (50)
Distance to nearest railroad line	meters	n/a
Length of railroads	meters	100-1000 (50)
Length of bus routes	meters	100-1000 (50)
Physical Features		
Elevation	meters above sea level	n/a
Average elevation	meters above sea level	100-1000 (50)
Standard deviation of elevation	meters	100-1000 (50)
Fraction of elevation points $>$ 20 m uphill	count	100-1000 (50)
Fraction of elevation points $<$ 20 m downhill	count	100-1000 (50)
Community Characteristics		
Population count	count	n/a
Population density	count/meters^2	500-2500 (250)
Deprivation index	n/a	100-1000 (100)
Greenspace		
Average NDVI value	n/a	100-1000 (100)
Land Cover		
Open water	%	100-1500 (100)
Developed open	%	100-1500 (100)
Developed low	%	100-1500 (100)
Developed medium	%	100-1500 (100)
Developed high	%	100-1500 (100)
Barren	%	100-1500 (100)
Deciduous forest	%	100-1500 (100)
Evergreen forest	%	100-1500 (100)
Mixed forest	%	100-1500 (100)
Shrub	%	100-1500 (100)
Grassland	%	100-1500 (100)
Pasture	%	100-1500 (100)
Crops	%	100-1500 (100)
Woody wetlands	%	100-1500 (100)
Herbaceous wetlands	%	100-1500 (100)
NEI Point Sources*		
Distance to nearest point source	meters	n/a
Point source count	meters	1000-10000 (1000)
Point source total emissions	tons	1000-10000 (1000)
Point source average emissions	tons	1000-10000 (1000)
Point source emissions weighted by distance	tons/meters	1000-10000 (1000)

*PM2.5, PM10 (all models) and Ni, Pb, Mn (element specific models)

Table 2.3: Summaries of final LUR models for each element. Each *model* R^2 is from the final regression model and the model predictors are from the final models and denoted as in Table 2.2 with a buffer radius in meters, if applicable.

Element	Model R^2	Model Predictors
Al	0.90	Developed high (1200), Length of bus routes (100), Length of roads: Class 4 (1000), Pasture (1500), Distance to nearest Class 4 road, Length of roads: Class 2 (1000)
Cu	0.99	Developed high (1000), Shrub (1500), Distance to nearest PM2.5 point source, Average daily truck count on interstates (800), Length of roads: Class 3 (500), Developed medium (400), Distance to nearest Class 4 road, Evergreen forest (1100)
Fe	0.96	Developed high (1000), Average daily truck count on interstates (800), Length of roads: Class 2 (1000), Developed medium (400)
K	0.76	Developed high (1300), Shrub (700), Length of bus routes (150), Distance to nearest Class 2 road
Mn	0.92	Developed high (1000), PM2.5 point source count (2000)
Ni	0.49	Barren (1100), Mixed forest (1100)
Pb	0.98	Developed high (1500), Length of railroads (1000), Pasture (800), Deprivation index (700)
S	0.75	Average daily truck count on highways (350), Length of bus routes (350), Open water (900), Distance to nearest Class 2 road
Si	0.91	Developed high (1100), Length of bus routes (100), Crops (1500), Open water (1500)
V	0.80	Developed high (1500), Shrub (1300), Fraction of elevation points more than 20 meters uphill (150), Deciduous forest (1400)
Zn	0.94	Developed high (1500), Length of roads: class 3 (1000), Length of bus routes (850)
TRAP	0.91	Developed high (1000), Average daily truck count on interstates (800), Open water (1500), Length of class roads: Class 3 (900), Average daily truck count on highways (850)
PM25	0.81	Length of bus routes (350), Length of roads: Class 2 (950), Barren (1500), Pasture (500), Woody wetlands (1300), PM2.5 point source average emissions (1000)

Table 2.4: Summaries of final LURF models for each element. Each *model pseudo* R^2 is from the final random forest model and the model predictors are from the final models and denoted as in Table 2.2 with a buffer radius in meters, if applicable.

Element	Model Pseudo R^2	m_{try}	Model Predictors
Al	0.54	3	Developed high (1200), Deciduous forest (1000), Deprivation index (700), Population count, Developed low (800), Developed open (1100), Average NDVI value (1000), Developed medium (400), Length of bus routes (100)
Cu	0.83	2	Average NDVI value (1000), Developed high (1000), Distance to nearest PM10 point source, Distance to nearest railroad line
Fe	0.88	3	Average NDVI value (1000), Developed high (1000), Deprivation index (800), Developed medium (400), Developed open (1100), Distance to nearest PM10 point source, Average elevation (400), Average daily truck count on interstates (800)
K	0.40	6	Shrub (700), PM10 point source count (7000), PM25 point source average emissions (7000), PM25 point source count (7000), PM10 point source average emissions (7000), Length of roads: Class 3 (350), Pasture (500), Distance to nearest PM25 point source, Length of bus routes (150), Evergreen forest (600), Distance to nearest Class 2 roads
Mn	0.88	2	PM10 point source count (2000), Mn point source count (2000), PM25 point source count (2000), Distance to nearest PM2.5 point source, Distance to nearest PM10 point source, Average NDVI value (1000), Developed high (1000), Developed medium (1400), Population count
Ni	0.27	2	Distance to nearest Ni point source, Developed medium (1400), Deprivation index (1000), Distance to nearest PM10 point source (2000), Ni point source average emissions (2000), PM25 point source count (2000), Ni point source count (2000), Average NDVI value (700), Average elevation (600), Distance to nearest PM25 point source (2000), PM25 point source total emissions (2000), Ni point source total emissions (2000), Length of railroads (100), PM10 point source count (2000), Deciduous forest (1000), PM10 point source total emissions (2000), Fraction of elevation points more than 20 meters downhill (100), Distance to nearest Class 1 road, Distance to nearest Ni point source (2000), Average daily truck count on interstates (800), Length of roads: Class 3 (900), Mixed forest (1100), PM10 point source average emissions (2000), Grassland (1200)
Pb	0.89	5	Average NDVI value (1000), Pasture (800), Developed open (1100), Developed medium (400), Length of bus routes (900), Deprivation index (700), Developed low (900), Population density (500), Number of major intersections (1000), Developed high (1500), Length of roads: Class 4 (1000), Average daily truck count on interstates (300)
S	0.51	2	Developed high (1500), Average daily truck count on highways (350)
Si	0.59	2	Developed high (1100), Average NDVI value (1000), Deciduous forest (900), Deprivation index (800), Developed low (800), Developed open (1100), Developed medium (400), Length of bus routes (100), Elevation
V	0.47	1	Deciduous forest (1400), Distance to nearest railroad line, PM10 point source count (6000)
Zn	0.83	3	Developed medium (400), Deprivation index (1000), Developed high (1500), Number of major intersections (1000), Length of bus routes (850), Average NDVI value (1000), Length of roads: Class 4 (1000), Developed low (900), Average daily truck count on interstates (300), Population density (500), Length of roads: Class 1 (1000), Developed open (1500), Deciduous forest (1500)
TRAP	0.80	1	Average elevation (100), Developed high (1000), Average daily truck count on interstates (800)
PM25	0.53	1	Average daily truck count on interstates (300), Deprivation index (500), Length of bus routes (350), Developed high (1500), Average daily truck count on highways (350), Herbaceous wetlands (1500), Length of roads: Class 1 (1000), Barren (1500), Standard deviation of elevation (1000), Average NDVI value (1000), Shrub (600), Length of roads: Class 2 (950), Length of railroads (150)

Table 2.5: Cross validated root mean squared error (RMSE) and mean absolute predictive error (MAPE) of LUR and LURF elemental PM models. RMSE units are ng/m^3 and MAPE is expressed as a ratio.

Element	RMSE (LUR)	RMSE (LURF)	MAPE (LUR)	MAPE (LURF)
Al	27.11	25.08	0.35	0.32
Cu	0.85	1.11	0.24	0.23
Fe	40.80	46.36	0.19	0.21
K	26.97	15.48	0.29	0.17
Mn	1.17	1.11	0.19	0.20
Ni	1.18	1.29	1.00	1.17
Pb	1.85	2.63	0.17	0.17
S	744.70	647.42	0.38	0.37
Si	102.74	80.67	0.41	0.22
V	0.53	0.24	0.67	0.45
Zn	44.48	27.74	0.32	0.24
TRAP	563.93	150.04	0.52	0.21
PM25	9160.99	3863.34	0.34	0.20

Figures

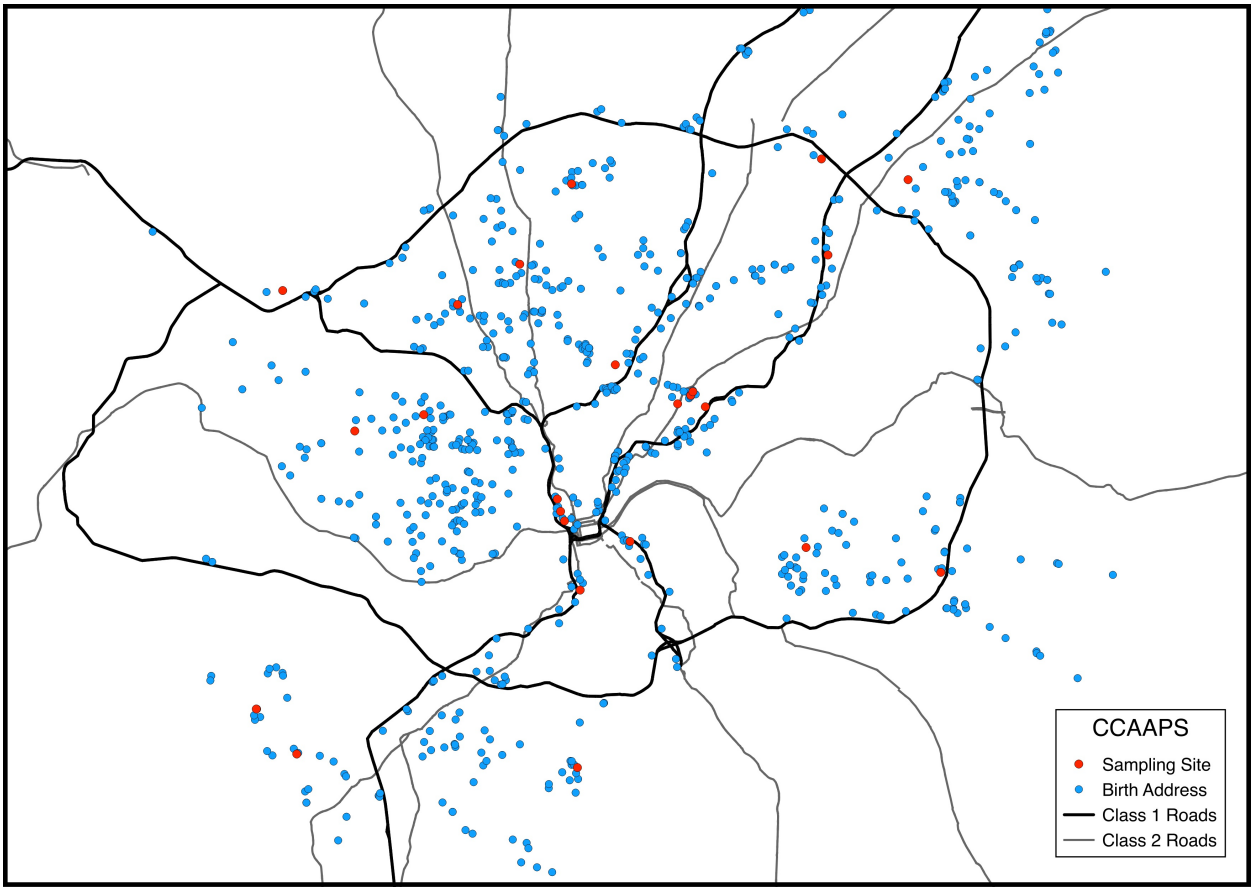

Figure 2.1: The location of the CCAAPS sampling sites in red and the birth addresses of the CCAAPS cohort in black.

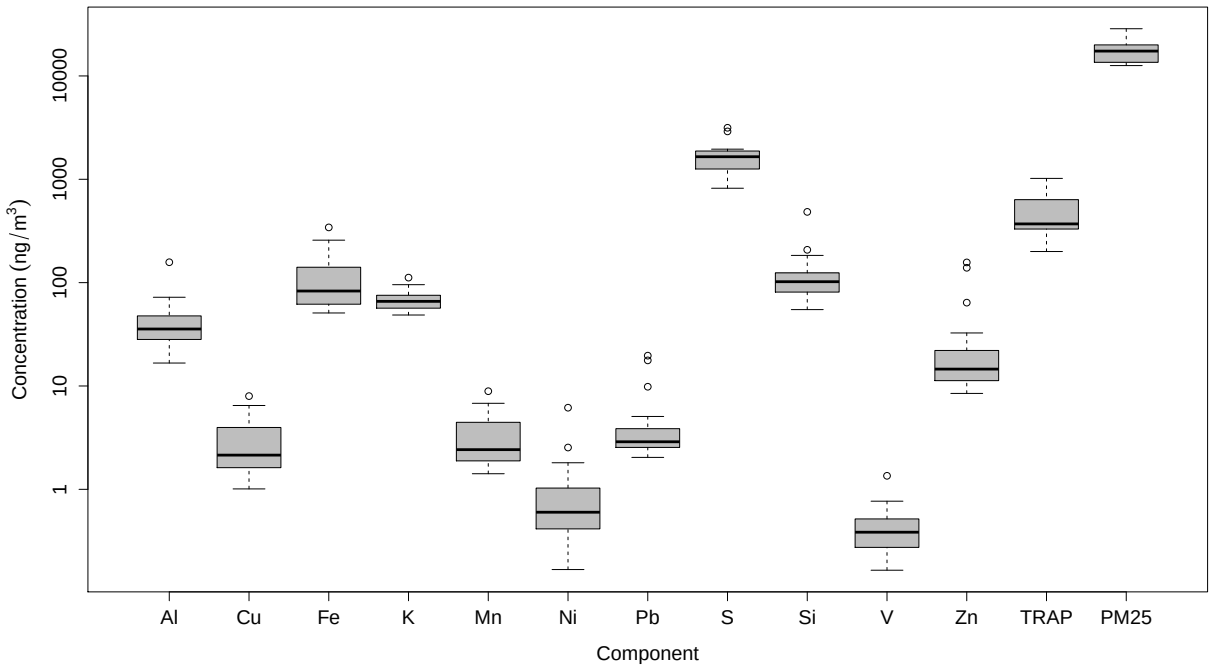

Figure 2.2: Box plot of measured elemental concentrations, TRAP, and total PM2.5 used to train the land use models.

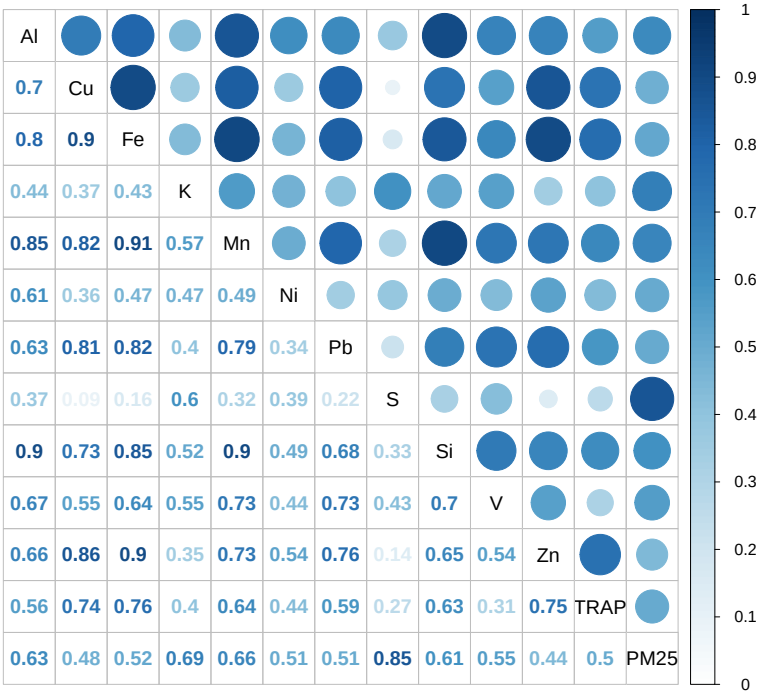

Figure 2.3: Spearman correlation matrix of measured elemental concentrations, TRAP, and total PM2.5. A darker blue and larger circle in the upper triangle of the grid corresponds to a larger Spearman's rho statistic shown in the lower triangle of the grid.

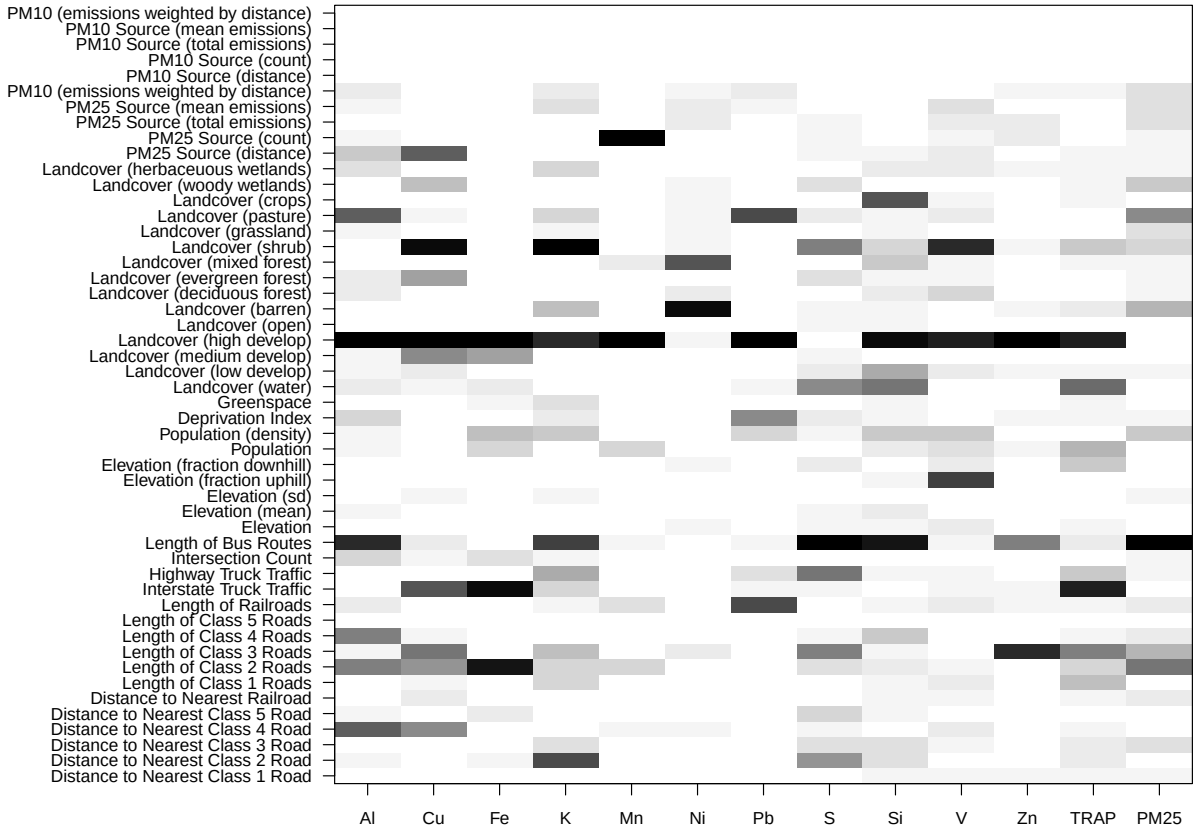

Figure 2.4: Selection frequency of land use predictors for all LUR models. A darker cell means that the land use predictor was selected more often for use in the final model across all folds of the cross validation.

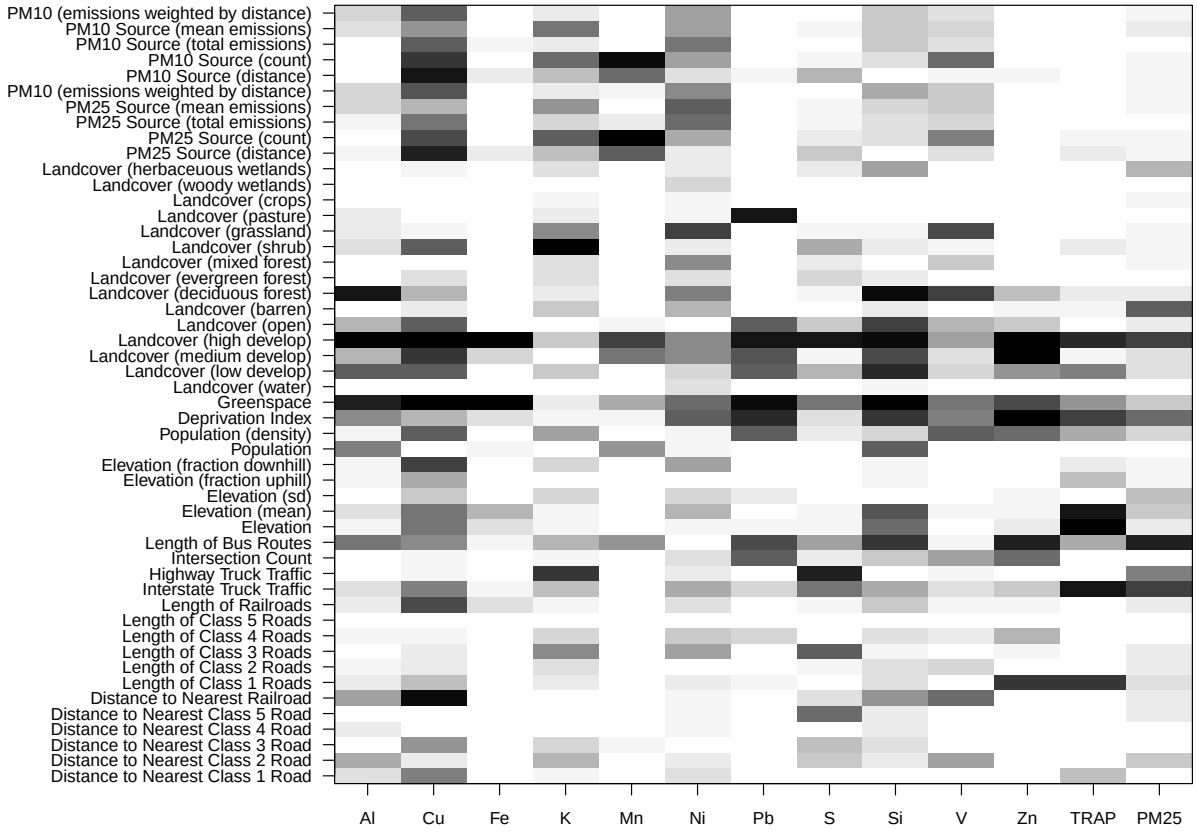

Figure 2.5: Selection frequency of land use predictors for all LURF models. A darker cell means that the land use predictor was selected more often for use in the final model across all folds of the cross validation.

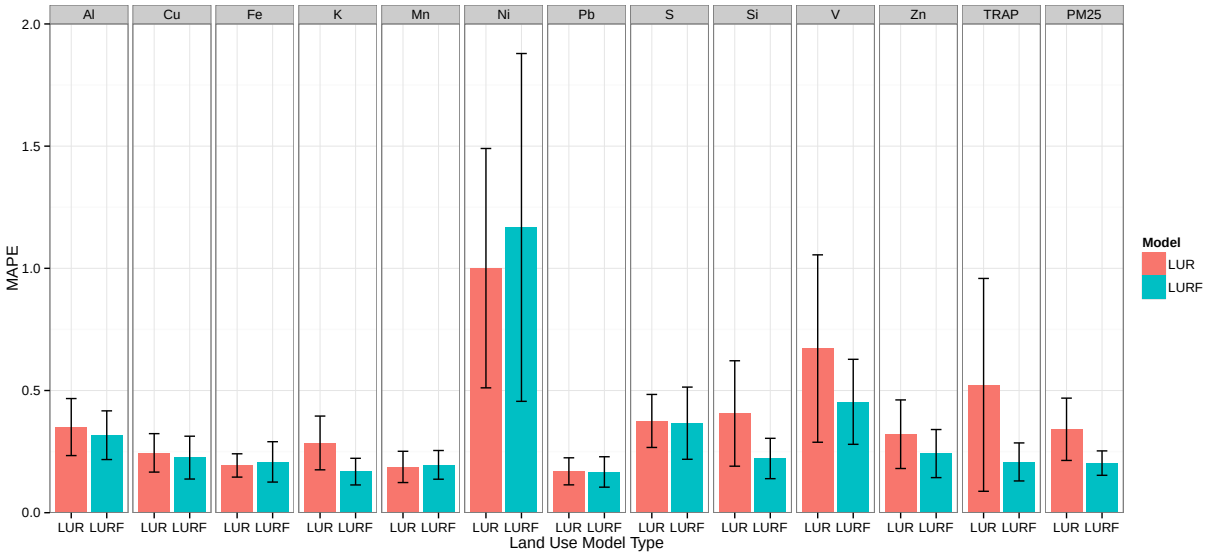

Figure 2.6: Cross validated absolute predictive error and 95% confidence interval for each elemental model, each built both using a LUR model and a LURF model.

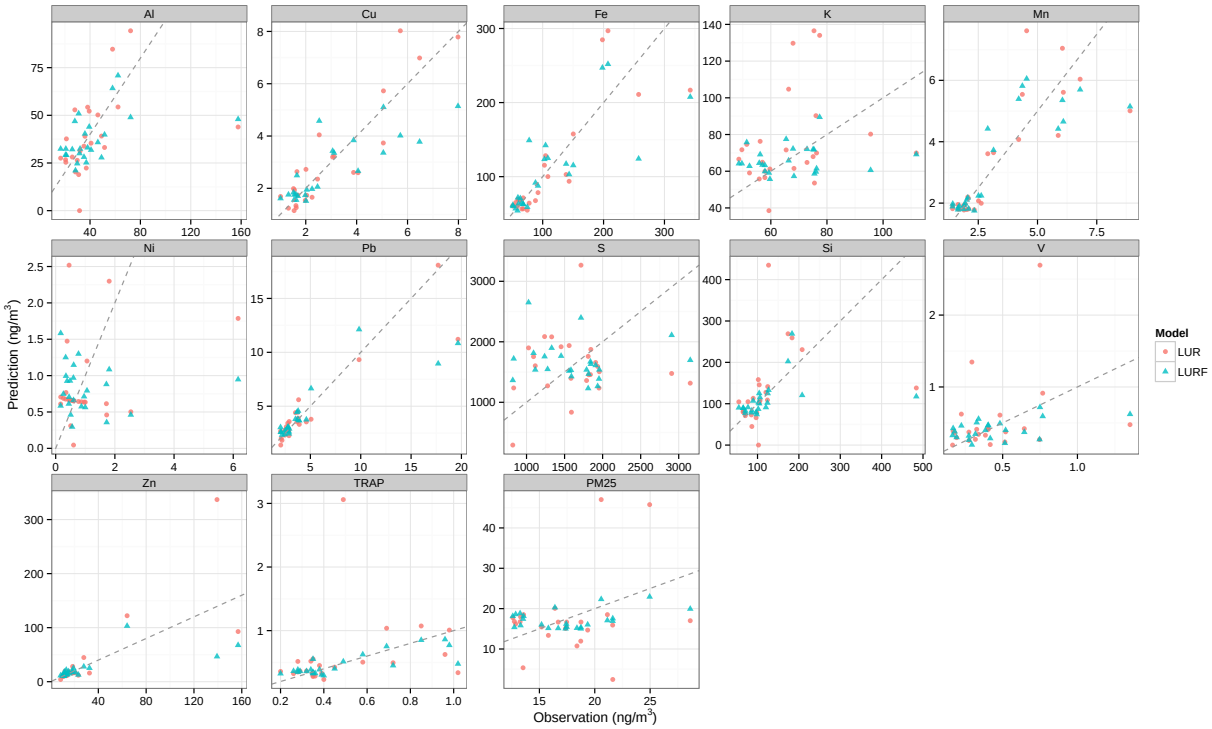

Figure 2.7: LOOCV predictions from the LURF and LUR land use models according to the true observed values. The dotted line represents the perfect agreement between observed and predicted concentrations.

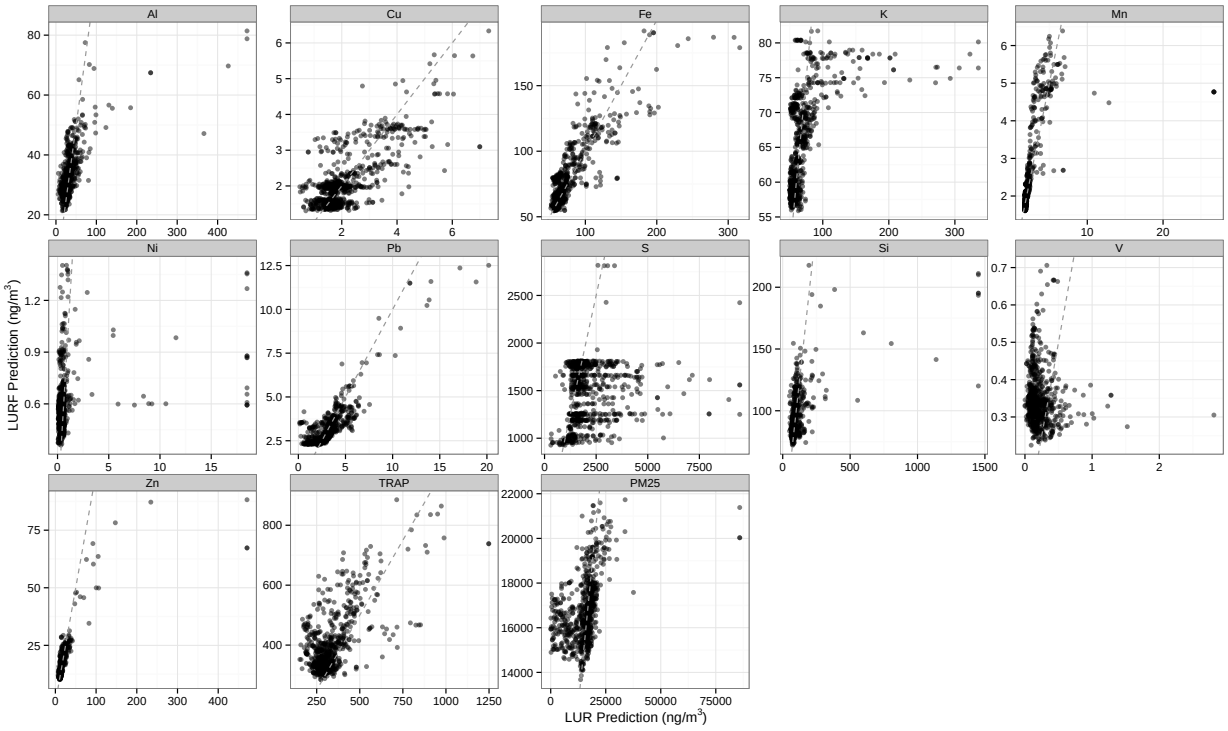

Figure 2.8: Agreement of the LURF and LUR exposure predictions for the 762 children in the CCAAPS cohort. The dotted line represents the perfect agreement between RF and LM predictions.

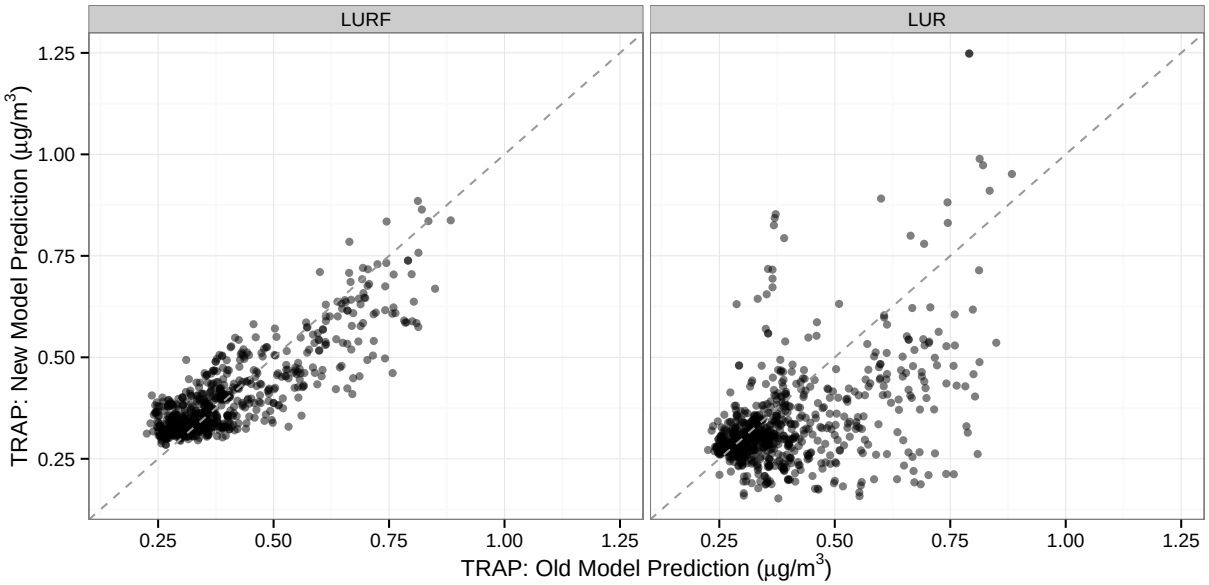

Figure 2.9: Agreement of the old land use model [3] TRAP concentration predictions and the new land use model TRAP concentration predictions, both LURF and LUR models, for the 762 children in the CCAAPS cohort. The dotted line represents the perfect agreement between predictions.

Example Land Use Predictor Figures

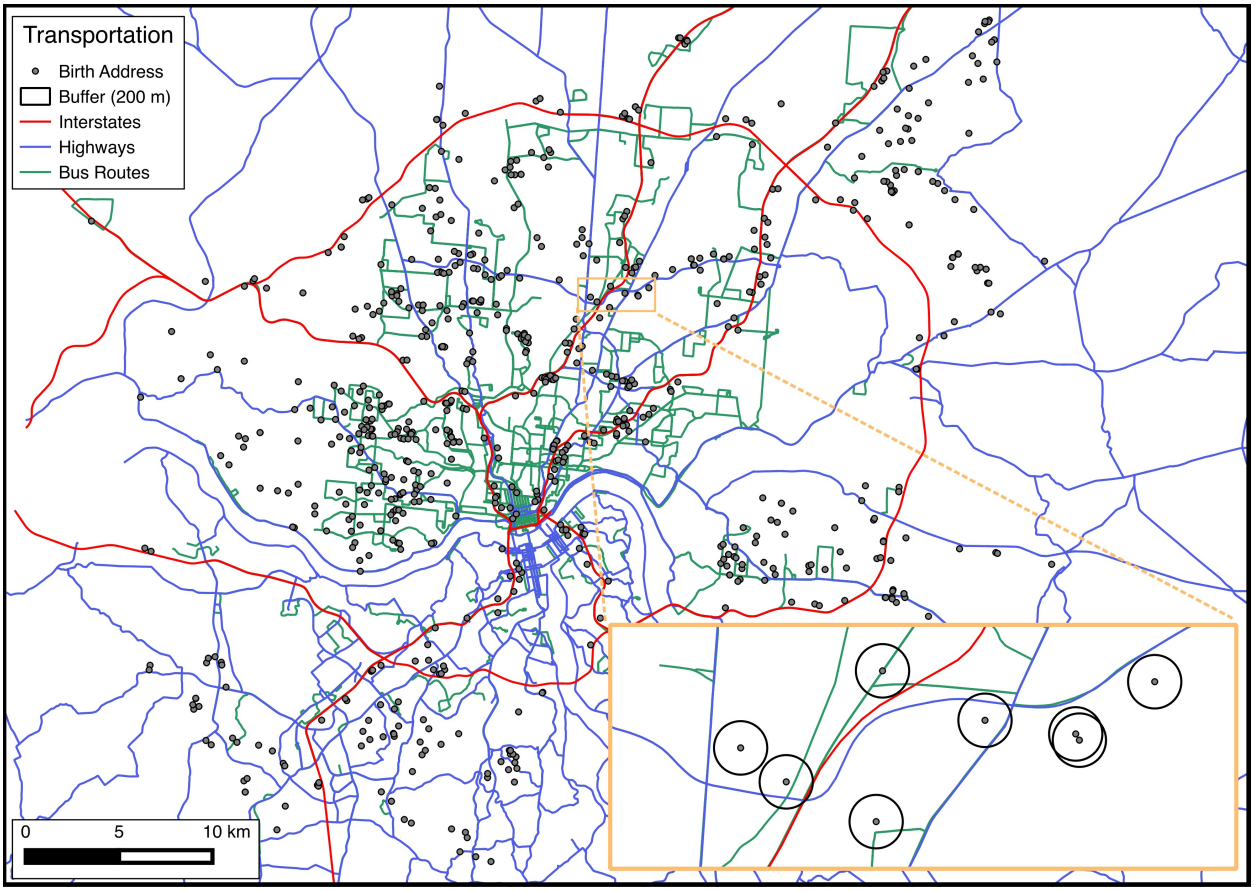

Figure 2.10: Location of CCAAPS birth record addresses, interstates, highways, and bus routes. The inset shows a magnified view of several locations with accompanying 200 meter fixed buffers.

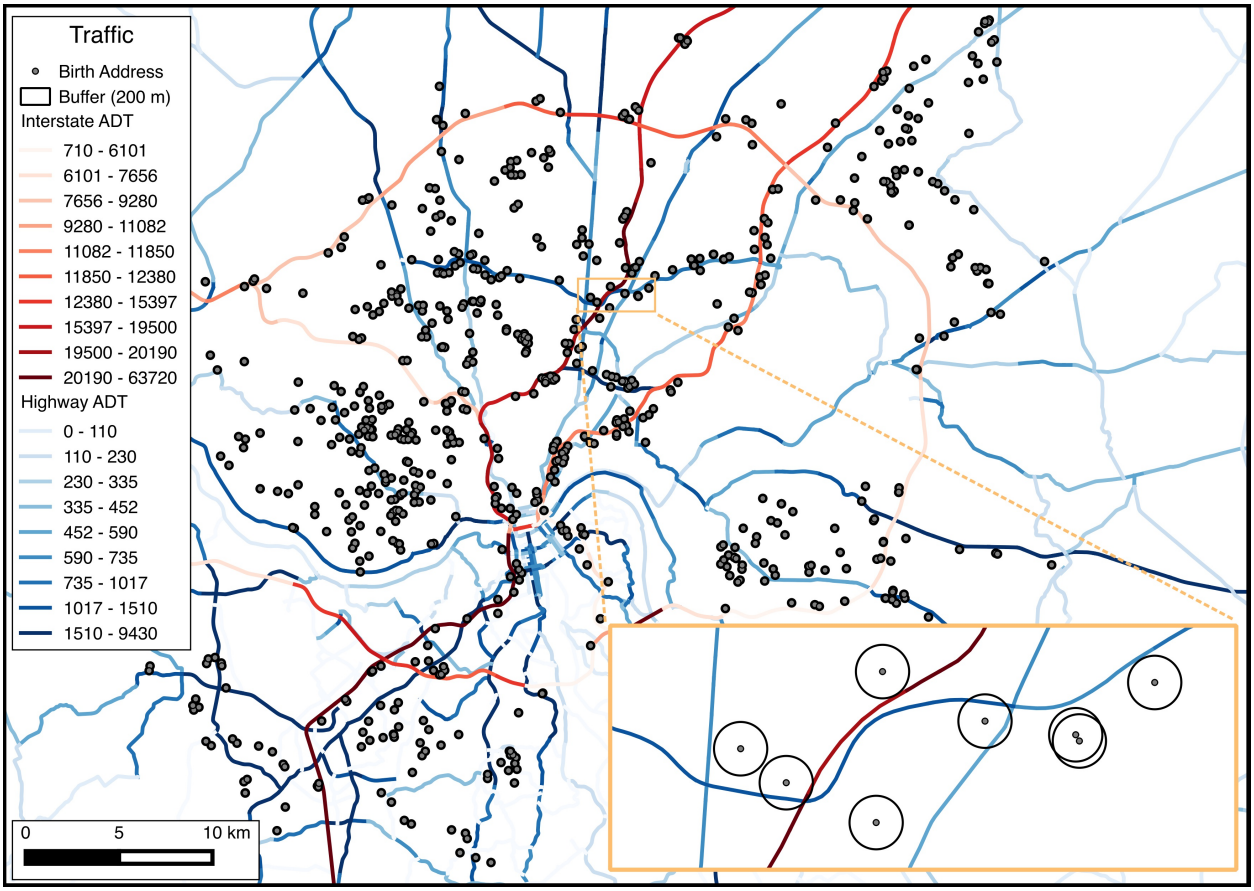

Figure 2.11: Location of CCAAPS birth record addresses and the intensity of traffic on interstates and highways as average daily truck (ADT) count. The inset shows a magnified view of several locations with accompanying 200 meter fixed buffers.

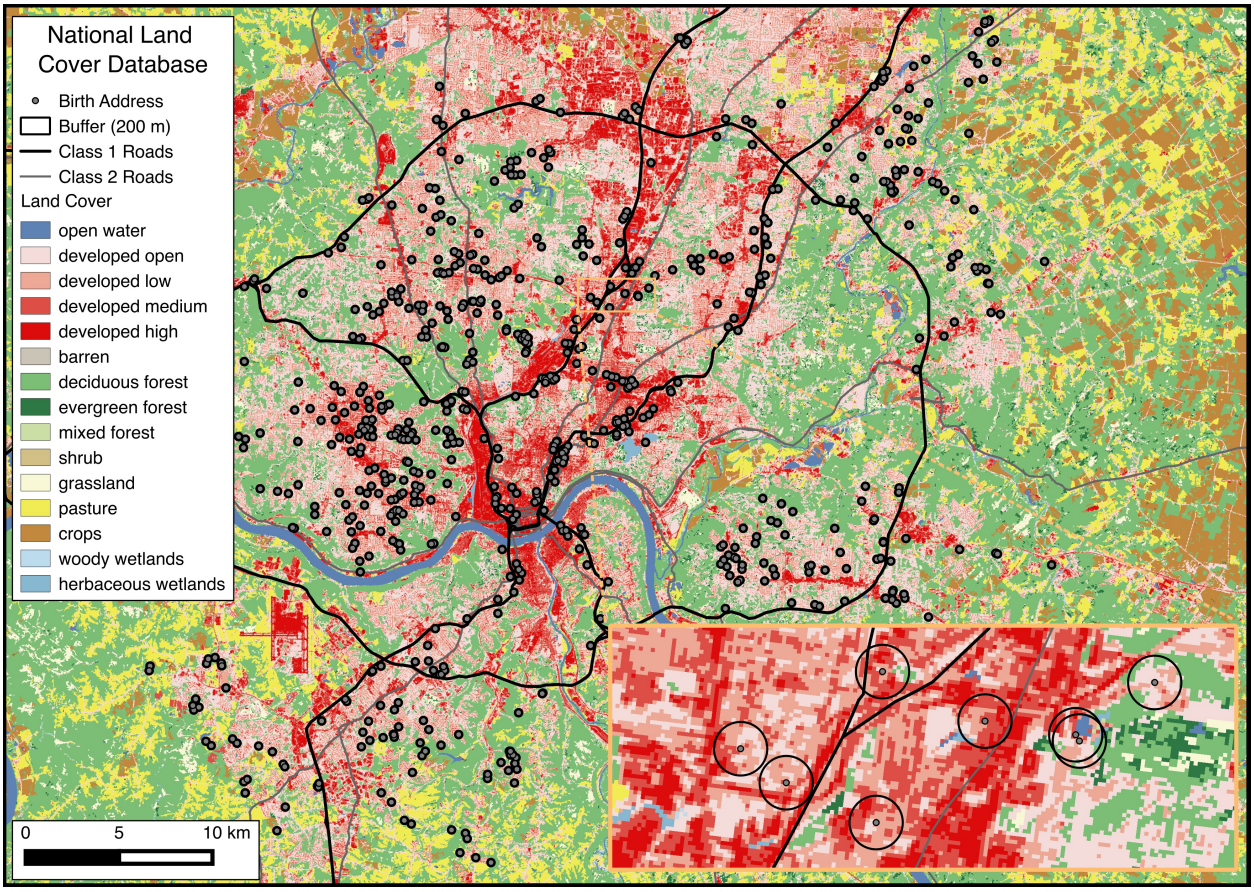

Figure 2.12: Location of CCAAPS birth record addresses and the land use classification for the study area. The inset shows a magnified view of several locations with accompanying 200 meter fixed buffers.

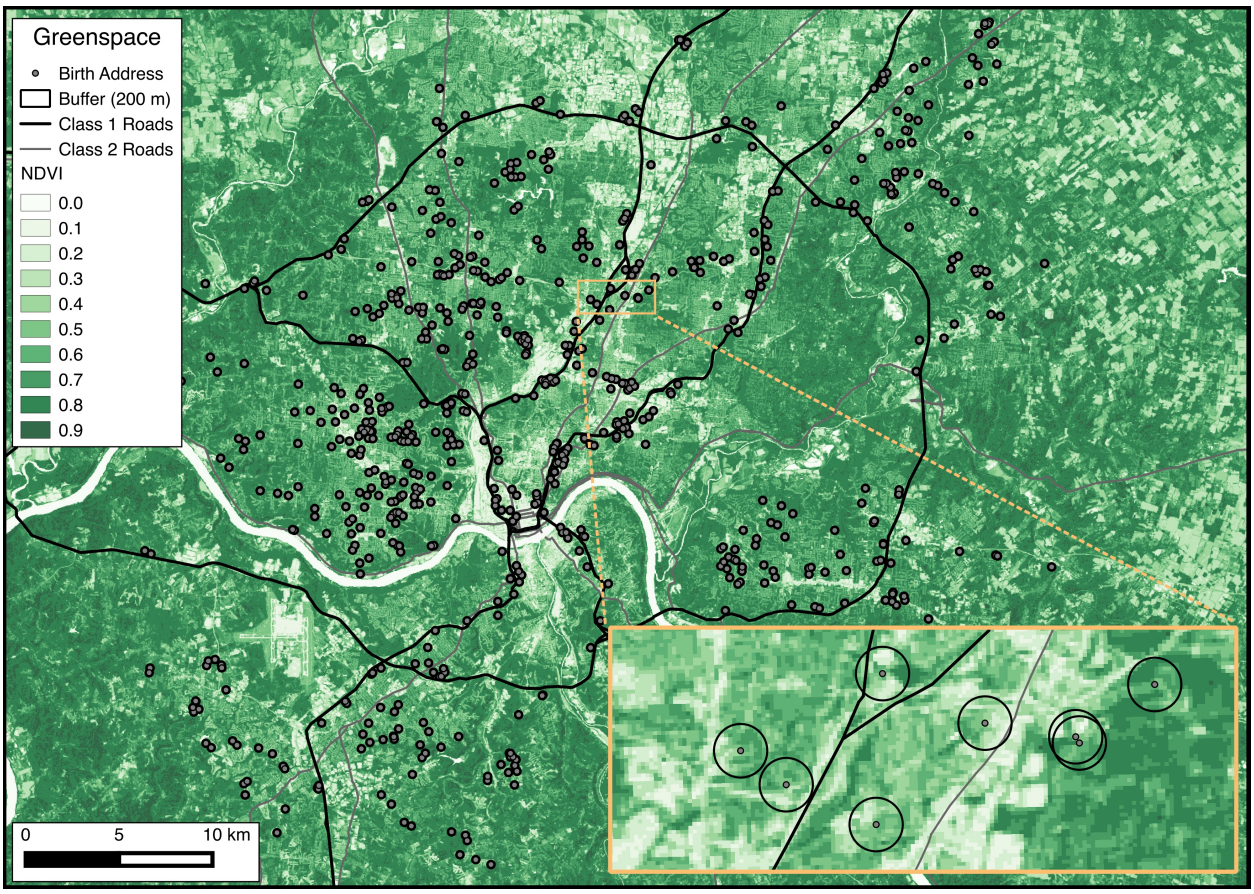

Figure 2.13: Location of CCAAPS birth record and NDVI values. The inset shows a magnified view of several locations with accompanying 200 meter fixed buffers.

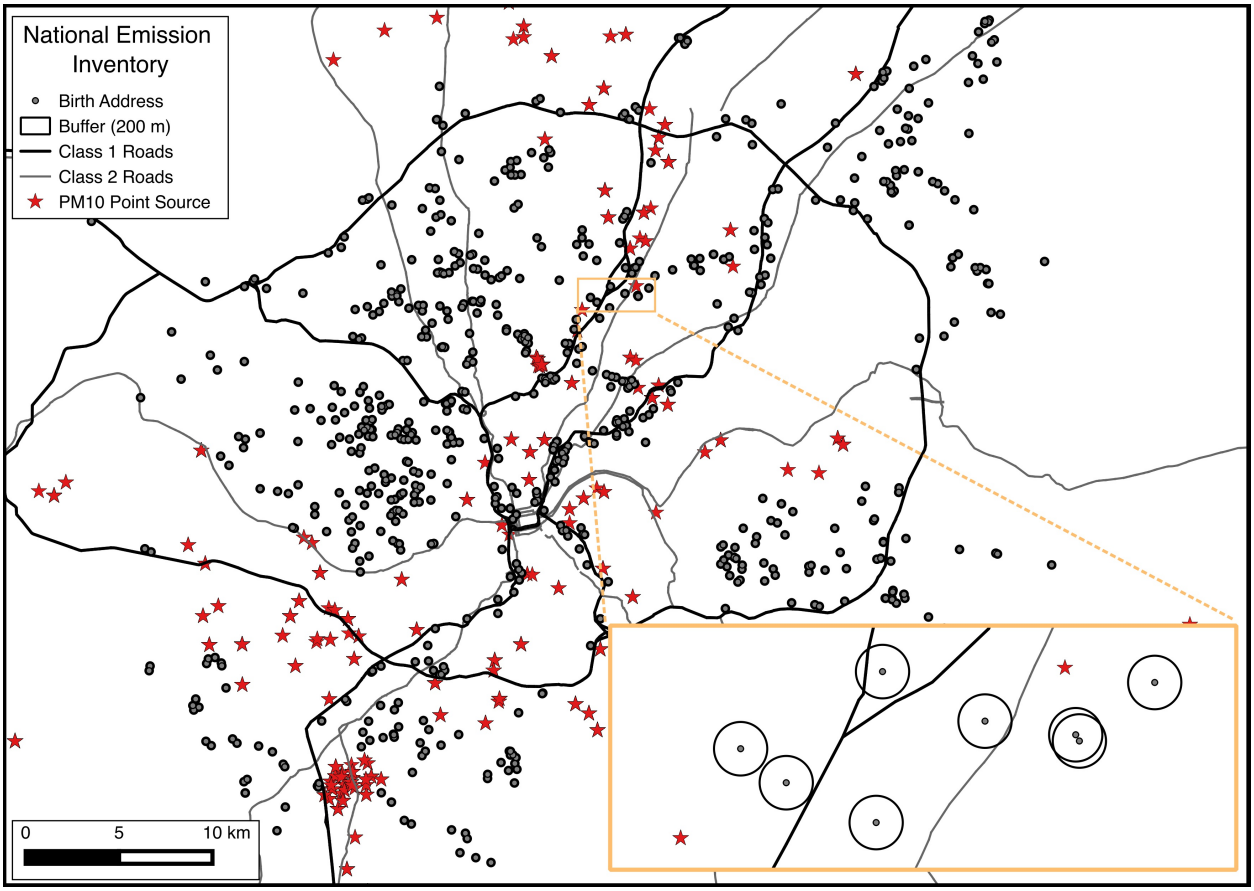

Figure 2.14: Location of CCAAPS birth record addresses and PM10 point source from the National Emission Inventory. The inset shows a magnified view of several locations with accompanying 200 meter fixed buffers.

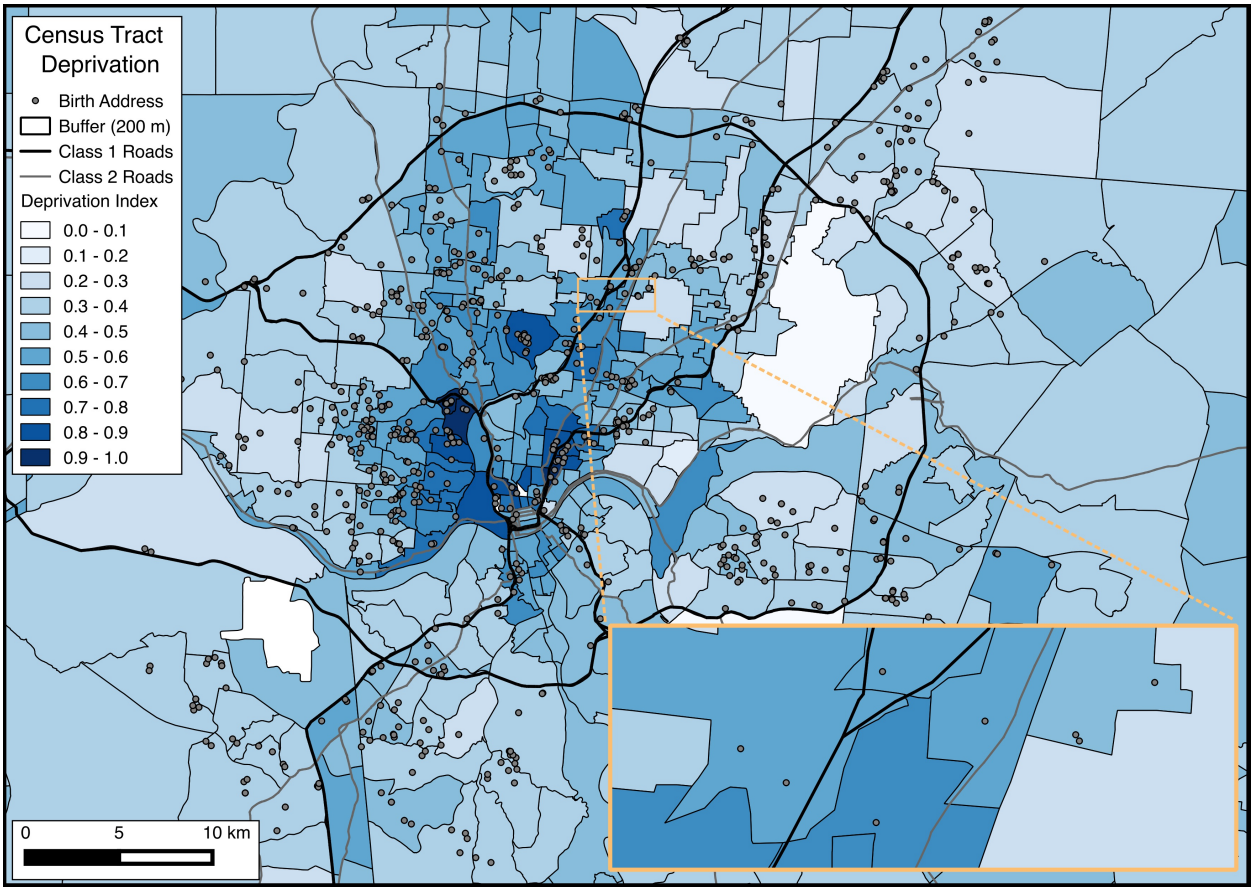

Figure 2.15: Location of CCAAPS birth record addresses and the deprivation index for each census tract. The inset shows that the deprivation value was extracted based on the census tract of each location.

Chapter 3

Estimating the Variance of Random Forest Predictions

A Comparison of Resampling and Recursive Partitioning Methods in Random Forest for Estimating the Asymptotic Variance Using the Infinitesimal Jackknife

Cole Brokamp[1], MB Rao[1], Patrick Ryan[1,2], Roman Jandarov[1]

[1]Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA

[2]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Corresponding Author:

Cole Brokamp

University of Cincinnati

Kettering Lab Building, 160 Panzeca Way, Cincinnati, OH 45267-0056

E-mail: brokamrc@mail.uc.edu

The authors declare no competing financial interests.

Abstract

Background: The infinitesimal jackknife (IJ) has recently been applied to the random forest to estimate the variance of their predictions. Additionally, the predictions have been shown to be asymptotically normal under a few conditions. These theorems were verified under a traditional random forest framework which uses CART trees and bootstrap resampling. However, random forests using conditional inference (CI) trees and subsampling have been found to be more accurate and not prone to variable selection bias. Whether or not the IJ will hold under these random forest variations is unknown. Here, we conduct simulation experiments to explore the applicability of the IJ to random forests using variations on the resampling method and base learner.

Methods: Data were simulated using eleven different functions and three different sample sizes. One hundred test points were generated and each trained on one hundred data sets using random forests with combinations of CART trees or CI trees, bootstrap sampling or subsampling, and three different values of m_{try}. The mean absolute bias for all variations were calculated as the absolute difference between the empirical variance of predictions and the IJ variance prediction with a suggested bootstrap correction.

Results: Using CI trees instead of traditional CART trees and using subsampling instead of bootstrap sampling resulted in a much more accurate estimation of prediction variance when using the IJ. The effect of m_{try} and sample size differed with respect to the empirical variance and individual data simulation distributions. The random forest variations here have been incorporated into an open source software package for the R programming language. We present an applied example of the IJ using the package on a data set of environmental ozone measurements. Our results suggest that random forests should be constructed using conditional inference trees instead of CART trees and subsampling instead of bootstrap sampling for both increased accuracy in predictions and increased accuracy in

variance predictions with the IJ.

3.1 Introduction

Although random forests have been proven to be more accurate than common parametric techniques like regression [10], they still remain underused because most researchers use them only for prediction and not for interpretation. Utilizing the infinitesimal jackknife (IJ) to estimate the variance of random forest's predictions will help to generate insight into random forests. The characterization of the statistical distribution of bagged tree ensemble predictions brings these type of learners out of the "black box" and into the realm of statistical inference. This will allow researchers to answer questions like "How much would predictions change if a different data set was used to train it?" and "Which predictions are the random forest more confident about?".

As an example, Figure 3.1 shows what this would look like for usage with the land use models (Chapter 2). In the figure, predictions for total PM2.5 from the land use random forest (LURF) models along with 95% confidence intervals, calculated using the IJ, are plotted against the actual total PM2.5 concentrations at all 24 CCAAPS sampling sites. For comparison, the same predictions and accompanying 95% confidence intervals from the land use regression (LUR) models are also plotted. Here, it can be seen that the random forest is less confident about its inaccurate predictions and that the LURF confidence intervals overall are much smaller than those produced by the LUR models. Although this is only a brief demonstration of plotting predictions, the plotting of random forest predictions confidence intervals with respect to variation in predictor variables could allow for elucidation of the complex interactions and relationships that the current implementation of random forests are so good at detecting but not so good at explaining.

Although Wager et al. have proven that the IJ can be used to estimate the prediction variances of traditional random forests [47, 48], their methods have not been tested using alternative individual tree types used in random forest, such as conditional inference trees. This variation is widely used and is important for eliminating variable selection bias and increasing the predictive accuracy of traditional random forests. Because their formulas are

proven under the assumption of traditional random forests, they are hard to verify under other random forest variations and whether or not their proof will hold is unknown. Here, we explore the applicability of the IJ to random forest variations, including resampling method and individual tree type, and compare the accuracy of their estimates of prediction variances. Furthermore, we implement them as an open-source software tool that can be an invaluable statistical tool used in almost any supervised statistical learning situation.

3.1.1 Random Forests

Random Forests are often implemented in prediction analyses because of their increased accuracy and resistance to multi-collinearity and complex interaction problems as compared to linear regression [10]. In a recent study, random forest was found to be the most accurate classification algorithm among 179 classifiers, based on 121 different data sets [12]. The technique itself is an ensemble learning method that builds on bagging – specifically the bootstrapped aggregation of several regression trees – to predict an outcome. Bagging is most often used to reduce the variance of an estimated prediction function and is most useful for models which are unbiased but have a high variance, like regression trees [10]. Random forests, first proposed by Breiman [13], modify the bagging technique by ensuring that the individual trees are de-correlated by using a bootstrap sample for each tree and also randomly selecting a subset of predictors for testing at each split point in each tree. The random forest comes with the advantages of tree-based methods, namely the ability to capture complex interactions and maintain low bias, while at the same time alleviating the problem of high variance of predictions usually associated with tree-based methods by growing the individual trees to a very deep level (usually one observation per terminal node) and averaging their predictions.

Algorithm

The specific algorithm for random forest as used for regression is as follows:

1. For $b = 1$ to B total trees:

    - Draw a bootstrap sample from the training data
    - Grow tree T_b by repeating the following steps for each terminal node of the tree until the desired node size is reached:
        - Randomly select m_{try} of the total p variables
        - Pick the best variable and split-point from the m_{try} variables based on the best reduction in the sum of the squared errors of the predictions
        - Split the node into two daughter nodes

2. Output the total ensemble of all trees.
3. To predict at a new point x, average the prediction of all trees: $\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Tuning Parameters

The performance of random forests can be tuned using two parameters, m_{try} and B. B is the total number of trees and is set to 500 by default in the `randomForest` package within `R`. The number of trees should be large enough so that the error rate is stabilized. Since the random forest is grown one tree at a time, the error rate can be plotted as a function of the number of trees to visually ensure that enough trees are being used. m_{try} usually has more effect on the ensemble accuracy and is set to $max\{floor(\frac{1}{3}p), 1\}$ as the default in the `randomForest` package within `R`. Variations in m_{try} can be auditioned and the value producing the lowest error can be used in the final random forest model.

Accuracy Metric

The most commonly used metric to measure the accuracy of random forests in regression settings is *pseudo* R^2. It is calculated as $1 - \frac{MSE}{var(Y)}$ where Y is a vector of the outcomes and MSE is the mean of the squared errors for all prediction points. The error for each prediction point is calculated based on predictions using trees where the sample point was

not used to train the tree. This generally represents the amount of variance explained by the random forest model and is analogous to the coefficient of determination, here denoted as *model* R^2, in ordinary least squares regression. However, since this accuracy metric is based only on data not used to train the model, it is a conservative estimate and tends to not overestimate the true cross validated model accuracy, such as often occurs when using the *model* R^2 in regression settings. Note also that the *pseudo* R^2 can be negative because it is possible for MSE to be greater than $var(Y)$.

Variable Importance

A metric used to determine the influence of each predictor on the accuracy of the random forest is called "variable importance". In a regression setting, this is calculated separately for each variable. After fitting each tree in the random forest, the prediction MSE for samples not included in the training set of that tree, called "out of bag" (OOB) samples, is calculated. This is called the OOB error rate. The values of the predictor variable under focus are randomly permuted and the tree is fit again and the OOB error rate is calculated again. The difference in OOB error rates are averaged over all trees which utilize that predictor in the ensemble and then normalized by the standard deviation of the differences. This allows for comparison of the variable importance measure between variables with different ranges.

3.1.2 Estimating the Variance of Bagged Tree Predictions

Bootstrap sampling, subsampling, and the jackknife all rely on estimating the variance of a statistic by using the variability between resamples rather than using statistical distributions. This section will cover the jackknife and how it is applied to the resampling distribution to generate variance estimates for random forest predictions.

Ordinary Jackknife

The ordinary jackknife is a resampling method useful for estimating the variance or bias of a statistic. The jackknife estimate of a statistic can be found be by repeatedly calculating the statistic, each time leaving one observation from the sample out and averaging all estimates. The variance of the estimate can be found by calculating the variance of the jackknifed estimates:

$$\hat{V}_J = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)} \right)^2 \tag{3.1}$$

where n is the total sample size, $\hat{\theta}_{(-i)}$ is the statistic estimated without using the i^{th} observation, and $\hat{\theta}_{(\cdot)}$ is the average of all jackknife estimates.

Jackknife After Bootstrap

The ordinary jackknife is extended for use with bagging by applying it to the bootstrap distribution [49]. Instead of leaving out one observation at a time, the existing bootstrap samples are used and the statistic is calculated based on all resamples which do not use the i^{th} observation:

$$\hat{V}_{JB} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\bar{t^*}_{(-i)}(x) - \bar{t^*}_{(\cdot)}(x) \right)^2 \tag{3.2}$$

where $\bar{t^*}_{(-i)}(x)$ is the average of $t^*(x)$ over all bootstrap samples not containing the i^{th} example and $\bar{t^*}_{(\cdot)}(x)$ is the mean of all $\bar{t^*}_{(i)}(x)$.

Infinitesimal Jackknife and Resampling

As opposed to $\hat{V}_J$ and $\hat{V}_{JB}$, where the behavior of a statistic is studied after removing one or more observations at a time, the IJ looks at the behavior of a statistic after down-weighting each observation by an infinitesimal amount. [50]. Adapted from Efron [49], the following is a gentle introduction to the idea.

Define

$$t_i^* = t(\mathbf{y}_i^*) \qquad \mathbf{y}_i^* = (y_{i1}^*, y_{i2}^*, ..., y_{ik}^*, ..., y_{in}^*) \tag{3.3}$$

as the i^{th} calculation of a statistic based on the i^{th} bootstrap sample, and

$$Y_{ij}^* = \#\{y_{ik}^* = y_j\} \tag{3.4}$$

as the number of times that the original data point y_j appears in the i^{th} bootstrap sample $\mathbf{y}_i^*$. Then, the count vector $\mathbf{Y}_i^* = (Y_{i1}^*, Y_{i2}^*, ..., Y_{ik}^*, ..., Y_{in}^*)$ forms a multinomial distribution with n draws on n categories, each having a probability of $1/n$. Using the mean and variance of the multinomial distribution, we can say that

$$\mathbf{Y}_i^* \sim (\mathbf{1}_n, \mathbf{I} - \mathbf{1}_n\mathbf{1}_n^T/n). \tag{3.5}$$

By fixing the original data and writing the bootstrap replication statistic as function of the count vector, we can define the ideal smoothed bootstrap estimate S_0 as the multinomial expectation of $T(\mathbf{Y}^*)$.

$$S_0 = E[T(\mathbf{Y}^*)], \quad \mathbf{Y}^* \sim \text{Mult}_n(n, \mathbf{p}_0), \tag{3.6}$$

with $\mathbf{p}_0 = (1/n, 1/n, ..., 1/n)$. Extending this to a probability vector of $\mathbf{p} = (p_1, p_2, ..., p_n)$ leads to

$$S(\mathbf{p}) = E[T(\mathbf{Y}^*)], \quad \mathbf{Y}^* \sim \text{Mult}_n(n, \mathbf{p}), \tag{3.7}$$

Using the delta method, we can define the directional derivative as

$$\dot{S}_j = \lim_{\epsilon \to 0} \frac{S(\mathbf{p}_0 + \epsilon(\boldsymbol{\delta}_j - \mathbf{p}_0)) - S(\mathbf{p}_0)}{\epsilon} \tag{3.8}$$

where $\boldsymbol{\delta}_j$ is the j^{th} coordinate vector with all zeros except for a one in the j^{th} place. We can

then use the delta method to estimate the standard deviation of s_0

$$\frac{\left(\sum_{j=1}^{n} \dot{S}_j^2\right)^{1/2}}{n} \tag{3.9}$$

In order to reduce this back into terms of $\mathbf{Y}^*$, we define $w_i(\mathbf{p})$ as the probabilities of $\mathbf{Y}^*$ in Equation 3.7 divided by the probabilities of $\mathbf{Y}^*$ in Equation 3.6,

$$w_i(\mathbf{p}) = \prod_{k=1}^{n} (np_k)^{Y_{ik}^*}, \tag{3.10}$$

such that

$$S(\mathbf{p}) = \sum_{i=1}^{B} w_i(\mathbf{p}) t_i^* / B \tag{3.11}$$

For $\mathbf{p}(\epsilon) = \mathbf{p}_0 + \epsilon(\delta_j - \mathbf{p}_0)$ as in Equation 3.8,

$$w_i(\mathbf{p}) = (1 + (n-1)\epsilon)^{Y_{ij}^*} (1-\epsilon)^{\sum_{k \neq j} Y_{ik}^*} \tag{3.12}$$

Letting $\epsilon \to 0$ results in

$$w_i(\mathbf{p}) \doteq 1 + n\epsilon(Y_{ij}^* - 1) \tag{3.13}$$

Substituting this back into Equation 3.11 gives

$$S(\mathbf{p}(\epsilon)) \doteq \sum_{i=1}^{B} [1 + n\epsilon(Y_{ij}^* - 1)] t_i^* / B = s_0 + n\epsilon \ \mathrm{cov}_j \tag{3.14}$$

Using Equation 3.8 defines $\dot{S}_j = n \ \mathrm{cov}_j$. Thus, the IJ estimated variance of a bagged predictor is

$$\hat{V}_{IJ} = \sum_{j=1}^{n} \mathrm{cov}_j \tag{3.15}$$

or the covariance between the predictions and the number of times each sample was used in the resamples.

Random Forest Prediction Variance

Wager et al. [47] have recently extended this idea by applying the IJ to random forest predictions. Based on using subsamples rather than bootstrap samples, they have shown that the variance of random forest predictions can be consistently estimated. Here the IJ variance estimator is applied to the resampling distribution for a new prediction point:

$$\hat{V}_{IJ} = \sum_{i=1}^{n} Cov_* \left[T(x; Z_1^*, ..., Z_n^*), N_i^* \right] \tag{3.16}$$

where $T(x; Z_1^*, ..., Z_n^*)$ is the prediction of the tree T for the test point x based on the subsample $Z_1^*, ..., Z_n^*$ and N_i^* is the number of times Z_i appears in the subsample. Furthermore, random forest predictions are asymptotically normal given that the underlying trees are based on subsampling and that the subsample size s scales as $s(n)/n = o(log(n)^{-p})$, where n the is number of training examples and p is the number of features [48].

Because $\hat{V}_{IJ}$ is calculated in practice with a finite number of trees B, it is inherently associated with Monte Carlo error. Although this error can be decreased by using a large B, a correction has been suggested [47]:

$$\hat{V}_{IJ}^B = \sum_{i=1}^{n} C_i^2 - \frac{s(n-s)}{n} \frac{\hat{v}}{B} \tag{3.17}$$

where $C_i = \frac{1}{B} \sum_{b=1}^{B} (N_{bi}^* - s/n)(T_b^* - \bar{T}^*)$ and $\hat{v} = \frac{1}{B} \sum_{b=1}^{B} (T_b^* - \bar{T}^*)^2$. This is essentially a Monte Carlo estimate of Equation 3.16 with a bias correction subtracted off. These estimates are asymptotically normal given a few key conditions, one of which is that the underlying trees are honest (see Section 3.1.3). Simulation experiments using sub bagged random forests have shown that these variance estimates are biased [48] and this is likely due to the fact that the underlying trees are not honest. The implementation of honest trees within a sub bagged tree ensemble and its resulting prediction variance has not been studied.

Here, we conduct simulation experiments to characterize the differences in the estimates

of the variances of bagged tree ensemble predictions when using variations of the random forest algorithm based around the type of sampling and regression tree.

3.1.3 Variations on Random Forests

Traditional random forests have a bias which favors splitting on variables with more levels or a larger continuous range [51]. The bias has been shown to come from two distinct sources [51]: (1) bootstrap resampling and (2) CART trees. We hypothesize that these two bias sources may also cause biased estimation of $\hat{V}_{IJ}^{B}$ and explore variations on random forests which eliminate the variable selection bias to see if they perform well with the IJ variance estimator.

Resampling

Each tree in the random forest algorithm is built on a resample of the original sample. By convention, the random forest uses a bootstrap sample (sampling with replacement), with size equal to the original sample size, n. It has been shown that a bootstrap sample distribution is different from the null hypothesis distribution, even if the original data set is distributed according to the null hypothesis [51]. Thus, statistical tests on the resamples used for choosing split points will be biased. This bias has been shown to be eliminated when using a subsample without replacement ("sub bagging") instead of bootstrapped samples [51]. Here, we implement subsampling of size $n^{0.7}$ as recommended by Wager [48] (Section 3.1.2) and compare the performance of $\hat{V}_{IJ}^{B}$ to using bootstrap resampling.

Tree Type

Wager defines an honest tree as one in which the distribution of the predicted outcome, conditional on the explanatory variables, does not depend on the training labels [48]. The most popular recursive partitioning algorithm and the one used in the random forest algorithm (Section 3.1.1) is CART [13]. In the case of regression, this algorithm performs a search for

the best possible split over all split points of all variables by minimizing the sum of squared errors between the predicted and actual values. These type of trees tend to select variables that have many possible split points, like continuous instead of categorical variables [51] due to the multiple testing problem. Conditional inference (CI) trees [52] are trees that are honest because they use outcomes to make predictions but use another method to find split points. CI trees implement a test statistic, like the Spearman correlation coefficient, student's t test, or F statistic from ANOVA to pick the predictor that is most associated with the outcome based on the smallest p-value. P-values are generated using a permutation test framework first laid out by Strasser [53] in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. To find the best split point, the standardized test statistic is then maximized. Implementing CI trees has been shown to alleviate the variable selection bias [51]. Here, we compare the performance of $\hat{V}_{IJ}^{B}$ when using CI trees in random forests compared to using CART trees.

3.2 Methods

3.2.1 Data Simulation

Ten different predictor variables $(X_1, ..., X_{10})$ were generated by sampling from the normal distribution, with $X_1, ..., X_5$ having mean zero and unit variance and $X_6, ..., X_{10}$ having a mean of ten and variance of five. Eleven different simulation functions were then used to generate eleven different synthetic outcomes. Table 3.1 shows the name and corresponding simulation function used to generate each simulated dataset. Here, AND and OR are used to denote the unique characteristics of these simulated datasets derived from using the indicator function, $I(\cdot)$ (1 if the argument is true; 0 otherwise). Similarly, SUM and SQ are based on the summing and summing of the squares of the predictor variables, respectively. The number in each data simulation name corresponds to the number of predictor variables

included in the simulation functions. Note that less than the ten total predictor variables are used in each data simulation although all ten predictor variables are used in the construction of the random forests. To simulate the data, 100 random test points were generated from each distribution and then 100 random training sets of varying size (n = 200, 1000, 5000) were generated from each distribution. Eleven different distributions, each with three different sample sizes, resulted in 33 total types of simulated data sets.

3.2.2 Random Forest Variations

All possible combinations of the proposed variations on random forests were implemented on the simulated data (both CI and CART trees, as well as bootstrap and subsampling). For each variation, m_{try} was set to three different default values: (1) $max\{floor(\frac{1}{3}p), 1\}$, (2) $max\{floor(\frac{2}{3}p), 1\}$, and (3) $max\{p, 1\}$ (where p is the total number of variables). For the current case of the simulations where $p = 10$, this resulted in m_{try} sizes of 3, 6, and 10. Because $p = m_{try} = 10$ is the trivial case of bagged trees, we instead opted for an m_{try} of 9. For subsampling random forest implementations, a subsample size of $n^{0.7}$ was used and all forests used $B = 5n$ total trees, as recommended by Wager [48]. Table 3.2 shows the specific numbers for subsample and resample size for each corresponding total sample size.

3.2.3 Simulation Experiments

Figure 3.2 contains a diagram depicting an overview of the simulation experiments. Specifically, the four combinations of CI or CART trees with bootstrap or subsampling were implemented on all simulated data sets (Table 3.1), each with sample sizes of 200, 1000, and 5000. $\hat{V}_{IJ}^{B}$ (Equation 3.17) was calculated for all 100 test points, each using all 100 training samples. The empirical prediction variance for each test point was calculated as the variance of all 100 predictions of each test point on the training samples. The absolute bias in $\hat{V}_{IJ}^{B}$ was calculated as the absolute difference of each variance estimate and the empirical prediction variance. The average absolute bias for each test point was calculated by averaging

the absolute bias across all 100 data sets. The average absolute bias was normalized by the empirical variance and termed the "absolute predictive bias" in order to help interpretation and compare biases across different distributions and prediction ranges. Averaging the absolute predictive bias across all 100 test points resulted in the mean absolute predictive bias (MAPB) for each combination of tree type, resampling type, distribution, and sample size:

$$MAPB = \frac{1}{100}\sum_{k=1}^{100}\frac{\frac{1}{100}\sum_{r=1}^{100}|\hat{V}_{IJ}\left(x^{(k)};Z^{(r)}\right) - \mathrm{Var}_r\left[RF_s\left(x^{(k)};Z^{(r)}\right)\right]|}{\mathrm{Var}_r\left[RF_s\left(x^{(k)};Z^{(r)}\right)\right]} \tag{3.18}$$

where k represents the index of each test point $x^{(k)}$, r is the index of each training sample $Z^{(r)}$, and $RF_s\left(x^{(k)};Z^{(r)}\right)$ is the ensemble prediction of the k^{th} test point using the r^{th} training set.

3.2.4 Statistical Computing

All statistical computing was done in `R`, version 3.1.2 [42], using the `randomForest` [54] and `Party` packages [55] wrapped into the custom package, `RFinfer` (available online at `https://github.com/cole-brokamp/RFinfer`). Section 3.5 describes the custom `R` package, including installation and usage examples.

3.3 Results

3.3.1 Data Simulation

Ten predictor variables were generated from the normal distribution in order to simulate the data. $X_1, ..., X_5$ were drawn from the standard normal distribution but $X_6, ..., X_{10}$ were drawn from a normal distribution with a mean of ten and a variance of five. The last five predictor variables were created to have a larger range in order to observe the effects of the random forest variations compared to data with only small ranging predictors. 100 data sets

were simulated for each of the eleven different simulation functions (Table 3.1) and different sample sizes (n = 200, 1000, 5000). Furthermore, 100 test points used for prediction were generated for each simulation function.

3.3.2 Empirical Variance

For each of the 100 test points, the variance of their predictions over all 100 test sets was calculated and termed the empirical variance. Table 3.3 contains the median of these empirical variances for each distribution and sample size. As expected, the empirical variance increased within each type of distribution as more variables were used to generate the synthetic outcome and also decreased with increasing sample size. The OR and AND empirical variances were relatively small, all with a median less than 0.005. This is likely because the distributions utilized the indicator function, reducing the synthetic outcome to only a few possible levels and defeating the effect of using predictors $X_6, ..., X_{10}$, which had a larger range than $X_1, ..., X_5$. In contrast, the SUM and SQ distributions had a relatively larger empirical variance, especially $SUM5$ and $SQ5$, which utilized the predictors with a larger range and variation.

3.3.3 Bias in Variance Predictions

The mean absolute predictive bias (MAPB) was calculated for each combination of resample type, tree type, m_{try}, sample size, and distribution by calculating the absolute difference in the variance estimate and the empirical variance, normalizing this bias by the empirical variance, and averaging over all 100 data sets and all 100 prediction points (Equation 3.18). See Figure 3.2 for a diagram depicting the simulation experiments. Table 3.4 and Figure 3.3 show the MAPB for all variations and Figures 3.4, 3.5, and 3.6 show the MAPB for sample sizes of $n = 200$, $n = 1000$, and $n = 5000$, respectively. CI trees were not created for $n = 5000$ due to computational limitations. Each of the simulation factors are explored in detail below.

3.3.4 Distribution

Overall, the SUM and SQ distributions performed well, with MAPB of mostly less than one. The AND and OR distributions, however, performed much worse, especially with increasing sample size and using bootstrap resampling and high m_{try} values.

m_{try}

Increasing m_{try} caused a large increase in MAPB for the OR and AND distributions, but caused a smaller effect with varied directions on the SUM and SQ distributions. Using subsampling with the OR and AND datasets generally exhibited a small increase in MAPB with increasing m_{try}, whereas using bootstrap resampling with the OR and AND datasets exhibited a very large increase in MAPB with increasing m_{try}. Within the SUM and SQ distributions, m_{try} had a much larger impact when bootstrap was used as the resampling method rather than subsampling. Here, m_{try} had a varied effect when using bootstrap resampling that depended on the number of variables used in each distribution and the total sample size.

Sample Size

For the SUM and SQ distributions, increasing sample size decreased the MAPB for all combinations of m_{try}, tree type and resample type. However, the OR and AND distributions showed the opposite effect of increasing sample size, with a higher MAPB. This effect was especially large with the bootstrap resampling method; for example, the MAPB of the random forests with an m_{try} of 9 trained on the $OR1$ distribution increased by an average of 3 fold when using $n = 1000$ instead of $n = 200$.

Tree Type

The effect of tree type was consistent no matter the sample size, resampling method, or m_{try} used; CI trees always had a lower MAPB than CART trees in every case. The decrease in

MAPB when using CI trees instead of CART trees did not seem to differ with respect to sample distribution.

Resampling Method

The best improvement in MAPB resulted from utilizing subsampling rather than bootstrap sampling. In fact, the worst performing simulation type using subsampling always performed better than the best simulation type using bootstrap sampling for every distribution. This was again the case for all combinations of sample size, m_{try}, and tree type. The difference was inflated when using a higher m_{try} in the bootstrapped OR and AND distributions.

3.4 Discussion

Here we have shown that using the IJ to estimate the variance of random forest predictions is much more accurate when using conditional inference trees instead of traditional CART trees and when using subsampling instead of bootstrap sampling. These simulation experiments show that Wager's proof [48] holds when using CI trees instead of traditional CART trees under various simulated data sets of different distributions and sizes.

3.4.1 Resample Method

The factor with the largest impact on MAPB was by far the resample method. Implementing sub bagging resulted in a more accurate estimation of the prediction variance, and this eclipsed the change in MAPB caused by any other variations in m_{try}, sample size or tree type. Although using CI trees was better than using CART trees, the difference between these two was highlighted the most when using bootstrap resampling with the nonlinear OR and AND functions. However, the magnitude of improvement in MAPB was not increased for distributions utilizing the predictors with a wider range. This was not expected given the known bias of CART trees utilizing wider ranging predictors [51].

3.4.2 Sources of Increased Bias

The nonlinear distributions, OR and AND had an extremely small empirical variance compared to the SUM and SQ distributions. Furthermore, the empirical variance decreased more rapidly with increasing sample size compared to the SUM and SQ distributions too. Thus, the increase in MAPB is likely due more to the decreased empirical variance rather then an increase in the absolute error of $\hat{V}_{IJ}^{B}$ and this is likely why the MAPB increased with increasing sample size for the OR and AND distributions, but decreased with increasing sample size for the SUM and SQ distributions.

The key to the random forest model is decorrelation of the individual trees using m_{try} and resampling. Bootstrap resampling does decorrelate trees, with each resample showing an average number of distinct observations of $0.632n$ [10]. However, using subsampling with a subsample size of $n^{0.7}$ results in a far lower number of distinct original observations per resample (see Table 3.2 for example). Thus, subsampling creates more decorrelation in individual trees than bootstrap sampling and so m_{try} makes a large difference in the performance of bootstrapped random forests because there is room for additional decorrelation, but not in subsampled random forests. Similarly, the effects of m_{try} are greater in the AND and OR distributions when using bootstrap resampling and not subsampling because the variance of the resamples are already so small that bootstrap resampling does not sufficiently decorrelation the individual trees, unlike when using subsampling. Overall, this is why the worst performing subsampled simulation still outperformed the best bootstrapped simulation. Subsampling is likely more resistant to the correlation problems found in data with a lower variance and forests built with a higher m_{try} value.

3.4.3 Conclusion

These results suggest that random forests should be constructed using CI trees instead of CART trees. It has already been shown that CI trees produce more accurate predictions [51] and we show here that they produce more accurate estimations of the prediction variance

too. However, it is most important to use subsampling instead of bootstrap sampling as this has the largest impact on the accuracy of $\hat{V}_{IJ}^{B}$.

3.4.4 Future Directions

These simulations extend those performed previously by Wager [47, 48] by using different distributions, varying m_{try} values, and including auxiliary noise variables in the training sets. However, in the future it would be valuable to evaluate the performance of $\hat{V}_{IJ}^{B}$ on correlated or multivariately distributed data, as well as on data with complex interactions because this is where the advantage of random forest truly lies.

The `RFinfer` package makes the code used in the simulations within this manuscript freely available and makes the experiments reproducible. We hope to extend this package in the future with functions that take advantage of the prediction confidence intervals, allowing researchers to take advantage of the novel insights into these "black box" models.

3.5 `RFinfer` package for `R`

3.5.1 Introduction

`RFinfer` is a package for `R` designed as a useful set of add on tools for the `randomForest` and `Party` packages. Currently, it produces prediction variances based on the IJ from random forests with the option to use bootstrap or subsampling as well as CART or conditional inference trees. In the future, we plan to expand this package to include other tools that will help in drawing statistical inferences from random forests like predictive comparisons and visualization tools.

3.5.2 Installation

The package is currently not available on CRAN, but can be installed from GitHub by running `devtools::install_github('cole-brokamp/RFinfer')` in `R`.

Dependencies

The IJ algorithm requires the number of times that each data point is included in each resample, but the `randomForest` package on CRAN only provides an indicator if each data point was included or not. To install a modified version of the package that includes the number of uses for each data point in each resample when specifying the `keep.inbag=TRUE`, option run `devtools::install_github('cole-brokamp/randomForest')` in `R`. Note that the package will not work without this modified version of the `randomForest` package.

Computation Time

Currently, the CI methods are much more computationally intensive because there is no C implementation of the CI random forest method that indicates the number of times that each sample is included in each resample. In order to carry out our simulations using $\hat{V}_{IJ}^{B}$, we had to use a pure R implementation of CI random forests. This is different for CART random forests, where a C implementation already exists in the `randomForest` package. However, it should be noted that the difference in computational times is due to the random forest creation step, not the implementation of $\hat{V}_{IJ}^{B}$. This should not be an issue in the future when a C implementation of CI random forests is created.

3.5.3 `rfPredVar()`

The main function of the package is `rfPredVar()`, which takes a supplied random forest and returns its predictions and prediction variances. The user can specify if the forest should be built using conditional inference trees instead of the traditional CART trees, if a progress bar

should be shown, and if the 95% confidence intervals should also be returned. The specific arguments are:

`rf.data`: Original data used to train the random forest

`pred.data`: Data used to predict with the forest, defaults to `rf.data` if not given

`CI`: Should 95% confidence intervals based on the central limit theorem be returned?

`tree.type`: Either `'ci'` for conditional inference tree or `'rf'` for traditional CART tree

`prog.bar`: Should a progress bar be shown? (only applicable when `tree.type='ci'`)

`rf`: A random forest trained with `keep.inbag = TRUE`

This function takes a random forest in order to extract m_{try}, the number of trees, and if the forest was trained using subsampling or bootstrap resampling. If subsampling was specified in the original random forest call, but `rfPredVar` is called with conditional inference trees, the algorithm will match the exact subsamples for each tree, allowing for direct comparison of the effect of tree type without variation in the resampling. Instead of the default predict method for forests from `cforest`, the predictions from a conditional inference tree are the direct averages of all tree predictions, *not* using the observation weights. Therefore, predictions from this function will likely differ from `predict.cforest` when using subsampling. Finally, this function only works with regression forests and not classification forests.

3.5.4 Example

First, load the package and some example data.

```
> library(RFinfer)
> data(airquality)
```

```
> d <- na.omit(airquality)
```

Next, train a random forest using subsampling with the `keep.inbag` option specified.

```
> rf <- randomForest(Ozone ~ .,data=d,keep.inbag=T,sampsize=30,replace=FALSE)
```

Extract the prediction variances along with their 95% confidence intervals.

```
> rfPredVar(rf,rf.data=d,CI=TRUE,tree.type='rf')
```

This will result in a data frame with the following structure:

```
## 'data.frame':  111 obs.  of 4 variables:
##   $ pred :  num 42.2 30.7 24.5 26.8 32.3 ...
##   $ pred.ij.var:  num 6.74 8.93 3.81 3.58 13.98 ...
##   $ l.ci :  num 28.95 13.19 17.09 19.79 4.92 ...
##   $ u.ci :  num 55.4 48.2 32 33.8 59.7 ...
```

If we wanted to compare this to a forest built using conditional inference trees instead of the traditional CART trees, we could make another call to `rfPredVar`:

```
> rfPredVar(rf,rf.data=d,CI=TRUE,tree.type='ci')
```

However, if we wish to use bootstrap sampling instead of subsampling, we would specify that in the original call to the random forest:

```
> rf <- randomForest(Ozone ~ .,data=d,keep.inbag=T,replace=TRUE)
```

and proceed as above.

Tables

Table 3.1: Ten functions used to simulate data. $X_1, ..., X_5$ were sampled from the normal distribution with mean 0 and variance 1. $X_6, ..., X_{10}$ were sampled from the normal distribution with mean 10 and variance 5.

Name	Simulation Function
SUM1	X_1
SUM3	$X_1 + X_3 + X_5$
SUM5	$X_1 + X_3 + X_5 + X_6 + X_7$
SQ1	X_1^2
SQ3	$X_1^2 + X_3^2 + X_5^2$
SQ5	$X_1^2 + X_3^2 + X_5^2 + X_6^2$
OR1	$I(X_1 > 0.4)$
OR3	$I(X_1 > 0.4) * I(X_3 > 0.6) * I(X_5 > 0.4)$
OR5	$I(X_1 > 0.4) * I(X_2 > 0.6) * I(X_3 > 0.4) * I(X_5 > 0.4) * I(X_6 > 6)$
AND3	$\frac{1}{3}[I(X_1 > 0.4) + I(X_2 > 0.6) + I(X_3 > 0.4)]$
AND5	$\frac{1}{5}[I(X_1 > 0.4) + I(X_3 > 0.6) + I(X_5 > 0.4) + I(X_6 > 6)]$

Table 3.2: Three sample sizes (n) used for the simulated data sets and their corresponding subsample size (s), subsample fraction (s/n), and total resamples (B).

n	$s = n^{0.7}$	s/n	$B = 5n$
200	41	0.20	1,000
1,000	126	0.13	5,000
5,000	388	0.08	25,000

Table 3.3: Median of the empirical variances (Var) for each distribution and sample size.

Distribution	Var ($n = 200$)	Var ($n = 1000$)	Var ($n = 5000$)
SUM1	0.0055	0.0007	0.0001
SUM3	0.0531	0.0192	0.0061
SUM5	0.8661	0.2588	0.0946
SQ1	0.0501	0.0088	0.0018
SQ3	0.3512	0.1236	0.0404
SQ5	78.8994	17.1922	6.1274
OR1	0.0018	0.0004	0.0000
OR3	0.0036	0.0007	0.0001
OR5	0.0048	0.0009	0.0001
AND3	0.0008	0.0001	0.0000
AND5	0.0009	0.0002	0.0000

Table 3.4: The mean absolute predictive bias (MAPB) for each simulation, each the combination of a distribution, m_{try}, sample size, tree type, and resampling method.

Distribution	m_{try}	CART						CI					
		Bootstrap			Subsample			Bootstrap			Subsample		
		200	1000	5000	200	1000	5000	200	1000	5000	200	1000	5000
SUM1	3	0.91	1.05	1.21	0.47	0.37	0.32	0.71	0.79		0.40	0.31	
	6	1.14	1.56	1.99	0.50	0.46	0.45	0.89	1.17		0.41	0.32	
	9	1.04	1.22	1.23	0.48	0.46	0.43	0.81	0.89		0.39	0.34	
SUM3	3	0.59	0.50	0.43	0.39	0.28	0.20	0.51	0.42		0.36	0.26	
	6	0.52	0.47	0.43	0.36	0.27	0.21	0.48	0.44		0.32	0.25	
	9	0.57	0.56	0.53	0.35	0.27	0.22	0.54	0.53		0.31	0.25	
SUM5	3	0.62	0.52	0.43	0.40	0.28	0.21	0.53	0.46		0.39	0.27	
	6	0.55	0.44	0.37	0.38	0.26	0.20	0.49	0.41		0.36	0.25	
	9	0.56	0.46	0.42	0.37	0.27	0.21	0.51	0.44		0.35	0.25	
SQ1	3	0.98	1.16	1.30	0.55	0.45	0.38	0.78	0.94		0.45	0.31	
	6	1.24	1.60	2.19	0.63	0.57	0.52	0.88	1.05		0.46	0.34	
	9	1.34	1.49	2.17	0.66	0.62	0.54	0.89	1.05		0.49	0.34	
SQ3	3	0.67	0.60	0.50	0.43	0.30	0.25	0.56	0.50		0.37	0.26	
	6	0.63	0.52	0.46	0.42	0.29	0.24	0.54	0.48		0.37	0.27	
	9	0.65	0.56	0.50	0.41	0.30	0.23	0.59	0.54		0.37	0.28	
SQ5	3	0.95	1.09	1.03	0.51	0.42	0.36	0.76	0.88		0.42	0.29	
	6	1.14	0.91	0.64	0.60	0.52	0.39	0.97	0.87		0.42	0.35	
	9	0.82	0.60	0.54	0.56	0.41	0.29	0.71	0.54		0.42	0.33	
OR1	3	1.48	2.60	4.79	0.50	0.51	0.63	1.21	2.04		0.42	0.40	
	6	3.16	8.68	20.57	0.75	1.03	1.65	2.36	6.89		0.52	0.70	
	9	7.12	22.94	45.56	1.31	1.71	2.09	5.70	19.24		0.91	1.44	
OR3	3	1.30	2.64	4.82	0.44	0.46	0.61	0.95	1.90		0.37	0.35	
	6	1.87	5.21	16.36	0.50	0.72	1.10	1.31	3.56		0.37	0.48	
	9	2.72	11.83	40.21	0.60	1.01	1.72	2.06	8.89		0.40	0.70	
OR5	3	1.12	2.45	4.71	0.43	0.46	0.63	0.84	1.98		0.38	0.35	
	6	1.38	3.97	11.75	0.43	0.60	0.97	1.03	2.84		0.38	0.42	
	9	1.72	7.58	24.36	0.46	0.79	1.31	1.33	4.53		0.37	0.52	
AND3	3	1.03	1.85	3.55	0.41	0.38	0.40	0.86	1.48		0.36	0.32	
	6	1.31	3.71	9.69	0.41	0.52	0.78	1.03	2.60		0.35	0.34	
	9	2.41	8.51	32.16	0.48	0.89	1.37	1.69	4.58		0.35	0.46	
AND5	3	0.78	1.24	2.57	0.40	0.33	0.32	0.69	1.11		0.39	0.29	
	6	0.87	2.25	6.68	0.38	0.39	0.52	0.70	1.66		0.37	0.29	
	9	1.38	6.27	19.41	0.40	0.54	0.96	0.98	2.82		0.37	0.33	

Figures

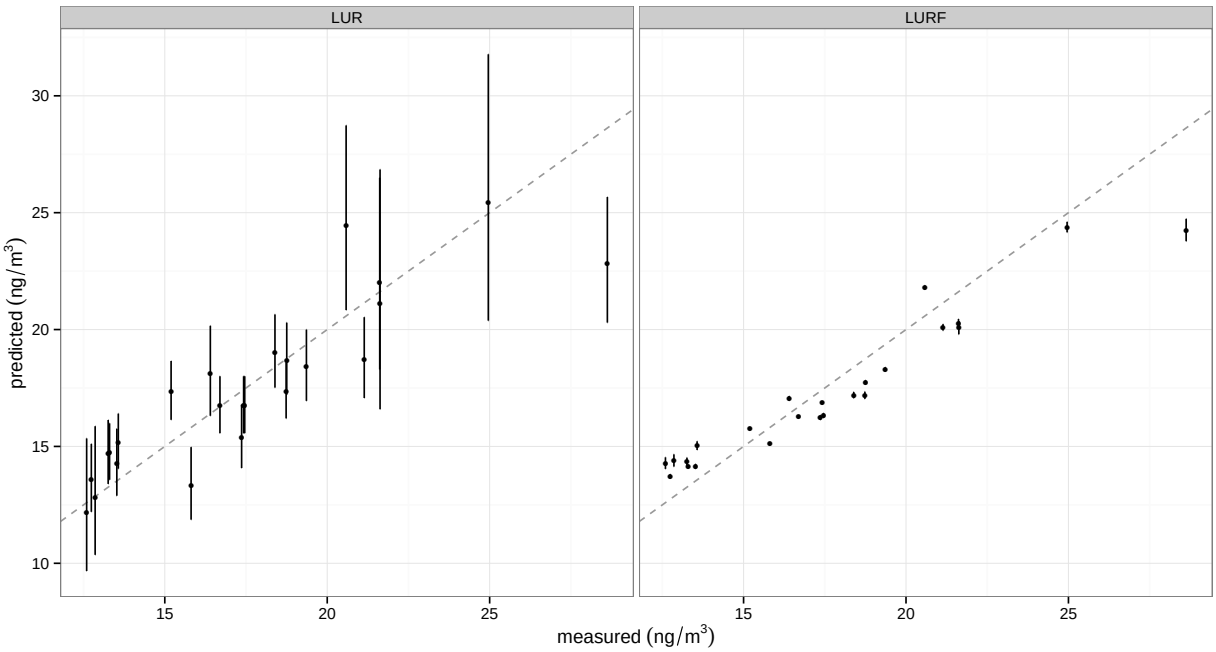

Figure 3.1: 95% confidence intervals for predictions from land use regression (LUR) and land use random forest (LURF) models according to the actual concentrations at all 24 CCAAPS sampling sites for PM2.5. The dotted line represents a perfect agreement between measured and predicted.

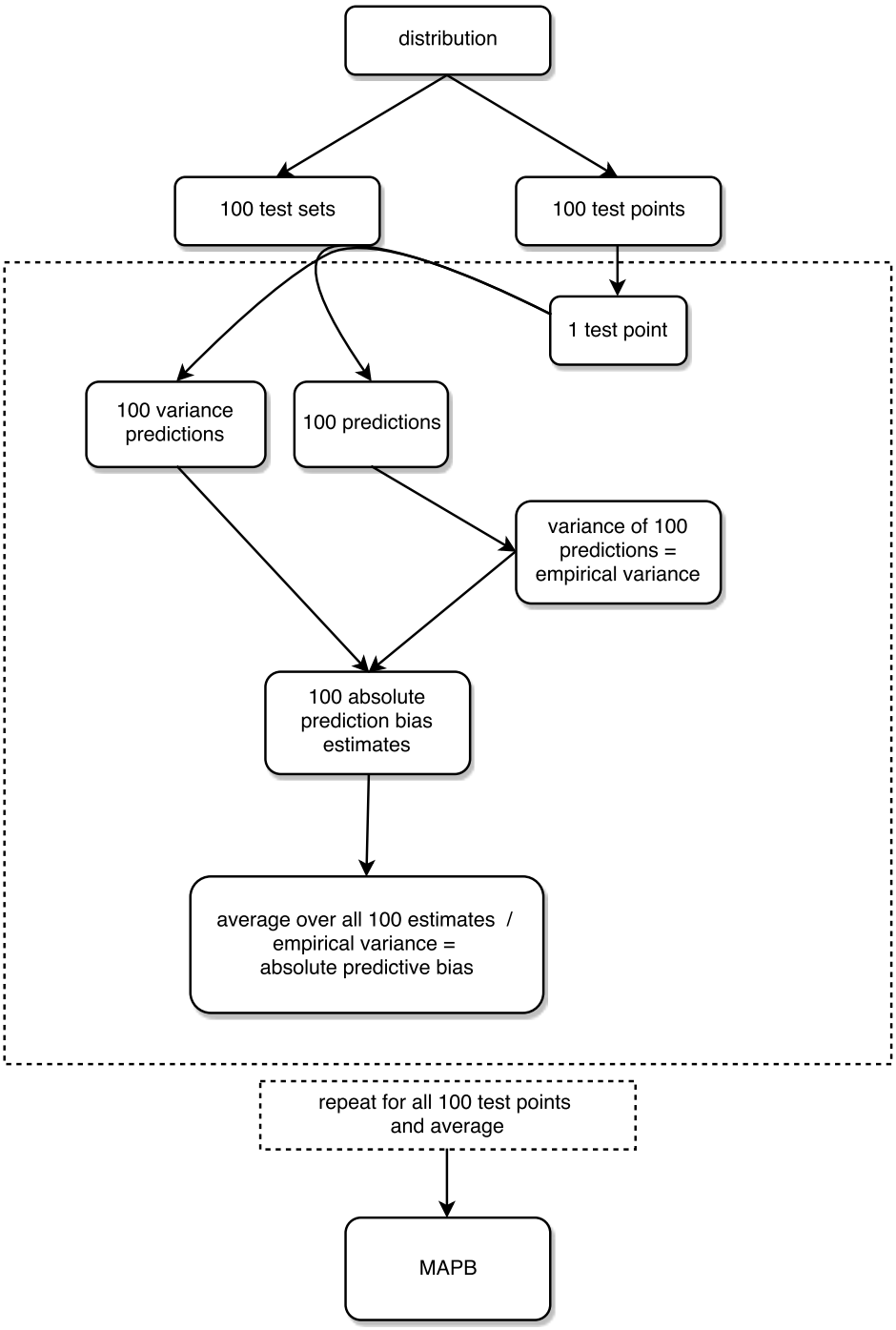

Figure 3.2: A diagram depicting the simulation experiments. 100 test points and 100 test tests were generated for each distribution and these were used to calculate the mean absolute predictive bias (MAPB).

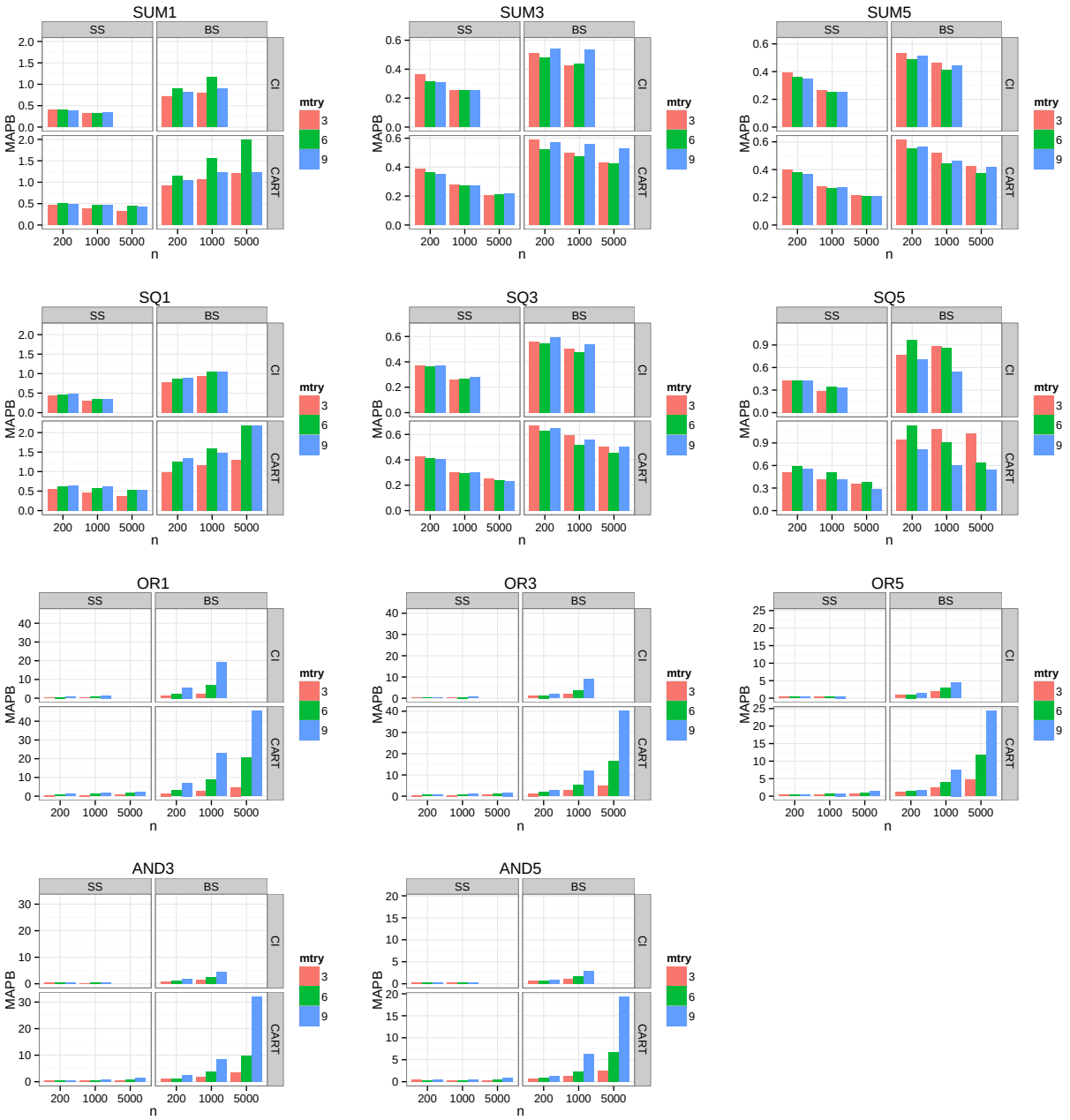

Figure 3.3: An overview of the MAPB for all simulation variations. Figures 3.4, 3.5, and 3.6 each show the results in more detail for each sample size.

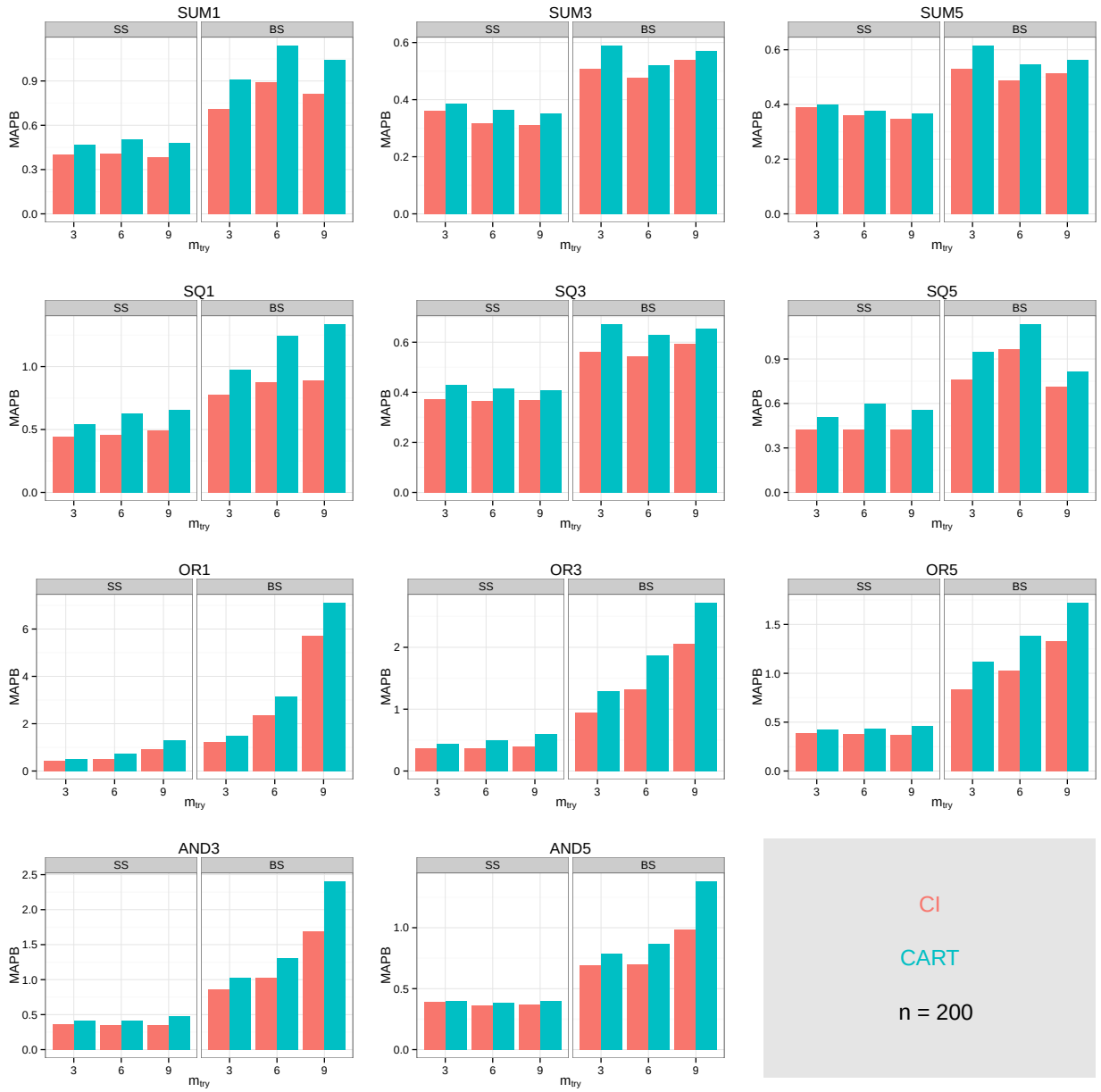

Figure 3.4: MAPB for each combination of subsample (SS) or bootstrap (BS) resampling, conditional inference (CI, red bars on left) or traditional CART (CART, green bars on right) trees, and m_{try} for $n = 200$.

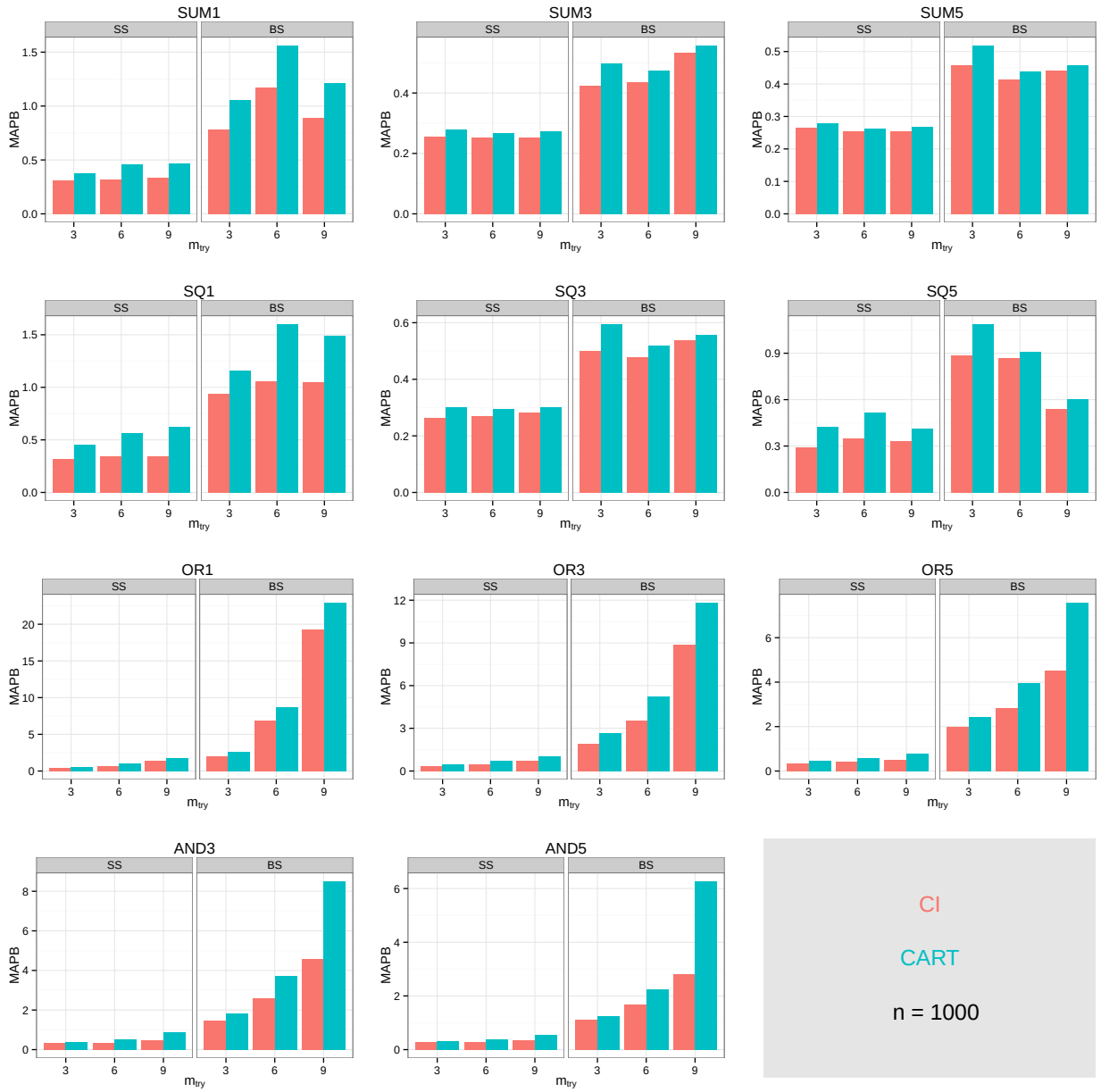

Figure 3.5: MAPB for each combination of subsample (SS) or bootstrap (BS) resampling, conditional inference (CI, red bars on left) or traditional CART (CART, green bars on right) trees, and m_{try} for $n = 1000$.

Figure 3.6: MAPB for each combination of subsample (SS) or bootstrap (BS) resampling, traditional CART (CART, green bars) trees, and m_{try} for $n = 5000$.

Chapter 4

Elemental Components of Particulate Matter and Respiratory Health

Association of Elemental Components of Particulate Matter with Childhood Lung Function and Asthma Development

Cole Brokamp[1], Roman Jandarov[1], MB Rao[1], David Bernstein[2], Sergey A. Grinshpun[1], Gurjit K. Khurana Hershey[3], James Lockey[1], Grace LeMasters[1,3], Patrick Ryan[1,4]

[1]Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA

[2]Department of Internal Medicine, University of Cincinnati, Cincinnati, Ohio, USA

[3]Division of Asthma Research, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[4]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Corresponding Author:

Cole Brokamp

University of Cincinnati

Kettering Lab Building, 160 Panzeca Way, Cincinnati, OH 45267-0056

E-mail: brokamrc@mail.uc.edu

Acknowledgments:

The authors thank the study and clinic staff for their efforts in study coordination, subject recruitment, data management, and data collection. They also thank the participating CCAAPS families for their time and effort. This work was supported by grants from the National Institute of Environmental Health Sciences (5R01ES011170 and R01ES019890).

The authors declare no competing financial interests.

Abstract

Background: Particulate matter (PM) has long been known to have a negative health on many aspects of public health. Although PM has traditionally been defined as all particles smaller than 2.5 μm or 10 μm, recent epidemiological studies have shown that specific elemental constituents of PM and their sources are associated with adverse health outcomes. The effects of individual PM components are important determinants of public health, but have not yet been evaluated in a largely urban city, nor in an American location.

Objectives: Our objective was to estimate the exposure of a cohort of children to elemental components of PM and find its effect on their respiratory health.

Methods: The exposure of children from the Cincinnati Childhood Allergy and Air Pollution Study to Al, Cu, Fe, K, Mn, Ni, Pb, S, Si, V, Zn, traffic related air pollution and total PM2.5 were estimated using a previously validated land use random forest model based on residential address history. A causal diagram was created and used to determine that models required adjustment for neighborhood characteristics. These were defined using census tract level socioeconomic characteristics and a residential traffic proximity metric. Linear models were used to relate the estimated exposures to diagnosis of asthma at age seven and percent predicted forced vital capacity (FVC) and forced exhaled volume in the first second (FEV_1).

Results: The exposure of the cohort to individual PM components did not necessarily reflect their estimated exposure to TRAP or total PM2.5. Al, Fe, Pb, Si, Zn, and total PM2.5 were associated with decreased lung function on their own, but after controlling for confounding by neighborhood characteristics, only Al and Se were significantly associated with FVC. Although not statistically significant all other elemental components except K and S remained negatively associated with lung function and asthma after adjustment. Future health models should consider elemental components of total PM when associating them with health effects.

4.1 Introduction

4.1.1 Health Effects of Particulate Matter Components

Particulate matter (PM) has long been known to have a negative effect on public health [56]. At a cellular level, it causes inflammation in the brain [57] and lungs [58]. At an individual level, it has been associated with increased risk for lung cancer [59, 60], asthma exacerbation [61], elevated blood pressure [62], and increased cardiovascular mortality from both short term [63] and long term [64] exposures. At a population level, it is associated with increased daily mortality [65, 66, 67, 68, 69, 70] and an overall shortened life expectancy [71, 72]. Although PM has traditionally been defined as the fraction of all particles smaller than 2.5 μm or 10 μm [73], recent epidemiological studies have shown that specific elemental constituents of PM and their sources are associated with adverse cardiovascular and respiratory health outcomes in adults [14, 15, 16]. Specifically, epidemiological studies have identified elemental carbon, organic carbon, and nitrates as being associated with increased risk for cardiovascular and respiratory hospital admissions [74, 75] and mortality [76]. Elemental components of PM2.5, including Ni, Zn, Si, Al, V, Cr, As, Br, have also been associated with increased cardiovascular and respiratory hospital admissions [14, 75], increased mortality [77], and lower birth weight [78]. Characterizing the health effects of PM components has been identified as a research priority by the National Research Council for the National Academies [17]. Recently, LUR models have been developed for particle composition in twenty areas in Europe as a part of the ESCAPE study [18] and exposure to PM2.5 sulfur was found to be associated with an increased risk for natural cause mortality in adults [20]. The same LUR model was also used to relate Ni and S exposure to decreased lung function in five cohorts of children [21]. Furthermore, the same group has associated long term exposure to PM2.5 copper and PM10 iron with increased levels of inflammatory blood markers [22]. Clearly, the effects of individual components of PM are important determinants of public health, but land use exposure models have not yet been created in an American location.

4.1.2 Causal Structure

Causal diagrams are useful in epidemiological research for identifying variables that have to be controlled for and measured in order to obtain the total and unconfounded effect estimates [79]. There exists a suggested framework for using causal diagrams to determine the unbiased effect estimate for an exposure on an outcome [80] and without an explicit causal model, "unprincipled covariate adjustment may fail to remove all confounding bias or even introduce new biases through over control or endogenous selection" [81]. These biases may include selection bias [82] or confounding and collider bias [83]. Covariates included in a model may be a confounder, which affects both the exposure and the outcome; a mediator, which is affected by the exposure and affects the outcome; a proxy confounder, which is affected by a confounder and affects the exposure or outcome; or a competing exposure, which affects the outcome but is not related with the exposure. Although these confounder definitions are not universally accepted, they provide a quantitative framework that links causal diagrams to statistical modeling. See [84] for a review of the types of confounding and different definitions.

Because of the difference between exposure covariates and other confounder covariates, effect estimates from a single model do not all have identical interpretations. Effects can be primary, which are the causal effects of the primary exposure; secondary, which are the causal effects of covariates; total, which are the net of all associations of a variable through all causal pathways; or direct, which are the association of a variable after blocking or controlling some causal pathways. Presenting effect estimates of exposure and confounders from the same model will likely lead to incorrect interpretations, such as confusion of direct effects for total effects, and this has been termed the "Table 2 fallacy" [85, 86]. Here, we use a causal inference structure to carefully determine the necessary modeling adjustments to unconfound socioeconomic status from air pollution exposure and respiratory health.

4.2 Methods

4.2.1 Study Population

The Cincinnati Childhood Allergy and Air Pollution Study (CCAAPS) is an ongoing prospective birth cohort that focuses on high-risk atopic children [24, 25]. Children in the Greater Cincinnati and Northern Kentucky region were screened by birth record between October 2001 and July 2003 and enrolled in the study if they lived at least 1,500 m or less than 400 m from the nearest major road. Figure 2.1 shows the birth locations of the CCAAPS cohort. In addition, each child must have had a parent with symptoms of asthma, eczema, or rhinitis and allergic sensitization by a positive skin prick test result to at least one of a panel of 17 aeroallergens. Informed consent was obtained and the study was approved by the University of Cincinnati Institutional Review Board.

4.2.2 Asthma Diagnosis and Lung Function

At the age seven study visit, CCAAPS children completed spirometric testing according to ATS criteria described elsewhere [87]. Percent predicted values of forced expiratory volume in one second (FEV_1) and total forced vital capacity (FVC) were calculated for children $<$ eight years [88]. Children with either a $FEV_1 < 90\%$ predicted, a physician diagnosis of asthma, asthma symptoms in the last 12 months (tight chest or throat, difficulty breathing or wheezing after exercise, wheezing and/or whistling in the chest), or an eNO level of $>$ 20 ppb received 2.5 mg levalbuterol through a nebulizer followed 15 minutes afterwards by repeat spirometry [89]. Children with $< 12\%$ increase in FEV_1 had a methacholine challenge test (MCCT). Children were defined as having asthma if they reported having symptoms of asthma and had either bronchial hyper-reactivity ($> 12\%$ increase in FEV_1 following bronchodilation) or a positive MCCT (PC20 $\leq$ of 4 mg/ml methacholine concentration) [90]. Of the total 762 children initially enrolled in the study, an asthma diagnosis was available for 598 of them and the remaining 164 were excluded from the analysis.

4.2.3 Exposure Assessment

The exposure assessment model is explained in full in Chapter 2, but briefly land use random forest (LURF) models were used to assess exposure to total PM2.5, traffic related air pollution (TRAP), and the following elemental components of PM2.5: Al, Cu, Fe, K, Mn, Ni, Pb, S, Si, V, and Zn. The LURF was based on ambient air sampling conducted in the study area at 24 different sites between 2001 and 2006. Land use predictors such as traffic intensity, roadway density, community-level socioeconomic variables, elevation gradient, greenspace, land cover, and emission point sources were used to build a random forest model for each PM2.5 component. The models were validated and used to assess the exposure of each child based on their annual residential address history. Their estimated exposures across all seven years were averaged to generate their total estimated exposure to each component.

4.2.4 Causal Structure

A causal diagram was used to quantitatively document the causal structure of PM exposure causing asthma diagnosis and lung function at age 7. Based on the currently available literature, we identified race, parental income, parental education, neighborhood characteristics, and other unobserved neighborhood stressors as confounders of the effect of PM on respiratory health. Secondhand smoke exposure, allergies, and unobserved genetic characteristics were identified as competing exposures. We chose the neighborhood effect because a lower socioeconomic status does not necessarily mean that a child will reside closer to sources of air pollution, but it can mean that they will live in a certain neighborhood in which a higher fraction of the population resides close to sources of air pollution. Adjustment for neighborhood level variables has been suggested in order to determine if any exposure relationship with health effects are due to neighborhood level confounding, indirect pathway biases, or collider bias [91, 92, 93]. To quantify this confounder, we created neighborhood characteristic indices.

4.2.5 Neighborhood Characteristics

In order to adjust for the neighborhood characteristics, we used census tract level measurements of socioeconomic status and traffic proximity. These were created for all census tracts within the counties in which CCAAPS subjects resided as well as the bordering counties (Ohio: Hamilton, Clermont, Butler, Warren, Preble, Montgomery, Greene, Clinton, Brown; Kentucky: Boone, Kenton Campbell, Gallatin, Grant, Pendleton, Bracken; Indiana: Switzerland, Ohio, Dearborn, Franklin, Union).

Socioeconomic Status

Eight census tract level variables (fraction that graduated high school, fraction of households in poverty, median household income, fraction of population receiving public assisted income, fraction of houses that are vacant, median home value, white fraction of population, and black fraction of population) were obtained from the 2010 Census 5-year American Community Survey for all counties in which CCAAPS children resided.

Residential Traffic Proximity

Residential traffic proximity (RTP) was calculated for each census tract by calculating the total length of Class 4 roads that were located within 400 meters of a Class 1 road and then dividing by the total length of Class 4 roads. See Section 2.2.1 for details on the roadway definitions. RTP essentially provides the fraction of each census tract that lives in close proximity to primary highways. Figure 4.7 shows the RTP for each census tract alongside the Class 1 and Class 4 roads.

Neighborhood Characteristic Indices

Neighborhood characteristics were calculated as described previously [32], except with the addition of RTP. Specifically, RTP and the eight census variables were included in a principal components analysis (PCA) to extract representative measures for each census tract. The

first three components explained 0.54, 0.15, and 0.11 of the total variance each and combined accounted for 0.8 of the total variation in the nine total variables. Each census tract was assigned a value for each of the three principal components using the individual variables and the PCA loadings. The collection of the three component measurements on all census tracts in the study area were then each normalized to a range of $[0, 1]$ by subtracting the minimum and dividing by the difference of the resulting range. Each child was assigned a value for each of the three neighborhood characteristic indices based on the census tract of their birth residential address.

4.2.6 Statistical Modeling

A small number of children had missing information on exposures to PM2.5 components and neighborhood characteristic principal components because of missing land use variables and census tract level characteristics, respectively. Similarly, 17 of the children each had missing lung function measurements. Table 4.1 contains the number of missing measurements for each variable for the total 598 children. Only children without missing information on variables used in each model were used.

Logistic regression and linear regression were used to associate the exposure of PM components with asthma diagnosis and lung function, respectively. The adjusted models included the three neighborhood characteristic principal components in order to estimate the total effect of the PM components on the outcomes. Pollutant concentrations were not transformed for use in the models and odds ratios and regression coefficients were calculated based on an increase of exposure equal to one interquartile range of each component. All statistical computing was done in `R`, version 3.1.2 [42]. The code used to calculate the land use predictors and generate exposure estimates for the cohort has been made into an R package and is available online at `https://github.com/cole-brokamp/aiRpollution`. The code used for the association with respiratory health effects is available upon request.

4.3 Results

4.3.1 Cohort Characteristics

Table 4.1 describes the cohort, which had an age 7 asthma prevalence of about 16% and average percent predicted FVC and FEV_1 values of 101.8 and 102.2, respectively. In addition to the data shown in the table, 22% of the cohort is African American, 55% is male, and 94% have mothers with at least a completed high school level education. All children resided in the Greater Cincinnati area, which is mostly urban. Figure 2.10 shows the location of the CCAAPS residential birth addresses alongside the highways, interstates, and bus routes in the study area of Greater Cincinnati.

4.3.2 PM Exposure

Land use random forest (LURF) models (Chapter 2) were used to estimate the exposure of each child to PM components as an average daily concentration based on their annual residential history from birth to age 7 for total PM2.5, traffic related air pollution (TRAP), and the following elemental components of PM2.5: Al, Cu, Fe, K, Mn, Ni, Pb, S, Si, V, and Zn. Table 4.2 and Figure 4.1 describe the estimated exposures for the cohort as a whole. The concentrations were the highest the highest for S, which had a median concentration of $1,521\ ng/m^3$ and was an order of magnitude larger than any other component. Other components with higher concentrations ($> 50\ ng/m^3$) were Fe, K, and Si, while Cu, Mn, Ni, Pb, and V all had lower concentrations ($< 3\ ng/m^3$). Of the PM components, S and Fe had the largest variation, followed by Si, Al, Zn, and K. Components with very small variation included Cu, Mn, Ni, Pb, and V.

Figure 4.2 illustrates the correlation matrix of the elemental PM exposures for the cohort. The upper diagonal and lower diagonal represent the Spearman correlation coefficient as a colored dot and numeric value, respectively. Most of the elemental components were highly correlated with one another, except for K and S. K had relatively small correlations

with the other elemental components, while S was slightly negatively correlated with all of the elemental components (except for K). TRAP was more highly correlated with all of the elemental components than PM2.5 was, although both were highly correlated with all components and one another.

In addition to Spearman correlation coefficients, elemental PM components were plotted according to the estimated exposure to TRAP (Figure 4.3) and total PM2.5 (Figure 4.4). Each figure is overlaid with a local second degree polynomial regression line fit and 95% confidence interval spanned over 75% of the closest data points. These plots highlight that exposure to PM components does not necessarily reflect the total PM2.5 or TRAP for all of the CCAAPS cohort. Most elements are linearly correlated with TRAP, but K, Mn, Ni, and S are exceptions. K, Mn, and Ni follow a linear relationship with higher levels of TRAP ($> 0.4\ \mu g/m^3$), but show a steeper decrease in concentrations for lower concentrations of TRAP. S behaves differently, with its concentration averaging about $1,500\ ng/m^3$ for all TRAP concentrations less than $0.6\ \mu g/m^3$. The difference between the elemental exposure estimates and the total PM2.5 exposure estimates are even greater. All of the elements show a nearly stable mean concentration for lower PM2.5 values ($> 17\ \mu g/m^3$), but then exponentially increase for higher levels of total PM2.5.

To further highlight the difference between the panel of elemental PM components and total PM2.5, the exposure estimates for three subjects that each moved three times before age seven are plotted in Figure 4.5. The child in the top set of panels was exposed to relatively lower concentrations of PM2.5, with exposure to almost exactly the median amount of PM2.5 in the first three years of life. However, the child moved at age two and although their estimated exposure to total PM2.5 stayed nearly the same after the move, they experienced large decreases in Al, Cu, Ni, and TRAP as well as large increases in S and Mn. Similar trends can be seen for the child in the middle set of panels who was estimated to have the lowest exposure to total PM2.5 during the first two years of life, but estimated to have the highest exposure to TRAP during the first two years of life. Similarly, the child in the

bottom set of panels was estimated to have above median exposure to total PM2.5 for all ages except two; however, this child was estimated to have above median exposure to Mn, Ni, and Si for only age two.

Overall, comparison of the longitudinal estimated exposure estimates clearly shows that total PM2.5 or TRAP exposure estimates do not necessarily correspond with elemental exposure estimates.

4.3.3 Confounding by Neighborhood Characteristics

A causal diagram (Figure 4.6) was used to determine that accounting for confounding by neighborhood characteristics would allow for the estimation of the total effect of PM exposure on asthma development. The neighborhood characteristics were calculated based on eight census tract level socioeconomic variables and residential traffic proximity (RTP). RTP was calculated as the fraction of the total length of residential roads that were within 400 meters of major highways for each census tract (Figure 4.7). Three census tract level indices from these nine total variables were extracted using principal components analysis to represent the "neighborhood" confounder in our model. All three indices were included in the adjusted models based on the census tract of the child's residential address at birth.

4.3.4 Effect of PM Exposure on Asthma and Lung Function

All odds ratios (OR) and regression coefficients were calculated separately for an increase of exposure equal to the interquartile range (IQR) for each PM component; see Table 4.1 for the IQR of each PM component. Adjusted models included the addition of all three neighborhood characteristic indices.

Asthma

The odds ratios for the risk of asthma development by age seven and an IQR increase in PM component exposure are shown in Table 4.3 and Figure 4.8. Unadjusted models showed

an increased risk of asthma diagnosis at age 7 for increased exposure to Fe, Mn, V, and total PM2.5 (ORs of 1.26, 1.19, 1.25, and 1.38, respectively). Most of the other elements, including Al, Cu, Ni, Pb, Si, Zn, and TRAP had odds ratios much greater than one, but had confidence intervals that slightly overlapped one. Both K and S were not associated with asthma at all, with odds ratios of 0.91 and 0.98, respectively. After adjusting for the neighborhood characteristics, asthma was not statistically associated with exposure to any PM component. The addition of the neighborhood characteristics did not change the estimate variances, but rather drove the estimate itself towards a null association.

$\mathbf{FEV_1}$ and FVC

The regression coefficients for the change in percent predicted FEV_1 and FVC at age seven due to an IQR increase in PM component exposure are shown in Table 4.4, Table 4.5, and Figure 4.9. In general, most of the PM components were associated with decreased lung function; however, as in their relationship with asthma, S and K did not follow the general trend. The regression coefficients for S were very close to zero and the regression coefficients for K were positive. In the unadjusted models for FEV_1, exposures to Al, Fe, Pb, Si, Zn, and total PM2.5 were significantly associated with decreased lung function ranging from -0.8% for Zn to -1.46% for total PM2.5. However, after the addition of the neighborhood characteristic indices, none of the components were significantly associated with FEV_1 as all coefficient estimates again moved towards the null association. In the unadjusted models for FVC, exposures to Al, Cu, Fe, Pb, Si, V, Zn, TRAP, and total PM2.5 were significantly associated with decreased lung function ranging from -0.83% for Zn to -1.51% for total PM2.5. Again, the addition of the neighborhood characteristic indices shifted most of the estimates toward the null association, but Al (-1.34%) and Si (-1.19%) remained significantly associated with decreased lung function. In this case, unlike with asthma, the addition of the neighborhood characteristics did inflate the variance of the regression coefficient estimates.

4.4 Discussion

Here, we have shown that in the urban area of Cincinnati, Ohio, exposure to total PM2.5 does not necessarily represent exposure to individual components of PM. We found that the components were differentially associated with asthma and lung function. Using a causal diagram, we were able to control for confounding neighborhood characteristics in order to estimate the total effect of PM exposure on respiratory health. Although Al, Fe, Pb, Si, Zn and total PM2.5 were associated with decreased lung function on their own, after controlling for confounding, only Al and Si were significantly associated with FVC. Although not statistically significant all other elemental components except K and S remained negatively associated with lung function and asthma. The statistically insignificant results after adjusting for confounding by neighborhood level characteristics were due to increased variance estimates and not associations shifting towards the null. The increased variance is likely due to the uneven distribution of exposures and outcomes, with children with poor respiratory health very much more likely to live in neighborhoods with poorer community characteristics.

4.4.1 Previous Work

Only one other study has previously associated exposure to elemental PM components with childhood respiratory health [21]. Exposure assessment for Cu, Fe, K, Ni, S, Si, V, and Zn were based on LUR models for five different European birth cohorts. Although our study only included one cohort, it is interesting to note that Eeftens et al. observed widely differing within-cohort variability of estimated PM concentrations. Their cohorts were estimated to have levels of exposure to elemental components similar in magnitude as our study found, expect for S and K. These were the two components that were not correlated with the other estimated exposures in our study and were not negatively associated with respiratory health in our cohort. Specifically, our estimates of S were much higher than those from Eeftens et al and this could be due the high use of S in diesel fuel during our sampling campaign

(2001 - 2006) compared to lower utilization of S in fuels in Europe during their sampling campaign (2008 - 2010). Furthermore, K is known to be associated with biomass burning and could be another exposure that is specific to the geography and community characteristics in Cincinnati, Ohio compared to the communities in Europe. The association between exposure and lung function was conducted individually for each site and then pooled using a meta analysis. Eeftens et al. found that the association between elemental concentrations and both FVC and FEV_1 differed among cohorts, but overall none of the elements were significantly associated with lung function in their confounder adjusted models. Instead of using a causal structure for adjustment in their models, they instead adjusted for all individual confounders that were not on the causal pathway but were univariately associated. As discussed in the introduction (Section 4.1.2), it is known that this can lead to biased effect estimates. Instead of using average lifetime exposure, the estimates calculated by Eeftens et al. were based on the address collected for age 6 or 8 years. Using only the most recent residential address has been shown to underestimate exposure to air pollution by missing high early life exposures because children usually move to areas with lower air pollution levels [32].

4.4.2 Elemental Exposure Signature

Differences in the "elemental exposure signature" are likely due to specific elemental sources that are not captured in total PM2.5 models. See Table 2.4 for a list of the final predictors used in the LURF models. Although total PM2.5 mass may not differ, its composition may differ widely and aggregating total PM2.5 mass into one measurement likely disguises the nuances of differing elemental exposures, which clearly affect health differently.

4.4.3 Conclusion

Here we showed that estimated exposures to total PM2.5 and TRAP do not necessarily correspond with estimated exposures to individual components and this should be considered in future epidemiological studies association PM with health effect. A disadvantage of our study

(and almost all studies utilizing exposure estimate models) is that the exposure estimates are considered fixed and not random. This ignores uncertainty in the exposure estimates and may bias the estimation of effects with health outcomes. In the future, exposure assessment and health effects could be combined into one hierarchical model. Furthermore, it will be important to consider multivariate and mixture effect models using the entire panel of PM components instead of associating each component individually with health effects because interactions between elemental concentrations could play important roles in their negative health effects. Finally, further research on different cohorts and geographical areas should be conducted on the association of elemental PM components and childrens' respiratory health.

Tables

Table 4.1: Summary (mean or number and percentage yes) of elemental PM2.5 exposures, neighborhood characteristic principal components, and health outcomes for the CCAAPS cohort with a complete asthma diagnosis at age 7 ($n = 598$). PM2.5 components are expressed as the average daily concentration in ng/m^3 and FVC and FEV_1 are expressed as a percent of the predicted value. Also included are the interquartile ranges (IQR) for the PM2.5 components, used to generate the odds ratios and regression coefficients in the models.

PM2.5 Component	Summary	Number Missing	IQR
Al	32.30	10	6.05
Cu	2.12	10	0.69
Fe	75.87	11	21.10
K	65.53	10	10.38
Mn	2.37	10	0.77
Ni	0.59	10	0.13
Pb	3.08	10	0.80
S	1478.83	10	415.20
Si	91.82	11	14.29
V	0.34	12	0.06
Zn	15.45	10	5.09
TRAP	384.34	11	79.01
Total PM2.5	16,530.25	10	1,400.69
Neighborhood Characteristic PC			
PC1	0.43	2	
PC2	0.65	2	
PC3	0.74	2	
Respiratory Health Outcomes			
Asthma	95 (16%)	0	
FVC	101.80	17	
FEV_1	102.21	17	

Table 4.2: Summary of elemental concentrations, TRAP, and total PM2.5 estimated for the CCAAPS cohort ($n = 598$). All units are ng/m^3.

PM2.5 Component	Minimum	25th Percentile	Median	Mean	75th Percentile	Maximum	SD
Al	21.35	28.45	31.02	32.30	34.50	71.70	6.04
Cu	1.31	1.63	1.97	2.12	2.32	5.73	0.65
Fe	55.01	61.26	69.04	75.87	82.37	190.37	21.20
K	56.06	59.83	64.88	65.53	70.21	80.52	6.15
Mn	1.62	1.79	1.97	2.37	2.55	6.14	0.92
Ni	0.38	0.50	0.57	0.59	0.63	1.32	0.14
Pb	2.27	2.52	2.83	3.08	3.32	11.50	0.99
S	936.19	1,255.47	1,521.42	1,478.83	1,670.67	2,685.12	249.14
Si	72.12	82.47	87.10	91.82	96.75	213.23	15.33
V	0.24	0.30	0.33	0.34	0.36	0.70	0.06
Zn	10.12	11.84	13.78	15.45	16.92	72.48	6.70
TRAP	288.08	331.79	359.35	384.34	410.80	846.27	79.07
Total PM2.5	13,676.18	15,745.89	16,367.55	16,530.25	17,146.58	21,042.41	1147.05

Table 4.3: Unadjusted and neighborhood adjusted odds ratios (OR) with lower and upper 95% confidence interval (CI) bounds for an interquartile increase in PM 2.5 component exposure and asthma diagnosis at age 7. Rows that have a confidence interval excluding one are bolded.

PM Component	Unadjusted OR (95% CI)	Adjusted OR (95% CI)
Al	1.22 (0.99,1.48)	1.03 (0.78,1.34)
Cu	1.13 (0.9,1.41)	0.94 (0.71,1.23)
Fe	**1.26 (1.03,1.53)**	1.07 (0.83,1.37)
K	0.91 (0.62,1.32)	0.84 (0.56,1.25)
Mn	**1.19 (1.01,1.4)**	1.05 (0.85,1.27)
Ni	1.14 (0.93,1.38)	0.98 (0.77,1.22)
Pb	1.16 (0.99,1.34)	1.01 (0.79,1.23)
S	0.98 (0.68,1.42)	1.06 (0.72,1.57)
Si	1.19 (0.99,1.42)	1.03 (0.8,1.3)
V	**1.25 (1.01,1.53)**	1.02 (0.77,1.34)
Zn	1.12 (0.97,1.29)	0.98 (0.76,1.19)
TRAP	1.19 (0.97,1.45)	0.96 (0.72,1.26)
Total PM2.5	**1.38 (1.06,1.78)**	1.17 (0.86,1.57)

Table 4.4: Unadjusted and neighborhood adjusted regression coefficients (Coef) with lower and upper 95% confidence interval (CI) bounds for an interquartile increase in PM 2.5 component exposure and FEV_1 at age 7. Estimates that have a confidence interval excluding zero are bolded.

PM Component	Unadjusted Coef (95% CI)	Adjusted Coef (95% CI)
Al	**-1.13 (-2.18,-0.09)**	-0.87 (-2.22,0.49)
Cu	-0.96 (-2.08,0.15)	-0.66 (-1.97,0.66)
Fe	**-1.34 (-2.38,-0.3)**	-1.14 (-2.44,0.16)
K	1.33 (-0.44,3.1)	1.54 (-0.25,3.33)
Mn	-0.58 (-1.45,0.29)	-0.24 (-1.26,0.78)
Ni	-0.41 (-1.41,0.59)	0.08 (-1.06,1.22)
Pb	**-0.91 (-1.75,-0.08)**	-0.64 (-1.76,0.48)
S	-0.22 (-1.97,1.54)	-0.3 (-2.08,1.49)
Si	**-1.27 (-2.24,-0.3)**	-1.15 (-2.4,0.1)
V	-0.93 (-1.99,0.12)	-0.44 (-1.82,0.93)
Zn	**-0.8 (-1.58,-0.01)**	-0.5 (-1.56,0.56)
TRAP	-0.92 (-1.97,0.13)	-0.54 (-1.95,0.88)
Total PM2.5	**-1.46 (-2.73,-0.19)**	-1.11 (-2.57,0.35)

Table 4.5: Unadjusted and neighborhood adjusted regression coefficients (Coef) with lower and upper 95% confidence interval (CI) bounds for an interquartile increase in PM 2.5 component exposure and FVC at age 7. Estimates that have a confidence interval excluding zero are bolded.

PM Component	Unadjusted Coef (95% CI)	Adjusted Coef (95% CI)
Al	**-1.37 (-2.36,-0.37)**	**-1.34 (-2.63,-0.05)**
Cu	**-1.12 (-2.18,-0.06)**	-1.04 (-2.29,0.21)
Fe	**-1.36 (-2.35,-0.37)**	-1.2 (-2.44,0.05)
K	0.82 (-0.87,2.51)	0.93 (-0.78,2.64)
Mn	-0.78 (-1.61,0.05)	-0.58 (-1.55,0.39)
Ni	-0.85 (-1.81,0.11)	-0.49 (-1.58,0.59)
Pb	**-0.95 (-1.75,-0.15)**	-0.7 (-1.77,0.37)
S	0.16 (-1.52,1.84)	0.08 (-1.62,1.79)
Si	**-1.24 (-2.16,-0.31)**	**-1.19 (-2.38,0)**
V	**-1.02 (-2.03,-0.01)**	-0.6 (-1.92,0.71)
Zn	**-0.83 (-1.58,-0.08)**	-0.56 (-1.57,0.45)
TRAP	**-1.06 (-2.07,-0.06)**	-0.94 (-2.29,0.4)
Total PM2.5	**-1.51 (-2.73,-0.3)**	-1.17 (-2.56,0.22)

Figures

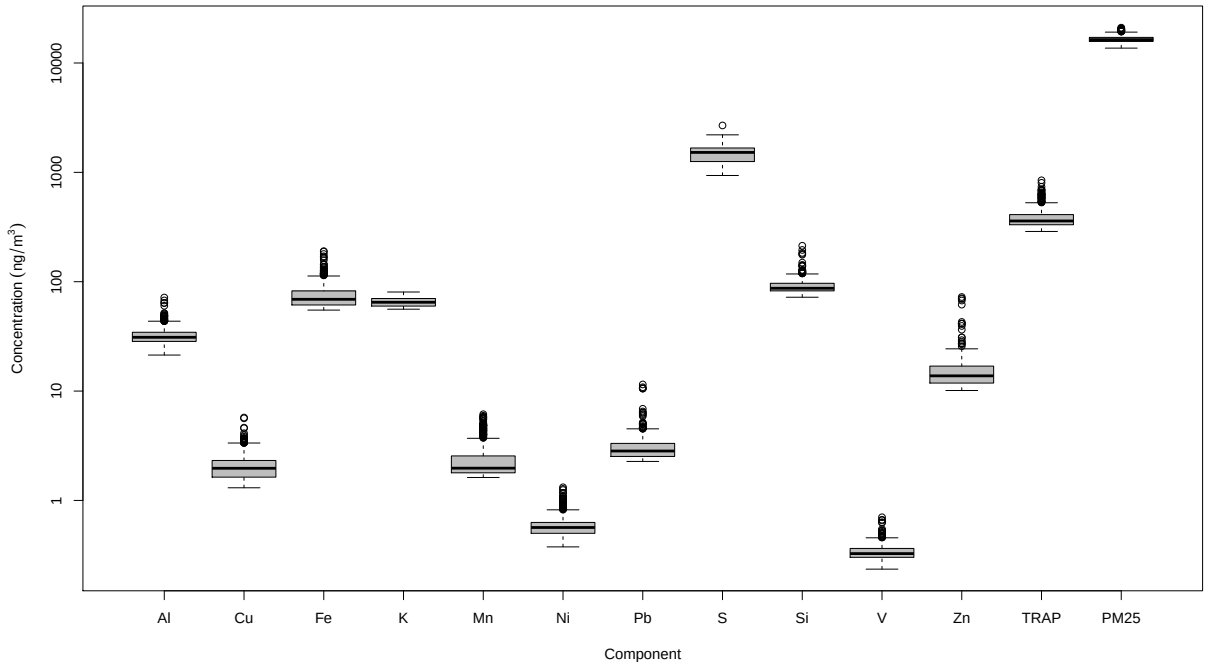

Figure 4.1: Boxplot of estimated elemental concentrations, TRAP, and total PM2.5 for the CCAAPS cohort.

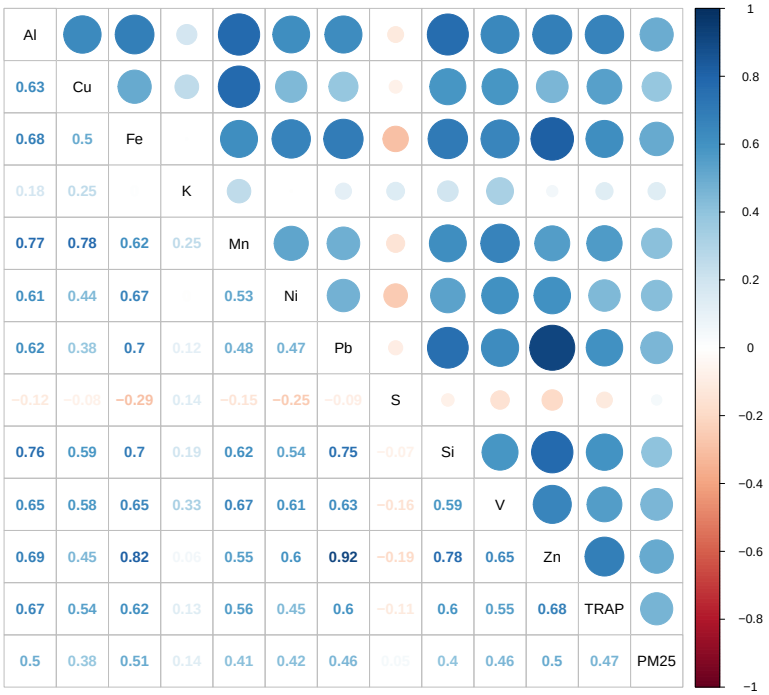

Figure 4.2: Spearman correlation matrix of estimated elemental concentrations, TRAP, and total PM2.5 for the CCAAPS cohort. A darker blue or red and larger circle in the upper triangle of the grid corresponds to a larger positive or negative Spearman's rho statistic shown in the lower triangle of the grid.

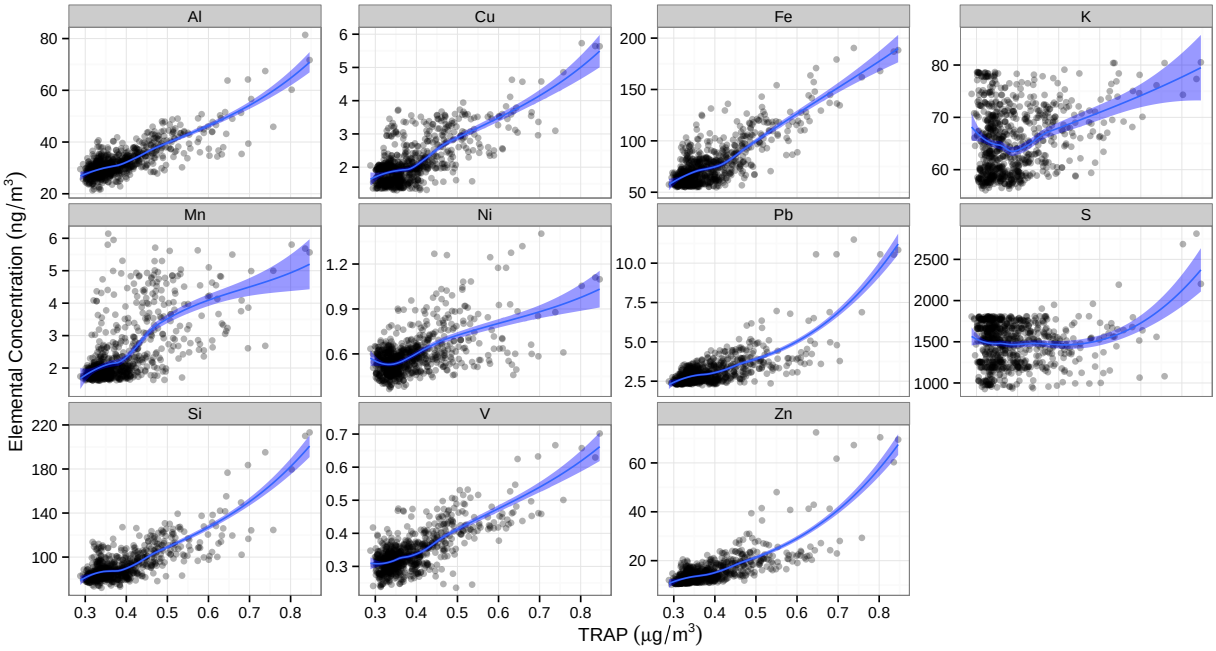

Figure 4.3: The estimated exposure of the CCAAPS cohort to TRAP plotted in relation to each PM component exposure estimate. The overlay is a local second degree polynomial regression line fit and 95% confidence interval spanned over 75% of the closest data points.

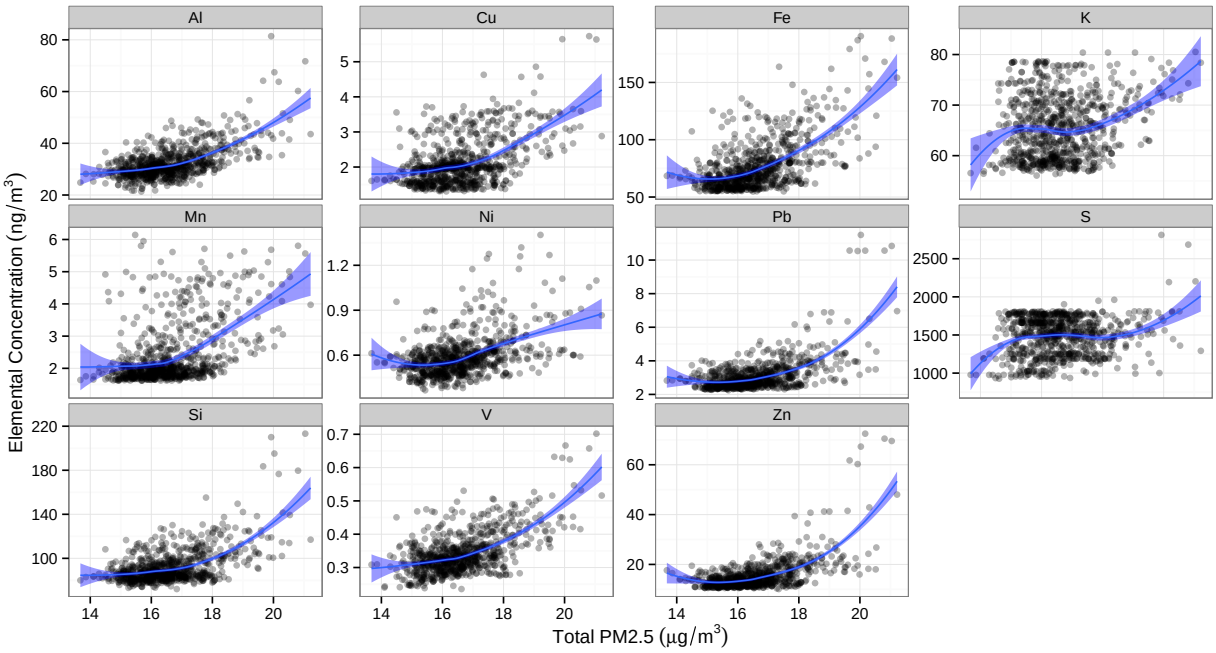

Figure 4.4: The estimated exposure of the CCAAPS cohort to total PM2.5 plotted in relation to each PM component exposure estimate. The overlay is a local second degree polynomial regression line fit and 95% confidence interval spanned over 75% of the closest data points.

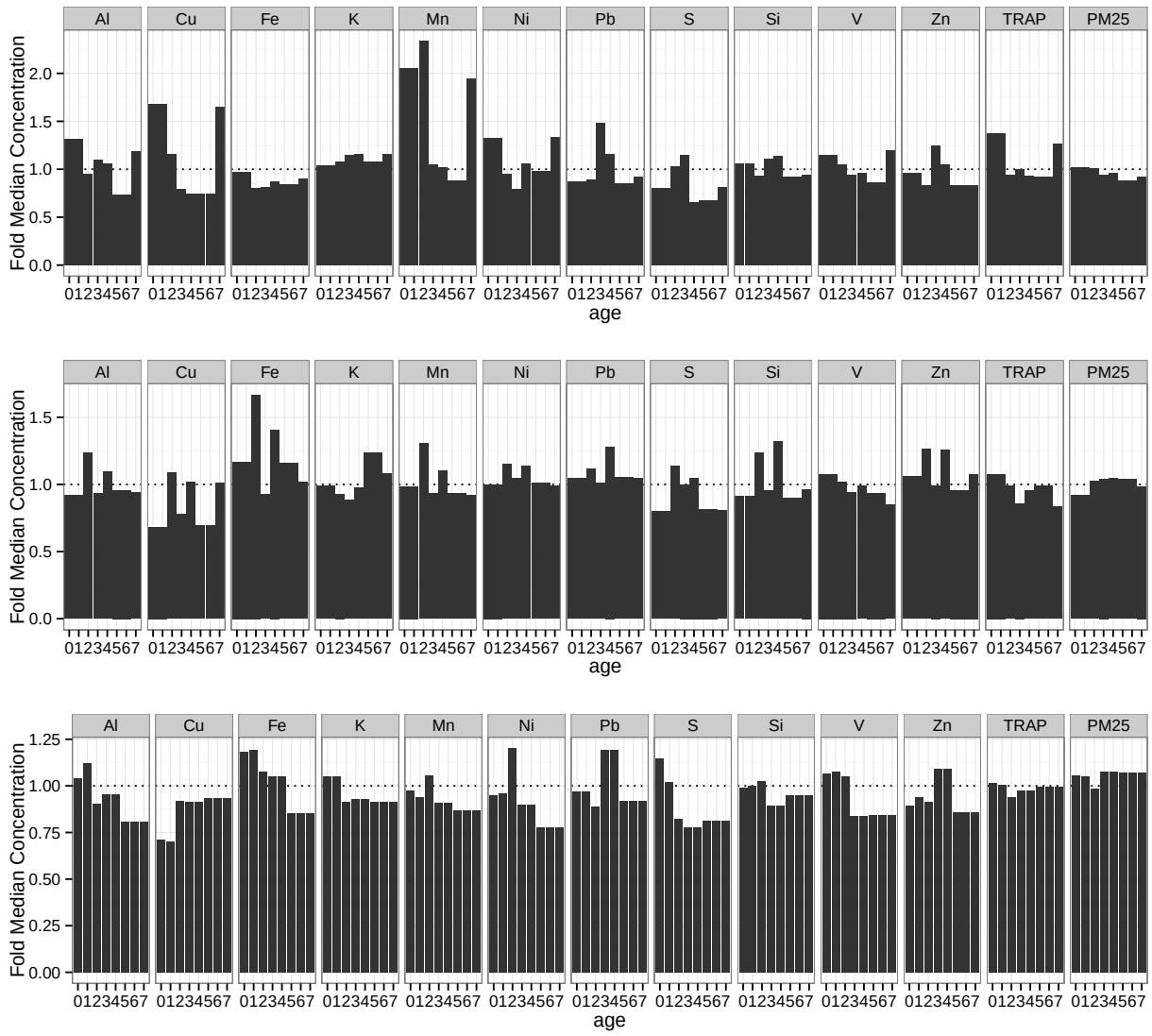

Figure 4.5: The exposure assessment for PM components from birth through age seven for three subjects who moved three times. Each concentration is plotted as the fold of the overall median concentration for the CCAAPS cohort.

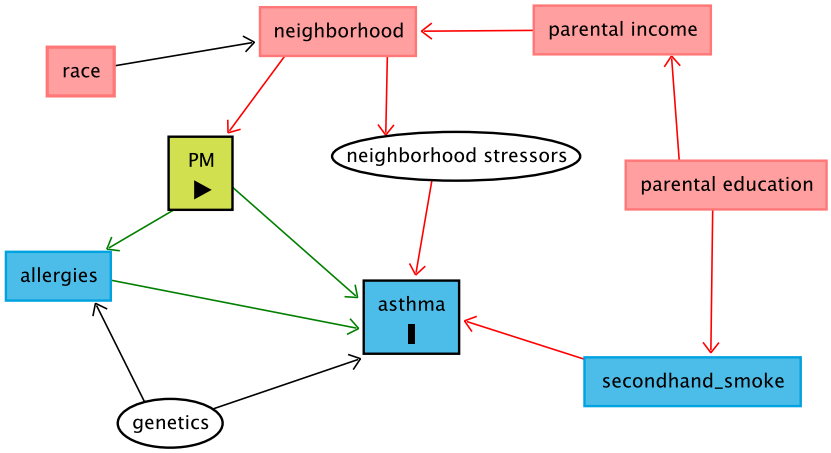

Figure 4.6: A causal diagram for the development of asthma with PM as the primary effect. Here, race, neighborhood, parental income, parental education are all ancestors of both the exposure and the outcome. Allergies and secondhand smoke exposure are ancestors of only the outcome. Neighborhood stressors and genetics are unobserved variables.

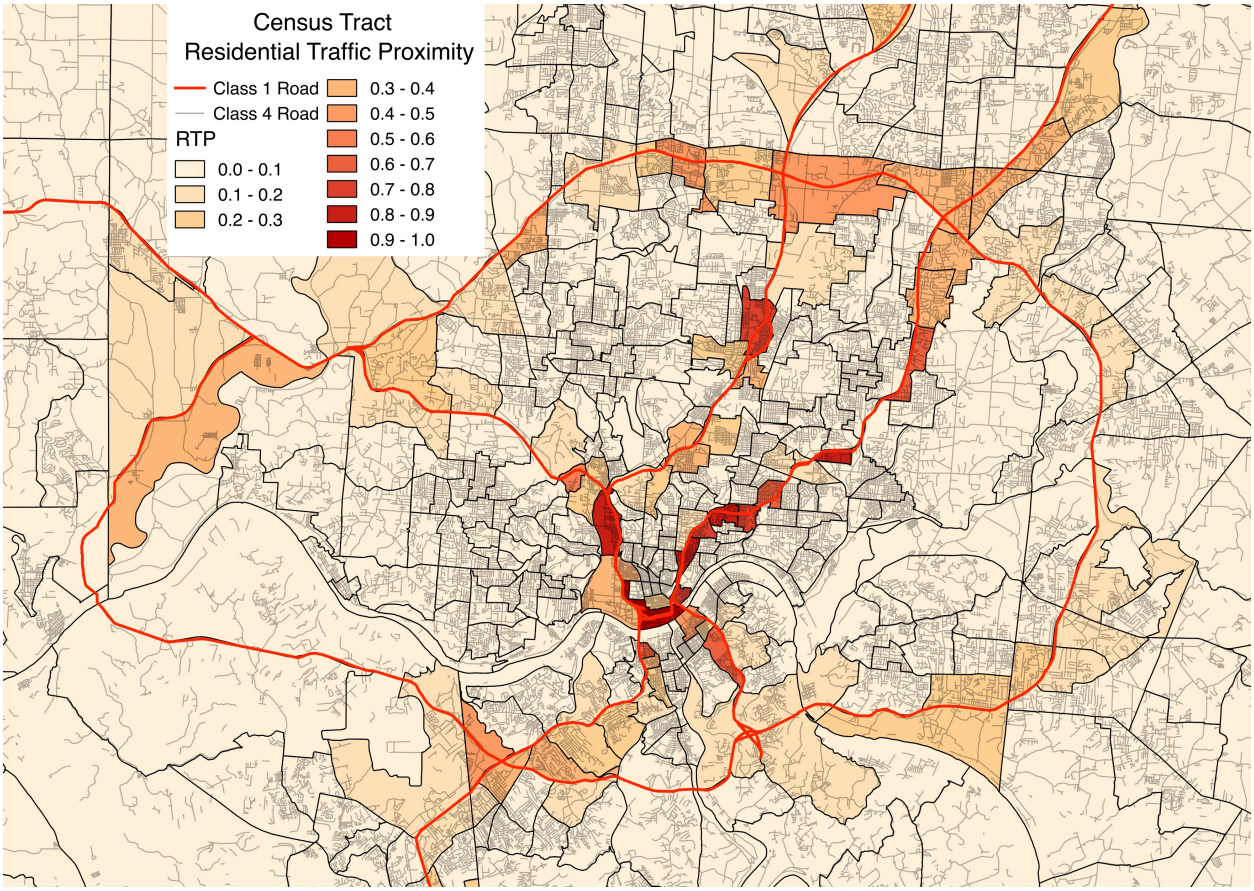

Figure 4.7: Residential Traffic Proximity (RTP) measurement for each census tract in the Greater Cincinnati area along with the Class 1 and Class 4 roadways. RTP was calculated as the fraction of Class 4 roads that were located within 400 meters of a Class 1 road.

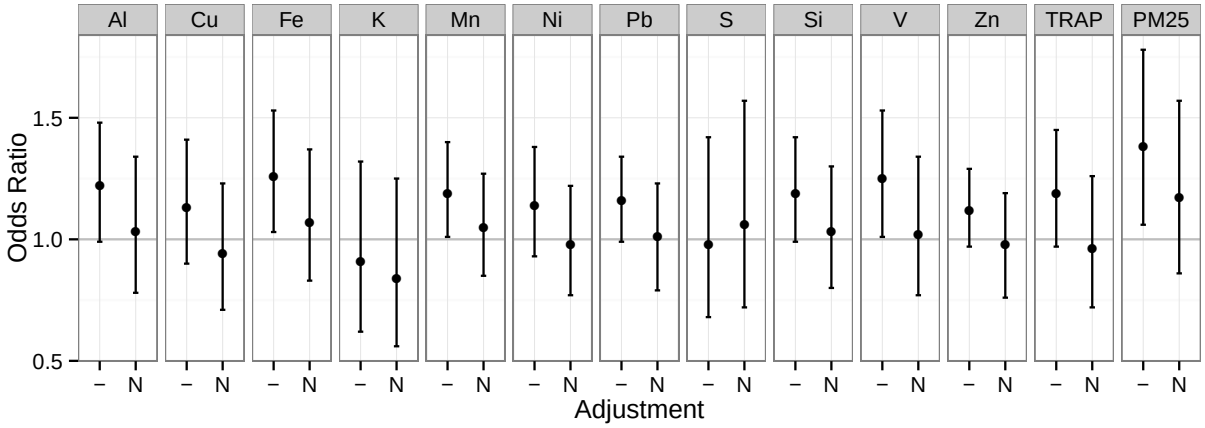

Figure 4.8: Odds Ratios for development of Asthma at age 7 and interquartile increase in exposure for each PM component. Models are unadjusted (−) or adjusted for neighborhood characteristics (N).

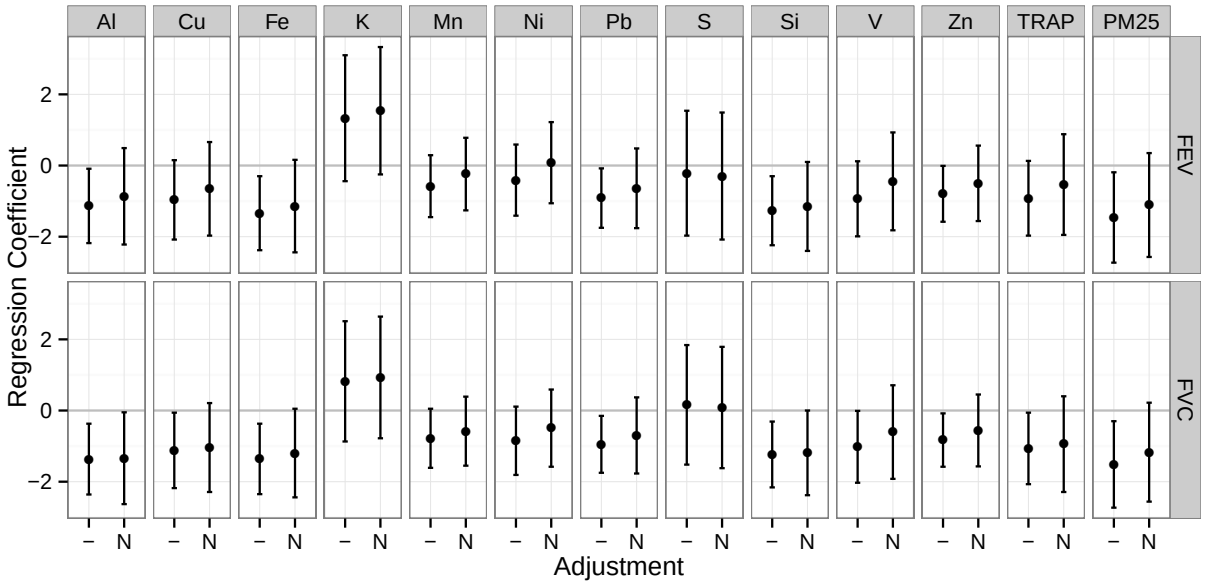

Figure 4.9: Regression coefficients for percent predicted FEV_1 and FVC at age 7 and interquartile increase in exposure for each PM component. Models are unadjusted (−) or adjusted for neighborhood characteristics (N).

Chapter 5

Discussion

In a two stage model that associates spatial pollutants with health effects, it is often standard practice to measure exposures at different locations than are needed for health analysis. Usually, the exposure model is selected based on accuracy, exposures are identified as known and are plugged into a health model disregarding measurement error. However, more accurate exposure prediction does not necessarily improve health effect estimates [94] and ignoring measurement error introduces bias in the estimation of health effects [95, 96]. This measurement error has been broken down into two components [97]: (1) a Berkson-like component that is due to the smoothing of the exposure surface and (2) a classical-like component from the variability in estimating exposure model parameters. This classical-like error differs from classical error in linear regression because it is heteroskedastic and correlated across study subjects. A methodology that corrects for finite-sample bias and correctly estimates standard errors has been implemented [97] which involves an asymptotic bias correction and nonparametric bootstrap. Because this framework is not viable without uncertainty estimates for exposures, the application of the infinitesimal jackknife to random forest to estimate the prediction variance of LURF models is a valuable contribution. Comparing LURF to LUR models within this framework is a promising avenue for future research.

References

[1] David Briggs. The role of gis: coping with space (and time) in air pollution exposure assessment. *Journal of Toxicology and Environmental Health, Part A*, 68(13-14):1243–1261, 2005.

[2] Alexander Kolovos, Andre Skupin, Michael Jerrett, and George Christakos. Multi-perspective analysis and spatiotemporal mapping of air pollution monitoring data. *Environmental science & technology*, 44(17):6738–6744, 2010.

[3] Patrick H Ryan, Grace K LeMasters, Pratim Biswas, Linda Levin, Shaohua Hu, Mark Lindsey, David I Bernstein, James Lockey, Manuel Villareal, Gurjit K Khurana Hershey, et al. A comparison of proximity and land use regression traffic exposure models and wheezing in infants. *Environmental health perspectives*, pages 278–284, 2007.

[4] Sarah B Henderson, Bernardo Beckerman, Michael Jerrett, and Michael Brauer. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental science & technology*, 41(7):2422–2428, 2007.

[5] Saori Kashima, Takashi Yorifuji, Toshihide Tsuda, and Hiroyuki Doi. Application of land use regression to regulatory air quality data in japan. *Science of the Total Environment*, 407(8):3055–3062, 2009.

[6] Zev Ross, Paul B English, Rusty Scalf, Robert Gunier, Svetlana Smorodinsky, Steve Wall, and Michael Jerrett. Nitrogen dioxide prediction in southern california using

land use regression modeling: potential for environmental health analyses. *Journal of Exposure Science and Environmental Epidemiology*, 16(2):106–114, 2006.

[7] David J Briggs, Susan Collins, Paul Elliott, Paul Fischer, Simon Kingham, Erik Lebret, Karel Pryl, Hans van Reeuwijk, Kirsty Smallbone, and Andre Van Der Veen. Mapping urban air pollution using gis: a regression-based approach. *International Journal of Geographical Information Science*, 11(7):699–718, 1997.

[8] Rob Beelen, Marita Voogt, Jan Duyzer, Peter Zandveld, and Gerard Hoek. Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a dutch urban area. *Atmospheric Environment*, 44(36):4614–4621, 2010.

[9] Patrick H Ryan and Grace K LeMasters. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation Toxicology*, 19(S1):127–133, 2007.

[10] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[11] Alexandre Champendal, Mikhail Kanevski, and Pierre-Emmanuel Huguenot. Air pollution mapping using nonlinear land use regression models. In *Computational Science and Its Applications–ICCSA 2014*, pages 682–690. Springer, 2014.

[12] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[13] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[14] Antonella Zanobetti, Meredith Franklin, Petros Koutrakis, Joel Schwartz, et al. Fine particulate air pollution and its components in association with cause-specific emergency admissions. *Environ Health*, 8(58):58, 2009.

[15] Boris Z Simkhovich, Michael T Kleinman, and Robert A Kloner. Air pollution and cardiovascular injury: epidemiology, toxicology, and mechanisms. *Journal of the American College of Cardiology*, 52(9):719–726, 2008.

[16] Douglas W Dockery. Health effects of particulate air pollution. *Annals of epidemiology*, 19(4):257–263, 2009.

[17] National Research Council (US). Committee on Research Priorities for Airborne Particulate Matter. *Research Priorities for Airborne Particulate Matter: Continuing Research Progress. IV*. National Academies Press, 2004.

[18] Kees de Hoogh, Meng Wang, Martin Adam, Chiara Badaloni, Rob Beelen, Matthias Birk, Giulia Cesaroni, Marta Cirach, Christophe Declercq, Audrius Dedele, et al. Development of land use regression models for particle composition in twenty study areas in europe. *Environmental science & technology*, 47(11):5778–5786, 2013.

[19] Joyce JY Zhang, Liu Sun, Olesya Barrett, Stefania Bertazzon, Fox E Underwood, and Markey Johnson. Development of land-use regression models for metals associated with airborne particulate matter in a north american city. *Atmospheric Environment*, 106:165–177, 2015.

[20] Rob Beelen. Natural cause mortality and long-term exposure to particle components: An analysis of 19 european cohorts within the multi-center escape project. *Environmental health perspectives*, 2015.

[21] Marloes Eeftens, Gerard Hoek, Olena Gruzieva, Anna Mölter, Raymond Agius, Rob Beelen, Bert Brunekreef, Adnan Custovic, Josef Cyrys, Elaine Fuertes, et al. Elemental

composition of particulate matter and the association with lung function. *Epidemiology*, 25(5):648–657, 2014.

[22] Regina Hampel, Annette Peters, Rob Beelen, Bert Brunekreef, Josef Cyrys, Ulf de Faire, Kees de Hoogh, Kateryna Fuks, Barbara Hoffmann, Anke Hüls, et al. Long-term effects of elemental composition of particulate matter on inflammatory blood markers in european cohorts. *Environment international*, 82:76–84, 2015.

[23] Michelle L Bell, Francesca Dominici, Keita Ebisu, Scott L Zeger, and Jonathan M Samet. Spatial and temporal variation in pm2. 5 chemical composition in the united states for health effects studies. *Environmental health perspectives*, pages 989–995, 2007.

[24] Grace K LeMasters, Kimberly Wilson, Linda Levin, Jocelyn Biagini, Patrick Ryan, James E Lockey, Sherry Stanforth, Stephanie Maier, Jun Yang, Jeff Burkle, et al. High prevalence of aeroallergen sensitization among infants of atopic parents. *The Journal of pediatrics*, 149(4):505–511, 2006.

[25] Patrick H Ryan, Grace LeMasters, Jocelyn Biagini, David Bernstein, Sergey A Grinshpun, Rakesh Shukla, Kimberly Wilson, Manuel Villareal, Jeff Burkle, and James Lockey. Is it traffic type, volume, or distance? wheezing in infants living near truck and bus traffic. *Journal of allergy and clinical immunology*, 116(2):279–284, 2005.

[26] Shaohua Hu, Rafael McDonald, Dainius Martuzevicius, Pratim Biswas, Sergey A Grinshpun, Anna Kelley, Tiina Reponen, James Lockey, and Grace LeMasters. Unmix modeling of ambient pm 2.5 near an interstate highway in cincinnati, oh, usa. *Atmospheric environment*, 40:378–395, 2006.

[27] Ronald C Henry. Multivariate receptor modeling by n-dimensional edge detection. *Chemometrics and intelligent laboratory systems*, 65(2):179–189, 2003.

[28] Ronald C Henry. Unmix version 2 manual. *Prepared for the US Environmental Protection Agency*, 2000.

[29] 2002 tiger/line files [machine-readable data files]. ua 2003. us department of commerce, geography division, us census bureau.

[30] 2002 tiger/line files technical documentation. ua 2003. us department of commerce, geography division, us census bureau.

[31] 2000 census of population and housing. ua 2004. us census bureau.

[32] Cole Brokamp, Grace LeMasters, and Patrick Ryan. Residential mobility impacts exposure assessment and community socioeconomic characteristics in longitudinal epidemiology studies. *Journal of Exposure Science and Environmental Epidemiology*, 00:1–7, 2016.

[33] Eric F Vermote, Didier Tanré, Jean Luc Deuzé, Maurice Herman, and J-J Morcette. Second simulation of the satellite signal in the solar spectrum, 6s: An overview. *Geoscience and Remote Sensing, IEEE Transactions on*, 35(3):675–686, 1997.

[34] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[35] Sara Alvarez, Ramon Diaz-Uriarte, Ana Osorio, Alicia Barroso, Lorenzo Melchor, Maria Fe Paz, Emiliano Honrado, Raquel Rodríguez, Miguel Urioste, Laura Valle, et al. A predictor based on the somatic genomic changes of the brca1/brca2 breast cancer tumors identifies the non-brca1/brca2 tumors with brca1 promoter hypermethylation. *Clinical Cancer Research*, 11(3):1146–1153, 2005.

[36] Grant Izmirlian. Application of the random forest classification algorithm to a seldi-tof proteomics study in the setting of a cancer prevention trial. *Annals of the New York Academy of Sciences*, 1020(1):154–174, 2004.

[37] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical meth-

ods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003.

[38] Erik C Gunther, David J Stone, Robert W Gerwien, Patricia Bento, and Melvyn P Heyes. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proceedings of the national academy of sciences*, 100(16):9608–9613, 2003.

[39] Michael Z Man, Greg Dyson, Kjell Johnson, and Birong Liao. Evaluating methods for classifying expression data. *Journal of biopharmaceutical statistics*, 14(4):1065–1084, 2004.

[40] Holger Schwender, Manuela Zucknick, Katja Ickstadt, Hermann M Bolt, GENICA network, et al. A pilot study on the application of statistical classification procedures to molecular epidemiological data. *Toxicology letters*, 151(1):291–299, 2004.

[41] Vladimir Svetnik, Andy Liaw, Christopher Tong, and Ting Wang. Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In *Multiple Classifier Systems*, pages 334–343. Springer, 2004.

[42] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[43] Roger Bivand, Tim Keitt, and Barry Rowlingson. *rgdal: Bindings for the Geospatial Data Abstraction Library*, 2014. R package version 0.8-16.

[44] Roger Bivand and Colin Rundel. *rgeos: Interface to Geometry Engine - Open Source (GEOS)*, 2014. R package version 0.3-4.

[45] RS Bivand, EJ Pebesma, and V Gomez-Rubio. Classes and methods for spatial data in r. *R News*, 5(9), 2005.

[46] Bénédicte Jacquemin. Ambient air pollution and adult asthma incidence in six european cohorts (escape). *Environmental health perspectives*, 2015.

[47] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

[48] Stefan Wager. Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*, 2014.

[49] Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

[50] L Jaeckel. The infinitesimal jackknife, memorandum. Technical report, MM 72-1215-11, Bell Lab. Murray Hill, NJ, 1972.

[51] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[52] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

[53] Helmut Strasser and Christian Weber. On the asymptotic theory of permutation statistics. 1999.

[54] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[55] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

[56] Regina Rückerl, Alexandra Schneider, Susanne Breitner, Josef Cyrys, and Annette Peters. Health effects of particulate air pollution: a review of epidemiological evidence. *Inhalation toxicology*, 23(10):555–592, 2011.

[57] A Campbell, M Oldham, A Becaria, SC Bondy, D Meacher, C Sioutas, C Misra, LB Mendez, and M Kleinman. Particulate matter in polluted air may increase biomarkers of inflammation in mouse brain. *Neurotoxicology*, 26(1):133–140, 2005.

[58] S Ebersviller, K Lichtveld, Kenneth G Sexton, J Zavala, Y-H Lin, I Jaspers, and HE Jeffries. Gaseous vocs rapidly modify particulate matter and its biological effects–part 1: Simple vocs and model pm. *Atmospheric Chemistry and Physics*, 12(24):12277–12292, 2012.

[59] Johanna Lepeule, Francine Laden, Douglas Dockery, and Joel Schwartz. Chronic exposure to fine particles and mortality: an extended follow-up of the harvard six cities study from 1974 to 2009. *Environmental health perspectives*, 120(7):965, 2012.

[60] Michelle C Turner, Daniel Krewski, C Arden Pope III, Yue Chen, Susan M Gapstur, and Michael J Thun. Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers. *American journal of respiratory and critical care medicine*, 184(12):1374–1381, 2011.

[61] E Samoli, PT Nastos, AG Paliatsos, K Katsouyanni, and KN Priftis. Acute effects of air pollution on pediatric asthma exacerbation: evidence of association and effect modification. *Environmental Research*, 111(3):418–424, 2011.

[62] Robert D Brook and Sanjay Rajagopalan. Particulate matter, air pollution, and blood pressure. *Journal of the American Society of Hypertension*, 3(5):332–350, 2009.

[63] Joel Schwartz and Douglas W Dockery. Increased mortality in philadelphia associated with daily air pollution concentrations. *American review of respiratory disease*, 145(3):600–604, 1992.

[64] C Arden Pope III, Richard T Burnett, Michael J Thun, Eugenia E Calle, Daniel Krewski, Kazuhiko Ito, and George D Thurston. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama*, 287(9):1132–1141, 2002.

[65] HR Anderson, RW Atkinson, SA Bremner, J Carrington, and J Peacock. Quantitative systematic review of short term associations between ambient air pollution (particulate matter, ozone, nitrogen dioxide, sulphur dioxide and carbon monoxide), and mortality and morbidity. *Report to Department of Health revised following first review*, 2007.

[66] Francesca Dominici, Aidan McDermott, Scott L Zeger, and Jonathan M Samet. Airborne particulate matter and mortality: timescale effects in four us cities. *American Journal of Epidemiology*, 157(12):1055–1065, 2003.

[67] Francesca Dominici, Aidan McDermott, Michael Daniels, Scott L Zeger, and Jonathan M Samet. Revised analyses of the national morbidity, mortality, and air pollution study: mortality among residents of 90 cities. *Journal of Toxicology and Environmental Health, Part A*, 68(13-14):1071–1092, 2005.

[68] Sandra Mallone, Massimo Stafoggia, Annunziata Faustini, Gian Paolo Gobbi, Achille Marconi, and Francesco Forastiere. Saharan dust and associations between particulate matter and daily mortality in rome, italy. *Environmental health perspectives*, 119(10):1409, 2011.

[69] Roger D Peng, Michelle L Bell, Alison S Geyh, Aidan McDermott, Scott L Zeger, Jonathan M Samet, and Francesca Dominici. Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental health perspectives*, 117(6):957, 2009.

[70] Joel Schwartz, Francine Laden, and Antonella Zanobetti. The concentration-response relation between pm (2.5) and daily deaths. *Environmental health perspectives*, 110(10):1025, 2002.

[71] Particulate Matter Air Pollution. Cardiovascular disease: An update to the scientific statement from the american heart association brook, robert d. *Rajagopalan, Sanjay*, pages 2331–2378.

[72] ZS Boritz, TK Michael, and AK Robert. Air pollution and cardiovascular injury. *J. Am. Coll. Cardiol*, 52:719–726, 2008.

[73] World Health Organization. Regional Office for Europe and World Health Organization. *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*. World Health Organization, 2006.

[74] Roger D Peng, Michelle L Bell, Alison S Geyh, Aidan McDermott, Scott L Zeger, Jonathan M Samet, Francesca Dominici, et al. Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environ Health Perspect*, 117(6):957–963, 2009.

[75] Michelle L Bell, Keita Ebisu, Roger D Peng, Jonathan M Samet, and Francesca Dominici. Hospital admissions and chemical composition of fine particle air pollution. *American journal of respiratory and critical care medicine*, 179(12):1115–1120, 2009.

[76] Bart Ostro, Wen-Ying Feng, Rachel Broadwin, Shelley Green, and Michael Lipsett. The effects of components of fine particulate air pollution on mortality in california: results from calfine. *Environmental health perspectives*, pages 13–19, 2007.

[77] Meredith Franklin, Petros Koutrakis, and Joel Schwartz. The role of particle composition on the association between pm2. 5 and mortality. *Epidemiology (Cambridge, Mass.)*, 19(5):680, 2008.

[78] Michelle L Bell, Kathleen Belanger, Keita Ebisu, Janneane F Gent, Hyung Joo Lee, Petros Koutrakis, and Brian P Leaderer. Prenatal exposure to fine particulate matter and birth weight: variations by particulate constituents and sources. *Epidemiology (Cambridge, Mass.)*, 21(6):884, 2010.

[79] Sander Greenland, Judea Pearl, and James M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

[80] Ian Shrier and Robert W Platt. Reducing bias through directed acyclic graphs. *BMC medical research methodology*, 8(1):70, 2008.

[81] Felix Elwert. Graphical causal models. In *Handbook of causal analysis for social research*, pages 245–273. Springer, 2013.

[82] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

[83] Sander Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003.

[84] Sander Greenland and Hal Morgenstern. Confounding in health research. *Annual review of public health*, 22(1):189–212, 2001.

[85] Daniel Westreich and Sander Greenland. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4):292–298, 2013.

[86] Tyler J Vanderweele and Nancy Staudt. Causal diagrams for empirical legal research: a methodology for identifying causation, avoiding bias and interpreting results. *Law, probability and risk*, 10(4):329–354, 2011.

[87] Martin R Miller, JATS Hankinson, V Brusasco, F Burgos, R Casaburi, A Coates, R Crapo, P Enright, CP Van der Grinten, P Gustafsson, et al. Standardisation of spirometry. *Eur respir J*, 26(2):319–38, 2005.

[88] Xiaobin Wang, Douglas W Dockery, David Wypij, Diane R Gold, Frank E Speizer, James H Ware, and Benjamin G Ferris Jr. Pulmonary function growth velocity in

children 6 to 18 years of age. *American Review of Respiratory Disease*, 148:1502–1502, 1993.

[89] RO Crapo, R Casaburi, AL Coates, PL Enright, JL Hankinson, CG Irvin, NR MacIntyre, RT McKay, JS Wanger, SD Anderson, et al. Guidelines for methacholine and exercise challenge testing-1999. this official statement of the american thoracic society was adopted by the ats board of directors, july 1999. *American journal of respiratory and critical care medicine*, 161(1):309, 2000.

[90] Tiina Reponen, Stephen Vesper, Linda Levin, Elisabet Johansson, Patrick Ryan, Jeffery Burkle, Sergey A Grinshpun, Shu Zheng, David I Bernstein, James Lockey, et al. High environmental relative moldiness index during infancy as a predictor of asthma at 7 years of age. *Annals of Allergy, Asthma & Immunology*, 107(2):120–126, 2011.

[91] Basile Chaix, Cinira Leal, and David Evans. Neighborhood-level confounding in epidemiologic studies: unavoidable challenges, uncertain solutions. *Epidemiology*, 21(1):124–127, 2010.

[92] Jeffrey D Morenoff, James S House, Ben B Hansen, David R Williams, George A Kaplan, and Haslyn E Hunte. Understanding social disparities in hypertension prevalence, awareness, treatment, and control: the role of neighborhood context. *Social science & medicine*, 65(9):1853–1866, 2007.

[93] Robert J Sampson, Stephen W Raudenbush, and Felton Earls. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328):918–924, 1997.

[94] Adam A Szpiro, Christopher J Paciorek, and Lianne Sheppard. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology (Cambridge, Mass.)*, 22(5):680, 2011.

[95] Alexandros Gryparis, Christopher J Paciorek, Ariana Zeka, Joel Schwartz, and Brent A Coull. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10(2):258–274, 2009.

[96] Adam A Szpiro, Lianne Sheppard, and Thomas Lumley. Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12(4):610–623, 2011.

[97] Adam A Szpiro and Christopher J Paciorek. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*, 24(8):501–517, 2013.