Adapting Statistical Learning Methods for Spatial Applications

Travis Hee Wai

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Adam Szpiro, Chair

Ali Shojaie

Paul Sampson

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Adapting Statistical Learning Methods for Spatial Applications

Travis Hee Wai

Chair of the Supervisory Committee:
Dr. Adam Szpiro
Biostatistics

In this dissertation, we develop new principled applications of statistical learning methods in spatial applications. In the first chapter, we consider a modified regression tree approach allowing for spatial correlation for applications in spatially indexed datasets. In the second chapter, we consider incorporating penalized regression estimators into universal kriging models. In the third and final chapter, we propose a class of flexible, additive regression tree models for joint estimation across multiple domains of interest.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables v

Chapter 1: Introduction 1

Chapter 2: Random Spatial Forests 4
2.1 Introduction 4
2.2 Methods 5
2.3 Simulation Study 15
2.4 Application to Sub-Species of $PM_{2.5}$ 19

Chapter 3: Penalized Universal Kriging 30
3.1 Introduction 30
3.2 Methods 33
3.3 Simulation Study 40
3.4 Comparison of Methods for Pollutants across Continental United States in 2009-2010 47

Chapter 4: Multi-City Data-Enriched Regression Trees 62
4.1 Introduction 62
4.2 Methods 66
4.3 Simulation Studies 76
4.4 NOx Concentrations in MESA Air Communities 79

Chapter 5: Discussion 101
5.1 Summary 101
5.2 Future Work and Extensions 102

Appendix A: Appendix 117
A.1 Proof of Equivalence for Update of Ω^{k+1} 117

LIST OF FIGURES

Figure Number — Page

2.1 Simulation Study comparing different approaches to estimating additive models of Random Forests and Spatial Smoothing 18
2.2 Predicted EC concentration across the continental United States 22
2.3 Predicted OC concentration across the continental United States 23
2.4 Predicted Si concentration across the continental United States 24
2.5 Predicted S concentration across the continental United States 25
2.6 Random Forests and Spatial Smoothing Components for EC 26
2.7 Random Forests and Spatial Smoothing Components for OC 27
2.8 Random Forests and Spatial Smoothing Components for S 28
2.9 Random Forests and Spatial Smoothing Components for Si 29

3.1 Example Variogram including Geographic Covariates 31
3.2 Simulation Study comparing Universal Kriging Model Selection and Regularization Approaches 43
3.3 Comparing Run Time of Penalized and Likelihood Methods for Universal Kriging 46
3.4 Scatter Plots of Predicted vs Observed EC Concentrations for Penalized Universal Kriging 50
3.5 Scatter Plots of Predicted vs Observed OC Concentrations for Penalized Universal Kriging 51
3.6 Scatter Plots of Predicted vs Observed S Concentrations for Penalized Universal Kriging 52
3.7 Scatter Plots of Predicted vs Observed Si Concentrations for Penalized Universal Kriging 53
3.8 Scatter Plots of Predicted vs Observed NO_2 Concentrations for Penalized Universal Kriging 54
3.9 Scatter Plots of Predicted vs Observed $PM_{2.5}$ Concentrations for Penalized Universal Kriging 55

3.10 Predicted EC across the continental United States using Penalized Universal Kriging 56
3.11 Predicted OC across the continental United States using Penalized Universal Kriging 57
3.12 Predicted S across the continental United States using Penalized Universal Kriging 58
3.13 Predicted Si across the continental United States using Penalized Universal Kriging 59
3.14 Predicted NO_2 across the continental United States using Penalized Universal Kriging 60
3.15 Predicted $PM_{2.5}$ across the continental United States using Penalized Universal Kriging 61
4.1 Observed NO_x Concentrations in MESA Air Cities 63
4.2 Example of a Multi-City Regression Tree 67
4.3 Simulation Study comparing Prediction Accuracy when Pooling Estimates . 78
4.4 Percent change in cross-validated RMSPE across MESA Air cities 84
4.5 Scatter Plots of NO_x Predictions in New York City 87
4.6 Scatter Plots of NO_x Predictions in Los Angeles 88
4.7 Scatter Plots of NO_x Predictions in Chicago 89
4.8 Scatter Plots of NO_x Predictions in Baltimore 90
4.9 Scatter Plots of NO_x Predictions in St. Paul 91
4.10 Scatter Plots of NO_x Predictions in Winston-Salem 92
4.11 Scatter Plots of NO_x Predictions in Rockland County 93
4.12 Cross-Validated NO_x Residuals in New York 94
4.13 Cross-Validated NO_x Residuals in Los Angeles 95
4.14 Cross-Validated NO_x Residuals in Chicago 96
4.15 Cross-Validated NO_x Residuals in Baltimore 97
4.16 Cross-Validated NO_x Residuals in St. Paul 98
4.17 Cross-Validated NO_x Residuals in Winston-Salem 99
4.18 Cross-Validated NO_x Residuals in Rockland County 100

LIST OF TABLES

Table Number Page

2.1 Cross-Validated Estimates of RMSPE for Random Spatial Forests on components of $PM_{2.5}$ 20

3.1 Cross-Validated RMSPE for Penalized Universal Kriging Models on EC, OC, Si, S, $PM_{2.5}$, and NO_2 48

4.1 Cross-Validated estimates of RMSPE for individual and multi-city models for NO_x in MESA Air cities 83

ACKNOWLEDGMENTS

DEDICATION

Chapter 1

INTRODUCTION

A fundamental problem in environmental epidemiology studies on the association of air pollution exposure with health outcomes is identifying exposure levels for each individual in a cohort study. Measurements are not made at each study participants place of residence, therefore individual-specific exposure levels are estimated using observations from regulatory monitors.

Estimation of the underlying spatial process often focuses on spatial smoothing by kriging [54], where the observed surface is modeled as a realization of a Gaussian process with spatial correlation that depends on the distance between observed sites. Statistical analyses typically incorporate geographic covariates into kriging models, which is known as universal kriging; and the use of these covariates has been shown to work well in practice [33]. This is structured as an additive model,

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\nu}(\mathbf{s}),$$

where the observed process, $\mathbf{Y}(\mathbf{s})$, observed at n locations, $\mathbf{s} = \{s_1, \ldots, s_n\}$, is modeled a linear function of covariates, $\mathbf{X}$, which are subject to variation from the mean, $\mathbf{X}\boldsymbol{\beta}$, by a realization of a spatial process, $\boldsymbol{\nu}(\mathbf{s})$.

Technological advancements have increased the ease and scope of data collection for these spatially-indexed covariates. For example, geographic information system (GIS) covariates from programs such as ArcGIS provide users with hundreds of covariates describing proximity variables to significant geographical features and buffer variables measuring geographic features within some radius. In addition, researchers have examined including additional sources such as satellite data [86], traffic data [71], and meteorology data [3].

In order to leverage these data sources in spatial applications, many studies have ex-

amined applying statistical learning methods instead of universal kriging to leverage these high-dimensional covariate sets [35, 11, 2]. However, some studies suggest that applying statistical learning methods do not yield any noticeable advantages over traditional geostatistical approaches such as kriging [7, 21]. Although statistical learning methods are designed to efficiently use large sets of covariates and observations, they do not allow for spatial correlation, which can be a significant factor in modeling the underlying process. In these applications, spatial correlation is a statistical characterization of spatial variation not explained by the covariates, for example topography, climatological, and meteorological patterns, that are difficult to model explicitly. In order to maximize variability explained through the additive model, it is desirable to leverage geographic covariates by using statistical learning methods to model systematic variation which cannot be modeled in the spatial process. By ignoring the spatial correlation, statistical learning methods may model spatial structure that could have otherwise been included in the spatial process, and little has been done to explore the degree to which the predictive power of these models can be improved by incorporating spatial information into the statistical learning techniques themselves. In this dissertation, we examine using statistical learning techniques with spatially indexed data and propose modifications to allow for spatial correlation for principled applications of statistical learning in spatially indexed datasets.

In Chapter 2, we consider a principled approach to apply the random forests algorithm in spatial applications by building regression trees adjusted for spatial correlation. Our main contribution is the development of a computationally efficient tree building algorithm which selects each split of the tree adjusting for spatial correlation. We evaluate two different approaches for estimation of random spatial forests, a pseudo-likelihood approach combining random forests with kriging and a non-parametric version for a general class of spatial smoothers. We show improved prediction accuracy of our method compared to existing two-step approaches combining random forests and kriging across a range of numerical simulations and demonstrate its performance on elemental carbon, organic carbon, silicon, and sulfur measurements across the continental United States from 2009-2010.

In Chapter 3, we examine the large p problem in spatial applications, where many covariates are collected, such as satellite and mapping data, to be used in a universal kriging model. A popular approach for dealing with large numbers of covariates in linear models is penalized regression, but this approach has seen little traction in spatial statistics due to complications by the spatial covariance. By taking advantage of the relationship between ridge regression and linear mixed models, we present a penalized regression framework for universal kriging models that scales in both n and p, simplifies estimation as the target of optimization is convex, and can be applied using existing software. We demonstrate improved prediction accuracy and computational efficiency across a variety of scenarios in simulations and show that these methods do as well or better than existing approaches across a wide range of pollutants collected across the continental United States in 2009-2010.

In Chapter 4, we take a look at the Multi-Ethnic Study of Atherosclerosis (MESA), which began in 2004 and included participants in six metropolitan areas across the United States: Baltimore, MD; Chicago, IL; Winston Salem, NC; Los Angeles, CA; New York, NY; and St. Paul, MN. A snapshot campaign was conducted in each city during 2006-2007 to estimate pollutant concentrations for each individual involved in the study, but constructing a single model using observations across all cities is difficult in practice as air pollutant concentrations can vary widely from city to city. Constructing a separate model for each city individually is an unsatisfying solution, as factors contributing to air pollution are expected to be similar across cities. We propose multi-city data-enriched regression trees, an approach to building additive regression tree models that combine information across separate domains to select the structure of the regression tree and the associated contrasts. We demonstrate by simulation the scenarios under which combining information across cities improves prediction accuracy, and show on MESA NO_x measurements from Winter 2006-2007 that our multi-city data-enriched regression trees result in superior prediction accuracy than individual city models alone.

Chapter 2
RANDOM SPATIAL FORESTS

2.1 Introduction

Random forests [9] has been shown to be effective for prediction in high-dimensional scenarios and some have examined applying it to spatially-indexed covariates in order to estimate spatial processes [35, 11]. However, random forests does not incorporate information about the spatial locations of the data, and some studies suggest random forests alone does not compare favorably to traditional geostatistical approaches such as kriging [21, 7]. In order to add spatial information to the random forests estimate, two-step approaches have been proposed where a spatial smoother is fit to the residuals from the random forests estimate. Rolf et al. [68] provide a simple set of conditions under which this approach improves estimation accuracy and it has been shown to perform better than either using either method alone in practice [51].

Two-step optimization approaches are an inefficient optimization scheme. Combining random forests and kriging can be viewed as an additive model,

$$\mathbf{Y}(\mathbf{s}) = f(\mathbf{X}) + \boldsymbol{\nu}(\mathbf{s}),$$

with $f(\mathbf{X})$ a random forests estimate of the spatially-indexed covariates and a spatially correlated mean-zero stochastic term, $\boldsymbol{\nu}(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where the spatial covariance is known up to parameters $\boldsymbol{\theta}$. In these applications, the stochastic process is a statistical characterization of spatial variation not explained by the covariates, for example topography, climatological, and meteorological patterns, that are difficult to model explicitly. In order to maximize variability explained through the additive model, it is desirable to use random forests to model systematic variation which cannot be modeled in the spatial process. By

ignoring the spatial correlation, random forests may model spatial structure that could have otherwise been included in the spatial process. Little has been done to explore the degree to which the predictive power of these models can be improved by incorporating spatial information into random forests itself. Hengl et al. [31] proposed random forests for spatial data, where they explored adding geographic proximity as a covariate before applying the random forests algorithm but only found similar prediction accuracy to the two-step random forests kriging approach. We emphasize that our goal is to use random forests to utilize geographic covariates which model variation which could not be modeled by the spatial process, and including geographic proximity as a covariate does not help in that regard.

Our main contribution is a novel algorithm to construct spatially adjusted trees which allow for spatial correlation in sub-$\mathcal{O}(n^3)$ run time, and evaluate two different procedures for constructing random forests estimates from spatially adjusted trees. In Section 2.2, we describe a summary of random forests and universal kriging, describe our modified tree building algorithm allowing for spatial correlation, and examine different approaches to constructing random forests estimates from spatially adjusted trees. In Section 2.3, we provide simulation results demonstrating the advantage of our approach over two-step estimation strategies. In Section 2.4, we apply our method to annual average elemental carbon (EC), organic carbon (OC), Silicon (Si), and Sulfur (S) across the continental United States for 2009-2010.

2.2 Methods

2.2.1 Regression Trees

Regression trees have gained popularity for their ability to approximate a wide variety of non-linear functions. Trees are built through an iterative process called recursive binary splitting, which aim to minimize *tree impurity*, traditionally mean-squared error, through a greedy optimization approach. At each iteration, a new terminal node of the tree is created by an exhaustive search selecting the branch which minimizes tree impurity at the current step. Although trees are often described as segmenting the data into terminal nodes by

following decision rules in an attempt to sort observations with similar values together, a regression tree can also be formulated as a linear model.

A tree, $\mathbf{t}(\mathbf{X})$ with k terminal nodes can be written as $\mathbf{t}(\mathbf{X}) = \mathbf{C}^k\boldsymbol{\pi}^k$. Each column t of $\mathbf{C}^k$ is a vector indicating observations in the new terminal node created in iteration t, and its entries are

$$C_{it} = \begin{cases} 1, & X_{j_t} \leq r_{j_t}^t \text{and } i \text{ in terminal node being split} \\ 0 & \text{else} \end{cases}$$

with X_{j_t} the covariate the splitting rule is created on,and $r_{j_t}^t$ the associated cutpoint.

Similarly to a binary tree, each of the k terminal nodes of the tree is encoded by a unique combination of the k columns of $\mathbf{C}^k$ and the tree estimate for that terminal node is a unique linear combination of the corresponding entries of the k vector $\boldsymbol{\pi}^k$. One particular advantage of treating a regression tree as a linear model is that it allows the parameter estimates to be profiled out as:

$$\hat{\boldsymbol{\pi}}^k = \left(\left(\mathbf{C}^k\right)^T \mathbf{C}^k\right)^{-1} \left(\mathbf{C}^k\right)^T \mathbf{Z}(\mathbf{s}),$$

and the total tree impurity,

$$\|\mathbf{Z}(\mathbf{s}) - \mathbf{t}(\mathbf{X})\|_2^2 = \mathbf{Z}(\mathbf{s})^T \left(\mathbf{I}_n - \mathbf{C}^k \left((\mathbf{C}^k)^T\mathbf{C}^k\right)^{-1} (\mathbf{C}^k)^T\right) \mathbf{Z}(\mathbf{s}),$$

depends only on the structure of the tree design matrix $\mathbf{C}^k$ leading to efficient computational methods since $\hat{\boldsymbol{\pi}}^k$ does not need to be optimized for every possible new branch.

2.2.2 Random Forests

While regression trees are able to approximate a wide variety of non-linear functions, they are often not good predictors alone due to their high variance. One method of variance reduction to improve prediction performance is bootstrap aggregation (bagging), an ensemble method of averaging over trees constructed on bootstrapped samples. The bagged estimate can be

written as

$$\hat{\mathbf{f}}(\mathbf{X}) = \frac{1}{B}\sum_{i=1}^{B}\mathbf{t}^i(\mathbf{X}^i),$$

with B is the number of bootstrap replicates, $\mathbf{t}^i(\mathbf{X}^i)$ the tree built on bootstrapped sample i. Optimal variance reduction occurs when each of the trees is independent, but in many cases trees built on bootstrapped sample tend to be similar. In order to minimize correlation between trees, random forests only uses a random subset of the covariates when creating a new terminal node for each tree. The process of bagging over trees in combination with the added randomization used in building a tree enables random forests to approximate a large class of functions while maintaining low generalization error.

2.2.3 Universal Kriging

Universal kriging is a widely used geostatistics method which incorporates spatial information available in the monitoring data with a linear function of the geographic covariates by adding a spatial correlation model. The universal kriging model can be structured as an additive model,

$$\mathbf{Z}(\mathbf{s}) = \mathbf{Y}(\mathbf{s}) + \boldsymbol{\epsilon},$$
$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\nu}(\mathbf{s}),$$

where the underlying process, $\mathbf{Y}(\mathbf{s})$, contains a linear mean structure on the covariates with observations subject to variation from the linear model by a realization of a spatial process $\boldsymbol{\nu}(\mathbf{s})$. The kriging approach models the spatial process, $\boldsymbol{\nu}(\mathbf{s})$, as a realization of a Gaussian process with an independent error term, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I}_n), \boldsymbol{\epsilon} \perp \boldsymbol{\nu}(\mathbf{s})$, thus, $\mathbf{Z}(\mathbf{s}) \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\rho}, \sigma^2, \tau^2))$, where $\mathbf{V}(\boldsymbol{\rho}, \sigma^2, \tau^2) = \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\rho}) + \tau^2\mathbf{I}_n$. We define the set of spatial covariance parameters as $\boldsymbol{\theta} = \{\boldsymbol{\rho}, \sigma^2, \tau^2\}$ and estimate the unknown spatial covariance parameters and fixed effects $\boldsymbol{\theta}, \boldsymbol{\beta}$ by maximization of the log-likelihood.

$$\underset{\boldsymbol{\theta},\boldsymbol{\beta}}{\operatorname{argmax}} -\frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{Z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\boldsymbol{\theta})(\mathbf{Z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta}).$$

For any fixed set of spatial covariance parameters $\boldsymbol{\theta}_0$, $\boldsymbol{\beta}$ which maximizes $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}_0|\mathbf{Z}(\mathbf{s}))$ is easily shown to be the generalized least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\theta}_0)\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\theta}_0)\mathbf{Z}(\mathbf{s}).$$

The method of eliminating $\boldsymbol{\beta}$ from the log likelihood by profiling is commonly employed, and universal kriging models are estimated by optimizing (Eq. 2.1):

$$\begin{aligned}\underset{\boldsymbol{\theta}}{\text{argmax}} \; -\frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{Z}(\mathbf{s}) - \mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\boldsymbol{\theta})(\mathbf{Z}(\mathbf{s}) - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \text{s.t}\,\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{Z}(\mathbf{s}). \end{aligned} \tag{2.1}$$

2.2.4 Efficient Estimation Strategies for Large Spatial Datasets

Optimization of the log-likelihood in a universal kriging model involves inverting the covariance matrix, which is $\mathcal{O}(n^3)$. Recent work in making spatial statistics computationally feasible has relied on clever ways of structuring the covariance matrix to reduce the computational complexity in calculating its inverse [5, 41]. Following Cressie and Johannesson [16], we note that any valid symmetric postive semidefinite spatial correlation matrix can be broken down as $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where $\mathbf{D}$ is a $m \times m$ positive definite matrix. By the Woodbury matrix identity, inverting the spatial covariance matrix is:

$$\mathbf{V}^{-1}(\boldsymbol{\theta}) = \frac{1}{\tau^2}\mathbf{I}_n - \frac{1}{\tau^2}\mathbf{U}\left(\mathbf{U}^T\mathbf{U} + \frac{\tau^2}{\sigma^2}\mathbf{D}^{-1}\right)^{-1}\mathbf{U}^T. \tag{2.2}$$

Computation of Eq. 2.2 is reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(m^3)$, and by constructing the spatial covariance so that $m \ll n$ large gains in computational efficiency may be gained.

For any covariance matrix, $\boldsymbol{\Sigma}$, we can construct an equivalent *spatial mixed effects model*

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{S}(\mathbf{s})\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{2.3}$$

by defining the *spatial basis functions* as, $\mathbf{S}(\mathbf{s}) = \mathbf{U}\mathbf{D}^{1/2}$ and letting $\boldsymbol{\eta} \sim (0, \sigma^2\mathbf{I}_k), \boldsymbol{\eta} \perp \boldsymbol{\epsilon}$ be the *spatial random effects*. The relationship between spatial processes and smoothers is well established [60], and since orthogonality of $\mathbf{U}$ is not required a wide variety of spatial basis

functions may be used in this setting, such as smoothing spline basis functions [80]. Further details regarding the class of basis functions which may be used are detailed in Section 3.1 of Cressie and Johanssen [16], and it has been demonstrated that they are able to approximate covariance functions often used in spatial statistics [59].

Under the spatial mixed effects model,

$$\mathbf{Z}(\mathbf{s}) \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\sigma^2, \tau^2)), \quad \mathbf{V}(\sigma^2, \tau^2) = \sigma^2 \mathbf{S}(\mathbf{s})\mathbf{S}^T(\mathbf{s}) + \tau^2 \mathbf{I}_n. \tag{2.4}$$

Predictions at unobserved locations follow from the conditional expectation of the spatial mixed effects model given the realization of the spatial random effect

$$\mathrm{E}[\mathbf{Z}(\mathbf{s_0})|\hat{\boldsymbol{\eta}}] = \mathbf{X}_0\hat{\boldsymbol{\beta}} + \mathbf{S}(\mathbf{s_0})\hat{\boldsymbol{\eta}},$$

where $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator, and $\hat{\boldsymbol{\eta}}$ is the best linear unbiased predictor. The spatial random effect $\hat{\boldsymbol{\eta}}$ can also be interpreted as a penalized regression estimator ([69] 4.5.3). By Henderson's justification [67], optimizing $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$ leads to minimizing the criteria

$$\|\mathbf{Y}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta} - \mathbf{S}(\mathbf{s})\boldsymbol{\eta}\|_2^2 + \frac{\tau^2}{\sigma^2}\|\boldsymbol{\eta}\|_2^2,$$

which can be interpreted as a penalized regression estimate on $\hat{\boldsymbol{\eta}}$ with tuning parameter $\lambda = \frac{\tau^2}{\sigma^2}$.

2.2.5 Spatially Adjusted Trees

Additive models combining regression trees and kriging can be formulated as

$$\mathbf{Y}(\mathbf{s}) = \mathbf{t}(\mathbf{X}) + \boldsymbol{\nu}(\mathbf{s}), \tag{2.5}$$

with $\mathbf{t}(\mathbf{X})$ the regression tree constructed from the covariates and $\boldsymbol{\nu}(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\rho}))$ a realization of a Gaussian process.

Under the additive model formulation of the underlying process combining trees and kriging (2.5), $\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{t}(\mathbf{X}), \mathbf{V}(\boldsymbol{\theta}))$. By maximum likelihood, we wish to find a regression

tree estimate $\hat{\mathbf{t}}(\mathbf{X})$ and covariance parameters $\hat{\boldsymbol{\theta}}$ such that

$$\{\hat{\mathbf{t}}(\mathbf{X}), \hat{\boldsymbol{\theta}}\} = \underset{\mathbf{t}(\mathbf{X}),\boldsymbol{\theta}}{\operatorname{argmax}} \left[-\frac{1}{2} \log |\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2} (\mathbf{Y}(\mathbf{s}) - \mathbf{t}(\mathbf{X}))^T \mathbf{V}^{-1}(\boldsymbol{\theta}) (\mathbf{Y}(\mathbf{s}) - \mathbf{t}(\mathbf{X})) \right]. \quad (2.6)$$

We propose a principled likelihood-based optimization motivated by profile likelihood. The regression tree is profiled out of the optimization problem as

$$\begin{aligned} &\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[-\frac{1}{2} \log |\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2} (\mathbf{Z}(\mathbf{s}) - \hat{\mathbf{t}}(\mathbf{X}|\mathbf{V}(\boldsymbol{\theta})))^T \mathbf{V}^{-1}(\boldsymbol{\theta}) (\mathbf{Z}(\mathbf{s}) - \hat{\mathbf{t}}(\mathbf{X}|\mathbf{V}(\boldsymbol{\theta}))) \right] \\ &\quad \text{s.t } \hat{\mathbf{t}}(\mathbf{X}|\mathbf{V}(\boldsymbol{\theta})) = \underset{\mathbf{t}(\mathbf{X}|\mathbf{V}(\boldsymbol{\theta}))}{\operatorname{argmin}} (\mathbf{Z}(\mathbf{s}) - \mathbf{t}(\mathbf{X}|\mathbf{V}(\boldsymbol{\theta})))^T \mathbf{V}^{-1}(\boldsymbol{\theta}) (\mathbf{Y}(\mathbf{s}) - \mathbf{t}(\mathbf{X}|\mathbf{V}(\boldsymbol{\theta}))). \end{aligned} \quad (2.7)$$

The dependence of the profiled spatially adjusted regression tree on the spatial correlation matrix is emphasized as $\hat{\mathbf{t}}(\mathbf{X}|\mathbf{V}(\boldsymbol{\theta}))$. We note similarities of this optimization problem to universal kriging and traditional regression trees. In universal kriging, $\mathbf{t}(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and the profile likelihood optimization criteria selects $\boldsymbol{\beta}$ which maximizes $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}_0|\mathbf{Y}(\mathbf{s}))$ for some fixed $\boldsymbol{\theta}_0$. On the other hand, if we ignore the spatial process and let $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{I}_n$, there are no covariance parameters to maximize over and we would build a normal regression tree which minimizes mean squared error. Thus, a spatially adjusted regression tree should be built to minimize (Eq. 2.7) for a given $\boldsymbol{\theta}_0$.

2.2.6 Spatially Adjusted Tree Building Algorithm

We propose a novel, computationally feasible *spatially adjusted tree building algorithm* to construct a spatially adjusted tree which aims to minimize (2.7) by recursive binary splitting. Note that this algorithm is a greedy approach and does not guarantee convergence to the true minimizer. In Section 2.2.1, we showed that each tree can be written as a linear combination of the tree design matrix $\mathbf{C}^k$ and their corresponding weights $\boldsymbol{\pi}^k$ and the optimization criteria for the regression tree in Eq. 2.7 becomes:

$$\ell(\mathbf{C}^k, \boldsymbol{\pi}^k) = \left(\mathbf{Z}(\mathbf{s}) - \mathbf{C}^k\boldsymbol{\pi}^k\right)^T \mathbf{V}^{-1}(\boldsymbol{\theta}) \left(\mathbf{Z}(\mathbf{s}) - \mathbf{C}^k\boldsymbol{\pi}^k\right). \quad (2.8)$$

By profile likelihood, we define the "characteristic matrix" for the spatial tree building algorithm $\mathbf{\Omega}^k$ (Eq. 2.9) which depends only on the tree design matrix.

$$\mathbf{\Omega}^k = \mathbf{V}^{-1}(\boldsymbol{\theta}) - \mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{C}^k \left((\mathbf{C}^k)^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{C}^k\right)^{-1}(\mathbf{C}^k)^T\mathbf{V}^{-1}(\boldsymbol{\theta}). \tag{2.9}$$

The loss is solely a function of $\mathbf{\Omega}^k$, the characteristic matrix, and the observations $\mathbf{Z}(\mathbf{s})$. For any new branch, $\mathbf{c}^A$, a vector noting which observations are in the new terrminal node, updating $\mathbf{\Omega}^{k+1}$ depends only on $\mathbf{c}^A$ and the previous characteristic matrix $\mathbf{\Omega}^k$ as

$$\mathbf{\Omega}^{k+1} = \mathbf{\Omega}^k - \mathbf{\Omega}^k\mathbf{c}^A\left(\left(\mathbf{c}^A\right)^T\mathbf{\Omega}^k\mathbf{c}^A\right)^{-1}\left(\mathbf{c}^A\right)^T\mathbf{\Omega}^k \tag{2.10}$$

Details of equality for Eq. 2.10 are included in the Appendix. Using this fact, the change in loss between $\mathbf{C}^k$ and $\mathbf{C}^{k+1} = \begin{bmatrix}\mathbf{C}^k & \mathbf{c}^A\end{bmatrix}$ is easily shown to be

$$\nabla\ell(\mathbf{c}^A) = \ell(\mathbf{C}^k) - \ell(\mathbf{C}^{k+1}) = \mathbf{Y}(\mathbf{s})^T\left(\mathbf{\Omega}^k\mathbf{c}^A\left(\left(\mathbf{c}^A\right)^T\mathbf{\Omega}^k\mathbf{c}^A\right)^{-1}\left(\mathbf{c}^A\right)^T\mathbf{\Omega}^k\right)\mathbf{Y}(\mathbf{s}). \tag{2.11}$$

The spatially adjusted tree building algorithm is summarized in Algorithm 1.

2.2.7 Computational Complexity for Spatially Adjusted Trees

We first note that $\left(\left(\mathbf{c}^A\right)^T\mathbf{\Omega}^k\mathbf{c}^A\right)^{-1}$ is a scalar and the change in loss for any candidate split is

$$\nabla\ell(\mathbf{c}^A) = \frac{\left\|\left(\mathbf{c}^A\right)^T\mathbf{w}\right\|_2^2}{\left(\mathbf{c}^A\right)^T\mathbf{\Omega}^k\mathbf{c}^A} = \frac{\mathrm{Num}(\mathbf{c}^A)}{\mathrm{Den}(\mathbf{c}^A)}, \quad \mathbf{w} = \mathbf{\Omega}^k\mathbf{Y}(\mathbf{s}).$$

For any covariate there are at most $n-1$ possible new cutoff values across all terminal nodes. Define the set of possible splits, $\mathbf{c}_{ij}^A \in \mathbf{C}_i^A$ as the set of possible splits in terminal node j, for the i^{th} cutpoint. If the new candidate split is the first possible split of a terminal node, then it is a vector with a single one in position l_1 and zeros everywhere else. In this case, $\nabla\ell(\mathbf{c}_1^A) = \frac{w_{l_1}^2}{\Omega_{l_1l_1}^k}$ and is $\mathcal{O}(1)$. Otherwise, the next candidate split adds a single one in position l_{i+1} to the previous split and:

$$\nabla\ell(\mathbf{c}_{(i+1)j}^A) = \frac{\mathrm{Num}(\mathbf{c}_{ij}^A) + \mathbf{w}_{l_{i+1}}^2}{\mathrm{Den}(\mathbf{c}_{ij}^A) + \mathbf{\Omega}_{l_{i+1}l_{i+1}}^k + 2\sum_{m=1}^{i}\mathbf{\Omega}_{l_ml_{i+1}}^k}.$$

Algorithm 1 Spatially Adjusted Tree Building Algorithm

1. set $\mathbf{C}^1 = \left[\mathbf{1}_n\right]$

2. Given $\mathbf{V}(\boldsymbol{\theta})$, set the initial value for
   $\boldsymbol{\Omega}^1 = \mathbf{V}^{-1}(\boldsymbol{\theta}) - \mathbf{C}^1\left((\mathbf{C}^1)^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{C}^1\right)^{-1}(\mathbf{C}^1)^T\mathbf{V}^{-1}(\boldsymbol{\theta})$.

3. For $k = 2, 3, \ldots$

   (a) Take a random sample of the covariates $\mathbf{X}_r \subset \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p\}$.

   (b) Check each of the j existing terminal nodes for a new terminal node created by a decision rule based on the sampled covariates $\mathbf{X}_r$, to create a candidate set of possible splits $\mathbf{C}_i^A$.

   (c) Find the candidate split $\mathbf{c}^A \in \mathbf{C}_i^A$ which maximizes the change in loss
   $$\mathbf{Y}(\mathbf{s})^T\left(\boldsymbol{\Omega}^k\mathbf{c}^A\left(\left(\mathbf{c}^A\right)^T\boldsymbol{\Omega}^k\mathbf{c}^A\right)^{-1}\left(\mathbf{c}^A\right)^T\boldsymbol{\Omega}^k\right)\mathbf{Y}(\mathbf{s}).$$

   (d) Update
   $$\boldsymbol{\Omega}^{k+1} = \boldsymbol{\Omega}^k - \boldsymbol{\Omega}^k\mathbf{c}^A\left(\left(\mathbf{c}^A\right)^T\boldsymbol{\Omega}^k\mathbf{c}^A\right)^{-1}\left(\mathbf{c}^A\right)^T\boldsymbol{\Omega}^k.$$

4. Repeat step 3 until the maximum number of splits are exceeded or there are no more branches can be split without creating a branch with less than m observations.

The loss for each additional split can be computed in $\mathcal{O}(n)$ (compared to $\mathcal{O}(1)$ for standard regression trees) and the worst case run time for our spatially adjusted tree is $\mathcal{O}(pn^2 \log(n) + h)$, where h is the run time to needed to invert the spatial covariance matrix a single time.

2.2.8 Random Spatial Forests: Pseudo-Likelihood Approach

Our algorithm describes a process for constructing spatially adjusted trees for known $\boldsymbol{\theta}_0$, and we can construct spatial random forests estimates by aggregating over these trees. In practice however, $\boldsymbol{\theta}$ is unknown. Since random forests estimates the expectation of an infinite tree [27], we propose a pseudo-likelihood approach where we replace the regression tree with its bagged random forests estimate in Eq. 2.7.

Joint optimization of the random forests and the covariance parameters characterizing the spatial process is difficult for a number of reasons. For Matern covariance functions, the likelihood function is a non-convex function of the covariance parameters. Gradient based approaches for finding local minima/maxima cannot be applied for random forests since no closed form gradient exists. Further, numerical gradients are complicated by randomness in resampling of observations and covariates from random forests creating a stochastic function evaluation.

In order to simplify estimation of the covariance parameters we approximate the spatial correlation using spatial basis functions as in Equation 2.4, and define an alternative set of parameters, $\kappa = \sigma^2 + \tau^2$, $\delta = \sigma^2 / (\sigma^2 + \tau^2)$. Using these parameters, the spatial covariance becomes:

$$\mathbf{V}(\kappa, \delta) = \kappa \mathbf{R}(\delta), \quad \mathbf{R}(\delta) = \left(\delta \mathbf{S}(\mathbf{s}) \mathbf{S}^T(\mathbf{s}) + (1 - \delta) \mathbf{I}_n\right).$$

Now, the profile log-likelihood can easily be shown to be written as a function of a single

parameter δ by profiling out $\hat{\mathbf{f}}(\mathbf{X}|\mathbf{R}(\delta))$ and $\hat{\kappa}$ as

$$\ell(\delta) = -\frac{n}{2}\log(\hat{\kappa}) - \frac{1}{2}|\mathbf{R}(\delta)| - \frac{1}{2\hat{\kappa}}\left(\mathbf{Y}(\mathbf{s}) - \hat{\mathbf{f}}(\mathbf{X}|\mathbf{R}(\delta))\right)^T \mathbf{R}^{-1}(\delta)\left(\mathbf{Y}(\mathbf{s}) - \hat{\mathbf{f}}(\mathbf{X}|\mathbf{V}(\hat{\kappa},\delta))\right)$$

$$\text{s.t } \hat{\mathbf{f}}(\mathbf{X}|\mathbf{R}(\delta)) = \underset{\mathbf{f}(\mathbf{X}|\mathbf{R}(\delta))}{\operatorname{argmin}} \left(\mathbf{Y}(\mathbf{s}) - \mathbf{f}\left(\mathbf{X}|\mathbf{R}(\delta)\right)\right)^T \mathbf{R}^{-1}(\delta)\left(\mathbf{Y}(\mathbf{s}) - \mathbf{f}\left(\mathbf{X}|\mathbf{R}(\delta)\right)\right)$$

$$\text{and } \hat{\kappa} = \frac{1}{n}(\mathbf{Y}(\mathbf{s}) - \hat{\mathbf{f}}(\mathbf{X}|\mathbf{R}(\delta)))^T\mathbf{R}^{-1}(\delta)(\mathbf{Y}(\mathbf{s}) - \hat{\mathbf{f}}(\mathbf{X}|\mathbf{R}(\delta))).$$

This parameterization makes optimization simpler, as we only need to optimize over $\delta \in [0, 1]$. Since we have a single parameter restricted to a small search space, we optimize the model by performing a grid search and selecting δ which minimizes the pseudo-likelihood $\ell(\delta)$.

2.2.9 Random Spatial Forests: Non-Parametric Estimation

In the previous section, we derived an additive model using a pseudo-likelihood approach to integrate random forests into a likelihood model. However, it is not easy to interpret $\boldsymbol{\eta}$ as a random effect since it is difficult to imagine the data generating mechanism that might give rise to such fields [32]. In this case, modeling the spatial process using a random effect is a form of regularization and pseudo-likelihood gives us a way to estimate the parameter δ. The goal of modeling the air pollution surface in epidemiological studies is to produce accurate estimates for individuals at unobserved locations, which can be viewed as a prediction problem. An alternative criterion when prediction accuracy is desired, which is often the case in many statistical learning applications, is to minimize the expected mean squared test error:

$$\underset{\delta}{\operatorname{argmin}}\, \mathrm{E}\left[\|\mathbf{Y} - \hat{\mathbf{Y}}(\mathbf{s},\delta)\|_2^2\right].$$

Noting the relationship between linear mixed models and penalized regression in section 2.2.4, a natural non-parametric approach would be to select the tuning parameter δ for the additive model by k-fold cross-validation in order to find δ which minimizes the out of sample test error.

The addition of running k-fold cross-validation to estimate expected test error for each candidate δ results in a substantial increase in computational time. But a unique property of

random forests is that since random forests only uses "out-of-bag" samples in its estimation, the resulting function $\hat{\mathbf{Y}}(\mathbf{s})$ is equivalent to its cross-validated estimate [27]. This is desirable since the mean squared error of $\hat{\mathbf{Y}}(\mathbf{s})$ on the training set is equivalent to its expected test error and makes k-fold cross-validation is unnecessary, reducing the computational burden. In order to leverage this property, we propose applying the random forests algorithm to aggregate our spatially adjusted trees and each their associated spatial smoothers as:

$$\hat{\mathbf{Y}}_t(\mathbf{s}, \delta) = \frac{1}{B} \sum_{i=1}^{B} \left[\hat{\mathbf{t}}^i \left(\mathbf{X}^i | \mathbf{R}^i(\delta) \right) + \mathbf{S}^i(\mathbf{s}) \hat{\boldsymbol{\eta}}^i \right].$$

For each bootstrap sample, we can use our tree building algorithm in Section 2.2.6 to estimate both the spatially adjusted tree $\mathbf{t}^i \left(\mathbf{X}^i | \mathbf{R}^i(\delta) \right)$ and its associated spatial random effect $\hat{\boldsymbol{\eta}}^i$ by its best linear unbiased predictor. Over a grid of $\delta \in [0, 1]$, we fit $\hat{\mathbf{Y}}(\mathbf{s}, \delta)$ using our spatially adjusted tree building algorithm and select

$$\underset{\delta}{\operatorname{argmin}} \, \| \mathbf{Z}(\mathbf{s}) - \hat{\mathbf{Y}}(\mathbf{s}, \delta) \|_2^2.$$

2.3 Simulation Study

We conduct a set of simulations to compare different methods of combining random forests with a spatial smoother. Datasets for simulations are created on a grid of points over the continental United States spaced at 25km intervals and GIS covariates at these locations are provided from ArcGIS 10.2.

2.3.1 Generating the Observed Surface

For each simulation, we constructed a fixed exposure surface from an additive model of a function of GIS covariates, $\mathbf{f}(\mathbf{X})$, and a fixed realization of a Gaussian process with exponential covariance process, $\boldsymbol{\nu}(\mathbf{s})$, with range randomly generated between $10\% - 20\%$ of the maximum distance between points. A variety of different generating functions were used for the function of GIS covariates, sparse signals where only a few coefficients were non-zero,

some where all coefficients were non-zero, and some which included interaction terms between the GIS covariates to induce non-linearity, but our simulations did not suggest the type of generating function had a significant impact on the results.

The observed surface is constructed as

$$\mathbf{Y}(\mathbf{s}) = \gamma \mathbf{f}(\mathbf{X}) + \boldsymbol{\nu}(\mathbf{s}),$$

where the parameter γ controls the proportion of variance attributable to the GIS covariates. We consider two scenarios for this parameter: (1) *Strong Covariates*: 65% of the generated process is due to the covariates (2) *Weak Covariates*: 35% of the generated process is due to the covariates.

2.3.2 Methods combining Random Forests with Spatial Smoothing

For our examples we formulate the spatial basis functions using TPRS following Olives et al. [61]. This choice is arbitrary, and as noted in Section 2.2.4 one could consider selecting alternative spatial basis functions. We selected TPRS as an alternative to kriging as there is an equivalence between thin plate regression splines (TPRS) and kriging with a Matern-class covariance with infinite range [60]. We compared the following six methods:

1. Random Forests (RF)—implemented using the `randomForests` package.

2. Spatial Smoothing (TPRS)—implemented by the `mgcv` package.

3. Random Forests plus Spatial Smoothing (RF-TPRS)—Two step approach where first RF is run, then TPRS if fit to the RF residuals.

4. Spatial Smoothing plus Random Forests (TPRS-RF)—Two step approach where first TPRS is run, then RF is applied to the residuals from TPRS.

5. Random Spatial Forests— Pseudo-Likelihood (SpatRF-PL), Section 2.2.8.

6. Random Spatial Forests—Non-Parametric (SpatRF-NP), Section 2.2.9.

2.3.3 Evaluating Reconstruction Accuracy

We generate a single observed surface $\mathbf{Y}(\mathbf{s})$ and hold out 200 points for validation. For each simulation, 150 points are randomly sampled to train six different models to compare on. Training points are observed with independent measurement error

$$\mathbf{Z}(\mathbf{s}_{\text{train}}) = \mathbf{Y}(\mathbf{s}_{\text{train}}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \tau^2 \mathbf{I}_{n_{\text{train}}}\right),$$

where τ^2 is randomly generated to be between 10% and 25% of the total variance. Model reconstruction accuracy is estimated by their root mean square prediction error (RMSPE) at validation points and we report the average RMSPE on the validation points for each method from 30 different randomly sampled training points. This is repeated for 180 different observed surfaces. Density plots of average RMSPE for each method are shown in Figure 2.1.

2.3.4 Simulation Results

In the *strong covariates* scenario, RF does better than TPRS alone while this relationship is reversed in the *weak covariates* scenario. This demonstrates that when a large percentage of the observed surface can be explained by the covariates, constructing a surface using a function of the covariates by RF performs better than ignoring the covariates and applying TPRS. On the other hand, when the covariates can only explain a small percentage of the total variation, using only the covariates via RF leads to worse prediction accuracy than simply applying TPRS alone.

Additive models combining RF and TPRS (RF-TPRS and TPRS-RF) do better than either RF or TPRS alone. Comparing RF-TPRS and TPRS-RF highlights the importance of the optimization approach. Although RF-TPRS and TPRS-RF are both composed of a random forests and thin plate regression spline, the order of estimation can have a large impact on the models prediction accuracy. When the covariates are responsible for a large percentage of variability in the observed surface RF-TPRS performs noticeably better than TPRS-RF, and vice versa when the covariates explain a small portion of the variance. Our

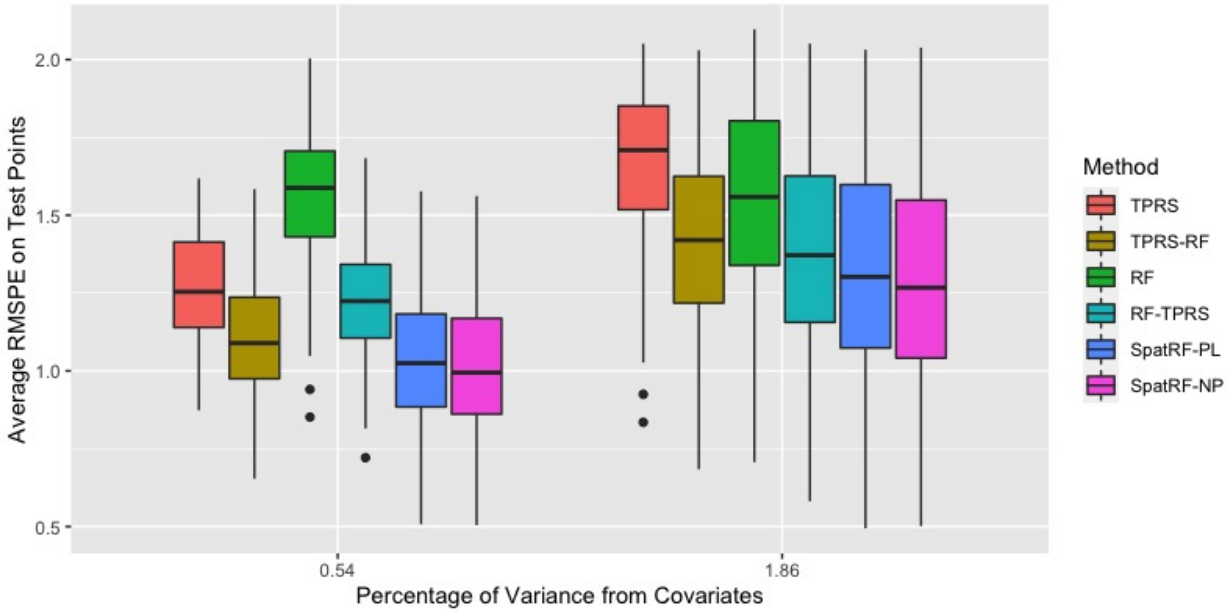

Figure 2.1: Simulation Results: Each point in the boxplot represents the average RMSPE at the 200 validation points for one of the 180 generated surfaces over 30 repeated samples. The box and whisker summarizes the prediction accuracy of the method on different simulated surfaces from a variety of generating functions.

methods demonstrate how constructing random forests allowing for spatial correlation leads to more accurate predictions than either two-step approach regardless of how much variability can be explained by the covariates. In our simulations, SpatRF-PL and SpatRF-NP have better prediction accuracy than RF-TPRS and TPRS-RF in all scenarios. Comparing our two methods, SpatRF-NP performs slightly better than SpatRF-PL in both cases.

2.4 Application to Sub-Species of $PM_{2.5}$

We develop air pollution models for annual averages of four $PM_{2.5}$ (particulate matter less than 2.5 μm in aerodynamic diameter) sub-species: elemental carbon (EC), organic carbon (OC), silicon (Si), and sulfur (S) using Environmental Protection Agency (EPA) Interagency Monitoring for Protected Visual Environments (IMPROVE) and Chemical Speciation Network (CSN) monitoring data from 2009-2010. Following Bergen et al. [6], we only include CSN and IMPROVE monitors with at least 10 data points per quarter and no more than 45 days between consecutive measurements. Si and S measurements were averaged over 01/01/2009–12/31/2009, while EC/OC consisted of measurements from 204 IMPROVE and CSN monitors averaged over 01/01/2009–12/31/2009, and measurements from 51 CSN monitors averaged over 05/01/2009 - 04/30/2010. Annual averages were square-root transformed prior to modeling.

In addition to methods used in the simulations, we include universal kriging estimates which deal with the high dimensionality of the covariates by pre-processing the covariates by partial least squares (UK-PLS) and use an exponential covariance matrix. This technique was employed in the original analysis [6] and is commonly employed in many land-use regression settings. Additionally, we examine random forests with spatial information included (RF w/ TPRS). This approach included spatial basis functions are included as covariates accounting for geographic proximity between observations and is a heuristic for including spatial information into normal random forests.

Surface reconstruction accuracy of each methods is assessed by comparing predictions generated from ten-fold cross-validation. Performance of each model is based on their average

	UK-PLS	TPRS	RF	RF w/ TPRS	RF-TPRS	TPRS-RF	SpatRF - PL	SpatRF - NP
EC	0.163	0.272	0.143	0.140	0.135	0.162	0.134	0.131
OC	0.591	0.777	0.589	0.553	0.554	0.589	0.553	0.535
Si	0.073	0.070	0.080	0.069	0.074	0.069	0.069	0.068
S	0.079	0.087	0.154	0.088	0.107	0.079	0.078	0.078

Table 2.1: Ten-fold cross-validated prediction accuracy, summarized by RMSPE, of each method for $PM_{2.5}$ components Elemental Carbon (EC), and Organic Carbon (OC), Silicon (Si), Sulfur (S) collected by AQS and IMPROVE monitoring networks from 2009-2010.

RMSPE over ten separate cross-validation runs in Table 2.1. Cross-validated prediction accuracy over the different components of PM2.5 show similar findings to our simulation results, with SpatRF-NP consistently performing at least as well as any of the alternative methods. Predictions using RF-TPRS, TPRS-RF, UK-PLS, and SpatRF-NP across the continental United States for EC, OC, Si, and S are shown in Figures 2.2, 2.3, 2.4, and 2.5, respectively.

When a large proportion of the variance can be explained by the covariates, demonstrated by EC and OC, RF performs better than TPRS and RF-TPRS has improved cross-validated accuracy over TPRS-RF. SpatRF-NP and SpatRF-PL show small but noticeable improvements over RF-TPRS and are more accurate for both pollutants. In these examples using random forests instead of using a linear model with dimension reduction on the covariates by PLS can yield noticeable improvements in prediction accuracy as SpatRF-PL and SpatRF-NP have noticeably smaller RMSPE than UK-PLS.

Si and S are examples where spatial smoothing is able to model a larger proportion of the variance then the covariates alone. In both of these cases, RF-TPRS has higher RMSPE than TPRS-RF. Interestingly, TPRS alone has better cross-validated accuracy than RF-TPRS suggesting that applying spatial smoothing to the residuals from random forests does not guarantee that the combined approach is more accurate than either individual method

alone. For S, TPRS-RF, UK-PLS, SpatRF-PL, and SpatRF-NP all do quite well (CV R^2 0.94-0.95), while SpatRF-NP has the highest cross-validated R^2 for Si.

In order to closely examine how efficient combining random forests and spatial smoothing aids in surface reconstruction, we break down the additive model estimates into its parts. From the additive RF and TPRS components of the model (Figure 2.6), it is clear the smoothing spline is unable to pick up the sharp fluctuations in EC that occur in large cities, for example, Los Angeles and New York. In these cases, the use of GIS covariates can lead to large improvements in prediction accuracy. RF-TPRS performs better than TPRS-RF, and RF-TPRS shows a noticeable improvement over UK-PLS. However, by modifying the random forest algorithm to account for estimation of the spatial process we get slight improvements over RF-TPRS using our spatially adjusted random forest algorithms.

In cases like silicon and sulfur, the overall patterns appear to have a large scale structure. Sulfur has a large east to west relationship, peaking near Pittsburgh and gives a clear example of when attempting to use machine learning methods performs worse than traditional geostatistical approaches where spatial smoothing is employed. Figure 2.8 shows the RF and TPRS components for each of the additive models. Although much of the variation is explained by geographic location, RF aims to use its geographic covariates to explain the east to west relationship of the model. Since the standard RF algorithm does not account for the spatial process, it attempts to model large scale east to west variation in a sub-optimal matter.

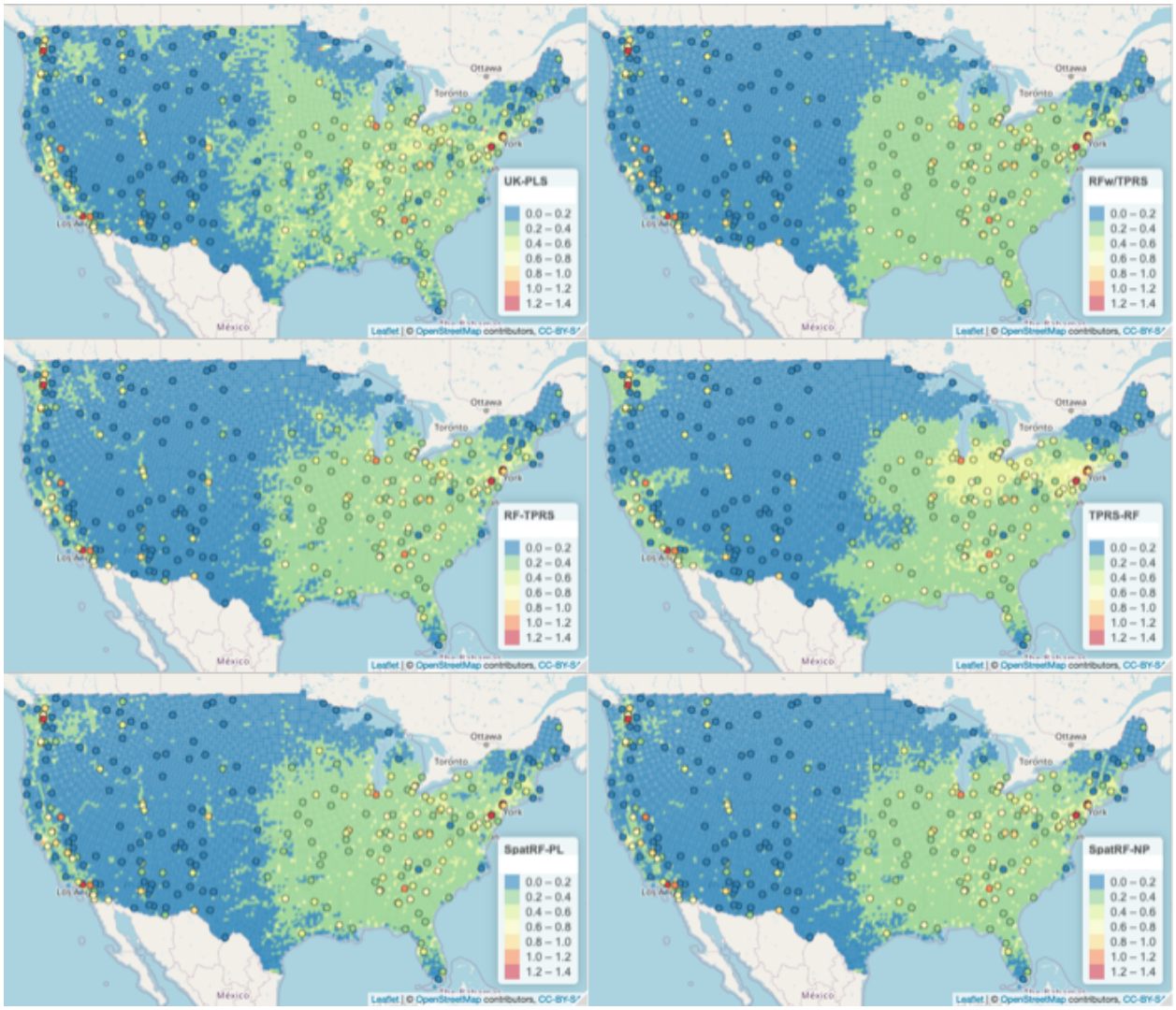

Figure 2.2: Predicted Elemental Carbon concentrations (in $\mu g/m^3$) across the continental United States. Points are observed annual averages at monitoring locations. Top Left: Universal Kriging with dimensions reduction by Partial Least Squares, Top Right: Random Forests with Spatial Information included as Covariates, Middle Left: Thin Plate Regression Splines followed by Random Forests, Middle Right: Random Forests followed by Thin Plate Regression Splines, Bottom Left: Random Spatial Forests by Pseudo-Likelihood, Bottom Right: Random Spatial Forests by Non-Parameteric Estimation

Figure 2.3: Predicted Organic Carbon concentrations (in $\mu g/m^3$) across the continental United States. Points are observed annual averages at monitoring locations. Top Left: Universal Kriging with dimensions reduction by Partial Least Squares, Top Right: Random Forests with Spatial Information included as Covariates, Middle Left: Thin Plate Regression Splines followed by Random Forests, Middle Right: Random Forests followed by Thin Plate Regression Splines, Bottom Left: Random Spatial Forests by Pseudo-Likelihood, Bottom Right: Random Spatial Forests by Non-Parameteric Estimation

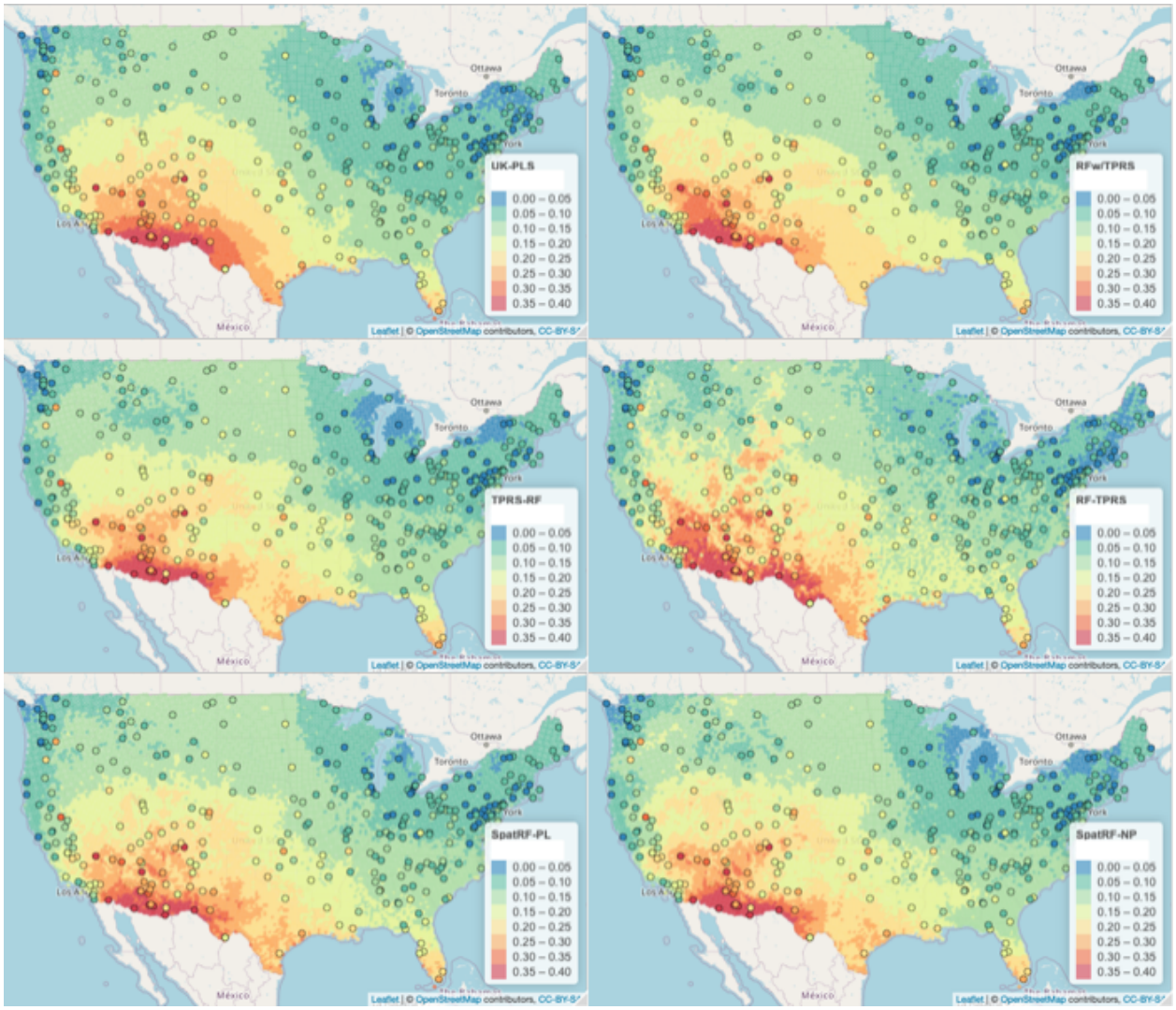

Figure 2.4: Predicted Silicon concentrations (in ng/m^3) across the continental United States. Points are observed annual averages at monitoring locations. Top Left: Universal Kriging with dimensions reduction by Partial Least Squares, Top Right: Random Forests with Spatial Information included as Covariates, Middle Left: Thin Plate Regression Splines followed by Random Forests, Middle Right: Random Forests followed by Thin Plate Regression Splines, Bottom Left: Random Spatial Forests by Pseudo-Likelihood,Bottom Right: Random Spatial Forests by Non-Parameteric Estimation

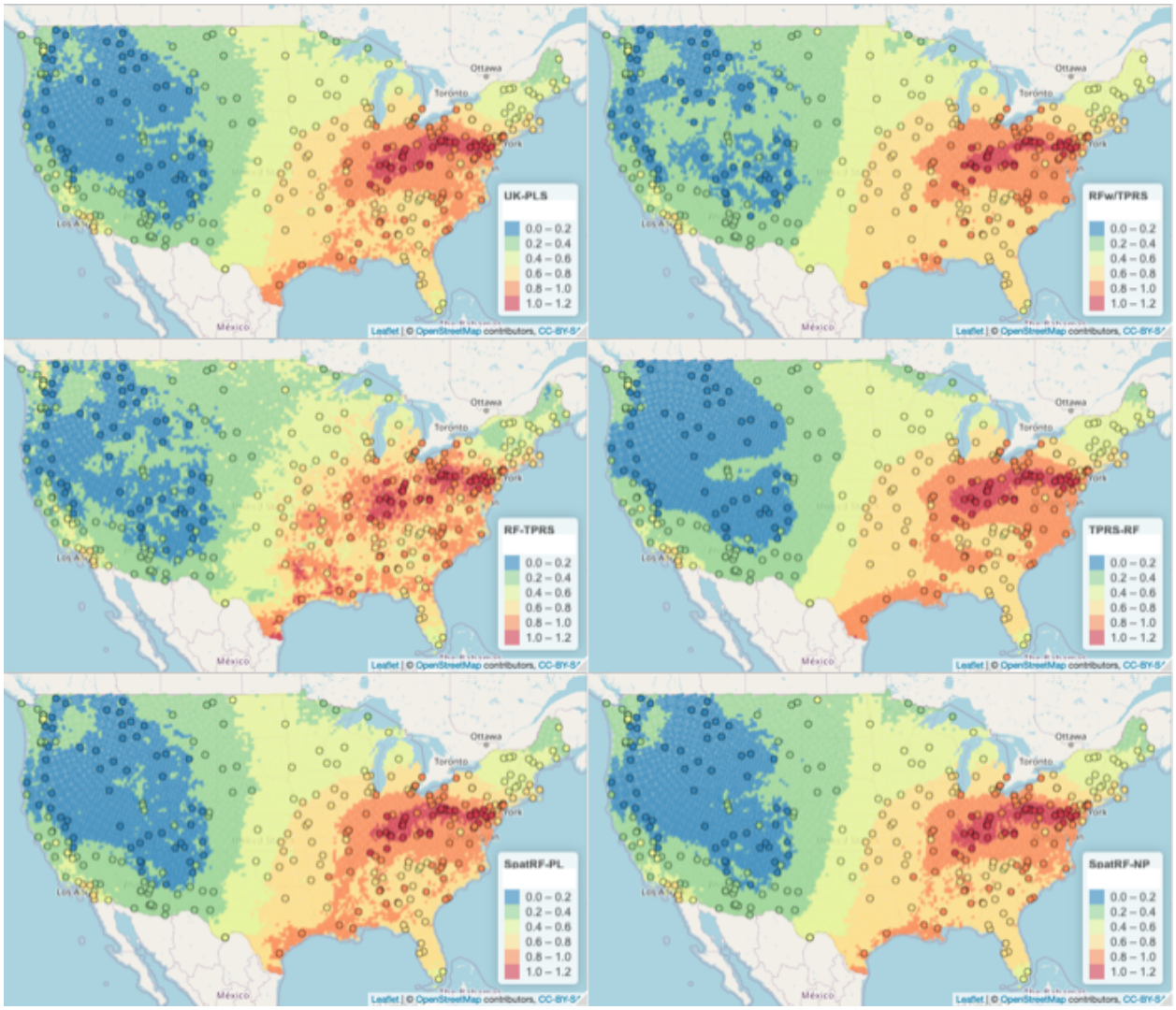

Figure 2.5: Predicted Sulfur concentrations (in $\mu g/m^3$) across the continental United States. Points are observed annual averages at monitoring locations. Top Left: Universal Kriging with dimensions reduction by Partial Least Squares, Top Right: Random Forests with Spatial Information included as Covariates, Middle Left: Thin Plate Regression Splines followed by Random Forests, Middle Right: Random Forests followed by Thin Plate Regression Splines, Bottom Left: Random Spatial Forests by Pseudo-Likelihood, Bottom Right: Random Spatial Forests by Non-Parameteric Estimation

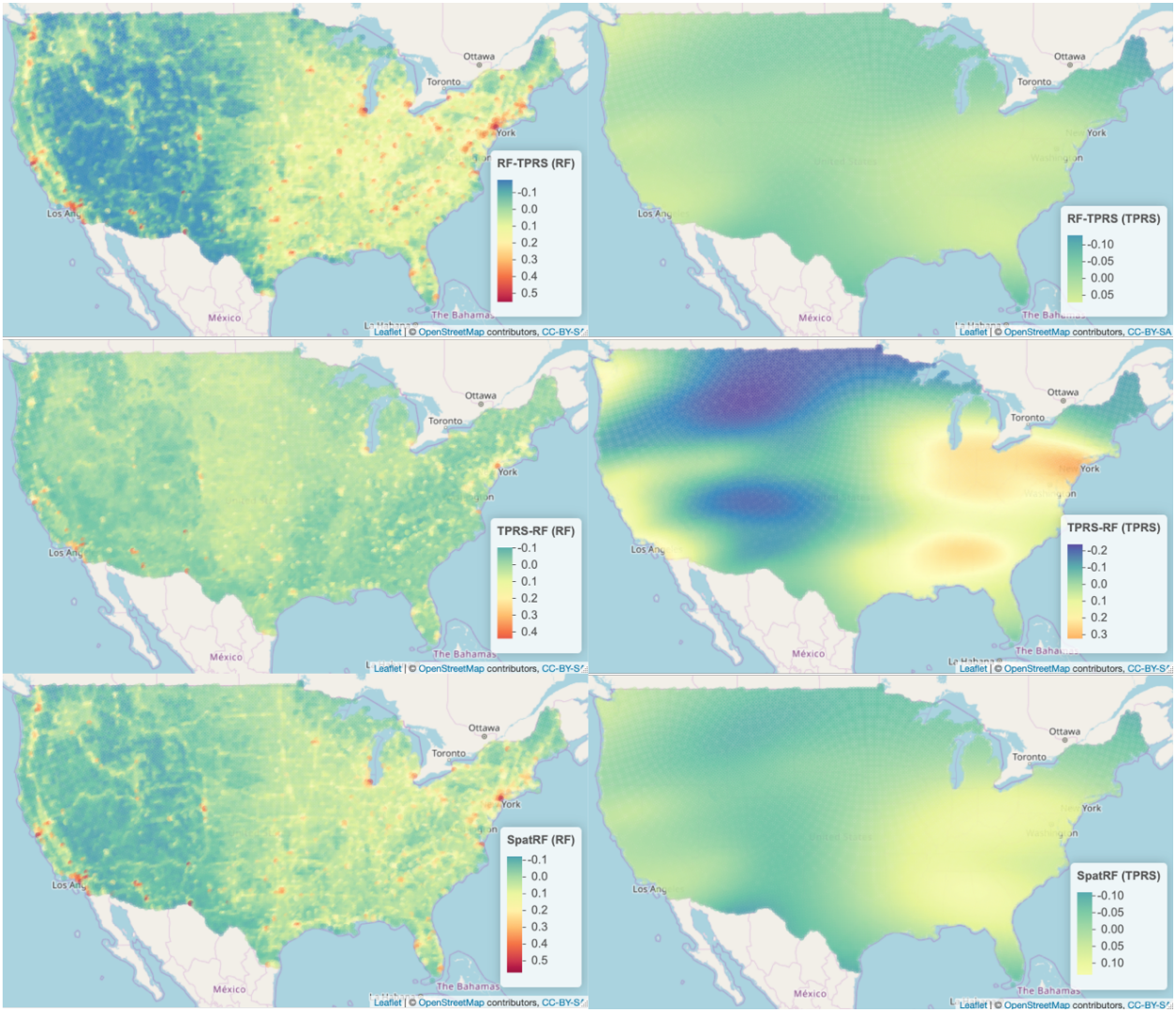

Figure 2.6: RF and TPRS components for Elemental Carbon concentrations (in $\mu g/m^3$). Maps on the left side show the RF component of the additive model while maps on the right show the TPRS estimate. Each row is a different estimation order, top: RF first, then TPRS, middle: TPRS first, then RF, bottom: SpatRF-NP which jointly estimates the RF and TPRS components

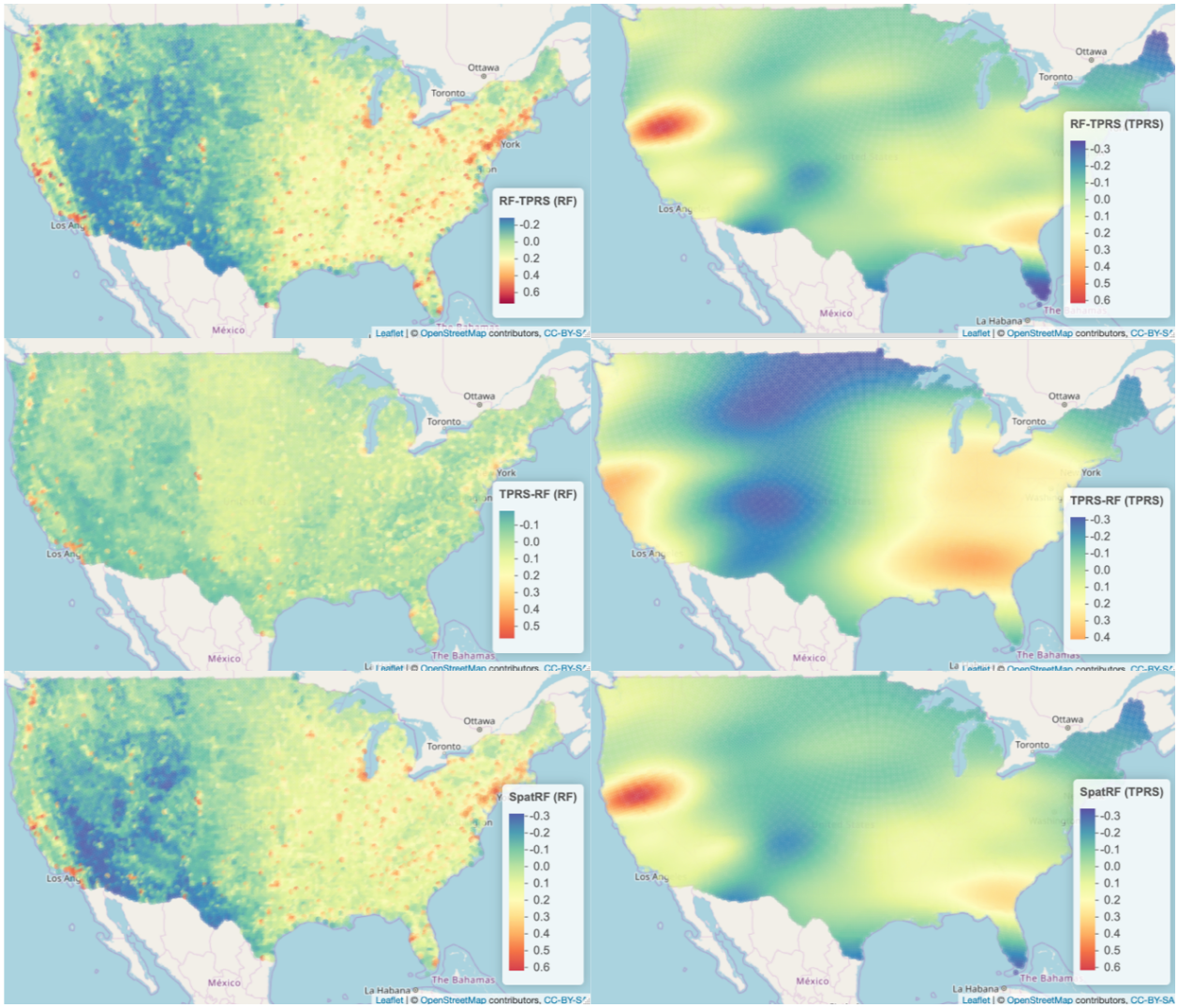

Figure 2.7: RF and TPRS components for Organic Carbon concentrations (in μg/m^3). Maps on the left side show the RF component of the additive model while maps on the right show the TPRS estimate. Each row is a different estimation order, top: RF first, then TPRS, middle: TPRS first, then RF, bottom: SpatRF-NP which jointly estimates the RF and TPRS components

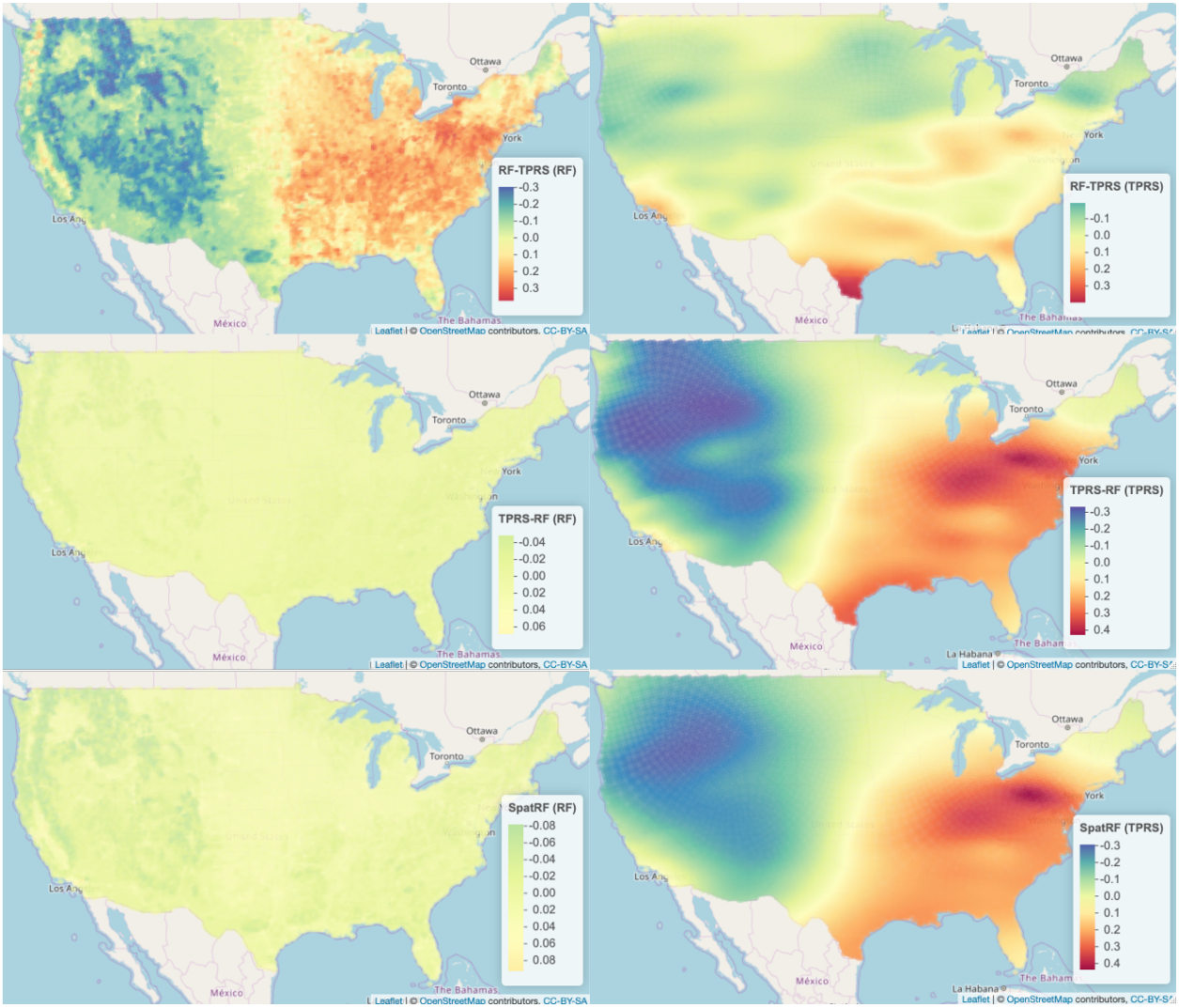

Figure 2.8: RF and TPRS components for Sulfur (in μg/m^3). Maps on the left side show the RF component of the additive model while maps on the right show the TPRS estimate. Each row is a different estimation order, top: RF first, then TPRS, middle: TPRS first, then RF, bottom: SpatRF-NP which jointly estimates the RF and TPRS components

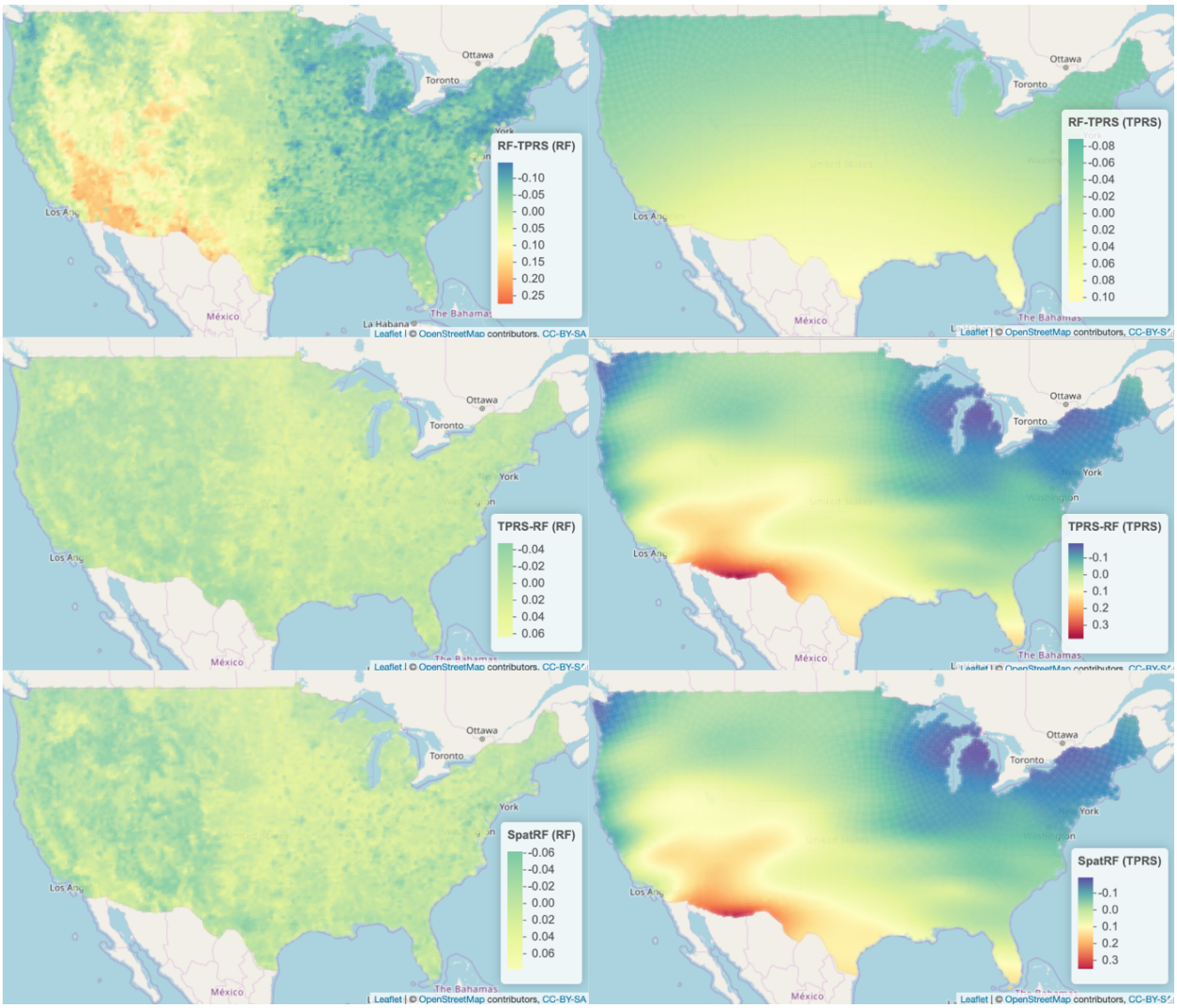

Figure 2.9: RF and TPRS components for Organic Carbon concentrations (in μg/m^3). Maps on the left side show the RF component of the additive model while maps on the right show the TPRS estimate. Each row is a different estimation order, top: RF first, then TPRS, middle: TPRS first, then RF, bottom: SpatRF-NP which jointly estimates the RF and TPRS components

Chapter 3

PENALIZED UNIVERSAL KRIGING

3.1 Introduction

Universal kriging contains a linear mean structure on covariates where observations are subject to variation from the linear model by a realization of a spatial process, and the unknown parameters are estimated by maximum likelihood. In recent years, it has become common to obtain large numbers of geographic covariates, often of greater numbers than observations. In this case, optimization of the universal kriging model is complicated as the likelihood becomes degenerate and requires pre-processing or regularization of the covariates.

Model selection with large numbers of co-linear predictors typically involve: (a) dimension reduction to a smaller number of composite covariate scores, (b) variable or subset selection, or (c) shrinkage or regularization. A number of studies have explored dimension reduction techniques such as partial least squares (PLS) [88] or principal component analysis (PCA) [62]. Dimension reduction techniques are employed to create a small number of composite covariate scores. Since the goal is high quality predictions, supervised approaches like PLS would seem to be preferred over unsupervised methods such as PCA, as identifying features which are associated with pollutant levels would suggest better predictions. However, in practice PLS often performs no better than PCA (Section 6.3.2, [37]).

A disadvantage of dimension reduction approaches is that constructing composite covariates scores must be done prior to estimating the residual spatial process. The set of composite covariates scores which maximizes the predictive ability of universal kriging model finds the optimal combination of the fixed effects and spatial process. This point is illustrated through a variogram in Figure 3.1, where the desired combination of the fixed effects and the residual spatial process minimize the size of the unexplained variance (or nugget

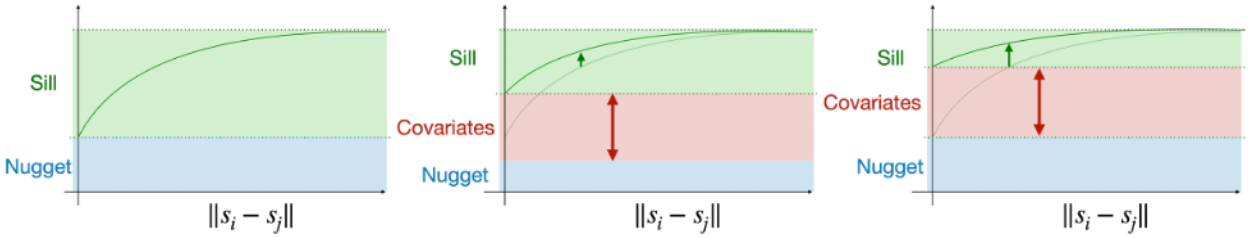

Figure 3.1: **Left**: A typical variogram used in spatial statistics, with the distance between observations on the x-axis, and the variance of the process on the y-axis. The sill represents the proportion of variance attributable to the spatial process, while the nugget effect represents the proportion of unexplained variance. **Center**: The optimal combination of the fixed effects and the spatial process for prediction minimizes the size of the nugget. **Right**: Pre-processing approaches, which are unable to adjust for the spatial process to be estimated, are *greedy* in that they select covariates which maximize the proportion of variance attributable to the covariates but do not adjust for the spatial process, can result in sub-optimal estimates of the additive model.

effect). Pre-processing approaches, such as PLS or PCA, are not able to adjust for the residual spatial process since the spatial process is jointly estimated with the fixed effects and is unknown a priori. This would not be a problem if the covariates were independent of the residual spatial process, but this assumption is violated in many applications. GIS covariates, such as meters to airports, elevation, or population buffers, clearly have spatial structure as they define proximity measures. Because of this, ignoring the residual spatial process when pre-processing GIS covariates can lead to an inefficient estimates.

An alternative method of dealing with large numbers of co-linear predictions is variable or subset selection. While this approach is often viewed as the "holy grail" for sparse models, it is computationally intractable at $\mathcal{O}(2^p n^3)$ operations as subset selection approaches are a brute force search over all possible subsets of selected covariates. Instead, forward stepwise selection is often employed [58, 1]. While this approach is not nearly as computationally intensive as best subset selection, it still requires searching through all p possible variables to determine which variable to add. The process of searching through all possible variables is repeated until some pre-determined number of covariates are selected, and this approach

can be computationally difficult for large datasets.

These types of variable selection procedures can be difficult to deploy in practice, as every new possible subset of covariates to be considered requires optimization of the associated residual spatial process as well. This is a well-known problem for a few reasons ([22], 4.2). First, Matern class covariance functions are over-parameterized and non-identifiable [42]. Second, optimization over Matern class covariance parameters are non-convex and optimization approaches often used, such as BFGS [50], are run with multiple starting parameters. While theoretically, one could guarantee convergence to the global optimum with an infinitely dense set of starting parameters, this is impractical and in practice these algorithms do not guarantee, and for majority do not attain, the global optimum value. Finally, these approaches are computationally intensive since inverting the spatial covariance matrix is $\mathcal{O}(n^3)$ and becomes computationally infeasible for moderately large values of n. This problem is an active area of research and a large number of methods have been proposed to address this problem [29].

Comparing subset selection and forward stepwise selection with shrinkage, Hastie et al. [28] showed that shrinkage approaches performed better than either subset selection or forward stepwise selection over a wide range of simulations. Shrinkage has become popular in statistical learning applications, although it has not seen wide use in spatial analysis due to the presence of spatial correlation. Unknown parameters for penalized regression are usually estimated by convex optimization and including a Matern class spatial covariance matrix results in a loss function that is no longer convex with respect to the unknown parameters. In Section 3.2, we detail the equivalence between likelihood based universal kriging models using low rank approximations to the spatial covariance and penalized regression estimators. We demonstrate how this approach can incorporate a wide variety of shrinkage techniques from the statistical learning literature and be applied using existing software. In Section 3.3, we provide simulation results comparing shrinkage to dimension reduction and demonstrate data generating scenarios when each method is expected to do better or worse. In Section 3.4, we compare these different approaches in predicting annual averages for a variety of air

pollutants, elemental carbon (EC), organic carbon (OC), Silicon (Si), and Sulfur (S), particulate matter less than 2.5 microns in diameter ($PM_{2.5}$), and nitrous oxides (NO_2), across the continental United States for 2009-2010. We end the chapter with a discussion of the advantages of our method and aspects for future work.

3.2 Methods

3.2.1 Low Rank Approximations of the Spatial Covariance for the Large n Problem

Consider n observations at monitoring locations $\mathbf{s} = \{s_1, \ldots, s_n\}$ across the domain of interest and let $\mathbf{Z}(\mathbf{s})$ be the observed process. We are interested in modeling the underlying surface $\mathbf{Y}(\mathbf{s})$ on the basis that observations are made with measurement error (or nugget effect), $\boldsymbol{\epsilon}(\mathbf{s}) \sim N(0, \tau^2\mathbf{I}_n)$

$$\mathbf{Z}(\mathbf{s}) = \mathbf{Y}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}). \tag{3.1}$$

Universal kriging models the underlying surface by combining the fixed effects, a linear function on geographic covariates, $\mathbf{X}(\mathbf{s})$, with a spatially correlated zero mean Gaussian process $\boldsymbol{\nu}(\mathbf{s}) \sim N\left(0, \sigma^2\boldsymbol{\Sigma}\right)$ as

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \boldsymbol{\nu}(\mathbf{s}). \tag{3.2}$$

In Section 2.2.4 we demonstrated that universal kriging can be equivalently formulated as a spatial mixed model,

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{S}(\mathbf{s})\boldsymbol{\eta} + \boldsymbol{\epsilon}(\mathbf{s}). \tag{3.3}$$

Under Eq. 3.3, the unknown parameters to be estimated are $(\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2, \tau^2)$. The fixed effects coefficients, $\boldsymbol{\beta}$, are estimated by their best linear unbiased estimate (BLUE), the generalized least squares estimate:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T(\mathbf{s})\mathbf{V}^{-1}(\sigma^2, \tau^2)\mathbf{X}(\mathbf{s})\right)^{-1}\mathbf{X}^T(\mathbf{s})\mathbf{V}^{-1}(\sigma^2, \tau^2)\mathbf{Z}(\mathbf{s}),$$

while the random effects coefficients , $\boldsymbol{\eta}$, are estimated by their best linear unbiased predictor (BLUP):

$$\hat{\boldsymbol{\eta}} = \sigma^2\mathbf{S}^T(\mathbf{s})\mathbf{V}^{-1}(\sigma^2, \tau^2)\left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}}\right).$$

When the goal is to predict values of the underlying process at unobserved locations, we can reduce optimization of the *covariance parameters* (σ^2, τ^2) to a single parameter,$\lambda_{\boldsymbol{\eta}}$ [23]. To demonstrate this, define the ratio of the variance attributable to the spatial process to the nugget as $\lambda_{\boldsymbol{\eta}} = \frac{\tau^2}{\sigma^2}$.

$$\begin{aligned}\mathbf{V}(\sigma^2, \tau^2) &= \sigma^2 \mathbf{S}(\mathbf{s})\mathbf{S}^T(\mathbf{s}) + \tau^2 \mathbf{I}_n \\ &= \sigma^2 \left(\mathbf{S}(\mathbf{s})\mathbf{S}^T(\mathbf{s}) + \lambda_{\boldsymbol{\eta}} \mathbf{I}_n\right)\end{aligned}$$

Define $\boldsymbol{\Omega}(\lambda_{\boldsymbol{\eta}}) = \mathbf{S}(\mathbf{s})\mathbf{S}^T(\mathbf{s}) + \lambda_{\boldsymbol{\eta}}\mathbf{I}_n$, thus $\mathbf{V}(\sigma^2, \tau^2) = \sigma^2 \boldsymbol{\Omega}(\lambda_{\boldsymbol{\eta}})$. It is now easy to see that the BLUE for the fixed effects and the BLUP for the random effects are equivalently:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^T(\mathbf{s})\boldsymbol{\Omega}^{-1}(\lambda_{\boldsymbol{\eta}})\mathbf{X}(\mathbf{s})\right)^{-1} \mathbf{X}^T(\mathbf{s})\boldsymbol{\Omega}^{-1}(\lambda_{\boldsymbol{\eta}})\mathbf{Z}(\mathbf{s}) \\ \hat{\boldsymbol{\eta}} &= \mathbf{S}^T(\mathbf{s})\boldsymbol{\Omega}^{-1}(\lambda_{\boldsymbol{\eta}}) \left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}}\right).\end{aligned}$$

and the only parameter relevant for prediction is $\lambda_{\boldsymbol{\eta}}$.

As noted in [60, 69] using spatial processes for prediction has a connection to smoothing splines, and the spatial random effect $\hat{\boldsymbol{\eta}}$ can also be interpreted as a penalized regression estimator with penalty $\lambda_{\boldsymbol{\eta}}$. By Henderson's justification [67], optimization of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$ is equivalent to finding:

$$\underset{\boldsymbol{\beta},\boldsymbol{\eta}}{\operatorname{argmin}} \|\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta} - \mathbf{S}(\mathbf{s})\boldsymbol{\eta}\|_2^2 + \lambda_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_2^2, \tag{3.4}$$

The loss function for the penalized form of the spatial mixed effects model (Eq. 3.4), is a convex function of the parameters $\boldsymbol{\beta}, \boldsymbol{\eta}$ with a single parameter to be estimated $\lambda_{\boldsymbol{\eta}}$, where the random effects are now equivalently fixed effects with tuning parameter $\lambda_{\boldsymbol{\eta}}$. Stein [74] suggests caution when using when using low rank approaches with likelihood based methodology, as the behavior becomes dominated by fine scale behavior. Viewing Eq. 3.4 as penalized regression we estimate $\lambda_{\boldsymbol{\eta}}$ by cross-validation instead of maximum likelihood; Gerber and Nychka[23] showed that similar parameter estimates are obtained using either approach. While maximum likelihood for mixed models are traditionally estimated using the EM algorithm, which can require many iterations to converge to a local optimum and has highly

conditional convergence properties [4], the penalized formulation allows us to take advantage of the wide array of efficient convex optimization approaches that are guaranteed to converge to the global optimum.

3.2.2 Model Selection and Regularization for the Large p Problem

Pre-Processing Approaches for Universal Kriging

Both dimension reduction and subset selection approaches handle large number of covariates by reducing the dimensionality of $\mathbf{X}(\mathbf{s})$ to a smaller number of features or feature scores by pre-processing the covariates using a linear transformation:

$$\tilde{\mathbf{X}}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\mathbf{P}, \quad \mathbf{P} \in \mathbb{R}^{p \times r}.$$

Dimension reduction approaches find r composite feature scores which preserve the largest among of information in $\mathbf{X}(\mathbf{s})$ (PCA), or maximize the correlation with the observations (PLS). In either approach, we apply a transformation matrix $\mathbf{P}$ (commonly known as *principal loadings*) where each of the r columns of $\mathbf{P}$ takes a linear combination of the p covariates. The number of components to use, r, is commonly selected by cross-validation.

In subset selection, $\mathbf{P}$ selects the "best" subset of features of $\mathbf{X}(\mathbf{s})$. Since best subset selection is computationally infeasible, a number of alternative optimization approaches to subset selection have been proposed. The most common approach used is forward stepwise selection [1, 58], although other heuristic methods have been proposed [56]. While these approaches reduce the computational burden, these techniques are still computationally intensive and are not included for comparison in this paper.

Comparing the two approaches, dimension reduction techniques create a lower dimensional representation of the full covariate set by constructing composite features scores, while subset selection allows for parsimonious and simpler to interpret models. Both dimension reduction and subset selection require the transformation matrix $\mathbf{P}$ to be pre-specified before estimation, then the regularized form of the covariates $\tilde{\mathbf{X}}(\mathbf{s})$ is used as fixed effects in a universal kriging model.

Adaptive Estimation using Shrinkage

Both dimension reduction and subset selection require the covariates to be pre-processed prior to parameter estimation in a universal kriging model. In these scenarios, shrinkage can be desirable - rather than requiring regularization of the covariates to be performed prior to estimating the model, shrinkage applies regularization of the covariates concurrently with estimation of the model parameters. Shrinkage estimators can also address the same problems pre-processing approaches do. Dimensions reduction aims to leverage large numbers of weak predictors by constructing $r < p$ composite features which maintain as much of the information as possible from the full dataset. Alternatively, ridge regression can handle large numbers of covariates by shrinking the coefficients towards a common value (often zero). Subset selection aims to find a small number of the most important predictors. Analogously, LASSO [78] applies shrinkage to the coefficient estimates which encourages the majority of them to be zero so that only a small subset of important features have non-zero coefficients.

While shrinkage has seen wide use in the statistical learning literature, it is not commonly employed in spatial applications because of the presence of the spatial process. By taking advantage of low-rank approximations to the spatial covariance matrix in Section 3.2.1, we provide a framework which can be built on to apply a wide variety of penalized regression estimators while allowing for the spatial process. Since the penalized form of the spatial mixed model (Eq. 3.4) is convex in $(\boldsymbol{\beta}, \boldsymbol{\eta})$ penalization of the fixed effects may be added as:

$$\underset{\boldsymbol{\beta},\boldsymbol{\eta}}{\operatorname{argmin}} \|\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta} - \mathbf{S}(\mathbf{s})\boldsymbol{\eta}\|_2^2 + \lambda_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_2^2 + \lambda_{\boldsymbol{\beta}} P(\boldsymbol{\beta}). \tag{3.5}$$

By composition rules of convex functions, for any convex penalty function on $P(\boldsymbol{\beta})$, Eq. 3.5 is still convex and standard convex optimization approaches, for example gradient descent, may be applied. A wide variety of penalties can be applied, for example LASSO [77], group LASSO [72], elastic net [91], and selection of penalties for specific scenarios should be determined by prior knowledge on the covariates collected and hypothesized relationships between them and the spatial quantity to be predicted.

In environmental epidemiological applications, the purpose of these models is to predict

air pollutant exposures for study participants rather than examine the association between land use covariates and pollutant levels in the presence of spatial correlation. Because of this, we select tuning parameters by cross-validation to maximize out-of-sample prediction accuracy. The penalized formation of the spatial linear mixed model in Eq. 3.5 is similar to generalized additive model selection [53, 13]. Marra and Wood [53] examined a wide variety of ways to estimate $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\eta}}$ and found that treating both as penalties and selecting both their values by cross-validation were optimal for predictive mean squared error, and is the approach used in this paper.

3.2.3 Alternative Optimization for Compatibility with Existing Software

Any convex penalty on $\boldsymbol{\beta}$ or $\boldsymbol{\eta}$ can be added in these scenarios, for example, elastic net [91] and sparse group LASSO [72]. However, many of the penalization techniques already have existing libraries and packages which are optimized to run efficiently. For ridge and LASSO regression, the widely used `glmnet` package is highly optimized for computational convenience and employs warm starts to allow for cross-validation over a range of candidate $\lambda_{\boldsymbol{\beta}}$ values in just about the same amount of time it takes to optimize for the single, smallest $\lambda_{\boldsymbol{\beta}}$ [19]. It is often the case that one would like to use existing software, but modify it to allow for joint estimation of the residual spatial process. Here, we propose a simple approach to use existing penalized regression software to examine different penalties on $\boldsymbol{\beta}$ for universal kriging models.

Note that, for a fixed value of $\lambda_{\boldsymbol{\eta}}$, the objective function in Eq. 3.5 is equivalent to:

$$\|\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta} - \mathbf{S}(\mathbf{s})\boldsymbol{\eta}\|_2^2 + \lambda_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_2^2 + \lambda_{\boldsymbol{\beta}}P(\boldsymbol{\beta}) = (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})^T \boldsymbol{\Omega}(\lambda_{\boldsymbol{\eta}})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})^T + \lambda_{\boldsymbol{\beta}}P(\boldsymbol{\beta}),$$

As $\boldsymbol{\Omega}(\lambda_{\boldsymbol{\eta}})$ is a symmetric positive semidefinite covariance matrix, we take its singular value decomposition $\boldsymbol{\Omega}(\lambda_{\boldsymbol{\eta}}) = \mathbf{TDT}^T$, where the columns of $\mathbf{T}$ are orthonormal and $\mathbf{D}$ is a diagonal $k \times k$ matrix with diagonal elements $d_1, \ldots, d_k > 0$. We define the square root of

$\mathbf{D}$ as $\mathbf{D}^{-1/2}$ with diagonal elements $\frac{1}{\sqrt{d_1}}, \ldots, \frac{1}{\sqrt{d_k}}$ and let $\boldsymbol{\Omega}^{-1/2}(\lambda_{\boldsymbol{\eta}}) = \mathbf{D}^{-1/2}\mathbf{T}$. Now,

$$\begin{aligned}(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})^T \boldsymbol{\Omega}^{-1}(\lambda_{\boldsymbol{\eta}}) (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}} P(\boldsymbol{\beta}) &= \left\| \boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}}) (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}) \right\|_2^2 + \lambda_{\boldsymbol{\beta}} P(\boldsymbol{\beta}) \\ &= \left\| \tilde{\mathbf{Z}}(\mathbf{s}) - \tilde{\mathbf{X}}(\mathbf{s})\boldsymbol{\beta} \right\|_2^2 + \lambda_{\boldsymbol{\beta}} P(\boldsymbol{\beta}), \\ &\quad \tilde{\mathbf{Z}}(\mathbf{s}) = \boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}})\mathbf{Z}(\mathbf{s}), \tilde{\mathbf{X}}(\mathbf{s}) = \boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}})\mathbf{X}(\mathbf{s}).\end{aligned}$$

This approach is equivalent to "de-correlating" the observations, $\mathbf{Z}(\mathbf{s})$, and the GIS covariates, $\mathbf{X}(\mathbf{s})$ before applying penalization [36] and motivates an alternative optimization approach which leverage existing software packages. Over a range of $\lambda_{\boldsymbol{\eta}}$, first construct the "de-correlating matrix" $\boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}})$ and transform the observations, $\tilde{\mathbf{Z}}(\mathbf{s}) = \boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}})\mathbf{Z}(\mathbf{s})$, and the covariates, $\tilde{\mathbf{X}}(\mathbf{s}) = \boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}})\mathbf{X}(\mathbf{s})$. Using the transformed version of the observation and covariates, any existing software for different penalties on $\boldsymbol{\beta}$ can now easily be applied. This process is summarized in Algorithm 2.

Algorithm 2 Penalized Universal Kriging Optimization using Existing Software

1. For $i = 1, 2, \ldots$
   (a) For each $\lambda_{\boldsymbol{\eta}}$, construct $\boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}})$.
   (b) "De-correlate" the observations and covariates as
   $$\tilde{\mathbf{Z}}(\mathbf{s}) = \boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}})\mathbf{Z}(\mathbf{s}), \quad \tilde{\mathbf{X}}(\mathbf{s}) = \boldsymbol{\Omega}^{-\frac{1}{2}}(\lambda_{\boldsymbol{\eta}})\mathbf{X}(\mathbf{s}).$$
   (c) For $j = 1, 2, \ldots$
      i. Use the existing software package to optimize
      $$\left\| \tilde{\mathbf{Z}}(\mathbf{s}) - \tilde{\mathbf{X}}(\mathbf{s})\boldsymbol{\beta} \right\|_2^2 + \lambda_{\boldsymbol{\beta}}^j P(\boldsymbol{\beta}).$$
2. Select the pair $(\hat{\lambda}_{\boldsymbol{\eta}}, \hat{\lambda}_{\boldsymbol{\beta}})$ which minimizes cross-validated prediction error.

3.2.4 Relationship between Penalized Regression and Linear Mixed Models

In Section 3.2.1, we used the relationship between the spatial linear mixed model and ridge regression to transform it into a penalized regression problem. For certain shrinkage penalties, this relationship between penalized regression estimators and mixed models means we can construct parametric model based "equivalents", in that the estimating equations are the same for both. For example, instead of treating $\boldsymbol{\beta}$ as a fixed effect in (3.3), let $\boldsymbol{\beta} \sim N(\mathbf{0}, \zeta^2 \mathbf{I}_p)$. In this case, the estimating equations for $(\boldsymbol{\beta}, \boldsymbol{\eta})$ are:

$$\hat{\boldsymbol{\beta}} = \left(X^T(s)\boldsymbol{\Omega}^{-1}X(s) + \frac{\sigma^2 + \tau^2}{\zeta^2}\mathbf{I}_p \right)^{-1} X(s)^T \boldsymbol{\Omega}^{-1} Y(s)$$

$$\hat{\boldsymbol{\eta}} = \left(\mathbf{S}^T(\mathbf{s})\mathbf{S}(\mathbf{s}) + \frac{\tau^2}{\sigma^2}\mathbf{I}_k \right)^{-1} \mathbf{S}^T(\mathbf{s}) \left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}} \right)$$

This is the "Bayesian ridge" model, in that the minimizers for $(\boldsymbol{\beta}, \boldsymbol{\eta})$ in Eq. 3.4 are equivalent with $\lambda_{\boldsymbol{\beta}} = \frac{\sigma^2+\tau^2}{\zeta^2}$ and $\lambda_{\boldsymbol{\eta}} = \frac{\tau^2}{\sigma^2}$. By careful selection of the prior on $\boldsymbol{\beta}$, one can, in a sense, induce a penalized regression estimator. For ridge regression, this prior is a normal prior on $\boldsymbol{\beta}$, for LASSO, a double exponential prior can be used leading to the "Bayesian LASSO" [78, 63].

There are a few subtle, but important, differences between the linear mixed model and penalized regression estimator. In a linear mixed model parameters are selected by maximum likelihood or the maximum a priori estimate, while in penalized regression they are selected by cross-validation. In many applications of universal kriging, the goal is to provide area-specific pollutant concentrations rather than to estimate the association between pollutants and covariates [81]. We wish to model the observed process with a "process model" comprised of a surface modeled using covariates and a spatially stochastic term [15]. While likelihood estimates have many desirable properties, such as theoretical consistency under the correctly specified model, here we are not interested in learning about the association between the covariates and pollutant concentrations in the presence of spatially correlated noise, rather, we use the geographic covariates and the spatially correlated residual process to model the observed process. In this case, parameters selected by cross-validation select

their "prediction-optimal" values, which are not necessarily the same as the consistent likelihood based estimates. In a related problem, Yang [87] showed that the Bayseian Information Criteria, which is model consistent, and Akaike Information Criterion, which is mean average squaerd error optimal, cannot be shared; that is, for any model selection criterion to be consistent it must be sub-optimal in terms of mean average squared error.

Another important difference is that likelihood based approaches are able provide measures of uncertainty. However, it is not easy to interpret the residual spatial process as a "random error" since it is difficult to imagine the data generating mechanism that might give rise to such a process [32], thus inference on the variance of the spatial random effects is of little concern. While it is often desirable to provide uncertainty estimates to describe the degree of confidence one has in estimates from their model, our focus is on the development of models to be used in environmental epidemiology studies. In this setting, there are no natural ways of carrying through the error from a prediction model. Although measurement error models are typically employed, measurement error is the distribution of the estimated pollutant levels around the truth rather than uncertainty of the predictions themselves. For environmental epidemiology applications these measures of uncertainty are not useful and thus the only quantity we can measure is the raw prediction accuracy of the estimates themselves. In this setting, we believe penalized regression to be the preferred approach over likelihood based optimization. If the goal is strictly prediction accuracy, penalized approaches are much easier to compute and are targeted to minimize the quantity of interest.

3.3 Simulation Study

We conduct a set of simulations to compare the different model selection and regularization approaches for geographic covariates in universal kriging. Datasets for simulations are created on a grid of points over the continental United States spaced at 25km intervals and GIS covariates at these locations are provided from ArcGIS 10.2.

For each simulation, we constructed a fixed exposure surface, $\mathbf{Z}(\mathbf{s})$, from an additive

model of a linear function of GIS covariates, $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$, and a fixed realization of a Gaussian process with exponential covariance process, $\boldsymbol{\nu}(\mathbf{s})$. We considered three different types of generating functions on the covariates

1. *sparse*: a linear combination of 10 randomly sampled covariates where each covariate is scaled to have the same contribution to the overall variance of the signal.

2. *dense*: a linear combination of all the covariates where the coefficients $\boldsymbol{\beta} \sim (0, 1)$.

3. *lowrank*: apply PCA to the full GIS covariate matrix, then each PC is scaled to have the same contribution to the overall variance of the signal.

For our simulations we compared the following six methods:

1. OK: Ordinary Kriging without covariates

2. UK-PCA: Universal Kriging with dimension reduction on covariates by PCA. The number of PC's are chosen by cross-validation.

3. UK-PLS: Universal Kriging with dimension reduction on covariates by PLS. The number of PC's are chosen by cross-validation.

4. SpatRidge-Lklhd: Universal Kriging with covariates treated as normally distributed random effects (Section 3.2.4).

5. SpatRidge-CV: Penalized form of spatial linear mixed model with ridge penalty on the covariates. Both tuning parameters are chosen by cross-validation. Optimization follows Section 3.2.3 using the R package `glmnet`.

6. SpatLasso-CV: Penalized form of spatial linear mixed model with lasso penalty on the covariates. Both tuning parameters are chosen by cross-validation. Optimization follows Section 3.2.3 using the R package `glmnet`.

Three of these methods are commonly used in the literature (OK, UK-PCA, and UK-PLS), and the other three are detailed in this paper (SpatRidge-Lklhd, SpatRidge-CV, SpatLasso-CV). We include the likelihood and penalized versions of ridge regression in order to compare how equivalent estimators with parameters selected by cross-validation and maximum likelihood perform against each other. The likelihood based lasso is not included as studies suggest that superior methods exist for model based variable selection [17, 8]. Further, including the likelihood based lasso would require a complete Bayesian analysis to be run in each iteration of our simulations and would require considerable computational resources.

For our simulations we specified an exponential covariance matrix in all methods. For the likelihood based optimization approaches, (OK, UK-PCA, UK-PLS, SpatRidge-Lklhd), the range parameter is estimated by maximum likelihood along with the covariance parameters (σ^2, τ^2). For the penalized versions (SpatLasso-CV, SpatRidge-CV), the range is set to the maximum distance between observed points as suggested in Kammann and Wand [39]. This choice of range parameter is arbitrary and one could consider alternative approaches to selecting the range parameter. However, this parameter is purposefully mis-specified, as the true range used to generate the spatial process is one-fourth the maximum distance between points, to prevent differences in prediction accuracy of the various approaches be affected by user input.

To estimate the expected prediction test error of each method in predicting pollutant concentrations at a new point, we sample 200 unobserved points, $\mathbf{Y}(\mathbf{s}_{\text{test}})$, and, averaged over different samples of observed training points, report the average RMSPE. Density plots of average RMSPE for each method in each of these scenarios are shown in Figure 3.2.

Using OK as a baseline, all methods that incorporate geographical covariates have better prediction accuracy than OK, which does not use covariates. Across each of the generating functions, the same relationship between prediction accuracy holds. SpatLasso-CV consistently has the lowest prediction error, followed by the likelihood and penalized version of ridge regressions (SpatRidge-Lklhd and SpatRidge-CV), with Universal Kriging after dimension

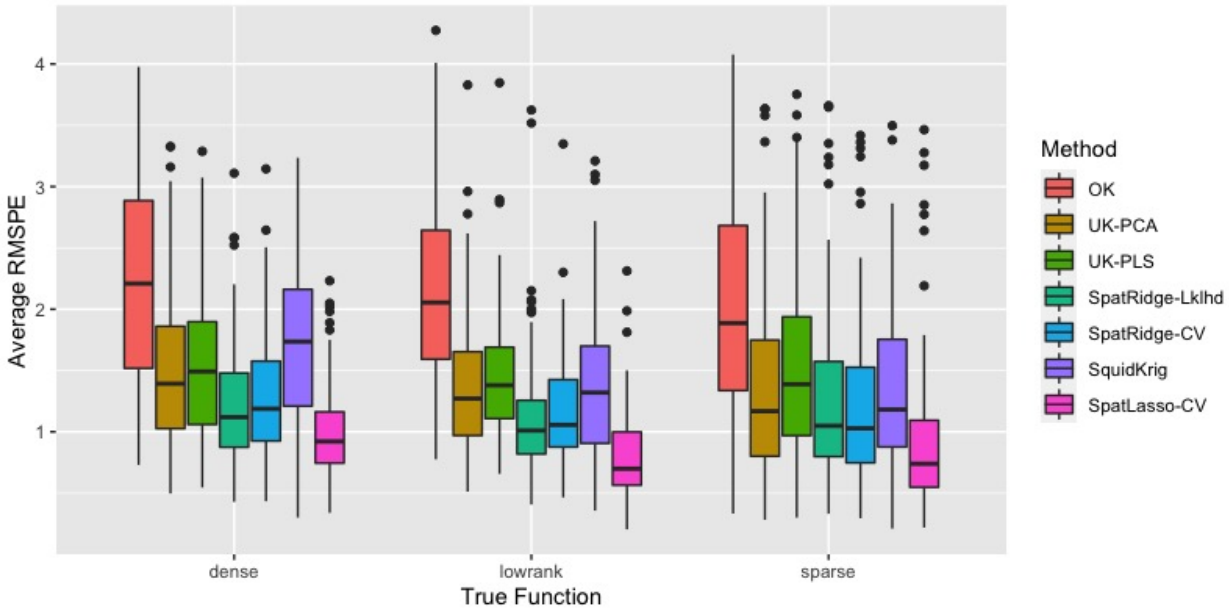

Figure 3.2: Simulation Results: Each point represents the average RMSPE on the validation points from different sampled training points. The box and whisker summarizes the prediction accuracy of the method on different simulated surfaces for each of the different generating functions, with dense functions on the left, low rank functions in the middle, and sparse functions on the right. The box shows the 25^{th}, median, and 75^{th} percentiles, and the whisker shows the most extreme value within 1.5 times the range of the quartiles. Each point in the boxplot represents the average RMSPE on the fixed 200 validation points over 50 repeated samples.

reduction by PCA (UK-PCA) or PLS (UK-PLS) having slightly lower prediction accuracy.

Comparing SpatRidge-Lklhd and SpatRidge-CV, the same model with different ways of estimating model parameters, there are little to no differences in prediction accuracy between these two approaches. These results support the findings of Gerber and Nychka [23], where they reported similar estimates for kriging when estimating parameters by maximum likelihood or cross-validation. In our simulations, we did not find fixing the (mis-specified) range parameter to affect prediction accuracy significantly, since the range parameter is estimated in the likelihood version while it is fixed in the cross-validation approach. These results are in line with Olives et al. [61], where they experimented with different rules for picking the range parameter and did not report any significant differences.

Looking at the dimension reduction methods, dimension reduction by PCA results in better prediction accuracy than PLS. PCA is unsupervised, in that dimension reduction is done without looking at the observations, while PLS selects components aimed at explaining the most variance in the observations. While PLS would appear to be desirable, since ideally it would identify components that are informative of features associated with the pollutant of interest, in our simulations it appears that it is unable to identify any ore informative features then simply using PCA alone. We note that this observation is not unique as it is also noted in James et al.[37].

Our intention was to demonstrate that each method performed best when the assumptions used to regularize the covariates matched the data generating mechanism, but this did not hold in our simulations across a wide variety of settings. In the *dense* scenario, we note that the covariates here are actual GIS covariates, and not independently drawn random covariates. Because of this, it is likely that the covariates are often highly correlated, for example, population within 10km is likely correlated with higher numbers of roads and less green space. In this case, even though our simulation generated each of the covariates to have a small effect, correlation between covariates means a small number of covariates can capture most of the variation in land use. For our *lowrank* simulations, we note that while the covariates are generated from a low rank matrix of the full covariate set, we only draw

a smaller sample of covariates. There are no known consistency results for PCA, in other words, there are no guarantees that the same principal components should be estimated in a sample of the full matrix. Further, principal components are a linear combination of the covariates so a function of the principal components is a dense function of the covariates. Our results do suggest that prediction accuracy and relative differences in prediction accuracy between methods are very similar in both the *dense* and *lowrank* simulations.

Further, we analyze the run time to demonstrate the scalability of the penalized approaches. We compare the same six methods, but this time vary the sample size from 100, 250, 500, 1000, and 2000. For fairness, each method using covariates is parallelized across 10 nodes. For the penalized approaches (SpatRidge-CV and SpatLasso-CV), the grid search among covariance parameters is parallelized. For dimension reduction (UK-PCA and UK-PLS), cross-validation to select the number of components is parallelized. The remaining likelihood based approaches (SpatRidge-Lklhd and OK) are parallelized for multiple starts of the optimization, which can be necessary as all optimum are checked to ensure the estimated parameters are not a saddle point and since there may be multiple local maximums. Figure 3.3 shows run time for each of the six methods, with run times shown on the $\log_{10}$ scale. For small sample sizes (100 or less), the likelihood based approaches are slightly faster than the penalized approaches, likely due to the fact that penalized approaches perform a grid search where 10-fold CV is performed for each point on the grid. By a sample size of 250, penalized approaches are already faster than likelihood based dimension reduction approaches, UK-PCA and UK-PLS, which are nearly identical and the two lines are nearly indistinguishable. As the sample size reaches 1000, the penalized approaches are as fast as computing the maximum likelihood estimates (OK and SpatRidge-Lklhd). SpatRidge-Lklhd is always slower than OK, as it involves estimating an additional parameter, the variance for the covariate effects. As the sample size increases to 2000, we start to see large gains using the penalized methods compared to the likelihood based approaches, as the run time is an order of magnitude lower. Comparing the two penalization approaches, SpatRidge-CV is always faster than SpatLasso-CV. The ridge penalty is smooth and closed form solutions

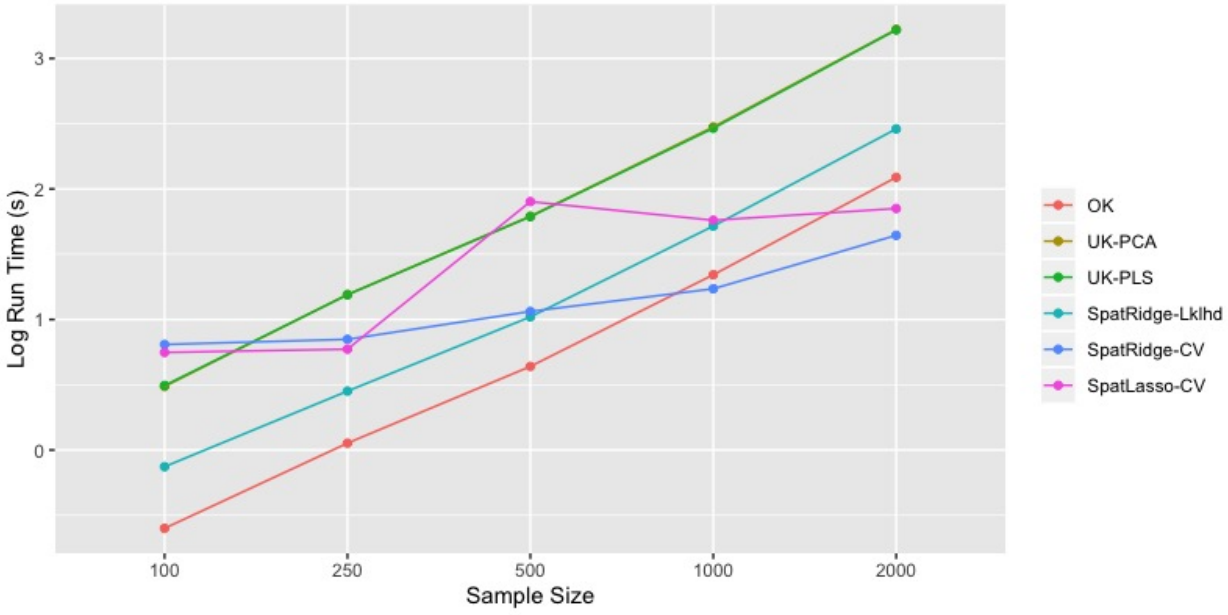

Figure 3.3: r

un-times of Penalized Universal Kriging to Likelihood Approaches]Simulation Results: Each point represents the average run time (in log(s)) to select the optimum model. UK-PCA and UK-PLS are nearly identical in run time, and the two lines lie on top one another.

exist for any fixed $\lambda_{\boldsymbol{\beta}}$. In contrast, LASSO regression is sensitive to the chosen tuning parameters to optimize over, and can become unstable for small values of $\lambda_{\boldsymbol{\beta}}$, likely causing the large variation in run time for smaller sample sizes. While the likelihood based approaches will become computationally infeasible rather quickly, the penalized approaches are able to scale for large samples as convex optimization algorithms are much more efficient then non-convex optimization approaches. The added computational burden of performing a two parameter grid search becomes negligible compared to the time needed to maximize the over-parameterized, non-convex likelihood surfaces, which require multiple starting values for optimization to find a local maximum.

3.4 Comparison of Methods for Pollutants across Continental United States in 2009-2010

We develop air pollution exposure models to reconstruct annual averages of NO_2, $PM_{2.5}$, and $PM_{2.5}$ sub-species EC, OC, Si, S. NO_2 annual measurements are obtained from the AQS network in 2010, and all included sites were operating for 244 or more days of the year following [88]. $PM_{2.5}$ were collected from AQS and Interagency Monitoring for Protected Visual Environments (IMPROVE) network in 2010, and all sites that had greater than 14 measurements were included following [70]. EC, OC, Si, S annual averages come from IMPROVE and Chemical Speciation Network (CSN) monitoring data from 2009-2010. Only monitors with at least 10 data points per quarter and no more than 45 days between consecutive measurements were included following Bergen et al. [6]. Si and S measurements were averaged over 01/01/2009–12/31/2009, while EC/OC consisted of measurements from 204 IMPROVE and CSN monitors averaged over 01/01/2009–12/31/2009, and measurements from 51 CSN monitors averaged over 05/01/2009 - 04/30/2010. Annual averages for all pollutants were square-root transformed prior to modeling.

Surface reconstruction accuracy of each method is assessed by comparing predictions generated from ten-fold cross-validation. Monitoring sites were randomly assigned to one of ten cross-validation groups. Each group is held out as a "test set" and observations in the remaining groups are used as a "training set" to fit the model and generate test set predictions using each of the seven methods to compare. In each fold, parameter selection is performed using only the training data, and this may result in different models and parameters being estimated in each cross-validation fold. Each group is used as a test set once to obtain predicted values for the entire data set. Performance of each model based on their cross-validated RMSPE is shown in Table 3.1.

Scatter plots showing observed concentrations against predictions for each method and pollutant is shown in Figures 3.4, 3.5, 3.6, 3.7, 3.8, 3.9. Further, predictions for each pollutant across the continental United States using every method is shown in Figures 3.10, 3.11, 3.12,

	Cross-Validated R^2					
	EC	OC	Si	S	$PM_{2.5}$	NO_2
OK	0.266	0.753	0.070	0.085	1.621	3.968
UK-PCA	0.151	0.568	0.069	0.076	1.136	2.369
UK-PLS	0.159	0.575	0.070	0.075	1.176	2.601
SpatRidge-Lklhd	0.143	0.550	0.070	0.077	1.140	2.618
SpatRidge-CV	0.144	0.552	0.070	0.075	1.138	2.285
SpatLasso-CV	0.142	0.559	0.070	0.077	1.089	2.397

Table 3.1: Ten-fold cross-validated prediction accuracy, summarized by RMSPE, of each method for EC, OC, Si, S, $PM_{2.5}$, and NO_2 collected by AQS, CSN, and IMPROVE monitoring networks from 2009-2010.

3.13, 3.14, 3.15.

The use of covariates does not seem to improve prediction accuracy for Si and S over simply kriging without covariates (OK), but leads to large gains in prediction accuracy for EC, OC, NO_2, and $PM_{2.5}$. Cross-validated RMSPE suggests SpatLasso-CV and SpatRidge-CV lead to a noticeable improvement in prediction accuracy over UK-PCA and UK-PLS for EC, although the differences in prediction accuracy is almost negligible for all other pollutants.

Focusing on estimates of EC in the midwest (Figure 3.10), estimates using SpatLasso-CV and SpatRidge-CV suggests EC levels in the region are lower and more similar to rural mountain west areas except for interstate highways, while UK-PLS and UK-PCA seem to suggest levels in that area are more similar to the southeast. Looking at the scatterplot (Figure 3.4), we note that predictions in the 0.5-1.0 range are tighter to the 1-1 line for SpatLasso-CV and SpatRidge-CV compared to UK-PLS and UK-PCA.

In all pollutants where the use of covariates shows large improvements in CV RMSPE, the penalized approaches always have the highest cross-validated prediction accuracy. For EC, OC, and $PM_{2.5}$, SpatLasso-CV is slightly higher than all competing methods, while

SpatRidge-CV is best for NO_2.

UK-PCA and UK-PLS again perform very similarly, supporting our findings from simulations and James et al. [37], where even though one might suspect that PLS would have result in prediction accuracy as it is a supervised approach and PCA is unsupervised, in our examples it does not seem to provide any additional benefit.

Comparing estimation of penalties/covariance parameters by cross-validation (SpatRidge-CV) against via likelihood (SpatRidge-Lklhd), the penalized version of ridge regression here always has better prediction accuracy than the likelihood based estimation approach. Although the likelihood based approach allowed for the range parameter to be estimated, fixing the range parameter in the penalized approach did not seem to affect prediction accuracy at all. These findings are in line with our simulations and the results of Olives et al. [61], where we also did not see much of a difference between these two methods even though the range parameter was purposefully mis-specified in the penalized approach.

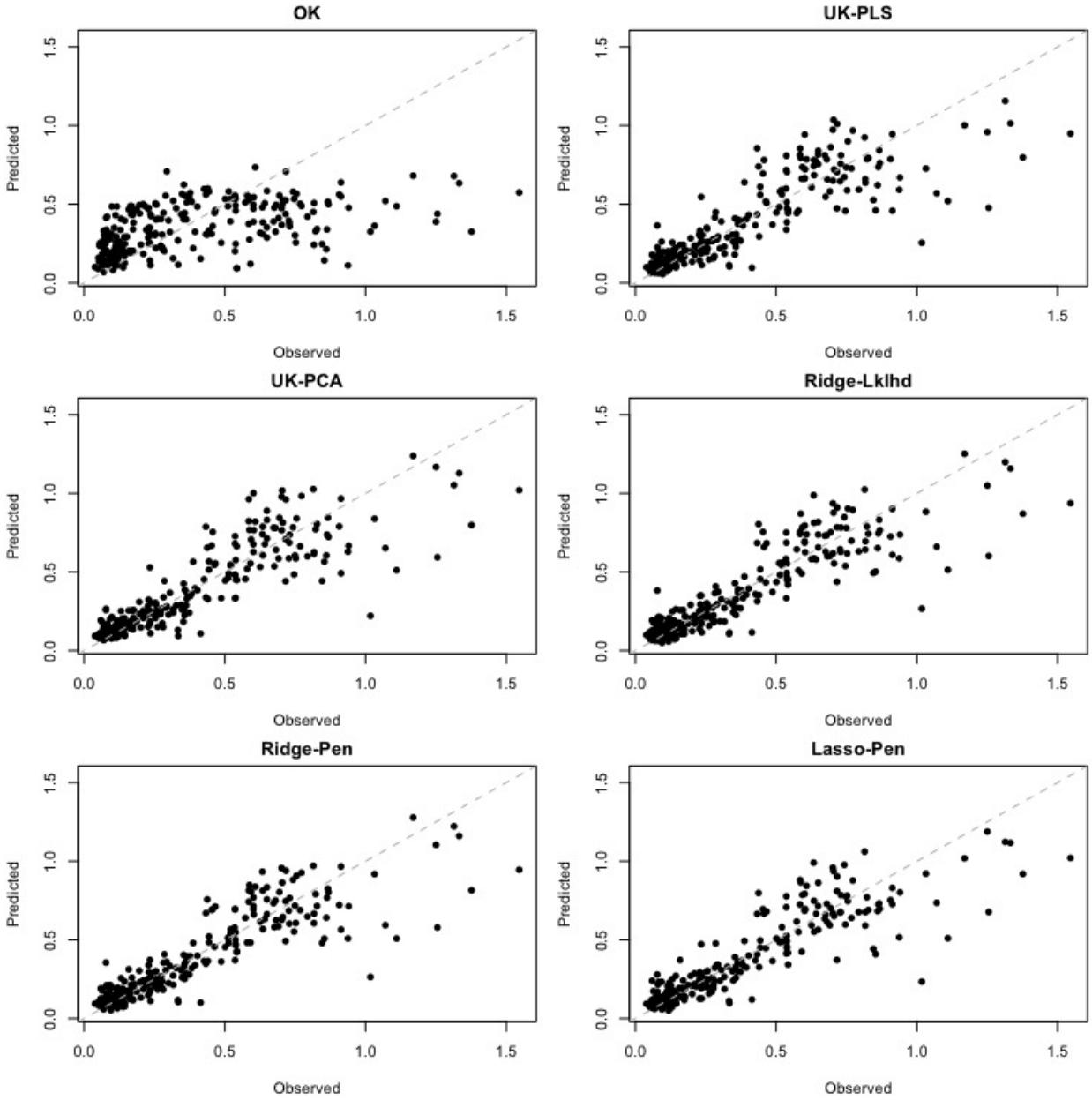

Figure 3.4: Cross-Validated Estimates against observed values for elemental carbon (EC), clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

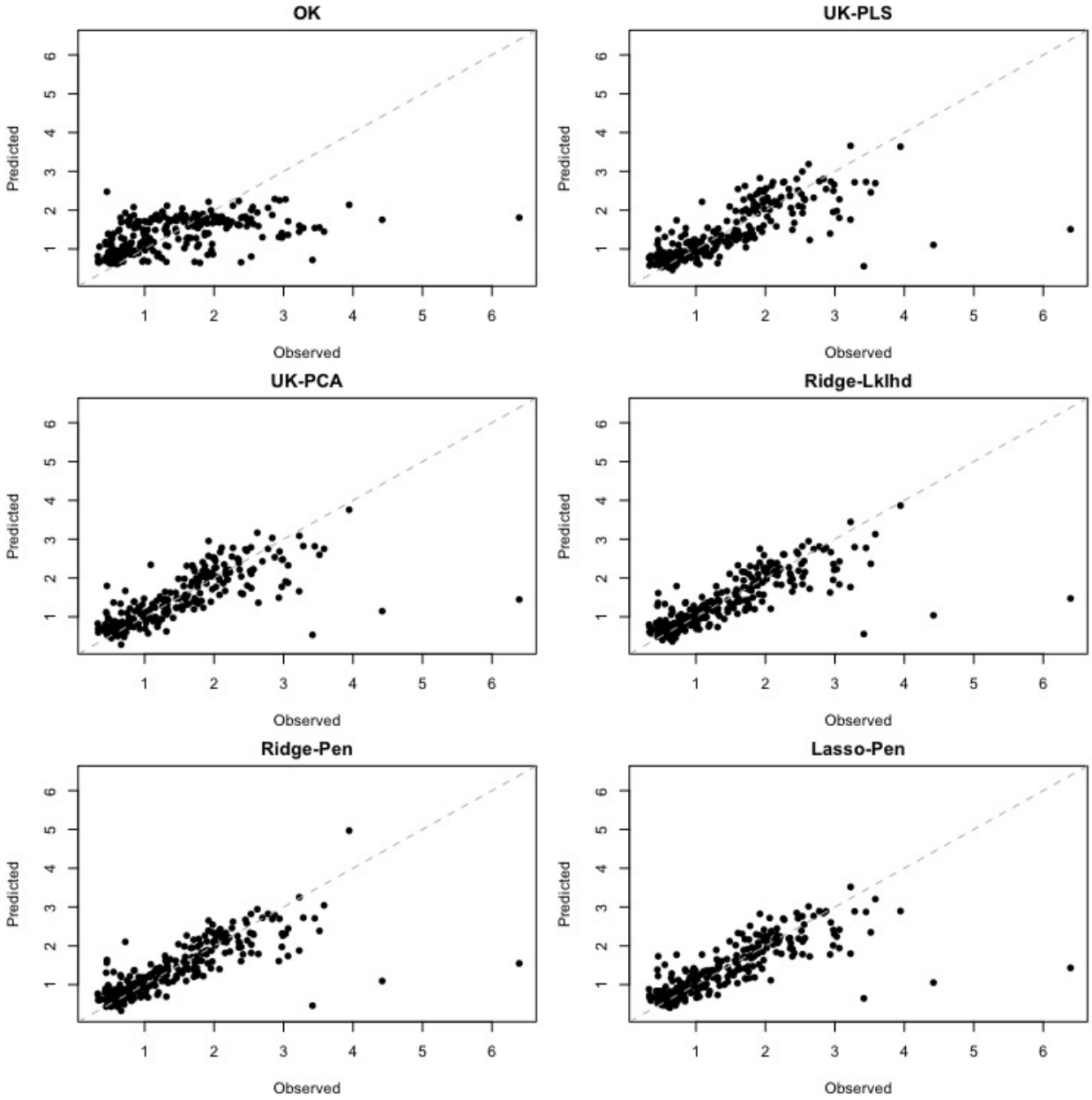

Figure 3.5: Cross-Validated Estimates against observed values for organic carbon (OC, clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

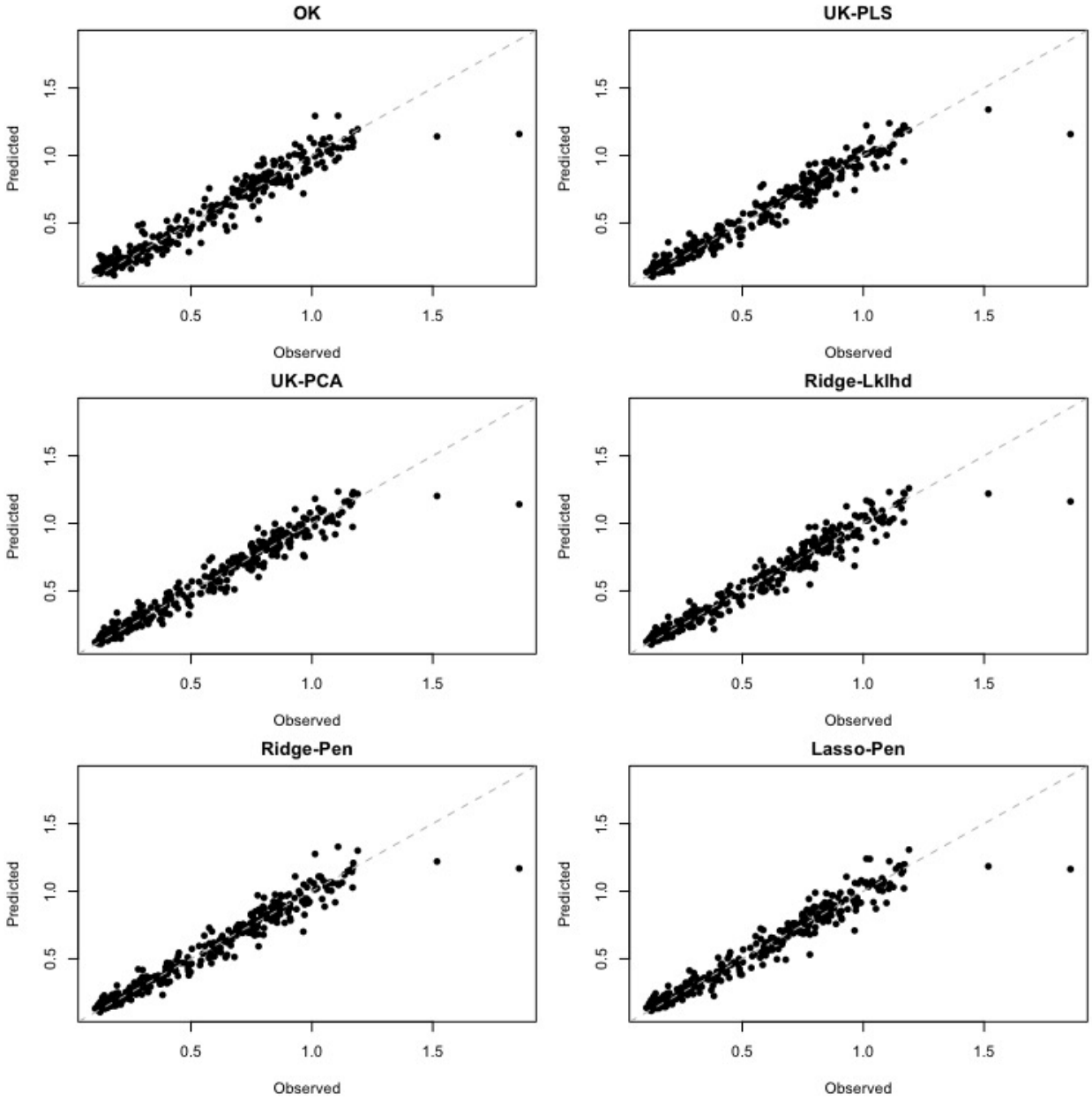

Figure 3.6: Cross-Validated Estimates against observed values for sulfur (S), clockwise from clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

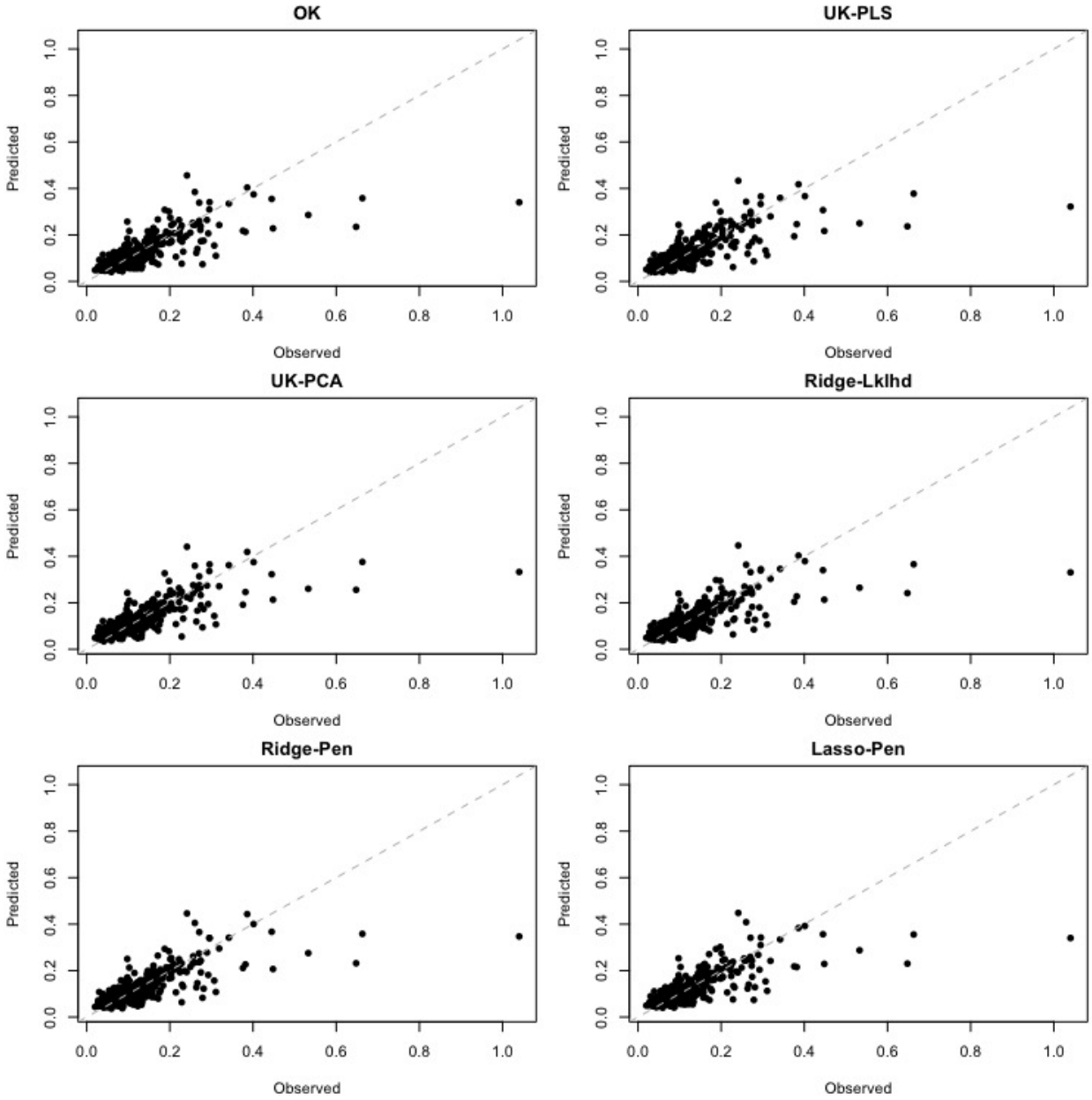

Figure 3.7: Cross-Validated Estimates against observed values for silicon (Si), clockwise from clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

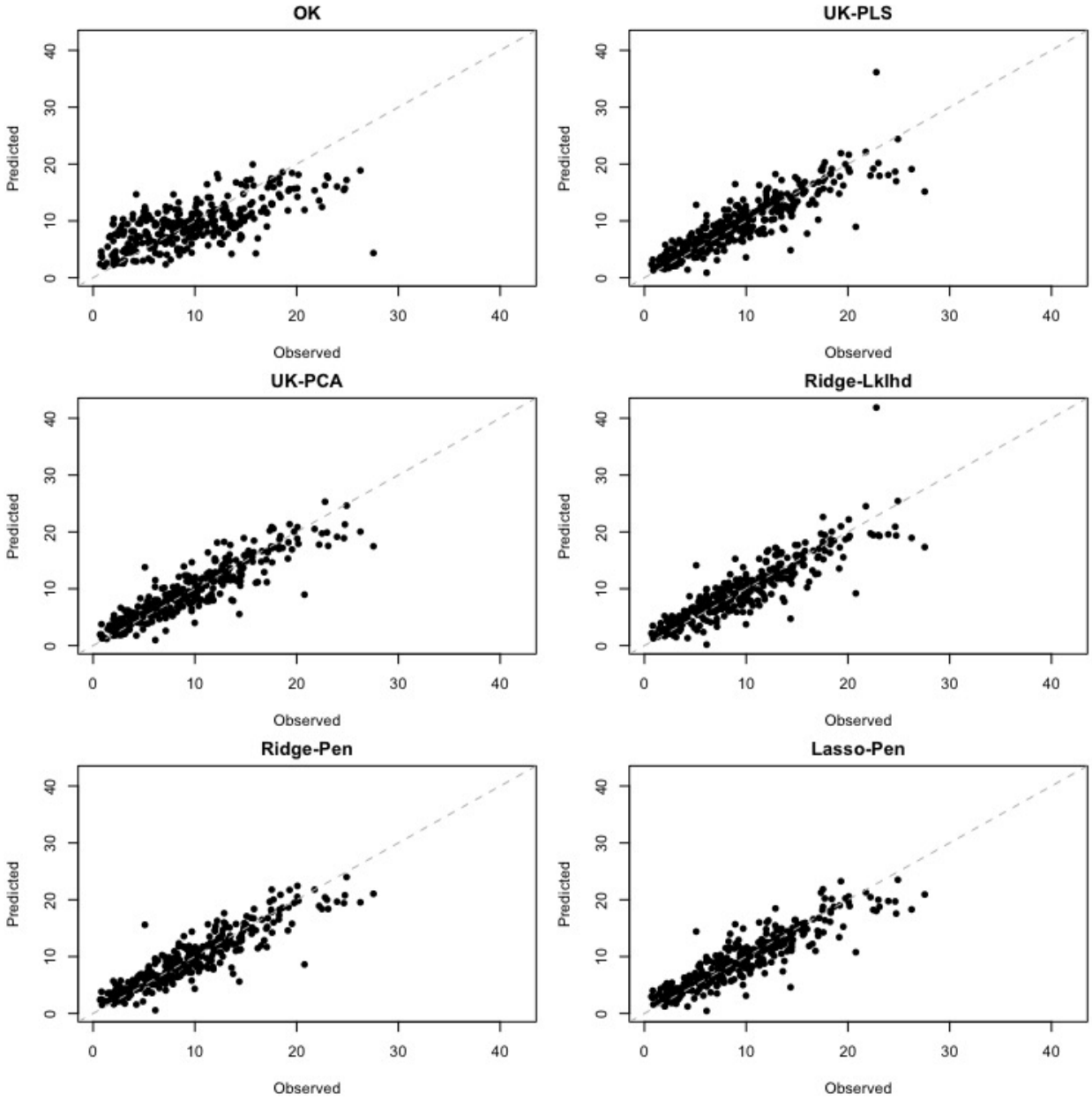

Figure 3.8: Cross-Validated Estimates against observed values for nitrogen oxides (NO_2) clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

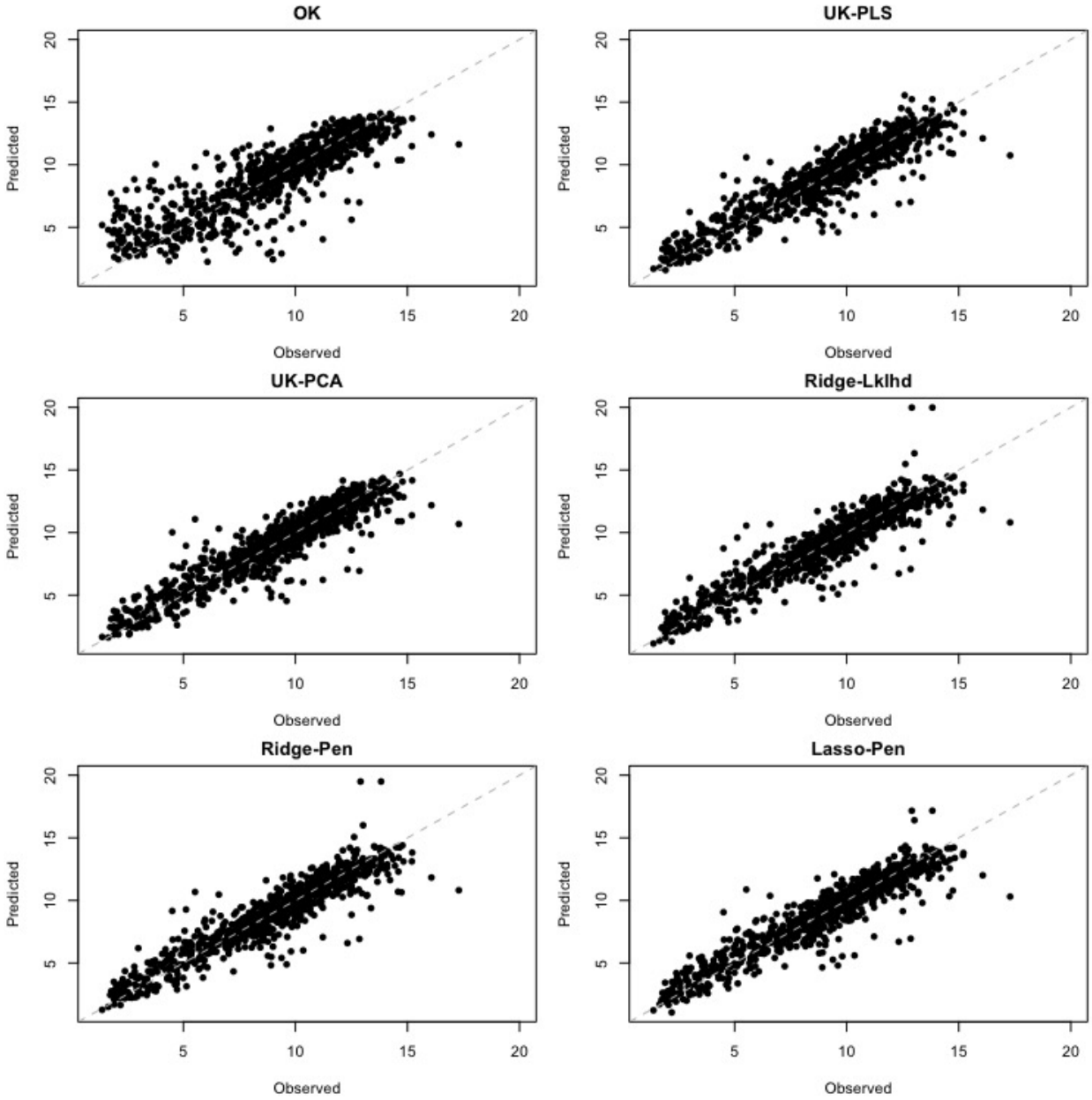

Figure 3.9: Cross-Validated Estimates against observed values for particulate matter 2.5 microns or less ($PM_{2.5}$),clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

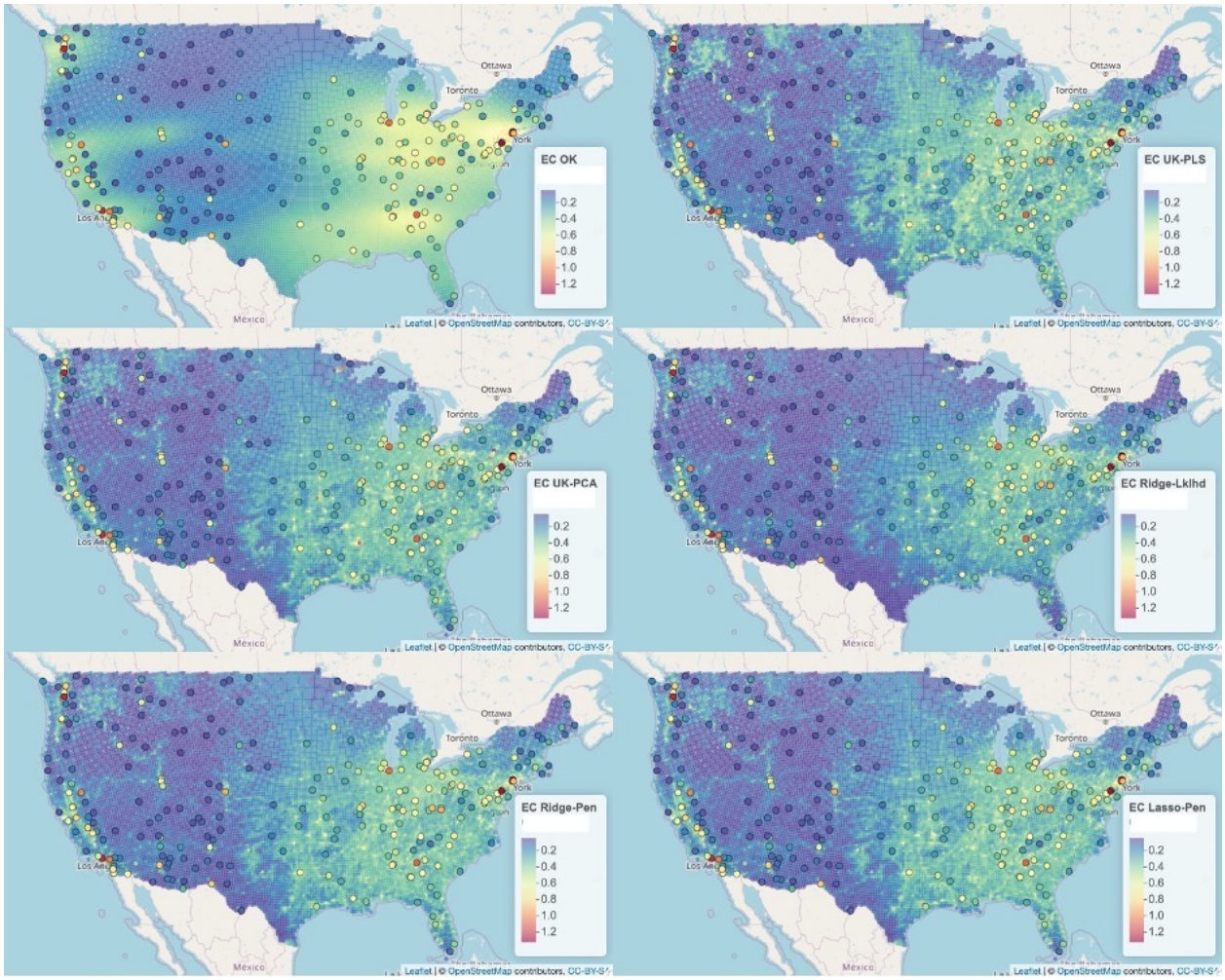

Figure 3.10: Predicted elemental carbon (EC) concentrations across the continental United States, clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

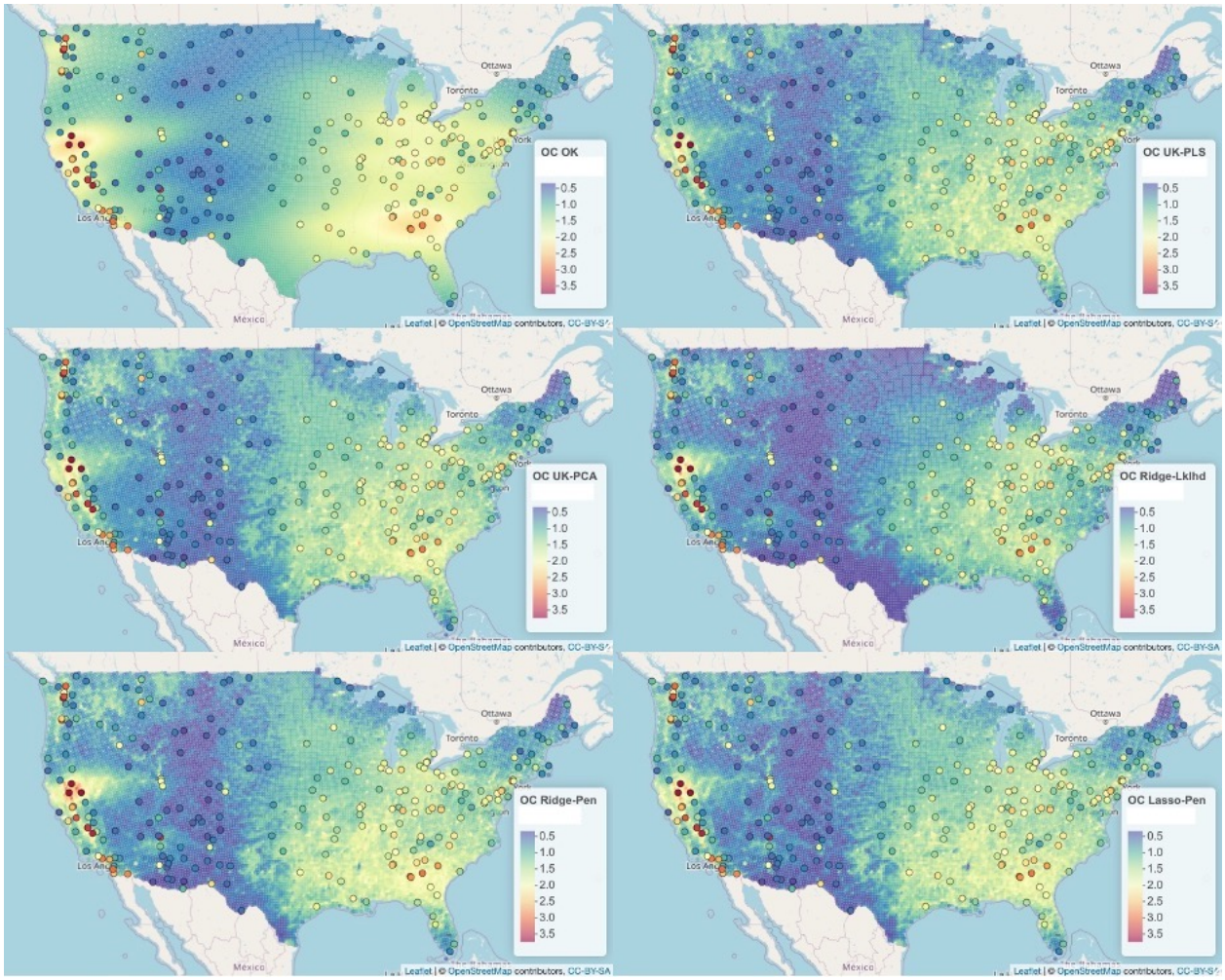

Figure 3.11: Predicted organic carbon (OC) concentrations across the continental United States, clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

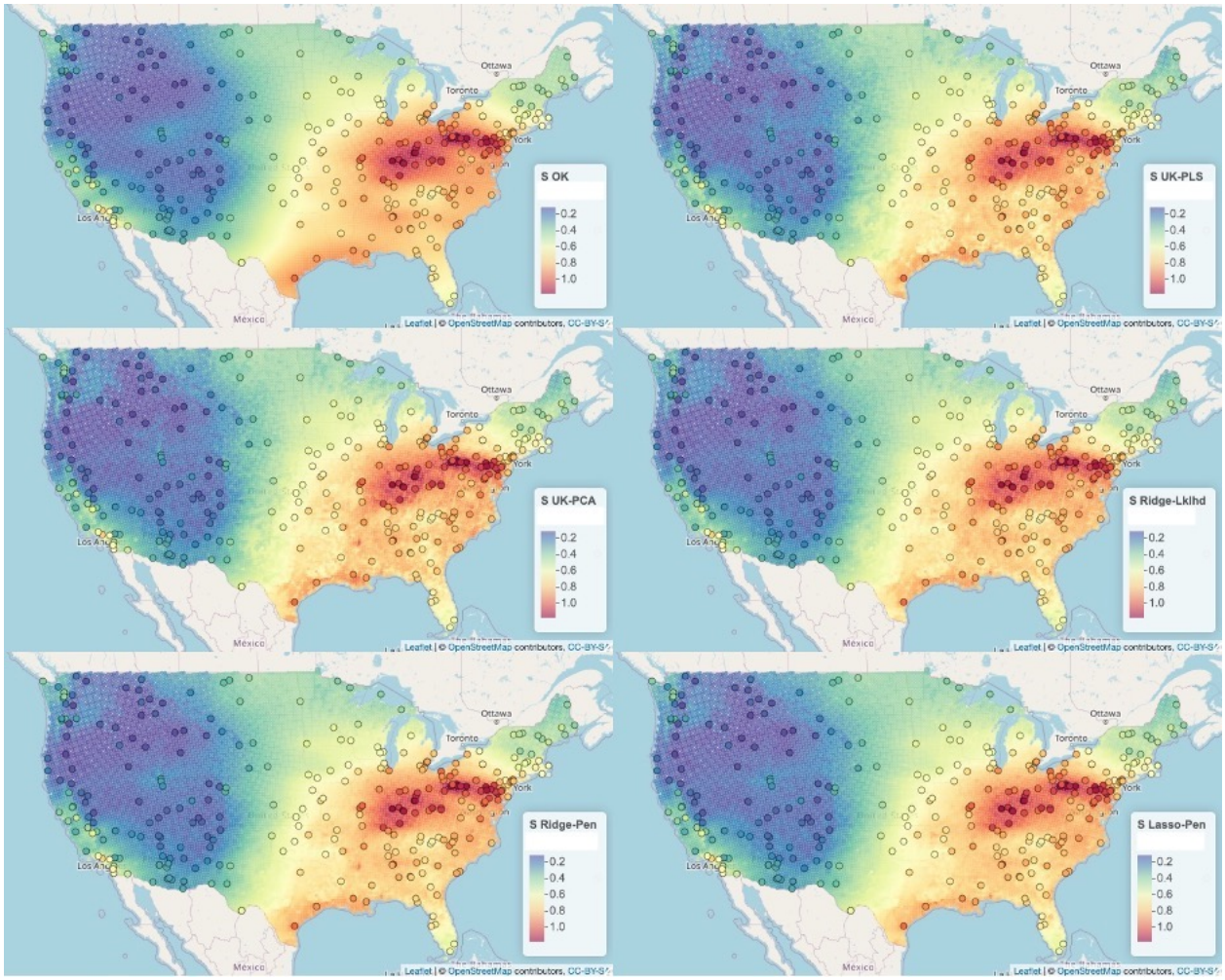

Figure 3.12: Predicted sulfur (S) concentrations across the continental United States, clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

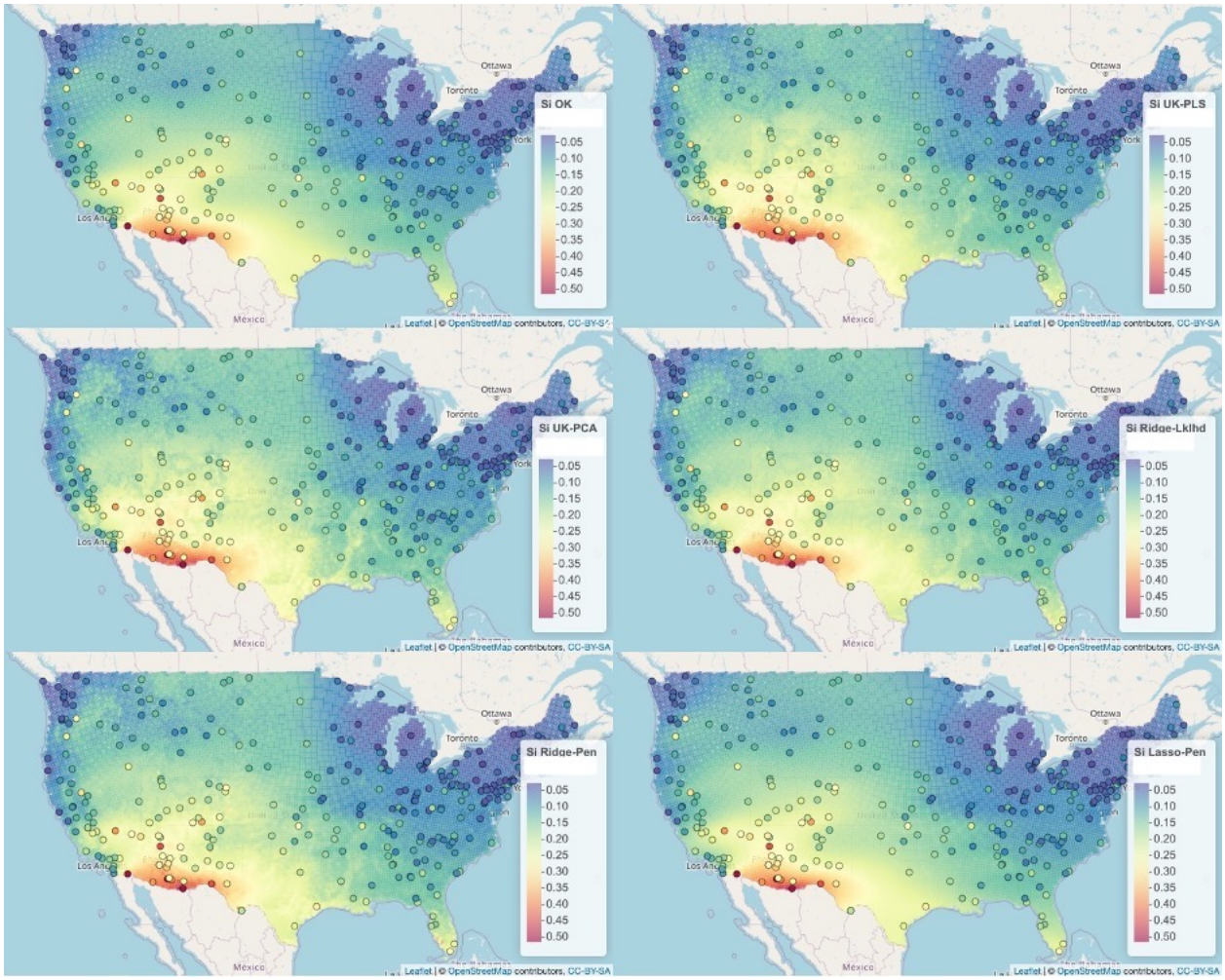

Figure 3.13: Predicted silicon (Si) concentrations across the continental United States, clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

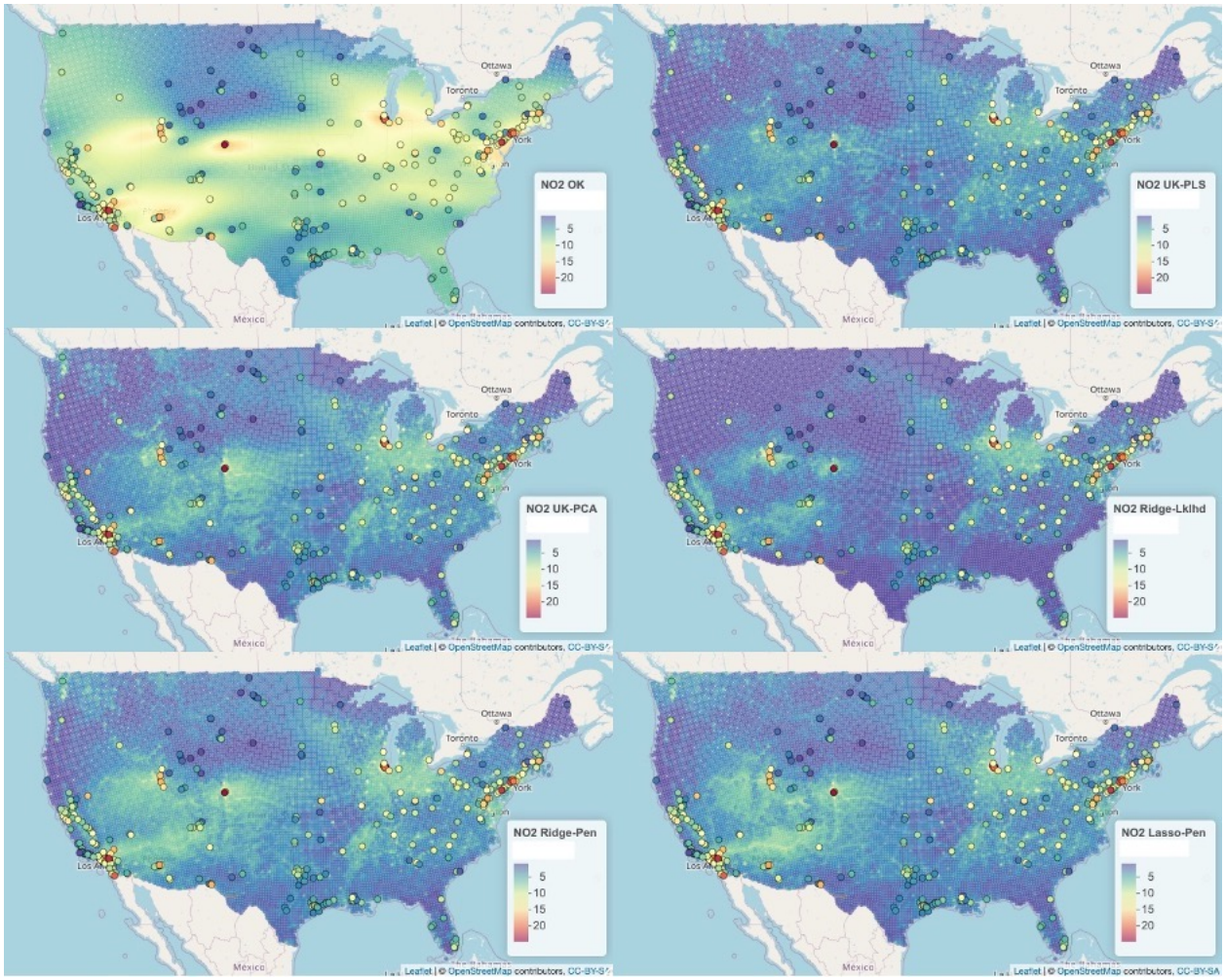

Figure 3.14: Predicted nitrogen oxides (NO_2) concentrations across the continental United States, clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

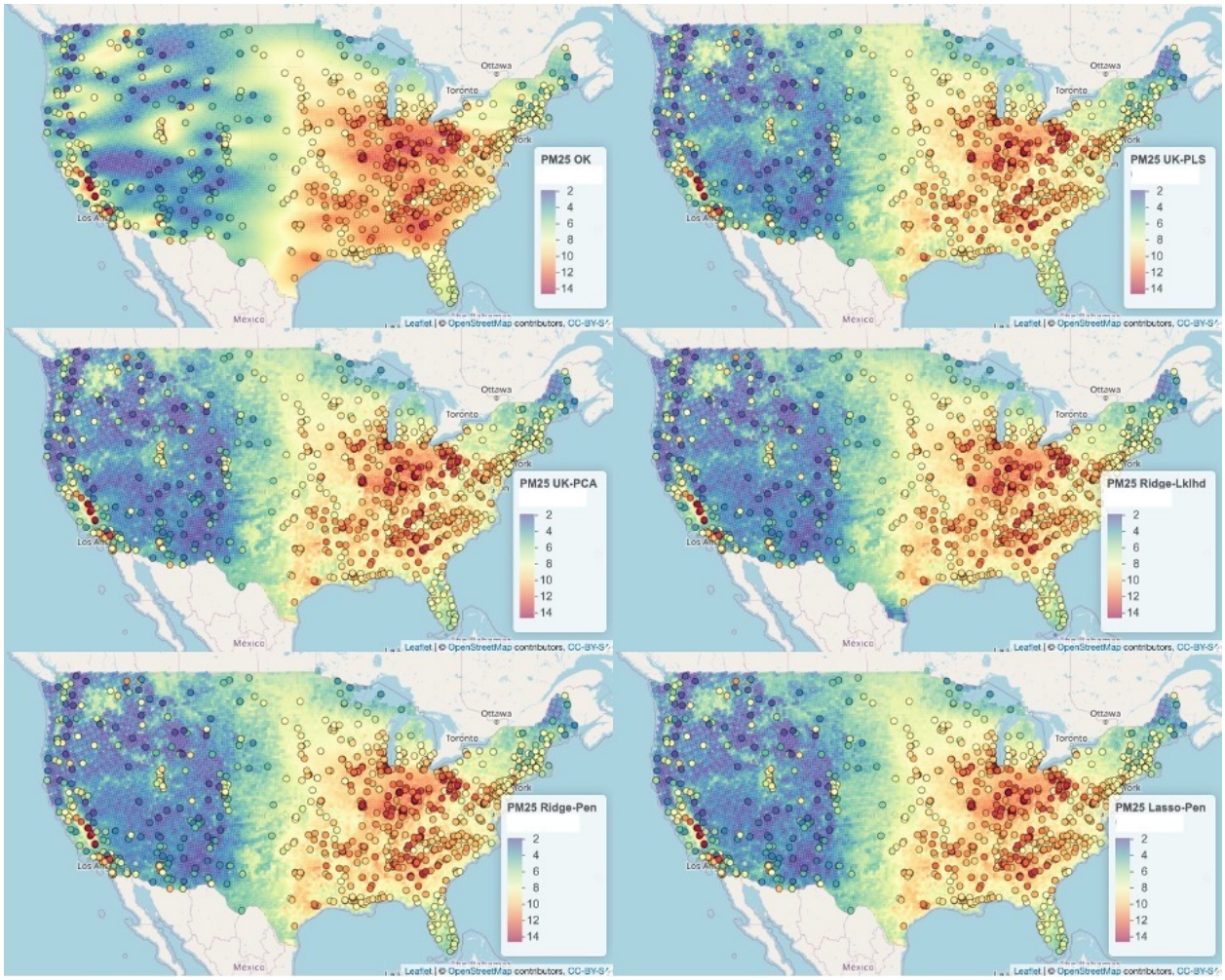

Figure 3.15: Predicted particulate matter 2.5 microns or less ($PM_{2.5}$) concentrations across the continental United States, clockwise from top left: Ordinary Kriging (OK), Universal Kriging with dimension reduction by PLS (UK-PLS), Ridge regression with parameters estimated by maximum likelihood (Ridge-Lklhd), LASSO with parameters estimated by cross-validation (Lasso-Pen), Ridge regression with parameters estimated by cross-validation (Ridge-Pen),Universal Kriging with dimension reduction by PCA (UK-PCA),

Chapter 4

MULTI-CITY DATA-ENRICHED REGRESSION TREES

4.1 Introduction

There is considerable evidence linking nitrogen oxides (NOx) with a variety of health effects, such as cardiovascular risk factors [45, 46, 43], stress hormones [25], lung function [52, 82], cognitive decline [44], and physical disability [83], and pollutant concentrations are dis-proportionally higher in disadvantaged and minority communities [24]. The Multi-Ethnic Study of Atherosclerosis (MESA) began in 2004 and is currently ongoing, following a large cohort of individuals to further examine the relationship between air pollution and cardiovascular disease. This study involved participants in six metropolitan areas across the United States: Baltimore, MD; Chicago, IL; Winston Salem, NC; Los Angeles, CA; New York, NY; and St. Paul, MN. Data sources for air pollution modeling include monitors deployed by MESA Air at stationary locations, at participants' homes, and community-based "snapshot" monitoring campaigns and is described in detail in Cohen et al[14]. The purpose of these snapshot campaigns was to model the within-city differences in order to capture the local concentration gradients.

In Chapter 2, we proposed Random Spatial Forests, a method of bagging spatially-adjusted trees, and in our examples it had superior cross-validated prediction accuracy over traditional kriging approaches. It is thus desirable to apply them in this scenario as well, but the MESA air snapshot campaign presents a few additional problems that were not previously addressed. Ensemble methods of regression trees are effective in scenarios where large numbers of covariates are used, often with the association between them and the outcome of interest unknown a priori. Combustion of fossils fuels is a known source of NO_x and it is expected that pollutant concentrations should be larger on highways and decrease

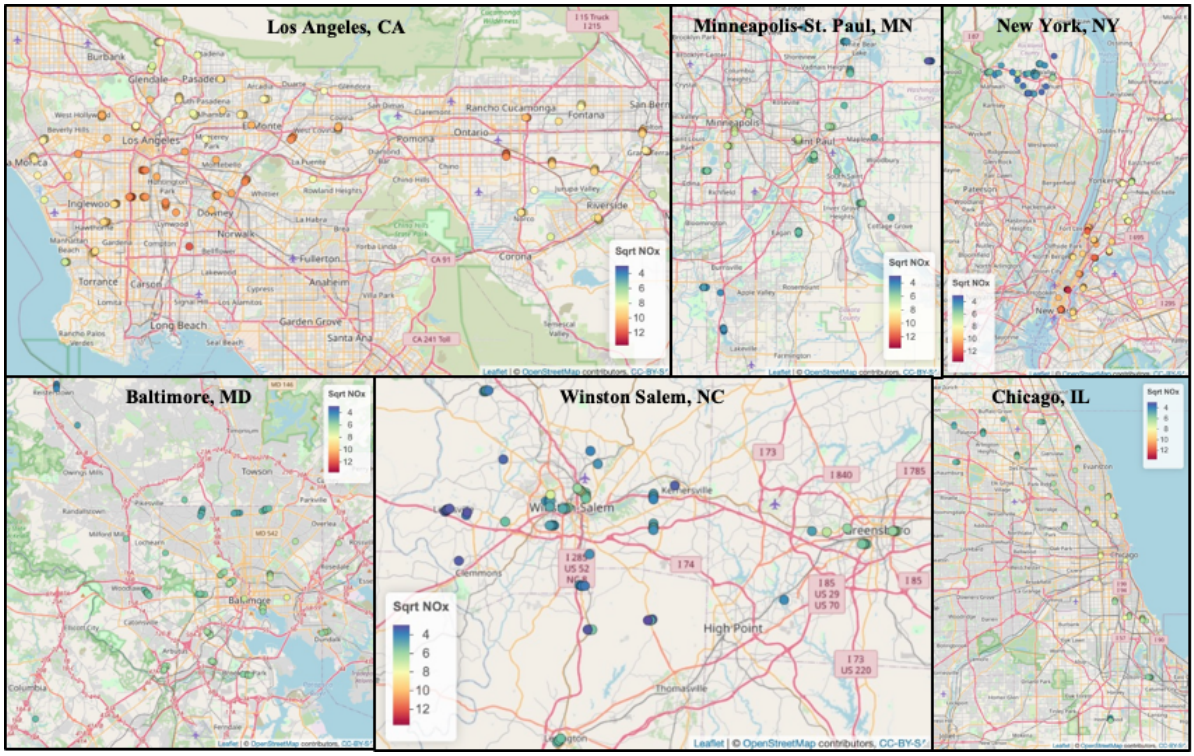

Figure 4.1: Observed NO_x levels in (clockwise from top left): Los Angeles, CA, St. Paul, MN, New York, NY, Baltimore, MD, Winston Salem, NC, and Chicago, IL

away from the road. The MESA air snapshot campaign specifically focused on evaluating these near roadway gradients as monitors were strategically placed at varying distances from roadways in order to examine how quickly pollutant concentrations disperse [56]. In this scenario, it would be more sensible to regress on distance to roadways as a fixed effect and use regression trees to model association between residual variation not attributable to roadways and the remaining geographic features.

A natural approach to use measurements from the MESA air snapshot campaign to predict air pollution exposures for study participants is to fit a single model across the six MESA cities: Los Angeles, CA Chicago, IL, Baltimore, MD, St. Paul, MN, Winston Salem, NC, and New York, NY. However, this approach does not work well in practice as overall levels and the magnitude of the association between covariates and pollutant concentrations

may vary within cities. The MESA NO_x concentrations demonstrate this problem (Figure 4.1). In the large cities, Los Angeles and New York City, NO_x concentrations are high throughout the city. Medium sized cities Baltimore, Chicago and Minneapolis-St. Paul have overall concentrations much lower than Los Angeles and New York City, but higher than the small city of Winston Salem. Constructing national models using data collected in the six cities often result in solely modeling how overall concentrations vary between cities as these are the largest sources of variation. Matern class covariance functions with a single range parameter are forced to choose between a large range, modeling overall averages across the cities but ignoring spatial gradients at finer resolutions, and a small parameter, which does the opposite.

An alternative approach is to fit a separate model in each city individually [89]. While constructing a different air pollutant model for each city results in more accurate models than a single model, it can be an unsatisfying solution. Regression trees are a flexible modeling technique that describe the association between pollutant levels and geographic features while allowing for non-linearity and interactions by partitioning observations into disjoint terminal nodes, or subgroups. The ability of regression trees to correctly identify informative subgroups depends on the sample size, and by constructing separate models in each city individually we use only a portion of observations that are available.

Collecting pollutant monitoring data is not cheap and it is difficult to deploy large networks in practice. Air pollution monitors are expensive to deploy, maintain and upkeep. Further, cohort studies such as MESA depend on estimating past exposures to air pollutants for study participants and it is now impossible to obtain additional measurements to improve accuracy of existing pollutant models. Even though existing models cannot be improved by collecting new data, pollutant concentrations were measured in multiple cities and it is natural to think that factors associated with increases/decreases in pollutant levels should be similar across cities. In this case, the most reasonable way to improve existing models is to leverage observations across cities to enhance each city-specific pollutant model individually.

In the MESA air study, we see similar patterns in pollutant concentrations within cities.

In Los Angeles, concentrations are highest downtown, but decrease outwards towards both the San Gabriel mountains and the Pacific Ocean. In midwest cities Chicago and Minneapolis-St. Paul, we see a similar trend of concentrations being highest downtown, and decreasing with increasing distance to the city center. This pattern where NO_x concentrations are highest in the city center and decrease as the city density decreases are also observed in cities along the East Coast, Baltimore and Winston Salem. Since pollutant concentrations show similar patterns across cities, one way to improve tree-based estimates is to combine observations across cities in order to determine splits in the regression tree. This can be thought of as a form of variable selection where the variables being selected are the split rules in the regression tree.

Another method to improve estimates by combining models across cities is through data-enriched regression [10] where they explored a hybrid of pooling and fitting separate models to datasets with similar but not identical statistical characteristics. We may expect, for example, that highways are associated with larger increases in NO_x in Los Angeles where there are greater numbers of cars compared to St. Paul or Winston Salem. Although we would expect the magnitude of the association to be larger in Los Angles than in St. Paul or Winston Salem, we would still expect highways to be associated with increased concentrations of NO_x in St. Paul and Winston Salem. Thus, it may be desirable to encourage estimates to be similar, but not identical, to each other across cities.

In Section 4.2, we propose a flexible additive regression tree framework for the MESA air snapshot campaign which allows for fixed effects, multiple city-specific regression trees, city-specific spatial processes, and data enriched city-specific tree estimates using ℓ_2 shrinkage. By extending the tree building algorithm developed by Chapter 2, we show that these multi-city data- enriched regression trees are computationally feasible and illustrate how these trees can be used in ensemble models through a random forests inspired bootstrap aggregation approach. In Section 4.3, we provide simulation results demonstrating that combining information across cities improves prediction accuracy over separate city-specific models when the association between features and pollutant concentrations are similar between cities. In

Section 4.4, we apply our method to average winter NO_x measurements across the six different MESA cities in 2006-2007 and demonstrate that combining information across cities leads to more accurate models in each city individually.

4.2 Methods

4.2.1 Adding Fixed Effects to Spatially Adjusted Regression Trees

The MESA air snapshot campaign focused on understanding near roadway gradients by sampling roadways at distances of 0-50m, 10m-100m, and 100m-350m. This sampling scheme of the snapshot campaign was designed to assess how quickly pollutant levels drop as distance from roads increases. The use of the regression trees in these settings is to identify subgroups, which are unknown a priori, among geographic covariates where pollutant levels differ. Since roads are known to be associated with NO_x it would be preferrable to regress explicitly on roads as a fixed effect rather than including roads as a covariate in the regression tree.

Noting the additive model formulation of spatially adjusted regression trees, fixed effects based on covariates, $\mathbf{Z} \in \mathbb{R}^{N \times p}$, can be included in Eq. 2.5 as

$$\mathbf{Y}(\mathbf{s}) = \mathbf{Z}\boldsymbol{\beta} + \mathbf{t}(\mathbf{X}) + \boldsymbol{\nu}(\mathbf{s}). \tag{4.1}$$

4.2.2 Multi-City Regression Trees

We describe multi-city regression trees, a method which constructs a city-specific regression trees for each city individually, but incorporates observations across all cities in order to determine a shared tree structure describing the relationship between geographic factors and pollutant levels across cities. The shared tree structure is analogous to variable selection as each of the terminal nodes are categorical variables defined by geographic features. Instead of picking variables or categories which seem to be associated with pollutant concentrations in each city individually, multi-city regression trees determine which categories are the most associated with pollutant concentrations across *all* cities.

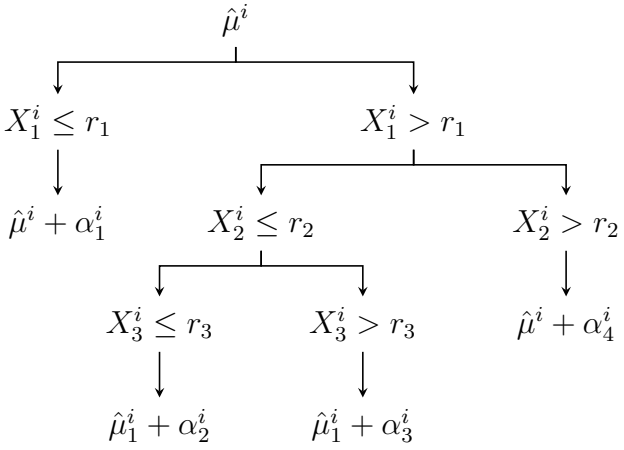

Figure 4.2: An example of a single tree structure which can be used for each of the i cities

To illustrate, a multi-city regression tree is shown in Figure 4.2. For each city-specific tree, all terminal nodes are defined by splitting on the same set of covariates $X_j, j = 1, 2, 3$ and the same set of splitting values $r_j, j = 1, 2, 3$. In the initialization of the tree, α_i^0 is the initial intercept for each city i. This parameter is estimated for each city individually, and is designed to model between-city pollutant concentration differences so that the subsequent splits are chosen to model within-city contrasts. While the structure of the tree is the same in each city, each city-specific tree has its own set of estimates, $\boldsymbol{\alpha}^i$, which allow for the size of the effect to differ across cities. For example, if the covariate were distance to highways we may think that the increase in air pollution is larger in big cities with large interstates such as Los Angeles compared to smaller cities like Winston Salem.

We define a multi-city regression tree with k terminal nodes in matrix form for d cities. First, let $\mathbf{Y}(\mathbf{s})$, the stacked vector of observations, ordered by city as:

$$\mathbf{Y}(\mathbf{s}) = \left[\mathbf{Y}^T(\mathbf{s}_1), \ldots, \mathbf{Y}^T(\mathbf{s}_d)\right]^T.$$

The tree design matrix, $\mathbf{C}^k$ is then constructed as:

$$\mathbf{c}_i^j = \begin{cases} 1, & \mathbf{X}_j^i(\mathbf{s}_i) \leq r_j \text{ and is in terminal node being split} \\ 0 & \text{else} \end{cases}$$

$$\mathbf{C}^j = \begin{bmatrix} \mathbf{c_1^j} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{c_2^j} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & & \mathbf{c_d^j} \end{bmatrix} \in \{0,1\}^{N \times d}$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}^0 & \mathbf{C}^1 & \mathbf{C}^2 & \dots & \mathbf{C}^k \end{bmatrix} \in \{0,1\}^{N \times d(k+1)}.,$$

where $\mathbf{c_j^i}$ is an indicator vector denoting observations in the terminal node created by j^{th} split for city i. $\mathbf{C}$ is the stacked tree design matrix with k terminal nodes for d cities, where each block $\mathbf{C}^j$, $j \geq 1$ corresponds to the split matrix for the j^{th} terminal node with a separate column for each city. In order to initialize the tree, for $\mathbf{C}^0$ each city indicator vector $\mathbf{c}_i^j = \mathbf{1}_{n_i}$, which is equivalent to fitting city-specific intercepts.

Define the associated contrast vector, $\boldsymbol{\alpha}$, as

$$\boldsymbol{\alpha}^j = \begin{bmatrix} \alpha_1^j & \alpha_2^j & \dots & \alpha_d^j \end{bmatrix} \in \mathbb{R}^{d \times 1}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} (\boldsymbol{\alpha^0})^T & (\boldsymbol{\alpha}^1)^T & \dots & \left(\boldsymbol{\alpha}^k\right)^T \end{bmatrix}^T \in \mathbb{R}^{d(k+1) \times 1},$$

where $\boldsymbol{\mu}^i$ is a d-vector of overall levels in each city, and is the parameters corresponding to $\mathbf{C}^0$, the city-specific intercepts. For each split matrix $\mathbf{C}^j, , j \geq 1$, $\boldsymbol{\alpha}^j$ is the associated vector of city-specific contrasts for the terminal node created by split j.

To allow for fixed effects, we define

$$\mathbf{Z}^l = \begin{bmatrix} \mathbf{Z_1^l} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2^l} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & & \mathbf{Z_d^j} \end{bmatrix} \in \{0,1\}^{N \times d}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}^1 & \mathbf{Z}^2 & \dots & \mathbf{Z}^p \end{bmatrix} \in \mathbb{R}^{N \times dp}$$

and the associated coefficients β as:

$$\boldsymbol{\beta}^l = \begin{bmatrix} \beta_1^l & \beta_2^l & \dots & \beta_d^l \end{bmatrix} \in \mathbb{R}^{d\times 1}$$
$$\boldsymbol{\beta} = \begin{bmatrix} (\boldsymbol{\beta^0})^T & (\boldsymbol{\beta}^1)^T & \dots & (\boldsymbol{\beta}^p)^T \end{bmatrix}^T \in \mathbb{R}^{dp\times 1}.$$

In order to allow for a city-specific spatially correlated residual process, we include $\boldsymbol{\Sigma}^i$ for city i, where each city is allowed to have its own set of correlation parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_d$:

$$\boldsymbol{\Sigma}(\mathbf{s}, \boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Sigma}(\mathbf{s_1}; \boldsymbol{\theta_1}) & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Sigma}(\mathbf{s_2}; \boldsymbol{\theta_2}) & 0 & \dots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & & \\ 0 & 0 & \dots & & \boldsymbol{\Sigma}(\mathbf{s_d}; \boldsymbol{\theta_d}) \end{bmatrix} \in \mathbb{R}^{N\times N}$$

Let the spatially correlated residual process be a realization of a Gaussian process, thus

$$\mathbf{Y}(\mathbf{s}) \sim N\left(\mathbf{Z}\boldsymbol{\beta} + \mathbf{C}^k\boldsymbol{\pi}^k, \boldsymbol{\Sigma}(\mathbf{s}, \boldsymbol{\theta})\right).$$

The set of city-specific trees are constructed by maximum likelihood. Assuming the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$ are known, maximizing the likelihood of the observed process is equivalent to minimizing

$$\ell(\mathbf{C}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{Y}(\mathbf{s}) - \mathbf{Z}\boldsymbol{\beta} - \mathbf{C}\boldsymbol{\alpha})^T \boldsymbol{\Sigma}^{-1}(\mathbf{s}, \boldsymbol{\theta}) (\mathbf{Y}(\mathbf{s}) - \mathbf{Z}\boldsymbol{\beta} - \mathbf{C}\boldsymbol{\alpha}). \tag{4.2}$$

4.2.3 Multi-City Data-Enriched Regression Trees

Multi-city regression trees make use of observations across trees to determine informative subgroups from large sets of geographic covariates, and here we address different approaches to constructing estimates for each of these subgroups across the MESA cities. There are two simple approaches to estimating the contrasts. First, let each city have its own set of city-specific contrasts using observations only in that city. This would be ideal if the contrasts associated with each subgroup are very different in each city individually. Alternatively, if the

contrasts in each city are identical the ideal approach would be to take the average contrast across all cities, ignoring city membership. In practice, the optimum approach likely lies somewhere between these two extremes. We would expect the association between geographic features and pollutant concentrations to be similar across cities, but not necessarily the same as all cities. The problem of determining how much weight to put on estimates from separate datasets is addressed in [10], which they describe as data-enriched regression. Given multiple datasets modeling the association between features and an outcome, data-enriched regression uses a penalty to adaptively combine parameter estimates between datasets.

Following Chen et al.[10], we propose a penalized method in order to induce similarity among the contrasts vectors between cities. Define the penalty matrix for contrasts pertaining to the j^{th} split as:

$$\mathbf{L}^j = d\mathbf{I}_d - \vec{\mathbf{1}}_d^T \vec{\mathbf{1}}_d, j = 1, 2, \ldots,$$

and define the full penalty matrix as

$$\mathbf{L} = \begin{bmatrix} \mathbf{L^0} & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{L}^1 & 0 & \ldots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & & \\ 0 & 0 & \ldots & & \mathbf{L}^k \end{bmatrix} \in \mathbb{R}^{d(k+1)\times d(k+1)}$$

and add the penalty matrix to Eq. 4.2 as:

$$\left(\mathbf{Y}(\mathbf{s}) - \mathbf{Z}\boldsymbol{\beta} - \mathbf{C}\boldsymbol{\alpha}\right)^T \boldsymbol{\Sigma}^{-1}(\mathbf{s}) \left(\mathbf{Y}(\mathbf{s}) - \mathbf{Z}\boldsymbol{\beta} - \mathbf{C}\boldsymbol{\alpha}\right) + \lambda \boldsymbol{\alpha}^T \mathbf{L} \boldsymbol{\alpha}. \tag{4.3}$$

This type of penalty is similar to *graph constrained estimation* [47], and a variety of penalties have been proposed to induce similarities between "connected" groups [48, 65]. For our application, we set $\mathbf{L^0} = \mathbf{0}^{d\times d}$ to allow each city to start with its own mean. The form of $\mathbf{L}^j$ encourages similarity between predictions in different cities for the same split, and is equivalent to penalizing the deviations to the average effect across cities.

$$\lambda \boldsymbol{\alpha}^T \mathbf{L} \boldsymbol{\alpha} = \lambda \sum_{i=1}^{k} \sum_{j=1}^{d} \sum_{k \neq j} \left\| \alpha_j^i - \alpha_k^i \right\|_2^2 = d\lambda \sum_{i=1}^{k} \sum_{j=1}^{d} \left\| \alpha_j^i - \bar{\alpha}^i \right\|_2^2$$

Our penalized formulation of multi-city regression trees, which we call *multi-city data-enriched regression trees*, is analagous to a linear mixed model. The un-penalized initial estimate in each tree is equivalent to city-specific intercepts, while the penalized contrasts in each terminal node are random slopes, where the variance of the random slope is inversely proportional to the tuning parameter λ.

The tuning parameter λ has the same role of the variance parameter in a linear mixed model and controls how much information we share between cities. As $\lambda \to 0$ (infinite variance of the random slopes), we ignore estimates in other cities and estimate the contrasts in each city individually. As $\lambda \to \infty$ (random slopes with zero variance), we simply take the average of the contrast estimates across cities.

4.2.4 Optimization of Multi-City Data-Enriched Regression Trees

Optimization of multi-city spatially data-enriched regression trees follows a similar approach taken in Algorithm 1. The addition of fixed effects can be incorporated into the spatially-adjusted tree building algorithm by appending the fixed effects to the initialization of the tree design matrix as

$$\begin{aligned}\tilde{\mathbf{C}}_0 &= \begin{bmatrix}\mathbf{Z} & \mathbf{C}_0\end{bmatrix} \\ \tilde{\boldsymbol{\alpha}}^0 &= \begin{bmatrix}\boldsymbol{\beta}^T & (\boldsymbol{\alpha}^0)^T\end{bmatrix}^T.\end{aligned}$$

In addition, any penalty on the fixed effects, $\boldsymbol{L}_{\mathbf{Z}}$, can be included by appending the initial penalty matrix as

$$\tilde{\mathbf{L}}^0 = \begin{bmatrix}\mathbf{L}^{\mathbf{Z}} & \mathbf{L}^{\mathbf{Z}0} \\ \mathbf{L}^{0\mathbf{Z}} & \boldsymbol{L}^0,\end{bmatrix}$$

where $\mathbf{L}^{\mathbf{Z}}$ is a penalty on the fixed effects, $\boldsymbol{L}^0$ the penalty on the city-specific intercepts, and $\mathbf{L}^{0\mathbf{Z}}$ and $\mathbf{L}^{\mathbf{Z}0}$ optional penalties between the fixed effects and the city-specific intercepts.

We further define

$$\tilde{\mathbf{C}} = \begin{bmatrix} \tilde{\mathbf{C}}^0 & \mathbf{C}^1 & \mathbf{C}^2 & \dots & \mathbf{C}^k \end{bmatrix}$$
$$\tilde{\boldsymbol{\alpha}} = \begin{bmatrix} (\tilde{\boldsymbol{\alpha}}^0)^T & (\boldsymbol{\alpha}^1)^T & \dots & \left(\boldsymbol{\alpha}^k\right)^T \end{bmatrix}$$
$$\tilde{\mathbf{L}} = \begin{bmatrix} \tilde{\mathbf{L}}^0 & 0 & 0 & \dots & 0 \\ 0 & \mathbf{L}^1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & & \\ 0 & 0 & \dots & & \mathbf{L}^k \end{bmatrix}$$

where $\tilde{\mathbf{C}}$, $\tilde{\boldsymbol{\alpha}}$, and $\tilde{\mathbf{L}}$ are the tree design matrix, contrasts, and penalties with the fixed effects included. The penalized form of the loss function (Eq. 4.3) becomes

$$\left(\mathbf{Y}(\mathbf{s}) - \tilde{\mathbf{C}}\tilde{\boldsymbol{\alpha}}\right)^T \boldsymbol{\Sigma}^{-1}(\mathbf{s}, \boldsymbol{\theta}) \left(\mathbf{Y}(\mathbf{s}) - \tilde{\mathbf{C}}\tilde{\boldsymbol{\alpha}}\right) + \lambda \tilde{\boldsymbol{\alpha}}^T \tilde{\mathbf{L}} \tilde{\boldsymbol{\alpha}},$$

and the closed form optimal estimate for $\tilde{\boldsymbol{\alpha}}$ is:

$$\hat{\boldsymbol{\alpha}} = \left(\tilde{\mathbf{C}}^T \boldsymbol{\Sigma}^{-1}(\mathbf{s}, \boldsymbol{\theta}) \tilde{\mathbf{C}} + \tilde{\mathbf{L}}\right)^{-1} \tilde{\mathbf{C}}^T \boldsymbol{\Sigma}^{-1}(\mathbf{s}, \boldsymbol{\theta}) \mathbf{Y}(s).$$

Since a closed-form solution exists for the contrasts, the loss function is characterized by the matrix $\boldsymbol{\Omega}$ as

$$\ell(\tilde{\mathbf{C}}) = \mathbf{Y}(\mathbf{s})^T \boldsymbol{\Omega} \mathbf{Y}(\mathbf{s})$$
$$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{C}} \left(\tilde{\mathbf{C}}^T \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{C}} + \tilde{\mathbf{L}}\right)^{-1} \tilde{\mathbf{C}}^T \boldsymbol{\Sigma}^{-1}.$$

This formulation of the loss through the matrix $\boldsymbol{\Omega}$ leads to a computationally desirable property; given the current state of the tree $\mathbf{C}$ and adding a new split by appending a new contrast matrix $\mathbf{C}^A$ as

$$\begin{bmatrix} \mathbf{C} & \mathbf{C}^A \end{bmatrix},$$

computing the new $\boldsymbol{\Omega}^{k+1}$ relies only on the previous one, $\boldsymbol{\Omega}^k$, the new contrast matrix $\mathbf{C}^A$ and (if desired) the penalty on the new contrast matrix $\mathbf{L}^{k+1}$ by

$$\boldsymbol{\Omega}^{k+1} = \boldsymbol{\Omega}^k - \boldsymbol{\Omega}^k \mathbf{C}^A \left(\left(\mathbf{C}^A\right)^T \boldsymbol{\Omega}^k \mathbf{C}^A + \mathbf{L}^{k+1}\right)^{-1} \left(\mathbf{C}^A\right)^T \boldsymbol{\Omega}^k.$$

By this approach, the worst case run time for the multi-city data-enriched regression trees is $\mathcal{O}\left(pn\log(n)(n+d^3)\right)$ and can be run for reasonably sized datasets. The multi-city data-enriched regression tree building procedure is summarized in Algorithm 3.

Algorithm 3 Multi-City Data-Enriched Regression Tree Building Algorithm

1. Start with $\mathbf{C} = \tilde{\mathbf{C}}^0$.
2. Given $\boldsymbol{\Sigma}$, set the initial value for $\boldsymbol{\Omega}^0 = \boldsymbol{\Sigma}^{-1} - \tilde{\mathbf{C}}^0 \left(\left(\tilde{\mathbf{C}}^0 \right)^T \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{C}}^0 + \tilde{\mathbf{L}}^0 \right)^{-1} \left(\tilde{\mathbf{C}}^0 \right)^T \boldsymbol{\Sigma}^{-1}$.
3. For $t = 1, 2, \ldots$
   (a) Check each of the k existing terminal nodes for a new terminal node created by a decision rule based on the selected covariates $\mathbf{X}_r$, to create a candidate set of possible splits.
   (b) Find the set of candidate split $\mathbf{C}^A$ which maximizes the change in loss
   $$\mathbf{Y}(\mathbf{s})^T \left(\boldsymbol{\Omega}^k \mathbf{C}^A \left(\left(\mathbf{C}^A \right)^T \boldsymbol{\Omega}^k \mathbf{C}^A + \mathbf{L}^{k+1} \right)^{-1} \left(\mathbf{C}^A \right)^T \boldsymbol{\Omega}^k \right) \mathbf{Y}(\mathbf{s})$$
   (c) Update
   $$\boldsymbol{\Omega}^{k+1} = \boldsymbol{\Omega}^k - \boldsymbol{\Omega}^k \mathbf{C}^A \left(\left(\mathbf{C}^A \right)^T \boldsymbol{\Omega}^k \mathbf{C}^A + \mathbf{L}^{k+1} \right)^{-1} \left(\mathbf{C}^A \right)^T \boldsymbol{\Omega}^k.$$
4. Repeat step 3 until the maximum number of splits are exceeded or there are no more branches can be split without creating a branch with less than m observations.

4.2.5 Random Forests of Multi-City Data-Enriched Regression Trees

Tree-based methods alone are not often used for prediction as they suffer from high variance. Instead, ensemble method of trees are often employed, the most prevalent of which random forests [9] and boosting [12]. Either of these method could be employed using multi-city

data-enriched regression trees in place of normal regression trees for the MESA air snapshot campaign. Here, we propose a modified version of random forests specifically designed to take advantage of the clustered sampling design in MESA air.

Random Forests is a type of bootstrap aggregation (bagging), an ensemble method of averaging over trees constructed on bootstrapped samples. The difference between random forests and bagging is that random forests adds another layer of randomization, where only a random subset of the covariates is considered for each new terminal node for each tree, in order to minimize correlation between trees.

In the MESA Air snapshot campaign, monitors were located in clusters of six near major roadways, with three on both sides located approximately 50, 100, and 300 m from the road. Bootstrapping each monitor individually can be problematic in this setting. Our multi-city data-enriched regression trees are additive models consisting of both a regression tree and a spatial process. If monitors are sampled individually, it is highly unlikely that a whole cluster of mointors will be held out and each cluster will contain at least one monitor. In this case, the optimal estimates attributes all the variation to the spatial smoother, and estiamtes held out monitors by simply smoothing from nearby monitors in the same cluster, and we do not learn about geographic features that are associated with higher or lower concentrations.

Our goal in the MESA air snapshot campaign is to predict pollutant levels at participants place of residence at sites not near monitoring locations. Rather than re-sampling with individual monitors as the independent units, we propose a modified version of random forests where re-sampling of monitors is done with each cluster of near roadway monitors as the independent units instead of individual observations.

4.2.6 Estimation of Unknown Parameters

Algorithm 1 allows us to build spatially-adaptive data enriched regression trees given $\boldsymbol{\theta}, \lambda$. However, these parameters are unknown a priori and must be estimated. Optimization of these tuning parameters can be difficult since no gradient based approaches can be applied. In Chapter 2, we employed a grid search, but the computational cost of optimizing large

numbers of parameters concurrently by grid search is restrictively large.

Rather than estimate all the parameters concurrently, we take a two-step optimization approach. Recall that the design of the spatial covariance matrix $\mathbf{\Sigma}$ is block diagonal, and each city-specific spatial process, $\nu_i(s_i) \sim (0, \mathbf{\Sigma}(s_i, \boldsymbol{\theta}_i))$, $i = 1, \ldots, d$, is a statistical characterization of spatially smooth residual variance not attributable to the fixed effects or regression tree in that specific city. The spatial process being independent, city-specific estimates is taken because of the well known screening effect [75], where conditioning on nearby observations reduces the influence of distant ones. In a study such as the MESA air snapshot campaign, the contribution from spatial smoothing of observations in different cities is likely to be very small, and treating them as being independent instead of including the miniscule correlation between these points is fundamentally similar to covariance tapering [41]. Rather than estimating all covariance parameters jointly, we propose fitting individual city-specific models to estimate each city-specific set of spatial covariance parameters $\hat{\boldsymbol{\theta}}_i, i = 1, \ldots, d$. As each individual city-specific model does not include any observations from other cities, λ does not factor into any of these individual models. The spatial covariance parameters may be estimated either by maximum likelihood or cross-validation, and in Chapter 2 we found cross-validated estimates to have better prediction accuracy than maximum likelihood, and is the approach taken here.

Multi-city data-enriched regression tree models take individual city-specific estimates and use the tuning parameter λ to determine how similar estimates between cities should be. Our two-step optimization approach uses the plug-in estimates as known values for the covariance parameters $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_d$, and then selects the tuning parameter λ by cross-validation using all the data across MESA cities.

4.2.7 Standardization across Multiple Cities

A common issue when observations are made in multiple domains is choosing a method of standardization. Our goal is to develop models that borrow information across cities to build representative, highly accurate models in each city individually, and we do not wish to

have the structure of our multi-city trees be dominated by cities with the most number of observations or largest variance.

Many methods for standardization have been proposed, such as dividing by the number of observations to make the contribution to the total loss be "equally-sized", or to standardize observations in each specific city to have unit variance. There is no consensus on a "best" method, and the choice of standardization is arbitrary. While both of these approaches are reasonable, we take a third approach

Our approach to standardization is motivated by the relationship between penalized regression and linear mixed models in Section 2.2.4. First, we note that due to the block structure of the $\mathbf{\Sigma}$, we can re-write the loss (Eq. 4.3) as

$$\sum_{j=1}^{d} \left[\left(\mathbf{Y}(\mathbf{s}_j) - \mathbf{Z}_j\boldsymbol{\beta}_j - \mathbf{C}_j\boldsymbol{\alpha}_j \right)^T \mathbf{\Sigma}_j^{-1}(\boldsymbol{\theta}) \left(\mathbf{Y}(\mathbf{s}_j) - \mathbf{Z}_j\boldsymbol{\beta}_j - \mathbf{C}_j\boldsymbol{\alpha}_j \right) \right] + \lambda \boldsymbol{\alpha}^T \mathbf{L} \boldsymbol{\alpha}.$$

In Section 2.2.4, we demonstrated that for any valid semi-definite spatial correlation matrix can be expressed we can re-write the spatial covariance $\mathbf{\Sigma}_j = \sigma_j^2 \mathbf{S}(\mathbf{s}_j)\mathbf{S}^T(\mathbf{s}_j) + \tau_j^2 \mathbf{I}_n$, and in spatial mixed model form, this is equivalent to

$$\sum_{j=1}^{d} \frac{1}{\tau_j^2} \left[\left\| \mathbf{Y}(\mathbf{s}_j) - \mathbf{Z}_j\boldsymbol{\beta}_j - \mathbf{C}_j\boldsymbol{\alpha}_j - \mathbf{S}(\mathbf{s}_j)\boldsymbol{\eta}_j \right\|_2^2 + \frac{\tau_j^2}{\sigma_j^2} \left\| \boldsymbol{\eta} \right\|_2^2 \right] + \lambda \boldsymbol{\alpha}^T \mathbf{L} \boldsymbol{\alpha}.$$

The loss in mixed model form standardizes observations by τ_j^2, the size of the nugget, or unexplained variance, in each city. Our approach uses the empirical estimates of τ_j^2 from fitting the individual city models first, and uses each τ_j^2 as a plug-in estimate for the standardization of each city in multi-city regression trees.

4.3 Simulation Studies

We conduct a simulation study to demonstrate the benefits of combining information in regression trees across cities. We simulate pollutant levels using three covariates across three different cities as

$$Y^i(s) = \beta_0^i + x_1^i\beta_1^i + x_2^i\beta_2^i + x_3^i\beta_3^i + \nu^i(s), \nu^i(s) \sim \left(0, \mathbf{\Sigma}^i(\delta^i)\right).$$

We generate the city-specific intercepts, $\beta_0^i \overset{i.i.d}{\sim} N(100, 50)$, to model cities with widely varying overall levels. Next, we create within city contrasts by using three covariates, x_1^i, x_2^i, and x_3^i to represent geographic features associated with changes in pollutant concentrations. The size of the effect for each of the three geographic features is allowed to vary as $\beta_i \sim (b^i, \sigma_\beta^2)$, $i = 1, 2, 3$. The purpose of letting this effect vary is to demonstrate cities where the effect may be larger or smaller, and the variance parameter σ_β^2 controls how similar or different the effectds in each city are. When σ_β^2 is small, the size of the effects is mostly similar in all cities. As σ_β^2 gets larger, the effects in each city are likely to be different and thus the association between $Y^i(s)$ and $x_j^i, j = 1, 2, 3$ varies cities. Additionally, we generate a separate realization of a Gaussian process $\boldsymbol{\Sigma}^i(\boldsymbol{\theta}^i)$ from an isotropic exponential covariance function for each city individually. The spatial process is designed to model excess variance that is spatially smooth that is not attributable to the geographic covariates. Because our simulations are designed to examine the utility of combining estimates between cities, we let the covariance parameters of the spatial process be known a priori rather than estimating them, as one would do in practice.

Using the generate surfaces, $Y_i(s)$, we compare three different methods:

1. Individual: Random Forests of spatially adjusted regression trees, fit in each city separately.

2. Multi-City: Random Forests of multi-city regression trees, fit for all cities concurrently.

3. Multi-City Data-Enriched : Random Forests of multi-city data-enriched regression trees, fit for all cities concurrently. The tuning parameter is selected by cross-validation.

Simulation results in Figure 4.3 show the relative mean square prediction error of each of the three different methods as a function of σ_β^2 using the individual city regression trees as a baseline for prediction accuracy had no information been shared across cities. Our simulations suggest that combining information to learn the structure of the tree leads to better prediction accuracy when the effect is truly similar between cities, or when σ_β^2 is small.

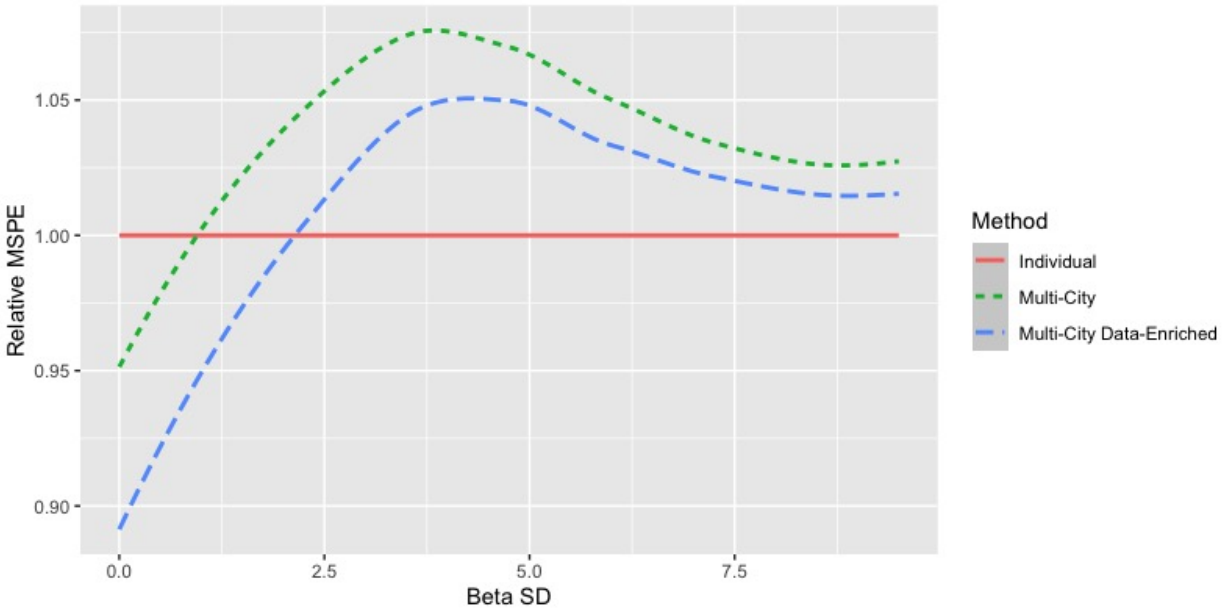

Figure 4.3: Simulation results showing mean squared prediction error (MSPE) of each of the methods scaled using Individual as a baseline. The green short-dashed line shows MSPE of multi-city trees, which estimates the coefficients in each city individually. The blue long-dashed line displays MSPE of multi-city data-enriched trees, which uses an ℓ_2 penalty to shrink coefficient estimates towards each other.

As σ_β^2 increases, we see that that our multi-city trees become inefficient and lead to worse prediction accuracy than simply building individual models in each city. In our simulations, sharing information improves on MSPE when the standard deviation of the coefficients is less than ≈ 1 for multi-city trees, and ≈ 2 for multi-city data-enriched trees. Comparing the multi-city approaches, we see that the shrinkage based data-enriched estimates are strictly better than the multi-city trees with no shrinkage. This difference is most noticeable when σ_β^2 is small, and the differences between the two methods shrinks as σ_β^2 increases.

These results make sense intuitively; if the effects are similar across cities combining information aids in attaining better estimates. If the effects are widely different in each city, trying to combine information becomes inefficient and better estimates are attained by constructing models in each city individually. Our results suggest that simply combining models across cities is not guaranteed to improve estimates in each city individually, and applications of our method should be restricted to cities where the association between features and pollutant concentrations are similar to maximize predictive accuracy of our multi-city data-enriched regression tree models.

4.4 NOx Concentrations in MESA Air Communities

We apply our data-enriched multi-city regression trees to the MESA air snapshot campaign, which collected measurements in the six MESA cities during the winter of 2006-2007. Winter NO_x concentrations were estimated using the average of two week sampling periods at 747 locations across the continental United States. In total, the number of monitors in each city ranged from 95 to 203. For each of the monitoring locations, 642 geographic information system (GIS) covariates to use in the model including proximity measures (distance to nearest major road, intersection, truck route, railway, railyard, coastline, airport, and port) and buffer measures (major road length, truck route length, land-use category, long-term vegetation index, population density, and emission sources).

The MESA snapshot campaign sampled six different cities. In addition to the original six sampling regions, two additional regions were added: Riverside County, CA and Rockaland,

NY, in order to enhance gradient exposures among the cohorts. While Los Angeles is a sprawling metropolis and no distinct differences between the original sampling area and Riverside County, CA are noticeable looking at Figure 4.1, the two sampling areas exhibit large differences in overall level in New York with levels generally very high in New York City, and very low across west of the Hudson River in Rockland County. Although these measurements are all from the same state, there is little reason to believe that NO_x models in Rockland County, NY should be similar to New York City, NY and we build separate models for each of these regions individually. We divided up our observations into seven distinct regions: (1) Los Angeles, CA, (2) New York City, NY, (3) Rockland, NY, (4) St. Paul, MN, (5) Baltimore, MD, (6) Winston-Salem, NC, and (7) Chicago, IL.

Since the MESA snapshot campaign specifically focused on near road gradients, we selected three fixed covariates ot use, $\mathbf{Z}_l, l = 1, 2, 3$, meters to A1 roads, meters to A2 roads, and meters to A3 roads. In a similar manner to Mercer et al.[56], these covariates were transformed to the $\log_{10}$ scale to model dispersion to exponentially decrease away from major roadways. The remaining 639 GIS covariates were used in constructing the regression tree component of the additive model.

For each of the 7 individual MESA snapshot campaign regions, we fit an exponential spatial correlation model with range fixed to the max distance between points in that particular city, as suggested by Kammann and Wand [39]. This choice of range parameter is arbitrary, but in related studies on air pollution no major differences in prediction accuracy were observed [61]. Rather than estimate a separate sill, σ_i^2, and nugget in each city, τ_i^2, we define a single covariance parameter for each city, the proportion of variance attributable to the sill as $\delta_i = \sigma_i^2/(\sigma_i^2 + \tau_i^2)$. This approach is taken in Chaper 2, and Gerber and Nychka [23] showed that similar parameter estimates can be obtained for this parameter either by cross-validation or maximum likelihood. The city-specific covariance parameters, $\hat{\delta}_i$ are selected by fitting random spatial forests in each city individually and selecting δ_i which minimizes the cross-validation error.

Since we would expect the effects of roads on air pollution concentrations across MESA

cities, we penalized the fixed effects, $\mathbf{L}^{\mathbf{z}_1} = d\mathbf{I}_d - \vec{\mathbf{1}}_d^T\vec{\mathbf{1}}_d, l = 1, 2, 3$. We assumed the fixed effects to be independent of the city-specific intercepts and let $\mathbf{L}^{\mathbf{z}0}+ = \mathbf{L}^{0\mathbf{z}} = \mathbf{0}^{d\times d}$. In order to let each city have its own city-specific intercept, we set $\mathbf{L}^0 = \mathbf{0}^{d\times d}$. For each successive contrast, we penalized the size of the contrasts to be similar across cities as $\mathbf{L}^i = d\mathbf{I}_d - \vec{\mathbf{1}}_d^T\vec{\mathbf{1}}_d, i = 1, \ldots, k$.

Three different methods of building multi-city data-enriched models are examined. First, we apply our multi-city data-enriched regression trees to all cities at once. Second, two separate multi-city data-enriched regression trees are constructed; one for the largest cities Los Angeles and New York City, and another for remaining cities. The second approach is motivated by our simulation study, where we found that combining information across cities is most efficient when the effects across cities are similar. Los Angeles and New York City are two major cities in the United States with large populations and infamously congested streets, and it would be reasonable to expect that NO_x models built in these regions would be unlike those constructed in the remaining United States. Third, we consider fitting an alternative method to dividing up the cities using a single multi-city data-enriched regression tree model. Here, we change the penalty to only shrinking estimates in New York and Los Angeles to be similar, and separately shrinking contrasts in the remaining five cities to be similar. This can be written as

$$\lambda\boldsymbol{\alpha}^T\mathbf{L}\boldsymbol{\alpha} = \lambda\sum_{i=1}^{k}\left(\sum_{j\in\{LA,NY\}}\sum_{k\neq j}\left\|\alpha_j^i - \alpha_k^i\right\|_2^2 + \sum_{j\in\{CHI,BAL,MSP,WS,RC\}}\sum_{k\neq j}\left\|\alpha_j^i - \alpha_k^i\right\|_2^2\right).$$

In total, four different method are compared:

1. Individual: Random Spatial Forests (Chapter 2), fit in each city separately.

2. Multi-City (All): Random Forests of multi-city data-enriched regression trees, fit for all cities concurrently.

3. Multi-City (Split): Random Forests of multi-city data-enriched regression trees, one model for New York and Los Angeles and one for Chicago, Baltimore, St. Paul, and

Winston-Salem and Rockland County.

4. Multi-City (Divided): Random Forests of multi-city data-enriched regression trees, with all cities fit in a single model. The penalty is adjusted so that estimates in New York and Los Angeles are encouraged to be similar, and separately, estimates in Chicago, Baltimore, St. Paul, and Winston-Salem and Rockland County are encouraged to be similar to each other.

Performance of each approach was estimated by ten-fold cross validation. Because of the nature of how our sampling locations were selected, monitors were grouped by cluster and all monitors in a cluster were kept in the same cross validation fold. Cross-validated cRMSPE of each method is shown in table 4.4.

We first compare applying multi-city data-enriched regression trees to all cities at once (Multi City (All)) to the approach which shares no information across cities and constructs a separate model for each city individually (Individual). Combining information across cities leads to lower cRMSPE on average across the seven different regions, but this approach works better in some regions than others. Multi-city data-enriched regression trees are shown to have lower cRMSPE in New York, Winston-Salem and Rockland County, but perform worse in Chicago, Baltimore, St. Paul, and Los Angeles.

Our simulation study demonstrated that the gain in efficiency of combining observations across cities is maximized when the cities are actually similar. Comparing the single multi-city models with different versions of penalties (All vs Divided), splitting cities up based on prior knowledge always has lower cRMSPE than regularizing all cities together. Alternatively, our approach which splits Los Angeles and New York into one model, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County into a separate model improves prediction error over individual city approach in every subregion. Comparing the two alternative ways of dividing cities (Split vs Divided), neither approach strictly dominates the other. Building separate models (Split) has lower cRMSPE averaged across the seven subregions, and performs better in about half the cities, Los Angeles, New York, Chicago, and

Winter 2006-2007	Individual	Multi-City (All)	Multi-City (Split)	Multi-City (Divided)
Average	15.40	14.53	**14.08**	14.37
Low Angeles	18.18	18.05	**17.47**	18.07
New York City	27.24	23.43	**22.60**	23.03
Chicago	7.85	8.41	**6.56**	7.88
St. Paul	6.04	6.02	5.94	**5.55**
Baltimore	5.70	6.67	**5.19**	6.24
Winston-Salem	10.11	8.98	8.91	**8.77**
Rockland County	6.22	5.55	5.69	**5.53**

Table 4.1: Comparing expected mean square prediction error at a cluster (cRMSPE) for the three different approaches of applying random forests to spatially adjusted regression trees in the MESA air snapshot campaign cities.

Baltimore, while splitting cities up by the penalty (Divided) performs better in St. Paul, Winston-Salem, and Rockland County. The divided and split methods are actually very similar with a few key differences: The divided approach uses a single set of tree designs for all cities, while the split approach uses one set of tree designs for Los Angeles and New York, and another set for the remaining five cities. Further, the split approach allows for two separate penalties, while the divided approach uses a single penalty, and it needs to be decided a priori how to scale the two groups of cities against each other.

In New York City (Figure 4.5), pooling information using multi-city (all, divided, and split) improves estimates most notably when the observed NO_x levels are low. Individual city models tend to badly overestimate the pollutant levels in these locations, while our multi-city models mitigates the size of these errors. Residual plots (Figure 4.12) show north from Manhattan, the size of the errors using individual city models tends to be more extreme in these areas where pollutant levels are lower overall. A similar relationship is seen when crossing over the river from Manhattan to Brooklyn, where the individual city model badly

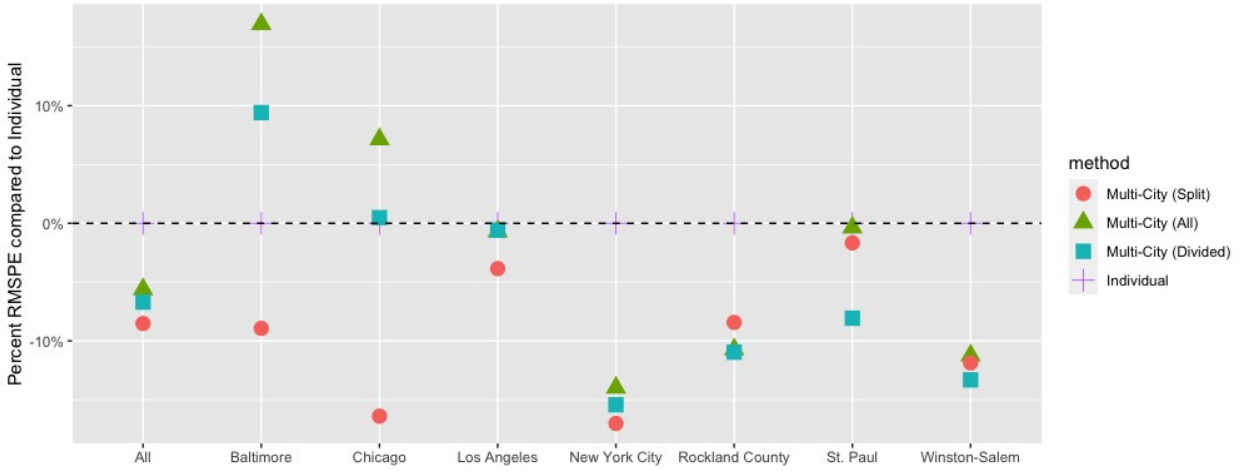

Figure 4.4: Percent change in cross-validated RMSPE across MESA Air cities by using multi-city trees, compared to using individual city specific trees.

overestimates pollutant levels directly across the river in Brooklyn, while the magnitude of these overestimated levels is lower using the multi-city models.

In Los Angeles, pooling information using all cities results in worse prediction accuracy than an individual model, but pooling only with New York City results in improved cross-validated prediction accuracy. Scatter plots of predicted against observed pollutant levels in Figure 4.6 demonstrate this problem. By combining information across all cities, multi-city (all) over regularizes our prediction, and our predictions underestimate the size of the effects within Los Angeles. By combining Los Angeles only with another large metro area in New York City, we are still able to model the wide varying range of pollutant levels in large cities, but with reduced variance. We do note that across the city center of Los Angeles the magnitude of errors appears to be slightly smaller throughout using all multi-city methods (all, split, and divided) when compared to the individual city models..

Improvements in prediction accuracy in Chicago are readily seen from scatter plots in Figure 4.7. Examining the residual plots (Figure 4.14) demonstrate that pooling information across cities using both of our multi-city models helps most noticeably near the suburbs to the

northwest and southeast of the city of Chicago, as the magnitude of errors in these locations is lower by pooling. The benefit of splitting the large cities and smaller cities up here is most noticeable in the monitor downtown by the waterfront; while the multi-city (all) model that combines all cities badly overestimates NO_x levels at this location, by not including city contrasts from Los Angeles and New York City we mitigate the size of pollutant estimates in this city center monitor in both the split and divided approaches. Here, the split approach, which constructs a different set of regression trees, seems to mitigate the overestimation in the city center better than the divided approach, and has lower average cRMSPE in Chicago.

In Winston-Salem and Rockland County, both methods combining information across cities using our multi-city models improve on cross-validated prediction accuracy over individual city trees. From the scatter plots in Figures 4.10 and 4.11, this advantage is most noticeable seen in areas where pollutant levels are low, as all multi-city regression tree methods tend to predict more accurately across locations where the true NO_x levels are less than ≈ 30.

Scatter plots for St. Paul and Baltimore are shown in Figures 4.9 and 4.8 and residual plots are included in Figures 4.16 and 4.15. We note that in both of these cities the size of improvements gained by pooling information is rather small, and predicted pollutant NO_x levels and residual maps are similar across the different methods.

Across all cities, we see similar patterns showing the benefits of combining information across cities using multi-city models. Scatterplots demonstrate that by combining observations across cities, we obtain better predictions in areas where lower pollutant levels are observed. These lower pollutant levels are often seen in areas farther away from city centers, and the number of monitors in these areas are more sparse. Residual plots demonstrate this effect northwest and southeast of the city center of Chicago, to the north and southeast of Manhattan, and west to Santa Monica and north to Pasadena in Los Angeles, where the multi-city models overpredict NO_x levels less badly than individual city models. By combining contrast estimates across cities, we are often able to improve predictions in the city outskirts, where the MESA air snapshot campaign placed fewer monitors.

The advantage of splitting up cities into large and small cities is most obvious in the city centers of Baltimore and Chicago, where the multi-city (all) model badly overestimates pollutant levels in the downtown location nearest to the waterfront. This shows how pooling estimates across cities that are not similar can cause erroneous predictions. When estimates in Chicago are combined with Los Angeles and New York City, pollutant estimates are estimated to be vastly too high in the city centers of Chicago.

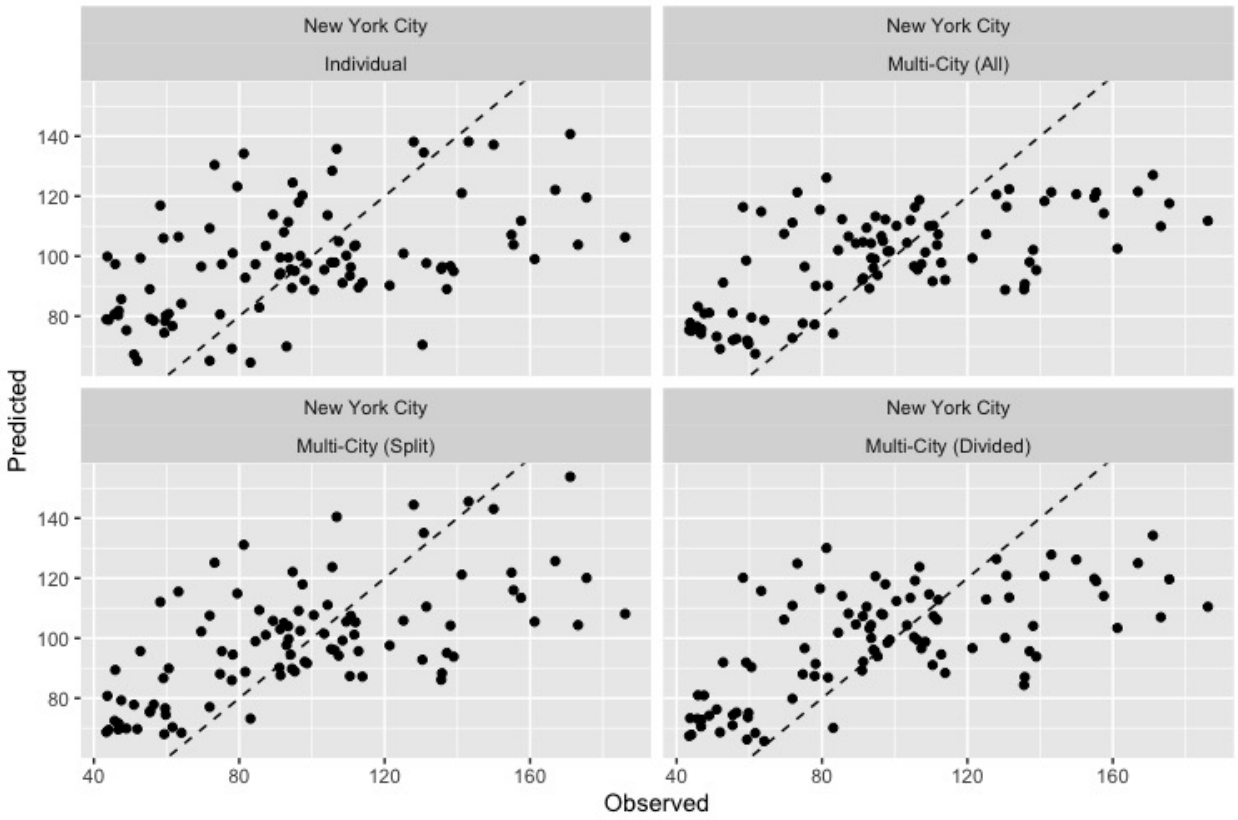

Figure 4.5: Scatter plots of predicted (y-axis) vs observed (x-axis) pollutant concentrations in New York City using: **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

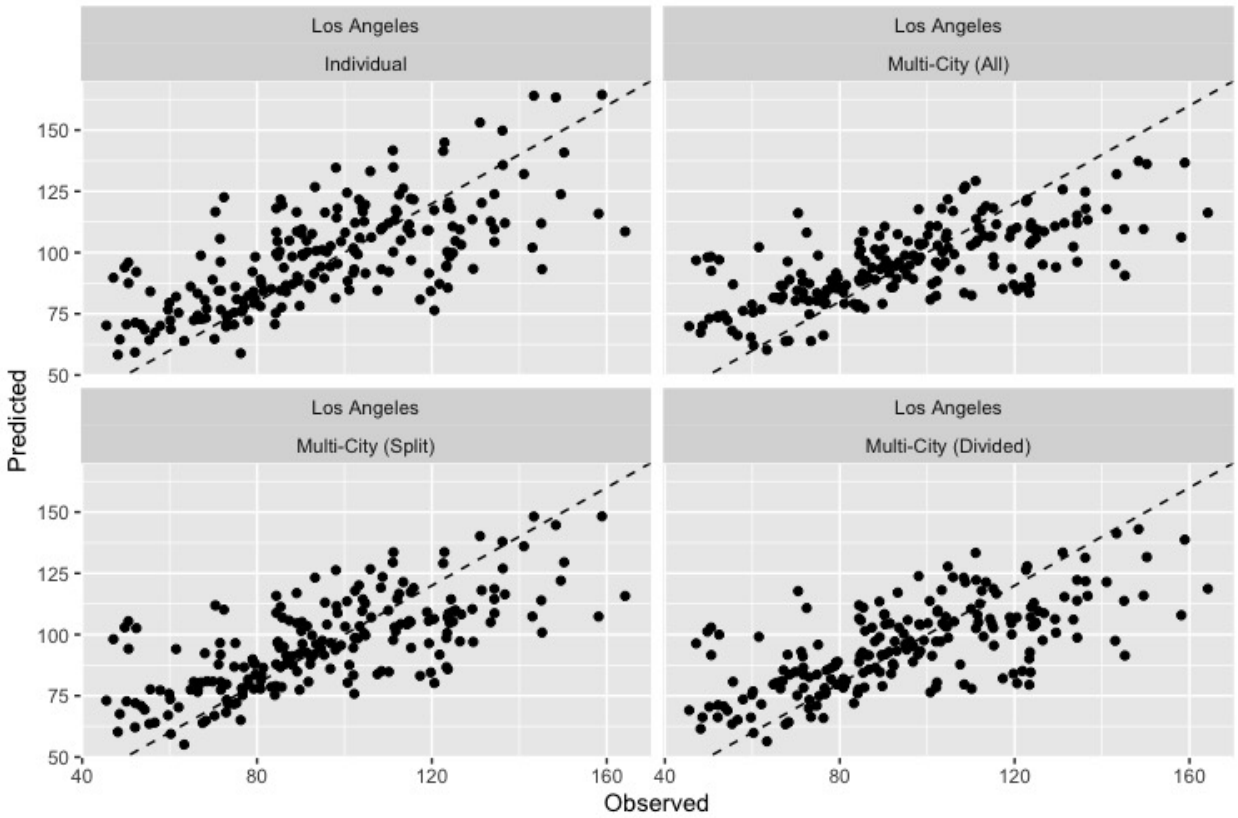

Figure 4.6: Scatter plots of predicted (y-axis) vs observed (x-axis) pollutant concentrations in Los Angeles using: **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

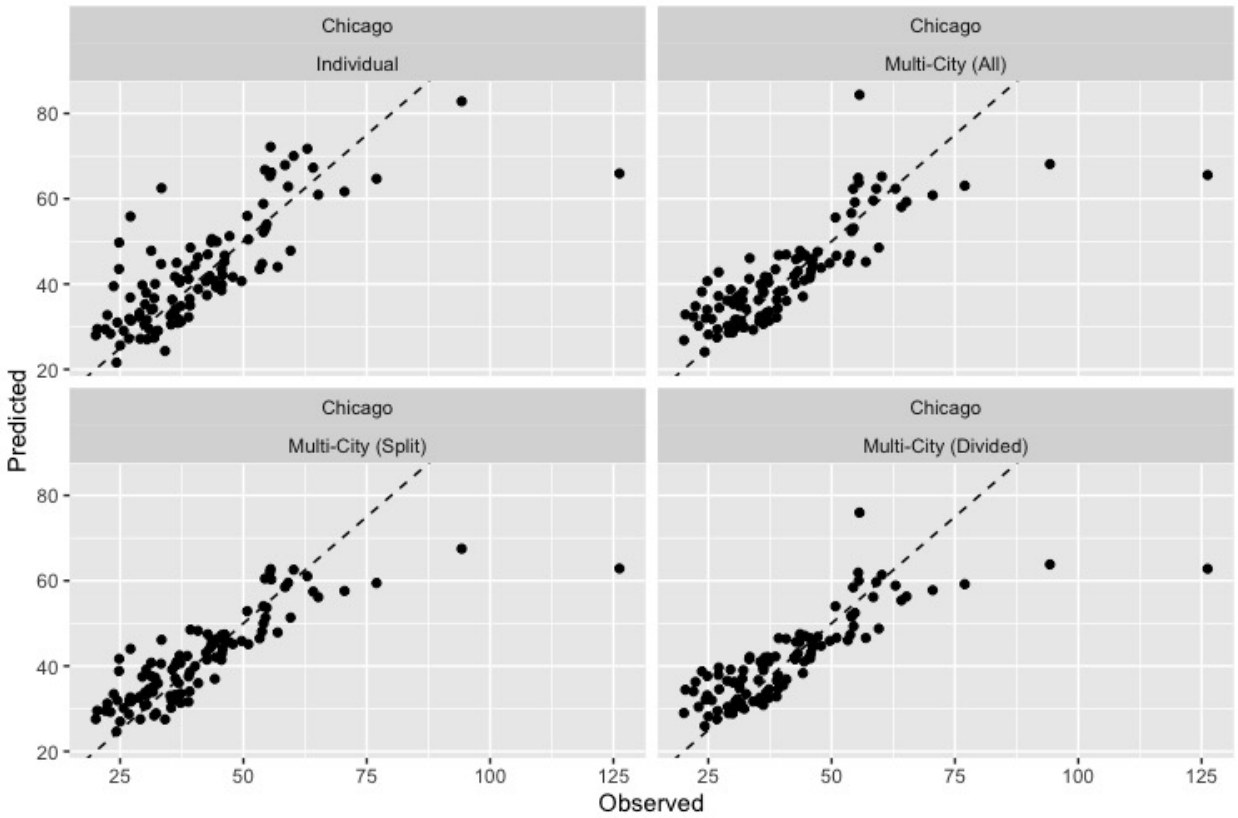

Figure 4.7: Scatter plots of predicted (y-axis) vs observed (x-axis) pollutant concentrations in Chicago using: **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

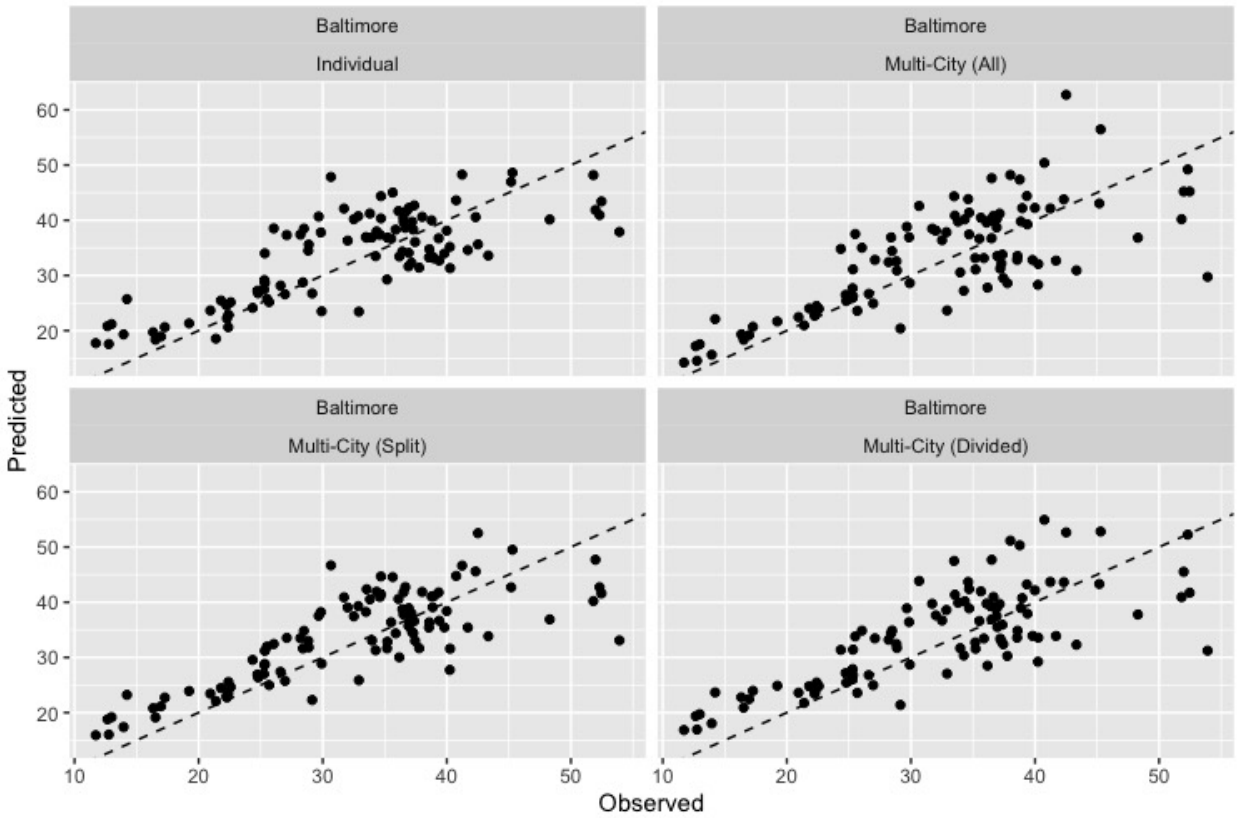

Figure 4.8: Scatter plots of predicted (y-axis) vs observed (x-axis) pollutant concentrations in Baltimore using: **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

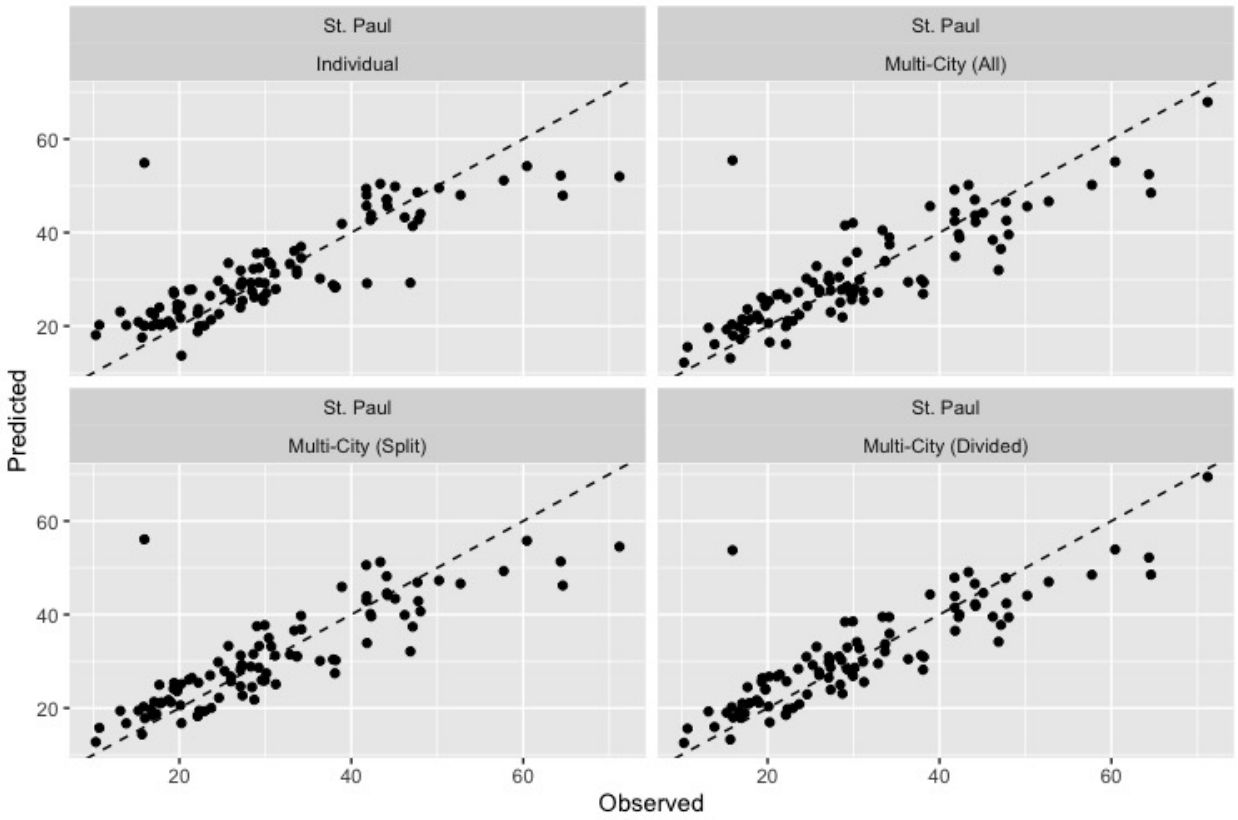

Figure 4.9: Scatter plots of predicted (y-axis) vs observed (x-axis) pollutant concentrations in St. Paul using: **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

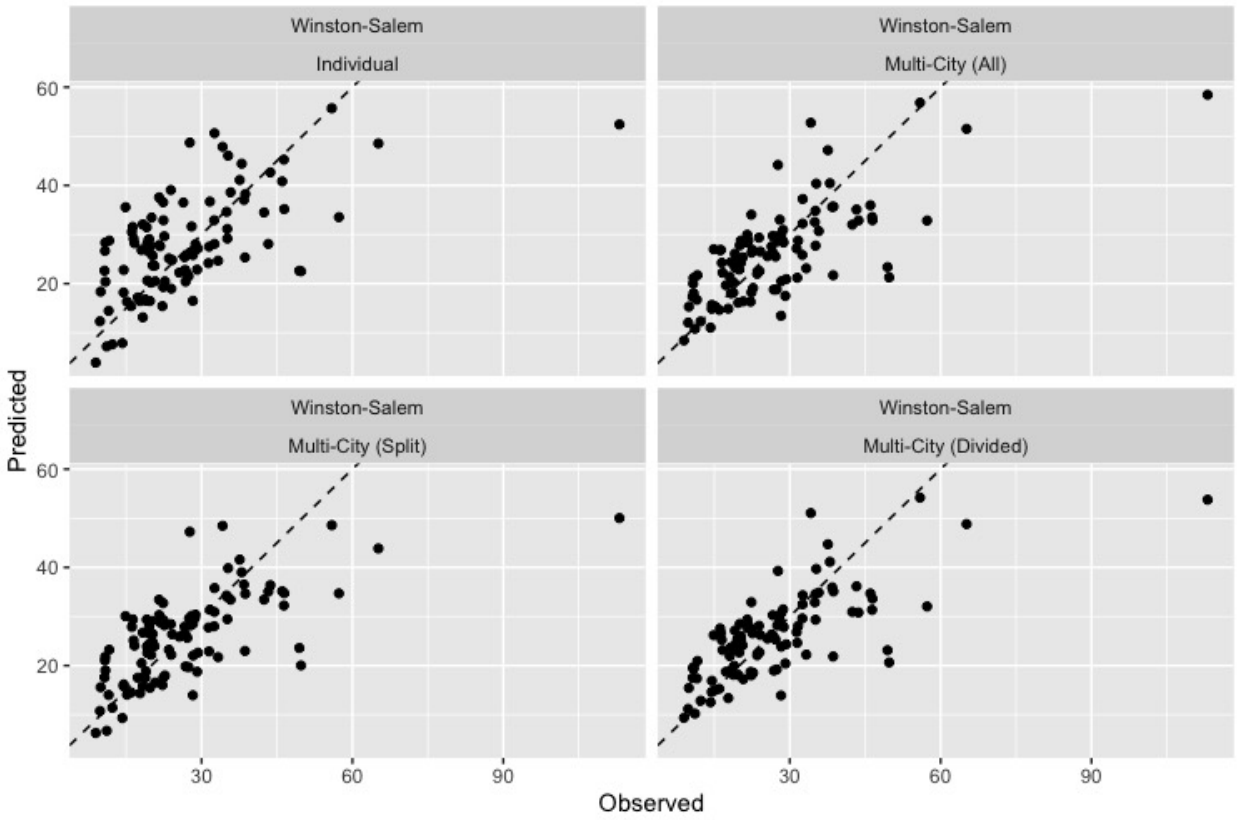

Figure 4.10: Scatter plots of predicted (y-axis) vs observed (x-axis) pollutant concentrations in Winston-Salem using: **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

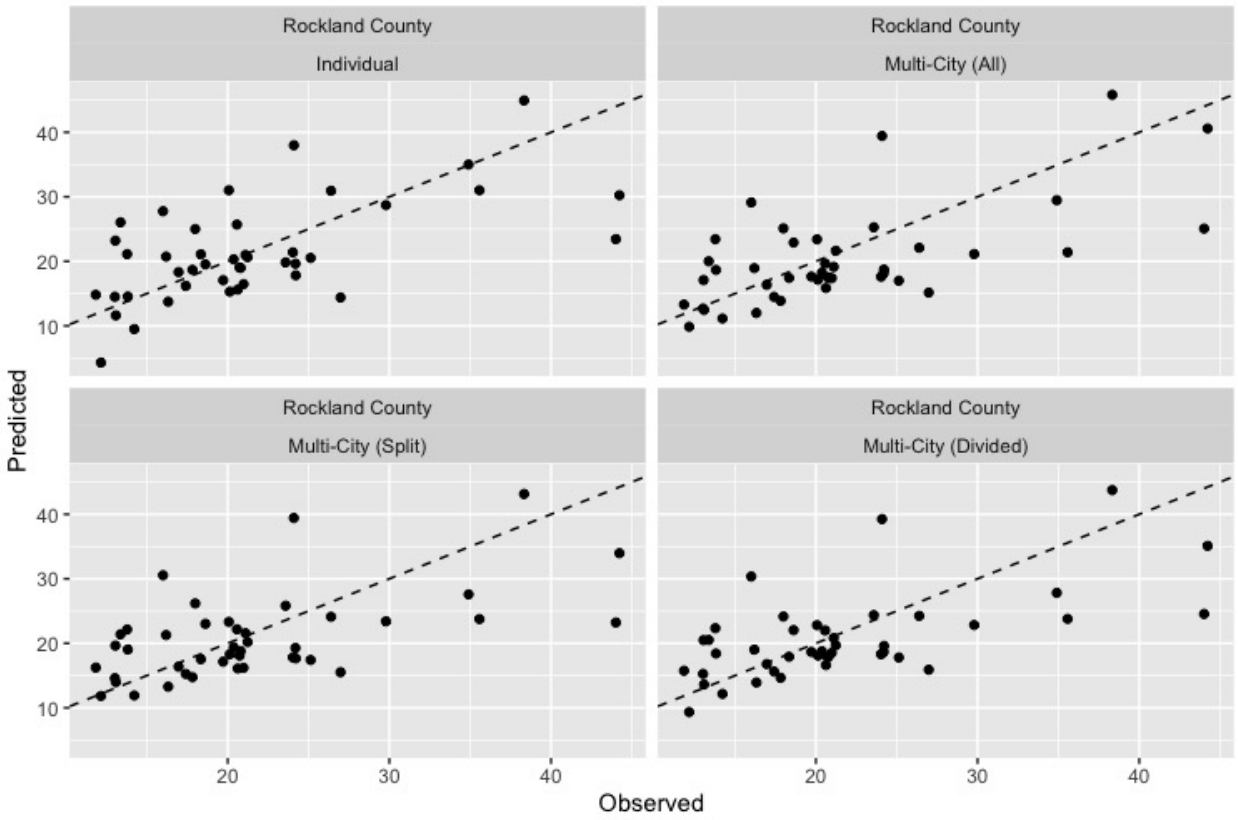

Figure 4.11: Scatter plots of predicted (y-axis) vs observed (x-axis) pollutant concentrations in Rockland County using: **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

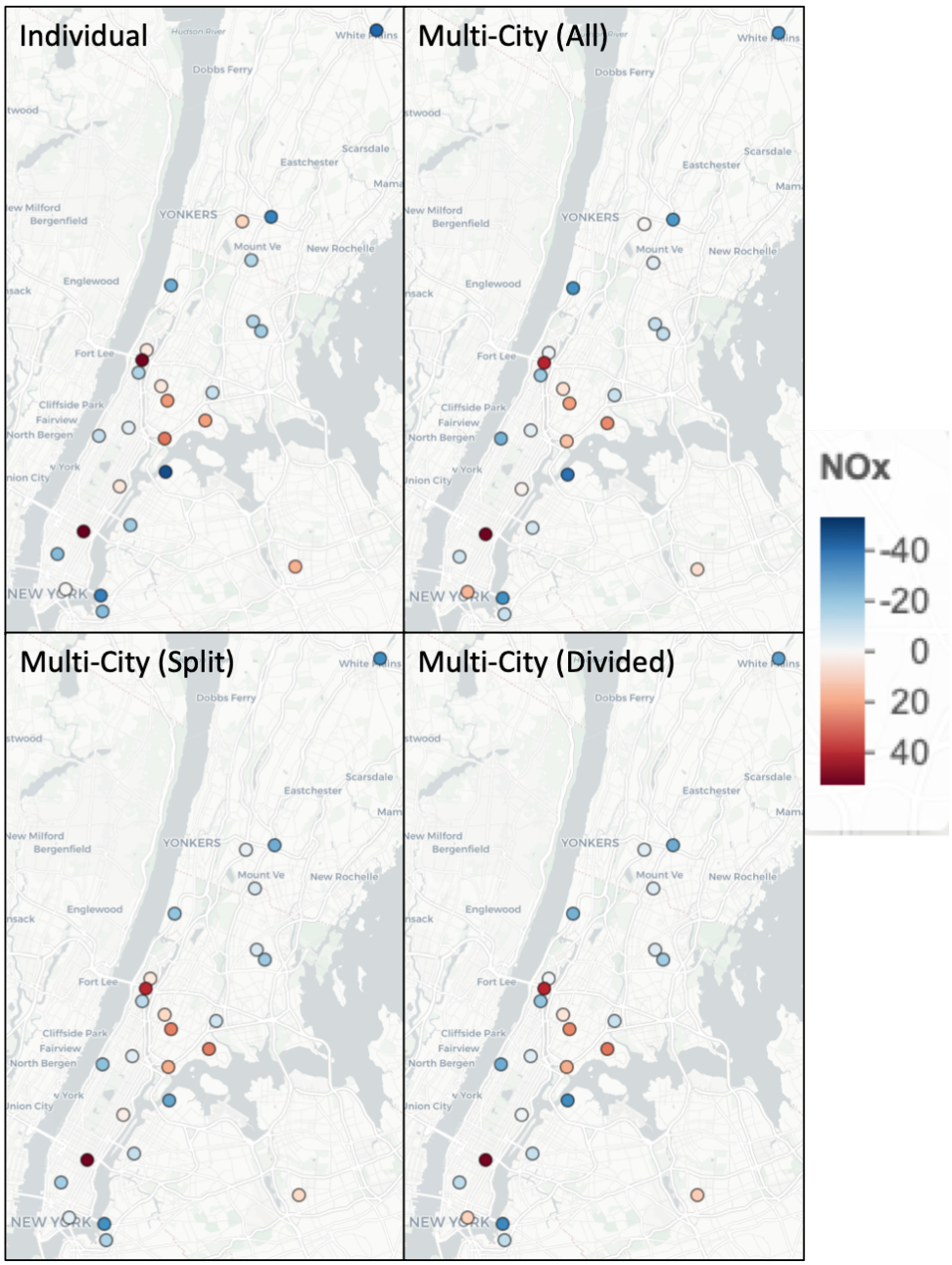

Figure 4.12: Cross-Validated NO_x residuals, presented as the average in each cluster, in New York City using **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

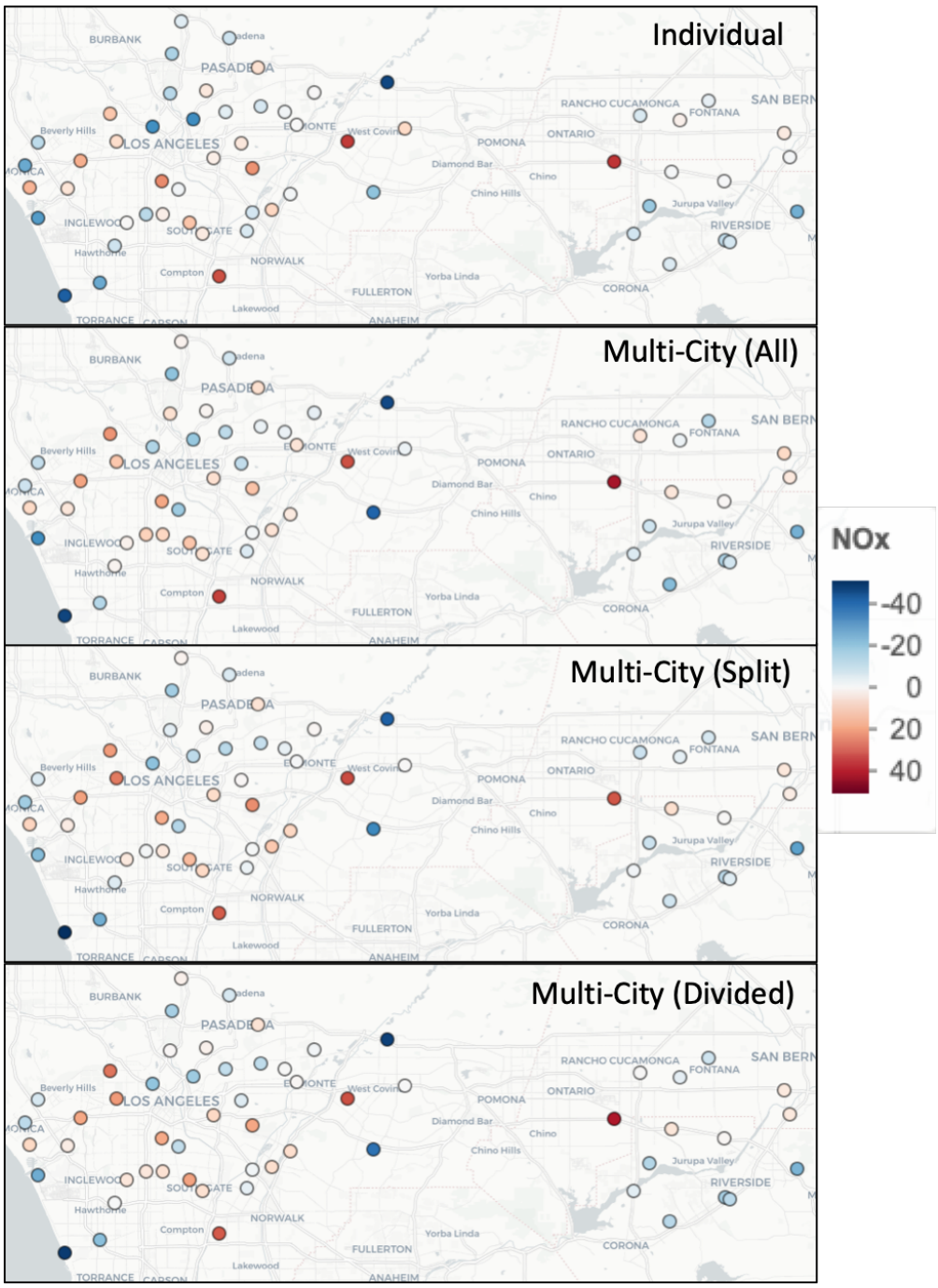

Figure 4.13: Cross-Validated NO_x residuals, presented as the average in each cluster, in Los Angeles using **Top:** Individual trees, **Upper Middle:** Multi-city data-enriched regression trees with all cities as a single model, **Lower Middle:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

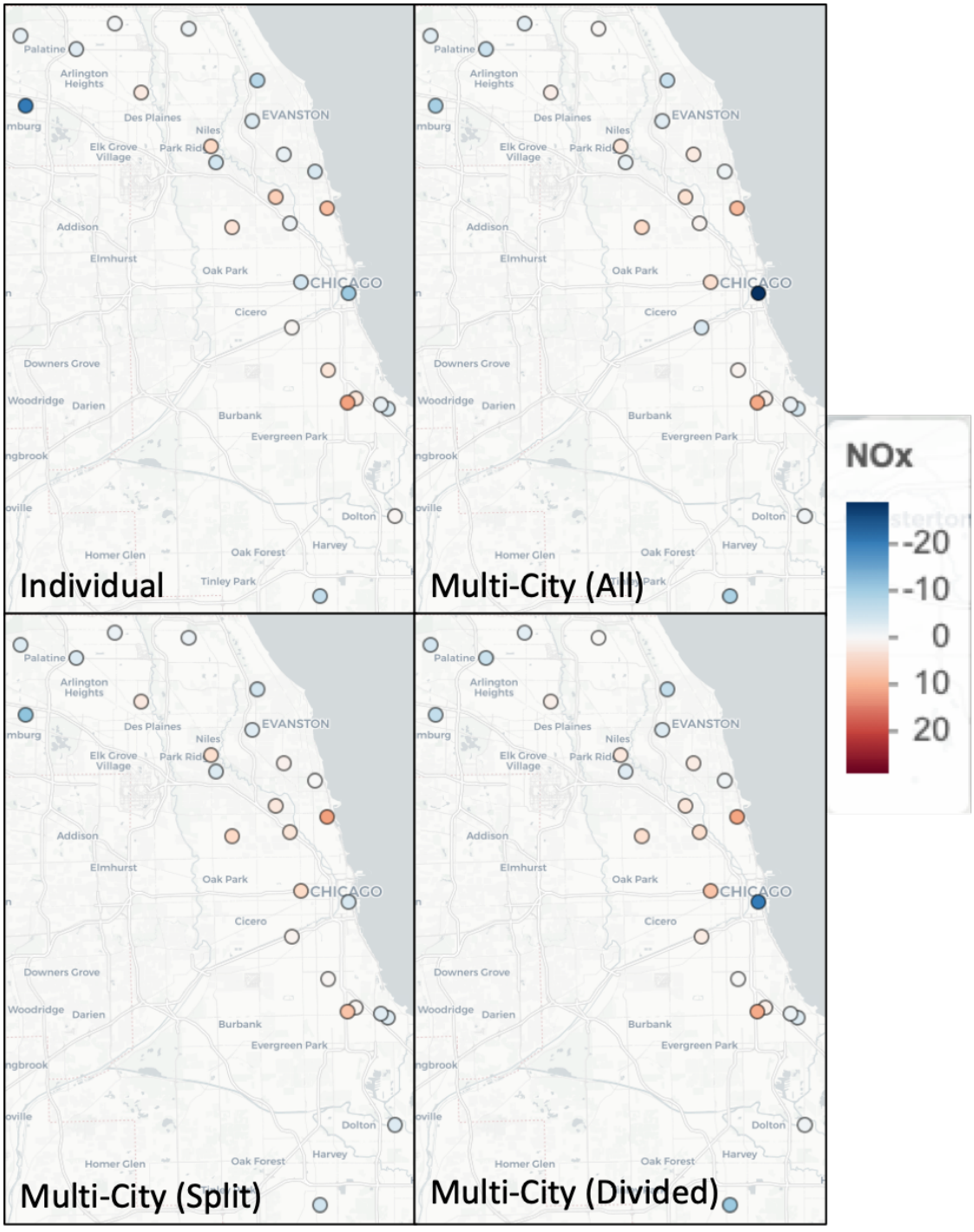

Figure 4.14: Cross-Validated NO_x residuals, presented as the average in each cluster,in Chicago using **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

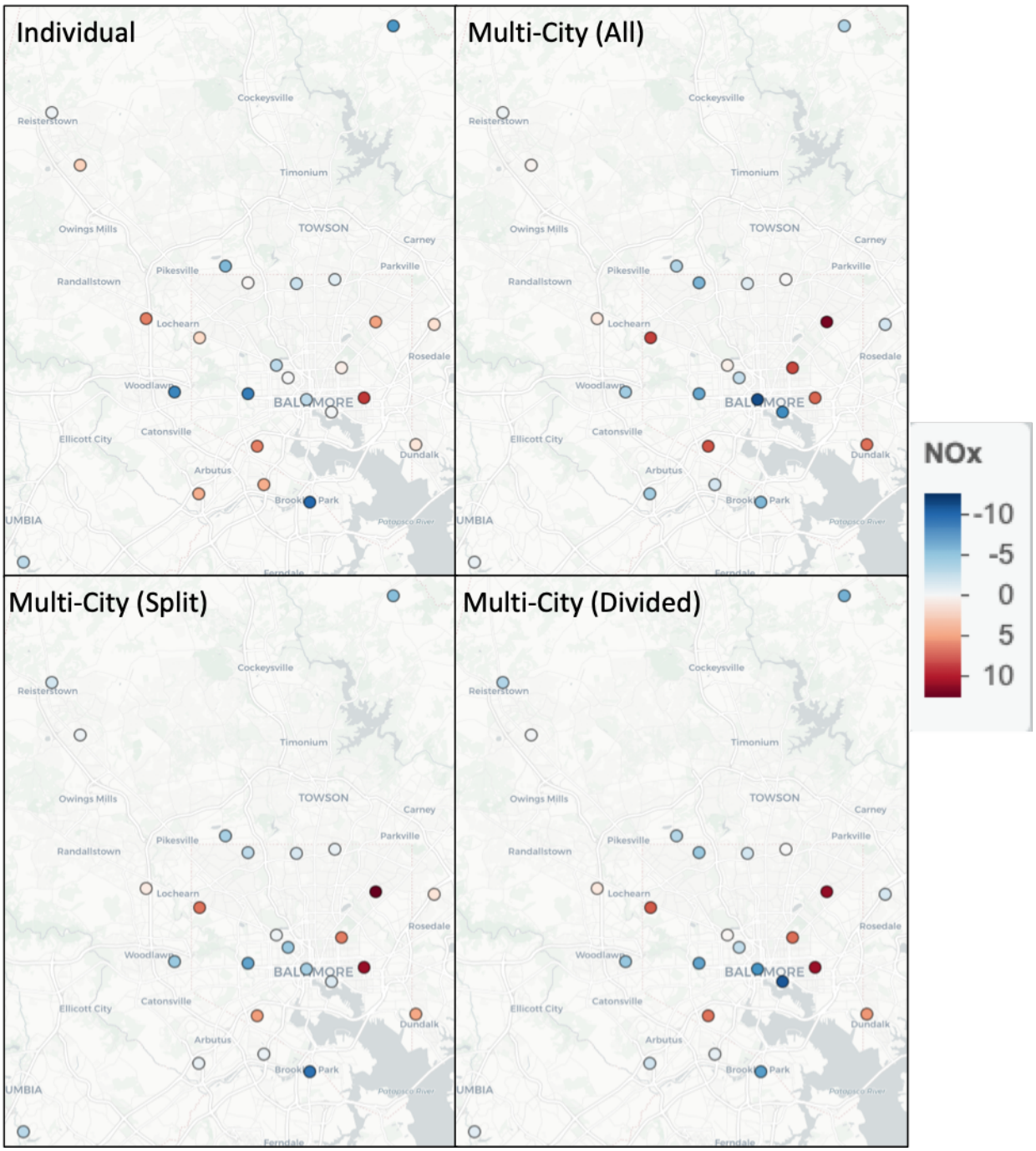

Figure 4.15: Cross-Validated NO_x residuals, presented as the average in each cluster, in Baltimore using **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

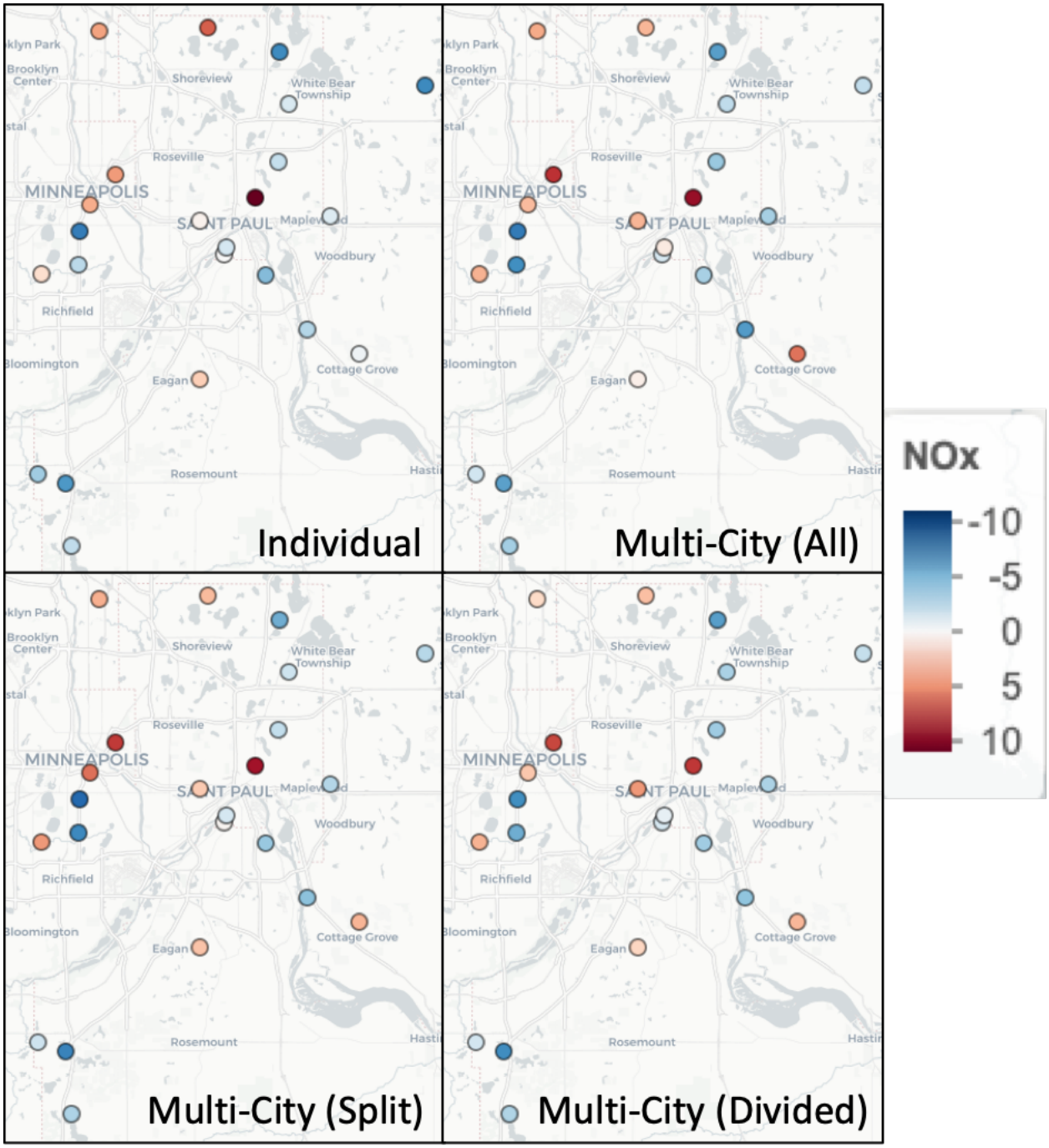

Figure 4.16: Cross-Validated NO_x residuals, presented as the average in each cluster, in St. Paul using **Left:** Individual trees, **Center:** Multi-city data-enriched regression trees with all cities as a single model, and **Right:** Multi-city data-enriched regression trees with cities divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

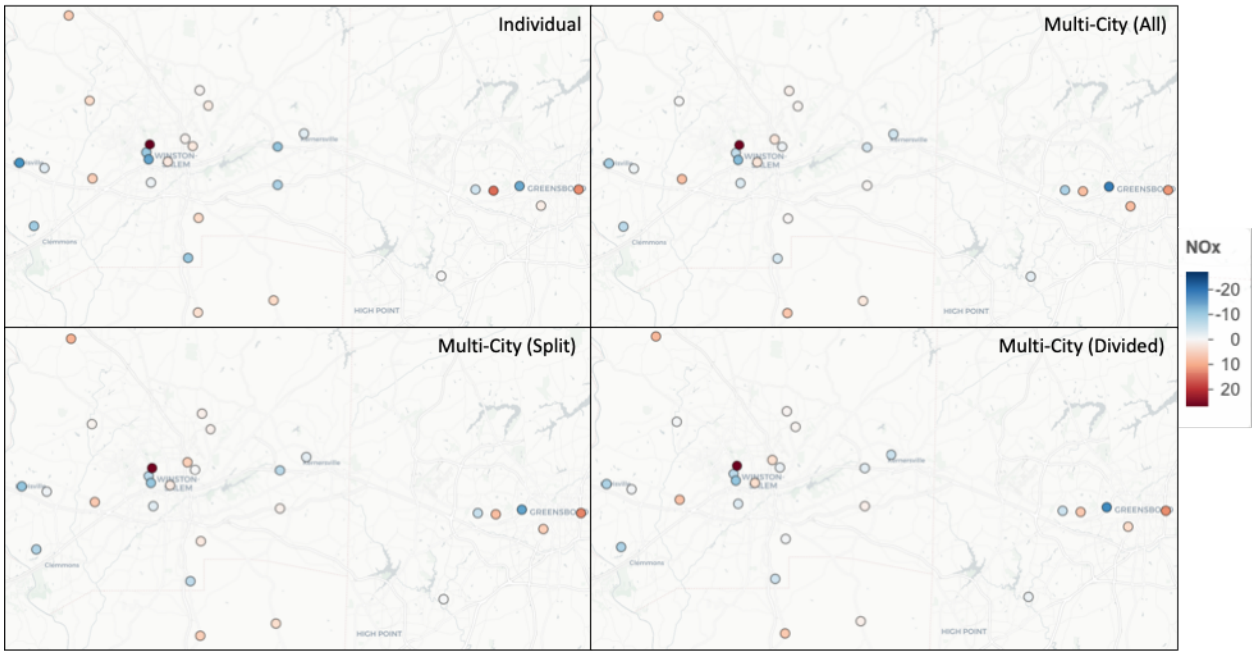

Figure 4.17: Cross-Validated NO_x residuals, presented as the average in each cluster, in Winston-Salem using **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

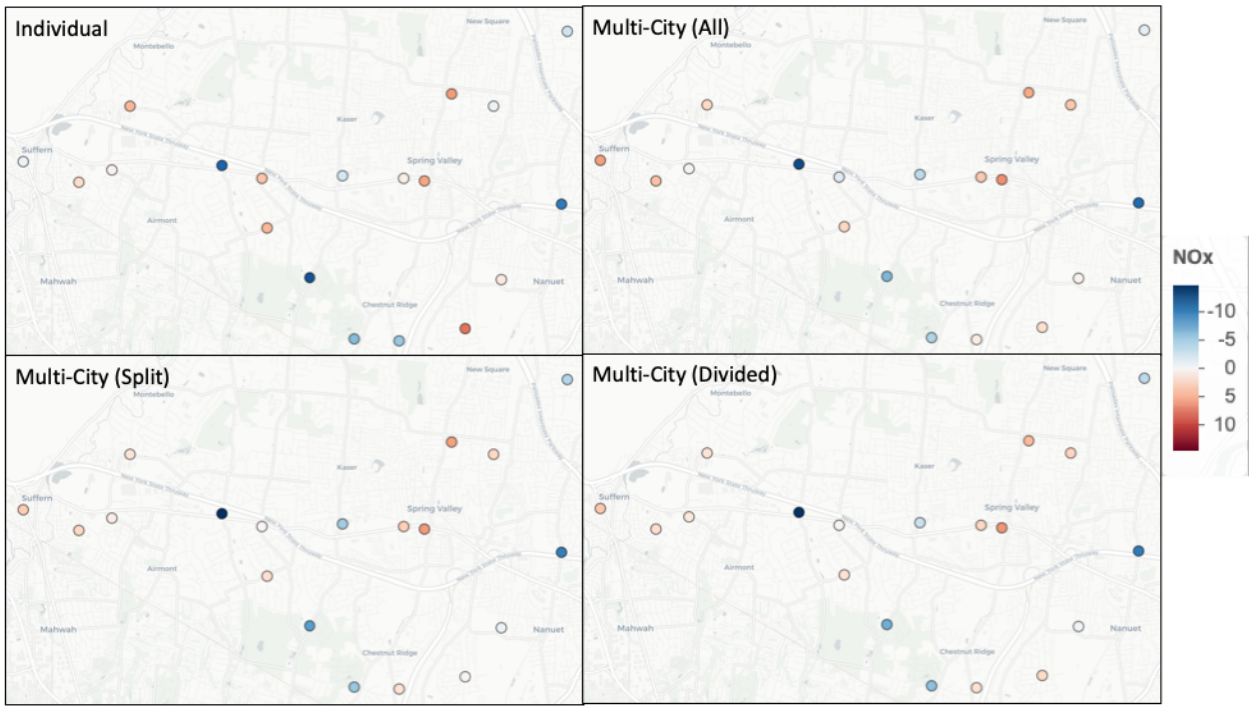

Figure 4.18: Cross-Validated NO_x residuals, presented as the average in each cluster, in Rockland County using **Top Left:** Individual trees, **Top Right:** Multi-city data-enriched regression trees with all cities as a single model, **Bottom Left:** Multi-city data-enriched regression trees with cities split into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County, and **Bottom Right:** Multi-city data-enriched regression trees with the penalty divided into New York City and Los Angeles, and Chicago, St. Paul, Baltimore, Winston-Salem, and Rockland County.

Chapter 5

DISCUSSION

5.1 Summary

In Chapter 2, we presented a novel interpretation of regression trees in the form of a linear model, suggesting a principled approach to estimating regression trees which allow for correlation. By carefully constructing the tree design matrix, we show that this approach lends itself to efficient computation by taking advantage of its block structure. Through simulation results and on observed annual average EC, OC, Si, and S from 2009-2010, we demonstrate that this approach results in more accurate predictions than two-step estimation methods.

In Chapter 3, we proposed a framework to incorporate shrinkage approaches from the statistical learning literature into a universal kriging model which does not require pre-processing the covariates and scales for both large numbers of observations and covariates. Our simulations suggest that shrinkage offers large gains in prediction accuracy over existing approaches under a correctly specified mean and optimization of the penalized models by cross-validation scale demonstrably better than likelihood based approaches and are significantly easier to optimize. Across a range of pollutants collected in 2009-2010, shrinkage does at least as well as existing approaches for all pollutants, with a noticeable advantage for elemental carbon.

While our approach does not require pre-processing of the covariates, it does requires a pre-specification of the spatial basis functions for the residual spatial process. In a typical spatial analysis, selecting the spatial correlation function involves fitting a parametric model to the sample variogram by "eye" [66]. However, this approach is not straightforward when using geographic covariates as the residual covariance parameters (partial sill, nugget, and range) do not represent quantities describing the observed spatial process of interest, but

rather the residual spatial process, or the excess variance which cannot be attributed to the fixed effects. Since the fixed effects also depend on the selected residual spatial process, it is difficult to determine the variogram which the residual spatial process is to be fit to. In our simulations, mis-specifying the residual spatial process has no noticeable effect on prediction accuracy.

In Chapter 4, we demonstrated that by using multi-city data-enriched regression trees, we are able incorporate similar observations across cities and construct more accurate predicted pollutant exposures in each city individually and overall. Our simulations demonstrate that when the effects in each city are truly similar, building a common mean structure improves on constructing a separate model in each city, although this approach becomes inefficient as the effects between cities differs.This effect is demonstrated by the MESA air snapshot campaign NO_x measurements. If a single model aiming to find a shared mean structure between all six cities is fit, multi-city data enriched regression trees improves performance in Los Angeles, New York, and Winston Salem but results in worse prediction accuracy in Chicago, Baltimore and St. Paul. By dividing mega-cities Los Angeles and New York into one model and the other four cities into another, we see that performance is at least as good or better than individual models in all cities. By using the linear model formulation of regression trees, we showed that a flexible class of additive regression tree models which allow for the inclusion of fixed effects, multiple clustered observations, shrinkage between parameter estimates and allowing for correlation. These regression trees in these settings can still be estimated by recursive binary splitting in sub-$\mathcal{O}(n^3)$ time. These additive models of regression trees can be used in popular statistical learning such as random forests, and incorporating classical statistics components such as fixed effects and spatial correlation can improve prediction accuracy over naive applications of these ensemble approaches.

5.2 Future Work and Extensions

In Chapter 2, our optimization approach took advantage of recent computational developments in spatial statistics to reduce the parameterization of the covariance to a single

parameter δ and selecting its value by grid search. We note here that this is not required, for example, one could consider using a normal kriging covariance with an exponential covariance function and select the parameters by grid search, but adding additional parameters becomes computationally expensive since the number of points to consider scales exponentially. Bayesian optimization and covariance matrix adaptation—evolution strategy have been used in the machine learning literature for gradient free optimization of "black-box" prediction models with stochastic function evaluation where multiple tuning parameters need to be selected. Both of these methods can be applied to random spatial forests and are easily parallelizable to make optimization feasible for more complex covariance function parameterizations.

We examined using random forests algorithm using our novel tree building algorithm to adjust for spatial correlation, but another popular tree based ensemble method is boosting, and it would be straightforward to apply our tree building algorithm to boost spatially adjusted trees. Our tree building algorithm adjusting for correlation is also not restricted to estimation in spatial applications. For example, prediction problems where it is desirable to adjust for correlation occurs in other application such as network-linked data [48].

The general approach of formulating a tree as a linear model would suggest that we could extend this method to generalized linear models (GLM) by adjusting the tree impurity metric to the negative log-likelihood of the selected GLM. However, this approach is computationally difficult. For the identity link, parameter estimates for the contrast vector $\boldsymbol{\pi}$ are profiled out, leading to a search only over candidate split vectors. For GLMs, there are no general closed form estimates for the corresponding parameters, thus each candidate split would require an inner optimization to obtain estimates for $\boldsymbol{\pi}$, which we suspect would make this approach prohibitively computationally intensive.

In Chapter 3, we looked at modeling annual averages. Rather than report annual averages at individual sites, a complete characterization of the pollutant levels at each site over time using a spatio-temporal model is often desired [49, 76]. Estimation in these settings is difficult, and there have been a number of proposals to make large spatio-temporal models

computationally feasible [40, 38, 57]. Penalized regression methods have also been shown to work well in practice for time series estimation [64], and future extensions for research would examine principled constructions which combine penalized approaches to both spatial smoothing and time series estimation to create penalized spatio-temporal models. These models would have a number of tuning parameters to select, but this may be addressed by optimization approaches from statistical learning [20]. As the goal in environmental epidemiological applications is strictly air pollution mapping, we can treat the unknown variables as tuning parameters in penalized regression models instead of latent variables in mixed models and take advantage of the growing body of statistical learning literature designed to address prediction in high-dimensional datasets.

In Chapter 4, we defined multi-city regression trees as a method for modeling structure between both the observations and features. We did not examine the use of penalized regression methods used in Chapter 3 in these settings, but these methods may be extended by a simple application of glm-funk [73]. This approach can be used for a wide variety of data types using GLMs, while our multi-city regression trees are restricted to continuous outcomes only.

The motivation for multi-city regression trees is that it identifies features which contribute to pollutant levels across cities in order to improve estimates. It is thus desirable to evaluate variable selection/variable importance measures to examine what the multi-city regression trees identifies as useful features. There are two common ways variable importance in random forests is typically defined, first by percentage of tree impurity and second using a permutation based approach. Percentage of tree impurity is difficult to define as our spatially-adjusted regression trees are additive models of two separate components, the regression tree and the spatial process. Further, the estimates in each terminal node change in every step, so the variance of each terminal node changes depending on the full structure of the regression tree. Permutation based approaches would seem to be a sensible solution to use instead, but this is also not the case. This approach is not defensible due to correlation between features, as the independence assumption is badly violated [34]. For example, it is

difficult to imagine urban land use and traffic intersections being independent, and thus even defining groups of independent features is hard to accomplish. Variable importance for non-parametric methods is explored in Williamson et al. [85], but even here they showed that identifying important features under correlation is difficult in practice. A possible extension to compute variable importance with correlated features is through the use of Shapley values [84], and future work may examine their use for variable importance of highly correlated geographic features in spatial applications.

In this dissertation we compared various statistical learning approaches in spatial applications by MSPE. While accurate predictions at un-monitored locations are desirable, it is often of interest to produce uncertainty estimates as well [18]. Inference for penalized regression methods (Chapter 3) is an active area of research [73, 90], and future work may examine coverage of these uncertainty estimates in spatial applications. Quantile random forests [55] is a widely used technique for producing uncertainty estimates for random forests estimates. This approach is not generalizeable to our spatially adjusted regression trees, as the non-negativity requirement of the weights for each observation are violated due to the correlation induced by the spatial process. Another direction which has been proposed is the use of "honest" trees [79], however this requires sample splitting to generate unbiased estimates. This is a form of a bias/tradeoff, and requires a large number monitors to get reasonable point estimates and uncertainty bounds, but collecting a sufficiently large number of unique monitoring sites is extremely difficult in spatial studies.

BIBLIOGRAPHY

[1] Rebecca C Abernethy, Ryan W Allen, Ian G McKendry, and Michael Brauer. A land use regression model for ultrafine particles in vancouver, canada. *Environmental science & technology*, 47(10):5217–5225, 2013.

[2] A Alimissis, K Philippopoulos, CG Tzanis, and D Deligiorgi. Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmospheric Environment*, 191:205–213, 2018.

[3] MA Arain, R Blair, N Finkelstein, JR Brook, T Sahsuvaroglu, B Beckerman, L Zhang, and M Jerrett. The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. *Atmospheric Environment*, 41(16):3453–3464, 2007.

[4] Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[5] Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

[6] Silas Bergen, Lianne Sheppard, Paul D Sampson, Sun-Young Kim, Mark Richards, Sverre Vedal, Joel D Kaufman, and Adam A Szpiro. A national prediction model for pm2. 5 component exposures and measurement error–corrected health effect inference. *Environmental health perspectives*, 121(9):1017–1025, 2013.

[7] Veronica J Berrocal, Yawen Guan, Amanda Muyskens, Haoyu Wang, Brian J Reich,

James A Mulholland, and Howard H Chang. A comparison of statistical and machine learning methods for creating national daily maps of ambient $pm_{2.5}$ concentration. *arXiv preprint arXiv:1904.08931*, 2019.

[8] Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, Brandon Willard, et al. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019.

[9] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[10] Aiyou Chen, Art B Owen, Minghui Shi, et al. Data enriched linear regression. *Electronic journal of statistics*, 9(1):1078–1112, 2015.

[11] Gongbo Chen, Shanshan Li, Luke D Knibbs, Nicholas AS Hamm, Wei Cao, Tiantian Li, Jianping Guo, Hongyan Ren, Michael J Abramson, and Yuming Guo. A machine learning method to estimate pm2. 5 concentrations across china with remote sensing, meteorological and land use information. *Science of the Total Environment*, 636:52–60, 2018.

[12] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[13] Alexandra Chouldechova and Trevor Hastie. Generalized additive model selection. *arXiv preprint arXiv:1506.03850*, 2015.

[14] Martin A Cohen, Sara D Adar, Ryan W Allen, Edward Avol, Cynthia L Curl, Timothy Gould, David Hardie, Anne Ho, Patrick Kinney, Timothy V Larson, et al. Approach to estimating participant pollutant exposures in the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Environmental science & technology*, 43(13):4687–4693, 2009.

[15] Noel Cressie. Comment: When is it data science and when is it data engineering? *Journal of the American Statistical Association*, 115(530):660–662, 2020.

[16] Noel Cressie and Gardar Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, 2008.

[17] Jyotishka Datta, Jayanta K Ghosh, et al. Asymptotic properties of bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132, 2013.

[18] Tracy Qi Dong and Jon Wakefield. Modeling and presentation of vaccination coverage estimates using data from household surveys. *arXiv preprint arXiv:2004.03127*, 2020.

[19] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[20] Jean Feng and Noah Simon. Gradient-based regularization parameter selection for problems with nonsmooth penalty functions. *Journal of Computational and Graphical Statistics*, 27(2):426–435, 2018.

[21] Eric W Fox, Jay M Ver Hoef, and Anthony R Olsen. Comparing spatial regression to random forests for large environmental data sets. *arXiv preprint arXiv:1812.10236*, 2018.

[22] Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

[23] Florian Gerber and Douglas W Nychka. Parallel cross-validation: a scalable fitting method for gaussian process models. *arXiv preprint arXiv:1912.13132*, 2019.

[24] Anjum Hajat, Ana V Diez-Roux, Sara D Adar, Amy H Auchincloss, Gina S Lovasi, Marie S O'Neill, Lianne Sheppard, and Joel D Kaufman. Air pollution and individual and neighborhood socioeconomic status: evidence from the multi-ethnic study of atherosclerosis (mesa). *Environmental health perspectives*, 121(11-12):1325–1333, 2013.

[25] Anjum Hajat, Ana Diez Roux, Cecilia Castro-Diehl, Kristen Cosselman, Sherita Hill Golden, Adam Szpiro, Sverre Vedal, and Joel D Kaufman. The association between air pollution and stress hormones: Evidence from the multi-ethnic study of atherosclerosis. In *ISEE Conference Abstracts*, 2015.

[26] David A Harville. Matrix algebra from a statistician's perspective, 1998.

[27] 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.

[28] Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.

[29] Matthew J Heaton, Abhirup Datta, Andrew O Finley, Reinhard Furrer, Joseph Guinness, Rajarshi Guhaniyogi, Florian Gerber, Robert B Gramacy, Dorit Hammerling, Matthias Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019.

[30] Travis Hee Wai, Michael T Young, and Adam A Szpiro. Random spatial forests. *arXiv preprint arXiv:2006.00150*, 2020.

[31] Tomislav Hengl, Madlene Nussbaum, Marvin N Wright, Gerard BM Heuvelink, and Benedikt Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, 2018.

[32] James S Hodges. *Richly parameterized linear models: additive, time series, and spatial models using random effects*. Chapman and Hall/CRC, 2016.

[33] Gerard Hoek, Rob Beelen, Kees De Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment*, 42(33):7561–7578, 2008.

[34] Giles Hooker and Lucas Mentch. Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

[35] Xuefei Hu, Jessica H Belle, Xia Meng, Avani Wildani, Lance A Waller, Matthew J Strickland, and Yang Liu. Estimating pm2. 5 concentrations in the conterminous united states using the random forest approach. *Environmental science & technology*, 51(12):6936–6944, 2017.

[36] Hsin-Cheng Huang, Nan-Jung Hsu, David M Theobald, and F Jay Breidt. Spatial lasso with applications to gis model selection. *Journal of Computational and Graphical Statistics*, 19(4):963–983, 2010.

[37] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[38] Marcin Jurek and Matthias Katzfuss. Multi-resolution filters for massive spatio-temporal data. *arXiv preprint arXiv:1810.04200*, 2018.

[39] EE Kammann and Matthew P Wand. Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):1–18, 2003.

[40] Matthias Katzfuss and Noel Cressie. Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics*, 23(1):94–107, 2012.

[41] Cari G Kaufman, Mark J Schervish, and Douglas W Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.

[42] CG Kaufman and Benjamin Adam Shaby. The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484, 2013.

[43] Joel D Kaufman, Sara Adar, R Graham Barr, Matthew Budoff, Gregory L Burke, Cynthia L Curl, Martha Daviglus, Ana V Diez-Roux, Amanda Gassett, David R Jacobs, et al. Air pollution and progression of coronary artery calcification: Mesa air. In *ISEE Conference Abstracts*, 2014.

[44] Jason Kilian and Masashi Kitazawa. The emerging risk of exposure to air pollution on cognitive decline and alzheimer's disease–evidence from epidemiological and animal studies. *Biomedical journal*, 41(3):141–162, 2018.

[45] Peter J Leary, RG Barr, David A Bluemke, Catherine L Hough, Joel D Kaufman, Adam A Szpiro, Steven M Kawut, and Victor C Van Hee. The relationship of roadway proximity and nox with right ventricular structure and function: the mesa-right ventricle and mesa-air studies. In *C28. RIGHT VENTRICULAR PATHOBIOLOGY*, pages A3976–A3976. American Thoracic Society, 2013.

[46] Peter J Leary, Joel D Kaufman, R Graham Barr, David A Bluemke, Cynthia L Curl, Catherine L Hough, Joao A Lima, Adam A Szpiro, Victor C Van Hee, and Steven M Kawut. Traffic-related air pollution and the right ventricle. the multi-ethnic study of atherosclerosis. *American journal of respiratory and critical care medicine*, 189(9):1093–1100, 2014.

[47] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[48] Tianxi Li, Elizaveta Levina, Ji Zhu, et al. Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132–164, 2019.

[49] Johan Lindström, Adam A Szpiro, Paul D Sampson, Assaf P Oron, Mark Richards, Tim V Larson, and Lianne Sheppard. A flexible spatio-temporal model for air pollution

with spatial and spatio-temporal covariates. *Environmental and ecological statistics*, 21(3):411–433, 2014.

[50] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[51] Ying Liu, Guofeng Cao, Naizhuo Zhao, Kevin Mulligan, and Xinyue Ye. Improve ground-level pm2. 5 concentration mapping using a random forests-based geostatistical approach. *Environmental pollution*, 235:272–282, 2018.

[52] Jaime Madrigano, Sara D Adar, Joseph Schwartz, Eric A Hoffman, Paul Enright, John Austin, Adam Szpiro, Karen Hinckley Stukovsky, Patrick Kinney, Sverre Vedal, et al. Ambient fine particulate matter (pm2. 5) exposure and longitudinal change in percent emphysema on computed tomography (ct). the multi-ethnic study of atherosclerosis (mesa) lung and air pollution studies. In *B15. AIR POLLUTION IN WOMEN, CHILDREN, AND ELDERLY: RISK SUSCEPTIBILITY AND REDUCTION*, pages A2438–A2438. American Thoracic Society, 2014.

[53] Giampiero Marra and Simon N Wood. Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387, 2011.

[54] Georges Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.

[55] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

[56] Laina D Mercer, Adam A Szpiro, Lianne Sheppard, Johan Lindström, Sara D Adar, Ryan W Allen, Edward L Avol, Assaf P Oron, Timothy Larson, L-J Sally Liu, et al. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (nox) for the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Atmospheric Environment*, 45(26):4412–4420, 2011.

[57] Kyle P Messier and Matthias Katzfuss. Scalable penalized spatiotemporal land-use regression for ground-level nitrogen dioxide. *arXiv preprint arXiv:2005.09210*, 2020.

[58] Eric V Novotny, Matthew J Bechle, Dylan B Millet, and Julian D Marshall. National satellite-based land-use regression: No2 in the united states. *Environmental science & technology*, 45(10):4407–4414, 2011.

[59] Douglas Nychka, Christopher Wikle, and J Andrew Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331, 2002.

[60] Douglas W Nychka. Spatial-process estimates as smoothers. *Smoothing and regression: approaches, computation, and application*, 329:393, 2000.

[61] Casey Olives, Lianne Sheppard, Johan Lindström, Paul D Sampson, Joel D Kaufman, and Adam A Szpiro. Reduced-rank spatio-temporal modeling of air pollution concentrations in the multi-ethnic study of atherosclerosis and air pollution. *The annals of applied statistics*, 8(4):2509, 2014.

[62] Hector A Olvera, Mario Garcia, Wen-Whai Li, Hongling Yang, Maria A Amaya, Orrin Myers, Scott W Burchiel, Marianne Berwick, and Nicholas E Pingitore Jr. Principal component analysis optimization of a pm2. 5 land use regression model with small monitoring network. *Science of the total environment*, 425:27–34, 2012.

[63] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[64] Collin A Politsch, Jessi Cisewski-Kehe, Rupert AC Croft, and Larry Wasserman. Trend filtering–i. a modern statistical tool for time-domain astronomy and astronomical spectroscopy. *Monthly Notices of the Royal Astronomical Society*, 492(3):4005–4018, 2020.

[65] Timothy W Randolph, Sen Zhao, Wade Copeland, Meredith Hullar, and Ali Shojaie. Kernel-penalized regression for analysis of microbiome data. *The annals of applied statistics*, 12(1):540, 2018.

[66] Paulo J Ribeiro Jr, Peter J Diggle, Maintainer Paulo J Ribeiro Jr, and MASS Suggests. The geor package. *R news*, 1(2):14–18, 2007.

[67] George K Robinson et al. That blup is a good thing: the estimation of random effects. *Statistical science*, 6(1):15–32, 1991.

[68] Esther Rolf, Michael I Jordan, and Benjamin Recht. Post-estimation smoothing: A simple baseline for learning with side information. *arXiv preprint arXiv:2003.05955*, 2020.

[69] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

[70] Paul D Sampson, Mark Richards, Adam A Szpiro, Silas Bergen, Lianne Sheppard, Timothy V Larson, and Joel D Kaufman. A regionalized national universal kriging model using partial least squares regression for estimating annual pm2. 5 concentrations in epidemiology. *Atmospheric environment*, 75:383–392, 2013.

[71] Apolline Saucy, Martin Röösli, Nino Künzli, Ming-Yi Tsai, Chloé Sieber, Toyib Olaniyan, Roslynn Baatjies, Mohamed Jeebhay, Mark Davey, Benjamin Flückiger, et al. Land use regression modelling of outdoor no2 and pm2. 5 concentrations in three low income areas in the western cape province, south africa. *International journal of environmental research and public health*, 15(7):1452, 2018.

[72] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

[73] Arjun Sondhi. *Statistical miscellany: causality, networks, and bandits*. PhD thesis, 2019.

[74] Michael L Stein. Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19, 2014.

[75] Michael L Stein et al. The screening effect in kriging. *The Annals of Statistics*, 30(1):298–323, 2002.

[76] Adam A Szpiro, Paul D Sampson, Lianne Sheppard, Thomas Lumley, Sara D Adar, and Joel D Kaufman. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*, 21(6):606–631, 2010.

[77] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[78] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[79] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[80] Grace Wahba. *Spline models for observational data*. SIAM, 1990.

[81] Jon Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183, 2007.

[82] Meng Wang, Carrie Pistenmaa Aaron, Jaime Madrigano, Eric A Hoffman, Elsa Angelini, Jie Yang, Andrew Laine, Thomas M Vetterli, Patrick L Kinney, Paul D Sampson, et al. Association between long-term exposure to ambient air pollution and change in quantitatively assessed emphysema and lung function. *Jama*, 322(6):546–556, 2019.

[83] Jennifer Weuve, Joel D Kaufman, Adam A Szpiro, Cynthia Curl, Robin C Puett, Todd Beck, Denis A Evans, and Carlos F Mendes de Leon. Exposure to traffic-related air pollution in relation to progression in physical disability among older adults. *Environmental health perspectives*, 124(7):1000–1008, 2016.

[84] Brian Williamson and Jean Feng. Efficient nonparametric statistical inference on population feature importance using shapley values. In *International Conference on Machine Learning*, pages 10282–10291. PMLR, 2020.

[85] Brian D Williamson, Peter B Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 2019.

[86] Hao Xu, Matthew J Bechle, Meng Wang, Adam A Szpiro, Sverre Vedal, Yuqi Bai, and Julian D Marshall. National pm2.5 and no2 exposure models for china based on land use regression, satellite measurements, and universal kriging. *arXiv preprint arXiv:1808.09126*, 2018.

[87] Yuhong Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[88] Michael T Young, Matthew J Bechle, Paul D Sampson, Adam A Szpiro, Julian D Marshall, Lianne Sheppard, and Joel D Kaufman. Satellite-based no2 and model validation in a national prediction model based on universal kriging and land-use regression. *Environmental science & technology*, 50(7):3686–3694, 2016.

[89] Kai Zhang, Timothy V Larson, Amanda Gassett, Adam A Szpiro, Martha Daviglus, Gregory L Burke, Joel D Kaufman, and Sara D Adar. Characterizing spatial patterns of airborne coarse particulate (pm10–2.5) mass and chemical components in three cities: the multi-ethnic study of atherosclerosis. *Environmental health perspectives*, 122(8):823–830, 2014.

[90] Sen Zhao and Ali Shojaie. A significance test for graph-constrained estimation. *Biometrics*, 72(2):484–493, 2016.

[91] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Appendix A

APPENDIX

A.1 Proof of Equivalence for Update of Ω^{k+1}

Lemma A.1.1 *Blockwise Inversion, [26] Thm 8.5.11*

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \\ -\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & \left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \end{bmatrix}$$

Assume that only $\mathbf{\Omega}^k$ is known. For each new possible split $\mathbf{C}^A$, finding the additional change in loss depends only on

$$\mathbf{\Omega}^{k+1} = \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{C}^{k+1}\left(\left(\mathbf{C}^{k+1}\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^{k+1} + \mathbf{L}^{k+1}\right)^{-1}\left(\mathbf{C}^{k+1}\right)^T\mathbf{\Sigma}^{-1}$$

First, we examine the inverse $\left(\left(\mathbf{C}^{k+1}\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^{k+1} + \mathbf{L}^{k+1}\right)^{-1}$. Since $\mathbf{C}^{k+1} = \begin{bmatrix} \mathbf{C}^k & \mathbf{C}^A \end{bmatrix}$, and $\mathbf{L}^{k+1} = \begin{bmatrix} \mathbf{L}^k & \mathbf{0} \\ \mathbf{0} & \mathbf{L}^A \end{bmatrix}$,

$$\left(\left(\mathbf{C}^{k+1}\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^{k+1} + \mathbf{L}^{k+1}\right)^{-1} = \begin{bmatrix} \left(\mathbf{C}^k\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^k\mathbf{L}^k & \left(\mathbf{C}^k\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^A \\ \left(\mathbf{c}^A\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^k & \left(\mathbf{C}^A\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^A\mathbf{L}^A \end{bmatrix}^{-1}$$

By Lemma A.1.1,

$$\begin{bmatrix} \left(\mathbf{C}^k\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^k + \mathbf{L}^k & \left(\mathbf{C}^k\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^A \\ \left(\mathbf{C}^A\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^k & \left(\mathbf{C}^A\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^A + \mathbf{L}^A \end{bmatrix}^{-1} = \begin{bmatrix} \left(\left(\mathbf{C}^k\right)^T\mathbf{\Sigma}^{-1}\mathbf{C}^k + \mathbf{L}^k\right)^{-1} + \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}$$

where,

$$\mathbf{M}_{11} = \left(\left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k + \mathbf{L}^k\right)^{-1} \left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^A\mathbf{M}_{22} \left(\mathbf{C}^A\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k \left(\left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k + +\mathbf{L}^k\right)^{-1}$$

$$\mathbf{M}_{12} = -\left(\left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k + \mathbf{L}^k\right)^{-1} \left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^A\mathbf{M}_{22}$$

$$\mathbf{M}_{21} = -\mathbf{M}_{22} \left(\mathbf{C}^A\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k \left(\left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k + +\mathbf{L}^k\right)^{-1}$$

$$\begin{aligned}\mathbf{M}_{22} &= \left(\left(\mathbf{C}^A\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^A - \left(\mathbf{C}^A\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k \left(\left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k + \mathbf{L}^k\right)^{-1} \left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^A\right)^{-1} \\ &= \left(\left(\mathbf{C}^A\right)^T \mathbf{\Omega}^k\mathbf{C}^A\right)^{-1}\end{aligned}$$

Now,

$$\begin{aligned}\mathbf{\Omega}^{k+1} &= \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\begin{bmatrix}\mathbf{C}^k & \mathbf{C}^A\end{bmatrix}\begin{bmatrix}\left(\left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k + \mathbf{L}^k\right)^{-1} + \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{C}^k & \mathbf{C}^A\end{bmatrix}^T \mathbf{\Sigma}^{-1} \\ &= \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{C}^k \left(\left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1}\mathbf{C}^k + +\mathbf{L}^k\right)^{-1} \left(\mathbf{C}^k\right)^T \mathbf{\Sigma}^{-1} \\ &\quad - \mathbf{\Sigma}^{-1}\left(\mathbf{C}^k\mathbf{M}_{11}\left(\mathbf{C}^k\right)^T + \mathbf{C}^A\mathbf{M}_{21}\left(\mathbf{C}^k\right)^T + \mathbf{C}^k\mathbf{M}_{12}\left(\mathbf{C}^A\right)^T + \mathbf{C}^A\mathbf{M}_{22}\left(\mathbf{C}^A\right)^T\right)\mathbf{\Sigma}^{-1} \\ &= \mathbf{\Omega}^k - \mathbf{\Omega}^k\mathbf{C}^A\left(\left(\mathbf{C}^A\right)^T \mathbf{\Omega}^k\mathbf{C}^A + \mathbf{L}^A\right)^{-1} \left(\mathbf{C}^A\right)^T \mathbf{\Omega}^k\end{aligned}$$