Title
On Model Determination, Prediction and Statistical Learning: The Case of Space-Time Data

Author
Watson, Gregory Luke

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

On Model Determination, Prediction and Statistical Learning: The Case of Space-Time Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Gregory Luke Watson

2021

ABSTRACT OF THE DISSERTATION

On Model Determination,
Prediction and Statistical Learning:
The Case of Space-Time Data

by

Gregory Luke Watson
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2021
Professor Donatello Telesca, Chair

Problems of model determination, prediction and statistical learning for space-time data arise in many fields. Evaluation metrics developed for independent data are overly optimistic in the dependent context of space-time data and are unreliable for model selection, tuning and averaging. This dissertation makes three contributions to the field of space-time prediction with applications in air pollution exposure modeling. First, it formalizes the prediction error associated with the spatial interpolation of space-time data and investigates a variety of cross-validation (CV) procedures for estimating that error. Consistent with recent best practice, location-based CV is shown to be appropriate for estimating spatial interpolation error as in our analysis of California wildfire data. Interestingly, commonly held notions of bias-variance trade-off of CV fold size do not trivially apply to dependent data, and we recommend leave-one-location-out (LOLO) CV as the preferred prediction error metric for spatial interpolation.

Second, this evaluation framework was applied to compare the predictive accuracy of ten machine learning algorithms on the spatial interpolation of maximum daily 8-hour average ozone during California wildfires in 2008, the first such analysis of ozone during a wildfire event. Gradient boosting and random forest, both ensembles of tree-based models, were the

best performing models, having the lowest LOLO CV estimates of prediction error.

Third, it introduces treeging, a new machine learning algorithm for space-time prediction that enriches the flexible mean structure of regression trees with a kriging covariance model into the base learner of an ensemble algorithm. The combination of a flexible mean structure with the dependence structure of a traditional spatial statistics model yields a prediction tool that performs well across a variety of simulation scenarios and several case studies.

The dissertation of Gregory Luke Watson is approved.

Sudipto Banerjee

Michael Jerrett

Christina M. Ramirez

Donatello Telesca, Committee Chair

University of California, Los Angeles

2021

To Jennifer, Mirabelle and my parents.

I love you all very much.

Table of Contents

List of Figures . . . . . **ix**

List of Tables . . . . . **x**

Acknowledgments . . . . . **xi**

Vita . . . . . **xiii**

1 Introduction **1**

1.1 Space-Time Data 2

1.2 Space-Time Prediction 2

1.2.1 Dependence Models 3

1.2.2 Machine Learning 4

1.2.3 Model Evaluation 5

1.3 Air Pollution Epidemiology 5

1.4 Overview 7

2 Prediction & Model Evaluation for Space-Time Data **8**

2.1 Introduction 8

2.2 Motivating Application 10

2.3 Prediction Error: Estimands and Estimators 11

2.3.1 The Case of Independent Replication 11

2.3.2 Prediction Error Estimands for Space-Time Data 13

2.3.3 Prediction Error Estimators for Space-Time Data 14

2.4 Simulation Study 17

2.4.1 Simulation Results & Discussion . . . 18
2.5 Case Study . . . 24
2.5.1 Case Study Results & Discussion . . . 26
2.6 Discussion . . . 28
2.A Appendix . . . 31
2.A.1 Case Study Covariates . . . 31
2.A.2 Simulation Procedure . . . 32
2.A.3 Simulation Parameters . . . 33
2.A.4 Conditional Variance . . . 33

3 Machine Learning Models Accurately Predict Ozone Exposure during Wildfire Events . . . 34
3.1 Introduction . . . 35
3.2 Materials and Methods . . . 40
3.2.1 Data . . . 40
3.2.2 Statistical Analysis . . . 42
3.3 Results and Discussion . . . 45
3.A Appendix . . . 54
3.A.1 WRF-Chem Simulation Inputs . . . 54
3.A.2 CV Estimates of RMSE and R^2 . . . 54
3.A.3 Tuning Parameters . . . 56

4 Treeging . . . 58
4.1 Introduction . . . 58
4.2 Treeging . . . 61
4.2.1 Mean Structure . . . 61

4.2.2 Covariance Estimation . . . 62
4.2.3 Treeging Algorithm . . . 63
4.3 Spatial Treeging . . . 63
4.3.1 Spatial Simulation Study . . . 64
4.3.2 Spatial Case Studies . . . 67
4.4 Space-Time Treeging . . . 68
4.4.1 Space-Time Simulation Study . . . 69
4.4.2 Space-Time Case Studies . . . 71
4.5 Parameter Tuning . . . 72
4.6 Large Data Set Approximations . . . 75
4.7 Discussion . . . 75
4.A Appendix . . . 77
4.A.1 Spatial Simulations . . . 77
4.A.2 Case Study Covariates . . . 78
4.A.3 Spherical Covariance Estimation . . . 79
4.A.4 Space-Time Simulations . . . 79

5 Discussion . . . 81

References . . . 84

List of Figures

2.1 Prediction error and its estimators, simulation study 19

2.2 Prediction error estimators plotted against true error, simulation study 21

2.3 Prediction error as a function of conditional variance, simulation study 22

2.4 Prediction error as a function of training data size, case study 25

2.5 Estimated prediction error by week, case study 27

3.1 Maximum daily 8-hour average ozone 41

3.2 10-fold and LOLO CV estimates of RMSE and R^2 45

3.3 LOLO CV prediction heat maps, gradient boosting & random forest 46

3.4 10-fold & LOLO CV predicted vs observed, gradient boosting & random forest . 48

3.5 Variable importance, gradient boosting vs. random forest 51

3.6 Extrapolation error, gradient boosting & random forest 52

3.7 Nearest neighbor model tuning 56

3.8 GAM spline basis tuning 57

4.1 Spatial simulation model comparison 66

4.2 Spatial case studies model comparison 68

4.3 Space-time simulation model comparison 70

4.4 Space-time case studies model comparison 72

4.5 Tuning the number of treeges 73

4.6 Tuning the subsample proportion 74

List of Tables

3.1 Covariates for ozone prediction 39

3.2 CV comparison of ozone prediction models 55

4.1 Mean and dependence structures 62

Acknowledgments

I am immensely grateful to the many professors, colleagues, friends and family who have supported and assisted me in the work that has culminated in this dissertation. My advisor and committee chair, Professor Donatello Telesca, has been an inexhaustible source of statistical knowledge, sound advice, good ideas and patience. I am so very fortunate and grateful that Professors Christina Ramirez, Michael Jerrett and Sudipto Banerjee have given so generously of their time in serving on my committee.

I am also thankful for the opportunity to have collaborated with Professor Ramirez on several COVID-19 related projects and with Professor Jerrett on a number of studies of wildfire air pollution. Professors Thomas Belin and William Cumberland have also provided valuable advice and assistance during my time at UCLA.

I would also like to thank my colleagues on the California Simulation of Insurance Markets (CalSIM) team at the UCLA Center for Health Policy Research and the UC Berkeley Labor Center. I am especially grateful for the opportunity to collaborate with Dr. Xiao Chen and Professors Dylan Roby and Jerry Kominski.

I gratefully acknowledge the many sources of funding that have supported this work and made my doctoral studies possible, including the UCLA Department of Biostatistics under the leadership of Professors William Cumberland and Sudipto Banerjee, the Raymond D. Goodman Scholarship Fund, the UCLA Graduate Summer Research Mentorship Program and the Graduate Student Researcher positions provided by Professors Telesca, Jerrett and the CalSIM team. I also acknowledge the grant from the Bureau of Land Management [grant number L14AC00173] that partially funded the work in chapter three.

I am grateful for the research contributions of Professors Telesca and Jerrett as well as Professor Colleen Reid of the University of Colorado Boulder and Dr. Gabriele Pfister at the National Center for Atmospheric Research to the work presented in this dissertation. Chapter three of this dissertation was previously published in *Environmental Pollution* (Watson et al., 2019), chapter two was posted as an arXiv preprint (Watson et al., 2020b) and is

under consideration for peer-review publication, and chapter four is in preparation for peer review submission.

Finally, I would like to express my profound and utterly inadequate gratitude to my family for the sacrifices they have made and support they have given me, without which I would have been unable to complete this work, including my parents, my brothers, and most especially my wonderful spouse, Jennifer, and darling daughter, Mirabelle.

Vita

2000 – 2004	B.S. in Biochemistry, Mathematics, University of Notre Dame, Notre Dame, Indiana
2007 – 2009	M.T.S. in Biblical Studies, Emory University, Atlanta, Georgia
2009 – 2011	M.S. in Biostatistics, University of California, Los Angeles, California
2011 – 2013	Research Analyst, UCLA Center for Health Policy Research
2013 – Present	Graduate Student Researcher, UCLA Center for Health Policy Research

Publications

Gregory L. Watson, Colleen E. Reid, Michael Jerrett, and Donatello Telesca, "Treeging," (2021). In preparation.

Gregory L. Watson, Di Xiong, Lu Zhang, Joseph A. Zoller, John Shamshoian, Phillip Sundin, Teresa Bufford, Anne W. Rimoin, Marc A. Suchard, and Christina M. Ramirez, "Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model," *PLoS Computational Biology* 17 (2021).

Gregory L. Watson, Colleen E. Reid, Michael Jerrett, and Donatello Telesca, "Prediction & Model Evaluation for Space-Time Data," (2020), arXiv:2012.13867. Under consideration, *Annals of Applied Statistics*.

Gregory L. Watson, Donatello Telesca, Colleen E. Reid, Gabriele G. Pfister, and Michael Jerrett, "Machine learning models accurately predict ozone exposure during wildfire events," *Environmental Pollution* 254 (2019).

CHAPTER 1

Introduction

Problems of model determination, prediction and statistical learning for space-time data arise in many fields. Space-time data by definition are indexed by spatial location and time, and a nearly universal feature of such data is dependence across the space-time domain. This dependence is of profound importance for the application of statistical learning procedures for two reasons. First, many statistical learning techniques assume data are independent (conditional on any covariates). Violations of this independence assumption may render models inefficient, and worse can render model evaluation procedures and estimators of prediction error inconsistent and therefore inappropriate for use on space-time data. Second, dependence makes prediction ambiguous as a concept, because different relationships between the space-time index of the training data and the data to be predicted are possible. It is critical in this context to clearly articulate the type of prediction desired for a particular application and select an evaluation procedure that targets that prediction error.

The contributions of this dissertation are motivated by applications in air pollution exposure modelling. Air pollution is a major threat to human health, and epidemiological studies of its health effects rely on predictions of air pollution concentrations when and where they are not observed. Machine learning algorithms have become an attractive alternative to traditional statistical prediction models for this predictive task. However, understanding how well they perform and comparing them to traditional models requires the development of appropriate model evaluation metrics. This dissertation introduces such a framework and compares the performance of a number of prediction tools using simulations and air pollution data collected during California wildfires. It also proposes a new machine learning algorithm for space-time prediction that combines the strengths of traditional statistical models with

those of the best performing machine learning algorithms.

1.1 Space-Time Data

Space-time data are indexed by spatial location and time. This indexing is important, because observations of many space-time phenomena exhibit dependence across space and time. Quite often, for example, observations taken close together in space or time tend to be more similar than those taken farther apart. As a consequence, knowledge of the value of an observable at one space-time location imparts information on the value at another location. In contrast, observations that are independent supply no information on each other. The amount of information provided by another observation is the strength of the dependence.

This dependence generally persists after the effects of other related quantities (i.e., covariates) have been taken into account. In many contexts, observations are marginally dependent, but are reasonably assumed to be independent conditional on the covariates upon which they depend. Space-time data are usually not independent conditional on covariates, a critical distinction from many other types of data. Unless otherwise noted, throughout this document when data, random variables or stochastic processes are described as dependent or independent data, we mean conditional dependence or independence given any relevant covariates.

1.2 Space-Time Prediction

Various types of prediction may be desired in the context of space-time data. These include replication, imputation, forecasting, spatial extrapolation and interpolation. Each type of prediction is defined by a different relationship between the prediction cases and the observed training data. For example, forecasting is the prediction of cases at times subsequent to those at which the training data were observed. Replication is the prediction of observations from an entirely new realization of the space-time field. For independent data, in contrast, only one type of relationship exists between the training data and the prediction cases:

independence. We focus primarily on spatial interpolation within the space-time domain motivated by wildfire air pollution data. Repeated measurements of air pollution were taken by stationary monitors. The predictive (interpolation) goal is to predict the air pollution time series at unmonitored locations within the spatial domain of the observed data.

In addition to observations of the quantity of interest, there is often ancillary data available which can be exploited for prediction, e.g., as covariates in a regression approach. To be useful in this manner, covariates must be observed at both the prediction and training locations. It may be difficult to observe covariates at prediction locations in some contexts, e.g., forecasting, but for interpolation they are often available. Land use, meteorological data, satellite retrievals, and the output of atmospheric chemistry numerical simulation models are some of the covariates that may be useful for interpolating air pollution data.

1.2.1 Dependence Models

Traditional approaches to interpolation of spatial or space-time data have focused on the dependence between observations (Cressie, 2015a). The most common of these is kriging, named after South African engineer Danie Krige, who published the method for estimating ore grades at gold mines in 1951 (Krige, 1951; Matheron, 1963). Kriging is the best linear unbiased prediction (BLUP) given a covariance model that describes the dependence between observations and the prediction case. It is a special case of the BLUP for generalized least squares (Goldberger, 1962) and is the maximum a priori (MAP) prediction of a Gaussian process (GP) model with suitably chosen priors (Banerjee et al., 2014). Its strength lies in its ability to exploit the spatial or space-time dependence between observations to predict at unobserved locations. The complexity of the dependence models used for kriging varies tremendously from very simple covariance functions assuming isotropy to quite complicated models.

While kriging is the most commonly used method of spatial interpolation, many other approaches have been proposed. Thiessen triangulation, an example of Voronoi tesselation, partitions a spatial region into polygons so that every point in the spatial domain is assigned

to the polygon circumscribing the observed data point to which it is closest (Okabe et al., 2000). Another type of interpolating model is inverse distance weighting, in which the value of the process at at a point is predicted as a weighted sum of the other observations. The weights are assigned proportional to the inverse distance of an observation from the point at which prediction is desired (Lu and Wong, 2008). Nearby observations thus receive higher weight than those farther apart. Thin-plate splines interpolate across the spatial domain with smooth functions (Hutchinson and Gessler, 1994). One advantage of kriging over these approaches is the probability model it supplies for estimating prediction error (Jerrett et al., 2005).

An important drawback to kriging and most Bayesian GP models is the linear mean structure. More flexible models have been proposed to allow nonlinear effects. These include geographically weighted regression (Brunsdon et al., 1996), thin-plate splines (Paciorek et al., 2009), varying coefficient models (Berrocal et al., 2010; Szpiro et al., 2010) and treed Gaussian process (Gramacy and Lee, 2008). These approaches have conceptual appeal, but each has drawbacks. The principal difficulty is the lack of available software that can accommodate space-time data sets of the size used in epidemiological studies of air pollution.

1.2.2 Machine Learning

More recently, machine learning algorithms with flexible mean structures, have become popular for spatial and space-time prediction. Their popularity is no doubt motivated by the well known successes of machine learning algorithms on a variety of prediction problems. These models generally assume observations are independent of each other (conditional on the covariates) because they were developed for use on data where this is a reasonable assumption. This assumption is obviously wrong on most spatial and space-time data sets, but these algorithms still perform very well. Many different techniques have been employed for air pollution prediction models, including generalized additive models (Liu et al., 2009), support vector machines (Weizhen et al., 2014), gradient boosting (Zhan et al., 2017), random forest (Hu et al., 2017), neural networks (Di et al., 2016, 2017), and generative adversarial

networks (Gao et al., 2020). Boosting and random forest made the best predictions of particulate matter in two comparison of machine learning algorithms (Reid et al., 2015; Watson et al., 2019). Both are ensembles of trees, each of which models the data using random splits of the covariate space. This provides a very flexible mean structure and naturally accounts for interactions between covariates and model selection.

1.2.3 Model Evaluation

Estimating prediction error is critical for comparing, tuning and averaging models. However, assessing model predictive performance is not straightforward in the context of space-time data. For independent data, cross-validation (Stone, 1974) and bootstrap (Efron, 1979) provide powerful nonparametric estimators of prediction error. The naive application of these techniques to dependent data results in overly optimistic estimates (Roberts et al., 2017; Watson et al., 2020b). When the data can be partitioned into groups that are independent of other groups, this bias can be overcome by choosing the group as the sampling unit. Spatial and space-time data, however, cannot be partitioned into independent units. Each type of space-time prediction (see section 1.2) corresponds to a particular definition of prediction error. Only by clearly conceptualizing the predictive task and selecting an estimator that targets the appropriate measure of prediction error is it possible to evaluate space-time prediction models.

1.3 Air Pollution Epidemiology

Air pollution is a major threat to human health, animal health and the environment. Millions of people die each year from airborne pollutants (Chuang et al., 2011; Kim et al., 2015). Among the most toxic airborne pollutants are particulate matter and ground-level ozone. Particulate matter (PM) is simply liquid or solid particles suspended in the air (World Health Organization, 2013) and is the most dangerous air pollutant to human health (Kim et al., 2015). There are many natural sources of PM, including volcanoes, wild fires and dust

storms (Misra et al., 2001). The primary anthropogenic sources are fuel combustion and wearing of roadways and vehicles from motor vehicle traffic (De Kok et al., 2006).

The toxicity of PM depends largely upon the size of the particles, because that dictates how far they can penetrate into the lungs of living organisms (Esworthy, 2013). Consequently particles are usually classified based on their diameter as coarse PM (diameter less than 10 µm), i.e., PM_{10} or fine PM (diameter less than 2.5 µm), i.e., $PM_{2.5}$ (Atkinson et al., 2010). Particulate matter has been linked with increased mortality (Samoli et al., 2008; Laden et al., 2000), a host of pulmonary conditions including decreased lung function (Brauer et al., 2012), shortness of breath (Guaita et al., 2011), exacorbated asthma (Reid et al., 2019; Cadelis et al., 2014; McCormack et al., 2011; Silverman and Ito, 2010). PM has also been linked with increased risk of cardiovascular disease hospitalization and mortality (Brook et al., 2010, 2004), heart attack (Dominici et al., 2006), cardiac arrithmyia (Kim et al., 2012; Zanobetti et al., 2013) and diabetes (Pearson et al., 2010). It has also been associated with increased school absences in children (Ransom and Pope, 2013; Marcon et al., 2014) and negative economic impact (Hou et al., 2012).

Ozone at ground level is also toxic to human and animal health (Bell et al., 2004; Stocker et al., 2013). It has been linked with increased mortality (Bell et al., 2004), decreased respiratory function, exacerbation of chronic obstructive pulmonary disease (COPD), bronchitis, emphysema, and asthma (Reid et al., 2019; Silverman and Ito, 2010; Guarnieri and Balmes, 2014; Lee et al., 2014; National Research Council, 2008). The effects of long term exposure are less well understood, but likely include increased respiratory and cardiopulmonary mortality (Khaniabadi et al., 2017; Crouse et al., 2015; Jerrett et al., 2009; Turner et al., 2016; Sousa et al., 2013), decreased lung function (Hwang et al., 2015), the progression of emphysema (Barr et al., 2016) and increased susceptibility to influenza (Kesic et al., 2012).

Inferring the long-term health effects of air pollution is difficult, but critical to ameliorating its negative consequences. Experimental studies are unethical, and so studies are necessarily observational. The goal is typically to infer the effect of exposure to a pollutant on health outcomes using a regression analysis. Population health outcome data may be

available from administrative data, but the amounts of pollutants to which individuals have been exposed is usually unknown. Information on pollutant exposure is usually restricted to measurements made by sparsely distributed air-pollution monitors and numerical simulations that attempt to model atmospheric processes computationally, usually reporting their output over a regular grid (Jerrett et al., 2005).

Epidemiological analyses of air pollution require estimates of pollution concentration at the locations at which health outcome data are available (Ryan and LeMasters, 2007). These locations generally do not coincide with the locations of the monitoring network or the simulation grid. This difference of spatial domains is referred to as spatial misalignment (Gryparis et al., 2008). Quantifying pollution concentrations at locations where it is not observed is challenging on account of its small-scale spatial variability (Hoek et al., 2008). To reconcile this misalignment between the locations at which pollution and health outcome data are available, researchers often construct a pollution surface over the entire spatial domain of interest, which includes the spatial domain of the health outcome data.

1.4 Overview

This dissertation makes three contributions to the field of space-time prediction with applications in air pollution epidemiology. Chapter two formalizes the prediction error associated with spatially interpolating space-time data and investigates a variety of cross-validation (CV) procedures for estimating that error. Illustrating these techniques on California wildfire data, it identifies leave-one-location-out (LOLO) CV as the preferred prediction error metric for spatial interpolation. Chapter three applies the LOLO CV framework to evaluating a set of ten machine learning algorithms for the spatial interpolation of ozone data recorded during California wildfires in 2008. Gradient boosting and random forest were the best performing models. Chapter four introduces treeging, a new machine learning algorithm for space-time prediction that enriches the flexible mean structure of regression trees with a kriging covariance model into the base learner of an ensemble algorithm. Finally, chapter five closes with a discussion.

CHAPTER 2

Prediction & Model Evaluation for Space-Time Data

Evaluation metrics for prediction error, model selection and model averaging on space-time data are understudied and poorly understood. The absence of independent replication makes prediction ambiguous as a concept and renders evaluation procedures developed for independent data inappropriate for most space-time prediction problems. Motivated by air pollution data collected during California wildfires in 2008, this manuscript attempts a formalization of the true prediction error associated with spatial interpolation. We investigate a variety of cross-validation (CV) procedures employing both simulations and case studies to provide insight into the nature of the estimand targeted by alternative data partition strategies. Consistent with recent best practice, we find that location-based cross-validation is appropriate for estimating spatial interpolation error as in our analysis of the California wildfire data. Interestingly, commonly held notions of bias-variance trade-off of CV fold size do not trivially apply to dependent data, and we recommend leave-one-location-out (LOLO) CV as the preferred prediction error metric for spatial interpolation.

2.1 Introduction

Space-time interpolation is an important task in air pollution epidemiology for estimating exposure over a continuous space-time domain within which pollutants have been observed at discrete locations and times. This task is often cast as an exercise in prediction, but in certain contexts goes by other names, including downscaling, data fusion, land use regression and infill prediction. A wide variety of deterministic, statistical and machine learning models have been used to interpolate spatial or space-time data, indicative of the ongoing and diverse

interest in the problem (Van Donkelaar et al., 2015; Reid et al., 2015; Watson et al., 2019; Berrocal et al., 2020). The selection, tuning and evaluation of interpolation models typically uses cross-validation (CV) to estimate prediction error. CV is appealing, because it is easy to perform and allows for the comparison of different classes of prediction models.

While CV is widely used, its behavior in the context of space-time dependence is inadequately understood. Efron famously characterized CV estimation of prediction error for independent data (Efron, 1983), but little has been done to extend this formalization to the case of dependent data, for which independent replication in not observable. Space-time dependence complicates model evaluation in two ways: (1) evaluation procedures developed for independent data are generally inappropriate, because they underestimate the true error (Roberts et al., 2017); (2) multiple types of prediction are possible due to different relationships between the training data and the data to be predicted, (e.g., forecasting as opposed to spatial interpolation), whereas prediction is unambiguous in the context of independent data. There is growing awareness in the literature that independent model evaluation procedures should not be naively applied to models fit to space-time data, prompting various alternatives. Motivated by air pollution data collected during California wildfires in 2008, we provide a formalization of the true prediction error associated with spatial interpolation and investigate a variety of CV procedures using simulations and case studies. Particulate and gaseous air pollution from wildfires often display extreme spatiotemporal heterogeneity, which makes this a particularly germane example with which to interrogate prediction error.

We restrict our discussion to approaches that estimate prediction error per se, notably excluding information criteria, which are popular tools for model comparison. We note that some information criteria require a likelihood to be computed, which may not exist for some prediction algorithms, and that the large number of proposed information criteria makes a thorough discussion infeasible. We also operate without concern for consistency of model selection, focusing on evaluation metrics that estimate prediction error. In a model selection context our goal is to select the best predicting model, i.e., that with the lowest expected prediction error. We do not require that such a procedure converge to the true model if

indeed a true model were to exist.

The remainder of this chapter proceeds as follows. Section 2.2 introduces a motivating case study on wildfire pollution; Section 2.3 reviews prediction error estimands and estimators first for independent data and then for space-time dependent data, formalizing the true prediction error associated with spatial interpolation; Sections 2.4 and 2.5 investigate the behavior of these estimators using a battery of simulations and two case studies; and Section 2.6 summarizes estimator performance and offers recommendations for best practices.

2.2 Motivating Application

Daily measurements of ozone (8-hour maximum) and 24-hour average $PM_{2.5}$ (particulate matter smaller than 2.5μm in diameter) were recorded at monitors across the state of California from May 1, 2008, through September 30, 2008. Within this time period a large complex of wildfires burned for several weeks in northern California, shrouding much of the state in smoke and exposing its populace to toxic pollutants.

Inferring the health consequences of exposure to these pollutants is typically done in two stages: (1) a space-time prediction model trained on the monitor data and fine-resolution covariates translates pollution measurements from the coarse space-time resolution of the monitors to the much finer spatial resolution of health outcome data, and (2) health outcomes are regressed on predicted pollution exposure adjusting for potentially confounding covariates. The two-stage approach breaks a challenging problem into two more manageable tasks, although uncertainty or inaccuracy in exposure predictions can induce bias in the subsequent epidemiological analysis (Szpiro and Paciorek, 2013). Avoiding or correcting for this bias is an active area of research. This manuscript concerns the first stage.

2.3 Prediction Error: Estimands and Estimators

2.3.1 The Case of Independent Replication

Let $\mathbf{z} = \{z_1, \ldots, z_n\}$ denote n independent and identically distributed (iid) samples from an unknown probability distribution F, such that $z_1, \ldots, z_n \sim_{iid} F$. We assume $z_i = \{y_i, \mathbf{x}_i\}$, where $y_i \in \mathbb{R}$ denotes a real-valued outcome, and $\mathbf{x}_i \in \mathbb{R}^p$ denotes a random vector of p covariates. In this context, a common framework often used to formalize the notion of prediction error relies on the existence of an independent and identically distributed test point $z_0 = \{y_0, \mathbf{x}_0\} \sim F$.

Let $\eta(\mathbf{x}_0, \mathbf{z}) : \mathbb{R}^{np+n+p} \mapsto \mathbb{R}$ define a prediction rule that maps the training data $\mathbf{z}$ and a covariate vector $\mathbf{x}_0$ to a real number that serves as a prediction for y_0. Discrepancies between y_0 and its predicted value may be quantified by a loss function, $Q\{y_0, \eta(\mathbf{x}_0, \mathbf{z})\} : \mathbb{R}^2 \mapsto \mathbb{R}^+$, mapping y_0 and its prediction $\eta(\mathbf{x}_0, \mathbf{z})$ to a non-negative real number. This loss function is easily generalized to quantify prediction across multiple test points, $\mathbf{z}_0 = \{z_{01}, ..., z_{0m}\}$, by summing the loss at individual points, i.e., $Q\{\mathbf{y}_0, \eta(\mathbf{X}_0, \mathbf{z})\} = \sum_{j=1}^{m} Q\{y_{0j}, \eta(\mathbf{x}_{0j}, \mathbf{z})\}$. Given the training data $\mathbf{z}$, the *true error* of a prediction rule trained on $\mathbf{z}$ is

$$Err \coloneqq E_F[Q\{y_0, \eta(\mathbf{x}_0, \mathbf{z})\}]; \tag{2.1}$$

the expected loss incurred by $\eta(\cdot)$, where E_F is an expectation taken with respect to probability distribution F, over the state space of z_0 (Efron, 1983).

We operate under the assumption that F is unknown, as is typical in many applications. In this context, the expected loss in equation 2.1 cannot be computed in closed form, and the definition of estimators of Err with good operating characteristics becomes a surprisingly formidable problem (Efron, 2004). Some progress is often made by constructing estimators which rely on the separation of training and testing through data partitions.

For example, K-fold cross-validation (CV) (Stone, 1974) partitions the training data $\mathbf{z}$ into K disjoint subsets often referred to as folds, $\mathbf{z}^{(1)} = \{\mathbf{y}^{(1)}, \mathbf{x}^{(1)}\}, \ldots, \mathbf{z}^{(K)} = \{\mathbf{y}^{(K)}, \mathbf{x}^{(K)}\}$, usually of equal or nearly equal size, where K is an integer between 2 and n, so that $\mathbf{z} =$

$\cup_{k=1}^{K} \mathbf{z}^{(k)}$, and $\mathbf{z}^{(k)} \cap \mathbf{z}^{(\ell)} = \varnothing$ for $k \neq \ell$. The prediction rule η is trained K times, each time excluding one subset, with the withheld subset serving as a test set on which to evaluate the prediction rule trained on the rest of the data. The K-fold CV estimator of prediction error averages the loss function evaluated over each of the withheld subsets,

$$\widehat{err}^{(CV_K)} \coloneqq \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} Q\left\{\mathbf{y}^{(k)}, \eta\left(\mathbf{x}^{(k)}, \mathbf{z}^{(-k)}\right)\right\}, \tag{2.2}$$

where $\mathbf{z}^{(-k)}$ is the training data excluding the k-th fold, i.e., $\mathbf{z}^{(-k)} \coloneqq \mathbf{z} \setminus \mathbf{z}^{(k)}$, and n_k is the number of observations in $\mathbf{z}^{(k)}$.

In a similar fashion, one may rely on the bootstrap (Efron, 1979). The procedure generates B data sets, $\mathbf{z}_1^*, \ldots, \mathbf{z}_B^*$, by resampling n observations with replacement from the training data $\mathbf{z}$. For each bootstrap dataset $\mathbf{z}_b^*$ $(b = 1, \ldots, B)$, let $\mathbf{z}_{-b}^* \coloneqq \mathbf{z} \setminus \mathbf{z}_b^*$ denote the training data withheld from the b-th bootstrap sample. The *out-of-bag* estimator of Err is defined as follows,

$$\widehat{err}^{(boot)} \coloneqq \frac{1}{B} \sum_{b=1}^{B} Q\left\{\mathbf{y}_{-b}^*, \eta\left(\mathbf{x}_{-b}^*, \mathbf{z}_b^*\right)\right\}.$$

When the training data, $\mathbf{z}$, are independent samples from F and have finite, positive variance, the bootstrap estimator converges in distribution to the true error (Bickel and Freedman, 1981). Further refinements of these two basic estimators are reported in (Zhang, 1995) and (Efron and Tibshirani, 1997).

A more direct estimator of Err is often used in practice when n is large. In this setting, one relies on a validation set $\mathbf{z}^*$ consisting of m samples from F independent of $\mathbf{z}$. The expectation in equation 2.1 may be approximated by averaging the loss function over the validation set, yielding the *validation error*,

$$err^* \coloneqq \frac{1}{m} \sum_{i=1}^{m} Q\left\{y_i^*, \eta\left(\mathbf{x}_i^*, \mathbf{z}\right)\right\},$$

where $\{y_i^*, \mathbf{x}_i^*\}$ is the i-th validation sample. Because the validation set consists of independent samples from F, it is an unbiased estimator of the true error with a variance dependent on the number of validation samples m.

Reliance on the basic idea of training a prediction rule on a subset of the data and evaluating it on a different (i.e., disjoint or independent) subset might be proscribed in the absence of independent replication. The use of cross-validation is, however, prevalent in several applications of machine learning to spatial and space-time data. Therefore, in the following section we attempt to clarify the role of data-partition strategies in space-time analysis.

2.3.2 Prediction Error Estimands for Space-Time Data

We characterize the sampling structure of space-time data using the formalism of marked point processes (Møller and Waagepetersen, 2007). Specifically, given a bounded spatial region $\mathcal{S} \subset \mathbb{R}^2$, we assume that data is observed at a spatial point pattern $\mathbf{s} = \{s_1, s_2, \ldots, s_n\}$, $s_i \in \mathcal{S}$, $(i = 1, 2, \ldots, n)$. Associated with each spatial location s_i, we observe a time series of real-valued outcomes $\mathbf{y}(s_i) = \{y(s_i, t_1), \ldots, y(s_i, t_{m_i})\}$, as well as a matrix of real-valued covariates $\mathbf{X}(s_i) = \{\mathbf{x}(s_i, t_1), \ldots, \mathbf{x}(s_i, t_{m_i})\} \in \mathbb{R}^{p \times m_i}$, where m_i is the length of the time series observed at s_i. The collection of outcomes and covariate information at location s_i is denoted as $z(s_i) \coloneqq \{\mathbf{y}(s_i), \mathbf{X}(s_i)\}$ and may be formally characterized as a point-process mark at location s_i.

Within this framework, a space-time dataset is composed of a spatial point pattern $\mathbf{s}$ with associated marks $\mathbf{z}(\mathbf{s}) = \{z(s_1), \ldots, z(s_n)\}$, and sampling distribution $F\{\mathbf{s}, \mathbf{z}(\mathbf{s})\} \coloneqq F_s(\mathbf{s}) F_{z|s}\{\mathbf{z}(\mathbf{s}) \mid \mathbf{s}\}$, which may be represented hierarchically as the marginal distribution of the point pattern $F_{\mathbf{s}}$ times the conditional distribution of the marks $F_{\mathbf{z}|\mathbf{s}}$.

In this setting, the notion of prediction is inherently ambiguous until one qualifies how a test data point is to be obtained. In particular, *spatial interpolation* is formally interpreted as the conditional prediction at location $s_0 \sim F_s$ of $\mathbf{y}(s_0)$, given $\mathbf{X}(s_0)$ and $\mathbf{z}(\mathbf{s})$. Let $\eta\left(\mathbf{X}(s_0), \mathbf{z}(\mathbf{s})\right)$ represent a prediction rule, and $Q\{\mathbf{y}(s_0), \eta\left(\mathbf{X}(s_0), \mathbf{z}(\mathbf{s})\right)\}$ be a loss function, both analogous to our definitions in Section 2.3.1. Given a training marked spatial pattern $\mathbf{z}(\mathbf{s})$, the *interpolation error* of $\eta(\cdot)$ trained on $\mathbf{z}(\mathbf{s})$, is naturally defined as

$$Err_s \coloneqq E_{F\{s_0, z(s_0)\}}\left[Q\{\mathbf{y}(s_0), \eta(\mathbf{X}(s_0), \mathbf{z}(\mathbf{s}))\}\right]; \tag{2.3}$$

the expected loss incurred by $\eta(\cdot)$, where $E_{F\{s_0,z(s_0)\}}$ is an expectation taken with respect to probability distribution F, over the joint state space of $\{s_0, z(s_0)\}$.

We note that, in the definition of Err_s, the spatial point pattern distribution $F_s(\cdot)$, need not coincide with the one defining the observed point pattern $\mathbf{s}$. It could be useful, for instance, to use a sampling schedule for s_0 proportional to the density of the population at risk. In this example, the use of exogenous information may be critical when the goal is one of providing spatial interpolation for environmental stressors, because predictive accuracy is much more important in populated areas.

2.3.3 Prediction Error Estimators for Space-Time Data

The formalism introduced in Section 2.3.2 will serve as a guiding framework for the characterization of structured data partition strategies defining prediction error estimators for space-time data.

Standard K-fold CV (equation 2.2) is still commonly used on space-time data (Reid et al., 2015; Guo et al., 2017) despite acknowledgement in the spatial statistics community that its estimates are too optimistic (Watson et al., 2019; Roberts et al., 2017). In light of the estimand defined in (2.3), unstructured data partitions are unlikely to exclude data associated with specific locations, making it very likely that locations represented in the test set also appear in the training data. Rather than targeting the spatial interpolation error, naïve CV estimators appear to target the *imputation error*, i.e. the expected loss incurred by a prediction rule $\eta(\cdot)$ for unobserved time points at observed locations.

We are not the first to point out these problems with standard K-fold CV. For example, Roberts et al. (2017) advocate for the use of spatial or space-time blocking as structured data partition strategies targeting prediction errors in spatial and space-time data. Specifically, the spatial (or space-time) domain is divided into rectangular (or cubic) regions that serve as a data-partition unit. The sizing of data blocks aims to reduce or eliminate the impact of spatial (or space-time) dependence on estimates of prediction error, in an effort to replicate estimators defined in iid settings. This approach thus appears to target a form of *replication*

error, i.e. the expected loss incurred by a prediction rule $\eta(\cdot)$ for a new realization of the marked space-time process, which is critically different from the spatial interpolation error defined in (2.3).

Alternatively, spatially buffered CV has been proposed for mitigating the effect of spatial dependence for prediction error estimation and model selection (Le Rest et al., 2014; Pohjankukka et al., 2017). These strategies pick one or more spatial locations for each test fold and train the model on observations located beyond a buffer distance from the test locations. The usual goal is to pick a buffer distance that excludes observations from the training data that have residual spatial correlation with the test data. This is the same goal as spatial blocking and appears to similarly target the replication error. Smaller buffers that do not entirely remove spatially dependent training data appear to target a *spatial extrapolation error*, i.e., the error associated with predicting at a spatially distant (but not independent) location.

The key to defining a structured CV strategy for the estimation of the interpolation error in (2.3), is naturally implied by the sampling structure of the marked point process introduced in Section 2.3.2. More precisely, let $\mathbf{z}\left(\mathbf{s}_{(i)}\right)$ be the set of observations across all locations, except for s_i. We define the leave-one-location-out cross-validation (LOLO-CV) estimator of Err_s as,

$$\widehat{err}_s^{(LOLO)} := \frac{1}{n}\sum_{i=1}^{n} Q\left\{\mathbf{y}(s_i), \eta\left(\mathbf{X}(s_i), \mathbf{z}\left(\mathbf{s}_{(i)}\right)\right)\right\}, \tag{2.4}$$

in which the CV fold left out of model training is the entire time series of observations taken at a particular location. This is analogous to the leave-one-out CV estimator in the iid setting, except the CV resampling unit is no longer individual observations but rather every observation taken at an individual location. Defining each CV fold as the observations taken at one location ensures that predictions are always being evaluated at a location excluded from the data on which the prediction rule $\eta(\cdot)$ was trained. This intuitively mimics interpolation in which the goal is to predict the time series at an unobserved location. This intuition is presumably the motivation for a number of authors using LOLO-CV (Keller et al., 2017; Watson et al., 2019), even though, to our knowledge, our analysis is the first to

attempt a formalization of the target estimand in (2.3).

LOLO-CV can be generalized to partition data over multiple locations. Specifically, we partition the observed point pattern $\mathbf{s}$ into K disjoint location-subsets or spatial folds, $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(K)}$, with associated marks $\mathbf{z}(\mathbf{s}^{(1)}), \ldots, \mathbf{z}(\mathbf{s}^{(K)})$, so that $\mathbf{s} = \cup_{k=1}^{K} \mathbf{s}^{(k)}$, and $\mathbf{s}^{(k)} \cap \mathbf{s}^{(\ell)} = \emptyset$, for $k \neq \ell$. The K-fold leave-location-out (LLO_K)-CV estimator of Err_s is defined as,

$$\widehat{err}_s^{(LLO_K)} := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in \mathbf{s}^{(k)}} Q\{\mathbf{y}(s_i), \eta(\mathbf{X}(s_i), \mathbf{z}(\mathbf{s}^{(-i)}))\},$$

where $\mathbf{z}(\mathbf{s}^{(-i)})$ is the set of observations at all locations excluding the spatial points in $\mathbf{s}^{(i)}$, and n_k is the number of locations in spatial fold k.

Setting $K = n$ yields the leave-one-location-out (LOLO) cross-validation estimate of prediction error, $\widehat{err}_s^{(LOLO)}$ in (2.4). LLO_K-CV trains and evaluates K models so that when K is smaller than n, LLO_K is computationally expedient compared to LOLO-CV. Setting $K = 10$ is a common choice, presumably on analogy to the iid setting (Lee et al., 2016).

As in the independent case, there is a bias-variance tradeoff between these estimators. In the context of independent data, the LOO estimator has lower bias, but higher variance than K-fold CV estimators for $K > 1$. In the dependent setting, however, removing locations from the data can result in hard to predict sources of bias, and the standard wisdom that suggests 10-fold CV as a rule of thumb estimator (Kohavi et al., 1995) for balancing bias and variance may not apply. Estimator bias depends on the dependence structure, the density of observed locations and the number of observations in each time series. We elucidate some of these aspects in our simulation studies, showing that LOLO CV seems to be the best for typical studies in which the observed data include a relatively small number of locations and a relatively large number of time points.

LOLO-CV and LLO_K-CV use the empirical distribution of locations $\hat{F}(\mathbf{s})$ in computing the average loss of η. The observed locations provide an estimate of the spatial intensity function, which weights the average loss more heavily in areas with more observed locations. If prediction error is to be taken as the expectation over a uniform spatial field, then using the empirical spatial intensity may introduce sampling bias (Diggle et al., 2010; Gelfand et al.,

2012). In epidemiological applications, however, it is often desirable to define prediction error over a non-uniform sampling intensity, with the preferential sampling pattern of pollution monitors providing critical information into the relative importance of different locations. If it is desirable to define prediction error as the expected loss over a different location sampling intensity, it is foreseeable to apply corrections to $\hat{err}^{(LOLO)}$ and $\hat{err}^{(LLO_K)}$. This could be used to estimate the expected loss uniformly across the spatial region or to incorporate ancillary information, for example, direct estimates of population density if prediction error were to be more heavily weighted where the population were more dense. These corrected estimators are not investigated in this manuscript.

2.4 Simulation Study

To illustrate the behavior of these prediction error estimators in the context of dependent data, we perform a series of simulation studies. Each simulation is conducted over a dense space-time grid with a small subset of the grid locations sampled as observed locations. Six covariates and the outcome were simulated over the space-time grid. The covariates were simulated from a Gaussian process with a separable covariance function and a sinusoidal mean function. The outcome was simulated from a Gaussian process with a separable covariance function and a mean function that depended on the first three covariates with linear, sigmoidal and threshold effects respectively. The remaining three covariates are spurious. The precise form of each mean and covariance function is detailed in section 2.A.1.

For a given model, the true prediction error is computed as the mean square error (MSE) of predictions made over the unobserved portion of the grid by a model trained on the observed data. Each estimator of prediction error is then computed over the observed data and compared to the true prediction error. To help elucidate the impact of sample size and dependence structure on estimator behavior and motivated by air pollution data sets, we chose two different numbers of observed monitors, 50 and 150, and two sets of spatial dependence parameters, corresponding to low and moderate dependence. Table 2.A.3 in section 2.A.3 lists the parameters of each simulation scenario. Points near the boundary of

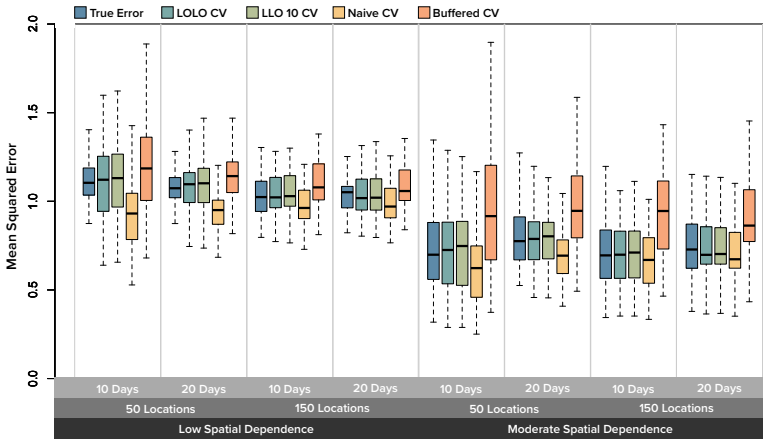
the grid have fewer nearby points, which can result in unexpected edge effects. To avoid this, only points sufficiently far from the grid boundary were eligible for selection as observed locations. We evaluated four estimators of prediction error: (1) naive 10-fold CV, (2) LLO_{10} CV, (3) LOLO CV and (4) buffered leave-location-out CV on three different models: linear regression, random forest and space-time Kriging. These models were selected as examples of the three main classes of models used in the literature: (1) simple regression models, (2) machine learning models with flexible mean structures that assume independent data, and (3) Kriging models with rigid mean structures but relatively flexible covariance models. These three models are far from a comprehensive list and are intended to illustrate the behavior of different CV procedures on simple, but different classes of prediction strategies.

2.4.1 Simulation Results & Discussion

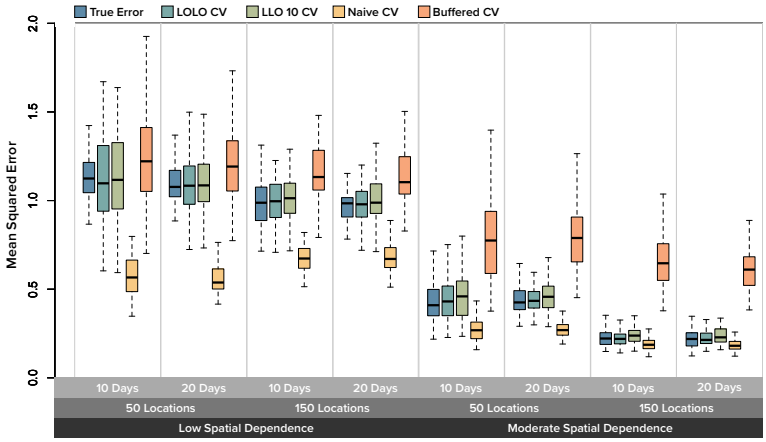
Figure 2.1 depicts the true error and four estimates for each model across the 100 replicates of each simulation scenario. Across all three models, LOLO CV (turquoise) and LLO_{10} CV (green) appear to target the true error (blue).

Naive 10-fold CV (yellow) is very optimistic for random forest and Kriging, especially when spatial dependence is low or relatively few locations are observed. The information provided by observations at other time points at the location of the data point being predicted is apparently very informative for models flexible enough to exploit it. Naive CV is less optimistic for simple linear regression, because it is much less flexible than random forest or Kriging and consequently worse at exploiting this information.

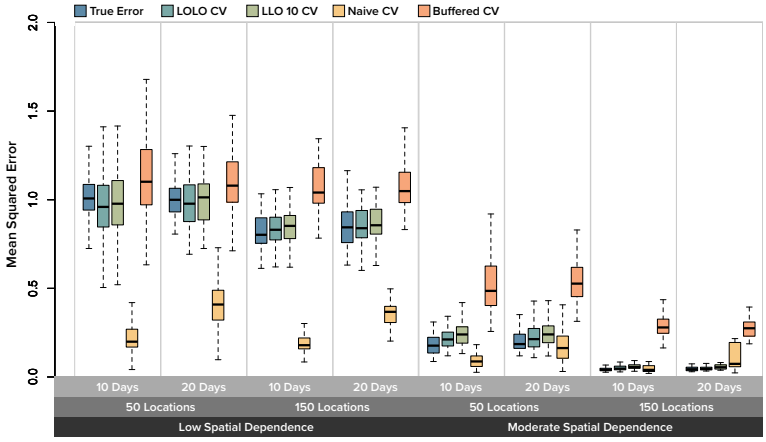
The optimism of naive 10-fold CV is less pronounced as the number of observed locations or the spatial dependence increases for all 3 models. In these cases, the other observed locations are more predictive of the test location, making naive CV's access to observations for other timepoints at the test location less advantageous and thereby reducing its optimistic bias. Interestingly, the naive 10-fold CV estimates of prediction error for Kriging become less stable for longer time series despite the corresponding additional data, suggesting that naive



(a) Simple Linear Regression



(b) Random Forest



(c) Kriging

Figure 2.1: Boxplots of the true error and its estimators for each model across the 100 replicates of each simulation scenario.

10-fold CV is particularly unreliable for techniques that rely upon the dependence structure for prediction.

The LOLO estimates of prediction error tend to be slightly smaller and less biased (i.e., closer to the true error) than LLO_{10} for two reasons. First, LLO_{10} generally uses smaller data sets for training because its folds are larger, and the additional training data available to LOLO reduces its bias. This result is well-known in cross-validated studies of independent data. Second, withholding more locations further hinders the LLO_{10} training models because it removes the information they provide on other locations due to the spatial dependence between locations. This does not occur in the independent setting where there is no dependence between observations.

This effect makes the bias-variance trade-off of CV fold size more complicated than in the independent case. For independent data, as the fraction of the data used in the training data increases, the bias decreases, but the variance increases. Leave-one-out CV includes all but 1 observation in the training data and has smaller bias but greater variance than 10-fold CV, which includes 90% of the data in the training data. For data exhibiting space-time dependence, this relationship need not hold. Whether LOLO CV has higher variance than LLO_{10} CV depends on the dependence structure, the number of observed locations and the length of the time series observed at each location.

Spatially buffered CV (orange) is pessimistic (i.e., biased upward) in all cases, which is consistent with the suggestion that it may target a spatial extrapolation error rather than the interpolation error. For a fixed buffer size, its pessimism is greater for higher spatial dependence, since the locations within the buffer that are withheld from the training model are more highly correlated with the observations being predicted. Its pessimism is lower when more locations are observed, because the additional locations beyond the buffer may partially make up for the loss of information associated with withholding observations within the buffer.

Figure 2.2 plots MSE against CV estimates of MSE for each of the 3 models for an example simulation scenario. The CV estimates generally correlate fairly well with the true

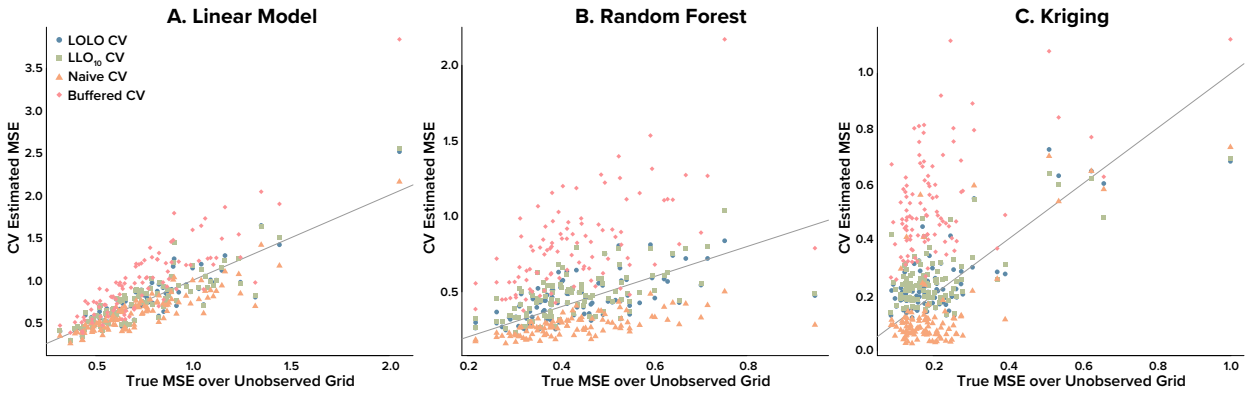
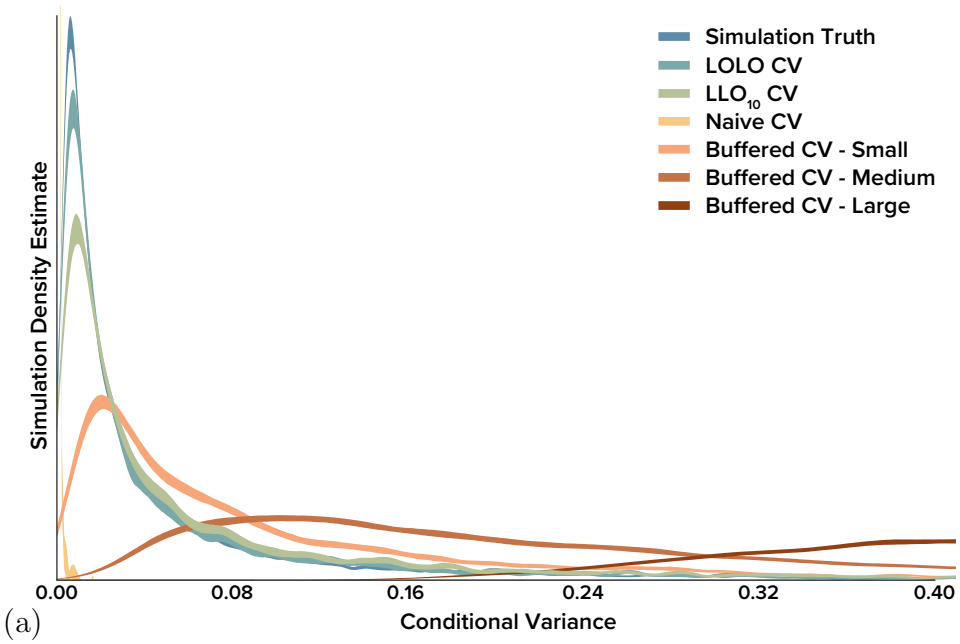


Figure 2.2: Scatter plots of the true MSE computed over the unobserved grid against CV estimates of MSE for (a) Simple Linear Regression, (b) Random Forest, and (c) Kriging. The grey line depicts $y = x$, i.e., perfect correlation.

MSE, suggesting they are not simply accurate on average, but provide useful estimates for a particular data set. The optimistic bias of naive 10-fold CV and the pessimistic bias of buffered CV are evident from their deviation from the line of perfect correlation. As in figure 2.1, these biases are more pronounced for random forest and Kriging on account of their greater flexibility.

It is useful to characterize prediction error as a function of the conditional variance of the outcome to be predicted given the observed data. This conditional variance effectively quantifies the information that the training data provides on the point to be predicted and can be computed exactly in a simulation setting (see section 2.A.4 for detail). Figure 2.3 summarizes our findings for simulation scenario 3. In panel (a), we report density estimates and 95% pointwise intervals of the conditional variance for the true error and each cross-validated estimator, using 300 bootstraps of the simulation replicates.

How closely the conditional variance distribution for an estimator matches the true distribution (blue) corresponds to how accurately it estimates the true prediction error. LOLO CV (turquoise) is most similar to the true error with a slightly greater distribution due to the slightly smaller size of the LOLO training data. LLO_{10} CV (green) is also similar to the true distribution, but the larger fold size increases the conditional variance because of the



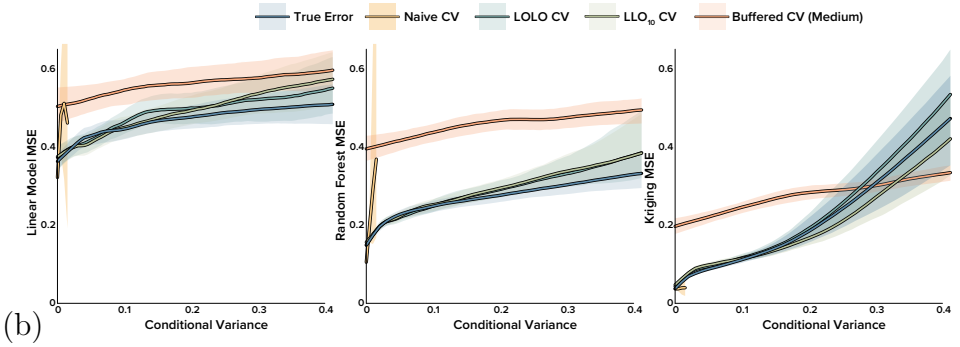


Figure 2.3: Conditional variance of test points vs. MSE. (a) Density estimates and 95% pointwise confidence bands of test points given the training data for the 100 replicates of simulation scenario 3. (b) Median and 95% pointwise confidence bands for the true and estimated MSE as a function of the conditional variance of test points given the training data for the 100 replicates of simulation scenario 3 (LOWESS smoothed).

correspondingly smaller size of the training data. The conditional variance for naive CV is much lower than that of the true distribution and it peaks so sharply near 0 that it is difficult to make out in the figure. This difference in conditional variance describes the optimistic bias of naive CV: the training data is much more informative for prediction points than the truth, corresponding to a much lower conditional variance and overly optimistic estimates of interpolation error.

Figure 2.3 includes the conditional variance for three spatially buffered CV estimators of increasing buffer size. As the buffer size increases so does the conditional variance, because a larger buffer corresponds to more spatially distant and therefore less informative training data. This explains the pessimistic bias of buffered CV in estimating interpolation error. As the buffer size is reduced and less distant locations are included in the training data, buffered CV approaches LOLO CV, becoming equivalent when the buffer size is smaller than the closest spatial distance between observations.

Figure 2.3 (b) depicts the median and 95% pointwise confidence bands for LOWESS (locally weighted smoothing) curves estimating the mean true and estimated errors as a function of conditional variance for each model over 300 bootstraps of the simulation replicates. As expected, the true error increases with conditional variance. LOLO and LLO_{10} CV also exhibit this relationship, approximating the true error curve quite well for all three models. At larger values of conditional variance there are relatively few data points and consequently greater uncertainty around the estimated curves. This figure shows quite clearly why naive 10-fold CV is such a poor estimator. Because naive 10-fold CV includes observations taken at the same locations as the outcomes being predicted, the conditional variances are very small compared to the conditional variance of the outcome at unobserved locations.

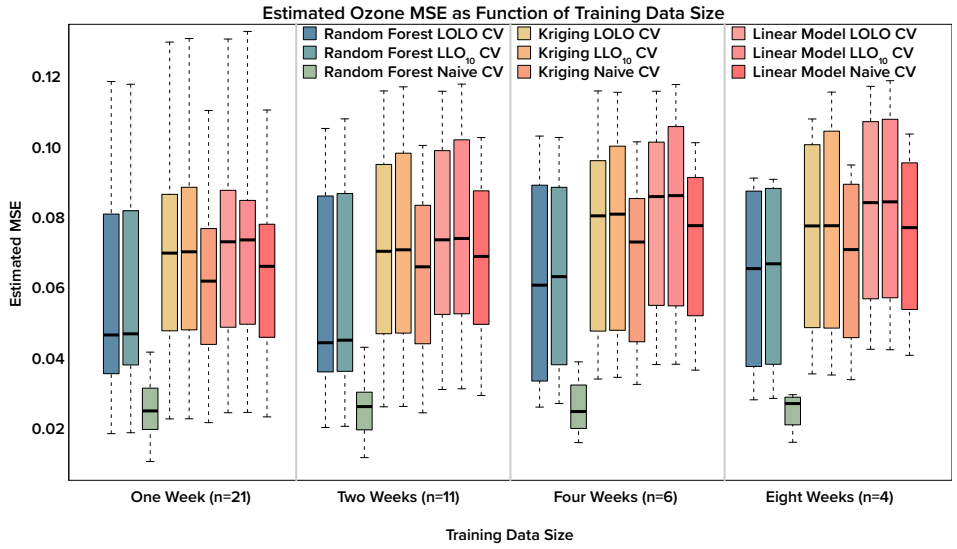
The medium buffer size spatially buffered CV is also depicted, which consistently overestimates the true error except for Kriging at high conditional variance. This may simply result from the greater uncertainty due to the relatively few observations with high conditional variance, but it may also reflect the sensitivity of Kriging to nearby observations, which are withheld from model training in spatially buffered CV.

Our results in Figure 2.3 (b) also allow for an interesting comparison of predictive models. Kriging predicts very well (i.e., has low MSE) when the conditional variance is small, because it explicitly uses the space-time dependence between data points. Its accuracy falls off substantially, however, as the conditional variance increases. Random forest and simple linear regression ignore this dependence, which reduces their accuracy when the conditional variance is small (especially for simple linear regression), but their performance falls off less quickly at higher conditional variance than does Kriging.

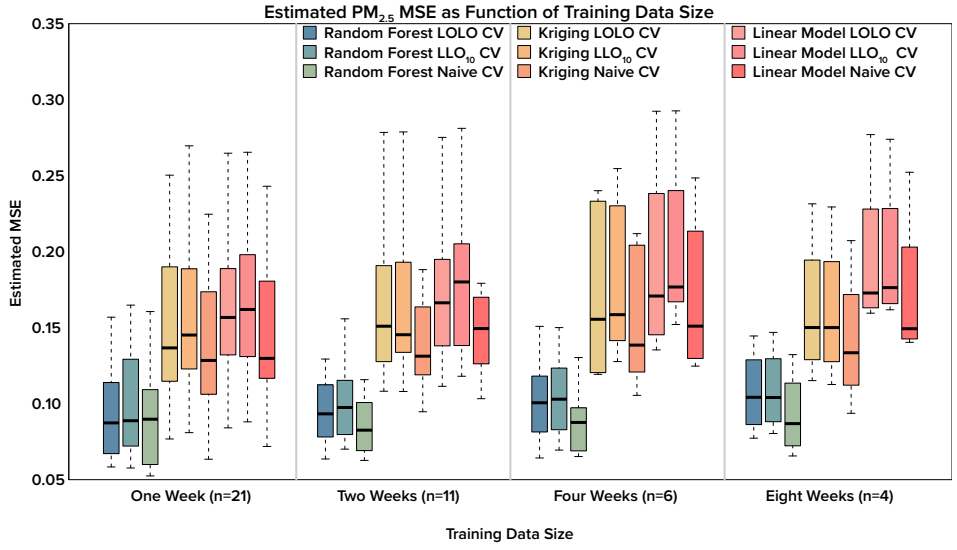
2.5 Case Study

Over the weekend of June 20–21, 2008, thousands of lightning strikes in northern California ignited a complex of wildfires that would burn for several weeks, shrouding much of the state in smoke and exposing its populace to toxic pollutants. We set out to showcase how different model evaluation strategies compare, when the goal is producing fine resolution interpolation of pollutant exposure across the state of California.

More precisely, 186 ozone monitors recorded a total of 25,236 measurements of 8-hour maximum ozone, and 136 $PM_{2.5}$ monitors made 14,247 $PM_{2.5}$ measurements from May 1, 2008, through September 30, 2008, with most monitors recording one observation per day. Linear regression, random forest and universal Kriging (i.e., Kriging with covariates) interpolating prediction models for ozone and $PM_{2.5}$ were fit separately to each week of the data using land use, meteorological, satellite and numerical simulation covariates, which are detailed in table 4.A.2 in section 2.A.1. The LOLO, LLO_{10} and naive 10-fold CV estimates of prediction error were computed for each model. To assess the impact of time series length on the prediction error estimators, this procedure was repeated on progressively longer, non-overlapping intervals of 1, 2, 4 and 8 weeks.



(a) Ozone



(b) $PM_{2.5}$

Figure 2.4: Prediction error as a function of training data size for random forest, Kriging and simple linear regresson on (a) ozone and (b) $PM_{2.5}$ monitoring data.

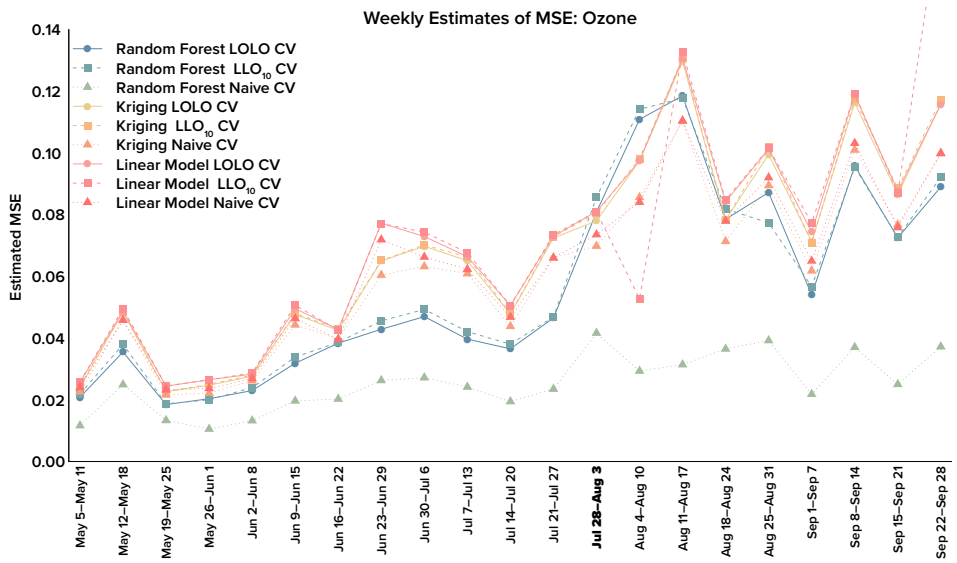
2.5.1 Case Study Results & Discussion

Figure 2.4 depicts LOLO, LLO_{10} and naive CV prediction error estimates on the ozone (a) and $PM_{2.5}$ (b) monitor data for increasingly long subsets of the data. Just as in the simulations, the LOLO and LLO_{10} estimates are similar, and LLO_{10} tends to be slightly more pessimistic. Estimated prediction error generally increases slightly with time series length. While additional timepoints add information, they also increase the prediction burden. This is particularly true if the temporal process is nonstationary, which figure 2.5 suggests may be the case at least for the ozone data.

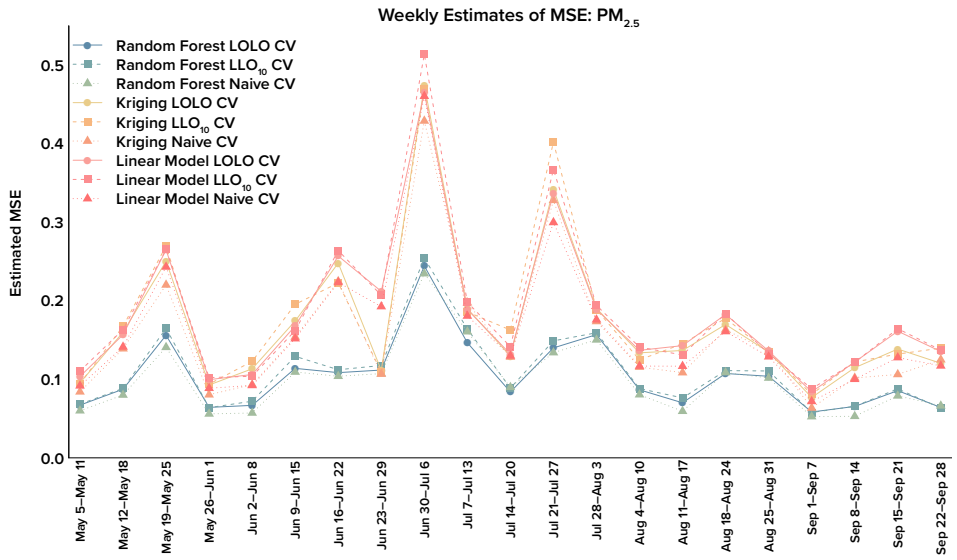
The naive 10-fold CV estimates are consistently optimistic compared to the location-based CV estimates. This optimism is considerably more pronounced for random forest than for Kriging or simple linear regression on the ozone data. The naive 10-fold CV estimates are apparently less biased for simple linear regression, because it lacks the flexibility to fully exploit the more highly dependent folds. Kriging, on the other hand, is very flexible, but its performance depends upon accurately estimating the covariance function, which may be quite difficult especially if the process is nonstationary. The optimism of the naive 10-fold CV ozone estimates for random forest increases with time series length, suggesting that the additional training data at the location where prediction is being made exacerbates the bias.

Figure 2.5 depicts prediction error estimates for models trained and evaluated separately on each weeklong subset of the data. Most of the estimates for ozone increase dramatically in the second half of the study period. This suggests that the process is nonstationary and that the WRF-Chem model may be miscalibrated for this time period, which may both be due to the wildfires that occurred during this time. Interestingly, the naive CV estimates for random forest do not exhibit this trend with estimates staying in approximately the same range across the entire study, clearly indicating once more the dangers of using naive CV for flexible models fit to dependent data.

Figure 2.5 (a) also demonstrates the perils of using naive CV for model selection, averaging or tuning (including variable selection). Consider the week of July 28, 2008, which is



(a) Ozone



(b) $PM_{2.5}$

Figure 2.5: Weekly estimates of prediction error for ozone (a) and $PM_{2.5}$ (b) monitoring data.

bolded along horizontal axis of the figure. Comparing models via naive CV, random forest is the best model by a wide margin with an estimated MSE of 0.042, compared to 0.070 and 0.074 for Kriging and the simple linear regression respectively. In contrast, using LOLO or LLO_{10} Kriging has the lowest MSE. Model ranking is thus not invariant to choice of evaluation metric, which means that naive CV is inappropriate for any procedure that depends upon the relative ranking of predictive performance. These includes model and variable selection, model parameter tuning and model averaging, which often uses relative performance to compute model weights (Van der Laan et al., 2007).

2.6 Discussion

The simulation and case study results confirm that location-based CV techniques are fairly robust estimators of interpolation error over several classes of models, consistent with recent best practice. Naive CV is shown to be highly unreliable for interpolation error estimation as well as for model comparison on space-time data. Buffered CV, which excludes observations within a buffer distance from the training data, is overly pessimistic and appears to target the replication error—the expected error for predicting a new instance of the entire space-time process—rather than the interpolation error, which explicitly conditions upon the observed portions of the space-time field. It also requires selecting the buffer distance, which may be difficult and prompt ad hoc decisions regarding the search window. In addition, even for independent data, CV procedures that include a tuning parameter (e.g., the buffer distance) may not be consistent (Dudoit and van der Laan, 2005).

LOLO and LLO_{10} CV both appear to target the true interpolation error with LLO_{10} generally displaying a greater pessimistic bias. This is partially due to the larger amount of data in the LOLO training folds, which is analogous to the well-known bias-variance tradeoff in CV fold size for independent data. Reserving more locations from model training as the test fold in LLO_{10} CV eliminates portions of the space-time field from the training data. How this impacts estimates of prediction error depends upon the strength and distribution of the space-time dependence, time series length and spatial sampling intensity. It can be the case

that LOLO CV has lower bias *and* lower variance than LLO_{10}. Consequently we recommend LOLO CV for estimating interpolation error unless it is computationally infeasible, in which case LLO_{10} is a reasonable alternative.

This judgment is an empirical one, as there is a lack of theoretical characterization of these procedures in an asymptotic setting. Asymptotics are challenging for CV even in the case of independent data, which admits to replication. In a space-time setting asymptotics may be characterized in multiple ways increasing the difficulty of the problem.

Random forest, Kriging and simple linear regression were selected because they are popular examples of 3 model classes frequently used for spatial or space-time interpolation. This gives us confidence that the behavior of the estimators observed here will hold across spatially interpolating models more generally. The models may not be the optimal models for these or other interpolation problems. Gradient boosting (Watson et al., 2019; Reid et al., 2015) and sometimes deep learning neural networks have outperformed random forest in previous studies. Kriging models with more flexible mean structures (Berrocal et al., 2010) or less stringent assumptions on the form of the covariance function may outperform the Kriging models considered here. The ability to reliably compare these and any other interpolation techniques is precisely the motivation behind this work, and the consistent relationship between CV estimators of prediction error demonstrated here will likely hold across a wide variety of models. Crucially this includes methods that accommodate high dimensional data. This is particularly important for Kriging models, which have traditionally been limited to relatively small data sets due to the matrix inversion involved, which does not scale well either in computations ($O(n^3)$) or memory usage ($O(n^2)$) Cressie (2015a). There are now a variety of computationally expedient approaches that have improved the scaleability of Kriging-style models, including Gaussian process approximations (Katzfuss, 2017) and nearest neighbor Gaussian process models (Datta et al., 2016). See Heaton et al. (2019) for a recent summary and comparison of these and other approaches.

More useful than the specific guidance this work suggests for the appropriate choice of CV techniques for evaluating interpolation models is its formal discussion of prediction

error estimation for space-time data. While there has been growing acknowledgment in the literature that naively applying unstructured CV on space-time data is ill advised, there is a lack of clarity on what prediction error means in the presence of space-time dependence. Nowhere is this more obvious than in the recent trend towards reporting naive, location-based and timepoint-based CV estimates of prediction error while offering no guidance or discussion on which is appropriate for the problem at hand (Brokamp et al., 2018; Hu et al., 2017).

This work fills this void, distinguishing estimators that target interpolation error from those that appear to target replication, imputation or extrapolation errors. The appropriate space-time prediction error depends entirely upon the predictive problem at hand. Identifying the relevant space-time prediction error for a particular problem and then selecting an evaluation method for that specific prediction error is the appropriate model evaluation procedure.

We have not dealt with issues related to consistency of model selection, focusing strictly on prediction error estimation. A model selection procedure is generally described as consistent if the probability it selects the true model goes to 1 as $n \to \infty$. Leave-one-out (LOO) CV is known to be an inconsistent criterion for model selection on independent data, although it can be adapted to reclaim consistency for at least certain classes of models (Shao, 1993). We do not expect location-based CV procedures to provide consistent selection of the true model. Nevertheless they may be quite useful for selecting/averaging models on the basis of a principled assessment of predictive performance.

2.A Appendix

2.A.1 Case Study Covariates

Covariate	Data Source
Monitor Latitude	U. S. Environmental Protection Agency
Monitor Longitude	U. S. Environmental Protection Agency
Elevation (m)	National Digital Elevation Model
Date	U. S. Environmental Protection Agency
Dew Point (°K)	Rapid Update Cycle
Boundary Layer Height (m)	Rapid Update Cycle
Surface Pressure (Pa)	Rapid Update Cycle
Relative Humidity (%)	Rapid Update Cycle
Temperature at 2 m (°K)	Rapid Update Cycle
U-Component of Wind Speed (m/s)	Rapid Update Cycle
V-Component of Wind Speed (m/s)	Rapid Update Cycle
Inverse Distance to Nearest Fire (m^{-1})	Fire Inventory from NCAR v1.5
Annual Average Traffic within 1 km	Dynamap 2000, TeleAtlas
Agricultural Land Use within 1 km (%)	2006 National Land Cover Database
Urban Land Use within 1 km (%)	2006 National Land Cover Database
Vegetation Land Use within 1 km (%)	2006 National Land Cover Database
Normalized Difference Vegetation Index	Landsat Data
Nitrogen Dioxide (log molecules/cm^2)	Ozone Monitoring Instrument Satellite
WRF-Chem Carbon Monoxide (log moles/day)	WRF-Chem
WRF-Chem $PM_{2.5}$ (log kg/day)	WRF-Chem
WRF-Chem Ozone (log 8 Hour Maximum)	WRF-Chem

2.A.2 Simulation Procedure

Covariates for the simulations described in section 3 were generated from a 6-dimensional Gaussian process (GP) with mean function $\mu_X(s,t)$ and covariance function $\Sigma(s,t)$, i.e.,

$$X_1, ..., X_6 \sim GP(\mu_X(s,t), \Sigma(s,t)),$$

where s indexes spatial locations and t indexes time. A combination of sinusoidal functions was selected for the mean function to induce some periodicity across space and time,

$$\mu_X(s,t) = \sin(t)\frac{\sin(s_1)}{250u_1}\frac{\cos(s_2)}{125u_2} + \cos(t)\left(\frac{\cos(s_1)}{125u_3} + \frac{\sin(s_2)}{250u_4}\right),$$

where $u_1, ..., u_4$ are uniformly distributed random variables, i.e., $u_1, ..., u_4 \sim U[0,1]$.

The space-time covariance function $\Sigma(s,t)$ was constructed as the product of a spacial covariance function Σ_s and a temporal covariance function Σ_t, i.e., $\Sigma(s,t) = \Sigma_s\Sigma_t$. These were parameterized as squared exponential covariance functions,

$$\Sigma_s(s_1, s_2) = \exp\left(-\frac{\|s_1 - s_2\|^2}{v_s^2}\right) + \epsilon I_{n_s},$$
$$\Sigma_t(t_1, t_2) = \exp\left(-\frac{\|t_1 - t_2\|^2}{v_t^2}\right) + \epsilon I_{n_t},$$

where $\|\cdot\|$ is the Euclidean norm, v_s and v_t scale dependence as a function of distance, and ϵI_{n_s} and ϵI_{n_t} are diagonal matrices that add a small constant for numerical stability.

The simulation outcome y was simulated from a GP with mean function $\mu_y(s,t)$ and covariance function $\Sigma(s,t)$,

$$y \sim GP(\mu_y(s,t), \Sigma(s,t)), \tag{2.5}$$

where

$$\mu_y(s,t) = 0.5x_1 - 0.5I(x_2 > 0.1) + (1 + \exp(-x_3))^{-1}.$$

The outcome is independent of the remaining 3 covariates, x_4, x_5, and x_6, which were included as spurious covariates to assess the robustness of prediction models.

2.A.3 Simulation Parameters

Scenario	Replicates	Observed Locations	Spatial Dependence	Temporal Dependence	Locations	Days
1	100	50	Low	Moderate	400	10
2	100	50	Low	Moderate	400	20
3	100	50	Moderate	Moderate	400	10
4	100	50	Moderate	Moderate	400	20
5	100	150	Low	Moderate	400	10
6	100	150	Low	Moderate	400	20
7	100	150	Moderate	Moderate	400	10
8	100	150	Moderate	Moderate	400	20

2.A.4 Conditional Variance

Simulation outcome $y(s,t)$ was simulated over a space-time lattice from a Gaussian Process with mean function $\mu_y(s,t)$ and covariance function $\Sigma(s,t)$ (see Section S2). Let $\mathbf{y} = \{y(s_1,t_1),\ ...,\ y(s_n,t_m)\}$ denote y over the n spatial locations and m timepoints of the simulation lattice, and $\Sigma_{\mathbf{y}}$ its covariance matrix. Let $\mathbf{y}^*$ denote any subset of $\mathbf{y}$ and $\mathbf{y}^c$ its complement, i.e., $\mathbf{y}^c = \mathbf{y} \setminus \mathbf{y}^*$. The covariance matrices of $\mathbf{y}^*$ and $\mathbf{y}^c$ are $\Sigma_{\mathbf{y}^*}$ and $\Sigma_{\mathbf{y}^c}$ respectively, which are submatrices of $\Sigma_{\mathbf{y}}$ with the rows and columns corresponding to $\mathbf{y}^*$ and $\mathbf{y}^c$ respectively. The conditional covariance matrix of $\mathbf{y}^*$ given $\mathbf{y}^c$ is

$$\Sigma_{\mathbf{y}^*|\mathbf{y}^c} = \Sigma_{\mathbf{y}^*} - \Sigma_{\mathbf{y}^*\mathbf{y}^c}\Sigma_{\mathbf{y}^c}^{-1}\Sigma'_{\mathbf{y}^*\mathbf{y}^c},$$

where $\Sigma_{\mathbf{y}^*\mathbf{y}^c}$ is the submatrix of $\Sigma_{\mathbf{y}}$ with rows corresponding to $\mathbf{y}^*$ and columns to $\mathbf{y}^c$. The conditional variance of $\mathbf{y}^* \mid \mathbf{y}^c$ is the diagonal of $\Sigma_{\mathbf{y}^*|\mathbf{y}^c}$.

CHAPTER 3

Machine Learning Models Accurately Predict Ozone Exposure during Wildfire Events

Epidemiologists use prediction models to downscale (i.e., interpolate) air pollution exposure where monitoring data is insufficient. This study compares machine learning prediction models for ground-level ozone during wildfires, evaluating the predictive accuracy of ten algorithms on the daily 8-hour maximum average ozone during a 2008 wildfire event in northern California. Models were evaluated using a leave-one-location-out cross-validation (LOLO CV) procedure to account for the spatial and temporal dependence of the data and produce more realistic estimates of prediction error. LOLO CV avoids both the well-known overly optimistic bias of k-fold cross-validation on dependent data and the conservative bias of evaluating prediction error over a coarser spatial resolution via leave-k-locations-out CV. Gradient boosting was the most accurate of the ten machine learning algorithms with the lowest LOLO CV estimated root mean square error (0.228) and the highest LOLO CV $\hat{R}^2$ (0.677). Random forest was the second best performing algorithm with an LOLO CV $\hat{R}^2$ of 0.661. The LOLO CV estimates of predictive accuracy were less optimistic than 10-fold CV estimates for all ten models. The difference in estimated accuracy between the 10-fold CV and LOLO CV was greater for more flexible models like gradient boosting and random forest. The order of estimated model accuracy depended on the choice of evaluation metric, indicating that 10-fold CV and LOLO CV may select different models or sets of covariates as optimal, which calls into question the reliability of 10-fold CV for model (or variable) selection. These prediction models are designed for interpolating ozone exposure, and are not suited to inferring the effect of wildfires on ozone or extrapolating to predict ozone in other spatial or temporal domains. This is demonstrated by the inability of the best performing models to accurately predict ozone during 2007 southern California wildfires.

3.1 Introduction

Ground-level ozone is toxic to humans, animals and plants and contributes significantly to climate change as the third most important greenhouse gas (World Health Organization, 2013; Menzel, 1984; Bell et al., 2004; Stocker et al., 2013; Smith et al., 2009; Silva et al., 2017; Fowler et al., 2008). Short-term exposure is linked to increased mortality (Bell et al., 2004), decreased respiratory function, exacerbation of chronic obstructive pulmonary disease (COPD), bronchitis, emphysema and asthma (National Research Council, 2008; Sousa et al., 2013; Guarnieri and Balmes, 2014; Lee et al., 2014). Long-term exposure has been linked with respiratory and cardiovascular mortality (Turner et al., 2016; Crouse et al., 2015), decreased lung function (Hwang et al., 2015) and the progression of emphysema (Barr et al., 2016).

Wildfires contribute to the formation of ozone in the lower atmosphere (troposphere) by releasing volatile organic compounds (VOCs) and nitrogen oxides (NOx), which react in the presence of sunlight to form ozone (Val Martin et al., 2006; Jaffe and Wigder, 2012). Fires upwind of large population centers can expose millions or even tens of millions of people to ozone and other pollutants (Wu et al., 2006). Climate change is expected to intensify wildfires, which will likely increase the prevalence of wildfire-related ozone exposure Jaffe and Wigder (2012); Confalonieri et al. (2007).

The health effects of wildfire-induced ozone exposure are poorly understood. A study of hospital admissions in Porto, Portugal, in 2005 while wildfires were burning nearby, indicated ozone was significantly associated with cardiovascular disease admissions, but not with respiratory admissions (Azevedo et al., 2011). This analysis, however, did not control for weather or land-use covariates. During a bushfire in southeastern Australia, respiratory emergency department visits were significantly associated with PM_{10} (particulate matter 10 μm or smaller in diameter), but not with ozone (Tham et al., 2009).

A key challenge facing epidemiological analyses of air pollution exposure is quantifying pollution concentrations where people live, which may be distant from regulatory monitoring sites. This is particularly difficult for wildfire-related pollution, because wildfires often ignite far from urban regulatory monitoring sites, and satellite evidence indicates that traditional monitoring networks are too sparse to capture smoke plume variation and dynamics (Reid et al., 2015). Epidemiologists

attempt to overcome this difficulty by constructing exposure models to predict pollution concentration at unmonitored locations and times. The prediction of a quantity across a domain, such as a spatial region, based on observations of that quantity at discrete locations within that domain is referred to as downscaling or interpolation, and is an example of infill prediction.

The simplest downscaling exposure models rely upon the tendency of nearby observations to be more similar than those farther apart to interpolate between pollution monitor observations without the use of additional information (i.e., without covariates). This tendency is an example of spatial (or space-time) dependence. Kriging is a very commonly used method for interpolating observations, modeling air pollution concentrations as the best linear unbiased prediction (BLUP) given the data and mean and covariance functions selected by the researcher and estimated from the data (Stein, 2012). Kriging tends to perform relatively well when monitoring data is dense, but model accuracy degrades at locations or times distant from monitor observations.

To improve accuracy, especially when monitoring density is sparse, researchers have employed regression models that incorporate ancillary information as covariates. These models are referred to as land use regression in the literature, but the covariates need not pertain to land use, and increasingly include satellite retrievals, meteorological data, and less frequently the output of atmospheric chemistry numerical simulation models. Land use regression models have been used for modeling air pollution exposure at least since the Small Area Variations in Air quality and Health (SAVIAH) study in 1997 (Briggs et al., 1997) with numerous examples appearing subsequently. While these regression models include covariate information, they have often assumed linear, additive covariate effects. This assumption makes the effects easy to interpret but is too stringent to predict air pollution concentrations accurately. These models cannot accommodate nonlinear effects or interactions between covariates unless they are specified by the analyst a priori. They also lack a mechanism for variable selection, requiring the analyst to manually select covariates or employ a separate variable selection procedure.

These limitations have prompted the development of more flexible models that allow for nonlinear effects including spatially or spatiotemporally varying-coefficient models (Gelfand et al., 2003). These models generally fit covariate effects with smooth functionals that need not be linear and may be indexed by space or space-time, which allows the covariate effects to differ across space and time. These models are an improvement over the very stringent restrictions set on covariate

effects in linear, additive models, but they often rely on research code, making them less accessible to other researchers. In our experience we have also found that these more sophisticated models do not scale well to realistic space-time data settings.

These challenges have motivated researchers to turn to more flexible models that do not require such stringent assumptions, including a variety of machine learning algorithms, which have been shown to be very useful for prediction (Robert, 2014), especially random forest (Caruana and Niculescu-Mizil, 2006), gradient boosting (Ferreira and Figueiredo, 2012) and neural networks (Goodfellow et al., 2016). They may lack the straightforward interpretability of linear regression, but this is of secondary concern when prediction is the primary objective, and variable importance scores have been developed for many such methods to quantify the contribution of each covariate.

Generalized additive models (Liu et al., 2009), support vector machines (Lu and Wang, 2005; Weizhen et al., 2014), gradient boosting (Zhan et al., 2017) and deletion/substitution/addition (Beckerman et al., 2013) have been used to model particulate matter exposure. Neural networks (Di et al., 2016, 2017) and random forest (Hu et al., 2017; Zhan et al., 2018a; Brokamp et al., 2018) have been used to model both particulate matter and ozone concentrations. A comparison of 11 machine learning models indicated that random forest, gradient boosting and bagged trees predict $PM_{2.5}$ (particulate matter smaller than 2.5 μm in diameter) concentrations well during a wildfire event (Reid et al., 2015). Random forest, boosting and Cubist performed well in a comparison of 8 machine learning tools predicting $PM_{2.5}$ in British Columbia (Xu et al., 2018). Here we conduct a similar analysis using ten machine learning algorithms to model ozone exposure during a wildfire air pollution event for the first time, evaluating their predictive accuracy for use as land use regression models.

Comparing and evaluating prediction models for dependent data is challenging. Cross-validation (CV) and the bootstrap are commonly used model evaluation procedures that repeatedly fit a model to a training subset of the data and evaluate the accuracy of its predictions on a different, test subset, combining the performance across multiple test subsets into a nonparametric estimate of prediction error. For data that are spatially and temporally dependent (i.e., autocorrelated), however, these procedures can be overly optimistic, because of the dependence between training and test subsets (Roberts et al., 2017).

In the case of daily air pollution monitoring observations, we wish to estimate the average error made by a downscaling model when predicting at a new location within the spatial domain of the data. Including observations in the training data that were taken at the same monitors as the test set observations provides an unrealistic amount of information on the test data, because of the strong correlation between observations taken at the same location. This produces estimates of prediction error that are biased downward, especially for flexible models which tend to overfit to a greater degree than less flexible models when trained on dependent data. The true prediction error associated with predicting at a new location would be greater than these estimates, because the model could not have been trained on any observations recorded at a new location.

When dependence is restricted to observations within the same group or cluster, consistent (i.e., asymptotically unbiased) estimates of prediction error can be recovered by resampling groups rather than individual observations. It is unrealistic, however, to assume that spatially dependent data are nested within independent groups of observations. Modified cross-validation schemes have been used on pollution exposure data that partition the data into spatial grid cells (Zhan et al., 2017) or monitor locations (Lee et al., 2016; Zhan et al., 2018b). Such approaches attempt to reduce the dependence between training and test data sets by placing all observations at a particular location or within a particular region into the same cross-validation fold. In this vein, we use leave-one-location-out (LOLO) cross-validation, which defines each CV fold as the observations recorded at a single monitor location (Just et al., 2015). This does not partition the data into independent groups, but estimates the error associated with predicting the time series of ozone observations at a new location, conditioning upon the observed monitor data. By using all the observations at a single location as the test set, LOLO CV ensures that no observations from this location appear in the training set, which would result in unrealistically low estimates of prediction error. It also avoids the overly conservative bias of leave-k-locations-out CV, which tends to overestimate prediction error because it uses substantially fewer observations for model training.

Table 3.1: Covariates for ozone prediction.

Covariate	Data Source
Monitor Latitude	U. S. Environmental Protection Agency
Monitor Longitude	U. S. Environmental Protection Agency
Elevation (m)	National Digital Elevation Model
Date	U. S. Environmental Protection Agency
Dew Point (°K)	Rapid Update Cycle
Boundary Layer Height (m)	Rapid Update Cycle
Surface Pressure (Pa)	Rapid Update Cycle
Relative Humidity (%)	Rapid Update Cycle
Temperature at 2 m (°K)	Rapid Update Cycle
U-Component of Wind Speed (m/s)	Rapid Update Cycle
V-Component of Wind Speed (m/s)	Rapid Update Cycle
Inverse Distance to Nearest Fire (m^{-1})	Fire Inventory from NCAR v1.5
Annual Average Traffic within 1 km	Dynamap 2000, TeleAtlas
Agricultural Land Use within 1 km (%)	2006 National Land Cover Database
Urban Land Use within 1 km (%)	2006 National Land Cover Database
Vegetation Land Use within 1 km (%)	2006 National Land Cover Database
Normalized Difference Vegetation Index	Landsat Data
Nitrogen Dioxide (log molecules/cm^2)	Ozone Monitoring Instrument Satellite
WRF-Chem Carbon Monoxide (log moles/day)	WRF-Chem
WRF-Chem $PM_{2.5}$ (log kg/day)	WRF-Chem
WRF-Chem Ozone (log 8 Hour Maximum)	WRF-Chem

3.2 Materials and Methods

3.2.1 Data

One hundred ground-based ozone monitors administered by the United States Environmental Protection Agency (EPA) made hourly observations from which the daily maximum 8-hour averages were computed across northern California between May 6, 2008 and September 26, 2008 for a total of 13,487 observations. We selected this time period with the goal of estimating ozone exposures before, during, and after a spate of wildfires that afflicted northern California in late June and July of 2008. The mean maximum 8-hour average concentration was 36.2 ppb, and the standard deviation was 13.6 ppb. During the study, the maximum 8-hour average exceeded 70 ppb 236 times and exceeded 75 ppb 107 times. Most exceedances occurred while the fires were burning (153 and 75 respectively), although this time period is one of high solar intensity when high concentrations of ozone are expected. It is not our objective, however, to quantify the contribution of wildfires to ozone formation, but simply to downscale ozone concentrations during a wildfire event for subsequent epidemiological analysis. Figure 3.1 depicts the temporal evolution of monitor observations throughout this time period.

Twenty-one covariates were also collected for the monitor locations, including location, elevation, date, atmospheric weather data (dew point, boundary layer height, surface pressure, relative humidity, temperature, and wind speed), inverse distance to the nearest fire, traffic, land use information (agricultural, urban, and vegetation), tropospheric nitrogen dioxide (NO_2) vertical column density and predictions of daily total carbon monoxide concentration (CO), particulate matter ($PM_2.5$) and daily maximum 8-hour average ozone. Table 4.A.2 lists the covariates and their sources.

Monitor elevation was determined from the 2010 National Elevation Dataset for California. The date of each observation was encoded as the continuous covariate, Julian date. The U.S. National Centers for Environmental Prediction's Rapid Update Cycle atmospheric prediction model provided hourly predictions of dew point, planetary boundary layer height, surface pressure, relative humidity, temperature, and the U and V components of wind speed, which were averaged into daily values U.S. Department of Commerce National Oceanic and Atmospheric Administration (2018).

Inverse distance to the nearest fire was included as a covariate. The Fire Inventory from

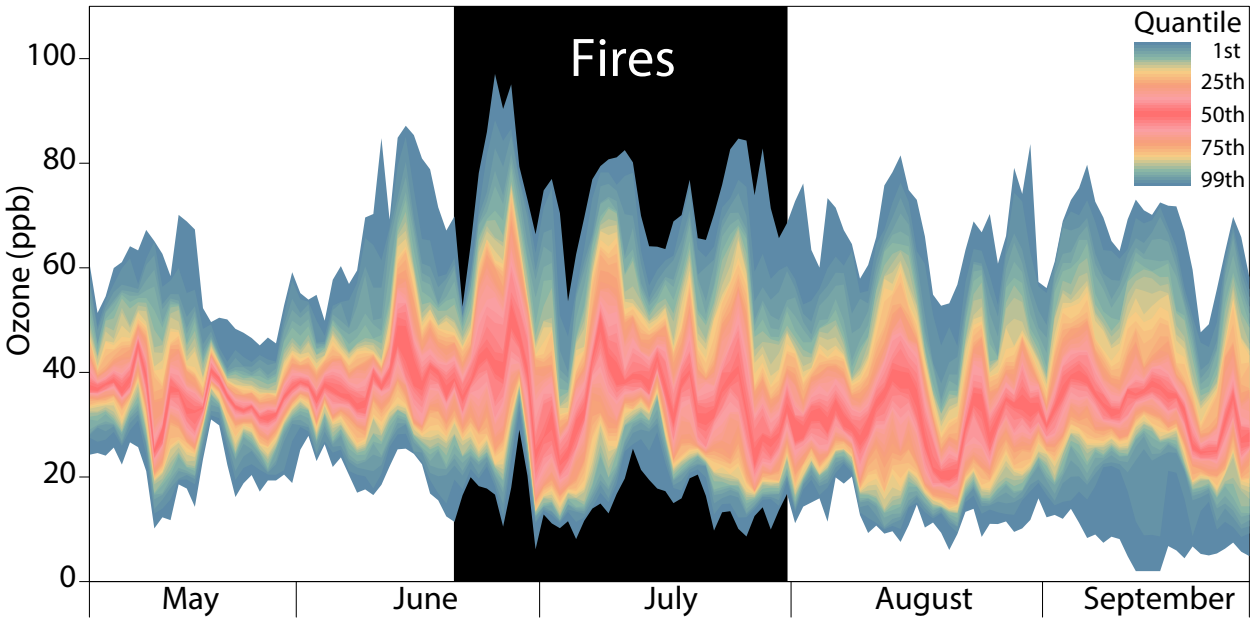


Figure 3.1: The daily empirical distribution of maximum daily 8-hour average ozone between May 6 and September 26, 2008 at 100 northern California monitoring sites.

NCAR (FINN) v1.5 provided estimates of fire point locations in California during the study period (Wiedinmyer et al., 2011). Fire points occurring within 5 km of each other were clustered and circumscribed by a polygon using the ArcGIS Aggregate Points tool, and the distance between each monitoring site and the closest point on the nearest fire cluster polygon was determined on each day using the ArcGIS Near tool. On days with no fire in California, distance to the nearest fire was undefined. Conceptualizing this undefined distance as equivalent to the nearest fire cluster being infinitely far away, inverse distance to fire was defined as 0 for observations taken on days with no fires in California and as the inverse of the distance to the nearest fire cluster otherwise.

Dynamap 2000, a TeleAtlas product, was used to compute the annual average of roadway traffic within 1 km of each monitor (Reid et al., 2015). The National Land Cover Database for 2006 (Fry et al., 2011) was used to calculate the percentage of urban development (codes 22, 23, and 24), agriculture (codes 81 and 82) and other vegetation (codes 21, 41, 42, 43, 52, and 71) within 1 km of each monitor.

The normalized difference vegetation index (NDVI) quantifies the density of green vegetation on a scale between -1 and 1 by measuring the visible and near-infrared light reflected at a location via

remote sensing. The chlorophyll in healthy vegetation absorbs most of the visible light and reflects much of the near-infrared light to which it is exposed, giving locations with more vegetation a higher NDVI score. NDVI for each monitor location was extracted from the NDVI remote sensing raster surface and included as a covariate.

Nitrogen dioxide (NO_2) was estimated on each day at monitor locations (if available) using the Berkeley High-Resolution (BEHR) NO_2 tropospheric column density retrieved from NASA's Ozone Monitoring Instrument (OMI) satellite, which has an overpass time of 1:30 local time (Levelt et al., 2006) and a resolution varying between 13 x 24 km to 42 x 162 km.

Predictions of daily total carbon monoxide (CO) and $PM_{2.5}$ and the maximum daily 8-hour average ozone concentration were extracted for each day from the Weather Research and Forecasting with Chemistry (WRF-Chem) 3.2 model. WRF-Chem is a regional chemical transport model that simulates meteorology and behavior of atmospheric gases and aerosols (Powers et al., 2017; Pfister et al., 2011b). Appendix 3.A.1 details the WRF-Chem inputs and options used for our simulations.

3.2.2 Statistical Analysis

Each observation comprises an outcome, y_i, the log maximum 8-hour average ozone on a given day at a given monitoring location, and a vector of covariates, $\mathbf{x}_i = (x_{i1}, ..., x_{ip})', i = 1, ..., n$, where n is the number of observations, and p is the number of covariates. The vector of outcomes, $\mathbf{y} = (y_1, ..., y_n)'$, and the matrix of covariates, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)'$, together compose the data, $D = \{\mathbf{X}, \mathbf{y}\}$. Ozone observations were log transformed to reduce the impact of heteroscedasticity (non-constant variance across the range of a variable), as data exploration revealed the variance was substantially greater than the mean at high values. The maximum daily 8-hour average ozone from the WRF chemical transport model (WRF-Chem) was also log transformed to have the same scale as the outcome. All other covariates were transformed to have a mean of 0 and variance of 1.

Ten predictive algorithms were trained and evaluated on these data: elastic net regression, generalized additive models (GAM), gradient boosting, k-nearest neighbor regression, lasso regression, linear models, multivariate adaptive regression splines (MARS), neural network, random forest, and support vector machines with a radial basis kernel (SVM). All of the models except for neural networks were fit using models available in version 6.0 of the caret R package (Kuhn et al.,

2018). Neural networks were given special consideration on account of their growing popularity as machine learning prediction tools and especially the recent publication of papers using a neural network with inverse distance weighted convolutional layers to predict ozone and particulate matter (Di et al., 2016, 2017). We tested a neural network that mimicked those models, employing inverse distance weighting to create convolutional spatial, temporal and space-time layers using the keras R package (Allaire and Chollet, 2015). These models were less accurate than a standard feedforward neural network, and so we have reported the results of that network here. Training each prediction model produces a prediction rule $\eta(\mathbf{x}, D_T)$, which is a function of D_T, the data on which it was trained, and a vector of covariates, $\mathbf{x}$, mapping them to a prediction for $y \mid \mathbf{x}$, which is often used as an estimator of $E(y \mid \mathbf{x})$, the conditional expectation of y given $\mathbf{x}$.

The models were tuned, selected, and evaluated using cross-validated estimators of root mean square error (RMSE) and R^2, which are both functions of the mean square error (MSE). The MSE of a prediction rule $\eta(\mathbf{x}, D_T)$, where D_T is the data with which η was trained, may be estimated using a test data set D_W as

$$\hat{MSE}(D_W, \eta(\mathbf{x}, D_T)) = \frac{1}{n_w} \sum_{j \in D_W} (y_j - \eta(\mathbf{x}_j, D_T))^2, \tag{3.1}$$

where n_w is the number of data points in D_W. If D_W and D_T are disjoint, (i.e., if η was not trained using any part of D_W), then this is an out-of-sample estimator of the MSE. RMSE may be estimated by the square root of $\hat{MSE}$, and R^2 is estimated by

$$\hat{R}^2(D_W, \eta(\mathbf{x}, D_T)) = 1 - \frac{\hat{MSE}(D_W, \eta(\mathbf{x}, D_T))}{n_w^{-1} \sum_{j \in D_W} (y_j - \bar{y}_w)^2}, \tag{3.2}$$

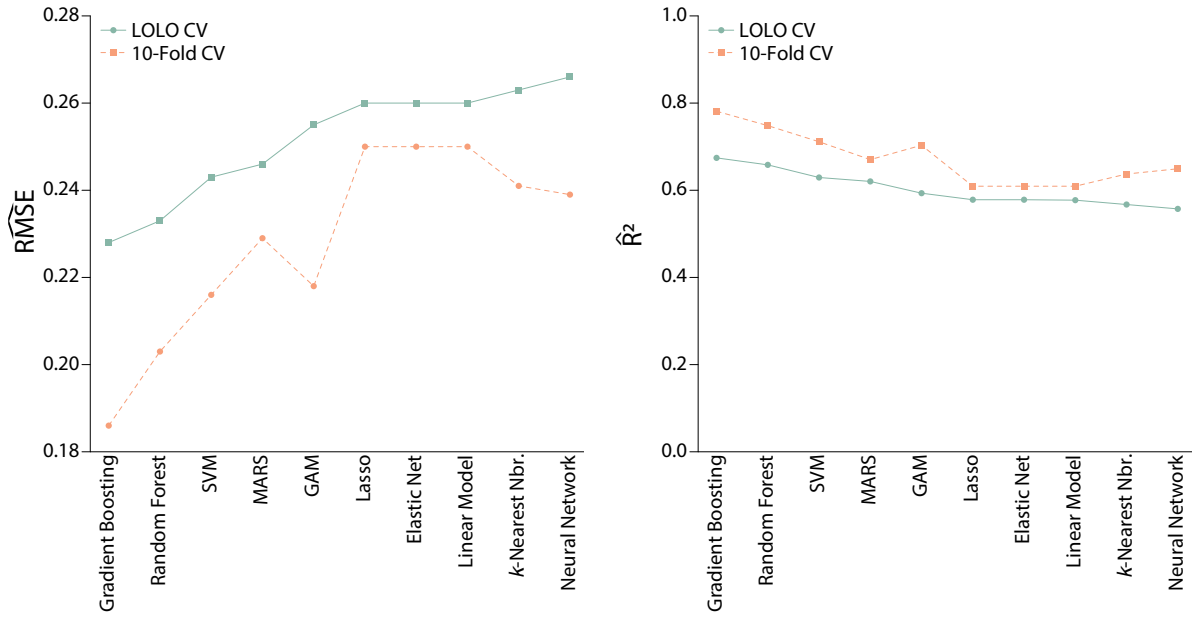
where $\bar{y}_w = n_w^{-1} \sum_{j \in D_W} y_j$ is the mean outcome in D_W. For ease of notation, the function arguments for $\hat{MSE}$, $\hat{RMSE}$, and $\hat{R}^2$ are hereafter suppressed.

Two different cross-validation (CV) strategies were employed for model evaluation: 10-fold cross-validation and leave-one-location-out (LOLO) cross-validation. For 10-fold CV, the data were randomly partitioned into 10 non-overlapping subsets, each containing one tenth of the data. Each subset served as the test data for models trained on the other nine tenths of the data, resulting in ten different pairs of training and test sets, with each observation appearing in one test set and the nine training sets not paired with that test set. This yielded 10 estimates of MSE for each model, which were averaged into an overall estimate of MSE, from which the 10-fold CV estimates of RMSE and R^2 were computed.

Ten-fold cross-validation is widely used for estimating prediction error; however, it is known to be overly optimistic for dependent data (Roberts et al., 2017). Data recorded by air pollution monitors are expected to exhibit spatial or space-time dependence. To more accurately estimate the downscaling error associated with predicting ozone at an unobserved location, RMSE and R^2 were estimated using LOLO CV, in which a model is trained on data from all but one location, and its prediction error is computed for the observations at the withheld location. This process is repeated with observations at each location serving as the withheld test set once, and the resulting errors are averaged into the LOLO CV estimate of prediction error. Unlike 10-fold CV in which observations are distributed among folds uniformly at random, LOLO CV ensures that no observations from the test location may appear in the training data. This provides a realistic estimate of the downscaling prediction error associated with predicting ozone observations at a new location within the same region as the monitoring data.

Most predictive machine learning algorithms depend upon one or more parameters whose values must be set prior to fitting the model. Algorithm performance can vary greatly depending on these parameter values, and it is often desirable to select values that optimize some criteria in an attempt to improve model performance. The process of choosing values for these parameters is often referred to as tuning and the parameters themselves as tuning parameters (or hyperparameters). In our analysis, most tuning parameter values were selected by comparing the performance of candidate values on 25 bootstrap samples of the data using the caret R package (Kuhn et al., 2018). Parameters for k-nearest neighbors and GAM were specifically tuned for LOLO CV in an attempt to stabilize the LOLO CV prediction error, as these models made extremely poor LOLO predictions using bootstrap-selected tuning parameter values. Appendix 3.A.3 in the supplemental material details these tuning procedures and their results. Substantial effort was taken in selecting the number of layers, nodes, activation functions and distance weighting functions for the neural network. The most accurate model was a feedforward neural network with one hidden layer of 21 nodes using a rectified linear unit activation function without the inverse distance weighting functions and convolutional layers employed in previously published models (Di et al., 2016, 2017).

To investigate the transferability of a model trained on data in one region to another, i.e., its ability to extrapolate rather than downscale, the predictive performance of the two best models trained on the 2008 northern California wildfire period—those two with the lowest LOLO CV



(a) $\hat{RMSE}$ (b) $\hat{R}^2$

Figure 3.2: 10-fold and leave-one-location-out cross-validated estimates of RMSE and R^2 for downscaling ozone prediction models.

estimates of RMSE—was evaluated on data collected during a 2007 wildfire event in southern California. The southern California data consisted of 5,978 daily 8-hour maximum ozone values recorded at 72 monitors between September 1, 2007 and November 28, 2007.

3.3 Results and Discussion

Figure 3.2 graphically depicts the cross-validated estimates of RMSE and R^2 for each algorithm using 10-fold CV and LOLO CV. In every case, the 10-fold CV $\hat{RMSE}$ was lower than the LOLO CV $\hat{RMSE}$, and the 10-fold CV $\hat{R}^2$ was higher than the LOLO CV $\hat{R}^2$. Gradient boosting had the lowest 10-fold CV $\hat{RMSE}$ (0.186 log ppm), lowest LOLO CV $\hat{RMSE}$ (0.228 log ppm), highest 10-fold CV $\hat{R}^2$ (0.784), and highest LOLO CV $\hat{R}^2$ (0.677). Random forest placed second in all four categories. The table in Appendix B lists exact values for $\hat{RMSE}$ and $\hat{R}^2$ for each model. These results answer the two primary questions posed by this study, demonstrating that machine learning methods can downscale ozone during a wildfire with reasonable accuracy and identifying gradient boosting and random forest as performing particularly well.

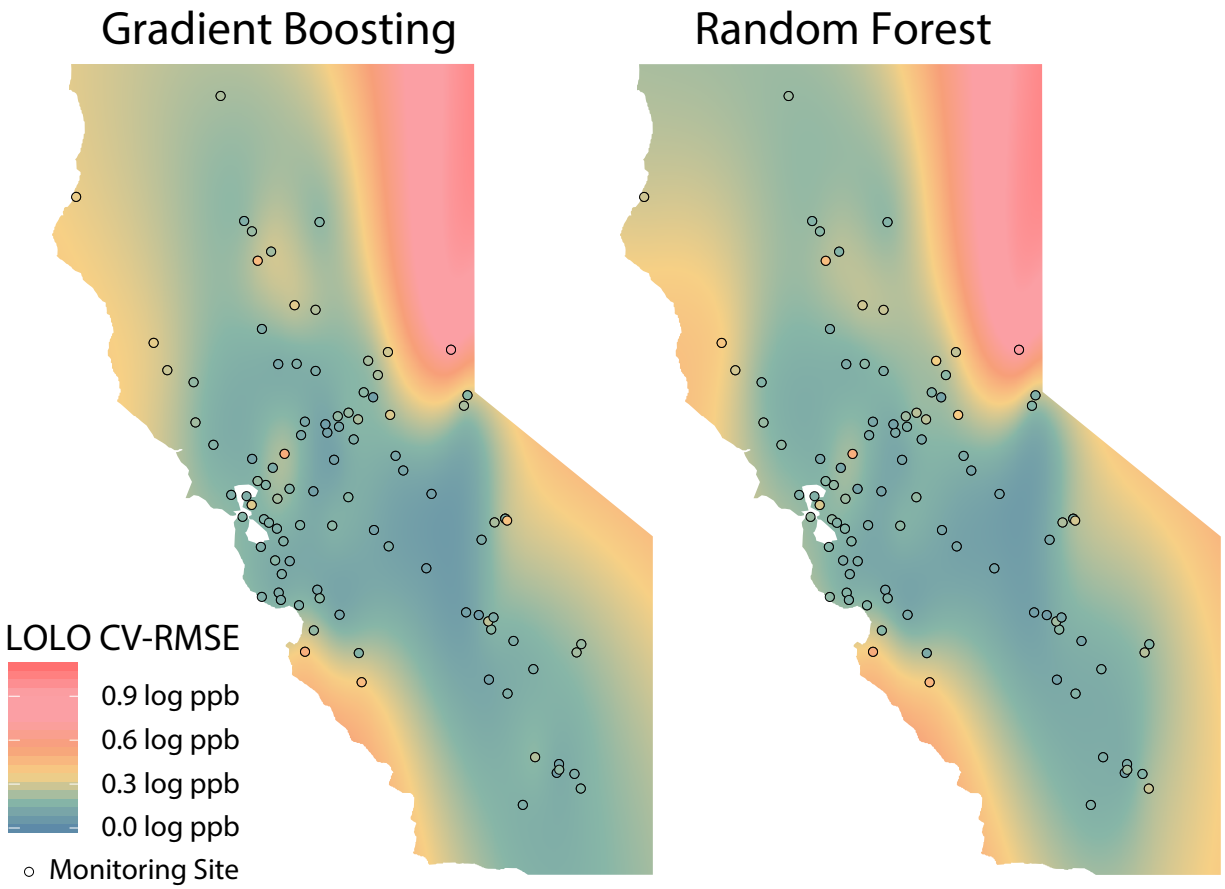


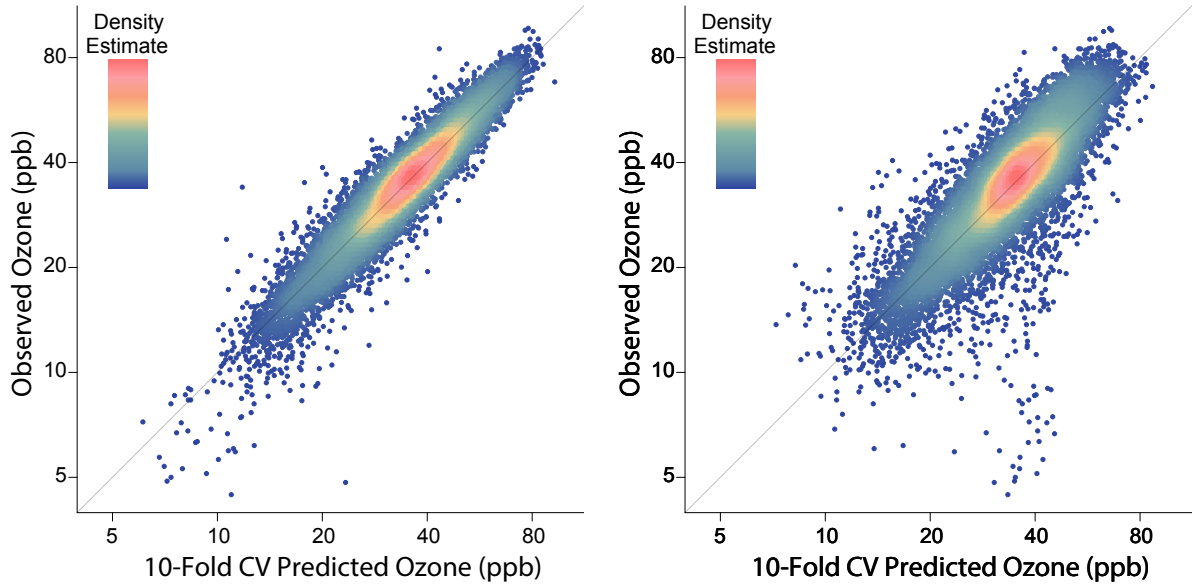
Figure 3.3: The leave-one-location-out (LOLO) cross-validated estimates of RMSE averaged over the study period (May 6, 2008–September 26, 2008) are plotted at each monitor location for gradient boosting (left) and random forest (right). These average prediction error estimates were then smoothed throughout the study region using a two-dimensional spline-on-sphere smoother to provide a visual estimate of how downscaling predictive performance may vary across the spatial domain.

The 10-fold CV estimates of RMSE and R^2 were optimistic compared to those of LOLO CV for all ten models. This over optimism of 10-fold CV is intuitive, because of the strong dependence between test and training observations from the same monitor location. The 10-fold CV estimators are also unreliable for model selection. The ordering of model performance is not invariant to the choice of evaluation criterion, which is demonstrated here by the evaluation of the neural network. It is the worst performing model when evaluated by LOLO CV, but is sixth best according to 10-fold CV. The ordering of GAM, MARS and k-nearest neighbors also differ, though the magnitude of those differences is not as substantial.

The difference between the 10-fold and LOLO cross-validated estimates of performance was smaller for relatively inflexible models like lasso, elastic net, and linear regression than for the other models, whose greater flexibility enabled them to better exploit the more highly dependent folds of 10-fold CV. The large difference for k-nearest neighbor regression is due to the strong dependence between observations recorded at the same monitor location. In 10-fold CV, the nearest neighbors of an observation are very likely to be other observations taken at that location. In LOLO CV, no observations taken at the test location appear in the training data. It is no surprise that this yields substantially higher estimates of prediction error.

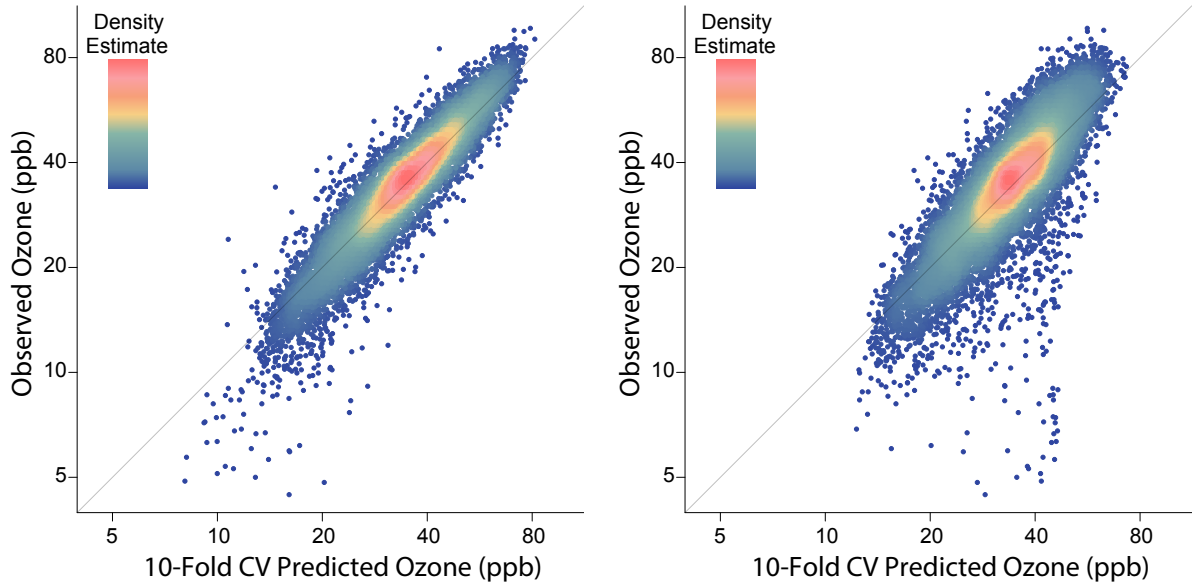
Figure 3.4 plots 10-fold and LOLO CV predictions against observed daily 8-hour maximum average ozone for gradient boosting and random forest. The CV prediction for each point is the predicted value when it appears in the test fold of the CV procedure. Points that fall on the grey diagonal line are perfectly predicted. The tighter clustering of points around this line in the 10-fold CV plots corresponds to the more accurate predictions made when test set monitor locations are included in the training data.

The neural network performed less well than in previous applications for predicting ozone and particulate matter over a spatial grid across the continental United States (Di et al., 2016, 2017). The recent popularity of convolutional neural networks is largely due to their performance on image-processing problems. Gridded spatial (or space-time) data bear a much greater resemblance to image data than do the point process monitor data upon which they were evaluated here. It is also possible that alternate specifications of the network architecture could improve performance, but developing such a model goes well beyond the scope of this comparison, which is limited to readily available algorithms that do not require substantial expertise in model specification or



(a) Gradient Boosting 10-Fold CV

(b) Gradient Boosting LOLO CV



(c) Random Forest 10-Fold CV

(d) Random Forest LOLO CV

Figure 3.4: 10-fold and leave-one-location-out (LOLO) cross-validated gradient boosting and random forest predictions plotted against observed daily 8-hour maximum average ozone on the log scale.

implementation. At least in this context neural networks do not succeed as an automatically regularized statistical learning tool (i.e., as a black box), but their performance in other studies suggests they may work well as a highly specialized tool designed using domain knowledge specific to a particular application.

The magnitude of the difference between the LOLO and 10-fold CV estimates of prediction error has meaningful consequences for estimating exposures for subsequent epidemiological analyses. Downscaled exposure is often used as the covariate of interest in analyses seeking to infer the health consequences of air pollution without accounting for prediction uncertainty. The more realistic estimates of prediction error provided by LOLO CV offer better insight into whether it is reasonable to ignore this uncertainty. This may motivate improvements to epidemiological models to account for exposure measurement error.

The increased accuracy of LOLO CV comes at a computational cost. When the number of locations exceeds 10, LOLO CV is more computationally expensive than 10-fold CV. In this analysis, there are 100 monitor locations and therefore 100 folds in LOLO CV, corresponding to approximately 10 times the computational burden of 10-fold CV. Grouping monitor locations into folds is an appealing strategy to alleviate this burden (Zhan et al., 2018b), however, it estimates prediction error over a different spatial resolution than LOLO CV, and will result in overly conservative estimates.

The top two models, gradient boosting and random forest, are both ensembles of tree-based models that provide very flexible mean structures. Their excellence suggests that the mean structure characterizing the relationship between covariates and ozone likely includes interactions, nonlinearities and possibly discontinuities. These results do not demonstrate that the underlying chemical processes by which ozone forms are similarly complicated, but that seems likely. The 10-fold CV estimates of RMSE and R^2 were similar to, although slightly lower than, those reported in a similar analysis of machine learning exposure models for $PM_{2.5}$ during the same wildfire time period (Reid et al., 2015). Tree-based ensembles were also the best performing models in that study, suggesting that algorithms with flexible mean structures can produce useful exposure models for ozone and $PM_{2.5}$ during wildfire events. Traditional exposure models have focused on modeling the dependence between observations, while employing a simple mean structure. The machine learning models evaluated here assume independent observations, but offer much greater flexibility in mod-

eling the mean. This approach is expected to provide more accurate predictions distant from the observations on which the model was trained than methods that rely upon the dependence between observations. Combining the flexible mean structure of tree-based ensembles with the dependence structures of traditional spatial statistics models is a promising avenue for future work.

An interesting related analysis would be assessing the effect wildfires have on ozone formation. Doing so would require additional numerical simulations from the WRF-Chem model that exclude wildfire emissions from its inputs. With the data currently available to us, it is impossible to disentangle the effect of wildfires from the other inputs of the WRF-Chem model. This inferential problem is also quite different from the downscaling task which is the focus of this study and may require an entirely different modeling approach. A model that provides excellent downscaling predictions may not be useful for drawing scientific inference.

Another interesting question is whether ozone formation was NOx-limited or VOC-limited. During a wildfire the chemical regime is primarily determined by the amount and conditions of the wildfire fuel, leading to rapid changes in NOx and VOC sensitivity from day to day and even within the course of a day. Understanding this would require a fully separate chemical analysis of air quality conditions in northern California that is beyond both the scope of this paper and the scope of our data, as we lack data on VOC concentrations. One previous study, however, examined the changes that occurred in atmospheric chemistry when wildfire plumes interacted with urban pollution during these fires (Singh et al., 2012).

We also lack information on chlorofluorocarbon (CFC) emissions, which break down ozone and thus influence ozone concentrations. If data on these compounds were available during the study period, we could include them as covariates in an attempt to improve predictions. The absence of data on VOC and CFC emissions does not invalidate the downscaling enterprise. Downscaling models are constructed to combine the available information, whether from observational processes or complex computational models like WRF-Chem, into accurate predictions within a particular domain. They are not scientific models and do not necessarily imply anything about the chemical or physical mechanisms by which pollutants form and move. Using flexible models strictly for downscaling also protects us from biases in the WRF-Chem output. We need not validate the accuracy of the WRF-Chem simulations; we simply rely on the models to learn the relationship between this output and observed ozone concentrations.

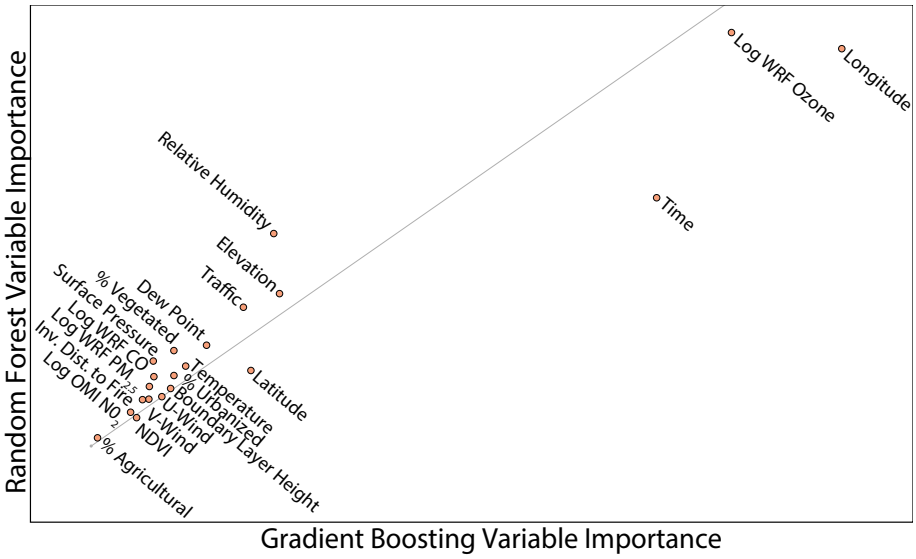


Figure 3.5: Pairwise normalized gradient boosting and random forest variable importance scores for models trained on the full data. The light grey line denotes equal importance in the two models.

Figure 3.5 plots covariate importance scores for random forest and gradient boosting models fit to the full data. Random forest variable importance was calculated as the mean decrease in residual sum of squares resulting from splitting on that covariate averaged across all the trees of the forest. Variable importance for gradient boosting was calculated by permuting that covariate's values and computing the average difference in MSE between predictions made with permuted and un-permuted values (Breiman, 2001). The variable importance scores for the two models were normalized to sum to one for ease of comparison. Most covariates are close to the grey diagonal, which indicates equal importance in the two models. Longitude, WRF-Chem ozone and time were the three most important covariates for both models. WRF-Chem is constructed to estimate atmospheric ozone and so it is not surprising that it has a high importance score. Longitude, latitude and time can proxy for unobserved factors, but also calibrate the effect of other covariates similar to space- or time-varying coefficient models. This calibration may be especially important for the numerical outputs of the WRF-Chem model. Longitude is likely particularly important because it can be used to index many of the significant geographical features of northern California that run approximately North-South, including the coast, San Joaquin Valley, coastal and Sierra Nevada mountain ranges. These geographical features may be associated with important, unobserved

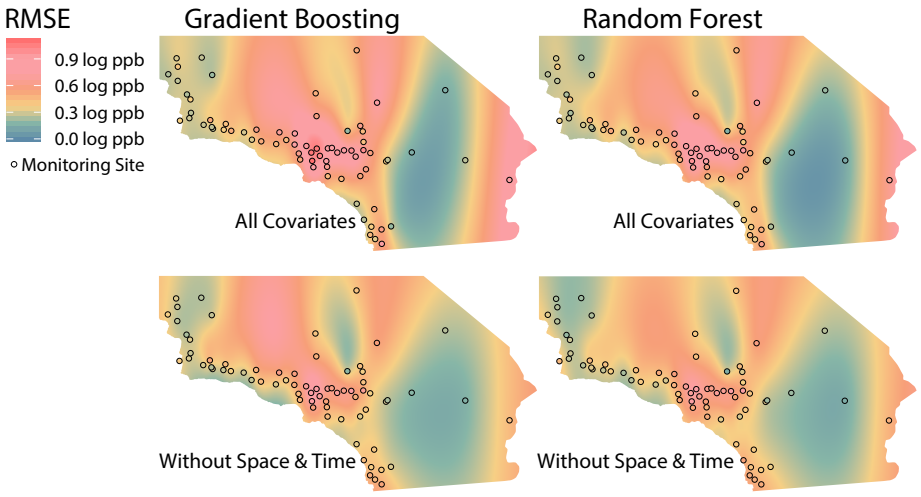


Figure 3.6: The extrapolation error made by gradient boosting and random forest models trained on the 2008 data and predicting ozone during the 2007 southern California fire. The mean RMSE at each location is smoothed throughout the study region using a two-dimensional spline-on-sphere smoother.

information that is not captured by the other covariates including VOC and CFC concentrations.

Neither model when trained on the northern California 2008 wildfire data accurately predicted ozone exposure in southern California in 2007. The predictions from both models had negative $\hat{R}^2$, indicating that their predictions were less accurate (i.e., had higher estimated MSE) than the sample mean of the southern California ozone monitors, which by definition has an $\hat{R}^2$ of 0. In fairness to gradient boosting and random forest, in an out-of-domain prediction problem, the sample mean is unknown, and therefore cannot be used as a prediction rule. When downscaling gradient boosting and random forest models were fit to the 2007 southern California wildfire data, they had LOLO CV errors comparable to the LOLO CV errors reported above. These models accurately downscaled ozone observations during the southern California wildfire (just as they did for the northern California data), but they did not extrapolate well outside of the domain in which they were trained.

This is not surprising and illustrates the proper interpretation of our modeling efforts, which is statistical downscaling (i.e., interpolating) within the observed space-time domain. The substantial decrease in predictive accuracy between within-domain (i.e., downscaling) and out-of-domain (i.e., extrapolating) predictive performance suggests that the relationships between covariates and

ozone exposure differ in space and time and demonstrates the dangers of using a downscaling model for extrapolation. Within the observed space-time domain, space and time can proxy for unobserved spatially-indexed covariates in flexible models like gradient boosting and random forest, improving downscaling predictions, but impeding straightforward extrapolation to different space-time domains where the relationship between these unmeasured covariates and space-time may be different. One referee suggested that the difference in ozone precursors between regions, specifically whether ozone formation is NOx-limited or VOC-limited, is likely one such unobserved, spatially-indexed covariate hindering spatial extrapolation.

As a further check, we repeated this extrapolating procedure excluding latitude, longitude and time from the covariates. The domains of the other covariates were comparable between the two data sets, with the exception of a few low values for surface pressure in the southern California data. Excluding spatial and temporal covariates improved predictive accuracy, reducing LOLO CV $\hat{RMSE}$ from 0.587 to 0.499 for gradient boosting and 0.544 to 0.462 for random forest. As expected, space-time covariates improve downscaling predictions but worsen extrapolating predictions, because the unobserved information indexed by these covariates differs in other regions and at different times.

The scale on which prediction is performed is also important and application specific. In this analysis we performed prediction on the log scale to normalize the variance, but also because it balances large and small errors allowing accurate predictions to be made at both large and small concentrations. We believe this is a natural scale for the intended subsequent applications of our predictions in epidemiological analyses of pollution health effects. Prediction on the untransformed, original scale would more heavily weight predictions at high ozone concentrations. This may be useful for some applications, but we prefer to balance predictive accuracy at low and high concentrations for our application.

The comparison performed here demonstrates that machine learning prediction algorithms, especially ensembles of tree models like gradient boosting and random forest, can accurately downscale ozone concentrations during wildfire events. We believe they would downscale ozone similarly well in the absence of wildfire events. The models we consider here, however, did not accurately extrapolate beyond the space-time domain on which they were trained. This analysis also demonstrates that the choice of evaluation metric is critical to understanding predictive performance.

Metrics that ignore the dependent structure of the data, including k-fold CV, are overly optimistic and unreliable for model selection. LOLO CV is a superior alternative that accounts of the spatial dependence of the data in evaluating model predictive performance, resulting in more reliable estimates of predictive performance.

3.A Appendix

3.A.1 WRF-Chem Simulation Inputs

The WRF-Chem model estimates concentrations of trace species in the atmosphere by simulating meteorological conditions and chemical behavior of atmospheric gases and aerosols (Powers et al., 2017; Pfister et al., 2011b). Simulations for this analysis were conducted with version 3.2 of the WRF-Chem model using the RRTMG longwave and shortwave radiation scheme (Iacono et al., 2008), Grell 3D Cumulus Parameterization (Grell and Dévényi, 2002), Bougeault and Lacarrere (BouLac) TKE boundary layer scheme (Bougeault and Lacarrere, 1989), unified Noah land surface model (Ek et al., 2003) and the Monin-Obukhov (Janjic Eta) Surface Layer scheme (Janjić, 1996). Regional transport of airmasses from regions not included in the WRF-Chem domain were accounted for by using output from the global Model for OZone And Related chemical Tracers (MOZART)-4 chemical transport model for chemical and upon National Centers for Environmental Prediction (NCEP) Final Analysis (FNL) for meteorological spatial and temporal boundary conditions (Pfister et al., 2011a). Other model inputs included a 2008 California anthropogenic emissions inventory supplied by Jeremy Avise at the California Air Resources Board via private correspondence in 2010 and also referenced in Pfister et al. (2011), online biogenic emissions using the Model of Emissions of Gases and Aerosols from Nature (MEGAN) (Guenther et al., 2006) and fire emissions estimated by FINN v1.5 (Wiedinmyer et al., 2011).

3.A.2 CV Estimates of RMSE and R^2

Table 3.2 reports LOLO CV and 10-fold CV estimates of RMSE and R^2 for each of the 10 models.

Table 3.2: Comparison of downscaling ozone prediction models using 10-fold and leave-one-location-out cross-validation for the 2008 California wildfire period (May 6–Sept 26, 2008).

Model	LOLO CV $\hat{RMSE}$ (log ppb)	10-fold CV $\hat{RMSE}$ (log ppb)	LOLO CV $\hat{R}^2$	10-fold CV $\hat{R}^2$
Gradient Boosting	0.228	0.186	0.677	0.784
Random Forest	0.233	0.203	0.661	0.751
Support Vector Machines	0.243	0.216	0.632	0.714
Multivariate Adaptive Regression Splines	0.246	0.229	0.623	0.673
Generalized Additive Model	0.255	0.218	0.596	0.706
Lasso	0.260	0.250	0.581	0.612
Elastic Net	0.260	0.250	0.581	0.612
Linear Model	0.260	0.250	0.580	0.612
k-Nearest Neighbors	0.263	0.241	0.570	0.640
Neural Network	0.266	0.239	0.560	0.652

3.A.3 Tuning Parameters

The tuning parameter value for k chosen by the default tuning procedure in the caret R package for k-nearest neighbor regression provided very poor LOLO predictions. To find a value for k, which denotes the number of neighboring observations averaged for prediction, that gave lower LOLO estimates of prediction error, we evaluated the LOLO CV prediction error with k set at every integer from 1 to 50, selecting 19 as the best choice for k, as it corresponds to the lowest LOLO CV $\hat{RMSE}$. Figure 3.7 graphically depicts the results of this tuning procedure.

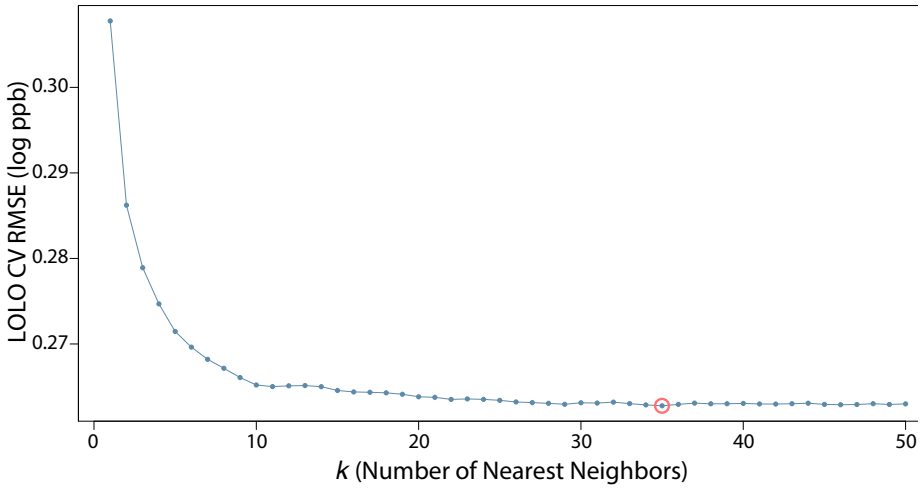


Figure 3.7: LOLO CV $\hat{RMSE}$ for k-nearest neighbor regression models for values of k ranging from 1 to 50. The model with $k = 19$ was selected as best, because it had the lowest LOLO CV $\hat{RMSE}$.

Similarly for GAM, the default caret procedure provided very poor LOLO CV predictions, motivating a similar tuning procedure for the basis dimension for the thin plate regression splines which relate covariates to the outcome. This is argument k to the s() function of the R mgcv package. We evaluated the LOLO CV $\hat{RMSE}$ for the GAM model with basis dimensions ranging from 3 to 15. Three is the smallest basis dimension allowed for a second-order penalty term, which is the default for thin plate regression splines (Wood, 2003). A basis dimension of 3 resulted in the lowest LOLO CV $\hat{RMSE}$, and consequently we selected this as the value of the tuning parameter.

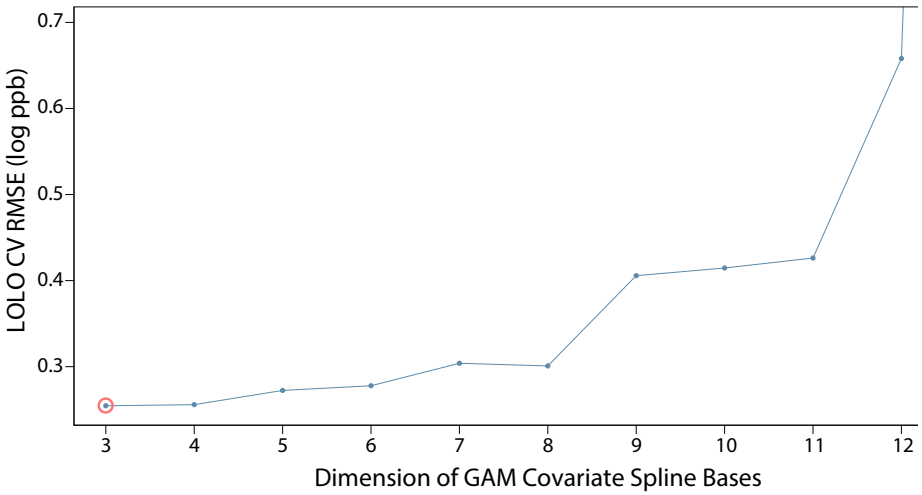


Figure 3.8: LOLO CV $\hat{RMSE}$ for GAM regression models with thin plate regression spline basis dimension varying between 3 and 15. Three was selected as, because it corresponded to the model with the lowest LOLO CV $\hat{RMSE}$.

CHAPTER 4

Treeging

Treeging combines the flexible mean structure of regression trees with the covariance-based prediction strategy of kriging into the base learner of an ensemble prediction algorithm. In so doing, it combines the strengths of the two primary types of spatial and space-time prediction models: (1) models with flexible mean structures (often machine learning algorithms) that assume independently distributed data, and (2) kriging or Gaussian Process (GP) prediction models with rich covariance structures but simple mean structures. We investigate the predictive accuracy of treeging across a thorough and widely varied battery of spatial and space-time simulation scenarios, comparing it to ordinary kriging, random forest and ensembles of ordinary kriging base learners. Treeging performs well across the board, whereas kriging suffers when dependence is weak or in the presence of spurious covariates, and random forest suffers when the covariates are less informative. Treeging also outperforms these competitors in predicting atmospheric pollutants (ozone and $PM_{2.5}$) in several case studies. We examine sensitivity to tuning parameters (number of base learners and training data sampling proportion), finding they follow the familiar intuition of their random forest counterparts. We include a discussion of scaleability, noting that any covariance approximation techniques that expedite kriging (GP) may be similarly applied to expedite treeging.

4.1 Introduction

Spatial or space-time prediction of a quantity based on observations of it at other locations or times (or both) is a common problem in fields as diverse as environmental health, real estate and mining. Formally, the problem is to predict the value of a random field $Y(d)$ at a new spatial or space-time location $d_0 \in \mathcal{D}$ given n observations of the field at other locations, $\mathbf{y}(\mathbf{d}) = [y_1(d_1), ..., y_n(d_n)]'$, where d is the spatial ($\mathcal{D} \subset \mathcal{R}^2$) or space-time ($\mathcal{D} \subset \mathcal{R}^3$) coordinates that index the field.

Ancillary information in the form of covariates $\mathbf{X}(d) = [X_1(d), ..., X_p(d)]$ is generally available, in which case the goal is to predict $Y(d_0) \mid \mathbf{X}(d_0), \mathbf{z}(\mathbf{d})$, where $\mathbf{z}(\mathbf{d}) = \{\mathbf{y}(\mathbf{d}), \mathbf{X}(\mathbf{d})\}$ is the observed data. The conditional expectation $E[Y(d_0) \mid \mathbf{X}(d_0), \mathbf{z}(\mathbf{d})]$ is the usual predictive target, because it is the minimal mean squared error (MSE) prediction of $Y(d_0) \mid \mathbf{X}(d_0), \mathbf{z}(\mathbf{d})$.

If we suppose $Y(d) \mid \mathbf{X}(d)$ is composed of a mean function f of the covariates at d plus an error $\epsilon(d)$ that is independent of $\mathbf{X}(d)$ and has expectation 0, i.e., $E[\epsilon(d)] = 0$, then

$$Y(d) \mid \mathbf{X}(d) = f[\mathbf{X}(d)] + \epsilon(d). \tag{4.1}$$

The second moment of $\epsilon(d)$ encodes the spatial or space-time dependence between values of the random field. In particular, let $\Sigma[d_1, d_2] \coloneqq \text{Cov}[\epsilon(d_1), \epsilon(d_2)]$, where Σ is a positive semidefinite function defining the covariance between $\epsilon(d_1)$ and $\epsilon(d_2)$ (and therefore between $Y(d_1)$ and $Y(d_2)$) for any $d_1, d_2 \in \mathcal{D}$. In this setting, the conditional predictive target is decomposed as the sum of the mean function and the expected conditional error,

$$E[Y(d_0) \mid \mathbf{X}(d_0), \mathbf{z}(\mathbf{d})] = f[\mathbf{X}(d_0)] + E[\epsilon(d_0) \mid \mathbf{z}(\mathbf{d})]. \tag{4.2}$$

It will be useful to characterize alternative prediction techniques based on the assumptions they make regarding these two terms.

The dominant traditional approaches to spatial and space-time prediction, especially for interpolation, use simple mean structures to model f and rely primarily on exploiting the spatial or space-time dependence of the random field for prediction. The mean function is usually assumed to be linear in the covariates, i.e., $f[\mathbf{X}(d_0)] = \mathbf{X}(d_0)'\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathcal{R}^p$. There is a rich literature on covariance estimation in this context, ranging from simple parametric functions of the distance between points (Cressie, 2015b; Gelfand et al., 2010) to sophisticated anisotropic functions (Paciorek and Schervish, 2006; Sang and Huang, 2012). Given an estimate of the covariance function, $\hat{\Sigma}$, prediction proceeds by computing the best linear unbiased predictor (BLUP),

$$\hat{Y}(d_0)|\mathbf{X}(d_0), \mathbf{z}(\mathbf{d}) = \mathbf{X}(d_0)'\hat{\boldsymbol{\beta}} + \hat{\Sigma}_0\hat{\Sigma}(\mathbf{d})^{-1}[\mathbf{y}(\mathbf{d}) - \mathbf{X}(\mathbf{d})\hat{\boldsymbol{\beta}}], \tag{4.3}$$

where $\hat{\Sigma} \coloneqq \hat{\Sigma}(\mathbf{d}, \mathbf{d})$ is the $n \times n$ estimated covariance matrix of the observed data, $\hat{\Sigma}_0 \coloneqq \hat{\Sigma}(d_0, \mathbf{d})$ is the estimated covariance of the prediction location with the locations of the observed data, and $\hat{\boldsymbol{\beta}} = [\mathbf{X}(\mathbf{d})'\hat{\Sigma}(\mathbf{d})^{-1}\mathbf{X}(\mathbf{d})]^{-1}\mathbf{X}(\mathbf{d})'\hat{\Sigma}(\mathbf{d})^{-1}\mathbf{y}(\mathbf{d})$. This prediction procedure is widely known as kriging after Danie Krige (Krige, 1951), the mining engineer who proposed it, and is a special case of the

BLUP for generalized least squares (Goldberger, 1962). It is also the maximum a posteriori (MAP) prediction of $Y(d_0)$ of a Bayesian Gaussian process (GP) model with an uninformative prior on $\boldsymbol{\beta}$ and an equivalent covariance structure. The common types of kriging in the literature (ordinary, universal, with external drift, etc.) amount to differences in which if any covariates are included in $\mathbf{X}(d)$. We remain general in notation but primarily ensivision a universal kriging model with the coordinates and additional covariates in $\mathbf{X}(d)$ (Cressie, 2015a; Gelfand et al., 2010).

While the dependence structure in equation 4.3 is potentially very flexible, the mean structure is quite rigid, and $\mathbf{X}(d_0)'\hat{\boldsymbol{\beta}}$ may be a very poor estimator of the true mean $f[\mathbf{X}(d_0)]$. In particular, assuming f is linear and additive in the covariates is dubious, and the lack of regularization promotes overfitting.

In stark contrast stand a bevy of machine learning algorithms with flexible mean structures that have become popular for spatial and space-time prediction (Watson et al., 2019; Xu et al., 2018; Reid et al., 2015). These models estimate f with a flexible, possibly nonparametric $\tilde{f}$, but usually assume that, observations of the random field are independent of each other conditional on the covariates, i.e., that $\text{Var}[\epsilon(d_1), \epsilon(d_2)] := 0$ for $d_1 \neq d_2$. From this it immediately follows that the expected conditional error in equation 4.2, $E[\epsilon(d_0) \mid \mathbf{z}(\mathbf{d})]$ must be 0, an obviously false assumption in many spatial or space-time contexts. So while $\tilde{f}[\mathbf{X}(d_0)]$ may be flexible enough to approximate $f[\mathbf{X}(d_0)]$, it is a biased estimator of $E[Y(d_0) \mid \mathbf{X}(d_0), \mathbf{z}(\mathbf{d})]$. There are certain exceptions to this characterization in the literature, and we defer a discussion of their merits to section 4.7.

The linear mean structure often assumed in kriging (and many other GP-based procedures) and the working independence assumpion often made by machine learning tools are simply not realistic. However, their complementary strengths suggest they might be combined into an algorithm with flexible mean *and* covariance structures. We take precisely this approach, enriching regression trees with a kriging covariance model, and then ensembling them into a new prediction algorithm which we call *treeging*, a portmanteau of *tree* and *kriging*. Treeging is implemented in a freely available R package at https://github.com/gregorywatson/treeging.

The remainder of this manuscript proceeds as follows. In the next section, we present the treeging algorithm. In sections 4.3 and 4.4, we illustrate its use on spatial and space-time data respectively, including extensive simulations and case studies. Section 4.5 investigates how performance depenends upon tuning parameters, and section 4.6 discusses treeging in the context of large

data sets. Finally, we close with a discussion.

4.2 Treeging

4.2.1 Mean Structure

We define treeging as an ensemble of base learners named *treeges*, which are regression trees enriched with kriging. A regression tree is a heuristically defined model that recursively partitions the covariate space according to the values of covariates, modeling the conditional expectation of $Y(d)$ given $\mathbf{X}(d)$ as

$$E[Y(d)|\mathbf{X}(d)] = f[X(d), \boldsymbol{\theta}] = \sum_{\ell=1}^{L} f_\ell[\mathbf{X}(d), \boldsymbol{\theta}_\ell] 1[\mathbf{X}(d) \in \mathcal{X}_\ell], \tag{4.4}$$

where $\mathcal{X}_1, ..., \mathcal{X}_L$ partition the covariate space, and $f_\ell[\mathbf{X}(d), \boldsymbol{\theta}_\ell]$ gives the conditional expectation $Y(d)|\mathbf{X}(d)$ for observations with covariates that are in $\mathcal{X}_\ell$ (the ℓ-th leave). When a constant leaf model is employed, i.e., when $f_\ell[\mathbf{X}(d), \boldsymbol{\theta}_\ell] = \theta_\ell$, the regression tree corresponds to a linear model with an $n \times L$ design matrix $\mathbf{X}_\tau(\mathbf{d})$ with a column for each leaf in the tree. The elements of each row of $\mathbf{X}_\tau(\mathbf{d})$ are 0 except for a 1 at the element corresponding to the leaf of the tree in which the covariates associated with that row reside, i.e., $[\mathbf{X}_\tau(\mathbf{d})]_{i\ell} = 1[\mathbf{X}(d_i) \in \mathcal{X}_\ell]$. Substituting $\mathbf{X}_\tau(d)$ for $\mathbf{X}(\mathbf{d})$ in equation 4.3 provides an alternative BLUP,

$$\tilde{Y}(d_0)|\mathbf{X}_\tau(d_0), \mathbf{z}(\mathbf{d}) = \mathbf{X}_\tau(d_0)'\hat{\boldsymbol{\beta}}_\tau + \hat{\Sigma}_0\hat{\Sigma}(\mathbf{d})^{-1}[\mathbf{y}(\mathbf{d}) - \mathbf{X}_\tau(\mathbf{d})\hat{\boldsymbol{\beta}}_\tau]. \tag{4.5}$$

The superlative designation of the BLUP as *best* suggests uniqueness but obscures the fact that the BLUP is specific to the choice of mean and covariance functions. Selecting an alternate mean function in equation 4.5 thus provides an alternate BLUP.

Like regression trees, treeges exhibit low bias and high variance in their predictions on account of their tendency to overfit. Random forest improves upon the predictive performance of a single tree by combining many in an ensemble using bootstrap samples and random subsets of the covariates to reduce the variance of the predictions (Breiman, 2001). Motivated by this straightforward yet highly effective strategy, we ensembled treeges by training each one on a random subset of the training data sampled without replacement, combining their predictions via a simple average. The size of the subsample is a parameter that can be tuned, and by default we set to be $m = \text{round}(0.632n)$, because this is approximately the expected proportion of observations that appear in a bootstrap

Table 4.1: True and estimated mean and dependence structures for spatial or space-time prediction.

	Mean Structure	Dependence Term
Truth	$f[\mathbf{X}(d_0)]$	$E[\epsilon(d_0) \mid \mathbf{z}(\mathbf{d})]$
Kriging	$\mathbf{X}(d_0)'\hat{\boldsymbol{\beta}}$	$\hat{\Sigma}_0\hat{\Sigma}(\mathbf{d})^{-1}[\mathbf{y}(\mathbf{d}) - \mathbf{X}(\mathbf{d})\hat{\boldsymbol{\beta}}]$
1 Regression Tree	$\mathbf{X}_\tau(d_0)'\hat{\boldsymbol{\beta}}_\tau$	0
1 Treege	$\mathbf{X}_\tau(d_0)'\hat{\boldsymbol{\beta}}_\tau$	$\hat{\Sigma}_0\hat{\Sigma}\left(\mathbf{d}\right)^{-1}\left[\mathbf{y}(\mathbf{d}) - \mathbf{X}_\tau(\mathbf{d})\hat{\boldsymbol{\beta}}_\tau\right]$
Random Forest	$\frac{1}{B}\sum_{b=1}^{B}\mathbf{X}_\tau^{(b)}(d_0)'\hat{\boldsymbol{\beta}}_\tau^{(b)}$	0
Treeging	$\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_\tau^{(t)}(d_0)'\hat{\boldsymbol{\beta}}_\tau^{(t)}$	$\frac{1}{T}\sum_{t=1}^{T}\hat{\Sigma}_0^{(t)}\hat{\Sigma}\left(\mathbf{d}^{(t)}\right)^{-1}\left[\mathbf{y}(\mathbf{d}^{(t)}) - \mathbf{X}_\tau^{(t)}\left(\mathbf{d}^{(t)}\right)\hat{\boldsymbol{\beta}}_\tau^{(t)}\right]$

sample of size n (Efron and Tibshirani, 1997). We also mimic random forest in considering a subset of the covariates at each split. The resulting treeging prediction for $Y(d_0)$ is

$$\tilde{Y}(d_0) \mid \mathbf{X}(d_0), \mathbf{z}(\mathbf{d}) = \frac{1}{T}\sum_{t=1}^{T}\left\{\mathbf{X}_\tau^{(t)}(d_0)'\hat{\boldsymbol{\beta}}_\tau^{(t)} + \hat{\Sigma}_0^{(t)}\hat{\Sigma}\left(\mathbf{d}^{(t)}\right)^{-1}\left[\mathbf{y}(\mathbf{d}^{(t)}) - \mathbf{X}_\tau^{(t)}\left(\mathbf{d}^{(t)}\right)\hat{\boldsymbol{\beta}}_\tau^{(t)}\right]\right\}, \quad (4.6)$$

where $t = 1, ..., T$ indexes treeges, and a superscript (t) indicates the data subset, design matrix or estimate for the t-th treege.

Table 4.1 lists the mean and dependence structures for kriging, random forest, a single treege and treeging. The limitations of kriging and random forest are apparent as is the manner in which they have been combined into treeging.

4.2.2 Covariance Estimation

Because a treege uses the same covariance structure as kriging, it accommodates any covariance estimation procedure that works for kriging. This makes available the wide variety of covariance structures developed for kriging, including procedures that iteratively estimate the mean and covariance, and importantly allows for the incorporation, of domain specific knowledge (Stein, 2005; Lloyd, 2010; Boisvert et al., 2009). It is also possible to use different covariance function estimators for different treeges within a treeging ensemble. The default covariance estimation strategy used

for the results in this manuscript employs spherical covariance functions. Briefly, for each treege we fit a spherical variogram function to the empirical variogram of the regression tree residuals to estimate the parameters of a spherical covariance function. For space-time data, we fit separable spherical covariance functions to the spatial and temporal distances. The details of these procedures may be found in the appendix. One convenient feature of the spherical covariance function is that covariance decreases to 0 at a finite distance unlike other common choices, such as the exponential function. This can be exploited for computational expediency as we discuss in section 4.6.

4.2.3 Treeging Algorithm

Algorithm 1 describes the treeging algorithm in pseudocode for predicting a vector of outcomes, $\mathbf{y}(\mathbf{d}_0)$. For each of the T treeges, a random subsample of size round(np) of the observed data is used to train a regression tree. Using an optimization procedure, the parameters of a spherical semi-variogram, $\boldsymbol{\varphi}$ are fit to the residuals of the regression tree, $\mathbf{e}_\tau^{(t)}\left(\mathbf{d}^{(t)}\right)$, and their pairwise distances, $h_{ij} = \|d_i^{(t)} - d_j^{(t)}\|$, for $i, j = \{1, ..., \text{round}(np)\}$. The estimated spherical parameters $\boldsymbol{\varphi}^{(t)}$ provide an estimate of the covariance function, $\hat{\Sigma}^{(t)}$. Using this covariance function and the regression tree design matrix $\mathbf{X}^{(t)}(\mathbf{d})$, the treege prediction can be computed according to equation 4.5. Averaging the predictions of the T treeges yields the treeging prediction, $\tilde{\mathbf{y}}(\mathbf{d}_0)$.

4.3 Spatial Treeging

We investigated the suitability of treeging for spatial interpolation by comparing it to random forest, kriging and a kriging ensemble in a series of simulations and two case studies. The kriging ensemble is constructed using the same subsampling and ensembling procedures as treeging, but with a linear mean structure for the base learners rather than the tree-based mean structure of treeges. Kriging ensembles are not commonly used, however, we included it in the comparison to discern the predictive gains or losses associated with treeging's flexible mean structure from those of ensembling.

We used R^2 as our primary measure of predictive accuracy, which is the ratio of the mean

Algorithm 1 Treeging

procedure TREEGING($\mathbf{z}(\mathbf{d}), \mathbf{d}_0, \mathbf{X}_0(\mathbf{d}_0), p, T$)

  for $t \leftarrow 1$ to T **do**

    $\mathbf{z}\left(\mathbf{d}^{(t)}\right) \leftarrow$ A random subset of size round(np) from $\mathbf{z}(\mathbf{d})$

    $\left\{\mathbf{X}_\tau^{(t)}(d), \boldsymbol{\beta}_\tau^{(t)}\right\} \leftarrow$ Train a regression tree on $\mathbf{z}\left(\mathbf{d}^{(t)}\right)$

    $\mathbf{e}_\tau^{(t)}\left(\mathbf{d}^{(t)}\right) \leftarrow \mathbf{y}\left(\mathbf{d}^{(t)}\right) - \mathbf{X}_\tau^{(t)}\left(\mathbf{d}^{(t)}\right)' \boldsymbol{\beta}_\tau^{(t)}$

    $\gamma^{(t)}\left(h; \boldsymbol{\varphi}^{(t)}\right) \leftarrow$ Fit spherical semi-variogram to $\mathbf{e}_\tau^{(t)}\left(\mathbf{d}^{(t)}\right)$ and pairwise distances

    $\hat{\Sigma}_{ij}^{(t)} \leftarrow \Sigma\left(d_i^{(t)}, d_j^{(t)}; \boldsymbol{\varphi}^{(t)}\right), i, j = 1, ..., \text{round}(np)$

    $\hat{\Sigma}_0^{(t)} \leftarrow \Sigma\left(\mathbf{d}_0, \mathbf{d}^{(t)}; \boldsymbol{\varphi}^{(t)}\right)$

    $\tilde{\mathbf{y}}^{(t)}(\mathbf{d}_0) \leftarrow \mathbf{X}_\tau^{(t)}(d_0)\hat{\boldsymbol{\beta}}_\tau^{(t)} + \hat{\Sigma}_0^{(t)}\hat{\Sigma}^{(t)^{-1}}\left[\mathbf{y}\left(\mathbf{d}^{(t)}\right) - \mathbf{X}_\tau^{(t)}\left(\mathbf{d}^{(t)}\right)\hat{\boldsymbol{\beta}}_\tau^{(t)}\right]$

  end for

  $\tilde{\mathbf{y}}(\mathbf{d}_0) \leftarrow \frac{1}{T}\sum_{t=1}^{T} \tilde{\mathbf{y}}^{(t)}(\mathbf{d}_0)$

end procedure

squared prediction error to the variance,

$$R^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4.7}$$

where $\bar{y}$ is the mean of $\mathbf{y}$. This is often interpreted as the proportion of the variation in $\mathbf{y}$ explained by predicted values $\hat{\mathbf{y}}$, but the R^2 can be negative if it is a worse predictor than $\bar{y}$, which is quite possible in out-of-sample prediction in which case the sample mean is unknown. In the simulation setting, the true R^2 can be computed. In the case studies, we estimate the R^2 using 10-fold cross-validation (CV), denoting the estimate as $\hat{R}^2_{CV}$.

4.3.1 Spatial Simulation Study

We simulated 1,029 spatial random fields across a variety of covariate effect sizes and levels of spatial dependence. In each case, the field was generated from a Gaussian process with a mean function depending on 3 covariates, one with a linear effect, one a threshold effect, and one a sigmoidal effect to mimic what is often encountered in environmental pollution studies. The mean function was scaled up or down using an effect size multiplier to elucidate the relationship between predictive performance and covariate effect size. The appendix provides the mean function and simulation

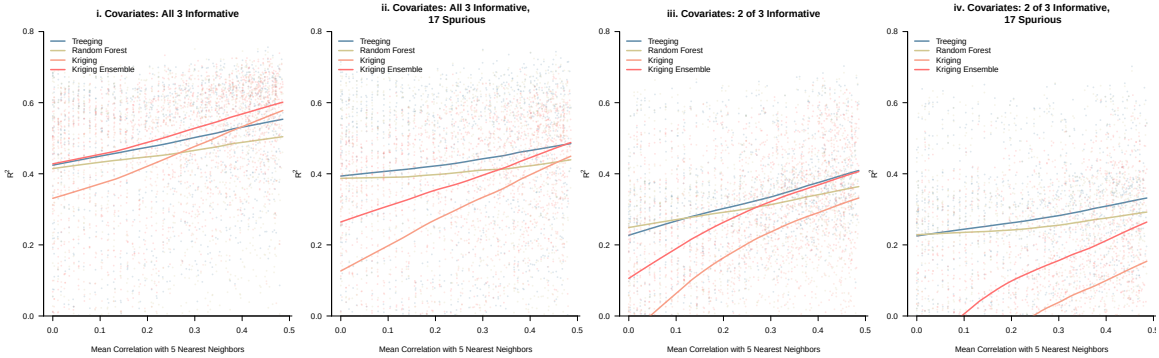
parameters. We randomly selected 100 locations in a spatial domain as the observed data on which to train the models, and assessed their predictive accuracy over a 21 $\times$ 21 grid of points across the spatial field under 4 different covariate scenarios: (i) the 3 informative covariates, (ii) the 3 informative covariates plus 17 spurious covariates, (iii) 2 of the 3 informative covariates selected at random, and (iv) 2 of the 3 informative covariates plus the 17 spurious covariates. In addition, the spatial coordinates were included as covariates in all cases. These 4 scenarios were constructed to reflect some of the common challenges encountered in spatially dependent data. Often there are potentially informative ancillary data, but which covariates are informative is unknown a priori, and not all relevant covariates may be available.

Figure 4.1A depicts the predictive R^2 for each model under each covariate scenario as a function of the spatial dependence, which is assessed as the mean correlation between the prediction locations and the five nearest training data points. The results are smoothed using a LOWESS curve to facilitate comparison. Not surprisingly, there is a general increasing trend in R^2 as spatial dependence increases for all models across every scenario, as increased spatial dependence corresponds to additional information with which to predict $Y(d)$ over the unobserved grid. This increase is less pronounced for random forest, which ignores dependence, although it is able to make some gains, because the spatial coordinates are included as covariates.

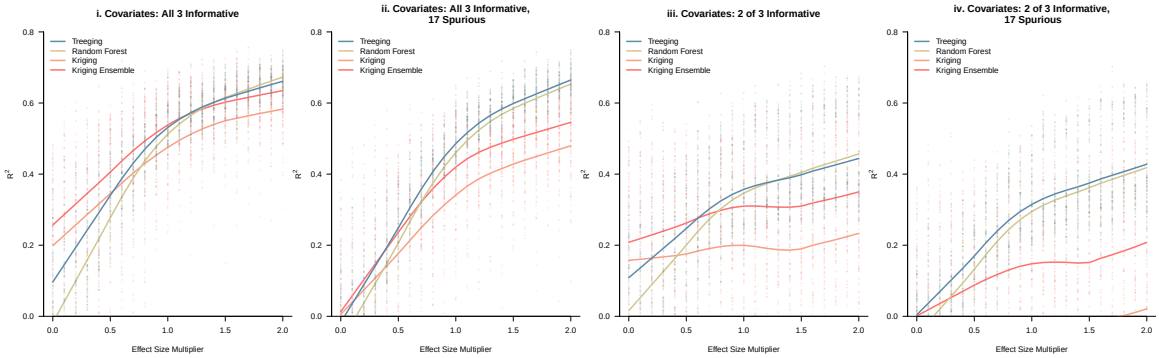
The kriging ensemble is the best performing model when trained on all three informative coviarates without any spurious covariates (scenario i), with treeging second best until it is overtaken by kriging at high levels of dependence. In the presence of spurious coviarates (scenario ii), treeging is the best model with the kriging ensemble performing comparably at high levels of dependence. In scenarios iii and iv, kriging and the kriging ensemble perform very poorly. Across all four scenarios, random forest performs respectably, but treeging outpferforms it across the board with a small margin between them at low spatial correlation that increases as dependence gets stronger.

Figure 4.1B plots the marginal effect size multiplier against R^2 for the same set of simulations. Across all four covariate scenarios there is again a consistent relationship between treeging and random forest. When the effect size multiplier is 0, the covariate effects are 0 and random forest performs poorly. Treeging performs noticeably better, although it is also not particularly accurate. As covariate effect strengths increase, the R^2 for both treeging and random forest increases rapidly

A. Predictive Accuracy as a Function of Spatial Dependence



B. Predictive Accuracy as a Function of Covariate Effect Size



C. Most Accurate Model as a Function of Spatial Dependence and Covariate Effect Size

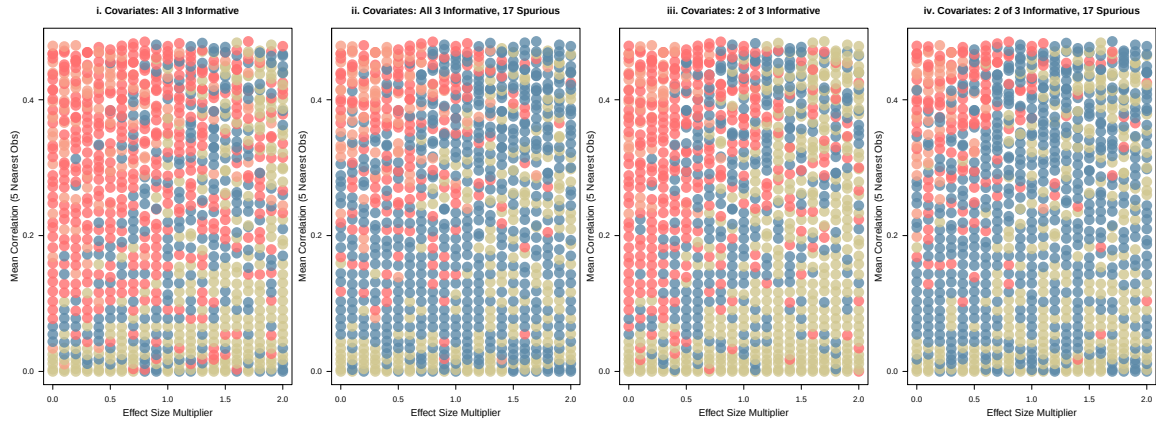


Figure 4.1: Spatial Simulation Model Comparison: Predictive accuracy (R^2) as a function of (A.) spatial dependence, (B.) covariate effect size, and (C.) both. In (C.) the color of each dot depicts the highest R^2 model for the simulation at that combination of the effect size multiplier and mean correlation.

with random forest gradually closing the gap between them. Kriging and ensembled kriging perform fairly well when trained on all the informative covariates (scenarios i and ii) with the ensemble generally superior. At 0 effect size they have higher R^2 than treeging and random forest, but are overtaken as effect size increases. In scenarios iii and iv, however, they perform very poorly with R^2 decreasing as covariate effect size increases. Despite this additional predictive information, the mean structure of kriging is apparently too rigid when an informative covariate is withheld.

Figure 4.1C plots effect size against mean correlation with the color of each dot depicting the best (i.e., highest R^2) model for that simulation. At low effect sizes (the left edge), kriging and ensembled kriging are generally best except when there is very little or no spatial dependence. When there is low dependence (the bottom edge), random forest and treeging are best. With the exception of scenario i, when there is substantial spatial dependence and effect size, treeging tends to be best.

4.3.2 Spatial Case Studies

We compared the predictive accuracy of the same set of models on daily 8-hour maximum average ozone and 24-hour average fine particulate matter ($PM_{2.5}$) concentrations during wildfires in California in 2008 (Watson et al., 2019). Data were collected from 100 ozone and 45 $PM_2.5$ monitors over 118 and 120 days respectively. In addition to these data, 15 covariates including atmospheric weather data, land use information and chemical transport model predictions were collected along with the spatial coordinates. These covariates and their sources are listed in the appendix. To compare their accuracy at strictly spatial interpolation, we considered each day of the data as a distinct data set and estimated predictive accuracy using 10-fold CV, taking the average of each day's 10 folds as the estimated $\hat{R}^2_{CV}$ for that day. In Section 4.4.2 we evaluate interpolation on the entire data set in our discussion of space-time treeging, which is a more realistic approach to interpolation on these data.

Figure 4.2 depicts the average $\hat{R}^2_{CV}$ on each day for each of the models on $PM_{2.5}$ and ozone as box and whisker plots. For both pollutants, treeging was the most accurate model (ozone $\hat{R}^2_{CV}$: 0.62, $PM_{2.5}$ $\hat{R}^2_{CV}$: 0.58), performing slightly better than random forest (ozone $\hat{R}^2_{CV}$: 0.61, $PM_{2.5}$ $\hat{R}^2_{CV}$: 0.56). For ozone, the kriging models are somewhat less accurate on average (kriging $\hat{R}^2_{CV}$: 0.43, ensemble $\hat{R}^2_{CV}$: 0.48), and worryingly have substantially higher variance. For $PM_{2.5}$,

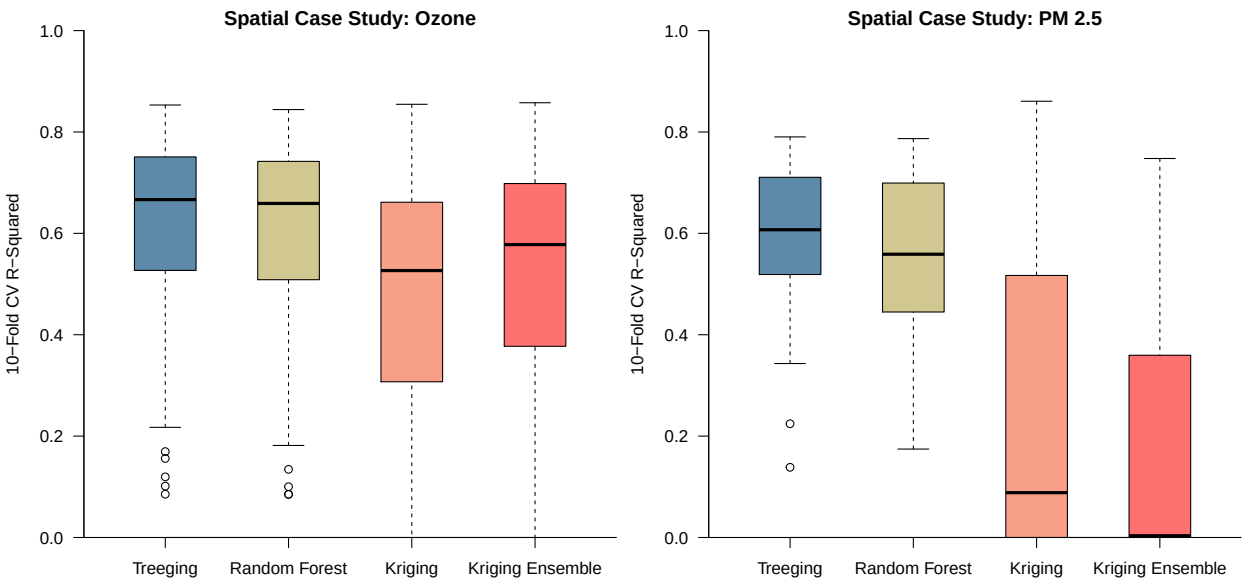


Figure 4.2: Spatial Case Studies Model Comparison: 10-fold CV estimates of R^2 were produced for each day, treating each day as an independent spatial data set. The resulting distribution of estimates is summarized for each model as a boxplot, with the darkened central line indicating the median $\hat{R}^2_{CV}$ across all days for that model.

this variation is even greater, and the CV $\hat{R}^2_{CV}$ is also much lower (kriging $\hat{R}^2_{CV}$: 0.00, ensemble $\hat{R}^2_{CV}$: -0.24). These results suggest that kriging models may be less stable than models with built-in feature selection like treeging and random forest.

4.4 Space-Time Treeging

Space-time treeging considers prediction of a random field $Y(d)$ indexed over a 3-dimensional space-time domain, $\mathcal{D} \subset \mathcal{R}^3$. In the context of space-time data, multiple types of prediction are possible (Watson et al., 2020b). We consider here the suitability of treeging for space-time interpolation, a common type of space-time prediction in which repeated measurements are taken at a set of locations and the predictive goal is to predict the entire time series at new spatial locations. This specific predictive target is motivated by typical applications in environmental pollution studies, which require the estimation of pollutant concentration as a potential source of exposure at unobserved spatial locations (Watson et al., 2020b).

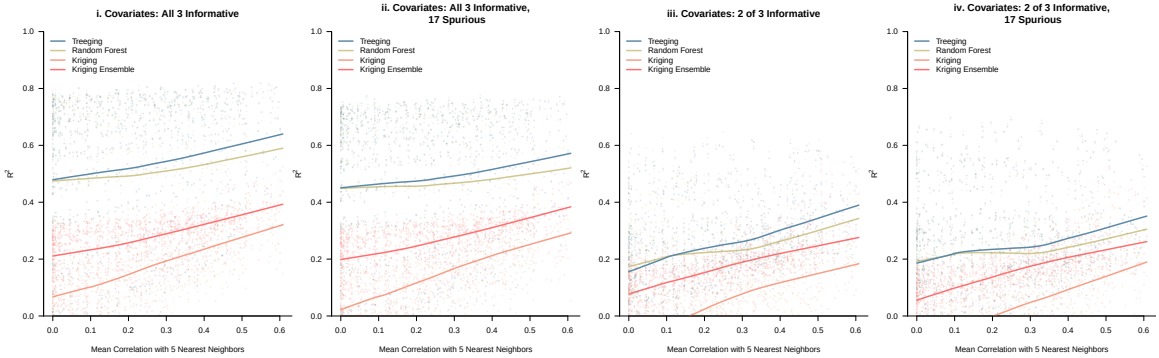
4.4.1 Space-Time Simulation Study

We conducted a second battery of simulations to investigate the predictive accuracy of space-time treeging. The random field $Y(d)$ was sampled as a Guassian process over a time series of length 30 at each of 40 training locations and an 11 $\times$ 11 grid of test locations. The spatial dependence, temporal dependence and covariate effect sizes were varied to investigate their impact on predictive accuracy. The precise details and the functional forms of the covariates may be found in the appendix. Similarly to the spatial simulation study described above, we considered four covariate scenarios.

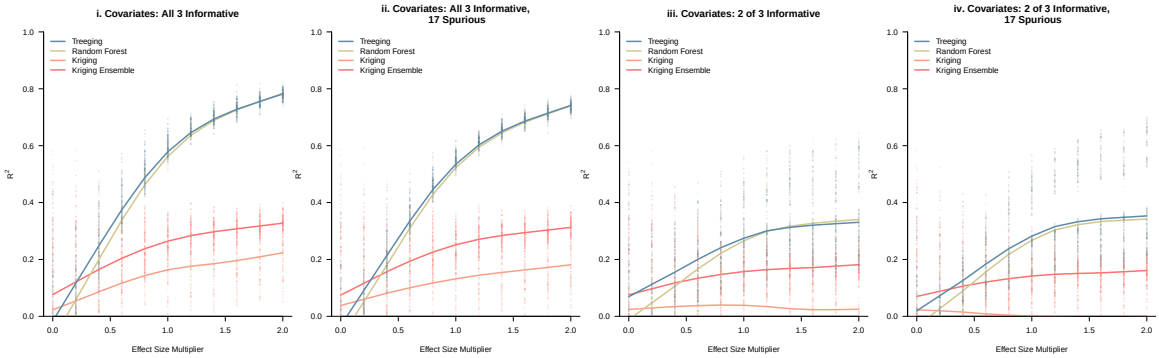
Figure 4.3 shows the results of the simulation study. Figure 4.3A depicts predictive R^2 as a function of the mean correlation between test points and the five training points with which they are most highly correlated. The functional relationship between mean correlation and R^2 is estimated using a LOWESS curve for each of the four models. In all four of the covariate scenarios, treeging (blue) and random forest (green) have nearly identical R^2 when the mean correlation is 0, i.e., when there is no dependence, which is the left hand extreme of the plots. As the mean correlation increases, the R^2 for random forest increases slightly, while that of treeging improves more rapidly, resulting in increasing divergence between these two curves. Treeging better exploits the additional information in the increasing dependence between training and test points, because it expliticly models this dependence, while random forest assumes there is none. Kriging (orange) and ensembled kriging (red) also improve as the mean correlation increases, since they too model the space-time dependence. Kriging is consistently below the kriging ensemble, which is not surprising given the well known improvements in predictive accuracy associated with ensembling base learners. As the dependence increases, the kriging ensemble overtakes random forest in each of the covariate scenarios around a mean correlation of 0.2. At sufficiently high dependence in the spurious covariate scenarios (ii and iv), its performance is comparable to treeging.

Figure 4.3B depicts predictive R^2 as a function of covariate effect size for the same set of simulations. Again kriging lags the kriging ensemble by a consistent amount. When the effect size multiplier is 0, the covariates have no effect on the outcome, aside from possible marginal associations with space or time resulting from the space-time dependence, random forest has the worst R^2 in all four scenarios, because it relies exclusively on the relationship between the covariates and the outcome. Not surprisingly it is at this extreme that treeging excels random forest by the

A. Predictive Accuracy as a Function of Spatial Dependence



B. Predictive Accuracy as a Function of Covariate Effect Size



C. Most Accurate Model as a Function of Spatial Dependence and Covariate Effect Size

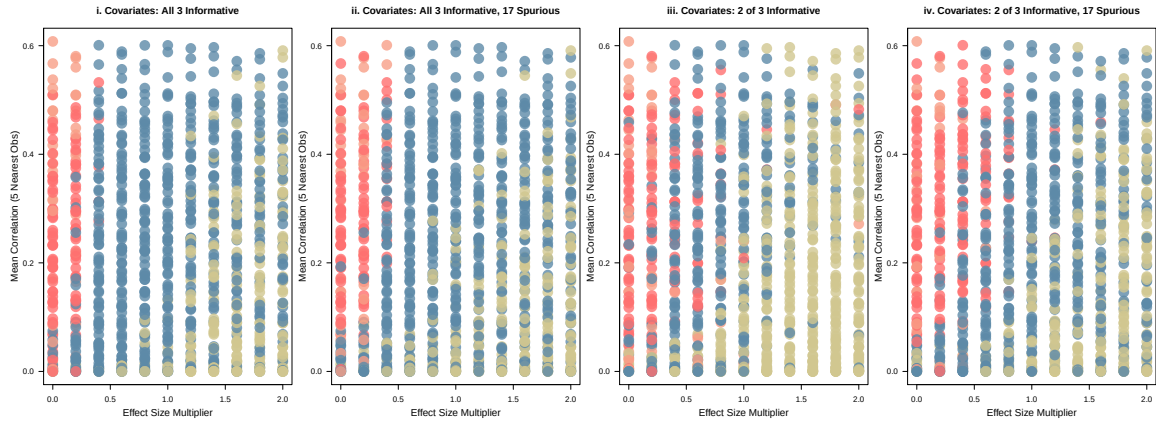


Figure 4.3: Space-Time Simulation Model Comparison: Predictive accuracy (R^2) as a function of (A.) spatial dependence, (B.) covariate effect size, and (C.) both. In (C.) the color of each dot depicts the highest R^2 model for the simulation at that combination of the effect size multiplier and mean correlation.

largest margin. It is also at this extreme that the kriging models tend to predict better than treeging. It would seem that treeging's flexible mean structure hurts its accuracy in the absence of informative covariates, likely on account of the greater potential for overfitting. As the effect size multiplier increases, the covariates exert a stronger effect, treeging and random forest surpass the kriging models on the strength of their flexible mean structures, and random forest gradually closes the gap with treeging as the information in the covariates dominates that of the space-time dependence.

Figure 4.3C combines the marginal plots of figures 4.3A and 4.3B, plotting effect size against mean correlation with each simulation as a single dot. The color of the dot indicates which of the four models had the highest R^2 for that simulation. While this depiction ignores the margin of victory by which a model is best, it nonetheless offers a useful graphical summary of model performance. In all four covariate scenarios, the kriging models tend to perform best at very low effect sizes (the left edge of the plots). Random forest, and to a lesser extent, treeging tend to be best at very low mean correlation (the bottom edge of the plots). In between, treeging tends to dominate across a wide swath of the plots. This nicely illustrates the benefits of treeging—while it may not be the best model in all cases, it is the generally the best when both the mean structure and dependence structures are informative and competitive when only one is.

4.4.2 Space-Time Case Studies

We evaluated the space-time predictive performance of treeging using the same $PM_{2.5}$ and ozone data used in section 4.3.2 for spatial case studies. To provide insight into the potential impact of time series length, we estimated predictive accuracy on increasingly long subsets of the data from one to eight weeks. We used 10-fold location-based CV to estimate predictive accuracy in which the monitor locations were partitioned into 10 folds, so that all observations taken at a particular spatial location fall in the same fold. This is critical to avoid overly optimistic estimates of spatial interpolation prediction error (Watson et al., 2020b).

Figure 4.4 graphically depicts the $\hat{R}^2_{CV}$ case study results. For both $PM_{2.5}$ and ozone, treeging is the best performing model, followed by random forest then ensembled kriging and lastly kriging. For $PM_{2.5}$, the difference in $\hat{R}^2_{CV}$ between treeging and random forest is very slight. For ozone, however, treeging does provide a substantial improvement over random forest. This suggests that

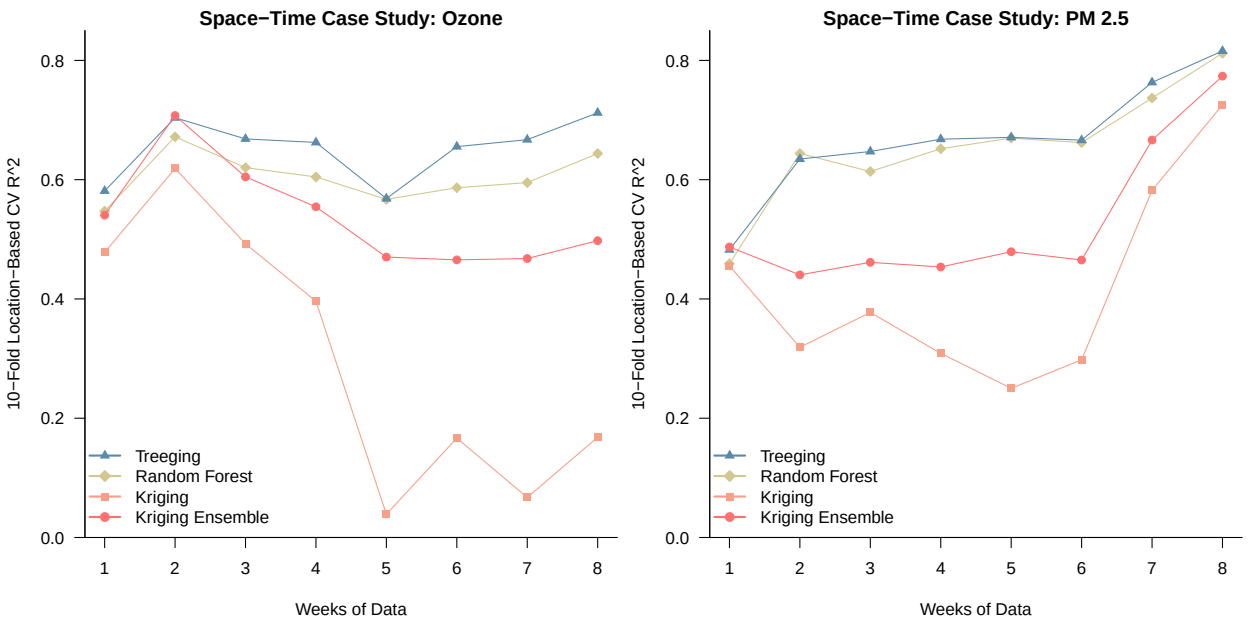


Figure 4.4: Space-Time Case Studies Model Comparison.

the covariates may be more informative for interpolating $PM_{2.5}$, similar to simulation scenario iii, while there may be informative covariates for ozone that are not contained in this data, as in simulation scenario iv.

The $\hat{R}^2_{CV}$ for kriging and to a lesser extent ensembled kriging are both lower and more highly variable than treeging and random forest in figure 4.4. The accuracy of kriging could potentially be improved with the use of a more sophisticated covariance estimation procedure. This may benefit kriging relative to random forest, but a more accurate covariance model would also benefit treeging, which can use any covariance structure that kriging can. The greater variability of kriging estimates reflects the potential for overfitting with a rigid mean structure that lacks a mechanism for feature selection. This is mitigated to some extent by the model averaging in the ensemble, but there is still greater variability than the tree-based ensembles.

4.5 Parameter Tuning

The two primary tunable parameters of treeging are the number of treeges and the subsample proportion. Figures 4.5 and 4.6 depict predictive R^2 (simulation) or $\hat{R}^2_{CV}$ (case studies) for a variety of values of these parameters. In all cases, there is a substantial improvement in predictive

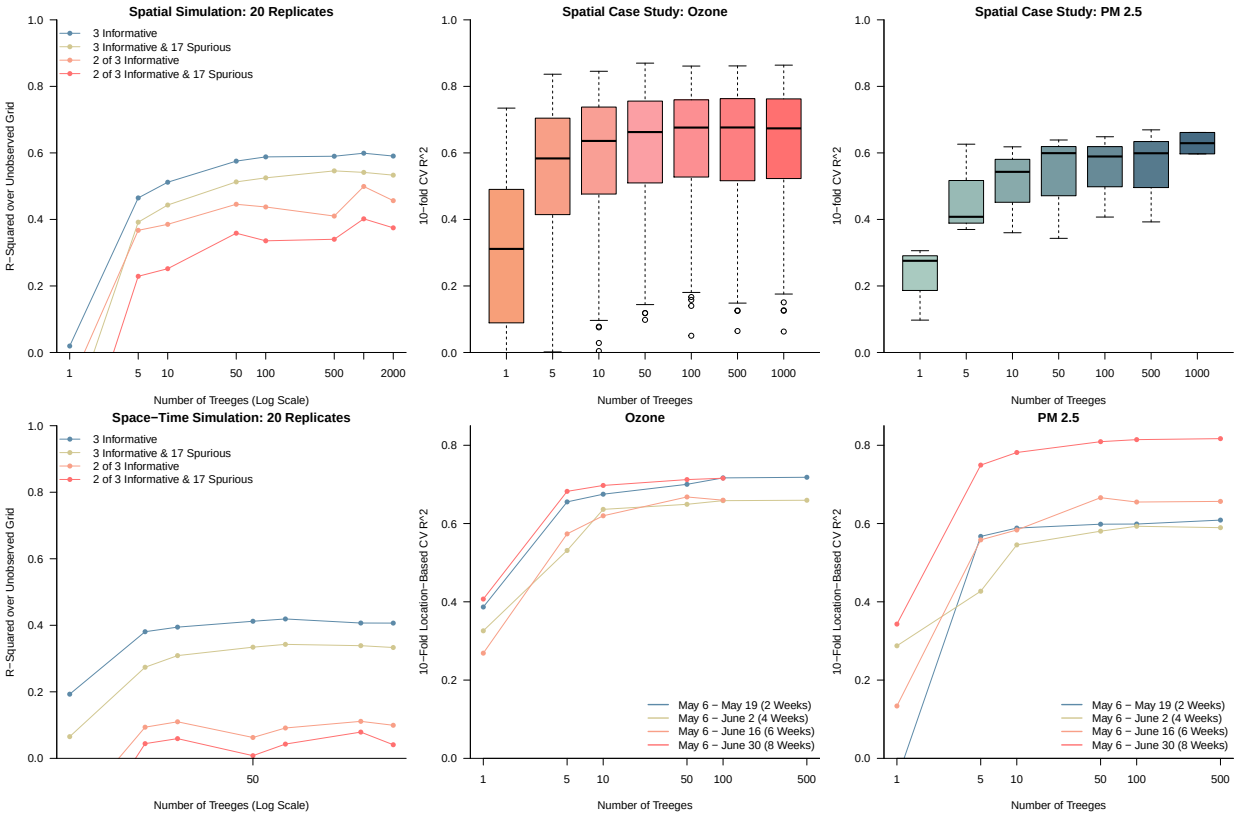


Figure 4.5: Tuning the Number of Treeges

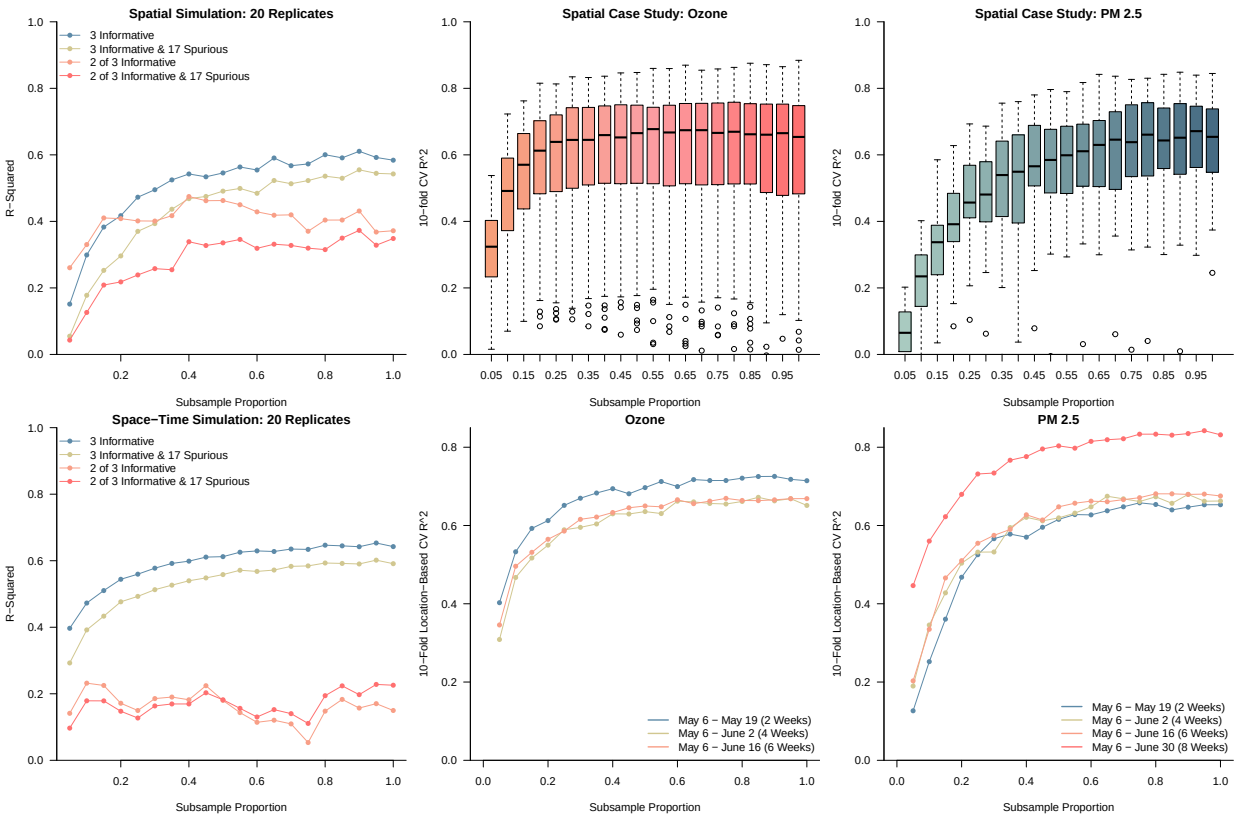


Figure 4.6: Tuning the Subsample Proportion

accuracy from 1 to 5 treeges, after which it generally increases until leveling off at 50 treeges. In some cases there may be evidence of subsequent improvement, but based on these results we set 50 as the default number of treeges. There is, of course, no guarantee this choice will be optimal for a particular problem. Computing time increases with the number of treeges, which may be an important consideration in the context of medium or large data.

Similarly, predictive accuracy increases initially with subsample proportion p and levels off. In both the spatial and space-time case studies, it seems to level off at a higher proportion for $PM_{2.5}$ than for ozone. In the spatial simulations there may be some evidence of an interesting divergence between scenarios in which all the informative covariates are observed (i and ii) and those in which an informative covariate is withheld (iii and iv). There appears to be a decrease in predictive accuracy for iii and iv for $p > 0.75$, whereas accuracy continues to increase for i and ii. It may be that higher subsample proportions can exacerbate the effects of overfitting when the covariates are misspecified. These results reinforce our default choice of $p = 0.632$, motivated

by 0.632 being the expected proportion of the observations that appear in a bootstrap sample. There is a computational discount on the order of p^3, which again may be relevant for medium or large data.

4.6 Large Data Set Approximations

Treeging is computationally intensive. However, it is feasible whenever kriging is. The computational complexity of kriging is $o(n^3)$, making each treege approximately $o(p^3n^3)$ and a naive, sequential implementation of treeging $o(Tp^3n^3)$, where p is the subsample proportion and T is the number of treeges in the ensemble. Considerable gains can be made by parallelization, since treeges are fit independently.

Additional gains can be made by exploiting sparsity in $\mathbf{X}_\tau(\mathbf{d})$ and spherical covariance matrices. The treege design matrix $\mathbf{X}_\tau(\mathbf{d})$ has one nonzero element per row, a 1 at the column corresponding to the leaf in which the covariates associated with that row reside. Consequently, only only L of its nL entries are nonzero, where L is the number of leaves in the treege. The estimated covariance matrix $\hat{\Sigma}$ also tends to be sparse when a spherical covariance model is used, because the spherical covariance function decreases to 0 at a finite distance. Multiplying sparse matrices is generally substantially faster than dense matrices. The treeging R package uses the sparse matrix multiplication routines of the Basic Linear Algebra Subprograms (BLAS) via the Rcpp package (Dongarra, 2002; Eddelbuettel et al., 2011).

Treeging can also take advantage of strategies that have been developed to improve the scaleability of kriging or Bayesian Gaussian process regression, such as Nearest-Neighbor Gaussian Processes (Datta et al., 2016).

4.7 Discussion

Treeging is an effective, new tool for spatial and space-time prediction. By enriching the flexible mean structure of regression tree base learners with the covariance structures of Gaussian process prediction and ensembling them, it combines the strengths of the two major classes of space-time prediction algorithms. While it is not the optimal model in all scenarios, it is clearly the best

performing overall, whereas kriging suffers when dependence is weak or the mean structure is misspecified and random forest suffers when the covariates are uninformative.

Treeging offers clear advantages over alternatives that often rely on assumptions of smoothness and structural encoding of covariate effects (Berrocal et al., 2010; Szpiro et al., 2010). These do not offer the black box benefits of ensembling strategies that automatically learn regression surfaces, requiring subject level knowledge on the number of basis functions and how to encode covariates. The generative adversarial networks that have recently been applied to spatial interpolation problems (Zhu et al., 2020; Manepalli et al., 2020; Watson et al., 2020a; Gao et al., 2020) similarly require some level of subject knowledge, with expected performance of deep learning networks depending on a large number of tuning structures.

Treed Gaussian processes (TGP) (Gramacy and Lee, 2008) and RF-GLS (Saha et al., 2021) are the most closely related approach to treeging in the literature. TGP is a Bayesian regression tree with a GP at each leaf. This provides a very flexible probability model, but practically does not scale even to small air pollution space-time data sets like the approximately 13,000 observation ozone case study considered here, and convergence of MCMC over the space of regression trees remains challenging. It is likely and the goal of future work to show that treeging approximates the MAP (maximum a posteriori) under a particular, restrictive version of TGP. Work on RF-GLS explores some of the theoretical considerations related to treeging.

A number of extensions and improvements to treeging present themselves as useful avenues for future work. Naive ensembling could be extended to include a formalized notion of additivity and a formal optimization strategy, motivated by the well-known successes of boosting including on space-time air pollution problems (Watson et al., 2019; Reid et al., 2015). Localized subsampling may provide a useful alternative for treege construction, pariticularly in the precense of training data sampling bias. Random forest variable importance metrics may provide insight into the contribution of particular covariates within a treeging ensemble. One could also explore bootstrap strategies for uncertainty quantification and possibly investigate the role of out-of-bag (OOB) error as estimator of interpolation error. Some work in IID cases has been published making use of the OOB error for random forest uncertainty quantification (Zhang et al., 2019). Finally, one could consider different flexible models for the mean structure.

4.A Appendix

4.A.1 Spatial Simulations

The three informative and 17 spurious spatial simulation covariates were sampled from a Gaussian process with mean 0 and an independent covariance function. The spatial field, $Y(d)$, was sampled from a GP with a mean function $\mu_Y(d)$ depending on three covariates:

$$\mu_Y(d) = \eta * \{X_1(d) + 1[X_2(d) \geq 0] - 1[X_2(d) < 0] + 3[1 + \exp(-2X_3(d) + 3)]^{-1}\}, \tag{4.8}$$

where a is the covariate effect size multiplier that scales up or down the strength of the covariate effects. An isotropic exponential covariance function was employed, $\Sigma(d_1, d_2) = \exp(-\|d_1 - d_2\|/\nu)$, where ν is a parameter determines the strength of spatial dependence for points on the random field that are a distance of $\|d_1 - d_2\|$ apart.

The values of η and ν were varied to investigate model performance under a variety of circumstances. The values of the effect size multipler, η, were $\{0, 0.1, 0.2..., 2\}$, and ν varied from zero to 1.25, taking the following values, {0, 0.05, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25, 0.275, 0.3, 0.325, 0.35, 0.375, 0.4, 0.425, 0.45, 0.475, 0.5, 0.525, 0.55, 0.575, 0.6, 0.625, 0.65, 0.675, 0.7, 0.725, 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9, 0.925, 0.95, 0.975, 1, 1.025, 1.05, 1.075, 1.1, 1.125, 1.15, 1.175, 1.2, 1.225, 1.25}.

The unobserved grid of spatial locations at which $Y(d)$ was simulated was the cartesian product of 21 evenly spaced points begining at 0 and ending at 10. The observed data were simulated at 100 locations randomly selected with uniform probability $\{(s_1, s_2) : s_1, s_2 \in (0, 10)\}$.

4.A.2 Case Study Covariates

Covariate	Data Source
Monitor Latitude	U. S. Environmental Protection Agency
Monitor Longitude	U. S. Environmental Protection Agency
Date[1]	U. S. Environmental Protection Agency
Elevation (m)	National Digital Elevation Model
Boundary Layer Height (m)	Rapid Update Cycle
Surface Pressure (Pa)	Rapid Update Cycle
Relative Humidity (%)	Rapid Update Cycle
Temperature at 2 m (°K)	Rapid Update Cycle
U-Component of Wind Speed (m/s)	Rapid Update Cycle
V-Component of Wind Speed (m/s)	Rapid Update Cycle
Inverse Distance to Nearest Wildfire (m^{-1})[2]	Fire Inventory from NCAR v1.5
Annual Average Traffic within 1 km	Dynamap 2000, TeleAtlas
Agricultural Land Use within 1 km (%)	2006 National Land Cover Database
Urban Land Use within 1 km (%)	2006 National Land Cover Database
Vegetation Land Use within 1 km (%)	2006 National Land Cover Database
Normalized Difference Vegetation Index	Landsat Data
Nitrogen Dioxide (log molecules/cm^2)	Ozone Monitoring Instrument Satellite
WRF-Chem $PM_{2.5}$ (log kg/day))[3]	WRF-Chem
WRF-Chem Ozone (log 8 Hour Maximum)[4])	WRF-Chem

[1] Date was included for space-time simulations only.

[2] Proximity to wildfires was incorporated as inverse distance to the nearest wildfire. See Watson et al. (2019) for details.

[3] Only included as covariate for $PM_{2.5}$ case studies.

[4] Only included as covariate for ozone case studies.

4.A.3 Spherical Covariance Estimation

The default covariance estimation procedure for treeging fits a spherical variogram to the empirical variogram of the spatial or temporal residuals. For space-time treeging, we assume separability and separately fit spatial and temporal spherical covariances, multiplying them together to estimate the space-time covariance. The residuals are the difference between the fitted and observed training outcomes, i.e.,

$$\mathbf{e}(\mathbf{d}) = \mathbf{y}(\mathbf{d}) - \hat{\mathbf{y}}(\mathbf{d}), \tag{4.9}$$

where $\hat{\mathbf{y}}(\mathbf{d}) = T(\mathbf{X}(\mathbf{d}), \boldsymbol{\theta})$ for treeging and $\hat{\mathbf{y}}(\mathbf{d}) = \mathbf{X}(\mathbf{d})\boldsymbol{\beta}$ for kriging. The empirical variogram is

$$\hat{\gamma}(h \pm \delta) = \frac{1}{2|N(h \pm \delta)|} \sum_{|d_i - d_j| \leq \delta} |y(d_i) - y(d_j)|^2, \tag{4.10}$$

and the spherical variogram is

$$\gamma(h, r, s, a) = \begin{cases} 0 & h = 0 \\ a + (s - a)\left(\frac{3h}{2r} - \frac{h^3}{2r^3}\right) & 0 < h \leq r \\ s & h > r, \end{cases} \tag{4.11}$$

and spherical covariance function is

$$C(h, r, s, a) = \begin{cases} s & h = 0 \\ (s - a)\left(1 - \frac{3h}{2r} + \frac{h^3}{2r^3}\right) & 0 < h \leq r \\ 0 & h > r. \end{cases} \tag{4.12}$$

4.A.4 Space-Time Simulations

Covariates for the space-time simulations were generated from a Gaussian process (GP) with mean 0 and a separable covariance function $\Sigma_X(d) = \Sigma_X(s)\Sigma_X(t)$, where $d = (s, t)$. Exponential covariance functions with randomly selected range parameters were used for $\Sigma_X(s)$ and $\Sigma_X(t)$. The random field mean $\mu_Y[\mathbf{X}(d)]$ was a nonlinear function of the first 3 covariates,

$$\mu_Y[\mathbf{X}(d)] = \eta\{\tfrac{1}{2}X_1(d) + \tfrac{3}{4}[X_2(d) \leq 0.1] - \tfrac{3}{4}[X_2(d) > 0.1] + 2\{1 + \exp[-2X_3(d)]\}^{-1} + X_i(d)Xj(d)\}, \tag{4.13}$$

where η is the covariate effect size multiplier that can be scaled up or down to simulate stronger or weaker covariate effects, and i and j are 2 of the 3 informative covariates selected at random.

Values of the effect size multiplier η were $\{0, 0.2, 0.4, ..., 2\}$. The additional covariates were included as spurious covariates in scenarios B and D to assess the robustness of prediction models.

A separable covariance function was used for $Y(d)$ with independent exponential spatial and temporal covariance functions. The ranges were varied across the following ranges to investigate the effects of varying levels of spatial and temporal dependence. Values for the spatial covariance range were $2 - \sqrt{u}$ for $u \in \{0, 0.25, 0.5, ..., 4\}$. Values for the temporal covariance range were $\{0, 25., 5, 7.5, 10\}$.

The unobserved grid of spatial locations at which $Y(d)$ was simulated was the cartesian product of 11 evenly spaced points begining at 0 and ending at 10. The observed data were simulated at 40 spatial locations randomly selected with uniform probability $\{(s_1, s_2) : s_1, s_2 \in (0, 10)\}$. Both the observed and unobserved data were simulated over 30 consecutive timepoints, yielding a total of 1,200 observed data points and 3,630.

CHAPTER 5

Discussion

Model determination, prediction and statistical learning in the context of space-time data can be challenging. Predictive goals should be clearly articulated so that appropriate evaluation metrics may be selected for model training, tuning and evaluation. This dissertation presents a framework for model evaluation and prediction error estimation for the spatial interpolation of space-time data and proposes a new prediction algorithm.

Chapter two formalized the prediction error associated with spatial interpolation of space-time data and investigated cross-validation (CV) procedures to elucidate the estimands they target. Random forest, kriging and a simple linear model were used as diverse, representative models for illustrating evaluation metric performance. Consistent with recent best practices, the simulation study indicated that the naive application of model evaluation metrics developed for independent data is overly optimistic while location-based CV procedures are appropriate for estimating spatial interpolation error. The standard intuition regarding the bias-variance trade-off of CV fold size does not extend trivially to space-time data, making leave-one-location-out (LOLO) CV the recommended metric, although 10-fold location-based CV is an acceptable substitute when LOLO CV is too computationally burdensome. This determination is made empirically, as theoretical guarantees for CV procedures are difficult to establish even in the case of independent data, which admits to replication. There may be room for future work in this direction, although it is not immediately obvious how one might proceed.

Using this model evaluation framework, chapter three assessed the predictive performance of ten machine learning algorithms for spatially interpolating ozone data during a major wildfire event. The earlier version of the work in this chapter published in *Environmental Pollution* was the first such analysis of ozone concentrations during a wildfire event using machine learning algorithms (Watson et al., 2019). In this comparison, models with flexible mean structures tended

to perform better than those with rigid means. Gradient boosting and random forest had the lowest LOLO-CV estimates of prediction error, suggesting that ensembled tree-based models work particularly well. This is not altogether surprising, as these models are known to perform well in a variety of settings. Their ability to automatically accommodate interactions, nonlinear effects and spurious covariates seems relevant for predicting atmospheric phenomena like air pollution.

Consistent with the findings of chapter two, naive 10-fold CV provided unreliable estimates of prediction error and a different ranking of model performance. Its over optimism was more pronounced for more flexible models apparently on account of the ability of these models to exploit the additional information on the prediction case present in the naive CV folds. The difference in model ranking indicates that naive CV is also unreliable for selecting, tuning or averaging models.

The spatial interpolation provided by the gradient boosting model tested on this data was used in two subsequent epidemiological analyses of the health consequences of air pollution exposure during the wildfire event. The first of these found that ozone concentration was associated with increased asthma emergency department visits (Reid et al., 2019). The second of these studies is forthcoming and has identified differential impacts of ozone and $PM_{2.5}$ by race/ethnicity and socioeconomic status. These findings are of increasing significance as wildfire severity and frequency increase.

Chapter four introduced treeging, a new machine learning algorithm for space-time prediction that enriches regression trees with a kriging covariance structure to form the base learner of an ensemble tool. This approach was motivated by the differing strengths of traditional models, which focus on space-time dependence, and machine learning algorithms, which offer flexible mean structures. The choice of a tree-based ensemble for the mean function was suggested by the excellent performance of random forest and boosting in chapter two and other analyses. In extensive spatial and space-time simulation studies that included four different covariate scenarios, treeging performed well. While kriging tends to perform poorly in the presence of spurious covariates, non-linear effects or non-additive mean structures, and machine learning tools like random forest suffering when the covariates are not informative, treeging performs well across the board and in many case was the best (lowest MSE) model. It also performed well in spatial and space-time case studies using the ozone and $PM_{2.5}$ wildfire data.

A number of possible extensions and promising avenues for future work suggest themselves.

Formalizing prediction error estimands for other types of space-time prediction including forecasting and spatial extrapolation and investigating the appropriateness of potential estimators will likely prove useful as it has for spatial interpolation.

Incorporating the location-based resampling strategies explored here into model tuning and averaging could provide improved predictive performance. For example, procedures like random forest that use bagging to construct dissimilar base learners for the purposes of reducing the variance of an ensembled prediction could use location in selecting bootstrap samples. Similarly, location-based CV estimators of prediction error will provide more reliable weights for combining prediction models than naive estimators.

The use of spatially interpolated predictions in two-step epidemiological analyses of the health effects of air pollution suggests there may be utility in a combined, one-step procedure. Predicting air pollution as a first step and then using it as a covariate in a subsequent regression model ignores the uncertainty in the prediction model. Combining the procedure into a single analysis may allow for better propagation of uncertainty, but also increases the complexity of the problem and precludes multiple secondary analyses from a single set of interpolating predictions. An alternative could be to incorporate prediction uncertainty into the secondary regression model using techniques for addressing measurement error.

There are a number of potential directions for future work involving treeging. Assessing whether variable importance metrics and uncertainty quantification approaches developed for random forest translate to treeging may be useful. More sophisticated approaches to model averaging such as developing a formal optimization strategy or introducing localized sampling as an alternative to naive subsampling. Establishing the precise connection with treed Gaussian processes would be interesting. In addition, a number of computational improvements may be possible to improve its computational efficiency and thereby increase the size of the data sets on which it may be used.

References

Allaire, J. and Chollet, F. (2015). *keras: R Interface to 'Keras'*. R package version 2.2.0.9001.

Atkinson, R. W., Fuller, G. W., Anderson, H. R., Harrison, R. M., and Armstrong, B. (2010). Urban ambient particle metrics and health: a time-series analysis. *Epidemiology*, 21(4):501–511.

Azevedo, J. M., Gonçalves, F. L., and de Fátima Andrade, M. (2011). Long-range ozone transport and its impact on respiratory and cardiovascular health in the north of portugal. *Int. J. Biometeorol.*, 55(2):187–202.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Barr, R., Hoffman, E., Madrigano, J., Aaron, C., Sampson, P., Sheppard, L., Vedal, S., Kaufman, J., and Wang, M. (2016). Long-term exposure to ozone and accelerated progression of percent emphysema and decline in lung function: The mesa air and lung studies. *Medicine*, 1(2):3.

Beckerman, B. S., Jerrett, M., Serre, M., Martin, R. V., Lee, S.-J., Van Donkelaar, A., Ross, Z., Su, J., and Burnett, R. T. (2013). A hybrid approach to estimating national scale spatiotemporal variability of $pm_{2.5}$ in the contiguous united states. *Environ. Sci. Technol.*, 47(13):7,233–7,241.

Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2004). Ozone and short-term mortality in 95 us urban communities, 1987-2000. *JAMA*, 292(19):2,372–2,378.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. Agr. Bio. Environ. Stat.*, 15(2):176–197.

Berrocal, V. J., Guan, Y., Muyskens, A., Wang, H., Reich, B. J., Mulholland, J. A., and Chang, H. H. (2020). A comparison of statistical and machine learning methods for creating national daily maps of ambient pm2. 5 concentration. *Atmospheric Environment*, 222:117130.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Stat*, pages 1,196–1,217.

Boisvert, J., Manchuk, J., and Deutsch, C. (2009). Kriging in the presence of locally varying anisotropy using non-euclidean distances. *Mathematical Geosciences*, 41(5):585–601.

Bougeault, P. and Lacarrere, P. (1989). Parameterization of orography-induced turbulence in a mesobeta–scale model. *Monthly Weather Review*, 117(8):1872–1890.

Brauer, M., Amann, M., Burnett, R. T., Cohen, A., Dentener, F., Ezzati, M., Henderson, S. B., Krzyzanowski, M., Martin, R. V., Van Dingenen, R., et al. (2012). Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. *Environ. Sci. Technol.*, 46(2):652–660.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Van Reeuwijk, H., Smallbone, K., and Van Der Veen, A. (1997). Mapping urban air pollution using gis: a regression-based approach. *Int. J. Geog. Inf. Sci.*, 11(7):699–718.

Brokamp, C., Jandarov, R., Hossain, M., and Ryan, P. (2018). Predicting daily urban fine particulate matter concentrations using a random forest model. *Environmental science & technology*, 52(7):4173–4179.

Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., Luepker, R., Mittleman, M., Samet, J., Smith, S. C., et al. (2004). Air pollution and cardiovascular disease: a statement for healthcare professionals from the expert panel on population and prevention science of the american heart association. *Circulation*, 109(21):2655–2671.

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., et al. (2010). Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the american heart association. *Circulation*, 121(21):2,331–2,378.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geog. Anal.*, 28(4):281–298.

Cadelis, G., Tourres, R., and Molinie, J. (2014). Short-term effects of the particulate pollutants contained in saharan dust on the visits of children to the emergency department due to asthmatic conditions in guadeloupe (french archipelago of the caribbean). *PloS One*, 9(3):e91136.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM.

Chuang, K.-J., Yan, Y.-H., Chiu, S.-Y., and Cheng, T.-J. (2011). Long-term air pollution exposure and risk factors for cardiovascular diseases among the elderly in taiwan. *Occup. Environ. Med.*, 68(1):64–68.

Confalonieri, U., Menne, B., Akhtar, R., Ebi, K. L., Hauengue, M., Kovats, R. S., Revich, B., and Woodward, A. (2007). Human health (2007). In Parry, M., Canziani, O., and Palutikof, J., editors, *Climate change 2007: impacts, adaptation and vulnerability*, volume 4. Cambridge University Press Cambridge, Cambridge.

Cressie, N. (2015a). *Statistics for spatial data*. John Wiley & Sons.

Cressie, N. (2015b). *Statistics for spatial data*. John Wiley & Sons.

Crouse, D. L., Peters, P. A., Hystad, P., Brook, J. R., van Donkelaar, A., Martin, R. V., Villeneuve, P. J., Jerrett, M., Goldberg, M. S., Pope III, C. A., Brauer, M., Brook, R. D., Robichaud, A., Menard, R., and Burnett, R. T. (2015). Ambient $pm_{2.5}$, o_3, and no_2 exposures and associations with mortality over 16 years of follow-up in the canadian census health and environment cohort (canchec). *Environ. Health Perspect.*, 123(11):1180.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *J. Am. Stat. Assoc.*, 111(514):800–812.

De Kok, T. M., Driece, H. A., Hogervorst, J. G., and Briedé, J. J. (2006). Toxicological assessment of ambient and traffic-related particulate matter: a review of recent studies. *Mutat. Res. Rev. Mutat. Res.*, 613(2):103–122.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., and Schwartz, J. (2016). Assessing $pm_{2.5}$ exposures with high spatiotemporal resolution across the continental united states. *Environ. Sci. Technol.*, 50(9):4,712–4,721.

Di, Q., Rowland, S., Koutrakis, P., and Schwartz, J. (2017). A hybrid model for spatially and temporally resolved ozone exposures in the continental united states. *J. Air Waste Manage. Assoc.*, 67(1):39–52.

Diggle, P. J., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *J. Am. Med. Assoc.*, 295(10):1,127–1,134.

Dongarra, J. (2002). Basic linear algebra subprograms technical forum standard. *International Journal of High Performance Applications and Supercomputing*, 16(2):119–141.

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40(8):1–18.

Efron, B. (1979). Bootstrap methods: another look at the jackknife annals of statistics 7: 1–26. *View Article PubMed/NCBI Google Scholar*.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–642.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Stat. Assoc.*, 92(438):548–560.

Ek, M., Mitchell, K., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. (2003). Implementation of noah land surface model advances in the national centers for environmental prediction operational mesoscale eta model. *Journal of Geophysical Research: Atmospheres*, 108(D22).

Esworthy, R. (2013). Air quality: Epa's 2013 changes to the particulate matter (pm) standard. *Congressional Res. Serv*, pages 7–5700.

Ferreira, A. J. and Figueiredo, M. A. (2012). Boosting algorithms: A review of methods, theory, and applications. In *Ensemble machine learning*, pages 35–85. Springer.

Fowler, D., Amann, M., Anderson, F., Ashmore, M., Cox, P., Depledge, M., Derwent, D., Grennfelt, P., Hewitt, N., Hov, O., and Jenkin, M. (2008). Ground-level ozone in the 21st century: future trends, impacts and policy implications. *R. Soc. Sci. Policy Rep.*, 15(08).

Fry, J., Xian, G. Z., Jin, S., Dewitz, J., Homer, C. G., Yang, L., Barnes, C. A., Herold, N. D., and Wickham, J. D. (2011). Completion of the 2006 national land cover database for the conterminous united states. *Photogramm. Eng. Remote Sens.*, 77(9):858–864.

Gao, Y., Liu, L., Zhang, C., Wang, X., and Ma, H. (2020). Si-agan: Spatial interpolation with attentional generative adversarial networks for environment monitoring. In *ECAI 2020*, pages 1786–1793. IOS Press.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *J. Am. Stat. Assoc.*, 98(462):387–396.

Gelfand, A. E., Sahu, S. K., and Holland, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics*, 23(7):565–578.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298):369–375.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed gaussian process models with an application to computer modeling. *J. Am. Stat. Assoc.*, 103(483):1,119–1,130.

Grell, G. A. and Dévényi, D. (2002). A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophysical Research Letters*, 29(14):38–1.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J., and Coull, B. A. (2008). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10(2):258–274.

Guaita, R., Pichiule, M., Maté, T., Linares, C., and Díaz, J. (2011). Short-term impact of particulate matter ($pm_{2.5}$) on respiratory mortality in madrid. *Int. J. Environ. Health Res.*, 21(4):260–274.

Guarnieri, M. and Balmes, J. R. (2014). Outdoor air pollution and asthma. *The Lancet*, 383(9928):1581–1592.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P., and Geron, C. (2006). Estimates of global terrestrial isoprene emissions using megan (model of emissions of gases and aerosols from nature). *Atmospheric Chemistry and Physics*, 6(11):3181–3210.

Guo, Y., Tang, Q., Gong, D.-Y., and Zhang, Z. (2017). Estimating ground-level pm2. 5 concentrations in beijing using a satellite-based geographically and temporally weighted regression model. *Remote Sensing of Environment*, 198:140–149.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.*, 42(33):7561–7578.

Hou, Q., An, X., Wang, Y., Tao, Y., and Sun, Z. (2012). An assessment of china's pm_{10}-related health economic losses in 2009. *Sci. Total Environ.*, 435:61–65.

Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., and Liu, Y. (2017). Estimating $pm_{2.5}$ concentrations in the conterminous united states using the random forest approach. *Environ. Sci. Technol.*, 51(12):6,936–6,944.

Hutchinson, M. and Gessler, P. (1994). Splines—more than just a smooth interpolator. *Geoderma*, 62(1-3):45–67.

Hwang, B.-F., Chen, Y.-H., Lin, Y.-T., Wu, X.-T., and Lee, Y. L. (2015). Relationship between exposure to fine particulates and ozone and reduced lung function in children. *Environ. Res.*, 137:382–390.

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., and Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Calculations with the aer radiative transfer models. *Journal of Geophysical Research: Atmospheres*, 113(D13).

Jaffe, D. A. and Wigder, N. L. (2012). Ozone production from wildfires: A critical review. *Atmos. Environ.*, 51:1–10.

Janjić, Z. (1996). The surface layer in the ncep eta model, paper presented at 11th conference on numerical weather prediction. *Am. Meteorol. Soc., Norfolk, Va.*

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., and Giovis, C. (2005). A review and evaluation of intraurban air pollution exposure models. *J. Exposure Sci. Environ. Epidemiol.*, 15(2):185.

Jerrett, M., Burnett, R. T., Pope III, C. A., Ito, K., Thurston, G., Krewski, D., Shi, Y., Calle, E., and Thun, M. (2009). Long-term ozone exposure and mortality. *N. Engl. J. Med.*, 360(11):1,085–1,095.

Just, A. C., Wright, R. O., Schwartz, J., Coull, B. A., Baccarelli, A. A., Tellez-Rojo, M. M., Moody, E., Wang, Y., Lyapustin, A., and Kloog, I. (2015). Using high-resolution satellite aerosol optical depth to estimate daily pm2. 5 geographical distribution in mexico city. *Environmental science & technology*, 49(14):8576–8584.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214.

Keller, J. P., Chang, H. H., Strickland, M. J., and Szpiro, A. A. (2017). Measurement error correction for predicted spatiotemporal air pollution exposures. *Epidemiology*, 28(3):338.

Kesic, M. J., Meyer, M., Bauer, R., and Jaspers, I. (2012). Exposure to ozone modulates human airway protease/antiprotease balance contributing to increased influenza a infection. *PloS One*, 7(4):e35108.

Khaniabadi, Y. O., Hopke, P. K., Goudarzi, G., Daryanoosh, S. M., Jourvand, M., and Basiri, H. (2017). Cardiopulmonary mortality and copd attributed to ambient ozone. *Environ. Res.*, 152:336–341.

Kim, J.-B., Kim, C., Choi, E., Park, S., Park, H., Pak, H.-N., Lee, M.-H., Shin, D. C., Hwang, K.-C., and Joung, B. (2012). Particulate air pollution induces arrhythmia via oxidative stress and calcium calmodulin kinase ii activation. *Toxicol. Appl. Pharmacol.*, 259(1):66–73.

Kim, K.-H., Kabir, E., and Kabir, S. (2015). A review on the human health impact of airborne particulate matter. *Environ. Int.*, 74:136–143.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145. Montreal, Canada.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *J. South Afr. Inst. Min. Metallc.*, 52(6):119–139.

Kuhn, M., Wing, C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt, T. (2018). *caret: Classification and Regression Training*. R package version 6.0-80.

Laden, F., Neas, L. M., Dockery, D. W., and Schwartz, J. (2000). Association of fine particulate matter from different sources with daily mortality in six us cities. *Environ. Health Perspect.*, 108(10):941.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., and Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, 23(7):811–820.

Lee, J. Y., Lee, S.-B., and Bae, G.-N. (2014). A review of the association between air pollutant exposure and allergic diseases in children. *Atmos. Pollut. Res.*, 5(4):616–629.

Lee, M., Kloog, I., Chudnovsky, A., Lyapustin, A., Wang, Y., Melly, S., Coull, B., Koutrakis, P., and Schwartz, J. (2016). Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the southeastern us 2003–2011. *J. Exposure Sci. Environ. Epidemiol.*, 26(4):377.

Levelt, P. F., van den Oord, G. H., Dobber, M. R., Malkki, A., Visser, H., de Vries, J., Stammes, P., Lundell, J. O., and Saari, H. (2006). The ozone monitoring instrument. *IEEE Trans. Geosci. Remote Sens.*, 44(5):1093–1101.

Liu, Y., Paciorek, C. J., and Koutrakis, P. (2009). Estimating regional spatial and temporal variability of $pm_{2.5}$ concentrations using satellite data, meteorology, and land use information. *Environ. Health Perspect.*, 117(6):886.

Lloyd, C. D. (2010). *Local models for spatial analysis*. CRC press.

Lu, G. Y. and Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Comput. Geosci.*, 34(9):1,044–1,055.

Lu, W.-Z. and Wang, W.-J. (2005). Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. *Chemosphere*, 59(5):693–701.

Manepalli, A., Singh, A., Mudigonda, M., White, B., and Albert, A. (2020). Generalization properties of machine learning based weather model downscaling. In Rush, A., editor, *Proceedings of the International Conference on Learning Representations*, pages 1–8.

Marcon, A., Pesce, G., Girardi, P., Marchetti, P., Blengio, G., de Zolt Sappadina, S., Falcone, S., Frapporti, G., Predicatori, F., and de Marco, R. (2014). Association between pm_{10} concentrations and school absences in proximity of a cement plant in northern italy. *Int. J. Hyg. Environ. Health*, 217(2-3):386–391.

Matheron, G. (1963). Principles of geostatistics. *Econ. Geol.*, 58(8):1246–1266.

McCormack, M. C., Breysse, P. N., Matsui, E. C., Hansel, N. N., Peng, R. D., Curtin-Brosnan, J., D'Ann, L. W., Wills-Karp, M., and Diette, G. B. (2011). Indoor particulate matter increases asthma morbidity in children with non-atopic and atopic asthma. *Ann. Allergy Asthma Immunol.*, 106(4):308–315.

Menzel, D. B. (1984). Ozone: an overview of its toxicity in man and animals. *J. Toxicol. Environ. Health*.

Misra, C., Geller, M. D., Shah, P., Sioutas, C., and Solomon, P. A. (2001). Development and evaluation of a continuous coarse (pm_{10}–pm_{25}) particle monitor. *J. Air Waste Manage. Assoc.*, 51(9):1309–1317.

Møller, J. and Waagepetersen, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684.

National Research Council (2008). *Estimating mortality risk reduction and economic benefits from controlling ozone air pollution*. National Academies Press.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. (2000). Spatial tessellations: Concepts and applications of voronoi diagrams, john wiley & sons. *Chichester, UK*.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of non-stationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17(5):483–506.

Paciorek, C. J., Yanosky, J. D., Puett, R. C., Laden, F., and Suh, H. H. (2009). Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann. Appl. Stat.*, pages 370–397.

Pearson, J. F., Bachireddy, C., Shyamprasad, S., Goldfine, A. B., and Brownstein, J. S. (2010). Association between fine particulate matter and diabetes prevalence in the us. *Diabetes Care*, 33(10):2196–2201.

Pfister, G., Avise, J., Wiedinmyer, C., Edwards, D., Emmons, L., Diskin, G., Podolske, J., and Wisthaler, A. (2011a). Co source contribution analysis for california during arctas-carb. *Atmos. Chem. Phys.*, 11(15):7,515–7,532.

Pfister, G., Parrish, D., Worden, H., Emmons, L., Edwards, D., Wiedinmyer, C., Diskin, G., Huey, G., Oltmans, S., Thouret, V., and Weinheimer, A. (2011b). Characterizing summertime chemical boundary conditions for airmasses entering the us west coast. *Atmos. Chem. Phys.*, 11(4):1769–1790.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., and Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019.

Powers, J. G., Klemp, J. B., Skamarock, W. C., Davis, C. A., Dudhia, J., Gill, D. O., Coen, J. L., Gochis, D. J., Ahmadov, R., Peckham, S. E., and Grell, G. A. (2017). The weather research and forecasting model: Overview, system efforts, and future directions. *Bull. Am. Meteorol. Soc.*, 98(8):1717–1737.

Ransom, M. and Pope, I. (2013). Air pollution and school absenteeism: Results from a natural experiment. *Manuscript, Department of Economics, Brigham Young University*.

Reid, C. E., Considine, E. M., Watson, G. L., Telesca, D., Pfister, G. G., and Jerrett, M. (2019). Associations between respiratory health and ozone and fine particulate matter during a wildfire event. *Environment international*, 129:291–298.

Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., Raffuse, S. M., and Balmes, J. R. (2015). Spatiotemporal prediction of fine particulate matter during the 2008 northern california wildfires using machine learning. *Environ. Sci. Technol.*, 49(6):3887–3896.

Robert, C. (2014). *Machine learning, a probabilistic perspective*. Taylor & Francis.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Ryan, P. H. and LeMasters, G. K. (2007). A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation toxicology*, 19(sup1):127–133.

Saha, A., Basu, S., and Datta, A. (2021). Random forests for spatially dependent data. *Journal of the American Statistical Association*, (just-accepted):1–46.

Samoli, E., Peng, R., Ramsay, T., Pipikou, M., Touloumi, G., Dominici, F., Burnett, R., Cohen, A., Krewski, D., Samet, J., et al. (2008). Acute effects of ambient particulate matter on mortality in europe and north america: results from the aphena study. *Environ. Health Perspect.*, 116(11):1,480.

Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Silva, R. A., West, J. J., Lamarque, J.-F., Shindell, D. T., Collins, W. J., Faluvegi, G., Folberth, G. A., Horowitz, L. W., Nagashima, T., Naik, V., and Rumbold, S. T. (2017). Future global mortality from changes in air pollution attributable to climate change. *Nat. Clim. Change*, 7(9):647.

Silverman, R. A. and Ito, K. (2010). Age-related association of fine particles and ozone with severe acute asthma in new york city. *J. Allergy Clin. Immun.*, 125(2):367–373.

Singh, H., Cai, C., Kaduwela, A., Weinheimer, A., and Wisthaler, A. (2012). Interactions of fire emissions and urban pollution over california: Ozone formation and air quality simulations. *Atmospheric environment*, 56:45–51.

Smith, K. R., Jerrett, M., Anderson, H. R., Burnett, R. T., Stone, V., Derwent, R., Atkinson, R. W., Cohen, A., Shonkoff, S. B., Krewski, D., and Pope, C. A. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: health implications of short-lived greenhouse pollutants. *The Lancet*, 374(9707):2091–2103.

Sousa, S., Alvim-Ferraz, M., and Martins, F. (2013). Health effects of ozone focusing on childhood asthma: what is now known–a review from an epidemiological point of view. *Chemosphere*, 90(7):2051–2058.

Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association*, 100(469):310–321.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stocker, T., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. (2013). Fifth assessment report of the intergovernmental panel on climate change. *The Phys. Sci. Basis*.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B Methodol.*, pages 111–147.

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

Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D., and Kaufman, J. D. (2010). Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*, 21(6):606–631.

Tham, R., Erbas, B., Akram, M., Dennekamp, M., and Abramson, M. J. (2009). The impact of smoke on respiratory hospital outcomes during the 2002–2003 bushfire season, victoria, australia. *Respirology*, 14(1):69–75.

Turner, M. C., Jerrett, M., Pope III, C. A., Krewski, D., Gapstur, S. M., Diver, W. R., Beckerman, B. S., Marshall, J. D., Su, J., Crouse, D. L., and Burnett, R. T. (2016). Long-term ozone exposure and mortality in a large prospective study. *Am. J. Respir. Crit. Care Med.*, 193(10):1,134–1,142.

U.S. Department of Commerce National Oceanic and Atmospheric Administration (2018). Rapid update cycle.

Val Martin, M., Honrath, R., Owen, R. C., Pfister, G., Fialho, P., and Barata, F. (2006). Significant enhancements of nitrogen oxides, black carbon, and ozone in the north atlantic lower free troposphere resulting from north american boreal wildfires. *J. Geophys. Res. D: Atmos.*, 111(D23).

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Van Donkelaar, A., Martin, R. V., Spurr, R. J., and Burnett, R. T. (2015). High-resolution satellite-derived pm2.5 from optimal estimation and geographically weighted regression over north america. *Environmental Science & Technology*, 49(17):10482–10491.

Watson, C. D., Wang, C., Lynar, T., and Weldemariam, K. (2020a). Investigating two super-resolution methods for downscaling precipitation: Esrgan and car. *arXiv preprint arXiv:2012.01233*.

Watson, G. L., Reid, C. E., Jerrett, M., and Telesca, D. (2020b). Prediction & model evaluation for space-time data. *arXiv preprint arXiv:2012.13867*.

Watson, G. L., Telesca, D., Reid, C. E., Pfister, G. G., and Jerrett, M. (2019). Machine learning models accurately predict ozone exposure during wildfire events. *Environmental Pollution*, 254:112792.

Weizhen, H., Zhengqiang, L., Yuhuan, Z., Hua, X., Ying, Z., Kaitao, L., Donghui, L., Peng, W., and Yan, M. (2014). Using support vector regression to predict pm_{10} and $pm_{2.5}$. In *IOP Conference Series: Earth and Environmental Science*, volume 17, page 012268. IOP Publishing.

Wiedinmyer, C., Akagi, S., Yokelson, R. J., Emmons, L., Al-Saadi, J., Orlando, J., and Soja, A. (2011). The fire inventory from ncar (finn): a high resolution global model to estimate the emissions from open burning. *Geosci. Model Dev.*, 4(3):625.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

World Health Organization (2013). *Health effects of particulate matter. Policy implications for countries in eastern Europe, Caucasus and central Asia. Copenhagen: WHO Regional Office for Europe; 2013*.

Wu, J., Winer, A. M., and Delfino, R. J. (2006). Exposure assessment of particulate matter air pollution before, during, and after the 2003 southern california wildfires. *Atmos. Environ.*, 40(18):3333–3348.

Xu, Y., Ho, H. C., Wong, M. S., Deng, C., Shi, Y., Chan, T.-C., and Knudby, A. (2018). Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level pm2. 5. *Environmental Pollution*, 242:1417–1426.

Zanobetti, A., Coull, B. A., Gryparis, A., Kloog, I., Sparrow, D., Vokonas, P. S., Wright, R. O., Gold, D. R., and Schwartz, J. (2013). Associations between arrhythmia episodes and temporally and spatially resolved black carbon and particulate matter in elderly patients. *Occup. Environ. Med.*, pages oemed–2013.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Zhu, L., and Zhang, M. (2017). Spatiotemporal prediction of continuous daily $pm_{2.5}$ concentrations across china using a spatially explicit machine learning algorithm. *Atmos. Environ.*, 155:129–139.

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., and Di, B. (2018a). Spatiotemporal prediction of daily ambient ozone levels across china using random forest for human exposure assessment. *Environmental Pollution*, 233:464–473.

Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M. L., and Di, B. (2018b). Satellite-based estimates of daily no2 exposure in china using hybrid random forest and spatiotemporal kriging model. *Environmental science & technology*, 52(7):4180–4189.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2019). Random forest prediction intervals. *The American Statistician*.

Zhang, P. (1995). Assessing prediction error in non-parametric regression. *Scandinavian Journal of Statistics*, pages 83–94.

Zhu, D., Cheng, X., Zhang, F., Yao, X., Gao, Y., and Liu, Y. (2020). Spatial interpolation using conditional generative adversarial neural networks. *International Journal of Geographical Information Science*, 34(4):735–758.