Random Spatial Forests

Travis Hee Wai
Department of Biostatistics, University of Washington
and
Michael T. Young
Department of Environmental & Occupational Health Sciences, University of Washington
and
Adam A. Szpiro
Department of Biostatistics, University of Washington

June 2, 2020

Abstract

A fundamental problem in environmental epidemiology studies on the association of air pollution exposure with health outcomes is identifying exposure levels for individuals in a cohort study. Measurements are not made at each study participants place of residence, thus individual specific exposure levels are estimated using observations from regulatory monitors. In recent years, it has become desirable to use random forests and other statistical learning techniques to model air pollution exposure at unobserved sites. However, these methods do not exploit the spatial structure of the data. We propose a computationally efficient algorithm to build regression trees allowing for spatial correlation and use these trees to construct random spatial forests. Simulations show that our method outperforms existing approaches on spatially indexed data, and we demonstrate its improved accuracy on elemental carbon, organic carbon, silicon, and sulfur measurements across the continental United States.

Keywords: Random Forests, Regression Trees, Geostatistics, Exposure Prediction, Universal Kriging
1 Introduction

Recent environmental epidemiology studies have found air pollution exposure to be associated with a variety of health effects, such as heart disease, (Kaufman et al. 2016), cognitive impairment (Power et al. 2016), and lung disease (Sack et al. 2017). In these types of studies, exposure levels for individuals are not observed directly. Instead, individual exposure levels are estimated using statistical models constructed off of measurements at regulatory monitoring sites. Here, we develop air pollution exposure models which accurately reconstruct annual averages of four PM$_{2.5}$ (particulate matter less than 2.5 $\mu$m in aerodynamic diameter) sub-species: elemental carbon (EC), organic carbon (OC), silicon (Si), and sulfur (S) using Environmental Protection Agency (EPA) Interagency Monitoring for Protected Visual Environments (IMPROVE) and Chemical Speciation Network (CSN) monitoring data from 2009-2010. These two networks are fairly evenly dispersed throughout the continental United States where IMPROVE monitors are located mostly in remote areas, while CSN monitors are in more urban areas.

It has been demonstrated that including geographic covariates in land-use regression models to predict air pollutant concentrations works well in practice (Hoek et al. 2008). For our application, we obtained geographic information systems (GIS) covariates representing land-use characteristics. These covariates include proximity variables, measures of proximity to the nearest geographic feature, and buffer variables, measuring geographic features within a given radius. Proximity variables include measures of distance to major roadways, commercial zones, airports, ports, railroads, and railyards. Buffer variables measure the sum or average of various features within a given radius. For every feature, we estimated several buffer variables of different radii, ranging from 50m to 15km. Specific buffer variables included measurements of the area of each USGS land-use classification (both 1990
and 2000) and the average Normalized Difference Vegetation Index (NDVI) value, sum of roadways lengths, to name a few.

These sets of covariates are often quite large, in our case we obtained 473 GIS covariates but only have data from 323 monitoring locations. In addition, technological advancements have increased the ease and scope of data collection for geographic land-use covariates such as satellite data (Xu et al. 2018), traffic data (Saucy et al. 2018), and meteorology data (Arain et al. 2007). In order to leverage all of these different information sources, researchers often use high dimensional methods from the statistical learning literature. In particular, random forests (Breiman 2001) has been shown to be effective in prediction problems and recently many have examined using it to estimate pollutant levels (Hu et al. 2017, Stafoggia et al. 2019, Chen et al. 2018, Li, Cui, Meng, Zhao & Fu 2019). However, some studies suggest that applying random forests and other machine learning methods, which do not utilize spatial information, to spatial data do not yield any noticeable advantages over traditional geostatistical approaches such as kriging (Berrocal et al. 2019, Fox et al. 2018). In order to correct for this deficiency, some have proposed a two-step approach applying kriging or smoothing of the residuals from the statistical learning technique in attempt to add spatial structure into predictions (Rolf et al. 2020). In many cases, this two-step approach has been shown to perform better than either using either method alone (Liu et al. 2018).

Applying random forests without allowing for the spatial process is an inefficient optimization strategy for the two-step approach. Estimates combining random forests and kriging can be viewed as an additive model. Here, the spatial process is a statistical characterization of spatial variation not explained by covariates, for example topography, climatological, and meteorological patterns, that are difficult to model explicitly. In order
to maximize variability explained through the additive model, we would like to use the
random forests part to model systematic variation which cannot be modeled in the spatial
process. By ignoring spatial correlation, random forests may model spatial structure that
could have otherwise been modeled in the spatial process.

Little has been done to explore the degree to which the predictive power of these models
can be improved by incorporating spatial information into random forests itself. Hengl et al.
(2018) proposed random forests for spatial data, where they explored adding geographic
proximity as a covariate before applying the random forests algorithm but only found
similar prediction accuracy to the two-step random forests kriging approach. We emphasize
that our goal is to use random forests to utilize geographic covariates which model variation
which could not be modeled by the spatial process, and including geographic proximity as
a covariate does not help in that regard.

Our main contribution is a novel algorithm to construct spatially adjusted trees which
allow for spatial correlation, and evaluate two different procedures for constructing random
forests estimates from spatially adjusted trees. In section 2, we describe a summary of ran-
dom forests and universal kriging, describe our modified tree building algorithm allowing
for spatial correlation, and examine different approaches to constructing random forests
estimates from spatially adjusted trees. In section 3, we provide simulation results demon-
strating the advantage of our approach over two-step estimation strategies. In section 4, we
apply our method to annual average elemental carbon (EC), organic carbon (OC), Silicon
(Si), and Sulfur (S) across the continental United States for 2009-2010. We end the paper
with a discussion of the advantages of our method and aspects for future work.
2 Methods

2.1 Tree-Based Methods

2.1.1 Regression Trees

Regression trees have gained popularity for their ability to approximate a wide variety of non-linear functions. Trees are built through an iterative process called recursive binary splitting, which aim to minimize total tree impurity, traditionally mean-squared error, through a greedy optimization approach. At each iteration, a new terminal node of the tree is created by an exhaustive search selecting the branch which minimizes tree impurity for the current iteration. Although trees are often described as segmenting the data into terminal nodes by following decision rules in an attempt to sort observations with similar covariates together, a regression tree can also be formulated as a linear model.

Consider \( n \) observations at locations \( s_1, \ldots, s_n \in A \subset \mathbb{R}^d \) from the process

\[
Z(s) = Y(s) + \epsilon, \quad \epsilon \stackrel{i.i.d}{\sim} (0, \tau^2 I_n)
\]

A tree, \( t(X) \) with \( k \) branches and \( k+1 \) terminal nodes can be represented as \( t(X) = C^k \pi^k \), with

\[
C^k = \begin{bmatrix} c_{ij} \end{bmatrix} \in \{0, 1\}^{n \times (k+1)}.
\]

In each iteration \( t \) of the algorithm, a column vector \( c_t \) is augmented to \( C^{t-1} \) marking which observations are in the new terminal node created during that step

\[
c_{it} = \begin{cases} 
1, & \text{observation } i \text{ in terminal created by step } t \\
0, & \text{else}
\end{cases}
\]
Similarly to a binary tree, each of the \( k + 1 \) terminal nodes of the tree is then encoded by a unique combination of the \( k + 1 \) columns of \( C^k \) and the tree estimate for that terminal node is a unique linear combination of the corresponding entries of the \( k + 1 \) vector \( \pi^k \). One particular advantage of treating a regression tree as a linear model is that by profiling out \( \hat{\pi}^k \), the total tree impurity only depends on the structure of the tree design matrix

\[
\|Z(s) - t(X)\|_2^2 = Z(s)^T \left( I_n - C^k \left( (C^k)^T C^k \right)^{-1} (C^k)^T \right) Z(s),
\]

leading to efficient computational methods since \( \hat{\pi}^k \) does not need to be optimize for every possible new branch.

### 2.1.2 Random Forests

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

\[
\hat{f}(X) = \frac{1}{B} \sum_{i=1}^{B} t^i(X^i)
\]

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

2.2.1 Universal Kriging

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

\[ Y(s) = X\beta + \nu(s) \]

which contains a linear mean structure on the covariates and observations are subject to variation from the linear model by a spatially correlated zero mean stochastic term \( \nu(s) \sim (0, \Sigma(\theta)) \) where the spatial covariance is known up to parameters \( \theta \). The kriging approach models the error term \( \nu(s) \) as a realization of a Gaussian process and estimate \( \theta, \beta \) by maximization of the log-likelihood.

\[
\arg\max_{\theta, \beta} -\frac{1}{2} \log |\Sigma(\theta)| - \frac{1}{2} (Y(s) - X\beta)^T \Sigma^{-1}(\theta)(Y(s) - X\beta)
\]

For any fixed \( \theta_0 \), \( \beta \) which maximizes \( \ell(\beta, \theta_0|Y(s)) \) is easily shown to be the generalized least squares estimator

\[
\hat{\beta} = (X^T \Sigma^{-1}(\theta_0) X)^{-1} X^T \Sigma^{-1}(\theta) Y(s).
\]

The method of eliminating \( \beta \) from the log likelihood by profiling is commonly employed, and universal kriging models are estimated by optimizing (1).

\[
\arg\max_{\theta} -\frac{1}{2} \log |\Sigma(\theta)| - \frac{1}{2} (Y(s) - X\hat{\beta})^T \Sigma^{-1}(\theta)(Y(s) - X\hat{\beta})
\]  
\[ \text{s.t } \hat{\beta} = (X^T \Sigma^{-1}(\theta) X)^{-1} X^T \Sigma^{-1}(\theta) Y(s) \]

(1)
2.2.2 Efficient Estimation Strategies for Large Spatial Datasets

Optimization of the log-likelihood in a universal kriging model involves inverting the covariance matrix which is \(O(n^3)\). Recent work in making spatial statistics computationally feasible has relied on clever ways of structuring the covariance matrix to reduce the computational complexity in calculating its inverse \cite{Banerjee2008, Kaufman2008}. Here, we take an approach following \cite{Cressie2008} and decompose the spatial covariance matrix \(\Sigma(\sigma^2) = \sigma^2 S(s)S^T(s)\). This decomposition leads to the spatial mixed effects model (2).

\[
Z(s) = X\beta + S(s)\eta + \epsilon \tag{2}
\]

where \(S(s) \in \mathbb{R}^{n \times k}\) are spatial basis functions and \(\eta \sim \text{i.i.d.} (0, \sigma^2 I_k)\), \(\eta \perp \epsilon\) are the spatial random effects. The class of basis functions which can be constructed by this procedure are detailed in section 3.1 of \cite{Cressie2008} and are able to approximate covariance functions often used in spatial statistics \cite{Nychka2002}. Under the spatial mixed effects model,

\[
Z(s) \sim (X\beta, \sigma^2 S(s)S^T(s) + \tau^2 I_n).
\]

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

\[
E[Z(s_0)|\hat{\eta}] = X_0\hat{\beta} + S(s_0)\hat{\eta},
\]

where \(\hat{\beta}\) is the best linear unbiased estimator, and \(\hat{\eta}\) is the best linear unbiased predictor.

The spatial random effect \(\hat{\eta}\) can be interpreted as a penalized regression estimator \cite[4.5.3]{Ruppert2003}. By Henderson’s justification \cite{Robinson1991}, optimizing \(\hat{\beta}\) and \(\hat{\eta}\) leads to minimizing the criteria.
\[ \| Y(s) - X\beta - S(s)\eta \|_2^2 + \frac{\tau^2}{\sigma^2} \| \eta \|_2^2, \]

which can be interpreted as ridge regression on \( \hat{\eta} \) with penalty \( \lambda = \frac{\tau^2}{\sigma^2} \).

### 2.3 Spatially Adjusted Trees

Additive models combining regression trees and kriging can be formulated as

\[ Y(s) = t(X) + \nu(s), \tag{3} \]

with \( t(X) \) the regression tree constructed from the covariates and \( \nu(s) \sim N(0, \Sigma(\theta)) \) a realization of a Gaussian process.

Under the additive model (3), \( Y(s) \sim N(t(X), \Sigma(\theta)) \). By maximum likelihood, we wish to find a regression tree estimate \( \hat{t}(X) \) and covariance parameters \( \hat{\theta} \) such that

\[ \{ \hat{t}(X), \hat{\theta} \} = \arg\max_{t(X), \theta} \left[ -\frac{1}{2} \log |\Sigma(\theta)| - \frac{1}{2} (Y(s) - t(X))^T \Sigma^{-1}(\theta) (Y(s) - t(X)) \right] \tag{4} \]

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

\[ \arg\max_{\theta} \left[ -\frac{1}{2} \log |\Sigma(\theta)| - \frac{1}{2} (Y(s) - \hat{t}(X|\Sigma(\theta)))^T \Sigma^{-1}(\theta) (Y(s) - \hat{t}(X|\Sigma(\theta))) \right] \]

s.t \( \hat{t}(X|\Sigma(\theta)) = \arg\min_{t(X|\Sigma(\theta))} (Y(s) - t(X|\Sigma(\theta)))^T \Sigma^{-1}(\theta) (Y(s) - t(X|\Sigma(\theta))) \) \tag{5} \]

The dependence of the profiled spatially adjusted regression tree on the spatial correlation matrix is emphasized as \( \hat{t}(X|\Sigma(\theta)) \). We note similarities of this optimization problem to universal kriging and traditional regression trees. In universal kriging, \( t(X) = X\beta \) and the profile likelihood optimization criteria selects \( \beta \) which maximizes \( \ell(\beta, \theta_0|Y(s)) \) for some
fixed $\theta_0$. On the other hand, if we ignore the spatial process and let $\Sigma(\theta) = I_n$, there are no covariance parameters to maximize over and we would build a normal regression tree which minimizes mean squared error. Thus, a spatially adjusted regression tree should be built to minimize $[5]$ for some fixed $\theta_0$.

2.3.1 A Computationally Feasible Spatially Adjusted Tree Building Algorithm

We propose a novel, computationally feasible spatially adjusted tree building algorithm, which modifies the normal tree building procedure in order to construct a spatially adjusted tree which aims to minimize $[5]$. Note that this algorithm is not guaranteed to find the actual minimizer but instead aims to minimize $[5]$ by a greedy approach analogous to recursive binary splitting. In Section 2.1.1, we showed that each tree can be written as a linear combination of the tree design matrix $C$ and their corresponding weights $\pi$ so we can re-write $[5]$ as:

$$\ell(C^k, \pi^k) = (Y(s) - C^k \pi^k)^T \Sigma^{-1} (Y(s) - C^k \pi^k)$$

(6)

By profile likelihood, we define the ”characteristic matrix” for the spatial tree building algorithm $\Omega^k$ [7] which depends only on the tree design matrix.

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

(7)

The loss depends only on $\Omega^k$, the characteristic matrix, and the observations $Y(s)$. By blockwise inversion [Harville (1998), Thm 8.5.11], updating $\Omega^{k+1}$ only depends on the indicator vector noting which observations are in the terminal node created by the $k + 1^{st}$ branch, $c^A$ and the existing sufficient statistic $\Omega^k$ [8].

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

(8)
Using this fact, the change in loss between $C^k$ and $C^{k+1} = \begin{bmatrix} C^k & c^A \end{bmatrix}$ is easily shown to be

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

Since $((c^A)^T \Omega^k c^A)^{-1}$ is a scalar, calculating the change in loss for each candidate vector for each possible split and updating $\Omega^{k+1}$ can be evaluated in $O(n^2)$ and bypasses the $O(k^3)$ matrix inversion from the least squares estimator.

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

### 2.4 Random Spatial Forests: Pseudo-Likelihood Approach

Our algorithm describes a process for constructing spatially adjusted trees for known $\theta_0$, and we can construct spatial random forests estimates by aggregating over these trees. In practice however, $\theta$ is unknown. Since random forests estimates the expectation of an infinite tree (Hastie et al. (2005), 15.3.4), we propose a pseudo-likelihood approach where we replace the regression tree with its bagged random forests estimate in (5).

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

In order to simplify estimation of the covariance parameters we approximate the spatial covariance as in section 2.2.2. Let

$$V(\kappa, \delta) = \kappa R(\delta), \quad R(\delta) = (\delta S(s)S^T(s) + (1 - \delta)I_n)$$

11
Algorithm 1 Spatially Adjusted Tree Building Algorithm

1. set $C^0 = \begin{bmatrix} 1_n \end{bmatrix}$

2. Given $\Sigma$, set the initial value for $\Omega^0 = \Sigma^{-1} - C^0 \left( (C^0)^T \Sigma^{-1} C^0 \right)^{-1} (C^0)^T \Sigma^{-1}$

3. For $i = 1, 2, \ldots$
   (a) Take a random sample of the covariates $X_r \subset \{X_1, X_2, \ldots, X_p\}$
   (b) Check each of the $i$ existing terminal nodes for a new terminal node created by a decision rule based on the sampled covariates $X_r$, to create a candidate set of possible splits $c^A$.
   (c) Find the candidate split $c^A \in c^A$ which maximizes the change in loss
   \[ Y(s)^T (\Omega^k c^A \left( (c^A)^T \Omega^k c^A \right)^{-1} (c^A)^T \Omega^k) Y(s) \]
   (d) Update $\Omega^{k+1} = \Omega^k - \Omega^k c^A \left( (c^A)^T \Omega^k c^A \right)^{-1} (c^A)^T \Omega^k$
   (e) Repeat steps (a)-(c) until no more the maximum number of splits is exceeded or there are no more branches can be split without creating a branch with less than $m$ observations

12
The profile log-likelihood can easily be shown to be written as a function of a single parameter $\delta$ by profiling out $\hat{f}(X|R(\delta))$ and $\hat{\kappa}$ as

$$
\ell(\delta) = -\frac{n}{2} \log (\hat{\kappa}) - \frac{1}{2} |R(\delta)| - \frac{1}{2\hat{\kappa}} \left( Y(s) - \hat{f}(X|R(\delta)) \right)^T R^{-1}(\delta) \left( Y(s) - \hat{f}(X|R(\delta)) \right) \\
\text{s.t. } \hat{f}(X|R(\delta)) = \arg\min_{f(X|R(\delta))} \left( Y(s) - f(X|R(\delta)) \right)^T R^{-1}(\delta) \left( Y(s) - f(X|R(\delta)) \right) \\
\text{and } \hat{\kappa} = \frac{1}{n} (Y(s) - \hat{f}(X|R(\delta)))^T R^{-1}(\delta) (Y(s) - \hat{f}(X|R(\delta)))
$$

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

### 2.5 Random Spatial Forests: Non-Parametric Estimation

In the previous section, we derived an additive model using a pseudo-likelihood approach to integrate random forests into a likelihood model. However, it is not easy to interpret $\eta$ as a random effect since it is difficult to imagine the data generating mechanism that might give rise to such fields [Hodges (2016)]. In this case, modeling the spatial process using a random effect is a form of regularization and pseudo-likelihood gives us a way to estimate the parameter $\delta$. The goal of modeling the air pollution surface in epidemiological studies is to produce accurate estimates for individuals at unobserved locations, which can be viewed as a prediction problem. An alternative criterion when prediction accuracy is desired, which is often the case in many statistical learning applications, is to minimize the expected mean squared test error:

$$\arg\min_\delta \text{E} \left[ \| Y - \hat{Y}(s, \delta) \|^2 \right].$$
Noting the relationship between $\delta$ and a penalty in section 2.2.2, a natural non-parametric approach would be to select the tuning parameter $\delta$ for the additive model by $k$-fold cross-validation in order to find $\delta$ which minimizes the out of sample test error.

The addition of running $k$-fold cross-validation to estimate expected test error for each candidate $\delta$ results in a substantial increase in computational time. But a unique property of random forests is that since random forests only uses "out-of-bag" samples in its estimation, the resulting function $\hat{Y}(s)$ is equivalent to its cross-validated estimate [Hastie et al. (2005)]. This is desirable since the mean squared error of $\hat{Y}(s)$ on the training set is equivalent to its expected test error and makes $k$-fold cross-validation is unnecessary, reducing the computational burden. In order to leverage this property, we propose applying the random forests algorithm to aggregate our spatially adjusted trees and each their associated spatial smoothers as:

$$\hat{Y}_i(s, \delta) = \frac{1}{B} \sum_{i=1}^{B} \left[ \hat{t}_i(X_i | R_i(\delta)) + S_i(s)\hat{\eta}_i \right]$$

For each bootstrap sample, we can use our tree building algorithm in section 2.3.1 to estimate both the spatially adjusted tree $t_i(X_i | R_i(\delta))$ and its associated spatial random effect $\hat{\eta}_i$ by its BLUP. Over a grid of $\delta \in [0, 1]$, we fit $\hat{Y}(s, \delta)$ using our spatially adjusted tree building algorithm including the spatial smoother in each sample and select

$$\arg\min_\delta \|Z(s) - \hat{Y}(s, \delta)\|_2^2.$$  

3 Simulation Study

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

3.1 Generating the Observed Surface

For each simulation, we constructed a fixed exposure surface from an additive model of a function of GIS covariates, $f(X)$ and a fixed realization of a Gaussian process with exponential covariance process, $\nu(s)$, with range randomly generated between 10%−20% of the maximum distance between points. We considered generating three different functions of the covariates (1) sparse: a linear combination of 10 randomly sampled covariates where each covariate is scaled to have the same total contribution to the overall signal, (2) dense a linear combination of all the covariates where the coefficients $\beta \sim (0, 1)$, and (3) nonlin: $f(X)$ is a combination of 10 interactions of 2-4 covariates.

The observed surface is constructed as

$$Y(s) = \gamma f(X) + \nu(s),$$

where the parameter $\gamma$ controls the proportion of variance attributable to the GIS covariates. We consider two scenarios:

1. Strong Covariates: 65% of the generated process is due to the covariates.

2. Weak Covariates: 35% of the generated process is due to the covariates.

3.2 Methods combining Random Forests with Spatial Smoothing

For our examples we formulate the spatial basis functions using TPRS following Olives et al. (2014). This choice is arbitrary, and as noted in 2.2.2 one could consider selecting...
alternative spatial basis functions. We selected TPRS as an alternative to kriging as there
is an equivalence between thin plate regression splines (TPRS) and kriging with a Matern-
class covariance with infinite range \( \text{Nychka2000} \). We compare seven methods:

1. Random Forests (RF)- implemented using the \texttt{randomForests} package

2. Spatial Smoothing (TPRS) - implemented by the \texttt{mgcv} package

3. Random Forests plus Spatial Smoothing (RF-TPRS): 2 step approach where first RF
   is run, then TPRS if fit to the RF residuals.

4. Spatial Smoothing plus Random Forests (TPRS - RF): 2 step approach where first
   TPRS is run, then RF is applied to the residuals from TPRS

5. Random Spatial Forests - Pseudo-Likelihood (SpatRF-PL), section 2.4

6. Random Spatial Forests - Non-Parametric (SpatRF-NP), section 2.5

3.3 Evaluating Reconstruction Accuracy of the Different Methods

For each of the six combinations for type of function and proportion of variance explainable
by the covariates, we generate a single observed surface \( Y(s) \) and hold out 200 points for
validation. 30 times each, 150 points are randomly sampled to train seven different models
to compare on. Training points are observed with independent measurement error

\[
Z(s_{\text{train}}) = Y(s_{\text{train}}) + \epsilon, \quad \epsilon \sim N(0, \tau^2 I_{n_{\text{train}}})
\]
where $\tau^2$ is randomly generated to be between 10% and 25% of the total variance. Prediction accuracy of each model is evaluated on the held-out points $Y(s_{test})$ as $R^2$ defined as

$$R^2 = \max \left\{ 0, 1 - \frac{RMSE(\hat{Y}(s_{test}), Y(s_{test}))^2}{\text{Var}(Y(s_{test}))} \right\}$$

We report the average $R^2$ for each method from 30 different randomly sampled training points. This is further repeated 30 times for each of the six scenarios for generating the observed surface. Density plots of average $R^2$ for each method in each of these scenarios are shown in Figure 1 [Figure 1 about here]

### 3.4 Simulation Results

In the *strong covariates* scenario, RF does better than TPRS alone while this relationship is reversed in the *weak covariates* scenario. This demonstrates, rather unsurprisingly, that when a large percentage of the observed surface can be explained by the covariates, constructing a surface using a function of the covariates by RF performs better than ignoring the covariates and applying a spatial smoother. On the other hand, when the covariates are responsible for a small percentage of the total variation using only the covariates via RF leads to worse prediction accuracy than simply applying TPRS alone. RF-TPRS and TPRS-RF both do better than either RF or TPRS, showing that in these scenarios fitting an additive model with both RF and TPRS does better than fitting either alone. Comparing RF-TPRS and TPRS-RF highlights the importance of the optimization approach. Although RF-TPRS and TPRS-RF are both composed of a random forests and thin plate regression spline, the order of estimation can have a large impact on the models prediction accuracy. When the covariates are responsible for a large percentage of variability in
the observed surface RF-TPRS performs noticeably better than TPRS-RF, and vice versa when the covariates explain a small portion of the variance. Our methods demonstrate how constructing random forests allowing for spatial correlation leads to more accurate predictions than either two-step approach regardless of how much variability can be explained by the covariates. In our simulations, SpatRF-PL and SpatRF-NP have better prediction accuracy than RF-TPRS and TPRS-RF in all scenarios. Comparing our two methods, SpatRF-NP performs slightly better than SpatRF-PL in both cases.

4 Application to Sub-Species of PM$_{2.5}$

Following Bergen et al. (2013), we only include CSN and IMPROVE monitors with at least 10 data points per quarter and no more than 45 days between consecutive measurements. Si and S measurements were averaged over 01/01/2009–12/31/2009, while EC/OC consisted of measurements from 204 IMPROVE and CSN monitors averaged over 01/01/2009–12/31/2009, and measurements from 51 CSN monitors averaged over 05/01/2009 – 04/30/2010. Annual averages were square-root transformed prior to modeling.

In addition to methods used in the simulations, we include universal kriging estimates which deal with the high dimensionality of the covariates by pre-processing the covariates by partial least squares (UK-PLS) and use an exponential covariance matrix. This technique was employed in the original analysis by Bergen et al. (2013) and is commonly employed in many land-use regression settings. The reported $R^2$ values for UK-PLS are not identical to those reported in by Bergen et al. (2013) since they calculated $R^2$ on the square root scale, while we transform predictions back to the original scale before computing $R^2$ values. Additionally, we examine random forests with spatial information included (RF w/ TPRS).
This approach is similar to Hengl et al. (2018), where spatial basis functions are included as covariates accounting for geographic proximity between observations.

Surface reconstruction accuracy of each method is assessed by comparing predictions generated from ten-fold cross-validation. Monitoring sites were randomly assigned to one of ten cross-validation groups. Each group is held out as a test set and observations in the remaining groups are used as a training set to fit the model and generate test set predictions using each of the seven methods to compare. Each group is used as a test set once to obtain predicted values for the entire data set. Performance of each model is based on their average cross-validated $R^2$ over ten separate cross-validation runs in Table 1. [Table 1 about here] [Figures 2, ??, ?? about here]

Cross-validated prediction accuracy over the different components of PM$_{2.5}$ show similar findings to our simulation results. When a large proportion of the variance can be explained by the covariates, demonstrated by EC and OC, RF performs better than TPRS and RF-TPRS has improved cross-validated accuracy over TPRS-RF. SpatRF-NP and SpatRF-PL show small but noticeable improvements over RF-TPRS and are more accurate for both pollutants. In these examples using random forests instead of using a linear model with dimension reduction on the covariates by PLS can yield noticeable improvements in prediction accuracy as SpatRF-PL and SpatRF-NP show noticeable increases in cross-validated $R^2$ over UK-PLS.

Si and S are examples where spatial smoothing is able to model a larger proportion of the variance than the covariates alone. In both of these cases, RF performs worse $R^2$ than TPRS and RF-TPRS has lower cross-validated accuracy TPRS-RF. Interestingly, TPRS alone has better cross-validated accuracy than RF-TPRS suggesting that applying spatial smoothing to the residuals from random forests does not guarantee that the combined
approach is more accurate than either individual method alone. For S, TPRS-RF, UK-PLS, SpatRF-PL, and SpatRF-NP all do quite well (CV $R^2$ 0.941-0.947). Evaluating the different methods on Si, SpatRF-NP has the highest cross-validated $R^2$.

In order to closely examine how efficient combining random forests and spatial smoothing aids in surface reconstruction, we break down the additive model estimates into its parts. From the additive RF and TPRS components of the model (Figure 4), it is clear the smoothing spline is unable to pick up the sharp fluctuations in EC that occur in large cities, for example, Los Angeles and New York. In these cases, the use of GIS covariates can lead to large improvements in prediction accuracy. RF-TPRS performs better than TPRS-RF, and RF-TPRS shows a noticeable improvement over UK-PLS. However, by modifying the random forest algorithm to account for estimation of the spatial process we get slight improvements over RF-TPRS using our spatially adjusted random forest algorithms.

In cases like silicon and sulfur, the overall patterns appear to have a large scale structure. Sulfur has a large east to west relationship, peaking near Pittsburgh and gives a clear example of when attempting to use machine learning methods performs worse than traditional geostatistical approaches where spatial smoothing is employed. Figure 5 shows the RF and TPRS components for each of the additive models. Although much of the variation is explained by geographic location, RF aims to use its geographic covariates to explain the east to west relationship of the model. Since the standard RF algorithm does not account for the spatial process, it attempts to model large scale east to west variation in a sub-optimal matter.

For all PM$_{2.5}$ sub-species, SpatRF-NP and SpatRF-PL perform as well or better than any of the two step approaches, RF w/TPRS, and UK-PLS. SpatRF-NP and SpatRF-PL perform similarly for EC and S, but for pollutants where cross-validated $R^2$ is lower in
general (OC and Si), we see a noticeable improvement in using SpatRF-NP over SpatRF-PL. Based off our findings on PM$_{2.5}$ sub-species and simulations, of the two proposed approaches we recommend using SpatRF-NP.

5 Discussion

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

Here, we examined a random forests algorithm using our novel tree building algorithm to adjust for spatial correlation. Another popular tree based ensemble method is boosting, and it would be straightforward to apply our tree building algorithm to boost spatially adjusted trees. Our tree building algorithm adjusting for correlation is also not restricted to estimation in spatial applications. For example, prediction problems where it is desirable to adjust for correlation occurs in other application such as network-linked data Li, Levina, Zhu et al. (2019).

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

In this paper, we took advantage of recent computational developments in spatial statistics to reduce the parameterization of the covariance to a single parameter $\delta$, leading to a simple optimization routine by grid search. We note here that this is not required, for example, one could consider using a normal kriging covariance with an exponential covariance function and select the parameters by grid search, but adding additional parameters becomes computationally expensive since the number of points to consider scales exponentially. Bayesian optimization and covariance matrix adaptation - evolution strategy have been used in the machine learning literature for gradient free optimization of ”black-box” prediction models with stochastic function evaluation where multiple tuning parameters need to be selected. Both of these methods can be applied to random spatial forests and are easily parallelizable to make optimization feasible for more complex covariance function parameterizations.

Extensions for future work include the use of random spatial forests for more complex spatial misalignment problems. Bose et al. (2018) consider dimension reduction of multiple pollutants by principal components which are adaptively estimated to be predictable by a pre-specified set of covariates. Similarly, Keller et al. (2017) proposed predictive k-means clustering which generates clusters that are predictable by covariates. Szpiro et al. (2010) developed a spatio-temporal model which decouples the space and time components by first deriving temporal basis functions in order to account for temporal correlation and estimates spatially continuous fields associated with each of these basis functions. All of these methods jointly estimate a dimension reduced summary of a complex problem, in the
form of clusters, principal components, or temporal basis functions, and associated predictive function of the covariates used to predict their values at unobserved locations, and it would be of interest to examine whether these methods could be modified to incorporate spatial random forests in the prediction step.

References

Arain, M., Blair, R., Finkelstein, N., Brook, J., Sahsuvaroglu, T., Beckerman, B., Zhang, L. & Jerrett, M. (2007), ‘The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies’, *Atmospheric Environment* 41(16), 3453–3464.

Banerjee, S., Gelfand, A. E., Finley, A. O. & Sang, H. (2008), ‘Gaussian predictive process models for large spatial data sets’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(4), 825–848.

Bergen, S., Sheppard, L., Sampson, P. D., Kim, S.-Y., Richards, M., Vedal, S., Kaufman, J. D. & Szpiro, A. A. (2013), ‘A national prediction model for pm2.5 component exposures and measurement error–corrected health effect inference’, *Environmental health perspectives* 121(9), 1017–1025.

Berrocal, V. J., Guan, Y., Muyskens, A., Wang, H., Reich, B. J., Mulholland, J. A. & Chang, H. H. (2019), ‘A comparison of statistical and machine learning methods for creating national daily maps of ambient pm$_{2.5}$ concentration’, *arXiv preprint arXiv:1904.08931*. 

23
Bose, M., Larson, T. & Szpiro, A. A. (2018), ‘Adaptive predictive principal components for modeling multivariate air pollution’, *Environmetrics* **29**(8), e2525.

Breiman, L. (2001), ‘Random forests’, *Machine learning* **45**(1), 5–32.

Chen, G., Li, S., Knibbs, L. D., Hamm, N. A., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M. J. & Guo, Y. (2018), ‘A machine learning method to estimate pm2.5 concentrations across china with remote sensing, meteorological and land use information’, *Science of the Total Environment* **636**, 52–60.

Cressie, N. & Johannesson, G. (2008), ‘Fixed rank kriging for very large spatial data sets’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 209–226.

Fox, E. W., Hoef, J. M. V. & Olsen, A. R. (2018), ‘Comparing spatial regression to random forests for large environmental data sets’, *arXiv preprint arXiv:1812.10236* .

Harville, D. A. (1998), ‘Matrix algebra from a statistician’s perspective’.

Hastie, T., Tibshirani, R., Friedman, J. & Franklin, J. (2005), ‘The elements of statistical learning: data mining, inference and prediction’, *The Mathematical Intelligencer* **27**(2), 83–85.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. & Gräler, B. (2018), ‘Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables’, *PeerJ* **6**, e5518.

Hodges, J. S. (2016), *Richly parameterized linear models: additive, time series, and spatial models using random effects*, Chapman and Hall/CRC.
Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P. & Briggs, D. (2008), ‘A review of land-use regression models to assess spatial variation of outdoor air pollution’, *Atmospheric environment* **42**(33), 7561–7578.

Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J. & Liu, Y. (2017), ‘Estimating pm2.5 concentrations in the conterminous united states using the random forest approach’, *Environmental science & technology* **51**(12), 6936–6944.

Kaufman, C. G., Schervish, M. J. & Nychka, D. W. (2008), ‘Covariance tapering for likelihood-based estimation in large spatial data sets’, *Journal of the American Statistical Association* **103**(484), 1545–1555.

Kaufman, J. D., Adar, S. D., Barr, R. G., Budoff, M., Burke, G. L., Curl, C. L., Daviglus, M. L., Roux, A. V. D., Gassett, A. J., Jacobs Jr, D. R. et al. (2016), ‘Association between air pollution and coronary artery calcification within six metropolitan areas in the usa (the multi-ethnic study of atherosclerosis and air pollution): a longitudinal cohort study’, *The Lancet* **388**(10045), 696–704.

Keller, J. P., Drton, M., Larson, T., Kaufman, J. D., Sandler, D. P. & Szpiro, A. A. (2017), ‘Covariate-adaptive clustering of exposures for air pollution epidemiology cohorts’, *The annals of applied statistics* **11**(1), 93.

Li, R., Cui, L., Meng, Y., Zhao, Y. & Fu, H. (2019), ‘Satellite-based prediction of daily so2 exposure across china using a high-quality random forest-spatiotemporal kriging (rf-stk) model for health risk assessment’, *Atmospheric Environment* **208**, 10–19.

Li, T., Levina, E., Zhu, J. et al. (2019), ‘Prediction models for network-linked data’, *The Annals of Applied Statistics* **13**(1), 132–164.
Liu, Y., Cao, G., Zhao, N., Mulligan, K. & Ye, X. (2018), ‘Improve ground-level pm2. 5 concentration mapping using a random forests-based geostatistical approach’, *Environmental pollution* **235**, 272–282.

Nychka, D. W. (2000), ‘Spatial-process estimates as smoothers’, *Smoothing and regression: approaches, computation, and application* **329**, 393.

Nychka, D., Wikle, C. & Royle, J. A. (2002), ‘Multiresolution models for nonstationary spatial covariance functions’, *Statistical Modelling* **2**(4), 315–331.

Olives, C., Sheppard, L., Lindström, J., Sampson, P. D., Kaufman, J. D. & Szpiro, A. A. (2014), ‘Reduced-rank spatio-temporal modeling of air pollution concentrations in the multi-ethnic study of atherosclerosis and air pollution’, *The annals of applied statistics* **8**(4), 2509.

Power, M. C., Adar, S. D., Yanosky, J. D. & Weuve, J. (2016), ‘Exposure to air pollution as a potential contributor to cognitive function, cognitive decline, brain imaging, and dementia: a systematic review of epidemiologic research’, *Neurotoxicology* **56**, 235–253.

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

Rolf, E., Jordan, M. I. & Recht, B. (2020), ‘Post-estimation smoothing: A simple baseline for learning with side information’, *arXiv preprint arXiv:2003.05955*.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003), *Semiparametric regression*, number 12, Cambridge university press.
Sack, C., Vedal, S., Sheppard, L., Raghu, G., Barr, R. G., Podolanczuk, A., Doney, B., Hoffman, E. A., Gassett, A., Hinckley-Stukovsky, K. et al. (2017), ‘Air pollution and subclinical interstitial lung disease: the multi-ethnic study of atherosclerosis (mesa) air–lung study’, *European Respiratory Journal* **50**(6), 1700559.

Saucy, A., Röösli, M., Künzli, N., Tsai, M.-Y., Sieber, C., Olaniyan, T., Baatjies, R., Jeebhay, M., Davey, M., Flückiger, B. et al. (2018), ‘Land use regression modelling of outdoor no2 and pm2.5 concentrations in three low income areas in the western cape province, south africa’, *International journal of environmental research and public health* **15**(7), 1452.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., De’Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M. et al. (2019), ‘Estimation of daily pm10 and pm2.5 concentrations in italy, 2013–2015, using a spatiotemporal land-use random-forest model’, *Environment international* **124**, 170–179.

Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D. & Kaufman, J. D. (2010), ‘Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies’, *Environmetrics* **21**(6), 606–631.

Xu, H., Bechle, M. J., Wang, M., Szpiro, A. A., Vedal, S., Bai, Y. & Marshall, J. D. (2018), ‘National pm2.5 and no2 exposure models for china based on land use regression, satellite measurements, and universal kriging’, *arXiv preprint arXiv:1808.09126*.
Figure 1: Simulation Results: Each point represents the average $R^2$ on the validation points from different sampled training points. The box and whisker summarizes the prediction accuracy of the method on different simulated surfaces from a variety of generating functions. The box shows the 25th, median, and 75th percentiles, and the whisker shows the most extreme value within 1.5 times the range of the quartiles. Each point in the boxplot represents the average $R^2$ on the fixed 200 validation points over 30 repeated samples.
Figure 2: Predicted Elemental Carbon concentrations (in \( \mu g/m^3 \)) across the continental United States. Points are observed annual averages at monitoring locations. Top Left: Universal Kriging with dimensions reduction by Partial Least Squares, Top Right: Random Forests with Spatial Information included as Covariates, Middle Left: Thin Plate Regression Splines followed by Random Forests, Middle Right: Random Forests followed by Thin Plate Regression Splines, Bottom Left: Random Spatial Forests by Pseudo-Likelihood, Bottom Right: Random Spatial Forests by Non-Parameteric Estimation.
Figure 3: Predicted Sulfur concentrations (in $\mu g/m^3$) across the continental United States. Points are observed annual averages at monitoring locations. Top Left: Universal Kriging with dimensions reduction by Partial Least Squares, Top Right: Random Forests with Spatial Information included as Covariates, Middle Left: Thin Plate Regression Splines followed by Random Forests, Middle Right: Random Forests followed by Thin Plate Regression Splines, Bottom Left: Random Spatial Forests by Pseudo-Likelihood, Bottom Right: Random Spatial Forests by Non-Parameteric Estimation
Figure 4: RF and TPRS components for Elemental Carbon concentrations (in $\mu$g/m$^3$). Maps on the left side show the RF component of the additive model while maps on the right show the TPRS estimate. Each row is a different estimation order, top: RF first, then TPRS, middle: TPRS first, then RF, bottom: SpatRF-NP which jointly estimates the RF and TPRS components.
Figure 5: RF and TPRS components for Sulfur (in $\mu g/m^3$). Maps on the left side show the RF component of the additive model while maps on the right show the TPRS estimate. Each row is a different estimation order, top: RF first, then TPRS, middle: TPRS first, then RF, bottom: SpatRF-NP which jointly estimates the RF and TPRS components.
|     | UK-PLS | RF w/ TPRS | RF-TPRS | TPRS-RF | SpatRF - PL | SpatRF - NP |
|-----|--------|------------|---------|---------|-------------|-------------|
| EC  | 0.730  | 0.799      | 0.817   | 0.713   | 0.828       | 0.827       |
| OC  | 0.581  | 0.599      | 0.593   | 0.569   | 0.594       | 0.627       |
| Si  | 0.557  | 0.572      | 0.534   | 0.586   | 0.583       | 0.598       |
| S   | 0.947  | 0.928      | 0.889   | 0.941   | 0.943       | 0.943       |

Table 1: Ten-fold cross-validated prediction accuracy, summarized by $R^2$, of each method for PM$_{2.5}$ components Elemental Carbon (EC), and Organic Carbon (OC), Silicon (Si), Sulfur (S) collected by AQS and IMPROVE monitoring networks from 2009-2010.