PLOS ONE

RESEARCH ARTICLE

Comparing spatial regression to random forests for large environmental data sets

Eric W. Fox[1]*, **Jay M. Ver Hoef**[2], **Anthony R. Olsen**[3]

1 Department of Statistics and Biostatistics, California State University East Bay, Hayward, California, United States of America, **2** National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Oceanic and Atmospheric Administration, Seattle, Washington, United States of America, **3** National Health and Environmental Effects Research Laboratory, Western Ecology Division, United States Environmental Protection Agency, Corvallis, Oregon, United States of America

* eric.fox@csueastbay.edu

Citation: Fox EW, Ver Hoef JM, Olsen AR (2020) Comparing spatial regression to random forests for large environmental data sets. PLoS ONE 15(3): e0229509. https://doi.org/10.1371/journal.pone.0229509

Editor: Yang Li, University of Minnesota Duluth, UNITED STATES

Received: September 15, 2019

Accepted: February 7, 2020

Published: March 23, 2020

Data Availability Statement: The macroinvertebrate multimetric index (MMI) and StreamCat data sets, which were used estimate the models in this paper, are available from a GitHub repository: (https://github.com/ericwfox/slmrf) The data can be accessed by installing the R package using the install_github() command. For users that are not familiar with R, the data are also available in CSV format in the \inst folder of the repository (data files are named mmi_data.csv and streamcat_preds.csv). Details about the 2008/09 National Rivers and Streams Assessment MMI data are provided in the technical report: USEPA.

Abstract

Environmental data may be "large" due to number of records, number of covariates, or both. Random forests has a reputation for good predictive performance when using many covariates with nonlinear relationships, whereas spatial regression, when using reduced rank methods, has a reputation for good predictive performance when using many records that are spatially autocorrelated. In this study, we compare these two techniques using a data set containing the macroinvertebrate multimetric index (MMI) at 1859 stream sites with over 200 landscape covariates. A primary application is mapping MMI predictions and prediction errors at 1.1 million perennial stream reaches across the conterminous United States. For the spatial regression model, we develop a novel transformation procedure that estimates Box-Cox transformations to linearize covariate relationships and handles possibly zero-inflated covariates. We find that the spatial regression model with transformations, and a subsequent selection of significant covariates, has cross-validation performance comparable to random forests. We also find that prediction interval coverage is close to nominal for each method, but that spatial regression prediction intervals tend to be narrower and have less variability than quantile regression forest prediction intervals. A simulation study is used to generalize results and clarify advantages of each modeling approach.

Introduction

As we enter the age of "big data" [1, 2], innovative statistical methods are required for insights from massive data sets [3]. While big data is an abstract concept [4], data for a statistical analysis are generally prepared as tables, with records down the rows, and variables across the columns [5]. While we immediately recognize big data problems for large numbers of records, there are also issues when there are large numbers of columns [6]. We are interested in a spatial data set in the United States of national importance based on an aquatic health index, where there are thousands of rows and hundreds of columns. Two leading candidates for analyzing such data are random forests [7], because it handles large numbers of covariates with nonlinear relationships, and spatial regression [8], because it accounts for possible spatial

National Rivers and Streams Assessment 2008-2009 Technical Report (EPA/841/R-16/008); 2016. URL: (https://www.epa.gov/national-aquatic-resource-surveys/nrsa) Details about the StreamCat data set are provided in the paper: Hill RA, Weber MH, Leibowitz SG, Olsen AR, Thornbrugh DJ. The Stream-Catchment (StreamCat) Dataset: A database of watershed metrics for the conterminous United States. Journal of the American Water Resources Association. 2016;52(1):120–128. URL: (https://www.epa.gov/national-aquatic-resource-surveys/streamcat).

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

autocorrelation. The ultimate goal of the analysis is to predict the aquatic health index at over one million watersheds throughout the country, while also gaining some understanding of which covariates are important. The goal of this article is to use best practices for analyzing the data with spatial regression and random forests, and then compare the methods.

Motivating data set

For this study, we use a data set containing the biological condition of 1859 sites from the US Environmental Protection Agency's 2008/09 National Rivers and Streams Assessment (NRSA) [9]. The NRSA used a generalized random-tessellation design [10] to collect a spatially-balanced and representative sample of stream sites across the conterminous United States (CONUS). The target population for the design consisted of all rivers and streams within the CONUS that had flowing water during the study period, which extended between April to September of 2008/09. Benthic macroinverabrates (e.g., aquatic insects, crustaceans, and worms) were sampled to determine the biological condition of stream sites. A multimetric index (MMI) was developed for the NRSA to summarize several measures of the condition of macroinvertebrate assemblages (e.g., taxonomic composition, diversity, tolerance to disturbance, etc.) into a combined index. The reported MMI values were calculated by summing six individual measures, or 'metrics', and then normalized to a 0-100 scale [11]. The individual metrics used for the MMI were selected separately for each of the nine ecoregions [12] to account for some of the natural variation in climate, geology, hydrology and soils among stream sites. The ecoregion boundaries and locations of the 2008/09 NRSA stream sites are shown in Fig 1. A comprehensive description of the MMI developed for the NRSA can be found in [11, 13, 14].

For modeling the MMI we use a large suite of 209 covariates from the Stream-Catchment (StreamCat) data set [15] (publicly available at https://www.epa.gov/national-aquatic-resource-surveys/streamcat). StreamCat contains upstream landscape features (e.g., topograpy,

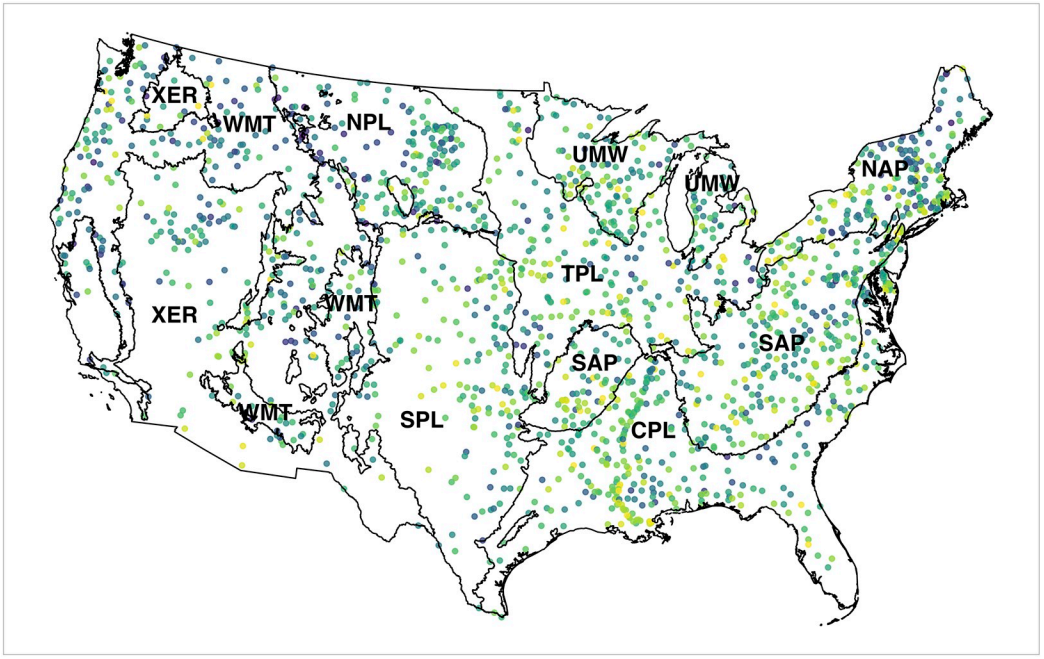

Fig 1. Locations of 2008/09 NRSA stream sites with point colors corresponding to sampled MMI scores. Ecoregions: Coastal Plains (CPL), Northern Appalachians (NAP), Northern Plains (NPL), Southern Appalachians (SAP), Southern Plains (SPL), Temperate Plains (TPL), Upper Midwest (UMW), Western Mountains (WMT), and Xeric (XER).

https://doi.org/10.1371/journal.pone.0229509.g001

precipitation, landscape imperviousness, urban and agricultural land use) for 2.6 million stream reaches across the CONUS and allows for spatially explicit prediction. The variables in StreamCat can be linked to the National Hydrography Dataset Plus Version 2 (NHDPlusV2) [16] and are available at two spatial scales: local catchment and full contributing watershed. [15] defines a "catchment" as the local drainage area for an individual NHDPlusV2 stream segment, excluding upstream drainage area; and a "watershed" as the set of hydrologically connected catchments that contribute flow to a given catchment (i.e., catchment plus upstream catchments). Note that we only make MMI predictions for the 1.1 million perennial stream reaches (as designated in NHDPlusV2) since the sample frame for the NRSA is limited to these types of streams. Descriptions of StreamCat covariates used in this study are provided in the Supplement (S1 File).

The MMI response data and StreamCat covariates, which were used to estimate the models in this study, are available at https://github.com/ericwfox/slmrf.

Literature review

There have been surprisingly few attempts to compare random forests and spatial regression. Most found random forests superior to various forms of linear regression with autocorrelated errors [17–21], although [22], and [23] found spatial regression outperformed random forests. All comparisons used either root-mean-squared prediction error (RMSPE) or mean absolute prediction error (MAPE) on some form of K-fold cross-validation, and only [23] evaluated the estimated prediction errors to examine whether prediction intervals contained the true values with the correct proportion. None of the papers simulated spatially autocorrelated data to evaluate and compare the different modeling approaches.

The comparisons in our review used random forests and spatial regression mostly as black box methods. By black box methods, we mean that data are used without much examination, and methods that rely on default values and little user interaction. Generally, random forests seems to outperform spatial regression, but only as a black box method. Can practitioners with more experience make each of these methods work better, and then how will they compare? It will be (or should be) rare that data, collected at great expense, are subjected to black box methods. The history of regression has taught statisticians to use best practices for their data analyses, including exploring data, making transformations, checking residuals, using model diagnostics, and then possibly refitting models. We suggest that the results given in the previously mentioned literature are not necessarily reflective of a considered approach to many data analyses. In contrast, we will confine ourselves to a single data set, and a single response variable, but we take considerable effort to make each method work as well as possible, and discuss the ramifications after the analyses. We will also use simulations to investigate properties not seen in the real data.

Our objectives are to compare random forest and spatial regression modeling approaches for predicting and mapping the MMI for all 1.1 million perennial stream reaches across the CONUS. In a related study, [14] used random forest modeling to predict the binary "good" and "poor" MMI condition classes with StreamCat predictor variables. The random forest models developed in [14] were used to map the predicted probability of good stream condition for all perennial CONUS stream reaches. In this article we instead model the MMI scores directly, and include both random forests and spatial regression. We also evaluate each method's ability to quantify the uncertainty of the MMI predictions. Previous studies have produced maps of random forest model uncertainties by interpolating the residuals [19, 24], or taking the standard deviations of the predictions made by each tree in the ensemble [25]. In this article we take a different approach, and formally construct random forest prediction

intervals using the method of quantile regression forests [26], which has been studied primarily in the context of non-spatial data. We also consider a hybrid random forest regression-kriging approach, in which a simple-kriging model is estimated for the random forest residuals, and simple-kriging predictions of residuals are added to random forest predictions. Although we focus on a particular data set, we generalize the concepts through simulations, and our overall goal is application to other large environmental data sets.

Spatial regression model

Here, we introduce the spatial regression model, likelihood-based estimation methods, and kriging prediction and variance equations that we apply to the MMI and StreamCat data. The reduced rank method, and covariate transformation and selection procedures will be discussed in subsequent sections. A thorough review of the geostatistical modeling approach discussed in this study is provided in [27] and [28].

Suppose that $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \cdots, Y(\boldsymbol{s}_n))'$ is a response vector that is spatially referenced at locations $\boldsymbol{s}_i \in D \subset \mathbb{R}^2$. The spatial regression model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z} + \epsilon, \tag{1}$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix for the covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, $\boldsymbol{z}$ is an $n \times 1$ vector of spatially autocorrelated random variables, and ϵ is an $n \times 1$ vector of independent random errors. The $n \times n$ covariance matrix for the model can be expressed as

$$\boldsymbol{\Sigma} = \mathrm{var}(\boldsymbol{Y}) = \mathrm{var}(\boldsymbol{z}) + \mathrm{var}(\epsilon) = \boldsymbol{R} + \sigma_\epsilon^2 \boldsymbol{I}. \tag{2}$$

To simplify estimation of (2), we assume a stationary covariance function that depends on Euclidean distance and takes an exponential form. That is, the (i, j) entry of $\boldsymbol{R}$ is given by

$$\mathrm{cov}(z(\boldsymbol{s}_i), z(\boldsymbol{s}_j)) = \sigma_z^2 \exp(- \parallel \boldsymbol{s}_i - \boldsymbol{s}_j \parallel / \alpha), \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean distance metric, and σ_z^2 and α are parameters to be estimated. In the geostatistical literature, the parameters σ_z^2, α, and σ_ϵ^2 are, respectively, called the partial sill, range, and nugget. The nugget parameter models residual variation in the response when the separating distance is zero.

The model in (1) is also commonly referred to as the spatial linear model (SLM), and as the universal-kriging model when used for spatial prediction, with the ordinary-kriging model being the special case when $\boldsymbol{X}$ is a $n \times 1$ column vector of 1's. While numerous types of covariance functions have been proposed for the SLM [29, pp. 80–93], we only consider the exponential form in (3) since estimation of the SLM is computationally demanding with large data sets. Moreover, for modeling MMI, we focus instead on estimating covariate transformations in $\boldsymbol{X}$. Regionally varying intercept terms are also included in $\boldsymbol{X}$ to account for differences in MMI development in the nine ecoregions.

The parameters of a spatial regression model for a particular data set can be estimated using maximum likelihood (ML) [27, p. 92] or restricted maximum likelihood (REML) [30, 31] estimation. The negative log-likelihood can be expressed as

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = 0.5\{n \log(2\pi) + \log(|\boldsymbol{\Sigma}(\boldsymbol{\theta})|) + (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + c\}, \tag{4}$$

where, for ML, $c = 0$, and for REML, $c = -p \log(2\pi) + \log|\boldsymbol{X}' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{X}|$; the covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, is written here to emphasize dependence on the unknown covariance parameters $\boldsymbol{\theta}$ (i.e.,

nugget, partial sill, and range). Minimizing (4) with respect to $\boldsymbol{\beta}$ gives

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Y}. \tag{5}$$

The ML or REML estimators of the covariance parameters, $\hat{\boldsymbol{\theta}}$, are obtained by substituting $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ into (4) and minimizing with respect to $\boldsymbol{\theta}$; estimators of the regression coefficients are consequently found by substitution of $\hat{\boldsymbol{\theta}}$ back into (5), i.e., $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$. In practice, we obtain ML or REML estimates of $\boldsymbol{\theta}$ numerically using the general purpose optimization function `optim()` provided in the statistical software package R [32]. Note that REML estimators tend to have less bias and better performance (in terms of mean squared error) than ML estimators, especially when p is large relative to n [27, 33].

Once the parameters are estimated for the spatial regression model, we use universal kriging to make predictions and construct prediction intervals [28, pp. 148–148]. The universal-kriging prediction and variance equations for the response at a new location $\boldsymbol{s}_0$ are given by

$$\hat{Y}(\boldsymbol{s}_0) = \boldsymbol{x}(\boldsymbol{s}_0)'\hat{\boldsymbol{\beta}} + \boldsymbol{c}(\boldsymbol{s}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{6}$$

$$\text{var}(\hat{Y}(\boldsymbol{s}_0)) = C(\boldsymbol{s}_0, \boldsymbol{s}_0) - \boldsymbol{c}(\boldsymbol{s}_0)'\boldsymbol{\Sigma}^{-1}\boldsymbol{c}(\boldsymbol{s}_0) + \boldsymbol{t}(\boldsymbol{s}_0)'(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{t}(\boldsymbol{s}_0), \tag{7}$$

where $\boldsymbol{t}(\boldsymbol{s}_0) = \boldsymbol{x}(\boldsymbol{s}_0) - \boldsymbol{X}'\,\Sigma^{-1}\,\boldsymbol{c}(\boldsymbol{s}_0)$, $\boldsymbol{c}(\boldsymbol{s}_0) = \text{cov}(Y(\boldsymbol{s}_0), \boldsymbol{Y}) = (C(\boldsymbol{s}_0, \boldsymbol{s}_1), \cdots, C(\boldsymbol{s}_0, \boldsymbol{s}_n))'$, and $\boldsymbol{x}(\boldsymbol{s}_0)$ is the covariate vector at $\boldsymbol{s}_0$. Note that the covariance function is defined here as $C(\boldsymbol{u}, \boldsymbol{v}) = \sigma_z^2 \exp(||\boldsymbol{u} - \boldsymbol{v}||/\alpha) + I(\boldsymbol{u} = \boldsymbol{v})\sigma_\epsilon^2$ for all locations $\boldsymbol{u}, \boldsymbol{v} \in D$. Also, note that (6) is derived as the homogeneously linear combination of the data, $\boldsymbol{\lambda}'\boldsymbol{Y}$ where $\lambda \in \mathbb{R}^n$, that minimizes the mean-squared-prediction error, $E(Y(\boldsymbol{s}_0) - \boldsymbol{\lambda}'\boldsymbol{Y})^2$, subject to the unbiasedness constraint $E(\boldsymbol{\lambda}'\boldsymbol{Y}) = E(Y(\boldsymbol{s}_0)) = \boldsymbol{x}(\boldsymbol{s}_0)'\boldsymbol{\beta}$; and (7) is the minimized mean-square-prediction error, often referred to as the kriging variance.

Reduced rank methods

For data sets with a large number of records, inverting the covariance matrix when optimizing the log-likelihood function (4) can be computationally burdensome. For example, the motivating 2008/09 NRSA data set contains nearly 2000 records; thus there is a computational cost when estimating a spatial regression model for this data set. To accelerate parameter estimation for the SLM we consider reduced rank methods [34, 35]. In this section we specify the reduced rank method used in the study; in a subsequent section we will discuss the application of this method to a computationally efficient covariate selection routine.

Consider a set of r knot locations $\{\boldsymbol{k}_i: i = 1, \cdots, r\}$, distributed over the same domain as the observed data, such that $r \ll n$. Instead of modeling the covariance matrix for $\boldsymbol{Y}$ in terms of the Euclidean distances between the observed locations, we can alternatively model the covariance matrix in terms of the knot locations as

$$\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{K}^{-1}\boldsymbol{S}' + \sigma_\epsilon^2\boldsymbol{I}, \tag{8}$$

where $\boldsymbol{S}$ is an $n \times r$ matrix with (i, j) element $\sigma_z^2 \exp(- \parallel \boldsymbol{s}_i - \boldsymbol{k}_j \parallel /\alpha)$; $\boldsymbol{K}$ is an $r \times r$ matrix with (i, j) element $\sigma_z^2 \exp(- \parallel \boldsymbol{k}_i - \boldsymbol{k}_j \parallel /\alpha)$; and σ_z^2, α, and σ_ϵ^2 are parameters to be estimated. An advantage of the specification in (8) is that application of the well-known Sherman-Morrison-Woodbury formula (see [36] for a review) yields the following decomposition:

$$\boldsymbol{\Sigma}^{-1} = \sigma_\epsilon^{-2}[\boldsymbol{I} - \boldsymbol{S}(\sigma_\epsilon^2\boldsymbol{K} + \boldsymbol{S}\boldsymbol{S}')^{-1}\boldsymbol{S}']. \tag{9}$$

Since (9) only involves inverting an $r \times r$ matrix computation speed is greatly improved.

Note that (8) can be viewed as the covariance matrix for $\boldsymbol{Y}$ in the spatial mixed effects model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{S\gamma} + \epsilon$, where $\boldsymbol{\gamma}$ is an $r \times 1$ vector of random effects such that $\text{var}(\boldsymbol{\gamma}) = \boldsymbol{K}^{-1}$. Also, one property of the covariance matrix specified in (8) is that if the knots are the observed data locations $\{\boldsymbol{s}_i\}$, then $\boldsymbol{S} = \boldsymbol{K} = \boldsymbol{R}$, and consequently (8) is equivalent to the full rank covariance matrix defined in (2).

Covariate transformations

Many of the covariates in the StreamCat data set have nonlinear relationships with MMI. For example, Fig 2(a) and 2(c) shows highly skewed relationships between MMI and two StreamCat covariates: watershed area in square km (WsAreaSqKm); and the percent of watershed area classified as developed, medium intensity land use within a 100-m buffer of a stream reach (PctUrbMd2006WsRp100). The log-transformation helps linearize the relationship between MMI and PctUrbMd2006WsRp100 (Fig 2d), and also reveals a quadratic relationship between MMI and WsAreaSqKm (Fig 2b). While not shown here, many of the other StreamCat covariates exhibit similar types of nonlinearities, and further motivate considering covariate transformations.

To linearize relationships between the covariates and response variable, we estimate Box-Cox transformations [37] for the covariates. Specifically, we estimate transformations of the form

$$g(x;\lambda_1,\lambda_2) = \begin{cases} \dfrac{(x+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(x+\lambda_2) & \lambda_1 = 0, \end{cases} \tag{10}$$

where $x > -\lambda_2$. Note that Box-Cox transformations were first proposed as a way to transform the response variable [37]; however, these types of power transformations have also been applied to the independent variables in regression modeling [38, pp. 50–63].

Fig 2 suggests that StreamCat covariates can be zero-inflated (e.g., PctUrbMd2006WsRp100), or have quadratic relationships with the response (e.g., log-transformed WsAreaSqKm). Different types of transformation effects are considered depending on whether or not the StreamCat covariate is zero-inflated. To estimate transformations for the zero-inflated covariates, we estimate the following linear models for varying values of λ_1 and λ_2:

$$y = \beta_0 + \beta_1 I(x_i \neq 0) + \epsilon, \tag{11}$$

$$y = \beta_0 + \beta_1 g(x_i;\lambda_1,\lambda_2) I(x_i \neq 0) + \epsilon, \tag{12}$$

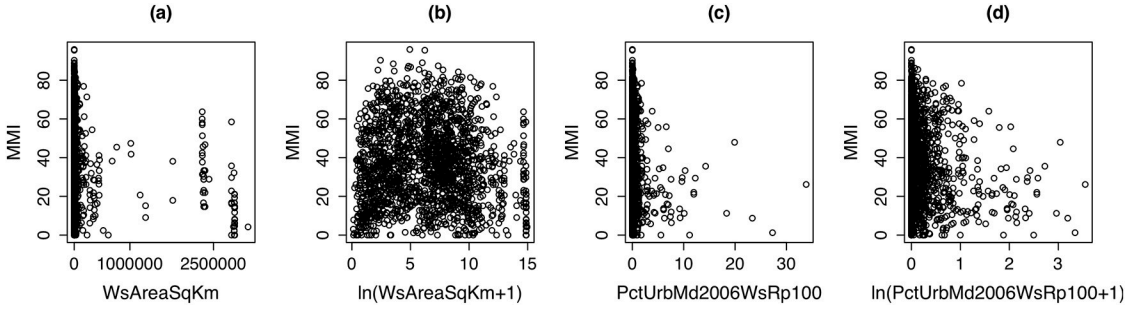

Fig 2. Scatter plots of MMI versus StreamCat covariates. (a) watershed area in square kilometers (WsAreaSqKm); (b) natural logarithm of WsAreaSqKm; (c) percent of watershed area classified as developed, medium intensity land use in 2006 within a 100 meter buffer of a stream reach (PctUrbMd2006WsRp100); (d) natural logarithm of PctUrbMd2006WsRp100.

https://doi.org/10.1371/journal.pone.0229509.g002

$$y = \beta_0 + \beta_1 I(x_i \neq 0) + \beta_2 g(x_i; \lambda_1, \lambda_2) I(x_i \neq 0) + \epsilon. \quad (13)$$

Here y is MMI, x_i is the i^{th} StreamCat covariate, ϵ is the error term, and $I(a)$ is the indicator function, equal to one if the argument a is true, and zero otherwise. For each zero-inflated covariate, we use the Aikaike Information Criterion (AIC) [39] to select optimal (λ_1, λ_2) and the type of transformation: zero/nonzero indicator (11), interaction between the indicator and transformed covariate (12), or both (13). That is, out of several candidate values for (λ_1, λ_2), we select the Box-Cox parameter values and type of transformation effect that corresponds to the linear model (11, 12, or 13) with the lowest AIC. To estimate transformations for the other covariates (not zero-inflated), we estimate the following linear models for varying values of λ_1 and λ_2:

$$y = \beta_0 + \beta_1 g(x_i; \lambda_1, \lambda_2) + \epsilon, \quad (14)$$

$$y = \beta_0 + \beta_1 g(x_i; \lambda_1, \lambda_2) + \beta_2 (g(x_i; \lambda_1, \lambda_2))^2 + \epsilon. \quad (15)$$

For each covariate that is not zero-inflated, we again use the AIC to select optimal (λ_1, λ_2) and the type of transformation: linear (14) or quadratic (15) polynomial. In practice, we vary the values of the exponent parameter λ_1 between 0 and 3, and try several shifting parameters λ_2 to ensure that (10) is well defined. We also define a covariate x_i as being zero-inflated if the proportions of zeros is greater than 2%. Note that the transformations are estimated separately for each covariate, and that spatial autocorrelation is ignored while estimating transformations because otherwise the procedure would be too slow computationally.

Once transformations have been selected, a new design matrix containing the transformations for each StreamCat covariate can be used to fit either a multiple regression or spatial regression model. Since we include indicator variables for zero-inflated covariates and quadratic polynomial effects, the transformed design matrix is larger than the design matrix without transformations. In the next section, we discuss a covariate selection approach for reducing the number of parameters in a spatial regression model.

Covariate selection

Covariate selection for the spatial regression model is implemented in two phases. For the first phase, we fit a multiple linear regression model (LM) with the full set of covariates from the StreamCat data set. Dummy variables for the ecoregions (Fig 1) are also included as additional covariates in the LM to account for regional variations in MMI development. Variables are then selected using a backwards stepwise algorithm (i.e., the `step()` function from [32]). Note, we start by selecting variables for an LM rather than an SLM since the LM can be rapidly estimated, and software is readily available for variable selection.

For the second phase, we estimate an SLM with the variables selected for the LM. We then conduct further variable selection for the SLM since some covariates may no longer be significant once spatial autocorrelation is taken into account [8, 40]. Specifically, we repeatedly remove the covariate with the largest absolute t-statistic and re-fit the SLM until the model's AIC score starts to increase. To speed up ML estimation of the SLM during this procedure we use the reduced rank method. Once all variables are selected for the SLM, REML with the full rank covariance matrix is used to estimate the final model. A step-by-step description of the covariate selection procedure is provided in the Supplement (S1 File).

Comparison with LASSO. As a comparison, we also fit an MMI model using the LASSO, a modern variable selection technique, which potentially performs better than stepwise

methods when the number of covariates is large [41, pp. 68–69]. The parameters for the LASSO model are estimated by minimizing the residual sum of squares, $(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})$, subject to the constraint $\sum_{j=1}^{p} |\beta_j| \leq s$, where s is a tuning parameter. The constraint has the effect of shrinking the coefficients towards zero, and setting some coefficients exactly equal to zero (thereby performing variable selection).

We estimate a LASSO model for MMI using the R package `glmnet` [42]. The software selects the value of the tuning parameter, s, that minimizes an internal cross-validation error. Note that incorporation of spatial autocorrelation parameters into a LASSO regression model is not supported by the `glmnet` package, and is therefore not considered in this work. Our motivation for including a LASSO model is to compare its performance with the stepwise methods used to select covariates for the LM and SLM.

Random forest model

Random forest (RF) modeling has become a popular technique for regression and classification with complex environmental data sets [25, 43–47]. In contrast to multiple regression, RF is an algorithmic procedure that makes no *a priori* assumptions about the relationship between the predictor variables and the response. RF has a reputation for good predictive performance when the data contain a large number predictor variables, and when there are complex nonlinearities and interaction effects in the relationship between the predictors and response variable [44, 48, 49]. In addition, RF provides several measures of variable importance that allow for interpretation of the fitted model [41, p. 593].

An RF model can be defined as a collection of regression trees $\{T_b: b = 1, \cdots, B\}$ each built from a bootstrap sample of the data set $\{\boldsymbol{Y}, \boldsymbol{X}\}$. When growing each tree T_b, at each parent node a subset of m of the p predictor variables are randomly selected, and the best split-point is found among those m variables to form two daughter nodes. The trees in the RF ensemble are grown deep with no pruning. Bagging trees [50] are the special case when $m = p$ (i.e., all predictor variables are used as candidates for splitting at each node). An RF prediction at a new site with predictor values $\boldsymbol{x} = (x_1, \cdots, x_p)$ is found by averaging the predictions made by each tree in the ensemble:

$$\hat{f}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\boldsymbol{x}).$$

RF is also commonly used for classification, in which case $T_b(\boldsymbol{x})$ takes on discrete values (e.g., 0/1 for binary classification) and the RF prediction can be defined as the majority vote from the collection of class predictions $\{T_b(\boldsymbol{x})\}$. However, in this study we only consider RF for regression. Note that the RF algorithm produces individual tree estimators with high variance. That is, a regression tree fit to different portions of the data can yield very different predictions at an unsampled site. The main idea behind RF is that averaging over many tree models is a way to reduce variance, and thereby improve predictive performance relative to a single tree model [7, 50]. Moreover, only considering a random subset of $m < p$ predictors at each node has the effect of decorrelating the trees in the ensemble, which can further improve the performance of the RF model relative to bagging. See [41] for a comprehensive review of RF and relevant theory.

We implement RF using the R package `randomForest` [51]. The main tuning parameters when estimating an RF model with this package: the number of trees, `ntree`; the number of predictor variables randomly selected at each node, `mtry`; and the minimum number of cases in a tree's terminal node, `nodesize`. However, RF is generally insensitive to choice of

tuning parameters, and the defaults provided by the `randomForest` package perform adequately for most data sets [25, 44, 51]. In practice, we use `ntree` = 1000, and the defaults `mtry` = $p/3$ and `nodesize` = 5. We use more trees than the default value (500) since this is recommended for data sets that have a large number of predictors [51]. For the RF model of MMI, we also use the full set of StreamCat predictor variables, as well as an additional categorical predictor for the ecoregions. We do not perform additional subset selection since the RF algorithm is robust to handling large sets of predictor variables [7, 48, 49].

Quantile prediction intervals

Prediction intervals for RF can be computed using quantile regression forests (QRF) [26]. While RF provides information on the conditional mean of the response, QRF instead provides information on the conditional distribution function of the response. Approximation of the conditional distribution function is useful for making quantile predictions and forming associated prediction intervals. For instance, a 90% prediction interval for the response at a new site with predictors $\boldsymbol{x}$ can be formed as the 0.05 and 0.95 quantile predictions denoted by $[\hat{Q}_{0.05}(\boldsymbol{x}), \hat{Q}_{0.95}(\boldsymbol{x})]$. Here $Q_\alpha(\boldsymbol{x})$ defines the α-quantile, that is, the value of the random response variable Y such that the probability Y is less than $Q_\alpha(\boldsymbol{x})$, for given $\boldsymbol{x}$, is exactly equal to α; $\hat{Q}_\alpha(\boldsymbol{x})$ denotes the QRF estimator of this quantity. A main distinction between the RF and QRF algorithms is that RF only keeps track of the mean value of the response data at each leaf (terminal node) of each tree. QRF, on the other hand, keeps track of all the response data at each leaf of each tree and approximates the full conditional distribution with this additional information. In practice, we implement the QRF method using the R package `quantregForest` [52], which builds on the `randomForest` package also used in this study.

Random forest regression kriging

Random forest regression kriging (RFRK) has been proposed in a number of studies as a way to account for spatial autocorrelation in RF modeling [18, 20, 21]. For RFRK, a prediction at a new site is given by summing the RF prediction and the kriging prediction of the RF residual. Formally, an RFRK prediction of $Y(\boldsymbol{s}_0)$, at a new site $\boldsymbol{s}_0$, is given by

$$\hat{Y}(\boldsymbol{s}_0) = \hat{f}_{RF}(\boldsymbol{x}(\boldsymbol{s}_0)) + \hat{e}(\boldsymbol{s}_0)$$

where $\hat{f}_{RF}$ is the RF prediction with covariates $\boldsymbol{x}(\boldsymbol{s}_0)$, and $\hat{e}(\boldsymbol{s}_0)$ is the kriging prediction of the residual. In practice, we use ML to estimate a simple-kriging model for the RF residuals that assumes a zero mean (i.e, $E(e(\boldsymbol{s})) = 0$) and exponential model for the covariance matrix. Additionally, we use the simple-kriging variances for the residuals to construct prediction intervals. Computational details are provided in the Supplement (S1 File).

Performance measures

We assess the performance of different models for MMI using 10-fold cross-validation. Nine models of MMI are considered for the comparison: (1) an ordinary-kriging model, i.e., an SLM with no covariates and a single intercept term ($\boldsymbol{X}$ is a $n \times 1$ column vector of 1's); (2) an LM with no transformations; (3) an SLM with no transformations; (4) an LM with transformations; (5) an SLM with transformations; (6) a LASSO model; (7) a LASSO model with transformations; (8) an RF model; and (9) an RFRK model.

The root-mean-square prediction error (RMSPE) and coverage of the prediction intervals are used to evaluate the cross-validation performance of the different models. Let Y_i denote the i^{th} observed value and $\hat{Y}_i$ the 10-fold cross-validation prediction. The RMSPE is computed as

the square root of the mean of $(Y_i - \hat{Y}_i)^2$ and coverage of the 90% prediction intervals is computed as the mean of $I(\hat{Y}_i - 1.645\text{se}(\hat{Y}_i) < Y_i < \hat{Y}_i + 1.645\text{se}(\hat{Y}_i))$ for all i in $1, \cdots, n$, where I is the indicator function and $\text{se}(\hat{Y}_i)$ is the prediction standard error. Since we estimate quantile prediction intervals for RF, the coverage of the 90% prediction intervals is computed as the mean of $I(\hat{Q}_{0.05}(\boldsymbol{x}_i) < Y_i < \hat{Q}_{0.95}(\boldsymbol{x}_i))$ for all i in $1, \cdots, n$, where $\boldsymbol{x}_i$ are the predictor variables at the i^{th} site and $\hat{Q}_\alpha(\boldsymbol{x}_i)$ is the 10-fold cross-validation prediction of the α-quantile.

Note that the covariate selection and transformation procedures are embedded in the 10-fold cross-validation of the LM and SLM. Specifically, at each iteration, the data is split into 10 equally sized folds, and one fold is held out as a test set. The remaining 9 folds are used for training, i.e., to select covariates using the stepwise procedure, estimate Box-Cox transformations, and estimate the parameters of the LM or SLM. Therefore, a different set of covariates and transformations are selected using the training data at each iteration of the cross-validation. This ensures that the data used to validate the models are completely independent from the data used to select covariates and estimate transformations. See [41, pp. 241–249] for a comprehensive review of K-fold cross-validation.

Results

The 10-fold cross-validation performance measures for modeling MMI are presented in Table 1. In terms of RMSPE, RFRK and RF resulted in the best performance, respectively followed by the SLM, LASSO, and LM with transformations. The SLM, LASSO, and LM without transformations did not perform as well; and not surprisingly, the ordinary-kriging model had the highest RMSPE. The results show that covariate transformations for the LM, LASSO, and SLM were necessary to obtain performance comparable with RF. Accounting for spatial autocorrelation also improved the performance of the SLM relative to the LM for both the transformed and untransformed cases. Moreover, additional covariate selection for the SLM resulted in a more parsimonious model than the LM. RFRK also performed better than RF, although the difference was not substantial (the Pearson correlation between RF and RFRK cross-validation predictions was greater than 0.98).

Table 1. Cross-validation performance results.

Model	k	RMSPE	PIC90	PIC95
OK	4	18.55	0.900	0.962
LM	49	18.20	0.882	0.946
SLM	37	17.65	0.886	0.948
LM-TF	67	17.41	0.883	0.944
SLM-TF	51	16.82	0.882	0.937
LASSO	72	18.04		
LASSO-TF	134	16.98		
RF		16.52	0.914	0.955
RFRK		16.41	0.905	0.960

k is the number of parameters from the model fit using the entire data set (i.e., all $n = 1859$ observations); PIC90 and PIC95 are the coverages of the 90% and 95% prediction intervals; OK is the ordinary-kriging model; LM-TF, SLM-TF, and LASSO-TF are the LM, SLM, and LASSO with covariate transformations, respectively; and other abbreviations are defined in the text. Computation of prediction intervals for LASSO is not supported by the `glmnet` package.

https://doi.org/10.1371/journal.pone.0229509.t001

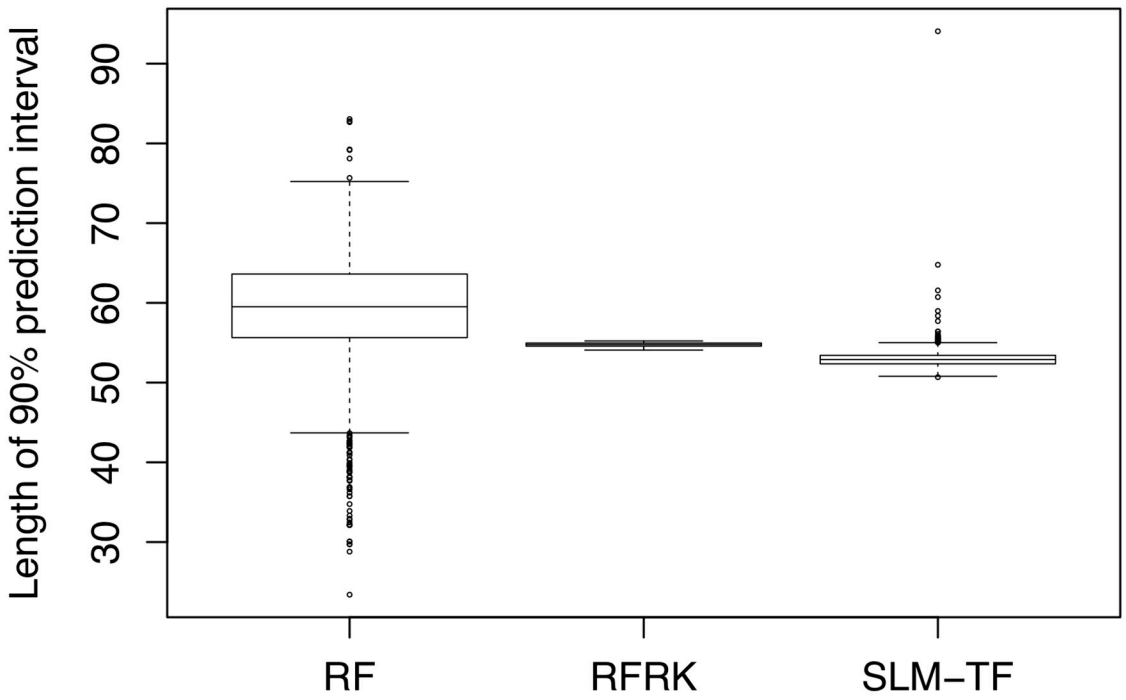

Fig 3. Boxplots of the lengths of 90% prediction intervals. Prediction intervals for RF were computed using the quantile regression forest method, while prediction intervals for RFRK and the SLM with transformations (SLM-TF) were computed using the kriging variances.

https://doi.org/10.1371/journal.pone.0229509.g003

The coverages of the 90% and 95% prediction intervals were close to nominal for all models in Table 1. That is, for each method, the prediction intervals computed during cross-validation contained the true observed MMI values with approximately the correct proportion (within ±2%). However, even though the coverages were similar, there were considerable differences between the lengths of the prediction intervals from the different methods (Fig 3). The median length of the RF quantile prediction intervals was larger than the SLM and RFRK predictions intervals. This is reasonable since the coverages of the predictions intervals were slightly over the nominal level for RF, and slightly under the nominal level for the SLM (Table 1). Additionally, Fig 3 shows much greater variability in the lengths of the RF quantile prediction intervals than the SLM and RFRK prediction intervals. The lengths of the prediction intervals for the SLM also have a positive skew and outliers; this can be explained by the universal-kriging variances getting larger for sites that fall away from than bulk of the data in the covariate space [53]. The distribution of the prediction interval lengths for RFRK were narrower and more symmetric, in comparison, since the simple-kriging variances only account for the uncertainty due to relative geographic distances between points, and not possible quantitative extrapolation in the covariates.

Scatter plots of predicted versus observed values are presented in Fig 4. The scatter plots reveal that the predictions from the different models tend towards the mean of the observed MMIs (36.9). This effect was most pronounced for the ordinary-kriging, RF, and RFRK models; in comparison, the LM and SLM had wider distributions of predicted values. While most of the predictions from the LM and SLM (with and without transformations) were within the defined range of the MMI (0–100), a small percentage (<0.5%) of predictions were negative, and set to zero. The predictions from the RF and RFRK models, on the other hand, were contained within the observed MMI range. Note that, by definition, RF models cannot predict

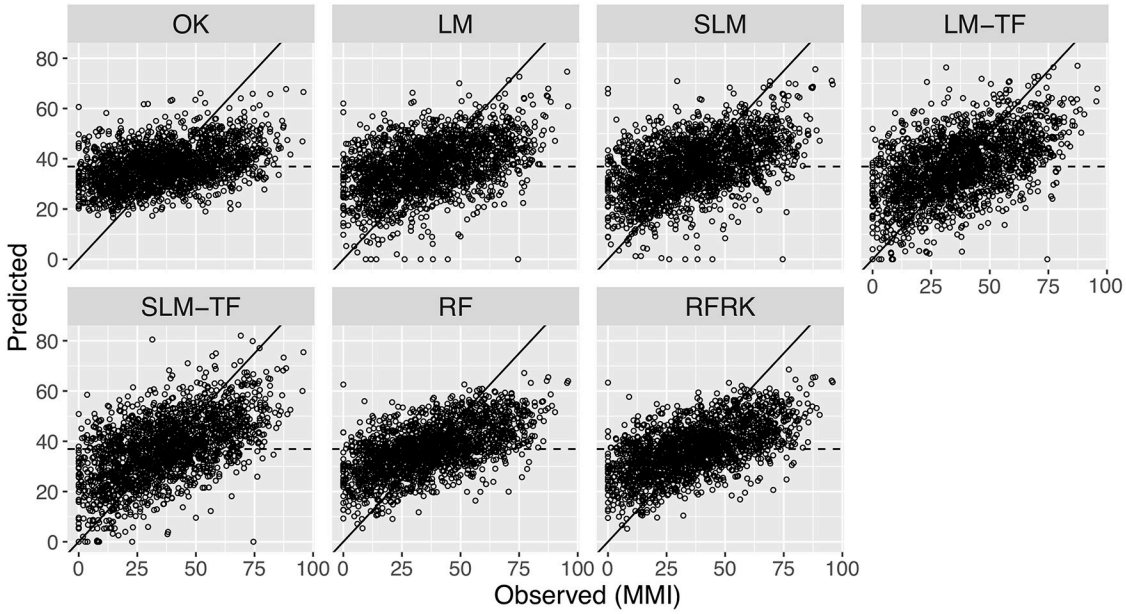

Fig 4. Scatter plots of predicted versus observed MMI values. Predictions are from 10-fold cross-validation. The 1-1 line (solid) and mean observed MMI value (36.9; dashed horizontal line) are also shown in each panel. Labels for the seven models are defined in Table 1 and the text.

https://doi.org/10.1371/journal.pone.0229509.g004

outside the range of the observed data since each tree model within the forest makes predictions by taking the mean of the response data falling within a given leaf (terminal) node.

Covariance parameter estimates (nugget, partial sill, and range) for the spatial regression models are presented in Table 2. The ordinary kriging model has a larger estimated nugget parameter, $\hat{\sigma}^2_\epsilon$, and smaller nugget-to-sill ratio, $\hat{\sigma}^2_\epsilon/(\hat{\sigma}^2_z + \hat{\sigma}^2_\epsilon)$, than the SLM. This is expected since the covariates in the SLM explain additional variation in MMI not accounted for by ordinary kriging. Moreover, the spatially-referenced StreamCat covariates in the SLM account for some spatial autocorrelation in MMI. To further assist interpretation, Table 2 also presents the effective range, $-\hat{\alpha}\log[0.01 * (\hat{\sigma}^2_z + \hat{\sigma}^2_\epsilon)/\hat{\sigma}^2_z]$, which is defined here as the distance beyond which spatial autocorrelation is less than 0.01 (i.e., the distance h found by solving $\rho(h) = C(h)/C(0) = 0.01$; [54]). The effective ranges for ordinary-kriging and the SLM reveal that spatial autocorrelation in the data is close to zero for distances beyond 480–580km. Additionally, for the RFRK model, both the effective range (160km) and nugget-to-sill ratio (0.951) indicate little, perhaps negligible, spatial autocorrelation in the RF residuals. Note that, for the effective

Table 2. Estimated covariance parameters.

	OK	SLM	SLM-TF	RFRK
Nugget	278.08	257.17	226.78	261.08
Partial Sill	135.05	68.59	53.03	13.52
Range	139.09	189.31	167.98	100.66
Effective Range	485.03	576.87	494.19	160.44
Nugget-to-Sill Ratio	0.67	0.79	0.81	0.95

The effective range is the distance (km) beyond which spatial autocorrelation is less than 0.01, and the nugget-to-sill ratio is given by $\hat{\sigma}^2_\epsilon/(\hat{\sigma}^2_z + \hat{\sigma}^2_\epsilon)$.

https://doi.org/10.1371/journal.pone.0229509.t002

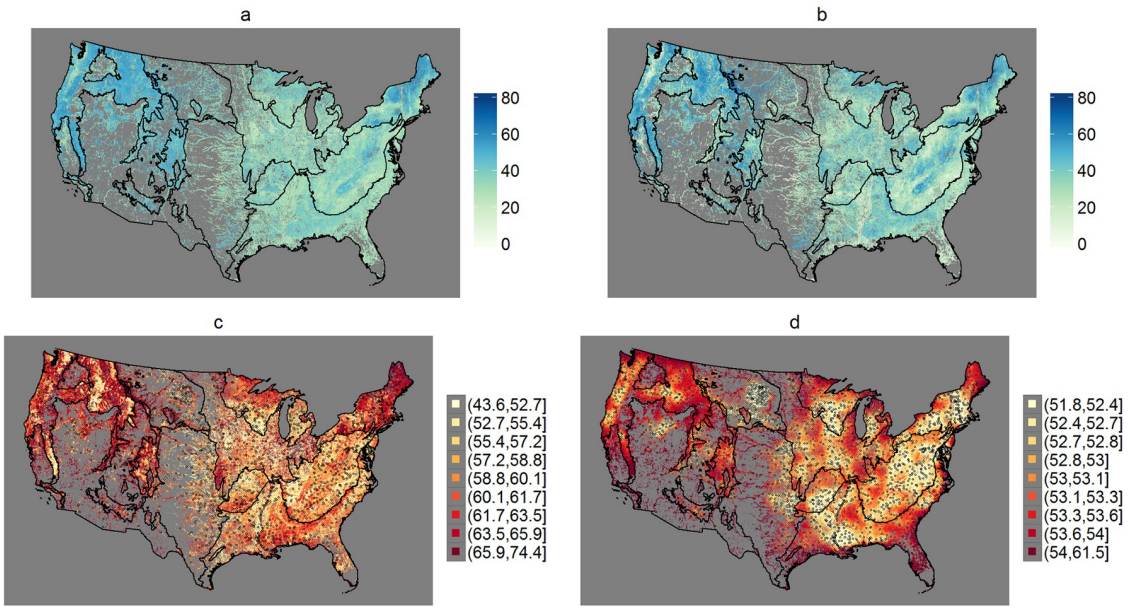

Fig 5. Maps of MMI predictions and prediction errors. The maps show the MMI predictions and lengths of 90% prediction intervals for RF (a,c) and SLM with transformations (b,d). The prediction sites are the 1.1 million perennial stream reaches (catchments) in the NRSA sampling frame. Note the different scales in the maps of prediction interval lengths (c,d).

https://doi.org/10.1371/journal.pone.0229509.g005

range calculations we use 0.01 instead of 0.05, which is more common, since the nugget-to-sill ratio for RFRK is greater that 0.95.

Maps of the RF and SLM predictions of MMI are presented in Fig 5(a) and 5(b). The maps show MMI predictions for 1.1 million perennial stream reaches within the CONUS. Again, predictions were made only for perennial stream reaches since the sampling frame for the 2008/09 NRSA is limited to these types of streams. Overall, the maps show similar spatial patterns in the MMI predictions from the two models. As expected, regions dominated by urban or agricultural land use tend to have lower MMI predictions than more remote regions. The most noticeable difference between the prediction maps is that the SLM shows a wider distribution of predicted values than RF, with sharper differences between regions with low and high MMI predictions (e.g., Willamette Valley versus Cascades in Oregon; Piedmont versus Blue Ridge Mountains in Georgia, S. and N. Carolina, and Virginia). Also note that the RF predictions never reach zero, while about 1.2% of the SLM predictions are negative and set to zero in Fig 5 since MMI is defined between 0–100.

In contrast to the prediction maps, the maps of the RF and SLM prediction errors (lengths of 90% prediction intervals) are strikingly different (Fig 5c and 5d). There is much greater variability in the lengths of prediction intervals in the map for the RF model than the SLM (note, this is consistent with the cross-validation results in Fig 3). Moreover, the SLM prediction intervals show greater precision in regions that are more densely sampled (e.g., Mississippi Basin). In contrast, the precision of the RF quantile prediction intervals do not appear to scale with the density of sampling locations.

Both the SLM and RF models provide measures of variable importance. For the SLM, covariates can be ranked in terms of the absolute t-statistics for the coefficients. For RF, covariates can be ranked using a permutation-based measure (i.e., the average increase in mean-square error when each covariate is permuted in the out-of-bag data; [41, p. 593]). The highest ranked variables for modeling MMI were similar for the RF model and SLM with

transformations. Specifically, the covariates for watershed area, average topographic wetness index, and ecoregions were ranked in the top five for both models. An SLM regression coefficient summary and RF variable importance plot are provided in the Supplement (S1 File; Table S1 and Fig S2).

Last, some diagnostics for the SLM are provided in the Supplement (S1 File). A map of the SLM squared residuals (Fig S3) reveals little geographic pattern, which suggests that the assumption variance stationarity is reasonable. Directional semivariograms of the SLM residuals (Fig S4) look similar, with a little autocorrelation at short ranges, and do not reveal any obvious directional dependence.

Simulation

The motivating MMI data set for comparing spatial regression to RF had a high amount of nonlinearity in the covariates, and apparently little autocorrelation in the residuals. The results from modeling this data indicated a slight advantage to the RF approach in terms of predictive performance. However, the performance of the SLM was improved, and made competitive to RF, by considering linearizing transformations of the covariates. Since this is but a single applied data set, in this section we use simulations to generalize the comparison to spatial data sets with different characteristics. Specifically, we use simulations to explore the effect of nonlinear relationships between response and covariates, R^2, and varying amounts of autocorrelation. The goal of this simulation study is to illustrate the relative strengths of each modeling approach when generating spatial data with specified characteristics.

Data for this simulation study are generated from the following model:

$$\begin{aligned} y(\boldsymbol{s}) &= f(x_1, \cdots, x_4) + \delta(\boldsymbol{s}) = c[ag(x_1, x_2) + h(x_3, x_4)] + \delta(\boldsymbol{s}) \\ &= c[a \sin(5\pi x_1 x_2) + 2x_3 - x_4] + \delta(\boldsymbol{s}). \end{aligned} \tag{16}$$

Here $\delta(\boldsymbol{s}) = z(\boldsymbol{s}) + \epsilon$ is a spatially autocorrelated error term such that $\text{cov}(z(\boldsymbol{s}), z(\boldsymbol{s} + \boldsymbol{h})) = \sigma_z^2 \exp(- \parallel \boldsymbol{h} \parallel / \alpha)$ and $\text{var}(\epsilon) = \sigma_\epsilon^2$ is the nugget effect. The parameter c governs the proportion of variance in y explained by the covariates in the systematic component of the model f. The parameter a governs amount of nonlinearity in f, which is decomposed into a nonlinear term g and linear term h. Note that the sine function, with multiplicative interaction between x_1 and x_2, in (16) was chosen since it is difficult to recover with a linear model, and so RF is expected to have advantages if the data are generated from this type of nonlinear function.

The following characteristics of the simulated data are varied by adjusting the values of the parameters in (16):

- The amount of spatial autocorrelation in the error term $\delta(\boldsymbol{s})$. We set $\sigma_z^2 = 1$ and $\sigma_\epsilon^2 = 9$ for a low amount of autocorrelation, and $\sigma_z^2 = 9$ and $\sigma_\epsilon^2 = 1$ for a high amount of autocorrelation. The range parameter is always $\alpha = 0.5$.
- The empirical R^2, i.e., the proportion of variation in y explained by f. The value of parameter c is adjusted in each simulation run to give an empirical R^2 which is either high (0.9) or low (0.1).
- Whether the linear or nonlinear term dominates. The value of parameter a is adjusted in each simulation run so that the proportion of variance in f explained by the nonlinear term g is either high (0.9) or low (0.1).

This gives a total of $2^3 = 8$ cases since there are 2 levels (high/low) for each characteristic (spatial autocorrelation, empirical R^2, and amount of nonlinearity). The 8 cases are summarized in Table 3.

Table 3. Simulation results.

		R^2	σ^2_ϵ	σ^2_z	a	c	RMSPE			
							LM	SLM	RF	RFRK
1	NL	0.1	9	1	2.94	0.51	3.28 (0.028)	**3.22** (0.020)	3.32 (0.027)	3.26 (0.022)
2	NL	0.1	1	9	2.94	0.43	2.77 (0.145)	**1.57** (0.029)	2.80 (0.146)	1.64 (0.033)
3	NL	0.9	9	1	2.94	4.56	9.03 (0.074)	9.02 (0.073)	**7.45** (0.067)	7.45 (0.065)
4	NL	0.9	1	9	2.94	3.87	7.65 (0.394)	7.40 (0.369)	6.29 (0.322)	**5.96** (0.290)
5	L	0.1	9	1	0.33	1.52	3.16 (0.025)	**3.09** (0.018)	3.24 (0.026)	3.18 (0.021)
6	L	0.1	1	9	0.33	1.29	2.66 (0.140)	**1.34** (0.010)	2.74 (0.143)	1.53 (0.024)
7	L	0.9	9	1	0.33	13.69	4.23 (0.038)	**4.19** (0.033)	4.54 (0.036)	4.51 (0.032)
8	L	0.9	1	9	0.33	11.61	3.58 (0.186)	**2.82** (0.116)	3.85 (0.201)	3.16 (0.138)

The first column indicates the case number and the second column indicates whether the linear (L) or nonlinear (NL) structural component of the model in (16) dominates. Values for the RMSPE, and parameters *a* and *c*, were averaged over 20 independent simulation runs. The standard errors of the RMSPE scores are shown in parenthesis. Note that the standard errors were computed as $SD_{RMSPE}/\sqrt{20}$, where SD_{RMSPE} is the standard deviation of the RMSPE scores over the 20 runs.

https://doi.org/10.1371/journal.pone.0229509.t003

For each simulation case, we generated 20 data sets $[\boldsymbol{y};\boldsymbol{x}_1, \cdots, \boldsymbol{x}_4]$ of 1500 points with locations randomly generated over the unit square. For each data set, 500 points were used for training, and the other 1000 as a test set. Values for the covariates $\boldsymbol{x}_i$ were drawn from Unif [0, 1]. Data from the spatially autocorrelated error term $\delta(\boldsymbol{s})$ in (16) were generated using the Cholesky decomposition method [27, p. 201]. Values of parameters *c* and *a* were selected to fix the empirical R^2 and amount of nonlinearity in the simulated data sets generated for each case. Since values of *c* and *a* varied, Table 3 presents the averaged values.

The LM, SLM, RF, and RFRK models are compared in the simulations. The SLM is fit using REML with the full rank covariance matrix, and no covariate transformations are considered. Model performance measures (RMSPE and prediction interval coverage) are averaged over the 20 simulated data sets generated for each case. The simulation code is provided in an R package available at https://github.com/ericwfox/slmrf.

Simulation results

Simulation results for the RMSPE are presented in Table 3. When there was a high amount of autocorrelation in the error term and $R^2 = 0.1$ (case 2,6), SLM and RFRK performed substantially better than LM and RF. When $R^2 = 0.9$ and the nonlinear component dominated (case 3,4), RF and RFRK performed substantially better than LM and SLM; RFRK also performed better than RF when there was a high amount of autocorrelation in the error term (case 4). When $R^2 = 0.9$ and the linear component dominated (case 7,8), the SLM had the best performance among all methods; RFRK also performed better than LM when there was a high amount of autocorrelation in the error term (case 8). When the nugget effect dominated (case 1,5), all models performed similarly in terms of RMSPE, and the SLM had slightly better performance than other methods. For all cases, SLM performed better than LM, and RFRK performed at least as well as RF. However, this is reasonable since there was some amount of autocorrelation in the data generated for each case. Moreover, when there was only a small amount of autocorrelation in the error term and $R^2 = 0.9$ (case 3,7), the spatial models performed approximately as well as the non-spatial models (i.e., RF performed as well as RFRK in case 3, and LM performed approximately as well as SLM in case 7).

The coverages of the 90% prediction intervals for LM, SLM, and RFRK were close to nominal for all simulation cases (Table 4). The quantile prediction intervals for RF showed over-

Table 4. Simulation results for coverage of 90% prediction intervals.

	LM	SLM	RF	RFRK
1	0.897	0.893	0.869	0.892
2	0.901	0.894	0.874	0.897
3	0.916	0.914	0.914	0.897
4	0.914	0.917	0.916	0.894
5	0.897	0.892	0.866	0.892
6	0.903	0.894	0.869	0.894
7	0.900	0.899	0.959	0.899
8	0.899	0.907	0.955	0.905

The first column indicates the different cases, which are summarized in Table 3. Values were averaged over 20 independent simulation runs.

https://doi.org/10.1371/journal.pone.0229509.t004

coverage when the linear term dominated and $R^2 = 0.9$ (case 7,8), but were otherwise reasonable, although not as precise as the other methods.

Discussion

In this article we compared spatial regression and RF methods for modeling stream condition (MMI) with over 200 potential covariates. We used the models for prediction and uncertainty quantification of MMI at 1.1 million perennial stream reaches across the CONUS. Initial exploratory analysis revealed highly nonlinear relationships between StreamCat covariates and MMI scores, which motivated the application of Box-Cox transformations for the covariates in the spatial regression model. To summarize the modeling results: First, the SLM with transformations and RF model performed comparably well in terms of cross-validation RSMPE, with RF having a slight advantage (0.3 difference in RMSPE; Table 1). Second, the SLM performed better than the multiple linear regression and LASSO models, which did not account for spatial autocorrelation and used more covariates. Third, the maps of the SLM and RF predictions showed similar spatial trends in stream condition, although the RF predictions were smoother and had greater tendency to concentrate around the mean. Fourth, many of the top predictors identified by the t-statistics for the coefficients in the SLM and the variable importance measures for RF were similar.

A novel contribution of this study was the assessment and comparison of prediction intervals for the spatial regression and RF methods. The construction of prediction intervals is not yet common practice in RF modeling. In contrast to geostatistics, there is no consensus in the RF literature on best practices for uncertainty quantification. We investigated two ways to construct prediction intervals for RF models: first, by using the quantile regression forest method [26]; and second, by fitting the RFRK model and using the simple-kriging variances for the RF residuals to form intervals. We found that coverages of the prediction intervals for the SLM, RF, and RFRK models of stream condition were close to nominal (Table 1). However, the lengths of the RF prediction intervals, computed using quantile regression forests, had much greater variability than the SLM and RFRK prediction intervals. One explanation for these differences is that the kriging variances are optimized by minimizing the mean-square-prediction error, whereas the RF quantile prediction intervals are not found by directly minimizing a loss function. The large amount of variability in the prediction interval lengths for quantile regression forests was also acknowledged in [26] in applications to a variety of data sets.

The results of this study indicate that there are several trade-offs when deciding between an RF or spatial regression approach to modeling a large environmental data set. We summarize these below:

- RF performed slightly better than the SLM with covariate transformations for modeling stream condition on a national scale. One explanation is that tree-based methods account for a wider range of nonlinearities than Box-Cox transformations. There was also a low amount of spatial autocorrelation in the MMI response data since the sampled stream sites were far apart. To quantify this, the average distance between each sampling location and its closest, neighboring sampling location was 30.07 km.
- The SLM stands out as a better descriptive model for ecological processes than RF. The initial exploratory analysis and transformation procedure provided insights into the functional relationships between the covariates and MMI response variable. In contrast, the ways in which RF deals with nonlinearities and interactions are hidden in the ensemble of trees and difficult to interpret.
- RF has a computational edge over the SLM. The RF algorithm is easy to implement using the `randomForest` package, and RF models are generally insensitive to values of the tuning parameters. However, computational considerations for fitting an SLM are also not overly-demanding with modern approaches such as REML and reduced rank methods.
- Predictions from RF will always be within the range of the observed data, whereas spatial regression can extrapolate outside this range. In the context of modeling MMI, this was an advantage of the RF approach since the MMI is bounded between 0-100; the SLM also generated some negative MMI predictions that needed to be set to zero in the prediction map (Fig 5). However, in applications to other data sets, predicting within the range of sampled values is not, in general, an advantage. For instance, for an unbounded normally distributed response variable, if only 1% of the data were sampled, then we would expect that, in the other 99% unsampled sites, there will be values both greater and less than the values in the sample.
- The MMI modeling results suggest advantages to the spatial regression approach for uncertainty quantification. The SLM prediction intervals were narrower, on average, than the RF quantile prediction intervals. Moreover, the prediction intervals for the SLM are more suitable for spatial data since they scale with sampling density.

It is important to note that the cross-validation RMSPE of all models considered in Table 1 were close (ranging between 16.41-18.55) in relation to the MMI response scale (0-100). The RMSPE was just one criterion we used to compare the different models. As summarized above, there are other strengths and weaknesses of each modeling approach in terms of computational efficiency, assessing covariate relationships, and uncertainty quantification. Generally, in applications to similar types of large environmental data sets, RF has considerable advantages over the SLM when treating each approach as a black box method, as we discussed in the Introduction. We recommend the SLM if the practitioner takes a more careful approach to modeling by exploring the data and estimating transformations to account for nonlinearities. Alternatively, the RFRK model provides a compromise between the two methods by combining the RF approach for handling nonlinearities and high-order interactions, and a residual-kriging approach for constructing prediction intervals and mapping prediction errors.

The simulations demonstrated that no single type of model performs best under all conditions, and that each method is designed for specific purposes. For data sets with a high amount

of nonlinearity in the covariates, and transformations to linearity are difficult or impossible, RF or RFRK are superior methods that can uncover patterns and interactions that are difficult to recover with an LM or SLM. However, for data sets with a high amount of spatial autocorrelation and linear structure in the covariates, the SLM is the superior method. Also, if the data have a small amount spatial autocorrelation, then it may not be worth the computational effort to fit a spatial model, and either RF or LM are sufficient. The simulation results also indicate that the SLM prediction intervals generally have better coverage than the RF quantile prediction intervals.

Alternative approaches and future directions

SSN modeling. The spatial regression models for MMI stream condition can potentially be improved by using a Spatial Stream Network (SSN) [55] approach that uses stream distance, instead of straight-line, Euclidean distance, to construct a valid covariance function. SSN models also account for the topology of the stream network; i.e., whether pairs of sites are "flow-connected" (water from the upstream site flows into the downstream site) or "flow-unconnected" (water from one site does not flow into the other site). However, one limitation is that the software for fitting SSN models [56, 57] has been primarily developed for stream network data sets on small to moderately sized geographic scales, with a limited number of records ($<$2000). A substantial amount of additional work would be required to prepare the spatial information (shape files, hydrologic distances, and topological relationships) necessary to fit an SSN model to the nationally-scaled MMI stream data. Moreover, given the low amount of spatial autocorrelation found when modeling MMI using Euclidean distance, it may not be worthwhile to also consider more complex SSN models.

Spatial regression with large *n*. In this study we used a reduced rank method, based on the Sherman-Morrison-Woodbury matrix decomposition, to speed up estimation and covariate selection for the SLM. However, there are many viable alternatives to this method [58]. In fact, recent studies have found shortcomings to reduced rank methods such as a tendency to over-smooth predictions [59]. Alternatives, discussed in [58], are methods that introduce sparsity in the covariance matrix such as spatial partitioning or covariance tapering. For instance, the approach taken in spatial partitioning is to divide the spatial domain into subregions, and assume independence between observations in each subregion. This creates a block-diagonal structure in the covariance matrix that allows for parallelization during likelihood estimation. Further research is needed to determine the extent to which these types of methods can improve the performance and computational efficiency of the SLM in comparison to algorithmic modeling approaches such as RF.

RF prediction intervals. The development of prediction intervals for random forests is currently an active area of research. In this study we focused on the QRF method to construct prediction intervals, which is one of the most commonly used approaches. However, a reviewer has made us aware of several other approaches [60–62]. In particular, [61] proposed a method based on the empirical distribution of the out-of-bag predictions errors. Through simulations and the analysis of real data sets, [61] found that the intervals constructed using their method have coverage rates close to the nominal levels and tend to be narrower than competing methods such as QRF. Therefore, we recommend that practitioners also consider these more recent approaches to constructing prediction intervals for random forest models.

Concluding remarks

Going back to our introductory paragraph, environmental data may be large in n (rows) or p (columns). For spatial models, research for large n is very active, including the reduced rank

approaches, among others. Less attention is given to large *p*, although many proven techniques can be combined into an overall strategy. We suggest that modelers of spatial data carefully consider transformations to linearity and subsequent removal of covariates to obtain a parsimonious set of (transformed) covariates, including possible interactions. We provided one such example, creating indicator variables for covariates with excessive zeros, using Box-Cox transformations on nonzero covariate values, and creating their interaction. Model selection was possible for large *p* in the presence of large *n* using reduced rank methods. We stress that this is not the only strategy, but rather an example of how to proceed for both large *n* and *p*, which has received little attention. Given such a strategy, we created an SLM model that performed comparably well with RF, and had some advantages for uncertainty quantification. Conclusively, there is no correct way to statistically analyze and model large environmental data sets. The results of this study suggest that a variety of modeling approaches can be considered, and that each approach can lead to different insights into the data set and applied problem. By comparing spatial regression to RF we ultimately found ways to improve both techniques.

Supporting information

S1 File. PDF containing supplementary information. This supplement contains: (1) a step-by-step description of the covariate selection procedure; (2) computation details of the random forest regression-kriging computations; and (3) additional tables and figures.
(PDF)

S1 Data and Code. R package with data sets and simulation code. Link to GitHub repository: https://github.com/ericwfox/slmrf.
(DOCX)

Acknowledgments

We thank Dave Holland (US EPA, National Exposure Research Laboratory, Exposure Methods and Measurements Division) for providing valuable comments that improved this paper. This manuscript has been subjected to review by the Western Ecology Division of ORD's National Health and Environmental Effects Research Laboratory and approved for publication. Approval does not signify that the contents reflect the views of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use. The data from the 2008/09 NRSA used in this paper resulted from the collective efforts of dedicated field crews, laboratory staff, data management and quality control staff, analysts and many others from EPA, states, tribes, federal agencies, universities and other organizations. For questions about these data, please contact nars-hq@epa.gov.

Author Contributions

Data curation: Anthony R. Olsen.

Methodology: Eric W. Fox, Jay M. Ver Hoef, Anthony R. Olsen.

Supervision: Jay M. Ver Hoef, Anthony R. Olsen.

Writing – original draft: Eric W. Fox, Jay M. Ver Hoef.

Writing – review & editing: Eric W. Fox, Jay M. Ver Hoef, Anthony R. Olsen.

References

1. McAfee A, Brynjolfsson E, Davenport TH, Patil D, Barton D. Big data: the management revolution. Harvard Business Review. 2012; 90(10):61–67.
2. Lohr S. The age of big data. New York Times. 2012; 11 (2012).
3. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 2015; 35(2):137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
4. Chen M, Mao S, Liu Y. Big data: A survey. Mobile Networks and Applications. 2014; 19(2):171–209.
5. Hand DJ. Statistics and data mining: intersecting disciplines. ACM SIGKDD Explorations Newsletter. 1999; 1(1):16–19. https://doi.org/10.1145/846170.846171
6. Tukey JW. Use of many covariates in clinical trials. International Statistical Review/Revue Internationale de Statistique. 1991; 59(2):123–137.
7. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324
8. Ver Hoef JM, Cressie N, Fisher RN, Case TJ. Uncertainty and spatial linear models for ecological data. In: Hunsaker CT, Goodchild MF, Friedl MA, Case TJ, editors. Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications. New York: Springer-Verlag; 2001. p. 265–282.
9. USEPA. National Rivers and Streams Assessment 2008-2009: A Collaborative Survey (EPA/841/R-16/007); 2016. Available from: https://www.epa.gov/national-aquatic-resource-surveys/nrsa.
10. Stevens DL, Olsen AR. Spatially balanced sampling of natural resources. Journal of the American Statistical Association. 2004; 99(465):262–278. https://doi.org/10.1198/016214504000000250
11. USEPA. National Rivers and Streams Assessment 2008-2009 Technical Report (EPA/841/R-16/008); 2016.
12. Omernik JM. Ecoregions of the conterminous United States. Annals of the Association of American Geographers. 1987; 77(1):118–125. https://doi.org/10.1111/j.1467-8306.1987.tb00149.x
13. Stoddard JL, Herlihy AT, Peck DV, Hughes RM, Whittier TR, Tarquinio E. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society. 2008; 27(4):878–891. https://doi.org/10.1899/08-053.1
14. Hill RA, Fox EW, Leibowitz SG, Olsen AR, Thornbrugh DJ, Weber MH. Predictive mapping of the biotic condition of conterminous-USA rivers and streams. Ecological Applications. 2017; 27(8):2397–2415. https://doi.org/10.1002/eap.1617 PMID: 28871655
15. Hill RA, Weber MH, Leibowitz SG, Olsen AR, Thornbrugh DJ. The Stream-Catchment (StreamCat) Dataset: A database of watershed metrics for the conterminous United States. Journal of the American Water Resources Association. 2016; 52(1):120–128. https://doi.org/10.1111/1752-1688.12372
16. McKay L, Bondelid T, Dewald T, Johnston J, Moore R, Rea A. NHDPlus Version 2: User Guide; 2012. Available from: http://www.horizon-systems.com/NHDPlus/NHDPlusV2_home.php.
17. Li J, Heap AD, Potter A, Daniell JJ. Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software. 2011; 26(12):1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004
18. Li J, Heap AD, Potter A, Huang Z, Daniell JJ. Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin. Continental Shelf Research. 2011; 31(13):1365–1376. https://doi.org/10.1016/j.csr.2011.05.015
19. Appelhans T, Mwangomo E, Hardy DR, Hemp A, Nauss T. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Statistics. 2015; 14:91–113. https://doi.org/10.1016/j.spasta.2015.05.008
20. Hengl T, Heuvelink GB, Kempen B, Leenaars JG, Walsh MG, Shepherd KD, et al. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLOS One. 2015; 10(6):e0125814. https://doi.org/10.1371/journal.pone.0125814 PMID: 26110833
21. Fayad I, Baghdadi N, Bailly JS, Barbier N, Gond V, Hérault B, et al. Regional scale rain-forest height mapping using regression-kriging of spaceborne and airborne LiDAR data: application on French Guiana. Remote Sensing. 2016; 8(3):240. https://doi.org/10.3390/rs8030240
22. Parmentier I, Harrigan RJ, Buermann W, Mitchard ET, Saatchi S, Malhi Y, et al. Predicting alpha diversity of African rain forests: models based on climate and satellite-derived data do not perform better than a purely spatial model. Journal of Biogeography. 2011; 38(6):1164–1176. https://doi.org/10.1111/j.1365-2699.2010.02467.x
23. Temesgen H, Ver Hoef JM. Evaluation of the spatial linear model, random forest and gradient nearest-neighbour methods for imputing potential productivity and biomass of the Pacific Northwest forests. Forestry. 2014; 88(1):131–142. https://doi.org/10.1093/forestry/cpu036

24. Oliveira S, Oehler F, San-Miguel-Ayanz J, Camia A, Pereira JM. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. Forest Ecology and Management. 2012; 275:117–129. https://doi.org/10.1016/j.foreco.2012.03.003

25. Freeman EA, Moisen GG, Coulston JW, Wilson BT. Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Canadian Journal of Forest Research. 2015; 45:1–17.

26. Meinshausen N. Quantile regression forests. Journal of Machine Learning Research. 2006; 7 (Jun):983–999.

27. Cressie N. Statistics for spatial data. John Wiley & Sons; 1993.

28. Cressie N, Wikle CK. Statistics for spatio-temporal data. John Wiley & Sons; 2011.

29. Chiles JP, Delfiner P. Geostatistics: modeling spatial uncertainty. New York: John Wiley & Sons; 1999.

30. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika. 1971; p. 545–554. https://doi.org/10.1093/biomet/58.3.545

31. Harville DA. Bayesian inference for variance components using only error contrasts. Biometrika. 1974; 61(2):383–385. https://doi.org/10.1093/biomet/61.2.383

32. R Core Team. R: A Language and Environment for Statistical Computing; 2016. Available from: https://www.R-project.org/.

33. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association. 1977; 72(358):320–338. https://doi.org/10.2307/2286798

34. Cressie N, Johannesson G. Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70(1):209–226. https://doi.org/10.1111/j.1467-9868.2007.00633.x

35. Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70(4):825–848. https://doi.org/10.1111/j.1467-9868.2008.00663.x

36. Henderson HV, Searle SR. On deriving the inverse of a sum of matrices. Siam Review. 1981; 23(1):53–60. https://doi.org/10.1137/1023004

37. Box GE, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological). 1964; 26(2):211–252.

38. Fox J. Applied regression analysis and generalized linear models. 2nd ed. SAGE Publications; 2008.

39. Akaike H. A new look at the statistical model identification. IEEE transactions on automatic control. 1974; 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705

40. Hoeting JA, Davis RA, Merton AA, Thompson SE. Model selection for geostatistical models. Ecological Applications. 2006; 16(1):87–98. https://doi.org/10.1890/04-0576 PMID: 16705963

41. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer Series in Statistics. Springer New York; 2009.

42. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010; 33(1):1–22. https://doi.org/10.18637/jss.v033.i01 PMID: 20808728

43. Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems. 2006; 9(2):181–199. https://doi.org/10.1007/s10021-005-0054-1

44. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. Ecology. 2007; 88(11):2783–2792. https://doi.org/10.1890/07-0539.1 PMID: 18051647

45. Carlisle DM, Falcone J, Meador MR. Predicting the biological condition of streams: use of geospatial indicators of natural and anthropogenic characteristics of watersheds. Environmental Monitoring and Assessment. 2009; 151(1):143–160. https://doi.org/10.1007/s10661-008-0256-z PMID: 18493861

46. Evans JS, Murphy MA, Holden ZA, Cushman SA. Modeling species distribution and change using random forest. In: Drew CA, Wiersma YF, Huettman F, editors. Predictive species and habitat modeling in landscape ecology. Springer New York; 2011. p. 139–159.

47. Hill RA, Hawkins CP, Carlisle DM. Predicting thermal reference conditions for USA streams and rivers. Freshwater Science. 2013; 32(1):39–55. https://doi.org/10.1899/12-009.1

48. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods. 2009; 14(4):323–348. https://doi.org/10.1037/a0016973 PMID: 19968396

49. Biau G. Analysis of a random forests model. Journal of Machine Learning Research. 2012; 13 (Apr):1063–1095.

50. Breiman L. Bagging predictors. Machine Learning. 1996; 24(2):123–140. https://doi.org/10.1007/BF00058655

51. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002; 2(3):18–22.

52. Meinshausen N. quantregForest: Quantile Regression Forests; 2016. Available from: https://CRAN.R-project.org/package=quantregForest.

53. Hengl T, Heuvelink GB, Stein A. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma. 2004; 120(1):75–93. https://doi.org/10.1016/j.geoderma.2003.08.018

54. Irvine KM, Gitelman AI, Hoeting JA. Spatial designs and properties of spatial correlation: effects on covariance estimation. Journal of Agricultural, Biological, and Environmental Statistics. 2007; 12 (4):450–469. https://doi.org/10.1198/108571107X249799

55. Ver Hoef JM, Peterson EE. A moving average approach for spatial statistical models of stream networks. Journal of the American Statistical Association. 2010; 105(489):6–18. https://doi.org/10.1198/jasa.2009.ap08248

56. Peterson EE, Ver Hoef JM. STARS: An ArcGIS toolset used to calculate the spatial information needed to fit spatial statistical models to stream network data. Journal of Statistical Software. 2014; 56(2):1–17. https://doi.org/10.18637/jss.v056.i02

57. Ver Hoef JM, Peterson EE, Clifford D, Shah R. SSN: An R package for spatial statistical modeling on stream networks. Journal of Statistical Software. 2014; 56(3):1–45.

58. Heaton MJ, Datta A, Finley AO, Furrer R, Guinness J, Guhaniyogi R, et al. A case study competition among methods for analyzing large spatial data. Journal of Agricultural, Biological and Environmental Statistics. 2019; 24(3):398–425. https://doi.org/10.1007/s13253-018-00348-w

59. Stein ML. Limitations on low rank approximations for covariance matrices of spatial data. Spatial Statistics. 2014; 8:1–19. https://doi.org/10.1016/j.spasta.2013.06.003

60. Tung NT, Huang JZ, Nguyen TT, Khan I. Bias-corrected quantile regression forests for high-dimensional data. In: 2014 International Conference on Machine Learning and Cybernetics. vol. 1; 2014. p. 1–6.

61. Zhang H, Zimmerman J, Nettleton D, Nordman DJ. Random forest prediction intervals. The American Statistician. 2019; 0(0):1–15.

62. Zhu L, Lu J, Chen Y. HDI-Forest: Highest Density Interval Regression Forest. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; 2019.