Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Random Forests for Spatially Dependent Data

Arkajyoti Saha, Sumanta Basu & Abhirup Datta

To cite this article: Arkajyoti Saha, Sumanta Basu & Abhirup Datta (2021): Random Forests for Spatially Dependent Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1950003

To link to this article: https://doi.org/10.1080/01621459.2021.1950003

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–19: Theory and Methods
https://doi.org/10.1080/01621459.2021.1950003

Taylor & Francis
Taylor & Francis Group

Random Forests for Spatially Dependent Data

Arkajyoti Saha[a], Sumanta Basu[b], and Abhirup Datta[a]

[a]Department of Biostatistics, Johns Hopkins University, Baltimore, MD; [b]Department of Statistics and Data Science, Cornell University, Ithaca, NY

ABSTRACT

Spatial linear mixed-models, consisting of a linear covariate effect and a Gaussian process (GP) distributed spatial random effect, are widely used for analyses of geospatial data. We consider the setting where the covariate effect is nonlinear. Random forests (RF) are popular for estimating nonlinear functions but applications of RF for spatial data have often ignored the spatial correlation. We show that this impacts the performance of RF adversely. We propose RF-GLS, a novel and well-principled extension of RF, for estimating nonlinear covariate effects in spatial mixed models where the spatial correlation is modeled using GP. RF-GLS extends RF in the same way generalized least squares (GLS) fundamentally extends ordinary least squares (OLS) to accommodate for dependence in linear models. RF becomes a special case of RF-GLS, and is substantially outperformed by RF-GLS for both estimation and prediction across extensive numerical experiments with spatially correlated data. RF-GLS can be used for functional estimation in other types of dependent data like time series. We prove consistency of RF-GLS for β-mixing dependent error processes that include the popular spatial Matérn GP. As a byproduct, we also establish, to our knowledge, the first consistency result for RF under dependence. We establish results of independent importance, including a general consistency result of GLS optimizers of data-driven function classes, and a uniform law of large number under β-mixing dependence with weaker assumptions. These new tools can be potentially useful for asymptotic analysis of other GLS-style estimators in nonparametric regression with dependent data.

ARTICLE HISTORY
Received December 2020
Accepted June 2021

KEYWORDS
Gaussian processes; Generalized least squares; Random forests; Spatial

1. Introduction

Geo-referenced data, exhibiting spatial correlation, are commonly analyzed in a mixed-model framework consisting of a fixed-effect component for the covariates and a spatial random-effect (Banerjee, Carlin, and Gelfand 2014). If Y, X, and ℓ, respectively, denote the response, the D-dimensional vector of covariates, and the spatial location, then the spatial linear mixed model can be expressed as $Y = X^\top \boldsymbol{\beta} + w(\ell) + \text{error}$. The linear regression term $\mathbf{x}^\top \boldsymbol{\beta}$ parsimoniously models the fixed covariate effect $m(\mathbf{x}) = E(Y \mid X = \mathbf{x})$, while the spatial random effect $w(\ell)$ is often modeled flexibly using Gaussian Processes (GPs) that encode the dependence in the data across locations. The mixed model allows for inference on the covariate effect as the estimate of β, while adjusting for the spatial dependence via w. At the same time, it leverages the conveniences of a GP to seamlessly predict the outcome at a new location via kriging. This flexibility has made the spatial linear mixed model with GP-distributed random effects the flag-bearer in geo-statistics.

This article focuses on applications where the linearity assumption for the covariate effect $m(\mathbf{x})$ is inappropriate. We consider the spatial nonlinear mixed-effect model $Y = m(X) + w(\ell) + \text{error}$. A nonlinear $m(\mathbf{x})$ can be modeled in terms of splines or other basis function expansions, while still modeling the spatial effect using GP. However, smooth and continuous basis functions are not ideal for modeling nonsmooth and possibly discontinuous (like piecewise constant) covariate effects. Also, basis functions on multi-dimensional covariate domains experience curse of dimensionality even for only 3 or 4 covariates as the knots are usually too far apart to adequately represent the covariate space (Taylor and Einbeck 2013).

Random forest (RF) (Breiman 2001) estimates nonlinear regression functions using an ensemble of nonsmooth, data-adaptive family of basis functions, which have more expressive powers than traditional fixed basis methods (Lin and Jeon 2006; Scornet 2016). RF has become one of the most popular method for flexible nonlinear function estimation, with wide applications in different scientific and engineering fields. While RF have also been used for a number of geospatial applications (Ahijevych et al. 2016; Di et al. 2019; Lim et al. 2019; Viscarra Rossel, Webster, and Kidd 2014; Fayad et al. 2016), most of the aforementioned work do not make procedural considerations for the RF algorithm to address the spatial correlation in the data. This is fundamentally at odds with the tenets of spatial modeling where the spatial correlation is explicitly accommodated using GP-distributed random effects.

Very little attention has been paid to how ignoring this spatial correlation affects function estimation using RF. The criterion (loss function) used to recursively split the nodes of decision trees of RF is essentially an ordinary least-square (OLS) loss. It is well known that OLS is suboptimal for dependent data as it ignores data correlation. Also, RF creates and consolidates the trees using "bagging" (bootstrap aggregation)

CONTACT Abhirup Datta abhidatta@jhu.edu Department of Biostatistics, Johns Hopkins University, Ithaca, NY, 14853-0001.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

(Breiman 1996). As spatial data are correlated, this resampling violates the assumption of independent and identically distributed (iid) data units, fundamental to bootstrapping. We provide empirical evidence that these limitations manifest in inferior estimation/prediction performance of RF under spatial dependence.

There have been two ways of accounting for spatial dependence in RF. The first approach performs spatial adjustment through kriging *after* fitting an RF without accounting for the spatial dependence. For prediction at new locations, it uses the spatial mixed model framework to combine this vanilla RF estimate of the covariate effect $m(X)$ with kriging from the GP part (using the residuals from the RF fit) (Fayad et al. 2016; Fox, Ver Hoef, and Olsen 2020). This is referred to as RF residual kriging (RF-RK). The other approaches attempt to explicitly use spatial information in RF. The spatial RF (RFsp) of Hengl et al. (2018) adds all pairwise spatial-distances as additional covariates. Georganos et al. (2019) proposed geographically local estimation of RF. These approaches abandon the spatial mixed model framework, only focusing on prediction, modeling the response as a joint function g of the covariates and the locations, that is, $E(Y) = g(X, \ell)$, and are thus not able to isolate or estimate the covariate effect $m(\mathbf{x})$. Also, to our knowledge, there is no asymptotic theory justifying these approaches.

In this article, we bridge the gap between spatial mixed-effect modeling using GP and regression function estimation using RF by proposing a well-principled rendition of RF to explicitly incorporate the spatial correlation structure implied by GP. *This enables using RF for estimating a nonlinear covariate effect in the spatial mixed-effect model while explicitly accounting for spatial correlation via the GP-distributed random effects.* Our approach is motivated by the fact that in a GP-based spatial linear mixed model, marginal likelihood maximization for β is equivalent to a generalized least-square (GLS) optimization. Under data dependence, GLS is more efficient than OLS for linear regression, leading to better finite sample performance.

GLS has been previously used for parameter estimation in nonlinear regression under data dependence (Kimeldorf and Wahba 1971; Diggle and Hutchinson 1989). Since regression with polynomials, splines or other *known* basis functions are essentially linear in the parameters, the extension of these approaches to a GLS-style quadratic form global optimization is immediate and the solution is available in closed form. However, RF uses *basis functions obtained by a data-adaptive, greedy algorithm*, so a GLS formulation for estimating the regression part using RF in a spatial GP regression is not immediate and has not been explored before.

An observation, central to our GLS extension of regression trees and forests is that to split a node of a decision tree in RF, the local (intra-node) loss can be equivalently represented as a global (in all nodes) OLS linear regression problem with a binary design matrix. The subsequent node representative assignment is simply the OLS fit given the chosen split. For dependent data, we can replace this OLS step with a GLS optimization problem at every node split and grow the tree accordingly. The node representatives also become the GLS fits instead of node means. The global GLS loss-based splitting and fits thus use information from all current nodes as is desirable under data-dependence and are more efficient than the OLS analogs.

Our GLS-style RF also naturally accommodates data resampling or sub-sampling used for creating a forest of trees. The key observation for this is that in a linear regression between response $\mathbf{Y}$ and covariates $\mathbf{X}$, the GLS loss with covariance matrix $\boldsymbol{\Sigma}_0$ coincides with an OLS loss with $\tilde{\mathbf{Y}} = \boldsymbol{\Sigma}_0^{-1/2}\mathbf{Y}$, $\tilde{\mathbf{X}} = \boldsymbol{\Sigma}_0^{-1/2}\mathbf{X}$. Thus GLS can be thought of as OLS with the decorrelated responses $\tilde{\mathbf{Y}}$. Analogously, we introduce resampling after the decorrelation in our algorithm, essentially resampling the uncorrelated contrasts $\tilde{\mathbf{Y}}$ instead of the correlated outcomes $\mathbf{Y}$.

We refer to our method as RF-GLS and present a computationally efficient algorithm for implementing it. RF-GLS reduces exactly to RF when the working correlation matrix used in the GLS-loss is the identity matrix. For spatial mixed model regression, RF-GLS estimates the nonlinear covariate effect while using the GP to model the spatial random effect. Hence, traditional spatial tasks like kriging (spatial predictions) can be performed easily. For large spatial data, RF-GLS also avoids onerous big GP computations by harmonizing with the nearest neighbor GPs (Datta et al. 2016a) to yield an algorithm of linear time-complexity.

We show that RF-GLS has distinct advantages over competing methods for both estimation and prediction. For estimation, RF-GLS, accounting explicitly for the spatial correlation via GLS loss, provides improved estimates of the covariate effect $m(X)$ in a wide range of simulation scenarios over RF which completely ignores the spatial information. Other methods like RFsp do not even produce a separate estimate of the covariate effect and cannot be considered for the task of estimation. For prediction, RF-GLS outperforms both RF-RK and RFsp. RF-GLS conveniently uses the spatial mixed model framework where the structured spatial dependence is parsimoniously encoded via GP. Methods like RFsp abandon this mixed-model framework and uses a large number of additional distance-based covariates in RF. A downside to this unnecessary escalation of the problem to high-dimensional settings is that the several additional distance-based covariates far outnumber the true set of covariates thereby biasing the covariate selection at each node-splitting toward the spatial covariates. This leads to poor prediction performance of RFsp when the spatial noise is small relative to the covariate signal. On the other hand, when the spatial noise is large, RF-RK, estimating the covariate effect $m(X)$ without accounting for the spatial dependence, performs substantially poorly while RFsp (dominated by spatial covariates) performs comparably to RF-GLS. RF-GLS performs best at all ranges of the covariate-signal-to-spatial-noise-ratio (SNR), where each of RF-RK and RFsp suffers at opposite ends of this SNR spectrum (see Figure 3).

1.1. Theoretical Contributions

Another major contribution of this article is a thorough theoretical analysis of RF-GLS under dependent error processes like GP. For iid data, Scornet et al. (2015) proved the consistency of Breiman's RF, which allows the node splitting and node representation to be based on the same data. The other strand of RF theory consider "honest" trees (Mentch and Hooker 2016; Wager and Athey 2018) which use disjoint data subsets for splitting and representation. To our knowledge, study of RF under dependent processes has not been conducted in either

paradigms. RF-GLS, like Breiman's RF, uses the same data for partitioning and node representation. Hence we adopt the framework of Scornet et al. (2015) to study consistency.

Our main result (Theorem 3.1) proves RF-GLS is $\mathbb{L}_2$ consistent if the error process is absolutely-regular (β-) mixing (Bradley 2005). As corollary, we prove RF-GLS is consistent for nonlinear mean estimation in the spatial mixed model regression under the ubiquitous Matérn family of covariance functions. As a byproduct of the theory, we also establish consistency of the vanilla CART (Breiman et al. 1984) and of RF under β-mixing error processes. To our knowledge, this is the first result on consistency of these procedures under dependence.

The theoretical analysis, besides being novel in terms of establishing consistency for forest estimators under dependence, required addressing several new challenges that do not arise for the analysis of classic RF for iid settings. These complications were introduced by the global GLS-style quadratic loss and the dependent errors. We summarize the new contributions here:

1. *Limits of GLS regression-trees:* Node splitting and node representation in regression trees of classic RF uses local (intra-node) sample means and variances which trivially asymptote to their population analogs. For RF-GLS, to adjust for data correlation, we use a global GLS (quadratic form) loss for node-splitting and the set of node representatives is the global GLS estimate. Both steps now depend on data from all nodes (weighted by their spatial dependence) and their asymptotic limits are non-trivial. We meticulously track these data-weights from all nodes when splitting a given node to show in Lemma 2.1 and Theorem 2.1 that the RF-GLS split-criterion and node representatives converge to the same desirable local population limits as in RF while being empirically more efficient under dependence. These findings may seem similar to the well-known fact that GLS and OLS estimates both converge to the same limits with the GLS being more efficient under dependence. However, those classic results assume a true linear model whereas for us the true function is nonlinear $\mathbb{E}(y) = m(\mathbf{X})$, and is being estimated by a mis-specified tree estimate. To our knowledge, limits of the global GLS tree estimates as population-level local node means for that tree is a new contribution.
2. *Equicontinuity of split-criterion:* A center-piece in the proof of consistency of RF in Scornet et al. (2015) was stochastic equicontinuity of the CART split criterion, such that if two set of splits are close, their corresponding empirical split-criterion values are close, irrespective of the location of the splits. For RF-GLS, the split criterion (GLS loss with plugged in GLS estimate) is is a complicated matrix-based function of the design matrix representing a set of splits. Hence, the scalar techniques of Scornet et al. (2015) do not apply here. Our equicontinuity proof is quite involved and is built on viewing GLS predictions as oblique projections on the design matrices corresponding to the splits and subsequently using results on perturbations of projection operators. More details about the technique used can be found in Section 5.2.1.
3. *New uniform laws of large number (ULLNs):* Controlling the quadratic form estimation error for RF-GLS under dependence by an ULLN is key to establishing consistency. The standard technique of first proving an ULLN that replaces the dependent errors by iid ones fails here as the iid error sequence does not preserve the distribution of the cross-product terms of the quadratic form. We use a novel strategy of creating separate bivariate iid error sequences for each of the off-diagonal bands of cross-product terms in the quadratic form. We then establish a new ULLN (Proposition S2.1) for these cross-product terms having the same concentration bound (in terms of random $\mathbb{L}_p$ norm entropy numbers) as the squared (diagonal) terms. Next, to generalize the ULLNs from iid to dependent settings, existing results require Lipschitz continuity or an uniform bound on the estimators none of which are satisfied by regression-trees. We established a general ULLN (Proposition 5.3) for β-mixing processes that uses a weaker $(2+\delta)$th moment assumption that is satisfied by regression-trees. Neither of the two ULLNs were needed in the asymptotic study of RF for iid data as there was no quadratic form loss or cross-product terms, and the error process was independent thereby not requiring any dependent ULLNs.
4. *General tools for machine learning for dependent data:* Using these ULLNs described above we established a general consistency result (Theorem 5.1) for GLS estimates for a broad class of functions under dependent error. This result generalizes Theorem 10.2 of Györfi et al. (2002) (which was for iid data and OLS estimates) to β-mixing dependent processes and GLS losses. We believe this general result would be of widespread interest. Just like the iid result of Györfi et al. (2002) was used to establish consistency of a wide range of OLS estimators (including piecewise polynomials, univariate and multivariate splines, data-driven partitioning based estimators like RFs and trees), our Theorem 5.1 can potentially be used to establish consistency of the analogs of each of these methods that explicitly accounts for dependence (spatial/serial correlation) in the data by switching to a GLS procedure.
5. *Consistency for Matérn GPs:* We prove consistency of RF-GLS when the true dependent error process comes from the Matérn Gaussian family (the staple choice for geospatial analysis due to interpretability of the parameters in terms of spatial surface smoothness). The proof combines several results spread over the fields of time-series, stochastic differential equations, information theory, and spatial statistics. To our knowledge, this is the first consistency result for an RF-type algorithm under Matérn spatial correlation.

While RF-GLS is motivated by the spatial GP-based mixed model framework, the method and the theory is developed much more broadly and can be used for regression function estimation in many other dependent data settings. The method only relies on knowledge of the residual covariance matrix and thus can work with any dependent error process with a valid second moment. The theory is applicable for the large class of β-mixing processes. We briefly discuss how RF-GLS can be used for function estimation under serially correlated errors (time series). In particular, we show that for autoregressive errors, one of the mainstays of time-series analysis, RF-GLS would yield a consistent estimate of the regression function.

The rest of the article is organized as follows. In Section 2 we present our methodology. The theory of consistency is presented in Section 3, including theory for common spatial and time-series processes in Section 3.3. Results from a variety of simulation experiments validating our method is presented in Section 4. While the formal proofs of all the results are provided in the supplementary materials, an outline of the proof of the main consistency result is presented in Section 5 that highlights the new technical tools developed in the process. We discuss future extension of the presented work in Section 6.

1.2. Notations

$|S|$ denotes the cardinality of any set S. $\mathbb{I}(\cdot)$ is the indicator function. The null set is $\{\}$. For any matrix $\mathbf{M}$, $\mathbf{M}^+$ denotes its generalized Moore–Penrose inverse, and $\|\mathbf{M}\|_p$, for $1 \le p \le \infty$, denotes its matrix $\mathbb{L}_p$ norm. For any $n \times n$ symmetric matrix $\mathbf{M}$, $\lambda_{\min} = \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n = \lambda_{\max}$ denotes its eigenvalues. A sequence of numbers $\{a_n\}_{n\ge 1}$ is $O(b_n)$ (or $o(b_n)$) when the sequence $|a_n/b_n|$ is bounded above (or goes to 0) as $n \to \infty$. A sequence of random variables is called $O_b(1)$ if it is uniformly bounded almost surely, $O_p(1)$ if it is bounded in probability, and $o_p(1)$ if it goes to 0 in probability. $X \sim Y$ implies X follows the same distribution as Y. $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the set of real numbers, integers, and natural numbers respectively. For $M \in \mathbb{R}^+$, T_M is the truncation operator, that is, $T_M(u) = \max(-M, \min(u, M))$.

2. Method

2.1. Review of Spatial GP Regression

The standard geospatial data unit is the triplet (Y_i, X_i, ℓ_i) where Y_i is the univariate response, X_i is a D-dimensional covariate (feature) vector, and ℓ_i is the spatial location (often, geographical co-ordinates). A spatial linear mixed-model for such data is specified as $Y_i = X_i^\top \boldsymbol{\beta} + w(\ell_i) + \epsilon_i^*$, where $X_i^\top \boldsymbol{\beta}$ is the linear mean (covariate effect), $w(\ell_i)$ models the spatial structure in the response not accounted for by the covariates, and $\epsilon_i^* \overset{\text{iid}}{\sim} N(0, \tau^2)$ is the random noise. The spatial effect $w(\ell_i)$ is often posit to be a smooth surface across space and is modeled as a centered GP $w(\cdot) \sim \text{GP}(0, C(\cdot,\cdot \mid \theta))$ with covariance function C. This essentially means that for any finite collection of locations $\ell_1, \ldots, \ell_n$, we have $\mathbf{w} = (w(\ell_1), \ldots, w(\ell_n))^\top \sim N(\mathbf{0}, \mathbf{C})$ where $\mathbf{C}$ is the $n \times n$ matrix with entries $\text{cov}(w(\ell_i), w(\ell_j)) = C(\ell_i, \ell_j \mid \boldsymbol{\theta})$.

We can marginalize over the spatial random effects $w(\ell_i)$ and write $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\Sigma}_0 = \mathbf{C} + \tau^2\mathbf{I}$. For known $\boldsymbol{\Sigma}_0$, maximizing this marginalized likelihood of $\mathbf{Y}$ is equivalent to minimizing the quadratic form $\min_{\boldsymbol{\beta}} \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_0^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. This expression can be recognized as the GLS loss which replaces the squared error loss $(1/n)\sum_{i=1}^n (Y_i - X_i^\top \boldsymbol{\beta})^2$ of OLS when the data are correlated.

We focus on estimating a general mean function in the spatial nonlinear mixed model

$$Y_i = m(X_i) + w(\ell_i) + \epsilon_i^*;\ w(\cdot) \sim GP(0, C(\cdot,\cdot \mid \theta)). \quad (1)$$

When $m(\mathbf{x})$ is modeled in terms of a fixed basis expansion, the marginalized model becomes $\mathbf{Y} \sim N(B(\mathbf{X})\boldsymbol{\gamma}, \boldsymbol{\Sigma}_0)$ where $B(\mathbf{X})$ are the bases and $\boldsymbol{\gamma}$ are the coefficients. Hence, the model remains linear in the unknown coefficients $\boldsymbol{\gamma}$ which can be once again estimated using GLS.

As discussed in the introduction, the scope of fixed basis function expansions are limited due to their inability to model regression functions with discontinuities, and curse of dimensionality in multi-dimensional covariate domains. RF are particularly suitable for such general regression function estimation due to their use of regression trees, and natural accommodation of higher covariate-dimensionality using random selection. However, unlike fixed basis expansion models which are linear in the parameters, RF estimation is a nonlinear greedy algorithm for which a GLS extension is not straightforward. In the next sections, we will develop a GLS-style RF algorithm for estimating $m(X)$ under dependence.

2.2. Revisiting the RF Algorithm

We first review the original RF algorithm. Given data $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^D$, $i = 1, \ldots, n$, the RF estimate of the mean function $m(\mathbf{x}) = \mathbb{E}(Y \mid X = \mathbf{x})$ is the average of n_{tree} regression tree estimates of m. In a regression tree, data are split recursively into nodes of a tree starting from a root node. To split a node, a set of $M_{try} (\ll D)$ features are chosen randomly and the best split point is determined with respect to each feature d by searching over all the "gaps" in that feature of the data as cutoff candidate $c := c(d)$. Here, "best" is determined as the feature-cutoff combination (d, c) maximizing the CART (classification and regression trees) split criterion (Breiman et al. 1984)

$$v_n^{\text{CART}}((d,c)) = \frac{1}{n_P}\left[\sum_{i=1}^{n_P}(Y_i^P - \bar{Y}^P)^2 - \sum_{i_r=1}^{n_R}(Y_i^R - \bar{Y}^R)^2 - \sum_{i_l=1}^{n_L}(Y_i^L - \bar{Y}^L)^2\right], \quad (2)$$

where, Y_i^P, Y_i^R, and Y_i^L denote the responses in parent node, right child, and left child, respectively. $\bar{Y}^P$, $\bar{Y}^R$, and $\bar{Y}^L$ denote the respective node means, $n_P = n_R + n_L$, n_R, and n_L are the respective node cardinalities. The feature and cut-off value combination that minimizes the CART-split criterion (2) is chosen to create the child nodes. As each split is only based on one feature, all nodes are hyper-rectangles. Each newly created node is assigned a node representative—the mean of the responses of the node members.

The nodes are iteratively partitioned this way till a prespecified stopping rule is met—we arrive at leaf nodes having single element (fully grown tree) or the number of data points in each leaf node is less than a prespecified number or the total number of nodes reaches a prespecified threshold. The regression tree estimate for an input feature $\mathbf{x} \in \mathbb{R}^D$ is given by the representative value of the leaf node containing $\mathbf{x}$. RF is the average of an ensemble of regression trees with each tree only using a resample or subsample of the data.

2.3. Dependency Adjusted Node-Splitting

The RF algorithm of Section 2.2 does not utilize any information on the locations ℓ_i or the ensuing spatial correlation.

The fundamental unit of the algorithm is splitting of a given node, and it is inherently local in nature (in the covariate domain). The split criterion (2) and subsequent assignment of the node representatives are both based only on members within the parent node. For iid data, this local approach is reasonable as members of one node are independent of the others. However, geo-referenced data units data can be distant in the covariate-domain while being close in the spatial domain, and therefore strongly correlated.

The spatial nonlinear mixed model from Section 2.1 can be written more generally as $Y_i = m(X_i) + \epsilon_i$ where $\epsilon_i = w(\ell_i) + \epsilon_i^*$ is not an iid process but a stochastic (Gaussian) Process capturing the spatial dependence. Hence, $\text{cov}(Y_i, Y_j) = \text{cov}(\epsilon_i, \epsilon_j) \neq 0$ for most or all (i, j) pairs implying that members of the other nodes can be highly correlated with those of a node-to-be-split and operating locally (intra-node) leaves out this information. We now propose a new global split-criterion using all the correlations in the data, as is desirable and in line with the common practice in spatial analysis.

To explore generalizations of RF for dependent settings, we first recall an equivalent global representation of the CART-split criterion (2). Consider a regression tree grown up to the set of leaf nodes $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ forming a partition of the feature space. The node representatives $\widehat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)^\top$ are the corresponding means. To split the next node, say $\mathcal{C}_K$ without loss of generality, the CART-split criterion (2) determines the best feature-cutoff combination (d^*, c^*) and creates child nodes $\mathcal{C}_K^{(L)} = \mathcal{C}_K \cap \{\mathbf{x} \in \mathbb{R}^D | x_{d^*} < c^*\}$ and $\mathcal{C}_K^{(R)} = \mathcal{C}_K \cap \{\mathbf{x} \in \mathbb{R}^D | x_{d^*} \geq c^*\}$ and the node representatives $\hat{\beta}_K^{(L)}$ and $\hat{\beta}_K^{(R)}$ (simply the means of the left and right child nodes respectively). We can write the optimal split-direction (d^*, c^*) and node representatives $\hat{\beta}_K^{(L)}, \hat{\beta}_K^{(R)}$ as the maximizer

$$(d^*, c^*, \hat{\beta}_K^{(L)}, \hat{\beta}_K^{(R)}) = \underset{d,c,(\beta^{(L)},\beta^{(R)})\in\mathbb{R}^2}{\arg\max} \frac{1}{|\mathcal{C}_K|}\Big[\sum_{X_i\in\mathcal{C}_K} (y_i - \hat{\beta}_K)^2 - \Big(\sum_{X_i\in\mathcal{C}_K, X_{id}<c} \left(y_i - \beta^{(L)}\right)^2 + \sum_{X_i\in\mathcal{C}_K, X_{id}\geq c} \left(y_i - \beta^{(R)}\right)^2 \Big)\Big]. \tag{3}$$

After the split, the new set of nodes is $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{K-1}, \mathcal{C}_K^{(L)}, \mathcal{C}_K^{(R)}\}$ and the set of representatives is updated to $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{K-1}, \hat{\beta}_K^{(L)}, \hat{\beta}_K^{(R)})^\top$. Let $\mathbf{Z}_0 = (\mathbb{I}(X_i) \in \mathcal{C}_j)$ be the $n \times K$ membership matrix before the split, and $\mathbf{Z}$ denote the $n \times (K+1)$ membership matrix for a split of $\mathcal{C}_K$ using (d, c), that is, $\mathbf{Z}_{ij} = \mathbf{Z}_{ij}^{(0)}$ for $j < k$, $\mathbf{Z}_{iK} = \mathbb{I}(\{X_i \in \mathcal{C}_K\} \cap \{X_{id} < c\})$ and $\mathbf{Z}_{i(K+1)} = \mathbb{I}(\{X_i \in \mathcal{C}_K\} \cap \{X_{id} \geq c\})$. We note but suppress the dependence of $\mathbf{Z}$ on (d, c). Then the optimization in Equation (3) can be rewritten in the following way.

$$(d^*, c^*, \hat{\boldsymbol{\beta}}) = \underset{d,c,\boldsymbol{\beta}\in\mathbb{R}^{K+1}}{\arg\max} \frac{1}{n}\left(\|\mathbf{Y} - \mathbf{Z}^{(0)}\widehat{\boldsymbol{\beta}}^{(0)}\|_2^2 - \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2 \right). \tag{4}$$

This connection of OLS regression with RF is well-known (see, e.g., Friedman et al. 2008 where RF rules are used in downstream Lasso fit). To see why Equations (3) and (4) are equivalent, note that, for a given (d, c) the $\hat{\boldsymbol{\beta}}$ optimizing (4) is simply $\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Y}$, the minimizer of the OLS loss $\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2$. Since the membership matrices have orthogonal columns consisting of only 1s and 0s, the components of $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ are simply the node means. Hence, the last two components of $\boldsymbol{\beta}$ are $\hat{\beta}_K^{(L)}$ and $\beta_K^{(R)}$. Since the first $(K-1)$ columns of $\mathbf{Z}$ are same as that of $\mathbf{Z}^{(0)}$, this implies that the first $(K-1)$ components of $\hat{\boldsymbol{\beta}}$ are the same as that of $\hat{\boldsymbol{\beta}}^{(0)}$, that is, the means of the first $(K-1)$ nodes. So although $\hat{\boldsymbol{\beta}}$ in (4) is a global optimizer of a linear regression re-estimating all the node representatives, in practice, only the representatives of the child nodes of $\mathcal{C}_K$ are updated and this is equivalent to the local optimization (3).

The global formulation of node-splitting offers an avenue for GLS-style generalization of the split criterion for the spatial GP regression. Let $\boldsymbol{\Sigma}_0 = \text{cov}(\mathbf{Y}) = \text{cov}(\boldsymbol{\epsilon})$ denote the covariance matrix of the marginalized response, and $\mathbf{Q} = \boldsymbol{\Sigma}_0^{-1}$. Then à la GLS we can simply replace the squared error loss $\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2$ with a quadratic loss $(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})^\top\mathbf{Q}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})$, and propose a GLS-style split criterion

$$\begin{aligned} &v_{n,\mathbf{Q}}^{\text{DART}}((d, c)) \\ &= \frac{1}{n}\Big[\left(\mathbf{Y} - \mathbf{Z}^{(0)}\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z}^{(0)})\right)^\top \mathbf{Q} \left(\mathbf{Y} - \mathbf{Z}^{(0)}\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z}^{(0)})\right) \\ &\quad - \left(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z})\right)^\top \mathbf{Q} \left(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z})\right) \Big]. \end{aligned} \tag{5}$$

We refer to Equation (5) as the dependency-adjusted regression tree (DART) split criterion. For a split (d^*, c^*) maximizing (5), the new set of node representatives are given by the GLS estimate

$$\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z}) = \hat{\boldsymbol{\beta}} = \left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Y}\right). \tag{6}$$

Both the loss used to choose the optimal split and the node representatives now depend on data from all nodes, weighted by the precision (inverse covariance) matrix $\mathbf{Q}$, akin to the spatial linear mixed-model. Even though we preserve the greedy recursive partitioning strategy of RF, for each node split our loss function and node representation are now global in the sense that they consider all the data points, not just the ones inside the node to be split.

To demonstrate why the DART loss function and the GLS node representatives are respectively more appropriate for dependent data than the CART loss and node means used in the original RF, we now present a theoretical result and an empirical example. To split a given parent node $\mathcal{B}$ into left and right child nodes $\mathcal{B}^L$ and $\mathcal{B}^R$, respectively, based on the split direction (d, c), it is almost immediate that the CART split criterion (2), being the difference between the sample variance of the parent node with the those of the children nodes, is an estimator of its asymptotic limit

$$\begin{aligned} \text{Vol}(\mathcal{B})\Big[&\mathbb{V}(Y|\mathbf{X}\in\mathcal{B}) - \mathbb{P}(\mathbf{X}\in\mathcal{B}^R|\mathbf{X}\in\mathcal{B})\mathbb{V}(Y|\mathbf{X}\in\mathcal{B}^R) \\ &-\mathbb{P}(\mathbf{X}\in\mathcal{B}^L|\mathbf{X}\in\mathcal{B})\mathbb{V}(Y|\mathbf{X}\in\mathcal{B}^L)\Big], \end{aligned} \tag{7}$$

that is, the population difference between the variance in the parent node and the pooled variance in the potential children nodes. Indeed, if we were privileged to infinite amount of data, we should use Equation (7) to split a node as it minimizes the total intra-node variances in the children. Hence, for iid data, it is reasonable to use the CART criterion (2) which is the finite sample analogue of Equation (7). Under dependence, however, it is unclear if Equation (2) is an efficient estimator of Equation (7).

Our construction of the DART split criterion is guided by the same principles of GLS used in linear regression for dependent data. The resulting loss function (5) depends on data from not only the node $\mathcal{B}$ to be split, but on all data, and it is not immediately clear what its population limit is. Optimization of Equation (5) for the optimal split (d^*, c^*) relies on the GLS estimate (6) which is also a function of all data. Hence, we first provide a result about its asymptotic limit. We use assumptions on the dependence structure of the error process (Assumption 1) and regularity of the precision matrix (Assumption 2) discussed in more details later in Section 3 on consistency of our method. The proofs of these results can be found in Section S2.1 (supplementary materials).

Assumption 1 (*Mixing condition*). $Y_i = m(X_i) + \epsilon_i$ where the error process $\{\epsilon_i\}$ is a stationary, *absolutely regular* (β-mixing) process (Bradley 2005) with finite $(2+\delta)$th moment for some $\delta > 0$.

Assumption 2 (*Regularity of the working precision matrix*). The working precision matrix $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ admits a regular and sparse lower-triangular Cholesky factor $\mathbf{\Sigma}^{-\frac{1}{2}}$ such that

$$\mathbf{\Sigma}^{-\frac{1}{2}} = \begin{pmatrix} \mathbf{L}_{q\times q} & 0 & 0 & \cdots & \cdots \\ \rho^{\top}_{1\times(q+1)} & & 0 & \cdots & \cdots \\ 0 & \rho^{\top}_{1\times(q+1)} & & 0 & \cdots \\ \vdots & \ddots & & & \vdots \\ \cdots & 0 & 0 & \rho^{\top}_{1\times(q+1)} & \end{pmatrix},$$

where $\boldsymbol{\rho} = (\rho_q, \rho_{q-1}, \cdots, \rho_0)^{\top} \in \mathbb{R}^{q+1}$ for some fixed $q \in \mathbb{N}$, and $\mathbf{L}$ is a fixed lower-triangular $q \times q$ matrix .

Lemma 2.1 (*Limit of GLS estimate for fixed partition regression-tree*). Under Assumptions 1 and 2, for a regression-tree based on a fixed (data-independent) partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$, the GLS estimate $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)^{\top}$ from Equation (6) has the following limit.

$$\hat{\boldsymbol{\beta}}_l \xrightarrow{\text{a.s.}} \mathbb{E}(Y|X \in \mathcal{C}_l) \text{ as } n \to \infty \text{ for } l = 1, \ldots, K.$$

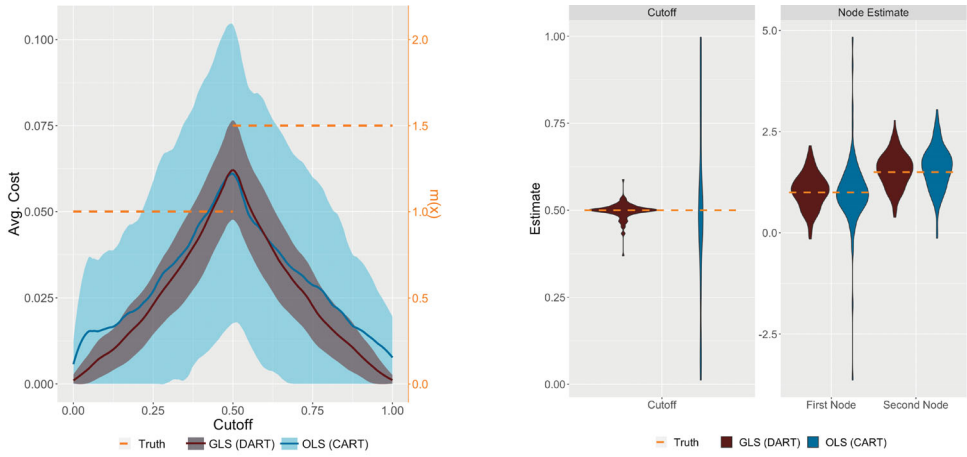
Lemma 2.1 shows that the GLS-estimate for the node representatives, although using data from all the nodes, asymptote to within node population means – the same limit as that of the sample node means used in RF. This is a new result of independent importance showing that the GLS estimate for a fixed-partition regression-tree, is a consistent estimator of the node-specific conditional means. This in turn leads to the following result about the DART loss.

Theorem 2.1 (*Theoretical DART-split criterion*). Under Assumptions 1 and 2, for a tree built with a fixed (data-independent) set of splits $\mathcal{C}_1, \ldots, \mathcal{C}_K$, as $n \to \infty$ the empirical DART-split criterion (5) to split a node $\mathcal{B} = \mathcal{C}_l$ into left and right child nodes $\mathcal{B}^L$ and $\mathcal{B}^R$ respectively, converges almost surely to the following for some constant $\alpha = \alpha(\mathbf{Q})$:

$$\begin{aligned} v^*_{\mathbf{Q}}((d,c)) = \alpha \text{Vol}(\mathcal{B}) \Big[& \mathbb{V}(Y|\mathbf{X} \in \mathcal{B}) \\ & - \mathbb{P}(\mathbf{X} \in \mathcal{B}^R|\mathbf{X} \in \mathcal{B}) \mathbb{V}(Y|\mathbf{X} \in \mathcal{B}^R) \\ & - \mathbb{P}(\mathbf{X} \in \mathcal{B}^L|\mathbf{X} \in \mathcal{B}) \mathbb{V}(Y|\mathbf{X} \in \mathcal{B}^L) \Big]. \end{aligned} \tag{8}$$

The aforementioned result shows that remarkably the limit of the the DART-split criterion converge to the same respective limit (7) of the CART-split criterion up to a constant α. This result is intriguing as although the empirical DART split criterion depends on the entire set of splits and data from all the nodes, its asymptotic limit is simply the population variance difference (7) between the parent node and the children nodes and does not depend on the other nodes. As discussed before, the quantity (7) would be the ideal one to use for splitting a node if one had knowledge of the population distribution, and it is reassuring that the DART criterion asymptotes to it.

These asymptotic results are in line with the conventional GLS wisdom. GLS estimators are known to have the same asymptotic limit as the OLS estimators but are more efficient under dependence. To demonstrate how these asymptotic results translate to finite sample performance, we conduct a simple experiment using data generated from $Y_i = m(X_i) + \epsilon_i$, where ϵ_i is an GP with exponential covariance function on the regularly spaced one-dimensional lattice, and $m(x)$ is a function of a single covariate supported on $[0, 1]$ such that $m(x) = 1$ for $x \leq 0.5$, and $m(x) = 1.5$ for $x > 0.5$. If we use a two-node decision tree to estimate m, then it is obvious that a good loss criterion should be maximized near the cut-off value of 0.5 where there is a discontinuity in the true regression function. In Figure 1(a), we plot the average CART and DART split criterion as a function of the choice of cutoff, and the point-wise confidence bands. The DART split criterion was scaled by α to have the same asymptotic limit as the CART split criterion. For reference we also plot the true regression function on the secondary y-axis. We see that the average curves for both the CART and DART criterion are quite identical, both peaking near the true cutoff of 0.5. However, the CART criterion using the OLS loss has large uncertainty as reflected by the much wider confidence bands, than the DART criterion. This in turn affects the estimates of the cutoff and the node representatives as reflected in Figure 1(b). While both losses lead to similar mean estimates of the cutoff, and node representatives, the variability is substantially higher for the OLS loss. This large variability is especially evident for the choice of the cutoff where the CART loss can choose cutoff far away from



(a) CART and DART criterion (b) Estimates of split-cutoff and node representatives.

Figure 1. Comparison between OLS loss (CART criterion) and our GLS loss (DART criterion) for node-splitting in a regression tree under GP correlated errors. Left Figure (a) plots the average CART (blue) and DART (brown) criterion and point-wise 95% confidence bands over 100 replicate datasets. The true regression function $m(x)$ is plotted in orange in the secondary axis. Right figure (b) plots the densities (violin plots) of the estimates of the cutoff used for node-splitting and the node representatives from the two loss functions.

0.5 with high probability whereas the DART loss chooses a cutoff near 0.5 almost always. We will show in Section 4 over a wide range of simulation studies how this leads to poor finite sample estimation and prediction performance of RF for dependent data.

2.4. GLS-Style Regression Tree

The DART split criterion (5) and node representation (6) constitute the fundamental node-splitting operation of a GLS-style regression tree algorithm that incorporates the spatial information via the correlation matrix. We now introduce additional notation to formally detail the algorithm of this GLS-style regression tree. Creation of a forest from the trees will be discussed in Section 2.5.

For ease of presentation, so far we have talked about the case of splitting the last node (Kth) at a given level of the tree. More generally, we can denote the complete set of nodes in level $k-1$ by $\mathfrak{C}^{(k-1)} = \{\mathcal{C}_1^{(k-1)}, \mathcal{C}_2^{(k-1)}, \cdots, \mathcal{C}_{g^{(k-1)}}^{(k-1)}\}$ which is a partition of the feature space. To split the l_1th node $\mathcal{C}_{l_1}^{(k-1)}$, we consider the following membership matrices: $\mathbf{Z}^{(0)}$, which corresponds to the nodes in parent level, with the column for the node-to-be-split pushed to the last column, and $\mathbf{Z}$ which corresponds to the membership of the potential child nodes

$$\begin{aligned} \mathcal{C}_{l_1^{(1)}}^{(k)} &= \mathcal{C}_{l_1}^{(k-1)} \cap \{\mathbf{x} \in \mathbb{R}^D | x_d < c\}, \\ \mathcal{C}_{l_1^{(2)}}^{(k)} &= \mathcal{C}_{l_1}^{(k-1)} \cap \{\mathbf{x} \in \mathbb{R}^D | x_d \geq c\} \end{aligned} \tag{9}$$

based on a split (d, c). Note that $\mathbf{Z}$ above is a function of the previous set of nodes $\mathfrak{C}^{(k-1)}$, the node to be split l_1 and the split (d, c) all of which is kept implicit. We denote the GLS-style split criterion (5) as $v_{n,\mathbf{Q}}^{\text{DART}}(\mathfrak{C}^{(k-1)}, l_1, (d, c)) := v_{n,\mathbf{Q}}^{\text{DART}}(d, c)$ for the node $\mathcal{C}_{l_1}^{(k-1)}$ at (d, c).

Equipped with the notation, the GLS-style regression tree algorithm is presented below:

Algorithm 1 GLS-style random regression tree

Input: Data $\mathcal{D}_n = (Y_1, X_1, \ldots, Y_n, X_n)$, working precision matrix $\mathbf{Q}$, stopping rules t_n (maximum number of nodes) and t_c (minimum number of members per node), number of features considered for each split M_{try}, randomness R_Θ (some probability distribution to choose a subsample of size M_{try} from $\{1, \ldots, D\}$), output point $\mathbf{x}_0$.

Output: Estimate $m_n(\mathbf{x}_0; \Theta)$ of the mean function m at $\mathbf{x}_0$.

procedure
  Initialize $k \leftarrow 1$; $\mathfrak{C}^{(1)} \leftarrow \{\mathbb{R}^D\}$, $g^{(1)} \leftarrow 1$;
  while $g^{(k)} < t_n$ and $|\mathcal{C}_l^{(k)}| > t_c$ for at least one $l \in 1, 2, \ldots, g^{(k)}$ **do**
    Update $k \leftarrow k + 1$;
    Initialize $\mathfrak{C}^{(k)} \leftarrow \{\}$; $g^{(k)} \leftarrow 0$;
    for $l_1 \in 1 : g^{(k-1)}$ **do**
      if $|\mathcal{C}_{l_1}^{(k-1)}| \leq t_c$ or $g^{(k)} \geq t_n$ **then**
        $\mathfrak{C}^{(k)} \leftarrow \mathfrak{C}^{(k)} \cup \mathcal{C}_{l_1}^{(k-1)}$; $g^{(k)} \leftarrow g^{(k)} + 1$;
      else
        $R \overset{\text{i.i.d.}}{\sim} R_\Theta$
        for $d \in R$, $c \in \text{gaps}(\{X_{id} | 1 \leq i \leq n-1\})$[1] **do**
          $v_{n,\mathbf{Q}}^{\text{DART}}((d, c)) \leftarrow$ Equation (5)
        $(d^*, c^*) \leftarrow \arg\min_{(d,c)} v_{n,\mathbf{Q}}^{\text{DART}}((d, c))$
        $\mathcal{C}_{l_1^{(1)}}^{(k)}$ and $\mathcal{C}_{l_1^{(2)}}^{(k)} \leftarrow$ Equation (9) with (d^*, c^*)
        $\mathfrak{C}^{(k)} \leftarrow \mathfrak{C}^{(k)} \cup \mathcal{C}_{l_1^{(1)}}^{(k)} \cup \mathcal{C}_{l_1^{(2)}}^{(k)}$;
        $g^{(k)} \leftarrow g^{(k)} + 2$;
  Representatives $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{g^{(k)}})^\top \leftarrow$ Equation (6) with $\mathbf{Z}_{n \times g^{(k)}} = \left(\mathbb{I}(X_i \in \mathcal{C}_l^{(k)})\right)$;
  Output $m_n(\mathbf{x}_0; \Theta) = \sum_{l=1}^{g^{(k)}} \hat{\beta}_l \mathbb{I}(\mathbf{x}_0 \in \mathcal{C}_l^{(k)})$;

[1] For any set of real numbers $S = \{r_1 < \cdots < r_s\}$, $\text{gaps}(A) = \{(r_i + r_{i+1})/2 : i = 1, \ldots, s-1\}$.

Note that, if the true covariance matrix $\mathbf{\Sigma}_0$ is unknown, as in practice, $\mathbf{Q}$ will be $\mathbf{\Sigma}^{-1}$ where $\mathbf{\Sigma}$ is some working covariance matrix (estimate and/or computational approximation of $\mathbf{\Sigma}_0$). We discuss estimation of $\mathbf{\Sigma}_0$ and choice of $\mathbf{Q}$ in Section 2.7. The algorithm works with any choice of the working precision matrix $\mathbf{Q}$. For example, with $\mathbf{Q} = \mathbf{I}$, the algorithm is identical to the usual regression-tree used in RF.

2.5. RF-GLS

We now focus on growing a RF from n_{tree} number of GLS-style trees. In RF, each tree is built using a resample or subsample of the data $\mathcal{D}_{n,t} = (\mathbf{Y}_t, \mathbf{X}_t)$ where $\mathbf{Y}_t$ is the resampled (or subsampled) vector of the responses and $\mathbf{X}_t$ is the design matrix using the corresponding rows. For our GLS-style approach, naive emulation of this is not recommended. To elucidate, if one resorts to resampling, under dependent settings, this would mean resampling correlated data units (Y_i, X_i) thereby violating the principle of bootstrap. Also, it is unclear what the working covariance matrix would be between the resampled points. If subsampling is used, this is avoided, as one can use the submatrix $\mathbf{\Sigma}_t$ corresponding to the subsample. However, each tree will use a different subsample and hence a different $\mathbf{\Sigma}_t$. Inverting covariance matrices of dimension $O(n)$ require $O(n^3)$ operations, and this approach would require inverting n_{tree} such matrices, thereby substantially increasing the computation.

Interestingly, the GLS-loss itself offers a synergistic solution to resampling of dependent data. To motivate our approach, we once again revisit RF, and note that the CART-split criterion (4) for a re(sub)sample can be expressed using the squared error loss $\|\mathbf{P}_t\mathbf{Y} - \mathbf{P}_t\mathbf{Z}\boldsymbol{\beta}\|_2^2$, where $\mathbf{P}_t$ is the selection matrix for the resample. Now GLS loss with $\mathbf{Y}$ and $\mathbf{Z}\boldsymbol{\beta}$ coincides with an OLS loss with $\tilde{\mathbf{Y}} = \mathbf{\Sigma}^{-1/2}\mathbf{Y}$, $\tilde{\mathbf{Z}} = \mathbf{\Sigma}^{-1/2}\mathbf{Z}$. Hence, the immediate extension for the resample (subsample) in our setup would be using the loss

$$\begin{aligned}\|\mathbf{P}_t\tilde{\mathbf{Y}} - \mathbf{P}_t\tilde{\mathbf{Z}}\boldsymbol{\beta}\|_2^2 &= \|\mathbf{P}_t\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{Y} - \mathbf{P}_t\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{Z}\boldsymbol{\beta}\|_2^2 \\ &= (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})^\top\mathbf{\Sigma}^{-\top/2}\mathbf{P}_t^\top\mathbf{P}_t\mathbf{\Sigma}^{-1/2}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}).\end{aligned} \tag{10}$$

Thus, to use a GLS-loss in RF, we essentially resample the contrasts $\tilde{\mathbf{Y}}$ instead of the outcomes $\mathbf{Y}$. This principle of contrast resampling has been used in parametric bootstrapping of spatial data (Pardo-Igúzquiza and Olea 2012; Saha and Datta 2018a). In our algorithm the resampling amounts to simply replacing the $\mathbf{Q}$ in the DART-split criterion (5) and node representative calculation (6) with $\mathbf{Q}_t = \mathbf{\Sigma}^{-\top/2}\mathbf{P}_t^\top\mathbf{P}_t\mathbf{\Sigma}^{-1/2}$. Computationally, our approach has the advantage of only requiring a one-time evaluation of the Cholesky factor $\mathbf{\Sigma}^{-1/2}$ that is used in all trees. As in RF, subsequent to growing n_{tree} trees corresponding to each different resample we take the average to get the forest estimate. This completes the specification of a novel GLS-style RF for dependent data. We refer to the algorithm as *RF-GLS* and summarize it in Algorithm 2. It is clear when $\mathbf{Q} = \mathbf{I}$, the node-split criterion, the node representatives, and the resampling step all become identical to RF. Hence, RF is simply a sub-case of RF-GLS with an identity working correlation matrix.

Algorithm 2 RF-GLS

Input: Data $\mathcal{D}_n = (Y_1, X_1, \ldots, Y_n, X_n)$, working correlation matrix $\mathbf{\Sigma}$, number of trees n_{tree}, stopping rules for the trees: t_n (maximum number of nodes) and t_c (minimum number of members per node), number of features considered for each split M_{try}, randomness $R_\Theta^{(1)}$ (some probability distribution to choose a resample of size n from $\{1, \ldots, n\}$), randomness $R_\Theta^{(2)}$ (some probability distribution to choose a subsample of size M_{try} from $\{1, \ldots, D\}$), output point $\mathbf{x}_0$.

Output: Estimate $m_n(\mathbf{x}_0; \Theta)$ of the mean function m at $\mathbf{x}_0$.

procedure
  Calculate $\mathbf{\Sigma}^{-1/2}$;
  Initialize $t = 1$;
  for $t = 1 : n_{\text{tree}}$ **do**
    Generate $R_t \overset{\text{i.i.d.}}{\sim} R_\Theta^{(1)}$
    $\mathbf{P}_t \leftarrow (I(i = R_t[j]))$
    $\mathbf{Q}_t = \mathbf{\Sigma}^{-\top/2}\mathbf{P}_t^\top\mathbf{P}_t\mathbf{\Sigma}^{-1/2}$
    $m_n(\mathbf{x}_0; \Theta_t) \leftarrow$ Algorithm 1 with $\mathcal{D}_n, \mathbf{Q}_t, t_n, t_c, M_{try}, R_\Theta^{(2)}, \mathbf{x}_0$;
  Output $\hat{m}_n(\mathbf{x}_0) = \frac{1}{n_{\text{tree}}}\sum_{t=1}^{n_{\text{tree}}} m_n(\mathbf{x}_0; \Theta_t)$

2.6. Kriging Using RF-GLS With GPs

Subsequent to estimating the regression function $m(\mathbf{x})$ using RF-GLS in the spatial nonlinear mixed model (1), we can seamlessly perform traditional spatial tasks like predictions (kriging) and recovery of the latent spatial random surface $w(\ell)$. This is because RF-GLS only estimates the mean part, and the covariance is still being modeled using a GP. This facilitates easy formulation of the predictive or latent distribution conditional on the data as standard conditional normal distributions. If $\mathcal{L}$ denotes the training data locations, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, and $\mathbf{m} = (\widehat{m}(X_1), \ldots, \widehat{m}(X_n))^\top$ where $\widehat{m}$ is the estimate of m from RF-GLS, then prediction at a new location ℓ_{new} with covariate $\mathbf{x}_{\text{new}}$ will simply be given by the kriging estimate

$$\widehat{y}_{\text{new}}(\mathbf{x}_{\text{new}}, \ell_{\text{new}}) = \widehat{m}(\mathbf{x}_{\text{new}}) + \mathbf{v}^\top\mathbf{\Sigma}^{-1}(\mathbf{Y} - \widehat{\mathbf{m}}) \tag{11}$$

where $\mathbf{v}^\top = \text{cov}(\epsilon(\ell_{\text{new}}), \epsilon(\mathcal{L}))$, $\mathbf{\Sigma} = \text{cov}(\epsilon(\mathcal{L}), \epsilon(\mathcal{L}))$. The prediction equation (11) possess the advantage of nonparametrically estimating the mean function of the covariates while retaining the spatial structure encoded in the GP covariance function which adheres to the philosophy of *first law of geography*, that is, proximal things are more correlated than distant ones. The prediction framework is completely agnostic to the choice of the covariance function and can work with nonstationary or multi-resolutional covariance functions if deemed appropriate. One can also obtain estimates of the latent surface using the conditional distributions $w(\ell) \mid \mathbf{Y}, \mathbf{X}$ akin to the spatial linear model.

Prediction equations of the form of Equation (11) has been used in applications of RF to spatial settings and has been termed as random forest *residual kriging* (RF-RK) (Viscarra Rossel, Webster, and Kidd 2014; Fayad et al. 2016). The estimate $\widehat{m}$ in these applications come from a naive application of RF without accounting for the dependence. Our empirical studies in Section 4 will demonstrate how residual kriging using RF-GLS improves over use of RF in dependent settings.

RF-GLS also has several advantages over the RFsp of Hengl et al. (2018), which does not use GP but include pairwise distances between a location and all other locations, as additional covariates. For n locations, this adds n covariates. RF-GLS avoids such unnecessary escalation of the problem to high-dimensional settings. Direct use of the GP and the mixed-model framework helps model the spatial structure parsimoniously via the covariance function parameters. Our simulation studies in Section 4 demonstrates the improved prediction performance of RF-GLS over RFsp. Most importantly, RF-GLS separates out the contribution of the covariates and the spatial component, thereby allowing estimation of the regression function m. Estimate of m is not available from the RFsp of Hengl et al. (2018) which only offers a prediction given the covariates and the location.

2.7. Practical Considerations

The development of the RF-GLS method has been presented using a working covariance matrix $\boldsymbol{\Sigma}$. In practice, the true correlation $\boldsymbol{\Sigma}_0$ will not be known and needs to be estimated assuming a parametric form $\boldsymbol{\Sigma}$ based on the covariance function used. In Section 3 we present the result that both RF and RF-GLS are consistent under dependent errors (akin to both OLS and GLS being consistent for linear models). Hence, for data analysis, we recommend running the first pass of RF on the data to get a preliminary estimate of m, use it to obtain the residuals from which the parameters of $\boldsymbol{\Sigma}$ can be estimated using a maximum likelihood approach. This once again parallels the practice in linear models where the *oracle GLS* assuming knowledge of the true covariance matrix is replaced by a *feasible GLS* where the covariance matrix is estimated based on residuals from an initial OLS estimate.

The second consideration concerns computational scalability of the approach. RF-GLS requires computing the Cholesky factor $\boldsymbol{\Sigma}^{-1/2}$. It is clear from Algorithm 2, that this is only a one-time cost, unlike the possible alternate approach discussed in Section 2.5 where subsampling is conducted before decorrelation, which would lead to computing a different Cholesky factor for each tree. However, spatial covariance and precision matrices arising from GP are dense and for large n, evaluating $\boldsymbol{\Sigma}^{-1}$ even once still incurs the computational cost of $O(n^3)$ and storage cost of $O(n^2)$ both of which are taxing on typical personal computing resources.

Over the last decade, the inventory of approximation techniques attacking the computational weakness of GP has grown and become increasingly sophisticated (see Heaton et al. 2019, for a review). NNGP (Datta et al. 2016a; Finley et al. 2019; Datta et al. 2016b,c) has emerged as one of the leading candidates. Centered on the principle that a few nearby locations are enough to capture the spatial dependence at a given location (Vecchia 1988), NNGP replaces the dense graph among spatial locations with the nearest neighbor graphical model. This was shown to directly yield a sparse Cholesky factor $\widetilde{\boldsymbol{\Sigma}}^{-1/2}$ that offers an excellent approximation to the original dense $\boldsymbol{\Sigma}^{-1/2}$ (Datta et al. 2016a). Software packages implementing NNGP are also publicly available (Finley, Datta, and Banerjee 2021; Saha and Datta 2018b). Hence, for very large data, we recommend using this NNGP sparse Cholesky factor $\widetilde{\boldsymbol{\Sigma}}^{-1/2}$ instead of $\boldsymbol{\Sigma}^{-1/2}$. This will reduce both the computation and storage cost from cubic to linear in sample size. We will show in Proposition 3.1, that using NNGP for RF-GLS ensures a consistent estimate of m even when the true data generation is from a full Matérn GP.

3. Consistency

In this section, we present the main theoretical result on consistency of RF-GLS for a very general class of dependent error processes. The outline of the proof highlighting the new theoretical challenges addressed are presented in Section 5 along with some general results of independent importance. The formal proofs are provided in the supplementary materials.

3.1. Assumptions

We first discuss Assumptions 1 and 2 and make additional assumptions required for the proof of consistency. In Assumption 1, we focus on absolutely regular or β-mixing processes, since this class of stochastic processes is rich enough to accommodate many commonly used dependent error processes like ARMA (Mokkadem 1988), GARCH (Carrasco and Chen 2002), certain Markov processes (Doukhan 2012) and GPs with Matérn covariance family. At the same time, uniform law of large numbers (ULLN) from independent processes can be extended to this dependent process under moderate restriction on the class of functions under consideration. No additional assumption is required on the decay rate of the β-mixing coefficients (which are often hard to check).

Assumption 2 requires the Cholesky factor of the precision matrix to be sparse and regular. Such structured Cholesky factors routinely appear in time series analysis for AR(p) process. For spatial data, exponential covariance family on a 1-dimensional grid satisfies this. Other covariances like the Matérn family (except the exponential covariance) do not generally satisfy this assumption. However, NNGP covariance matrices satisfy this and are now commonly used as an excellent approximation to the full GP covariance matrices (Datta et al. 2016a). Since this assumption is on the working covariance matrix and not on the true covariance of the process, we can always use an approximate working covariance matrix like ones arising from NNGP to satisfy this. We discuss these examples in Section 3.3. Under Assumption 2, for any two vectors $\mathbf{x}$ and $\mathbf{y}$, defining $x_i = y_i = 0$ for $i \leq 0$, we have

$$\mathbf{x}^\top \mathbf{Q}\mathbf{y} = \alpha \sum_i x_i y_i + \sum_{j \neq j'=0}^{q} \rho_j \rho_{j'} \sum_i x_{i-j} y_{i-j'} + \sum_{i \in \tilde{\mathcal{A}}_1} \sum_{i' \in \tilde{\mathcal{A}}_2} \tilde{\gamma}_{i,i'} x_i y_{i'}, \tag{12}$$

where $\alpha = \|\rho\|_2^2$, $\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_2 \subset \{1, 2, \ldots, n\}$ with $|\tilde{\mathcal{A}}_1|, |\tilde{\mathcal{A}}_2| \leq 2q$, $\tilde{\gamma}_{i,i'}$'s are fixed (independent of n) functions of $\mathbf{L}$ and ρ. The expression of the quadratic form in Equation (12) makes it evident that $\lambda_{\max}(\mathbf{Q})$ is bounded as $n \to \infty$. As the third term is a sum of fixed (at most $4q^2$) number of terms, it is $O(1)$ as long as $\mathbf{x}$ and $\mathbf{y}$ are bounded.

Assumption 3 (Diagonal dominance of the working precision matrix). $\mathbf{Q}$ is diagonally dominant satisfying $\mathbf{Q}_{ii} - \sum_{j \neq i} |\mathbf{Q}_{ij}| > \xi$ for all i, for some constant $\xi > 0$.

Diagonal dominance implies $\lambda_{\min}(\mathbf{Q})$ is bounded away from zero as $n \to \infty$ which is needed to ensure stability of the GLS estimate. We will discuss in Section 3.3 how working correlation matrices from popular time series and spatial processes with regular design satisfy this Assumption. Note that under Assumption 2, checking that the first $(q+1)$ rows of $\mathbf{Q}$ are diagonally dominant is enough to verify Assumption 3.

Assumption 4 (Tail behavior of the error distribution).

1. $\exists\{\zeta_n\}_{n\geq 1}$ such that
$$\zeta_n \to \infty,\ \frac{t_n(\log n)\zeta_n^4}{n} \to 0,\ \text{and}\ \mathbb{E}\left[\left(\max_i \epsilon_i^2\right)\mathbb{I}\left(\max_i \epsilon_i^2 > \zeta_n^2\right)\right] \to 0 \text{ as } n \to \infty.$$
2. $\exists$ constant $C_\pi > 0$ and $n_0 \in \mathbb{N}^*$ such that with probability $1-\pi, \forall n > n_0$,
$$\max_i |\epsilon_i| \leqslant C_\pi \sqrt{\log n}.$$
3. Let $\mathcal{I}_n \subseteq \{1, 2, \cdots, n\}$ with $|\mathcal{I}_n| = a_n$ and $a_n \to \infty$ as $n \to \infty$. Then $\frac{1}{a_n}|\sum_{i\in\mathcal{I}_n}\epsilon_i| > \delta$ with probability at most $C\exp(-ca_n)$ and $\frac{1}{n}|\sum_i \epsilon_i^2| > {\sigma^2}_0$ with probability at most $C\exp(-cn)$ for any $\delta > 0$, and some constants $c, C, {\sigma^2}_0 > 0$.

We show for Gaussian errors, ζ_n needs to be $O(\log n)^2$ which makes the scaling condition in Assumption 4(a) as $t_n(\log n)^9/n \to 0$. This is the same scaling used in Scornet et al. (2015) for Gaussian errors and using the entire sample. In general, the choice of ζ_n will be dependent on the error distribution. Assumption 4(a), (b) and (c) will all be satisfied by sub-Gaussian errors.

Assumption 5 (Additive model). The true mean function $m(\mathbf{x})$ is additive on the coordinates x_d of $\mathbf{x}$, that is, $m(\mathbf{x}) = \sum_{d=1}^{D} m_d(x_d)$, where each component m_d is continuous.

As demonstrated in Scornet et al. (2015), additive models provide a rich enough environment to address the asymptotic properties of nonparametric methods like RF sans the additional complexities in controlling asymptotic variation of m in leaf nodes. Since RF is invariant to monotone transformations of covariates (Friedman, Hastie, and Tibshirani 2001; Friedman 2006), without loss of generality, the covariates can be distribution function transformed to be Unif$[0,1]$ distributed. Hence we assume that the components (functions) m_d are supported on $[0,1]$, implying m is uniformly bounded by some constant M_0.

3.2. Main Result

For the tth tree, the predicted value from our method at a new point $\mathbf{x}_0$ in covariate space is denoted by $m_n(\mathbf{x}_0; \Theta_t, \mathbf{\Sigma}, \mathcal{D}_n)$ where $\mathcal{D}_n = \{(\mathbf{x}_i, y_i) \,|\, i = 1, \ldots, n\}$ denote the data. Note that the ith data unit corresponds to location ℓ_i and that the covariance matrix $\mathbf{\Sigma}$ is based on the covariance function evaluated at pairs of locations ℓ_i and ℓ_j. But unless otherwise needed we suppress the locations ℓ_i and simply use the subscript i.

Θ_t indicates the randomness associated with each tree. In practice, Θ_t will include both the resampling of data-points used in each tree as well as the choice of random splitting variable for iterative splitting in the tree. For tractability, Scornet et al. (2015) considered subsampling instead of resampling for the theoretical study. In our theoretical study, for analytical tractability of the GLS weights, we consider trees that use the entire set of samples and the randomness Θ_t in each tree is only used to choose the candidate set of features for each split. The randomness for each tree are iid, that is, $\Theta_t \overset{\text{iid}}{\sim} \Theta, \Theta \perp \mathcal{D}_n, \forall t \in \{1, \cdots, n_{\text{tree}}\}$. The finite RF-GLS estimate $\widehat{m}_{n,n_{\text{tree}}}(\mathbf{x}_0; \Theta_1, \cdots, \Theta_{n_{\text{tree}}}, \mathbf{\Sigma}, \mathcal{D}_n)$ that will be used in practice is given by the sample average of the individual tree estimates. Conceptually, n_{tree} can be arbitrarily large; hence following Scornet et al. (2015), we focus on "infinite" RF-GLS estimate given by $\bar{m}_n(\mathbf{x}_0; \mathbf{\Sigma}, \mathcal{D}_n) = \mathbb{E}_\Theta m_n(\mathbf{x}_0; \Theta, \mathbf{\Sigma}, \mathcal{D}_n)$ where the expectation w.r.t Θ is conditional on $\mathcal{D}_n$. For notational convenience, we hide the dependence of $m_n, \widehat{m}_{n,n_{\text{tree}}}, \bar{m}_n$ on $\mathbf{\Sigma}$ and $\mathcal{D}_n$ throughout the rest of this article. Our main result on $\mathbb{L}_2$-consistency is stated next, the proof is deferred to Section 5.1.

Theorem 3.1. Under Assumptions 1-5 and if for some $\delta > 0$, $\lim_{n\to\infty} \mathbb{E}\frac{1}{n}\sum_i |m_n(X_i)|^{2+\delta} < \infty$, then RF-GLS is $\mathbb{L}_2$-consistent, that is, $\lim_{n\to\infty} \mathbb{E}\int (\bar{m}_n(X) - m(X))^2\, dX = 0$,.

The uniformly bounded $(2+\delta)$th moment assumption in Theorem 3.1 is needed to generalize uniform laws of large number bounding the GLS estimation error from the iid setting to the dependent setting. We discuss this in details in Section 5.3. The following corollaries discuss three specific cases where this assumption is met. Their proofs are deferred to Section S2.3.

Corollary 3.1. Under Assumptions 1-5, RF-GLS is $\mathbb{L}_2$ consistent if either:

1. Case 1: The errors are bounded.
2. Case 2: The working precision matrix $\mathbf{Q}$ satisfies $\min_i \mathbf{Q}_{ii} > \sqrt{2}\max_i \sum_{j\neq i} |\mathbf{Q}_{ij}|$.

For bounded errors (part (a)), the $(2+\delta)$th moment-bound of Theorem 3.1 is immediately satisfied, and hence consistency can be established without further assumptions. For unbounded errors, a stronger form of diagonal dominance condition is needed in Corollary 3.1 Part (b). This is used to control the $(2+\delta)$th moment of the data weights arising from the gram-matrix $(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z})^{-1}$ which in turn ensures the moment-bound. We discuss examples and specific parameter choices ensuring this in Section 3.3. Also note that, the assumption of diagonal dominance is not on the true correlation matrix of the error process and hence is not a restriction on the data-generation mechanism, but rather on the working correlation matrix which is chosen by the user. One can always use parameters in the working correlation matrix that satisfies this (although enforcing this is not needed in practice).

RF is RF-GLS with $\mathbf{Q} = \mathbf{I}$. Hence the assumption of Corollary 3.1 part (b) is trivially satisfied. This proves consistency of RF under β-mixing dependence.

Corollary 3.2. Under Assumptions 1, 4, and 5, RF (Breiman 2001) is $\mathbb{L}_2$-consistent.

To our knowledge, Corollary 3.2 is the first result on consistency of RF under a dependent (β-mixing) error process. Since RF is simply RF-GLS with the working correlation matrix $\mathbf{\Sigma} = \mathbf{I}$, Assumptions 2 and 3 are automatically satisfied, and hence we only need the Assumptions of β-mixing process, tail bounds, and additive model. The consistency result is analogous to the ordinary least-square estimate being consistent even for correlated errors. Besides its own importance, Corollary 3.2 also heuristically justifies the first step used in practical implementation of RF-GLS. The parameters in the working correlation matrix is unknown, and as highlighted in Section 2.7, we use the RF to get a preliminary estimate of m, estimate the spatial parameters using the residuals, and use these estimated parameters in the working correlation matrix for RF-GLS. This is again, analogous to feasible GLS which estimates the working correlation matrix using residuals based on OLS. Corollary 3.2 guarantees that the initial estimator used to obtain the residuals is consistent.

3.3. Examples

In this section, we give examples of two popular dependent error processes under which a consistent estimate of m can be obtained using RF-GLS.

3.3.1. Spatial Matérn GPs

Our main example focuses on the spatial nonlinear mixed model using GPs as described in Section 2.6. While many candidates exist for the covariance function of GP, the class of Matérn covariances enjoy hegemonic popularity in the spatial literature owing to its remarkable property of characterizing the smoothness of the spatial surface $\epsilon(\ell)$ (Stein 2012). The stationary (isotropic) Matérn covariance function is specified by

$$C(\ell_i, \ell_j \,|\, \boldsymbol{\phi}) = C(\|\ell_i - \ell_j\|_2) = \sigma^2 \frac{2^{1-\nu}\left(\sqrt{2}\phi\|\ell_i - \ell_j\|_2\right)^{\nu}}{\Gamma(\nu)} \mathcal{K}_{\nu}\left(\sqrt{2}\phi\|\ell_i - \ell_j\|_2\right), \quad (13)$$

where $\boldsymbol{\phi} = (\sigma^2, \phi, \nu)^\top$ and $\mathcal{K}_\nu$ is the modified Bessel function of second kind.

We consider a Matérn process sampled on one-dimensional regular lattice. This regular design is considered both for tractability of the Matérn GP likelihood but also for ensuring stationarity of the process in the sense required in Theorem 3.1 as for irregular spaced data $\text{cov}(\epsilon_1, \epsilon_2) \neq \text{cov}(\epsilon_2, \epsilon_3)$ whenever $\|\ell_1 - \ell_2\|_2 \neq \|\ell_2 - \ell_3\|_2$. Such assumptions on the dimensionality and/or regularity of design has been widely used for theoretical studies of spatial processes (Du et al. 2009; Stein et al. 2002). By keeping the gap in the lattice fixed, we are also essentially using increasing-domain asymptotics, as parameters are generally not identifiable in fixed domain asymptotics for Matérn GPs (Zhang 2004).

The error process arising from the marginalization of (1) is the sum of a Matérn process and a nugget (random error) process. We consider half-integer $\nu \in 1.5, 2.5, \ldots$. This class of processes are popularly studied and used owing to their convenient state-space representation (Hartikainen and Särkkä 2010) which in turn leads to efficient computation of these Matérn GP likelihoods. The state-space representation of half-integer Matérn GP is equivalent to that of a stable AR(q_0) process on the continuous one-dimensional domain with $q_0 = \nu + 1/2$. However, unlike an AR(q_0) time series, the Matérn GP when sampled on the discrete integer lattice is no longer an AR(q_0) process. Consequently, unlike covariance matrices from AR processes, covariance matrices $\mathbf{\Sigma}$ generated from Matérn GP (expect for exponential GP), do not satisfy the sparsity and regularity of the working correlation matrix of Assumption 2.

Instead, we consider the working correlation $\mathbf{\Sigma}$ to come from the nearest neighbor Gaussian process (NNGP) (Datta et al. 2016a) based on the Matérn covariance. As discussed in Section 2.7, NNGP covariance matrices are one of the most successful surrogates for full GP covariances for large spatial data, reducing likelihood computations from $O(n^3)$ to $O(n)$. What is important for the theoretical study is that an NNGP is constructed by sequentially specifying the conditional distributions as $\epsilon_i \,|\, \epsilon_{1:i-1} \sim \epsilon_i \,|\, \epsilon_{N_q(i)}$ where $N_q(i) \subset \{1, \ldots, i-1\}$ is the set of q-nearest neighbors of ℓ_i among $\ell_1, \ldots, \ell_{i-1}$. When the locations are the integer grid, $N_q(i)$ becomes $\{i-1, \ldots, i-q\}$, and the NNGP construction is akin to an AR(q) process. Consequently, the Cholesky factor $\mathbf{\Sigma}^{-1/2}$ from NNGP on an integer lattice satisfies Assumption 2 with $\rho = (1, -\mathbf{c}^\top\mathbf{C}^{-1})^\top/\sqrt{1 - \mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c}}$ and $\mathbf{L}$ such that $\mathbf{L}^\top\mathbf{L} = \mathbf{C}^{-1}$ where $\mathbf{C} = \text{cov}(\epsilon_{1:q})$, $\mathbf{c} = \text{cov}(\epsilon_{1:q}, \epsilon_{q+1})$ (Finley et al. 2019). This ensures the following consistency result of RF-GLS fitted with NNGP for data generated using Matérn GP. The proof is in Section S2.6.

Proposition 3.1. Consider a spatial process $y(\ell_i) = m(X_i) + \epsilon(\ell_i)$ from (1) where m is an additive model as specified in Assumption 5, $\epsilon(\ell_i) = w(\ell_i) + \epsilon^*(\ell_i)$ where $\epsilon^*(\ell)$ denote iid $N(0, \tau_0^2)$ noise, and $w(\ell)$ be a Matérn GP, sampled on the integer lattice, with parameters $\boldsymbol{\phi}_0 = (\sigma^2{}_0, \phi_0, \nu_0)^\top$, ν_0 being a half-integer. Let $\mathbf{\Sigma}$ denote a covariance matrix from the nearest neighbor Gaussian process (NNGP) derived from a Matérn covariance with parameters $\boldsymbol{\phi} = (\sigma^2, \phi, \nu)^\top$ and τ^2. Then there exists some $K > 0$ such if $\phi > K$, then RF-GLS using $\mathbf{\Sigma}$ yields an $\mathbb{L}_2$ consistent estimate of m.

One observation is central to the proof. The half-integer Matérn GP, which is an AR(q_0) process in the continuous domain, when sampled on a discrete lattice becomes an ARMA process ((Ihara 1993) Theorem 2.7.1). This will establish absolutely regular mixing of these Matérn processes using the result of Mokkadem (1988) on ARMA processes, and subsequently consistency of RF-GLS by Theorem 3.1.

3.3.2. Autoregressive Time Series

Our main focus in this article, is estimation of nonlinear regression function in the spatial mixed model (1). However, the scope of our RF-GLS algorithm is much broader. It can be used for functional estimation in the general nonlinear regression model $Y_i = m(X_i) + \epsilon_i$ where ϵ_i is a dependent stochastic process with valid second moment. The method only relies on knowledge of an estimate of the residual covariance matrix $\mathbf{\Sigma} = \text{cov}(\boldsymbol{\epsilon})$. The

general consistency result (Theorem 3.1) is also not specific to the spatial GP setting, and only relies on general assumptions on the nature of the dependence, tail bounds of the error, and structure of the working correlation matrix. In this section, we demonstrate that RF-GLS can be used for consistent function estimation for time series data, that is, where ϵ_i models the serial (temporal) correlation. In particular, we discuss consistency of RF-GLS for Autoregressive (AR) error processes, one of the mainstays of time series studies. An AR(q) model can be written as follows:

$$\epsilon_i = a_1\epsilon_{i-1} + a_2\epsilon_{i-2} + \cdots + a_q\epsilon_{i-q} + \eta_i, \qquad (14)$$

where η_i is a realization of a white-noise process at time i. AR processes are β-mixing (Mokkadem 1988), they also produce a banded Cholesky factor of the precision matrix as required in Assumption 2. Hence, we have the following assertion. Its proof is in Section S2.6.

Proposition 3.2. Consider a time series $Y_i = m(X_i) + \epsilon_i$ where i is the time, m satisfies Assumption 5, ϵ_i denote a sub-Gaussian stable $AR(q_0)$ process. Let $\mathbf{\Sigma}$ denote a working correlation matrix from a stationary $AR(q)$ process. Then RF-GLS using $\mathbf{\Sigma}$ produces an $\mathbb{L}_2$ consistent estimate of m if

1. $q = 1$ and the working autocorrelation parameter ρ used in $\mathbf{\Sigma}$ satisfies $|\rho| < 1$ (for bounded errors) or $|\rho| < 1/(2\sqrt{2})$ (for unbounded errors).
2. $q > 1$ and the AR(q) working precision matrix $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ satisfies Assumption 3 (for bounded error) or $\min_i \mathbf{Q}_{ii} > \sqrt{2}\max_i \sum_{j\neq i} |\mathbf{Q}_{ij}|$ (for unbounded error).

We separate the results for $q = 1$ and $q \geq 2$ since unlike AR(1), for general AR(q) it is challenging to derive closed form expressions of the constraints on the parameter space needed to satisfy Assumption 3 or the stronger diagonal dominance condition in Proposition 3.2 part 2 (for unbounded errors). However, verifying these conditions for a given AR precision matrix $\mathbf{Q}$ is straightforward due to stationarity and banded structure. One only needs to check the first $q + 1$ rows of $\mathbf{Q}$ irrespective of the sample size n. The necessary condition for row $q + 1$ is

$$\|\rho\|^2 > 2\kappa \sum_{j=1}^{q} |\sum_{j'=j}^{q} \rho_{j'}\rho_{j'-j}|, \qquad (15)$$

where κ equals 1 for bounded errors and equals $\sqrt{2}$ for unbounded errors. Additional checks are only needed for the first q rows of $\mathbf{Q}$.

In practical implementation of AR processes, the order of the autoregression is often chosen based on analysis of the autocorrelation function of the residuals and may not equal the true order of the autoregression. Proposition 3.2 accommodates this scenario by not restricting the working autoregressive covariance to be of the same order or have the same coefficients as the ones generating the data.

4. Illustrations

We conduct simulation experiments to demonstrate the advantages of the RF-GLS over competing methods for both estimation and prediction in finite samples. We simulate data from the spatial nonlinear mixed model of Equation (1). We consider the following choices for the true mean function $m_1(x) = 10\sin(\pi x)$ and $m_2(\mathbf{x}) = (10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5)/6$ (Friedman function, Friedman 1991). $w(\ell)$ is an exponential GP on a two-dimensional spatial domain with 27 different combinations of covariance parameters spatial variance σ^2, spatial decay ϕ and error variance τ^2 as a % of the spatial variance. For each setting, we perform simulations for 100 replicate datasets. To evaluate prediction performance, we keep 10% hold-out data. Details of the parameter choices and hold-out design are in Section S1.1.

For evaluating estimation performance, we consider RF (which does not use spatial information), RF-GLS (Oracle) using true covariance parameter values, and RF-GLS which obtains an estimate of the covariance parameters from RF residuals using the BRISC package (Saha and Datta 2018b) and uses them in RF-GLS. The estimation performance of the methods are evaluated based on a discrete Mean Integrated Square Errors (MISE) for m, evaluated over uniformly generated data points from the covariate space that provides a good coverage of the entire space. MISE for the estimated function $\hat{m}$ is given as:

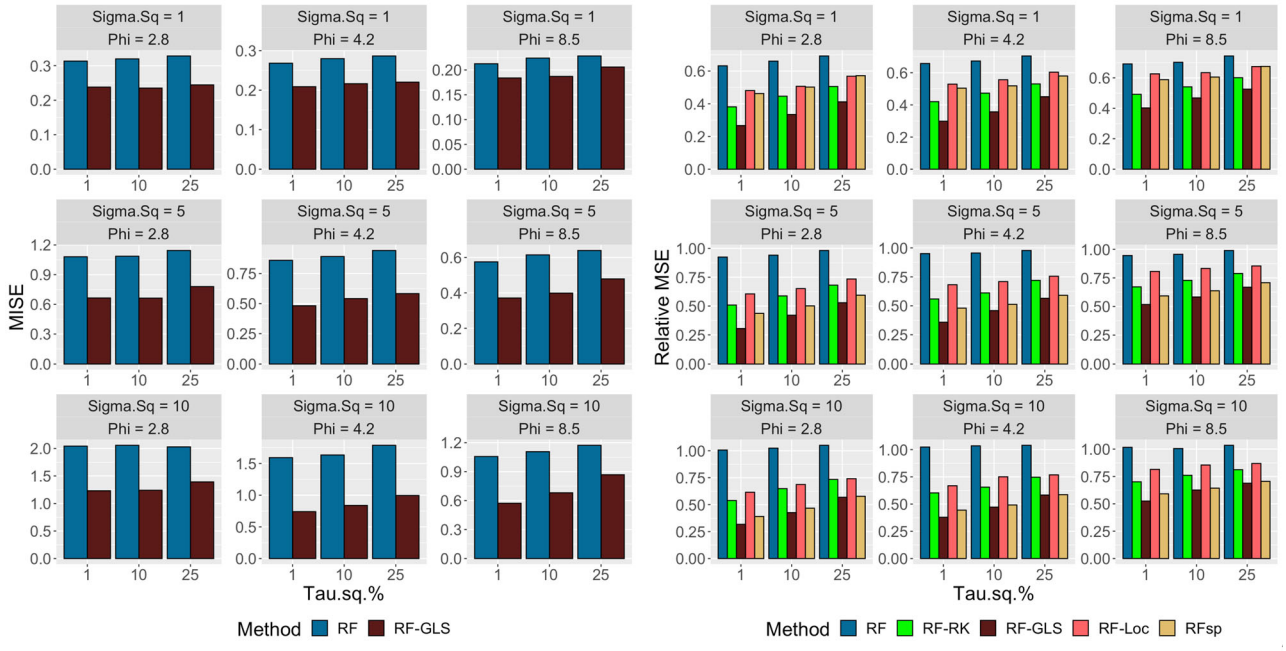
$$\text{MISE} = \int (m(\mathbf{x}) - \hat{m}(\mathbf{x}))^2 d\mathbf{x} \approx \frac{1}{n_0}\sum_{i=1}^{n_0}(m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i))^2$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_{n_0}$ are a dense set of points in the covariate domain (See Section S1.1 for details).

We observe that the RF-GLS performs at par with RF-GLS (Oracle) which assumes knowledge of the true spatial parameters (supplementary material, Section S1.2). As in reality, we won't have knowledge of the true parameters, for the rest of the article we will be comparing the competing methods only with RF-GLS, where model parameters are unknown and estimated from RF residuals. We present the estimation and prediction results for $m = m_2$. The results for $m = m_1$ are similar, and are provided in the supplementary materials, Section S1.3.

The median MISE over 100 simulations for all the 27 setups are shown in Figure 2(a). RF-GLS outperforms RF across all the scenarios demonstrating that exploiting the spatial information substantially improves estimation performance over the vanilla RF. Additionally, we also notice that as the strength of the spatial variation increases (σ^2 goes from 1 to 10), the ratio of MISE of RF and RF-GLS increases indicating a greater gain in terms of MISE.

We consider five candidates for evaluating prediction performance: RF, RF-RK (Fox, Ver Hoef, and Olsen 2020), RF-GLS, RF-Loc (the 2-D spatial locations of the data are used as additional covariates in RF to account for the spatial structure), and RFsp (Hengl et al. 2018). Among these, RF is the only one that does not use spatial information and only accounts for the mean. For RFsp, we used unordered pairwise Euclidean distances as additional covariates in RF to account for the spatial structure in the data. The prediction performance are evaluated based on the relative mean squared error (MSE) for the test data. MSE measures the average squared difference between the estimated values and the actual value of the response. Relative MSE helps compare MSE across different simulation setups,



(a) Estimation performance (b) Prediction performance

Figure 2. Comparison between competing methods on (a) estimation and (b) spatial prediction when the mean function is $m = m_2$.

while standardizing by the variance of the response.

$$\text{MSE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2;$$

$$\text{Relative MSE} = \frac{\text{MSE}}{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \bar{y}_{\text{test}})^2}$$

Figure 2(b) shows the median relative MSE over 100 simulations for all the 27 setups. Expectedly, all methods performed better than RF, as unlike the other methods, RF does not use any spatial information. RF-GLS performs best or at par with the best across all settings.

RF-Loc and RFsp belong to the class of approaches that add extra spatial covariates to RF. RF-Loc always performed worse than RFsp and RF-GLS. However, when the spatial variance (σ^2) is high, RFsp performs comparably to that of RF-GLS, both of which outperform the others. For lower σ^2, RF-GLS significantly outperforms all the other methods. In this case, RFsp performs worse than RF-RK which is now the second best method.

We conducted an in-depth study in Section S1.4 to understand the performance of RFsp. We summarize the findings here. Mentch and Zhou (2020) showed that for iid errors, adding additional noise covariates can improve prediction performance of RF for low (covariate-)signal-to-(random-)noise ratio (SNR). For high SNR, the trend was reversed and using extra noise covariates (when uncorrelated with the true covariates) did not help. For our setting of spatially correlated errors, akin to how noise covariates can be added to RF to explain random variation, RFsp adds distance-based covariates to explain spatial variation. The appropriate quantity to understand the performance of RFsp would be the (covariate-)signal-to-(spatial-)noise ratio SNR $= \text{var}(m(\mathbf{X}))/\text{var}(w(\ell)) = \text{var}(m(\mathbf{X}))/\sigma^2$. This SNR is low when σ^2 is large and vice versa (Figure S3(a)).

RFsp adds n pairwise distance-based covariates to RF to explain the spatial variation. If D denotes the number of true covariates, then RFsp uses a total of $n + D$ covariates in RF and the probability to include a true covariate in the M_{try} set of candidates for splitting a node becomes vanishingly small when $n \gg D$. When σ^2 is small (high SNR), the true covariates predominantly dictates the variation in the outcome but is rarely selected in RFsp (Figure S3(b)) resulting in its poor performance. In fact for high SNR, RFsp performs even worse than RF-RK which can estimate m reasonably well in this setting despite ignoring the spatial dependence as the covariate signal dominates. For low SNR ($\sigma^2 = 10$), as the spatial contribution increases, RFsp (being naturally equipped to capture the spatial correlation in the data) performs comparably to RF-GLS (Figure 3). For low SNR, RF-RK, ignoring the dominant spatial variation when estimating m, performs worse than both RF-GLS and RFsp.

RF-RK and RFsp outperforms each other in two ends of the SNR spectrum. RF-GLS combines the strengths of both, using RF to estimate a nonlinear covariate effect while parsimoniously modeling the spatial effect using a GP specified by only 2 or 3 parameters thereby avoiding introduction of a large number of spatial covariates. Consequently, RF-GLS produce comparable or better results to both RF-RK and RFsp across the SNR spectrum (Figure 3).

Section S1 of the supplementary materials file contains a number of additional simulation studies. These include results for larger sample size (Section S1.5, supplementary material), mean function with more covariates (Section S1.6, supplementary material), misspecified GP smoothness (Section S1.7, supplementary material), and misspecified entire spatial effect (not generated from a GP, Section S1.8, supplementary material). For each study, and all choices of data-generation parameters, RF-GLS performed as the best or comparably with

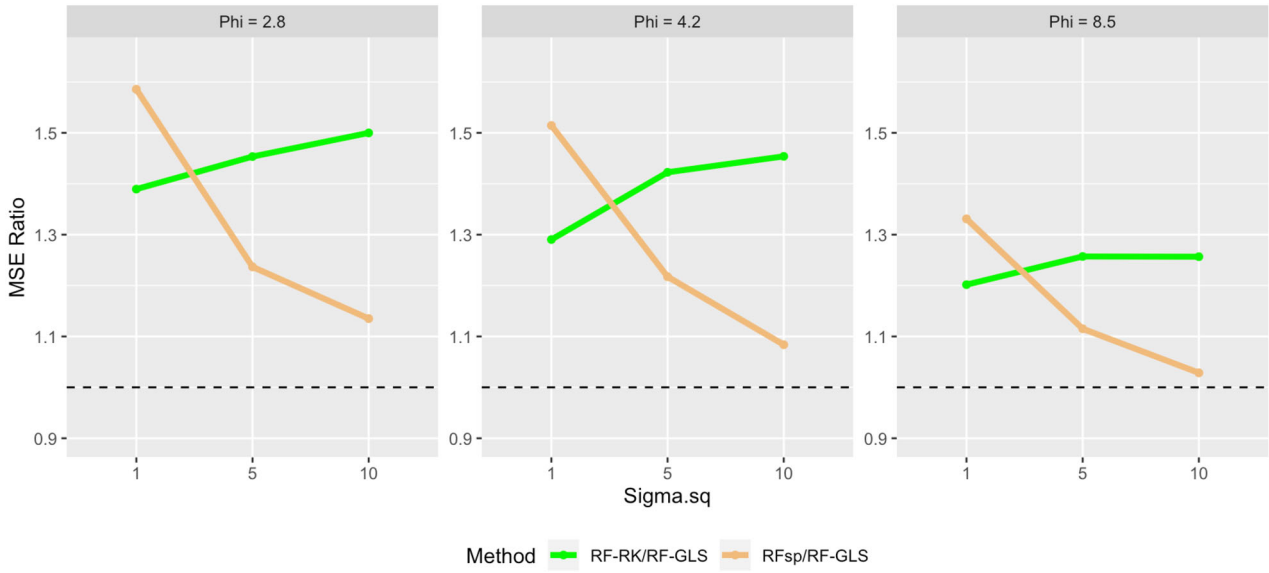


Figure 3. MSE ratio for RF-RK & RF-GLS ($\frac{\text{MSE(RF-RK)}}{\text{MSE(RF-GLS)}}$) and RFsp & RF-GLS ($\frac{\text{MSE(RFsp)}}{\text{MSE(RF-GLS)}}$) for $\tau^2 = 10\%\sigma^2$ when the mean function is $m = m_2$.

the best method across all scenarios for both estimation and prediction.

5. Proof of Consistency

For studying RF-GLS with dependent data, we adopt the framework of consistency analysis of nonparametric regression (introduced in Nobel et al. 1996, generalized in Györfi et al. 2002). In Scornet et al. (2015), the authors also adopted this framework to prove consistency of RF for iid errors. After presenting an informal outline of the consistency argument in Györfi et al. (2002), we provide a road map of how we extend different pieces of this argument for RF-GLS with dependent data, and highlight additional technical challenges that we resolved in this work. All formal proofs are provided in the supplementary materials.

Györfi et al. (2002) considered a least-square estimator of the form

$$m_n(,\Theta) \in \arg\min_{f\in\mathcal{F}_n} \frac{1}{n}\sum_i \big[f(X_i) - Y_i\big]^2$$

where $\mathcal{F}_n$ is a carefully chosen, data-dependent function class which is large enough to control *approximation error* (i.e., how well the true function is estimated by the function class), and small enough to control *estimation error* (i.e., how far the estimate m_n is from the representative of $\mathcal{F}_n$ closest to m) in the sense that a Uniform Law of Large Number (ULLN) holds on this class. m_n is consistent if both errors vanish asymptotically.

In order to use powerful exponential inequalities which hold for classes of bounded functions, the proof of $\mathbb{L}_2$-consistency in Györfi et al. (2002) uses a standard truncation argument in probability theory with a diverging sequence of truncation thresholds $\{\zeta_n\}$. They first show that if *uniformly over the class* $T_{\zeta_n}\mathcal{F}_n$ *of truncated functions* in $\mathcal{F}_n$,

1. *Approximation Error* of m by a data-driven class $\mathcal{F}_n$ containing m_n is small, that is,

$$\mathbb{E}\left[\inf_{f\in T_{\zeta_n}\mathcal{F}_n} \mathbb{E}_X\big[f(X) - m(X)\big]^2\right] \to 0;$$

2. *Estimation Error* is small so that a ULLN holds over $\mathcal{F}_n$ for squared error loss, that is,
$\mathbb{E}\left[\sup_{f\in T_{\zeta_n}\mathcal{F}_n} \left|\frac{1}{n}\sum_i (f(X_i) - Y_i))^2 - \mathbb{E}[f(X) - Y]^2\right|\right] \to 0;$

then the truncated estimator $T_{\zeta_n} m_n$ is $\mathbb{L}_2$-consistent for f. Then they extended the consistency guarantee from truncated to the original estimators by showing that the *truncation error* vanishes under suitable tail decay assumptions on the error distribution.

In the analysis of RF-GLS, there are two main challenges. The error process ϵ_i is no longer an iid process but a stochastic process capturing the dependence. Also, we work with a quadratic loss $(\mathbf{Y} - f(\mathbf{X}))^\top \mathbf{Q}(\mathbf{Y} - f(\mathbf{X}))$ instead of $\|\mathbf{Y} - f(\mathbf{X})\|^2$ where, $f(\mathbf{X}) - \mathbf{Y} = \big(f(X_1) - Y_1, f(X_2) - Y_2, \ldots, f(X_n) - Y_n\big)^\top$. Addressing these require nontrivial generalizations of each of the above pieces.

5.1. Consistency of Quadratic Loss Optimizers in Data-Driven Function Classes Under Dependent Errors

Our main technical statement (Theorem 5.1) is a generalization of (Györfi et al. 2002, Theorem 10.2) to the setting of dependent error processes and quadratic loss functions. Let $\mathcal{D}_n = \big\{(X_1, Y_1), \ldots, (X_n, Y_n)\big\}$ be the data where $Y_i = m(X_i) + \epsilon_i$. With randomness parameter Θ, let $\mathcal{F}_n = \mathcal{F}_n(\mathcal{D}_n, \Theta)$ be a data-dependent class of functions. We will consider an optimal estimator $m_n \in \mathcal{F}_n$ with respect to quadratic loss:

$$m_n \in \arg\min_{f\in\mathcal{F}_n} \frac{1}{n}(f(\mathbf{X}) - \mathbf{Y})^\top \mathbf{Q}(f(\mathbf{X}) - \mathbf{Y}). \tag{16}$$

This simply states that m_n is the GLS estimate with respect to the working precision matrix $\mathbf{Q}$ in the class $\mathcal{F}_n$. This is analogous to the OLS assumption used in Györfi et al. (2002). When $\mathcal{F}_n$ is the class of piecewise constant functions on the partitions

generated by a regression tree, m_n is our GLS-style regression-tree estimate. Hence studying estimators of the form (16) more generally suffices to prove consistency of GLS-style regression tree and henceforth of RF-GLS.

We now state a general technical result establishing $\mathbb{L}_2$-consistency of such GLS estimators m_n under β-mixing (absolutely regular) error processes. This is a sufficiently large class of processes that includes spatial Matérn GP and autoregressive time series as discussed in Sections 3.3.1 and 3.3.2. The result is applicable beyond RF to more general nonparametric GLS estimators from dependent data using other suitable function classes.

Theorem 5.1. Let $\{\epsilon_i\}$ be a stationary β-mixing process satisfying Assumption 1, and the matrix $\mathbf{Q}$ satisfies Assumptions 2 and 3. Let $m_n(.,\Theta) : \mathbb{R}^D \to \mathbb{R}$ denote a quadratic-loss optimizer (with respect to $\mathbf{Q}$) of the form (16) in a data-dependent function class $\mathcal{F}_n$. If m_n and $\mathcal{F}_n$ satisfies the following conditions:
(C.1) (Truncation error) $\exists\ \{\zeta_n\}$ such that $\lim_{n\to\infty} \zeta_n = \infty$ and $\zeta_n^2/n \to 0$, such that we have,

$$\lim_{n\to\infty} \mathbb{E} \max_i \left[m_n(X_i) - T_{\zeta_n} m_n(X_i)\right]^2 = 0$$

(C.2) (Approximation error) $\lim_{n\to\infty} \mathbb{E}_\Theta\big[\inf_{f\in T_{\zeta_n}\mathcal{F}_n} \mathbb{E}_X|f(X) - m(X)|^2\big] = 0$
(C.3) (Estimation error) Let $\dot{X}_i$, $\dot{\epsilon}_i$ and $\dot{Y}_i = m(\dot{X}_i) + \dot{\epsilon}_i$ be such that $\dot{\mathcal{D}}_n = \{(\dot{X}_i, \dot{Y}_i)|i = 1,\ldots,n\}$ be identically distributed as $\mathcal{D}_n$ but independent of $\mathcal{D}_n$. Define $f(\dot{\mathbf{X}})$ and $\dot{\mathbf{Y}}$ similar to $f(\mathbf{X})$ and $\mathbf{Y}$. Then, we have for all arbitrary $L > 0$

$$\lim_{n\to\infty} \mathbb{E}\Bigg[\sup_{f\in T_{\zeta_n}\mathcal{F}_n} \Big|\frac{1}{n}(f(\mathbf{X}) - \mathbf{Y})^\top \mathbf{Q}(f(\mathbf{X}) - \mathbf{Y}) - \mathbb{E}\frac{1}{n}(f(\dot{\mathbf{X}}) - \dot{\mathbf{Y}})^\top \mathbf{Q}(f(\dot{\mathbf{X}}) - \dot{\mathbf{Y}})\Big|\Bigg] = 0.$$

Then we have

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}\left[\mathbb{E}_X(m_n(X,\Theta) - m(X))^2)\right] = 0, \text{ and} \\ &\lim_{n\to\infty} \mathbb{E}_X(\bar{m}_n(X) - m(X))^2 \qquad\quad = 0; \end{aligned}$$

where $\bar{m}_n(X) = \mathbb{E}_\Theta m_n(X,\Theta)$ and X is a new sample independent of the data.

The proof is deferred to Section S2.3. Theorem 5.1 is a result of independent importance as it is a general statement on $\mathbb{L}_2$ consistency of a wide class of GLS estimates under β-mixing dependent errors. Besides data-driven-partitioning-based estimates like RF or RF-GLS, it can be used to study properties of histograms, kernel-density estimates, local polynomials, etc., under β-mixing error processes. The second part of the theorem states that the consistency also holds for an ensemble estimator $\bar{m}_n$ that averages many such estimates m_n each specified with random parameters Θ. This will be used to show consistency of the RF-GLS forest estimate subsequent to showing consistency of each RF-GLS tree estimate.

5.2. Approximation Error

The condition C.2 of asymptotically vanishing approximation error ensures that as sample size increases, the growing class of approximating functions (e.g., piece-wise constant functions in the case of regression trees) is rich enough to approximate the target function. In earlier works on consistency of RF, approximation error was controlled under a stringent assumption of vanishing diameter of leaf nodes. Scornet et al. (2015) replaced this by a condition that the variance of Y (or equivalently, variation of m) within a leaf node of a regression-tree vanishes asymptotically, and verified this condition for RF. There are two steps to show this.
(i) Establish a *theoretical or population-level split-criterion*—an asymptotic limit of the empirical split-criterion used in practice, such that variation of m in the leaf-nodes of a hypothetical regression tree generated using the theoretical criterion is small.
(ii) Establish stochastic equicontinuity of the empirical split-criterion, such that if two set of qualifying splits $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ are close, their corresponding empirical split-criterion values are close, irrespective of the location of the splits.

For our RF-GLS trees, the partitioning is driven by the DART criterion (5) and is different from the CART (2) criterion used in the RF trees. Since the GLS loss and estimator involves the matrix $\mathbf{Q}$, they are not available in simple scalar expressions unlike the OLS loss (sum of squares) and estimator (mean response within a node). So we address a number of technical challenges for steps (i) and (ii) that do not appear in the analysis of RF.

For (i), Lemma 2.1 and Theorem 2.1, establishes that the DART split-criterion remarkably has the same limit of as that for the CART criterion. Hence, variation of m in trees generated by this theoretical criterion is controlled in the same way as for RF.

For (ii), we require an entirely new and involved proof of equicontinuity for the DART-split criterion of RF-GLS as the previous arguments of Scornet et al. (2015) do not immediately generalize for RF-GLS loss function (5). We discuss the new contributions in Section 5.2.1.

5.2.1. Equicontinuity of the Split Criterion

Equicontinuity of the CART-split criterion $\frac{1}{n_P}[\sum_{i=1}^{n_P}(Y_i^P - \bar{Y}^P)^2 - \sum_{i_r=1}^{n_R}(Y_i^R - \bar{Y}^R)^2 - \sum_{i_l=1}^{n_L}(Y_i^L - \bar{Y}^L)^2]$ was the centerpiece of the theory in Scornet et al. (2015), requiring involved but elegant arguments on the geometry of splits. Since the CART criterion only concerns the parent node to be split and its potential two child nodes, the equicontinuity essentially boiled down to showing closeness of the respective means and variances of these three nodes for the two sets of splits. These three scalar mean and variance differences are functions of the difference in volumes of the respective nodes which goes to zero uniformly as the splits come closer.

For RF-GLS, to update each node, the entire set of node representatives get updated via the GLS-estimate (6) which is analytically intractable due to the matrix inversion. Also, the DART-split criterion (5)

$$\frac{1}{n}\Bigg[\left(\mathbf{Y} - \mathbf{Z}^{(0)}\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z}^{(0)})\right)^\top \mathbf{Q}\left(\mathbf{Y} - \mathbf{Z}^{(0)}\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z}^{(0)})\right) - \frac{1}{n}\left(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z})\right)^\top \mathbf{Q}\left(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{\text{GLS}}(\mathbf{Z})\right)\Bigg]$$

is a quadratic form of the plugged-in GLS-estimate and thus a function of the representatives of all nodes and not just the 3

nodes as in RF. Our equicontinuity proof is built on viewing GLS predictions as oblique projections on the design matrices corresponding to the splits.

We consider two scenarios: R1—where for at least one set of split, both potential child nodes have substantial volumes, and R2—where for each set of split, one potential child node have ignorable volume. Under R1, all nodes for at least one set of split have substantial representation ensuring that the gram matrix has norm bounded away from zero and equicontinuity can be established using perturbation bounds on orthogonal projection operators (Chen, Chen, and Li 2016). Under R2, this will not be the case as for both set of splits one child node will have small volume. Instead, we first show that difference of the DART-split criterion at the previous level of (parent) splits is small using the same perturbation argument as the parent node volumes are bounded away from zero. Subsequently we argue that the creation of the children nodes do not change the split criterion substantially as one of the child nodes is essentially empty. This new matrix-based proof for equicontinuity also circumvents the need to invoke mathematical induction as required in Scornet et al. (2015).

Proposition 5.1 (Equicontinuity of empirical DART-split criterion). Under Assumptions 2–4(b) and (c), the DART split criterion (5) is stochastically equicontinuous with respect to the set of splits.

The proof is involved, and is deferred to Section S2.1 where the Proposition is more technically phrased using additional notation on splits. Subsequent to proving equicontinuity, we can prove the following result of approximation error

Proposition 5.2. Let $\mathcal{F}_n = \mathcal{F}_n(\Theta)$ is the set of all functions $f : [0,1]^D \to \mathbb{R}$, piecewise constant on each cell of the partition obtained by an RF-GLS tree. Then, under Assumptions 1–5, the class $T_{\zeta_n}\mathcal{F}_n$ satisfies the approximation error condition (C.2).

5.3. Estimation Error

Theorem 5.1 shows that for GLS estimators like Equation (16), one needs to control the quadratic form estimation error (ULLN C.3). A common technique for proving ULLNs under dependence is

(i) to prove an analogous ULLN for iid error processes, and then
(ii) use mixing conditions to generalize the result for the dependent process of interest.

Both steps require addressing new challenges for RF-GLS which we discuss in the next two subsections.

5.3.1. Cross-Product Function Classes

For step (i), it is difficult to directly state an iid analogue of the ULLN **C.3** as it is not possible to have an error process $\{\epsilon_i^*\}$ which is simultaneously iid, satisfies $\epsilon_i^* \sim \epsilon_i$ and $E(\mathbf{y}^\top \mathbf{Q}\mathbf{y}) = E(\mathbf{y}^{*\top}\mathbf{Q}\mathbf{y}^*)$. To see this, simply note that as $\text{cov}(\epsilon_i^*, \epsilon_{i-j}^*) = 0 \neq \text{cov}(\epsilon_i, \epsilon_{i-j})$, with $y_i^* = m(X_i) + \epsilon_i^*$ we will have $E(q_{i,i-j} y_i y_{i-j}) \neq E(q_{i,i-j} y_i^* y_{i-j}^*)$ where $\mathbf{Q} = (q_{ii'})$. Instead, we create separate iid analogues of ULLN for each term in the expansion of the quadratic form, that is, both the squared error and the cross-product terms. These ULLN are stated as condition C.3.iid in the supplementary materials. Condition C.3.iid(a) for the squared (diagonal) terms is the standard squared error ULLN using the iid error processes, and has been proved in Györfi et al. (2002) Theorem 10.2 generally, and in Scornet et al. (2015) in particular for RF.

For the cross-product (off-diagonal) terms, the ULLN is stated in Condition C.3.iid(b) of the supplementary material, and is to our knowledge a novel strategy. The analysis of RF under iid settings in Scornet et al. (2015) only used a squared error loss which does not involve any cross-product terms and hence did not need to prove any ULLN for cross-product terms. We construct separate ULLN for the cross-product terms $\sum_i q_{i,i-j}\epsilon_i\epsilon_{i-j}$ for each lag $j = 1, \ldots, q$. As mentioned earlier, use of the univariate copy $\{\epsilon_i^*\}$ will not allow to generalize to the corresponding term in C.3. This is because $\epsilon_i^* \perp \epsilon_{i-j}^*$ but ϵ_i and ϵ_{i-j} are correlated. Instead, for each lag $j = 1, \ldots, q$, we create bivariate iid sequences $(\tilde{\epsilon}_i, \ddot{\epsilon}_{i-j})$ that are identically distributed as the joint (bivariate) distribution of the pairs $(\epsilon_i, \epsilon_{i-j})$, but are independent over i. We thus exploit the banded nature of the working precision matrix $\mathbf{Q}$ to prove iid ULLNs for cross-product terms at each of the q lags. Combining these with the ULLN for the squared terms gives us an ULLN for the entire quadratic form.

Formulation and establishing this cross-product ULLN using lag-specific bivariate iid copies is a new contribution and is of independent importance for establishing vanishing limits of any estimation error that involves interaction terms. We prove this ULLN in Proposition S2.1 of the supplementary material by showing that cross-product function classes has the same concentration rate as that for squared-error function classes with respect to the random $\mathbb{L}_p$ norm entropy number.

5.3.2. ULLN for β-Mixing Processes

For step (ii), we need to go from Conditions (C.3.iid)(a) and (C.3.iid)(b) for iid processes to their analogs for dependent error processes, which would then immediately establish C.3. It has been shown that the mixing condition of the stochastic process determines the assumptions required on the class $\mathcal{F}_n$ (Dehling and Philipp 2002). If we look at the "hierarchy" of dependence structures, strong-mixing or α-mixing (Bradley 2005), is one of the broadest family of dependent processes accommodating dependent structures "furthest" from independence. However, existing ULLN results for α-mixing processes require the class of functions in $\mathcal{F}_n$ to be Lipschitz continuous (Dehling and Philipp 2002). As regression-tree estimates are inherently discontinuous due to nature of discrete partitioning, this will not be satisfied here.

Hence, we focus on absolutely regular or β-mixing process. This class of mixing processes is rich enough to include a number of commonly used spatial or time-series structures as discussed in the examples of Section 3.3. Our main challenge here is that no existing ULLN for β-mixing processes apply to the class of functions on partitions of the RF-GLS trees. ULLN for Glivenko-Cantelli classes under β-mixing was established in Nobel and Dembo (1993). Similar results have been established for a class of $\tilde{\phi}$-mixing processes in Peligrad (2001). Both results,

do not need any convergence rate on the mixing coefficients, but require the class $\mathcal{F}_n$ to have an envelope (dominator) F (free of n). For data-driven partitioning based estimates like RF or RF-GLS trees such uniform envelopes are not available (as the envelop is the truncation threshold $\zeta_n \to \infty$). Instead, we propose an ULLN for dependent processes that uses a weaker assumption of a n-varying envelop with a moment-bound. The proof is deferred to Section S2.2 (supplementary material).

Proposition 5.3. (A general ULLN for β-mixing processes) Let $\{U_i\}$ be an $\mathbb{R}^d$-valued stationary β-mixing process. Let $\mathcal{G}_n(\{U_i\}_{i=0}^{n-1})$ be a class of functions $\mathbb{R}^d \to \mathbb{R}$ with envelope $G_n \geq \sup_{g\in\mathcal{G}_n} |g|$, such that G_n is "uniformly mean integrable", that is,

$$\lim_{C\to\infty} \lim_{n\to\infty} \mathbb{E}\frac{1}{n}\sum_i |G_n(U_i)|\mathbb{I}(|G_n(U_i)| > C) = 0. \tag{17}$$

Let $\{U_i^*\}$ be such that U_i^* is identically distributed as U_i, $\forall\, i$ and $U_i^* \perp U_j^*$; $\forall i \neq j$. Then, $\sup_{g\in\mathcal{G}_n} \left|\frac{1}{n}\sum_i (g(U_i^*) - \mathbb{E}g(U_i^*))\right| \xrightarrow{\mathbb{L}_1} 0 \implies \sup_{g\in\mathcal{G}_n} \left|\frac{1}{n}\sum_i (g(U_i) - \mathbb{E}g(U_i))\right| \xrightarrow{\mathbb{L}_1} 0$.

Proposition 5.3 ensures that ULLN for iid errors is enough to generalize to β-mixing error processes as long as the function classes are contained within a sequence of *mean uniform integrable* envelopes in the sense of Equation (17). Next, we show that the ULLN holds for RF-GLS trees under a $(2+\delta)$th moment assumption that is sufficient for mean uniform integrability.

Proposition 5.4 (Estimation error for RF-GLS). Let $\mathcal{F}_n = \mathcal{F}_n(\Theta)$ is the set of all functions $f : [0,1]^D \to \mathbb{R}$, piecewise constant on each cell of the partition obtained by an RF-GLS tree. If any subset $\tilde{F}_n \subseteq \mathcal{F}_n$ satisfies the following condition:
(C.4) (Moment bound:) $\exists$ an envelope $F_n \geqslant \sup_{f\in\tilde{\mathcal{F}}_n} |f|$, such that $\lim_{n\to\infty} \mathbb{E}\frac{1}{n}\sum_i |F_n(X_i)|^{2+\delta} < \infty$ for some $\delta > 0$,
then $T_{\zeta_n}\tilde{\mathcal{F}}_n$ satisfies the ULLN (C3) for β-mixing error processes, and under Assumption 2.

The proof is in Section S2.2. The $(2+\delta)$th moment condition C.4 is easier to verify for RF or RF-GLS as discussed in Corollaries 3.1 and 3.2.

5.4. Proof of Theorem 3.1

Equipped with Theorem 5.1, we can prove Theorem 3.1 by showing that RF-GLS meets the conditions C.1–C.3. Proposition S2.3 in the supplementary material) shows that the truncation error condition C.1 is met for any ζ_n satisfying the scalings of Assumption 4(a) required for proving the approximation and estimation error conditions.

We have already shown conditions C.2 (Proposition 5.2) and C.3 (Proposition 5.4) for two separate choices of function classes. The last step of the proof is choosing the function class that satisfies both conditions. Let $\mathcal{F}_n = \mathcal{F}_n(\Theta)$ be the set of all functions $f : [0,1]^D \to \mathbb{R}$ piece-wise constant on each cell of the partition $\mathcal{P}_n(\Theta)$ created by an RF-GLS tree with data $\mathcal{D}_n$ and randomization Θ. We have already shown in in Proposition 5.2 that $\mathcal{F}_n$ satisfies C.2. To apply Proposition 5.4, $\mathcal{F}_n$ also needs to satisfy the moment condition of C.4. As $T_{\zeta_n}\mathcal{F}_n$ is only bounded by ζ_n which goes to ∞, clearly $\mathcal{F}_n$ will not satisfy C.4 .

We carefully carve out a subclass $\tilde{\mathcal{F}}_n \subseteq \mathcal{F}_n$ which is still wide enough to satisfy the approximation error condition (C.2), while satisfying the additional restriction (C.4). For a given partition $\mathcal{P}_n(\Theta)$, we define $\tilde{\mathcal{F}}_n$ as follows:

$$\tilde{\mathcal{F}}_n = \tilde{\mathcal{F}}_n(\Theta) = \{m_n\} \cup \left\{ \cup_{\mathbf{x}_\mathcal{B}\in\mathcal{B}\in\mathcal{P}_n(\Theta)} \sum_{\mathcal{B}\in\mathcal{P}_n(\Theta)} m(\mathbf{x}_\mathcal{B})\mathbb{I}(\mathbf{x}\in\mathcal{B}) \right\} \subseteq \mathcal{F}_n(\Theta). \tag{18}$$

Since by construction of RF-GLS, m_n is the optimizer over a much larger set $\mathcal{F}_n(\Theta)$, trivially m_n is also the optimizer in $\tilde{\mathcal{F}}_n$. The first step of the proof of Proposition 5.2 makes it evident why Condition C.2 will also hold for this smaller class $\tilde{\mathcal{F}}_n$. To apply Proposition 5.4 and show Condition C.3, the final piece is to show that the condition (C.4) is satisfied by $\mathcal{F}_n$. Since apart from m_n, $\mathcal{F}_n$ consists of functions that are bounded by M_0, we can have the envelope to be $F_n = |m_n| + M_0$. Hence, for Condition (C.4) to hold, it is enough to show $\lim_{n\to\infty} \frac{1}{n}\sum_i \mathbb{E}|m_n(X_i)|^{2+\delta} < \infty$ which is an assumption of the Theorem. □

6. Discussion

We considered nonlinear regression function estimation in the spatial mixed model and developed a random forest method (RF-GLS) for estimating the nonlinear covariate effect, while still modeling the spatial effect using GPs, as is conventional. Retaining the GP framework facilitates parsimonious encoding of structured spatial dependence, and conducting all traditional spatial tasks like kriging (prediction at a new location), and recovery of the latent spatial random effect surface. We show in Section 4 how these advantages of RF-GLS manifest into superior estimation performance over naive RF that does not use any spatial information, and superior predictive performance over many existing RFsp methods.

Our method RF-GLS, more generally, can be used for functional estimation under many types of dependence. RF-GLS uses the same fundamental principle that generalizes OLS to GLS. We show how adapting the concept of GLS in RF synergistically mitigates all the issues encountered in naive application of RF to dependent settings. While simple in principle, RF-GLS algorithm differs inherently from RF, by optimizing globally (across all nodes) for each split, to account for dependence across all data points. We show RF is a special case of RF-GLS with an identity working correlation matrix, and is substantially outperformed by RF-GLS for both estimation and prediction under dependence.

We present a thorough theoretical study establishing consistency of RF-GLS under a general assumption of β-mixing dependence. In particular, we establish consistency of function estimation by RF-GLS for the spatial nonlinear mixed-model using the ubiquitous Matérn GP. We also establish consistency of RF-GLS for functional estimation under auto-regressive time-series errors. Finally, as a byproduct of the theory, we also establish consistency of RF for dependent settings, which to our knowledge, is the first such result.

The theoretical results required involved proofs to address the challenges of accommodating dependent error processes

and use of a quadratic form loss for node-splitting. In the process, we developed a number of tools of independent importance. The general result (Theorem 5.1) on $\mathbb{L}_2$ consistency of data-driven partitioning-based GLS estimates under β-mixing dependence extends the analogous result in Györfi et al. (2002) Theorem 10.2 which was for iid settings and OLS estimates. This will be useful for studying other function estimators like local polynomials under dependence. Proposition S2.2 establishes random-norm entropy number concentration bounds for general cross-product function classes, which can find its use in establishing ULLN for any class of functions containing interaction terms. Finally, Proposition 5.3 proposes a general ULLN for β-mixing processes requiring less restrictive assumptions than existing results.

For future work, extension to multivariate outcomes is an important direction. The spatial community is increasing shifting toward multivariate analysis, as GIS systems are empowered to collect data on many variables. Multivariate extension of RF-GLS can leverage the rich literature of multivariate cross-covariance functions for GP (Genton and Kleiber see 2015, for a review). When working with a very large number of variables p, a potential computational roadblock for RF-GLS will be evaluation of the $np \times np$ Cholesky factor. Strategies like graphical models may need to be adopted to enforce sparsity and effectuate computational speedup.

Supplementary Materials

An online supplementary materials file contains additional data analyses, and proofs of all the theoretical results.

Acknowledgments

We thank to the editor, the associate editor and two anonymous reviewers for their suggestions which helped to improve the article. Arkajyoti Saha and Abhirup Datta were supported by NSF award DMS-1915803 and the Johns Hopkins Bloomberg American Health Initiative Spark Award. Sumanta Basu was supported by NSF award DMS-1812128, and NIH awards R01GM135926 and R21NS120227.

References

Ahijevych, D., Pinto, J. O., Williams, J. K., and Steiner, M. (2016), "Probabilistic Forecasts of Mesoscale Convective System Initiation Using the Random Forest Data Mining Technique," *Weather and Forecasting*, 31, 581–599. [1]

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: CRC Press. [1]

Bradley, R. C. (2005), "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," arXiv: math/0511078. [3,6,16]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [2]

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1,11]

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Boca Raton, FL: CRC Press. [3,4]

Carrasco, M. and Chen, X. (2002), "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 17–39. [9]

Chen, Y. M., Chen, X. S., and Li, W. (2016), "On Perturbation Bounds for Orthogonal Projections," *Numerical Algorithms*, 73, 433–444. [16]

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a), "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets," *Journal of the American Statistical Association*, 111, 800–812. [2,9,11]

——— (2016b), "On Nearest-Neighbor Gaussian Process Models for Massive Spatial Data," *Wiley Interdisciplinary Reviews: Computational Statistics*, 8, 162–171. [9]

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016c), "Nonseparable Dynamic Nearest Neighbor Gaussian Process Models for Large Spatio-Temporal Data With an Application to Particulate Matter Analysis," *Annals of Applied Statistics*, 10, 1286–1316. [9]

Dehling, H., and Philipp, W. (2002), "Empirical Process Techniques for Dependent Data," in *Empirical Process Techniques for Dependent Data*, Boston, MA: Springer, pp. 3–113. [16]

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., et al. (2019), "An Ensemble-Based Model of PM2. 5 Concentration Across the Contiguous United States with High Spatiotemporal Resolution," *Environment International*, 130, 104909. [1]

Diggle, P. J., and Hutchinson, M. F. (1989), "On Spline Smoothing With Autocorrelated Errors," *Australian Journal of Statistics*, 31, 166–182. [2]

Doukhan, P. (2012), *Mixing: Properties and Examples*, Vol. 85, New York: Springer Science & Business Media. [9]

Du, J., Zhang, H., Mandrekar, V., et al. (2009), "Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators," *The Annals of Statistics*, 37, 3330–3361. [11]

Fayad, I., Baghdadi, N., Bailly, J.-S., Barbier, N., Gond, V., Hérault, B., El Hajj, M., Fabre, F., and Perrin, J. (2016), "Regional Scale Rain-Forest Height Mapping Using Regression-Kriging of Spaceborne and Airborne LiDAR Data: Application on French Guiana," *Remote Sensing*, 8, 240. [1,2,8]

Finley, A. O., Datta, A., and Banerjee, S. (2021), "spNNGP: R package for Nearest Neighbor Gaussian Process Models," *Journal of Statistical Software*, (accepted conditionally). [9]

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019), "Efficient Algorithms for Bayesian Nearest Neighbor Gaussian Processes," *Journal of Computational and Graphical Statistics*, 28, 401–414. [9,11]

Fox, E. W., Ver Hoef, J. M., and Olsen, A. R. (2020), "Comparing Spatial Regression to Random Forests for Large Environmental Data Sets," *PloS One*, 15, e0229509. [2,12]

Friedman, J., Hastie, T., and Tibshirani, R. (2001), *The Elements of Statistical Learning*, Vol. 1, New York: Springer Series in Statistics. [10]

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 1–67. [12]

——— (2006), "Recent Advances in Predictive (Machine) Learning," *Journal of Classification*, 23, 175–197. [10]

Friedman, J. H., Popescu, B. E., et al. (2008), "Predictive Learning via Rule Ensembles," *The Annals of Applied Statistics*, 2, 916–954. [5]

Genton, M. G. and Kleiber, W. (2015), "Cross-Covariance Functions for Multivariate Geostatistics," *Statistical Science*, 147–163. [18]

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S. (2019), "Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling," *Geocarto International*, 1–16. [2]

Györfi, L., Kohler, M., Krżyzak, A., and Walk, H. (2002), *A Distribution-Free Theory of Nonparametric Regression* (Vol. 1), New York: Springer. [3,14,16,18]

Hartikainen, J., and Särkkä, S. (2010), "Kalman Filtering and Smoothing Solutions to Temporal Gaussian Process Regression Models," in *2010 IEEE International Workshop on Machine Learning for Signal Processing*, Kittilä, Finland, IEEE, pp. 379–384. [11]

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D., Sun, F., Zammit-Mangion, A. (2019), "A Case Study Competition Among Methods for Analyzing Large Spatial Data," *Journal of Agricultural, Biological and Environmental Statistics*, 24, 398–425. [9]

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., and Gräler, B. (2018), "Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-Temporal Variables," *PeerJ*, 6, e5518. [2,9,12]

Ihara, S. (1993), *Information Theory for Continuous Systems*, Vol. 2, World Scientific. [11]

Kimeldorf, G., and Wahba, G. (1971), "Some Results on Tchebycheffian Spline Functions," *Journal of Mathematical Analysis and Applications*, 33, 82–95. [2]

Lim, C. C., Kim, H., Vilcassim, M. R., Thurston, G. D., Gordon, T., Chen, L.-C., Lee, K., Heimbinder, M., and Kim, S.-Y. (2019), "Mapping Urban air Quality Using Mobile Sampling With Low-Cost Sensors and Machine Learning in Seoul, South Korea," *Environment International*, 131, 105022. [1]

Lin, Y., and Jeon, Y. (2006), "Random Forests and Adaptive Nearest Neighbors," *Journal of the American Statistical Association*, 101, 578–590. [1]

Mentch, L., and Hooker, G. (2016), "Quantifying Uncertainty in Random Forests Via Confidence Intervals and Hypothesis Tests," *The Journal of Machine Learning Research*, 17, 841–881. [2]

Mentch, L., and Zhou, S. (2020), "Getting Better from Worse: Augmented Bagging and a Cautionary Tale of Variable Importance," arXiv: 2003.03629. [13]

Mokkadem, A. (1988), "Mixing Properties of ARMA Processes," *Stochastic Processes and Their Applications*, 29, 309–315. [9,11,12]

Nobel, A., and Dembo, A. (1993), "A Note on Uniform Laws of Averages for Dependent Processes," *Statistics & Probability Letters*, 17, 169–172. [16]

Nobel, A. et al. (1996), "Histogram Regression Estimation Using Data-Dependent Partitions," *The Annals of Statistics*, 24, 1084–1105. [14]

Pardo-Igúzquiza, E. and Olea, R. A. (2012), "VARBOOT: A Spatial Bootstrap Program for Semivariogram Uncertainty Assessment," *Computers & Geosciences*, 41, 188–198. [8]

Peligrad, M. (2001), "A Note on the Uniform Laws for Dependent Processes Via Coupling," *Journal of Theoretical Probability*, 14, 979–988. [16]

Saha, A. and Datta, A. (2018a), "BRISC: Bootstrap for Rapid Inference on Spatial Covariances," *Statistics*, 7, e184. [8]

——— (2018b), *BRISC: Fast Inference for Large Spatial Datasets Using BRISC*, r package version 0.1.0. [9,12]

Scornet, E. (2016), "Random Forests and Kernel Methods," *IEEE Transactions on Information Theory*, 62, 1485–1500. [1]

Scornet, E., Biau, G., and Vert, J.-P. (2015), "Consistency of Random Forests," *The Annals of Statistics*, 43, 1716–1741. [2,3,10,14,15,16]

Stein, M. L. (2012), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Science & Business Media. New York: Springer-Verlag. [11]

Stein, M. L. (2002), "The Screening Effect in Kriging," *The Annals of Statistics*, 30, 298–323. [11]

Taylor, J., and Einbeck, J. (2013), "Challenging the Curse of Dimensionality in Multivariate Local Linear Regression," *Computational Statistics*, 28, 955–976. [1]

Vecchia, A. V. (1988), "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society*, Series B, 50, 297–312. [9]

Viscarra Rossel, R. A., Webster, R., and Kidd, D. (2014), "Mapping Gamma Radiation and Its Uncertainty From Weathering Products in a Tasmanian Landscape With a Proximal Sensor and Random Forest Kriging," *Earth Surface Processes and Landforms*, 39, 735–748. [1,8]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [2]

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261. [11]