Principled Spatial Machine Learning with Random Forests and Gaussian Processes

Abhi Datta
Johns Hopkins University
Department of Biostatistics

Outline of talk

Geospatial modeling using spatial linear mixed models

Success of machine learning methods like Random Forests (RF) for non-linear regression

Issues of RF for spatial (dependent) data

RF-GLS: Extending RF to model spatial dependence

Theoretical and empirical evaluations

Extensions to binary spatial data

Geospatial/point-referenced data

Data: $(Y_i, X_i, s_i) \ : i = 1, \ldots, n$

- Y_i : scalar response
- X_i : d-dimensional covariate
- s_i : location

Objectives:

Understand relationship between X and Y

Predict at a new location s_0

Spatial linear mixed models

$$Y_i = X_i^\top \beta + w(s_i) + \epsilon^*(s_i)$$

- Linear fixed effect
- Smooth spatial effect
  - Process-level modeling: Allow predictions at any location
- Measurement error / microscale variation $\epsilon^*(s_i) \overset{iid}{\sim} N(0, \tau^2)$

Spatial linear mixed models

$$Y_i = X_i^\top \beta + w(s_i) + \epsilon^*(s_i)$$

- $w(s_i)$ often modeled as a *Gaussian Process*

  $w(\cdot) \sim GP(0, C(\cdot, \cdot))$

  $w = (w(s_1), \ldots, w(s_n))' \sim N(0, C),$

  $C_{ij} = Cov(w(s_i), w(s_j)) = C(s_i, s_j \mid \theta)$

- C encodes the Tobler's First law of geography, i.e.,

  "*everything is related to everything else, but near things are more related than distant things.*"

- Example: Exponential covariance function $C(s_i, s_j \mid \theta) = \sigma^2 \exp(-\phi \|s_i - s_j\|_2)$

Spatial mixed models

Non-linear mean function

- $E(Y_i) = m(X_i)$

- Many recent spatial applications where m is estimated using Random forests (RF)
  - RF is an ensemble of regression trees
  - Trees essentially use data-driven basis functions
  - Can model higher order interactions

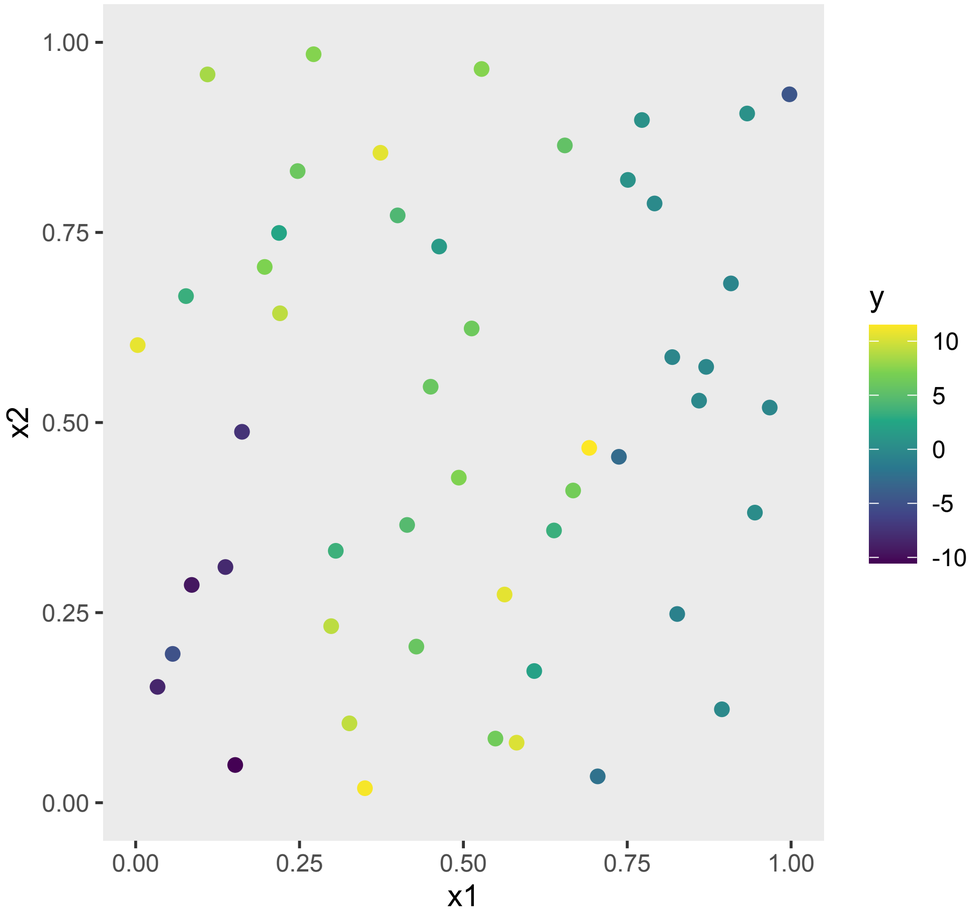
Random Forests

Leo Breiman

Machine Learning **45**, 5–32(2001)

Review of Regression Tress and Random Forests

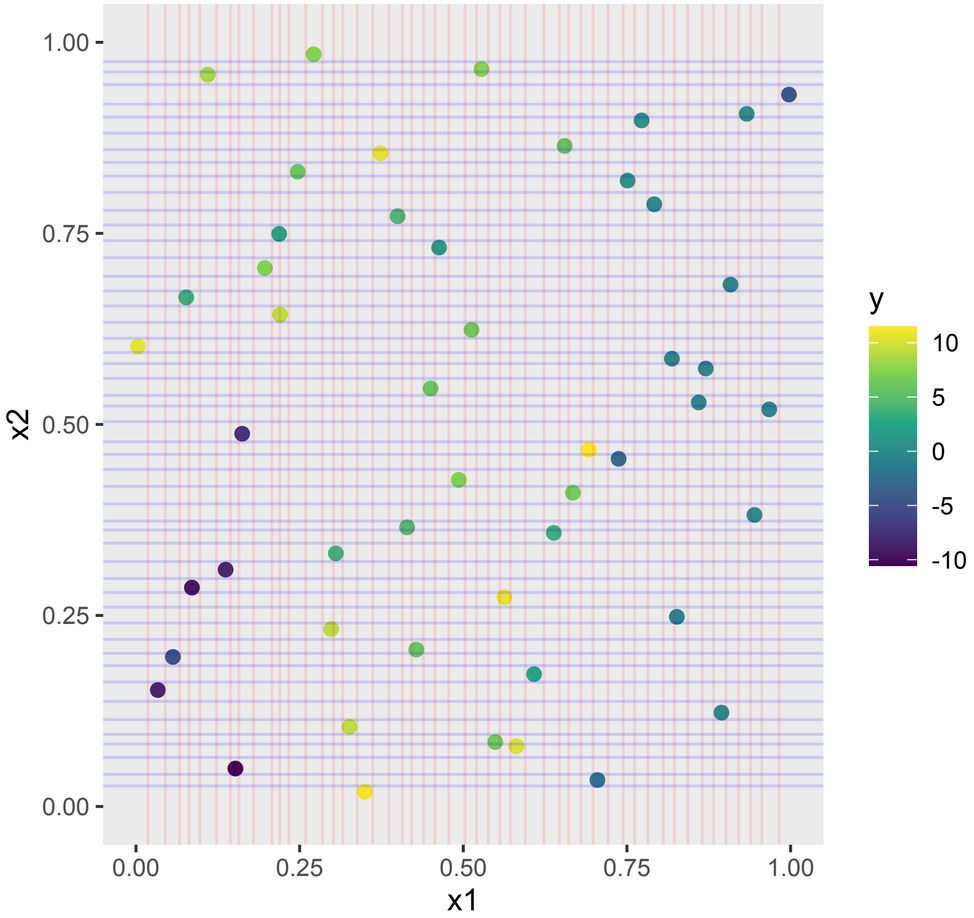
Regression trees



$[0,1]^2$

Review of Regression Tress and Random Forests

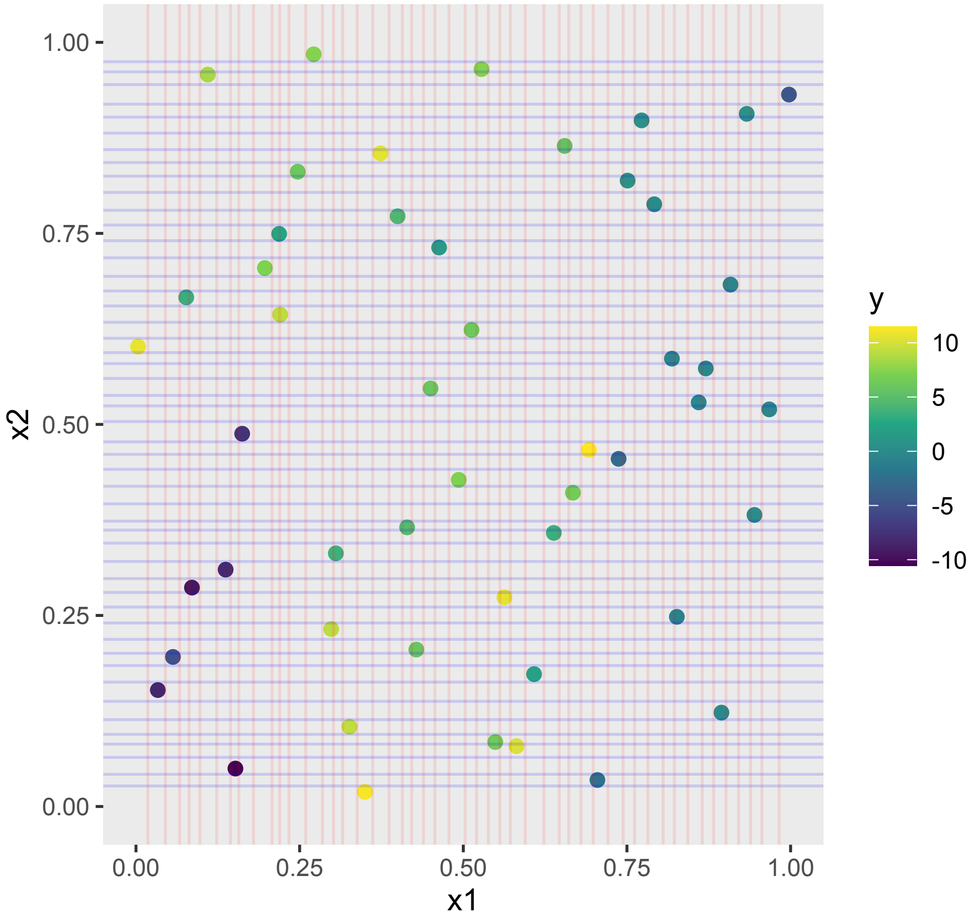
Regression trees



$[0,1]^2$

Review of Regression Tress and Random Forests

Regression trees



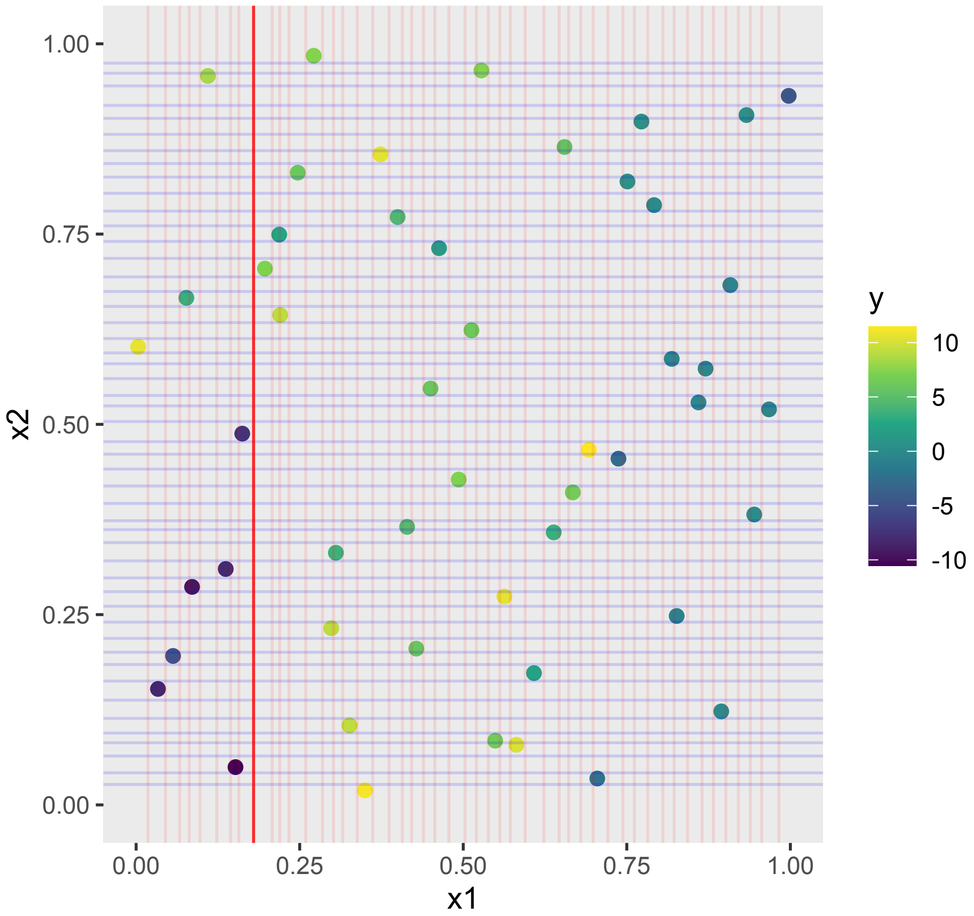
$[0,1]^2$

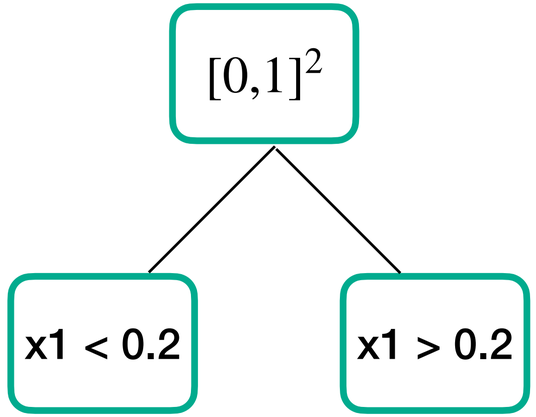
Regression tree (RT) Split criterion: Maximize

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

Review of Regression Tress and Random Forests

Regression trees

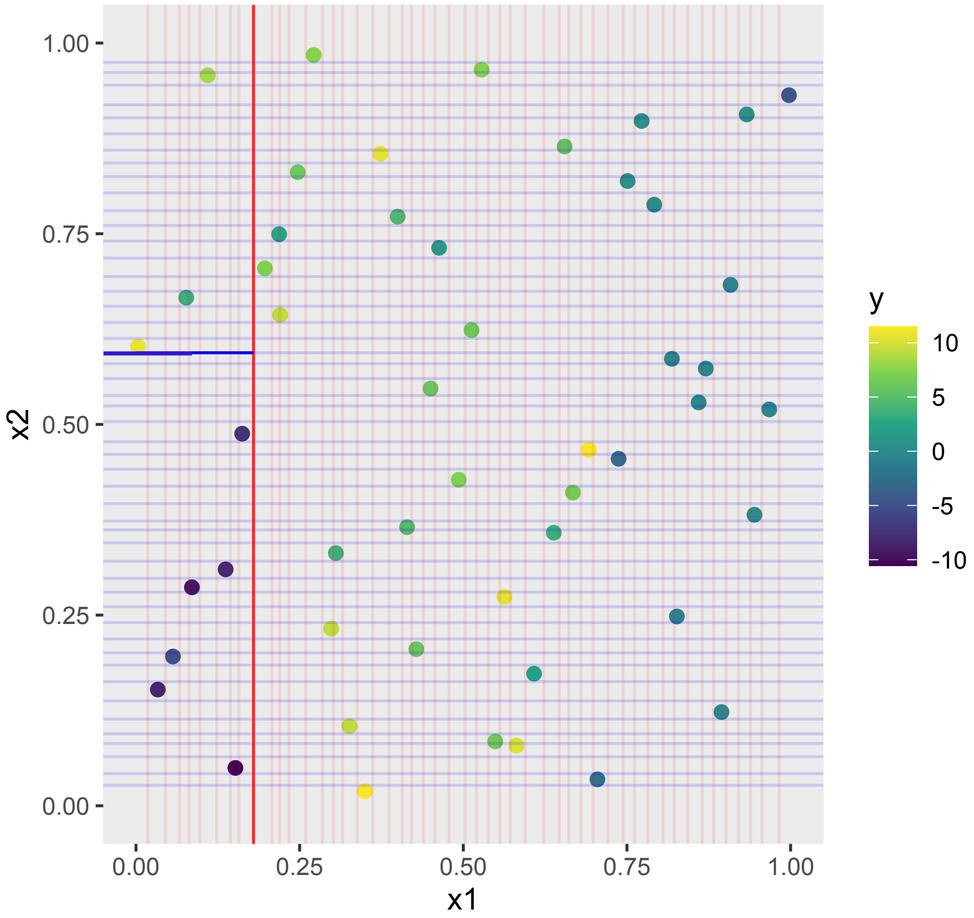


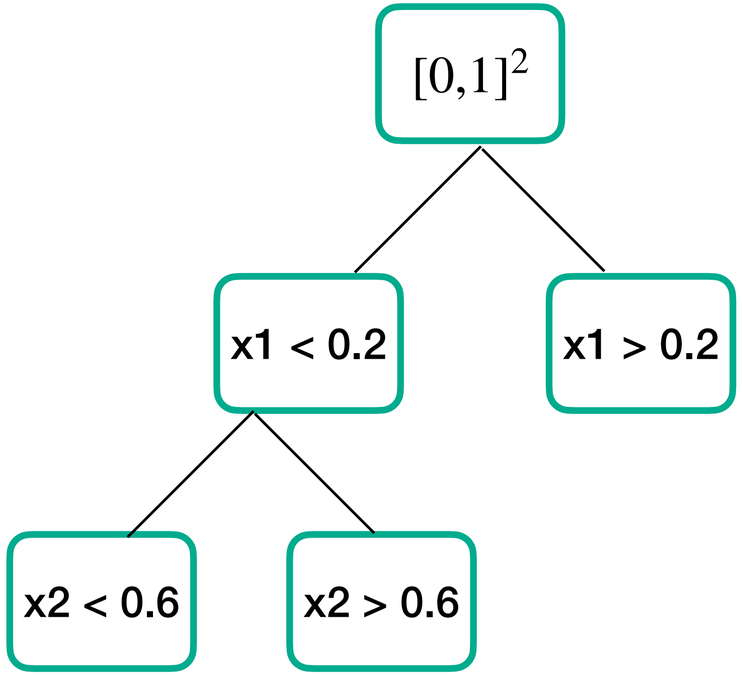


Regression tree (RT) Split criterion: Maximize $\frac{1}{n_P}\left[\sum_{i=1}^{n_P}(Y_i^P - \bar{Y}^P)^2 - \sum_{i_r=1}^{n_R}(Y_i^R - \bar{Y}^R)^2 - \sum_{i_l=1}^{n_L}(Y_i^L - \bar{Y}^L)^2\right]$

Review of Regression Tress and Random Forests

Regression trees



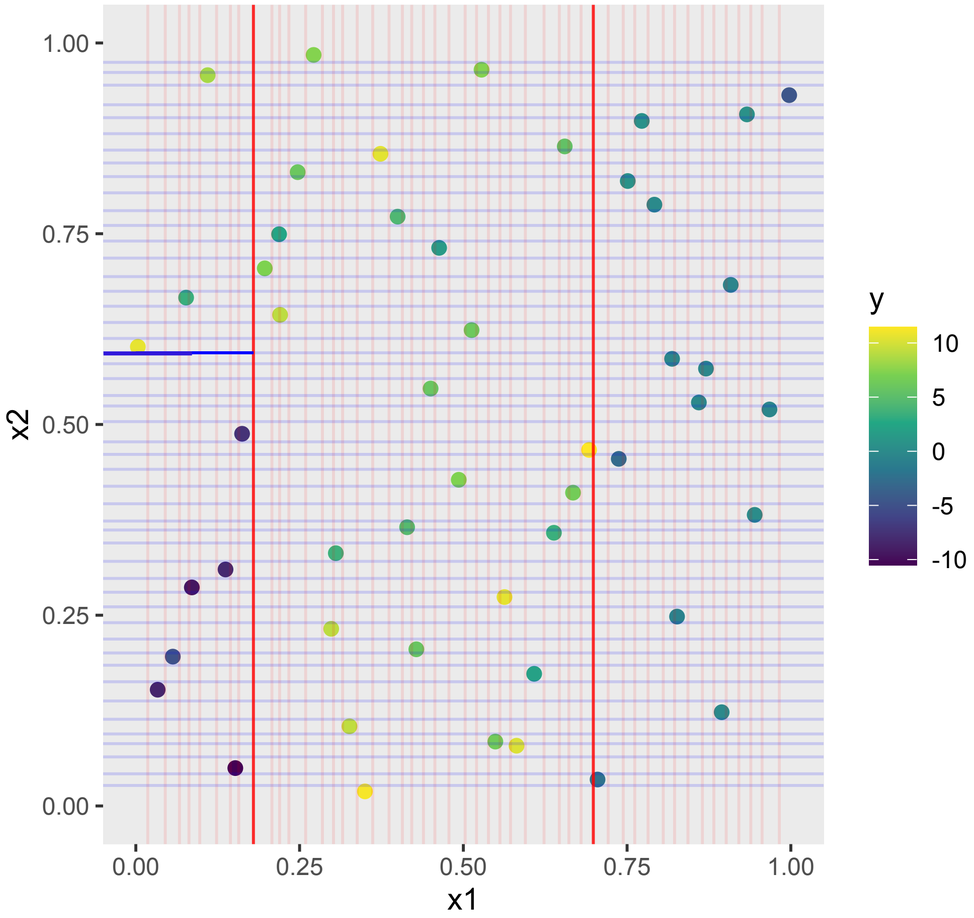


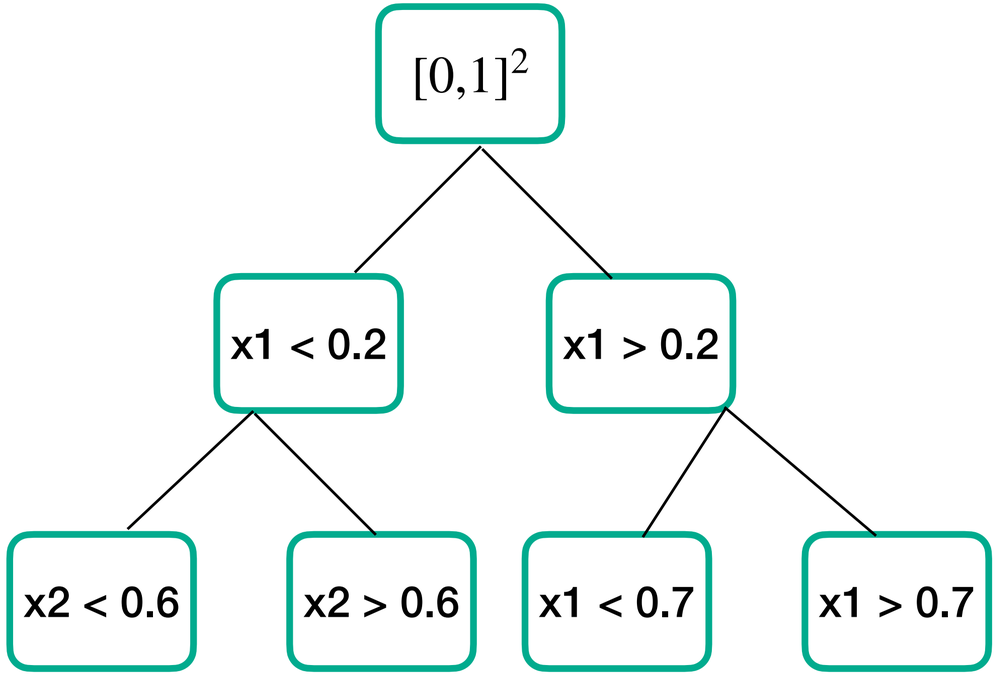
Regression tree (RT) Split criterion: Maximize

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

Review of Regression Tress and Random Forests

Regression trees





Regression tree (RT) Split criterion: Maximize

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

Review of Regression Tress and Random Forests

Data: $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^D, i = 1, \ldots, n$

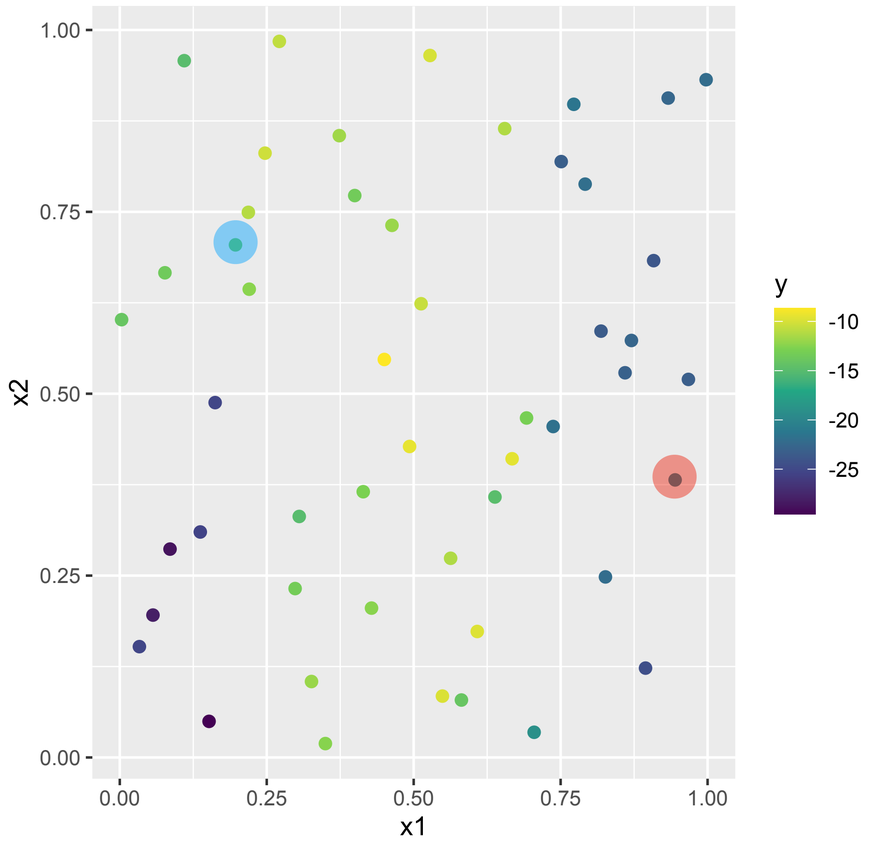
Node creation: Sequentially maximizing the RT-split criterion within each node

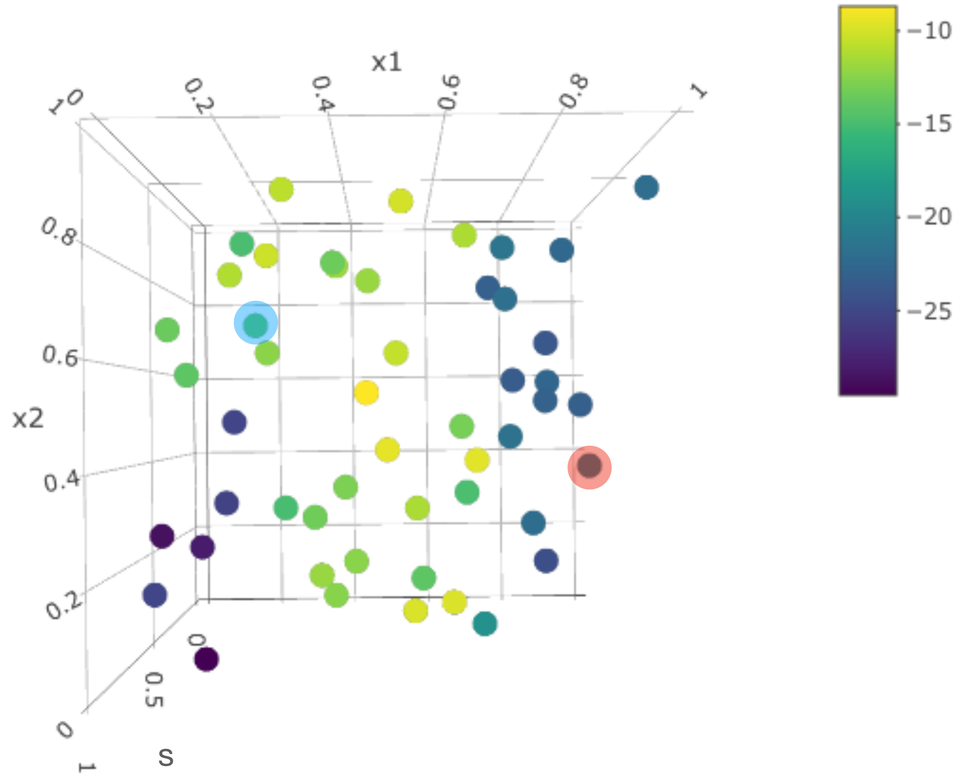
Leaf node estimates: The value of the tree estimator at each leaf node is the mean of the responses of the node members

Random feature selection: For each split, only consider a randomly chosen (M_try << D) subset of the features as candidate split direction

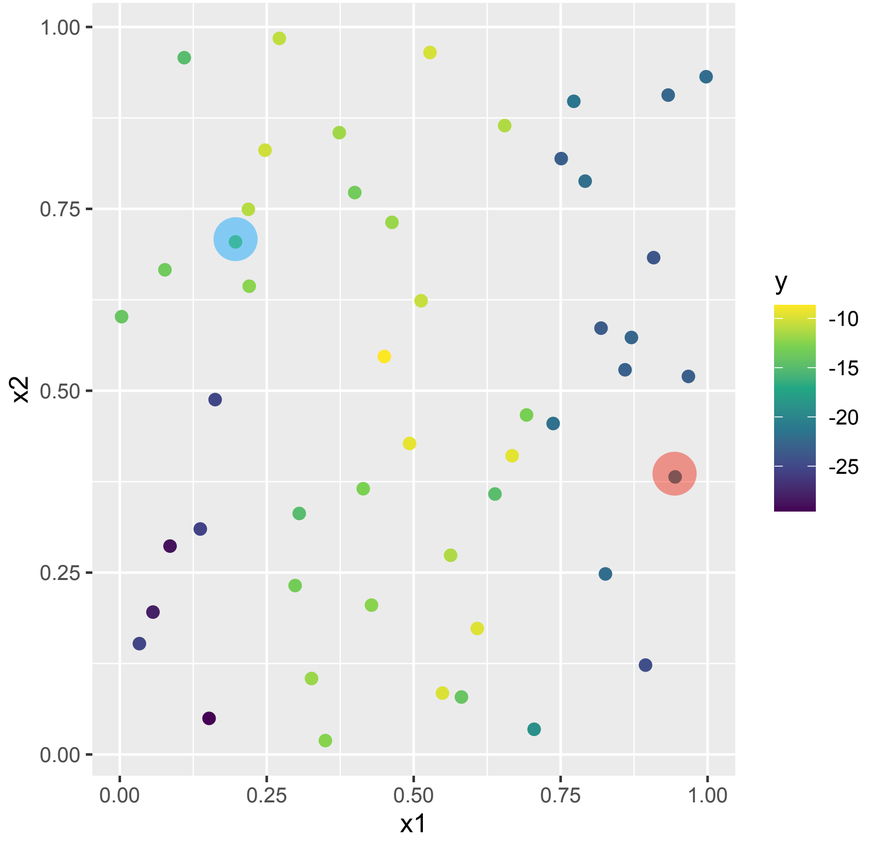
Bagging / subsampling: RF estimate = average of a large number of regression trees, each tree grown with a resample/subsample of the data

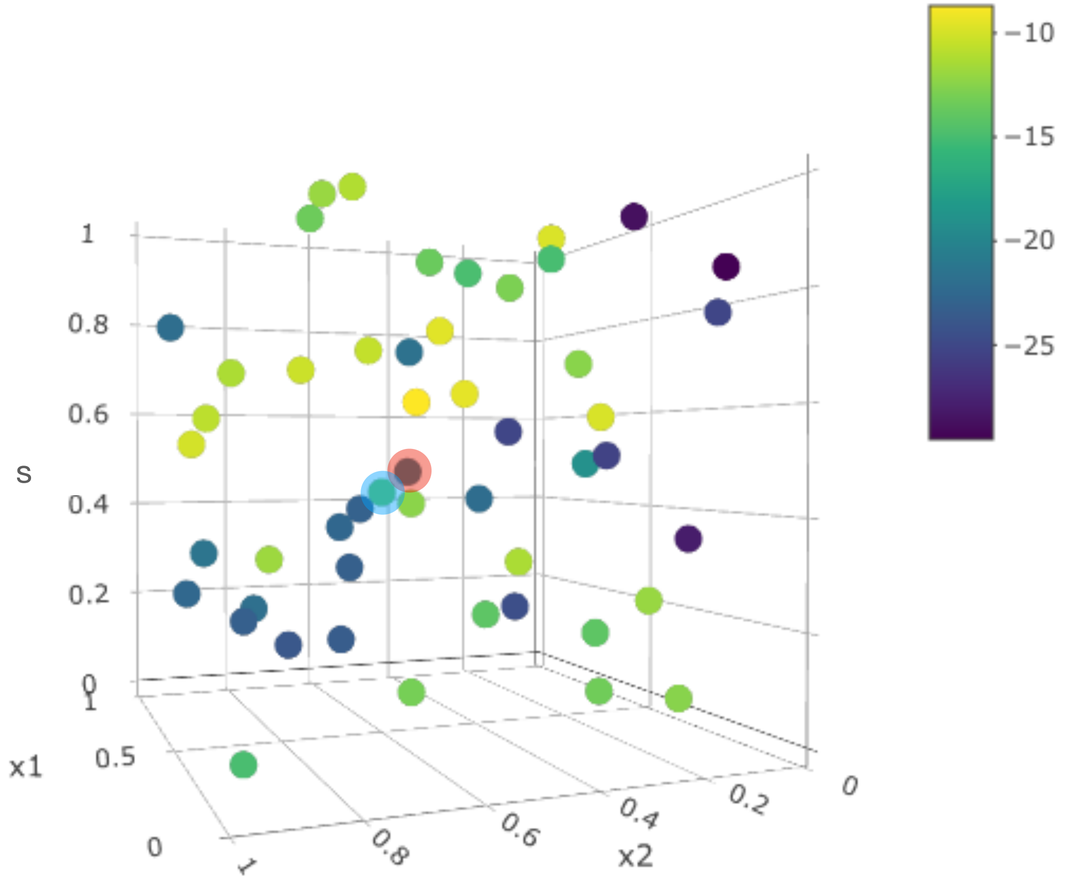
Issues of RF for dependent data





Issues of RF for dependent data





Issues of RF for dependent data

Tree growing and leaf node estimates ignore data correlation

Resampling of correlated data to create a forest of trees is not ideal

Spatial non-linear mixed model

Model:

$$Y_i = m(X_i) + w(s_i) + \epsilon^*(s_i);$$
$$w(\cdot) \sim GP(0, C(\cdot, \cdot));$$
$$\epsilon^*(s_i) \sim N(0, \tau^2)$$

- Model the spatial effect as a GP
- Estimate the non-linear m using RF (while accounting for the dependence)

Marginal model:

$$Y_i = m(X_i) + \epsilon(s_i);$$
$$\epsilon(s_i) = w(s_i) + \epsilon^*(s_i);$$
$$Cov(\epsilon) = C + \tau^2 I = \Sigma$$

Generalized least squares

Revisiting linear mixed models

Data generation: $Y_i = X_i^\top \beta + \epsilon_i$, where the errors ϵ_i is a dependent process.

If $\Sigma = Cov(\epsilon)$, then a common estimator of the mean function is $x^\top \hat{\beta}_{GLS}$ where

$$\hat{\beta}_{GLS} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Y = \arg\max_\beta (Y - X\beta)^\top \Sigma^{-1} (Y - X\beta)$$

The GLS loss is widely used to replace the OLS loss for time-series/spatial data

Revisiting RT-split criterion

Split k^{th} parent node into left and right child nodes by maximizing

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

New set of leaf nodes: $\{\mathcal{C}_1,\mathcal{C}_2,\cdots,\mathcal{C}_{K-1},\mathcal{C}_K^{(L)},\mathcal{C}_K^{(R)}\}$

New set of representatives: Leaf-node means

RT-split criterion as OLS optimization

To split the parent node $\mathcal{C}_K$ next, the RT-split criterion is equivalent to the following optimization

$$(d^*, c^*, \hat{\beta}) = \arg\max_{d,c,\beta} \frac{1}{n} \left(\|y - Z^{(0)}\hat{\beta}^{(0)}\|_2^2 - \|y - Z\beta\|_2^2 \right)$$

$Z^{(0)} = (I(X_i \in C_j))$ and Z are membership matrices for the old and new set of nodes

New node representatives: $\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}$

DART-split criterion using GLS loss

Replace RT split criterion, a global OLS loss with *Dependency-adjusted Regression Tree (DART)-split criterion* a global GLS loss

$$\arg\max_{d,c,\beta} \frac{1}{n}\left[(y - Z^{(0)}\hat{\beta}^{(0)})^{\top} Q (y - Z^{(0)}\hat{\beta}^{(0)}) - (y - Z\beta)^{\top} Q (y - Z\beta)\right]$$

$Q = \Sigma^{-1}$ where Σ is the working covariance matrix (typically an estimate of the true covariance matrix Σ_0)

GLS-style regression tree

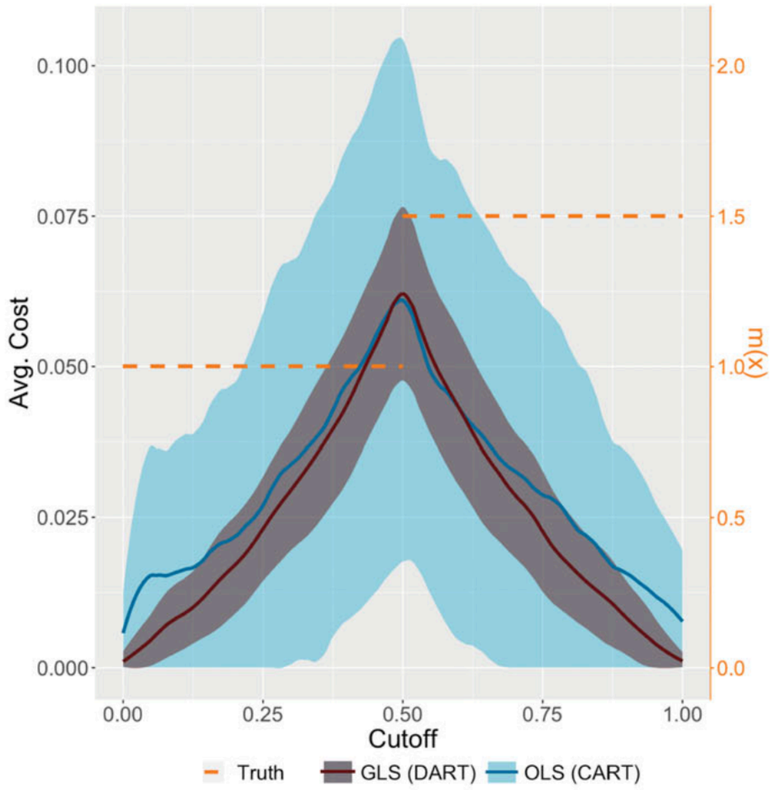
Maximize DART criterion to sequentially find the best directions (feature) d^* and cutoff c_d^*

New set of nodes: $\{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_{K-1}, \mathcal{C}_K^{(L)}, \mathcal{C}_K^{(R)}\}$

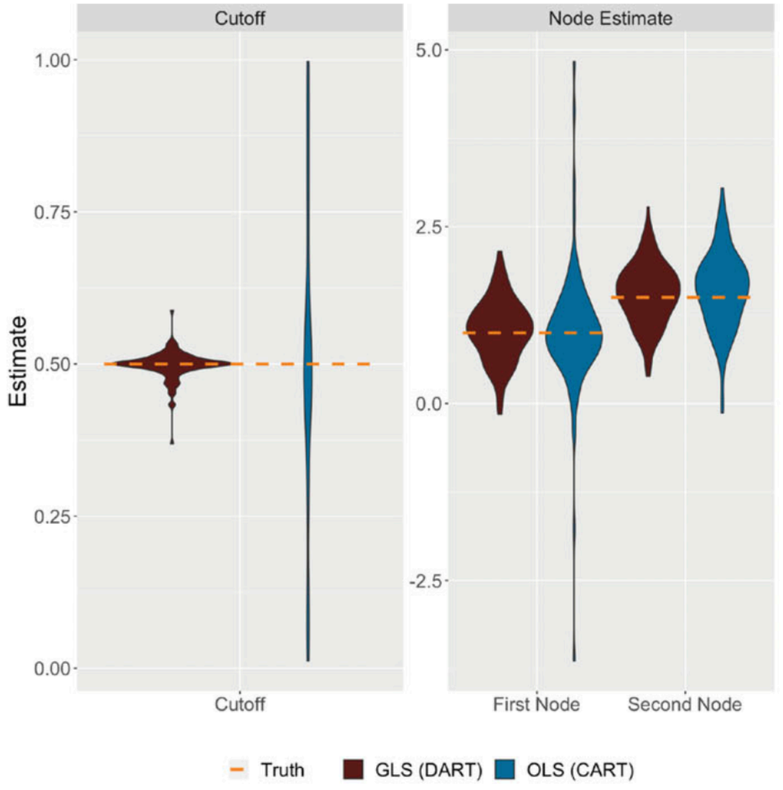
Leaf-node estimates: $\hat{\boldsymbol{\beta}}_{GLS}(\mathbf{Z}) = \hat{\boldsymbol{\beta}} = \left(\mathbf{Z}^\top \mathbf{Q} \mathbf{Z}\right)^{-1} \left(\mathbf{Z}^\top \mathbf{Q} \mathbf{y}\right)$

Global in nature: Both splitting of nodes and leaf-node estimates use correlation among all data points.

GLS vs OLS tree for dependent data



(a) CART and DART criterion



(b) Estimates of split-cutoff and node representatives.

Theory: DART criterion, although global in nature, converges to the same local variance difference as the CART criterion

DART is more efficient, yields better cut-off and leaf node estimates

Trees to forest

GLS loss with Y and Z is equivalent to OLS loss with $\tilde{Y} = \Sigma^{-1/2}Y$ (decorrelation) and $\tilde{Z} = \Sigma^{-1/2}Z$.

Immediate extension for resampling: Decorrelate, then resample

We essentially resample contrasts (prewhitened data)

- Example: For AR(1) data: $\Sigma^{-1/2}y = (y_1, y_2 - \rho y_1, y_3 - \rho y_2, \ldots)'$

RF-GLS summary

Dependency-adjusted Regression Tree (DART)-split criterion using GLS loss

Leaf node estimates: GLS estimates given the final set of splits

Resampling: Decorrelate, then resample

All 3 steps become identical to the RF algorithm using $Q = I$

- RF is a special case to RF-GLS with $Q = I$

Kriging with RF-GLS and Gaussian Process

For a linear model, $m(x) = x^\top \beta$, kriging (prediction) at a new location s_{new} is given by

$$\hat{y}_{new}(x_{new}, s_{new}) = x_{new}^\top \hat{\beta} + C(s_{new}, S)\Sigma^{-1}(y - X\hat{\beta})$$

$$\Sigma = C + \tau^2 I$$

For RF-GLS: Immediate extension to kriging by replacing $x^\top \hat{\beta}$ with $\widehat{m}(x)$.

Review of RF Theory

Data: $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^D, i = 1, \ldots, n$

Ann. Statist.
Volume 43, Number 4 (2015), 1716-1741.

Consistency of random forests

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert

Scornet et al. (2015):
If $Y_i = m(X_i) + \epsilon_i$, where the errors ϵ_i are iid,
Breiman's RF produces an $\mathbb{L}_2$ consistent estimator of m.

Separate line of theory by Wager and Athey (2018), S. Athey, J. Tibshirani, S. Wager, et al. (2019) - assume different data splits used in node creation and representation

No theory under dependent errors in either paradigms.

Theory for RF-GLS

If the errors $\epsilon \sim$ sub-Gaussian stationary β- (absolutely regular) mixing process, then under regularity conditions on the working precision matrix Q, RF-GLS is $\mathbb{L}_2$ consistent for m.

Theory for RF-GLS

Examples where the consistency holds:

Spatial Matérn GP with half-integer smoothness ν on a one-dimensional lattice

Theory for RF-GLS

Examples where the consistency holds:

Stationary autoregressive time-series with Gaussian errors.

Theory for RF-GLS

Examples where the consistency holds:

Classic RF

- $Q = I$, i.e., classic RF is $\mathbb{L}_2$ consistent under β-mixing dependence
  (to our knowledge, first result on consistency of RF for dependent errors)

- Analogous to OLS being consistent under dependence

Simulations

Data generation process:

$$Y_i = m(X_i) + w(s_i) + \epsilon^*(s_i);$$
$$w(\cdot) \sim GP(0, C(\cdot, \cdot));$$
$$\epsilon^*(s_i) \sim N(0, \tau^2)$$

Exponential GP: $C(s_i, s_j) = \sigma^2 \exp(-\phi \|s_i - s_j\|_2)$

Two choices of m:

1. $m(x) = 10 \sin(\pi x)$
2. $m(x) = 10(\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5)/6$ (Friedman function)

Simulations

Candidate methods:

Estimation:
RF, RF-GLS

Prediction:

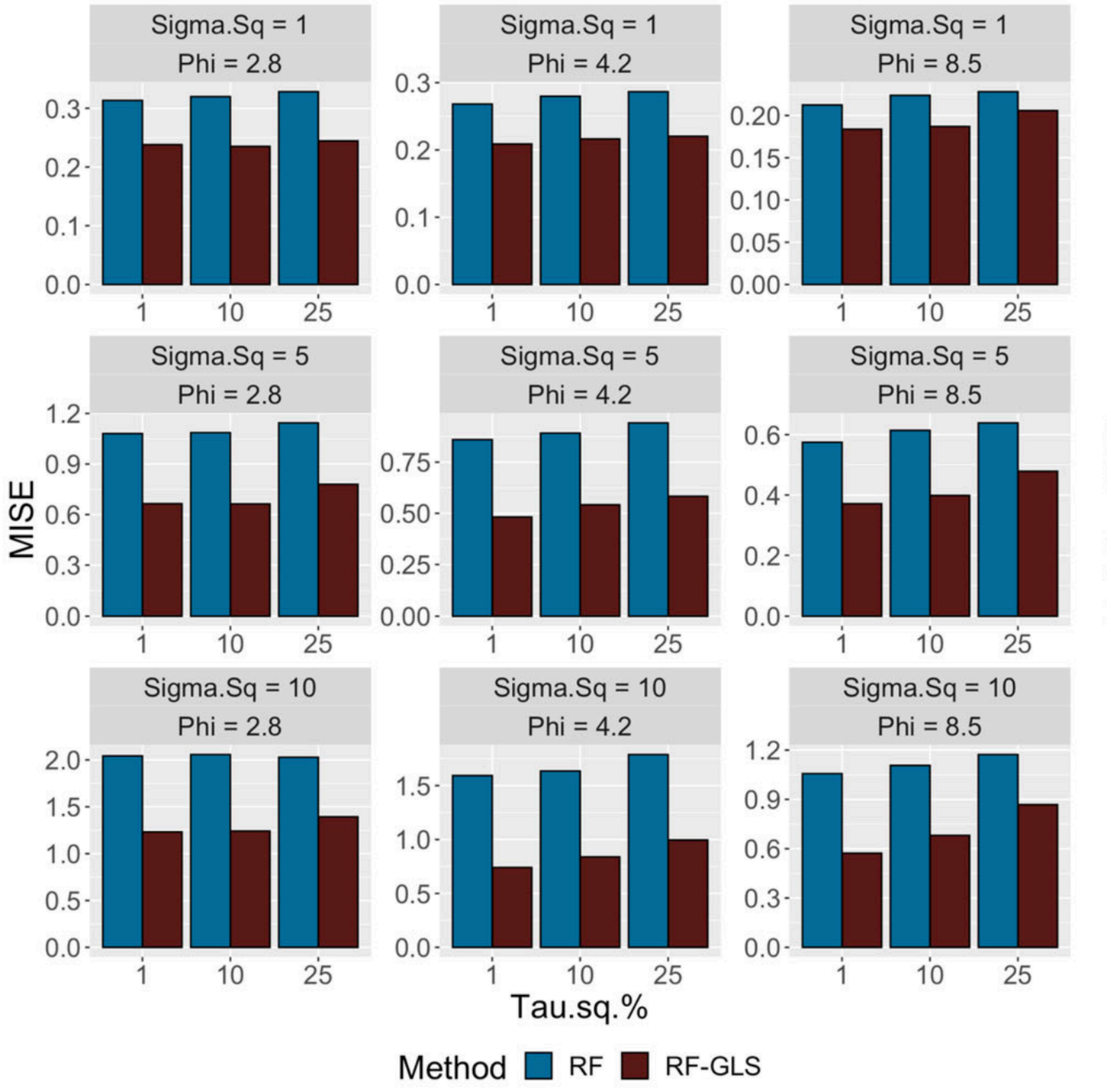
1. RF
2. RF-RK (Fox et al. 2020, Viscarra et al. 2014, Fayad et al. 2016): Random forest residual kriging
3. RF-GLS
4. RF-loc: Random forest with the locations added to the covariate list
5. RF-sp (Hengl et al. 2018): Random forest with all pairwise distances added to the covariate list

Simulations

Estimation performance

Comparison metric

$$\text{MISE} = \int (m(x) - \hat{m}(x))^2 dx$$

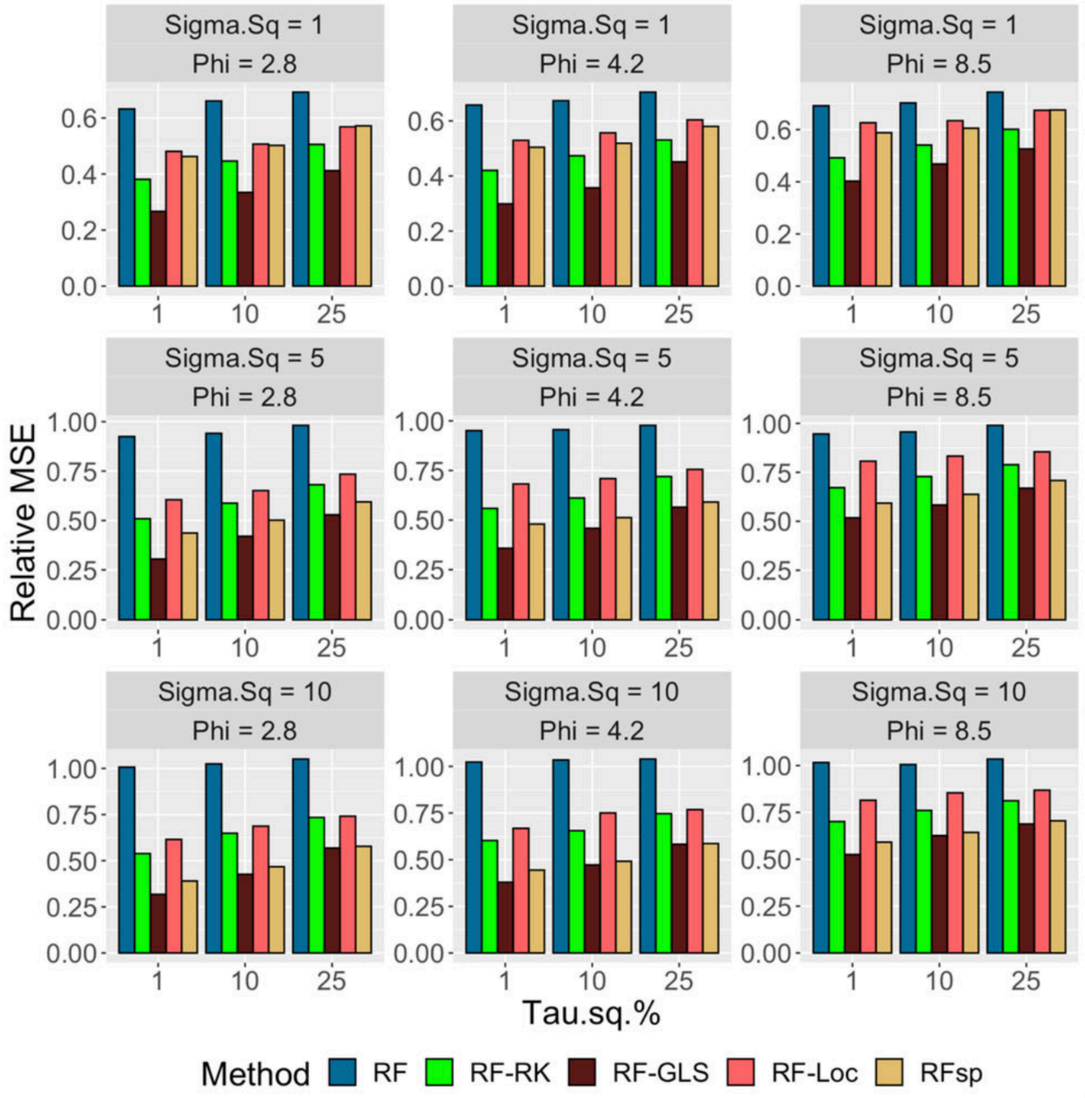


MISE when m = Friedman function

Simulations

Prediction performance

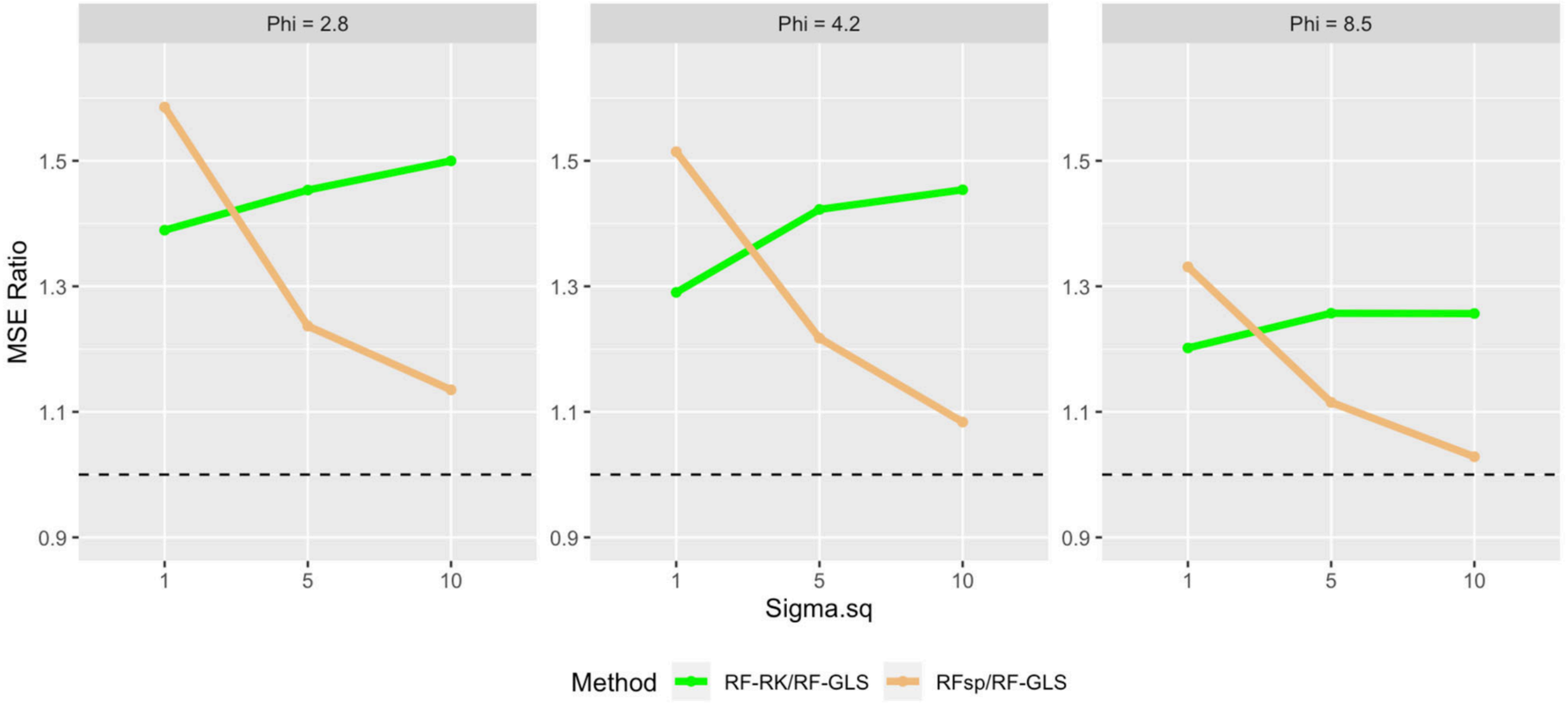
Comparison metric:
Mean Square Prediction Error (MSPE) on hold-out data



MSPE when m = Friedman function

Simulations

Prediction performance as a function of covariate signal to spatial noise ratio (SNR)

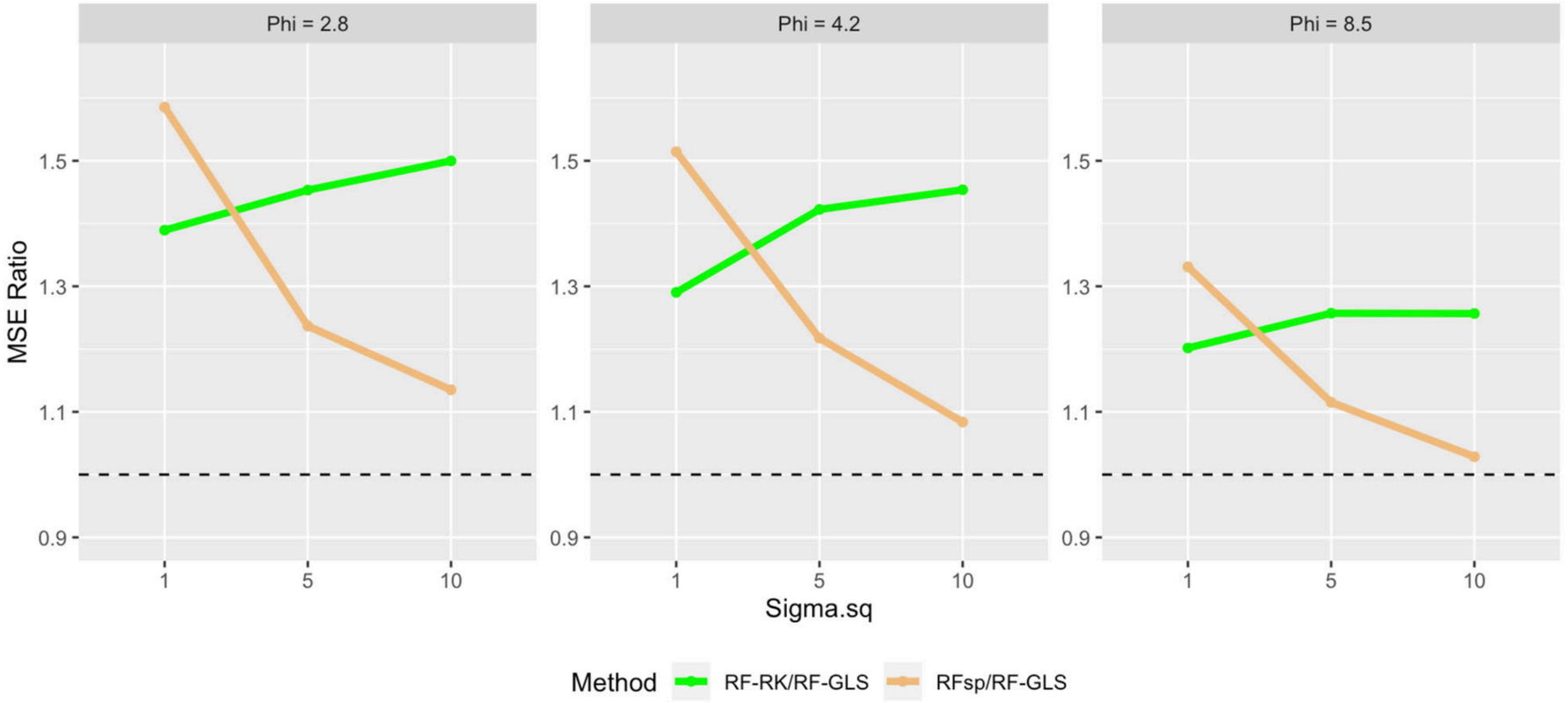


MSPE ratios for RF-RK and RFsp with respect to RF-GLS

Low SNR (large σ^2): RF-RK, ignoring spatial correlation during estimation, performs poorly

Simulations

Prediction performance as a function of covariate signal to spatial noise ratio (SNR)

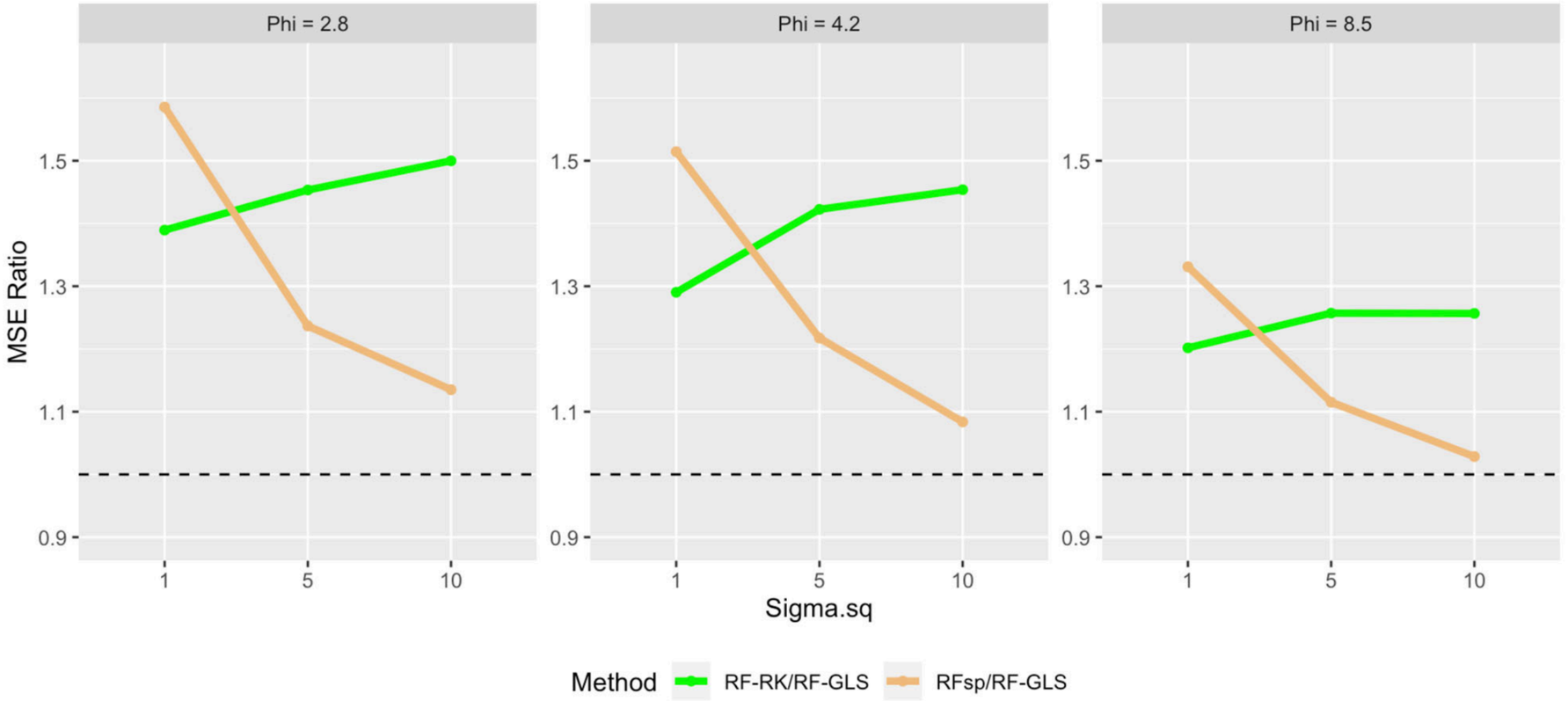


MSPE ratios for RF-RK and RFsp with respect to RF-GLS

High SNR (small σ^2) : RF-sp performs poorly as the high-dimensional set of spatial covariates drowns out the true covariate effect

Simulations

Prediction performance as a function of covariate signal to spatial noise ratio (SNR)



MSPE ratios for RF-RK and RFsp with respect to RF-GLS

RF-GLS performs well at both ends of the SNR spectrum by parsimoniously accounting for the spatial dependence using GP

Extension to binary spatial data (RF-GLMM)

Spatial generalized linear mixed model: $E(Y_i) = h(X_i^\top \beta + w(s_i))$

- h is a suitable link function
- same GP specification for spatial random effects w

Extension to binary spatial data (RF-GLMM)

Extension to non-linear covariate effect : $E(Y_i) = h(m(X_i) + w(s_i))$

- h is a suitable link function
- same GP specification for spatial random effects w

Extension to binary spatial data (RF-GLMM)

Extension to non-linear covariate effect : $E(Y_i) = h(m(X_i) + w(s_i))$

- h is a suitable link function
- same GP specification for spatial random effects w

Challenges in estimating m using RF while accounting for dependence

- Marginal model for Y not available in closed form
- For binary data, RF uses the *Gini impurity measure* for node splitting
  - no dependent analog of *Gini impurity*
- RF will estimate the mean function $p(x) = E(Y \mid X = x)$, it is different from the covariate effect $m(x)$ for non-linear h

Extension to binary spatial data (RF-GLMM)

GLMM with non-linear covariate effect : $E(Y_i) = h(m(X_i) + w(s_i))$

Main ideas:

- *Gini impurity measure split criterion* $\propto$ OLS split criterion
- Use the GLS split criterion to account for dependence to obtain leaf-node estimates $\widehat{p}(X)$
- Link-inversion to estimate $m(x)$ from $p(x)$ and predict at new locations via kriging as in GLMM

Theory:

Consistency of estimators of both mean function $p(x)$ and covariate effect $m(x)$.

Discussion

RF-GLS, a well principled extension of RF for dependent data
- same reasoning that replaces OLS with GLS for linear models under-dependence

GLS synergistically addresses issues of RF for dependent data
- accounts for data correlation in tree growing and leaf node estimates
- resampling of uncorrelated contrasts instead of correlated responses

General result for consistency of RF and RF-GLS under β-mixing error processes
- includes AR time-series and Matern GPs

Better finite sample performance of RF-GLS over RF for both estimation and prediction

Extension to binary data by connecting Gini measure to OLS loss, link inversion

Papers and Software

Joint work with Arkajyoti Saha (U. Washington) and Sumanta Basu (Cornell U.)

Saha, A., Basu, S., & Datta, A. (2021). Random forests for spatially dependent data. *Journal of the American Statistical Association*, 1-19.

Saha, A., Basu, S., & Datta, A. (2022). RandomForestsGLS: An R package for Random Forests for dependent data. *Journal of Open Source Software*, *7*(71), 3780.

Saha, A., & Datta, A. Random forests for binary spatial data (Preprint coming soon)

Work supported by NSF award DMS-1915803 and NIH National Institute of Environmental Health Sciences (NIEHS) grant R01 ES033739.

Thank you

abhidatta.com ✉abhidatta@jhu.edu @datta_science