Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data

Patrick Schratz[1], Jannes Muenchow[1], Jakob Richter[2], Alexander Brenning[1]

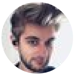
Kolloquium (Department of Statistics), LMU Munich, 20 Jun 2018

[1] Department of Geography, GIScience group, University of Jena
[2] Department of Statistics, TU Dortmund

https://pjs-web.de @pjs_228 @pat-s @pjs_228
patrick.schratz@uni-jena.de Patrick Schratz

Outline

1. Introduction
2. Data and study area
3. Methods
4. Results
5. Discussion



Patrick Schratz
@pjs_228

Slides of my upcoming talk at LMU Munich on Jun 20th: bit.ly/2M0Gqjv

10:06 PM - Jun 18, 2018

2 See Patrick Schratz's other Tweets

Introduction

Introduction

Whoami

- "Data Scientist/Analyst"
- B.Sc. **Geography** & M.Sc. **Geoinformatics** at University of Jena
- Self-taught programmer
- Interested in model optimization, R package development, server administration
- Arch Linux package maintainer
- PhD student (since 2016)

Contributions to `mlr`

- Integrated new sampling scheme for CV: Spatial sampling
- Redesigned the tutorial site (`mkdocs` -> `pkgdown`)
- Added getter for inner resampling indices
- more to come ;)

Introduction

LIFE Healthy Forest

Early detection and advanced management systems to reduce forest decline by invasive and pathogenic agents.

Main task: Spatial (modeling) analysis to support the early detection of various pathogens.

Pathogens

- Fusarium circinatum
- **Diplodia sapinea** (→ needle blight)
- Armillaria root disease
- Heterobasidion annosum

Fig. 1: Needle blight caused by **Diplodia pinea**

Introduction

Motivation

- Find the model with the **highest predictive performance**.
- Results are assumed to be representative for data sets with similar predictors and different pathogens (response).
- Be aware of **spatial autocorrelation** ⚠
- Analyze differences between spatial and non-spatial hyperparameter tuning (no research here yet!).
- Analyze differences in performance between algorithms and sampling schemes in CV (both performance estimation and hyperparameter tuning)

Data & Study Area

Data & Study Area

```
## Skim summary statistics
##  n obs: 926 
##  n variables: 12 
## 
## Variable type: factor
## 
##  variable     missing     n     n_unique                    top_counts                
## -----------  ---------  -----  ----------  ---------------------------------------------
##   diplo01        0       926       2                 0: 703, 1: 223, NA: 0          
##  lithology       0       926       5         clas: 602, chem: 143, biol: 136, surf: 32 
##    soil          0       926       7          soil: 672, soil: 151, soil: 35, pron: 22 
##    year          0       926       4         2009: 401, 2010: 261, 2012: 162, 2011: 102
## 
## Variable type: numeric
## 
##    variable       missing     n      mean       p0       p50       p100       hist   
## ---------------  ---------  -----  --------  -------  --------  --------  ----------
##       age            0       926    18.94       2        20        40              
##    elevation         0       926    338.74    0.58     327.22    885.91            
##    hail_prob         0       926     0.45     0.018     0.55       1               
##      p_sum           0       926    234.17    124.4    224.55    496.6             
##       ph             0       926     4.63     3.97      4.6       6.02             
##      r_sum           0       926    -4e-05    -0.1     0.0086    0.082             
##  slope_degrees       0       926    19.81     0.17     19.47     55.11             
##      temp            0       926    15.13     12.59    15.23      16.8             
```

Data & Study Area

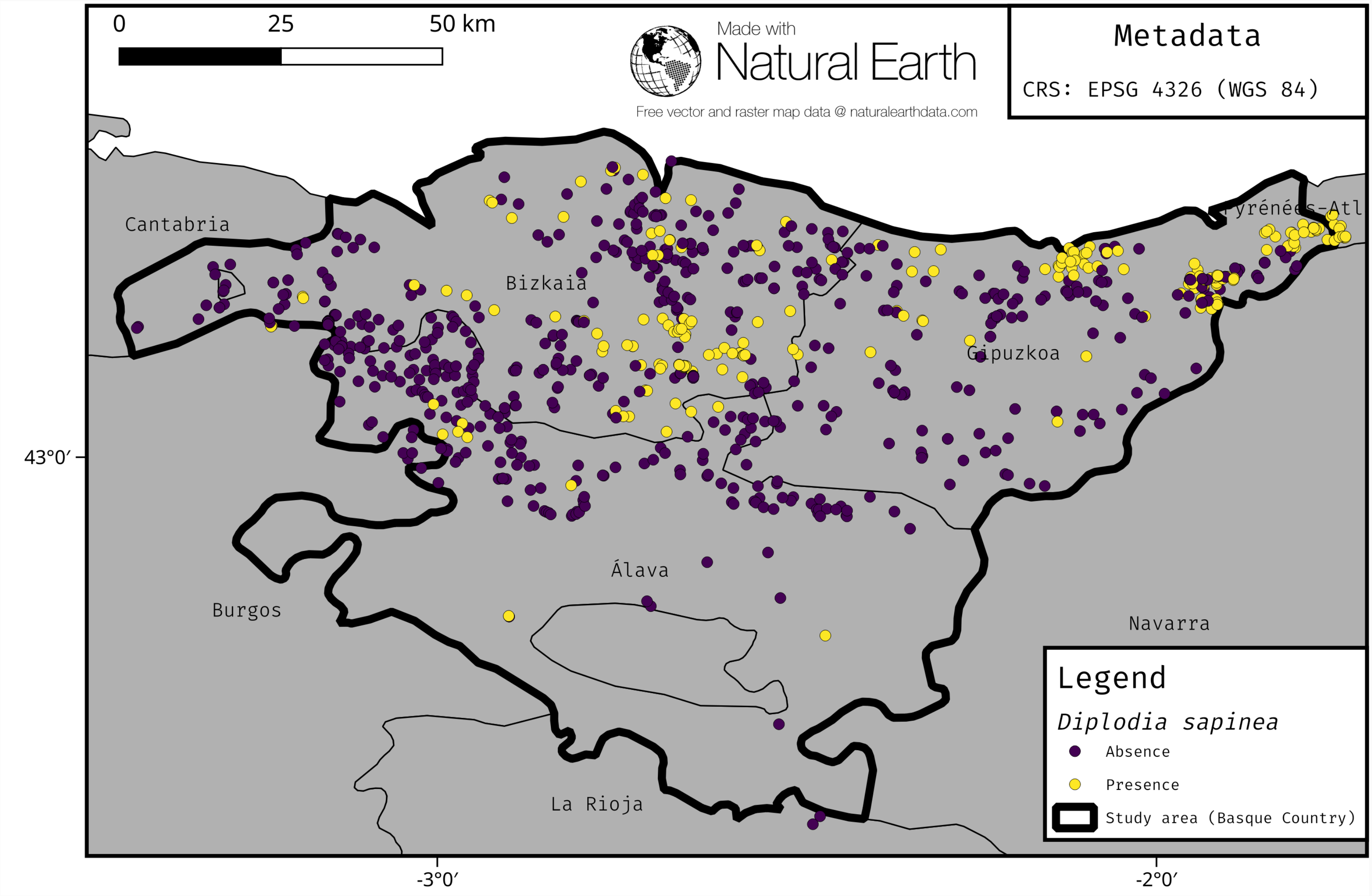


Fig. 2: Study area (Basque Country, Spain)

Methods

Methods

Machine-learning models

- Boosted Regression Trees (`BRT`)
- Random Forest (`RF`)
- Support Vector Machine (`SVM`)
- k-nearest Neighbor (`KNN`)

Parametric models

- Generalized Addtitive Model (`GAM`)
- Generalized Linear Model (`GLM`)

Performance Measure

Brier Score

Nested Cross-Validation

- Cross-validation for **performance estimation**
- Cross-validation for **hyperparameter tuning** (sequential model based optimization, Bischl, Richter, Bossek, Horn, Thomas, and Lang (2017))

Different sampling strategies (Performance estimation/Tuning):

- Non-Spatial/Non-Spatial
- Spatial/Non-Spatial
- Spatial/Spatial
- Non-Spatial/No Tuning
- Spatial/No Tuning

Methods

Nested (spatial) cross-validation

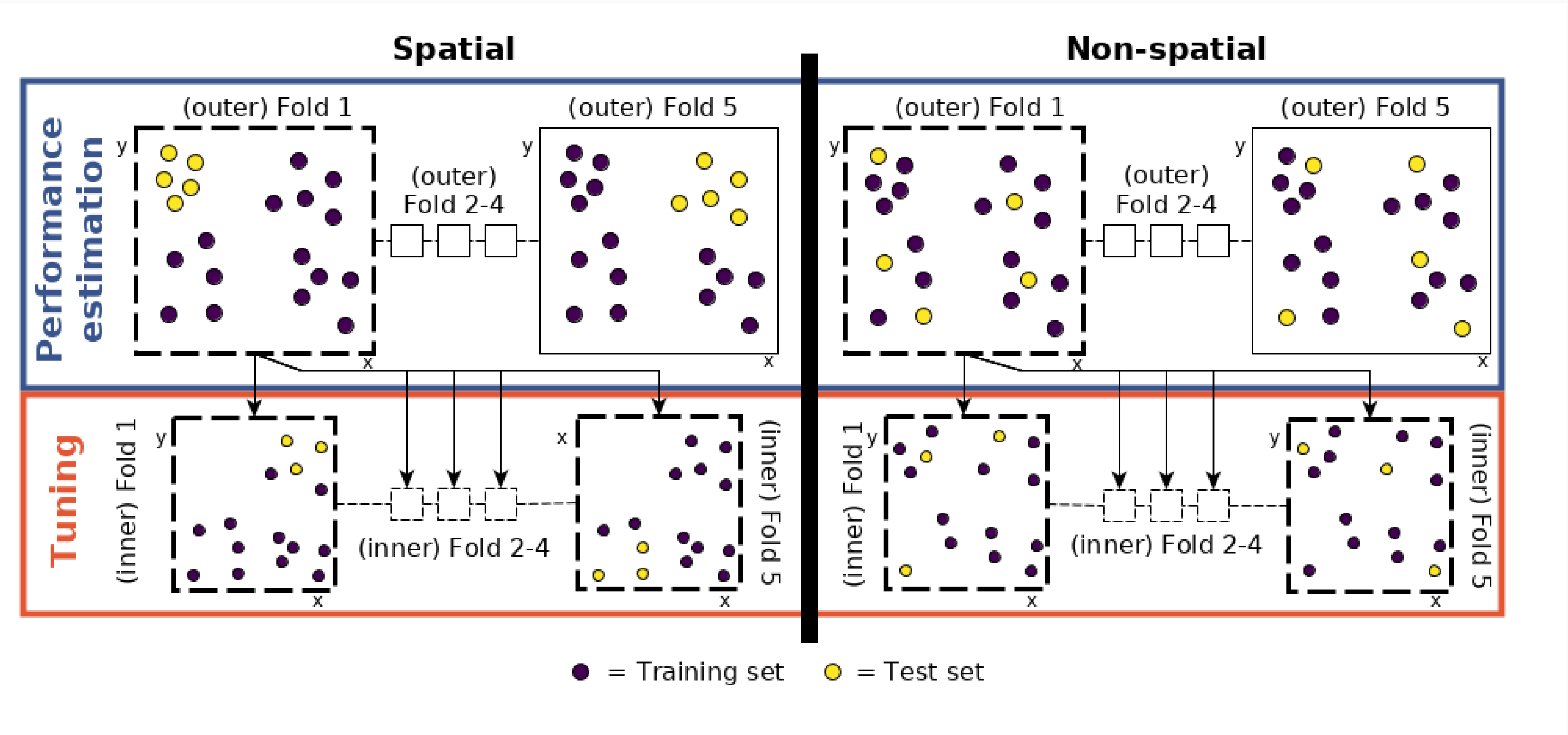


Fig. 3: Nested spatial/non-spatial cross-validation

Nested (spatial) cross-validation

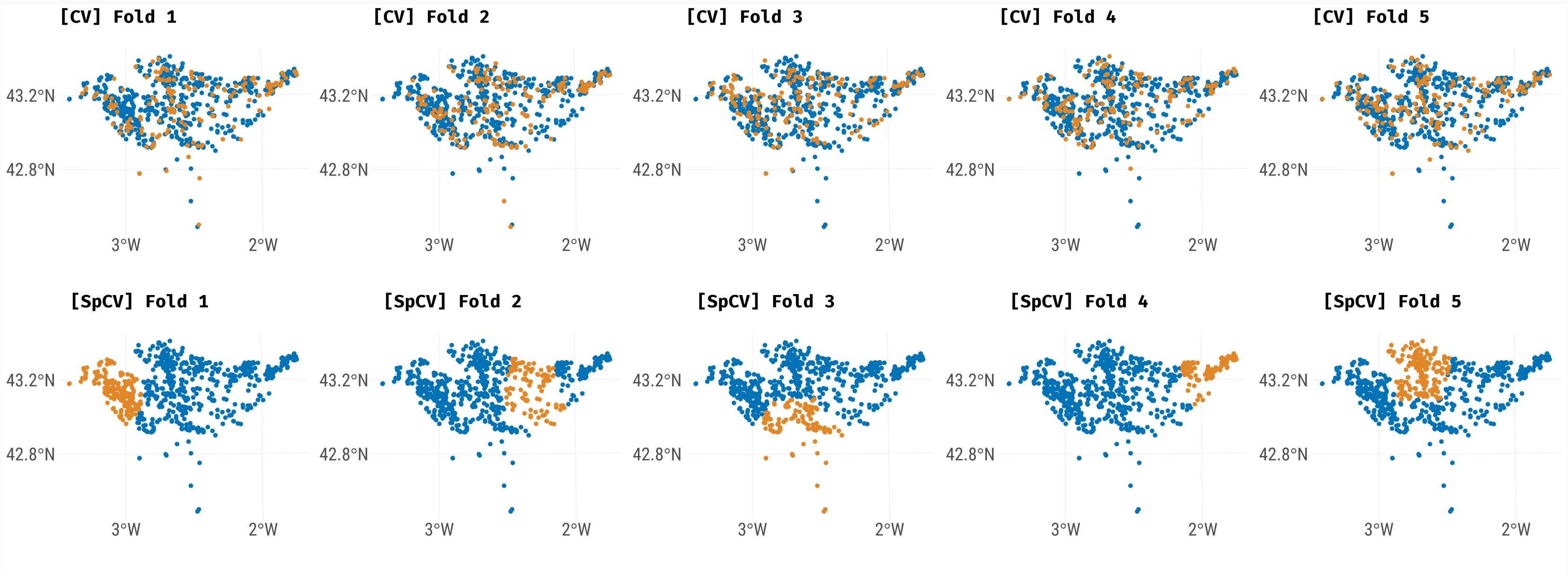


Fig. 4: Comparison of spatial and non-spatial partitioning of the data set.

Methods

Hyperparameter tuning search spaces

RF : Probst, Wright, and Boulesteix (2018)

BRT, SVM, KNN: Self-defined limits based on evaluation of estimated hyperparameters

Algorithm (package)	Hyperparameter	Type	Value	Start	End	Default
BRT (gbm)	`n.tree`	integer	-	100	15000	100
	`shrinkage`	numeric	-	0	1.0	0.001
	`interaction.depth`	integer	-	1	20	1
RF (ranger)	`mtry`	integer	-	1	11	$\sqrt{p}$
	`min.node.size`	integer	-	1	10	1
	`sample.fraction`	numeric	-	0.2	0.9	1
SVM (kernlab)	`C`	numeric	-	2^{-15}	2^{15}	1
	σ	numeric	-	2^{-15}	2^{15}	1
KNN (kknn)	`k`	integer	-	1	250	7
	`distance`	integer	-	1	300	2

Table 1: Hyperparameter limits and types of each model.
Notations of hyperparameters from the respective R packages were used.
p = Number of variables.

Results

Hyperparameter tuning

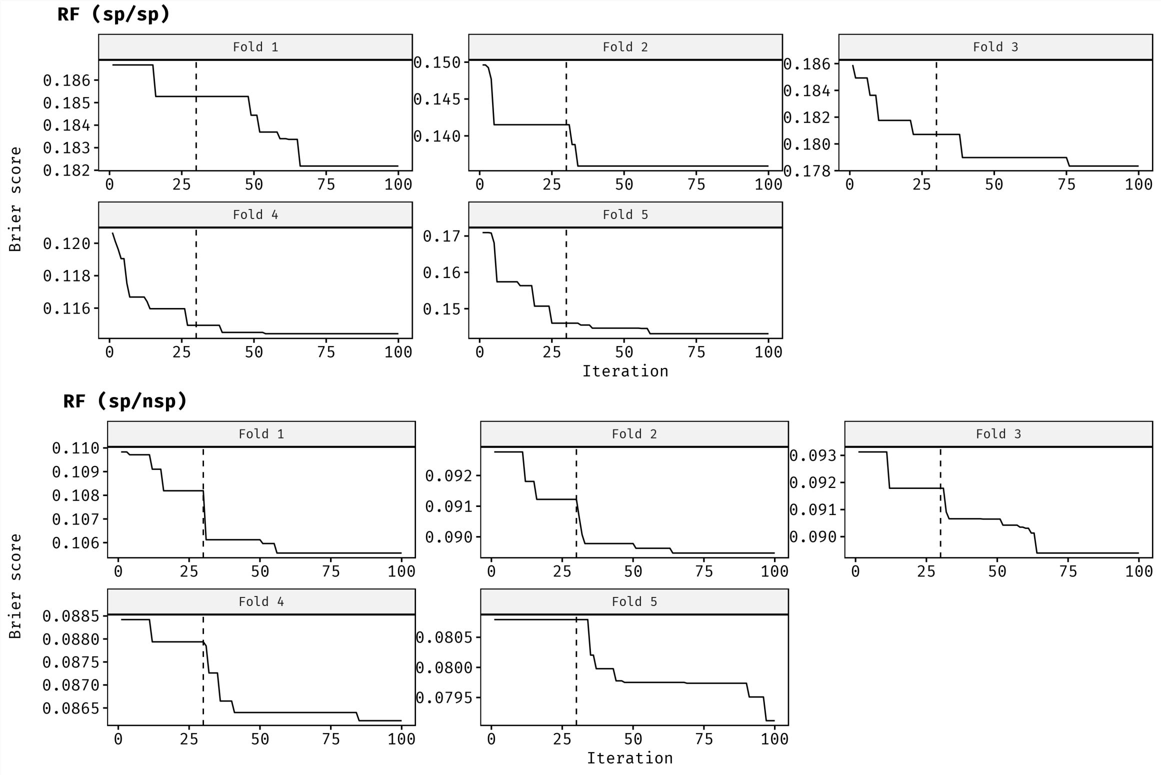


Fig 4: SMBO optimization paths of the first five folds of the **spatial/spatial** and **spatial/non-spatial** CV setting for RF. The dashed line marks the border between the initial design (30 randomly composed hyperparameter settings) and the sequential optimization part in which each setting was proposed using information from the prior evaluated settings.

Results

Hyperparameter tuning

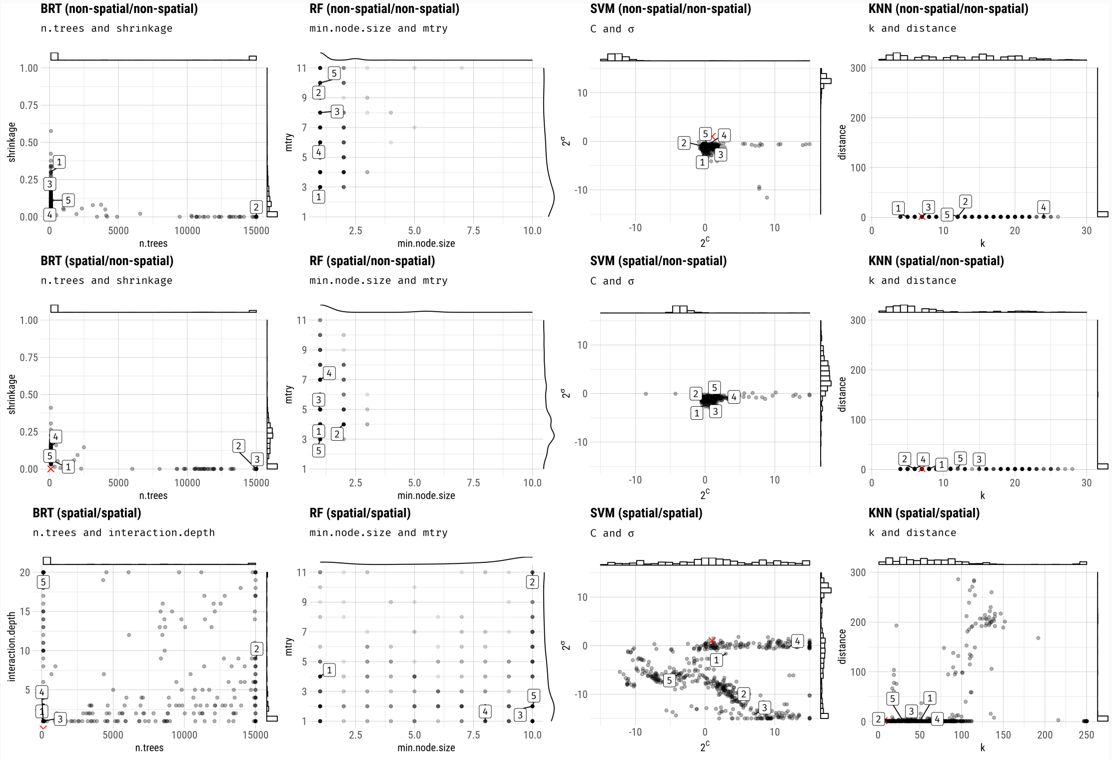


Fig 5: Best hyperparameter settings by fold (500 total) each estimated from 100 (30/70) SMBO tuning iterations per fold using five-fold cross-validation. Split by spatial and non-spatial partitioning setup and model type. Red crosses indicate the default hyperparameters of the respective model. Black dots represent the winning hyperparameter setting of each fold. The labels ranging from one to five show the winning hyperparameter settings of the first five folds.

Predictive Performance

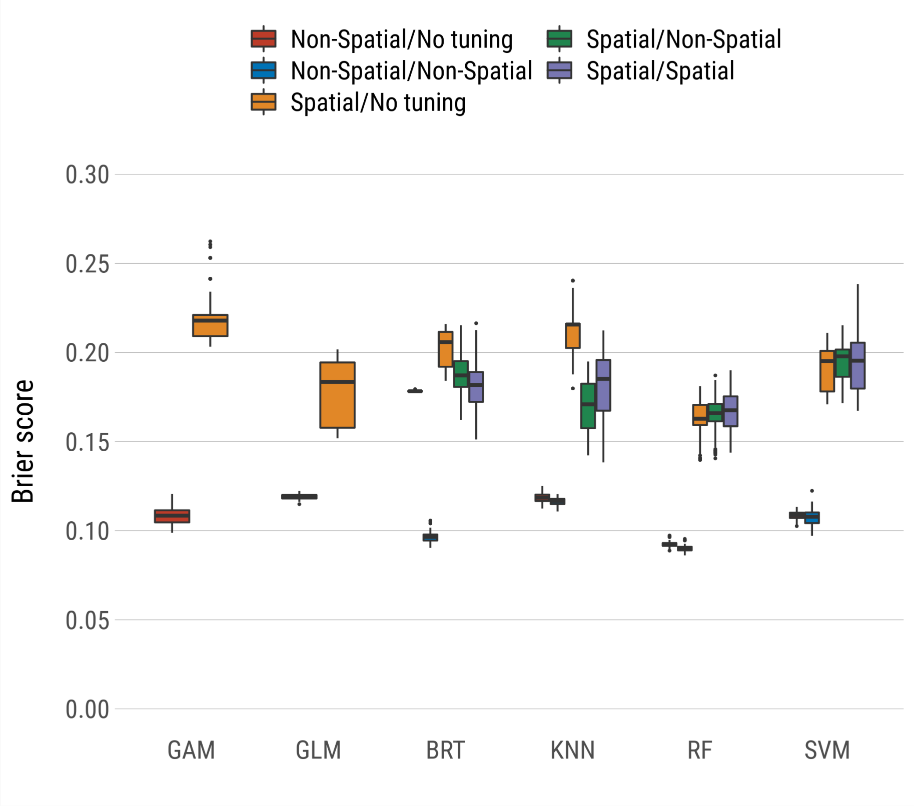


Fig 6: (Nested) CV estimates of model performance at the repetition level using 100 SMBO iterations for hyperparameter tuning. CV setting refers to performance estimation/hyperparameter tuning of the respective (nested) CV, e.g. "Spatial/Non-Spatial" means that spatial partitioning was used for performance estimation and non-spatial partitioning for hyperparameter tuning.

Discussion

Discussion

Predictive performance

- `RF` showed the best predictive performance

Discussion

Predictive performance

- `RF` showed the best predictive performance
- High bias in performance when using non-spatial CV

Discussion 💬 (Performance)

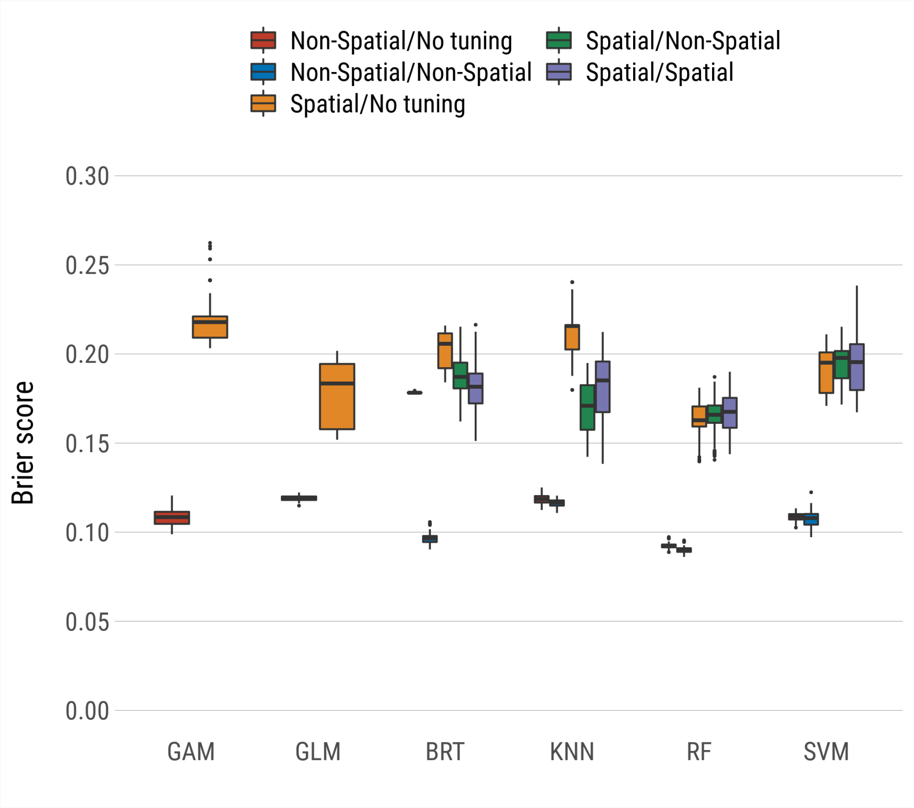


Fig 6: (Nested) CV estimates of model performance at the repetition level using 100 SMBO iterations for hyperparameter tuning. CV setting refers to performance estimation/hyperparameter tuning of the respective (nested) CV, e.g. "Spatial/Non-Spatial" means that spatial partitioning was used for performance estimation and non-spatial partitioning for hyperparameter tuning.

Discussion

Predictive Performance

- `RF` showed the best predictive performance
- High bias in performance when using non-spatial CV

Discussion

Predictive Performance

- `RF` showed the best predictive performance
- High bias in performance when using non-spatial CV
- The `GLM` shows an equally good performance as BRT, KNN and SVM

Discussion

Predictive Performance

- `RF` showed the best predictive performance
- High bias in performance when using non-spatial CV
- The `GLM` shows an equally good performance as BRT, KNN and SVM
- The `GAM` suffers from overfitting

Hyperparameter tuning

- Almost no effect on predictive performance.

Hyperparameter tuning

- Almost no effect on predictive performance.
- Differences between algorithms are higher than the effect of hyperparameter tuning

Hyperparameter tuning

- Almost no effect on predictive performance.
- Differences between algorithms are higher than the effect of hyperparameter tuning
- Spatial hyperparameter tuning has no substantial effect on predictive performance compared to non-spatial tuning

Hyperparameter tuning

- Almost no effect on predictive performance.
- Differences between algorithms are higher than the effect of hyperparameter tuning
- Spatial hyperparameter tuning has no substantial effect on predictive performance compared to non-spatial tuning
- Optimal parameters estimated from spatial hyperparameter tuning show a wide spread across the search space

Discussion

Tuning

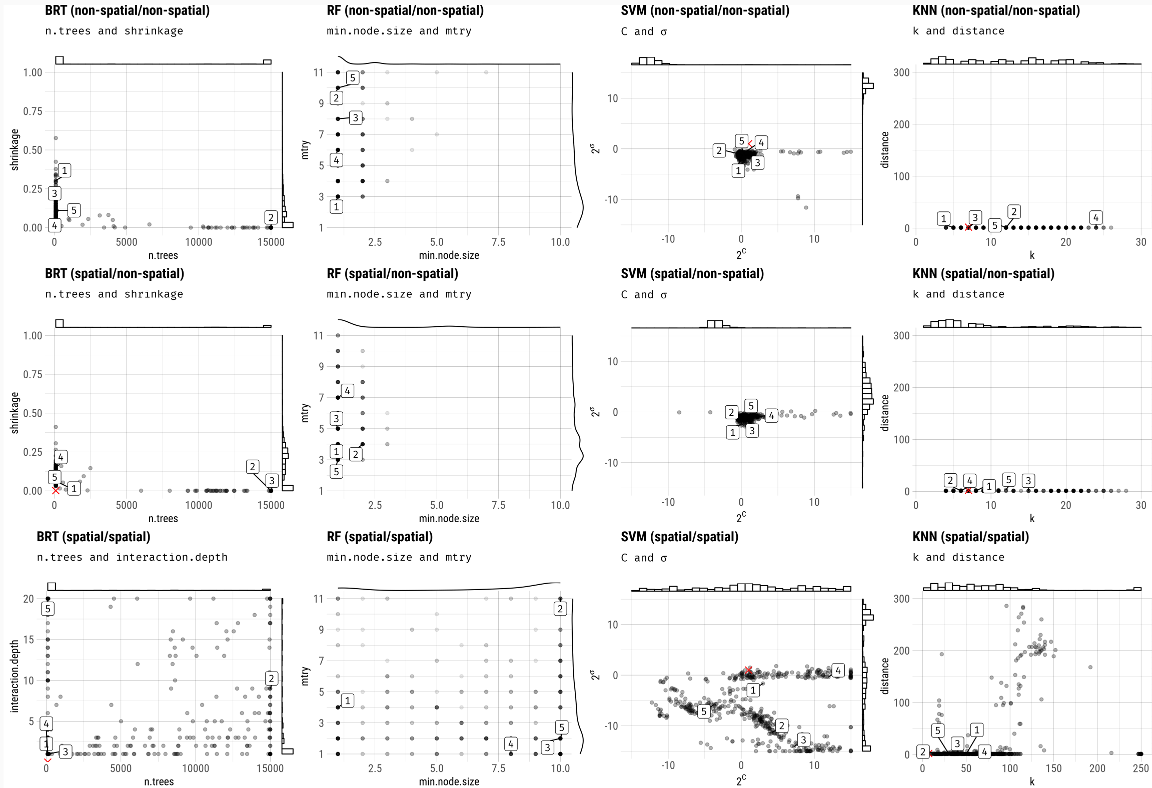


Fig 5: Best hyperparameter settings by fold (500 total) each estimated from 100 (30/70) SMBO tuning iterations per fold using five-fold cross-validation. Split by spatial and non-spatial partitioning setup and model type. Red crosses indicate the default hyperparameters of the respective model. Black dots represent the winning hyperparameter setting of each fold. The labels ranging from one to five show the winning hyperparameter settings of the first five folds.

Hyperparameter tuning

- Almost no effect on predictive performance.
- Differences between algorithms are higher than the effect of hyperparameter tuning.
- Spatial hyperparameter tuning has no substantial effect on predictive performance compared to non-spatial tuning.
- Optimal parameters estimated from spatial hyperparameter tuning show a wide spread across the search space.

Spatial hyperparameter tuning should be used for spatial data sets to have a consistent resampling scheme.

References

Bischl, B, J. Richter, J. Bossek, et al. (2017). "mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions". In: *ArXiv e-prints*. arXiv: 1703.03373 [stat].

Probst, P, M. Wright and A. Boulesteix (2018). "Hyperparameters and Tuning Strategies for Random Forest". In: *ArXiv e-prints*. arXiv: 1804.03515 [stat.ML].

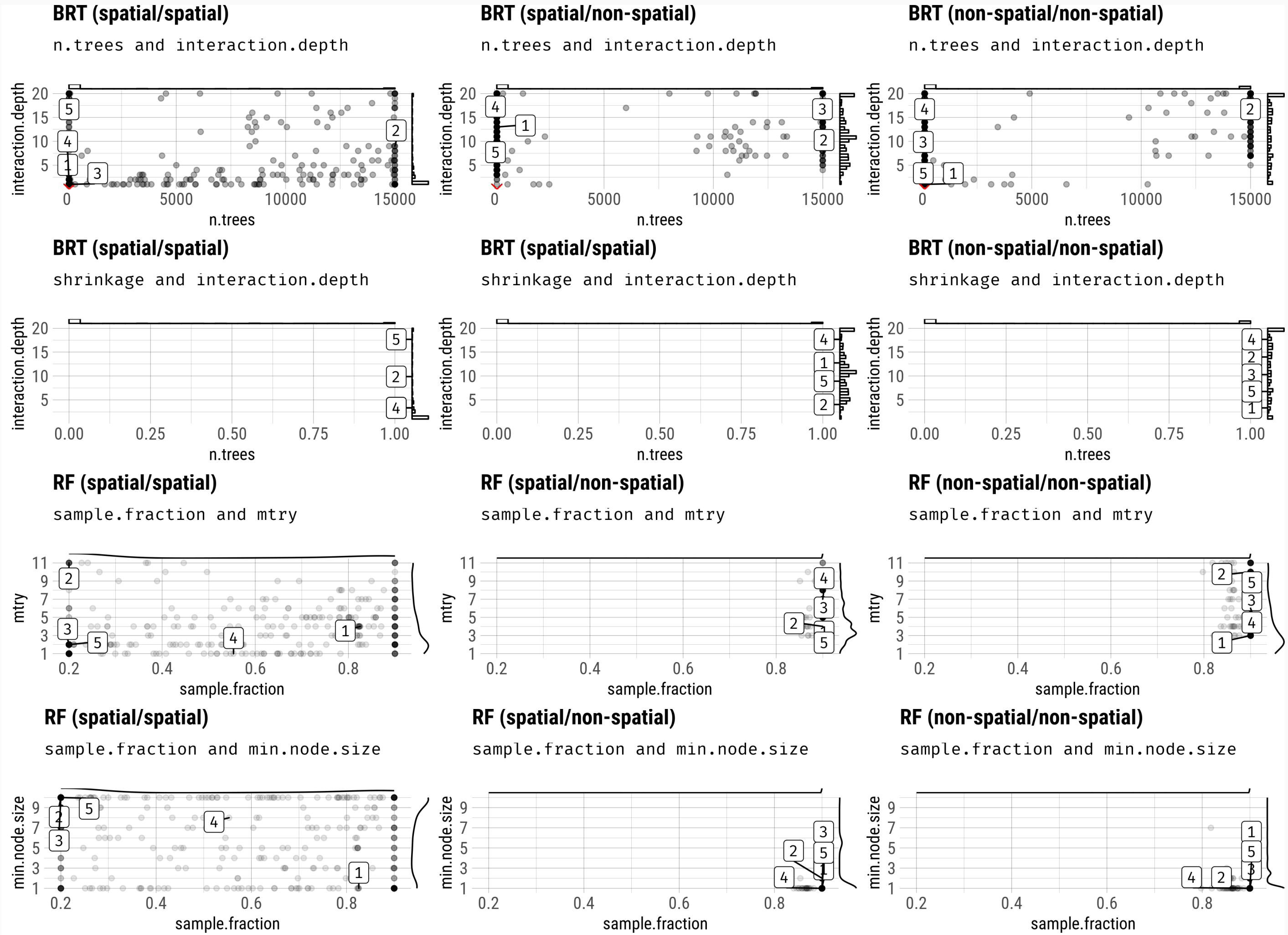
Thanks for listening!

Questions? Slides can be found here: https://t.co/SyWRky6sGn

And now, let's have a 🍺 ;)

Backup ➕

Backup


BRT (spatial/spatial)
n.trees and interaction.depth
BRT (spatial/non-spatial)
n.trees and interaction.depth
BRT (non-spatial/non-spatial)
n.trees and interaction.depth
BRT (spatial/spatial)
shrinkage and interaction.depth
BRT (spatial/spatial)
shrinkage and interaction.depth
BRT (non-spatial/non-spatial)
shrinkage and interaction.depth
RF (spatial/spatial)
sample.fraction and mtry
RF (spatial/non-spatial)
sample.fraction and mtry
RF (non-spatial/non-spatial)
sample.fraction and mtry
RF (spatial/spatial)
sample.fraction and min.node.size
RF (spatial/non-spatial)
sample.fraction and min.node.size
RF (non-spatial/non-spatial)
sample.fraction and min.node.size
interaction.depth
n.trees
mtry
sample.fraction
min.node.size