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]

GIScience Seminar Series, Jena, 14 Feb 2018

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

https://pat-s.github.io @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

Introduction

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 and prediction to other areas.

Pathogens

- Fusarium circinatum
- **Diplodia pinea** (➔ needle blight)
- Armillaria root disease
- Heterobasidion annosum

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

Motivation

- Find the model with the **highest predictive performance** for our data set.
- Results are assumed to be representative for data sets with similar predictors and different pathogens as response.
- Be aware of **spatial autocorrelation** ⚠
- Conduct "optimal" hyperparameter tuning for machine-learning models.
- Show and analyze differences in performances between spatial cross-validation and non-spatial cross-validation.

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      median   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
## ph              0         926   4.63       3.97    4.6      6.02
## p_sum           0         926   234.17     124.4   224.55   496.6
## r_sum           0         926   -0.00004   -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

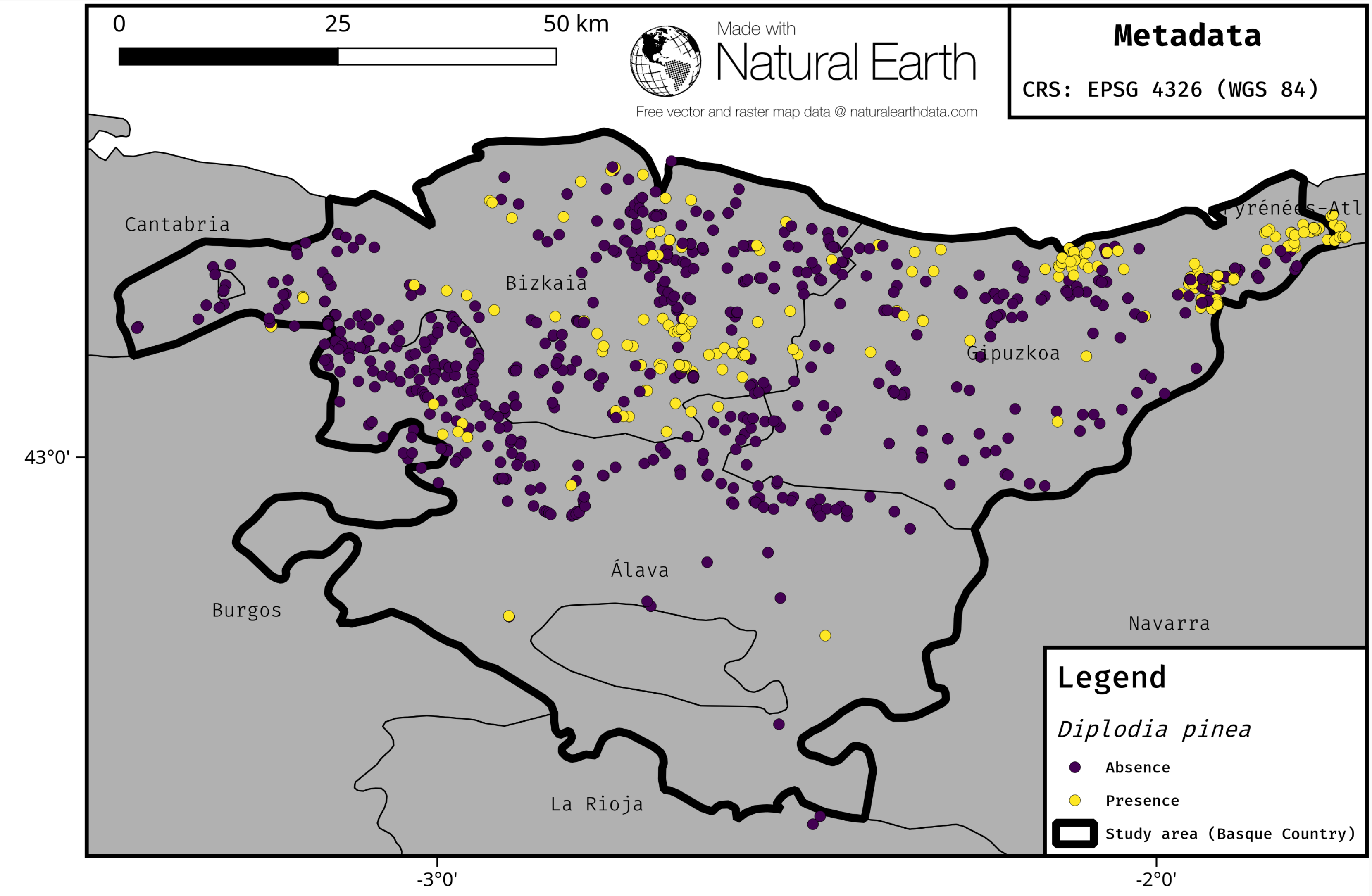


Fig. 2: Study area (Basque Country, Spain)

Methods

Methods

Machine-learning models

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

Parametric models

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

Performance Measure

Area under the Receiver Operating Curve (AUROC)

Nested Cross-Validation

- Cross-validation for **performance estimation** *[outer level]*
- Cross-validation for **hyperparameter tuning** (random search) *[inner level]*

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

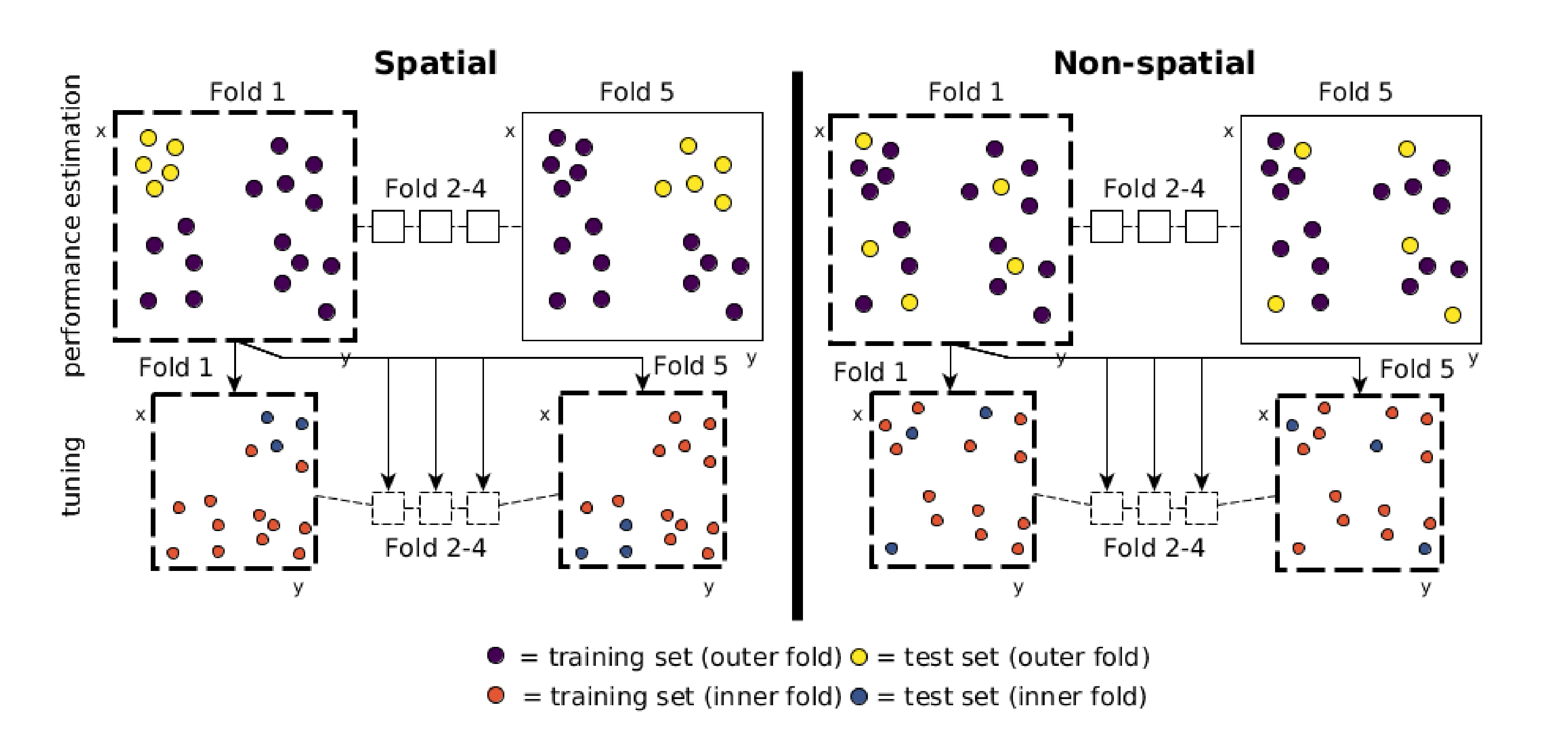


Fig. 3: Nested spatial/non-spatial cross-validation

Methods

Nested (spatial) Cross-Validation

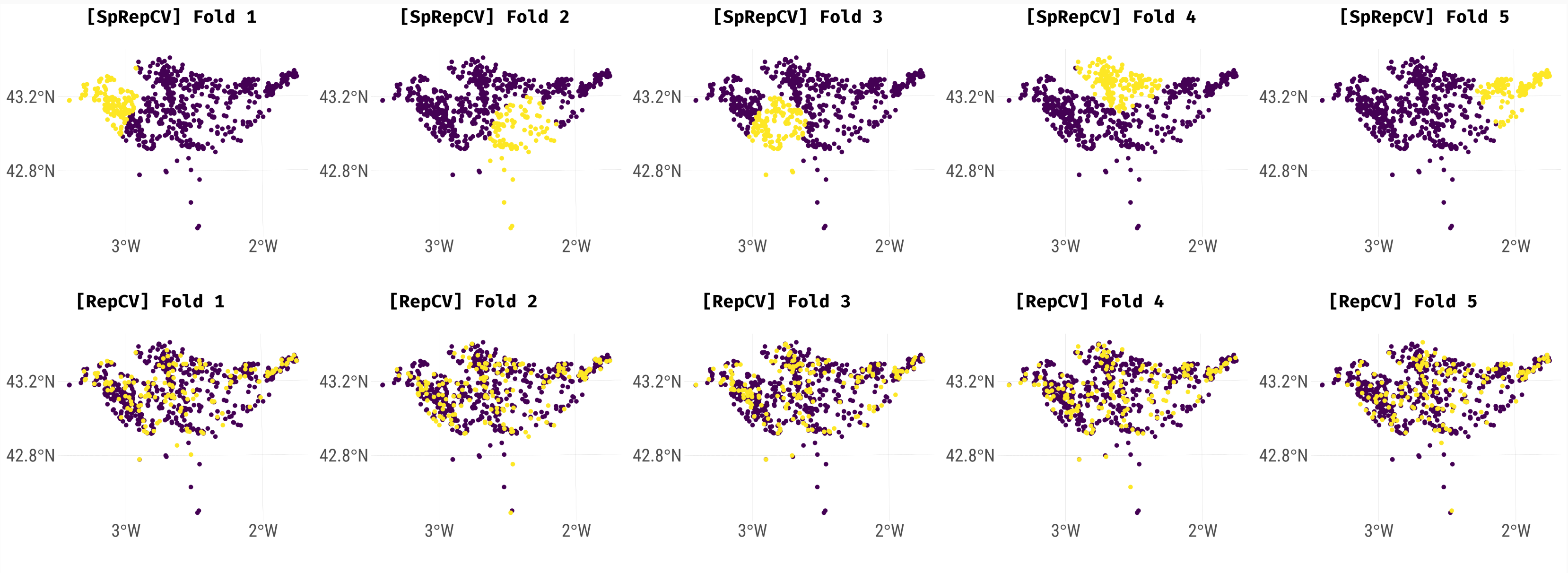


Fig. 4: Comparison of spatial and non-spatial partitioning of the data set.

Hyperparameter tuning

Random search has desirable properties in high dimensional and no disadvantages in low dimensional situations compared to **grid search** (Bergstra & Bengio, 2012).

Algorithm (package)	Hyperparameter	Type	Value	Start	End
BRT (gbm)	`n.tree`	integer	-	100	10000
	`shrinkage`	numeric	-	0	1.5
	`interaction.depth`	integer	-	1	40
RF (ranger)	`mtry`	integer	-	1	11
	`num.trees`	integer	-	10	10000
SVM (kernlab)	`C`	numeric	-	2^{-12}	2^{15}
	σ	numeric	-	2^{-15}	2^{6}
WKNN (kknn)	`k`	integer	-	10	400
	`distance`	integer	-	1	100
	`kernel`	nominal	*		

* triangular, Epanechnikov, biweight, triweight, cos, inv, Gaussian, optimal

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

Results

Results

Hyperparameter tuning

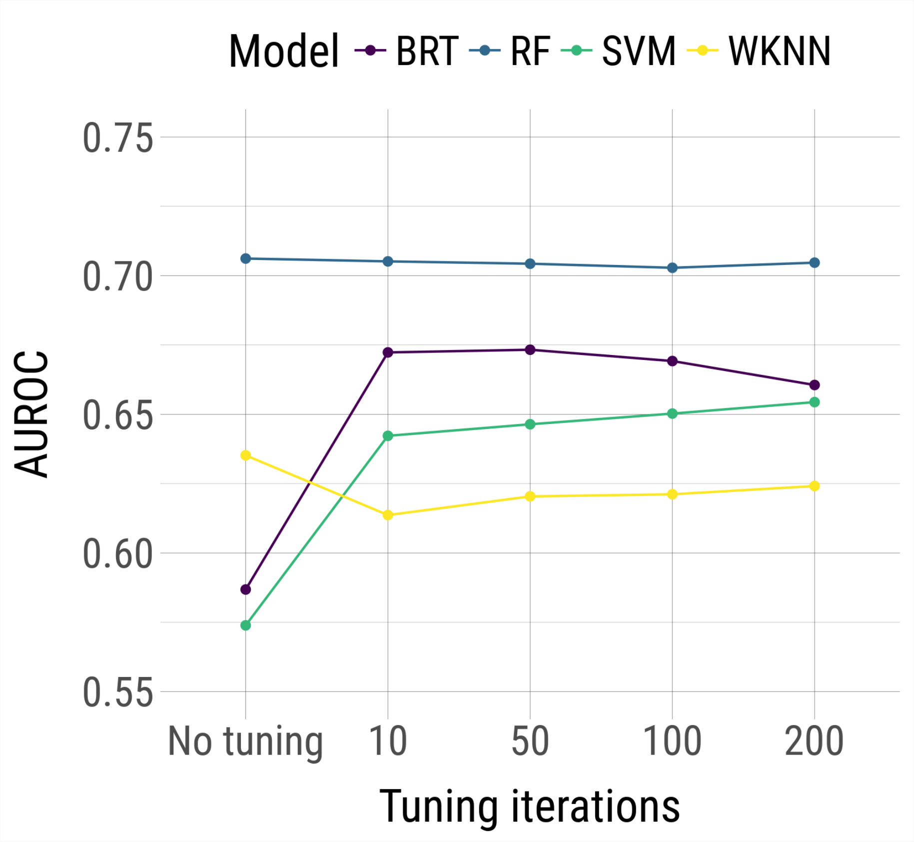


Fig 4: Hyperparameter tuning results of the spatial/spatial CV setting for BRT, WKNN, RF and SVM: Number of tuning iterations (1 iteration = 1 random hyperparameter setting) vs. predictive performance (AUROC).

Results (Predictive Performance)

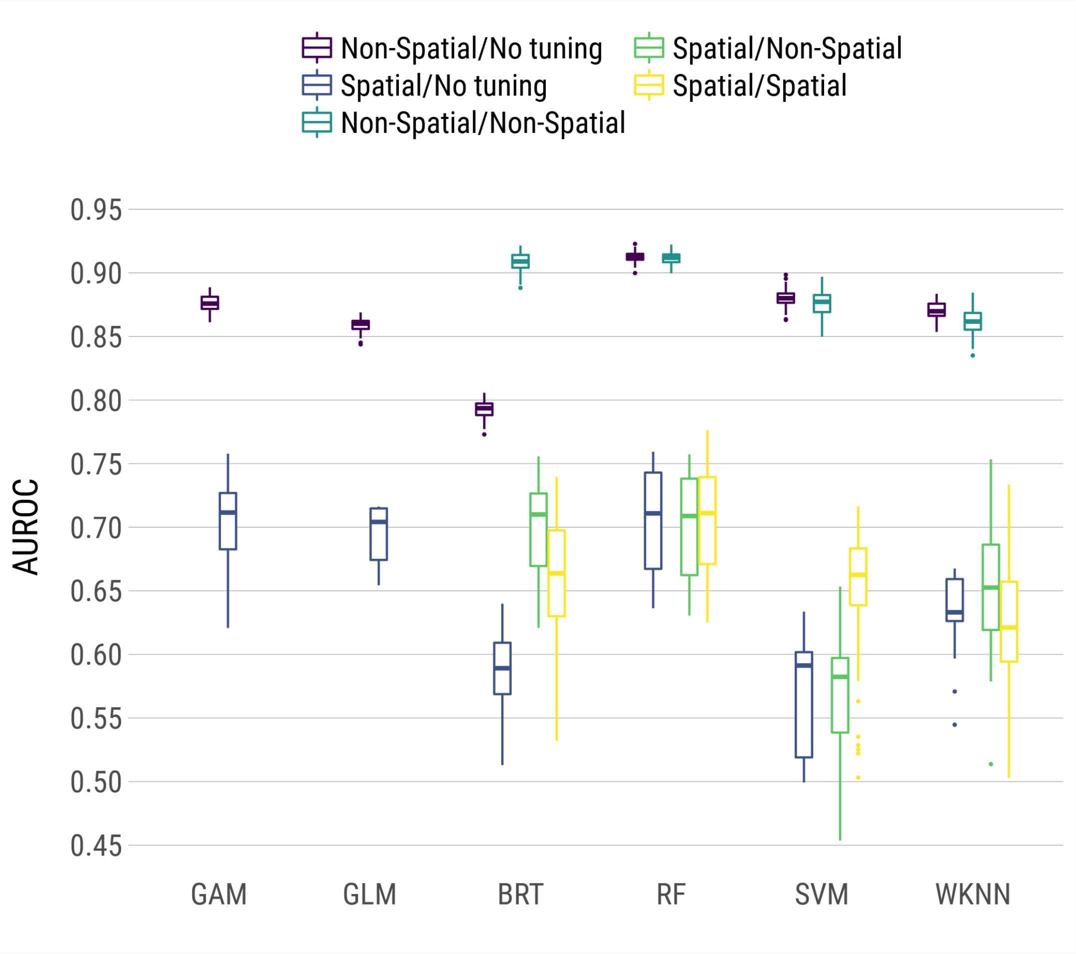


Fig 5: (Nested) CV estimates of model performance at the repetition level using 200 random search iterations. CV setting refers to perfomance estimation/hyperparameter tuning of the respective (nested) CV, e.g. ”Spatial/Non-Spatial” means that spatial partitioning was used for performance estimation and non-spatial partitioning for hyperparameter tuning.

Discussion

Discussion

Predictive performance

- `RF` and `GAM` showed the best predictive performance

Discussion

Predictive performance

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

Discussion (Performance)

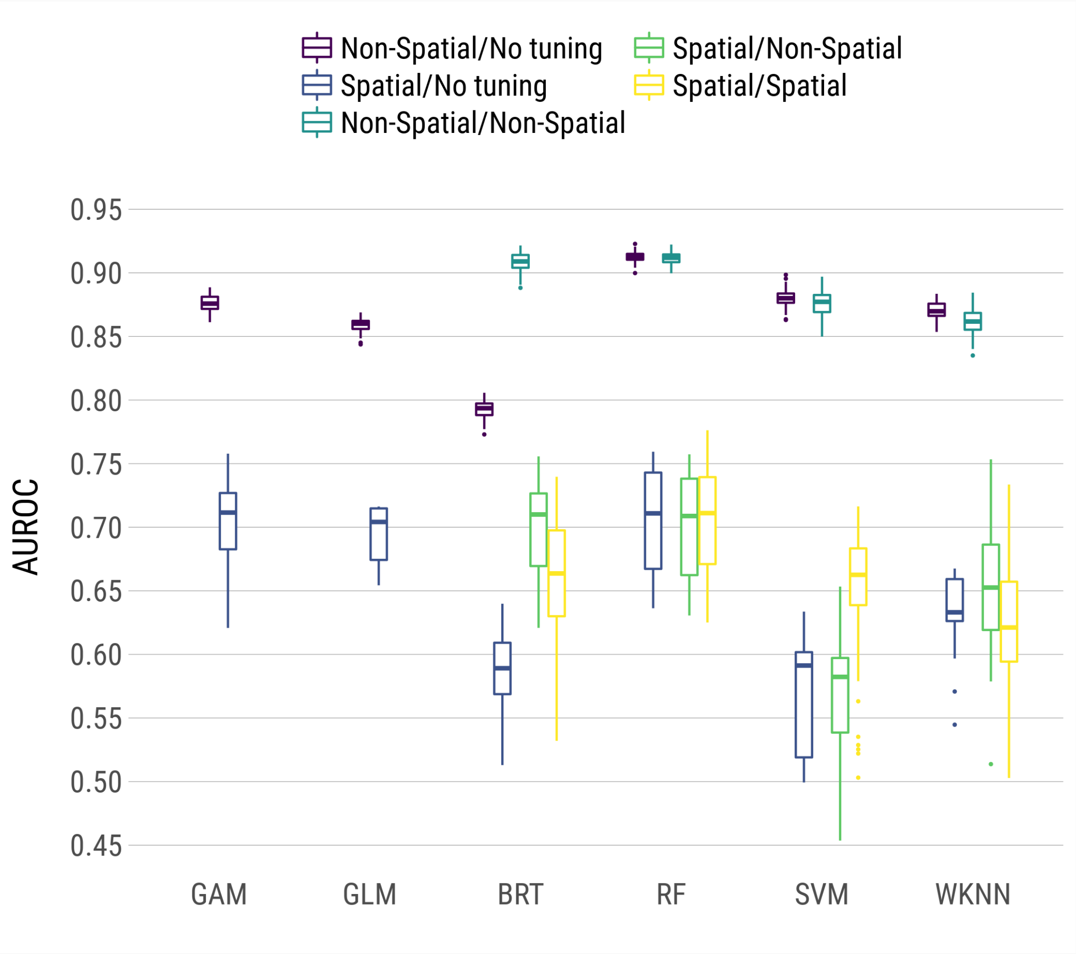


Fig 6: (Nested) CV estimates of model performance at the repetition level using 200 random search iterations. CV setting refers to perfomance estimation/hyperparameter tuning of the respective (nested) CV, e.g. "Spatial/Non-Spatial" means that spatial partitioning was used for performance estimation and non-spatial partitioning for hyperparameter tuning.

Predictive Performance

- `RF` and `GAM` showed the best predictive performance
- High bias in performance when using non-spatial CV
- Parametric models (`GLM`, `GAM`) show equally good performance estimates as the best ML algorithm (`RF`)

Discussion

Iturritxa et al. (2014)

GLM: 0.65 AUROC (without predictor `hail`)

GLM: 0.96 AUROC (with predictor `hail`)

This work

GLM: 0.66 AUROC (without predictor `hail_prob`) + slope, pH, lithology, soil

GLM: 0.694 (with predictor `hail_prob`) + slope, pH, lithology, soil

Discussion

Hyperparameter tuning

- Saturates at 50 repetitions and has a small effect for `SVM` and `BRT` (arbitrary defaults).

Discussion

Hyperparameter tuning

- Saturates at 50 repetitions and has a small effect for `SVM` and `BRT` (arbitrary defaults).
- Almost no effect on predictive performance for WKNN and RF (reasonable defaults).

Discussion

Hyperparameter tuning

- Saturates at 50 repetitions and has a small effect for `SVM` and `BRT` (arbitrary defaults).
- Almost no effect on predictive performance for WKNN and RF (reasonable defaults).
- Default hyperparameters of `RF` (and all other learners) are not suitable for spatial data

Discussion (Tuning)

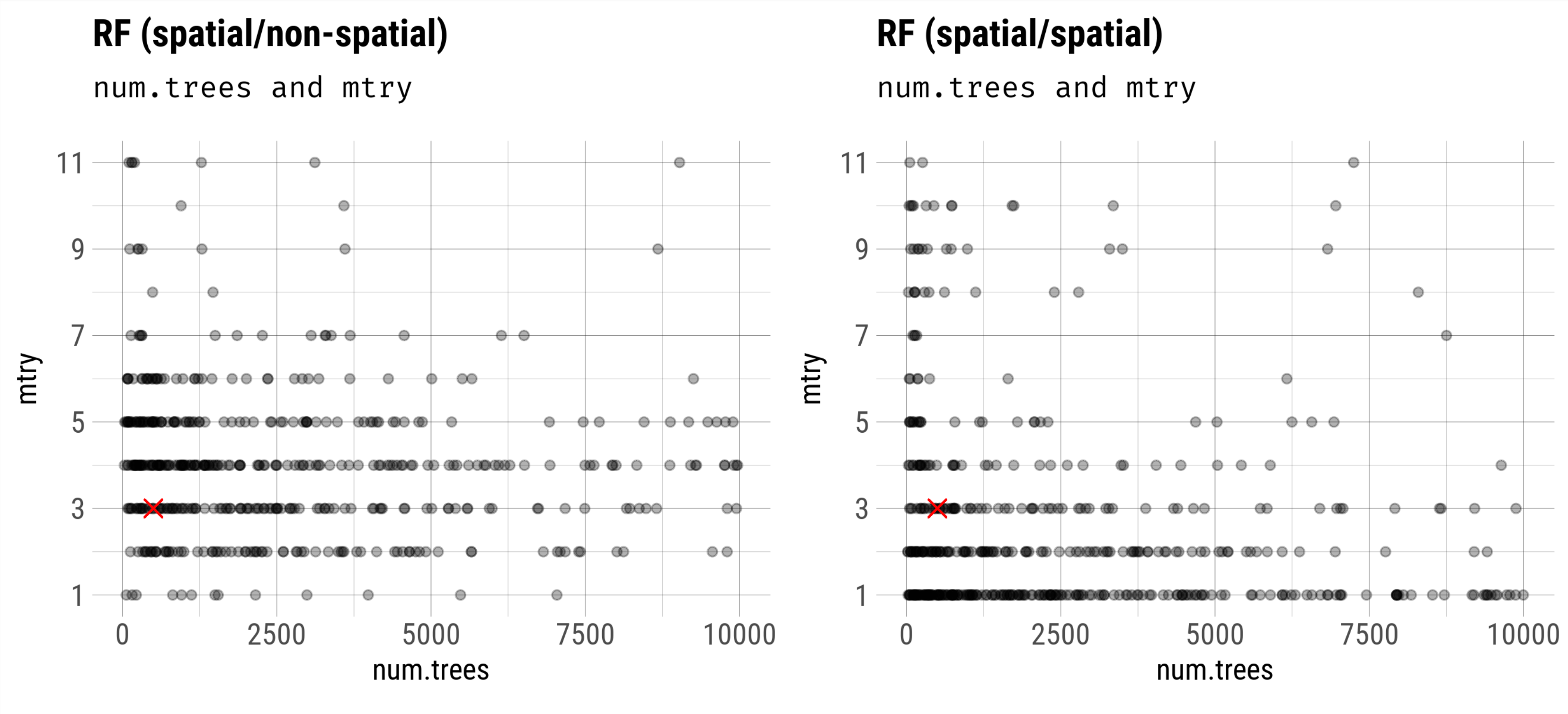


Fig 7: Best hyperparameter settings by fold (500 total) each estimated from 200 random search tuning iterations per fold using five-fold cross-validation. Split by spatial and non-spatial partitioning setup. Red crosses indicate default hyperparameter values. Black dots represent the winning hyperparameter setting out of each random search tuning of the respective fold.

Discussion

Hyperparameter tuning

- Saturates at ~ 50 repetitions and has a small effect for `SVM` and `BRT` (arbitrary defaults).
- Almost no effect for `WKNN` and `RF` (reasonable defaults).
- Default hyperparameters of `RF` (and all other learners) are not suitable for spatial data
    - They **possibly** lead to biased performance estimates as they cause fitted models to make use of the remaining spatial autocorrelation in the data.
    - Meaningful default values (`RF`, `WKNN`) have been estimated on non-spatial data sets.

❗ Always perform a spatial hyperparameter tuning for spatial data sets, even if it does not improve accuracy ❗

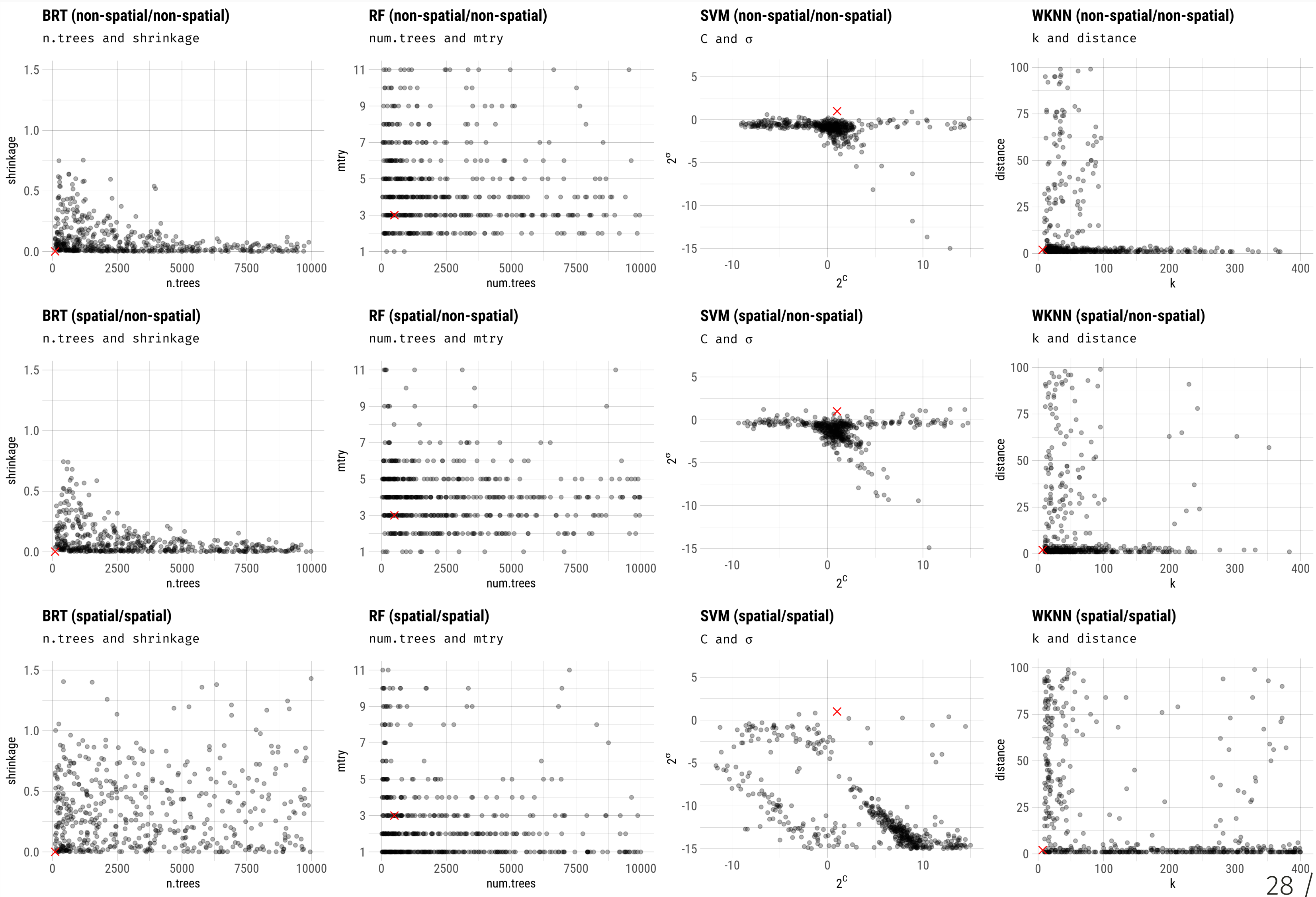
References

Bergstra, J., & Bengio, Y. (2012). Random search for hyperparameter optimization. J. Mach. Learn. Res., 13, 281–305. URL: http://dl.acm.org/citation.cfm?id=2188385.2188395.

Iturritxa, E., Mesanza, N., & Brenning, A. (2014). Spatial analysis of the risk of major forest diseases in Monterey pine plantations. Plant Pathology, 64, 880–889. doi:10.1111/ppa.12328.

Backup ➕

Backup



BRT (non-spatial/non-spatial)
n.trees and shrinkage
shrinkage
n.trees
RF (non-spatial/non-spatial)
num.trees and mtry
mtry
num.trees
SVM (non-spatial/non-spatial)
C and σ
2^{σ}
2^{C}
WKNN (non-spatial/non-spatial)
k and distance
distance
k
BRT (spatial/non-spatial)
n.trees and shrinkage
shrinkage
n.trees
RF (spatial/non-spatial)
num.trees and mtry
mtry
num.trees
SVM (spatial/non-spatial)
C and σ
2^{σ}
2^{C}
WKNN (spatial/non-spatial)
k and distance
distance
k
BRT (spatial/spatial)
n.trees and shrinkage
shrinkage
n.trees
RF (spatial/spatial)
num.trees and mtry
mtry
num.trees
SVM (spatial/spatial)
C and σ
2^{σ}
2^{C}
WKNN (spatial/spatial)
k and distance
distance
k

Backup

Table 2: Mean AUROC (repetition level) for different 5-fold 100 times repeated cross-validation settings. Settings with tuning are based on 200 random search iterations. Highest values of each column are highlighted in bold.

Performance estimation	Non-Spatial		Spatial		
Hyperparameter tuning	Non-Spatial	Non	Non-Spatial	Spatial	Non
GLM	-	0.859	-	-	0.694
GAM	-	**0.876**	-	-	0.705
BRT	0.909	-	0.699	0.661	0.587
RF	**0.911**	-	**0.702**	**0.705**	**0.706**
SVM	0.876	-	0.566	0.654	0.574
WKNN	0.861	-	0.652	0.624	0.635