optRF

Optimising random forests for reliable genomic selection in wheat breeding

Dr. Thomas Martin Lange

Postdoctoral researcher
Breeding Informatics
Georg-August University of Göttingen

What is genomic selection?

- Building a prediction model that describes the relationship between the phenotype and the genotype using a training population
- Using the model to predict the unknown phenotype in a test population
- Selecting the best individuals in the test population based on the predicted phenotypes

Training data:

SNP_1	SNP_2	SNP_3	...	SNP_p	Yield
AA	CG	AT	...	AC	8.31
CC	GG	TT	...	CC	6.69
AA	CG	TT	...	AA	7.42

Prediction model

Test data:

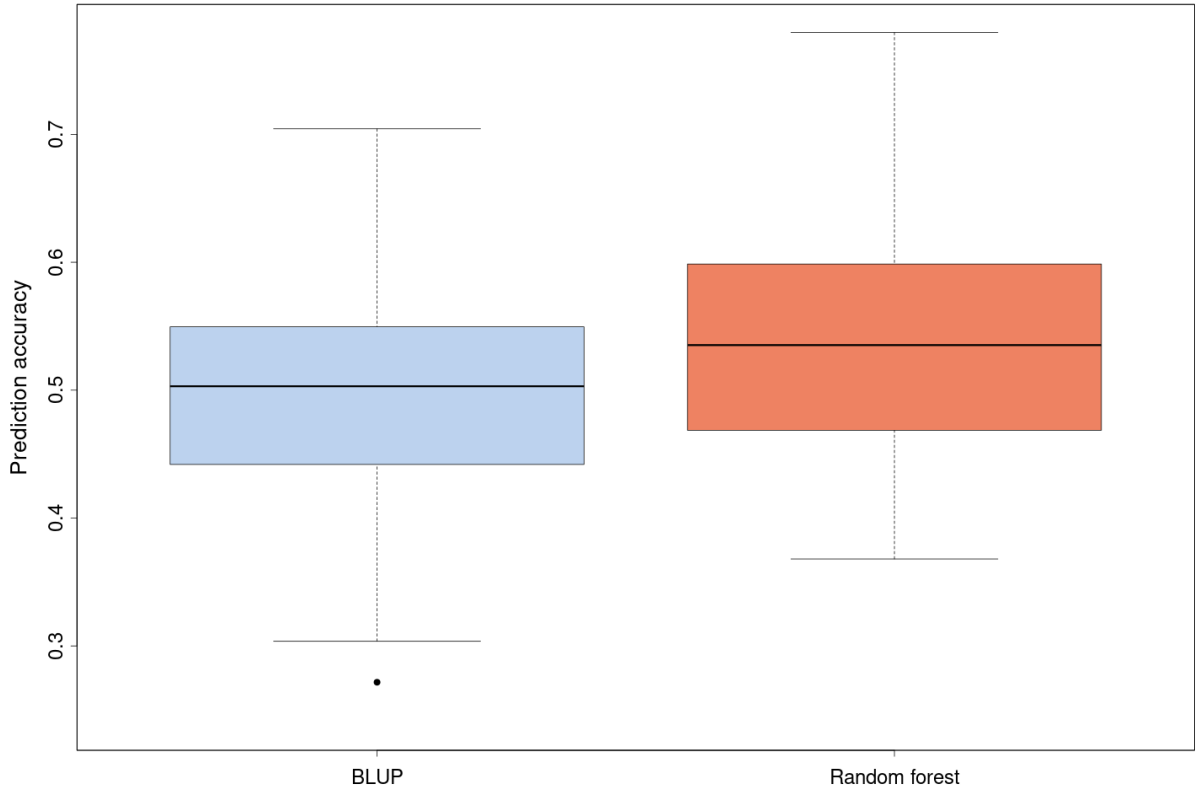
SNP_1	SNP_2	SNP_3	...	SNP_p	Predicted yield	Selection
AC	GG	TT	...	AA	**6.36**	✗
CC	CC	AA	...	AA	**8.82**	✓

Random forest prediction models

- Random forest is well-suited for genomic prediction
  - Non parametric method
  - Advantageous when data contain large number of SNPs but few observations
  - Includes non additive interactions between genes
- Random forest often outperforms other prediction models
- Research question: Does random forest perform better than the BLUP model in predicting the yield in a wheat trial?
  - Download of a publicly available data set from wheat breeding
  - Genomic prediction using random forest and BLUP and comparing the prediction accuracy

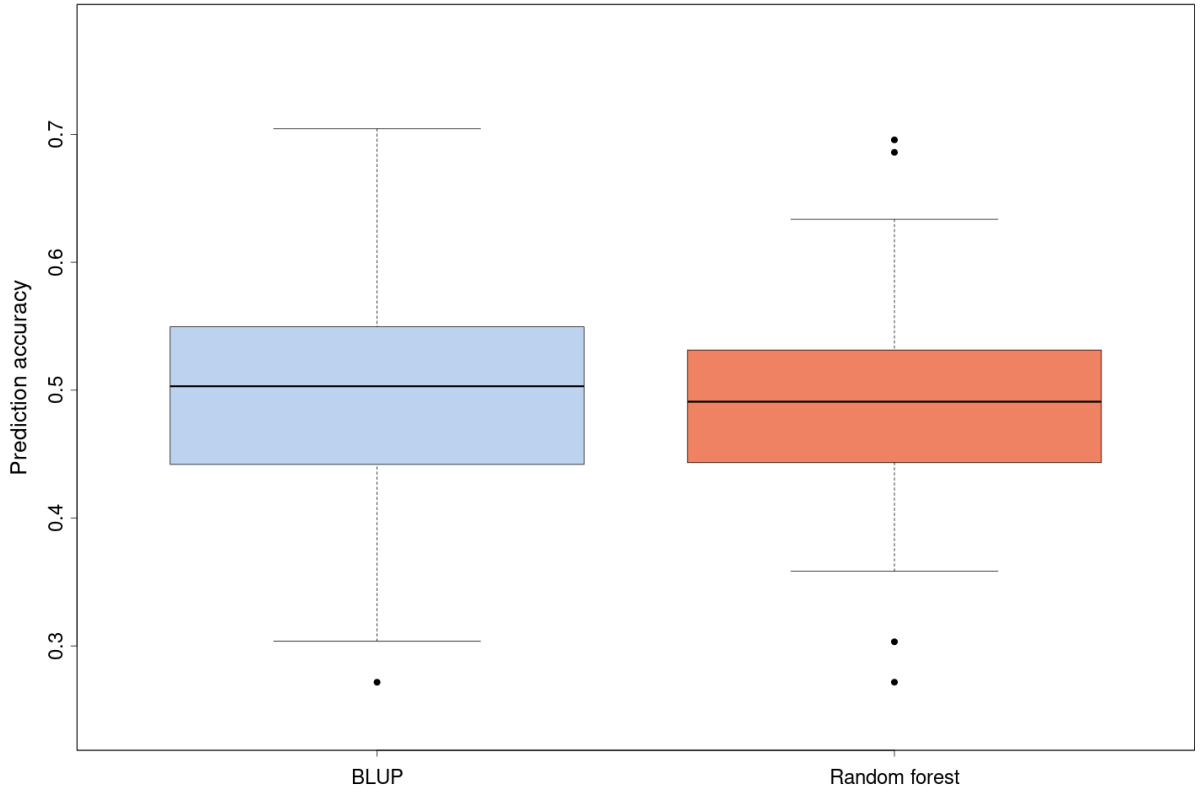
Random forest vs BLUP (I)

- First result: Random forest performed significantly better than BLUP
- Second result: Random forest and BLUP performed equally
- Third result: Random forest performed significantly worse than BLUP



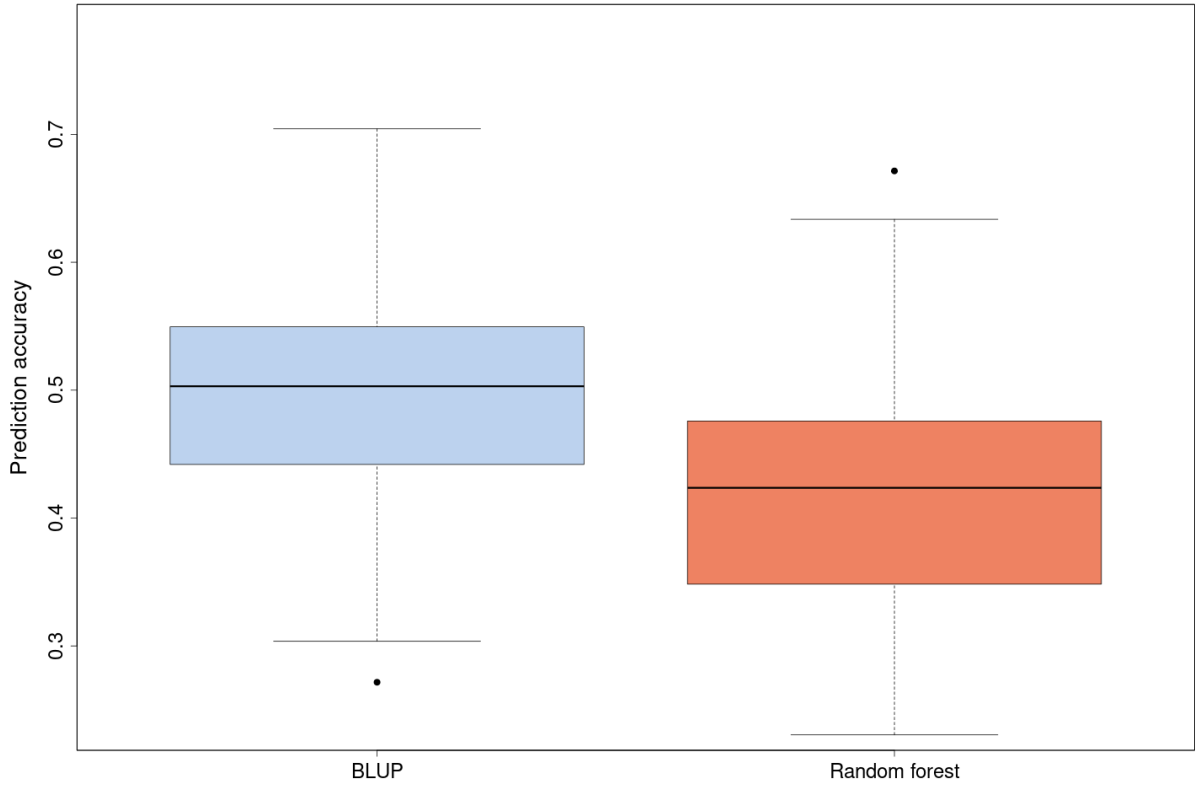
Random forest vs BLUP (II)

- First result: Random forest performed significantly better than BLUP
- Second result: Random forest and BLUP performed equally
- Third result: Random forest performed significantly worse than BLUP

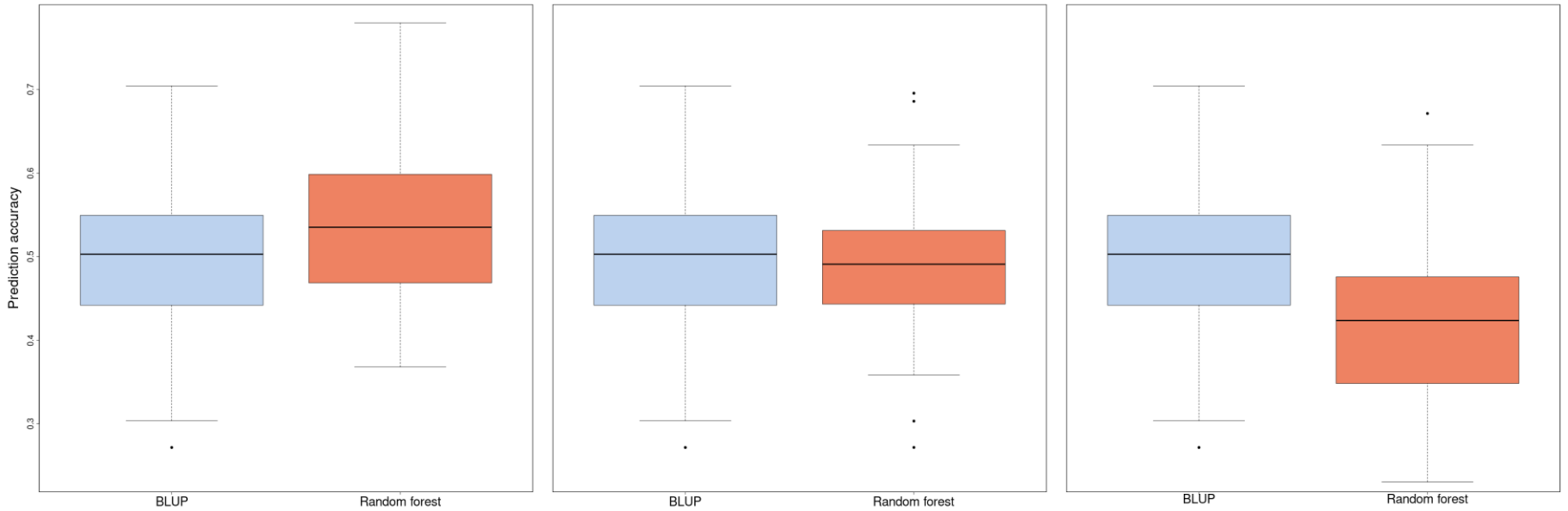


Random forest vs BLUP (III)

- First result: Random forest performed significantly better than BLUP
- Second result: Random forest and BLUP performed equally
- Third result: Random forest performed significantly worse than BLUP

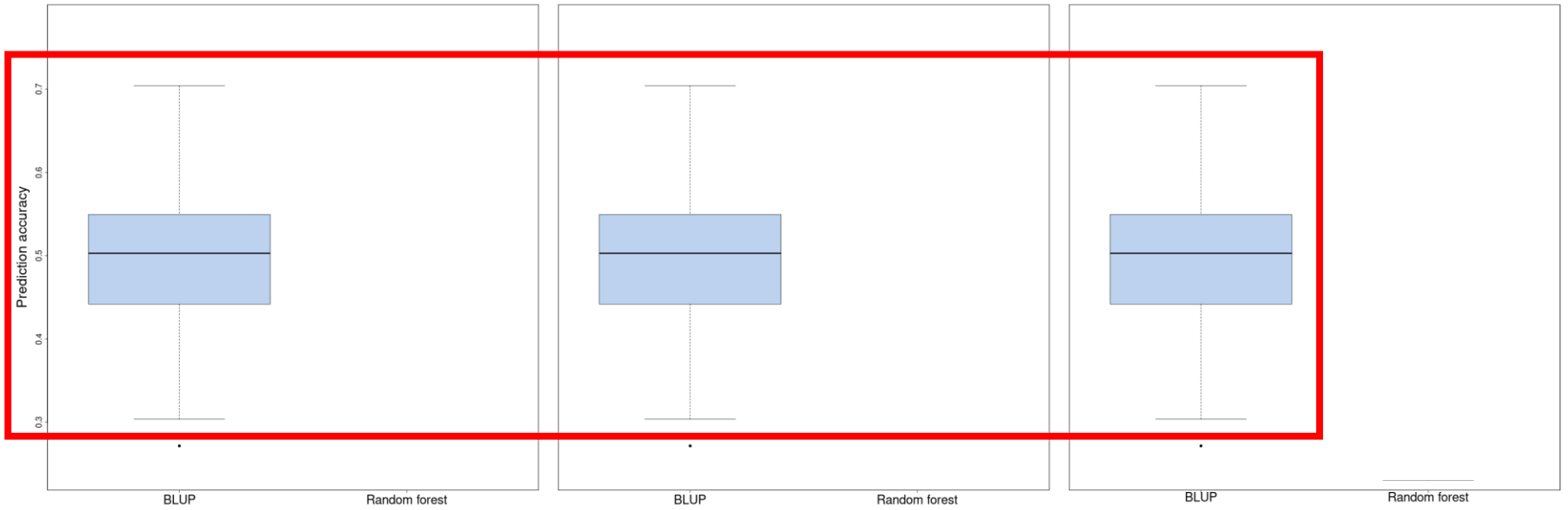


The effect of non-determinism on genomic prediction



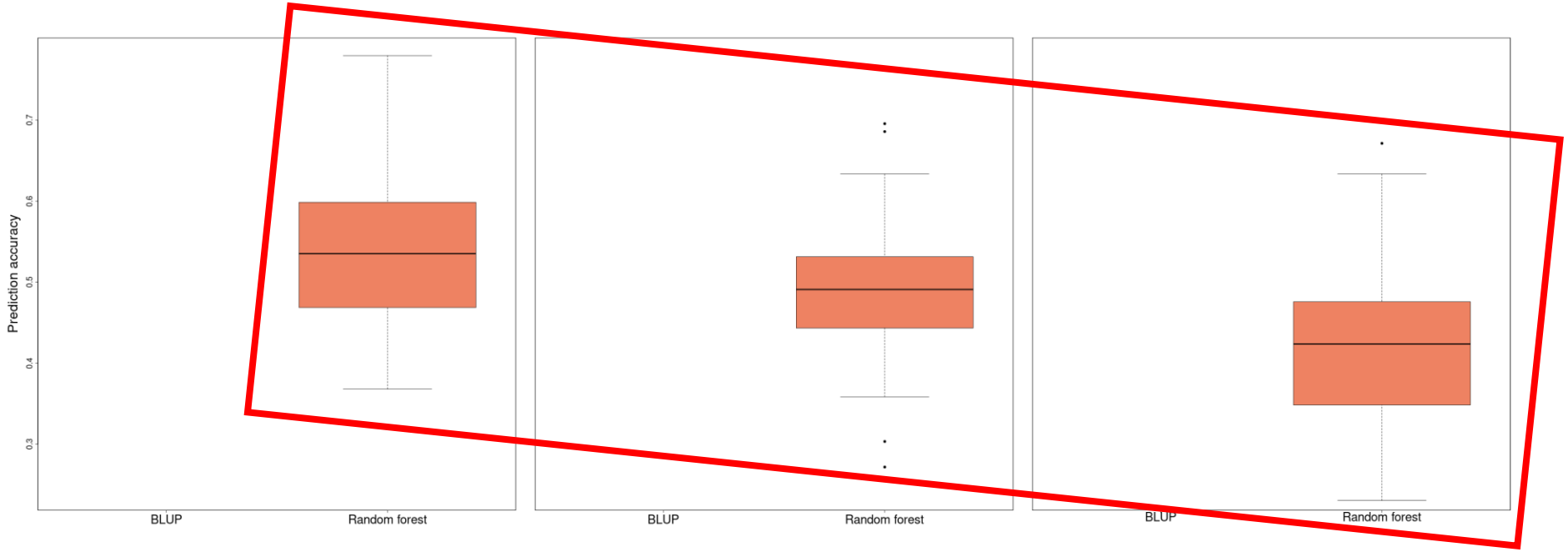
- The results vary in repetitions of the analysis

The effect of non-determinism on genomic prediction



- The prediction accuracy of the BLUP model is constant in the three repetitions
- BLUP is a deterministic method
  - Given the same input data, the same prediction model will be build and the same predictions will be made

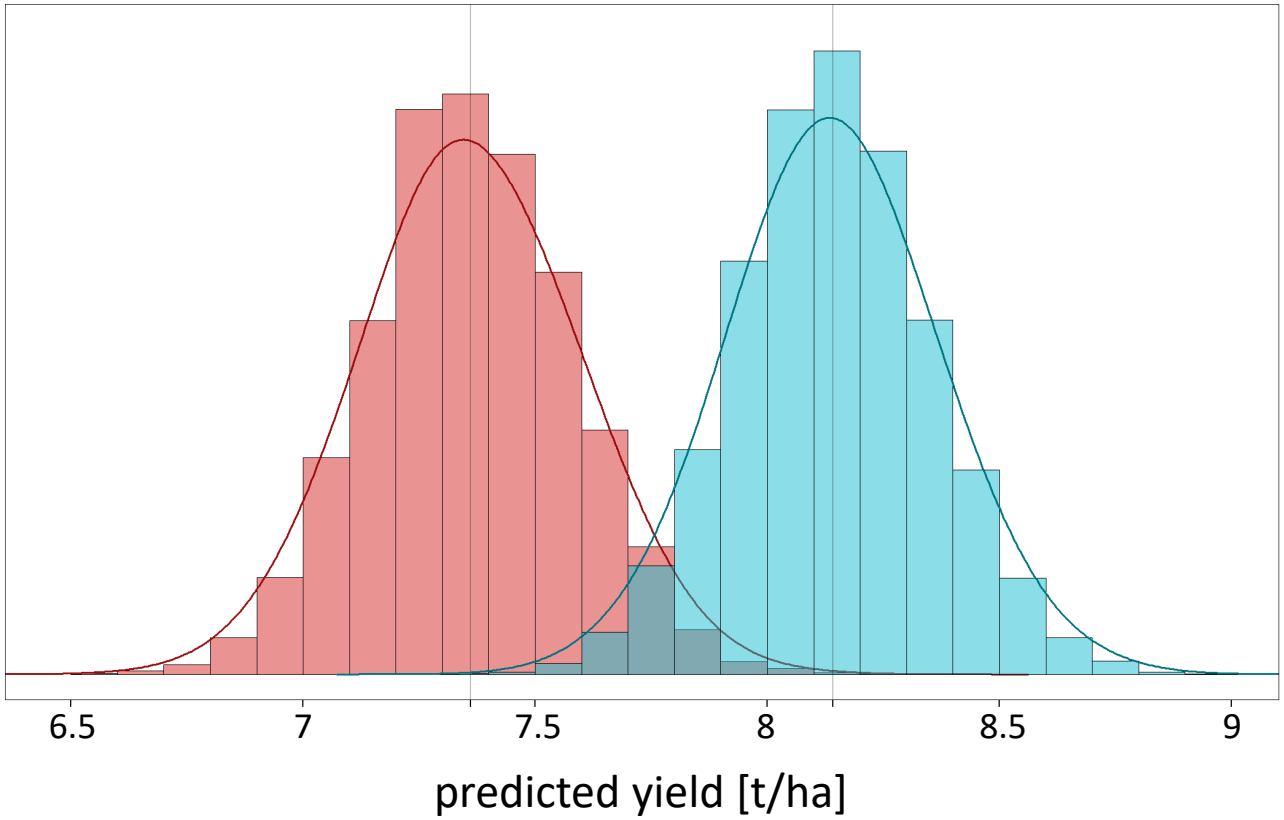
The effect of non-determinism on genomic prediction



- The prediction accuracy of random forest change in the three repetitions
- Random forest is a non-deterministic method
  - Even with the same input data, random forest may build different prediction models which lead to different predictions and can lead to different prediction accuracies

How non-determinism can affect decisions

- For example: The predicted yield of two varieties using random forest 10,000 times
- The predictions are normally distributed around a mean prediction per variety
- The predictions show some overlap: It could be possible to make a wrong decision just by chance

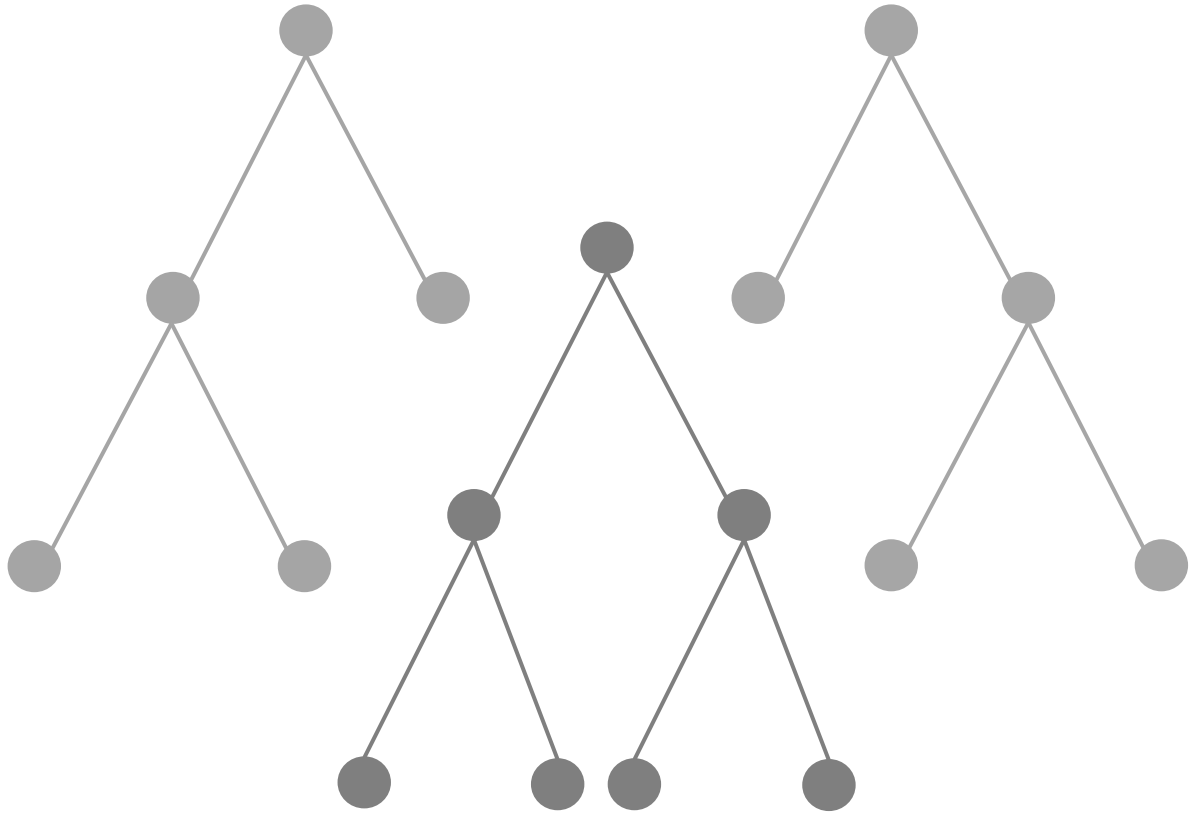


Why is Random Forest non-deterministic?

- Random forest works by growing multiple decision trees and averaging their predictions as the final prediction of random forest
- To increase accuracy, randomness is introduced in the decision trees

  → Variable selection

  → Bootstrapping

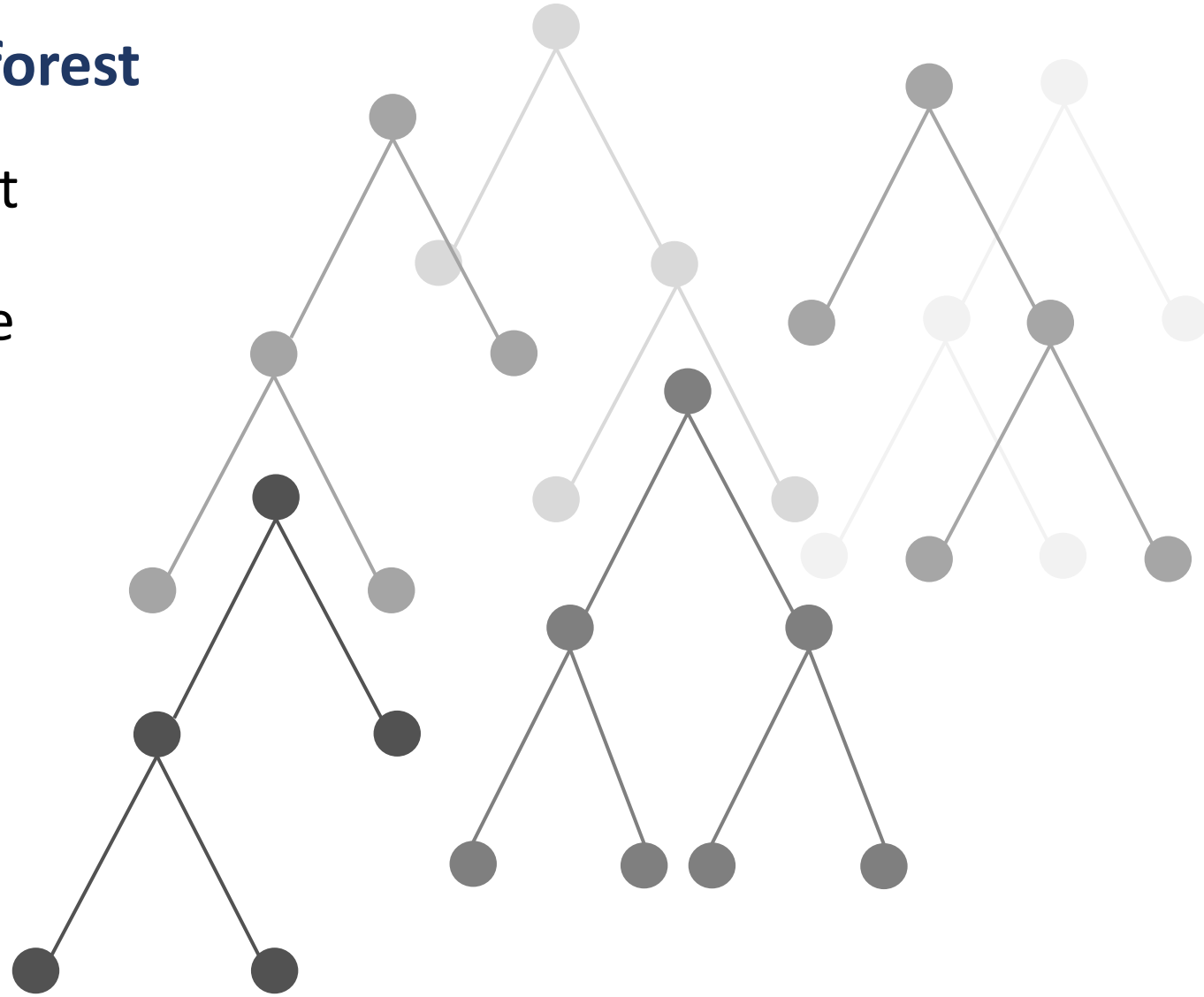


Increasing the stability of random forest

- Since the prediction of random forest is based on decision trees, more decision trees will lead to more stable predictions
- Disadvantage:

  Growing a decision tree requires computation time

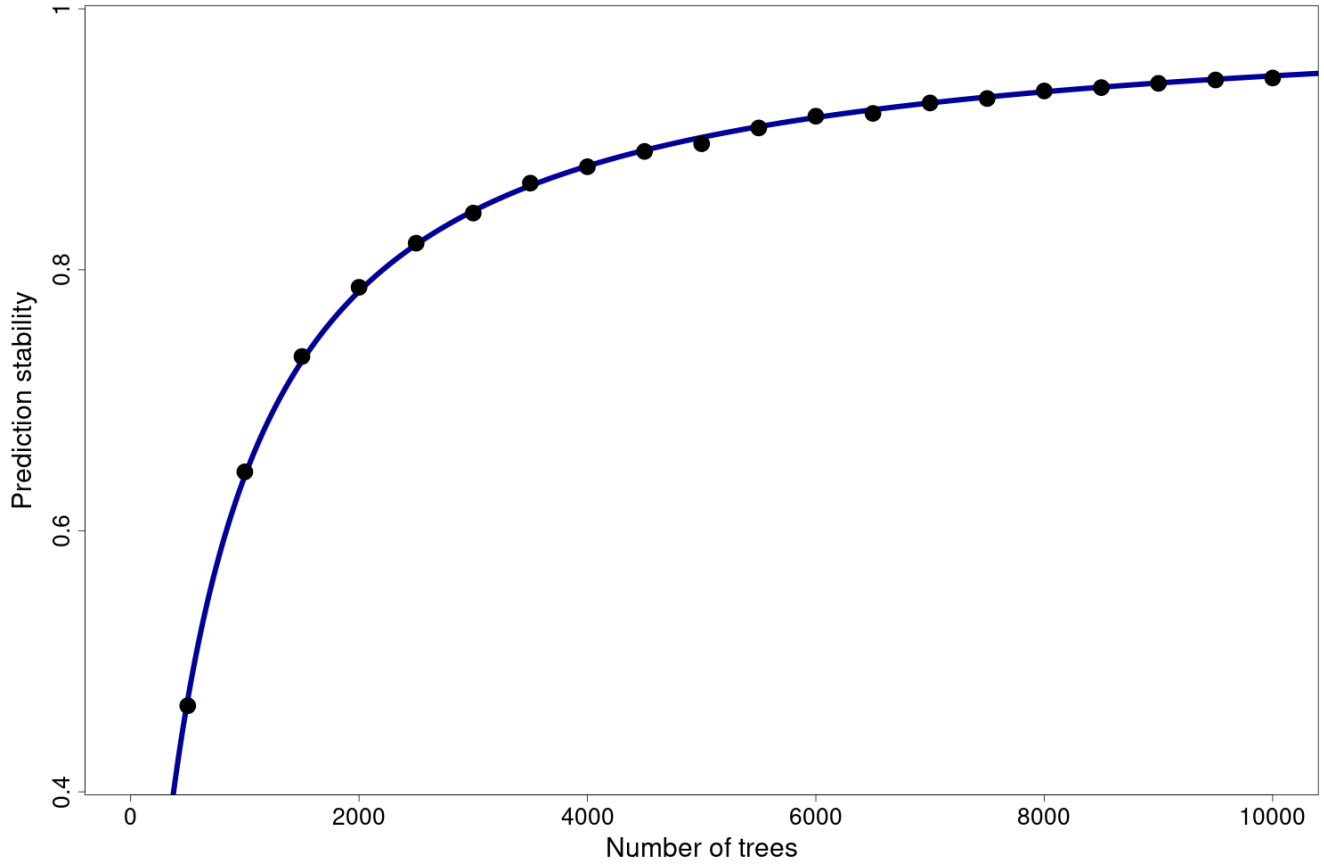
  → Growing 1,000 decision trees takes double as much time as growing 500 decision trees



Relationship between stability and the number of trees

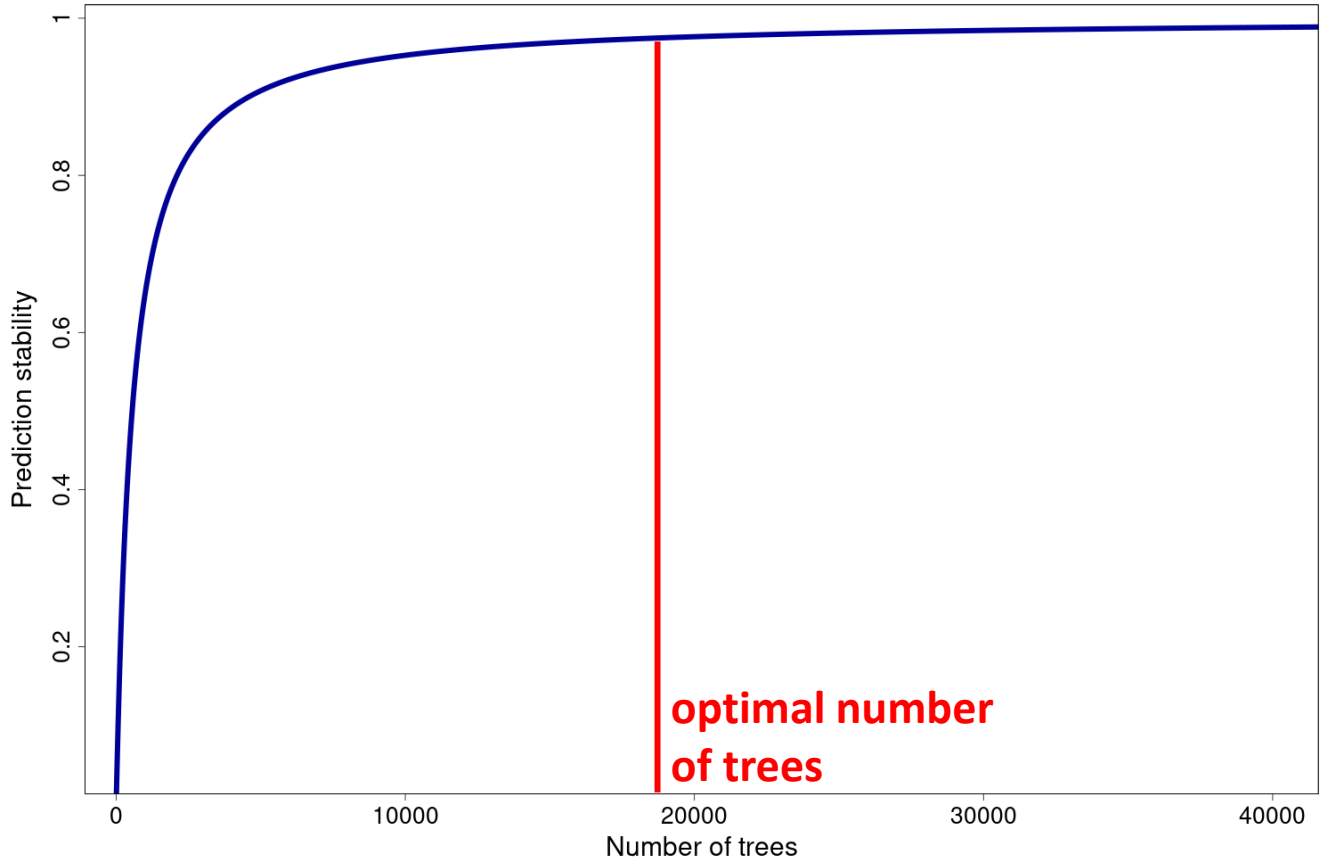
- The variability of predictions can be measured as the *prediction stability* by repeating the model fitting process
- The prediction stability increases non-linearly with higher number of trees
- The relationship can be modelled using a two parameter logistic (2PL) regression model:

$$\widehat{ICC_j} = \frac{1}{1 + \left(\frac{\theta_1}{t_j}\right)^{\theta_2}}$$



Determining the optimal number of trees

- The 2PL model allows estimating the prediction stability for very high numbers of trees
- The optimal number of trees is where further trees lead to minimal gains in prediction stability
- But the optimal number of trees is data set dependent



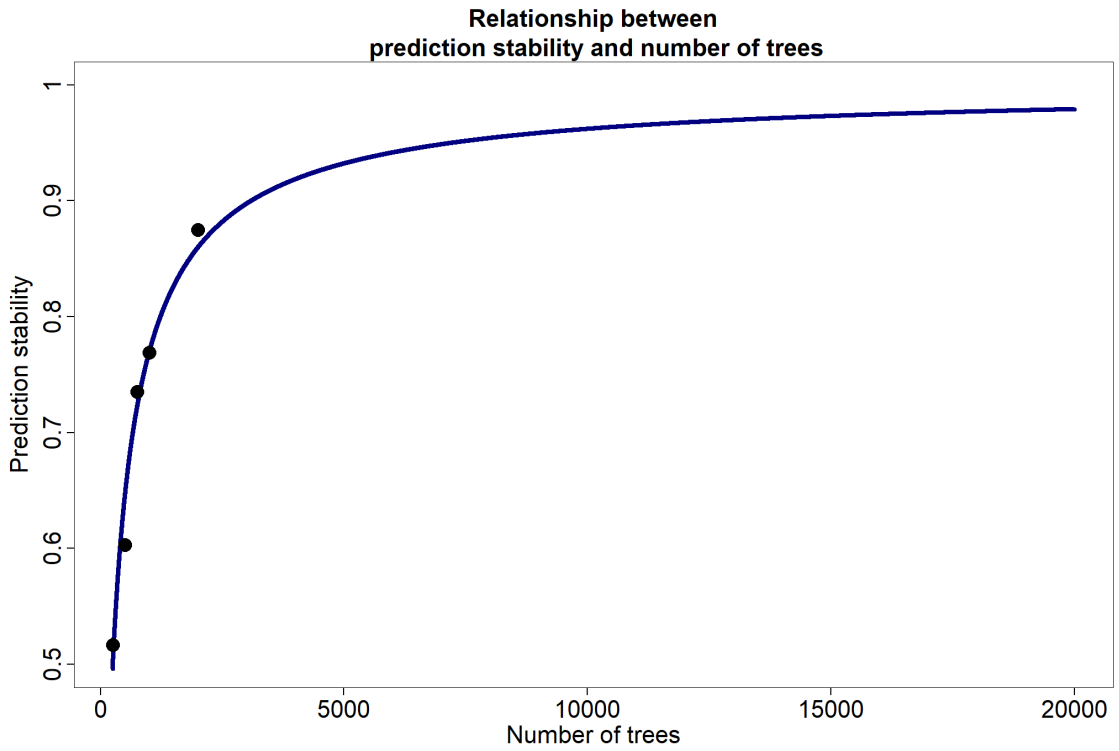
The optRF package

- The optRF package does all these calculations automatically, giving as a result a clear recommendation for the optimal number of trees

```
> opt_prediction(y = Response,
+                X = GenomicMarkers)
```

Recommended number of trees: 19000

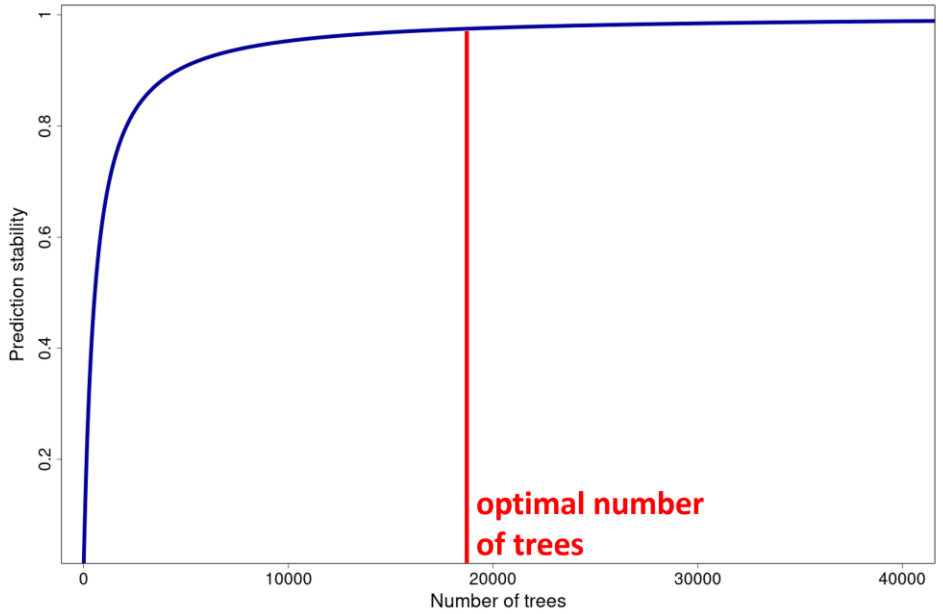
```
> randomForest(y = Response,
+              x = GenomicMarkers,
+              ntree = 19000)
```



→ Use the optimal number of trees to make accurate predictions and reliable selection decisions using random forest

Take home messages

- Random forest is a non-deterministic prediction model
  - → Predictions and selection decisions can change when repeating the analysis
- Increasing the number of trees increases the stability but also the computation time



- The R package optRF determines the optimal number of trees and estimates the prediction stability for your data
- Whenever working with random forest or any other non-deterministic prediction model, the prediction stability must be calculated and published with the results