Introduction to machine learning for spatial mapping of the environment

Hanna Meyer
Remote Sensing & Spatial Modelling,
Institute of Landscape Ecology, University of Münster

Before we start...

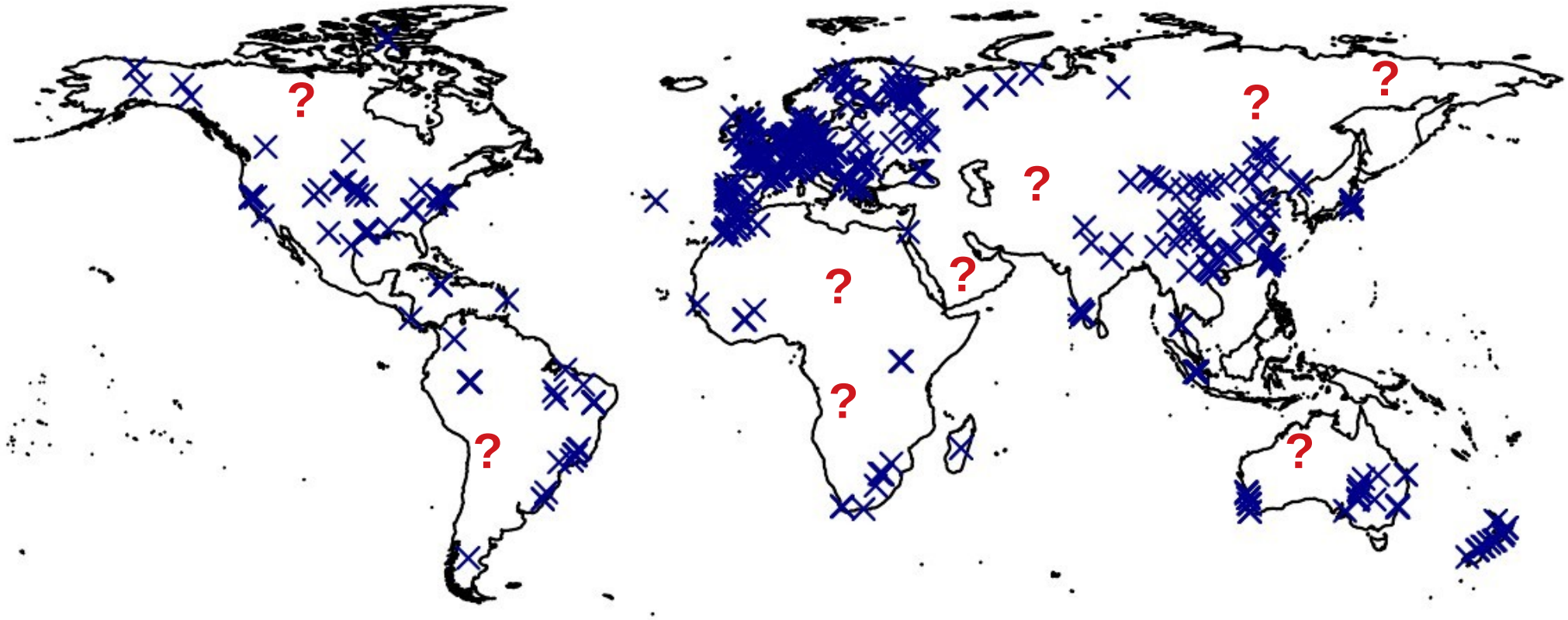
What is your

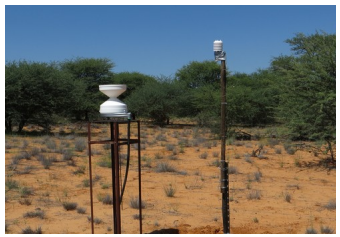
- Name
- Study program / PhD work etc ?
- Experience with R, spatial data (GIS), machine learning
- Expectations for this course ?

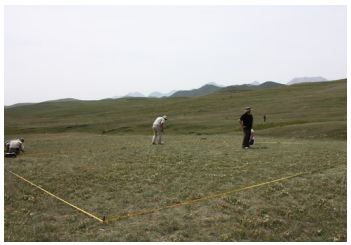
Aim of this workshop

- After this course you should ideally be able to:
  - Know and apply the basic workflow of spatial mapping via machine learning
  - Assess the quality of your predictions
  - Understand major risks and pitfalls
  - Critically analyze the results
  - Communicate the quality and limitations of predictions
- Organization:
  - Lecture + hands-on
  - 2 breaks (more if needed :))
  - Please interrupt me any time if you have questions
- Material:
  - Data: https://github.com/HannaMeyer/mlsp_GOE24
  - Slides + code will be added to this repository after the workshop

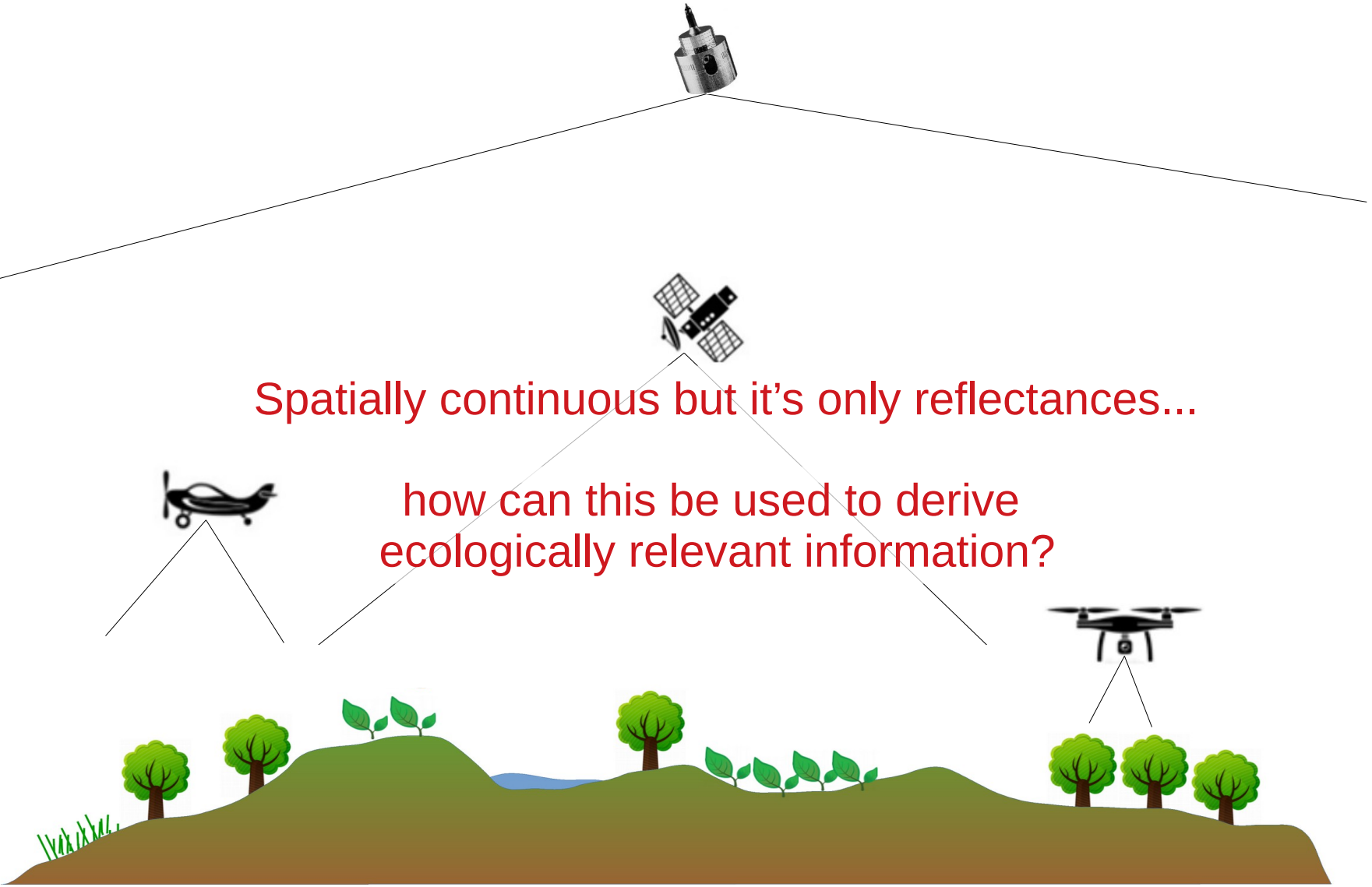
Problem: We only have limited (point) information about the environment





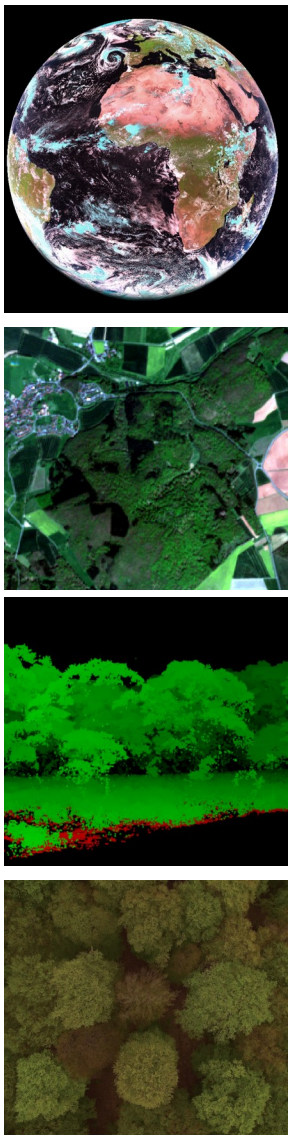


Remote Sensing to derive continuous information



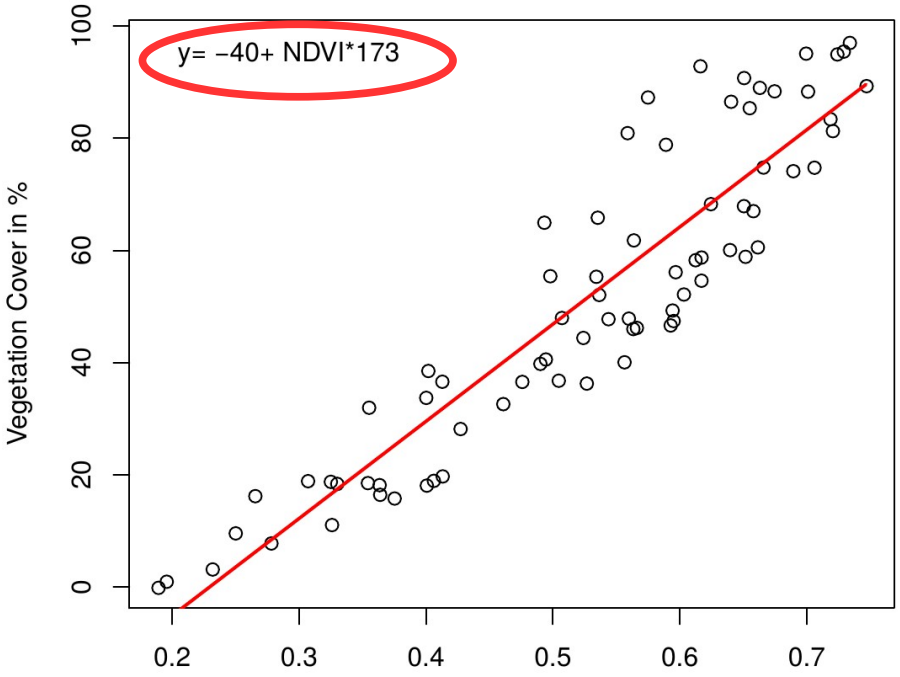
Spatially continuous but it's only reflectances...

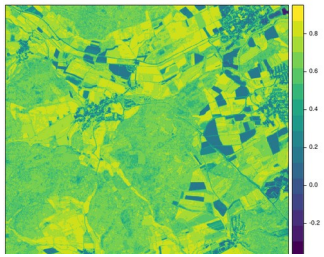
how can this be used to derive ecologically relevant information?



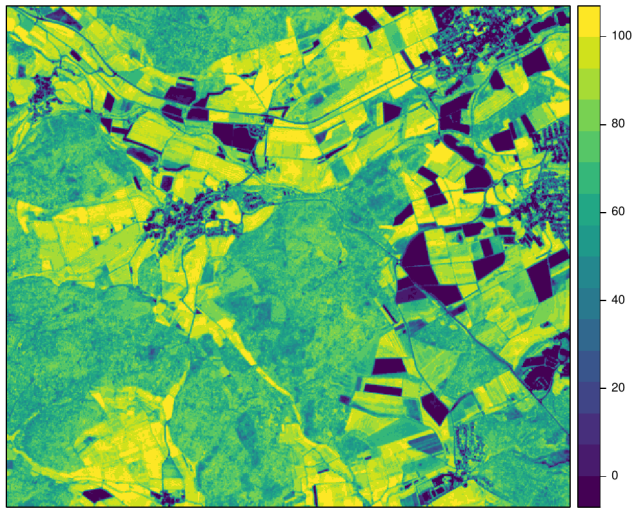
How can we translate the remote sensing information to the ecological variable?

e.g. vegetation cover from satellite data



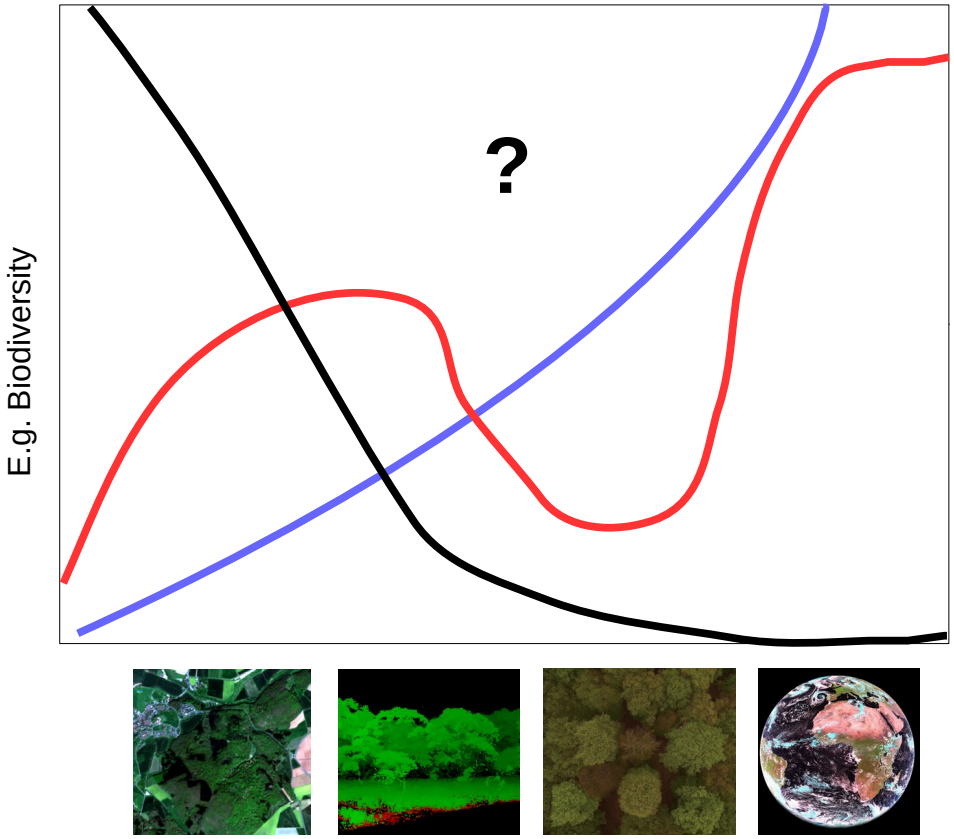


Modelled vegetation cover



How can we translate the remote sensing information to the ecological variable?

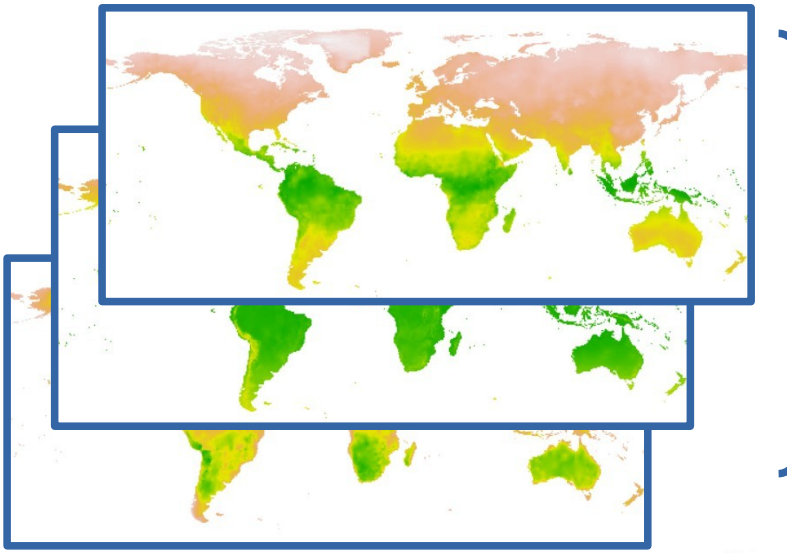
...but what about more complex variables ?



Models that can deal with complex nonlinear relationships are required!

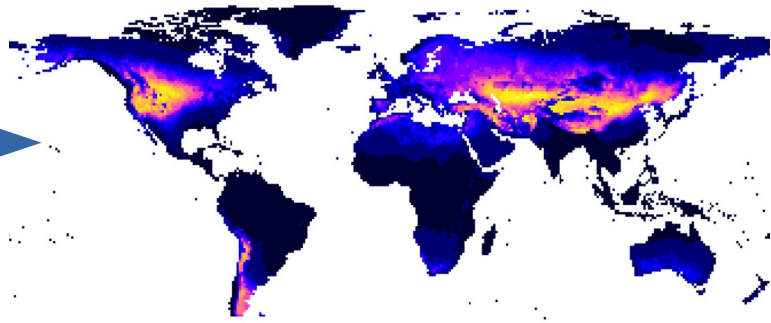
The machine learning way

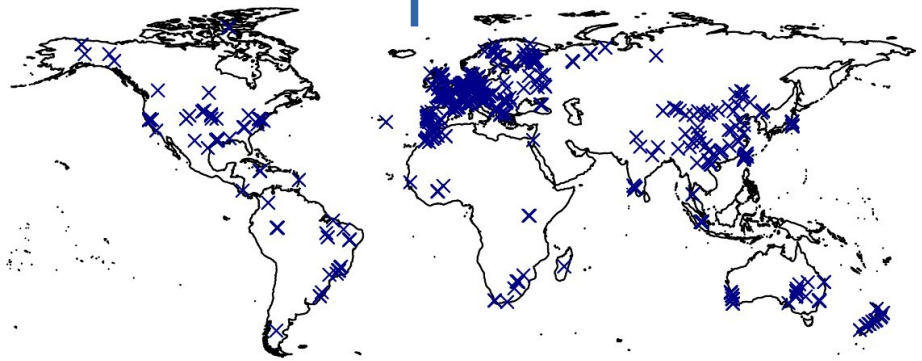
Predictors (remote sensing, climate, terrain,...)



Machine learning
algorithm

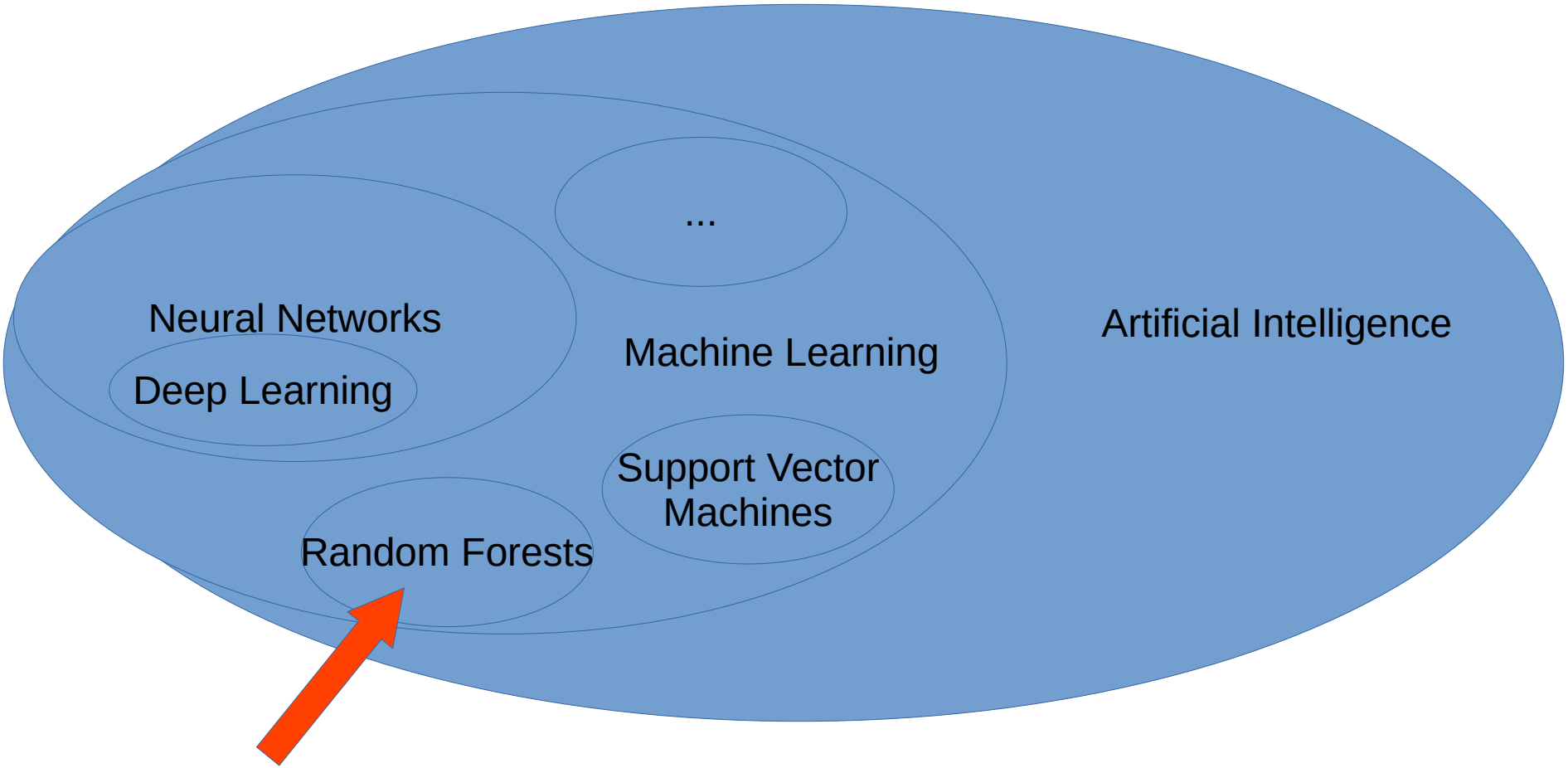
Spatial prediction





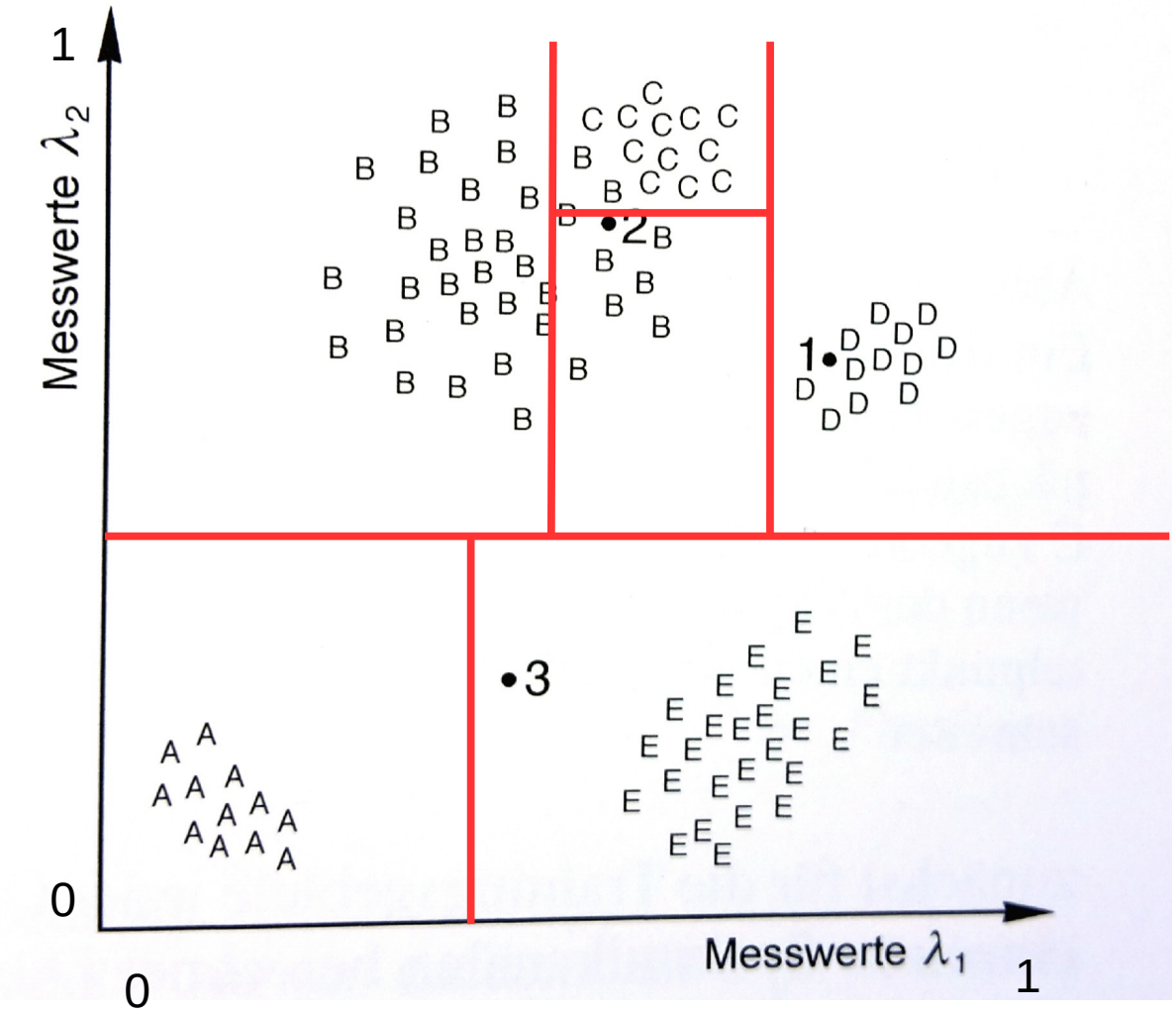
Response

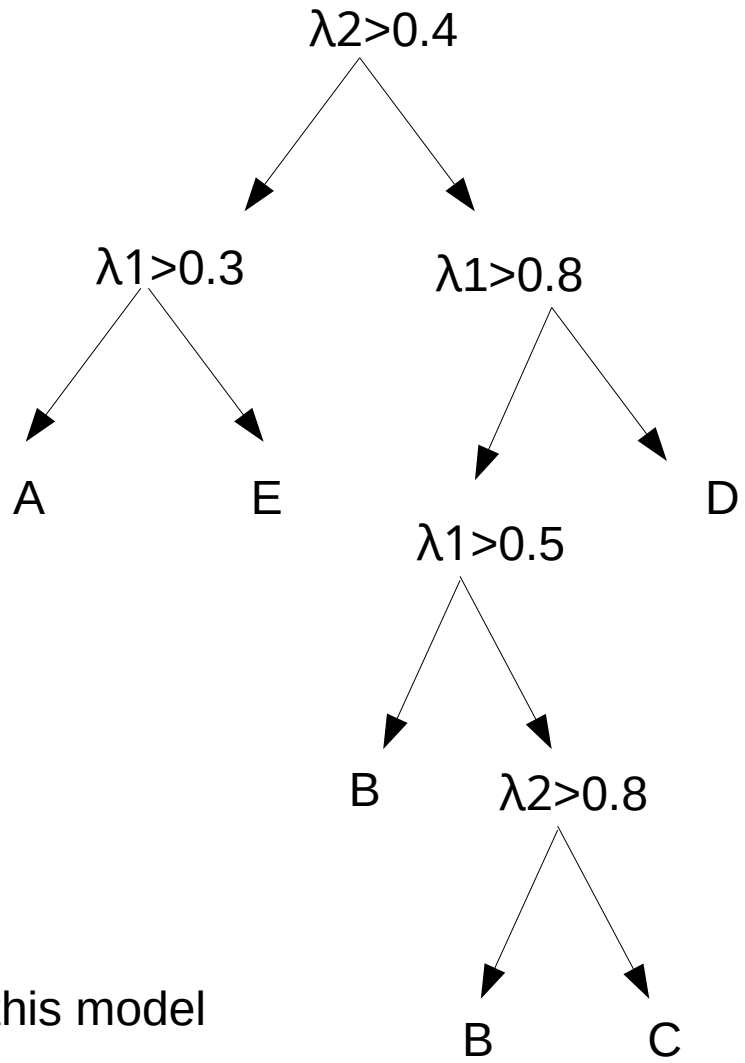
What is this “algorithm” ?



Most frequently used in ecology! Used for this workshop

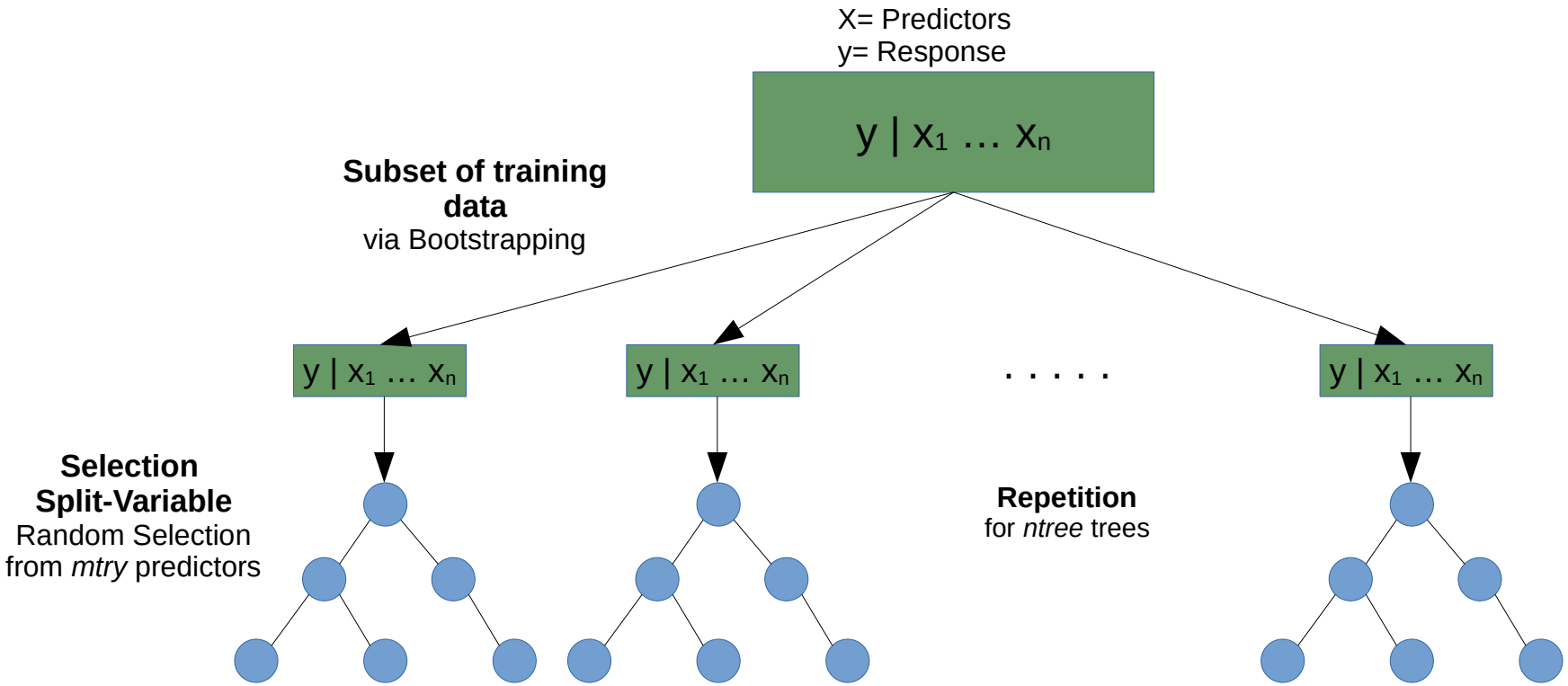
Random Forests: basis are decision trees



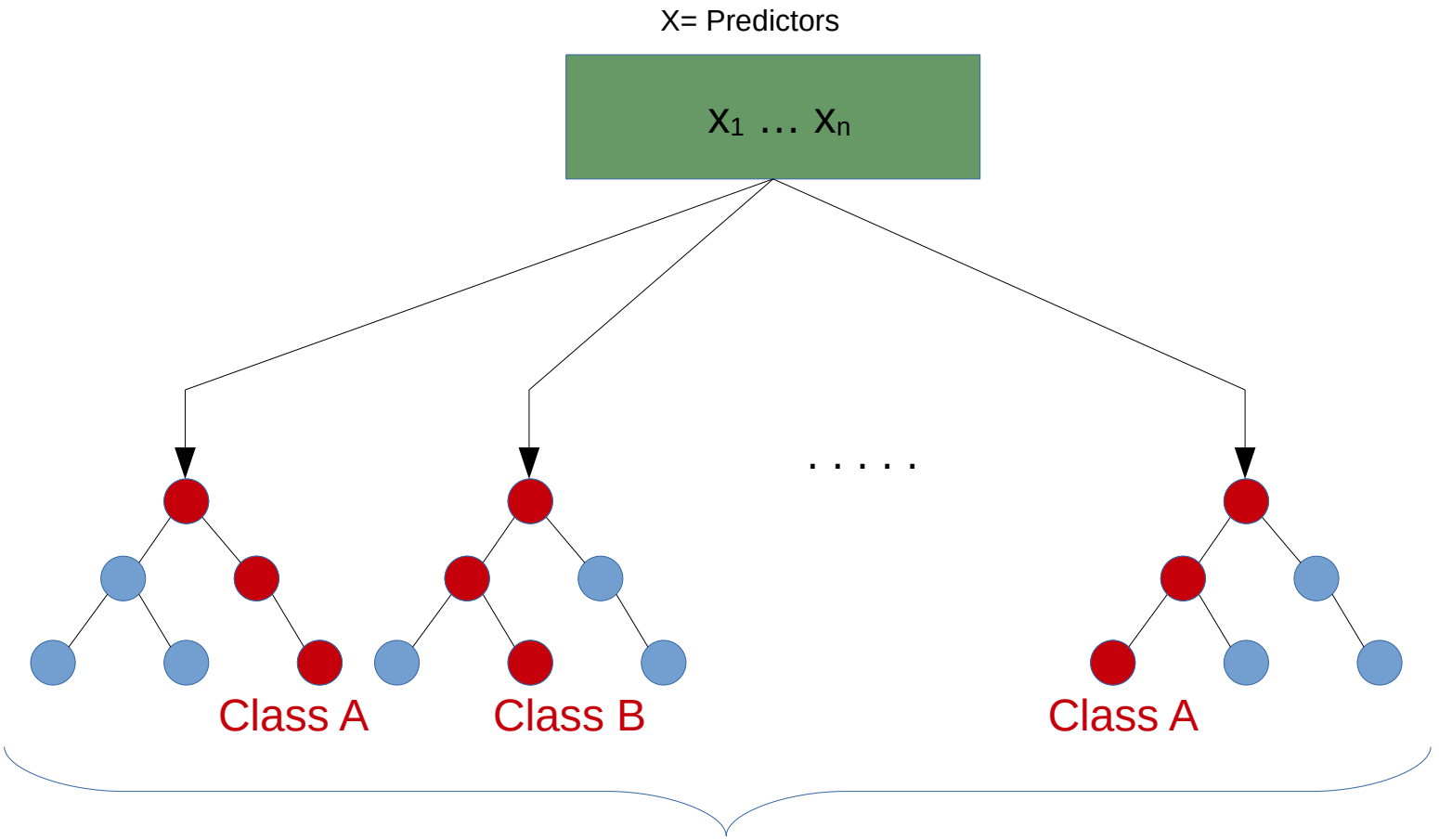


→ Prediction: classify new data points based on this model

Random Forest: Training



Random Forest: Prediction



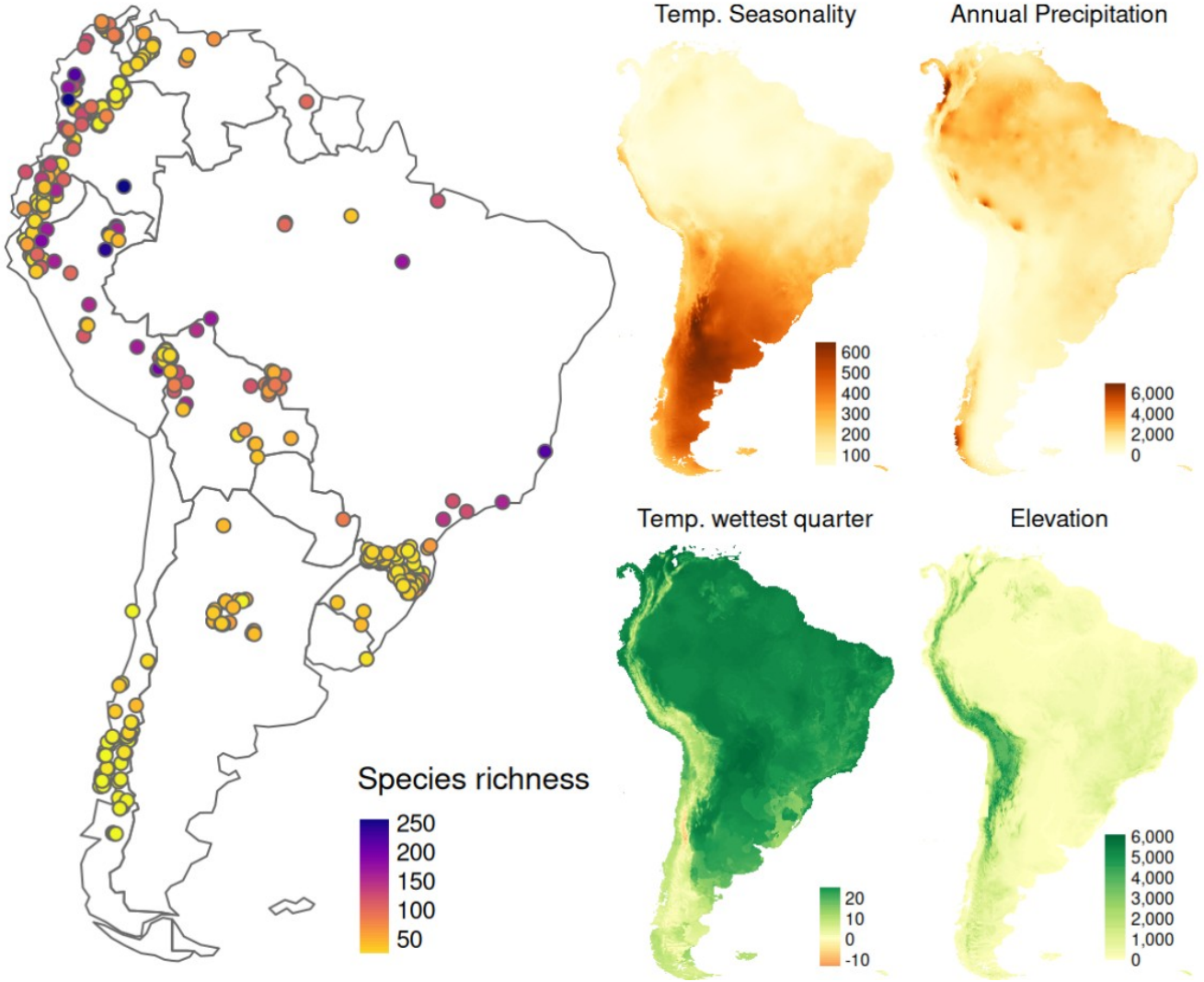
Decision by combining all trees (ensemble)

Classification: Majority; Regression: Mean value

Let’s try this for a case study

Aim: Produce a **spatial continuous map** of plant species richness for South America

- based on **reference data from sPlotOpen**
- assuming that **climate and elevation are predictors** of species richness
- Assuming that relationships are **complex,** therefore apply machine learning



Definition Bioclim Variablen:
https://www.worldclim.org/data/bioclim.html

Time for practice!

Task one: Download the data and explore the dataset in R.
E.g. is there a relationship between temperature and species richness?
Required packages: CAST, sf, terra, mapview (optional)

Get the species data in R

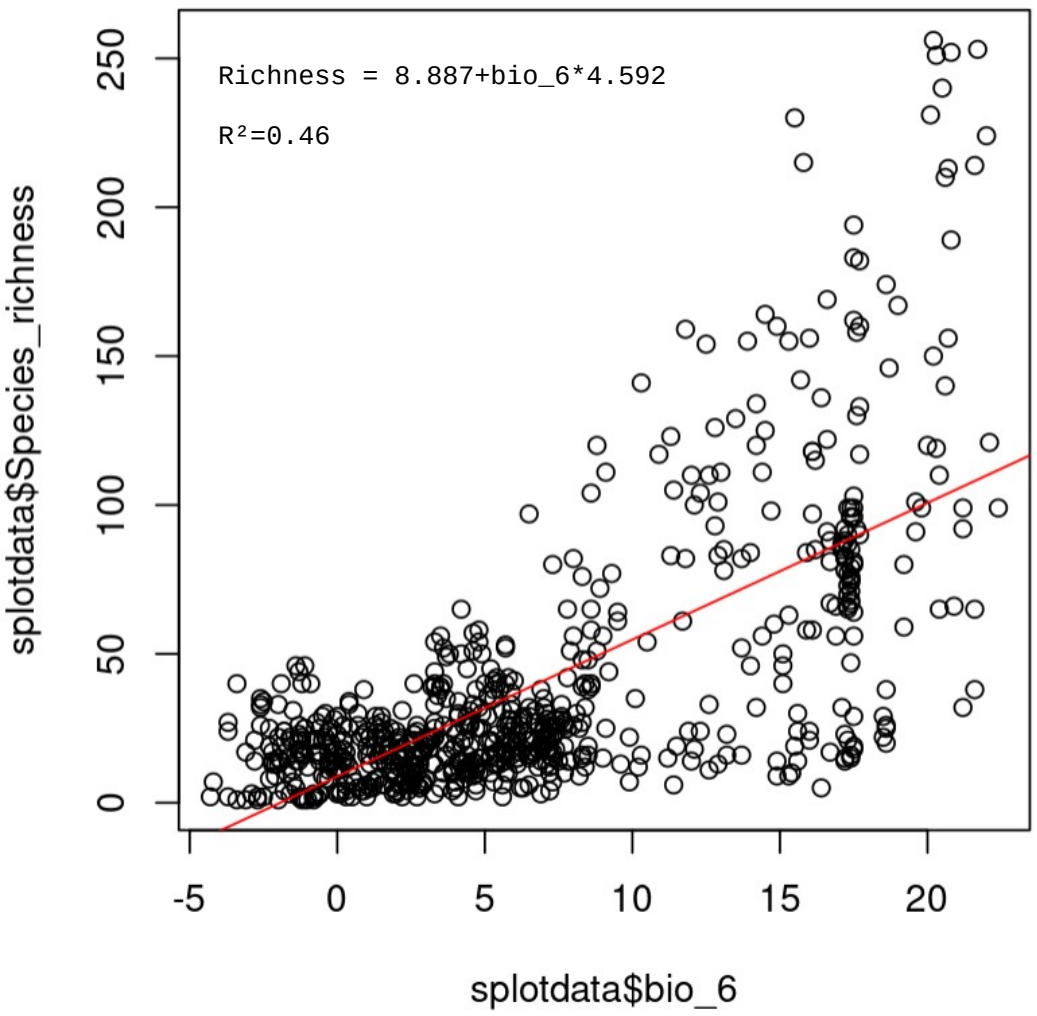
```
library(CAST)
data(splotdata)
```

...and the predictors as raster data

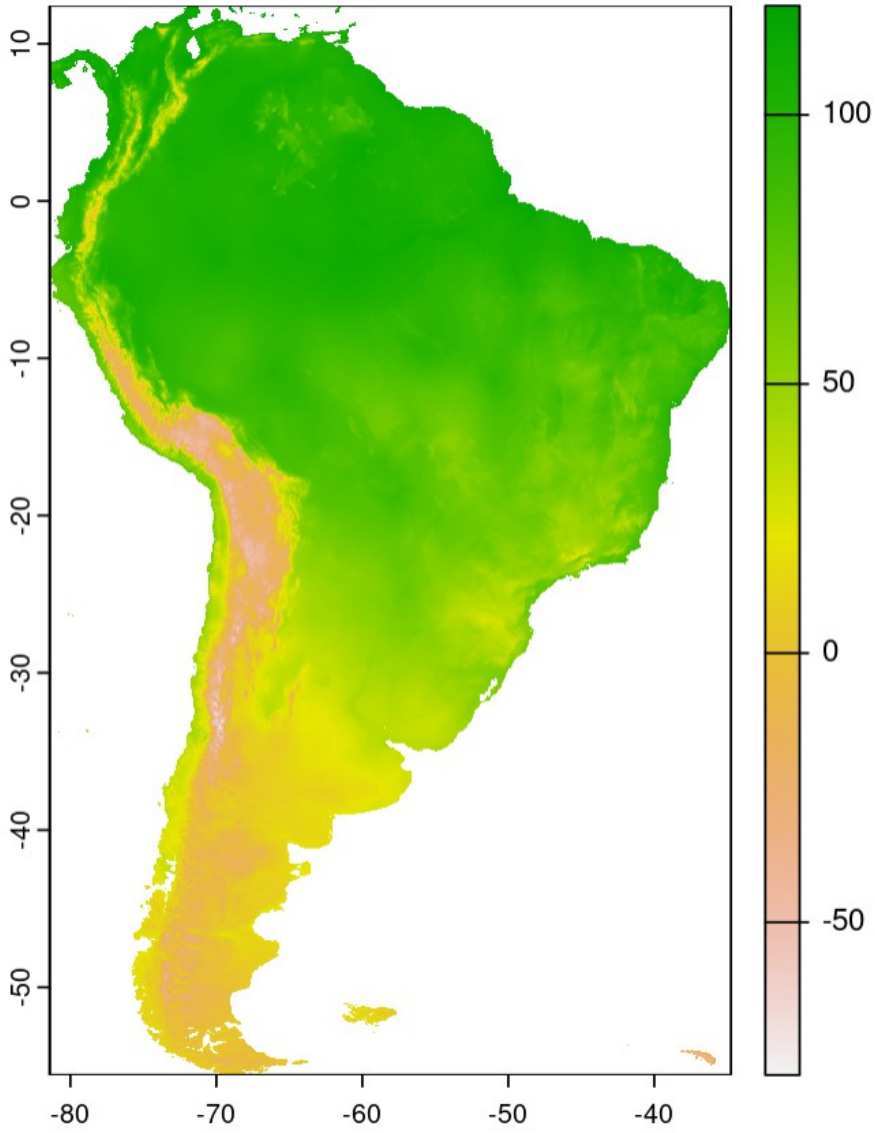
```
library(terra)
predictors_sp <- rast("data/predictors.tif")
```

First try: a linear model

Linear model



Predicted species richness

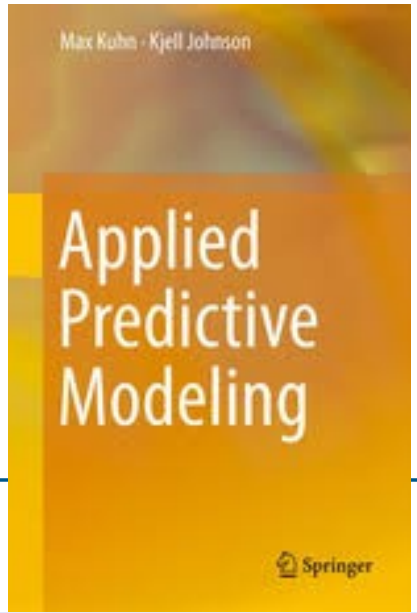


Machine learning in R

- Many packages for different ML algorithms (e.g. Random Forests, Neural Networks, Support Vector Machines, ...)
- For classification and regression problems
- Wrapper packages
  - allowing access to many algorithms via a unified syntax
  - Supporting functionality for cross-validation etc.
  - Caret (Classification And REgression Training)
  - Mlr3 (Machine Learning in R)
  - Tidymodels

For today's session

Extension for spatial modelling:
R-package CAST (Caret Applications for Spatio-Temporal models)
https://hannameyer.github.io/CAST/



Step 1: Model training in R

Predictors* Response

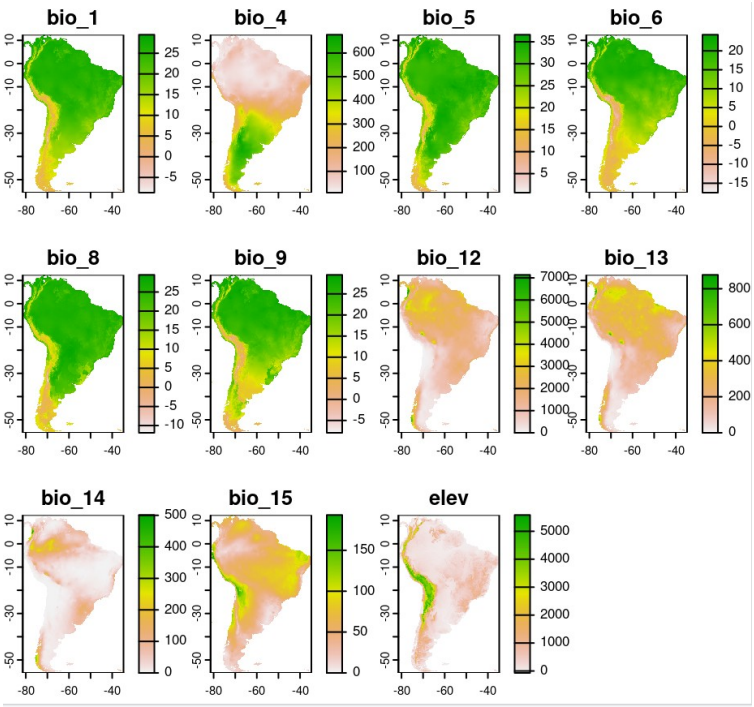
bio_1	bio_4	bio_5	...	richness
17.65000	463.9651	30.5		52
17.35417	459.5525	30.1		56
18.31667	473.3216	31.4		65
18.04167	485.8116	31.2		50
18.79167	478.4959	32.0		45
18.92083	478.9594	32.2		31

How to do it in R

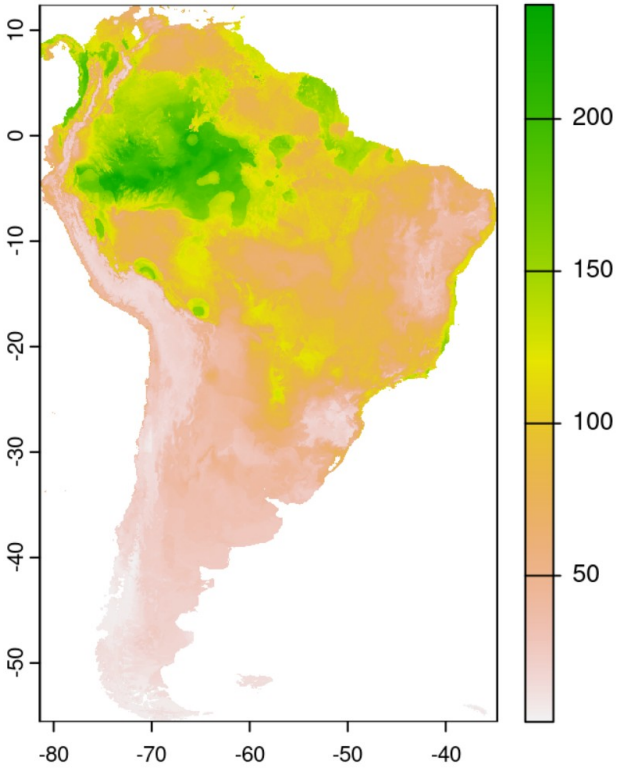
```
library(caret)
model <- train(predictors,
                response,
                method="rf")
```

*Explanation of variables:
https://www.worldclim.org/data/bioclim.html

Step 2: Model prediction in R



+ trained model =



How to do it in R

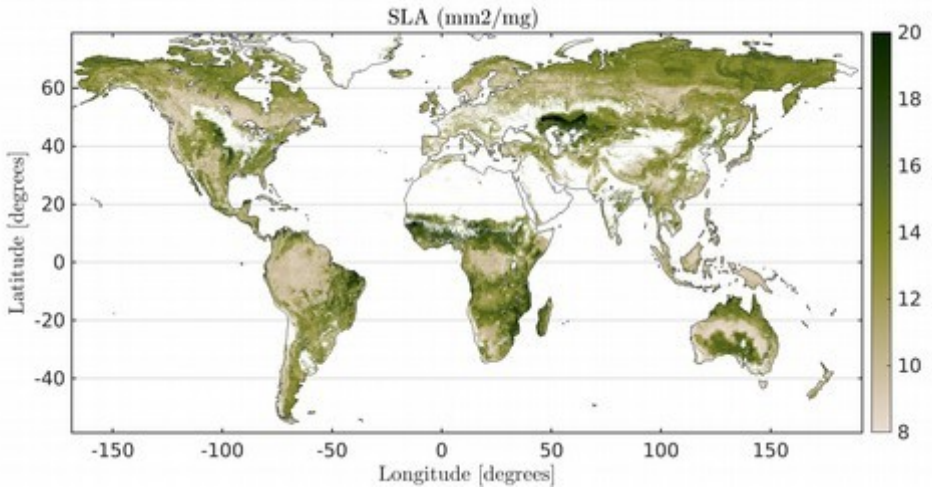
```
predictors_sp <- rast("predictors.tif")
prediction <- predict(predictors_sp,model,na.rm=TRUE)
plot(prediction)
```

Time for practice

Train a random forest model using all available bioclimatic and terrain variables and deploy it to entire South America

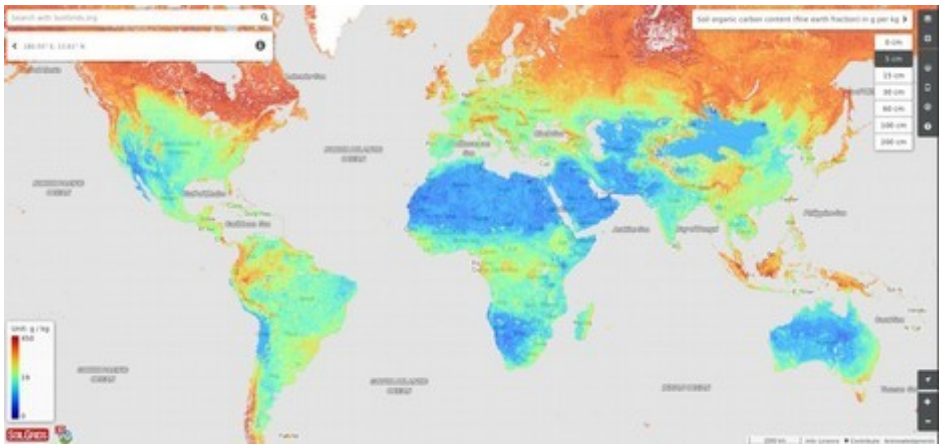
Coffee break

Global maps of ecosystem variables based on machine learning (a few examples)

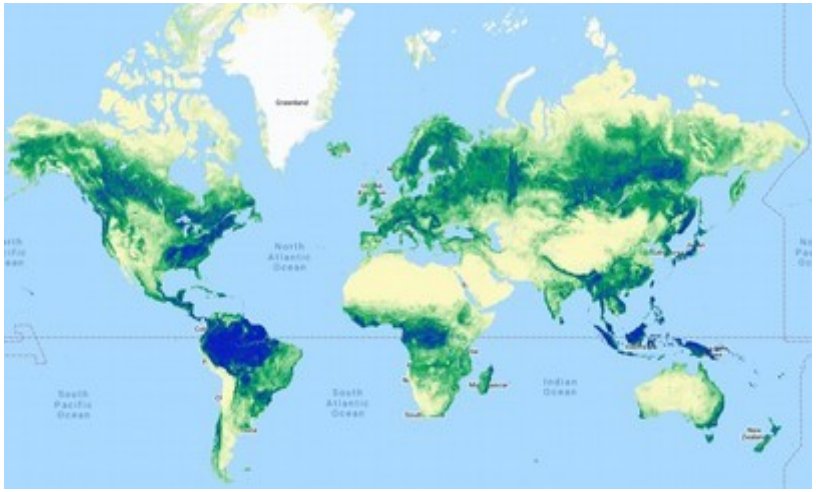


Moreno-Martínez et al., 2018

Based on van den Hoogen et al., 2019



Hengl et al., 2017



Bastin et al. 2019

Machine learning as a magic tool to map everything ?

...but there are increasingly doubts about the quality of these results

Wissenschaft

Wenn die KI daneben liegt

Welche Fehler drohen, wenn Forscher Wissenslücken per Computer schließen wollen, zeigen zwei aktuelle Klimastudien.

Von **Tin Fischer**

6. November 2019, 16:44 Uhr / Editiert am 9. November 2019, 17:42 Uhr / DIE ZEIT Nr. 46/2019, 7. November 2019 / 9 Kommentare

AUS DER ZEIT NR. 46/2019

DEEP TROUBLE FOR DEEP LEARNING

BY DOUGLAS HEAVEN

Nature 574, 163-166 (2019)

Home / News & Opinion

Researchers Find Flaws in High-Profile Study on Trees and Climate

Four independent groups say the work overestimates the c
global forest restoration, but the authors insist their origina

Oct 17, 2019
KATARINA ZIMMER

Comment | Published: 23 August 2021

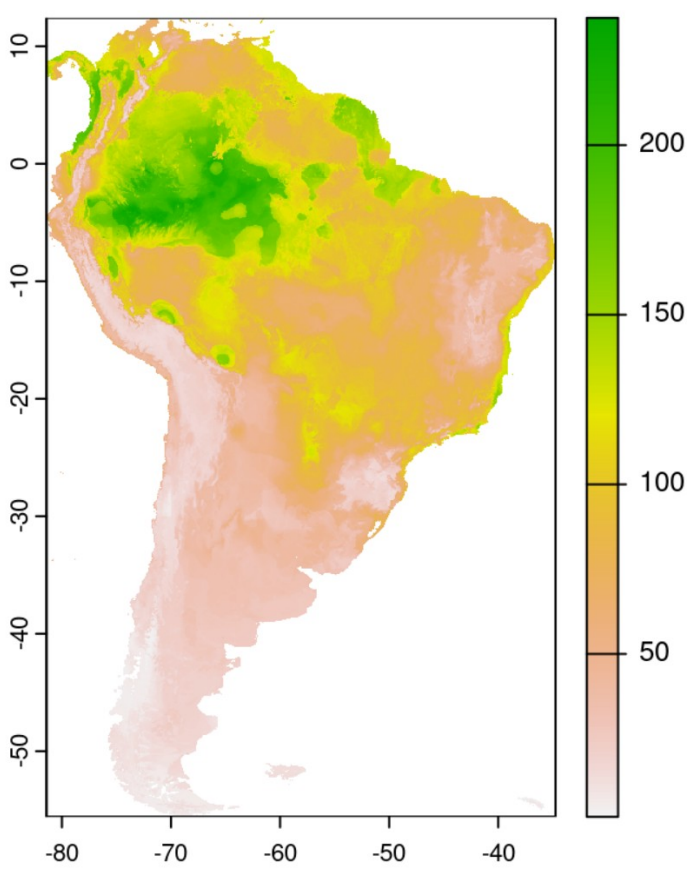
Conservation needs to break free from global priority mapping

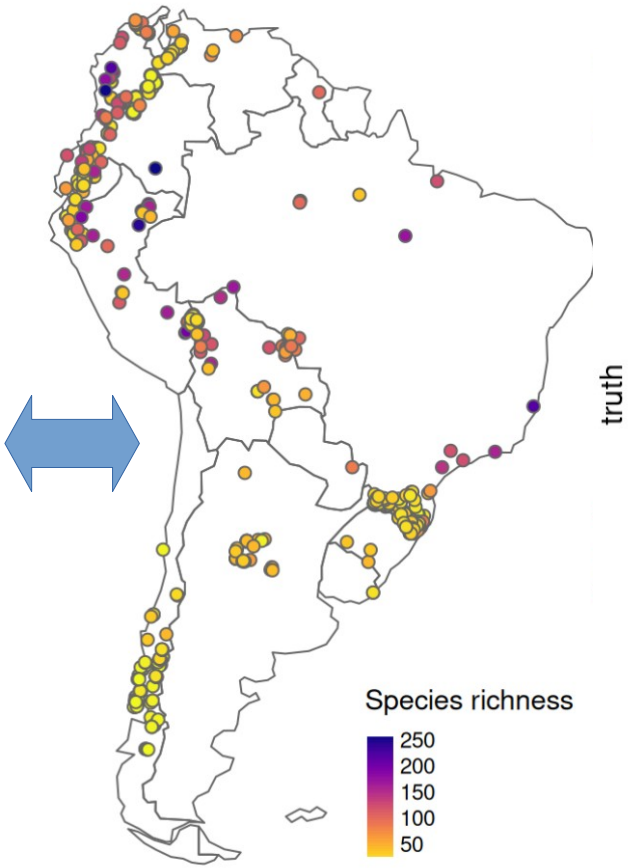
Carina Wyborn & Megan C. Evans

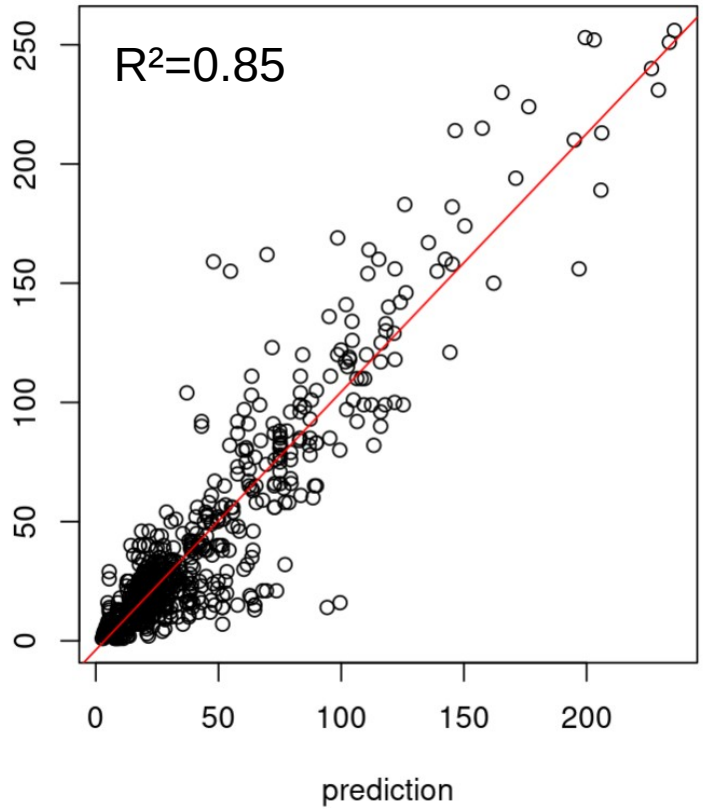
Nature Ecology & Evolution (2021) | Cite this article

Have we been too ambitious? What is the quality of our predictions?

Step 3: Quality assessment
First idea: How well can we predict our reference data?

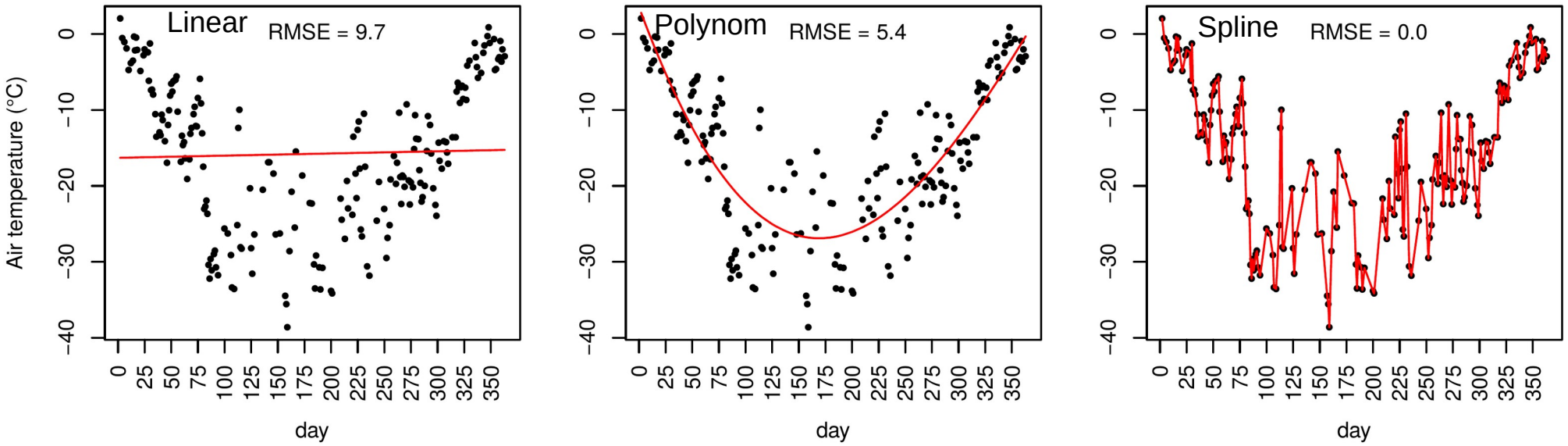






Is that a fair way of model validation?

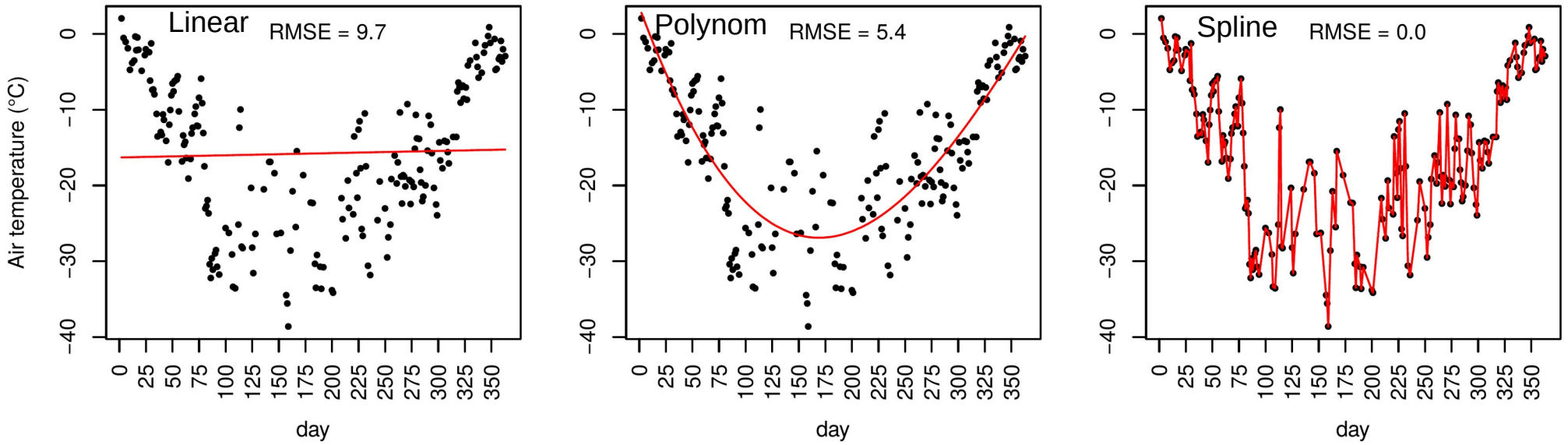
Problem: Overfitting



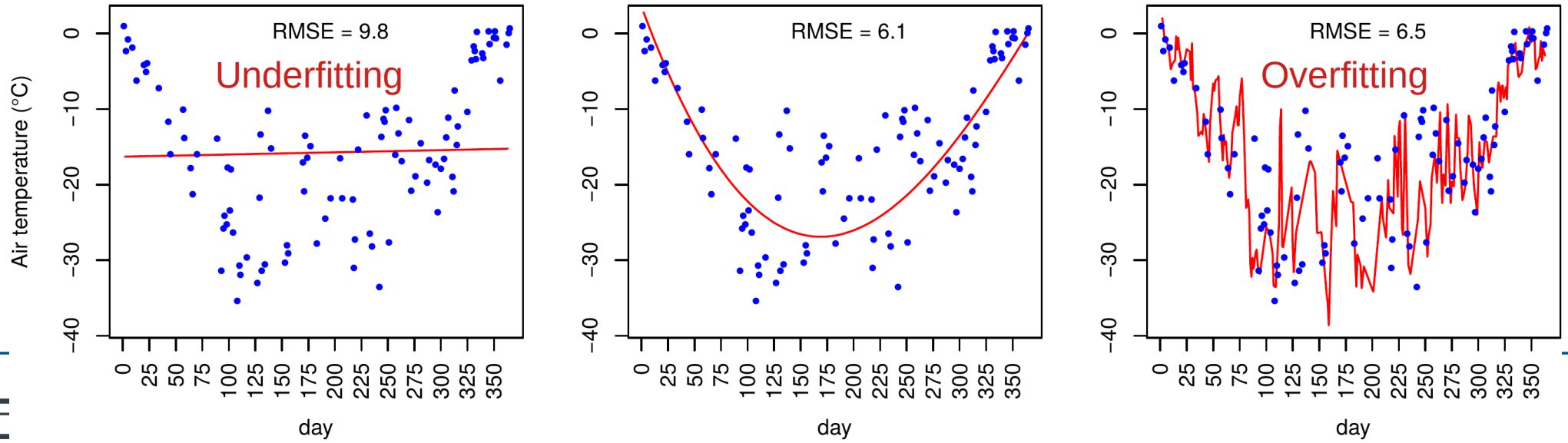
Which one is the best model?

Problem: Overfitting

Training (2/3 of the data)



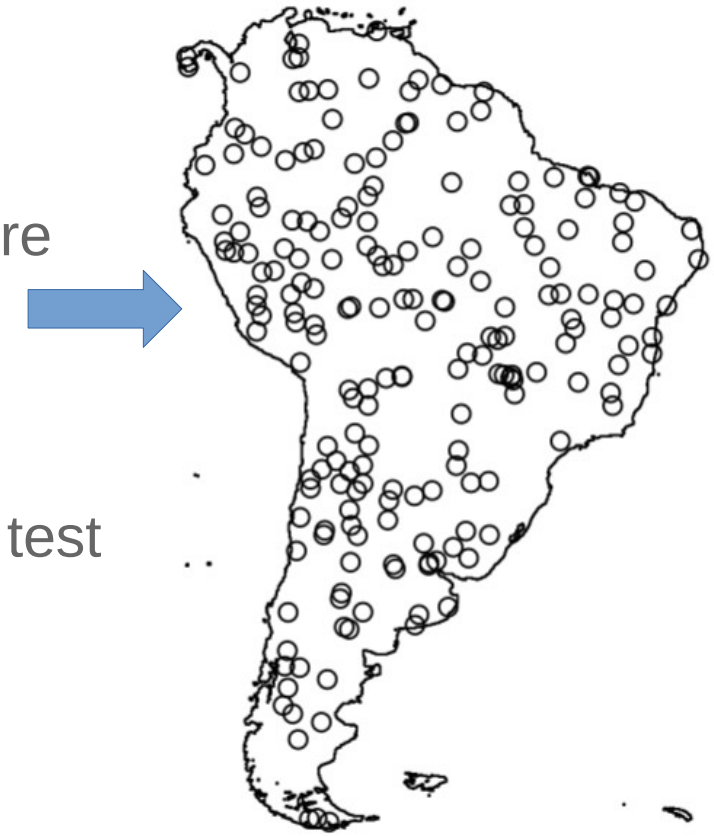
Validation (1/3 of the data)



We need to look at independent data instead!

How to do this?

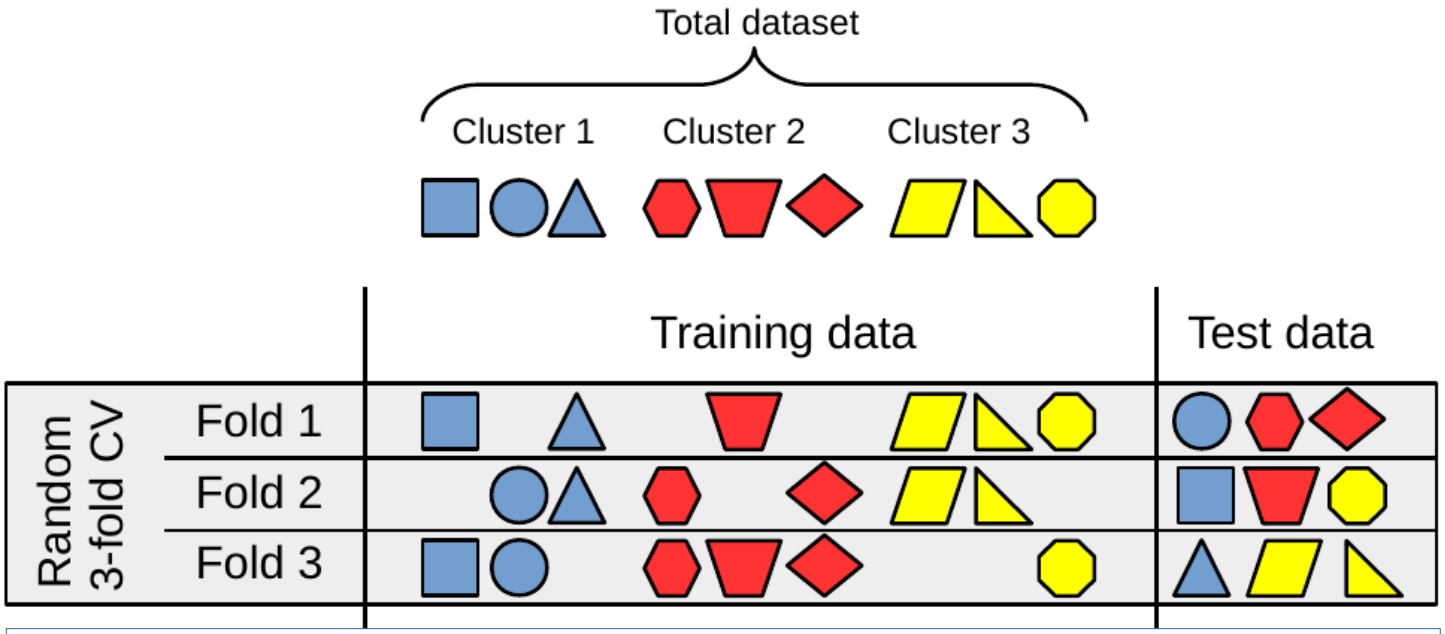
- Ideally: Use a representative sample of the entire prediction area, e.g. a random sample



- If this is not available/possible:
  - Split the available data into training and test
  - Or use cross-validation if the data set is small (main purpose: tuning!)

Either way: How do we split the data?

Quality assessment via random cross-validation



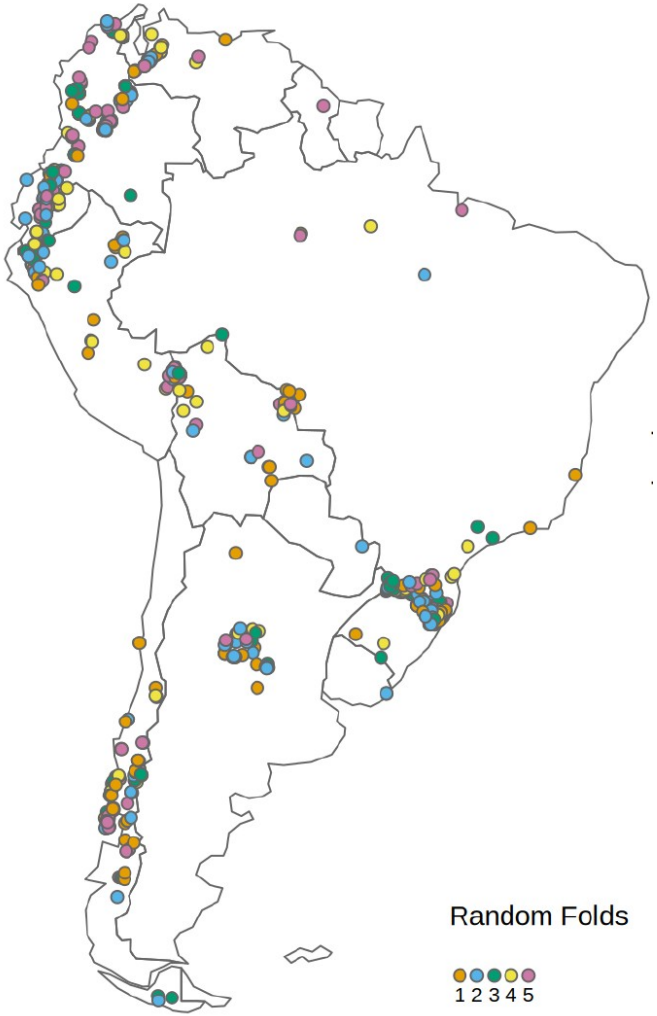
How to do it in R

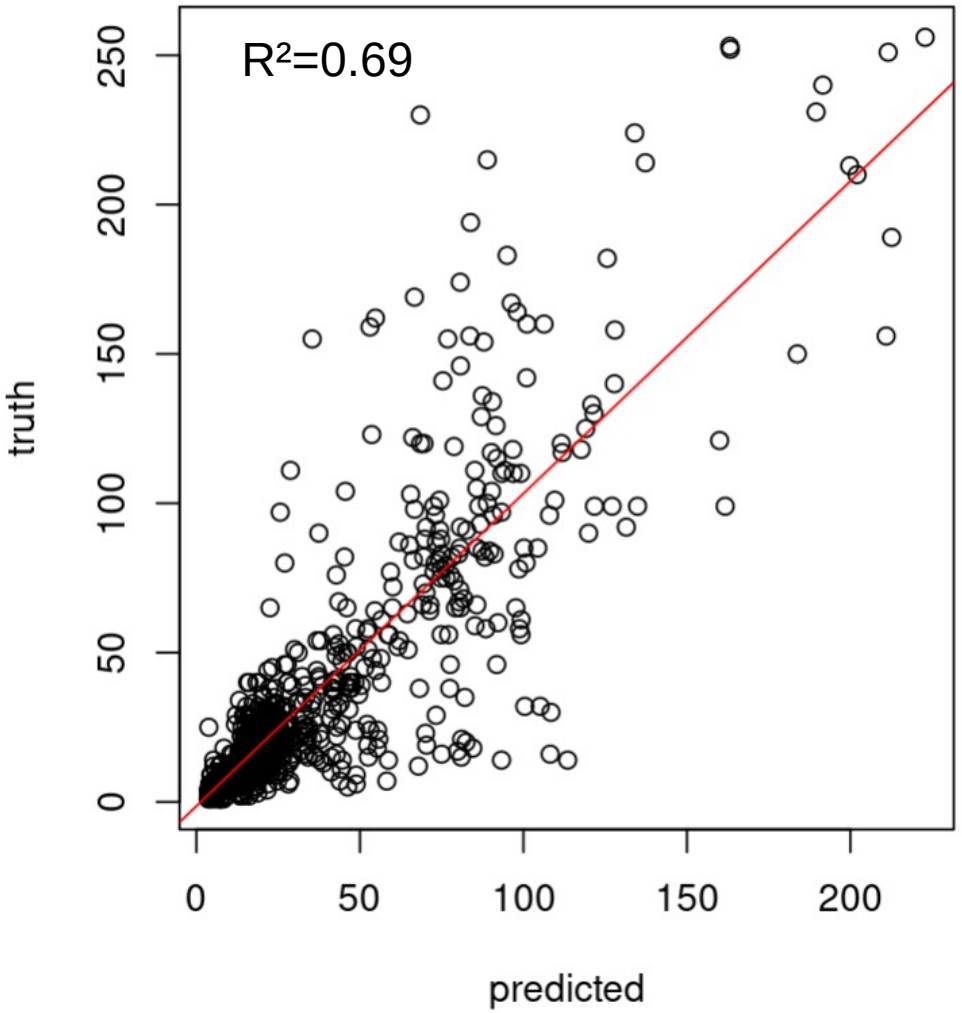
```
ctrl <- trainControl(method="cv",
                     number=5,
                     savePredictions="final")

model <- train(trainDat[,predictors],
                       trainDat$Species_richness,
                       method="rf",
                       trControl = ctrl)
```

Time for practice

Re-train your random forest model using random cross-validation.
What is the estimated model performance?

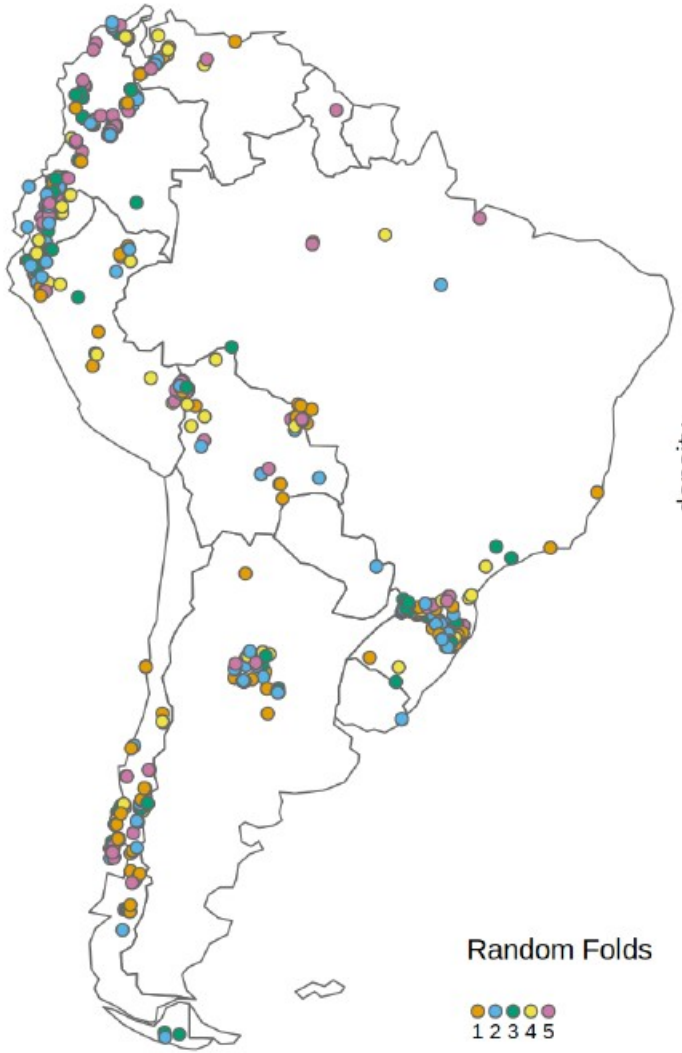
Quality assessment via random cross-validation

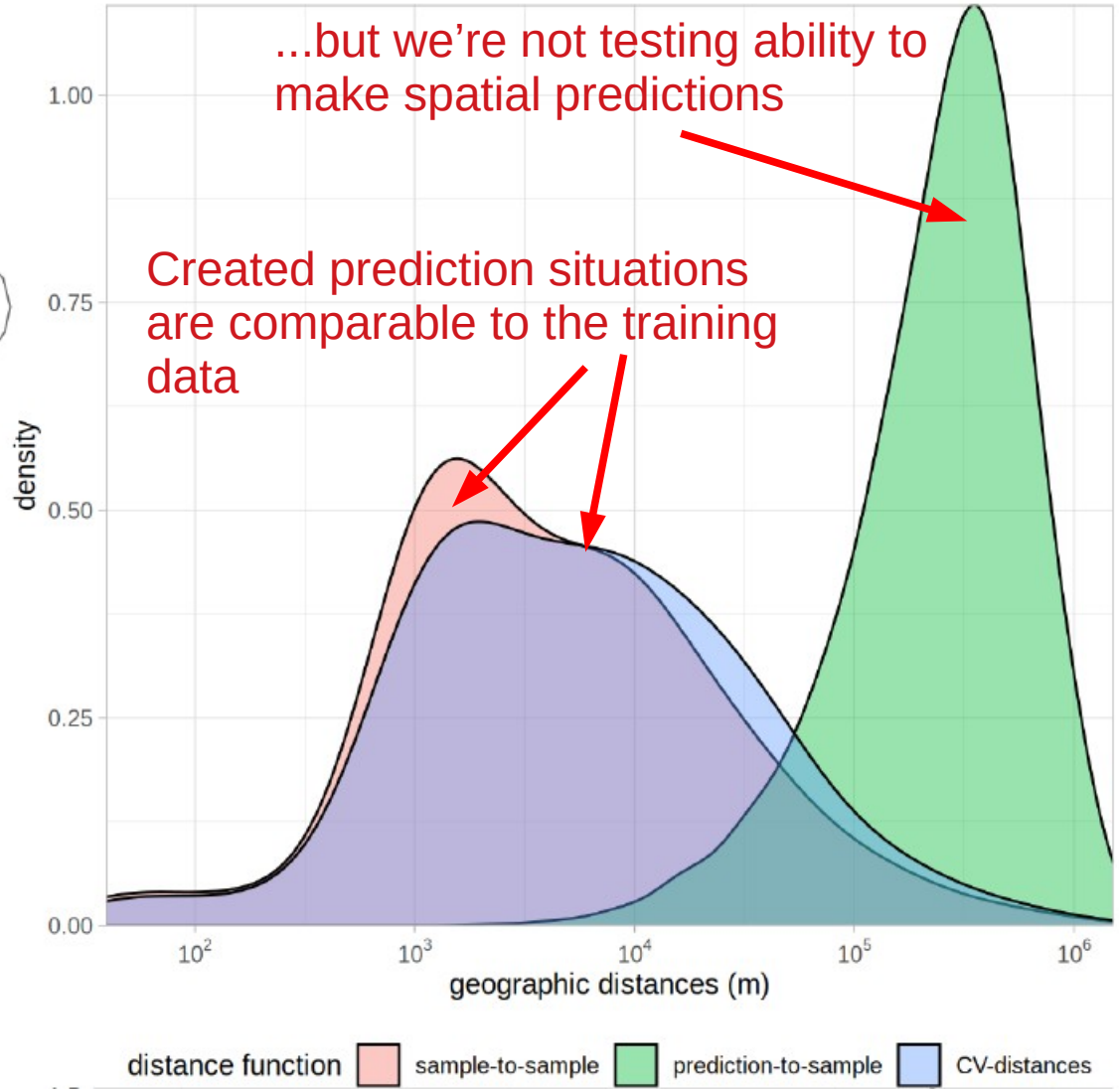




But can we consider this an independent validation?

… but is this representative for our map?





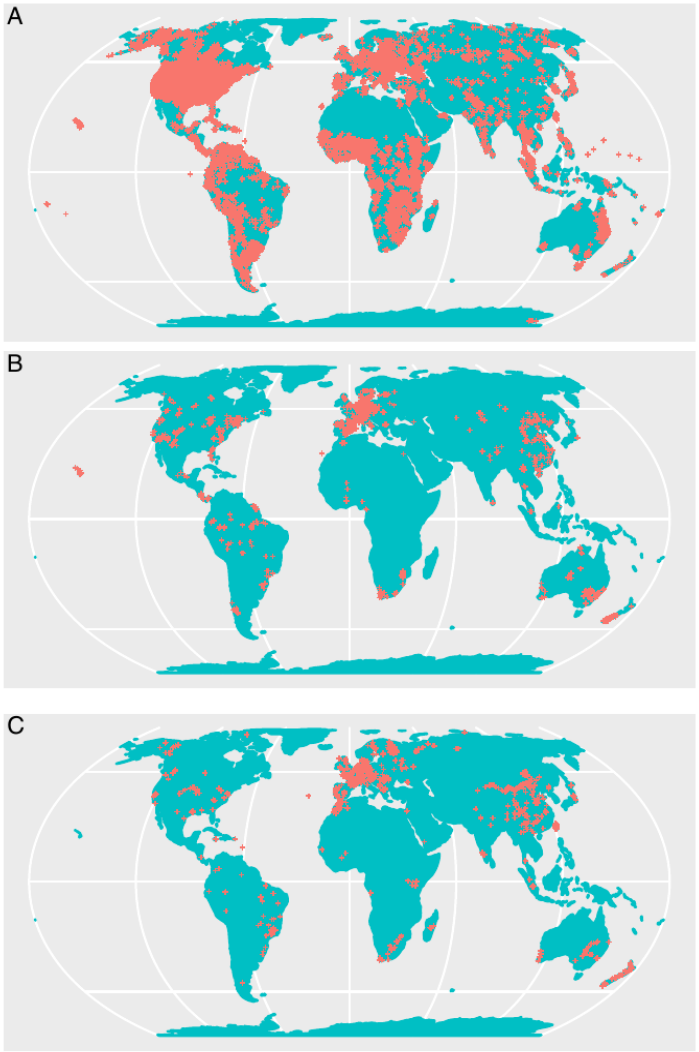
→ **We need spatial cross-validation!**

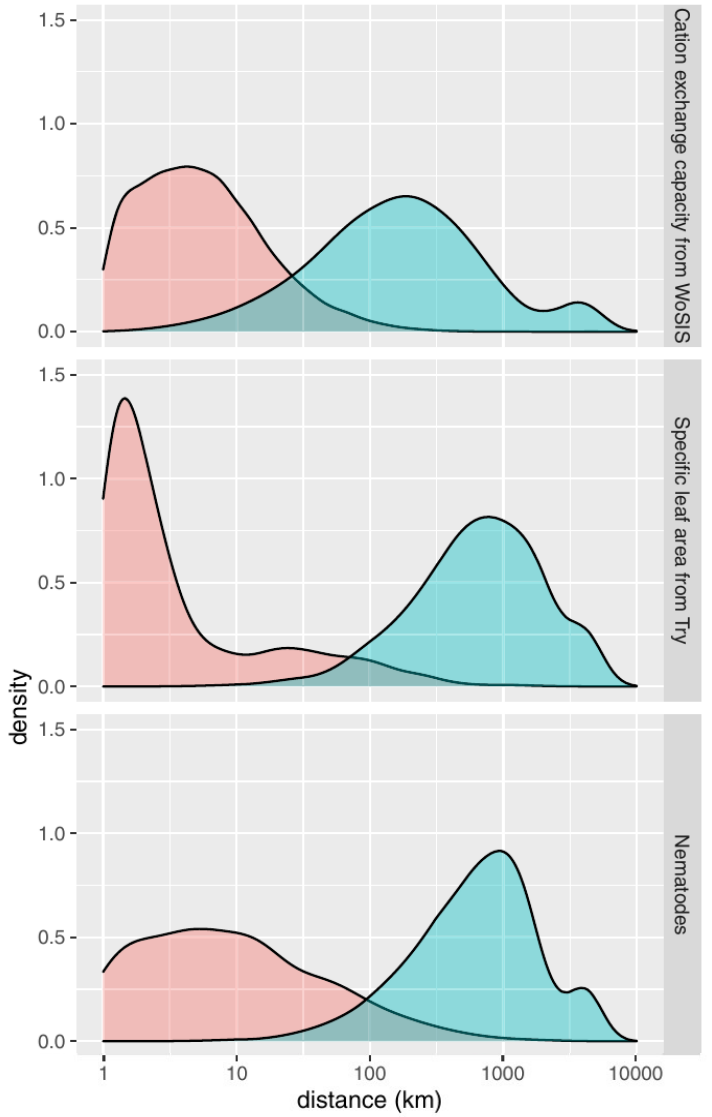
Clustered reference data is an issue in many ecological studies

Soil maps

Plant traits

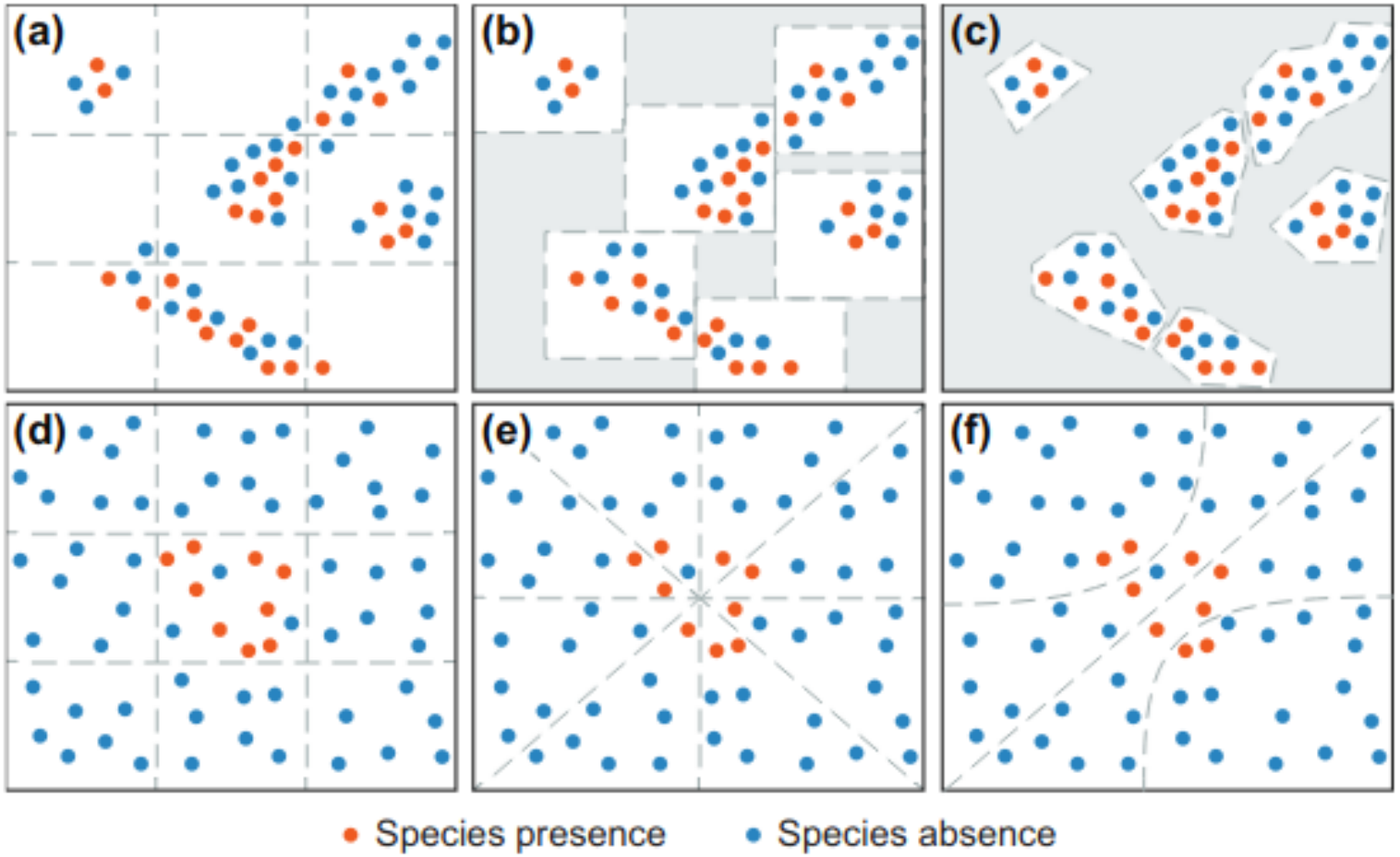
Nematodes





Meyer & Pebesma (2022)

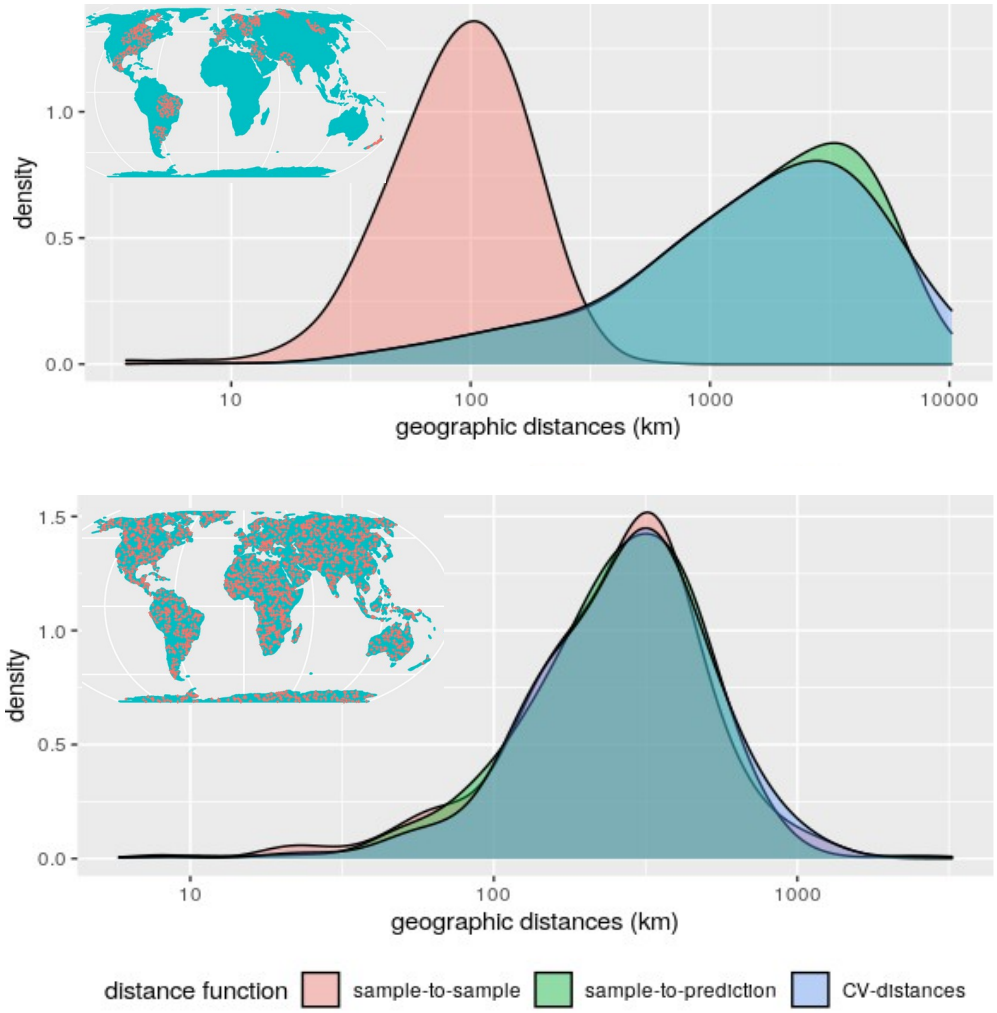
Suggestion: Spatial cross-validation



Roberts et al., 2017

...but which one to choose?

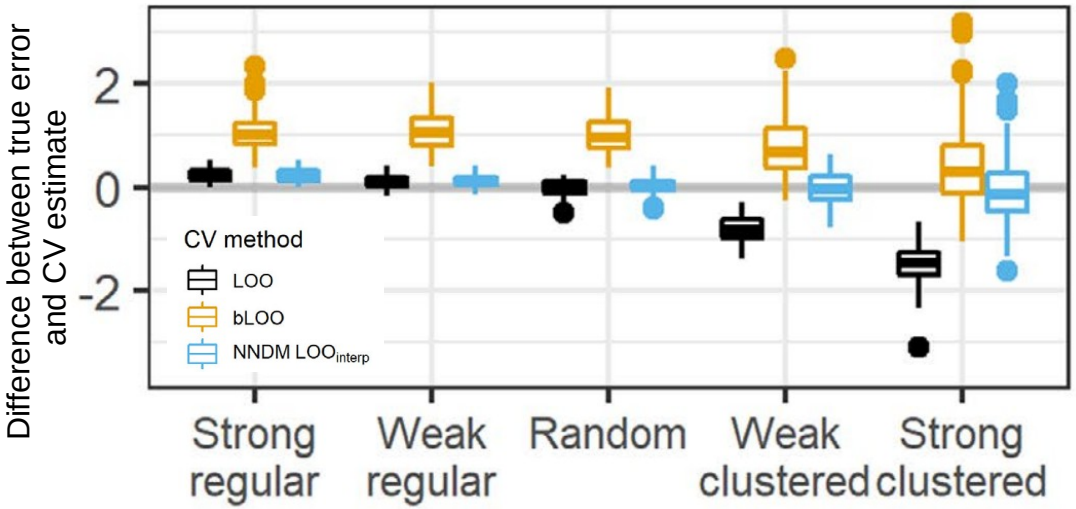
Suggestion: “k-fold nearest neighbor distance matching CV”



Received: 20 September 2021 | Accepted: 8 March 2022
DOI: 10.1111/2041-210X.13851
Methods in Ecology and Evolution
RESEARCH ARTICLE

Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation

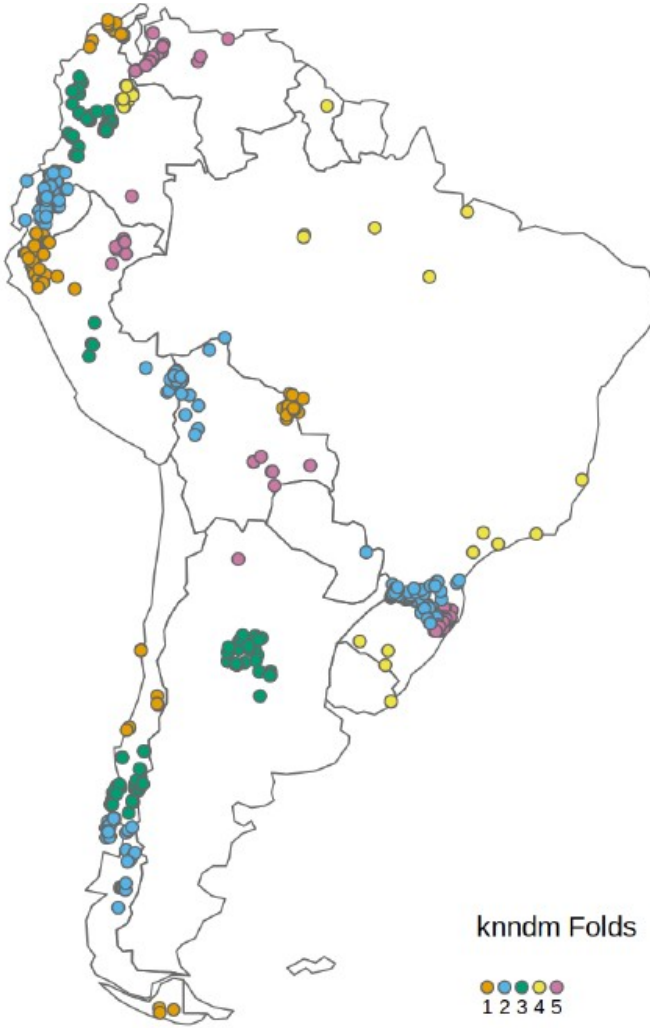
Carles Milà[1] | Jorge Mateu[2] | Edzer Pebesma[3] | Hanna Meyer[4]



Mila et al., 2022

Reproduce figures: hannameyer.github.io/CAST/articles/cast04-plotgeodist.html

Better approach: knndm cross-validation



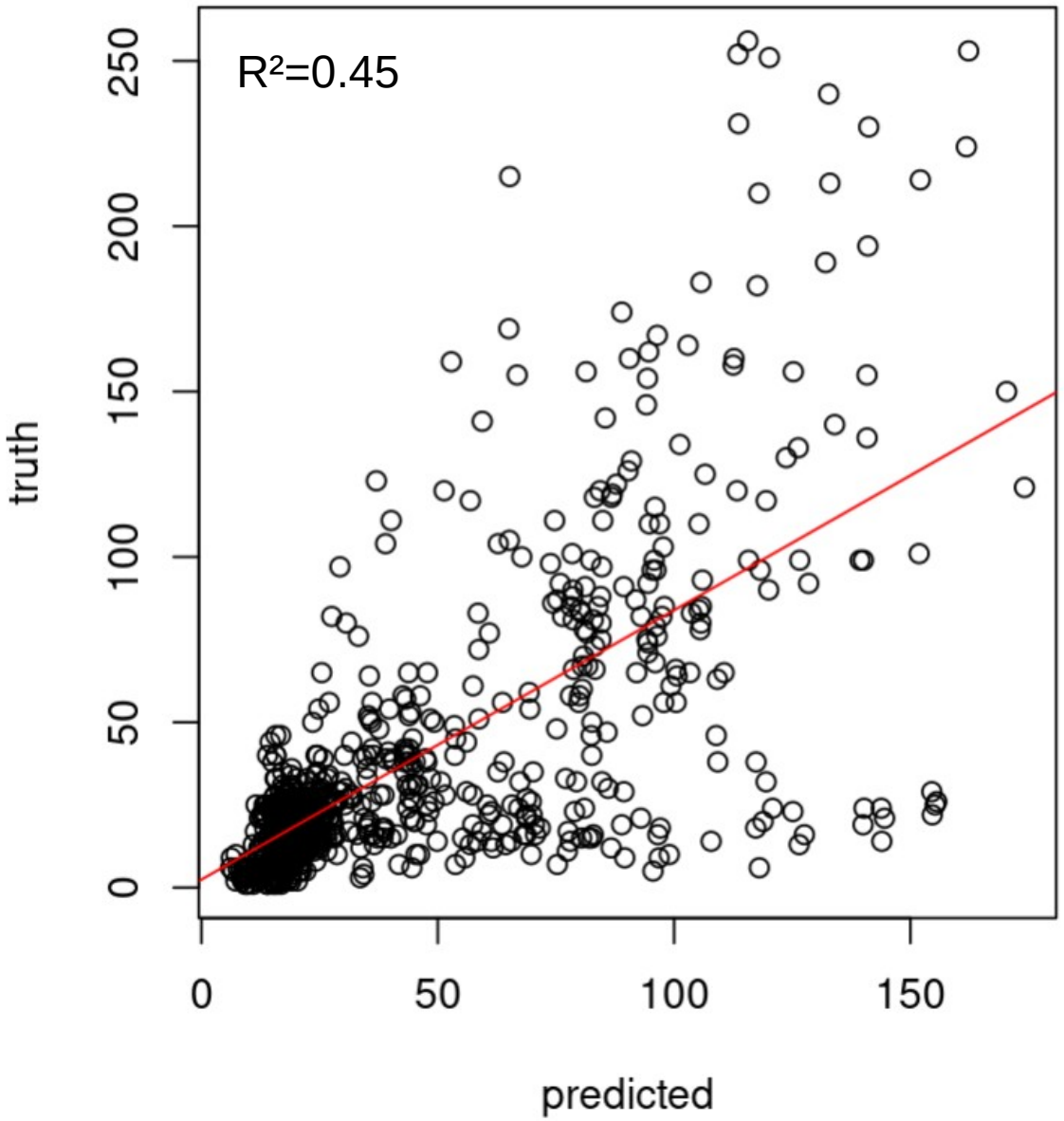
How to do it in R

```
knndm_folds = knndm(tpoints = splotdata,
                    modeldomain = modeldomain,
                    k = 5)

ctrl <- trainControl(method="cv",
                     index = knndm_folds$indx_train,
                     indexOut = knndm_folds$indx_test,
                     savePredictions = "final")
```

Time for practice

Re-train your random forest model using knndm spatial cross-validation.
What is the estimated model performance?

Summary cross-validation



Validation strategy	Performance
Training data	0.85
Random CV	0.69
“Spatial” knndm CV	0.44

- Standard validation procedures lead to an overoptimistic view on prediction performance!
- Prediction situations created during CV need to resemble those encountered while predicting the map from the reference data

Coffee break!

...but the relevance of spatial validation is still highly underestimated

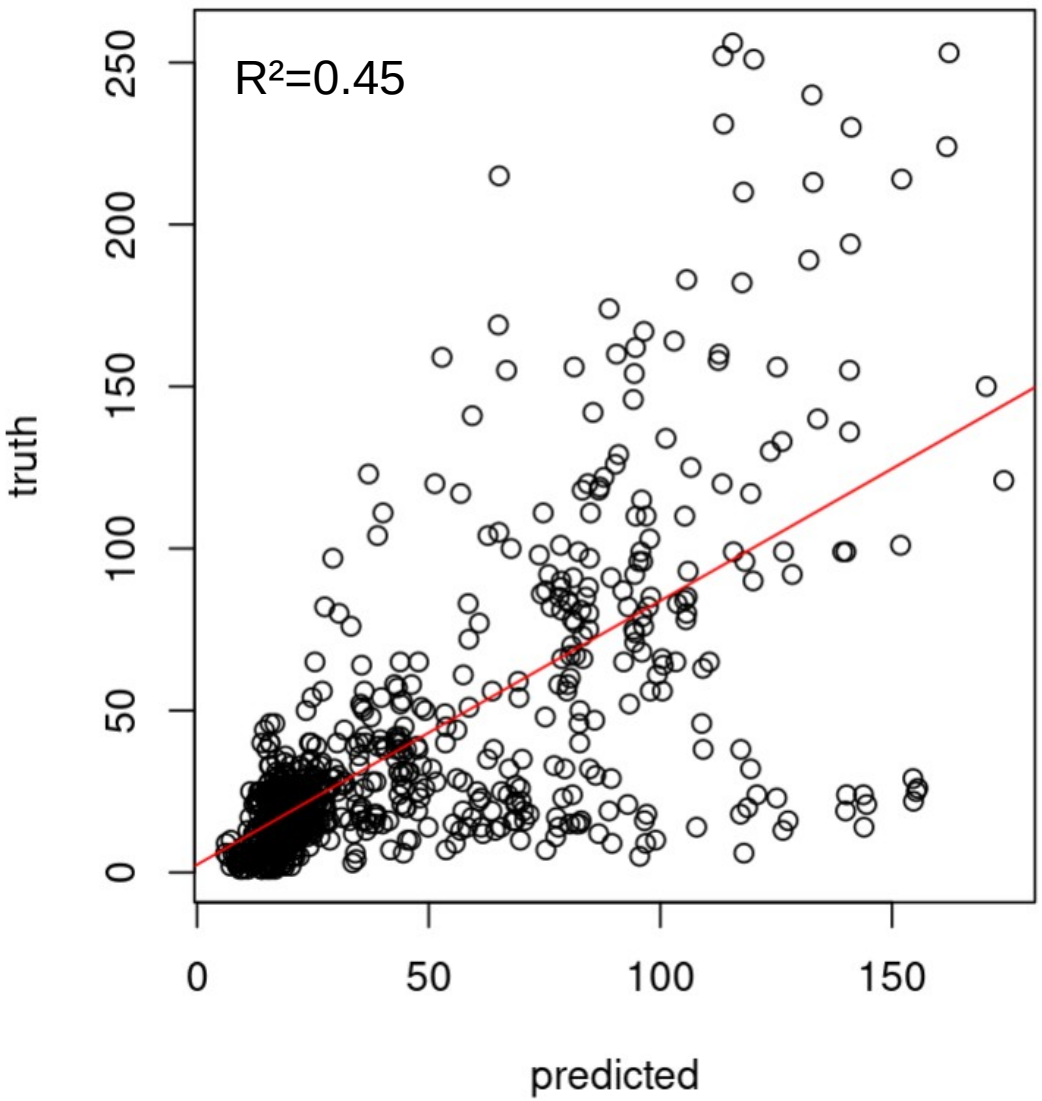
"*I am actually surprised to see the poor performance of your NN approach[...]. Typically with sufficient training data a NN approach can often reproduce the predicted variable very well even if the underlying reasons are unknown*"
(an editor from a high impact journal in the remote sensing community)

Data reproduction is not the same as data prediction!

Random cross-validation!

Spatial cross-validation!

How can we improve the spatial predictions?

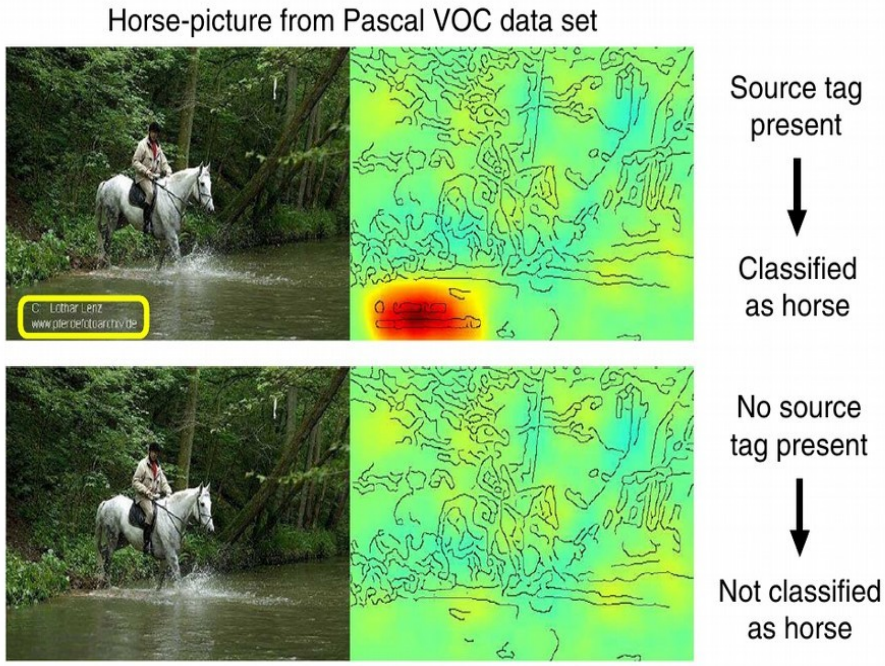


https://xkcd.com/1838/

Problem: The model cannot well generalize...but why?

Is the model behaving like the “clever Hans” ?

https://commons.wikimedia.org/wiki/File:Osten_und_Hans.jpg#/media/Datei:Osten_und_Hans.jpg



“Unmasking Clever Hans predictors and assessing what machines really learn” (Lapuschkin et al., 2019, Nature communications)

Models that are not able to learn the scientifically meaningful relationships → not transferable!

Can we improve the model by removing "clever Hans predictors" ? → Spatial variable selection

for *each resampling iteration* **do**
- Partition the data into training and test data
- Tune and train models using all possible 2-variable combinations
- Predict on test data and calculate model performance

end

Spatial cross-validation!

Which 2 variables lead to the best model?

Keep the best performing 2-variable model ($model_{best}$)

for *each additional number of variables i, i=3...N* **do**

- **for** *each remaining variable* V_R **do**
  - **for** *each resampling iteration* **do**
    - Partition the data into training and test data
    - Tune and train models using the variables of $model_{best}$ and V_R
    - Predict on test data and calculate model performance
  - **end**
- **end**

Spatial cross-validation!

Which further variables improve the model?

- **if** *mean(error of* $model_i$*) > mean(error of* $model_{best}$*)* **then**
  - break
- **end**
- Keep the best performing i-variable model ($model_{best}$)

end

Step 4: Improve the model via spatial variable selection

How to do it in R

```
ffsmodel <- ffs(trainDat[,predictors],
                trainDat$Species_richness,
                method="rf",
                metric="Rsquared",
                tuneGrid=data.frame("mtry"=2),
                ntree=50,
                trControl=ctrl_knndm)
```

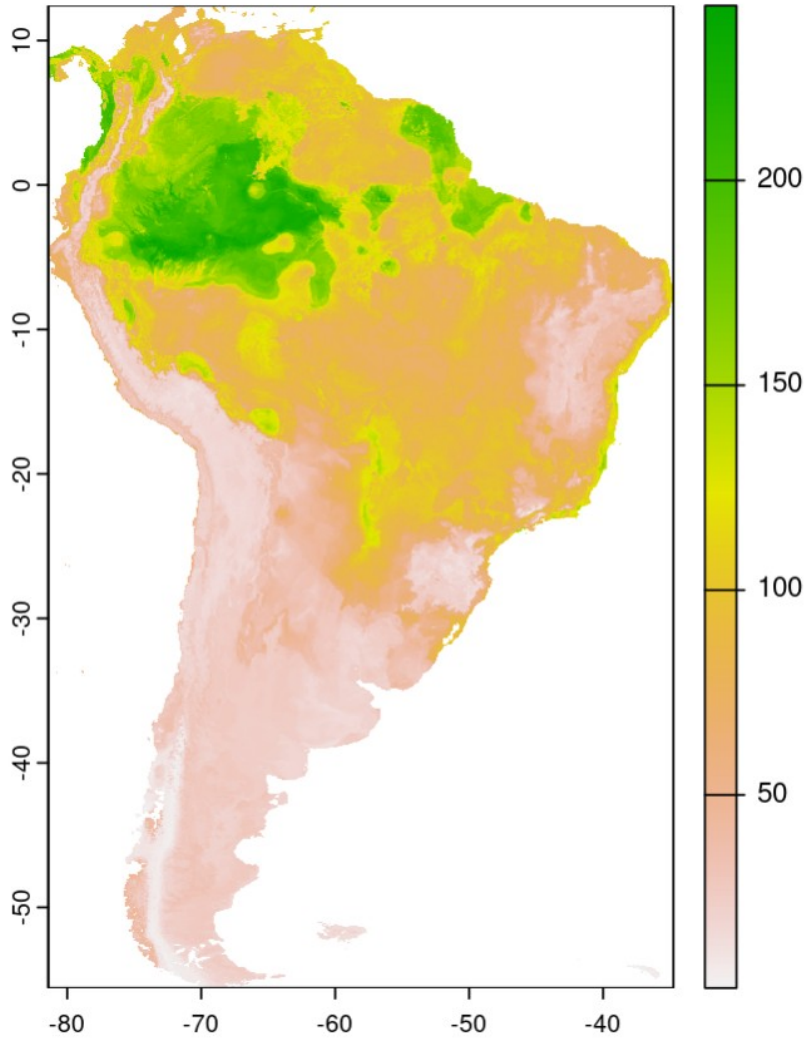
...after variable selection which includes final model training: predict!

Time for practice

Can we improve our predictions by spatial variable selection?

Step 4: Improve the model via spatial variable selection

Validation strategy	Performance
Training data	0.85
Random CV	0.69
“Spatial” knndm CV	0.44
Improved model “Spatial” knndm CV	0.49

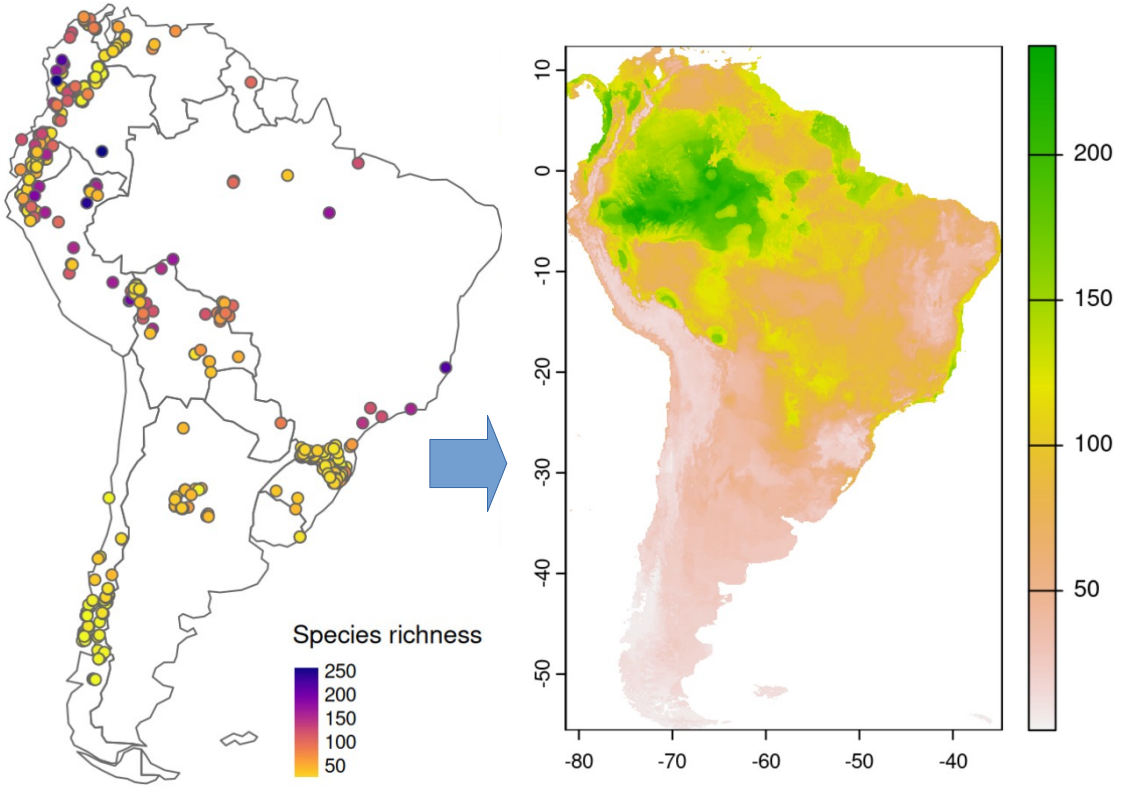


What we have learned so far...

- Cross-validation strategy affect:
  - Performance estimate
  - Selected hyperparameters (not discussed here)
  - Variable selection
- Consequences of using an unsuitable CV:
  - Unreliable performance estimates
  - Models that can well reproduce but not necessarily predict ("clever Hans effect")
- Hence, CV strategies that fit the prediction task are required during model selection and validation!

But is this sufficient for reliable mapping ?

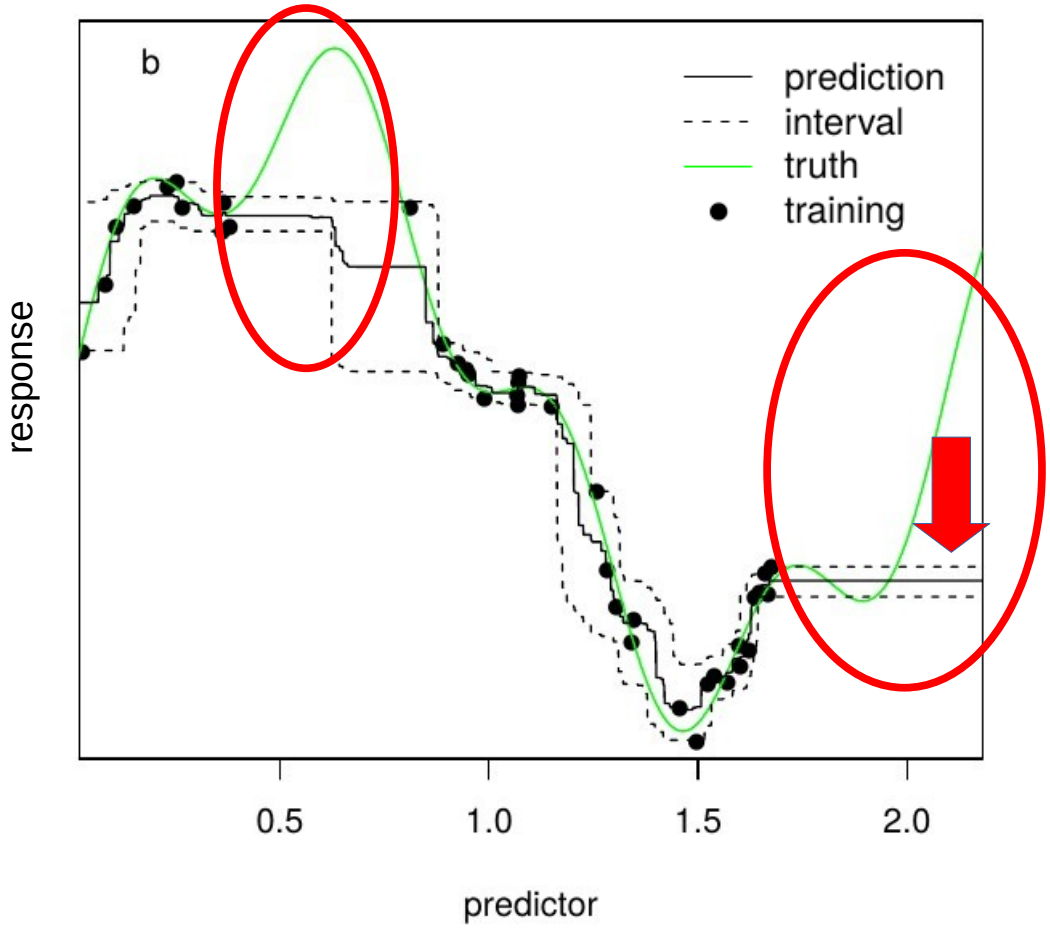
Limits to accuracy assessment



- Mapping requires prediction far beyond clustered reference data
- Transfer to new space required
- New space might differ in environmental properties

What happens if the model has never “seen” such new environments?

Machine learning models are weak in extrapolations



- Machine learning can fit very complex relationships.
- But gaps in predictor space are problematic (the model has no knowledge about these areas!)
- A measure for the “unknown” is needed!

Meyer & Pebesma (2021)

Suggestion: Area of Applicability (AOA)

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE | Open Access

Predicting into unknown space? Estimating the area of applicability of spatial prediction models

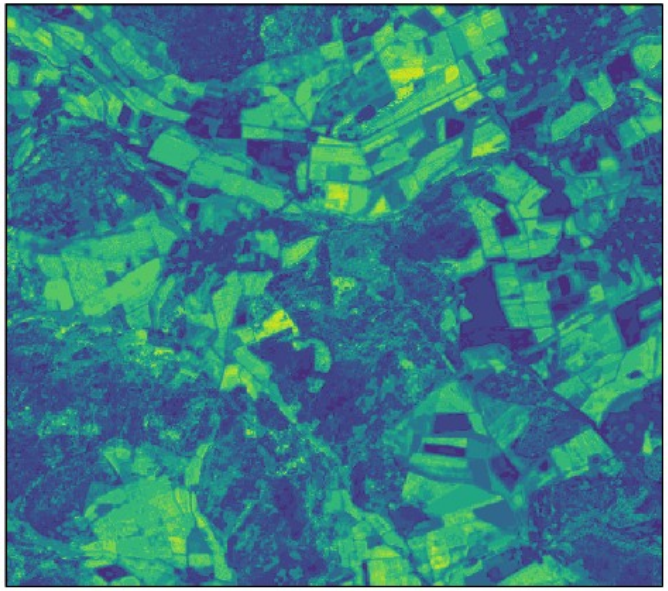
Hanna Meyer, Edzer Pebesma

We try to derive the area...

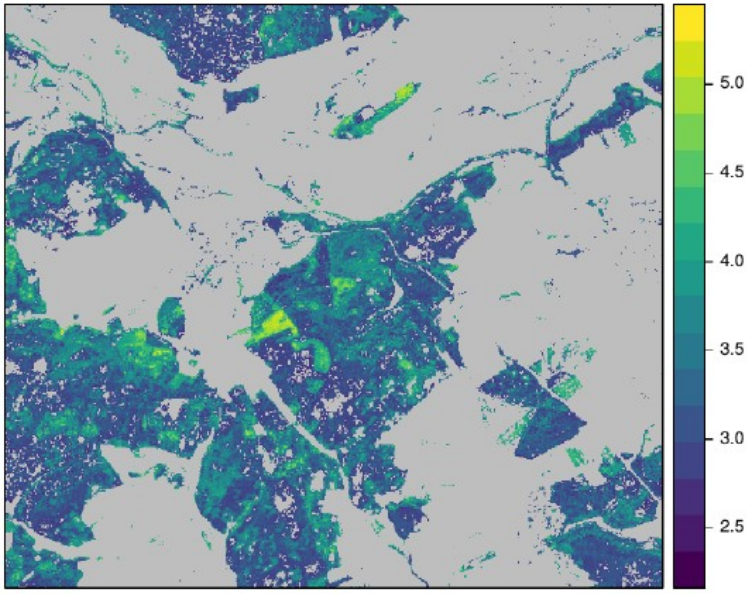
- to which the model can be applied because it has been enabled to learn about relationships
- where the estimated performance holds

Sentinel-2 scene and training data points of leaf area index

Predictions



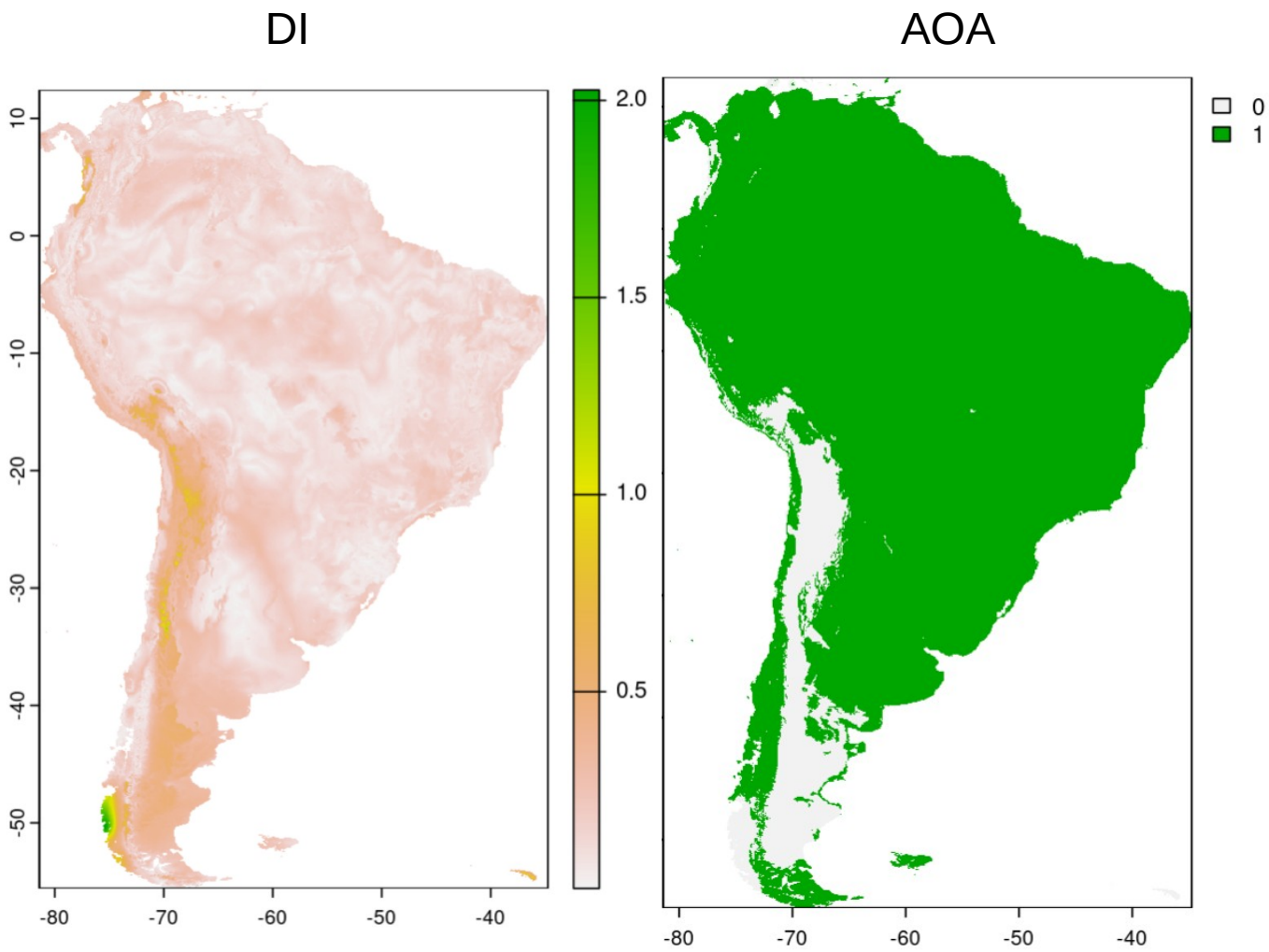
Predictions limited to the AOA



Step 5: Assessment of the Area of Applicability

How to do it in R

```
AOA <- aoa(predictors_sp,
           model)

plot(AOA$DI)
plot(AOA$AOA)
```

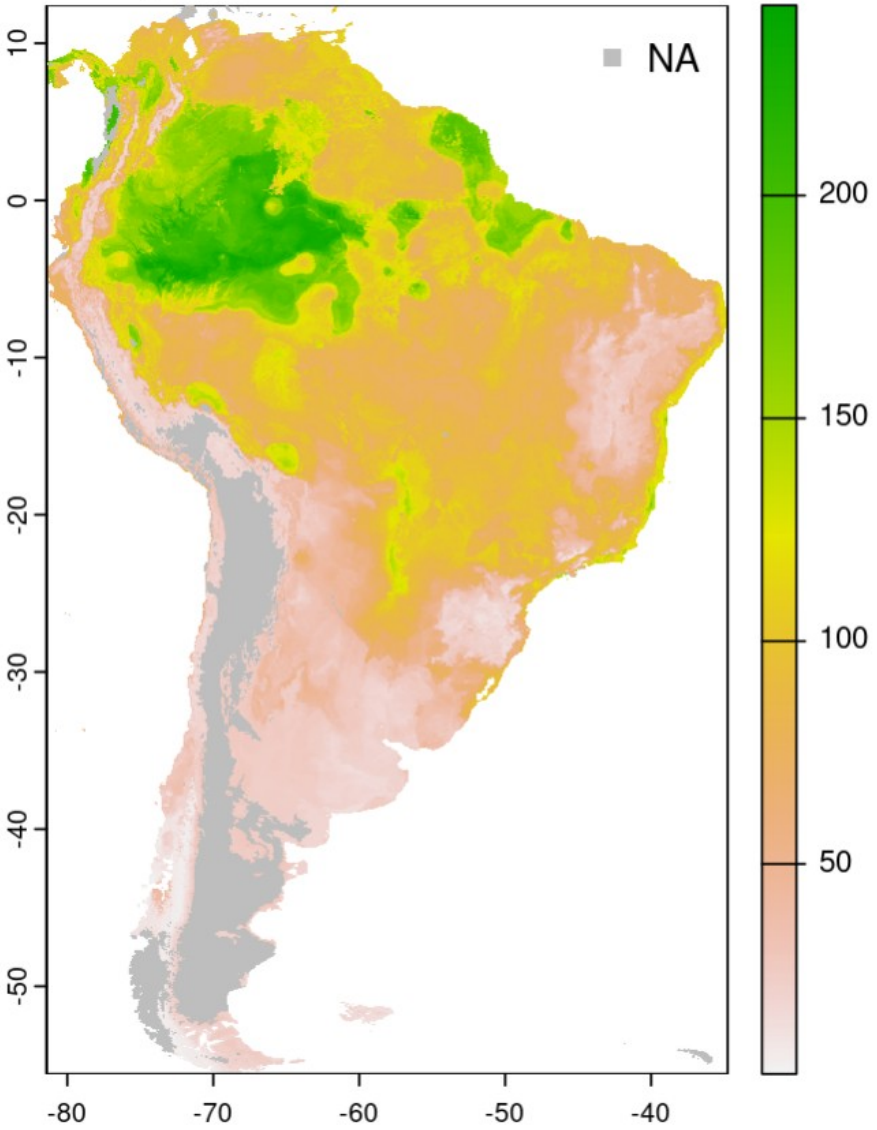


Time for practice

Calculate the area of applicability of your model.
Present your final predictions of species richness for South America for the AOA and with the corresponding cross-validation estimates.

Present final results

Predicted species richness



Estimated accuracy based on spatial cross-validation:

$R^2 = 0.49$
RMSE = 21.3

Why is it relevant to map “unknown space”?

Results are not just nice maps but used for...

- subsequent modeling
- nature conservation
- risk assessment
- …

COMMENT

https://doi.org/10.1038/s41467-022-29838-9 OPEN

Machine learning-based global maps of ecological variables and the challenge of assessing them

Hanna Meyer [1]✉ & Edzer Pebesma [2]✉

Our opinion: predictions should only be presented for the area of applicability to avoid error propagation or misplanning

Coming back to the aims of this workshop

- After this course you should ideally be able to:
  - Know and apply the basic workflow of spatial mapping via machine learning
  - Assess the quality of your predictions
  - Understand major risks and pitfalls
  - Critically analyze the results
  - Communicate the quality and limitations of predictions

Train and predict

Cross-validation

Clever Hans effect

Area of applicability

Further reading

- Textbooks for machine learning / statistical learning:
  - Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. (First ed., pp. 141 – 145). New York: Springer
  - James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R (First ed.). New York: Springer.
  - Book chapter “The CAST package for training and assessment of spatial prediction models”. https://arxiv.org/abs/2404.06978
- Talk and tutorial from the OpenGeoHub 2022 summer school on Machine learning-based maps of the environment - challenges of extrapolation and overfitting, including discussions on the area of applicability and the nearest neighbor distance matching cross-validation (https://doi.org/10.5446/59412).
- See https://hannameyer.github.io/CAST/index.html for more tutorials and links to scientific papers on the presented methods