Universität Münster

Assessing the area of applicability of spatial prediction models through a local data point density approach

Fabian Schumacher[1], Christian Knoth[2], Marvin Ludwig[3], and Hanna Meyer[4]

[1,2] University of Münster, Institute for Geoinformatics, Heisenbergstraße 2, 48149 Münster
[3,4] University of Münster, Institute of Landscape Ecology, Heisenbergstraße 2, 48149 Münster

Institute of Landscape Ecology

Remote Sensing & Spatial Modelling

ifgi
Institute for Geoinformatics
University of Münster

Previous slide

Next slide

Universität Münster

Abstract

Contents

Universität Münster

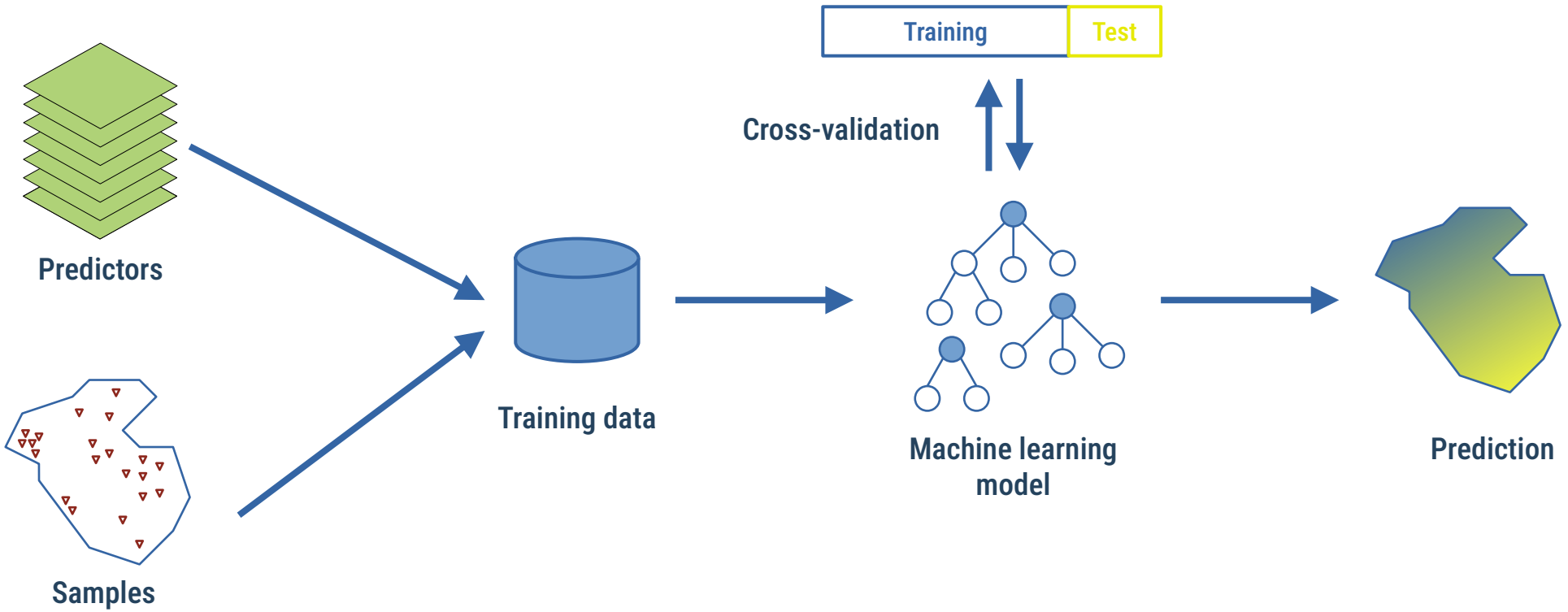
1 Introduction (2-minute oral)
2 Methods
3 Results
4 Discussion
5 Conclusion
6 How to use?

Universität Münster

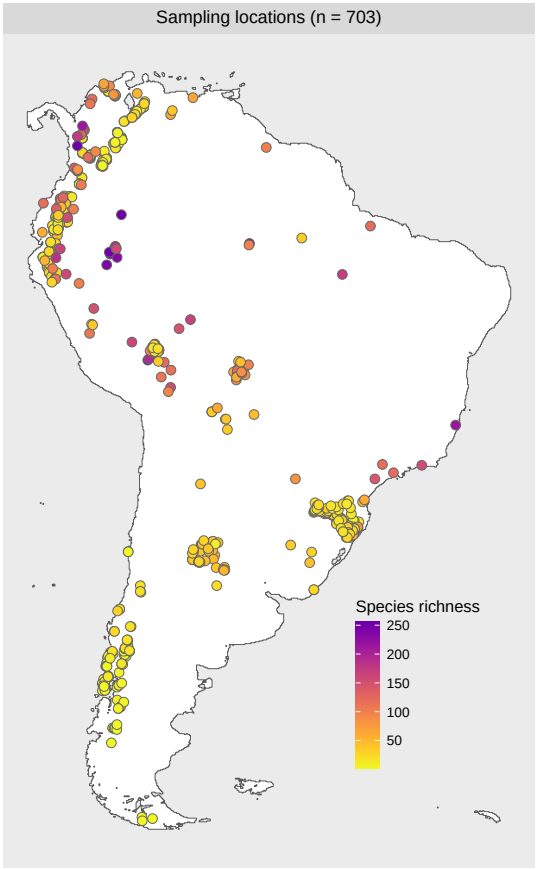
Introduction

Spatial modelling workflow

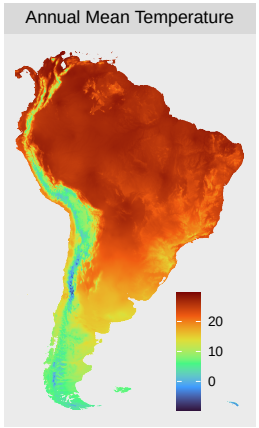
Universität Münster

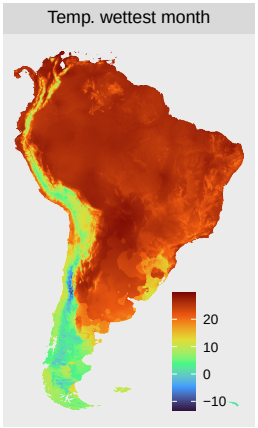


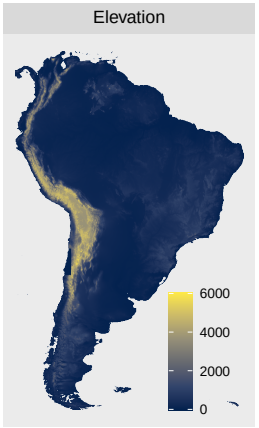
Samples

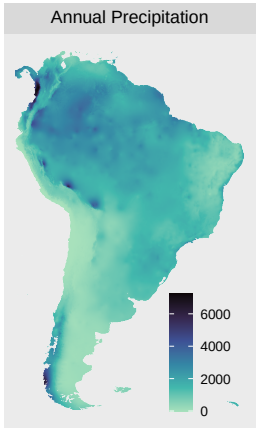


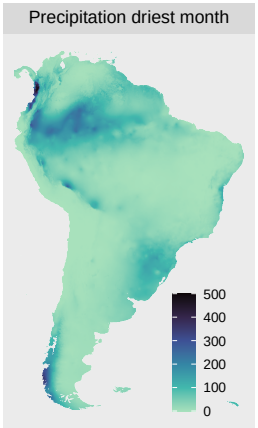
Predictors



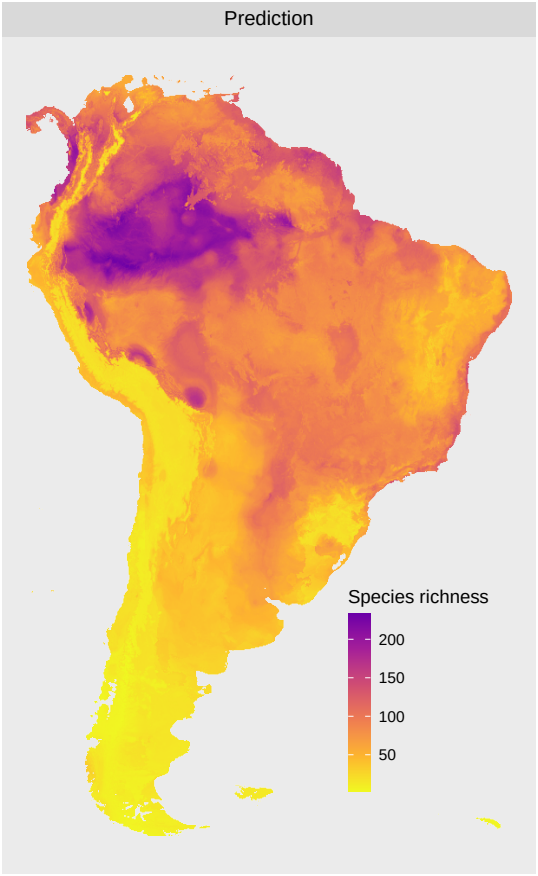




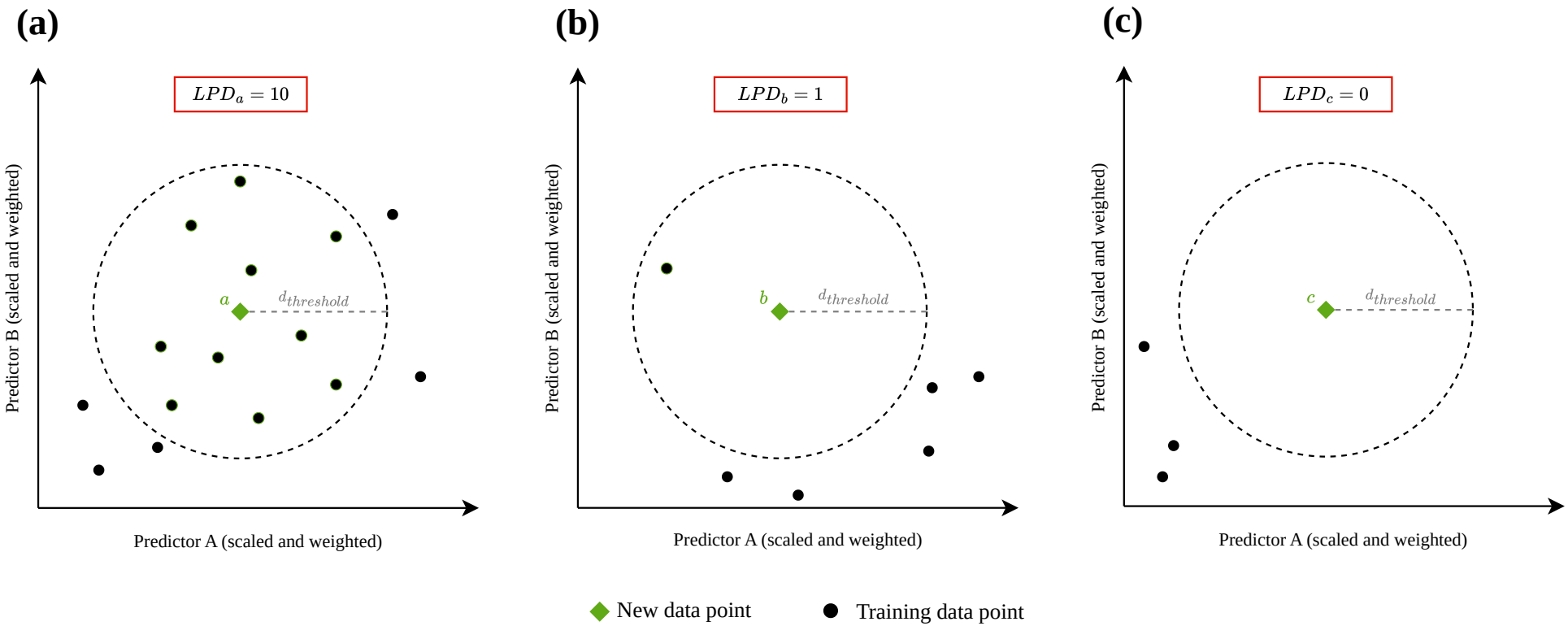




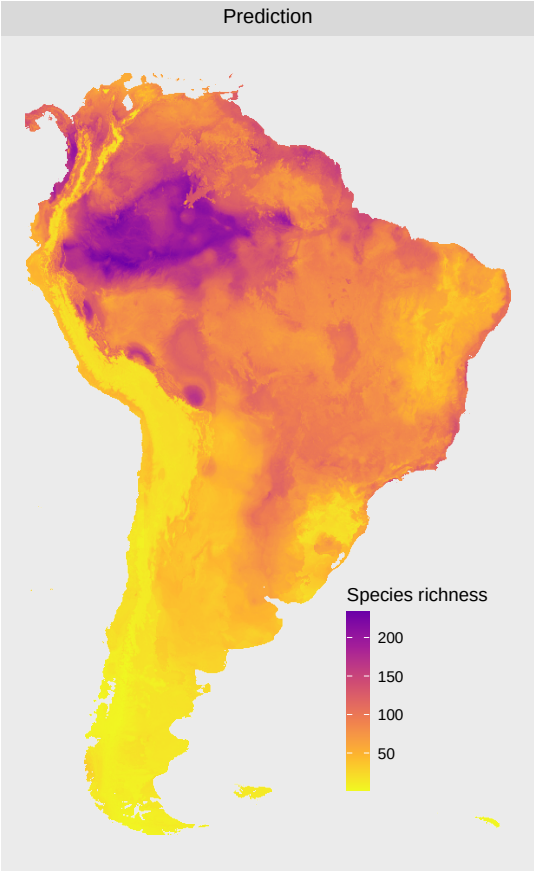
Prediction



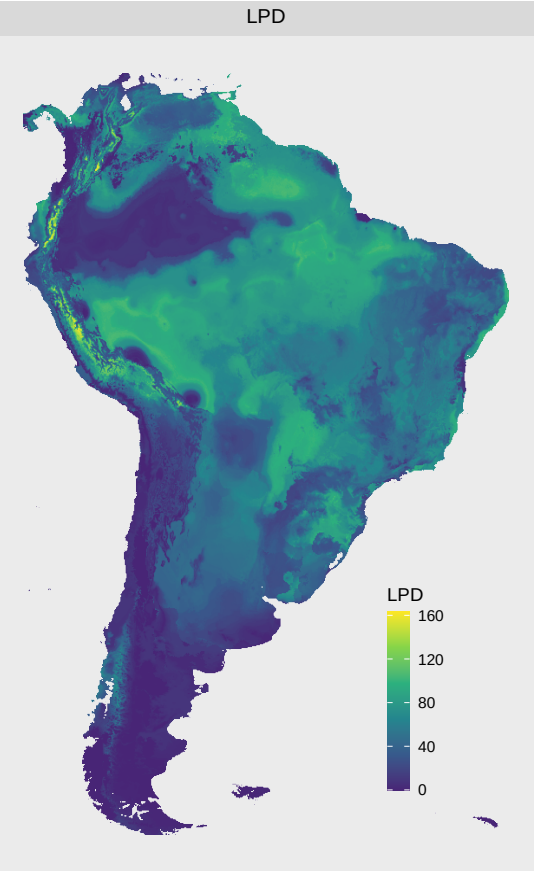
Source: Own representation.


(a)
$LPD_a = 10$
a
$d_{threshold}$
Predictor B (scaled and weighted)
Predictor A (scaled and weighted)
(b)
$LPD_b = 1$
b
$d_{threshold}$
Predictor B (scaled and weighted)
Predictor A (scaled and weighted)
(c)
$LPD_c = 0$
c
$d_{threshold}$
Predictor B (scaled and weighted)
Predictor A (scaled and weighted)
New data point
Training data point

Prediction



Local data point density



Universität Münster

Methods

Area of Applicability (AOA)[1]

Universität Münster

- The **area of applicability (AOA)** builds the baseline for the conception of the **local data point density (LPD)** approach
- Methodology in the context of spatial prediction models by Meyer and Pebesma (2021)
- **Goal:** Outline the regions in a target area where a model's predictions can be considered reliable according to the cross-validation performance of the model

➔ Inside the AOA the cross-validation performance of the model holds on average

- **How it works:** The authors calculate a dissimilarity index (DI) that measures how different a new location is from the training data points in the multidimensional predictor data space

➔ The AOA is derived by applying a threshold on this DI based on the models cross-validation performance

Area of Applicability (AOA) - Example

Universität Münster

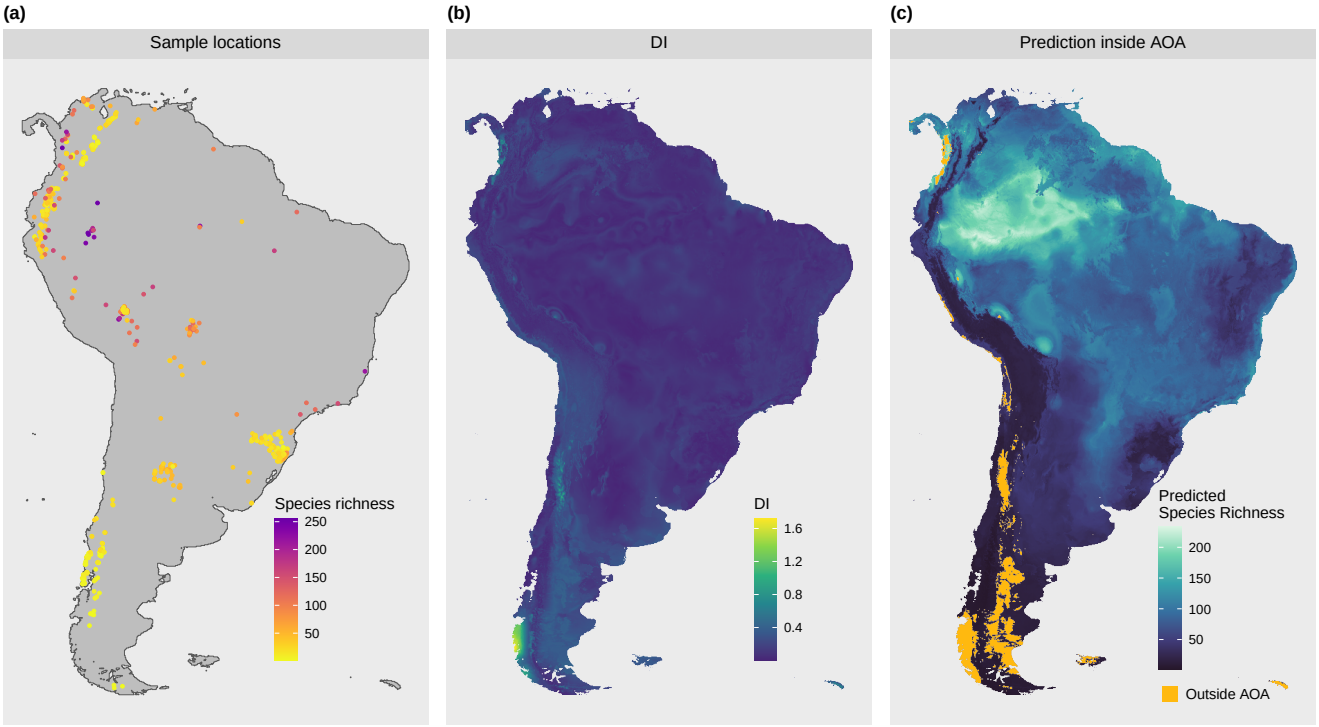


Figure 1: (a) Training samples that were incorporated in the machine learning model. (b) DI of the prediction locations in the target area. (c) `Species richness` predicted for South America based on a machine-learning model. Orange regions are outside the AOA.

(Source: Own representation)

Why LPD?

Universität Münster

- Dissimilarity index (DI) of a new prediction data point only calculated based on the nearest-neighbor in the training data

  → Local training data point densities not considered

  Problem:

  Inside the AOA is not possible to discriminate between areas in the predictor space where few, or even one isolated training data point is nearest and a predictor space location that is densely covered by training data points (see Figure 1)

  Assumption:

  We assume that local training data point densities can be highly decisive for the uncertainty of predictions

Why LPD?

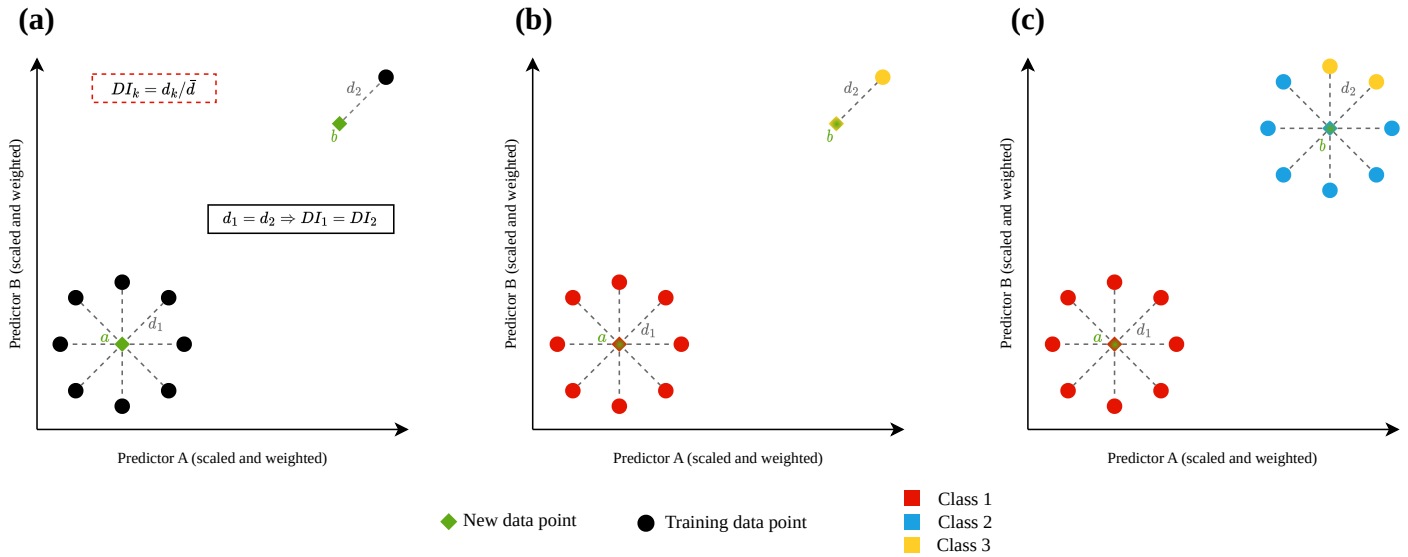


Figure 2: Hypothetical 2-dimensional scaled and weighted predictor data space with training data and new prediction locations to show limitations of the AOA using a classification example. It is assumed, that all new prediction locations fall within the AOA, i.e. their DI value is smaller than the threshold calculated from cross-validated training data.

(Source: Own Representation)

Our method - LPD

Universität Münster

- Newly developed **local data point density approach (LPD)** based on the given concepts of the **AOA** method
  - ➔ Allows for a better assessment of the area of applicability of a model and the training data coverage of different regions in the target area
- Quantitative measure for a new prediction data point that indicates how many similar training data points (in terms of predictor values) have been included in the model training (see Fig. 2)
  - ➔ A training data point is considered similar if it defines a new data point as being within the AOA, i.e. the model is considered applicable for the corresponding prediction location
- Using the relationship between the LPD and the models cross-validation performance to express prediction uncertainty

Can be seen as a first approach towards including local data point densities in the validation of spatial mapping workflows

Our method - LPD

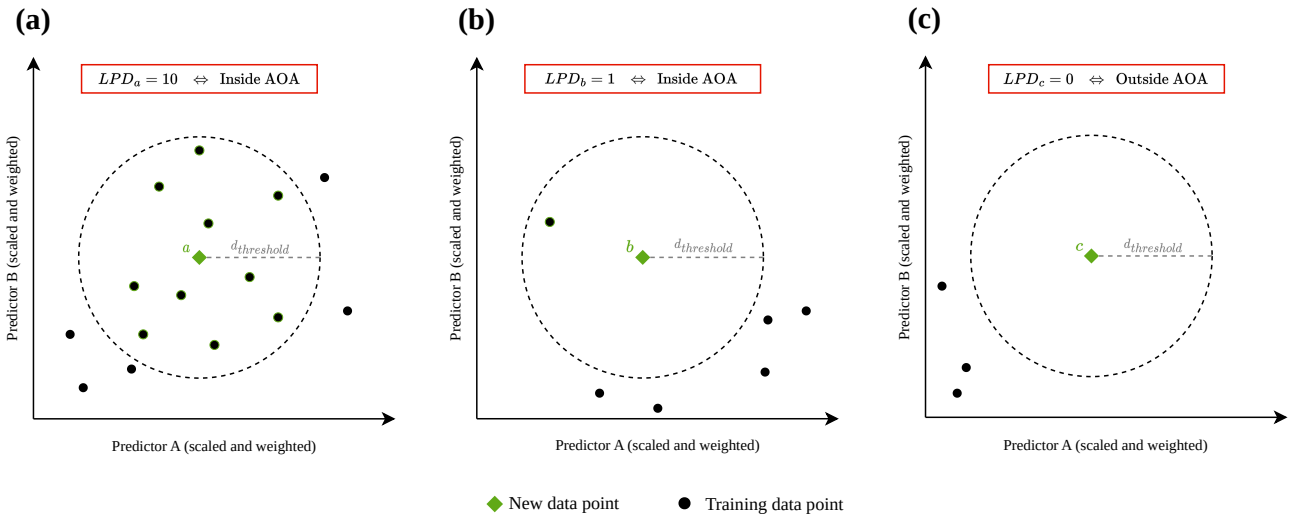
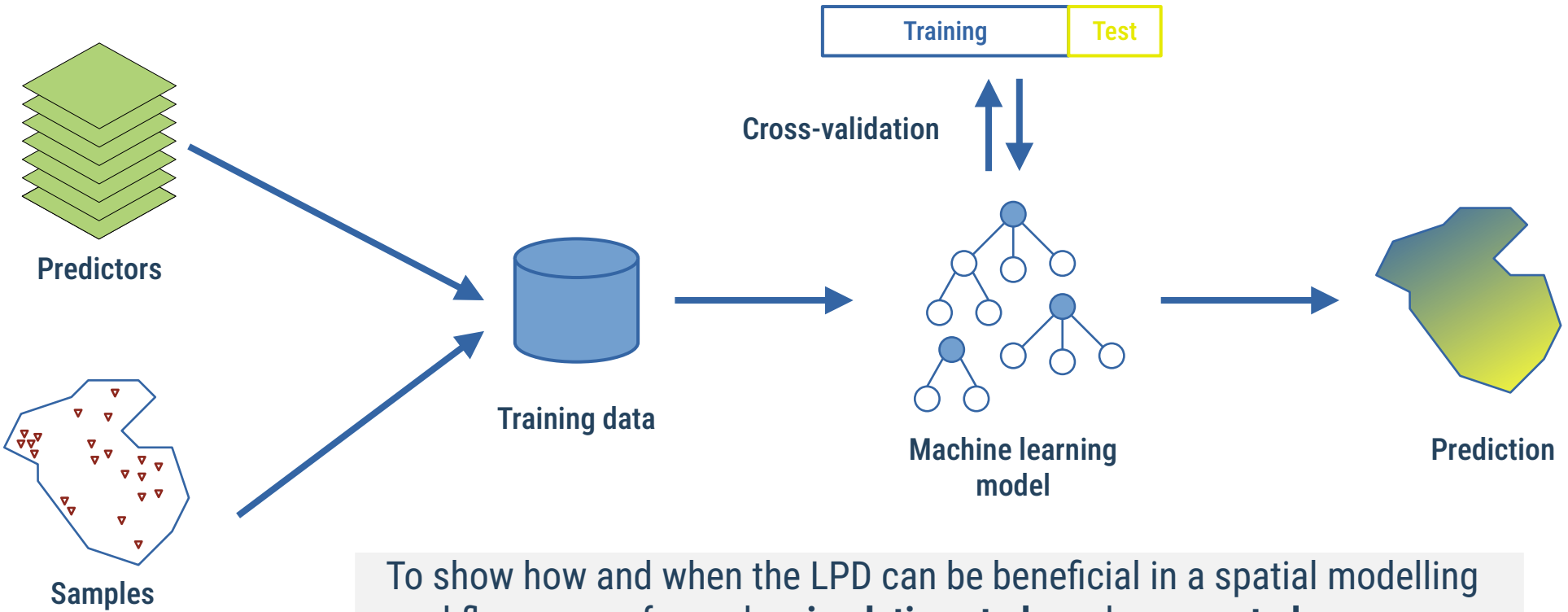


Figure 3: The DI is calculated by dividing a distance in the multidimensional predictor data space through the average mean distance between the cross-validated training data. Since the AOA threshold is defined as the outlier-removed maximum DI of the training data, we can transform it back into a respective distance in the predictor space. (a), (b), and (c) clarify when a training data point is included in the LPD.

(Source: Own representation)

How did we test the method?

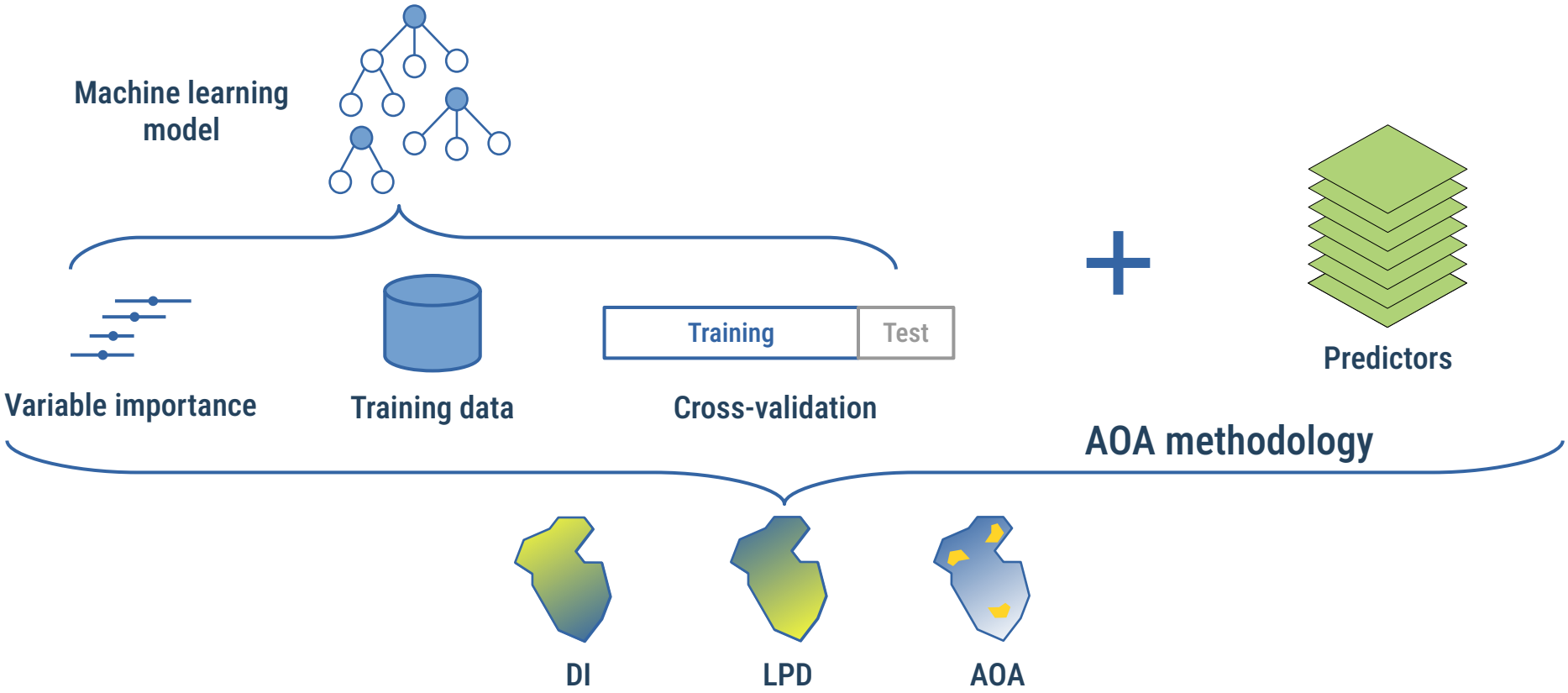
Universität Münster



To show how and when the LPD can be beneficial in a spatial modelling workflow, we performed a **simulation study** and a **case study**

How did we test the method?

Universität Münster



Simulation study

Universität Münster

3 prediction tasks were simulated with a **known** response variable

Predictors: 19 bioclimatic variables from the Worldclim[2] dataset

Response: Simulated via principle component analysis (PCA) with the variables bio2, bio5, bio10, bio13, bio14, and bio19

Sampling: 3 sampling scenarios (random, clustered, biased with outlier) with 100 samples

Models: Random Forest (RF) model for each sampling scenario

Validation: Cross-validation (CV) designed in line with the sampling distributions (Random CV, Spatial-Leave-One-Cluster-Out CV, and kNNDM[3] CV)

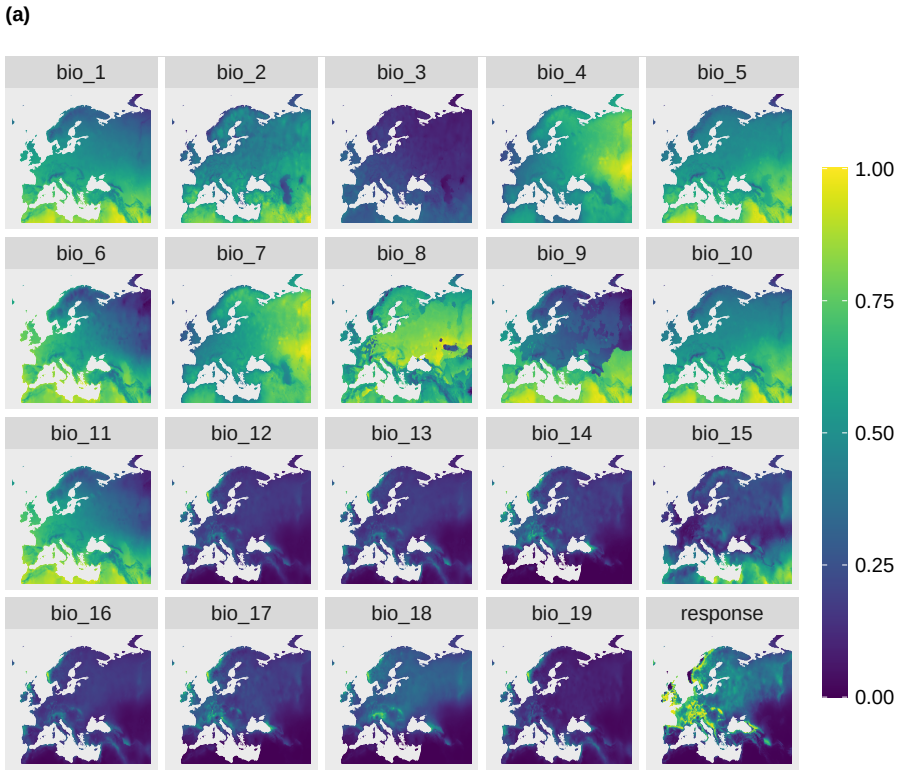
→ make predictions and calculate DI, LPD and AOA

Analytics:

- LPD ~ DI
- LPD ~ True absolute error/RMSE
- LPD ~ Standard deviations of the random forest ensemble
- DI/LPD from CV ~ CV performance (RMSE)

Simulation study

Universität Münster



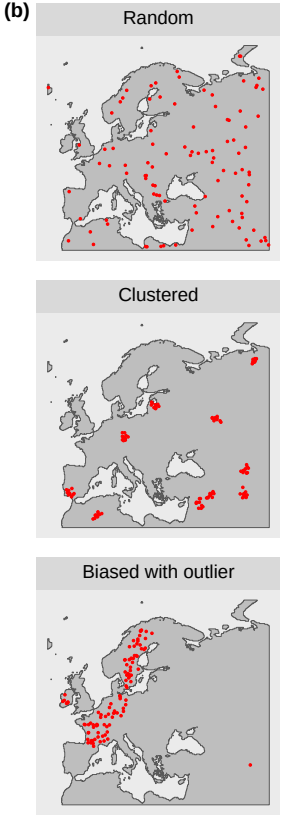
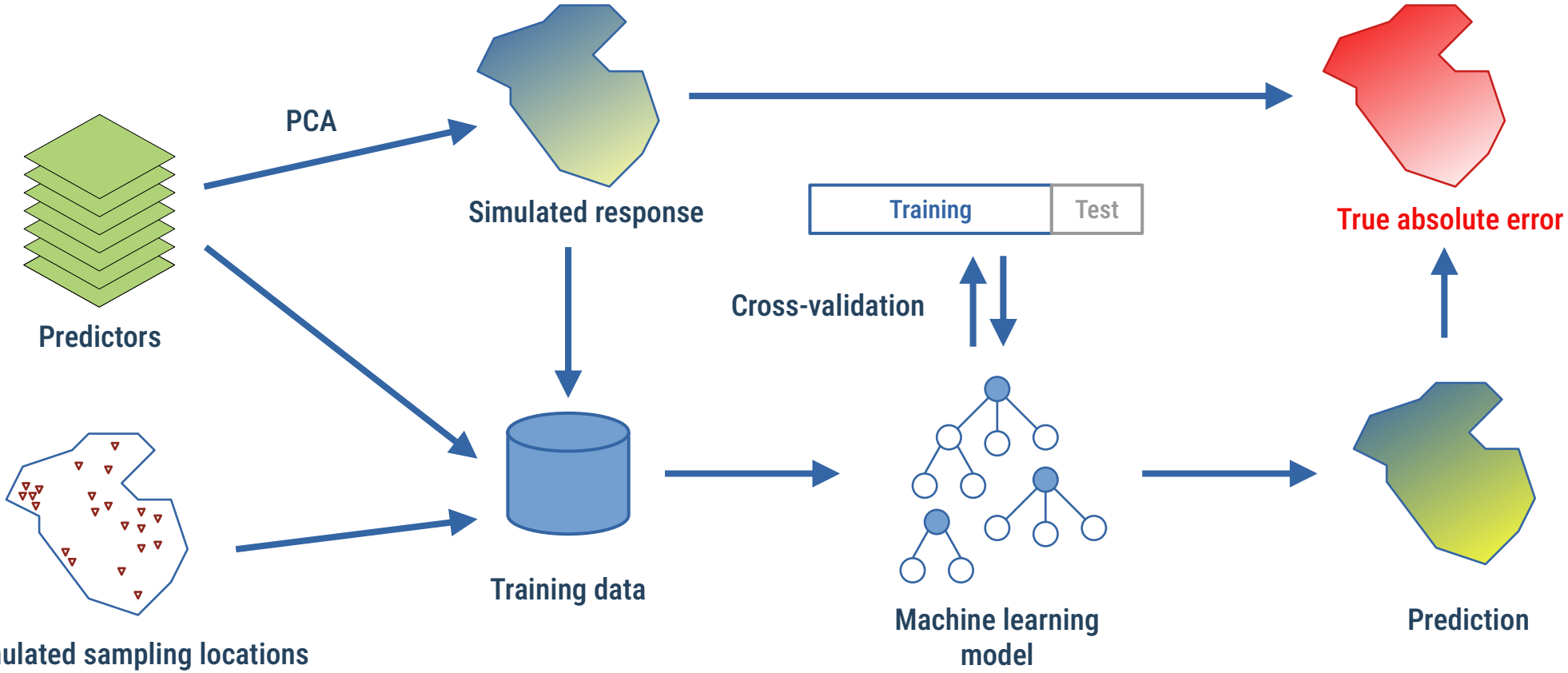


Figure 4: (a) The 19 predictor variables from the Worldclim dataset and the response variable used in the simulation study. All grids were cropped to the area of Europe and stretched from 0 to 1 for better visualization. (b) The three different sampling designs used in the simulation study (random, clustered, biased with outlier).

(Source: Own representation)

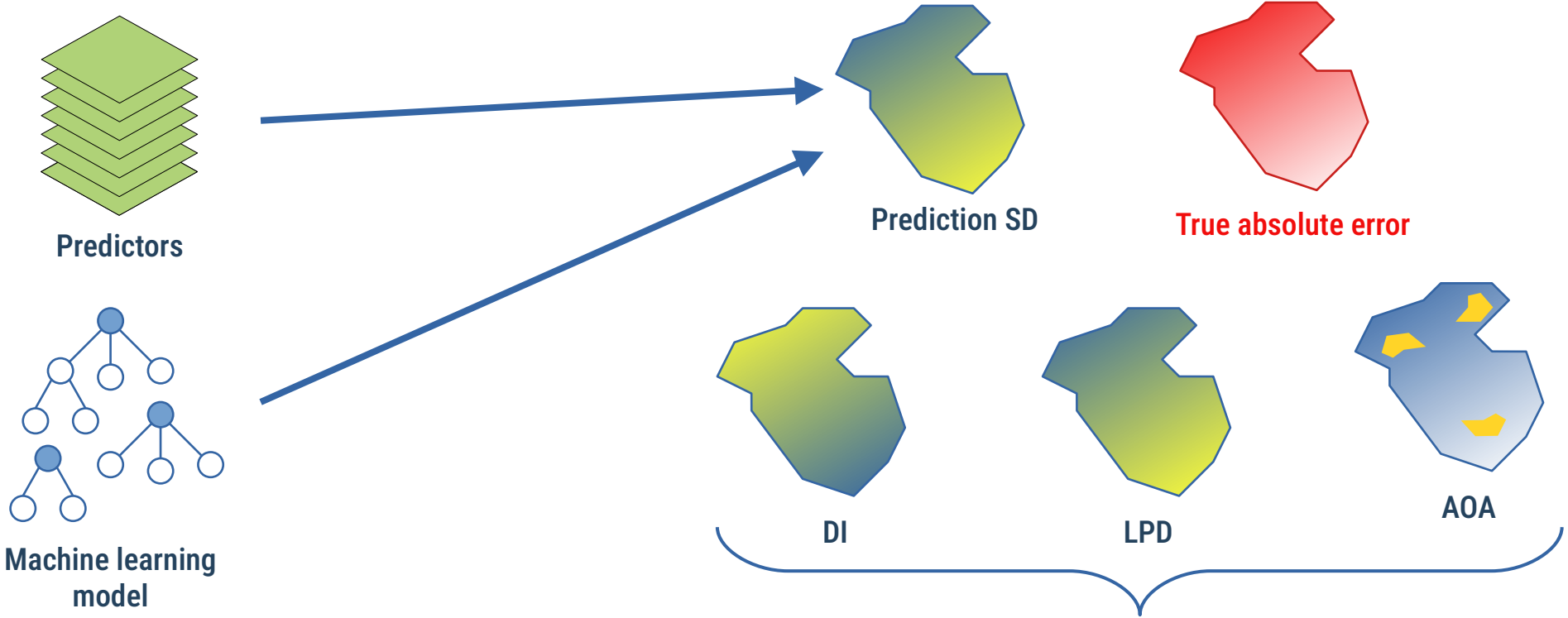
Simulation study - workflow

Universität Münster



Simulation study – workflow

Universität Münster



Case study

Universität Münster

Predicting ‘species richness’ for the whole of South America

Predictors: 4 bioclimatic variables from the Worldclim[2] dataset and elevation selected by forward feature selection (FFS)

Response: Only available for sampling locations – not the for the whole target area

Sampling: 703 sampling locations derived from the sPlotOpen[5] dataset

Models: Random Forest (RF) model with forward feature selection (FFS)[4]

Validation: kNNDM[3] CV

→ make predictions and calculate DI, LPD and AOA

Analytics:

- Visual and value-based assessment of the LPD
- LPD ~ Standard deviations of the random forest ensemble
- DI/LPD from CV ~ CV performance (RMSE) relationship
- Performance prediction from the relationship

Case study

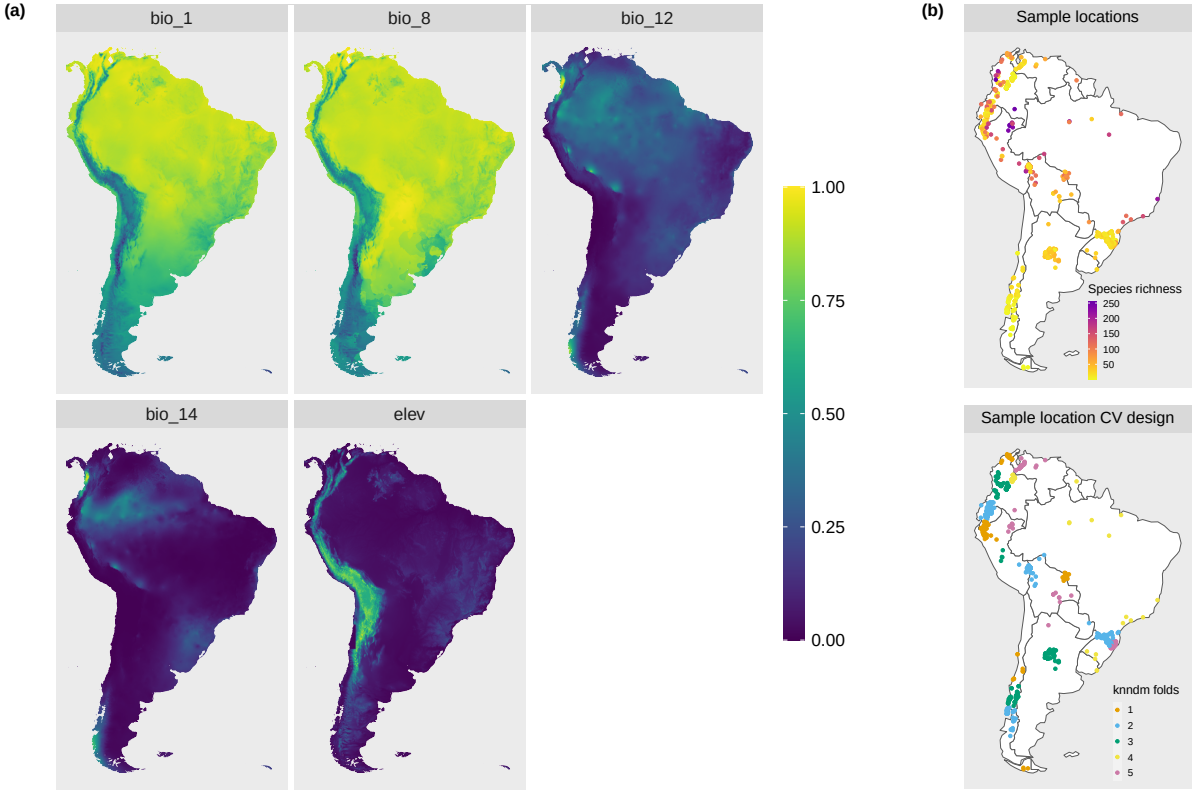
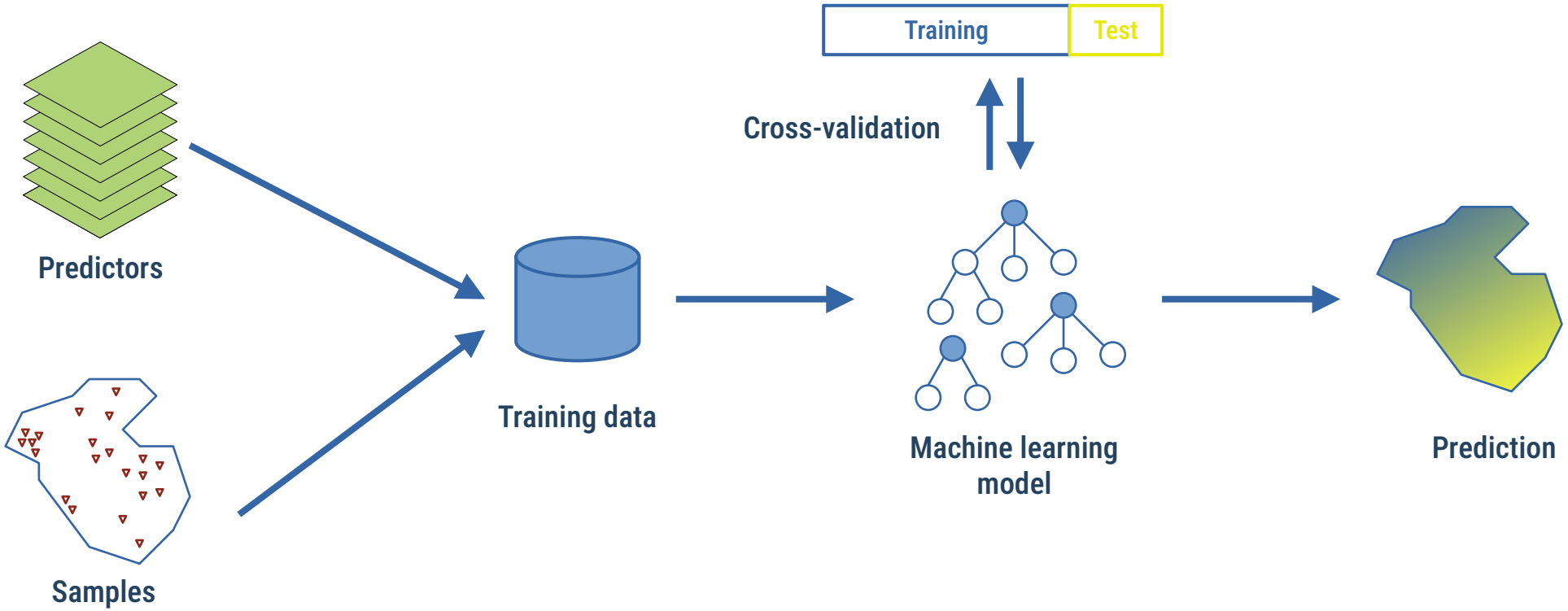


Figure 5: (a) The predictor variables chosen with by FFS during the model training. (b) The training locations visualized with the 'species richness' response values (top) and the computed knndm folds (bottom).

(Source: Own representation)

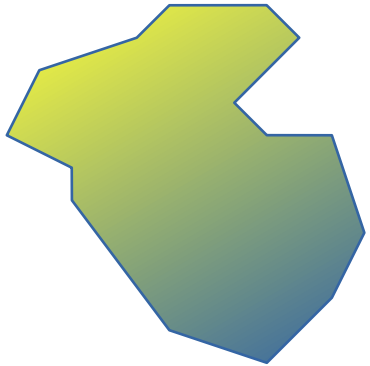
Case study - workflow

Universität Münster

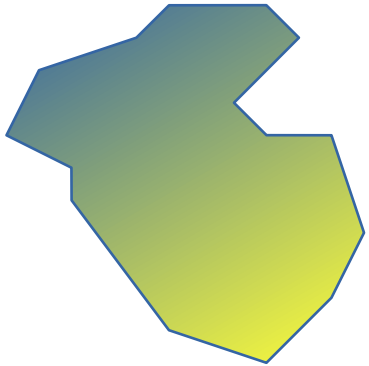


Case study - workflow

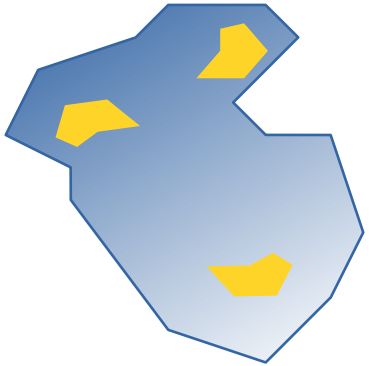
Universität Münster



DI



LPD



AOA

Visual and value-based comparison and interpretation

Universität Münster

Results

Simulation study - results

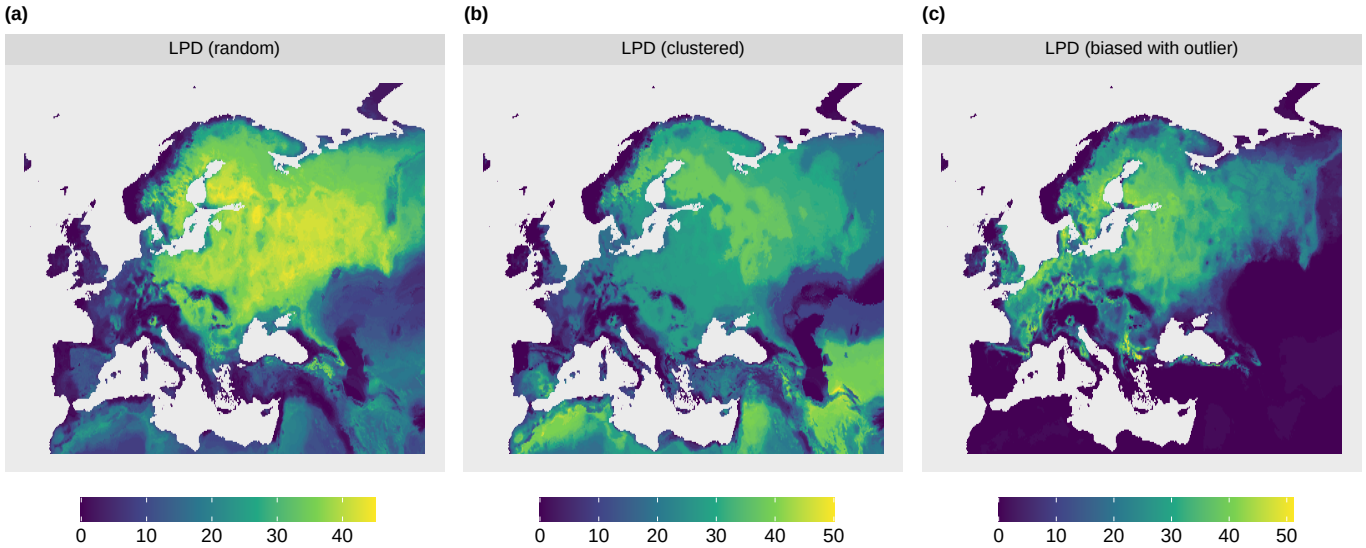


Figure 6: Local data point density (LPD) for the (a) random, (b) clustered, and (c) biased with outlier sampling design.

(Source: Own representation)

Simulation study - results

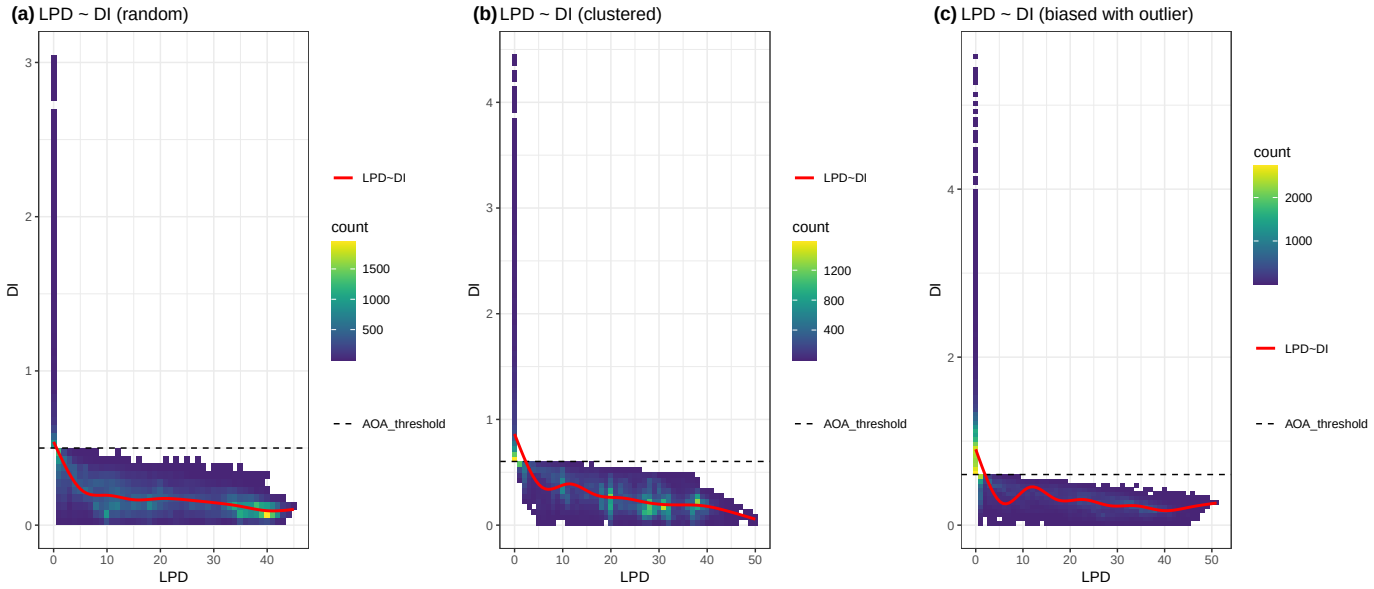


Figure 7: The calculated LPD values of the prediction locations for each scenario plotted in data bins against their respective DI values; A generalized additive model is fitted to the data and shown as a red line for better visualization of the relationship. (a) is referred to the random, (b) to clustered, and (c) to biased with outlier sampling design.

(Source: Own representation)

Simulation study - results

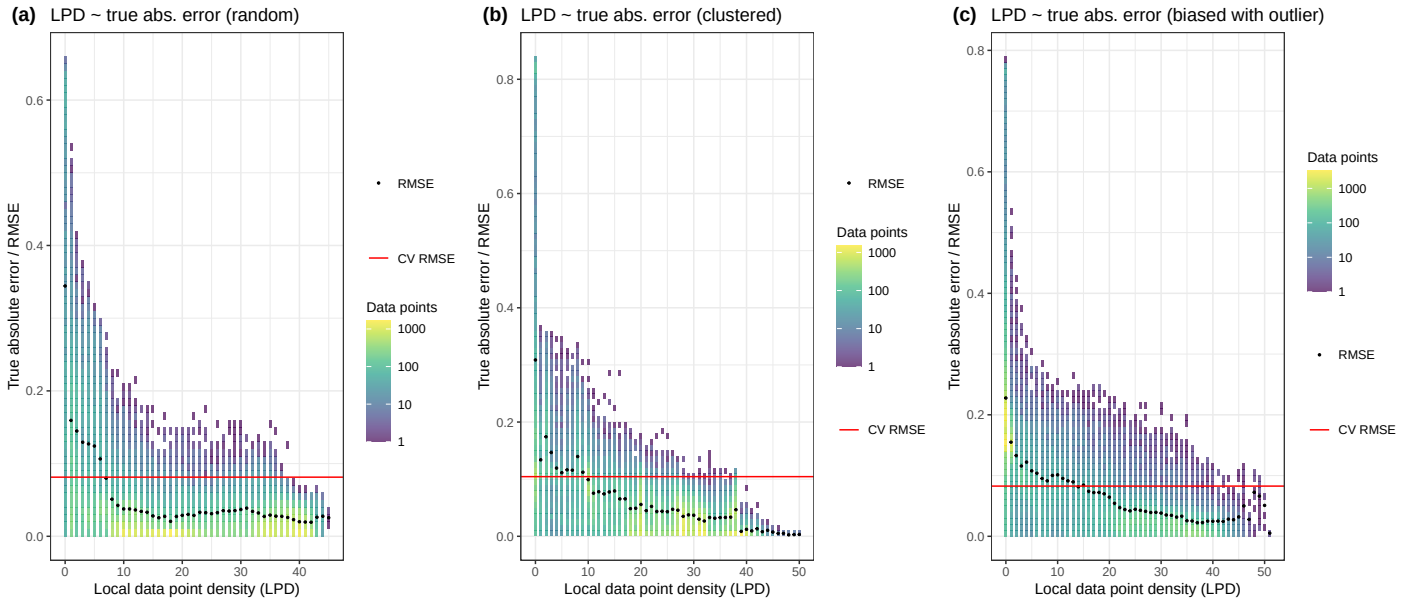


Figure 8: LPD values of the prediction location for each scenario plotted in data bins against the respective true absolute error values; The red line signifies the RMSE obtained from CV in each scenario and the black points show the RMSE for all prediction locations with the specific LPD value (see Table A1). (a) is referred to the random, (b) to clustered, and (c) to biased with outlier sampling design.

(Source: Own representation)

Simulation study - results

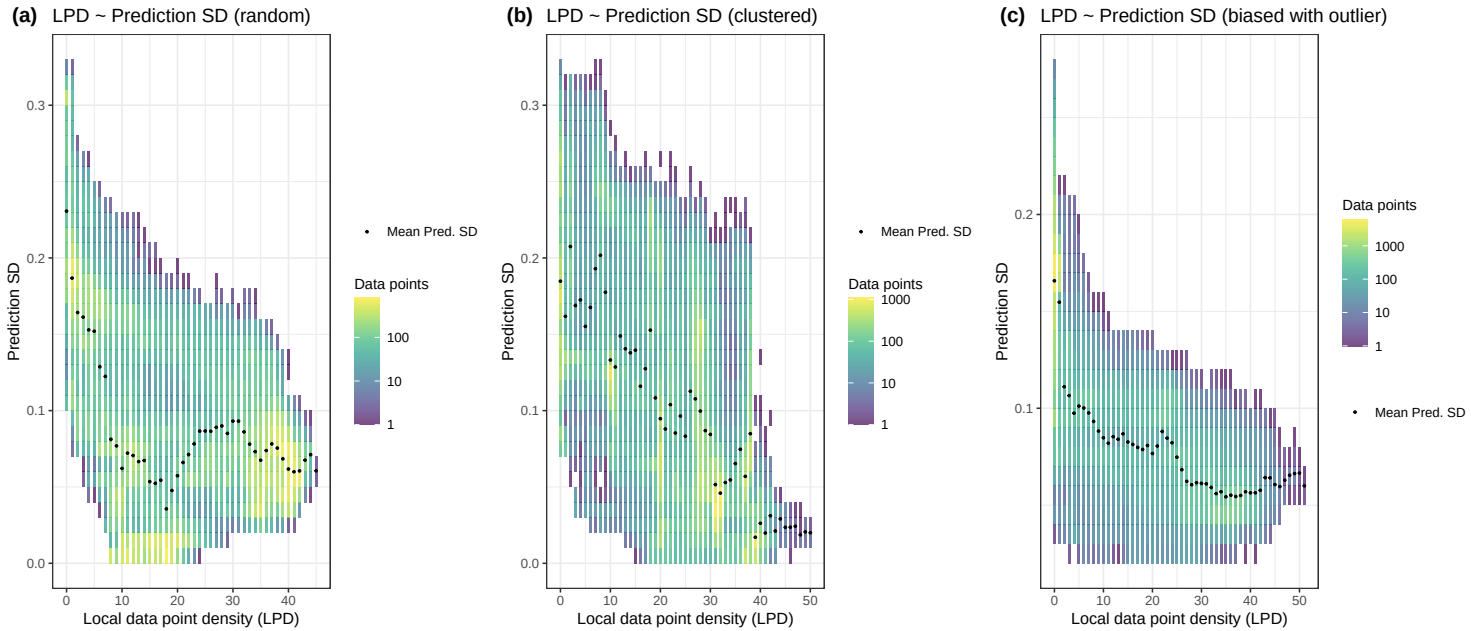


Figure 9: LPD values of the prediction locations for each scenario plotted in data bins against the respective prediction standard deviations for the 500 trees in the RF model; The black points show the mean prediction standard deviation for all locations with a specific LPD value (see Table A2). (a) is referred to the random, (b) to clustered, and (c) to biased with outlier sampling design.

(Source: Own representation)

Simulation study - results

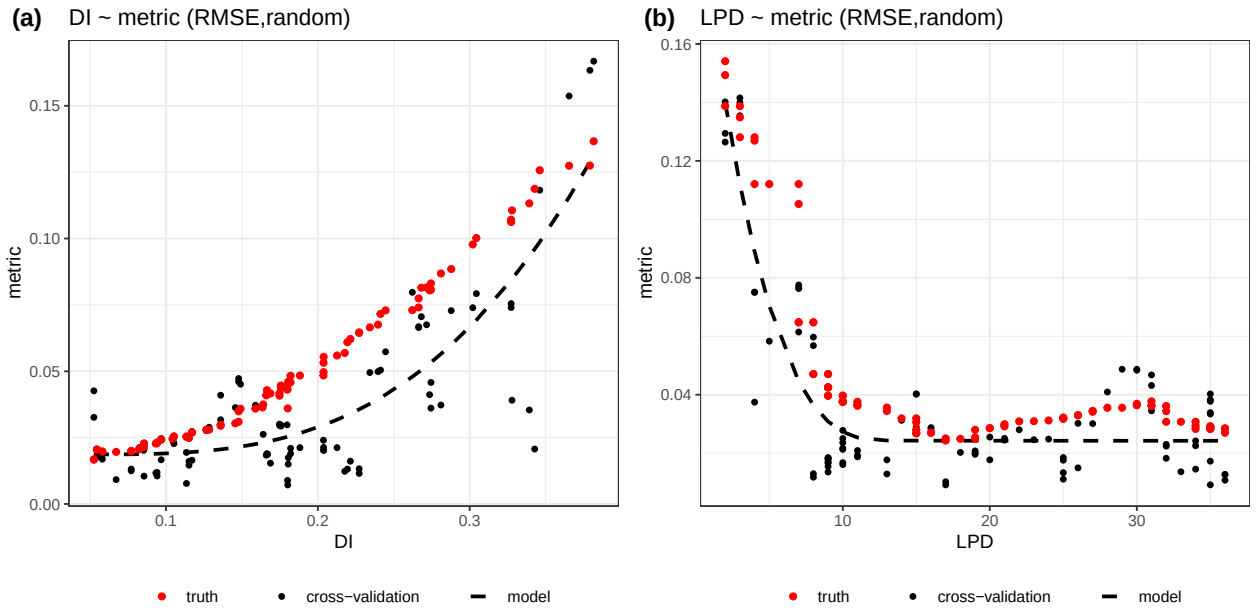


Figure 10: (a) Relationship between the error metric (RMSE) and the dissimilarity index (DI), (b) relationship between the error metric (RMSE) and the local data point density (LPD) all for the random sampling scenario; Each single data point corresponds to the RMSE from a sliding window of size 5, either along the DI axis or along the LPD axis. The 2- dimensional models are shown as a blue line. The true RMSE which was calculated using the reference map and corresponding predictions within the identical windows of DI and LPD values is shown in red.

(Source: Own representation)

Simulation study - summary

Universität Münster

- For low LPD values the cross-validation performance of the model still tends to be exceeded, whereas for high LPD values the cross-validation performance is highly under cut (see Figure
  - → **Assumption:** Local data point densities have an impact on the models performance to a certain degree
- The standard deviations of the random forest ensemble decrease on average with increasing LPD values
  - → **Assumption:** Prediction uncertainty decreases with increasing local training data point densities
- There seems to be a relationship between the LPD and the models performance
  - → **Assumption:** We can use the LPD to carry out model performance estimation

Case study – results

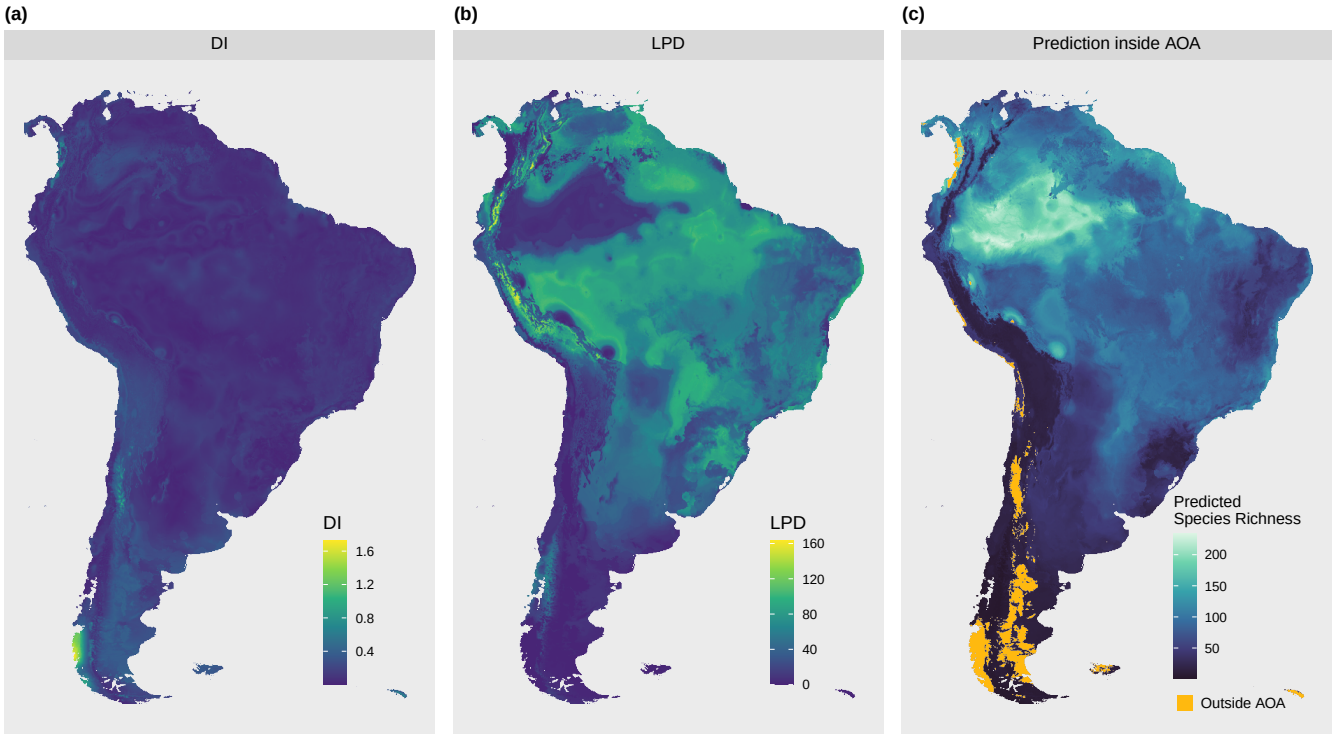


Figure 11: Comparison between the (a) DI, (b) LPD and (c) prediction inside the AOA. Areas outside the AOA are shown in orange

(Source: Own representation)

Case study – results

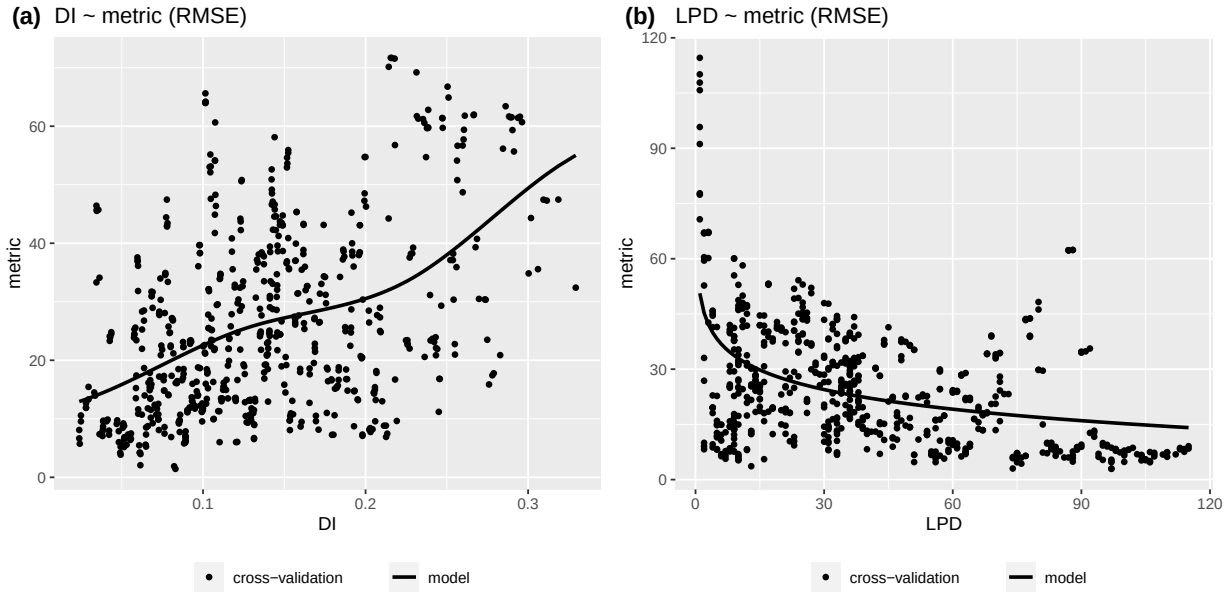


Figure 12: (a) Relationship between the error metric (RMSE) and the dissimilarity index (DI), (b) relationship between the error metric (RMSE) and the local data point density (LPD) for the random sampling scenario; Each single data point corresponds to the RMSE from a sliding window of size 5, either along the DI axis, along the LPD axis, or both axes (in this case mean RMSE is used). The 2-dimensional models are shown as a red line.

(Source: Own representation)

Case study – results

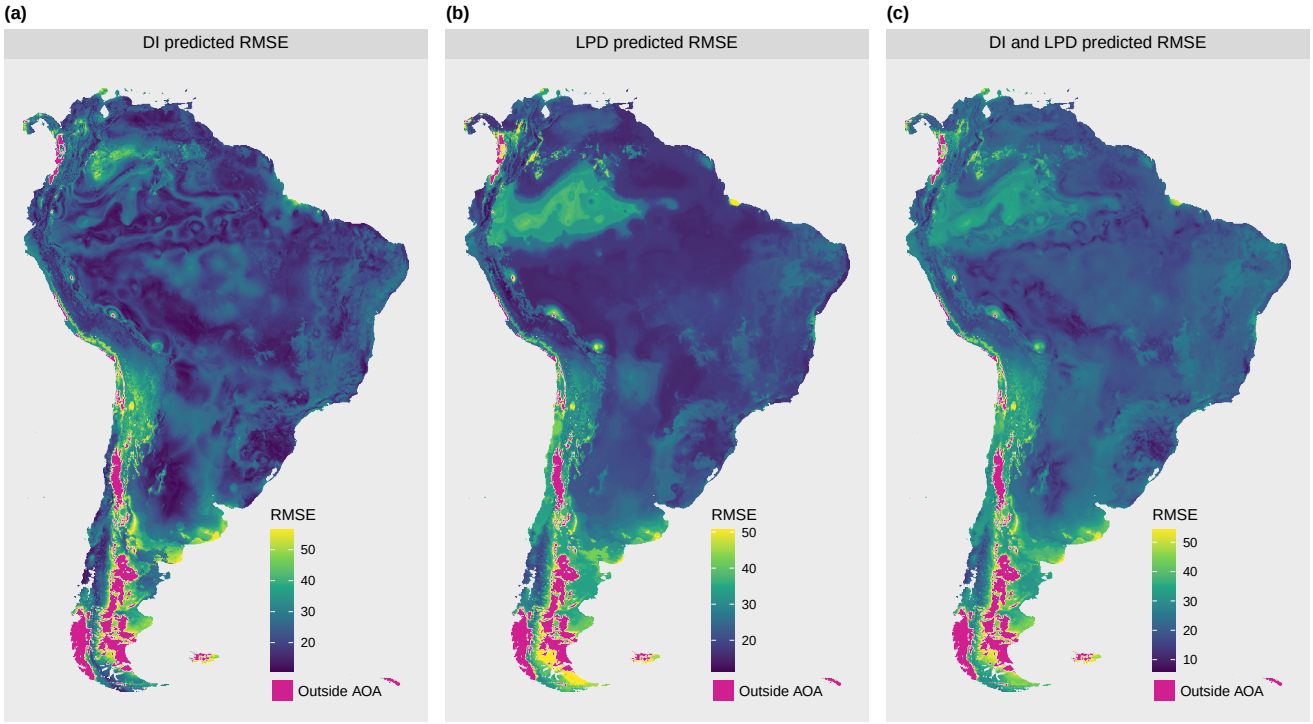


Figure 13: Comparison of the model performance (RMSE) predictions inside the AOA by the (a) DI, (b) LPD, and (c) DI + LPD error model.

(Source: Own representation)

Case study - summary

Universität Münster

- Low local training data point densities (low coverage by the training data) are not reflected with the dissimilarity index (DI)
- A relationship between LPD and model performance is also noticeable for ‘real-world’ data which enables area-wide performance estimation

Universität Münster

Discussion

Discussion

- Low training data point densities do not necessarily lead to poor prediction results , but merely increase the probability of a poorer result (higher uncertainty)
  - ➔ The presence of a single or a few data similar data points can be sufficient
- DI is dependent on the cross-validation strategy of the model, hence also the LPD
  - ➔ Cross-validation needs to be designed in line with the sampling distribution
  - ➔ Relationship between LPD and model performance can be determined
- Including LPD for the delineation of the AOA or using it for outlier-caused AOA detection?

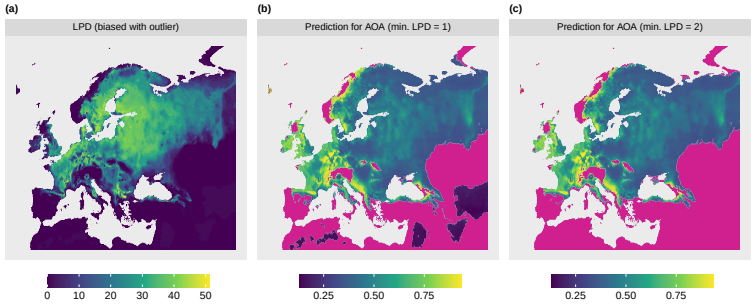


Figure 14: Detection of outlier-caused AOA on the example of the biased with outlier sampling scenario; (a) shows the LPD with the sampling locations, (b) shows the regular AOA, and (c) shows an LPD-dependent AOA with a minimum LPD of 2 as an additional AOA condition.

Discussion

Universität Münster

- Deriving an combined uncertainty measure from the DI and LPD
  - ➔ Needs further investigation to see if a combined uncertainty measure from DI and LPD values is meaningful to estimate model performance
- Computation times:
  - ➔ For large training and prediction data sets the LPD has high computation times
  - ➔ Limiting the LPD to user defined maximum of neighbours to be considered can help not having to forego density information, but disables the estimation of the model performance

Universität Münster

Conclusion

Conclusion

Universität Münster

- The LPD is a quantitative density measure that can provide important additional information, where in the target area the training data coverage is too low or significantly lower than in surrounding areas
- The LPD can be used to provide estimates of model performance in the target area
- We suggest communicating the LPD in a spatial prediction mapping workflow and alongside the DI and AOA

→ **Can be a further gainful factor in the critical assessment of overly optimistic data-driven prediction maps**

Universität Münster

How to use?

How to use the LPD?

Universität Münster

- Methods are implemented inside the `aoa` function in the `CAST` R package on CRAN

```
AOA = aoa(newdata = newdata, model = model, LPD = TRUE)
```

- A method for visualizing and exploring the DI, LPD and AOA interactively is implemented in the `CASTvis` R package on GitHub

```
exploreAOA(aoa = AOA)
```

- Further information can be found under:
  - https://cran.r-project.org/web/packages/CAST/index.html
  - https://github.com/HannaMeyer/CAST
  - https://hannameyer.github.io/CAST/articles/cast04-AOA-tutorial.html
  - https://github.com/fab-scm/CASTvis

Universität Münster

References

References

1 Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution, 12 (9), 1620–1633. https://doi.org/10.1111/2041-210X.13650

2 Fick, S. E., & Hijmans, R. J. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 37 (12), 4302–4315. https://doi.org/10.1002/joc.5086

3 Linnenbrink, J., Milà, C., Ludwig, M., & Meyer, H. (2023). kNNDM: k-fold Nearest Neighbour Distance Matching CrossValidation for map accuracy estimation. EGUsphere, 1–16. https://doi.org/10.5194/egusphere-2023-1308

4 Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software, 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

5 Sabatini, F. M., Lenoir, J., Hattab, T., Arnst, E. A., Chytrý, M., Dengler, J., De Ruffray, P., Hennekens, S. M., Jandt, U., Jansen, F., Jiménez-Alfaro, B., Kattge, J., Levesley, A., Pillar, V. D., Purschke, O., Sandel, B., Sultana, F., Aavik, T., Aćić, S., . . . Bruelheide, H. (2021). sPlotOpen – an environmentally balanced, open-access, global dataset of vegetation plots. Global Ecology and Biogeography, 30 (9), 1740–1764. https://doi.org/10.1111/geb.13346