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

Hanna Meyer (1), Edzer Pebesma (2)

(1) Westfälische Wilhelms-Universität Münster, Institute of Landscape Ecology, Münster, Germany
(2) Westfälische Wilhelms-Universität Münster, Institute for Geoinformatics, Münster, Germany

Introduction

Predictive modelling using machine learning has become very popular for spatial mapping of the environment. Models are often applied to make predictions far beyond sampling locations where new geographic locations might considerably differ from the training data in their environmental properties. However, areas in the predictor space without support of training data are problematic. Since the model has no knowledge about these environments, predictions have to be considered highly uncertain. **Estimating the area to which a prediction model can be reliably applied is required.**

Problem & Definition of the AOA

Problem of predicting beyond training data with machine learning (e.g. Random Forests): predictions become uncertain when moving away from training data (Fig. 1b). In contrast to simple models (Fig. 1a) this is not reflected by the prediction interval width outside the data range.

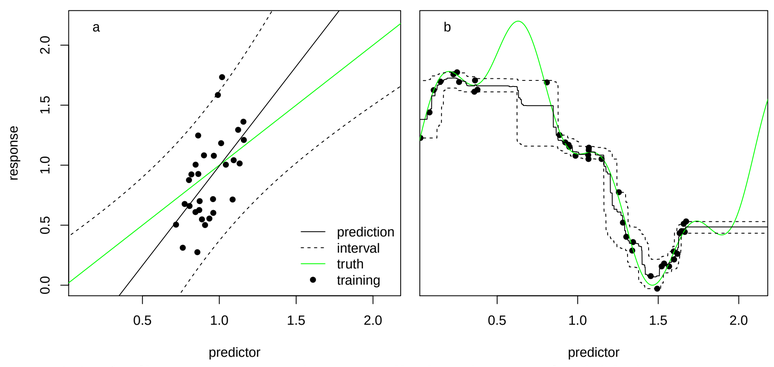


Figure 1 Left: linear regression prediction interval width increases with distance from the center of the training data, right: a more complex relationship fitted with Random Forest.

We suggest a methodology that delineates the "area of applicability" of a prediction model, that we define as the area, for which, in average, the cross-validation error of the model applies.

Suggested Methodology

1) Standardize predictor variables, and weight them by the variable importance scores of the trained model

2) Define a distance measure in predictor space (Euclidean)

3) Compute dissimilarity index (DI) for prediction locations as the distance to the nearest training data point divided by the average distance between all training points

4) Define the area of applicability (AOA) by thresholding DI; the threshold is taken as a quantile of all DI values of the training data, with distances calculated between points that do not occur in the same cross-validation fold.

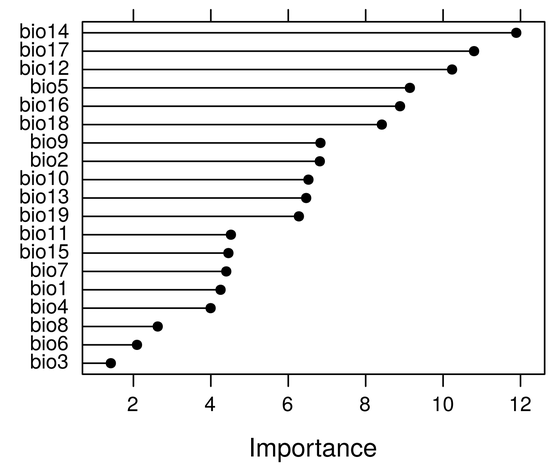


Figure 2 (top-right) : Example of a variable importance ranking from a trained machine learning model

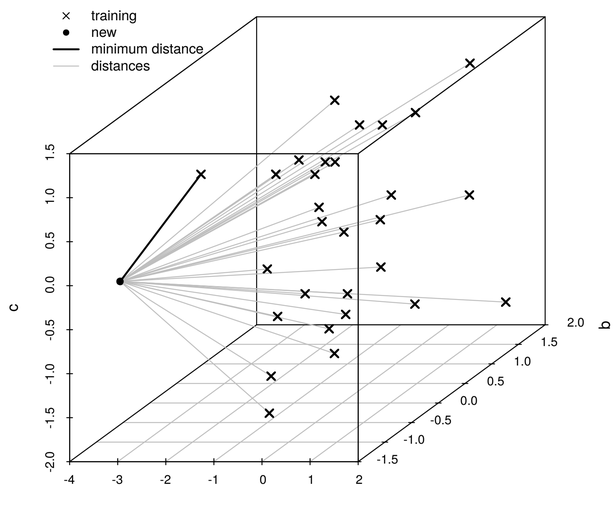


Figure 3 (bottom-right): Example of how the dissimilarity index (DI) for a new data point is calculated: The distance in the normalized and weighted predictor space to each training data point is calculated to identify the closest data point. The distance to the closest data point is then divided by the average of the mean distances between the training data.

Case Study: AOA of prediction model based on simulated data

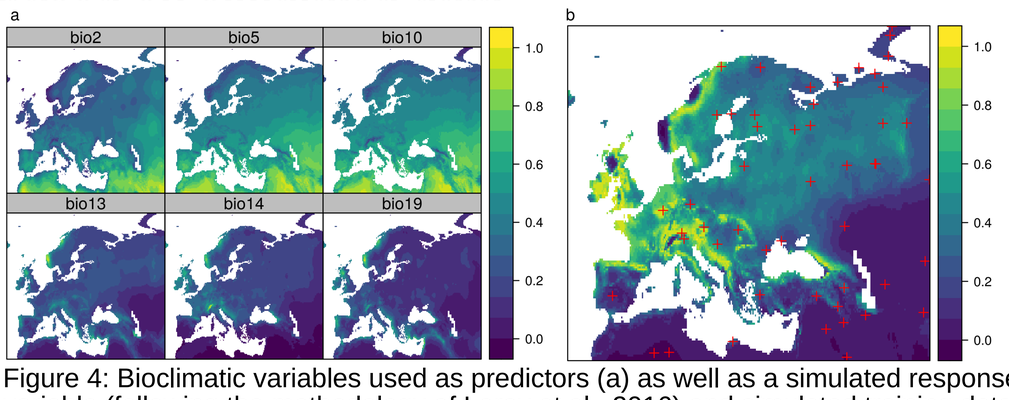


Figure 4: Bioclimatic variables used as predictors (a) as well as a simulated response variable (following the methodology of Leroy et al., 2016) and simulated training data (b).

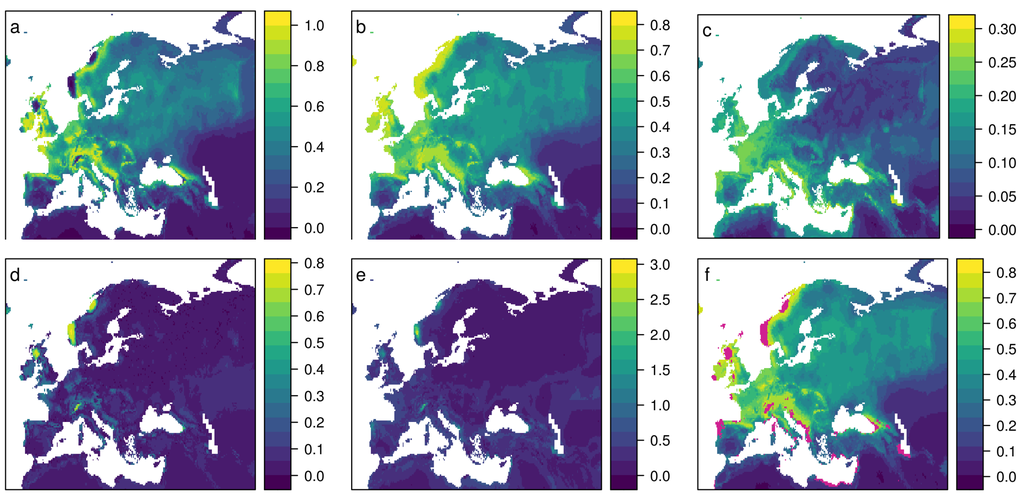


Figure 5: The simulated response (a), predictions made by Random forest (b), for comparison the standard deviations of predictions made by 500 trees (c), the true prediction error (d), the newly suggested dissimilarity index (e) and the predictions for the AOA, where areas outside the AOA are shown in pink (f)

Case Study II: Estimating the AOA for spatially clustered data

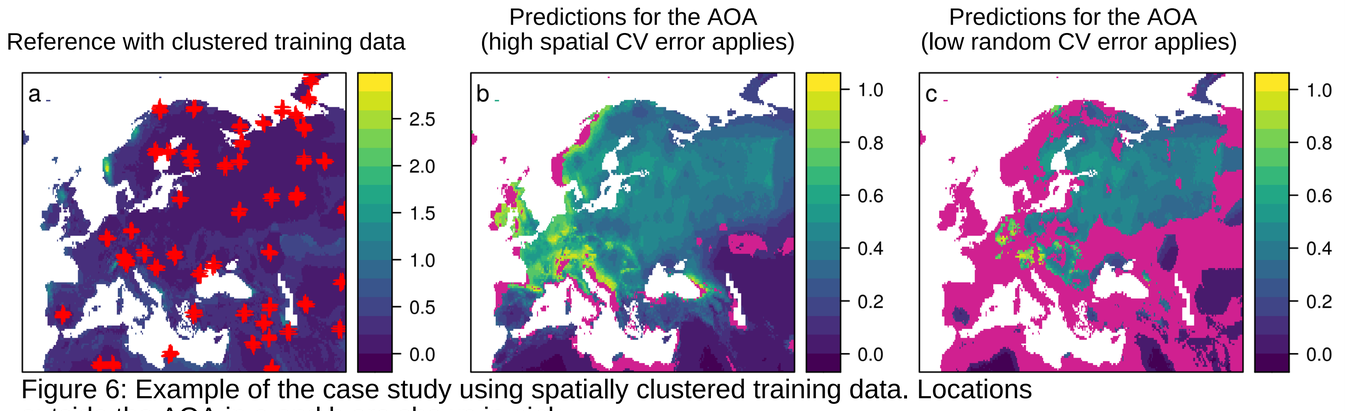
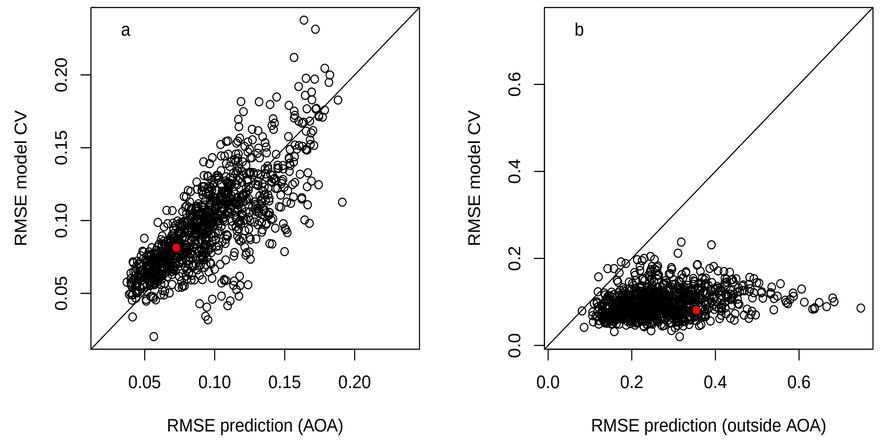


Figure 6: Example of the case study using spatially clustered training data. Locations outside the AOA in a and b are shown in pink

Approx. 1000 case study settings: Does the AOA really reflect the area for which the cross-validation error applies?



- High agreement between DI and prediction error
- Cross validation error applies within the AOA, but not outside
- Applicable across sampling designs (and appropriate cross-validation strategies)

Figure 7: Comparison between cross-validation RMSE and the prediction error within (a) and outside (b) of the AOA based on ~1000 simulations. Red dot represents the results for the presented case study.

Discussions & Conclusions

- The AOA is the area for which the model can make reliable predictions, where “reliable” means that in average, the estimated model error applies
- Knowledge on the AOA is important when predictions are used as a baseline for decision making or subsequent environmental modelling to avoid error propagation
- We suggest that the AOA should be provided alongside the predictions, and complementary to validation measures
- The method to estimate the AOA as presented here should be considered a first attempt and contains a number of aspects that are up for discussion

Further Information

- Methodology and tutorial on this method: Currently in the developer version of the CAST package: https://github.com/HannaMeyer/CAST
- Code to reproduce the case study: https://github.com/HannaMeyer/AOA_CaseStudy
- Paper describing the methodology: Meyer&Pebesma (to be submitted shortly): “Predicting into unknown space Estimating the area of applicability of spatial prediction models“

References

Leroy, B., Meynard, C.N., Bellard, C. & Courchamp,F. (2016): virtualspecies, an R package to generate virtual species distributions. Ecography 39(6), 599–607.