Method of AOA explained in figures

Supplementary to the paper 'Predicting into unknown space? Estimating the area of applicability of spatial prediction models'

Hanna Meyer

18/10/2020

Create Sample data, scale and weight them

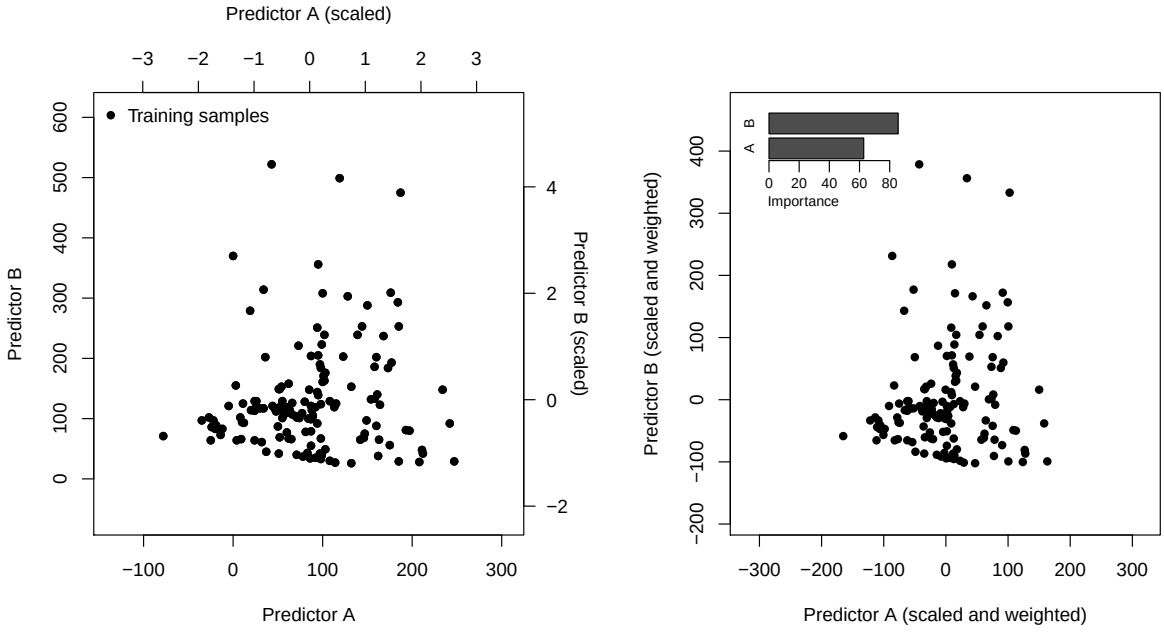


Figure 1: Initial situation: Training samples in a multidimensional (here 2) predictor space. First, the predictor space is scaled and weighted (second x and y axis) according to the estimated variable importance shown in the topright corner.

Calculate the DI within the training data to get the threshold for the AOA

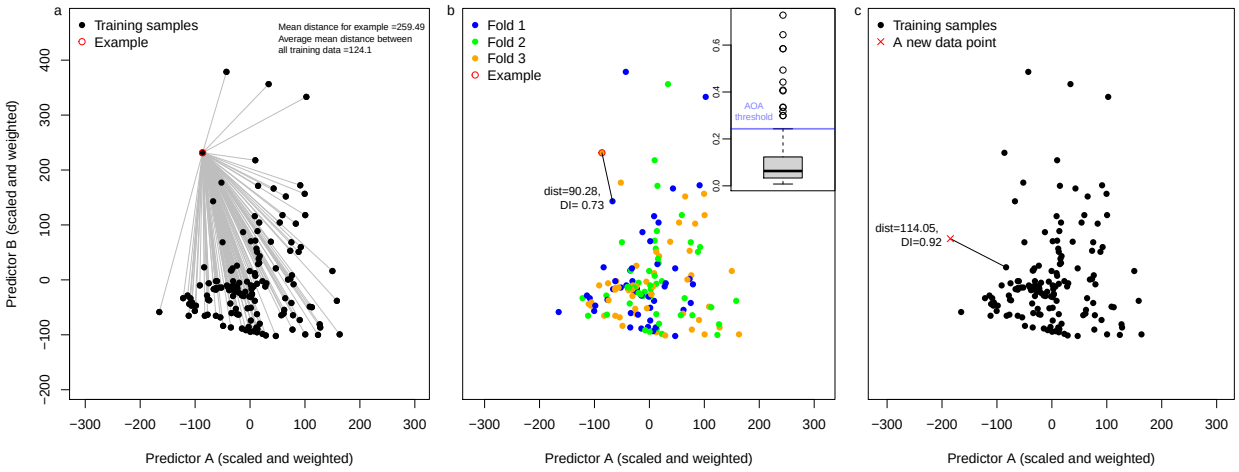


Figure 2: To answer the question if the trained model can be applied to a new data point, a threshold is used. This is the outlier-removed maximum DI of the cross-validated training data. The calculation of the DI of the training data is shown here for one example: The example data point is in fold 3 of the model. Therefore the distance to the nearest training data point NOT located in fold 3 is calculated (b) and this distance divided by the average of the mean distances between all training data (a). The DI is calculated for each training data point (boxplot) and the threshold is then derived from these DI values, so that a data point is outside the AOA if it is more dissimilar than the dissimilarity observed within the training data.

Estimate the AOA for each new potential data point

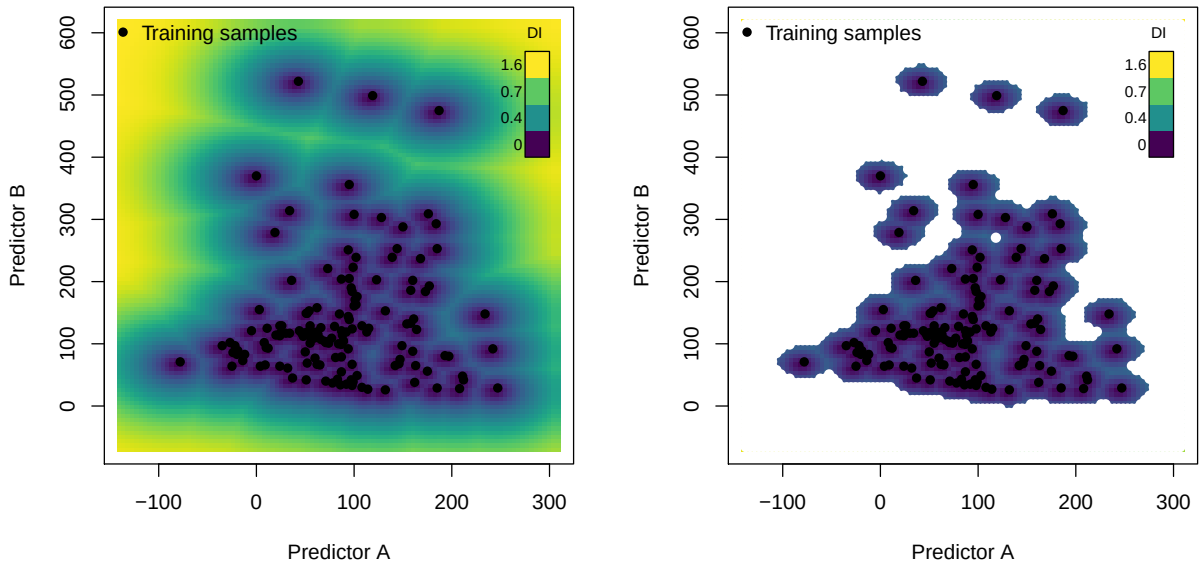
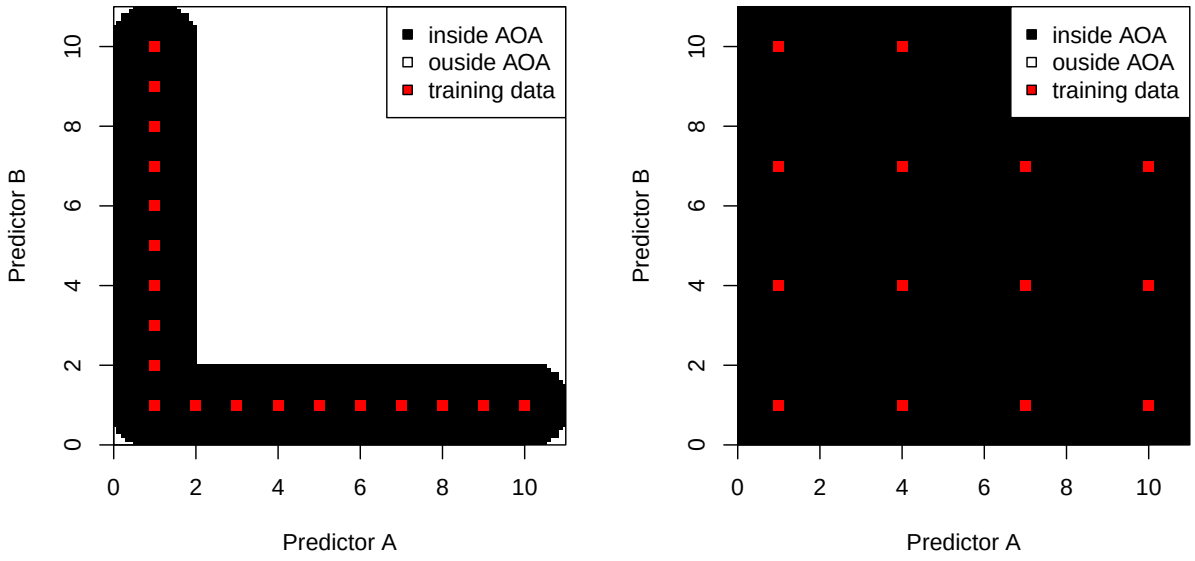


Figure 3: This threshold can be applied to the entire predictor space to derive the AOA (b) from the DI (a) of each new data point.

Additional figures

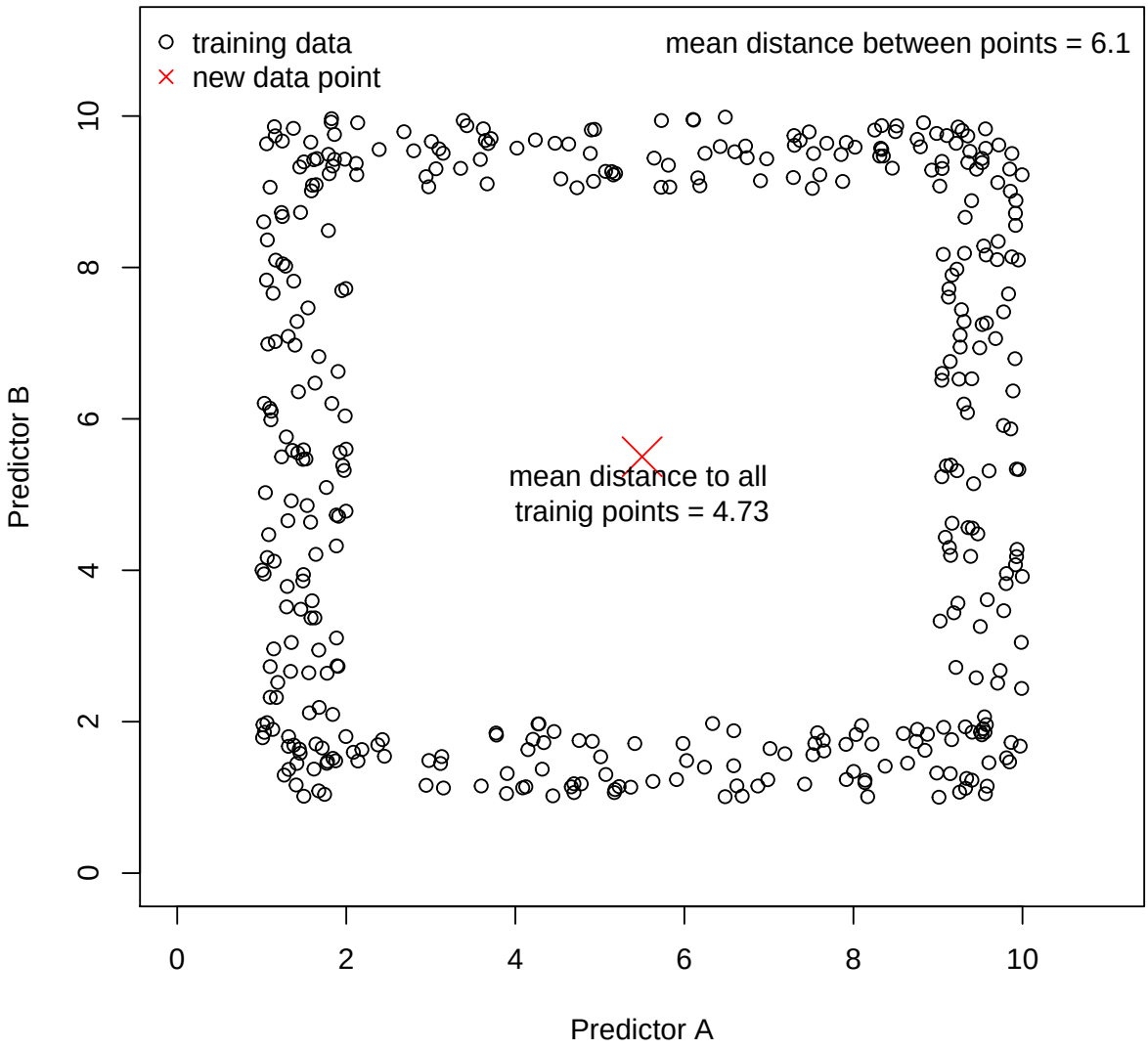
Full range of training data covered

When the full range of environmental conditions is covered the result is that the model can be applied to the entire study area. However, covering the full range of each variable is not enough (figure left) but the combined ranges need to be covered if the model should be applicable to the entire study area (figure right).



Why can't we use the distance to all data points instead of distance to the nearest neighbor?

The distance to all training points is not helpful here because then gaps could not be considered. But gaps are equally problematic (see our first figure in the manuscript) and using the average distance to training points cannot account for that. In the example below the new data point even has a lower average distance to training data although it is clearly in an area not covered by training data.


training data
new data point
mean distance between points = 6.1
mean distance to all
trainig points = 4.73
Predictor A
Predictor B
0
2
4
6
8
10