EGU2020-8492
https://doi.org/10.5194/egusphere-egu2020-8492
EGU General Assembly 2020

Mapping (un)certainty of machine learning-based spatial prediction models based on predictor space distances

Hanna Meyer[1] and Edzer Pebesma[2]
[1]University of Münster, Institute of Landscape Ecology, Münster, Germany (hanna.meyer@uni-muenster.de)
[2]University of Münster, Institute for Geoinformatics, Münster, Germany (edzer.pebesma@uni-muenster.de)

Spatial mapping is an important task in environmental science to reveal spatial patterns and changes of the environment. In this context predictive modelling using flexible machine learning algorithms has become very popular. However, looking at the diversity of modelled (global) maps of environmental variables, there might be increasingly the impression that machine learning is a magic tool to map everything. Recently, the reliability of such maps have been increasingly questioned, calling for a reliable quantification of uncertainties.

Though spatial (cross-)validation allows giving a general error estimate for the predictions, models are usually applied to make predictions for a much larger area or might even be transferred to make predictions for an area where they were not trained on. But by making predictions on heterogeneous landscapes, there will be areas that feature environmental properties that have not been observed in the training data and hence not learned by the algorithm. This is problematic as most machine learning algorithms are weak in extrapolations and can only make reliable predictions for environments with conditions the model has knowledge about. Hence predictions for environmental conditions that differ significantly from the training data have to be considered as uncertain.

To approach this problem, we suggest a measure of uncertainty that allows identifying locations where predictions should be regarded with care. The proposed uncertainty measure is based on distances to the training data in the multidimensional predictor variable space. However, distances are not equally relevant within the feature space but some variables are more important than others in the machine learning model and hence are mainly responsible for prediction patterns. Therefore, we weight the distances by the model-derived importance of the predictors.

As a case study we use a simulated area-wide response variable for Europe, bio-climatic variables as predictors, as well as simulated field samples. Random Forest is applied as algorithm to predict the simulated response. The model is then used to make predictions for entire Europe. We then calculate the corresponding uncertainty and compare it to the area-wide true prediction error. The results show that the uncertainty map reflects the patterns in the true error very well and considerably outperforms ensemble-based standard deviations of predictions as indicator for uncertainty.

The resulting map of uncertainty gives valuable insights into spatial patterns of prediction uncertainty which is important when the predictions are used as a baseline for decision making or subsequent environmental modelling. Hence, we suggest that a map of distance-based uncertainty should be given in addition to prediction maps.