Geophysical Research Abstracts
Vol. 21, EGU2019-1621, 2019
EGU General Assembly 2019

Spatial autocorrelation in geoscientific machine learning applications: moving from data reproduction to spatial prediction

Hanna Meyer (1,2), Christoph Reudenbach (1), and Thomas Nauss (1)

(1) Department of Geography, Philipps Universität Marburg, Marburg, Germany (hanna.meyer@geo.uni-marburg.de), (2) Institute for Geoinformatics, University of Münster, Münster, Germany

Machine learning algorithms find frequent application for spatial prediction tasks of various environmental variables. However, the characteristics of geographic data especially the spatial autocorrelation is widely ignored in such applications. We hypothesize that this is problematic and results in models that can reproduce training data but are unable to make spatial predictions beyond the locations of the training samples. We assume that spatial validation and especially strategies of spatial variable selection are essential to make reliable spatial predictions of environmental data.

To approach this assumption, we use two case studies aiming at remote sensing based prediction of land cover as well as leaf area index for the "Marburg Open Forest", an open research and education site of Marburg University, Germany. As predictor variables we use spectral information from the remote sensing data but we also present terrain-related and geolocation variables to the models. The frequently used Random Forest algorithm is used as machine learning algorithm to train models using non-spatial and spatial cross-validation strategies and we compare how spatial variable selection affects the predictions.

The results show that spatial cross-validation is essential to avoid an overoptimistic view on the model performance. This finding is increasingly and consistently supported in recent literature. However, we could also show that certain predictor variables (especially geolocation variables, e.g. latitude, longitude) can lead to considerable overfitting and result in models that can reproduce the training data but fail in making spatial predictions. The problem becomes apparent in a visual assessment of the spatial predictions that show clear linear artefacts. The spatial variable selection could automatically detect and remove such variables that lead to overfitting, resulting in reliable spatial prediction patterns and improved statistical spatial model performance. Therefore, not only spatial cross-validation but also spatial variable selection must be considered in spatial predictions of environmental data.