Geophysical Research Abstracts
Vol. 20, EGU2018-8365, 2018
EGU General Assembly 2018

How to detect and avoid overfitting in spatio-temporal machine learning applications

Hanna Meyer (1), Christoph Reudenbach (1), Tomislav Hengl (2), Marwan Katurji (3), and Thomas Nauss (1)

(1) Philipps Universität Marburg, Geography, Umweltinformatik, Marburg, Germany (hanna.meyer@geo.uni-marburg.de), (2) ISRIC - World Soil Information, Wageningen, The Netherlands, (3) Center for Atmospheric Research, University of Canterbury, Christchurch, New Zealand

Machine learning algorithms are nowadays well established in environmental sciences and find increasing application for modelling spatio-temporal dynamics. In this context, machine learning algorithms learn from spatio-temporal observations to predict a certain variable at unknown locations and at an unknown point in time. However, it has been shown that the estimated performance of the models highly depends on the validation strategy and that standard approaches can lead to considerable misinterpretations.

We use two case studies of environmental spatio-temporal estimation tasks (air temperature and soil moisture) to demonstrate the importance of target-oriented validation strategies for spatio-temporal prediction models. We therefore compare Random Forest model performances using a standard random k-fold cross validation (CV) with a Leave-Location-Out, Leave-Time-Out, and Leave-Location-and-Time-Out CV as target oriented strategies. The results indicate that considerable differences between random k-fold CV (low RMSE) and target-oriented CV (high RMSE) exist, highlighting the need for target-oriented validation to avoid an overoptimistic view on model results.

We assume that the differences between random k-fold CV and target-oriented CV are attributed to overfitting caused by misleading predictor variables. To approach this problem, we introduce a forward feature selection method that selects variables in view to the target-oriented performance. Using this method we could decrease the degree of overfitting and simultaneously improve target-oriented performances.

The results highlight the importance of target-oriented validation strategies and a careful selection of predictor variables when working with space-time data to obtain valuable results for environmental monitoring.