Geophysical Research Abstracts
Vol. 12, EGU2010-12757, 2010
EGU General Assembly 2010

A review of model-error in digital soil mapping: Confronting statistical soil landscape models with large-scale field validation data

Tim Häring (1,2) and Boris Schröder (2,3)

(1) Bavarian State Institute of Forest Research (LWF), Hans-Carl-von-Carlowitz-Platz 1, 85354 Freising, Germany, (tim.haeringt@lwf.bayern.de), (2) University of Potsdam, Institute of Geoecology, Karl-Liebknecht-Str. 24/25, 14476 Potsdam, Germany, (3) Leibniz-Centre for Agricultural Landscape Research (ZALF) e.V., Soil Landscape Modelling, Eberswalder Str. 84, 15374 Müncheberg, Germany

In the last years a lot of new developed and popular algorithms from the field of machine learning were used for spatial prediction models in digital soil mapping, i.e. support vector machines or ensemble tree methods such as random forests or boosted regression trees. These models have proven as very powerful and demonstrated promising results in many case studies.

One great pitfall of these models is that they tend to overfit the training data. So the challenge when using these models in digital soil mapping is to choose the right values for the tuning parameters and evaluate the prediction ability of the fitted model on an independent test set. A successful prediction of soil types or soil properties depends on a sound and representative training and test-dataset.

In this study, we present a comparison of the statistical model quality (prediction error on an independent test dataset and the out-of-bag error) with findings of an extensive field campaign. We trained and tested random forest models for the prediction of soil map units in four different study areas over an area of altogether 4200 km2 in Bavaria (Germany). In a subsequent field campaign, we validated our predictions at 4500 locations. Our main focus was to validate map units with a high model error in order to improve their prediction. Our attention was not only to validate in strict true or false (like a statistical model validation), but also to evaluate the taxonomic distance of our results, i.e. the soil type, layering, or substrate. This approach provides further insights into possible reasons and spatial patterns of wrong predictions.