EGU21-15211, updated on 16 May 2021
https://doi.org/10.5194/egusphere-egu21-15211
EGU General Assembly 2021

Using machine learning to model the distribution of Vegetation Types across Norway

Lasse Torben Keetz[1,2], Anders Bryn[2,3], Peter Horvath[2], Olav Skarpaas[2], Lena Merete Tallaksen[1], and Indrė Žliobaitė[4]

[1]Department of Geosciences, University of Oslo, Oslo, Norway (l.t.keetz@geo.uio.no)
[2]Natural History Museum, University of Oslo, Oslo, Norway (olav.skarpaas@nhm.uio.no)
[3]Division of Survey and Statistics, Norwegian Institute of Bioeconomy Research, Ås, Norway (anders.bryn@nibio.no)
[4]Department of Computer Science, University of Helsinki, Helsinki, Finland (indre.zliobaite@helsinki.fi)

Machine learning (ML) provides a powerful set of tools that can improve the accuracy of distribution models by automatically representing underlying ecological relationships empirically captured from large sets of data. However, more recent methodological advances in ML that have been less frequently applied, e.g. in the field of deep learning, may yield additional potential for Distribution Modeling. In this project, we use two ML algorithms, Random Forest (RF) and multi-layer feed-forward artificial neural networks (ANN), to predict the occurrences of Vegetation Types (VT) across Norway. Accurate predictions may support environmental management or the validation of earth system models. The VT data (derived from the AR18x18 data set; n=31 classes) covers the entire spatial scope in 0.9 ha plots on a systematically sampled 18 km grid (n = 1,081 plots, n = 22,154 observations). It was obtained through a field-based survey by a group of trained experts between 2004 and 2014. We use the cloud-based platform "Google Earth Engine" to generate a set of remotely sensed predictor variables based on SENTINEL-2 satellite imagery (i.e. surface reflectance from 12 spectral bands and six vegetation indices). These are then combined with ancillary environmental rasters used previously to model Norwegian VT distribution, e.g. representing climate, land cover, or geological properties (n=55, before one-hot encoding). Preliminary results suggest that in both modeling approaches, the generated SENTINEL-2 variables, particularly the Normalized Difference Vegetation Index (NDVI), have the highest predictive power as measured by permutation importance. The mean overall accuracy using 5-fold cross-validation shows only minor differences between the two methods (approx. 0.45 for ANN vs. 0.44 for RF; the respective F1-scores are 0.35 for ANN and 0.34 for RF; most frequent class baseline accuracy = 0.136). The modeling challenges we currently face include a class imbalance in the VT data set, reconciling the different spatial resolutions of the environmental predictors, and discrepancies in the timing of data acquisition. The next steps in the project will be to incorporate spatial cross-validation into the workflow and to analyze the differences between the ML methods in detail (e.g. regarding the ability to model rare VTs or differences in variable importance). Moreover, we will evaluate the possibility to include additional satellite data sources. This work is a contribution to the Strategic Research Initiative ‘Land Atmosphere Interaction in Cold Environments’ (LATICE) of the University of Oslo.