CAN FREELY AVAILABLE GLOBAL AUXILIARY DATASETS IMPROVE ACCURACY OF OBJECT-BASED REGIONAL LAND USE / LAND COVER CLASSIFICATION IN HETEROGENOUS SAVANNA LANDSCAPES?

Pekka Hurskainen[1 2], Andreas Hemp[3], Hari Adhikari[1 2]

[1] Earth Change Observation Laboratory, Department of Geosciences and Geography, University of Helsinki
[2] Institute for Atmospheric and Earth System Research, Faculty of Science, University of Helsinki
[3] Department of Plant Systematics, University of Bayreuth, Germany

Classifying land use / land cover (LULC) with sufficient accuracy in heterogeneous landscapes can be challenging using only satellite data. Since early 1980s, one approach to improve classification accuracy has been the inclusion of features from auxiliary geospatial datasets in classification models. Although this approach has been effective, its use has been mostly limited to pixel-based classifications. Another complication of its use in the Sub-Saharan Africa context has been that auxiliary datasets of reasonable quality, accuracy and resolution have been scarce until recent years.

We wanted to test whether some of the new global geospatial datasets can improve object-based LULC classification accuracy by including them as explanatory features in random forest (RF) classification models. Our study area, 1300 km^2 in size, is located on the southern slopes and surrounding savanna of Mt. Kilimanjaro in Tanzania. The area is characterized by varying topography and heterogeneous mosaic of disturbed savanna vegetation, croplands and built-up areas.

We included features only from freely available datasets with global coverage. These included topographic features from *Alos World 3D* DEM, population features from *WorldPop* and *Global Urban Footprint*, soil features from *SoilGrids*, canopy cover from *Global Tree Canopy Cover*, distance to watercourses from *OpenStreetMap* and statistical seasonal features from *Landsat-8* time series.

The classification was based on image objects (n = 47010) derived from segmentation of four geometrically co-registered, calibrated, atmospherically corrected and mosaicked *Formosat-2* scenes from 2012 with 8-meter spatial resolution. We used ground reference data (n = 1370) for training and validation, as well as for defining 17 LULC classes following the *Land Cover Meta Language* protocol. We trained six different classification models with different set of features in each. The baseline classification model, to which we compared the other models, consisted of only spectral and texture features from *Formosat-2*. To account for the variability of random forest predictions, we iterated the algorithm 50 times for each model.

Based on our experimental setup, we found that inclusion of auxiliary features significantly improved classification accuracy when comparing any of the models to the baseline. The baseline model gave only a moderate median overall accuracy (OA) of 60.7 %. Inclusion of auxiliary features in the next four models, however, increased median OA between 6.1 and 9.5 percentage points. The best overall accuracy, 77.2 %, an improvement of 16.5 percentage points to the baseline, was achieved with the sixth model that included all available features. According to RF unscaled feature importance measure, the three most important auxiliary features of the best performing model were elevation, followed by EVI range (amplitude) and slope.

We have shown that *a priori* geospatial information about local topography, soil characteristics, settlement patterns and vegetation phenology can help to differentiate and classify LULC types in heterogeneous savanna landscapes with significantly better accuracy than using only single-date satellite data. These results are in line with similar studies using pixel-based classification models. Bearing in mind that the auxiliary datasets tested in this study are freely available from the internet for any geographical area in the world, their usefulness in improving classifications of complex heterogeneous landscapes, especially in developing countries, should be further investigated.

Keywords: Image classification, classification accuracy, land use, land cover, auxiliary features, random forest