**2.6 Spectal classification**

For the spectral classification of the study area we exclusively used the “CAST” (Hanna Meyer) and “caret” (Max Kuhn et. al.) R-packages to process a “RandomForest” model with a four folds cross-validation for spectral image classification and a forward feature selection. The “LEGION” R-Package (A. Schönberg) was used to compute RGB and infrared based vegetation indices. The “RStoolbox” R-package was used to perform a Principal Component Analysis (PCA) of the generated vegetation indices.

**2.6.1 Train prediction model**

2.6.1.1 Identify spectral training areas

The first step is to find suitable spectral training areas with the help of aerial RGB-image of the study area. Because we have no reliable ground truth for the study area we are fully dependent on aerial observation. The results must be considered under this point of view as well. We selected five training areas for each class (tree, shrub, grasslands, soil, shadow).

2.6.1.2 Compute vegetation indices for training area

At first we clip the available RGB- and infrared aerial images of the study area to the extent of the generated training areas. The LEGION package was used to compute RGB (VVI, VARI, NDTI, RI, CI, BI, SI, HI, TGI, GLI, NGRDI) and infrared (NDVI, TDVI, SR, MSR) based vegetation indices for the training area. The resulting rasters are now stacked and combined with the three bands of the RGB-image.

2.6.1.3 Principal Component Analysis

The previously generated vegetation indices stack is now used to perform a Principal Component Analysis (PCA) which gathers all needed information from the relatively large raster stack and saves it to three raster-files (PCA1, PCA2, PCA3) containing all necessary information. PCA1 contains approximately 75% of the raster stack information. PCA2 and PCA3 contains approximately 10%. Additional PCA rasters were rejected because of the low percentage (<5%) of explained variables.

2.6.1.4 Merge training dataset

Now the training areas shapes need to be rasterized in order to join the data to the RGB- and previously generated PCA-rasters. The Raster stack now contains the following datasets: RGB-Red, RGB-Green, RGB-Blue, PCA1, PCA2, PCA3 and the rasterized training areas. Finally we cut the raster stack by the extent of the training areas to ensure only the selected areas are used for the RandomForest machine learning algorithm.

2.6.1.5 Train RandomForest

With the previously generated dataset we train the RandomForest machine learning algorithm with a four folds cross-validation and a forward feature selection. The produced model has a mediocre RMSE of 0.21 and is used to predict the study area in the following steps.

**2.6.2 Predict study area**

Following the RandomForest training we use the resulting model to predict the study area. The workflow is basically the same as seen before (see 2.6.1.2&3) except the adding of training areas. The final raster stack contains now all data of the study area (RGB-Red, RGB-Green, RGB-Blue, PCA1, PCA2, PCA3). The result is a classified raster with the classes (1 = tree, 2 = shrub, 3 = grasslands, 4 = soil, 5 = shadow). For the last step we extract the raster values from the prediction raster and assign them to the generated polygons of the five segmentation algorithms. Each polygon contains now a value (classes 1-5) of the majority of containing pixel values and can thereby be classified.