University of Cape Town
Statistical Sciences Department
Honours Project

Assessing the Predictive Performance of Statistical Methods and Machine Learning Algorithms Using Spatial Cross Validation

---

Senamile Dubazane (DBZSEN001)

Shamiso Chikamhi (CHKSHA007)

Supervisors: Sulaiman Salau & Şebnem Er

August 27, 2021

1 Deliverable 4

This week we experimented on the data set to and selecting the covariates. The data set has different covariates adopted from

- developers.google.com
- worldclim.com
- cgiarcsi.community

This data set tries to model the live woody biomass which is a by-product of management, restoration, and hazardous fuel reduction treatments and the end product of natural disasters.

The covariates selected are:

- clay content of the top soil
- organic carbon stocks
- sand content of the topsoil
- slope
- soil organic content
- annual mean solar radiation
- annual mean temperature
- annual mean water vapour

We were able to write a small R script that could load the raster data into R and stacking the different layers to match the locations using the stack function. A challenge we are still facing with this data for now is the format of the x and y coordinates that are not the usual longitude and latitude making it difficult to subset to a specific area.

We attempted to plot the study area on an spplot before sampling. Finally sampled using random sampling without replacement, 500, 1000 and 5000 points. For each sample, we plot the points on the spplot of the data area to show the distribution of the points on the study area.

Finally we fit a random Forrest tree to the data and measure root mean square error using the convectional hold one out cross validation. We observe RMSE increasing from a smaller sample size which may show increase in variability of the data.

2 Plan

Our plan with the data from this point is to:

- understand the data set more and the covariates
- sample a smaller study region

- fit more machine learning models on the random samples
- perform the different kind of cross validations including Spatial cross validation.