Crop Type Mapping in Kenya

Sukhada Ghewari sghewari@ucsd.edu
Shardul Deshpande sdeshpan@ucsd.edu

Abstract

Crop type mapping at the field level is critical for a variety of applications in agricultural monitoring, and satellite imagery is becoming an increasingly abundant and useful raw input from which to create crop type maps. Mapping crop types in smallholder regions faces challenges of small fields, sparse ground truth labels, intercropping, and highly heterogeneous landscapes.[2] To address this, we have worked on developing Machine Learning(ML) techniques for determining crop types with help of Sentinel-2 satellite dataset of Kenya. Overall, the study illustrates comparative analysis and use of various ML algorithms to advance understanding of smallholder agricultural systems.

1 Introduction

Smallholder farms are the most common form of agriculture in the world, particularly in food insecure regions of South Asia and SubSaharan Africa.[4] Recent analysis of sub-national census data suggests that over 50% in Sub-Saharan Africa originate from farms smaller than 5 ha in size. Basic information about these regions is still not well-known and hence investment in ground surveys as well as use of satellites to understand key aspects has gained traction. In this study, we have used time series dataset of Sentinel-2 which is publicly available and is also high resolution to map smallholder fields in Western Province of Kenya.

Table 1: Sentinel-2 Spectral Bands

Spectral Bands			
Band Name	Abbreviation	Central Wavelength (µm)	Resolution (m)
Band 1 - Coastal aerosol	AEROS	0.443	60
Band 2 - Blue	BLUE	0.490	10
Band 3 - Green	GREEN	0.560	10
Band 4 - Red	RED	0.665	10
Band 5 - Vegetation Red Edge	RDED1	0.705	20
Band 6 - Vegetation Red Edge	RDED2	0.740	20
Band 7 - Vegetation Red Edge	RDED3	0.783	20
Band 8 - NIR	NIR	0.842	10
Band 8A - Vegetation Red Edge	RDED4	0.865	20
Band 9 - Water vapour	VAPOR	0.945	60
Band 10 - SWIR - Cirrus	CIRRU	1.375	60
Band 11 - SWIR	SWIR1	1.610	20
Band 12 - SWIR	SWIR2	2.190	20

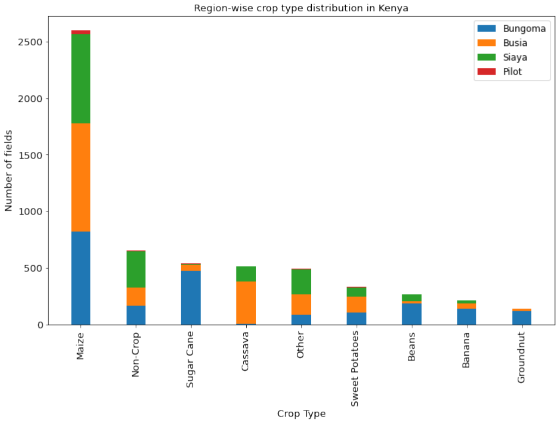
The Sentinel-2 satellites capture thirteen spectral bands ranging from blue to short-wave infrared as mentioned in table 1. The time between observations varied between 5 to 10 days aperiodically with average time of 4.6 days in two consecutive data points. The Green Chlorophyll Vegetation Index is calculated using the formula $GCVI = \frac{NIR}{GREEN1}$. This index is designed to capture the chlorophyll content in the leaves of the different crops[1] and has been shown to be an important feature in crop classification settings.

We have done comparative analysis of several Machine Learning algorithms: Logistic Regression, Random Forest Classifier, Support Vector Machine(SVM), Multi-Layer Perceptron(MLP) and Gradient Boosting as given in 4

2 Related Work

Semantic Segmentation of Crop Type in Africa is previously studied using remote sensing image datasets with help of machine learning specifically deep learning techniques. The studies are done in regions of South Sudan and Ghana with overall accuracy of crop type mapping to be 59.9 and 88.7 respectively for the best deep learning models: a 2D U-Net + CLSTM approach, and a 3D CNN.[3] Similarly, a study is done to perform pixel level data analysis for the region of Kenya using Transfer Learning Techniques[2].

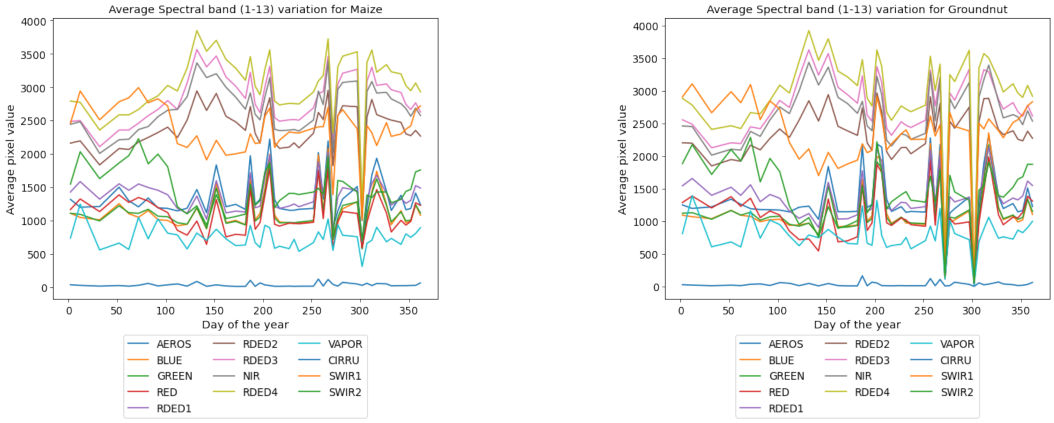
Figure 1: Region-wise crop distribution in Kenya



3 Exploratory Data Analysis

The dataset we are using is provided by Sustainbench [2]. The dataset has 5738 npy files, each conaining 365 days (1 year) time series data for a field. Each land field is represented by a single pixel of Sentinel-2 Satellite image. Hence, each field is represented as an array of shape (365, 1).

Figure 2: Thirteen spectral band signatures

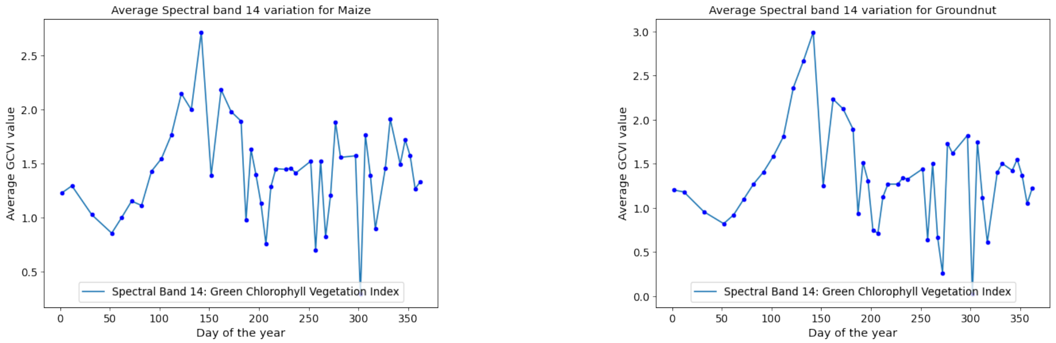


(a) 13 Spectral bands for Maize

(b) 13 Spectral bands for Groundnut

The data is collected from 3 regions in Kenya, namely Bungoma, Busia, and Siaya. The fourth 'Pilot' region is a small region for which data was gathered to check measurement instrumentation. There are 9 classes namely, banana, beans, cassava, groundnut, maize, non-crop, other, sugarcane, and sweet potatoes.

Figure 3: Fourteenth spectral band signatures



(a) 14th Spectral band for Maize

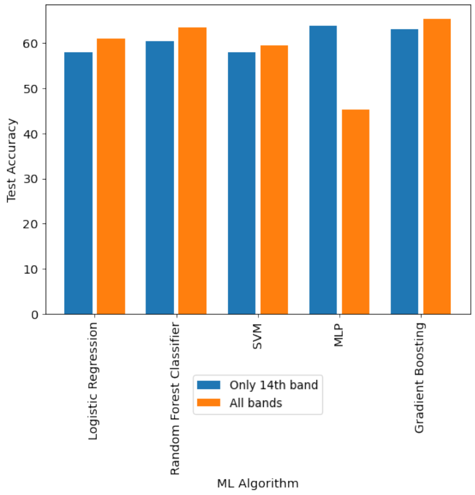
(b) 14th Spectral band for Groundnut

Figure 1 shows the region-wise distribution of crop fields in Kenya. We can see that Maize is the most grown crop in Kenya, followed by Sugarcane and Cassava. We can also see that maize is produced in all regions about equally; sugarcane, banana, beans, and groundnut is mostly produced in Bungoma; while Busia produces a large portion of Cassava.

We have plotted the average spectral band values for a particular crop throughout the year. That forms the time series signature for that crop and it will be used to distinguish one crop from another. First 13 spectral bands have values in the range of (0, 4095). The last spectral band (Green Chlorophyll Vegetation Index) is in the range of (0,3). Figure 2a and figure 2b show the daily average spectral band values for the first 13 spectral bands for Maize and Groundnut respectively. Figure 3a and figure 3b show the daily average spectral band values for the 14th spectral band (GCVI) for Maize and Groundnut respectively. We can see that the difference is significant in spectral band distribution which can be utilised by the learning algorithm to distinguish two crops.

4 Methodology and Results

Figure 4: Test accuracies for different ML algorithms



We used Machine Learning based methods as the preliminary method to train the model to identify the crop type based on the Sentinel-2 Satellite images of the crop fields. It is a multi-class classification problem. We tried Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Gradient Boosting. The train-test split was 80-20 %. In figure 4, the test accuracy for these 5 algorithms is shown for 2 cases:

- When only 14th band (GCVI) is used in the training and testing.
- When all bands are used in training and testing.

In case where all bands were considered, we flattened the 2 dimensional data for 14 bands into a single dimensional array to make it amenable for the ML algorithm training. We observed that the test accuracy increases when all 14 bands are used, except for MLP. We believe that the number of features when we consider all bands is large, and the MLP is not able to extract useful information from large feature space. The highest test accuracy for case 1 was achieved by MLP (63.85 %). For the second case, gradient boosting achieved the highest test accuracy (65.33 %) but at the cost of long training time.

5 Work in Progress

In the initial experiments, we have not made use of the time-series data. We simply treated the data points as individual features with no information about time. We are planning to use LSTM to incorporate the time-series data. We are also planning to use dimensionality reduction techniques to facilitate faster training, especially in case of tree boosting. We also intend to use deep-learning based models to learn the feature space.

References

[1] T.J. Gitelson A.A. Vina A. Ciganda V. Rundquist D.C. Arkebauer. “Remote estimation of canopy chlorophyll content in crops”. In: *Geophys* (2005).

[2] Dan M. Kluger, Sherrie Wang, and David B. Lobell. “Two shifts for crop mapping: Leveraging aggregate crop statistics to improve satellite-based maps in new regions”. In: *Remote Sensing of Environment* 262 (2021), p. 112488.

[3] Rose Rustowicz Robin Cheong Lijing Wang Stefano Ermon Marshall Burke David Lobell. “Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods”. In: *Geophys* (2005).

[4] T Lowder S.K. Skoet J. Raney. “The number, size, and distribution of farms, smallholder farms, and family farms worldwide”. In: *World Dev.* 87.3 (2016), pp. 16–29.