IMPROVE CROP AREA ESTIMATION BY INTEGRATING SPATIAL-TEMPORAL REMOTELY SENSED DATA AND HOUSEHOLD DATA AT LARGE SCALE

Zhe Guo
International Food Policy Research Institute
2033 K Street, NW, Washington DC, 2006 USA

Introduction

Knowing cropland distribution and allocation of crop types is important for the monitoring and planning of agricultural resources at different levels from landscape studies to regional and continental studies. While the global population grows fast, the location specific information on planting area, harvested area, yield and production are vital for food security planning. Knowing "what" is being grown is even more important in developing countries as it will allows decision makers to locate populations that are most vulnerable to food insecurity and poverty. Teff is the most important staple crop by area and value in Ethiopia. In 2011/12, it was estimated that teff made up 20 percent of all the cultivated area in Ethiopia, covering about 2.7 million hectares and grown by 6.3 million farmers. The second most important crop was maize at 15 percent of all cultivated area. When we look at the value of production of teff—using a simple average of producer prices collected by the Central Statistical Agency (CSA) in a large number of producer markets in the country—and compare it to other crops, teff production in 2012 was valued at 1.6 billion USD, again the most important crop in the country. By any standards, teff is an important crop, for farm income as well as food security. On the consumption side, 65 percent of Ethiopia's 85 million people get their "daily bread and livelihood" from it. This research will focus on developing a new approach to produce accurate and reliable teff distribution map in Ethiopia using remotely sensed data and household survey data.

Traditionally, the annual estimation of crop land area and production are conducted by government agencies. In Ethiopia, such information is collected by ESA (Ethiopia Statistical Agency) at subnational level. The data is lack of spatial information and usually lies at the district level and woreda level. On the other land, conventional methods of land use and land cover mapping are labor-intensive and time consuming and usually expensive also which results that land use maps are infrequently prepared with often insufficient details. It is a luxury to produce such map annually and the agricultural related classes in the map are usually generalized (e.g. crop land mosaic). Moreover it become out-of-date soon particularly in rapidly changing environments.

Recent decades, remotely sensed data and its analysis has become a valuable tool for estimating and mapping cropland area and crop types on the ground especially the time frequency and spatial resolution being improved dramatically in recent years1. The advantages of remote sensing can be interpolated in many aspects. It allows us to collect information repeatedly. The cost of using remote sensing technology is relatively low compared to traditional way (e.g. land survey). It also provides an access to map and monitor land productivity and production where has no access or low access to land survey crews in the country. It also allows to be applied on a range of scales from household, local landscape, to regional and country level agriculture development. In addition, together with GIS, it could be combined with other information such as physical, biological, and socio-economic data to analyze the changes in the countries and access the potential impacts.

The normalized difference vegetation index (NDVI) as derived from moderate resolution satellite imagery is widely used to derive land cover and land use product and changes of agriculture productivity because it provides observations at a daily time step, allowing for frequent updating of the vegetation status. Although NDVI is affected by soil background, atmospheric scattering and it relatively insensitive to high biomass levels, it provides sufficient stability to capture the vegetation growth status and conditions, and vegetation phenology. The time series satellite imageries also allow capturing the seasonal and inter-annual changes for a relatively long time period and thus characterize the general vegetation behavior within its spatial footprint. For example the NDVI time series have been applied successfully in the studies of early warning of potential food production problems in African countries. Many researchers have explored the use of time-series NDVI for ecosystem monitoring, land cover or crop identification, and change detection.

The correlation between time series signals of NDVI crop characteristics described above provide information on differentiate land and vegetation objects. Vegetation indices have been extensively used for land over mapping and land use change detections. Even though it is hard to detect crop types using only moderate resolution images, many research have showed promising solutions on integrate remotely sensed data with other spatial ancillary datasets to improve the detection accuracy. This research intend to derive a method to improve the teff mapping and area estimation by combining satellite earth observations and subnational statistics from national household survey. The method is based on 16-day temporal resolution MODIS vegetation data to disaggregate tabular statistical data on cropped area per administrative unit. The aim is to contribute to the development of methods to improve crop density distribution by combining spatial and temporal disaggregated remote sensing data and subnational statistics.

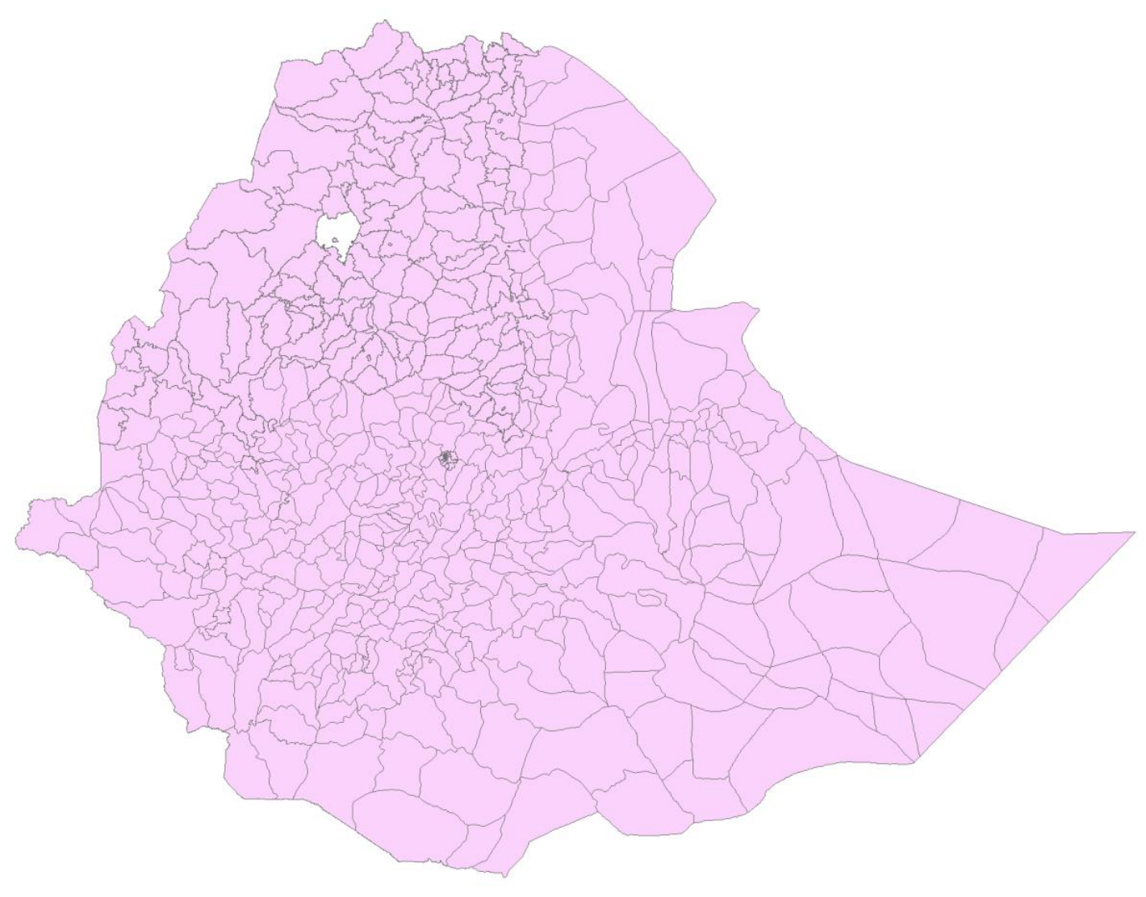
.

Objectives

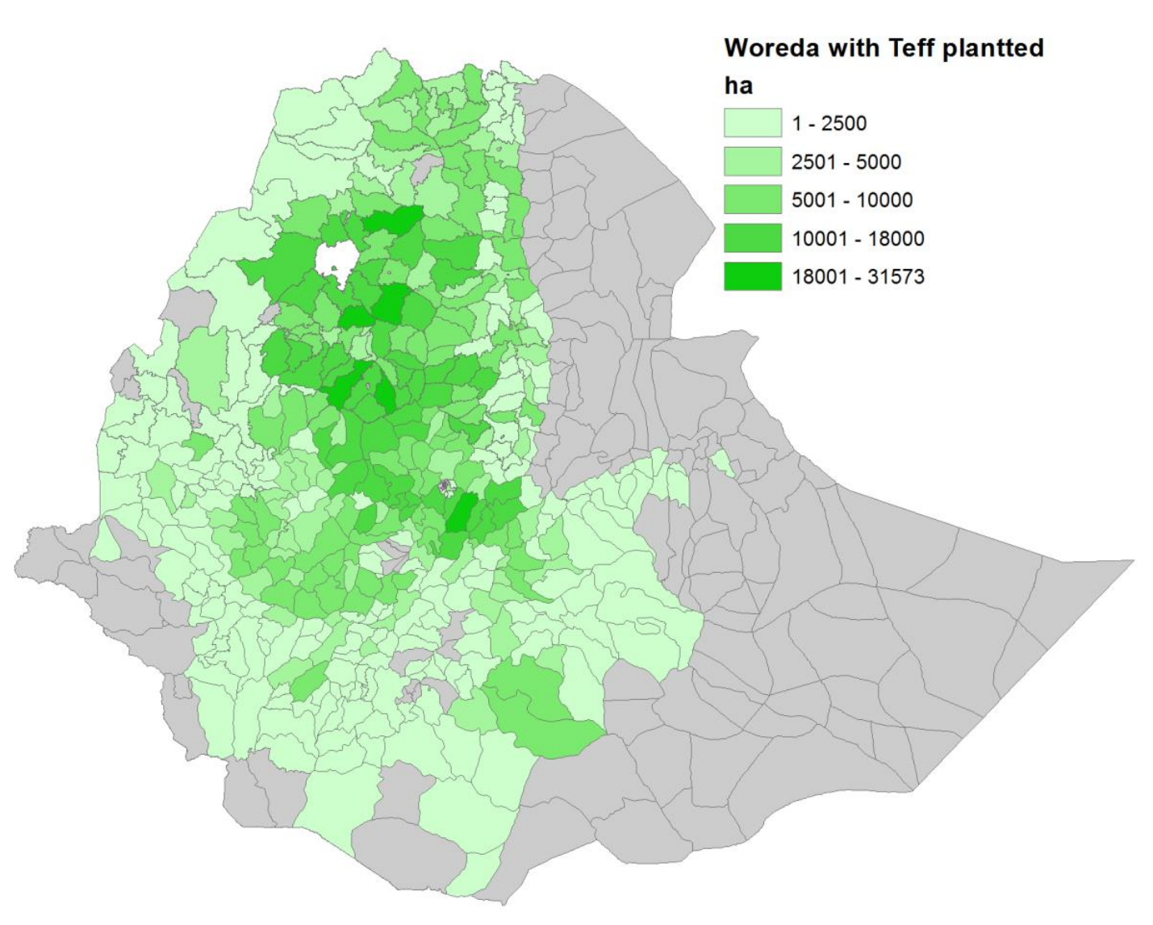
Mapping crop distributions are expensive and resource intensive in traditional way. Crop land area and production information are either collected at coarse resolution, usually taking administrated units as the reporting unit at large scale (e.g. region or countries) or the crop mapping practice is done at landscape level over a relative small spatial extent. In the first case, it is hard to know where exactly the crops are distributed within the units. In the second case, even though the map could contain very detailed agriculture information, the approaches are hard to be scale up and to be applied at a large scale. Using remotely sensed data could be a solution but it is hard to identify crops using satellite data at moderate resolution especially when the land parcel are relative small with a common land parcel size less than 2 ha. The objectives of this research are to develop a method to estimate crop planting area by integrating time series remotely sensed data with household survey information. To be specific, the research questions include: (1) How to use remote sensing data to develop homogeneous clusters at large scale? (2) How to use unsupervised classification framework to quantify the clustering process (3) Develop a statistical approach to combine crop statistics with remote sensing data so the crop specific area could be estimated.

Data and Materials

The NDVI data from Moderate Resolution Imaging Spectroradiometer (MODIS) satellite is processed for the time period of January 1st to December 31, 2001. MOD13Q1 product provided NDVI observations every 16 days at 250-meter spatial resolution in the Sinusoidal projection. Four tiles (h22v08, h21v07, h21v08, and h22v07) in every 16-day period in 2001 are acquired from the NASA data gateway. The images are then mosaic together to cover the entire Ethiopia with a range of 19200*19200 250 meter resolution pixels. The Ethiopia country boundary is then applied to mask out the pixels lies outside the country of each time period. The individual layer is eventually stacked together for further analysis.



The administrative of Woreda in Ethiopia



Teff harvested area at woreda level

The teff area and production statistics are obtained from Atlas of the Ethiopian Rural Economy which is developed jointly by International Food Policy Research Institute (IFPRI), Ethiopian Development Research Institute (EDRI), and Central Statistical Agency (CSA). The 2001/2002 Ethiopian Agricultural Sample Enumeration (EASE) is the major source of the atlas. The EASE is aimed to enable a better understanding of the structure of agriculture for planning, policy making and one of the basic information that is being collected in the atlas is agricultural production and planting/harvested area information. EASE questionnaires were administered to more than 450,000 households across Ethiopia or around 1000 households in each woreda in order to be representative at the woreda level. With a focus on agricultural households, EASE is expected to show higher reliability and consistent with the complete census of the rural population. Harvested area of teff at woreda level in 2011/12 are obtained and processed in the research. The time period of the processed remote sensing data and household survey data are consistent.

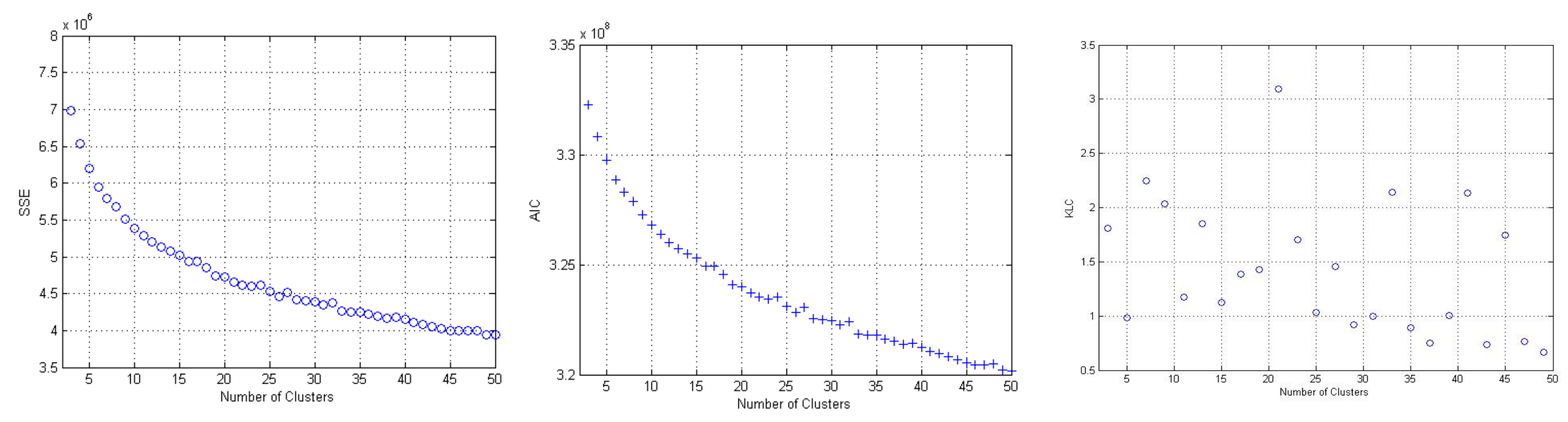
Methods

- K-means unsupervised classification

The K-mean unsupervised classification algorithm was used to derive spectral classes from the 23 NDVI time series data. In this studies, the spectral/NDVI distance is used in the sequential method. It iteratively classifies the pixels, redefines the criteria for each class, and classifies again so that the spectral distance patterns in the data gradually emerge. By iterating this procedure, the means of clusters are shifting and eventually stabled. The optimal iterations and the optimal resulting clusters are selected based on the statistically measurements. Because K-means clustering is a nonlinear optimization process, the results could be different depending on the initial value of the cluster centroid. To avoid the initial value effect, the same k-means routine has replicated 10 times with a set of randomly picked initial values in each time. In order to make sure the partition obtained is stable, the maximum of iteration time are set to 1000 which are way beyond the default 100 in order to make sure the algorithm converge toward an optimal solution in all the k-means runs.

A major challenge in unsupervised classification is the estimation of the optimal number of classes. The number of classes need to be pre-decided before conducting the classification procedure. Theoretically, the number could be range from 1 to totally amount of pixels. Based on the methodology, the K-means approach is designed to minimize the sum of the distances from individual pixel to their corresponding clustering centroids which is the pooled within-cluster sum of squares around the cluster means. There are commonly two ways to locate the optimal number of classes. In the first approach, when plotting the pooled within cluster sum of squares versus the number of clusters k, the square errors decreases monotonically as clusters k increase, but from some k onwards the decrease flattens markedlyThe second method is to develop statistical indicators to examine the variances of different number of clusters. The AIC (Akaike information criterion) and BIC (Bayesian information criterion) are used in examining optimal number of clusters. The AIC and BIC have a better performance since the regularization part is taken into account in the model. Besides criterions mentioned above, Krzanowski and Lai (KL) proposed to derive a criterion for use with the within-group sum of squares objective function trace (W). The KL criterion shows superior performance compare to AIC or BIC especially when the data show heterogeneity in nature.

K-means classification is applied to twenty three 16-date composite NDVI layers with each layer contains over 41 million pixels in order to cover the entire nation. In order to avoid developing too many classes so the grouping information could be practical useful, the maximum number of classes during the process of examining the optimal number of classes are set to 50. The K-means algorithm is applied to these dataset with a resulting number classes from 1 to 50 with a step of 2 in order to save computational time. The pooled within-cluster sum of squares are calculated for each iteration with the predefined number of classes. Both visual check of the "elbow" in the plot of variances and number of plots and criterions are applied together in order to make a consistent decision on optimal number of clusters. In the figure 3, the number of clusters from 1 to 50 are plotted with SSE and AIC. Although the "elbow" point is not very clearly demonstrated in both SEE and AIC plot, the compensation are smaller when the number of clusters are over 25. The variance decreases more when the total number of clusters are small. The trend line of the plot are become flatter when the total number of numbers over 25.



The plot of number of clusters with SSE (left), AIC (middle) and KLC (right)

- Disaggregate subnational statistics using optimized regression

The K-means unsupervised classification algorithm categorized the pixels into 21 classes based on the information derived from time series of NDVI values with each pixel associated with one of the classes. The gridded map of 21 clusters are then converted into vector format. The physical area of each cluster in a woreda is calculated through geospatial analysis by intersecting the cluster polygons with the woreda map of Ethiopia. The physical area of the woreda is equal to the summary of area of all the clusters located in it. Meanwhile, the total harvested area of teff in a woreda is equal to the summary of the physical area of all the clusters times the area percentage of teff in each cluster. If we treated area of cluster in every woreda as explanatory variables and the teff harvested area in a woreda as dependent variable, the area percentage of teff in a cluster could be estimated with an additive multiple linear function. The model could be described as below:

$$Y = \sum_{i=1}^{n} (b_i x_i \quad + \varepsilon)$$

Where
Y = Teff cropped area per woreda (ha) in 2001
bi = regression coefficient
xi = Area of NDVI ith classes per woreda (ha) in 2001
n = Number of NDVI classes in a corresponding woreda
Ɛ = Residual error

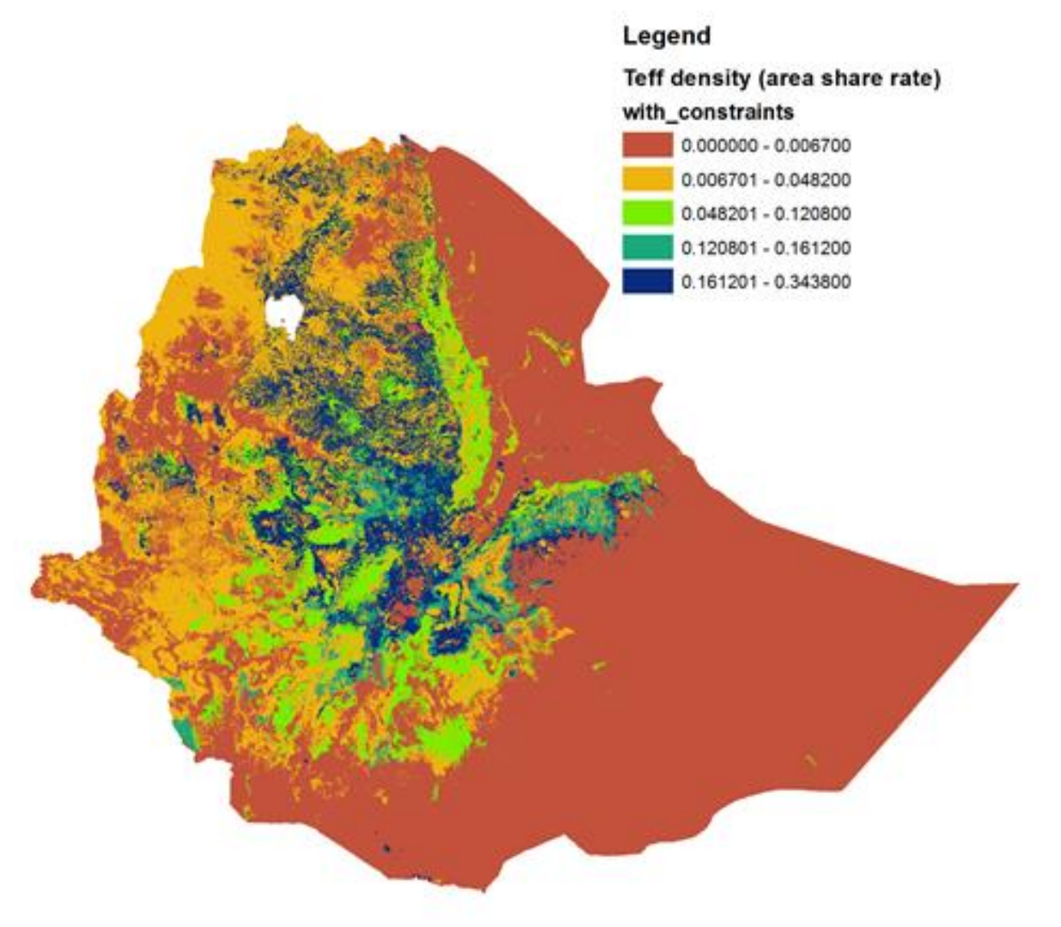
At woreda level, the Y and Xi is known and the only unknown is the bi which is representing the crop density of the clusters in a specific woreda. By definition, the bi value should be constrained from 0.0 to 1.0 (100%), since area density can only be ranged range from 0% to 100% by nature. In each woreda, a series of equations similar to formula above are established. When putting all the equations among the woredas together, the multiple linear regression is set up to resolve the equations for the country. The shares of the teff harvested area in the clusters are displayed in table 1 (left column). It is almost impossible to control the coefficient in a multiple linear process as the coefficients could be negative number or values above 1.

The non-linear constrained minimization optimization (CMO) is introduced in the process in order to resolve the regression problems mentioned above. The CMO is designed to resolve the problem of finding a vector x that is local minimum to a scalar function f(x) subject to constraints on the allowable x which is perfect to this applicatoin. The Trust-Region methods for nonlinear minimization solvers are used in the analysis. The basic idea is to approximate f with a simpler function q which reasonably reflects the behavior of function f in a neighborhood N around the point x. Using CMO, the user can defined the constraints in the optimization process. In this application, the constraints are that the estimated coefficients have to be ranged from 0 to 1. The results from the CMO are shown in Table 1 (right column). Both the additive multiple regression and CMO method give consistent results on estimating the model coefficients which means a relative reliable results given the same input dataset. The estimates from the CMO provide a series of teff area density value with a rate from 0 to 1 and will be used to map teff density in Ethiopia.

Results

The results from the CMO are shown in Table below. Both the additive multiple regression and CMO method give consistent results on estimating the model coefficients which means a relative reliable results given the same input dataset. The estimates from the CMO provide a series of teff area density value with a rate from 0 to 1 and will be used to map teff density in Ethiopia. The estimated shares of teff density in the woreda (bethas) will be linked back to the clusters in the woredas. The final results of teff density map at 250 meter resolution is produced and displayed in figure below.

Share of teff area	Multiple linear regression without constraints	non-linear constrained minimization optimization with constraints (0<= betha <=1)
Betha 1	-0.0255	-
Betha 2	0.1242	0.1208
Betha 3	0.1639	0.1612
Betha 4	0.0336	0.0368
Betha 5	0.2946	0.2924
Betha 6	0.046	0.0482
Betha 7	0.0008	-
Betha 8	0.1018	0.0868
Betha 9	0.0311	0.027
Betha 10	-0.0192	-
Betha 11	-0.0004	-
Betha 12	0.0323	0.0236
Betha 13	0.2725	0.2739
Betha 14	-0.0014	-
Betha 15	0.0007	0.0004
Betha 16	-0.0031	0.0027
Betha 17	0.3474	0.3438
Betha 18	0.0105	0.0067
Betha 19	0.0321	0.0305
Betha 20	0.1018	0.1123
Betha 21	0.056	0.0248


Legend
Teff density (area share rate)
with_constraints
0.000000 - 0.006700
0.006701 - 0.048200
0.048201 - 0.120800
0.120801 - 0.161200
0.161201 - 0.343800