Mapping Cropland in Ethiopia Using Crowdsourcing

Linda See¹, Ian McCallum¹, Steffen Fritz¹, Christoph Perger¹, Florian Kraxner¹, Michael Obersteiner¹, Ujjal Deka Baruah², Nitashree Mili³, Nripen Ram Kalita⁴

¹Ecosystems Services and Management Program, International Institute for Applied Systems Analysis, Laxenburg, Austria
²Department of Geography, Gauhati University, Guwahati, India
³Centre for Geographical Studies, Dibrugarh University, Dibrugarh, India
⁴Department of Geography, B. Borooah College, Guwahati, India
Email: see@iiasa.ac.at

Received June 15, 2013; revised July 15, 2013; accepted August 8, 2013

ABSTRACT

The spatial distribution of cropland is an important input to many applications including food security monitoring and economic land use modeling. Global land cover maps derived from remote sensing are one source of cropland but they are currently not accurate enough in the cropland domain to meet the needs of the user community. Moreover, when compared with one another, these land cover products show large areas of spatial disagreement, which makes the choice very difficult regarding which land cover product to use. This paper takes an entirely different approach to mapping cropland, using crowdsourcing of Google Earth imagery via tools in Geo-Wiki. Using sample data generated by a crowdsourcing campaign for the collection of the degree of cultivation and settlement in Ethiopia, a cropland map was created using simple inverse distance weighted interpolation. The map was validated using data from the GOFC-GOLD validation portal and an independent crowdsourced dataset from Geo-Wiki. The results show that the crowdsourced cropland map for Ethiopia has a higher overall accuracy than the individual global land cover products for this country. Such an approach has great potential for mapping cropland in other countries where such data do not currently exist. Not only is the approach inexpensive but the data can be collected over a very short period of time using an existing network of volunteers.

Keywords: Cropland Mapping; Crowdsourcing; Interpolation; Validation

1. Introduction

Climate change will have far reaching effects on agricultural production in the future, where many studies have shown that crop yields, particularly in Africa, will be compromised under a warmer climate [1,2]. With pressures to increase agricultural production by as much as 70% by 2050 in order to feed the predicted population of 9 billion people [3], global food security issues are of pressing importance. Monitoring food security and evaluating the ability of countries to respond to food shortages requires good baseline information on the spatial distribution of cropland [4], while this spatial information is also a critical input to economic land use models that predict future competition for land across multiple sectors including agriculture [5].

Information on the spatial distribution of cropland comes from two main sources. The first is global land cover maps, which usually have a category for cultivated and/or managed land as well as mosaics of cropland and natural vegetation. By extracting these classes from the global products, it is possible to produce a simple crop mask. The second source of information is national land cover maps, which are usually produced by national mapping or government agencies. However, these maps are not always available or, in some cases, have not even been produced, particularly for some developing countries. For this reason, initiatives such as AFRICOVER have been developed in which the Food and Agriculture Organization of the United Nations (FAO) has worked with member governments in selected countries in Africa to produce land cover and land use maps at a resolution of 100 m [6].

Land cover maps are created using a top down approach in which remotely-sensed data from satellites are classified. For example, the GLC-2000 (Global Land Cover 2000) was created from 14 months of SPOT-VEGETATION at a 1 km resolution for the environmental reference year 2000 [7], while GlobCover 2005, at a 300 m resolution, was created from the MERIS sen-
citizens and interested experts to crowdsource this information. It is possible to use alternative classification algorithms to determine the type of land cover visible from Google Earth, where a sample was crowd-sourced via a Geo-Wiki competition; and 2) data used to validate the map, which originate from multiple sources as explained below.

2.1. Crowdsourced Data on Cropland

USAID (United States Agency for International Development) held a Food Security Open Data Challenge (“Hacking for Hunger”) in the middle of September 2012 where different problems requiring a solution were presented to the hacking community. The Geo-Wiki team proposed a challenge calling for individuals to help collect information on cropland and human settlement across Ethiopia using a simplified version of Geo-Wiki as shown in Figure 1. The blue box on the interface represents a 1 km² pixel, where a random sample of pixels was generated across Ethiopia.

Users were asked to examine the Google Earth image in the 1 km² area and to indicate the degree of visible settlements and cropland from “none” present to a “high” degree. Instructions with examples were provided to help users gain experience in interpreting Google Earth images. Users were encouraged to contribute as many of
these pixels as possible and to share interesting findings via facebook.

The idea behind the challenge was to collect evidence for cross-referencing the crowdsourced information with data from the Land Matrix (http://landmatrix.org/). This project collects the locations of land acquisitions or “land grabbing” so the idea was to see whether areas targeted for land acquisitions are areas of existing cropland and settlement, where Ethiopia is one of the worst affected countries [22]. Some evidence of this has been found [21], which means that population displacement may occur if these land acquisitions take place, and as a consequence, local livelihoods could be negatively affected.

During the “Hacking for Hunger” event, more than 2000 pixels of 1 km² were collected. The site was then opened up to the Geo-Wiki network of volunteers in the form of a three week competition to collect as much information as possible. By the end of the three week period, more than 77,000 pixels were collected where the coverage is shown in Figure 2.

Figure 1. Geo-Wiki interface for data collection.

Figure 2. Distribution of crowdsourced data collected for Ethiopia by cropland category.
Information on the degree of cultivation was collected in four categories: none, low, medium and high. These were reclassified to numerical values as follows: 0%, 20%, 50% and 90% respectively.

2.2. Data for Map Validation

Validation data were available from the GOFC/GOLD validation portal [23], which includes data used to validate the GLC-2000, the STEP (System for Terrestrial Ecosystem Parameterization) database, which is used to train and validate MODIS land cover, and the Visible Infrared Imaging Radiometer Suite (VIIRS) database, which is being developed to validate a new land surface product. Validation data from the Chinese 30 m land cover map were also used [24]. These validations are only at a single point rather than a pixel so they were first reviewed for homogeneity across a larger area, and those points which fell in complex landscapes were removed from the validation exercise. Finally, crowdsourced data from the first Geo-Wiki competition [25] provided an independent source of validation data. The validation data from these different sources were extracted for Ethiopia and each area was then reviewed using Google Earth to ensure quality. After data clean-up, there were 493 validation points available for the accuracy assessment (see Section 3.3).

3. Methods

3.1. Interpolation

A simple inverse distance weighted interpolation method was used to create the Ethiopian cropland map. This interpolation method is based on Tobler’s first law of geography, i.e. things that are close together are more related to one another than things further away [26]. For each grid point to be interpolated, the algorithm identifies all the other points within a certain neighborhood and calculates a weighted vector, w, based on a simple inverse power function:

$$ w(d) = \frac{1}{d^x} $$

where d is the distance and x governs the rate of distance decay. A value of 2 is most commonly chosen for x. Each interpolated point is then calculated as a weighted average of its neighbors. For this exercise the default values in ArcGIS were used, i.e. a power of 2 and a neighborhood of 12 points. Although different settings and interpolation methods could be employed, the point of this study was not to optimize the method of interpolation but rather to demonstrate how even simple interpolation can effectively be used to create a cropland map based on crowdsourcing.

3.2. Difference Maps

The crowdsourced cropland map was compared to the GLC-2000, MODIS and GlobCover where the cropland classes were extracted and the images were reclassified to produce maps showing areas with the presence and absence of cropland in Ethiopia. The GLC-2000, MODIS and GlobCover were resampled to match the resolution of the crowdsourced cropland map (i.e. 1 km). The images were then subtracted to produce difference images in order to highlight the main areas of disagreement.

3.3. Map Validation

The crowdsourced cropland map was validated using the dataset described in Section 2.2 by extracting the presence or absence of cropland at each of the validation locations. A confusion matrix was then populated (Table 1) and the overall accuracy was calculated as follows:

$$ \text{Accuracy} = \frac{\sum_{i,j} x_{i,j}}{\sum_{i} \sum_{j} x_{i,j}} \times 100 $$

where i is the class from the map of interest, e.g. the crowdsourced cropland map, j is the class from the validation data set and n is the total number of classes.

In addition, user’s and producer’s accuracies were calculated as follows:

$$ \text{User’s accuracy}_i = \frac{x_{i,j}}{\sum_{j} x_{i,j}} $$

$$ \text{Producer’s accuracy}_j = \frac{x_{j,i}}{\sum_{i} x_{j,i}} $$

The user’s accuracy reflects errors of commission while the producer’s accuracy refers to errors of omission. The same accuracy measures were then applied to the GLC-2000, MODIS and GlobCover for Ethiopia and the cropland class using the same validation dataset.

4. Results

The interpolated cropland map for Ethiopia is provided

<table>
<thead>
<tr>
<th>Table 1. Confusion matrix.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class from the cropland map</td>
</tr>
<tr>
<td>Class 1</td>
</tr>
<tr>
<td>Class 1</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>Class n</td>
</tr>
</tbody>
</table>
in Figure 3 while a difference image between this map and the GLC-2000 is provided in Figure 4.

The GLC-2000 shows much more cropland than the interpolated map although there are some areas (shown in red) where the interpolated map indicates cropland but the GLC-2000 does not.

Figures 5 and 6 contain images showing the spatial differences between the interpolated map (Figure 3) and MODIS and GlobCover respectively. In contrast, MODIS shows much less cropland than the interpolated map, missing quite a significant area of cropland in the central Eastern part of the country known as the Harerghe Highlands where rainfed agriculture is definitely reported [27]. GlobCover, like the GLC-2000, shows more cropland than the interpolated map throughout most of the country but also misses areas in the central part where the interpolated map indicates cropland. Such a simple visual comparison serves to highlight the large spatial differences between each of the land cover maps and the interpolated cropland map, but it also highlights the differences among the different products.

Table 2 contains the accuracy measures for the three
global land cover products for Ethiopia and the interpolated map. Overall accuracies range between 74.5% for Globcover to 89.3% for the interpolated map, showing just under an 8% increase in accuracy over the second best product, i.e. MODIS. Thus, the map produced through interpolation of crowdsourced data has the best overall accuracy.

In terms of user’s accuracy, all the maps have high values for the category “No crop” but lower values for the presence of cropland. This indicates that identification of areas without cropland is easier than the opposite case, which is not surprising for the global products since

<table>
<thead>
<tr>
<th>Maps</th>
<th>Overall accuracy</th>
<th>User’s accuracy</th>
<th>Producer’s accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLC-2000</td>
<td>77.3</td>
<td>90.5</td>
<td>48.1</td>
</tr>
<tr>
<td>MODIS</td>
<td>81.8</td>
<td>83.2</td>
<td>67.5</td>
</tr>
<tr>
<td>GlobCover</td>
<td>74.5</td>
<td>89.3</td>
<td>43.9</td>
</tr>
<tr>
<td>Crowdsourced cropland map</td>
<td>89.3</td>
<td>91.7</td>
<td>78.8</td>
</tr>
</tbody>
</table>

Table 2. Accuracy assessment.
they are poor at detecting croplands in areas of low agricultural intensification. This is because the spectral signatures and temporal profiles are similar to grasslands, which would include areas of Ethiopia. The values for identification of the “Crop” class in the global land cover products range from 67.5% for MODIS to 43.9% for GlobCover. However, from a user’s perspective the interpolated map has the highest accuracy for the presence of cropland at 78.8%.

For the producer’s accuracy, both MODIS and the interpolated map performed very well in terms of labeling areas as having “No crop” while the GLC-2000 and GlobCover performed less well. However, MODIS performed very poorly in falsely labeling cropland while the other three products performed similarly, with producer’s accuracies for the presence of cropland varying between 66.3% and 69.6%.

5. Discussion and Conclusions

A cropland map for Ethiopia was created using crowdsourced data collected via Google Earth and Geo-Wiki, which was shown to have higher accuracy than global land cover products in the cropland domain. However, the user’s and producer’s accuracy for the presence of cropland clearly indicates that there is still room for improvement in the crowdsourced map. In this regard there are three main issues that deserve further discussion. The first concerns the ability of the volunteers to identify cropland from Google Earth images. Although some basic training materials were provided, more could be done to control for quality, e.g. control points could be used throughout the competition to show volunteers where mistakes have been made. Images that were difficult to interpret, and which were flagged by the confidence that the volunteers placed on their interpretation, could be discussed interactively so that others could benefit from feedback on a variety of landscapes. These are features that are currently not implemented but are planned for future versions of Geo-Wiki.

A second source of error could arise from the density of samples and the interpolation method used. Although the samples collected through crowdsourcing covered roughly 5% of the area of Ethiopia, there will be areas that require a higher density of samples in order to characterize them with higher accuracy. Rather than randomly sampling pixels across all of Ethiopia, we could have sampled more frequently in areas where cropland is thought to occur, using the three global land cover maps as a basis for driving this sampling. Moreover, as mentioned above, the interpolation method chosen was one of the simplest available in order to demonstrate the feasibility of this approach. Other interpolation methods and additional data layers, e.g. a digital elevation model, slope, rainfall and temperature, could be used to improve the interpolation of cropland.

A third issue concerns the validation data used to calculate the accuracy measures. Rather than creating a stratified validation sample from the crowdsourced map, existing validation data were used from a range of different sources, which will reflect different resolutions and different temporal windows. However, each validation point was verified using Google Earth. We would also argue that the change over time is a very small component of what is a much higher uncertainty due to misclassification error. The validation sample consisted of 18.6% cropland and the remaining points were non-cropland. Based on FAO statistics, the area harvested in Ethiopia has varied between 11.2% and 13.5% over the period 2005 to 2011 [28] so the validation dataset is only slightly higher in terms of cropland than the FAO figures.

Although the accuracy measures indicated that the crowdsourced cropland map performed better than the global products, they only represent one way of judging map quality. The difference images served to highlight that there are large spatial differences between the crowdsourced map and the three global land cover products in the cropland domain as well as between the three global land cover maps themselves. Ultimately all of the products must be judged by the end user, which requires their use in different applications, ideally feeding back to the producers where there are problems. The crowdsourced cropland map for Ethiopia is freely available for downloading from the following website: http://beta-hybrid.geo-wiki.org. We would encourage users to experiment with the commenting tools on the website to provide us with feedback.

The bottom up approach to mapping cropland that was demonstrated in this paper has considerable potential in areas where cropland maps do not currently exist. Using a motivated network of volunteers and a more targeted sampling scheme, it would be possible to map the entire world in this way.

6. Acknowledgements

This research was supported by the Austrian Research Funding Agency (FFG) via the LandSpotting (No. 828332) and Farm Support (No. 833421) projects and by the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 262937 (ISAC: Information Service on Agricultural Change).

REFERENCES


