The Use of Crowd Sourced Data to Validate Global Land Cover Datasets

Alexis Comber[1], Linda See[2], Steffen Fritz[2], Marijn van der Velde[2], Christoph Perger[2]

[1]Department of Geography, University of Leicester, UK, LE1 7RH
Tel. (+44 116 252 3812) Fax (+44 116 252 3854)
Email: ajc36@le.ac.uk
[2]International Institute of Applied Systems Analysis (IIASA), Laxenburg, A-2361, Austria

KEYWORDS: VGI, geographically weighted model, accuracy, cropland

1. Introduction

Land cover and land cover change are important variables in understanding land-atmosphere interactions and the impacts of climate changes. Land cover changes for example are one the major drivers and consequences of climate change and have been described as the greatest threat to biodiversity (Feddema et al., 2005).

A number of different global datasets describe land cover. However, there is considerable disagreement between tem regarding the amount and spatial distribution of land cover features particularly in relation to forest and cropland. Differences of as much as 20% have been found in the amount of land classified as arable or cropland when global land cover products have been compared (Fritz and See, 2005; See and Fritz, 2006; Fritz et al., 2011). The uncertainties in these products are so great that their input into climate models is problematic and they cannot be used for land cover change detection. Whilst, formal approaches for determining the reliability of land cover data have been proposed (e.g. Strahler, 2006), many land cover datasets are not validated using these protocols (Foody, 2002).

This paper describes a method for using volunteered land cover data, collected through a Geo-Wiki approach, to contribute to land cover validation (Fritz et al., 2012).

The Geo-wiki incorporates a web-based interface using Google Earth (Perger et al., 2012) and invites volunteers to record the land cover at locations throughout the world. The Geo-wiki campaign was undertaken in the autumn of 2011 using the Human Impact Geo-Wiki (http://humanimpact.geo-wiki.org). Each volunteer was asked to complete an online tutorial to demonstrate the process and then they were asked to record the land cover at a random sample of locations. Based on their interpretation of the landscape they assigned each location to one of 10 predefined land cover classes, including a cropland class of Cultivated and managed.

2. Methods

The land cover at a total of 17382 locations were recorded by the volunteers for an African case study. These are shown in Figure 1.

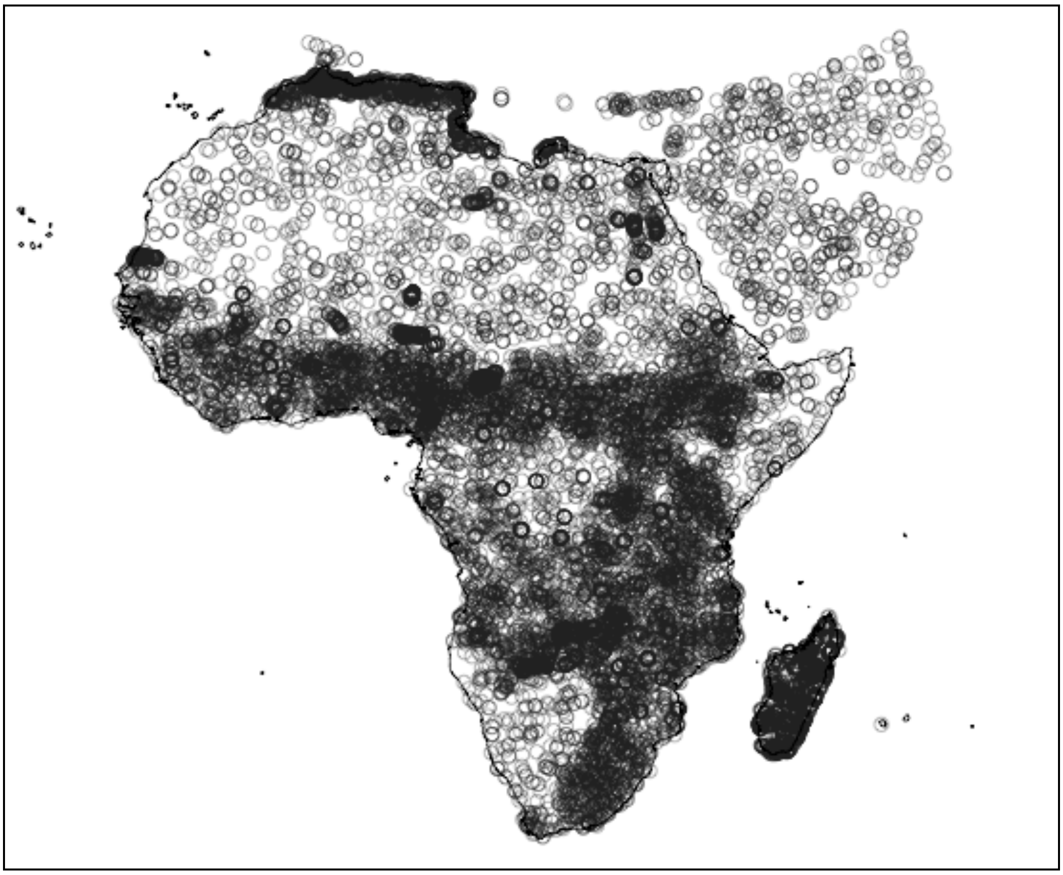


Figure 1. The density of VGI on land cover in the study area

At each location the proportion of cropland recorded in 6 global land cover datasets was extracted. The global land cover datasets were JRC, GlobCover, GLC, MODIS, GeoCover and Hansen.

The analysis evaluated VGI cropland presence / absence against the presence / absence of cropland, as determined by a threshold, in each of the global datasets in turn.

Measures of spatially distributed correspondence between VGI cropland and cropland data from the global datasets were generated using the methods described in Comber (2012), Comber et al (2012a) and Comber et al (2012b). In brief, these approaches use a geographically weighted kernel to generate local correspondence measures (user's, producer's, portmanteau and partial portmanteau) based on the calculation of local confusion matrices throughout the study area. These allow the spatial variations in correspondences to be mapped.

The volunteered data on cropland describing Cultivated and managed were converted to a binary dataset indicating the presence of either of these classes. The cropland proportions in each of the global land cover datasets were converted to binary data, indicating the presence [1]or absence [0] of cropland, where presence was defined by applying a threshold of greater than 50% cropland.

A geographically weighted regression was applied to each of the global datasets in turn to determine the probability that the presence VGI cropland classes, $P(y = 1)$, was predicted by the presence of cropland in each of the global datasets together, in the following way:

Equations should be centred on the page and numbered consecutively in the right-hand margin, as below. They should be referred to in the text as Equation 1.

$$P(y_i = 1) = \mathrm{logit}(b_{0(u_i,v_i)} + b_1 x_{1(u_i,v_i)} \ldots + b_n x_{n(u_i,v_i)}) \quad (1)$$

where $P(y_i = 1)$ is the probability that the VGI cropland cover class, y at location i is correctly predicted, b_0 is the intercept term, $x_{1..n}$ are the proportions of cropland indicated by the 6 global datasets under consideration, $b_{1..n}$ are the slopes and (u_i, v_i) is a vector of two dimensional co-ordinates describing the location of i over which the coefficient estimates are assumed to vary.

The basic idea here is that all of the global dataset cropland proportions are considered as a series of independent variable in the logistic regression. The analysis returns a coefficient for each of those, the highest of which indicates the strongest effect in predicting the presence of VGI cropland.

The logit function is defined as

$$\mathrm{logit}(q) = \frac{exp(q)}{1 + exp(q)} \quad (2)$$

where q is any value.

The bandwidth was set to include 1% of the data points for the local geographically weighted analyses computed at each location in the study area as defined by 100km grid cover the study area.

From the coefficients of this analysis, it is possible to develop 4 measures of geographically weighted correspondence as described in Comber (2012), Comber et al (2012a) and Comber et al (2012b):

1) User's accuracy: estimate the probability that a location is correctly classified (ie the same as the volunteer class). It indicates errors of commission and inclusion.

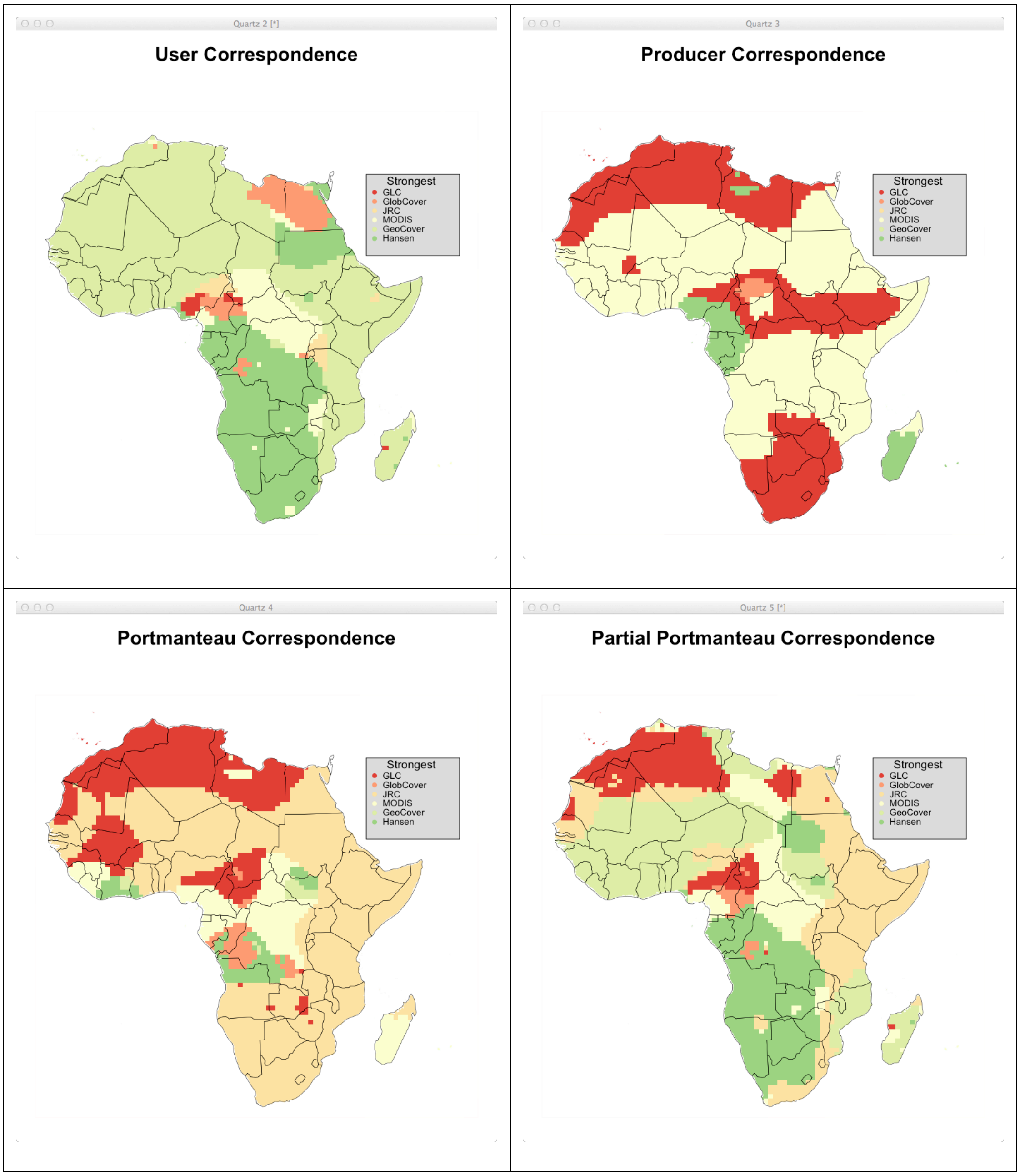
2) Producer's accuracy: estimate of probability that a reference location is correctly classified as indicated in the global data. It indicates errors of omission and exclusion. It indicates the probability that cropland locations are omitted from the global datasets.

3) Portmanteau accuracy: provides an estimate of the probability that the volunteered data and the global datasets are the same. It includes both the presence and absence of the class in this evaluation and as such it includes both specificity and sensitivity.

4) Partial portmanteau accuracy: provides an estimate of the probability that the volunteered data and the global datasets are the same, under the proviso that at least of the datasets indicates the presence of cropland.

3. Results

At each location, for each correspondence measures, the highest coefficient value indicated the global dataset with the strongest correspondence to the VGI data on cropland. These are shown in Figure 2.



User Correspondence

Producer Correspondence

Portmanteau Correspondence

Partial Portmanteau Correspondence

Figure 2. The global datasets with the highest correspondence to volunteered information on land cover.

4. Discussion

Crowd sourced data and Volunteered Geographical Information (VGI – Goodchild, 2007) have the potential to support analyses and research in many areas of science and social science. This research has shown that VGI on land cover can be used to make some inferences on the quality of global land cover products and how that varies spatially.

One of the problematic issues in incorporating volunteered data in scientific analyses relates to its unknown provenance, specifically its unknown quality and reliability. The need to understand the quality of VGI is critical if volunteered information is going to be used for scientific research. Without such measures there will always be a lack of trust or credibility in these data. On-going research is developing novel methods for determining the quality of VGI.

References

Comber A.J., (2013). Geographically weighted methods for estimating local surfaces of overall, user and producer accuracies. *Remote Sensing Letters*, **4(4)** pp373-380 DOI: 10.1080/2150704X.2012.736694

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., Foody, G.M. (in press). Using volunteered geographic information to evaluate global land cover datasets. Paper accepted for publication in *International Journal of Applied Earth Observation and Geoinformation*

Comber, A., Fisher, P.F., Brunsdon, C. and Khmag, A. (2012). Spatial analysis of remote sensing image classification accuracy. *Remote Sensing of Environment*, **127** pp237–246.

Foody, G.M. (2002). Status of land cover classification accuracy assessment, *Remote Sensing of Environment*, **80** pp185-201.

Fritz S, McCallum I, Schill C, Perger C, See L, Schepaschenko D, van der Velde M, Kraxner F and Obersteiner M., (2012). Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling and Software*, **31** pp110-123.

Fritz, S. and See, L. (2005), Comparison of land cover maps using fuzzy agreement. *International Journal of Geographical Information Science*, **19(7)** pp787-807.

Fritz, S., See, L., McCallum, I., Schill, C., Obersteiner, M., van der Velde, M., Boettcher, H., Havlik, P. and Achard, F. (2011), Highlighting continued uncertainty in global land cover maps to the user community. *Environmental Research Letters*, **6** 044005.

Goodchild M.F. (2007). Citizens as sensors: the world of volunteered geography. *Geojournal* **69** pp211-221.

Perger C, Fritz S, See L, Schill C, Van der Velde M, McCallum I and Obersteiner M 2012 A campaign to collect volunteered geographic Information on land cover and human impact. In: Jekel T, Car A, Strobl J and Griesebner G (Eds.) *GI_Forum 2012: Geovizualisation, Society and Learning*. Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach, pp.83-91.

See, L. and Fritz, S. (2006), Towards a global hybrid land cover map for the year 2000. *IEEE Transactions on Geosciences and Remote Sensing*, **44** pp1740-1746.

Strahler A H, Boschetti L, Foody G M, Friedl M A, Hansen M C, Herold M, Mayaux P, Morisette J T, Stehman S V and Woodcock C E 2006 *Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps*, European Commission, Joint Research Centre, Ispra, Italy, EUR 22156 EN, 48pp.

Biography

Lex Comber is a Reader in Geographic Information at the University of Leicester
Linda See, Steffen Fritz, Christoph Perger and Marijn van der Velde are all Research Scholars in the Ecosystems Services and Management research program at IIASA