Evaluating Global Land Cover Datasets: Comparing VGI on Cropland with Formal Data

Alexis COMBER[1], Chris BRUNSDON[2], Linda SEE[3] and Steffen FRITZ[3]
[1]University of Leicester/UK · ajc36@le.ac.uk
[2]University of Liverpool/UK
[3]IIASA, Laxenburg/Austria

This contribution was double-blind reviewed as extended abstract.

Abstract

This paper explores the use of crowd sourced data on land cover. It compares data collected through the Geo-Wiki system describing cropland land cover with cropland data recorded in global land cover datasets. Using an African case study, local confusion matrices are calculated at discrete locations. At location the strongest correspondences are used to infer which global dataset best describes cropland at each location. Future research areas are suggested.

1 Introduction

Land cover and land cover change have been found to be important variables in understanding land-atmosphere interactions and particularly the impacts of climate changes (FEDDEMA et al. 2005). A number of different global datasets describe land cover but with considerable disagreement between them in the amount and distribution of different types of land cover features particularly in relation to forest and cropland. This is a long-standing problem, one that has been recognised since the emergence of different global datasets in the early 1990s which differ, for example, by as 20% differences in the amount of land classified as arable or cropland (FRITZ et al. 2011). The potential errors and uncertainties associated with these products mean that their input into applications such climate models is questionable. They certainly cannot be used to model land cover change.

There is considerable interest in crowdsourced data, also referred to volunteered geographical information (VGI – GOODCHILD 2007) when it includes a locational reference. The European Commission has funded projects to determine how such data may be used to manage crises or emergencies[1], to monitor deforestation[2] and the head of GMES has noted

[1] http://projects.jrc.ec.europa.eu/jpb_public/act/publicexportworkprogramme.html?actId=453&d-2325611-p=6

[2] http://www.gmes-masters.com/sites/default/files/media/inline/gmes_results_booklet_2012.pdf

Jekel, T., Car, A., Strobl, J. & Griesebner, G. (Eds.) (2013): GI_Forum 2013. Creating the GISociety.
© Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach. ISBN 978-3-87907-532-4.
© ÖAW Verlag, Wien. eISBN 978-3-7001-7438-7, doi:10.1553/giscience2013s91.

crowdsourcing to be critical for future environmental monitoring[3]. Many other research areas are aware of the benefits of employing the crowd[4]. The amount and types of VGI have increased dramatically in recent years, taking the form of posts, blogs, geotagged tweets, geotagged pictures, data on specific events (e.g. weather) due to the increased availability of location-enabled digital data capture devices. There are considerable opportunities for many areas of research and investigation if the potential of information collected by volunteers can be harnessed (COHN 2008, COLEMAN 2010, HAKLAY et al. 2010, HAND 2010, PERGER et al. 2012, COMBER et al. 2013) and recent work has shown how VGI can be used to generate forest inventories (VAN DER VELDE et al. 2012) and to validate global land cover maps (COMBER et al. 2013).

This research explores the utility of crowd-sourced or volunteered information about land cover, specifically cropland. Crowdsourced data on cropland are used to determine which of the many global datasets best describe cropland land use in different areas, using an African case study. The aims of this research were to:

1) Estimate spatially distributed measures of correspondence between volunteered information on cropland with cropland data from 7 global land cover datasets.
2) Use the estimated measures to infer which of the global datasets best describes cropland in each location.

2 Methods

Local measures of correspondence between volunteered land cover and land cover from global datasets were generated from geographically weighted measures of correspondence (COMBER 2012, COMBER et al. 2012, COMBER et al. 2013) and calculated at regular locations through the study area. These methods are described in more detail below.

Volunteered data on cropland was captured through a Geo-Wiki campaign incorporating a web-based interface using Google Earth (PERGER et al. 2012). Volunteers were recruited informally for different campaigns involving both the remote sensing community and the wider public. The campaign was undertaken in the autumn of 2011 using the Human Impact Geo-Wiki (http://humanimpact.geo-wiki.org). Based on their interpretation of the landscape, each volunteer assigned sample locations to one of 10 predefined land cover classes, including the class of *Cultivated and managed.* The land cover at a total of 17382 locations were recorded by the volunteers for an African case study. At each location the proportion of cropland recorded was extracted, from the following global datasets: JRC, GlobCover, GLC, MODIS, GeoCover and Hansen.

The analysis sought to identify correspondences between the proportions of cropland indicated in the global datasets and the presence of cropland. The volunteered data on cropland describing *Cultivated and managed* were converted to a binary dataset indicating the presence of either of these classes. These were logistically regressed against the proportions of cropland indicated at each sample location by the global land cover datasets. Only the

[3] http://www.spacenews.com/sites/spacenews.com/files/print_issue_pdfs/SPN_20121001_Oct_2012.pdf

[4] http://techland.time.com/2011/09/19/foldit-gamers-solve-aids-puzzle-that-baffled-scientists-for-decade/

Globcover dataset was not a significant predictor of VGI cropland. Then, a geographically weighted logistic regression of VGI cropland against the proportions of cropland indicated by the global datasets, except for Globcover, was used to determine which of the global datasets best predicted the VGI data.

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

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; $x_{1 \ldots n}$ are the proportions of cropland indicated in the 6 global datasets ($n = 6$) under consideration; $b_{1 \ldots n}$ 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 *logit* function is defined by:

$$\operatorname{logit}(q) = \frac{\exp(q)}{1 + \exp(q)}$$

where q is any value.

A 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 covering the study area. 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.

3 Results

The results of the GWR analysis, identifying the best predictor of VGI cropland, identify for each location in the study area, the independent variable representing the global land cover datasets. That is, the global dataset with the largest coefficient was considered as the strongest predictor of the volunteered information on cropland use. The spatial distributions and the global land use datasets they suggest are shown in figure 1.

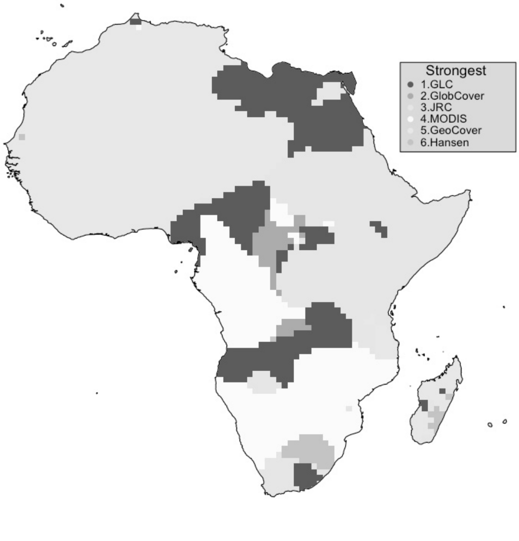

Fig. 1:
The spatial distribution of the global land cover datasets that best predict VGI on cropland

4 Discussion

There are obviously many assumptions embedded in this research. However, it is very much at an early stage and is part of a wider suite of research activity that is variously:

- Developing methods for determining VGI quality (BRUNSDON & COMBER 2012, COMBER et al. 2013, FOODY & BOYD 2012);
- Considering formal informatics methods for combining evidence (Bayesian, Probability, Dempster-Shafer, Possibility, Endorsement theories) to determine at what point does the weight of evidence indicate that the VGI is 'believable';

Exploring different thresholds of user reliability moving away from some of the more naïve methods such as 'Linus Law' which is a simple form of 'accepting the crowd' (e.g. HAKLAY et al. 2010).

References

BRUNSDON, C. & COMBER, A. J. (2012), Experiences with Citizen Science: Assessing Changes in the North American Spring. Geoinformatica DOI 10.1007/s10707-012-0159-6.

COHN, J. P. (2008), Citizen science: can volunteers do real research? BioScience, 58 (3), 192-197. doi:10.1641/B580303.

COLEMAN, D. (2010), The potential and early limitations of volunteered geographic information. Geomatica, 64 (2), 27-39.

COMBER A. J. (2013), Geographically weighted methods for estimating local surfaces of overall, user and producer accuracies. Remote Sensing Letters, 4 (4), 373-380.

COMBER, A., FISHER, P. F., BRUNSDON, C. & KHMAG, A. (2012), Spatial analysis of remote sensing image classification accuracy. Remote Sensing of Environment, 127, 237-246.

COMBER, A., SEE, L., FRITZ, S., VAN DER VELDE, M., PERGER, C. & FOODY, G. M. (2013), Using control data to determine the reliability of volunteered geographic information about land cover. International Journal of Applied Earth Observation and Geoinformation, 23, 37-48.

FEDDEMA, J. J., OLESON, K. W., BONAN, G. B., MEARNS, L. O., BUJA, L. E., MEEHL G. A. & WASHINGTON W. M. (2005), The importance of land-cover change in simulating future climates. Science, 310, 1674-1678.

FOODY, G. M. & BOYD, D. S. (2012), Exploring the potential role of volunteer citizen sensors in land cover map accuracy assessment, in Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science (Accuracy 2012), Florianopolis, Brazil, 203-208.

FRITZ, S., SEE, L., MCCALLUM, I., SCHILL, C., OBERSTEINER, M., VAN DER VELDE, M., BOETTCHER, H., HAVLIK, P. & ACHARD, F. (2011c), Highlighting continued uncertainty in global land cover maps to the user community. Env. Research Letters, 6, 044005.

GOODCHILD, M. F. (2007), Citizens as sensors: the world of volunteered geography. Geojournal, 69, 211-221.

HAKLAY, M., BASIOUKA, S; ANTONIOU, V & ATHER, A, (2010), How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. Cartographic Journal, 47 (4), 315-322.

HAND, E. (2010), Citizen science: people power. Nature, 466 (7307), 685-687.

PERGER, C., FRITZ, S., SEE, L., SCHILL, C., VAN DER VELDE, M., MCCALLUM, I. & OBERSTEINER, M. (2012), A campaign to collect volunteered geographic Information on land cover and human impact. In: JEKEL T, CAR A, STROBL J. & GRIESEBNER G (Eds.), GI_Forum 2012: Geovisualization, Society and Learning. Berlin/Offenbach, Wichmann Verlag, 83-91.

VAN DER VELDE, M., SEE, L. & FRITZ, S. (2012), Conservation: Citizens add to satellite forest maps. Nature, 490, 342.