COMPARING NATIONAL DIFFERENCES IN WHAT PEOPLE PERCEIVE TO BE *THERE*: MAPPING VARIATIONS IN CROWD SOURCED LAND COVER

A. Comber [a], P. Mooney [b], R.S. Purves [c], D. Rocchini [d], A. Walz [e]

[a] School of Geography, University of Leeds, Leeds, LS2 9JT, UK, a.comber@leeds.ac.uk
[b] Department of Computer Science, National University of Ireland Maynooth, Ireland Peter.Mooney@nuim.ie
[c] Department of Geography, University of Zurich, 8057 Zurich, Switzerland ross.purves@geo.uzh.ch
[d] Fondazione Edmund Mach, 38010 S. Michele all'Adige, Italy duccio.rocchini@fmach.it
[e] Potsdam Institute for Climate Impact Research, 14412, Potsdam, Germany ariane.walz@pik-potsdam.de

Commission VI, WG VI/4

KEY WORDS: Semantics, Geo-Wiki, VGI, Land cover, Citizen Science

ABSTRACT:

This paper describes a simple comparison of the distributions of land cover features identified from volunteered data contributed by different social groups – in this case comparing two groups of Geo-Wiki campaigns. Understanding the impacts on analyses of citizen science data contributed by different groups is critical to ensure robust scientific outputs and to fully realise the potential benefits to formal scientific research. It is well known that different people, with different backgrounds and subject to different cultural factors, hold varying landscape conceptualisations. This paper analyses volunteered geographical information on land cover to generate land cover maps. It uses a geographically weighted approach to generate land cover mappings. The mappings generated by different groups (in this case a from a specific unnamed country) are compared and the results show how the predicted land cover distributions vary, with large differences in some classes (e.g. Barren land, Shrubland, Wetland) and little difference in others (e.g. Tree cover). This suggests that for some landscape features cultural and national differences matter when it comes to using crowdsourced data in formal scientific analyses and highlights the potential problems of *not* considering contributor backgrounds in citizen science. This is important because such data re now routinely being used to develop global land cover data, to generate uncertainty estimates of existing global land cover products and to generate global forest inventories. These in turn are being suggested as suitable inputs to such things as global climate models. A number of critical research directions arising from these findings are discussed.

1. INTRODUCTION

There is much interest in using crowd-sourced data, data generated through citizen science activities or what Goodchild (2007) referred to as *volunteered geographical information* to support formal scientific endeavours. As a result the scientific community has explored different opportunities arising from crowd-sourced data collection and analysis (Cohn, 2008; Coleman, 2010; Haklay et al, 2010; Hand, 2010) and there has been an explosion of applications underpinned by crowd-sourced data in nearly all areas of scientific investigation: from astronomy (Raddick et al, 2010) to zoology (Silvertown, 2009). In the domain of land cover / land use, the European Commission has funded a number of projects considering how such data may be used to help manage crises and emergencies[1], to develop Citizen Observatories for Land Cover and Land Use[2] and to monitor deforestation[3]. The reasons for these initiatives in the context of land cover are various but include uncertainties over future funding of remote sensing in Europe[4] and cost benefits (for example LUCAS sampling is expensive, costing €6.42m)[5]. As a result a number of crowd-sourced land cover data collection systems have been initiated and perhaps the best known is the Geo-Wiki system developed by Perger et al (2012) at IIASA, Austria although others exist (Pistorius & Poona, 2014; Vaz & Arsanjani 2015; Kinley, 2013). Geo-Wiki has been used for a number of campaigns and has seen some system development in order to increase data collection and contribution.

The rise of activities such as Geo-Wiki within mainstream scientific investigation provides a critical research context to the research presented in this paper. Geo-Wiki collects land cover data from volunteers and a number of applications have been developed to, for example, assess the quality of existing land cover products (Fritz et al, 2009), determine their uncertainties (Fritz et al., 2011) and generate hybrid global land cover maps (See et al., 2015). However, one of the critical issues associated with the use of citizen data relates to its quality (Foody et al, 2013; Comber et al, 2013a). One key problem is that different contributors or volunteers may have different underlying conceptualisations of the features that observed and thus recorded in crowd-sourced data. In the context of land cover, variation in concepts result in different interpretations of the boundaries between classes and so in the land cover data that recorded. This issue is illustrated by ethnophysiography (Derungs et al, 2013; Mark and Turk 2003) and by linguistic and cultural factors (Smith and Mark, 1998) and is well known in the context of *formal* land cover creation (for example, Comber et al, 2005; Comber et al, 2008). It is important to note that some differences may reflect real variation in the way landscapes are perceived, and there may be no “right” answer in terms of land cover or land use, without consideration of context. In formal land cover creation, divergent conceptualisations are mitigated by the inclusion of experimental designs: data collection protocols, training, sampling designs, QA procedures etc. These ensure that the inferences from any data analysis are statically robust.

This paper evaluates the potential impacts of using volunteered data by comparing the land cover data contributed by two groups, one formed of volunteers solely from one country,

[1] http://projects.jrc.ec.europa.eu/jpb_public/act/publicexportworkprogramme.html?actId=453

[2] https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/2196-sc5-17-2015.html

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

[4] http://news.eoportal.org/policy/121001_pol1.html

[5] http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/en/lan_esms.htm

This contribution has been peer-reviewed.
Editors: A.-M. Olteanu-Raimond, C. de-Runz, and R. Devillers
doi:10.5194/isprsarchives-XL-3-W3-71-2015

named *Gondor* to avoid making inferences based on national stereotypes, the other containing all other nationalities.

2. DATA AND CASE STUDY

The research uses data collected through the Geo-Wiki project. Geo-Wiki is an open, web and app interface. As part of the registration process, volunteers are asked to describe their experience and where they are from. Once registered, volunteers contribute to different campaigns in which they describe the land cover at a series of randomly selected locations with Google Earth providing background imagery. In this research, data from two of these, capturing land cover using the same 10 classes were combined and the data for North and South America selected. The distributions of the data are summarised in Table 1.

Class	Group	
	Others	*Gondor*
Tree	3100	4144
Shrub	673	1860
Grass	1607	1554
Crop	1305	1176
Wetland	543	245
Urban	91	107
Snow	368	256
Barren	856	593
Water	555	364

Table 1. The land cover data collected by different volunteers

It is evident that despite a random sample of locations there are large differences between Gondor and Others especially in the number of images classified as *Shrub* and *Barren* land cover. The distribution of the data points is shown in Figure 1.

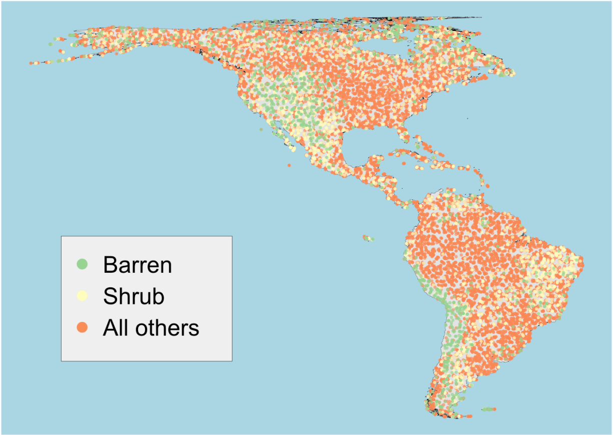

Figure 1. The distribution of the data

A Kernel Density Estimation (KDE) was used to generate surfaces for visualising the data trends and to provide a visual reference for later analyses. The KDEs in Figure 2 describe the probability of a class being present at each location.

In order to examine the potential impacts of using data operationally that was contributed by volunteers from different countries, with no consideration of the number of volunteers, their experience, background and training, the data were separated into 2 subsets each with groups. The first contained *Professional* and *Non-Professional*, the second contained contributed volunteers from *Gondor* and *Others* (i.e. data from volunteers from all other countries).

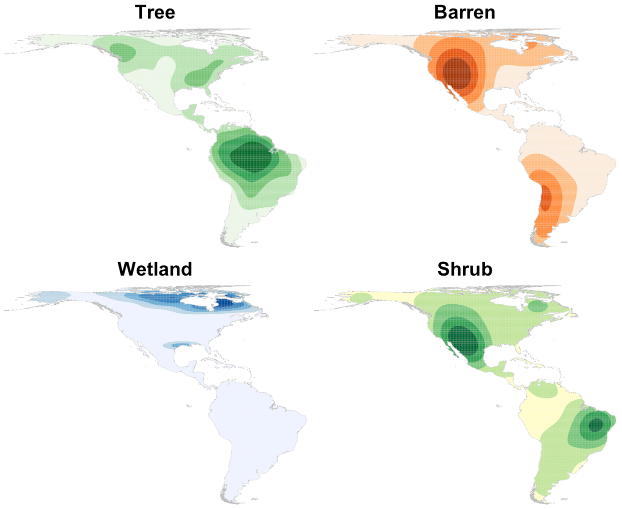

Figure 2. The KDE surfaces for 4 classes

To compare the impacts of different groups, geographically weighted averages were computed for each class at each point on 50km grid under a 50km kernel. This generated a value in the range [0, 1] at each of the 80,073 grid locations for each class, and the class with the greatest value was assigned. This approach is a smoothing approach similar to that used by Comber (2013a) to determine fuzzy memberships distributions.

3. RESULTS

Generating KDEs of the data contributed by different groups provides a convenient way to summarise the potential impacts. Figure 3 compares the surfaces generated for 2 land covers of data contributed volunteers from *Gondor* and data contributed by *Others*. Evidently different features are mapped in different locations (*Wetland* in Louisiana) and to different degrees (*Barren* land in Chile) by different groups.

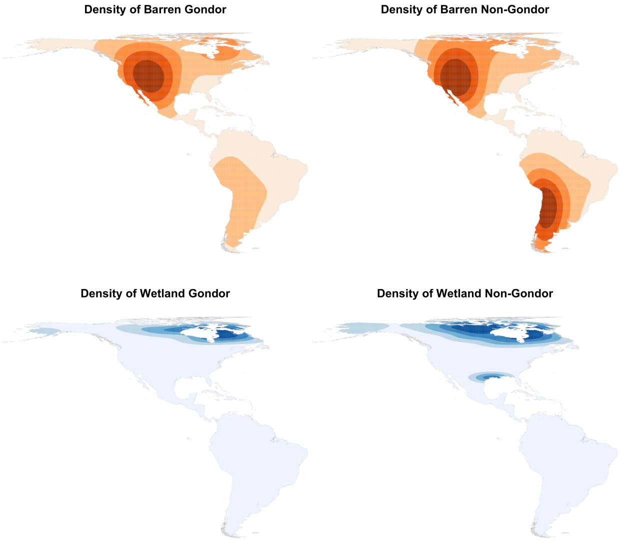
Density of Barren Gondor
Density of Barren Non-Gondor
Density of Wetland Gondor
Density of Wetland Non-Gondor

Figure 3. KDEs from data contributed by different groups

The Geo-Wiki data was analysed in the following way. At each location in a 50km grid and for each class, a local probability measure was generated, describing the probability of that class being present at that location. The local probability was calculated using a geographically weighted regression model that analysed the number of data points of that class under a 50km kernel, weighting each point's contribution to the model by its distance to the kernel centre. This generates probability surfaces for each class and at each location, the class with the greatest probability was assigned. Figure 4 shows the land cover map generated in this way using data from all contributors. This land cover mapping is generated in the same way the operational data described by See et al (2015) and Schepaschenko et al (2015). Figure 4 provides a baseline against which to compare the impacts of only using data from specific groups (Figure 5 and Figure 6).

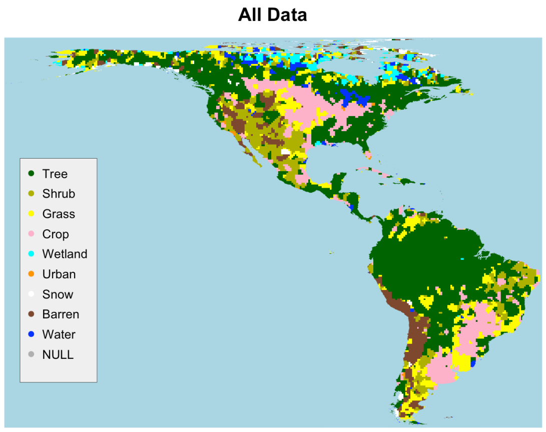

Figure 4. The land cover data generated by all contributors

Comparing Figures 4, 5, and 6 there are clear differences between the groups, although the maps generated from the data contributed by *Others* has much smaller differences to the map in Figure 4. Interesting and potentially significant differences are the subtle but important differences in the distributions of the *Wetland* class, *Shrub* and *Barren* and *Crop* and *Grass*.

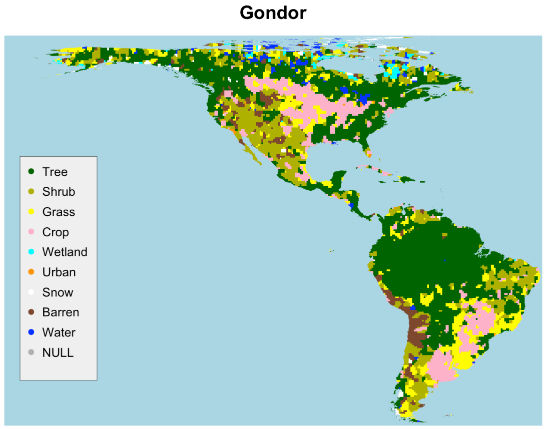

Figure 5. The land cover data generated by contributors from Gondor

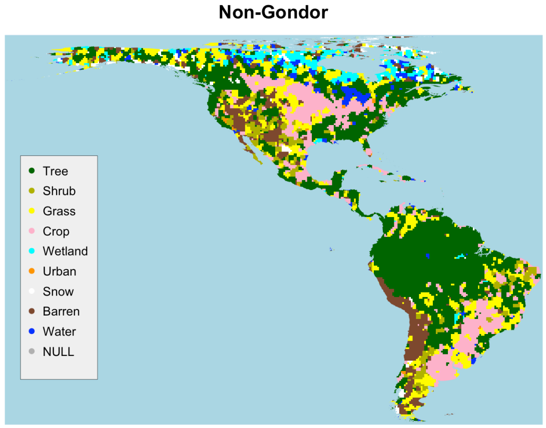

Figure 6. The land cover data generated by people from other countries

4. DISCUSSION

The analysis in this paper presents an initial analysis comparing differences between data contributed by 2 groups of volunteers, 21 from *Gondor* and 70 people of other nationalities or whose background was unknown. These numbers are very typical for a Geo-Wiki campaign. These preliminary results for a North and South American case study suggest that the well known differences in *what is perceived to be there* by different groups **matters**, even with a very simple 10 class nomenclature. This has profound implications for a number of on-going research activities that are using crowdsourced data to generate hybrid global land cover datasets from existing (but uncertain) global datasets and from crowdsourced data (See et al., 2015; Schepaschenko et al., 2015). These researches have not considered the impacts of contributor cultural or national background but their datasets they are creating are being suggested as suitable and improved inputs to global climate change models.

The methods used to develop the land cover maps apply a simplified approach. For example, a bandwidth of 50km was chosen rather than the optimal bandwidth being determined, as recommended in the GWR literature, and no comparisons with formal data were made. This objective of this work was to evaluate the potential impacts of contributors with different backgrounds. The number of contributors does allow for the possibility that the any observed differences (for example comparing Figure 5 and Figure 6) may be statistical noise / variation rather than representing any underlying Additionally, the land cover was allocated to the single class with the greatest probability at each location. In many locations many classes have similar levels of (high) probability suggesting the need for spatially distributed measures of uncertainty in the class allocation.

However, the geographically weighted approach to analysing Geo-Wiki data is at the core of many current activities which are using the data to construct land cover maps for operational usage (See et al., 2015; Schepaschenko et al., 2015). The research presented in this paper suggests that, as well thinking about developing measures of crowd-sourced data quality (Comber, 2013a; Foody et al 2014), there is a critical need to consider some of the 'COSIT' considerations related to spatial cognition[6] and who the volunteers are, where they come from,

[6] http://www.cosit.info

what their background (cultural and professional) is and so on. There is also a need to consider how volunteers are recruited and whether that can be done in a more representative way or even a targeted way, where for example, data contributed by an individual who fails to meet some criteria are excluded from analysis.

There are a number of areas for further work. These include the need to compare the interaction of professional experience, identified as an important factor (Comber et al, 2013b; Foody et al, 2014), with nationality and the degree to which differences between the groups are ameliorated when this is considered. Second, to explore whether the differences and similarities between groups persists for different classes in different areas: this study focused on a particular study area where one of the groups may have greater knowledge and experience. Third, to develop methods to consider how to integrate citizen science data with formal data and to develop quantitative measures related to citizen semantics. For example, there may be large difference in the way that different groups of citizens resolve any ambiguity they may have around the labelling of features – consider the example of Forest illustrated by Comber et al (2005). The Geo-Wiki volunteers attached measures of their confidence in the class labels that were attached to each data point. This may capture their uncertainty for example around their understanding of the affordances associated with the different land cover types. Fourth, recent activity in citizen science is using data from an increasingly wide range of data sources. Features may be labelled in different and novel ways depending on how the crowd-sourced data are captured or mined. Fifth, there is a need to consider the impact of digital and other divides on the nature of the information that is contributed and the potential for biases towards western, developed populations in recording and describing the world and the formalisation of that set of cultural perceptions. This is also influenced by the nature of the technologies used to capture and share such information. On-going work will consider these issues.

ACKNOWLEDGEMENTS

Thanks to Steffen Fritz and his team at IIASA for the use of the Geo-Wiki data. The data and the code will be available to researchers on request.

REFERENCES

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., See, L., Fritz, S., Van der Velde, M., Perger, C., Foody, G.M. (2013a). 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

Comber, A., Brunsdon, C., See, L., Fritz, S. and McCallum, I. (2013b). Comparing expert and non-expert conceptualisations of the land: an analysis of crowdsourced land cover data. *Lecture Notes in Computer Science: Spatial Information Theory*, 8116: 243-260, doi: 10.1007/978-3-319-01790-7_14

Comber, A.J., Fisher, P.F. and Wadsworth, R.A., (2008). Semantics, Metadata, Geographical Information and Users. Editorial *Transactions in GIS,* 12(3): 287–291

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (2005). What is land cover? *Environment and Planning B: Planning and Design*, 32:199-209.

Derungs, C, Wartmann, F, Purves, R S, and Mark, D M, (2013). The meanings of the generic parts of toponyms: use and limitations of gazetteers in studies of landscape terms. *Spatial Information Theory* (pp. 261-278).

Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., & Boyd, D. S. (2013). Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Transactions in GIS*, *17*(6), 847-860.

Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., ... & Comber, A. (2014). Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality. *The Cartographic Journal*, 1743277413Y-0000000070.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., ... & Obersteiner, M. (2009). Geo-Wiki. Org: The use of crowdsourcing to improve global land cover. *Remote Sensing*, *1*(3), 345-354.

Fritz, S., See, L., McCallum, I., Schill, C., Obersteiner, M., Van der Velde, M., ... & Achard, F. (2011). Highlighting continued uncertainty in global land cover maps for the user community. *Environmental Research Letters*, *6*(4), 044005.

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

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of openstreetmap and ordnance survey datasets. *Environment Planning B*, 37(4):682–703

Hand, E. (2010). Citizen science: people power. *Nature* 466(7307):685–687.

Kinley, L. (2013). Towards the use of Citizen Sensor Information as an Ancillary Tool for the Thematic Classification of Ecological Phenomena. *Proceedings of the 2nd AGILE (Association of Geographic Information Laboratories for Europe) PhD School 2013*.

Mark, DM and Turk, AG (2003). Landscape categories in yindjibarndi: Ontology, environment, and language. In: Kuhn, W., Worboys, M.F., Timpf, S. (eds.) *COSIT 2003*. LNCS, vol. 2825,pp. 28–45. Springer, Heidelberg

Perger C, Fritz S, See L, Schill C, Van der Velde M, et al. (2012). A campaign to collect volunteered geographic Information on land cover and human impact. In: Jekel T, Car A, Strobl J, Griesebner G, editors. GI_Forum 2012: Geovisualisation, Society and Learning. Berlin / Offenbach: Herbert Wichmann Verlag. pp. 83–91.

Pistorius, T., & Poona, N. (2014). Accuracy assessment of game-based crowdsourced land-use/land cover image classification. In *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International* (pp. 4780-4783). IEEE.

Raddick, M.J, Bracey G, Gay P L, Lintott C J, Murray P, Schawinski K, Szalay AS & Vandenberg, J. (2010). Galaxy zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review*, *9*(1), 010103.

Schepaschenko, D., See, L., Lesiv, M., McCallum, I., Fritz, S., Salk, C., ... & Ontikov, P. (2015). Development of a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics. *Remote Sensing of Environment*, *162*, 208-220.

See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., Comber, A. Perger C, Schill C, Zhao Y, Maus V, Siraj MA, Albrecht F, Cipriani A, Vakolyuk M, Garcia A, Rabia AH, Singha K, Marcarini AA, Kattenborn T, Hazarika R, Schepaschenko M, van der Velde M, Kraxner F, Obersteiner, M (2015). Building a hybrid land cover map with crowdsourcing and geographically weighted regression. *ISPRS Journal of Photogrammetry and Remote Sensing*, *103*, 48-56.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, *24*(9), 467-471.

Smith, B. and Mark, D.M. (1998). Ontology and Geographic Kinds. In *Proceedings of 8th International Symposium on Spatial Data Handling*, editors T. K. Poiker and N. Chrisman, (International Geographical Union, Vancouver) pp308-320

Vaz, E., & Jokar Arsanjani, J. (2015). Crowdsourced mapping of land use in urban dense environments: An assessment of Toronto. *The Canadian Geographer/Le Géographe Canadien*, DOI: 10.1111/cag.12170