The Impact of Contributor Confidence, Expertise and Distance on the Crowdsourced Land Cover Data Quality

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

This contribution was double-blind reviewed as extended abstract.

Abstract

There is much interest in the opportunities for formal scientific investigations afforded by crowdsourcing and citizen sensing activities. However, one of the critical research issues relates to the 'quality' of the data collected in this way. This paper uses volunteer data on land cover collected under the Geo-Wiki system, where contributors label the land cover class at a series of locations, with expert labels at the same locations. It examines the statistical relationships between the accuracy of volunteer labels, their self assessed confidence in labelling, their 'experiential distance' to the location under consideration and the level of their domain expertise. The results show that distance has a minor effect on the reliability of land cover labelling, and that generally expertise has a greater effect, but not for all land cover classes.

1 Introduction

A number of ongoing research activities have a specific focus on crowd sourced data, so called volunteered geographical information (GOODCHILD 2007). For example, under the EU-COST programme there are currently two actions considering these topics – 'Mapping and the Citizen Sensor'[1] and ENERGIC[2]. One of the critical issues in the use of crowdsourced data relates to the quality and reliability of the information that is contributed by volunteers and members of the public, and a number of methods have been suggested for assessing this (BRUNSDON & COMBER 2012; COMBER et al. 2013, FOODY & BOYD 2012, FOODY et al. (in press)).

This research explores and examines the relationships between the accuracy and reliability of the crowd sourced labelling of land cover, as well as the contributors' labelling

[1] http://www.cost.eu/domains_actions/ict/Actions/TD1202
[2] http://www.cost.eu/domains_actions/ict/Actions/IC1203

Vogler, R., Car, A., Strobl, J. & Griesebner, G. (Eds.) (2014): GI_Forum 2014. Geospatial Innovation for Society.
© Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach. ISBN 978-3-87907-545-4.
© ÖAW Verlag, Wien. eISBN 978-3-7001-7652-7, doi:10.1553/giscience2014s309.

confidence, their level of domain expertise, and their geographic distance to the location under consideration (experiential distance). It uses volunteer data captured through the Geo-Wiki Human Impact campaign[3] that incorporates a web-based interface using Google Earth (PERGER et al. 2012), undertaken in the autumn of 2011. The aims of this research were to examine:

1) the statistical relationships between self-assessed measures of confidence, geographic, and experiential distance, with the quality and reliability of volunteered data;
2) how the relationships varied for different land cover types.

2 Methods

The Geo-Wiki Human Impact campaign asked users to indicate the land cover class at a series of randomly sampled locations. It had a number of interesting additional features. First, it captured a number of variables additional to the land cover label at each location including contributor confidence in their land cover label, scored as *Sure*, *Quite sure*, *Less sure* and *Unsure*. Second, as part of the registration process, volunteers were asked to answer some background questions and to include information about their experience, profession and expertise. The expertise and home country of each volunteer were inferred from this information. Third, the campaign randomly introduced 300 'control' points, locations at which the land cover had been agreed on by a panel of experts. The information recorded at the control locations is the subject of the analyses described below.

In total the 297 control locations were scored 7,363 times by 65 volunteers, with each control point labelled on average 25.7 times, and each labelling 117.5 control locations. The locations of the control data and the volunteers are shown in Figure 1. The locations of the 65 volunteers and the variations in expertise are shown in Table 1.

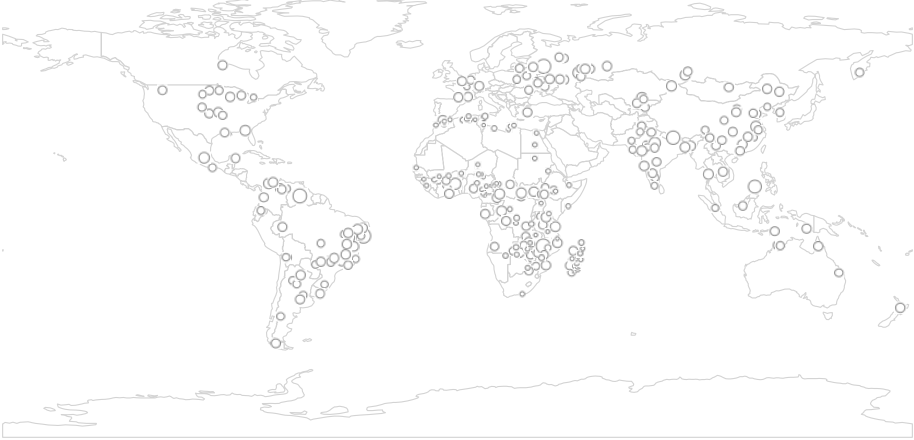

Fig. 1: The 297 control points with the size of the plot character in relation to the number of occasions it was labelled

[3] http://humanimpact.geo-wiki.org/

A series of regressions were undertaken to explore the relationships between the likelihood of volunteers to correctly identify the land cover class and the geographic distance to the location being considered, calculated using the great circle distance, the level of volunteer confidence and volunteer expertise. That, is

$$P(y_i = 1) = b_0 + b_1 x_1 \dots + b_n x_n$$

where *P(y$_i$ = 1)* is the probability that the land cover class at location *i* was correctly identified, b_0 is the intercept term, $x_{1 \dots n}$ are the variables describing volunteer distance, volunteer self-reported confidence in the land cover class label and their inferred expertise, which were sequentially included in the analysis.

These analyses were then extended to consider how the relationships vary for different land cover classes.

Tab. 1: Location and expertise of the volunteers

Country	Number of volunteers
Argentina	6
Austria	8
Italy (FAO)	4
Germany	4
India	2
Italy	1
Russia	1
UK	19
USA	1
Unknown	19

Expertise	Number
Expert	29
Novice	12
Some Expertise	21
Unknown	3

3 Results

The results of the regressions, sequentially including additional terms, are summarised in Table 2. These results suggest a number of statements based on the exponentials of the model coefficients:

1) The relative odds of volunteers correctly predicting the land cover label decrease by 2% for each additional 1000km between the users location and the site under consideration;
2) The relative odds of confident volunteers correctly predicting the land cover label are around 2.01 times greater than for those who are not confident;
3) The relative odds of being correct are 1.66 time greater for volunteers who are experts than for others non-experts.

Tab. 2: The relationships between volunteer Distances, Expertise and Confidence and correctly predicting land cover

Model	Variable	Exponential of the Coefficient Estimate	2.50% CI	97.50% CI
1	Distance (km)	0.980	0.966	0.993
2	Distance (km)	0.981	0.967	0.995
	Confident	2.093	1.729	2.534
3	Distance (km)	0.980	0.966	0.993
	Confident	2.098	1.731	2.543
	Expert	1.656	1.462	1.877

*CI – Confidence Interval

Table 3 shows how these relationships vary when each land cover class is considered individually, where the number of control points for that class was greater than 50. The models that consider each land cover class in turn suggest a number of modifications to the statements above:

1) The effect of 'experiential distance' does not vary considerably between classes, although the exponential so of the coefficient estimates suggest that for *Shrub cover* the relative odds of volunteers correctly predicting the land cover label decrease by 8% for each additional 1000km between the users location and the site under consideration;
2) The relative odds of confident volunteers correctly predicting the land cover label are 1.2 times greater than for those who are not confident for *Tree cover*, similar for *Shrub cover*, 1.4 times greater for *Herbaceous vegetation / Grassland*, 2.2 times greater for *Cultivated and managed* and 1.8 times greater for mosaic classes;
3) The relative odds of being correct are 1.3 times greater for volunteers who are experts than for others non-experts for *Tree cover*, 2.1 times greater for *Shrub cover*, 4.1 times greater for *Cultivated and managed* and 1.9 times greater for *Mosaics*. The exception is the class of *Herbaceous vegetation / Grassland*, where experts were less reliable than non-experts.

Tab. 3: The relationships between volunteer Distances, Expertise and Confidence and correctly predicting land cover, broken down by class

		Distance			Expert			Confident		
class	count	Estimate	2.5% CI	97.5% CI	Estimate	2.5% CI	97.5% CI	Estimate	2.5% CI	97.5% CI
1	975	0.986	0.952	1.021	1.231	0.932	1.623	1.354	0.906	2.014
2	400	0.923	0.875	0.973	1.044	0.681	1.599	2.186	1.188	4.136
3	464	1.022	0.979	1.067	1.382	0.923	2.079	0.718	0.464	1.108
4	1799	0.982	0.958	1.007	2.214	1.747	2.803	4.140	2.408	7.171
5	1412	1.010	0.982	1.038	1.791	1.415	2.269	1.920	1.348	2.747

Classes: 1 Tree cover; 2 Shrub cover; 3 Herbaceous vegetation / Grassland; 4 Cultivated and managed; 5 Mosaic: cultivated and managed / natural vegetation

4 Discussion

This paper extends the research described in See et al (2013) by considering the impact of distance between the volunteer's location and the location being analysed. It shows that distance has a minor effect on the reliability of the labelling performed by volunteers, and that expertise matters generally but not for all classes such as *Shrub Cover*. One of the key issues raised by this work relates to the use of 'experiential distance' and geographic distance. The assumption is that people have more experience of nearer places than of far away ones. This is patently not the case as people may regularly visit certain places on holiday. Future work will apply these confidences to the ~53,000 locations that were labeled by the volunteers in order to generate surfaces of reliability, and will consider other data collected as part of the Human Impact Geo-Wiki campaign including measures of human impact and measures of land abandonment, both parameterized with volunteer self-assessment of confidence. It will also consider the structure of mental maps and of volunteer cognitive experiences.

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.

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.

FOODY, G. M., SEE, L., FRITZ, S., VAN DER VELDE, M., PERGER, C., SCHILL, C., BOYD, D. S. & COMBER, A. (in press), Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality. The Cartographic Journal (October 2013). DOI: http://dx.doi.org/10.1179/1743277413Y.0000000070.

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.

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

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: Geovizualisation, Society and Learning. Berlin/Offenbach, Wichmann, 83-91.

SEE, L., COMBER, A. J., SALK, C., FRITZ, S., VAN DER VELDE, M., PERGER, C., SCHILL, C., MCCALLUM, I., KRAXNER, F. & OBERSTEINER M. (2013), Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. PLoS ONE, 8 (7): e69958. DOI:10.1371/journal.pone.0069958.