Research Article *Transactions in GIS*, 2013, 17(6): 847–860

Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project

G. M. Foody,* L. See,† S. Fritz,† M. Van der Velde,† C. Perger,‡ C. Schill§ and D. S. Boyd*

*School of Geography, University of Nottingham
†International Institute of Applied Systems Analysis (IIASA), Austria
‡International Institute of Applied Systems Analysis (IIASA), University of Applied Sciences, Wiener Neustadt
§University of Freiburg

Abstract

The recent rise of neogeography and citizen sensing has increased the opportunities for the use of crowdsourcing as a means to acquire data to support geographical research. The value of the resulting volunteered geographic information is, however, often limited by concerns associated with its quality and the degree to which the contributing data sources may be trusted. Here, information on the quality of sources of volunteered geographic information was derived using a latent class analysis. The volunteered information was on land cover interpreted visually from satellite sensor images and the main focus was on the labeling of 299 sites by seven of the 65 volunteers who contributed to an Internet-based collaborative project. Using the information on land cover acquired by the multiple volunteers it was shown that the relative, but not absolute, quality of the data from different volunteers could be characterized accurately. Additionally, class-specific variations in the quality of the information provided by a single volunteer could be characterized by the analysis. The latent class analysis, therefore, was able to provide information on the quality of information provided on an inter- and intra-volunteer basis.

1 Introduction

Crowdsourcing has become a popular means of acquiring information in a wide variety of subject areas from astronomy (Raddick and Szalay 2010) to zoology (Dickinson et al. 2010, Wiersma 2010). The power of the crowd is well known and it can sometimes be used to derive information that was impossible or at least impractical to obtain by other means. The potential of crowdsourcing has been noted in geography and associated geospatial sciences with numerous examples of its use (Basiouka and Potsiou 2012, Davis and de Alencar 2011, Girres and Touya 2010, Goodchild 2007, Heipke 2010, Neis et al. 2012). There are, of course, major limitations to crowdsourced data, not least those connected with their quality and the level of trust that may be placed upon them (Flanagin and Metzger 2008, Hudson-Smith et al.

Address for correspondence: Giles M. Foody, School of Geography, University of Nottingham, Nottingham, NG7 2RD, UK. E-mail: giles.foody@nottingham.ac.uk

Acknowledgements: We are extremely grateful to all who kindly contributed to the project, notably the volunteers who provided the class label information used. The work reported in this article benefits in part from funding to GMF from the British Academy (reference SG112788) and EPSRC (reference EP/J0020230/1). Additionally, the research was partly supported by the Austrian Research Promotion Agency (FFG) via the Land Spotting project (No. 828332). Finally, we are grateful to the editor and three referees for their constructive comments on the original manuscript.

 doi: 10.1111/tgis.12033

2009, Goodchild and Glennon 2010, Haklay et al. 2010, Wiersma 2010, Doan et al. 2011). The accuracy of crowdsourced data is explored here in relation to its potential role in the validation of land cover maps, especially those derived from satellite remote sensing, although the methods are of a more general nature and of broad applicability.

Land cover is one of the most important environmental variables. Land cover exerts considerable influence over a range of environmental and human systems. For example, land cover change is both a cause and a consequence of climate change, is the greatest threat to biodiversity, and, among other things, can cause soil erosion, flooding and altered local climates, all of which can also greatly impact human health and well-being (Briassoulis 2003, Feddema et al. 2005, Freitas 2006). Land cover and land cover change, therefore, are key variables, central to major research priorities such as the strategic research directions for the geographical sciences identified by the US National Academies (CSDGSND 2010). That being the case, there is clearly a need for accurate and up-to-date land cover information.

Fortunately, satellite remote sensing has the potential to provide spatially and temporally detailed information on the Earth's surface to inform land cover map production. Indeed, land cover maps are now widely produced at a range of spatial and temporal scales via remote sensing. However, for a map to have value it must be accurate and hence the validation of land cover maps is a major issue. While 'best practices' for map validation exist (Strahler et al. 2006) the methods proposed are typically based upon established design-based inferential methods using rigorous probability sampling which may often be impractical to implement fully. Many maps, therefore, are either validated not at all or only sub-optimally (Foody 2002, Olofsson et al. 2013); but it is important that a map be validated otherwise it is no more than one, untested, hypothesized scenario of the land cover (Strahler et al. 2006).

There is great potential for crowdsourcing to contribute constructively to land cover map validation activities (Iwao et al. 2006, Fritz et al. 2012). Indeed, contemporary mapping activity benefits greatly from recent advances in geoinformation technologies beyond remote sensing, notably through citizen sensors fostered by the proliferation of inexpensive and highly mobile location-aware devices able to provide spatial data of value to the mapping community. Mapping, therefore, can draw upon not only the work of authoritative agencies but also the broader, amateur, community, often through contributions to Internet hosted collaborative activities (e.g. Iwao et al. 2006). A key problem with the volunteered geographic information (VGI) provided by the amateur community is that it can be of variable, and typically unknown, quality. The volunteers providing the data may vary greatly, from enthusiastic but naïve and untrained hobbyists to highly skilled professionals of considerable expertise. While the latter group of people may often volunteer (Brabham 2012) the quality of the data provided may still be a concern given the known concerns with expert labelers (Foody 2002, 2010). There is also a danger that VGI may sometimes be corrupted by errors deliberately introduced with malicious intent. It is often valuable, therefore, to have data provided by multiple observers, especially as this enables a consensus-based approach to be used in labeling cases (Haklay et al. 2010, Tang and Lease 2011). Even with such approaches there are concerns about the variation in quality of data acquired by different contributors as well as variations within the set of data provided by a single contributor. If VGI are to be used it is important that information on the quality of the data and of the data sources be made available.

The aim of this article is to explore the potential to estimate the accuracy of VGI. Specifically, the aim is to characterize the accuracy of each volunteer or data source, such that the relative accuracy of each can be determined to help potential users of the VGI evaluate its suitability for the intended application. Attention is focused on VGI collected as part of a major

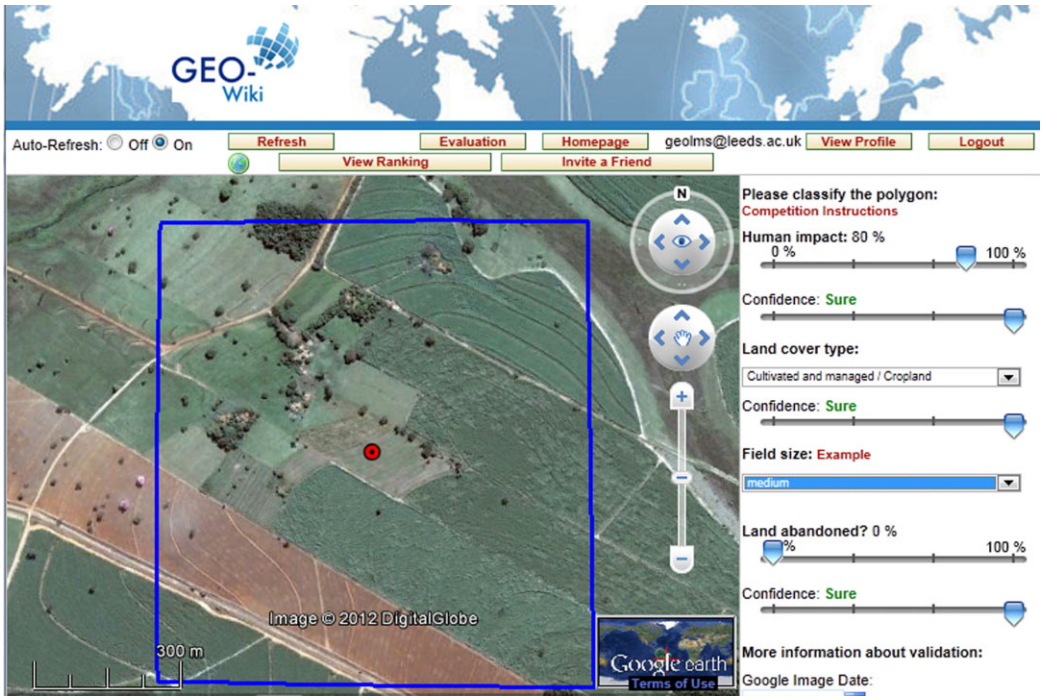

Figure 1 An example of the Geo-Wiki interface used by the volunteers to provide land cover (and other) information for the highlighted area. The volunteer is invited to label the land cover of the region contained by the blue box, the centre of which is indicated by the red dot

international open call for data collection via Geo-Wiki (Fritz et al. 2012) to aid the validation of satellite sensor derived land cover products. Hence the work presented in this article lies at the interface of remote- and citizen-sensing.

2 Data

The Geo-Wiki project (http://www.geo-wiki.org) was initiated to encourage a global network of volunteers to help improve the quality of global land cover maps through crowd-sourcing (Fritz et al. 2012). In brief, an open call for contributions to the project was made in September 2011 and closed in November of the same year. The project invited each volunteer to look at a series of satellite sensor images and assign a land cover label from a defined list of classes to each highlighted region as shown in Figure 1; the volunteer is asked to label the area contained within the blue box. More details of the competition can be found in Perger et al. (2012). The legend comprised 10 classes: tree cover, shrub cover, herbaceous vegetation/grassland; cultivated and managed, mosaic: cultivated and managed/natural vegetation, regularly flooded/wetland, urban/built-up, snow and ice, barren, and open water. A brief on-line tutorial was available to aid the volunteers through the labeling process and there were no constraints on involvement. Hence, the volunteers could vary greatly in their ability to accurately label the land cover from the images presented to them. The volunteers could vary from enthusiastic amateurs to experts in image interpretation and little or no information about the

nature of the volunteers was available to indicate possible data quality and trust levels. A total of 65 volunteers contributed to the project and each was presented with up to 299 images for labeling. The same set of images was used throughout and so the cases contained in the derived data set were labeled multiple times. Three volunteers labeled all 299 cases and 18 labeled at least 200 cases while others labeled fewer, with one volunteer labeling just a single case. Thus, not only was the set of class labels derived by the volunteers expected to be of unknown and uncertain quality, it was also often incomplete.

A subset of the available data was used here. Specifically, the data were extracted from the 10 volunteers who provided the fullest set of labels. This data set included the set of labels derived from three experts who also revisited the entire set of 299 images to derive a ground reference data set. The latter is essentially a set of labels defined by consensus amongst the experts, similar to that used in the production of reference data in major mapping programs (e.g. Scepan et al. 1999, Bicheron et al. 2008). The production of the reference data was undertaken by the experts as a group and after each had undertaken the labeling task individually. Although derived in a normal way and based on expert opinion, these ground reference data are still likely to be imperfect as the experts are not infallible. The reference data, therefore, do not represent the perfect gold standard reference that should ideally be used in accuracy assessment (Foody 2010) but form what is believed to be a very high quality reference. To maintain a focus on VGI from non-experts, the data acquired by the three experts were excluded from further analysis, leaving the labels provided by seven volunteers, labeled A-G, for evaluation. Critically, however, these data were acquired independently of the production of the ground reference data. The seven volunteers each labeled at least 289 of the 299 cases, reducing the potential for complications caused by missing data, which is a topic of current research. The location of the 299 cases and the seven volunteers whose labels are used in the research are illustrated in Figure 2.

3 Methods

The degree of agreement between the set of labels derived by a volunteer with those from each other volunteer was assessed and measured using the kappa coefficient of agreement (Cohen

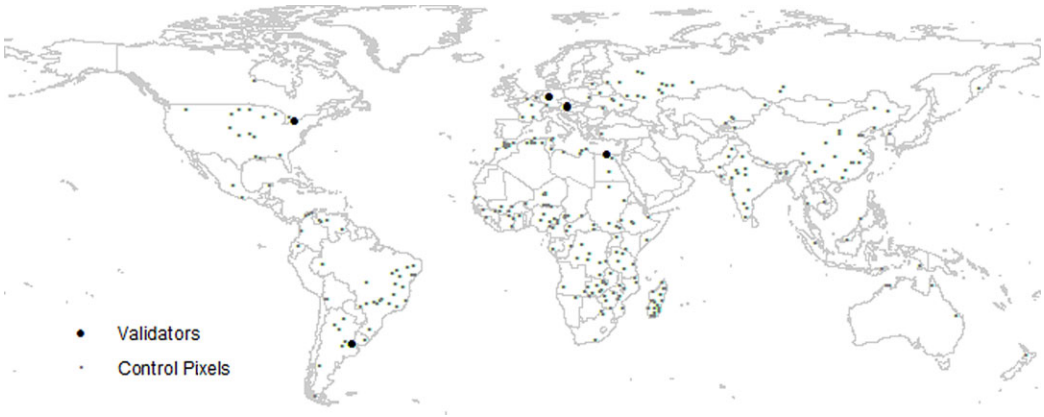

Figure 2 The location of the 299 cases (control pixels) and the seven volunteers (validators) whose contributions are used in the research; note two volunteers were located in Buenos Aires, Argentina and two are located very close together in Austria

1960). The latter is a popular measure of inter-rater agreement and indicates the level of agreement between a pair of classifications, specifically the level of agreement above that which might be expected by chance; cases with missing values were excluded from the calculations. Kappa was calculated from:

$$\hat{\kappa} = \frac{p_o - p_c}{1 - p_c} \tag{1}$$

where p_o represents the observed proportion of agreement and p_c the proportion of agreement expected to occur by chance. Although the kappa coefficient is not really suited to accuracy assessment (Foody 1992, 2008, Pontius and Millones 2011) it was also calculated for each volunteer relative to the ground reference data to provide an indicator of agreement to the reference labels.

The accuracy of each volunteer's labeling was also explored using a latent class analysis. This type of analysis has been widely used to derive information on the accuracy of classification in the absence of a gold standard reference data set (Hui and Zhou 1998). In essence, the latent class analysis uses the observed associations between the set of labels derived by the volunteers as imperfect indicators of the unobserved (latent) classes of a latent variable that equate, in this study, to the actual land cover. Here, it was assumed that the true land cover class was a latent variable T, indexed t, and that the images were classified into an exhaustively defined set of q mutually exclusive classes.

A standard latent class model is based on the probability of observing the patterns of class allocation made by the set of volunteers. The set of labels provided by the volunteers represent the observed or manifest variables of the analysis and are used to provide information on the unobserved (latent) variable. The set of volunteers, V, each label the images, yielding a set of predicted class labels for each case. Here, M_v represents one of the set of V manifest variables indexed $1 \leq v \leq V$, and its values are class labels represented by m_v which lie in the range r (1-q). Using vector notation $\mathbf{M}$ and $\mathbf{m}$ to represent the complete response patterns (i.e. $\mathbf{M}$ denotes (M_1, . . ., M_V) and $\mathbf{m}$ denotes (m_1, . . . m_q)), the latent class model is that the probability of obtaining the response pattern $\mathbf{m}$, represented as $P(\mathbf{M} = \mathbf{m})$, is a weighted average of the q class-specific probabilities $P(\mathbf{M} = \mathbf{m}|T = t)$ (Magidson and Vermunt 2004). On the assumption that the labels derived from each volunteer are conditionally independent of those from all other volunteers, the latent class model may be written as:

$$P(\mathbf{M} = \mathbf{m}) = \sum_{t=1}^{q} P(T = t) \prod_{v=1}^{V} P(M_v = m_v | T = t) \tag{2}$$

where $P(T = t)$ is the proportion of cases belonging to latent class t (Yang and Becker 1997, Vermunt 1997, Vermunt and Magidson 2003); if necessary the model can be adjusted to allow for violation of the conditional independence assumption. The fit of this model to the data is typically evaluated with regard to the likelihood ratio chi-squared statistic, L^2. The latter compares the observed frequencies in the multi-dimensional contingency matrix that illustrates the pattern of class allocation made by the volunteers with those expected from the latent class model. A perfect fit occurs if $L^2 = 0$ and a model is normally viewed as fitting the data if the calculated value of L^2 is sufficiently small to be attributable to the effect of chance (Magidson and Vermunt 2004). Conventionally, therefore, a model would be viewed as providing a good fit to the data if the probability of observing the derived L^2

statistic by chance was high (i.e. $p > 0.05$). Further details on latent class modeling may be found in Rindskopf and Rindskopf (1986), Yang and Becker (1997), Vermunt and Magidson (2003), and Magidson and Vermunt (2004).

A critical feature of the latent class model defined in Equation (2) is that its only parameters are the latent class probabilities. Assuming that the latent classes are identifiable and have been labeled to match the actual land cover class labels, the model's parameters may be used to indicate the accuracy of each volunteer's contributions. For example, the model parameter $P(T = t)$ represents the probability of a case being a member of latent class t which, for a simple random sample, directly reflects the proportion of that class in the data set which can be valuable in support of class area estimation applications and non-site specific accuracy assessment (Foody 2012, Foody and Boyd 2012, 2013). Additionally, the model parameters representing the conditional probabilities include measures that equate to the popular site-specific measure of producer's accuracy (Foody, 2010, 2012, Yang and Becker 1997). Note that the latter is, more formally, described as the conditional probability that a pixel classified as belonging to a class is also labeled as a member of that class in the ground reference data (Stehman 1997, Liu et al. 2009). Thus, when $m_v = t$ the expression $P(M_v = m_v|T = t)$ represents the producer's accuracy for class t in the volunteered data represented by M_v. Thus, the parameters of a suitable latent class model describe key measures of the accuracy of the volunteer data sources and can even be derived in the absence of reference data (Foody 2012). Additional measures of accuracy may be derived from the latent class analysis but are not considered in this work.

The latent class model may also be used to derive a new a classification for each case. This classification is derived following Bayes theorem with each case allocated to the class with which it displays the largest posterior probability of membership (Vermunt and Magidson 2003, Magidson and Vermunt 2004) using:

$$P(T = t|\mathbf{M} = \mathbf{m}) = \frac{P(T = t)P(\mathbf{M} = \mathbf{m}|T = t)}{P(\mathbf{M} = \mathbf{m})}. \quad (3)$$

The basic latent class modeling approach was adapted here to allow for missing cases. This required the identification of volunteers who provided an incomplete set of labels. It was evident from the contributions received that there were unlabelled cases in the data provided by volunteers C-G. Additionally, it was observed that a pair of volunteers sometimes failed to label a case. This latter situation occurred with volunteers C and F, D and F, E and F, and F and G. As the full pattern of class labels across all seven volunteers cannot be formed for the cases with one or more missing labels, sub-models that focused on only the possible set of patterns given the missing cases were defined to enable the analysis to be undertaken with the LEM software (Vermunt 1997). The latter is a freely available and widely used software system that allows the modeling of latent variables (the software and further details can be downloaded from http://www.tilburguniversity.edu/nl/over-tilburg-university/schools/socialsciences/organisatie/departementen/mto/onderzoek/software/).

The central issue of relevance to this article is that the parameters of a latent class model that fits the observed pattern of labeling arising from the set of volunteers may be used to indicate the accuracy of the data contributed by each volunteer. To allow a critical evaluation of the estimates of volunteer accuracy derived from the parameters of the fitted latent class model, the results were compared against accuracy values derived in the conventional manner relative to the high quality ground reference data set generated by three experts. The reference data were also used to help label the latent classes.

Thus the accuracy of the labeling by each of the seven selected volunteers was assessed relative to the ground reference data via a cross-tabulation of the volunteers' labels with those depicted in the ground reference data. The accuracy of labeling was quantitatively described by the producer's accuracy estimated from the cross-tabulation matrix and compared against the corresponding estimate from the latent class model. The mean producer's accuracy calculated over all classes was also used as a simple guide to overall classification accuracy.

4 Results and Discussion

Although the thematic resolution of the land cover classification used contained 10 classes, only eight were actually identified within the set of images used; the regularly flooded/wetland and the snow and ice classes were absent. Additionally, three of the classes were rare, occurring, for example, just once or twice within the ground reference data set. These rare classes were the urban/built-up, barren, and open water. Given the very low number of cases of these classes, the absolute magnitude of the producer's accuracy for each was very sensitive to minor variations in the number of cases predicted by each volunteer. The other five classes were much more abundant, with, for example, 20–119 cases recorded for each in the ground reference data set.

The 299 cases varied greatly in appearance and labeling challenge. For some, the label was relatively obvious and all seven volunteers provided the same, correct label (Figure 3). For other cases there was considerable uncertainty and the volunteers did not agree on the labeling (Figure 4).

Figure 3 Two cases upon which all seven volunteers agreed on the class label: (a) urban/built-up; and (b) cultivated and managed

Figure 4 Two cases upon which the volunteers did not unanimously agree on labelling: (a) a site labelled as mosaic by the experts but which was labelled as tree cover by three volunteers, mosaic by three volunteers and shrub by one volunteer; and (b) a site labelled as mosaic by the experts but which was labelled as mosaic by four volunteers, herbaceous by two volunteers and barren by one volunteer

Table 1 Kappa coefficient (sample size) for evaluation of inter-rater agreement

	B	C	D	E	F	G	Reference
A	0.535 (299)	0.557 (294)	0.397 (292)	0.391 (292)	0.321 (290)	0.434 (289)	0.564 (299)
B		0.565 (294)	0.497 (292)	0.455 (292)	0.273 (290)	0.423 (289)	0.603 (299)
C			0.492 (287)	0.481 (287)	0.238 (286)	0.547 (284)	0.663 (294)
D				0.412 (285)	0.253 (284)	0.387 (282)	0.536 (292)
E					0.267 (284)	0.473 (282)	0.503 (292)
F						0.227 (281)	0.227 (290)
G							0.503 (289)

There was a relatively low level of agreement between the labels derived by the set of volunteers, with the value of the kappa coefficient of agreement varying from 0.227 to 0.565 (Table 1). In addition the value of the kappa coefficient derived for evaluations of volunteer labels against the ground reference data labels were low, ranging from 0.227 to 0.663. These results indicate low levels of agreement between the different sets of class labels derived. With no means to distinguish between the volunteers in terms of relative accuracy and with low levels of agreement to the reference data, the results provide little to suggest, for example, that information from a single volunteer may have an immensely useful role in aiding the validation of land cover maps. The potential to enhance the value of the VGI by utilizing the labels provided by the multiple contributors together with a latent class analysis was evaluated.

The latent class analysis used the inter-relationships between the volunteered data to uncover the hidden (latent) information on land cover. A latent class model informed by the

Table 2 Estimate of producer's accuracy (%) for volunteers A-G derived from the model

Class	Volunteer						
	A	B	C	D	E	F	G
Tree cover	100.00	86.27	74.73	62.60	73.23	67.43	66.51
Shrub cover	64.44	74.54	83.47	71.13	50.81	69.65	60.61
Herbaceous vegetation / Grassland	69.54	71.22	73.27	45.03	64.65	47.52	24.79
Cultivated and managed	94.16	92.66	100.00	70.31	87.14	20.17	91.82
Mosaic: cultivated and managed / natural vegetation	54.87	73.80	95.34	97.75	67.90	64.74	67.50
Regularly flooded / wetland	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Urban / built-up	50.00	25.00	50.00	50.00	50.00	50.00	25.00
Snow and ice	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Barren	38.70	0.00	11.99	0.00	50.90	30.25	0.00
Open water	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Mean	49.67	44.85	51.38	42.18	46.96	37.48	36.12

labels from the seven selected volunteers was constructed. This model had an $L^2 = 1341.4$ ($p >> 0.05$), suggesting a good fit to the observed data. The probabilities of class membership derived from the model were used to allocate each case to one of the latent classes. Cross-tabulation of the derived labels with the ground reference data provided a means to label the latent classes.

The parameters from the latent class model that equate to the producer's accuracy of each volunteer were derived for each class. These model-based estimates of producer's accuracy provided estimates of the quality of the labeling by each volunteer for each class and, when averaged, overall (Table 2). The model based estimates were compared against the estimates derived from use of the ground reference data set (Table 3). It was apparent that the model based estimates for the five most abundant classes were different in absolute terms, often considerably over-estimating accuracy. However, closer inspection shows that there were strong relationships between the estimated and actual producer's accuracies. This was apparent on both a class-specific (Figure 5) and overall basis (Figure 6). It was also evident that there was a tendency for the class-specific accuracies to be over-estimated for the five most abundant classes while the overall accuracies, influenced by poor estimation in relation to the relatively rare classes (Tables 2 and 3), was under-estimated.

Focusing on the five land cover classes for which a relatively large sample was available and hence the producer's accuracy not overly influenced by a small number of cases, it was evident that there was typically a strong positive relationship between the estimated and actual accuracy values. Additionally, the model-based estimates typically indicated correctly the relative performance of a volunteer with regard to the accuracy with which the different classes were identified. For example, volunteer A provided the most accurate classification of the tree cover class (Figure 5a) but also the least accurate classification of the mosaic class (Figure 5e). Volunteer D, however, provided extremely accurate labeling of the mosaic class (Figure 5e) but was much less accurate with the herbaceous vegetation/grassland (Figure 5c). Additionally, it was evident that volunteer B was, in terms of accuracy, always a good labeler

Table 3 Actual producer's accuracy (%) for volunteers A-G, determined relative to the ground reference data set

	Volunteer						
Class	A	B	C	D	E	F	G
Tree cover	78.72	72.34	63.83	61.70	60.87	48.94	45.65
Shrub cover	55.00	65.00	70.00	65.00	45.00	65.00	70.00
Herbaceous vegetation / Grassland	62.50	62.50	58.33	37.50	65.22	41.67	41.67
Cultivated and managed	87.39	73.95	88.03	52.54	74.56	14.16	84.48
Mosaic: cultivated and managed / natural vegetation	36.47	67.06	68.29	94.94	49.41	51.22	45.57
Regularly flooded / wetland	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Urban / built-up	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Snow and ice	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Barren	100.00	50.00	100.00	50.00	100.00	100.00	50.00
Open water	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Mean	62.01	59.08	64.85	56.17	59.51	52.10	53.74

but never the best. Critically, the accuracy with which a land cover class was identified varied greatly between the volunteers. Each volunteer also typically varied in ability to label the different classes; and classes accurately labeled by one volunteer might be inaccurately identified by another. These inter- and intra-volunteer variations in labeling accuracy may reflect variations in background education, skills and experience as well as issues connected with the acquisition of the volunteer's labels via the website. It is evident that each volunteer typically found some aspects of the labeling task more difficult than others and so the accuracy with which each labeled a class varied. It was also evident that this type of information may help users, for example, to select VGI for an application or perhaps to help identify which volunteers to approach for a future study and possibly those who might benefit most from training. Critically, the approach appears able to indicate the relative performance of each volunteer in terms of the accuracy of labeling and the performance of a single volunteer with regard to each class.

The strong positive relationship between the estimated and actual producer's accuracy was also evident in relation to the overall accuracy, expressed here as the mean producer's accuracy calculated over all classes (Figure 6). Although the sample size was small and the data points in Figure 6 lie away from the 1:1 line there is a clear trend that suggests an ability to characterize the relative accuracy of each contributor. Users of VGI may find this useful in sifting large VGI data sets to allow, for example, inaccurate data to be discarded to allow a focus on the higher quality data and data sources. It should be noted, however, that the approach presented here is applicable to situations in which multiple contributors are used in a manner that ensures a set of labels are available for each case. The approach would not be useful in situations in which only the results of one volunteer (e.g. the most recent editor of an online system) were presented. In such situations it may be desirable to make the labels from other contributors available as an aid to assessing the accuracy of the data provided.

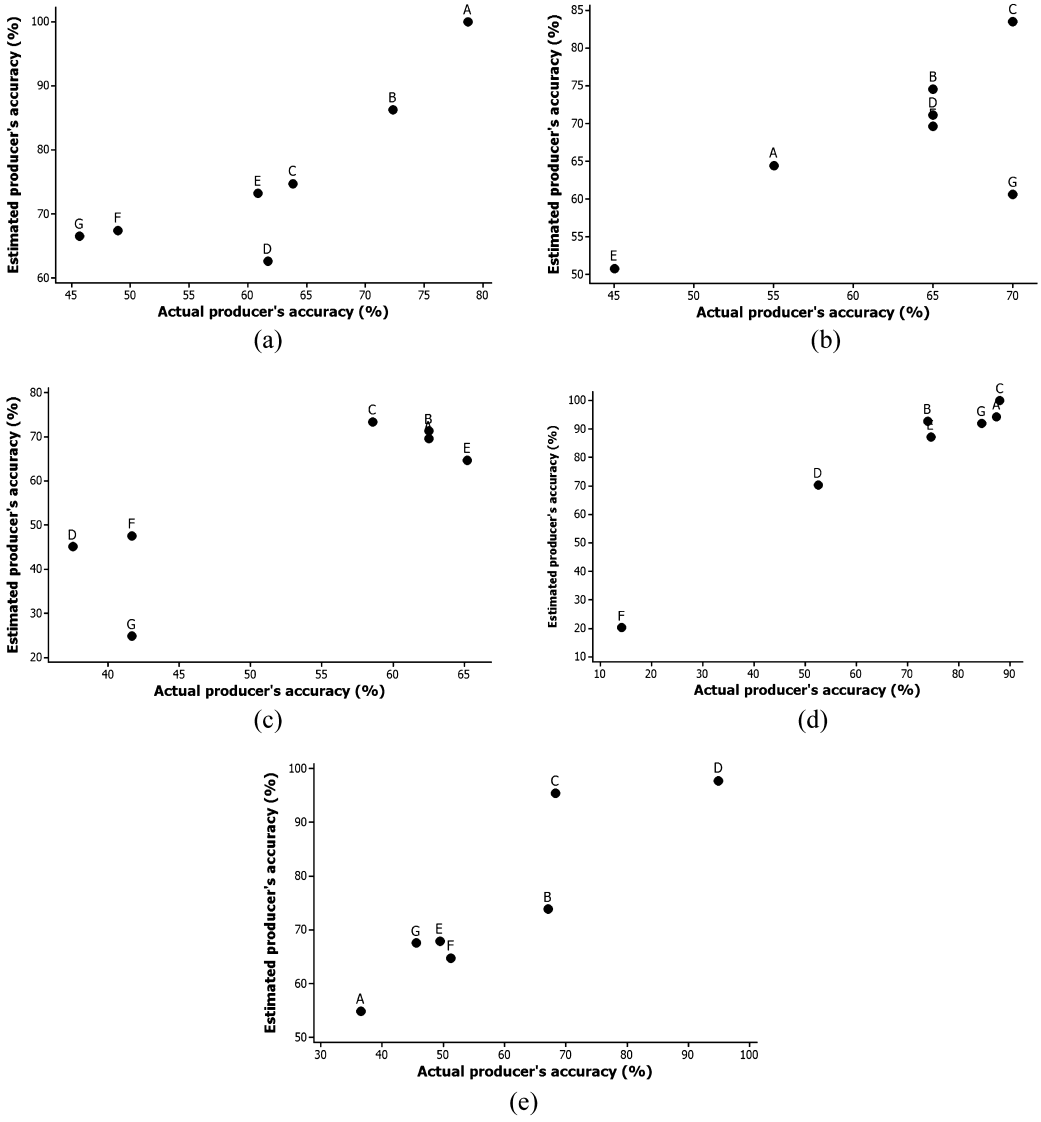

Figure 5 Relationships between the estimated and actual producer's accuracy for the five most abundant land cover classes: (a) tree cover; (b) shrub cover; (c) herbaceous vegetation/grassland; (d) cultivated and managed; and (e) mosaic: cultivated and managed/natural vegetation. Note that the range of both x and y axis scales varies between the graphs shown

5 Conclusions

An open call for contributions to a land cover class labeling exercise was undertaken. A set of 65 volunteers generously contributed to the project. Attention here was focused on the data provided by seven of the volunteers who labeled nearly all of the 299 sample cases provided for labeling plus a reference data set derived from three expert labelers who contributed to the exercise. The labeling task varied greatly, with some cases relatively straightforward to label

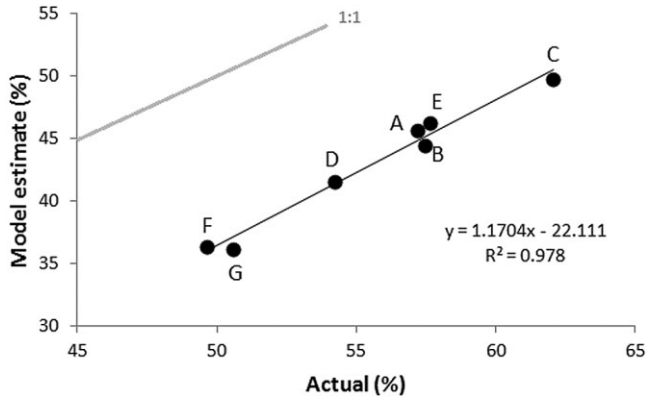

Figure 6 Relationship between estimated and actual mean producer's accuracy determined over all classes for each volunteer contributing data to the analysis. A standard regression line (and equation) is shown to help illustrate the relationship between the estimated and actual accuracy values and the 1:1 line is shown in grey

while others were more ambiguous and challenging to label. Each volunteer provided a set of labels that was only poorly related to the reference data (kappa <0.67) but the data from multiple volunteers could be constructively combined in the latent class analysis to yield useful information on the quality of the VGI. The key finding is that the latent class model provided a means to correctly characterize the relative performance of each volunteer in terms of class-specific and overall classification accuracy.

The main conclusions of the work reported here are:

- The latent class modeling approach provided a means to directly estimate the producer's accuracy of the volunteer data sources without reference data.
- The model based estimates of producer's accuracy were typically incorrect in absolute terms, with the accuracy often underestimated on an overall basis or overestimated for the abundant classes. However, in relative terms the model-derived estimates typically characterized the relative performance of the volunteers correctly in terms of the accuracy with which they labeled the cases. The latter was evident in relation to both class-specific and overall accuracy.
- The model provided a means to evaluate the accuracy of volunteered data both between and within data sources. For example, the model allowed the relative performance of volunteers to be assessed in terms of the accuracy of their labeling. It also allowed the performance of a single volunteer in labeling the different classes to be assessed.

Critically, the latent class modeling approach applied allowed a means to characterize key aspects of the quality of VGI. There is potential to estimate additional measures of accuracy (e.g. user's accuracy) or to enhance the modeling (e.g. possibility of additional latent classes and variables or use of additional contributors) and future work is exploring some of the possibilities to further enhance the value of VGI.

References

Basiouka S and Potsiou C 2012 VGI in Cadastre: A Greek experiment to investigate the potential of crowdsourcing techniques in cadastral mapping. *Survey Review* 44: 153–161

Bicheron P, Defourny P, Brockmann C, Schouten L, Vancustem C, Huc M, Bontemps S, Leroy M, Achard F, Herold M, Ranera F, and Arino, O 2008 *Globcover: Products Description and Validation Report*. Toulouse, France, Medias

Brabham D C 2012 The myth of amateur crowds: A critical discourse analysis of crowdsourcing coverage. *Information Communications and Society* 15: 394–410

Briassoulis H 2003 Land use changes and global aggregate impacts. *EOLSS (Online UNESCO Encyclopedia of Life Support Systems)* (available at http://www.eolss.net/)

Cohen J 1960 A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46

CSDGSND (Committee on Strategic Directions for the Geographical Sciences in the Next Decade) 2010 *Understanding the Changing Planet: Strategic Directions for the Geographical Sciences*. Washington, D.C., National Academies Press

Davis Jr C A and de Alencar R O 2011 Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city. *Transactions in GIS* 15: 851–68

Dickinson J L, Zuckerberg B, and Bonter D N 2010 Citizen science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution and Systematics* 41: 149–72

Doan A, Ramakrishnan R, and Halevy A Y 2011 Crowdsourcing systems on the World-Wide Web. *Communications of the ACM* 54: 86–96

Feddema J J, Oleson K W, Bonan G B, Mearns L O, Buja L E, Meehl G A, and Washington W M 2005 The importance of land-cover change in simulating future climates. *Science* 310: 1674–78

Flanagin A J and Metzger M J 2008 The credibility of volunteered geographic information. *GeoJournal* 72: 137–48

Freitas H 2006 Land-use/land-cover changes and biodiversity loss. *EOLSS (Online UNESCO Encyclopedia of Life Support Systems)* (available at http://www.eolss.net/)

Foody G M 1992 On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing* 58: 1459–60

Foody G M 2002 Status of land cover classification accuracy assessment. *Remote Sensing of Environment* 80: 185–201

Foody G M 2008 Harshness in image classification accuracy assessment. *International Journal of Remote Sensing* 29: 3137–58

Foody G M 2010 Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment* 114: 2271–85

Foody G M 2012 Latent class modelling for site and non-site specific classification accuracy assessment without ground data, *IEEE Transactions on Geoscience and Remote Sensing* 50: 2827–38

Foody G M and Boyd D S 2012 Exploring the potential role of volunteer citizen sensors in land cover map accuracy assessment. In *Proceedings of the Tenth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2012)*, Florianopolis, Brazil: 203–8

Foody G M and Boyd D S 2013 Using volunteered data in land cover map validation: Mapping West African forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6: in press

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: 110–23

Girres J-F and Touya G 2010 Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS* 14: 435–59

Goodchild M F 2007 Citizens as sensors: The world of volunteered geography. *GeoJournal* 69: 211–21

Goodchild M F and Glennon J A 2010 Crowdsourcing geographic information for disaster response: A research frontier. *International Journal of Digital Earth* 3: 231–41

Haklay M, Basiouka S, Antoniou V, and 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: 315–22

Heipke C 2010 Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing* 65: 550–57

Hudson-Smith A, Batty M, Crooks A, and Milton R 2009 Mapping for the masses: Accessing Web 2.0 through crowdsourcing. *Social Science Computer Review* 27: 524–38

Hui S L and Zhou X H 1998 Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* 7: 354–70

Iwao K, Nishida K, Kinoshita T, and Yamagata Y 2006 Validating land cover maps with Degree Confluence Project information. *Geophysical Research Letters* 33: L23404

Liu C, White M, and Newell G 2009 Measuring the accuracy of species distribution models: A review. In *Proceedings of the Eighteenth World IMACs/MODSIM Congress*, Cairns, Australia: 4241–47

Magidson J and Vermunt J K 2004 Latent class models. In Kaplan D (ed) *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA, Sage Publications: 175–98

Neis P, Zielstra D, and Zipf A 2012 The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet* 4: 1–21

Olofsson P, Foody G M, Stehman S V, and Woodcock C E 2013 Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation, *Remote Sensing of Environment* 128: in press

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 *Proceedings of the GI Forum Conference 2012*, Salzburg, Austria

Pontius R G and Millones M 2011 Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing* 32: 4407–29

Raddick M J and Szalay A S 2010 The universe online. *Science* 329: 1028–29

Rindskopf D and Rindskopf W 1986 The value of latent class analysis in medical diagnosis. *Statistics in Medicine* 5: 21–7

Scepan J, Menz G, and Hansen M C 1999 The DISCover validation image interpretation process. *Photogrammetric Engineering and Remote Sensing* 65: 1075–81

Stehman S V 1997 Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment* 62: 77–89

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*. Ispra, Italy, European Commission, Joint Research Centre

Tang W and Lease M 2011 Semi-supervised consensus labeling for crowdsourcing. In *Proceedings of the SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval*, Beijing, China

Vermunt J K 1997 *Log-linear Models for Event Histories*. Thousand Oaks, CA, Sage Publications

Vermunt J K and Magidson J 2003 Latent class analysis. In Lewis-Beck M, Bryman A E, and Liao T F (eds) *The Sage Encyclopedia of Social Science Research Methods* (Volume 2). Thousand Oaks, CA, Sage Publications: 549–53

Wiersma Y F 2010 Birding 2.0: Citizen science and effective monitoring in the web 2.0 world. *Avian Conservation and Ecology* 5: 13

Yang I and Becker M P 1997 Latent variable modelling of diagnostic accuracy. *Biometrics* 53: 948–58

