The Cartographic Journal *Vol. 52 No. 4 pp. 336–344 November 2015*

REFEREED PAPER

Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality

G.M. Foody[1], L. See[2], S. Fritz[2], M. van der Velde[2], C. Perger[2,3], C. Schill[4], D.S. Boyd[1] and A. Comber[5]

[1] School of Geography, University of Nottingham, Nottingham NG7 2RD, UK
[2] International Institute of Applied Systems Analysis (IIASA), Laxenburg, Austria
[3] University of Applied Sciences, Wiener Neustadt, Austria
[4] University of Freiburg, Freiburg, Germany
[5] Department of Geography, University of Leicester, Leicester LE1 7RH, UK
Email: giles.foody@nottingham.ac.uk

Crowdsourcing is a popular means of acquiring data, but the use of such data is limited by concerns with its quality. This is evident within cartography and geographical sciences more generally, with the quality of volunteered geographic information (VGI) recognized as a major challenge to address if the full potential of citizen sensing in mapping applications is to be realized. Here, a means to characterize the quality of volunteers, based only on the data they contribute, was used to explore issues connected with the quantity and quality of volunteers for attribute mapping. The focus was on data in the form of annotations or class labels provided by volunteers who visually interpreted an attribute, land cover, from a series of satellite sensor images. A latent class model was found to be able to provide accurate characterizations of the quality of volunteers in terms of the accuracy of their labelling, irrespective of the number of cases that they labelled. The accuracy with which a volunteer could be characterized tended to increase with the number of volunteers contributing but was typically good at all but small numbers of volunteers. Moreover, the ability to characterize volunteers in terms of the quality of their labelling could be used constructively. For example, volunteers could be ranked in terms of quality which could then be used to select a sub-set as input to a subsequent mapping task. This was particularly important as an identified subset of volunteers could undertake a task more accurately than when part of a larger group of volunteers. The results highlight that both the quantity and quality of volunteers need consideration and that the use of VGI may be enhanced through information on the quality of the volunteers derived entirely from the data provided without any additional information.

Keywords: neogeography, thematic mapping, accuracy, citizen science

INTRODUCTION

Recent advances in geoinformation technologies have contributed to the emergence and rapid acceleration of citizen sensing (Goodchild, 2007; Sui *et al.*, 2013). This has had many impacts on a diverse array of activities. Within the geographical sciences, there are numerous examples of data generated by volunteers being acquired and used across the human and physical dimensions of the subject area. Volunteered geographic information (VGI) is now, for example, commonly encountered within contexts as diverse as conservation (Newell *et al.*, 2012), hydrology (Lowry and Fienen, 2013), meteorology (Siegel, 2013) and urban planning (Brabham, 2009) encompassing also activities such as contributions to contemporary problems and crises such as post-disaster damage assessments (Goodchild and Glennon, 2010; Zook *et al.*, 2010; Gao *et al.*, 2011; van Aardt *et al.*, 2011). In many cases, this activity is coordinated through Internet-based collaborative projects, a feature that acts to ensure barriers to involvement are low and so open up contribution to the populace worldwide, although access to technology can still sometimes be a constraint (Haklay, 2013). In relation to cartography, VGI has commonly been acquired via Internet-based projects and widely used in relation to map production (e.g. Mooney and Corcoran, 2012a) and map validation (Foody and Boyd, 2013) activities.

There is tremendous potential for VGI to aid mapping activities. VGI can, for example, be acquired quickly over large areas and cheaply, if not freely, and there is considerable scope for the information to be acquired to support

DOI: 10.1080/00087041.2015.1108658

basic tasks such as map production and updating. In this way, VGI has revolutionized aspects of mapping and increased the engagement of the general public in mapping-related activity. However, VGI has, of course, limitations and by its very nature presents numerous concerns. Indeed, the limitations of VGI may in some cases be more apparent than its positive attributes. Many concerns arise because volunteering is relatively unconstrained; the broad reach of the Internet and proliferation of location enabled devices and other geoinformation technologies means that barriers to involvement are often low and almost anyone, anywhere, can potentially contribute. The great advantages that this freedom to contribute conveys in terms of the ability to acquire data for areas, large or small, are tempered by concerns that it also allows inputs from the utterly incompetent or even malicious individual who wishes to damage the activity. Indeed, the volunteers contributing to a project can vary greatly in nature and ability from the naive novice to authoritative expert (Raykar *et al.*, 2010; Brabham, 2012) as well as undesirable people such as spammers (Vuurens *et al.*, 2011; Hirth *et al.*, 2012; Neis *et al.*, 2012; Raykar and Yu, 2012), with each type possibly contributing to differing and unknown extents to a project. There are, therefore, numerous concerns with the quality of VGI (Flanagin and Metzger, 2008).

Often there is no obvious means to distinguish between contributors and hence to perhaps differentially weight the information from contributors or attach confidence or trust levels to the information each provides. Critically, unlike authoritative data collection, VGI is typically ac-quired without accord to strict protocols and standards (Haklay *et al.*, 2010). VGI is fundamentally imperfect but also generally of unknown and heterogeneous imperfection. Given the immense potential of citizen sensing and its obvious limitations, it is common, therefore, to encounter extreme views on VGI and its practical value. Indeed, VGI is often viewed either very favourably, with a focus on its positive features, an emphasis on the wisdom of the crowd and its future potential, or negatively with a focus on its limitations and a fear of 'mob rule' (Roman, 2009). Between these two ends of the spectrum are perspectives that consider the potential for both VGI and authoritative data to be used together, perhaps blended in some way or used in a manner that takes advantage of their relative strengths. These activities may, therefore, sometimes straddle the amateur and professional communities for mutual benefit in support of a specific objective, but are constrained by concerns with the quality of the information.

The use of VGI in mapping is recognized as a major challenge (Haklay *et al.*, 2010). A central issue of concern is the quality of the information that is volunteered as error and uncertainty in VGI will limit the trust that can be placed upon it and its practical value. So, while VGI has great potential in a diverse array of mapping applications its value and use will be tempered by the concerns. Not least among the latter are problems such as the provision of information that may not only be inconsistent but also conflicting which can ultimately act to hinder the very application the volunteers are seeking to aid (Voigt *et al.*, 2011). For the full potential of VGI to be realized, it is important to be able to characterize its quality and to be able to rate volunteers in terms of the accuracy of their contributed data. This not only helps the users of VGI, but could be seen as providing feedback to the volunteers to help them enhance their skills and understanding.

For many concerned with map production and evaluation, a key objective is to make good use of the VGI through awareness of limitations connected to its quality. Without direct information on quality of the VGI, it may be tempting to seek contributions from a large number of volunteers. The basis for this is founded on observations from crowdsourcing projects that show a positive relationship between the accuracy of contributed data and number of volunteers (Snow *et al.*, 2008; Welinder *et al.*, 2010) as well as Linus' law (Haklay *et al.*, 2010). However, the practical implementation of this strategy can be a challenge. For example, while Linus' law may suggest that a correct contribution will be made eventually as more and more volunteers take part, it is difficult to see how the one correct contribution would be selected against the majority view provided by the other volunteers. In standard ensemble or consensus-based approaches to using data that are popular when the contributions from multiple volunteers are available, a common approach is to follow the majority or dominant view. Thus, the information provided by the one in a million volunteer who, for example, labels a particular case correctly will normally be lost within the much larger sea of alternative labels. Additionally, there are associated concerns with seeking increasing numbers of volunteers as this may actually degrade rather than enhance the quality of the data set. Simply providing more data may be unhelpful if, for example, the data are of poor quality; the value of gaining more contributions from volunteers who provide inaccurate data is questionable and acts to dilute the value of the useful data. Note also that even if the data provided by additional volunteers are of a high quality, it can sometimes be unhelpful to add it. This relates to a well-known problem encountered in remote sensing, the curse of dimensionality or Hughes effect, in which the accuracy of a mapping project may decline as the volume of data increases, even if the data provided are of high quality and often places a demand on researchers to actively seek to reduce the size of their data sets (Pal and Foody, 2010).

Assessing the quality of volunteered or crowdsourced information has been a subject of considerable research within the general realm of citizen science for some time. A variety of approaches to rating data have been developed (Raykar and Yu, 2012), although some of the methods have major limitations such as poor performance when a volunteer provides only a small amount of data (Raykar and Yu, 2011). Alternatively, sometimes it is possible to adopt a 'social' approach to quality assurance which may, for example, draw on inputs from trusted individuals who act as gatekeepers (Goodchild and Li, 2012). Additional approaches to quality assurance use geographical contextual information as a means to check the sense or reasonableness of the contributed information given existing knowledge (Goodchild and Li, 2012). It may not, however, always be possible to adopt such approaches and they too, of course, have limitations such as the quality of the gatekeepers or editors and the contextual information available. As a result, there is considerable interest in intrinsic measures of data

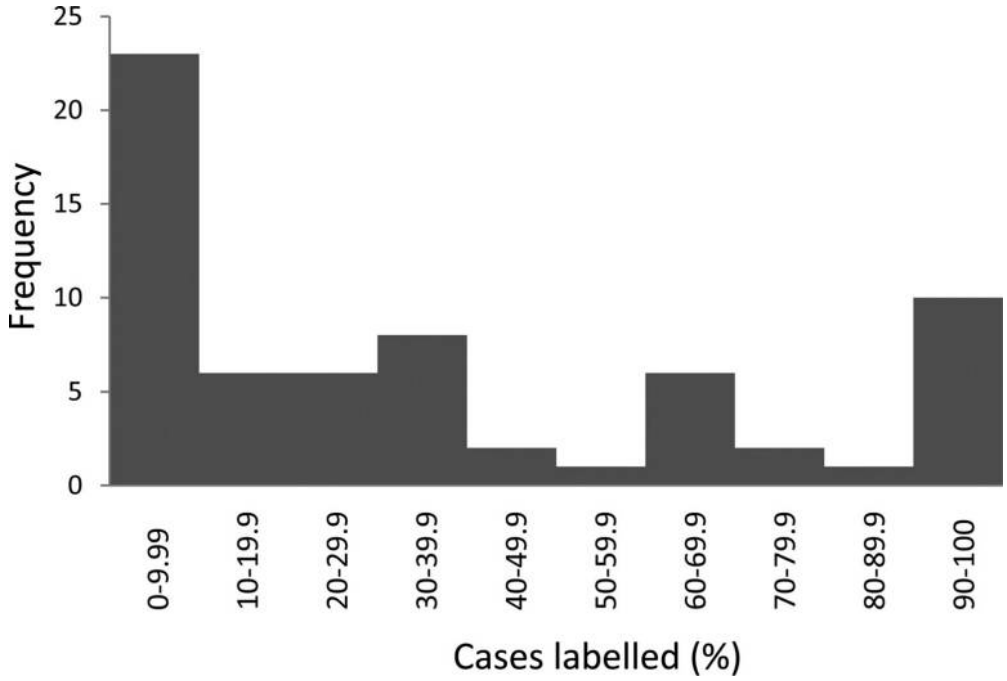

Figure 1. Histogram depicting the number of cases labelled (as a percentage of the total 299 cases available) by the 65 volunteers to the project

quality, which is in measures of quality that are derived from the data set itself and without an independent reference.

In a recent article, Haklay *et al.* (2010) highlight a set of issues relating to VGI in mapping applications. They stress that while much research has addressed aspects of positional accuracy and completeness, other aspects such as attribute accuracy require attention. They also suggest the need for research into issues such as how many volunteers are needed, how good the volunteers need to be at the task, the amount of information they contribute and assessments of whether the quality of information provided by a volunteer varies with the amount of data they have provided (e.g. do volunteers get better as they do more?). This article seeks to contribute to research on these topics. Specifically, it uses an intrinsic measure of volunteer data quality to explore issues of volunteer quantity and quality in relation to map attribute accuracy. Thus, we aim to explore some issues connected with the quantity (number of volunteers and number of cases contributed by a volunteer) and quality (accuracy of labelling) of VGI in the context of thematic mapping. The focus is on the accuracy of attribute labelling and draws on a real VGI programme run through the Geo.wiki system (Fritz *et al.*, 2012; Perger *et al.*, 2012). In essence, we seek to develop answers to three inter-related questions. First, can the quality of volunteers in terms of the accuracy of their labelling be characterized from only the data they provide and does the quality of this characterization vary with the number of volunteers contributing labels? Second, is the number of cases contributed by a volunteer related to the quality of the data volunteered and does it impact upon the characterization of volunteer quality? Third, can information on volunteer quality be used in a way to enhance mapping applications?

AQUISITION OF VGI

VGI was acquired through the Geo-wiki project (http://www.geo-wiki.org). This provides an Internet-based means for volunteers to contribute to a mapping task (Fritz *et al.*, 2012). An unrestricted call for contributions to the project was made with volunteers contributing over the period September to November 2011. The volunteers were invited to view a series of up to 299 satellite sensor images and assign each a land cover label from a defined list of 10 classes; more details are given in Perger *et al.* (2012). The set of labels from each volunteer were available for this study with no additional information. The end product was a data set that comprised a matrix of class labels for 65 volunteers over the 299 cases. After the contributions were received, three experts, who had also contributed to the project, met to discuss the task and agree on an authoritative label for each of the 299 cases. The latter set of labels was used as the reference data in exploring the accuracy of allocations made by the volunteers.

Critically, for each of the 299 images available, a set of class labels was derived by volunteers from around the world. As in other projects based on VGI, there was an unequal contribution by the volunteers (Mooney and Corcoran, 2012b). Here, a total of 65 volunteers contributed data. On average, a volunteer labelled ~110 (~36.8%) of the 299 cases. The variation in contribution was, however, very large (Figure 1). One volunteer labelled just a single case and 25% of the volunteers labelled 19 (~6.3%) or fewer of the cases. While most volunteers labelled only a small proportion of cases, it was evident that some valiantly contributed to the project with 25% of volunteers labelling at least 203 (~67.9%) cases. Indeed,

10 volunteers labelled at least 90% of the cases with three labelling all 299.

Initial analyses used all of the data acquired to explore the ability to characterize the accuracy of the volunteers in terms of their labelling. To determine if the accuracy characterizations could be used constructively to aid mapping applications, further analyses were undertaken but these focused on subdivisions of the data set. Unknown to the contributors, the 299 images had been presented in three batches or sections to allow investigation into a range of factors that affect the quality of labelling; the first batch comprised 99 cases and the remaining two batches each comprised 100 cases. The main difference between these batches of images was that the third comprised images acquired by a finer spatial resolution sensor than used in the first two batches and as such might be believed to be easier to label accurately. The nature of the data set allowed an investigation into whether volunteers appeared to get better with experience, by focusing on the difference in labelling between the first and second batch of cases. The design also allowed an initial assessment of the effect of image spatial resolution on labelling, by focusing on differences in labelling between the second and third batches of cases.

In many studies, attention is focused upon on sub-set of the classes present, often on a single class. This is common in remote sensing applications, from projects such as those focused on crops (Hill *et al.*, 1980), habitat monitoring in support of national and international directives (Sanchez-Hernandez *et al.*, 2007) or studies of change such as those linked to major international policies such as the United Nations programme on Reducing Emissions from Deforestation and Forest Degradation (Gibbs *et al.*, 2007). Here, the concern was focused on only one class: cultivated and managed land. This was the most abundant class, with 119 of the 299 images actually of this class. This class was selected given interest in human use of the land and as a means to reduce sensitivity to sample size problems in the analyses.

AN INTRINSIC MEASURE OF QUALITY

In situations where a set of volunteers label the same set of cases, it may be possible to derive an intrinsic measure of volunteer quality via a latent class model (Foody *et al.*, 2013). As an intrinsic measure, it is derived entirely from the data itself without any additional information such as reference data. The method is described in detail elsewhere (e.g. Vermunt and Magidson, 2003a; Magidson and Vermunt, 2004), but as it is used here, a brief overview of the salient features is provided.

The basis of the approach is that the set of labels derived from the volunteers for each case labelled conveys information on the actual or true label in the unobserved (latent) variable. The relationship between the observed and latent variables may be expressed in the form of a latent class model such as

$$f(y_i) = \sum_{x=1}^{K} P(x) \prod_{t=1}^{T} f(y_{it}|x)$$

in which $f(y_i)$ is a vector representing the entire set of responses from the T volunteers ($1<t<T$) contributing to the project for case i, K is the number of classes and x the latent variable (Vermunt and Magidson, 2003a, b). The key feature of relevance to this article is that if the model fits the observed data adequately, then its parameters may be used to indicate the quality of labels provided by each volunteer. This is because the $f(y_{it}|x\cdot)$ parameters of the model above represent the conditional probabilities of membership. From the model, it is, therefore, possible to extract the conditional probability that a case allocated the label of the cultivated and managed land cover class is actually a member of that class in the ground reference data. In the terminology used widely within the remote sensing community, this latter value is the producer's accuracy for the class (Liu *et al.*, 2009) and an estimate may be derived for each volunteer from the model's parameters. Thus, the parameters of the latent class model may be used to generate for each volunteer a value that indicates their accuracy in labelling the class of interest. This value is generated without the use of reference data and so represents an intrinsic measure of data quality. The accuracy of the estimates derived in this way can be evaluated through comparison against the ground reference data generated by the set of expert labellers.

CHARACTERIZING THE QUALITY OF VOLUNTEERS

A minimum requirement for developing an identifiable latent class model is that there should be at least data from three volunteers. Leaving the labels from the experts used to generate the reference data aside for the moment, only two volunteers labelled all 299 cases. Thus, to fit a model to the data volunteered by unknown individuals, it would be necessary to include some unlabelled or missing cases in the analysis. Recognizing that while the latent class model can operate with missing cases (Vermunt, 1997), but is not completely unaffected by this situation (Foody, 2013), an initial starting point might reasonably be to use the labels provided by the five volunteers who labelled the most cases. This fits with other studies of annotations obtained by crowdsourcing that show high accuracy can be achieved from a small number of volunteers (Snow *et al.*, 2008), ensures sufficient data for the specification of an identifiable model and reduces problems associated with missing cases. Additionally, it might be speculated that those volunteers who labelled the most cases were highly committed and possibly highly skilled for the task; the validity of this contention will be explored later. In the data generated by these five volunteers, there were just 19 unlabelled cases (i.e. 1.27% of cases had missing labels).

The data from the five volunteers who labelled the most cases were used to form a latent class model. The outputs of this model provided only a poor characterization of the volunteers in terms of their labelling accuracy, expressed in terms of an estimate of their producer's accuracy for the cultivated and managed land cover class. The accuracy with which the initial five volunteers labelled the class was very low, with a weak, near zero and insignificant correlation between the actual and estimated values ($r=0.03$); it should

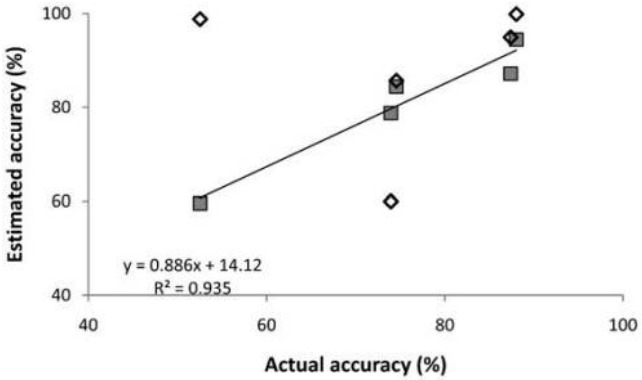

Figure 2. Relationship between the actual and estimated accuracy for the five volunteers who contributed the most data obtained when using the data from only those five volunteers (open diamonds) and, with regression line and equation, when using the data from all 65 volunteers (filled squares)

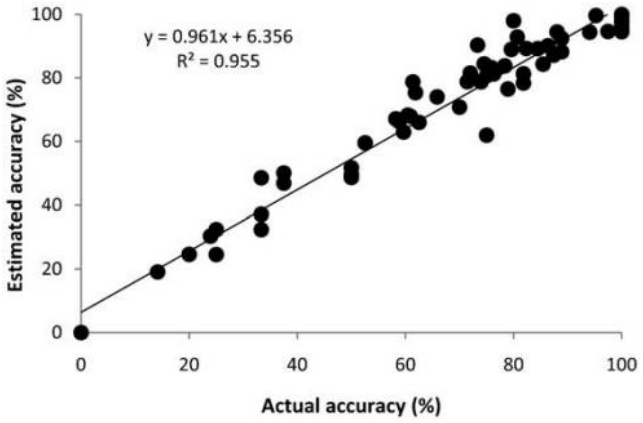

Figure 3. Relationship between the actual and estimated accuracy for each of the 65 volunteers who contributed labels

be noted that the small sample size limits value of this analysis. Previous work with this data set (Foody *et al.*, 2013) suggested that accurate characterization was possible with a slightly larger number of volunteers, seven. Additionally, Haklay *et al.* (2010) suggests that if Linus' law holds accuracy might be expected to increase should the number of volunteers increase until a threshold is reached after which accuracy no longer continues to increase. Indeed, studies have shown that labelling accuracy can vary positively with the number of volunteers (Welinder *et al.*, 2010). Randomly selecting from the as yet unused volunteers, the number of volunteers was increased to 10, 15, 25 and 35 before using the data from all 65. The latent class analyses were repeated for each group of volunteers and the accuracy with which the quality of the five initial volunteers was characterized was evaluated. The results showed that the quality of the characterization of each volunteer, reflected in the R^2 statistics obtained, tended to increase as the number of volunteers contributing data increased. The trend was not perfect, with a relatively poor estimate of producer's accuracy derived from the analysis using data from 35 contributors. This may be because the set of contributors selected to increase the number of volunteers up from 25 to 35 included four contributors who labelled fewer than 27 cases (i.e. less than 10% of the total number of cases available for labelling) and all labelled less than 52% of the cases. It was possible that the incompleteness of the labelling by this set of volunteers may contribute to the poor estimation. The trend between the accuracy of the estimates and the number of volunteers, however, does suggest that while quantity of contributors is important, there are additional variables affecting the ability to characterize the accuracy of the volunteers in terms of their labelling. The latter might include issues such as the quality of the individual volunteers, the consistency of the volunteers within the task and data concerns such as the missing cases. Of particular note, however, is that when the data from all 65 volunteers were used, the correlation between the predicted and actual accuracy of the volunteers in terms of their labelling was very high (R^2=0.9359; Figure 2). This result shows that VGI sources can be well characterized from the volunteered data alone and confirms the potential of the latent class model approach as a means to estimate the quality of volunteers when cases are labelled multiple times. It also appears that the ability to characterize the quality of the volunteers tends to be high except when only a small number of volunteers is used (Table 1). Thus far, the attention has been on the labelling provided by the five volunteers who labelled the most cases. The data set, however, also allows a deeper investigation into the effect of variations in the number of cases contributed by volunteers on the ability to rate their accuracy. Of particular interest, was the entire data set that comprised a relatively large number of volunteers, 65, who also contributed to differing degrees, from labelling just one to all 299 cases (Figure 1). Analyses were undertaken to determine if it was possible to estimate the accuracy of each of the 65 volunteers. This analysis would also reveal if it was possible to characterize correctly the accuracy of a volunteer irrespective of the number of cases he or she labelled, and would address the challenge of rating contributors who provide few labels (Raykar and Yu, 2012). A standard latent class model was applied to the entire data set and appeared to fit the data closely. The parameter that expressed the producer's accuracy for the labelling of each volunteer was extracted from the model and compared against the actual value derived relative to the reference data (Figure 3). It was evident that the estimated accuracy was close to the actual accuracy of the labelling of each volunteer and that there was a close relationship between the estimated and actual producer's accuracy values across the entire range of

Table 1. Variation in the strength of the relationship between the estimated and actual accuracy for the five volunteers who contributed the most to the project for different numbers of volunteers contributing

Number of volunteers	R^2
5	0.0009
10	0.8194
15	0.8594
25	0.8579
35	0.7279
65	0.9359

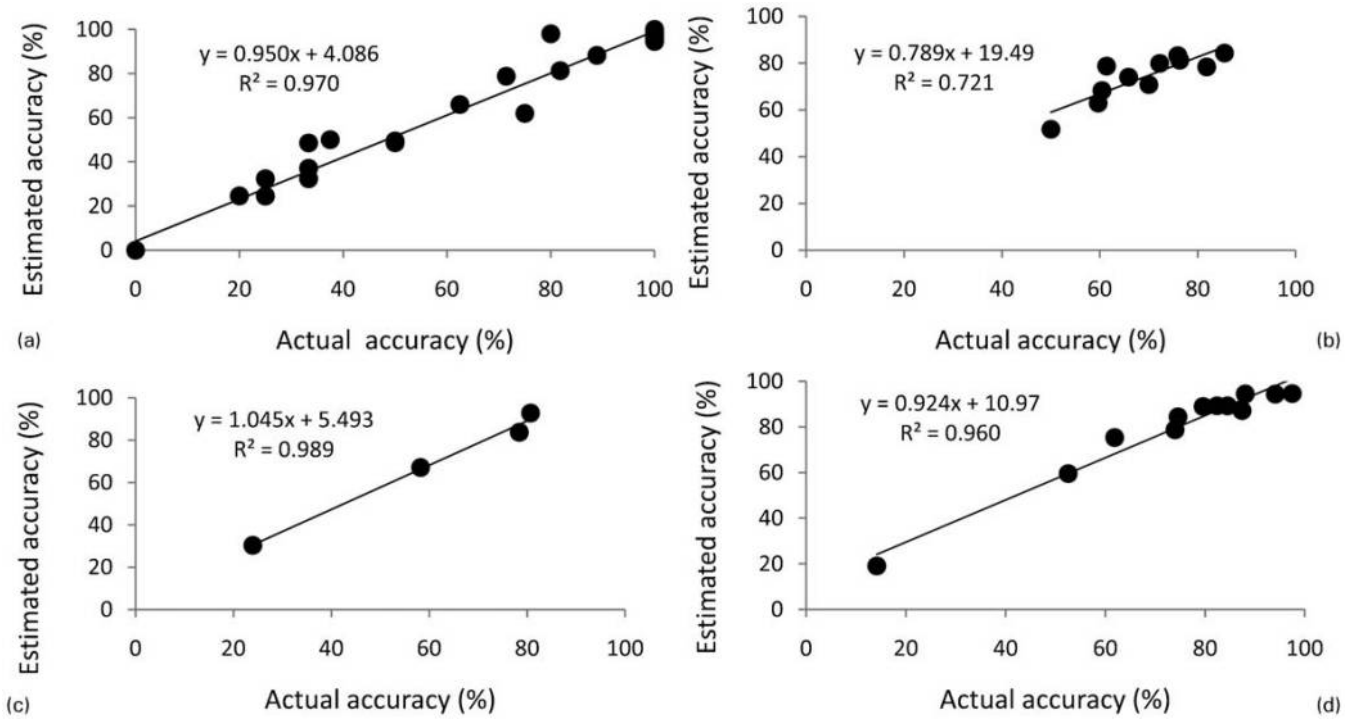

Figure 4. Relationships between the actual and estimated accuracy for volunteers contributing different amounts of data based on the first two batches of images, labelling (a) 1–29 cases, (b) 30–59 cases, (c) 60–89 cases and (d) 90–119 cases

accuracy values. Additionally, as the volunteers contributed to very different degrees, the relationship indicates that the amount of data contributed had no major effect on the analysis and the ability to rate volunteers in terms of their quality of labelling. Thus, even though the number of cases contributed by a volunteer varied greatly, the latent class model approach was able to characterize the accuracy of each volunteer very well irrespective of the number of cases labelled. To further illustrate this, the data set was divided in quartiles defined by the number of cases labelled. Within each quartile, there was a strong relationship between the actual and estimated accuracy values (Figure 4). Critically, even when a volunteer contributed a few labels, the latent class analysis was able to characterize the accuracy of that volunteer. It was also evident that in each quartile, a wide range of accuracies was encountered. The latter illustrates that there was no simple relationship between the number of cases labelled and the accuracy of the labelling. Thus, it was quite possible for a volunteer to contribute few cases but to be highly accurate (Figure 4a) and equally a volunteer could contribute many labels but be relatively inaccurate (Figure 4d). The latter shows that it is inappropriate to assume that those volunteers who provide the most labels are accurate. Indeed, those who labelled the most images varied greatly in quality, spanning a range in accuracy from 14.15% to 97.45% (Figure 4d); only the volunteers contributing the fewest cases had a greater range which spanned the full scale from 0% to 100% (Figure 4a). It was, however, evident that only one of the volunteers who contributed a large number of labels had an accuracy of $<$50% with the other volunteers of markedly higher accuracy.

USING INFORMATION ON VOLUNTEER QUALITY TO ENHANCE MAPPING

The results above confirmed an ability to characterize the quality of the volunteers in terms of the accuracy with which they labelled cases provided that there was a reasonably large number of volunteers labelling the same set of cases. Now we seek to determine if this ability to rate volunteers can be used constructively in mapping applications.

Armed with a means to estimate the quality of the volunteers in terms of labelling accuracy via a latent class model, we investigated if a sub-set of the volunteers can be identified in a way that helps a mapping application. Critically, we seek to identify the best volunteers in terms of labelling accuracy, so that they, and they alone, may be used to undertake further labelling or to highlight volunteers whose contributions might be down-weighted, even ignored, in later analyses or perhaps might be targeted for training to enhance future performance. We also seek an appreciation of how many volunteers are actually needed and if information on the quality of the volunteers in terms of their labelling can be used to keep the quantity of volunteers required to a small number.

Focusing on the 14 volunteers who labelled all 199 cases in the first and second batches of images, a latent class model was fitted to the data on the first set of cases (n=99). This model was used to characterize the quality of the 14 volunteers in terms of the producer's accuracy for the cultivated and managed class. This information was then used to select a sub-set of the 14 volunteers to label the second set of cases (n=100). Attention focused on the accuracy with which the cases in the second set of cases were classified, with each case labelled following a basic ensemble

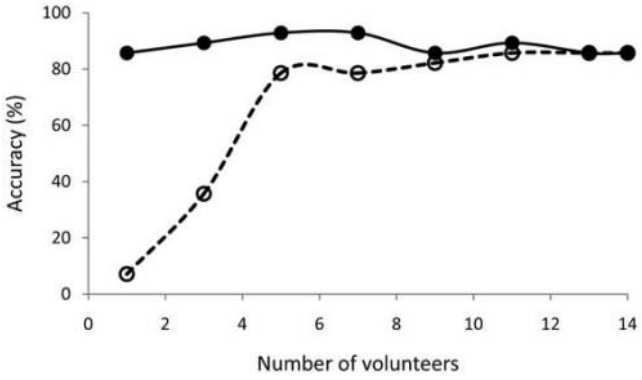

Figure 5. Relationship between the producer's accuracy of labelling and the number of volunteers when starting with the most accurate volunteer (solid circles and line) and least accurate volunteer (open circles and dashed line)

allocation (i.e. the dominant class label over the set provided by the volunteers was allocated to each image) and the accuracy of the resulting allocation was evaluated relative to the label depicted in the ground reference data set generated by the experts. Two contrasting approaches were adopted to explore the value of information on volunteer quality on the mapping task. Both of these approaches selected volunteers based on their rank order of producer's accuracy derived from the latent class model fitted to the data from the first batch of images. The first set of analyses began with the best individual volunteer, with the highest estimated producer's accuracy, and added data from the next best pair of volunteers (to allow an odd number of volunteers and so aid identification of the dominant class label) until the data from all 14 were used. The second approach started at the opposite end of the spectrum and began with the volunteer who had the lowest producer's accuracy and added progressively stronger volunteers until all 14 had been used.

The best volunteer had an estimated producer's accuracy of 85.71%. Adding data from the next best pair of volunteers resulted in accuracy rising to 89.29%. The addition of data from a further pair of volunteers also resulted in an increase in accuracy to 92.86% (Figure 5). This latter value was, however, the highest accuracy obtained as the addition of data from further volunteers ultimately resulted in the accuracy with which the second batch of cases were classified declining back to 85.71% when the data from all 14 volunteers were used, the value achieved by the best individual volunteer. The trend was, therefore, one in which accuracy initially rose with the addition of more volunteers, reached a peak and then declined. This confirms that there is a danger in following a simple 'the more the better' philosophy in relation to the number of volunteers as simply increasing the number of volunteers may actually degrade the accuracy of the classification.

The worst volunteer had an estimated producer's accuracy of just 7.14%. Adding data from additional volunteers, of increasing quality, was found to generally increase the accuracy with which the cases in the second batch of images were labelled (Figure 5). Indeed, it was evident that there was a general trend for accuracy to rise from the low starting point as volunteers were added. The size of the increase in accuracy obtained at each step, however, declined. Moreover, the maximum value achieved with the use of all 14 volunteers was, as noted above, the same as the accuracy achieved if only the labelling of the single best volunteer had been used. The shape of the relationship in Figure 5 also suggests that the addition of further volunteers may not be expected to have a major effect on mapping accuracy.

The results of the two approaches to adding volunteers highlight again that the number of volunteers is not the only issue to consider in using VGI and that the quality of the data provided is important. Critically, if the quality of the contributions from volunteers can be characterized, this provides a means to select a strong sub-set and this may allow the accuracy of a mapping task to be undertaken more accurately than if a larger set of volunteers had been used. It is also important to note that the two approaches are just a guide to the trends and need not represent the best or worst scenarios. Note, for example, that the accuracy of labelling would have been 100% had the data from three specific volunteers been used. Thus, while the latent class model enabled a strong sub-set of volunteers to be selected, it was not, in this case, the best possible sub-set.

Finally, a series of analyses were undertaken to reveal within-task variation in labelling by the volunteers. As

Table 2. The actual and estimated producer's accuracy for the volunteers who provided the most data obtained using the entire data set (299 cases) and each batch of images alone. The correlations between the actual and estimated accuracy values were all large and significant: the values of R^2 calculated using all of the data, the first, second and third batches were 0.9767, 0.9819, 0.9817 and 0.8149 respectively

	All		First		Second		Third	
Volunteer	Actual	Estimated	Actual	Estimated	Actual	Estimated	Actual	Estimated
1	94.06	94.42	100.00	99.65	88.88	86.69	94.20	96.81
2	87.39	87.18	45.45	47.19	100.00	99.72	95.65	99.94
3	82.35	89.28	72.72	83.85	82.14	87.24	85.50	96.78
4	97.45	94.62	86.36	83.88	100.00	99.72	100.00	99.94
5	84.48	89.24	95.45	94.41	92.85	95.57	77.27	96.97
6	14.15	19.09	22.72	26.21	7.14	8.31	14.28	29.42
7	88.03	94.47	86.36	89.15	89.28	95.56	88.06	99.88
8	52.54	59.56	36.36	36.69	46.42	53.99	60.29	88.72
9	74.56	84.42	95.45	99.68	85.71	95.56	62.50	96.52
10	73.94	78.77	50.00	52.42	85.71	91.40	76.81	93.77
Average	74.89	79.10	69.08	71.31	77.81	81.37	75.45	89.87

noted above, the images had been presented to the volunteers in three batches. The first two batches were of similar nature and so allowed an exploration of the impact of experience on the volunteers. The third batch was made up of images with a fine spatial resolution and so allowed insight into the effect of variations in image quality on the accuracy with which volunteers could label images. Using the data from the 10 volunteers who provided the most labels, thus reducing complications linked to concerns such as missing cases, a series of latent class models were generated using all of the data as well as that generated with the first, second and third batches of images alone to explore variations in labelling accuracy with experience and image type.

The estimates of producer's accuracy from the latent class models and the actual values determined relative to the reference data were close for analyses based on the entire data set and for the individual batches of images (Table 2). However, it was noticeable that the second set of cases were generally classified more accurately than the first which might suggest that some contributors improved in terms of labelling accuracy and learned to provide better labels. There were, however, some contributors whose performance actually declined over the three sets of data, perhaps indicating a problem such as tiredness (e.g. volunteers 5 and 9 in Table 2). It was also evident that the third set of cases were typically classified more accurately than the first set, suggesting that the finer spatial resolution imagery may have enabled more accurate labelling. The small difference in average accuracy between the second and third sets of data, however, suggests that the effects arising from the use of finer spatial resolution images may have been small, at least for the class of interest and data sets used. It was also evident that the latent class model tended to over-estimate the accuracy of the labelling for the third set of cases. This latter result highlights a need for more research on the potential and limitations of the latent class model for characterizing the quality of volunteers.

CONCLUSIONS

VGI is fundamentally flawed, but by explicitly recognizing this issue and working intelligently with it, the cartographic community can make use of it for mapping applications. To aid this activity, there is a need to be able to characterize the accuracy of the volunteered data. Here, we used an intrinsic method of quality characterization based on a latent class model to indicate the accuracy of VGI. This method allows the characterization of volunteers in terms of the accuracy of their labelling without any reference data. The quality characterizations also allow the user of the VGI to selectively choose volunteers for further data collection in a way that should help achieve a high accuracy. Thus, following a 'bigger is better' policy with regard to the number of volunteers may be inappropriate as higher accuracy can be obtained by using a carefully selected, and relatively small, set of volunteers identified on the basis of their estimated quality.

The key conclusions linked to the questions raised earlier are:

1. The latent class model appears to offer a means to characterize the quality of sources of VGI. The quality of the characterization of volunteer labelling accuracy was also generally high except when only a small number of volunteers was used.
2. The size of a volunteer's contribution was not a good guide to the quality of the data provided. Large variations in the accuracy of the data provided by volunteers were evident, but those contributing few cases could be as accurate as those contributing many and *vice versa*. It was also evident that an accurate characterization of the accuracy of a volunteer in terms of the quality of the data provided could also be made irrespective of the number of cases contributed by that volunteer.
3. The information on volunteer quality revealed by the latent class model may be used constructively. It can, for example, be used to rank volunteers so that only the data from high quality sources are used and this can increase the accuracy of later analyses. Moreover, as a higher accuracy can be obtained from a sub-set of the volunteers, the ability to rate and rank volunteers helps in the selection of an appropriate number of high quality volunteers to use in preference to a simple 'bigger is better' approach to using volunteers. The model was also able to highlight within-task variations in volunteer performance. The latter may help both the volunteers and users of the data (e.g. in identifying suitable working practices or training needs, etc.).

BIOGRAPHICAL NOTES

Giles M. Foody is Professor of Geographical Information Science at the University of Nottingham, UK. His main research interests focus on the interface between remote sensing, ecology, and informatics. He is chair of the European Union COST Action on 'Mapping and the Citizen Sensor' (TD1202) which involves researchers from over 30 countries.

ACKNOWLEDGEMENTS

We gratefully thank 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 EPSRC (reference EP/J0020230/1), British Academy (reference SG112788) and European Union COST Action TD1202. The latent class analyses were undertaken with the LEM and LatentGold packages.

REFERENCES

Brabham, D. C. (2009). 'Crowdsourcing the public participation process for planning projects', **Planning Theory**, 8, pp. 242–262.

Brabham, D. C. (2012). 'The myth of amateur crowds: a critical discourse analysis of crowdsourcing coverage', **Information, Communication and Society**, 15, pp. 394–410.
Flanagin, A. and Metzger, M. (2008). 'The credibility of volunteered geographic information', **GeoJournal**, 72, pp. 137–148.
Foody, G. M. (2013). 'Rating crowdsourced annotations: evaluating contributions of variable quality and completeness', **International Journal of Digital Earth**.
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 Observation and Remote Sensing**, 6, 1305–1312.
Foody, G. M., See, L., Fritz, S., van der Velde, M., Perger, C., Schill, C. and Boyd, D. S. (2013). 'Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project', **Transactions in GIS**.
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, pp. 110–123.
Gao, H., Barbier, G. and Goolsby, R. (2011). 'Harnessing the crowdsourcing power of social media for disaster relief', **IEEE Intelligent Systems**, 26, pp. 10–14.
Gibbs, H. K., Brown, S., Niles, J. O. and Foley, J. A. (2007). 'Monitoring and estimating tropical forest carbon stocks: making REDD a reality', **Environmental Research Letters**, 2, 045023.
Goodchild, M. F. (2007). 'Citizens as sensors: the world of volunteered geography', **GeoJournal**, 69, pp. 211–221.
Goodchild, M. F. and Glennon, J. A. (2010). 'Crowdsourcing geographic information for disaster response: a research frontier', **International Journal of Digital Earth**, 3, pp. 231–241.
Goodchild, M. F. and Li, L. (2012). 'Assuring the quality of volunteered geographic information', **Spatial Statistics**, 1, pp. 110–120.
Haklay, M. (2013). 'Neogeography and the delusion of democratisation', **Environment and Planning A**, 45, pp. 55–69.
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', **The Cartographic Journal**, 47, pp.315–322.
Hill, J. D., Strommen, N. D., Sakamoto, C. M. and Leduc, S. K. (1980). 'LACIE – application of meteorology for United-States and foreign wheat assessment,' **Journal of Applied Meteorology**, 19, pp. 22–34.
Hirth, M., Bobfeld, T. and Tran-Gia, P. (2012). 'Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms', **Mathematical and Computer Modelling**, 57, 2918-2932.
Liu, C. White, M. and Newell, G. (2009). 'Measuring the Accuracy of Species Distribution Models: A Review', in **18th World IMACs/ MODSIM Congress**, pp. 4241–4247, Cairns, Jul 13–17.
Lowry, C. S. and Fienen, M. N. (2013). 'Crowdhydrology: crowdsourcing hydrologic data and engaging citizen scientists', **Ground Water**, 51, pp. 151–156
Magidson, J. and Vermunt, J. K. (2004). 'Latent class models', in **The SAGE Handbook of Quantitative Methodology for the Social Sciences**, ed. by Kaplan, D., pp. 175–198, Sage, Thousand Oaks, CA.
Mooney, P. and Corcoran, P. (2012a). 'The annotation process in OpenStreetMap', **Transactions in GIS**, 16, pp. 561–579.
Mooney, P. and Corcoran, P. (2012b). 'Characteristics of heavily edited objects in OpenStreetMap', **Future Internet**, 4, pp. 285–305.
Newell, D. A., Pembroke, M. M. and Boyd, W. E. (2012). 'Crowd sourcing for conservation: Web 2.0 a powerful tool for biologists', **Future Internet**, 4, pp. 551–562.
Neis, P., Goetz, M. and Zipf, A. (2012). 'Towards automatic vandalism detection in OpenStreetMap', **ISPRS International Journal of Geoinformation**, 1, pp. 315–332.
Pal, M. and Foody, G. M. (2010). 'Feature selection for classification of hyperspectral data by SVM', **IEEE Transactions on Geoscience and Remote Sensing**, 48, pp. 2297–2307.
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 **GI_Forum 2012: Geovizualisation, Society and Learning**, ed. by Jekel, T., Car, A., Strobl, J. and Griesebner, G., pp. 83–91, Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/ Offenbach.
Raykar, V. C. and Yu, S. (2011). 'An Entropic Score to Rank Annotators for Crowdsourced Labelling Tasks', in **Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics**, pp. 29–32, Hubli, Karnataka, India, Dec 15–17.
Raykar, V. C. and Yu, S. (2012). 'Eliminating spammers and ranking annotators for crowdsourced labelling tasks', **Journal of Machine Learning Research**, 13, pp. 491–518.
Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L. and Moy, L. (2010). 'Learning from crowds', **Journal of Machine Learning Research**, 11, pp. 1297–1322.
Roman, D. (2009). 'Crowdsourcing and the question of expertise', **Communications of the ACM**, 52, p. 12.
Sanchez-Hernandez, C., Boyd, D. S. and Foody, G. M. (2007). 'One-class classification for monitoring a specific land cover class: SVDD classification of fenland', **IEEE Transactions on Geoscience and Remote Sensing**, 45, 1061–1073.
Siegel, R. P. (2013). NOAA starts crowdsourcing weather data. http://www.triplepundit.com/2013/02/noaa-starts-crowdsourcing-weather-data-2/
Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. Y. (2008). 'Cheap and Fast – But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks', in **2008 Conference on Empirical Methods in Natural Language Processing**, pp. 254–263, Hawaii, HI, Oct 25–27.
Sui, D., Elwood, S. and Goodchild, M (Ed.). (2013). **Crowdsourcing Geographic Knowledge**, Springer, New York.
van Aardt, J. A. N., McKeown, D., Fauiring, J., Raqueno, N., Caterline, M., Renschler, C., Eguchi, R., Messinger, D., Krzaczek, R., Cavilla, S., Antalovich, J., Philips, N., Bartlett, B., Salvaggio, C., Ontiveros, E. and Gill, S. (2011). 'Geospatial disaster response during the Haiti earthquake: a case study spanning airborne deployment data collection, transfer, processing and dissemination', **Photogrammetric Engineering and Remote Sensing**, 77, pp. 943–952.
Vermunt, J. K. (1997). **Log-linear Models for Event Histories**, Sage, Thousand Oaks, CA.
Vermunt, J. K. and Magidson, J. (2003a). 'Latent class analysis', in **The Sage Encyclopaedia of Social Science Research Methods**, ed. by Lewis-Beck, M., Bryman, A. E. and Liao, T. F., Vol. 2, pp. 549–553, Sage Publications, Thousand Oaks, CA.
Vermunt, J. K. and Magidson, J. (2003b). 'Latent class models for classification', **Computational Statistics and Data Analysis**, 41, pp. 531–537.
Voigt, S., Schneiderhan, T., Tweie, A., Gahler, M., Stein, E. and Mehl, H. (2011). 'Rapid damage assessment and situation mapping: learning from the 2010 Haiti earthquake', **Photogrammetric Engineering and Remote Sensing**, 77, pp. 923–931.
Vuurens, J., de Vries, A. P. and Eickhoff, C. (2011). 'How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy', in **SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR)**, Beijing, Jul 28, pp. 48–55.
Welinder, P., Branson, S., Belongie, S. and Perona, P. (2010). 'The multi-dimensional wisdom of crowds', in **Advances in Neural Information Processing Systems 23**, ed. by Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R. and Culotta, A., pp. 2424–2432, Curran Associates, Red Hook, NY.
Zook, M., Graham, M., Shelton, T. and Gorman, S. (2010). 'Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake', **World Medical and Health Policy**, 2, pp. 7–33.