International Journal of Geo-Information

MDPI

Article

Increasing the Accuracy of Crowdsourced Information on Land Cover via a Voting Procedure Weighted by Information Inferred from the Contributed Data

Giles Foody [1,*], Linda See [2], Steffen Fritz [2], Inian Moorthy [2], Christoph Perger [2], Christian Schill [3] and Doreen Boyd [1]

1 School of Geography, University of Nottingham, Nottingham NG7 2RD, UK; doreen.boyd@nottingham.ac.uk
2 International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria; see@iiasa.ac.at (L.S.); fritz@iiasa.ac.at (S.F.); moorthy@iiasa.ac.at (I.M.); pergerch@iiasa.ac.at (C.P.)
3 Faculty of Environment and Natural Resources, Albert-Ludwig University, 79085 Freiburg, Germany; christian.schill@felis.uni-freiburg.de
* Correspondence: giles.foody@nottingham.ac.uk; Tel.: +44-115-951-5430

Received: 22 January 2018; Accepted: 21 February 2018; Published: 25 February 2018

Abstract: Simple consensus methods are often used in crowdsourcing studies to label cases when data are provided by multiple contributors. A basic majority vote rule is often used. This approach weights the contributions from each contributor equally but the contributors may vary in the accuracy with which they can label cases. Here, the potential to increase the accuracy of crowdsourced data on land cover identified from satellite remote sensor images through the use of weighted voting strategies is explored. Critically, the information used to weight contributions based on the accuracy with which a contributor labels cases of a class and the relative abundance of class are inferred entirely from the contributed data only via a latent class analysis. The results show that consensus approaches do yield a classification that is more accurate than that achieved by any individual contributor. Here, the most accurate individual could classify the data with an accuracy of 73.91% while a basic consensus label derived from the data provided by all seven volunteers contributing data was 76.58%. More importantly, the results show that weighting contributions can lead to a statistically significant increase in the overall accuracy to 80.60% by ignoring the contributions from the volunteer adjudged to be the least accurate in labelling.

Keywords: crowdsourcing; volunteered geographic information (VGI); ensemble; classification accuracy; latent class analysis

1. Introduction

Members of the general public have for centuries made substantial contributions to science. The inputs range greatly and include the observations of environmental features by an individual and the processing of vast datasets by teams of citizens working in parallel in subjects ranging from astronomy to zoology. Technological developments such as the internet have greatly facilitated the recent strong rise in citizen science activity [1]. Additional technological advances, such as those that have allowed inexpensive and location-aware devices to become commonplace, have been associated with a substantial increase in citizen science activity within geography for which spatial data sets are important. This type of activity has been described in a variety of ways including neogeography, volunteered geographic information, user-generated content, and crowdsourcing [2]. The latter term will be used in this article. Crowdsourcing has become a popular means of acquiring geographic

information. Indeed, the rise of the citizen sensor and growth of volunteered geographic information has revolutionised aspects of contemporary geoinformatics and mapping [3–6]. The power of the crowd has been harnessed in a wide range of mapping applications such as building damage mapping to aid post-disaster humanitarian aid [7,8] through scientific studies of the Earth [9] to the provision of complete open mapping at local to global scales such as OpenStreetMap [10]. Crowdsourcing has greatly changed mapping practice and also allows information that was otherwise impossible or at least impractical to obtain by other means to be acquired. One growth area in geoinformatics has been crowdsourcing as a source of ground reference data on land cover to inform analyses of satellite remote sensing imagery [11]. This is an important and growing application area, with citizens having the potential to provide the ground reference data that are needed to fully exploit the potential of remote sensing as a source of information on land cover.

A major problem with the volunteered geographic information (VGI) on land cover provided by the citizen community is that it can be of variable and typically unknown quality, resulting in concern over data accuracy and fitness for purpose [12–15]. The volunteers providing the data may, for example, vary greatly in their skill and ability to provide accurate class labels. Some contributors may simply be enthusiastic but unskilled while others, and quite commonly so [16], may actually have considerable relevant expertise [17,18]. Nonetheless, the power of the crowd is such that its combined wisdom helps generate a final high quality crowdsourced product.

The collective view of the crowd can be obtained in a variety of ways. Commonly, a simple democratic voting procedure is used to bring together the individual inputs from the volunteers and determine a single crowdsourced view. As such, it is common to find that a consensus or ensemble approach to labelling is used with crowdsourced data [14,15]. In these approaches, the contributions from each volunteer are often equally weighted. While ensemble approaches often appear to work well there are still concerns on the variation in quality of data acquired by citizens [19]. This is apparent in relation to performance relative to other citizens, but also within an individual's own set of contributions as performance might vary within given task. For example, in labelling-based tasks, a volunteer may be able to accurately label a sub-set of the classes but not the rest and so contribute quite differently to another volunteer with a different skill-set. A common concern is that a basic ensemble approach weights each contributor's inputs equally even though the volunteers may be of very different ability. This can give rise to a range of potential problems. For example, one volunteer, who may have considerable relevant expertise, may correctly label a case but this lone voice could be lost among the contradictory labelling provided by less informed members of the crowd who may be very numerous. As such, the composition of the crowd is important [17] and there may be a desire to weight contributions unequally to avoid problems of mob rule.

A variety of ways to facilitate effective use of VGI have been proposed. It is, for example, possible for trusted contributors to act as gatekeepers or to check the credibility of a contribution in relation to its known geographic context [20]. These various approaches to try and assure the quality of VGI are, however, not a panacea. It would, for example, be perfectly possible for a gatekeeper acting in good faith to be a barrier to the provision of accurate information from a new but presently untrusted contributor who actually has more skill and knowledge than the gatekeeper. Other means to try and enhance the quality of VGI have included the acquisition of information on the confidence of labelling. For example, volunteers may be asked when labelling cases to indicate for each one their confidence in the class allocation made [18]. This might then allow cases labelled with considerable uncertainty to be filtered out so that only cases labelled with high confidence are used. However, this type of approach has problems. Volunteers may have inflated views on their ability and in some instances, for example, ignorant people will still confidently label cases [21]. An enhancement of this basic method could be based on the surprisingly popular approach that focuses on labelling that is more popular than predicted [22]. Variations in volunteer performance would still be expected. If, however, this variation could be quantified then it may be possible to use this information to enhance analyses. For example, information on the performance of volunteers in terms of their ability to label cases obtained from the

data may be used to enhance the accuracy of land cover maps [23]. Estimates of volunteer performance could also be used to weight simple voting procedures, perhaps acting to amplify the contributions from volunteers deemed skilled while down-weighting or even ignoring contributions from volunteers deemed to be inaccurate data sources. Thus, it would be possible to recognise that contributions vary in value and seek to weight them unequally within an ensemble approach. In previous work, it was shown that it is possible to characterise the quality of volunteers in terms of the accuracy of their labelling for each class using only the contributed data [24,25]. Here, the aim is to go beyond the characterisation of the quality of the volunteered data and show how this information, and other information inferred from the contributed data, may be used to enhance the final crowdsourced label that may be applied to VGI.

The key aim of this paper is to explore some simple scenarios for enhancing the accuracy of crowdsourced data on land cover obtained via visual interpretation of satellite sensor images provided via an internet based collaborative project. The paper seeks to show that useful information to inform an ensemble classification that employs a weighted voting strategy can be inferred from the volunteered data and this can be used to increase the overall accuracy of the ensemble classification.

2. Data

The data used comprised land cover class labels obtained from a group of volunteers for a set of 299 satellite sensor images of locations selected randomly over the global land mass. These data were acquired via an open call for data collection through the Geo-Wiki project [26,27] and were used in earlier research [24,25]. The data are available for downloading from the PANGAEA repository as documented in [28]. Each volunteer was invited to view the series of satellite sensor images and assign each a land cover label from a defined list of 10 classes: tree cover, shrub cover, herbaceous vegetation/grassland, cultivated and managed, mosaic of cultivated and managed/natural vegetation, regularly flooded/wetland, urban/built-up, snow and ice, barren, and open water. The volunteers were aided in this task by a brief on-line tutorial and no constraints were put upon contribution. An example of the interface used to collect the data is shown in Figure 1.

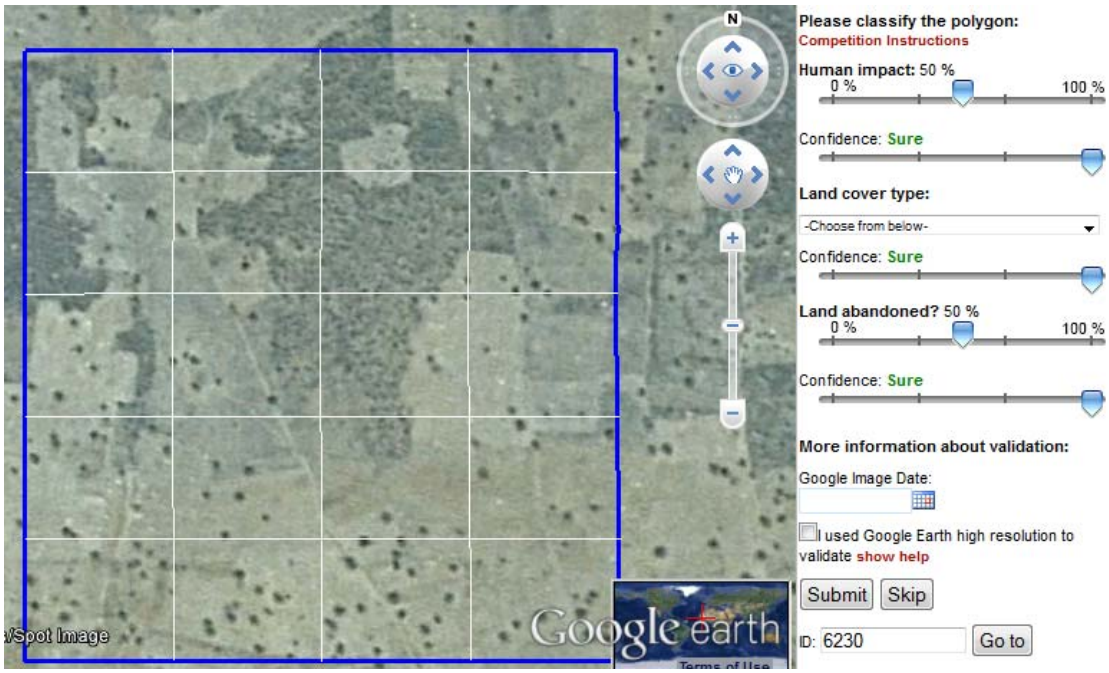

Figure 1. The Geo-Wiki interface used to collect information on land cover type among other features visible from the satellite sensor imagery.

In total, 65 volunteers contributed to the project but their contributions varied greatly in completeness. The amount of images labelled spanned the full spectrum possible, with one volunteer labelling a single image while a few labelled all 299; the average number of images labelled by a volunteer was approximately 110 images. Here, attention is focused on the labels provided by the 10 volunteers who labelled most if not all of the 299 images; these 10 volunteers labelled at least 289 images each. Consequently, this group of volunteers annotated broadly the same set of images reducing the potential for problems such as optimistic bias in their labelling that could occur by skipping the complex to label cases and focusing on only the easier images. The focus on a relatively small group of volunteers is also in keeping with suggestions in the literature [15,25] as well as a means of balancing the competing pressures of seeking multiple annotations but wishing to label many cases [29].

Although a key focus of this article is on information obtained directly from the crowdsourced data without any independent reference data set, a reference data set was formed to help demonstrate and confirm the approach used. Thus, a reference data set was generated simply to confirm the value of the approaches to be adopted, ensuring that the results and interpretations are credible. Three of the 10 selected contributors were experts who also revisited the entire set of 299 images to derive a ground reference data set after discussion amongst themselves informed by their own set of labellings. Although this reference data set is unlikely to be perfect and represent a true gold standard reference which can lead to misestimation [30] it is, however, of a type that is common in major mapping programmes (e.g., [31]). These ground reference data were used to assess the accuracy of the labelling generated from the data contributed by the remaining seven volunteers. This approach reduced the potential for complications caused by missing data and meant that for most of the 299 images, a set of seven class labels were defined. Each label was treated here as a vote for the relevant class and used in simple ensemble methods to obtain a single crowdsourced land cover class label for each image. To maintain anonymity these seven volunteers were labelled A–G.

3. Methods

The work focuses on four scenarios. The first scenario is a benchmark test of the value of the crowd. In this, the accuracy with which individual volunteers classified the images is compared to the accuracy of the classification obtained from the volunteers as a whole using a basic majority voting approach to label each image from the set of labels generated for it by the seven volunteers. Here, accuracy was measured relative to the reference data set generated from the three expert contributors and expressed as the percentage of correctly allocated cases.

All additional analyses sought to use information inferred from the data contributed by the volunteers to weight the voting procedure. Here, the weighting focused on the skill of the volunteers in terms of their ability to label each class and on the relative abundance of the classes in the data set. Information on both of the latter variables was inferred from the results of a latent class analysis of the volunteered data.

The latent class analysis uses the observed data contributed by the volunteers to provide information on an unobserved or latent variable which in this case is the actual land cover. A standard latent class model to describe the relationship between the observed and latent variables was used and can be written as

$$f(y_i) = \sum_{x=1}^{C} P(x) \prod_{v=1}^{V} f(y_{iv}|x)$$

where $f(y_i)$ is a vector representing the complete set of responses obtained from the V volunteers ($1 < v < V$) contributing data for the case i, C is the number of classes, and x the latent variable [32,33]. Assuming that the model is found to fit with the observed data, the parameters of this model provide the information to inform weighted voting approaches. Specifically, the $f(y_{it}|x)$ parameters of the model represent the conditional probabilities of class membership. Thus, for example, these model parameters indicate the conditional probability that a case allocated a class label by a volunteer is

actually a member of that class; in the geoinformatics community, this probability is often referred to as the producer's accuracy for the specified class. Critically, for each volunteer, it is possible to obtain a conditional probability of class membership for each class, indicating the volunteer's skill in labelling each class. The average conditional probability calculated over all classes was also used as a measure of the volunteer's overall skill. In addition, the other latent class model parameter, $P(x)$, indicates the prevalence or abundance of the classes. A feature to note here is that the information on both volunteer skill and class abundance is inferred from only the contributed data.

The information on per-class producer's accuracy for each class and each volunteer could be used to weight the contributions from the volunteers. Of the many ways to approach this task, in Scenario 2 any label (i.e., vote) for a class from a volunteer whose accuracy in labelling of that class was estimated to be substantially less than the maximum accuracy observed for that class was deleted. Here, the focus was on instances for which there was a very large difference in the accuracy relative to that observed for the most accurate volunteer. The approach was implemented here by ignoring the label provided by a volunteer if that volunteer's estimated accuracy for that specific class, rounded to a whole number, was more than 30% less than the highest estimated accuracy for that class associated with another volunteer. This, in effect, was seeking to determine if removing votes from volunteers known to be inaccurate on a specific class would help the overall labelling task. Note that while the labels for a class may be ignored, the other class labels provided by a volunteer would still be used, it is only the labels for class(es) on which the volunteer's performance was viewed as insufficiently high that are removed.

In Scenario 3, the entire contribution from a volunteer with low overall accuracy, expressed as the mean of the producer's accuracy estimated over all classes, were down-weighted to zero by their removal. In essence this was seeking to explore the effect of 'silencing' an inaccurate contributor. Here, this was undertaken twice: the contributions from the volunteer deemed least accurate were removed (Scenario 3a) and the contributions from the two volunteers deemed least accurate were removed (Scenario 3b).

The measure of overall accuracy used in Scenario 3 weights each class equally but accounts for variations in class abundance could further enhance the analysis. This approach would, for example, reduce the effect of poor performance on classes that are rare and so have little impact on the overall proportion of cases correctly classified. Given this context, Scenario 4 sought to extend the analysis one step further and weight the per-class producer's accuracy values estimated for the volunteers by class abundance information estimated from the latent class model. Here, the contributions from the most inaccurate contributor were again removed. In addition, the research sought to explore the effect of magnifying the input of the most accurate contributor, here achieved by duplicating their contributions, effectively making a vote count twice. This weighting is relatively arbitrary and different results could be expected at other settings. In total three different approaches were explored: the magnification of the contributions of most accurate contributor (Scenario 4a), the magnification of the contributions from the most accurate contributor and the removal of the data from the least accurate contributor (Scenario 4b) and the removal of the contributions from least accurate contributor (Scenario 4c).

The overall accuracy of a crowdsourced set of class labels was expressed as the percentage of cases whose labelling agreed with that in the reference data set. The statistical significance of differences in overall accuracy was calculated using the McNemar test. The latter focuses on the discordant cases, the cases which were allocated correctly in only one of the pair of classifications compared. The test is based on the normal curve deviate, z, and the null hypothesis of no significant difference is rejected if the value of z obtained is greater than the critical value of $|1.96|$; the sign is important for a hypothesis with a directional component for which the critical value of z at the 95% level of confidence is 1.645.

4. Results and Discussion

A reference data set, to be used purely for illustrative purposes and ensure credibility of the results, was obtained from the three expert contributors who allocated labels after reaching a consensus.

The labelling from these contributors showed moderate levels of pairwise agreement (with 66.6–69.9% pairwise agreement; kappa coefficients varied from 0.55–0.61) and final class allocations were made after discussion amongst the experts informed by their own initial labelling. It was apparent that the classes varied greatly in abundance. Two classes (regularly flooded/wetland and snow and ice) were determined to be absent in the reference data set, although some cases were sometimes incorrectly labelled as belonging to these classes.

The accuracy with which each volunteer classified the set of satellite sensor images is highlighted in Table 1. The accuracy of the classifications from each and every volunteer was less than that obtained by combining their contributions with a simple majority vote approach. The most accurate individual, for example, provided a set of labels with an overall accuracy of 73.91% while the ensemble classification obtained via the use of the majority vote procedure applied to the volunteered data had an accuracy of 76.58%. This result confirms the oft-stated view that the crowd can be more accurate than the individuals in it.

Table 1. Per-class and overall classification accuracies (%) for the seven volunteers obtained from the latent class model

Class	A	B	C	D	E	F	G
Tree cover (T)	100	86.27	74.73	62.60	73.23	67.43	66.51
Shrub cover (S)	64.44	74.54	83.47	71.13	50.81	69.65	60.61
Herbaceous vegetation/Grassland (H)	69.54	71.22	73.27	45.03	64.65	47.52	24.79
Cultivated and managed (C)	94.16	92.66	100	70.31	87.14	20.17	91.82
Mosaic (M)	54.87	73.8	95.34	97.75	67.9	64.74	67.5
Regularly flooded/wetland (R)	0	0	0	0	0	0	0
Urban/built-up (U)	50	25	50	50	50	50	25
Snow and ice (I)	0	0	0	0	0	0	0
Barren (B)	38.7	0	11.99	0	50.9	30.25	0
Open water (O)	25	25	25	25	25	25	25
Overall (mean)	49.67	44.85	51.38	42.18	46.96	37.48	36.12
Overall (mean weighted by class size)	59.1	59.27	64.86	51.98	55.12	34.93	50.57

Although the simple majority voting approach provided a basic ensemble approach to classification that was more accurate than its component parts, the testing of the three other scenarios sought to explore the possibility to raise the accuracy of the crowdsourced labelling further by weighting the contributions from the volunteers, notably by their skill or accuracy inferred from the latent class analysis.

The estimates of producer's accuracy obtained from the latent class analysis for each volunteer with regard to each class (Table 1) highlight that volunteers vary greatly in their skill and ability to label the imagery. In addition to the variation between volunteers there was variation in the accuracy of classes within the set of data contributed by the volunteers. For example, it was evident that an individual could be very highly accurate with regard to one class but inaccurate with another. For example, Volunteer A had estimated accuracy values of 100% and 54.87% for the tree cover and mosaic classes. In relation to the latter, note also that Volunteer D's estimated accuracy values were almost the direct opposite with 62.60% and 97.75% for the tree cover and mosaic classes, respectively. In addition, it was evident that a volunteer with generally low accuracy could still be highly accurate on a specific class. This was evident for Volunteer G who was only highly accurate on one class: cultivated and managed, which also was a relatively abundant class.

In Scenario 2, the vote for a class by a volunteer was removed if that volunteer was highly inaccurate in the labelling of that specific class in comparison to the other volunteers. The effect of removing the votes for a class from a volunteer deemed to be unskilled for the labelling of that class increased the accuracy of the overall ensemble approach using the majority voting procedure to 78.26%.

An alternative approach to using the estimated information on volunteer labelling accuracy is to remove all contributions from volunteers adjudged to provide labels of low or insufficient accuracy. This was explored in Scenario 3. It was evident in Scenario 3a that by dropping the entire set of contributions of the least accurate volunteer (Volunteer G, with a mean producer's accuracy of 36.12%)

the accuracy of the ensemble classification could increase to 77.92%. Moreover, the largest ensemble accuracy observed in Scenario 3, 79.59%, was obtained in Scenario 3b when the contributions from the two least accurate volunteers (Volunteers F and G) were ignored. It was also evident that the accuracy of the contributions by these two volunteers were noticeably less accurate than from the other volunteers (Table 1).

In addition to information on the accuracy with which each volunteer can classify the classes, the latent class model also indicates the prevalence or abundance of the classes. This information on class abundance inferred from the analysis was used to adjust the estimates of overall volunteer accuracy, here expressed as the average producer's accuracy. The weighted overall accuracy values (Table 1) revealed that one volunteer (Volunteer C) was noticeably more accurate and one noticeably less accurate (Volunteer F) than the remaining set; note that after weighting for class abundance Volunteer F rather than G is associated with the lowest labelling accuracy. Increasing the weight of the accurate volunteer by duplicating their contributions (i.e., giving each vote a weight of two) increased accuracy. For example, increasing the vote for the most accurate volunteer in Scenario 4a raised the accuracy of the ensemble from the benchmark value of 76.58% to 78.59%. Furthermore, ignoring the labels from the least accurate volunteer in addition further increased accuracy to 79.59% in Scenario 4b. However, it was also apparent that a more accurate ensemble could be achieved in Scenario 4c by solely removing the contributions of the least accurate volunteer, which yielded an ensemble classification with an accuracy of 80.60%. It should be noted that at the 95% level of confidence, this latter ensemble classification was also significantly more accurate than that achieved by increasing the weighting for the most accurate volunteer ($z = 4.31$) and by additionally ignoring the least accurate volunteer's data ($z = 3.90$). The ensemble classification arising through the removal of the contributions from the least accurate volunteer (Scenario 4c) was also the most accurate of all classifications reported in the study and significantly different at the 95% level of confidence to the benchmark classification based on the standard majority voting rule ($z = 5.54$).

The ability to increase the accuracy of the crowdsourced labelling by weighting the voting process is highlighted in Figure 2 which shows the overall accuracy of classifications relative to the reference data for individuals and from each of the four scenarios for ensemble classification discussed. Additional summary data for each of the classifications arising from the scenarios reported is provided in Table 2 and the full confusion matrix provided for the classifications arising from the basic ensemble (Table 3) and Scenario 4c (Table 4).

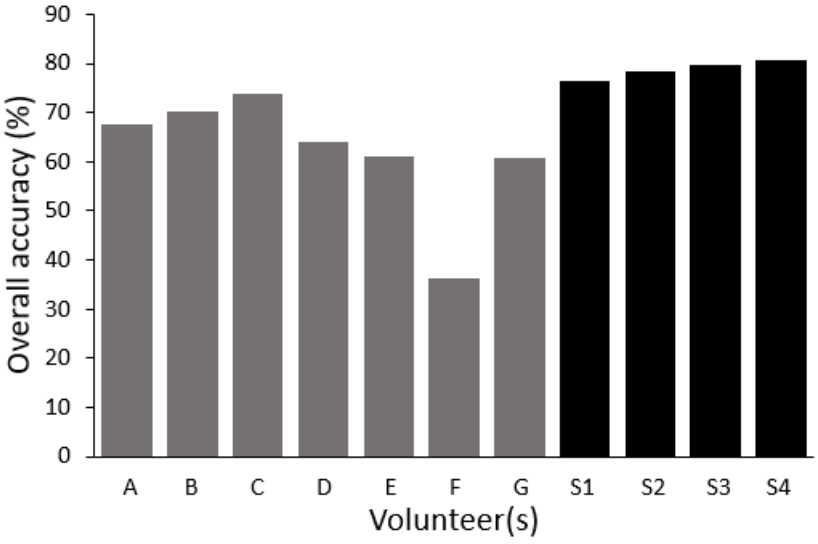

Figure 2. Overall classification accuracy determined relative to the reference data set for each of the individual volunteers (grey bars) and the highest accuracy from each of the four scenarios for an ensemble classification (S1–S4, black bars).

Table 2. Summary of the correct allocations observed in each scenario (S1–S4c) reported for each class and the class size (n) in the reference data set; complete confusion matrices for two key classifications are given in Tables 3 and 4.

Class	n	S1	S2	S3a	S3b	S4a	S4b	S4c
T	47	35	32	34	36	34	35	35
S	20	15	15	15	15	16	16	16
H	24	17	19	16	17	16	16	17
C	119	98	108	99	101	104	106	106
M	85	60	55	65	65	61	61	63
R	0	0	0	0	0	0	0	0
U	1	1	1	1	1	1	1	1
I	0	0	0	0	0	0	0	0
B	2	2	2	2	2	2	2	2
O	1	1	1	1	1	1	1	1
Total	299	229	233	233	238	235	238	241

Table 3. Confusion matrix for the benchmark classification of Scenario 1; columns show the class label in the reference data set and rows the label determined in the scenario.

Class	T	S	H	C	M	R	U	I	B	O	Total
T	35	1	1	0	10	0	0	0	0	0	47
S	8	15	3	0	2	0	0	0	0	0	28
H	2	3	17	0	6	0	0	0	0	0	28
C	0	0	0	98	5	0	0	0	0	0	103
M	2	0	1	20	60	0	0	0	0	0	83
R	0	0	0	0	0	0	0	0	0	0	0
U	0	0	0	0	1	0	1	0	0	0	2
I	0	0	0	0	0	0	0	0	0	0	0
B	0	1	2	1	1	0	0	0	2	0	7
O	0	0	0	0	0	0	0	0	0	1	1
Total	47	20	24	119	85	0	1	0	2	1	299

Table 4. Confusion matrix for the classification of Scenario 4c; columns show the class label in the reference data set and rows the label determined in the scenario.

Class	T	S	H	C	M	R	U	I	B	O	Total
T	35	1	1	0	7	0	0	0	0	0	44
S	8	16	3	0	4	0	0	0	0	0	31
H	2	3	17	1	3	0	0	0	0	0	26
C	1	0	0	106	7	0	0	0	0	0	114
M	1	0	1	12	63	0	0	0	0	0	77
R	0	0	1	0	0	0	0	0	0	0	1
U	0	0	0	0	1	0	1	0	0	0	2
I	0	0	0	0	0	0	0	0	0	0	0
B	0	0	1	0	0	0	0	0	2	0	3
O	0	0	0	0	0	0	0	0	0	1	1
Total	47	20	24	119	85	0	1	0	2	1	299

Figure 2 highlights that each ensemble approach yielded a classification that was more accurate than that arising from the individual contributors alone. It also highlights that the relative accuracy of the classifications weighted by class abundance obtained from the individuals inferred from the latent class analysis (Table 1) corresponds with the actual accuracy assessed relative to the reference data (Figure 2). In particular, the relatively low accuracy of the labelling provided by Volunteer F is evident and it is the removal of these data in Scenario 4c that resulted in the largest, and statistically significant, increase in accuracy over the benchmark classification of Scenario 1. From earlier research [25],

the accuracy with which the data contributed by each of the volunteers may be characterised could increase if the number of volunteers also increased, paving the way for further refinement of the analysis.

The results, especially from Scenario 4c, show that, for the data set used, the removal of inaccurate data is of more value than the enhancement or amplification of more accurate data sources. It should be noted that this latter issue may reflect the composition of the volunteers used in this study. Given that all seven volunteers had contributed labels for virtually all of the images, it may be that these people have a high level of motivation which could be used as a proxy variable to indicate high skill sets so it was the removal of the occasional anomalously poor inputs that was important rather than efforts to amplify good quality contributions. Had the set of volunteers been of more mixed ability, notably if made up of a large number of true amateurs, then less expertise might be present and different trends may have been observed. Similarly, it should be noted that the results may, of course, be specific to the data set used and the information inferred could be used in other ways (e.g., to inform labelling in tie-break situations by allocating to the class indicated by the relatively more accurate labellers).

Finally, it should be stressed that the information on volunteer skill and class abundance to weight the voting procedure were all inferred from the contributed data alone. In many applications there may be little or no reference data available to allow a standard assessment of the accuracy of labelling and comparison of classifications such as that provided by Figure 2. Critically, however, all of the information contained in Table 1 was obtained from the set of contributed crowdsourced labels only; this includes the information on class size which was obtained from the latent class model. Thus, the information on per-class and overall classification accuracy needed to enhance the voting method is inferred entirely from just the contributed data; the reference data were only used in this study to provide supporting evidence that the approaches discussed actually did impact on accuracy. The quality of the crowd-sourced estimates may also increase if data from additional volunteers are available [25]. As well as providing an intrinsic approach to the assessment of contributed data quality, the approach has additional advantages. Since only the contributed data are required, there is, therefore, no need to use a proportion of the crowdsourced data to measure the variables directly, perhaps via some dedicated ground based research or use of additional experts. There is also no use of external auxiliary information. Further enhancements could be made by expressing skill in different ways; the measure of accuracy used may not always be ideal and other approaches could be used to focus more directly on the objectives of a specific study (e.g., weighting by unequal costs of errors). Critically, however, this article has gone beyond earlier work to show that the quality of contributed data can be estimated from the data alone to demonstrate how crowdsourced labelling can be enhanced via simple weighted voting methods without any additional data.

5. Conclusions

The results have highlighted that the wisdom of the crowd can be used to generate a single crowdsourced set of land cover annotations that are more accurate than those achieved by any individual in the crowd. More importantly, estimates of the skill of each individual in terms of classifying classes and on the abundance of the classes that were inferred from the contributed data may be used to increase the accuracy of the crowdsourced labels; reference data were used here to confirm the validity of the approach, but are not required for its implementation. In this study, a significant increase in the accuracy of labelling of land cover from satellite sensor imagery was obtained by down-weighting the contributions adjudged to be of relatively low overall accuracy for the task. This was most apparent when the estimation of the volunteer's skill, expressed here as the average producer's accuracy calculated over all classes, was weighted by class abundance. It is evident that very simple methods may be used to increase the quality of crowdsourced data which should hopefully further facilitate the use of crowdsourcing of geographic data.

Acknowledgments: This work benefitted from funding from the EU COST Action TD1202 and the EU Horizon 2020 funded project LandSense (No. 689812) as well as founding work funded by the British Academy (reference SG112788) and EPSRC (reference EP/J0020230/1). We are also grateful to the editor and the two referees who provided constructive comments to enhance this article.

Author Contributions: G.F., L.S., S.F., and I.M. discussed the idea; L.S., S.F., C.P., C.S., and D.S. contributed underpinning research; G.F. undertook the analyses and wrote the paper with inputs from all authors.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bonney, R.; Shirk, J.L.; Phillips, T.B.; Wiggins, A.; Ballard, H.L.; Miller-Rushing, A.J.; Parrish, J.K. Next steps for citizen science. *Science* **2014**, *343*, 1436–1437. Available online: http://science.sciencemag.org/content/343/6178/1436 (accessed on 12 February 2018). [CrossRef] [PubMed]
2. See, L.; Mooney, P.; Foody, G.; Bastin, L.; Comber, A.; Estima, J.; Fritz, S.; Kerle, N.; Jiang, B.; Laakso, M.; et al. Crowdsourcing, citizen science or Volunteered Geographic Information? The current state of crowdsourced geographic information. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 55. Available online: http://www.mdpi.com/2220-9964/5/5/55 (accessed on 12 February 2018). [CrossRef]
3. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. Available online: https://link.springer.com/article/10.1007/s10708-007-9111-y (accessed on 11 December 2017). [CrossRef]
4. Pullar, D.; Hayes, S. Will the future maps for Australia be published by 'nobodies'. *J. Spat. Sci.* **2017**, 1–8. Available online: http://www.tandfonline.com/doi/abs/10.1080/14498596.2017.1361873 (accessed on 11 December 2017). [CrossRef]
5. Capineri, C.; Haklay, M.; Huang, H.; Antoniou, V.; Kettunen, J.; Ostermann, F.; Purves, R. *European Handbook of Crowdsourced Geographic Information*; Ubiquity Press: London, UK, 2016. Available online: https://doi.org/10.5334/bax (accessed on 12 February 2018).
6. Foody, G.; See, L.; Fritz, S.; Mooney, P.; Olteanu-Raimond, A.M.; Fonte, C.C.; Antoniou, V. *Mapping and the Citizen Sensor*; Ubiquity Press: London, UK, 2017. Available online: https://doi.org/10.5334/bbf (accessed on 12 February 2018).
7. Goodchild, M.F.; Glennon, J.A. Crowdsourcing geographic information for disaster response: A research frontier. *Int. J. Dig. Earth* **2010**, *3*, 231–241. Available online: http://www.tandfonline.com/doi/full/10.1080/17538941003759255 (accessed on 11 December 2017). [CrossRef]
8. Kerle, N.; Hoffman, R.R. Collaborative damage mapping for emergency response: The role of Cognitive Systems Engineering. *Nat. Hazards Earth Syst. Sci.* **2013**, *13*, 97–113. [CrossRef]
9. Goodchild, M.F.; Guo, H.; Annoni, A.; Bian, L.; de Bie, K.; Campbell, F.; Craglia, M.; Ehlers, M.; van Genderen, J.; Jackson, D.; et al. Next-generation digital earth. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 11088–11094. Available online: http://www.pnas.org/content/109/28/11088.full (accessed on 11 December 2017). [CrossRef] [PubMed]
10. Mooney, P.; Minghini, M. A review of OpenStreetMap data. In *Mapping and the Citizen Sensor*; Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C.C., Antoniou, V., Eds.; Ubiquity Press: London, UK, 2017; pp. 37–59, ISBN 978-1-911529-17-0. Available online: https://doi.org/10.5334/bbf (accessed on 11 December 2017).
11. Foody, G.M. Citizen science in support of remote sensing research. In *International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2015; pp. 5387–5390. Available online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7326952 (accessed 24 February 2018).
12. Flanagin, A.J.; Metzger, M.J. The credibility of volunteered geographic information. *GeoJournal* **2008**, *72*, 137–148. Available online: https://link.springer.com/article/10.1007/s10708-008-9188-y (accessed on 11 December 2017). [CrossRef]
13. Koswatte, S.; McDougall, K.; Liu, X. VGI and crowdsourced data credibility analysis using spam email detection techniques. *Int. J. Dig. Earth* **2017**, 1–13. Available online: http://www.tandfonline.com/doi/abs/10.1080/17538947.2017.1341558 (accessed on 11 December 2017). [CrossRef]

14. Fonte, C.C.; Antoniou, V.; Bastin, L.; Bayas, L.; See, L.; Vatseva, R. Assessing VGI data quality. In *Mapping and the Citizen Sensor*; Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C.C., Antoniou, V., Eds.; Ubiquity Press: London, UK, 2017; pp. 137–163, ISBN 978-1-911529-17-0. Available online: https://doi.org/10.5334/bbf (accessed on 11 December 2017).
15. Haklay, M.; Basiouka, S.; Antoniou, V.; Ather, A. How many volunteers does it take to map an area well? The validity of Linus' Law to volunteered geographic information. *Cartogr. J.* **2010**, *47*, 315–322. Available online: http://www.tandfonline.com/doi/abs/10.1179/000870410X12911304958827 (accessed on 7 December 2017). [CrossRef]
16. Brabham, D.C. The myth of amateur crowds: A critical discourse analysis of crowdsourcing coverage. *Inf. Commun. Soc.* **2012**, *15*, 394–410. Available online: http://www.tandfonline.com/doi/abs/10.1080/1369118X.2011.641991 (accessed on 11 December 2017). [CrossRef]
17. Comber, A.; Mooney, P.; Purves, R.S.; Rocchini, D.; Walz, A. Crowdsourcing: It matters who the crowd are. The impacts of between group variations in recording land cover. *PLoS ONE* **2016**, *11*, e0158329. Available online: https://doi.org/10.1371/journal.pone.0158329 (accessed on 5 December 2017). [CrossRef] [PubMed]
18. See, L.; Comber, A.; Salk, C.; Fritz, S.; van der Velde, M.; Perger, C.; Schill, C.; McCallum, I.; Kraxner, F.; Obersteiner, M. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLoS ONE* **2013**, *8*, e69958. Available online: https://doi.org/10.1371/journal.pone.0069958 (accessed on 5 December 2017). [CrossRef] [PubMed]
19. Salk, C.F.; Sturn, T.; See, L.; Fritz, S. Limitations of majority agreement in crowdsourced image interpretation. *Trans. GIS* **2017**, *21*, 207–223. [CrossRef]
20. Goodchild, M.F.; Li, L. Assuring the quality of volunteered geographic information. *Spat. Stat.* **2012**, *1*, 110–120. Available online: http://www.sciencedirect.com/science/article/pii/S2211675312000097 (accessed on 7 December 2017). [CrossRef]
21. Kruger, J.; Dunning, D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* **1999**, *77*, 1121–1134. Available online: http://dx.doi.org/10.1037/0022-3514.77.6.1121 (accessed on 5 December 2017). [CrossRef] [PubMed]
22. Prelec, D.; Seung, H.S.; McCoy, J. A solution to the single-question crowd wisdom problem. *Nature* **2017**, *541*, 532–535. Available online: https://www.nature.com/articles/nature21054 (accessed on 7 December 2017). [CrossRef] [PubMed]
23. Gengler, S.; Bogaert, P. Integrating crowdsourced data with a land cover product: A Bayesian data fusion approach. *Remote Sens.* **2016**, *8*, 545. Available online: http://www.mdpi.com/2072-4292/8/7/545/htm (accessed on 12 February 2018). [CrossRef]
24. Foody, G.M.; See, L.; Fritz, S.; Van der Velde, M.; Perger, C.; Schill, C.; Boyd, D.S. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Trans. GIS* **2013**, *17*, 847–860. Available online: http://onlinelibrary.wiley.com/doi/10.1111/tgis.12033/full (accessed on 7 December 2017). [CrossRef]
25. Foody, G.M.; See, L.; Fritz, S.; Van der Velde, M.; Perger, C.; Schill, C.; Boyd, D.S.; Comber, A. Accurate attribute mapping from volunteered geographic information: Issues of volunteer quantity and quality. *Cartogr. J.* **2015**, *52*, 336–344. Available online: http://www.tandfonline.com/doi/abs/10.1080/00087041.2015.1108658 (accessed on 7 December 2017). [CrossRef]
26. Fritz, S.; McCallum, I.; Schill, C.; Perger, C.; See, L.; Schepaschenko, D.; van der Velde, M.; Kraxner, F.; Obersteiner, M. Geo-Wiki: An online platform for improving global land cover. *Environ. Modell. Softw.* **2012**, *31*, 110–123. Available online: http://www.sciencedirect.com/science/article/pii/S1364815211002787, (accessed on 11 December 2017). [CrossRef]
27. See, L.; Fritz, S.; Perger, C.; Schill, C.; McCallum, I.; Schepaschenko, D.; Duerauer, M.; Sturn, T.; Karner, M.; Kraxner, F.; et al. Harnessing the power of volunteers, the internet and Google Earth to collect and validate global spatial information using Geo-Wiki. *Technol. Forecast. Soc. Chang.* **2015**, *98*, 324–335. [CrossRef]
28. Fritz, S.; See, L.; Perger, C.; McCallum, I.; Schill, C.; Schepaschenko, D.; Duerauer, M.; Karner, M.; Dresel, C.; Laso-Bayas, J.-C.; et al. A global dataset of crowdsourced land cover and land use reference data. *Sci. Data* **2017**, *4*, 170075. [CrossRef] [PubMed]
29. Boyd, D.; Jackson, B.; Wardlaw, J.; Foody, G.; Marsh, S.; Bales, K. Slavery from space: Demonstrating the role for satellite remote sensing to inform evidence-based action related to UN SDG number 8. *ISPRS J. Photogramm. Remote Sens.* **2018**, in press.

30. Foody, G.M. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sens. Environ.* **2010**, *114*, 2271–2285. Available online: https://www.sciencedirect.com/science/article/pii/S0034425710001434 (accessed on 12 February 2018). [CrossRef]
31. Scepan, J.; Menz, G.; Hansen, M.C. The DISCover validation image interpretation process. *Photogramm. Eng. Remote Sens.* **1999**, *65*, 1075–1081. Available online: https://www.asprs.org/wp-content/uploads/pers/1999journal/sep/1999_sept_1075-1081.pdf (accessed on 11 December 2017).
32. Vermunt, J.K.; Magidson, J. Latent class analysis. In *The Sage Encyclopedia of Social Science Research Methods*; Lewis-Beck, M., Bryman, A.E., Liao, T.F., Eds.; Sage Publications: Thousand Oaks, CA, USA, 2003; Volume 2, pp. 549–553.
33. Vermunt, J.K.; Magidson, J. Latent class models for classification. *Comput. Stat. Data Anal.* **2003**, *41*, 531–537. Available online: http://www.sciencedirect.com/science/article/pii/S0167947302001792 (accessed on 7 December 2017). [CrossRef]