Assessing quality of volunteer crowdsourcing contributions:

Lessons from the Cropland Capture game

*Carl F Salk[1,2], Tobias Sturn[1], Linda See[1], Steffen Fritz[1], Christoph Perger[1]

(1) Ecosystems Services and Management Program, International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria

(2) Southern Swedish Forest Research Center, Swedish University of Agricultural Sciences, Box 52, S-23053 Alnarp, Sweden

*Corresponding author: salk@iiasa.ac.at; +43(0) 2236 807 293

Other authors' contact information:

TS: sturn@iiasa.ac.at; +43(0) 2236 807 518

LS: see@iiasa.ac.at; +43(0) 2236 807 423

SF: fritz@iiasa.ac.at; +43(0) 2236 807 353

CP: pergerch@iiasa.ac.at; +43(0) 2236 807 357

Word count: 7704

Abstract

Volunteered Geographical Information (VGI) is the assembly of spatial information based on public input. While VGI has proliferated in recent years, assessing the quality of volunteer-contributed data has proven challenging, leading some to question the efficiency of such programs. In this paper, we compare several quality metrics for individual volunteers' contributions. The data was the product of the 'Cropland Capture' game, in which several thousand volunteers assessed 165,000 images for the presence of cropland over the course of six months. We compared agreement between volunteer ratings and an image's majority classification with volunteer self-agreement on repeated images and expert evaluations. We also examined the impact of experience and learning on performance. Volunteer self-agreement was nearly always higher than agreement with majority classifications, and much greater than agreement with expert validations, although these metrics were all positively correlated. Volunteer quality showed a broad trend toward improvement with experience, but the highest accuracies were achieved by a handful of moderately active contributors, not the most active volunteers. Our results emphasize the importance of a universal set of expert-validated tasks as a gold standard for evaluating VGI quality.

Keywords: crowdsourcing; volunteered geographic information; cropland; data quality; image classification; Geo-Wiki

Introduction

Citizen involvement in natural and social science data collection and processing has grown quickly over the last decade. Although citizens have been involved in scientific research and conservation activities for a considerable time now (Miller-Rushing, Primack and Bonney

2012), the recent widespread participation has been facilitated through the interactivity of Web 2.0, open access to high resolution satellite imagery and the proliferation of mobile devices which can record photographic and location-based information. Citizen science has benefitted from Digital Earth innovations which provide geographical context and base data for citizen tasks. Citizen science can in turn enhance Digital Earth products, for instance by refining maps with tasks that machine learning algorithms struggle to perform. However, accompanying the growth of citizen science has been a growth in the need for tools to evaluate it and in the need for potentially costly expert validations as an external standard for evaluating the quality of such data. In this paper, we address several pressing questions about citizen science data evaluation and provide guidance not just on how best to evaluate data after collection, but also on how to guide its collection so that eventual analysis can be carried out more effectively.

There are many terms in the literature to refer to this citizen involvement in science. These including 'crowdsourcing' (Howe 2006) which is often used for commercial micro tasks, 'volunteered geographic information' (VGI; Goodchild 2007) for the collection of georeferenced information, and 'citizen science' (Bonney *et al.* 2009) which is the broader involvement of citizens in a range of scientific activities from data collection to data analysis and research design. However, regardless of the specific terminology, all of these activities share in common the distributed completion of small, clearly-defined tasks.

Citizen science projects have successfully contributed to many fields of research such as the classification and discovery of new galaxies via Galaxy Zoo (Clery 2011), the identification of bird species via eBird (Sullivan *et al.* 2014), understanding the three-dimensional structures of proteins through the Fold-It game (Khatib *et al.* 2011) and land-cover classification of satellite imagery in Geo-Wiki, an application built using the Google Earth API (Fritz *et al.* 2009, 2012). Yet the quality of data collected by non-specialists remains an

overarching concern (Flanagin and Metzger 2008), especially given the desire to integrate citizen-collected data with more authoritative sources (Coleman 2013) and the expected growth in this source of data in the future (See, Fritz and de Leeuw 2013). Data quality considerations depend on the specific goal of the assigned task, such as whether they are from a commercial crowdsourcing platform such as Amazon Turk, for species identification and environmental monitoring, or for the collection of VGI, where the emphasis is on the spatial aspects of data and the mapping of objects. In this paper, we focus on quality considerations in a simple VGI task, although the results presented here are relevant to many other types of simple crowdsourced tasks.

Data quality of VGI can be assessed via a number of different attributes. These include credibility of the data based on the existence of metadata or past performance of the contributors, the positional accuracy of data, the thematic quality or the tags associated with georeferenced objects, the spatial and attribute completeness of the data, how up-to-date the data is, and the logical consistency of data (Fonte *et al.* Submitted). Many recent studies have focussed on quality assessment of OpenStreetMap, a very successful VGI initiative to map many different types of features such as roads and points of interest around the world (Ramm, Topf and Chilton 2011). These studies have mainly examined the positional accuracy (Haklay 2010, Haklay *et al.* 2010, Neis, Zielstra and Zipf 2011, Canavosio-Zuzelski, Agouris and Doucette 2013), the completeness (Haklay 2010, Neis, Zielstra and Zipf 2011, Hecht, Kunze and Hahmann 2013) and the currency (Jokar Arsanjani *et al.* 2013) of OpenStreetMap, with the study by Girres and Touya (2010) covering a broader range of accuracy assessment measures of French OpenStreetMap data. Other studies have examined the thematic quality of image classifications in Geo-Wiki (Comber *et al.* 2013, Foody *et al.* 2013, See, Comber, *et al.* 2013) and shown varying levels of performance across contributors and across land cover types. However, one of the issues with this dataset was not having sufficient data from multiple

contributors at individual locations to develop statistically robust relationships between contributor performance, land cover type and other factors such as image resolution, location, etc. For this reason, we developed a simplified game version of Geo-Wiki called 'Cropland Capture' in which many images were rated by multiple contributors over a 6 month period.

There are many approaches to quality control for volunteer-contributed data. Allahbakhsh et al. (2013) group these into eight classes. Some of these categories involve managing tasks to reduce the risk of poor-quality work (e.g. 'real-time support' and 'workflow management'). Others, like 'contributor evaluation' assess the quality of a reviewer and assign this rating to all of their work. However, if no information other than the reviewer's work is available for this assessment, then this quickly becomes a chicken and egg problem. The remaining five approaches directly assess data quality, rather than managing it or using proxies. However, some of these approaches are essentially the same, at least in the context of land-cover validation. 'Expert review' is necessarily the source of gold standards for 'ground truth'. 'Output agreement' is simply 'majority consensus' between two workers. The final category, 'input agreement,' is defined as 'Independent workers receive an input and describe it to each other. If they all decided its' a same input, it's accepted as a quality answer' (Allahbakhsh et al., 2013). In our view, this approach is not relevant to land-cover classification. If two workers are independently given images and rate them as cropland, they could be analysing the same image, but this certainly doesn't prove it. Thus, the list of Allahbakhsh et al. (2013) contains two approaches, 'expert review' and 'majority consensus' that are relevant to land-cover validation. To this, we add a third measure: the consistency of a volunteer with their previous ratings when an image is rated additional times.

In this paper, we use a simple binary crowdsourcing task to assess how best to evaluate volunteer quality and accuracy and use these results to provide guidance on the design of

online games for VGI and other types of tasks. While schemes have long been sought for evaluation of task accuracy in the absence of external/expert validation (Dawid and Skene, 1979; Bachrach et al., 2012, Digital Globe, 2014), it is thought that, at least in some cases, reference data is required for true standardization of volunteer-contributed data (Bird et al., 2014). We show that the latter view is true, at least in our example of land-cover validation, by turning the approaches outlined in the previous paragraph into quantitative metrics and comparing their performance. Before we address these questions, we begin with an overview of the Cropland Capture game and the data that were collected during the game and the methods by which we analyse it.

The Cropland Capture game

Because games are now the most common smartphone application type (dotMobi 2014), and serious games (games with a purpose) and gamification of existing applications are becoming more common (Michael and Chen 2005, Deterding *et al.* 2011), Geo-Wiki was moved into a gaming environment in an attempt to increase participant numbers and to collect more data. As with previous Geo-Wiki crowdsourcing projects (Perger *et al.* 2012, Fritz *et al.* 2013), the goal of this game is to provide data that can improve global land cover maps, in this case focusing on cropland. Although cropland cover is available globally from remotely-sensed global land cover products such as GLC-2000 (Fritz *et al.* 2003), MODIS (Friedl *et al.* 2010), GlobCover (Defourny *et al.* 2006) and the recent 30m Chinese land cover product (Yu *et al.* 2013), these products are not accurate enough for many applications. The ultimate goal of the game was to improve global cropland mapping by gathering data for training and validation, something that will be a part of future research.

In Cropland Capture, volunteer players (hereafter called ‘volunteers’) labeled imagery

(either from satellites or ground-based photographs) to gain points and to become eligible for a prize drawing at the end of the game. The game ran for a period of six months from mid-November 2013 until early May 2014. Incentives for participation were prizes awarded at the end of the game, which included smartphones and tablets. To become part of the final draw, a volunteer had to rank among a week's top three scorers; scores were reset to 0 on a weekly basis at midnight each Friday. Some individuals made it into the top three in many different weeks so they increased their chances of winning the prizes at the end. Special weekly prizes were added in the last five weeks of the game to motivate additional participation. Although prizes are no longer offered, the game can still be played online at www.geo-wiki.org/games/croplandcapture. The data analyzed in this paper comes only from the six month period noted above.

Figure 1 shows the game interface with an example image. The user is asked 'Is there cropland in the red box?' and chooses from three choices: 'cropland', 'not cropland' and 'maybe cropland'. If even a small fraction of the image contains cropland then it should be rated as 'cropland.' Cropland is defined as arable land and permanent crops based on the FAO definition (http://faostat.fao.org/site/375/default.aspx). Arable land consists of temporary agricultural crops, land under market and kitchen gardening and land temporarily fallow (less than five years) while permanent crops include cocoa and coffee but not forest plantations. Permanent meadows and pastures for grazing are not included in this definition, but land regularly mowed for hay is included. Volunteers were provided with a gallery of images to illustrate different types of cropland that they might see and to emphasize that the presence of any cropland at all, no matter how small, should result in a rating of 'cropland.'

Figure 1. An example of the Cropland Capture user interface. Users swipe the image toward 'yes', 'no' or 'maybe', or click on one of these options, depending on the type of device they are using.

The bulk of the images came from locations in the global validation data set of Zhao *et al.* (2014). These images were satellite-derived and roughly square, ranging from 100 m to 1 km on a side. Regardless of the scale of the presented scene, all images showed up as about the same size on the volunteer's device. Most of these images were high resolution, but some Landsat images were included where better resolution was not available. In addition to these satellite-based images, we used ground-based photos from the Degree Confluence Project (http://confluence.org). The dates associated with the imagery were recorded separately.

In this paper we take the term 'majority classification' to mean a decision made about the content of an image. For our purposes, this can be based either on expert validation, or on the majority of volunteer responses (excluding ratings of 'maybe'). For clarity, we use the term 'rating' to refer to a single decision by a single volunteer about a single image, in contrast to the collective, majority classification. We recognize that the majority classification is not necessarily correct and that in extreme cases some images may not even be classifiable, for example because of insufficient resolution or clouds obscuring the landscape. We address these issues of external validity using expert validations in the following section.

For each correct rating, the volunteer receives a single point. In this paper, we use the word 'rating' to mean the decision made by an individual volunteer about whether an image is cropland or not. For incorrect answers, the volunteer loses one point. If a volunteer answers 'maybe' they do not gain or lose any points. For the purpose of awarding points, correct answers were defined solely by the game's participants and gave some benefit of the doubt when there was not a strong majority classification. If ≥80% of an image's ratings (excluding responses of 'maybe') were cropland, only responses of 'yes' were considered correct. Similarly, if ≤20% of all non-maybe answers were 'cropland', then 'no' was the only correct answer. However, if the proportion of cropland ratings was between 20-80%, or if the image

had never previously been rated, either ‘yes’ or ‘no’ was credited as a correct response. While not having an explicit empirical basis, the values of 20% and 80% represent a tradeoff between identifying correct answers and penalizing wrong answers. Had these values been looser (i.e. closer to 0% and 100%), a few careless ratings of an easy image would result in all ratings being considered correct. On the other hand, if the values had been more stringent (i.e. closer to 50%), many potentially correct ratings of confusing and controversial images would be penalized.

We took several steps to reduce the possibility of playing the game in a way that provided little useful information. First, the proportion of images with and without cropland was approximately balanced (Table 2). This reduced the possibility for point accrual from random guessing. However, since both cropland and non-cropland were rewarded as correct answers for certain images (see previous paragraph), this problem could not be ruled out completely. To ensure that random play did not bias our results, we manually examined patterns of agreement with majority classifications for the most active participants (those rating at least 1000 images), and found no evidence of insincere participation. All participants took breaks and worked at uneven rates, so it is unlikely that any were bots. It is also possible that volunteers would over-use the ‘maybe’ rating, providing information only on the easiest images. This did not happen. Because image acquisition was automated, and it was impossible to check the quality of 165,000 images, some images entered rotation that were later decided by expert review to be unclassifiable. Examples include clouds, extremely low resolution and a few that were blank due to failed downloads. Even in these extreme cases, the proportion of ‘maybe’ ratings almost never exceeded 50%, suggesting that most volunteers erred toward guessing on hard images rather than toward caution, in spite of the risk of losing points.

We also took measures to reduce other sources of bias in our findings. It is possible that volunteers would perform better in evaluating images from familiar landscapes, particularly from regions where they live or have lived. While a ‘home field advantage’ cannot be ruled out, it

should have little or no impact on our findings as all participants were given randomly-selected images from all regions of the world; thus, all volunteers enjoyed this advantage occasionally, unless they came from a place with absolutely no agriculture. Similarly, variation in the difficulty of images is unlikely to affect outcomes of this research. While we show that such variation certainly exists, the random assignment of images again prevents systematic bias. It is possible that chance drawing of particularly easy or difficult sets of images could bias the metrics for a volunteer who contributed only a small number of ratings. However, our results were quite robust to inclusion of only contributors with >1000 images rated, suggesting that task difficulty does not bias our findings.

Analytical Methods

To compute volunteer quality metrics, it was first necessary to compute some image-specific metrics. We summed the number of times each image was rated in each of the three categories, and computed the proportion of responses in each category. Each image was then classified either as cropland or non-cropland based on which of these two possibilities received the most ratings from the volunteers. Note that these classifications were on the basis of a simple majority vote, and that this is different from the procedure described above for awarding points. Some images had an equal number of ratings in the cropland and no cropland categories so were classified as ties and not used in subsequent analyses. In determining the majority classification of an image, ratings of 'maybe' were omitted, no matter how frequently they were used.

For each volunteer, we computed several performance measures which form the basis for these analyses. A summary of these metrics is seen in Table 1. Total output was measured with the number of images rated (including repeats) or the number of images receiving a non-

maybe (i.e. cropland or non-cropland) rating. The quality of volunteer output was assessed in several ways. Each contributor's rate of agreement with majority-based classifications (see above) was calculated. For images rated more than once by a volunteer, a self-agreement rate was computed as the proportion of subsequent ratings agreeing with initial ratings. Ratings of 'maybe' were not counted in either the majority-agreement or self-agreement ratings. However, the ratio of 'maybe' ratings to all ratings was computed as a metric of caution. Ability to detect the two different cover types was assessed by separately calculating the proportion of cropland and non-cropland ratings that were correct.

Variable	Description
Maybe rate	The proportion of a volunteer's ratings that are 'maybe' rather than cropland or not cropland. All images were included in this variable, regardless of the majority classification.
Self-agreement rate	The proportion of a volunteer's ratings of a previously seen image that agree with his/her initial rating of that image.
Majority-agreement rate	The proportion of a volunteer's ratings that agree with the volunteers' majority-based classification of an image.
Expert-agreement rate	The proportion of a volunteer's ratings of images validated by experts that agree with the expert ratings.
Cropland identification rate	The proportion of images with a majority classification of 'cropland' that a volunteer rated as 'cropland.'
Non-cropland	The proportion of images with a majority classification of 'non-

identification rate	cropland' that a volunteer also rated as 'non-cropland.'
Images rated	The total number of images a volunteer has rated. Ratings of 'maybe' and ratings of images for which the majority classification was tied between cropland and non-cropland are included in this metric.

Table 1. The volunteer-specific metrics used in this article. Unless otherwise mentioned, these metrics exclude ratings of 'maybe' and all ratings of images for which the majority classification was a tied vote. Ratings of 'maybe' were excluded so that these metrics include only 'hard' errors (i.e. cropland classified as non-cropland or vice versa). Soft errors (classification changes from 'maybe' to 'cropland' or 'non-cropland') could be considered errors when they result from repeated rating of a single image by the same volunteer, but were not included to make self-agreement and majority-agreement rates more comparable.

We used regression analysis to test the relationship between different aspects of contributors' performance. For certain variable pairs, we have reasons to hypothesize a causal relationship, for instance, the hypothesis that accuracy increases with experience. Because the independent variable (total images rated) is measured without error, it is appropriate to use standard ordinary least squares (OLS) regression in this case. However, for other variable pairs, there is no theoretical reason to expect most metrics to be more accurately measured than others as we believe that all such metrics are reflections of an underlying (and unobserved) quality variable. Thus, it is inappropriate to use OLS regression because assumes that all error is contained in the dependent variable. Instead, we employ major axis regression (also known as type II regression) as implemented in the R package *lmodel2*. This method does not assume any underlying differences between the variables being analyzed, and unlike OLS regression,

returns the same result when the identity of the variables is switched (Legendre and Legendre, 1988). For some variables, log transformations were used to improve homoscedasticity. A side effect of this fix is that it eliminates values of zero. To circumvent this problem, we added one-half of the smallest non-zero value to variables containing values of zero before log transformation. Because neither of these methods is perfectly suited to our data, we report both as complementary outcomes.

In some cases we were interested to uncover not just patterns among typical volunteers, but also patterns among top volunteers as they contribute disproportionately to the eventual land cover classification goals of Cropland Capture. For this purpose we used quantile regressions (Cade and Noon, 2003), implemented in the R package *quantreg*. Quantile regression does not rely on the distributional assumptions of OLS regression so is particularly suited to the heteroskedastic patterns seen between many of the quality metrics. However, quantile regression still assumes error only in the dependent variable, so is best suited to relationships where one variable is precisely known. Another limitation of quantile regression is that it is not possible to compute an R^2 value in the classical sense, so instead we compute a pseudo-R^2, known as 'ρ', as an estimate of goodness of fit (Koenker and Machado, 1999).

To analyze patterns of learning we plotted learning curves based on the proportion of each 100 ratings that agreed with the majority classification (ratings of 'maybe' and images for which the classification was a tie vote between 'cropland' and 'non-cropland' were omitted). While a few participants showed visibly recognizable learning curves, most did not, either showing consistent performance or occasional valleys of very poor performance. Because these curves were so heterogeneous, they do not lend themselves to quantitative analysis and we do not report further on them.

As a more quantitative approach to the question of contributor learning, we examined the direction of change (relative to the majority classification) when a participant changed their

rating from ‘cropland’ to ‘non-cropland’ or vice versa between the first and second viewing of the same image. In total, this happened 38,344 times (ratings of ‘maybe’ were omitted). To assess possible learning over time, we evaluated whether switches toward agreement with the majority classification were more common than changes away from the majority classification. This was evaluated statistically with a binomial test with an expected probability of 50% (i.e. that changes were random).

In addition to volunteer ratings, 342 images were selected at the end of the game for expert validation to provide a baseline for evaluation of volunteers. These were not selected at random, but rather chosen to include different types of easy and difficult images. Easy images were roughly split between those where the majority classification overwhelmingly agreed on ‘cropland’ and ‘non-cropland.’ Difficult images came from several categories: images with many ratings in both the non-cropland and cropland categories, images with many ‘maybe’ ratings, images for which top volunteers (as assessed by majority-agreement rate) did not agree with the majority classification, and images where top raters disagreed with one another.

In the expert validation process, two remote sensing specialists (authors LS and SF) independently evaluated the 342 selected images in the same way as game volunteers, i.e. giving a response of ‘yes’, ‘no’ or ‘maybe’. Images for which the two specialists disagreed in their rating, or for which they both responded ‘maybe’, were reviewed in a group to determine whether one of them failed to notice some feature on the landscape. In some cases, input from regional experts was sought to interpret unexplained landscape features or additional local imagery was viewed either to provide landscape context or view questionable features in more detail. However, the ultimate determination for each image was made based on whether a skilled viewer could reasonably be expected to rate that image with no external information, and if so, what the correct rating would be.

Results

A total of 2,783 volunteers contributed ratings to Cropland Capture between November 2013 and May 2014. The game included 165,439 different images. Of these, 50.3% were satellite images, and 49.7% were ground-based photographs. Combining all volunteers and images, a grand total of 4,547,038 ratings were delivered. This total figure includes images that were seen more than once by a particular user to test repeatability of their ratings. For a typical contributor the percentage of repeat images was small, although the total value was inflated to 38.3% of ratings due to a small number of volunteers who contributed more ratings than the number of available images. The number of ratings performed by individual volunteers is log-normally distributed and ranges from 2 to 593,572. Volunteers came from all regions of the world, but with a bias toward wealthier countries.

Overall, volunteers disagreed with the majority classification of images 5.6 % of the time (Table 2). Majority-agreement rate for contributors with more than 1000 ratings ranged from 83.4% to 98.8%. When volunteers rated an image more than once, they agreed with their initial rating 96.5% of the time. Among those who rated more than 1000 images, self-agreement with previous ratings ranged from 88.8% to 100%. The average user gave a response of ‘maybe’ on 4.3% of ratings. For those rating >1000 images, this rate ranged from 0.0 % to 15.4 %.

Volunteer rating		Majority classification: **Cropland**	Majority classification: **Not cropland**	Accuracy
	Cropland	2,059,002 (46.8 %)	123,841 (2.8 %)	94.3 %
	Not cropland	121,677 (2.7 %)	2,094,747 (47.6 %)	94.5 %
	Accuracy	94.4 %	94.4 %	94.4 %

Table 2. Error matrix for agreement of individual volunteer ratings with the majority classification of all volunteers for images in the Cropland Capture game. This matrix does not include individual ratings of 'maybe' or images for which the vote was tied between 'cropland' and 'non-cropland.'

Across contributors, ratings showed very little bias toward either cropland or non-cropland (Figure 2, Table 2). The rates of correctly identifying cropland and non-cropland were nearly identical (Table 2). User's and producer's accuracies were virtually identical for both image types; all values fell between 94.3 and 94.5 % (Table 2). However, many individual volunteers showed some bias toward cropland or non-cropland. Those with very many images rated showed both lower bias and higher overall accuracy (Figure 2; note the concentration of large circles in the upper right corner of the figure). Among volunteers who rated fewer images, biases were more strongly exhibited (Figure 2).

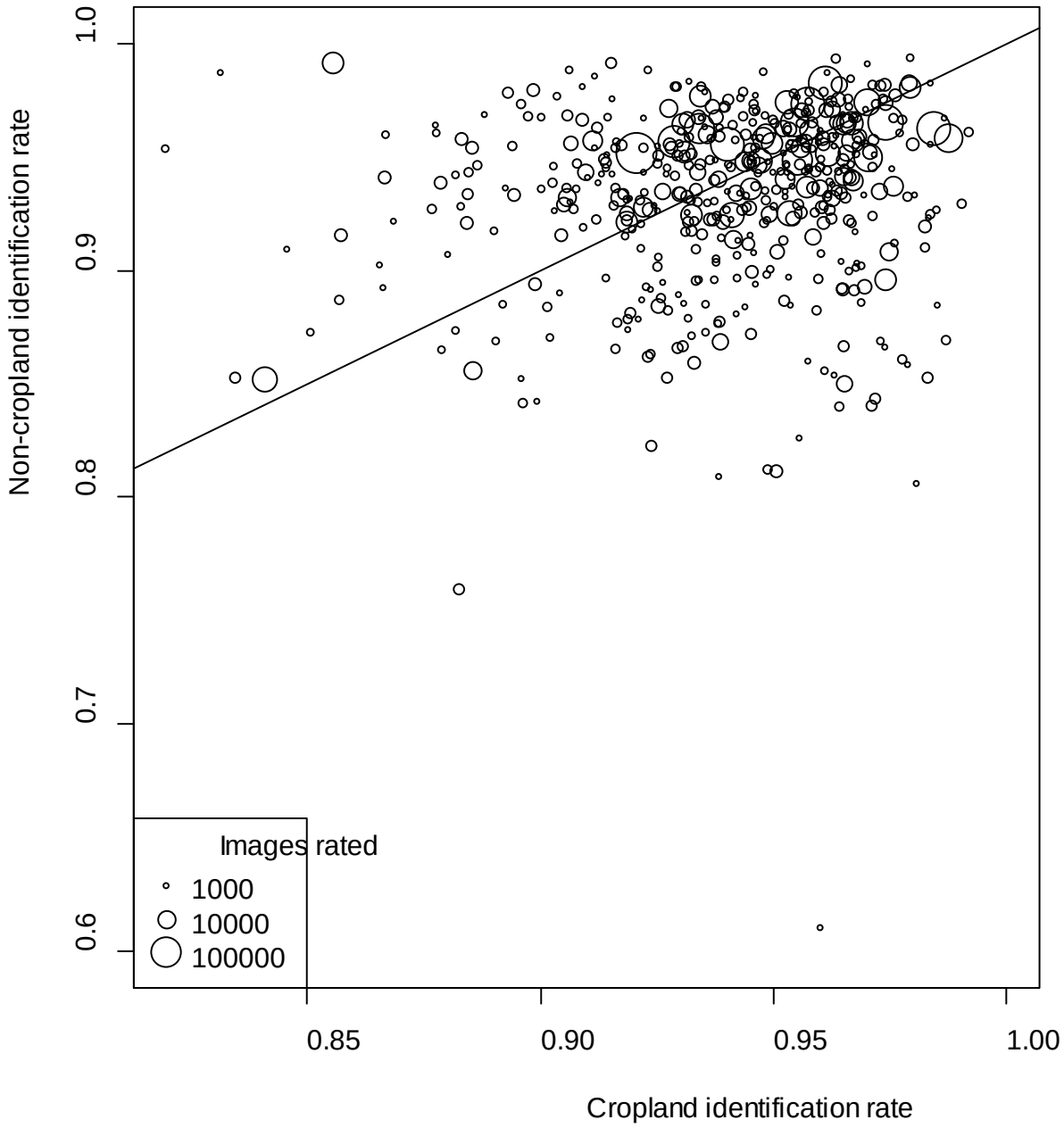

Figure 2. The relationship between volunteers' rates of correct identification of cropland and non-cropland. The axes represent the proportion of images identified as 'cropland' that the majority classified as 'cropland' (x) and the same metric for 'non-cropland' ratings (y). Each circle corresponds to an individual volunteer and its diameter indicates the total number of images rated. The diagonal line is a 1:1 line indicating equal rates of cropland and non-cropland identification.

Nearly all volunteers showed greater self-agreement than majority-agreement (Figure 3). Even so, there was a strong positive relationship between these variables, with volunteers who are more self-consistent showing greater majority-agreement (Figure 3; major-axis regression; $p<.0001$; $R^2=.634$). Volunteer self-agreement increased with use of the 'maybe' response (Figure 4). This trend was seen regardless of whether we used the raw variables ($p<.0001$, $R^2=.0864$) or log-transformations of the variables ($p<.0001$, $R^2=.0362$).

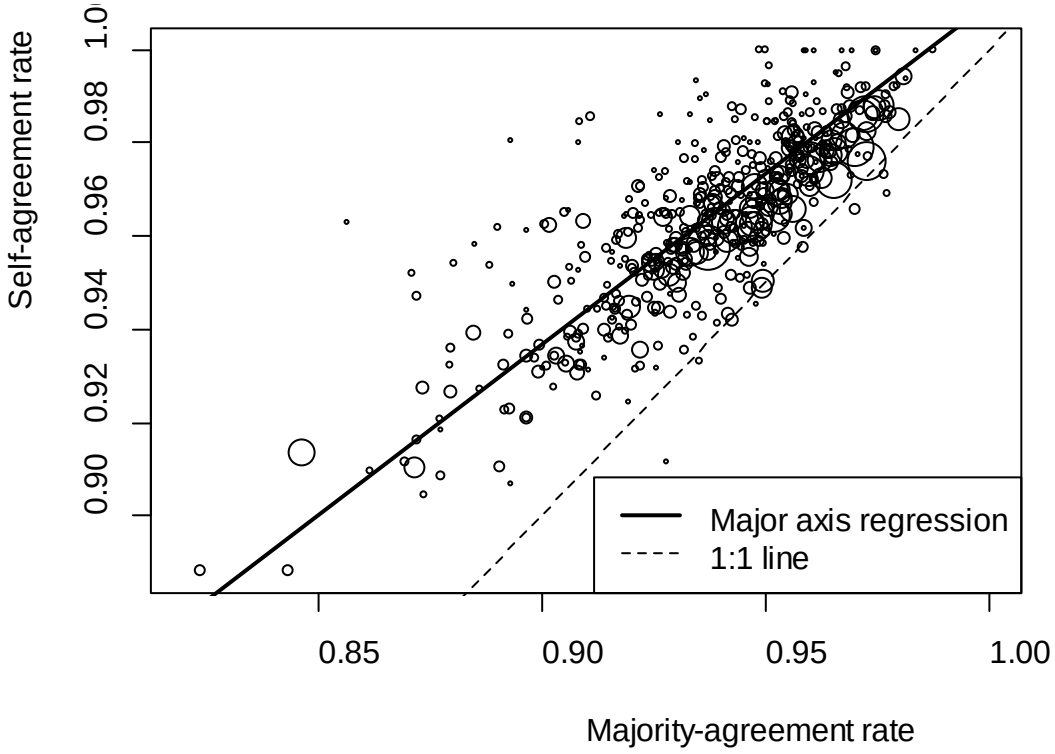

Figure 3. The relationship between volunteers' self-agreement rate as a function of their majority-agreement rate in the Cropland Capture game. Each point corresponds to a single volunteer. Only volunteers who have rated more than 1000 images are included in this figure. The solid line is a major axis regression which treats variables equally, rather than assuming all error is in the dependent variable. Circle size is proportional to number of images rated by a

volunteer. The 1:1 line shows where self-agreement rate equals majority-agreement rate.

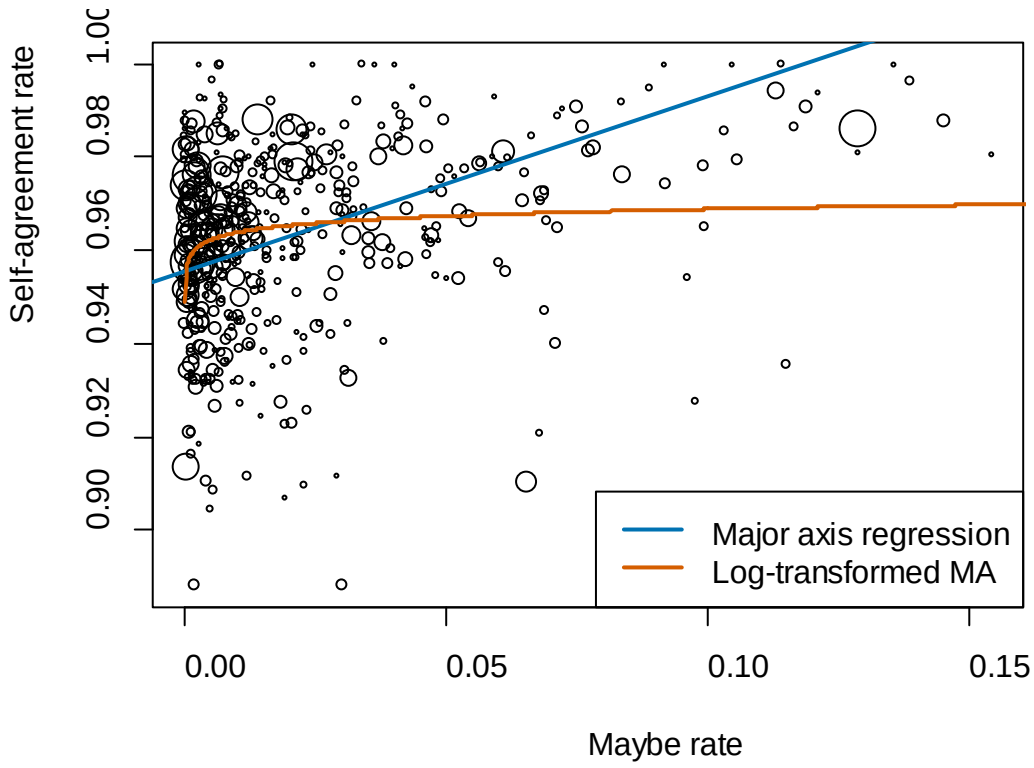

Figure 4. Volunteer agreement with their own previous ratings of the same image as a function of their willingness to admit uncertainty about presence of cropland in an image (maybe rate) in the Cropland Capture game. Each point corresponds to a single volunteer and circle size is proportional to the total number of images rated. Only those who rated more than 1000 images are included in this figure. The straight line is a major axis regression which treats variables equally, rather than assuming all error is in the dependent variable. The curved line is a major axis regression with the maybe rate log transformed. Because it was difficult to meet the assumptions of regression, both methods were applied, giving qualitatively similar results that were statistically significant in both cases (see main text).

The total number of images rated by a volunteer shows a complex relationship with

rating quality. Volunteers' median rate of majority-agreement increases significantly with total images rated in the game (quantile regression with τ=.5; p<.0001, ρ=.65; Figure 5; note that 'τ' is the quantile level of the regression and that 'ρ' is a measure of goodness of fit, analogous to R^2 in ordinary least squares regression). However, among top performing volunteers, this relationship is reversed (quantile regression with τ=.9; p=.0005, ρ=.93; Figure 5). When only contributors with substantial experience (>1000 images rated) were considered, some of these relationships changed considerably. Median volunteers still showed improved majority-agreement rate as images rated increased (p<.0001, ρ=.05), but the slope of 90th percentile volunteers became significantly positive (p=.0002; ρ=.09).

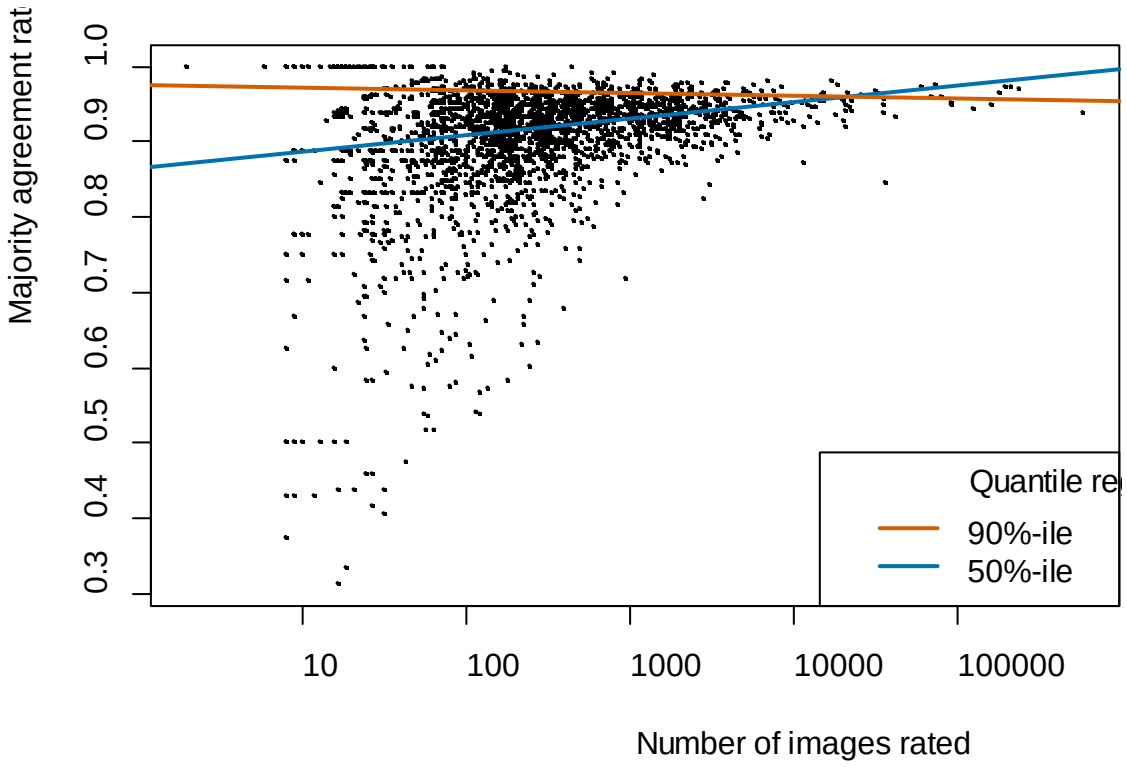

Figure 5. Volunteer agreement with the majority-based classification of images as a function of total images rated in the Cropland Capture game. Each point corresponds to a single volunteer. The upward-sloping line is a median (50th percentile) quantile regression, and the downward-

sloping line a 90th percentile quantile regression. Both slopes differ significantly from zero (see main text).

Volunteer self-agreement rate showed similar patterns to majority-agreement rate with increasing experience (Figure 6). Because volunteers received only occasional repeat images (to reduce the likelihood of them remembering having seen it before), this calculation was only possible for contributors who performed large numbers of ratings. As above, we used a minimum of 1000 ratings as the cutoff. In contrast to majority-agreement rate, the median user's self-agreement rate decreased slightly but non-significantly with experience ($p=.728$). The top volunteers (90th percentile of self-agreement rate) showed a more strongly negative trend toward less self-agreement with increasing experience ($p=.018$, $\rho=.019$). Only the lowest ranking contributors (10th percentile of self-agreement rate) showed an increase in self-agreement rate with experience ($p=.0015$, $\rho=.014$). More direct evidence for volunteer learning was provided by the directional analysis of self-contradictions. Overall, 54.7% of changed ratings between the first and second viewing of an image were in the direction of agreement with the majority classification and 45.3% were in the opposite direction. This trend is statistically strong (binomial test, $p<.0001$).

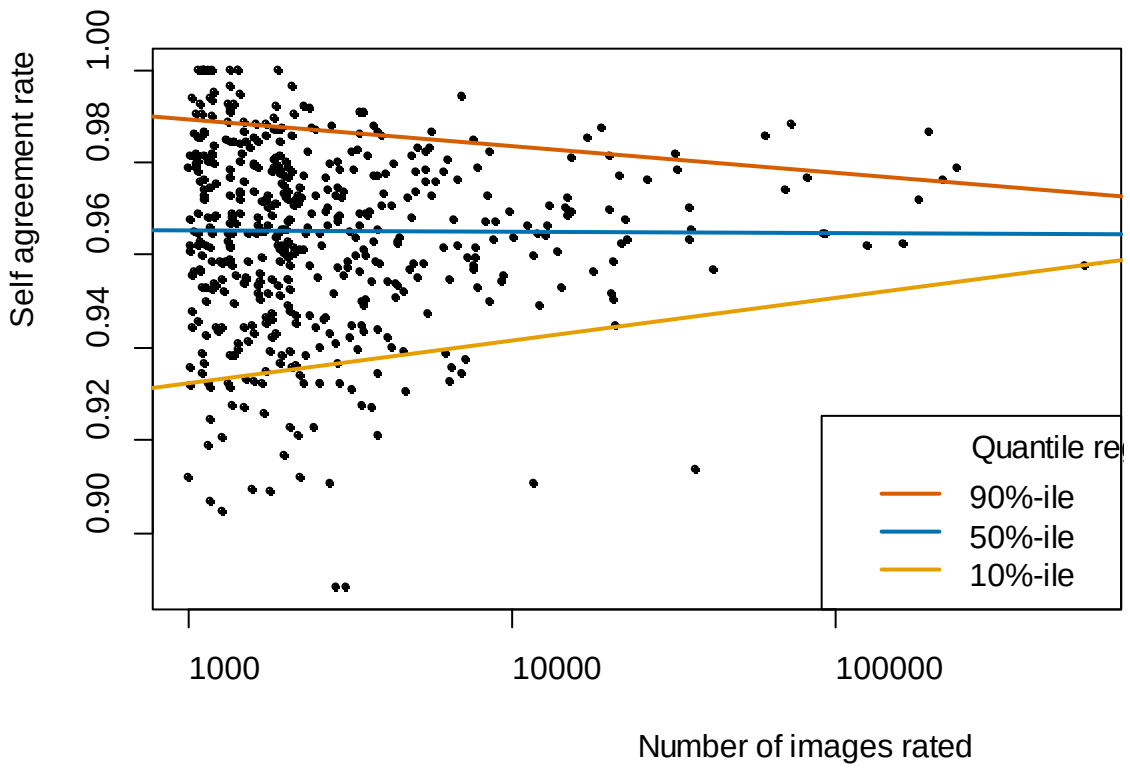

Figure 6. Volunteer rate of self-agreement on repeatedly rated images as a function of total number of images rated in the Cropland Capture game. Each point corresponds to a single volunteer. Only volunteers who have rated more than 1000 images are included in this figure. The lines represent regressions through different quantiles of the distribution. The slopes of the 90th and 10th percentile lines, but not the 50th percentile line, differ significantly from zero (see main text).

Among volunteers who rated more than 15 images that were validated by experts, expert-agreement rate ranged from 5% to 93%. This value was consistently less than the majority-agreement rate. Each volunteer disagreed with the experts more than with the majority classification, typically by a very wide margin (Figure 7). However, there was still a positive association between volunteers' majority-- and expert-agreement rates, although with little

predictive power (logit-transformed major axis regression, p=.001, R^2=.113). The rate of disagreeing with one's previous ratings of the same image also underestimated expert-validated accuracy rates; all volunteers disagreed with experts more than with themselves, and in nearly all cases this discrepancy was large.

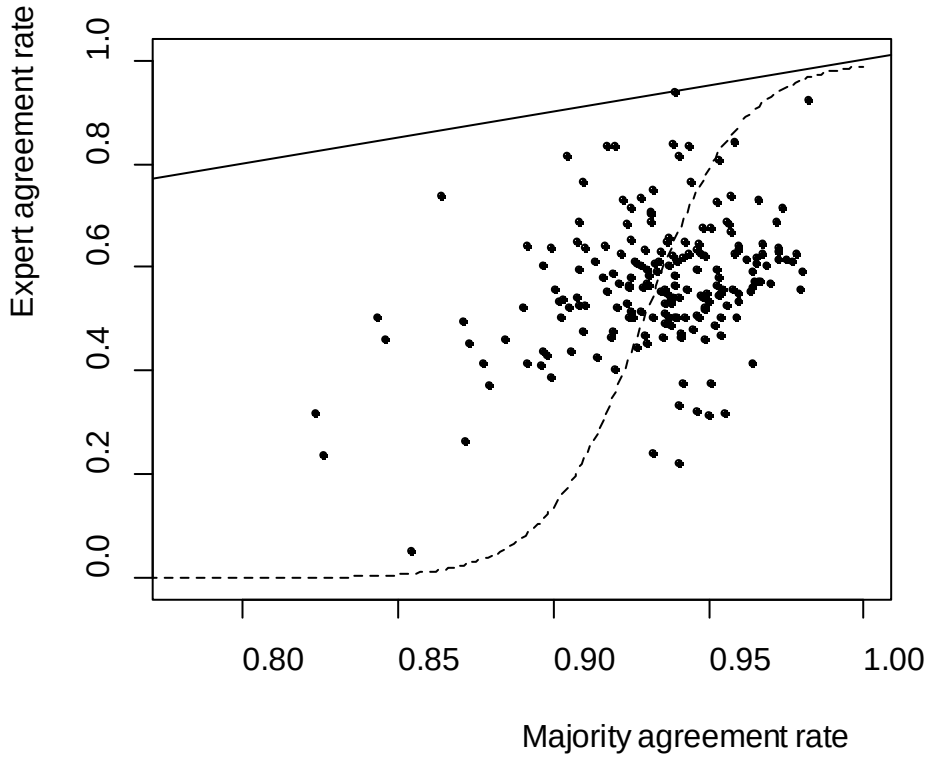

Figure 7. The relationship between the majority-agreement rate and expert-agreement rate of volunteers in the Cropland Capture game. Each point represents a single volunteer who rated at >15 images that were also classified by experts. The solid line is a 1:1 line, showing that all volunteers agreed with the majority classification more often than with experts. The dashed line shows the correlation between the two variables using major axis regression. The line is not straight because expert agreement rate was logit transformed before regression to avoid predictions beyond the possible range of [0,1].

Discussion

This paper has compared agreement with experts, agreement the majority of other volunteers and consistency in repeated ratings of an image as metrics to compare the quality of performance on a simple land-cover identification task. While all of these metrics are somewhat correlated with one another, they give insight into different facets of volunteers' performance, a subject elaborated on below. We have shown that many volunteers exhibit a clear and quantifiable bias in their ratings, and are more likely to correctly identify 'cropland' than 'non-cropland' or vice-versa. We have also demonstrated limited effects of improved quality of responses with increased game play. Finally, this work has shown that majority-agreement and self-agreement are not fully able to substitute for experts to validate the quality of crowdsourced tasks. In the process, we have uncovered certain design habits that promote easier downstream extraction of information from games for the solicitation of VGI. All of this information has important implications for the design of serious games and the choice of metrics used to evaluate and compensate their volunteers.

The independence of volunteers' true positive and true negative rates has consequences for devising schemes for volunteer scoring and compensation. We don't provide specific guidance here as this is a decision that depends on the goals of a game and the relative benefit of correct answers and cost of incorrect answers. However, it is worth keeping in mind that because individuals can show very different patterns of these errors, not taking into account their strengths and weaknesses may inadvertently lead to less efficient accrual of the information that game managers seek. This would waste the time of volunteers, expert validators and game managers.

That the direction of the relationship between the number of images rated to majority-agreement rate changes with quantile suggests that more than one process governs this

relationship. Our interpretation is that volunteers do learn to better identify cropland with experience, as shown by the analysis of switched responses when re-rating images. However in order to rate a really big number of images, it is not possible to spend much time on individual images. Therefore, the most active raters can be expected to have a somewhat higher error rate than the best intermediate-experience volunteers. However, this conclusion should be drawn with caution as these patterns seem to weaken when only contributors who have rated >1000 images are included.

At first glance, the comparison of the different metrics of volunteer quality suggests that volunteers are highly effective at rating photographs and satellite imagery for the presence of cropland. Consistency was high relative both individual volunteers' previous ratings of the same images, and with other volunteers' ratings of those images. Further, self-agreement rate and majority-agreement rate are positively correlated. The subtle differences between these two metrics seem to indicate that volunteers are generating meaningful data. Self-agreement rate can be interpreted in two ways. It could be seen as a direct indicator of undesirable sloppiness, or at least indecisiveness. It could also be an indicator of learning; in reality, it is likely a combination of both. As an indicator of learning, we found direct evidence – disagreement with one's previous ratings of an image shows a significant tendency toward agreement with the majority classification. In spite of the statistical strength of this relationship, the split between switching to and from agreement with the majority classification was less than 10 percentage points (54.7% vs. 45.3%). That there is much more noise than signal in this shift, suggest that self-contradiction can be useful as a measurement of sloppiness or indecisiveness, in spite of the toward-crowd bias. Self-agreement is an assessment of volunteer quality that can be made independently of any knowledge of the correctness of responses, but the rate of agreement with the majority classification is not a judgment made in a vacuum; it depends on the responses of many other volunteers. Taken together, the correlation among these two metrics shows that

volunteers who are careful (i.e. rarely contradict their previous answers to repeated images) are also more likely to agree with the majority classification. That volunteers who are unequivocally better in one respect (they are self-consistent) are also in greater agreement with the majority classification suggests that the crowd is providing useful information.

Unfortunately, results from expert validation temper this apparent good news. In spite of self-consistent volunteers' also agreeing better with majority classifications of images, few volunteers exhibited anything close to the level of agreement with experts that they shared with the other volunteers in the game (Figure 7). This shows that there is a collective behavior of the crowd that, in spite of being shared among many volunteers, is at odds with the explicit goal of the game. The incoherence of results derived in these two ways suggests that extracting a reliable signal from crowdsourced data without guidance from expert validations is not possible for this type of task. This is not a difficult problem to surmount, but it does mean that a certain degree of effort is required to plan campaigns so that the crowd is channeled into the desired way of rating images.

The need for expert validations calls to attention another problem that arises when many raters evaluate a huge number of tasks. The problem is that in some situations, it is impossible to choose validation images at the end of a competition such that most volunteers have looked at enough of them for volunteer-specific comparison with expert ratings. In our study, only 194 of the 2,783 participants performed at least 15 ratings of images that were chosen for validation by experts (shown in Figure 7). While these images were selected to include a range of difficulties rather than to maximize the number of expert-validated images rated by volunteers, selecting for the latter goal would not have greatly increased the overlap between expert- and volunteer-viewed images. When tasks from a very large pool are more or less randomly assigned to a much smaller number of raters, it is inevitable that the number of validated tasks

necessary to robustly compare volunteers is prohibitively large. The solution is having a pre-defined set of tasks that are assigned to most or all volunteers. While it may seem pointless to have a group of individual images that are each rated by thousands of different people, this pattern of play is necessary to robustly evaluate contributor quality, and also simplifies implementing the other metrics discussed in this paper.

References

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-nezhad, H. R., Bertino, E. and Dustdar, S., 2013. "Quality control in crowdsourcing systems." *IEEE Internet Computing*, 17(2), 76-81.

Bachrach, Y., Minka, T., Guiver, J. and Graepel, T., 2012. "How to grade a test without knowing the answers – a Bayesian graphical model for adaptive crowdsourcing and aptitude testing." *ArXiv*, 1206.6386, http:// http://arxiv.org/abs/1206.6386.

Bird, T.J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., et al., 2014. "Statistical solutions for error and bias in global citizen science datasets." *Biological Conservation*, 173, 144-154.

Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., and Shirk, J., 2009. "Citizen science: A developing tool for expanding science knowledge and scientific literacy." *BioScience*, 59 (11), 977–984.

Cade, B. S. and Noon, B. R., 2003. "A gentle introduction to quantile regression for ecologists". *Frontiers in Ecology and the Environment*, 1(8): 412-420.

Canavosio-Zuzelski, R., Agouris, P., and Doucette, P., 2013. "A photogrammetric approach for assessing positional accuracy of OpenStreetMap© roads." *ISPRS International Journal of Geo-Information*, 2 (2), 276–301.

Clery, D., 2011. "Galaxy Zoo volunteers share pain and glory of research." *Science*, 333 (6039), 173–175.

Coleman, D., 2013. "Potential contributions and challenges of VGI for conventional topographic base-mapping programs." In *Crowdsourcing Geographic Knowledge*, edited by D. Sui, S. Elwood, and M. Goodchild, 245–263. Springer Netherlands.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., and Foody, G., 2013. "Using control data to determine the reliability of volunteered geographic information about land cover." *International Journal of Applied Earth Observation and Geoinformation*, 23, 37–48.

Dawid, A. P. and Skene, A. M., 1979. "Maximum likelihood estimation of observer error-rates using the EM algorithm." *Applied Statistics*, 28, 20-28.

Defourny, P., Vancustem, C., Bicheron, P., Brockmann, C., Nino, F., Schouten, L., and Leroy, M., 2006. "GLOBCOVER: A 300m global land cover product for 2005 using ENVISAT MERIS time series." *In*: *Proceedings of the ISPRS Commission VII Mid-Term Symposium: Remote Sensing: from Pixels to Processes*. Enscede NL.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., and Dixon, D., 2011. "Gamification. using game-design elements in non-gaming contexts." ACM Press, 2425.

Digital Globe, 2014. "DigitalGlobe Crowdsourcing." https://www.digitalglobe.com/sites/default/files/Crowdsourcing-DS-CROWD.pdf

dotMobi, 2014. "Global mobile statistics 2013 Section E: Mobile apps, app stores, pricing and failure rates." Accessed 7 Mar 2014. Available from: http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/e#popularappcatagories.

Flanagin, A. and Metzger, M., 2008. "The credibility of volunteered geographic information." *GeoJournal*, 72, 137–148.

Fonte, C.C., Bastin, L., See, L., Foody, G.M., Lupia, F., and Vatseva, R., Submitted. "Usage of VGI for validation of land cover maps." *International Journal of Geographical Information Science*.

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*, 17 (6), 847–860.

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., and Huang, X., 2010. "MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets." *Remote Sensing of Environment*, 114 (1), 168–182.

Fritz, S., Bartholomé, E., Belward, A., Hartley, A., Stibig, H.-J., Eva, H., and Mayaux, P., 2003. *Harmonisation, mosaicing and production of the Global Land Cover 2000 database (Beta Version)*. Luxembourg: Office for Official Publications of the European Communities, No. EUR 20849EN.

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

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 & Software*, 31, 110–123.

Fritz, S., See, L., van der Velde, M., Nalepa, R.A., Perger, C., Schill, C., McCallum, I. et al., 2013. "Downgrading recent estimates of land available for biofuel production." *Environmental Science & Technology*, 47 (3), 1688–1694.

Girres, J.-F. and Touya, G., 2010. "Quality assessment of the French OpenStreetMap dataset." *Transactions in GIS*, 14 (4), 435–459.

Goodchild, M.F., 2007. "Citizens as sensors: the world of volunteered geography." *GeoJournal*, 69 (4), 211–221.

Haklay, M., 2010. "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets." *Environment and Planning B: Planning and Design*, 37, 682–703.

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 (4), 315–322.

Hecht, R., Kunze, C., and Hahmann, S., 2013. "Measuring completeness of building footprints in OpenStreetMap over space and time." *ISPRS International Journal of Geo-Information*, 2 (4), 1066–1091.

Howe, J., 2006. "The rise of crowdsouring." *Wired Magazine*, 14 (6).

Jokar Arsanjani, J., Helbich, M., Bakillah, M., and Loos, L., 2013. "The emergence and evolution of OpenStreetMap: a cellular automata approach." *International Journal of Digital Earth*, 1–15.

Khatib, F., DiMaio, F., Group, F.C., Group, F.V.C., Cooper, S., Kazmierczyk, M., Gilski, M., et al., 2011. "Crystal structure of a monomeric retroviral protease solved by protein folding game players." *Nature Structural & Molecular Biology*, 18 (10), 1175–1177.

Koenker, R. and Machado, J. A. F., 1999. "Goodness of fit and related inference processes for quantile regression". *Journal of the American Statistical Association*, 94, 1296-1310.

Legendre P., Legendre L., 1988. *Numerical ecology*. Number 20 in Developments in Environmental Modelling, 2nd edition. Amsterdam: Elsevier.

Michael, D.R. and Chen, S.L., 2005. *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade.

Miller-Rushing, A., Primack, R., and Bonney, R., 2012. "The history of public participation in ecological research." *Frontiers in Ecology and the Environment*, 10 (6), 285–290.

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

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*: T. Jekel, A. Car, J. Strobl, and G. Griesebner, eds. *GI_Forum 2012: Geovisualisation, Society and Learning*. Berlin / Offenbach: Herbert Wichmann Verlag, 83–91.

Ramm, F., Topf, J., and Chilton, S., 2011. *OpenstreetMap: Using and Enhancing the Free Map of the World*. Cambridge, England: UIT Cambridge.

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F., and Obersteiner, M., 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLoS ONE*, 8 (7), e69958.

See, L., Fritz, S., and de Leeuw, J., 2013. "The rise of collaborative mapping: Trends and future directions." *ISPRS International Journal of Geo-Information*, 2 (4), 955–958.

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., et al., 2014. "The eBird enterprise: An integrated approach to development and application of citizen science." *Biological Conservation*, 169, 31–40.

Yu, L., Wang, J., Clinton, N., Xin, Q., Zhong, L., Chen, Y., and Gong, P., 2013. "FROM-GC: 30 m global cropland extent derived through multisource data integration." *International Journal of Digital Earth*, 6 (6), 521–533.

Zhao, Y., Gong, P., Yu, L., Hu, L., Li, X., Li, C., Zhang, H. et al., 2014. Towards a common validation sample set for global land-cover mapping. *International Journal of Remote Sensing*, 35 (13), 4795–4814.