Sparks, K., Klippel, A., Wallgrün, J. O., & Mark, D. M. (2015). Citizen science land cover classification based on ground and aerial imagery. In S. I. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. M. Freundschuh, & S. Bell (Eds.), Proceedings, Conference on Spatial Information Theory (COSIT 2015), Santa Fe, NM, USA, Oct. 12-16, 2015 . Berlin: Springer.

Citizen Science Land Cover Classification Based on Ground and Aerial Imagery

Kevin Sparks[1], Alexander Klippel[1], Jan Oliver Wallgrün[1], David Mark[2]

[1]The Pennsylvania State University, University Park, State College, PA 16801
{kas5822, klippel, wallgrun} @psu.edu
[2]NCGIA & Department of Geography, University at Buffalo, Buffalo, NY 14228
dmark@buffalo.edu

Abstract. If citizen science is to be used in the context of environmental research, there needs to be a rigorous evaluation of humans' cognitive ability to interpret and classify environmental features. This research, with a focus on land cover, explores the extent to which citizen science can be used to sense and measure the environment and contribute to the creation and validation of environmental data. We examine methodological differences and humans' ability to classify land cover given different information sources: a ground-based photo of a landscape versus a ground and aerial based photo of the same location. Participants are solicited from the online crowdsourcing platform Amazon Mechanical Turk. Results suggest that across methods and in both ground-based, and ground and aerial based experiments, there are similar patterns of agreement and disagreement among participants across land cover classes. Understanding these patterns is critical to form a solid basis for using humans as sensors in earth observation.

Keywords: land cover, citizen science, classification

1 Introduction

Land cover data is often a critical parameter in geophysical research. Climate modeling, food security, and biodiversity monitoring are a few areas that have recently increased in importance, all of which land cover are central to. With the recent availability of different types of earth observation data (e.g., high resolution aerial photos, ground-based photos), and the growth of citizen science and crowdsourcing, new opportunities have opened up for environmental monitoring and data creation from non-authoritative sources (i.e. novice citizens). One of the main advantages of using citizen science in any area of research is its efficiency. It is critical however that we ensure this new source of monitoring and data creation is as reliable and consistent as possible. This paper follows up results found in Sparks et al. [1], and will examine how citizen science can best be used for the purposes of environmental monitoring and land cover classification.

There are multiple global land cover datasets, many of which are created for a specific purpose or bias in mind. These specialized purposes often lead to disagreements between datasets and their classification schemes. This variation is illustrated by the Geo-Wiki project [2] showing the differences between GLC-2000, MODIS land cover products, and GlobCover. Variation in the classification schemes can occur via differing land cover classes (e.g., the inclusion or exclusion of a Grassland class), changed meanings of shared terminology (e.g., defining a Grassland class in different ways), and differences in the interpretation and perception of the land cover classes among users. Non-standardized data is a challenge to be faced in all fields of research. In our specific example of land cover, Ahlqvist [3] and Comber et al. [4] discuss the need for a more standard interpretation in light of the subjectivity in the dataset creation process, and in the interpretation of that data by the user. Comber et al. [4] specifically discusses the point how the perception of geographic terms in a land cover classification scheme will differ depending on the purpose of why that dataset was created. Likewise, the users' perception is often influenced by cultural and individual differences. If citizen science is to be incorporated in the environmental dataset/classification creation process, than there is a need for standardizing class definitions [5]. This need is proven through Robbins [6] discussion on land cover and land use classification choices by foresters and herders in a local Indian community. He shows how their classification choices for a surrounding area in a local community are influenced by these people's cultural and political roles in that community. This example further illustrates the point of class interpretation variation between producers and users of the data.

In order to solve this challenge of interpretability of land cover classification schemes, we must gain a deeper understanding of how humans perceive land cover classes [7]. This means exploring users' natural concepts of the environment, and being concerned with cognitive models about the geographic world. Coeterier [8] discusses that even when asking citizens to compare landscapes that are vastly different from each other, there is agreement among the importance of higher-level attributes of those landscapes. These high-level attributes are not necessarily features or objects within the landscape (i.e., trees, grass, water), but instead the landscape's use, naturalness, and spaciousness. These types of high-level attributes can nonetheless be quantified and used in a classification scheme. Continuing with these higher-level attributes of landscapes, Habron [9] analyzes the perceptual variation of what is considered Wild Land in Scotland. He concludes that while there is variation among sections of the population, there is a general agreement on the core definition of what Wild Land is. Also, by identifying that human impact has a large effect on what is considered Wild Land, he implicitly notes that citizens can consistently identify and distinguish non-natural environmental features from natural environmental features. These types of cognitive model processes discussed by Coeterier [8] and Habron [9] are the issues we must be aware of when attempting to increase interpretability on land cover classes.

In addition to interpretation variation and accuracy related issues, unknown variation gets introduced when a given dataset changes its methodology for the creation process. Comber et al. [10] uses the example of the Great Britain Datasets LCM1990 and LCM2000 to illustrate how a new methodology in the dataset creation process can create uncertainty between either observing land cover change, or simply observing a

change in how the land cover is represented semantically. Furthermore, Foody [11] reiterates that land cover is dynamic. The earth's surface will change in the time it takes to update datasets especially for developed regions. This alone encourages a method of collecting data that is quick and reliable in order to keep up with the changing earth surface.

While there is a lot of interest surrounding the opportunities of the crowd, there is a high demand for systematic evaluations of how much improvement in environmental monitoring can be achieved using crowd-based assessments. In response to these challenges, this paper follows up on experiments described in Sparks et al. [1], and analyzes the consistency and reliability of humans' classification of land cover given ground-based and aerial-based photos of landscapes. If citizen science is to be incorporated into the evaluation of land cover data, there needs to be a more rigorous understanding of how humans perceive and conceptualize land cover types and a more detailed assessment of how well humans perform in recognizing predefined land cover classes. We are reporting on two experiments that provide insights on the relationships between human conceptualizations of land cover and land cover classifications using novices. Our findings suggest inter-participant agreements are not random but rather systematic to unique land cover stimuli and unique land cover classes, but are not greatly influenced by additional information such as aerial photos.

2 Background

Recent advancements in engineering and technology have created an opportunity for citizen science to have a significant impact on scientific research. We have seen examples of this impact across many research fields through the discovery of protein structures [12], the identification of galaxies [13], and the validation of land cover classes [2]. This last reference [2], referring to the Geo-Wiki project, is the most recent example of a crowdsourcing effort to assist in environmental monitoring. The Geo-Wiki project identifies locations where global land cover datasets disagree on a given land cover classification. It then solicits crowdsourced participants, provides them with aerial imagery, and asks them to make a classification choice for that location of disagreement. This data shows a lot of promise in validating land cover datasets, but like most sources of citizen science and crowdsourced data, it fails to assure reliability and consistency.

In order to ensure reliability and consistency, most attempts to gather data come from more authoritative sources. The Land Use/Cover Area Frame Survey (LUCAS) [14] is an example of a more authoritative source that attempts to capture land use/cover data. LUCAS, commissioned by Eurostat, uses trained surveyors to collect and create land cover data rather than relying on novice citizens. These land surveyors personally visit many locations, recording land transects, taking photos, and determining land use/cover at a given location. In this example coming from a more authoritative source, efficiency is sacrificed for reliability and consistency.

Citizen science needs to be able to guarantee a relatively high amount of reliability and consistency, along with being efficient. In the context of using citizen science for environmental monitoring, the Citizen Observatory Web (COBWEB) [15] uses citizens

living in biosphere reserves across Europe to collect environmental data using mobile devices. Like LUCAS, COBWEB uses humans to collect data in the field, with the difference coming from trained (LUCAS) versus novice (COBWEB) sensors. The project's aim is to gain a deeper understanding of environmentally crowdsourced data by working with citizens throughout the process of data creation. By quality controlling this information, COBWEB hopes to impact environmental policy formation and more general societal and commercial benefits through the use of citizen science.

Data from Geo-Wiki project [2] has been analyzed to measure the quality of humans' classification of land cover given aerial photos [16, 17, 18, 19, 20]. See et al. [17] and Comber et al. [19] focus on the differences between expert Geo-Wiki participants and non-expert Geo-Wiki participants when classifying land cover given aerial imagery. See et al. [17] reports averaged agreement rates between participants of 66%-76% agreement when classifying land cover, noting experts generally having a higher maximum agreement than non-experts. Comber et al. [19] concludes with a similar result of experts being different than non-experts, but still calls for "…further investigation into formal structures to allow such differences to be modeled and reasoned with" ([19], pg 257). And while expertise has a general influence, that influence is varied across land cover classes, with expertise playing a larger role in certain types of classes [20].

While aerial photos have been available for some time, access to quality datasets of ground-based photos have recently emerged. The Geo-Wiki campaigns offer insight on humans' land cover classification using aerial photos. Others have attempted to test the effectiveness of using ground-based photos for humans' land cover classification [21] [22]. While no research has tested these ground-based photos on a large number of crowdsourced participants, research has concluded that ground-based photos are a valid data source when attempting to classify land cover [21] [22].

In combination with the success of the Geo-Wiki project in contributing to the growth of land cover datasets, OpenStreetMap [23] has also shown success in the contribution of environmental information from citizen science. OpenStreetMap is an open source dataset that is built from citizens volunteering and creating geographic information. Arsanjani et al. [24] analyzed OpenStreetMap land use/cover contributions to measure the accuracy of participants. He concludes that OpenStreetMap, and in general other forms of crowdsourced geographic data, can be reliable and consistent sources for mapping land use.

To summarize, citizen science and crowdsourced geographic data, while being efficient, are largely critiqued for being unreliable and inconsistent. In the context of land cover data, there has been preliminary research that suggests citizen science is promising for monitoring, validating, and creating environmental data. However, as projects like COBWEB show, there is a need to further understand how humans perceive and classify environmental features in order to determine reliable practices. Furthermore, various environmental information channels (ground-based versus aerial-based photos) and methods of classification need to be tested to determine best practices for citizen science involvement in environmental monitoring.

3 Experiments

To further advance our understanding of the potentials and limits of the human sensory and conceptual system in contributing to earth observations, we systematically extend our previous experiments [1] in two ways: first, we replicate an experiment on land cover classes but use a different methodology; second, participants received multiple perspectives on the same environment, that is, ground-based photos were complemented by aerial photos. The rational for these changes are explained in more detail below. Experiments in previous work [1] tested varying levels of participant expertise when classifying ground-based photos into land cover classes. The overall task in the experiments reported here remains the same: Participants are asked to classify photos into land cover classes.

3.1 Experiment 1 – Ground-based photos

The first question we address is a methodological one: Do we change the results of previous studies when the experimental setup is changed. Specifically, instead of using CatScan [25] the experimental setup was switched to Qualtrics [26]. CatScan is a card sorting tool that presents participants with stimuli/icons on the left half of the screen, and empty groups on the right half of the screen. Participants are asked to click and drag these stimuli/icons from the left half of the screen into groups on the right half of the screen based on their similarity. Qualtrics is an online survey platform with a lot of customizability. The driving force behind this change is the greater flexibility for non-free classification of land cover photos that Qualtrics offers (see Figure 1): (a) Images can be presented individually allowing for higher resolution, (b) additional information for individual photos can be obtained such as how certain or uncertain a selected land cover class is, finally, (c) in preparation for experiment 2, the display real estate can be used to provide additional information for solving the classification task.

The first experiment asks participants to choose a land cover class, based on the National Land Cover Dataset (NLCD) [27] classification scheme, for ground-based photos of land cover. The experiment is a replication of the experiment of Sparks et al. [1] with the methodological change mentioned above.

Materials. Two datasets were used in experiment 1: First, ground-based photos of landscapes provided by the Degree Confluence Project (DCP) (confluence.org). Second, the National Land Cover Dataset (NLCD) 2006 provided by the United States Geological Survey (USGS) Land Cover Institute [27].

The DCP is a website which provides a platform for collecting crowdsourced photos of the environment at confluence points across the world. The word confluence as defined for the purposes of the DCP is the location where two integer latitude and longitude coordinate lines meet. A total of 799 photos were collected across the continental United States, which we sampled from for experiment 1 and experiment 2.

The NLCD 2006 is thematic land cover data for the United States prepared from Landsat 7 Enhanced Thematic Mapper Plus and Landsat 5 Thematic Mapper imagery collected between 2001 and 2006. We use NLCD as an authoritative dataset to measure

participants' classification against. NLCD data however is not being used as ground truth, or being used to determine accuracy of participant classification. The data is only being used to see how much participants agree with an authoritative dataset. NLCD also provides the scheme from which participants can choose land cover classes in the experiments.

Latitude and longitude coordinates from the DCP data were used to spatially join their corresponding land cover class from the level II NLCD 2006. The level II NLCD classification scheme has a total of 16 land cover classes, but after spatially joining the 799 DCP photos, only 11 were returned. We aggregated Deciduous Forest, Evergreen Forest, and Mixed Forest into one Forest class (removing 2 options from the 16), and Developed Medium Intensity, Developed High Intensity, and Perennial Ice/Snow did not return enough photos (removing 3 options from the 16). The DCP images, now each assigned to 1 of 11 land cover classes, were sorted into bins based on their land cover class. 7 photos were randomly selected within each class, totaling 77 images. These 11 land cover classes make up the categorical choices for each question in the experiment.

Participants. 20 lay participants (non-experts, 11 female) were recruited through the crowdsourcing platform Amazon Mechanical Turk (AMT); average age 32.4 years; reimbursement: $1.80. Eight participants have postsecondary degrees. Participants were asked to provide the type of landscape they live in, given the options of Rural, Sub-urban, and Urban. Participants were not provided with definitions of Rural, Sub-urban, and Urban, and we did not verify their response. Of the three options for the currently lived in landscape, 3 participants live in Rural, 9 in Sub-urban, and 8 in Urban.

We believe that it is important to note that crowd science does not necessarily mean large samples. We have looked into calculating effect size but this does not seem to be straightforward for classification tasks. While it is possible to pick up smaller effects with larger samples, the goal is not to show that there is a difference even if it takes 10,000 participants in each experiment. Given that the patterns that we observed are present not only across methods but also across different participant groups (experts versus novices), we believe that it is not pertinent for this paper to increase the sample size.

While this paper is not a review on the validity of using AMT to solicit participants for academic studies, we will provide the following comments on our use of AMT for this research. As seen in Sparks et al. [1], AMT participants and expert participants solicited personally from a university campus performed the classification task with very similar results (statistically speaking, significantly not different). This would suggest the lack of any influence from AMT "super workers" that perform similar tasks multiple times on AMT. To reinforce this, our experimental surveys, being environmental classification tasks, are relatively unique from other AMT HITS.

Procedure. Qualtrics is an online survey service that we used to build our surveys and record data. Participants are solicited through AMT and directed to the Qualtrics survey via a link. Once participants begin the survey, they are asked basic demographic questions (age, gender, level of education). After providing this personal information, participants are given the definitions of each of the 11 land cover classes. The definitions

of each class are taken directly from the NLCD classification scheme. Participants must confirm that they have read and understood each definition before they can progress to the main experiment. They are also given prototypical photos of what each land cover type looks like. Each participant has access to these definitions and prototypical photos at any point throughout the experiment. The participant is then shown each of the 77 ground-based photos, one at a time, and asked to make a classification decision given the 11 land cover class options. The participant is also asked to give their level of confidence about their choice of land cover: Sure (most confident), Quite sure, Less sure, and Unsure (least confident). The directions were explicit in informing the participants that Sure meant most confident, and Unsure meant least confident. Once a selection had been made, the participants could not go back and revisit previous questions. The participants had to finish the experiment in one sitting (i.e. they could not stop, exit the survey, and revisit the survey at a later time to finish). An example of what a question in the survey looks like can be seen in Figure 1.

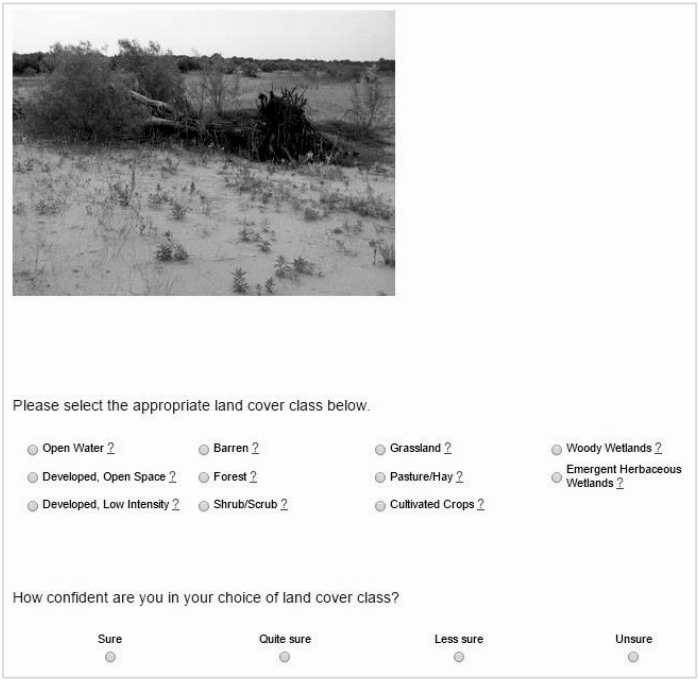

Fig. 1. Qualtrics interface showing one land cover photo participants are asked to classify (the image in the actual experiment is in color). Each photo is shown with the 11 potential land cover classes and a scale for the confidence of the selected classification. The blue question marks next to the land cover class options are links to the definition and prototypical photo of that land cover class.

Results. Results are reported as the level of agreement between participants' classification choice and the classification from the NLCD for a given photo. Once again, we are not claiming this NLCD classification as accuracy or ground truth, rather we are

using it to measure percent agreement between an authoritative dataset and participants. Overall agreement with NLCD is 45.97%. The average length of an experiment was 21 minutes 13 seconds. Time data for each individual question was not recorded.

After performing a Chi Square analysis, the following land cover classes significantly agreed with NLCD more frequently than expected by having a standardized residual value greater than 1.96 (Table 1): Developed, Low Intensity (dL), Forest (FO), Open Water (OW), and Shrub Scrub (SS). The following land cover classes significantly disagreed with NLCD more frequently than expected by having a standardized residual value less than -1.96: Barren (BA), Emergent Herbaceous Wetlands (EW), Pasture/Hay (PH), and Woody Wetlands (WW).

Table 1. Standardized residuals for experiment 1.

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
Correct	-6.11	-0.06	4.20	-0.95	-7.17	9.18	-0.59	13.09	-6.46	2.06	-7.17

Confusion matrices were created to allow us to see participant agreement across the 11 land cover classes (see Figure 2). The confusion matrices' classes (columns and rows) come from the NLCD classification scheme, and illustrate the grouping behavior between participants across classes. For example, looking across the first row (BA), 21.43% of the photos that NLCD classified as Barren (BA), participants also agreed were Barren (BA). However, 55% of the photos NLCD classified as Barren (BA), participants thought were Shrub/Scrub (SS). Most importantly, these matrices allow us to visualize participants' classification patterns. As previously mentioned, this is the most important metric.

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	21.4	0.71	0	0	14.3	5	1.43	0	2.14	55	0
CC	6.43	45.7	0	0	4.29	0	22.9	0	16.4	4.29	0
dL	0	0	62.9	30	0	0	5	0	2.14	0	0
dO	0	0	45.7	42.1	0	0	2.86	0	9.29	0	0
EW	0.71	2.14	0	0	17.1	30.7	5	0	15	22.1	7.14
FO	0	0	0	0	2.86	82.9	0.71	0	0	9.29	4.29
GS	7.14	13.6	0	0	2.86	0	43.6	0	15.7	16.4	0.71
OW	0	0	0	0	0.71	0	0	98.6	0	0	0.71
PH	2.86	4.29	0	2.14	2.14	14.3	34.3	0	20	17.9	2.14
SS	22.9	0	0	0	2.86	1.43	15	0	3.57	54.3	0
WW	0	0	0	0	7.86	72.1	0	0	0	2.86	17.1

Fig. 2. Confusion matrix for experiment 1 (ground-based photos) showing percentages of participant agreement. Agreement between 5% and 25% is indicated by light grey, agreement between 25% and 50% is grey, and agreement above 50% is dark grey.

Along with the relatively low overall agreement with NLCD (45.97%), we can see from the confusion matrix in Figure 2 that overall, the level of participant agreement varies across land cover classes. Classes like Open Water (OW), Forest (FO), and the Developed classes (dL, dO) show a relatively large amount of agreement among participants. Otherwise, participants show a relatively large amount of disagreement with each other among the rest of the land cover classes. Specifically, participants are classifying a variety of photos into classes like Emergent Herbaceous Wetlands (EW) and Grasslands (GS), suggesting these classes are more heterogeneous compared to classes like Open Water (OW).

We also created a confusion matrix for photos for which participant indicated a high level of certainty (Sure) in their responses (Figure 3). The Total Confident column at the end of the matrix shows how many Sure confidence responses that particular land cover class received across the 20 participants out of a total of 140.

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW	Total Confident
BA	33.33	0	0	0	16.66	0	0	0	0	50	0	30
CC	0	68.08	0	0	0	0	17.02	0	10.63	4.25	0	47
dL	0	0	67.18	28.12	0	0	3.12	0	1.56	0	0	64
dO	0	0	46.42	41.07	0	0	1.78	0	10.71	0	0	56
EW	0	2.56	0	0	10.25	53.84	0	0	12.82	17.94	2.56	39
FO	0	0	0	0	0	100	0	0	0	0	0	39
GS	3.33	26.66	0	0	0	0	50	0	3.33	16.66	0	30
OW	0	0	0	0	0	0	0	100	0	0	0	114
PH	8.69	4.34	0	0	0	0	43.47	0	17.39	26.08	0	23
SS	26.66	0	0	0	0	0	6.66	0	0	66.66	0	30
WW	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 3. Confusion matrix for only *Sure* confidence responses.

Overall participant agreement with NLCD for Sure confidence responses is 71.19%. Participants who indicated to be Sure of classifying Forest (FO) and Open Water (OW) were in 100% agreement with each other, and NLCD. No one was confident in classifying any Woody Wetlands (WW) photos. When comparing this Sure confusion matrix (Figure 3) with the previous confusion matrix (Figure 2), the low percentages of agreement (light pink squares) disappear for the Sure confusion matrix, and the high percentages of agreement intensify. Higher percentage of agreement can be seen specifically for the Pasture/Hay (PH) row and Emergent Herbaceous Wetlands (EW) row. Simply looking at the last column, Total Confident, provides insight into the most interpretable land cover classes. As perhaps expected, participants were more frequently confident when classifying a photo as Open Water (OW). Both Developed classes showed frequent confidence too. Conversely, no participants were confident when classifying a

photo as Woody Wetlands (WW), and were not frequently confident when classifying a photo as Barren (BA), Grassland (GS), Pasture/Hay (PH), or Shrub/Scrub (SS).

One of the motivations to switch from CatScan to Qualtrics was the possibility to obtain additional information such as the confidence a participant has making a classification. As mentioned previously, we repeated the analysis above considering only Sure responses. The agreement between NLCD and participants increases significantly ($x2 = 24.1902$, $df = 1$, p-value < 0.001). Overall agreement with NLCD (71.91%) starts to approach the overall accuracy of Level II NLCD (78%) [28]. A similar pattern can be seen between the two confusion matrices, only with a less amount of low disagreement, and more high agreement in the Sure confusion matrix. Participants are Sure most often when classifying Open Water (OW) and, to a lesser extent, the Developed land cover classes. This is perhaps expected as water and developed features are more easily distinguishable from natural features.

Discussion. The change in the experimental setup has allowed us to confirm results obtained previously but also add to our understanding of how humans might aid in earth observations. Overall percent agreement of Sure responses with NLCD (71.19%) is comparable to the overall range of accuracy reported in Geo-Wiki campaigns of 64-84% [16] and 66-76% accuracy [17]. The overall results, not considering the certainty of the classification, are very similar to the results seen in previous experiments [1] that use the same ground-based photos, but a different experimental interface (CatScan). The overall agreement with NLCD differs only 1.62% from Sparks et al. [1] to the experiment above. More so, the pattern in the confusion matrices from Sparks et al. [1] and the experiment above are very similar. Barren (BA) and Shrub/Scrub (SS) frequently are confused with one another, participants are generally in high agreement on what is Developed but vary between what is Open Space and what is Low Intensity, and participants are in high agreement on what is a Forest (FO) and Open Water (OW) photo. This suggests that participants' classification choices are not being influenced differently from the CatScan interface to the Qualtrics interface.

3.2 Experiment 2 – Ground and aerial based photos

To further advance our understanding of the human potential as an earth observer, experiment 2 extends experiment 1 by allowing participants access to not only ground-based photos but also corresponding aerial photos showing the area in question. The aerial photos contribute additional context information that might aid in achieving consistent classifications.

Materials. The National Agricultural Imagery Program (NAIP) imagery provided by the United States Department of Agriculture (USDA) is used in combination with the materials described in experiment 1. Using the latitude and longitude coordinates from each of the 77 ground-based photos, NAIP imagery of those same locations were downloaded. The NAIP imagery has a spatial resolution of 1 meter and is taken during the agricultural growing season across the continental United States. The NAIP images are shown to the participants at a 1:2000 scale, and cover an extent of 300 meters (~984

feet) by 300 meters. We added a white square centered on the image that is 30 meters (~98 feet) by 30 meters. This white square represents where the corresponding ground-based photo should be located. Participants are asked to only make their classification choice based on both information sources, the ground-based photo and what is inside the white square (see Figure 4). Everything outside of the white square is there to provide context of the surrounding area. Participants are encouraged to consider this surrounding area when making their classification choice.

Participants. 20 lay participants (non-experts, 6 female) were recruited through AMT; average age 32.9 years; reimbursement: $1.80. Seven participants have postsecondary degrees. Of the three options for currently lived in landscape, 3 participants live in Rural, 9 Sub-urban, and 8 Urban. These 20 participants for experiment 2 were a separate group from the previous 20 participants in experiment 1.

Procedure. The NAIP photos were added to each question in the survey. When considering the NAIP photo, participants are asked to make their decision based off of the region inside the white square in the center of the photo. Otherwise the procedure is the same as experiment 1. An example of what a question in the survey looks like can be seen below (Figure 4).

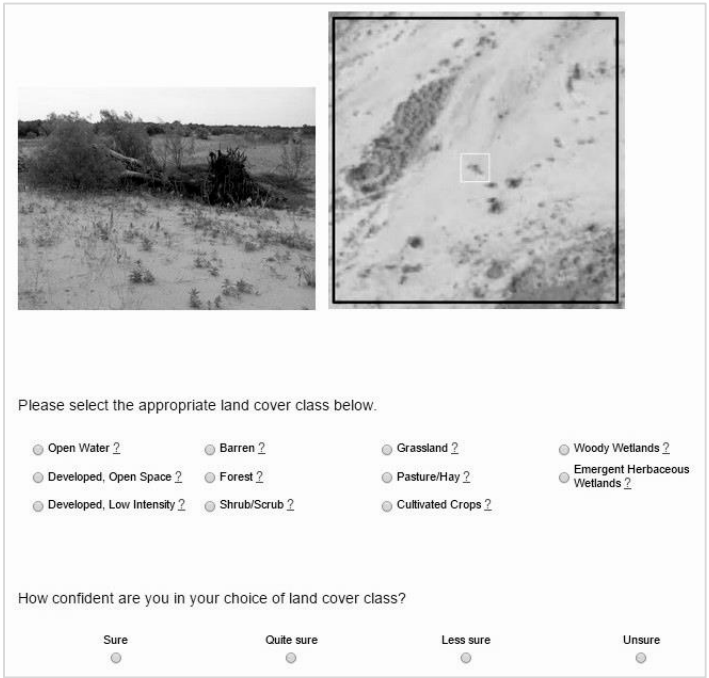

Fig. 4. Qualtrics interface showing one land cover photo participants are asked to classify (the images in the actual experiment are in color).

Results. Overall agreement with NLCD is 42.79%. The average length of an experiment was 21 minutes 16 seconds. The drop in agreement with NLCD from experiment 1 (45.97%) to experiment 2 (42.79%) is not statistically significant ($x2 = 3.0305$, df = 1, p-value = 0.081). Agreement for Barren (BA) increased significantly from experiment 1 to experiment 2 ($x2 = 13.7708$, df = 1, p-value < 0.001), and agreement for Shrub/Scrub (SS) and Woody Wetlands (WW) significantly decreased from experiment 1 to experiment 2 ($x2 = 4.1362$, df = 1, p-value = 0.04, and $x2 = 4.702$, df = 1, p-value = 0.03).

After performing a Chi Square analysis, the following land cover classes significantly agreed with NLCD more frequently than expected by having a standardized residual value greater than 1.96 (Table 2): developed, Low Intensity (dL), Forest (FO), and Open Water (OW). The following land cover classes significantly disagreed with NLCD more frequently than expected by having a standardized residual value less than -1.96: Emergent Herbaceous Wetlands (EW), Pasture/Hay (PH), and Woody Wetlands (WW).

Table 2. Standardized residuals for experiment 2.

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
Correct	0.01	-0.70	2.70	0.37	-7.50	8.97	-0.87	13.45	-7.32	-0.34	-8.76

Similar to experiment 1, we can see from the confusion matrix (Figure 5) that participant agreement varies across land cover classes. Once again, classes like Open Water (OW), Forest (FO), and Developed (dL, dO) are the exception, showing a relatively large amount of agreement among participants. Otherwise, participants show a relatively large amount of disagreement among each other across rest land cover classes, specifically in classes like Emergent Herbaceous Wetlands (EW) and Grasslands (GS).

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	42.86	0	0	0	11.43	2.86	1.43	0	0.71	40	0.71
CC	5.71	40	0	2.86	1.43	0	32.86	0.71	11.43	4.29	0.71
dL	0.71	0	53.57	35.71	0	0	7.14	0	2.86	0	0
dO	0	4.29	38.57	44.29	0	0.71	4.29	0	7.14	0.71	0
EW	2.86	2.14	0	0	12.86	37.14	8.57	0	12.86	20.71	2.86
FO	0.71	0	0	0	0.71	78.57	1.43	0	2.14	14.29	2.14
GS	12.14	11.43	0.71	0	3.57	0	39.29	0	17.14	15.71	0
OW	0	0	0	0	0.71	0	0	96.43	0	0	2.86
PH	7.14	5.71	0	5.71	3.57	2.14	42.86	0	13.57	17.86	1.43
SS	32.86	0	0	0	1.43	0.71	15.71	0	7.86	41.43	0
WW	0	0	2.14	0.71	0.71	85	0.71	0	0	2.86	7.86

Fig. 5. Confusion matrix for experiment 2 (ground and aerial-based photos).

Like in experiment 1, we performed the analysis again using only those images that participants indicated they were Sure about (Figure 6). Overall participant agreement with NLCD for Sure confidence responses is 59.55%. No land cover class had full agreement among participants who were Sure. Similar to experiment 1, no one was confident in classifying any Woody Wetlands (WW) photos. Participants classified more images as Barren (BA) and Grassland (GS) land cover classes in experiment 2 (for both overall and Sure responses).

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW	Total Confident
BA	64.51	0	0	0	6.45	0	0	0	0	29.03	0	31
CC	0	59.52	0	2.38	0	0	33.33	0	4.76	0	0	42
dL	0	0	58.2	40.29	0	0	1.49	0	0	0	0	67
dO	0	0	55.55	40.74	0	0	1.85	0	1.85	0	0	54
EW	7.31	0	0	0	2.43	68.29	7.31	0	4.87	9.75	0	41
FO	0	0	0	0	0	95.65	0	0	0	4.34	0	46
GS	17.24	27.58	0	0	0	0	44.82	0	3.44	6.89	0	29
OW	0	0	0	0	0	0	0	99.14	0	0	0.85	117
PH	6.45	6.45	0	3.22	0	0	70.96	0	0	12.9	0	31
SS	44.11	0	0	0	2.94	0	14.7	0	0	38.23	0	34
WW	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 6. Confusion matrix for only *Sure* confidence responses.

Discussion. The introduction of corresponding aerial photos significantly increased Barren (BA) agreement with NLCD from experiment 1, but otherwise agreement with NLCD slightly decreased with the introduction of aerial photos, with Shrub Scrub (SS) and Woody Wetlands (WW) dropping significantly.

These drops in agreement, along with a relatively unchanged number of Sure confidence responses from experiment 1 to experiment 2 across each class, suggests that aerial photos do not provide any more clarity when perceiving and classifying land cover types.

The introduction of aerial photos could possibly be contradicting what its corresponding ground-based photo is showing. For example, if a given ground-based photo portraying something a participant might classify as Shrub/Scrub (SS) is contradicted by its corresponding aerial photo that portrays something a participant might classify as Barren (BA), perhaps the aerial photo takes precedence for the participant when determining a land cover. The participant might also be considering the surrounding area too much outside of the location of interest in the aerial photo, i.e. the white square. The introduction of a larger region could be leading to more heterogeneous surfaces, making the classification of the location more ambiguous.

Barren environments perhaps benefit from a larger context of the surrounding area. In experiment 1, participants may have been influenced more by the presence of one or two shrubs in the ground-based photo in a mostly barren environment. Whereas in experiment 2, when provided with more of the surrounding area, the participant can see how much of the land cover is truly barren, with sparse shrubs potentially having less of an impact with a larger area. Barren (BA) is also generally a more homogenized land cover class, as is Grassland (GS). This homogenization is perhaps intensified and considered more when an aerial perspective is included. This possibly explains why participants chose Barren (BA) and Grassland (GS) more frequently.

The significant decrease in Woody Wetlands (WW) agreement could possibly be explained by the influence of, in most cases, a homogenized looking canopy from an aerial perspective in the Woody Wetlands (WW) photos. In experiment 1, participants did not have this homogenized canopy influence, and are perhaps more focused on the potential source of water seen from ground-based photos.

Similarly to the reasoning for the increase in participant agreement for Barren (BA), Shrub/Scrub (SS) may have experienced a significant decrease in agreement when considering the larger surrounding area. Participants' classification choice may have been more heavily influenced from the aerial photos rather than the ground-based photo.

4 General Discussion/Outlook

Looking at the results from experiment 1 and experiment 2, as well as the three experiments outlined in Sparks et al. [1], a consistent parameter in all these experiments has been the categorical 11-class classification method. The land cover class semantics are potentially the largest influence on participant agreement. These categorical land cover classification tasks are difficult. A possible explanation for why, is that this categorical classification scheme is generalizing land covers too much, and these classes are at too high of a level that subjectivity overrides objectivity. As discussed in Section 2, the earth's surface is often complex and heterogeneous. Forcing this complexity into relatively high-level categorical classes is prone to errors and disagreements.

We have now run experiments using different interfaces (CatScan vs Qualtrics), different users (Amazon Mechanical Turk vs live Experts), and different stimuli (ground vs ground and aerial). Yet with all these changes, a similar pattern seen in the confusion matrices persists (Figure 7).

Future research will begin to explore non-categorical classification methods, such as free classification of landscape images [29] as well as mimicking a decision tree process of classification. In the case of the latter, participants will make an initial decision regarding the presence of environmental features in the image (e.g., is this image primarily vegetated or primarily non-vegetated) that will then lead to another unique set of choices, and so on. Humans will more likely agree on the presence of environmental features versus higher-level categories shown in the experiments above.

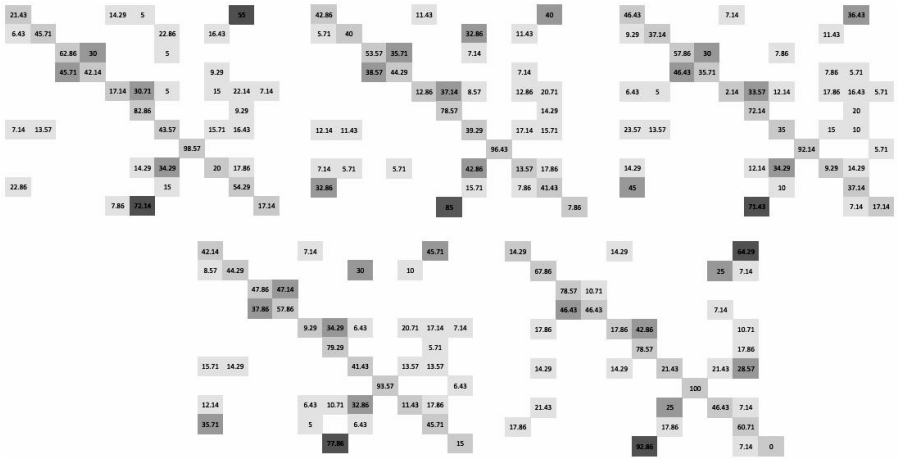

Fig. 7. All 5 confusion matrices.

Confusion matrices were also created for each subcategory of landscape each participant said they currently lived in (Rural, Sub-urban, Urban) for both the ground-based, and ground and aerial-based experiments. Data collected is preliminary as each sub-class of participant included less than 10 participants. Generally, rural dwelling participants showed relatively low agreement on the Developed classes and relatively higher agreement on Woody Wetlands (WW) and Cultivated Crops (CC). Urban dwelling participants showed generally higher agreement on Developed classes. While this is preliminary, we intend to explore the influence of lived in landscape for land cover classification in future research.

In the context of projects like Geo-Wiki and COBWEB, we need to be aware of the classification task design and try to make it as objective as possible. As discussed in section 2, the earth's surface is complex and often heterogeneous. In order to use citizen science for environmental monitoring and be as reliable and consistent as possible, we need to continue to explore how humans perceive land cover classes and environmental features, and make sure that the classification task is designed in a way to encourage objectivity and discourage subjectivity.

Acknowledgments. This research is funded by the National Science Foundation (#0924534). We would like to thank the Degree Confluence Project for permission to use photos from the confluence.org website for our research. We would like to thank the members of the Human Factors in GIScience laboratory.

References

[1] K. Sparks, A. Klippel, J. O. Wallgrun and D. Mark, "Crowdsourcing landscape perceptions to validate land cover classifications," in *Land Use and*

Land Cover Semantics - Principals, Best Practices and Prospects, CRC Press / Taylor & Francis, in press.

[2] S. Fritz, I. McCallum, C. Schill, C. Perger, R. Grillmayer, F. Achard, F. Kraxner and M. Obersteiner, "Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover," *Remote Sensing,* vol. 1, no. 3, pp. 345-354, 2009.

[3] O. Ahlqvist, "In search of classification that supports the dynamics of science: the FAO Land Cover Classification System and proposed modifications," *Environment and Planning B: Planning and Design,* vol. 35, no. 1, pp. 169-186, 2008.

[4] A. Comber, P. Fisher and R. Wadsworth, "What is land cover?," *Environment and Planning B,* vol. 32, pp. 199-209, 2005.

[5] M. R. Jepsen and G. Levin, "Semantically based reclassification of Danish land-use and land-cover information," *International Journal of Geographical Information Science,* vol. 27, no. 12, pp. 2375-2390, 2013.

[6] P. Robbins, "Beyond Ground Truth: GIS and the Environmental Knowledge of Herders, Professional Foresters, and Other Traditional Communities," *Human Ecology,* vol. 31, no. 2, pp. 233-253, 2003.

[7] O. Ahlqvist, "Semantic issues in Land Use and Land Cover Studies – Foundations, Application and Future Directions," in *Remote Sensing of Land Use and Land Cover: Principles and Applications*, Boca Raton, FL, Taylor and Francis, 2012, pp. 25-36.

[8] J. F. Coeterier, "Dominant attributes in the perception and evaluation of the Dutch landscape," *Landscape and Urban Planning,* vol. 34, no. 1, pp. 27-44, 1996.

[9] D. Habron, "Visual perception of wild land in Scotland," *Landscape and Urban Planning,* vol. 42, no. 1, pp. 45-56, 1998.

[10] A. Comber, P. Fisher and R. Wadsworth, "Integrating land-cover data with different ontologies: identifying change from inconsistency," *International Journal of Geographical Information Science,* vol. 18, no. 7, pp. 691-708, 2004.

[11] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sensing of Environment,* vol. 80, pp. 185-201, 2002.

[12] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, M. Jaskolski and D. Baker, "Crystal structure of a monomeric retroviral protease solved by protein folding game players," *Nature Structural & Molecular Biology,* vol. 18, no. 10, pp. 1175-1177, 2011.

[13] D. Clery, "Galaxy Zoo Volunteers Share Pain and Glory of Research," *Science,* vol. 333, pp. 173-175.

[14] "LUCAS: Land Use/Cover Area frame Statistical Survey," [Online]. Available: http://www.lucas-europa.info/.

[15] "COBWEB: Citizen Observatory Web," [Online]. Available: https://cobwebproject.eu.

[16] C. Perger, S. Fritz, L. See, C. Schill, van der Velde, Marijn, I. McCallum and M. Obersteiner, "A Campaign to Collect Volunteered Geographic Information on Land Cover and Human Impact," *GI_Forum 2012: Geovizualisation, Society and Learning*, 2012.

[17] L. See, A. Comber, C. Salk, S. Fritz, van der Velde, Marijn, C. Perger, C. Schill, I. McCallum, F. Kraxner, M. Obersteiner and T. Preis, "Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts," *PLoS ONE*, vol. 8, no. 7, p. e69958, 2013.

[18] G. M. Foody, L. See, S. Fritz, Van der Velde, M., C. Perger, C. Schill and D. S. Boyd, "Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project," *Transactions in GIS*, vol. 17, no. 6, pp. 847-860, 2013.

[19] A. Comber, C. Brunsdon, L. See, S. Fritz and I. McCallum, "Comparing Expert and Non-expert Conceptualisations of the Land: An Analysis of Crowdsourced Land Cover Data," *Spatial Information Theory*, pp. 243-260, 2013.

[20] A. Comber, L. See and S. Fritz, "The Impact of contributor Confidence, Expertise and Distance on the Crowdsourced Land Cover Data Quality," *GI_Forum 2014 - Geospatial Innovation for Society*, 2014.

[21] K. Iwao, K. Nishida, T. Kinoshita and Y. Yamagata, "Validating land cover maps with Degree Confluence Project information," *Geophysical Research Letters*, vol. 33, 2006.

[22] G. M. Foody and D. S. Boyd, "Using Volunteered Data in Land Cover Map Validation: Mapping West African Forests," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 3, pp. 1305-1312, 2013.

[23] "OpenStreetMap," [Online]. Available: https://www.openstreetmap.org.

[24] J. J. Arsanjani, M. Helbich, M. Bakillah, J. Hagenauer and A. Zipf, "Toward mapping land-use patterns from volunteered geographic information," *International Journal of Geographical Information Science*, vol. 27, no. 12, pp. 2264-2278, 2013.

[25] A. Klippel, "Spatial information theory meets spatial thinking - Is topology the Rosetta Stone of spatio-temporal cognition?," *Annals of the Association of American Geographers*, vol. 102, no. 6, pp. 1310-1328, 2012.

[26] "Qualtrics: Online Survey Software & Insight Platform," [Online]. Available: http://www.qualtrics.com/.

[27] J. Fry, G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold and J. Wickham, "Completion of the 2006 national land cover database for the conterminous United States," *Photogrammetric Engineering & Remote Sensing*, vol. 77, no. 9, 2011.

[28] J. D. Wickham, S. V. Stehman, L. Gass, J. Dewitz, J. A. Fry and T. G. Wade, "Accuracy assessment of NLCD 2006 land cover and impervious surface," *Remote Sensing of Environment,* vol. 130, pp. 294-304, 2013.

[29] A. Klippel, D. Mark, J.O. Wallgrün, and D. Stea, "Conceptualizing landscapes: A comparative study of landscape categories with Navajo and English-speaking participants," In S. I. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. M. Freundschuh, & S. Bell (Eds.), *Proceedings, Conference on Spatial Information Theory (COSIT 2015),* Santa Fe, NM, USA, Oct. 12-16, 2015 . Berlin: Springer.