K.A. Sparks, A. Klippel, J.O. Wallgrun, D. Mark. (in press) "Crowdsourcing landscape perceptions to validate land cover classifications", In O. Ahlqvist, K. Janowicz, D. Varanka, S. Fritz (Eds.), *Land Use and Land Cover Semantics - Principles, Best Practices and Prospects*, CRC Press / Taylor & Francis *(in press).*

Crowdsourcing Landscape Perceptions to Validate Land Cover Classifications

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

[1]The Pennsylvania State University, University Park, State College, PA 16801
Email: {kas5822; klippel; wallgrun} @psu.edu

[2]NCGIA & Department of Geography, University at Buffalo, Buffalo, NY 14228
Email: dmark@buffalo.edu

> "If every object and event in the world were taken as distinct and unique---a thing in itself unrelated to anything else---our perception of the world would disintegrate into complete meaninglessness. The purpose of classification is to give order to the thing we experience."
> --- Abler, R., Adams, J. S., & Gould, P., 1971

> Of all the countless possible ways of dividing entities of the world into categories, why do members of a culture use some groupings and not use others? What is it about the nature of the human mind and the way that it interacts with the nature of the world that gives rise to the categories that are used?
> --- Malt 1995

1. Introduction

The way humans understand their natural environments—landscapes—either as an individual or as a collective, frames prominent research topics in several disciplines. Landscape perception for instance has been analyzed for land management and planning purposes to characterize landscape aesthetics and objective scenic beauty. Understanding earth surface processes through terrain analysis has been relevant to military and civil engineering. From a geographic perspective, it can be argued that the *man-land tradition* or in more recent terms *human-environment relations* is nothing less than one of the four intellectual cores of geography (Pattison 1964, see also Mark et al. 2011).

Land cover data has been used for much more than looking up land cover at a given location, including uses for climate modeling, food security, and biodiversity monitoring. The focus of this book chapter is on how the abundance of freely available high resolution imagery of the earth’s surface and the maturing of crowd science offers new opportunities for an unprecedented access to environmental information. We will examine how crowd science and human perceptions can be used for the purposes of improving overall quality of land cover datasets. First we will discuss variation between land cover classifications and explore why they exist. Second we will discuss issues and debates surrounding the accuracy of land cover datasets. Last we will discuss shortcomings with current assessment processes and opportunities for new methods to assist in the overall quality of land cover datasets.

Although widely used, many global land cover datasets have unique characteristics that lead to differences between classifications. Variety is represented through differing land cover classes, changed meanings of shared terminology, and differences in interpretation and perception of the land cover classes. Ahlqvist (2008) discussed the importance of standardizing terminologies in science, using the topic of variation in land cover classifications as an example. He stresses the need for more interpretability, reflecting on the subjectivity not only in the creation process of land cover classifications, but also in the interpretation of the classification from the user. This point builds off of research by Comber et al. (2005) who discusses the varying conceptualizations of the world that geographic data are mapped into. They list examples of how terms such as *Forest* and *Beach* have varying meanings based on the purpose that the land cover dataset was created for, and how these terms are interpreted differently across cultures and among users. There is a growing demand for harmonization of data, especially in class descriptions (Jepsen and Levin, 2013). Recognizing this becomes more important as local environmental knowledge is increasingly being incorporated into land use and land cover analysis. Robbins (2003) exemplifies this through differences shown in land cover and land use classification choices made by foresters and herders, enforced by their respective cultural and political role in their community. This creates variation between the classifications generated by the producers, and variation between how a unique land cover class is perceived by the users.

Gaining a deeper understanding of the perception of land cover classes addresses significant challenges in the effectiveness of classification interpretability (Ahlqvist 2012). Comber et al. (2004) uses the example of the Great Britain Datasets LCM1990 and LCM2000 to illustrate how changes in methodology and semantics cause unknown variation between datasets, creating uncertainty between either observing land cover change, or simply observing a change in how the land cover is represented.

In order to reduce variability in interpretation of classification, land cover classifications must be concerned with users' natural concepts and perceptions of the land cover, and be aware of formal cognitive models about the common-sense geographic world. Coeterier (1996) concludes that even when comparing landscapes of great differences between inhabitants of those landscapes, there is agreement among the importance of higher level attributes of the landscape. Some of these attributes include the unity of the landscape, its use, maintenance, naturalness, and spaciousness. More so, these attributes are not necessarily independent from each other. Habron (1998) analyzes the perceptual differences of Wild Land across varying demographics in Scotland. He concludes that human presence/influence has a large effect on the perception of Wild Land. Furthermore, what is considered Wild Land varies between sections of the population, with a consensus on a core definition and variation at the periphery.

Along with classification and interpretation variation, land cover datasets have accuracy related issues. Foody (2002, 2008) has discussed the state of land cover dataset quality along with their corresponding accuracy assessments and has noted the debate surrounding accuracy expectations. Once accuracy assessments are performed, of which there are multiple different methods of assessment, the vast majority of land cover datasets do not meet the commonly recommended target of 85% accuracy. He further discusses that 85% is perhaps unrealistically high. This number was historically specified by Anderson et al. (1976) for mapping general land cover classes (Anderson, Level 1). Additionally that accuracy rate was inspired by work associated with USDA's Census of Agriculture in where 85%, "would be comparable to the accuracy of land-cover maps derived from aerial photograph interpretation." (Foody, 2008, pg. 3140). Yet in the face of potentially harsh critiques of accuracy, land cover nonetheless can benefit from new approaches of classification and assessment. Wilkinson's (2005) 15 year survey of published papers on satellite image classification revealed no upward trend in classification accuracy. This creates an opportunity for unconventional classification approaches to assist in a severe lack of advancing accuracy rates.

As previously mentioned, once land cover datasets are created, accuracy assessments are often performed to measure the quality of the dataset. Comber et al. (2012) notes that a common approach of measuring land cover accuracy is to compare the dataset with corresponding data that is considered to be of a higher accuracy. This could mean a collection of relatively sparse ground data acquired in the field used as control points. Comber discusses that these approaches of assessment overlook the spatial distribution of errors, leading to possible localized sub-regions of high inaccuracy that distort the global accuracy. Furthermore, Foody (2002) reiterates that land cover is dynamic. The earth's surface will inevitably change in the time it takes to update datasets. This creates an opportunity for new methods of assessment and classification that are flexible, has the capability of covering large areas, and are quick relative to classic approaches.

New (unconventional) methods need evaluation. While there is a lot of interest surrounding the opportunities of the crowd, examples of which will be explained in detail in section 2, there is a high demand for systematic evaluations of how much improvement in land cover classification can be achieved using crowd-based assessments. In response to these challenges, this chapter analyzes the correspondence between human conceptualizations of land cover and spectrally derived land cover datasets. If crowdsourced human participants are 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 analyze crowdsourced human participants' ability to recognize existing land cover classes given on the ground photographs. We are reporting on three experiments that provide insights on the relationships between human conceptualizations of land cover and land cover classifications using novices, educated novices, and experts. Our findings suggest misclassifications are not random but rather systematic to unique landscape stimuli and unique land cover classes. By comparing novices and experts we are able to evaluate the potential for using crowdsourcing in aiding the advancement of land cover classifications.

2. Background

Recent growth and improvement to online crowdsourcing platforms now allow for large-scale contributions to scientific research. These crowd contributions have shown to be successful across many disciplines, including the discovery of protein structures (Khatib et al. 2011) and identification of new galaxies (Clery 2011). In the context of land cover, using crowdsourced human participants to validate global land cover datasets has been recognized by previous research. This use of crowdsourcing shows promise, especially as the use of more than one dataset often provides more accurate land cover mapping (Aitkenhead and Aalders, 2011). The Geo-Wiki project (Fritz et al. 2009) asks online participants to use aerial imagery via Google Earth as well as any local knowledge they may have to make classification choices on which land cover type they are observing given a predefined classification scheme. This volunteer geographer approach complements the classification and accuracy assessments in use, but at the moment fails to guarantee a level of quality in the volunteered data.

The Land Use/Cover Area Frame Survey (LUCAS) (http://www.lucas-europa.info/) is an example of a more authoritative, non-crowd based attempt at capturing land cover data. LUCAS, commissioned by Eurostat, deploys land surveyors to many locations across the European Union to determine land cover/land use, record transects, and take photographs of the landscape. By virtue of LUCAS's means of data collection, creating a comprehensive dataset using this method would be highly improbable. Using these data as a means of validation however is more likely.

For the purposes of measuring land cover, due to the complexity of the earth's surface, all measurements contain error to some unknown extent. It is thus very difficult to precisely describe and categorize features of land cover. This error is true for both remote sensing classification and classification via human interpretation of aerial imagery. Foody (2002) discusses this from the perspective of remote sensing to the degree that ground truth measurements are still a classification and thus contains some degree of error. Kinley (2013) and Hoffman and Pike (1995) discuss this from the perspective of volunteered data and terrain analysis, stating that these data and terrain descriptions are often critiqued harshly for not meeting an impossible ideal. Yet in the face of inescapable error in land cover data and volunteered data, steps must be taken to ensure the methods for collecting data allow for the opportunity of the highest quality products. This means understanding humans' concepts and perceptions of land cover in order to assist in the classification process.

The majority of experiments measuring quality of crowdsourced volunteered land cover classifications come from experiments run through the Geo-Wiki project (Perger et al. 2012, See et al. 2013, Foody et al. 2013a, Comber et al. 2013, Comber et al. 2014). See et al. (2013) most notably reports on an experiment which expert and non-expert participants during a Geo-Wiki campaign were asked to classify land cover given aerial imagery for the purposes of measuring participant accuracy rates, and comparing expert and non-expert results. Control points generated by three experts visually classifying land cover from aerial imagery were used to measure how accurate the crowdsourced participants' classifications were. Averaged accuracy rates range from 66%-76% for the full set of participants, with experts reaching a maximum of 84%, and non-experts reaching a maximum of 65%. Comber et al. (2013) also uses crowdsourced classification data gathered from Geo-Wiki but focuses on the level of agreement between expert and non-expert classification of land cover type, rather than reporting accuracy rates measured against control points. They conclude by illustrating map outputs that show obvious visual differences between expert and non-expert classification choices, and call for "…further investigation into formal structures to allow such differences to be modelled and reasoned with" (Comber et al. 2013, pg 257). Comber et al. (2014) further states that expertise in classification has a general influence but is varied across land cover classes.

Similarly to Geo-Wiki, the OpenStreetMap (http://www.openstreetmap.org) dataset is comprised of crowdsourced geographic information that research has identified as potential data to assist, support, and validate other land use mapping projects. Arsanjani et al. (2013) has analyzed OpenStreetMap contributions to analyze the accuracy of participants' land use (opposed to land cover) classifications in an urban setting compared to other non-crowdsourced land use datasets. He concludes that OpenStreetMap, and in general other forms of crowdsourced geographic data, can be valid data sources for mapping land use.

Perger et al. (2012) notes how land cover can be difficult to classify when only given aerial imagery. Deviating from classification via aerial imagery, others have attempted to measure the effectiveness of using on-the-ground photographs for the purposes of land cover classification (Iwao et al. 2006, Foody et al. 2013b). The data source of these ground based photographs come from the Degree Confluence Project (DCP) which will be explained in detail later in the chapter. While the Iwao and Foody papers both report land cover classification accuracy rates, the main intention of their research was to test the validity of using DCP data to classify land cover. Both conclude DCP data is a valid data source when attempting to classify land cover.

To summarize, land cover classifications have not experienced significant accuracy improvements in the past 20 years. Research has recognized an opportunity to benefit from advancing technologies and improvements in crowd science to assist in the evaluation of land cover. This has largely been experimented through providing crowdsourced participants aerial imagery of the earth surface and asking for their classifications of the land cover. Aerial imagery however can sometimes provide a lack of information when distinguishing between similar land cover classes. Other research has proven the validity of using on-the-ground photographs for land cover classification, but has failed to test it with crowdsourced participants and a wide range of land cover classes.

3. Experiments

We conducted three experiments to shed light on humans' understanding of and ability to classify land cover according to official National Land Cover Dataset (NLCD) 2006 classes (Fry et al. 2011). The first two experiments involve lay people (without and with intervention), while the third uses experts.

3.1 Experiment 1 - Lay people, no intervention

The first experiment addresses the question whether lay people can classify images of land cover according to existing land cover classes. While the ground truth itself, that is, the NLCD 2006, only has a level II accuracy of 78% (Wickham et al. 2013), it serves as a starting point for improving the understanding of how humans perceive land cover classes.

3.1.1 *Materials*

Two datasets were used for this experiment: on-the-ground-photographs of landscapes provided by The Degree Confluence Project (DCP) (confluence.org), and the NLCD 2006 provided by the Multi-Resolution Land Characteristics Consortium (http://www.mrlc.gov).

The DCP is a site that provides a platform for collecting crowdsourced photographs of landscapes at confluence points across the world in a systematic way. The word confluence as defined for the purposes of the DCP is the location where two integer latitude and longitude coordinate lines meet. An example of this would be 'latitude 42 N, Longitude 100 W' as opposed to 'latitude 42.65 N, longitude 100.23 W'. Users are encouraged to visit these locations, take photographs of the landscape, and upload the images with metadata such as date visited and travel information.

For the scope of these experiments, we constrained our data collection to the contiguous United States. A total of 799 photographs were collected out of a possible 856. In an attempt to be consistent in data collection, north facing photographs were collected when at all possible. Two sampling criteria restricted the data collection process: First, scenes that included snow in the photograph were excluded as this is not reflective of the land cover but rather temporal weather conditions. Second, images that included human presence were excluded. Outside of these sampling restrictions, few confluences do not have photographs uploaded to the website, and as such, could not be collected.

Latitude and longitude coordinates from the DCP dataset were extracted and converted into a point shapefile to be used in ESRI's ArcGIS software (Figure 1). This allowed for the extraction of the corresponding land cover class from NLCD level II (16 land cover classes) for each confluence point and its corresponding image. Out of the 16 possible classes from NLCD level II, 11 were used in the experiments: we aggregated *deciduous forest*, *evergreen forest*, and *mixed forest* into one *forest* class. Additionally, the following three classes did not provide sufficient sampling points to ensure balanced class representation, and a suitable number of total images: *developed medium intensity*, *developed high intensity*, and *perennial ice/snow*. From the remaining 11 classes 7 locations and associated images were randomly selected (stratified random sampling), resulting in 77 images shown partly in figure 2. To ensure that confluences were not on the boundary of two land cover classes, confluences were selected when located in a homogenous land cover region of at least 90 meters (3 NLCD pixels) in the direction the photo was taken. Although land cover change has the possibility of influencing incorrect land cover extraction, each of the 77 images were analyzed together with their corresponding land cover class to ensure consistency between land cover features in the images, and assigned land cover classes. It is important to note that Wickham et al. (2013) accuracy assessment of NLCD 2006 for the contiguous United States concludes that level I (8 aggregated land cover classes) accuracy is equal to 84% and level II accuracy is equal to 78%.

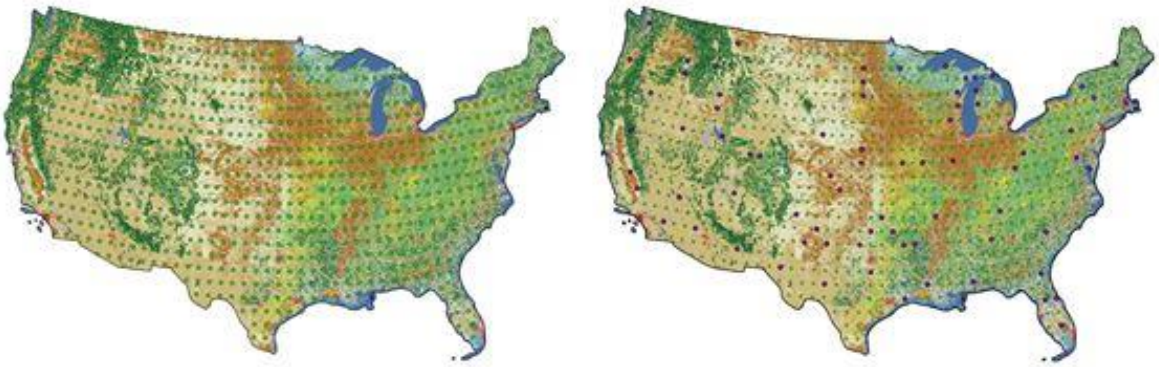


Figure 1. The NLCD 2006 overlaid by confluence points (left). Stratified random sampled confluence points, 77 total sampled, 7 in each land cover class (right).

3.1.2 *Participants*

20 lay participants (non-experts, 5 female) were recruited through the crowdsourcing platform Amazon Mechanical Turk (AMT); average age 32.2 years; reimbursement: $1.25.

3.1.3 *Procedure*

The experimental software CatScan (Klippel et al. 2008) used for the experiment has been designed to be serviceable in combination with AMT (Figure 2). In the experiment, each participant performed a non-free classification task. During the non-free classification, all images were initially displayed on the left panel of the screen. On the right side of the screen, the 11 land cover classes were displayed into which participants were able to drag icons from the left panel into the classes on the right panel. It was possible to leave classes empty.

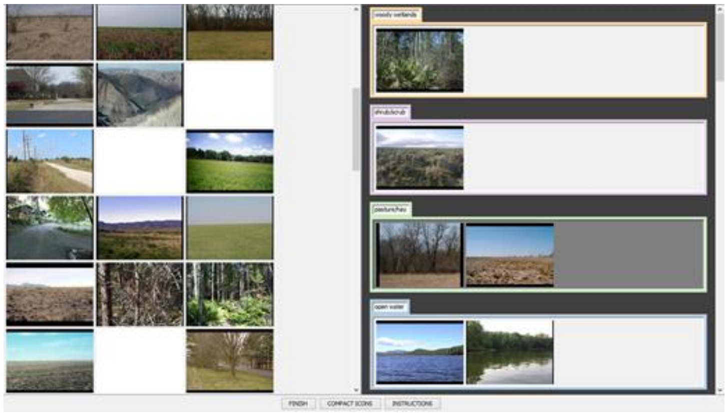


Figure 2. Screenshot of the CatScan interface of an ongoing mock-up experiment.

3.1.4 *Results*

The classification results should be interpreted with consideration to Wickham's (2013) accuracy assessment in mind. To reiterate, our sample from the NLCD was taken from the level II classification, which Wickham concludes is 78% accurate. There exists, however, accuracy variation among classes in the NLCD, and Wickham stresses the need for improved distinction among grass-dominated classes (*develop, open space* (dO), *grassland* (GS), *pasture/hay* (PH), *cultivated crops* (CC), and *emergent herbaceous wetland* (EW)), as they account for higher classification error relative to the other classes.

Participants used an average of 10.25 classes (out of the possible 11) with a standard deviation of 1.07. The average grouping time was 665.86 seconds (11 minutes 5 seconds) with a standard deviation of 263.73 seconds (4 minutes 23 seconds).

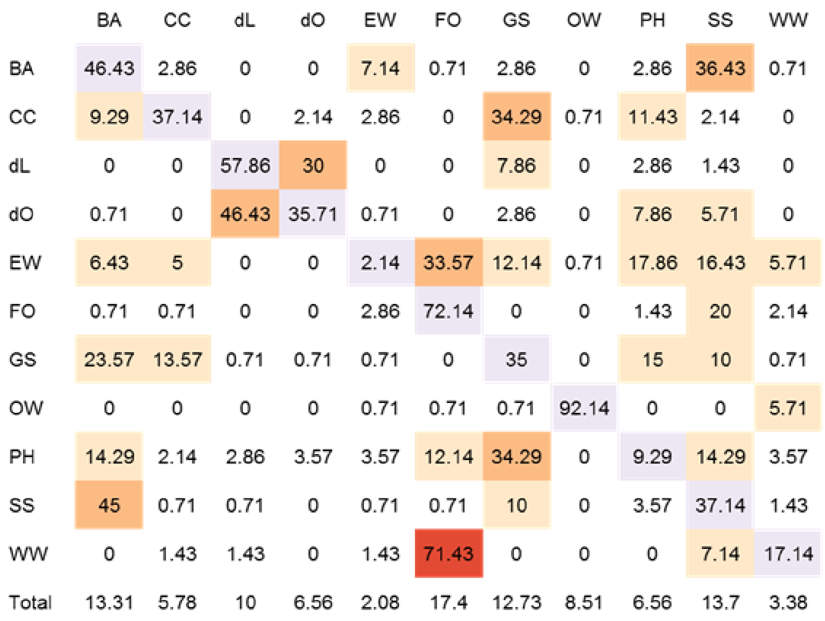
To analyze the classification results, we created a confusion matrix (Figure 3) that not only shows the number of correctly classified land cover images but additionally reveals how images were misclassified; the confusion matrix shows in which class an image was placed and whether or not this was the correct class. We performed chi square tests to corroborate the interpretation statistically. Several main observations can be summarized as follows.

Overall classification accuracy for experiment 1 is approximately 40.19%. Against the relatively low overall accuracy of the classification task, the following land cover classes were significantly classified correctly 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). In contrast, the following land cover classes were significantly classified less correctly than expected by having a standardized residual value less than -1.96: *emergent herbaceous wetlands* (EW), *pasture*/*hay* (PH), *woody wetlands* (WW).

Table 1. Standardized residuals for experiment 1

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
correct	1.57	-0.77	4.47	-1.13	-9.63	8.08	-1.31	13.14	-7.82	-0.77	-5.83

As participants proceed through the experiment, CatScan records the land cover class that an image is placed in. Correct classification is assumed based on the land cover class the image is sampled from (see Figure 1). Organizing this data in form of confusion matrices allows for reviewing the classification behavior of all participants and assessing both correct and incorrect classifications. The confusion matrix below (Figure 3) shows the classification behavior in percentages; results can be summarized as follows: The *Woody wetlands* (WW) class is almost exclusively confused with *forest* (FO). Participants are generally successful in recognizing *developed* land cover but confuse *developed, open space* (dO) and *developed, low intensity* (dL), having more success classifying *developed, low intensity* (dL). Participants almost exclusively confuse *barren* (BA) and *shrub/scrub* (SS) with each other. The *emergent herbaceous wetlands* (EW), *grassland* (GS), and *pasture/hay* (PH) classes are confused across many classes.



	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	46.43	2.86	0	0	7.14	0.71	2.86	0	2.86	36.43	0.71
CC	9.29	37.14	0	2.14	2.86	0	34.29	0.71	11.43	2.14	0
dL	0	0	57.86	30	0	0	7.86	0	2.86	1.43	0
dO	0.71	0	46.43	35.71	0.71	0	2.86	0	7.86	5.71	0
EW	6.43	5	0	0	2.14	33.57	12.14	0.71	17.86	16.43	5.71
FO	0.71	0.71	0	0	2.86	72.14	0	0	1.43	20	2.14
GS	23.57	13.57	0.71	0.71	0.71	0	35	0	15	10	0.71
OW	0	0	0	0	0.71	0.71	0.71	92.14	0	0	5.71
PH	14.29	2.14	2.86	3.57	3.57	12.14	34.29	0	9.29	14.29	3.57
SS	45	0.71	0.71	0	0.71	0.71	10	0	3.57	37.14	1.43
WW	0	1.43	1.43	0	1.43	71.43	0	0	0	7.14	17.14
Total	13.31	5.78	10	6.56	2.08	17.4	12.73	8.51	6.56	13.7	3.38

Figure 3. Confusion matrix for experiment 1 (lay participants with no intervention) showing percentages of correct (diagonal) and misclassified landscape images (rows). Misclassified classes between 5% and 25% are indicated by light pink, misclassifications between 25% and 50% are light orange, and misclassifications above 50% are red. The 'Total' row indicates the percentage of classification choices made in each class.

3.1.5 *Discussion*

Comparing experiment 1 to Wickham's (2013) analysis of the NLCD 2006 accuracy, both human classification, and NLCD classification have relative difficulty in classifying *emergent herbaceous wetlands* (EW) and *pasture/hay* (PH). From this we can speculate that both visual stimuli, and spectral characteristics of the land cover features in *emergent herbaceous wetlands* (EW) and *pasture/hay* (PH) are not well defined and cause confusion between classes.

Using the *woody wetlands* (WW) class as an example, visually classifying certain land cover features cannot be done with relatively high levels of confidence, whereas spectrally it can. This could be a case of remote sensors' ability to collect data outside the visual spectrum leading to clear distinctions between, say, forest and woody wetland via soil and vegetation moisture. Visually recognizing this distinction from DCP data proves to be very difficult for human classification (17% accuracy for WW).

Although human classification and NLCD classification may start to have similar relative inaccuracies for *developed*, *open space* (dO), the confusion matrix shows humans being very successful in generally identifying *developed*. While remote sensors may have difficulty distinguishing spectral characteristics between developed features and natural features, this may not be as difficult a task for human classification. Developed features, although potentially spectrally similar to certain natural land cover features surrounding it, become easily identifiable for humans to visually interpret and distinguish from surrounding natural features.

3.2 Experiment 2 - Lay people, intervention

Intrigued by the findings of experiment 1, especially by the overall low number of correctly classified images, we designed an intervention described in Section 3.2.3. The goal of this intervention was to reduce confusion between land cover classes and increase classification accuracy.

3.2.1 *Materials*

Same as experiment 1.

3.2.2 *Participants*

20 new lay participants (non-experts, 11 female) were recruited through AMT; average age 34.2 years; reimbursement: $1.25.

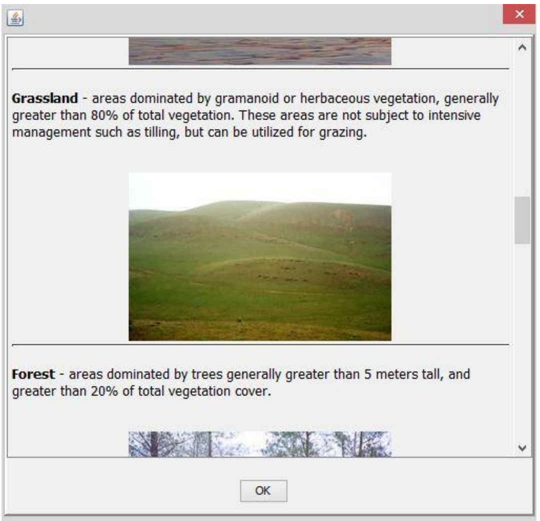
3.2.3 *Procedure*

The main procedural difference between experiment 1 and experiment 2 was the inclusion of the NLCD land cover class definitions as defined on the Multi-Resolution Land Characteristics Consortium website (http://www.mrlc.gov), and associating prototypical images for each land cover class with the definition (Figure 4). The images were sourced from the DCP, and were assigned to each definition based off of their associating extracted NLCD class. These definitions and prototypical images were shown to the participants before they began the experiment and were available to revisit throughout the entire experiment.

3.2.4 *Results*

Participants used an average of 10.65 classes (out of the possible 11) with a standard deviation of 0.59. The average grouping time was 822.01 seconds (13 minutes 42 seconds) with a standard deviation of 326 seconds (5 minutes 26 seconds).

The overall accuracy is 44.35%. The improvement in classification by lay participants after the intervention is statistically significant ($\chi^2 = 5.2807$, df = 1, p = .02), with *developed*, *open space* (dO) specifically benefiting from the intervention, increasing its accuracy 22.15%



Grassland - areas dominated by gramanoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.

Forest - areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover.

Figure 4. An example of what the lay people see before and during the experiment.

from experiment 1. This relatively high accuracy of *developed, open space* (dO) contrasts with the confusion between grass-dominated classes in NLCD that Wickham (2013) notes is relatively inaccurate.

Against the relatively low overall accuracy of the classification task, the following land cover classes were significantly classified correctly more frequently than expected by having a standardized residual value greater than 1.96 (Table 2): *developed, open space* (dO), *forest* (FO), *open water* (OW). In contrast, the following land cover classes were significantly classified less correctly than expected by having a standardized residual value less than -1.96: *emergent herbaceous wetlands* (EW), *pasture/hay* (PH), *woody wetlands* (WW).

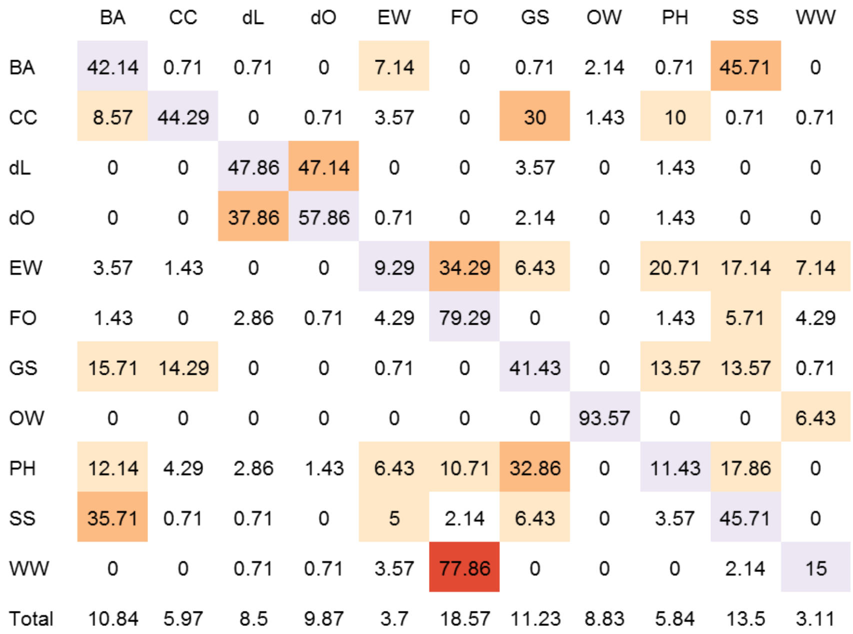
Table 2. Standardized residuals for experiment 2

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
correct	-0.55	-0.01	0.87	3.37	-8.75	8.72	-0.72	12.29	-8.22	0.34	-7.33

When examining the confusion matrix below (Figure 5) in comparison to the confusion matrix for experiment 1 (Figure 3), general relationships between classes persist but changes occur in magnitudes of accuracy. As in experiment 1, experiment 2 also results in almost exclusive confusion of *woody wetland* (WW) being misclassified as *forest* (FO), *barren* (BA) and *shrub/scrub* (SS) being confused with each other, the *developed* classes being confused with each other, and the *emergent herbaceous wetlands* (EW), *grassland* (GS), and *pasture/hay* (PH) confused across many classes. Differences between the experiments were as follow: Participants classified *developed, open space* (dO) more accurately than *developed, low intensity* (dL) in experiment 2, compared to participants classifying *developed, low intensity* (dL) more accurately than *developed, open space* (dO) in experiment 1. Participants confused *barren* (BA) with *shrub/scrub* (SS) more often, and confused *shrub/scrub* (SS) with *barren* (BA) less often in experiment 2, compared to experiment 1.

3.2.5 *Discussion*

Referring to the grass dominated classes that Wickham (2013) notes are the cause for most confusion in NLCD 2006, while all grass dominated classes increase in varying degrees of accuracy from experiment 1 to experiment 2, *developed*, *open space* (dO) by far benefits the most from the intervention, increasing 22.15%. The inclusion of the intervention changes *developed*, *open space* (dO) from a cause of confusion in experiment 1, similar to NLCD relative confusion, to a class that is relatively accurate. We can speculate then that even though *developed*, *open space* (dO) may need more distinction to decrease confusion for NLCD, human classification of this class is relatively accurate when provided land cover class definitions. This further would indicate that the land cover class definition for *developed*, *open space* (dO) creates more clarity, whereas the land cover class definition for *developed, low intensity* (dL) introduces more confusion.



	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	42.14	0.71	0.71	0	7.14	0	0.71	2.14	0.71	45.71	0
CC	8.57	44.29	0	0.71	3.57	0	30	1.43	10	0.71	0.71
dL	0	0	47.86	47.14	0	0	3.57	0	1.43	0	0
dO	0	0	37.86	57.86	0.71	0	2.14	0	1.43	0	0
EW	3.57	1.43	0	0	9.29	34.29	6.43	0	20.71	17.14	7.14
FO	1.43	0	2.86	0.71	4.29	79.29	0	0	1.43	5.71	4.29
GS	15.71	14.29	0	0	0.71	0	41.43	0	13.57	13.57	0.71
OW	0	0	0	0	0	0	0	93.57	0	0	6.43
PH	12.14	4.29	2.86	1.43	6.43	10.71	32.86	0	11.43	17.86	0
SS	35.71	0.71	0.71	0	5	2.14	6.43	0	3.57	45.71	0
WW	0	0	0.71	0.71	3.57	77.86	0	0	0	2.14	15
Total	10.84	5.97	8.5	9.87	3.7	18.57	11.23	8.83	5.84	13.5	3.11

Figure 5. Confusion matrix for experiment 2 (lay participants with intervention).

Participants confused *shrub/scrub* (SS) with *barren* (BA) less, but confused *barren* (BA) as *shrub/scrub* (SS) more. This indicates that the intervention convinced participants that *shrub/scrub* (SS) includes more land cover possibilities than perhaps initially thought, while the intervention narrowed the possibilities of what might be considered *barren* (BA).

3.3 Experiment 3 - Experts

Given the potential for errors based on the accuracy of the level II NLCD data (78%) we also investigated how experts would classify the images we sampled.

3.3.1 *Materials*

Experts were provided the class definitions and visual prototypes of each land cover class, just like in experiment 2, but on printed out sheets of paper.

3.3.2 *Participants*

Four experts were solicited that have ecological and geographic information science backgrounds with experience in working with land cover data.

3.3.3 *Procedure*

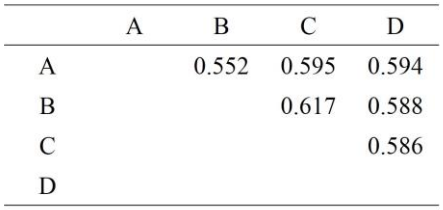
Each expert viewed the original DCP images on a computer screen, one at a time. As previously mentioned in the materials section, they were each given a printed out copy of the class definitions and visual prototypes of each land cover class. Each expert viewed the original DCP images on a computer screen one at a time, and recorded their classification choice on a sheet of paper.

3.3.4 *Results*

The classifications by each expert were compared against those from each other expert in order to establish levels of agreement between experts. We represent agreement as Cohen's kappa coefficient (Figure 6) and percent agreement (Figure 7). Cohen's kappa coefficient is a measure of inter-rater agreement for categorical objects. It expands on general percentage agreement and takes into account the likelihood of random agreement. The coefficient is defined by the following equation:

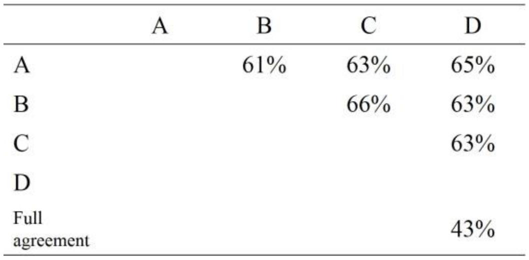
$$\kappa = \frac{\rho_o - \rho_c}{1-\rho_c} \qquad (13.1)$$

Where ρ_o is the observed proportion of agreement and ρ_c is the proportion of agreement expected by chance. If the raters are in perfect agreement then κ = 1. If the raters agreement is what would be expected by chance then κ = 0. Foody (2013b) uses this coefficient as an index of the level of inter-rater agreement in an experiment of classifying presence of forest (forest, or non-forest) given DCP images.



	A	B	C	D
A		0.552	0.595	0.594
B			0.617	0.588
C				0.586
D				

Figure 6. Cohen's Kappa coefficient values between the experts, A-D.



	A	B	C	D
A		61%	63%	65%
B			66%	63%
C				63%
D				
Full agreement				43%

Figure 7. Percent agreement between the experts, A-D. Full agreement indicates the percentage that all 4 experts agreed on the same classification given a DCP image.

The overall accuracy is 48.37%. There is no statistically significant difference between educated lay participants (experiment 2) and experts ($\chi^2 = 1.52$, df = 1, p = .22). The most notable change from experiment 2 to experiment 3 is the increase of *cultivated crop* (CC) accuracy (23.57%).

Against the relatively low overall accuracy of the classification task, the following land cover classes were significantly classified correctly more frequently than expected by having a standardized residual value greater than 1.96 (Table 3): *cultivated crops* (CC), *developed, low intensity* (dL), *forest* (FO), *open water* (OW). In contrast, the following land cover classes were significantly classified less correctly than expected by having a standardized residual value less than -1.96: *barren* (BA), *emergent herbaceous wetlands* (EW), *grassland* (GS), *woody wetlands* (WW).

Table 3. Standardized residuals for experiment 3

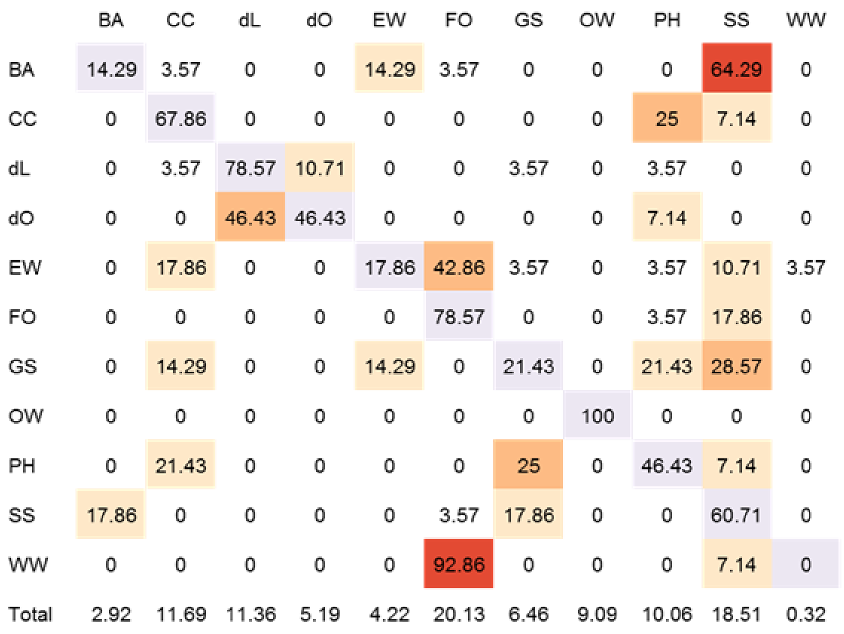
	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
correct	-3.78	2.16	3.35	-0.21	-3.38	3.35	-2.99	5.73	-0.21	1.37	-5.37

Referencing Wickham's (2013) analysis of NLCD accuracy, we see experts performed relatively well where NLCD, experiment 1, and experiment 2 did not, in classifying *cultivated crops* (CC). Conversely, the experts match the NLCD and are relatively inaccurate in other grass dominated classes such as *emergent herbaceous wetlands* (EW) and *grassland* (GS).

When examining the confusion matrix below (Figure 8), the following relationships between classes found in experiment 1 and 2 persist in experiment 3: *woody wetlands* (WW) is almost exclusively confused as *forest* (FO), the confusion of *barren* (BA) as *shrub/scrub* (SS) continues to increase, the *developed* classes are confused between each other, and the *emergent herbaceous wetlands* (EW), *grassland* (GS), and *pasture/hay* (PH) confused across many classes. Differences between experiment 3 and the previous experiments are as follows: Experts were more successful in classifying *developed, low intensity* (dL) than *developed, open space* (dO), which is more similar to experiment 1, and had little confusion when classifying *developed, low intensity* (dL). Although accuracy for *open water* (OW) was high in the previous two experiments, experts were perfect in correctly classifying, and not confusing another class as *open water* (OW). Experts significantly classified *cultivated crops* (CC) correctly more frequently than expected, which was not accomplished in the previous experiments. Experts significantly classified *barren* (BA) and *grassland* (GS) less correctly than expected which was not accomplished in the previous experiments.

3.3.5 *Discussion*

Referring to the grass dominated classes that Wickham (2013) notes are the cause for most confusion in NLCD 2006, experts are relatively successful in classifying *cultivated crops* (CC), and relatively poor at classifying *grassland* (GS). The *cultivated crops* (CC) success differs from NLCD, experiment 1, and experiment 2's relative successes. This could indicate that experts are uniquely capable in recognizing anthropogenically induced patterns relating to crop fields that lay participants are unable to visually recognize, and remote sensors are unable to spectrally identify. Conversely, experts are unsuccessful in recognizing *grassland* (GS), almost equally confusing the class with 4 other classes. While the previous 2 experiments struggled with *grassland* (GS), this



	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	14.29	3.57	0	0	14.29	3.57	0	0	0	64.29	0
CC	0	67.86	0	0	0	0	0	0	25	7.14	0
dL	0	3.57	78.57	10.71	0	0	3.57	0	3.57	0	0
dO	0	0	46.43	46.43	0	0	0	0	7.14	0	0
EW	0	17.86	0	0	17.86	42.86	3.57	0	3.57	10.71	3.57
FO	0	0	0	0	0	78.57	0	0	3.57	17.86	0
GS	0	14.29	0	0	14.29	0	21.43	0	21.43	28.57	0
OW	0	0	0	0	0	0	0	100	0	0	0
PH	0	21.43	0	0	0	0	25	0	46.43	7.14	0
SS	17.86	0	0	0	0	3.57	17.86	0	0	60.71	0
WW	0	0	0	0	0	92.86	0	0	0	7.14	0
Total	2.92	11.69	11.36	5.19	4.22	20.13	6.46	9.09	10.06	18.51	0.32

Figure 8. Confusion matrix for experiment 3 (experts).

indicates that experts uniquely conceptualize *grassland* (GS) as a broader class that includes many other land covers that previous experiments do not consider.

The experts' success in recognizing *developed, low intensity* (dL) could indicate their ability to successfully recognize anthropogenic influences in land cover as also shown in the accuracy of *cultivated crops* (CC), and their overall ability to recognize *developed* classes. This is further indicated by confusion of *cultivated crops* (CC) mostly with *pasture/hay* (PH) which is another class that has some degree of anthropogenic influence by definition.

4. Conclusions / Outlook

The overall match between participants' classifications and NLCD is rather low (40.19 - 48.37%). Accuracy increased statistically significantly using an intervention of providing definitions and prototypical images as examples as mentioned previously. The misclassifications are not random but rather systematic. This is the case on the level of land cover classes as well as on the level of individual images.

Classification accuracy naturally increases the more land cover classes are aggregated. The Anderson Level 1 classification groups *pasture/hay* (PH) and *cultivated crops* (CC) as a single land cover class, all of the *developed* classes as a single land cover class, and *woody wetlands* (WW) and *emergent herbaceous wetlands* (EW) as a single land cover class. Even though other research that analyzes the quality of human classification of land cover (Perger et al. 2012, See et al. 2013) gives the human participants a similar amount of land cover classes to choose from (10 classes compared to our 11), accuracy results are either presented after some level of aggregation to account for potential confusion between similar land cover classes (Perger et al. 2012), or some land cover classes' accuracy results omitted (See et al. 2013). When using humans to classify land cover, the level of aggregation in class representation becomes a heavily influencing factor. As seen in the results above, humans are much more accurate in discerning specific land cover classes, and naturally more accurate overall when distinguishing between fewer land cover classes.

See et al. (2013) shows results of *shrub cover*, *grassland*, and *mosaiced cropland* as having the lowest accuracies. They thus argue that there is a need to provide more examples of how classes that are often confused are represented specifically within Google Earth. When comparing experiment 3 (experts) results to the land cover classes that were most often confused in See's study (most specifically *shrub cover* and *mosaiced cropland*), human classification accuracy is relatively high in our experiment for those land cover classes when using on the ground photographs. This perhaps indicates the necessity for more contextual information when classifying particular land cover classes, such as shrub and crop type land cover.

When assigning complex tasks to be performed by the crowd, one must ensure that the volunteered data quality is appropriate and sustainable. In the context of land cover validation, humans are very successful in correctly classifying certain land cover via on the ground photographs, and poor in classifying others. Lessons learned from these three experiments are currently integrated in additional experiments that will, among other things, provide additional information about the area to be classified in form of aerial images, ask participants to perform classifications along individual dimensions, and allow for an indication of uncertainty of classifications.

Acknowledgments

This research is funded by the National Science Foundation under grant #0924534. We would like to thank the Degree Confluence Project for permission to use photos from the confluence.org website for our research.

References

Abler, R., J. S. Adams, and P. Gould. *Spatial organization: the geographer's view of the world*: Prentice-Hall, (1971).

Aitkenhead, M. J., and I. H. Aalders. "Automating land cover mapping of Scotland using expert system and knowledge integration methods." *Remote Sensing of Environment* 115.5 (2011): 1285–95.

Ahlqvist, Ola. "In search of classification that supports the dynamics of science: the FAO Land Cover Classification System and proposed modifications." *Environ. Plann. B* 35.1 (2008): 169–86.

Ahlqvist, Ola. . "Semantic issues in land cover studies – representation and analysis". *Remote Sensing of Land Use and Land Cover: Principles and Applications* Edited by Giri, C. Boca Raton, FL: Taylor and Francis. (2012): 25-36.

Anderson, J. R., et al. "A Land Use And Land Cover Classification System For Use With Remote Sensor Data." *Geological Survey Professional Paper 964* (1976).

Arsanjani, Jamal J., et al. "Toward mapping land-use patterns from volunteered geographic information." *International Journal of Geographical Information Science* 27.12 (2013): 2264–78.

Clery, D. "Galaxy Zoo Volunteers Share Pain and Glory of Research." *Science* 333: (2011): 173–75.

Coeterier, J. F. "Dominant attributes in the perception and evaluation of the Dutch landscape." *Landscape and Urban Planning* 34.1 (1996): 27–44.

Comber, Alexis, Peter Fisher, and Richard Wadsworth. "Integrating land-cover data with different ontologies: identifying change from inconsistency." *International Journal of Geographical Information Science* 18.7 (2004): 691–708.

Comber, Alexis, et al. "What is land cover?" *Environment and Planning B* 32 (2005): 199–209.

Comber, Alexis, et al. "Spatial analysis of remote sensing image classification accuracy." *Remote Sensing of Environment* 127 (2012): 237–46.

Comber, Alexis, et al. "Comparing Expert and Non-expert Conceptualisations of the Land: An Analysis of Crowdsourced Land Cover Data." *Spatial Information Theory*. Ed. Thora Tenbrink, et al. Springer International Publishing, (2013). 243–60.

Comber, Alexis, Linda See, and Steffen Fritz. "The Impact of contributor Confidence, Expertise and Distance on the Crowdsourced Land Cover Data Quality." *GI_Forum 2014 - Geospatial Innovation for Society* (2014).

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

Foody, Giles M. "Harshness in image classification accuracy assessment." *International Journal of Remote Sensing* 29.11 (2008): 3137–58.

Foody, G. M., et al. "Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project." *Transactions in GIS* 17.6 (2013a): 847–60.

Foody, Giles M., and Doreen S. Boyd. "Using Volunteered Data in Land Cover Map Validation: Mapping West African Forests." *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 6.3 (2013b): 1305–12.

Fritz, Steffen, et al. "Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover." *Remote Sensing* 1.3 (2009): 345–54.

Fry, J., et al. "Completion of the 2006 national land cover database for the conterminous United States." *Photogrammetric Engineering & Remote Sensing* 77.9 (2011).

Habron, Dominic. "Visual perception of wild land in Scotland." *Landscape and Urban Planning* 42.1 (1998): 45–56.

Hoffman, Robert, and Richard Pike. "On the Specification of the Information Available for the Perception and Description of the Natural Terrain." *Local applications of the ecological approach to human machine systems*. Ed. Peter Hancock. Hillsdale, New Jersey: Lawrence Erlbaum Associates, (1995).

Iwao, Koki, et al. "Validating land cover maps with Degree Confluence Project information." *Geophys. Res. Lett.* 33 (2006).

Jepsen, Martin, and Gregor Levin. "Semantically based reclassification of Danish land-use and land-cover information" *International Journal of Geographical Information Science* 27.12 (2013): 2375-2390.

Khatib, Firas, et al. "Crystal structure of a monomeric retroviral protease solved by protein folding game players." *Nat Struct Mol Biol* 18.10 (2011): 1175–77.

Kinley, Laura. "Towards the use of Citizen Sensor Information as an Ancillary Tool for the Thematic Classification of Ecological Phenomena." *Proceedings of the 2nd AGILE (Association of Geographic Information Laboratories for Europe) PhD School 2013* (2013).

Klippel, A., M. Worboys, and M. Duckham. "Identifying factors of geographic event conceptualisation." *International Journal of Geographical Information Science* 22.2 (2008): 183–204.

Malt, B. C. “Category Coherence In Cross-Cultural Perspective.” *Cognitive Psychology* 29.2 (1995): 85–148.

Mark, D., et al. “Landscape in language: An introduction.” *Landscape in language: An introduction*. Ed. D. Mark, et al. 4th ed. John Benjamins Publishing Company, (2011).

Pattison, William. “The Four Traditions of Geography.” *Journal of Geography* 63 (1964): 211–16.

Perger, Christoph, et al. “A Campaign to Collect Volunteered Geographic Information on Land Cover and Human Impact.” *GI_Forum 2012: Geovizualisation, Society and Learning* (2012).

Robbins, Paul. “Beyond Ground Truth: GIS and the Environmental Knowledge of Herders, Professional Foresters, and Other Traditional Communities.” *Human Ecology* 31.2 (2003): 233–53.

See, Linda, et al. “Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts.” *PLoS ONE* 8.7 (2013): e69958.

Wickham, James D., et al. “Accuracy assessment of NLCD 2006 land cover and impervious surface.” *Remote Sensing of Environment* 130 (2013): 294–304.

Wilkinson, G. G. “Results and implications of a study of fifteen years of satellite image classification experiments.” *IEEE Trans. Geosci. Remote Sensing* 43.3 (2005): 433–40.