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

1. Introduction

This paper analyzes the correspondence between human conceptualizations of landscapes and spectrally derived land cover classifications. Although widely used, global land cover data have known discontinuities in accuracy across different datasets. Computational accuracy assessments are performed to correct for errors, yet inaccuracies and disagreement persist (Foody 2002). With the emergence of crowdsourcing platforms large-scale contributions to validate land cover classification are now possible and practical. The potential use of crowdsourcing methods for validation purposes by having human volunteers check for inconsistencies in global land cover datasets has been recognized by previous research. The Geo-Wiki project (Fritz 2009), for instance, asks online participants to use aerial imagery via Google Earth as well as their local knowledge to validate whether or not the land cover/land use is being accurately represented by the land cover classification in question. This volunteer geographer approach complements the accuracy assessments in use, but fails to guarantee a level of quality in the volunteered data. If crowdsourced human participants are to be incorporated into accuracy assessments of land cover types, there needs to be some understanding of how humans perceive and conceptualize land cover types and rigorously measure how well humans perform in recognizing predefined land cover classifications.

We are reporting on three experiments that provide insights on the relationships between human conceptualizations of landscapes 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 land cover classifications.

2. Methods

Two datasets were used for this experiment: on-the-ground-photographs of landscapes provided by The Degree Confluence Project (DCP) (confluence.org), and the National Land Cover Dataset (NLCD) 2006 (Fry 2011) provided by the Multi-Resolution Land Characteristics (MRLC) consortium.

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. Research methods have shown a successful level of reliability in using DCP data for validating land cover classifications (Iwao 2006). Confluence as defined for the purposes of the DCP is the location where two integer latitude and longitude coordinates 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.

We constrained our data collection to the continental United States. A total of 799 photographs where collected out of a possible 856. 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 landscape or 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. This allowed for the extraction of the corresponding land cover class from NLCD level II (16 land cover classes) for each confluence point and the corresponding image. Although land cover change has the possibility of influencing incorrect land cover extraction, each of the 77 images were visually analyzed with their corresponding land cover class to ensure logical consistency. It is important to note that Wickham's (2013) accuracy assessment of NLCD 2006 for the conterminous United States concludes that level II accuracy = 78%. For the scope of this experiment we aggregated Deciduous-Forest, Evergreen-Forest, and Mixed-Forest into one "Forest" class, leaving a possible 14 land cover classes to sample from.

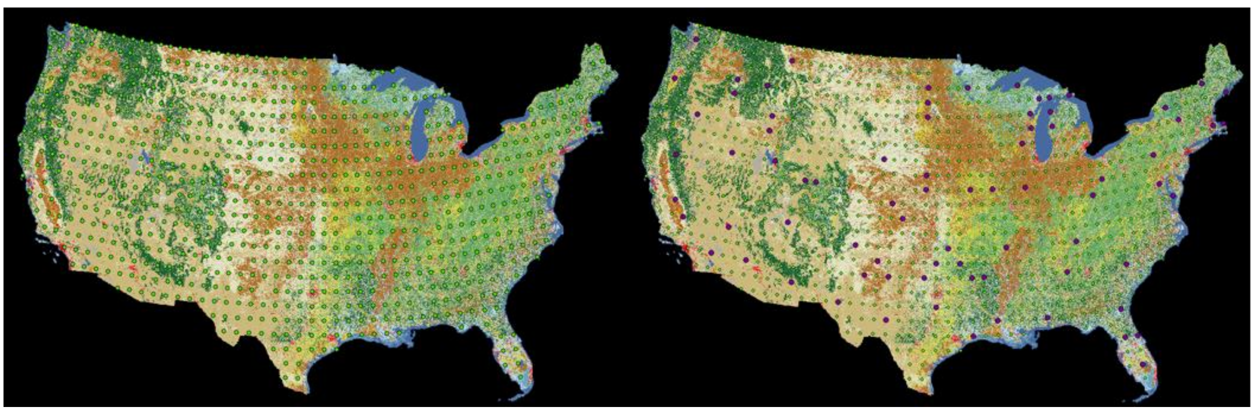

Figure 1. The NLCD 2006 overlaid by confluence points (left). Stratified random sampled confluence points, 77 total sampled, 7 in each land cover class (right).

The images, now each defined with a land cover class, were sorted into bins based on their land cover class. A total of 11 land cover classes were sampled from a possible 14 with Developed, Medium-Intensity, Developed High-Intensity, and Perennial Ice/Snow not represented. Seven images from each class were randomly selected, totaling 77 images.

The experimental software CatScan (Klippel 2008) used for the experiment has been designed to be serviceable in combination with Amazon's Mechanical Turk (AMT). In each experiment, participants performed a non-free classification task. 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. It was possible to leave classes empty.

Three experiments were conducted to provide insight on the relationships between human perceptions/conceptualizations of landscapes and land cover classifications. The first experiment solicited 20 lay participants (5 female) to perform the non-free classification test. The second experiment (20 lay participants, 11 female) included an intervention of definitions and visual examples for each land cover class. The third experiment used expert participants only (4 experts, ecological and GIS backgrounds with experience in working with land cover).

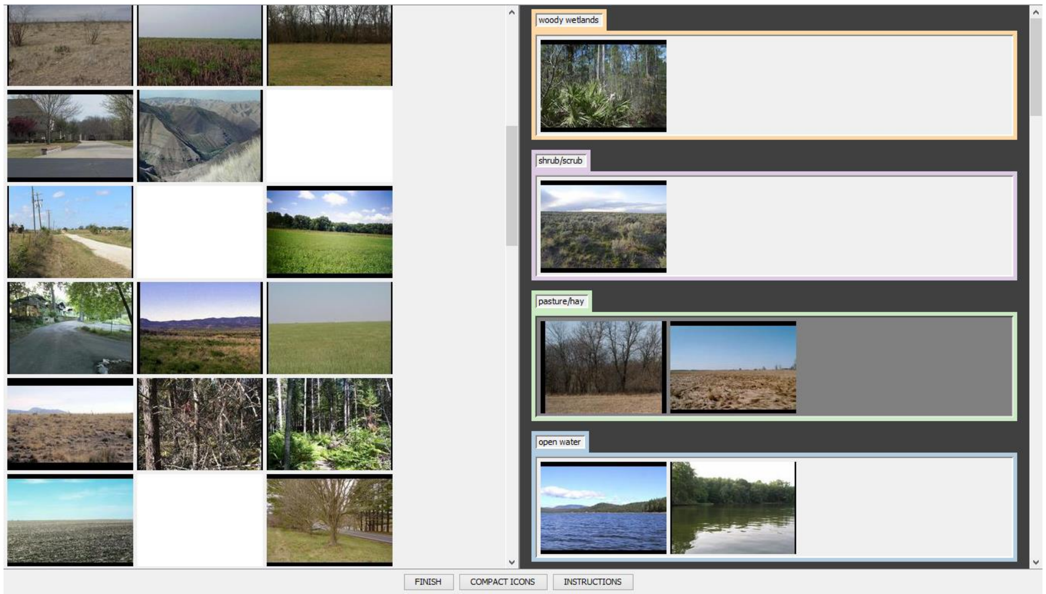

Figure 2. Screenshot of the CatScan interface of an ongoing mock-up experiment.

3. Results

To analyze the classification results, we created confusion matrices (Figure 3-5) that show the number of correctly classified land cover images and in which classes the misclassifications occur. We performed chi square tests to corroborate the interpretation statistically. Several main observations can be summarized as follows: First, there are statistically significant differences between the number of ‘correctly’ (the NLCD class is considered ‘correct’) identified land cover images across all 11 land cover classes in all three experiments. Second, the improvement in classification of lay participants after the intervention is statistically significant ($\chi^2 = 5.2807$, df = 1, p = .02). Third, there is no statistically significant difference between educated lay participants and experts ($\chi^2 = 1.52$, df = 1, p = .22). Forth, the overall match between participants’ classifications and NLCD is rather low (40.19 - 48.37%). This accuracy rate range still indicates the difficulty of human land cover classification even in the face of measuring it against the inaccuracies of NLCD 2006. If human classification was near perfect, then we would expect to see accuracy rates of approximately 78%, matching the NLCD 2006 level II accuracy rate.

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	46.43				7.14					36.43	
CC	9.29	37.14					34.29		11.43		
dL			57.86	30			7.86				
dO			46.43	35.71					7.86	5.71	
EW	6.43	5			2.14	33.57	12.14		17.86	16.43	5.71
FO						72.14				20	
GS	23.57	13.57					35		15	10	
OW								92.14			5.71
PH	14.29					12.14	34.29		9.29	14.29	
SS	45						10			37.14	
WW						71.43				7.14	17.14

Figure 3. Confusion matrix for experiment 1 showing percentages of correct (diagonal) and misclassified landscape images (rows). Total number of misclassified images smaller/equal 5% are blackened out, values between 5% and 25% are indicated by light pink areas; values between 25% and 50% are light orange and, misclassifications above 50% are red.

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	42.14				7.14					45.71	
CC	8.57	44.29					30		10		
dL			47.86	47.14							
dO			37.86	57.86							
EW					9.29	34.29	6.43		20.71	17.14	7.14
FO						79.29				5.71	
GS	15.71	14.29					41.43		13.57	13.57	
OW								93.57			6.43
PH	12.14				6.43	10.71	32.86		11.43	17.86	
SS	35.71				5		6.43			45.71	
WW						77.86					15

Figure 4. Confusion matrix for experiment 2.

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	14.29				14.29					64.29	
CC		67.86							25	7.14	
dL			78.57	10.71							
dO			46.43	46.43					7.14		
EW		17.86			17.86	42.86				10.71	
FO						78.57				17.86	
GS		14.29			14.29		21.43		21.43	28.57	
OW								100			
PH		21.43					25		46.43	7.14	
SS	17.86						17.86			60.71	
WW						92.86				7.14	0

Figure 5. Confusion matrix for experiment 3.

4. Discussion / Outlook

The overall 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.

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 landscapes 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

References

Foody G M, 2002, Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1):185-201.

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

Fry J, Xian G, Jin S, Dewitz J, Homer C, Yang L, Barnes C, Herold N, and Wickham J, 2011, Completion of the 2006 national land cover database for the conterminous United States. *PE&RS*, 77(9):858-864.

Iwao K, Nishida K, Kinoshita T, and Yamagata Y, 2006, Validating land cover maps with degree confluence project information. *Geophysical Research Letters*, 33(23).

Klippel A, Worboys M, and Duckham M, 2008, Identifying factors of geographic event conceptualisation. *International Journal of Geographic Information Science*, 22:2, 183-204

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