PITFALLS AND POTENTIALS OF CROWD SCIENCE: A META-ANALYSIS OF CONTEXTUAL INFLUENCES

A. Klippel*, K. Sparks, J. O. Wallgrün

Department of Geography, The Pennsylvania State University
University Park, Pennsylvania, USA
(klippel, kas5822, wallgrun)@psu.edu

Commission II, WG II/4

KEY WORDS: Crowd Science, Language, Methods, Evaluation

ABSTRACT:

Crowd science is becoming an integral part of research in many disciplines. The research discussed in this paper lies at the intersection of spatial and behavioral sciences, two of the greatest beneficiaries of crowd science. As a young methodological development, crowd science needs attention from the perspective of a rigorous evaluation of the data collected to explore potentials as well as limitations (pitfalls). Our research has addressed a variety of contextual effects on the validity of crowdsourced data such as cultural, linguistic, regional, as well as methodological differences that we will discuss here in light of semantics.

1. INTRODUCTION

Crowd science (here interchangeably used with *crowdsourcing*) is becoming an integral part of current research in many disciplines (Khatib et al., 2011, Clery, 2011). Two of the greatest beneficiaries of crowd science are the spatial and behavioral sciences. As a young methodological development, crowd science needs attention from the perspective of a rigorous evaluation of the data collected to explore potentials as well as limitations (pitfalls). Conceptually, crowdsourcing can be distinguished into being either *active* or *passive*. Active crowdsourcing involves a software platform and the active elicitation of input from the crowd. Active crowdsourcing can occur either 'in situ' via mobile devices (e.g., Citizens as Sensors (Goodchild, 2007)) or address any kind of (geographic) topic that can be communicated electronically via a computer such as a mapping or data collection project (such as OpenStreetMap[1] or Ushahidi[2]) or any kind of behavioral experiment that can be deployed electronically. There are numerous applications for active crowdsourcing; we will discuss Amazon's Mechanical Turk[3] as the most prominent yet recently controversial platform. Passive crowdsourcing is targeting information that has been made publicly available but not as a response to a particular request or to a request different from the research question at hand. In other words, the information collected is unsolicited. Web sites, Twitter feeds, or Facebook entries are examples of such information. Advances, especially in natural language processing (Woodward et al., 2010, Socher et al., 2013) and georeferencing (Hu and Ge, 2007, Gelernter and Balaji, 2013) are enabling access to an immense reservoir of data, information, and knowledge that potentially is related to specific aspects of geographical space.

Our research has addressed a variety of contextual effects on the validity of crowdsourced data such as cultural, linguistic, regional, as well as methodological differences that we will discuss here in light of semantics.

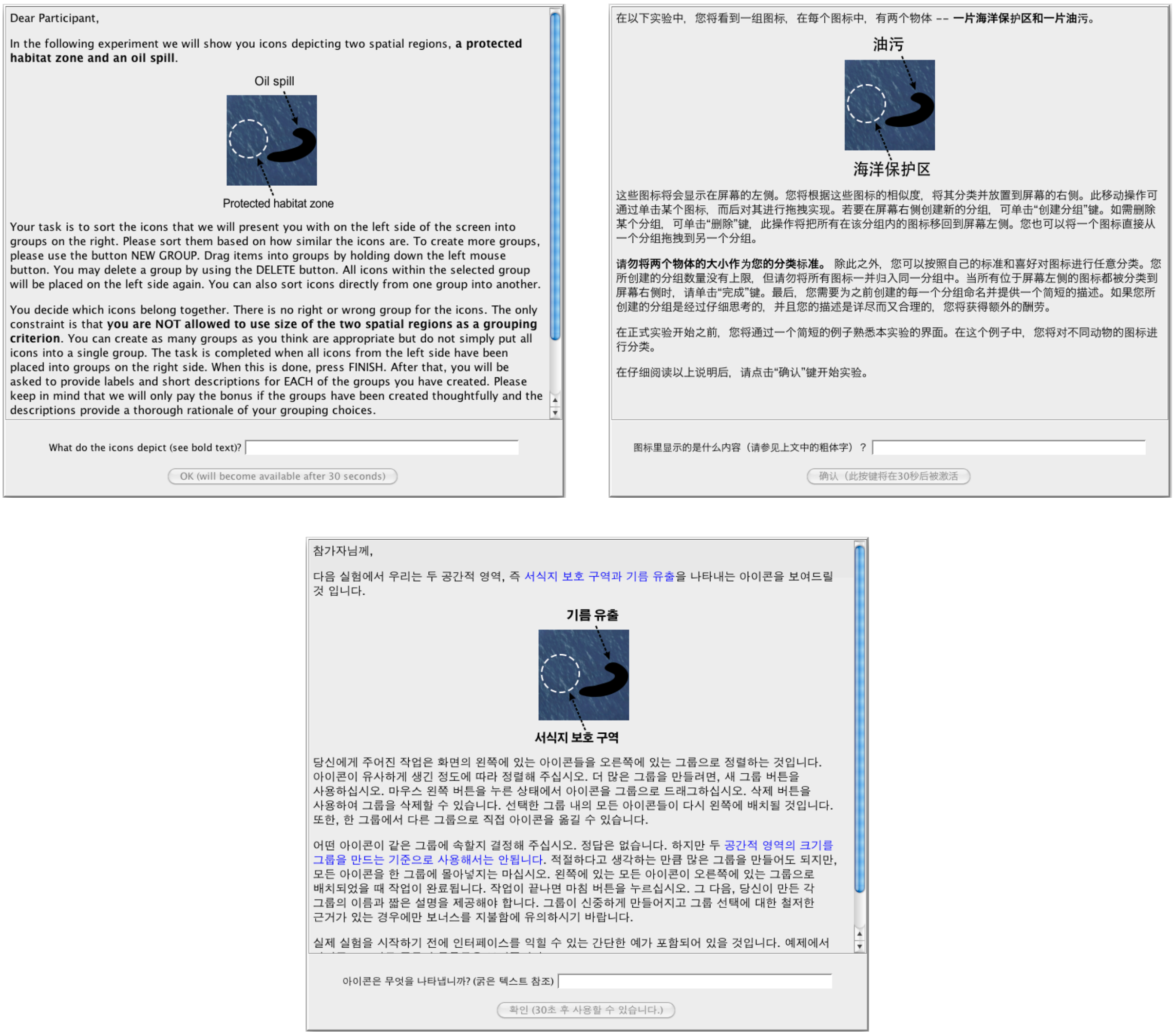
2. CULTURAL, LINGUISTIC, AND REGIONAL CONTEXTS

Sourcing from the crowd opens up the possibility for using diverse cultural and linguistic backgrounds to make contributions to the question of, for example, *linguistic relativity* (Gumperz and Levinson, 1996, Boroditsky, 2000). Linguistic relativity, as a major research area in the cognitive sciences, is addressing the influence of language on cognitive processing. The theory is that someone's native language has substantial influences on the way information is processed. While usually performed in rather expensive field or lab studies, being able to target groups in the crowd that share characteristics such as speaking the same languages or speaking the same language but in different environmental contexts, is an exciting possibility. Though running experiments via the Internet has great appeal, the downside is that certain crowd science platforms may not be globally available, such as Amazon Mechanical Turk, or that the infrastructure in a country is not sufficiently developed. Additionally, it may be more difficult to control unwanted influencing parameters and prevent participants from cheating, for example, using translation services to pretend mastery of a foreign language.

We have run several experiments in this context, using active crowdsourcing (eliciting responses from the crowd after forming a research question), using passive crowdsourcing (re-using information the crowd made publicly available for potentially a different purpose), and using a mixed approach in which we compared results from a field study against a crowd sourced experiment; results were largely positive. In (Klippel et al., 2013), we demonstrated the feasibility of using crowdsourcing via the Internet platform Amazon Mechanical Turk (AMT) as a means to address questions of linguistic relativity by comparing responses of English, Chinese, and Korean speaking participants (cmp. Figure 1). The question addressed was: How many spatial relations between overlapping extended spatial entities people intuitively distinguish (see Figure 5 (left) for examples of the stimulus material used in this study). While participants share a current cultural and linguistic environment (their computers were located in the US), their mother tongue of English, Chinese, and Korean was different. The research approach mimicked studies in the psychological sciences trying to replace expensive overseas field

*Corresponding author
[1] https://www.openstreetmap.org/
[2] www.ushahidi.com/
[3] https://www.mturk.com/

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
Editors: A.-M. Olteanu-Raimond, C. de-Runz, and R. Devillers
doi:10.5194/isprsannals-II-3-W5-325-2015



Dear Participant,

In the following experiment we will show you icons depicting two spatial regions, **a protected habitat zone and an oil spill**.

Oil spill

Protected habitat zone

Your task is to sort the icons that we will present you with on the left side of the screen into groups on the right. Please sort them based on how similar the icons are. To create more groups, please use the button NEW GROUP. Drag items into groups by holding down the left mouse button. You may delete a group by using the DELETE button. All icons within the selected group will be placed on the left side again. You can also sort icons directly from one group into another.

You decide which icons belong together. There is no right or wrong group for the icons. The only constraint is that **you are NOT allowed to use size of the two spatial regions as a grouping criterion**. You can create as many groups as you think are appropriate but do not simply put all icons into a single group. The task is completed when all icons from the left side have been placed into groups on the right side. When this is done, press FINISH. After that, you will be asked to provide labels and short descriptions for EACH of the groups you have created. Please keep in mind that we will only pay the bonus if the groups have been created thoughtfully and the descriptions provide a thorough rationale of your grouping choices.

What do the icons depict (see bold text)?

OK (will become available after 30 seconds)

在以下实验中，您将看到一组图标，在每个图标中，有两个物体 -- **一片海洋保护区和一片油污。**

油污

海洋保护区

这些图标将会显示在屏幕的左侧。您将根据这些图标的相似度，将其分类并放置到屏幕的右侧。此移动操作可通过单击某个图标，而后对其进行拖拽实现。若要在屏幕右侧创建新的分组，可单击“创建分组”键。如需删除某个分组，可单击“删除”键，此操作将把所有在该分组内的图标移回到屏幕左侧。您也可以将一个图标直接从一个分组拖拽到另一个分组。

请勿将两个物体的大小作为您的分类标准。 除此之外，您可以按照自己的标准和喜好对图标进行任意分类。您所创建的分组数量没有上限，但请勿将所有图标一并归入同一分组中。当所有位于屏幕左侧的图标都被分类到屏幕右侧时，请单击“完成”键。最后，您需要为之前创建的每一个分组命名并提供一个简短的描述。如果您所创建的分组是经过仔细思考的，并且您的描述是详尽而又合理的，您将获得额外的酬劳。

在正式实验开始之前，您将通过一个简短的例子熟悉本实验的界面。在这个例子中，您将对不同动物的图标进行分类。

在仔细阅读以上说明后，请点击“确认”键开始实验。

图标里显示的是什么内容（请参见上文中的粗体字）？

确认（此按键将在30秒后被激活

참가자님께,

다음 실험에서 우리는 두 공간적 영역, 즉 서식지 보호 구역과 기름 유출을 나타내는 아이콘을 보여드릴 것 입니다.

기름 유출

서식지 보호 구역

당신에게 주어진 작업은 화면의 왼쪽에 있는 아이콘들을 오른쪽에 있는 그룹으로 정렬하는 것입니다. 아이콘이 유사하게 생긴 정도에 따라 정렬해 주십시오. 더 많은 그룹을 만들려면, 새 그룹 버튼을 사용하십시오. 마우스 왼쪽 버튼을 누른 상태에서 아이콘을 그룹으로 드래그하십시오. 삭제 버튼을 사용하여 그룹을 삭제할 수 있습니다. 선택한 그룹 내의 모든 아이콘들이 다시 왼쪽에 배치될 것입니다. 또한, 한 그룹에서 다른 그룹으로 직접 아이콘을 옮길 수 있습니다.

어떤 아이콘이 같은 그룹에 속할지 결정해 주십시오. 정답은 없습니다. 하지만 두 공간적 영역의 크기를 그룹을 만드는 기준으로 사용해서는 안됩니다. 적절하다고 생각하는 만큼 많은 그룹을 만들어도 되지만, 모든 아이콘을 한 그룹에 몰아넣지는 마십시오. 왼쪽에 있는 모든 아이콘이 오른쪽에 있는 그룹으로 배치되었을 때 작업이 완료됩니다. 작업이 끝나면 마침 버튼을 누르십시오. 그 다음, 당신이 만든 각 그룹의 이름과 짧은 설명을 제공해야 합니다. 그룹이 신중하게 만들어지고 그룹 선택에 대한 철저한 근거가 있는 경우에만 보너스를 지불함에 유의하시기 바랍니다.

실제 실험을 시작하기 전에 인터페이스를 익힐 수 있는 간단한 예가 포함되어 있을 것입니다. 예제에서

아이콘은 무엇을 나타냅니까? (굵은 텍스트 참조)

확인 (30초 후 사용할 수 있습니다.)

Figure 1: Screenshot of the three start screens of CatScan in three different languages (English, Chinese, Korean)

studies (Papafragou and Selimis, 2010). We were able to demonstrate a) that it is possible to elicit feedback from diverse linguistic background through platforms such as AMT; b) that results are reliable; and c) that in this particular experiment the main distinction between non-overlapping, overlapping, and proper-part relations outweighs potential language specific differences. We were able to confirm the validity of individual responses by having native language speakers on the research team. However, collecting data from native Korean speakers proved challenging due to smaller numbers of AMT workers which forced us to lower AMT approval rates. This led to an invitation for cheaters: some participants used translation services to read the instructions and provide answers to the questions. While we were able to identify these participants, the lesson learned is that high AMT approval rates are essential for ensuring the validity of research results obtained through AMT.[4]

Another study (Xu et al., 2014) used a framework for passive crowdsourcing (Jaiswal et al., 2011) and collected a corpus of >11,000 instances of route directions from web pages covering the entire USA at the granularity of states. Through various tools that aided data processing, we were able to show regional differences in the way that people give route directions, that is, whether they have a preference for cardinal or relative directions (see Figure 2). The data validation in this case is challenging as there is no way to access the ’participants’ or to learn anything about their backgrounds (e.g., are the people who put route directions on web pages people who lived in an area for a long time or grew up there such that they absorbed enough regional specificities?). One aspect in favor of this crowdsourcing approach is the large number of participants, that is, the size of the corpus. In this sense this study fulfills a classic promise of crowd science, that is, that large numbers (classic interpretation of the crowd) exhibit intelligence (Surowiecki, 2005). This is a valid point here as it is fair to assume that the majority of people who put route directions on the web need to have had some substantial exposure to their environments, especially at the level of states and regions (it was not the point of this study to prove that people in Connecticut differ from people in New Hampshire in the way they give route directions). In addition, we were able to confirm the results by theoretical considerations such as different environments (mountains versus plains) as well as historical planning strategies. While one point of big data is that it may mean the end of theory (Anderson, 2008), we believe that we are not there yet in terms of data quality and reliability, especially with respect to behavioral studies using

[4] There is a lively discussion about the validity of AMT for research results that cannot be discussed in detail here. For an overview see, for example, (Crump et al., 2013). Amazon also recently announced changes to their financial model which potentially will lower the attractiveness of AMT for academic research.

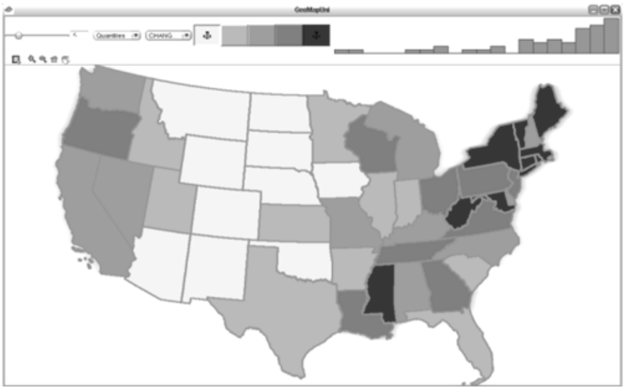


Figure 2: Proportion of relative direction vs. cardinal direction usage for expressing change of direction in the U.S. (Dark: more relative direction usage; light: more cardinal direction usage)

the crowd.

We were able to show that the patterns that emerged in our study (see Figure 2) were in line with many regional characteristics and planning aspect of cities across the US. The bottom line is that without improved methods of accessing background information of the crowd many behavioral studies benefit from theoretical grounding of their findings as well as large numbers.

Crowdsourcing can also be used to complement field studies. In a recent study (Klippel et al., 2015), we addressed emerging topics in the area of landscape conceptualization and explicitly used a diversity fostering approach to uncover potentials, challenges, complexities, and patterns in human landscape concepts. Based on a representation of different landscapes (see Figure 5 (right) for examples of the images used as stimulus material), responses from two different populations were elicited: Navajo and the (US) crowd. Data from Navajo participants was obtained through field studies while data from English-speaking participants was collected via AMT. Results support the idea of conceptual pluralism, that is, even within a linguistically homogeneous group of participants different conceptualizations of reality (geographic landscapes) exist (see also Section 4.).

3. EXPERTS VERSUS LAY PEOPLE VERSUS DIFFERENT INPUT DATA SOURCES

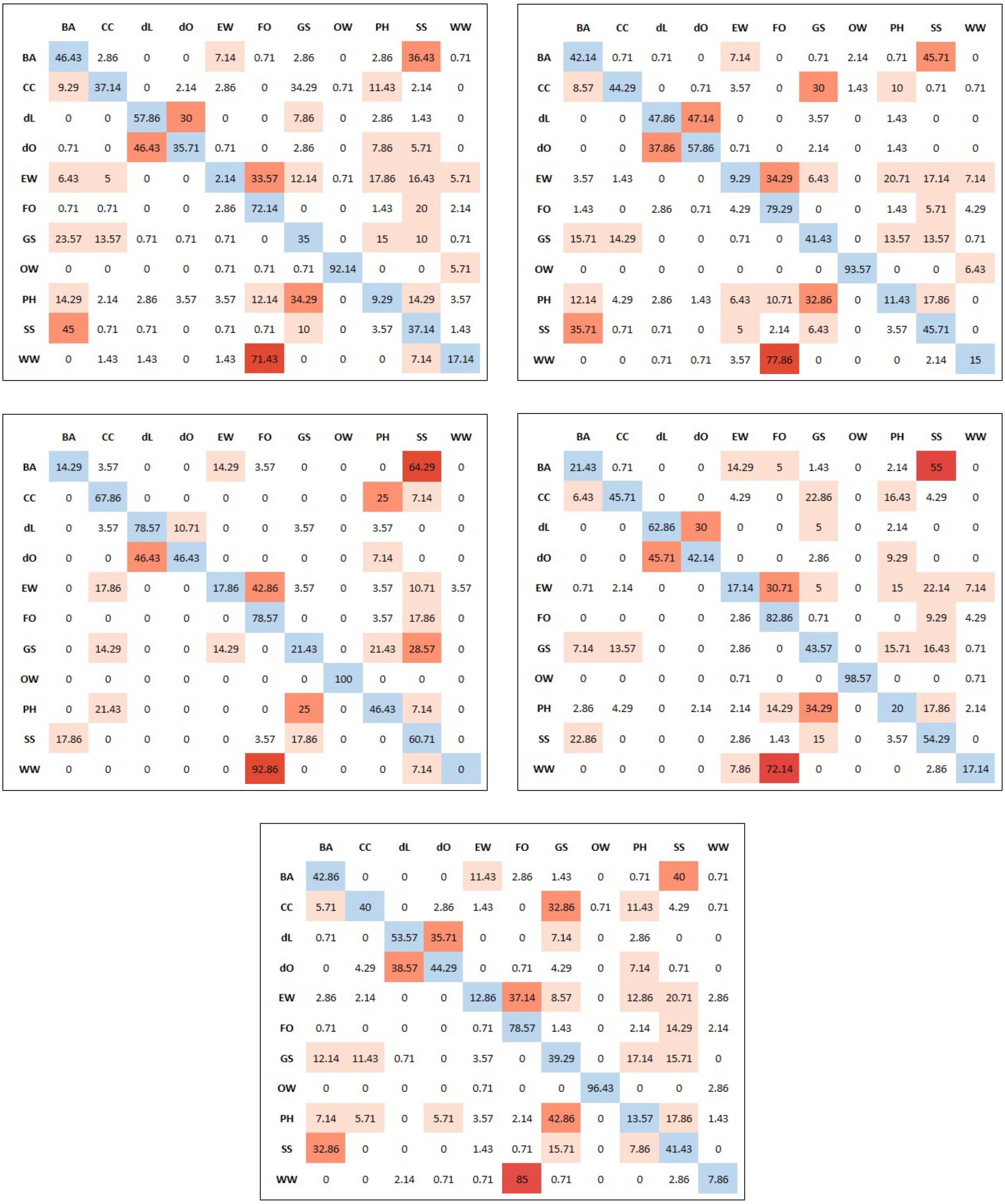
One of the potentially most exciting developments in crowd science is the possibility of extending earth observations beyond artificial sensors and use the crowd to aid in unprecedented extensive data collection (Salk et al., 2015, Comber et al., 2013, Goodchild and Glennon, 2010). There are excellent reasons to use the crowd as human sensors: In certain situations, the crowd outperforms artificial sensors. One of the best examples are birding applications in which volunteers contribute tremendous and reliable insight into the distribution and migration patterns of birds[5]. This data would be impossible to collect through current sensor networks. In other areas such as land cover data, human sensors complement artificial sensors to, potentially, increase the availability of ad hoc data (Heinzelman and Waters, 2010) or improve artificial sensors (Comber et al., 2013). The Geo-Wiki Project (Fritz et al., 2009) provides aerial photos of the earth's surface to online participants and asks them to classify these patches of land into various land cover classes. While there are studies that explore the accuracy and reliability of this Geo-Wiki data (Foody and Boyd, 2013, Perger et al., 2012, See et al., 2013), there is a need for further understanding citizens' perception and their classification process of the environment. The Citizen Observatory Web[6], for example, aims to have citizens create environmental data through mobile devices in and around the area where the citizens live. By working with them throughout this process, one of their goals is to better understand the citizens' environmental perception and learn how citizens go about the data creation process. Although the community is making progress, we are far from understanding humans' abilities to sense environmental information reliably.

While a lot of excitement has been spread through projects such as Geo-Wiki, a comprehensive set of studies we performed on humans' abilities to reliably identify land cover types shows that the claimed high accuracy of human land cover classifications in other studies is only possible at a coarse level of granularity or for specific land cover types. Figure 3 shows the results of five experiments we conducted (Sparks et al., 2015a, Sparks et al., 2015b), which tested the effect of participant expertise, methodological design, and the influence of different input data sources and perspectives (i.e., ground-based photos and aerial photos) when classifying land cover, in the form of confusion matrices. Correctly classified images are along the diagonal (top-left to bottom-right). All experiments asked participants to classify photos of land covers into one of 11 possible categories. The two methodological designs varied the size of photos, and the visual availability of those photos. The first methodological design presented the participant with a series of ground-based photos all at once, side by side, as relatively small icons. This allowed the participant to see all the images at all times throughout the classification process. The second, presented the participant with a ground-based (and aerial) photo one at a time, and thus were larger images than shown in the previous methodological design. Thus, the participant could not simultaneously view all the images in the second methodological design. These categorical classification tasks have proven to be difficult for participants. The experiments demonstrated that a) experts are not significantly different from educated lay participants (i.e., participants given definitions and prototypical images of the land cover classes before the experiment) when classifying land cover, b) methodological changes in classification tasks did not significantly affect participants' classification, and c) the addition of aerial photos (plus ground images) did not significantly change participants' classification.

The earth's surface can be complex and heterogeneous so asking crowdsourced participants to take this complexity and classify it into relatively low-level categories is perhaps not the most effective method, that is, the level of granularity at which humans are able to classify land cover might be rather coarse. This is especially the case when understanding these low-level categories rely so much on participants' interpretation of class names. This interpretation perhaps has the largest influence on classification outcome as we see variation in expertise, methodological design, and different input data sources has little influence on classification outcome. Some land cover classes are more challenging to interpret than others, with participants classifying land cover classes like Forest, Developed, and Open Water more consistently. Conversely, participants classified more challenging classes like Grassland and Pasture less consistently. As previously mentioned, this pattern persisted in light of participant expertise differences, and varying input data sources/perspectives (ground- based photos versus aerial photos).

[5] see http://www.birds.cornell.edu/

[6] https://cobwebproject.eu



	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

	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

	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

	BA	CC	dL	dO	EW	FO	GS	OW	PH	SS	WW
BA	21.43	0.71	0	0	14.29	5	1.43	0	2.14	55	0
CC	6.43	45.71	0	0	4.29	0	22.86	0	16.43	4.29	0
dL	0	0	62.86	30	0	0	5	0	2.14	0	0
dO	0	0	45.71	42.14	0	0	2.86	0	9.29	0	0
EW	0.71	2.14	0	0	17.14	30.71	5	0	15	22.14	7.14
FO	0	0	0	0	2.86	82.86	0.71	0	0	9.29	4.29
GS	7.14	13.57	0	0	2.86	0	43.57	0	15.71	16.43	0.71
OW	0	0	0	0	0.71	0	0	98.57	0	0	0.71
PH	2.86	4.29	0	2.14	2.14	14.29	34.29	0	20	17.86	2.14
SS	22.86	0	0	0	2.86	1.43	15	0	3.57	54.29	0
WW	0	0	0	0	7.86	72.14	0	0	0	2.86	17.14

	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

Figure 3: Comparison of patterns of responses of five experiments. Row/column names of each matrix represent unique land cover classes the participants could choose from (Barren, Cultivated Crops, Developed Low Intensity, Developed High Intensity, Emergent Herbaceous Wetlands, Forest, Grassland, Open Water, Pasture/Hay, Shrub/Scrub, Woody Wetlands). The first three matrices (left to right) represent the first three experiments, testing the influence of expertise in classification. The last two represent the last two experiments, testing the influence of added aerial photos. Results show agreement against NLCD data, more precisely the numbers represent how often images of the class given by the row have been categorized as the class given by the column (i.e., a confusion matrix). Darker (red) colors indicate higher error rates. More important is the comparison of similarity between patterns.

4. THE COMPLEXITY OF THE HUMAN MIND—COGNITIVE SEMANTICS

The final aspect to discuss in this short paper are competing conceptualizations humans may have of the same set of stimuli (Foucault, 1994, Wrisley III, George Alfred, 2008, Barsalou, 1983, Gärdenfors, 2000). We have made substantial progress in analyz-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
Editors: A.-M. Olteanu-Raimond, C. de-Runz, and R. Devillers
doi:10.5194/isprsannals-II-3-W5-325-2015

ing crowdsourced data in depth and provide a statistical measure on the agreement of participants with respect to the task they perform (in most of our experiments participants create categories for stimuli they are presented with such as landscape images). While this is a rather specific task, it does reveal some important aspects about the human mind (cognitive semantics) that sound straight forward but are difficult to quantify: the more complex the stimulus/task is, the more varied are participants responses. This is particularly true for unrestricted sampling from the crowd. To quantify this relation, we developed, for example, a cross-method-similarity-index (CMSI, see (Wallgrün et al., 2014)). The CMSI measures agreement between the results of applying different hierarchical clustering methods (cf. (Kos and Psenicka, 2000)) to the data collected in category construction experiments for a given number of clusters (c). The value is computed for different values of c. Analyses from two experiments are provided in Figure 4 with examples of the icons used in the respective experiment shown in Figure 5. Without going into too much detail: Consistency of human conceptualizations (cognitive semantics) is established in a bootstrapping approach by sampling from a participant pool (actual responses) with increasing sample sizes. The average CMSI values are then plotted over the sample size. The top part of Figure 4 shows results for the above mentioned experiment on overlap relations. It is clear that even a small number of participants converge at the most reliable solution, that is, a separation into three categories (non-overlapping, overlapping, and proper part relations). This is indicated by the line for three clusters in the graph approaching 1 (ideal solution) quickly and for low numbers of participants. In contrast, data from a recent experiment on landscape concepts (Klippel et al., 2015) shows that there is no universally acceptable category structure that, on an abstract level, would work for all participants, that is, no number of clusters converges to 1.

This finding, partially in combination with results discussed above, has resulted in three lines of current research:

- The quantification of how complex individual stimuli are.
- The statistical identification of conceptually consistent subgroups of participants.
- The definition of conceptual pluralism as a means to statistically determine the complete set of intuitive conceptualizations the crowd may have on the stimulus used.

5. CONCLUSIONS

We focused on a meta-discussion of the lessons learned so far on different aspects of semantics on crowd science. Crowd science is still a young discipline and as such requires discussions about pitfalls and potentials. We argue that the semantic diversity of the crowd is an opportunity rather than a downside. It does require, however, attention to detail to harvest the full potential of this diversity. First, there needs to be some quality control either in form of reliability scores (AMT), hands-on validation, or a thorough theoretical underpinning against which the results can be evaluated. Additionally, we need statistical methods that allow for identifying relevant semantic contexts, that is, we need new methods that intelligently process data collected from the crowd and identify consistent views/performance by sub-groups.

When the crowd is used to assist in earth observation, it is important to make the crowd's task as objective as possible. As seen in the land cover classification experiments described above, when subjective interpretation of terms is allowed, the consistency and reliability of responses drop and the variety of unique responses increases. Having a relatively high number of classes to classify from, and those classes being relatively broad in their interpretation allows for much more subjectivity than objectivity. To address this problem, we are currently designing experiments that replace a categorical land cover classification scheme with a feature-based classification scheme. This feature-based scheme mimics a decision tree process, continually asking the participant a series of 'either-or' questions (e.g. Is this photo either primarily vegetated or primarily non-vegetated?). Our hypothesis is that participants are more likely to agree on the presence or absence of environmental features compared to agreeing on lower-level categorical classifications. This scheme reduces the variety of class name interpretation in the classification task and creates a more objective approach.

REFERENCES

Anderson, C., 2008. The end of theory: The data deluge makes the scientific method obsolete. `http://archive.wired.com/science/discoveries/magazine/16-07/pb\textunderscoretheory`. Accessed: 01/01/2015.

Barsalou, L. W., 1983. Ad hoc categories. Memory and Cognition 11, pp. 211–227.

Boroditsky, L., 2000. Metaphoric structuring: understanding time through spatial metaphors. Cognition 75(1), pp. 1–28.

Clery, D., 2011. Galaxy evolution. Galaxy zoo volunteers share pain and glory of research. Science (New York, N.Y.) 333(6039), pp. 173–175.

Comber, A., Brundsdon, C., See, L., Fritz, S. and McCallum, I., 2013. Comparing expert and non-expert conceptualisations of the land: An analysis of crowdsourced land cover data. In: T. Tenbrink, J. Stell, A. Galton and Z. Wood (eds), Spatial Information Theory, Springer, Berlin, pp. 243–260.

Crump, M. J. C., McDonnell, J. V., Gureckis, T. M. and Gilbert, S., 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PLoS ONE 8(3), pp. e57410.

Foody, G. M. and Boyd, D. S., 2013. Using volunteered data in land cover map validation: Mapping west african forests. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6(3), pp. 1305–1312.

Foucault, M., 1994. The order of things: An archaeology of the human sciences. Vintage books ed edn, Vintage Books, New York.

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), pp. 345–354.

Gärdenfors, P., 2000. Conceptual Spaces. The Geometry of Thought. The MIT Press, Cambridge and MA.

Gelernter, J. and Balaji, S., 2013. An algorithm for local geoparsing of microtext. GeoInformatica 17(4), pp. 635–667.

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

Goodchild, M. F. and Glennon, J. A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth 3(3), pp. 231–241.

Gumperz, J. J. and Levinson, S. C. (eds), 1996. Rethinking linguistic relativity. Cambridge University Press, Cambridge, UK.

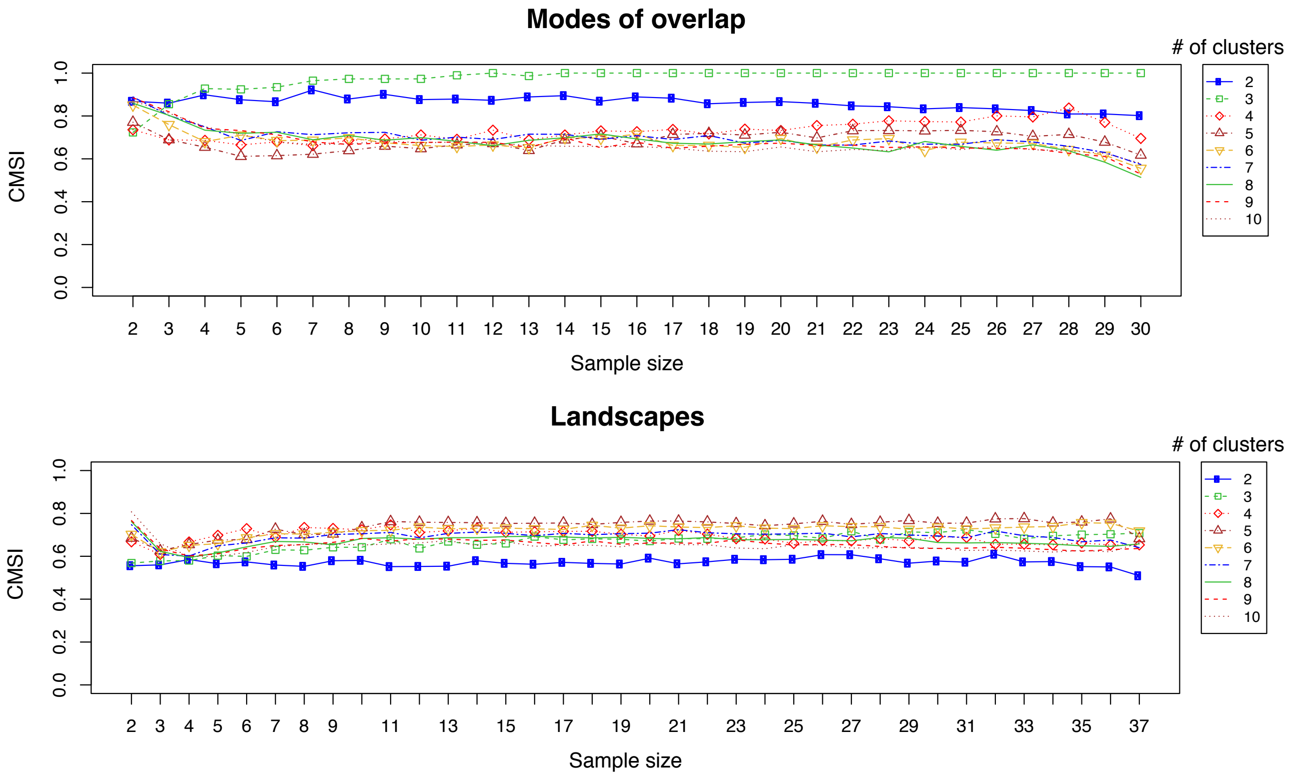


Figure 4: Results of cluster validation using CMSI for two experiments. Top: experiment on overlap relations (see Figure 5 (left)); bottom: experiment on landscape conceptualizations (see Figure 5 (right)).

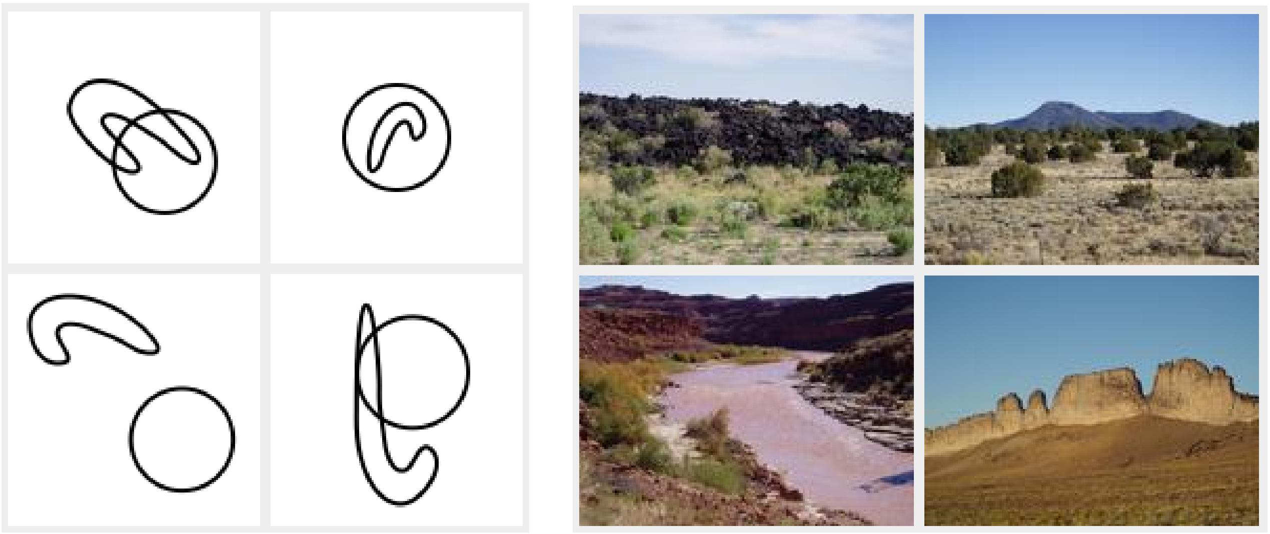


Figure 5: Example of the icons used in the Mode of Overlap experiment (left) and the Navajo Landscape Concepts (right).

Heinzelman, J. and Waters, C., 2010. Crowdsourcing crisis information in disaster-affected Haiti. Special Report 252, United States Institute of Peac, Washington, DC.

Hu, Y. and Ge, L., 2007. A supervised machine learning approach to toponym disambiguation. In: The Geospatial Web – How geobrowsers, social software and the Web 2.0 are shaping the network society, Springer, pp. 117–128.

Jaiswal, A., Zhang, X., Mitra, P., Pezanowski, S., Turton, I., Xu, S., Klippel, A. and MacEachren, A. M., 2011. GeoCAM: A Geovisual Analytics Workspace to Contextualize and Interpret Statements about Moveme. Journal of Spatial Information Science (3), pp. 33–69.

Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., Jaskolski, M. and Baker, D., 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nature structural & molecular biology 18(10), pp. 1175–1177.

Klippel, A., Mark, D. M., Wallgrün, J. O. and Stea, D., 2015. 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 and S. Bell (eds), Proceedings, Conference on Spatial Information Theory (COSIT 2015), Santa Fe, NM, USA, Oct.

12-16, 2015, Springer, Berlin.

Klippel, A., Wallgrün, J. O., Yang, J., Mason, J. and Kim, E.-K., 2013. Fundamental cognitive concepts of space and time: Using cross-linguistic, crowdsourced data to cognitively calibrate modes of overlap. In: T. Tenbrink, J. Stell, A. Galton and Z. Wood (eds), Spatial Information Theory, Springer, Berlin, pp. 377–396.

Kos, A. J. and Psenicka, C., 2000. Measuring cluster similarity across methods. Psychological Reports 86, pp. 858–862.

Papafragou, A. and Selimis, S., 2010. Event categorisation and language: A cross-linguistic study of motion. Language and Cognitive Processes 25(2), pp. 224–260.

Perger, C., Fritz, S., See, L., Schill, C., van der Velde, Marijn, McCallum, I. and Obersteiner, M., 2012. A campaign to collect volunteered geographic information on land cover and human impact. In: T. Jekel, A. Car, J. Strobl and G. Griesebner (eds), GI_Forum 2012: Geovizualisation, Society and Learning, Berlin: Wichmann, pp. 83–91.

Salk, C. F., Sturn, T., See, L., Fritz, S. and Perger, C., 2015. Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game. International Journal of Digital Earth pp. 1–17. published online, http://dx.doi.org/10.1080/17538947.2015.1039609, 02 Jun 2015.

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

SG 2007, n.d.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y. and Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Vol. 1631, Citeseer, p. 1642.

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

Sparks, K., Klippel, A., Wallgrün, J. O. and Mark, D. M., 2015b. Crowdsourcing landscape perceptions to validate land cover classifications. In: O. Ahlqvist, K. Janowicz, D. Varanka and S. Fritz (eds), Land Use and Land Cover Semantics, CRC Press, pp. 296–314.

Surowiecki, J., 2005. The wisdom of crowds. 1st anchor books ed edn, Anchor Books, New York, USA.

Wallgrün, J. O., Klippel, A. and Mark, D. M., 2014. A new approach to cluster validation in human studies on (geo)spatial concepts. In: K. Stewart, E. Pebesma, G. Navratil, P. Fogliaroni and M. Duckham (eds), Extended Abstract Proceedings of the GIScience 2014, Hochschülerschaft, TU Vienna, Vienna, pp. 450–453.

Woodward, D., Witmer, J. and Kalita, J., 2010. A comparison of approaches for geospatial entity extraction from Wikipedia. In: IEEE Fourth International Conference on Semantic Computing (ICSC), IEEE Computer Society, Washington, DC, USA, pp. 402–407.

Wrisley III, George Alfred, 2008. Realism and Conceptual Relativity. PhD thesis, The University of Iowa, USA.

Xu, S., Klippel, A., MacEachren, A. M. and Mitra, P., 2014. Exploring regional variation in spatial language using spatially-stratified web-sampled route direction documents. Spatial Cognition and Computation 14(4), pp. 255–283.