Towards evaluating crowdsourced image classification on mobile devices to generate geographic information about human settlements

Benjamin Herfort
GIScience Research Group, Heidelberg University
herfort@uni-heidelberg.de

Marcel Reinmuth
GIScience Research Group, Heidelberg University
reinmuth@stud.uni-heidelberg.de

João Porto de Albuquerque
Centre for Interdisciplinary Methodologies, University of Warwick
J.Porto@warwick.ac.uk

Alexander Zipf
GIScience Research Group, Heidelberg University
zipf@uni-heidelberg.de

Abstract

Geographic information crowdsourcing is an increasingly popular approach to derive geographic data about human settlements from remotely sensed imagery. However, crowdsourcing approaches are frequently associated with uncertainty about the quality of the information produced. Although previous studies have found acceptable quality of crowdsourced information in some application domains, there is still lack of research about the quality of information produced with mobile crowdsourcing tools. This paper aims to contribute towards filling this gap by presenting an initial analysis of the contributions of two crowdsourcing projects based on the MapSwipe mobile app, in Madagascar and South Sudan. Our results show, that there is substantial agreement amongst volunteers thus suggesting that mobile crowdsourcing is a viable approach to support the mapping of human settlements. Nevertheless, this study also identifies several factors that may cause disagreement between volunteers (e.g. bad imagery, dependence on individual users) and thus reduce the reliability of the information they produce.

Keywords: crowdsourcing, mapping, agreement, intrinsic, quality

1 Introduction

In the past couple of years, crowdsourcing approaches have been increasingly used to produce geographic information from remotely sensed imagery to complement official information regarding human settlements (Albuquerque, Herfort, & Eckle, 2016). Previous research has shown, that crowdsourced classification can be feasible to produce high quality geographic data for several use cases like land use monitoring or disaster response (Albuquerque et al., 2016; Fritz et al., 2009; Imran, Castillo, Meier, & Diaz, 2013; Schepaschenko et al., 2015). Nevertheless, some authors also highlight difficulties in specific areas, e.g. for damage assessment (Kerle & Hoffman, 2013; Westrope, Banick, & Levine, 2014). The quality of information produced by volunteers may vary not only across different application domains, but also in relation to the several types of crowdsourcing tasks that volunteers can undertake (Albuquerque et al., 2016).

The Missing Maps project is one example that shows how a community of volunteers and humanitarian organizations can work together and produce crowdsourced geographic information. In this context, the MapSwipe App was recently introduced to enable volunteers to classify satellite imagery from their mobile phones. The biggest advantage of this new generation of crowdsourcing tools for mobile devices is to lower the entry barrier for volunteers, who do not need special mapping skills and can use their smartphones to contribute whenever they may have some time to spare (e.g. during commuting). However, this also raises concerns about the quality of the data produced by volunteers in these variable conditions: are they as reliable as it was proven to be the case on other crowdsourcing applications?

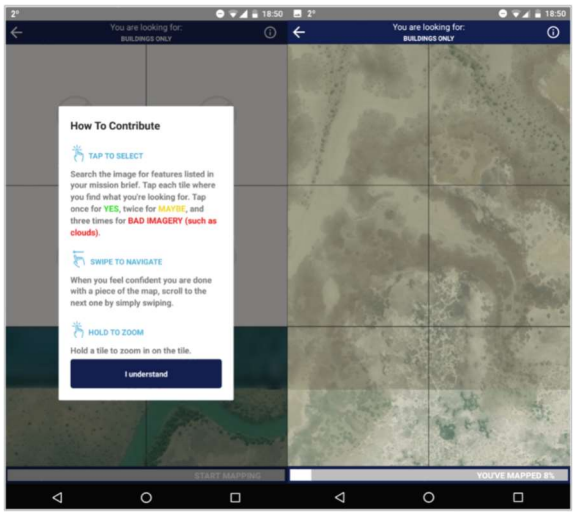
To answer this question, we present in this paper an initial analysis of the quality of the data produced by volunteers using the MapSwipe app. Since reference data is often not available to compare with crowdsourced results, we apply an intrinsic approach based on the agreement between volunteers and on indicators of inter-rater reliability. Furthermore, we investigate factors that may have an influence on the agreement between volunteers, and thus negatively impact the quality of the produced information.

The remainder of this paper is organized as follows. In Section 2 we describe our case study and give detailed information on the crowdsourced data collection using the MapSwipe App. Section 3 presents our methodology. In Section 4 we present the results of our analysis. Finally, the results are discussed and final remarks are presented to conclude this paper in section 5.

2 Case Study

In November 2014, the American Red Cross, the British Red Cross , the Humanitarian OpenStreetMap Team (HOT) and Doctors Without Borders/Médecins sans Frontiers (MSF) established the Missing Maps project. This project is aimed at

Figure 1: The MapSwipe App



"putting the most vulnerable people on the map"[1]. Within this project, MSF developed the MapSwipe App, a mobile crowdsourcing application to derive geographic information from satellite imagery. In MapSwipe, volunteers are asked to classify tiles of satellite imagery into four different classes ("No", "Yes", "Maybe", "Bad Imagery"). By tapping and swiping on the mobile device they can signal whether they were able to spot human settlements. The users can choose between several projects, e.g. in different countries. For each project a short introduction and guidance are given. Figure 1 shows two example screenshots from the MapSwipe project "Madagascar 9".

The image on the left-hand side shows how the users can classify the imagery by tapping. If the volunteer doesn't interact at all, this will be regarded as a "No" classification. On the right hand side the mapping interface is shown. On top of the interface volunteers are instructed what features they are looking for (for this project they search for buildings). The main part of the screen is divided into 6 squares, each square representing a single classification task with a width and height of approximately 150 meters.

Another important concept within the MapSwipe app are "groups". When volunteers are using the app, they will always work on a group of tasks. Each group has a height of three tiles and a width of approximately 60-70 tiles depending on the shape of area of interest. Overall, each group is represented as an elongated east-west band containing about 200 tiles. At the bottom of the screen, a progress bar indicates how many tiles have already been classified within the specific group. The results will be uploaded not until the user completed the group.

In our case study, we use MapSwipe data from two projects. The project "Madagascar 1" was created on 30th July 2016 and finished by 2nd November 2016 and covers the northernmost part of the state. The project was managed by the French NGO CartONG[2]. The project "South Sudan" was created on 23rd November 2016 and finished by 19th January 2017. It was managed by MSF Czech Republic. The data produced by the volunteers is released under the CC-BY-4.0 license and was obtained from the MapSwipe-API.[3] Through the MapSwipe-API we downloaded information on every single classification (result, user name, timestamp) and the completed count for each group.

Table 1 provides an overview of both projects. The table shows that both projects cover approximately the same area (circa 6000 km^2) and a comparable number of users (~850) contributed to each project. While the project "Madagascar 1" is in a rural area, the project "South Sudan" covers a more densely settled area. In the projects used in this case study each group was assessed by at least 3 individuals. For both projects the median of evaluations per task is 4. The variance for the project "Madagascar 1" is 224.0, the corresponding value for the project "South Sudan 1" is 3.3.

Table 1: MapSwipe projects information.

Name	Madagascar 1	South Sudan 1
Area (km^2)	6270.5	5800.3
Groups	1,388	1,286
Tasks	278,688	257,789
Contributions	989,193	874,418
Users	867	837

3 Methodology

In this section the overall methodology will be presented. In the first step, we compute the agreement between different volunteers on the individual task level. For each task, we assign the total number of classifications and the number of the distinct count for each individual class ("No", "Yes", "Maybe", "Bad Imagery"). The agreement level among volunteers is calculated using Equation (1), as the proportion of agreeing pairs of classifications out of all the possible pairs of

[1] http://wiki.openstreetmap.org/wiki/Missing_Maps_Project 26.01.2017

[2] http://www.cartong.org/ 26.01.2017

[3] http://mapswipe.org/index.html 26.01.2017

assignments, following (Fleiss, 1971). Accordingly, n is the number of ratings per subject (i.e., a tile in our case), k is the number of categories into which assignments are made (four in our case), and n_{ij} is the number of raters which assigned the i-th subject to the j-th category. This method allows us to compare tasks for which we obtained a different number of classifications.

$$P_i = \frac{1}{n*(n-1)} * \sum_{j=1}^{k} n_{ij}^2 - n_{ij} \qquad (1)$$

In the next step, we compute the inter-rater reliability for tasks within the corresponding groups of the MapSwipe project. The inter-rater reliability is a statistical measure to analyze the agreement between multiple raters by comparing to the agreement between these raters that they could obtain by chance. Fleiss kappa is calculated according to Fleiss (1971) as presented in Equation (2). In this equation $\bar{P}$ corresponds to the mean of the P_i's and $\overline{P_e}$ to the sum of the squared proportions of all assignments which were to each individual class. The higher the kappa value (max. 1) the stronger the agreement between the raters.

$$\kappa = \frac{\bar{P} - \overline{P_e}}{1 - \overline{P_e}} \qquad (2)$$

Furthermore, we analyze the spatial distribution of agreement per task and perform Moran's I for all tasks to test for spatial autocorrelation of the agreement values (Moran, 1950). This analysis is conducted to investigate whether the observed disagreement is randomly distributed or spatially clustered. By calculating Moran's I, we are able to analyze to what degree disagreement is caused by systematic factors of the underlying geographical phenomena (e.g. patterns represented in the satellite imagery).

In the next step, we present insights on all tasks, where volunteers disagreed (all tasks where agreement < 1). In our analysis, we differentiate between seven cases of disagreement. An investigation of the observed quantities can give first information on common false conclusions or misunderstandings by different volunteers. The cases consist of tasks where:

a) Only "Yes" and "No" contributions were captured (referred to as: "Yes-No")
b) Only "Yes" and "Maybe" contributions were captured (referred to as: "Yes-Maybe")
c) Only "Yes" and "Bad Imagery" contributions were captured (referred to as: "Yes-Bad")
d) Only "No" and "Maybe" contributions were captured (referred to as: "No-Maybe")
e) Only "No" and "Bad Imagery" contributions were captured (referred to as: "No-Bad")
f) Only "Maybe" and "Bad Imagery" contributions were captured (referred to as: "Maybe-Bad")
g) More than two different classes were observed (referred to as: "other")

Finally, we provide first qualitative details on the most common cases of disagreement and we show several examples. We then discuss to what degree clouds, large settlements and missing satellite imagery are patterns that can be associated with specific cases of disagreement.

4 Results

In this section, we present the results of our initial analysis. The average agreement among volunteers for the Madagascar project is 0.900 with a standard deviation of 0.218. For South Sudan, it's 0.868 with 0.252. Figure 2 and Figure 3 depict the distribution of agreement per task using violin plots (which combine a density chart with a box plot). In both projects, the agreement among volunteers was very high (0.8 – 1.0) for the clear majority of all tasks. However, there are also tasks with an agreement lower than 0.4.

Figure 2: Agreement "Madagascar 1"

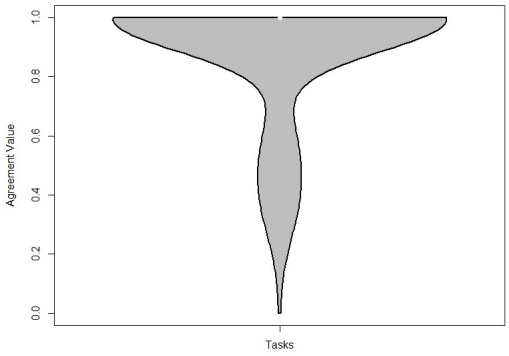


Figure 3: Agreement "South Sudan 1"

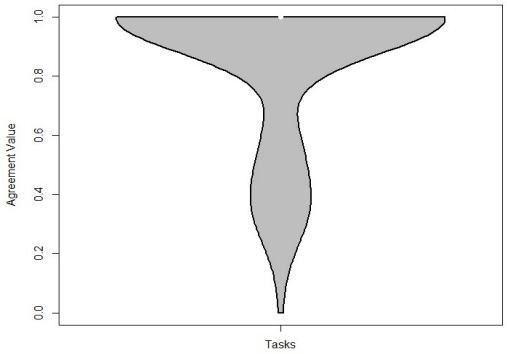


Figure 4: Inter-rater reliability per group ("Madagascar 1")

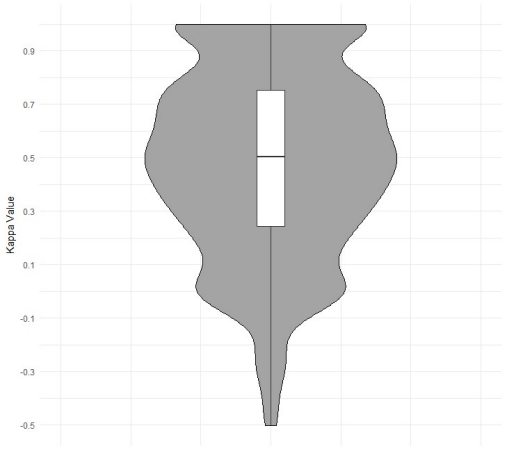
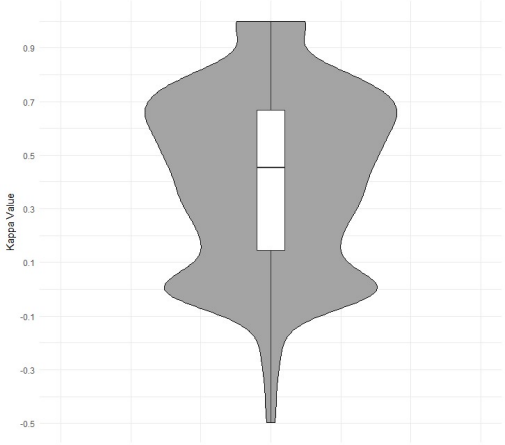


Figure 5: Inter-rater reliability per group ("South Sudan 1")



The inter-rater reliability of the analyzed groups shows a reasonable performance. As depicted in Figure 4 and Figure 5 the average kappa is around 0.5 per group and the majority of groups has a kappa greater than 0.4 indicating a moderate agreement among volunteers for most groups applying the scale presented by Landis & Koch (1977). Only few groups show a poor agreement lower than 0.2, while there is a considerable number of groups where the kappa is even higher than 0.6. These characteristics also apply for the project "South Sudan 1" although there is a stronger bias between groups with good agreement and groups showing poor agreement.

Figure 6 visualizes the spatial distribution of the agreement per task for both MapSwipe projects. The visual interpretation of the map suggests that tasks where volunteers disagree are not distributed randomly, but rather systematically. There are areas with a high concentration of disagreement such as in the northern part of the study area of the project "Madagascar 1". These high concentrations of disagreement are also present for the project "South Sudan 1", e.g. south eastern part. Furthermore, the map depicts bands with a width of about 1.5 – 2.5 kilometers where users disagreed. One example for this can be found in the project "Madagascar 1", where a band of disagreement is spanning through the whole study area from north to south. Finally, the visual interpretation also reveals vertical stripes of disagreement that overlap entirely with individual group geometries. This may be an indicator for a systematic mismatch among volunteers within specific groups.

The analysis of spatial autocorrelation demonstrates a Moran's I of 0.508 with a z-value of 377.488 and a p-value < 0.001. The high z-value and low p-value confirm the visual interpretations of a spatially autocorrelated distribution of agreement. The data from South Sudan also shows a significant spatial autocorrelation indicated by a Moran's I of 0.442 (z-score 244.40; p-value < 0.001). For both projects the agreement is significantly clustered and thus we conclude that the disagreements among volunteers are not randomly distributed, but caused by the underlying geographical phenomena depicted in the satellite imagery.

We differentiated the cases of disagreement for all tasks, where the agreement value was lower than 1 (consensus tasks). The observed task numbers regarding the different cases (a)-(f) of disagreement are presented in table 2. The table indicates that there are three common cases of disagreement: a), d) and e). Together these different cases make up to more than 80% of the tasks where volunteers disagreed for both projects. For the project "Madagascar 1" about 50% of the tasks are of type e). This indicates that there is no clear understanding among volunteers what's the difference between "No" and "Bad Imagery" classification. For the project "South Sudan 1" most tasks are of type a).

Table 2: Cases of disagreement

		Madagascar		South Sudan	
Case	Name	#	%	#	%
(a)	"Yes-No"	9,439	18.3	23,098	38.9
(b)	"Yes-Maybe"	849	1.6	1,498	2.5
(c)	"Yes-Bad"	601	1.2	661	1.1
(d)	"No-Maybe"	9,230	17.9	11,643	19.6
(e)	"No-Bad"	24,257	47.0	14,969	25.2
(f)	"Maybe-Bad"	1,134	2.2	482	0.8
(g)	"other"	6,077	11.8	7,047	11.9
	total	51,587	100	59,398	100

Figure 7 visualizes the spatial distribution of the different cases of disagreement. The visual interpretation of the data indicates that the different cases of disagreement come along with different spatial pattern and characteristics.

For "No-Bad" cases three different subcases exist. Bands from north to south indicate tasks for which no satellite imagery was available. Nevertheless, a considerable number of volunteers didn't classify accordingly and chose "No" instead of "No/Bad Imagery". Furthermore, the same applies for clouds. This subcase of "No-Bad" shows a rather concentrated and not stretched distribution. Finally, some volunteers misunderstood the meaning of tapping 3-times on the screen as "No" and therefore classified whole groups by mistake as "No/Bad Imagery", although they may just have wanted to indicate that there is no building. This is the cause of the vertical, group-wise stretch of "No-Bad" cases. Figure 8 provides examples for these cases.

"Yes-No" cases are often associated with tasks showing large settlements. Accordingly, the number of tasks with "Yes-No" disagreement rises the more settlements exist within the project area . This may be an indicator why the numbers are higher for the project "South Sudan".

Finally, "No-Maybe" cases show only slight spatially clustered characteristics, but not as strong along the bands as the other cases. This is indicated by a Moran's I of 0.227 (z-score 168.69; p-value < 0.001) for "Madagascar 1" and by a Moran's I of 0.145 (z-score 78.92; p-value < 0.001) for "South Sudan 1". From examples, we conclude that some of these cases consist of small geographical features that appear to be houses, but missing some of the typical characteristics of houses like rectangular shape or a clearly identifiable rooftop. Others consist of typical spatial indicators for nearby housing like tracks and flattened areas, but without objects that can undoubtedly identified as houses.

Figure 6: Spatial distribution and autocorrelation of agreement per task

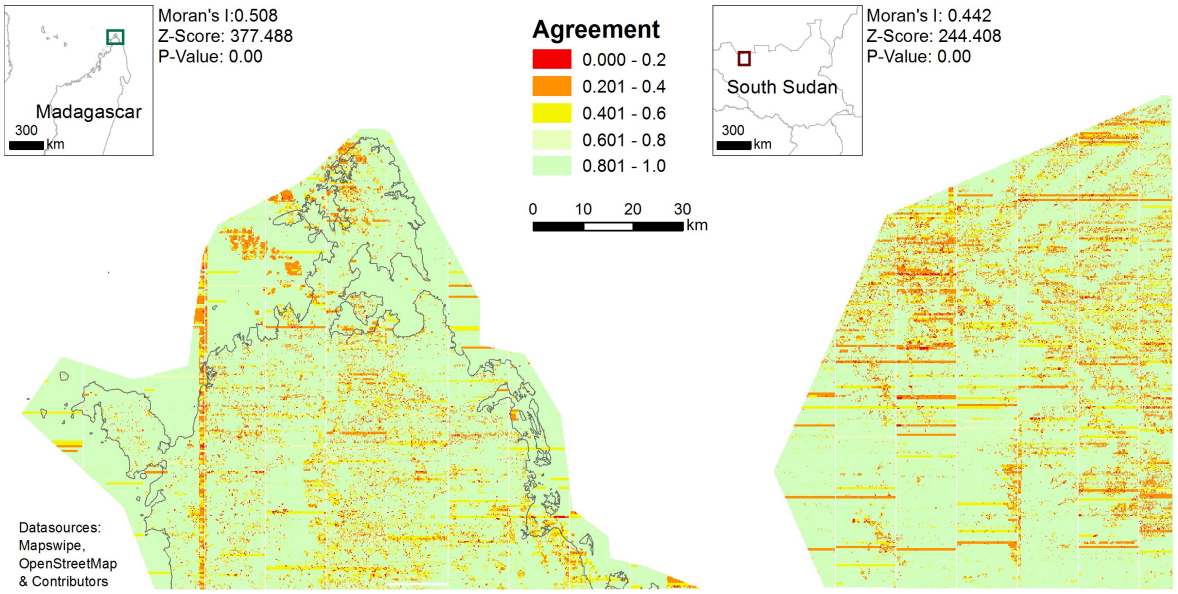


Figure 7: Spatial distribution of different cases of disagreement

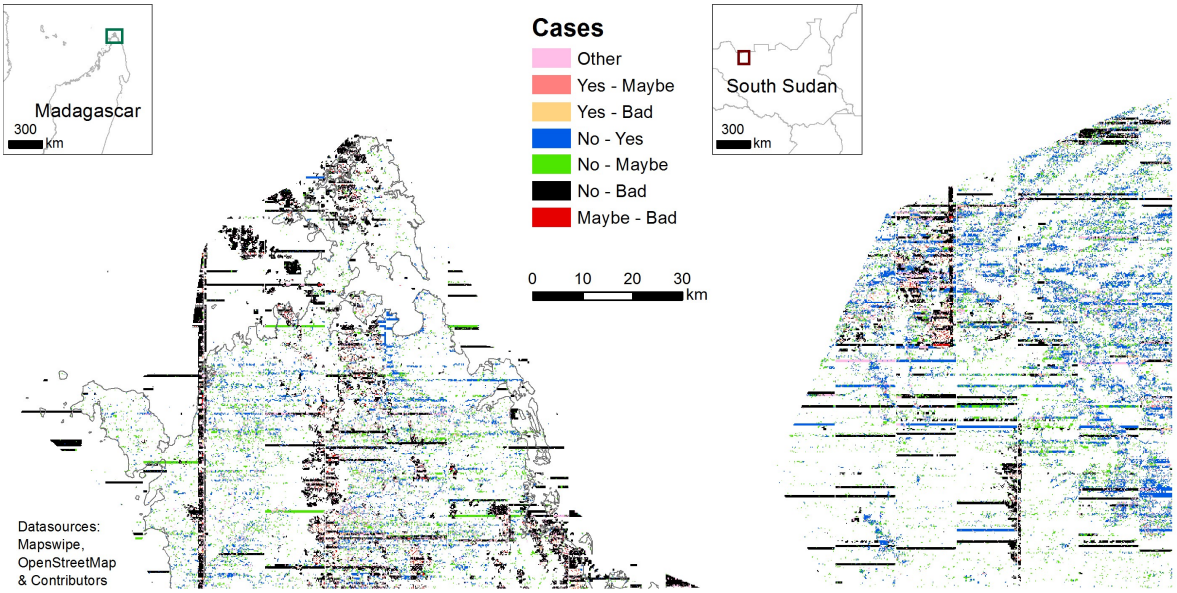
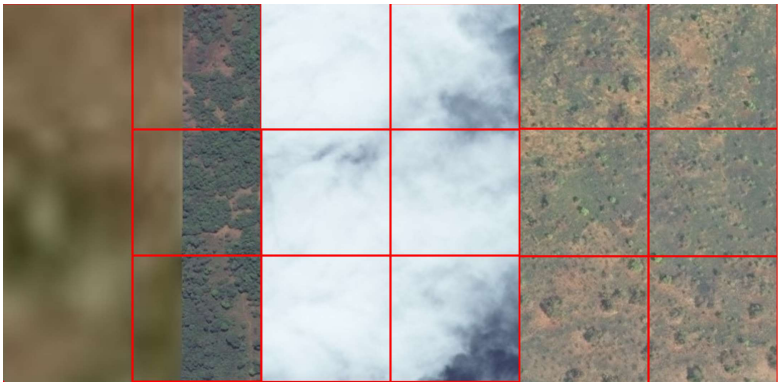


Figure 8: “No-Bad” disagreement (Imagery: Microsoft Bing)



5 Discussion

The results of this study show, that crowdsourced classification of satellite imagery using the mobile application MapSwipe produces reasonable information on human settlements, which seem to be similar to other crowdsourcing approaches and projects, e.g. as described by Porto de Albuquerque et al. (2016) or Arcanjo et al. (2016).

Nevertheless, the results also emphasize that there was disagreement between volunteers for a significant number of tasks. However, such disagreement cases appear not to be random. Their spatially clustered distribution suggests that they are systematically caused by underlying factors. In this study, we observed that clouds, missing satellite imagery and the behavior of individual users are common causes for disagreement among volunteers. The resulting disagreement cases may potentially reduce the quality of the crowdsourced information and thus need to be addressed in future applications.

The insights of this initial study may be used to indicate which types of classification tasks are not well understood by volunteers. Thus, the results may be useful to improve the instructions and the design of the mobile crowdsourcing applications, such as MapSwipe. This is especially important for ambiguous cases like tasks in which a large part of the imagery contains clouds, but some single buildings are also identifiable. Further enhancement of the application (e.g. different interfaces for different screen sizes) or more comprehensive instructions (e.g. an interactive tutorial with messages and example images) for volunteers could minimize the impact of these factors of disturbance and thus improve the quality of the resulting geographic information.

The overall inter-rater reliability for individual groups of the MapSwipe projects has shown fair agreement among volunteers. In the field of building damage assessment Westrope et al. (2014) found a low kappa of 0.22, while David et al. (2016) reach a value greater than 0.8 for the majority of classifications regarding twitter content analysis for a disaster event. Lue et al. (2014) apply Krippendorff's alpha to evaluate the quality of video-based damage assessment and obtain values greater than 0.6 for the majority of tasks.

Nevertheless, there are some groups for which the kappa value is low, although the average agreement of the individual tasks is high. This kind of paradox is well described by several authors, e.g. Feinstein & Cicchetti (1990); Sim & Wright (2005), and is influenced by the prevalence of the some classes within one group. Regarding the MapSwipe data, kappa values for groups where there are only few or no settlements (many "No" classifications) are strongly affected. Future research should therefore consider alternative statistical measures of agreement that offer solutions for this kind of problem.

The lack of an external, reliable reference dataset against which to verify the crowdsourced information is the main limitation of this study. Although the intrinsic indicators employed (i.e. based on the agreement level among volunteers) has been shown to be a valid indicator for data quality (Porto de Albuquerque et al., 2016), it is still necessary to quantify the impact of the factors of disturbance. The results presented here should thus be seen as initial indicators of data quality, which still need to be further analyzed. For instance, future work should investigate the impact of individual users on the overall agreement and how to minimize their influence on the outcomes.

Since this is only an initial study, further research is still needed to improve the existing crowdsourcing approaches and to understand their results better. Future steps in this direction should consider the integration of information generated through crowdsourcing and generated using automated approaches. This could also lead to a better understanding in which fields crowdsourced information can complement and support existing approaches. Expanding the applied crowdsourcing approach to a region where detailed reference data exists or using global data sets such as from WorldPop or Global Human Settlement Layer, and considering the requirements of the data users (e.g. MSF, Red Cross, HOT), would allow a more elaborated comparison of the different approaches.

References

Albuquerque, J., Herfort, B., & Eckle, M. (2016). The Tasks of the Crowd: A Typology of Tasks in Geographic Information Crowdsourcing and a Case Study in Humanitarian Mapping. *Remote Sensing*, *8*(10), 859. https://doi.org/10.3390/rs8100859

Arcanjo, J. S., Luz, E. F. P., Fazenda, Á. L., & Ramos, F. M. (2016). Methods for evaluating volunteers' contributions in a deforestation detection citizen science project. *Future Generation Computer Systems*, *56*, 550–557. https://doi.org/10.1016/j.future.2015.07.005

David, C. C., Ong, J. C., & Legara, E. F. T. (2016). Tweeting supertyphoon Haiyan: Evolving functions of twitter during and after a disaster event. *PLoS ONE*, *11*(3), 1–19. https://doi.org/10.1371/journal.pone.0150190

Feinstein, A. ., & Cicchetti, D. V. (1990). High Agreement But Low Kappa : I . the Problems of Two Paradoxes *. *J Clin Epidemiol*, *43*(6), 543–549.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*. https://doi.org/10.1037/h0031619

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., … Obersteiner, M. (2009). Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. *Remote Sensing*, *1*(3), 345–354. https://doi.org/10.3390/rs1030345

Imran, M., Castillo, C., Meier, P., & Diaz, F. (2013). Extracting Information Nuggets from Disaster- Related Messages in Social Media, (May), 791–801.

Kerle, N., & Hoffman, R. R. (2013). Collaborative damage mapping for emergency response: The role of Cognitive Systems Engineering. *Natural Hazards and Earth System Science*, *13*(1), 97–113. https://doi.org/10.5194/nhess-13-97-2013

Landis, J. R., & Koch, G. G. (2008). The Measurement of Observer Agreement for Categorical Data Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2529310. *Society*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lue, E., Wilson, J. P., & Curtis, A. (2014). Conducting disaster damage assessments with Spatial Video, experts, and citizens. *Applied Geography*, *52*, 46–54. https://doi.org/10.1016/j.apgeog.2014.04.014

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, *37*(1–2), 17–23. https://doi.org/10.1093/biomet/37.1-2.17

Schepaschenko, D., See, L., Lesiv, M., McCallum, I., Fritz, S., Salk, C., … Ontikov, P. (2015). Development of a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics. *Remote Sensing of Environment*, *162*, 208–220. https://doi.org/10.1016/j.rse.2015.02.011

Sim, J., & Wright, C. C. (2005). Interpretation, and Sample Size Requirements The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Phys Ther*, *85*(3), 257–268. https://doi.org/15733050

Westrope, C., Banick, R., & Levine, M. (2014). Groundtruthing OpenStreetMap Building Damage Assessment. *Procedia Engineering*, *78*, 29–39. https://doi.org/10.1016/j.proeng.2014.07.035