AGILE: GIScience Series, 4, 5, 2023. https://doi.org/10.5194/agile-giss-4-5-2023
Proceedings of the 26th AGILE Conference on Geographic Information Science, 2023.
Editors: P. van Oosterom, H. Ploeger, A. Mansourian, S. Scheider, R. Lemmens, and B. van Loenen.
This contribution underwent peer review based on a full paper submission.

Exploring MapSwipe as a Crowdsourcing Tool for (Rapid) Damage Assessment: The Case of the 2021 Haiti Earthquake

Simon Groß[1], Benjamin Herfort[1,2], Sabrina Marx[1], and Alexander Zipf[1,2]

[1]Heidelberg Institute for Geoinformation Technology, 69120 Heidelberg, Germany
[2]GIScience Chair, Institute of Geography, Heidelberg University, 69120 Heidelberg, Germany

Correspondence: Benjamin Herfort (benjamin.herfort@heigit.org)

Abstract. Fast and reliable geographic information is vital in disaster management. In the late 2000s, crowdsourcing emerged as a powerful method to provide this information. Base mapping through crowdsourcing is already well-established in relief workflows. However, crowdsourced post-disaster damage assessment is researched but not yet institutionalized. Based on MapSwipe, an established mobile application for crowdsourced base mapping, a damage assessment approach was developed and tested for a case study after the 2021 Haiti earthquake. First, MapSwipe's damage mapping results are assessed for quality by using a reference dataset in regard to different aggregation methods. Then, the MapSwipe data was compared to an already established rapid damage assessment method by the Copernicus Emergency Management Service (CEMS). Crowdsourced building damage mapping achieved a maximum F1-score of 0.63 in comparison to the reference data set. MapSwipe and CEMS data showed only slight agreement with Cohen's Kappa values reaching a maximum of 0.16. The results highlight the potential of crowdsourcing damage assessment as well as the importance for a scientific evaluation of the quality of CEMS data. Next steps for further integrating the presented workflow into MapSwipe are discussed.

Keywords. Volunteered Geographic Information, Crowdsourcing, Disaster and Risk Management, Open Geo Data, Building Damage Mapping

1 Introduction

To support disaster preparedness and response activities, volunteered geographic information (VGI) or data obtained through crowdsourcing is already used in many situations and since 2012 humanitarian mapping efforts such as post-disaster mapping campaigns improved the spatial coverage of OpenStreetMap (OSM) considerably (de Albuquerque et al., 2016; Scholz et al., 2018; Herfort et al., 2021).

However, most of the current research about humanitarian mapping has focused on general base mapping (e.g. building, roads and land cover). In the following we will provide a more detailed perspective on crowdsourced damage mapping approaches.

1.1 Crowdsourced Damage Assessment

The European Macroseismic Scale issued in 1998 by Grunthal (1998) (EMS-98 scale), which rates the damage degree (as in how strongly a single building is damaged) in five categories from slight damage to complete destruction, is commonly used in post-disaster damage assessment. However, as pointed out by Huynh et al. (2014) this scale is highly limited when using remote sensing for damage assessment, since damage degree is often not sufficiently distinguishable from a bird's-eye perspective. Studies regarding crowdsourced damage assessment therefore moved to focus on mapping damage extent (e.g. number of damaged or destroyed buildings) instead. First attempts regarding crowdsourced damage assessment were analysed by Kerle (2011), calling for universal standards since damage assessment was performed with different aggregation methods and damage scales.

In the aftermath of 2013 Taifun Haiyan, Westrope et al. (2014) analysed satellite based building damage mapping in OSM with respect to three classes (no damage, major damage, destroyed). Further case studies acknowledged the difficulty of mapping damage degree, for example in regard to the 2010 Yushu Earthquake in China (Xie et al., 2016). Several authors acknowledged that mapping damage is a fundamentally more difficult task for volunteers than base mapping, where only buildings themselves are mapped (Kerle, 2011; de Albuquerque et al., 2016; Westrope et al., 2014).

In contrast to approaches solely powered by satellite data, Khajwal and Noshadravan (2021) combined ground and aerial pictures with crowdsourced information gathered through surveys attached to the pictures for the case of 2017 Hurricane Harvey. In their work the authors could go beyond the limited damage scale and collected information on the condition of walls, roof structures and windows. The limits of the approaches based on satellite imagery point towards the potential of using higher resolution imagery for damage mapping, e.g. obtained from drones or ground level pictures, which are albeit their strongly improved availability not yet sufficiently utilized for crowdsourcing applications.

Using crowdsourcing as a means to create baseline map data in OSM for disaster response is nowadays an established approach and a community of practice has formed (Soden and Palen, 2016; Herfort et al., 2021). The open-source mobile application MapSwipe has proven to be a valuable crowdsourcing tool to derive data on human settlements (Scholz et al., 2018; Herfort, 2018). Besides baseline mapping, MapSwipe shows great potential to support (rapid) building damage assessment. However, a scientific evaluation of the crowdsourced damage information with respect to data quality is missing so far.

1.2 Institutionalized Damage Assessment

Institutionalizing crowdsourced damage assessment systematically began in 2008 after the Wenchuan Earthquake in China. Experts from different kinds of backgrounds coordinated a joint effort and formed the Global Earth Observation Catastrophe Assessment Network (GEO-CAN), which helped coordinate efforts during the next major earthquake (Barrington et al., 2012).

The GEO-CAN initiative helped coordinate a widespread damage assessment effort, where more than 600 volunteers from different stakeholders (governments, companies, universities, NGOs) were notified and instructed via email (Barrington et al., 2012).

As of 2022, the GEO-CAN initiative to our knowledge does not operate and Tomnod, a crowdsourcing coordination platform, retired their operations in August 2019 (Maxar Technologies, 2019). Crowdsourcing approaches have still been used after 2011 (Kuzin et al., 2021; Xie et al., 2016; Khajwal and Noshadravan, 2021), but these efforts where not institutionally connected to each other and data was collected independently in these cases. Efforts were also made in 2017 when the Humanitarian OpenStreetMap Team (HOT) conducted research on a rapid damage assessment approach on their crowdsourcing platform (Giovando, 2017). But as of today, crowdsourced damage assessment is not sufficiently institutionalized.

Institutionalized but not crowdsourced is the European Union's Copernicus Emergency Management Service (CEMS) (Copernicus Emergency Management Service, 2022a). CEMS operates since 2012 and produces reports, maps and (geographic) data sets, which are publicly available on their website for download. These include rapid mapping and damage assessment as well as data for risk and recovery. Havas et al. (2017) studied the possible integration of VGI into the CEMS workflow.

As CEMS maps are expected to be available within a few hours after the event, image acquisition primarily relies on satellite data and the mapping itself continues to be carried out manually (Kerle et al., 2019). Furthermore, services such as CEMS which provide rapid mapping products are not always available or activated, especially when the events are too small to be considered relevant (Notti et al., 2018).

1.3 Automated Damage Assessment

Whereas automated damage assessment based on satellite imagery has been proven to be rather difficult, the growing availability of unmanned areal vehicles (UAVs) and drones has opened up new potential for automated and detailed damage assessment (Kerle et al., 2019). When imagery is captured using UAVs damage detection is not limited to a two-dimensional representation of the ground, but can further exploit the potential of three-dimensional point clouds derived through structure from motion approaches (de Gélis et al., 2021). This also leads to a declining need to rely on formal damage mapping products such as the one produced by CEMS, and facilitates on-site mapping (Kerle et al., 2019).

However, automated approaches for building mapping in general and damage mapping in particular face challenges in regard to the transfer learning capacity and are relying on accurate and sufficient amount of training data (de Gélis et al., 2021; Li et al., 2022; Kerle et al., 2019). To overcome the constraints of UAV flight planning Kerle et al. (2019) imagine a two-step approach where first hotspot and damage candidates are identified. In a second step, a more detailed and multi-perspective survey is conducted.

To facilitate the initial rapid damage mapping Zahs et al. (2021) designed a tool which produces data which can either functions as training or validation data for the machine learning powered point cloud-based damage classification. The authors propose an interdisciplinary approach for timely and reliable assessment of building-specific damage grades (0-5) from UAV images (and point clouds) with high resolution (centimetre pixel size). The approach further relies on the combination of expert knowledge of earthquake engineers with fully automatic damage classification and human visual interpretation from crowdsourcing.

For the case of building mapping, it has been shown that this combination of human visual interpretation skills and automated mapping approaches can boost quality and efficiency at the same time (Herfort et al., 2019). Nevtherless, the potential of crowdsourced data collection for au-

tomated building damage detection from UAV imagery is yet to be explored.

1.4 Research Questions

Whereas there is a plethora of research about crowdsourcing as a method for damage assessment from satellite imagery, these approaches are not yet institutionalized, which prevents open data to be forwarded to relief organizations quickly and efficiently. Hence, in this paper we want to investigate to what extent the MapSwipe app could be used as a tool to engage volunteers in that envisioned workflow by examining the following specific research question.

- How well does MapSwipe perform as a tool for crowdsourced building damage classification? (RQ1)
- How does crowdsourced building damage information from MapSwipe compare to institutionalized information obtained through CEMS? (RQ2)

The remainder of this paper is organised as follows. The next section presents the basis for this research, explaining the case setting and used data sets. Section 3 shows the methodology used for analysing MapSwipe data for a building damage mapping project and how it is compared to CEMS data. Section 4 presents the results achieved from this analysis, whilst Section 5 provides a discussion of the results and Section 6 concludes this paper by making recommendations for practitioners and future research.

2 Case Study

2.1 Haiti Earthquake, 2021

On the 14th of August 2021 at 08:29 local time a 7.2 magnitude earthquake hit Haiti with a depth of approximately 10 km. The epicentre was located in the department Nippes, central on Haiti's southern peninsula. Approximately 2.3 Million people were exposed within a 100 km radius, 800,000 people were directly affected, 650,000 were in need of humanitarian assistance, more than 137,000 homes were damaged or destroyed including critical infrastructure, more than 12,200 people were injured and at least 2,248 people were killed. Economic damage is estimated to be at least 1.5 billion USD, approximately 10 percent of Haiti's GDP (GDACS, 2021; OCHA, 2022).

This study focuses on the city of Les Cayes, which is located approximately 40 km south west from the epicentre (Figure 1).

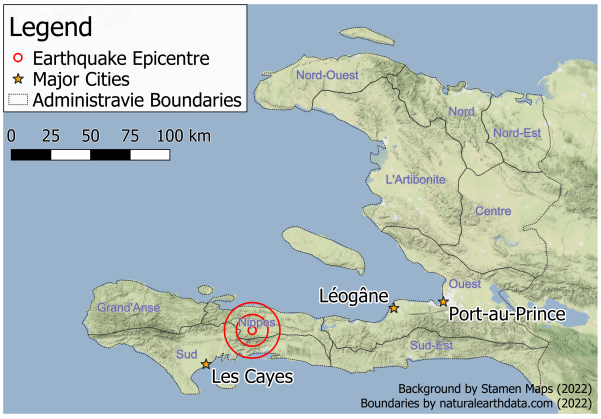

Figure 1. Overview on Haiti, the earthquake epicentre and surrounding major cities. Background layer provided by Stamen Maps (2022), terrain style.

2.2 Datasets

2.2.1 MapSwipe

MapSwipe's default change detection project type, which was used here to assess damage, is based on the tile level. For this project type, instead of only looking at six tiles at once (see MapSwipe's original project type Figure 2 left), users now compare a pair of satellite or aerial imagery tiles (Figure 2 right). Both tiles depict the same area, the upper tile shows the situation before the earthquake and the second tile shows the situation after the earthquake. In our case, the before image is an orthophoto from Haiti's open data portal (HaitiData, 2021) with an unknown resolution and the after image can be found on OpenAerialMap (Gastaminza, 2021) with an original resolution of 2 cm. The volunteers can then decide if buildings inside the respective tile have been damaged. For this study, a tile corresponded to the definition applied by tile map services (TMS) at zoom level 20 (Maso et al., 2010). The respective edge lengths for the tiles are ~ 40 Meters, that gives the pictures a maximum resolution of 15.6 cm in the app (initial resolution of the drone imagery was 2 cm). In total, the entire project area was covered by 1,806 mapping tasks.

The MapSwipe data was directly obtained from the projects page (MapSwipe, 2022b), where the results for all finished mapping projects can be found. For this study we selected the project named "Earthquake - Experimental Damage Assessment - Haiti (2) HOT", which was requested by the Humanitarian OpenSteetMap Team. It saw contributions from 116 volunteers and was completely assessed within 7 days.

Data concerning these projects can be downloaded in different formats and aggregations, here we selected "Aggregated Results (with Geometry)". The aggregated results are the individual tiles in geojson format with different attributes shown in Table 1. Each tile has been classified by

at least 7 volunteers into one of the following categories: "no damage" (class 0), "damaged" (class 1), "maybe damaged" (class 2), "bad imagery" (class 3).

Figure 2. Different MapSwipe interfaces. **Left:** Original base mapping interface for MapSwipe with six tiles, green for "building", yellow for "maybe", red for "bad imagery" and no color for "no building". **Right:** Damage assessment interface with before picture on the top and after picture on the bottom. The single tile can be marked with the same colors, except that "building" means "damaged".

2.2.2 CEMS Damage Mapping

CEMS provided multiple damage assessment layers for different regions in Haiti. In this study the product "[EMSR536] Les Cayes: Grading Product, version 1" was used (Copernicus Emergency Management Service, 2022b). CEMS was activated less than 4 hours after the earthquake and the data regarding Les Cayes was published about 35 hours after the event. For this study, we utilized a layer with points, placed approximately in the middle of the damaged building. Each point contains three different damage grading possibilities "destroyed", "damaged" and "possibly damaged". The first two were considered together, since a MapSwipe task does not differentiate between damaged or destroyed. Since a point can only be inside one MapSwipe tile even if the damaged building extends over multiple tiles, we used OSM building footprints to join the points with their nearest building (OpenStreetMap and contributors, 2017).

It is important to note that this dataset does not represent a ground truth. According to the overview map, the Copernicus analysis is based on an Airbuses Pléiades 1 A/B satellite picture, which has an approximate resolution of around 50 cm (Copernicus Emergency Management Service, 2021). This, albeit being considered "very high resolution" for satellite imagery, is magnitudes lower than the 2 cm (or 16 cm in MapSwipe) drone imagery utilized in the MapSwipe project.

Hence, CEMS data is used to compare the MapSwipe damage assessment approach to already established methods, but is not a sufficient data set to assess MapSwipe's performance and data quality.

2.2.3 Reference Data

In order to analyze MapSwipe's data quality, the authors of this study re-mapped the entire MapSwipe project using the private development instance of MapSwipe. Each task was carefully processed by two persons and disagreements between the raters were eliminated by discussing the respective tiles and using the best available resolution (2 cm instead of 16 cm) and visual assessment in QGIS.

3 Methodology

3.1 MapSwipe Data Quality Assessment

In crowdsourcing approaches which rely on redundant classification there is a particular importance to define the "correct" threshold when aggregating the contributions of individual users into a consensus answer (de Albuquerque et al., 2016). To account for this, different thresholds and their effect on data quality have been explored in this study. A threshold describes the share of volunteers that tag a tile as damaged. So if 5, 10, ..., 95 percent of the users tagged a tile as damaged it will be classified as damaged. By choosing a threshold-based approach we aim at a good balance between precision and sensitivity. Based on these aggregated MapSwipe results and the reference dataset (see 2.2.3), we used the scikit-learn library for Python to create a confusion matrix. This returns the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) and calculates the True Positive Rate (TPR) and False Positive Rate (FPR), further described in Table 2. These values are then used to calculate different data quality measures: accuracy, precision, sensitivity (same as TPR) and the F1-score, also described in Table 2. The detailed workflow is depicted in Figure 3.

3.2 MapSwipe Copernicus Comparison

Agreement analyses conducted between MapSwipe and Copernicus data were accomplished using confusion matrices based on the MapSwipe tiles and Copernicus buildings footprints. Again, the whole workflow is depicted in Figure 4.

The comparison is then first shown in a confusion matrix, where one column/row shows the number of damaged tiles and the other one non-damaged cells. The number of agreed tiles is therefore the diagonal sum from top-left to bottom-right. For further agreement analysis, Cohen's

Table 1. Excerpt from the aggregated MapSwipe data. Some columns were excluded or shortened (MapSwipe, 2022a). "0_count" refers to the number of users who labelled a task as no damage. "1_count" refers to the number of users who labelled a task that contained a damaged building.

idx	task_id	0_count	1_count	2_count	3_count	total_count	shares	agreement	geometry
644	19-154739-235159	1	2	4	0	7	...	0.333	MULTIPOLYG...
746	19-154718-235162	0	7	0	0	7	...	1	MULTIPOLYG...
1140	19-154744-235166	1	3	3	0	7	...	0.286	MULTIPOLYG...
1779	19-154738-235181	5	2	1	0	8	...	0.393	MULTIPOLYG...
1811	19-154748-235183	7	1	0	0	8	...	0.75	MULTIPOLYG...

Table 2. Different quality parameters and their explanations.

Measure	Formular	Description
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$	How many of all not damaged tiles were incorrectly classified as damaged?
True Positive Rate (TPR) or Sensitivity	$\frac{TP}{TP+FN}$	How many of all damaged tiles were correctly identified?
Precision	$\frac{TP}{TP+FP}$	How many of all damaged tiles were correctly labeled?
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	How many tiles of all tiles were correctly classified?
F1-Score	$\frac{2*Sensitivity*Precision}{Sensitivity+Precicion}$	The harmonic mean between sensitivity and precision

Kappa was used (Cohen, 1960). This measure is used to compare two different labelling methods for the same problem. The formula is as follows (Watson and Petrie, 2010):

$$\kappa = \frac{ObservedAgreement - ChanceAgreement}{MaximumAgreement - ChanceAgreement} \quad (1)$$

$$\kappa = \frac{p_0 - p_E}{1 - p_E} \quad (2)$$

Cohen's Kappa accounts for consensus that occurs randomly. Values range from $-1 < \kappa < 1$. 1 represents full consensus, $\kappa > 0$ agreement, $\kappa = 0$ no better than chance agreement and $\kappa < 0$ disagreement. Arbitrary levels are proposed by Landis and Koch (1977) where the values 0.00 – 0.2, 0.21 – 0.4, 0.41 – 0.6, 0.61 – 0.8 and 0.81 – 1 correspond to slight, fair, moderate, substantial, and almost perfect agreement. Ranganathan et al. (2017) also provide an interpretation for $\kappa < 0.6$ as "significant level of disagreement". Using this metric, agreement scores between different damaged definitions of MapSwipe are compared with either all Copernicus buildings or excluding the "possibly damaged" buildings. It must be noted that Cohen's Kappa and its variants are also criticized in the context of remote sensing because they are complicated to compute, difficult to understand and interpret (Pontius and Millones, 2011).

3.3 Data and Software Availability

All code and data that has been produced for this manuscript can be examined on the following GitHub page: https://github.com/simsi44/mapswipe_rapid_damage_assessment.

4 Results

4.1 MapSwipe Quality Assessment

Figure 5 shows the aggregated MapSwipe results and the spatial distribution of tasks for which MapSwipe users identified damaged buildings. The map provides a first visual impression that damage buildings in Les Cayes were primarily located in the eastern part of the study area.

Figure 6 shows the MapSwipe threshold definitions plotted against the four different quality measures. Precision rises with a higher threshold since it is more likely for damaged labels to be correct with a higher threshold, and sensitivity falls since more and more damaged tiles are overlooked. Accuracy rises until a threshold of 45 percent (0.83) and then, after a slight fall, stays about the same until the end. This illustrates the weakness for accuracy as a quality measure, since with high thresholds the high amount of TNs overshadow the FNs. Therefore, the F1-score is the better indicator here. It peaks at a threshold

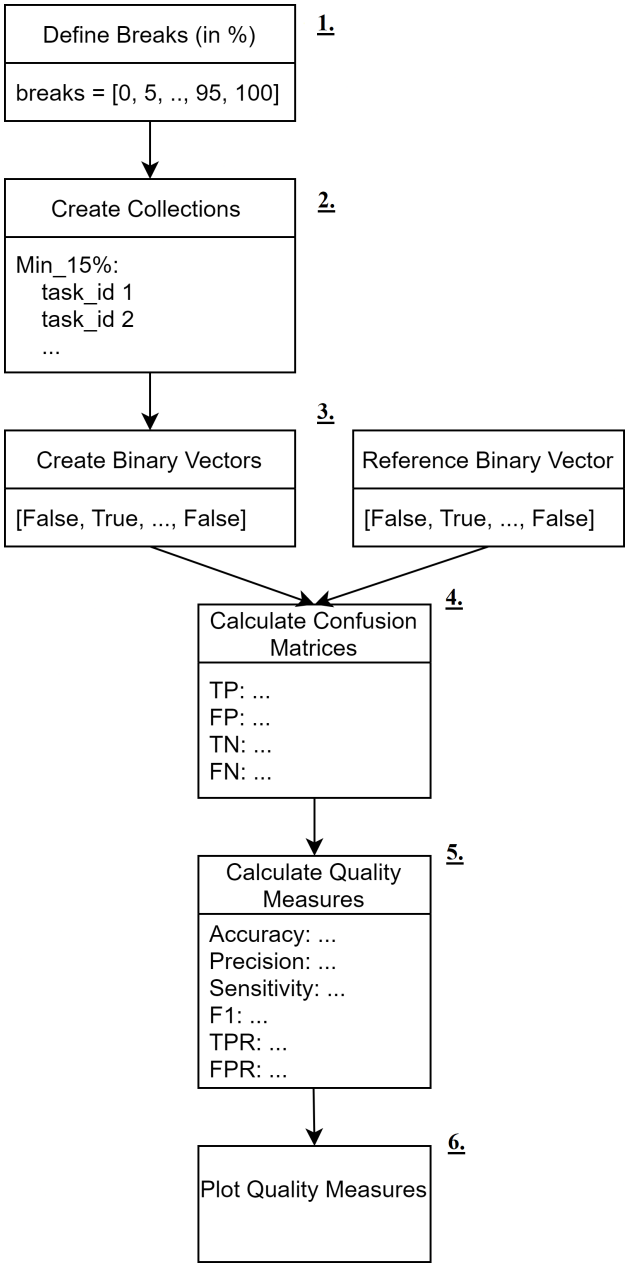

Figure 3. Workflow diagram for quality assessment (RQ1): **1.** Define step granularity **2.** Create collection for each step **3.** Create binary vectors for the collections and the reference dataset **4.** Calculate confusion matrices for each collection **5.** Derive quality measures **6.** Visualization.

of 35 percent (0.63) but only differs slightly at 30 or 40 percent (both 0.62). Based on this, simply relying on majority voting is not enough for this use case, since better results can be achieved setting the threshold lower than 50 percent. For the case of Les Cayes our results show that the aggregated result should be classified as "damage" if at least 35 percent of the individual raters identified damage in the aerial imagery.

Figure 7 shows the confusion matrix in a spatial context.

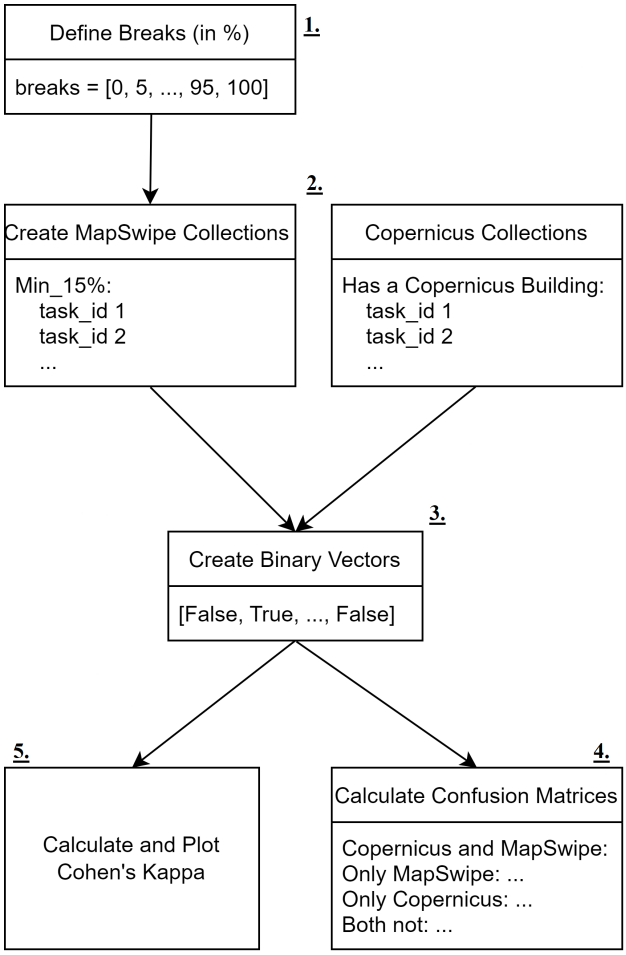

Figure 4. Workflow diagram for comparison with Copernicus data (RQ2): **1.** Define step granularity **2.** Create collections for each step and for Copernicus buildings **3.** Create binary vectors for each collection **4.** Calculate confusion matrices for each collection **5.** Calculate Kappa and Visialization.

4.2 MapSwipe Copernicus Comparison

After exploring how MapSwipe data performs, it can be compared to CEMS data. Figure 8 provides an overview for which tasks both approaches detected damage (54 tasks in total). Furthermore there were 34 tasks for which only CEMS detected damaged buildings, whereas there were 586 tasks for which damage was detected by MapSwipe users, but not CEMS. This might already suggest, that minor damage that was overlooked by CEMS, could be identified on the MapSwipe pictures due to the increased imagery resolution.

Examples of this confusion matrix' cells are depicted in Figure 9 with the respective MapSwipe answers. In the first picture, both MapSwipe and CEMS found a destroyed building. It is however still unclear if all MapSwipe users tagged the picture because of the destroyed building at the bottom or because of the debris on the top-right. The second picture is almost unanimously tagged without damage in MapSwipe, Copernicus may have found damaged here

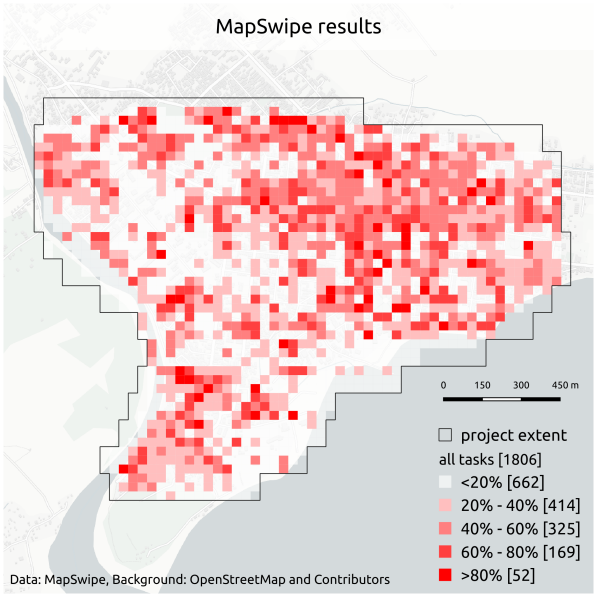

Figure 5. MapSwipe aggregated results for damage assessment. Percentages correspond the the share of volunteers that tagged the respective tile as damaged.

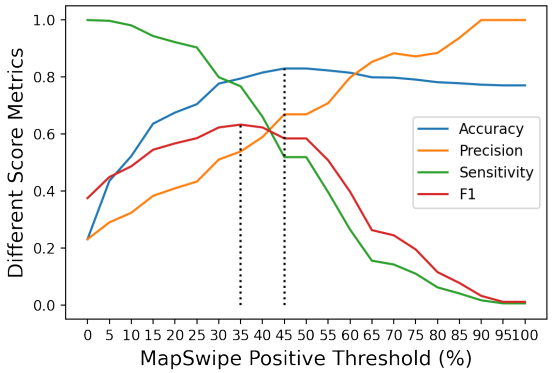

Figure 6. Quality parameters for different damage definition thresholds for MS-small. F1 peaks with sensitivity above precision. The accuracy high level for greater thresholds indicates its weaknesses in this context.

on the day after the event that is not recognizable anymore after ten days, when the MapSwipe image was taken, or it was a false positive tag. The third picture shows an example where there is clearly damage visible and it is tagged by the volunteers. However, CEMS could not find this one, maybe due to the worse resolution. On the fourth picture in both cases no damage could be found.

Cohen's Kappa was plotted with different thresholds for a positive definition in MapSwipe compared to the damaged buildings tagged by CEMS (Figure 10). The dashed line represents all CEMS tagged buildings, the solid line only the CEMS buildings tagged as "damaged" or "destroyed" (excluding "possibly damaged" buildings"). In both cases Kappa is almost monotonically increasing until 65 percent

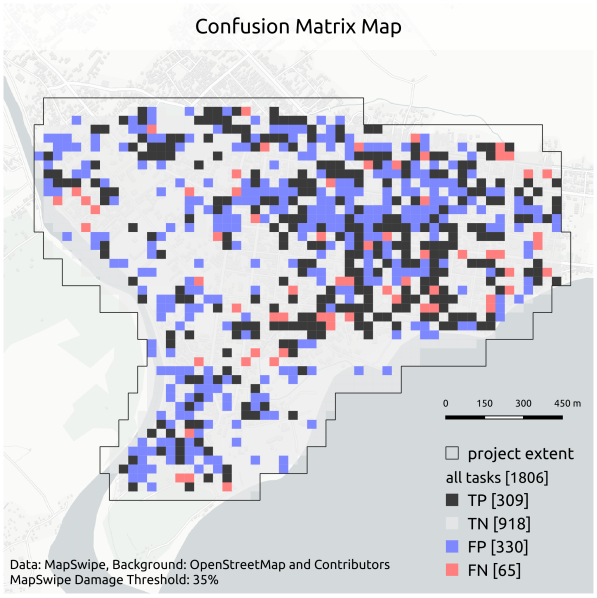

Figure 7. Spatial distribution of false positives, true positives, false negatives and true negatives considering a threshold >= 35 %.

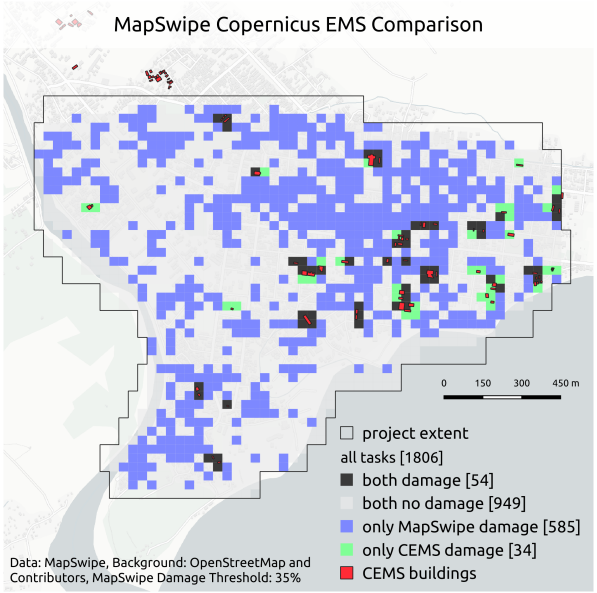

Figure 8. Spatial distribution of agreement between MapSwipe results and CEMS results, considering a threshold >= 35 %.

and then decreasing. That means that in order to reach maximum agreement with Copernicus data, 65 percent of users have to declare a tile damaged. Compared with the results from the quality assessment, the MapSwipe threshold has to be defined much stricter.

Kappa itself is astoundingly low, with only slight agreement in all relevant cases. That suggests, that generally CEMS data, and MapSwipe data do not compare well. When ignoring possibly damaged buildings in CEMS

Figure 9. Example tiles to assess agreement between Copernicus and MapSwipe. The white box gives information about the MapSwipe answers. **Top-left:** Tagged as damaged by MapSwipe and Copernicus. **Top-right:** Only tagged by Copernicus. **Bottom-left:** Only tagged by MapSwipe. **Bottom-right:** Not tagged by both.

agreement was even lower. This indicates that possibly damaged Copernicus buildings are often not tagged as damaged be MapSwipe users.

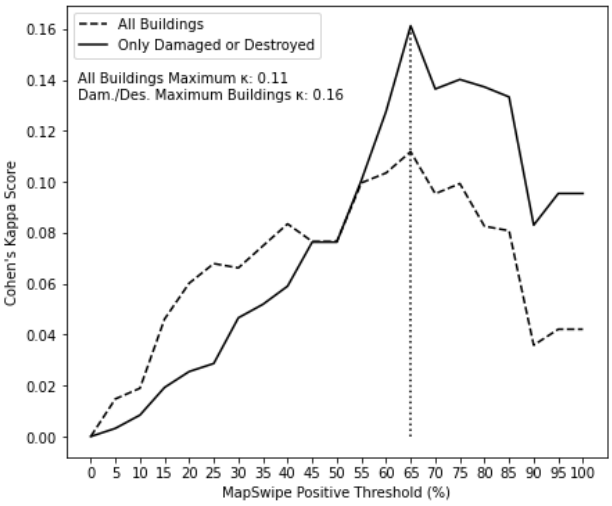

Figure 10. Cohen's Kappas between MapSwipe and CEMS labels by damage definition threshold. In both cases Kappa peaks at a threshold value of 65 %. Agreement is very low, which could be explained by different picture acquisition times and resolutions. For the MapSwipe threshold with maximum F1-score (35 %) Kappa values are 0.075 and 0.052 for all and damaged/destroyed buildings.

5 Discussion

First of all, it is to note that the results from damage assessment have to be discussed under different premises compared to previous MapSwipe projects. It is well known that damage assessment or change detection generally is a harder task for volunteers than base mapping, which is the original intent of MapSwipe (de Albuquerque et al., 2016; Kerle, 2011). Even though F1 scores using the reference data set of more than 0.60 pale in comparison to the classic MapSwipe approach (F1 >= 0.90; (Herfort, 2018)), reaching these numbers was neither the claim nor the expectation. Generally, the achieved F1 scores of up to 0.63 provide a positive outlook for this method being used in the future and hint at potential of this approach.

Arguably the most important factor influencing data quality is the underlying imagery. A lot of damage tagged in the MapSwipe project is simply only visible because of the picture's astounding quality. Kerle (2011) already argued that improving picture resolution from 50 to 15 cm, increases crowdsourced damage tags up to ten times.

The results of the CEMS data comparison also provide interesting implications. The first idea during the beginning of this study was to use CEMS data more as a verification dataset instead of a comparison. Unfortunately, CEMS data turned out not to be of sufficient quality. Some reasons can be theorized, for example, the quick creation of the datasets or the comparatively lower resolution of the satellite imagery utilized. Information about quality control protocol defined by CEMS can be found on their website Copernicus Emergency Management Service (2022c). After working with this data during this study, it can be assumed that CEMS did not reflect the actual extent of the damage on the ground correctly. However, analysing CEMS data quality in detail is beyond the scope of this study, but should be addressed in future research.

This has been highlighted by Elia et al. (2018) as well, who identify the strong necessity for (on-the-ground) data verification of professional and crowdsourced damage assessment products. The work of Mulder et al. (2016) in regard to the Haiti 2010 and the Nepal 2015 earthquakes also reminds us of the importance of incorporating local knowledge into data creation and the need to discuss social factors and their implication for disaster relief.

Another future research domain should investigate the potential of automatized damage detection. For example, Resch et al. (2018) are enhancing georeferenced Twitter data through semantic information extraction with spatial and temporal analysis for hot spot detection to assess the footprint of and the damage caused by the 2014 Napa, California Earthquake. Kuzin et al. (2021) take rapidly acquired and weighted crowdsourced labels and use them to train a neural network to identify approximate areas of damage for the case of 2017 hurricanes Irma and Maria, which impacted multiple islands in the Caribbean. Recent work highlights the potential of convolutional neural net-

works in comparison to traditional approaches especially when considering the wide variety of sensors and spatial resolution of imagery, e.g. captured from space, aerial and UAV platforms (Nex et al., 2019). Novel methods, which do not only rely on two-dimensional images, but exploit the rich geometric characteristics which can be derived from dense 3D point clouds, broaden the practical use of automated damage detection (Vetrivel et al., 2018).

The potential that lies in the combination of crowdsourcing and deep learning approaches to improve data quality and mapping performance is highlighted by Herfort et al. (2019), however they apply their method to crowdsourced settlement mapping. Future work should explore whether this potential is also met for more complex tasks, such as damage mapping.

6 Conclusion

In this study, a new use case for MapSwipe was explored. It was theorized that an app that was built for crowdsourcing settlement detection could also be used to assess damage after a natural disaster. This was framed by two research questions: (1) How well is the resulting dataset suited for building damage classification and (2) how does this approach perform compared to CEMS data?

The quality measures derived for MapSwipe show a clear potential for this new use case. However, it is dependent on two factors: aggregation method and underlying pictures. The research shows that moving the threshold for a damage definition can substantially influence the data's quality measures. Thresholds should therefore be defined, depending on the use case. Typical mistakes of MapSwipe users could be identified, which could be incorporated into tutorial before a user starts mapping. Furthermore, MapSwipe does not harness the full potential of the aerial or satellite pictures due to the download resolution, which leads to a reduction in potential data quality from 2 cm to 16 cm.

Apart from the quality perspective, the question remains if this use case will work in real conditions. The work by Zahs et al. (2021) and Kerle et al. (2019) clearly outline the trend towards on-site UAV based damage mapping. Whether MapSwipe will function as a tool for such damage assessment approaches depends on many more factors than just data quality, including how quickly imagery can acquired and made available for crowdsourcing. However, MapSwipe already provides two critical assets: an established community and infrastructure. Up to this point, MapSwipe projects were not very time sensitive, but if urgency can be communicated correctly (for example through notifications on the smartphone), processing times could potentially be sped up further. Whereas MapSwipe has proved that it works 'in production' for large scale base mapping from satellite imagery, it has to be explored further for damage mapping using drone imagery.

References

Barrington, L., Ghosh, S., Greene, M., Har-Noy, S., Berger, J., Gill, S., Yu-Min, A., and Huyck, C.: Crowdsourcing earthquake damage assessment using remote sensing imagery, Annals of Geophysics, 54, https://doi.org/10.4401/ag-5324, 2012.

Cohen, J.: A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, 20, 37–46, https://doi.org/10.1177/001316446002000104, 1960.

Copernicus Emergency Management Service: Earthquake-Situation as of 15/08/2021: Grading-Overview map 01, https://emergency.copernicus.eu/mapping/system/files/components/EMSR536_AOI02_GRA_PRODUCT_r1_RTP01_v1.pdf, 2021.

Copernicus Emergency Management Service: Copernicus Emergency Management Service, https://emergency.copernicus.eu/, 2022a.

Copernicus Emergency Management Service: Copernicus Emergency Management Service - Mapping, https://emergency.copernicus.eu/mapping/list-of-components/EMSR536, 2022b.

Copernicus Emergency Management Service: Quality control, https://emergency.copernicus.eu/mapping/ems/quality-control-0, 2022c.

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

de Gélis, I., Lefèvre, S., and Corpetti, T.: Change detection in urban point clouds: An experimental comparison with simulated 3d datasets, Remote Sensing, 13, https://doi.org/10.3390/rs13132629, 2021.

Elia, A., Balbo, S., and Boccardo, P.: A quality comparison between professional and crowdsourced data in emergency mapping for potential cooperation of the services, European Journal of Remote Sensing, 51, 572–586, https://doi.org/10.1080/22797254.2018.1460567, 2018.

Gastaminza, P.: Centre ville les cayes part1 Post image aout 2021, https://map.openaerialmap.org/#/-73.7483024597168,18.195434461776465,14/square/0322103132021330/61266ef1b7012200056b5cbb?_k=1s9qgp, 2021.

GDACS: Overall Red Earthquake alert in Haiti from 14 Aug 2021 12:29 UTC to 12:29, https://www.gdacs.org/report.aspx?eventtype=EQ&eventid=1281677, 2021.

Giovando, C.: Humanitarian OpenStreetMap Team | Rapid Mapping of Damage Extent after a Disaster, https://www.hotosm.org/updates/2017-05-03_rapid_mapping_of_damage_extent_after_a_disaster, 2017.

Grunthal, G.: European Macroseismic Scale 1998 (EMS-98), European Seismological Commission, Subcommission on Engineering Seismology, Working Group Macroseismic Scales. Conseil de l'Europe, Cahiers du Centre Europeen de Geodynamique et de Seismologie, 15, 1998.

HaitiData: Images "ortho" avant-après le séisme du 14-08-2021, https://haitidata.org/clip/ortho, 2021.

Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J. L., van Achte, T., Zeug, G., Mondardini, M. R. R., Grandoni, D., Kirsch, B., Kalas, M., Lorini, V., and Rüping, S.: E2mC: Improving Emergency Management Service Practice through Social Media and Crowdsourcing Analysis in Near Real Time, Sensors (Basel, Switzerland), 17, https://doi.org/10.3390/s17122766, 2017.

Herfort, B.: Understanding MapSwipe: Analysing Data Quality of Crowdsourced Classifications on Human Settlements, Ph.D. thesis, Heidelberg University Library, https://doi.org/10.11588/heidok.00024257, 2018.

Herfort, B., Li, H., Fendrich, S., Lautenbach, S., and Zipf, A.: Mapping Human Settlements with Higher Accuracy and Less Volunteer Efforts by Combining Crowdsourcing and Deep Learning, Remote Sensing, 11, 1799, https://doi.org/10.3390/rs11151799, 2019.

Herfort, B., Lautenbach, S., de Albuquerque, J. P., Anderson, J., and Zipf, A.: The evolution of humanitarian mapping within the OpenStreetMap community, Scientific Reports, 11, https://doi.org/10.1038/s41598-021-82404-z, 2021.

Huynh, A., Eguchi, M., Lin, A. Y.-M., and Eguchi, R.: Limitations of crowdsourcing using the EMS-98 scale in remote disaster sensing, in: 2014 IEEE Aerospace Conference, edited by IEEE, pp. 1–7, IEEE, Piscataway, NJ, https://doi.org/10.1109/AERO.2014.6836457, 2014.

Kerle, N.: Remote Sensing Based Post-Disaster Damage Mapping – Ready for a Collaborative Approach? - Earthzine, Earthzine, https://earthzine.org/remote-sensing-based-post-disaster-damage-mapping-ready-for-a-collaborative-approach/, 2011.

Kerle, N., Nex, F., Gerke, M., Duarte, D., and Vetrivel, A.: UAV-based structural damage mapping: A review, ISPRS International Journal of Geo-Information, 9, https://doi.org/10.3390/ijgi9010014, 2019.

Khajwal, A. B. and Noshadravan, A.: An uncertainty-aware framework for reliable disaster damage assessment via crowdsourcing, International Journal of Disaster Risk Reduction, 55, 102 110, https://doi.org/10.1016/j.ijdrr.2021.102110, 2021.

Kuzin, D., Isupova, O., Simmons, B. D., and Reece, S.: Disaster mapping from satellites: damage detection with crowdsourced point labels, https://arxiv.org/pdf/2111.03693, 2021.

Landis, J. R. and Koch, G. G.: The Measurement of Observer Agreement for Categorical Data, Biometrics, 33, 159–174, https://doi.org/10.2307/2529310, 1977.

Li, H., Herfort, B., Lautenbach, S., Chen, J., and Zipf, A.: Improving OpenStreetMap missing building detection using few-shot transfer learning in sub-Saharan Africa, Transactions in GIS, https://doi.org/10.1111/tgis.12941, 2022.

MapSwipe: MapSwipe | Every swipe helps put families on the map: Earthquake - Experimental Damage Assessment - Haiti (1) HOT (finished), https://mapswipe.org/en/project.html?projectId=-MhK2rqJYEKpSGMy7nVs, 2022a.

MapSwipe: MapSwipe | Every swipe helps put families on the map: Earthquake - Experimental Damage Assessment - Haiti (2) HOT (finished), https://mapswipe.org/en/project.html?projectId=-MiR5RA-m87bPmNwsbFR, 2022b.

Maso, J., Pomakis, K., and Julia, N.: OpenGIS web map tile service implementation standard, Open Geospatial Consortium Inc, pp. 4–6, 2010.

Maxar Technologies: In the blink of an eye – looking back on nine years with Tomnod, https://blog.maxar.com/leading-the-industry/2019/in-the-blink-of-an-eye-looking-back-on-nine-years-with-tomnod, 2019.

Mulder, F., Ferguson, J., Groenewegen, P., Boersma, K., and Wolbers, J.: Questioning Big Data: Crowdsourcing crisis data towards an inclusive humanitarian response, Big Data & Society, 3, 205395171666 205, https://doi.org/10.1177/2053951716662054, 2016.

Nex, F., Duarte, D., Tonolo, F. G., and Kerle, N.: Structural building damage detection with deep learning: Assessment of a state-of-the-art CNN in operational conditions, Remote Sensing, 11, https://doi.org/10.3390/rs11232765, 2019.

Notti, D., Giordan, D., Caló, F., Pepe, A., Zucca, F., and Galve, J. P.: Potential and limitations of open satellite data for flood mapping, Remote Sensing, 10, https://doi.org/10.3390/rs10111673, 2018.

OCHA: Haiti: Earthquake Situation Report No. 4 (7 September 2021) - Haiti, https://reliefweb.int/report/haiti/haiti-earthquake-situation-report-no-4-7-september-2021, 2022.

OpenStreetMap and contributors: Planet dump retrieved from https://planet.osm.org, 2017.

Pontius, R. G. and Millones, M.: Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment, International Journal of Remote Sensing, 32, 4407–4429, https://doi.org/10.1080/01431161.2011.552923, 2011.

Ranganathan, P., Pramesh, C. S., and Aggarwal, R.: Common pitfalls in statistical analysis: Measures of agreement, Perspectives in clinical research, 8, 187–191, https://doi.org/10.4103/picr.PICR_123_17, 2017.

Resch, B., Usländer, F., and Havas, C.: Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment, Cartography and Geographic Information Science, 45, 362–376, https://doi.org/10.1080/15230406.2017.1356242, 2018.

Scholz, S., Knight, P., Eckle, M., Marx, S., and Zipf, A.: Volunteered Geographic Information for Disaster Risk Reduction—The Missing Maps Approach and Its Potential within the Red Cross and Red Crescent Movement, Remote Sensing, 10, 1239, https://doi.org/10.3390/rs10081239, 2018.

Soden, R. and Palen, L.: Infrastructure in the wild: What mapping in post-earthquake Nepal reveals about infrastructural emergence, pp. 2796–2807, https://doi.org/10.1145/2858036.2858545, 2016.

Stamen Maps: Stamen Maps, http://maps.stamen.com/#toner/12/37.7706/-122.3782, 2022.

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., and Vosselman, G.: Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning, ISPRS Journal of Photogrammetry and Remote Sensing, 140, 45–59, https://doi.org/10.1016/j.isprsjprs.2017.03.001, 2018.

Watson, P. F. and Petrie, A.: Method agreement analysis: a review of correct methodology, Theriogenology, 73, 1167–1179, https://doi.org/10.1016/j.theriogenology.2010.01.003, 2010.

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

Xie, S., Duan, J., Liu, S., Dai, Q., Liu, W., Ma, Y., Guo, R., and Ma, C.: Crowdsourcing Rapid Assessment of Collapsed Buildings Early after the Earthquake Based on Aerial Remote Sensing Image: A Case Study of Yushu Earthquake, Remote Sensing, 8, 759, https://doi.org/10.3390/rs8090759, 2016.

Zahs, V., Herfort, B., Kohns, J., Ullah, T., Anders, K., Stempniewski, L., Zipf, A., and Höfle, B.: 3D point cloud-based assessment of detailed building damage through a combination of machine learning, crowdsourcing and earthquake engineering, in: EGU General Assembly Conference Abstracts, pp. EGU21–1304, 2021.

Acknowledgements

The authors would like to thank the MapSwipe community and volunteers for their inspiring work.