Department of Civil and Environmental Engineering

Stanford University

The Development and Uses of Crowdsourced Building Damage Information based on Remote-Sensing

By

Sabine Loos, Karen Barns, Gitanjali Bhattacharjee, Robert Soden, Benjamin Herfort , Melanie Eckle, Cristiano Giovando, Blake Girardot, Gregory Deierlein, Anne Kiremidjian, Jack Baker, and David Lallemant

Report No. 197
December 2018

The John A. Blume Earthquake Engineering Center was established to promote research and education in earthquake engineering. Through its activities our understanding of earthquakes and their effects on mankind's facilities and structures is improving. The Center conducts research, provides instruction, publishes reports and articles, conducts seminar and conferences, and provides financial support for students. The Center is named for Dr. John A. Blume, a well-known consulting engineer and Stanford alumnus.

Address:

The John A. Blume Earthquake Engineering Center
Department of Civil and Environmental Engineering
Stanford University
Stanford CA 94305-4020

(650) 723-4150
(650) 725-9755 (fax)
racquelh@stanford.edu
http://blume.stanford.edu

STANFORD UNIVERSITY

The Development and Uses of Crowdsourced Building Damage Information based on Remote-Sensing

Sabine Loos, Karen Barns, Gitanjali Bhattacharjee, Robert Soden, Benjamin Herfort , Melanie Eckle, Cristiano Giovando, Blake Girardot, Gregory Deierlein, Anne Kiremidjian, Jack Baker, and David Lallemant

Stanford Urban Resilience Initiative
Blume Earthquake Engineering Center

Abstract

The Development and Uses of Crowdsourced Building Damage Information based on Remote-Sensing

Crowdsourced analysis of satellite and aerial imagery has emerged as a new mechanism to assess post-disaster impact in the past decade. Compared to standard ground-based damage assessments, crowdsourcing initiatives rapidly process extensive data over a large spatial extent and can inform many important emergency response and recovery decisions. We test three approaches to crowdsourcing post-earthquake building damage using 50cm resolution satellite imagery from the 2010 Haiti earthquake. Approach 1 further develops the predominant building-level map-based assessment method that has been implemented in earlier crowdsourcing initiatives. Two novel area-based assessment approaches were also developed, where users rate the level of building damage in an image (Approach 2) and compare building damage between two images (Approach 3). The results of the two area-based approaches show a trend between crowdsourced and "true" field damage severity, which can be improved by weighting high-performing volunteers. Alternative methods, including Bayesian updating and network analysis, are proposed to analyze the paired comparison data from Approach 3. However, Approach 1 did not reach completion, because of the time intensive nature of building-level assessments.

Parallel to the crowdsourcing tests, an extensive 'demand survey' of interviews with post-disaster practitioners was conducted to develop a timeline of six key decisions that are dependent on post-earthquake building damage data. The resulting framework can guide future research concerning rapid damage estimates to address decision-makers' specific, and cross-cutting needs. Considering the results of the crowdsourcing tests and the demand survey, area-based approaches are promising methods to crowdsource building damage, because of its user-simplicity and ability to address specific post-disaster decisions.

Acknowledgements

This work would not have been possible without the support and collaboration of many interested and enthusiastic team members and volunteers.

First, thank you to our team from outside of the Stanford Urban Resilience Initiative. Benni Herfort and Melanie Eckle, we appreciate your countless hours developing and configuring our platforms for the crowdsourcing experiments. Blake Girardot and Cristiano Giovando from Humanitarian OpenStreetMap Team, for mobilizing volunteers, setting up the OpenStreetMap tasking managers and back-end, and providing useful experience and ideas for obtaining volunteered geographic information. And Keiko Saito, your excitement and feedback from a stakeholder perspective were invaluable. The opinions and expertise from all of our team members were crucial for carrying out this project, proving the benefits of interdisciplinary teams.

We thank the multiple volunteer communities who participated in our crowdsourcing tests, including the Humanitarian OpenStreetMap Team, Stanford community, Earthquake Engineering Research Institute, Southern California Earthquake Center, QuakeCore, American Society of Civil Engineers, and Structural Engineers Association of Northern California.

This research was supported by the National Science Foundation under Grant No. 1645335/EAGER "A dynamic, reliability-weighted, multi-pass probabilistic framework to reduce uncertainty in crowd-sourced post-disaster damage assessments". Satellite imagery is kindly provided by DigitalGlobe through the Open Data Program.

Contents

List of Figures

List of Tables

1 Introduction of Post-Earthquake Building Damage Information

Earthquakes radically impact the natural, social, and built environments of affected regions. From the moment the earthquake is over, knowledge on the impact to buildings, specifically, facilitates response and recovery decisions made by local governments, international agencies, and non-governmental organizations. As an example, the coordination of urban search and rescue teams in the days following an event relies on information about the locations and severity of building damage and potential casualties. Another example are the post-disaster impact analyses carried out during the response and recovery phases by the locally affected government, with the aid of the United Nations, World Bank, and European Commission. These Post-Disaster Needs Assessments (PDNA) aim at quantifying damages and losses to multiple economic sectors, with the housing sector often making up more than half of the total economic loss. Generally, these decisions are based on ad-hoc information obtained from whatever sources become available.

Numerous technological and organizational advances have made remotely sensed data rapidly available soon after an earthquake. Now, earth observation data, from satellite or aircraft sensors, are able to actually observe the regional impact of an earthquake over a large spatial scale within a relevant time-scale (Joyce 2016). Furthermore, the growth of "digital humanitarianism" has led to an increase in valuable information delivered by online communities after a crisis. Such communities, including Humanitarian OpenStreetMap Team (HOT), Standby Task Force, and Tomnod, organize after a disaster to rapidly map relevant characteristics in affected areas for on-the-ground decision-makers. The combination of remote sensing data and digital humanitarianism thereby is an extraordinary opportunity to crowdsource the assessment of earthquake-induced building damage.

As opposed to field-based assessments of damage which take extensive amounts of time to cover the entire affected region, remote-sensing based observations can be produced within a couple of weeks. The Haiti 2010 earthquake is a remarkable example of the widespread usage of satellite imagery to rapidly map induced building damage, since the public access could access very high

resolution satellite data one day after the event (G. Lemoine et al. 2013) (G. Lemoine et al. 2013). Through manual visual interpretation of this imagery, multiple private organizations and online communities seized this opportunity to produce over 2,000 maps of the damage to all of Port-au-Prince (Norman Kerle 2013). The most prominent of which is the crowdsourced building damage map produced by the Global Earth Observation Catastrophe Assessment Network (GEO-CAN) (Ghosh et al. 2011).

Even with the unprecedented amount of data after the Haiti earthquake, maps based on manual interpretation of building damage still had low accuracy. One reason is because of the evident difficulties of identifying certain types of building damage, such as soft-story collapse, from above (Corbane, Saito, et al. 2011). In the case of the Haiti 2010 earthquake, the utilization of crowdsourced damage maps remained limited due to uncertainty in the reliability of the final crowdsourced estimates (Lallemant and Kiremidjian 2015), (Westrope, Banick, and Levine 2014). This is because many post-earthquake decisions, especially in the case of damage and loss estimates in the PDNA, inform the trajectory of the affected region's recovery, and therefore need information that is sufficiently accurate.

Hence, the purpose of this study is to explore the utility of crowdsourced building damage assessments from both a user needs and data provider perspective through a two-fold approach. First, we carried out an extensive "demand survey" to gain insight on the stakeholder uses of building damage information. Second, we designed, tested, and analyzed three crowdsourcing experiments to obtain spatially aggregated damage assessments from an online crowd of non-expert evaluators.

1.1 Motivation and Objectives for Crowdsourcing Project

Given the practicality of crowdsourcing for post-earthquake building damage detection and the uncertainties surrounding the reliability of these assessments, this study is prompted from both a stakeholder application and a scientific basis. During previous examples of the usage of remote-sensing data for building damage detection, there was a division between the damage map products from data-providers and the information needed from response and recovery decision makers (Kerle and Hoffman 2013). One known application of post-earthquake damage estimation data is the Post Disaster Needs Assessment (PDNA). However, it is crucial to succinctly define the exhaustive set of potential uses of damage maps in order to provide effective decision support – information that we have insufficient knowledge about currently (Kerle and Hoffman 2013).

Therefore, we hope to address the gap between data-providers and end-users through the following objectives:

1. Identifying stakeholder needs for damage data through an extensive set of interviews and a "demand survey" of both data-providers and end-users
2. Collaborating with both Human OpenStreetMap Team (data providers) and The World Bank Global Facility for Disaster Reduction and Recovery (end-users)

In addition to understanding end-user needs of building damage information, there is also the need to improve the crowdsourcing task for obtaining damage assessments. There is significant uncertainty in the reliability of the final crowdsourced damage maps. This could be due to uncertainty linked to omissions of visible damage, the systematic over or underestimation by evaluators, or the misclassification of "damage" when viewed from above. This study endeavors to address these uncertainties by:

1. Developing various crowdsourcing tasks which reduce evaluator error, through multi-pass and area-based damage assessment approaches
2. Performing corresponding post-processing routines which acknowledge the variability in evaluator responses

Finally, current studies on crowdsourced building damage information focus on data collected from real-time applications, such as the GEO-CAN effort from the 2010 Haiti earthquake or that of the American Red Cross and HOT after Typhoon Haiyan (Ghosh et al. 2011), (Westrope, Banick, and Levine 2014). Operational scenarios do not allow data providers to test the full life cycle of crowdsourcing building damage from initial data collection to final analysis. Therefore, the design and analysis of real-time crowdsourcing approaches is constrained by the time necessary to deploy the assessments and develop final map products. Since we could carry out our crowdsourcing experiments from a purely scientific setting, we hope to address operational challenges by:

1. Testing multiple options during the life cycle of crowdsourced building damage, in terms of image pre-processing, experimental design, and post-processing statistical analysis

1.2 Project Scope

The project consisted of two principal efforts: 1) experiments to test different approaches to crowdsourcing post-disaster damage assessment and 2) a demand survey to better understand user information needs in post-disaster situations. The experimental testing effort was conducted in two

stages: a pre-experiment phase and the final experiments, while the demand survey ran in parallel.

1.2.1 Crowdsourcing Experiments

The goal of the experimental portion of this project was to develop three methods of asking for building damage information from the crowd and accordingly analyze the results relative to a ground-validation dataset from the 2010 Haiti earthquake. These three experiments included one traditional building-level approach and two area-based approaches. The experimental portion was carried out in several phases:

1. Developing and testing of user interface of three crowdsourcing experiments
2. Obtaining crowdsourcing results through deployment of experiments in full
3. Analyzing experimental crowdsourcing results compared to ground-validation dataset

The first phase was dedicated to the user experience and user interface (UI/UX) design of the experimental platform. Before launching the experiments in full, we carried out "pre-experiments", or tests of the experiment interface with a smaller sample of volunteer evaluators. In these pre-experiments, we tested numerous factors that we expected (based on a literature review) might influence how well volunteers could assess building damage in satellite images. These factors included imagery grid size, building density, varying amounts of vegetation, comparison types (combinations of urban and rural densities), and the presence of shadows.

We used the feedback from the pre-experiments to modify the user interface for the final experimental interfaces. We then deployed the experiments in full to obtain a set of crowdsourcing results for each experiment. The volunteer population included university, Humanitarian OpenStreetMap Team, and earthquake engineering communities.

The final phase consisted of analyzing the results of each experiment in comparison with a full ground-validation set of field surveys of building-level damage from after the 2010 Haiti earthquake.

1.2.2 Demand Survey

Our goal in conducting the demand survey was to develop a holistic and more comprehensive understanding of how groups involved in post-disaster activities use building damage information.

In particular, we focused on understanding how those groups use building damage information to make decisions, and what they need to consider the information actionable. The demand survey comprised two parts: a series of in-depth interviews with stakeholders involved in post-disaster activities and a short online form in which respondents were asked to react to a timeline of building damage information uses. The scope of the interviews exceeded the scope of this project, in that we gained an understanding of post-disaster activities beyond those predicated upon building damage information.

1.3 Project Team and Coordination

The scope of this project needed team members who were not only familiar with earthquake engineering and risk analysis, but also remote sensing and disaster recovery. Therefore, the project team was quite diverse - spanning across continents and disciplines.

1.3.1 Project Participants

The project team consisted of members from Stanford University's Urban Resilience Initiative (SURI), Humanitarian OpenStreetMaps Team (HOT), the World Bank's Global Facility for Disaster Reduction and Recovery (GFDRR), Heidelberg University and the University of Colorado, Boulder.

Members from Stanford University and UC Boulder led and coordinated the project through setting the initial project scope and objectives, data analysis and documentation. Dr. David Lallemant and Robert Soden supervised the Stanford team and were supported by Sabine Loos, Gitanjali Bhattacharjee and Karen Barns.

HOT leveraged their expertise and experience in mobilizing a large network of volunteers to conduct post-disaster mapping and assessments. HOT was involved in the initial study design and planning workshops, developing and maintaining the user interface for the building-by-building experiment on the OpenStreetMap tasking manager and recruitment of volunteers. The HOT team consisted of Cristiano Giovando and Blake Girardot,

Members form Heidelberg University's GIScience Research Group were also involved in the initial planning and study design phase. In addition to providing insight from their experience with crowdsourcing tasks, these members handled developing and maintaining the user interface for

the area-based assessments (pre- and final experiments 2 and 3) on the Pybossa platform. Members from Heidelberg University included Benjamin Herfort and Melanie Eckle.

Dr. Keiko Saito from the World Bank provided domain expertise and guidance from an end-user perspective, specifically regarding the required inputs for the PDNA process.

1.3.2 Project Coordination

This project was unique, because of the coordination required to maintain such a large and diverse team. The project team convened on two occasions for workshops, a major focus of which was developing the project scope and experimental design.

The first kickoff workshop was held at Stanford University in October 2016. This was attended by members from all teams. During this first workshop, we finalized the initial project scope was set and timeline.

Following the first workshop, Stanford team members designed the user-interface for the pre-experiments, coordinating with HOT and the team from Heidelberg University to discuss technical matters related to imagery and platform constraints. The pre-experiments for the two area-based approaches were launched in April for only volunteers at Stanford.

A second workshop was held at Stanford University in April 2017 which was attended by members from Stanford, Heidelberg and HOT. During this workshop, the team discussed learnings from running the pre-experiments and analyzed its results. Based on this discussion, we decided upon the final details of the final experiments. After the workshop, the Heidelberg University team modified the two pre-experiment platforms for the area-based approaches, and the HOT team created the OSM platform for the building-level approach. By June 2017, we launched the final experiments. Each sub-group of the team publicized the project within their respective communities (i.e. HOT shared the project on the HOT listserv).

The final experiments were live for approximately 4 months, with two of the three (the two area-based assessments) completed by September. The building-by-building assessment was closed at 10 % completion. The Stanford University team analyzed the results of the two area-based approaches, included in Section 6 of this report.

Stanford led the demand survey in parallel to the experiments, beginning with interviews of post-disaster practitioners, professionals and researchers to understand their roles, information needs and key decisions. HOT and GFDRR provided many contacts for these interviews. Stanford analyzed the information collected from these interviews, presented in Section 3 of this report.

1.4 Report Outline

This report reviews the process and conclusions from each major portion of this project. Chapter 2 reviews literature on 1) prior implementations of crowdsourced building damage assessments and 2) collaborative mapping suggestions which informed the designs of our crowdsourcing experiments. Chapter 3 outlines the process and results from the series of interviews and online surveys carried out with disaster response and recovery practitioners. Regarding the crowdsourcing experiments, Chapter 4 summarizes the steps taken to design the building-level and two area-based crowdsourcing tests. Chapters 6 and 7 describe the methods to analyze the responses from the two area-based approaches.

Each of the above chapters includes an extended description of their respective conclusions, which are then summarized in Chapter 8. Finally, we close the report with final thoughts on this project experience and ideas for future extensions.

2 Literature Review of Methods to Crowdsource Building Damage

Before designing the crowdsourcing tests executed in this study, it was necessary to research prior implementations of crowdsourced building damage assessments and suggested techniques on collaborative mapping. Therefore, the literature review is divided into two focus areas: earlier crowdsourced building damage implementations and studies reviewing crowdsourcing recommendations and techniques. A summary chart of the most relevant studies from the literature review, which we presented at the first workshop of this collaborative project, is included at the end of this section.

2.1 Previous Implementations of Manual Interpretation of Building Damage using Remote-Sensing Data

The use of remote-sensing technologies (satellite and aerial imagery, radar, other) for post-disaster building damage assessment has increased significantly over the past two decades. Our study focuses on the use of crowdsourcing for visual interpretation of satellite imagery for post-earthquake building damage assessments, an activity that was first operationalized at a large scale following the 2010 earthquake in Haiti. This literature review therefore focuses on this event. Other examples of such crowdsourcing have been carried out in disasters following the 2010 Haiti earthquake, including the February 2011 earthquake in Christchurch, New Zealand; Typhoon Haiyan in the Philippines in 2013, and the 2015 earthquake in Gorkha, Nepal (Westrope, Banick, and Levine 2014; Elia, Boccardo, and Balbo 2016; Foulser-Piggott et al. 2016).

The literature review chart at the end of this chapter lists multiple approaches used for assessing building damage with remote-sensing data during earlier implementations. This includes studies that relied both on crowdsourced and expert visual interpretation. The remainder of this section will elaborate on the information on previous crowdsourcing studies included in the literature

review chart: the organizations involved, the details of the tasks provided to the evaluators, the resolution of the imagery, the damage grades used to assess buildings, and outcomes or suggestions provided by the authors.

Common groups who have used remote-sensing data to assess post-disaster building damage include funding or governmental agencies (e.g. The World Bank), professional map developers or data providers (DLR-ZKI, ImageCat), raw imagery providers (DigitalGlobe), and "digital humanitarian" organizations (Humanitarian OpenStreetMap Team). Many of these organizations are also end-users of any produced building damage maps who we interviewed as part of the Demand Survey introduced in Section 3 of this report. The response from the 2010 earthquake in Haiti brought together the full diversity of organizations taking part in remote assessment of building damage, through the crowdsourcing effort carried out by over 600 trained volunteers as part of the Global Earth Observation-Catastrophe Assessment Network (GEO-CAN). This crowdsourcing implementation was led through a joint effort between the Haitian government, The World Bank, The European Commission, and the United Nations Institute for Training and Research who also commissioned ImageCat and the Rochester Institute of Technology to obtain aerial imagery and the Earthquake Engineering Research Institute to recruit volunteers (Ghosh et al. 2011).

The spatial resolution of the imagery provided to volunteer evaluators is largely related to the type of sensor used to collect imagery of an affected region after an earthquake. Most commonly, the availability of satellite imagery after a disaster has been facilitated through the International Charter "Space and Major Disasters" (UN-SPIDER 2018). Additionally, many organizations deploy manned or unmanned aerial vehicles (UAV) to obtain very high resolution (VHR) imagery. Spatial resolution is characterized in terms of centimeter specificity, which describes the size of the smallest object on the ground that is identifiable in a single pixel. For example, in a 50 cm resolution image, one pixel can depict an object that is 50 cm large. Most commercial satellite imagery providers, such as DigitalGlobe, can capture 30-50 cm resolution images, where one would be able to identify the details of a car in a parking lot from above. VHR aerial imagery can reach greater levels of detail, with a spatial resolution of 15 cm per pixel or higher. Research has shown that higher resolution imagery has greater accuracy for visual interpretation of damage than low resolution imagery, though both had relatively low accuracy rates compared to field-based assessments after the Haiti earthquake (Norman Kerle 2013; Corbane, Carrion, et al. 2011). Damage maps derived from aerial imagery provided a more accurate spatial representation of damage following the Haiti earthquake (Corbane, Carrion, et al. 2011; G. Lemoine et al. 2013). Therefore, it is expected that the quality of visually interpreted building damage will improve as higher resolution imagery becomes more routinely available in the future.

Even with higher resolution imagery, certain forms and severity levels of structural damage are not visible from nadir satellite imagery. For example, a structure that has experienced first floor soft-story collapse may seem completely intact when viewing from above, but may be completely collapsed in actuality (Foulser-Piggott et al. 2016; Ghosh et al. 2011; G. Lemoine et al. 2013). Figure 2.1 depicts a structure that was tagged as having no structural damage from the GEO-CAN initiative using satellite (nadir) imagery in Haiti, but was marked collapsed when viewing from pictometry (oblique) imagery and from the ground (Kerle and Hoffman 2013).Because of the observational difficulties of viewing building damage, oblique remote sensing data is useful for validating initial nadir-based damage assessments if it is collected (Corbane and Guido Lemoine 2011).

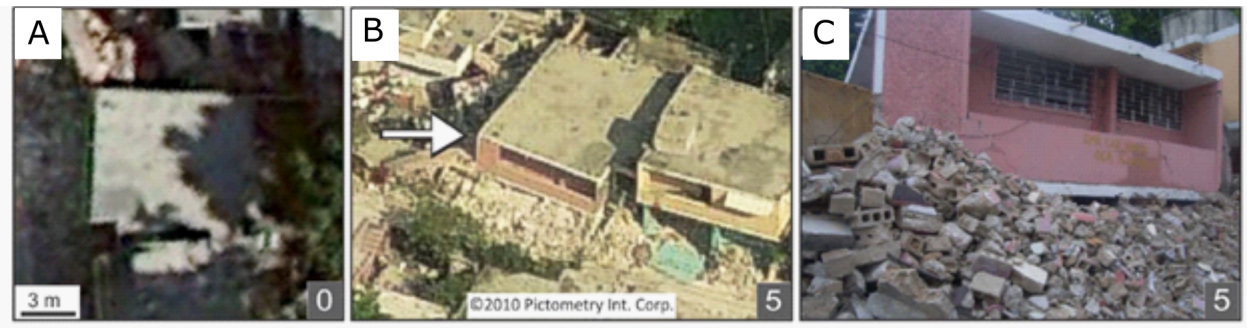


FIGURE 2.1: Example of a type of structural damage, soft-story collapse, that is invisible from above in satellite imagery (A), but can be more clearly seen in pictometry imagery (B), and from the ground (C) (Kerle and Hoffman 2013)

Given the apparent differences between damage that is visible from the ground versus remote-sensing images, the damage grades used in both cases hold different meanings and may not be directly comparable (Norman Kerle 2013). Since these remote-sensing based assessments are not precise damage assessments, but are rather proxies, or representations, of damage, damage grading schemes can vary between different crowdsourcing implementations. In fact, one of the key differences between previous crowdsourcing examples is the damage grading system that volunteers referenced to identify building damage in an image. Grading systems could be as simple as marking a binary indicator of damage/no damage or reach up to five levels of damage.

Initially, during the 2010 GEO-CAN initiative in Haiti, the typical damage grading system used for crowdsourced assessments was the European Macroseismic Scale of 1998 (EMS-98) shown in Figure 2.2 (Grünthal 1998). This is a five-level system designed for engineers to carry out in-situ structural damage assessments. Therefore, EMS-98 encompasses the full range of earthquake damage, from slight damage, such as facade cracking, to total collapse. However, volunteers for

the GEO-CAN initiative included remote sensing experts with no earthquake assessment experience, causing an inherent inability for many to understand structural characteristics they were assessing (Norman Kerle 2013). Furthermore, since the lower damage levels (Grades 1-3) of the EMS-98 scale represent types of structural damage that are inherently more difficult to view in image-based assessments, like cracking, many assessments in the 2010 Haiti earthquake focused only on heavy structural damage (Grade 4) and complete collapsed (Grade 5) (Ghosh et al. 2011; Corbane, Saito, et al. 2011). Even if the EMS-98 scale is subset to Grade 4 and Grade 5, distinguishing between very high damage and collapse still proved to be difficult when using low resolution imagery (Westrope, Banick, and Levine 2014). In fact, it was found that the EMS-98 scale provides no clear advantage over a simpler binary scale, even with higher resolution imagery (Huynh et al. 2014).

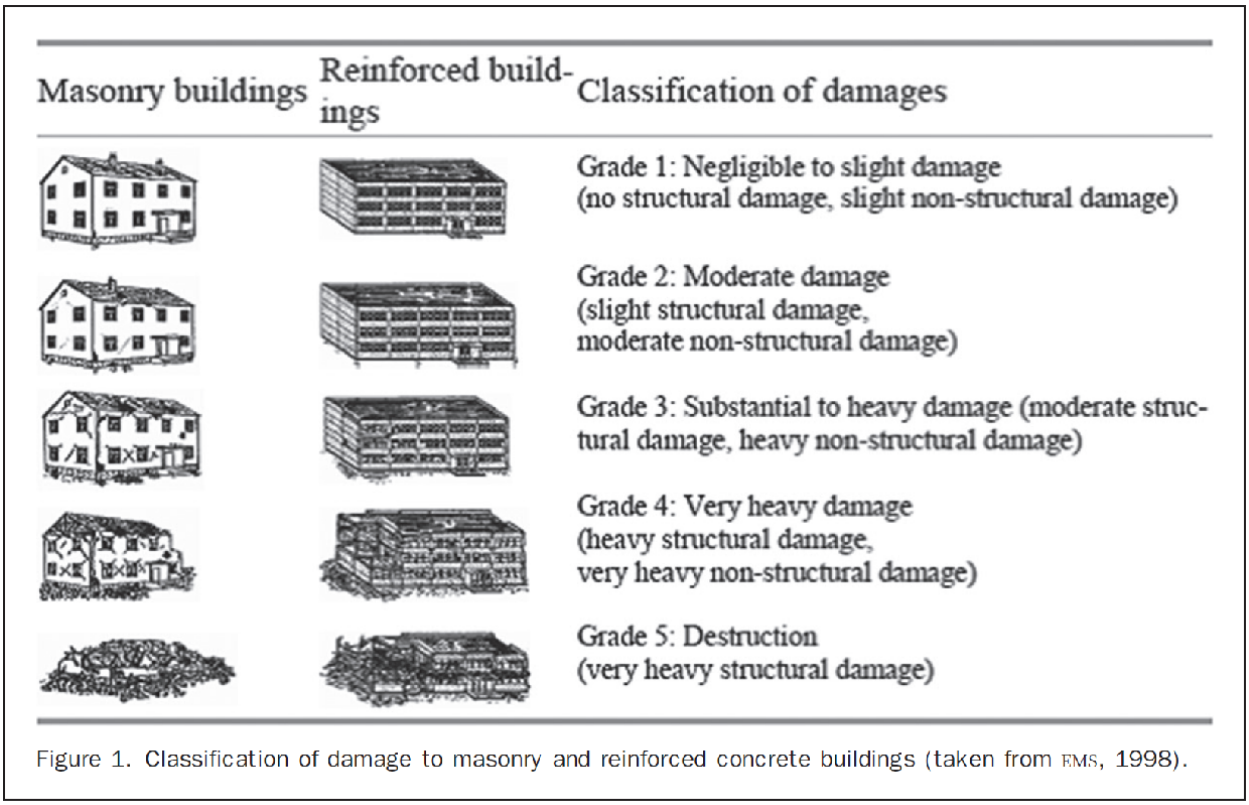


FIGURE 2.2: Damage Grading system of European Macroseismic Scale (EMS-98) which was commonly used in the 2010 Haiti GEO-CAN initiative (Ghosh et al. 2011; Grünthal 1998)

Following the issues of using the EMS-98 after the 2010 Haiti earthquake, image-based damage

assessments made use of simpler damage grading schemes. The initiative carried out by Humanitarian OpenStreetMap Team after Typhoon Haiyan in 2013 used a three-level system of "no damage", "damaged", and "destroyed" (Westrope, Banick, and Levine 2014). Similarly, the GEO-CAN initiative after the February 2011 Christchurch, NZ earthquake used a 3-level system of "substantial damage", "very heavy damage", and "complete destruction", each having associated visual attributes (Foulser-Piggott et al. 2016).

Given the plethora of options for remote-sensing data sources and damage grading schemes, numerous approaches to damage mapping exist, resulting in several maps with varying depictions of building damage. In fact, over 2,000 damage maps were produced and published on Reliefweb (www.reliefweb.int) after the 2010 Haiti earthquake, though this number includes both crowdsourced and expert manual damage interpretation (Norman Kerle 2013). These maps demonstrate the diversity of methods used for remote-sensing-based damage assessment: diversity in remote sensing technology used (e.g. optical vs radar imagery), diversity in resolution, diversity in assessment method (automated vs manual), diversity of damage measurement and interpretation (e.g. damage grades). As an example, Figure 2.3 shows how the final visualization of building damage varies greatly, ranging from individual building tags to aggregated heat maps of damage (Kerle and Hoffman 2013).

Specifically, during the 2010 GEO-CAN initiative, individual collapsed buildings were inspected using a point-based and building footprint analysis method, in Phases 1 and 2, respectively (Ghosh et al. 2011). The point-based assessment method of Phase 1 asked volunteer contributors to place a point on all the completely collapsed buildings visible in a 500m $\times$ 500m satellite image. Upon availability of newly acquired aerial imagery, the Phase 2 assessment method required each contributor to assign a damage grade to pre-event building footprints in 500m $\times$ 500m image. This approach resulted in individual buildings marked with a level of damage, which could then be aggregated into numbers of buildings damaged per an area of 500m $\times$ 500m. A similar method was implemented after the 2011 February earthquake in Christchurch by the GEO-CAN community, in collaboration with Tomnod, in which assessors were asked to demarcate the building footprints of damaged buildings (Foulser-Piggott et al. 2016). The literature review chart include other methods used in Haiti, Nepal, the Philippines, and elsewhere.

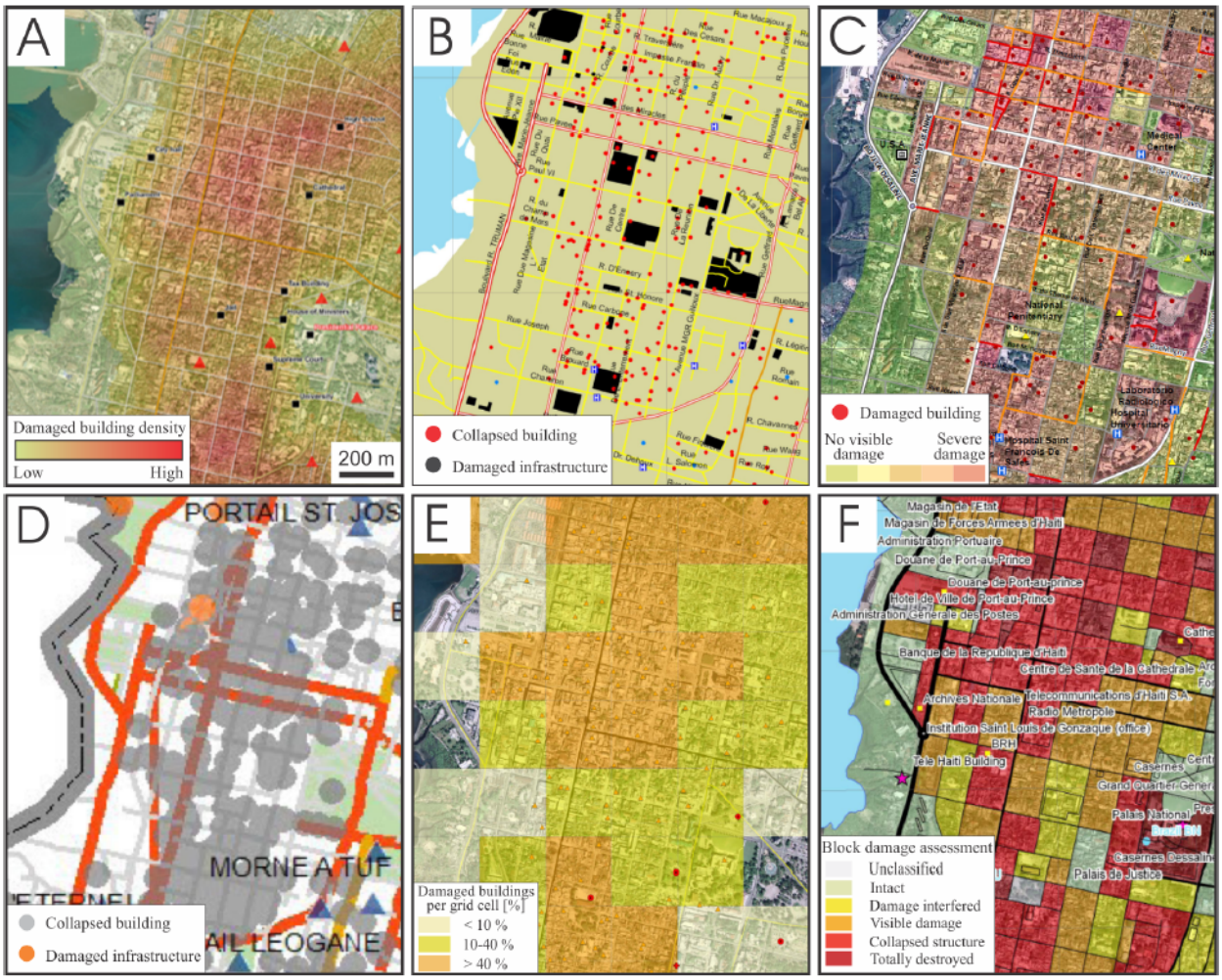


FIGURE 2.3: Examples of various forms of building damage maps in areas of Port-au-Prince derived from remote-sensing data after the 2010 Haiti earthquake. Original map sources described in (Kerle and Hoffman 2013)

2.1.1 Summary of Overarching Findings and Suggestions

The earlier implementations of remote-sensing based damage assessments resulted in specific conclusions and areas of improvement for future studies. These include suggestions for methodological improvement of the damage assessment task, recommended analysis techniques, and overarching organizational needs. This study aimed to incorporate as many of these suggestions as possible.

Many studies conclude that damage visible in remote-sensing imagery is highly different from

that visible from the ground (e.g. soft-story collapse and liquefaction-induced damage) (Foulser-Piggott et al. 2016). Related to this issue is the usage of damage grading systems geared towards field surveys of structural damage, such as the EMS-98 scale, which are not directly applicable to image-based assessments (Huynh et al. 2014). Therefore, it has been suggested that a simpler damage grading system to be used in image-based assessments, perhaps even reducing the specificity to a binary scale of "no damage" or "damage".

Few studies have addressed methods to complete real-time validation and extrapolation methods for estimating regional loss that could be useful for rapid post-disaster decisions, such as the Post Disaster Needs Assessment. Using the results from the 2010 Haiti GEO-CAN initiative, a Bayesian analysis method was proposed using a binomial prior distribution of regional building damage (Booth et al. 2011). However, the binomial distribution was found to be inaccurate for lower levels of damage, and it was suggested that a beta prior distribution may improve such inaccuracies (Lallemant and Kiremidjian 2015). In Section 7 of this study, we attempt different techniques of in-situ analyses of image-based assessments, including Bayesian updating.

Perhaps one of the greatest needs highlighted in many studies is the need for a standard operating procedure for coordination between organizations to collect and interpret this information (Voigt et al. 2011). As noted earlier, over 2,000 maps damage maps became available after the 2010 Haiti earthquake, all of which depict damage very differently. Some work has been done to develop standard operating procedures for crowdsourced assessments, including the BAR Methodology proposed by the Harvard Humanitarian Initiative and the Collaborative Spatial Assessment led by the European Commission Joint Research Centre (Al Achkar, Baker, and Raymond 2016; Corbane and Guido Lemoine 2011). However, there exists a disconnect between these map products and the needs of various end-users (Norman Kerle 2013). To the authors' knowledge, limited research has been carried out which addresses the divergence between the data produced and end-user needs. Thus, part of this study involves an extensive demand survey to understand the building damage information needs of relevant stakeholders Section 3.

In fact, for many rapid post-disaster decisions, such as the PDNA, stakeholders do not need damage information at a building-specific level, but rather over larger geographic regions. However, most of the earlier implementations of remote-sensing damage assessments focus on a building-specific approach, by observing and tagging specific buildings with a certain level of damage in an image. This is a highly time-consuming process, and the usability of area-based assessments of damage has been shown to be reasonably effective at exhibiting spatial patterns of damage, especially when assessing damage over an entire urban block (Corbane, Carrion, et al. 2011). Therefore,

the crowdsourced experiments in this study will aim to address the needs of end-users of building damage information through the design and analysis of an area-based damage assessment method.

2.2 Crowdsourcing and Collaborative Mapping Studies

Beyond researching prior approaches to manual building damage interpretation using remote-sensing, it was also necessary to understand how to optimally obtain volunteered-geographic information through crowdsourcing. The studies included in the literature review chart for this portion focus either on crowdsourcing specifically for building damage information or on general collaborative mapping for other purposes, such as for mapping roads (Albuquerque, Herfort, and Eckle 2016). This portion of the literature review informed our experimental design in terms of volunteer training and expertise, and are also included in the Literature Review Chart.

Given that volunteer contributors have varying backgrounds and skillsets, it is important to provide useful and instructive training, so they are prepared to complete a crowdsourcing task to the best of his or her ability. As expected, it has been found that volunteers are more easily able to detect higher levels of damage, such as severe damage or collapse, in a satellite image than lower levels of damage (Ghosh et al. 2011). A volunteer's ability to see highly damaged buildings is improved when they are given both positive and negative examples of building damage in a training module (Norman Kerle 2013). It has also been argued that volunteers need an explicit description of damage states, if using a damage grading system to mark specific building damage (Norman Kerle 2013). Enabling volunteers to correct their own work, plus providing a general description of the purpose of the task at hand, improves a volunteer's overall motivation to complete the crowdsourcing task (Norman Kerle 2013).

Many studies have also suggested ways to account for differences in expertise between volunteers. For example, multiple assessments (multi-pass) can be carried out for the same task and aggregated or ranked according to a volunteer's skill-level. The organization Tomnod implements a "CrowdRank" methodology to gauge the reliability of individual contributors for final estimation of a crowdsourcing task (Kerle and Hoffman 2013). Other suggestions include assigning tasks based on volunteer-indicated experience or giving contributors the ability to mark the confidence of their assessment (Ghosh et al. 2011). Collaborative mapping techniques include allowing volunteers to communicate between each other and forcing lesser experienced contributors to be accompanied by more professional counterparts during the actual crowdsourcing task (Norman Kerle 2013).

In this crowdsourcing study, selected suggestions from this portion of the literature review were incorporated into our final experimental design, volunteer training, and analysis of the volunteers' responses. This includes allowing multi-pass assessments and the development of experiment-specific training modules with visual examples of building damage, which will be discussed further in Section 4.

Specific Remote-Sensing Based Damage Assessment Implementations

Reference	Focus Area	Groups Involved	Methodology	Resolution & Damage Grades	Study Outcomes/Suggestions
Crowdsourcing for Rapid Damage Assessment: The Global Earth Observation Catastrophe Assessment Network (GEO-CAN) *Ghosh, Huyck, Greene, Gill, Bevington, Svekla, DesRoches, and Eguchi, (2011)*	Haiti EQ 2010	- GEO-CAN (ImageCat) - World Bank - UNITAR - UNOSAT - European Commission Joint Research Centre	- Multiple assessment techniques - 600 **trained** volunteers <u>Phase 1 detection</u>: - Just Grade 5 (highest) damage - **Satellite** GEO-CAN <u>Phase 2 detection</u>: - Grades 4&5 - Supplement with Aerial Imagery - Delineate building to get total floor area - Different land use types - Statistical extrapolation for lower damage states - Validate with **groundtruth** data (over 3 months)	<u>Resolution:</u> - Satellite: 50cm, 500m x 500m grid - UAV: 15cm, 500m x 500m grid <u>Damage Grades:</u> EMS-98 Scale - Grades 1-5 - For different land use types. Phase 1 and 2 focus on higher damage levels.	- **Polygon building delineation useful for estimating area of damage for specific building types** - Can have multiple "phases" of specificity - Phase 1 just assessed grade 5 - damaged or not - Necessary to have multiple assessment methodologies to assess different damage states (satellite-->aerial-->statistical-->ground data) - Satellite data omits damage detected from aerial and ground survey - Large sized data files require computing resources
Validating Assessments of Seismic Damage Made from Remote Sensing *Booth, Saito, Spence, Madabhushi, Eguchi, (2011)*	Haiti EQ 2010	- GEO-CAN - Cambridge Architectural Research Ltd - Earthquake Engineering Field Investigation Team - ImageCat	- Compared assessment data after Haiti EQ: - Remote sensing (GEO-CAN) - Pictometry (CAR) - Groundtruth (EEFIT) - Bayesian updating for damage state probability - GEO-CAN for prior dist (beta) - Update with Pictometry for likelihood fnt (beta) - Combine to get posterior and compare w/ groundtruth data (beta) - **Estimate lower damage states (D2 or D3) using knowledge of proportion of high damage states and distribution across damage states (ex. Binomial distribution)**	<u>Resolution:</u> - Vertical Imagery - 107,000 bldgs - 0.5 km2 squares - 15-25 cm resolution - Pictometry - 60 locations, 20 buildings each (1251 bldgs) - 10 cm spatial resolution - 4 orthog. directions - Groundtruth - 142 bldgs <u>Damage Grades:</u> - EMS-98, 3 Damage Levels for GEO-CAN: - D1, D4, D5 - EEFIT Ground Obs 5 levels: - D1-D5	- **Overestimation of high damage levels (D4 and D5)** - **Binomial distribution not good** for predicting lightly damaged bldgs (lower damage levels) from proportion of heavily damaged buildings - Defining damage states suggestions: - Assess needs for various applications (PDNA, civic response, and civic planning/insurance) - Set of damage descriptions for comprehensive range of building construction types - Web photos of damaged bldgs w/ damage assts, annotations, and classifications - **Bayesian analysis** is straight-forward if beta distributions are assumed - Adding Pictometry analysis reduces spread and uncertainties for commercial/downtown zone - Using ground observations could be valuable, bc removes epistemological variability and leaves only spatial variability

Comparison of Damage Assessment Maps Derived from Very High Spatial Resolution Satellite and Aerial Imagery Produced for the Haiti 2010 Earthquake
 	
Corbane, Carrion, Lemoine, Broglia, (2011)
	Haiti EQ 2010	- European Commission Joint Research Centre	- Compared area-based point-based remote-sensing damage maps produced after the Haiti EQ
			- Four already produced area-based damage maps generated by comparing pre and post-EQ imagery
			- Point-based assessment from GEO-CAN intiative
			  - Building centroids marked
			  - 300,000 labeled points
			  - Absolute damage density - number of damaged buildings per sampling unit
			  - Relative damage density – percentage of damaged of all buildings per sampling unit
			- Completed a pairwise comparison of each area-based assessment with the point-based assessment using composition and configuration analysis
				Resolution:
				- Area-based assessments
				  - 50 cm resolution
				  - Map 1: urban block area
				  - Map 2: urban block area
				  - Map 3: 250 m^2 grids
				  - Map 4: 200 m^2 grids
				Damage Grades:
				- Area-based assessments:
				  - Map 1:
				  - Map 2: Percentage of urban block showing visible damage
				  - Map 3: Increasing scale of damage grades, low to high visible damage
				  - Map 4: Number of buildings in grid showing visible damage
				- Point-based assessments:
				  - EMS-98 scale levels 1-5
					- Very high resolution satellite imagery has potential to capture overall spatial pattern of building damage
					- Urban block-based maps have greater reliability than grid-based maps in terms of damage patterns
					- Low consistency between grid-based and block-based maps
					- While satellite-derived damage assessments can capture the general damage patterns, it is difficult to interpret them accurately as overall damage estimates
					- Working on standard operating procedures and creating field sampling techniques

Earthquake Damage Assessment Based on Remote Sensing Data. The Haiti Case Study
 	
Ajmar, Boccardo, Tonolo, (2011)
	Haiti EQ 2010	- ITHACA
		- UN WFP
			- ITHACA performed 1st damage assessment few hours after eq. using
			  - experienced volunteers
			  - free software
			  - minimal technical specifications for images
			- Analysis was based on multi-temporal change detection activity between pre and post satellite data
				- Used GeoEye Satellite (50cm)
				- Updated with 15 cm when available
					- Good Quick Overview of Haiti events
					- Typical output for **WFP** staff is "cartographic product, that helps the decision makers to answer questions such as how much food aid is needed, and support the WFP staff in the field in finding the best way to deliver it to the hungry population"
					- **Important to identify features of interest for diff. stakeholders** (SAR teams need information on collapsed buildings, WFP needs to know road network accessibility)
					- Suggests Spatial Data Infrastructure to Improve Efficiency

Rapid Damage Assessment and Situation Mapping: Learning from the 2010 Haiti Earthquake	Haiti EQ 2010	- DLR/SKI	- Reviews collaboration in mapping after Haiti Earthquake and their specific damage assessment (visually inspected)
			- Focuses on collaboration between remote sensing/mapping groups
				Resolution:
				- 250 km x 250 km square grids assessed
				
				Damage Grades:
				- EMS-98 Scale
					- **Standard Operating Procedures Required**
					  - Rules of engagement
					  - Basic mapping standards
					  - Slim coordination tools
					  - Visual interpretation and qualitative damage is preferred for rapid assessment over automated

Voigt, Schneiderhan, Twele, Gahler, Stein, Mehl, (2011)					algorithms because of: interpretability, reliability, and timeliness
Groundtruthing OpenStreetMap Building Damage Assessment *Westrope, Banick, Levine,(2014)*	Philippines TC Haiyan 2013	- REACH - American Red Cross - OpenStreetMap	– Validating damage assessment from satellite imagery from OSM using field surveys conducted by REACH & ARC – Field survey specifically done to validate OSM building data - Trained, but not experienced - Used android assessment tool during field survey	<u>Damage Grades:</u> - OpenStreetMap: - 3-level (red, yellow, green) - Field: - 4-level Scale developed for Shelter Cluster Rapid Needs Assessment - No damage to completely destroyed	- **Satellite imagery resolution too low to differentiate btwn destroyed and damaged bldgs** - Use UAV to compensate - Volunteers left unclear bldgs untouched - Unclear pre-event data - Over estimation of "totally destroyed" buildings, possibly due to "media effect" - **Suggestions**: - Upgrade Java OSM Editor to include pre and post imagery interface - Satellite imagery should provide imagery to contributors within 24-48 hrs of disaster - Disaster-specific guidance materials are required (5-10 pg visual guide or youtube video) Example document in REACH final report - Validation- have OSM contributors validate assessments - Triangulate results automatically (similar to Tomnod)
Using Remote Sensing for Building Damage Assessment: GEO-CAN Study and Validation for 2011 Christchurch Earthquake *Foulser-Piggott, Spence, Eguchi, King, (2016)*	Christchurch, NZ EQ 2011	- GEO-CAN	- GEO-CAN damage evaluation using Tomnod - Pre-event **Satellite** imagery (Google Maps) and post-event satellite/**aerial** imagery validated w/ **field** surveys (FS) by NZ - <1% satellite imagery used in analysis - Draw polygon to delineate buildings using aerial imagery - Land use classified as residential and commercial	<u>Damage Grades:</u> - GEO-CAN: - Type_IDs 1-3 - Field Survey: - Green, Yellow, Red	- GEO-CAN analysts **did not mark every single building** (only damaged ones)--> omission errors (64%), some FS damaged buildings not identified in GEO-CAN - Damage levels would be underestimated using satellite imagery as compared to aerial imagery - Key factors affecting imagery: building density, contrast, tree cover, cloud cover, shadows from high rises, building use - Damage underestimated bc of invisible damage (internal, soft story, etc), liquefaction, aftershocks, poor accuracy for timber-framed buildings - **Suggestions**:

					- Create field survey with specific intent of validating GEO-CAN (same damage levels etc.) - Mark ALL buildings with NVD or damage level. Could reduce damage levels to visible and nonvisible

Visual Damage Assessment Using High- Resolution Satellite Images Following the 2003 Bam , Iran , Earthquake
 	
Saito, Spence, Foley, (2005)
	Bam, Iran EQ 2003	- Geological Survey of Iran
		- Saito
		- Spence
			- 2 experienced interpreters assessed imagery (pre vs post and just post) for damage in Bam
			  - Marked percentage of damage in each grid
			- Compared damage map results with that used by aerial photos by GSI by overlaying damage maps
				Resolution:
				- High Res Optical Images
				- Multi spectral 2.8 m resolution
				- Panchromatic 0.6 m resolution
				- 100m x 100m grids
				Damage Grades:
				- 4 damage levels based on percentage of damage (satellite)
				- 3 damage levels (20-50, 50-80, 80-100) (aerial)
					- Use of pre-event imagery makes assessment easier
					  - More grid squares are assigned higher damage levels bc can easily identify which buildings are damaged
					- Worth considering the number of buildings in each cell when interpreting damage

Current Methods And Future Advances For Rapid, Remote-Sensing-Based Wind Damage Assessment *Womble, Wood, Eguchi, (2016)*	Tornados in Birmingham, AL, Tuscaloosa ,AL and Joplin, MO	- ImageCat	- Used LiDAR and UAV to acquire 800 images	Resolution: - Baseline imagery resolution must be 25 cm or better Damage Grades: Enhanced Fujita Scale	- **Nearly 100% accuracy of completely destroyed structures if high resolution baseline imagery is available** - Key product is distribution of damaged buildings by occupancy - important for housing assistance program

Assessing Wind Disaster Damage to Structures
 	
Al Achkar, Baker, Raymond, (2016)
	Republic of Vanuatu TC Pam 2015	Harvard Humanitarian Initiative	**BAR Methodology Steps:**
			1. Set parameters w/ alphanumeric grid frame
			  a. Satellite imagery georeferenced in ArcMap
			2. Assign structure categories (light, medium, and heavy - dependent on locn)
			  a. Uploaded pre and post disaster imagery into ERDAS Imagine and manually compared
			  b. UAV imagery used to check (was not
				Resolution:
				- Very high resolution satellite imagery (Google Earth Pro)
				- UAV oblique imagery
				Damage Grades:
				- 0 = 0 no visible damage
				- 1 = 1 visible partial roof damage
				- 2 = roof sig. damaged but walls intact
				- 3 = roof and walls completely down
					- Proof of concept, not proof of accuracy
					- Get satellite imagery before reconstruction occurs
					- **Understanding cultural preferences/ construction techniques necessary to properly categorize structures**

			georeferenced, used landmarks) 3. Assign damage scale to every object a. Used markers/colors. No outlining of structures 4. Calculate point totals for each structure category/grid **Prerequisites for BAR**: - Pre-disaster imagery available (google earth ok). Should be as close to the event as possible - Volunteers have basic fluency		22

General Remote-Sensing Based Damage Assessment Studies

Reference	Focus Location	Overview	Study Outcomes/Suggestions
Limitations of crowdsourcing using the EMS-98 scale in remote disaster sensing *Huynh, Eguchi, Lin, Eguchi, (2014)*	- Japan tsunami 2011 - Wenchuan, China EQ 2008 - Haiti EQ 2010 - Christchurch, NZ EQ 2011	- Assesses validity of EMS-98 scale for crowdsourced damage assessment - Builds general platform to assess building damage from tsunami **(Disaster Response Platform)** - Delineate building - Volunteers split into 2 groups: experienced and not experienced	- **EMS-98 scale provides no clear advantage over simpler binary scale (even with high resolution images)**
Collaborative Spatial Assessment – CoSA *Corbane, Lemoine (2011)*	N/A	- Collaborative Spatial Assessment (coSA) – Standard Operating Procedures (SOP)	- SOP that is aimed at the use of remote sensing based assessments to address the recovery perspectives of the Post-Disaster Needs Assessment and the Human Recovery Needs Assessment - Defines procedural steps and who is responsible (As part of specialized team, not general crowd) - Includes sample design for field data collection
The growing role of web-based geospatial technology in disaster response and support *Kawasaki, Berman, Lex, Guan (2013)*	- Sichuan, China EQ 2008 - Haiti EQ 2010	- Overview of role of geospatial technology during specific disasters with respect to experiences in Harvard Center for Geographic Analysis	- **Provides good table of data provider/satellite/date acquired of satellite images post Haiti EQ 2010** - Suggests incorporating datasets from people on the ground (e.g. texts of images) like that provided from the Ushahidi Project [24] into remote analysis of building damage
Crowdsourcing Tools for Disaster Management: A Review of Platforms and Methods *Poblet, García-cuesta, Casanovas (2014)*	N/A	- Overview of ALL crowdsourcing tools used in disaster management cycle	- Develops taxonomy for disaster crowdsourcing tools → damage tagging falls under "Crowd as a microtasker" - **Two mobile apps that allow editing of OSM data: Pushpin and Vespucci**

Crowdsourcing Volunteer Recommendations

Reference	Focus Location	Overview	Study Outcomes/Suggestions

Collaborative damage mapping for emergency response: the role of Cognitive Systems Engineering
 	
Kerle, Hoffman (2013)
	N/A	- Review of existing collaborative mapping platforms:
		  - Proprietary (Google map maker)
		  - Open Content (OpenStreetMap)
		  - Closed systems for image analysis experts (Virtual Disaster Viewer)
		- Review of current challenges:
		  - How to convey instructions
		  - Do instructions result in useful results?
		  - Are instructions interpreted correctly?
		  - How to merge contributions
			- In the case of mapping building damage:
			  - EMS98 scale is ambiguous: designed for in situ damage assessment by structural engineers
			  - Individual damage elements (roof, facades) don't add linearly to given damage scale
			  - Must identify indicators shown in images that are related to certain types of damage
			  - Pay attention to: images with poor quality, poor resolution, locations where buildings are closely clustered, construction/demolition areas
			- Volunteer training requirements:
			  - Clear instructions
			  - Clear method: symbology, color codes, etc
			  - Easily learnable software environment
			  - Corrective feedback helpful for learning
			  - Volunteers need to know what work will be used for
			- Mappers need to be able to correct work (paralyzing if permanent)
			- Polygons for total floor space adds too much complexity
			- Merging of information options:
			  - More than 1person map a given grid cell (Tomnod and Crowdrank approach)
			  - Collaborative mapping (volunteers can communicate with each other)
			  - Less experienced volunteers are accompanied by professional
			- Apply cognitive task analysis tools to convey mapping instructions unambiguously
			  - Case walkthroughs with volunteer mappers
			  - Envisioning Desirements method to describe work features to make work easier or better
			- Redesign training method

Remote Sensing Based Post-Disaster Damage Mapping with Collaborative Methods
 	
Kerle (2013)
	Haiti 2010 EQ	- Review of Damage Mapping needs:
		  - Damage is depicted in various scales and categories
		  - Maps are not based on user needs
		  - Duplication of efforts and disagreement from multiple produced maps
		  - Traditional maps remain static
		  - Validation rarely takes place
		- Review of volunteer training needs in collaborative damage mapping
			- To consider for volunteer training:
			  - How volunteer understood instructions and examples
			  - Image analysis expertise
			  - Learning process during mapping task
			  - Time allocated to mapping
			- Training ambiguity cannot be reduced by expanding instruction manual, but should use corrective feedback during mapping activity
			- Expert collaboration needed

Using Remote Sensing for Building Damage	Christchurch, NZ EQ 2011	- 200 volunteers divided into 3 groups based on self-assessment of experience	- **Low experience volunteer delineates blocks of buildings instead of individual buildings → less accurate analyses**

Assessment: GEO-CAN Study and Validation for 2011 Christchurch Earthquake *Foulser-Piggott, Spence, Eguchi, King (2016)*		- Compared assessments between high and low level of experience	- More low experience volunteers needed to do the same level of assessments as high experience - Low experience can delineate same building multiple times - **Damage training material should incorporate multiple construction types (i.e. masonry, timber, RC) and different urban environments (commercial vs. residential)**

The Tasks of the Crowd: A Typology of Tasks in Geographic Information Crowdsourcing and a Case Study in Humanitarian Mapping
 	
Albuquerque, Herfort, Eckle (2016)
	South Kivu in Democratic Republic of Congo	–Used PyBossa crowdsourcing tool with satellite imagery input to assess whether areas are inhabited (Classification)
		  - Binary: yes or no
		- PyBossa Crowdsourcing framework:
		  - Bing satellite imagery
		  - Validation using OSM data
			- Easier to assess longer roads and larger settlements (larger objects)
			- **Suggestions:**
			  - Reduce error with volunteer training
			  - Don't divide area too small. Allow volunteer to see larger area (zoom in and out), gives clues to assessment by surrounding context clues.
			- **20% of highly skilled volunteers produced 80% of the assessments**
			  - Design improve interaction mechanism (mapswipe app) to reach more volunteers and motivate using badges

A Quality Comparison Between Expert and Crowdsourced Data in Emergency Mapping for a Potential Service Integration
 	
Elia, Boccardo, Balbo (2016)
	Nepal 2015 EQ	- Focus on Emergency Mapping
		- Tomnod General Approach:
		  - Crowdrank computer algorithm --> shows tags w/ most consensus
		  - Most "reliable" data (with most agreement) goes thru QC w/ expert analysts
		- HOT Nepal Building Damage Approach:
		  - Trace individual bldg and tag w/ bldg=yes
		  - Do NOT trace entire areas of bldgs, must be individual
		  - Compare OSM data to Copernicus data
		  - Comparable when bldg was delineated, but OSM had omission error (55373 bldgs delineated in OSM vs 74899 bldgs in Copernicus)
		- Infrastructure damage approach (by bldg block)
		  - HOT traced polygon around block of damaged houses w/ no damage levels
		  - Copernicus delineation maps
		  - Large overestimation of damaged areas by HOT compared to copernicus bc of poor lighting in images and very dated pre-event imagery
			- **Create relationship between expert and volunteer analysts:**
			  - OSM requires comprehensive validation process (see Tomnod Approach left)
			  - Perform QC with same AOI analyzed by expert imagery (Copernicus) and crowdsourced imagery (OSM)
			- Increase image brightness for OSM contributors by "acting on displayed brightness values on a reduced radiometric resolution of 8 bit"
			- **Implement "Image Analysis Toolbox" in OSM interface** so contributors can improve image quality (ex. Contrast enhancement, sharpen, etc.)
			- Provide more guidance to contributors that is continuously visible during mapping
			  - Include overview of local construction types

3 Demand Survey

In the aftermath of a disaster, numerous and complex processes take place as decision-makers move rapidly to gather, exchange, and act upon information. The set of processes includes assessments of the economic losses to various sectors, whether industrial, infrastructural, or social. Building damage cuts across sectors, and factors into estimates of both direct and indirect losses. Thus, many of the decisions made in the aftermath of a disaster have to do with the built environment, and buildings in particular. Understanding the damage to the housing and commercial building stock in an area affected by disaster is of importance to a variety of stakeholders, from international search and rescue teams to local housing authorities. Therefore, various post-disaster building damage assessments may be carried out at different spatial scales, precisions, and times.

With continuing advances in imaging technology and classification algorithms, the methods by which interested parties may conduct remote building damage assessments have multiplied. While previous assessments have relied on building-by-building field surveys, novel methods include using change-detection algorithms on various types of imagery. Advances in imaging and damage identification are undoubtedly valuable. However, previous work has indicated the need for damage assessment methodologies that account for the diverse information needs of end-users (Lallemant, Soden, et al. 2017). That is, the information produced by any damage assessment should be produced and disseminated in ways that serve the needs of its users. This is of particular importance in post-disaster contexts, in which decision-making processes may be accelerated and time for thorough review or interpretation of information is minimal.

In an effort to better understand the needs of end-users, we conducted a two-part survey of post-disaster building damage information needs. Given the highly complex nature of post-disaster information flows, we constrained our focus in this study to user needs for building damage information in the aftermath of a sudden-onset disaster. Most of the disasters cited by our interviewees were earthquakes.

3.1 Intent

The goal of this survey was to first propose a user needs-based framework for thinking about post-disaster damage information and to then test its utility with regard to:

- informing academicians, researchers, and other data producers by identifying unmet information needs;
- identifying information needs common to different decision-makers or stakeholders, thus indicating areas in which improvements would have high impact;
- contributing toward a holistic understanding, shared by data users and providers, of what information is needed when, and at what precision;
- reframing the production of damage information by focusing on user needs, rather than technical or technological capabilities.

We therefore focus on identifying post-earthquake decisions that rely on information about building damage, the qualities required for that information to be an actionable basis for decision-making, and the decision-makers involved. By considering decisions, we simplify information flows and center the end-user.

3.2 Process

3.2.1 Previous Work on Post-Disaster Information Needs

While some post-disaster processes have been studied in the literature, documentation of specific decision-makers' information needs in the aftermath of a disaster remains comparatively limited (Gralla, Goentzel, and Van de Walle 2013). Research in information systems has indicated that a shared understanding of how individual organizations within the disaster response community operate would positively impact coordination within that community (Bharosa, Lee, and Janssen 2010). Effective coordination after a disaster has, in turn, long been understood to underpin effective emergency response(Bharosa, Lee, and Janssen 2010; Chen et al. 2008). We propose that elucidating decision-makers' information needs is crucial to broader operational goals, since decision-makers within an organization rely on information to act.

In 2013, the Digital Humanitarian Network published a *Report From the Workshop on Field-Based Decision-Makers' Information Needs* , which served as a starting point for our study (Gralla, Goentzel,

and Van de Walle 2013). This report highlights the vastness of the scope of any general inquiry into post-disaster information needs. Consideration of the breadth of the post-disaster information needs space, the particular expertise of the research team, and the recent call for more user-centered post-disaster damage assessments prompted the limitation of our current scope to post-disaster building information needs in particular (Lallemant, Soden, et al. 2017).

3.2.2 First Draft of a Framework

Based on conversations in the first workshop, described in Section 1.3.2, we theorized that two qualities of the data required by each stakeholder would determine its utility to that stakeholder: (1) the time at which it was available to the stakeholder and (2) the spatial precision to which it was produced. Thus, our initial concept of the demand survey involved better understanding when stakeholders needed particular types of information and the spatial precision at which they considered that information actionable, i.e. usable as a basis for decisions.

Figure 3.1 represents our first draft of a framework in which to situate stakeholders' information needs. We proposed that shaded circles should indicate individual datasets, and that the positioning of those circles should indicate the time and the minimum spatial precision at which a stakeholder required that dataset to be reported for it to be actionable. As shown in Figure 3.2, we hypothesized that understanding the information needs of multiple stakeholders could prove useful in identifying areas in which any improvement could result in maximal gains. We also hypothesized that by clearly laying out stakeholders' information needs, unmet needs would become obvious and could inform the work of academicians, researchers, and data producers. These hypotheses became two of the four principal objectives of this study, as noted in Section 3.1.

FIGURE 3.1: The first draft of the proposed output of this demand survey. This figure charts the information needs of a single stakeholder in a two-dimensional space defined by the time since the disaster and the minimum spatial precision to which the data should be reported in order for the stakeholder to consider it actionable.

This draft framework differs in three principal ways from our final output, described in Section 3.4. First, in this framework, we do not contextualize user information needs - that is, we offer no insight into how datasets underlie or inform the particular decisions stakeholders make. In the process of conducting interviews, it quickly became apparent that the volume of information exchanged in the aftermath of a disaster would make cataloging individual datasets onerous. In addition, we questioned the longevity of such an analysis, given the changing ways in which datasets are produced. Second, this draft framework treats each stakeholder separately. However, it became apparent through our interviews that close collaboration between government agencies and non-governmental organizations in the aftermath of major disasters makes such delineation difficult, and rather uninformative. Finally, this draft framework doesn't clearly communicate that information needs often persist, i.e. are not instantaneous.

To address these shortcomings, we designed our final framework to focus explicitly on decisions – rather than stakeholders or datasets – to account for the highly collaborative nature of post-disaster processes.

FIGURE 3.2: The first draft of one proposed output of this demand survey. This figure charts the information needs of multiple stakeholders.

3.3 Survey Design

In the first part of the survey, we interviewed 11 members of the disaster-oriented community. These practitioners include both data providers and data users, and work within organizations that are active in disaster preparedness as well as response, recovery, and reconstruction. The interviews ranged widely within the theme of post-disaster processes, often addressing questions beyond how decision-makers use building damage information. Questions included how information is produced, in what formats information may be communicated, who is involved in the production of information, whether information is updated, and what level of confidence end-users require to consider information actionable.

Based on the interviews, we identified six post-earthquake decisions that rely on building damage information. In this context, a decision may refer to the outcome of a process upon which subsequent processes rely – such as a needs assessment – or a group of similar and repeated decisions. We characterized each decision by (1) the stakeholders involved in making the decision (2) the time period during which the decision is made (3) the information upon which the decision is

based (4) the format in which that information is needed in order for it to be actionable and (5) the minimum level of spatial precision (e.g. building-, block-, city-, or region-level) at which the stakeholders consider that information actionable.

In the second part of the survey, we solicited comments on the accuracy of a version of Figure 3.3 from members of the broader disaster-oriented community, including academics. Eleven respondents submitted their comments on the figure through an online form, in which they were required to respond to three questions: (1) What other moments or activities would you include on the chart? (2) Are the activities on the chart characterized accurately? If not, please describe the changes you would make. (3) What other feedback or suggestions would you share?

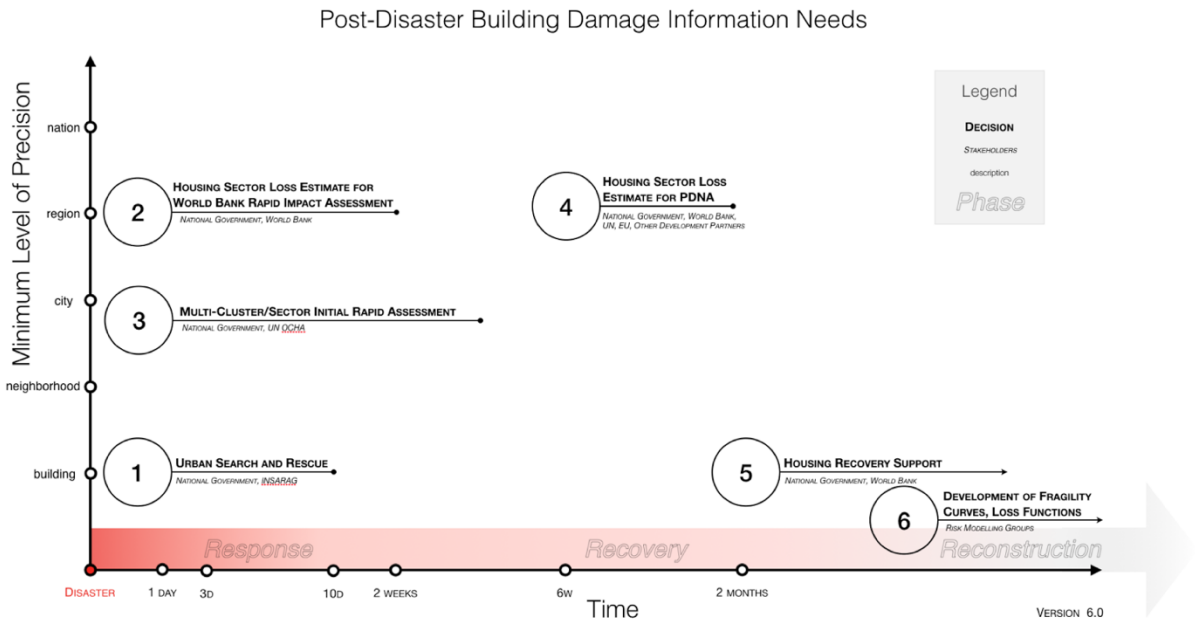


FIGURE 3.3: A case study of the building damage information needs for six key decisions made in the aftermath of a sudden-onset disaster.

3.4 Results

We focus on identifying post-earthquake decisions that rely on building damage information, the qualities needed for that information to be an actionable basis for decision-making, and the decision-makers involved.

We contextualize six decisions using a novel framework that describes their timing and the minimum level of spatial precision required of the information upon which they are based. We present these decisions graphically in Figure 3.3 and also summarize them in Table 3.1. We describe these decisions in detail as follows:

1. Urban search and rescue (USAR) operations locate and rescue live victims of building collapse, in coordination with the affected country's government. The International Search and Rescue Advisory Group (INSARAG) is a secretariat of the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) and facilitates coordisnation between international USAR teams, who may deploy within six hours of a disaster. USAR operations may continue up to ten days post-disaster and require damage information at the individual building level, in the form of maps and geotagged imagery (drone, aerial, and/or satellite). USAR teams focus on buildings with high population density and of particular structural types that are conducive to collapse survival. Knowledge of building inventory and local construction details is important to establish likelihood of survival and inform rescue operations.

2. The World Bank Rapid Impact Assessment is a coarse internal assessment of the direct or indirect losses to a subset of the affected country's economic sectors. Sectors typically included are housing, transport, agriculture, schools, and hospitals, as they generally account for 60-70% of the total losses in a disaster. The assessment is put together between one day and two weeks post-disaster. Losses are estimated regionally; building information in any format – including building types and distributions, construction costs, and damage statistics – is needed at the same level of spatial precision.

3. The Multi-Cluster/Sector Initial Rapid Assessment (MIRA) focuses on identifying and creating a common understanding of humanitarian needs after a disaster. The national government of the affected country leads the MIRA, in conjunction with the UN Resident/Humanitarian Coordinator. It has two principal outputs: a Preliminary Scenario Definition (PSD), produced within 72 hours of the disaster, and a MIRA Report, produced within two weeks of the disaster. The MIRA requires maps of housing damage at the neighborhood level, and the types and locations of shelters for displaced people.

4. The Post-Disaster Needs Assessment (PDNA) is led by the affected country's government, in conjunction with the World Bank Global Facility for Disaster Risk Reduction (GFDRR), United Nations (UN), European Union (EU), and development partners. Assembled between two weeks and three months after the disaster, the PDNA summarizes the disaster's

impacts and resulting financial needs, and is used a basis for recovery/reconstruction planning and requests for external assistance. The PDNA requires building damage information at the regional level, including damage to different building types, construction and replacement costs, and maps and/or tables of building damage.

5. During recovery and reconstruction, the national government of the affected country and/or the World Bank may offer housing recovery support to property owners. Such support requires damage information at the individual building level.

6. In the longer-term, the development of fragility curves and loss functions is one way in which risk modeling and disaster risk financing groups use individual building damage data. This is more common in areas with unique or less well-documented building types.

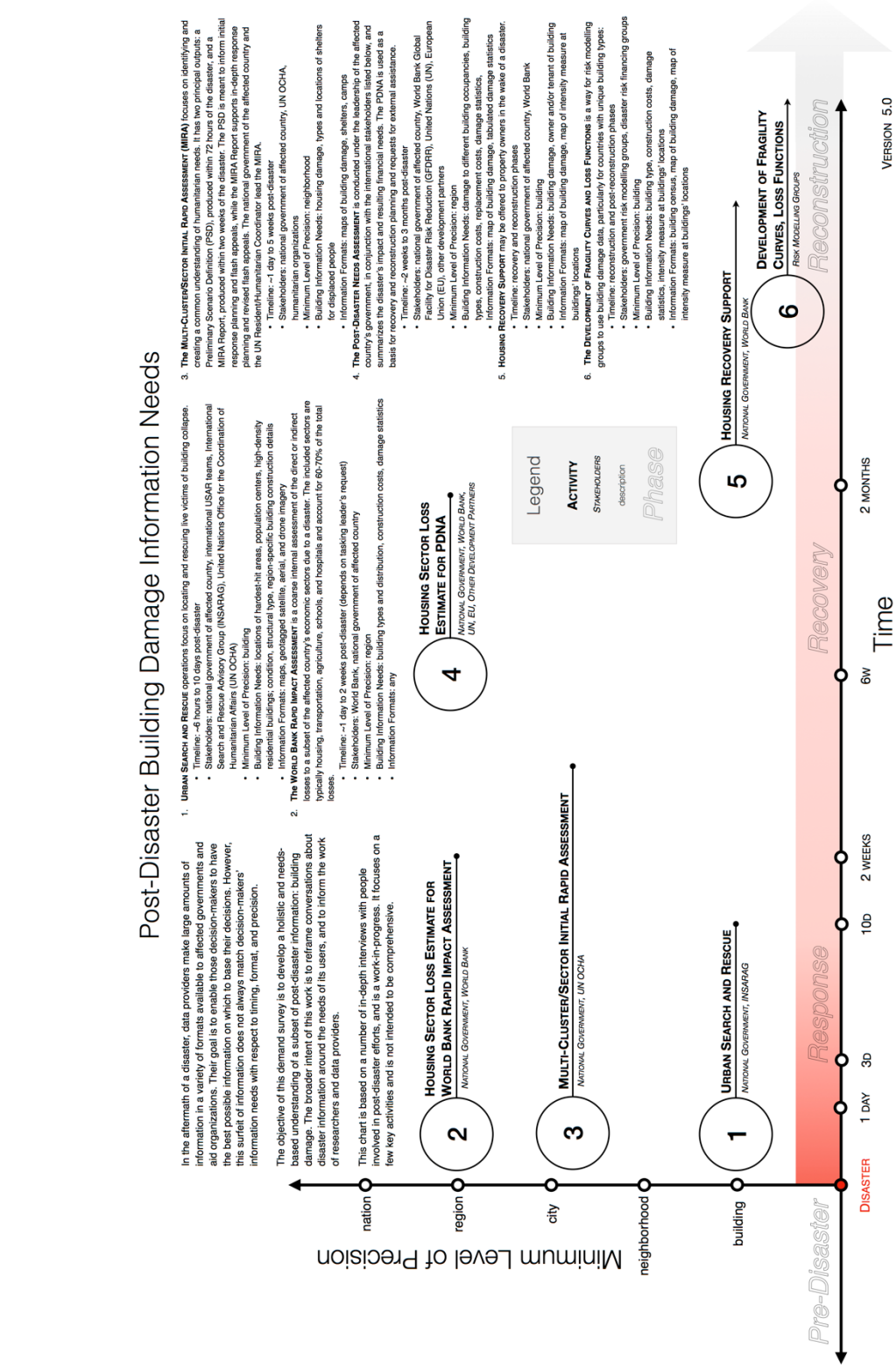


FIGURE 3.4: Six post-disaster decisions that rely upon building damage information, situated in our framework according to the earliest time post-disaster at which the decision is made and the minimum level of spatial precision to which the building damage information should be reported to be considered actionable by the decision-makers involved.

TABLE 3.1: Summary of building damage information needs for six post-disaster decisions.

Decision	**Decision-Makers**	**Building Damage Information Needs**	**Formats**
Urban search and rescue (USAR)	International Search and Rescue Advisory Group (INSARAG) international USAR teams national government of affected country	Damaged high-density buildings Region-specific construction details	Maps Geotagged imagery
Housing sector loss estimate for World Bank Rapid Impact Assessment	World Bank national government of the affected country	Building types Building type distribution Damage statistics	Any
Multi-cluster/sector Initial Rapid Assessment (MIRA)	United Nations Office for the Coordination of Humanitarian Affairs (OCHA) national government of affected country	Housing damage Types and locations of shelters	Maps
Housing sector loss estimate for Post-Disaster Needs Assessment (PDNA)	World Bank United Nations European Union national government of affected country	Damage to building types Damage to building occupancies Construction and replacement costs Damage statistics	Maps Tabulated damage statistics
Housing recovery support	national government of affected country World Bank	Building damage Building owner and/or tenant Intensity measure at building locations	Maps
Development of fragility curves, loss functions	Risk modelling groups	Damage to building types Damage statistics Intensity measure at building locations	Building census Maps

3.5 Conclusions

The application of our proposed framework to the post-disaster decisions described in the previous section meets at least three of the objectives of this study, specifically (1) informing academicians, researchers, and other data producers by identifying unmet information needs; (2) identifying information needs common to different decision-makers or stakeholders, thus indicating areas in which improvements would have high impact; and (4) reframing the production of damage information by focusing on user needs, rather than technical or technological capabilities. Whether this work also contributes toward a holistic understanding, shared by data users and

providers, of what information is needed when, and at what precision (the third objective listed in Section 3.1) remains to be seen.

We note that a comprehensive and detailed pre-disaster building census is one of the principal unmet information needs from which many of the decisions included in this study could benefit. The level of detail required from such a building census will differ by the particular decision and the decision-makers' involved. However, including information about the structural system of buildings within any such census would prove useful to predicting building damage, and would serve the information needs of all the decision-makers included in this study, though to different extents.

Our framework highlights that what we refer to as decisions are more often collaborative processes, or groups of highly related decisions, that unfold over particular segments of the post-disaster timeline. Figure 3.3 not only highlights the timing of these decisions but also suggests how information gathered or used by one organization could benefit decision-makers who are concurrently or subsequently active. For example, our results suggest that loss estimators could benefit from urban search and rescue teams collecting and sharing detailed damage information about the buildings they inspect – such information would be more detailed than damage estimates available from image-based surveys, and would become available prior to the results of building-by-building surveys. Since USAR teams typically focus on densely populated areas, the damage information they might collect could inform the work of teams assessing housing losses, setting up temporary shelters, and working to support housing recovery.

By mapping decisions according to the critical features of the information upon which they rely, our framework highlights concrete opportunities for information-sharing. Identifying these very granular opportunities for inter-agency coordination accords with the cluster approach already endorsed by the United Nations Inter-Agency Standing Committee (IASC), adopted in the United Nations Disaster Assessment and Coordination Field Handbook, and overseen by the United Nations Office for the Coordination of Humanitarian Affairs (The United Nations Office for the Coordination of Humanitarian Affairs 2013). These mutual information needs suggest not only benefits from improved information-sharing and coordination but also reinforce the need for rapid, reliable, and secure information-sharing systems already noted in the literature (Meissner et al. 2002).

3.6 Future Work

The 11 in-depth interviews conducted in the first part of this survey contained information well beyond the scope of the results discussed above. Immediate future work will include a full qualitative analysis of the post-disaster information flows (including and beyond those centered on building damage) described in those interviews.

Given the many and complex post-disaster information flows and stakeholders involved in response/recovery/reconstruction, we anticipate that applying the same framework used to contextualize the results presented above may prove useful to better understanding other post-disaster information needs and highlighting opportunities for closer coordination and information-sharing. Specifically, mapping decisions by the critical features of the information upon which they rely – in this case, the time and minimum spatial precision at which the information is needed – may be a sufficiently flexible and relatively straightforward methodology to apply to many other information problems.

This study highlights the benefits of collaborative, multi-disciplinary work that brings practitioners and researchers together. This effort in particular included researchers with expertise in structural engineering and firsthand experience with post-disaster building damage assessments, as well as practitioners with experience in producing building damage data, using it to inform a wide array of decisions, and coordinating post-disaster information-sharing and response/recovery/reconstruction efforts. This backgrounds of this team informed our decision to focus on building damage information needs; teams with other relevant expertise could conduct similar surveys on other information needs identified as important to decision-makers, e.g. existing laws and policies in the affected area, the status of telecommunications or other infrastructure systems, the capacity of existing systems to respond, and the legibility of information to end-users (Gralla, Goentzel, and Van de Walle 2013; Meissner et al. 2002).

4 Design of Crowdsourcing Experiments

This section details the three experiments developed in the study: one building-by-building approach and two area-based approaches. The study involved two rounds of experiments, first a "pre-experiment" to test the novel area-based approaches, followed by the final experiments of all three approaches. The outreach and recruitment process is also detailed in this section.

Each experiment used satellite imagery obtained from DigitalGlobe's OpenData program. This imagery is of 50cm resolution and was made available in the days after the 2010 earthquake through the International Charter "Space and Major Disasters" (UN-SPIDER 2018). Overall, this imagery covered the affected areas of Port-au-Prince and its surrounding neighborhoods.

Each experiment presented users a satellite image containing a 125 $\times$ 125m section of land within a defined area of interest, described in Section 5. The first experiment was a building-by-building approach where users were asked to record the observed damage level to each individual *building* in the image. This was referred to as experiment 1 and was implemented on an OpenStreetMap tasking manager. In contrast, experiments 2 and 3 were area-based approaches where users were asked to evaluate the overall level of damage for an entire *image*. In experiment 2, users were asked to rate the damage shown in an image on a scale of 1-5. In experiment 3, users were asked to compare two images and select the image that shows more damage. Both experiments 2 and 3 were implemented on the crowdsourcing platform Pybossa.

4.1 Pre-experiments

The first round of experiments completed in the study were the pre-experiments for the area-based approaches (experiments 2 and 3), both implemented in the Pybossa crowdsourcing platform. The pre-experiments were carried out only for the area-based approaches because of their relative novelty, so the research team had no prior experience with the two area-based approaches. This stage made use of a subset of approximately 100 images.

The goals for the pre-experiments included:

- Reducing the number of variables tested in the final experiment
- Testing the user interface to make the final experiment as user friendly as possible
- Identifying any technical or back-end issues

The pre-experiments tested several different variables. One variable common to both pre-experiments was the grid size, or size of the image shown to volunteers. During the pre-experiments, two different grid sizes were tested: 125 $\times$ 125m and 250 $\times$ 250m, corresponding to optical zoom level 17 and 18, respectively. The physical size of the image on a computer screen is the same for both zoom levels, meaning the 250 $\times$ 250m grid showed a more zoomed out image compared to the 125 $\times$ 125m grid. Volunteers during the pre-experiment indicated that it was more difficult to evaluate building damage from the 250 $\times$ 250m images due to the quality of the satellite imagery. Therefore, the 125 $\times$ 125m grid system was chosen for the final experiments, since it provided the user with better detail.

The pre-experiments provided volunteers with a draft of the training material as a PDF which they could access via a link on the pre-experiment pages. It was apparent that not all users realized the training material was available or took the time to read it. Therefore, the final experiments incorporated the training material in the platform itself as a click through tutorial before the volunteer could complete the crowdsourcing task in addition to an reference link accessible throughout the experiment.

While grid-size was one variable common to both pre-experiments, the specific methodologies for each presented their own challenges. The next two sections outline the unique variables tested for the pre-experiments for both area-based approaches.

4.1.1 Pre-experiment 2: Damage Rating

In pre-experiment 2, the damage rating approach, users were asked to choose the level of damage shown in each image on a scale of 0 (corresponding to no damage) to 100 (corresponding to complete damage). A reference scale with images depicting different levels of damage, in an urban context, was provided as a guide, as shown in 4.1. This approach also included a pre-disaster image to compare with the image to be assessed. The user entered their response by either moving the slider bar shown beneath the image or by directly entering the number in a box. Pre-experiment 2 consisted of a total of 32 images to be assessed, using the full range of test variables.

FIGURE 4.1: Interface of pre-experiment 2 in the Pybossa platform. A 5-image, damage reference scale is shown at the top. The image to be assessed is shown on the right, while the pre-disaster image is on the left. Users enter their response by typing a number in the box next to the save button or moving the slider below the target image

Reference Scale

Various reference scales of damage were tested (shown at the top of 4.1). Firstly, a scale using urban (high building density) images was compared to a scale showing rural (low density) images. Volunteers indicated that it was very difficult to compare the level of damage between images from different contexts, i.e. assessing an urban image against a rural reference scale was difficult and vice-versa.

Secondly, using only urban images, the pre-experiment tested the granularity of the reference scale. Users were presented with two different reference scales, one showing 5 levels of damage and the second showing 10 levels of damage. The goal with testing the two scales was to determine if the benefit gained from the increased level of detail warranted the more complicated task for

the user. The pre-experiments results were inconclusive for this test, so it was decided that both the 5 scale and 10 scale would be carried forward for the final round of experiments.

Even when the context was the same, the images shown in the reference scale showed only one example of each damage level, when in fact there may be many. For example, a damage level of 50 out of 100 could correspond to an image with 100 buildings, where half are completely destroyed and half are undamaged. Or, it could be that all 100 buildings have sustained moderate damage. The images for these two scenarios would look quite different. For this reason, it was decided that providing the user with one reference scale was misleading, so the images were removed and replaced with buttons on a numerical scale for the final experiment.

The extent of damage referred to each level of the damage scale was kept deliberately vague. The team discussed different ways to measure damage, such as providing the percentage of buildings collapsed, or the percentage of buildings showing any damage. Ultimately, we decided that providing explicit damage descriptions was too prescriptive, since the experiment is trying to capture a qualitative assessment of damage based on a user's perception. We also wanted to avoid the situation where a user was counting the number of buildings in an image, since this would resemble a traditional building-by-building approach and would increase the amount of time necessary to complete the crowdsourcing task. Therefore, users were provided with examples of what each damage level may refer to, but it was left open to user interpretation. The meanings of the values in the reference scale could then be associated with actual damage levels in the post-processing analysis of the results, described in Sections 6 and 7.

Pre-event Imagery

Another variable tested during pre-experiment 2 was the provision of pre-event imagery. Figure 4.1 shows an example where pre-event imagery was provided. Recognizing that recent pre-event imagery may not always be available, the goal was to determine if this additional information could improve user performance. User feedback from the pre-experiment indicated that being able to reference pre-event imagery was useful, so it was kept in the final experiments 1 and 2. Having two large images made the interface overly cluttered, though, so the final experiment made pre-event imagery available through a button which reveal it if the volunteer wanted.

User Interface

Volunteers in the pre-experiment indicated that the slider bar and number system was difficult to use. The slider bar was replaced in the final experiment, by instead having volunteers record their response through clicking on one of the buttons in the reference scale.

4.1.2 Pre-experiment 3: Damage Comparison

In pre-experiment 3, the damage comparison approach, volunteers were asked to compare two images, clicking on the image showing the higher level of damage. They were also provided with two further options, a "not sure" button as well as a "same damage" button. An example of the Pybossa interface is shown in Figure 4.2.

FIGURE 4.2: Interface of pre-experiment 3 in the Pybossa platform. Volunteers could click on the image showing a higher level of damage, choose the "same damage in both images", or "not sure"

Pre-experiment 3 consisted of 36 different comparisons covering the various test variables. These comparisons were selected to test how user performance is affected by comparing the following:

- Two adjacent images versus two non-adjacent images
- Two images from different contexts versus the same context, i.e. a rural image with an urban image versus two urban images

- Two images with similar versus very different damage levels (e.g. two images showing moderate damage versus one with no damage and one with severe damage)

Image Adjacency and Building Density Comparisons

Results from pre-experiment 2 indicated that comparing adjacent tiles may provide some useful additional context. This is most likely because adjacent tiles often exhibit similar building densities. For example, if one image shows an urban area, the adjacent image is also likely to be of an urban area. Similar to the findings from pre-experiment 2, comparing images of the same context was considered less difficult than comparing images from different contexts. Prioritizing adjacent images for comparison was considered but ultimately not carried forward to the final experiment, because the time required to implement versus random comparisons outweighed the benefit.

Damage Level Comparisons

The results also showed that there was no significant difference in user performance when comparing images with similar damage levels versus images with very different damage levels. However, the pre-experiment results showed that users tended to assess images with more buildings as showing higher levels of damage. Since the actual damage levels would not be known in a live scenario, no effort was made to order comparisons based on known damage levels.

User Interface

After examining the results from pre-experiment 3 responses, it was decided that the "not sure" option would be removed from the final experiment. If a user is unsure, this is likely because there is no discernable difference in visible damage and therefore the "same" button would capture this. Removing this option also simplified the resulting dataset.

4.2 Final Experiments

The final versions of experiments 2 and 3 on the Pybossa platform incorporated the findings from their respective pre-experiments, while the HOT team set up experiment 1 using the OpenStreetMap tasking manager. In addition, a survey was appended to experiments 2 and 3, to obtain information about each volunteer's experience and background. The survey asked the volunteer

for information about their occupation and if they had any previous experience with satellite imagery or assessing building damage. Completing the survey was optional, and it appeared after the user completed 15 questions, placed here to not deter users from beginning the experiments. Unfortunately, the survey completion rate was very low and not used in the analysis of the crowdsourcing results; it may have been better to incorporate it at the beginning alongside the tutorial.

4.2.1 Experiment 1: Building-level Approach

Experiment 1, the building-level approach, was designed to emulate earlier crowdsourcing initiatives, like those used in the Haiti earthquake or Typhoon Haiyan (Ghosh et al. 2011; Foulser-Piggott et al. 2016). For each building, users were asked to assess the level of damage as either "none", "some" or "destroyed". This three damage-level grading system was chosen, based on previous studies which suggested a simpler tagging system, mentioned in Section 2 (Huynh et al. 2014). Every buildings in an image were pre-identified to enable a direct comparison of user responses by avoiding errors of omission, as shown by the white markers in Figure 4.3. Users recorded this assessment by clicking on a building marker and selecting one of the three damage levels. Pre-event imagery was also available as a separate background layer that the volunteer could choose to show.

By nature, the building-level approach is very time intensive, since it requires a user to look at every single building in an image. A target of three volunteer assessments per image was set for experiment 1. Due to its time intensiveness, progress on this building-level approach was much slower than that of the two area-based experiments. After a month of slow response rates, the area of interest was reduced to from 281 images to 50 to achieve the three volunteer assessments per building within a smaller area. Unfortunately, experiment 1 was still not completed in the allotted time frame and therefore the results from this experiment have not been included in the analysis of this study. The area-based approach used in experiments 2 and 3 was developed to address this time demand, thus the lack of completion for experiment 1 further confirmed the need for an alternative approach.

4.2.2 Experiment 2: Damage Rating Approach

The interface for experiment 2 was updated based on the findings from the pre-experiment, as shown in Figure 4.4. After initial recruitment efforts, it was decided that only the 5-image scale

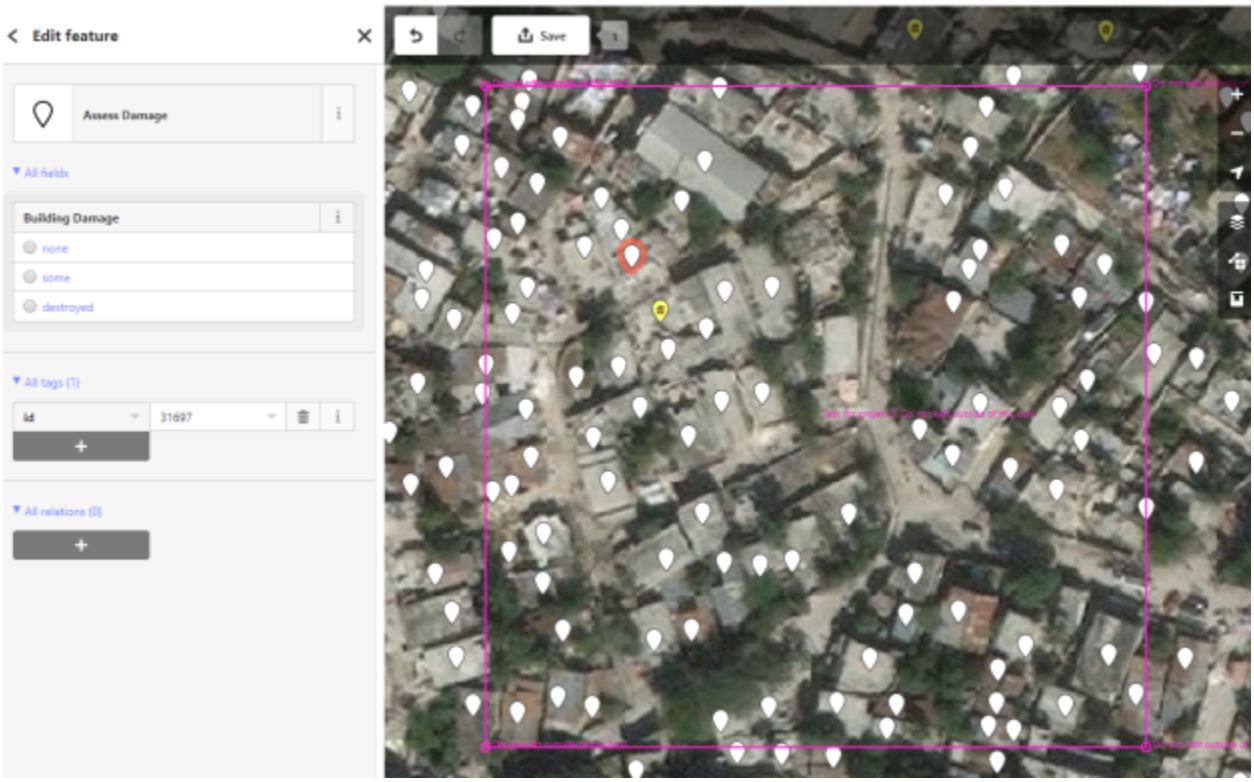


FIGURE 4.3: User interface for experiment 1 in the OpenStreetMap tasking manager. White markers indicate locations of buildings to tag with a level of damage: "none", "some", or "destroyed"

would be used for this experiment to focus volunteer efforts on completing all three experiments, rather than two versions of experiment 2.

A minimum target of three assessments per image was set for experiment 2, with the goal of achieving 10 for a smaller subset, so that we could develop more detailed statistics for user-to-user variability.

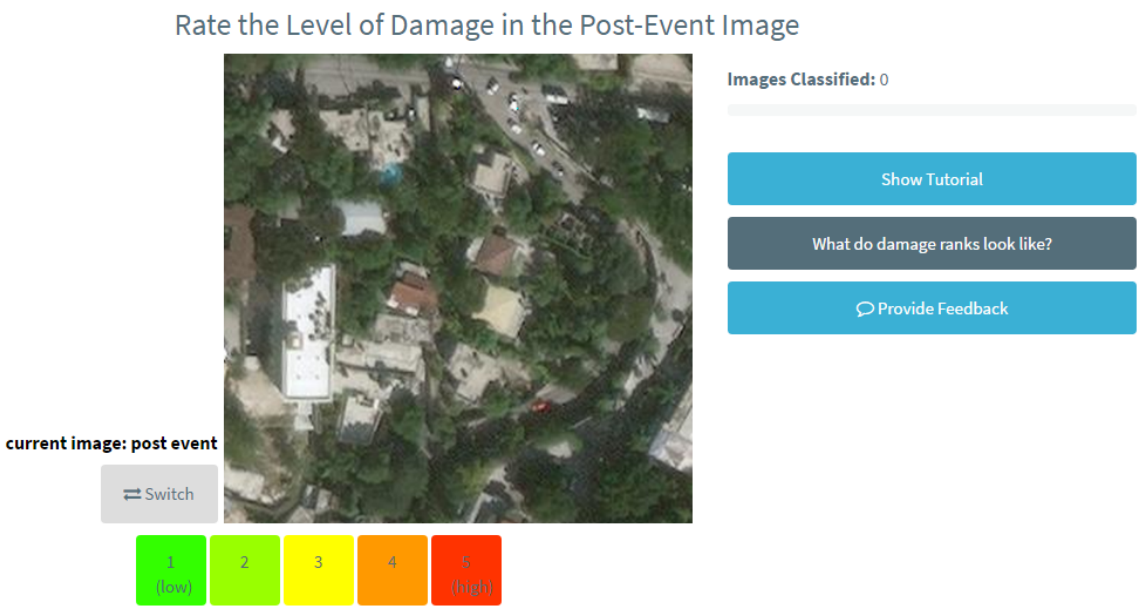


FIGURE 4.4: Final user interface for experiment 2

4.2.3 Experiment 3: Damage Comparison Approach

The interface for experiment 3 was also updated based on the findings from its pre-experiment and is shown in Figure 4.5. A minimum target of three assessments per image was also set for experiment 3.

FIGURE 4.5: Final user interface for experiment 2

4.3 Outreach and Participant Recruitment

The entire research team was involved in recruiting participants for the pre and final experiments, with efforts led by Stanford and HOT. Recruitment activities at Stanford involved Stanford publicity emails, mapathons, social media posts, reaching out to professional organizations, and developing a webpage on the Stanford Urban Resilience Initiative website. HOT's efforts focused on reaching out to their existing community by publicizing via their mailing list, social media, and blog posts.

The Stanford team held three mapathons on campus. The first was co-sponsored by Facebook, the Stanford Urban Resilience Initiative and the Stanford Geospatial Center and held in June 2017. The first part of the event involved an introduction to OpenStreetMaps and presentation of the project task. Following this, attendees started contributing to our pre-experiments, with the goal of testing out the OpenStreetMap platform for experiment 1. Due to some back-end issues with experiment 1, most effort shifted to pre-experiments 2 and 3. This event gathered a range of participants, including Facebook employees, Stanford students, staff, and faculty. Team members could observe attendees interacting with the platform and obtained valuable feedback regarding the user interface and structure of the questions.

The second and third mapathons focused on the final experiments. The second was held in August 2017 and was led by the Stanford Geospatial Center. The third and final mapathon was led by the Stanford team members in October 2017. By this time, the two area based assessments had been completed, so this effort focused specifically on experiment 1, the building-level approach. Unfortunately, this event did not have a significant impact on the progress for experiment 1, again because its time intensive nature.

In addition to the mapathons, the Stanford team emailed a number of different organizations on campus such as lists for: civil and environmental engineering, geospatial information science, and engineering student societies. Team members publicized the study on various social media platforms. The Stanford team also contacted various professional organizations, resulting in the study being publicized in newsletters from EERI, SEAONC, SCEC, and others. A project webpage was also created on the Stanford Urban Resilience Initiative's website.

Stanford's recruitment efforts heavily focused on those with a structural engineering or earthquake background. In future it would be preferable to reach a wider cross section of the population as the intended volunteer network in a live setting would encompass people from all backgrounds. Engaging someone with experience in survey design as well as marketing or outreach may be one way to address this.

HOT's efforts involved emailing a call to action regarding the project to their mailing list of active and interested volunteers, publishing two blogposts and posting to twitter. Despite HOT's large network of volunteers, we had difficulty recruiting participants, particularly for experiment 1. A large component of HOT's work is in response to current disasters and their volunteers are motivated by this immediate need and impact. As this study was a research experiment using the Haiti 2010 earthquake as its case study, it was difficult to convince volunteers to participate in a non-active task. In addition, the final experiments were run during a period when multiple hurricanes made landfall on the Caribbean islands and Gulf Coast of the US in 2017 (Hurricane Irma, Maria, Harvey). Since HOT was responding to several live emergencies, volunteers were understandably less interested to participate in this research-oriented task.

5 Ground-Validation and Experimental Results Datasets

The crowdsourcing experiments resulted in two forms of damage estimation data: a damage indicator and damage comparison dataset. Both area-based experiments 2 and 3 produced a damage indicator dataset, while only experiment 3 produced a comparison dataset. The accuracy of these two crowdsourcing datasets could be determined by comparing results with an extensive ground-validation dataset obtained from field surveys of building damage completed after the 2010 Haiti earthquake.

5.1 Ground-Validation Data from the 2010 Haiti Earthquake

In the months following the 12 January 2010 earthquake, the Haitian Ministry of Public Works led and conducted extensive field assessments of over 400,000 affected buildings by nearly 550 trained structural engineers (MTPTC 2010). These assessments were carried out using a modified version of the ATC-20 methodology, in which each building was assigned one of seven damage states originally defined in ATC-13 and described in Figure 5.1 (Applied Technology Council 1989; ATC-13 1985). Each damage state has an associated central damage factor (CDF), which is the midpoint of a range of damage ratios for a given structure.

TABLE 5.1: Description of ATC-13 damage states used to evaluate buildings in ground-validation dataset (ATC-13 1985)

Damage State	Damage Factor Range (%)	Central Damage Factor (%)	Damage Definitions
1 – None	0	0.0	No damage
2 – Slight	0-1	0.5	Limited localized minor damage not requiring repair
3 – Light	1-10	5.0	Significant localized damage of some components, generally not requiring repair
4 – Moderate	10-30	20.0	Significant localized damage of many components, warranting repair
5 – Heavy	30-60	45.0	Extensive damage requiring major repairs
6 – Major	60-100	80.0	Major widespread damage that may result in the facility being razed
7 – Destroyed	100	100.0	Total destruction of the majority of the facility

The field surveys of damage from Port-au-Prince and the surrounding region were used in this study to validate results from the crowdsourcing experiments. In total, there are 404,404 data points in this dataset with an associated latitude, longitude, and CDF.

In order to validate the crowdsourcing experiments, a metric of ground-validation building damage must be defined for each 125 $\times$ 125m grid defined for both area-based approaches. In this study, a mean CDF value was calculated , by taking the average of the CDF's for every building in each grid. Alternative metrics, such as the percentage of buildings in a grid cell with damage exceeding a given damage state could also be defined, but are highly collinear with the mean CDF value. Therefore, the mean CDF was used to compare with results from the crowdsourcing experiments to ground-validation information, due to its simple interpretation. Figure 5.1 depicts the mean CDF values for 125 $\times$ 125m grids containing field assessed buildings in Port-au-Prince and the surrounding region.

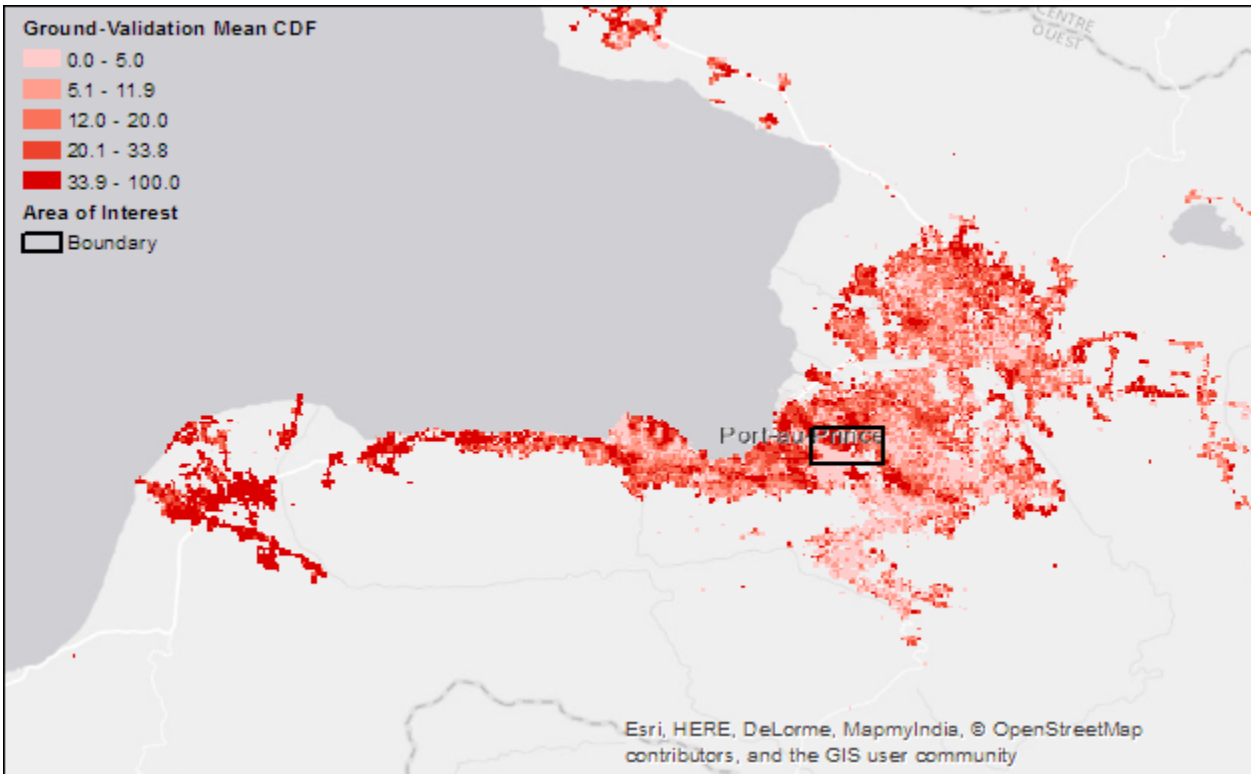


FIGURE 5.1: Map showing mean central damage factor (CDF) values of surveyed structures within 125m $\times$ 125m grid cells and the area of interest (AOI) boundary used for the crowdsourcing experiments

To efficiently implement the three crowdsourcing experiments for at least three multi-pass assessments, it was necessary to reduce the entire field surveyed region into a rectangular area of interest (AOI) of 281 grids, roughly 3.6 $\times$ 1.8km in size (delineated in Figure 5.1). This AOI was chosen because of its diversity in mean CDF values and building density, as shown in Figure 5.2. There is a clear spatial pattern of damage in the selected AOI, with higher mean CDF values located in the top left region. Blank areas in the AOI are grids without defined mean CDF values, either because field surveys were not completed in these regions or the imagery contained clouds.

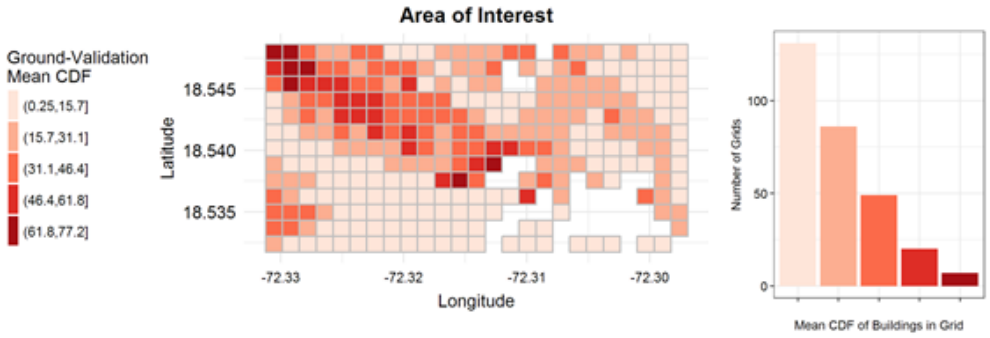


FIGURE 5.2: Distribution of mean central damage factor (CDF) values per grid in area of interest used in experiments

5.2 Damage Indicator Dataset

As mentioned previously, the damage indicator dataset resulted from both experiments 2 and 3, which could be easily downloaded from the Pybossa platform using a python script. In this dataset, each image or observation (corresponding to a 125 $\times$ 125 grid cell) was labeled by a unique "UID". The observations had an associated numeric "indicator" of damage, which was provided by individual volunteers, or users. The pertinent features in the damage indicator are described in more detail in Table 5.2.

TABLE 5.2: Features Included in Damage Indicator Datasets

Feature Name	Description	Example
UID	unique ID; The numerical ID of the grid (image) with an associated damage indicator	61190, 61191, ...
user_id	The IP address for the volunteer who provided a damage indicator (sometimes, partial IP address)	111.22.333.44
answer	The damage indicator value. Ranges from 1-5 for experiment 2 and 1-11 for experiment 3	1,2,3... 11
finish_time	The time when the final indicator value is recorded for an image	2017-06-28T14:45:25.282362
runtime	The amount of time (in seconds) taken for user to provide indicator value	4.637
latitude	Latitude of grid (image)	18.548
longitude	Longitude of grid (image)	-72.332

The damage indicator for a grid from experiment 2 is defined after a user clicks on a value between 1-5 in the numerical scale described in 4. Recalling that users compared the damage between two images rather than directly inputting a numerical indicator value in experiment 3, the process of obtaining a damage indicator value for an image is less straightforward for this approach. A damage indicator is instead defined for an image after a series of iterative comparisons between the image of interest and 10 "anchor" images with known damage representing 10 mean CDF levels. Unbeknownst to the users, each image is initially compared to the midpoint anchor image with the 6th lowest damage level, and binned between two images after a maximum of four comparisons with each subsequent anchor image. The iterative comparison process to reach a final damage indicator value in experiment 3 is visually detailed in Figure 5.3.

As shown in Figure 5.3, the only way images in experiment 3 are marked as having damage indicator values shown in the first three comparisons (6, 4, 8, 2, or 10) is if the user answers "same". This impacts the distribution of responses for each damage indicator value in experiment 3, which is discussed in the following section.

5.3 Damage Comparison Dataset

The contents of the damage comparison dataset, resulting from experiment 3, resembled those of the damage indicator dataset described in the previous section. Table 5.3 summarizes the useful features captured through the Pybossa interface. Notably absent are the results of each individual comparison between an anchor image of known damage with a comparison image of unknown damage (labeled by UID). Instead, each observation is only the final result of a set of comparisons – that is, the number of the final bin (damage range) into which the user unknowingly placed the image with unknown damage. This omission arose from communication errors between the experimental design team and the team members responsible for developing and operationalizing the interfaces.

However, as shown in Figure 5.3 in the previous section, each bin has a single, unique path of comparisons leading up to that bin. That is, there is only one way for a volunteer to classify a comparison image into any of the bins. Therefore, we could deduce the full set of comparisons implicit in each of the final observations.

For example, if the final classification of an image with unknown damage were bin 7, we could deduce the following:

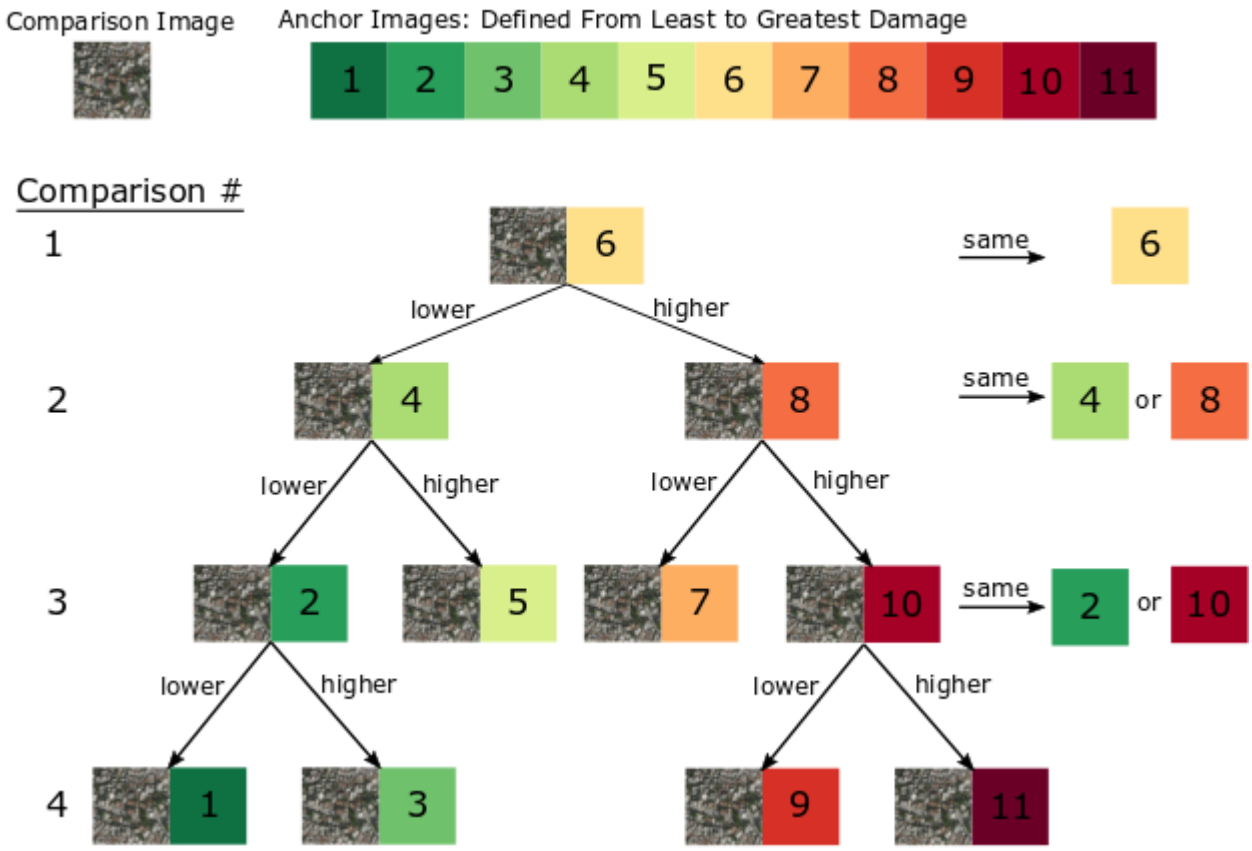


FIGURE 5.3: Iterative comparisons between image of interest and defined anchor images to achieve a final damage indicator value in experiment 3

1. In the first comparison, the volunteer indicated that the image with unknown damage had more damage than an anchor image with a mean central damage factor (CDF) of 50%
2. In the second comparison, the volunteer indicated that an anchor image in bin 8 had more damage than the image with unknown damage

If the final classification of an image with unknown damage were bin 2, we could deduce the following set of comparisons:

1. In the first comparison, the volunteer indicated that the anchor image, with a mean CDF of 50%, had more damage than the image with unknown damage
2. In the second comparison, the volunteer indicated that an anchor image in bin 4 had more damage than the image with unknown damage

3. In the third comparison, the volunteer indicated that an anchor image in bin 2 had the same level of damage as the image with unknown damage

Thus, while the raw dataset appeared to have the results of on the order of 800 comparisons, it actually included the data from over 1500 comparisons.

The anchor images for each comparison were randomly selected from a carefully predefined set for experiment 3. Due to their random selection, we could not retrieve the unique ID of the anchor image that was shown to the volunteer in each of the retrieved comparisons. However, we knew the level of damage in the anchor image for each comparison, since that was a pre-defined parameter.

Volunteers classified a large proportion of the images with unknown damage into bin 6, apparently indicating that they believed the image had the same level of damage as the anchor image with mean CDF of 50%. Based on anecdotal feedback, we believe that this is indicative of participants' difficulty in completing the task rather than an unbiased selection of bin 6. Bin 6 was in fact the easiest of all the bins to reach, in that it required only a single comparison.

TABLE 5.3: Features Included in Damage Indicator Datasets

Feature Name	**Description**	**Example**
task_id	the numerical ID of the task, uniquely associated with the UID	63463, 63464, ...
user_id	the volunteer's IP address (sometimes, partial IP address)	111.22.333.44
finish_time	the time at which the user finished the comparisons for a single image to be classified	2017-06-28T14:45:25.282362
UID_b	unique ID of image b; the numerical ID of the grid (image) that was randomly selected for the user to compare with the grid (image) to be classified ID of the	63673, 62217, ...
random_comparison	the result of a comparison between the image to be classified and a randomly selected image, also with unknown damage	63673 > 61190 , 63673 = 61190, 63673 < 61190
UID	unique ID; the numerical ID of the grid (image) to be classified through sequential comparisons with anchor images	61190, 61191, ...
answer	the number of the bin in which the user unknowingly classified the grid (image) through sequential comparisons with anchor images	1, 2, 3, ... 11
runtime	the time a user took to complete a task (i.e., bin an image) in seconds	4.637

6 Results from Damage Indicator Dataset

The 281 images investigated in experiments 2 and 3 received a minimum of three assessments by three separate volunteers. The goal of the analysis described in this section is to determine if it is possible to interpret the directly input damage indicator values in such a way that it represents the spatial distribution of ground-validation damage. The analysis of the damage indicator datasets for experiments 2 and 3 was carried out in four main steps:

1. Initial exploration of the damage indicator dataset, to determine potential relationships between variables, trends in user responses, and pertinent variable transformations
2. Development of simple regression models to explore the correlation between crowdsourced damage assessment results and real damage obtained from ground-validation data
3. Exploration of weighting individual responses according to volunteer performance or image characteristics
4. Aggregation of multi-pass assessments for spatial representation of damage distribution

6.1 Exploratory Analysis and Cleansing of User Responses

While volunteers provided damage indicator values for all 281 grids in both area-based experiments, the two datasets have different numbers of total responses and volunteers who contributed to each experiment. Experiment 2 (damage ranking) resulted in 1,306 responses from 65 users, found by the number of unique user ID's. Conversely, experiment 3 (damage comparison) had fewer users and responses, with 51 users providing 852 total responses.

The length of time to provide a damage indicator value for an image was recorded as the "runtime". For experiment 2, it took users between 1.2 seconds and 952 seconds (15.8 minutes) to input a damage indicator value for an image. Two outlier contributions took more than 30 minutes to

complete. It was found that images users believed to have lower damage had a wider range of runtimes to provide a damage ranking, and thus the task was more difficult for images with perceived lower visible damage. Conversely, runtimes for experiment 3 are the length of time to complete the entire series of comparisons to reach a final damage indicator value, not the time to complete a single comparison. Therefore, we did not incorporate experiment 3 runtimes in this study.

Before performing further analysis, the dataset was initially cleansed of unreliable user responses. We defined unreliable users as those who provided only one response or who provided the same response for all their contributions. All responses from unreliable users were removed from the damage indicator dataset. This data cleaning process reduced the total number of responses to 1,112 data points from 51 users in experiment 2 and 836 data points from 43 users in experiment 3. The reduction in responses for each damage indicator value when removing "one and same response users" is shown in Figure 6.1.

It should be noted that there were some users that input an inordinate amount of responses. For example, one user from experiment 2 had over 100 responses, and two users from experiment 3 had over 200 responses. While this number of responses is possible for a user to complete, it is not probable. Since user IDs were recorded as the IP address of the volunteer's computer, this substantial number of responses is most likely due to overlapping IP addresses when multiple volunteers are physically located in close proximity and connected to the same network. This could be a result of the various mapathons that were hosted at Stanford University to obtain results for the two experiments. However, if the responses from "excessive" users are removed, it severely reduces the datasets to 888 and 397 data points for experiments 2 and 3, as shown in Figure 6.1. Thus, the results from the "excessive" users were included in the analysis of this dataset.

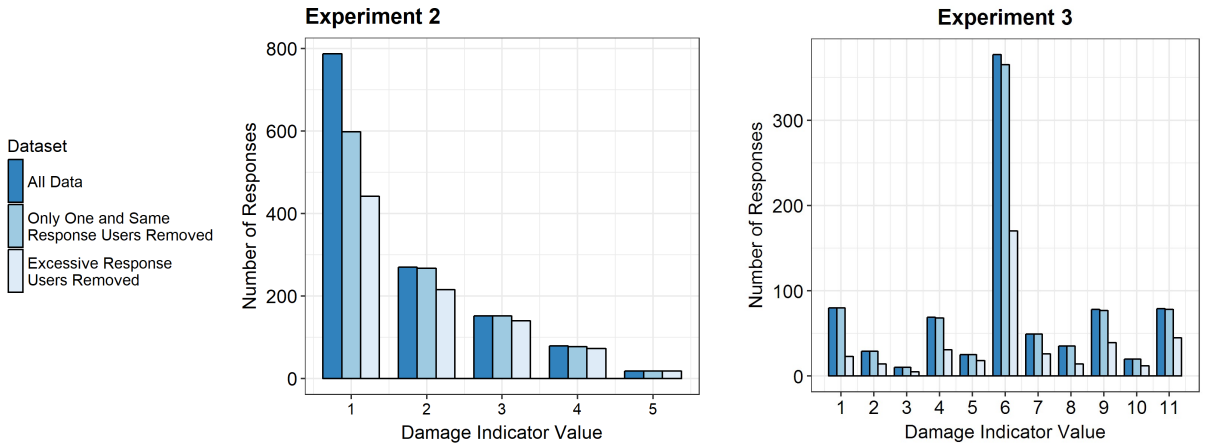


FIGURE 6.1: Distribution of the number of user responses for each damage indicator value for experiments 2 and 3 before and after the initial data cleansing of nonsensical responses and unreliable users. Also shown is the potential reduction in responses if removing users with an excessive number of responses

Cleaning the data slightly alters the distribution of the number of responses for each damage indicator value. Before cleaning the data, there is the greatest number of damage indicator values of "1" for experiment 2 and "6" for experiment 3, which can be observed by the "All Data" bar in Figure 6.1. The shape of the distribution in experiment 2 closely matches the expected distribution of mean CDF values in the area of interest from Figure 5.2 - the distribution of responses decreases with increasing indicator value. However, the shape of the distribution of responses from experiment 3 is more symmetrical, due to the iterative comparison method used to bin images Figure 5.3. The disproportionate number of responses with a damage indicator value of "6" for experiment 3 confirms that many users immediately chose "same" damage between the comparison and anchor image, as stated in Section 5.3. Given the features in the dataset, it was not possible to remove only the users who chose "same" for their responses.

Once the damage indicator dataset was cleaned, an exploratory analysis was performed to discover potential relationships between the damage indicator values obtained from crowdsourcing and the actual damage level from the ground-validation surveys. The mean central damage factor, defined from the ground-validation data, represents the "true" average building damage per grid. Scatterplots of the mean CDF versus the individual damage indicator value a single contributor provided for an image is shown in Figure 6.2. The gray violin plots exhibit the frequency distribution of responses and the horizontal red lines highlight the average of the mean CDF values of the images classified at each damage indicator level in Figure 6.2.

The volunteers' responses varied widely for each experiment's indicator values, shown by the individual data points plotted in black in Figure 6.2. Nonetheless, the average of the mean CDF values for user responses from experiment 2 exhibit a slight positive trend with increasing damage indicator value. The minor decrease in the average mean CDF value for damage indicator "5" is due to its significantly fewer number of responses. The unimodal shape of the frequency distributions for experiment 1 also show large agreement between users at the two lowest damage indicator values. However, with damage indicator values 3-5, this distribution becomes bimodal, implying that a concentration of responses aligns with higher mean CDF values. The increase in average mean CDF values fluctuates more with the increase in damage indicator value for experiment 3. Conversely, there is not a clear trend in frequency distribution shapes for experiment 3 due to the disproportionate number of responses for indicator value "6".

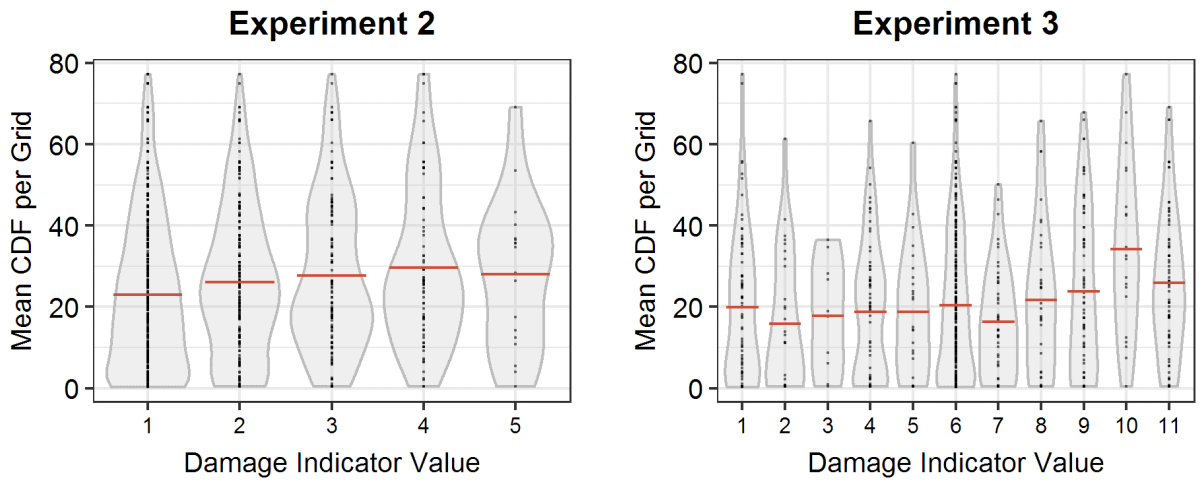


FIGURE 6.2: Scatterplots exhibiting the relationship between user-provided damage indicator value and the mean CDF per grid. The gray violin plots show the frequency distribution of responses and the horizontal red lines highlight the average mean CDF value for each damage indicator value.

It is initially unclear whether users are detecting the damage in an image or the building density in an image. Therefore, similar scatterplots were developed to compare the building density and the user-provided indicator value from experiments 2 and 3. As shown in Figure 6.3, building density also exhibits a slight positive trend with the user-provided damage indicator values for both experiments 2 and 3. This implies that volunteers are detecting a combination of both damage and building density.

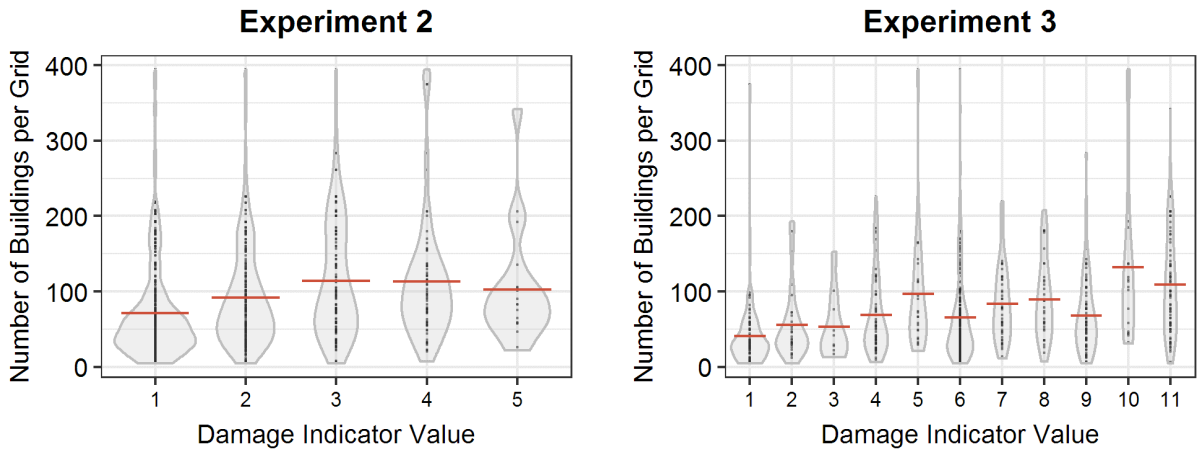


FIGURE 6.3: Scatterplots exhibiting the relationship between the damage indicator value and the building density in a grid. The gray violin plots show the frequency distribution of responses and the horizontal red lines highlight the average mean CDF value for each damage indicator value.

We can directly compare the relationship between the user-provided damage indicator values and the mean CDF or the building density per image by investigating their respective correlation coefficients. Indeed, the higher correlation coefficients for damage indicator versus building density, shown in Table 6.1, highlights how the crowd may be indicating that images with more buildings have higher damage.

TABLE 6.1: Correlation between user-provided damage indicator value and ground-validation mean CDF or building density

	Correlation Coefficient	
Experiment	Damage indicator value vs. Building Density	Damage indicator value vs. Mean CDF
Damage Ranking (2)	0.222	0.115
Damage Comparison (3)	0.247	0.124

While there is a higher correlation between damage indicator and ground-validation building density, the positive correlation between damage indicator and ground-validation damage is still an indication that the crowd is visually detecting a form of damage in the grids of satellite imagery provided to them to assess. The relationship between ground-validation and user-provided damage, however, is slight with the average of the mean CDF per grid ranging between 23.1-29.7 for experiment 2 and 15.8 - 34.2 for experiment 3. The slight positive relationship between the average mean CDF values and numerical damage indicators will be further analyzed in the following

section by incorporating the variation in performance of different contributors and aggregating multi-pass assessments. The urban density of an image will also be considered to account for the relationship between the damage indicator value provided and the number of buildings per grid.

6.1.1 Inferring Ground-Validation Damage from Crowdsourced Indicators

Several regression models were developed to explore the relation between the ordinal damage indicator assessments provided by the crowd and true damage from the field surveys. Because we would like to determine whether we can improve the overall ability of the crowd to infer "true" damage, linear regression models were used as the simplest and most interpretable parametric option. However, alternative regression models that could improve overall performance should be explored in the future.

Through the regression analysis of the damage indicator results, the following questions will be addressed:

- Does ordinal scaling of damage indicator values improve the performance of the crowd?
- Does weighting the responses based on users or images attributes improve the performance of the crowd?
- Does the resulting weighted multi-pass assessment from the crowdsourcing experiments properly reproduce spatial distribution of damage?

Linear regression between ground-validation and crowdsourced damage severity

Linear regression is a common method for assessing the strength of a relationship between two variables (James et al. 2013). The baseline regression developed for each experiment is a linear model with mean CDF as the dependent variable and the directly provided damage indicators as the independent variable:

$$Y \cong \beta_0 + \beta_1 X \tag{6.1}$$

$$Y = \text{Mean CDF}$$

$$X = \text{damage indicator values}$$

The most common method of determining the model coefficients, β, is by minimizing the least squares criterion, the sum of squared residuals or error (SSE), so the regression line most closely fits n data points. This results in the two baseline ordinary least squares regression models, shown in blue in Figure 6.4.

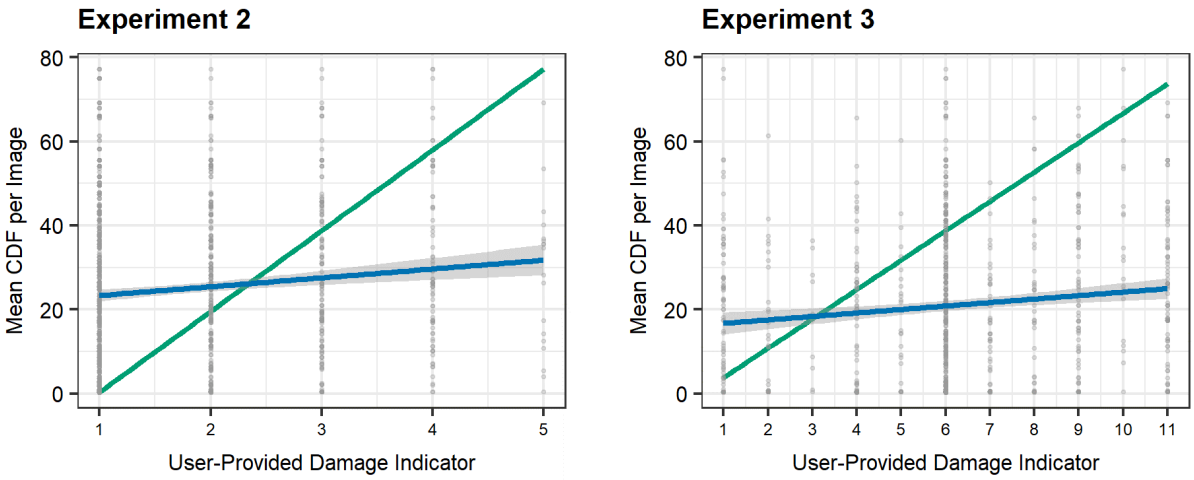


FIGURE 6.4: Baseline linear regression models (blue) with crowdsourcing results for experiments 2 and 3 showing standard error. The "true damage" line (green) shows a linear regression between the lowest and highest mean CDF values representing the extreme damage indicator values.

Defining error metric for comparing models

Assessing the performance of the regression model using a metric that defines its goodness of fit with individual data points is not the most applicable approach for two reasons. First, the wide variability in the crowd's performance would lead to substantially large errors between the data and the model. More importantly, we are not interested in the mean CDF value for ground-validation damage that most users associate with a given damage indicator, which is represented by the linear regression model. Rather, we are more interested in how close the mean CDF value defined by the crowdsourcing regression model agrees with the ground-validation mean CDF from our area of interest.

A new error metric was defined to compare the linear regression model and the anticipated linear trend between the damage indicators and ground-validation damage. Assuming a linear trend exists between these two variables, the anticipated linear trend can be defined by connecting the highest and lowest damage indicators for each experiment with the most extreme mean CDF

values in the area of interest. The mean CDF values for the interior damage indicators are then linearly interpolated. Figure 6.4 also shows the anticipated linear trend (green) in comparison with baseline regression models for experiments 2 and 3 (blue).

The goal of this error metric is to calculate the difference between the anticipated linear trend and every tested regression model - a lower error metric signifies better performance of the regression model. Therefore, the metric compares two linear models, not the specific data points to the anticipated linear trend, by comparing the coefficient β_1 for each model:

$$ErrorMetric = \left| \frac{\beta_{1,anticipated} - \beta_{1,regression}}{\beta_{1,anticipated}} \right| \tag{6.2}$$

$$\beta_{1,anticipated} = \beta_1 \text{ for the anticipated linear relationship}$$
$$\beta_{1,regression} = \beta_1 \text{ for the regression model}$$

This error metric allows for each tested model to be compared to the same line that represents what we would expect to be the relationship between the experimental damage indicators and ground-validation damage. It also allows for the regression models to be compared between experiments. One caveat to this metric is that it does not incorporate the differences in the y-intercept, β_0. Also, it assumes that ground-validation and crowdsourced damage are linearly related. However, this assumption was followed to match the form of the parametric regression model chosen to analyze the damage indicator dataset.

Testing Alternative Damage Indicators using Ordinal Scaling

The damage indicator values ranging from 1-5 for experiment 2 and 1-11 for experiment 3 are ordinal rather than numerical values of damage. Such data can be ordered, but their relationship is not necessarily evident (e.g. is damage level 4 twice as much as level 2?). Directly assuming the user-defined indicator values as numerical values would implicitly assume that the intervals between each indicator value is equally spaced. However, it is unclear that volunteers interpreted and responded as such.

We can therefore treat the user-provided numerical damage values as ordinally scaled categorical values and test if this improves the performance of the crowdsourced regression model. Variables with ordinal scales exhibit a clear ordering of levels, without knowing the absolute distance between these levels (Agresti 2002). It is implicit in the way these indicator values were obtained

from both experiments that they represent magnitudes of ground-validation damage ordered from least to greatest. Treating the user-defined damage indicators as ordinally scaled quantitative values allows for the damage indicators to have unequal intervals, and thus represent different magnitudes of ground-validation damage. Describing the data ordinally still allows for statistical analysis typical for quantitative variables (Agresti 2002).

A common method of ordinal scaling is to transform the provided damage indicators into "Ridit" scores. This transformation scales an ordinal variable to numerical values between 0-1, with intervals between values relative to the proportion of responses for each value. Determining the Ridit score transformation of the damage indicator value a_j, or the average cumulative proportion for a given value is

$$a_j = \sum_{k=1}^{j-1} p_k + \frac{1}{2} p_j \tag{6.3}$$
$$j = 1, 2, \ldots, c$$

Where p is the sample proportion of responses for each category (damage indicator value) j which ranges from 1 to the total number of categories c.

The transformed Ridit score damage indicator values are shown in Table 6.2. The original user-provided damage indicators are transformed into cumulative proportion values from 0 to 1, which can be used as alternatives to the original damage indicator values in the crowdsourced regression models.

TABLE 6.2: Ordinal Scaling of user-provided damage indicators by Ridit transformation (Agresti 2002)

Experiment	User-Provided Damage Indicator	Number of Responses	Sample Proportion	Ridit Score Damage Indicator (Cumulative Proportion)
Damage Ranking (2)	1	598	0.538	0.269
	2	267	0.24	0.658
	3	152	0.137	0.846
	4	77	0.069	0.949
	5	18	0.016	0.992
Damage Comparison (3)	1	80	0.096	0.048
	2	29	0.035	0.113
	3	10	0.012	0.136
	4	68	0.081	0.183
	5	25	0.03	0.239
	6	365	0.437	0.472
	7	49	0.059	0.719
	8	35	0.042	0.77
	9	77	0.092	0.837
	10	20	0.024	0.895
	11	78	0.093	0.953

Another transformation of the damage indicators that we tested was aggregating responses for certain adjacent values. Since the damage indicator values of "4" and "5" in experiment 2 comprised 1.6% and 6.9% of all the responses, respectively, the responses for these values were aggregated into one damage indicator value of "4" so the scale of damage indicators ranges from 1-4. In experiment 3, the responses for indicator values that are reached through the option of indicating "same" damage for both images were aggregated with their adjacent values that are reached without indicating "same" throughout the iterative comparison process shown in Figure 5.3. The responses for value "2" were aggregated with the responses for "1", "4" with "3", and so on; the responses for "11" were not aggregated with any other responses. Therefore, the aggregated damage indicator scale for experiment 3 ranges from 1-6. Subsequently, the aggregated damage indicator values were also ordinally scaled into Ridit scores using Equation 6.3.

Four separate linear regression models were then developed and compared. These models used four different values as the independent damage indicator: (1) the untransformed damage indicators directly provided by the users, (2) the ordinally scaled "Ridit" score damage indicators, (3) the aggregated untransformed damage indicators, and (4) the ordinally scaled aggregated indicators. The four models were fit using the full dataset without the one and same response users, but still containing the excessive users.

Understanding which form of damage indicator results in the best fit of the data can be quantified by determining which corresponding regression model results in the lowest error metric. The error metric for the linear models using the tested damage indicators are shown in Table 6.3.

TABLE 6.3: Comparison of linear regression models using different damage indicator transformations

Experiment	Damage Indicators used in Model	Error Metric (%)
Damage Ranking (2)	Untransformed (1-5)	0.890
	Ordinally Scaled (0-1)	0.920
	Aggregated (1-4)	0.912
	Aggregated, Ordinally Scaled (0-1)	0.924
Damage Comparison (3)	Untransformed (1-11)	0.892
	Ordinally Scaled (0-1)	0.910
	Aggregated (1-6)	0.897
	Aggregated, Ordinally Scaled (0-1)	0.913

In both experiments, ordinal scaling and aggregation of damage indicators increases the error metric of the regression models. Ordinal scaling thus provides no additional benefit in terms of improving a model which employs user-provided damage indicators to predict ground-validation damage. Therefore, the untransformed damage indicators are still employed for the additional regression analyses in the following sections.

Weighting User and Image Characteristics to Improve Crowd Performance

The exploratory data analysis of the untransformed damage indicator dataset highlighted the wide variation in the accuracy of different users in assessing the damage severity in the provided images. This variability in user accuracy has been addressed in recent crowdsourcing platforms, such as TomNod who uses a "crowdrank" algorithm to rank the reliability of their assessors on a numerical scale of 0-1 (www.tomnod.com). Instead of removing users who are particularly poor at

assessing damage, performance differences are addressed by using weighted least-squares (WLS) regression.

With WLS regression, the previous baseline linear regression models are modified by assigning weights to data points based on certain parameters of interest. Weighting data has many applications, especially with survey data, including rating responses of specific survey takers, addressing sampling bias, or treating heteroskedasticity in regression residuals (David and Sutton 2011; Hausser n.d.). The primary goal for weighting with user-based data is to increase the analytic capabilities of survey responses post-collection. In this study, weighting is applied to incorporate the differences in user performance, similar to rating responses of survey takers to emphasize key personal characteristics (Hausser n.d.).

Similar to ordinary linear regression, the β parameters of WLS regression are again estimated by minimizing the least squares criterion (sum of squared error) between the observed data and the regression output. However, the criterion is now modified by weighting the squared deviation of each data point with an additional weight w_i that governs the contribution of each response (NIST and SEMATECH 2018). The weighted fitting criterion is defined as

$$WSSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \tag{6.4}$$

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$
$$w_i = weight$$

Defining the weights w_i depends on the specific application of using WLS regression. In all cases of WLS regression, the weights' values relative to one another for a given weighting parameter influence the final model parameter estimates (β).

For the analysis of the damage indicator results, weights were defined to assess the impact of incorporating the following differences in responses based on user and image characteristics:

1. A metric to rate the "performance" of the user
2. Weight the user's "performance" equally for each of their cumulative contributions
3. User median runtime (only for experiment 2)
4. Building density in an image

When implementing WLS regression for the above weights, the weights were first defined for each user or image characteristic. The validity of each weighting parameter was then evaluated by using the defined error metric for weighted linear models regressed using subsets of the data. Finally, valid weighting parameters were cumulated to achieve a final linear regression model that can be used to plot the spatial distribution of damage.

Defining Weights for User and Image Characteristics

Perhaps the most important user attribute to incorporate into the baseline linear model is the differences between users' abilities to identify building damage. There were many users who successfully provided higher damage indicator values for images with larger ground-validation damage, and vice versa. However, a portion of the users rated damage in an opposite manner, providing higher damage indicator values for images with lower mean CDF values. To incorporate the variation in performance between all k users, we defined a "user performance metric", based on the coefficient β_1 of a linear model between the mean CDF and untransformed damage indicator values provided by each j^{th} user :

$$w_{performance,j} = \begin{cases} 0 & \text{if } \beta_{1,j} \leq 0 \\ \frac{\beta_{1,j}}{\sum_{j=1}^{k} \beta_j > 0} & \text{if } \beta_{1,j} > 0 \end{cases} \tag{6.5}$$

Each response for an individual user was weighted by this user performance weighting metric. Weighting users with negative individual β_1 coefficients with $wperformance, j = 0$ effectively removes their contribution in the overall WLS regression. Less than half of the users in experiments 2 and 3 were assigned a weight of zero; 21 out of 51 total users in experiment 2 and 15 out of 43 users in experiment 3. The weights for positively performing users is based on their individual β_1 coefficient, normalized by the sum of the β_1 coefficient for all users. This was done to scale the weighting parameters to be between 0 and 1, so the final weights for all the valid weighting parameters could be finally cumulated together.

We used this approach of applying individual β_1 coefficients to define a performance metric, because a metric did not previously exist to rate users on their reliability. As mentioned previously, if a previous metric did exist, such as Tomnod's "CrowdRank" rating system, this could be used in place of the defined weighting metric to carry out the WLS regression. Another option could be to provide a "training" set of the same set of images to users before they complete the actual damage assessment experiments. A user's performance on the same pre-defined set of images

with known ground-validation damage could then be used to define a weighting metric that is directly comparable between all users. In the future, we recommend incorporating a training set of assessment images for this purpose.

The next tested user performance metric ensures that the responses from positively-performing users who completed a greater number of tasks, and therefore have more data points, do not have a disproportionate weight in the overall WLS regression. This "equally distributed performance" weighting metric thus divides the user performance weighting metric by the number of tasks n_j completed by each user j, which was also scaled between 0 and 1:

$$w_{equal\ performance,j} = \begin{cases} 0 & \text{if } \beta_{1,j} \leq 0 \\ \frac{w_{performance,j}}{n_j} & \text{if } \beta_{1,j} > 0 \end{cases} \tag{6.6}$$

Another user attribute that was tested for weighting parameters was runtime, based on the idea that users who spend a longer time on a task will perform better. Again, the runtime is not the length of time for an individual comparison in experiment 3, so this weighting parameter was only tested for experiment 2. Weights were applied for those users whose median runtime, $m(t_i)$, was greater than the overall median runtime ($m(t)$ = 21.1 seconds) for all users:

$$w_{runtime,j} = \begin{cases} 0 & \text{if } m(t_j) < m(t) \\ \frac{m(t_j)}{\sum_{j=1}^{k} m(t_j) \geq m(t)} & \text{if } m(t_j) \geq m(t) \end{cases} \tag{6.7}$$

A final weighting parameter was tested to address the differences in image characteristics resulting from the initial data exploration showing an increase in the number of buildings per image b with the value of damage indicators provided by users. This weighting metric was created to simulate a real-time post-disaster scenario, in which it would be known whether images are "urban" or "rural", but the specific building count per image is unknown. This was done by weighting urban images with double the weight of rural images, then scaling all weights to be between 0 and 1 for all 281 images (p). Urban images were defined as those with more than 50 buildings per image.

$$w_{building\ count,p} = \begin{cases} w_{rural} & \text{if } b_p < 50 \\ 2 \times w_{rural} & \text{if } b_p \geq 50 \end{cases} \tag{6.8}$$

$$w_{rural} = \frac{0.5}{\sum_{i=1}^{p} w_{building\,count,i}}$$

Performance of Weighted Least-Squares Regression

To justify the use of the weighting parameters, it was necessary to ensure WLS regression models incorporating these weights improve the agreement of the baseline linear regression with the anticipated linear trend. Improvement, in this case, is quantified by a reduction in the error metric defined in Section 6.1.1. The weighting parameters' validity also depends on whether it applies for all subsets of users or images. For example, a weighting parameter could decrease the error metric for a certain group of users but increase it for another group of users. Therefore, the four weighting parameters were validated by examining both a reduction in the error metric and assessing whether this reduction is consistent for different subsets of users or images.

This validation was carried out by randomly grouping the damage indicator dataset into five mutually-exclusive subsets of users or images and recording the error metric for a WLS regression developed using only the data in each fold. An example of this process for one experiment and one weighting parameter is shown in Figure 6.5. The mean and the standard deviation of the error metric found for the five folds were then recorded and can be compared to that of the baseline (unweighted) regression model, shown in Table 6.5.

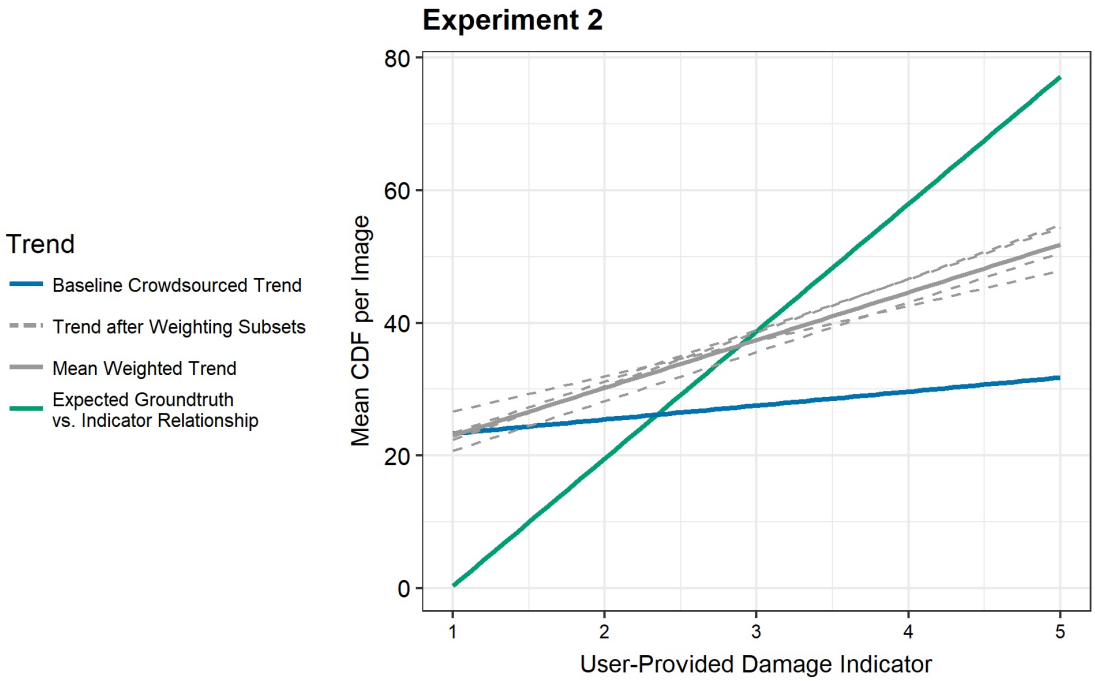


FIGURE 6.5: Example of the validation process using five random of subsets of the damage indicator dataset (split by user or image) to develop a WLS regression model

It is clear from Table 6.5 that the error metric of the baseline model for both experiments are large, with the β_1 coefficients being nearly 90% lower than anticipated linear trend for both experiments. However, by incorporating the two user performance weighting metrics, the baseline error metric reduces by about 30% for experiment 2 and 50% for experiment 3. This reduction shows that weighting users by their performance during the experiment improves the overall linear regression. However, the standard deviation for the error metric significantly increases when weighting user performance equally per contribution, meaning there is greater variability in the models using this weighting parameter. On the other hand, the error metric values for weighting users by their runtimes or images by their building density did not significantly reduce the error metric for either experiments. Also given the large standard deviation for the error metric when weighting by runtime or building density, using these two weighting parameters is not beneficial.

TABLE 6.4: Mean and standard deviation of error metric of five least squares linear regression models using the untransformed damage indicator of five random data subsets to validate each weighting parameter

		Performance Metric for 5 - subsets	
Experiment	**Weighting parameters applied in model**	**Mean**	**Standard Deviation**
Damage Ranking (2)	Unweighted (OLS)	0.890	0.022
	User performance weighted	0.727	0.037
	User performance weighted equally per contribution	0.628	0.056
	Runtime weighted	0.889	0.091
	Building density weighted	0.899	0.021
	Unweighted (OLS)	0.892	0.021
Damage Comparison (3)	User performance weighted	0.667	0.055
	User performance weighted equally per contribution	0.449	0.115
	Building density weighted	0.88	0.022

Comparing the results from this validation is also useful to assess the model performance between the two experiments. The baseline unweighted regression model for experiment 2 has a lower error metric than that of experiment 3, which signifies that the initial OLS linear regression for experiment 2 is closer to the anticipated linear trend between ground-validation and crowdsourced damage. This is also true for the WLS linear regression using the basic user performance metric. However, when equally distributing the user performance weights by the number contributions, the WLS regression for experiment 3 has a lower error metric, implying that this weighting parameter had greater impact on experiment 3. However, this may be due to the number of responses from "excessive" users in experiment 3, which was initially shown in Figure 6.1.

Overall Model Improvement Using Weighted Least Squares Regression

Considering the results of the validation for the weighting parameters, the final weighted least squares regression models were developed using the entire damage indicator dataset. The improvement in the inference of ground-validation damage by weighting the performance of different users with linear regression can be visually compared in Figures 6.6 and 6.7 for experiments 2 and 3, respectively.

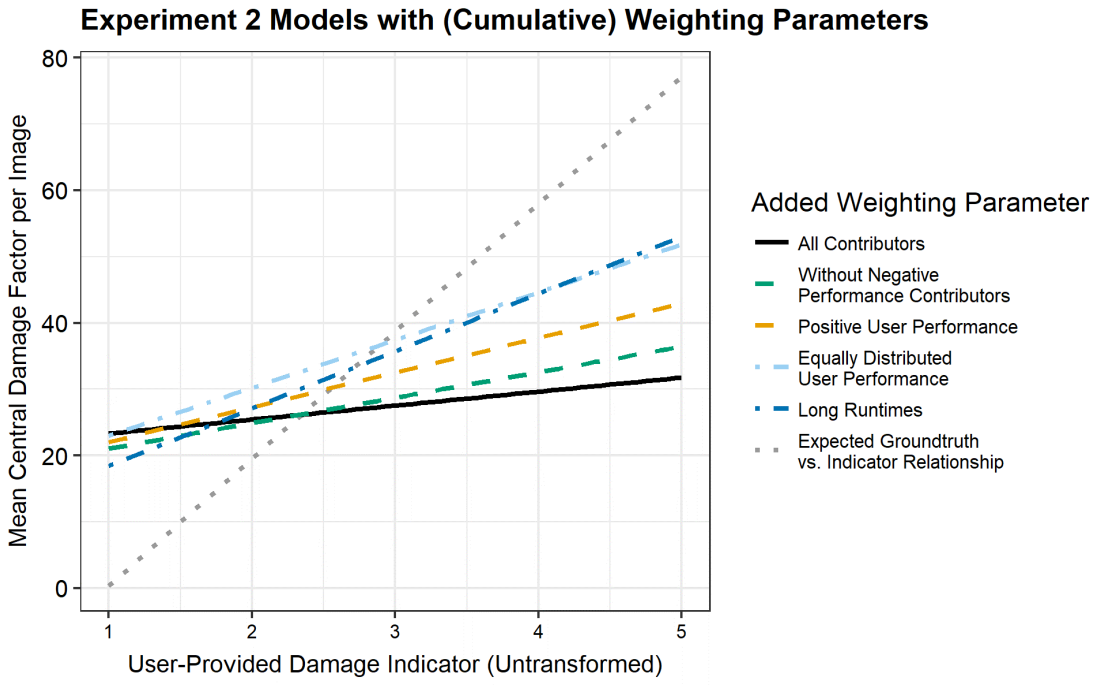


FIGURE 6.6: Improvement in Experiment 2 baseline linear regression model (Black) by removing negatively performing users (Green), weighting positive performing users (Yellow), and equally distributing positive user performance (Light Blue) compared to anticipated linear trend (Gray). Weighting long runtimes (dark blue) is shown, but does not improve performance

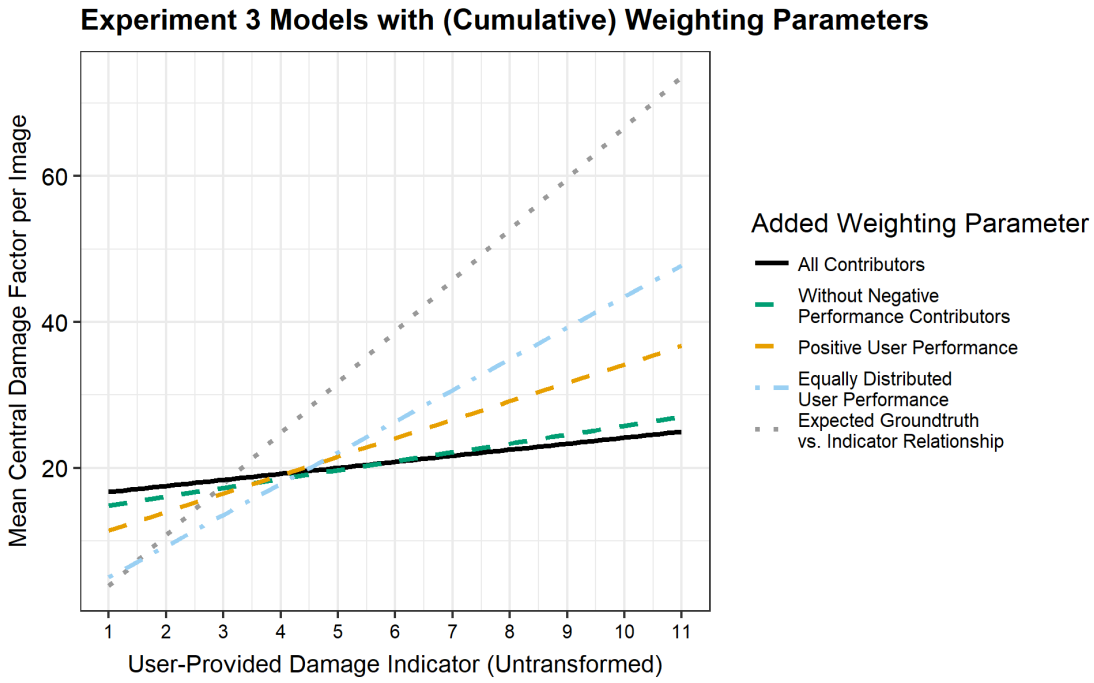


FIGURE 6.7: Improvement in Experiment 3 baseline linear regression model (Black) by removing negatively performing contributors (Green), weighting positive performing users (Yellow), and equally distributing positive user performance (Light Blue) in comparison with the anticipated linear trend (Gray).

The weighted least squares regression models shown in Figures 6.6 and 6.7 highlight some key results from the described weighting methodology. First, removing the negatively performing users slightly improves the baseline regression models (shown in black) for experiments 2 and 3, as the (green) regression model moves closer to the anticipated ground-validation versus crowdsourced damage indicator trend (gray). This shift closer to the anticipated line is quantified by the reduction in the error metric described in Section 6.1.1. The models are further improved by weighting positively performing users with the two user error metrics $w_{performance}$ and $w_{equal\ performance}$. Figure 6.6 also shows the WLS regression model with the runtime weighting metric multiplied with the user error metric $w_{runtime} \times w_{equal\ performance}$, as a visual confirmation that there is no clear improvement when incorporating user runtimes into the regression.

Considering these results, the WLS regression model used to assess the spatial distribution of damage uses the $w_{equal\ performance}$ metric. This WLS model will overestimate ground-validation damage at lower damage indicator values and underestimate ground-validation damage at higher damage indicator values with experiment 2 damage indicator results. Conversely, experiment 3

will *generally* underestimate ground-validation damage for all damage indicator values. However, the term generally is applied here, as there is still variation because of the multiple volunteer assessments carried out for each image, which will be discussed in the following section.

6.1.2 Spatial Distribution of Damage with Multi-Pass Aggregation

The spatial distribution of crowdsourced damage indicators versus the ground-validation damage can be visually compared through aggregating the results of the regression analysis. Recall that each image was assessed a minimum of three times by different users, resulting in at least three associated damage indicator values per image. The regression predicted mean CDF values from these damage indicators must be aggregated into one value in order to plot the spatial distribution of crowdsourced damage.

Three methods of aggregation for the predicted mean CDF values were tested and compared with the ground-validation mean CDF values:

1. The predicted mean CDF value of the highest performing user per image
2. The maximum of the multiple regression predictions of mean CDF per image
3. The average of the multiple regression predictions of mean CDF per image

Using the regression prediction of the highest performing user means we map the predicted mean CDF from the damage indicator provided by the user with the highest user performance, or individual regression coefficient $\beta_{1,j}$. Taking the maximum of the multiple assessments for an image ensures that the predicted and actual ground-validation damage (Figure 6.8) are as close as possible, as many users underestimate building damage.

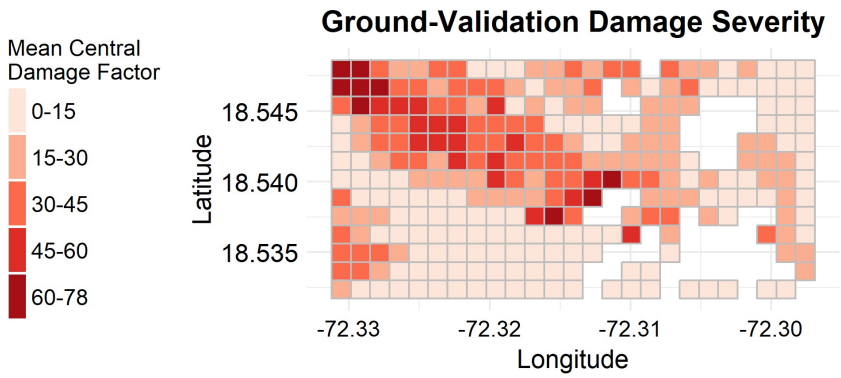


FIGURE 6.8: Spatial Distribution of Ground-validation Damage Quantified as Mean Central Damage Factor per image

The spatial distribution of crowdsourced damage aggregated by using the prediction of the highest performance users is shown in Figure 6.9. Initially, the predicted ground-validation damage from experiment 3 encompasses a wider range of mean CDF values. The measure of interest, though, is whether these predicted Mean CDF values are close to the ground-validation Mean CDF values shown in Figure 6.8.

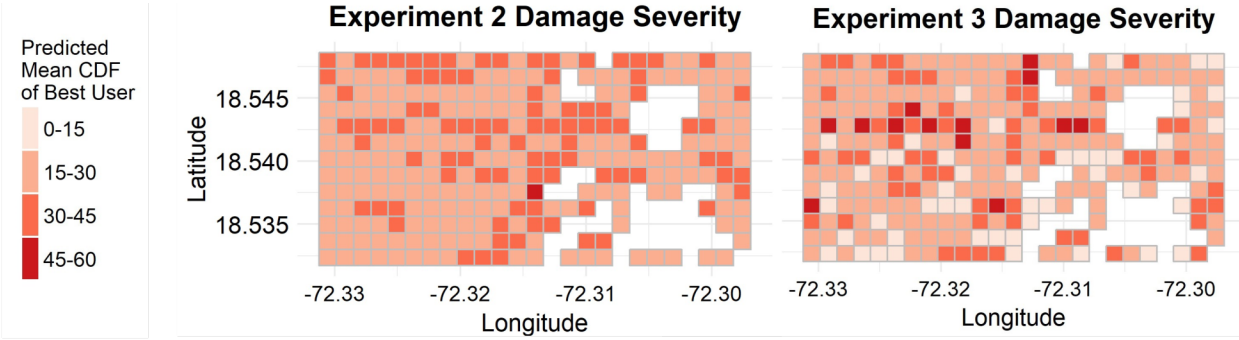


FIGURE 6.9: Spatial Distribution of Crowdsourced Damage if Using the Predicted Mean CDF value of the Highest Performing User

The maps of crowdsourced damage using the three aggregation methods were thus compared with the ground-validation damage map shown in Figure 6.8 to determine which aggregation method is most suitable. This was done by first determining the residual between the aggregated predicted mean CDF and actual mean CDF value for an image. The distribution of residuals for one aggregation method, using the prediction of the best user, is shown in Figure 6.10.

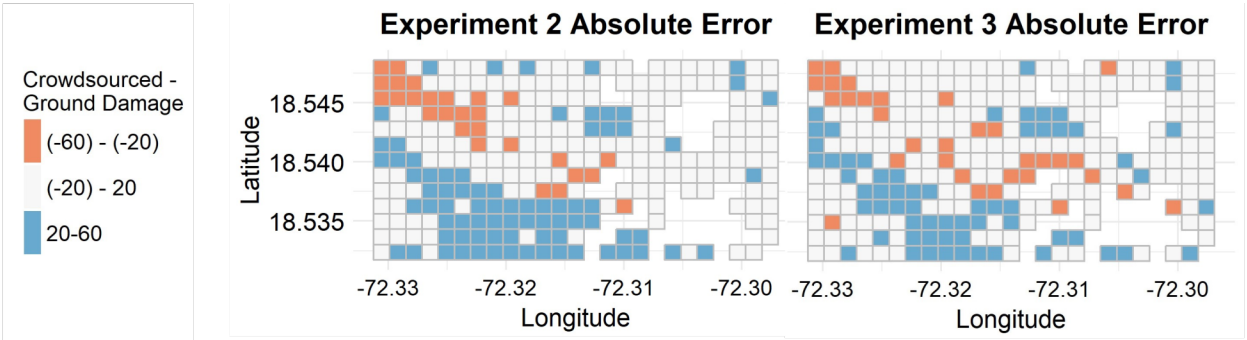


FIGURE 6.10: Spatial Distribution of Residuals Between Predicted Mean CDF value of the Highest Performing User and Actual Mean CDF from Ground-Validation Data

The aggregation methods could then be compared by evaluating the mean and standard deviation of the residual, shown in Table 5. Aggregation by employing the prediction made from the contribution of the best performance user per image results in the lowest mean residual between predicted and actual ground-validation damage. Generally, the standard deviation in the residual is similar for the three aggregation methods for each experiment. This measure of variance is also less than that of comparing residuals without aggregation, which signifies that aggregation reduces the uncertainty in the deviance between predicted and actual ground-validation damage.

TABLE 6.5: Mean and standard deviation of error metric of five least squares linear regression models using the untransformed damage indicator of five random data subsets to validate each weighting parameter

		Residuals (Crowdsourced - Ground Mean CDF)	
Experiment	**Aggregation Method**	**Mean**	**Standard Deviation**
Damage Ranking (2)	All Users (no aggregation)	3.7	19.4
	Maximum	13.2	17.4
	Mean	7.8	17.4
	Best Performance User	5.7	17.8
	All Users (no aggregation)	6.0	20.0
Damage Comparison (3)	Maximum	14.1	18.7
	Mean	6.1	18.1
	Best Performance User	4.4	18.9

6.1.3 Discussion of Results

Throughout this exploration and analysis of the damage indicator dataset, we addressed three main objectives. We first wanted to discover which attributes of a post-earthquake satellite image that the crowd is visually able to identify, and if user performance is consistent throughout the

crowd. Understanding the basic abilities of the crowd to detect building damage was addressed with the initial exploratory analysis of the damage indicator dataset. Second, we determined whether the relation between the user-provided indicators and ground-validation damage could be improved by building several regression models. Third, a comparison between the results of experiments 2 and 3 can be made with the regression models and the spatial distribution of user-provided building damage.

Based on solely the initial exploration of the damage indicator results, we can conclude that contributors viewing an image are indeed detecting a combination of building damage and building density in an image. However, with all the user's responses there is still a relatively low positive correlation between the user-provided indicators and both ground-validation building damage and density per image (values between 0.1-0.3). This is because of the the wide distribution in ground-validation values compared to each damage indicator value. Overall, the crowd generally underestimates damage, but there are some users who assess damage more accurately.

Using several weighted least squares regression models, we tested whether characteristics like a user's performance (learned through their assessments), the time taken to complete a task, or the building density in an image could be used to improve the overall performance of the crowd. For both experiments, weighting by user performance, especially weighting each user's contribution equally, improves the relationship between crowdsourced and ground-validation damage severity. However, weighting also increases the variability in how close the regression model is to the anticipated relationship. On the other hand, weighting by runtime for experiment 2 and building density for both experiments does not improve this relationship.

In addition, we tested specific transformations of the damage indicator values, through aggregation or ordinal scaling. In all cases, transforming the user-provided indicators with ordinal scaling does not provide any benefit.

The aggregation of the multi-pass assessments reduces the uncertainty in the residual between the crowdsourced predicted and actual damage. The lowest average residual occurs when using the prediction of the highest performing user for each image. It can be seen from the spatial distribution of the residuals, that the regression and aggregation still result in an overestimation of low damaged areas and underestimation of high damaged areas.

Comparisons between experiments 2 and 3 can be made in each step of the above analysis. The shape of the distribution of the crowdsourced damage indicator values from experiment 2 more closely matches that of the ground-validation damage values in the AOI. Furthermore, the baseline OLS regression without weighting for experiment 2 exhibits a lower error for experiment 2.

When incorporating the user performance weighting parameters, though, experiment 3 encompasses a slightly more representative range of ground-validation mean CDF values. Therefore, the results of experiment 2 are more representative of ground-validation damage initially, but experiment 3 improves when weighting by user performance.

In the future, this analysis could be improved with some minor changes. Given the wide variation in results from the crowd, linear regression models will inherently over and underestimate damage. Therefore, it would be prudent to test other nonlinear modeling methods, such as nonlinear, parametric splines, or some clustering methods. The described weighting method would still be applicable for alternative parametric models. Furthermore, collecting performance data on each user from a set of training images would make the performance weighting metric more consistent for every user. It would also be interesting to collect other attributes about contributors, such as work experience or age, to determine if that has any effect on user performance. If other user attributes are collected, they could then be incorporated with a multivariate interaction model, as opposed to the univariate regression models presented in this study.

Finally, this analysis should be benchmarked against data from the experiment 1 results, which would represent the typical method of tagging damage levels for individual buildings in an image. This would be the next step, if more responses are obtained.

7 Results from Damage Comparison Dataset

The damage comparison dataset was analyzed through two methods: network analysis and bayesian updating. The goal of the network analysis was to provide an example of using this method to sort a set of paired comparison data (two compared images) into an entire group of images ordered by damage severity. Bayesian updating was used to incorporate prior distributions of the damage severity of an image into the final estimate of damage using the series of comparisons for each image.

7.1 Network Analysis

Network analysis was one method used to structure and analyze the damage comparison dataset. Networks are useful for representing relationships between components, in this case a relationship is defined by a user's comparison of two images.

Networks and graphs consist of nodes (also called vertices) and edges. They can be weighted or unweighted (i.e. the edges may have a weight associated with them, such as capacity in a road network) and directed or undirected (e.g. traffic can travel in both directions or just one). In our network, images are represented by nodes and each paired comparison can be represented by a directed edge.

Directed networks can contain cycles and self-loops. A cycle in a directed network is a "closed loop of edges with the arrows on each of the edges pointing the same way around the loop" (Newman 2010). A self-loop consists of an edge which begins and ends at the same node, without passing through any intermediate nodes. A directed network without cycles or self-loops is called a directed acyclic graph or a DAG. Acyclicity usually denotes a sequence in a different dimension (e.g. time) and this can be found using standard sorting algorithms. In the network created from

the damage comparison dataset, the sequence represents increasing or decreasing levels of visible damage.

Networks can vary in density, ranging from dense (many edges between pairs of nodes) to sparse (few edges between pairs of nodes). The density of a network is an indication of the level of connectivity within the network. Numerous measures can be used to define the level of connectivity, such as the degree of a node (number of incoming or outgoing edges from a node) to the network density (mean degree divided by number of nodes in network). Before any image comparisons are completed, the network for this dataset would be sparse, containing nodes for each image in the dataset but no edges. As comparisons are completed, edges are added, and the network becomes denser with more information about the relationship between pairs of images in the dataset.

7.1.1 Network Construction

In this network, images are represented by nodes and each comparison of a pair of images is represented by an edge from the most damaged to least damaged image. Figure 7.1 illustrates how edges represent the results from paired comparisons. An edge from A to B indicates A is more damaged than B.

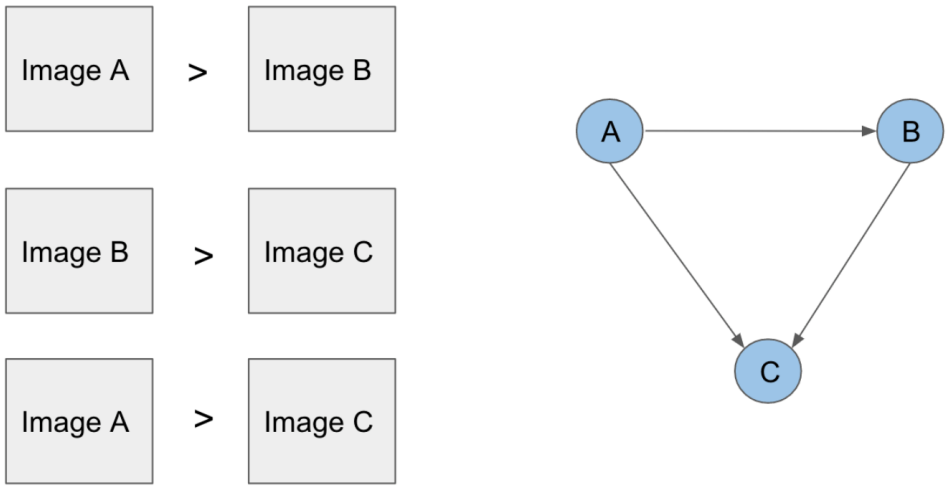


FIGURE 7.1: How edges are created from paired comparisons. An edge begins at the node with higher damage and points toward the node with less damage.

The edges can be weighted in a number of different ways. Where there are multiple assessments of the same pair of images, the weight could represent the total number of assessments completed. Where there are conflicting responses, the weight could represent the percentage of responders that have chosen image A as more damaged than image B. The weight could also be some combination of other relevant metrics or information. For example, a distance metric related to the spatial distance between two images. For this analysis, however, the edges were kept unweighted.

Networks can be visualized in different ways. Figure 7.2 shows three networks: networks using subsets of 21 images and 37 images as well as a network created from the entire dataset. These networks only contain edges that represent correct comparisons.

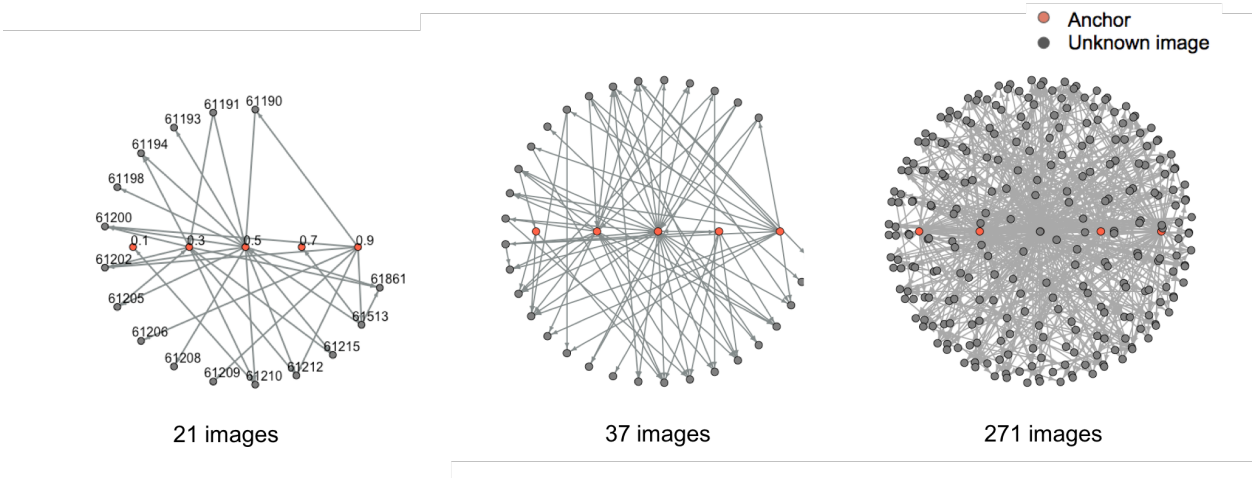


FIGURE 7.2: Networks created from damage comparison data. The first two networks are from subsets of 21 and 37 images. The network on the right shows the entire dataset. Anchors are represented by the pink nodes and unknown images are shown in grey. Refer to Section 5.3 for description of the dataset used, including anchors and unknown images.

Networks are a convenient way to structure this dataset as it captures the relationship between the different images. There are also a number standard analysis techniques and algorithms that can be applied to analyze data contained in a network.

For the purposes of this study, we aim to understand the overall distribution of damage (i.e. how many images show high, moderate, low and no damage) as well as where the damage is located. To do this, we sort the images in the network into order from most damaged to least damaged, based on the individual comparisons. Then, by using the "anchors" with known levels of damage (see Chapter 5 for detailed description of anchors and the images used in the study), the number

of images in each "bin" (i.e. damage range) are calculated. The results from sorting and binning analyses are detailed below. All results have excluded responses where the user selected "same".

7.1.2 Sorting Images Based on Damage

The images can be sorted by inferring their order based on the constructed network using standard network topological sorting algorithms. This study used the *toposort* function from the textsfR *igraph* package. Figure 7.3 explains the logic behind this sorting algorithm. The function can use an "in" or "out" method for sorting: "in" prioritizes nodes with no outgoing edges (i.e. images with less damage), while "out" prioritizes nodes with no incoming edges (images with more damage) - recall that an edge goes *from* a more damaged image *to* a less damaged image. The "out" method proved the most accurate and was used for all analyses.

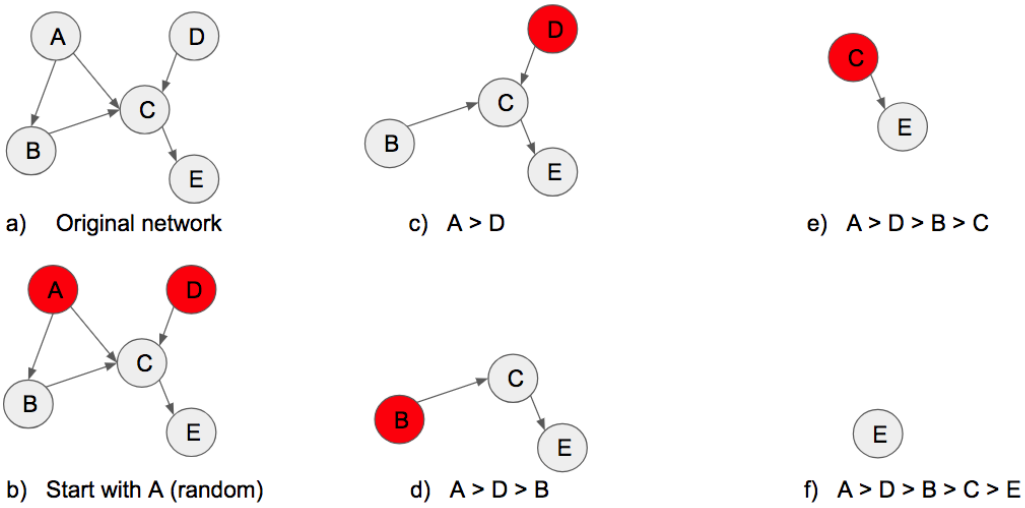


FIGURE 7.3: Example of how the topological sorting algorithm works using an example network (a). The algorithm prioritizes nodes without any incoming edges ("in" method), in this case nodes A and D (b). Node A is randomly chosen to begin, and it is removed from the network along with all attached edges. Node D is then removed as it has no incoming edges (c), followed by Node B (d), C (e) and finally E (f), resulting in the sorted order for this network.

The sorted order is plotted in the subsequent figures to visually show how well the sorting performed. These figures show a sequence plot where the horizontal axis represents the image's position in the sorted order (from greatest to least damage) and the vertical axis shows the ground-validation mean CDF of the image.

True sorted order

Before running the sorting algorithm, the true sorted order was calculated to use as a baseline. This is shown for the entire network as well as a 37-image subset in Figure 7.4 below.

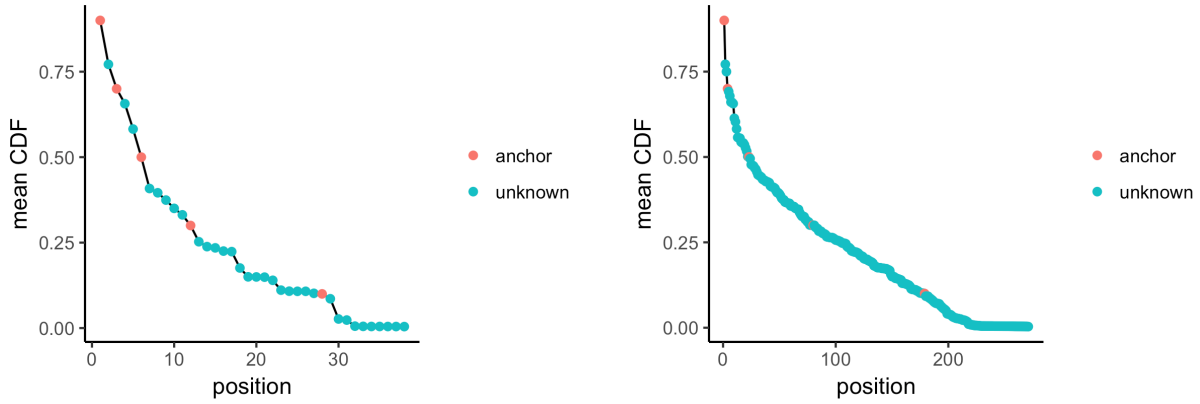


FIGURE 7.4: The true sorted order (from greatest to least damage) for all images in the experiment 3 results files. The figure on the left is for a 37-image subset while the figure on the right is the entire network. Each dot represents an image showing the ground-validation mean CDF and its position. Anchors are shown in pink.

Predicted Sorted Order Results

Using the topological sorting algorithm, the following predicted sorted orders were obtained. The following results are for networks containing only correct responses and all "same" responses have also been excluded.

Figure 7.5 shows the results for both the 37-image subset and the entire network with the true sorted order shown for reference. The entire network figure (right) only contains comparisons with anchors while the subset (left) includes comparisons with anchors as well as some 'random' comparisons.

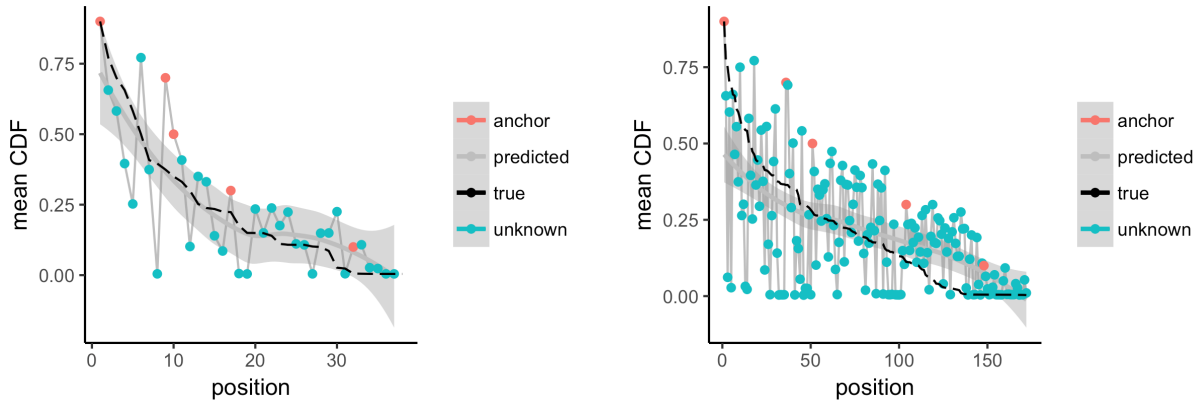


FIGURE 7.5: The true sorted order (from greatest to least damage) for all images in the experiment 3 results files. The figure on the left is for a 37-image subset while the figure on the right is the entire network. Each dot represents an image showing the ground-validation mean CDF and its position. Anchors are shown in pink.

The general trend is correct (i.e. the ground-validation mean CDF decreases as the image position increases), however there are a number of images in the incorrect position. This is due to the way the sorting algorithm works. As described in Figure 7.3b, when there are multiple nodes with no incoming edges, the algorithm selects one of these by random. To illustrate this, consider two example comparisons: 1) an image with a mean CDF of 0.25 (let's call it image Y, represented by node Y in the network) correctly compared to anchor 1 (mean CDF of 0.1) and 2) an image with a mean CDF of 0.95 (let's call it image Z represented by node Z) correctly compared only to anchor 5 with a mean CDF of 0.9 (represented by node A5). The edges from these two comparisons would originate from the unknown images (Y and Z) and end at the anchors (A1 and A5). Neither nodes Y or Z would have any incoming edges, only outgoing edges. Based on this information, we know that image Z shows more damage than image Y because we know the level of damage in the anchor images, however, the algorithm would consider them in the same way since neither have an incoming edge. These errors would reduce as more comparisons are made and the network density increases. For example, if the image represented by node Y were also compared to anchor 5, then there would be an edge from node A5 to node Y. Since it now has an incoming edge, it would not be selected until much later in the sorting process (i.e. at least until node A5 is removed from the network).

As this algorithm does not differentiate between anchors and unknowns it is encouraging to note that the anchors (shown in pink) have been placed in the correct order in both figures of Figure 7.5.

The above results only contain correct responses. If a network is created using only incorrect responses, the predicted order shown in Figure 7.6 is obtained. As expected, the trend is no longer correct. This highlights the impact of incorrect user responses for this method of analysis.

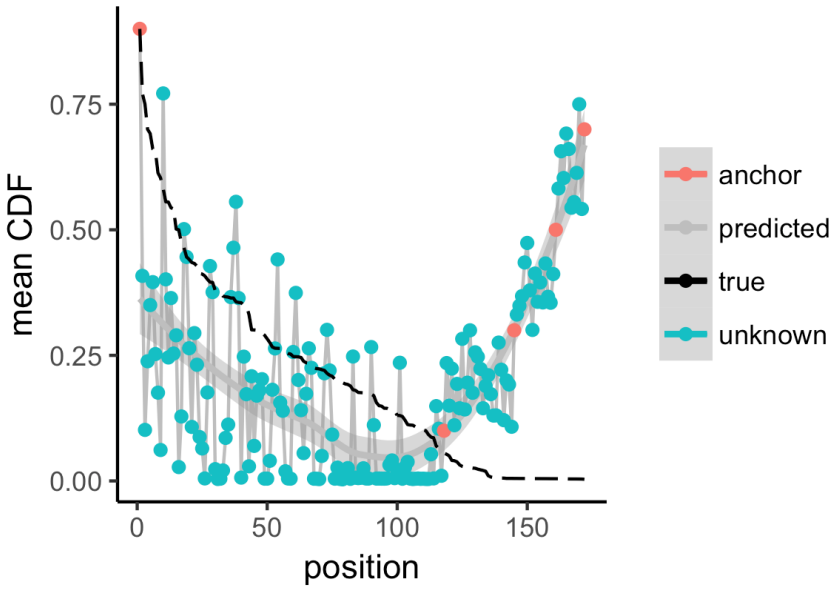


FIGURE 7.6: The predicted sorted order (from greatest to least damage) for the entire dataset using only incorrect responses.

Cycles

In the above results, user responses were filtered for correct or incorrect responses. When all responses are combined to create a network, the sorting algorithm no longer works because the network now contains one or more cycles.

Sorting algorithms only work for acyclic networks because when a cycle exists, there can be no real or underlying sorted order. When the sorting algorithm is run on a cyclic network, it only returns the images prior to the cycle before failing. Figure 7.7 below shows the output for this entire network.

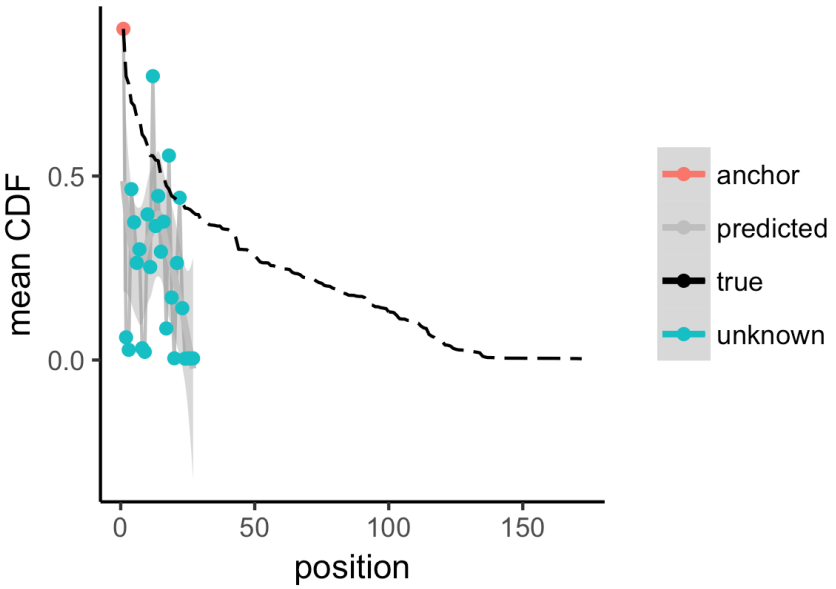


FIGURE 7.7: The predicted sorted order (from greatest to least damage) for the entire dataset using all responses (except same). The algorithm only returns 27 images as it is no longer acyclic, causing the algorithm to fail.

In reality, assuming there are no images with exactly the same level of damage, the true network should not contain any cycles since the images can be sorted into a true order from most to least damage. Even if there are images with the same level of damage, they could be assigned an arbitrary order as the exact sorted order does not matter.

Cycles are introduced due to errors in user response. This may be caused by two different users disagreeing on a single comparison, creating a cycle between two nodes. A cycle may also be created due to one incorrect response. For example, if images A, B, C and D have the following relationship: A > B > C > D, and images A, B and C have been correctly assessed and placed in this order, one incorrect assessment that D > A would create a cycle.

Aside from the difficulty in assessing these images, error or disagreement in user responses may indicate that the images contain similar levels of damage. This would be useful information. It may also indicate that more comparisons are required. This is something that could be dynamically identified and used to prioritize images for further comparison during a live project.

The level of error in user responses highlights that correctly identifying the level of damage in a satellite image is a difficult task, therefore efforts to improve user accuracy alone are unlikely to

resolve this issue. The issue of cycles in the network must be addressed in order for this approach to be useful in a real-world scenario.

Attempts were made to address this by finding cycles, merging them into a mega node and then running the sorting algorithm again. This is possible, however, as the networks tested contained cycles with 100 nodes, this was of limited value. Further work to address this could involve defining rules for conflicting assessments (e.g. rather than incorporating every user response as an edge, include the edge that represents the majority of user responses). To solve larger cycles, other information could be incorporated to refine assessments. Examples of such information could include: the physical location or distance between images (if another model or source that provides some information about spatial distribution of damage is available), user reliability metrics to prioritize experienced and reliable user responses, or simply highlighting the area for more refined assessment.

Random vs anchor comparisons

The majority of the image comparisons contained on anchor image of known damage. In addition to comparisons with anchors, each unknown image was also compared to a random unknown image. Figure 7.8 shows predicted sorted orders based on these different types of comparisons.

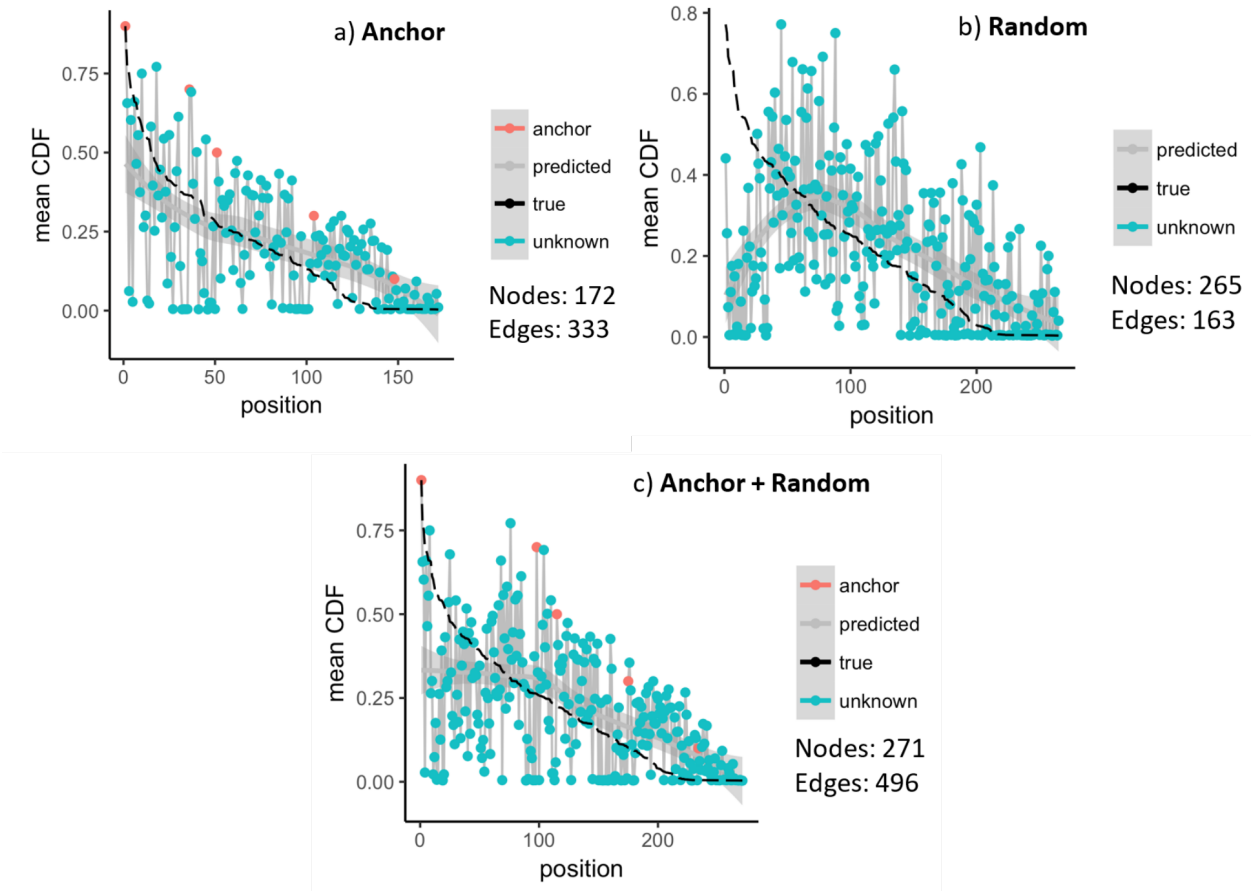


FIGURE 7.8: The predicted sorted order (from greatest to least damage) for the entire dataset of correct responses, using only comparisons with anchors (a), only random comparisons (b) and both anchor and random comparisons (c).

It should be noted that Figures 7.1a, b, and c contain different numbers of nodes and edges as describe in the figure. These figures show that incorporating random comparisons resulted in a less accurate sorted order prediction, compared to only using comparisons with anchors. This is because the network composed of only random comparisons is much less connected than the anchor only network, with approximately 0.6 edges per node, compared to almost 2 edges per node in the anchor only network. As previously discussed, fewer edges in a network due to fewer comparisons leads to a greater number of images placed in the incorrect order, reducing the overall accuracy of the predicted sorted order. Figure 7.1c is a combination of 7.1a and 7.1b so it follows that adding the random comparisons with a less accurate predicted sorted order would decrease the overall accuracy of the prediction compared to only using anchor comparisons as shown in 7.1a.

7.1.3 Estimating the Distribution of Damage

One of the key objectives of the study is to estimate the overall distribution of damage. Placing the images in a relative sorted order alone does not provide this information. By using the known anchors to delineate different levels of damage, the overall distribution of damage can be estimated. The different levels of damage are referred to as 'bins' and are described below.

True bins

To align with experiment 3 modified data (i.e. the comparisons that were back calculated, see Section 5.3, images were assigned to one of six bins. True bins were defined as shown in Table 7.1 and the true distribution of damage is shown in Figure 7.9. Note that there were no images in bin 6.

TABLE 7.1: Bins and corresponding mean CDFs

Bins	Mean CDF Range
1	X <= 0.1
2	0.1 <X <= 0.3
3	0.3 <X <= 0.5
4	0.5 <X <= 0.7
5	0.7 <X <= 0.9
6	0.9 <X

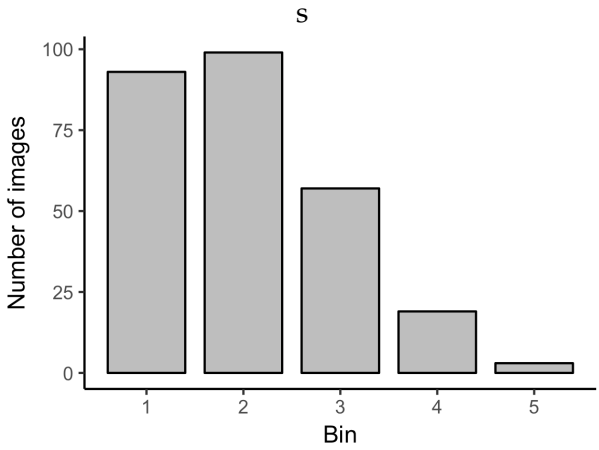


FIGURE 7.9: True distribution of damage across bins 1 to 5 (least to greatest damage). The majority of images showed minor to no damage.

Methodology to Predict Bins

An image is placed into a predicted bin based on its position in the sorted order, with respect to the other anchors. For example, using the example positions of anchors shown in Table 7.2, any image in position 1 to 9 would be assigned to bin 6, any image between 10 and 39 would be assigned to

bin 5, and so on. Recall the sorted order outputs images from most to least damage, so the image with the lowest position would be predicted to have the greatest damage.

TABLE 7.2: Example anchor positions

Anchor	Position
0.9	10
0.7	40
0.5	80
0.3	100
0.1	170

Since this method of assigning predicted bins relies on the anchors being in the correct order, this analysis was only completed for networks where this criterion was satisfied. These networks included only anchor comparisons filtered for correct responses, anchor and random comparisons filtered for correct responses and the 37-image subset with both anchor and random comparisons filtered for correct responses.

Predicted Bin Results

The predicted bin results are shown in Figures 7.10 - 7.12. The histograms on the left show the predicted overall distribution (blue) compared to the true distribution (pink). The heat maps on the right show the accuracy of the bin prediction at the image level (i.e. whether the correct images were placed in the correct bins) by showing the number of images binned in each predicted bin versus their true bin.

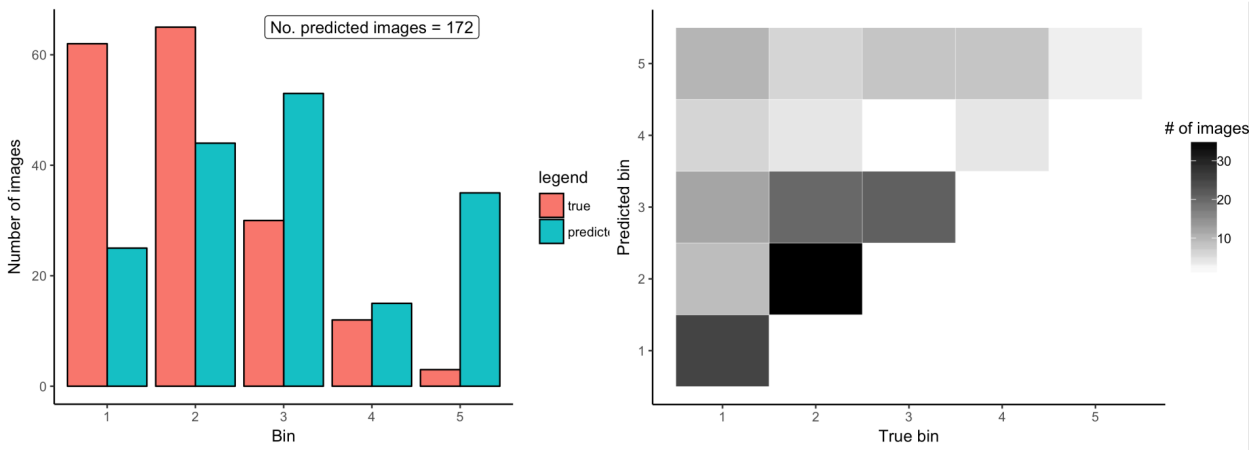


FIGURE 7.10: Histograms and heatmaps for network containing correct anchor only comparisons. 51% of images were correctly binned.

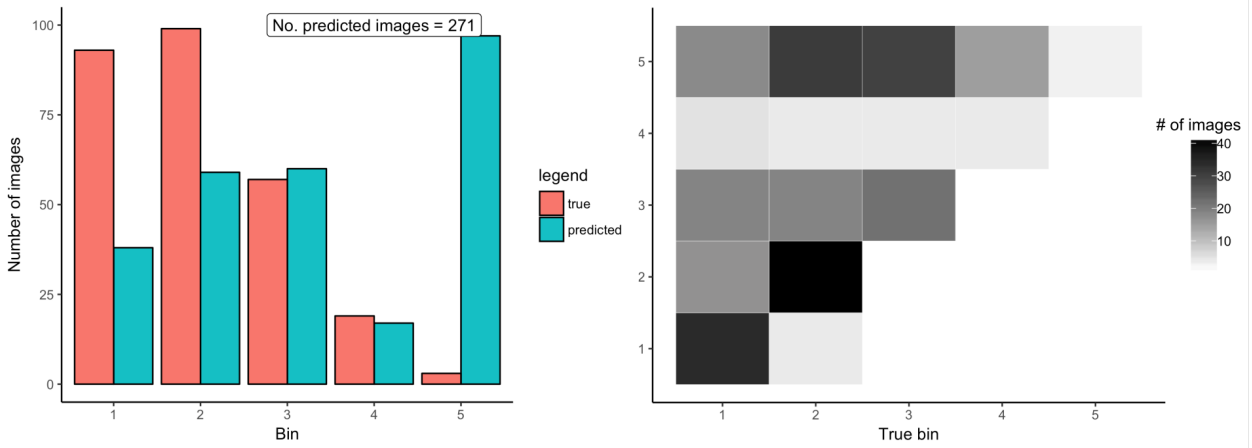


FIGURE 7.11: Histograms and heatmaps for network containing correct anchor and random comparisons. 38% of images were correctly binned.

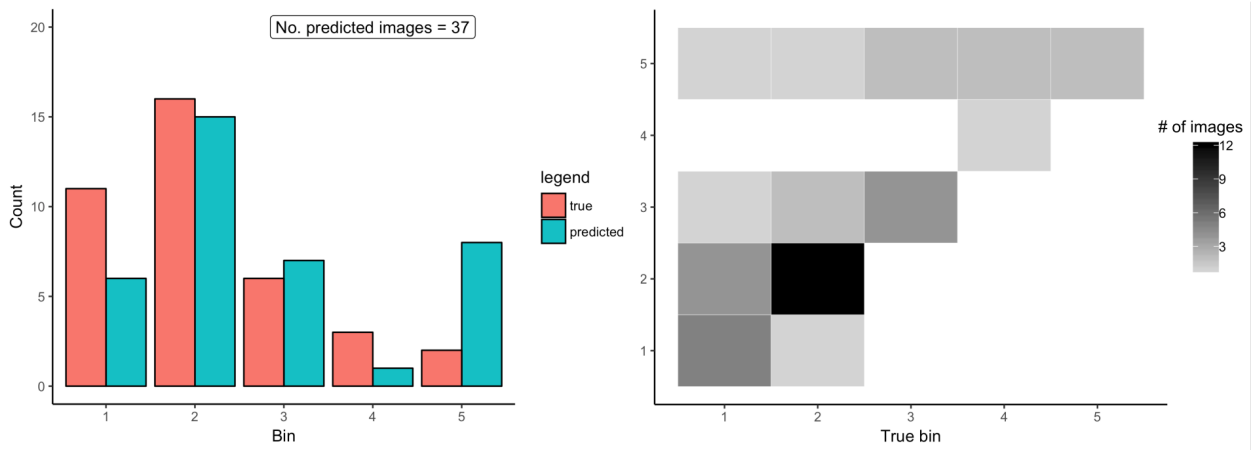


FIGURE 7.12: Histograms and heatmaps for network containing correct anchor and random comparisons for the 37-image subset. 63% of images were correctly binned.

From the histograms presented for the three different networks, the results overpredict the overall distribution of damage. This is most significant in bin 5 where the predicted number of images is much higher than the true number of images. This may indicate a systemic overestimation of damage by users (perhaps they are misinterpreting signs of damage or perhaps they are unsure and choose higher levels of damage to be conservative), or it may be an artifact of the sorting algorithm. One way to reduce the influence of the sorting algorithm is to try to ensure the majority of nodes have both an incoming and outgoing edge (i.e. they are compared to images that show more and less damage). This would partially address the issue outlined in the example of nodes Y and Z described above, where the algorithm prioritizes nodes with no incoming edges.

While understanding the overall distribution of damage is a key outcome, it is also important to analyze whether the correct images have been placed in the correct bins. If all images were placed in the correct bin, the heatmaps shown above would only show a diagonal line. Instead it reflects the general trend of overestimation of damage apparent from the histograms.

The network containing only correct anchor comparisons (Figure 7.10) correctly bins 51% of images, while the networking containing both correct anchor and random comparisons (Figure7.11) correctly bins only 38% of images. This aligns with the sorting results presented in Figure 7.8, since the binning is dependent on the sorted order. The random comparison network is much less dense than the anchor only comparisons and this influences the accuracy of the prediction.

It is interesting to note that the 37-image subset resulted in the highest percentage of correctly

binned images (63% versus 51% and 38%). This subset was selected to be relatively well connected based on correct comparisons so this likely influenced the results. In addition, the number of correctly binned images in all networks includes the 5 anchors that were by design correctly binned.

7.1.4 Limitations and Future Work

For this method to be useful in a real-life setting (i.e. running the project to assess damage after a disaster, rather than a test case where the actual damage is known), the issue of cycles due to user error or disagreement needs to be addressed. Potential ways to address this may include: assigning all in the cycle the same position and prioritizing for further assessment or incorporating other information do resolve conflicting assessments such as user reliability information or known information about the spatial distribution of damage.

The sorting algorithm prioritizes nodes with no incoming edges irrespective of whether the edge is attached to a known anchor. Ensuring the majority of nodes contain both an incoming and outgoing node would reduce the impact of this prioritization. Another approach would be to develop a custom sorting algorithm that incorporates the known levels of damage in the anchors to improve the accuracy of the sorting prediction. Perhaps the simplest way to do this would be to separate the network into subnetworks based on each node's relationship to anchor image nodes and run the sorting algorithm over the subnetworks, before then combining the outputs of each subnetwork.

More sophisticated network analysis techniques could be tested. One example could be to use community detection or clustering methods to identify groups of images with similar levels of damage.

There is likely an optimal range for the number of comparisons required to achieve a desired or suitable level of accuracy. Many simulations could be run where the network is updated as increasing numbers of assessment are received to determine these optimal thresholds (these may be based on the number of assessments or the number of assessments in agreement).

Perhaps the most promising aspect to this approach is the ease in which it handles new information and the speed at which analyses can be completed. This lends itself to being dynamically updated during a real-life project and further work should be undertaken to understand how the various network metrics (such as density, node degree etc.) could be used to prioritize pairs of

images for assessment or reassessment to speed up and reduce the overall number of assessments required.

7.1.5 Conclusion

Networks are an effective way to structure and analyze the damage comparison data. Standard network analysis algorithms can be applied to analyze the dataset, one example is the topological sorting algorithm used in this study.

The results show that as the number of comparisons increases, and the network becomes better connected, the accuracy of the predicted sorted order increases. There is likely an optimal range for the number of comparisons required, beyond which the increase in accuracy is no longer warranted. Running simulations of user responses would provide insight to these optimal conditions.

While there are a number of obstacles that need to be addressed for this approach to be useful in a real-world setting, this approach shows promise. One particular advantage is that this model is easily updated as new information is received (edges and nodes can be added very easily). Being able to add to the network and quickly rerun the analysis could be useful during the data collection phase of a live project allowing for images to be prioritized for assessment based on measures such as the density of the network or level of user agreement.

7.2 Bayesian Updating

We present a method, predicated on Bayesian updating of a prior damage distribution, for assimilating individual contributors' assessments of damage in an area. The result is a posterior damage distribution that (1) indicates the most likely level of damage in the area and (2) makes explicit the level of uncertainty in that damage estimate. Decision-makers' need for both an estimate of damage and of the uncertainty in that estimate emerged from conversations in the team workshops described in Section 1.3.2, as well as through interviews in the demand survey, described in Section **??**.

7.2.1 Motivation

Conversations within the project team and with practitioners made clear that estimates of the damage in an area need to be accompanied with measures of the uncertainty in that estimate to

properly inform decision-makers. Bayesian updating of a prior damage distribution allows for uncertainty quantification in addition to a mean estimate of damage. Uncertainty quantification is important especially at the early stages of the post-disaster timeline since damage estimates may factor heavily into loss estimates used for recovery and reconstruction planning, as well as requests for external assistance. Bayesian updating of a prior damage distribution with evidence – in this case, volunteer damage assessments – is also a straightforward method for aggregating crowdsourced data (Booth et al. 2011).

7.2.2 Methodology

The metric of interest is the mean building damage in an area, henceforth referred to as the mean central damage factor (mean CDF), which is described in Section 5.1. Earlier work has shown that the beta distribution is a good model for the distribution of building damage within an area (Lallemant and Kiremidjian 2015). Thus, we will define a prior beta distribution on the mean CDF of each area of interest. This prior distribution can be informative or uninformative. From experience, structural engineers who work in a particular region tend to have a general sense of the expected damage after a particular earthquake. We can leverage that understanding to define a prior distribution on building damage in an area. For example, if we generally don't expect high levels of damage in an area, the informative prior may have a mean closer to 0 than to 1. We can also use an uninformative prior, e.g. beta(1,1).

There are several ways to update the prior damage distribution. In this report, we implement and compare two approaches. In the first approach, we implement a simple scheme in which we treat alpha and beta (the parameters of the beta distribution) as pseudocounts, as described in Section 2.3.2 of Kochenderfer et al (Kochenderfer et al. 2015). We increment alpha by 1 if the respondent indicates that an image has higher damage than the anchor image. Conversely, if the respondent indicates that an image has lower damage than the anchor image, we increment beta by 1. If the respondent indicates that the two images show the same level of damage, we increment alpha and beta each by 0.5.

There are two drawbacks to this simple updating scheme. First, and perhaps most seriously, it does not take into account the information we know about the anchor image. For applications in which the anchor image shown to volunteers is random – that is, a setup in which an image with unknown damage is compared to a series of anchor images in no particular order – this updating scheme may not be effective. However, our experimental setup results in each image with unknown damage being compared in a logical sequence to anchor images. A second drawback

of this approach is that – as implemented here – it does not weight the contributions of different volunteers with varying skill levels. Weighting the responses of different volunteers according to their performance on a training set or assessment may be a promising direction for future work.

In the second approach, we implement a more complex updating scheme in which we carry out the following general steps for each image with an unknown damage metric (in this case, mean CDF):

1. Define the image's prior damage distribution, a beta distribution with parameters ai and bi. If using an uninformative prior, let a_i, $b_i = 1$.
2. Truncate the image's prior probability density function at the mean CDF of the anchor image to which it has been compared.
3. Determine the median mean CDF of the upper or lower half (depending on the particular response) of the truncated distribution, x_i.
4. Update alpha and beta so that $a_i + 1 = a_i + w_j x_i$ and $b_i + 1 = b_i + wj(1 - x_i)$, where w_j is an optional response-weighting parameter. These updated parameters define the posterior beta distribution over the unknown damage in the image of interest.
   (a) If including responses in which the volunteer indicates that the level of damage in the image of interest is the "same" as in the anchor image to which it is being compared, for those "same" responses, use the following update: $a_i + 1 = a_i + 0.5 w_j x_i$ and $b_i + 1 = b_i + 0.5(1 - x_i)$.
5. Repeat steps 1 through 4 for all responses involving the image of interest.

Figure 7.13 depicts how this approach might look for one response in which the volunteer indicated that the image of interest showed more damage than the anchor image to which it was being compared.

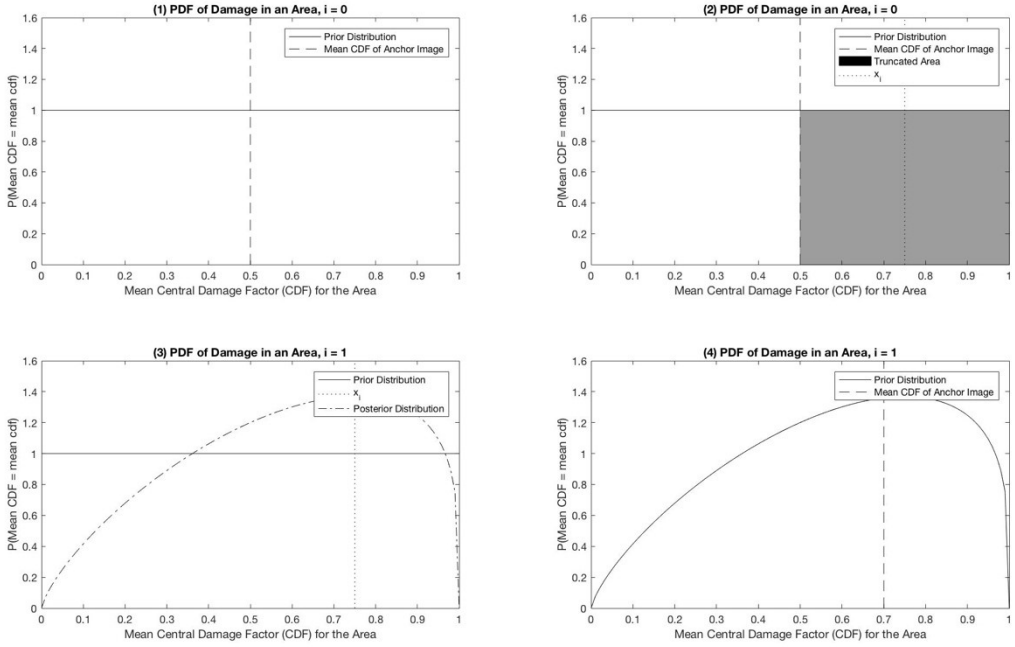


FIGURE 7.13: Basic sketch of the steps involved in the more complex Bayesian updating approach. Subplot 1 (top left) shows that the mean CDF of the anchor image is known to be 0.5. The prior distribution over the unknown damage in the image of interest is uninformative, i.e. beta(1,1). In subplot 2 (top right), we see that the volunteer response indicated that the image of interest had a higher level of damage than that in the anchor image. In subplot 3 (bottom left) we can compare the prior and posterior distributions over the damage in the image of interest. And in subplot 4 (bottom right), we see that the posterior distribution from subplot (3) has become the prior distribution for the next comparison, in which the new anchor image will have a mean CDF of 0.7.

As we show in the following section, this more complicated Bayesian updating approach yields better results than the simpler approach described earlier. That is, for a given image, the posterior damage distribution that results from the more complex updating approach tends to have a mean closer to that of the image's actual mean CDF. The underlying reason for such improvement is likely that this approach takes into account the information we know about the anchor images, whereas the simpler approach does not leverage this information.

One complication that we dealt with in both approaches was that volunteers had the option to respond that two images showed the same level of damage – and often availed themselves of

this option. When discussing results, we specify whether they were derived from the full set of responses (that is, including "sames") or from a reduced set (that is, excluding "sames").

We intend in this and following sections to present the outlines of a potential method for aggregating crowd-sourced damage assessments in a useful way and note that developing a statistically rigorous Bayesian updating methodology for this particular application remains a topic for future research. Our preliminary results indicate, however, the promise of such an approach.

7.2.3 Results of a Simple Bayesian Updating Scheme

In this section, we briefly present the results that follow from using a simple Bayesian updating scheme, described in the previous section. In this approach, the parameters alpha and beta are treated as pseudocounts.

One way to consider the accuracy of this approach is to compute the percentage of images whose actual damage level is within some number of standard deviations of their predicted damage levels, taken as the mean of their posterior beta distributions. These metrics are presented in Table 7.3. In short, these results are comparable, but not as good as those obtained from the more complex Bayesian updating scheme, discussed in the following section.

TABLE 7.3: The percentage of images whose actual damage level was within 1, 1.5 or 2 standard deviations of their predicted damage levels, taken as the mean of their posterior beta distributions.

Actual damage relative to predicted damage	**Images**
Within 1 standard deviation	16.30%
Within 1.5 standard deviations	31.50%
Within 2 standard deviations	40.20%

Another way in which we can consider the accuracy of this approach is to compute the raw error in the damage estimate of each image, computed as the actual damage level in the image minus the predicted damage level in the image. The predicted damage level of an image is the mean of its posterior beta distribution. In Figure 7.14, we can see that this approach generally overestimates damage but without any particular pattern, in contrast to the results shown in the following section. That is, this approach would suggest that the crowd's ability to distinguish damage in an image does not improve as the level of damage in the image increases, in marked contrast to the results shown in the following section.

While this approach is extremely straightforward, we find that the more complex updating scheme yields more accurate and meaningful results. Therefore, we move to a discussion of those results.

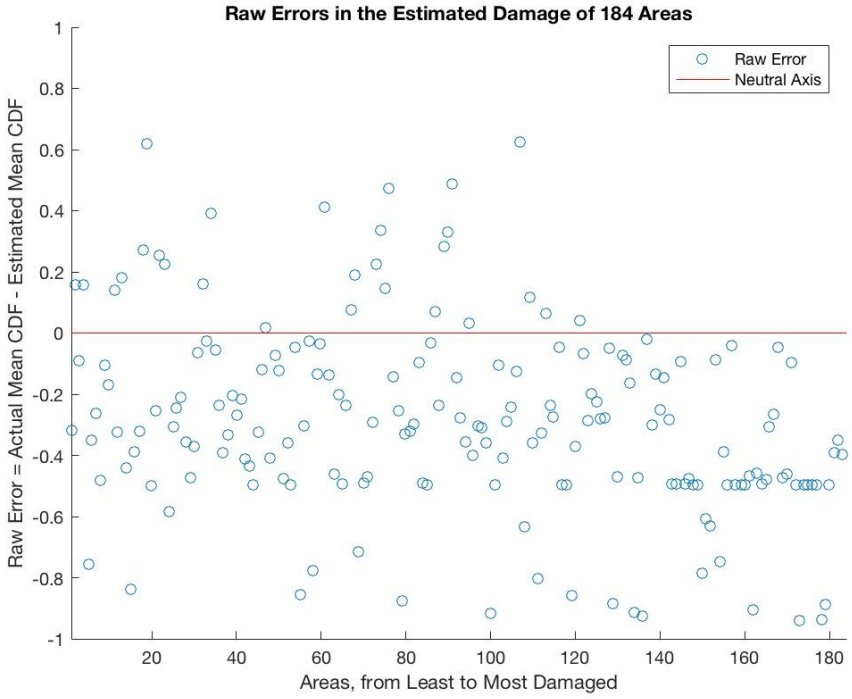


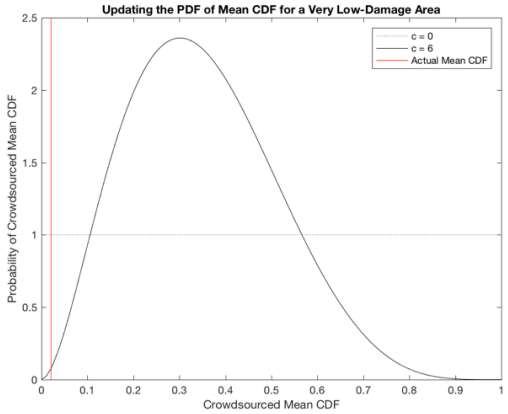
FIGURE 7.14: The raw error in the predicted damage level for each image, including responses of "same". The raw error was computed as the actual damage level in the image minus the predicted damage level, taken as the mean of the image's posterior beta distribution. Negative values indicate overestimation of the damage level in an image.

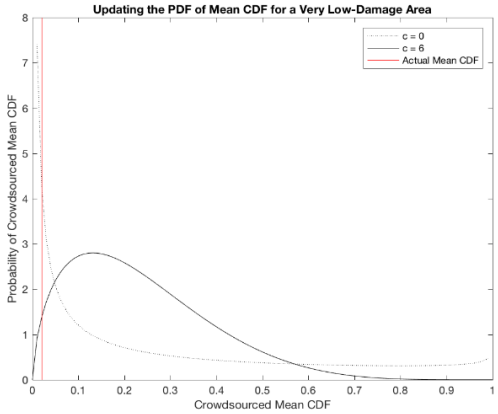
7.2.4 Results of a More Complex Bayesian Updating Scheme

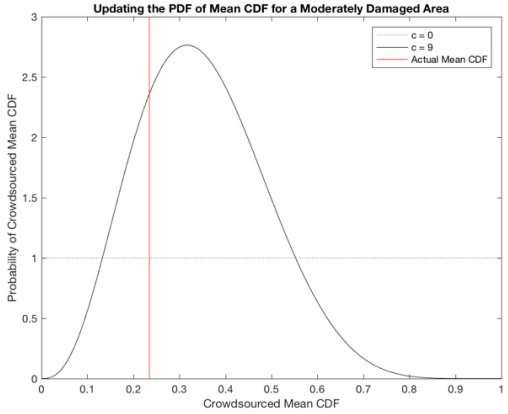
In this section, we first present results for three images – one with minimal damage, one with moderate damage, and one with a high level of damage – to illustrate nuances of the approach described in the previous section. We then discuss the results of various approaches when applied to the entire set of 184 images.

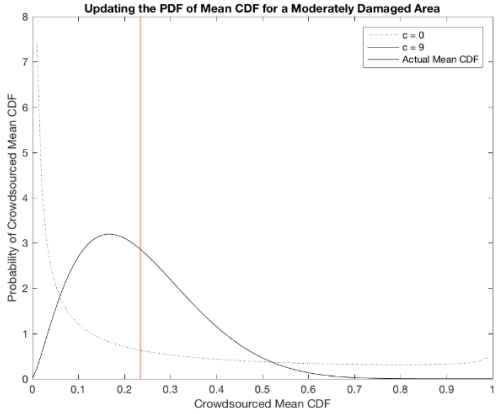
In Figure 7.17, we present illustrative results for three images using the more complex updating scheme, including "sames" as discussed in the previous section. The plots on the left show the results – that is, the posterior distribution after some number c of comparisons – for each of the

three images using an uninformative prior, i.e. beta(1,1). The plots on the right show the results for the same images, but using an informative prior, i.e. beta(0.2,0.8). A comparison of the two columns highlights the importance of using an informative prior, if possible, to improve accuracy and reduce uncertainty in the final result.









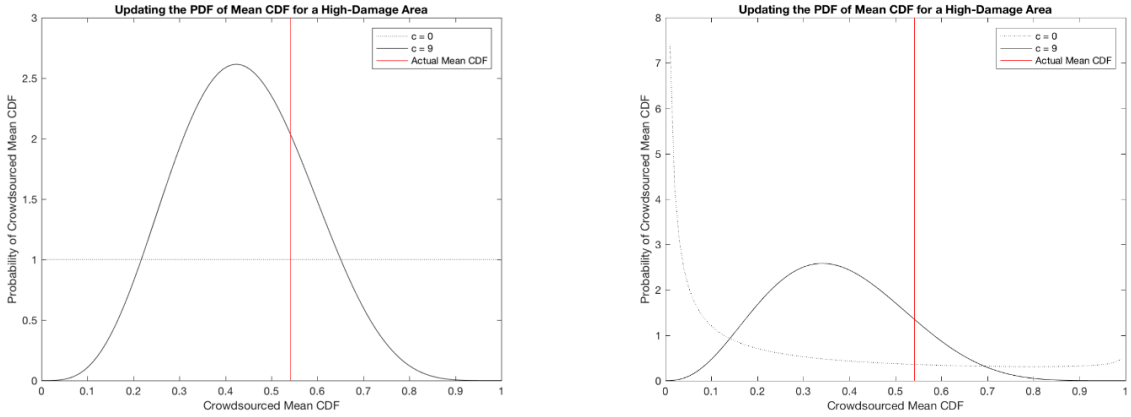


FIGURE 7.17: Results from complex updating scheme for low (top), moderate (middle), and high (bottom) damaged images. The results on the left use an uninformative prior and the results on the right use an informative prior.

Furthermore, the use of an informative prior improves results for the set of images considered as a whole. When using an informative prior, for example, an image's actual damage was within one standard deviation of its predicted damage for 23.4% of the images – a significant improvement over the 15.2% of the images whose actual damage was within a single standard deviation of their predicted damage when using an uninformative prior. Table 7.4 presents the accuracy of the Bayesian updating method, with and without an informative prior.

TABLE 7.4: The percentage of images whose actual damage level was within 1, 1.5 or 2 standard deviations of their predicted damage levels, taken as the mean of their posterior beta distributions. This table compares the overall accuracy of the two priors used in the more complicated Bayesian updating approach.

Actual damage relative to predicted damage	**Uninformative prior**	**Informative prior**
Within 1 standard deviation	15.20%	23.40%
Within 1.5 standard deviations	31.00%	40.20%
Within 2 standard deviations	45.70%	54.40%

Figure 7.18 highlights that the crowd proved better at estimating the damage in areas with higher overall damage, and that it generally failed to capture the level of damage in the approximately 50 areas with the lowest damage levels. While the overestimation of damage at the lower end of the damage spectrum was dramatic, the underestimation of the highest damage levels proved small in comparison.

Whether these limitations and levels of uncertainty are tolerable depends in large part on the tolerance of the particular decision-maker or agency using this information. We emphasize here that the imagery used in these series of experiments was from 2010, and that the quality of post-disaster imagery available now is significantly better. Further work may be required to accurately characterize the crowd's ability to distinguish damage in the higher-quality imagery that is now widely available.

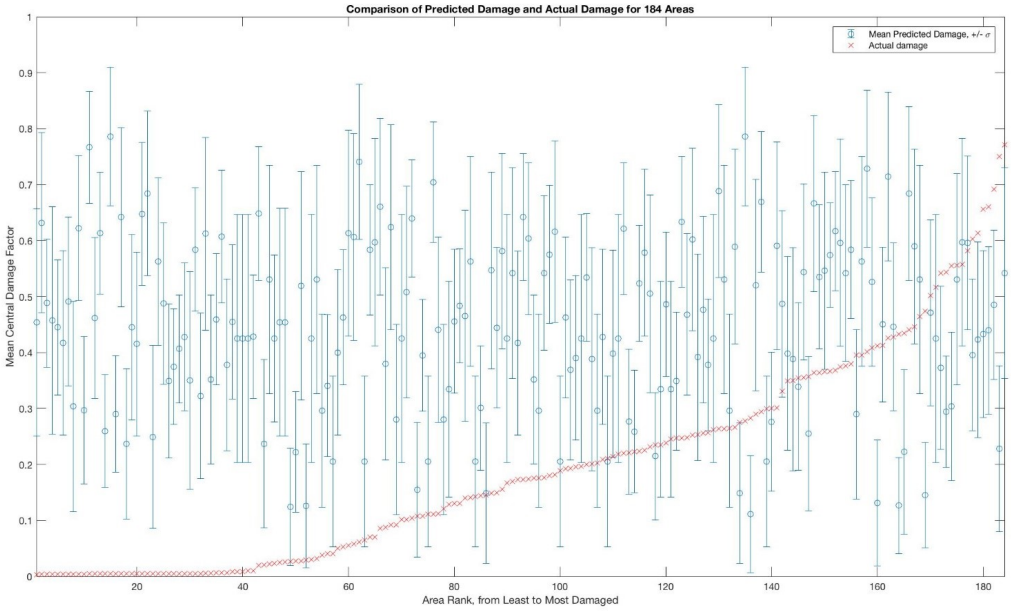


FIGURE 7.18: A box-and-whisker plot of the mean predicted damage, plus or minus one standard deviation, after implementation of the complex Bayesian updating scheme and including "sames". Comparing the predicted damage and the actual damage in all images shows that this approach does not result in reliable identification of low damage levels, and generally underestimates damage. The approach shows more promise when damage levels are moderate to high.

We now discuss the results of the same approach described above, except that in this case, we exclude all responses in which the respondent indicated that the image of interest and the anchor image to which it was being compared had the "same" level of damage. Because some images had only "same" responses, this portion of the analysis pertains to a reduced set of 174 images. our total set of images for this portion of the analysis is reduced to 174 images. The ten images not included in this portion of the analysis showed no apparent distinguishing characteristics to explain why respondents always considered them to have the "same" level of damage as the

anchor image to which they were being compared; the mean CDF of these ten images ranged from 0.0045 to 54.39 on a scale of 0 to 100.

By excluding "sames", we find that the accuracy of our Bayesian updating approach improves significantly in all six categories, as shown in Table 7.5. As when including "sames", an informative prior results in more accuracy when considering the totality of images. This suggests that there is no benefit to giving volunteers the option to indicate that an image has the same level of damage as an anchor image, especially given that including that option requires a not insignificant amount of additional work to be done at the front- and back-ends of the experimental platform.

TABLE 7.5: The percentage of images whose actual damage level was within 1, 1.5 or 2 standard deviations of their predicted damage levels, taken as the mean of their posterior beta distributions. This table compares the overall accuracy of the two priors used in the more complicated Bayesian updating approach.

Actual damage relative to predicted damage	**Uninformative prior**	**Informative prior**
Within 1 standard deviation	23.00%	32.80%
Within 1.5 standard deviations	36.20%	43.70%
Within 2 standard deviations	49.40%	56.30%

Excluding "sames" does not result in a noticeable change in the raw error, as is apparent when comparing the plots in Figure 7.19.

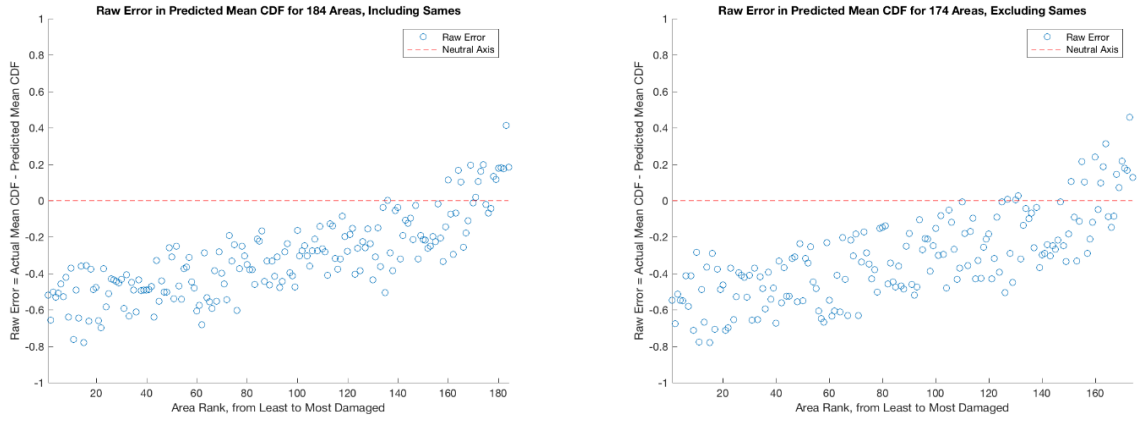


FIGURE 7.19: A comparison of the raw error in the predicted damage level for each image, including responses of "same" (left) and excluding responses of "same" (right), shows little difference. The raw error was computed as the actual damage level in the image minus the predicted damage level, taken as the mean of the image's posterior beta distribution

7.2.5 Conclusions

We demonstrate that the use of Bayesian updating to assimilate area-based damage assessments offers significant benefits to end-users of building damage information. Explicit quantification of the uncertainty in the damage estimates produced ranks first among the benefits. Other benefits include (1) computationally simple and efficient aggregation of assessments from different contributors (2) the option to straightforwardly weight contributors' assessments according to their performance on a training set (3) the option to incorporate expert knowledge of building damage distribution simply by choosing an informative prior distribution. In fact, choosing an informative prior distribution is arguably more appropriate, since lower levels of damage are generally known to be more common. Bayesian updating is thus a robust and flexible method well-suited to aggregating disparate damage assessments into a single damage estimate with associated variance.

We tested two Bayesian updating schemes. The first was a straightforward scheme in which the parameters alpha and beta were treated as pseudocounts. While this method had the advantage of extreme simplicity, we found that it yielded poor results when compared to the second, more complex updating method. To maximize the benefits of a Bayesian approach, end-users must be involved in setting up two parameters of the crowdsourcing platform: (1) the metric to be estimated (2) the maximum uncertainty tolerable in a final estimate. First, end-users should define the (continuous) damage summary statistic most relevant to their purposes; in the crowdsourcing experiments described in this report, we used the mean central damage factor over an area. Second, end-users must define an acceptable level of uncertainty in the final estimates of damage. This maximum level of uncertainty will determine the minimum number of assessments required for the damage in an image to be sufficiently certain for the end-user's purposes.

Using Bayesian updating to crowdsource damage estimates thus requires the involvement of the estimates' end-users, perhaps more explicitly than other methods discussed in this report. While this may appear restrictive in the context of rapid post-disaster information gathering, producing information that of immediate use to stakeholders in post-disaster processes serves the best interests of all parties involved. To facilitate effective coordination between information producers and users in post-disaster contexts, those stakeholders should coordinate in non-urgent contexts. For example, if crowdsourced building damage assessments were required, the basic parameters necessary to set up Bayesian updating (namely, the summary damage statistic of interest and the acceptable level of uncertainty) should be mutually agreed upon prior to a sudden-onset disaster.

We note that in the context of Bayesian updating, allowing the contributor to indicate that two

images have the same level of damage is inimical to gathering useful information, in large part from a user experience perspective rather than a mathematical one. From anecdotal feedback, we found that users defaulted to choosing "same" when really they were unsure how to judge the relative damage in a pair of images, for reasons including poor image quality, differing building densities, or inability to visually distinguish damage. This adapted use of the "same" option resulted in a large proportion of comparisons – for some areas, the majority of comparisons – in which the two images were rated by users as showing the same level of damage.

Inclusion of the "same" button on the user interface resulted from a series of discussions within team workshops, detailed in Section 1.3.2 of this report. Future experiments or implementations of this methodology would benefit from adhering to a binary response format – that is, forcing users to choose which image has greater damage. While it may sound reasonable to include an additional option by which contributors may indicate that they are unsure of the relative damage in the areas pictured, we caution that such an option would likely run counter to the information-gathering goals of the experiment. Anecdotal evidence indicates that our crowd included few contributors confident in their abilities to judge building damage – that is, most of our contributors seemed to be unsure most of the time. Whether an "unsure" response could be assimilated into an image's damage estimate remains unclear, and the costs of having such a response option currently outweigh any potential benefits.

Future Work

Future research in this direction should prioritize the development of statistically rigorous methods for incorporating damage information encoded in comparisons using anchor images. Other research directions that may prove fruitful include finding ways to weight the contributions of individual volunteers, e.g. based on their performance in an assessment or on a training set of images; investigating how this methodology may be useful when applied to other types of damage; and conducting similar experiments using the higher-quality imagery so widely available today.

8 Summary and Conclusions

This project aimed to address crowdsourcing for assessing post-earthquake building damage information in its entirety. This meant not only developing the platforms for three different crowdsourcing tasks, but also using novel approaches to analyze their results and making marked progress towards understanding stakeholder needs of this information.

Throughout this nearly two yearlong project, the team designed three different crowdsourcing experiments: one building-level and two novel area-based assessments. The building-level approach was designed to be analogous to previous crowdsourcing assessments, in which a volunteer would view a post-earthquake satellite image and rate the individual level of damage for each identifiable building. Because of the time intensive nature of this task, plus the requirement for volunteers to be trained in OpenStreetMap, only 85 images were assessed out of the total 843 images required (281 images $\times$ 3 passes). The two area-based approaches, on the other hand, were designed to simplify the crowdsourcing task by allowing volunteers to assess the damage for an entire region in an image as opposed to a single building. These two area-based assessments were fully completed within the four-month deployment period.

The crowdsourcing results could be classified into a damage indicator and damage comparison dataset. The damage indicator dataset from experiments 2 and 3 showed initially low correlation with the true damage in the area of interest but was improved by weighting the responses of "high performing" users. For the experiment 3 comparison dataset, network analysis and Bayesian updating were shown to be methods that could be used to obtain damage levels for a set of paired comparison data. Through these analysis techniques, we've put forth multiple methods that can be used when working with crowdsourced building damage data. From the stakeholder needs perspective, shifting the crowdsourcing task from a building-level to area-based approach was motivated by the team's understanding of how crowdsourced building damage information could be used. By convening multiple disciplines together, including both data providers, map developers, and disaster risk management specialists, we cold target our design and analysis to inform regional loss estimates, such as those reported in the Post Disaster Needs Assessment (PDNA).

However, the team wanted to understand uses of post-earthquake building damage information outside of the Post Disaster Needs Assessment. Hence, an extensive demand survey was carried out through multiple interviews with response and recovery practitioners to map out six key post-disaster activities that require building damage data.

The conclusions from each portion of this project will be summarized throughout this section, from initial experimental design to the results of the demand survey. Overall, we assert that the tested area-based methods are viable approaches to crowdsourcing building damage information. This method proved to be simpler and quicker to complete than previously implemented building-level approaches. In addition, the two area-based approaches can address multiple stakeholder needs, as identified by the demand survey.

However, much can be done to extend this project in the future, both in the experimental design and analysis of the results. Thus, the report will close with our final thoughts on the entire project process and ideas for future extensions.

8.1 On the Experimental Design

The experimental portion of the study consisted of two phases: a pre-experiment phase for experiments 2 and 3, and the final experiments. This section outlines the conclusions and recommendations from the overall experiment design process.

The pre-experiment phase was extremely valuable. Running the pre-experiments provided valuable feedback that was incorporated into the final experiments. Sitting alongside users as they completed the questions gave the research team insight into their thought process and highlighted any difficulties in the interface and how information was presented. Completing as much pre-testing before launching the final version of the experiment is highly recommended.

Perhaps the only downside to completing the pre-experiments was the decreased participation rate in the final experiment. Users who were eager to participate in the pre-experiments showed less enthusiasm when the final experiments were launched. Care should be taken to not prematurely exhaust the user base before the final experiments are online. Offering incentives for participation may help to address this issue.

Improve methods to collect user data. The intent behind collecting user data was to track a user's responses across each question, and to gain an understanding of their background and experience. To address the former, IP addresses were recorded with each response. As initial results were received it became apparent that there were instances where users in the same geographic location shared the same IP address. This meant that it was not possible to analyze the entire dataset by user. It is recommended that an alternative method, such as requiring a login, be implemented in the future to enable a more complete analysis of responses by user.

Similarly, the dataset received from the questionnaires about user background and experience was incomplete. The optional questionnaires appeared after the user had completed 15 questions in experiments 2 and 3, meaning many of the users did not even see the questionnaire. Placing the questionnaire upfront as part of the tutorial (and making it compulsory) is one way this could be addressed. The tradeoff between more user data and a potential decrease in user participation would need to be considered.

Refine comparison method for experiment 3. The intent behind experiment 3 was to dynamically update the images presented to users for comparison based on previous user responses, as this will reduce the overall number of required comparisons. This was implemented in part as the subsequent anchor image shown to a user was determined by their previous answer as outlined in section 5.2. Due to constraints of the Pybossa platform, dynamic updating was limited to individual users, such that user A's responses were not able to influence the questions shown to user B, but user A's first response did influence their subsequent questions. Testing this approach using a platform that enables full dynamic updating would provide a better understanding of the number of comparisons required to obtain a sufficient level of reliability in the results.

A number of improvements could also be made in the way images are assigned to a damage bin. The damage indicator dataset contained a large number of images placed in bin 6, indicating a bias that does not reflect actual damage. As outlined in section 5.3, images were placed in bin 6 if they clicked on the "same" button during the first comparison. In future, it is recommended that the same option be removed to avoid this bias from premature binning. Instead, additional comparisons could be implemented as a check. For example, if a user responds that the image shows more damage than anchor 6, then they will be asked to compare the image against anchor 8. If they then respond that the image shows less damage it will be placed in bin 7. An additional check could present another anchor 6 image to test if the user again selects the image shows more damage than anchor 6. If so, it could be binned. If not, then the comparisons could continue in the other direction until there is confirmation from both directions that the image is in the correct

bin. The increased number of comparisons this would require could be offset by implementing the dynamic tasking approach. In addition, an extra question could be added to each assessment, asking the user how confident they are in their response. A low level of confidence could be used as an indicator that the images show similar damage.

Improve training section for users and collect a training dataset. The final experiments included a click through tutorial with more descriptive examples of damage compared to the tutorial provided during the pre-experiments. This tutorial could be further improved in a number of ways. More informative descriptions and examples of damage could be provided, including how image characteristics could influence the assessment (e.g. shadows). Tutorial development would benefit from suggestions from outside disciplines, such as cognitive systems engineering, as described by Kerle and Hoffman (Kerle and Hoffman 2013).

As part of the tutorial, it would be helpful to both users and researchers to provide a training set of questions. Users would be asked a set of questions where they would be provided with feedback on their response (whether they were correct, where the visible damage is located, features that may look like damage but in fact are not etc.). This would also provide researchers with a training dataset for each user so that user reliability could be calculated and factored into a damage prediction (such as that in Section 6).

Design and usability of platform interface affects user participation. During the first mapathon, participants from Facebook attempted to edit and improve the user interface for experiments 2 and 3 in real time. The experiments were implemented on the Pybossa platform as it was an existing platform for running crowdsourcing experiments, dramatically reducing the time and effort required to set up the experiments. This provided some constraints to the design of the user interface, however, the study would have benefited from the experience of someone with a user experience or user interface background.

In addition, users were required to complete the experiments at a computer. If users were able to participate via an app on a mobile device, it is likely that participation rates would have been higher. As mentioned above, offering incentives may also help to increase user participation

Engage experts to guide outreach and recruitment efforts. Stanford's recruitment efforts heavily focused on those with a structural engineering or earthquake background. In future it would be preferable to reach a wider cross section of the population as the intended volunteer network in

a live setting would encompass people from all backgrounds. Engaging someone with experience in survey design as well as marketing or outreach may be one way to address this. The study may have also benefited from someone with a marketing or publicity background to handle the outreach communication.

8.2 From the Crowdsourcing Results

The analysis of the damage indicator and damage comparison results dataset aimed to:

1. Explore any trends relating to the crowd's ability to ascertain building damage through the developed area-based approaches
2. Demonstrate multiple methods that can be used to analyze both ordinal and paired comparison crowdsourcing datasets

The simple weighted regression models used for the indicator dataset could be used for any crowdsourcing method that results in a numerical scale of damage, associated with a building or image. Thus, this regression analysis draws conclusions on both area-based experiments, which could be easily extrapolated to the building-level experiment if more results are obtained in the future.

The network analysis and Bayesian updating methods proposed for the damage comparison dataset draw conclusions specifically for the paired comparison data resulting from experiment 3. These methods offer additional information beyond validation, such as the ability to iteratively update through the data collection process or to include prior knowledge on building damage.

8.2.1 The Damage Indicator Dataset

The damage indicator dataset is the most comparable between the three experiments, since each experiment produced results that could be interpreted as a numerical value of damage severity per image. Consequently, similar exploratory and predictive data analysis was carried out, leading to overarching conclusions for the two area-based approaches.

Volunteers are detecting a combination of building damage and density per image Upon initial exploration of the raw results from the experiments 2 and 3, there was a slight positive correlation with the crowdsourced damage indicators and ground-validation data. The damage indicators showed a positive correlation with the ground-validation metric of building damage (mean central damage factor), but also the number of buildings per image. This implies that volunteers were choosing damage indicator values based on a combination of building damage and building density in an image. However, it was impossible to retroactively understand the thought-process behind users' decisions. In future experiments, it is advised to conduct more in-person experiments where volunteers speak aloud their decision-making process throughout the survey.

Weight "good" users to improve overall performance and combine multiple passes The raw results from both experiments 2 and 3 showed a significantly wide distribution of the ground-validation damage severity for each user-provided damage indicator. Through further exploration, we found that numerous users incorrectly identified damage (21/51 users for experiment 2 and 15/43 users for experiment 3). However, many users also performed exceedingly well at this area-based assessment by choosing high damage indicator values for high damaged. The overall performance of the crowd was improved by weighting users that performed well on this assessment, indicated by an increase in the slope of the initial regression model through the raw results. Also, by weighting the performance of users, the response from the highest performing user could be used to produce a final map of crowdsourced damage.

Area-based approaches should be pursued further and benchmarked against a building-level approach If more data was obtained from experiment 1, the results of the regression analyses for both area-based approaches and the building-level approach could be compared directly. It is recommended to benchmark the two area-based experiments against the building-level experiment, to complete a final accuracy comparison between the two crowdsourcing methodologies. Based on our team's understanding of earlier building-level implementations, we expect that the results of the building-level approach would have similar accuracy to that of the two area-based experiments. Furthermore, the two area-based assessments were completed on a much quicker timescale.

Experiments 2 and 3 results could also be directly compared. The shape of the distribution of the number of responses from experiment 2 more closely matches that of the true distribution of damage in the area of interest. Furthermore, the initial correlation between crowdsourced and

ground-validation damage was slightly higher for experiment 2 and 3. After weighting good performing users, though, experiment 3 performed slightly better (with a lower error metric) than experiment 2 . Again, these conclusions are impacted by the inordinate amount of "same" responses from experiment 3. Thus, it is highly recommended to remove the option for choosing "same damage" between two images in future iterations of a image-comparison experiment for crowdsourcing.

8.2.2 The Damage Comparison Dataset

Two different approaches were used to analyze the damage comparison dataset: Bayesian updating and network analysis.

Network analysis can be used to sort paired comparisons into an ordered list of images We demonstrate that networks are an effective way to structure and analyze the damage comparison data as it captures the relationships between pairs of images defined by user assessments. Standard network analysis algorithms can be applied to analyze the dataset, as shown through the use of the topological sorting algorithm. The results show that the accuracy of the prediction using on a network is dependent upon the network density, which is directly related to the number of user assessments. As the network density increases, the accuracy of predictions increases however, there is likely an optimal range for the number of comparisons required, beyond which the increase in accuracy is no longer warranted. Running simulations of user responses would provide insight to these optimal conditions. While there are a number of obstacles that need to be addressed for this approach to be useful in a real-world setting, this approach shows promise. One particular advantage is that this model is easily updated as new information is received. Being able to add to the network and quickly rerun the analysis could be useful during the data collection phase of a live project, allowing for images to be prioritized for assessment based on measures such as the density of the network or level of user agreement.

Bayesian updating provides measures of uncertainty and incorporates prior expert knowledge Similarly, we demonstrate that the use of Bayesian updating to assimilate area-based damage assessments offers significant benefits to end-users of building damage information. Explicit quantification of the uncertainty in the damage estimates produced ranks first among the benefits. Other benefits include (1) computationally simple and efficient aggregation of assessments from

different contributors (2) the option to straightforwardly weight contributors' assessments according to their performance on a training set (3) the option to incorporate expert knowledge of building damage distribution simply by choosing an informative prior distribution. In fact, choosing an informative prior distribution is arguably more appropriate, since lower levels of damage are generally known to be more common. Bayesian updating is thus a robust and flexible method well-suited to aggregating disparate damage assessments into a single damage estimate with associated variance.

Confirming the negative effect of the "same" option Similar to the damage indicator results, one finding from both the Bayesian updating and network analysis studies is that the inclusion of the "same" response detracted rather than added to the results. Future experiments or implementations of this methodology would benefit from adhering to a binary response format – that is, forcing users to choose which image has greater damage. While it may sound reasonable to include an additional option by which contributors may indicate that they are unsure of the relative damage in the areas pictured, we caution that such an option would likely run counter to the information-gathering goals of the experiment. Anecdotal evidence indicates that our crowd included few contributors confident in their abilities to judge building damage – that is, most of our contributors seemed to be unsure most of the time. Whether an "unsure" response could be assimilated into an image's damage estimate remains unclear, and the costs of having such a response option currently outweigh any potential benefits.

8.3 Future Work

The results and limitations described previously suggest numerous avenues for promising future work. The experiments confirmed the initial hypothesis that area-based assessments are significantly easier and faster than building-level assessments. As such, future work can focus on further developing such methods and improving their performance. Proposed future work includes:

Area-based crowd-sourced damage assessment with higher-resolution imagery The experiments were conducted using 50 cm resolution satellite imagery. It was found that identifying damage was very difficult at that resolution. Using higher-resolution imagery (<10cm) would enable volunteers to much more easily assess damage. Conducting sensitivity-analysis to imagery resolution would provide valuable information for future crowd-sourced damage assessment efforts relying on optical imagery. Extending the Bayesian inference method: Bayesian updating is a

very promising method for combining multiple volunteer assessments as well as prior estimates. Further work in this area will be explored.

Extending the network analysis methods Network analysis is a novel method for ranking images based on volunteer damage assessments. This could be explored further with a larger data-set, cleaner data (no "same" answers) and more multi-pass assessments.

Benchmarking to current state of practice While originally planned, it was not possible to replicate current state-of-practice crowd-sourced based analysis, as we were unable to gather enough volunteers to conduct building-level assessments. While this demonstrated the motivation for our research (i.e. exploring methods that would be simpler and quicker), we were unable to compare results of our proposed new methods to that of current state of practice. Future work could make use of paid services such as mechanical turk in order to develop such a benchmark, or use a data-set already collected.

Repeated training While all volunteers had to run through training materials in order to advance to the experiment, it is unclear how effective the training was, or whether it was forgotten over time. Regular training is expected to improve overall performance while also preventing gradual drift / bias as volunteers conduct numerous assessments.

Improvements to experiment set-up As described previously, future work will involve re-designing several of the experiment set-up. This includes: (i) developing a single training-set for all volunteers to better weight user performance, (ii) removing the option to select "same" in the damage comparison experiment, (iii) using a single training-set for all volunteers would make it easier to weight their performance more accurately, (iv) requiring a login to identify single users (addressing the non-unique IP issue) etc.

8.4 From the Demand Survey

Through a series of in-depth interviews and an online survey, we identified six post-disaster decisions. Using a novel framework, we contextualized these decisions by their timing and the minimum level of spatial precision required of the information upon which they are based. We

presented our contextualization of these decisions graphically in Figure 3.4 and summarized the key features of these decisions in Table 3.1.

Our framework highlights that what we refer to as decisions are more often collaborative processes, or groups of highly related decisions, that unfold over particular segments of the post-disaster timeline. Figure 3.4 not only highlights the timing of these decisions but also suggests how information gathered or used by one organization could benefit decision-makers who are concurrently or subsequently active. By mapping decisions according to the critical features of the information upon which they rely, our framework also highlights concrete opportunities for information-sharing, which aligns with cluster approach to inter-agency coordination already adopted by the United Nations.

The application of our proposed framework to the post-disaster decisions reframes the production of post-disaster information by focusing on the needs of specific users, rather than on ever-advancing technical and technological capabilities. We hope that our work also contributes toward a holistic understanding, shared by data users and providers, of what information is needed when, and at what precision. This framing of post-disaster decisions may, for example, help to inform academicians, researchers, and other data producers of unmet information needs, and identify information needs common to different decision-makers, thus indicating areas in which improvements would have high impact. One of the principal information needs that is often unmet is a comprehensive and detailed pre-disaster building census, from which many of the decisions included in this study could benefit.

8.5 Final Thoughts

Accurate information on the extent and spatial distribution of damage quickly after a large disaster is critical input to numerous decision-making processes for disaster relief, early recovery and reconstruction. This worked seeked to explore methods to produce such assessment by leveraging the crowd and their willingness to contribute to humanitarian relief remotely through micro-tasked damage assessment. While not new, current state-of-practice involves volunteer tagging individual buildings from satellite or aerial imagery, and assigning damage levels to each building. This process has had limited success in practice due to issues of accuracy and the tedious nature of the task.

In this work we explored two new methods of volunteer tasking, and several new methods for data post-processing to get an accurate damage assessment with less effort. These have shown a lot of promise, though more testing is necessary.

It was found that simplifying the task from damage assessment at the building level to assessment of damage in entire images (showing dozens of buildings) is very promising. Volunteers were generally able to identify damage properly, though more work is needed to improve overall accuracy, and is described in the future work section.

Specific contributions of this work include:

1. Extended stakeholder survey, providing better understanding of the needs of various user for damage data along with its accuracy, resolution and timeframe.
2. A novel volunteer tasking approach focusing on assessing damage to an entire image (with dozens of buildings), rather than building level assessment.
3. A novel volunteer tasking approach focused on pairwise ranking of images based on visible damage.
4. Proposed method for data post-processing including and combining ordinal-scaling, user-performance weighting and multi-pass integration.
5. Proposed method for damage ranking using network analysis methods.
6. Proposed method for combining multiple assessments through Bayesian updating process.

Overall, this project has opened up new directions for research into post-disaster impact assessment. Such research is important both for supporting humanitarian and disaster recovery work, but also for developing novel approaches to crowd-tasking and data analysis beyond the field of disaster relief.

Bibliography

Agresti, Alan (July 2002). *Categorical Data Analysis*. Wiley-Interscience.

Ajmar, Andrea et al. (2011). "Earthquake damage assessment based on remote sensing data. The Haiti case study". In: *Italian Journal of Remote Sensing* 43.January 2010, pp. 123–128. ISSN: 11298596. DOI: 10.5721/ItJRS20114329. URL: http://www.aitjournal.com/articleView.aspx?ID=227.

Al Achkar, Ziad, Isaac L. Baker, and Nathaniel A. Raymond (2016). *Assessing Wind Disaster Damage to Structures*. Tech. rep. viii. Harvard Humanitarian Initiative.

Albuquerque, João, Benjamin Herfort, and Melanie Eckle (Oct. 2016). "The Tasks of the Crowd: A Typology of Tasks in Geographic Information Crowdsourcing and a Case Study in Humanitarian Mapping". In: *Remote Sensing* 8.10, p. 859.

Applied Technology Council (1989). *Procedures of Postearthquake Safety Evaluation of Buildings*. Tech. rep. Redwood City, CA: Applied Technology Council.

ATC-13 (1985). "Earthquake Damage Evaluation Data for California". In: *Applied Technology Council*. URL: https://www.atcouncil.org/pdfs/atc13.pdf.

Bharosa, Nitesh, Jinkyu Lee, and Marijn Janssen (2010). "Challenges and obstacles in sharing and coordinating information during multi-agency disaster response: Propositions from field exercises". In: *Information Systems Frontiers* 12.1, pp. 49–65. ISSN: 13873326. DOI: 10.1007/s10796-009-9174-z.

Booth, Edmund et al. (Oct. 2011). "Validating Assessments of Seismic Damage Made from Remote Sensing". In: *Earthquake Spectra* 27.S1, S157–S177.

Chen, Rui et al. (2008). "Coordination in emergency response management". In: *Communications of the ACM* 51.5, pp. 66–73. ISSN: 00010782. DOI: 10.1145/1342327.1342340. URL: http://portal.acm.org/citation.cfm?doid=1342327.1342340.

Corbane, Christina, Daniela Carrion, et al. (2011). "Comparison of damage assessment maps derived from very high spatial resolution satellite and aerial imagery produced for the Haiti 2010 earthquake". In: *Earthquake Spectra* 27.SUPPL. 1, pp. 199–218. ISSN: 87552930. DOI: 10.1193/1.3630223.

Corbane, Christina and Guido Lemoine (2011). *Collaborative Spatial Assessment - CoSA*. Tech. rep. Luxembourg: European Union, pp. 1–186. DOI: 10.2788/87600.

Corbane, Christina, Keiko Saito, et al. (Oct. 2011). "A Comprehensive Analysis of Building Damage in the 12 January 2010 M(W)7 Haiti Earthquake Using High-Resolution Satellite- and Aerial Imagery". In: *Photogrammetric Engineering Remote Sensing* 77.10, pp. 997–1009. ISSN: 0099-1112.

David, Matthew. and Carole D. Sutton (2011). *Social research : an introduction*. 2nd ed. London: SAGE Publications, p. 665. ISBN: 1847870139. URL: https://books.google.com/books?id=tJUx2nEscG4C%7B%5C&%7Ddq=weighting+reliable+data%7B%5C&%7Dsource=gbs%7B%5C_%7Dnavlinks%7B%5C_%7Ds.

Elia, Agata, Piero Boccardo, and Simone Balbo (2016). "A Quality Comparison Between Expert and Crowdsourced Data in Emergency Mapping for a Potential Service Integration". PhD thesis. Politecnico Di Torino.

Foulser-Piggott, Roxane et al. (Feb. 2016). "Using Remote Sensing for Building Damage Assessment: GEOCAN Study and Validation for 2011 Christchurch Earthquake". In: *Earthquake Spectra* 32.1, pp. 611–631.

Ghosh, Shubharoop et al. (May 2011). "Crowdsourcing for Rapid Damage Assessment: The Global Earth Observation Catastrophe Assessment Network (GEO-CAN)". In: *Earthquake Spectra* 27.S1, S179–S198.

Gralla, Erica (George Washington University), Jarrod (Massachusetts Institute of Technology) Goentzel, and Bartel(Tilburg University) Van de Walle (2013). *Field-Based Decision Makers' Information Needs in Sudden Onset Disasters*. Tech. rep. October, pp. 1–51.

Grünthal, Gottfried (1998). *European Macroseismic Scale 1998*. Vol. 15, p. 100. ISBN: 2879770084. URL: http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:European+Macroseismic+Scale+1998%7B%5C#%7D0.

Hausser, George (n.d.). *Situational Use of Data Weighting*. Tech. rep. TRC Market Research, pp. 1–3.

Huynh, Andrew et al. (2014). "Limitations of crowdsourcing using the EMS-98 scale in remote disaster sensing". In: *IEEE Aerospace Conference Proceedings*, pp. 1–7. ISSN: 1095323X. DOI: 10.1109/AERO.2014.6836457.

James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Ed. by G Casella, S Fienberg, and I Olkin. New York, NY: Springer, pp. 1–440. ISBN: 9781461471370.

Joyce, Karen (2016). "Remote Sensing and the Disaster Management Cycle". In: *Agricultural and Biological Sciences Grain Legumes*. ISSN: 9789533070865. DOI: 10.5772/711. arXiv: 0803973233.

Kerle, N and R R Hoffman (2013). "Collaborative damage mapping for emergency response : the role of Cognitive Systems Engineering". In: *Natural hazards and earth system sciences* 13.1, pp. 97–113.

Kerle, Norman (2013). "Remote Sensing Based Post-Disaster Damage Mapping with Collaborative Methods". In: *Intelligent Systems for Crisis Management*, pp. 121–133. DOI: 10.1007/978-3-642-33218-0.

Kochenderfer, Mykel J. et al. (2015). *Decision Making Under Uncertainty: Theory and Application*. 1st. The MIT Press. ISBN: 0262029251, 9780262029254.

Lallemant, David and Anne Kiremidjian (Aug. 2015). "A Beta Distribution Model for Characterizing Earthquake Damage State Distribution". In: *Earthquake Spectra* 31.3, pp. 1337–1362.

Lallemant, David, Robert Soden, et al. (2017). "Post-Disaster Damage Assessments as Catalysts for Recovery: A Look at Assessments Conducted in the Wake of the 2015 Gorkha, Nepal, Earthquake". In: *Earthquake Spectra* 33.S1, S435–S451. ISSN: 8755-2930. DOI: 10.1193/120316EQS222M. URL: http://earthquakespectra.org/doi/10.1193/120316EQS222M.

Lemoine, G. et al. (2013). "Intercomparison and validation of building damage assessments based on post-Haiti 2010 earthquake imagery using multi-source reference data". In: *Natural Hazards and Earth System Sciences Discussions* 1.2, pp. 1445–1486. ISSN: 2195-9269. DOI: 10.5194/nhessd-1-1445-2013. URL: http://www.nat-hazards-earth-syst-sci-discuss.net/1/1445/2013/.

Meissner, A et al. (2002). "Design Challenges for an Integrated Disaster Management Communication and Information System". In: *First IEEE Workshop on Disaster Recovery Networks* Diren. ISSN: 03064573. DOI: 10.1016/j.ipm.2009.07.002.

MTPTC (2010). *Evaluation des Batiments*. URL: http://www.mtptc.gouv.ht/accueil/actualites/article%7B%5C_%7D7.html.

Newman, Mark (2010). *Networks: An Introduction*. Oxford: Oxford University Press, pp. 1–720.

NIST and SEMATECH (2018). *Weighted Least Squares*. URL: http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd432.htm (visited on 01/15/2018).

Poblet, Marta, Esteban García-cuesta, and Pompeu Casanovas (2014). "Crowdsourcing Tools for Disaster Management : A Review of Platforms and Methods". In: Ccc, pp. 261–274.

Saito, Keiko, Robin Spence, and Terence A De C Foley (2005). "Visual Damage Assessment Using High- Resolution Satellite Images Following the 2003 Bam , Iran , Earthquake". In: *Earthquake Spectra* 21.December, pp. 309–318. DOI: 10.1193/1.2101107.

UN-SPIDER (2018). *International Charter Space and Major Disasters | UN-SPIDER Knowledge Portal*. URL: http://www.un-spider.org/space-application/emergency-mechanisms/international-charter-space-and-major-disasters (visited on 01/03/2018).

The United Nations Office for the Coordination of Humanitarian Affairs (2013). *United Nations Disaster Assessment and Coordination (UNDAC)*. Tech. rep., p. 269. URL: https://docs.unocha.org/sites/dms/Documents/UNDAC%20Handbook%202013%7B%5C_%7Denglish%7B%5C_%7Dfinal.pdf.

Voigt, Stefan et al. (2011). "Rapid damage assessment and situation mapping: learning from the 2010 haiti earthquake". In: *Photogrammetric Engineering and Remote Sensing* 77.9, pp. 923–931.

Westrope, Clay, Robert Banick, and Mitch Levine (Jan. 2014). "Groundtruthing OpenStreetMap Building Damage Assessment". In: *Procedia Engineering* 78, pp. 29–39.

Womble, J Arn, Richard L Wood, and Ronald T Eguchi (2016). "Current Methods And Future Advances For Rapid, Remote-Sensing-Based Wind Damage Assessment". In: *Resilient Infrastructure London 2016*. Womble 2005, pp. 1–11.