A Comparative Case Study Analysis of Data Production and Maintenance in Humanitarian Mapping Campaigns

Hannah Ker

This dissertation is submitted in partial fulfillment
of the requirements for the degree of MSc
in the Centre for Advanced Spatial Analysis,
Bartlett Faculty of the Built Environment, UCL

CASA0012

11,763 words

August 23, 2020

Supervised by Dr. Sarah Wise

Abstract

Humanitarian mapping campaigns have the capacity to produce large volumes of openly available geospatial data in OpenStreetMap (OSM) following a crisis. For regions lacking in authoritative geospatial data, OSM can be a valuable resource for first responders. However, OSM's crowd-produced nature and lack of formal mechanisms for quality control often raise questions around its trustworthiness and credibility. This work focuses particularly on the issue of temporal accuracy (or up-to-dateness) of OSM data. This work empirically evaluates the extent to which data produced during humanitarian mapping campaigns has been maintained over time. A comparative case study approach is employed, whereby four humanitarian cases; Port au Prince, Bangui, Tacloban, and Kathmandu; are compared between each other and against a reference case study of known high data quality; Heidelberg, Germany. The newly developed OpenStreetMap History Database (OSHDB) framework, developed by Raifer et al. (2019), is applied to filter and process large volumes of historical OSM data. When compared against Heidelberg, it is found that the data produced during the humanitarian campaigns is generally poorly maintained over time. Four years after the conclusion of each campaign, the majority of all data has not been updated or deleted, leaving it at risk of being out of date. These findings suggest that formal mechanisms or incentives for data maintenance should be integrated into humanitarian mapping processes.

Declaration

I hereby declare that this dissertation is all my own original work and that all sources have been acknowledged. It is 11,763 words in length.

Contents

List of Figures

List of Tables

Acknowledgements

I would firstly like to acknowledge my supervisor, Dr. Sarah Wise, for her invaluable expertise and guidance throughout my entire research process. Her enthusiasm for this topic has been incredibly inspirational and motivating.

This work was also greatly supported by various members of the humanitarian mapping community. In particular, Jorieke Vyncke, of Médecins Sans Frontières, provided ongoing feedback and subject matter expertise. Raphael Brechard, of Médecins Sans Frontières, and Andrew Braye, of the British Red Cross, also provided valuable feedback as I developed my research ideas. I am also grateful for the work that has been conducted by the Missing Maps community, Humanitarian OpenStreetMap Team, and wider OpenStreetMap community, to provide the data and infrastructure that forms the basis for this research.

My efforts to implement the OSHDB API were significantly aided by assistance from Rafael Trolio, of the Heidelberg Institute for Geoinformation Technology.

I am also grateful for the support from CASA faculty and fellow students, and from Professor Anahid Basiri and Dr. Andrew MacLachlan in particular, for guiding me in selecting an appropriate topic for this dissertation.

List of Acronyms and Abbreviations

API	Application Programming Interface
CAR	Central African Republic
ETL	Extract, Transform, Load
HOT	Humanitarian OpenStreetMap Team
JDBC	Java Database Connectivity
MSF	Médecins Sans Frontières
OSHDB	OpenStreetMap History Database
OSM	OpenStreetMap
POI	Point of Interest
VGI	Volunteered Geographic Information

Chapter 1

Introduction

Accurate and up-to-date geospatial data is an important resource that enables an effective response to a humanitarian crisis (Cowan 2011; Poser and Dransch 2010; Soden and Palen 2016; Zook et al. 2010). Such data can be critical for functions such as distributing aid, identifying affected regions, and coordinating response between humanitarian organizations (Soden and Palen 2016). However, as many regions with humanitarian need do not have official geospatial data, alternative sources of crowdsourced data are often needed (Zook et al. 2010).

OpenStreetMap (OSM) is a valuable source of open and freely available geospatial data that is often used in humanitarian operations (Palen et al. 2015; Soden and Palen 2016). However, as OSM is an example of what Goodchild (2007) terms "volunteered geographic information" (VGI), it does not have any formal mechanisms for quality control and so may be considered less trustworthy by data users.

This work furthers the existing body of literature on data quality in OSM. Situated within the humanitarian context, this work considers the temporality that is inherent to geospatial data production in OSM. Just as real-world geographic features and their associated attributes change over time, so must their digital representations within OSM. The concept of temporal data quality is explored by investigating practices of data maintenance, which is considered to be the necessary process by which data is kept up-to-date.

Data maintenance is particularly relevant in humanitarian mapping contexts

due to the unique modes of data production in this domain. The response effort following the 2015 Kathmandu earthquake demonstrates how event-based humanitarian mapping campaigns have been able to quickly produce large volumes of up-to-date geospatial data (Soden and Palen 2016). However, much of this data is produced by remote volunteers (Eckle and Albuquerque 2015) who may not be invested in the quality of the data over time. This primarily remote nature of contribution coupled with the large volume of data may mean that the OSM data produced during humanitarian mapping campaigns is at a risk of quickly becoming out of date. However, this has yet to be empirically evaluated. While one cannot deny the value of this data in the wake of a crisis, the humanitarian mapping community also acknowledges the importance of longer-term sustainability of this data and its value as a community resource after a crisis subsides (Soden and Palen 2014).

In this research, I aim to explore the characteristics of geospatial data production during selected humanitarian mapping campaigns in OSM, and the extent to which this data is maintained following each campaign. It is hoped that the results of this analysis will contribute to a greater understanding of the quality of data that is produced in humanitarian mapping efforts, particularly relating to the data's ongoing temporal accuracy. I employ a comparative case study approach whereby I investigate four selected humanitarian mapping campaigns and one reference period of mapping activity in a region of known high data quality. Humanitarian case studies are selected from Port au Prince, Bangui, Tacloban, and Kathmandu; and Heidelberg, Germany is selected as the reference case study. It is intended for the results of this analysis to provide a foundation for future work in this emerging research domain.

This work explores the following specific research questions:

- **RQ 1:** What are the characteristics of data production in the selected humanitarian mapping campaigns and how does this compare with the reference case study?
- **RQ 2:** To what extent is the data produced during the selected campaigns maintained over time and how does this compare with the reference case

study?

- **RQ 3:** What insight do these results offer into potential relationships between characteristics of data production and levels of data maintenance in each of the case studies?

These research questions are addressed throughout this document as follows: In **Chapter 2**, I review existing academic literature relating to OSM data quality and humanitarian applications. I critically examine this past literature to identify a key research gap that this work addresses. In **Chapter 3**, I provide a brief description of the OSM data model and outline relevant computational challenges in processing historical OSM data. I highlight the OSHDB framework (Raifer et al. 2019) as the state-of-the-art in managing this data. I outline my methodology in **Chapter 4**, describing my case study approach, techniques in data collection and processing, and procedures for investigating data production and maintenance. I present the results of this analysis in **Chapter 5** and discuss their significance and limitations in **Chapter 6**. I conclude this work in **Chapter 7** and provide recommendations for future research efforts.

Chapter 2

Literature Review

In this literature review, I introduce the OpenStreetMap (OSM) project, situating it within the broader phenomena of volunteered geographic information, neogeography, and the Web 2.0. I address key issues relating to data quality in OSM and review the large volume of past work that has addressed this topic. I next focus more closely on the issue of temporal accuracy in OSM and discuss the dynamics of editing. I then focus on the case of humanitarian mapping and discuss the applications of OSM in humanitarian contexts and the unique modes of data production in this domain. I conclude by situating the work of this thesis in the research gap that exists at the intersection of data maintenance, as a dimension of temporal data quality, and humanitarian mapping efforts.

2.1 Introduction to OSM

OSM is an example of what Goodchild (2007) terms, "volunteered geographic information" (VGI). VGI sits under the umbrella of "neogeography", in which the democratization of tools for geospatial data production and consumption lead to a wealth of citizen-generated geospatial datasets (Goodchild 2009; Haklay, Singleton, and Parker 2008). More broadly, neogeography is enabled by the rise of the Web 2.0, in which the lines between content production and content consumption on the web are blurred (O'Reilly 2009).

Popularly framed as the "Wikipedia of maps", OSM seeks to empower individuals to share their local spatial knowledge to create a crowdsourced map of the

world (K. Fox 2012; Lu 2019). Theoretically, anyone with access to the internet can contribute to OSM. At the time of writing, OSM has over 6 million registered users (although it is likely that not all users have contributed data) and almost 8 billion uploaded GPS points (*OpenStreetMap Statistics* 2020). OSM offers a free alternative to proprietary geospatial datasets, and is used for purposes such as vehicle routing (Graser, Straub, and Dragaschnig 2015; Luxen and Vetter 2011) and POI searching (Ruta et al. 2015). The OSM contribution landscape is also very heterogeneous. Large-scale, existing geospatial datasets can be imported into OSM, such as the US TIGER import in 2008 (Zielstra, Hochmair, and Neis 2013). Moreover, corporate entities; such as Facebook, Microsoft, and Apple; are increasingly involved in mapping efforts (Anderson, Sarkar, and Palen 2019). Recent efforts, such as Facebook's *mapwith.ai* tool[1], have also sought to employ machine learning techniques to automatically (or with minimal human supervision) create OSM features from satellite imagery (Albrecht et al. 2020; Yadav et al. 2020).

2.2 Data Quality and OSM

As a crowdsourced dataset, one of the primary potential issues with OSM is its quality. OSM does not provide any assurances of its quality; unlike traditional, authoritative geospatial datasets. Moreover, its contributors do not require any formal training or qualifications. Questions of data quality are also particularly relevant and challenging to address in this context because of the highly diverse nature of contributions and contributors, leading to variable quality throughout (Girres and Touya 2010; Gröchenig, Brunauer, and Rehrl 2014a; Haklay 2010; Neis and Zipf 2012).

Existing literature identifies numerous dimensions of geospatial data quality, such as positional accuracy, completeness, logical consistency, temporal accuracy, and usability (Antoniou and Skopeliti 2015; C. Fox, Levitin, and Redman 1994; Oort 2006). Questions of VGI data quality are also framed around the concepts of trust and credibility, reminding one of the presence of the data user who must

[1] `https://mapwith.ai/`

evaluate the fitness of the data for their task at hand (Flanagin and Metzger 2008; Severinsen et al. 2019). While a characteristic such as positional accuracy can be empirically evaluated, trust and credibility are perceptual qualities of a dataset that relate to its "believability" in the eyes of the data consumer (Flanagin and Metzger 2008). Barron, Neis, and Zipf (2014) also address the notion of fitness for use in their consideration of geospatial data quality, demonstrating how different applications of a given dataset will require different quality needs.

The framework laid out by Goodchild and Li (2012) is particularly useful in understanding the mechanisms for quality control in VGI projects. The authors outline the following three approaches: 1) the *crowdsourcing approach*, as evaluated by Haklay, Basiouka, et al. (2010), by which a community of contributors will converge on the "truth" and correct the errors of others; 2) the *social approach*, by which contributors are organized in a hierarchy with those at the top acting as content moderators or gatekeepers; and the 3) *geographic approach*, by which common-sense rules about the nature of geographic phenomena are used to filter out clear errors (Goodchild and Li 2012). This framework demonstrates how the community structure of VGI projects such as OSM can contribute to enhanced quality control, however additional technical checks and evaluation may be needed.

Efforts to empirically assess the quality of OSM data began with so-called "extrinsic" approaches, whereby OSM data is compared against an authoritative dataset of assumed high quality. Haklay (2010) compares the completeness and positional accuracy of OSM data in England with that from the Ordnance Survey. Girres and Touya (2010) extend this analysis to the French OSM dataset and Zielstra and Zipf (2010) compare OSM data in Germany to that from the TeleAtlas MultiNet dataset. Overall, these works find OSM data to be of relatively high quality, however quality can also be quite variable across both geographic space and across the different elements of geospatial data quality (Girres and Touya 2010; Haklay 2010). For example, Zielstra and Zipf (2010) find significant differences in completeness between urban and rural areas in Germany, with rural areas needing greater coverage.

More recent efforts to assess OSM data quality are trending towards intrinsic

approaches. Such efforts can be defined as "process-based measures focusing on pragmatic or contextual 'authority' by examining the processes generating information" (Anderson, Soden, et al. 2018, p. 297). Intrinsic efforts may be preferable due to the potential high cost of obtaining proprietary datasets, or the lack of availability of such reference datasets (Estes and Mooneyhan 1994). Intrinsic quality assessment may also be more appropriate as one acknowledges that authoritative, reference datasets may not be of sufficiently high quality themselves, as suggested by Goodchild and Li (2012, p. 112). One of the first efforts at intrinsic quality assessment was conducted by Haklay, Basiouka, et al. (2010), who empirically evaluate the positive (although non-linear) relationship between positional accuracy and the number of contributors for a given local region. The "Crowd Quality" framework suggested by Exel, Dias, and Fruijtier (2010) demonstrates how many intrinsic approaches will use characteristics of OSM contributors and history of the data as an indication of quality. Barron, Neis, and Zipf (2014) offer one of the most comprehensive frameworks for intrinsic quality assessment through their *iOSMAnalyzer* tool, which combines over 25 data quality indicators that are tailored to different application areas.

2.3 Temporal data quality and the evolution of OSM data

Temporal accuracy, or up-to-date-ness, is an important element of geospatial data quality (Oort 2006). As stated by Barron, Neis, and Zipf (2014, p. 884), "Ideally the process of updating the OSM features' geometries and attributes is carried out continuously, homogeneously, throughout and is not limited to specific features." This statement leads one to consider temporal accuracy as the "validity" or "currency" of data at a given point in time (Oort 2006, p. 17). Put simply, when the real world changes but its associated geospatial data does not, it can be considered as out of date (or temporally inaccurate).

Dimensions of temporal accuracy, however, have received comparatively little direct focus in existing work on OSM. Early quality assessments of OSM largely

focused on completeness and positional accuracy of the data, as asserted by Neis and Zielstra (2014, p.83), and exemplified by the work of Haklay, Basiouka, et al. (2010), Haklay (2010), and Helbich et al. (2012). The temporal dimension of OSM (ie. the evolution of its data over time) has been given attention, but this focus has mostly been to look at aspects of data quality such as completeness (Gröchenig, Brunauer, and Rehrl 2014b). While completeness and positional accuracy of OSM are undeniably critical for many applications, one must not disregard the many other elements of geospatial data quality, which may need to be prioritized in some applications.

Nevertheless, some selected works have made an effort to address temporal accuracy in OSM data. These efforts are largely intrinsic and investigate factors such as the relationship between volume of active contributors in a region and the frequency of feature updates (Girres and Touya 2010). Past work has also used a feature's date of last edit as an indication of its recency (Minghini, Brovelli, and Frassinelli 2018; Roick, Loos, and Zipf 2012). However, intrinsically assessing a feature's currency by looking at its date of last edit may be an inaccurate assessment, as there is no indication of whether or not the corresponding real-world feature has changed since that date (Barron, Neis, and Zipf 2014). In areas where reference data is not available, this shortcoming suggests a need develop more in-depth understandings of the dynamics of contributions and the lineage of the dataset.

When considering temporal accuracy, it is valuable to take a step back to think more deeply about the dynamics of editing OSM. As in the real world, geospatial entities in OSM do not necessarily remain consistent over time. The preservation of historical OSM data has allowed the characteristics and dynamics of change to be the subject for various research efforts. As the speed of mapping varies significantly over space, one can consider OSM as containing a multitude of overlapping maps, each along different trajectories to completeness (Chuang et al. 2013). The lineage of OSM data thus comes to the forefront of research efforts, such as in the case of Mooney and Corcoran (2012), who investigate the characteristics of features that have histories of significant revision; or with Trame and Keßler (2011), who use

heat maps to visualize the heterogeneous spatial patterns of feature edits over time. While this work reveals highly dynamic processes of editing and contributing, there has been evidence of stagnation in some parts of OSM (Gröchenig, Brunauer, and Rehrl 2014a). While reductions in editing activities may indicate saturation in the dataset (and thus suggest high completeness), this may also indicate places with low contributor engagement, as people are unwilling or unable to maintain the data (Gröchenig, Brunauer, and Rehrl 2014a).

Efforts to understand the prevalence of data maintenance offer a promising approach for helping us to evaluate the temporal accuracy of OSM data for a given area. McConchie (2013) provides a powerful metaphor, likening data maintenance in OSM to a process of "map gardening". Accordingly, one can understand maintenance as the activity by which data is kept up-to-date. Ongoing data maintenance is thus necessary for ensuring that OSM data is of high quality. The identification of data maintenance activities is grounded in the distinction between different forms of OSM contributions, used to identify different phases in the development of the map for a given region (Anderson, Soden, et al. 2018). It is thought that a data maintenance phase is reached once the map for a given area reaches a sufficient level of maturity, when the bulk of editing activity shifts from the addition of new features to the modification of existing ones (Anderson, Soden, et al. 2018). Quattrone, Dittus, and Capra (2017) offer the most in-depth assessment of data maintenance of OSM, looking globally at maintenance practices of point of interest (POI) data. The need for maintenance is also not unique to OSM, and has been identified in other crowd-produced knowledge repositories, such as Wikipedia (Kittur et al. 2007).

2.4 OSM data production and use in humanitarian contexts

It is acknowledged that quality assessments of OSM data should be grounded in specific application area(s) (Barron, Neis, and Zipf 2014). In this work, I focus specifically on the context of humanitarian mapping contexts.

High-quality information is critical to successful humanitarian and disaster re-

sponse work (Cowan 2011; Poser and Dransch 2010; Soden and Palen 2016; Zook et al. 2010). Geospatial information is necessary for functions such as promoting situational awareness during a crisis (for example, to alert responders to the locations of flood damage, as in Poser and Dransch (2010)) and ensuring access to resources (for example, by helping to manage the deliveries of supplies to remote villages in Nepal following the 2015 earthquake, as in Soden and Palen (2016)). OSM data is particularly relevant in this context as many of the locations in need of humanitarian assistance do not have any authoritative geospatial data available (Zook et al. 2010).

Following the 2010 Haiti earthquake, for example, responders needed more detailed and accurate geospatial information about the affected area (Meier 2012; Soden and Palen 2016; Zook et al. 2010). Following rapid, remote volunteer efforts to trace recent satellite imagery, OSM became the most complete and up-to-date map of the impacted area (Soden and Palen 2014). This OSM data was used by many aid agencies, including UNICEF and OCHA (Soden and Palen 2014). These efforts led to the formation of the Humanitarian OpenStreetMap Team (HOT), which has since led numerous humanitarian mapping efforts in response to natural disasters such as earthquakes and epidemics (Dittus, Quattrone, and Capra 2017).

Such cases of crisis mapping constitute a distinct sub-community within OSM, with unique modes of community organization and data production. Firstly, mapping efforts are largely driven by discrete tasks (Vyncke 2020). This is exemplified in the HOT Tasking Manager, which is a web-based tool that hosts mapping tasks, where each mapping task is accompanied by a detailed and specific set of instructions (Humanitarian OpenStreetMap Team 2020a). These tasks are often driven by the needs of organizations such as Médecins Sans Frontières (MSF) and the Red Cross who are directly engaged in humanitarian work in the field (Humanitarian OpenStreetMap Team 2020a). Secondly, the majority of contributors are remote (Eckle and Albuquerque 2015; Vyncke 2020). Remote contributors add basic geometries (such as building footprints or roads) by tracing satellite imagery (Vyncke 2015). While it is a best-practice in the humanitarian community to build and en-

gage communities of local stakeholders in the mapping process (as Soden and Palen (2014) describe in the Haiti example), this may not always be feasible to the extent desired (Vyncke 2020). Thirdly, the data is often more collaboratively produced than in traditional OSM contexts (Poiani et al. 2016; Vyncke 2020). While all of OSM is a collaborative effort between the entire community of contributors, geographically dispersed humanitarian mapmakers must often be more explicitly coordinated through tools such as the HOT Tasking Manager and mailing list (Palen et al. 2015). Lastly, areas under humanitarian focus are likely to show a distinct spatial pattern, whereby an area of interest has much more coverage than nearby regions on the map (Anderson, Soden, et al. 2018).

These distinct characteristics of data production in OSM during humanitarian mapping efforts have implications for the quality of the data. When one considers data quality as a dataset's "fitness for use", the rapidly changing circumstances in crisis scenarios means that elements of temporal data quality are particularly relevant (Chen et al. 2008). The bursts of mapping work done by volunteers in the wake of a crisis, as discussed by Dittus, Quattrone, and Capra (2017), may provide more up-to-date spatial data than previously existed (Soden and Palen 2014). For this reason, the Government of Nepal's Survey Department published maps containing OSM data on official platforms following the 2015 earthquake (Soden and Palen 2016). However, one can speculate that this rapid creation of new data by remote volunteers may mean that it is challenging to be maintained over time (especially if it is very detailed), and is thus more likely to be out-of-date when the crisis subsides. Efforts that directly assess the quality of data contributed to OSM in humanitarian mapping efforts have been limited. Moreover, there is need for research that directly addresses issues of data maintenance during these mapping efforts.

2.5 Summary of research opportunity

This literature review demonstrates how OSM is a valuable data source that is particularly relevant in humanitarian and disaster relief operations. However, OSM's lack of official mechanisms for quality assurance and crowdsourced nature may

raise concerns around its quality. As evidenced by the large volume of past work, thinking about geospatial data quality in OSM is not new. However, there is a need for more in-depth considerations of temporal accuracy in OSM data, particularly those that consider the unique modes of data production in humanitarian contexts. The concept of data maintenance offers a means to investigate temporal accuracy, and is thus the primary focus of this work. To our knowledge, this work is the first effort to explicitly quantify data maintenance at a local scale following humanitarian mapping campaigns.

Chapter 3

Data Description

This brief chapter provides a summary of the ways that spatial data is structured within OSM. I focus particularly on historical OSM data, which is used in this analysis. This review offers background information that is useful for understanding the methodology described in the following chapter.

While OSM is commonly engaged with as a user-facing map, it in fact functions as a highly flexible and detailed geospatial database (OpenStreetMap Wiki 2018a). A diagram of the technical components can be found in Appendix A. The OSM data model uses nodes, ways, and relations to represent the geometry of all real-world geographic features (OpenStreetMap Wiki 2020a). Nodes represent point features, such as shops, healthcare facilities, and bus stops (OpenStreetMap Wiki 2019a). Ways are ordered collections of nodes, commonly used to represent features such as roads (OpenStreetMap Wiki 2020d). Closed ways (where the start and end node are the same) are used to represent areal features, such as buildings (OpenStreetMap Wiki 2020d). Relations are used to represent relationships between multiple data elements, and may be used to represent entities such as turn restrictions between road sections (OpenStreetMap Wiki 2019b).

The attributes for all OSM elements are stored using tags, which consist of text key, value pairs. An element can have multiple tags, however each key for a single element must be unique (OpenStreetMap Wiki 2020c). While OSM does not impose any restrictions on the contents of a tag (aside from being a 255-character Unicode string), it is a best-practice within the community to follow established tagging

conventions for commonly occurring elements (OpenStreetMap Wiki 2020c). For example, the `highway=residential` tag is used to describe roads that provide access to homes (OpenStreetMap Wiki 2020a). The Taginfo website allows one to see commonly used tags across the world (OpenStreetMap Contributors 2020). Each OSM element also contains metadata such as the timestamp of last edit, version number, and user ID of the contributor (OpenStreetMap Wiki 2020a).

OSM data is commonly downloaded from services such as Geofabrik[1] or Planet OSM[2] in SHP, PBF, or OSM XML format (Mooney and Corcoran 2011). An example of the OSM XML format can be found on the relevant OSM Wiki page (OpenStreetMap Wiki 2017). From these services, one can download a copy of the database for the full planet, or from selected continents or countries. These files are usually frequently updated (eg. daily, in the case of Geofabrik (GeoFabrik 2020)) to ensure that new contributions or edits to OSM are reflected in the downloadable data. One can download both current snapshots and full historical OSM data from these services. The historical data provides access to all edits that were ever made in OSM and can allow one to understand various aspects of how the map has matured over time (Corcoran, Mooney, and Bertolotto 2013; Mooney and Corcoran 2012). Researchers are able to understand the evolution of individual features on the map, as each feature has an associated version number that allows one to track all revisions (OpenStreetMap Wiki 2020a).

Analyzing historical OSM data has traditionally been challenging, in part due to the incredibly large file size (Raifer et al. 2019; Mooney and Corcoran 2011). The processing described in Mooney and Corcoran (2012), for example, took 305h. At the time of writing, the full history planet OSM XML file is 140 GB, which is too large for many users or researchers to manage effectively without significant technical expertise. As is surveyed by Raifer et al. (2019), researchers and practitioners have developed numerous tools and algorithms that allow for simpler processing and querying of this large historical dataset. Examples include the *Is OSM Up-To-Date?* website developed by Minghini, Brovelli, and Frassinelli (2018), and the

[1] `https://www.geofabrik.de/`
[2] `https://planet.osm.org/`

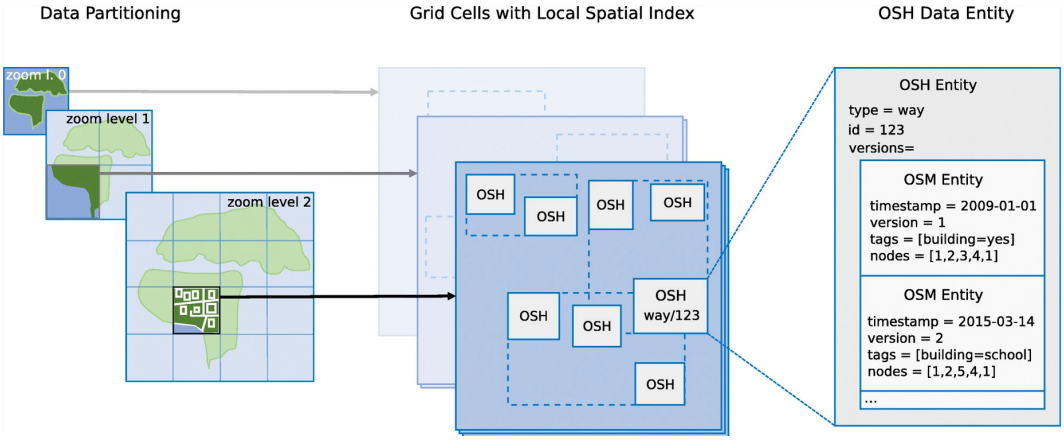

Figure 3.1: Summary of the OSHDB data model, provided by Raifer et al. (2019).

OSMatrix website developed by Roick, Loos, and Zipf (2012).

The current state-of-the-art in this domain is the recently developed OSHDB framework, which provides a flexible and fast way to perform spatio-temporal analyses on historical OSM data (Raifer et al. 2019). The OSHDB framework allows for the extracted historical OSM data to be stored in any JDBC (Java database connectivity) compliant database system and accessed through an API (Raifer et al. 2019). The authors developed a custom data model to allow for more efficient access and parallel processing, as is shown in Figure 3.1. Each version of an OSM Entity is grouped by a common ID into a parent OSH Entity. Each OSH Entity has a type that corresponds to one of the OSM data types (nodes, ways, or relations) (Raifer et al. 2019). Application of the OSHD framework for this analysis will be described in the following chapter.

Chapter 4

Methodology

This section describes the procedures used to collect and analyze historical OSM data to address the reseach questions identified in Chapter 1. I begin by describing the process of selecting case studies and then detail my procedures for collecting and cleaning relevant historical OSM data. I then describe the empirical approach used to analyze and visualize this data to understand characteristics of data production and subsequent data maintenance efforts. I considered each case study (mapping campaign) to be a *unit of analysis*, within which a *unit of observation* was a unique OSM entity that was created during a given mapping campaign. All data processing and analysis was done using the Java and R programming languages. Data visualization was done using R and QGIS. For reproducibility purposes, all code used for data collection and analysis can be accessed from a public GitHub repository.[1] This repository also includes samples of the data that was used in this analysis.

4.1 Case study approach

I applied a comparative case study analysis throughout this work, investigating four humanitarian mapping campaigns and one reference case study. Guided by the approach set out in Kaarbo and Beasley (1999), I conducted a focused and structured comparison of cases and looked for patterns in variables within and across cases. I investigated similarities and differences within the humanitarian cases, and between the humanitarian cases and the reference case. This case study approach is appro-

[1] `https://github.com/hannahker/osm-maintenance`

priate for this project's research aims as it allows us to refine and develop existing theories relating to humanitarian mapping practices (Kaarbo and Beasley 1999). Given the lack of existing methodological framework for empirically assessing data maintenance in OSM, this approach was largely exploratory and iterative.

I selected humanitarian case studies that consisted of mapping efforts in response to a humanitarian need, were constrained to subnational geographic areas, drove a sufficient volume of mapping activity on OSM (as defined by volume of unique contributors and volume of edits over time), and were sufficiently documented on community organizing pages such as the OSM Wiki[2], HOT Projects page[3], or HOT Tasking Manager[4]. Case studies were also selected with the intent to capture practices of humanitarian mapping at various stages in the history of the humanitarian OSM community. Case studies were also required to have been completed at least four years ago, to allow for the study of maintenance practices in the years following a mapping campaign.

Following from the above criteria, I focused on humanitarian mapping activities in 1) Port au Prince, Haiti, following the 2010 earthquake, 2) Tacloban, Philippines, following the 2015 typhoon, 3) Bangui, Central African Republic, following 2013 rebellions, and 4) Kathmandu, Nepal, following the 2015 earthquake. In addition to these four humanitarian case studies, I have also selected a reference case study as a contrasting example of mapping activities in a region with established high quality data (Arsanjani et al. 2013). Following Anderson, Soden, et al. (2018), I selected Heidelberg, Germany as the reference case.

4.2 Processing historical OSM data

This analysis is based on historical OSM data, as described in the previous chapter. Unprocessed OSM historical data extracts for each case study area were downloaded from Geofabrik (GeoFabrik 2020). This "raw" data was then processed using the OSHDB framework (Raifer et al. 2019). OSHDB was selected due to the

[2] https://wiki.openstreetmap.org/wiki/Main_Page
[3] https://www.hotosm.org/projects/
[4] https://tasks.hotosm.org/

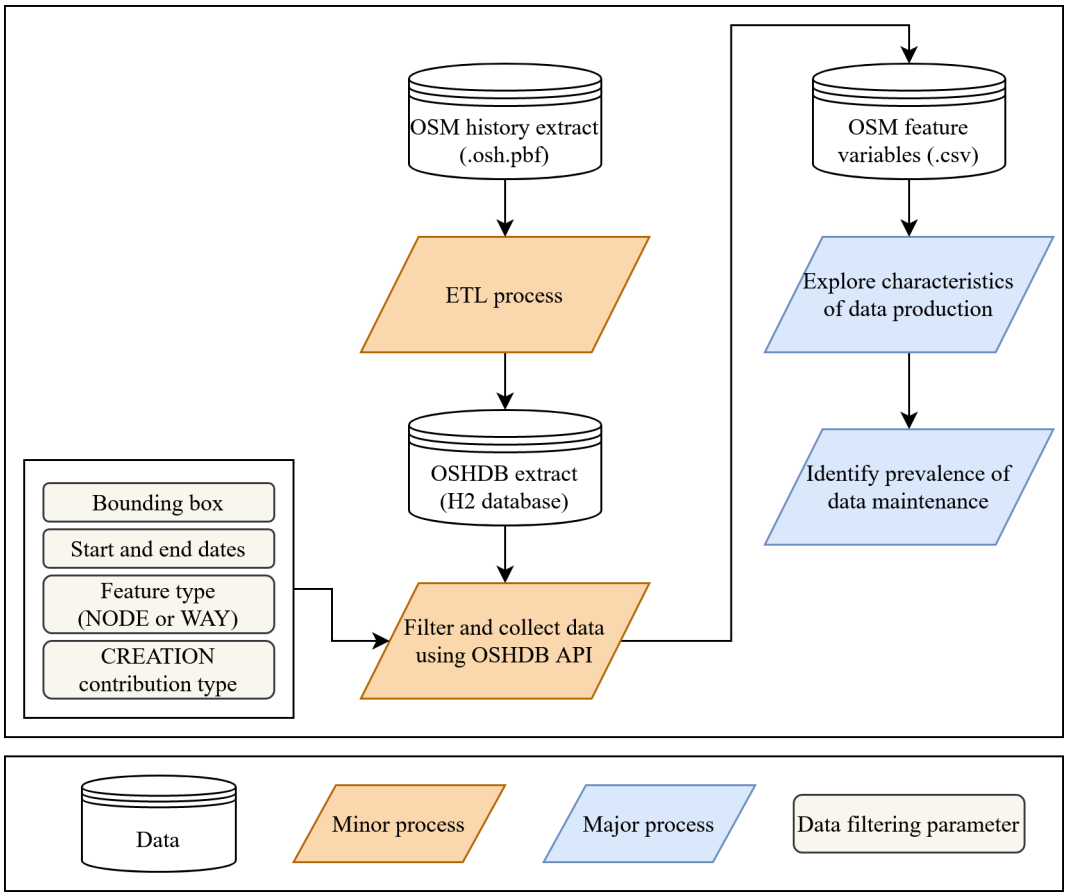

Figure 4.1: Summary of data processing pipeline.

speed and flexibility that it provides in processing and filtering historical OSM data. Despite these advantages offered by the OSHDB framework, the size and complexity of OSM history data means that the extraction of relevant variables constituted a significant portion of this methodology. The technical approach to processing data, including specific programming approaches, was significantly aided by the examples and documentation provided on the OSHDB GitHub page (Heidelberg Institute for Geoinformation Technology 2020a).

To begin, each OSM history extract, in *.osh.pbf* format, was converted to a local OSHDB instance, following the Extract, Transform, Load (ETL) process described in the OSHDB documentation (Heidelberg Institute for Geoinformation Technology 2020b). This process loads the data from each extract into separate local H2 databases. Some of the historical extracts were too large to be processed locally in this manner and so technical assistance in generating some extracts was

Table 4.1: Details of the spatial and temporal extents used to filter data for each case study.

Name	Start	End	Bounding box
Heidelberg	2008-12-31	2009-12-31	8.57, 49.35, 8.79, 49.46
Port au Prince	2010-01-12	2011-10-31	-72.57, 18.34, -72.16, 18.63
Bangui	2013-03-23	2015-12-31	18.49, 4.32, 18.59, 4.49
Tacloban	2013-11-10	2014-01-31	124.89, 11.18, 125.08, 11.34
Kathmandu	2015-04-25	2015-12-31	85.27, 27.67, 85.38, 27.75

provided by a researcher from the Heidelberg Institute for Geoinformation Technology (HeiGIT) team.

The OSHDB API (Raifer et al. 2019) was then used to filter and process the historical data to obtain variables of interest. Implemented in the Java programming language, this API allows for data filtering and aggregation based on the *MapReduce* programming framework (Raifer et al. 2019). This framework is designed for use with large datasets and contains a *map* function whereby data is filtered and sorted, followed by a *reduce* function whereby data is summarized and returned as aggregated values (Dean and Ghemawat 2008).

I collected data for each case study using the spatial and temporal extents specified in Table 4.1. Bounding boxes correspond to the smallest square area encompassing the city of interest (here with coordinates rounded to two decimal places for greater legibility). The start and end date for each of the humanitarian featuress was collected from the associated HOT project page, which includes a brief summary of each campaign, along with basic details such as the dates and completion status (Humanitarian OpenStreetMap Team 2020b). The start and end dates for the Heidelberg reference case were selected to cover a year-long period that was relatively early in the development of the map for this area, to be better compared against the humanitarian cases.

For each of the case studies, I focused solely on collecting the attributes from new data that was produced in each campaign. I also focused only on nodes and ways, given that they represent features of humanitarian interest, such as roads, buildings, and healthcare facilities. Using the OSHDB's *stream()* functionality, numerous variables were collected from all OSM features that were created during

Table 4.2: Summary of variables collected from OSM entities.

Variable	Description	Data type
ID	Unique identifier for OSM/OSH entity	Numeric
User ID	The user ID of the OSM contributor	Numeric
Bounding box	Bounding box for the OSH entity	Coordinates
Type	Node or way	Text
Version number	Version number for the OSM entity	Numeric
Timestamp	Time of feature creation	Date/time
Tags	Tags and keys associated with the OSM entity	Text
Visibility	Indicating whether or not the node has been deleted	Boolean

this time, and all subsequent versions of these features that were created in the time following the mapping campaigns. For example, details would be collected about OSM entity, *Node X*, that was created during the mapping efforts in *Case Study Y*. At its initial creation, *Node X* would have a version #1. By grouping together all versions of a given OSM entity within a parent OSH entity, the OSHDB data model also allowed us to collect information from all subsequent versions of *Node X* (which are distinct OSM entities themselves). These subsequent versions reflect modifications that were made to *Node X* after it was initially created, such as modifications to its geometry or tags. The variables collected for each OSM entity are summarized in Table 4.2.

The information captured by these variables is diverse, encompassing numerical, spatial, temporal, and attributional information. While most variables are highly structured, the information contained within the *Tags* variable requires significant cleaning to be useful. Across all case studies, I have collected data from 495,891 OSM entities. Many of these entities reflect subsequent versions of the entities created during each campaign.

4.3 Exploring the characteristics of data production

This historical OSM data was then analyzed to understand the basic characteristics of data production for each mapping campaign. Numerous data processing steps

were performed to understand how the data was produced over time, where the data was produced over space, what sources contributed to data production, and what types of geographic features were produced. Numerous summary statistics for each case study were also calculated to allow for quantitative comparison.

To understand the dynamics of data production over the duration of each mapping campaign, all features were aggregated by the day that they were produced. I calculated the total number of features produced and the unique number of contributors each day. I then calculated the relationship between these two variables, by day, using the Pearson correlation coefficient.

I investigated the patterns of data production over space by visualizing the density of new features created within each study area. I created a dot-density map for each case study, as this effectively shows the geographic distribution of contributions in each study area (Kimerling 2009). I calculated the centroid from each feature's bounding box to create this point-based density visualization.

I parsed the tags associated with each feature to understand the types of features that were mapped and their associated sources. While tagging conventions within the OSM community result in many standardized tags used across features, some basic text preprocessing was conducted to ensure that all text was normalized as much as possible, allowing for more accurate aggregation. I converted all characters to lowercase and removed non alpha-numeric characters. For each case study, I calculated the most frequently occurring tag keys (and not the associated values) to provide an indication of the types of features that were mapped. I also analyzed the values that are associated with the `source` tag key across features. While no longer a common tagging practice, the `source` tag key has frequently been used within the OSM community as a way to indicate the information source behind a given contribution (OpenStreetMap Wiki 2020b). For example, features that were added to OSM by tracing Bing satellite imagery might be accompanied by the `source=Bing` key-value pair. I calculated the most frequently occurring source values for each case study to provide an indication of the common information sources. As in Ahmouda, Hochmair, and Cvetojevic (2018), this information

Table 4.3: Metrics calculated to compare mapping activity across case studies.

Metric	Unit	Formula
Duration	*days*	End date - start date
Area	km^2	Bounding box length * width
Total features created	*features*	NA
Total unique contributors	*contributors*	NA
Burstiness	*days*	Number of days until 50% of all contributions were made (Dittus, Quattrone, and Capra 2017)
Campaign style	*EVENT* or *MISSION*	Event if burstiness 60, or Mission if burstiness = 60 (Dittus, Quattrone, and Capra 2017)

can contribute to an understanding of whether or not a contributor is remote, based on the presence of satellite imagery sources.

In addition to the above analyses, various metrics were calculated to provide a basis for empirical comparison between case studies. These metrics are summarized in Table 4.3. I calculated the total duration and size of study area, as well at the total number of unique features mapped and unique contributors. I also follow Dittus, Quattrone, and Capra (2017) in calculating the "burstiness" of each case study and classifying each case study as an event or mission. The burstiness measure provides insight into the dynamics of data production and is particularly relevant in the context of humanitarian mapping, as data is often produced very quickly over time in response to a crisis (Dittus, Quattrone, and Capra 2017).

The analysis described in this section will allow for a greater understanding of the many dimensions of data production during humanitarian mapping campaigns. Comparison between the humanitarian case studies and the Heidelberg reference will allow for the identification of difference and similarity between mapping in humanitarian and non-humanitarian contexts. This understanding will provide valuable context, allow for a more meaningful interpretation of the results of the following section on data maintenance following mapping campaigns.

Table 4.4: Classification scheme of maintenance frequency.

Number of versions	Maintenance frequency
1	Never
2	Once
3-10	Moderate
11-19	Frequent
20+	Extreme

4.4 Identifying data maintenance activities

I analyse the historical OSM data for each case study to understand the extent to which features are maintained following each case study mapping campaign. I define data maintenance in OSM to be the practice by which a given feature already existing within the database is updated, presumably to reflect a real-world change that has taken place in the corresponding geographic feature. For example, following a building renovation, the building's footprint may have changed, requiring updating to the geometry of the building's polygon in OSM. I operationalize this definition by following Quattrone, Dittus, and Capra (2017), in considering data maintenance to have occurred any time that an OSM entity has a version greater than 1. This definition leads to a binary consideration of maintenance, in that a given feature is considered to be *unmaintained* if it only has one version (indicating that it has not been modified since its initial creation), and is considered to have been *maintained* if it has at least two versions (indicating that it has been modified at least once since its initial creation). Note that entitiy deletions, in addition to modifications, also result in a new entity version being created.

In addition to this binary perspective, I consider different degrees of maintenance across features by examining the total number of versions that they each contain. I am thus able to make a distinction not only between maintained and unmaintained features, but also between those that are more frequently maintained than others (ie. those that have a greater number of versions). Based on the distribution of version numbers across all features in the dataset, I create a classification of maintenance frequency, as detailed in Table 4.4.

I also evaluated data maintenance over time, aiming to better understand *when*

maintenance occurs following a mapping campaign. To normalize across each of the case studies, I consider the four-year period following each of the mapping campaigns. At cumulative yearly intervals, I calculate the percentage of all features that have been maintained at least once and the distribution of features belonging to each maintenance frequency class in Table 4.4.

I then conduct a disaggregated evaluation of data maintenance by investigating the types of features that are maintained more frequently than others. I consider differences in the maintenance frequency between nodes and ways. I also consider the differences between features with `building`, `highway`, and all other features without either of these tags (all grouped together). I calculate the maintenance frequency distribution for each type of feature after four years have passed since each mapping campaign.

Following these efforts to quantify data maintenance across each of the case studies, the results are interpreted with respect to the findings from the previous section. I consider the results from Sections 4.3 and 4.4 to develop informed hypotheses about how the modes of data production during humanitarian mapping campaigns may have an impact on the prevalence of data maintenance. Given the limited scope of this work, these hypothesis cannot be empirically validated, and so should be explored in greater depth in future work.

4.5 Ethical considerations

Following UCL guidelines[5], this project was deemed to be exempt from a formal ethics review as it constitutes a non-invasive and non-interactive observation of public behaviour (editing processes on OSM). This methodology does not involve identifying individuals in any way that could place them at risk of harm, stigma, or prosecution. While this research includes OSM User IDs which may uniquely identify individuals, this analysis considers this data from an entirely aggregated perspective. The sample data that is published on GitHub for reproducibility purposes has been entirely anonymized (all User IDs have been replaced with the integer, 1).

[5] `https://ethics.grad.ucl.ac.uk/exemptions.php`

The historical OSM data used in this research was accessed in accordance with the guidelines set out on the *Geofabrik* download server. This research constitutes a quality assurance effort that is supported by the OSM community (Missing Maps), and does not present any derived works based on these files.

Chapter 5

Results

In this chapter, I present the results of the analysis conducted in this research project. I begin by providing background context for each of the selected case studies. I then summarize the characteristics of the data produced in each case study and provide an overview of the dynamics of data production over the duration of each mapping campaign. I present results that demonstrate the prevalence of data maintenance following each campaign.

5.1 Case study context

Prior to providing the empirical results of this analysis, I briefly address the humanitarian context in each of the case studies and their relevance to the humanitarian mapping community. The geographic location and spatial distribution of mapping activity for each case study can be seen in Figure 5.1. Full-size versions of each inset map can be found in Appendix B.

5.1.1 Port au Prince, Haiti

Haiti experienced a magnitude 7.0 earthquake on January 12, 2010, which caused an estimated 300,000 deaths, and widespread building damage and population displacement (DesRoches et al. 2011). It is estimated that this event has caused USD $8.1bn damage (Cavallo, Powell, and Becerra 2010). Humanitarian mapping efforts in Haiti following this earthquake have been well researched and discussed in past academic literature (Zook et al. 2010; Soden and Palen 2014; Palen et al. 2015; Meier 2012). This disaster has been described as a "catalyzing event" for

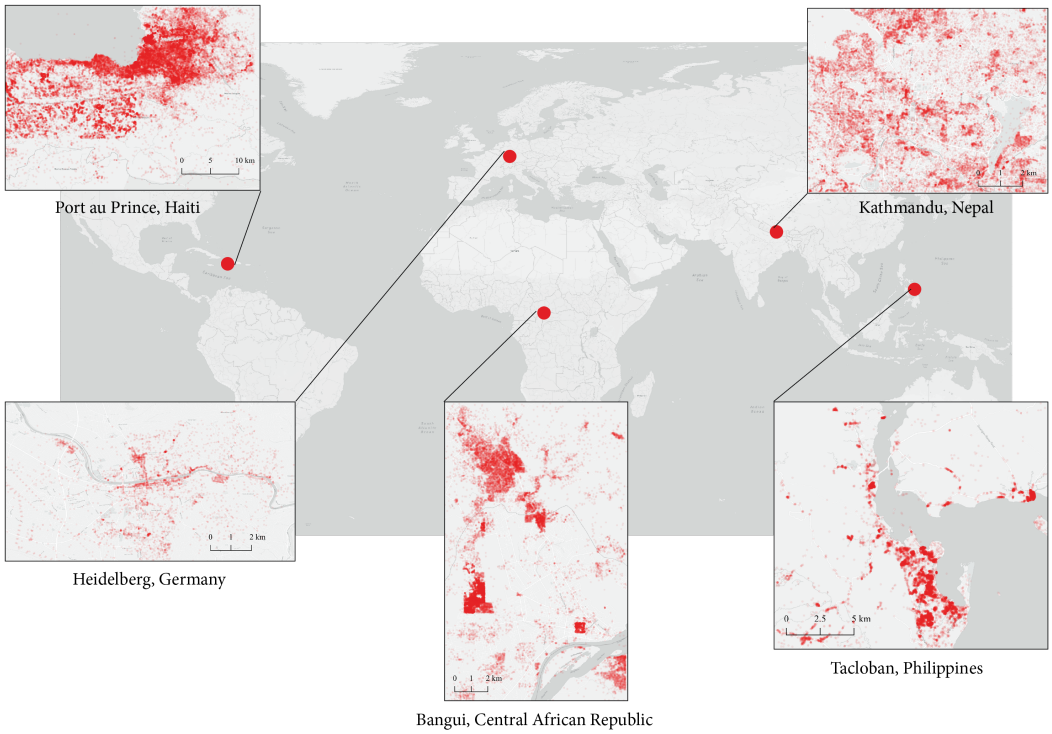

Figure 5.1: Map of case study locations and spatial distribution of features added. Greater intensity of colour corresponds to a greater density of features added to OSM during each mapping campaign.

many digitally-focused volunteer communities (Soden and Palen 2014, p. 314). Mapping efforts around this event also led to the formalization of the Humanitarian OpenStreetMap Team (HOT), the process of which is described in further detail by Soden and Palen (2014). Throughout their post-disaster efforts to raise awareness of the value of OSM and mobilize a community of mappers, one of HOT's primary goals was to "embed" OSM within the local community and further local ownership of this data (Soden and Palen 2014, p. 319). This effort was intended to allow for the long-term use of OSM data beyond this humanitarian response (Soden and Palen 2014). As shown by Figure 5.1, the features mapped within this study area are largely clustered around the coast of Port au Prince.

5.1.2 Tacloban, Philippines

The Philippines was greatly impacted by a tropical cyclone, Typhoon Haiyan (or Typhoon Yolanda), on November 8, 2013. This typhoon is said to be one of the strongest ever recorded (Lum and Margesson 2014). USAID estimates that this

disaster has caused over 6,000 deaths and the destruction or damage of over 1 million homes (*Typhoon Haiyan/Yolanda Fact Sheet #20* 2014). The city of Tacloban was one of the areas that faced greatest impact and was thus where much relief effort was focused (Lum and Margesson 2014), as shown by the clustering of features on the inset map in Figure 5.1. Following this crisis, mapping efforts in OSM were coordinated by HOT, with high-volume, remote mapping efforts organized by the newly developed Tasking Manager (OpenStreetMap Wiki 2018b). Palen et al. (2015) note that the mapping efforts in the Philippines were facilitated by these new tools for technical collaboration, which incorporated lessons learned from previous humanitarian mapping efforts, such as in Haiti.

5.1.3 Kathmandu, Nepal

Nepal was hit with a magnitude 7.6 earthquake on April 25th, centered approximated 76 km northwest of Kathmandu, which was followed by over 300 aftershocks of over 4.0 magnitude (*Nepal Earthquake 2015: Post Disaster Needs Assessment* 2015). It is estimated that over 9,000 people died in these disasters and over half a million homes were destroyed or damaged (*Nepal Earthquake 2015: Post Disaster Needs Assessment* 2015). As is described by Soden and Palen (2016) this crisis can be viewed as a turning point in the history of post-disaster mapping in the OSM community. Whereas in the Haiti case HOT needed to conduct notable outreach to spread awareness of the applicability of OSM data, interviews with GIS practitioners in the field found that up-to-date OSM data came to be an "expected resource" in Nepal (Soden and Palen 2016, p. 2801). Figure 5.1 shows how the mapping activity is decentralized throughout Kathmandu.

5.1.4 Bangui, Central African Republic

Violence and instability in CAR mounted in March 2013 when the Seleka rebel group seized the capital city, Bangui (Global Conflict Tracker 2020). This event launched a humanitarian mapping campaign that aimed to provide baseline geospatial data for the country (OpenStreetMap Wiki 2020e). Mapping the country's road network was a priority of this campaign, as well as mapping affected cities

and towns, as identified by local humanitarian stakeholders (OpenStreetMap Wiki 2020e). The features mapped within Bangui are clustered within the periphery of the area, as shown in Figure 5.1. UNICEF data for health facilities, water points, and schools was also imported into OSM as part of this campaign (OpenStreetMap Wiki 2020e).

5.1.5 Heidelberg, Germany

Heidelberg serves as the reference case study, allowing us to compare humanitarian mapping activities with those from a part of the map that has been established as high quality (researched by Arsanjani et al. (2013) and previously applied as a reference case study by Anderson, Soden, et al. (2018)). While I generalize and refer to all case studies as "mapping campaigns", I acknowledge that this Heidelberg case does not refer to a distinct campaign, as with the other humanitarian case studies. Figure 5.1 shows how the features mapped in this case study are distributed around the centre of this study area.

5.2 Characteristics of OSM data production

In this section, I provide details of the data that was produced in each of the case studies.

Table 5.1: Summary statistics for data produced in each case study.

Case Study	Duration *Days*	Area km^2	Unique Contributors	Features Created	Burst	Style
Port au Prince	658	1397	313	31962	17	event
Tacloban	82	356	199	19801	1	event
Kathmandu	250	98	881	37587	5	event
Bangui	1012	203	170	36788	164	mission
Heidelberg	365	192	108	3786	146	mission

Table 5.1 provides a basic summary of the data produced in each case study. Each case study covers varying temporal and spatial extents. The mapping campaign in Bangui, for example, is over ten times longer than that in Tacloban. The mapping campaign in Port au Prince covers an area that is nearly 15 times larger

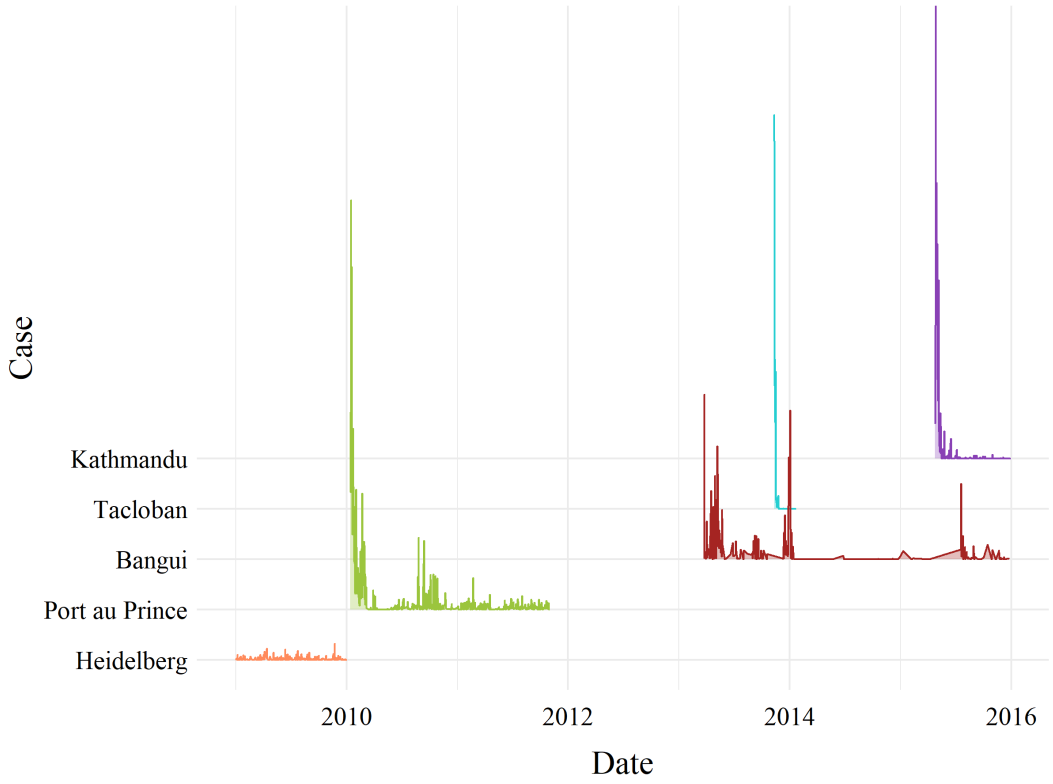

Figure 5.2: Daily volume of OSM contributions throughout duration of each mapping campaign. The implied z-axis indicates the relative daily contribution volume.

than that in Kathmandu. Despite these differences, all humanitarian mapping campaigns have produced volumes of data that are of the same order of magnitude (approximately 20,000 to 40,000 new features created). The reference case study in Heidelberg has produced notably less data. The case study in Kathmandu stands out when considering the number of unique contributors (nearly three-times the case study with the next largest number). Both Bangui and Heidelberg are classified as "missions", while Port au Prince, Tacloban, and Kathmandu are "events".

The dynamics of contributing patterns for each case study are further demonstrated in Figure 5.2, which shows the daily magnitude of contributions for each mapping campaign. The campaigns in Tacloban, Port au Prince, and Kathmandu follow a pattern where mapping activity is front-loaded and decays quickly after the beginning of the campaign. The magnitude of early mapping efforts in Tacloban, Kathmandu, and Port au Prince, shown by the burstiness values from Table 5.1 and the size of the peaks in Figure 5.2, set these cases apart from others. The mapping in

Bangui and the Heidelberg reference show a pattern where activity is more evenly sustained over a longer period of time, although Bangui still shows some notable peaks.

Further differences between the two mapping campaign styles are shown in Figure 5.3, where one sees that the 'event-style' campaigns have a stronger relationship between the number of daily contributors and contributions. In all cases, however, the relationship between daily contributor volume and daily contribution volume is statistically significant. Heidelberg and Bangui, the two 'mission-style' campaigns, have a notably shorter range of daily unique contributors (no more than 10 in a day), while campaigns such as that in Kathmandu have reached over 150 unique contributors in a day.

Both Figures 5.2 and 5.3 show a clear difference in the daily volume of data contributed between the Heidelberg reference and the humanitarian campaigns. Daily contribution volume in Heidelberg during this time barely exceed 100 new features, while humanitarian campaigns such as those in Tacloban and Kathmandu reach over 4,000 and 6,000 new features in a day, respectively, at their peaks.

Table 5.2, below, provides insight into the type of data that is contributed in each mapping campaign. Similarities are visible across all cases as the `building`, `highway`, and `source` tags are frequently occurring in each. The humanitarian campaigns show crisis-specific tags, such as `typhoon:damage` and `damage:event`.

Table 5.3 provides insight into the data sources for each mapping campaign. A notable proportion of the data from the humanitarian cases was generated from satellite imagery (eg. Bing or Worldview). Very little of the Heidelberg data has been tagged with a source, however none of the sources listed are from satellite imagery.

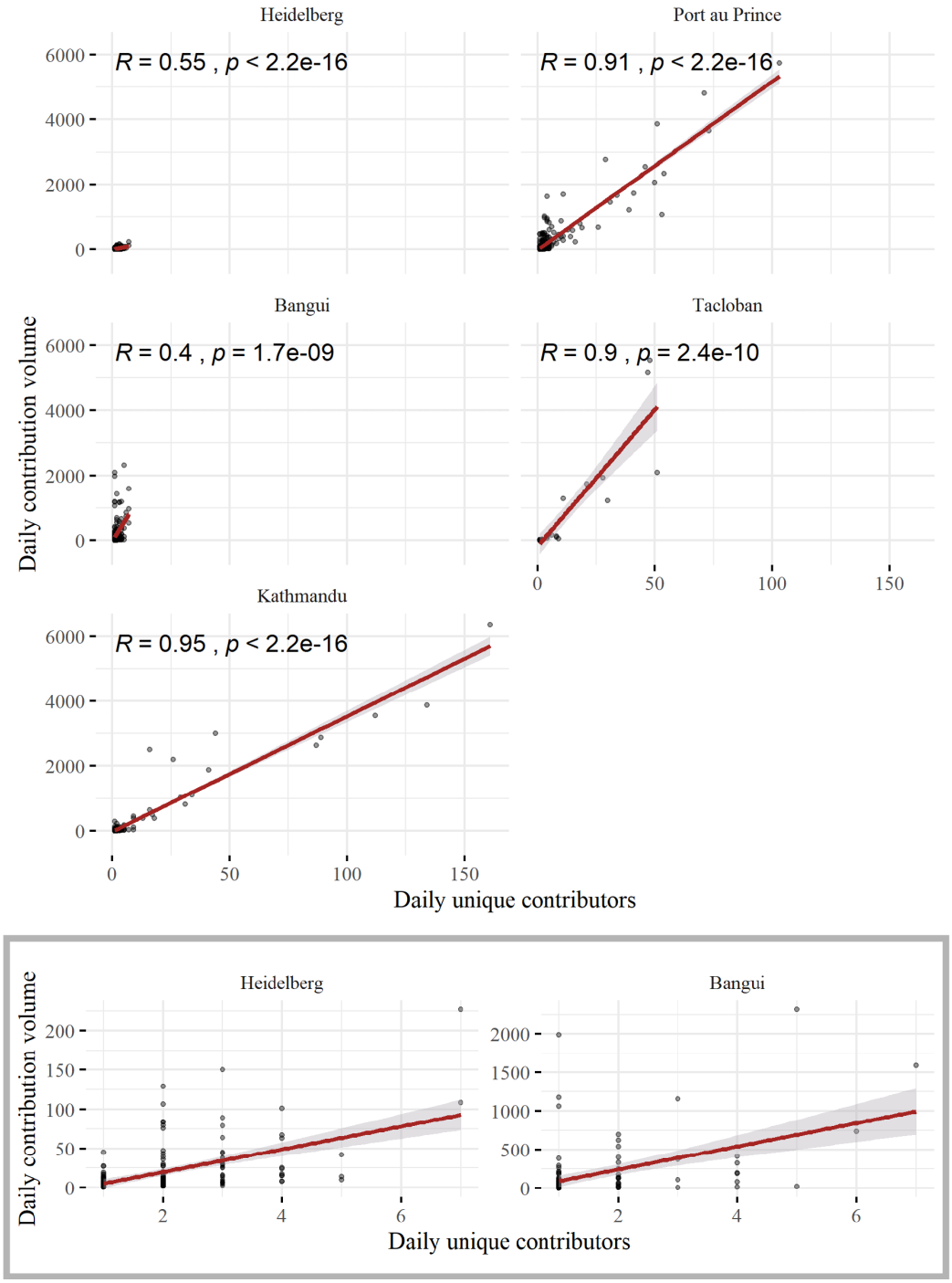

Figure 5.3: Scatterplots of daily unique contributors against daily contribution volume (in number of contributions) for each mapping campaign. Includes inset of Heidelberg and Bangui with rescaled axes.

Table 5.2: Top five most frequently occurring tag keys across each case study. Keys highlighted in grey appear across at least 4/5 case studies.

Port au Prince		Tacloban	
key	*percent*	*key*	*percent*
source	0.76	building	0.89
highway	0.33	source	0.27
building	0.29	typhoon:damage	0.11
attribute_source_date	0.17	highway	0.27
name	0.11	typhoon:damaged	0.15
Bangui		**Kathmandu**	
key	*percent*	*key*	*percent*
building	0.58	building	0.74
source	0.35	source	0.24
highway	0.11	idp:camp_site	0.14
source:date	0.06	damage:event	0.13
project:eurosha_2012	0.06	highway	0.04
Heidelberg			
key	*percent*		
highway	0.46		
name	0.23		
tracktype	0.14		
created_by	0.13		
amenity	0.10		

Table 5.3: Top five most frequently occurring source tag values for data from each case study.

Port au Prince		Tacloban	
source	*percent*	*source*	*percent*
geoeye	0.230	bing	0.131
google; 2010-01-21	0.150	Worldview-2; digitalglobe; nextview; 2013/11/09	0.100
yahoo	0.043	bing; 2010-11	0.018
NA	0.042	gsi/kiban 2500; naro	0.004
google 2010-01-17	0.033	hot task 355 image (arcgis)	0.021
Bangui		**Kathmandu**	
source	*percent*	*source*	*percent*
bing	0.186	pleiades 2015-04-27; cnes;airbus ds	0.135
bing 2012	0.132	nextview	0.079
worldview1	0.015	bing	0.014
bing et hotosm	0.012	gsimaps/std	0.005
NA	0.003	bing imagery	0.001
Heidelberg			
source	*percent*		
survey	0.017		
http://wiki.openstreetmap .org/wiki/import/catalogue /kreisgrenzen_deutschland _2005	0.002		
gps	0.001		
estimation	0.001		
rectified_map:837	0.001		

5.3 Prevalence of data maintenance

Following a review of the characteristics of data production across case studies, I present results that provide insight into the lineage of this data over time. Specifically, these results indicate the prevalence of data maintenance efforts in the years following each mapping campaign.

Figure 5.4, below, provides an initial look at the prevalence of data mainte-

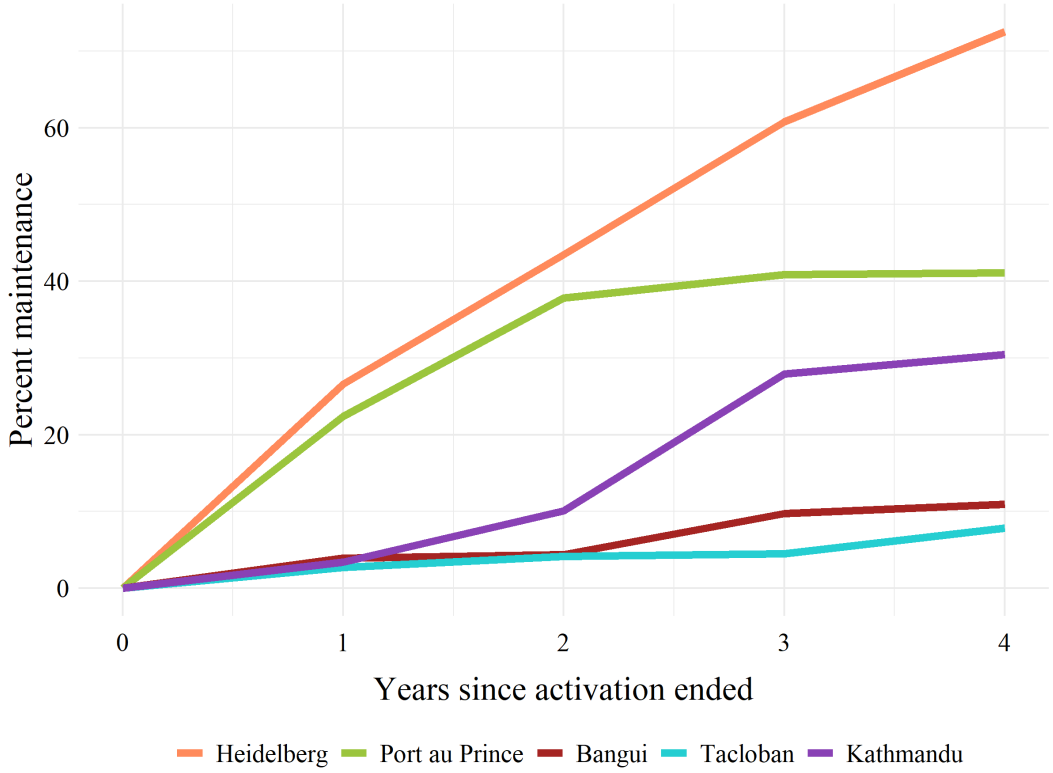

Figure 5.4: Cumulative percentage of maintained data (deleted or modified at least once) in years following end of mapping campaign

nance in each case study. This figure shows the cumulative percentage of features, created during the mapping campaign period, that have been modified at least once in the years after the campaign has ended. This and all subsequent figures include entity deletion within the definition of maintenance.

The reference case study, Heidelberg, has a significantly higher prevalence of data maintenance, as over 70% of this data has been modified or deleted in the four years after it was originally created. Among the humanitarian case studies, the Port au Prince and Kathmandu campaigns stand out for both reaching above 30% maintenance by the end of this four-year period. The other humanitarian case studies in Tacloban and Bangui only reach approximately 10% maintenance during this time. Figure 5.4 also shows how the rate of maintenance in Heidelberg appears to be approximately consistent during this time period, while maintenance efforts in the humanitarian case studies have plateaued in the fourth year after the campaign has ended. Maintenance efforts in Kathmandu increased notably in the third year

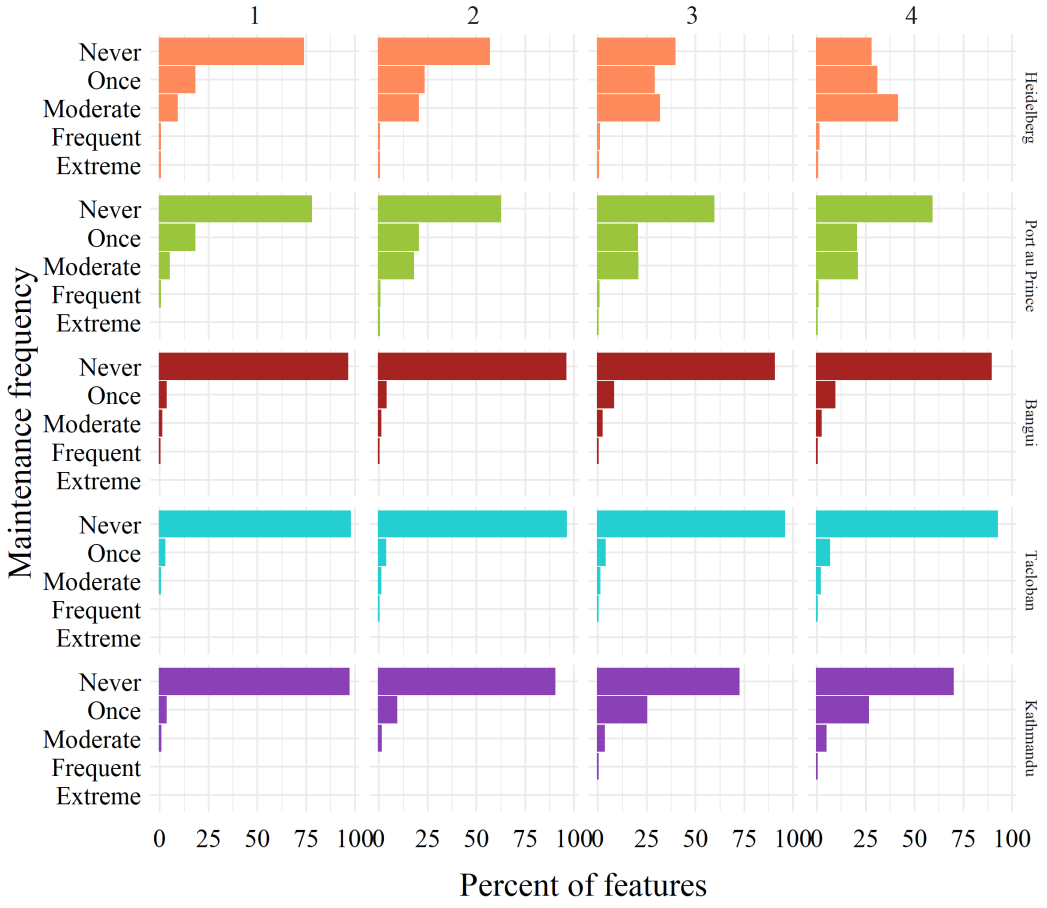

Figure 5.5: Distribution of maintenance frequency for features created during each case study, from 1-4 years (cumulatively) since the end of each campaign.

after the campaign ended.

Figure 5.5 presents data maintenance efforts by looking at the distribution in the maintenance frequency over time of all features that were created during each campaign. The x-axis of this figure is broken down by cumulative number of years since the end of the mapping campaign, and the y-axis is broken down by case study. In the first year after each mapping campaign, the majority of features in all case studies have never been maintained. Changes in the distribution of maintenance frequency are the most noticeable in the Heidelberg reference case, where the passing of time leads to a distribution that is less positively skewed, and towards higher frequency of maintenance in features. After four years, many features have been maintained more than once, with the majority of features classified within the *moderate* category (3-10 versions). Conversely, across all humanitarian cases,

the majority of features have still not been maintained after four years has passed. However, as is also reflected in Figure 5.4, the data from the Port au Prince and Kathmandu campaigns has been more frequently maintained than the other humanitarian cases.

Figures 5.6 and 5.7 illustrate the maintenance frequency distribution for features from each case study after four years have passed since the end of the mapping campaign. Figure 5.6 is disaggregated to show differences between node and way data types, and Figure 5.7 is disaggregated to show differences between features that are tagged with the `building`, `highway`, and all other tags. Figure 5.6 shows that nodes and ways follow roughly the same distribution of maintenance frequency across case studies. Similar results are shown in Figure 5.7. Interestingly, however, it appears as though nodes have been relatively well-maintained in the Port au Prince case study. In Kathmandu, the majority of the *other* features have been maintained at least once. Across all cases, buildings are notable as the most poorly maintained feature.

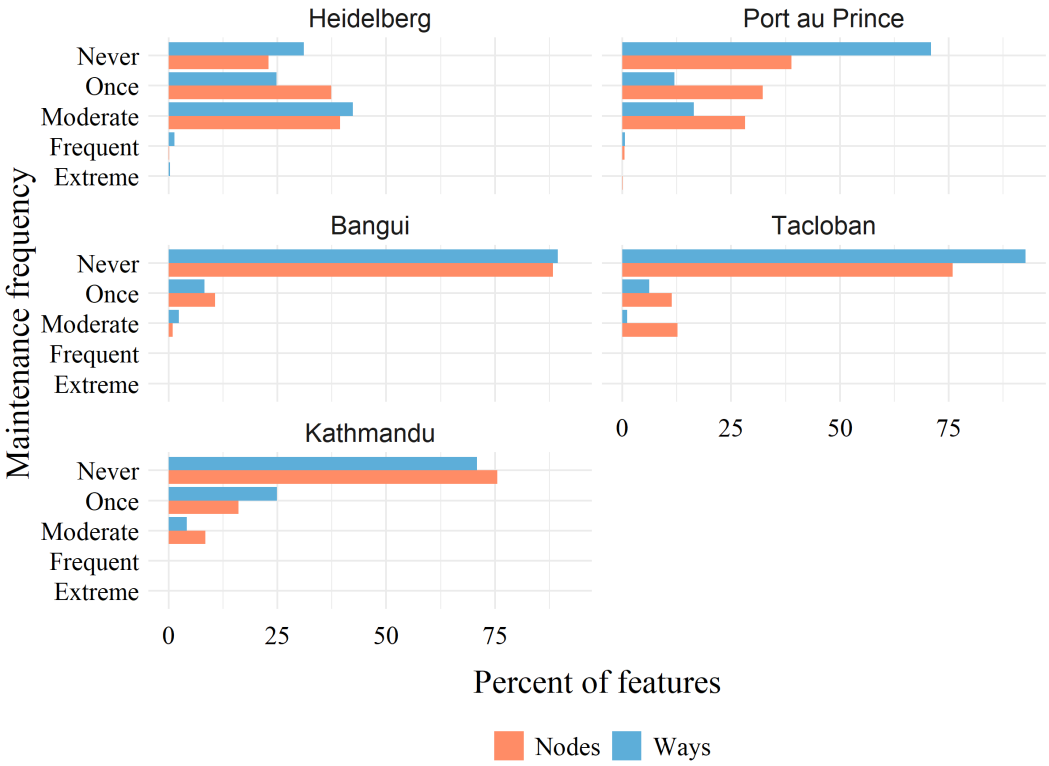

Figure 5.6: Maintenance frequency of nodes and ways for each case study, four years following the end of each mapping campaign.

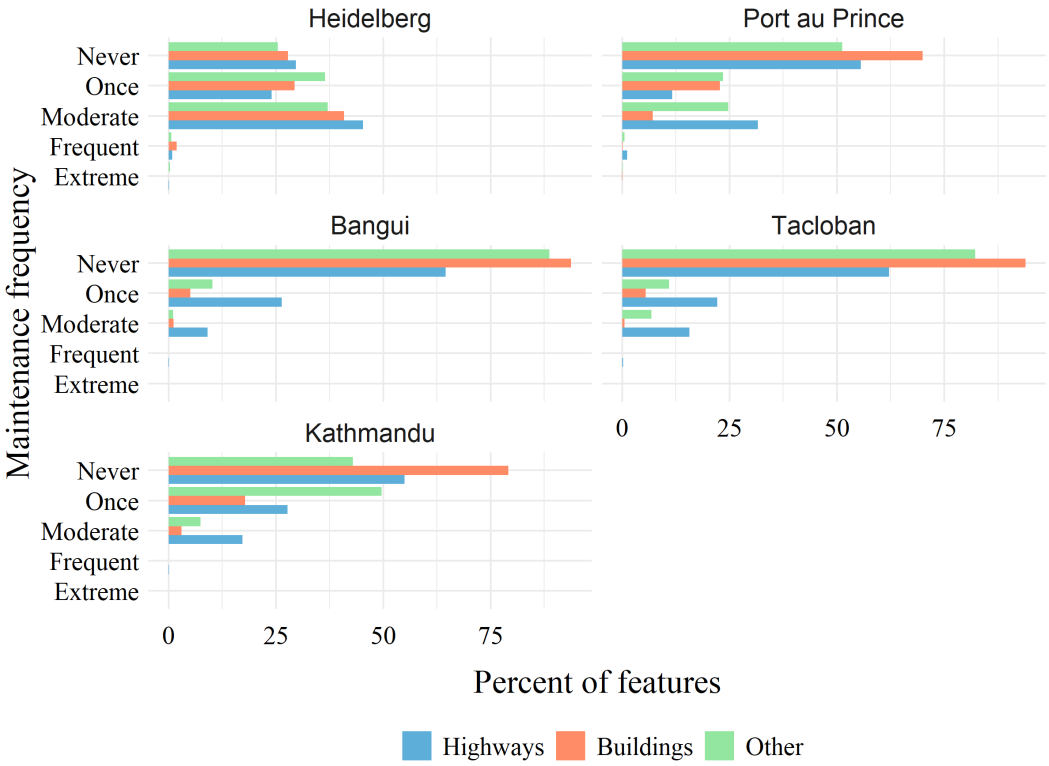

Figure 5.7: Maintenance frequency of features with `building`, `highway`, and all other tags for each case study; four years following the end of each mapping campaign.

Chapter 6

Discussion

In this chapter, I interpret the results of the analysis described in the previous chapter and consider their implications with respect to the existing body of literature discussed in Chapter 2. I then reflect on the limitations of this research approach.

6.1 Addressing research questions

RQ 1: What are the characteristics of data production in the selected humanitarian mapping campaigns and how does this compare with the reference case study?

These results show distinct differences between the humanitarian mapping campaigns and the Heidelberg reference. It is immediately apparent that the daily volume of data created reaches much higher levels during the humanitarian campaigns than in Heidelberg. For example, all humanitarian cases have at least one day where over 2000 features were created, while Heidelberg's maximum is only little over 100 contributions in a single day. Mapping efforts in Kathmandu, Port au Prince, and Tacloban also attracted large volumes of contributors on some days (reaching over 150 unique contributors in a single day in Kathmandu, for example), while efforts in Bangui and Heidelberg had less than 10 unique contributors each day.

The event/mission classification scheme proposed by Dittus, Quattrone, and Capra (2017) offers a useful framework for considering data production in humanitarian mapping campaigns. The early peaks for the event-style campaigns (Port

au Prince, Tacloban, and Kathmandu) shown in Figure 5.2 clearly demonstrate the dramatic burstiness that occurs when mapping is done urgently in response to an immediate natural disaster. The urgency of these events leads to a rapid and significant decay in contribution volume as time passes. As shown in Figure 5.3, this decay in contribution volume is correlated with a decay in contributor numbers (as is intuitive). This pattern in mapping activity from event style campaigns echoes the findings from Dittus, Quattrone, and Capra (2017, p. 1294).

While both Heidelberg and Bangui are classified as mission-style campaigns, Figures 5.2 and 5.3 show notable differences in the dynamics of how data is produced over time. Figure 5.2 shows that the volume of features created over time in Heidelberg remains relatively stable, while mapping in Bangui has significant peaks throughout the duration of the campaign. It is likely that some of these peaks are the result of data imports. This hypothesis is supported by Figure 5.3, where there is a relatively weak relationship between the number of daily contributors and the daily volume of contributed data, as some days have over 1000 features produced by fewer than four contributors.

A review of the commonly-occurring tags keys from Table 5.2 shows how features such as buildings and roads are common across all case studies. This is not an unexpected finding, as these tags are among the most frequently used in all of OSM. The TagInfo website indicates that 58% of all ways are tagged with `building` and 24% of all ways are tagged with `highway` (OpenStreetMap Contributors 2020). The presence of disaster-related tags in the humanitarian cases, such as `typhoon:damage, idp:camp_site`, and `damage:event`, is also notable. These disaster tags correspond to temporary attributes, suggesting that they will likely need to be removed or updated in the future.

The review of the commonly-used `source` tags in Table 5.3 shows that many of the features from the humanitarian cases were produced from satellite sources, indicating that many contributors were likely remote. The prevalence of remote contributors during humanitarian mapping campaigns is well-understood within the literature (Dittus, Quattrone, and Capra 2017; Eckle and Albuquerque 2015). While

very few of the features from Heidelberg were tagged with a `source`, there are not any frequently occurring satellite sources, indicating that this data was more likely to be produced by local mappers.

RQ 2: To what extent is the data produced during the selected campaigns maintained over time and how does this compare with the reference case study?

Figures 5.4, 5.5, 5.6, and 5.7 all show how the data from the Heidelberg reference has been maintained to a significantly greater extent than the data from the humanitarian case studies. Heidelberg is the only case study where over 50% of the features have been updated or deleted at least once since being created. Many of the features in Heidelberg have also been maintained with notable frequency (over 30% of features have been updated 3-10 times), while the vast majority of *maintained* features from humanitarian cases have only been updated or deleted once since creation. This finding is in line with expectations, as Heidelberg was selected as a reference due to its comparatively complete and accurate data (Arsanjani et al. 2013). While I am careful not to generalize these findings beyond the selected case studies, this result suggests that OSM data produced from humanitarian mapping efforts may be less maintained than other regions of the database. Greater effort may thus be needed in ensuring that the data produced in response to humanitarian need is maintained in OSM in the years following the campaign.

These findings may have implications for our understanding of the temporal accuracy of OSM data in areas with humanitarian mapping campaigns. Chapter 2 addressed how data maintenance can be considered as the process by which OSM data is kept up-to-date. Thus, it can be assumed that the case studies found to have poorly maintained data, such as in Tacloban and Bangui, may be more likely to have data that is out of date. With reference to the values in Figure 5.4 and Table 5.1, one could estimate that over 17,000 features that are currently on the map in Tacloban may be out of date.

This approach to understanding temporal accuracy advances existing work by considering not only the attributes of a given feature within OSM at a given point

in time (as in Barron, Neis, and Zipf (2014), who look at the date of last edit), but also the entire lineage of that feature over its history. This approach acknowledges the ongoing evolution of data within OSM, leading to a deeper understanding of its temporal dimensions.

However, it is difficult to identify the amount of data maintenance that is necessary in a given area to keep OSM data up to date. Theoretically, it can be understood that data only needs to be maintained if the associated geographic phenomena have changed in some way. It is assumed to be incredibly unlikely for all geographic phenomena in an area to remain the same over a long period of time, so given the passing of time, some data maintenance will always be necessary. However, given that this research has no "ground-truth" for how much change has occurred in a given area, it is challenging to know whether or not an apparent lack of data maintenance is a problem. However, the Heidelberg reference case offers an indication of data maintenance levels that might be appropriate. As the OSM data from Heidelberg is generally considered to be of high quality, one might assume that the levels of maintenance seen here are what other regions on the map should strive to achieve. Nevertheless, it is acknowledged that data maintenance needs may vary significantly across different locations (Quattrone, Dittus, and Capra 2017). The prevalence of temporally-sensitive tags such as, `typhoon:damage` and `damage:event`, in the data from Tacloban may suggest that OSM data produced during humanitarian mapping campaigns may in fact be in need of more maintenance than other parts of the map.

Ultimately, this work's findings suggest a potential shortcoming with current mechanisms of data production during humanitarian mapping campaigns. As is shown by the case studies, these campaigns may produce an incredibly large volume of data over short periods of time (eg. nearly 40,000 nodes and ways produced in Kathmandu alone in less than one year). While there is no denying that this data is immediately useful and likely up-to-date in the wake of a disaster, it also presents a challenge in that there is now more data that can potentially be out of date in the future if it is not well maintained. Ideally the data produced during a humanitarian

campaign is of lasting value to the local community. This work's results highlight a potential need for more formal mechanisms to ensure data maintenance takes place following humanitarian mapping campaigns.

RQ 3: What insight do the results offer into potential relationships between characteristics of data production and levels of data maintenance in each of the case studies?

Looking across the results previously discussed in this section, I generate informed hypotheses about potential relationships between characteristics of data production and levels of data maintenance. Given the limited volume of case studies under consideration, I acknowledge that these hypotheses may not generalize to other contexts. Rather, my intent is to synthesize these empirical findings with existing theory to offer well-considered suggestions for future research into factors that are associated with greater data maintenance.

Figures 5.2 and 5.3 demonstrate how the Heidelberg and Bangui case studies, the two mission-style campaigns, have notably different dynamics of data production and volumes of contributor engagement than the other case studies. One might then expect to see these two campaigns exhibit characteristics of data maintenance that similarly set them apart from the other event-style case studies. However, while Heidelberg is clearly distinct in its high levels of maintenance, Bangui shows characteristics of maintenance that are quite similar to Tacloban (as demonstrated in Figures 5.4 and 5.5). This finding may suggest that a mapping campaign's style (event or mission) may not be an indication of the extent to which the data will be maintained over time. However, future research here is needed.

The higher levels of data maintenance seen in Heidelberg can perhaps be explained by the low but sustained levels of editing activity conducted by likely local contributors, as shown in Figure 5.2 and Table 5.3. The overall volume of data produced in Heidelberg is also an order of magnitude less than in all humanitarian case studies. Practically speaking, it is likely easier to keep less data up-to-date. One can also assume that local contributors are more invested in the sustained accu-

racy of OSM data for their community, and so will work to keep it up-to-date. One might assume that the low but sustained activity shown in Figure 5.2 indicates the presence of an active local OSM community, which is then a potential indicator of future maintenance efforts taking place.

This hypothesis is also supported by the past work of Quattrone, Dittus, and Capra (2017), who find that OSM data is better maintained in areas with more experienced and active contributors. Areas subject to humanitarian mapping, such as the case studies here, have been found to have significant percentages of newcomer mappers (for example, over 80% in the Nepal earthquake) (Dittus, Quattrone, and Capra 2017).

Within the humanitarian case studies, Port au Prince and Kathmandu stand out for their higher levels of maintenance. Both the Kathmandu and Port au Prince cases have been documented as having significant local engagement (Soden and Palen 2014), which may in part explain why this data has been maintained more over time. Compared against mapping in Tacloban, Kathmandu and Port au Prince took place over longer periods of time and engaged more unique contributors. Tacloban had the lowest levels of data maintenance and was the case study with the shortest 'burst' value. This data was produced incredibly quickly, which may indicate that less effort was made in engaging with local mappers who would take care of it over time.

Our results in Figures 5.6 and 5.7 also show potential differences in the extent to which different feature types are maintained. These findings may suggest that some features are at a greater risk of being out of date than others. This hypothesis (also explored in Quattrone, Dittus, and Capra (2017) follows intuition, as one can acknowledge that features such as shops and other POIs may change frequently, while features such as parks and roads may be more consistent over time. However our analysis does not reveal any conclusive results and so more work here is needed.

6.2 Project limitations

While the results of this analysis offer insight into data production and maintenance during humanitarian mapping campaigns, it is acknowledged that this work is subject to a number of limitations.

Firstly, using OSM entity version numbers as indications of maintenance may not be wholly accurate. As indicated by the work of Mooney and Corcoran (2012), a higher number of versions for a given feature may not be entirely aligned with this work's definition of maintenance. The authors identify, for example, how disagreements between contributors may lead to a back and forth of revisions to a given feature in quick succession Mooney and Corcoran (2012). While such a conflict undoubtedly indicates that care is given to the quality of this feature, it might not be accurate to say that a feature with 20 versions as result of such an "edit war" is more maintained than a feature with only 3 versions. This shortcoming could be addressed in future work by more closely investigating the timing between revisions of a given OSM entity.

Secondly, this work does not distinguish between the different forms of data maintenance that can occur. As is described by Quattrone, Dittus, and Capra (2017), maintenance is constituted of activities including entity deletion, tag removal, and tag addition. A more detailed typology of feature changes in geographic datasets is also provided by Rehrl, Brunauer, and Gröchenig (2015). While it is out of the scope of this work to disaggregate the analysis by these different activities, such an effort may be a meaningful basis for future work.

Thirdly, the generalizability of these research findings is limited by this work's tightly-scoped case study approach. Given the lack of established methodological framework in this domain, a small number of cases were selected to allow for a more exploratory and descriptive investigation. This work's conclusions are thus only valid within the context of the four humanitarian case studies and one reference case study. It is hoped that this methodology can be repeated across a wider range of humanitarian mapping campaigns, leading to a stronger understanding data maintenance across the entire domain of humanitarian mapping.

Chapter 7

Conclusion

This research employs a comparative case study approach to investigate data production and maintenance in humanitarian mapping campaigns. This work responds to the need to more rigorously consider dimensions of temporal data quality in OSM, particularly within humanitarian mapping contexts. I focused specifically on humanitarian campaigns in Port au Prince, Bangui, Tacloban, and Kathmandu; and compare against mapping in Heidelberg as a reference. The recently developed OSHDB API (Raifer et al. 2019) was applied to efficiently process and filter large volumes of historical OSM data. In addition to the key results summarized below, this research builds off of Quattrone, Dittus, and Capra (2017) and offers a methodological approach for empirically assessing data maintenance in OSM.

Following the framework set out by Dittus, Quattrone, and Capra (2017), Bangui and Heidelberg were classified as mission-style campaigns; and Tacloban, Kathmandu, and Port au Prince as event-style campaigns. The humanitarian case studies differed from the Heidelberg reference in both the overall and daily volume of new contributions to OSM, with the humanitarian campaigns showing significantly more data added over shorter time frames.

The results of this work also show that the data produced during the selected humanitarian campaigns has been poorly maintained over time when compared against the maintenance levels seen in Heidelberg. Across all humanitarian case studies, the majority of data produced during the mapping campaign has not been modified or deleted after four years. This finding suggests that the OSM data in

these areas is at a greater risk of becoming out of date. Thus, the humanitarian mapping community may need to consider developing formal mechanisms or incentives for ensuring that the data produced during mapping campaigns is maintained over time, allowing for it to be a lasting resource for the affected communities.

It is hoped that this work will serve as a foundation for future investigations into data maintenance in humanitarian mapping. Future work should replicate this analysis on an expanded set of humanitarian case studies to test if these results generalize to other cases. Additionally, a disaggregated and larger-scale analysis into specific factors relating to data maintenance is recommended. Such efforts could help to identity different types of OSM data that are at a greater risk of becoming out-of-date. Future work should also more comprehensively investigate appropriate baseline levels of maintenance activity, for example by following the approach detailed here to quantify maintenance in areas of known high data quality.

Bibliography

Ahmouda, Ahmed, Hartwig H. Hochmair, and Sreten Cvetojevic (2018). Analyzing the effect of earthquakes on OpenStreetMap contribution patterns and tweeting activities. *Geo-spatial Information Science* **21**.3, pp. 195–212. DOI: 10.1080/10095020.2018.1498666.

Albrecht, Conrad M. et al. (2020). Change Detection from Remote Sensing to Guide OpenStreetMap Labeling. *ISPRS International Journal of Geo-Information* **9**.7, p. 427. DOI: 10.3390/ijgi9070427.

Anderson, Jennings, Dipto Sarkar, and Leysia Palen (2019). Corporate Editors in the Evolving Landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information* **8**.5, p. 232. DOI: 10.3390/ijgi8050232.

Anderson, Jennings, Robert Soden, et al. (2018). The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters. *International Journal of Human–Computer Interaction* **34**.4, pp. 295–310. DOI: 10.1080/10447318.2018.1427828.

Antoniou, V. and A. Skopeliti (2015). Measures and Indicators of VGI Quality: An Overview. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* **II-3/W5**, pp. 345–351. DOI: 10.5194/isprsannals-II-3-W5-345-2015.

Arsanjani, Jamal Jokar et al. (2013). Assessing the Quality of OpenStreetMap Contributors together with their Contributions. *16th AGILE International Conference on Geographic Information Science*. Leuven, p. 4.

Barron, Christopher, Pascal Neis, and Alexander Zipf (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS* **18**.6, pp. 877–895. DOI: 10.1111/tgis.12073.

Cavallo, Eduardo, Andrew Powell, and Oscar Becerra (2010). Estimating the Direct Economic Damages of the Earthquake in Haiti*. *The Economic Journal* **120**.546, F298–F312. DOI: 10.1111/j.1468-0297.2010.02378.x.

Chen, Rui et al. (2008). Coordination in emergency response management. *Communications of the ACM* **51**.5, pp. 66–73. DOI: 10.1145/1342327.1342340.

Chuang, Tyng-Ruey et al. (2013). The one and many maps: participatory and temporal diversities in OpenStreetMap. *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. GEOCROWD '13. Orlando, Florida: Association for Computing Machinery, pp. 79–86. DOI: 10.1145/2534732.2534737.

Corcoran, Padraig, Peter Mooney, and Michela Bertolotto (2013). Analysing the growth of OpenStreetMap networks. *Spatial Statistics* **3**, pp. 21–32. DOI: 10.1016/j.spasta.2013.01.002.

Cowan, Nuala M (2011). A Geospatial Data Management Framework for Humanitarian Response. *Proceedings of the 8th International ISCRAM Conference*, p. 5.

Dean, Jeffrey and Sanjay Ghemawat (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM* **51**.1, pp. 107–113. DOI: 10.1145/1327452.1327492.

DesRoches, Reginald et al. (2011). Overview of the 2010 Haiti Earthquake. *Earthquake Spectra* **27**.1_suppl1, pp. 1–21. DOI: 10.1193/1.3630129.

Dittus, Martin, Giovanni Quattrone, and Licia Capra (2017). Mass Participation During Emergency Response: Event-centric Crowdsourcing in Humanitarian Mapping. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: Association for Computing Machinery, pp. 1290–1303. DOI: 10.1145/2998181.2998216.

Eckle, Melanie and João Porto de Albuquerque (2015). Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. *Proceedings of the ISCRAM 2015 Conferen*, p. 9.

Estes, J. E. and D. W. Mooneyhan (1994). Of maps and myths. *Photogrammetric Engineering and Remote Sensing* **60**.5. URL: https://www.osti.gov/biblio/53462 (visited on 06/11/2020).

Exel, M van, E Dias, and S Fruijtier (2010). The impact of crowdsourcing on spatial data quality indicators. *Van Exel, M., E. Dias, and S. Fruijtier. "The impact of crowdsourcing on spatial data quality indicators." Proceedings of the GI-Science 2010 Doctoral Colloquium, Zurich, Switzerland.*, p. 5.

Flanagin, Andrew J. and Miriam J. Metzger (2008). The credibility of volunteered geographic information. *GeoJournal* **72**.3, pp. 137–148. DOI: 10.1007/s10708-008-9188-y.

Fox, Christopher, Anany Levitin, and Thomas Redman (1994). The notion of data and its quality dimensions. *Information Processing & Management* **30**.1, pp. 9–19. DOI: 10.1016/0306-4573(94)90020-5.

Fox, Killian (2012). OpenStreetMap: 'It's the Wikipedia of maps'. *The Guardian*. URL: https://www.theguardian.com/theobserver/2012/feb/18/openstreetmap-world-map-radicals (visited on 08/21/2020).

GeoFabrik (2020). OpenStreetMap Data Extracts. URL: https://download.geofabrik.de/.

Girres, Jean-François and Guillaume Touya (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* **14**.4, pp. 435–459. DOI: 10.1111/j.1467-9671.2010.01203.x.

Global Conflict Tracker (2020). Violence in the Central African Republic. URL: https://cfr.org/global-conflict-tracker/conflict/violence-central-african-republic (visited on 08/03/2020).

Goodchild, Michael (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal* **69**.4, pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

— (2009). NeoGeography and the nature of geographic expertise. *Journal of Location Based Services* **3**.2, pp. 82–96. DOI: 10.1080/17489720902950374.

Goodchild, Michael and Linna Li (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics* **1**, pp. 110–120. DOI: 10.1016/j.spasta.2012.03.002.

Graser, Anita, Markus Straub, and Melitta Dragaschnig (2015). Is OSM Good Enough for Vehicle Routing? A Study Comparing Street Networks in Vienna. *Progress in Location-Based Services 2014*. Ed. by Georg Gartner and Haosheng Huang. Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing, pp. 3–17. DOI: 10.1007/978-3-319-11879-6_1.

Gröchenig, Simon, Richard Brunauer, and Karl Rehrl (2014a). Digging into the history of VGI data-sets: results from a worldwide study on OpenStreetMap mapping activity. *Journal of Location Based Services* **8**.3, pp. 198–210. DOI: 10.1080/17489725.2014.978403.

— (2014b). Estimating Completeness of VGI Datasets by Analyzing Community Activity Over Time Periods. *Connecting a Digital Europe Through Location and Place*. Ed. by Joaquín Huerta, Sven Schade, and Carlos Granell. Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing, pp. 3–18. DOI: 10.1007/978-3-319-03611-3_1.

Haklay, Muki (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design* **37**.4, pp. 682–703. DOI: 10.1068/b35097.

Haklay, Muki, Sofia Basiouka, et al. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic

Information. *The Cartographic Journal* **47**.4, pp. 315–322. DOI: 10.1179/000870410X12911304958827.

Haklay, Muki, Alex Singleton, and Chris Parker (2008). Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass* **2**.6, pp. 2011–2039. DOI: 10.1111/j.1749-8198.2008.00167.x.

Heidelberg Institute for Geoinformation Technology (2020a). OSHDB - OpenStreetMap History Data Analysis. URL: https://github.com/GIScience/oshdb (visited on 08/19/2020).

— (2020b). Setup a Local OSHDB. URL: https://github.com/GIScience/oshdb/tree/master/oshdb-tool/etl (visited on 08/15/2020).

Helbich, Marco et al. (2012). Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. *Proceedings of GI_Forum 2012*. Berlin, p. 10.

Humanitarian OpenStreetMap Team (2020a). HOT Tasking Manager. URL: https://tasks.hotosm.org/ (visited on 08/13/2020).

— (2020b). Projects. URL: https://www.hotosm.org/projects/.

Kaarbo, Juliet and Ryan K. Beasley (1999). A Practical Guide to the Comparative Case Study Method in Political Psychology. *Political Psychology* **20**.2, pp. 369–391. DOI: 10.1111/0162-895X.00149.

Kimerling, A. Jon (2009). Dotting the Dot Map, Revisited. *Cartography and Geographic Information Science* **36**.2, pp. 165–182. DOI: 10.1559/152304009788188754.

Kittur, Aniket et al. (2007). He says, she says: conflict and coordination in Wikipedia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*. San Jose, California, USA: ACM Press, pp. 453–462. DOI: 10.1145/1240624.1240698.

Lu, Joanne (2019). After Hurricane Dorian, The 'Wikipedia Of Maps' Came To The Rescue. *NPR*. URL: https://www.npr.org/sections/goatsandsoda/2019/10/03/765783296/after-hurricane-

dorian-the-wikipedia-of-maps-came-to-the-rescue (visited on 08/21/2020).

Lum, Thomas and Rhoda Margesson (2014). TYPHOON HAIYAN (YOLANDA): U.S. AND INTERNATIONAL RESPONSE TO PHILIPPINES DISASTER. *Current Politics and Economics of South, Southeastern, and Central Asia* **23**.2, pp. 209–246. URL: https://search.proquest.com/docview/1622681509?pq-origsite=gscholar&fromopenview=true (visited on 08/03/2020).

Luxen, Dennis and Christian Vetter (2011). Real-time routing with OpenStreetMap data. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS '11. Chicago, Illinois: Association for Computing Machinery, pp. 513–516. DOI: 10.1145/2093973.2094062.

McConchie, Alan (2013). From Wiki Gardening to Map Gardening: Analyzing Contribution Patterns in OpenStreetMap. San Francisco, California, USA.

Meier, Patrick (2012). Crisis Mapping in Action: How Open Source Software and Global Volunteer Networks Are Changing the World, One Map at a Time. *Journal of Map & Geography Libraries* **8**.2, pp. 89–100. DOI: 10.1080/15420353.2012.663739.

Minghini, M., M. A. Brovelli, and F. Frassinelli (2018). An Open Source Approach for the Intrinsic Assessment of the Temporal Accuracy, Up-To-Dateness and Lineage of OpenStreetMap. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. XLII-4-W8. Copernicus GmbH, pp. 147–154. DOI: https://doi.org/10.5194/isprs-archives-XLII-4-W8-147-2018.

Mooney, Peter and Padraig Corcoran (2011). Accessing the history of objects in OpenStreetMap. *Proceedings AGILE 2011*. Utrecht, The Netherlands, p. 3.

— (2012). Characteristics of Heavily Edited Objects in OpenStreetMap. *Future Internet* **4**.1, pp. 285–305. DOI: 10.3390/fi4010285.

Neis, Pascal and Dennis Zielstra (2014). Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet* **6**.1, pp. 76–106. DOI: 10.3390/fi6010076.

Neis, Pascal and Alexander Zipf (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information* **1**.2, pp. 146–165. DOI: 10.3390/ijgi1020146.

Nepal Earthquake 2015: Post Disaster Needs Assessment (2015). Tech. rep. Government of Nepal.

O'Reilly, Tim (2009). What is Web 2.0. O'Reilly Media, Inc. URL: https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html.

Oort, Pepijn van (2006). Spatial data quality: from description to application. PhD Thesis. Wageningen University, NL.

OpenStreetMap Contributors (2020). OpenStreetMap Taginfo. URL: https://taginfo.openstreetmap.org (visited on 08/05/2020).

OpenStreetMap Statistics (2020). URL: https://www.openstreetmap.org/stats/data_stats.html (visited on 06/26/2020).

OpenStreetMap Wiki (2017). OSM XML. URL: https://wiki.openstreetmap.org/wiki/OSM_XML (visited on 08/04/2020).

— (2018a). Component overview. URL: https://wiki.openstreetmap.org/wiki/Component_overview (visited on 08/21/2020).

— (2018b). WikiProject Typhoon Haiyan. URL: https://wiki.openstreetmap.org/wiki/Typhoon_Haiyan (visited on 08/03/2020).

— (2019a). Node. URL: https://wiki.openstreetmap.org/wiki/Node (visited on 08/20/2020).

— (2019b). Relation. URL: https://wiki.openstreetmap.org/wiki/Relation (visited on 08/20/2020).

— (2020a). Elements. URL: https://wiki.openstreetmap.org/wiki/Elements (visited on 07/16/2020).

OpenStreetMap Wiki (2020b). Key:source. URL: https://wiki.openstreetmap.org/wiki/Key:source (visited on 08/06/2020).

— (2020c). Tags. URL: https://wiki.openstreetmap.org/wiki/Tags (visited on 08/04/2020).

— (2020d). Way. URL: https://wiki.openstreetmap.org/wiki/Way (visited on 08/20/2020).

— (2020e). WikiProject Central African Republic. URL: https://wiki.openstreetmap.org/wiki/WikiProject_Central_African_Republic (visited on 08/03/2020).

Palen, Leysia et al. (2015). Success & Scale in a Data-Producing Organization: The Socio-Technical Evolution of OpenStreetMap in Response to Humanitarian Events. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, pp. 4113–4122. DOI: 10.1145/2702123.2702294.

Poiani, Thiago Henrique et al. (2016). Potential of Collaborative Mapping for Disaster Relief: A Case Study of OpenStreetMap in the Nepal Earthquake 2015. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 188–197. DOI: 10.1109/HICSS.2016.31.

Poser, Kathrin and Doris Dransch (2010). Volunteered Geographic Information for Disaster Management with Application to Rapid Flood Damage Estimation. *GEOMATICA* **64**.1, pp. 89–98. DOI: 10.5623/geomat-2010-0008.

Quattrone, Giovanni, Martin Dittus, and Licia Capra (2017). Work Always in Progress: Analysing Maintenance Practices in Spatial Crowd-sourced Datasets. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: Association for Computing Machinery, pp. 1876–1889. DOI: 10.1145/2998181.2998267.

Raifer, Martin et al. (2019). OSHDB: a framework for spatio-temporal analysis of OpenStreetMap history data. *Open Geospatial Data, Software and Standards* **4**.1, p. 3. DOI: 10.1186/s40965-019-0061-3.

Rehrl, Karl, Richard Brunauer, and Simon Gröchenig (2015). Towards a Qualitative Assessment of Changes in Geographic Vector Datasets. *AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities*. Ed. by Fernando Bacao, Maribel Yasmina Santos, and Marco Painho. Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing, pp. 181–197. DOI: 10.1007/978-3-319-16787-9_11.

Roick, Oliver, Lukas Loos, and Alexander Zipf (2012). A technical framework for visualizing spatio-temporal quality metrics of volunteered geographic information. *Proceedings of the GEOINFORMATIK*.

Ruta, Michele et al. (2015). Indoor/Outdoor Mobile Navigation via Knowledge-Based POI Discovery in Augmented Reality. *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. Vol. 3, pp. 26–30. DOI: 10.1109/WI-IAT.2015.243.

Severinsen, Jeremy et al. (2019). VGTrust: measuring trust for volunteered geographic information. *International Journal of Geographical Information Science* **33**.8, pp. 1683–1701. DOI: 10.1080/13658816.2019.1572893.

Soden, Robert and Leysia Palen (2014). From Crowdsourced Mapping to Community Mapping: The Post-earthquake Work of OpenStreetMap Haiti. *COOP 2014 - Proceedings of the 11th International Conference on the Design of Cooperative Systems, 27-30 May 2014, Nice (France)*. Ed. by Chiara Rossitto et al. Cham: Springer International Publishing, pp. 311–326. DOI: 10.1007/978-3-319-06498-7_19.

— (2016). Infrastructure in the Wild: What Mapping in Post-Earthquake Nepal Reveals about Infrastructural Emergence. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, pp. 2796–2807. DOI: 10.1145/2858036.2858545.

Trame, Johannes and Carsten Keßler (2011). Exploring the Lineage of Volunteered Geographic Information with Heat Maps. *GeoViz*, p. 2.

Typhoon Haiyan/Yolanda Fact Sheet #20 (2014). Tech. rep. USAID. URL: https://www.usaid.gov/haiyan/fy14/fs20 (visited on 08/03/2020).

Vyncke, Jorieke (2015). Humanitarian Mapping and Local Communities. URL: https://www.openstreetmap.org/user/Jorieke%20V/diary (visited on 06/24/2020).

— (2020). Personal conversation with Jorieke Vyncke.

Yadav, Piyush et al. (2020). Human Assisted Artificial Intelligence Based Technique to Create Natural Features for OpenStreetMap. *arXiv:2007.02149 [cs]*. URL: http://arxiv.org/abs/2007.02149 (visited on 08/20/2020).

Zielstra, Dennis, Hartwig H. Hochmair, and Pascal Neis (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap – A United States Case Study. *Transactions in GIS* **17**.3, pp. 315–334. DOI: 10.1111/tgis.12037.

Zielstra, Dennis and Alexander Zipf (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. Guimarães, Portugal, p. 15.

Zook, Matthew et al. (2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy* **2**.2, pp. 7–33. DOI: 10.2202/1948-4682.1069.

Appendix A

OSM Component Overview

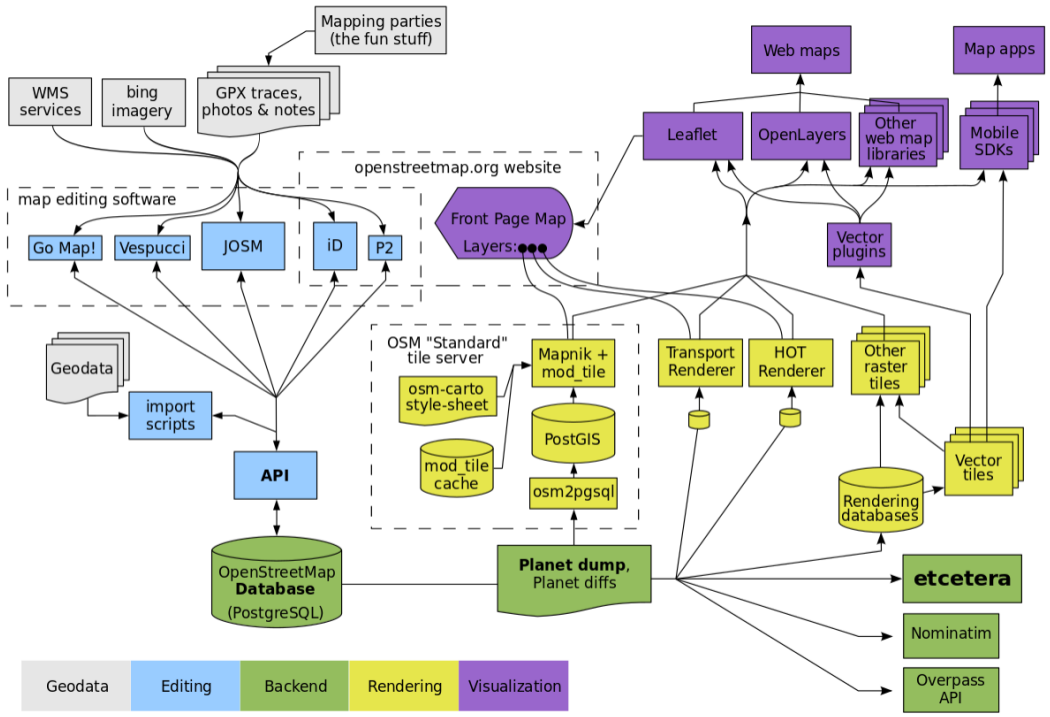

Figure A.1: Overview of the OSM's technical components (OpenStreetMap Wiki 2018a)

Appendix B

Enlarged Maps

This appendix includes enlarged versions of the inset maps from Figure 5.1.

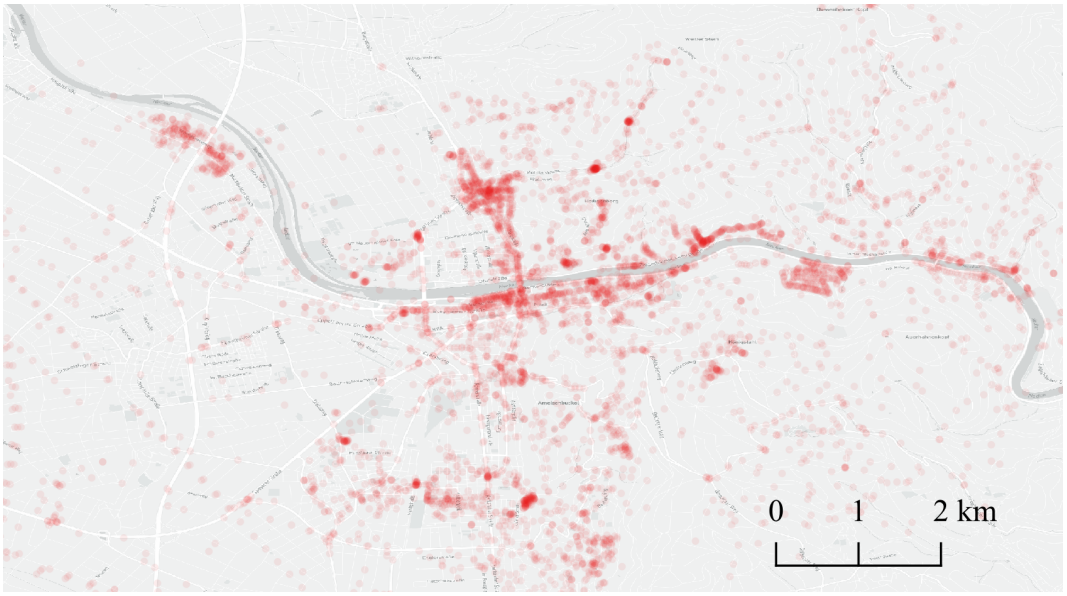

Figure B.1: Spatial distribution of features mapped in Heidelberg case study

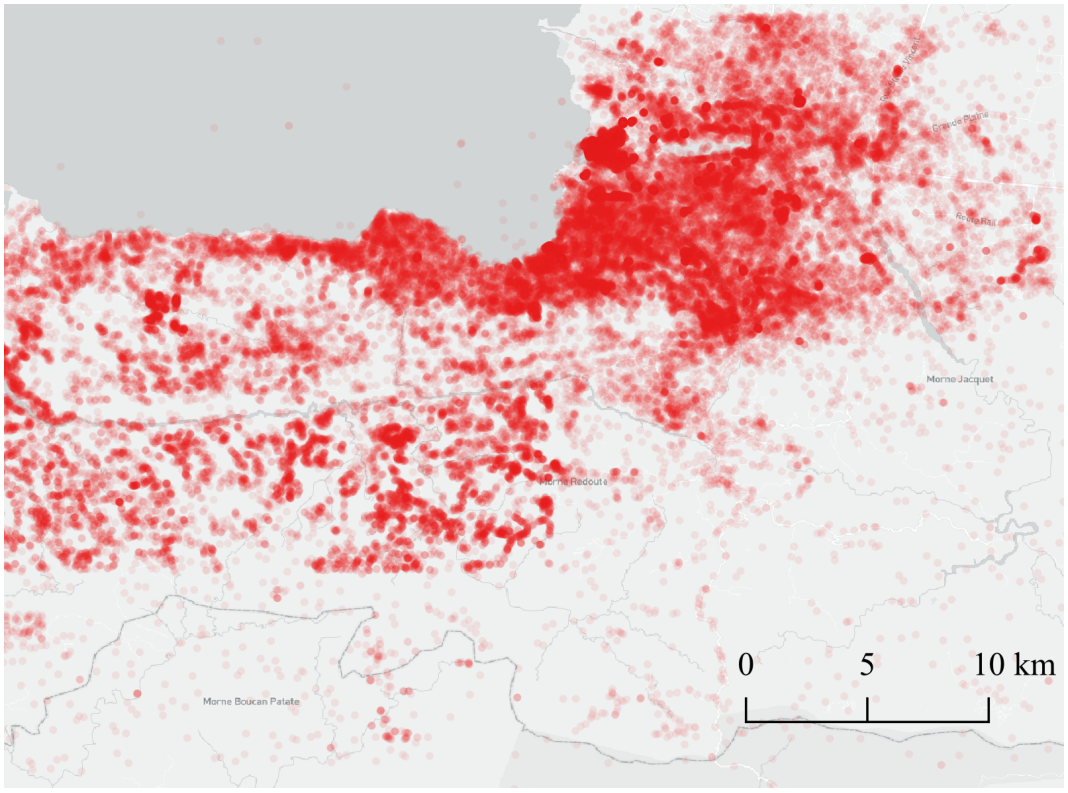

Figure B.2: Spatial distribution of features mapped in Port au Prince case study

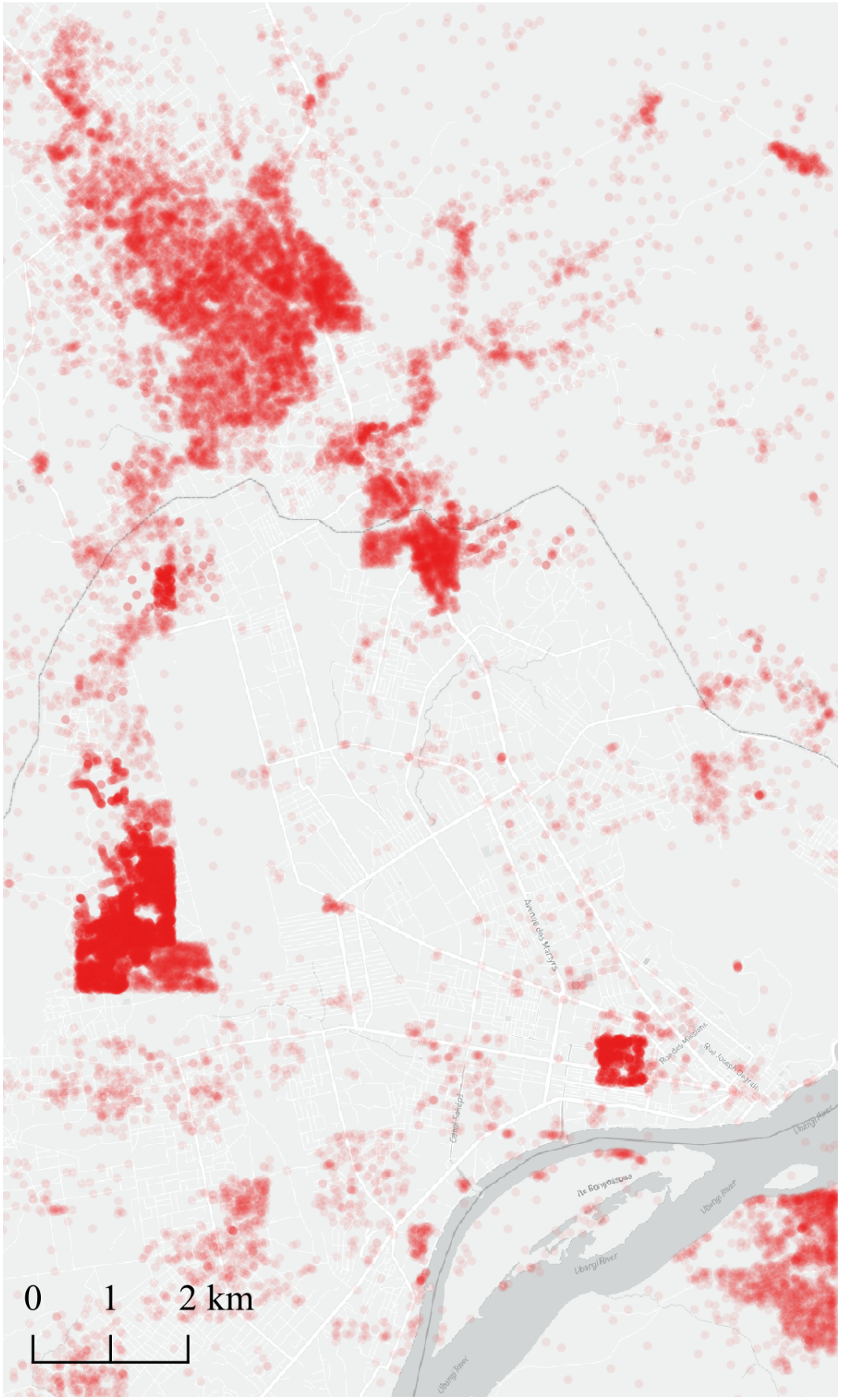

Figure B.3: Spatial distribution of features mapped in Bangui case study

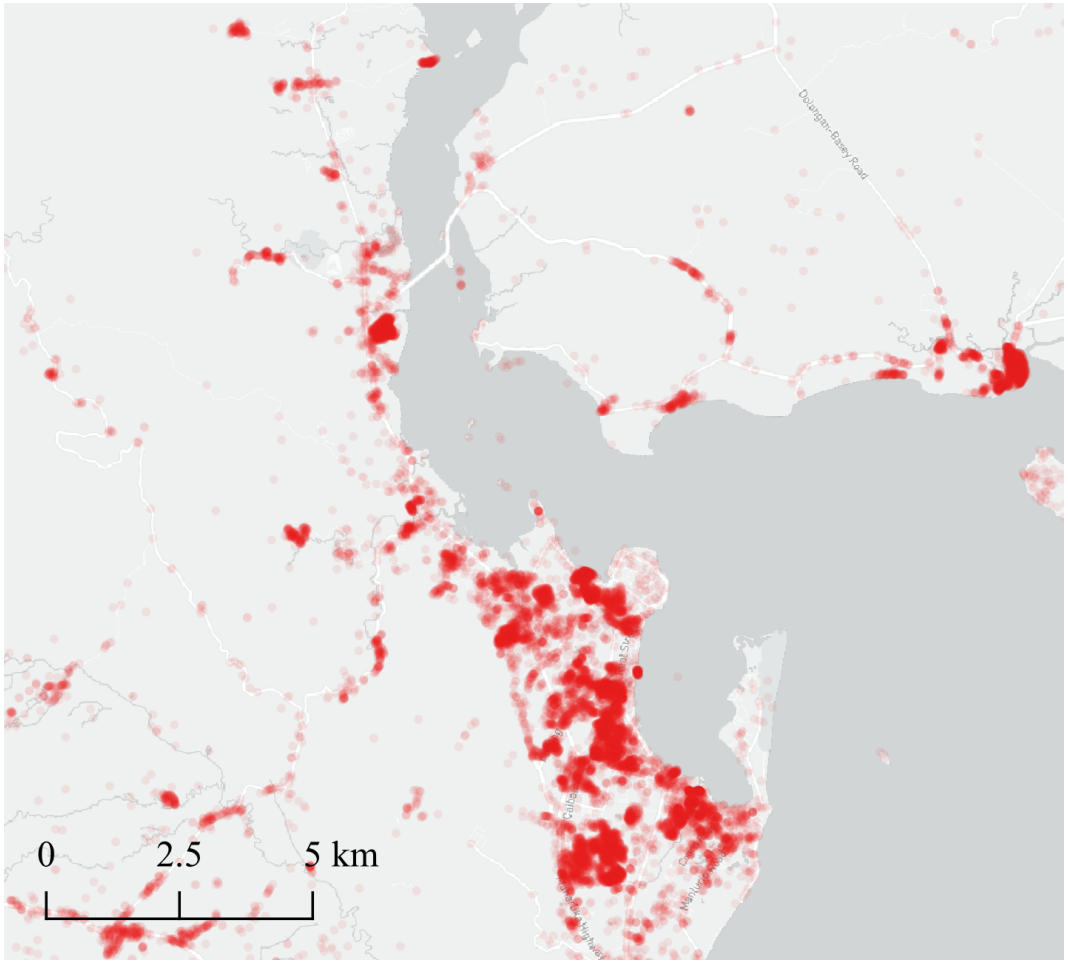

Figure B.4: Spatial distribution of features mapped in Tacloban case study

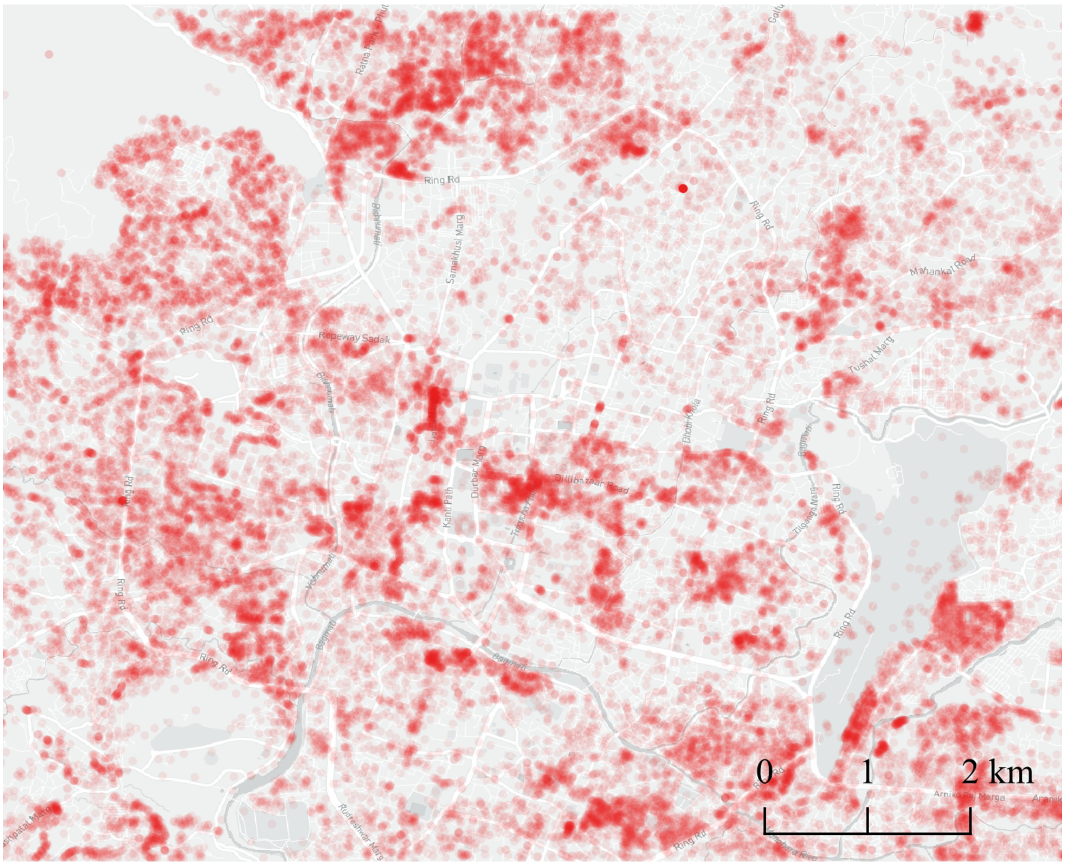

Figure B.5: Spatial distribution of features mapped in Kathmandu case study

Appendix C

Research Log

This appendix contains documentation of this research process.

Date	Notes
April	Begin to frame research topic; awaiting confirmation of support from MSF
April 16th	Initial check-in meeting with Sarah to discuss research aims
April 21st	Meeting with Sarah and Jorieke (of MSF) to discuss potential support
April 27th	Confirmation of support from MSF
May	Begin literature review and design methodology; challenges managing heavy workload with submissions from other coursework
May 4th	Review initial conceptual framework for project with Sarah
May 13th	Begin draft literature review
May 21st	Kick-off meeting with MSF to discuss project ideas
June	Prepare draft literature review and methods proposal; personal challenges in relocating back to UK from Canada
June 4th	Reviewed proposed Gantt chart for project with Sarah and discussed timeline for feedback
June 24th	Check in meeting with Jorieke (MSF) regarding initial findings from literature review
June 26th	Technical challenges in setting up local OSHDB database
June 30th	Meeting with Sarah to review key points from literature review and help to troubleshoot technical challenges

Date	Notes
July	Conduct analysis and draft content for results
July 3rd	Deliver draft literature review and methods roadmap to Sarah for review
July 9th	Received feedback from Sarah on draft material
July 29th	Meeting with Sarah to review preliminary results and discuss implications of findings
July 31st	Deliver draft results and discussion to Sarah for review
August	Focus on writing the dissertation document
August 3rd	Received feedback from Sarah on draft material
August 19th	Check in with Sarah to review document structure and style