COMPARING OPENSTREETMAP AND WIKIPEDIA EDIT PATTERNS DURING MAJOR EVENTS

by

Gary Esworthy
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geography and GeoInformation Science

Committee:

___________________________	Dr. Arie Croitoru, Thesis Director
___________________________	Dr. Anthony Stefanidis, Committee Member
___________________________	Dr. Andrew Crooks, Committee Member
___________________________	Dr. Anthony Stefanidis, Department Chair
___________________________	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
___________________________	Dr. Peggy Agouris, Dean, College of Science
Date: ______________________	Spring Semester 2016 George Mason University Fairfax, VA

Comparing OpenStreetMap and Wikipedia Edit Patterns During Major Events

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Gary Esworthy
Bachelor of Science
Indiana University of Pennsylvania, 2008

Director: Arie Croitoru, Professor
Department of Geography and GeoInformation Science

Spring Semester 2016
George Mason University
Fairfax, VA

DEDICATION

To B.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dr. Arie Croitoru, for all of his help, ideas, and insight. His guidance and feedback were key to the success of this thesis. I would also like to thank Dr. Andrew Crooks and Dr. Tony Stefanidis for being a part of my thesis committee.

I would also like to thank RadiantBlue Technologies for their financial support and allowing employees to continue their educations. I would especially like to thank my supervisors and team members for supporting my busy schedule while taking classes and writing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABSTRACT

COMPARING OPENSTREETMAP AND WIKIPEDIA EDIT PATTERNS DURING MAJOR EVENTS

Gary Esworthy, M.S.

George Mason University, 2016

Thesis Director: Dr. Arie Croitoru

During and in the aftermath of major natural disasters and other major events, data is rapidly contributed to online sources such as Wikipedia and OpenStreetMap. For Wikipedia, text, images, and sources are added to the event's page to. On OpenStreetMap, users attempt to update maps local to the area affected by the event by adding and updating roads, buildings, and other geospatial data. In both cases, users contribute data for public consumption. This thesis examines patterns of editing during and after such events, such as earthquakes and hurricanes, of both Wikipedia and OpenStreetMap. Using the Wikipedia Application Programming Interface and the various sources of OpenStreetMap history data, user contributions of these events will be processed to attempt to establish a pattern of edits. This data is analyzed in order to characterize and compare the patterns of activity in Wikipedia and OpenStreetMap and the results are compared with other events over time. The results indicate a positive

correlation between Wikipedia and OpenStreetMap contribution patterns for events that had organized mapping campaigns through OpenStreetMap. The contribution patterns for events that did not have organized mapping campaigns show less correlation.

1. INTRODUCTION

1.1 Background and Motivation

Crowdsourcing and sharing information on the web has been and continues to be a growing trend (Hudson-Smith et al. 2009). A prototypical example of online crowdsourcing of information is Wikipedia, the largest such online system that enables the sharing of knowledge on the internet (Xu and Li 2015). In recent years, OpenStreetMap has emerged as one of the most popular crowdsourcing communities for geospatial information, or volunteered geographic information (VGI; Goodchild, 2007). While the two sites share some similarities, such as the ability to contribute and freely search for information, they also differ in several ways. In view of this relationship, the objective of this thesis is to compare and analyze contribution patterns of user generated information in Wikipedia and OpenStreetMap. Focusing on periods during and after major earthquakes and other events, this research explores the patterns of user contributions in relation to such events, as well as compare and contrast these patterns across the two crowdsourcing platforms.

The selection of events for this study includes both events that have organized OpenStreetMap contribution campaigns and events that do not have these campaigns. This enables the study to add another point of comparison to see if there are any discernable contribution patterns and how they both compare to Wikipedia contribution patterns.

1.2 Overview of Approach

This thesis aims primarily to compare the trends of user-contributed data to Wikipedia and OpenStreetMap during and after a major event. The data has been gathered by means of official application program interface (API) and third-party datasets and parsed through custom scripts to generate change datasets that can be analyzed. Once these datasets have been acquired and processed, statistical models were run to determine what type of correlation could be found between the revision patterns of the two sources. Editing activity time series were created to visualize the editing patterns of the thirty day period starting when the event's Wikipedia article was generated. In addition to these time series, correlation values, and dynamic time warping graphs and distances (Keogh and Pazzani, 2001) were generated. These measures were then used to compare the Wikipedia revision patterns to the OpenStreetMap edit patterns.

1.3 Research Questions

The primary question that this thesis aims to address is how users contribute data to OpenStreetMap during a natural disaster and how it compares to editing patterns of Wikipedia for the same event. In particular, this research will explore editing activity in OpenStreetMap and Wikipedia following major events, such as earthquakes and weather events, in an attempt to compare and contrast these patterns across different crowdsourcing platforms. While OpenStreetMap has emerged in recent years as a prominent source of VGI data, currently there is limited research specifically on analyzing OpenStreetMap editing patterns.

Exploring these edit patterns will not only provide insight into how users contribute, but it may also provide the information necessary to improve the editing process. There have been many projects that have analyzed the quality and coverage of OpenStreetMap data (Forghani and Delavar 2014; Neis, Zielstra, and Zipf 2011; Girres and Touya 2010), but none have been found that address the editing patterns specifically. On the contrary, there have been studies that have looked at Wikipedia editing patterns. Specifically, Kämpf et al. (2012) studied the relationship of article access (or viewing) and article editing over time. Understanding the relationship across crowdsourcing platforms can also be beneficial for prediction purposes. For example, if a surge in Wikipedia editing activity is often followed by increased editing activity in OpenStreetMap, Wikipedia activity can serve as a possible predictor to the creation or enhancement of VGI in OpenStreetMap. Such analysis can only be carried out by quantifying and mining editing activity across such platforms rather than focusing on each source separately.

To further understand these patterns, other specific events will be analyzed from the OpenStreetMap Humanitarian project[1] to enable a comparison of organized crowd-sourcing efforts of OpenStreetMap to their Wikipedia pages and also to unorganized, or more organic, crowd sourcing.

1.4 Outline of the Thesis

The next chapter of this thesis, chapter two, will be a detailed literature review of the concepts, practices, and current state of the technology and services used in this study which will be vital to have an understanding of this study's research question and methodologies. Chapter two will briefly outline the problem statement and hypothesis of

[1] https://hotosm.org/

this study. Chapter three will provide information about the data used in this study and how it was acquired. Chapter four will detail the methodology used to process the data and obtain results for this study. Chapter five will present the results of the study and break them down into sections for each individual study. Chapter six will discuss the results as a whole and examine their implications. Chapter seven will serve as a concluding chapter and will examine any further work that can be done in relation to this study.

2. LITERATURE REVIEW

2.1 Introduction

The purpose of this chapter is to introduce the reader to the concepts of volunteered geographic information and the Wiki Model as well as the specific communities that contribute to Wikipedia and OpenStreetMap. The discussion of this topic begins with a review of the two platforms and their characteristics and introduces the so-called hype cycle pattern as a prototypical pattern of contributions in the wiki and the VGI paradigms. Section 2.1 will detail the history of the Wiki Model and VGI. Section 2.2 will look at research specifically related to VGI and disaster relief. Sections 2.3 and 2.4 will discuss Wikipedia and OpenStreetMap, respectively. Finally, section 2.5 will present information regarding hype cycles, curves, and trends that will be used to look at the results of this research.

2.2 The Wiki Model and VGI

The Wiki model is described as an online, virtual knowledge-sharing communities that are self-organized and rely heavily on volunteers to operate (Prasarnphanich and Wagner 2011). This model has evolved and spread to other areas, such as business and geospatial applications. In the business world, there is a shift toward using online communities such as blogs, wikis, and message boards to communicate with and learn from external sources (Prasarnphanich and Wagner 2011). There is also a recent shift towards this type of model with VGI. Sites like OpenStreetMap[2] and Wikimapia[3] have emerged that allow this type of open collaboration for geographic information (Goodchild, 2007). As the various online collaborative communities grow, there is an increasing desire to include geographic information with traditional 'non-geographic' data (Hudson-Smith et al., 2009).

Users' motivation to contribute to these communities has been the focus of many research efforts. These communities often lack direct incentive for participation and therefore much research has been done to determine what motivates users to contribute to these communities. Budhathoki and Haythornthwaite (2013) concluded that there are two main types of users in terms of the amount of contribution: lightweight and heavyweight. They found that lightweight users generally align their goals to the goals of the overall system (i.e. open source data should be available to all users). Heavyweight users, on the other hand, focused more on career and financial gains associated with the community and the data. Xu and Li (2015) found that content contribution is usually driven by the user's need for self-development while community participation in general is driven by the desire for a sense of community. Hargittai and Shaw (2014) found that a combination of gender and level of Internet skills contribute heavily to the motivating factors of contribution.

2.3 Wikipedia

Wikipedia is perhaps the most well-known example of a 'wiki model' community. Wikipedia was born as a 'fun' project by the creators of 'Nupedia', an

[2] https://www.openstreetmap.org
[3] http://www.wikimapia.org/

expert-driven online encyclopedia with much more rigor and seriousness than Wikipedia (König 2013). This 'fun' project would soon grow exponentially and become the world's most well-known example of online collaboration. Today, Wikipedia boasts 70,000 active contributors working on over 35 million articles in almost 300 different languages ("About Wikipedia", 2015).

A large body of research exists with regards to Wikipedia and its users, communities, and their editing patterns. Ortega and Gonzalez Barahona (2007) examined the editing patterns of the users that contribute the most to Wikipedia. They were able to find new editing patterns of these users and also presented a tool that can be used for quantitative analysis of these users' editing patterns.

Wikipedia can also be considered a graph, with each article representing a node and each link between articles representing an edge (Buriol et al. 2006). Using this idea, which they called the 'Wikigraph', Buriol et al. found that the number of articles, edits, and visitors were growing exponentially while the size of articles was growing linearly.

2.4 OpenStreetMap

OpenStreetMap is a free and open platform used to create, edit, and share geographic information (Budhathoki and Haythornthwaite 2013). This information includes roads, buildings, and other geographic data. At the time of this writing (March 2016), OpenStreetMap has over 2.5 million registered users and over 300 million 'ways', or geospatial features, in the database ("OpenStreetMap Stats", 2015).

The editing and annotation process is well documented and has been studied extensively. Mooney and Corcoran (2012) describe this process as open and collaborative, where users can use GPS and other mobile devices to make edits to the OpenStreetMap data by uploading data and even overlaying aerial and space-born imagery. This process also allows users to tag, or annotate, the data with freely chosen words which allows various GIS tools and databases to better understand the nature of the objects in the database.

While the amount of research regarding OpenStreetMap does not compare to the amount of research done with Wikipedia, OpenStreetMap has been the focus of a growing number of research projects. As documented in *OpenStreetMap in GIScience* (Arsanjani et al. 2015), OpenStreetMap is mentioned in over 400 papers in the ACM digital library and can be found in over 200 search results from search engines such as ScienceDirect and Scopus. However, within this growing body of literature there has been relatively few studies on contribution patterns in OpenStreetMap and how these patterns compare to other well-established crowdsourcing efforts such as Wikipedia. This is a primary motivation for the proposed study.

2.5 VGI and Disaster Relief

The major earthquake in Haiti that occurred on January 12, 2010 has been viewed as a major case study for how VGI can impact disaster relief. Zook et al. (2010) researched how information technology (IT) in general was used and also looked specifically at how VGI, such as user-generated maps and road networks, impacted the relief efforts.

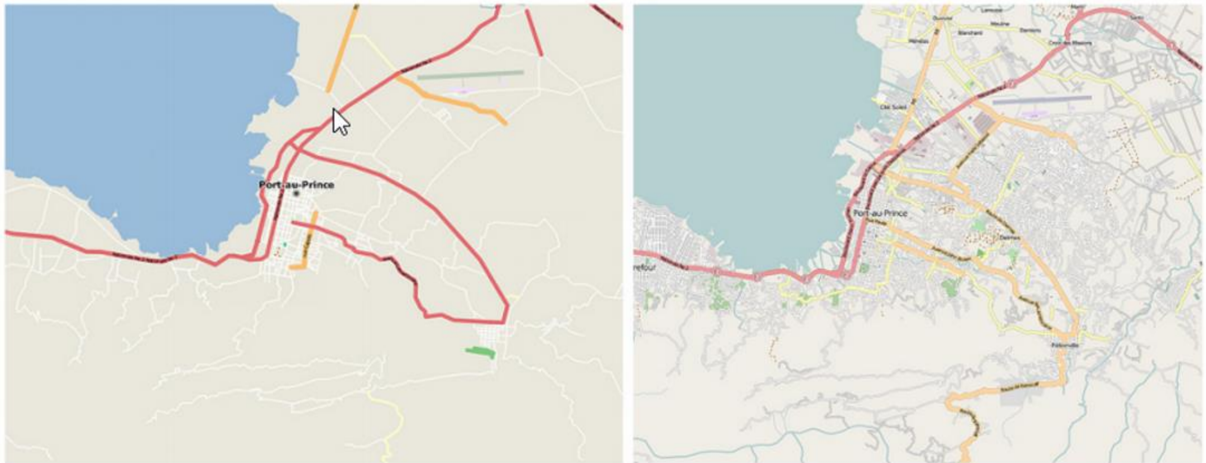

Figure 1 OpenStreetMap data in Haiti before (left) and after (right) the earthquake of 2010 (Zook et al., 2014).

Using VGI for disaster relief has a broad range of development and support. These efforts can range from massive government and commercially-backed ventures such as OpenStreetMap Haiti (Shemak 2014) to small two-person teams creating a website and asking friends to help (Singel 2005). The OpenStreetMap Haiti project utilized satellite imagery, aerial photographs, and global positioning system (GPS) data that was provided or acquired from different government and commercial entities. Figure 1 shows snapshots of the OpenStreetMap data in Haiti before and after the earthquake of 2010. A smaller example of using VGI for disaster relief is the Scipionus[4] project, which was created by a pair of young software engineers in Austin, Texas to enable the crowd-sourcing of data during the 2005 Hurricane Katrina. This small Wiki page utilized the Google Maps API to allow users to enter small text entries regarding property damage, missing persons, and other emergency information.

---

[4] http://gregstoll.dyndns.org/scipionus/

In addition to OpenStreetMaps's open contribution system, which is similar to Wikipedia's editing system, an organization called the Humanitarian OpenStreetMap Team (HOT) was formed to provide organized mapping campaigns ("Humanitarian OpenStreetMap Team" 2016).

> "The Humanitarian OpenStreetMap Team [HOT] applies the principles of open source and open data sharing for humanitarian response and economic development." – Humanitarian OpenStreetMap Team (2016)

The HOT mission statement, which precedes this paragraph, states that the goal of the organization is to provide open, free data for humanitarian response and economic development purposes. The HOT, which is a 501(c)(3) nonprofit organization, organizes campaigns to rally volunteer contributors to provide data during times of crisis. As of March 2016, the HOT currently has seven open mapping campaigns, which can be seen in Table 1.

Table 1. Current HOT campaigns as of March 2016 ("Humanitarian OpenStreetMap Team", 2016)

Event Year	**Action**
2015	Humanitarian Mapping Project, Salgar Landslide
2015	Activation, 2015 Nepal earthquake
2014-2015	Activation, 2014_West_Africa_Ebola_Response 2014, March - ongoing
2014	Humanitarian Mapping Project, Tharparkar Drought in Pakistan - ongoing
2014	Humanitarian Mapping Project, 2014 Paraguay floods
2013	Humanitarian Mapping Project, Central Africa Republic (CAR) Crisis
2013	Humanitarian Mapping Project, South Sudan Crisis Response - ongoing

With the rise of VGI disaster relief efforts, questions are raised about the validity and accuracy of this data. Camponovo and Freundschuh (2014) found that 50% of messages from victims, such as text, e-mail, and voice messages, were categorized incorrectly translated and categorized by volunteers. This can pose a major problem to emergency responders and crisis management personnel since this data is not validated the way that official data would be validated and raises issues with regard to data quality.

Many research projects have looked at the validity and accuracy issue of VGI and some opinions seem to favor the helpfulness of the data over the possibility of inaccuracies (Goodchild and Glennon, 2014) while others caution that it depends greatly on the use of the data (Forghani and Delevar, 2014; Girres and Touya, 2010). Other studies have found that the quality of VGI tends to rise over time. Neis et al. (2012) looked at the road networks of OpenStreetMap in France and determined that while not error-free, the trend was that the number of errors was going down and the quality of the data was rising. They also determined that in countries with large OpenStreetMap datasets, the data is becoming comparable to commercially provided data. Jackson et al. (2013) developed a methodology to compare VGI with reference datasets while other work has been done to try to associate demographics to VGI quality (Mullen at al, 2015).

2.6 Hype Cycles, Trends, and Correlation

This study will attempt to find patterns in user contributions after various events. An example of such a pattern is known as Garter's Hype Cycle (shown in Figure 2). Garter's Hype cycle can be defined as "the typical progression of an emerging technology from overenthusiasm through a period of disillusionment to an eventual understanding of the technology's relevance and role in a market or domain" (Linden and Fenn 2003).

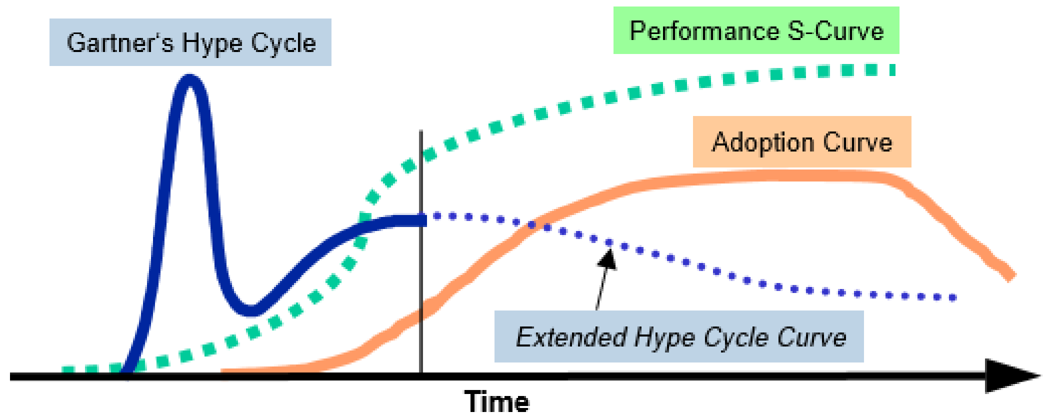

Figure 2. Garter's Hype Cycle compared to other curves (Linden and Fenn 2003).

While some work has been done to look at activity patterns of Wikipedia and wikis in general, there is a lack of such work for OpenStreetMap or similar VGI sites. For example, Arazy and Croitoru (2010) looked at activity patterns of corporate wiki systems rather than the larger, public Wikipedia. They found that these systems became inactive after a relatively short period, which can be likened to a hype cycle. On the other hand, there is a noticeable gap in similar research for OpenStreetMap or other VGI communities that this study and future work will aim to fill.

The Pearson correlation coefficient, or Pearson's *r* value, is used in many disciplines to determine correlation of datasets. Fields such as signal processing and computer science (Di Lena and Margara, 2010), agriculture (Taylor and Bates, 2013), healthcare (Graña et al., 2011), and others are highly researched using Pearson's *r* value as a basis. Pearson's *r* value, in short, simply shows the linear correlation between two variables where the result will be between -1 and 1, inclusive. Results close to -1 indicate a strong negative correlation while results close to 1 show a strong positive correlation. Results close to 0 indicate that there is little to no correlation between the two variables (Dowdy and Wearden, 1983).

$$r = r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\ \sqrt{(\sum y_i^2 - n\bar{y}^2)}}. \qquad \text{(Equation 1)}$$

Equation 1 shows how to determine the Pearson correlation coefficient, or *r*, for two datasets. The datasets are represented by $\{x_1 \ldots, x_n\}$ and $\{y_1 \ldots, y_n\}$ which each contain *n* values. There are many different forms of this equation to produce *r*, all with different benefits and potential computational issues.

In addition to basic checks for correlation, this study will also use a dynamic time warping (DTW) algorithm to look at the two series of data as they progress over time. Liao (2005) describes dynamic time warping as "the generalization of classical algorithms for comparing discrete sequences to sequences of continuous values". DTW attempts to align the given data series so minimize their difference in order to find the shortest path, or distance, between the two datasets.

$$d_{DTW} = \min \frac{\sum_{k=1}^{K} w_k}{K} \quad \text{(Equation 2)}$$

Equation 2 shows the DTW algorithm that starts with two time series, *Q* and *R*, in which *Q* contains *n* elements and *R* contains *m* elements. From there, an n x m matrix is constructed in which each cell (i, j) of the matrix contains the distance *d*, typically the Euclidian distance, between two points Q_i and R_j. A warping path *W* is then sought that will satisfy three constraints. First, a boundary condition requires the path to begin and end at opposite corners of the matrix, or at (1, 1) and (m, n). A continuity constraint requires that all steps be to adjacent cells. Finally, a time constraint requires the points on the path to be monotonically spaced in time. These warping paths are then applied to Equation 2 to find the minimum distance (Liao, 2005). To summarize, the algorithm attempts to step along each cell to determine the warping path in order to minimize the warping cost (Keogh and Pazzani, 2001).

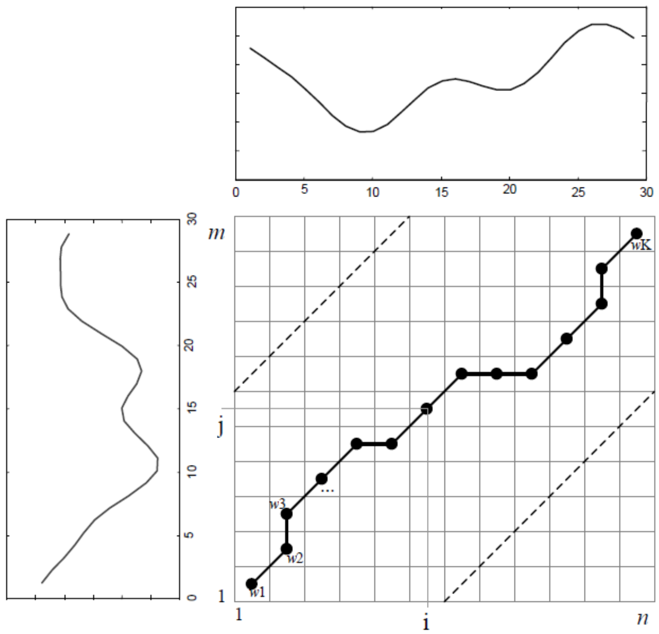

Figure 3. An example of a DTW warping path (Keogh and Pazzani, 2001).

Figure 3 shows an example of a DTW warping path that can be found using dynamic programming (Keogh and Pazzani, 2001). This is necessary due to the fact that there are an exponential amount of warping paths that could satisfy the conditions of the algorithm. For this study, the mlpy[5] Python library will be used to calculate the DTW shortest distances for each case study. This library, along with other Python graphing libraries, will also be used to generate the matrix graphs showing the paths that were used to determine the shortest distance.

[5] http://mlpy.sourceforge.net/

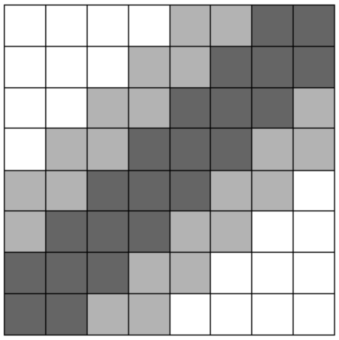

Figure 4. Example of dynamic time warping cell shading (Salvador and Chan, 2004)

The dynamic time warping distance results of this thesis will be shown on a shaded graph similar to the example shown in Figure 4. The shading of the cells simply shows which cells were calculated at each resolution using the dynamic time warping algorithm implemented in the mlpy library, which relies heavily on the work done by Salvador and Chen (2004).

3. PROBLEM STATEMENT AND HYPOTHESIS

This study aims to explore patterns in volunteered geographic information (VGI) edits during and immediately after a major earthquake event. This data is compared to editing patters of Wikipedia to determine if there is a common pattern and if not, how the patterns compare. These events are examined for a period of thirty days following the event to quantify the immediate and follow-up revision patterns of the different communities.

When reviewing literature for this study, a noticeable gap was found specifically regarding OpenStreetMap editing patterns. While there has been much research in the *quality* of OpenStreetMap data, little to no work has been done to look at the *quantity* and *frequency* of OpenStreetMap edits. The book *OpenStreetMap in GIScience* (Arsanjani et al. 2015) details the efforts of several studies that examined the quality and reliability of VGI contributions to OpenStreetMap.

The results of this study and future work regarding this study may also contribute to the understanding of activity patterns and motivation in crowdsourcing VGI. This study aims to understand the differences and similarities of Wikipedia and OpenStreetMap. While they are both crowdsourced, they are very different in nature, handle very different datasets, and potentially have very different user bases. This work also seeks to explore utilizing Wikipedia activity as a possible predictor of

OpenStreetMap activity, or vice-versa. For example, after a major event, does activity spike on one of the sites first? If so, can that be used to predict activity on the other site?

Considering that OpenStreetMap is a global, worldwide dataset that changes frequently and that Wikipedia pages are about a specific topic (i.e. a person, place, or event), this thesis hypothesizes that the patterns will not exactly match. Patterns of edits of a Wikipedia page about an event have a clear beginning (when the page was created) and a clear pattern following that event. There should be little to no 'noise' in the Wikipedia edit data due to the fact that the Wikipedia page is only about the event in question. OpenStreetMap, on the other hand, is edited by many people for many different purposes. It is impossible to know if a particular edit was caused by the event or if the edit would have happened anyway.

This study also focuses on looking at both events that have mapping efforts that are being coordinated by the Humanitarian OpenStreetMap team and those that do not. These patterns are compared directly to the Wikipedia contribution patterns to determine if there is any correlation between them and also if there are any pattern differences between those events that have coordinated mapping campaigns versus those that do not. This study hypothesizes that the revision datasets for those events that have organized mapping campaigns will be larger than those that do not have coordinated efforts. The question that will remain is what exactly these patterns look like and how they compare to other volunteered information.

Finding and comparing these patterns will be the primary focus of this thesis. The Wikipedia editing patterns can almost be seen as a 'baseline' pattern (clean data regarding only that event) while the OpenStreetMap data will require more effort to identify these patterns for comparison.

4. DATA

4.1 Earthquakes

For this research, major earthquakes will be the focus while other events will be used to supplement and for comparison. High magnitude earthquakes typically have a Wikipedia page created about them that describe the location, magnitude, damage, and other details. Earthquakes were chosen due to their relative frequency and overall media attention. There were 20 earthquakes in 2015 of at least 7 magnitude (“Earthquakes”, 2015) which can be seen in Table 2.

Table 2. Earthquakes in 2015 of at least 7.0 magnitude (“Earthquakes”, 2015).

Date	Magnitude	Location	Wikipedia Page?
2/13/2015	7.1	Northern Mid-Atlantic Ridge	No
2/27/2015	7	Indonesia	No
3/29/2015	7.5	Papua New Guinea	No
4/25/2015	7.8	Nepal	Yes
5/5/2015	7.5	Papua New Guinea	No
5/7/2015	7.1	Papua New Guinea	No
5/12/2015	7.3	Nepal	Yes
5/30/2015	7.8	Japan	No
6/17/2015	7	Southern Mid-Atlantic Ridge	No
7/18/2015	7	Solomon Islands	No
7/27/2015	7	Indonesia	No
9/16/2015	8.3	Chile	Yes
9/16/2015	7	Chile	No
10/20/2015	7.1	Vanuatu	No
10/26/2015	7.5	Afghanistan	Yes
11/18/2015	7	Solomon Islands	No
11/24/2015	7.6	Peru	No
11/24/2015	7.6	Peru	No
12/4/2015	7.1	Southeast Indian Ridge	No
12/7/2015	7.2	Tajikistan	Yes

Of the 20 major earthquakes of 2015, 5 have corresponding Wikipedia pages. Only earthquakes with Wikipedia pages have been selected for this study since the comparison between OpenStreetMap and Wikipedia edits is crucial. This study also includes an earthquake from January 2016 as well a sub-7.0 earthquake from 2015 that have a corresponding Wikipedia entry to add another event.

Table 3. List of earthquakes chosen for this study.

Date	Magnitude	Location
4/25/2015	7.8	Nepal
5/12/2015	7.3	Nepal
6/5/2015	6	Malaysia
9/16/2015	8.3	Chile
10/26/2015	7.5	Afghanistan
1/24/2016	7.1	Alaska (US)

The earthquakes that have been chosen, as seen in Table 3, each have a Wikipedia entry describing the event as well as a geographic area that can be studied for OpenStreetMap data. The following sections will detail how this data was found and acquired. In addition to the 7 earthquakes that were chosen for this study, 2 non-earthquake events were chosen with their corresponding Wikipedia entries to allow for a comparison of different types of events. Table 4 shows the non-earthquake events that were chosen for this research.

Table 4. List of non-earthquake events selected for this study.

Date	Event	Location
April 2014	West Africa Ebola outbreak	Guinea and Sierra Leone
October 2015	Hurricane Patricia	Mexico
November 2015	Paris attacks	Paris, France
May 2015	Texas flood and tornado outbreak	Texas, United States

Figure 5 shows a map of all data used for this study including both earthquakes and non-earthquake events. The polygons representing the OpenStreetMap data show that the entire state or country of interest was used for processing and analysis.

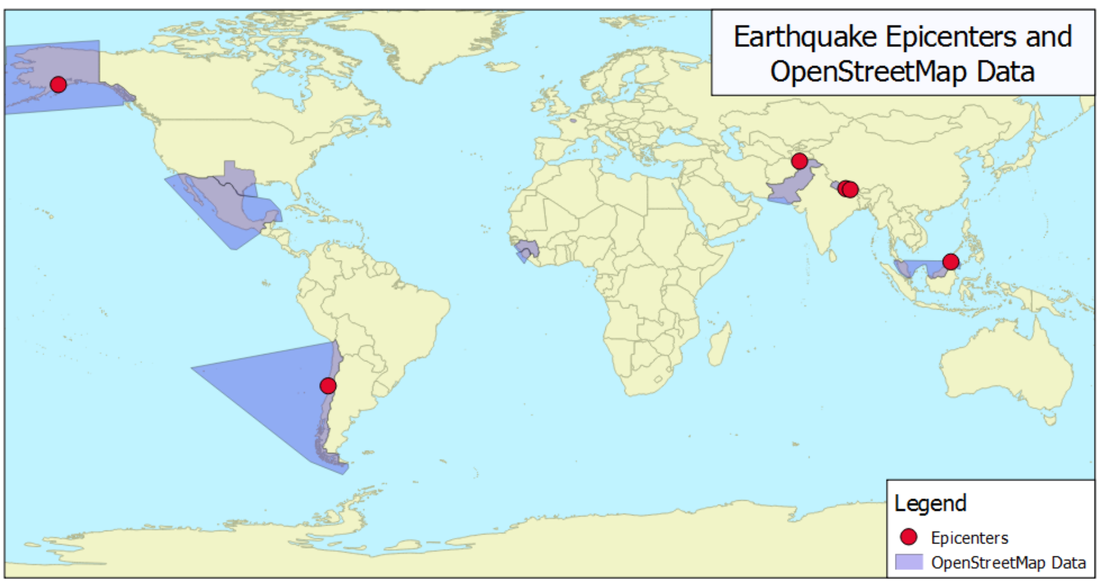

Figure 5. Map showing the epicenter of the earthquakes chosen and their corresponding OpenStreetMap dataset along with the additional events' OpenStreetMap datasets.

4.2 Wikipedia

Wikipedia data is acquired through the official Wikipedia API that is publicly accessible. A Representational State Transfer (REST) request will be made to the Wikipedia server that includes the page name, the beginning and end dates of interest, and what type of edit data should be returned. Specifically, the Wikipedia Revisions API will be called in order to obtain the data that will be used for this study ("MediaWiki Revisions API", 2015). The Wikipedia Revisions API query will return a collection of revision objects which contain:

- The name of the user that edited the page
- The timestamp of the revision
- The comment that the user attached to the revision
- The format of the revision
- The model of the revision
- The content of the revision
- The size of the revision in bytes

These fields were determined to be the relevant fields for this study to allow the analysis of revisions. Fields such as timestamp and size are crucial for data analysis while fields like format and model are used to be sure that the data is in the proper format. Some fields, such as the name of the user that made the edit, will likely not be used for this study but were included for data analysis and verification to see what types of users were making the edits.

```
https://en.wikipedia.org/w/api.php?action=query
    &prop=revisions
    &format=json
    &rvprop=ids%7Ctimestamp%7Cuser%7Csize
    &rvlimit=500
    &rvcontentformat=application%2Fjson
    &rvstart=2015-10-01T12%3A00%3A00Z
    &rvend=2015-10-31T12%3A00%3A00Z
    &rvdir=newer
    &rvtoken=rollback
    &titles=2015_Hindu_Kush_earthquake
```

Figure 6. Sample REST service call used to query Wikipedia revision data.

The query shown in Figure 6 will return the revisions for the given parameters in JavaScript Object Notation (JSON) format. JSON is a lightweight data format that is easy to generate and parse ("JSON", 2016). This type of query will be run for all earthquake pages for this study. The data that is retrieved from this method will be stored and used for later processing and analysis. To complete these queries, basic Python libraries will be used in order to allow the scripts to request, retrieve, and parse the JSON responses from the Wikipedia API server. These scripts will be used for each case study for this thesis in order to ensure that the data retrieved is in the exact same format. The processing of this data will be described in the Methodology chapter.

4.3 OpenStreetMap

Collecting revision data from OpenStreetMap is more challenging than collecting Wikipedia revision data. The public OpenStreetMap API was written primarily for editing content, not querying historical content. Due to the open nature of OpenStreetMap data, which is licensed under the Open Data Commons Open Database License (ODbL;

"OpenStreetMap Copyright and License" 2015), many third-party developers ingest and process this data to offer secondary products.

One such service providing secondary products is Geofabrik, which translates (from German) to "geo factory" ("OpenStreetMap Data Extracts, 2015"). Geofabrik specializes in processing OpenStreetMap data and provides daily, weekly, and monthly revision data for every region of the world. Using this service, OpenStreetMap snapshots will be downloaded for the geographical regions surrounding earthquake sites for processing.

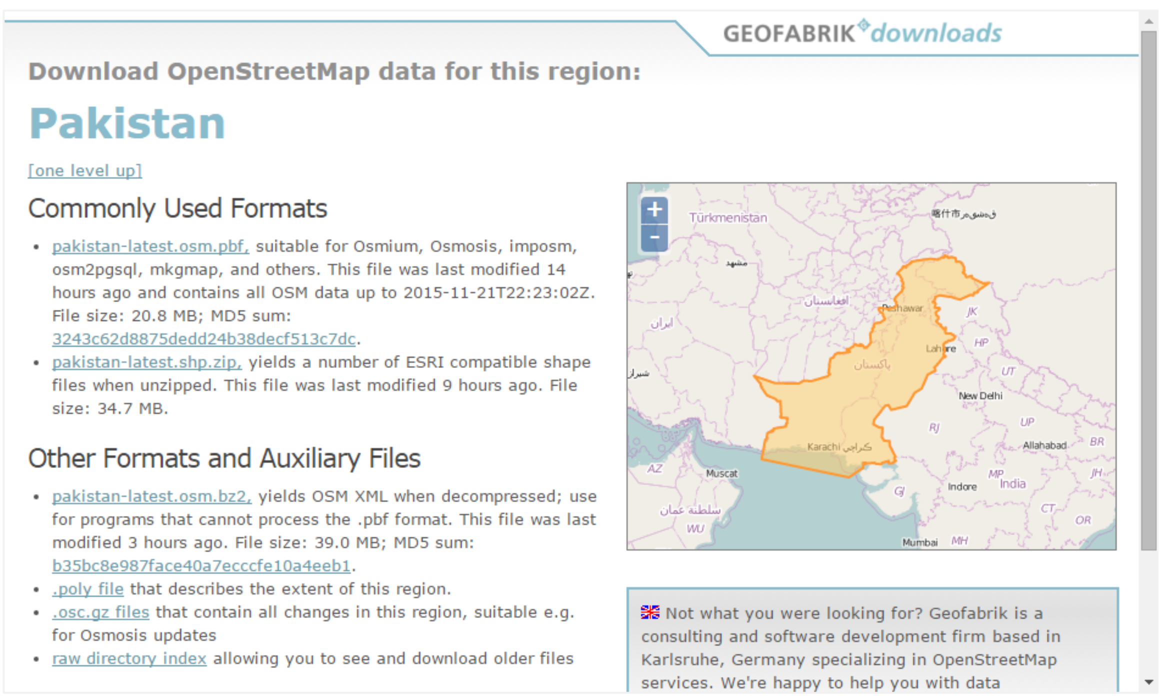

Figure 7. Screenshot of the GeoFabrik download site for the Pakistan region.

Figure 7 shows the ‘Pakistan’ page of GeoFabrik, which allows OpenStreetMap data to be downloaded in the given geographic region in a variety of formats. The geographic region pages of GeoFabrik contain download links for ESRI shapefiles (.shp), OpenStreetMap Extensible Markup Language (XML) files, polygon files (.poly), and OpenStreetMap Protocol Buffer Binary Format (.osm.pbf) files. Protocol buffers are an XML-like file format but is smaller in size and can be processed more efficiently (“Protocol Buffers”, 2016). Protocol buffers have been adopted by OpenStreetMap as a primary file format for transmitting and storing data. It is supported by a large number of third-party applications for reading, writing, and processing OpenStreetMap data (“PBF Software Compliance”, 2016).

For this research, a polygon file (.poly) will be downloaded and used to display the areas of study on a map for visualization purposes and two OpenStreetMap Protocol Buffer Binary Format files will be downloaded from each region. The first file will be downloaded and will be referred to as the ‘start’ file. This file shows the state of the geographic region before the earthquake occurred. The second file will be referred to as the ‘end’ file and will be a snapshot of the state of OpenStreetMap data in the geographic region after the earthquake. These two files will be processed to determine what has changed in the region between the two dates. The processing of this data will be described in chapter 6.

4.4 Defining an Edit

Since two different data sources are being used with different formats, the definition of an ‘edit’ must be clearly defined and used consistently throughout the case studies. For Wikipedia history data, the concept of a single ‘revision’ will be considered an edit. A revision in Wikipedia is any user-submitted change to an article (“MediaWiki Revisions API”, 2015). OpenStreetMap revision history data is more complex, so the lowest level data item will be used to track revisions. In OpenStreetMap revisions, a ‘node’ is added, removed, or modified (“OsmChange”, 2015). These node changes will be used to count revisions for the OpenStreetMap data. This will allow a comparison between every time the page is changed on Wikipedia to every time a node is changed in OpenStreetMap. The primary correlations for this study will compare the number of Wikipedia edits to the number of nodes changed in OpenStreetMap.

5. METHODOLOGY

5.1 Introduction

Once the data is retrieved as described in the chapter 5, it must be processed. The processing for this study is comprised of several steps for each type of data – Wikipedia revision data and OpenStreetMap data. The datasets that were retrieved are very different and require custom processing to gain meaningful insight. Custom scripts were written in the Python programming language for this study to achieve this goal. There are many free and open source libraries that can be used to process, parse, and generate statistics on a given dataset. This chapter will outline the steps taken on the different datasets in order to create a common, comparable dataset. Figure 8 shows the overall processing chain that was used for this study. The sections in this chapter will detail each step taken for both datasets. Section 6.2 will outline the steps taken to process Wikipedia revision data. Section 6.3 will detail how to process OpenStreetMap revision data. To conclude, section 6.4 will describe how the common datasets were brought together and how relevant statistical data was derived from them.

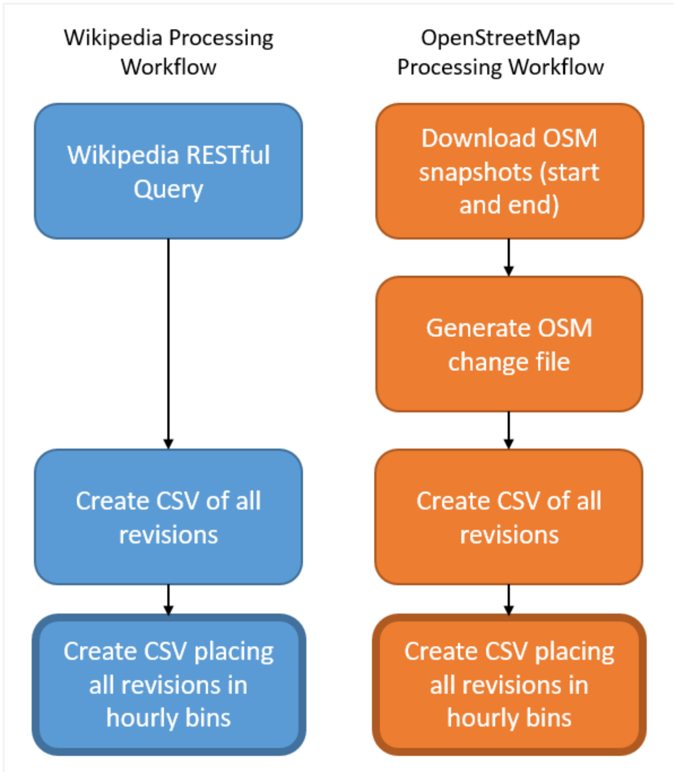

Figure 8. Flowchart showing the workflow of the data processing used in this study.

5.2 Wikipedia Revision Data

Wikipedia revision data, the simpler of the two data formats, will go through three specific steps in order to be ready for analysis. First, the data will be gathered via the Wikipedia REST API as described in the data chapter. Next, a comma-separated value file (CSV) will be created showing all revisions made to the page in the given time frame. Each row of the file will contain the timestamp, user name, and total edit size of the revision. Figure 9 outlines the overall algorithm used to obtain and process Wikipedia revision data for a given article and time frame.

```
START
        Create output files
        Request revision history

        WHILE (request response has revisions)
                Store revision
        END

        WHILE (request response has continue)
                Request next response
                WHILE (request response has revisions)
                        Store revision
                END
        END

        WHILE (stored revisions has revisions)
                Process each revision to determine min/max time
        END

        Create mapping of each hour between min and max

        WHILE (stored revisions has revisions)
                Process each revision to put each revision in an hour bin
        END

        Write output files
END
```

Figure 9. Algorithm used to query and process Wikipedia revision data.

The algorithm is derived mostly from the restrictions of the Wikipedia REST API. For example, the request must be repeated several times due to the results being limited to 500 by the API. For every 500 results, the algorithm will request the next batch of revision data and continue to do so until the response no longer contains the ‘continue’ flag, which is used to tell the end-user if there is more data to retrieve.

5.3 OpenStreetMap Revision Data

OpenStreetMap revision data is derived primarily from the Osmosis[6] command line tool. As described in the chapter 5, two OpenStreetMap snapshot files are downloaded for each impacted area from the Geofabrik website. These files are then processed through the Osmosis tool to generate an OpenStreetMap change file (.osc) that details every revision that was found between the two snapshots.

```
osmosis
    --read-xml file="july.osm"
    --read-xml file="june.osm"
    --derive-change
    --write-xml-change file="diff.osc"
```

Figure 10. Command example to generate an OpenStreetMap change file from OpenStreetMap snapshots.

Figure 10 shows how to use the Osmosis command line tool to generate an OpenStreetMap change file from two OpenStreetMap snapshot files. This change file is written in XML format to ensure that it can be processed cleanly. Once this file is generated, a Python script is run on the resulting change file to parse the revisions and output a CSV file for analysis. This process and output will be similar to the Wikipedia process and output to ensure that the files can be merged and analyzed consistently. Figure 11 shows the algorithm used to process the OpenStreetMap change files.

[6] http://wiki.openstreetmap.org/wiki/Osmosis

```
START
        Read OSM change file

        WHILE (there are more revisions)
                Process each revision to determine min/max time
        END

        Create mapping of each hour between min and max

        WHILE (there are more revisions)
                Process each revision to put each revision in an hour bin
        END

        Write output files
END
```

Figure 11. Algorithm used to process the OpenStreetMap change file.

5.4 Comparing the Datasets

Once both sets of data have been processed and the resulting CSV files have been created, they are ready to be merged and analyzed. First, the files are modified to show how many edits were completed on each platform during six hour periods. Six hour bins were chosen as a reasonable middle ground between the high temporal resolution of hourly bins versus the lower temporal resolution of daily bins. Due to significant time periods where edits were not happening on either platform, there were many hourly bins which contained 0 for several hours in a row. Moving to six hour bins proved to give a much cleaner dataset for analysis.

Once the data has been merged into six hour bins, the three datasets (Wikipedia edit counts, Wikipedia bytes changed, and OpenStreetMap node edits) are merged into a single spreadsheet based on their matching time tags. This is now the first time that the datasets are in the same file with the same time formatting and they can start to be compared. Both the SciPy[7] python library and the built-in Excel PEARSON function were used to determine the Pearson correlation coefficient between the datasets[8]. The same datasets that were used to calculate the Pearson correlation coefficient are then sent into a final Python script to calculate the dynamic time warping distance and generate the warping graph using the mlpy Python module. The documentation example of how to create these graphs was used (“Dynamic Time Warping”, 2016).

[7] http://www.scipy.org/

[8] The Microsoft Excel PEARSON function and the ScyPy PEARSONR function proved to give the same results.

6. RESULTS

6.1 Introduction and Overview

This chapter will provide a detailed breakdown of the results gathered from this study. Each following section will detail the results specific to each event that was studied. The overall results are presented in this section in Table 5 and Table 6. Table 5 shows the intermediate results of post-processing the data including the raw numbers of revisions for each platform. Table 6 is a summarized look at the correlation coefficients and dynamic time warping distances that were calculated for each case study and whether or not that event had a corresponding HOT mapping campaign. Each subsequent section will look at these values in depth and attempt to interpret their impacts. To bring in more events with humanitarian efforts, the final two sections of this chapter will look at non-earthquake events that had a significant humanitarian campaign associated with it.

Table 5. Post-processing results showing OpenStreetMap and Wikipedia revisions.

Event	**Start Date**	**End Date**	**OSM Node Edits**	**Wikipedia Edits**
Pakistan earthquake	10/26/2015 2:00	11/24/2015 20:00	31,900	494
Alaska earthquake	1/23/2016 4:00	2/21/2016 22:00	10,980	35
Malaysia earthquake	6/6/2015 5:00	7/5/2015 23:00	63,569	341
Paris attacks	11/14/2015 12:00	12/14/2015 6:00	50,776	3,154
Texas storms	5/25/2015 20:00	6/24/2015 14:00	266,633	212
Chile earthquake	9/16/2015 23:00	10/16/2015 17:00	449,047	312
Nepal earthquake	4/25/2015 6:00	5/25/2015 0:00	9,921,280	3,433
Nepal earthquake	5/12/2015 10:00	6/11/2015 4:00	4,598,933	278
West Africa Ebola outbreak	3/29/2014 20:00	4/28/2014 14:00	777,278	236
Hurricane Patricia	10/23/2015 2:00	11/21/2015 20:00	1,240,651	591

Table 6. Study results showing death totals, *r* values, DTW distances, and OSM campaign status (death totals via corresponding Wikipedia articles).

Event	Start Date	End Date	# Deaths	OSM to Wikipedia Edits (*r* value)	DTW Distance	OSM Campaign
Pakistan earthquake	10/26/2015 2:00	11/24/2015 20:00	399	0.08597	1.06133	No
Alaska earthquake	1/23/2016 4:00	2/21/2016 22:00	0	0.08231	0.95523	No
Malaysia earthquake	6/6/2015 5:00	7/5/2015 23:00	18	0.30637	0.93872	No
Paris attacks	11/14/2015 12:00	12/14/2015 6:00	137	0.14996	0.94710	No
Texas storms	5/25/2015 20:00	6/24/2015 14:00	47	-0.06815	1.12450	No
Chile earthquake	9/16/2015 23:00	10/16/2015 17:00	14	0.26158	0.83236	Yes
Nepal earthquake	4/25/2015 6:00	5/25/2015 0:00	8,964	0.42750	0.59798	Yes
Nepal earthquake	5/12/2015 10:00	6/11/2015 4:00	218	0.34682	0.85619	Yes
West Africa Ebola outbreak	3/29/2014 20:00	4/28/2014 14:00	11,325	0.36350	0.66705	Yes
Hurricane Patricia	10/23/2015 2:00	11/21/2015 20:00	13	0.72015	0.51862	Yes

6.2 Pakistan Earthquake

The 7.5 magnitude earthquake which impacted Pakistan and Afghanistan occurred on October 26, 2015, is known as the 2015 Hindu Kush Earthquake (USGS, 2015). This did not produce a mapping campaign for Pakistan, which was chosen as the OpenStreetMap data source for this study.

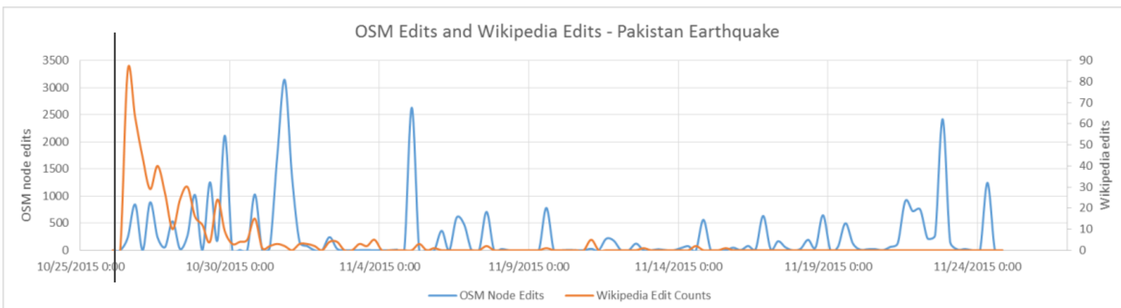

Figure 12. Chart showing OpenStreetMap and Wikipedia edits for the Pakistan earthquake.

Figure 12[9] shows the revision data for the thirty day period following the creation of the Wikipedia event page[10] for this earthquake. The left axis, represented by the blue dataset, shows the number of OpenStreetMap nodes that were added, modified, or deleted in this time period. The right axis, represented by the orange dataset, shows the number of times the Wikipedia event page was modified during this time period. For this event, there were 494 Wikipedia pages edits resulting in 33,734 bytes of data changed and 31,900 OpenStreetMap nodes revised.

[9] In this graph and others like it, the vertical line denotes the time of creation for the event's Wikipedia article

[10] https://en.wikipedia.org/wiki/October_2015_Hindu_Kush_earthquake

The calculated *r* values for this case study proved to be some of the lowest in this thesis. The *r* value for comparing Wikipedia edit counts to OpenStreetMap node edits was 0.08597 and the result for comparing Wikipedia bytes changed to OpenStreetMap node edits was 0.00518. These values are extremely close to zero and show little to no correlation between any of the datasets.

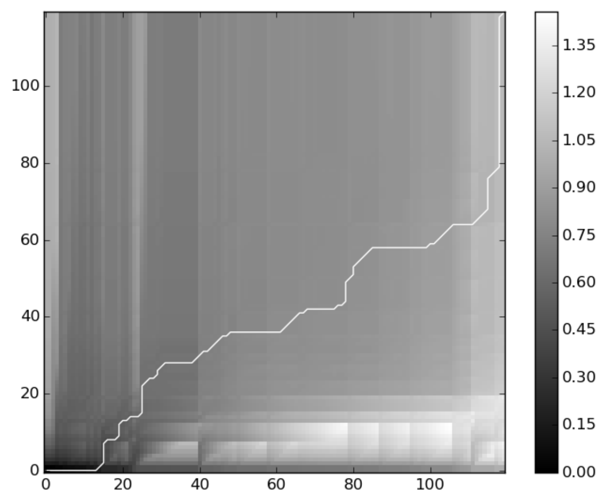

Figure 13. DTW distance graph of Pakistan OpenStreetMap to Wikipedia edits.

Figure 13 shows the dynamic time warping graph between Wikipedia page edits and OpenStreetMap node edits. The distance of 1.06133 was one of the highest generated by this study, showing a relatively long path. This can also be visually confirmed by the lack of a straight line between the bottom-left and top-right corners of the graph in Figure 13.

6.3 Alaska, United States Earthquake

The most recent event that this study analyzed was the earthquake that affected Alaska, United States, on January 24, 2016. This earthquake, known as the 2016 Old Iliamna Earthquake, was centered approximately 163 miles from Anchorage, the capital of Alaska and was recorded as a magnitude 7.1 event (USGS, 2015). This event did not produce a humanitarian mapping campaign.

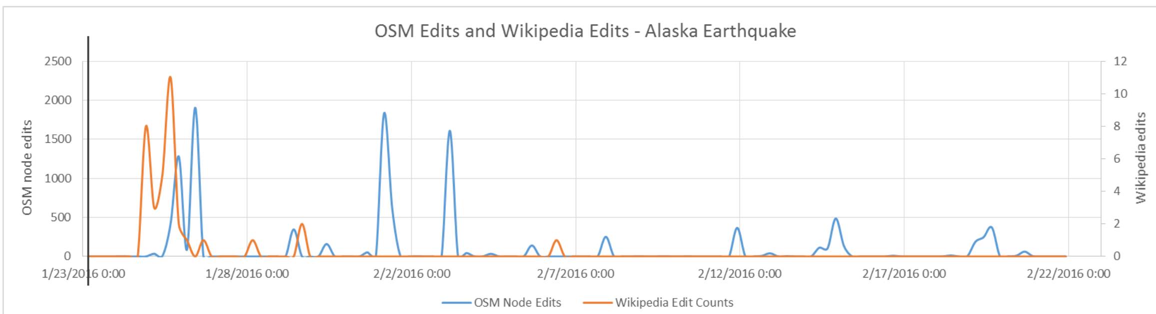

Figure 14. Chart showing OpenStreetMap and Wikipedia edits for the Alaska earthquake.

Figure 14 shows the thirty day period starting at the creation of the Wikipedia event page[11]. The blue dataset, which is shown on the left axis, shows the number of OpenStreetMap nodes that were added, modified, or deleted over this time period. The orange dataset, shown on the right axis, shows the number of times the Wikipedia page for the event was modified. Over this time period, there were 35 revisions to the Wikipedia page which resulted in 6,557 bytes of data being changed, and 10,980 OpenStreetMap nodes modified in the state of Alaska.

[11] https://en.wikipedia.org/wiki/2016_Old_Iliamna_earthquake

Bringing the datasets together, the resulting Pearson correlation coefficient, or *r* value, between Wikipedia revision count and OpenStreetMap node edits was 0.08231 while the *r* value between Wikipedia revision bytes changed and OpenStreetMap node edits was 0.01213. These values, which are very close to zero, indicate that there is little to no correlation between the datasets.

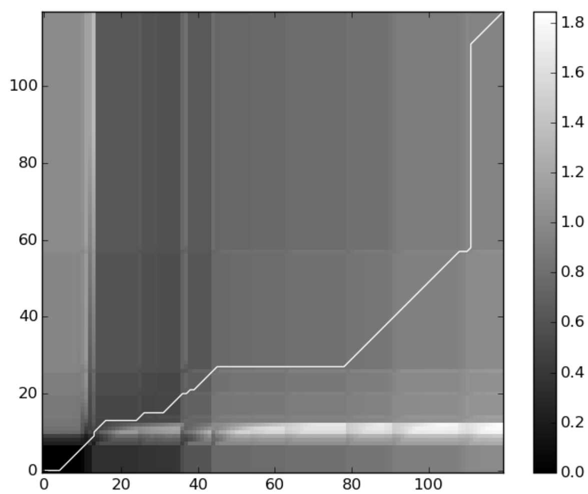

Figure 15. DTW distance graph of Alaska OpenStreetMap to Wikipedia edits.

Figure 15 shows the resulting dynamic time warping graph between the Wikipedia edit counts and the OpenStreetMap node edit counts as a distance of 0.99255. The lack of a more direct path from bottom-left to upper-right shows that the distance is not ideal and that the similarities between the datasets is not strong.

6.4 Malaysia Earthquake

The 2015 Sabah Earthquake, which occurred on June 5, 2015 in Sabah, Malaysia, was recorded as a 6.0 magnitude event (USGS, 2015). This event did not produce a humanitarian mapping effort from the Humanitarian OpenStreetMap team.

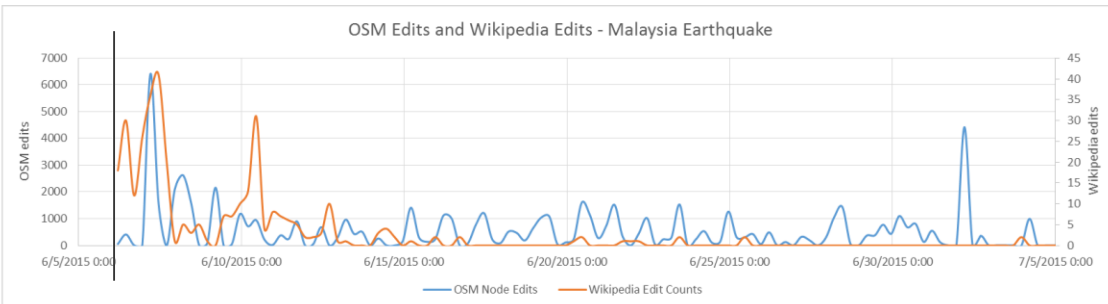

Figure 16. Chart showing OpenStreetMap and Wikipedia edits for the Malaysia earthquake.

Figure 16 shows the revision data during the thirty day period starting at the Wikipedia event page[12] creation time. The left axis, represented by the blue dataset, shows the number of OpenStreetMap nodes that were added, modified, or deleted during the time period. The right axis, represented by the orange dataset, shows the number of times the Wikipedia event page was modified during this time. There were 341 page revisions resulting in 48,889 bytes of data changed[13] for this study as well as 63,569 nodes that were added, modified, or deleted in the OpenStreetMap dataset.

[12] https://en.wikipedia.org/wiki/2015_Sabah_earthquake
[13] Two large revisions were removed from the dataset. The first was a vandalism revision that deleted the entire page and the second was the restoration of the page. These large revisions skewed the Wikipedia revision size data.

The r values generated for this dataset were 0.30637 for page edits and OpenStreetMap edits and 0.24542 when comparing page edit size to OpenStreetMap edits. These results are higher than previous studies, showing a moderate positive correlation between the datasets.

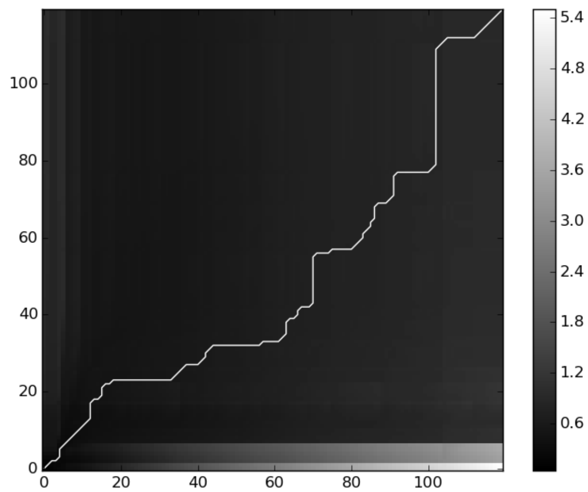

Figure 17. DTW distance graph of Malaysia OpenStreetMap to Wikipedia edits.

Figure 17 shows a dynamic time warping graph between the Wikipedia edit counts dataset and the OpenStreetMap node edits dataset. The DTW distance that was calculated for this dataset is 0.93872 and the graph implies a somewhat straight path between the bottom-left and top-right corners, implying somewhat similar datasets.

6.5 Paris, France Attacks

The 2015 Paris attacks, which occurred on November 15, 2015 in Paris, France, was a very high profile attack that was monitored worldwide. This event did not produce a humanitarian mapping effort from the Humanitarian OpenStreetMap team.

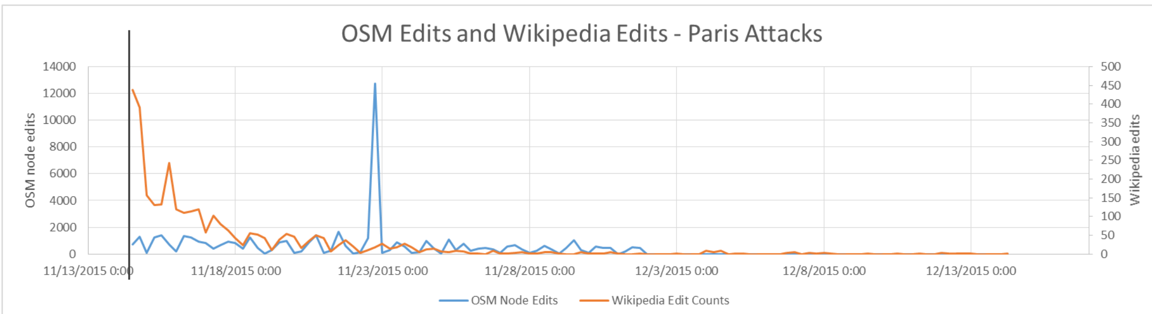

Figure 18. Chart showing OpenStreetMap and Wikipedia edits for the Paris attacks.

Figure 18 shows the revision data during the thirty day period starting at the Wikipedia event page[14] creation time. The left axis, represented by the blue dataset, shows the number of OpenStreetMap nodes that were added, modified, or deleted during the time period. The right axis, represented by the orange dataset, shows the number of times the Wikipedia event page was modified during this time. There were 3,154 page revisions resulting in 2,156,182 bytes of data changed for this study as well as 50,776 nodes that were added, modified, or deleted in the OpenStreetMap dataset.

[14] https://en.wikipedia.org/wiki/November_2015_Paris_attacks

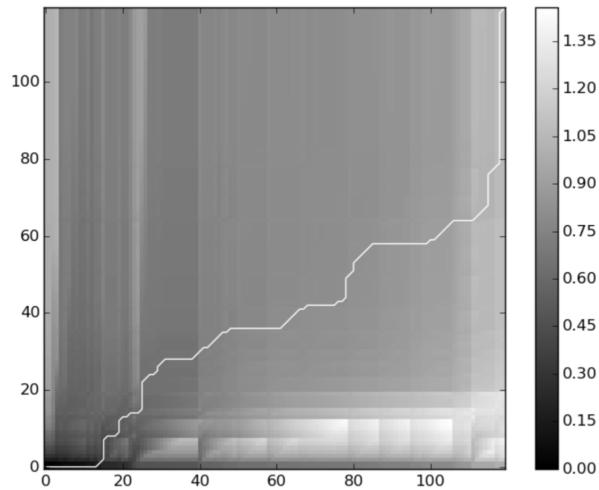

Figure 19. DTW distance graph of Paris OpenStreetMap to Wikipedia edits.

Figure 19 shows a dynamic time warping graph between the Wikipedia edit counts dataset and the OpenStreetMap node edits dataset. The DTW distance that was calculated for this dataset is 0.94710 and the graph implies a slightly straight path between the bottom-left and top-right corners, implying somewhat similar datasets.

6.6 Texas, United States Storms

The 2015 Texas storms and floods, which occurred in May and June of 2015 in Texas, United States, was a very long and devastating set of storms that produced tornados and record-breaking floods. This event did not produce a humanitarian mapping effort from the Humanitarian OpenStreetMap team.

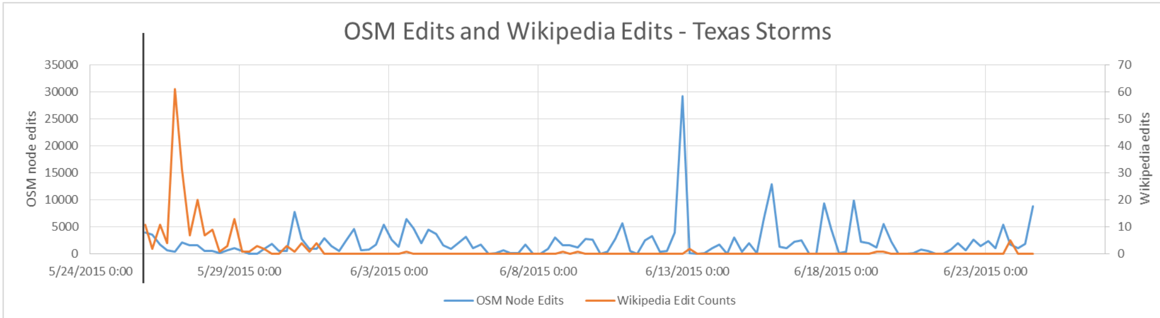

Figure 20. Chart showing OpenStreetMap and Wikipedia edits for the Texas storms.

Figure 20 shows the revision data during the thirty day period starting at the Wikipedia event page[15] creation time. The left axis, represented by the blue dataset, shows the number of OpenStreetMap nodes that were added, modified, or deleted during the time period. The right axis, represented by the orange dataset, shows the number of times the Wikipedia event page was modified during this time. There were 212 page revisions resulting in 59,925 bytes of data changed for this study as well as 266,633 nodes that were added, modified, or deleted in the OpenStreetMap dataset.

[15] https://en.wikipedia.org/wiki/2015_Texas-Oklahoma_flood_and_tornado_outbreak

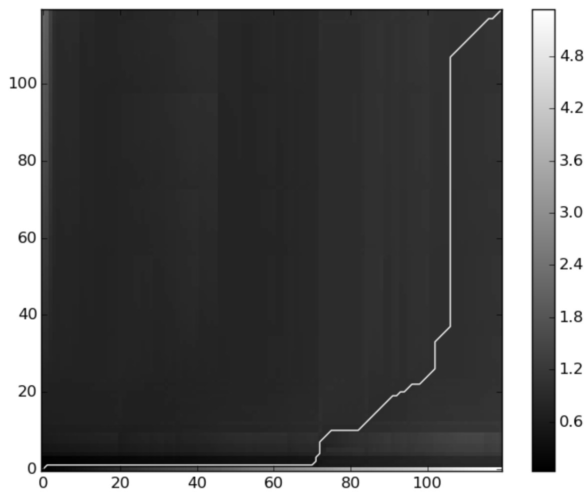

Figure 21. DTW distance graph of Texas OpenStreetMap to Wikipedia edits.

Figure 21 shows a dynamic time warping graph between the Wikipedia edit counts dataset and the OpenStreetMap node edits dataset. The DTW distance that was calculated for this dataset is 0.94710 and the graph implies a line that is far from a straight line between the bottom left and upper right corners.

6.7 Chile Earthquake

The earthquake that occurred in Chile on September 16, 2015, known as the 2015 Illapel Earthquake, was measured as an 8.3 magnitude event. The earthquake occurred 29 miles from the coast of Chile, near Illapel (USGS, 2015). This event had a small but successful humanitarian mapping campaign (“Humanitarian OpenStreetMap Team”, 2016).

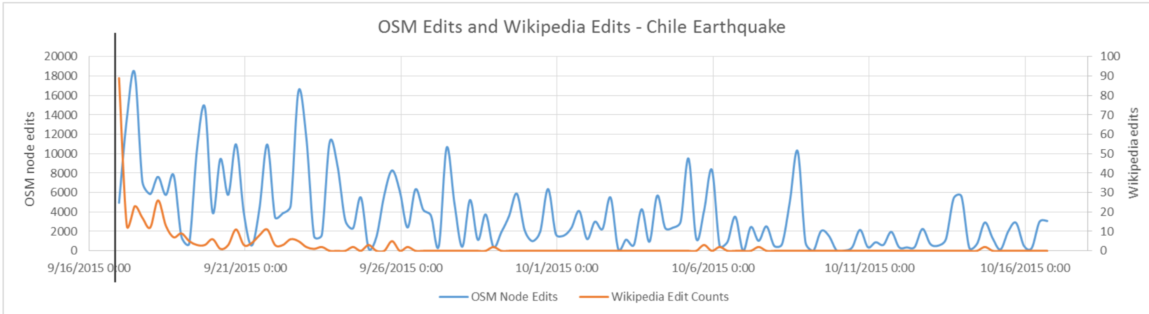

Figure 22. Chart showing OpenStreetMap and Wikipedia edits for the Chile earthquake.

Figure 22 shows the thirty day period starting at the Wikipedia event page[16] creation time. The blue dataset, which is shown on the left axis, shows the number of OpenStreetMap nodes that were added, modified, or deleted over this time period. The orange dataset, shown on the right axis, shows the number of times the Wikipedia page for the event was modified. Over this time period, there were 312 revisions to the Wikipedia page which resulted in 123,525 bytes of data being changed, and 449,047 OpenStreetMap nodes modified in the country of Chile.

Comparing the datasets generated an *r* value of 0.26158 for Wikipedia edit counts and OpenStreetMap node counts. The *r* value for Wikipedia bytes changed and OpenStreetMap node counts was 0.35194. These *r* values, which are noticeably higher than zero but still closer to zero than they are to one, show a somewhat significant positive correlation between the datasets.

[16] https://en.wikipedia.org/wiki/2015_Illapel_earthquake

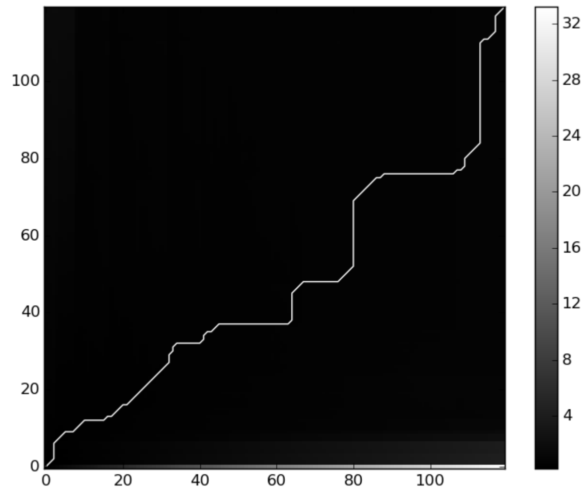

Figure 23. DTW distance graph of Chile OpenStreetMap to Wikipedia edits.

Figure 23 shows a dynamic time warping graph between the Wikipedia edit counts dataset and the OpenStreetMap node edits dataset. The DTW distance that was calculated for this dataset is 0.83236 and the graph implies a somewhat straight path between the bottom-left and top-right corners, implying somewhat similar datasets.

6.8 Nepal Earthquakes

The earthquake and subsequent aftershocks that occurred in April and May, 2015, in Nepal were the largest, highest impact earthquakes used in this study. The May aftershocks were so large that a Wikipedia page was generated for it, as well. This case study will look at both the 7.8 magnitude earthquake that occurred on April 25, 2015 and the 7.3 magnitude aftershock that occurred on May 12, 2015 in Nepal. These events generated a large humanitarian mapping effort from the HOT ("Humanitarian OpenStreetMap Team", 2016).

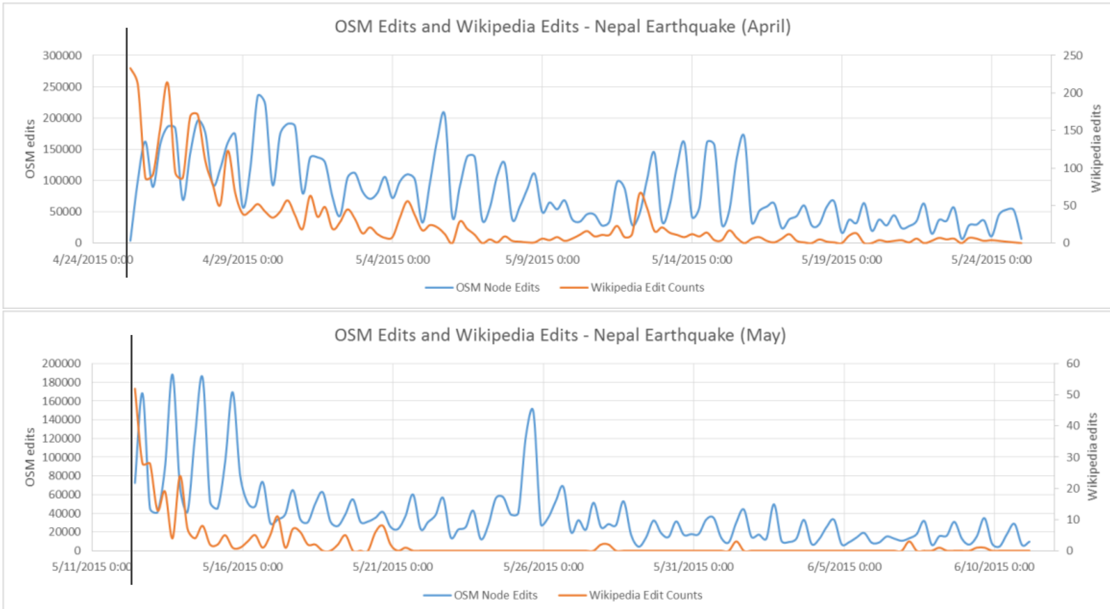

Figure 24. Chart showing OpenStreetMap and Wikipedia edits for the Nepal earthquakes.

Figure 24 shows the revision graphs for OpenStreetMap and Wikipedia page edits for the thirty day periods starting at the Wikipedia event page[17,18] creation time. The left axis, represented by the blue dataset, shows the number of OpenStreetMap nodes that were added, modified, or deleted during the time period. The right axis, represented by the orange dataset, shows the number of times the Wikipedia event page was modified during this time. For the April study, there were 3,433 Wikipedia edits resulting in 1,447,821 bytes of data being changed and 9,921,280 OpenStreetMap nodes modified. For the May case, there were 278 Wikipedia edits resulting in 45,430 bytes of data being changed and 4,598,933. By all accounts, the April event generated much more volunteered information than the May event.

[17] https://en.wikipedia.org/wiki/April_2015_Nepal_earthquake
[18] https://en.wikipedia.org/wiki/May_2015_Nepal_earthquake

The *r* values of the April datasets for comparing OpenStreetMap revisions to Wikipedia edits and number of bytes changed were 0.42750 and 0.14180, respectively. The May *r* values for comparing OpenStreetMap revisions to Wikipedia edits and number of bytes changed were 0.34682 and 0.15352, respectively. These results, specifically when comparing OpenStreetMap edits to Wikipedia edits, represent the highest correlation of any earthquake chosen in this study. These results also show a large discrepancy between the *r* values for each study when changing from Wikipedia edit counts to number of bytes changed.

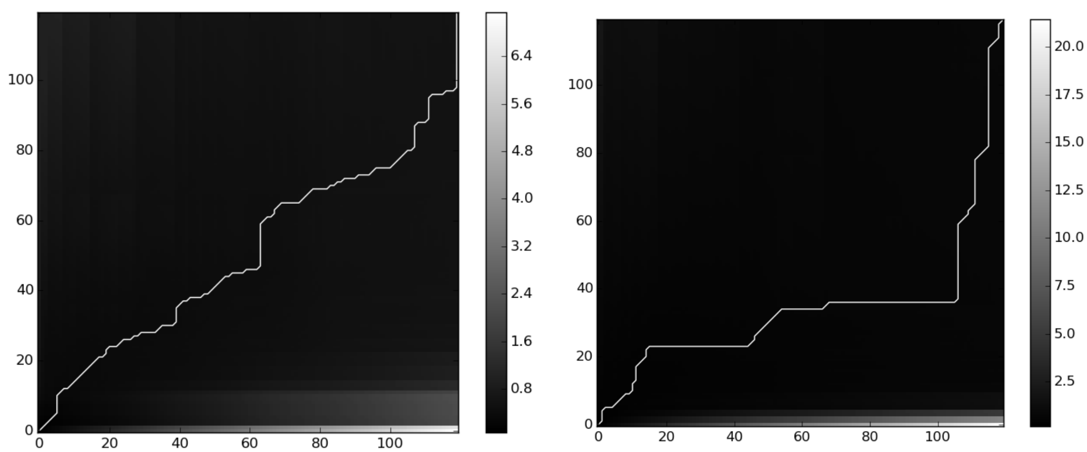

Figure 25. DTW distance graphs of April (left) and May (2015) Nepal OpenStreetMap to Wikipedia edits.

Figure 25 shows the dynamic time warping graphs for both the April (left) and May (right) datasets for comparing the OpenStreetMap edit data to the Wikipedia edit data. The April distance of 0.59798, like the *r* value, proved to be the strongest result for all earthquakes represented in this study. The May distance of 0.85619 showed much less similarity between the two datasets.

6.9 Africa – The 2014 West Africa Ebola Outbreak

Another non-earthquake event that was brought in to this study was the 2014 West Africa Ebola Outbreak. This event has a very large Wikipedia page[19] as well as a very large humanitarian mapping campaign being run by the Humanitarian OpenStreetMap team ("Humanitarian OpenStreetMap Team", 2016). OpenStreetMap data from both the Sierra Leone and Guinea was downloaded and processed for this study. Guinea was chosen as it is believed to be the starting point of the epidemic. Sierra Leone was chosen as it had the largest number of recorded cases.

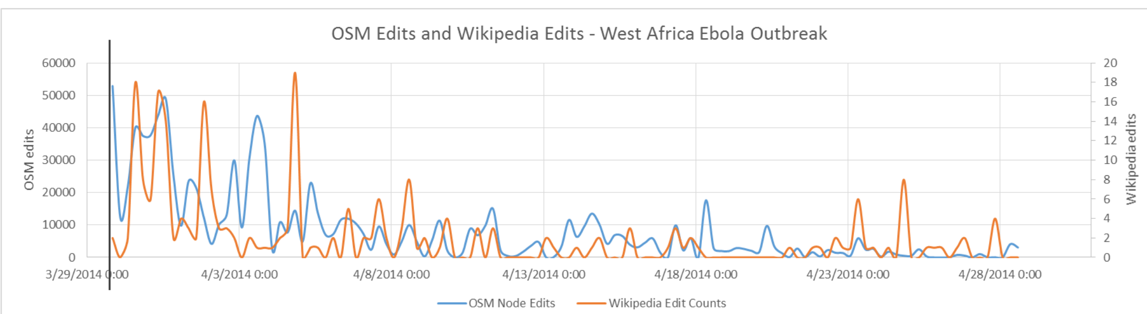

Figure 26. Chart showing OpenStreetMap and Wikipedia edits for the West Africa Ebola outbreak.

Figure 26 shows the thirty day period following the creation of the event page on Wikipedia. The left axis, represented by the blue dataset, shows the number of

[19] https://en.wikipedia.org/wiki/West_African_Ebola_virus_epidemic

OpenStreetMap nodes that were added, modified, or deleted while the right axis, represented by the orange dataset, shows the number of times the event page was revised on Wikipedia. For this event, there were 236 Wikipedia page edits resulting in 9,938 bytes changed and 1,003,374 nodes edited through OpenStreetMap.

The *r* values for this study show quite a strong positive correlation. The *r* value for OpenStreetMap nodes edited and Wikipedia edits is 0.49662 while the value for OpenStreetMap nodes edited and Wikipedia bytes changed is 0.34379. Both of these values show a positive correlation, but the *r* value of near 0.5 for Wikipedia edit counts shows one of the strongest positive correlations in this study.

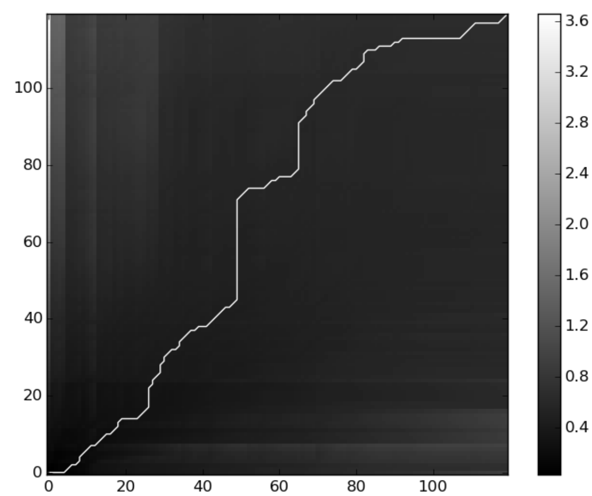

Figure 27. DTW distance graph of Guinea and Sierra Leone OpenStreetMap to Wikipedia edits.

Figure 27 shows the dynamic time warping graph between Wikipedia page edits and OpenStreetMap node edits. The distance of 0.65806 and the visually straight line between the lower-left and upper-right corners of the graph show a similar dataset relative to some of the others done in this study.

6.10 Mexico – Hurricane Patricia

The final non-earthquake event chosen for this study was Hurricane Patricia, which struck the western coast of Mexico in October of 2015 as a category 5 hurricane. This event, much like the West Africa Ebola Outbreak, has both a large Wikipedia page[20] and a humanitarian mapping campaign through the HOT ("Humanitarian OpenStreetMap Team", 2016).

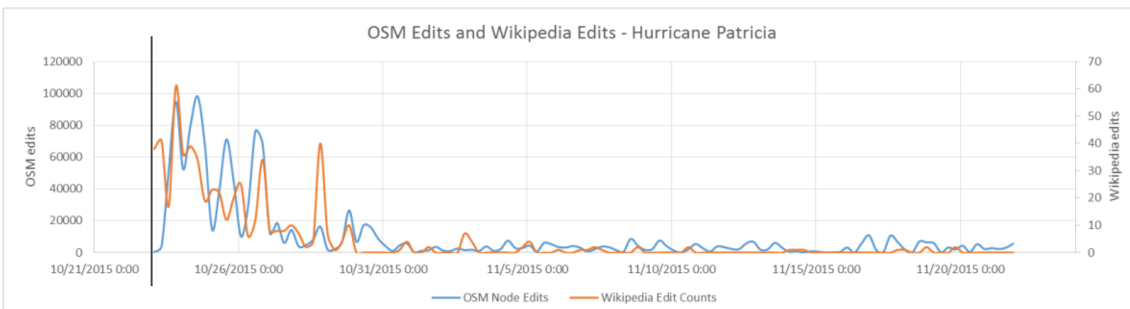

Figure 28. Chart showing OpenStreetMap and Wikipedia edits for Hurricane Patricia.

Figure 28 shows the thirty day period following the creation of the event page on Wikipedia. The left axis, represented by the blue dataset, shows the number of OpenStreetMap nodes that were added, modified, or deleted while the right axis, represented by the orange dataset, shows the number of times the event page was revised

[20] https://en.wikipedia.org/wiki/Hurricane_Patricia

on Wikipedia. For this event, there were 591 Wikipedia page edits resulting in 521,729 bytes changed and 1,240,651 nodes edited through OpenStreetMap.

The *r* values calculated for the Hurricane Patricia case proved to be the highest correlation found in this study. The *r* value for OpenStreetMap node edits to Wikipedia page counts was 0.72015 while the value for OpenStreetMap node edits to Wikipedia bytes changed was 0.35500, both the highest in their respective categories.

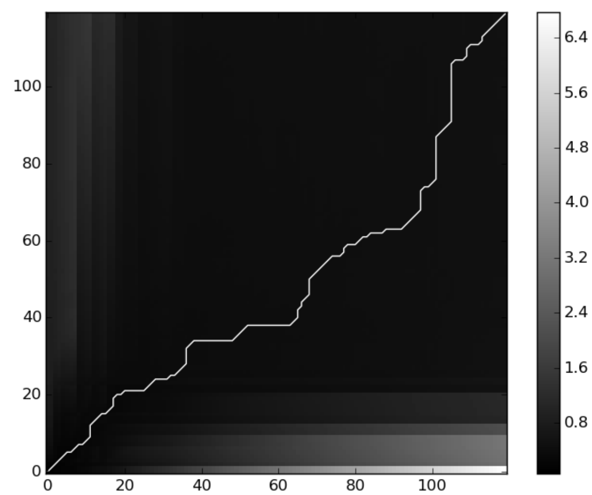

Figure 29. DTW distance graph of Mexico OpenStreetMap to Wikipedia edits.

Figure 29 shows the dynamic time warping graph for the Hurricane Patricia dataset. The calculated distance of 0.51862 was the lowest found in the entire study and can also be visually verified by the relatively straight path between the lower-left and upper-right corners of the graph in Figure 29.

7. DISCUSSION OF RESULTS AND CONCLUSION

This final chapter will provide an overall summary of the findings of this research as well as a discussion of potential implications of the results. Section 7.1 will summarize and discuss the overall results of the study. Section 7.2 will outline areas for possible future research. Section 7.3 will have concluding comments for the study.

7.1 Discussion of Results

The primary focus of this study was to compare the well-known structures and patterns of Wikipedia communities, specifically looking at user contribution patterns, and compare those results to OpenStreetMap communities and their contribution patterns during and after times of crisis. The main types of events that were chosen for this study were earthquakes, but to include more data points, other events were also considered. The events that were chosen had to meet a specific criteria to be included. This is due to the data and analysis requirements that are necessary to conduct the research. Events chosen must have a Wikipedia article generated (a user must choose to initially create an article about an event) and must also have available OpenStreetMap data. Once the events were identified, the data was gathered and processed for analysis.

When examining the contribution patterns of the Wikipedia articles, the events in this study typically followed Garter's Hype Cycle, as described by Linden and Fenn (2003) and reviewed in the chapter 2. This means that the articles saw a large spike in activity immediately after the event and the article were created. This spike then quickly decreased to near-zero, then leveled off into a slow and steady pattern of activity. This finding was consistent in all Wikipedia articles that were analyzed in this study.

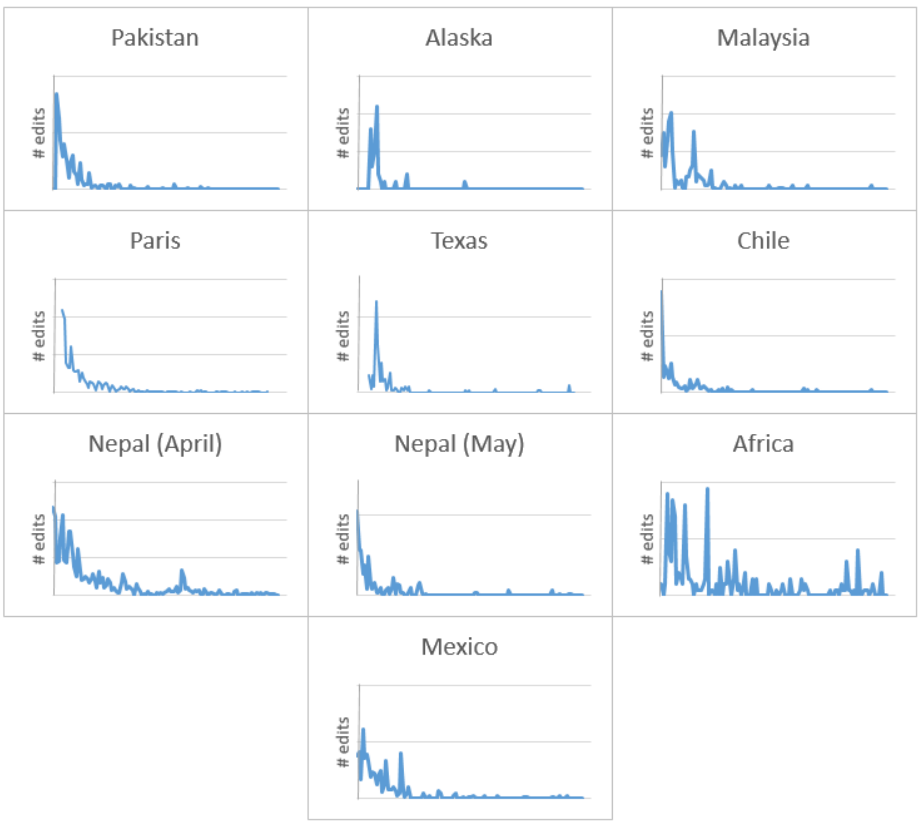

Figure 30. Graphs showing Wikipedia article edits counts for thirty day periods.

Figure 30 shows line graphs for all of the Wikipedia articles that were analyzed for this study. These graphs are thirty day snapshots from article creation. A clear pattern can be seen from looking at these graphs. The hype cycle, or some close variation of it, is clearly visible in these datasets. There is a large amount of activity at the beginning of the articles life cycle, the largest it will ever see, followed by a slow and steady decline in activity until the maintenance levels are reached. These maintenance levels are then maintained for an indefinite period of time. This pattern is similar to the consistently declining activity levels found by Arazy and Croitoru (2010). The outlier within this set is the West Africa Ebola Outbreak Wikipedia article (labeled ‘Africa’ in Figure 30). This particular article sees a slightly increased amount of activity as time goes on relative to the other pages in this study. This can likely be attributed to the fact that this article is much more of an ongoing event than the others. Earthquakes and hurricanes happen in a relatively short amount of time and can then be documented. The Ebola outbreak, however, was considered to be active from December 2013 until January 2016. It is interesting to note that even though this event lasted over two years, the initial thirty day period still resembled the hype cycle and still resembled the other articles in this study.

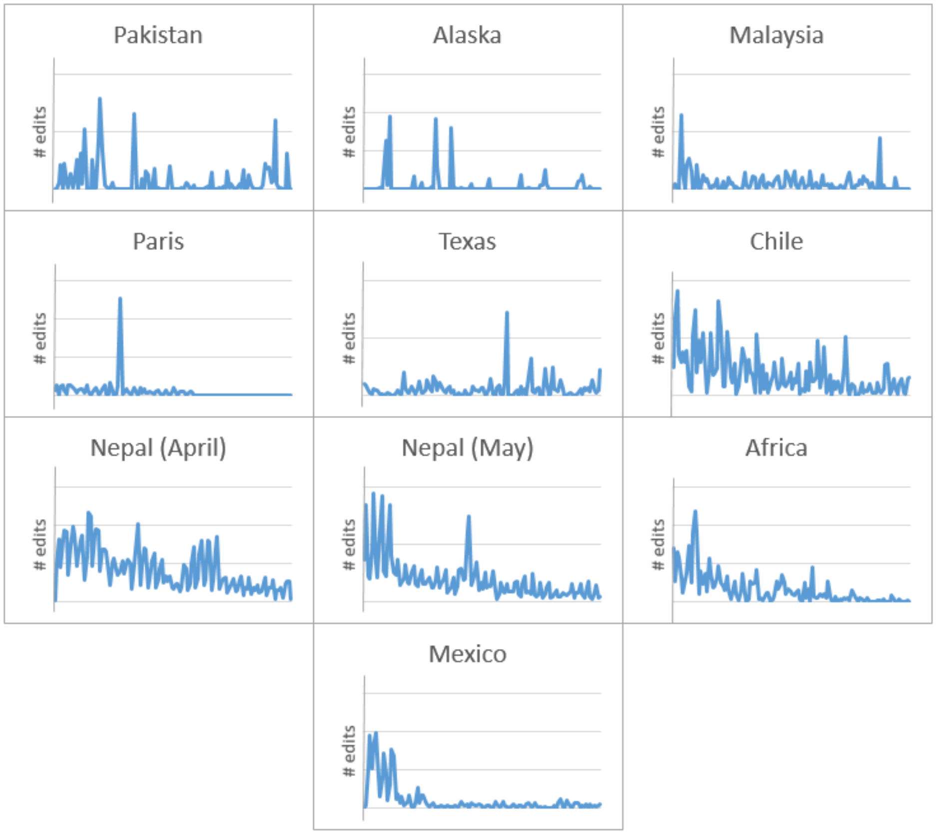

Figure 31. Graphs showing OpenStreetMap node edits counts for thirty day periods.

The contribution patterns of OpenStreetMap, which can be seen in Figure 31, are much less known and have had considerably less research done around them. As discussed in the literature review chapter, much of the OpenStreetMap research has focused on the quality of the user contributed data, rather than the patterns of those contributions. In recent years, there have been some emerging studies, such as the one conducted by Camponovo and Freundschuh (2014), that looked primarily at the response of OpenStreetMap communities to large scale crisis events. These responses are what eventually formed the HOT. The HOT, to summarize from chapter 2, is a nonprofit organization that organizes mapping contribution efforts after natural disasters and other crisis events. The focus of this portion of the study was to look at both events that were managed by the HOT and those that were not and see how they compare to their corresponding Wikipedia article's contributions. This study found that there was a strong correlation between the contribution patterns of Wikipedia articles and OpenStreetMap datasets only when the HOT managed a campaign. In all cases, this study observed both higher correlation (higher values show stronger correlation) and lower dynamic time warping minimum distance warping path (lower values show more similarity) when the event had a corresponding HOT campaign.

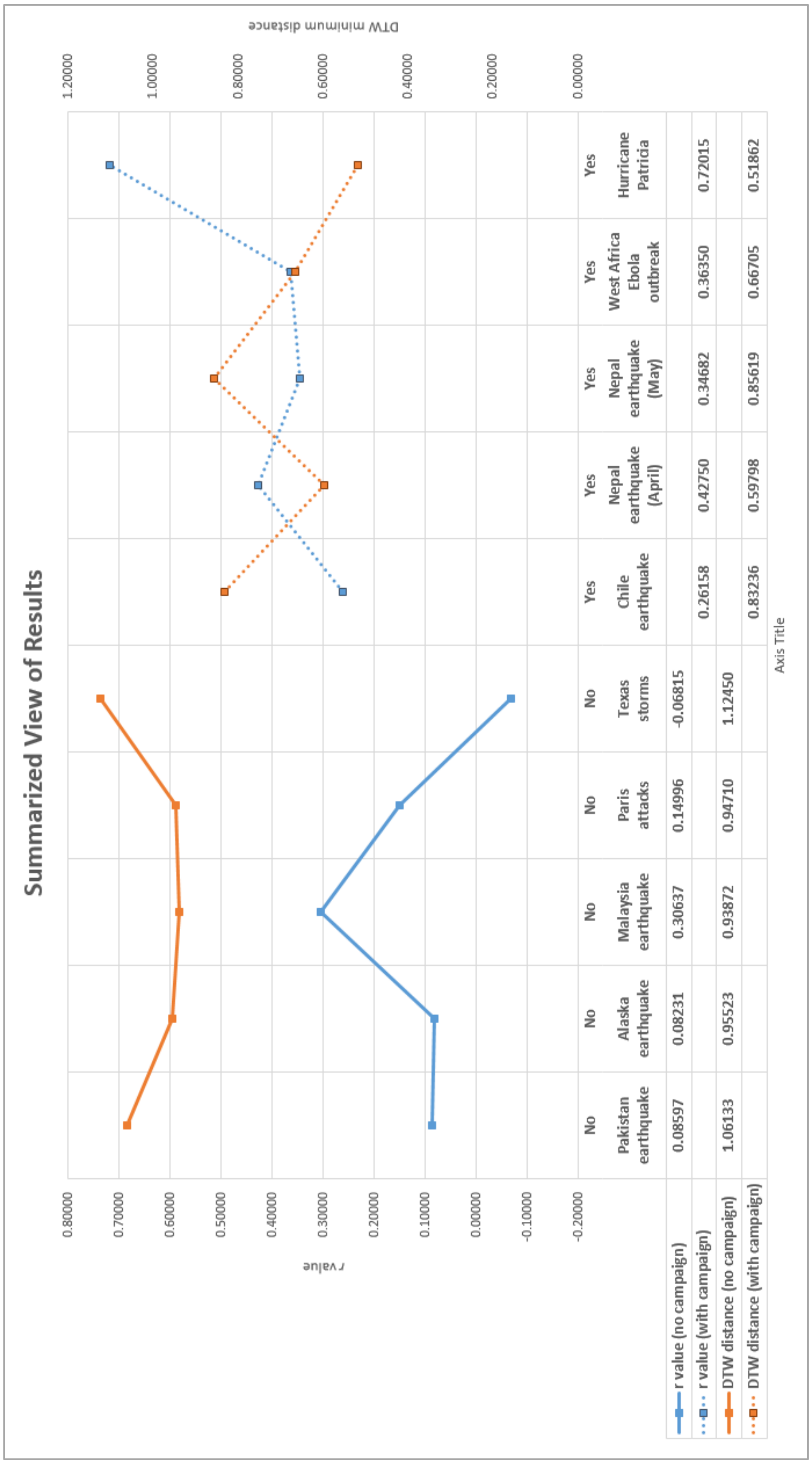

Figure 32. Summarized view of results.

Figure 32 shows a summarized view of the results of this study. The events with the lowest correlation between Wikipedia article edits and OpenStreetMap edits were the Pakistan and Alaska events, neither of which had a corresponding HOT campaign. The Malaysia earthquake, while having a higher correlation than the other non-campaign events, still had a relatively high DTW distance, showing less similarity than the other events. The Chile earthquake had a small HOT campaign which produced a higher correlation (though slightly lower than the Malaysia earthquake), but also produced a lower DTW distance. The Malaysia and Chile earthquakes produced what seem to be outliers relative to the other events that have or don't have HOT campaigns. The Malaysia earthquake produced the highest correlation for all non-HOT events while the Chile earthquake produces the smallest correlation for all HOT events. The April Nepal earthquake, the West African Ebola outbreak, and Hurricane Patricia all produced both the highest correlation and the lowest DTW distances. These events had large HOT campaigns created for them. The May Nepal earthquake, which was technically an aftershock of the April earthquake but was still large enough to warrant a Wikipedia article, also showed strong correlation and lower DTW distance, however the results were not quite as strong as the other HOT campaign events.

Table 7. Summarized view of average correlation and DTW distance for HOT and non-HOT events.

HOT Campaign	**Average *r* value**	**Average DTW distance**
Yes	0.42391	0.69444
No	0.11129	1.00538

The results of this study for the events that were analyzed suggest that organized mapping campaigns (such as those created by the HOT) have a large impact on the contribution patterns to OpenStreetMap. Without these campaigns, the contribution patterns do not seem to differentiate themselves from the contribution patterns of their corresponding Wikipedia articles. When these campaigns exist, patterns emerge from the OpenStreetMap data that are much more similar to the contribution patterns of the corresponding Wikipedia articles. Table 7 shows that the average *r* value for events with a corresponding HOT campaign is 0.45053 while the average for events without a corresponding HOT campaign is 0.11129. Similarly, the average DTW distance for events with HOT campaigns is 0.69264 while the average distance for events without HOT campaigns is 1.00538.

Wikipedia boasts a strong community with a solid structure that enables quality content creation (“Wikipedia: Contributing to Wikipedia”, 2016). OpenStreetMap, on the other hand, while having a strong sense of community and volunteering, does not quite have the structure or protocols that are seen on Wikipedia (“Contribute Map Data - OpenStreetMap Wiki”, 2016). This difference may be evident in the findings of this study that the contribution patterns in OpenStreetMap for those events that have organized campaigns tend to correlate stronger to their corresponding Wikipedia articles than those events that do not have organized campaigns. Using the Wikipedia communities as a baseline, the HOT creates a stronger community with their campaigns. Also, considering the validation tasks that are included in HOT campaigns, the data produced by these campaigns may be of higher quality and more in line with the data in Wikipedia articles due to the verification efforts present in that community.

In looking at these results in relation to the research question of using Wikipedia as a type of 'predictor' of OpenStreetMap data, it is clear that OpenStreetMap contribution patterns for events with HOT campaigns show much stronger correlation to Wikipedia contributions than those that do not have a corresponding HOT mapping campaign. The next section will outline some key areas where future research could help further define these predictors.

7.2 Areas of Future Research

There are several areas of potential future research for this effort. To start, more events could be studied and compared to further prove (or disprove) the correlation between HOT campaigns and Wikipedia article contribution patterns. The HOT website has many archived campaigns that can be analyzed in relation to their Wikipedia pages.

In addition to simply studying more events, other social media platforms can be analyzed in relation to OpenStreetMap contribution patterns. Social media platforms such as Twitter, Facebook, Instagram, and Flickr can be studied to determine if people are discussing, photographing, and sharing information about these events in a similar pattern to the users that are contributing map data to OpenStreetMap.

Another potential extension to this study could incorporate the amount of damage incurred by these events. In many cases, estimates of cost are associated with the damage of these events which could then be incorporated into the study to determine a potential impact to contribution patterns.

A further extension to this study could be to look specifically at the quality of the mapping data both for areas that are served by HOT campaigns and those that are not. This could argue further that these HOT campaigns produce a stronger community and also provide a higher quality dataset than those that do not have these campaigns.

Finally, an extension of this study could look at unique users and their contribution patterns to these platforms. While user data was available for this thesis, it was considered out of scope for the research questions that were presented. This idea can also be expanded to look at which demographics contribute to these platforms. Gender, education, and other demographic datasets can be looked at in conjunction with contribution data. A final expansion of studying user-specific contribution patterns would be to further understand the motivation behind the users' contributions.

7.3 Conclusion

This study has looked at the extensive research that has been conducted on the social networks of Wikipedia and their contribution patterns and compared them to the contribution patterns of users of OpenStreetMap. As a result, stronger correlations were found between Wikipedia articles contribution patterns and the contribution patterns of OpenStreetMap areas when there is an organized mapping campaign created by the Humanitarian OpenStreetMap team. As such this thesis paves a way to explore the relationship between volunteered information on Wikipedia and volunteered geographic information (i.e. map data) on the OpenStreetMap platform.

REFERENCES

"About Wikipedia." Accessed November 22, 2015. https://en.wikipedia.org/wiki/Wikipedia:About.

"A Disaster Map 'Wiki' Is Born." Accessed February 7, 2016. http://archive.wired.com/software/coolapps/news/2005/09/68743?currentPage=all

"About Humanitarian OpenStreetMap Team." Accessed March 12, 2016. https://hotosm.org/about.

Arazy, Ofer, and Arie Croitoru. "The Sustainability of Corporate Wikis: A Time-Series Analysis of Activity Patterns." ACM Transactions on Management Information Systems 1, no. 1 (December 1, 2010): 1–24. doi:10.1145/1877725.1877731.

Budhathoki, Nama R., and Caroline Haythornthwaite. "Motivation for Open Collaboration Crowd and Community Models and the Case of OpenStreetMap." American Behavioral Scientist 57, no. 5 (2013): 548–75.

Buriol, Luciana S., Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. "Temporal Analysis of the Wikigraph." In Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on, 45–51. IEEE, 2006. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4061340.

Camponovo, Michael E., and Scott M. Freundschuh. "Assessing Uncertainty in VGI for Emergency Response." Cartography and Geographic Information Science 41, no. 5 (October 20, 2014): 440–55. doi:10.1080/15230406.2014.950332.

"Contribute Map Data." Accessed March 24, 2016. http://wiki.openstreetmap.org/wiki/Contribute_map_data.

"Earthquakes." Accessed November 22, 2015. http://earthquake.usgs.gov/earthquakes/.

Di Lena, Pietro, and Luciano Margara. "Optimal Global Alignment of Signals by Maximization of Pearson Correlation." Information Processing Letters 110, no. 16 (July 2010): 679–86. doi:10.1016/j.ipl.2010.05.024.

Dowdy, S. M., and Stanley Wearden. Statistics for Research. Wiley Series in Probability and Mathematical Statistics. New York: Wiley, 1983.

“Dynamic Time Warping (DTW).” Accessed March 12, 2016. http://mlpy.sourceforge.net/docs/3.4/dtw.html.

Forghani, Mohammad, and Mahmoud Delavar. “A Quality Study of the OpenStreetMap Dataset for Tehran.” ISPRS International Journal of Geo-Information 3, no. 2 (May 22, 2014): 750–63. doi:10.3390/ijgi3020750.

Girres, Jean-François, and Guillaume Touya. “Quality Assessment of the French OpenStreetMap Dataset: Quality Assessment of the French OpenStreetMap Dataset.” Transactions in GIS 14, no. 4 (August 2010): 435–59. doi:10.1111/j.1467-9671.2010.01203.x.

Goodchild, Michael F. “Citizens as Sensors: The World of Volunteered Geography.” GeoJournal 69, no. 4 (November 30, 2007): 211–21. doi:10.1007/s10708-007-9111-y.

Goodchild, Michael F., and J. Alan Glennon. “Crowdsourcing Geographic Information for Disaster Response: A Research Frontier.” International Journal of Digital Earth 3, no. 3 (September 2010): 231–41. doi:10.1080/17538941003759255.

Graña, M., M. Termenon, A. Savio, A. Gonzalez-Pinto, J. Echeveste, J. M. Pérez, and A. Besga. “Computer Aided Diagnosis System for Alzheimer Disease Using Brain Diffusion Tensor Imaging Features Selected by Pearson’s Correlation.” Neuroscience Letters 502, no. 3 (2011): 225–29.

Hargittai, Eszter, and Aaron Shaw. “Mind the Skills Gap: The Role of Internet Know-How and Gender in Differentiated Contributions to Wikipedia.” Information, Communication & Society 18, no. 4 (April 3, 2015): 424–42. doi:10.1080/1369118X.2014.957711.

Hudson-Smith, Andrew, Andrew Crooks, Maurizio Gibin, Richard Milton, and Michael Batty. “NeoGeography and Web 2.0: Concepts, Tools and Applications.” Journal of Location Based Services 3, no. 2 (June 2009): 118–45. doi:10.1080/17489720902950366.

“Humanitarian OpenStreetMap Team.” Accessed March 12, 2016. http://wiki.openstreetmap.org/wiki/Humanitarian_OSM_Team.

Jackson, Steven, William Mullen, Peggy Agouris, Andrew Crooks, Arie Croitoru, and Anthony Stefanidis. “Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information.” ISPRS International Journal of Geo-Information 2, no. 2 (June 4, 2013): 507–30. doi:10.3390/ijgi2020507.

Jokar Arsanjani, Jamal, Alexander Zipf, Peter Mooney, and Marco Helbich, eds. OpenStreetMap in GIScience. Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing, 2015. http://link.springer.com/10.1007/978-3-319-14280-7.

“JSON.” Accessed February 14, 2016. http://www.json.org/.

Kämpf, Mirko, Sebastian Tismer, Jan W. Kantelhardt, and Lev Muchnik. “Fluctuations in Wikipedia Access-Rate and Edit-Event Data.” Physica A: Statistical Mechanics and Its Applications 391, no. 23 (December 2012): 6101–11. doi:10.1016/j.physa.2012.07.004.

Keogh, Eamonn J., and Michael J. Pazzani. “Derivative Dynamic Time Warping.” In Sdm, 1:5–7. SIAM, 2001. http://epubs.siam.org/doi/abs/10.1137/1.9781611972719.1.

König, René. “WIKIPEDIA: Between Lay Participation and Elite Knowledge Representation.” Information, Communication & Society 16, no. 2 (March 2013): 160–77. doi:10.1080/1369118X.2012.734319.

Linden, Alexander, and Jackie Fenn. “Understanding Gartner’s Hype Cycles.” Strategic Analysis Report Nº R-20-1971. Gartner, Inc, 2003. http://www.ask-force.org/web/Discourse/Linden-HypeCycle-2003.pdf.

“MediaWiki Revisions API.” Accessed December 3, 2015. https://www.mediawiki.org/wiki/API:Revisions.

Mullen, William F., Steven P. Jackson, Arie Croitoru, Andrew Crooks, Anthony Stefanidis, and Peggy Agouris. “Assessing the Impact of Demographic Characteristics on Spatial Error in Volunteered Geographic Information Features.” GeoJournal 80, no. 4 (August 2015): 587–605. doi:10.1007/s10708-014-9564-8.

Neis, Pascal, Dennis Zielstra, and Alexander Zipf. “The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011.” Future Internet 4, no. 4 (December 29, 2011): 1–21. doi:10.3390/fi4010001.

Ortega, Felipe, and Jesus M. Gonzalez Barahona. "Quantitative Analysis of The Wikipedia Community of Users." In Proceedings of the 2007 International Symposium on Wikis, 75–86. ACM, 2007. http://dl.acm.org/citation.cfm?id=1296960.

"OpenStreetMap Copyright and License." Accessed November 22, 2015. https://www.openstreetmap.org/copyright.

"OpenStreetMap Data Extracts." Accessed November 22, 2015. http://download.geofabrik.de/index.html.

"OpenStreetMap Stats" Accessed November 22, 2015. http://wiki.openstreetmap.org/wiki/Stats#Contributor_statistics_reports.

"OsmChange" Accessed December 3, 2015. http://wiki.openstreetmap.org/wiki/OsmChange.

"PBF Software Compliance." Accessed February 14, 2016. http://wiki.openstreetmap.org/wiki/PBF/Software_Compliance.

Prasarnphanich, Pattarawan, and Christian Wagner. "Explaining the Sustainability of Digital Ecosystems Based on the Wiki Model Through Critical-Mass Theory." IEEE Transactions on Industrial Electronics 58, no. 6 (June 2011): 2065–72. doi:10.1109/TIE.2009.2027248.

"Protocol Buffers" Accessed February 14, 2016. https://developers.google.com/protocol-buffers/.

Salvador, Stan and Chan, Phillip. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. 3rd Wkshp. on Mining Temporal and Sequential Data, ACM KDD '04, 2004.

Shemak, April. "The Cartographic Dimensions of Humanitarianism: Mapping Refugee Spaces in Post-Earthquake Haiti." Cultural Dynamics, 2014, 0921374014527920.

Taylor, J. A., and T. R. Bates. "A Discussion on the Significance Associated with Pearson's Correlation in Precision Agriculture Studies." Precision Agriculture 14, no. 5 (October 2013): 558–64. doi:10.1007/s11119-013-9314-9.

Warren Liao, T. "Clustering of Time Series Data—a Survey." Pattern Recognition 38, no. 11 (November 2005): 1857–74. doi:10.1016/j.patcog.2005.01.025.

“Wikipedia: Contributing to Wikipedia.” Wikipedia, the Free Encyclopedia, March 17, 2016. https://en.wikipedia.org/w/index.php?title=Wikipedia:Contributing_to_Wikipedia&oldid=710494020.

Xu, Bo, and Dahui Li. “An Empirical Study of the Motivations for Content Contribution and Community Participation in Wikipedia.” Information & Management 52, no. 3 (April 2015): 275–86. doi:10.1016/j.im.2014.12.003.

Zook, Matthew, Mark Graham, Taylor Shelton, and Sean Gorman. “Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake.” World Medical & Health Policy 2, no. 2 (January 21, 2010): 6–32. doi:10.2202/1948-4682.1069.

BIOGRAPHY

Gary Esworthy graduated from Center High School, Monaca, Pennsylvania, in 2004. He received his Bachelor of Science from the Indiana University of Pennsylvania in 2008. He has been employed as a software engineer in Northern Virginia for seven years working in the aerospace engineering industry.