THE PRODUCTION OF VOLUNTEERED GEOGRAPHIC INFORMATION:

A STUDY OF OPENSTREETMAP IN THE UNITED STATES

by

David A. Parr, B.S., M.S.

A dissertation submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in Geographic Information Science
August 2015

Committee Members:

Yongmei Lu, Chair

Ronald Hagelman

T. Edwin Chow

David Mark

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

I would firstly like to thank and acknowledge the efforts of my advisor Yongmei Lu. I am extremely grateful for the time and effort she put into my work as a doctoral student. Yongmei made time to discuss methods, teaching, writing and editing, and to write many letters of recommendation and provide feedback. With her guidance, my writing improved tremendously while at Texas State. I appreciate her high standard of research and her work in general. Thank you.

Additionally, thanks are extended to my dissertation committee, Edwin Chow, Ron Hagelman, and David Mark, for support throughout my doctoral work that culminated in the production of this dissertation. Thank you all for your time, effort, advice, and continued encouragement.

I would like to thank the Graduate College at Texas State University for providing financial support (a doctoral research stipend award) for this dissertation. Additionally, I would like to thank the Department of Geography for support my efforts while at Texas State University. This includes help provided by faculty (Sven Fuhrmann, Oswaldo Muniz, Ben Zhan, Fred Day, Rich Earl, Alberto Giordano, Niem Huynh, Jennifer Jensen, and Phil Suckling), staff (Pat Jones, Allison Glass-Smith, and Charles Robinson) and colleagues (Adam Mathews, Kathleen Seal, Deborah Hanh, David Yelacic, Laura Cano-Amaya, Ryan Schuermann, Nathaniel Dede-Bamfo, Ruojing Scholz, Carmen Brysch, Lindsey McKnight, Yan Lin, Christi Townsend, Shae Luther, Matt Patton, Mathew

Connolly, David Nicosia, Sara Eaves, Anthony Irwin, Bill Adams, Melanie Stine, Michael Scholz, Lauren Maples, Keith Bremer, Marty Wamsley, Clayton Whitesides, Todd Moore, Shelley Burleson, Aja Davidson, James Dietrich, Jane Atha, Mark Dondero, Hans Friedel, Eric Samson, Joey Ostling, Philip Julian, Bernie Fang, Tamara Biegas – and anyone whose name has escaped this list). Thank you all for your friendship and support.

I would like to acknowledge my immediate support network: my family and close friends. Thanks to my father, Henry, for his continued patience and support. Thanks in no particular order to Brian and Patricia Borowicz and their family, Gabe and Rachael Dagani and their family, Anthony Morales, Kristen Carney, Larsson and Vicki Omberg and their family, Jerry and Yi Zhao and their family, Vladimir Rozniatovsky, Priya Ponnapalli, and Ana Roberts.

Lastly, I have to thank the person whose patience, support, love, and considerable riling has given me the endurance and strength to complete this project: my lovely wife Abigail. It would not have happened if not for you. Your support has been vital to my success, and I hope that I can support you in your graduate school time half as well as you have supported me. I love you, sweetie.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation	Description
API	Application Programing Interface
CBSA	Core-based statistical area
CGI	Contributed Geographic Information
CROS	California Roadkill Observation System
GI	Geographic Information
GIScience	Geographic Information Science
GPS	Global Positioning System
ICT	Information Communications Technology
KDD	Knowledge Discovery in Database
OGC	Open GeoSpatial Consortium
OSM	OpenStreetMap
PCA	Principal Components Analysis
PCR	Principal Components Regression
PGIS	Participatory Geographic Information Systems
SDSS	Spatial Decision Support System
UGC	User-Generated Content
VGI	Volunteered Geographic Information

ABSTRACT

The arrival of the World Wide Web, smartphones, tablets and GPS-units has increased the use, availability, and amount of digital geospatial information present on the Internet. Users can view maps, follow routes, find addresses, or share their locations in applications including Google Maps, Facebook, Foursquare, Waze and Twitter. These applications use digital geospatial information and rely on data sources of street networks and address listings. Previously, these data sources were mostly governmental or corporate and much of the data was proprietary. Frustrated with the availability of free digital geospatial data, Steve Coast created the OpenStreetMap project in 2004 to collect a free, open, and global digital geospatial dataset. Now with over one million contributors from around the world, and a growing user base, the OpenStreetMap project has grown into a viable alternative source for digital geospatial information. The growth of the dataset relies on the contributions of volunteers who have been labeled 'neogeographers' because of their perceived lack-of-training in geography and cartography (Goodchild 2009b; Warf and Sui 2010; Connors, Lei, and Kelly 2012). This has raised many questions into the nature, quality, and use of OpenStreetMap data and contributors (Neis and Zielstra 2014; Neis and Zipf 2012; Estima and Painho 2013; Fan et al. 2014; Haklay and Weber 2008; Corcoran and Mooney 2013; Helbich et al. 2010; Mooney and Corcoran 2012b; Haklay 2010b; Budhathoki and Haythornthwaite 2013; Mooney, Corcoran, and Winstanley 2010; Mooney and Corcoran 2011; Haklay et al. 2010;

Mooney, Corcoran, and Ciepluch 2013; Stephens 2013).

This study aims to complement and contribute to this body of research on Volunteered Geographic Information in general and OpenStreetMap in particular by analyzing three aspects of OpenStreetMap geographic data. The first aspect considers the contributors to OSM by building a typology of contributors and analyzing the contribution quality through the lens of this typology. This part of the study develops the Activity-Context-Geography model of VGI contribution which uses three aspect dimensions of VGI contributions: the *Activity* (the amount and frequency of content creation, modification and deletion); *Context* (the technological and social circumstances that support a contribution); and *Geography* (the spatial dimensions of a contributor's pattern). Using the complete OpenStreetMap dataset from 2005 to 2013 for the forty-eight contiguous United States and the District of Columbia, the study creates twenty clusters of contributors and examines the differences in positional accuracy of the contributors against two datasets of public school locations in Texas and California. The second part of the study considers the questions of where mapping occurs by evaluating the spatial variability of OSM contributions and comparing mapping activity against population and socioeconomic variables in the US. The third part of the study considers the choices that OSM contributors make through the types of features that are most commonly mapped in different locations. Understanding the types of contributors, their differences in quality, the spatial variability in mapping activity, and their choices in types of features to provide data will provide insight into the credibility of users, the

trustworthiness of their contribution, and where there are gaps in mapping activity and feature representation.

CHAPTER I. INTRODUCTION

Overview

The arrival of the World Wide Web, smartphones, tablets and GPS-units has increased the use, availability, and amount of digital geospatial information present on the Internet. Users can view maps, follow routes, find addresses, or share their locations in applications including Google Maps, Facebook, Foursquare, Waze and Twitter. These applications use digital geospatial information and rely on data sources of street networks and address listings. Previously, these data sources were mostly governmental or corporate and much of the data was proprietary. Frustrated with the availability of free digital geospatial data, Steve Coast created the OpenStreetMap project in 2004 to collect a free, open, and global digital geospatial dataset. Now with over one million contributors from around the world, and a growing user base, the OpenStreetMap project has grown into a viable alternative source for digital geospatial information. The growth of the dataset relies on the contributions of volunteers who have been labeled 'neogeographers' because of their perceived lack-of-training in geography and cartography (Goodchild 2009b; Warf and Sui 2010; Connors, Lei, and Kelly 2012). This has raised many questions into the nature, quality, and use of OpenStreetMap data and contributors (Neis and Zielstra 2014; Neis and Zipf 2012; Estima and Painho 2013; Fan et al. 2014; Haklay and Weber 2008; Corcoran and Mooney 2013; Helbich et al. 2010; Mooney and Corcoran 2012b; Haklay 2010b; Budhathoki and Haythornthwaite 2013; Mooney, Corcoran, and Winstanley 2010; Mooney and Corcoran 2011; Haklay et al. 2010; Mooney, Corcoran, and Ciepluch 2013; Stephens 2013).

This study aims to complement and contribute to this body of research on

Volunteered Geographic Information in general and OpenStreetMap in particular by analyzing three aspects of OpenStreetMap geographic data. The first aspect considers the contributors to OSM by building a typology of contributors and analyzing the contribution quality through the lens of this typology. This part of the study develops the Activity-Context-Geography model of VGI contribution which uses three aspect dimensions of VGI contributions: the *Activity* (the amount and frequency of content creation, modification and deletion); *Context* (the technological and social circumstances that support a contribution); and *Geography* (the spatial dimensions of a contributor's pattern). Using the complete OpenStreetMap dataset from 2005 to 2013 for the forty-eight contiguous United States and the District of Columbia, the study creates twenty clusters of contributors and examines the differences in positional accuracy of the contributors against two datasets of public school locations in Texas and California. The second part of the study considers the questions of where mapping occurs by evaluating the spatial variability of OSM contributions and comparing mapping activity against population and socioeconomic variables in the US. The third part of the study considers the choices that OSM contributors make through the types of features that are most commonly mapped in different locations. Understanding the types of contributors, their differences in quality, the spatial variability in mapping activity, and their choices in types of features to provide data will provide insight into the credibility of users, the trustworthiness of their contribution, and where there are gaps in mapping activity and feature representation.

Dissertation Organization

Chapter two is an exploration of the background of the material and a review of pertinent literature. Chapter three lays out the conceptual framework on which this dissertation is based including the development of the Activity-Context-Geography (ACG) Model, discusses the methods of mapping in OpenStreetMap, and presents the research motivations, research questions and research objectives. Chapter four provides a discussion of the data sources, study areas, the method employed to classify OSM contributors using the ACG Model, the method to determine accuracy of the ACG Model groups, and the statistical methods employed to examine the spatial relationship between mapping activity and population. Chapter five presents the results of the ACG Model classification and statistical analyses of OpenStreetMap data quality. It also reports on the results of the research questions about where OpenStreetMap mapping occurs and what types of features are mapped. Chapter six provides a discussion of the results and places them in the context of the previous literature. Chapter seven provides a concluding summary of the dissertation and discusses possible future research topics that may follow this dissertation. All figures and tables in the dissertation are within the text near where they are first mentioned.

CHAPTER II. BACKGROUND AND LITERATURE REVIEW

The Emergence of Volunteered Geographic Information

In his argument for the creation of Geographic Information Science (GIScience) as a discipline, Michael Goodchild discusses how geographic information (also called spatial data) is unique. "What distinguishes spatial data is the fact that the spatial key is based on two continuous dimensions" (Goodchild 1992, 2). Geographic information (GI) is a digital representation of real geographic phenomena. How best to represent GI "is of such importance that one might go so far as to argue that the greatest challenge in GIScience is to find ways of building useful representations of the infinitely complex world around us in the almost absurdly limited, discrete digital environment of a computer" (Goodchild 1995, 1–2).

In addition to the problem of representation, geographic information poses other challenges. It requires a system capable of map projections to enable conversion between two-dimensional data and three-dimensional data. Goodchild argues that spatial data exhibits spatial dependence where nearby locations will tend to exhibit similar properties (Goodchild 1992). Spatial data requires analytical techniques different from other types of data. In other words, spatial data is special.

Traditionally, collecting, creating, storing, and using GI has required extensive technical resources and abilities. Collecting geographic information may require survey teams, global positioning systems (GPS), aerial data collection, or satellites. Even before the rise of Geographic Information Systems (GISystems), map data (non-digital GI) was collected and controlled by local, regional, and national governments (state agencies). A few large corporations might alone or in partnership with government agencies collect

GI, but the capability remained out of reach for the average citizen (Goodchild 2007c).

In the first decade of the 21st century, technological and social changes would begin a revolution in how GI is created. First, the advent of Web 2.0 technologies allowed greater participation from users to contribute to websites without expert knowledge (O'Reilly 2005). User-generated content is a pillar of 'Web 2.0', a model developed by Tim O'Reilly as interactive, social, distributed, scalable, and collaborative internet websites as opposed to static, author-driven, single-source websites (O'Reilly 2005). Web 2.0 is the internet of blogs, "mash-ups", and APIs (Application Programming Interfaces) which allow users to interact and contribute content, feedback, and commentary to websites (Chow 2008). Second, global positioning and locational devices became more affordable and ubiquitous (Turner 2006). Third, ubiquitous mobile computing allowed for quick and easy access to the Internet (Schuurman 2009). Fourth, a trend toward social and collaborative Internet projects provided the technical capability and social will necessary to enable citizen science and collaborative mapping projects using non-specialists computer users (Haklay, Singleton, and Parker 2008).

Researchers in geography have expanded on the Web 2.0 concept. Neogeography (Turner 2006) refers to advancement of web technologies to enable novice geographers and cartographers to create geographic information and maps without expert knowledge (Goodchild 2009b). 'DigiPlace' refers to the combination of virtual and physical spaces that are combined into one lived experience (Zook and Graham 2007). The GeoWeb refers to the geospatial world wide web, or the ability of websites to produce, distribute, and foster the creation of GI (Haklay, Singleton, and Parker 2008).

Michael Goodchild coined the term Volunteered Geographic Information (VGI)

as a special case of user-generated content primarily found on the Internet (Goodchild 2007a). The main distinction between VGI and other types of GI is that VGI is not produced by institutions or government agencies but by citizens attempting to share geographic knowledge (Elwood 2008). VGI may take many forms (further discussed in the next section), including having been used for collaborative mapping projects (Coleman, Georgiadou, and Labonte 2009; Hall et al. 2010), disaster response (Goodchild and Glennon 2010; Liu and Palen 2010; Roche, Propeck-Zimmermann, and Mericskay 2011), social networking (Newsam 2010; Stefanidis, Crooks, and Radzikowski 2011), urban management and planning (Song and Sun 2010; Madej et al. 2012; Knudsen and Kahlia 2012), citizen science (Tulloch 2008; Connors, Lei, and Kelly 2012), outdoor recreation (Parker, May, and Mitchell 2013) and farming (Kagoyire and de By 2012).

The research consensus on VGI is that it exhibits the following properties:

- VGI is produced by citizens, not institutions or state agencies (Goodchild 2007a; Elwood 2008; Goodchild 2007b);
- VGI is crowdsourced information from many contributors (Elwood, Goodchild, and Sui 2012; Haklay, Singleton, and Parker 2008);
- VGI is producing extremely large volumes of data (Sui, Goodchild, and Elwood 2013);
- VGI may take many forms, including citizen science projects (Tulloch 2008; Haklay 2010a), collaborative mapping efforts (Haklay, Singleton, and Parker 2008), or social networking communications (Graham, Hale, and Gaffney 2014; Stefanidis, Crooks, and Radzikowski 2011; Taylor,

Tsou, and Leitner 2013).

Types of VGI

What does it mean to be "volunteered" geographic information? The act of volunteering suggests a knowing effort of good will on the part of the contributor, but some GI may not be considered knowingly given by the contributor. Therefore, how voluntary the data is could be considered a gradient (Harvey 2013). "An individual who's use of a toll road is recorded is not volunteering geographic information" (Elwood, Goodchild, and Sui 2012, 575). Francis Harvey (2013) breaks VGI into two categories, VGI (opt-in participation such as OpenStreetMap or geocaching) and Contributed Geographic Information (opt-out). The microblogging platform Twitter, for example, originally included location information with an update ("tweet") by default (opt-out); it has since changed the policy to require contributors to choose to include location information (opt-in). Contributed geographic information (CGI) would be crowdsourced data gathered, for example, from cell phone tracking or RFID-equipped card tracking. In reported cases, software malfunctions or intentionally programmed software have tracked cell phone use and location information. In these cases, a cell phone's user's location was recorded and tracked in large detail. Research using CGI data is problematic as the data may be incomplete and ethically wrong to use personal information without consent. Locational privacy is a new and complex issue in the United States and some question the use of this data by the government or telephone companies for surveillance (Kar, Crowsey, and Zale 2013; Crampton 2010).

VGI has similarities to Participatory Geographic Information Science (PGIS) which has a longer history of research in the literature. PGIS (often Public PGIS or

PPGIS) projects use informed citizenry with local knowledge and GIS technology to solve a particular problem, often with the aim of empowering citizens (Tulloch 2008). PGIS has been used in water use planning (Nyerges et al. 2006), urban planning (Elwood 2002; Leitner et al. 2002), and environmental monitoring and decision making (Tulloch 2008; Young and Gilmore 2013; Balram, Dragićević, and Feick 2009). PGIS differs from VGI, however, in the scope of the user/contributor community and the level of involvement of the user/contributor in the project. PGIS projects look at local issues important to the community and stakeholders with participants at different levels (citizen, administrator, land owner, governmental authority). VGI projects, on the other hand, tend to have little or no hierarchy in theory with little difference between contributors. PGIS projects often require some interaction with GIS software or similar analogs in order to enable the decision making process. VGI, on the other hand, generally requires limited, if any, interaction with GIS technologies and not much more than placing points or drawing lines.

Volunteered Geographic Information (VGI) can be defined as user-generated content (UGC) contributed online that provides location information. VGI, as a special form of UGC, is limited to online content (Goodchild 2007a). The GI part of VGI may be topological vector structures or photo imagery, but it can also be latitude and longitude pairs or descriptive text. To follow Harvey's distinction between VGI and Contributed Geographic Information (Harvey 2013), VGI should not be coercively obtained; at a minimum, contributors should have the option to "opt-out" of contributing. For example, to geotag status updates in Twitter, the user must enable that feature.

Perhaps the best way to illustrate the properties of VGI is to compare them with

traditional, authoritative GI collected by government agencies, military groups, and private agencies. Agencies that produced traditional GI used procedures to control quality during the acquisition and compilation of geospatial data and procedures to document quality and methods in the form of metadata (Goodchild and Li 2012). VGI sources typically have little metadata, or when they do, data quality is not a concern. Instead, crowd-sourcing (aka, "the wisdom of crowds") is used as a technique to ensure quality and converge on truth (Raymond 2001). Table 1 explores the differences between VGI and traditional authoritative GI sources from government agencies, the military and corporations.

Table 1. Differences between VGI and Traditional Authoritative GI.

Topic	**Volunteered GI**	**Traditional Authoritative GI**
Data Type	Vector, raster, text, lat/long, imagery	Typically vector or raster
Metadata	Usually some type of accompanying metadata	Typically standardized; for U.S. federal agencies, following FGDC standards
Quality Control	Crowd-source error detection and fixes	Usually defined by the project
Access for Use	Often free with little or no restrictions in use	Varies by agency and type; corporate data is licensed and generally expensive; military data is typically confidential

Table 2. A typology of VGI Project Types.

VGI Project Type	Description	Example Sites
Citizen Science Projects	Users with little or no formal scientific training contribute location and other data for a project with a scientific goal. eBird, for example, allows users to contribute location, time, and species siting data for birds worldwide.	eBird.org, lowwater.org
Collaborative Mapping Projects	Users contribute geographic and attribute data with the intent to map the results. Wikimapia, for example, has users who outline or pinpoint features on aerial imagery and describe the features.	Wikimapia, OpenStreetMap
Social Media Data	Social media data that includes location information either within the text or as metadata to a contribution may be considered VGI. These can include photos with embedded location data, "check-ins" by users at stores and restaurants, and tags of locations in status updates.	Facebook, Foursquare, Twitter

The VGI project type most similar to PGIS are citizen science projects. In Table *2*, a typology of the VGI projects is detailed based on Goodchild's original article on VGI (Goodchild 2007c). Citizen science projects use the distributed efforts of individuals to

collect data, process data, or solve tasks. Citizen science projects using geographic information have a long history, going back to at least the annual Christmas bird count of the Audubon society begun in 1900. Survey maps in Britain in the 1930s and 1940s were carried out by schoolchildren and teachers (Elwood, Goodchild, and Sui 2012). Citizen science projects in the digital age include the SETI@Home project and Folding@Home, both of which used the distributed resources of home computers to help process and analyze data (Anderson et al. 2002). Citizen science projects that collect VGI include the Audubon Society and Cornell Lab of Ornithology eBird.org[1] site, the University of California oak disease mapping site[2], the University of California at Davis California Roadkill Observation System (CROS)[3], and the Low Water Crossing mapping site[4] of central Texas (Parr and Scholz 2015). Each of these sites allows contributors to submit location (either by submitting latitude and longitude of a point, indicating on a map, or submitting through a mobile device), a description of the feature, and optionally an accompanying photograph.

A benefit of citizen science projects is their potential to fill in the gaps in local knowledge at a lower cost than not using citizens (Connors, Lei, and Kelly 2012). Since the participant may not be formally trained, the ability to use a website or smartphone

[1] http://ebird.org/content/ebird/
[2] http://www.oakmapper.org/
[3] http://www.wildlifecrossing.net/california/
[4] http://www.lowwater.org

application is critical. Sites should be designed to be simple but ensure data quality (Newman et al. 2010). The reliability of the GI produced is a concern, as with all VGI (Harris 2012; Flanagin and Metzger 2008). The training, knowledge and ability of the contributor is key to the quality of the resulting information (Goodchild 2007a).

Citizen science blurs the distinction between a 'citizen' and a 'scientist' (Haklay 2010a). The concept of citizen science assumes a difference between the professionally-trained scientist and the lay-person which is a distinction that has arisen in the past two centuries (Crampton 2010). Citizen science projects require different levels of participation and engagement from their contributors. Haklay (2013) identifies four levels of participation. Level one uses 'citizens as sensors' (Goodchild 2007a) or volunteers their computers. Level two requires citizens to become data interpreters and volunteer thinking skills. Level three ('participatory science') requires participation in problem definition and data collection. Level four, which Haklay refers to as 'Extreme Citizen Science' (Haklay 2013), uses citizens to define the problem, collect data, and analyze for a solution.

Collaborative mapping projects represent a second type of VGI. Like citizen science projects, collaborative mapping projects collect and share data using the local knowledge of data producers to crowdsource GI. The difference, however, is the level of participation required, purpose and use of the GI collected. Collaborative mapping requires only the volunteering of information – the lowest level in Haklay's citizen science participation ranking (Haklay 2013). The purpose of a collaborative mapping project is to create a complete spatial dataset, generally on a theme. At the time of writing, many mapping projects are in progress at differing stages of maturity. Projects

include Wikimapia[5], Google Map Maker[6], OpenStreetMap[7] and its many offshoots (maps for cycling[8], topography[9], navigation[10], mountain biking[11], and aviation[12]).

Collaborative mapping projects using VGI began as a way to fill in map data that was not available freely or easily from state agencies or institutions (Coleman 2013; Haklay, Singleton, and Parker 2008). State agencies including the U.S. Geological Survey, U.S. Census and National Weather Service are now users of collaborative mapping project data or running VGI projects of their own (Devillers, Bégin, and Vandecasteele 2012; Coleman 2013). In times of crisis, mapping groups have been quick to respond to needs including the 2010 Haitian earthquake and the 2013 Colorado floods[13] (Liu and Palen 2010; Zook et al. 2010). Despite these benefits, questions remain to the quality of the map data, liability, and usefulness of the data compared to authoritative sources (Mooney et al. 2011; Flanagin and Metzger 2008; Goodchild 2009b). For more discussion on the issues of VGI, see the section *Research Issues in VGI.*

Collaborative mapping and other VGI projects use the technology of Web 2.0. Map mashups combining information from different websites began with Paul

---

[5] http://www.wikimapia.org
[6] http://www.google.com/mapmaker
[7] http://www.openstreetmap.org
[8] http://www.opencyclemap.org
[9] http://www.opentopomap.org
[10] http://www.openseamap.org
[11] http://www.openmtbmap.org
[12] http://www.openaviationmap.org
[13] https://lists.openstreetmap.org/pipermail/talk-us/2013-September/011789.html

Rademacher's desire to find affordable housing in San Francisco in 2004. Using information from Craigslist and the newly released Google Maps, Rademacher created a map showing apartments by rent that would update automatically from Craigslist.[14] Currently, the GeoWeb uses a variety of different technologies. These include on-line mapping tools from industry (Google Maps[15], Google Earth[16], Bing Maps[17], Yahoo! Maps[18]), visualization tools (OpenLayers[19]), spatially-enabled databases (PostGIS), communication protocols (Web Feature Service, tile servers, Web Mapping Service) and a suite of new data standards (GeoRSS, GeoXML, GeoJSON, Keyhole Markup Language) from the Open Geospatial Consortium (OGC) (Haklay, Singleton, and Parker 2008). The glue that enables all of these different technologies to work together is the Application Programming Interface (API), a set of standards that each technology uses to be able to communicate through a standardized programming language (Java, Javascript, perl, PHP, Python, or C) (Chow 2008). The API enables Information Communications Technologies (ICTs) to share data in a fast and automated fashion between sites.

This ability to share information from a large number of users from multiple sites has given rise to social networking. Social networks are ICTs (websites and mobile device applications) that allow users to share updates, photos, messages, news articles,

---

[14] http://www.housingmap.com
[15] http://maps.google.com
[16] http://earth.google.com
[17] http://www.bing.com/maps/
[18] http://maps.yahoo.com/
[19] http://openlayers.org/

web links, and game technologies across multiple platforms centered on a web portal. A 2914 US Census report found that nearly 75% of US households have some form of Internet access (File and Ryan 2014). While not ubiquitous, social media does have a large number of users. Sites include Facebook[20], Twitter[21], Weibo[22], MySpace[23], Instagram[24], Google Plus[25], Pinterest[26], LinkedIn[27] and many others. Facebook alone has one billion users – one out of every eight people living on the earth at of 2015.

Sharing is the heart of social networking, and sharing location information is a natural extension. Shared location information in social media represents a third type of VGI. Most social networking sites provide for opt-in location sharing and some sites (Fourquare) are built around location sharing. Facebook, Yahoo, Yelp Swarm, and Foursquare, etc., allow users to 'check-in', indicating their location to their shared users. Twitter tweets may include location information called geotags. Location information can also come from information *in context* of the shared data from language or technical codes (Crampton et al. 2013; Takhteyev, Gruzd, and Wellman 2011; Hardy 2008).

Shared location information in social media sites is notably different in the purpose of the data and the participation level of the contributor from other types of VGI.

---

[20] http://www.facebook.com
[21] http://www.twitter.com
[22] http://www.weibo.com
[23] http://www.myspace.com
[24] http://www.instagram.com
[25] http://plus.google.com
[26] http://www.pinterest.com
[27] http://www.linkedin.com

In some cases, contributors may be unaware that they are sharing location data. Privacy concerns also lead to security concerns (Elwood 2008; Kar, Crowsey, and Zale 2013; Elwood and Leszczynski 2011). Understanding GI, however, can lead to greater concern for privacy (Mathews et al. 2012). While the primary purpose of shared location information is to inform a social network, corporations and state agencies have also used this data to target advertisements, track behavior patterns, and monitor associations (Crampton 2009; Elwood and Leszczynski 2011). Despite these issues, socially shared location information has provided insights into the spread of disease (Cook et al. 2011), understanding collaborative authorship (Hardy, Frew, and Goodchild 2012; Hardy 2008), following the diffusion of ideas and social memes (Crampton et al. 2013; Graham, Hale, and Gaffney 2014), and better understanding the social connections online (Takhteyev, Gruzd, and Wellman 2011).

Note that within VGI, social media websites, citizen science projects, and collaborative mapping projects are not mutually exclusive. OpenStreetMap, for example, has registered users who can communicate with each other and share information, and it uses a social system to create the structure of its database. Likewise, a citizen science project can involve mapping or could have a social media aspect including sharing photos or commenting on contributed information. All of VGI is UGC – that is, User-Generated Content is online content and all of VGI falls into this category; yet citizen science projects and collaborative mapping projects have existed before the Internet came into prominence. Some social media information may be sponsored or used as advertisements which fall outside of the UGC definition. Participatory GIS (PGIS) has some overlap with the traits of VGI, collaborative mapping and citizen science (depending on the

project and goals of a PGIS project).

Research Issues in VGI

The National Research Council Committee on the Strategic Directions for the Geographical Sciences in the Next Decade identified the following three questions as the key research questions for VGI:

- "What are the characteristics of the producers of VGI and how should we evaluate the content and quality of what they have produced?" (NRC 2010, 108)
- "In what ways does participation in VGI have the unintended effect of increasing the digital divide?" (NRC 2010, 109)
- "What and where are the most significant threats to human privacy as presented by emerging geographical technologies and how can we design technologies to provide protection?" (NRC 2010, 110)

These questions form the basis of research topics in VGI, and the first question is central to the objectives of this thesis. Expanding on these issues, research in VGI from 2007-2015 has proved fruitful. Research areas include the motivations of contributors, the accuracy of VGI, how to trust VGI data, access and the digital divide in VGI, the changes in approaches to mapping, the relationship of VGI to other authoritative data, and social issues of participation such as the role of gender in exclusionary mapping practices. These issues are discussed in this section.

Traditionally, modern mapping required an advanced skillset and expensive equipment. As previously discussed, Web 2.0 technologies blur the distinction between consumer and producer and between novice and expert. The availability of technology

and growth of well-educated individuals has promoted the increase in interest in citizen science and collaborative mapping projects (Haklay 2013) The characteristics of VGI contributors such as level of training, interest, and motivation will help to explain how to evaluate the quality of their contributions. OpenStreetMap collaborators suggest a range of motivations. A contributor may have an idealistic of a free, open map or an anti-national mapping agency viewpoint. OpenStreetMap's creator, Steve Coast, suggest that participating in VGI projects is "addictive" as a contributor becomes part of a community (Haklay and Weber 2008, 16).

Coleman et al categorize VGI contributors into five classes: neophyte, interested amateur, expert amateur, expert professional, and expert authority (Coleman, Georgiadou, and Labonte 2009). A neophyte contributor would have little knowledge but some interest on a subject, whereas an expert would have training, and an authority would have practiced a field and established him/herself as an authority. Coleman et al also develop a set of motivating factors around VGI which could be altruistic, socially motivated, motivated by a pride of place or malevolent. Budhathoki and Haythornthwaite expand on the list presented by Coleman et al with a list of intrinsically and extrinsically motivating factors (Budhathoki and Haythornthwaite 2013). Intrinsic motivational factors include a *unique ethos, learning, personal enrichment, self-actualization, self-expression, self-image, fun, recreation, instrumentality, self-efficacy, meeting a need, the freedom to express,* and *altruism* (Budhathoki and Haythornthwaite 2013, 558). Extrinsic motivational factors include *career, social relations, the goal of the project, community, identity, reputation, monetary return, reciprocity, system trust, networking,* and *sociopolitical motivations* (Budhathoki and Haythornthwaite 2013, 558). They also

suggest that motivation may be tied to the level of commitment required, whether lightweight or heavyweight, to the project. In a 2009 survey of OpenStreetMap contributors (n=459), Budhathoki and Hawthornthwite found the highest motivations were the success of the OpenStreetMap community, the goals of the project, altruism, and usefulness of local knowledge. Very active contributors also consider the goals of open access data and career goals as motivators (Budhathoki and Haythornthwaite 2013).

While a contributor's motivation may explain why they produce VGI, how can a user determine if the data is trustworthy? Trust can refer to information credibility, where the GI itself is shown to be trustworthy, or it can refer to source credibility, where the contributor is known to provide reliable GI. Map data produced by state agencies and corporations puts the name of the institution on the data as a marker of trust and a clear point of liability. Whether VGI data can be trusted and who is liable for misinformation were questions from the outset of research on VGI (Goodchild 2007b). One possibility would be to include revision information within the dataset to show users where hot spots of change have occurred (Trame and Keßler 2010). Keßler expands on this idea of revision history to indicate levels of trust. If, as Haklay found, that more edits produce better accuracy, then the more edits that a feature has would indicate a more trustworthy feature (Haklay et al. 2010). Putting this idea into practice by field testing the positional accuracy of heavily edited features, only a weak positive correlation was found (Keßler and Anton de Groot 2013). Whether the general public is aware of any specific credibility issue with online geographic information is uncertain (Mathews et al. 2012).

Collaborative mapping has been hailed as the "democratization of cartography" (Crampton 2010, 37). Neogeography exchanges the expertise of cartographers,

mapmakers, and National Mapping Agencies (NMAs) with the distributed knowledge of citizens. The process of mapping is now as much of interest as the map itself (Crampton 2009). "To ask about the map and the mapping process is, then also to ask about the systems of social beliefs and practices that give rise to the mapping project" (Pickles 2004, 76). Collaborative mapping is a social process and is subject to the same biases and problems as society at large. Rather than being inclusive, the inequalities already present in society could be recreated in VGI (Elwood 2008). In a 2012 survey of online GeoWeb users, men were significantly more likely to have used OpenStreetMap than women. In particular the process of mapping may exclude minority groups and minority opinions. In an online discussion forum for OpenStreetMap, editors debated the labeling categories among different types of men's sexual entertainment facilities, but disregarded the differences among categories of schools for children (Stephens 2013).

Part of the issue with exclusion in the social construction of VGI may be related to issues of access and the digital divide. The digital divide may refer to the lack of computer skills, a lack of computers or network equipment, a lack of digital experience, or a lack of access (van Dijk and Hacker 2003). For example, having a smartphone may provide some Internet access, but it does not provide a full experience to gather skills in word processing or spreadsheet software. The digital divide may occur for possible contributors of VGI (who lack access to effectively contribute), for VGI users who cannot access the data, and in the geographic unevenness of VGI representation. Europe and North American have ten to seventy times more representation in VGI. VGI representation is lacking where it is needed most in Africa, Asia, and South America (Sui, Goodchild, and Elwood 2013).

OpenStreetMap

Frustrated by the restrictions of use, lack of updates, and lack of coverage of geospatial data from National Mapping Agencies, Steve Coast began the OpenStreetMap project in July 2004 as an effort to create a free map of the world (Ramm, Topf, and Chilton 2011). The term "free" here refers both to its zero cost and the right to use the information for any purpose (Stallman 2002). Its creation was inspired by the online encyclopedia Wikipedia. As with Wikipedia, anyone can add, change, or delete information in the OpenStreetMap dataset. Here, the mapping data rather than the map itself is the product of interest. As of 2013, there are over one half million contributors to the OpenStreetMap project and its data covers every country.

The analogy of OpenStreetMap to Wikipedia is an imperfect one. Both OpenStreetMap and Wikipedia maintain a voluntary board of editors, and procedures are agreed upon by consensus within the contributor community (Reagle 2010). OpenStreetMap requires a steeper learning curve to add information and to use it (Ramm, Topf, and Chilton 2011). Both Wikipedia and OpenStreetMap are social constructs for knowledge gathering and subject to the problems that exhibit biases against minorities (Stephens 2012; Reagle 2010). OpenStreetMap uses a collaboratively-authored wiki to document its structure – the same software used by Wikipedia and authored by the Mediawiki foundation. Unlike Wikipedia, only registered users can edit OpenStreetMap

(Haklay and Weber 2008).

Vandalism, which OpenStreetMap defines as "intentionally ignoring the consensus norms of the OpenStreetMap community"[28], is a problem on both OpenStreetMap and Wikipedia. Each has taken a social and technological approach to combat vandals: vandals are excluded from the discussion and editor process (a "virtual ban"), and robots (scripts and programs) check the databases for problem signs and flag the data in question. In Wikipedia, a page being vandalized may lead to that page being locked. In OpenStreetMap, one vandal introduced a fake town in an agricultural setting. Figure 1 shows a screenshot of a fake town in OSM – an example of vandalism in the data.

[28] http://wiki.openstreetmap.org/wiki/Vandalism

Figure 1. A screenshot of a fake town in OpenStreetMap. (Used under the Creative Commons license from http://wiki.openstreetmap.org/wiki/File:West_Harrisburg.jpg)

OpenStreetMap refers to multiple social and technical platforms. There is the database which houses the current working data and communicates with the website and API Server. There is the data itself, which is made of four primitive elements: nodes (points and nodes), ways (lines and areas), relations (multilines, multipoints, and multiareas) and tags (attribute data for each of the primitive elements). A series of servers housed in London and replicated around the world maintain the data. There is the API which allows a large variety of editors and viewers to authenticate and communicate with the central database. Finally, there is the community of authors, editors, and developers working on the data, map display, and software that enables OpenStreetMap.

Unlike government agencies, OpenStreetMap does not enforce any limitations on how spatial features are annotated in the database although some automated scripts

written by contributors do process and "clean" data (Haklay and Weber 2008).

As of April 2015, OpenStreetMap has over five hundred thousand contributors, although only five percent of these contribute in a meaningful way. Seventy-one percent of contributors are based in Europe, and another twelve percent are based in North America. There are over 150 million features represented in the OpenStreetMap dataset (Neis and Zipf 2012).

Spatial Data Quality

Understanding contributors' motivations and characteristics may lead to a better understanding how to evaluate the quality of VGI. Van Oort (2006) proposes seven aspects of spatial data quality to consider:

- Lineage – the history of a geographic dataset;
- Positional accuracy – accuracy of coordinate values;
- Attribute accuracy – the accuracy of all variables that are not positional or temporal;
- Logical consistency – the agreement of relationships between variables;
- Completeness – "a measure of the absence of data and the presence of excess data" (van Oort 2006, 23). This can refer to how much of the known features are represented in the data set;
- Usage, purpose and constraints – information to assist the user in understanding the quality of the dataset;
- Temporal Quality – the validity and accuracy of time measurements (van Oort 2006).

VGI projects may not adhere to the same rigid standards that traditional authoritative GI sources use in collecting data. A study comparing the positional accuracy and completeness in the OpenStreetMap road network of England found that while the data was not as complete or accurate as the British Ordinance Survey data, when the contributor was diligent the accuracy was within tolerances (Haklay 2010b). Completeness of data continued to improve over time. Haklay also found that an increase in contributors in an area improved positional accuracy. In other words, the palimpsetic nature of VGI (where data is constantly overwritten) should improve the positional accuracy (Haklay et al. 2010). A followup study using the French road network for comparison evaluated OpenStreetMap's GI with all of van Oort's spatial data quality variables. The study found that the lack of standards in gathering data, different processes of capture, and different data sources inhibited the quality of the data in each of the seven categories (Girres and Touya 2010). In 2012, a revised analysis method found that the OpenStreetMap road network positional accuracy and thematic accuracy was "very good" and approaching the Ordinance Survey quality (Koukoletsos, Haklay, and Ellul 2012).

Before VGI and the GeoWeb, geographic data creation was largely the domain of government agencies or large corporations (Goodchild 2007b). Organizations use data collection standards reported in the metadata to provide insight for the user of the data into its quality assurance (Goodchild 2002). Spatial data errors are often related to uncertainty in the data. Uncertainty may occur through error, vagueness in the data, or ambiguity in the definition of a feature (Fisher 1999). Error is the difference between a value in a database and its true value. Vagueness occurs when a range of possible values

are correct or exist in a fuzzy set. For example, a person with no hair is considered bald, but is a person with one hair? (Fisher 1999, 197). A single Boolean answer "yes" or "no" may not fit in this scenario. Ambiguity occurs out of non-specificity, or an imprecise definition of the phenomenon being described. Ambiguity may occur when different classification schemes use the same labels to refer to different degrees of a feature set's properties.

The Federal Geographic Data Committee (FGDC) is a government interagency that promotes the development, use, sharing, and dissemination of geographic information for the United States (Federal Geographic Data Committee 2015). In addition to promoting the National Spatial Data Infrastructure for sharing geographic information, the group also publishes standards for content, data transfer, positional accuracy, and metadata frameworks. These types of standards are largely missing from VGI data sources. When data quality is an issue with VGI, it uses crowd-sourcing to improve data quality (Haklay et al. 2010).

VGI presents new challenges in assessing spatial data quality. As the provenance and data collection methods will likely be unknown, other methods must be used to test the data for error (Haklay 2010b). VGI data creation does not use any of the quality controls that traditional GI use, nor does it report traditional error estimations. The most common model for quality control is crowd-sourcing corrections. More frequent changes may indicate higher quality (Haklay et al. 2010). Unlike other forms of GI, the context of VGI is critical to understanding its quality (Elwood, Goodchild, and Sui 2012). VGI contribution often includes context in the form of blog entries, wiki entries, photographs, and tweets and generally lacks formal metadata. This context may present clues to the

quality of the data.

"Big Data" Methods and Geographic Information

The rapid increase of available online data including VGI, point-of-sale data, location-aware technologies, and geosensor networks has been labeled the *exaflood* or *data avalanche* (Miller 2010). It's estimated that from 2012 to 2020 there will be an increase of fifty times the annual amount of data produced (Sui, Goodchild, and Elwood 2013). Colloquially, this rise in the amount of data produced and the methods associated with producing meaningful results from the data are called "big data" (Crampton et al. 2013).

"Big data" is defined by three features, commonly known as the 3V's: volume, velocity, and variety (Laney 2001; Beyer and Laney 2012). The volume of data is extremely large, although what constitutes large has been a changing value over time. The velocity of data refers to the frequency at which new data is produced. The variety of data refers to the different formats generated such as twitter geotags, Google Maps KML files, and embedded latitude-longitude coordinate data.

The generating of new methods to extract meaningful results from "big data" is currently becoming an areas of interest from the government, big business, and the academic research community (Executive Office of the President 2012; Lohr 2012). Big data is not without its critics, however, and using "big data" methods effectively may yet require expertise in the subject and a fine understanding of the data (Graham 2012; Snijders, Matzat, and Reips 2012).

Methods in "big data" are an extension of methods in knowledge discovery, often

called Knowledge Discovery in Database (KDD). "Big data" methods must extend KDD to handle the volume, velocity, and variety of data that is being produced. KDD is a set of techniques to find interesting patterns in massive databases that can be understood by humans and valid for generalization (Miller and Han 2009). The process of KDD involves:

- *Data selection*;
- *Data pre-processing*, or removing "noise" from the data and duplicate records as well as handling missing data;
- *Data reduction and projection*, or reducing the dimensionality of the data;
- *Data mining*, or finding patterns in the data;
- *Interpreting and reporting*, which involves evaluating and reporting any findings (Miller 2010, 189).

Data mining refers to a range of tasks and techniques to process the data. These tasks include: *clustering, classification, association, deviations, trends,* and *generalizations*. Clustering involves grouping data together without preset classes, which classification uses training data to prefer a particular set. Association tasks find relationships among data objects. Deviation tasks look for outlier or unusual data. Trending, which attempts to quantify the data, and generalization tasks look for a compact description of the data (Miller and Han 2009; Mennis and Guo 2009). Using KDD to explore, analyze, and visualize geographic data has been well established (Ester, Kriegel, and Xu 1995; Ballatore, Bertolotto, and Wilson 2013; MacEachren et al. 1999; Han, Cai, and Cercone 1992; Gahegan et al. 2001).

Whereas data mining and KDD techniques might sit on a single server, "big data"

methods require multiple computers to store, process, and visualize results given the size of the datasets (Lohr 2012). Tasks commonly include presorting the data into groups and then resducing the data using MapReduce or Hadoop programming techniques (White 2012). To perform these tasks, programmers must turn to cloud computing, a distributed computer model where computing is viewed as a service (Yang et al. 2011). Data, analysis engines, and visualizations may all be distributed on different machines spanning the globe. Cloud computing is possible due to fast communications speeds and large servers that can store many virtual computers. Because of this distribution, data and processing demands can be allocated on an 'as-needed' basis, which enables better performance costs in servers. Data can be centralized in a few large server rooms and decentralized across the globe at the same time. The capability of cloud computing for geographic 'big data' has yet to be fully realized (Goodchild 2009a).

Conclusion

In a short span of less than ten years since Michael Goodchild coined the phrase "Volunteered Geographic Information," (Goodchild 2007a) research interest in the topic has remained relatively high. The results have included a special issue of the journal *GeoJournal* (Elwood 2008), a pre-conference at the 2011 Annual Association of American Geographers Meeting in Seattle, Washington, an edited volume *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (Sui, Goodchild, and Elwood 2013), and numerous journal articles, dissertations, and specialty sessions at conferences such as ACM SIGSPATIAL and GEOCROWD, the ESRI User Conference, and other GIS research conferences.

OpenStreetMap, as a subtopic of VGI, has also been a keen topic for both researchers and the general public. In addition to yearly "mapathon" events where OSM enthusiasts meet to map out areas and an annual "State of the Map" conference for the OSM community, there are at least twenty active OSM community groups in the United States alone. Three books have been published on OSM: *OpenStreetMap* (Bennett 2010), *OpenStreetMap: Using and Enhancing the Free Map of the World* (Ramm, Topf, and Chilton 2011), and the edited research volume *OpenStreetMap in GIScience* (Jokar Arsanjani et al. 2015).

Despite the interest from researchers and the wider community, several key questions regarding the OpenStreetMap project remain which may inform the work of VGI research as a larger topic. Much of the work has looked at data quality issues (such as positional accuracy) in the data (Goodchild and Li 2012; Haklay et al. 2010; Girres and Touya 2010), and some work has looked at types of contributors (Budhathoki and Haythornthwaite 2012; Coleman, Georgiadou, and Labonte 2009), to date, there has not been an effort to tie the two together. Coleman et al and Budhathoki and Haythornwaite both use motivation as the lens of categorizing VGI contributors. Assessing motivation requires evaluating the motives of each individual contributor which would be nearly impossible for a contributor base as large as OpenStreetMap's. Using methods considering the *in situ* characteristics of the contributor inherent in the dataset should provide an insight into all contributors in the dataset.

A second area of concern with OpenStreetMap is the completeness of the dataset and the choices that contributors make when decided to produce data. Stephens (2013) has discusses how the social forces and the role of gender in OpenStreetMap. Men make up a majority of OpenStreetMap contributors and editors, and this reflects a bias in how

place is presented within the dataset. Hecht and Stephens (2014) found that urban biases give metropolitan areas larger footprints in VGI on a *per capita* basis. The biases reflected in the data are the consequences of the choices that contributors make when creating, editing, and deleting data. These choices may include how to represent features, what areas are of interest to map, and how attribute data is codified. Unlike the FGDC standards for accuracy and metadata, a project like OpenStreetMap relies on social consensus and a network of editors to ensure data quality, and these formal and informal decisions can have consequences on the types of features represented and the completeness of the data. As OpenStreetMap evolves towards being a viable product for use in government and corporate institutions, the positional and attribute accuracy, consistency, and completeness of the data should be topics keen to researchers.

CHAPTER III. RESEARCH MOTIVATIONS, GOALS AND OBJECTIVES

Research Motivations

Place of This Project in Geography

Of the four traditions of Geography (the Spatial Tradition; the Area Studies Tradition; the Earth Science Tradition; and the Man-Land Tradition) identified by Pattison (1990), mapping (according to Pattison) falls squarely within the Spatial Tradition. This tradition includes the broader outlook of spatial analysis, the nature of distance and location, geometry, and measurement.

I would argue that much of what constitutes recent advances in digital cartography, online mapmaking, and the programmatic functions that enable websites and applications that support Volunteered Geographic Information projects have fallen outside of the realm of Geography and into Computer Science and Software Engineering. Yet, fundamental to collaborative mapping and many spatial citizen science projects (and, to a lesser degree, social media location data) is a fundamental (and perhaps new) way that individuals interact with the world around them. The rise of 'user-centered design' in web and mobile cartography provides greater access to spatial information to users while at the same time demanding fewer map-reading skills (Tsou 2011). For those with access to digital technology, geographic information is ubiquitous, easy to find, and much easier to produce.

Software engineers at companies like Google may continue to determine the best practices in data storage for spatial data and user interfaces for mapping. Geography should play a key role in the questions of how society determines how to represent space, the geography of how VGI reflects our mapping needs, the issues of data quality, and

understanding how social processes impact VGI. New developments such as embedded, wearable technology, heads-up mapping displays in cars, virtual mapping in glasses, and 3D displays will, if and when they occur, continue to redefine the role between map, mapmaker, and mapreader. If technology is a lens through which humanity experiences the world, then the discipline of Geography should be a clear filter to understand how the technology changes our interaction with the world, and how the technology reflects the world back to us.

In that spirit, this research aims to understand how the mapmaker (in this case, the OSM contributor) reflects the world that they see through their contributions by analyzing the different types of contributors, considering how mapping activity is related to population and other characteristics of the spaces where mapping occurs, and by further elaborating on the process of VGI contribution.

Place of This Project in the Literature

As discussed in the Conclusion section of Chapter 2, both VGI and OSM have been active research topics within the past few years. In VGI research, topics ranged from applications of VGI projects, what constitutes VGI, and issues concerning privacy, access, and, later, social use and biases within VGI. For much of this time, research focused on the data quality of OpenStreetMap data, technical methods to quantify use and users, and applications for OSM data.

This research will address several topics not previously addressed in detail in the VGI or OSM literature. The first is to propose a model to build a typology of OSM (or any VGI project) contributors based on the users' contributions within the dataset compared to other contributors. The model should determine clusters of contributors

based on their contributions. This model is called the Activity-Context-Geography (ACG) Model, and it is described in detail in the next chapter. Researchers have developed typologies of contributors in VGI and OSM, but none have used the data itself to develop these typologies (Coleman, Georgiadou, and Labonte 2009; Budhathoki and Haythornthwaite 2012).

The second topic addressed in this research is to evaluate the ACG Model against an external dataset to test if clusters of contributors within the ACG Model have different data quality outcomes. The usefulness of the AGC Model may be determined by its ability to detect meaningful differences between clusters of contributors. The model should predict differences in positional accuracy between groups. Data quality is a lasting concern in OSM research (Haklay 2010b; Girres and Touya 2010), but comparing the quality of different groups contribution has not occurred to date.

The third topic addressed in this research considers the spatial patterns of the mapping activity from the contributors. Mapping activity in OSM involves a set of choices for each contributor that are reflected in the data produced. These choices include what areas to map, what types of features to map, how to represent these features spatially and thematically. This research intends to expand on work looking at the biases in the VGI contribution process (Stephens 2013; Hecht and Stephens 2014; Crutcher and Zook 2009).

Research Goals and Objectives

In this section, I list the three goals of the dissertation as they relate to the topics discussed in the section above. Each goal has one or more objectives which help to define

the research questions in the Research Methods section. After each goal, I expand on its rationale as a topic of research.

Research Goal 1

The first research goal is to better understand the types of VGI contributors through the patterns of OSM data and its context by building a model of contributor types.

Objective of this goal:

1. Develop and implement the Activity-Context-Geography Model of VGI Contribution to identify contributor clusters.

Rationale

With over one million registered data contributors, OSM has become one of the most popular collaborative mapping sites (Neis and Zipf 2012). Steve Coast began the project in 2004 with an aim to build a free-to-use, worldwide map dataset (Haklay & Weber, 2008; Ramm et. al, 2011). OSM is created and maintained by volunteers who often band together through mapping parties to fill-in local gaps in geographic data or correct errors (Haklay et al., 2008). Volunteers also fill in data gaps in remote countries that have restricted, expensive or non-existent map data. Although Europeans make up more than 70% of OSM data contributors (Neis and Zipf 2012), OSM has a true international scale and much of the work is created remotely (Haklay & Weber, 2008). OSM's data formats are unrestrictive, so that users can contribute any discrete object as map data, and any user (including automated programs) can edit or delete other people's data. Volunteer editors check modified data using automated and semi-automated

programs. Ownership of the data is retained by the contributor under the Open Data Commons Open Database License.

If the type of contributor and quality of contribution is a general concern for VGI, then it has become a key concern for OpenStreetMap in particular. OSM has generated exceptional interest from data users, the media, corporations, and researchers (Haklay et. al, 2010; Horita et. al, 2013; Neis & Zipf, 2012; Ramm et al., 2011). In times of humanitarian crisis including the 2010 Haitian earthquake and the 2013 Typhoon Haiyan in the Philippines, OSM has provided infrastructure support, and volunteers have used the OSM platform to update map information for disaster relief (Zook et al. 2010; Neis and Zielstra 2014). Corporations and government agencies like Apple (Schutzberg 2012) and the U.S. Census Bureau (Forrest 2010) are incorporating OSM data in their products or using OSM to verify their data.

It has been suggested that the quality of Wikipedia articles are believe to be closely tied to the type of contributor that writes or edits the article (Anthony, Smith, and Williamson 2009; Liu and Ram 2009). OSM, like Wikipedia, is a UGC site with a high degree of openness and a wide spectrum of participation. If the type of contributor in Wikipedia is related to the quality of their contribution, it may also be the case with OSM. Therefore, establishing a contributor typology and determining its relationship to data quality is an important research area that this dissertation explores. Previous attempts to build typologies in VGI include Coleman et al (2009), who grouped VGI data contributors into five categories based on their knowledge and experience: neophyte, interested amateur, expert amateur, expert professional, and expert authority. Others have built conceptual models for evaluating contributors based on the complex interactions

within the OSM dataset itself (Rehrl et al. 2013). This research builds a typology of contributors from the internal data of the OSM dataset using the Activity-Context-Geography Model.

Research Goal 2

The second research goal is to provide a method for understanding the spatial data quality of VGI contributors through the model of contributor types.

Objective of this goal:

1. Statistically analyze the differences in contributor data quality (specifically, positional accuracy).

Rationale

Because of the unrestrictive license and global dataset, the OpenStreetMap project has gained widespread use beyond the hobbyist community and into the governmental and corporate world. Companies such as Apple and Foursquare have integrated OSM into their map products (Duncan 2012). The U.S. Census uses OSM data to verify their own data sources (Forrest 2010). More companies (including ESRI) are using OSM's prepared map tiles to display underlying digital map information.

OpenStreetMap has also been a crucial resource in humanitarian projects where mapping locations of distress and trouble in a timely fashion are crucial. In Haiti after the

devastating 2010 earthquake, OSM provided the infrastructure for a team of volunteers to map areas in need of help quickly (Zook et al. 2010). A Humanitarian OSM Team was formed after the earthquake to respond to locations in need. Recently, they were involved in mapping the West Africa Ebola epidemic.[29]

The increasing use of OSM data raises the scrutiny of its credibility and, in particular, its data quality (Girres & Touya, 2010; Haklay, 2010; Mondzech & Sester, 2011; Mooney et. al, 2010; Neis & Zipf, 2012). Generally, credibility may be related to artefacts present in the dataset. Repeated edits of a feature by multiple contributors may suggest attempts to create more accurate data, and could be an indirect indicator for possible data quality (Trame and Keßler 2010). Proximity to other features edited later may suggest acceptance of the first feature (Keßler and Anton de Groot 2013). Previous studies on OSM data quality have examined the accuracy of its linear features (Girres and Touya 2010), the completeness of its dataset (Koukoletsos et. al, 2012), and the number of volunteers necessary to map a location (Haklay et al., 2010). An increase in the number of contributors covering an area improves the positional accuracy of OSM data for the area (Haklay et al., 2010). Furthermore, the overall accuracy of the OSM dataset has improved over time (Haklay, 2010). This research examines the data quality of OSM through the clusters of contributors in the Activity-Context-Geography Model.

[29] http://hot.openstreetmap.org/

Research Goal 3

The third research goal is to develop a model of OSM contribution and examine the choices that contributors make when producing VGI in OSM. The spatial variations in how these choices impact the OSM dataset are examined.

Objectives of this goal:

1. Statistically analyze the relationship between population, socioeconomic characteristics, and mapping activity in OSM.
2. Compare the differences in mapping activity at different geographic scales.
3. List the most commonly mapped feature types.

Rationale

In 2011, the OSM data for Germany was considered "completed" (Neis, Zielstra, and Zipf 2011). This was declared because the turn-by-turn road network and directions were comparable to the data from Tom Tom, a commercial data vendor. OSM is more than simply a road network, however, and it includes a variety of data including natural, man-made, and administrative (ie, county boundaries) features. Contributors have added hiking paths, trees, restaurants, and entire towns which outlines of buildings detailed.

While there may be no upper limit on the number of features worth mapping in an area, there is a geographic unevenness to the "completeness" of the OSM dataset. This is in large part due to the choices that contributors make when adding data to OSM. Previous studies have found that OSM users exhibit a gender bias (Stephens 2013) and a *per capita* urban bias (Hecht and Stephens 2014). As contributors make choices in how to contribute, other biases may appear that present spatially. In the Choices of an OSM

Contributor section in the next chapter, I elaborate on how these choices may impact the quality, consistency, and completeness of the OSM data. Understanding how these choices present data spatially should clue OSM contributors on where data quality or consistency issues are present and where more mapping needs to be focused.

CHAPTER IV. THEORETICAL FRAMEWORK OF THE RESEARCH

Theories of Volunteered Geographic Information

Conceptual Models of VGI

A conceptual framework is a set of broad ideas and theories from the related fields of inquiry used to structure ("scaffold") a project and assist in communicating its findings (Smyth 2004). The framework here builds on previous conceptual frameworks from Budhathoki (Budhathoki 2010), Jankowski and Nyerges (Jankowski and Nyerges 2001), and Nedović-Budić and Pinto (Nedovic-Budic and Pinto 1999).

To date, there are few conceptual frameworks directly related to VGI, but VGI research is informed by closely related fields. The Enhanced Adaptive Structuration Theory (EAST) builds on group decision support research to describe spatial decision support systems (SDSS) (Jankowski and Nyerges 2001). SDSS and its related field Public Participatory GIS (PPGIS) use GIS as a tool to make group spatial decisions such as where to place a park or how to manage water resources. SDSS/PPGIS have some similarities to VGI production in that there are usually multiple contributors with different types of motivation, skillsets, and interests. The EAST framework partitions the SDSS process into convening constructs, process constructs, and outcome constructs. The framework recognizes the importance of the structure of GIS software, the character of the participants, and the sources of structure (Jankowski and Nyerges 2001). A more general framework outlines the relation between interorganizational GIS users (Nedovic-Budic and Pinto 1999). Here, the context is placed before the motivation which drives the process enabling collaborative GIS work.

Budhathoki has developed a framework around the motivations of VGI

contributors (Budhathoki 2010). The key to the VGI process is motivation, which is the lens through which action takes place. Motivation is a key to understanding crowdsourcing projects such as OpenStreetMap or Wikipedia (Benkler and Nissenbaum 2006). Budhathoki is specifically addressing VGI as a collective project. Therefore, he argues, "motivation is a necessary, but not sufficient for the production of knowledge commons in cyberspace" (Budhathoki 2010, 33). The "action and interaction arena" considers how contributors interact, cooperate, creates rules and norms, and decide to produce VGI. The output has geospatial and non-geospatial components. Figure 2 displays Budhathoki's conceptual, motivation-based approach to VGI contribution from his 2010 dissertation.

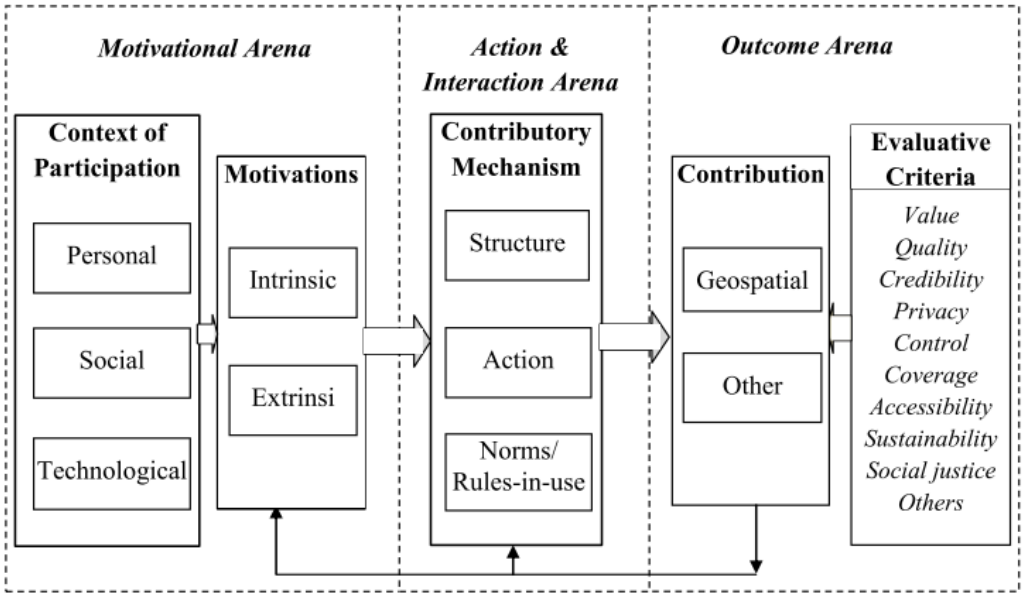

Figure 2. A conceptual framework of the motivations of VGI contributors. (From Budhathoki 2010).

Despite the viewpoint that motivation is the lens to understanding VGI production, it has been shown that motivation is not a necessary condition for creating

VGI (Harvey 2013). Applications may record location information without consent ("volunteering") or cameras may embed geographic coordinates automatically without any participation in the role of the contributor. Therefore, motivation is but one aspect of a contributor's construct.

Budhathoki is also only considering a certain type of VGI, namely VGI projects that are creating collective knowledge (Budhathoki 2010). VGI may come from a variety of projects from location sharing applications, social networking, shared images and photography, Wikipedia entries, collaborative mapping sites, and citizen science projects (Goodchild 2007c).

The Activity-Context-Geography Model

Motivation has been used to study VGI contributors, but motivation is difficult to assess directly from someone's data contribution. In order to group VGI data contributors, Rehrl et al (2013) consider the Actions (i.e., create, modify, delete) of contributor and the Domains (or geographic feature sets) of contribution, but they failed to consider the context in which one's data action takes place. Context is considered as a part of the Motivation arena by Budhathoki (2010). I propose a VGI participation model by considering three aspects of data contributors (Activity, Context, and Geography) to construct the *Activity-Context-Geography (ACG) model of VGI Contribution*.

The *Activity* aspect refers to the types and quantity of data contributions that produce VGI. This includes operations such as adding, editing, and deleting map and attribute data. It can also reflect the intensity of a contributor's data production. In the case of a citizen science project such as eBird.org, it may include the number of birds

sited by a contributor, the number of species overall, and the number of birds and species sited per birding event. In the case of a social networking site like Foursquare, it may include the number of check-ins, the number of locations, the number of check-ins per day, the number of business types frequented per contributor.

The *Context* aspect concerns the technological and social ("techno-social") circumstances that enable a contributor's contribution. The variables associated with the Context aspect include variables that relate to the contributor but do not produce VGI. These include how long a contributor has been a member of the VGI platform, how well-connected a contributor is in the platform's network, and if and how a contributor records his/her contribution activities. In the realm of OSM, this is related to details around the diaries and other annotations that contributors keep on the website which can be measured by variables including the number of diary or annotation entries, the word count, and the frequency of entries. In the case of a citizen science project, it may relate to similar diary variables. In the case of a social networking site, it could include the number of social network connections the contributor has or how frequently a contributor logs in or comments on other's entries.

The *Geography* aspect consists of the geographic components of the VGI data produced. VGI is unique from other forms of UGC in that it represents spatial patterns that reflect back to the ground. In some sense, the Geography aspect represents a view that a contributor has of the world, and it reflects the spatial interests or experience of a contributor. For the OSM dataset, I have chosen to represent the Geography aspect by the number of nodes per line or polygon feature, and the areal extent of features. Similar variables could be used for citizen science or social networking VGI.

The ACG model views contributors through their online persona. When personal data (motivation, training, experience) are typically unavailable for VGI data contributors, the ACG model can be used to examine contributors based on these three aspects. A VGI data contributor can fall on any point along the spectrum of any one of the three aspects. For example, a contributor with high Activity and low Geography variables may indicate a contributor with a strong connection to a relatively small area. The Geography aspect is not analogous to spatial extent, but spatial extent does have a strong impact on the Geography aspect variable. A higher Context aspect than other contributors may indicate a contributor with a stronger commitment to the VGI platform – an OSM power user or a social network user with many ties. This study is laid out to examine how effective the ACG model is to describe the different types of VGI data contributors, and if these different data contributors show different qualities in their VGI contribution. Figure 3 shows an example scatterplot of contributors plotted against the three dimensions of Activity, Context, and Geography. A complete list and description of the variables used in each aspect are included in Table 7, Table 8, and Table 9.

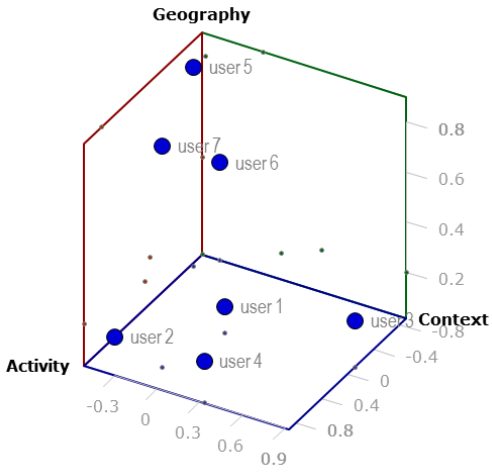

Figure 3. Example 3D Scatterplot of the ACG Model.

Mapping Activity in OpenStreetMap

OpenStreetMap, despite the name, is not a map, nor is it strictly a set of street data. Due to the open nature of the dataset, any spatial feature that can be designated as a point or a series or related nodes attached with attribute information can be stored in the OpenStreetMap database. OpenStreetMap, then, is not a map, but it is a world-wide, constantly-evolving set of spatial data that is designed to be easy to create and free to use(Haklay 2010b). While street and transportation network data roughly accounts for 54% of U.S. features in the dataset as of February 2013, contributors to OSM have added pharmacies and restaurants, hiking and bike paths, athletic fields and school grounds, political boundaries, rivers and streams, and even individual trees. I define mapping activity in OSM as the process of collecting, converting, modifying or deleting physical (trees, roads, buildings) and non-physical (political boundaries, thematic boundaries) entities into digital data represented as discrete features. Continuous features (eg, fields, rasters, elevation data) are not natively supported in OpenStreetMap.

The process by which data is collected varies. Originally, the creators OSM intended for hobbyists to "walk, hike, bike, or drive, recording their tracks using GPS devices. These recordings are then meticulously redrawn on a computer screen" (Ramm, Topf, and Chilton 2011, 3; Lin 2011). In practice, there are three methods that OSM data is collected. Using Global Positioning System (GPS) receivers in stand-alone units, in phones, or other devices, contributors can collect latitude and longitude coordinates or trace paths which are then loaded into features using an OSM data editor. Indeed, OSM has turned the sometimes dreary activity of field data collection into a social event

through regular mapping parties where OSM enthusiasts meet and map. Some contributors may mark this data as “field work” or “survey” data (Mooney and Corcoran 2012b). Contributors may take image data from Web Mapping Service tiles, orthophotos, or other sources and trace features using a mouse and computer screen in an OSM editor. OSM editors customarily add the image source (Bing Maps, Google Maps, etc) when adding data on screen. Both of these methods of generating map information may be considered “local knowledge” or “place knowledge” if the contributor has a connection to that location. Indeed, based on the attribute data in the OSM database, contributors attribute source information to “local/place knowledge” as often as they site GPS-created data. Figure 4 shows a comparison between GPS point traces of local buildings in Burlington, Vermont, and the (on-screen) traced buildings.

Figure 4. GPS Trace of Buildings in Burlington, VT (OpenStreetMap)

A third method of creating features in OSM is to import a dataset from another source. This could include a direct import of data, although importing restricted data is forbidden in the OSM terms of use. Contributors have added unrestricted datasets from

the U.S. Census and the U.S. Geographic Names Information System. The U.S. road network, rail network, water bodies, city boundary and place information was initially loaded from the U.S. Census TIGER files between August 2007 and January 2008. Loading data from a third-party could also include geocoding addresses in a third-party geocoding source.

Each of these three data source types can introduce errors into the OSM project. GPS data can have positional errors if the unit is not calibrated correctly or sensitive enough. It can be difficult to interpret the attribute information of buildings or landscapes from an aerial photo correctly. Cartographers have a long history of using map data from previous map sources. As recently as 2012, Sandy Island in the Pacific Ocean was shown to be a nonexistent island that had probably been produced by a mapping error and then propagated into multiple data sources when copied (BBC 2012). This begs the question – do different data source types exhibit differing spatial patterns of mapping?

To complicate matters, OSM is a heavily edited dataset. In the U.S. dataset used in this analysis, one feature had 289 distinct versions. Overall, 53% of the features in the datasets are revisions of previous features. Generally, a heavily edited feature suggests higher data quality in VGI, but more research needs to be done (Haklay et al. 2010; Mooney and Corcoran 2012a). Each revision of a feature may involve a separate data source. A feature initially loaded from a third-party data source could be verified using GPS trace data and then updated to match aerial imagery. Each revision may include a chaining effect of multiple data sources.

Table 3 shows a typology of data sources in OpenStreetMap. The asterisk (*) indicates that a contributor may verify Type B and C data on site. Mapping activity in

OSM can be categorized into these three data sources: A) data collected in field, B) data generated by tracing features on screen, and C) data imported from third-party sources. Of the three types, it is conceivable that each could be verified by field data collection or survey, but only *Type A: Data collected in the field* is certain to have had the physical presence of the data contributor with GPS. In this study, if a feature has multiple data source types, if one is GPS, then it is counted in Type A. If a feature has sources from aerial photos and imported data, it is counted in Type B. The reason for the hierarchical process is twofold: if a feature is noted as having multiple sources, it is unknown which information is derived from which source type. For the purposes of this project, data collected in field by an individual is considered primary as opposed to data collected from other means. In this study, Type A data was an order of magnitude smaller than other data source types.

Table 3. A Typology of Data Sources in OpenStreetMap. An * indicates that a contributor may verify Type B and C data on site.

Data Source Types	**Contributor verified information on site**	**Data initially created by contributor**
A. Data collected in the field	Yes	Yes
B. Data generated by tracing features on screen	No*	Yes
C. Data imported from third-party sources	No*	No

Table 4 lists the count of features from different data source types in OpenStreetMap. Data sources are determined by attribute tags within the dataset, including a key attribute tag ‘source.’

Table 4. Count of features in OpenStreetMap from Different Data Source Types

Data Source Type	**Count of Features**	**Percentage of Total Features**
All Features	27,133,894	100%
A: Data collected in field	384,081	1.4%
B: Data generated on screen	1,127,589	4.2%
C: Data imported	14,907,457	54.9%
Unknown data source	10,714,767	39.4%

The results of Table 3 reflect the nature of OpenStreetMap in the United States. The federal government in the US publishes most of its geospatial data without copyright restriction. This is different than other developed countries, where the percentage of Type C (imported) data would likely be much smaller compared to all data.

The goal of a project like OpenStreetMap is to provide a free and growing set of spatial data and maps. As the process of mapping is continuous, how can one identify when an area is ‘completed’? In 2011, the OpenStreetMap data for Germany was considered finished when comparing the road network in OSM to the road network in TomTom (Neis, Zielstra, and Zipf 2011). TomTom’s data is privately sourced. Of course,

OSM is not limited to a road network. In some cases, features down to individual houses and structures have been included in OSM (see Figure 5). In Pascagoula, Mississippi, the level of detail includes building numbers. This level of detail would be useful for routing and geocoding, which is a potential future use for OSM.

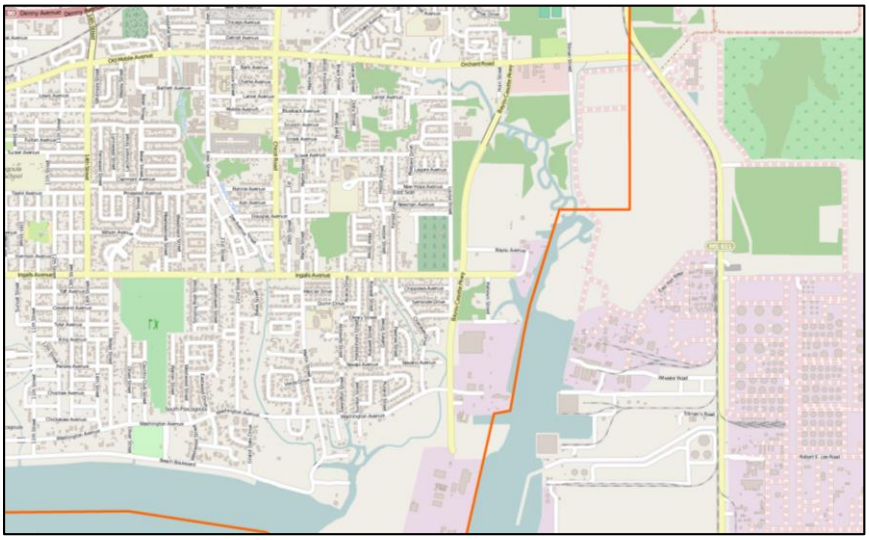

Figure 5. OpenStreetMap tile image of Pascagoula, MS. (OpenStreetMap)

What can be considered a "well-mapped" area in OSM? From the start of the OSM project, Germany has been one of the most active countries. In the United States, OSM participation is more varied geographically. There are some areas that have extremely detailed data (see Figure 5). If the goal of OSM is to have data this detailed, then we can identify these levels of detail and compare other areas accordingly. Even if there is not an end goal for the level of detail, using the concept of a "well-mapped" area may help to identify areas that need attention.

To match the definition from Neis, Zielstra and Zipf (2011), a "well-mapped" area, therefore should have a complete road network and turn-by-turn information for

drivers. This definition appears to match the original intention of OpenStreetMap as a road network dataset. OpenStreetMap, however, is more than simply a road network. Its dataset includes building outlines, walking paths, accessibility data, physical features like swamps and forests, coastlines, and even trees. A turn-by-turn road network for OSM as a definition of "well-mapped" should be a minimum. There may be no upper limit on the amount of detail that can be mapped in an area.

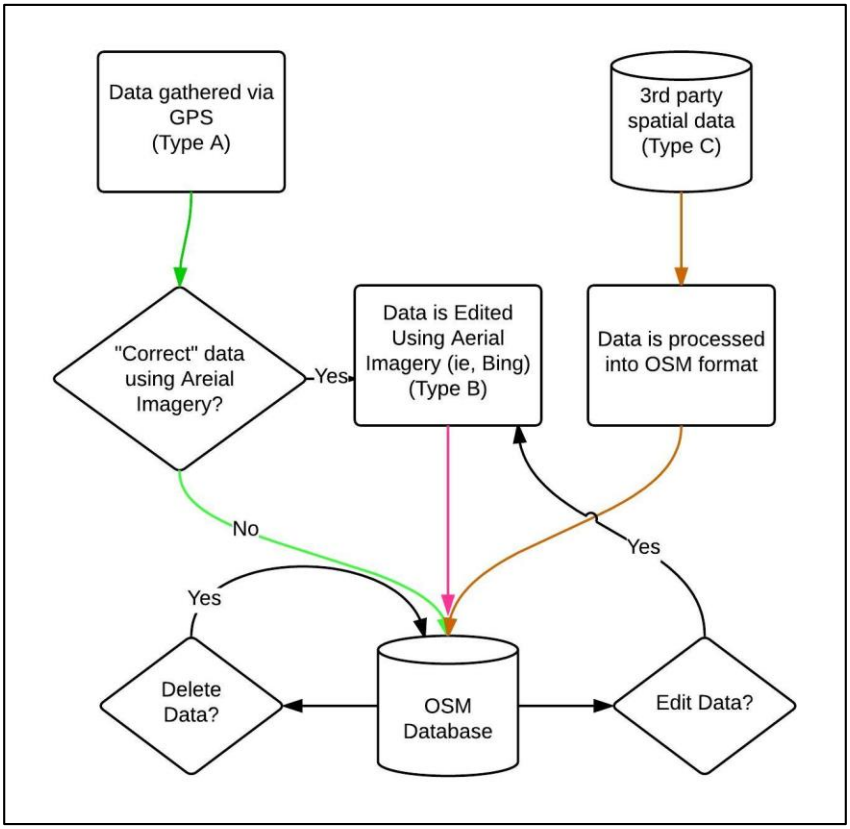

Figure 6. A Conceptual Model of Paths to Contribution in OSM.

Figure 6 presents a conceptual model and summary of the three Types of Contribution discussed in this section. Type A data is added by collecting information on the ground with GPS and follows the green line on the left. Type B data is added by

editing aerial imagery (usually under license from Bing) using an OSM editing tool like Osmosis or JOSM and follows a pink line. Some GPS data may also be edited or aligned using an OSM editing tool and aerial imagery before being uploaded into the database. Type C data is processed and brought in from a third party dataset and follows the orange line. In all cases, data can be edited or deleted from the database.

Choices of an OSM Contributor

The choice of how to contribute data into OSM is only one of the many choices that a contributor must decide. Contributors must decide what geographic areas they want to contribute to; what types of features they want to map; whether to create new features or edit existing ones; how those features should be represented (point, line, polygon); what attributes and descriptions of the features should be included; what level of accuracy is important (if it is a concern at all); whether to contribute with a group such as an OSM mapping party, a mapathon event or alone; whether to contribute all of their data as one changeset (a group of edits) or in separate changesets; which feature types to use for each feature; which and how much metadata to include; and which of the standards that OSM has set out to follow (if they are aware of them). Table 5 lists these choices in detail.

Table 5. Choices of the OpenStreetMap Contributor.

Choice Question	**Example Choice Answer**
What geographic areas should be mapped?	I will map in downtown Littleton, Colorado.
Should new features be added, or should existing ones be edited?	I will map businesses in the downtown main street.
Should GPS (Type A), aerial imagery tracing (Type B), or third-party data (Type C) be the source of new data?	I will walk the street and collect GPS points and then correct them using aerial imagery in JOSM.
How should features be represented (point, line, polygon)?	Businesses will be represented as point features.
What level of accuracy is needed?	I will place the point anyway overlaying the structure of the building.
Should data be added through a mapping party or other event or alone?	I will work with the Denver OSM group.
Should data be contributed as one large changeset or several separate changesets?	I will contribute all of today's data as one changeset.
What feature types (amenities) should be used to label features?	I will add a few "restaurant"s, a "café", and a "post office" feature.
Which standards from the OSM website should be followed?	I will use the prescribed list of amenities on the OSM website.

For other spatial data agencies, many of these questions would be codified and answered long before data collection began. A government survey and data collection team would employ standards and training to ensure data capture consistency and quality.[30] The data may also be checked and verified during when added into a spatial data repository or infrastructure. Who controls the data, processes, and edits the data will be strictly controlled. OSM allows the contributor as much flexibility as they want when creating or editing data with the provision that a critical mass of editors will correct errors that may arise in the data. Certainly, some locations in OSM have a high degree of detail and accuracy (see Figure 4 and Figure 5). The key to understanding the accuracy and completeness of OSM data, then, is to understand the process of how contributors add, edit, and delete data and how they decide to address the choices made during the process.

At each step, the choices that a contributor makes have an impact on the data. By choosing which geographic locations to map, other locations may be left out. There may be, for example, a bias towards geographic locations where people with more money (to purchase GPS equipment and computers), more free time (to enable mapping activity), more access (easier to reach over harder to reach), or are more frequently visited. Choosing to add a building as a point rather than a polygon may make the building more difficult to find or detect. The suggested list of amenities in OSM[31] lists "kindergarten," "university" and "school" but not "day care," "vocational school", "high school" or

[30] https://www.fgdc.gov/standards/standards_publications/, accessed April 12, 2015.
[31] http://wiki.openstreetmap.org/wiki/Map_Features, accessed April 12, 2015.

"elementary school." Interestingly, there are two separate categories for "library" and "public_bookcase" – a public bookcase being an outdoor library where anyone can leave or take a book. Of course, a contributor can create a new category, but it may not be recognized by OSM's tile-building software. These tiles are used by many online groups to display background mapping information.

If these choices in adding data to OpenStreetMap impact the data, they are also unevenly spatially represented in the data itself. Using the data in OSM and looking for differences in accuracy, attribute consistency, participation in OSM groups, and how mapping is related to the local population and socioeconomic characteristics of a region should present an idea of where mapping is incomplete and which areas require more attention in data collection than others.

CHAPTER V. DATA AND METHODS

Study Area

For this research, the analysis is limited to the contiguous forty-eight states and the District of Columbia of the United States of America. The primary reason for using the U.S. is that it is a large and diverse country. The U.S. Census data provides a consistent dataset with which to compare the OpenStreetMap data and eliminates most language problems in comparing tags and values related to the OpenStreetMap dataset. The Census provides data at multiple scales.

Data Sources

There are seven secondary data sources for this project:

- The 2010 U.S. Census (American Community Survey 5-year summary file) for socioeconomic data including geographic boundary data.
- The complete dataset of OpenStreetMap entries (including additions, deletes, and changes) from 2006 to 2013.
- The archive of OpenStreetMap email discussion lists.
- The OpenStreetMap user wiki.
- The state of Texas 2013 Department of Education school shapefile.
- The state of California 2012 Excel spreadsheet with latitude and longitude of school locations.

Data Processing

To answer the questions in the objective, I begin with the complete 2005-2013

OpenStreetMap dataset (including all changes) for the forty-eight contiguous states and the District of Columbia and the 2010 U.S. Census data. The 2010 U.S. Census data was chosen as it is a central dataset during the 2005-2013 timeframe. Every feature in the U.S. dataset for OpenStreetMap that fell completely within one of four U.S. Census geography types (block group, county, place, and core-based statistical areas) was counted. Choosing features that only fell within a Census area does eliminate some features that may overlap a County or Place, but it solves two problems: it eliminate overrepresentation; and it greatly simplifies the processing of the data. Choosing to represent the amount of a "partial feature" that fell within a given Census geometry required extensive computing power; a simple "within" lookup provides a simpler but still meaningful solution. A block group is a set of blocks, generally comprising a small neighborhood, and is the smallest statistical unit in the U.S. Census that includes socioeconomic data such as median household income. Counties are state-created administrative units. Census designated places are cities, towns, villages, or similar distinct jurisdictions. Core-based statistical areas (CBSAs) are defined by the U.S. Office of Management and Budget as adjacent areas with over 10,000 people that are economically linked. CBSAs replace the prior Metropolitan Statistical Area definitions. Block groups and counties cover every space in the forty-eight states and District of Columbia; CBSAs and places do not. CBSAs largely overlap with urban areas, although not strictly. Using multiple geographic scales for analysis is a potential way to recognize the effects of the Modifiable Areal Unit Problem, or MAUP, which will be discussed further in the Results. For the computation of population change, the American Community Survey 2000-2005 and 2010-2012 from the U.S. Census were tabulated to

compute county-level population change.

OpenStreetMap data is formatted in XML (eXtensible Markup Language). This file was processed using *perl* scripts to divide the data into different feature types (points, lines, polygons). Features were loaded into a Postgresql/PostGIS spatial database, which was used for storing both the spatial features of OSM and the Census. Because of the size of the dataset (over one hundred gigabytes of data), some analyses required using the parallel processing power of the Amazon Elastic Compute Cloud (EC2). This allowed faster processing but required more testing and verification. For most analysis, the statistical language *R* was used to create graphs and perform statistical calculations.

The complete OSM dataset including all accessible edits and changes were extracted from www.openstreetmap.org. All OSM data for this area from the beginning of OSM in 2004 until February of 2013 were included in the analyses. A total of 20,752 OSM contributors were included in this study.

To process the large amount of data from OSM, the eXtensible-Markup-Language OSM dataset was processed using MapReduce and the Amazon Elastic Compute Cluster (EC2). The MapReduce technique is a parallel computing model that breaks data into key-value pairs and then processes them based on key. Using contributor IDs as the keys, the OSM data were processed to derive values for individual variables for each contributor. This parallel processing greatly sped up the analysis as adding more data did not require recomputing the entire analysis.

Some data from the OpenStreetMap set was excluded from the analysis. Line and polygon features that did not include a valid topology were excluded. These may include features that had overlapped themselves without a node, or if a polygon did not close

back into the initial starting node.

Line and polygon features that did not fall completely within a Census boundary were not considered in the analysis of OpenStreetMap features compared to socioeconomic data. This was done for two reasons: it was unclear how to count lines and polygons that fell into multiple areas. Counting them twice would result in over-counting. Counting them based on the length or area that fell into a Census geographical area was also problematic, largely for programmatic reasons. This likely produced a bias against contributors who worked on long or large features. Many of these features, however, were likely imports from Census data. It should also be noted that both Census place and CBSA geographies do not encompass the entire United States area. Table 6 shows the total features for each Census geography.

Table 6. Count of features completely within each Census geography from OSM.

Geometry	OSM Points (% of total)	OSM Lines (% of total)	OSM Polygons (% of total)
OSM Features within 48 contiguous states	10,834,448 (100%)	15,030,286 (100%)	665,226 (100%)
OSM features within Census Block groups	10,782,928 (99.5%)	3,646,207 (24.3%)	621,328 (93.4%)
OSM features within Census Places	7,957,525 (73.4%)	10,512,908 (69.9%)	634,678 (95.4%)
OSM features within Census Counties	10,825,039 (99.9%)	14,452,885 (96.2%)	629,284 (94.6%)
OSM features within Census CBSAs	8,932,608 (82.4%)	12,052,312 (80.2%)	590,355 (88.7%)

The Activity-Context-Geography Model

Putting the ACG Model in Practice

Putting the ACG Model into practice begins with identifying variables related to each aspect and computing the variables for each contributor. For OSM, this means identifying the per-user variables relevant to the Activity, Context, and Geography aspects. The complete list of variables is included in Table 7, Table 8, and Table 9. For the Activity aspect, variables include the actions by a data contributor to create, modify, and delete features and their attributes. Features are individual spatial objects and are commonly represented as points, lines, or polygons. Other objects in OSM include

changesets, which are groups of related edits by one data contributor over a short period of time, and *relations*, which are logically related spatial objects. For example, a multipolygon object has two physically distinct but logically related parts (i.e. the state of Michigan) and would be brought together as a relation. Activity also includes action on attribute information, which in OSM involves key-value pairs known as *tags*. Key-value pairs may include attribute descriptors such as *key*="name", *value*="Main Street"; *key*="amenity", *value*="school"; or *key*="path", *value*="bike route." OSM was designed to be as simple as possible, and therefore there are no logical restrictions on creating tags within the database. To measure the Activity aspect, I count the total number of keys, the total number of values, and the number of different types of keys per data contributor. The Activity aspect also measures a data contributor's overall OSM data activity patterns, including the total number of days that a contributor is actively editing the database and the average number of contributions per day. There are thirty-four variables in the Activity aspect. Table 7 lists the variables in the Activity aspect with a description.

Table 7. Description of per-contributor variables used from the Activity Aspect.

Type	Per-Contributor Variables	Description
Activity	Count of points (created)	Number of point features created per user.
Activity	Count of lines (created)	Number of line features created.
Activity	Count of polygons (created)	Number of polygon features created.
Activity	Count of points (modified)	Number of point features modified (but not deleted).
Activity	Count of lines (modified)	Number of line features modified (but not deleted).
Activity	Count of polygons (modified)	Number of polygon features modified (but not deleted).
Activity	Count of keys in point features	Number of total attribute rows in point features.
Activity	Count of keys in line features	Number of total attribute rows in line features.
Activity	Count of keys in polygon features	Number of total attribute rows in polygon features.
Activity	Count of key types in point features	Count of types of attributes in point features.
Activity	Count of key types in line features	Count of types of attributes in line features.
Activity	Count of key types in polygon features	Count of types of attributes in polygon features.
Activity	Count of value types in point features	Count of different values in attributes in point features.
Activity	Count of value types in line features	Count of different values in attributes in point features.
Activity	Count of value types in polygon features	Count of different values in attributes in point features.
Activity	Count of changesets	Count of changesets, or groups of changes at one time.
Activity	Count of relations (created)	Number of relations (groups of related features) created.
Activity	Count of relations (modified)	Number of relations (groups of related features) modified.
Activity	Count of keys in changesets	Number of total attribute rows in changesets.
Activity	Count of days actively editing point features	Number of days that point features were created, modified, or deleted.
Activity	Count of days actively editing line features	Number of days that line features were created, modified, or deleted.
Activity	Count of days actively editing polygon features	Number of days that polygon features were created, modified, or deleted.
Activity	Count of days editing changesets	Number of days that changesets were created or deleted.
Activity	Count of days editing relations	Number of days that relations were created, modified, or deleted.
Activity	Points edited per day	Average number of point features edited (created, modified, or deleted) per day that the contributor was active.
Activity	Lines edited per day	Average number of line features edited (created, modified, or deleted) per day that the contributor was active.
Activity	Polygons edited per day	Average number of polygon features edited (created, modified, or deleted) per day that the contributor was active.
Activity	Changesets edited per day	Average number of changesets edited (created, or deleted) per day that the contributor was active.
Activity	Relations edited per day	Average number of relations edited (created, modified, or deleted) per day that the contributor was active.
Activity	Count of points (deleted)	Number of point features deleted.
Activity	Count of lines (deleted)	Number of line features deleted.
Activity	Count of polygons (deleted)	Number of polygon features deleted.
Activity	Count of changesets (deleted)	Number of changesets deleted.
Activity	Count of relations (deleted)	Number of relations deleted.

The Context aspect reflects the circumstances and actions that support a contributor's effort in producing VGI. These variables do not produce spatial features, but they may indicate the social and technological relationship between the contributor and the VGI platform and community. The Context variables are extracted through the OSM infrastructure including the OSM website. There are six variables in the Context aspect, such as the number of days since joining OSM, the word-count of an online diary that a contributor maintains, and the word-count of a contributor's comments on the features within the OSM dataset. High values for the context variables (above 1.5 standard deviations over the scaled average mean) may suggest contributors who are more vested or socially connected within the OSM community. There are six variables in the Context aspect. Table 8 lists the variables in the Context aspect with a description of each.

Table 8. Description of per-contributor variables used from the Context Aspect.

Type	Per-Contributor Variables	Description
Context	Days since contributor joined OSM	Number of days before March 1, 2013 that the contributor joined OSM.
Context	Avg. word count in comments of point feat.	The mean word count in comments of point features that have comments.
Context	Avg. word count in comments of line feat.	The mean word count in comments of line features that have comments.
Context	Avg. word count in comments of polygon feat.	The mean word count in comments of polygon features that have comments.
Context	Avg. word count in comments of changesets	The mean word count in comments of changesets that have comments.
Context	Avg. word count in comments of relations	The mean word count in comments of relations that have comments.
Context	Count of words in online diary	Word-count of entries in the online user diaries at http://www.openstreetmap.org/diary

The Geography aspect in the ACG model refers to the geographic extent, precision, and distribution of the features that have been acted upon by a data contributor.

This aspect represents the spatial nature of data contribution, and in the case of OSM, it is a quantitative description of the map data contributed. Example variables in the Geography aspect include the areal extent of the objects contributed by a contributor, and the average number of nodes used in creating lines and polygons. A smaller value (less than 1.5 standard deviations below the scaled mean) may indicate a specific focus on a particular region, while a larger value may indicate a wide area of interest. There are seven variables in the Geography aspect. Table 9 lists the variables in the Geography aspect with a description of each.

Table 9. Description of per-contributor variables used from the Geography Aspect.

Type	Per-Contributor Variables	Description
Geography	Areal extent of point features (km^2)	Areal extent of all of the contributor's point features.
Geography	Areal extent of line features (km^2)	Areal extent of all of the contributor's line features.
Geography	Areal extent of polygon features (km^2)	Areal extent of all of the contributor's polygon features.
Geography	Avg. nodes per line feature	The average number of nodes that a contributor used to create/modify line features.
Geography	Avg. nodes per polygon features	The average number of nodes that a contributor used to create/modify polygon features.

A total of forty-six variables were generated for each OSM data contributor following the ACG model. However, highly positively skewed data distribution dominates the OSM dataset for the forty-eight states. Ninety percent of the data contribution was made by 120 out of 20,752 users, or 0.5% of the contributors; 39.5% of the contributors had created, edited, or deleted five features or fewer. Of the forty-six variables, thirty-four belong to the Activity aspect, seven belong to the Context aspect, and five belong to the Geography aspect. Using Principal Component Analysis, I reduced

the variables to eight Activity components, four Context components, and two Geography components. For each aspect, the components were averaged together to form three aspect variables (Activity, Context, Geography) for each contributor. The process of averaging the data does reduce the explanatory power of the variables, but it allows the ACG Model to produce a three-dimensional representation of aspects for each contributor. For each contributor, the model produces one value for their Activity aspect, one value for their Context aspect, and one variable for their Geography aspect. Each data contributor can potentially be described as a point in a three-dimensional graph of the AGC Model.

Calculating areal extent was derived by the bounding box of a contributor's features, with separate values for point, line, and polygon features. The motivation behind using the bounding box of all features was to encompass the spread of contributions by an OSM user. This does introduce the potential to over represent contributors who produce a few features spread over a large area, but it does yield information regarding the spatial distribution of a contributor's input.

The variables average number of nodes per line and average number of nodes per polygon provide insight into a contributor's sophistication with GIS and the types of features that they choose to measure. Presumably, natural features such as hiking paths, rivers, and coastlines would have more nodes in a given area compared to man-made features. The average line had 18.4 nodes with a standard deviation of 45.3. The average polygon had 20.6 nodes with a standard deviation of 124.1.

Establishing Data Contributor Typology

Devising a method to identify clusters of contributors presents a challenge as

there are a small number of contributors who make an extremely large number of contributions. Traditional hierarchical clustering and k-means clusters both proved problematic as the small number of extreme cases are less similar to any other feature than themselves and tend to be grouped individually. A test run of hierarchical clustering on the raw data revealed several clusters with only one member. Since these extreme data contributors make important data contributions, dropping these cases is not an option. A second attempt to cluster contributors used a pseudolog transformation of the data. This led to reasonably-sized clusters, but some contributors with large aspect values were grouped with those with more moderate aspect values.

After a number of trials, standard deviation proved to be effective to reflect the distributions of the three aspect measures for the data contributors. An outlier in an aspect (Activity, Context, and Geography) is defined as being 1.5 standard deviations away from the mean. Using the outlier approach, there are potentially twenty-seven groups of OSM data contributors as each contributor may be a high outlier, not an outlier, or low outlier in each of the three aspects. Using an upper case to represent a high outlier, a lower case for a non-outlier, and a square-bracket letter represent a low-outlier, the groups can be labeled as acg, ac[g], acG, a[c]g, etc. This grouping scheme classifies OSM data contributors based on their measures in the three ACG aspects.

Note that the ACG model was established to grasp data contributors' characteristics that may be connected to their creditability and data quality. The next logical step is to examine if the grouping results are related to the contributors' data quality.

Assessing Data Quality

The quality of OSM data, in particular the positional accuracy, needs to be examined across the different groups of data contributors. Two government datasets were used as reference data. The first dataset is the 2011 public school location data obtained from the Texas Education Agency (TEA). It contains 8,360 school locations. The second dataset is the 2013 public school location data obtained from California Department of Education (CDOE) contains 11,234 schools. Both datasets contain point features.

To identify entries of schools from the OSM dataset, point objects with the word "School" in the title or keyed as a 'school' (usually with the key-value "amenity"="school") were extracted. A total of 14,288 point entries in California and 11,123 point entries in Texas were identified from OSM. The OSM dataset has more entries than the state government school datasets because OSM includes multiple versions of the same school, historical schools, religious schools, and commercial schools (e.g., 'yoga school') that are not administered by the state agencies.

To match schools between OSM and the reference data from TEA or CDOE, the Levenshtein distance between OSM attributed names and the names in the TEA/CDOE datasets. The Levenshtein distance measures the minimum number of changes necessary to match one string to a second string. For example, the word "pear" and the word "beat" have a Levenshtein distance of two as two letters would be changed to match the former to the latter. The Levenshtein ratio is computed as one minus the ratio of the Levenshtein length to the length of the longest name from two strings. This results in a value between zero and one on the similarity between a school name from OSM and one from the government data. A Levenshtein ratio of zero represents no similarity, while a value of 1

represents a perfect match. A school record from OSM is considered matching a record from the TEA or CDOE data when they have the highest Levenshtein ratio among all candidate matches, when the Levenshterin ratio is greater than 0.6, and when they are have a distance less than 1,340 meters apart. A preliminary data processing revealed that the closest mismatch of two schools from OSM school entries and TEA/CDOE records with exact same name are 1,340 meters apart (i.e. California Elementary in West Covina and California Elementary in Hacienda la Puente). Therefore, 1,340 meters is chosen to be the maximum possible distance between matching schools in the OSM and state agency datasets. A total of 10,744 matching schools were identified between OSM entries and government school data.

The positional accuracy of OSM school data can be assessed by the errors in school location data. The error in positional accuracy is calculated as the distance between an OSM school location and the corresponding TEA/CDOE school location. For the purpose of this report, the locations of the schools given by the state agencies are considered accurate. This was verified by some spot checking on schools through mapping against aerial photographs. Schools in this study are points, while school grounds will have a polygonal shape. This presents a possible false report of an error for small values of distance between the datasets. It is important to note that many OSM school locations were imported from GNIS (the U.S. Geographic Names Information

Systems) in the summer of 2009[32]. To better understand the role of secondary sources of data and their impact on OSM data quality, the positional accuracy was compared for the school datasets with and without secondary data. After excluding the secondary sourced OSM school entries, a total of 2,394 schools were left in the OSM dataset that were matched the schools in the TEA or CDOE dataset.

One-way analysis of variance (ANOVA) and a Tukey post-hoc test were performed to compare the positional accuracy of the OSM school entries across the different types of VGI contributors. The one-way ANOVA test was conducted for all the OSM school entries that matched a record in the government dataset and for only those OSM school entries that were generated by OSM data contributors as primary data and matched a record in the government dataset.

The Geographic Distribution of OpenStreetMap

Mapping Activity and Population

At a basic level, it's conceivable that mapping in an area is related to the population of that area. In urban areas, where population concentrations are higher, there may be more entities (restaurants, public buildings, roads, etc) considered worth mapping. Of course, man-made entities are only one type of feature to be mapped. Natural features and non-physical entities (ie, political borders) may also be mapped.

[32] http://wiki.openstreetmap.org/wiki/USGS_GNIS.

Increased population in an area also means that there are potentially more map contributors in that area either because of ease of access or sheer numbers.

If there is a significant positive relationship between mapping activity and population, this may suggest that mapping interest and "mappable" entities of interest are spread uniformly in the population. A strong correlation between population and features in OSM would also suggest where areas require more mapping activity when the ratio of entities to population is lower than the national average. This would be a convenient way to identify areas that require attention.

I use Pearson's product-moment correlation coefficient to compare the total 2010 population in each of the four Census geographies (block group, CBSA, county, and place) with the number of OSM features within that area, and the features from data sources Type A, B, and C (as in Table 3).

The hypothesis is that there is a linear relationship between the 2010 population size of a Census geographic area and the number of OSM features within that same area. Rejection of the hypothesis would suggest no relationship between population size and the coverage of ISM for a geographic area. A correlation coefficient of 0.7 or above with a significant level of $\alpha < 0.05$ would suggest a significant and strong relationship between the two.

Mapping Activity and Socioeconomic Characteristics

Mapping interest and ability may not be distributed evenly within the U.S. population. There are technological barriers to mapping for OpenStreetMap. A contributor needs a device (phone, GPS, computer) to input information; some level of education or training in mapping or using computers; the free time to spend on mapping

pursuits; and some motivation or interest in the material. In the case of pursuing field work, personal safety may also be a concern, and there may be questions of equal access for different populations.

In a survey of 426 OpenStreetMap contributors at a London gathering, Haklay and Budhathoki (2010) found that 96% were male (compared to 3% female), 32% were between the ages 20-30, 32% were between 31-40, and 22% were between 41-50. Of that group, 78% had a college degree. In an online survey of 641 readers of the academic mapping collective site *floating sheep*[33], Stephens (2013) found that male respondents were more than four times as likely to have used OSM and five times as likely to have contributed to OSM as female respondents. In the same survey, male and female respondents were equally likely to have used or contributed to Google Maps. In this survey, a majority of respondents who used OpenStreetMap (79.9%) were under 40 years old, and a majority (77.1%) had a college degree.

OpenStreetMap, then, tends to have contributors that are male, young, and educated. It may also be possible that these contributors have relatively higher incomes if they can afford computers and GPS devices to work with OpenStreetMap. To determine a relationship between OpenStreetMap and potential contributors in an area, I will perform a principal components regression using the following 2010 U.S. Census variables and the OSM feature count in each of the two U.S. Census geographies (i.e., CBSA, and

[33] http://www.floatingsheep.org

county):

- Persons with a Bachelor's Degree or higher (BS);
- Total Men (MEN);
- Total Women (WOMEN);
- Median Household Income (INC);
- Population age 25 to 54 (A25_54);
- Count of homeowners (HOWN);
- Count of persons whose primary transit to work is car (TCAR);
- Count of persons whose primary transit to work is public transit (bus, train) (TPUB);
- Count of persons whose primary transit to work is walking (TWLK).

I use principal components regression (PCR) to compare the above socioeconomic variables in the two Census geographies (CBSA and county) with the number of features within that area. The hypothesis states that there is a functional relationship between the magnitude of OSM mapping activities in a geographic area and the principal components of the socioeconomic-socioeconomic characteristics of that area. The magnitude of mapping in an area is measured as the total number of OSM mapping features. To understand how the different types of data sources are related to the socio-socioeconomic variables, a PCR model is also attempted for each data source type (A, B, and C). An R^2 above .70 suggests a model that provides a strong explanation of the relationship between the independent and dependent variables.

Principal components regression is a technique used for constructing explanatory models when there are multiple factors that are highly collinear with the independent

variables. The technique uses principal components analysis (PCA) to extract factors in orthogonal directions along the dataset. Each factor represents a combination of the original variables and explains some amount (i.e., 50%) of the correlation in the data. In order to reduce the dataset, the largest factors that account for over ninety percent of the correlation are including while other factors are removed. These factors are then used as the independent variables in an ordinary least squares linear regression.

Mapping Activity and Spatial Clustering

The previous tests have examined the spatial relationship between the persons in an area being mapped, but what are the spatial patterns within the OSM dataset itself? Mapping contributors in OSM are not bound to only map within one county or one neighborhood. Indeed, an active contributor or mapping community may impact any of the places or counties surrounding their preferred areas. This activity may result in a high degree of spatial autocorrelation. Spatial autocorrelation is a measure of the degree with which spatial features tend to cluster or be dispersed in space.

I will calculate a Moran's I statistic for spatial autocorrelation at the county level for feature density (features-per-square-mile) for each data source type (all features, types A, B, C). Feature density, rather than feature count, is preferred since counties have vastly different areas within the US. The null hypothesis is that there is no spatial autocorrelation in the distribution of OSM mapping features in the 48 states of the U.S. when examined at the county level.

If the z-score of the Moran's I falls within two standard deviations, or between -1.96 and 1.96 of the mean ($\alpha < 0.05$), the null hypothesis is accepted. In addition to the Moran's I test, I will generate a Getis-Ord Gi* hot-spot map to identify which counties

have high spatial autocorrelation when OSM features are grouped by the different data source types (all features, A, B, C, and A+B).

Mapping Activity and OSM Community Participation

Feature density, particularly the density of features that were generated from GPS or screen tracing, should indicate a high-level of OSM participation. I will identify the twenty metro areas in the U.S. that have the highest and lowest feature density. This list will be compared with the number of active contributors on the email lists of the major OSM groups in the U.S. These figures were derived from meetup.com's list of OpenStreetMap regional groups.[34] These lists are voluntary groups which hold scheduled meetings specifically for OpenStreetMap or include OpenStreetMap as a key interest of the group. The areas with highest feature density are examined in detail.

I will use a Spearman's rho rank correlation test to examine the rank of feature density in OSM and the rank of contributors in the OSM email lists in those areas. For this test, the hypothesis is that the OSM feature density in a geographic area is related to the number of the OSM data contributors in that area. A correlation coefficient of 0.7 or above with a significant level of $\alpha < 0.05$ would suggest the existence of such a relationship.

Mapping Activity and Feature Type Choices

Using the 2005-2013 OSM dataset, point, line, and polygon features that represent

[34] http://openstreetmap.meetup.com/, accessed March 15, 2015

entities are extracted and mapped at the county and CBSA area level. These features represent the most commonly mapped entity types in these geographies. The data was collected by summing all features with amenities in each county and CBSA geography and then counting the feature types (amenity) that had the highest representation in an area. For example, “parking”, “restaurant”, and “school” may be common feature types.

The OSM website lists the feature types that they consider standard and use to create tile maps. These tiles provide the background map imagery for many online sites, so having a feature type follow the OSM guidelines is important in representing features of a place. A place that has a very high occurrence of an uncommon feature type may indicate an area that is not well mapped or does not have the most common feature types.

CHAPTER VI. RESULTS

The Activity-Context-Geography Model

Typology of VGI Data Contributors

Applying Principal Components Analysis (PCA) on the forty-six variables that measure the three aspects of ACG revealed that eight factors were found to reflect 89.4% of the covariance for the Activity aspect. The variables ‘count_new_lines’ (the number of new line features created per contributor), ‘count_new_polygons’ (the number of new polygon features created per contributor), and ‘count_keys_values’ (the number of different key-value combinations in the attributes of the features) explain the greatest variance in the Activity aspect. Table 10 shows the factor loadings of the variables and the eight components in the Activity aspect.

Table 10. Factor loadings of variables in the Activity aspect. Numbers in parenthesis indicate percentage of variance explained.

Activity aspect variable	PC1 (32.4%)	PC2 (21.0%)	PC3 (11.1%)	PC4 (6.8%)	PC5 (6.1%)	PC6 (4.8%)	PC7 (4.1%)	PC8 (3.1%)
count_new_points	0.06	0.16	0.36	-0.08	0.10	-0.23	0.28	-0.09
count_new_lines	0.29	-0.11	-0.02	0.04	-0.03	0.06	-0.03	0.00
count_new_polygons	0.29	-0.07	0.01	-0.01	0.03	-0.02	0.03	0.00
count_changed_points	0.01	0.05	0.16	0.07	-0.09	0.11	-0.23	0.68
count_changed_lines	0.06	0.22	-0.07	-0.21	-0.24	0.06	-0.02	-0.01
count_changed_polygons	0.08	0.19	0.05	-0.23	0.20	-0.25	0.17	0.01
count_keys_points	0.02	0.08	0.46	0.12	-0.13	0.14	-0.09	-0.19
count_keys_lines	0.29	-0.09	-0.02	0.03	-0.05	0.06	-0.03	0.00
count_keys_polygons	0.29	-0.07	0.01	-0.01	0.02	-0.02	0.04	0.00
count_k_types_points	0.02	0.08	0.46	0.12	-0.13	0.14	-0.09	-0.19
count_k_types_lines	0.29	-0.09	-0.02	0.03	-0.05	0.06	-0.03	0.00
count_k_types_polygons	0.29	-0.07	0.01	-0.01	0.02	-0.02	0.04	0.00
count_v_types_points	0.02	0.08	0.46	0.12	-0.13	0.14	-0.09	-0.19
count_v_types_lines	0.29	-0.09	-0.02	0.03	-0.05	0.06	-0.03	0.00
count_v_types_polygons	0.29	-0.07	0.01	-0.01	0.02	-0.02	0.04	0.00
count_changesets	0.05	0.22	-0.13	0.46	0.01	0.01	0.20	0.00
count_new_relations	0.05	0.22	-0.06	-0.16	-0.36	-0.06	0.22	0.00
count_changed_relations	0.05	0.22	-0.12	-0.18	-0.43	0.09	0.04	-0.01
count_keys_changesets	0.04	0.23	-0.13	0.46	0.01	-0.01	0.17	0.00
active_days_points	0.08	0.29	-0.03	-0.11	0.23	0.09	-0.27	0.00
active_days_lines	0.08	0.30	-0.04	-0.10	0.21	0.10	-0.26	-0.01
active_days_polygons	0.10	0.29	0.00	-0.14	0.24	0.02	-0.18	-0.01
active_days_changesets	0.08	0.31	-0.05	0.02	0.22	0.10	-0.19	-0.01
active_days_relations	0.07	0.32	-0.10	0.01	0.03	0.12	-0.07	0.00
points_per_day	0.01	0.04	0.24	0.06	-0.08	-0.07	0.04	0.63
lines_per_day	0.28	-0.10	-0.02	0.05	-0.05	0.01	-0.06	0.00
polygons_per_day	0.11	-0.03	0.00	0.02	-0.02	-0.11	-0.10	-0.03
changesets_per_day	0.01	0.03	-0.02	0.18	-0.16	-0.53	-0.36	-0.07
relations_per_day	0.01	0.05	-0.02	0.10	-0.21	-0.55	-0.36	-0.07
count_deleted_points	0.06	0.15	0.22	-0.14	0.18	-0.33	0.36	0.13
count_deleted_lines	0.29	-0.09	-0.02	0.02	-0.05	0.06	-0.03	0.00
count_deleted_polygons	0.28	0.01	0.04	-0.08	0.10	-0.12	0.10	0.00
count_deleted_changesets	0.05	0.22	-0.13	0.46	0.01	0.01	0.20	0.00
count_deleted_relations	0.05	0.24	-0.10	-0.19	-0.43	0.04	0.11	-0.01

Four variables described 65.2% of the Context aspect. The variables

‘avg_word_count_comment_polygon’, ‘avg_word_count_point’, and ‘avg_word_count_line’ explained the greatest Context variance. These variables report the average word count in the optional comment section per feature per respective contributor. These are not attributes of an OSM feature but rather suggested by the OSM editors as a method for data contributors to explain their choices when adding, editing, or removing features. A contributor who is an active commenter represents someone familiar with the OSM ethos and infrastructure, and therefore would have a higher measurement for Context aspect. Table 11 shows the factor loadings of the variables in the Context aspect.

Table 11. Factor loadings of variables in the Context aspect. Numbers in parenthesis indicate percentage of variance explained.

Context Aspect Variables	**PC1 (18.6%)**	**PC2 (17.1%)**	**PC3 (15.3%)**	**PC4 (14.2%)**
days_since_join	0.32	-0.63	0.02	0.01
avg_word_count_comments_point	0.53	0.24	0.35	0.08
avg_word_count_comments_line	0.44	0.15	-0.48	0.06
avg_word_count_comments_polygon	0.57	0.27	0.26	0.05
avg_word_count_comments_changeset	-0.20	0.66	-0.05	-0.27
avg_word_count_comments_relation	0.18	0.05	-0.76	0.07
count_words_online_diary	0.17	-0.14	-0.03	-0.95

Two variables described 71.7% of the variance in the Geography aspect. The variables ‘areal_extent_lines’ and ‘areal_extent_polygons’ explained the greatest variance in the Geography aspect. Each of these variables represent the extent for all of the features a contributor has made. The values reflect the geographic extent that a contributor has mapped. Table 12 shows the factor loadings of the variables in the Geography aspect.

Table 12. Factor loadings of variables in the Geography aspect. Numbers in parenthesis indicate percentage of variance explained.

Geography aspect variables	PC1 (48.2%)	PC2 (23.5%)
area_extent_points	0.56	0.10
area_extent_lines	0.58	0.05
area_extent_polygons	0.57	0.10
avg_nodes_per_line	0.10	-0.70
avg_nodes_per_polygon	0.10	-0.70

The PCA generated scores for each of the three aspects were then used to classify the OSM data contributors. A high outlier score is one that is 1.5 standard deviations above the mean. A low outlier score is 1.5 standard deviations below the mean. OSM data contributors were also grouped into clusters based on being a low outlier, a high outlier, or not an outlier for each of the three ACG aspects. A total of twenty clusters emerged (Table 13). For the OSM dataset, 39.9% of the contributors had outlier characteristics in one or more aspect. The largest group was S1 (i.e. group *acg*) with 11,273 (54.3%) contributors; as would be expected, this is a no-outlier cluster in all three aspects. The smallest groups each had one contributor, including S6 (group *a[c]G*) indicating a Low Context and a High Geography aspect and S18 (group *A[c]g*) which indicates a high Activity aspect and a low Context aspect.

Table 13 shows the contributors per cluster type (n=20,752). “Low” represents low outliers (see text for explanation); “High” represents high outliers. A "-" represents a nonoutlier. The contributors account for OpenStreetMap data within the contiguous United States from the inception of OpenStreetMap until February, 2013. Combinations of aspect values not represented had no contributors.

Table 13. Contributors per Cluster Type.

Group ID	Key	Contributors (% of Total)	Activity Aspect	Context Aspect	Geography Aspect
S1	acg	11,273 (54.3%)	-	-	-
S2	ac[g]	3,722 (17.9%)	-	-	Low
S3	acG	1,688 (8.1%)	-	-	High
S4	a[c]g	50 (0.2%)	-	Low	-
S5	a[c][g]	33 (0.2%)	-	Low	Low
S6	a[c]G	1 (0.0%)	-	Low	High
S7	aCg	33 (0.2%)	-	High	-
S8	aC[g]	8 (0.0%)	-	High	Low
S9	aCG	46 (0.2%)	-	High	High
S10	[a]cg	386 (1.9%)	Low	-	-
S11	[a]c[g]	32 (0.2%)	Low	-	Low
S12	[a]cG	554 (2.7%)	Low	-	High
S13	[a][c]g	2 (0.0%)	Low	Low	-
S14	[a]Cg	4 (0.0%)	Low	High	-
S15	[a]CG	32 (0.2%)	Low	High	High
S16	Acg	1,599 (7.7%)	High	-	-
S17	AcG	1,184 (5.7%)	High	-	High
S18	A[c]g	1 (0.0%)	High	Low	-
S19	ACg	15 (0.1%)	High	High	-
S20	ACG	89 (0.4%)	High	High	High

Figure 7 shows a 3D plot with the average aspect variables for each cluster as the center of a circle where the diameter proportionally represents the number of contributors in that cluster. The aspect variables have been converted to z-scores for readability. The cluster *ACG*, which has outliers in each of the three aspects, is not shown for readability. Each cluster circle is centered on its average Activity, Context, and Geography aspect variables adjusted to a standardized z-score. The diameter of each circle corresponds to the number of contributors in each cluster. Cluster ACG is not shown. An upper-case letter indicates high-outlier aspect variables. An aspect in square brackets (ie, [a]) represents a low-outlier aspect variable. Lower-case letters represent aspect variables that

are not outliers. Table 14 reports the average number of features created, modified, and deleted in each of the 20 clusters.

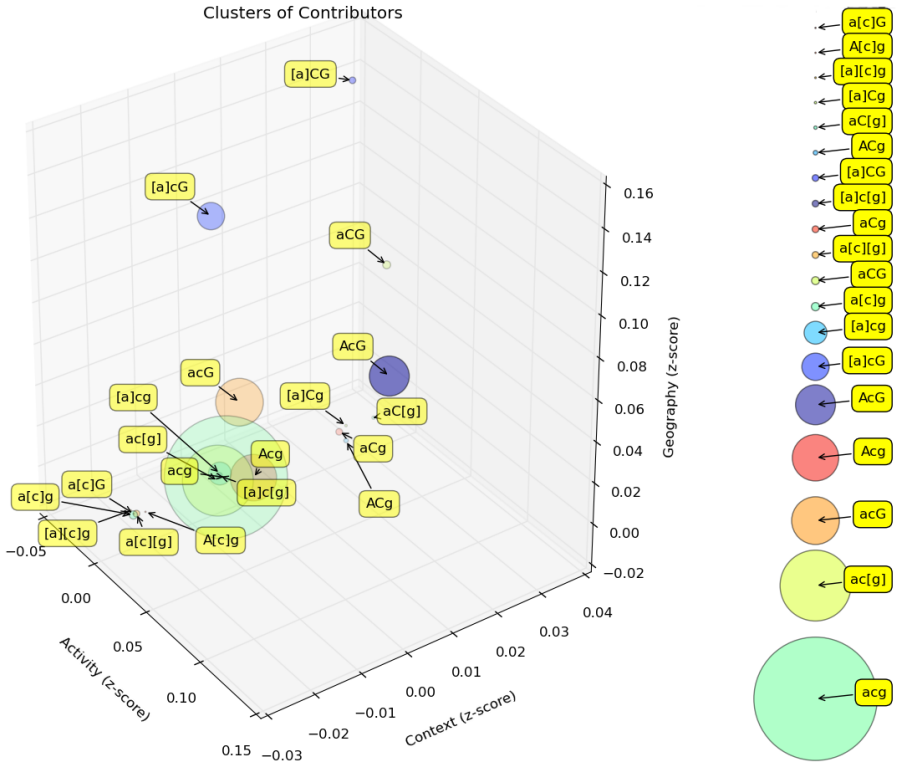

Figure 7. A scatterplot of OSM clusters of contributors.

Table 14. The average number of features created, modified, and deleted for each contributor type.

Group ID	Key	Avg. Features Created	Avg. Features Modified	Avg. Features Deleted
S1	acg	6.5	4.3	6.0
S2	ac[g]	0.8	0.3	0.4
S3	acG	44.5	54.4	72.5
S4	a[c]g	3.5	2.2	1.8
S5	a[c][g]	0.7	0.3	0.2
S6	a[c]G	1.0	1.0	2.0
S7	aCg	2.6	3.8	4.6
S8	aC[g]	0.6	0.4	1.0
S9	aCG	36.7	46.7	63.9
S10	[a]cg	435.3	27.3	408.3
S11	[a]c[g]	1.6	7.2	8.2
S12	[a]cG	8844.9	849.2	9299.0
S13	[a][c]g	10.5	3.0	13.5
S14	[a]Cg	44.3	28.8	53.5
S15	[a]CG	9197.9	1701.4	10564.4
S16	Acg	191.8	65.0	235.2
S17	AcG	11378.7	2312.4	8611.7
S18	A[c]g	4.0	14.0	8.0
S19	ACg	94.3	173.6	250.6
S20	ACG	29109.1	18607.7	35032.5

Data Quality and Positional Accuracy

For the 10,744 matching schools between the OSM dataset and the government datasets, a one-way analysis of variance (ANOVA) was run in two combinations: all matching schools grouped using the contributor type (S1 through S20); and all matching schools excluding any secondary data uploaded to OSM grouped by the contributor type (S1 through S20). For the analyses, groups S4, S5, S6, S7, S8, S9, S11, S13, S14, S18,

and S19 did not contribute any school data; these data contributors and their OSM contributions were excluded from the positional accuracy analyses.

Table 15 shows the mean distance and count of school features by Outlier cluster groups. Clusters without matching features in the dataset (S4, S5, S6, S7, S8, S9, S11, S13, S14, S18, and S19) have been omitted from the analysis. An asterix (*) indicates a p-value which is significant at or below the 0.5 level.

Table 15. The mean distance and count of school features by outlier cluster groups.

		Positional Error for All School Features (Error < 1340 meters, n=10774) One-way ANOVA F=2.2713*					**Positional Error with Imported Features Removed (Error < 1340 meters, n=2394)** One-way ANOVA F=2.3914*				
Group ID	**Key**	**Features**	**Mean Error (m)**	**SD**	**Min**	**Max**	**Features**	**Mean Error (m)**	**SD**	**Min**	**Max**
S1	acg	431	169.2	218.4	0.8	1302.7	429	169.1	218.7	0.8	1302.7
S2	ac[g]	26	130.7	147.7	3.9	611.2	26	130.7	147.7	3.4	611.2
S3	acG	172	198.0	253.7	3.9	1334.8	171	198.9	254.1	4.0	1334.8
S10	[a]cg	39	168.1	138.5	8.6	622.4	39	168.1	138.6	8.6	622.4
S12	[a]cG	394	146.6	173.9	1.7	1323.1	393	146.8	174.1	1.7	1323.1
S15	[a]CG	31	88.5	91.2	0.4	502.0	31	88.5	91.2	0.4	502.0
S16	Acg	69	145.2	140.4	4.2	593.5	69	145.2	140.4	4.2	593.5
S17	AcG	732	145.1	195.5	1.4	1245.3	729	145.4	195.8	1.4	1245.3
S20	ACG	8880	153.6	191.2	0.6	1333.1	507	143.7	181.1	0.6	1288.0

Table 16 shows the results of several one-way ANOVA tests. Each test is performed with features grouped within their outlier status within the aspect (High Outlier (H), Low Outlier (L), Not an Outlier (N)). A single asterisk (*) indicates a p-value which is significant at or below the 0.5 level. Double asterisks (**) indicates a p-value

which is significant at or below the .1 level. Note that in the data used for the positional accuracy testing, there were no contributors in the non-outlier Activity aspect nor in the low Context aspect who contributed school locations.

Table 16. One-way ANOVA results for each aspect grouped by outlier.

		Positional Error for All School Features (Error < 1340 meters, n=10774)					**Positional Error with Imported Features Removed (Error < 1340 meters, n=2394)**				
		One-way ANOVA					One-way ANOVA				
Aspect	**Outlier Group**	**Features**	**Mean Error (m)**	**SD**	**F**	**p-value**	**Features**	**Mean Error (m)**	**SD**	**F**	**p-value**
Activity	H	9681	152.9	191.1	2.1	.146	1305	144.7	187.5	4.8	.028*
	L	1093	162.3	204.0			1089	162.5	204.3		
Context	H	8911	153.4	190.9	0.3	.588	538	140.5	177.6	3.1	.078**
	N	1863	156.1	200.0			1856	156.4	200.3		
Geography	H	10209	153.3	192.0	1.3	.281	1831	149.3	193.0	1.7	.195
	L	26	130.7	147.7			26	130.7	147.7		
	N	539	166.0	205.0			537	165.9	205.3		

The Geographic Distribution of OpenStreetMap

Mapping Activity and Population

In Table 17, the results of the Pearson product-moment correlation between the 2010 U.S. Census population in four geographies and the number of features that fall within those geographies are shown. In addition to all features in each area, data source types A (data imported from GPS), B (data traced on screen), and C (data imported from

third-party source) are included in the analysis. Correlations marked with double asterisks (**) are significant below the 0.01 level. Only two correlations, between county and CBSA geographies and type A data source features, were not significant.

Three correlations met the criteria for strongly correlated results. The correlations between county geographies and CBSA with type C (imported) data source features, and the correlation between CBSA geographies and all features were above the .7 correlation cut-off.

Table 17. Results of Pearson correlation between mapping activity and population.

U.S. Census 2010 geographic area	Count (in analysis)	Avg. Population (2010 U.S. Census)	Avg. Number of Features (OSM 2005-2013)	Correlation with the number of features in area	Correlation with the count of Type A features in area (1.4% of all features)	Correlation with the count of Type B features in area (4.1% of all features)	Correlation with the count of Type C features in area (54.9% of all features)
Block group	216331	1395.7	181.0	.043**	.007**	.019**	.059**
County	3109	97118.2	8727.5	.517**	.017	.187**	.719**
Place	52756	7278.3	446.7	.483**	.016**	.101**	.643**
CBSA	933	302972.0	23124.6	.719**	.019	.108**	.843**

Mapping Activity and Socioeconomic Characteristics

Mapping activity for features that fall completely within counties and Core-Based Statistical Areas (CBSA) is tested using Principal Components Regression (PCR) with the socioeconomic variables taken from the 2010 American Community Survey / U.S. Census. Figure 8 (county level) and Figure 9 (CBSA level) show scatterplots between the independent variables in the regression analysis. The variables are listed diagonally.

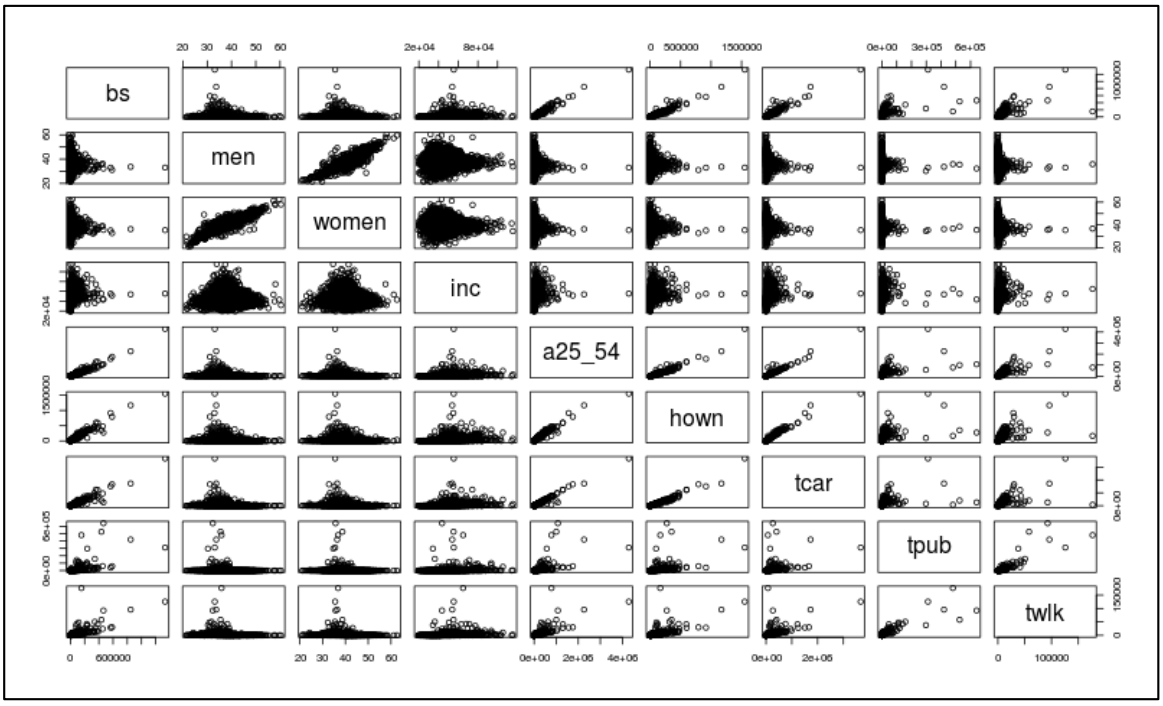

Figure 8. Scatterplots of the independent variables (county level).

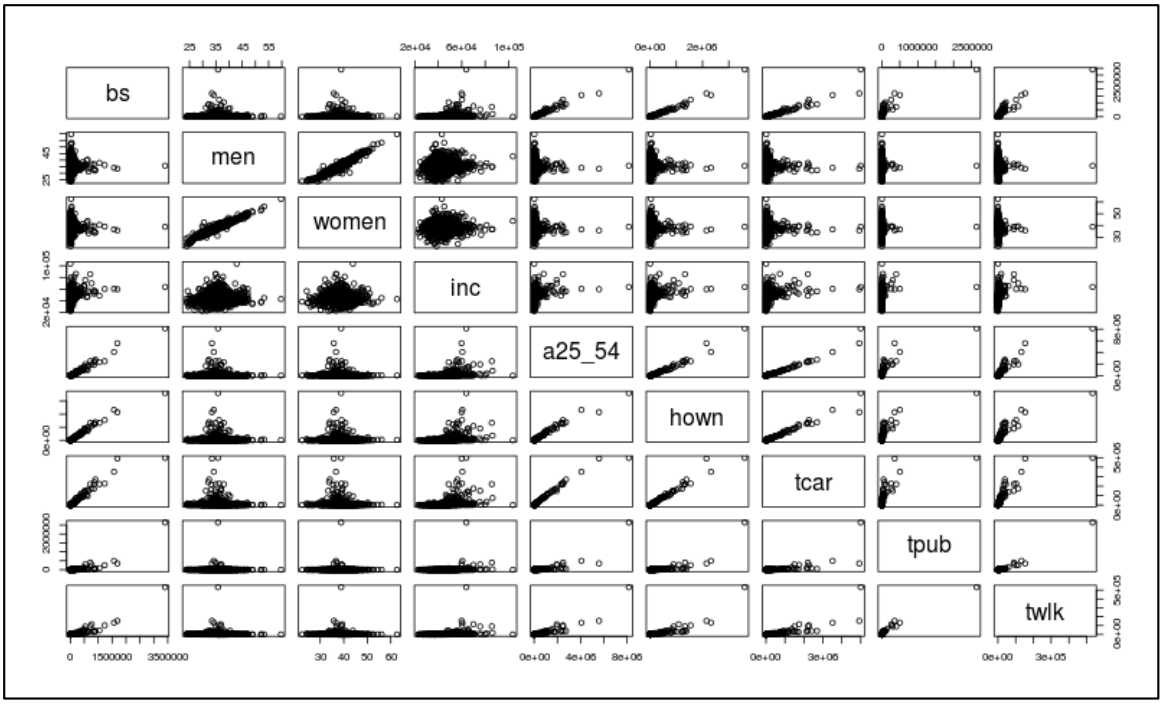

Figure 9. Scatterplots of the independent variables (CBSA level).

Table 18 shows the distributions (mean, min, max, standard deviation) of the variables used in the analysis at the County level. Table 19 presents the correlations between the independent variables in the analysis at the county level. The variables MEN and WOMEN are highly correlated, as are BS, A25_54, HOWN, and TCAR. Table 20 shows the relationship between the socio-socioeconomic variables and the four components that were derived from a Principal Components Analysis (PCA) for the County level. These four components account for 97.4% of the total variance in the dataset. Table 21 reports the results of four PCRs between the independent variables listed above the following dependent variables: the count of all features at county level; the count of features of Type A (features loaded from GPS) at county level; the count of features of Type B (features traced on screen) at county level; and the count of features of

Type C (features imported from other data sources) at county level. Features of unknown data source were included in "All Features" but not included in either Type A, B, or C. A single asterisk (*) indicates significance at .05 or below. Double asterisks (**) indicates significance at .01 or below. Note that Type A (GPS) features are not significantly explained by the PCR. Of the three source types, Type A appear to have a pattern that is least explained by socioeconomic data.

Table 18. Distribution of Variables in PCA (County level).

Variable	Mean	Min	Max	SD
BS	18502.39	15	1332186	48485.53
MEN	38.65	21.4	60.4	4.98
WOMEN	41.15	21.1	62.6	4.90
INC	44106.45	19351	115574	11439.00
A25_54	40478.09	35	4292605	134846.52
HOWN	24337.93	10	1552091	64335.25
TCAR	38421.50	29	3671019	118365.05
TPUB	2196.42	0	641106	21788.96
TWLK	1255.63	0	176569	5615.90

Table 19. Correlations between independent variables (county level).

	BS	MEN	WOMEN	INC	A25_54	HOWN	TCAR	TPUB
MEN	-.176							
WOMEN	-.178	.921						
INC	.260	-.089	-.160					
A25_54	.975	-.181	-.188	.270				
HOWN	.974	-.183	-.189	.320	.970			
TCAR	.959	-.189	-.196	.291	.981	.979		
TPUB	.601	-.081	-.082	.124	.607	.504	.448	
TWLK	.742	-.145	-.152	.194	.776	.694	.664	.866

Table 20. Relationship between independent variables and the Principal Components (county level).

Variable	**Component 1 (57.3% of variance)**	**Component 2 (20.3% of variance)**	**Component 3 (10.9% of variance)**	**Component 4 (8.9% of variance)**
BS	.425	.077	-.053	.190
MEN	-.125	.688	-.122	-.034
WOMEN	-.129	.690	-.062	.026
INC	.150	-.076	-.746	-.639
A25_54	.430	.074	-.046	.171
HOWN	.419	.059	-.169	.232
TCAR	.413	.050	-.184	.304
TPUB	.307	.129	.491	-.516
TWLK	.372	.099	.341	-.334

Table 21. Principal Components Regression results for County features.

PCR Coefficients	All Features	Type A Features	Type B Features	Type C Features
INTERCEPT	8727.53**	123.54	362.7	4794.9**
COMPONENT 1	5602.3**	53.32	965.6**	2645.1**
COMPONENT 2	-63.76	-48.11	164.6	99.7
COMPONENT 3	-3759.49**	-77.52	-469.6*	-1774.4**
COMPONENT 4	3388.29**	-113.77	1155.6**	1788.5**
R^2	.281**	.000	.035**	.556**
F-statistic	305.2	.861	28.24	972.1
Number of Observations	3109	3109	3109	3109

Table 22 shows the distributions (mean, min, max, standard deviation) of the variables used in the analysis at the CBSA level. Table 23 shows the relationship between the socio-socioeconomic variables and the three components that were derived from a Principal Components Analysis (PCA) for the County level. These three components account for 93.8% of the total variance in the dataset. Table 24 reports the make-up of the components that resulted in the PCA between the socioeconomic variables at the CBSA level. Table 25 reports the results of four Principal Components Regressions between the independent variables listed above the following dependent variables: the count of all features at CBSA level; the count of features of Type A (features loaded from GPS) at CBSA level; the count of features of Type B (features traced on screen) at CBSA level; and the count of features of Type C (features imported from other data sources) at CBSA level. Features of unknown data source were not included. A single asterisk (*) indicates significance at .05 or below. Double asterisks (**) indicates significance at .01 or below.

Table 22. Distribution of Variables in PCA (CBSA level).

Variables	Mean	Min	Max	SD
BS	56204.5	1288	3405436	173018.04
MEN	36.4	23.8	59.7	4.39
WOMEN	39.1	22.4	62.6	4.54
INC	44598.9	22881	103643	8844.56
A25_54	126936.1	4314	8146832	446372.22
HOWN	74976.4	2350	3609384	222458.72
TCAR	120419.1	3396	4993657	367155.80
TPUB	7280.2	0	2642480	91389.11
TWLK	3894.4	29	533170	20356.96

Table 23. Correlations between independent variables (CBSA level).

	BS	MEN	WOMEN	INC	A25_54	HOWN	TCAR	TPUB
MEN	-.041							
WOMEN	-.052	.966						
INC	.309	.056	-.015					
A25_54	.985	-.068	-.081	.326				
HOWN	.985	-.055	-.069	.349	.985			
TCAR	.960	-.077	-.092	.355	.982	.985		
TPUB	.798	-.017	-.018	.166	.775	.718	.646	
TWLK	.701	-.045	-.050	.257	.019	.850	.800	.965

Table 24. Relationship between independent variables and the Principal Components (CBSA level).

Variable	Component 1 (61.7% of variance)	Component 2 (21.8% of variance)	Component 3 (10.3% of variance)
BS	.419	-.026	.042
MEN	-.034	.706	-.012
WOMEN	-.040	-.703	.063
INC	.157	-.058	-.934
A25_54	.419	-.005	.011
HOWN	.413	-.014	-.039
TCAR	.403	.005	-.079
TPUB	.358	-.044	.288
TWLK	.400	-.027	.178

Table 25. Principal Components Regression results for CBSA features.

PCR Coefficients	All Features	Type A Features	Type B Features	Type C Features
INTERCEPT	23124.6**	411.3	1172.8	12535.1**
COMPONENT 1	17872.0**	104.1	961.7**	9065.7**
COMPONENT 2	1293.7	10.5	501.9	358.0
COMPONENT 3	-8215.9	-543.8	-1737.0*	-4543.3**
R^2	.531**	.003	.014**	.708**
F-statistic	350	.983	4.5	749.3
Number of Observations	933	933	933	933

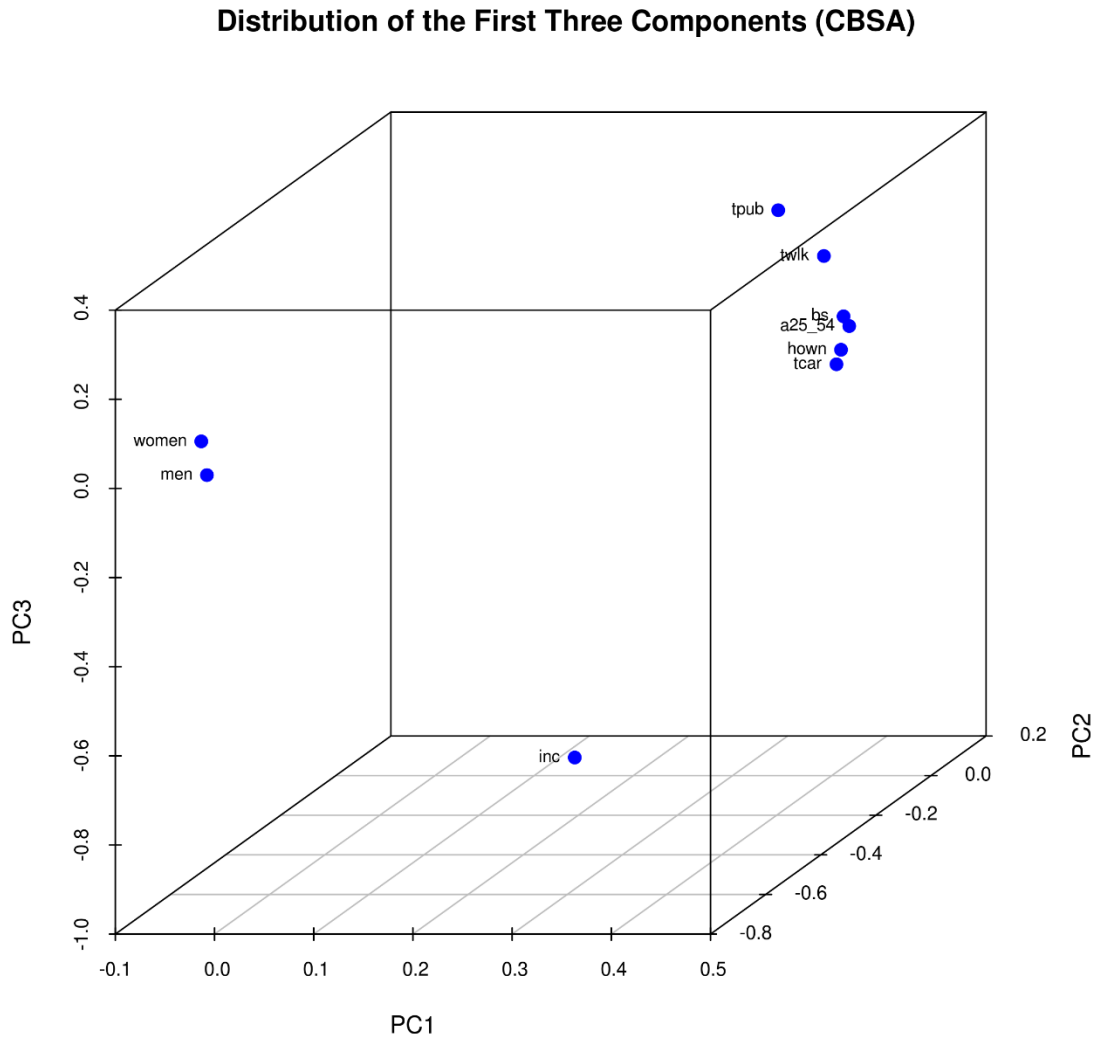

Figure 10. Distribution of the first three components with the variables of analysis. (CBSA level).

Figure 10 above shows the relationship between the variables of analysis and the first three components (dimensions) of the PCA. The variables MEN and WOMEN are highly correlated and therefore appear closely together. The variables HOWN, TCAR,

A25_54, and BS are closely related and also appear closely together. INC is unique compared to the other variables.

The following graphs (Figure 11, Figure 12, Figure 13, and Figure 14) display the fit of the comparison of responses in the Principal Components Regression for the predicted values and the original feature counts (all features, Type A features, Type B features, and Type C features) at the county level.

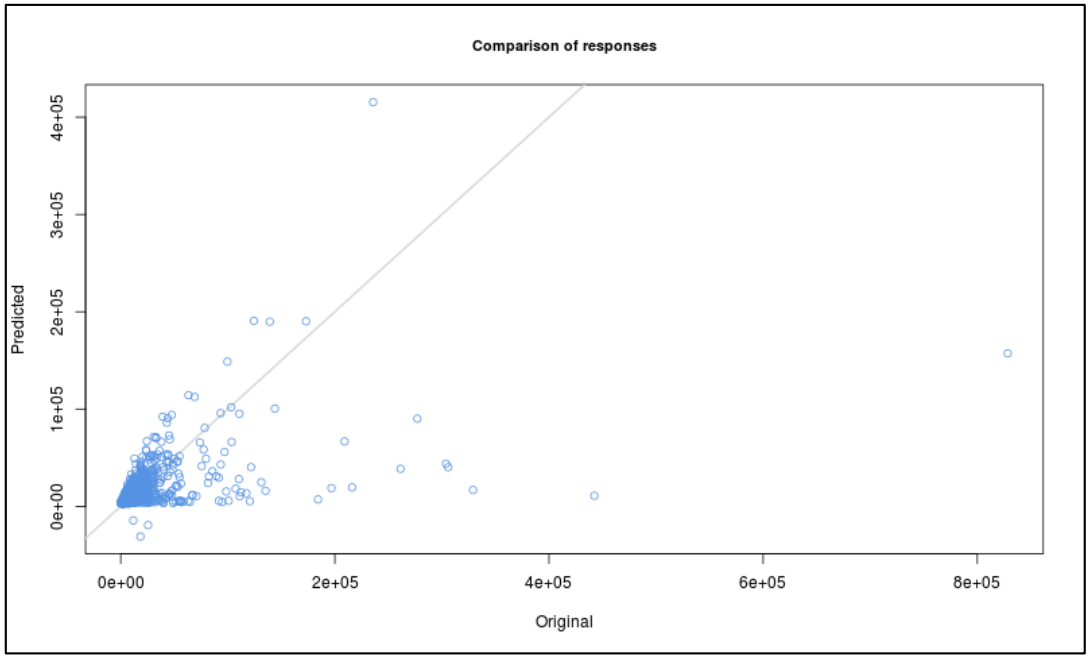

Figure 11. The comparison of responses between the predicted results of the Principal Components Regression and the original count of all features at the county level.

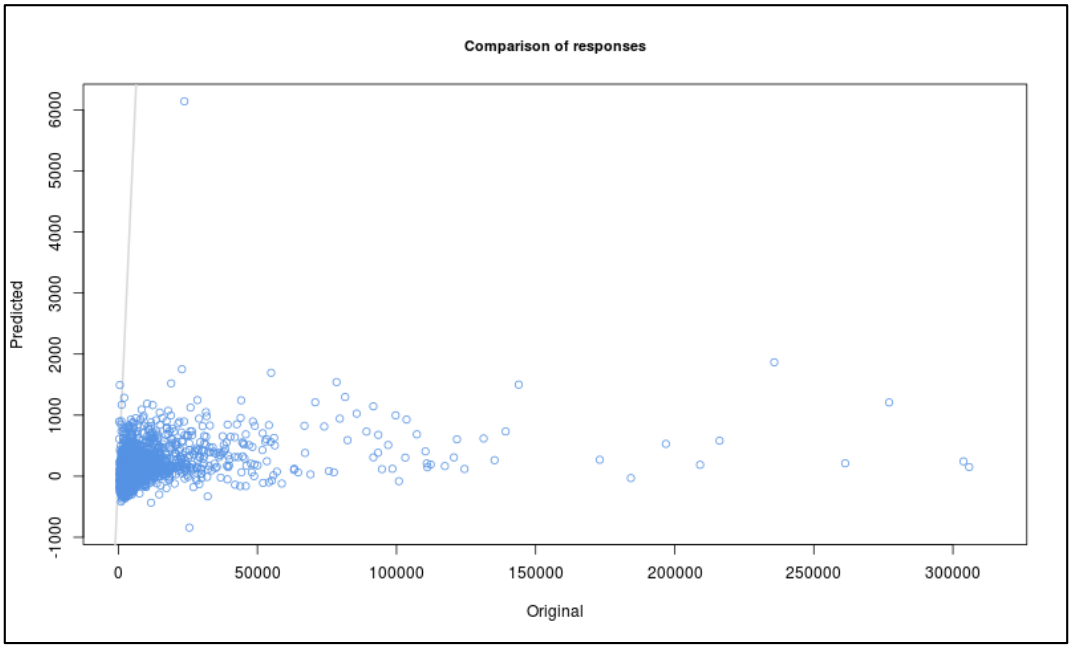

Figure 12. The comparison of responses between the predicted results of the Principal Components Regression and the original count of Type A features at the county level.

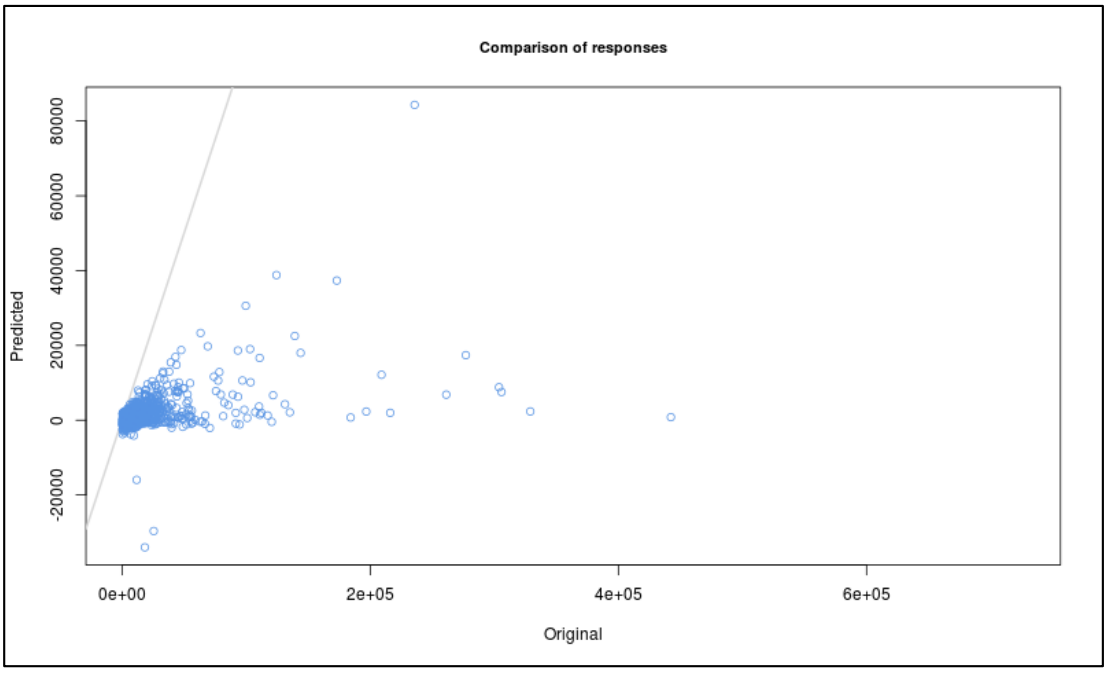

Figure 13. The comparison of responses between the predicted results of the Principal Components Regression and the original count of Type B features at the county level.

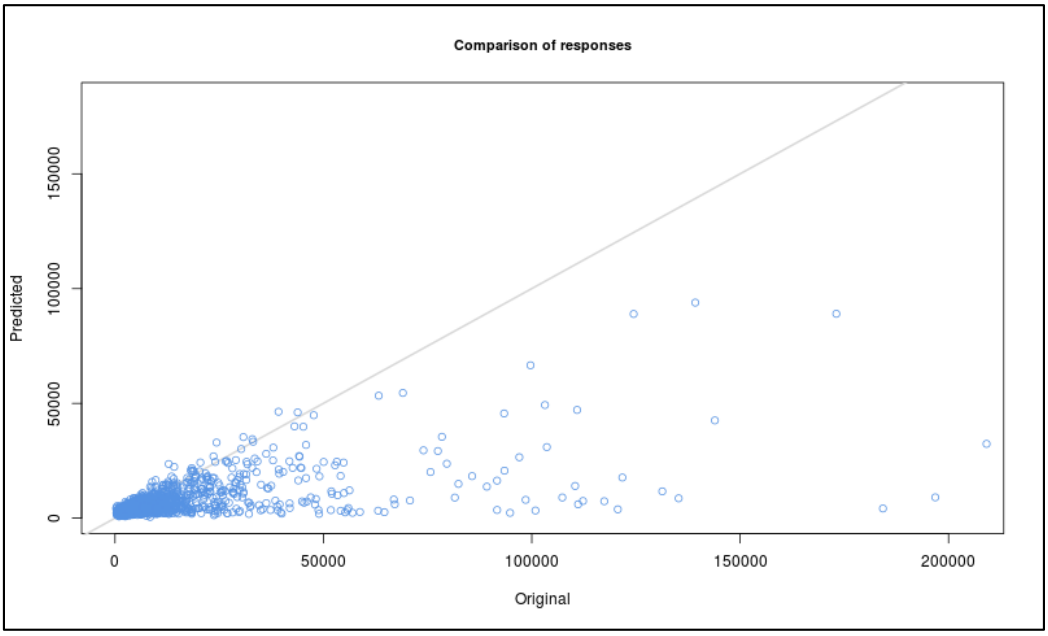

Figure 14. The comparison of responses between the predicted results of the Principal Components Regression and the original count of Type C features at the county level.

The following graphs (Figure 15, Figure 16, Figure 17, and Figure 18) display the fit of the comparison of responses in the Principal Components Regression for the predicted values and the original feature counts (all features, Type A features, Type B features, and Type C features) at the core-based statistical area (CBSA) level.

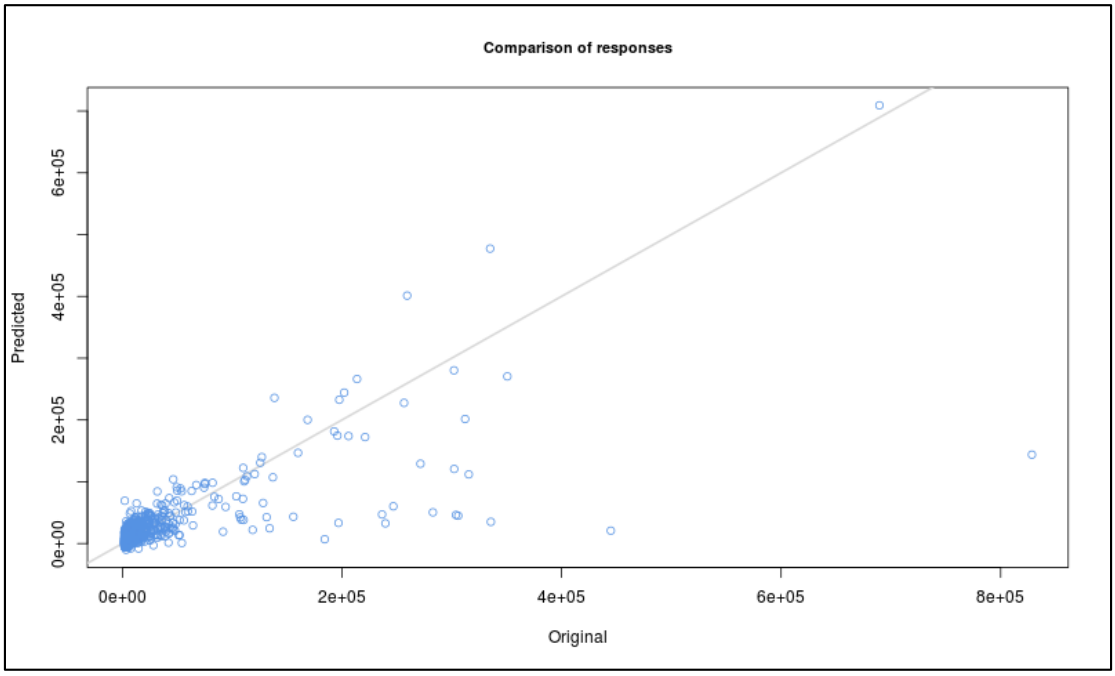

Figure 15. The comparison of responses between the predicted results of the Principal Components Regression and the original count of all features at the CBSA level.

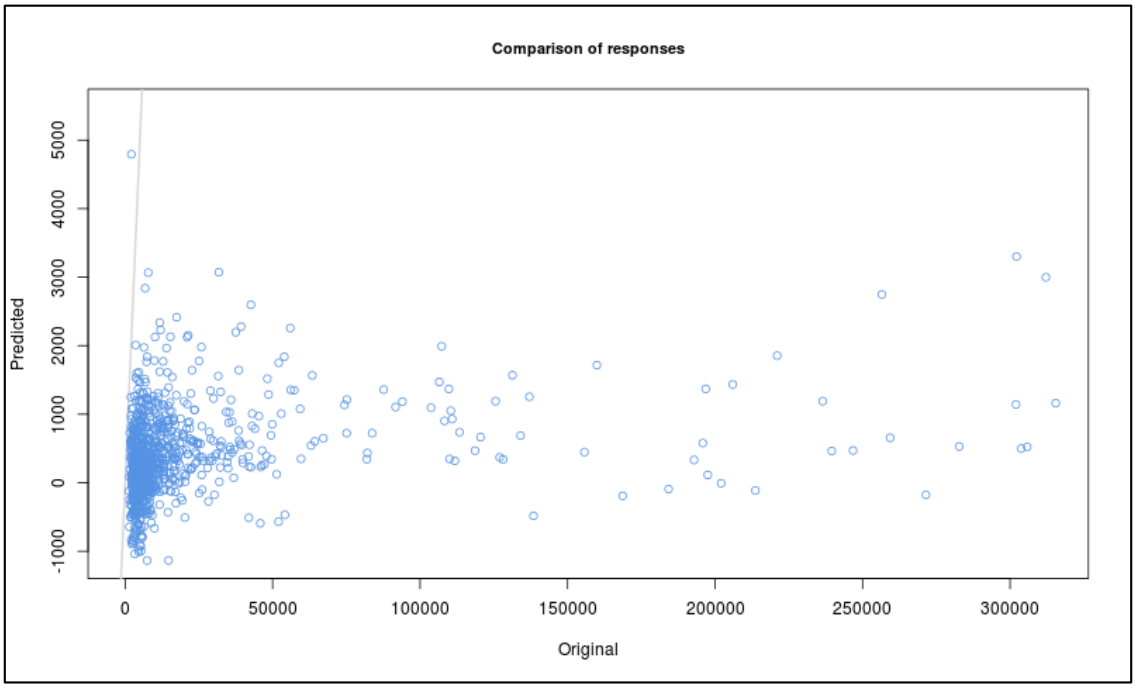

Figure 16. The comparison of responses between the predicted results of the Principal Components Regression and the original count of Type A features at the CBSA level.

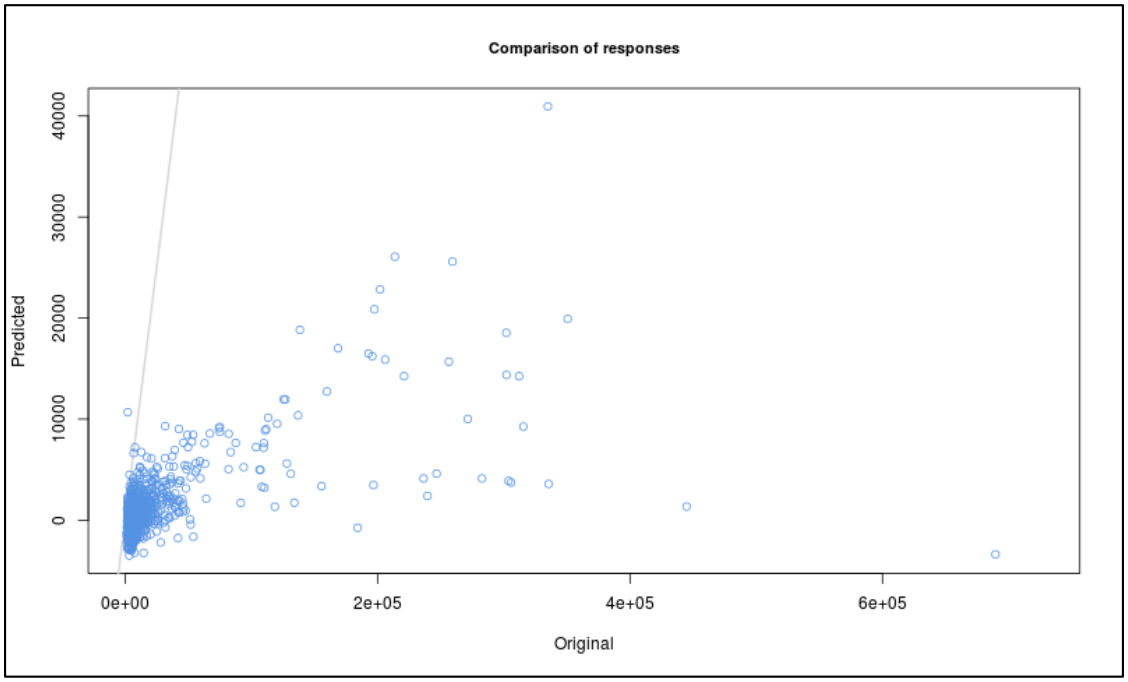

Figure 17. The comparison of responses between the predicted results of the Principal Components Regression and the original count of Type B features at the CBSA level.

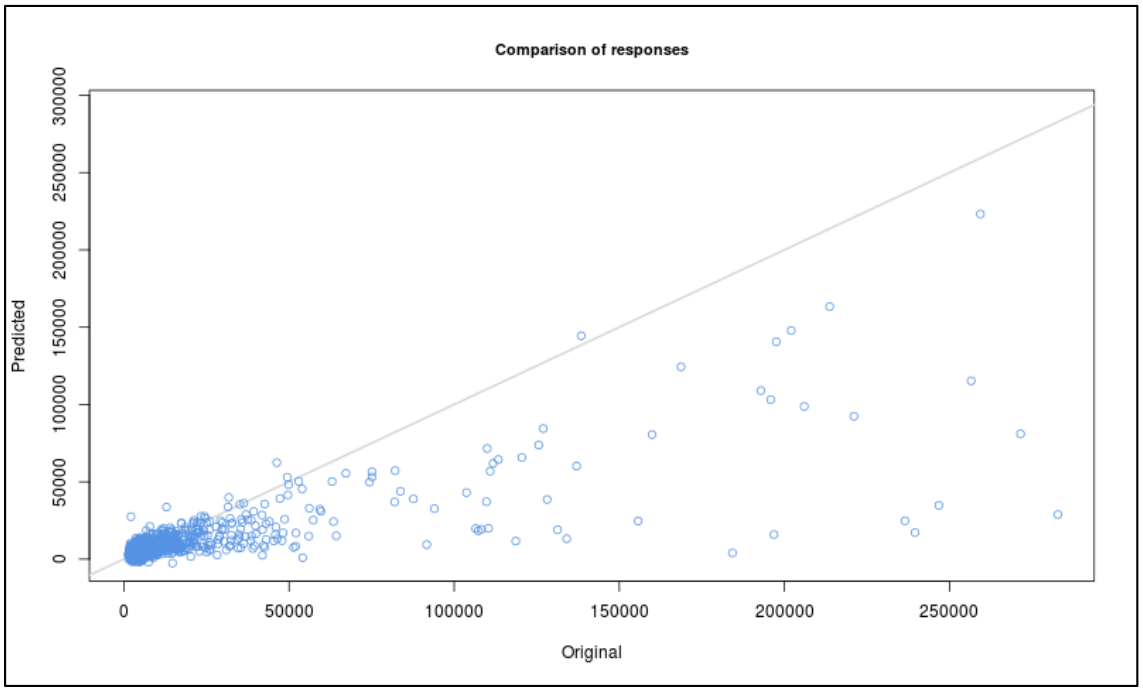

Figure 18. The comparison of responses between the predicted results of the Principal Components Regression and the original count of Type C features at the CBSA level.

Mapping Activity and Spatial Clustering

At the county level, a Moran's I index for spatial autocorrelation was run for all features, features of data source Type A (GPS created features), features of data source Type B (features traced from imagery), and features of data source Type C (features from a third-party data source).

Table 26 shows the results of the Morans I test for spatial autocorrelation at the county level. For three of the set of features (all features, Type B & C features), the results were significant below the .05 cutoff. Of these three, all are positively spatially correlated indicating a degree of clustering. Adjacent neighbors (including both edge and vertex neighbors) where used as the spatial relationships between features for grouping neighbors for the spatial autocorrelation test.

Table 26. Results of Moran's I Test for Spatial Autocorrelation at the county level.

Set of Features	Moran's I	Z-Score	p-value
All Features	.282**	28.834	0
Type A features	-.0002	.1823	.855
Type B features	.0277**	5.448	0
Type C features	.295**	29.407	0

Figure 19 shows four quantile maps of the density of features (per square mile) in each of 3109 counties in the US. The top-left map shows the quantiles of feature density for all features in OSM. The top-right map shows quantiles of feature density for Type A (GPS-traced) features. The bottom-left map shows quantiles of feature density for Type B

(traced from photos) features. The bottom-right map show quantiles of feature density for Type C (imported from other sources) features

Figure 20 shows four hot-spot (Getis-Ord Gi*) maps of feature density at the county level. The top-left map shows the hot-spot locations for all features within each county. The top-right map shows the hot-spot locations for features of Type A (GPS traces) within each county. The bottom-left map shows the hot-spot locations for features of Type B (photo traces) within each county. The bottom-right map shows the hot-spot locations for features of Type C (imported) within each county.

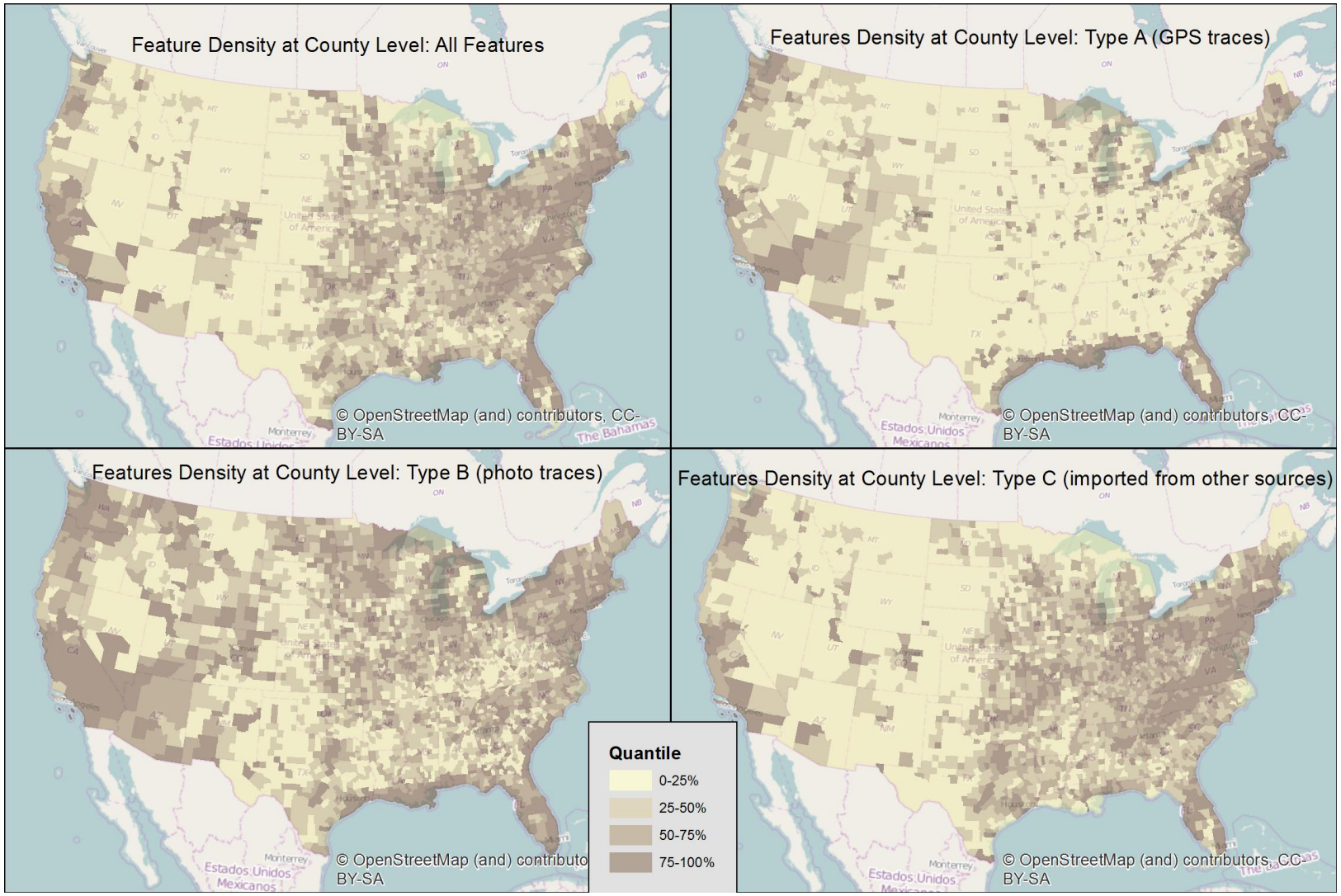

Figure 19. Density of Features at the County Level.

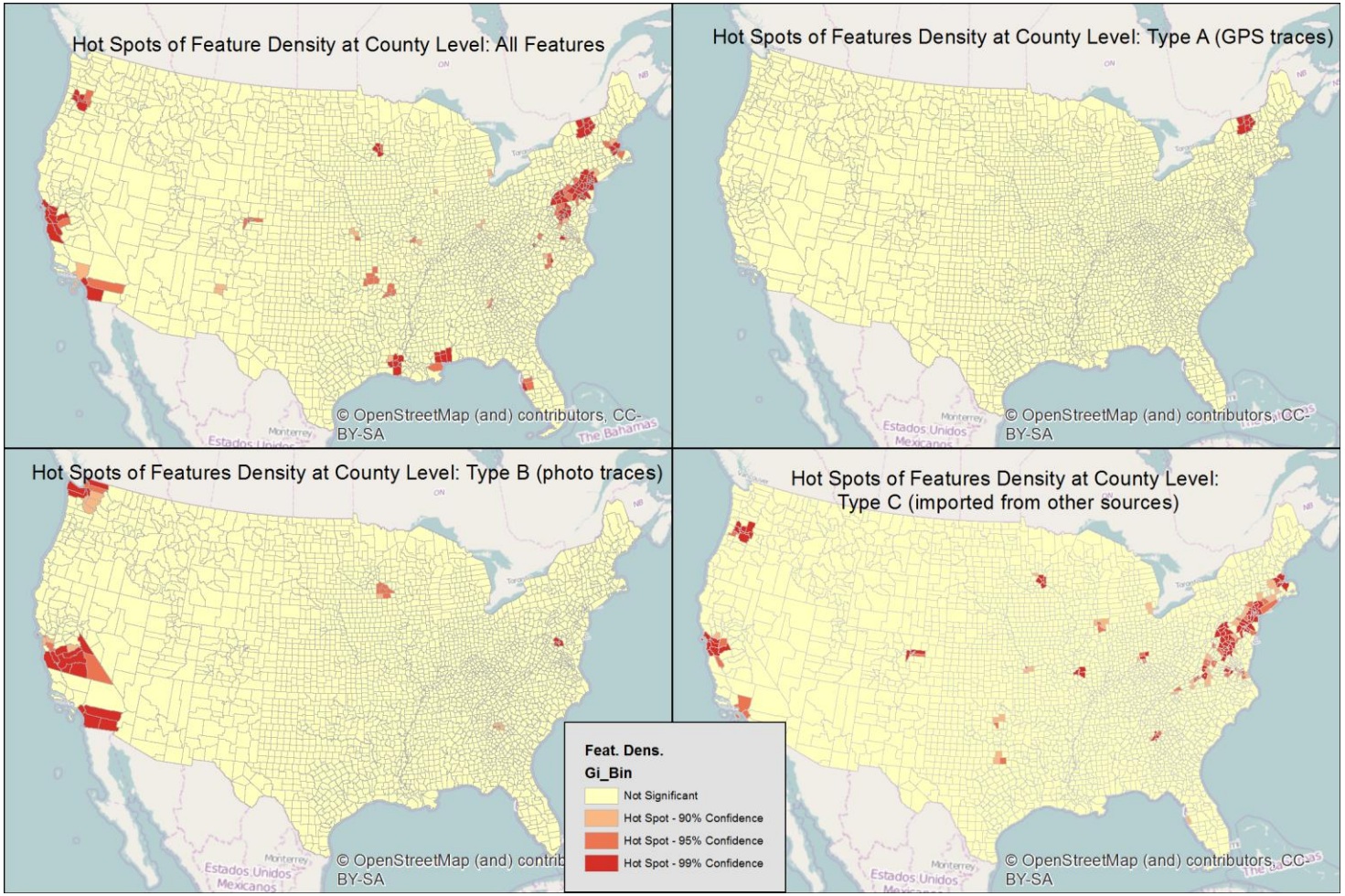

Figure 20. Hot Spot Maps of Feature Density at County Level.

Mapping Activity and OSM Community Participation

The twenty metropolitan CBSAs defined in the 2010 U.S. Census with the highest mapping activity (feature) density within their areas are shown in Table 27. The density in square kilometers is shown with the 2010 population and the total number of features in OSM (2005-2013). Table 28 shows the lowest twenty metro areas in feature density.

Table 27. Metro areas with highest density of mapping activity.

CBSA Metro Name	Mapping Features per km^2	Features in OSM	Mapping Features per capita	Population (2010)
Pascagoula, MS	123.5	444,864	2.8	159,143
Burlington, VT	97.1	335,608	1602.9	209.381
San Diego, CA	78.6	828,466	.27	3,022,460
Santa Cruz, CA	62.2	107,326	.42	256,901
Harrisburg, PA	55.8	236,486	.44	541,758
Reading, PA	48.8	110,399	.27	407,310
San Jose, CA	39.7	302,189	.17	1,793,888
San Francisco, CA	34.0	312,151	.07	4,244,889
Tampa Bay, FL	33.3	271,489	.10	2,745,350
Trenton, NJ	32.8	17.446	.00	364,445
New York, NY	32.4	689,549	.04	18,700,715
Fayetteville, AR	27.9	239,550	.54	445,626
Allentown, PA	25.9	94,009	.02	5,911,638
Philadelphia, PA	27.1	302,024	.05	5,911,638
Durham-Chapel Hill, NC	25.9	108,276	.22	488,508
Gulfport-Biloxi, MS	25.4	118,633	.49	241,122
Washington, DC	25.4	350,629	.06	5,416,691
Los Angeles, CA	25.4	334,924	.03	12,723,781
Boston, MA	24.8	256,529	.06	4,489,250
San Luis Obispo, CA	23.6	196,822	.74	265,577

Table 28. Metro areas with lowest density of mapping activity.

CBSA Metro Name	Mapping Features per km^2	Features in OSM	Mapping Features per capita	Population (2010)
Yuma, AZ	.466	5966	31.9	190,526
El Centro, CA	.720	7512	22.4	168,052
Rapid City, SD	.726	10435	11.8	123,078
Flagstaff, AZ	.844	36290	3.6	131,824
Lake Havasu City, AZ	1.03	31947	6.2	199,177
Billings, MT	1.09	11782	13.1	154,044
Lewiston, ID	1.10	3778	15.9	60,249
Cheyenne, WY	1.12	6848	13.0	89,221
Casper, WY	1.15	14200	5.2	73,520
Lake Charles, LA	1.16	8238	23.8	196,414
Wenatchee, WA	1.17	13141	8.2	108,155
Greeley, CO	1.23	11311	21.5	242,860
Prescott, AZ	1.27	23817	8.8	209,260
Visalia, CA	1.34	14929	28.8	429,404
Yakima, WA	1.39	13802	17.1	236,542
Bangor, ME	1.40	11488	13.3	152,934
Amarillo, TX	1.52	12947	18.9	245,177
Pueblo, CO	1.57	8626	18.1	156,244
Boise, ID	1.59	43010	13.9	598,730
Kennewick, WA	1.61	11221	21.2	238,406

Table 29 displays the metro areas with active OSM mapping clubs. The number of members in the club (as posted on their *meetup.com* website [35]) is shown. A Spearman correlation by ranks returns a .55 correlation between the number of members in the local OSM club and the feature density when excluding areas with no or an unknown number of mappers. The density of features of Type A (GPS traces), Type B (screen traces), and Type C (imported data) have Spearman correlations of .50, .46, and .46 with the number of members in OSM clubs, respectively.

[35] http://meetup.com, accessed March 1, 2015

Table 29. Metro Areas with OSM Mapping Clubs.

CBSA Metro Name	Features per km^2	Population (2010)	No. of Contributors in OSM database (2005-2013)	No. of Members of OSM Club (2015)	Rank in Metro Areas of Mapping Activity (Feature Density
San Francisco, CA	39.7	4,244,889	1674	430	8
Tampa Bay, FL	34.0	2,745,350	542	106	9
Philadelphia, PA	27.1	5,911,638	1178	170	14
Los Angeles, CA	25.4	12,723,781	1497	184	18
Boston, MA	24.9	4,489,250	1210	312	19
Baltimore, MD	22.5	2,683,160	605	79	21
Portland, OR	20.1	2,170,801	842	37	23
Detroit, MI	17.4	4,345,978	465	98	27
Seattle, WA	15.1	3,356,089	1288	309	37
Atlanta, GA	10.0	5,125,113	803	148	63
Oklahoma City, OK	9.95	1,218,920	251	31	64
Miami, FL	9.26	5,478,869	528	55	76
Austin, TX	7.35	1,627,571	574	193	108
Denver, CO	6.54	2,464,415	898	240	128
Phoenix, AZ	5.68	4,080,707	579	679	154
Portland, ME	5.49	513,139	210	25	162
Louisville, KY	5.17	1,261,825	212	39	171
Salt Lake City, UT	4.81	1,090,848	416	72	181
San Antonio, TX	4.69	2,057,782	383	16	186
Bend, OR	2.14	154,568	124	53	315

A Spearman correlation by ranks returns a .72 correlation between the number of mappers in an area and the feature density. The density of features of Type A (GPS traces), Type B (screen traces), and Type C (imported data) have Spearman correlations of .64, .62, and .71 with the number of members in OSM clubs, respectively. The correlation between number of members of an OSM club and the number of contributors in the same area is .83.

Metro areas (the 383 largest CBSAs) with OSM mapping clubs had four times the number of features mapped on average (174,583.1) than areas without mapping clubs (40,662.9). However, these metro areas had a lower rate of mapping per capita (.051 per person in areas with OSM mapping clubs compared to .075 per person in areas without OSM mapping clubs).

Mapping Activity and Feature Type Choices

Figure 21 is a map showing the most frequently mapped entity types per metro area in the U.S. using point features. Table 30 lists the details of the count of point feature representations at the metro level. Figure 22 is a map showing the most frequently mapped entity types per metro area in the U.S. using line features. Table 31 lists the details of the count of line feature representations at the metro level. Figure 23 is a map showing the most frequently mapped entity types per metro area in the U.S. using polygon features. Table 32 lists the details of the count of polygon feature representations at the metro level. In the case where an entity type is "Not Defined", it means that the most common feature in that area does not have an "amenity" tag defined. All tags in OpenStreetMap are optional. One such area was Abilene, Texas – as of February 2013, it had no polygons within its Census area that had amenity tags.

Table 30. Most Frequent Entities Mapped in Metro Areas (Point Features).

Entity Type	**Count of Metro Areas with this Type as Most Frequent**
Place of Worship	254
School	84
Parking	6
Restaurant	5
Bicycle Parking	4
Grave Yard	3
Fire Hydrant	2
Fuel	2
Fire Station	1
Library	1
Post Office	1

Table 31. Most frequent entities mapped in metro areas (Line features).

Entity Type	**Count of Metro Areas with this Type as Most Frequent**
Parking	138
Not Defined	135
School	37
University	30
College	6
Place of Worship	5
Grave Yard	3
Hospital	2
Public Building	2
Bank	1
Casino	1
Library	1
Marketplace	1
Theatre	1

Table 32. Most frequent entities mapped in metro areas (Polygon features).

Entity Type	**Count of Metro Areas with this Type as Most Frequent**
Parking	181
None Defined	90
University	25
School	22
Fast Food	10
Place of Worship	6
College	5
Restaurant	4
Bank	3
Grave Yard	2
Library	2
Post Office	2
Swimming Pool	2

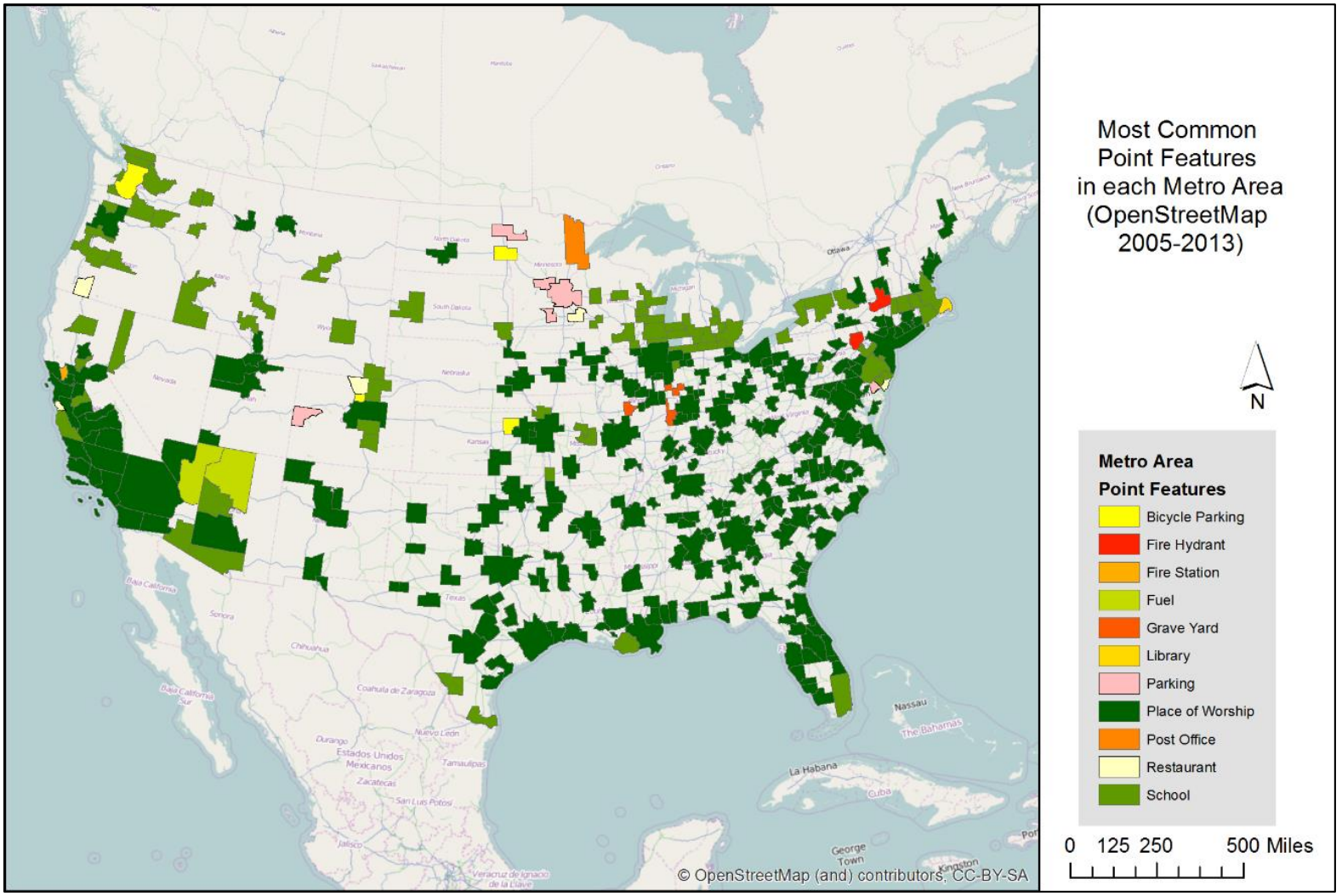

Figure 21. Map of Most Common Entities Mapped in Each Metro Area (Point Features).

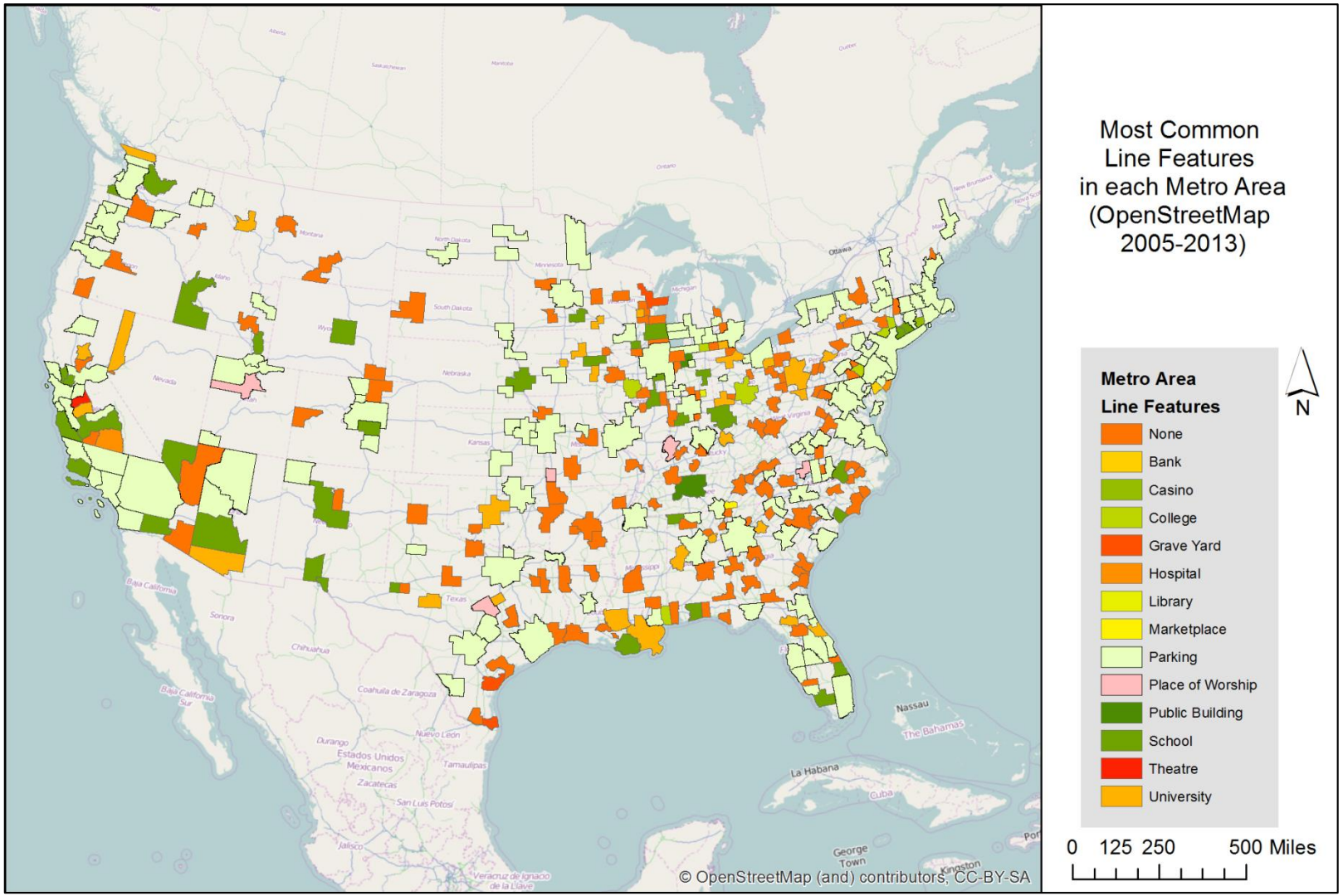

Figure 22. Most frequently mapped entities by metro area (Line features).

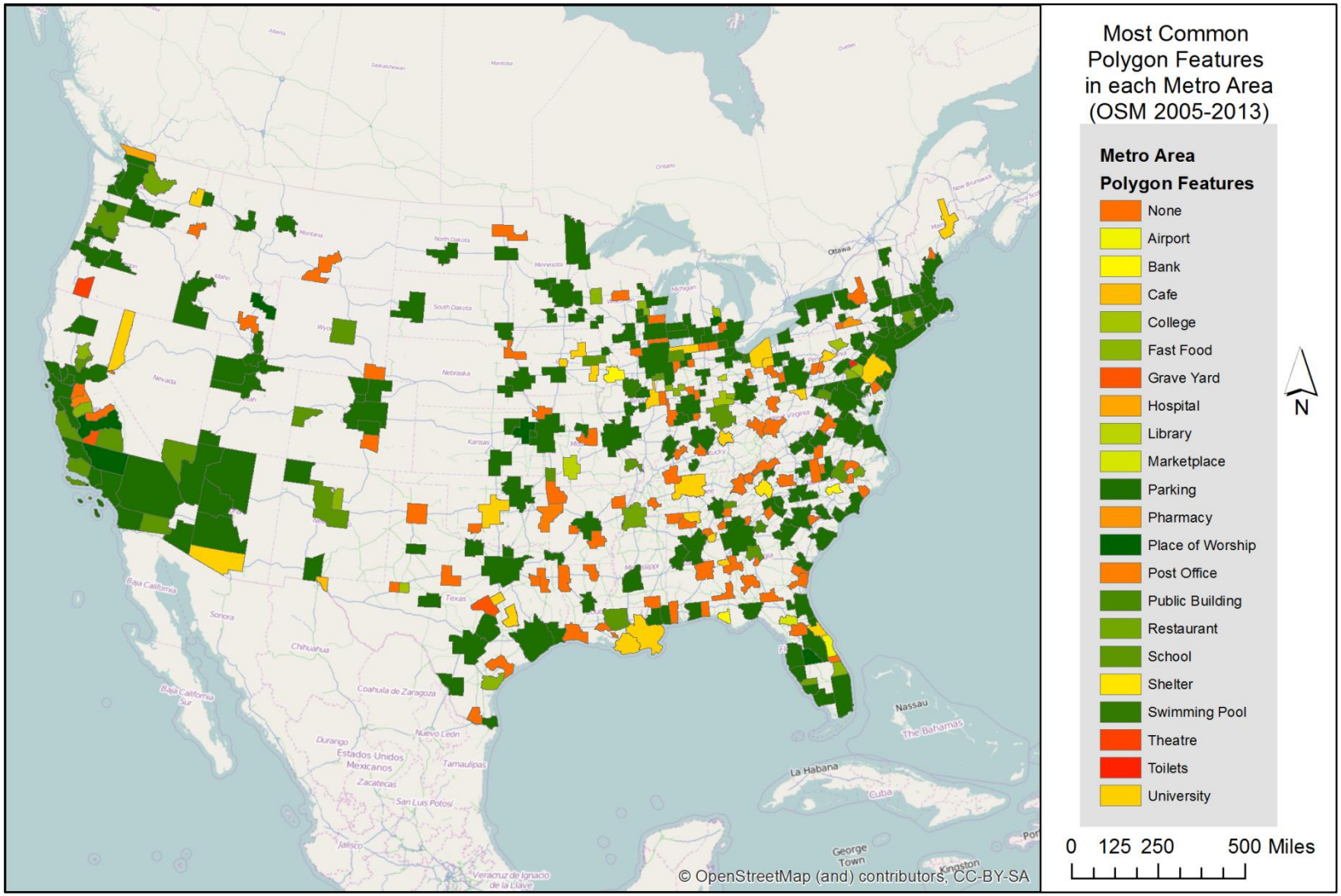

Figure 23. Most frequently mapped entities by metro area (Polygon features).

Figure 24 is a map showing the most frequently mapped entity types per county in the U.S. using point features. Table 33 lists the details of the count of point feature representations at the county level. Figure 25 is a map showing the most frequently mapped entity types per county in the U.S. using polygon features.

Table 33. Most frequent entities mapped in counties (Point features).

Entity Types	**Count of Counties with this Type as Most Frequent**
Place of Worship	2031
School	536
Grave Yard	335
Parking	77
Post Office	73
Restaurant	11
Fuel	9
Fire Station	7
Fire Hydrant	6
Bicycle Parking	5
Library	4
Toilets	4
Townhall	3

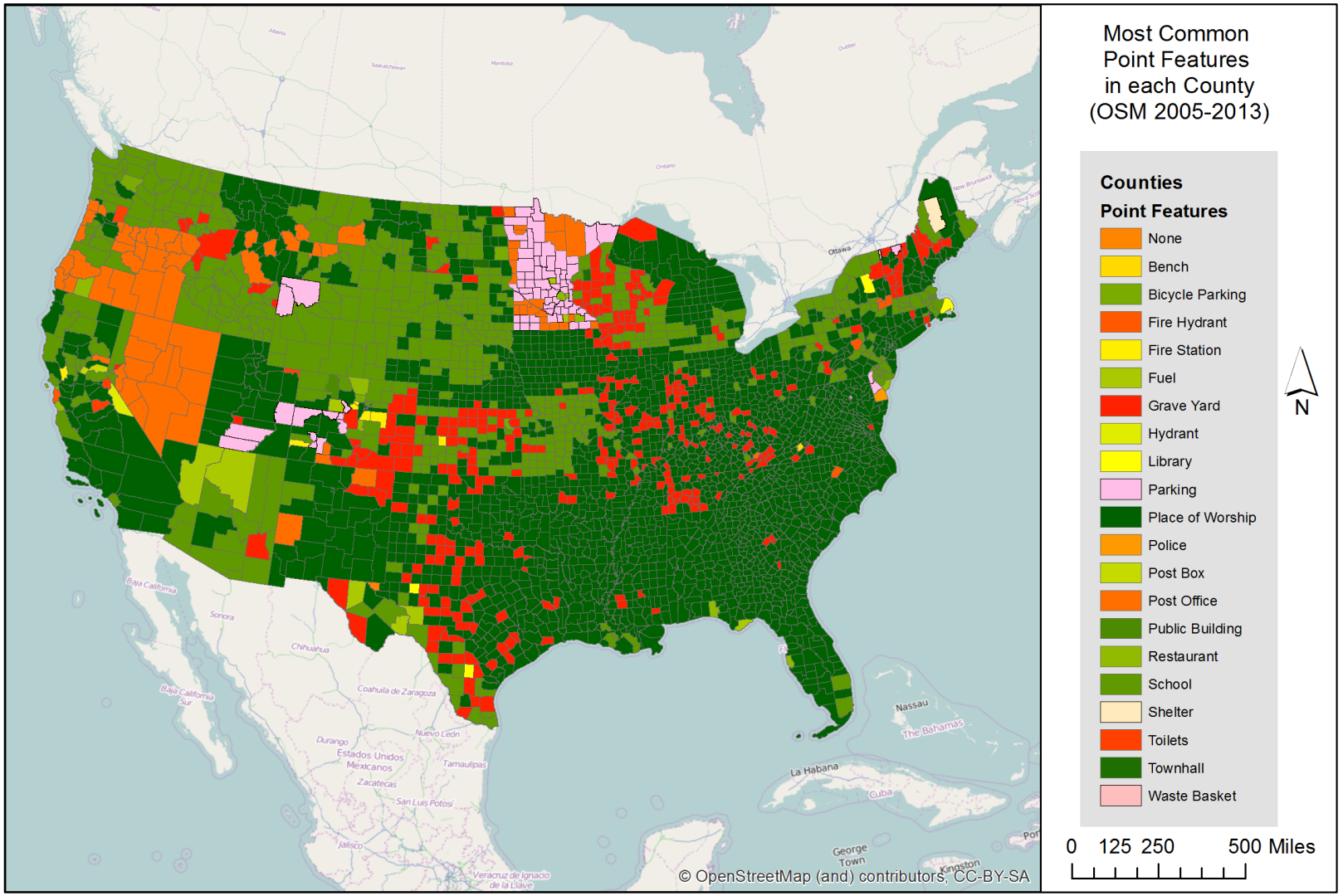

Figure 24. Most frequently mapped entities by county (Point features).

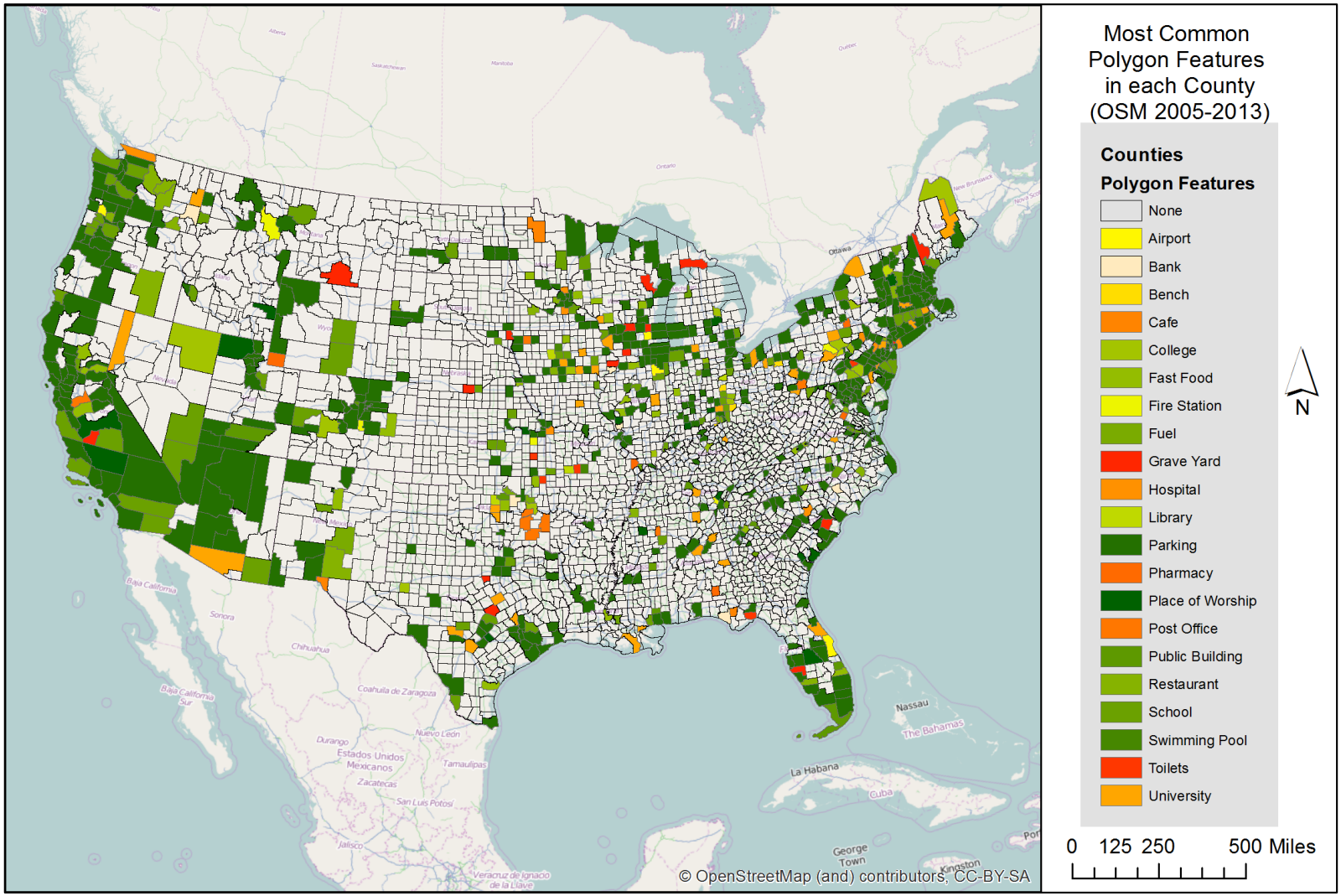

Figure 25. Most frequently mapped entities by county (Polygon features).

Table 34 lists the details of the count of polygon feature representations at the county level.

Table 34. Most frequent entities mapped in counties (Polygon features).

Entity Types	**Count of Counties with This Type as Most Frequent**
None Defined	2348
Parking	440
School	85
University	43
Fast Food	31
Place of Worship	18
Grave Yard	16
Restaurant	16
Public Building	14
College	12
Fuel	8
Hospital	8
Post Office	8
Bank	6
Library	6
Pharmacy	6
Fire Station	5
Swimming Pool	5
Toilets	6
Airport	4
Bench	4
Café	3
Shelter	2
Theatre	2
Townhall	2

CHAPTER VII. DISCUSSION OF RESULTS

The Activity-Context-Geography Model

Typology of VGI Contributors

This classification scheme for data contributors seems reflects how the contributors interact with OSM. Many of the 89 contributors in the high outlier group in all three aspects (Cluster S20, *ACG*) are robot account contributors who use automated scripts to batch-process data to conform to standards or upload data from other sources. Most contributors, however, fall into a cluster exhibiting no extreme aspect values (Clusters S1, *acg*). The data contributors in Cluster S1 made very few feature contributions in OSM and, after an initial interest, most did not become long-term or persistent contributors. The average contributor in the S1 cluster created 6.5 new features while the average contributor in S20 contributed 29,109 new features. The other clusters of high data contribution are S17 (*AcG*), S15 (*[a]CG*), and S12 (*[a]cG*).

To highlight an example from the variables used in the analysis, when comparing the values of the *avg_nodes_per_line* and *avg_nodes_per_polygon* variables, the values in the high outlier were an order of magnitude greater on average than the values in the not-an-outlier category. For details, see Table 35.

Table 35. Results for the average nodes per line and polygon between outlier groups.

Geography aspect	**Average avg nodes per line**	**Average avg nodes per polygon**
High outlier	31.3	31.5
Non-outlier	7.1	1.7
Low outlier	0	0

Previous efforts to categorize VGI contributors have considered different aspects of their contribution, motivations, and abilities (Coleman, Georgiadou, and Labonte 2009; Budhathoki and Haythornthwaite 2012). How does the ACG Model results of groups of contributors compare to earlier efforts? Considering the scale "Neophyte", "Interested Amateur", "Expert Amateur", "Expert Professional", and "Expert Authority", which of the ACG Model groups compare? A Neophyte is defined as having little or no experience, and therefore would have a low Activity or Context aspect *([a]cg, [a][c]g,* or *a[c]g*). Interest amateur, where most of the OSM contributors would be, should align to *acg* – no outliers in any aspect.

The three expert categories – Amateur, Professional, and Authority – may not align as well to the ACG Model. The Expert Amateur, which is "familiar with the strengths and weaknesses of " OSM (Coleman, Georgiadou, and Labonte 2009, 7), may be *Acg* or *ACg*. The Expert Professional has experience with GPS and mapping, which may show up in the Geography aspect (*acG, AcG*). The Expert Authority is a specialist in the field, which may align with *ACG*. The *ACG* and *AcG* clusters, however, appears to be robot contributors, or automated programs that check and correct data. They are also the

realm of contributors who import large amounts of data from other sources.

Budhathoki (2010, 8), aligns VGI participation in two dimensions: use-production as one dimension, and expert-amateur. This may also inform the clusters of the ACG Model. Registered users with little or no contribution may appear in the low or non-outlier form of the Activity aspect. Certainly the 33 *aCg* contributors who had a high Context but were not high in Activity or Geography are well connected to the OSM process – these may indicate Expert Users (as opposed to Expert Producers). Expert Producers may be indicated by the *ACG* group. Amateur Producers may be indicated by the *Acg* group.

Data Quality and Positional Accuracy

The one-way ANOVA results were significant at the 0.05 level for both the test of positional error for all school features and the test for positional error of only imported features compared against the contributor clusters. This suggests that there is a significant difference in data quality (in particular, positional accuracy) across the different clusters of OSM data contributors. This quality difference is observable for all of the OSM school entries in Texas and California as well as for only those school entries made as primary data contributions by the OSM contributors. The group S15 *([a]CG*) with a low Activity aspect and high Context and Geography aspects had the lowest mean positional error in both tests. The group S3 (*acG*) with a high Geography aspect had the highest mean positional error in both tests.

Nearly all of the features imported from other sources were created by the group S20 (*ACG*). Of the 8,380 matching school features in the OSM dataset identified as being imported from a secondary source, 8,373 (94%) of these were features created by cluster

S20 (*ACG*). Removing the features imported from other datasets lowered the mean positional error for cluster S20 by 6.5%.

To better understand how each aspect affects positional accuracy, there is no significant difference in explaining the accuracy when considering all the OSM schools that matched the government datasets. If the OSM schools from secondary sources are removed, there is a significant difference among the outlier groups in the Activity and Context aspects. High outliers in the Activity aspect have positional accuracy than those in the Low outlier group. In the Context aspect, matches within the High outlier group have a lower positional error than those that are not an outlier. It is clear that the inclusion of secondary sources has a large impact in data quality within the dataset. It also appears that in the Activity and Context aspects, contributors with more contributions and/or are better connected to the context of OSM have lower errors in positional accuracy.

The Geographic Distribution of OpenStreetMap

Mapping Activity and Population

At the county level and CBSA level, there is a strong, positive relationship between the number of features mapped in an area and the population of that area. At the block group and place level, this correlation was not strong. This may be partly due to the modifiable areal unit problem which is discussed later in this section. At the block group level, there should be a large disparity of "mappable features" from block group to block group, and this disparity may not related back to population. A block group of six or seven suburban blocks of ranch-houses, for example, would have a higher population than a block group of retail features, yet a block group of retail features would have more

"mappable" features than a suburban block group. This assumes that retail businesses, restaurants, and the like are preferred as entities over residential structures. While there is ample evidence of mapping residential structures in OSM, it is likely there is a clear preference towards mapping public infrastructure, commercial businesses, and recreational entities, although this would be an interesting area of study.

Mapping activity created from GPS data (Type A) and on-screen traces of aerial photos (Type B) was not strongly correlated with population. There may be several explanations for this. The amount of data of Type A is very small compared, and there are not enough contributors using GPS to cover enough area of the United States. Even with the plethora of cheap, GPS-enabled devices on the market, the accuracy needed on a phone and the skill needed to turn that data into accurate map information may be too much for most contributors.

The strong correlation between population and OSM features in urban areas suggests that there may be a ratio, or range of ratios, that fit to a "well-mapped" area. Other factors that may determine how many entities for mapping in an area include population change, urban density, natural features, and economic activity.

Mapping Activity and Socioeconomic Characteristics

The socioeconomic variables BS (having a Bachelor's Degree), A25_54 (ages 25-54), HOWN (homeownership) and TCAR (using a car as a primary means to get to work) have a high degree of multicollinearity. The variables MEN and WOMEN are also highly correlated. This weakens a linear regression analysis and introduces error, so I used Principal Components Regression to model the relationship between the independent variables and the OSM feature counts. The first consideration is that a model have a

reasonable correlation of coefficient, or R^2 explanatory power. If R^2 satisfies the minimum requirement, then the strength of coefficients is compared (Table 21 and Table 25).

For both the Census county and Census CBSA regressions (Table 21 and Table 25), Type A features (data collected by GPS) and Type B features (data created by tracing digital images) did not explain the relationship between the socioeconomic variables examined in the study and the count of OSM features ($R^2 < .05$). For both the Census county and Census CBSA regressions, Type C features (from third-party sources) had a large R^2 values (.560 and .708), respectively. The Census county and CBSA regressions for all features also had larger R^2 values than the Type A or Type B features but below the minimum set out as explanatory (.281 and .531, respectively).

Examining the PCR for Type C features further at the County level, the largest positive relationship for the county level was for Component 1 and Component 4, and Component 3 had a negative relationship. Component 1 is closely related to the variables BS, A25_54, HOWN, and TCAR; Component 3 is strongly negatively related to INC. Component 4 is negatively related to INC and TPUB. The variables MEN and WOMEN do not appear to impact this relationship. The highly correlated variables BS, A25_54, HOWN, and TCAR appear to explain the relationship best. A similar pattern exists for all features.

For Type C features in the PCR at the CBSA level (Table 25), Component 1 is significantly positively related while Component 3 is significantly negatively related to the model. Similar to the County level PCR, Component 1 is most related to the BS, A25_54, HOWN, TCAR, and TPUB variables. Component 3 is highly negatively related

to the INC variable. This suggests that INC (median household income) is strongly related, but independently from the other variables, to the mapping process. Usually, income would be considered highly correlated with education. It may be that, at least for OSM mapping, holding a Bachelor's degree (BS) is not the best choice for representative variable for education.

All features, which includes Types A, B, and C, at the CBSA level exhibit a similar pattern, although not as strong as the Type C features. This suggests that the relationships between features and socio-socioeconomic variables is largely an impact of the features pulled from 3rd party sources. In the United States, much of this data is from the U.S. Census TIGER data. The effect in OpenStreetMap may be a residual effect from the imported TIGER files, and it may not reflect the efforts of mappers in OSM.

Overall, the results of the PCR suggest that there is not a strong relationship between socioeconomic variables for features that were created using GPS (Type A) or by tracing remote images or aerial photography (Type B). This may make sense for Type B features – these features can be created from most computers at a remote distance. Indeed, an OpenStreetMap extension called Maproulette[36] allows authenticated OSM contributors to access a random location and verify, modify, and/or edit OSM data from anywhere in the world. For Type A features, however, the regression results suggest that there is not a relationship between the location of OSM features and the socioeconomic

[36] http://www.maproulette.org, accessed March 15, 2015.

variables of that area. This makes sense for two possible reasons – first, the number of OSM contributors is small enough that it does not represent the population as a whole, and second, the types of features that are of interest to map are not distributed evenly. Looking at the distribution of Type A and Type B feature types may be a topic for more research. I discuss the types of features that are mapped in more detail later in this chapter.

Mapping Activity and Spatial Clustering

From the results of the Morans I, two of the contribution types (Type B and Type C) and all of the features are clustered at a national scale. Mapping the quantiles of activity for each data source type (Figure 19) does show some interesting details. GPS traces appear to be most common on the coast and in the western national parks – that is, vacation spots.

The hot-spot map of Getis-Ord Gi* (Figure 20) suggests the locations of some outlier hot-spots of activity. Chittenden County, VT (seat: Burlington) is an outlier of Type A (GPS) activity. In fact, it is the only county where Type A data is higher than Type C (imported) data. At the local level, mapping activity is spatially autocorrelated around Los Angeles, San Francisco, and Seattle for Type B (on-screen trace) data sources feature density.

Mapping Activity and OSM Community Participation

Table 27 lists the highest density of mapping activity at the metro (CBSA) level, and two areas stand out at outliers: Pascagoula, Mississippi, and Burlington, Vermont. Upon inspection of the data, these communities have a small (less than five) number of contributors who have contributed a large (outlier) number of features to the dataset. In

general, OSM contribution is highly skewed – a small number of contributors account for most of the contributions. In smaller communities, this can have a dramatic effect on how "well-mapped" an area becomes. Clearly, a small group of contributors can have a huge influence in the quantity of map data in an area.

Beyond these two outliers are a series of large, urban cities many of which have active OSM groups (Table 29). These cities have a feature density between 23 and 78 features-per-square-kilometers. Table 28 indicates the areas with little mapping activity. These areas are smaller and may not have enough interest to dedicate to mapping activity. There is a moderate correlation between the size of OSM clubs and the density of mapping activity, but a strong correlation between the number of members in an OSM club and the number of contributors in that same area. The social aspect of OSM may drive the contribution of data in these areas. The communities that can support OSM mapping clubs may generate more mapping activity through mapping parties and mapathons. It should be noted that some metro areas might have OSM mapping clubs that are not affiliated with the OSM website and therefore have not been included in the study. Only clubs on the website had a count of contributors available.

Between these three tables, this suggests that there may be a "sweet spot" of feature density that could indicate a "well-mapped" area. The phrase "well-mapped" is highly subjective, and mapping is a continuous process. There is no fixed answer to what a well-mapped place would look like, yet a low density of mapping activity should indicate areas that need more attention in OSM.

Mapping Activity and Feature Type Choices

Table 30, Table 31, Table 32, Table 33, and Table 34 list the most common types

of features (ie, school, restaurant) identified and mapped in County and CBSA geographic areas in the United States. “Parking” features, “school” and “university” features, “restaurants,” “fast food,” and “place of worship” feature types appear commonly throughout the lists. Interestingly, some locations have “grave yard” as the most common feature type. Others have “swimming pool,” “library”, or “post office” – features which one might expect would occur less regularly on the landscape compared to “parking,” “schools,” or “restaurants.”

One of the areas that lists “post office” as the most common feature type is the metro area of Duluth, Minnesota. Examining the details of the city further, it’s clear that some features (trails, roads, transportation), very few building features have been imported or delineated in the data. This is a space that has a low OSM footprint. A quick Yellow Pages query found 357 restaurants in Duluth, outnumbering post offices by a wide margin. This suggests that looking at the prevalent feature types in an area should highlight areas with missing map data when infrequent feature types appear in the higher ranks of feature type frequency in OSM.

Putting OSM Activity in Place: A Tale of Four Cities

The results and maps tell an interesting story regarding the spatial distribution of OSM activity. To further elaborate on this, I will examine the activity in three cities: Burlington, VT; San Francisco, CA; and Yuma, AZ.

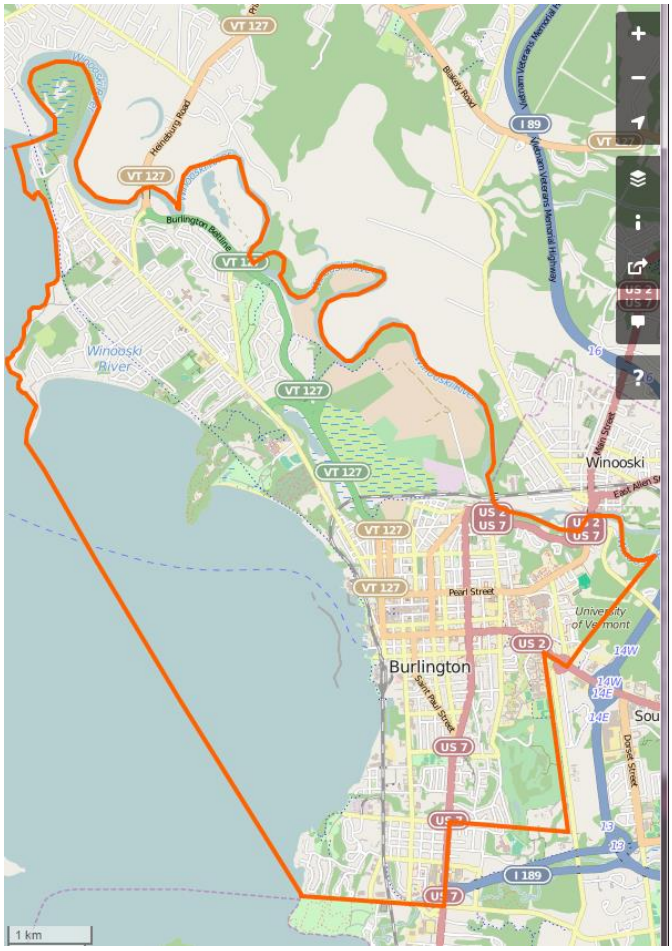

Figure 26. OpenStreetMap.org screenshot of Burlington, Vermont (from July 7, 2015).

Burlington, VT and Yuma, AZ are nearly identical in size and population. Burlington, as of the 2010 Census has 209,381 people; Yuma has 190,526 people. Both are within 50 miles of an international border. Economically, Burlington is a college town while Yuma is dominated by a military base and retirement community. In theory, both college students and retirees may have the time and skillset to produce OSM data. Yet Burlington has over 300,000 features in the OSM dataset while Yuma has a little over 6,000. Digging a little deeper in the data shows that Yuma has had 105 contributors while Burlington has had 175. While Burlington has more contributors, seventy

contributors should not automatically account for a 50-fold difference in features.

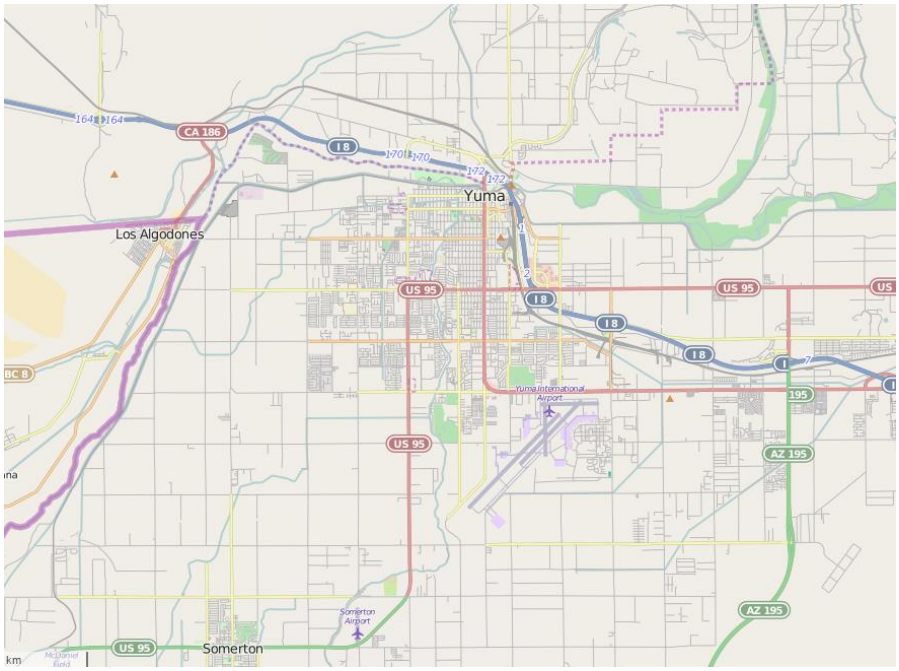

Figure 27. OpenStreetMap.org screen shot of Yuma, AZ (accessed July 7, 2015).

Indeed, seventy contributors is not the difference between Burlington and Yuma – one contributor is. The user *vtcraghead* (B. Morris) has contributed over 300,000 features to the Burlington area (where he resides) and is listed as one of the top 50 contributors in the US[37], while the contributors with the most features in Yuma has barely over 200. A second contributor in Burlington has over 4000 features. There is a similar pattern in some other outlier communities. Pascagoula, MS has one user with over 425,000

[37] http://openstreetmap.us/2013/04/special-invitation-to-top-mappers/, accessed July 7, 2015.

features, easily accounting for 99% of the data in the area. This user is also a top 50 contributors to the US. Yuma doesn't appear to have a local, passionate contributor as these other communities do.

The San Francisco-Oakland-Fremont metro area, wealthy, well-educated and attached at the hip to Silicon Valley, should be a natural hotbed of OSM activity. San Francisco does have a high average of activity compared to some other large US cities, and it has a large number of contributors. One contributor has over 20,000 features, while the next has over 8000. San Francisco also has an active OSM community group, with over 400 members. Unfortunately, it hasn't been possible to tie the OSM email list members to activity, but the indications suggest that there is a strong community with over 100 contributors who have each contributed 100 features. If Yuma is an area with low interest, and Burlington is an area with one dedicated user, then San Francisco represents an area with widespread interest. New York, Portland, and Denver also appear to follow these patterns. None of the contributors in San Francisco were in the top 50 of contributors to the US.

Some Suggestions Behind the Motivation of Contributors

The ACG Model is not designed to examine motivation. Previous work has examined this in depth (Budhathoki and Haythornthwaite 2012; Coleman, Georgiadou, and Labonte 2009; Stephens 2013). Using the results of the data, we can make two general findings about consistent (eg, over 100 features) contributors to OSM. The first is that one contributor in an area can have a large impact. This is clearly the case in some smaller communities that have a large number of contributors. This also explains why

socioeconomic variables mean little in relation to OSM activity. These contributors may be motivated by any number of factors (competition with other contributors, pride of place), but the correlation to the socioeconomic factors is weak. Elaborating further, the user *vtcraghead*, who contributes a large amount in Burlington, Vermont, is the head cartographer at a company that provides reports and data (including spatial data). This is someone who clearly has an expertise with GIS and perhaps a financial motivation to have good data. In some cases, contributing to OSM may be a financial benefit for the individual contributing. They can add and use data without the need for investing in a large data infrastructure.

The second finding is more subtle and perhaps rarer: that a well-developed community can build and contribute to a robust dataset. Several tech-savvy cities (San Francisco, Austin, Denver) have strong OSM communities that meet regularly and have discussions online. Community is a factor that drives the open source model (Raymond 2001), but in practice, it can be difficult to build and maintain – otherwise, every community would have a strong OSM presence.

Limitations of the Study

There are some caveats in the reported study. First, the size of the school campus may impact the positional accuracy assessment. The positional error is calculated as the distance between an OSM school point data and the matched point in government data. However, there is no standard throughout these two datasets regarding how a point is place on a school campus to represent the school. Furthermore, despite precautionary measures to identify and match schools between datasets, there is a possibility that a

school from the OSM dataset is incorrectly matched with one from the government data.

While the ACG Model attempts to organize contributors based on Activity, Context, and Geography, the aspect of Time, or how contributor's behavior patterns may change, is largely ignored. Several variables used to create the Activity context (features per day) and the Context aspect (length of time in OSM) do involve time. A modified model that works to encompass Time as an aspect (perhaps called GACT?) may explain how patterns of change occur both spatially and temporally.

The study is limited by the unavailability of detailed per-contributor socioeconomic information. This study was an attempt to view data trends over a large area, and previous studies (Haklay and Budhathoki 2010; Stephens 2013) have been limited to less than a thousand people. A study that considers both the socioeconomics and spatial contribution of activity of contributors may shed light on more details about the relationship between background and contribution to OSM.

The large size of the dataset in the study has required extensive use of the Amazon Elastic Compute Cloud (Amazon EC2) to store, process, and analyze the data. This allowed faster analysis and mapping of results, but it also required an extended amount of time for testing and verification.

The flexibility of the OpenStreetMap data format does not require any metadata from the contributor to provide details about data collection methods, time, provenance, or other information that would be useful to help gauge the quality of the information. Nearly 40% of the features in the dataset provided no information on how the data was collected. This impacts the analysis as the consideration on the method of data collection is central to the meaning of mapping activity in OSM.

The Modifiable Areal Unit Problem (MAUP)

The Census level geographies present a common problem in GIS analyses: the modifiable areal unit problem (Fotheringham and Wong 1991). The results of analysis are sensitive to the aggregation of information at different scales and units of analysis. Larger aggregates tend to produce higher correlations. To alleviate this, results from four geographic levels have been shown. Block groups, for the reasons discussed earlier, are too small and varied to related population details and mapping activity. Places (cities, towns, villages, etc) are varied in size, population, and density which limits how the mapping activity in these locations may present itself. Even so, additional, detailed work at a larger scale should be done to avoid some of the problems of the sensitivity to the MAUP.

CHAPTER VIII. CONCLUSION

Outcomes of Research Goals and Objectives

In this section, I summarize the outcomes of the research as it relates to the research goals spelled out in Chapter 3. This discussion is followed by some suggestions for future directions that the research may take.

Research Goal 1 Outcomes

The first research goal was to better understand the types of VGI contributors through the patterns of OSM data and its context by building a model of contributor types. The objective of this goal was to develop and implement the Activity-Context-Geography Model of VGI Contribution to identify contributor clusters.

With the growing presence of pervasive location acquisition technologies (Lu and Liu 2012), the data volume of VGI is growing drastically. However, it remains a challenge to assess the quality of VGI data (Flanagin and Metzger 2008). An ACG Model is proposed as a tool to define and group data contributors. Using the variables suggested by Activity, Context, and Geography arenas, twenty groups of VGI contributors to OpenStreetMap (OSM) were identified. The grouping of OSM contributors by the Activity, Context and Geography aspects into low, non-, and high outliers produces a new way to consider VGI contributors based on the data associated with their online presence.

This model is designed for any VGI endeavor. A study using the AGC Model to characterize Twitter users was presented at the International Conference on Location-based Social Media in 2015 (Parr and Lu 2015). The AGC Model could also be used to examine contributions from geographic citizen science projects like eBird.org or the

Central Texas Low Water Crossing project (lowwater.org). The AGC Model can be used to examine and characterize contributors' patterns in a VGI project.

Research Goal 2 Outcomes

The second research goal was to provide a method for understanding the spatial data quality of VGI contributors through the model of contributor types. The objective of this goal was to statistically analyze the differences in contributor data quality (specifically, positional accuracy). For this research goal, government datasets about public schools' locations were used as reference data to the OSM school location data for Texas and California. The positional accuracy of the school data in OSM was found to be significantly different across the different groups of data contributors by the ACG model. Examining the association between the creators of VGI and their data quality, this research is among the first to try to connect VGI data quality to the type of data contributor.

The outcome of this research suggests that the AGC Model may be useful beyond simply characterizing VGI contributors. Based on the data quality findings, the model may explain patterns of differences in data quality of contributions.

Research Goal 3 Outcomes

The third research goal was to develop a model of OSM contribution and examine the choices that contributors make when producing VGI in OSM. The spatial variations in how these choices impact the OSM dataset were examined. The objectives of this goal were to statistically analyze the relationship between population, socioeconomic characteristics, and mapping activity in OSM; to compare the differences in mapping activity at different geographic scales, and to list the most commonly mapped feature

types.

This study presented an exploratory analysis of the spatial relationship of OpenStreetMap data by examining the process and choices that OSM contributors make during the contribution process. The set of features was pulled from the complete 2005-2013 OSM dataset in the forty-eight contiguous states and the District of Columbia. Using tags within the data, features were sorted into three types: GPS traces (A), aerial photo traces (B), and data imported from other sources (C). Mapping activity was correlated with population which found that at some levels, mapping activity is strongly correlated with population. Using Morans I and Getis-Ord Gi* hot spot analysis, the spatial distribution of mapping activity identified outliers of mapping activity. Identifying the twenty highest and lowest metro areas with feature density identified a potential range of ratios that may indicate a "well-mapped" area. Correlating the number of members in OSM mapping groups with the number of contributors in their local metro area found a .83 correlation that these groups have a positive relationship. More mappers may lead to a higher number of features, although it may be related to population dynamics, as having an OSM mapping club did not produce a higher feature density than metro areas without an OSM mapping club.

The story of OpenStreetMap is one of a social process that uses mapping parties to collect and load information from GPS traces and on-screen aerial photo edits (Ramm, Topf, and Chilton 2011). In spite of this story, in the United States at least, most of the data comes from third-party sources. This may have had the unintended effect of lowering the participation of contributors in the United States; in other countries where spatial data is not freely available, generating and using data not available elsewhere is a

strong motivation (Haklay and Budhathoki 2010). This study does find, however, that overall, the urban areas that have active OSM groups have higher feature density than other areas. This suggests that a feature density of above 23 features per square kilometer may be a "sweet spot" indicating a well-mapped area Table 27. Looking at the feature density may also indicate areas that need more attention (Table 28).

Future Research

For future studies, a map of trust may be constructed based on the outcome from this study. Such a map would identify potentially less credible information. For the data contributors that belong to a contributor group that may tend to create relatively low quality VGI data, a map of their data contribution can be labeled as a baseline, calling for data the requires focused validation or a re-check. Of course, assigning the traits of a group to any individual data contributor is subject to Ecological Fallacy; so it should be with caution not to single out any data contributor.

Another direction for future studies would be to examine the other aspects of data quality as they relate to the AGC Model. The Federal Geographic Data Committee (FGDC) defines five aspects for geographic data quality, namely positional accuracy, attribute accuracy, completeness, logical consistency, and lineage. It is worthy further efforts to investigate the possible connection between these aspects of data accuracy with data contributor.

Future research should consider detailed analysis of areas of to explore in more detail what the extent of a "well-mapped" area would look like. As OSM provides per-user information of activity, context, and geography, another possible avenue of research

would be to use Markov Chain analysis to determine the probability of mapping activity and then model it using Agent-Based Modeling. Mapping is a continuous process in OSM, so areas that have even a smaller level of mapping activity will eventually have higher feature density than other areas. Identifying the areas that lack detail should be used to provide feedback on where to map next; this would also identify areas that may have quality or accuracy issues which is a crucial concern for the longevity of OSM (Ramm, Topf, and Chilton 2011; Mooney, Corcoran, and Winstanley 2010; Lin 2011; Haklay 2010b).

REFERENCES

Anderson, D. P., J. Cobb, E. Korpela, M. Lebofsky, and D. Wertheimer. 2002. SETI @ home: An experiment in Public-Resource Computing. *Communications of the ACM* 45 (11):56–61.

Anthony, D., S. W. Smith, and T. Williamson. 2009. Reputation and Reliability in Collective Goods: The Case of the Online Encyclopedia Wikipedia. *Rationality and Society* 21 (3):283–306.

Ballatore, A., M. Bertolotto, and D. C. Wilson. 2013. Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap. *Knowledge and Information Systems* 37:61–81.

Balram, S., S. Dragićević, and R. Feick. 2009. Collaborative GIS for spatial decision support and visualization. *Journal of Environmental Management* 90 (6):1963–5.

BBC. 2012. South Pacific Sandy Island "proven not to exist." *BBC News* :1–15. http://m.bbc.com/news/world-asia-20442487 (last accessed 2 March 2015).

Benkler, Y., and H. Nissenbaum. 2006. Commons-based Peer Production and Virtue. *Journal of Political Philosophy* 14 (4):394–419.

Bennett, J. 2010. *OpenStreetMap*. Birmingham, UK: PACKT.

Beyer, M. A., and D. Laney. 2012. The Importance of "Big Data": A Definition. *Gartner Publications* 21 June:1–9.

Budhathoki, N. R. 2010. Participants' Motivations to Contribute Geographic Information in An Online Community: Dissertation.

Budhathoki, N. R., and C. Haythornthwaite. 2013. Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist* 57:548–575.

Budhathoki, N. R., and C. Haythornthwaite. 2012. Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist* 57 (5):548–575.

Chow, T. E. 2008. The Potential of Maps APIs for Internet GIS Applications. *Transactions in GIS* 12 (2):179– 191.

Coleman, D. J. 2013. Potential Contributions and Challenges of VGI for Conventional Topographic Base-Mapping Programs. In *Crowdsourcing Geographic Information: Volunteered Geographic Information (VGI) in Theory and Practice*, eds. D. Z. Sui, S. Elwood, and M. F. Goodchild, 245–263. Dordrecht, Germany: Springer.

Coleman, D. J., Y. Georgiadou, and J. Labonte. 2009. Volunteered Geographic Information : the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research* 4:332–358.

Connors, J. P., S. Lei, and M. Kelly. 2012. Citizen Science in the Age of Neogeography: Utilizing Volunteered Geographic Information for Environmental Monitoring. *Annals of the Association of American Geographers* 102 (6):1267–1289.

Cook, S., C. Conrad, A. L. Fowlkes, and M. H. Mohebbi. 2011. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PloS one* 6 (8):e23610.

Corcoran, P., and P. Mooney. 2013. Characterising the metric and topological evolution of OpenStreetMap network representations. *The European Physical Journal Special Topics* 215 (1):109–122.

Crampton, J. 2010. *Mapping: A Critical Introduction to Cartography and GIS*. Oxford: Wiley-Blackwell.

Crampton, J. W. 2009. Cartography: performative, participatory, political. *Progress in Human Geography* 33 (6):840–848.

Crampton, J. W., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. W. Wilson, and M. Zook. 2013. Beyond the Geotag? Deconstructing "Big Data" and Leveraging the Potential of the Geoweb. *Cartography and Geographic Information Science* 40 (2):130–139.

Crutcher, M., and M. Zook. 2009. Placemarks and waterlines: Racialized cyberscapes in post-Katrina Google Earth. *Geoforum* 40 (4):523–534.

Devillers, R., D. Bégin, and A. Vandecasteele. 2012. Is the rise of Volunteered Geographic Information (VGI) a sign of the end of National Mapping Agencies as we know them ? *GIScience 2012* :0–2.

Van Dijk, J., and K. Hacker. 2003. The Digital Divide as a Complex and Dynamic Phenomenon. *The Information Society* 19:315–326.

Duncan, G. 2012. Why Are Companies Defecting from Google Maps? *Digital Trends*. http://www.digitaltrends.com/mobile/why-are-companies-defecting-from-google-maps/ (last accessed 1 March 2015).

Elwood, S. 2008. Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal* 72 (3-4):173–183.

Elwood, S. A. 2002. GIS use in community planning: a multidimensional analysis of empowerment. *Environment and Planning - Part A* 34 (5):905–922.

Elwood, S., M. F. Goodchild, and D. Z. Sui. 2012. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers* 102 (3):571–590.

Elwood, S., and A. Leszczynski. 2011. Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum* 42 (1):6–15.

Ester, M., H. Kriegel, and X. Xu. 1995. Knowledge Discovery in Large Spatial Databases : Focusing Techniques for Efficient Class Identification. In *Proceedings of the Fourth International Symposium on Large Spatial Databases*.

Estima, J., and M. Painho. 2013. Exploratory analysis of OpenStreetMap for land use classification. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '13*, 39–46. New York, New York, USA: ACM Press.

Executive Office of the President. 2012. *Big Data Across the Federal Government*.

Fan, H., A. Zipf, Q. Fu, and P. Neis. 2014. Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science* 28 (4):700–719.

Federal Geographic Data Committee. 2015. The Federal Geographic Data Committee Website. *The Federal Geographic Data Committee*. http://www.fgdc.gov.

File, T., and C. Ryan. 2014. *Computer and Internet use in the United States*.

Fisher, P. F. 1999. Models of uncertainty in spatial data. In *Geographical information Systems: Principles, Techniques, Management and Applications Vol. 1*, eds. P. A. Longley, M. F. Goodchild, D. Maguire, and D. W. Rhind, 191–205. New York: Wiley.

Flanagin, A. J., and M. J. Metzger. 2008. The credibility of volunteered geographic information. *GeoJournal* 72 (3-4):137–148.

Forrest, B. 2010. Base Map 2.0 : What Does the Head of the US Census Say to Open Street Map ? *O'Reilly Radar* :1–6. http://radar.oreilly.com/2010/03/base-map-20-what-does-the-head.html.

Fotheringham, S., and D. W. S. Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23 (7):1025–1044.

Gahegan, M., M. Wachowicz, M. Harrower, and T.-M. Rhyne. 2001. The Integration of Geographic Visualization with Knowledge Discovery in Databases and Geocomputation. *Cartography and Geographic Information Science* 28 (1):29–44.

Girres, J.-F., and G. Touya. 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14 (4):435–459.

Goodchild, M. F. 2007a. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69 (4):211–221.

———. 2007b. Citizens as Voluntary Sensors : Spatial Data Infrastructure in the World of Web 2 . 0. *International Journal of Spatial Data Infrastructure* 2:24–32.

———. 1995. Future Directions for Geographic Information Science. *Geographic Information Sciences* 1 (1):1–8.

———. 2007c. Geographic Information Systems. eds. J. P. Wilson and A. S. Fotheringham, 8–10. Blackwell Publishing.

———. 2009a. Geographic information systems and science: today and tomorrow eds. J. P. Wilson and A. S. Fotheringham. *GeoFocus* 7 (1):8–10.

———. 1992. Geographical Information Science. *International Journal of Geographical Information Systems* 6 (l):31–45.

———. 2002. Measurement-based GIS. In *Spatial Data Quality*, eds. W. Shi, P. F. Fisher, and M. F. Goodchild, 5–17. New York: Taylor & Francis Ltd.

———. 2009b. NeoGeography and the nature of geographic expertise. *Journal of Location Based Services* 3 (2):82–96.

Goodchild, M. F., and J. A. Glennon. 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth* 3 (3):231–241.

Goodchild, M. F., and L. Li. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics* 1:110–120.

Graham, M. 2012. Big data and the end of theory? *The Guardian* 9 March.

Graham, M., S. A. Hale, and D. Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *Professional Geographer* 66 (4):568–578.

Haklay, M. 2013. Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In *Crowdsourcing Geographic Information: Volunteered Geographic Information (VGI) in Theory and Practice*, eds. D. Z. Sui, S. Elwood, and M. F. Goodchild, 105–122. Dordrecht, Germany: Springer.

———. 2010a. Geographical Citizen Science – clash of cultures and new opportunities. In *GIScience workshop*, 1–6.

———. 2010b. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B Planning and Design* 37 (4):682–703.

Haklay, M., S. Basiouka, V. Antoniou, and A. Ather. 2010. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus Law to Volunteered Geographic Information. *Cartographic Journal, The* 47 (4):315–322.

Haklay, M., and N. Budhathoki. 2010. OpenStreetMap – Overview and Motivational Factors. In *Horizon Infrastructure Challenge Theme Day*.

Haklay, M., A. Singleton, and C. Parker. 2008. Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass* 2 (6):2011–2039.

Haklay, M., and P. Weber. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7 (4):12–18.

Hall, G. B., R. Chipeniuk, R. D. Feick, M. G. Leahy, and V. Deparday. 2010. Community-based production of geographic information using open source software and Web 2.0. *International Journal of Geographical Information Science* 24 (5):761–781.

Han, J., Y. Cai, and N. Cercone. 1992. Knowledge Discovery in Databases : An Attribute-Oriented Approach. In *Proceedings of the 18th VLDB Conference*.

Hardy, D. 2008. Discovering behavioral patterns in collective authorship of place-based information. In *9th International Conference of the Association of Internet Researchers*. Copenhagen, Denmark.

Hardy, D., J. Frew, and M. F. Goodchild. 2012. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science* 26 (7):1191–1212.

Harris, T. M. 2012. Interfacing archaeology and the world of citizen sensors: exploring the impact of neogeography and volunteered geographic information on an authenticated archaeology. *World Archaeology* 44 (4):580–591.

Harvey, F. 2013. To Volunteer or Contribute Locational Information? Towards Truth in Labeling for Crowdsourced Geographic Information. In *Crowdsourcing Geographic Information: Volunteered Geographic Information (VGI) in Theory and Practice2*, eds. D. Z. Sui, M. F. Goodchild, and S. Elwood, 31–41. Dordrecht, Germany: Springer.

Hecht, B., and M. Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *Proceedings of ICWSM 2014*.

Helbich, M., C. Amelunxen, P. Neis, and A. Zipf. 2010. Investigations on Locational Accuracy of Volunteered Geographic Information Using OpenStreetMap Data. In *GIScience 2010 Workshop*. Zurich, Switzerland.

Horita, F., L. Degrossi, L. Assis, A. Zipf, and J. P. de Albuquerque. 2013. The use of Volunteered Geographic Information and Crowdsourcing in Disaster Management : a Systematic Literature Review. In *Proceedings of the 19th Americas Conference on Information Systems*, 1–10. Chicago, Illinois.

Jankowski, P., and T. Nyerges. 2001. GIS-Supported collaborative decision making: results of an experiment. *Annals of the Association of American Geographers* 91 (1):48–70.

Jokar Arsanjani, J., A. Zipf, P. Mooney, and M. Helbich. 2015. *OpenStreetMap in GIScience: Experiences, Research and Applications (Lecture Notes in Geoinformation and Cartography)* eds. J. Jokar Arsanjani, A. Zipf, P. Mooney, and M. Helbich. Springer.

Kagoyire, C., and R. a. de By. 2012. Models for professional cyclic activities in VGI with a case in coffee farming. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '12*, 70. New York, New York, USA: ACM Press.

Kar, B., R. C. Crowsey, and J. J. Zale. 2013. The Myth of Location Privacy in the United States: Surveyed Attitude Versus Current Practices. *The Professional Geographer* 65 (1):47–64.

Keßler, C., and R. T. Anton de Groot. 2013. Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap. In *Geographic Information Science at the Heart of Europe: Lecture Notes in Geoinformation and Cartography*, Lecture Notes in Geoinformation and Cartography., eds. D. Vandenbroucke, B. Bucher, and J. Crompvoets, 21–37. Heidelberg: Springer International Publishing.

Knudsen, A. S., and M. Kahlia. 2012. Review The role of Volunteered Geographic Information in participatory planning : Examples from Denmark and Finland. *Perspektiv* (21):35–48.

Koukoletsos, T., M. Haklay, and C. Ellul. 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS* 16 (4):477–498.

Laney, D. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *META Group* 6 February.

Leitner, H., R. B. McMaster, S. Elwood, S. McMaster, and E. Sheppard. 2002. Models for making GIS available to community organizations: dimensions of difference and appropriateness. *Community participation and geographic information systems* (October):37–52.

Lin, Y.-W. 2011. A qualitative enquiry into OpenStreetMap making. *New Review of Hypermedia and Multimedia* 17 (1):53–71.

Liu, J., and S. Ram. 2009. Who Does What: Collaboration Patterns in the Wikipedia and Their Impact. In *19th Workshop on Information Technologies and Systems*. Phoeniz, Arizona.

Liu, S. B., and L. Palen. 2010. The New Cartographers: Crisis Map Mashups and the Emergence of Neogeographic Practice. *Cartography and Geographic Information Science* 37 (1):69–90.

Lohr, S. 2012. The Age of Big Data. *New York Times* 11 February:1–5.

Lu, Y., and Y. Liu. 2012. Pervasive location acquisition technologies : Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems* 36 (2):105–108.

MacEachren, A. M., M. Wachowicz, R. Edsall, D. Haug, and R. Masters. 1999. Constructing Knowledge From Multivariate Spatiotemporal Data: integrating geographical visualization and knowledge discovery in database methods. *International Journal of Geographical Information Science* 13 (4):311–334.

Madej, M., M. Soniat, L. Dupont, and M. M. Thompson. 2012. Assessing Street Conditions through Volunteer Spatial Mapping in Lakeview Assessing Street Conditions through Volunteer Spatial Mapping in Lakeview. *Planning and Urban Studies Reports and Presentations*.

Mathews, A. J., Y. Lu, M. T. Patton, N. Dede-Bamfo, and J. Chen. 2012. College students' consumption, contribution, and risk awareness related to online mapping services and social media outlets: does geography and GIS knowledge matter? *GeoJournal* 78 (4):627–639.

Mennis, J., and D. Guo. 2009. Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems* 33 (6):403–408.

Miller, H. J. 2010. The Data Avalanche Is Here. Shouldn'T We Be Digging? *Journal of Regional Science* 50 (1):181–201.

Miller, H. J., and J. Han. 2009. Geographic Data Mining and Knowledge Discovery: An Overview. In *Geographic Data Mining and Knowledge Discovery*, eds. H. J. Miller and J. Han, 9–21. Boca Raton, FL: CRC Press.

Mondzech, J., and M. Sester. 2011. Quality Analysis of OpenStreetMap Data Based on Application Needs. *Cartographica: The International Journal for Geographic Information and Geovisualization* 46 (2):115–125.

Mooney, P., and P. Corcoran. 2011. Accessing the history of objects in OpenStreetMap. In *AGILE 2011: the 14th AGILE international conference on geographic information science.*, 155. Utrecht: Springer.

———. 2012a. Characteristics of Heavily Edited Objects in OpenStreetMap. *Future Internet* 4 (4):285–305.

———. 2012b. How social is OpenStreetMap? In *Proceedings of AGILE 2012*.

Mooney, P., P. Corcoran, and B. Ciepluch. 2013. The potential for using volunteered geographic information in pervasive health computing applications. *Journal of Ambient Intelligence and Humanized Computing* 4 (6):731–745.

Mooney, P., P. Corcoran, and A. C. Winstanley. 2010. Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10*. New York, New York.

Mooney, P., H. Sun, P. Corcoran, and L. Yan. 2011. Citizen-generated spatial data and information: Risks and opportunities. *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics* :232–232.

Nedovic-Budic, Z., and J. K. Pinto. 1999. Understanding Interorganizational GIS Activities: A Conceptual Framework. *Journal of the Urban and Regional Information Systems Association* 11 (1):53–64.

Neis, P., and D. Zielstra. 2014. Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet* 6 (1):76–106.

Neis, P., D. Zielstra, and A. Zipf. 2011. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet* 4 (4):1–21.

Neis, P., and A. Zipf. 2012. Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information* 1 (3):146–165.

Newman, G., D. Zimmerman, A. Crall, M. Laituri, J. Graham, and L. Stapel. 2010. User-friendly web mapping: lessons from a citizen science website. *International Journal of Geographical Information Science* 24 (12):1851–1869.

Newsam, S. 2010. Crowdsourcing What Is Where : Community-contributed photos as volunteered geographic information. *IEEE World* :36–45.

NRC. 2010. What are the societal implications of citizen mapping and mapping citizens? In *Understanding the changing planet: strategic directions for the geographical sciences*, 105–112. National Academies Press.

Nyerges, T., P. Jankowski, D. Tuthill, and K. Ramsey. 2006. Collaborative Water Resource Decision Support: Results of a Field Experiment. *Annals of the Association of American Geographers* 96 (4):699–725.

O'Reilly, T. 2005. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. :5. http://oreilly.com/web2/archive/what-is-web-20.html.

Van Oort, P. 2006. Spatial data quality : from description to application, PhD Thesis. *Production*.

Parker, C. J., A. May, and V. Mitchell. 2013. The Role of VGI and PGI in Supporting Outdoor Activities. *Applied Ergonomics* 44 (6):886–894.

Parr, D., and Y. Lu. 2015. Classifying Twitter Users with the Activity-Context-Geography Model. In *International Conference for Location-based Social Media, March 13, 2015*, 1–11. Athens, GA.

Parr, D., and M. Scholz. 2015. Building a Low-Cost Geographic Website For Collecting Citizen Science Contributions. *Papers in Applied Geography*.

Pattison, W. D. 1990. The Four Traditions of Geography. *Journal of Geography* 89 (5):202–206.

Pickles, J. 2004. *A History of Spaces: Cartographic reason, mapping, and the geo-coded wrold*. New York: Routledge.

Ramm, F., J. Topf, and S. Chilton. 2011. *OpenStreetMap: Using and Enhancing the Free Map of the World*. Cambridge, UK: UIT Cambridge Ltd.

Raymond, E. S. 2001. *The cathedral and the bazaar*. O'Reilly Media, Inc.

Reagle, J. M. 2010. *Good Faith Collaboration: The Culture of Wikipedia*. Boston, MA: MIT Press.

Rehrl, K., S. Gröechenig, H. Hochmair, S. Leitinger, R. Steinmann, and A. Wagner. 2013. A Conceptual Model for Analyzing Contribution Patterns in the Context of VGI. In *Progress in Location-Based Services*, Lecture Notes in Geoinformation and Cartography., ed. J. M. Krisp, 373–388. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Roche, S., E. Propeck-Zimmermann, and B. Mericskay. 2011. GeoWeb and crisis management: issues and perspectives of volunteered geographic information. *GeoJournal* 78 (1):21–40.

Schutzberg, A. 2012. Ten Things You Need to Know About OpenStreetMap. *Directions Magazine* :1–9. http://www.directionsmag.com/articles/ten-things-you-need-to-know-about-openstreetmap/244904.

Schuurman, N. 2009. The new Brave New World: geography, GIS, and the emergence of ubiquitous mapping and data. *Environment and Planning D: Society and Space* 27 (4):571–572.

Smyth, R. 2004. Exploring the usefulness of a conceptual framework as a research tool : A researcher' s reflections. *Issues in Educational Research* 14 (167-180).

Snijders, C., U. Matzat, and U. Reips. 2012. "Big Data": Big Gaps of Knowledge in the Field of Internet Science. *International Journal of Internet Science* 7 (1):1–5.

Song, W., and G. Sun. 2010. The role of mobile volunteered geographic information in urban management. In *2010 18th International Conference on Geoinformatics*. Beijing, China: IEEE.

Stallman, R. 2002. *Free Software , Free Society : Selected Essays of Richard M . Stallman*. Boston, MA: GNU Press.

Stefanidis, A., A. Crooks, and J. Radzikowski. 2011. Harvesting ambient geospatial information from social media feeds. *GeoJournal* 78 (2):319–338.

Stephens, M. 2012. From Geo-Social to Geo-Local: The Flows and Biases of Volunteered Geographic Information: Dissertation.

———. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal* 78:981–996.

Sui, D. Z., M. F. Goodchild, and S. Elwood. 2013. Volunteered Geographic Information, the Exaflood, and the Growing Digital Divide. In *Crowdsourcing Geographic Information: Volunteered Geographic Information (VGI) in Theory and Practice*, eds. D. Z. Sui, M. F. Goodchild, and S. Elwood, 1–12. Dordrecht, Germany: Springer.

Takhteyev, Y., A. Gruzd, and B. Wellman. 2011. Geography of Twitter networks. *Social Networks* 34 (1):73–88.

Taylor, P., M. Tsou, and M. Leitner. 2013. Visualization of social media : seeing a mirage or a message ? *Cartography and Geographic Information Science* 40 (May):37–41.

Trame, J., and C. Keßler. 2010. Exploring the Lineage of Volunteered Geographic Information with Heat Maps. In *GeoViz 2011*, 6–7.

Tsou, M.-H. 2011. Revisiting Web Cartography in the United States: the Rise of User-Centered Design. *Cartography and Geographic Information Science* 38 (3):250–257.

Tulloch, D. L. 2008. Is VGI participation? From vernal pools to video games. *GeoJournal* 72 (3-4):161–171.

Turner, A. J. 2006. *Introduction to Neogeography* ed. A. J. Turner. O'Reilly Media, Inc.

Warf, B., and D. Sui. 2010. From GIS to neogeography: ontological implications and theories of truth. *Annals of GIS* 16 (4):197–209.

White, T. 2012. *Hadoop: the Definitive Guide* 3rd ed. San Francisco, CA: O'Reilly Media, Inc.

Yang, C., M. Goodchild, Q. Huang, D. Nebert, R. Raskin, Y. Xu, D. Fay, and I. Spatial. 2011. Spatial Cloud Computing: How geospatial sciences could use and help to shape cloud computing. *International Journal on Digital Earth* 4 (4):305–329.

Young, J. C., and M. P. Gilmore. 2013. The Spatial Politics of Affect and Emotion in Participatory GIS. *Annals of the Association of American Geographers* 103 (4):808–823.

Zook, M. A., and M. Graham. 2007. From cyberspace to DigiPlace : Visibility in an age of information and mobility. In *Socities and Cities in the Age of Instant Access*, 241–254.

Zook, M., M. Graham, T. Shelton, and S. Gorman. 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy* 2 (2):6–32.