Quality Assessment of Volunteered Geographic Information:

An Investigation into the Ottawa-Gatineau OpenStreetMap Database

by

Kent Thomas Jacobs

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Geography and Environmental Science

with a Specialization in Data Science

Carleton University
Ottawa, Ontario

ABSTRACT

Within the realm of Volunteered Geographic Information (VGI), reliability and quality of the geographic information continues to be a pressing concern. Many VGI projects do not have standard geospatial data quality assurance procedures and the reliability of such contributors remains in question. This study investigates the quality of VGI by analysing OpenStreetMap (OSM) data in Ottawa-Gatineau. First, a review of past publications into quality assessment of OSM data is examined. Next, a comparative analysis of OSM data is conducted relative to an authoritative dataset. The OSM historical information of map features and contributors is inspected to gain an understanding of how users are contributing to the database and their ability to do so accurately. Overall, OSM data in the context of Ottawa-Gatineau is comparable to or surpasses authoritative dataset quality and clustering contributors based on historical information can help identify tendencies within a contributor base.

DEDICATION AND ACKNOWLEDGEMENTS

I would like to first and foremost thank my caring and loving parents, Kathy and Tom, who have each supported me through this journey and encouraging me to study what I am interested in. Without my parents conveying to me a wealth of guidance, knowledge and wisdom, none of this would have been possible.

I would also like to thank Dr. Scott Mitchell, who has provided me with a plethora of support and guidance during my research and academic writing stages. His leadership allowed me to embrace a relatively new research topic and allowed for the opportunity to expand my knowledge and skillset. Dr. Murray Richardson and Dr. Andrew Davidson also provided me with further research and writing direction over the past two years.

I would like to recognize the Free and Open Source Software initiative which allowed me to automate and script portions of my research and edit the underlying source code when needed. The OpenStreetMap.org/OpenStreetMap Foundation played an integral role, given that much of the data used in this research is to their credit. I share the same acknowledgement to past academics in the OpenStreetMap and VGI fields who have inspired me to study this topic.

Credit is also attributed to Employment and Social Development Canada (ESDC). During my time as a student, the Geomatics division at ESDC allowed me to contribute towards GIS related projects and overlap some of my underlying research in the field of OSM/VGI.

Finally, I would like to give acknowledgement to Carleton University and the Department of Geography and Environmental Studies, where I have spent my undergraduate and graduate years. The staff and students at Carleton University have enhanced my learning experience and growth as a human being over the total of seven academic years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS/ABBREVIATIONS

Abbreviation	Meaning
AH	Amtliche Hausumrine
AOI	Area of Interest
ATKIS	Authoritative Topographic Cartographic Information System
BDLM	Base Digital Landscape Model
CSD	Census Subdivisions
DCMS	DMTI CanMap Suite
DEIL	Data Exploration and Integration Lab
DMTI	Digital Mapping Technologies Inc.
DWG	Data Working Group
ESDC	Employment and Social Development Canada
GIS	Geographic Information System
GPS	Global Positioning System
HOT	Humanitarian OpenStreetMap Team
ITN	Integrated Transport Network
ISO	International Organization for Standardization
JOSM	Java OpenStreetMap Editor
LBS	Location-Based Service
NRCan	Natural Resources Canada
OSM	OpenStreetMap
OSMF	OpenStreetMap Foundation
OS	Ordnance Survey
STATCAN	Statistics Canada
TC	Technical Committee
POI	Point of Interest
PoS	Point of Service
VGI	Volunteered Geographic Information
PCA	Principal Component Analysis

1 INTRODUCTION

1.1 Introduction

Digital spatial data is defined as information pertaining to a real-life feature located on the surface of the earth (Longley et al., 1999). Such information is crucial to help us understand the practical real-world issues, whether they are biological, environmental, social or real in time (Fisher et al., 2002). Digital spatial data are typically collected, gathered and stored by governmental departments or privately-owned enterprises. With advancements in geographic information technologies over the years, the methods of performing data acquisition and collection have changed significantly. Among these advancements is geographic information technologies which include the development of crowd-sourced georeferenced datasets.

Goodchild (2007) coined the term Volunteered Geographic Information (VGI), as the process of creating, collecting and circulating geographic information provided entirely by volunteers using various tools and applications. These various applications or tools may include Internet-based platforms, Global Positioning System (GPS) devices, or mobile devices. Some of these Internet-based platforms include Wikimapia, Google MyMaps, Instagram and Flickr (Senaratne et al., 2017). VGI emerged during the "Web 2.0" era in which individuals could carry out and utilize modern technologies and innovations to circulate VGI where they may not be qualified to do so. An example of this would be writing general articles for Wikipedia (Antoniou et al., 2010; Haklay, 2010).

Since 2007, one of the primary issues discussed with respect to VGI has been the quality of the geospatial information. The main issue surrounding quality is the fact that some of the information is collected by volunteers (who by definition act without compensation), who may have little to no experience working with geospatial information (Haklay, 2010). If VGI is being used to portray an accurate representation of the real-world, it must be improved in many ways and a process which involves quality assessment and assurance should be an integral part of this enhancement process. Assessing the accuracy of VGI can include spatial (positional) and thematic (attribute) error, as well as the reliability of the contributors and their specific mapping characteristics. This research also brings forward the potential to contribute to a new perspective on the models for "authoritative"

Mapping characteristics refers to how a VGI (OSM) contributor commits to a dataset. For example, a novice VGI contributor may be only comfortable adding hiking trails or simple Point of Interest (POI) nodes to OSM, whereas a more experienced VGI contributor may look to improve dataset attribute consistency across large spatial extents (e.g. Fixing business postal code or phone number formats across Canada). The terms "dataset evolution" or "chronological evolution" are used throughout this report and refers to the growth of a dataset's map features (buildings, roads, etc.) over time.

Given the methods of data collection and distribution, several questions have been raised, such as: how many contributors would make a crowd-sourced geographic dataset credible without authoritative endorsement; would data users be

willing to sacrifice spatial and thematic accuracy for temporal accuracy of a dataset; and how does VGI compare relative to "authoritative" datasets through benchmark evaluation processes?

1.2 OpenStreetMap Background

OpenStreetMap is one of the most prominent crowd-sourced web-based open mapping platforms. The OSM dataset is currently maintained by the OpenStreetMap Foundation (OSMF), which is an international non-profit organization supporting, but not totally governing, the OSM Project. The OSM Project was launched in 2004, by Steve Coast, and aims to develop a free map of the world available to everyone. There are several reasons for the initiation of the OSM project. It was partially a reaction to the large costs associated with proprietary datasets, but also to encourage the growth of geospatial data that are readily available for people to use and share (Ramm et al., 2011).

The inherent difference between OSM and proprietary dataset like Google Maps, brings forward an ethical debate as to how data are collected and shared. OSM has taken an "open" approach to the methods of data collection and distribution, whilst Google is a multi-national corporation which has chosen to take a "closed" approach to data collection and its propagation. Ultimately, both mapping applications try to answer the same spatial and thematic queries of "where" (e.g. Where is the nearest coffee shop? Where is the nearest hospital?)[1].

[1] http://geoawesomeness.com/why-would-you-use-openstreetmap-if-there-is-google-maps/

Today, the OSM project has accumulated the attention of millions of contributors within mapping communities globally which utilize local knowledge to maintain the temporal accuracy and currency of the available dataset. OSM users aim to achieve this goal by merging geographic data which is readily available from open government data portals with spatial and thematic data gathered from commercially available GPS devices, digitized aerial/satellite imagery or the local knowledge provided by simple paper maps. Due to the variety of methods surrounding the information and data collection used by the OSM project, various issues and concerns related to the quality of OSM data are brought forward.

1.3 OpenStreetMap in Canada

Since 2008, the federal government of Canada has released a significant amount of geographic data through open data portals which can then be imported into OSM. Open data from federal, provincial and municipal governments are continuously being imported into OSM due to such initiatives. Initial data imports for the Ottawa area involved uploading simple geographic data (administrative boundaries, roads, place names, water bodies) from GeoBase, a former government (federal, provincial, municipal) platform aimed to democratize geospatial data for Canadian citizens ('WikiProject Canada - OpenStreetMap Wiki', 2018). Natural Resources Canada (NRCan) is one of the government organizations that have participated in the open data initiative, aligning their internal CanVec dataset with an OSM-compatible license. CanVec contains much of the data included in GeoBase, plus extra data such as footprints of large commercial buildings ('WikiProject Canada - OpenStreetMap Wiki', 2018). While the Canadian OSM dataset lacks complete spatial

and thematic descriptors, many of Canada's populous regions are currently covered in OSM due to the GeoBase and CanVec imports, while rural and remote regions remain sparsely mapped due to the lack of an open data infrastructure or an active mapping community ('WikiProject Canada - OpenStreetMap Wiki', 2018).

1.4 Aims & Objectives

The aim of this research is to evaluate the quality and reliability of OSM map features (road networks, buildings, etc.) in the Ottawa-Gatineau region.

The general aim will be accomplished by fulfilling the following specific research objectives:

- Evaluate the effectiveness of contemporary methods to assess the accuracy of the Ottawa-Gatineau OSM dataset relative to benchmark datasets.
- Analyze the spatial and temporal variation of the results based on contemporary methods.
- Parse the Ottawa-Gatineau OSM history file to analyze the chronological dataset evolution (i.e. dataset growth over time) and attribute structure.
- Characterize and identify OSM users based on their contribution characteristics and tendencies.

1.5 Research Question(s)

The fundamental questions driving this research will be:

- *What quantitative metrics can be used to evaluate the accuracy (completeness, positional, thematic, temporal) of the Ottawa-Gatineau OSM dataset?*
- *In addition to evaluating the effectiveness of contemporary methods, what specific mapping characteristics and metrics can be extracted from the history of the OSM dataset which would help to gain an understanding of OSM contributor experience and temporal evolution map features?*

1.6 Context

Much of the research surrounding the field of VGI accuracy and reliability assessment has been conducted in the context of European areas of interest. The works of Ather (2009), Ciepłuch et al., (2010), Girres & Touya (2010), Haklay (2010), Koukoletsos et al. (2012), Helbich et al. (2012), Fan et al. (2014) and Hecht et al. (2013) examined the positional and thematic accuracy of OSM data. Haklay (2010) postulated the foundation of assessing the quality of OSM data by drawing on past analysis conducted by Goodchild & Hunter (1997) which measured positional accuracy of linear features in the United Kingdom. Goodchild & Hunter (1997) implemented a strategy to assess the positional accuracy of linear features by conducting a buffer zone analysis. This strategy relies on a benchmark dataset that is assumed to have a higher positional accuracy than that of the dataset being analyzed. Haklay (2010) utilized this approach to assess the spatial/positional accuracy of OSM road network data in the United Kingdom in contrast to trusted Ordnance Survey (OS)

data. Girres & Touya (2010) furthered the work of Haklay (2010) by examining other spatial data quality elements (i.e. geometric, thematic, temporal) in the context of the France OSM dataset.

Hecht et al. (2013) further outlined a variety of methodologies to analyze and assess the quality of OSM building features in Germany compared to official data from the German national mapping agencies. Their study demonstrated that a transitional phase occurred within the quality assessment of linear map features (Haklay, 2010) to polygonal features (building footprints) provided by the German OSM dataset and the challenges which accompany assessing building completeness and positional accuracy. Hecht et al. (2013) reviewed unit-based and object-based comparison methods which analyzed OSM building features and the disparities in modelling between the two comparison techniques. Like Haklay (2010) and Girres & Touya (2010), Hecht et al. (2013) extended the overlap analysis of Goodchild & Hunter (1997) and applied it to polygonal features (building footprints). There has been more research conducted applied to matching buildings from OSM to an authoritative reference source, than assessing the quality of other metrics such as positional, semantic, attribute, shape and size accuracy (Brovelli et al., 2016).

When OSM was founded in 2004, it was expected that open and free map data would benefit relief efforts in humanitarian crisis conditions. This was proven to be true immediately following the 2010 Haiti earthquake and again in 2015 in Nepal, when large volumes of geospatial data were contributed to both regions in OSM ('Humanitarian OSM Team - OpenStreetMap Wiki', 2018; Poiani et al., 2016).

Collaboration between the local actors in Nepal and remote contributors helped with the verification and accuracy of the OSM map features generated. The Humanitarian OpenStreetMap Team (HOT) is one organization that participates in such efforts. HOT applies its belief in open-source and open data sharing for humanitarian aid response and economic development. Das & Alam (2014) also used OSM data to develop a location-based emergency management system wherein healthcare centres and hospitals were mapped by extracting their exact geographic location.

Much of the research conducted in this field does not account for the evolutionary history of OSM data. Data regarding the mapping activities of a community can also provide understanding behind mapping progress, which in turn can be used to evaluate quality (Rehrl & Gröchenig, 2016). Activities and behaviours of a mapping community can refer to map feature additions or deletions (i.e. adding a building polygon to the OSM dataset then a different OSM user removing that building polygon). Rehrl & Gröchenig (2016) have designed a technical framework that assists with structuring and analyzing mapping activities and behaviours based on the idea of activity theory. Barron et al. (2014) have also developed a similar framework to make OSM quality assessments based on the input data's history. This framework is unique compared to past research, since Barron et al. (2014) can make statements of OSM accuracy and quality without reliance on a benchmark dataset. Using each of these frameworks developed by Rehrl & Gröchenig (2016) and Barron et al. (2014), it is possible to gain a thorough understanding of the historical (temporal) accuracy and evolution of the OSM dataset under analysis.

The International Organization for Standardization (ISO) Technical Committee 211 (ISO/TC 211) put forth a set of quality measures and indicators which are used to assess the quality of geographic information. Quality measures refer to quantitative assessment strategies that employ comparison to "benchmark" or "gold standard" reference datasets. The "benchmark" or "gold standard" reference dataset is assumed to be the higher quality product. Quality indicators refer to qualitative aspects that affect quality which cannot be measured. For the purposes of this research, I will be drawing on quality measures put forward by the ISO to assess the accuracy and quality of OSM data within Ottawa-Gatineau.

1.7 Justification

The rationale behind this research is to validate whether OSM (or VGI in general) will be able to ultimately complement authoritative or proprietary geographic data sources used by government agencies instead of solely relying on use of single data sets. There has been an identified problem of accuracy and quality surrounding VGI (OSM). Since this research is relatively new in the context of Canada, the aim is to fill the gap between VGI and its accuracy challenges by examining Ottawa-Gatineau as a case study region.

There are many "fitness-for-use" cases for OSM and crowd-sourced mapping data. Some of these use cases include humanitarian aid for disaster and emergency response, location-based services (LBS) (routing, navigation, geocoding) and monitoring socio-economic development. While my thesis does not directly examine

VGI for such use cases, it is still important to discuss them and how they are relevant to this report and the field of VGI.

Natural disasters often occur in areas that are not well mapped, and not under constant monitoring. The citizens on the ground at the scene of the disaster are usually the best observers to the natural disaster events in real time as they occur. Since much of the geographic information in disaster scenarios is provided by volunteers, there can be an unwillingness to use the crowd-sourced geographic information due to credibility or reliability issues.

The use of OSM data for emergency response in Canada is gaining in popularity. During the spring of 2017, the Ottawa-Gatineau region saw extreme flooding due to record rainfall amounts. The Canadian Red Cross used OSM building data to validate the number of homes affected by extreme flooding. In the summer of 2017, parts of British Columbia entered a state of emergency due to the rapid movement of wild fires affecting communities across the province. OSM building footprint data were used to provide impact assessments to gain a greater understanding of the damage extent caused by the wild fires[2].

With the further development of LBS, the quality of OSM data (or VGI in general) must be assessed. A LBS can be defined as "any service or application that extends spatial information processing, or Geographic Information System (GIS) capabilities, to end users via the Internet and/or wireless network" (Jiang & Yao, 2006; Koeppel, 2000). LBS can have a wide variety of features such as geocoding,

[2] https://blog.mapbox.com/osmgeoweek-mapathons-support-building-canada-2020-eb0ca4edf8fe

navigation and routing, and point of interest (POI) searches. Geocoding is the process of associating exact geographic location with data (road names, addresses, house numbers) (Amelunxen, 2010). Efficient routing and navigation applications require accurate geocoding results to the most specific level (building level) (Barron et al., 2014). If addresses of buildings are inaccurate or incomplete, this will directly impact the routing results.

As a student employee at Employment and Social Development Canada (ESDC), I evaluated the quality of Canadian (Atlantic-Eastern Canada) OSM road networks relative to benchmark reference datasets, for the purposes of geocoding and routing. One of the main data use cases for road networks for the Geomatics division at ESDC is the 90-50 buffer analysis. This procedure aims to capture 90% of the Canadian population within 50-kilometre driving distances of Service Canada Points of Service (PoS). Currently, this procedure uses a proprietary road network dataset to generate 50-kilometre driving distance buffer zones from Service Canada PoS. The use of OSM could be used in the procedure to help aid the process where proprietary datasets lack road network completeness.

In August 2016, Statistics Canada (STATCAN) initiated a two-year pilot project aimed at understanding the potential of crowd-sourced geographic information for statistical purposes. The project uses OSM as a web-based platform for gathering information on buildings (attribute information, address, etc.) in the Ottawa-Gatineau region. When users contribute data on buildings and locations, they are supporting the efforts of first responders, relief efforts and government to develop informed

policies and programs. STATCAN traditionally seeks data through citizen surveys. By using OSM as a platform for crowd-sourced information, STATCAN is exploring new and innovative methods of data collection that have not been previously used by the Government of Canada. The future goal of STATCAN is to have accurate national-level statistics on buildings and their attributes that can be used to compare specific regions[3]. While I am not officially affiliated with this project at STATCAN, the Data Exploration and Integration Lab (DEIL) have expressed interest in my research.

Thus, by developing methods to evaluate the accuracy (completeness, positional, thematic, temporal) and reliability (mapping characteristics and behaviours) of OSM data over a spatial and temporal scale, it is possible to justify "fitness-for-use" scenarios for OSM data and augmentation or placement of current data holdings in government organizations. This research will assist in the overall acceptance of VGI in general, both in the public and private sectors, and contribute to new approaches that reduce reliance on "authoritative" or "gold standard" datasets.

1.8 Thesis Structure

This thesis starts with a literature review in Chapter 2 that introduces existing VGI research, specifically classification types and quality assessment strategies of VGI. Methods used to conduct the research are outlined in Chapter 3, followed by results in Chapter 4 and discussion in Chapter 5. Finally, Chapter 6 includes concluding remarks, research limitations and potential future work in the field of OSM.

[3] https://www.statcan.gc.ca/eng/crowdsourcing

2 LITERATURE REVIEW

This review of relevant literature begins by examining Haklay et al.'s (2014) investigation into the use of VGI. Sections 2.2 through 2.4 outline the key work of Senaratne et al. (2017), reviewing two main classification categories of VGI and four approaches to assuring the quality of VGI being provided. Section 2.4 briefly outlines other relevant literature surrounding the field of VGI, including both ISO standardized quality assessment measures (quantitative) and indicators (qualitative). Hecht et al. (2013) departs from traditional quality assessment strategies of comparing linear features to that of comparing polygon geometries and how to quantify quality regarding building footprints; to signify the importance of such a paper and its importance surrounding OSM building quality assessment, Section 2.5 will be dedicated to discussing that research.

2.1 Volunteered Geographic Information Use in Government

With VGI platforms like OSM growing in public popularity, so is the interest of government departments in using these free and open data sources. Within the growing field of research being conducted into verifying the reliability and accuracy of VGI, a recurring theme is hesitation to make a transition from solely utilizing authoritative or proprietary datasets to including VGI. In order to provide a further understanding of this, Haklay et al. (2014) investigated the use of VGI in government. Throughout this case study, the authors provide possible direction for the successful implementation of VGI into government workflows.

One of the major areas for governments to transition using VGI was cases where government-held data resources were lacking. This was observed in the case of Indonesia following significant flooding events where disaster management agencies and HOT used the OSM project as a platform to crowdsource geographic information. Initially, the leaders of 267 urban villages of Jakarta, Indonesia were questioned about important infrastructure in their villages, which was then mapped into the OSM dataset by community members and university students. A notable secondary side effect of these efforts resulted in community members becoming interested in developing a detailed basemap of Jakarta which fostered a growing interest in crowdsourced mapping through OSM (Haklay et al., 2014).

The use of VGI in government has also been proven valuable within Canada (Coleman et al., 2009). With the increasing difficulties (financial and otherwise) being encountered by national mapping agencies to maintain the currency of government datasets, departments such as NRCan have explored the use of OSM. For this case, NRCan wanted to familiarize its employees with OSM data quality and workflows. The process involved NRCan releasing its in-house geographic data into OSM (CanVec) and developing a change detection procedure which would keep its national road network dataset up to date; However, there are incompatibilities between the license and OSM data terms of use and the intellectual property rights of Canadian government that need to be settled before OSM data is used directly by government organizations (Haklay et al.,2014).

Shifting from using authoritative or proprietary datasets to VGI has had its proven successes and challenges, which have shown that there needs to be a well-defined use case for the data and its sources. Rather than solely transiting from a government generated dataset to VGI, a more appropriate solution would be using VGI to complement or augment a commonly used dataset. Haklay et al. (2014) stated that maintaining user participation remains one of the greatest challenges. Typically, at the beginning of a VGI project, user participation and interaction are at its highest but at some point, people begin to lose interest. Maintaining user participation over longer periods of time will continue to be one of the challenges encountered by VGI projects which will require government and community interaction initiatives.

2.2 Volunteered Geographic Information Classification

VGI can be classified into two main categories: explicit and implicit volunteering or geography. Antoniou et al. (2010) outlined that spatially explicit applications of VGI are when the contributors are primarily focused on the spatial features of an area or mapping activities. Spatially explicit activities are most prominent within the OSM project (i.e. mapping roads or building footprints of a geographical location). Spatially implicit VGI is associated with types of media (images, text, videos) that have a geographic location associated with them. The main difference comes from the fact that spatially implicit VGI contains a geo-tagged location but this is not the primary source of information provided, rather it is an additional feature to the media form. Craglia et al. (2012) summarized spatially explicit and implicit VGI by stating "if a particular piece of information is about the

characteristics of a place it is explicitly geographic, if a piece of information is not about a place and can still be geo-tagged it is implicitly geographic."

2.2.1 Map-based VGI

Map-based VGI includes point, line and polygon features such as those representing the basic map features. As previously stated, OSM is one of the most prominent open map-based VGI platforms. Many map-based VGI platforms are used for navigation or point of interest (POI) searches. These POIs may include amenity locations such as hospitals, grocery stores, schools, daycares, or restaurants, and these are map features that are sometimes insufficient in authoritative datasets; whereas although VGI may have more such features, the quality of the information/data contained in the OSM dataset can be variable due to its open feature classification scheme and the inability to properly assert whether the data contributors are geospatial experts. It is the culmination of these two factors which typically generates the quality related issues surrounding thematic (attribute) accuracy in the dataset.

2.2.2 Image-based VGI

Image-based VGI is produced implicitly with web or mobile portals. Some examples of image-based VGI include the Flickr and Instagram platforms. Images are taken and uploaded to the portal with a geo-tagged spatial location associated with the image. Tagging an image simply means adding metadata to the content in the form of specific keywords or adding a geographic location to the image (Golder & Huberman, 2006; Senaratne et al., 2017; Valli & Hannay, 2010). Environmental

monitoring, human navigation and event trajectory analysis are some of the data use cases associated with image-based VGI (Senaratne et al., 2017); However, there are also quality issues associated with image-based VGI. Incorrectly geo-tagged photos can cause reduction in information quality since the photo could be tagged at a location several kilometres away from an event or where the picture was taken. These issues arise when the user is not zoomed in to the exact location view in their map portal when tagging the image (Senaratne et al., 2017).

2.2.3 Text-based VGI

Text-based VGI is another form of spatially implicit volunteering. Within text-based platforms, users can contribute geographic information in the form of text through blogs and social media networks. Like image-based VGI, text-based VGI can have geo-tagging capabilities where the geographic location is not the primary source of information. Detection of disease propagation, journalism, politics, and general event detection are some of the data use cases for text-based VGI (Senaratne et al., 2017). One of the key issues surrounding text-based VGI is that much of the information is considered "spam". Due to the high volume of information on platforms like Twitter, this exponential volume of VGI contribution adds to the difficulties associated with assessing the quality of this form of VGI (Senaratne et al., 2017).

2.3 Approaches to Volunteered Geographic Information Quality Assurance

Goodchild & Li (2012) propose three different approaches which can be used to aid in assuring the highest quality data is contained within VGI. These approaches are the crowd-sourcing, social, and geographic. This chapter also discusses a fourth approach to quality assurance of VGI involving data mining techniques.

2.3.1 The Crowd-Sourcing Approach

The crowd-sourcing approach is the act of validating and correcting contribution errors from past contributors. The OSM project relies heavily on crowd-sourcing activities to assure the quality of its geographic information. Crowd-sourcing can also refer to the ability of a crowd to eventually meet the truth (Goodchild & Li, 2012). This concept is also known as Linus' Law, an ideal practice in software engineering and states "given enough eyes, all bugs are shallow". Linus' Law is primarily used in respect to software development and relies on the practice that if many individuals use a software program continuously, the bugs and errors will be identified and can then be eliminated by software engineers (Goodchild & Li, 2012; Raymond, 2001). Haklay et al. (2010) employed Linus' Law as a basis for examining the positional accuracy of OSM map features, and it was found that accuracy improves as the number of editors increases, but only to a certain point (around 5-13 contributors) within a designated spatial area.

2.3.2 Social Approach

The second approach to quality assurance of VGI, as proposed by Goodchild & Li (2012), is the social approach. The social approach relies on a defined set of individuals to act as gatekeepers or moderators on VGI platforms. Goodchild & Li (2012) and have found that voluntary contributions follow a skewed frequency distribution with a select few individuals making large numbers of contributions and most individuals making very few contributions. Nielson (2006) refer to this behaviour as the 90:9:1 rule with regards to social media interaction, 90% of the users are individuals who never contribute, 9% contribute very little and 1% of the users almost always contribute. Mooney & Corcoran examined the characteristics of heavily edited features in OSM from the British Isles in 2012. The study by Mooney & Corcoran (2012) indicated that in the case of heavily edited features, 84% of the edits were found to be from only 12% of the contributors.

The social approach to quality assurance can be extended to the classification types of users contributing information. This approach has been implemented on platforms like Wikipedia with the introduction of the concept of administrators and moderators. Wikipedia administrators are charged with assuring the quality of articles by reviewing/editing them and removing copyrighted materials, abusive content and outright vandalism. The OSM project presented a similar classification of users. Contributors of OSM data are classified as regular users or Data Working Group (DWG) users. Much like Wikipedia administrators, the OSM DWG has

identified twelve active members (as of November 2017[4]) which deal with vandalism, copyright, bots and disputes. This contribution control model ensures that every user of the OSM dataset can contribute and edit geographic features but it is the DWG which will have the last call if a dispute over the information arises (Goodchild & Li, 2012).

2.3.3 Geographic Approach

The final VGI quality assurance approach proposed by Goodchild & Li (2012) is the geographic approach. This approach refers to (local) general geographic knowledge of an area and application of the fundamental laws of geography. It is applied under the understanding that an individual from Ottawa, Ontario contributing to the Ottawa portion of the OSM dataset would be more knowledgeable/have enhanced local knowledge and sense of features in the region than somebody not from Ottawa. The geographic approach primarily depends on existing geographic laws; specifically, Tobler's First Law of Geography states "All things are related, but nearby things are more related than distant things" (Goodchild & Li, 2012; Tobler, 1970).

2.3.4 Data Mining Approach

As discussed, a fourth approach was proposed by Senaratne et al. (2017) which is independent of crowd-sourcing, social and geographic VGI quality assurance strategies. This approach is known as the data mining approach. Data mining entails collecting copious amounts of data and analyzing the data to establish patterns and

[4] http://wiki.osmfoundation.org/wiki/Data_Working_Group

form heuristics about the data at hand. Basiri et al. (2016) focus on the data mining strategy to analyze user trajectory movements to form rules which are then used to check and validate an OSM dataset. For example, it is seen that if an entity's trajectory includes moving at an average speed of 50 km/hr, it can be concluded that the entity's method of travel is likely by car. This study also gathered information on trajectory intervals between stop and start times and determined that if a trajectory has a 15-30 second interval stop, it could be concluded that the travel mode could have been a bus route. The conclusions of this work, according to Senaratne et al. (2017), validate this approach of quality assurance and will eventually be used as a widespread measure among the VGI community in the future along with the historical evolution of a geographic dataset.

2.4 ISO Standardized Quality Measures and Indictors

As previously stated, the ISO/TC 211 put forth a set measurable of quality indicators which can be used to assess and assure the quality of VGI inputs. Section 2.4.1 reviews literature related to the standardized quality assessment of VGI (both quality measures and indices). Following this, Section 2.5 examines the work of Hecht et al. (2013) in greater detail.

2.4.1 Quality Measures

2.4.1.1 Completeness

Completeness of geographic information refers to the "relationship between actual represented objects and their conceptualizations". Completeness can be

measured in the absence of data (errors of omission) and an excess of data (errors of commission) as discussed in Senaratne et al. (2017).

Koukoletsos et al., (2012) created a feature-based automated method of matching linear data between VGI and a reference dataset. The VGI dataset used was OSM road networks and the dataset of measure was the Integrated Transport Network (ITN) layer provided by the United Kingdom's national mapping agency, Ordnance Survey (OS). The methods imposed by Koukoletsos et al. (2012) are unique to the quality assessment of VGI field because this method combines geometric and attribute constraints (road name and type) to deal with heterogeneity between the two datasets. Koukoletsos et al. (2012) found that OSM proves to be more complete in urban areas and sparser in rural, like the results of Haklay (2010) in the context of the United Kingdom (Senaratne et al., 2017).

2.4.1.2 Accuracy

Accuracy is the degree of closeness between a measurement and the accepted true value and is in the form of positional, thematic and temporal accuracy (Senaratne et al., 2017). Positional accuracy of geographic features refers to the accuracy of positional features (i.e. how close is an object to the real-world representation?). Thematic (or attribute) accuracy refers to the accuracy of quantitative and qualitative attributes of a geographic dataset or database (i.e. is the map feature or object exactly what we believe it is?) (Senaratne et al., 2017; van Oort, 2006).

Ather (2009), Girres & Touya (2010) Haklay (2010), Kounadi (2009), Ueberschlag (2010) all examined positional accuracy of OSM road network data

relative to a reference dataset. Each of these studies were based on the work of Goodchild & Hunter (1997), which developed an estimation algorithm to provide an estimation of overlap between reference datasets and a VGI dataset (i.e. OSM). It was found that there is on average 80% overlap with OSM road network data and the reference dataset, although these values can range from 50 to 100% based on the geographic region under review (Haklay et al., 2010).

Mooney & Corcoran (2012) stated that most of the errors contained in the OSM dataset are caused by mis-classification and misspelling of feature values. Girres & Touya (2010) also found that most attributional spelling errors are rooted in the mis-specification of road networks. An example of this would be when road networks were tagged as a secondary in the reference dataset and residential or tertiary in the OSM dataset (Senaratne et al., 2017).

There are several tools to help the average OSM contributor by assisting with classification tagging and editing of their contributions. For example, Vandecasteele & Devillers (2013) developed and trialed a tag recommendation application for OSM data, to reduce semantic and spelling errors when contributors are mapping OSM features.

2.4.1.3 Temporal Accuracy

Temporal accuracy refers to the accuracy of a dataset relative to historical changes in the real-world. Much of the current work surrounding quality assessment of VGI does not account for the temporal aspects of the dataset's content accuracy. Due to this fact, a paradigm shift is beginning to occur in the understanding of the

importance of this aspect of the contributions and the global audience is currently witnessing the creation of tools designed specifically to review VGI dataset evolution. Girres & Touya (2010) inspect the evolution of OSM map features over a three-month period by percentage of growth. The inherent problem with this assessment approach lies in the fact that OSM dataset growth and evolution refers to the addition of new objects, rather than the update or changes to existing map features. Rehrl & Gröchenig (2016) propose a technical framework for analyzing mapping theory in the context of VGI (OSM). The developed framework builds on a theory called "Activity Theory" as a framework for mapping activity for a map contributor level. This study benefits in bridging the gap between contributor driven models and GIS task-oriented models for VGI quality assessment.

2.4.1.4 Topological Consistency

Consistency refers to the coherence in the data structure of the digitized spatial data (Senaratne et al., 2017), while topological consistency can refer to the overlap of map features (i.e. building overlaps), as well as overshoots and undershoots (i.e. road network exaggerations or deprecations). With many moving parts and contributors to VGI platforms (OSM), maintaining consistency among a dataset remains a challenge. Researchers in the field of VGI have proposed methods to assess consistency mainly through intrinsic dataset reviews to detect problems in datasets (Senaratne et al., 2017). Corcoran et al. (2010) recommend methods to assess and check geometric inconsistencies of vector data through the definitions of planar and non-planar relationships across map features. Other publications have focused on evaluating topological consistency of road networks through the

development of automated matching algorithms to match data from OSM and reference datasets (Senaratne et al., 2017; Will, 2014).

2.4.2 Quality Indicators

2.4.2.1 Lineage, Usage, Purpose

Lineage refers to the evolution of a dataset; from data collection, acquisition, compilation and storage of the current data file extension (Guinée, 2002; Hoyle, 2001; Van Oort & Bregt, 2005), and as with temporal accuracy, qualitative accuracy assessment strategies surrounding VGI are lacking. It is difficult to account for intrinsic characteristics such as reputation, credibility and local knowledge in a large dataset like OSM. Many publications have outlined what VGI projects could do to keep track of data usage and lineage. Girres & Touya (2010) recommend that moderators keep track of map feature contributors for data source metadata (i.e. source=City of Ottawa, source=NRCan-CanVec-8.0, etc.). Similar methods have been suggested by Keßler & De Groot (2013), which examines trust and reputation as a proxy to assess consistency, thematic accuracy and completeness of a VGI dataset, as opposed to ground truthing or benchmarking methods traditionally used (Senaratne et al., 2017).

2.5 Hecht et al. (2013)'s Approach to Evaluating Building Completeness over Space and Time

Hecht et al. (2013) designed a variety of methodologies to analyze and assess the quality of OSM building features in Germany compared to official data from national mapping agencies. The areas of interest used for comparison in this study were the North Rhine-Westphalia and Saxony regions. Within North Rhine-Westphalia and Saxony, a city, medium-sized town, small town and rural district were selected. Leipzig, Chemnitz, Bautzen and Volgtlandkreis were focal areas selected for the Saxony region. Essen, Munster, Lemgo and Kreis Coesfeld were the focal areas selected for the North Rhine-Westphalia region. The OSM dataset was downloaded from the OSM service provider, Geofabrik, as of 17 November 2011, 24 May 2012, and 5 November 2012. The national building dataset, Amtliche Hausumrine (AH), was used as a reference building polygon dataset in the North Rhine-Westphalia region. Due to completeness issues in the Saxony region, the Authoritative Topographic Cartographic Information System (ATKIS) Base Digital Landscape Model (BDLM) was used as a reference.

This study demonstrated that a transitional phase has occurred from the quality assessment of linear map features (Haklay, 2010) to polygonal features (building footprints) in OSM and the challenges that are intrinsic with assessing building completeness and positional accuracy. Hecht et al. (2013) further elaborates on unit-based and object-based comparison methods of analyzing OSM building features and the disparities in modelling between the two comparison techniques.

The study conducted by Hecht et al. (2013) was one of the first to assess the completeness and positional accuracy of OSM buildings in any geographic area. The conventional approach of OSM quality assessment is to assess the quality of linear features, usually road network data; However, Hecht et al. (2013) went outside the typical norm to evaluate completeness on a much larger scale. Like Haklay (2010), Hecht et al. (2013) does not limit the comparative study to one homogeneous geographical location within Germany. There were two federal states selected, one of the East (Saxony) and West (North Rhine-Westphalia). During this process, they observed the spatial variability between urban and rural building completeness in two distinct locations.

Temporal variability is one dimension that is not accounted for in many publications surrounding quality assessment of OSM. By gathering OSM data from three different time periods (November 2011, May 2012, November 2012), Hecht et al. (2013) can observe the growth of the dataset over an extended period. This temporal dimension is not accounted for in the works of Ather (2009) and Haklay (2010), which examines OSM data at a static period. However, Girres and Touya (2010) examine the growth of the France OSM data over a 3-month period (June 2009 and October 2009). While Girres and Touya (2010) account for this temporal variability, they do not examine the spatial variability of the dataset growth (i.e. where the dataset growth is occurring). Girres and Touya (2010) simply examined unit-based map feature growth of road network segments over 3-months.

While Hecht et al. (2013) discover an increase in building completeness in each German region, will there eventually be a point where OSM building completeness is halted? Is there a point where OSM quality cannot improve any further to represent real world geographic features? As this is a living dataset which is updated in an irregular fashion, these are some of the underlying temporal questions that should be brought forward.

Overall, this study contributes significantly to the advancement of quality assessment and assurance of OSM data because Hecht et al. (2013) departs from the traditional methods to assess the quality of OSM data. Prior to 2013, most of the comparative studies used linear map features (road networks) to compare to a reference dataset. Hecht et al. (2013) proposed an innovative method of object-based comparison in modelling that was unheard of in previous studies of OSM quality assessment. This study was not limited to one small scale geographic location or one time. The OSM data was collected from a variety of different time periods over one year. Therefore, this publication feeds directly into the research objectives outlined in Section 1.4 in its ability to analyze the spatial and temporal variation of OSM building completeness.

2.6 Summary

Quality assessment and assurance of VGI is not entirely new to the field given the emergence of the term in 2007 by Goodchild. Shortly thereafter, quality assessment and assurance strategies surrounding VGI began to take precedence among the development of methodologies that aim to quantify quality and reliability

through benchmark comparison analysis (Haklay, 2010). With an increase in literature surrounding VGI, the widespread exploration of VGI datasets for governance will also rise in popularity. Haklay et al. (2014) found that there have been successful cases of public and government interaction involving VGI projects, which include land management, biodiversity and disaster response investigations. To assist in the transition between VGI datasets and current government data holdings, government agencies will have to continue to consider quality assessment, data maintenance and manipulation, and VGI user participation to augment or potentially replace authoritative datasets.

3 METHODOLOGY

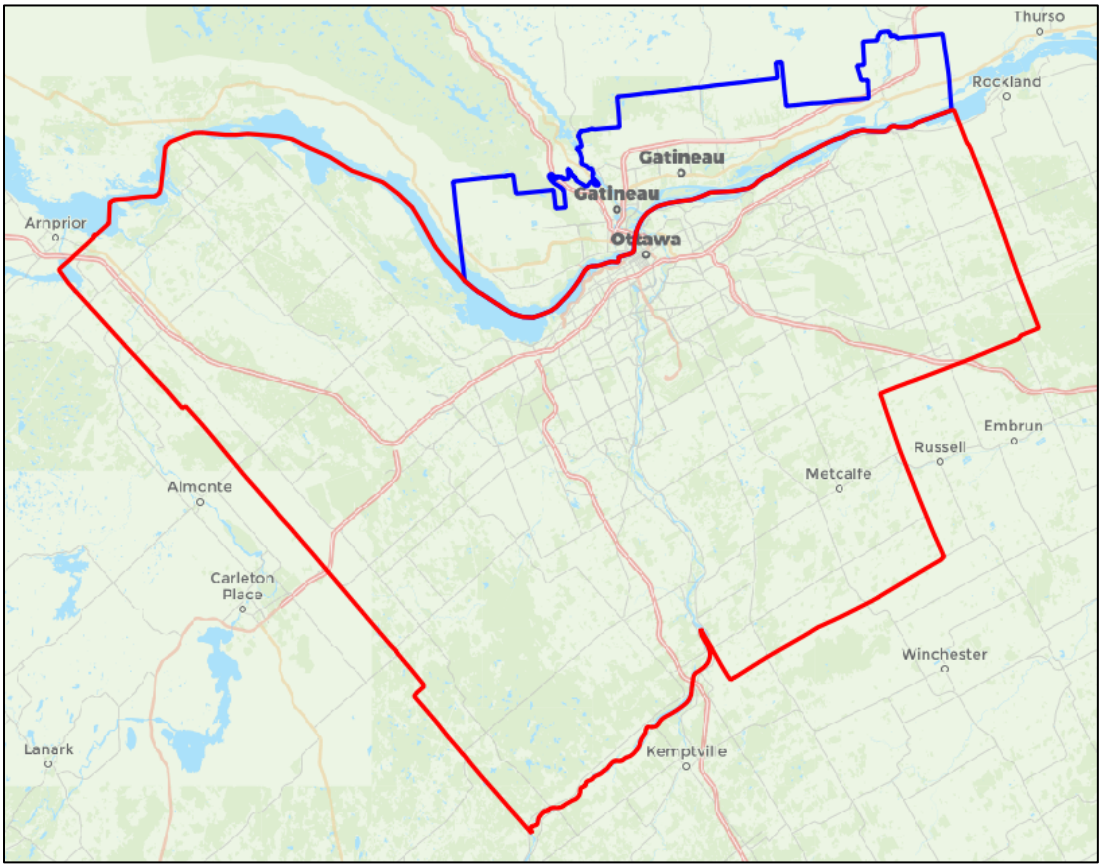
This chapter outlines the designed methodology used to address the research questions in Section 1.5. Section 3.1 provides contextual information on the study area selection and how the data used for this research was collected. Section 3.2 outlines the quality measures that were used to assess the quality of OSM data. Section 3.2.1, 3.2.2, 3.2.3 and 3.2.4 will include comprehensive definitions of each of the quality measures as defined by van Oort (2006) in the publication, "Spatial data quality: from description to application", and the procedures taken to quantify accuracy and quality. Please refer to Annex A for figures regarding QGIS models, scripts, etc.

3.1 Study Area & Data Collection

The Area of Interest (AOI) selected for this research was the Ottawa-Gatineau region within Canada. The level of geography used to extract OSM data from Ottawa-Gatineau was the corresponding Census Subdivisions (CSD). CSD is the general term for municipalities or areas treated as municipal regions for statistical purposes, as defined by Statistics Canada (2005). The Ottawa-Gatineau region was scrutinized due to its vast spatial area (Ottawa 2790.30 square kilometres, Gatineau 342.96 square kilometres) and heterogeneous population distribution (Ottawa 334.8 persons per kilometre squared; Gatineau 773.7 persons per kilometre squared) (Government of Canada, 2017a; Government of Canada, 2017b).

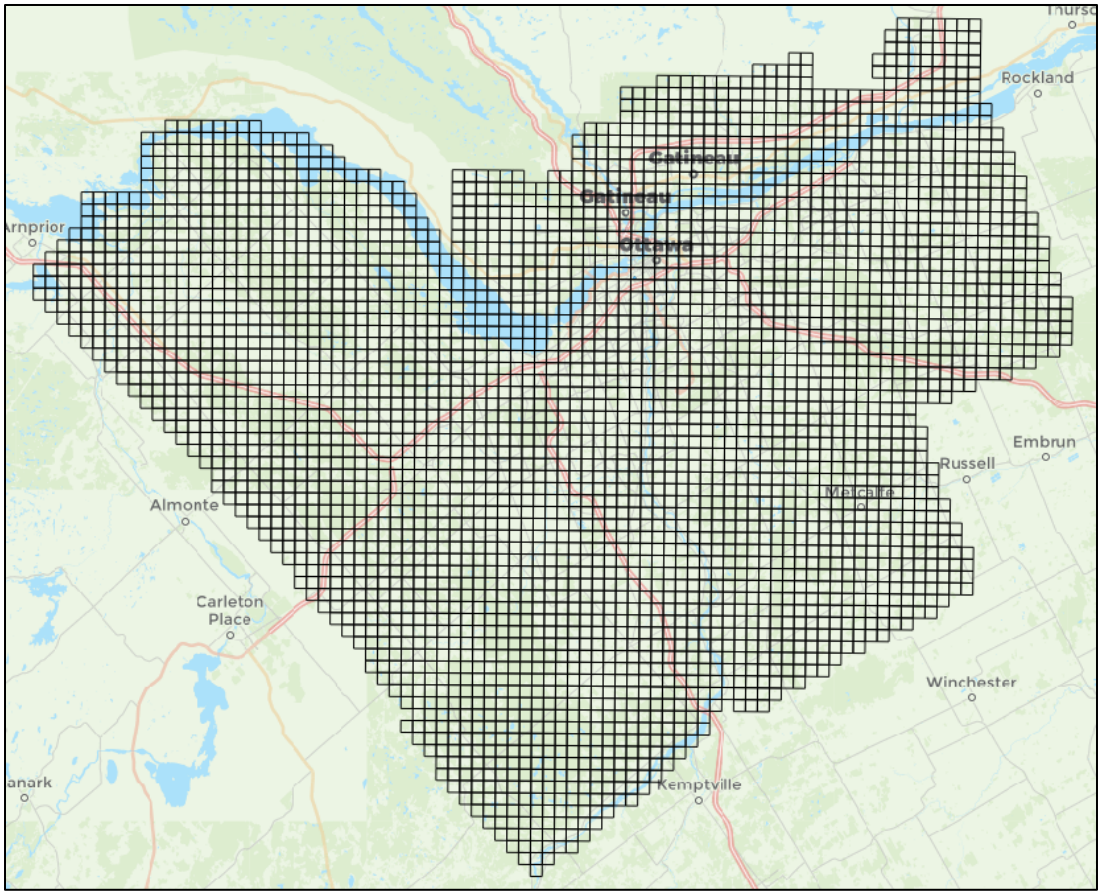
The investigation into the Ottawa-Gatineau OSM dataset was exploited to show the variance in OSM data quality between rural and urban communities. Figure

3.1 shows the respective Ottawa (red) and Gatineau (blue) CSD boundaries. Grid cell regions were also created using the Ottawa-Gatineau CSD boundaries to visualize spatial variation among certain quality measures; see Figure 3.2 for the 1-kilometre grid used.



Maps© Thunderforest, Data© OpenStreetMap contributors.

Figure 3.1: Ottawa (red) and Gatineau (blue) CSD geographies.



Maps© Thunderforest, Data© OpenStreetMap contributors.

Figure 3.2: 1-kilometre grid cell representation of the Ottawa-Gatineau CSD boundaries.

The data used for this research were collected from a variety of sources. The OSM data were collected from the Geofabrik gateway which generates daily OSM data snapshots from OpenStreetMap.org. Geofabrik is a web-based portal that provides OSM data from the continent, country and provincial/state administration levels. Geofabrik provides the snapshots of OSM data in a variety of formats; .osm.pbf, .osm, .osh.pbf, .poly, .osc.gz, and .shp. The OSM snapshots were downloaded from Geofabrik for 1 January 2016, 10 January 2017 and 20 June 2017. Downloading data from different time frames allows for examination of the temporal evolution of an OSM dataset. Once the OSM data were downloaded, the Osmium library was used to

process the OSM data using the Ottawa-Gatineau CSD boundaries as a clip file. The dataset that was used as a "gold-standard" or comparative reference dataset was the DMTI Spatial Inc.™ CanMap® Suite (DCMS) 2016 and 2017 dataset. DCMS 2016 was acquired through with the Maps, Data and Government Information Centre (MADGIC) at Carleton University. Quality assessment results using DCMS 2017 reference data were generated during the summer of 2017 at ESDC in Gatineau, Quebec. For further information on the quality or lineage of the DCMS reference dataset, please refer to the 2018 DMTI Spatial Inc.™ CanMap® Suite Data Dictionary[5]. The DCMS was used in this research as a reference dataset because it is a proprietary dataset that costs money, therefore, there is the assumption that it is of higher quality.

OSM data snapshots from 2017 (January and June) were compared to DCMS 2017, and OSM data snapshots from 2016 was compared to DCMS 2016. Table 3.1 outlines the quality measures and map features that were assessed, the reference datasets used and the corresponding OSM data that were evaluated.

[5] https://github.com/ktjaco/misc/raw/master/docs/DMTI_Data_Dictionary_2018.pdf

Table 3.1: Overview of quality assessment measures and OSM data evaluated relative to corresponding reference datasets.

Quality Measure	Map Feature	Reference Data Used	OSM Data Evaluated
Completeness	Road Networks	DCMS 2016	January 2016
		DCMS 2017	January 2017
			June 2017
	Geocoding	Ottawa 2014 Municipal Election Voting Locations	January 2016
			January 2017
			June 2017
Positional Accuracy	Road Networks	DCMS 2016	June 2017
		DCMS 2017	January 2017
			June 2017
	Buildings	DCMS 2016 (Ottawa only)	January 2017
		DCMS 2017	June 2017
	Geocoding	Ottawa 2014 Municipal Election Voting Locations	January 2016
			January 2017
			June 2017
Thematic Accuracy	Road Networks	DCMS 2016	January 2016
		DCMS 2017	January 2017
			June 2017
	Buildings	DCMS 2016	January 2016
		DCMS 2017	January 2017
			June 2017

3.2 Quality Measures

3.2.1 Completeness

Completeness of geographic information refers to the presence or absence of features in a dataset (i.e. how inclusive is the dataset of real-world features?). Completeness was examined in this research by comparing total street network length relative to the DCMS road network dataset and batch geocoding match rate completeness.

3.2.1.1 Road Networks

For the purposes of this research, the road networks from OSM and DCMS were categorized into major and minor road network types. Further analysis was conducted to also include all road network types from each of the datasets. Table 3.2 outlines the road network classification values used from each of the datasets.

Table 3.2: Road network classifications for OSM and DMTI datasets.

Type	Dataset	Field Name	Field Value
Major	OSM	Highway	Motorway, Primary, Secondary, Trunk
	DCMS	CARTO	1, 2, 3, 4
Minor	OSM	Highway	Tertiary, Unclassified, Residential, Living-Street, Pedestrian
	DCMS	CARTO	5

Equation 1 outlines the unit-based comparison calculations required to assess the completeness of OSM road networks relative to a reference dataset (DCMS). For

example, Road Network[1] represents the OSM Road Network and Road Network[2] is the benchmark dataset (DCMS).

$$C_{Length} = (\sum Road\ Network^1 - \sum Road\ Network^2) \qquad \textbf{(1)}$$

Figure 3.3 illustrates a flow diagram of the equation implemented to assess the completeness of each of the road networks. The AOI regions under investigation are the Ottawa-Gatineau CSD boundaries with the generated 1-kilometre grid cell regions.

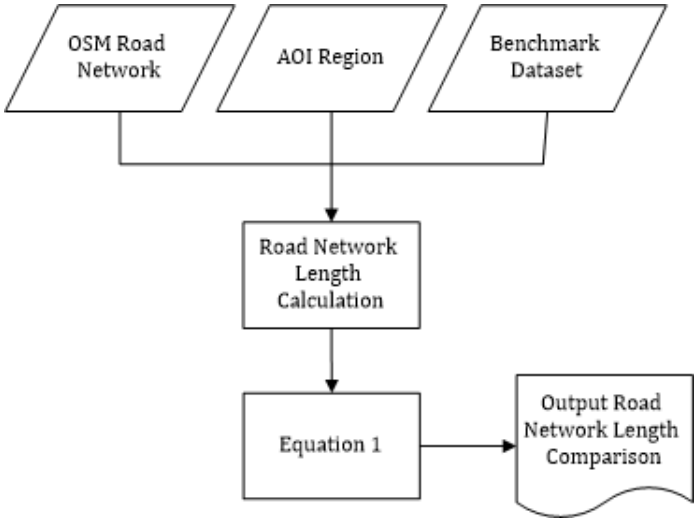


Figure 3.3: Flow diagram illustrating the procedure to compare road network length amongst two datasets.

3.2.1.2 Buildings

To gain an understanding of OSM building completeness, two object-based comparisons were made between the OSM data and DCMS data. The first method analyzes overlap proportion between OSM and DCMS building footprints with the threshold that at least 50% of the DCMS building footprints area overlapping an OSM building (Equation 2). Fan et al. (2014) applied a similar overlap method to their research based on the work of Rutzinger et al., (2009) which specified that matching

may be incorrect if the overlap is less than 30%. The 30% threshold accounts for the imagery offset from using Bing satellite imagery for the OSM dataset and for how larger buildings will have a larger area overlap. Figure 3.4 outlines the GIS computations required to calculate overlap proportion between two building footprint datasets.

$$C_{Overlap} = \left(\frac{\sum Building_{Ref\ in\ OSM}}{\sum Building_{Ref}}\right) \times 100 \qquad \textbf{(2)}$$

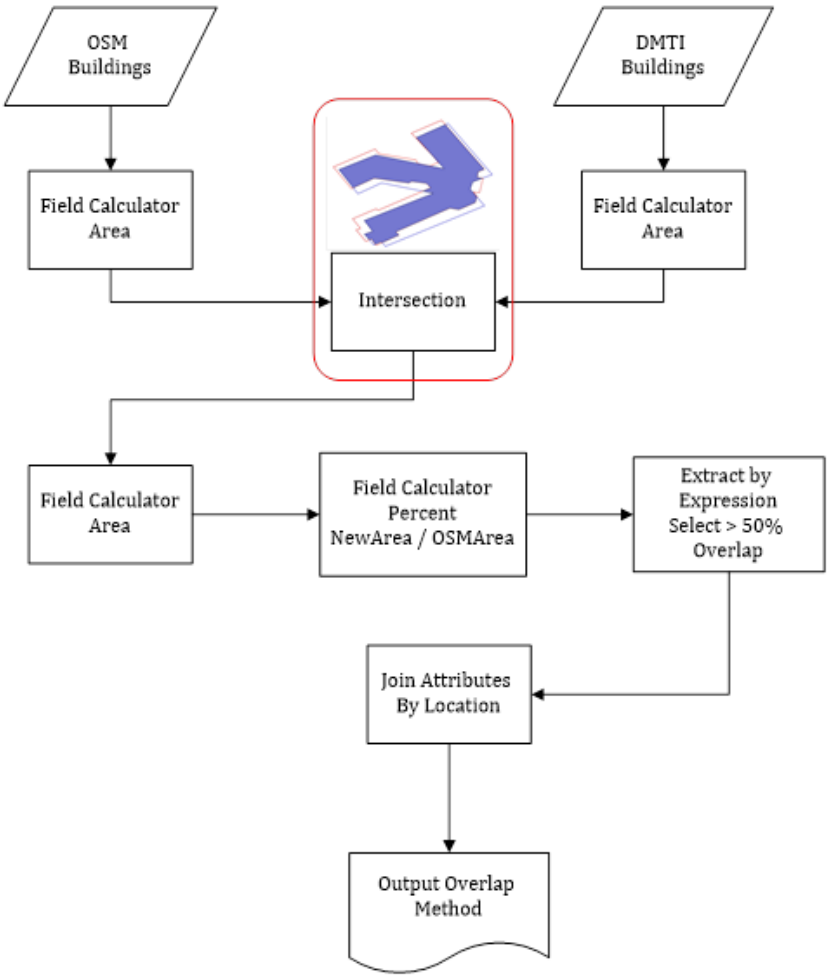


Figure 3.4: Flow diagram depicting the GIS computations used to calculate overlap proportion between intersecting OSM and DMTI Spatial Inc. building footprints. NewArea represents the area calculated from the Intersection algorithm. OSMArea represents the area calculated from the original OSM buildings dataset.

The second object-based approach to quantifying building positional accuracy follows Hecht et al.'s (2013) centroid comparison between two building datasets. The methodology involved computing the centroids for each of the DCMS building polygons, selecting the centroids that intersect OSM building polygons, then finally calculating a proportion percentage between selected centroids intersected by OSM buildings and total centroids of the DCMS building dataset. Equation 3 outlines the proportion percentage calculation used to for the centroid comparison approach. Figure 3.5 shows the GIS computations used to generate the centroid method output.

$$P_{Centr} = \left(\frac{\sum Centroid_{Ref\ in\ OSM}}{\sum Centroid_{Ref}}\right) \times 100 \qquad \textbf{(3)}$$

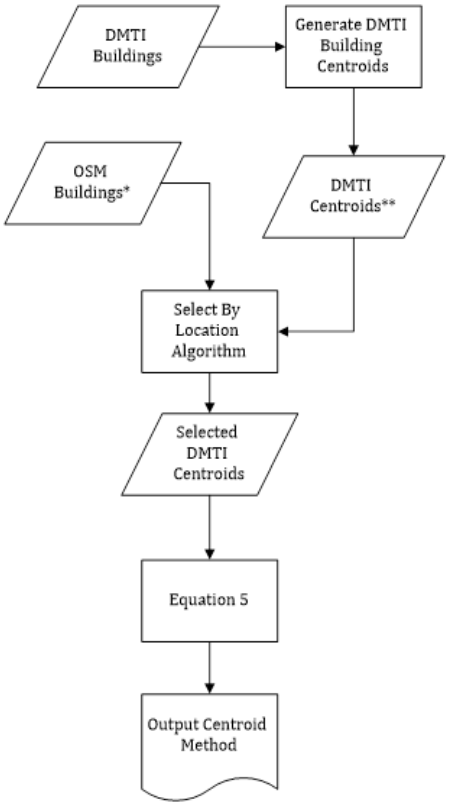


Figure 3.5: Flow diagram depicting the GIS computations used to calculate centroid proportion between reference building footprints within OSM buildings.

3.2.1.3 Geocoding

Geocoding completeness was analyzed by calculating match rates using OSM data from January 2016, January 2017 and June 2017. Nominatim is a tool developed to search OSM data by name and address (geocoding) and to generate addresses of OSM geographic coordinates (reverse geocoding). Using Oracle VM VirtualBox, three separate Ubuntu Linux 16.04 Nominatim instances were built with Docker using OSM snapshot data (.osm.pbf) from January 2016, January 2017 and June 2017. Figure 3.6 shows the architecture of the VM's that were used for batch geocoding. A Python script was written to batch geocode the voting locations for the 2014 City of Ottawa municipal election (Annex A – Appendix 5). The 2014 voting locations dataset was used as a benchmark because it of its generous size (571 voting locations) and the presence of "ground-truth" coordinates already appended to the address dataset file (Figure 3.7).

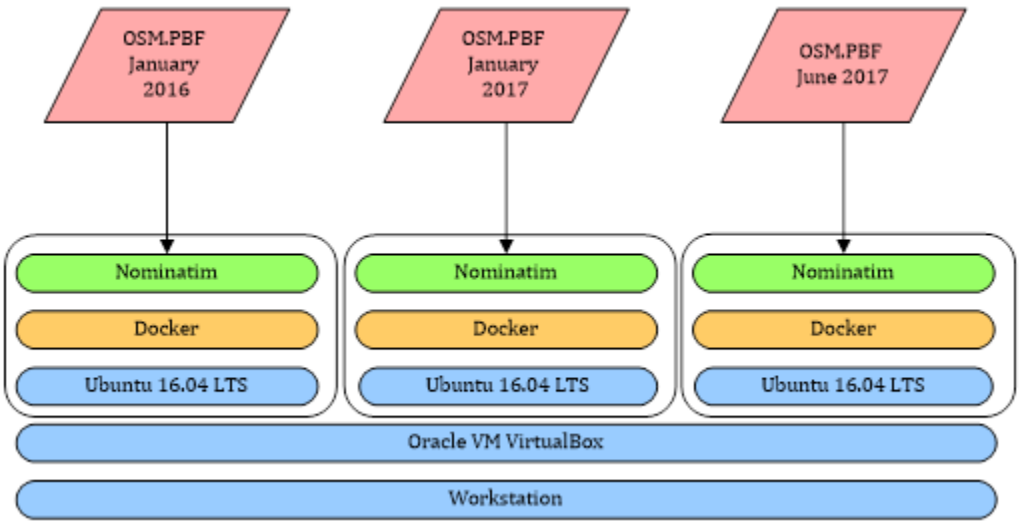
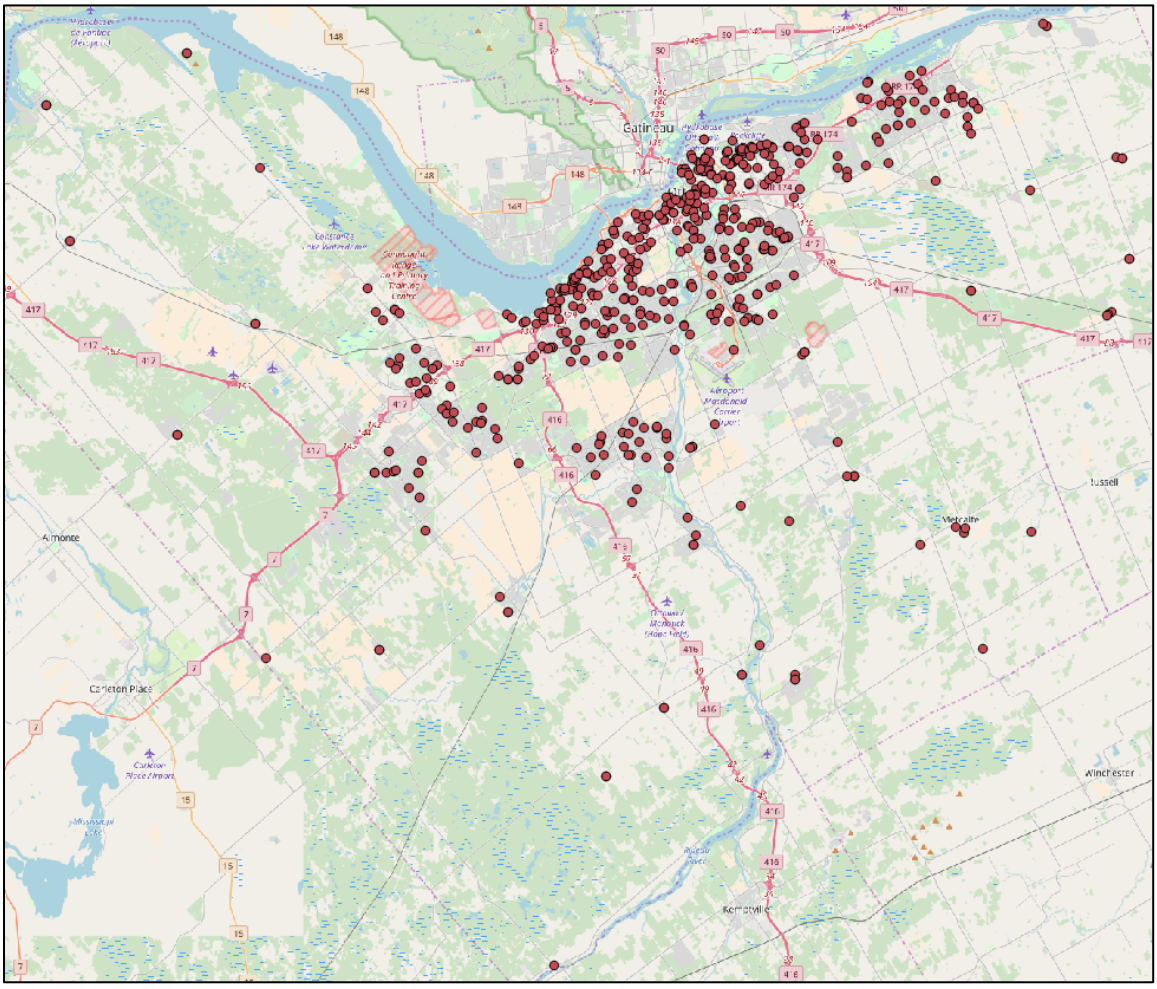


Figure 3.6: OSM Nominatim geocoding architecture at the operating system (blue), deployment (orange), application (green) and dataset (red) levels.

Equation 4 shows the approach taken to assess the completeness for geocoding 2014 voting locations. The match rate is represented as a percentage (i.e. 98% of submitted addresses returned a geographic coordinate result).

$$C_{Geocode} = \left(\frac{\sum Returned\ Addresses}{\sum Total\ Addresses}\right) \times 100 \qquad \textbf{(4)}$$



Data© OpenStreetMap contributors.

Figure 3.7: 2014 Ottawa municipal election voting locations.

3.2.2 Positional Accuracy

Positional accuracy of geographic information refers to the accuracy of geographic coordinate values (i.e. how close is an object to the real-world representation) (van Oort, 2006).

3.2.2.1 Road Networks

Goodchild & Hunter (1997) initially implemented a method to analyze digitized features in comparison to a "ground-truth" objective. Figure 3.8 illustrates the buffer analysis method designed by Goodchild & Hunter (1997) which is also used to quantify positional accuracy of OSM features as a percentage of overlap. The "coastline to be tested" in Figure 3.8 represents the OSM road network being evaluated. The "true coastline" represents the buffered benchmark dataset. The buffer width was justified by the width of the road network (5-10 m).

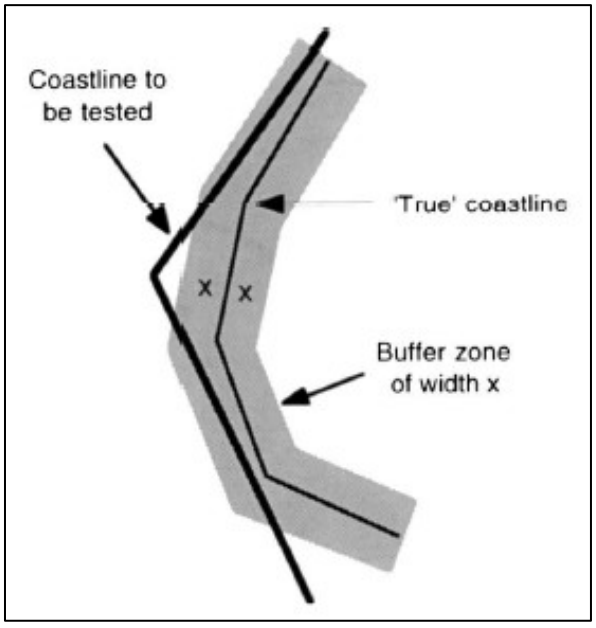


Figure 3.8: A buffer width of x around a "true coastline" that is being intersected by an evaluation coastline to be tested (Goodchild & Hunter, 1997).

For this research, OSM road networks in Ottawa-Gatineau were evaluated relative to DCMS using methodologies outlined in Goodchild & Hunter (1997), Haklay (2010) and Girres & Touya (2010). Equation 5 outlines the calculations used to determine overlap percentages between OSM and DCMS buffer widths.

$$P_{Road} = \left(\frac{\sum Length\ of\ line\ segment\ within\ Buffer}{\sum Total\ length\ of\ measured\ line}\right) \times 100 \qquad \textbf{(5)}$$

Figure 3.9 illustrates the GIS computations used to calculate the overlap percentage between OSM and DMTI road networks. For demonstration purposes, the example in Figure 3.9 is a small intersection found in Ottawa-Gatineau. Annex A – Appendix 4 outlines the QGIS Processing model workflow used to execute the GIS computations outlined in Figure 3.9. The output result of the model is a geospatial file (.shp, .GeoJSON, etc.) with corresponding road network lengths and positional accuracy assessments. This QGIS Processing model was originally created by Anita Graser out of the Center for Mobility Systems at the Austrian Institute of Technology in Vienna, Austria[6]. The model was altered to execute with QGIS version 2.18.

[6] https://anitagraser.com/2013/12/15/osm-quality-assessment-with-qgis-positional-accuracy/

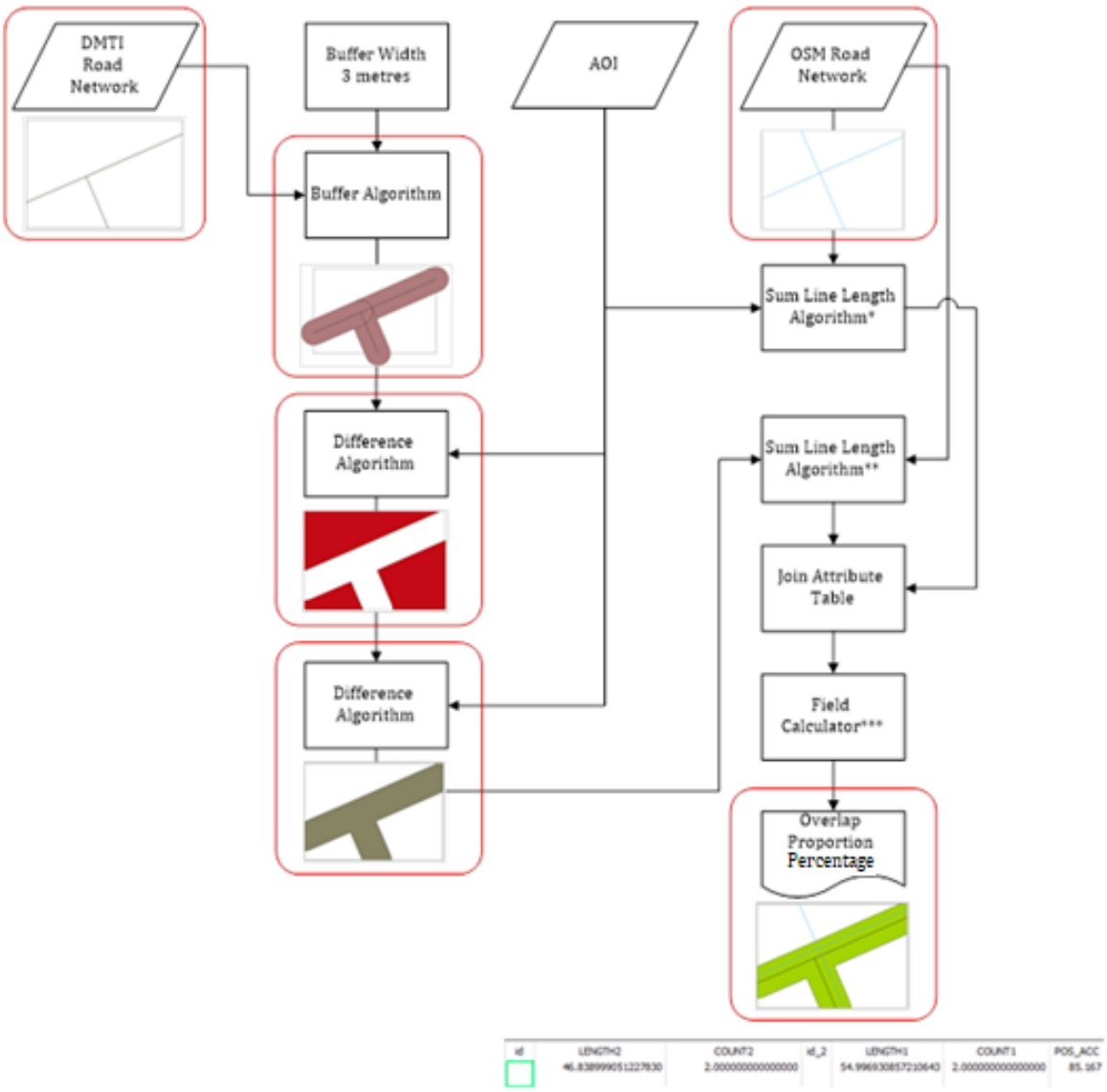


Figure 3.9: Flow diagram depicting the GIS computations used to calculate overlap percentage of OSM and DMTI road networks.[7]

[7]

**Sum line length of OSM road networks throughout entire AOI.*
***Sum line length of OSM road networks within 3 metre buffer width of DMTI road networks.*
****Execution of Equation 3 to calculate overlap proportion percentage.*

3.2.2.2 Geocoding

Using the "ground-truth" locations of the City of Ottawa voting locations, the geocoded results (geographic coordinates) were assessed for positional accuracy. The Haversine distance formulas (great circle distance) were used to find the distance between "ground-truth" geographic coordinates and returned coordinates from the deployed OSM Nominatim instances. The Haversine formula (Equation 6) was used instead of the Euclidean distance (straight line) because the Euclidean distance would be exaggerated over long distances from ground-truth, thus Haversine provides a more accurate depiction of length.

$$P_{Hav\left(\frac{d}{r}\right)} = Hav(\emptyset_1 - \emptyset_2) + \cos(\emptyset_1)\cos(\emptyset_2)\,Hav(\lambda_2 - \lambda_1) \qquad \textbf{(6)}$$

3.2.3 Thematic Accuracy

Thematic accuracy, also known as attribute accuracy, refers to the accuracy of all the attributes other than positional and thematic associated with a geospatial dataset (i.e. is the map feature what we believe it is?) (van Oort, 2006).

3.2.3.1 Road Networks

Using similar methods outlined by Girres & Touya (2010), OSM and reference datasets were assessed for attribute accuracy. This assessment accounts on whether an attribute fields are informed (complete) or uninformed (incomplete). For each of the datasets, road network types and names were assessed for completeness.

$$TA_{Roads} = \left(\frac{\sum Ftr\ Attr\ \textbf{Not Null}}{\sum Ftr}\right) \times 100 \qquad \textbf{(7)}$$

3.2.3.2 Buildings

Building footprint thematic accuracy will implement similar informed/uninformed feature completeness methods.

$$TA_{Bldg} = \left(\frac{\sum Ftr\ Attr\ \textbf{Not Null}}{\sum Ftr}\right) \times 100 \qquad \textbf{(8)}$$

3.2.4 Temporal Accuracy

Temporal accuracy refers to the accuracy of a geospatial dataset relative to historical changes in the real-world. Temporal accuracy of OSM data has been questioned because the disagreements among OSM community members about validity of any edit or feature addition can cause the OSM data to vary considerably between specific time periods. It is possible to inspect the OSM database by processing the generated OSM history file (.osh.pbf). This history file contains the current version of the OSM data of a region, plus all the history of changes to those data through time. For any map feature object (node[8], way[9], relation[10]) there can be zero or more previous versions in the history file. Deleted map objects are also included in this file format.

[8] https://wiki.openstreetmap.org/wiki/Node
[9] https://wiki.openstreetmap.org/wiki/Way
[10] https://wiki.openstreetmap.org/wiki/Relation

Using a framework developed by Oslandia[11], a company that specialized in open-source GIS architecture and software solutions, it is possible to inspect the historical evolution of the OSM database. Oslandia developed this framework in Python with *numpy*[12] and *pandas*[13] packages and the *Luigi*[14] Python library.

Oslandia outlines their OSM data analysis model into three main task categories; data parsing, metadata building and metadata analysis[15]. The data parsing involves processing the regular OSM file format (.osm.pbf) to obtain machine read-able .csv files. Some of the data parsing tasks include parsing the OSM entities (nodes, ways, relations), parsing OSM tags[16] (keys and values) and parsing OSM users directly from the contributor information. The metadata building phase entails generating the history of the OSM entities or "data behind the data".

The second group of tasks entail extracting the OSM element[17] metadata (creation dates, version numbers), extracting the OSM change set metadata (timestamp, number of modification) and extracting OSM user metadata (timestamps, number of opened change sets, number of modifications).

The last phase of the framework is the metadata analysis. By parsing and extracting the history and metadata elements behind the OSM entities, it is possible to obtain knowledge on how OSM users contribute to the OSM database and their

[11] http://oslandia.com/en/home-en/
[12] http://www.numpy.org/
[13] https://pandas.pydata.org/
[14] https://github.com/spotify/luigi
[15] http://oslandia.com/en/2017/06/19/openstreetmap-data-analysis-how-to-handle-it/
[16] https://wiki.openstreetmap.org/wiki/Tags
[17] https://wiki.openstreetmap.org/wiki/Elements

ability to do it properly. Therefore, by knowing how expert OSM users contribute to the OSM database, it is feasible to conclude that the OSM map feature is at its highest accuracy and quality without relying on an authoritative comparison dataset. The final phase of the framework concentrates on machine-learning-focused tasks to prepare the data, reduce dimensionality through Principal Component Analysis (PCA), and classifying users into groups through a k-means classification algorithm.

For this report, it was decided to take components that cover at least 70% of the cumulative variance (7 components). The PCA algorithm was run from the *sklearn* Python module and with a fit transformation to generate a new linear projection. Through unsupervised machine learning using inputs from the PCA dimensions, a k-means algorithm is implemented to cluster OSM users without any prior knowledge on OSM or geospatial experience. Prior to running the k-means clustering algorithm, the variables were normalized so a more accurate representation of user characteristics could be determined between the cluster groups.

While there are no defined set of rules to determine a suitable number of clusters for the k-means algorithm, the elbow and silhouette methods were considered. This method looks for a juncture indicating a drop in explained variance (i.e. an "elbow") in plots of principal components. However, "elbows" are not always easily distinguishable (Ketchen & Shook, 1996). It was determined that 4 clusters would be appropriate for this analysis and would allow for enough contributors in each of the cluster groups.

Temporal accuracy of OSM features can also be assessed by examining the OSM database evolution over a selected time period. Girres & Touya (2010) inspect the evolution of OSM map features over a three-month period by percentage of growth. This research used that approach to investigate the growth of the Ottawa-Gatineau OSM map objects (OSM entities, roads, buildings, etc.) through January 2016, January 2017 and June 2017 time periods.

4 RESULTS

This chapter includes a presentation of the outputs associated with the evaluation of OSM map features relative to authoritative datasets (road networks, buildings), quality of services that can be enabled with OSM data (geocoding) and the congregation of mapping characteristics unique to the Ottawa-Gatineau contributor base.

4.1 Quality Measures

4.1.1 Completeness

4.1.1.1 Road Networks

Across all road network types, it was found there was a surplus of OSM road network length relative to the DCMS. However, among certain micro-level road network types, results began to vary. From January 2016 to 2017 DCMS had a greater minor road network length than OSM (i.e. January 2016 minor OSM road network length is 91.5% of the DCMS road network dataset). A surplus of OSM road network length also exists for January and June 2017 among major road network types, excluding January 2016, when OSM length was 99% of DCMS. Breaking down length difference between minor road networks, DCMS is considered more complete. During January 2016, minor OSM road networks represented 91.5% of DCMS, while in January 2017, this proportion percentage increased to 93.13% and to 93.53% in June 2017.

Table 4.1: Length difference between OSM and DCMS road networks.

Date	OSM Length (metres)	DMTI Length (metres)	Length Difference (metres)	Percentage (%)
January 2016	11,615,720	9,747,577	1,868,143	-
January 2017	11,953,214	9,766,616	2,186,597	-
June 2017	12,308,824	9,766,616	2,542,208	-

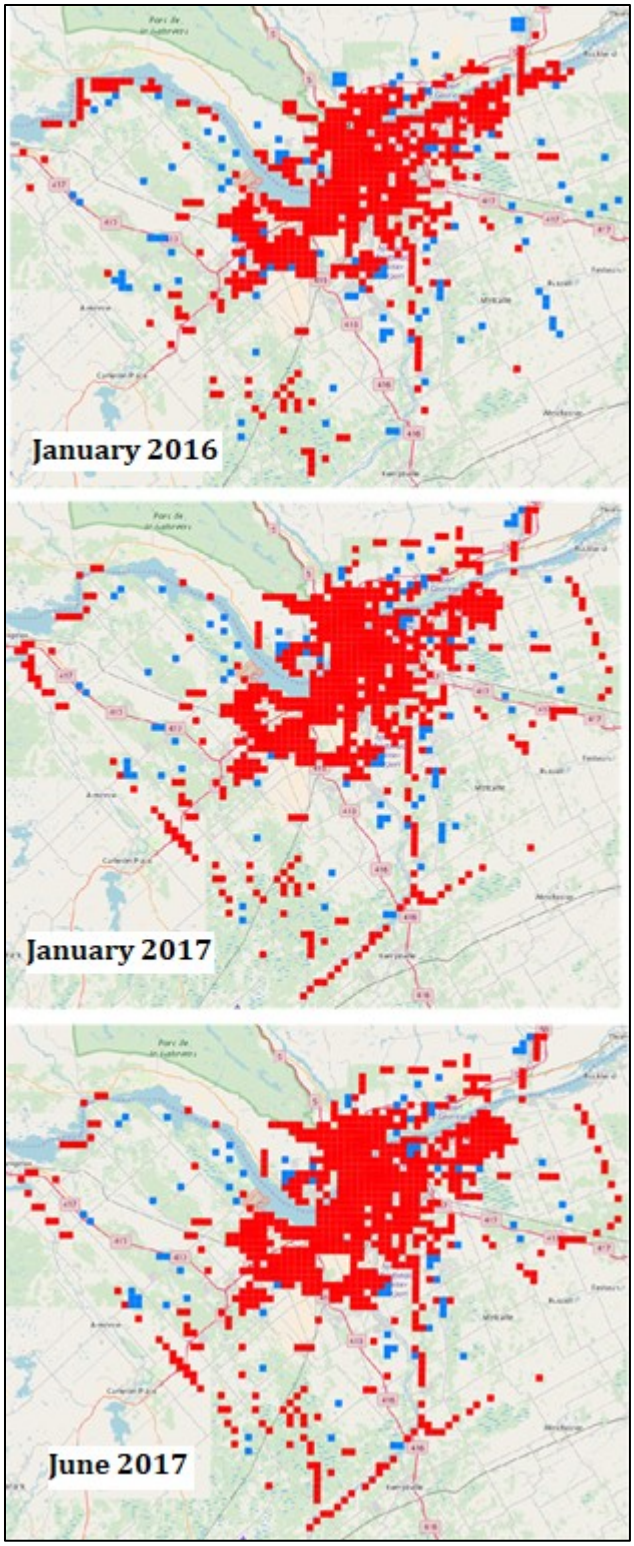
Table 4.2: Length difference between OSM and DCMS major road networks.

Date	OSM Length (metres)	DMTI Length (metres)	Length Difference (metres)	Percentage (%)
January 2016	2,394,485	2,396,547	-2,061	99%
January 2017	2,433,346	2,413,238	20,108	-
June 2017	2,466,084	2,413,238	52,846	-

Table 4.3: Length difference between OSM and DCMS minor road networks.

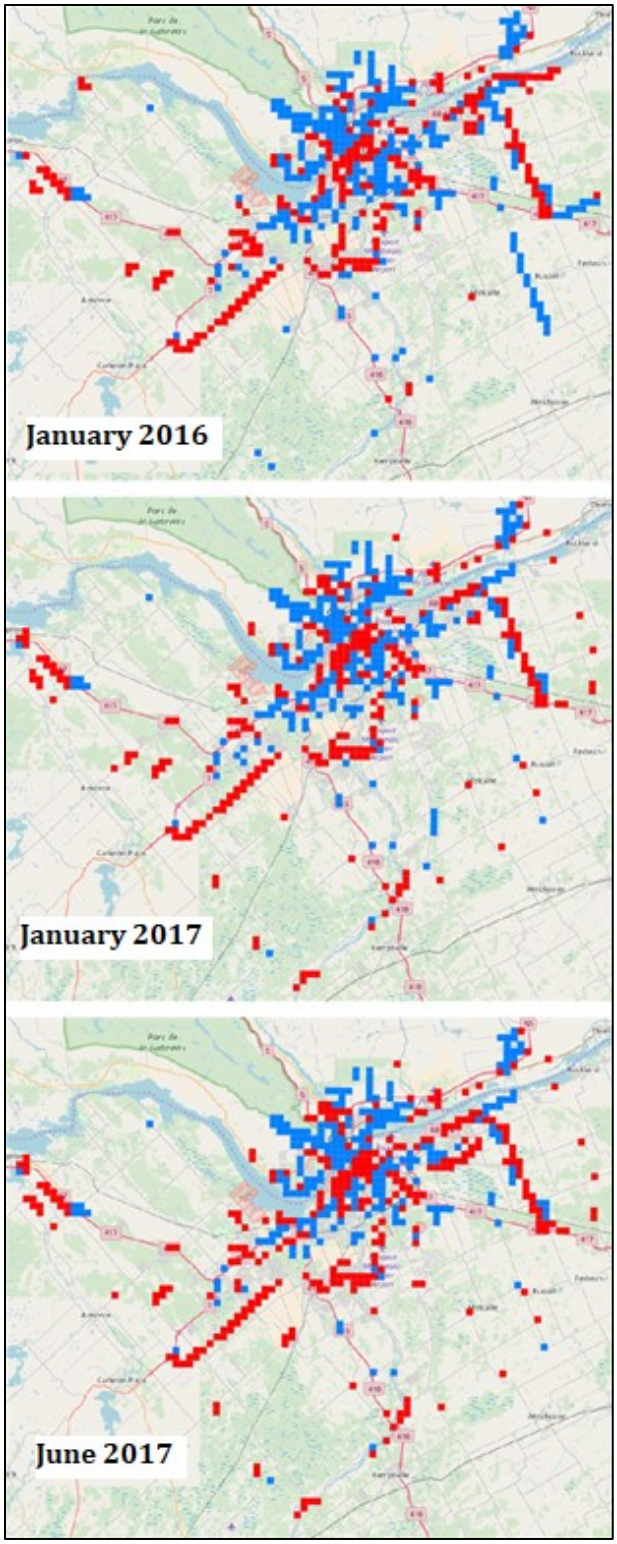
Date	OSM Length (metres)	DMTI Length (metres)	Length Difference (metres)	Percentage (%)
January 2016	6,317,370	6,905,388	-588,018	91.50%
January 2017	6,418,034	6,891,082	-473,048	93.13%
June 2017	6,445,008	6,891,082	-446,074	93.53%

Figures 4.1, 4.2 and 4.3 illustrate the spatial variation of road network length difference over 1-kilometre grid cell regions. Red cells represent more OSM road network length, blue represents less OSM road network length and blank regions represent neutral length difference or regions where no roads exist in the OSM or DCMS datasets.



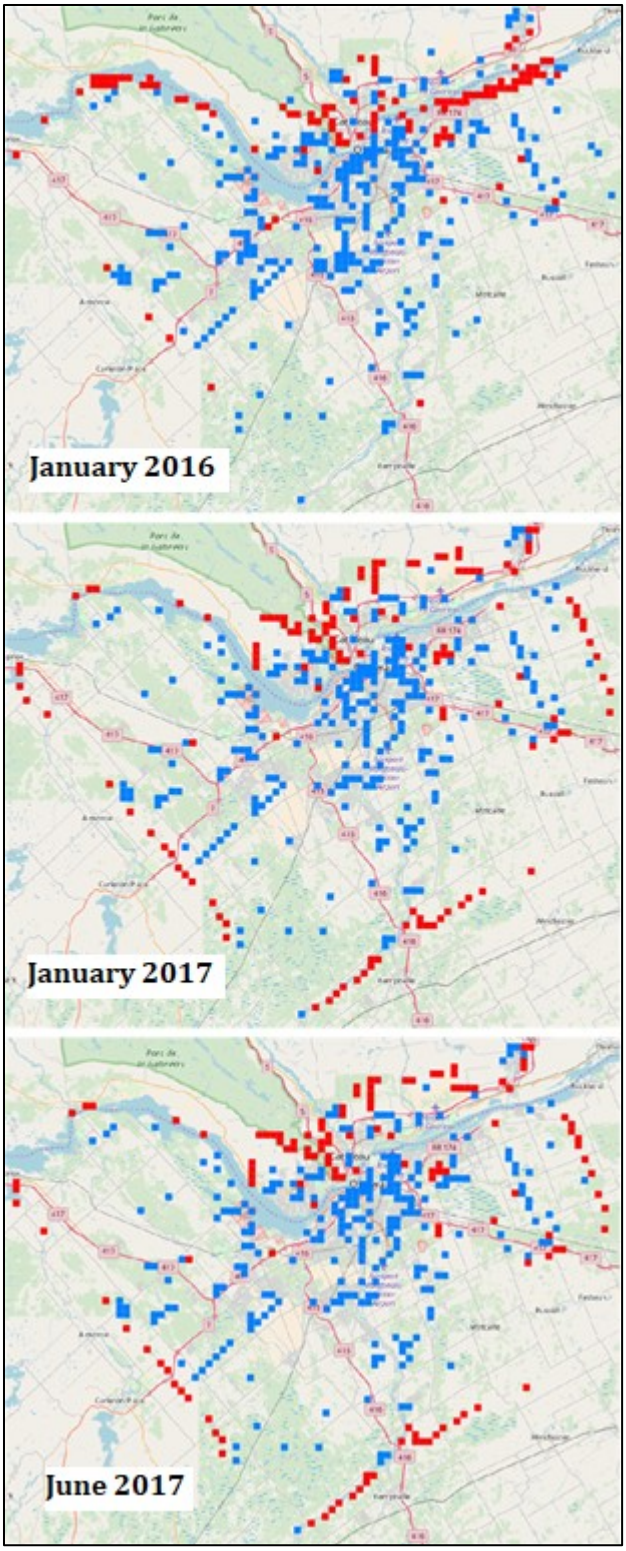
Blue: less OSM (< -1km difference). Red: more OSM (> 1km difference).

Figure 4.1: Road completeness difference between OSM and DMTI CanMap Suite over 1-kilometre grid cell regions from January 2016 (top), January 2017 (middle) and June 2017 (bottom).



Blue: less OSM (< -100m difference). Red: more OSM (> 100m difference).

Figure 4.2: Major road completeness difference between OSM and DMTI CanMap Suite over 1-kilometre grid cell regions from January 2016 (top), January 2017 (middle) and June 2017 (bottom).



Blue: less OSM (< -1km difference). Red: more OSM (> 1km difference).

Figure 4.3: Minor road completeness difference between OSM and DMTI CanMap Suite over 1-kilometre grid cell regions from January 2016 (top), January 2017 (middle) and June 2017 (bottom).

4.1.1.2 Buildings

4.1.1.2.1 Unit-Based Comparison

Over January 2016, January 2017 and June 2017, the Ottawa-Gatineau OSM dataset included a surplus of building outlines relative to DCMS. Due to this surplus, proportion percentages were not calculated. The building footprints found in the DCMS dataset mainly included large buildings (commercial, government buildings etc.) and excludes residential buildings.

Between January 2016 and January 2017, the OSM building footprints increased from 41,667 to 87,818, an increase of 46,151 buildings. From January 2017 to June 2017 there was an increase of 202,067 building outlines. Between the DCMS 2016 and 2017 datasets, there was an increase of 1,984 building footprints.

Table 4.4: Total number of OSM and DCMS building outlines.

Date	Total number of OSM building outlines	Total number of DCMS building outlines
January 2016	41,667	13,160 (Ottawa)
January 2017	87,818	15,144
June 2017	289,885	15,144

The total area of the Ottawa-Gatineau OSM building dataset increased from 21,876,292m^2 to 25,423,659m^2. This growth continues into June 2017, increasing to 67,239,748m^2. The DCMS dataset increased 3,088,181m^2 between January 2016 and January 2017, however the DCMS 2016 dataset excluded building footprints from Gatineau.

Table 4.5: Total area (m2) of OSM and DCMS building outlines.

Date	Total area of OSM building outlines (m^2)	Total area of DCMS building outlines (m^2)
January 2016	21,876,292	15,599,955 (Ottawa)
January 2017	25,423,659	18,688,136
June 2017	67,239,748	18,688,136

4.1.1.2.2 Object-Based Comparison

In January 2016, it was found that there were 6,170 DCMS building centroids within OSM building outlines. The number of DCMS building centroids increased to 7,017 in January 2017 and 12,222 in June 2017. Between January 2017 and June 2017 there is an increase of 34.37% of DCMS building centroids within OSM building outlines.

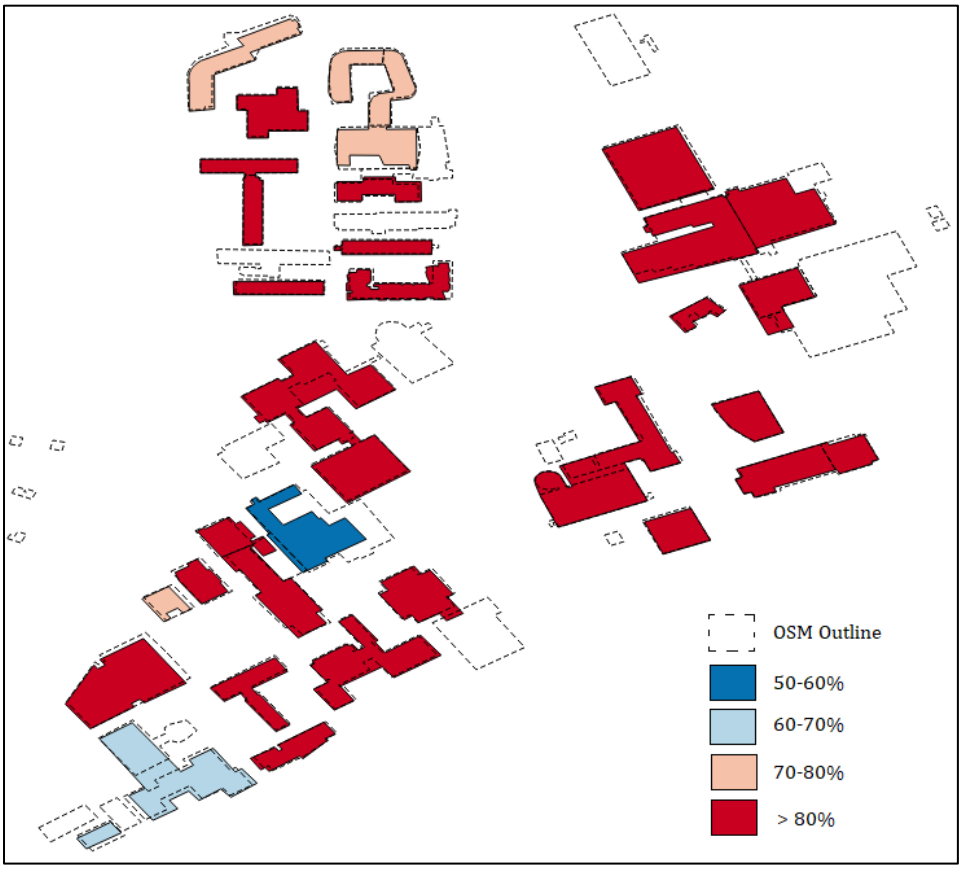
Table 4.6: Proportion between total number of DCMS building centroids in OSM building outline and total DCMS building centroids.

Date	Total number of DCMS centroids in OSM outline	Total number of DCMS building centroids	Percentage (%)
January 2016	6,170	13,160 (Ottawa)	46.88
January 2017	7,017	15,144	46.33
June 2017	12,222	15,144	80.70

Within the Ottawa CSD, 45.90% of the DCMS building footprints overlapped OSM footprints with at least a 50% or more area threshold. The overlap proportion increased 35.48% between January and June 2017 within the Ottawa-Gatineau regions. Figure 4.4 illustrates the degree of overlap between building outlines in OSM and DCMS across the Carleton University campus. Figure 4.5 outlines the mean overlap percentage across 1-kilometre grid cell regions (Red >80%, Blue 50-60% overlap).

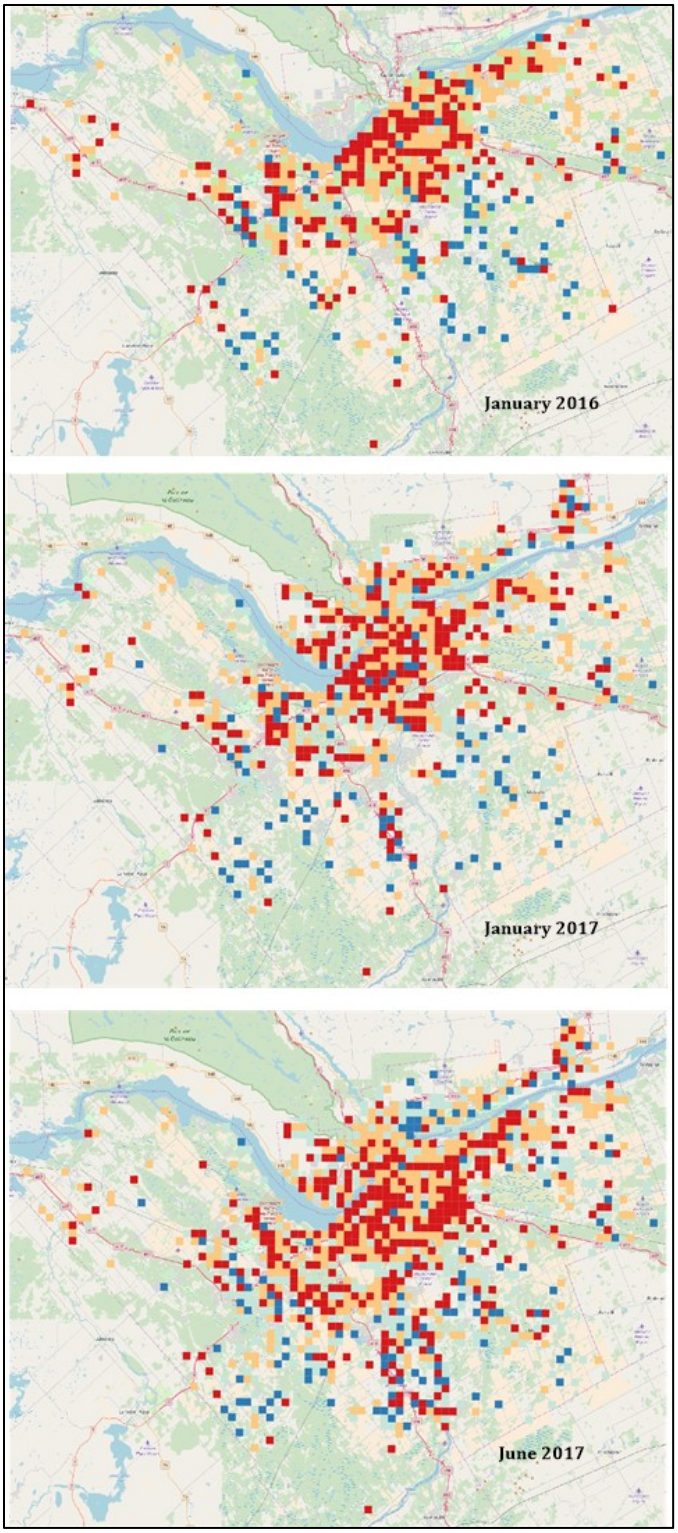
Table 4.7: Proportion between total number of DCMS buildings in OSM (> 50% overlap) and total number of DCMS buildings.

Date	Total number of DCMS Buildings in OSM (> 50% Overlap)	Total number of buildings in DCMS	Percentage (%)
January 2016	6,041	13,160 (Ottawa)	45.90
January 2017	6,902	15,144	45.57
June 2017	12,275	15,144	81.05



Red represents high degree of overlap (> 80%).
Blue represents low degree of overlap (50-60%).

Figure 4.4: Colour categorized illustration of degree of overlap between building outlines at Carleton University in OSM and DCMS datasets.



Red represents high degree of overlap (> 80%).
Blue represents low degree of overlap (50-60%).

Figure 4.5: Mean building overlap across 1-kilometre grid cell regions throughout January 2016, January 2017 and June 2017.

4.1.1.3 Geocoding

Geocoding completeness was evaluated by developing a Python script with the capability to batch geocode the 543 City of Ottawa voting locations for the 2014 municipal election. These 543 addresses were geocoded using the three Nominatim (OSM's geocoding service) instances installed. Throughout the three separate OSM snapshots (January 2016, January 2017, June 2017), each of the instances achieved over 96% completion ratios respectively (Table 4.8).

Table 4.8: Geocoding match rate (%) of Ottawa municipal polling locations.

Date	Total number of returned addresses	Total number of submitted addresses	Match Rate (%)
January 2016	543	562	96.61
January 2017	544	562	96.79
June 2017	544	562	96.79

4.1.2 Positional Accuracy

4.1.2.1 Roads Network

Using Equation 3 outlined in Section 3.2.2.1, it can be concluded that within a 10-metre buffer or greater of DCMS road networks, there is an overlap percentage of 76.90% or greater for OSM data captured in June 2017. This overlap percentage increased to 93.59% among major road networks and 92.60% among minor road networks at the 10-metre buffer interval. Table 4.9 outlines the percentage of overlap between the Ottawa-Gatineau OSM and DCMS road networks between 1 to 20-metres.

Line graph representations of these findings are depicted in Figure 4.6. Figure 4.7, 4.8 and 4.9 illustrate the spatial variation across 1-kilometre grid cell regions with regards to overlap percentage. Please refer to Annex B Appendix 1-5 for January 2016 and 2017 results.

Table 4.9: June 2016 – Overlap proportion between OSM and DCMS over 1-20 metres.

Road Type	**Width (metres)**	**Overlap Percentage (%)**
All	1	25.78
	5	69.61
	10	76.90
	15	78.27
	20	79.35
Major	1	30.48
	5	85.25
	10	93.59
	15	94.10
	20	94.54
Minor	1	32.29
	5	84.93
	10	92.60
	15	93.56
	20	93.90

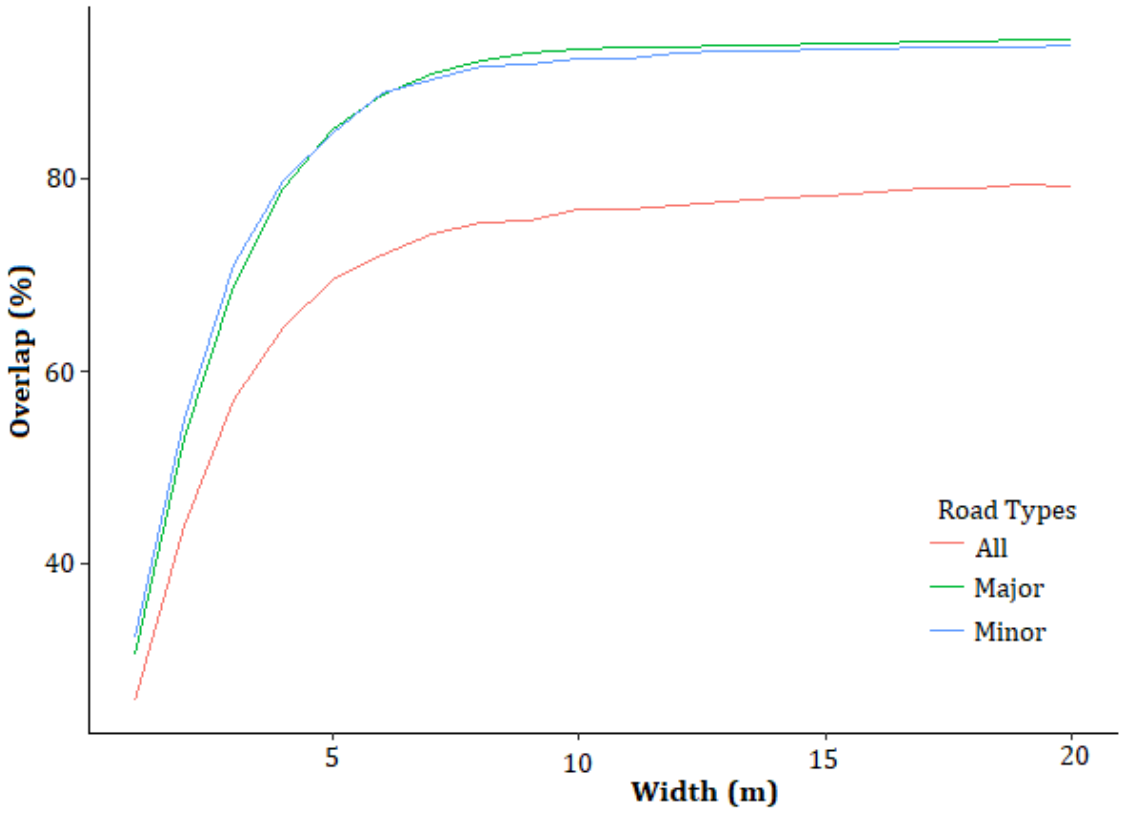
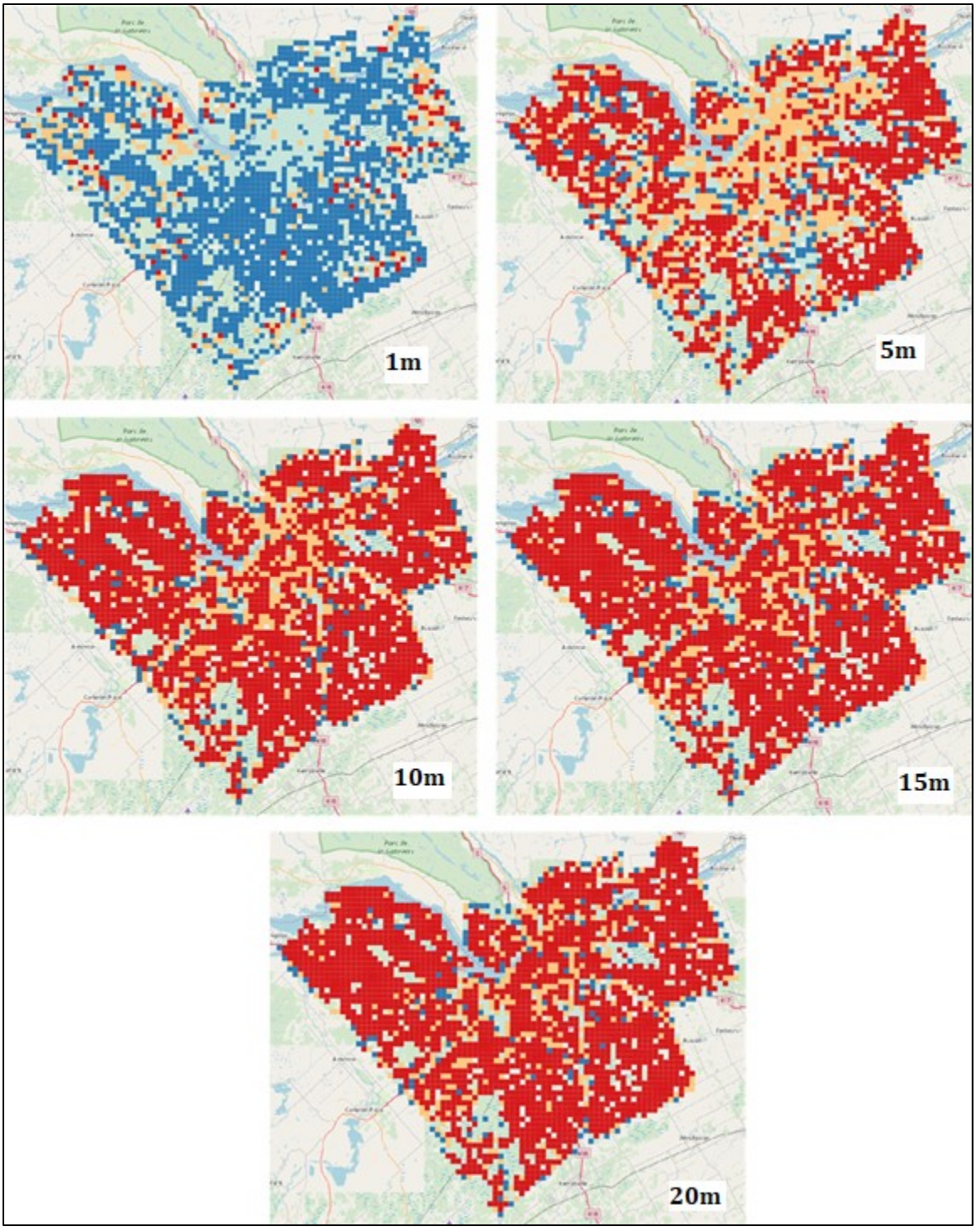
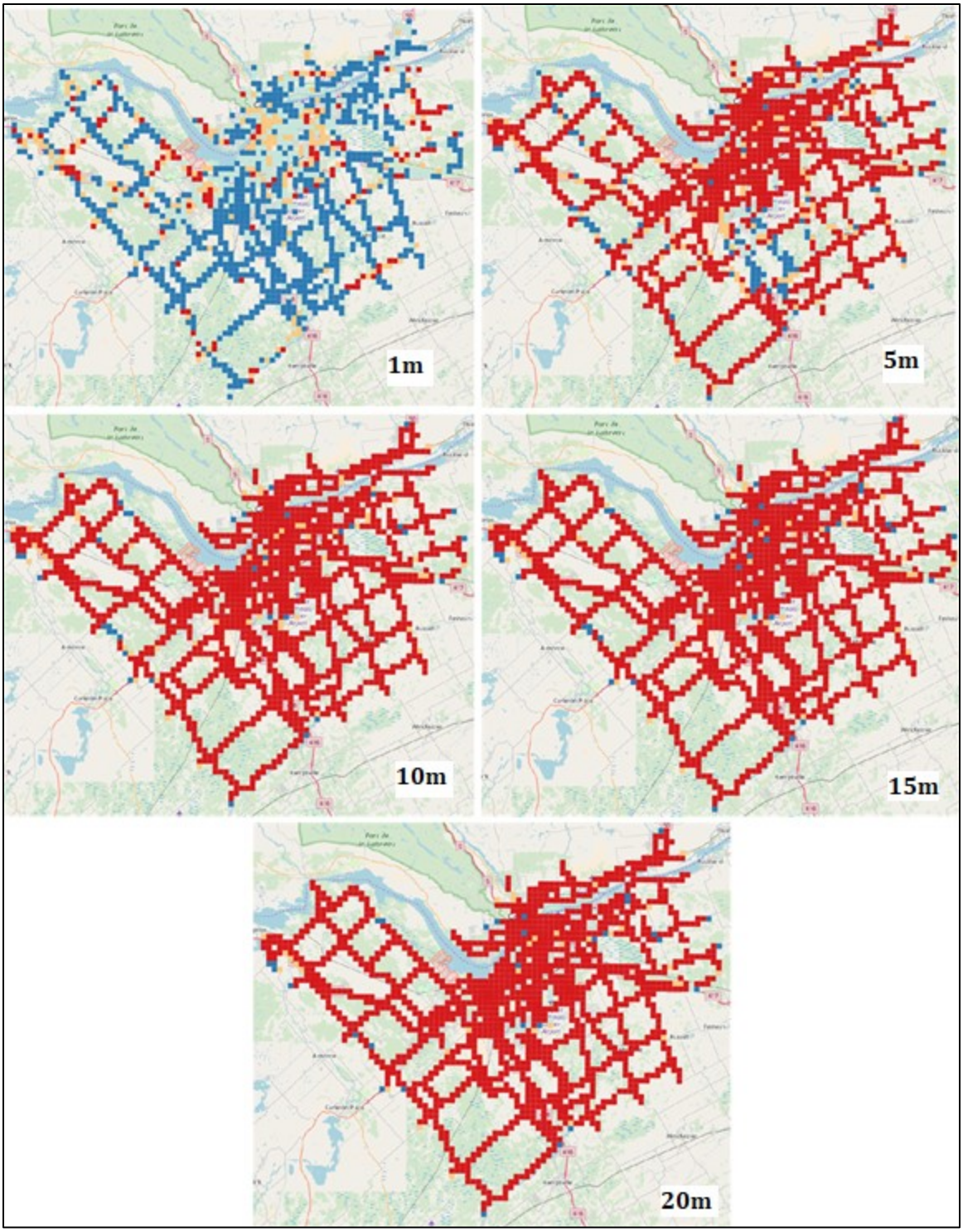


Figure 4.6: June 2017 – Overlap proportion (%) between road network types across 1 to 2 metre buffer widths.



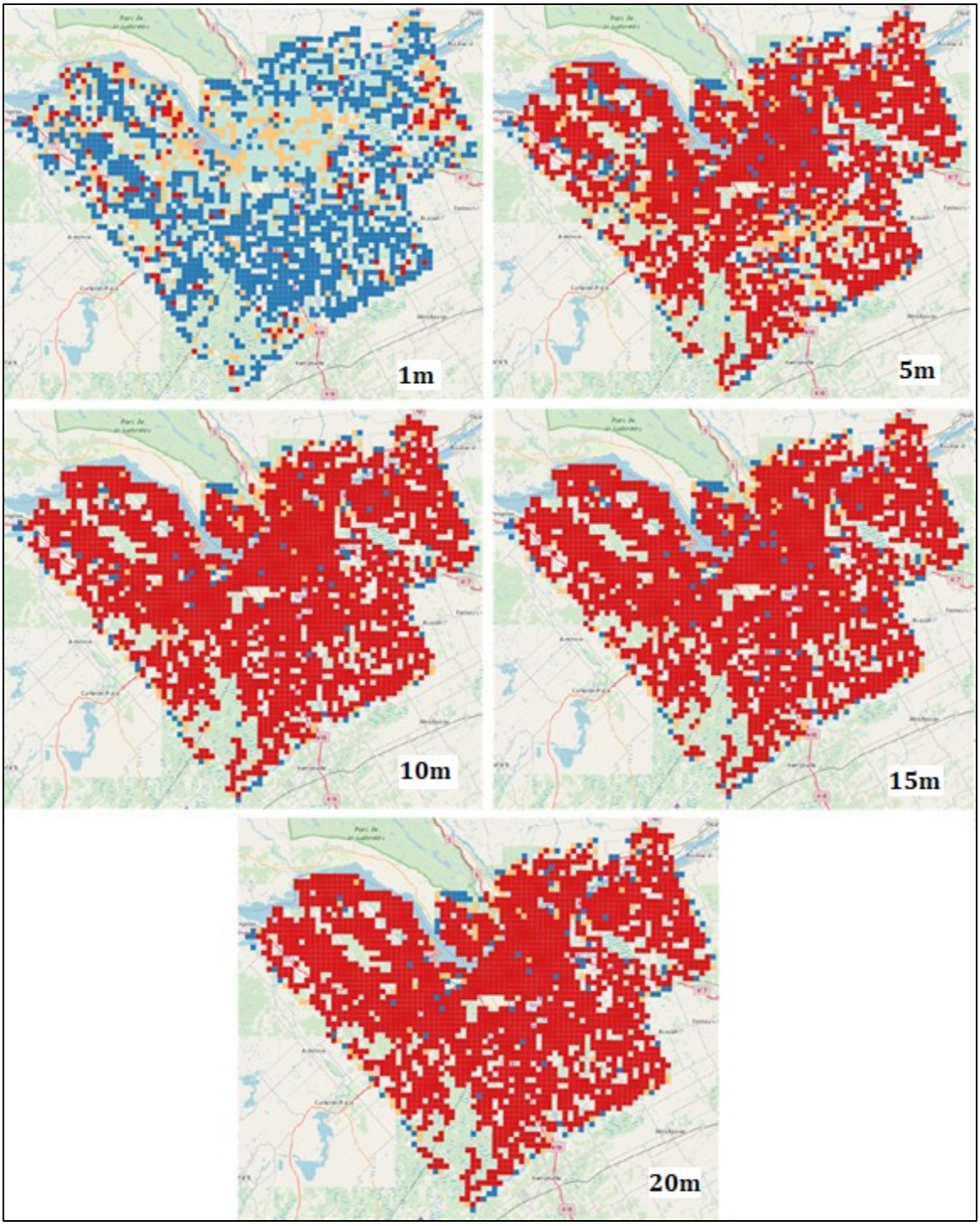
Dark Blue <25% overlap. Dark red >75% overlap.

Figure 4.7: June 2017 – OSM road network positional accuracy across 1m, 5m, 10m, 15m, and 20m buffer widths.



Dark Blue <25% overlap. Dark red >75% overlap.

Figure 4.8: June 2017 – Major OSM road network positional accuracy across 1m, 5m, 10m, 15m, and 20m buffer widths.



Dark Blue <25% overlap. Dark red >75% overlap.

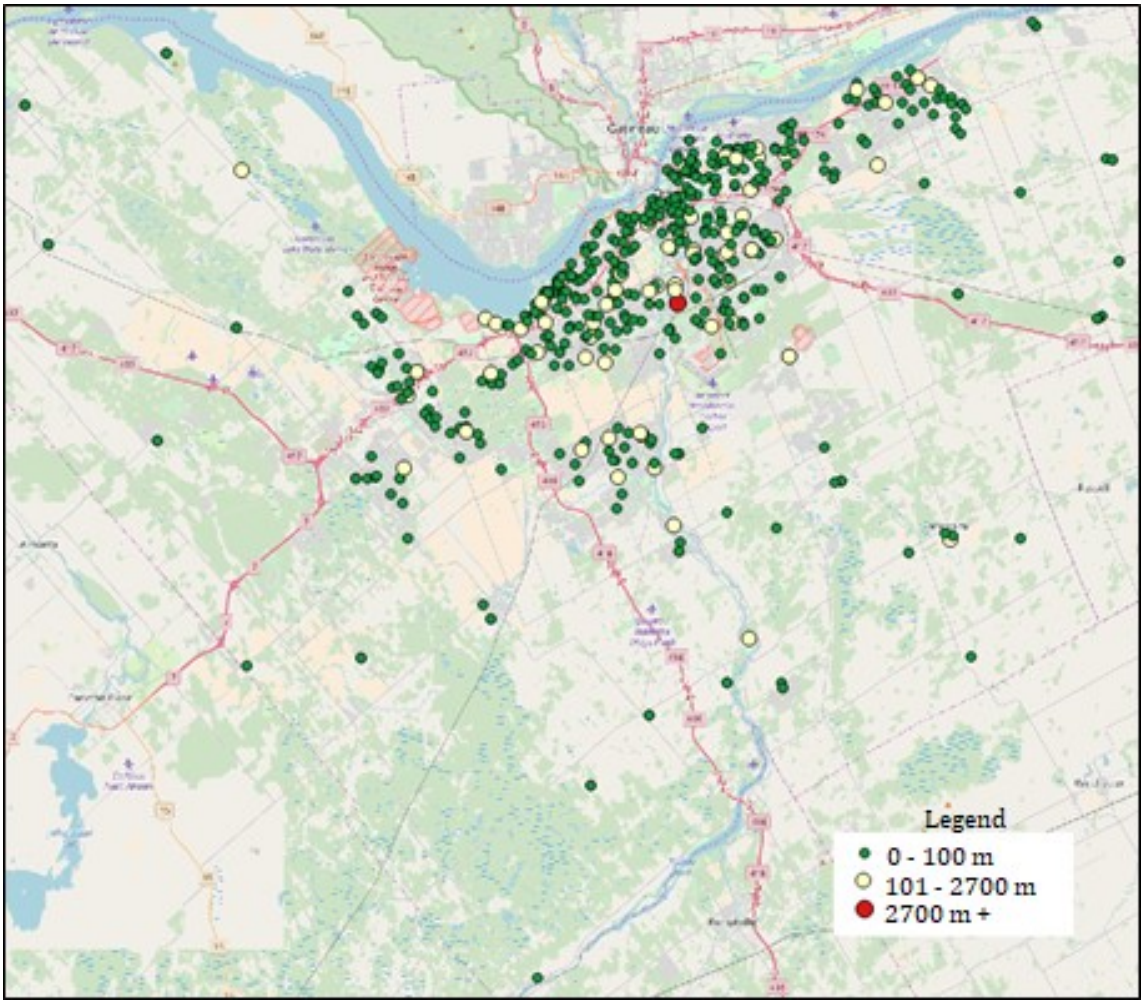
Figure 4.9: June 2017 – Minor OSM road network positional accuracy across 1m, 5m, 10m, 15m, and 20m buffer widths.

4.1.2.2 Geocoding

Table 4.10 summarizes differences between locations derived from geocoding compared to ground-truth locations. Minimum results ranged from 1.666 to 0.268 metres from the ground truth. June 2017 has the lowest difference geocoding result at 0.268 metres from the coordinates collected by the City of Ottawa. This narrative continues to hold true throughout all the summary statistics with geocoding results from June 2017 performing the lowest difference geocoding results from January 2016 and 2017 Nominatim instances. Figure 4.10 illustrates the geocoding results to ground truth as graduated symbology intervals. Thresholds indicated in Figure 4.10 were chosen to separate small deviations from ground truth that could be caused by the difference between geocoding at the building level versus street level interpolation geocoding between blocks (0-100 metres), and large deviations from ground truth that may result from geocoding by only street name (101 – 2,700 metres; 2,700 metres +).

Table 4.10: Summary statistics relative to ground truth coordinates.

Summary Statistics	January 2016	January 2017	June 2017
Min.	1.666	0.355	0.268
1st Qu.	28.86	27.55	13.16
Median	62.76	61.05	30.9
Std.Dev.	2.505	3.202	0.267
Mean	757.1	960.2	77.33
3rd Qu.	130.7	181.9	69.47
Max.	20,390	27,640	3,982.



Green: close to ground truth. Red: farther from ground truth.

Figure 4.10: Geocoding results for June 2017.

4.1.3 Thematic Accuracy

4.1.3.1 Road Networks

Road network attribute accuracy maintained a comparable completion ratio across January 2016 to June 2017, ranging from 59.10% to 62.30%. OSM Road networks in January 2016 included only 59.10% of the segments having names associated with them. This proportion increased to 62.30% in January 2017 then decreased to 60.70% in June 2017. DCMS maintained similar attribute completeness among road network names through 2016 and 2017 datasets. Tables 4.11 and 4.12

summarize the findings of attribute completeness associated with road network names.

Table 4.11: Thematic accuracy proportion between completed OSM "name" road segment tags and total number of OSM road segments.

Date	Number of road segments with "name" tags	Total OSM road segments	Percentage (%)
January 2016	27,313	46,212	59.10
January 2017	37,919	60,586	62.30
June 2017	39,476	65,035	60.70

Table 4.12: Thematic accuracy proportion between completed DCMS "name" road segment tags and total number of DCMS road segments.

Date	Number of road segments with "name" tags	Total DCMS road segments	Percentage (%)
DMTI CanMap Suite 2016 (Ottawa only)	45,488	53,472	85.07
DMTI CanMap Suite 2017	45,706	53,731	85.07

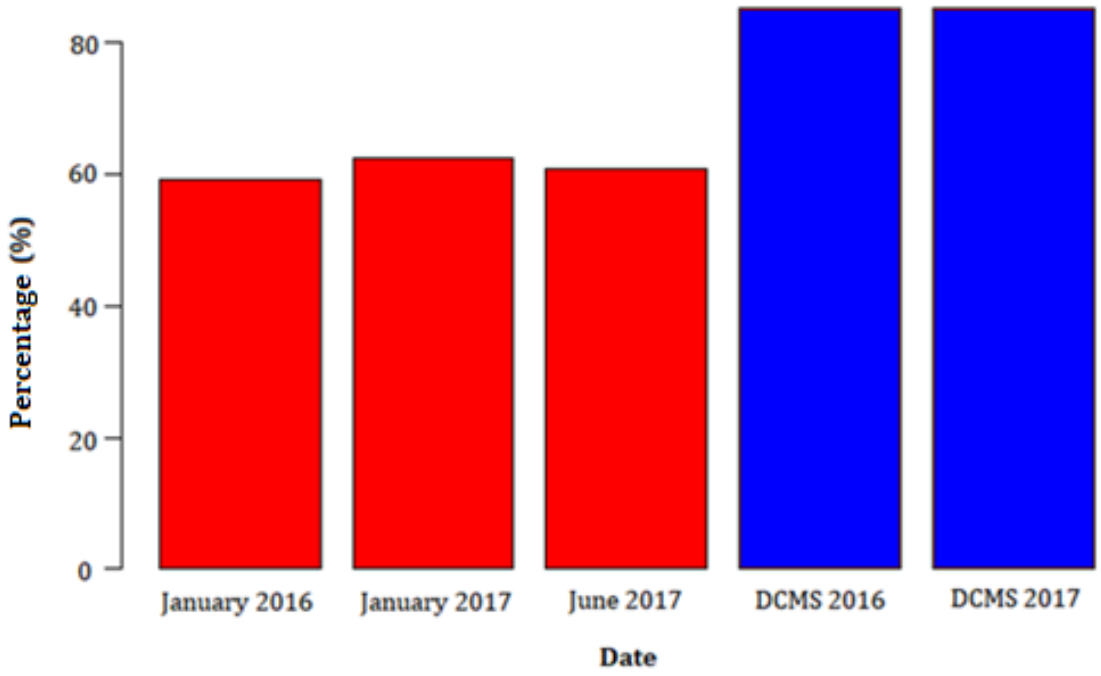


Figure 4.11: Bar chart of road network attribute name completeness represented as a percentage (%). Red: OSM. Blue: DCMS.

4.1.3.2 Buildings

January 2016, January 2017 and June 2017 OSM snapshots provide insight into the steady increase of OSM building completeness for the Ottawa-Gatineau region. In response, this also increased the attribute completeness among building tags (excluding *building = yes*). From January 2016 to June 2017 there is a 28.57% overall increase in OSM buildings across the Ottawa-Gatineau region. Building name attribute completion decreased from January 2016 to June 2017. Table 4.13 and 4.14 summarize the findings of building name and type attribute completeness.

Table 4.13: Thematic accuracy proportion between completed OSM "type" building tags and total number of OSM buildings.

Date	Number of buildings with "type" tags (not including building = yes)	Total OSM buildings	Percentage (%)
January 2016	23,334	41,557	56.15%
January 2017	53,386	87,818	60.80%
June 2017	245,600	289,889	84.72%

Table 4.14: Thematic accuracy proportion between completed OSM "name" building tags and total number of OSM buildings.

Date	Number of buildings with "name" tags	Total OSM buildings	Percentage (%)
January 2016	1,986	41,557	4.78%
January 2017	2,345	87,818	2.67%
June 2017	3,908	289,889	1.35%

DCMS building type completeness (excluding *"TYPE" = OTHER*) increased to 17.87% from 12.25% between 2016 and 2017 datasets. There is a minor increase among DCMS building name attributes from 5.63% to 5.92%. However simultaneously, the overall DCMS building dataset increased from 13,160 building outlines to 15,144, a difference of 1,984 buildings. Table 4.15 and 4.16 outline the findings for building attribute completeness for DCMS 2016 and 2017.

Table 4.15: Thematic accuracy proportion between completed DCMS "type" building tags and total number of DCMS buildings.

Date	Number of buildings with "type" tags (not type = other)	Total DCMS buildings	Percentage (%)
DMTI CanMap Suite 2016 (Ottawa)	1,612	13,160	12.25
DMTI CanMap Suite 2017	2,706	15,144	17.87

Table 4.16: Thematic accuracy proportion between completed DCMS "name" building tags and total number of DCMS buildings.

Date	Number of buildings with "name" tags	Total DCMS buildings	Percentage (%)
DMTI CanMap Suite 2016 (Ottawa)	741	13,160	5.63%
DMTI CanMap Suite 2017	897	15,144	5.92%

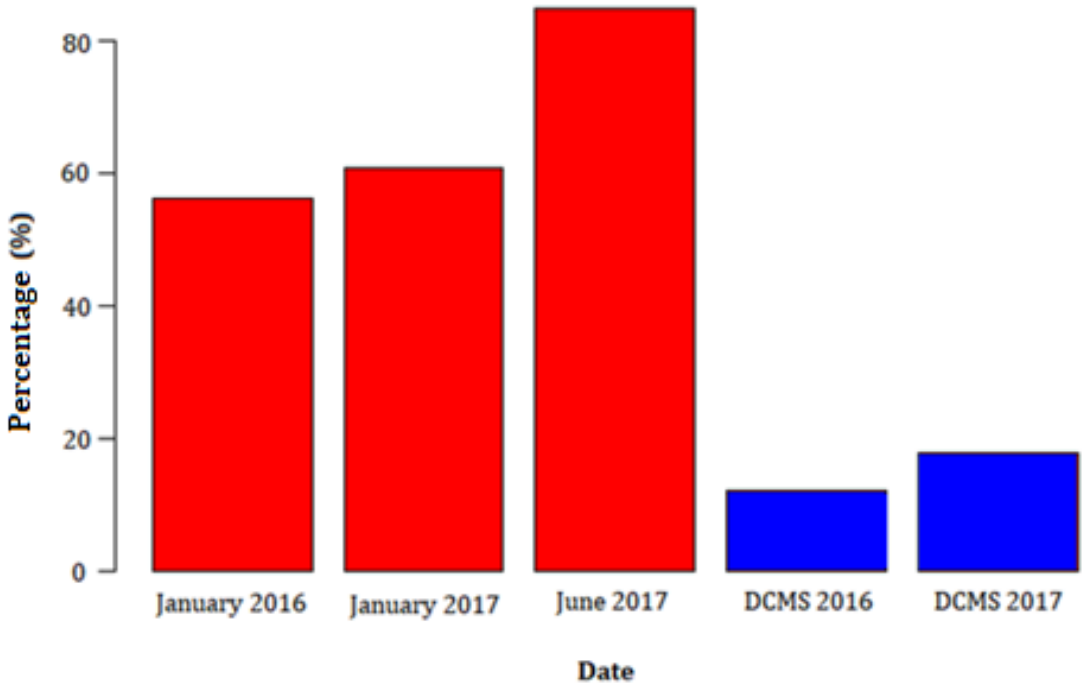


Figure 4.12: Building attribute name completeness represented as a percentage (%) of total polygons. Red: OSM. Blue: DCMS.

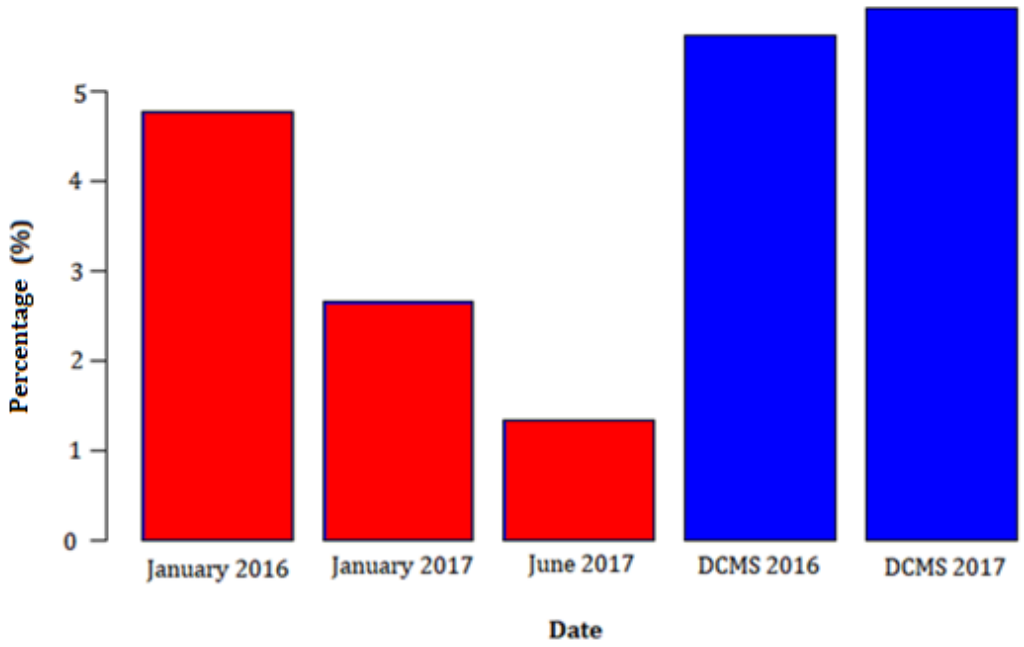


Figure 4.13: Road network attribute type completeness represented as a percentage (%) of total road segments. Red: OSM. Blue: DCMS.

4.1.4 Temporal Accuracy

4.1.4.1 Temporal Evolution

Figures 4.14, 4.15, 4.16 and 4.17 outline the temporal evolution of the Ottawa-Gatineau OSM database entities. Figure 4.14 illustrates the temporal evolution of the three OSM entity types (nodes, relations, ways) from 31 December 2007 to 30 September 2017. Figure 4.15 shows the evolution of users and changesets through this same period. In Figures 4. 16 and 4.17, the total number of users and nodes per kilometre squared where calculated using criteria from the 2016 census (Government of Canada, 2017a, 2017b).

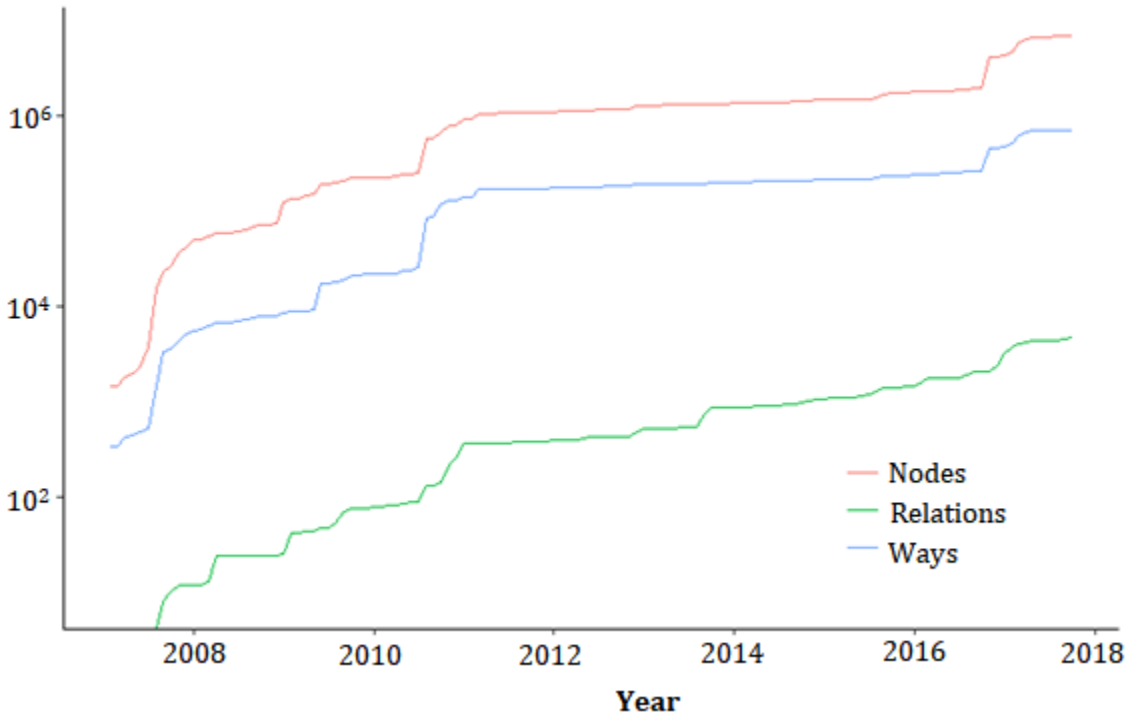


Figure 4.14: Log-scaled number of nodes, relations and ways between January 2007 to October 2017.

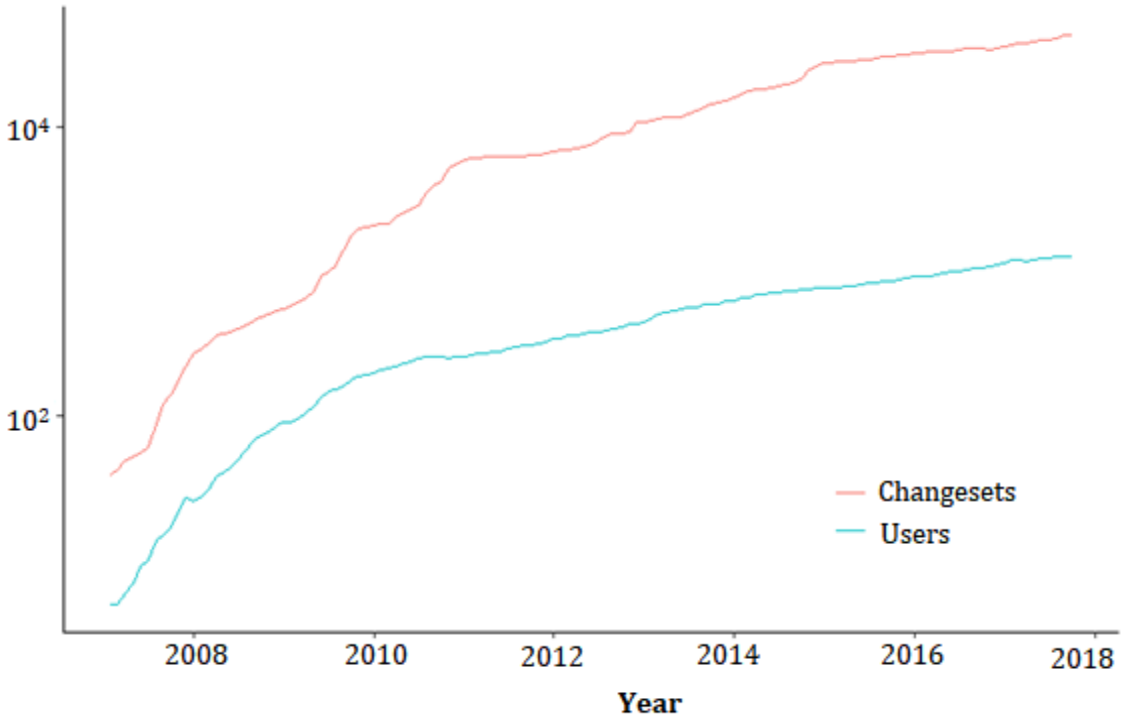


Figure 4.15: Log-scaled number of changesets and users between January 2007 to October 2017.

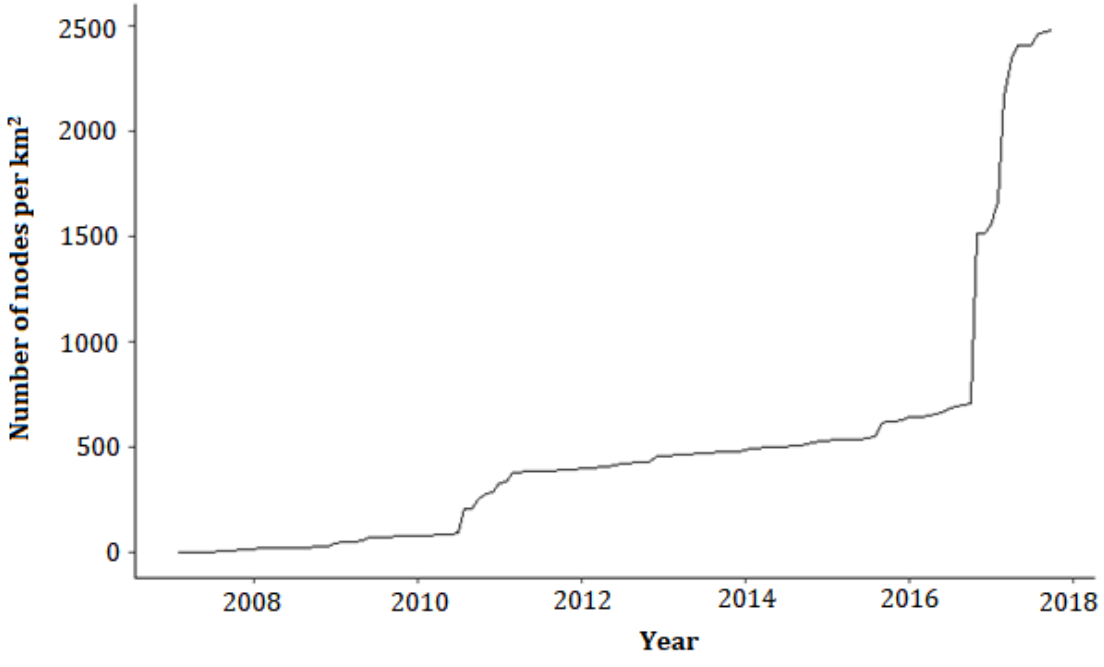


Figure 4.16: Number of nodes per km² between January 2007 to October 2017.

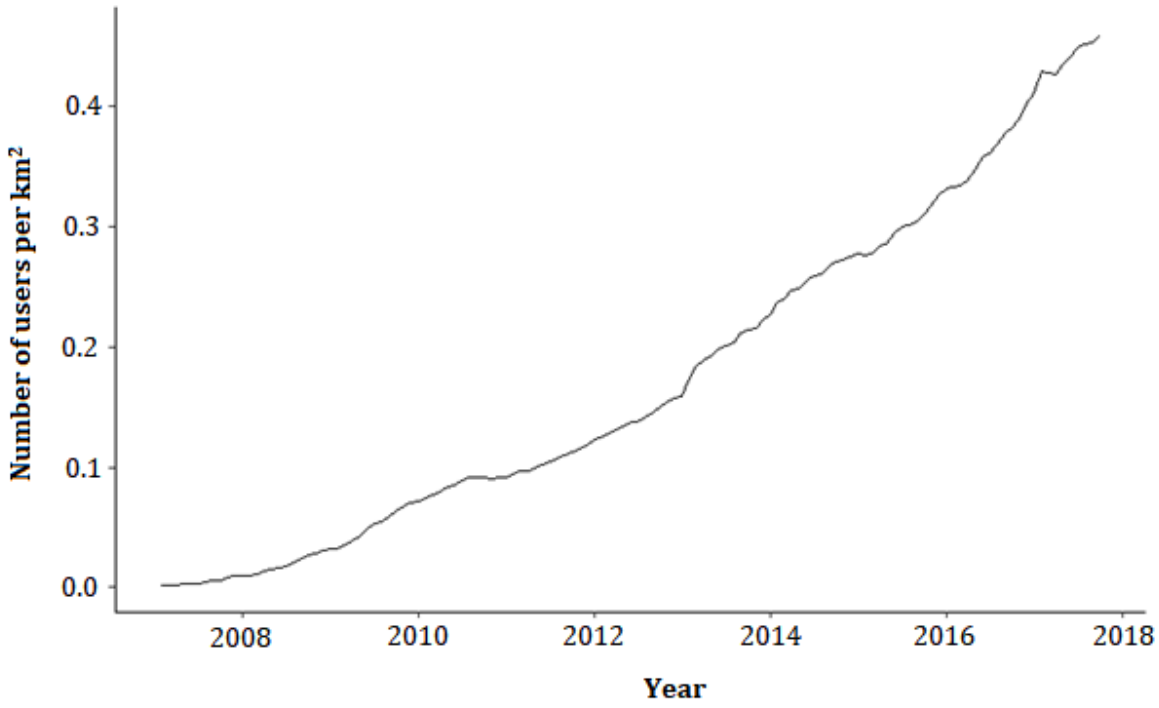


Figure 4.17: Number of users per km² between January 2007 to October 2017.

4.1.4.2 OSM Tag Structure: Object Classification

In the entire Ottawa-Gatineau OSM database, the *addr:housenumber* and *addr:street* tag keys are used the most frequently on map features (Figure 4.18). The overall frequency of these tag keys was over 1,600,000 in the Ottawa-Gatineau OSM dataset, regardless of their OSM map element type (node, relation, way). Out of the first versioned map elements, there are over 681,000 nodes tagged with the key *addr:street* and *addr:housenumber* (Figure 4.19). Out of the 906,000 nodes equal to version 1, this means that the tag *addr:street* appeared in 75% of these entities (Figure 4.20).

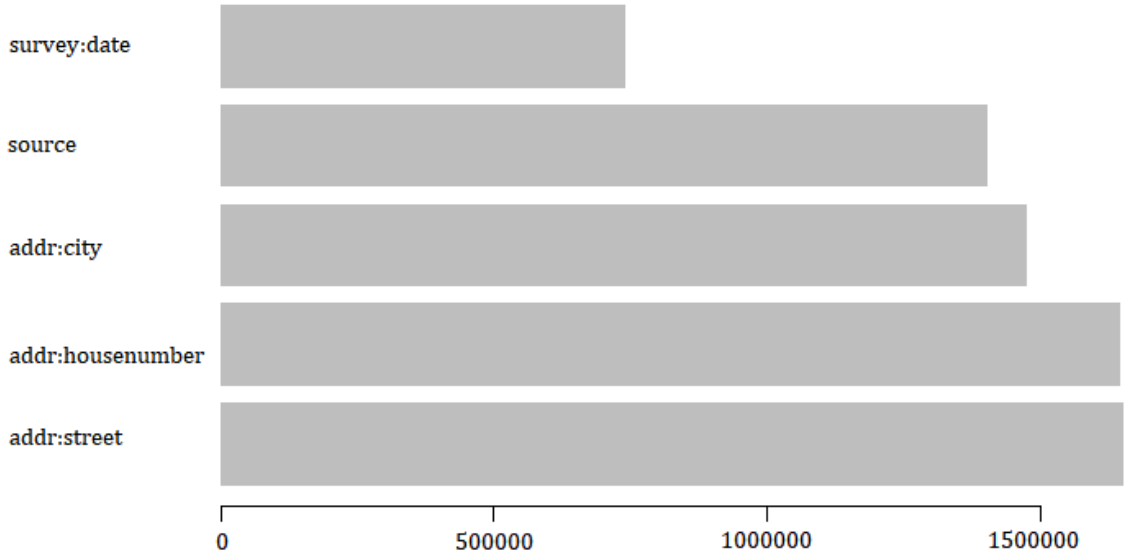


Figure 4.18: Most frequent tag keys used in the Ottawa-Gatineau OSM dataset (with respect to OSM types; nodes, relations, ways).

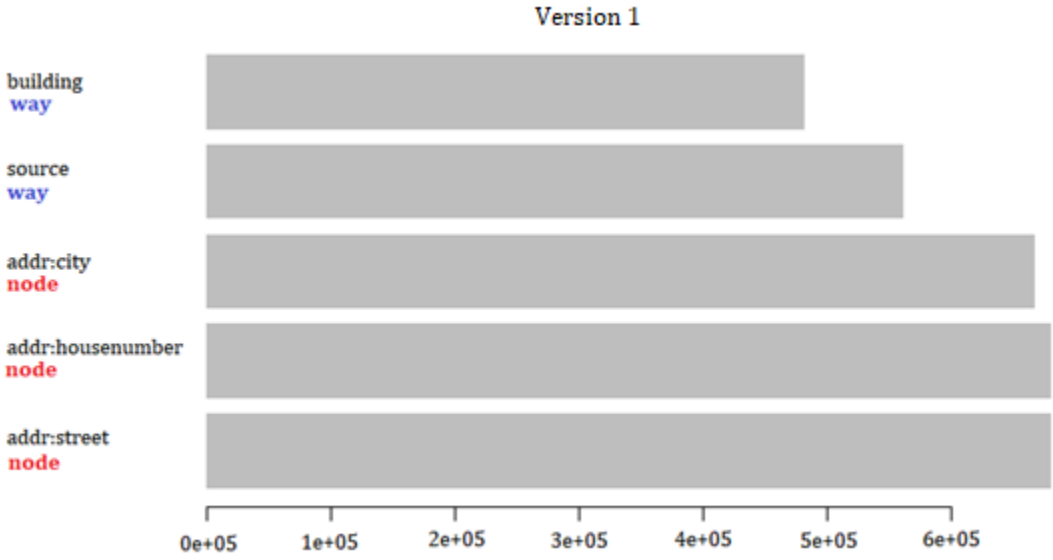


Figure 4.19: Most frequent tag keys used in the Ottawa-Gatineau OSM dataset with respect to OSM map elements at Version 1.

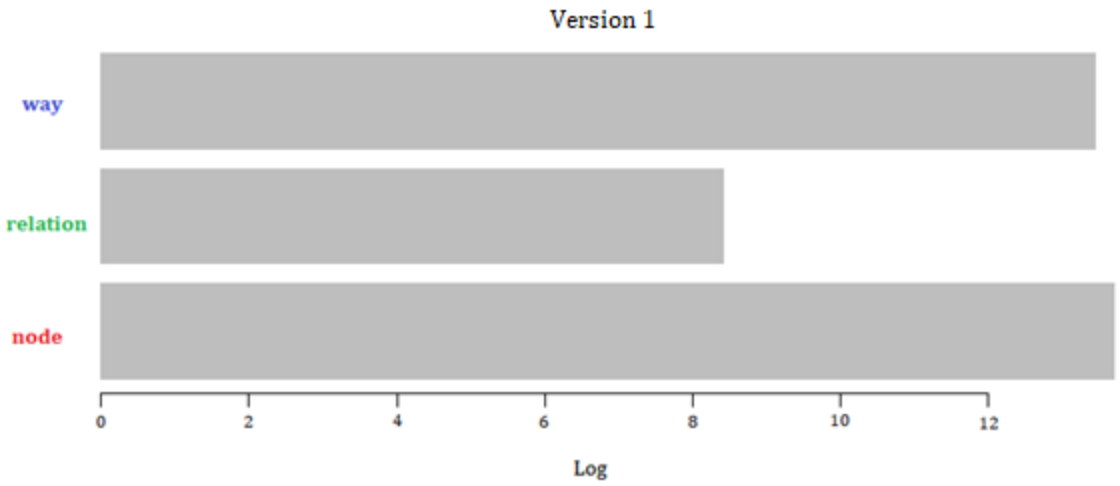


Figure 4.20: Total number of first version Ottawa-Gatineau map elements.

Figure 4.21 shows how OSM users are contributing to the database with respect to the tag frequency among map elements. *Source* tags are intensively used in the first version of map objects, but coverage decreases when the objects are updated. This trend is similar with respect to *building* tags. The opposite narrative is expressed regarding *name* tags where it is common to add the name of a map object after a few updates. This is evident with the addition of *name* (relation) tags added to map objects between versions 1 and 5 with an increase of *name* (relation) tags used added 18% to 73% of the time.

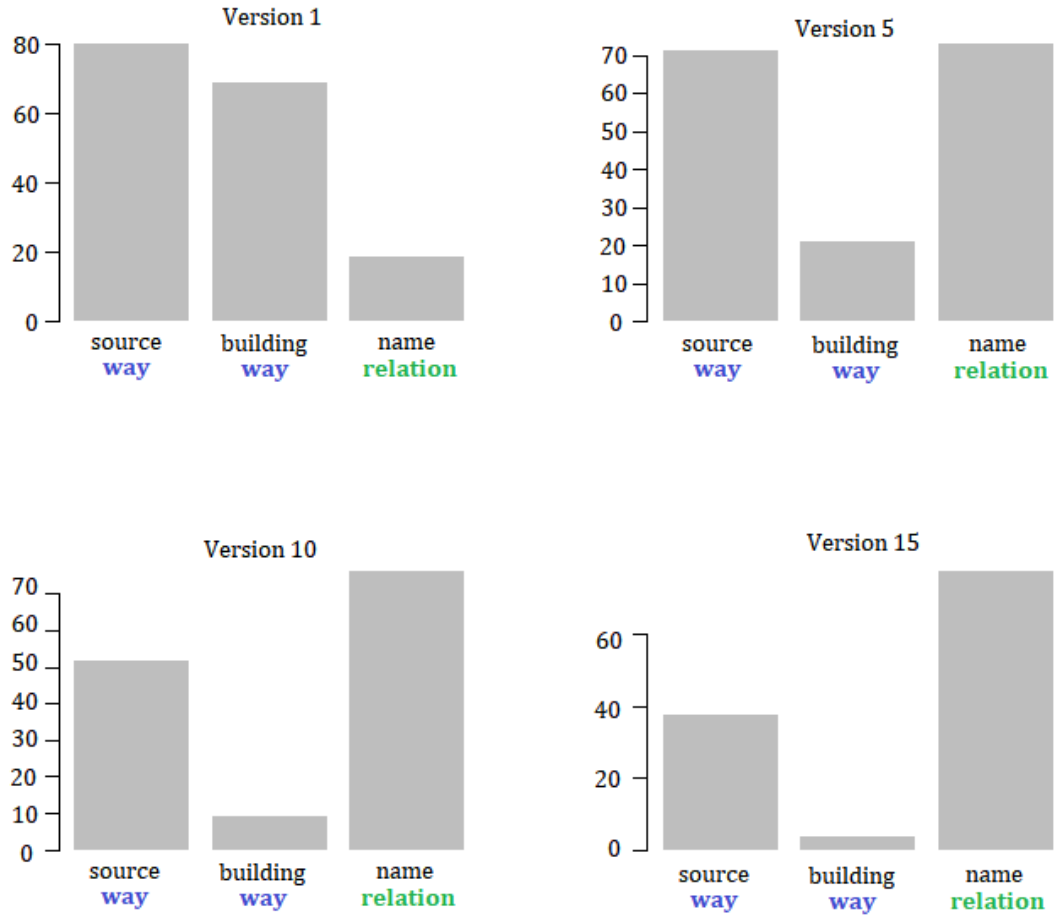


Figure 4.21: Tag frequency across map objects among all element types.

Figure 4.22 reveals most highway elements at version 1 either being nodes or ways. There are 17,627 nodes and 104,051 ways found among highway map elements. The total proportions among all map entity types (nodes, relations, ways) with respect to highway elements are shown in Figure 4.23, 4.24 and 4.25. Figure 4.23 illustrates that when users contributed a new node to the database that was a highway feature, the chosen value was *bus_stop* over a third of the time. Node elements tagged *traffic_signals* or *motorway_junction* tend to reach higher versions. Figure 4.24 shows highway way map elements and proportion of tags used across

version numbers. Among newly created highway ways, nearly half of them are tagged as residential. This proportion among highway ways tagged as residential remains relatively high across version numbers. Figure 4.25 shows highway relation map elements and proportion of tags used across version numbers. Across newly created highway relations, there are only 3 map relations and 2 have the tag value of service. There also appears to be no indication of a highway related relation with a high version number.

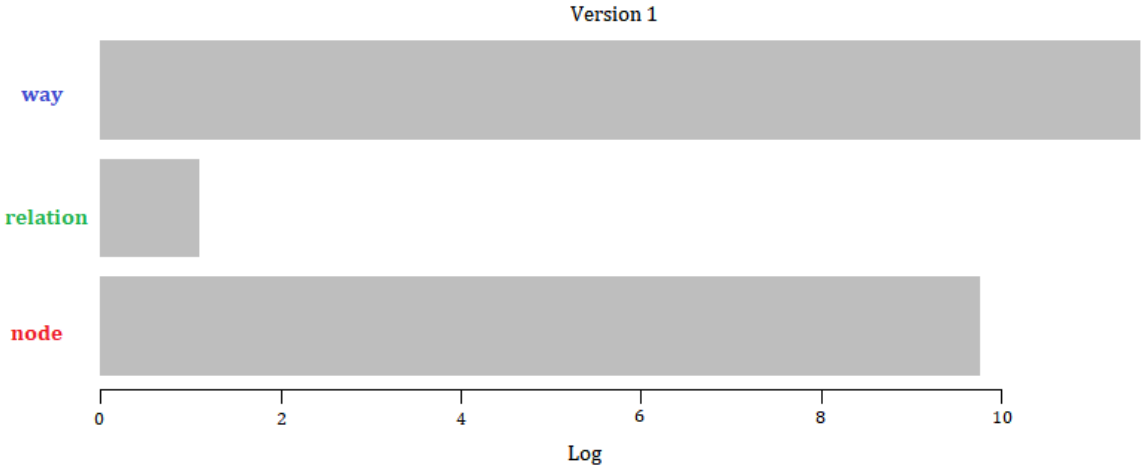


Figure 4.22: Number of highway elements at version 1.

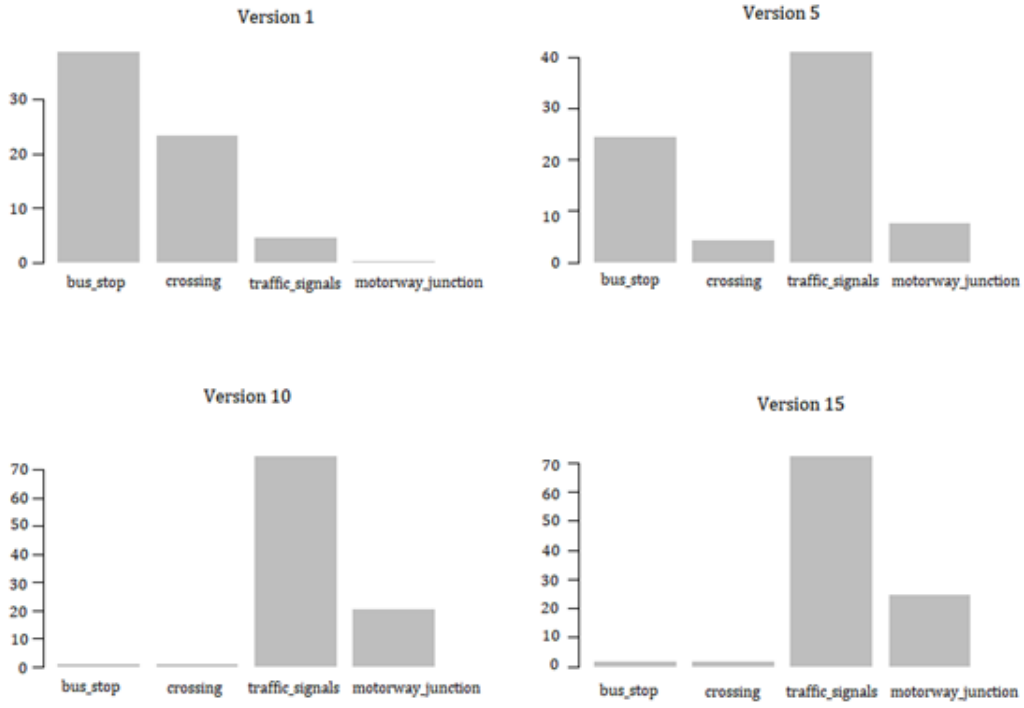


Figure 4.23: Proportion of highway related element node tag values added to map objects across versions 1, 5, 10 and 15.

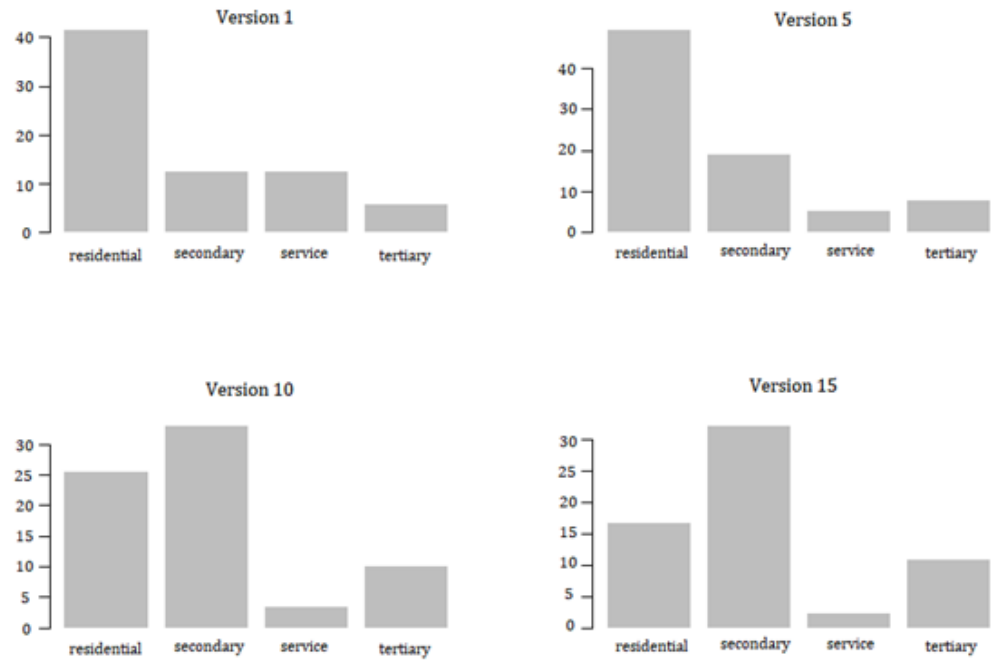


Figure 4.24: Proportion of highway related element way tag values added to map objects across versions 1, 5, 10 and 15.

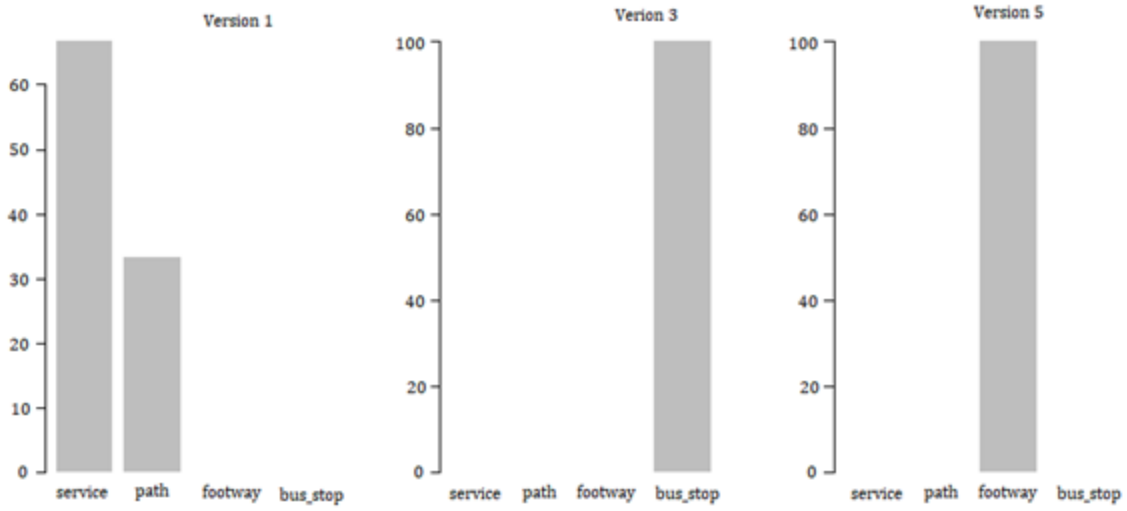


Figure 4.25: Proportion of highway related element relation tag values added to map objects across versions 1, 3 and 5.

Figure 4.26 reveals most building elements at version 1 either being nodes or ways. There are 2,060 nodes and 481,690 ways found among building related elements. The total proportions among all map entity types (node, relation, way) with respect to building elements are shown in Figure 4.26. Figure 4.27, 4.28 and 4.29 indicate when OSM users contribute a new node or way to the Ottawa-Gatineau OSM dataset as building related, the chosen value is *yes* 92% of the time for nodes, 51% of the time for ways and 68% of the time for relations. However, with respect to building related elements, there are only 102 first versioned relations.

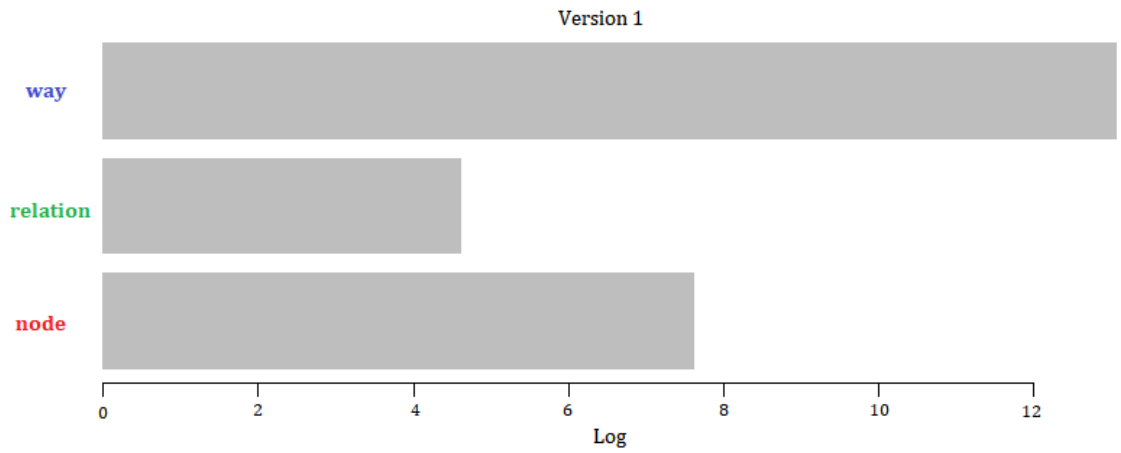


Figure 4.26: Number of building elements at version 1.

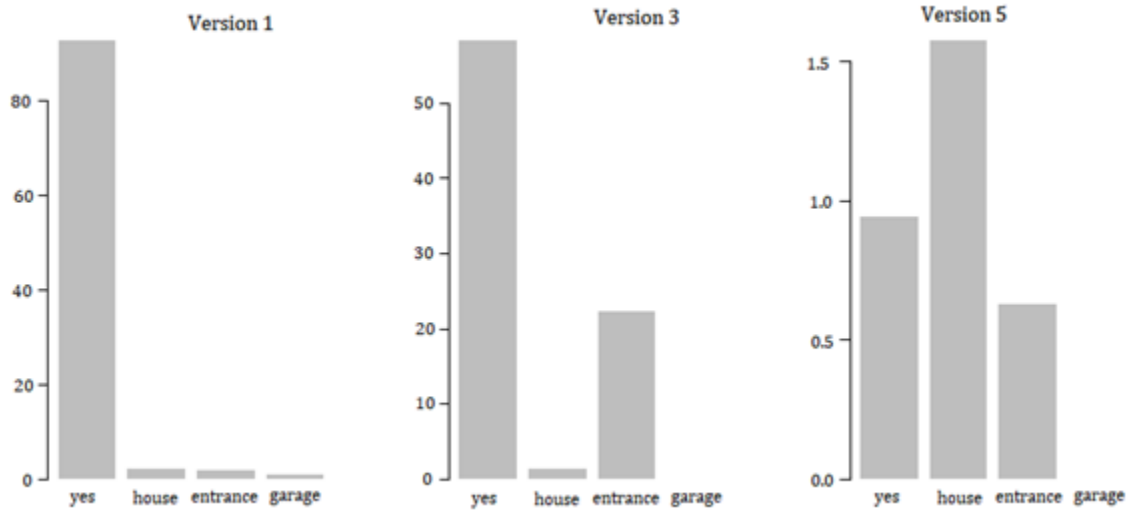


Figure 4.27: Proportion of building related element node tag values added to map objects across versions 1, 3 and 5.

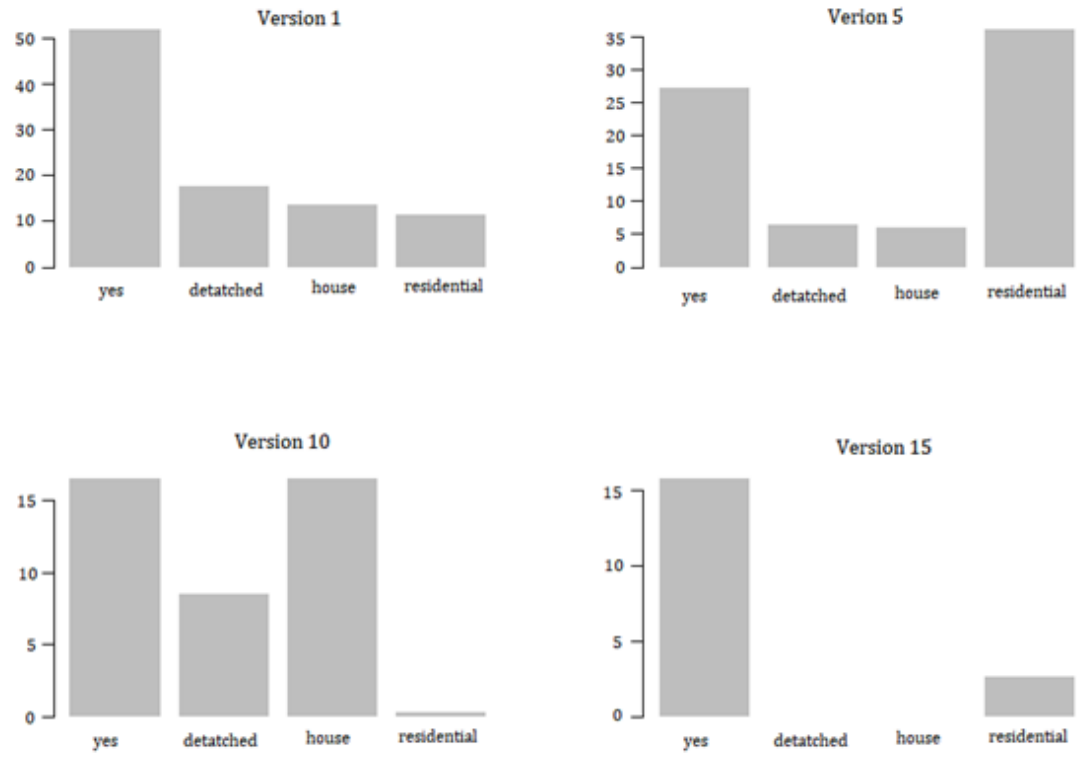


Figure 4.28: Proportion of building related element way tag values added to map objects across versions 1, 5, 10 and 15.

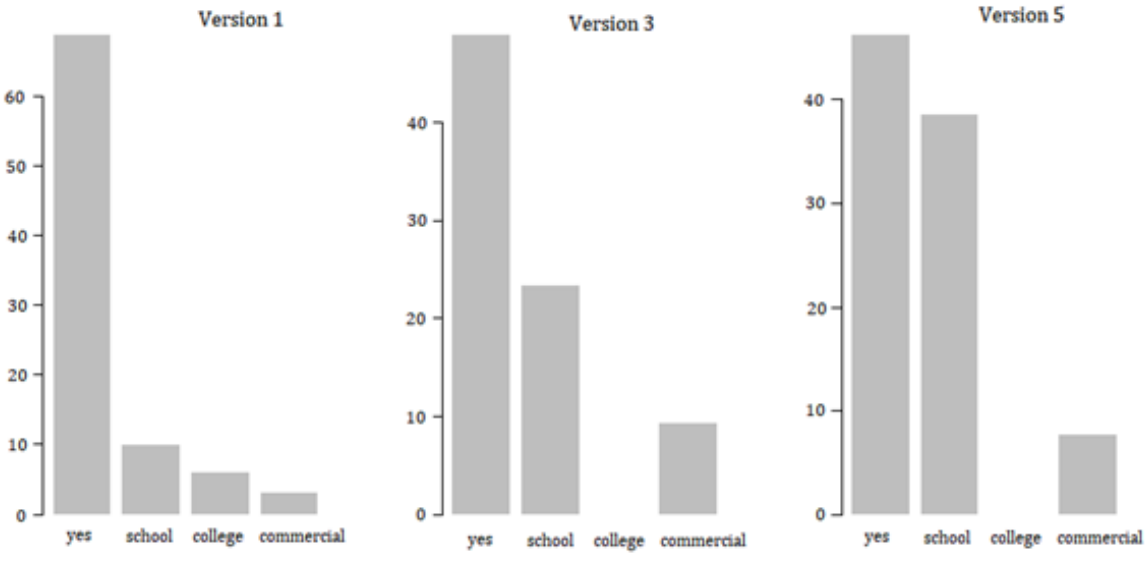


Figure 4.29: Proportion of building related element relation tag values added to map objects across versions 1, 3 and 5

4.1.4.3 OSM User Classification

For this report the DCMS dataset was used as an authoritative comparison dataset, but limitations exist surrounding the use of benchmark datasets with respect to availability, costs and the fact that there is no actual consensus on what defines a "perfect" or high-quality geographic dataset. As a result, it is possible to dive into the metadata of OSM elements or "the data behind the data" to gain an understanding of the OSM element quality. This is made possible by extracting information on element, changeset and user metadata. User ID 360 will be used as the example OSM contributor throughout Section 4.1.4 to demonstrate the metadata extracted for each of the Ottawa-Gatineau OSM contributors.

4.1.4.3.1 Temporal Characteristics

By extracting temporally related characteristics on OSM users it is possible gain an understanding on how OSM users contribute over time. Table 4.17 outlines that user ID 360 has been an OSM contributor for 1,717 days with their lifespan on the OSM website at 1,379 days. The OSM lifespan is defined as the time from the first OSM contribution to the time of the last OSM contribution. User ID 360 has made modifications to the OSM database on 3 different days. By analyzing temporal characteristics of OSM contributors, it is possible to hypothesize the contributor's experience in the OSM ecosystem. Given that user ID 360 has only contributed on 3 different days across 1,717 days (4.7 years), it is a clear indication that they are not an active OSM contributor and their contributions are sporadic.

Table 4.17: Example of extracted temporal user characteristics from the OSM history file.

UID	360
Lifespan	1379
Number of Inscription Days	1717
Number of Activity Days	3

4.1.4.3.2 Changeset Characteristics

OSM temporal metadata characteristics are then paired with changeset historical information. Every changeset in OSM had a start time and end time associated with them, along with the responsible user. Thus, the changeset quantities are calculated along with average duration for each of the changesets. Table 4.18 shows user ID 360 having a total of 4 changesets with the average changeset duration lasting slightly over 30 seconds (Average Changeset Time). This shows that user ID 360 is often opening a changeset (i.e. starting a new set of changes to OSM) and saving those edits only 30 seconds later, rather than working diligently on many changesets over a longer duration. By analyzing user changeset characteristics, it is possible to gain insight into how productive an OSM contributor may be given how long they have a changeset open.

Table 4.18: Example of extracted changeset metadata from the OSM history file.

UID	360
Lifespan	1379
Number of Inscription Days	1717
Number of Activity Days	3
Number of Changesets	4
Average Changeset Time (Minutes)	0.575

4.1.4.3.3 Contribution Intensity Characteristics

Contribution intensity examines how often OSM users are modifying OSM elements. For example, in Table 4.19, user ID 360 appears to modify OSM elements only once in about half a minute (Average Changeset Time) and very few modifications for the number of days registered as an OSM user. This implies that user ID 360 may not be as productive as some OSM contributors since they only have 1 modification on average by map element for each of their changesets (i.e. creating a node and then not changing anything else about that node). On the other hand, higher average number of modifications may suggest automated software (i.e. robot-like behaviour) that is contributing to the database with simple autocorrections.

Table 4.19: Example of extracted contribution intensity metadata from the OSM history file.

UID	360
Lifespan	1379
Number of Inscription Days	1717
Number of Activity Days	3
Number of Changesets	4
Average Changeset Time (Minutes)	0.575
Average Modifications By Element	1.045

4.1.4.3.4 Element Characteristics

To characterize how an OSM user contributes to the database, metadata on map element features must be extracted. Table 4.20 shows that user ID 306 has completed a total of 93 modifications to the Ottawa-Gatineau OSM database. Of the total 93 modifications, 86 of these were on nodes, 7 on ways and 0 on relations.

Table 4.20: Example of extracted element feature metadata from the OSM history file.

UID	360
Lifespan	1379
Number of Inscription Days	1717
Number of Activity Days	3
Number of Changesets	4
Average Changeset Time (Minutes)	0.575
Average Modifications By Element	1.045
Total Modifications	93
Total Node Modifications	86
Total Way Modifications	7
Total Relation Modifications	0

4.1.4.3.5 Modification Characteristics

Table 4.22 shows a significant difference between the number of changesets and number of modifications by user 306. This is a result from the fact that a changeset can include multiple modifications. Table 4.21 shows a short sample of the enhanced element metadata classification.

Table 4.21: Example of extracted modification feature metadata from the OSM history file.

Type	Node
ID	5330724
Version	2
Visible	False
UID	186592
Changeset ID	5710758
Initialization	True
Up To Date	True
Will Be Corrected	False
Will Be Auto-Corrected	False

Table 4.22 outlines a comprehensive description of OSM user contribution characteristics. For example, Table 4.22 depicts that user 360 has a total of 86 modifications to nodes, where they removed 3 existing nodes and added 83 new nodes to the OSM dataset. Of these 86 modifications on nodes, 70 of them were corrected by other OSM users, 3 of them were auto-corrected and 13 resulted in up-to-date nodes (meaning they have not been updated after user ID 360). A similar depiction of OSM element feature characteristics is shown for ways.

Table 4.22: User contribution characteristics with regards to OSM elements.

UID	360
Lifespan	1379
Number of Inscription Days	1717
Number of Activity Days	3
Number of Changesets	4
Average Changeset Time (Minutes)	0.575
Average Modifications By Element	1.045
Total Modifications	93
Total Node Modifications	86
Total Way Modifications	7
Total Relation Modifications	0
Node Modifications	86
Node Modifications Creation	83
Node Modifications Improvements	0
Node Modifications Deletion	3
Node Modifications Up To Date	13
Node Modifications Corrected	70
Node Modifications Auto-Corrected	3
Way Modifications	7
Way Modifications Creation	6
Way Modifications Improvements	0
Way Modifications Deletion	1
Way Modifications Up To Date	5
Way Modifications Corrected	1
Way Modifications Auto-Corrected	1
Relation Modifications	0
Relation Modifications Creation	0
Relation Modifications Improvements	0

Relation Modifications Deletion	0
Relation Modifications Up To Date	0
Relation Modifications Corrected	0
Relation Modifications Auto-Corrected	0

4.1.4.3.6 Metadata Normalization

To run the k-means algorithm to cluster OSM contributors based on certain criteria and characteristics, the OSM history metadata had to be normalized by some means. Figure 4.30 illustrates three histograms from the number of node, way and relation modification variables. The number of node, way and relation modifications for each user appear to be highly left skewed, resulting in an alternative normalization strategy. An alternative method to normalize the variables was to represent each of the variables as a percentage of another variable. For instance, the number of node, way and relation modifications can be represented as a percentage relative to all modifications, the number of created, deleted and improved elements among all modifications for each element type, and number of changesets opened by a user relative to all changesets.

However, some of the variables such as total number of node, way or relation modifications cannot be represented as percentages. These variables were represented by comparisons relative to other OSM contributors. For example, the total number of node modifications is now represented as percentage of users that did fewer node modifications. Table 4.23 shows the full contributor description of user ID 360. User ID 360 contributed more node and way modifications to the Ottawa-Gatineau OSM database than 80% and 67% of the users that contributed.

Among user ID 360's node modifications, 96% were node creations and 3% were node deletions.

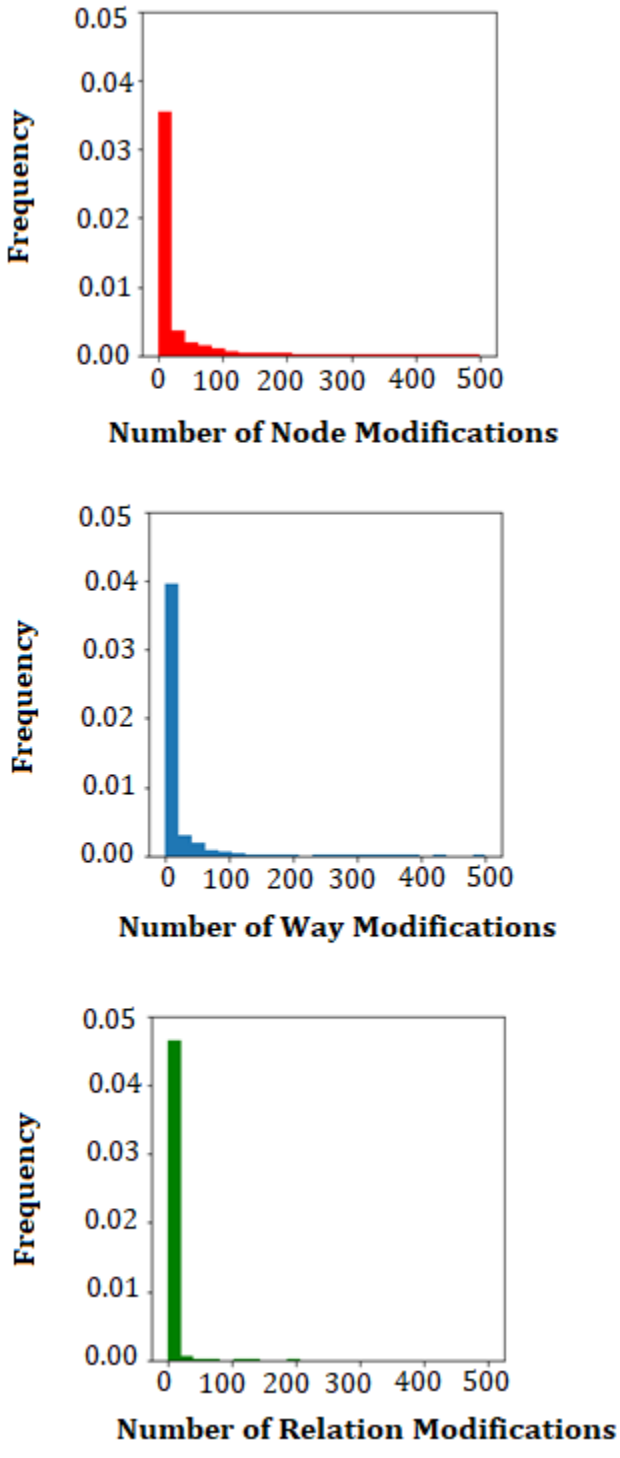


Figure 4.30: Number of node, way and relation modification variable histograms.

Table 4.23: Full user description of User ID 360.

UID	360
Lifespan	0.333
Number of Inscription Days	0.414
Number of Activity Days	0.792
User Changesets	0.705
Average Changeset Time	0
User Modifications By Element	0.704
User Total Modifications	0.773
Total Node Modifications	0.924
Total Way Modifications	0.075
Total Relation Modifications	0
User Node Modifications	0.798
Node Modifications Creation	0.965
Node Modifications Improvements	0
Node Modifications Deletion	0.034
Node Modifications Up To Date	0.151
Node Modifications Corrected	0.813
Node Modifications Auto-Corrected	0.034
User Way Modifications	0.674
Way Modifications Creation	0.857
Way Modifications Improvements	0
Way Modifications Deletion	0.143
Way Modifications Up To Date	0.714
Way Modifications Corrected	0.143
Way Modifications Auto-Corrected	0.143
User Relation Modifications	0.720
Relation Modifications Creation	0

Relation Modifications Improvements	0
Relation Modifications Deletion	0
Relation Modifications Up To Date	0
Relation Modifications Corrected	0
Relation Modifications Auto-Corrected	0
User Total Changesets	0.542
Local Changesets	0.10
Total Changesets iD	0.857
Total Changesets JOSM	0
Total Changesets Maps.me Android	0
Total Changesets Maps.me iOS	0
Total Changesets Other	0.024
Total Changesets Potlatch	0.122
Total Changesets Unknown	0

4.1.4.3.7 Principal Component Analysis (PCA) Development

4.1.4.3.7.1 Design

Figure 4.31 shows that about 74% of the cumulative variance can be explained throughout the first 7 principal components. Therefore, the latter principal components can be removed from the analysis given that they do not provide much information.

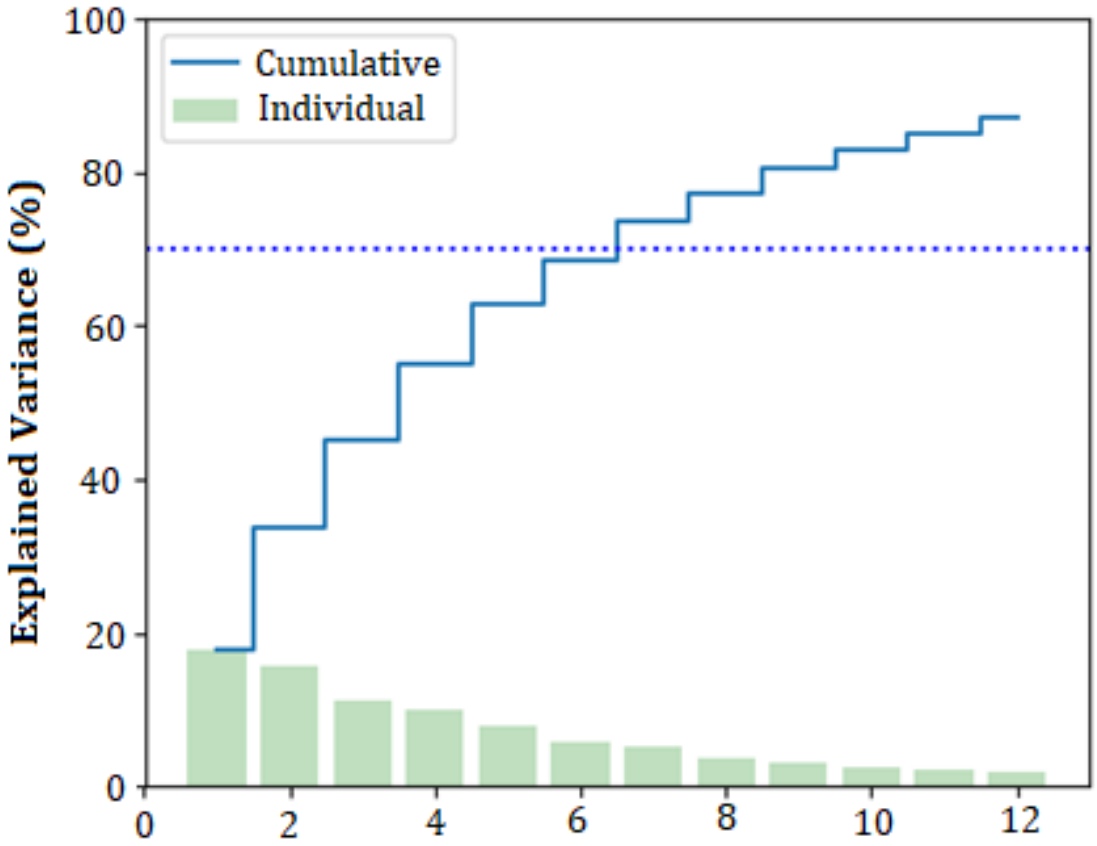


Figure 4.31: A plot that shows that most of the variance (73.8%) can be explained by the first 7 principal components. Subsequent principal components were discarded.

4.1.4.3.7.2 Implementation

The PCA algorithm is run from the *sklearn* Python module with 7 components and with a fit transformation to generate a new linear projection. After running the PCA algorithm, the metadata about the OSM user is sorted into 7 components (Table 4.25). Figure 4.32 illustrates the loadings between OSM contribution variables and each of the components (-1, a strong negative contribution; +1, a strong positive contribution). Table 4.26 outlines the variable abbreviations and their respective definitions.

Table 4.24: User ID 360 summarized by the 7 components. Please refer to Table 4.26 for the principal component definitions.

UID	360
PC1	-0.84
PC2	-0.37
PC3	-0.82
PC4	-0.11
PC5	-0.35
PC6	0.11
PC7	-0.03

For example, principal component 7 has a strong position contribution loading associated with OSM edits in the Java OpenStreetMap (JOSM) editor application (0.551). JOSM is a desktop application for editing OSM data, that has a traditional GIS user interface. According to the OSM Wiki page, JOSM has a "relatively steep learning curve and is a popular editor among more experience OSM contributors given the variety of plug-ins and stability" ('JOSM - OpenStreetMap Wiki', 2018); therefore, this characteristic can be used as an index of user experience.

Table 4.25 indicates a positive loading associated with PC6 for user ID 360. By visualizing all the principal components across each of the contribution variables (Figure 4.32), there are strong positive contribution loadings in PC6 associated with total number changesets in iD (0.503) and total number of node modification improvements (0.477). Furthermore, it is possible to conclude that user ID 360 tends to only contribute simplistic map features (nodes) using the iD editor. The iD editor is an OSM editor usable in any browser (without plug-ins or additional programs)

with the aim to be “simple and user friendly”[18]. Like the JOSM editor, this characteristic involving the iD editor can be used as an index of user experience.

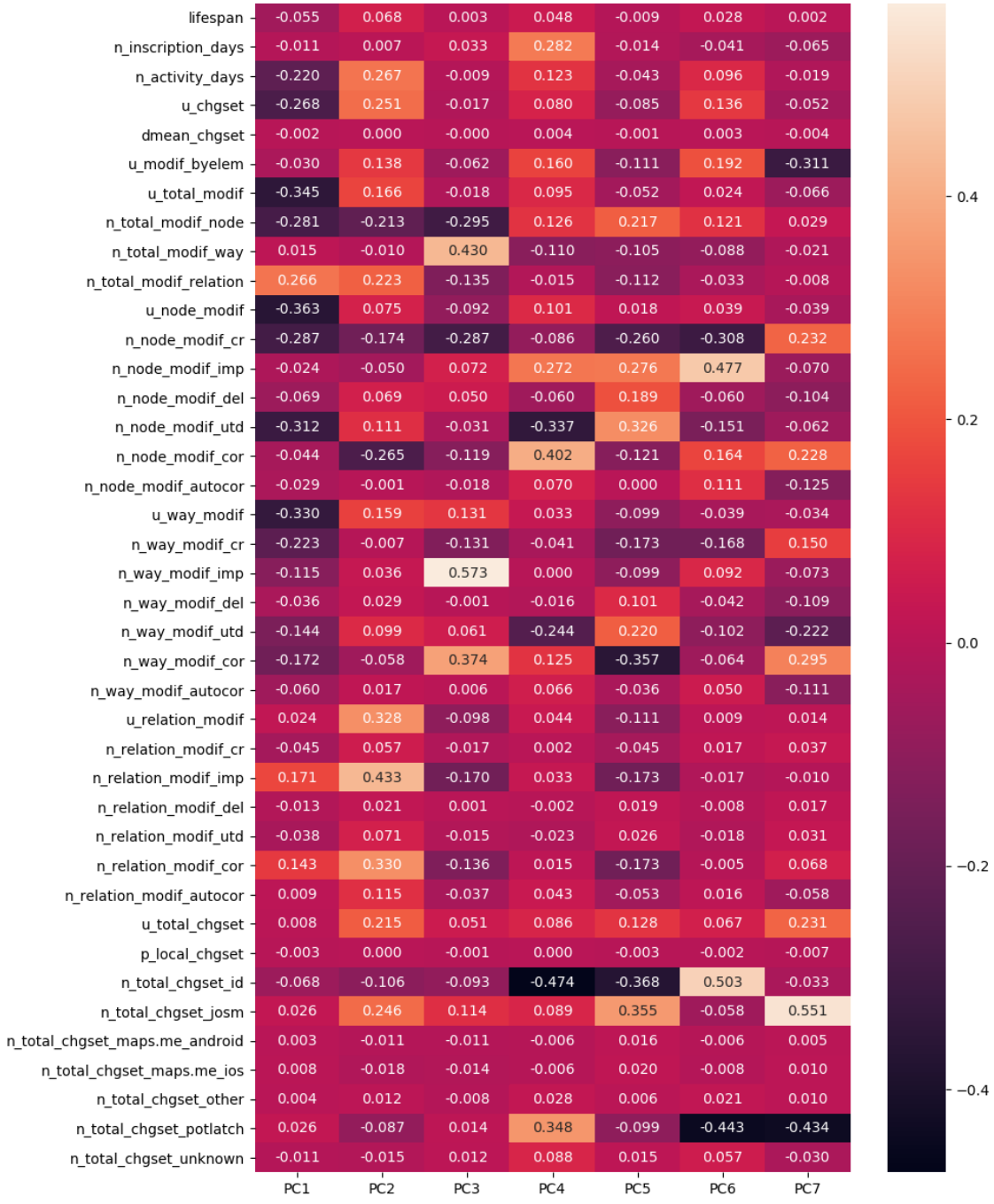


Figure 4.32: Plotted feature contribution loadings to each of the 7 principal components.

[18] https://wiki.openstreetmap.org/wiki/ID

Table 4.25: Feature contribution loading variables and their respective definitions.

Variable Abbreviation	Variable Definition
lifespan	Lifespan
n_inscription_days	Number of inscription days
n_activity_days	Number of activity days
u_chgset	User changesets*
dmean_chgset	Average changeset duration
u_modif_byelem	User modifications by element*
u_total_modif	User total modifications*
n_total_modif_node	Number of total modifications (node)
n_total_modif_way	Number of total modifications (way)
n_total_modif_relation	Number of total modifications (relation)
u_node_modif	User node modifications*
n_node_modif_cr	Number of node modifications (created)
n_node_modif_imp	Number of node modifications (improvements)
n_node_modif_del	Number of node modifications (deletions)
n_node_modif_utd	Number of node modifications (up to date)
n_node_modif_cor	Number of node modifications (corrected)
n_node_modif_autocor	Number of node modifications (autocorrected)
u_way_modif	User way modifications*
n_way_modif_cr	Number of way modifications (created)
n_way_modif_imp	Number of way modifications (improvements)
n_way_modif_del	Number of way modifications (deletions)
n_way_modif_utd	Number of way modifications (up to date)
n_way_modif_cor	Number of way modifications (corrected)
n_way_modif_autocor	Number of way modifications (autocorrected)
u_relation_modif	User relation modifications*
n_relation_modif_cr	Number of relation modifications (created)

n_relation_modif_imp	Number of relation modifications (improvements)
n_relation_modif_del	Number of relation modifications (deletions)
n_relation_modif_utd	Number of relation modifications (up to date)
n_relation_modif_cor	Number of relation modifications (corrected)
n_relation_modif_autocor	Number of relation modifications (autocorrected)
u_total_chgset	User total changesets*
p_local_chgset	Percent local changesets
n_total_chgset_id	Number of total changesets (with iD) *
n_total_chgset_josm	Number of total changesets (with JOSM) *
n_total_chgset_maps.me_android	Number of total changesets (with Maps.Me Android) *
n_total_chgset_maps.me_ios	Number of total changesets (with Maps.Me iOS) *
n_total_chgset_other	Number of total changesets (other editor) *
n_total_chgset_potlatch	Number of total changesets (with Potlatch) *
n_total_chgset_unknown	Number of total changesets (with unknown) *

*Some of the variables can be expressed as percentages of other variables:

- the number of node/way/relation modifications amongst all modifications;
- the number of created/improved/deleted elements amongst all modifications, for each element type;
- the number of changesets opened with a given editor, amongst all changesets.

4.1.4.3.8 K-Means Clustering

4.1.4.3.8.1 Design

There was no clear indication of an elbow in the plot suggesting an appropriate number of clusters. The line graph in Figure 4.33 does not outline a distinct elbow so it was decided that 4 clusters would be used.

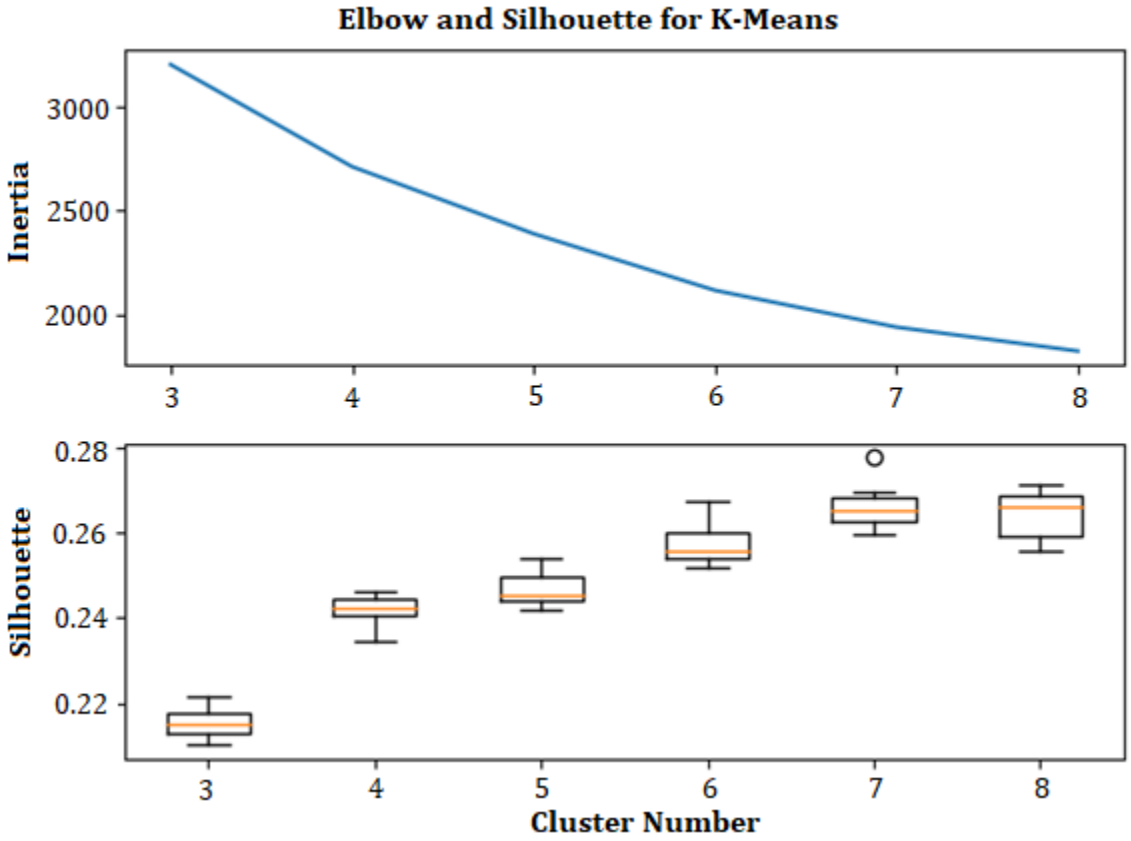


Figure 4.33: Clusters that were derived from the k-means algorithm.

4.1.4.3.8.2 Contributor Classification

The k-means clustering algorithm allocated each of the OSM contributors in the Ottawa-Gatineau region into 4 relatively balanced groups. Table 4.26 and Figure 4.34 illustrate each of the 4 clusters and their respective number of contributors associated with that cluster group. Figure 4.35 and 4.36 illustrate plot graphs for each

of the contributors according to their group. Figure 4.35 illustrates that PC1 and PC2 allow to differentiate C1 and C2, while Figure 4.36 allow C0 and C3 to differentiate from each other.

Table 4.26: Four cluster groups with their associated loading values across each of the seven principal components.

Cluster (C#)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Number of Contributors
0	0.23	-0.24	0.86	-0.12	0.02	0.01	-0.01	468
1	-0.76	0.39	-0.09	-0.03	-0.06	0.001	-0.02	669
2	1.42	0.98	-0.42	-0.03	-0.23	-0.05	-0.01	230
3	0.14	-0.71	-0.45	0.16	0.16	0.01	0.04	528

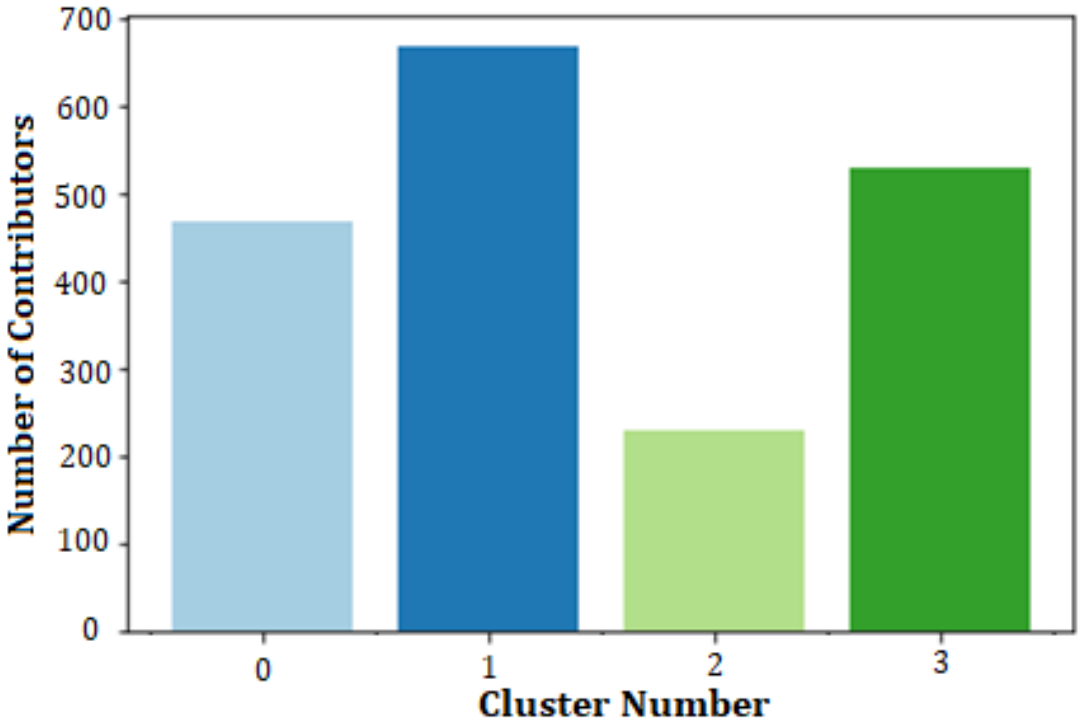


Figure 4.34: Bar graph illustrating the number of contributors for each cluster group.

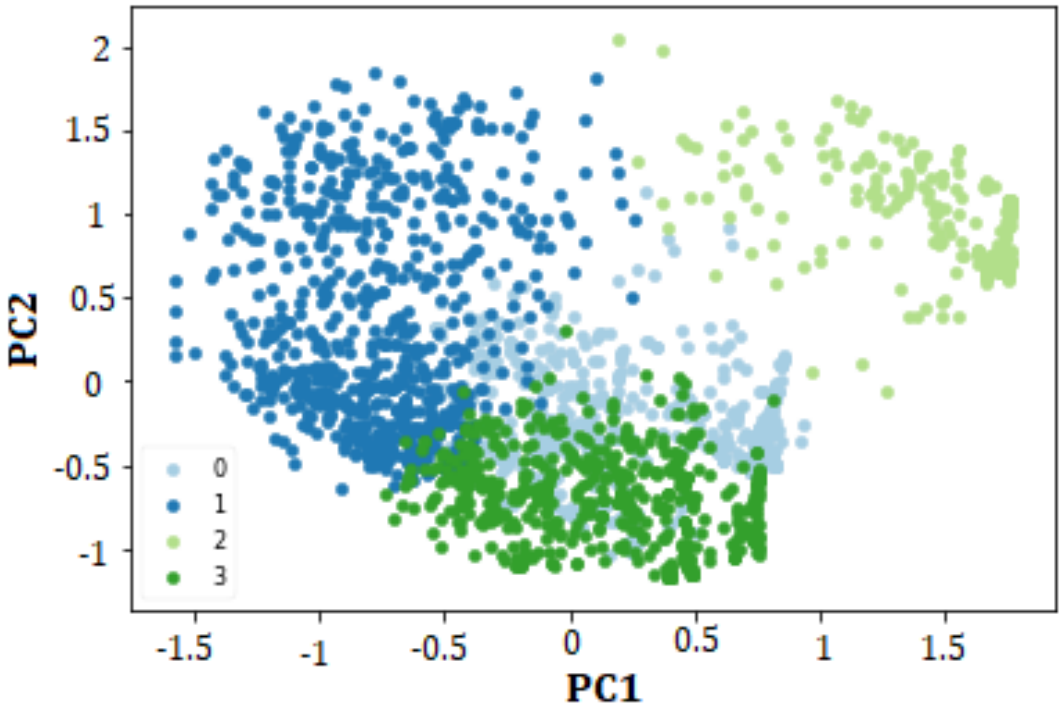


Figure 4.35: Biplots for PC1 and PC2.

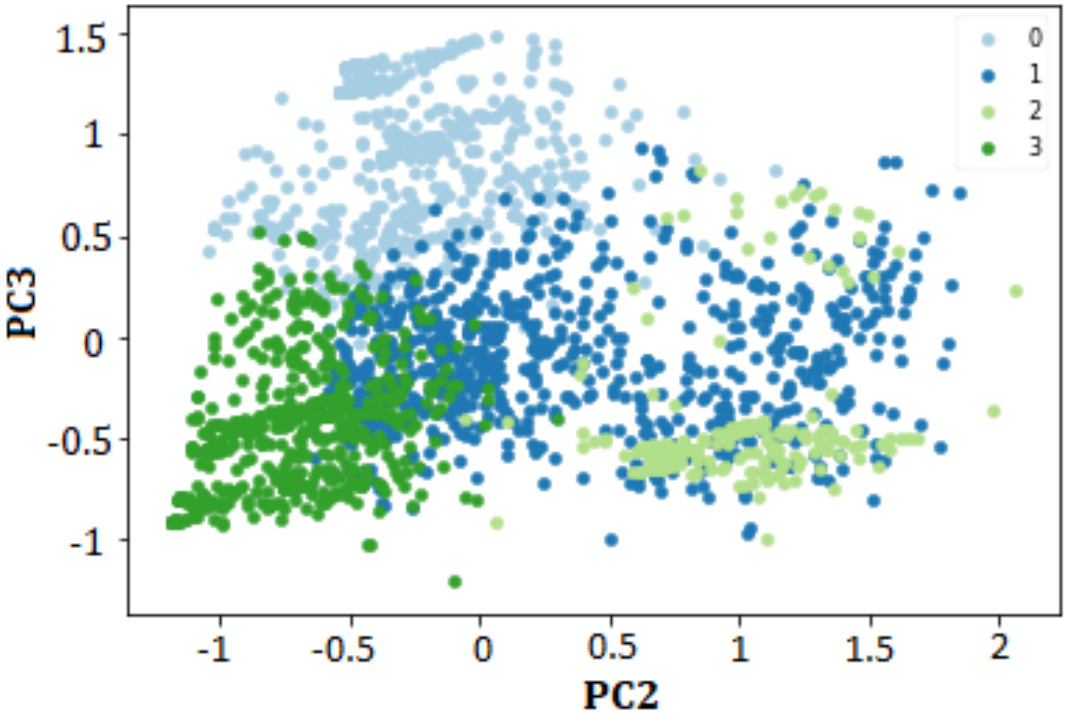


Figure 4.36: Biplots for PC2 and PC3.

4.1.4.4 Temporal Mapping

Using the OSM history file in conjunction with the OSM database existing map geometries, it is possible to visualize the metadata behind each of the OSM map objects. Figures 4.37 to 4.42 illustrate the spatial variation that exists in how an OSM database is constructed. Figures 4.37 to 4.39 examine OSM road networks in connection with the average number of active OSM contributors, version numbers and years since road network creation. Figures 4.40 to 4.42 examine OSM building footprints in connection with the average number of active OSM contributors, version numbers and years since building footprint creation. These three metadata characteristics are visualized over 1-kilometre grid cell regions.

4.1.4.4.1 Road Networks

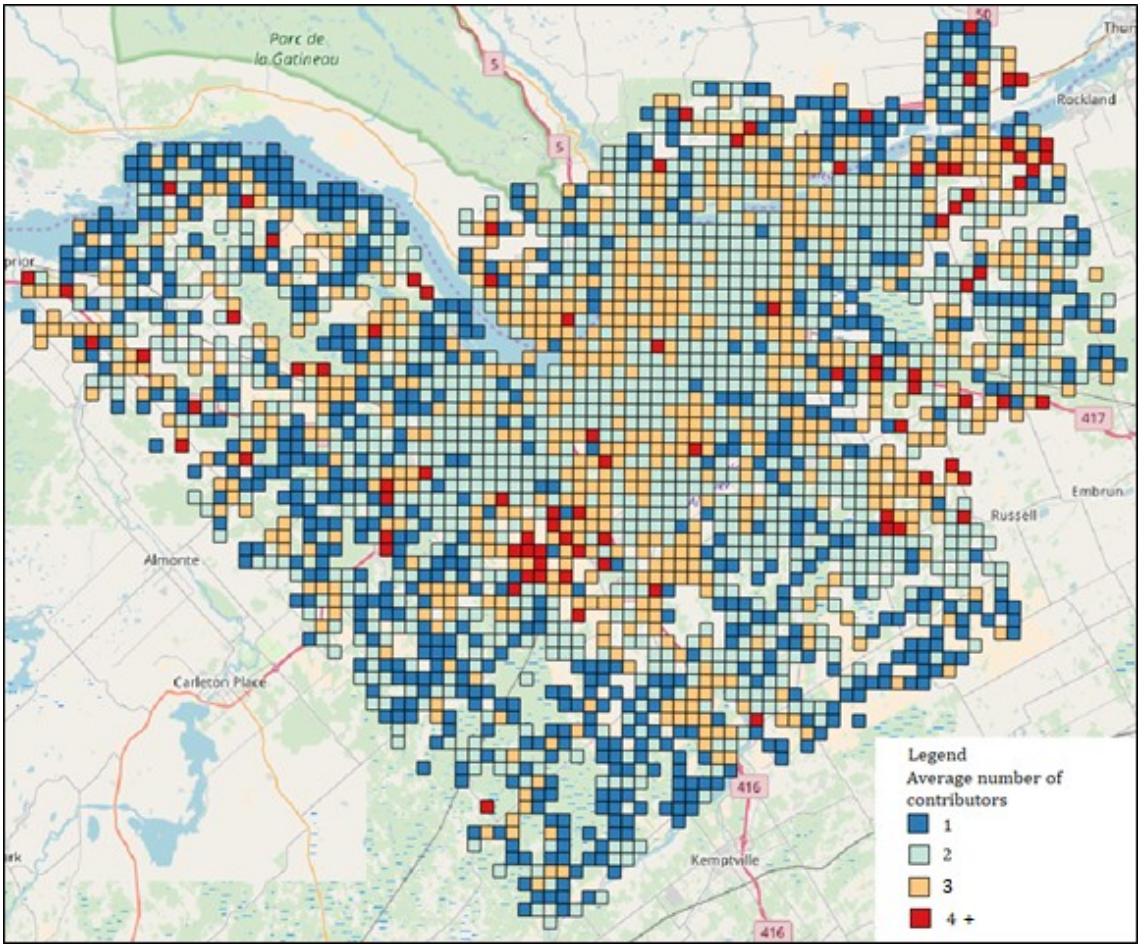


Figure 4.37: Average number of active contributors per OSM road segment in Ottawa-Gatineau over 1-kilometre grid cell regions.

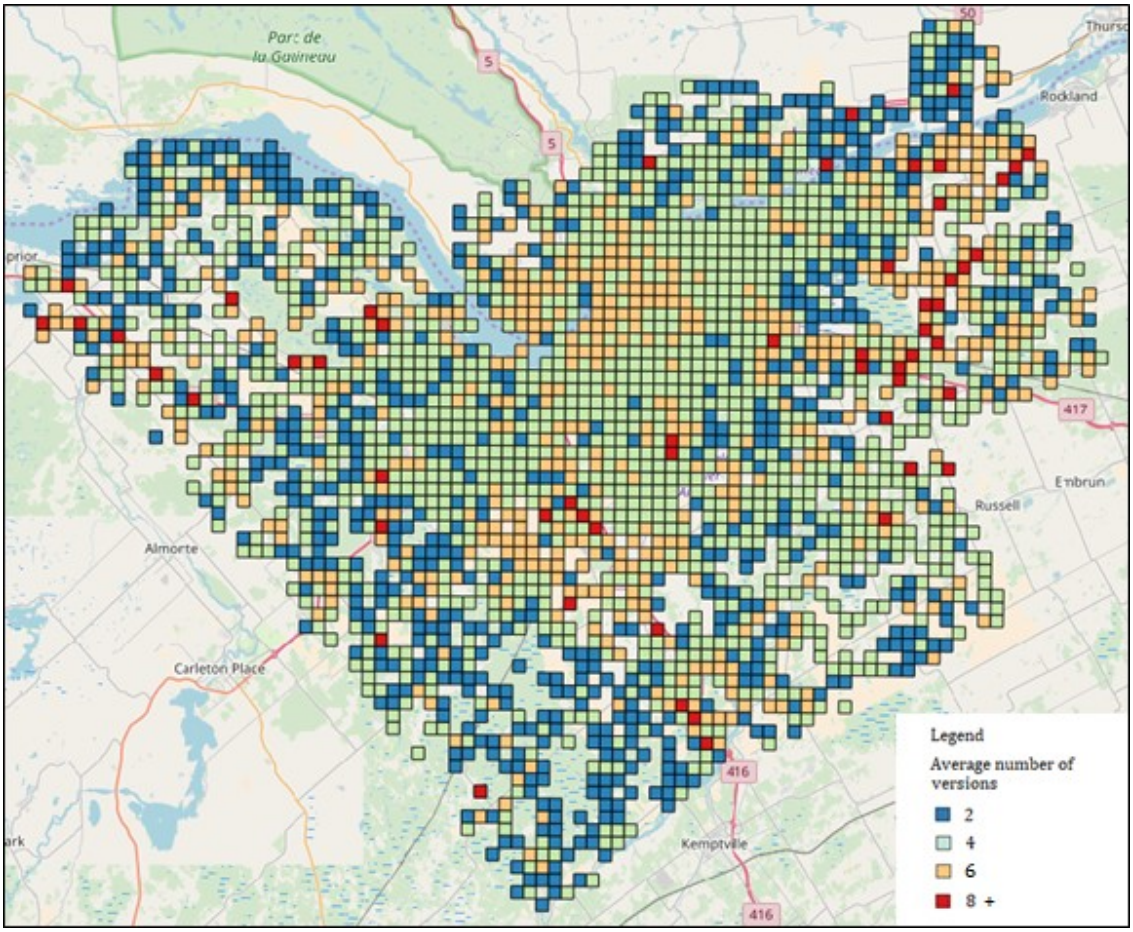


Figure 4.38: Average number of versions per OSM road segment in Ottawa-Gatineau over 1-kilometre grid cell regions.

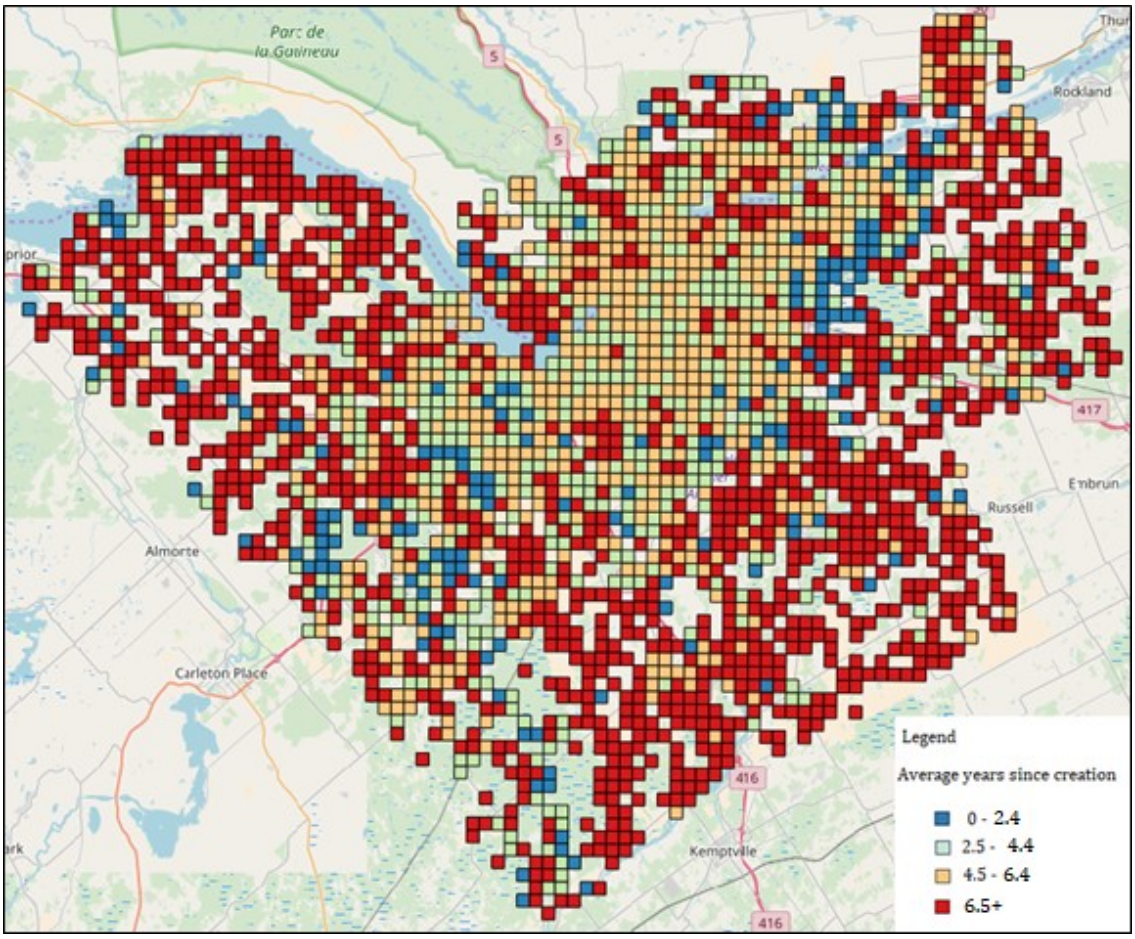


Figure 4.39: Average years since creation of OSM road segments in Ottawa-Gatineau over 1-kilometre grid cell regions.

4.1.4.4.2 Buildings

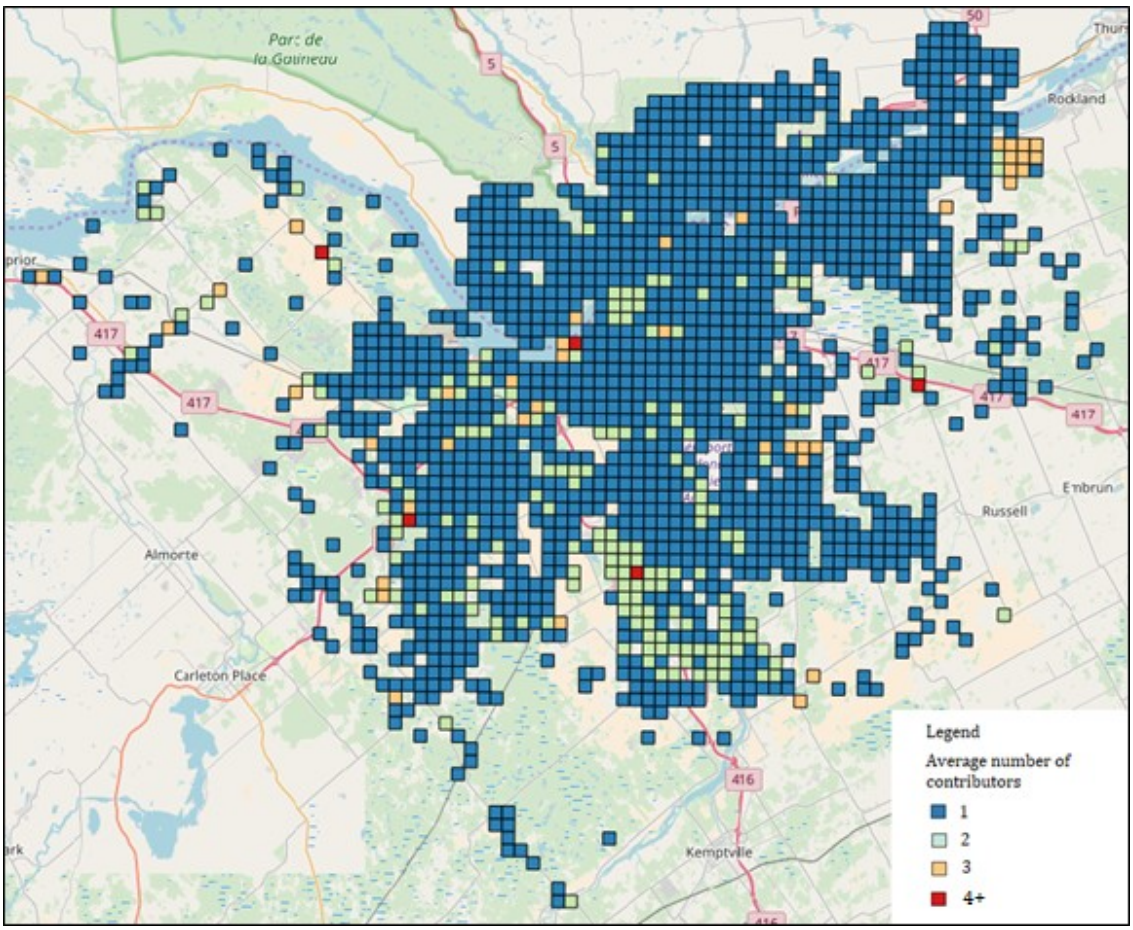


Figure 4.40: Average number of active OSM building contributors in Ottawa-Gatineau over 1-kilometre grid cell regions.

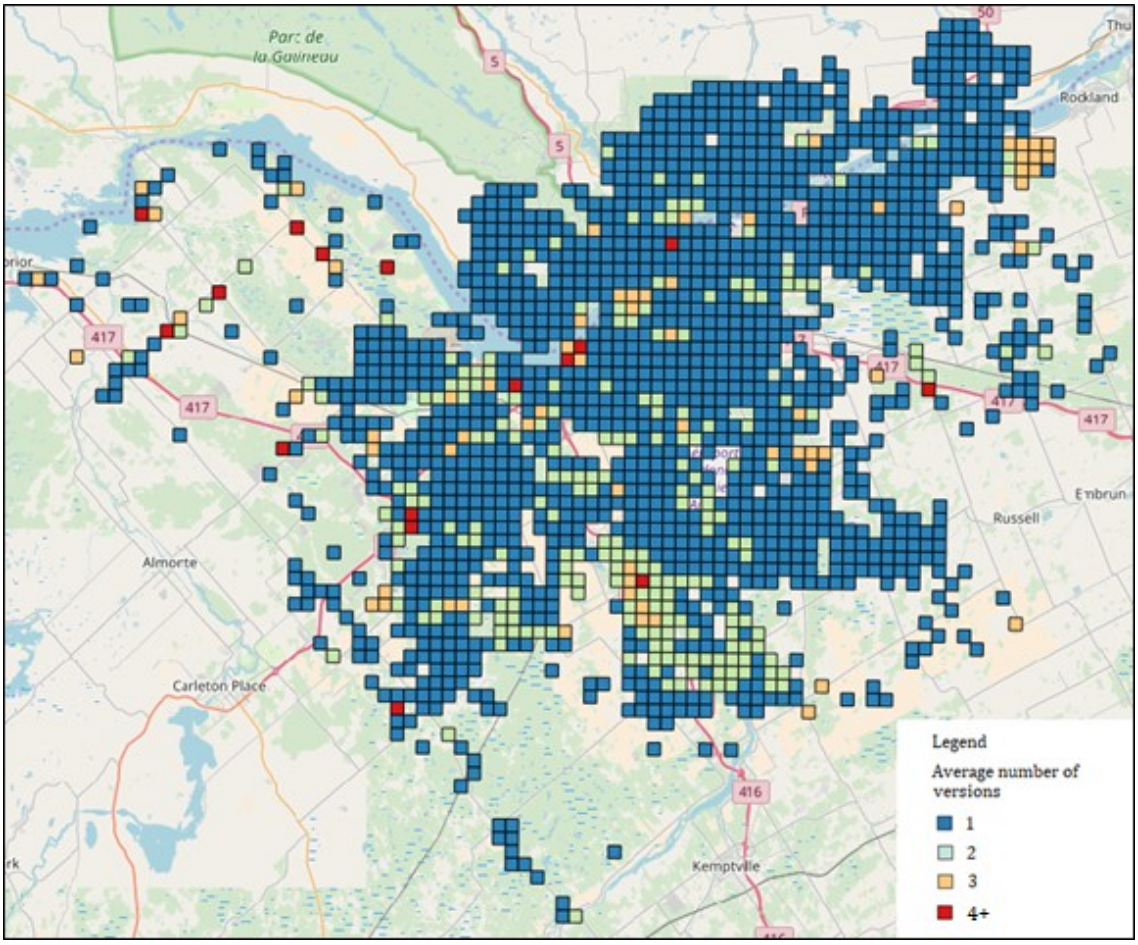


Figure 4.41: Average number of OSM building versions in Ottawa-Gatineau over 1-kilometre grid cell regions.

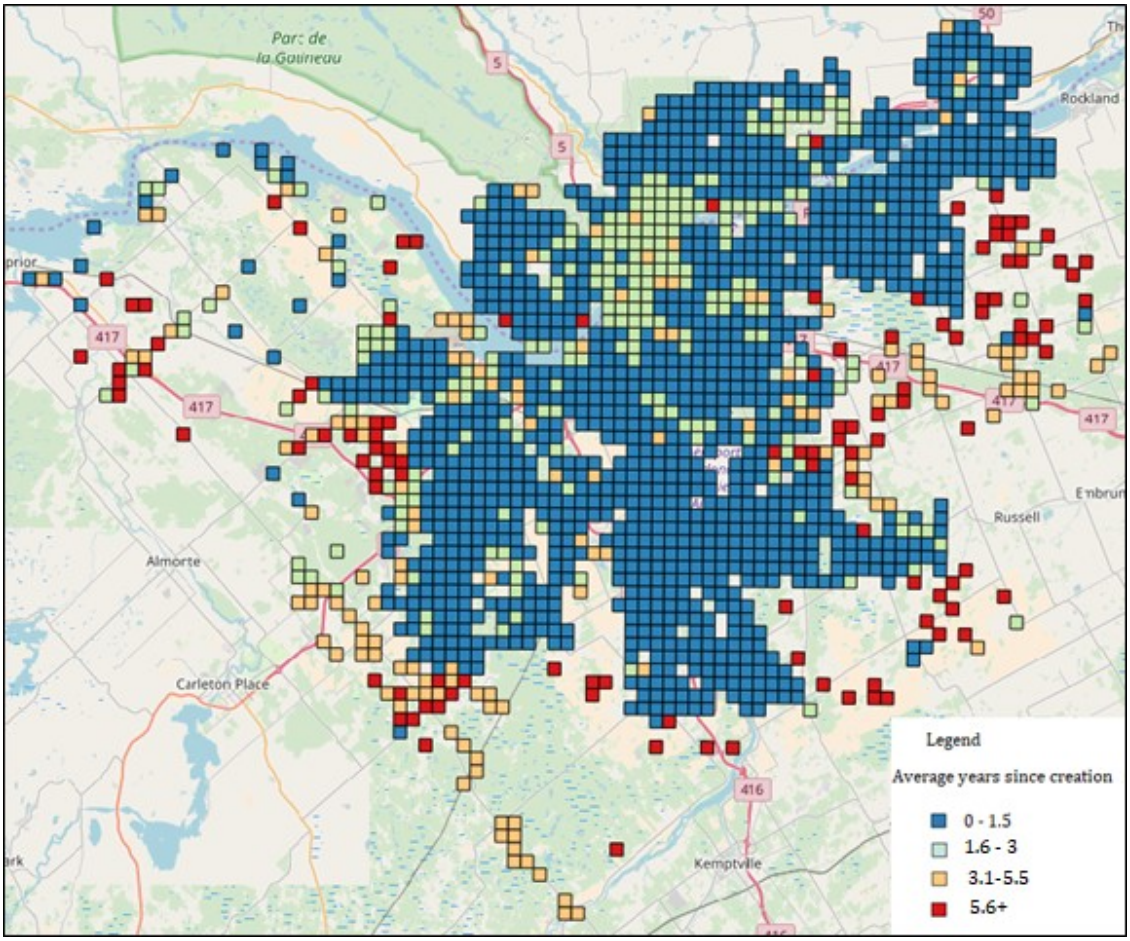


Figure 4.42: Average years since creation of OSM buildings over 1-kilometre grid cell regions.

4.1.4.4.3 User Classification Cluster Mapping

Knowing which group contributed to an OSM map feature can also help determine the quality of an OSM map feature or dataset. Figure 4.43 illustrates the road networks and buildings of the Ottawa core. Within this small AOI, the last user cluster to contribute to road networks and building features is C1. This overwhelming trend continues throughout Ottawa-Gatineau with C1 being the last identified user classification to contribute to road network or building features. Figure 4.43 shows that only 3 clusters contributed to building features in the last edits; the fourth cluster did not contribute any building information.

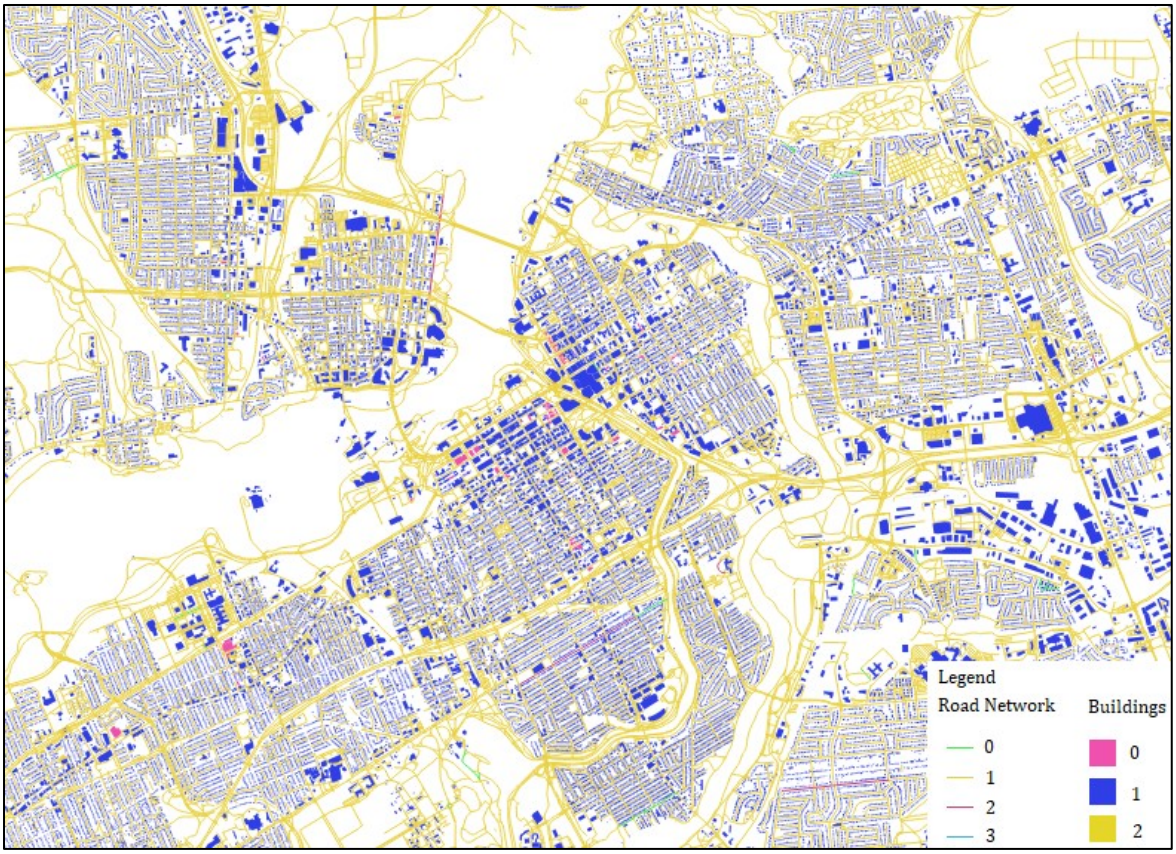


Figure 4.43: OSM road networks and buildings around Ottawa Centre, according to last user cluster classification (C0, C1, C2, C3). C3 was excluded from the building features because they did not associate themselves with the last group of users to update a feature.

5 DISCUSSION

5.1 Completeness

5.1.1 Road Networks

The evaluation of OSM road network data with respect to completeness suggests that OSM coverage is relatively complete throughout the Ottawa-Gatineau regions. However, there is a distinct disconnect between urban and rural OSM completeness as was also found in Haklay et al. (2010). OSM completeness is greater in urban Ottawa regions, specifically Ottawa Centre/Downtown. While there are sparse 1-kilometre vector grid cells with poor or neutral OSM coverage in Ottawa Centre, overall Figure 4.1 suggests positive spatial autocorrelation with the clustering of the red vector grid cells in urban and suburban regions of the Ottawa-Gatineau region.

The results outlined in Section 4.1.1.1 indicate minor change between the three OSM datasets in terms of road network length (January 2016, January 2017, June 2017). This is due to in large part the short temporal interval between the dataset downloads and that road networks within the Ottawa-Gatineau are already relatively complete. When imports of OSM compliant open data are initiated into the OSM database, typically the first map features imported are road networks. However, there appear to be regions that indicate growth of the Ottawa-Gatineau OSM dataset over the January to June 2017 time frame.

Figure 5.1 examines an instance of a grid cell in Kanata that was represented with neutral OSM coverage (-1-kilometre to +1 kilometre) to good OSM coverage (>

1-kilometre) in June 2017. Figure 5.2 indicates the DCMS representation of the road network in Kanata. This instance is indicative of the crowd contribution behind of the OSM database and the additional level of details that can be still be added to a complete dataset. The OSM road network in Figure 5.1 shows that the June 2017 OSM road network (purple) has additional connecting road networks in this subdivision that are not evident in January 2017 OSM road network (green). The DCMS (Figure 5.2) road network does not contain this level of detail.

While a time frame between January 2017 and June 2017 may not provide significant growth of an OSM dataset, it does show how the contributor base can provide additional detail to a dataset prior to changes being noticeable in proprietary datasets. The works of Girres and Touya (2011) also indicate a growth of OSM road networks over a short temporal scale of 3-months (June 2009 to October 2009).

Figure 5.1: An instance of OSM temporal growth from January 2017 to June 2017 from poor OSM coverage to good OSM coverage in a grid cell region. Green: January 2017. Purple: June 2017.

Figure 5.2: The DCMS representation of the road network from Figure 4.1.

5.1.2 Buildings

The assessment of OSM building data with respect to completeness suggests that OSM coverage surpasses building data found in the DCMS. However, it should be noted that the DCMS does not account for residential building footprints, but rather much larger commercial and government buildings. This substantial increase in OSM building data through Ottawa-Gatineau is a direct result of the STATCAN OSM Pilot Project and consultation between STATCAN and the City of Ottawa to import building data into the OSM database[19]. Throughout January 2016 to January 2017 and January 2017 to June 2017, OSM building footprints increased more than two-fold between these time frames. As a result, this allowed for the highest quality building footprint product available on an open data portal to be imported into the Ottawa-Gatineau OSM database. The results of the STATCAN initiative also raise the question of the validity of comparing an OSM dataset to an authoritative dataset if OSM has reached such a level of completeness and quality.

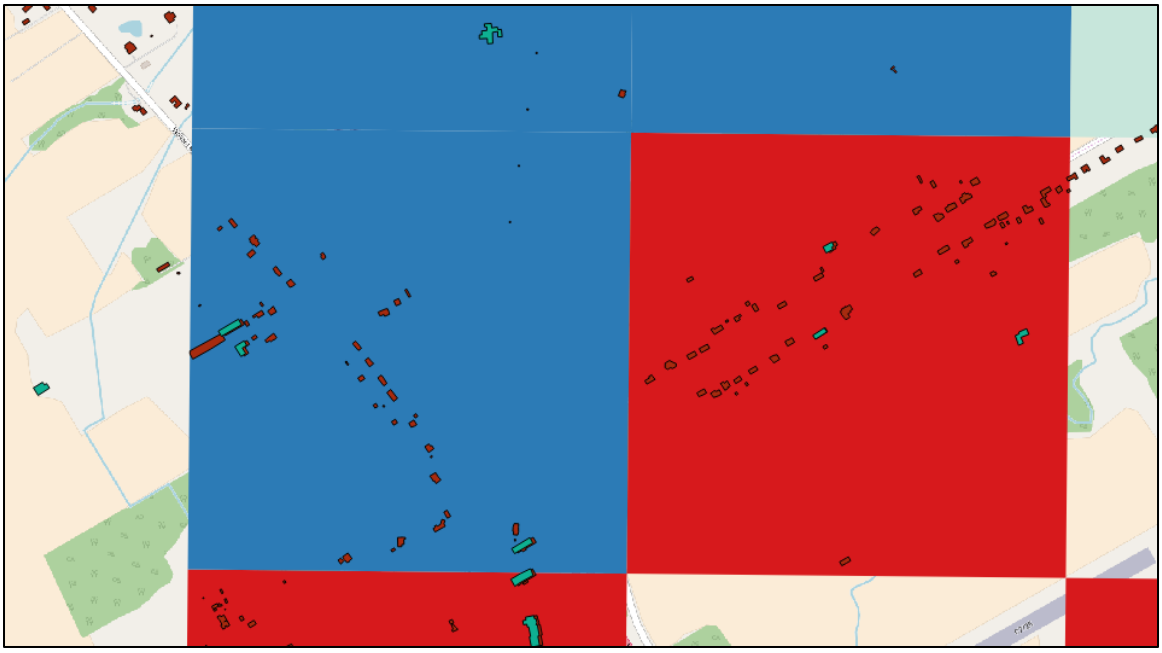
The object-based comparison results provide further suggestion of a surge in completeness as well as overall building footprint quality. From the launch of the City

[19] https://wiki.openstreetmap.org/wiki/Canada:Ontario:Ottawa/Import/Plan

of Ottawa import (October 2016) to the end (June 2017), there was an increase of 34.27% of DCMS centroids within OSM building footprints and an increase of 35.48% with respect to overlap between DCMS and OSM building footprints.

Similar trends of separation between urban and rural regions in terms of OSM completeness are also evident with respect to building data. Prior to January 2017, building data did not exist in rural regions of Ottawa (southern city limits). As the City of Ottawa import began to progress between January to June 2017, building overlap between DCMS began to increase. This planned import allowed for the rural OSM buildings to surpass DCMS rural buildings. Figure 5.3 illustrates a region of southern rural Ottawa where OSM building data now exists due to the City of Ottawa OSM import.

The comparison methodologies surrounding building footprints in this report would have benefitted greatly using a comprehensive authoritative building dataset. The works of Hecht et al. (2013) are exemplary of the in-depth analysis that can be accomplished with a comprehensive comparison dataset created by a national mapping agency. While this report does mirror the works of Hecht et al. (2013) through some of the designed methodologies, this report does not account for residential buildings that are present in the Ottawa-Gatineau OSM dataset, but merely the OSM building that matched the DCMS.



Teal: DCMS. Brown: OSM.

Figure 5.3: An instance in rural Ottawa where building data ceased to exist prior to January 2017.

The works of Fan et al. (2014) tend to agree with the disparities between urban and rural OSM building features. Like Hecht et al. (2013), Fan et al. (2014) calculated the total building area between ATKIS BDLM (reference) and OSM data. Following this calculation, Fan et al. (2014) applied a grid layer of cells over top each of the ATKIS and OSM datasets to calculate total area within each cell. This method helps to visualize the coverage and to identify which regions within a certain area are more complete (Ciepłuch et al., 2011; Haklay, 2010).

5.1.3 Geocoding

Geocoding match rates remained relatively high across January 2016 to June 2017. This is a result of the address attribute information already present in road network and building features. Ideally, address information is appended to each building feature within OSM, however, if building data do not exist, interpolation of

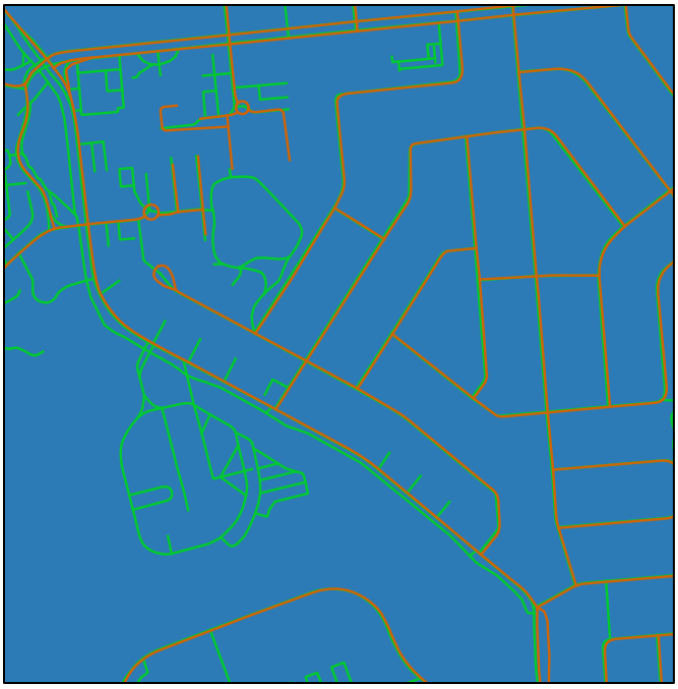
existing road networks generates geocoding results. With the City of Ottawa import, this should in turn produce higher match rates and positional accuracy of geocoding results over time.

5.2 Positional Accuracy

5.2.1 Road Networks

By implementing the Goodchild & Hunter (1997) buffer analysis methodology in the context of Ottawa-Gatineau road networks, it can be concluded that road networks maintain 78.47% of greater overlap on average within a 10 metre or greater buffer width for OSM road network data captured across January 2016, January 2017 and June 2017. The temporal scale between January 2016 to June 2017 does not allow for meaningful change between overlap percentage. This is due to the buffer analysis methodology being a measurement of dataset similarity rather than completeness.

Within the Ottawa Downtown core, overlap percentage remain relatively low (<25%), while in rural regions overlap remains high (>75 %). This is due to the OSM road work being more complete in Ottawa's Downtown core and urban centres, thus decreasing the overlap percentage. Figure 5.4 outlines an instance where a grid cell is represented as low overlap percentage due to the OSM road network length being greater and not lying within the DCMS 10 metre buffer zone.



Green: OSM. Brown: DMTI.

Figure 5.4: A grid cell in Gatineau that represents low overlap percentage.

5.2.2 Geocoding

Positional accuracy of geocoding results does improve relative to the "ground-truth" coordinates (Table 4.10). The median distance from ground-truth decreased two-fold between January 2016 and 2017 to June 2017. Results returned from June 2017 OSM data snapshots also decreased with respect to maximum, minimum and mean returned coordinates. The improved geocoding completeness and positional accuracy is due to a result of the City of Ottawa building and address imported into the OSM database. By associating an address with each building, this will in turn improve the overall accuracy of geocoding results and provide a more complete building dataset for multiple use cases.

Figure 5.5: Building level geocoding results from June 2017 OSM data.

5.3 Thematic Accuracy

5.3.1 Road Networks

With respect to thematic (attribute) accuracy, it was discovered that there was a greater percentage of DCMS road segments with road names relative to OSM. However, this percentage of road names in the DCMS remained stagnant at 85.04% between the 2016 and 2017 data products. The percentage of OSM road segments with names increased between January 2016 and January 2017, then decreased in June 2017. While the percentage of OSM road segments with names is lower than the percentage of DCMS road segments, it should be noted that not all OSM road segments need name fields linked with them.

For instance, *highway=service* is a common tag for road networks found in OSM. The tag *highway=service* is tagged to a road used for general access to a building, service station, commercial park, driveways etc. (which typically do no have names). There are approximately 13,503 road segments in the Ottawa-Gatineau region with *highway=service* tags, however, 606 of those have names tagged to them when they

necessarily might not need names because of their specific use of building access. Consequently, this would lead to discrepancies when calculating percentage of road networks that have road names tagged to them.

5.3.2 Buildings

The increase in building tags (excluding *building=yes*) between January 2016 to June 2017 is a direct result of the STATCAN pilot project and the Ottawa OSM community. From January 2016 to June 2017, the percentage of buildings with tags increased from 56.15% to 84.72%. The tag *building=yes* was excluded from this percentage calculation because tagging a *building=yes* in OSM is the equivalent of not tagging the building (default tag is *building=yes*). While the number of buildings with name tags increased from 1,986 to 3,908 between January 2016 to June 2017, the percentage decreased.

Consequently, this is a result of the bulk City of Ottawa import that contributed nearly 300,000 building footprints to Ottawa-Gatineau. Since residential homes do not have name tags associated with them, this in turn would decrease the percentage of buildings with name tags. An increase in DCMS building attribute accuracy also emerged between 2016 and 2017 data products. Table 4.15 outlines a 5.62% increase in building type tags (excluding *"TYPE" = OTHER*) and slight increase in building name tags.

5.4 Temporal Accuracy

5.4.1 Temporal Evolution

Parsing the OSM history file allowed for a full review of the Ottawa-Gatineau OSM dataset, by inspecting the total number of nodes, ways and relations contributed. Around late 2016 and late 2017 there is a significant increase in map elements and contributors added to the Ottawa-Gatineau OSM database. This is a direct response of the City of Ottawa OSM building import and launch of the STATCAN OSM pilot project.

5.4.2 OSM Tag Structure: Object Classification

Inspecting the overall OSM tag structure revealed that the most frequent tags used were *source*, *addr:city*, *addr:housenumber* and *addr:street* regardless of the OSM map element type. Each of those specific tags align with the typical tag sets used to describe a detached residential house (*building=detached*) as acknowledged by the TagInfo OSM web browser tool (Figure 5.6). Figure 5.6 shows that *addr:housenumber*, *addr:street*, *addr:city* and *source* tags are the most common tags for map features tagged as *building=detached*.

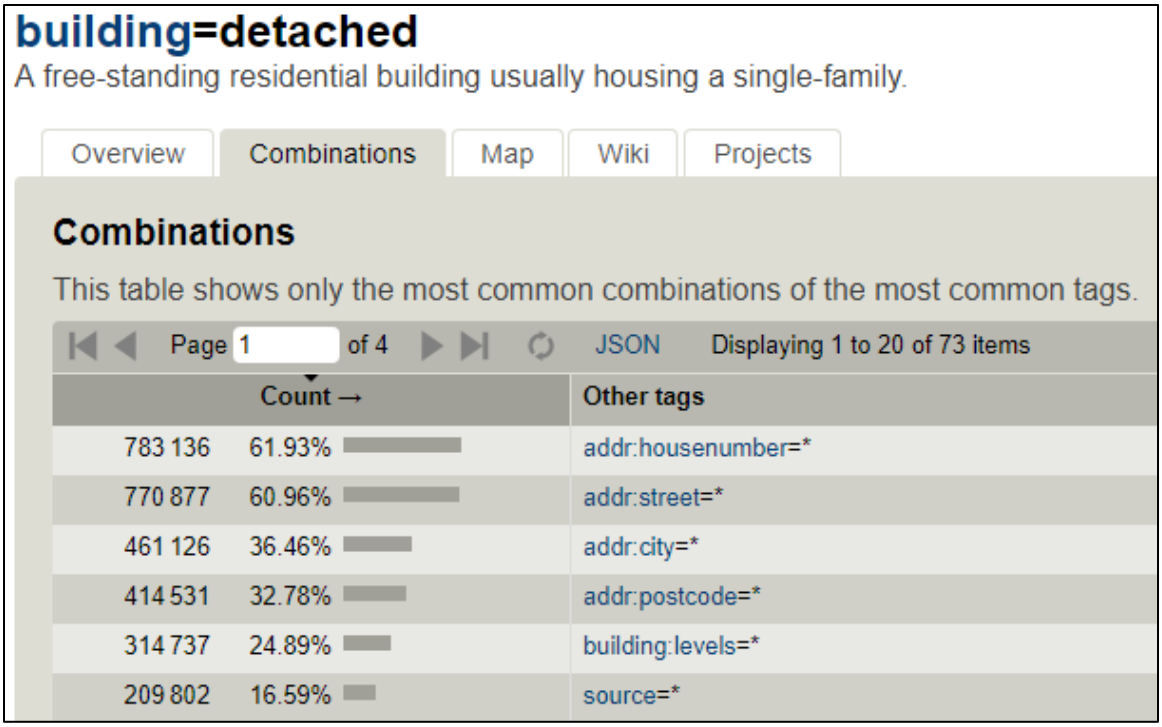


Figure 5.6: Most common tags for building=detached in the OSM dataset as indicated by taginfo.openstreetmap.org.

Figure 5.7 shows the tag structure of a detached residential building in Ottawa with four of the most frequent map tags used describing this map feature (except *addr:city*). Given the results indicated by TagInfo OSM in Figure 5.6, the tags associated with the building in Figure 5.7 are in fact compatible with what is typically associated with detached residential buildings in the OSM dataset globally.

Furthermore, Figure 5.7 indicates that the detached house is on version 2 with a comment noting that *addr:city* was removed from the element to clean up inconsistency. Prior to this detached home being edited it would have been on version 1, with the *addr:city* tag still present. The instance displayed in Figure 5.6 demonstrates the work of validation and clean-up present in the OSM community to ensure the highest quality and most consistent data.

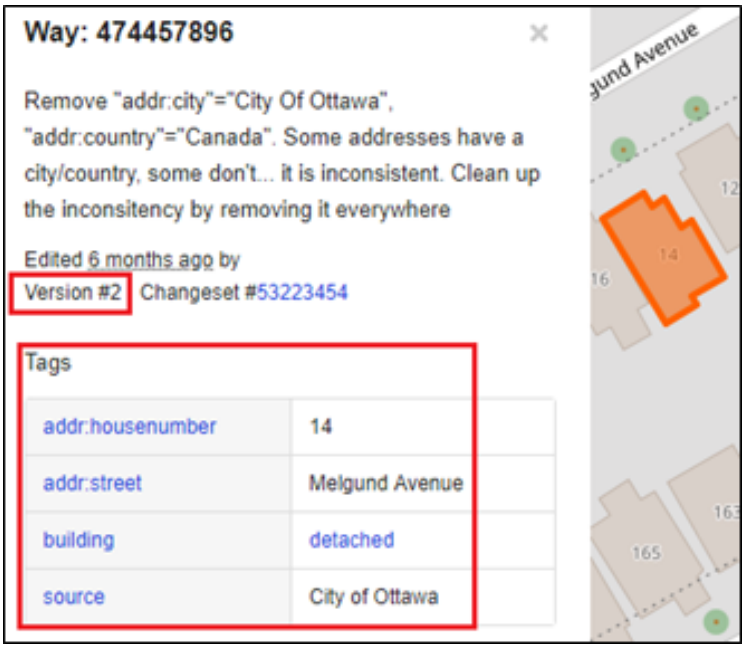


Figure 5.7: The tag structure of a detached residential home in Ottawa.

Tag values *house* and *detached* remain relatively consistent through version numbers 1 to 10. This is a direct result of the City of Ottawa OSM building import and some of the inconsistencies that may exist in the database as a result. Figure 5.7 shows an instance where this inconsistency may exist concerning the OSM tag structure. The building outlined in the top of Figure 5.7 shows a *building* tag value type of *detached* with the direct neighbour tagged as *yes*. Haklay (2010) and Girres and Touya (2010) highlight the inconsistencies that may exist in VGI (OSM) in terms of its data quality. Discrepancies in tagging structure, as shown in Figure 5.8, are a direct consequence for having unorganized groups of OSM contributors (Haklay, 2010).

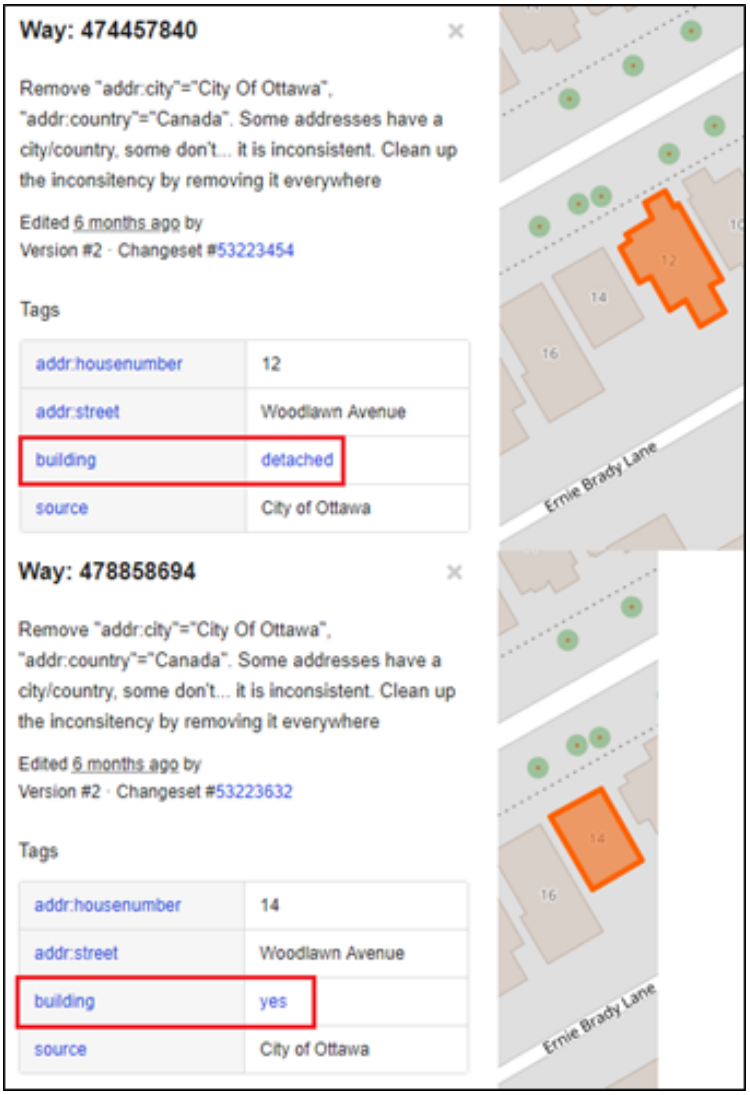


Figure 5.8: An instance of inconsistent tag values concerning residential homes in Ottawa.

5.5 Classifying OSM Contributors

While it is clear, through past studies (Ciepłuch et al., 2010; Girres & Touya, 2010; Haklay, 2010; Mooney & Corcoran, 2012; Zielstra & Zipf, 2010) and this thesis that snapshots of OSM data can be trusted in terms of quality, there are still concerns surrounding its reliability related to the contributor base. "Graffiti" or "vandalism" will continue to occur in OSM once exposure of the dataset begins to expand. For example, when Niantic Inc. transitioned from using Google Maps base map tiles to

using OSM data within their Pokémon Go game, unaccustomed map features began to appear in the OSM dataset due to this new contributor base (i.e. fictitious water bodies intentionally inserted in backyards to spawn certain Pokémon species, to enhance game performance)[20].

Nevertheless, quality assurance applications do exist to heavily monitor these types of map contributions and new applications are constantly being developed by organizations to adapt to users' behaviours (e.g., Osmose, OSMCha, etc.). That said, Niantic Inc. will be adjusting their algorithms so that areas with good or familiar map contributions to OSM get increasing spawn power, and areas where players try to cheat will be reprimanded[21]. Therefore, by diving into the OSM history file, it is possible to gain an accurate understanding of the OSM contributor base that tends to map in a region.

The k-means clustering algorithm classified each of the Ottawa-Gatineau OSM contributors into four cluster groups given the unique ways that they contributed to the dataset. Each of the clusters have varying experience with OSM and focused on different map features. Given C0's strong positive contribution loading values associated with PC1 and PC3, it can be determined that these users tend to modify OSM ways and nodes. C1 contributors could be identified as the "OSM validators" or "experts" given their stronger contribution loadings associated with overall activity (number of activity and inscription days), experience with the JOSM editor and weaker loading values associated with feature creation and frequency their

[20] https://wiki.openstreetmap.org/wiki/Pok%C3%A9mon_Go
[21] https://blog.openstreetmap.org/2018/04/01/niantic-openstreetmap-collaboration/

contributors are further corrected. C2 and C1 users are similar given that they both tend to contribute to OSM relations, except that C2 users are far less versed with OSM. C3 is associated mainly with inexperienced OSM users that tend to improve nodes in Potlach or iD editors.

5.5.1 Temporal Mapping

By visualizing the Ottawa-Gatineau OSM contributor characteristics (average number of contributors, versions, years since creation) over 1-kilometre grid cell regions, disparities between urban and rural areas emerge, trends that agree with observations in Haklay (2010) and Fan et al. (2014). Road networks across the Ottawa-Gatineau region tend to reach a higher number of version and contributors within Ottawa's city core relative to regions in the rural extent. Given that rural areas tend to grow at slower rates, much of the hinterland road networks were initialized into the OSM dataset 6-10 years ago. Within regions of eastern (Orleans-Blackburn Hamlet) and western Ottawa (Kanata-Stittsville) many of the grid cells are categorized as 0-2.5 years since created. Thus, the areas identified within these grid cells are likely new subdivision complexes. Building features also tend to follow many of these similar trends.

C1 contributors are associated with OSM validation and are the most active contributors to the Ottawa-Gatineau OSM dataset. The majority of the OSM road network and building features found within Ottawa-Gatineau tend to have C1 as the last group to contribute to features. This lends further evidence to the notion that C1

users are the “OSM validators” and most experienced contributors of the Ottawa-Gatineau dataset.

Furthermore, by knowing how OSM users contribute to the OSM database, it is possible to conclude that the OSM map feature is at its highest accuracy and quality without relying on a comparison dataset. The majority of OSM quality assessments have been relying on methods surrounding comparison datasets, although studies that involve intrinsic OSM contributor assessments have been gaining momentum (Barron et al., 2014; Gröchenig et al., 2014; Rehrl & Gröchenig, 2016). While conventional comparison methods provide beneficial contribution to the field of OSM quality assessment, questions begin to arise on the comparison dataset quality (Gröchenig et al., 2014). This is in large part due to the black box that many proprietary companies work in regarding their data acquisition and collection methods.

6 CONCLUSION

6.1 Reviewing Aims & Objectives

The overall aim of this research was to evaluate the quality and reliability of OSM map features (road networks, buildings, etc.) in the Ottawa-Gatineau region. This specific aim was to be accomplished by satisfying the following research objectives: (1) evaluate the effectiveness of contemporary methods used to assess the accuracy of the Ottawa-Gatineau OSM dataset relative to benchmark datasets; (2) analyze the spatial and temporal variation between the outlined quality measures; (3) parse the Ottawa-Gatineau OSM history file to analyze the chronological dataset evolution and attribute structure; and (4) characterize and identify OSM users based on their contribution characteristics and tendencies.

6.2 Research Findings

This research used a case study in Ottawa-Gatineau to examine the quality and reliability of OSM data, specifically road networks and building footprints. Overall, the quality of OSM road networks is comparable to or surpasses road networks found within the DCMS. Building footprints in the Ottawa-Gatineau OSM dataset were also more complete than that found in the DCMS. This is in large part since the DCMS only included large commercial building footprints. However, the City of Ottawa building footprint import instigated by the local OSM community and STATCAN signifies how civic good can come out of governmental organizations opening public datasets. In turn, this also raises the question surrounding the legitimacy comparing OSM to an

"authoritative" data source if certain OSM communities have reached such an elevated level of data quality.

By breaking down the OSM history file, it was discovered that contributors in the OSM database could be grouped based on their mapping behaviours. The results obtained in Chapter 4.1.4 exposed a group of experienced OSM contributors that could be regarded as "OSM validators". These experienced contributors were register users of OSM for much longer periods of time and tended to map more complex map features (i.e. relations), and with more advanced editing software (i.e. JOSM). It was discovered that these experienced OSM contributors were usually the last cluster group to contribute to building and road network map features. Thus, given their longevity exposed to the OSM ecosystem, tendencies that involve contributing and validating map features, this group could be viewed as the validators of the OSM Ottawa-Gatineau dataset.

6.3 Conclusions, Limitations & Future Research

The availability of OSM compliant data sources will continue to grow at a national level across Canada. In turn, this will only benefit the OSM project in Canada moving forward, ensuring that the highest quality data is included in the OSM database. The OSM project shows the achievements that can be accomplished through collaborative mapping efforts. As a result, crowdsourcing and OSM have instigated a culture change how government organization collect and gather

information as proved by the STATCAN crowdsourcing project and the Building Canada 2020 initiative[22].

This research has examined the accuracy and quality of OSM data through comparative analysis using an authoritative benchmark dataset. Methodologies of past VGI academic studies were explored, along with new ways to inspect the history of the OSM dataset. However, there were several limitations to the research that could be reported on here. While road network accuracy was examined, this research did not evaluate the use of those road networks for executing complex algorithms. The metadata associated with road networks (i.e. speed, type, width, etc.) provides crucial information for route optimization (Graser et al., 2013; Basiri et al., 2016). This research also did not examine the accuracy of OSM nodes, specifically POIs. By analyzing OSM POIs, it is possible to immerse deeper into the realm of OSM Nominatim geocoding result quality.

The majority of this research evaluated accuracy and quality of OSM data through quality measures outlined by the ISO/TC 11, while other factors introduced by Basiri et al. (2016) were not discussed; they examined the use of data mining strategies to investigate and validate OSM data, which in turn could greatly improve automated quality assurance processes and applications. With Niantic Inc. transitioning from Google Maps to OSM, the use of gamification could assist with quality assessment and assurance of the OSM database. Further exploration into the use of gamification or contributor training will assist with the concerns surrounding

[22] https://wiki.openstreetmap.org/wiki/WikiProject_Canada/Building_Canada_2020

OSM of retaining longstanding contributors as discussed in this research. The results of this research and initiatives such as the STATCAN Crowdsourcing initiative raise the question of validity of comparing an OSM dataset to an authoritative dataset if OSM has reached such a level of completeness and quality. As a result, further research should depart from comparing to an authoritative or benchmark dataset, to using machine learning methodologies to gain an understanding about the OSM contributor base and to develop additional indices of data quality.

7 REFERENCES

Amelunxen, C. (2010). On the suitability of volunteered geographic information for the purpose of geocoding.

Antoniou, V., Morley, J., & Haklay, M. (2010). Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica*, *64*(1), 99–110.

Ather, A. (2009). A quality analysis of openstreetmap data. *ME Thesis, University College London, London, UK*, *22*.

Barron, C., Neis, P., & Zipf, A. (2014). A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, *18*(6), 877–895.

Basiri, A., Jackson, M., Amirian, P., & Pourabdollah, A. (2016). Quality assessment of OpenStreetMap data using trajectory mining. *Geo-Spatial Information Science*, *19*(1), 56. https://doi.org/10.1080/10095020.2016.1151213

Brovelli, M. A., Minghini, M., Molinari, M. E., & Zamboni, G. (2016). Positional Accuracy Assessment of the Openstreetmap Buildings Layer Through Automatic Homologous Pairs Detection: the Method and a Case Study. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLI-B2*, 615–620.

Ciepłuch, B., Mooney, P., & Winstanley, A. C. (2011). Building Generic Quality Indicators for OpenStreetMap. Presented at the 19th annual GIS Research UK (GISRUK). Retrieved from http://www.port.ac.uk/special/gisruk2011/

Ciepłuch, B., Jacob, R., Winstanley, A., & Mooney, P. (2010). Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In

Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Vol. 337).

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered Geographic Information: The Nature and Motivation of Produsers. *International Journal of Spatial Data Infrastructures Research*, *4*, 27.

Craglia, M., Ostermann, F. O., & Spinsanti, L. (2012). Digital Earth from vision to practice : making sense of citizen-generated content. *International Journal of Digital Earth*, *5*(5), 398–416.

Das, R. C., & Alam, T. (2014). Location based emergency medical assistance system using OpenstreetMap. In *Informatics, Electronics & Vision (ICIEV), 2014 International Conference on* (pp. 1–5). IEEE.

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, *28*(4), 700–719. https://doi.org/10.1080/13658816.2013.867495

Fisher, P., Wenzhong, S., Goodchild, M., & Shi, W. (2002). *Spatial Data Quality*. Taylor & Francis.

Girres, J.-F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, *14*(4), 435–459.

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198–208.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Goodchild, M. F., & Hunter, G. J. (1997). A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, *11*(3), 299–306.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Government of Canada, S. C. (2017a, February 8). Focus on Geography Series, 2016 Census - Census subdivision of Gatineau, V (Quebec). Retrieved 13 January 2018, from http://www12.statcan.gc.ca/census-recensement/2016/as-sa/fogs-spg/Facts-csd-eng.cfm?LANG=Eng&GK=CSD&GC=2481017&TOPIC=10

Government of Canada, S. C. (2017b, February 8). Focus on Geography Series, 2016 Census - Census subdivision of Ottawa, CV (Ontario). Retrieved 13 January 2018, from http://www12.statcan.gc.ca/census-recensement/2016/as-sa/fogs-spg/Facts-csd-eng.cfm?LANG=eng&GK=CSD&GC=3506008

Graser, A., Straub, M., & Dragaschnig, M. (2013). Towards an Open Source Analysis Toolbox for Street Network Comparison: Indicators, Tools and Results of a Comparison of OSM and the Official Austrian Reference Graph. *Transactions in GIS*, *18*(4), 510–526. https://doi.org/10.1111/tgis.12061

Gröchenig, S., Brunauer, R., & Rehrl, K. (2014). Estimating Completeness of VGI Datasets by Analyzing Community Activity Over Time Periods. In *Connecting a Digital Europe Through Location and Place* (pp. 3–18). Springer, Cham. https://doi.org/10.1007/978-3-319-03611-3_1

Guinée, J. B. (2002). Handbook on life cycle assessment operational guide to the ISO standards. *The International Journal of Life Cycle Assessment*, *7*(5), 311.

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

Haklay, M., Antoniou, V., Basiouka, S., Soden, R., & Mooney, P. (2014). *Crowdsourced geographic information use in government*. World Bank Publications.

Haklay, M. (Muki), Basiouka, S., Antoniou, V., & Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, *47*(4), 315–322. https://doi.org/10.1179/000870410X12911304958827

Hecht, R., Kunze, C., & Hahmann, S. (2013). Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS International Journal of Geo-Information*, *2*(4), 1066. https://doi.org/10.3390/ijgi2041066

Helbich, M., Amelunxen, C., Neis, P., & Zipf, A. (2012). Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata. *Proceedings of GI_Forum*, 24–33.

Hoyle, D. (2001). ISO 9000: quality systems handbook.

Humanitarian OSM Team - OpenStreetMap Wiki. (2018). Retrieved 27 June 2018, from https://wiki.openstreetmap.org/wiki/Humanitarian_OSM_Team

Jiang, B., & Yao, X. (2006). Location-based services and GIS in perspective. *Computers, Environment and Urban Systems*, *30*(6), 712–725. https://doi.org/10.1016/j.compenvurbsys.2006.02.003

JOSM - OpenStreetMap Wiki. (2018). Retrieved 25 June 2018, from https://wiki.openstreetmap.org/wiki/JOSM

Keßsler, C., & De Groot, R. T. A. (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* (pp. 21–37). Springer.

Ketchen, D. J., & Shook, C. L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, *17*(6), 441–458.

Koeppel, I. (2000). What are location services?–from a GIS perspective. *Environmental Systems Research Institute (ERSI) White Paper*, *7*.

Koukoletsos, T., Haklay, M., & Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, *16*(4), 477–498. https://doi.org/10.1111/j.1467-9671.2012.01304.x

Kounadi, O. (2009). Assessing the quality of OpenStreetMap data. *Msc Geographical Information Science, University College of London Department of Civil, Environmental And Geomatic Engineering*.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (Eds.). (1999). *Geographical Information Systems, 2 Volume Set* (2 edition). New York: Wiley.

Mooney, P., & Corcoran, P. (2012). The annotation process in OpenStreetMap. *Transactions in GIS*, *16*(4), 561–579.

Nielson, J. (2006). Participation Inequality: The 90-9-1 Rule for Social Features. Retrieved 8 March 2018, from https://www.nngroup.com/articles/participation-inequality/

Poiani, T. H., dos Santos Rocha, R., Degrossi, L. C., & de Albuquerque, J. P. (2016). Potential of collaborative mapping for disaster relief: A case study of OpenStreetMap in the Nepal earthquake 2015. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on* (pp. 188–197). IEEE.

Ramm, F., Topf, J., & Chilton, S. (2011). *OpenStreetMap: using and enhancing the free map of the world*. UIT Cambridge Cambridge.

Raymond, E. S. (2001). *The Cathedral & the Bazaar : Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly Media.

Rehrl, K., & Gröchenig, S. (2016). A Framework for Data-Centric Analysis of Mapping Activity in the Context of Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, *5*(3), 37. https://doi.org/10.3390/ijgi5030037

Rutzinger, M., Rottensteiner, F., & Pfeifer, N. (2009). A Comparison of Evaluation Techniques for Building Extraction From Airborne Laser Scanning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *2*(1), 11–20. https://doi.org/10.1109/JSTARS.2009.2012488

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (Muki). (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, *31*(1), 139–167. https://doi.org/10.1080/13658816.2016.1189556

Statistics Canada. (2005). Census subdivision (CSD) - Census Dictionary. Retrieved 6 January 2018, from http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo012-eng.cfm

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(sup1), 234–240.

Ueberschlag, A. (2010). A first assessment of the OpenStreetMap quality in Switzerland. *Unpublished Manuscript, EPFL, Lausanne, Switzerland*.

Valli, C., & Hannay, P. (2010). Geotagging Where Cyberspace Comes to Your Place. In *Security and Management* (pp. 627–632).

van Oort, P. A. (2006). *Spatial data quality: from description to application*. Wageningen Universiteit.

Van Oort, P. A. J., & Bregt, A. K. (2005). Do Users Ignore Spatial Data Quality? A Decision-Theoretic Perspective. *Risk Analysis*, *25*(6), 1599–1610.

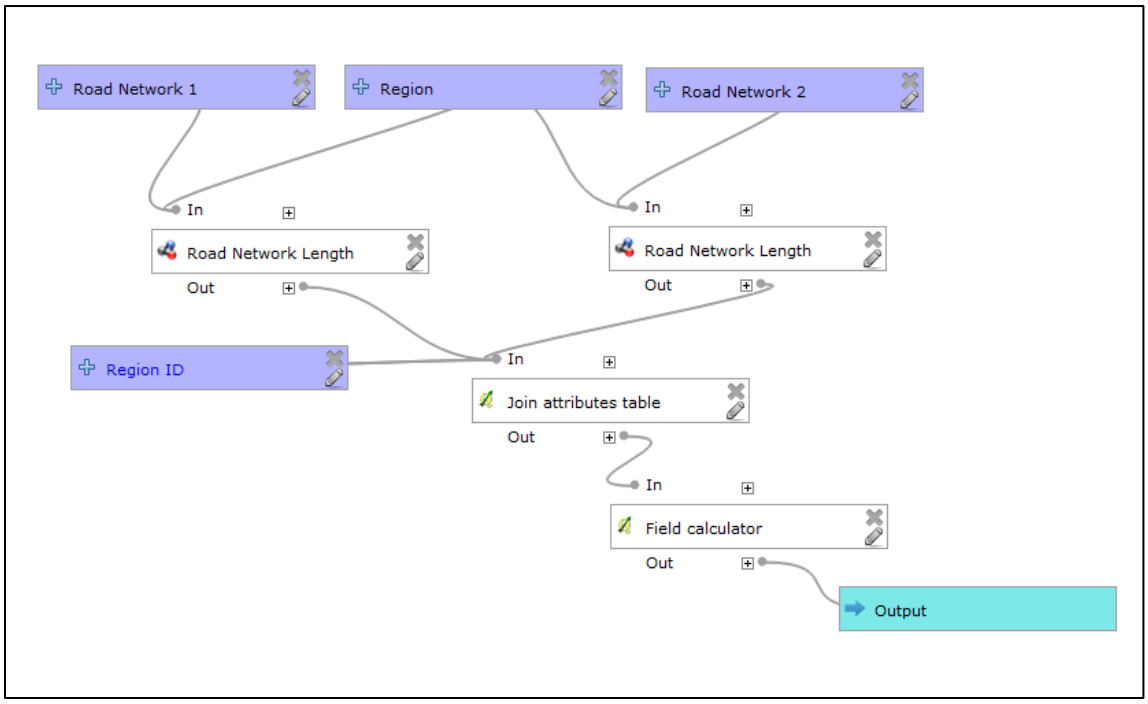
Vandecasteele, A., & Devillers, R. (2013). Improving volunteered geographic data quality using semantic similarity measurements. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *1*(1), 143–148.

WikiProject Canada - OpenStreetMap Wiki. (2018). Retrieved 12 June 2018, from https://wiki.openstreetmap.org/wiki/WikiProject_Canada

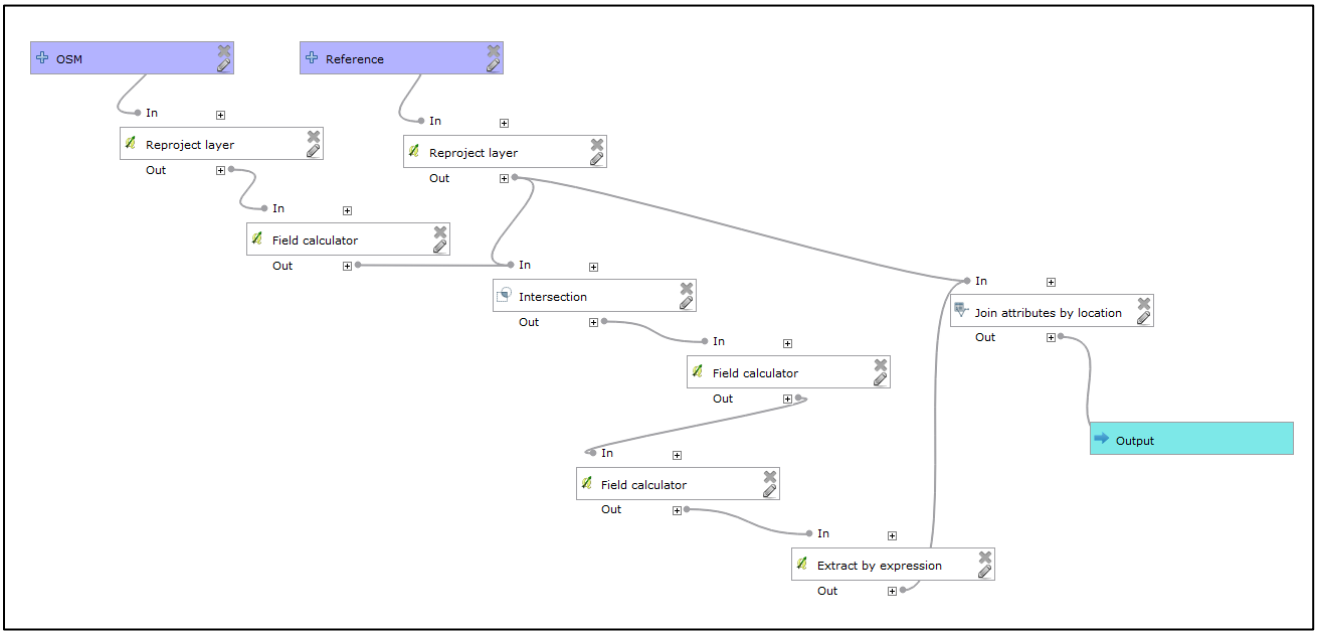
Will, J. (2014). Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network: a case study in Göteborg, Sweden. *Student Thesis Series INES*.

Zielstra, D., & Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany.

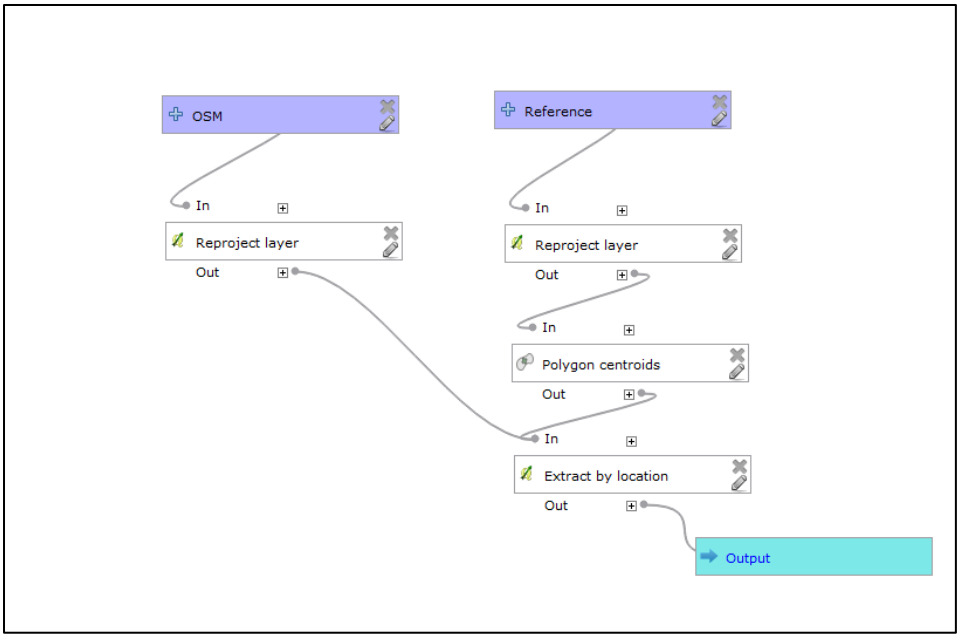
ANNEX A: METHODOLOGY



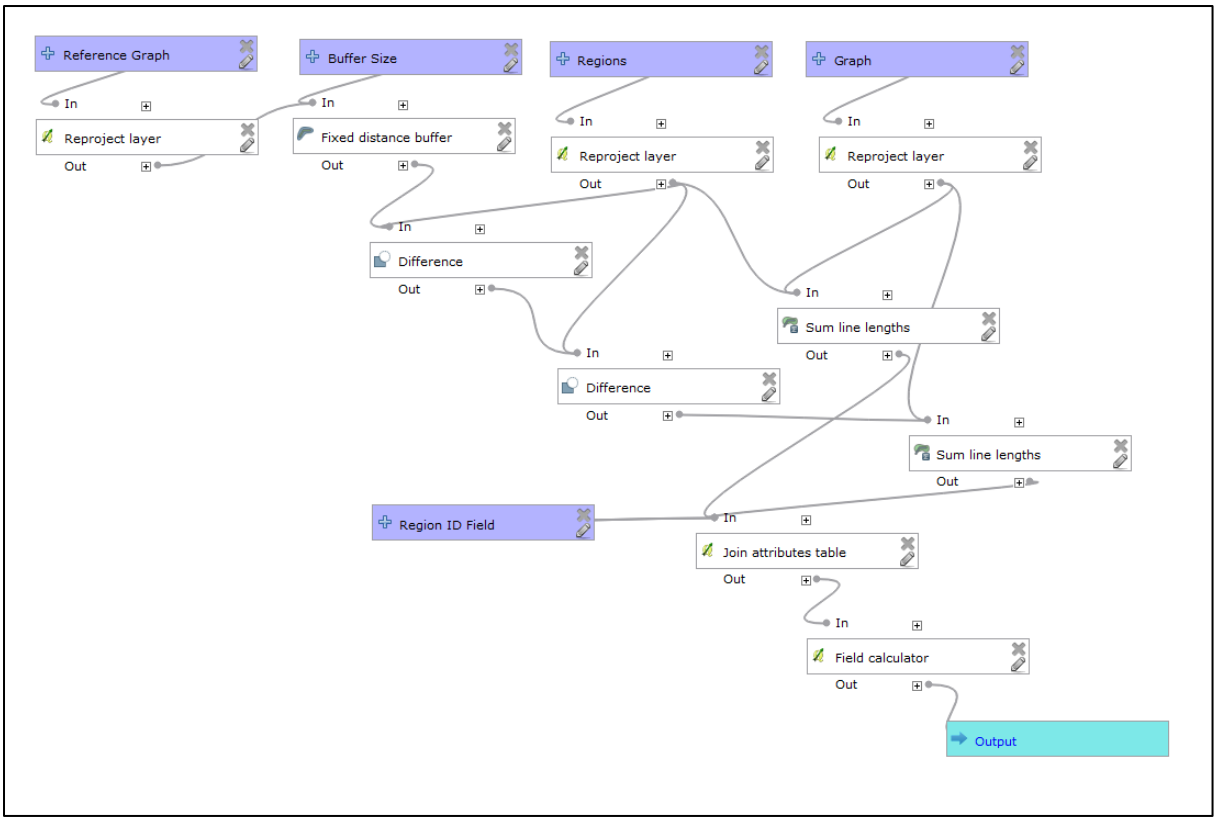
Appendix 1: Model to compare road network length amongst two datasets.



Appendix 2: Model to calculate overlap proportion between intersecting OSM and reference building footprints.



Appendix 3: Model to calculate centroid proportion reference building footprints and overlapping OSM buildings.



Appendix 4: Model to calculate overlap proportion between two linear features.

```
# import the geocoding services you'd like to try
# source credit: https://gist.github.com/rgdonohue/c4beedd3ca47d29aef01

from geopy.geocoders import Nominatim
import csv, sys
print ('Creating geocoding objects...')
nominatim = Nominatim(timeout=100)
# choose and order your preference for geocoders here
geocoders = [nominatim]
def geocode(address):
    i = 0
    try:
        while i < len(geocoders):
            # try to geocode using a service
            location = geocoders[i].geocode(address)

            # if it returns a location
            if location != None:

                # return those values
                return [location.latitude, location.longitude]
            else:
                # otherwise try the next one
                i += 1
    except:
        # catch whatever errors, likely timeout, and return null values
        print (sys.exc_info()[0])
        return ['null','null']

    # if all services have failed to geocode, return null values
    return ['null','null']
print ('Geocoding addresses!')
# list to hold all rows
dout = []
with open('input.csv', mode='r') as fin:
    reader = csv.reader(fin)
    j = 0
    for row in reader:
        print ('Processing #:',j)
        j+=1
        try:
            # configure this based upon your input CSV file
            street = row[0]
            city = row[1]
            province = row[2]
            address = street + ", " + city + ", " + province
            result = geocode(address)
            # add the lat/lon values to the row
            row.extend(result)
            # add the new row to master list
            dout.append(row)
        except:
            print ('Error')
print ('Writing the results to file')
# print results to file
with open('geocoded.csv', 'w') as fout:
    writer = csv.writer(fout)
    writer.writerows(dout)
print ('Done!')
```

Appendix 5: Batch geocoding script. Please visit my personal GitHub page for more thesis related source code. https://github.com/ktjaco/

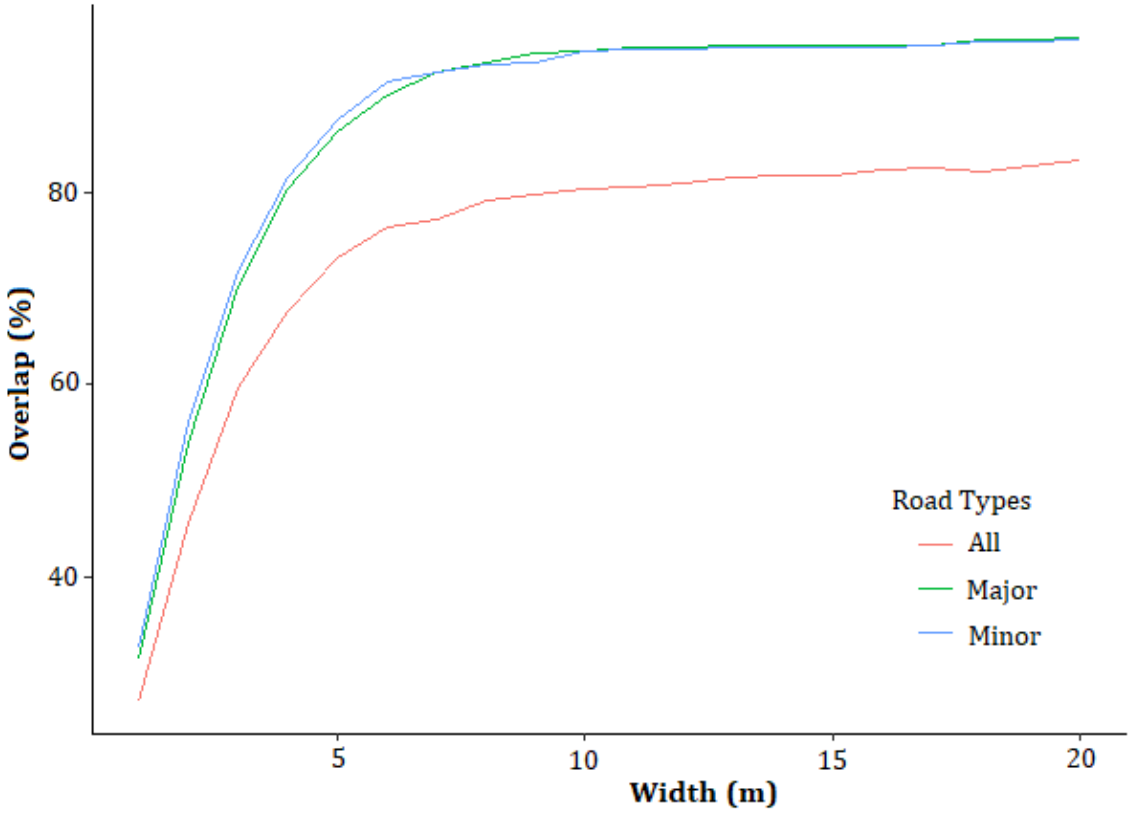
ANNEX B: RESULTS

All Roads		Major Roads		Minor Roads	
Buffer Width (m)	**Overlap Percentage (%)**	**Width**	**Percent**	**Width**	**Percent**
1	27.07	1	31.42	1	32.65
2	45.51	2	53.91	2	56.002
3	59.63	3	70.19	3	71.80
4	67.53	4	80.38	4	81.49
5	73.20	5	86.28	5	87.56
6	76.27	6	90.04	6	91.56
7	77.17	7	92.44	7	92.46
8	79.09	8	93.43	8	93.34
9	79.77	9	94.41	9	93.41
10	80.25	10	94.57	10	94.70
11	80.51	11	95.02	11	94.91
12	81.00	12	95.10	12	94.93
13	81.46	13	95.31	13	95.03
14	81.70	14	95.17	14	95.04
15	81.71	15	95.26	15	95.10
16	82.24	16	95.29	16	95.11
17	82.60	17	95.34	17	95.27
18	82.22	18	95.77	18	95.70
19	82.73	19	95.89	19	95.73
20	83.23	20	96.02	20	95.80
January 2017					
All Roads		**Major Roads**		**Minor Roads**	

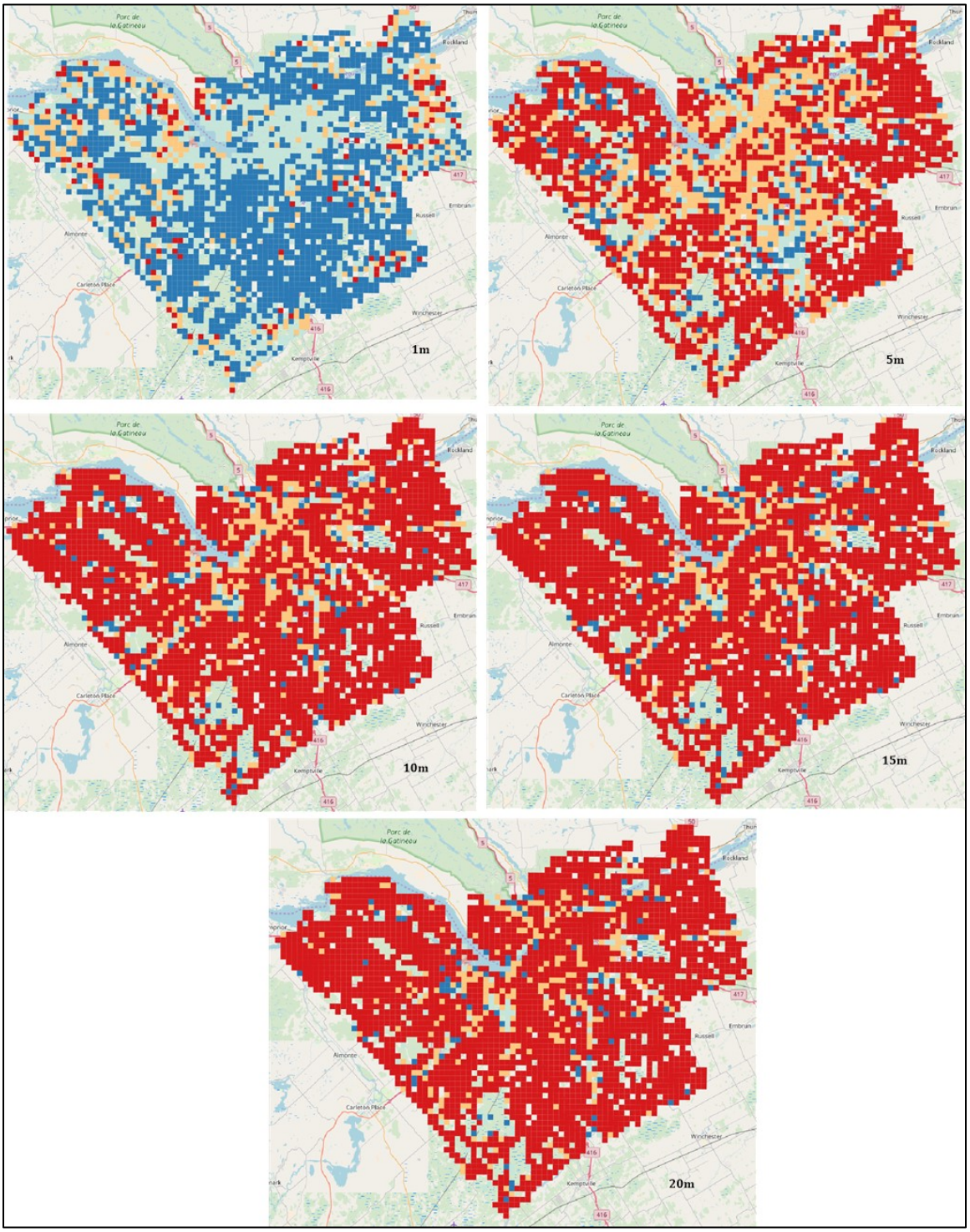
Buffer Width (m)	Overlap (%)	Buffer Width	Overlap (%)	Buffer Width	Overlap (%)
1	26.40	1	30.99	1	32.27
2	45.06	2	53.86	2	55.00
3	58.29	3	69.83	3	71.09
4	65.94	4	79.65	4	79.89
5	71.03	5	85.70	5	84.91
6	73.61	6	89.24	6	89.11
7	75.74	7	91.49	7	90.35
8	77.01	8	92.76	8	91.86
9	77.17	9	93.43	9	91.94
10	78.27	10	93.89	10	92.58
11	78.37	11	94.00	11	92.47
12	78.58	12	94.06	12	93.21
13	78.94	13	94.24	13	93.34
14	79.32	14	94.24	14	93.56
15	79.51	15	94.26	15	93.78
16	79.89	16	94.30	16	93.80
17	80.25	17	94.68	17	93.81
18	80.52	18	94.68	18	93.84
19	80.66	19	94.70	19	93.86
20	80.72	20	94.76	20	93.90
June 2017					
All Roads		**Major Roads**		**Minor Roads**	
Buffer Width	**Overlap (%)**	**Buffer Width**	**Overlap (%)**	**Buffer Width**	**Overlap (%)**
1	25.78	1	30.48	1	32.29
2	44.03	2	53.09	2	55.03

3	57.02	3	69.08	3	71.10
4	64.61	4	79.17	4	79.92
5	69.61	5	85.25	5	84.93
6	72.15	6	88.75	6	89.11
7	74.29	7	91.06	7	90.36
8	75.58	8	92.40	8	91.87
9	75.77	9	93.10	9	91.95
10	76.90	10	93.59	10	92.60
11	76.92	11	93.72	11	92.65
12	77.29	12	93.80	12	93.23
13	77.68	13	93.99	13	93.33
14	78.08	14	93.99	14	93.45
15	78.27	15	94.10	15	93.56
16	78.69	16	94.17	16	93.60
17	79.06	17	94.45	17	93.75
18	79.10	18	94.45	18	93.80
19	79.48	19	94.50	19	93.81
20	79.35	20	94.54	20	93.90

Appendix 6: Buffer analysis results for January 2016, 2017 and June 2017.

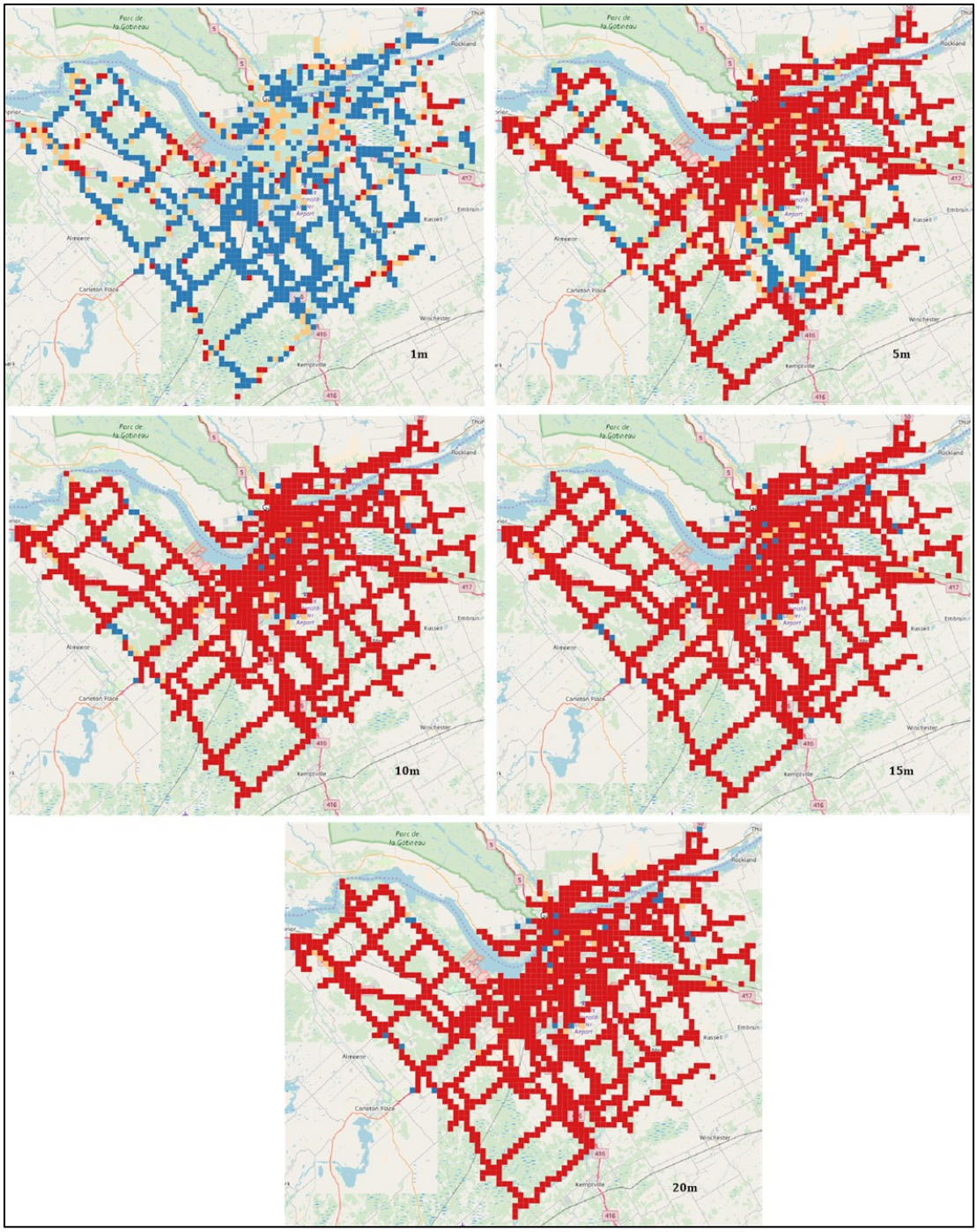


Appendix 7: January 2016 road overlap proportion (%) across 1 to 20 metre buffer widths.



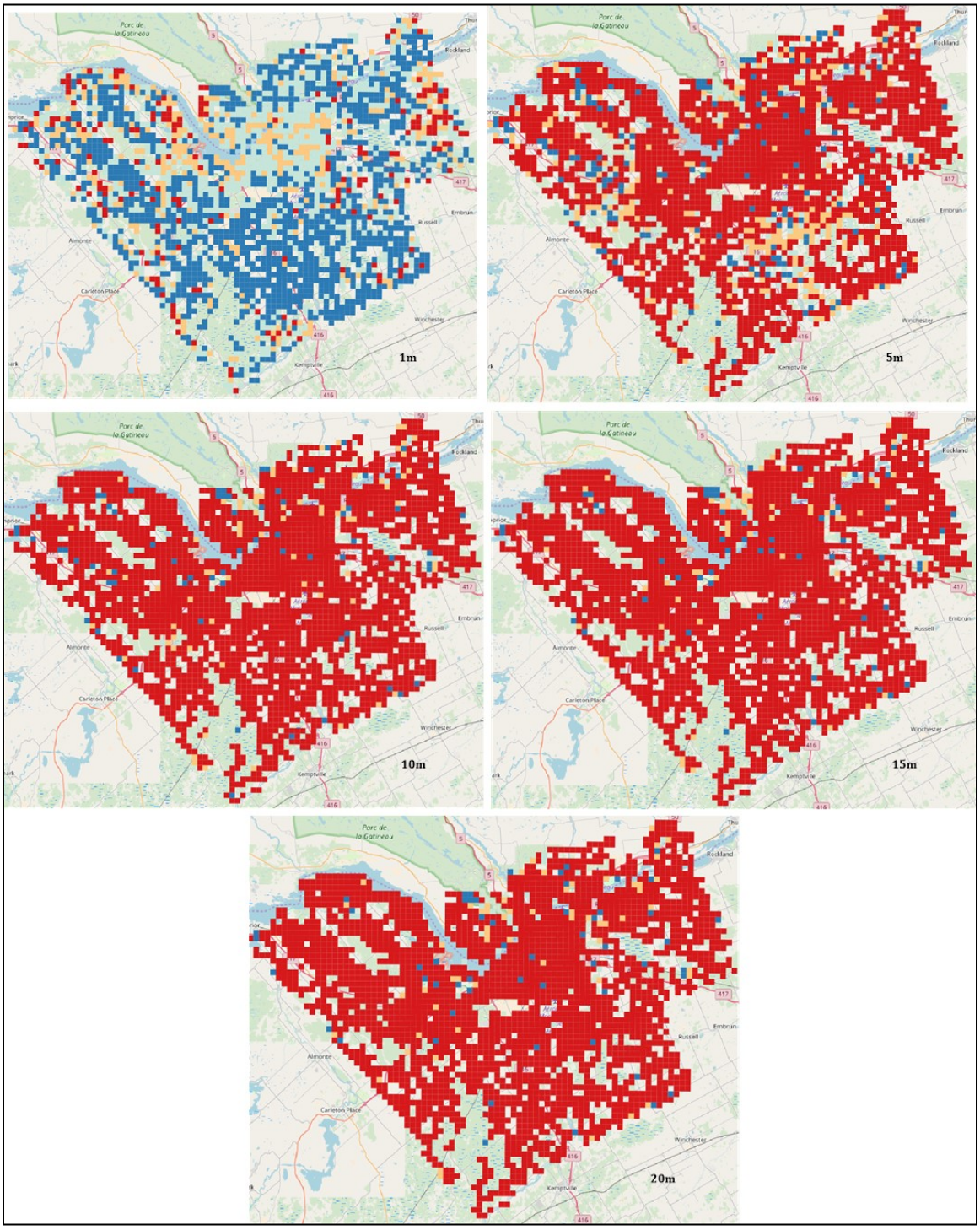
Dark Blue <25% overlap. Dark red >75% overlap.

Appendix 8: January 2016 OSM road network positional accuracy across 1 to 20m buffer widths.



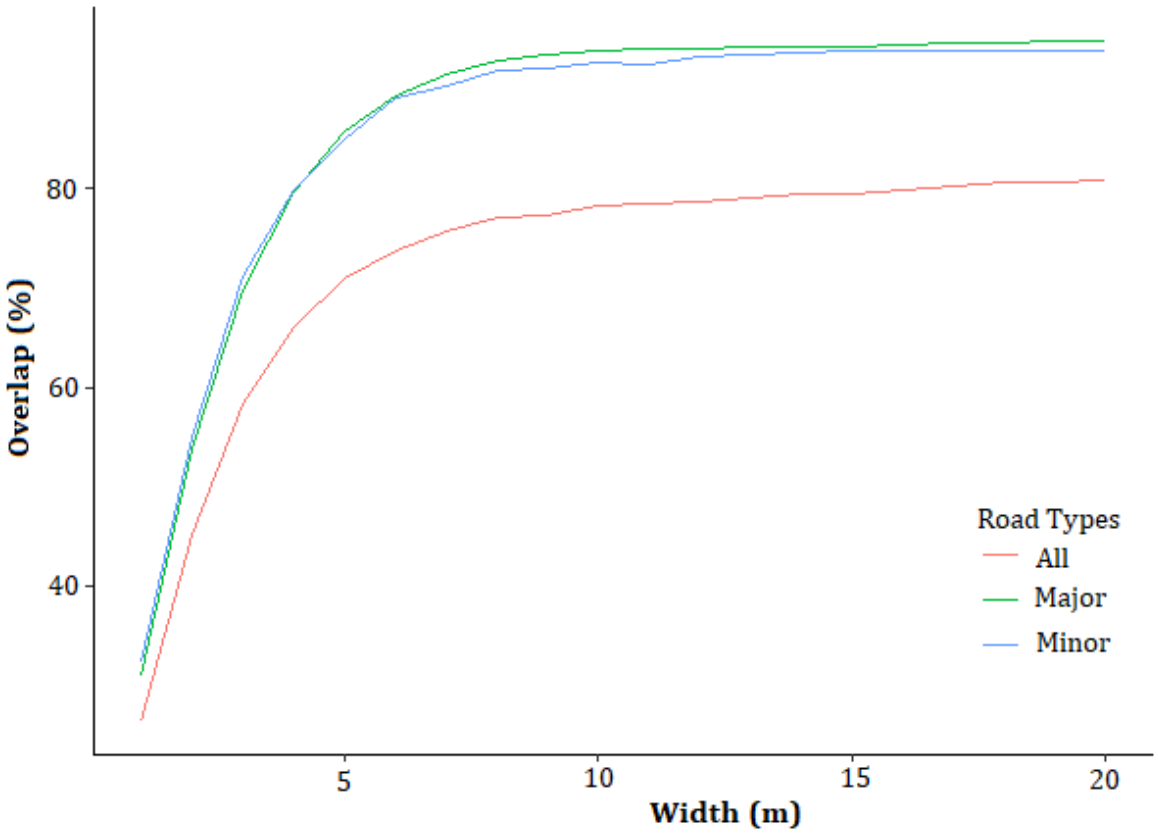
Dark Blue <25% overlap. Dark red >75% overlap.

Appendix 9: January 2016 major OSM road network positional accuracy across 1 to 20m buffer widths.

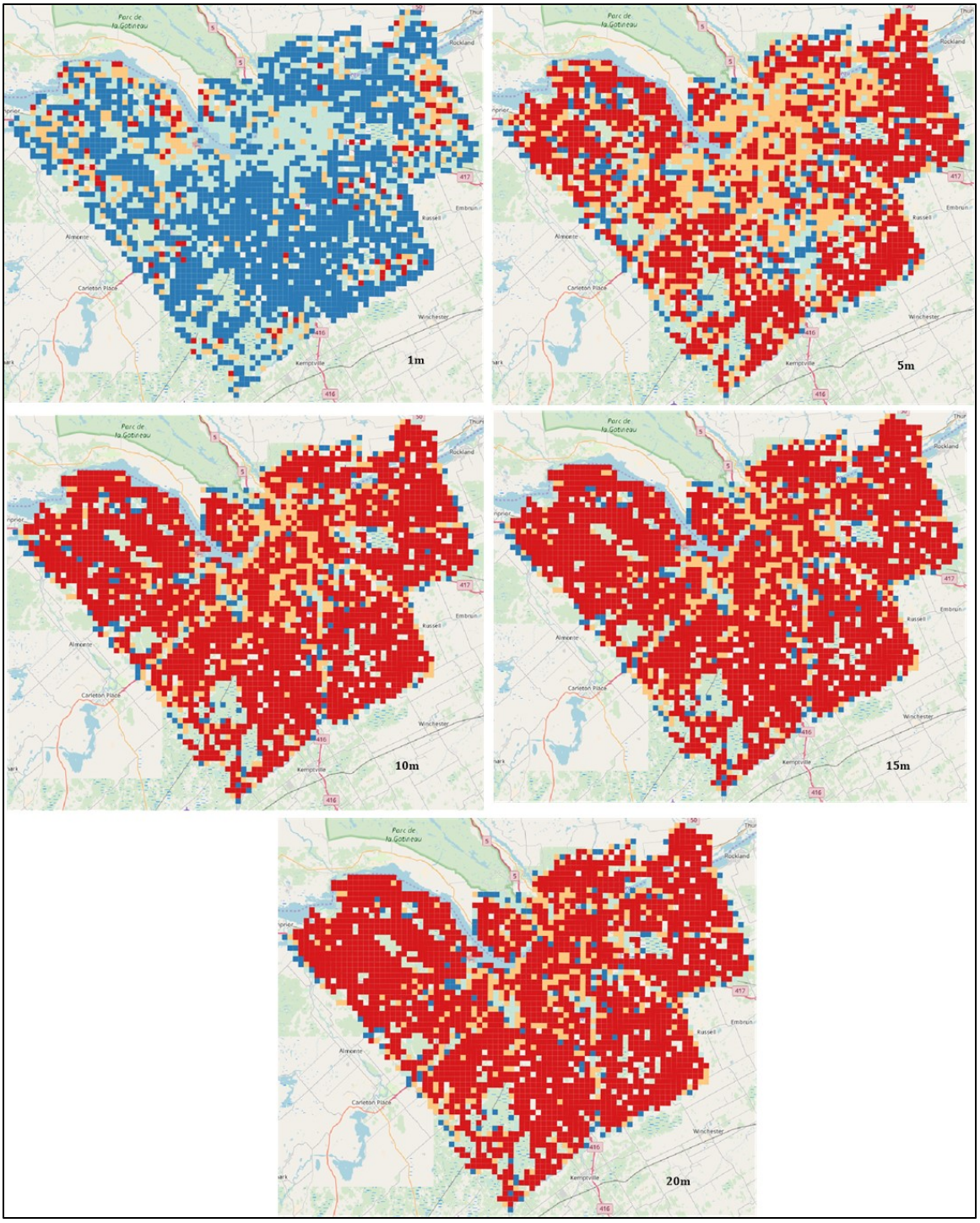


Dark Blue <25% overlap. Dark red >75% overlap.

Appendix 10: January 2016 minor OSM road network positional accuracy across 1 to 20m buffer widths.

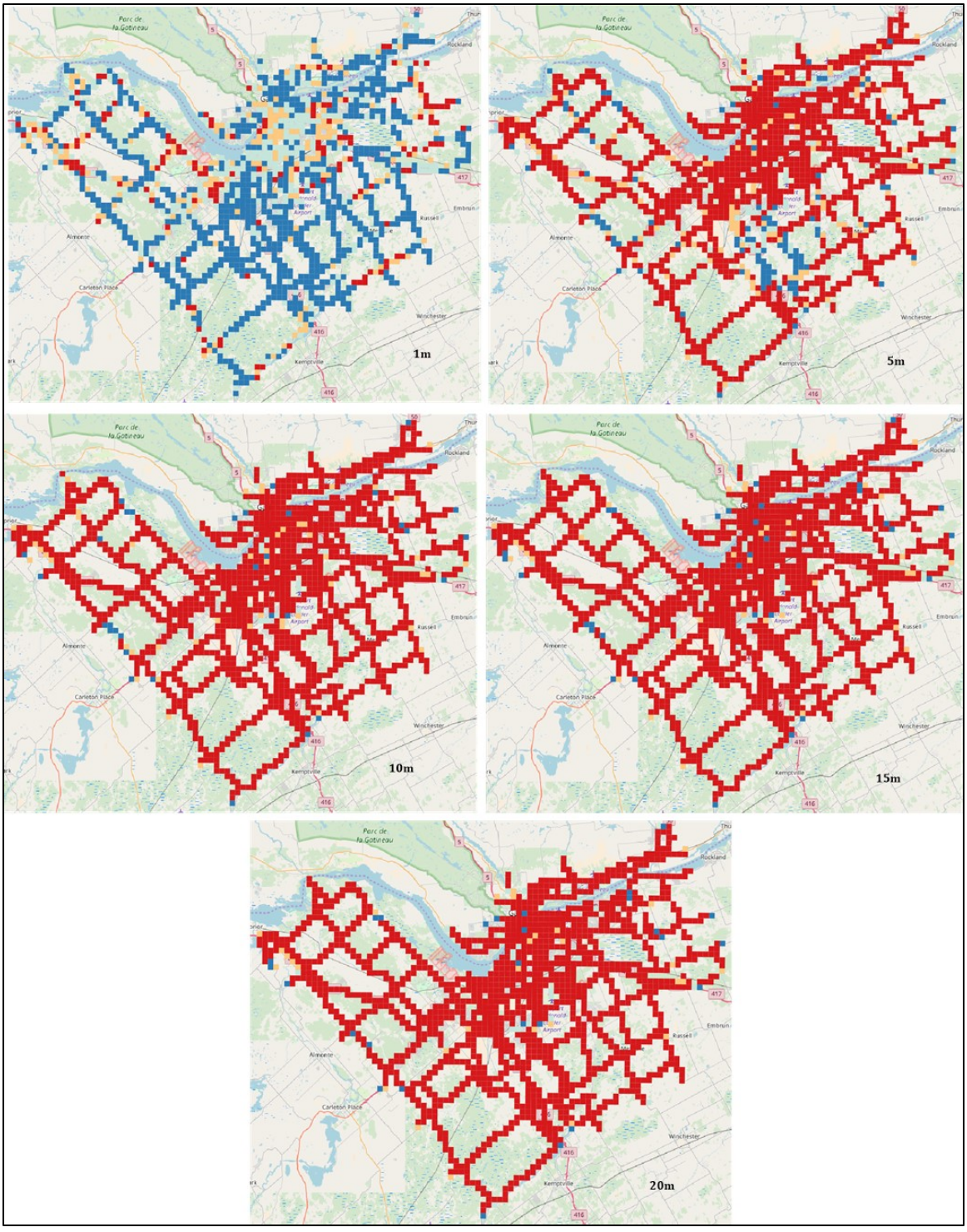


Appendix 11: January 2017 road overlap proportion (%) between road network types across 1 to 20 metre buffer widths.



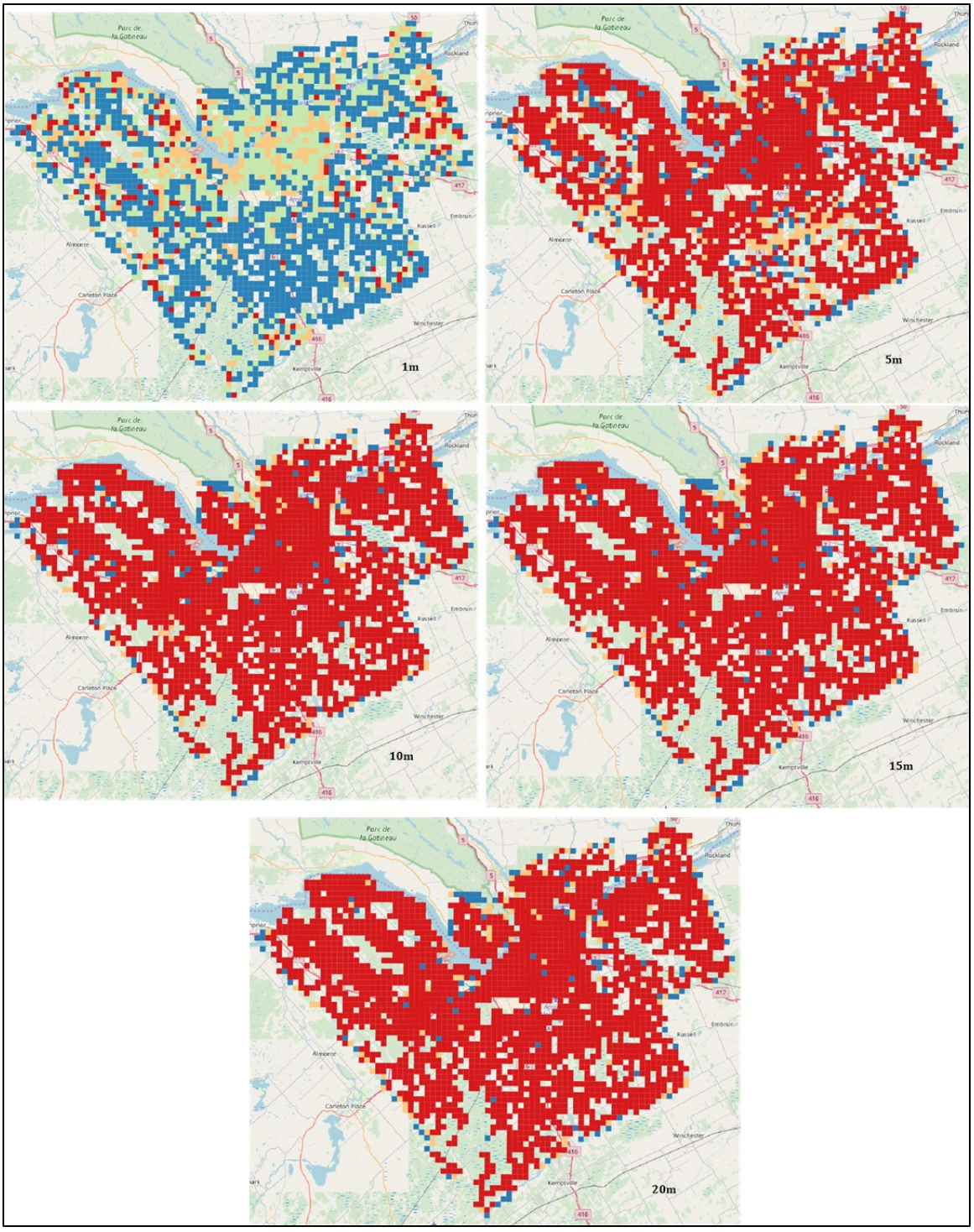
Dark Blue <25% overlap. Dark red >75% overlap.

Appendix 12: January 2017 OSM road network positional accuracy across 1 to 20m buffer widths.



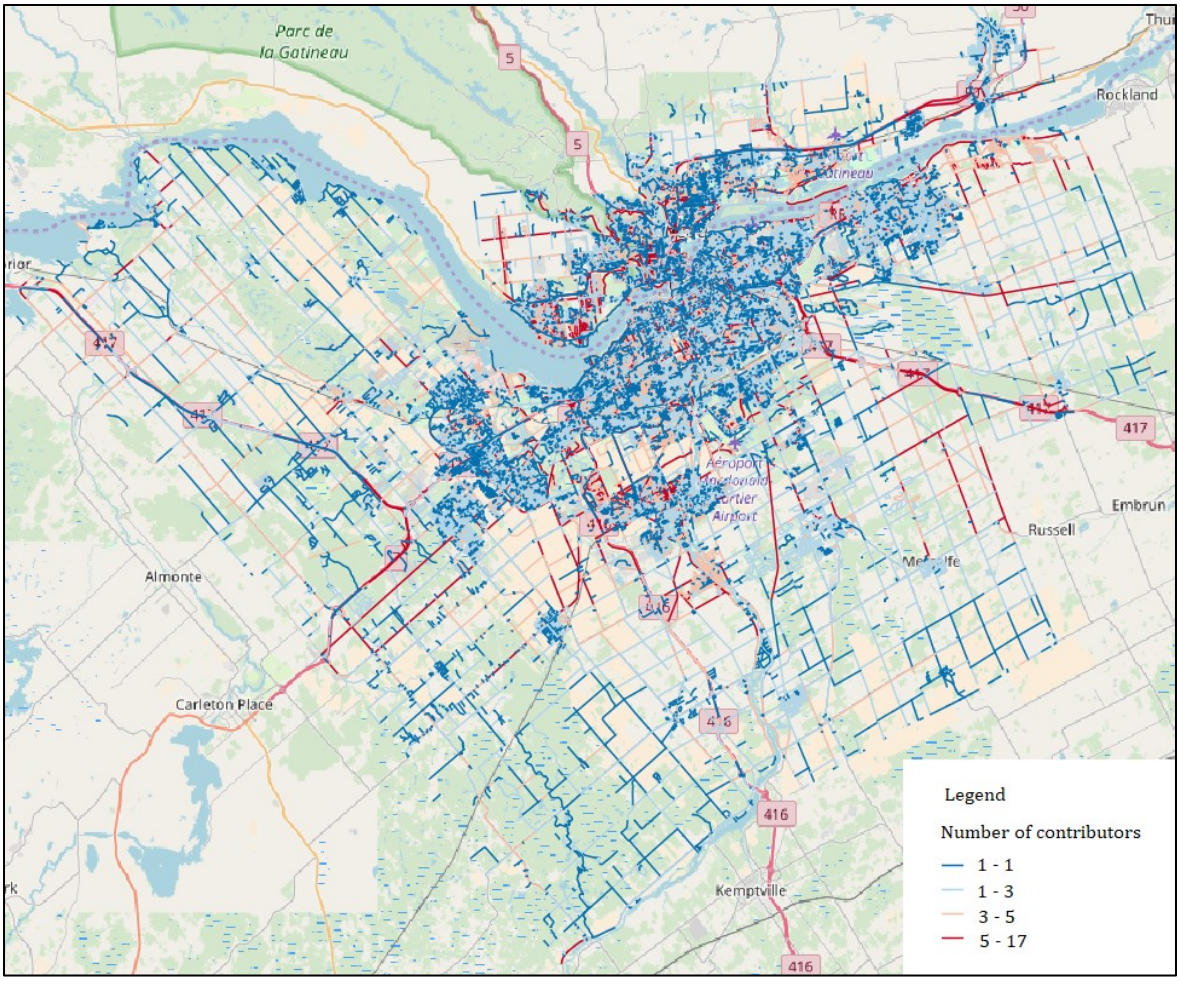
Dark Blue <25% overlap. Dark red >75% overlap.

Appendix 13: January 2017 major OSM road network positional accuracy across 1m, 5m, 10m, 15m, and 20m buffer widths.

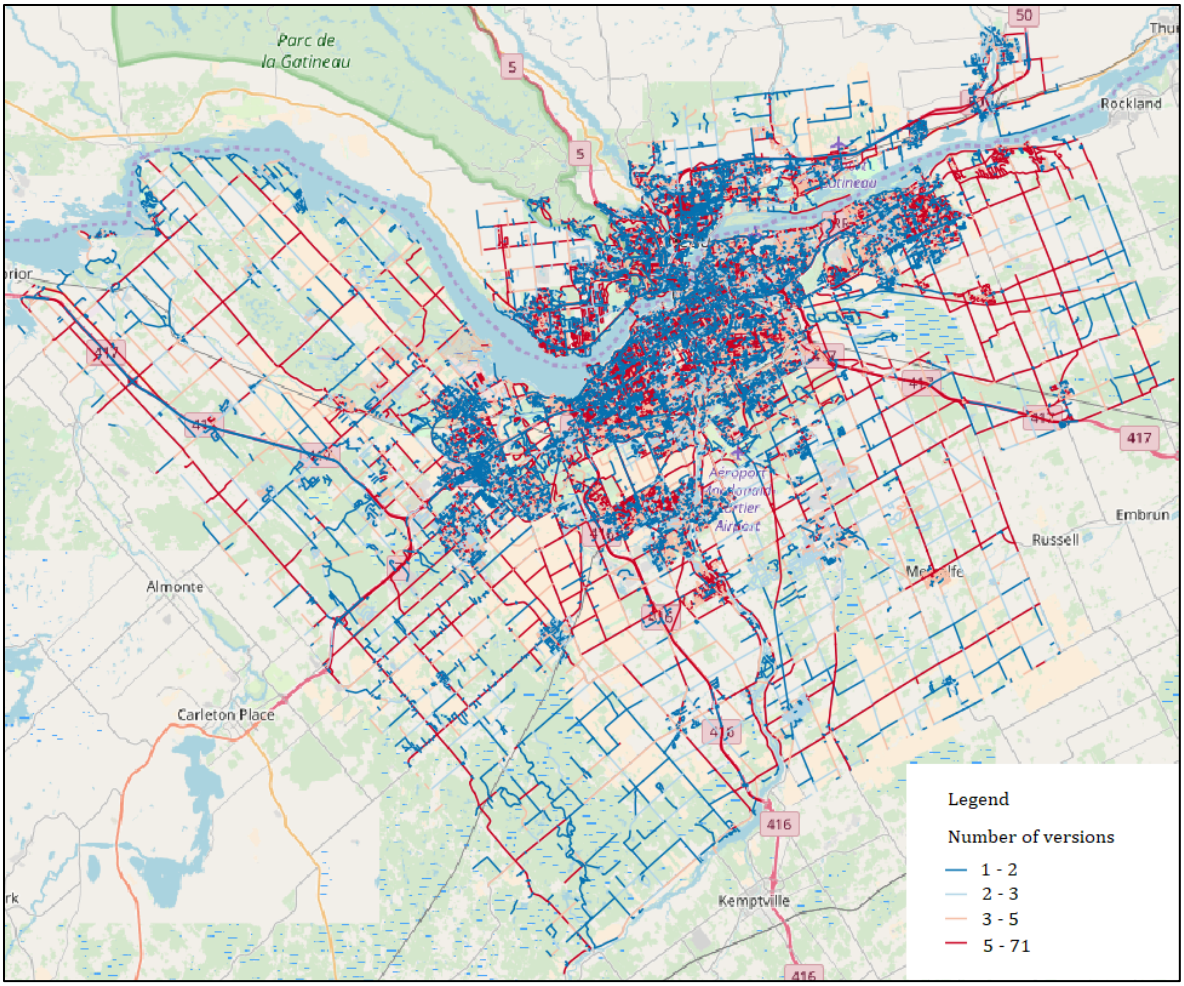


Dark Blue <25% overlap. Dark red >75% overlap.

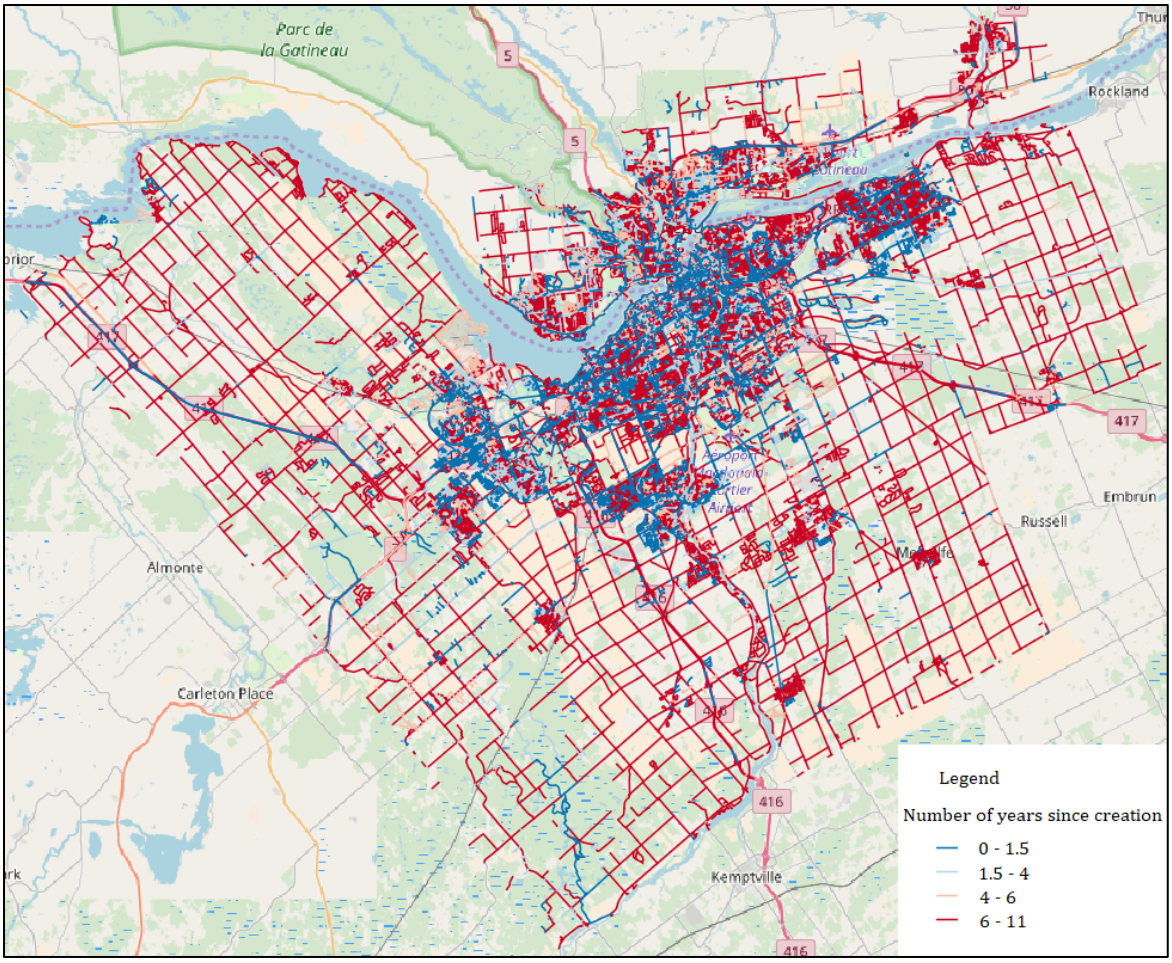
Appendix 14: January 2016 minor OSM road network positional accuracy across 1 to 20m buffer widths.



Appendix 15: Number of active contributors per OSM road segment in Ottawa-Gatineau.



Appendix 16: Number of versions per OSM road segment in Ottawa-Gatineau.



Appendix 17: Years since creation of OSM road segments in Ottawa-Gatineau