Using Geographic Relevance (GR) to contextualize structured and unstructured spatial data

by

Majuratan Sadagopan

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Masters of Science

in

Geography

Waterloo, Ontario, Canada, 2019

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Geographic relevance is a concept that has been used to improve spatial information retrieval on mobile devices, but the idea of geographic relevance has several potential applications outside of mobile computing. Geographic relevance is used measure how related two spatial entities are using a set of criteria such as distance between features, the semantic similarity of feature names or clustering pattern of features. This thesis examines the use of geographic relevance to organize and filter web based spatial data such as framework data from open data portals and unstructured volunteer geographic information generated from social media or map-based surveys. There are many new users and producers of geographic information and it is unclear to new users which data sets they should use to solve a given problem. Governments and organizations also have access to a growing volume of volunteer geographic information but current models for matching citizen generated information to locations of concern to support filtering and reporting are inadequate. For both problems, there is an opportunity to develop semi-automated solutions using geographic relevance metrics such as topicality, spatial proximity, cluster and co-location. In this thesis, two geographic relevance models were developed using Python and PostgreSQL to measure relevance and identify relationships between structured framework data and unstructured VGI in order to support data organization, retrieval and filtering. This idea was explored through two related case studies and prototype applications. The first study developed a prototype application to retrieve spatial data from open data portals using four geographic relevance criteria which included topicality, proximity, co-location and cluster co-location. The second study developed a prototype application that matches VGI data to authoritative framework data to dynamically summarize and organize unstructured VGI data. This thesis demonstrates two possible approaches for using GR metrics to evaluate spatial relevance between large data sets and individual features. This thesis evaluates the effectiveness of GR metrics for performing spatial relevance analysis and it demonstrates two potential use cases for GR.

Acknowledgments

This thesis has been an unconventional journey and I would like to thank my supervisor Dr. Rob Feick for his endless patience and support as I pursued my career and thesis simultaneously. I was given a lot of room to experiment in my thesis and I was fully supported when I decided to pursue a career opportunity through Mitacs that lead to a temporary break from the thesis. I would also like to thank the members of my defense committee, Dr. Peter Johnson and Dr. Su-Yin Tan, for their valuable insights and advice on improving this research.

I would like to thank the City of Kitchener and Josh Joseph for providing the data and case study information for this project. I acknowledge the GeoThink research project for supporting my research project and connecting me with many groups and scholars who have influenced my research and career. I would like to thank Dave Gerrard and the tram at GDA for supporting my research and career through the Mitacs program and later as a full-time employee.

I would like to thank my parents, brothers, grandparents and friends for their support in this process. A special thanks goes to Maria Cheng and Ratul Mehta who have been steadfast friends and unwavering supporters of my work over the years. Lastly, I would like to thank my current manager Xuejin Ruan and my team at Environics Analytics for supporting me as I balanced work at the firm with thesis work.

1 Table of Contents

List of Figures

List of Tables

List of Abbreviations

Geographic Information System **(GIS)**

Geographic Information **(GI)**

Volunteer Geographic Information **(VGI)**

Public Participation Geographic Information System **(PPGIS)**

Information Retrieval **(IR)**

Geographic Information Retrieval **(GIR)**

Geographic Relevance **(GR)**

Query footprint **(Q-footprint)**

Place footprint (P-footprint)

1 Introduction

The rate at which data is generated and collected is growing at an exponential rate. In 2017, IBM released a report stating that 90 percent of the data in the world had been created in the last two years (IBM Marketing Cloud, 2017). Research by Cisco projects that data storage capacity will quadruple between 2016 and 2021 (Cisco, 2018). Data generated by mobile phones, IoT devices, governments, private corporations, and individual citizens are contributing to the growing data trend (Centre for International Governance Innovation, 2018; Cisco, 2018). Big data has emerged as an area of study because our ability to collect, and sense data is outpacing our ability to process, store and manage collected data (Cisco, 2018). Furthermore, data from all dimensions of society are increasingly being linked to geographic locations; this includes public census databases, private customer databases, and community built open source databases (Purves, Clough, Jones, Hall, & Murdock, 2018). Large volumes of big data are spatial data (Burns & Thatcher, 2015) and researchers are increasingly finding that location data is an important and effective way to filter, categorize and organize data (Ivanova, Morales, de By, Beshe, & Gebresilassie, 2013; Purves et al., 2018; Spinsanti & Ostermann, 2013). As noted by Burns and Thatcher (2015), the ability to rapidly collect and analyze massive volumes of data has led to an increased focus on relationships between data and knowledge creation. There is a need to develop processes to manage and analyze web based spatial data in order to enable data driven decision making.

The literature indicates that growth in the web continues to drive growth in the creation and use of VGI data, but challenges associated with maintaining the integrity of VGI data remains an impediment to the use of VGI (Neis & Zielstra, 2014). Growth of web-based data such as VGI and open data has also driven research interest in the fitness of use problem due to the challenge of retrieving and organizing web-based spatial data (Neis & Zielstra, 2014; Wentz & Shimizu, 2018). Spatial data quality literature has discussed the use of contextual analysis to assure VGI data quality (Comber et al., 2016; Goodchild & Li, 2012). Researchers in geographic information retrieval and geographic relevance have developed sophisticated models for understanding the spatial context (Purves et al., 2018; Reichenbacher, De Sabbata, Purves, & Fabrikant, 2016). Geographic relevance researchers developed a multi-criteria model for evaluating spatial context that includes relevance measures such as topicality, proximity, directionality, and co-location (Reichenbacher et al., 2016). The metrics developed in GR have been primarily developed for mobile search, but the concepts developed in GR have similarities with concepts developed in the literature on fitness for use, assurance of VGI data quality and conflation (Goodchild & Li, 2012; McKenzie, Janowicz, & Adams, 2014; Wentz & Shimizu, 2018). There is an

opportunity to use GR metrics to evaluate spatial context and address challenges in spatial data quality literature, geographic information retrieval and VGI analysis. This thesis looks at using GR metrics to evaluate spatial relevance of open data in Chapter 3 and using GR metrics to match VGI data to framework data to support data analysis in Chapter 4.

Raper first introduced the idea of geographic relevance as a concept that links spatial data to a location-based query, Raper defines geographic relevance as follows.

> *'Geographic relevance (GR) of this kind can, therefore, be defined as "a relation between a geographic information need and "the spatio-temporal expression of the geographic information objects needed to satisfy it" in order to take some action'*(Raper, 2007, p. 846)

De Sabbata and Reichenbacher (2012) expand upon Raper's (2007) concept and define geographic relevance as a concept that improves spatial data retrieval by evaluating relationships of entities in the real world. As noted below, GR improves spatial data retrieval in databases by evaluating spatial relationships of geographic entities.

> "*GR refers to the relevance of a geographic entity, given a specific context of usage. That is, GR does not refer to the relevance of a geo-referenced document or a document reporting geographic information, it refers to the relevance of the real world entity or event by itself… GR is intended to assess the relevance of an object, that is a representation of a geographic entity within a computer system or database."* (De Sabbata & Reichenbacher, 2012, p. 1496)

The value of geographic relevance can be compared to the value of the PageRank algorithm powering the Google search engine. The PageRank algorithm made it possible for computers to understand which web sites are important to a user based upon a query (Page, Brin, Motwani, & Winograd, 1999). Research in geographic relevance can potentially make it possible for computers to understand what spatial data are relevant to a user. Google significantly improved public accessibility to the web using relevance metrics and research in geographic relevance has the potential to significantly improve public accessibility to location-based information.

Research interest in geographic relevance has been primarily driven by the growing importance of mobile phones, and the importance of spatial context in mobile search (Raper, 2007). The field of geographic relevance has borrowed extensively from research in geographic information retrieval which focused on location-based data retrieval on the web (Acheson, Wartmann, & Purves, 2018). The fields of GR and GIR placed a significant amount of effort into understanding spatial context and handling

ambiguity (Bordogna, Ghisalberti, & Psaila, 2012; Reichenbacher et al., 2016). GIS researchers studying diverse topics such as geotagged social media or spatial data quality have increasingly recognized the value of contextual analysis for analyzing geographic information from web-based data sources (Goodchild & Li, 2012; McKenzie et al., 2014; Spinsanti & Ostermann, 2013). Past research on spatial relevance and spatial data has tended to focus on two main characteristics of spatial data which include semantic attributes and spatial attributes; some studies have also analyzed topology (Adams, Li, Raubal, & Goodchild, 2007; McKenzie et al., 2014; Spinsanti & Ostermann, 2013). GR literature builds on existing metrics and proposes the use of new spatial relevance metrics that consider unique characteristics of spatial data sets, such as the distribution of features as well as the spatial context of the user (De Sabbata & Reichenbacher, 2012). GR researchers use mobile sensor information such as the direction of travel or velocity of travel to better understand the context of a user (Reichenbacher et al., 2016). There is an opportunity to expand the use of GR to address challenges such as spatial data quality in crowdsourced data or data retrieval in spatial databases.

This thesis explores the use of GR metrics to identify links between spatial data sets to filter, organize, and retrieve structured and unstructured spatial data such as framework data or VGI data. The research on GR metrics is intended to address two broad problems in GIS today. The first problem is improving accessibility to the growing volume of open spatial data on the web. Governments are increasingly making spatial data available to the public via open data portals. As a result, more non-expert users are using spatial data to address a diverse set of problems. However, retrieving and aggregating spatial data can be challenging for experts and non-experts alike (Ivanova et al., 2013). Several studies have attempted to develop models to evaluate context and relevance to improve spatial data retrieval (Ivanova et al., 2013; Jonietz, Zipf, Jonietz, & Zipf, 2016; Wentz & Shimizu, 2018).

The second problem is improving governments capability to use citizen-generated spatial data collected via the web. Due to the growing number of map-based web applications, citizens can author spatial data that pinpoints and describes locations of concern. The VGI data created by citizens are rich in information, but it is unstructured and difficult to analyze without manual interpretation. Several studies have explored the use of contextual analysis to summarize or enrich VGI data (Goodchild & Li, 2012; Spinsanti & Ostermann, 2013). Both of these research problems call for new models to be developed to better evaluate spatial relevance based upon context. This thesis explores the use of GR metrics to address these problems through two case studies.

1.1 Research Objectives

The concept of geographic relevance can be used improve fundamental data management tasks in GIS such as geographic information retrieval and spatial filtering of data. These basic tasks have become increasingly complex as GIS has moved to web and mobile environments. Modern GIS need to handle a wide array of spatial data sets that are generated at different scales and velocities by different users and organization with different perspectives (Neis & Zielstra, 2014; Noskov & Zipf, 2019; Rabari & Storper, 2015). The central research question of this thesis is to determine whether a generic set of geographic relevance criteria can be used to measure relevance between structured and unstructured spatial data sets and features. This research question is explored over three chapters which address the following research objectives.

1. Review literature on geographic relevance and identify relevant methods of relevance analysis in light of widespread authoring and use of spatial data by non-specialists
2. Identify a set of GR metrics that can be used to generate spatial relevance scores between large spatial data sets in an enterprise database for the purpose of helping inexperienced users find and access spatial data to support research and analysis
3. Use the identified GR metrics from the second objective to generate spatial relevance scores between unstructured VGI features and structured framework features.
4. Develop prototype software applications to evaluate the effectiveness of the proposed GR metrics and evaluate if GR can be defined using a generic set of criteria.

These objectives will address the central research question by developing two different GR models for two different case studies using a common set of GR criteria. This thesis will discuss how well the identified GR criteria translate to different study context and scales. The objective results should indicate if it is possible to identify a constant set of GR criteria that are applicable to general spatial relevance analysis.

1.2 Organization of Thesis

This thesis is compromised of 5 chapters. Chapter 1 is the introduction, Chapter 2 will review literature that is pertinent to the first research objective, Chapter 3 presents a case study on evaluating spatial relevance between framework data, Chapter 4 presents a case study on measuring geographic relevance between VGI features and spatial enterprise data, and Chapter 5 provides future direction for further research on this topic. This thesis will conduct a broad literature review on VGI, Open Data, GIR,

GR, and several related fields of study in Chapter 2. The literature review informs the development of two studies on the use of GR to address spatial data quality challenges associated with open data and VGI. The literature review has three objectives which are to; define characteristics of VGI, identify gaps in VGI data quality literature, and develop an understanding of geographic relevance. Review of VGI literature offers insight about the nature and characteristics of spatial data that is created on the web by amateur geographers. Review of Spatial Data Quality literature presents insights into the data quality challenges associated with web-based data such as VGI. It provides examples of studies that have used contextual spatial analysis to improve VGI data quality. Review of GR literature provides an overview of the tools and methodologies that researchers have used to automatically evaluate context and determine relevance between spatial data. Chapter 2 provides direction for the development of the case studies presented in Chapter 3 and 4.

Chapter 3 looks at the use of geographic relevance to measure relevance between framework data to retrieve spatial data that is relevant to a domain and location of study. The solution developed in this Chapter is directed at the fitness for use problem in SDQ literature, and it aims to improve data accessibility on open data portals. The Chapter is compromised of a literature review, a methodology, and a results section. Chapter 3 identifies and defines GR metrics and presents an evaluation model to create GR scores between spatial data. The Chapter also presents a prototype data processing application named UrbanData that was used to analyze relevance between data sets from a regional open data portal. The Chapter presents a set of analysis results that evaluate the effectiveness of GR metrics for evaluating the relevance of spatial data for a given domain and study context.

Chapter 4 looks at the use of geographic relevance to spatially contextualize volunteered geographic information. The methods developed in this Chapter are intended to contribute to VGI data quality literature and GR literature. This Chapter identifies GR metrics and presents an evaluation model to identify many to one relationship between unstructured VGI data and structured framework data. A prototype data processing application named UrbanContext and a data visualization application named UrbanContext Viz are also presented in this Chapter to demonstrate how GR metrics can be implemented in software. A case study is presented in this Chapter that demonstrates how the prototype GR application can be used to analyze citizen feedback in the planning process. The analysis results of this Chapter present insights into the effectiveness of the identified GR metrics. Chapter 5 discusses the results of the two case studies and summarizes the progress made on the thesis research objectives. This Chapter also identifies study limitations and provides future direction for research.

2 Literature Review and Research Framework

This thesis is motivated by spatial data quality challenges in volunteered geographic information (VGI) which is an emerging source of spatial data generated by web-based communities. In a 2012 paper, Goodchild and Li discussed the need for a context-based approach for data validation of volunteered geographic information (Goodchild & Li, 2012). Research on geographic relevance has significant potential to address this issue, but no studies exist on the application of GR on VGI data. The spatial data quality of VGI data is of increasing importance because the use of volunteered geographic information is increasing. Companies such as MapBox and Apple are selling navigation technologies on top of VGI databases like OpenStreetMap and researchers are exploring approaches for using geotagged social media to detect forest fires (Spinsanti & Ostermann, 2013). This literature review explores the idea of VGI contextualization in three sections that discuss Volunteered Geographic Information, Spatial Data Quality, and Geographic Relevance. Volunteered geographic information (VGI) is a spatial data trend that has been facilitated by the web. This section of the literature review looks at the concept of VGI and the nature of the data produced on the web. The second topic covered in the literature review is spatial data quality which includes a large corpus of work on how to compare spatial features and identify spatial relationships between VGI data. The last section of the literature review looks at geographic relevance which is a new field of research that has emerged from work in geographic information retrieval. This section explores new models for analyzing relevance and context.

The three topics of research discussed in this literature review provide a broad perspective of emerging trends in GIS such as the growing volume of spatial data on the web, the growing challenge of ensuring spatial data quality and the innovations that have been made to address these challenges. The three topics covered in this literature have significant differences. Volunteered Geographic Information is a social and technological phenomenon that has attracted interest from researchers in social sciences, urban planning, and GIS (Goodchild & Li, 2012; Seeger, 2008). Research in spatial data quality is primarily technical, with an emphasis on the development of methodologies to compare and validate spatial data. Research in geographic relevance is primarily focused on understanding user information needs and retrieving spatial data to meet user information needs (Raper, 2007; Reichenbacher et al., 2016). The three topics in this literature review are related because all three research fields explore challenges associated with handling heterogeneous spatial data and ambiguity. VGI literature looks at the production of heterogeneous spatial data on the web. Literature on VGI data quality looks at approaches to reduce the heterogeneity of VGI data through standards, validation models, and

conflation systems. Geographic relevance literature looks at approaches to retrieve spatial data that fits users' needs based upon ambiguous criteria.

The subsequent sections of the literature are organized as follows. The section on volunteered geographic information reviews the literature on VGI and related concepts such as Neogeography, PPGIS, and VGI. The section on spatial data quality discusses research on validating VGI data using methods like spatial data matching, conflation, and contextualization. The section on geographic relevance will review research on geographic information retrieval, geographic relevance and related methodologies that are used to measure relevance. This literature review discusses the concepts and methodologies that motivate this thesis and inform the methodologies developed in the case studies discussed in subsequent chapters.

2.1 Volunteered geographic information

Volunteer Geographic Information is a research concept that was developed as a result of advances in web technologies, the development of Web 2.0 made it possible to create and consume content on the web at a scale that was not possible before (Jonietz, Antonio, See, & Zipf, 2017). The concept of volunteer geographic information (VGI) was coined by Goodchild to describe the growing body of amateur geographers that are generating spatial data online using platforms such as OpenStreetMap (Goodchild, 2007a). VGI is important because it provides the GIS community with an alternative or supplement to the use of traditional authoritative data, which is significantly more expensive to acquire (Jonietz et al., 2017). Platforms such as OpenStreetMap makes it possible to access spatial data sets with global coverage at no cost. In contrast, comparable data sets from providers such as TomTom, ESRI, or Here maps are prohibitively expensive. In addition to making valuable spatial data available at a low cost, VGI has unique characteristics that make it effective for tasks like disaster response. Unlike traditional authoritative data that is released in quarterly cycles, VGI can be responsive, time-sensitive, and geosocial (Rob Feick & Roche, 2013). The field of VGI has evolved over time as researchers have strived to define VGI, improve the quality of VGI data, and apply VGI data to unique research problems.

The concept of VGI is related to other web 2.0 concepts such as Neogeography, crowdsourcing, citizen-science, and user-generated content (See et al., 2016). Several studies use these terms interchangeably (Kalvelage et al., 2018; Lin, 2018), others have written on the distinction between VGI and related terms (Goodchild, 2009; See et al., 2016). Review of the literature indicates several terms associated with VGI are distinct concepts that reflect differences in data collection methods, data

structure and data volume (Goodchild, 2007a; See et al., 2016; Seeger, 2008). These differences have significant implications on how data can be analyzed and used. Literature that discusses the definition of VGI tends to touch upon two main topics which include the characteristics of the people or sensors generating VGI data and the characteristics of the VGI data itself (Goodchild, 2009; Kalvelage et al., 2018; See et al., 2016). The social dimension of VGI includes research on Neogeography, collective intelligence, and public engagement. Discussions on VGI characteristics center on the typology and characteristics of VGI data as well as spatial data quality. The following sections review the literature on the social and technological trends related to VGI before reviewing the literature on the characteristics of VGI.

2.1.1 Neogeography, Geoweb and Public Engagement

Neogeography is a concept that was originally developed by Turner to describe trends in technology that make GIS and cartography tools accessible to users outside of the GIS industry (Turner, 2006). Goodchild described Neogeography as the proliferation of amateur geographers in the field of geography (Goodchild, 2009). For Neogeography describes the ability for the public to create and share location information using commonly accessible tools and applications such as Google Maps or mobile phones (Haklay, 2013; Turner, 2006). The concept of Neogeography is tightly linked to the concept of Web Mapping 2.0 because web 2.0 technologies made it possible for the public to access and create spatial data at little to no cost (Batty, Hudson-Smith, Milton, & Crooks, 2010; Haklay, Singleton, & Parker, 2008). Concepts such as crowdsourcing and citizen science are sometimes interchanged with Neogeography. These concepts overlap in the sense that they all rely on web technologies to collect or aggregate data using the concept of collective intelligence (Doan, Ramakrishnan, & Halevy, 2011) but Neogeography is unique because it focuses on the societal and technological trends related to the generation of open spatial data on the web (Connors, Lei, & Kelly, 2012). Neogeography has contributed to the emergence of VGI (Goodchild, 2009; Jonietz et al., 2017), the democratization of GIS (Haklay, 2013) and the growing use of citizen data in scientific research (Connors et al., 2012). Neogeography describes the set of web technologies and societal trends that enable the production of VGI; thus the concept of Neogeography is often presented alongside the concept of VGI in the literature (See et al., 2016). The web and mobile technologies that enabled the concept of Neogeography and drove growth in VGI have continued to evolve since the inception of the concept. Similarly, societal trends influencing Neogeography and VGI continue to evolve. Over time, advances in technology enable new forms of social interaction that drive new societal trends.

One of the biggest technology or software platforms that influenced web mapping, Neogeography and VGI was the development of Google Maps. The software application made it easy for anyone to share location data using maps with pushpins and it spurred the creation of numerous map mashup applications (Elwood, Goodchild, & Sui, 2012; Haklay et al., 2008). The development of Google Maps was followed by the development of other important mapping applications such as OpenStreetMap and Wikimapia which used web 2.0 technologies to build interactive web mapping applications driven by volunteers that created VGI (Goodchild, 2007a). Social media applications such as Flickr also played a role in the growth of VGI and Neogeography because they allowed users to store photos with location data (S. Gao, Li, Li, Janowicz, & Zhang, 2017; Haklay et al., 2008). The group of web mapping applications that were developed on top of web 2.0 technologies has been broadly referred to by researchers as the geoweb (Haklay et al., 2008; Jankowski, Czepkiewicz, Młodkowski, Zwoliński, & Wójcicki, 2019; Verplanke, McCall, Uberhuaga, Rambaldi, & Haklay, 2016). Over the years, the number of web-based and mobile location-based application has continued to grow and dimensions of the geoweb have changed. Google reduced support for some of its free web mapping services by shutting down Google Maps Engine. However, solutions such as ESRI's ArcGIS Online, CartoDB, and MapBox have developed solutions that fill the gap left by Google and provide numerous new mapping capabilities for developers. OpenStreetMap continues to remain a powerful platform for VGI data, but the platform is also increasingly supported by governments and corporations (OpenStreetMap, 2018) that donate data to the platform (DigitalGlobe, 2018) and fund the development of software tools for the platform (MapBox, 2019). Growth in social media has led to the development of several new location-based social media apps such as Snap and Instagram, while older platforms such as Twitter continue to make location data a part of the platform. Several companies such as ESRI and PingStreet are also building geoweb solutions for governments that allow staff to collect spatial data from citizens. Overall the number of geoweb applications is growing, and there are a broad set of VGI applications on the web that cover diverse topics such as ecology, public safety, feature mapping, gazetteers, hiking, search data and more (See et al., 2016). The number of location applications on the web continues to grow. As spatial data becomes more ubiquitous, questions arise about what motivates people to contribute spatial data and how the geoweb helps society.

The democratization of spatial data and GIS is a prime benefit of VGI and Neogeography that has been discussed in the literature (Haklay, 2013; Kleinhans, Ham, & Evans-Cowley, 2015). Neogeography tools and geoweb applications have made a large amount of diverse spatial data openly available on the web (Jonietz et al., 2017). VGI platforms such as OpenStreetMap rely on active communities to

contribute data (Goodchild, 2007a; Jonietz et al., 2017), and validate data (Goodchild & Li, 2012). In general, effective VGI platforms are developed to support social causes that attract participants interested in contributing to the greater good (See et al., 2016). However, some researchers have criticized VGI as a tool that is primarily used by the technological elite, and there are concerns that VGI is not a representative or equitable (Cinnamon & Schuurman, 2013; Haklay, 2013). It is difficult to determine if VGI platforms have a biased user base because many VGI platforms don't collect user or contributor data (See et al., 2016). As the web becomes more ubiquitous and digital literacy improves some of the concerns raised by Haklay (2013) will likely diminish. Literature shows that VGI tools have been successfully developed to address a diverse set of issues that include environmental monitoring, natural hazard analysis, land cover mapping, and more (Connors et al., 2012; Jankowski et al., 2019; Jonietz et al., 2017). In contrast, the potential of VGI has primarily been underwhelming when considering public engagement and the use of VGI in governance (Brown, 2012; Haklay, 2013).

The concept of public engagement has been of interest to the GIS community for a long time (Brown, 2012; Seeger, 2008; Tulloch, 2008), and the growth of the geoweb increased research interest in using GIS tools such as Neogeography and VGI to solve challenges around public engagement and shared governance (Brown & Raymond, 2014; Kalvelage et al., 2018). The term public engagement is often interchanged with the concept of public participation (Newfoundland Labrador Office of Public Engagement, 2013; Ross, Baldwin, & Carter, 2016). There is no clear consensus on how the two terms should be used, but the International Association of Public Participation uses the terms public engagement and public participation to refer to the process of engaging the public in government decision making (International Association of Public Participation, 2019). This thesis uses the definition of public participation that is used by IAP2. The concept of participation is rooted in the belief that people who are affected by a decision should be involved in the decision-making process (Planning Institute Australia, 2011). The process of public participation includes four key tasks that include informing the public about an issue, consulting the public for opinions, deliberating on possible solutions and co-creating a solution that reflects the values and needs of the community (Newfoundland Labrador Office of Public Engagement, 2013). Participation can vary from being a power-sharing process between governments and citizens to a less representative process of simply informing citizens on decisions (Schlossberg & Shuford, 2005). The field of public participation GIS (PPGIS) is a branch of GIS research that focuses on how the public can use geospatial technologies to share information with governments and participate in the decision-making process (Verplanke et al., 2016). Participatory GIS (PGIS) is another related term that describes the use of GIS tools to support participation, this thesis follows the

recommendation by Tulloch and considers PPGIS and PGIS as the same concept (Tulloch, 2008; Verplanke et al., 2016). The advance of the web provided researchers with new tools to address known challenges in PPGIS, such as collecting representative data and connecting with underrepresented communities (Cinderby, 2010; Johnson & Sieber, 2012). VGI has often been discussed in conjunction with PPGIS because it describes the process of capturing non-expert spatial information using the web (Brown & Kyttä, 2014). VGI and related concepts such as the geoweb reduce the cost of engaging and connecting with citizens, and it makes it possible to scale public participation programs (Jankowski et al., 2019). The geoweb can be used to facilitate bi-directional engagement between citizens and governments, and it provides an opportunity for governments to create a dialogue with citizens (Johnson & Sieber, 2012). Several researchers have specifically focused on the use of VGI to support participation in the planning process. Facilitated VGI (f-VGI) is a concept that has been defined to describe VGI data that is collected in a controlled environment to gather data on specific topics of study such as a new planning project (Kalvelage et al., 2018; Seeger, 2008). According to Kalvelage et al. (2018), the distinction between VGI and f-VGI is in the specificity of the collected data where VGI data is randomly generated at the discretion of participants while f-VGI data is created in response to questions raised by the researcher. The f-VGI concept is not widely used but researchers have discussed the relationship between PPGIS, PGIS and VGI (Verplanke et al., 2016). PPGIS processes tend to struggle with maintaining active user bases because integrating participant feedback into decision making remains a challenge (Brown, 2012; Brown & Kyttä, 2014). Furthermore, research is needed to understand how the broad public can be incentivized to contribute data to VGI platforms to support PPGIS processes (Verplanke et al., 2016). Overall, Neogeography, the geoweb, and public participation are important societal and technological trends that influence research in VGI and the development of VGI applications. These dimensions of VGI influence the characteristics of VGI and their potential application.

2.1.2 VGI Characteristics

Volunteers generate VGI using geoweb platforms such as OpenStreetMap or iNaturalist to collect spatial data on diverse topics such as land cover or ecology (Jonietz et al., 2017). The number of VGI projects and applications has grown as mobile sensors become cheaper, and web technologies become more ubiquitous (Elwood et al., 2012). Growth in the geoweb has resulted in the development of an increasingly diverse set of spatial applications that generate diverse sets of spatial data, VGI is often used as an umbrella term to describe diverse types of spatial data, but there are important functional differences between different data sources (See et al., 2016). The concept of the geoweb and

VGI often encompass purpose built VGI applications such as OpenStreetMap and general location-based applications such as Twitter (Hahmann, Purves, & Burghardt, 2014). VGI data can range from passively contributed geotagged photos to quasi-scientific spatial data (Rob Feick & Roche, 2013). Numerous terms have been used describe different types of citizen derived geographic information, these terms include crowdsourcing, citizen science, PPGIS, collaborative mapping and more (See et al., 2016). Discussion of all these terms is beyond the scope of this thesis but Figure 2.1 and the sections below provide a brief summary of relevant concepts.

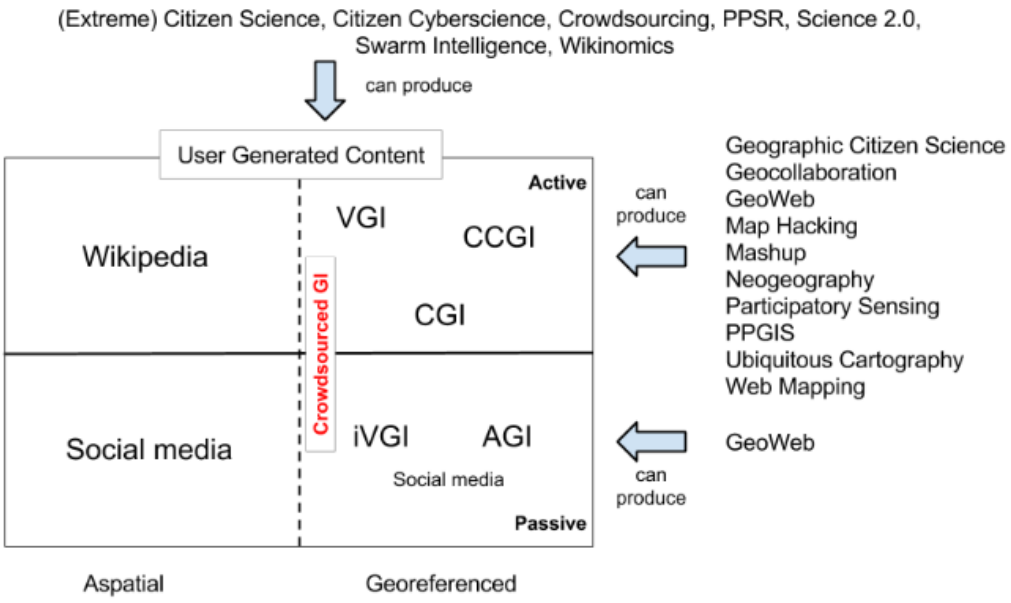

Figure 2.1: A diagram categorizing terms related to citizen derived spatial information (See et al., 2016).

As illustrated by See et al. (2016) in Figure 2.1, VGI can be broadly categorized as active VGI and passive VGI where active VGI involves conscious contributions to platforms such as OpenStreetMap, and passive VGI data is data that is generated as a by-product of digital activity such as Tweeting or contributing to Foursquare (See et al., 2016). Passive VGI has also been referred to as Ambient Geographic Information (AGI), and it refers to geographic information that is automatically collected when a user contributes spatially tagged data to a platform (Stefanidis, Crooks, & Radzikowski, 2013). AGI most commonly contains point data, and it is far less structured than VGI data collected from active VGI sources such as OpenStreetMap or Wikimapia (McKenzie et al., 2014; Spinsanti & Ostermann, 2013). Passive VGI is an incredibly rich source of data as platforms such as Twitter, Foursquare, Yelp, and others create millions of data points per day. Analyzing ambient information requires the analyst to handle challenges associated with unstructured data, big data and natural language processing but the

outcome of such analysis can provide invaluable insight into human trends (Roche, Propeck-Zimmermann, & Mericskay, 2013; Stefanidis et al., 2013). Active VGI refers to the form of VGI that is often associated with concepts such as collective intelligence, crowdsourcing, and citizen science (Connors et al., 2012; Jonietz et al., 2017; Kalvelage et al., 2018; See et al., 2016), participants of active VGI projects deliberately collect data on a topic of interest and attempt to follow community guidelines on data collection to ensure data quality (Connors et al., 2012; Goodchild & Li, 2012; Kalvelage et al., 2018). As shown in Figure 2.1, concepts such as the geoweb, Neogeography, public participation GIS (PPGIS), map hacking and others describe processes or applications that are used to generate active VGI data (See et al., 2016). The purpose of active VGI platforms is to generate spatial data that can be used directly in GIS applications or studies using known tools such as ETL software and mapping applications. When discussing active and passive VGI, it is useful to categorize the type of knowledge that is captured. Structured VGI, such as land cover data and environmental data, has been described as scientific knowledge, this data includes complex geometry data with a set of structured attributes (Connors et al., 2012). Platforms that collect scientific knowledge are also described using terms such as citizen science or geographic citizen science (Connors et al., 2012; See et al., 2016). Unstructured VGI data primarily collects the opinions of participants or local knowledge using custom applications or social media applications (Hall, Chipeniuk, Feick, Leahy, & Deparday, 2010; Tulloch, 2008). PPGIS and PGIS approaches tend to produce unstructured VGI that contains local knowledge about a specific location or topic of interest (See et al., 2016; Verplanke et al., 2016). Passive VGI from platforms such as Twitter can also be used to collect local knowledge (Spinsanti & Ostermann, 2013), but active VGI collected from PPGIS processes tend to contain less noise and more relevant information for a given topic of interest (See et al., 2016; Verplanke et al., 2016).

Another defining characteristic of VGI is the responsiveness of the medium. VGI platforms can be used to rapidly collect vital updated local data in response to disaster situations (Camponovo & Freundschuh, 2014). In some cases, VGI platforms have been more responsive and effective at collecting data than news organizations in disaster situations (Poser & Dransch, 2015). The large community of contributors on OpenStreetMap allows the platform to be responsive to major events. However, due to the volunteer-based nature of VGI it is rarely temporally responsive at all times (Rob Feick & Roche, 2013). In disaster situations, OpenStreetMap allows volunteers to update infrastructure data such as roads or buildings that may or may not be damaged or flooded. Passive VGI platforms such as Twitter make it possible to collect local knowledge about the location of people in affected areas or details about the nature of the disaster (Camponovo & Freundschuh, 2014; Spinsanti & Ostermann, 2013). The

temporal nature is a key differentiator between VGI and authoritative data sources such as TomTom or Google Maps. However, as passive sensors become more sophisticated, companies such as Google, MapBox, and Mapillary are increasingly developing time sensitive spatial data without the use of VGI. For the time being, VGI remains a key tool for disaster response and research interest in this field continues to grow (Jonietz et al., 2017).

The previous sections have introduced the concept of VGI and discussed several related terms such as Neogeography, the geoweb, and public participation. This section has discussed the characteristics of VGI and the various classifications that are used to describe different types of VGI data. Review of the literature has shown that VGI can be collected using active and passive platforms and that VGI data can be described as scientific knowledge or local knowledge (Connors et al., 2012; Hall et al., 2010; See et al., 2016). Differences in structure and format of the various VGI data have also been explained. The spatial-temporal nature of VGI has also been reviewed as it is a key characteristic of VGI. The following sections discuss the concept of spatial data quality in VGI and the solutions researchers have developed to address the challenge of data quality.

2.2 Spatial Data Quality

Spatial data quality has been a major focus of GIS research for several decades, and it continues to be a significant dimension of GIS research today (Devillers et al., 2010; Goodchild & Li, 2012; D. Li, Zhang, & Wu, 2012). Spatial data quality is defined as the measure of the difference between spatial data and the real-world spatial entity the data represents. High data quality indicates data closely resembles reality, and low data quality indicates a divergence between data and real-world entities (Devillers & Jeansoulin, 2006; Devillers et al., 2010; Vandecasteele & Devillers, 2013). Accuracy and quality of information are vital in GIS due to the role maps and spatial data play in the decision-making process (Camponovo & Freundschuh, 2014). There are inherent challenges to ensuring the quality of spatial data because generalization and simplification are required to conceptualize real-world spatial entities within a spatial database (Devillers et al., 2010). The issue of spatial data quality has become more important because larger groups of untrained users have access to geospatial data and map applications over the web (Devillers, Bédard, & Jeansoulin, 2013). With trends like VGI and Neogeography, amateurs are increasingly creating and consuming spatial data over the web with little organizational oversight or assurance of data quality (Ali & Schmid, 2014; Goodchild & Li, 2012). VGI applications are already being used to support vital operations such as monitoring environmental trends (Connors et al., 2012) and responding to natural disasters (Camponovo & Freundschuh, 2014; Spinsanti

& Ostermann, 2013). A growing body of research has focused on developing tools and frameworks to validate VGI data and improve VGI data quality to support the use of VGI in critical situations such as disaster response. Research in VGI data quality is important because in some cases, VGI is the only source of data that is available to support decision making (Stefanidis et al., 2013).

At times, VGI platforms such as OpenStreetMap can produce data that is comparable to spatial data produced by authoritative data sources such as the Ordnance Survey (Haklay, 2010). However, this observation is not the norm because OpenStreetMap generally lacks consistent structure due to inconsistencies in quality, standards, and perceptions between contributors (Ali & Schmid, 2014; Devillers et al., 2013). Overall contributors to VGI platforms lack the tools, training, and organizational oversight that authoritative data agencies have (Ali & Schmid, 2014). Therefore, ensuring VGI data quality requires the development of innovative social models, standards, and methodologies that are oriented towards ensuring quality on a distributed platform like OpenStreetMap (Goodchild & Li, 2012). The following sections review the progress made in spatial data quality (SDQ) literature before reviewing the literature on VGI data quality.

There is a large body of work in SDQ that has contributed to the development of international spatial data quality standards which are maintained by the International Organization for Standardization (ISO) (Devillers et al., 2010). Geographic data is regulated under ISO/TC 211 which is a specifications that aims to define standards for spatial data management, acquisition, processing, analysis, and more (Swedish Standards Institute, 2019). The specific standards that the ISO has published to monitor spatial data quality and reliability are defined by ISO 19157:2013 (Swedish Standards Institute, 2019). The ISO standards and the SDQ literature identify several metrics that evaluate different dimensions of spatial data quality. The literature commonly identifies five key metrics of quality which include completeness, logical consistency, positional accuracy, temporal accuracy and attribute accuracy (Devillers et al., 2010; Esmaeili & Karimipour, 2015; Goodchild & Li, 2012). The description of each data quality metric is as follows.

- Completeness: Measure of omissions of features, attributes, and relationships. Omission errors occur when spatial data is incomplete (Esmaeili & Karimipour, 2015).
- Logical consistency: A measure of how well spatial data conforms to relationships in the data structure, this measure can involve topological relationships (Esmaeili & Karimipour, 2015).
- Positional accuracy: The measure of how well the coordinate values match the real-world coordinates of the represented entity (Esmaeili & Karimipour, 2015).

- Temporal Accuracy: How a temporal attribute of GIS data matches the true presence of an entity at a given time (D. Li et al., 2012).
- Attribute Accuracy: How well attribute records match records considered to be true. (D. Li et al., 2012)
- Lineage: Description of the source material and methodologies used to produce spatial data. (D. Li et al., 2012)
- Fitness of Use: An external measure of quality that determines how well spatial data meets the information needs of the user. (Devillers et al., 2013, 2010; Devillers & Jeansoulin, 2006).

In practice, implementing the metrics described above is very complex and entire fields of research have been developed around topics such as positional error analysis, classification accuracy, conflation, spatial matching and error propagation (Adams et al., 2007; Koukoletsos, Haklay, & Ellul, 2012; D. Li et al., 2012). This thesis categorizes data quality research in three broad topics which include assuring data quality procedurally (Devillers et al., 2010; Goodchild & Li, 2012), measuring data quality (Koukoletsos et al., 2012; D. Li et al., 2012) and improving data quality through enrichment (Adams et al., 2007; McKenzie et al., 2014). The following sections review literature related to VGI data qualities from the three categories described above. The discussion in the following sections depicts the progress made in the field of VGI data quality and opportunities for future work.

2.2.1 Assuring Spatial Data Quality

Data quality is important to many VGI communities and various platforms have developed social and automated approaches to ensure data quality from a systematic perspective. Differences in the size and purpose of VGI platforms tend to dictate the approach used to ensure quality. Large platforms that track general knowledge can rely on crowds to validate data (Goodchild & Li, 2012) while smaller domain-specific VGI projects may rely on contributions of experts to validate data (Connors et al., 2012). Goodchild and Li categorized VGI data quality approaches into three major topics that are termed the crowdsourced approach, the social approach, and the geographic approach (Goodchild & Li, 2012). The crowdsourced approach relies on large groups of contributors to validate data collectively to arrive at a single source of truth. This approach is also referred to as the “Linus Law” approach based upon the approach used to maintain code quality on the Linux project (Goodchild & Li, 2012). This approach uses the idea of collective intelligence to validate data, but it is dependant on the perspective and expertise of users. The concept of collective intelligence assumes that groups generate more reliable information than individuals (Goodchild & Li, 2012). The crowd-sourced approach to data quality is not unique to VGI

because it is also used in platforms like Wikipedia to ensure the validity of entries (Wilkinson & Huberman, 2007). The challenge of this approach revolves around the diversity of contributors within VGI systems, as the expertise and background of VGI contributors vary, the perception of quality and accuracy also varies (Grira, Bédard, & Roche, 2010). The collective intelligence approach to data quality also suffers from a lack of centralized control where disagreements between contributors can result in constant changes in data as different contributors attempt to assert their views (Goodchild & Li, 2012). In contrast, the social approach to data quality relies on hierarchy and proven experience of contributors to ensure data quality. This system ranks contributors to determine user reliability and provides moderator privileges to contributors who are deemed to trustworthy (Goodchild & Li, 2012). The social approach to data quality assumes that the trustworthiness of a contributor is a proxy for the quality of contributed data (Fogliaroni, D'Antonio, & Clementini, 2018). At times the social approach to data validation can be quasi-professional where volunteers contribute data, but professionals act as the gatekeepers of the data, this approach to data validation has been demonstrated in several VGI and citizen science projects (Connors et al., 2012; Kalvelage et al., 2018). The geographic approach to data quality is a proposed concept where purported geographic data is compared to a known geographic area for validation (Goodchild & Li, 2012). The geographic approach proposed by Goodchild and Li is comparable to the use of topological rules to ensure spatial data quality in framework data (Ali & Schmid, 2014) and the use of metadata to standardize a common understanding of spatial entities and their relationships (Devillers et al., 2010). Ali and Schmid built on this concept and developed methodologies for evaluating the quality of data classification on OSM using hierarchical consistency analysis and classification plausibility analysis (Ali & Schmid, 2014). Hierarchical consistency analysis evaluates the hierarchy of boundaries in OSM and identifies inconsistent data classification based upon identified hierarchical rules of regional and state boundaries (Ali & Schmid, 2014). Classification plausibility determines the likelihood of classification being correct in a given location. This approach uses machine learning techniques to teach a program how data should be classified based on geometry and attributes, then the program is used to evaluate the validity of existing OSM tags (Ali & Schmid, 2014). Another study attempted to address semantic heterogeneity of VGI data by developing an OSM editor plugin that recommends feature tags based upon tags of nearby features (Vandecasteele & Devillers, 2015). The OSM semantic plugin in the study analyses the semantic properties of neighbourhood features by measuring the semantic similarity between pairs of tags (Vandecasteele & Devillers, 2015). When a contributor uses OSM semantic plugin and adds a point to OSM in a shopping district, the plugin automatically suggests resteraunts or shop tags to be associated with the point

(Vandecasteele & Devillers, 2015). There are numerous dimensions to spatial data control on VGI platforms, and a significant amount of research has been conducted on improving these processes. Social constructs are still a significant dimension of data quality control on VGI platforms, but there are a growing number of studies that are attempting to use geographic context to ensure data quality.

2.2.2 Measuring Spatial Data Quality

Measuring spatial data quality is a particularly challenging task because there are numerous dimension of spatial data quality and there are numerous approaches for measuring different dimensions of spatial data quality (Devillers et al., 2013, 2010; D. Li et al., 2012). Most approaches for quality measurement require comparisons between purported data and true value data (Koukoletsos et al., 2012). Studies on VGI data quality often involve comparisons of VGI such as OpenStreetMap (OSM) to authoritative framework data maintained by mapping agencies such as the Ordnance Survey (Girres & Touya, 2010; Haklay, 2010; Koukoletsos et al., 2012). Comparisons of VGI data to authoritative data often involves some process of data matching. Spatial matching refers to techniques that compare the geometric and semantic properties of spatial elements to identify elements that correspond to the same place in the physical world (McKenzie et al., 2014). Once spatial features from the VGI data set are matched to spatial features from an authoritative data set, discrepancies between the purported data and the authoritative data are identified. Data quality studies on VGI platforms such as OpenStreetMap have evaluated data quality metrics such as positional accuracy, completeness, logical consistency, lineage, attribute accuracy, semantic accuracy, and usage (Girres & Touya, 2010; Haklay, 2010). Some metrics such as logical consistency are internal measures of data quality where quality is assessed based on how often similar features are represented using dissimilar geometries or attributes (Grira et al., 2010). Other metrics of spatial data quality such as positional accuracy, attribute accuracy and semantic accuracy require direct comparisons between individual VGI features and their matching counterpart within an authoritative dataset (Girres & Touya, 2010; Haklay, 2010). When matching linear and polygonal geometries, properties such size, orientation, and shape can be used as unique identifiers of geometries (Doytsher, Filin, & Ezra, 2001; Koukoletsos et al., 2012), but matching point geometries largely relies on distance and attribute-based data matching. Haklay evaluated the quality of linear OSM street network data by buffering OSM street center lines and intersecting the buffered geometry with Ordnance Survey street segments to determine consistency in coverage and position between OSM and Ordnance Survey data (Haklay, 2010). Grira and Touya evaluated positional accuracy of line segments by measuring the average distance between matching line segments as well as the Hausdorff distance between two matching line segments (Grira et al., 2010). Koukoletsos et al. (2012) also conducted a

study on VGI data quality with a specific focus on linear data matching to determine data completeness, they developed a seven-step process for linear data matching. The seven-step process divides linear data into tiles and compares VGI data to reference data at a segment level using geometric and attribute properties before matching individual features from both data sets using geometric and attribute properties (Koukoletsos et al., 2012). Overall, these studies show that OSM VGI can reach high positional accuracies in some locations, but the studies also show that OSM tends to have a high degree of heterogeneity in positional accuracy, completeness, and semantics (Girres & Touya, 2010; Vandecasteele & Devillers, 2015). The OSM platform has been the focus of many VGI data quality studies due to the size of the platform and the availability of authoritative reference data such as the Ordnance Survey (Haklay, 2010). Data quality research for other sources of VGI data such as Flickr and Twitter have tended to focus on improving data quality rather than measuring quality. The following section describes select studies that have developed methodologies to improve the quality and functionality of VGI data through processes of data matching and conflation.

2.2.3 Improving Spatial Data Quality

Spatial data quality of VGI data can be improved by combining VGI data with authoritative VGI data or combining VGI data with other VGI data sets (Leibovici et al., 2015; McKenzie et al., 2014). Conflation is the process of combining spatial data from different sources to create a new data set that is of higher data quality or richer in information than its constituents (Abdollahi & Riyahi Bakhtiari, 2017; Touya, Coupé, Jollec, Dorie, & Fuchs, 2013). The literature on conflation generally focuses on the integration of spatial data, but there are studies on improving the quality of VGI data by integrating semantic data with spatial data (Ballatore & Bertolotto, 2018), this process is referred to as enrichment. The conflation of VGI data has been explored in the context of OSM data (Hacar & Gökgöz, 2019), citizen science (Connors et al., 2012; Leibovici et al., 2015) and mining of social media data (McKenzie et al., 2014; Spinsanti & Ostermann, 2013). Conflation can be necessary when working with point based VGI data due to the heterogeneity of data and the inconsistent data quality (Leibovici et al., 2015). There are two core processes associated with conflation, the first process involves spatial data matching to identify common features and the second step requires combining matching features to form an aggregate feature (Hacar & Gökgöz, 2019; McKenzie et al., 2014). Spatial data matching is handled through the development of similarity measures between spatial features (Adams et al., 2007; McKenzie et al., 2014). The spatial data matching process attempts to establish the identity between two features by comparing their positional, geometric and semantic characteristics (McKenzie et al., 2014). A significant number of studies focus on the linear matching of road network data due to the importance

of data quality for navigation, these studies tend to focus on comparing the geometric and positional attributes of spatial data in order to match and conflate data (Hacar & Gökgöz, 2019; Koukoletsos et al., 2012). Other studies have looked at the conflation of point and POI data of VGI data (McKenzie et al., 2014). The conflation of point data tends to require comparisons of semantic and attribute traits rather than spatial traits to establish identity between features (McKenzie et al., 2014; Ramos, Vandecasteele, & Devillers, 2014; Yu, West, Arnold, McMeekin, & Moncrieff, 2016). Another study used semantic matching approaches to enrich OpenStreetMap data with non-spatial DBpedia data (Ballatore & Bertolotto, 2018). This thesis used an intermediate service to match non-spatial DBpedia data to point coordinates before conflating the generated DBpedia point data with OpenStreetMap data. In general, approaches to conflation vary significantly based upon the characteristics of the data being conflated. However, researchers have attempted to identify some basic concepts that are common to the process of conflation and spatial data matching (Adams et al., 2007; McKenzie et al., 2014).

Most conflation and spatial matching methodologies have some consideration of geometry, proximity, feature type and attributes (Adams et al., 2007; Hacar & Gökgöz, 2019; McKenzie et al., 2014; Yu et al., 2016). Geometry dictates the methodology used to conflate spatial data. The linear road network has geometric and topological properties that make it possible to match features based solely on geometric and spatial characteristics (Koukoletsos et al., 2012). Linear spatial data has properties such as orientation, sinuosity, mean perpendicular distance, mean length of triangle edges and modified degree of connectivity that can be used measure similarity between features (Hacar & Gökgöz, 2019). Polygon geometries have a comparable set of properties that can be used to determine feature similarity (Hacar & Gökgöz, 2019). In contrast, point geometries do not have a set of defining characteristics that can be used to match features (McKenzie et al., 2014). Spatial proximity of features is a fundamental filtering mechanism that is used to identify feature matches regardless of geometry type (Hacar & Gökgöz, 2019; McKenzie et al., 2014). Semantic matching is the process of evaluating the similarity between features based upon attribute information such as road names, road category, or dataset name (Yu et al., 2016). Studies that discuss semantic matching processes tend to rely on ontologies to determine the semantic similarity between feature attributes or tags (Ramos et al., 2014). The process of conflation involves generating normalized similarity scores between features using the different characteristics of the input data sets, the normalized scores are then aggregated to a single similarity score using a score weighting system (Hacar & Gökgöz, 2019; McKenzie et al., 2014). The aggregate similarity score is used to determine if two features match and informs subsequent conflation processes (Hacar & Gökgöz, 2019). Once two features are matched, the data needs to integrate or

conflated. Data integration requires combining the geometry of two features and integrating the attributes of two features (Touya et al., 2013). Geometry data is integrated using algorithms that prioritize the preservation of different data characteristics such as topology and shape (Touya et al., 2013). Attribute data can be integrated using weighting algorithms that fill gaps in attribute data, conflicts of attribute data require one data source to be prioritized over another (McKenzie et al., 2014). When conflating VGI data with known data, researchers tend to prioritize the preservation of known data in the conflation process (Leibovici et al., 2015). Overall, numerous studies have attempted to improve the quality and value of VGI data through conflation or enrichment (Hacar & Gökgöz, 2019; McKenzie et al., 2014; Ramos et al., 2014). Studies on conflation have developed numerous methodologies to match data using multi-attribute similarity metrics that combine geometric criteria and semantic criteria. The heterogeneity of VGI data has driven increased interest in the development of semantic matching criteria (Ballatore & Bertolotto, 2018; Yu et al., 2016). Conflating VGI requires an improved understanding of spatial context and semantics. Researchers continue to look at models and metrics that can address these challenges.

2.3 Geographic Relevance

The concept of geographic relevance (GR) was developed to improve geographic information retrieval (GIR) systems. It is an area of research that codifies the metrics needed to understand the spatial context (Purves et al., 2018). Geographic relevance is defined as a measure of how well a given piece of spatial information meets the spatial information needs of a user, and it is often measured through analysis of spatial context (Raper, 2007; Reichenbacher et al., 2016). As noted in section 2.2, the geographic context is also explored in spatial data quality literature and several topics in SDQ literature share similarities with GR literature (Goodchild & Li, 2012; Spinsanti & Ostermann, 2013). Research in GR is closely linked to research in GIR and information retrieval (IR), and the concept of relevance is central to both fields of research. Research in Information Retrieval (IR) pioneered the concept of relevance used in GIR and GR, relevance in IR is a measure of similarity between query text and a set of documents (M. Li, Sun, & Fan, 2015). Information Retrieval (IR) is the field of research concerned with retrieving data from a database based upon relevance to a users' query (Dominich, 2008; M. Li et al., 2015). The difference between conventional SQL queries and IR systems is the ambiguity of user queries and the presence of probabilistic relevance rankings that estimate relevance between content and the query (Kunz, 2009). Information retrieval is probabilistic, and data retrieval is deterministic. Data retrieval returns exact matches of structured data in a random order. In contrast, information retrieval

systems only return relevant unstructured data in order of relevance (Merrouni, Frikh, & Ouhbi, 2019). The idea of relevance is a concept in IR that codifies the metrics needed to understand a user's information needs and retrieve relevant data (De Sabbata & Reichenbacher, 2012). Quantifying a user's information needs is a non-trivial task that is handled through the development of complex information retrieval models that evaluate context and user information (Hjørland, 2010; Merrouni et al., 2019). Location data in information retrieval have become more important as mobile phones and mobile search has grown (M. Li et al., 2015; Merrouni et al., 2019). The need to understand the spatial context in information retrieval has driven research interest in geographic information retrieval.

2.3.1 Geographic Information Retrieval Systems

GIR is predominantly concerned with identifying fuzzy or vague location references and understanding the spatial relationship between the user, the query and documents in the database (Kunz, 2009; Mata-Rivera, Torres-Ruiz, Guzmán, Moreno-Ibarra, & Quintero, 2015). Geographic Information Retrieval (GIR) systems are information retrieval systems that are designed to handle spatial data (Purves et al., 2018). In order to handle spatial data, GIR systems have to recognize spatial relationships and account for different representations of spatial entities (Mata-Rivera et al., 2015). Most GIR research focuses upon document-based queries from the web in order to match a single location reference to single physical location. GIR systems evaluate queries using semantic analysis, spatial analysis, and relevance analysis (Clough, Joho, & Purves, 2006; Purves et al., 2018).

Spatial references on the web are often incomplete and fuzzy. Therefore, semantic analysis is required to match vague location references to known locations (Mata-Rivera et al., 2015; Silva, Martins, Chaves, Afonso, & Cardoso, 2006). GIR expands IR by providing a means for IR systems to process toponyms, identify geographic footprints, and evaluate geographic relevance (Kunz, 2009). Bordogna et al. (2012) describe GIR as follows.

> *"Geographic information retrieval (GIR) is nowadays a hot research issue that involves the management of uncertainty and imprecision and the modeling of user preferences and context. Indexing the geographic content of documents implies dealing with the ambiguity, synonymy, and homonymy of geographic names in texts."* (Bordogna et al., 2012, p. 105)

GIR research aims to build models to understand place names, understand location references, and identify spatial relationships (Overell, 2009; Purves et al., 2007). GIR systems have to manage ambiguity and handle fuzzy location references such as "close", "near", "trail" or "Mid-Europe"

(Bordogna et al., 2012). GIR systems also have to handle multiple location reference in a single query so there is a need to maintain many to one relationships when evaluating relevance between features (Bordogna et al., 2012). Researchers in GIR have attempted to address these challenges using a variety of methodologies that handle semantic analysis, spatial analysis and data management (Acheson et al., 2018; Bordogna et al., 2012; Purves et al., 2018). Several GIR systems have been developed over the years both by academic and commercial groups. Some academic examples of GIR systems include the Geo-Finder application by Bordogna et al. (Bordogna et al., 2012) and the SPIRIT application developed by Purves et al. (Purves et al., 2007, 2018). Google and Bing are two major examples of commercial platforms that have GIR capabilities (Overell, 2009). The geo-finder system is modeled after a typical information retrieval system with an indexing module and a retrieval module, but the indexing module of the Geo-Finder system is customized to support the identification of location references (Bordogna et al., 2012).

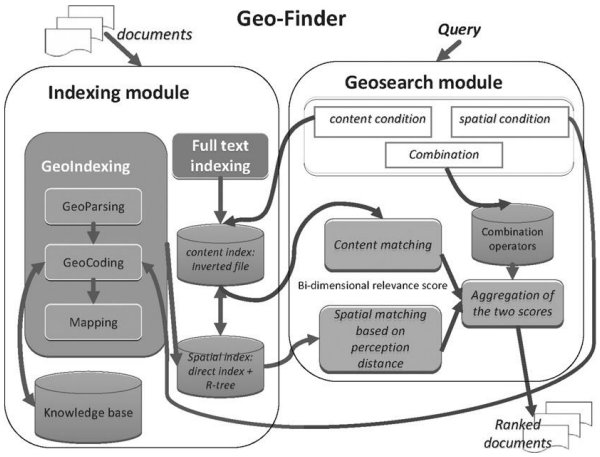

Figure 2.2: Geo-Finder Architecture (Bordogna et al., 2012)

Geo-finder was implemented using Postgres with the PostGIS extension. The geo-finder system was composed of several sub-modules that perform spatial query analysis and spatial query indexing. Geo-finder included a reference database which includes an English gazetteer populated with place names from open libraries such as Geonames and OpenStreetMap. The gazetteer also includes a list of distinct location-based markup terms such as city, mountain and others. A set of parameterized rules that disambiguate location references are also stored in the gazetteer. The geo-index module parses each document and identifies the fuzzy footprint of each document by applying the parametric rules defined in the knowledge base. The results of the geo-indexing module are then stored in a dual data structure which is composed of a direct index and a 2-D R-tree. (Bordogna et al., 2012)

The SPIRIT platform, developed by Purves et al. (2007), is another GIR system which attempts to model and calculate spatial relevance. The platform was developed to support several key features, which include; spatial query expansion, place name disambiguation, ambiguous location search, spatial concept recognition, spatial relevance ranking, and query visualization. The SPIRIT platform enabled search through spatial indexing of web documents and spatial disambiguation of queries. Documents in the SPIRIT platform are assigned a spatial footprint using geoparsing and geocoding. A list of known locations contained within the SPIRIT geo-ontology are used to identify candidate location terms within the web document. The SPIRIT platform used a gazetteer lookup methodology to compare terms in the document to known location names contained within the SPIRIT geo-ontology. The SPIRIT platform combined the gazetteer lookup methodology with several disambiguation heuristics to address ambiguous location references. A text index is also contained in the SPIRIT architecture to support relevance ranking based on topicality. The SPIRIT engine evaluated concepts using both the spatial index and the text index and used three different approaches to generate relevance scores. (Purves et al., 2007, 2018)

2.3.2 Geographic Relevance Metrics

Geographic relevance research is closely tied to GIR but the focus of GR research in mobile computing while GIR has focused on the web. Mobile phones have multiple spatial sensors that make it possible to collect important contextual needed to measure GR. As a result, several models for measuring GR have been based upon cell phone sensors. One of the earliest papers on GR was published by Raper (2007), he defined a measure of the geographical relevance that considered four core metrics that include geometry, time, spatial perception, locality, and manifestation. De Sabbata and Reichenbacher (2012) expanded on Rapers' work and offered an alternative measure of geographic relevance that includes hierarchy, cluster, co-location, and association rule.

- **Hierarchy:** Hierarchy in geographic relevance research is defined as the distance between the user and the location of a given entity within a geographic hierarchy.
- **Cluster:** Custer refers to the degree of membership between an entity and a spatial cluster of related or unrelated entities.
- **Co-location:** This GR metric evaluates the extent to which the target object follows a desired spatial pattern such as a theater close to resteraunts.
- **Association Rule:** This GR metric evaluates the extent to which two spatial entities are correlated, such as high hotel prices within a downtown core.

Another set of criteria to measure geographic relevance was developed by Reichenbacher et al. (2016) in a subsequent paper. The new set of criteria included seven metrics of relevance that are summarized to five scores. This model of GR relied significantly on mobile sensors to gather information about the directionality of the user's movement and the spatial-temporal proximity between users and locations of interest.

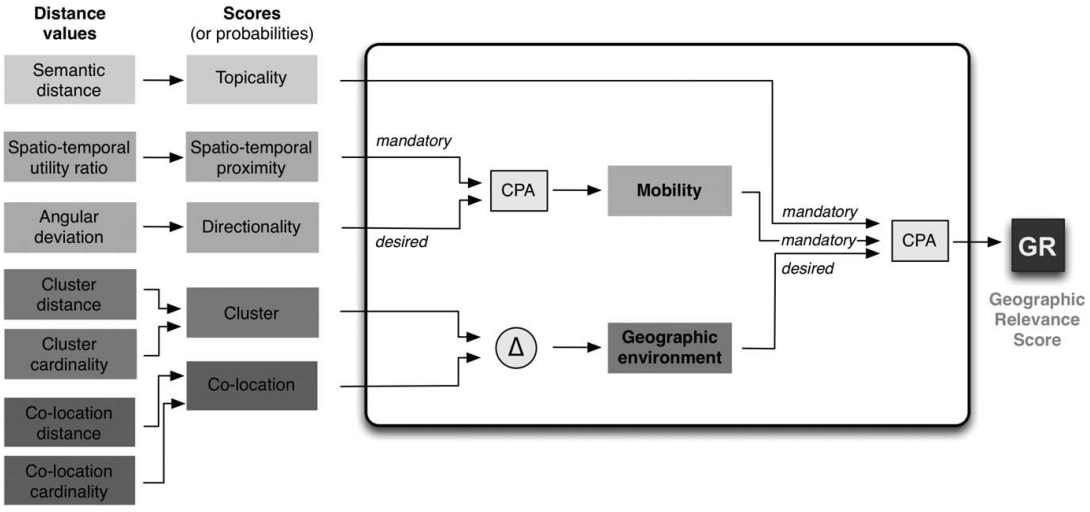

Figure 2.3: Computational GR model that incorporates mobility sensors (Reichenbacher et al., 2016)

The computational model in Figure 2.3 evaluates each GR criterion between a user query and a geographic entity (Reichenbacher et al., 2016), the individual criterion are defined as follows.

- **Topicality:** The semantic similarity between a user query and the entity type or category.
- **Spatio-Temporal Proximity:** A proximity measure that accounts for distance and travel time between the user and the geographic entity.
- **Directionality:** The angle between a user's direction of travel and the geographic entity.
- **Cluster:** The size of the cluster containing the geographic entity. Entities contained within a cluster are assumed to be more relevant to the user.
- **Co-location:** The extent to which the geographic entity satisfies a co-location pattern with entities that have been referenced in the query.

Researchers have started to explore the idea of using geographic relevance outside of mobile search to apply it to challenges like VGI contextualization (M. Li et al., 2015), and the field will likely continue to evolve beyond simple information retrieval.

3 Using Geographic Relevance to Retrieve Domain-Specific Spatial Data

3.1 Introduction

Location data is becoming more accesible to businesses, citizens, and government bodies. Increased use of personal computing (computers, mobile phones) and applications that allow citizens to use and create spatial data are contributing to a growing public awareness of spatial information. Furthermore, open data programs are publishing authoritative government data on the web at no cost. The growing availability of authoritative spatial data on the web is an important opportunity for organizations to leverage spatial data to support research and operations. The volume of freely accessible spatial data and the range of data sets that are made available vary significantly across levels of government and within departments but governments are increasingly adopting open data mandates. However, the heterogeneity of spatial data on the web makes it challenging to use open data at scale. Most GIS professionals know that the data needed to create a map is rarely limited to a single topic or provider. An urban planner that is studying the impacts of a new transit station in a city would need to access diverse sets of data from multiple organizations and levels of government. The planner would need road network data from the local municipality, transit data from the regional transit authority, environmental data from the regional conservation authority, demographic data from the national census and planning policy data from the regional planning authority. Finding and assembling data sets from such varied data sources is a tedious process that takes staff time and resources. Furthermore, in many cases researchers are not aware of all the data that is available on open data platforms (Corti, Lewis, & Kralidis, 2018) and the diversity of open data makes it difficult for an inexperienced researcher to find all the data they need across multiple data portals (Guidoin, Marczak, Pane, & McKinney, 2014; Ivanova et al., 2013). Researchers have described this challenge as the fitness of use problem (Devillers et al., 2010; See et al., 2016). Fitness of use has been described in spatial data quality literature as an evaluation of how well data fits the user needs; the criterion prevents misuse of spatial data by connecting users to the data they need (Devillers et al., 2010).

The challenge of retrieving and organizing spatial data is a consistent problem that has been noted in a variety of GIS literature, particularly in studies that focus on contextual analysis (Bordogna & Psaila, 2008; Spinsanti & Ostermann, 2013). Analysis tools such as gazetteers or spatial ontologies also often rely on the availability of large open data sets (Codescu, Horsinka, Kutz, Mossakowski, & Rau, 2011; Ramos et al., 2014). Accessibility to spatial data is a challenge that several researchers have tried to address through the development of spatial data registries, search engines (Corti et al., 2018;

Florance, McGee, Barnett, & McDonald, 2015), and fitness for use criteria(Jonietz et al., 2016; Wentz & Shimizu, 2018). As spatial data becomes freely available on the web, there is an opportunity to automate the task of assembling spatial data for research projects and address aspects of the fitness of use problem of spatial data on the web (Ivanova et al., 2013).

This thesis looks at developing a process for assembling spatial data using geographic relevance metrics. This Chapter aims to use geographic relevance metrics to assemble domain and location specific spatial data to support research. This Chapter presents a prototype spatial search platform that can dynamically generate a domain-specific spatial database. Geographic relevance metrics are used in this to determine spatial relevance between layers to identify data sets that are relevant to a topic of study. This subject is explored over five sections. Section 3.2 reviews literature related to open data, geographic information retrieval, and geographic relevance. Section 3.3 describes a methodology to measure geographic relevance between spatial data sets to identify spatial data that is relevant to a topic of study. Section 3.3 also describes UrbanData, a prototype application that implements the conceptual models developed in this thesis into a software system. Section 3.4 and 3.5 describe a case study for data acquired from the City of Kitchener and the subsequent results from the case study. Section 3.6 discusses future directions for research.

3.2 Background and Related Research

3.2.1 Open Data

Open Data is a growing trend among governments across the world. Leaders and analysts believe that open access to government data can improve citizen engagement, business growth and government operations (Gruen, Houghton, & Tooth, 2014). Open Governance is a term that describes an organizational commitment by governments to foster openness and accountability to increase citizen engagement, drive innovation, create economic opportunities and create a more efficient government (The Open Government Partnership, 2016). Open Governance is an idea that has garnered a lot of attention and commitment from local and national governments; many have committed to the international open government partnership which requires participating countries to develop and implement action plans on open government (The Open Government Partnership, 2016). A significant dimension of Canada's Open Governance Action Plan is to broaden access to data and information produced by governments, as a result, open data is often a key dimension of open governance initiatives (Government of Canada, 2014). Open data, within governments, refers to a set of initiatives focused on making government data, public sector information and data generated by government-funded research

accessible to the public at no cost (Gruen et al., 2014). Researchers have defined open data as non-private and non-confidential data generated using public funds and distributed with no restriction on usage (Janssen, Charalabidis, & Zuiderwijk, 2012).

Both open data and open governance are relatively new phenomena that have grown rapidly over the decade. The UK may have been one of the earliest governments to explore the idea of open governance and open data with its Power of Information initiatives in the mid-2000s, the US soon followed suit with the creation of open data mandates in 2009 (Gruen et al., 2014). Numerous governments across the world have followed suit, which has led to the creation of international committees for open data and the development of an open data charter (The World Wide Web Foundation, 2017). Western countries such as Canada, UK, USA, and France have been first movers on open data and open governance. However, Open Data is a global trend as shown by the Web Foundations open data rankings which lists Korea in the top 10, the Philippines in the top 25 and Saudi Arabia in the top 75 (The World Wide Web Foundation, 2017).

Reports estimate that open data will have a significant impact on driving innovation in the tech sector and improving government operations (City of Toronto, 2018; Gruen et al., 2014; Guidoin et al., 2014). Open data is valuable because governments are some of the largest and most reliable generators of authoritative information (Janssen et al., 2012). Making expensive and authoritative data open to the public can spur research and innovation that would not be possible otherwise (Gruen et al., 2014). Governments see open data as an avenue to promote and foster innovation and economic growth (City of Toronto, 2018). Many open governance action plans note that open data has little intrinsic value on its own; value is only created when people start to use open data (Janssen et al., 2012). As a result, accessibility and usability are two major mandates that are outlined in the International Open Data Charter (Calderon, Carfi, & De Luca, 2015). Implementing these mandates is a significant challenge for governments.

Open Data is often generated as a by-product of government activities such as budgeting, maintaining land registries, or tracking endangered species (Specht, 2015). In many cases, different levels of government are generating data that are not designed to be published or used by external organizations or applications. Some publicly generated data sets such as transit data have widely recognized standards such as the GTFS data format. Other data sets such as zoning by-laws do not have a widely adopted standard which leads to a significant amount of variance between data sets generated by different municipalities. There are opportunities to increase standardization of data content,

structure, and format, but there are cases where governments can not adopt external standards due to organizational needs. Data standardization is often a major internal challenge to governments implementing open data plan as noted in Toronto's open data master plan.

> *"The City of Toronto needs to focus on establishing an automated data release …Decreasing the manual effort required to publish and update open datasets is essential for modernizing and scaling up the City's Open Data program."* (City of Toronto, 2018)

This results in cases where similar data between two different government agencies differ in structure and content (Hacar & Gökgöz, 2019). As a result, the use of open data often requires domain knowledge or extensive research on the structure or content of data. This character of open data is a significant barrier to large scale use of open data, especially to researchers who don't have the staff resources needed to preprocess open data. In response to this issue, government staff and research groups have started to develop frameworks to better define and structure open data to promote interoperability across organizations and government sectors (Guidoin et al., 2014). Open data encompasses all data types; this includes financial data, records, permits, and licenses. A significant amount of spatial data has been released under open data programs which have instigated research interest around tracking and studying open spatial data. Researchers have noted that the number of open spatial data sets continues to grow and public and academic institutions continue to develop platforms and infrastructure to support the open data movement (Corti et al., 2018).

There are important parallels between the growth of Open Data and the growth of the web. Researchers and companies alike have recognized the opportunites presented by Open Data to enrich and augment research and data analysis. Companies and research groups such as ThinkDataWorks and Harvard Hypermap have developed open data search engines with a significant focus on spatial data (Corti et al., 2018). Large amounts of data are being created and released as open data, but the challenge of information overload and lack of standards across organizations is a major barrier to adoption and use of open data (Guidoin et al., 2014). Much like the early web, open data needs search and indexing technologies to make data accessible.

3.2.2 Open Data Search Engines

Thousands of spatial web services using open standards are being made available to the public through open data programs; but finding, accessing and using these data sets remains a significant challenge of open data (W. Li, Yanga, & Yang, 2010). According to ESRI, a global leader in GIS technology,

there are over 115 000 open spatial data sets available on the ArcGIS platform with over 6000 organizations sharing and maintaining this data (ESRI Inc, 2019). Harvard Hypermap, an open data search engine developed at Harvard, has over 200 000 data set within its open data registry (Corti et al., 2018). Researchers have identified the need to develop spatial data search engines or web crawlers that can find and retrieve spatial data (W. Li et al., 2010). Platforms such as the Harvard Hypermap (Corti et al., 2018) and Open Geoportal (Florance et al., 2015) have been developed by researchers to explore the challenge of making open data accessible. Comparable commercial projects have also been launched with products like ArcGIS Hub by ESRI and Namara.io by ThinkDataWorks. These search engines use web crawlers to collect metadata about layer extents, tags, taxonomies, and service links to find and retrieve spatial data that is available on the web.

Harvard Hypermap is an index of open spatial data endpoints; the experimental platform aims to make spatial data more accessible to the public by aggregating a list of open data endpoints across the web and tracking their health (Corti et al., 2018). The Hypermap platform is built on top of Django, Apache Solr, Memcached, Postgres, and PostGIS (Corti et al., 2018). The platform leverages the technologies above to develop a service that allows users to search for data using a keyword, source, layer type, map extent and date range (Corti et al., 2018). The Hypermap platform primarily relies on the semantic search capabilities of ElasticSearch coupled with an extent-based spatial query to retrieve spatial data sets that are of interest to the user. Open Geoportal is another spatial data search platform that was developed to improve search and retrieval of spatial data sets harvested by the Open Geoportal Federation research group (Florance et al., 2015). The Open Geoportal platform acts as a federated database for spatial data services provided by Harvard, MIT, Tufts, and Berkley. The platform also features sophisticated location-based search, particularly extent-based filtering and layer scoring based on the layer's similarity to the map (Florance et al., 2015). The Open Geoportal research team developed an extent-based approach to determine the spatial relevance of data layers. Open Geoportal uses three metrics to determine the relevance of spatial data. The metrics evaluate similar extent area, similar centroid location and extent intersection (Florance et al., 2015). OpenGeoportal also allows users to search open data on key metrics such as date, keywords, data type, and institution (Florance et al., 2015). Like the Hypermap project, the Open Geoportal project was motivated by the need to make open spatial data accessible.

3.2.3 Fitness for Use

The problem of finding and retrieving relevant data for a project or a study is not limited to open data, experts and non-experts working with spatial data often have to triage spatial data portals and metadata to find and acquire needed spatial data (Ivanova et al., 2013). The challenge of retrieving spatial data along with the growing volume and variety of spatial data has driven interest in evaluating fitness for use of spatial data (Ivanova et al., 2013; Jonietz et al., 2017, 2016; Wentz & Shimizu, 2018). Fitness for use is a measure of the suitability of spatial data for a set of application needs or requirements (Jonietz et al., 2016). Evaluating fitness for use of spatial data remains an open problem but several studies have attempted to develop frameworks to evaluate fitness for use of different spatial data sets in different contexts. Ivanova et al. (2013) have identified two major challenges associated with fitness for use that are described as communicating quality requirements and evaluating quality. Communicating quality requirements is a challenge because users may not know how to express data needs or understand a provider's data quality parameters (Ivanova et al., 2013). Evaluating fitness for use is a challenge because quality parameters and standards differ between vendors. Therefore users tend to determine fitness for use of data through comparisons between data within their possession and vendor data (Ivanova et al., 2013). Furthermore, evaluating fitness for use of VGI data is complicated by the fact that metadata for VGI data is often incomplete or inconsistent (Jonietz et al., 2016). Ivonova et al. (2013) developed a sample application named GUESS to evaluate fitness for use of spatial data found in a spatial data catalogue. The system accepts text queries from users and returns relevant data based upon a set of criteria that considers user needs as well as user characteristics. User profile information is stored in the application to determine user needs, the GUESS profile tracks user history such as previously used data resources and domains of interest (Ivanova et al., 2013). Candidate spatial data sets are retrieved based upon the domain of interest, the user profile and characteristics of other users with similar information needs (Ivanova et al., 2013). In contrast, Jonietz et al. (2016) developed an approach to evaluate fitness for use of VGI POI data by comparing VGI data to authoritative reference data. The study defined fitness for use as a measure of how well a POI data set met user-defined quality criteria such as positional accuracy or attribute accuracy (Jonietz et al., 2016). Wentz and Shimizu (2018) proposed a method to measure fitness for use of data using a multi-criteria decision-making model named DaFFU. The DaFFU model allows users to prioritize several measures of spatial data quality such as positional accuracy, completeness, logical consistency, credibility and more, the DaFFU model measures data quality scores for data sets using a weighting schema that is based upon user priorities (Wentz & Shimizu, 2018). In general, the literature indicates that multiple criteria

are needed to evaluate fitness for use of spatial data. Some studies focus on user information needs such as user preferences and domains of interest while others evaluate the internal quality characteristics of data such as logical consistency. Currently, there are no clear set of criteria that are used to evaluate fitness for use.

3.2.4 Geographic Information Retrieval and Geographic Relevance

The growth of location data across the web and mobile devices motivated research in geographic information retrieval systems (GIR). Understanding spatial relevance is a key research problem in GIR literature that has been approached in many different ways in the literature (Bordogna et al., 2012; Purves et al., 2007). Spatial relevance is a concept that is often broken down into a semantic and spatial dimension when measured within software systems. Semantic relevance in GIR refers to the presence of location references within a document. Spatial relevance in GIR is a measure of proximity between the query and location references within a document.

Research on geographic information retrieval (GIR) systems has led to the development of several search platforms that enable web document searches that rank results based upon spatial relevance. Development of search engines has also led to the development of algorithms and heuristics for measuring and evaluating spatial relevance. Purves et al. (2018) determined that the geographic relevance of web documents for a given query is best determined using a combination of spatial and topical analysis. Purves et al. (2018) measured topical relevance using three heuristics, which are defined as follows.

- **Document Frequency:** Terms that occur in a few documents are important than terms that occur in a few documents.
- **Term Frequency:** Terms that appear more frequently in a given document are more important to the given document.
- **Document Length:** Higher ratios of term frequency to document word count indicate higher levels of importance.

Gao et al. (2016) developed a geographic information retrieval system that developed a new model for model spatial footprints of documents and queries. This research approached the task of measuring geographical relevance through spatial and semantic similarity measurements using point-set based spatial footprints (Y. Gao, Jiang, Zhong, & Yu, 2016). Gao et al. (2016) determined spatial similarity scores between query footprints through the use of topological and distance-based tests. Bordogna et

al. (2012) developed a heuristic based methodology for measuring the relevance of web documents in the Geo-Finder platform. Bordogna et al. (2012) measured geographic relevance within the Geo-Finder using a series of heuristics that identified the spatial footprints of queries and web documents using a gazetteer. The system used a gazetteer to parse documents and identify the fuzzy footprints of documents and queries; the fuzzy footprints are used to determine spatial relevance scores which are combined with topological relevance scores to generate an aggregate relevance score (Bordogna et al., 2012; Bordogna & Psaila, 2008). The relatively simple spatial relevance evaluation model used in GIR has been expanded in geographic relevance (GR) literature to evaluate contextual spatial data when evaluating spatial relevance. Measures of geographic relevance are defined using a varying set of metrics in the literature; many of the metrics defined in GR literature are comparable to relevance metrics found in GIR literature. Reichenbacher et al. (2016) argue that GR is a measure of five associated measures, which include topicality, spatiotemporal proximity, directionality, cluster, and co-location. This model of evaluating spatial context and geographic relevance evaluates several spatial characteristics and properties that are not found in GIR or spatial data quality literature. The model developed by Reichenbacher et al. (2016) may be applicable to spatial relevance analysis in other GIS contexts and it may be more effective than current models.

3.3 Methodology

This section presents UrbanData, a prototype application that dynamically generates domain-specific spatial databases in response to user information needs. UrbanData is a data processing application that analyzes and compares spatial data sets within a database and assigns spatial relevance scores between data sets. The UrbanData application generates rankings that attempt to measure the fitness of use of spatial data sets based on a user-defined topic of study. This application is oriented towards addressing the challenges associated with retrieving and organizing spatial data when conducting a location-based study. The app is developed around the premise that a user is interested in exploring a known data set and is interested in finding relevant data sets that can add context and reference content to the data set of interest. Potential users of the UrbanData application include GIS analysts, urban planners, environmental researchers, health researchers, engaged citizens and neogeographers. With the exception of the GIS analyst, people from these backgrounds may not be familiar with spatial databases, GIS systems or spatial data standards. However, they often have a need to work with spatial data from multiple sources in order to support research. For example, a planner may need to use spatial data sourced from the provincial government, a local conservation authority

and a neighboring municipality in order to review a development application. Another example may have an environmental researcher attempting to use provincial soil maps and regional watershed maps to study the relationship between soil characteristics and watersheds. Currently open data portals and open data search engines rely on a combination of manually defined metadata and dynamically generated spatial indexes or layer extents to support search functions. Metadata can be incomplete at times which can make these search systems ineffective. Furthermore, using spatial indexes for search does not account for the spatial distribution of features within layers. UrbanData builds on these relatively basic topic and spatial linkages between data sets to include criteria that consider the distribution of spatial features. It demonstrates how data producers can provide finer-grain associations between data sets that are pertinent to specific areas or topics. These linkages make data retrieval easier for end-users

This study uses a simplified use case in order to develop a prototype model to address this challenge. When a user selects the data set of interest within the application, UrbanData dynamically retrieves and ranks a set of spatial data sets from the framework database that are relevant to the topic and location of study. It is envisioned that a production implementation of the UrbanData application would be deployed by a system administrator on an open data portal or geoportal and that users would interact with the application as a data search engine. In order to deploy the UrbanData software application a systems administrator would need to load a set of framework data sets into the UrbanData framework database. The systems administrator would also need to provide three threshold distance values in order to calibrate the relevance ranking system. The details of the relevance ranking system are explained in the following sections. Users can interact with UrbanData to retrieve domain specific spatial data. It is envisioned that a user would select a single domain data set that is of interest from a list or search interface much like current open data portals. Once the user selects a domain data set that is of interest the UrbanData application will compare the domain data to all the other data sets in the framework database and retrieve a set of data sets that are topically and spatialy relevant to the domain data set. The UrbanData application may present the set of domain specific spatial data sets as a list, a map or a downloadable package of layers that can be used for future analysis. Behind the scenes, UrbanData uses the methodology described in the following sections to retrieve and rank spatial data sets using geographic relevance criteria.

The UrbanData methodology builds on metrics defined in GR literature to determine relevance between spatial data sets. This methodology will expand on metrics defined by Reichenbacher and De Sabbata (2012), particularly the metrics of topicality, spatial proximity, cluster and co-location. The four geographic relevance metrics used in this thesis have been developed based on research in GIR (Jones & Purves, 2008), conflation (McKenzie et al., 2014) and geographic relevance (Reichenbacher et al., 2016). Research in GIR, conflation, and GIR have widely recognized the need for semantic and spatial metrics when evaluating relevance or context (Fu, Jones, & Abdelmoty, 2005a; Goodchild & Li, 2012; McKenzie et al., 2014; Spinsanti & Ostermann, 2013). In this methodology, the semantic relevance metric is defined as topicality and the spatial relevance metric is defined as spatial proximity. GR researchers have expanded the common understanding of relevance and context by identifying secondary metrics of spatial relevance which include directionality, cluster co-location and others (Raper, 2007; Reichenbacher et al., 2016). This thesis uses adapted versions of the cluster and co-location metric as described by Reichenbacher et al. The literature on GIR and GR have predominantly focused on evaluating relevance between a single location or entity of interest and its surrounding elements (Reichenbacher et al., 2016). This thesis evaluates relevance between spatial datasets which is computationally expensive and structurally complex. The following section will describe the novel approach used to adapt GR metrics intended for mobile phones to a spatial database. The assumptions made in the development of the GR metrics for UrbanData differ from assumptions made in GR literature and assumptions made in Chapter 4 of this thesis due to the scale and volume of the data being analyzed. The methodology begins with an overview of the modules and workflow used to analyze data and generate relevance metrics. Next, the architecture of the UrbanData platform and GR metrics are described at a high level. The subsequent sections provide a detailed description of the individual GR metrics used to measure relevance. The sections describing GR metrics will include code samples and sample data from the case study described in section 3.4. The sample data in the sections below are used to explain general concepts in the methodology. The code associated with the methods described in the following sections can be found in Appendix A.

3.3.1 UrbanData Overview

There are numerous approaches to measure geographic relevance (Raper, 2007; Reichenbacher et al., 2016). UrbanData uses four metrics developed by Reichenbacher et al. (2016) which includes topicality, spatial proximity, cluster, and co-location. The GR metrics defined by Reichenbacher et al. (2016) have been adjusted to detect characteristics of large spatial data sets and account for the characteristics of input data. The first adjustment is the use of the spatial proximity metric rather than

the spatial-temporal proximity metric because the spatial temporal metric is designed for moving mobile users. The second key change is to use a cluster-colocation metric rather than the cluster metric because the cluster metric is designed to evaluate the importance of points of interest. These changes are necessary to evaluate the general characteristics of spatial data. All four metrics are described in greater detail in subsequent sections. As shown in Figure 3.1, UrbanData measures geographic relevance between a single input layer and a set of framework layers. Layers can range from framework data sets such as roads, parks, trails and geo-tagged Tweets. Geographic relevance is analyzed by isolating and identifying traits within layers and between layers to determine relevance and association between layers. The diagram below illustrates how individual GR metrics are applied between input query layers and framework data sets.

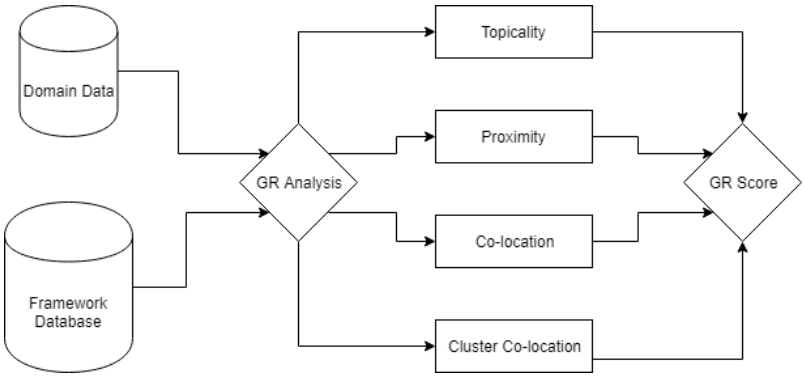

Figure 3.1: UrbanData computational model where domain data is compared to framework data.

Figure 3.1 illustrates how domain data is compared to framework data sets. The combination of the input dataset and a target framework dataset defines the query for the GR analysis. Topicality, proximity, cluster co-location, and co-location are measured between the input dataset and each framework data set and produces a normalized score between 0 and 1. The four relevance scores are then combined to generate a single GR score between the two query layers. This process is repeated between the input layer and every framework data set to generate a list of framework layers that are ranked by geographic relevance. The layers with the highest GR scores are then selected to form the domain-specific spatial database. The following section discusses how this workflow is implemented in the UrbanData software application.

The UrbanData platform is a data processing application that ingests spatial data and outputs analysis results in the form of spatial and non-spatial database tables. The UrbanData analysis results can be visualized using desktop and web mapping application like QGIS, ArcGIS Pro, or ArcGIS Online.

The UrbanData application is built using PostgreSQL, PostGIS, and Python. It is composed of a set of PL/SQL functions, Python scripts, SQL views, and SQL materialized views. PostgreSQL is an open source feature-rich object-relational database that has been in active development for several decades. The PostGIS spatial plugin enables spatial functionality within PostgreSQL for storing and managing spatial data and conducting spatial operations such as distance, within, and cluster. All spatial data are stored in a PostGIS database, and all the spatial operations (e.g. proximity) are executed using PostGIS functions. Python is used in conjunction with PostgreSQL to download spatial framework data from open data portals via the Python ogr2ogr library. The Python NLTK library is also used to measure topicality between spatial data sets. The multi-step data analysis process outlined Figure 3.1 generates GR scores by comparing the input data set against each data set in the framework database using the topicality, spatial proximity, cluster co-location, and co-location modules. GR scores calculated for each framework data layer are used to rank and filter framework data layers. The five layers with the highest geographic relevance score are selected to generate the spatial domain database. The domain database is then mapped with the input layer to verify if the UrbanData platform correctly identifies data sets that provide context for the input data set and if the UrbanData results are ranked appropriately. The following sections describe how each of the four GR metrics in UrbanData are implemented. The description of each metric is accompanied by examples from the case study described in section 3.4.

3.3.2 Topicality

Topicality is a relevance metric that evaluates the conceptual similarity of one spatial data set to the conceptual similarity of another data set. The semantic definition of spatial data sets has been described in the literature as a combination of feature type, attribute information and topological characteristics (Adams et al., 2007; Hacar & Gökgöz, 2019; Yu et al., 2016). UrbanData defines the semantic definition of a spatial data set based solely on the feature type. Attribute information is not considered in this analysis because the analysis of attribute data across large data sets is computationally expensive. Topological characteristics are handled by other spatial metrics defined in this methodology. This thesis defines topicality as a measure of the similarity between the feature types of two layers. Measuring semantic similarity between two words or concepts often relies on the use of an ontology (Ballatore & Bertolotto, 2011, 2018; Codescu et al., 2011). This methodology makes use of the WordNet ontology that is included in the Python WordNet library. To measure topicality, the feature type of the input layer and the target layer are matched to terms within the WordNet ontology. This process often requires the use of mapping or generalization techniques (Ramos et al., 2014). Once the layer descriptors are mapped to the ontology, the similarity is measured between the two concepts by

measuring the distance between nodes in the ontology graph (Liu, Bao, & Xu, 2012; Pedersen, Patwardhan, & Michelizzi, 2004). Topicality is measured in the UrbanData platform using the Python NLTK WordNet package. This library was developed to measure the similarity or relatedness of words within the WordNet database (Liu et al., 2012; Pedersen et al., 2004) and continues to be used widely for semantic analysis (Nguyen, Richards, Chan, & Liszka, 2016; Zhu & Iglesias, 2017). The NLTK similarity measure accepts two words as input and returns a similarity score between 0 and 1.

The NLTK similarity measure is limited to words within the WordNet library. The topicality metrics measure the semantic similarity of spatial data sets using layer names; however, layer names are retrieved from Open Data portals that often use names that are not found in the WordNet library. Therefore, every layer is manually assigned a layer name and a set of tags that describe the features within the layer; the use of multiple tags ensures that layers can be mapped to terms in the WordNet library. When comparing two layers, all the tags from the input layer are compared to all the tags in the target layer, and the highest NLTK relevance score is used. This approach to measuring similarity attempts to address gaps in data in the WordNet library and differences in the way terms are conceptualized in the WordNet library.

3.3.3 Spatial Proximity

The proximity metric (proximity) is the most well known and widely used relevance metric in GIS. This metric is based on the underlying assumption that "a purported fact should be consistent with its geographic context" (Goodchild & Li, 2012, p. 115). The proximity metric is also rooted in Tobler's first law of geography, which broadly states that objects nearby are more related than objects that are distant (Tobler, 1970). Measuring proximity between two layers requires some form of generalization because spatial data sets are composed of numerous features that vary in concentration and distribution, while distance is measured between two points. Measuring distance between the nearest points or features of two layers ignores the distribution of features in both layers and is prone to false positives. For example, road networks from neighboring cities are likely to connect which would result in a distance measure of 0, but the layers would have very different spatial coverage. In contrast, a tree data set and a road data set in the same city may not have any features that touch which would result in a distance measure greater than 0 but the data sets have similar spatial coverage. The UrbanData spatial proximity metric is intended to represent a measure of distance that best represents proximity between two layers based on the distribution of features. Aggregate geometries such as layer extents and layer centroids represent the general distribution and concentration of spatial features in a data set. Using

aggregate geometries to calculate proximity is an inexpensive approach to measure the distance between layers based on the spatial distribution of features within the layer. Proximity can be measured using the minimum cartesian distance between two layer extents, but this approach does not work well when layer extents intersect. For example, a bridge layer may have four individual features spread across a city, which gives the bridge layer a very large extent. In contrast, a trail layer may have a large number of features concentrated in the center of the city, thus creating a small extent that lies within the larger extent of the bridge layer. The closest distance between these two layers is 0 because one layer contains the other, but the individual features of the two layers may not be close to each other.

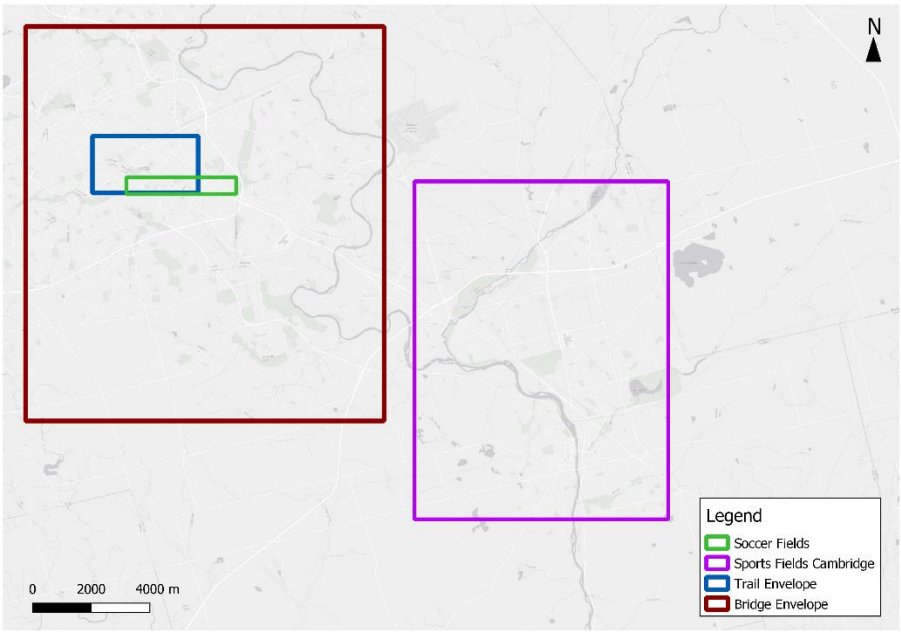

Figure 3.2 Centroid-based proximity measures differentiate trails from soccer fields.

Figure 3.2 above illustrates the challenge of measuring the proximity between three layers with different extents. In a predefined study extent, it is possible for all layers being analyzed to have a separation distance of 0. To better differentiate proximity scores between layers, UrbanData makes use of the layer extent centroid to measure distance. Measuring proximity using the centroid of layers ensures that layers that are within another layer still have a meaningful proximity score. The centroid based approach to measuring the proximity between layers has been explored in the Geoportal platform using a similar center metric (Florance et al., 2015). The use of centroids does have disadvantages when comparing layers with substantial differences in coverages an extent. For example, a national dataset may have a lot of features that are near a provincial data set, but the proximity score

will be low because the distance between layer centroids is high. Despite this caveat, the centroid-based proximity metric does provide an ideal balance between small computational expense and representativeness of proximity. Furthermore, the co-location and cluster co-location described in section 3.3.4 and 3.3.5 define relevance metrics that are more representative of the relationship between individual layer features. Thus, the spatial proximity score in UrbanData is a normalized measure of the distance between the centroids of two-layer extents where a distance of 0 is scored as 100% and distances beyond a user-defined threshold is scored as 0%.

The spatial proximity metric is not expected to generate false positive results in this case study because the study layers are municipal data sets that are reasonably comparable in size but vary in distribution. The proximity score is generated from distance measurements using an inverse distance function that relies on a threshold distance. This metric assumes that spatial objects beyond located beyond a given proximity are irrelevant. The threshold distance is determined by the user based on the study extent. The UrbanData application uses a threshold distance of 5 km because the study extent is the City of Kitchener which ranges between 10 km and 15 km; a proximity distance beyond 5 km likely indicates that the layer being analyzed is generated by a neighbouring municipality. These distance assumptions are rough approximations based primarily on the judgment of the author and an approximate measure of the study extent. The formula below is used to normalize the distance between two spatial data sets to a normalized score between 0 and 1.

$$SP_{AB} = (dist_{max} - dist_{AB}) / dist_{max}$$

SP_{AB} = Spatial Proximity score between layer A and B

$Dist_{AB}$ = Distance between layer A and B (meters)

$dist_{max}$ = Threshold distance (meters)

The function above is a simplification of the distance decay function used in the GR literature (De Sabbata, 2013). The GR literature uses an exponential distance decay function to generate spatial proximity score; an exponential distance decay function assumes that relevance decreases exponentially as distance increases. The purpose of this thesis is to provide a proof of concept for using GR metrics in the context of determining relevance between spatial data sets; thus, a simple implementation of distance decay is sufficient for the scope of this thesis. The formula shown above is implemented in the UrbanData platform using a PostGIS PL/SQL function that generates a spatial envelope for each layer to

identify each layer centroid. The distance between layer centroids is then determined using the St_Distance function.

```
//Generate layer centroids
SELECT
    st_envelope(st_union(st_transform(st_makevalid(geom),26917)))
FROM %I.%I
```

```
SELECT ST_Distance(input_layer.centroid, target_layer.centroid)
```

The St_Envelope function is used in conjunction with the St_Centroid and St_Distance function to determine the distance between two layers. The normalization function takes the distance between two layer centroids and returns a score between 0 and 1.

3.3.4 Co-location

Co-location is a measure of how often feature A can be found within a predefined distance of feature B (Reichenbacher et al., 2016). The same reasoning motivates the co-location metric as the spatial proximity metric where near objects are considered to be more relevant than distant objects. The co-location metric makes it possible to better understand how features within two different layers are positioned in relation to one another. The co-location metric also touches upon the semantic characteristics of a layer by detecting location patterns of a spatial data set that can be used to describe a spatial data set. For example, it can be said that features in the tree layer are commonly observed within parks features.

The literature defines co-location as a spatial pattern that may or may not have statistical significance. The co-location pattern identifies spatial features that are often located near one another (Barua & Sander, 2014; Yoo & Shekhar, 2006). Co-location is a rule that infers the presence of one category of features based on the presence of another category of features (Chen, Zhang, Deng, Nie, & Yi, 2017; Yoo & Shekhar, 2006). In the literature, co-location rules can be identified using data mining or statistical analysis (Barua & Sander, 2014; Huang, Shekhar, & Xiong, 2004). Co-location can infer correlation based on the methodology used to identify the co-location pattern. In UrbanData, co-location is not identified using statistical methods, and co-location is not used to infer correlation. This methodology also does not generate co-location scores. Instead, this methodology makes scores based on observed co-location patterns. Co-location patterns are identified between pairs of layers in this methodology. Co-location patterns are detected through data mining and the use of a threshold distance. The threshold distance is a user-defined distance that makes it possible to identify co-location

patterns (Deng, He, Liu, Cai, & Tang, 2017). The threshold distance is determined based on the study context and how the user perceives near objects. For a mobile user who is walking, a visible distance of 50 m would likely be considered near while a car driver would find 1 km near. In UrbanData, a threshold distance of 50 m is used to mine co-location patterns. This distance is assumed to be a small distance for most pedestrians, thus considered to be close. Co-location between two layers (A, B) is measured as the number of features in layer A that are within a threshold distance of layer B divided by the total number of features in A.

$$Co_{AB} = N_{AB}(d)/N_A$$

Co_{AB} = Co-location score between layer A and B

$N_{AB}(d)$ number of features in layer A within d (distance) of layer B

N_A number of features in layer A

The co-location metric is implemented in the UrbanData platform using a set of PL/SQL functions and views. In the formula above, co-location is mined on the domain layer with regard to a target layer within the framework database. The input layer is buffered using the PostGIS St_Buffer function with the threshold distance as input. The buffered input layer is then intersected with the target layer using the St_Intersects function to determine how many features from the input feature layer are within the threshold distance from features in the target layer. For example, if 2 out of 10 trail features are within 50 m of a park, the co-location score is 0.2 (20%).

```
st_buffer(input_layer.geom, %L)
st_intersects(input_layer.geom, target_layer.geom)
```

The co-location score is then calculated by dividing the number of input features within the threshold distance of the target layer by the total number of features in the input layer. The result of this analysis is a normalized score between 0 and 1. This score is then used to create the aggregate GR by combining it with other GR metrics. The following section describes the cluster co-location metric, which uses parts of the co-location metric to evaluate spatial relevance.

3.3.5 Cluster Co-location

The cluster co-location metric is a novel GR metric. The cluster co-location metric is based upon the cluster metric defined in the GR literature (Reichenbacher et al., 2016); but it adjusts the metric to enable comparisons between spatial data sets. The cluster metric is used to detect spatial location

patterns within a layer. The cluster metric in GR is developed on the idea that points of interest that are close to other points of interest are more important than isolated points of interest (Reichenbacher et al., 2016). For example, a restaurant in a mall that is located near other shops is more attractive to a shopper than an isolated restaurant. The cluster co-location metric proposed in this thesis evaluates the location of clusters in the input layer in relation to features in the target layer. This metric assumes that two layers are relevant to each other if features from the input layer tend to cluster around features in the target layer. Thus, the metric aims to measure how often clusters in the input layer are co-located with features in the target layer. For example, food stands may cluster around park entrances and baseball field; therefore, parks and food stands have a high cluster co-location score. Like the co-location metric, the cluster co-location metric aims to detect patterns in spatial data sets to determine relevance between two data sets.

The cluster co-location metric determines relevance by measuring the number of clusters in the input layer that are within a threshold distance of features in the target layer. The cluster co-location formula generates a normalized cluster co-location score by dividing the co-located clusters in the input layer by the total number of clusters in the input layer.

$$ClCo_{AB} = Cl_{AB}/Cl_A$$

$ClCo_{AB}$ = Cluster Co-location Score between layer A and layer B

Cl_{AB} = Clusters from layer A within distance **d** of features from layer B

Cl_A = Total clusters from layer A

Clusters are identified in the input layer using the DBSCAN algorithm, this approach has been used in GR and data contextualization studies (Reichenbacher et al., 2016; Spinsanti & Ostermann, 2013). The DBSCAN algorithm is effective for this use case because the algorithm requires little domain knowledge, creates clusters of arbitrary size and is efficient on databases (Sander, Ester, Kriegel, & Xu, 1998). Co-location between input layer clusters and target layer features is determined by generating an envelope with a one-meter buffer around clustered features. Input layer clusters that intersect target layer features are added to the list of co-located clusters; this number of co-located clusters is then divided by the total number of clusters in the data set. The UrbanData platform implements this metric by using the PostGIS st_clusterdbscan function to detect clusters in the input layer. The st_convexhull, st_collect and st_buffer function are then used to create geometries for clusters. The st_clusterdbscan function works on lines, points, and polygons to determine if a feature is part of a cluster. The st_buffer

function standardizes geometries to a polygon type, and the st_collect function aggregates geometries to a multipolygon. The st_convexhull generates a single polygon that envelopes all the geometries within the cluster. The st_intersects function is then used to detect clusters that are co-located with target layers features.

```
st_clusterdbscan(input_table.geom, 50)
```

```
st_convexhull(st_collect(st_buffer(ham.geom, (1)::double precision)))
```

The result of the cluster co-location module is a normalized score that ranges from zero to one where a score of zero indicates no clusters are co-located with target features while a score of 1 indicates that all input layer clusters are co-located with target layers. The following section describes how the metrics discussed in this section are aggregated to generate a GR score.

3.3.6 Geographic Relevance Score

The previous sections have discussed the relevance between an input domain layer and target framework layers. The topicality, spatial proximity, co-location, and cluster co-location metrics are used to measure dimensions of spatial and semantic relevance between a pair of data sets and produce a normalized score between 0 and 1. This section discusses the aggregation of these four metrics into a single GR score that represents the degree of relevance between the input layer and the target layer.

Comparable studies in GR, conflation, and GIR have approached the challenge of combining relevance metrics with varying strategies. McKenzie et al. (McKenzie et al., 2014) developed multiple weighting models to combine semantic and spatial metrics to maximize model accuracy. Bordogna et al. (2012) built two aggregation models to combine semantic and spatial data; where one model favored semantic relevance while the other model weighed semantic and spatial relevance equally. Reichenbacher et al. (2016) used a conjunctive score model that considered some geographic relevance metrics as essential. Therefore, metrics such as spatial proximity had to be higher than 0 for the GR score to be valid. This thesis uses an equal weighting model to generate GR scores. The GR metrics described in the previous sections have produced normalized scores between 0 and 1; an equal weighting model is essentially the mean of the four GR metrics. Current GR literature has advocated against equal weighting models for score aggregation as some metrics are more important than others (Reichenbacher et al., 2016). UrbanData uses an equal weighting model because it is difficult to prioritize one metric over another, and there are no clear essential metrics.

Each metric described in this methodology only describes a small dimension of the entities being compared. The topicality metric considers layer names and tags; it does not consider the attributes of the layer or individual features. The spatial proximity measures proximity based on extent centroids, which may create false positives in some contexts. The co-location and cluster co-location metric measure similarity of feature distribution between two layers but the validity of these metrics is dependent on the effective selection of threshold distances. Due to the scale of the analysis and the gaps in the individual GR metrics; this thesis assumes that an equal-weighted GR score provides the most reliable approach to measuring GR. Thus, the GR metric is defined as follows.

$$GR_{AB} = (T_{AB} + SP_{AB} + Co_{AB} + ClCo_{AB})/4$$

GR_{AB} = Aggregate geographic relevance score between layer A and B

T_{AB} = Topicality score between layer A and B

SP_{AB} = Spatial Proximity score between layer A and B

Co_{AB} = Co-location score between layer A and B

$ClCo_{AB}$ = Cluster Co-location score between layer A and B

In the UrbanData platform, the GR metric and the four underlying relevance component values are stored in a Postgres table that includes the input layer name and each target layer name.

3.4 Case Study: City of Kitchener Open Data

The UrbanData application was tested using open data from the City of Kitchener, a mid-sized city located in southern Ontario with a population of 220,000 and an emerging technology industry. The City of Kitchener maintains a joint open data portal that hosts data from the City of Kitchener, City of Waterloo, Cambridge and the Region of Waterloo (https://open-kitchenergis.opendata.arcgis.com). The Open Data portal is named the Kitchener GeoHub, and it contains approximately 300 open data sets with 267 spatial data sets as of May 29, 2019. The portal hosts data ranging over a wide array of topics which include the environment, infrastructure, municipal services, planning, points of interest, non-spatial records, sports and recreation, and transportation. The Kitchener GeoHub leverages ESRI's ArcGIS Online platform, particularly the ArcGIS Hub. The Kitchener GeoHub is unique from other open data portals such as the City of Toronto Open Data portal or the New York Open Data portal because three different municipal agencies are contributing to a single data portal. As a result, the Kitchener

GeoHub contains seemingly redundant data sets that are maintained by different agencies. A user that searches for roads will be matched to four distinct yet similar road data sets which contain similar data but contain different attribute data and cover different extents.

The UrbanData application was tested using a sample of 13 data sets retrieved from the Kitchener Geohub and one VGI planning data set provided by the City of Kitchener. The analysis evaluates the effectiveness of UrbanData at retrieving domain data for a planning study conducted by the City of Kitchener on improving major trails in 2015. The VGI data set used in this thesis was collected by the City of Kitchener to support the planning study on major trails. The analysis is constrained to 5 domain data sets related to the planning study because the GR metrics used by UrbanData are computationally expensive and time intensive; moreover, verifying the results of the study across a large number of data sets with a large number of outputs is not practical. The number of data sets used in this thesis is not entirely indicative of the computational complexity of this thesis as data sets with many features exponentially increase the complexity of analysis.

The test uses 14 data sets that include boundaries, roads, trails, VGI data, bridges, and related features. The data selected from the Kitchener GeoHub include point, line and polygon geometries, and the data vary in distribution across Kitchener and Cambridge. The data was selected based upon the perceived relevance to the Kitchener trails, but a few data sets were also selected to create noise in the analysis. The VGI data provided by the city are 505 georeferenced survey responses that were collected from residents using an interactive web mapping application that allowed users to make comments about specific locations across the city. The VGI data set contains points that are generally located near major trails in the core of the City of Kitchener. The data sets are shown in Table 3.1, the data steward is the organization that maintains and publishes the data. Data tags are used to describe each data set semantically. The data tags are used to conduct topicality analysis. The domain data column identifies layers that are used as domain data sets for the analysis. The feature count indicates the number of features in the layer. Larger layers are more expensive to analyze due to the number of features.

Table 3.1: Data sets from Kitchener Open Data Portal and City of Kitchener Planning department.

ID	Data Steward	Data Set	Data Tags	Feature Count	Domain Data
1	City of Cambridge	Business Improvement Areas	business, downtown, uptown	3	
2	City of Kitchener	Heritage Districts	heritage, district	4	
3	City of Kitchener	Soccer Fields	soccer, field	8	TRUE
4	City of Kitchener	Parking Public Lots	parking, public parking, lot	50	
5	City of Kitchener	Rivers Creeks Ponds and Lakes	rivers, creeks, ponds, lakes	122	
6	City of Cambridge	Sports Fields	sports, field	134	
7	City of Kitchener	Bridges	bridges	135	
8	City of Kitchener	Parking On Street	street parking, parking	157	
9	City of Kitchener	Railway Lines	rail, railway	181	TRUE
10	City of Kitchener	Parks	park	438	TRUE
11	Kitchener Planning	Trail Survey	trail, survey	505	TRUE
12	City of Kitchener	Cycling Infrastructure	cycling, bike, lane	1093	
13	City of Kitchener	Trails	trail	3530	TRUE
14	City of Kitchener	Roads	roads	6057	

Despite the small number of data sets being used, a large number of features are analyzed. There are over 12,000 features in the study database; for each analysis, the system has to traverse and analyze all the features in the database. The domain data sets in Table 3.1 are compared to each data set in Table 3.1 using the GR metrics described in section 3.3. The domain data sets used in the analysis are listed below.

- Soccer Fields
- Railway Lines
- Parks
- Trail Survey
- Trails

The trail survey is the VGI data set provided by the Kitchener planning department. This data set is primarily concentrated in the core of Kitchener and contains several clusters of points. The parks and trail data sets have large extents that spread across the City of Kitchener and contain several thousand features together. The railway data set is a smaller data set that is relevant to the trail because there are several major intersections between the rail and trail and several records in the trail survey reference

rails. The Kitchener soccer fields data set is not directly related to major trails or the trail survey, it is primarily used as noise to confuse the UrbanData system and test the effectiveness of GR metrics.

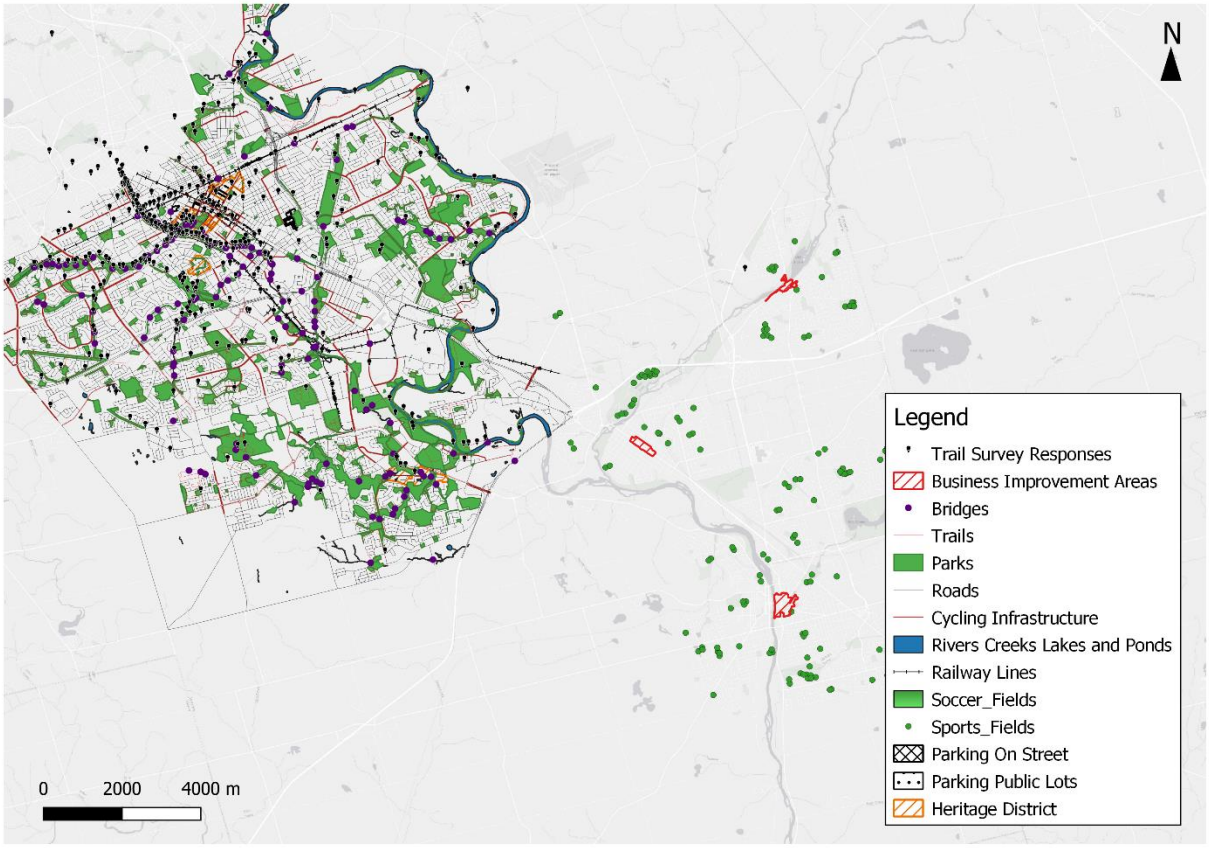

Figure 3.3 A map of all the layers in the framework database, layers are in Kitchener and Cambridge.

Each analysis requires aggregate and feature level analysis between the input data and the other 13 datasets in the database. The result of each analysis is a table with a ranked list of layers that include individual GR metric scores and aggregate metric scores. The output tables provide metadata that is used to rank the relevance of framework data with respect to domain data. In this test, the domain database is a set of the five layers with the highest GR ranks with respect to a domain data set. UrbanData is a data processing application that only outputs analysis tables into the Postgres database. Results can be visualized using mapping applications and framework such as QGIS and ArcGIS JS. This analysis uses QGIS to visualize GR rankings and domain databases to evaluate outputs. The following sections discuss the results of this analysis.

3.5 Results

Analysis of each data set resulted in the creation of an output table that contains analysis results between the input (domain) layer and target framework data sets. Each analysis produces 13 sets of output records that contain a topicality score, a spatial proximity score, a co-location score, a cluster co-location score, and an aggregate geographic relevance score. Section 3.5.1 will review individual analysis results for each domain data set. For each analysis, the output data scores will be presented in a table, and the five highest ranking datasets will be mapped as the domain database. The review will also discuss the distribution of GR scores. Review of individual analysis results is followed by a review of overall results and trends among the GR scores in section 3.5.2.

3.5.1 Domain Data Analysis

3.5.1.1 Parks

The UrbanData analysis of Kitchener Parks resulted in the selection of roads, trails, rivers, cycle routes, and bridges in the domain database. Figure 3.4 shows that features in the layer are often located near other features in the domain database, all the domain data sets have similar coverage and represent features that are relevant to each other such as creeks and parks or cycle routes and parks.

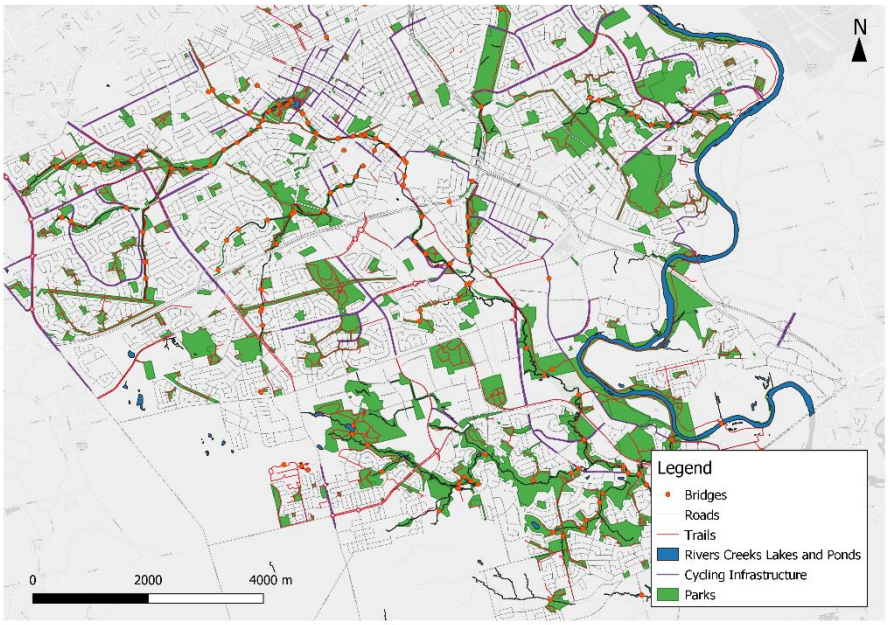

Figure 3.4 The domain database for parks includes trails, bridges, roads, rivers, and cycle routes.

The layers in the domain database consistently intersect the parks layer, layers such as rivers, bridges, and trails are often contained within park features, and the roads and cycle network layers connect

parks. The results in Table 3.2 show that the highest GR score is 0.71, and the lowest domain database score is 0.45, while the lowest overall score is -0.35. The two most relevant data sets, roads, and trails are large data sets that have high co-location and cluster co-location scores with parks which is reasonable because trails are often located within parks while parks are often connected to roads. The proximity, co-location, and cluster co-location scores generally have a high degree of variance between the most relevant layers and the least relevant layers, and they are good indicators of geographic relevance.

Table 3.2 UrbanData analysis for the parks layer against 13 framework data sets.

Rank	Domain Table (Input)	Framework Table (Target)	GR Score	Proximity	Co-location	Co-location Cluster	Topicality
1	Parks	Roads	0.71	0.77	0.97	0.94	0.14
2	Parks	Trails	0.70	0.99	0.76	0.94	0.13
3	Parks	Rivers Creeks Ponds and Lakes	0.55	0.92	0.41	0.78	0.10
4	Parks	Cycling Infrastructure	0.52	0.94	0.31	0.72	0.10
5	Parks	Bridges	0.45	0.83	0.22	0.61	0.13
6	Parks	Trail Survey	0.43	0.60	0.26	0.72	0.13
7	Parks	Railway Lines	0.38	0.95	0.11	0.39	0.08
8	Parks	Soccer Fields	0.31	0.86	0.00	0.06	0.33
9	Parks	Heritage Districts	0.25	0.51	0.08	0.22	0.20
10	Parks	Parking Public Lots	0.24	0.79	0.04	0.06	0.08
11	Parks	Parking On Street	0.20	0.65	0.02	0.06	0.08
12	Parks	Sports Fields	-0.31	-1.57	0.00	0.00	0.33
13	Parks	Business Improvement Areas	-0.33	-1.47	0.00	0.00	0.14

The relevance rankings of roads and trails are based on a marginal difference in GR scores, and it can be argued that the topicality score between parks and trails should be higher than the topicality score between parks and roads. The topicality score generally shows little useful variation between layers, and it is difficult to explain how the NLTK WordNet module assigns a score. The lower ranked data sets in the domain database have lower co-location scores, but proximity and topicality scores are comparable to the roads and trails layer. The analysis correctly considered cycling infrastructure and bridges more relevant to parks than on-street parking or public parking lots. The proximity score is particularly effective at reducing the relevance score of distant layers located in the City of Cambridge such as Cambridge Sports Fields and Business Improvement Areas. The proximity score is also very useful for

differentiating the relevance of layers when co-location and cluster co-location scores are 0 for multiple layers. Overall, the topicality score showed little discernible variance across the analysis while the proximity, co-location, and cluster co-location where effective tools to determine the relevance of layers. The rankings and scores of the layers in the analysis are reasonable.

3.5.1.2 Railway Lines

Relevance analysis of railway lines shows that parks, trails, cycling infrastructure, rivers, and roads are the most geographically relevant layers to Kitchener Railway lines. The railway data set is a smaller data set than parks and features are mainly located in key corridors in the core of the City of Kitchener. Figure 3.5 shows that layers in the domain data set intersect railways, and there are occasions where segments of the rail coincide with parks or roads. As shown in Figure 3.5, railway features are located along key corridors through the center of Kitchener so they don't touch or intersect framework layers as often as larger data sets like parks.

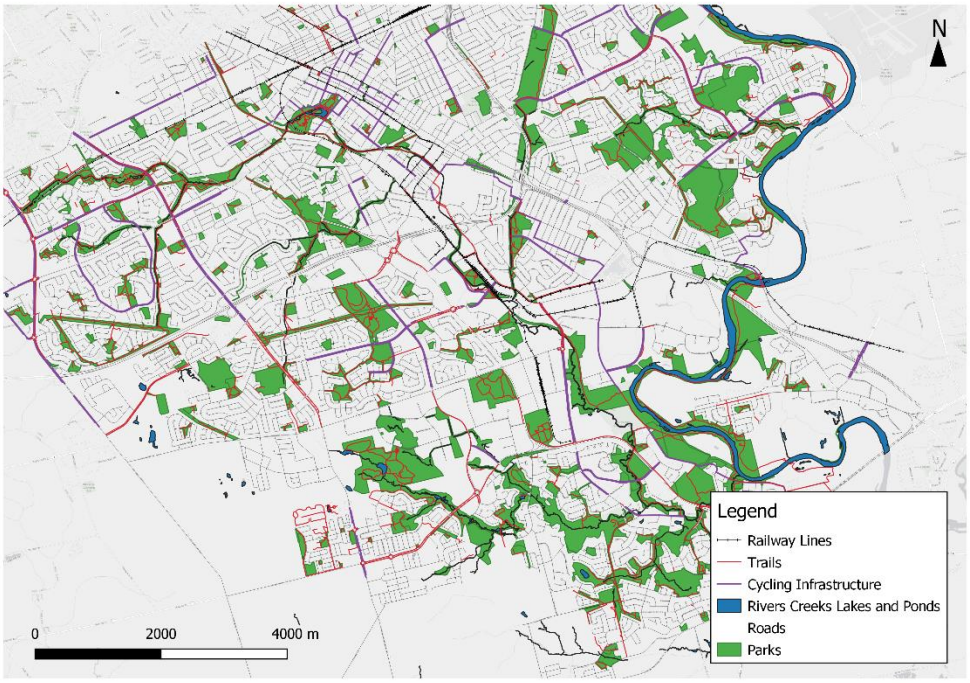

Figure 3.5 The domain database for rails includes parks, rivers, trails, roads, and cycle lanes.

The GR scores shown in Table 3.3 reflect the visual trends shown in Figure 3.5, the GR scores between the railway dataset and framework layers are generally lower than the GR scores between parks and framework data sets. The highest GR score from the analysis is 0.54, and the lowest domain database score is 0.51, the difference in relevance scores of the top five layers is marginal. The

UrbanData scores in Table 3.3 appear to be poor indicators of relevance. The low co-location scores shown in Table 3.3 reflects the concentration of railway features along central corridors within Kitchener. The co-location cluster score is 1 for all framework data sets in Kitchener. This likely occurred because UrbanData identified one large cluster in the railway dataset that happened to intersect all the framework layers from the City of Kitchener. The lack of variance in the co-location cluster scores makes it an ineffective relevance metric for the evaluation of this dataset. The topicality score also shows little variance across the analysis with 10 out of 13 layers receiving scores of 0.08 or 0.09. Thus, topicality is an ineffective relevance measure once again. Proximity is the only GR metric that shows notable variance across the data, the highest proximity score is 0.95, and the lowest proximity score for Kitchener data sets is 0.48.

Table 3.3 The UrbanData relevance scores for Kitchener railway lines.

Rank	Domain Table (Input)	Framework Table (Target)	GR Score	Proximity	Co-location	Co-location Cluster	Topicality
1	Railway Lines	Parks	0.59	0.95	0.32	1.00	0.08
2	Railway Lines	Trails	0.58	0.73	0.52	1.00	0.08
3	Railway Lines	Cycling Infrastructure	0.57	0.93	0.24	1.00	0.09
4	Railway Lines	Rivers Creeks Ponds and Lakes	0.56	0.89	0.20	1.00	0.14
5	Railway Lines	Roads	0.53	0.90	0.15	1.00	0.09
6	Railway Lines	Soccer Fields	0.49	0.79	0.10	1.00	0.08
7	Railway Lines	Parking Public Lots	0.49	0.87	0.00	1.00	0.08
8	Railway Lines	Bridges	0.49	0.84	0.04	1.00	0.08
9	Railway Lines	Trail Survey	0.48	0.62	0.22	1.00	0.09
10	Railway Lines	Parking on Street	0.45	0.67	0.03	1.00	0.08
11	Railway Lines	Heritage Districts	0.40	0.48	0.03	1.00	0.10
12	Railway Lines	Business Improvement Areas	-0.31	-1.47	0.00	0.00	0.25
13	Railway Lines	Sports Fields	-0.37	-1.57	0.00	0.00	0.08

The GR rankings of layers in relation to railways appear reasonable in some cases, but several layers are ranked incorrectly. The road data set should be considered the most relevant data set to

railways because railway features are often located near roads. The roads layer has the highest co-location score with rails and roads likely has the strongest topical relevance to layers due to their function as transportation networks. Layers such as bridges or parking lots should have higher relevance ranks than soccer fields. Railways intersect bridges while parking lots are often located near rail stations. In general, UrbanData was not effective at determining relevance between railways and framework layers.

3.5.1.3 Soccer Fields

Like the railway layer, the soccer fields layer does not touch or intersect features in the framework database very often. Figure 3.6 shows that the soccer field layer only contains a few features located near the center of Kitchener. The soccer fields seem to be related to the parks layer as several soccer fields are located within parks. Figure 3.6 shows that many of the layers selected in previous domain databases have been selected for this map. The roads layer is a notable omission in Figure 3.6 since roads likely connect to soccer fields.

Figure 3.6 The domain database for soccer fields includes rivers, trails, rails, parks, and cycle lanes.

The relevance scores in Table 3.4 contain the lowest relevance scores generated by UrbanData. The highest GR score is 0.30, the lowest domain database score is 0.22, and the lowest Kitchener data score is 0.14. The co-location and cluster co-location scores are 0 across the analysis. Proximity and topicality are the primary metrics used to determine relevance. The topicality score shows greater variance than previous results, and it helps UrbanData prioritize parks and bike routes over railway lines and rivers. The overall GR scores are low in this analysis, but the relevance metrics work as desired. Cluster and cluster co-location are not influential in this analysis, but that is reasonable considering the spatial relationship between soccer fields and other layers. The proximity and topicality metrics work well to prioritize relevant layers such as Kitchener Parks and filter out layers that are topically relevant but spatially irrelevant such as Cambridge Sports Fields.

Table 3.4 The UrbanData relevance scores between soccer fields and framework data.

Rank	Domain Table (Input)	Framework Table (Target)	GR Score	Proximity	Co-location	Co-location Cluster	Topicality
1	Soccer Fields	Parks	0.42	0.86	0.50	0.00	0.33
2	Soccer Fields	Cycling Infrastructure	0.35	0.64	0.63	0.00	0.14
3	Soccer Fields	Trails	0.25	0.81	0.00	0.00	0.20
4	Soccer Fields	Railway Lines	0.24	0.85	0.00	0.00	0.13
5	Soccer Fields	Rivers Creeks Ponds and Lakes	0.24	0.87	0.00	0.00	0.08
6	Soccer Fields	Parking On Street	0.22	0.78	0.00	0.00	0.10
7	Soccer Fields	Parking Public Lots	0.21	0.78	0.00	0.00	0.08
8	Soccer Fields	Roads	0.21	0.76	0.00	0.00	0.08
9	Soccer Fields	Bridges	0.19	0.70	0.00	0.00	0.08
10	Soccer Fields	Trail Survey	0.16	0.50	0.00	0.00	0.13
11	Soccer Fields	Heritage Districts	0.14	0.37	0.00	0.00	0.20
12	Soccer Fields	Sports Fields	-0.17	-1.70	0.00	0.00	1.00
13	Soccer Fields	Business Improvement Areas	-0.36	-1.60	0.00	0.00	0.14

3.5.1.4 Trail Survey

The trail survey data set is a VGI data set collected by the Kitchener planning department. The trail survey data set does not represent any spatial features or objects. Instead, it represents geotagged survey responses from citizens of the City of Kitchener. The survey was collected to gather feedback on trails, and many of the responses are located along major trails in the center of Kitchener, but a large number of responses are located around smaller trails, parks, and water bodies across Kitchener. Several responses are also located along roads and cycling networks because many of the survey respondents were interested in improving the connectivity of trails across the city. Figure 3.7 shows the domain database for the trail survey data set. The map shows that the trail survey features are often located within a park, near waterbodies or linearly grouped along roads, trails or bicycle networks.

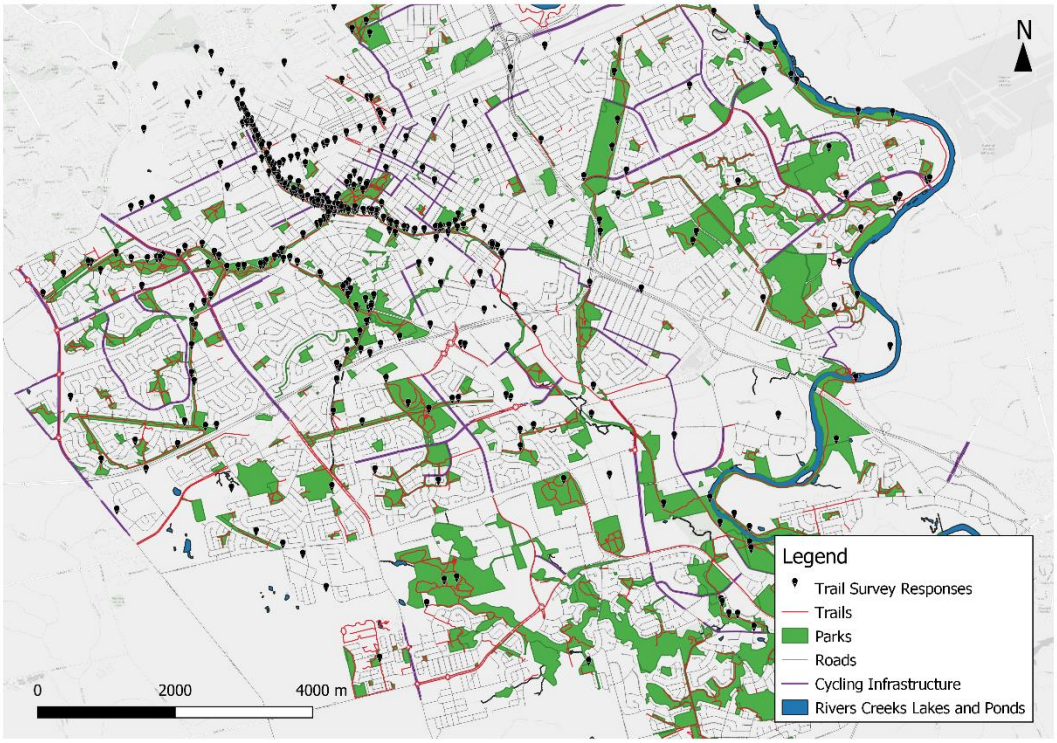

Figure 3.7 Trail survey is most relevant to trails, parks, roads, bikes, and rivers.

The relevance scores in Table 3.5 show more variance than previous results with the highest GR score being 0.79, and the lowest domain database score being 0.30. The lowest score for Kitchener data is 0.11. The variance in proximity scores is low in this analysis with 8 of 13 layers receiving a proximity score between 0.60 and 0.70; the proximity score also decreases as the GR score increases. The inverse relationship between proximity and GR is reasonable because topicality, co-location, and cluster co-

location scores can be high regardless of the distance between layers. The co-location cluster metric shows a high degree of variance, and they indicate that trail survey points are primarily located near parks and trails, but points can also be found near roads, rivers, and cycle lanes. The topicality score has little variance in this analysis except for the trail data set, which has a score of 1 because the trail survey and trail data set share a common tag. The topicality metric allowed UrbanData to prioritize trails, but it did little to determine the relevance of other layers. Overall, the co-location and the cluster co-location where the most important metrics for differentiating relevance of data set. The topicality metric was useful to a limited extent, and the proximity metric was useful for filtering out very distant data sets from the City of Cambridge, but it did not help with ranking domain database layers.

Table 3.5 UrbanData relevance scores for the Kitchener Trail Survey

Rank	Domain Table (Input)	Framework Table (Target)	GR Score	Proximity	Co-location	Co-location Cluster	Topicality
1	Trail Survey	Trails	0.76	0.60	0.64	0.82	1.00
2	Trail Survey	Parks	0.57	0.60	0.69	0.86	0.13
3	Trail Survey	Roads	0.49	0.61	0.54	0.68	0.13
4	Trail Survey	Rivers Creeks Ponds and Lakes	0.35	0.66	0.26	0.36	0.11
5	Trail Survey	Cycling Infrastructure	0.29	0.61	0.17	0.27	0.13
6	Trail Survey	Railway Lines	0.27	0.62	0.13	0.23	0.09
7	Trail Survey	Bridges	0.25	0.62	0.09	0.18	0.09
8	Trail Survey	Heritage Districts	0.23	0.53	0.10	0.14	0.17
9	Trail Survey	Parking Public Lots	0.20	0.70	0.01	0.00	0.08
10	Trail Survey	Soccer Fields	0.16	0.50	0.00	0.00	0.13
11	Trail Survey	Parking On Street	0.11	0.35	0.01	0.00	0.08
12	Trail Survey	Business Improvement Areas	-0.25	-1.12	0.00	0.00	0.13
13	Trail Survey	Sports Fields	-0.27	-1.22	0.00	0.00	0.14

3.5.1.5 Trails

The domain database for the trail and trail survey data are identical, the domain database map for trails is shown in Figure 3.7. The trail data set is a large data set like parks that cover the extent of the City of Kitchener. Figure 3.7 shows that trail features are distributed throughout the City along waterbodies or roads and within parks, trails often connect to cycle routes and trail survey features tend to cluster around trail features. Table 3.6 shows that GR scores between trails and framework data sets

are high for domain data sets, and there is a high degree of variance across the analysis. The three highest relevance scores are marginally different because scores range from 0.74 to 0.71. The lowest relevance score in the domain database is 0.58. The proximity scores generally increase as relevance increases, but there are several cases where layers with low GR scores such as railways have high proximity scores while layers with high GR scores like the trail survey have low proximity scores. This is not unreasonable because other relevance metrics may be more influential in this UrbanData analysis. The co-location metric indicates that six framework layers have perfect co-location scores with the trail's layers, this has not been observed in other analysis, and it does not seem reasonable. It is possible that trail geometries are stored as large continuous multi-polygon features that may skew the co-location metric. This is the first analysis where the co-location cluster metric does not follow the same score patter as the co-location metrics. The co-location cluster metrics has a high degree of variance and it helps UrbanData marginally prioritize roads over the trail survey data. The topicality once again shows little variance with the exception of the trail survey score and it appears to be an ineffective measure of relevance.

Table 3.6 The UrbanData relevance results for Kitchener trails.

Rank	Domain Table (Input)	Framework Table (Target)	GR Score	Proximity	Co-location	Co-location Cluster	Topicality
1	Trails	Parks	0.70	0.99	0.81	0.87	0.13
2	Trails	Roads	0.65	0.79	0.70	0.98	0.13
3	Trails	Trail Survey	0.50	0.60	0.15	0.26	1.00
4	Trails	Cycling Infrastructure	0.42	0.95	0.20	0.40	0.11
5	Trails	Rivers Creeks Ponds and Lakes	0.39	0.93	0.24	0.28	0.11
6	Trails	Bridges	0.32	0.84	0.12	0.21	0.09
7	Trails	Railway Lines	0.29	0.93	0.05	0.10	0.09
8	Trails	Heritage Districts	0.25	0.85	0.00	0.02	0.13
9	Trails	Parking Public Lots	0.23	0.78	0.01	0.06	0.08
10	Trails	Soccer Fields	0.20	0.52	0.06	0.05	0.17
11	Trails	Parking On Street	0.18	0.63	0.01	0.02	0.08
12	Trails	Business Improvement Areas	-0.34	-1.47	0.00	0.00	0.13
13	Trails	Sports Fields	-0.36	-1.56	0.00	0.00	0.13

UrbanData effectively prioritized parks and roads in relation to trails over parking lots and bridges. The high relevance score of the trail survey data set appears to be appropriate but the cycle infrastructure layer should likely receive a higher relevance score based on proximity, co-location, and cluster co-location scores. The GR metrics were generally effective for prioritizing the five most spatially relevant layers to the trail data set and the GR metrics have been consistently effective at assigning low relevance ranks to layer

3.5.2 Summary

Section 3.5.1 looked at how UrbanData analyzed 5 domain data sets against 13 other framework databases to generate a domain database of 5 layers for each domain dataset. The analysis showed how individual GR metrics such as topicality, co-location, cluster co-location, and proximity vary when comparing data sets. The analysis shows that the distribution GR metric can have a small variance between scores or a large variance between score depending on the study context and spatial data characteristics. When a GR metric shows a small range of scores such as a score of 1 for all 13 framework layers, the GR metric is not effective because it cannot be used to identify differences between data an evaluate relevance. When the GR metric has a large range of evenly distributed relevance scores it is possible to differentiate layers and evaluate relevance because the most relevant framework data set will have a score that is significantly higher than the least relevant data set.

The review of data showed that every single GR metric used in UrbanData was important for differentiating layers in at least one of the five analyses. The topicality metric was most important for measuring relevance when evaluating the Kitchener Soccer Field layer. This result supports the assumption that the topicality metric is important for measuring GR and it indicates that the implementation of the topicality metric can be improved to better represent relevance. When UrbanData analyzed the Kitchener Soccer Field data the cluster and co-location score were close to zero for all framework layers and showed little variation between layers. The proximity metric was the most reliable relevance metric in this analysis as it had a consistent variance that could be used to differentiate spatial data. The co-location and cluster co-location metric were very effective for determining relevance for the analysis of parks, trail survey data, and trails but these metrics showed little variation for railways and soccer fields. It also appeared that co-location and co-location cluster may be redundant metrics because both metrics tended to show similar scores in the parks, soccer fields, trail survey and trail analysis. The cluster co-location metric and the co-location metric only had dissimilar score distributions for the railway analysis. In the rail survey analysis, the cluster co-location

score was 1 for all framework data sets and the co-location score ranged between 0 and 0.3 across the analysis. The similarity in scores between co-location and cluster co-location are reasonable because they measure similar data distribution characteristics of spatial data. But, the redundancy of these metrics and the equally weighted aggregation of scores might skew GR scores towards unreasonably prioritizing the distribution of spatial data over other relevance metrics such as proximity and topicality. Table 3.7 and Table 3.8 show summary statistics for the scores generated in the UrbanData analysis of five domain data sets.

Table 3.7 The summary statistics of GR metrics for 65 pairs of layers.

	GR Score	Proximity	Co-location	Cluster Co-location	Topicality
Mean	0.28	0.40	0.22	0.36	0.17
St Dev	0.31	0.82	0.33	0.41	0.19
Min	-0.37	-1.70	0.00	0.00	0.08
Max	0.79	0.99	1.00	1.00	1.00
Median	0.30	0.70	0.04	0.14	0.13
Mode	0.16	0.99	0.00	0.00	0.13

Table 3.7 above shows average scores of each metric along with summary statistics that describe the range and distribution of relevance metrics. The mean score show the average score for each metric while standard deviation, min and max illustrate distribution of scores. High standard deviation figures indicate that the metric has a large distribution of scores that can be used to differentiate layers. Table 3.7 shows that the proximity metric had the highest average score and the largest standard deviation while topicality had the lowest mean and standard deviation. The cluster and cluster co-location metric have comparable mean and standard deviation values but these values don't indicate that the two metrics are redundant. The overall GR score shows an average score and a standard deviation that is most similar to the co-location metric and the cluster co-location metric; this indicates that the co-location and cluster co-location metrics are overweighed. However, as illustrated in Table 3.8, the co-location and cluster co-location metrics are not always identical so there is no conclusive evidence that these metrics are redundant. Future analysis would benefit from conducting analysis between GR metrics to attain a better understanding of the relationship between different GR metrics.

Table 3.8 Average GR metric scores for each domain layer in UrbanData

	GR Score	Proximity	Co-location	Co-location Cluster	Topicality
Parks	0.32	0.44	0.25	0.42	0.15
Railway Lines	0.36	0.43	0.06	0.85	0.10
Soccer Fields	0.14	0.36	0.00	0.00	0.21
Trail Survey	0.25	0.31	0.23	0.27	0.18
Trails	0.35	0.45	0.54	0.25	0.18

Table 3.8 shows the average score for GR metrics for each domain database analysis. The results show that the co-location and cluster co-location metric are important indicators of relevance when comparing large data sets such as trails or parks to other data sets that have a similar size and extent. The topicality metric scores tended to show little variation when evaluating semantic relevance between layers but it was an important for measuring relevance for soccer fields. The effectiveness of the topicality metric in the soccer fields analysis supports the assumption that topicality is an important GR metric. It also indicates that topicality can be a very good indicator of relevance if implemented well. The proximity metric was the most consistent indicator of relevance and it was very important for identifying and filtering out irrelevant data sets. Using layer centroids to measure proximity between layers allowed UrbanData to evaluate proximity between layers that intersected or contained the other layer. The centroid based proximity measure also contributed to generating negative scores for framework layers that were beyond the threshold distance from the study extent. Measuring proximity using distance between layer extents or layer features would have resulted in less variance among proximity scores because most layers had intersecting extents. Therefore, most proximity scores would be zero or close to zero. The results show the identified GR metrics help quantify different dimensions of geographic relevance to create relevance scores that account for differences in spatial data characteristics and the study extent. Proximity accounts for the distance between data sets, topicality accounts for conceptual similarity of data, co-location and cluster co-location account for distribution of spatial features. Each GR metric used in UrbanData is a good indicator of geographic relevance but there are opportunities to improve implementation of these metrics and the UrbanData model.

3.6 Discussion and Conclusion

The growing availability of open data is presenting new needs for methods that find and retrieve spatial data based on spatial criteria. The literature has shown that spatial data retrieval is an important issue, but current platforms do not have sophisticated approaches for determining spatial relevance. Researchers in spatial data quality have referred to this problem as the fitness of use criterion (Wentz & Shimizu, 2018). This Chapter argues that research on geographic relevance is applicable to the problem of spatial data retrieval on open data portals. This research problem has been formalized in this Chapter as generating a domain-specific spatial database using geographic relevance. This thesis adapted measures of geographic relevance which include topicality, spatial proximity, and cluster from Reichenbacher et al. (2016). Cluster co-location is a novel metric of geographic relevance that has been developed in this thesis to compare the spatial properties of large data sets.

The UrbanData platform has demonstrated an approach to retrieve and rank domain specific spatial data using geographic relevance scores. The UrbanData model and the GR metrics worked reasonably well at ranking different types of spatial data but review of the results indicate that there are opportunities to improve the GR model used in UrbanData. The results from this chapter provide some insights about how individual GR metrics behave in different scenarios but larger tests are needed to better understand how metrics like proximity, co-location and cluster co-location handle other edge cases. A better understanding of individual GR metrics can also inform the development of a better GR aggregation model. UrbanData used an equal weighted aggregation model where each GR metric was assigned an equal weight when generating the aggregate GR score. However, results seem to indicate that co-location and cluster co-location may be redundant therefore the co-location and cluster co-location metrics are being double counted in the final GR score. Reichenbacher et al. (2016) developed a GR model that aggregated the cluster and co-location metrics into a single score referred to as the Geographic Environment. The final GR score is generated by combining the Topicality, Proximity and Geographic Environment scores using an equally weighted average (Reichenbacher et al., 2016). The UrbanData model did not adopt this approach based on the assumption that co-location and cluster co-location are unique but results indicate that future models may need to aggregate these metrics to avoid overweighting these metrics and skewing results.

The UrbanData analysis also showed that the effectiveness of GR metrics varies across different study contexts and there may be opportunities to eliminate unnecessary GR metrics based upon the study context. A possible approach to improving the UrbanData model would be the use of step-by-step

analysis model where the UrbanData analysis is conducted with all four GR metrics and criterion are progressively removed from the GR aggregation model to maximize variance in GR scores and optimize explanatory power. The sequence of the analysis can also be reversed so that metrics are iteratively added to the GR score aggregation model. This model would require the UrbanData model to evaluate the distribution of scores for each GR metric in each analysis and to determine whether the distribution of scores meets threshold criteria such as a defined threshold range or standard deviation. This approach to GR score aggregation may contribute to GR scores that have a higher mean score and larger variance of scores which would improve relevance rankings. However, the UrbanData results showed that the identified GR metrics rarely provide false positives and all GR metrics appeared to be somewhat relevant in most of the test cases. A model that dynamically eliminates different GR metrics in the analysis may be prone to erroneously removing important GR criteria. Nonetheless, it is a model that is worth exploring. Another model that may be an effective improvement of the UrbanData model would be the dynamic calibration of individual GR metrics based upon characteristics of the domain data. As noted in earlier sections, GR metrics such as proximity, co-location and cluster co-location rely on predefined threshold values which are often selected by the application administrator based upon their understanding of the framework database and study extent. Dynamic calibration of threshold values could evaluate the extent of the domain data set and the distribution of features within the domain database to determine threshold distance for proximity, cluster and co-location. This approach would reduce the need for manual calibration of the UrbanData model and it is assumed it would create GR scores that are more representative of spatial relevance.

4 Using Geographic Relevance to Contextualize Unstructured VGI

4.1 Introduction

Public engagement, a term that is often used interchangeably with public participation (Ross et al., 2016), is an important dimension of urban planning that can be resource intensive and expensive. Nonetheless, it is an essential process for ensuring equitable and fair decision making (Abelson et al., 2003; Planning Institute Australia, 2011). Recent advances in internet technologies and changing trends in technology adoption are creating new opportunities to use technology to improve public engagement, governance, and urban planning (Kalvelage et al., 2018; Seeger, 2008).

Public participation within the planning process has three major goals which include informing the public about proposed development(s), gathering local information and knowledge from citizens regarding sites of interest and incorporating the opinions and desires of citizens into the decision-making process for the planned developments (Planning Institute Australia, 2011). Ensuring that public participation is representative of the diverse views of the community is a significant challenge(Abelson et al., 2003; Cinderby, 2010). Public engagement hearings and surveys are effective ways to communicate with residents and communities, but marginalized and less affluent parts of society tend to be underrepresented at these sessions (Cinderby, 2010). Practical challenges such as language barriers, time constraints, and lack of transportation can impede citizens from participating in a public engagement session. There is a need to create a model of engagement that is inclusive, accessible, and fosters collaboration (Innes & Booher, 2004). Research shows that geoweb tools can be effectively used to scale public participation efforts to reach a large set of participants that can be more representative of the views of the community (Jankowski et al., 2019).

The public is increasingly adopting digital forms of communication and they are willing and capable of using digital solutions to communicate and interact with the government (Zook, Graham, & Boulton, 2015). Digital solutions currently augment existing public participation forums such as town halls and seminars and they will continue to play a larger role in the engagement process as governments engage younger tech-savvy citizens. Planners need to be aware of this trend and they need to ensure that online engagement translates to offline relationships and actions (Kleinhans et al., 2015). Social media is widely recognized tool for public engagement that is used by governments to gather feedback on initiatives (Evans-Cowley & Griffin, 2012; Schweitzer, 2014), but social media is not geared

towards supporting engagement on complex location-based projects and it is biased towards younger segments of the population. As a result, several companies have developed geoweb solutions to address the challenge of public participation using location-based web applications. Platforms such as IdeaScale, CommunityRemarks, BangTheTable, and Neighbourly allow cities to post information about planning initiatives and solicit citizen feedback online with support for publishing documents and facilitating surveys. These platforms are great at facilitating discussion but they also generate complex unstructured location data that is difficult to analyze without manual review. In contrast, map-based reporting applications such as Community Remarks, ArcGIS Survey123, and SeeClickFix allow citizens to identify specific locations of concern by interacting with web maps or reporting information from their phone. These group of applications allows organizations to control the structure and content of the data being collected online and they are primarily focused on supporting reporting systems such as 311.

The location-based comment data that can be collected using the applications described above is a form of data that is described as facilitated volunteer geographic data (FVGI) (Seeger, 2008). This form of data is often rich in spatial and semantic information, but it is also unstructured and contains folksonomies and local knowledge that are difficult to process (Kalvelage et al., 2018). As a result, the use of web-based data for decision making is still a challenge in urban planning. Planners are collecting more georeferenced input from citizens using social media and VGI applications but interpreting the data into information is still a significant challenge. There are few consistent processes for integrating digital feedback into decision making (Brown, 2012). Furthermore, the cost of processing unstructured data collected from citizens can be prohibitive to the use of web tools in urban planning (McKenzie et al., 2014) . GIS researchers have recognized the need for automated processing of unstructured spatial data generated by volunteers, particularly through contextual analysis (Goodchild & Li, 2012). Researchers in geographic relevance have developed probabilistic methodologies to understand the spatial context in mobile information retrieval where information searches made on the phone retrieves data based on the user's location, velocity, and direction of travel and other criteria (M. Li et al., 2015; Reichenbacher et al., 2016). Some studies have used basic contextual analysis to mine data on Twitter to detect events such as wildfire (Spinsanti & Ostermann, 2013), these methods often rely on a combination of keyword search techniques coupled with proximity analysis to detect locations of interest. This thesis argues that there is an opportunity to use the metrics developed in geographic relevance to automate contextualization and analysis of VGI data. This thesis develops a prototype application named UrbanContext that uses metrics developed in geographic relevance research to contextualize unstructured VGI data by matching data to relevant framework data. The following

sections explore the literature used to motivate and inform this thesis before discussing the design, implementation, and application of the UrbanContext application.

4.2 Literature

4.2.1 Planning, VGI and Public Participation GIS

It is widely accepted that the best approach for planning a community is through consultation and stakeholder engagement (Hodge & Gordon, 2008). Community consultation is essential in modern planning because the issues addressed by planners are generally *"wicked"* problems. These problems are defined as problems with high complexity and multiple competing points of view that make it difficult to identify a single correct solution (Charalabidis, Gionis, Ferro, & Loukis, 2010). Therefore collaboration is needed to develop solutions that are representative of the community and equitable.

Traditionally, planning has been a top-down decision-making process were officials in power directed large scale development projects with little regard for local communities, this has historically been a problem in inner cities and neighbourhoods that were occupied by poor or marginalized communities (Hodge & Gordon, 2008). Over time planners started to acknowledge that a top-down approach to urban planning tended to hurt and disenfranchise small communities (Jacobs, 1992). The backlash against conventional planning theory resulted in major changes in the profession; over time the role of the planner has evolved from a decision maker to a facilitator who encourages communication and collaboration, and the law has adjusted to reflect this trend.

In Canada, public participation is now mandated by law in the planning process (Government of Ontario, 2019). In Ontario, provincial policy dictates the need to consult the community during the planning process. At every stage of the planning process, the city and developers must try to engage the local community and provide them with an opportunity to share their concerns. As part of the planning process municipalities and developers must notify residents within a given distance of a proposed plan of the proposed changes via mail and newspaper notifications. A public hearing must be scheduled as part of this process and any member of the public has an opportunity to voice concerns or provide a written comment at the public hearing to influence any decisions on the plan. Participation in the public hearing gives the participant a legal right to appeal any decision made on the proposed planning issue within the provided appeal period. Authorities must consider, and address comments and concerns raised through the appeal or public engagement process. (Government of Ontario, 2019)

Good public participation is defined as citizen engagement where citizens have the power to influence decisions, can access and interact with the process (Brown & Kyttä, 2014; Webler, Seth Tuler, 1999), despite legal mandates, researchers have widely recognized that public participation can be improved significantly (Kleinhans et al., 2015; Shipley & Utz, 2012). Planners and researchers recognize the challenges and flaws associated with the current public engagement process and each successive generation of urban planners have attempted to improve the process (Brown & Kyttä, 2014; Brown & Raymond, 2014; Kleinhans et al., 2015). Web and mobile technologies are a major point of focus among modern planning and GIS researchers; many have identified the web as a significant opportunity to improve public engagement and urban planning (Brown, 2012; Seltzer & Mahmoudi, 2012). Despite the consensus among researchers that the web has immense potential to improve urban planning, it has also been noted that the rate of adoption and the effectiveness of web technologies have been underwhelming (Brown, 2012; Kleinhans et al., 2015). There are numerous reasons behind the lack of technology adaption in this sector; two major barriers to adoption, as identified in the literature, are access to quality data and the ability to derive information from the data (Brown, 2012; Kleinhans et al., 2015). Planners also need a better understanding of emerging technologies, particularly technologies such as social media and volunteered geographic information which has the potential to empower small communities.

Public Participation GIS (PPGIS), is defined as a field of Geographic Information Science (GIS) that concerns itself with engaging the public using GIS platforms (Tulloch, 2008). Participation GIS (PGIS), a dimension of PPGIS that focuses on the collection of local knowledge using GIS platforms(Verplanke et al., 2016). PPGIS is defined by two dominant terms: a) public which refers to the community at large or a subset of the community at large, and b) participation which broadly refers to interaction with the public (Cinderby, 2010). PPGIS is a field of research that evaluates the use of GIS and web solutions to improve public participation processes in government. Though research in PPGIS is extensive; researchers have criticized the concept as impractical due to organizational, societal, and technological constraints. As noted by Brown, the field of PPGIS has failed to be an effective tool for democratization due to institutional (Brown, 2012). PPGIS and planners, in general, have struggled to ensure that participants in the planning and decision-making process are representative of the public (Cinderby, 2010). In recent years, the concept of PPGIS has been complemented by volunteered geographic information (VGI) a field of research that looks at the growing community of amateur geographers who generate large databases of spatial data using web 2.0 tools and applications (Verplanke et al., 2016).

Volunteered geographic information (VGI) is a web 2.0 trend that is characterized by a growing body of amateur geographers who actively contribute their spatial knowledge and observations to create spatial data (Basiouka & Potsiou, 2012; Goodchild, 2007b). VGI data is a unique source of GIS data because some it can be geosocial, heterogeneous, time-sensitive and responsive (Rob Feick & Roche, 2013); the responsive nature of VGI data has made it an invaluable tool for disaster response (Spinsanti & Ostermann, 2013) and the application of VGI data continue to grow. Many researchers have compared VGI, PPGIS, and PGI data, the three forms of data share similar characteristics. According to the literature, VGI is primarily concerned with collection and maintenance of data for broad consumption while PPGIS and PGIS are topics of research concerned with empowering communities by allowing them to collect local knowledge and have more influence on political decisions(Verplanke et al., 2016).

VGI is a significant opportunity for planners and cities to tap into local knowledge and collective intelligence to improve decision making. There is an increased focus on leveraging data in government, and there is a significant desire to make policy more data-centric. VGI can play a significant role in enabling such initiatives, but the planning profession has been slow to adopt such solutions. As a result, researchers have generally spearheaded the use of VGI within the planning sector through projects such as GeoActon, MapChat, and AdaptNS (Beaudreau, Johnson, & Sieber, 2012; Hall et al., 2010; Minano, Johnson, & Wandel, 2018). The research on web 2.0 and planning based VGI tool demonstrate clear value as planners have the opportunity to collect data they would not have access to otherwise, whether it's feedback or local knowledge (Hall et al., 2010). Studies on the use of the GeoWeb and VGI for use in planning make it apparent that gaps in expertise are major impediments to adoption. In general, planners don't have the technical skill set or time needed to set up and deploy custom community engagement applications; they also don't have the time to maintain and evaluate the data collected from such an application unless it is part of the project requirements (Beaudreau et al., 2012; Hall et al., 2010).

Despite the growing presence of technology and internet tools in our personal lives, barriers to adoption of web-based tools are still abundant. Many government staff are unaware of the latest web tools and even fewer are aware of geoweb tools (Johnson & Sieber, 2012). Furthermore, governments are extremely sensitive to information management and dissemination; governments want to control the content that is published through their platforms and VGI inherently forces governments to relinquish that control (Johnson & Sieber, 2012). Governments are trusted sources of information and

citizens rely on these institutions to provide them with reliable information. In contrast, VGI in the planning context can be highly unreliable; VGI researchers have identified this as spatial uncertainty (Grira et al., 2010).

VGI data collected by cities and urban planners differ significantly from the data collected through popular VGI platforms such as OpenStreetMap or iNaturalist. The concept of volunteered geographic information has many attributes that overlap and coincide with concepts such as public participatory GIS (PPGIS) but there are key differences in how data is collected and used that differentiates participatory GIS from VGI (See et al., 2016; Tulloch, 2008). Participatory approaches with a GIS term have generally fallen under several umbrella terms which include public participatory GIS (PPGIS), participatory GIS (PGIS), geo-questionnaire and neogeographic mapping (See et al., 2016; Verplanke et al., 2016). PPGIS, VGI all share the idea of Neogeography and the use of local knowledge from non-experts to generate spatial data (Goodchild, 2007b; Tulloch, 2008; Verplanke et al., 2016). Facilitated VGI (f-VGI) is a term that has been used to describe a class of VGI data that is collected in controlled settings by study facilitators (Kalvelage et al., 2018). Seeger initially defined the concept of f-VGI within the context of landscape planning as unstructured VGI that is provided in response to a set of predefined criteria (Seeger, 2008, p. 200). According to Kalvelage et al. (2018), f-VGI tends to be used in planning scenarios to collect unstructured citizen feedback in a controlled setting or system. However, the concept of f-VGI is not widely used in the research community and PPGIS is more commonly associated with the process of collecting VGI data from citizens to support urban planning or government decision making (See et al., 2016; Verplanke et al., 2016). The term f-VGI may not be necessary to describe VGI data gathered through PPGIS processes but it is useful to recognize differences in data characteristics between VGI collected on platforms such as OSM in contrast to data collected in PPGIS processes. Data created in community-based VGI platforms such as OpenStreetMap can grow rapidly to include thousands of entries and tend to stay active over numerous years with a living community of contributors (Goodchild & Li, 2012). Data created in PPGIS projects tend to be unstructured and VGI applications tend to be used as disposable tools that only exist for the lifespan of a government project (Beaudreau et al., 2012). VGI data collected in PPGIS processes is commonly characterized by spatial data in the form of a geo-referenced point, line or polygon that is coupled with a large unstructured body of text (Hall et al., 2010; Kalvelage et al., 2018; Seeger, 2008). Unlike geotagged social media or large scale VGI platforms, VGI data collected in PPGIS applications is constrained to an area of study and topic of interest (Kalvelage et al., 2018; Seeger, 2008). The study constraints inherent to PPGIS ensure that data contains less noise than other comparable data sources

such as geotagged social media which is characterized as geographically uneven and thematically dispersed (McKenzie et al., 2014; Spinsanti & Ostermann, 2013). In Chapter 4, references to VGI are directed at VGI data collected from PPGIS applications where data is defined by a geographic point that is associated with a body of texts that represents feedback and local knowledge provided by citizens. The processes discussed in this Chapter are also relevant to ambient VGI data collected from social media and other comparable sources of local knowledge. However, due to limited resources this Chapter looks at a constrained and focused use case for analyzing unstructured VGI using geographic relevance.

4.2.2 Geographic Relevance and Contextualization

Spatial context is an idea that has been formalized in Tobler's First Law of Geography (Tobler, 1970). Goodchild and Li expand on Tobler's first law by arguing that Tobler's first law insinuates that a proposed fact about a location should be consistent with what is already known about the vicinity of a location (Goodchild & Li, 2012). Contextual analysis of VGI data is the process of linking VGI data to authoritative data to enrich VGI data and support filtering and verification (Spinsanti & Ostermann, 2013). The challenge that researchers continue to face is understanding spatial context when working with data that are abstractions of the real world (Hahmann et al., 2014). Linking points and polygons that are intended to represent trees and parks is a non-trivial task that requires complex data models. This issue is further complicated when working with VGI data where data is conceptualized by untrained individuals rather than professional organizations. Variances in data make it difficult to develop a consistent data processing methodology due to a large number of edge cases. Geographic relevance is an emerging concept in the literature that attempts to formalize these ideas and theories around measuring spatial relevance and automating spatial contextualization.

Reichenbacher et al. (2016) developed a broad definition of geographic relevance that introduces the idea of implicit and explicit information need. It was argued that geographic relevance is a measure of how well spatial information matches a users' information needs (Reichenbacher et al., 2016). The idea of user information need is central to the concept of geographic relevance and spatial context. In the real world, a person or an object is surrounded by millions of objects, whether it's grass, trees, ponds, street lights, sidewalks or buildings. To a tourist the street lights, sidewalks, and buildings are likely objects of interest. To an ecologist the grass, trees and ponds are likely objects of interest. Spatial context is defined by the user and the spatial objects in an area of interest. Geographic relevance uses the concept of user information need or geographic information needs (GINs) (Raper, 2007) in order to

filter and prioritize probable objects of interest from a large list of possible objects of interest. Explicit user information needs refer to queries for spatial information made from a website or a mobile device. Herein, information need can be identified by parsing a query and identifying the location of the user (Reichenbacher et al., 2016). Implicit user information need refers to the identification of information need based on who the user is or the context in which the data was collected (Raper, 2007). GR models developed by Reichenbacher et al. (2016) account for these information needs through a set of metrics that generate a topical score, a geographic environment score, and a mobility score (Reichenbacher et al., 2016). Each of these relevance scores are generated by evaluating a variety of semantic and spatial criteria that represent relevance.

4.2.3 Semantic Relevance

Semantic relevance or topicality is a measure of similarity between a body of text (document) and a query (Reichenbacher et al., 2016), . Semantics refers to the study of meaning; it delves into relations between signifiers like words, phrases, signs, and symbols and what they stand for (W. Li, Goodchild, & Raskin, 2014). Semantic relatedness describes the strength of association between two concepts, which encompasses hypernymy, hyponymy, meronymy, antonymy, synonymy, and other nonclassical relationships (Zhang, Gentile, & Ciravegna, 2013). The semantic relatedness of two concepts is commonly measured using semantic distance which is an inverse measure of the distance between two lexical concepts, but the process of identifying and extracting lexical concept is a non-trivial challenge that has been addressed in multiple research papers (Zhang et al., 2013). Spatial semantics encompasses all the information related to a spatial object that describes what a given geographic element (Ramos et al., 2014). Semantic relevance looks at how lexically defined location references in documents or queries can be matched to the lexical representation of spatial entities stored in a GIS system (Bordogna et al., 2012; Purves et al., 2018). The following section reviews the methodologies and tools used to break down user queries and identify semantically relevant entities with respect to user queries.

The semantic representation of GIS data is defined by elements, attributes, and topological relationships. For example, a polygon element with a name attribute of “Toronto” is used to represent the City of Toronto. The City of Toronto polygon element is located within the province of Ontario, thus the city of Toronto is semantically related to the province of Ontario. In semantic analysis, It is important to recognize that location references for the same locations can vary between organizations and people due to differences in how location data is conceptualized. This is a phenomenon known as semantic

heterogeneity and it significantly increases the complexity of evaluating semantic similarity between features (Laurini, 2014; Ramos et al., 2014). For example, the City of Toronto has many local nicknames and aliases such as “TO” or “the six” that are place names derived from popular media rather than authoritative data. Measuring semantic relevance requires the development of systems that can match location references to spatial entities while accounting for semantic discrepancies (Ballatore, Bertolotto, & Wilson, 2013). Gazetteers and spatial ontologies are two of the most well-documented approaches to matching location references in the text to spatial entities in GIS systems.

Gazetteers precede computer technology (Graham & De Sabbata, 2015). For the purpose of this paper, gazetteers will refer to digital gazetteers which are stored and maintained on computers (S. Gao et al., 2017). Digital gazetteers are indexed databases that contain structured information about named places, where the place is defined as a geographic location that has been referenced and identified using a socially accepted name (Graham & De Sabbata, 2015). It is important to distinguish places from features in the context of gazetteers because it has implications on how data is stored and referenced in databases. All places can be features but not all features can be places; features refer to distinct physical elements with distinct boundaries such as parks, mountains, buildings and administrative boundaries (Goodchild & Hill, 2008; Graham & De Sabbata, 2015). From a GIS perspective, a feature is a spatial record in a database that can be referenced directly using an ID. Places can be distinct features like administrative boundaries or points of interest but they also encompass vague or generic location references such as “Southern Ontario”, “downtown” or “along Queen Street"(Derungs, Wartmann, Purves, & Mark, 2013). From a GIS perspective, places are not distinct database records and they can not be referenced using an ID unless they are linked to a feature. Gazetteers are indexed databases that hold information about a subset of places that have been assigned a proper name by society, organization or government. Records must contain a name, a spatial footprint or reference point and a category (S. Gao et al., 2017; Goodchild & Hill, 2008). Gazetteers lookups are relatively simple database operations that are effective at matching place names to locations (McKenzie et al., 2014). They are not effective at handling edge cases which may include ambiguous location references or the use of loca(S. Gao et al., 2017). Spatial ontologies are often used in GIR and GR systems to match location references that can not be matched by gazetteers (Purves et al., 2007; Reichenbacher et al., 2016).

Spatial ontologies make it possible to deal with ambiguous location references found in the document. Ambiguity is difficult to handle because many edge cases need to be handled. The most common types of ambiguity found in spatial data can be categorized into three dominant categories

which are termed multiple references, variant name, and geoname (Neuhaus, 2018). Multiple references refer to instances where multiple locations share the same name, the variant name refers to the use of multiple names for a given location, and geoname refers to place names with multiple meanings (Neuhaus, 2018).

Ontologies are defined as a representation of domain-specific knowledge that allows machines to work with the semantic content of an entity and handle ambiguity (Neuhaus, 2018) (Machado, de Alencar, Campos, & Davis, 2011). Ontologies are graph databases that are composed of classes and relations where classes define objects or concepts, and relations identify linkages between classes (N. Li, Raskin, Goodchild, & Janowicz, 2012). Classes are used to define and represent entities such as buildings, people or locations while relations are used to define relationships between classes to indicate whether objects represent the same object, similar object or the inverse object (N. Li et al., 2012). Spatial ontologies are customized ontologies that are developed to handle unique GI challenges such as place name disambiguation and tracking spatial relationships such as spatial containment or adjacency (Fu, Jones, & Abdelmoty, 2005b).

GIR and GR researchers use gazetteers and geographic ontologies to preprocess queries and documents to identify candidate place names; this process is termed as spatial query expansion in the literature (Bordogna et al., 2012; Fu et al., 2005b). Query expansion is an approach to running queries on large bodies of text where parts of the text can be used to form a query while other parts of the corpus simply produce noise (Krishnan, Deepak, Ranu, & Mehta, 2018). Spatial query expansion is the process of using a gazetteer or geographic ontology to extract location references within the text and identify a set of candidate locations that are relevant to the query (Fu et al., 2005b). Each candidate location is defined as a place footprint (p-footprint), which delineates a location reference, the set of candidate locations in a query defines a query footprint (q-footprint) which defines the spatial search extent of the query (Fu et al., 2005b). Spatial query footprints is a concept that has been used extensively in GIR (Acheson et al., 2018; Bordogna et al., 2012), but it has not been discussed to a great extent in GR literature despite its relevance to the GR problem.

4.2.4 Spatial Relevance

Spatial relevance is a measure of the strength of the relationship between two entities based on spatial and topological relationships between the two objects. Spatial relevance is a central idea of GI literature (Tobler, 1970) and it is a topic that has been widely explored in conflation (Haklay, 2010), spatial contextualization (Spinsanti & Ostermann, 2013), geographic information retrieval (Bordogna et

al., 2012; Purves et al., 2007) and geographic relevance (Reichenbacher et al., 2016). Reichenbacher et al. (2016) define spatial relevance as an aggregate measure of four metrics, which are spatiotemporal proximity, directionality, cluster, and co-location. The use of directionality, cluster and co-location measures for spatial relevance analysis are novel to the field of geographic relevance and the metrics were first proposed by Reichenbacher et al. (2016).

Spatio-temporal proximity is defined as a measure of Euclidian (spatial) distance or the network (spatiotemporal) distance between two spatial entities. Spatial proximity is a common measure of spatial relevance that has been used for the automated conflation of VGI data (Spinsanti & Ostermann, 2013), VGI data matching (McKenzie et al., 2014) and GIR research (Bordogna et al., 2012). Directionality is a concept that is specific to mobile computing, and it is closely related to spatiotemporal proximity. The directionality metric infers that a mobile user is more interested in locations that are in the user's current direction of movement (Reichenbacher et al., 2016). Directionality is a metric of geographic relevance that is specific to navigation using mobile devices and Reichenbacher et al. (2016) describe the directionality metric as a desirable measure of relevance rather than an essential measure of relevance. The cluster metric assumes that spatial objects found in a large cluster of related entities are more relevant to the user than objects found in a small cluster (Reichenbacher et al., 2016). This metric assumes that the size of the spatial cluster indicates spatial importance, and an important entity is more likely to be relevant to user information needs. Spinsanti and Ostermann describe this concept as the social confirmation heuristic, which is a pertinent idea when validating VGI data collected using social media (Spinsanti & Ostermann, 2013). Clusters can be detected using several well-established algorithms such as k-means clustering (Kanungo et al., 2002) or GDBSCAN (Sander et al., 1998). The co-location metric assumes that given an entity belonging to a given category, it is probable to find an entity belonging to the second category within a defined distance of the first category (Reichenbacher et al., 2016). This rule is comparable to the cluster rule, where entities that are co-located with known objects of interest are considered to be more relevant to the user's information needs. Several algorithms exist to measure co-location of spatial entities which include cluster based rules and classification based rules (Huang et al., 2004).

4.2.5 Score Aggregation

There is a broad set of literature that evaluates the process of semantic and spatial score aggregation, but there is little consensus among the literature (Bordogna et al., 2012; Koukoletsos et al., 2012; Purves et al., 2007; Reichenbacher et al., 2016; Spinsanti & Ostermann, 2013). GIR platforms such as SPIRIT and Geo-Finder generate isolated semantic and spatial scores which are aggregated using a weighting scheme (Bordogna et al., 2012; Purves et al., 2018), Spinsanti and Osterman also developed semantic and spatial score aggregation module in order to automate contextualization of VGI data (Spinsanti & Ostermann, 2013). The geo-retrieval module in GeoFinder parses user queries into two subsections; one is the content base which consists of content keywords and the other is the spatial content which includes a list of spatial keywords. The Geo-Retrieval module returns two scores which are the content retrieval score and the spatial retrieval score. These two scores are combined using asymmetric or symmetric aggregation schemas. The asymmetric aggregation model runs the content condition evaluation and the spatial condition evaluation in sequence. Documents are evaluated to verify if the document matches the content conditions of the query, only documents that meet the content conditions of the query are selected for spatial evaluation. This can create a result set where thematically irrelevant and spatially relevant documents are eliminated. In contrast, symmetric aggregation evaluates content conditions and spatial conditions in parallel and then creates an average score of the two to retrieve relevant documents. This means that spatially relevant documents may be retrieved even if the document content is not relevant to the query. (Bordogna & Psaila, 2008)

The SPIRIT search engine handles queries by generating a document spatial similarity score which is a composite score of textual and spatial document scores. This is handled by comparing four key metrics. A footprint similarity score is produced to compare each document footprint to the query footprint. A document spatial similarity score is produced based on the footprint similarity score of all footprints in a document. Document spatial similarity scores are combined with the textual similarity score to provide the general geographic relevance score. The footprint similarity score is calculated using multiple metrics which include inside analysis, near analysis and direction analysis. (Purves et al., 2018)

- **Inside:** (Binary) Evaluates if the query footprint is inside the document footprint.
- **Near:** (Distance Decay Function) Calculates a score based upon the distance between the centroids of the two footprints.

- **North-of, South-of, East-of, West-of:** An evaluation of the direction or angle between the centroids of query footprints.

This system uses the best match approach to determine a document's spatial relevance, the footprint within the document with the highest similarity score is used as the document's similarity score. Both the textual relevance score and the spatial relevance score are first normalized to a 0 to 1 score before being combined. (Purves et al., 2007, 2018)

4.3 Methodology

Herein we describe UrbanContext, a prototype application that automatically contextualizes VGI data. UrbanContext is a data processing application that detects patterns within the VGI data set to compare and match VGI features to features in the framework database. The first section of this methodology will describe the algorithm used to process VGI data. The second section will describe the software and database architecture developed to support the UrbanContext algorithm. The third section discusses a case study of the application using a sample data set from the City of Kitchener. The UrbanContext application is a data processing application that accepts VGI data and framework data as inputs and produces a set of foreign ID tables as output. The GR models that are developed in this thesis are expected to be generally relevant for GIS analysis of unstructured data. However, the use case for the UrbanContext application is primarily envisioned for GIS analysts who support Urban Planning applications. The UrbanContext application can also be connected to web applications that allow non-technical users such as policy planners and councillors to interact with the UrbanContext application and visualize results. A sample web application named UrbanContext Viz is demonstrated alongside the UrbanContext application in order to visualize results and review outputs. Currently, technology savvy planners and GIS analysts collect, transcribe, analyze, map and summarize citizen feedback in a report using an ad-hoc combination of software and manual processes. This process could include the use of Excel for classification, the use of ArcGIS for mapping and the use of Google Earth or web maps for background research. Policy planners and councillors tend to interact with summary reports about community feedback rather than the community data directly. The UrbanContext application is intended to assist planners and GIS analyst with exploring and summarizing content but it can also be used to create applications and visualizations that could potentially be used by policy makers and councillors directly to better understand complex planning issues. It is expected that more people will use and review data if it is easier to explore and summarize information.

The applications demonstrated in this chapter need to be manually implemented by a GIS analyst, but the design of these applications provide a blueprint for the development of a publicly accessible application. UrbanContext can be deployed by a GIS analyst by loading framework data sets and unstructured VGI data into the UrbanContext application. The GIS analyst also has to provide the UrbanContext application with three threshold distance values in order to calibrate the system. The threshold distance values are described in greater detail in sections 4.3.4, 4.3.5 and 4.3.7. The general assumption is that the GIS analyst is familiar with the study extent and input data and identifies parameter values based on their understanding of the study context. Once the input data is loaded into UrbanContext and threshold values are defined, UrbanContext creates a gazetteer using framework data. The application uses the gazetteer and the UrbanContext geographic relevance model to match VGI data points to framework features. The outputs of this analysis are provided as a foreign ID table that contains the ID of VGI features and IDs of framework features and their associated relevance scores. A GIS analyst can use the foreign ID table and the associated relevance scores to generate summary statistics, maps and figures about the VGI data set and their associated framework features. The UrbanContext Viz application is a web application built on ESRI's ArcGIS JS 4 API and it makes is possible for non-technical users to interact with the UrbanContext analysis results. It is preconfigured to support specific visualizations that are pertinent to a case study where VGI data consists of geotagged comments that can have multiple location references. The UrbanContext application matches each VGI data point to multiple features in the framework database using the GR analysis model. The UrbanContext Viz application allows users to interactively click on each VGI data point and visualize all the framework features and locations that are discussed in the selected VGI comment. This application allows analysts and data reviewers to quickly visualize and review the locations and objects that are important to each citizen that contributed feedback. It also makes it possible to review individual relevance and match scores generated by the UrbanContext analysis.

The methodology for UrbanContext builds on metrics defined in GR literature, particularly measures of topicality, spatial proximity, cluster, and co-location (Reichenbacher et al., 2016). This methodology expands on existent approaches and proposes the novel use of query footprints (q-footprints) as defined by Purves et al. (Acheson et al., 2018; Purves et al., 2007) to evaluate geographic relevance between unstructured and ambiguous VGI data and framework data. The use of query footprints has been proven to be an effective construct for disambiguating vague spatial references in web documents (Fu et al., 2005b; Purves et al., 2007, 2018). This methodology describes an approach for using q-footprints with GR metrics to match VGI data to framework data using many to one

relationships. The following sections describe the methodology using sample data from the case study discussed in section 4.4. These samples serve as examples to explain general concepts about the methodology. The specific functions and code used to implement the methodologies discussed in the following sections can be found in Appendix A.

4.3.1 UrbanContext

The UrbanContext application is a back-end data processing application that links VGI data to framework data using GR metrics. The UrbanContext application is solely focused on data processing but it is possible to integrate the UrbanContext modules into web applications to further automate data processing and to allow data visualization. This implementation of UrbanContext does not have an integrated data visualization module so output data must be visualized and verified using mapping visualization applications such as QGIS, ArcGIS Pro or ArcGIS JS. This implementation also relies heavily on VGI data and framework data from the City of Kitchener but the UrbanContext application can be used in other study contexts.

The UrbanContext application is built using Postgres with the PostGIS extension and Python with the NLTK extension and the ogr2ogr extension. Postgres is a feature-rich open-source relational database system that adheres to SQL standards, PostGIS is a spatial extension for Postgres that makes it possible to manage, manipulate, and analyze spatial data using SQL. The UrbanContext platform uses Postgres to store framework data and VGI data and to store output datasets. The application uses PostGIS for all spatial analysis function such as distance measurements or cluster creation.

The NLTK library, also known as the natural language toolkit, is a Python library that makes it possible for programs to work with human language data. It bundles lexical resources such as WordNet and supports language processing functions such as parsing, tokenization, tagging, and more. The UrbanContext application uses the NLTK WordNet interface to generate the topicality score by measuring semantic similarity scores between spatial concepts. The ogr2ogr library is also used in the UrbanContext platform to automate the management of spatial data such as downloading spatial data into Postgres from Open Data Portals.

The specific implementation of the UrbanContext application includes 16 PL/SQL functions to handle custom spatial calculation, two python functions to handle NLTK functions and data management, one materialized view for the gazetteer and a set of materialized views to handle data processing over multiple stages of the GR analysis process. Currently, the UrbanContext application is

not a fully automated system because full automation would require the development of numerous additional modules focused solely on data management. Development of these modules is ancillary to the purpose of this thesis. The exact implementation and functions used in these modules are discussed in the following sections of this methodology. Discussing the assumptions made for various metrics will clarify the implementations of different UrbanContext functions. Code samples can be found in Appendix A and table outputs can found in Appendix B for further reference.

The UrbanContext platform relies on two main categories of data which include VGI or VGI data and reference data. The VGI data is the data set being analyzed by the UrbanContext platform it is considered to be the input data for the UrbanContext platform. The VGI or VGI data must consist of a geometry which can be a point, line or feature and a body of text that makes references to locations in the vicinity of the geometry. This thesis focuses on the analysis of VGI data which has a defined study context and topic of study but comparable sources such as geotagged Twitter or Flickr data can also be used. Reference data is authoritative spatial data which represents key mapping features such as roads, forests, water bodies and more. Reference data is the data that the input data (VGI) is being compared to, the reference data is often referred to as the target data in this methodology. The reference data is selected based on the topic of study and generally contain data that could provide context for the study. For example, a study on municipal trails could include data sets such as trails, roads, bike lanes and intersections while a study on deforestation may contain layers such as forests, habitats, and waterbodies. The subject matter expertise primarily drives the selection of the reference data,

The UrbanContext application takes the VGI data as input data set and analyzes the data by comparing each VGI feature to framework (reference) data. The UrbanContext application looks for location references in the VGI data, and then it looks for features in the framework database that match the criteria of the location references. The output of the UrbanContext analysis is a foreign key table that links VGI data points to individual features in the framework database using many to one relationships. The foreign key table can then be used to organize and filter the VGI data based upon links to the framework data. The output foreign key table can then be used to identify key locations of concerns in the framework database based on comments in the VGI data.

4.3.2 Architecture

The UrbanContext application relies on five main metrics to analyze the input VGI data and links the data to the reference framework data. The five metrics are query footprints (q-footprints), topicality, spatial proximity, cluster, and co-location. The details of these metrics are described in subsequent

sections. The UrbanContext application is composed of four main sections that connect. The first section is the storage tier that contains reference framework data and input VGI data. The second section is the analysis metadata section, this includes generated data that is needed for the analysis such as a gazetteer and VGI cluster detection. The third section is the analysis section which includes modules to create q-footprints from the input VGI and to measure geographic relevance using co-location, topicality, proximity, and cluster. The last tier is the output tier which generates foreign id tables linking each VGI data points to a set of unique features in the framework database. The foreign id tables are referred to as footprints, footprints are foreign key tables that link a single input data point to multiple location references. Each record in the foreign key table represents a place reference and is referred to as a p-footprint. The entire output foreign key table is referred to as a GR footprint because it represents a set of geographic relevant place references for each VGI data point. Figure 4.1 below represents the UrbanContext architecture, as described above, the application starts using a database with VGI data and framework data and moves through the geographic relevance modules to produce a final GR footprint which can then be visualized in QGIS or ArcGIS JS. The analysis process produces several footprint tables that link each VGI data point to a set of location references.

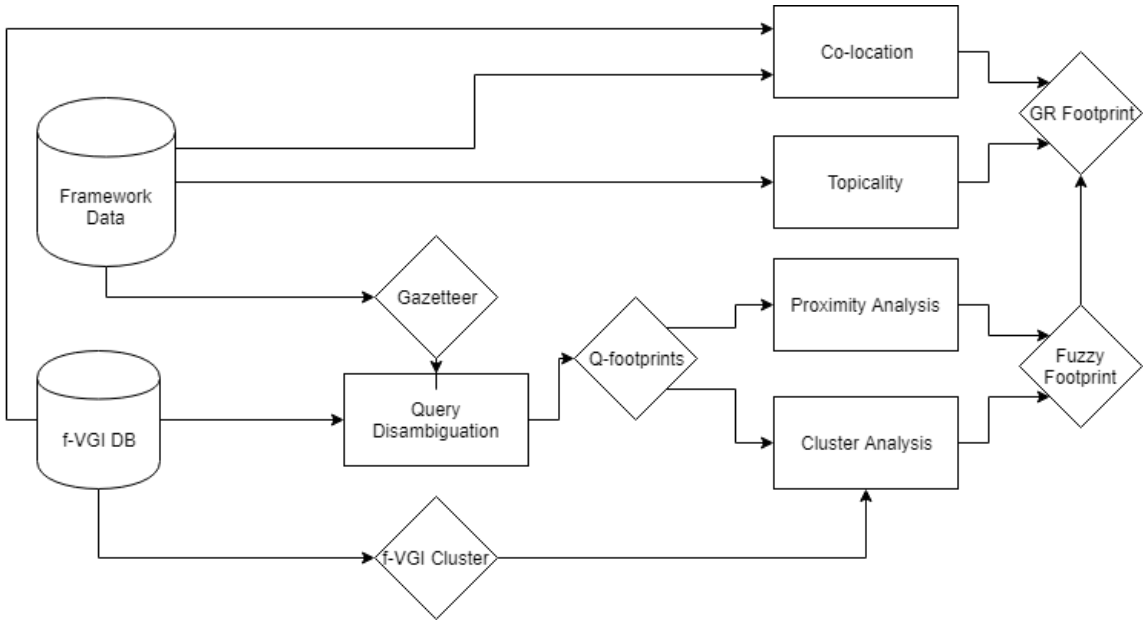

Figure 4.1: UrbanContext architecture depicts how VGI data is processed using GR metrics.

This assessment model uses q-footprints to pre-process unstructured data into manageable chunks of information that can be processed using geographic relevance criteria. The geographic relevance criteria used in this thesis are topicality, spatial proximity, cluster, and co-location. The inputs

to this assessment model include a framework database and an VGI data set and the output of this data set is a list of geographically relevant features which is referred to in this thesis as a GR-footprint. A place footprint (p-footprint) is a term that specifies the geographic location of a location reference (Fu et al., 2005b). A footprint (q-footprint) is a term that describes the geographic extent of a given query as defined by multiple location references, the q-footprint can be composed of multiple p-footprints (Fu et al., 2005b). This thesis uses two additional concepts of fuzzy footprints and GR footprints to describe concepts that are used in the GR assessment model. A fuzzy footprint is described in this methodology as a set of p-footprints that have been matched to spatial framework features without being validated using GR metrics. The GR footprint is a set of p-footprints that have been matched spatial framework features using GR metrics; the GR-footprint is a set of p-footprints that are geographically relevant to the input feature. As shown in the UrbanConext architecture diagram, creation of GR-footprints is a multi-step process. Framework data is used to generate a gazetteer. Each VGI data point in the input data set is put through a disambiguation process where location references within the VGI comment are identified using the gazetteer. Location references identified by the gazetteer are recorded as p-footprints generate a single q-footprint for the VGI data point. The location references identified in the p-footprint are matched to the proximity analysis module. This module finds features in the framework data that match the description of the p-footprint and filters out features based on spatial proximity to the input VGI point. The ranking process generates a set of candidate features for each p-footprint. The cluster analysis module calculates the cluster score of framework features. The proximity score and cluster score of candidate features are combined to create an aggregate score that is used to identify a single candidate feature that best matches a p-footprint based on spatial relevance and cluster relevance. This analysis is repeated for every footprint in the q-footprint to generate a fuzzy footprint for the VGI point. Topicality and Co-location are calculated between layers. The co-location module generates a co-location relevance score for each framework data set based upon co-location patterns between the VGI data and the framework data. The topicality score is generated between the VGI data layer and framework layers using an ontology that evaluates the similarity of semantic descriptions of VGI data and framework data. The co-location and topicality scores are attached to p-footprints in the fuzzy footprint to create aggregate GR scores for each p-footprint. Any p-footprints with negative GR scores are then filtered out, and the remaining p-footprints are ranked to create the GR-footprint for each VGI data point. A GR-footprint is generated for each point in the VGI data set. The resultant data set is a table that links VGI data to framework data based on geographic relevance.

The result can then be used to filter, categorize, and summarize the VGI data set based on relationships with the framework data. The following sections describe the specific implementation of q-footprints, topicality, spatial proximity, cluster, co-location, and GR scores. UrbanContext, the data processing application presented in this thesis, is also discussed in this section. The UrbanContext application is an example of how the metrics in this thesis can be implemented in a software system, but the metrics discussed in this thesis can be used in other software systems using different implementations. For explanation purposes, this thesis has coupled the discussion of GR metrics with the software implementation of UrbanContext to better illustrate results and methodologies. It is important to note that this methodology is only concerned with discussing approaches for measuring GR and generating GR-footprints. The visualization and interpretation of the results are up to the user, but section 4.5 will show examples of how to visualize and interpret the data using tools like Excel, QGIS or ArcGIS JS.

This thesis made use of the ArcGIS JS 4 API and ArcGIS Online to visualize results. Every layer in the framework database was loaded into ArcGIS Online using the shapefile importer. ArcGIS Online was also used to style and visualize layers. Layers were styled based on feature types such as parks or trails. The VGI data set was joined to the GR footprint foreign key table so that every VGI data point had an attribute that contained a set of foreign keys that linked to the reference data. The joined VGI table was then uploaded to ArcGIS Online. ArcGIS JS 4 was used to load the ArcGIS Online layers into a web map that allows users to click on VGI data points and see the associated GR-Footprint. The interactive analysis is supported by a click event listener that listens to click events, pulls the list of foreign keys from the VGI data and filters the reference data to only show geometries that are related to the VGI point. The ArcGIS JS web application is hosted using Glitch, a free JavaScript sandbox that is available online.

Figure 4.2 The UrbanContext Viz web application, found at https://urban-context.glitch.me/

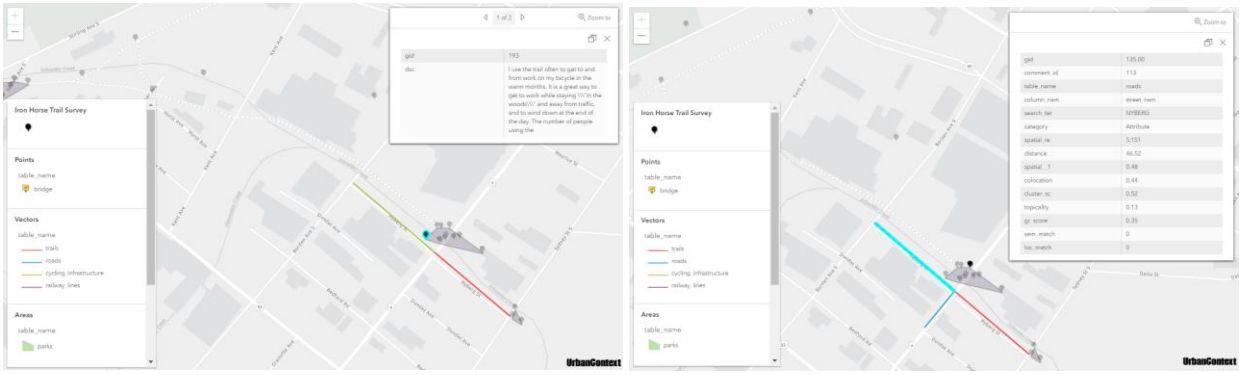

Figure 4.3: Users can click on VGI data points in UrbanContext Viz to display relevant locations.

Figure 4.2 and Figure 4.3 illustrate how users can interact with the web based UrbanContext Viz application to visualize individual VGI points their associated locations. Different features such as streets, parks and bike lanes are highlighted as users interactively select VGI data points that discuss different features and locations. Users can click on individual framework features to see a popup with the GR score assigned for the active VGI point and the selected framework feature. Planners or GIS analysts can use this application to review individual comments while the application highlights valuable contextual information about the locations that are being referenced in a comment. This visualization is particularly helpful if the Planner or GIS analyst is not familiar with the location that is being studied in the VGI dataset.

4.3.3 Query Footprints

Query expansion is a data processing methodology that extracts location references from an unstructured corpus of text (Krishnan et al., 2018) to generate a set of candidate location references that are referred to as place footprints (p-footprints) (Fu et al., 2005b). Location references are extracted from a body of text using a gazetteer or ontology (Fu et al., 2005b). Creating an ontology is beyond the scope of this thesis; therefore, a gazetteer was generated using framework data to support query disambiguation. Traditional gazetteers are simple data structures that are composed of three key elements; a place name, a place type, and a footprint (Machado et al., 2011). This methodology attempts to identify relationships with individual framework features. Therefore, the gazetteer built in this thesis has a data structure that makes it possible to trace p-footprints back to a source feature or layer. The custom gazetteer is generated using materialized views that aggregate place names from the framework data and it includes; a place name, the source table, the source column, the feature ID and the place name category. The gazetteer place names can be layer names, attribute categories or individual feature names from any framework data set in the framework database, the exact SQL query can be found in Appendix A.

Table 4.1: A sample of the gazetteer table that makes it possible to trace source features

id	table	column	place name	feature id	category
1949	parks		Parks		Layer
986	parks	park	PAIGE PARK NATURAL AREA (FLOOD PLAIN)	1868	Feature
1947	parks	park	DUKE PARK	1873	Feature
1705	roads	street_nam	HICKORY HOLLOW		Attribute
1704	roads	street_nam	BLACKHORNE		Attribute

The VGI data points are disambiguated using a join operation with the gazetteer that generates a set of p-footprints for each VGI data point. Once p-footprints are identified, GR metrics are used to match the VGI point to individual layer features with unique Ids. Each p-footprint is classified in one of three categories which delineate the granularity of the place reference. The three classifications are a layer, attribute, and feature, and they match the classification of gazetteer entries. The layer classification is used for p-footprints that reference spatial layers by name or category. This can include references such as “trails”, “parks” or “Starbucks Locations”. Attribute refers to p-footprints that

reference a category value within a layer such as “highways” in the streets layer or “banks” in the points of interest layer. Feature refers to p-footprints that reference distinct spatial objects that have a distinct geometry and can be uniquely identified, feature p-footprints can be directly referenced to spatial database records with a unique ID. The classification of each p-footprint identifies the level of ambiguity associated with each p-footprint, where layers are the most ambiguous footprint, and features are the least ambiguous. As shown in Table 4.2, as the granularity of the matched p-footprint increases, the number of spatial features that can be matched to p-footprints decrease. In some cases, GR metrics are not needed to match location references in VGI data to individual features. The GR metrics used to process p-footprints are also used to validate p-footprints as spatial relevance metrics can detect location reference anomalies based on spatial criteria rather than semantic criteria. The q-footprint acts as a filter to reduce noise, but measures of relevance are generated by the GR metrics described in the following sections.

Table 4.2: Hierarchy of feature matches

Category	Ambiguity	Example	Candidate Features
Layer	3	*Streets*	Any feature within the Streets layer
Attribute	2	*Highways*	Any feature within the Streets layer classified as a highway
Feature	1	*CN Tower*	Single Feature -> CN Tower, Toronto, ON, CA

The figures below illustrate how a sample comment from the VGI data set is disambiguated using the gazetteer described above. The comment in Figure 4.4 references multiple streets and features. Query disambiguation extracts seven location references from the comment text and creates a table of seven p-footprints, as shown in Table 4.3.

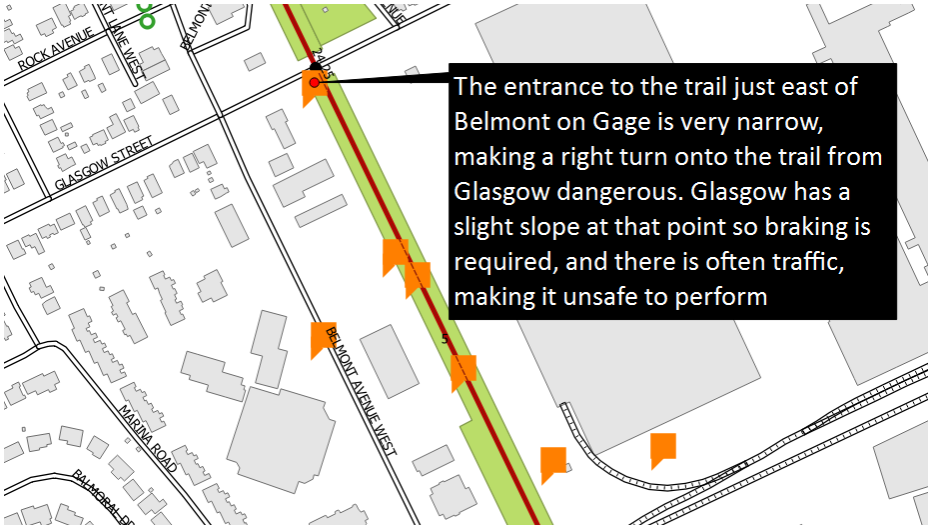

Figure 4.4: Example VGI data point provided by the City of Kitchener

Location references identified within the comment by the gazetteer are highlighted in yellow in the comment below.

> *"The entrance to the trail just east of Belmont on Gage is very narrow, making a right turn onto the trail from Glasgow dangerous. Glasgow has a slight slope at that point so braking is required, and there is often traffic, making it unsafe to perform"*

Table 4.3 illustrates how the highlighted location references below are stored and classified by the UrbanContext system. For every identified location reference, the UrbanContext system stores the source table, the source column and the type of location reference of the gazetteer place name.

Table 4.3: A sample p-footprint that has been generated for the comment in Figure 4.3

Source table	Source column	place name	Source category
roads	street_nam	BELMONT	Attribute
roads	street_nam	EAST	Attribute
roads	street_nam	GAGE	Attribute
roads	street_nam	GLASGOW	Attribute
trails		trails	Layer
trails		trail	Layer

The subsequent GR metrics described below use the data from Table 4.3 to create GR metric scores for proximity, cluster, topicality, and co-location. Metrics such as proximity and cluster are evaluated between the VGI point and candidate features while metrics such as topicality and co-location measure relationships between the VGI data layer and source layers of place names. The process of disambiguating comments and generating GR scores is repeated for every point in the VGI data set.

4.3.4 Spatial Proximity

Spatial proximity is an essential measure of relevance where an entity that is beyond a given threshold distance from the location of interest is considered irrelevant (De Sabbata & Reichenbacher, 2012). The threshold distance that is used to determine if a feature is relevant is a fuzzy concept that varies based on a user's perspective and the study extent (Bordogna et al., 2012). In this methodology, the spatial proximity score is measured using an inverse distance function, and it is used to identify candidate features based on p-footprints. The assumptions used to score spatial proximity are; spatial proximity is highest when the distance between two objects is 0, relevance decreases as distance increases (Reichenbacher et al., 2016). The spatial proximity score is 1 when the distance is 0, and the spatial proximity score is 0 if distance is greater than the threshold. The formula below is a measure of spatial proximity that generates a normalized score between 0 and 1. Given two spatial entities C and D, the spatial proximity score is as follows.

$$SP_{CD} = (dist_{max} - dist_{CD}) / dist_{max}$$

SP_{CD} = Spatial proximity score between feature C from layer A and feature D from layer B

$Dist_{CD}$ = Distance (meters) between feature C from layer A and feature D from layer B

$dist_{MAX}$ = The threshold distance (meters)

This function assumes that spatial relevance follows a linear distance decay pattern, this assumption follows the pattern of contextualization and spatial matching literature (McKenzie et al., 2014), but spatial relevance may follow an exponential distance decay function as well (De Sabbata, 2013). The threshold distance can be set as a parameter for each layer based on the density of features and distance between them. This approach for threshold distance creation is based on the assumption that some spatial features occur less often than other spatial features (Bordogna et al., 2012). This methodology makes use of a single threshold distance that is set based on the study context and the topic of interest in the VGI data set. The data set is analyzed in this thesis is an VGI survey on trail use;

this thesis assumes that a reasonable threshold distance for the given data set is 1000 meters which are commonly described as the ten-minute walking distance for most pedestrians.

The functions needed to measure spatial proximity for individual p-footprints varies based upon the type of p-footprint being analyzed. As noted in section 4.3.2, p-footprints can be categorized as a layer, attribute, or feature. If the p-footprint is a layer the spatial proximity function is run against every feature in the source table, and the closest feature is assigned the highest proximity score. If the p-footprint is an attribute, the spatial proximity function is run against feature in the source table that is of that category and the closest feature with the provided attribute is assigned the highest spatial proximity score. If the p-footprint is a feature, spatial proximity does not need to be evaluated.

```
SELECT  (1000 - st_distance(geom,geom)/1000 as distance
FROM  table_schema.table
WHERE column=attribute
```

The query above is an example of how the spatial proximity score is generated for each p-footprint; this methodology relies on the st_distance function provided by PostGIS to calculate the distance between entities. Features with high proximity scores are identified as candidate entities; candidate entities are stored and passed on to subsequent functions for further GR analysis. Each p-footprint can be matched to multiple candidate entities. In this methodology, three entities with the highest proximity scores are selected for each p-footprint. The three candidate entities are filtered down to a single candidate entity by combining the proximity score with the cluster score described in the following section.

4.3.5 Cluster

Cluster is described by De Sabbata and Reichenbacher (2012) as an area of importance that increases the importance of features contained within; the size or density of the cluster can be considered when assigning importance to clusters. The cluster metric is a spatial data metric that assigns importance to spatial groupings of features. As noted by Spinsanti and Ostermann “The Clusterer emulates the social confirmation heuristic and searches for patterns and confirmation in the VGI” (Spinsanti & Ostermann, 2013, p. 40). In this thesis, the cluster metric evaluates how many features in the target layer exist within clusters of the input layer with the assumption that features within clusters of the input layer are more relevant to features in the input layer. The cluster score is assigned to individual features based on their presence within a cluster and the size (by feature count) of the cluster that contains the feature. The importance of the cluster can also be measured by feature density rather than the pure count of features. The cluster score (Cl) is measured by generating a set of clusters from

an input layer using the PostGIS DBSCAN algorithm. Each cluster is assigned a normalized score based on the size of the cluster in comparison to other clusters in the layer where the largest cluster in the layer is given a score of 1 and every other cluster in the layer is assigned a relative score between 0 and 1. Cluster scores are assigned to target layer features by checking which cluster contains the target layer feature using the PostGIS st_intersects function. If a target feature intersects multiple clusters, the largest cluster score is assigned to the target feature.

$$Cl_{DA} = Cl_{DA\text{-}intersect}/Cl_{A\text{-}max}$$

ClR_{DA} = Cluster score for feature D from layer B for clusters in layer A

$Cl_{DA\text{-}intersect}$ = Size of the largest cluster in layer A that touches feature D from layer B

$Cl_{A\text{-}max}$ = Size of Largest Cluster in layer A

The formula above shows how normalized cluster scores are generated for clusters in the input layer. The calculated cluster score is then assigned to the target layer features that intersect the given cluster. This methodology leverages the DBSCAN (density-based clustering of applications with noise) algorithm to identify clusters within VGI data. The DBSCAN algorithm is used in GR and VGI literature to perform relevance analysis (De Sabbata, 2013; Spinsanti & Ostermann, 2013), and it is widely implemented in spatial analysis packages like PostGIS. DBSCAN is particularly effective for this use case because it requires minimal domain knowledge of the data; it can discover clusters of arbitrary shape, and it is efficient on large databases (Sander et al., 1998).

The cluster score is assigned to candidate features that are identified using the spatial proximity function. The spatial proximity function identifies three candidate features for every p-footprint, the cluster metric is assigned to each candidate feature, and an aggregate score of spatial proximity and cluster is generated in order to select most geographically relevant feature for a given p-footprint. Once a feature is matched for a given p-footprint, co-location and topicality scores are evaluated for the p-footprint to determine the overall geographic relevance of the matched feature.

4.3.6 Topicality

Topicality is an essential measure of relevance that determines if a given entity is relevant or irrelevant in a given context (De Sabbata & Reichenbacher, 2012). In this methodology, topicality is a measure of semantic relevance between layer type. As noted by De Sabbata “topicality is defined as the extent to which a piece of information …concerns the topic the user is interested in”(De Sabbata &

Reichenbacher, 2012, p. 1497). Thus, a measure of topicality is generated between two entities, a topic of interest and a targeted piece of information. The topicality metric is an inverse measure of semantic distance, which is measured using ontologies such as WordNet. Ontologies are graph data structures that track relationships between words (Machado et al., 2011). Semantic distance is measured by counting the number of nodes needed to traverse from an input word node to a target word node. This methodology leverages the semantic similarity function provided in the NLTK WordNet library (Liu et al., 2012; Pedersen et al., 2004). The semantic similarity function accepts two words as inputs and returns a score between 0 and 1, where 1 indicates that the two words are the same and a score of 0 indicates that the two words are unrelated. An example of the python functions used to generate topicality scores is shown below.

from nltk.corpus import wordnet as wn

park = wn.synset('park.n.01')

trail = wn.synset('trail.n.01')

Park.path_similarity(trail)

In this methodology, topicality scores are calculated for framework layers and then scores are joined to individual p-footprints. The table of topicality scores between the VGI data layer and the framework layers are shown in Table 4.4. Tags, created by the study author, are used to ensure that the NLTK library matches the layer with an appropriate concept in the WordNet library. When multiple tags are available for a layer, the best performing tag is used to measure topicality.

Table 4.4 The set of pre-processed topicality scores used for GR ranking

Input Table	Input Tags	Target Table	Target Tags	Topicality Score
trail_survey	trail	trails	trail	1
trail_survey	trail	parks	park	0.083
trail_survey	trail	roads	road	0.125
trail_survey	trail	bridge	bridge	0.111
trail_survey	trail	railway_lines	railway	0.062
trail_survey	trail	cycling_infrastructure	cycle, bike	0.076

The scores above are attached to p-footprints from the Q-footprint analysis using a layer join that matches the target table of the topicality analysis to the source table of each p-footprint. The NLTK library supporting topicality analysis is an external library that can not be calibrated. For example, according to the NLTK library, bridges are more strongly related to trails than parks. Due to the complexity of the English language, it is not feasible to adjust for all the semantic nuances that can be identified in an unstructured body of the text. As a result, some anomalies should be expected in the GR analysis result due to inconsistencies in the implementation of the topicality metric.

4.3.7 Co-location

The co-location pattern is defined as a set of spatial features that are often located in close geographic proximity (Deng et al., 2017; Huang et al., 2004). Co-location rules infer the presence of one feature based on the presence of another feature; co-location rules are generated by mining spatial data to detect patterns (Barua & Sander, 2014; Huang et al., 2004). Co-location looks at patterns such as resteraunts being located near movie theaters or parks being located near rivers. Co-location assumes that given an entity in the first category, it is probable to find an entity belonging in the second category within a defined distance (Reichenbacher et al., 2016). Co-location rules do not necessarily infer the correlation between two observed entities; rather, they are formalizations of patterns that are observed in the data. Co-location rules can be detected using statistical approaches or data mining approaches (Barua & Sander, 2014; Huang et al., 2004). This methodology uses a distance-based data mining approach to identify co-location patterns. Distance-based data mining approaches for identifying co-location have been discussed in several studies (Deng et al., 2017; Huang et al., 2004); many proposed approaches rely on the use of a user-defined threshold distance (Deng et al., 2017). This thesis implements a simple data mining heuristic to identify co-location rules between an input layer and a target layer using a user-defined threshold distance. Co-location between two layers (A,B) is measured as the number of features in A that are within a predefined distance of any feature in B divided by the total number of features in A. The input data set used in this methodology is the City of Kitchener VGI data set and the target layers are the framework data. The co-location mining heuristic is evaluated between the VGI data set and every framework layer. This methodology measures the percentage of features in the VGI data set (A) occurs within a distance (d) of a given framework layer (B). The co-location score is a percentage score that ranges between 0 and 1. Like topicality scores, co-location is measured between layers rather than individual features.

$$Co_{AB} = N_{AB}(d)/N_A$$

Co_{AB} = Co-location score between layer A and B

$N_{AB}(d)$ number of features in layer A within d (distance) of layer B

N_A number of features in layer A

The formula above is used to generate a ratio that is representative of co-location patterns between two layers, high co-location pattern (1) indicates that features from one layer can often be found near features from another layer. The co-location scores for different tables are assigned to p-footprint records using a join on the p-footprint source table and the co-location target table.

Table 4.5 Co-location scores between the survey data and framework data

Input Layer	Target Layer	Co-location
trail_survey	trails	0.584158
trail_survey	parks	0.659406
trail_survey	roads	0.435644
trail_survey	bridge	0.069307
trail_survey	railway_lines	0.089109
trail_survey	cycling_infrastructure	0.138614

This co-location mining heuristic is a simplification of the methods proposed by other research groups such as Huang et al. or Deng et al. The focus of this thesis is to evaluate the validity of co-location as a GR metric. Therefore a simplified and less accurate co-location mining heuristic is reasonable for this use case. The following section discusses the approach used to combine the GR metrics described in the previous sections in order to generate a GR footprint.

4.3.8 Geographic Relevance Scores

The previous sections have discussed the assessment model used to generate GR footprints. The previous sections have discussed how Q-footprints are generated for VGI data and how GR metrics are applied to p-footprints. The previous sections have discussed the reasoning behind the four GR metrics which are topicality, spatial proximity, cluster, and co-location. This section discusses the process of

combining the normalized scores generated by the four GR metrics. The previous sections have already discussed a significant part of how GR scores and GR footprints are generated; this section briefly discusses the logic and literature behind the approach taken in this methodology and it presents a formula for GR score generation that can be implemented in other systems.

The main problem discussed in this section is the aggregation of GR metric scores; several studies show that different weighting schemas can significantly affect the accuracy of relevance ranking systems (Bordogna et al., 2012; Koukoletsos et al., 2012). For this methodology, a general geographic relevance score can be generated by calculating the mean of all four scores as implemented in the GeoFinder system (Bordogna et al., 2012). However, an equal weighting of scores may overemphasize the importance of spatial relevance metrics such as proximity, cluster, and co-location (De Sabbata & Reichenbacher, 2012). A simple combination of metrics would also ignore the distinction between metrics that are mandatory for determining geographic relevance and metrics that are desirable for determining geographic relevance (De Sabbata & Reichenbacher, 2012). In this thesis, key relevance metrics are the Q-footprint, topicality, and spatial proximity. The q-footprint is the most important metric for determining relevance; it determines if two features have any semantic relevance, and it extracts important location from unstructured bodies of text. Topicality is another key semantic relevance metric that must be positive for two features to be relevant to each other. A topicality score of 0 indicative of significant differences between the data sets being compared. Spatial proximity is another important metric for determining relevance, particularly geographic relevance. Features that exist beyond a threshold distance of a feature are likely not relevant to the study (De Sabbata & Reichenbacher, 2012; Deng et al., 2017). In contrast, metrics such as cluster and co-location are indicative of geographic relevance, but they are not a prerequisite (De Sabbata & Reichenbacher, 2012). Given these assumptions and the differences in the importance of different GR metrics, this methodology implements a GR formula that is a conditional average of the four GR metrics discussed above. Different GR metrics are prioritized in this methodology using the continuous preference logic model that was developed by Dujmovic (Dujmović, 2007) and implemented by De Sabbata and Reichenbacher (2012) for the GR analysis.

The Continuous Preference Logic Model (CPL) is used in this methodology to combine geographic relevance metrics using conjunctive partial absorption (CPA) and disjunctive partial absorption operators (DPA) (Dujmović, 2007). The CPA operator allows the combination of “mandatory” inputs with “desired” inputs using the “and” operator where the mandatory input must be greater than

zero for the score to be greater than zero (De Sabbata, 2013). The DPA operator allows a combination of "sufficient" and "desired" inputs using an "or" operator where the score will be one if the sufficient input is one (De Sabbata, 2013).

The formula used to calculate GR scores does not incorporate q-footprints because they are considered a preprocessing step to evaluating geographic relevance. As discussed in the assessment model section, GR scores are only generated for features within the p-footprint. The GR formula for this methodology assumes that topicality and spatial proximity are essential metrics. If either of these metrics are 0 the entire GR score should be 0.

$$GR = T_{weight} * SP_{weight} * ((T_{AB} + SP_{GH} + Cl_{AH} + Co_{B})/4)$$

T_{AB} = Topicality Score between layer A and layer B

$T_{Weight} = T_{Weight} = T_{AB} > 0 \ ? \ 1 \ : 0.$

SP_{CD} = Spatial Proximity Score between feature C from layer A and feature D from layer B

$SP_{Weight} = SP_{GH} > 0 \ ? \ 1 : 0$

Cl_{AD} = Cluster Score for feature H from layer B based on clusters in layer A

Co_{AB} = Co-location Score between layer A and layer B

A = Input layer A

B = Target layer B

C = Input feature C from input layer A

D = Target feature D from target layer B

The formula above calculates GR scores for p-footprints within the GR footprint, the core metric is a mean average of topicality, spatial proximity, cluster, and co-location. In the formula, a conjunctive partial absorption score is created for topicality and spatial proximity. The formula ensures that if topicality or spatial proximity is 0 or negative, then the conjunctive partial absorption score is negative, if both scores are positive then the CPA score is 1. Both the CPA scores are then multiplied against the core GR score, thus the GR score is only altered by the CPA values if topicality or spatial proximity is 0.

Calculation of the aggregate GR score is the last step in the GR assessment model described in this methodology. An aggregate GR score is generated for each p-footprint within the fuzzy footprint; the aggregate GR score is then used to filter out p-footprints that are not geographically relevant to the input VGI data point. In the UrbanContext platform, the aggregate GR score is generated using the following PL/SQL query.

```
(greatest(0,topicality)/greatest((1*10^(-9)),topicality))*
(greatest(0,spatial_rel_score)/greatest((1*10^(-9)),spatial_rel_score))*
((spatial_rel_score+cluster_score+topicality+colocation_score)/4) as gr_score
```

The result of this calculation is a GR footprint that corresponds to an VGI data point. The GR footprint is composed of a set of p-footprints that have reference information to a feature in the framework database and a corresponding GR score.

Table 4.6: Sample GR-Footprint with aggregate GR scores

FVGI GID	P-Footprint ID	GR Score
371	5796	0.00
371	3071	0.39
371	2867	0.71
371	2867	0.71
371	2609	0.00
371	2120	0.39

Table 4.6 above is a sample GR footprint that has been generated for a single VGI data point, p-footprints with a score of 0 have not been filtered out in this example. This result set is generated for every VGI data point; the resultant data set can be used to filter and visualize the data based on key associations in the data set. The following sections describe how the methodology described in this section can be used to address data processing challenges associated with urban planning. The following section will describe a case study from an urban planning project in the City of Kitchener.

4.4 Case Study: City of Kitchener

In the summer of 2015, the City of Kitchener launched a study to improve a major multi-use trail known as the Iron Horse Trail. As part of this project, the City of Kitchener citizen feedback was gathered through a variety of methods. City staff decided to adopt a digital solution including survey, emails, direct conversation at public meetings, and a map-based online survey. The map-based online survey was hosted on the commercial Community Remarks platform and resulted in over 200 geotagged online responses.

These data were processed in UrbanContext to identify key locations and features that were most relevant to citizen comments within the study context. The following sections describe the Kitchener Iron Horse Trail Study use of UrbanContext and sample result of the UrbanContext analysis. The full set of analysis results are discussed in section 4.5 of the results.

4.4.1 Kitchener Iron Horse Trail Improvement Strategy

The City of Kitchener is a municipality of 255,00 people located in southeastern Ontario, Canada. The region is branded as Canada's technology hub with companies like Google and Square located in Kitchener while the neighbouring City of Waterloo is home to companies such as OpenText and Blackberry. The Iron Horse Trail is a 4.5 km trail that runs through the heart of Kitchener and Waterloo and is a key pedestrian transportation corridor that is used over 250,000 times per year (City of Kitchener, 2015). The Iron Horse Trail is a former railway line that has been repurposed into a trail, thus it intersects numerous streets across its 4.5 km span and the City of Kitchener and Waterloo have grown around the former railway line.

In 2015, the City of Kitchener conducted a study to inform future development of the Iron Horse Trail; the goal was to identify key locations of concern and identify opportunities to improve the trail. A major dimension of the Iron Horse Trail study was citizen engagement and feedback gathering. The City asked its citizens which improvements they wanted to see and what their key concerns where along or near the Kitchener Iron Horse Trail. The goal of the survey was to collect local knowledge about the Kitchener Iron Horse trail and to identify issues with the trail the City was not aware of (City of Kitchener, 2015). The City of Kitchener promoted the Iron Horse Trail consultation process using the municipal website, social media, signage and an interactive map based survey (City of Kitchener, 2015). Feedback was collected from citizen using trail feedback stations, public workshops, public meetings, e-mail correspondence, social media and an interactive online survey (City of Kitchener, 2015). The trail feedback stations, public workshops and public meetings allowed citizens to talk to City staff directly to raise concerns and citizens were also encouraged to leave written comments to review. Citizen feedback was collected by City staff using forms that allowed citizens to categorize the nature of the feedback and to submit unstructured comments about concerns or ideas related to the Kitchener Iron Horse trail. City staff and volunteers manually transcribed the comments into a central excel spreadsheet that included columns to organize and categorize collected data. The excel spreadsheet included an ID field, a Comment Source Field, a Comment field and several other columns to categorize the data based on categories such as lighting, trail surface, maintenance, signage, amenities and more.

The interactive online survey was a map-based application built on top of the Community Remarks platform. The Community Remarks platform is a web based public engagement and surveying tool that is used to support planning projects and processes in over fifty government organizations (PlaceVision Inc., 2019). As shown in Figure 4.5, citizens can see proposed projects on an interactive map and then click on different locations on the map to submit specific comments about locations of concern or interest within the proposed project. Citizens can submit categorized comments, respond to comments or vote on comments within the application. Government staff can use the application to identify where issues exist and broadly identify topics of concern such as safety or wayfinding.

Figure 4.5: Screenchot of the City of Kitchener Survey, built using Community Remarks.

The City of Kitchener created an interactive map based survey using Community Remarks and made the survey available to the public on the City website. As shown in Figure 4.5 and 4.6, the application allowed users to find an area of interest on Google Maps, drop pins on the map and identify pre-identified topics of concern such as signage, safety and more.

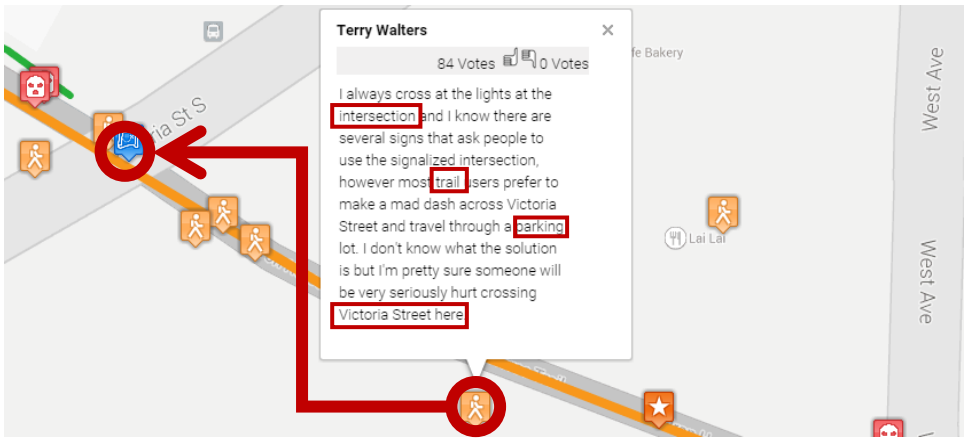

Figure 4.6: Sample comment from the Kitchener Community Remarks app that references multiple locations

As shown in Figure 4.6, the interactive survey allowed citizens to provide detailed feedback about locations and features of concern. Many comments from the app referenced multiple locations near the comment and responses often covered multiple topics such as discussing how citizens currently use the trail to identifying safety issues. The number of topics and locations discussed in a comment varied between users which contributed to heterogeneity of data. The comments from the interactive survey were manually reviewed by City staff by downloading the comment data from the Community Remarks application as a CSV and using Excel to manually review and categorize individual comments.

Table 4.7: Summary table of feedback received over various engagement channels (Josh Joseph, 2015).

Type of Public Consultation	Comments Received
Trail Feedback Station at Glasgow (May 27, 2015)	**84**
Trail Feedback Station at Courtland (June 3, 2015)	**20**
Trail Feedback Station at Queen (June 9, 2015)	**20**
Public Workshop (June 8, 2015)	**241**
Public Meeting (June 23, 2015)	**214**
Interactive Trail Survey (May 15 – July 2, 2015)	**243**
Email Correspondence	**14**
Social Media	**36**
Other (walking tours, group discussions, etc.)	**12**
Total Comments	**884**

The City received 884 remarks related to the Iron Horse Trail Improvement Strategy from all the engagement channels (Josh Joseph, 2015). Of the 884 community remarks, 243 geotagged remarks were collected using the Community Remarks platform (Josh Joseph, 2015). However, City staff provided 505 geotagged comments for this research project, it is assumed that the City only used a subset of the data collected on Community Remarks for the official staff report.

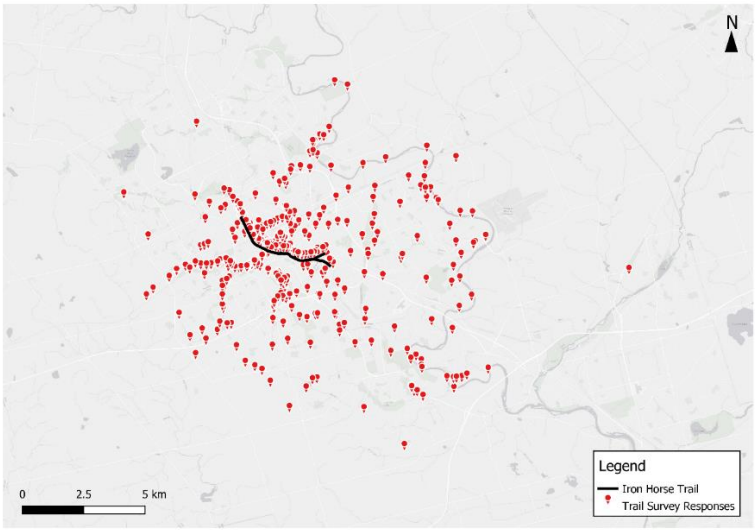

Figure 4.7: Geotagged comments (VGI data) collected by the City of Kitchener

Figure 4.7 illustrates some of the challenges associated with the survey data collected by the City. In Figure 4.7, the Iron Horse Trail is delineated by the black line in the center of the map and citizen responses are delineated by red pins which are distributed across the City of Kitchener. Many comments are spread across the City of Kitchener and many of these comments are not always focused on the Iron Horse trail. As shown in Figure 4.8, the Community Remarks application attempts to guide users to provide relevant information by centering the map application on the Kitchener Iron Horse trail and offering categorized comment pins so that users can focus on specific topics that are relevant to the survey. However, the Community Remarks application does not force users to conform to response standard. Therefore, the survey response data set contains a lot of information about locations, facilities and services across the City in addition to feedback on the Iron Horse Trail.

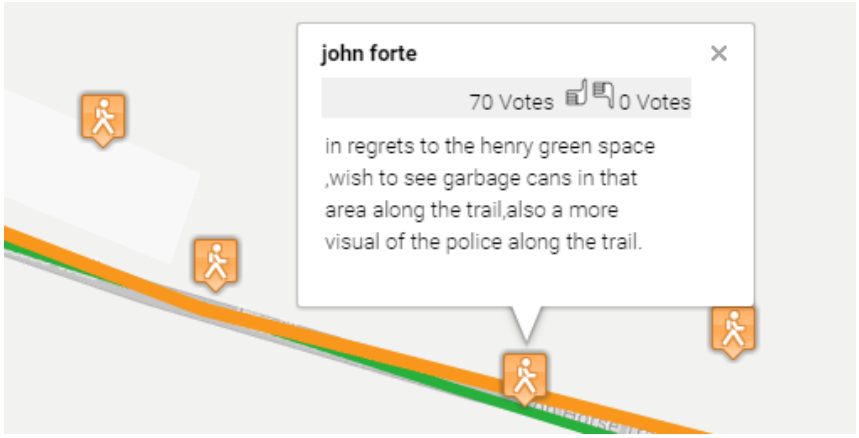

Figure 4.8: Screenshot of CommunityRemarks application used to collect citizen feedback

Figure 4.8 demonstrates how comments are not necessarily focused on the survey topic, this comment discusses a green space near the trail in addition to policing on the trail. City staff reviewing this comment would need to be aware of the amenities near this particular location to understand the concerns discussed in this comment. If staff are not familiar with the locale, reference maps have to be reviewed to understand the comment. Reviewing each comment using reference maps for over 800 comments is a time intensive task that is prone to errors. City staff reviewed the comments received over multiple engagement channels and created summary maps and figures that identified key areas along the trail and general topics of concern. Figure 4.9 shows a sample of the survey comments that were classified by the City, the comments with the green arrow icon refer to trail improvement opportunities while the purple exclamation mark indicates safety concerns.

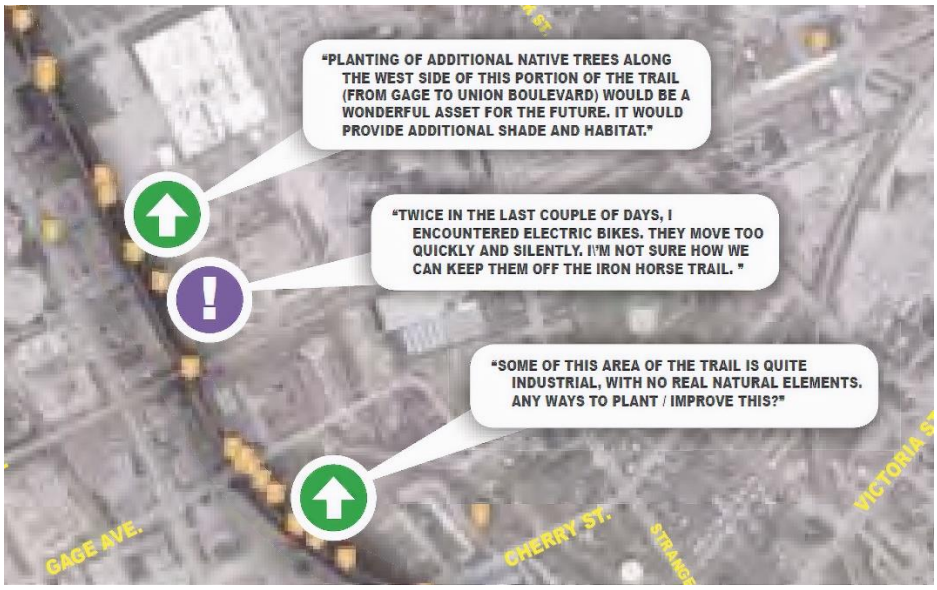

Figure 4.9: Sample of map-based survey results from the City (City of Kitchener, 2015)

The city identified four major categories to organize the data which include physical aspects of the trail, safety, and accessibility, improvement opportunities, and network connectivity. Some categories are references to spatial objects such as trail-road crossings or the LRT, while other comments are topical such as safety concerns. The results of this summary are included in a City of Kitchener staff report which was submitted to the council in 2015. City staff summarized feedback around three major portions of the trail and recommended that the City prioritize the central section of the Iron Horse Trail based on citizen feedback.

Figure 4.10 Summary figure of responses from the Kitchener Staff report (Josh Joseph, 2015)

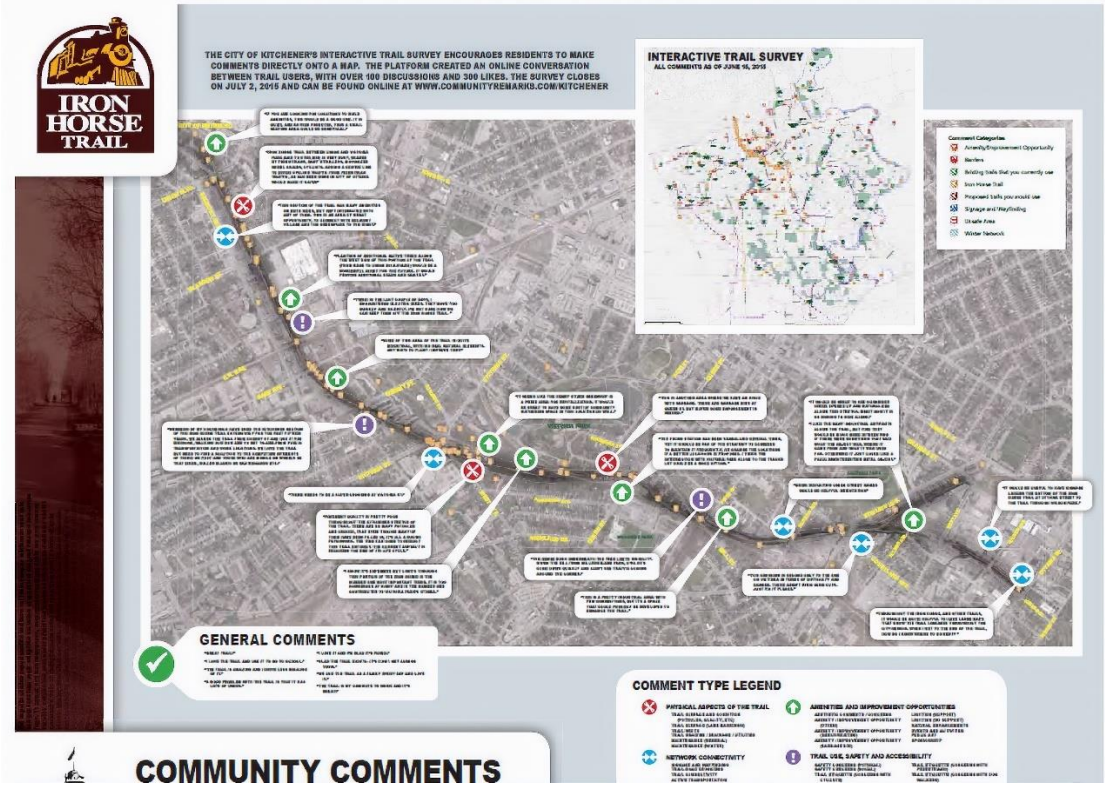

Figure 4.11 Summary map of manually classified citizen responses (City of Kitchener, 2015).

Figure 4.11 shows a more detailed map of categorized citizen comments generated by City staff. The official report used these maps to identify key topics of concerns and opportunities in order to present City council with a strategy that recomends where to allocate resources and how to allocate reources. The following sections discuss how UrbanContext was used to automate the process of finding key locations of concern using citizen response data provided by the City of Kitchener. The purpose of

this case study is not to propose a method to automate the process of reviewing citizen comments. The purpose of UrbanContext is to aid planners and City staff to organize and filter content as they review comments. The intent is to reduce the time needed to review comments and increase the accuracy of summary reports to better reflect the sentiments of the community.

4.4.2 UrbanContext Case Study

In this case study, the UrbanContext application processes the VGI data that was collected by the City of Kitchener for the Iron Horse Trail improvement strategy study. During the public engagement period of the study, the City of Kitchener collected over 500 geotagged comments which were provided to this thesis for analysis. Due to the compute-intensive nature of the GR analysis, only a subset of the VGI data centered around the Iron Horse Trail was used in this analysis. The analysis data set was selected by creating a 3km buffer around the Iron Horse Trail and selecting the survey responses that are contained by the buffer. The resultant analysis data set contains 200 VGI data points. Figure 4.12 below illustrates the distribution of the filtered data.

Figure 4.12: The filtered Iron Horse Trail Survey responses.

The filtered VGI data set makes data analysis more manageable and it also reduces the need for the use of large data sets that cover the entire extent of the survey. The filtered data creates a more

manageable study scenario, but it does not eliminate the complexity of the unstructured data or reduce the amount of noise and error in the data. This thesis analyzes 209 VGI data points by comparing the VGI data to six framework data sets acquired from the Kitchener Open Data portal. Each VGI data point has three attributes which include id, an unstructured body of text and geometry data, a sample of data records is shown in Table 4.8.

Table 4.8 A sample of three data points from the analyzed VGI data set

gid	comment	point
2	A better crossing at West and Victoria. Not many bikers are going to go to the lights at such a busy corner.	(540206.29,4810166.85)
4	a connection to the Laurel trail would be great	(538733.53,4812033.55)
5	Add an automated, diagonal (scramble) crossing here (Stirling/Courtland).	(541728.70,4809959.78)

The case study uses VGI data provided by the City of Kitchener and reference data acquired from the City of Kitchener's open data portal. The reference database used in this thesis contains the data sets (layers) described in Table 4.9. The reference data sets shown in the table below were selected due to their perceived relevance to trails in the City of Kitchener and references to these features within the VGI data set. Manual review of VGI comments showed that trails, parks, roads, railway lines, and bridges were often directly or indirectly referenced in the geotagged comments created by City residents. The data sets in the table include features that contain the Iron Horse Trail (parks) or intersect the Iron Horse Trail (roads, railways, etc.).

Table 4.9: Reference datasets from the City of Kitchener.

Name	Description	Record Count
parks	Parks data maintained by Kitchener GIS	438
railway_lines	Railway data maintained by Kitchener GIS	181
cycling_infrastructure	Cycling infrastructure data maintained by Kitchener GIS	1093
trails	Kitchener trail network data maintained by Kitchener GIS	3530
bridge	Bridge (point) location data maintained by Kitchener GIS	135
roads	Kitchener road network data maintained by Kitchener GIS	6057

This case study analyzed 209 VGI data points to generate 209 GR-footprints which are composed of 682 p-footprints. The GR-footprints are then used to create maps that identify which framework features and location are referenced in the VGI data set and which locations are most important in the framework data set. The following sections describe the results of the analysis as well as discuss the approaches used to review and validate the data.

4.5 Results

This section explores the results of the UrbanContext analysis through two main sections. The first section discusses the process used to validate the UrbanContext output and explores individual GR footprints to get a better understanding of findings. The second section discusses the results and the summary findings of the analysis. This includes a mapping the most important locations identified by the UrbanContext and graphing the composition of fuzzy footprints and GR footprints. This section will also provide a summary description of the observed accuracy of results generated by the UrbanContext platform. The tables and figures in the following sections depict fuzzy footprints. Therefore tables will contain negative proximity and GR scores. References to GR-footprints only apply to records that have positive GR scores. Verifying and summarizing the outputs of the UrbanContext is a non-trivial challenge, even with a small concentrated dataset there are a significant number of records that can only be verified through manual review. Reichenbacher et al. (2016) approached the challenge of reviewing data by enlisting volunteers, while Marzouki et al.(2018) reviewed analysis results using a qualitative review that involved selecting random samples of data and manually reviewing information using a proposed conceptual model. Spinsanti and Osterman evaluated the GeoCANVI system by comparing reports of fire incidents detected in social media against official reports of fire incidents (Spinsanti & Ostermann, 2013). Review of UrbanContext draws from approaches outlined by Marzouki and Spinsanti (Marzouki, Mellouli, & Daniel, 2018; Spinsanti & Ostermann, 2013). First, a random sample of 50 VGI data points and associated GR-footprints across the study site are reviewed for spatial and semantic accuracy. Section 4.5.2 will compare the aggregate results of the GR analysis to results reported by the City of Kitchener. This will compare the key locations identified by the UrbanContext application to key locations identified by the City of Kitchener to determine result coverage.

4.5.1 GR-Footprint Accuracy

The semantic and spatial accuracy of GR footprints were reviewed by selecting a sample of VGI data points and their associated GR-footprints. In this review, each VGI point is mapped with contextual data, and the comment is manually reviewed to identify location references within the text and to

match it to real features on the map. Then each GR-footprint is visualized using the UrbanContext web application to inspect the number of correct p-footprint matches, the number of missed p-footprints and the number of false p-footprints. A correct p-footprint is a p-footprint that is generated by a legitimate location reference found in the VGI comment such as *"Victoria Park"* rather than *"my friend Victoria"* which is also matched to the correct framework feature such as the Victoria Park polygon within the Kitchener parks data layer. As noted by Goodchild and Li (2012), spatial data validation is very subjective due to variances in perspective. This review focuses on validating p-footprints based on the GR concept of information need; where a p-footprint is considered correct if the spatial data feature is considered to be relevant to the VGI feature. The number of missed p-footprints is determined by identifying discrepancies between p-footprints that were manually identified and are missed by UrbanContext. False p-footprints are p-footprints that were generated using false location references, or they are p-footprints that failed to meet spatial relevance criteria. These are p-footprints that were not identified as relevant in the manual review but selected as relevant by the UrbanContext system.

To illustrate the review process, a detailed description of how three sample VGI points and their associated GR footprints were validated is provided in the following paragraphs. Accuracy of results vary across different data points, and the samples described below are not representative of all results but they provide some insight on results produced by UrbanContext. The paragraphs below will present the location of the VGI point on a map, the VGI comment text, and the GR-footprint table. The VGI point context map will also visualize GR-footprint features as well as the VGI clusters layer for context. A discussion will accompany each VGI point on observations regarding the footprint accuracy. The data review samples will be followed by a summary of all the data review results from the sample of 15 VGI points and their associated GR-footprint.

Figure 4.13 shows VGI data point 193 (gid), this VGI data point is located at the bottom edge of the study extent, and it discusses the connection between cycling routes along Nyberg Street and the Kitchener Iron Horse Trail. This VGI data point is part of a larger cluster of VGI data points that focus on issues regarding the connection between the Kitchener Iron Horse Trail and the bike trail on Nyberg Street.

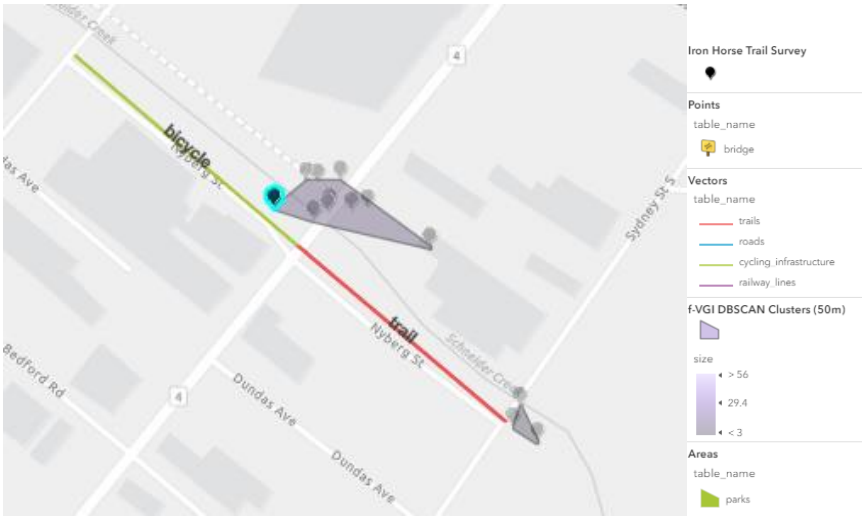

Figure 4.13: Comment (193) local bike trails and the Iron Horse Trail. (https://urban-context.glitch.me/)

The text associated with the VGI data point is shown below, the highlights indicate location references that were identified by manual review. The comment is generally vague and does not make direct references to any particular feature; the trail and bike trails are generally relevant to the comment. The location and feature references used to generate p-footprints for this comment are highlighted in yellow below.

> *"I use the trail often to get to and from work on my bicycle in the warm months. It is a great way to get to work while staying 'in the woods' and away from traffic and to wind down at the end of the day. The number of people using the …"*

UrbanContext matched the VGI point in Figure 4.13 above to two features. One feature is a trail segment that belongs to the Iron Horse Trail, this feature matches the "trail" entity reference in the comment above. The matched trail is shown in Figure 4.13 as a red line segment south of the highlighted comment point. The other match is a segment of the cycling network which is depicted as a green line segment in Figure 4.13, the bike network segment is matched due to the bicycle reference in the comment above. The trail and bike features are shown in Figure 4.13 were selected using the methodology described in section 4.3. The gazetteer scanned the comment and identified the "trail" and "bicycle" terms as part of the q-footprint. As shown in Table 4.10 the "trail" and "bicycle" term was matched to the trail and cycling infrastructure layers and then features from the layers were selected using GR criteria such as proximity co-location, cluster, and topicality. The scores for the two matches are shown in Table 4.10.

Table 4.10: GR Analysis scores for VGI comment 193

gID	Table	Column	Search Term	Search Category	Proximity	Co-location	Cluster	Topicality	Geographic Relevance
193	trails		trail	Layer	0.49	0.58	0	1.00	**0.69**
193	cycling infrastructure		bicycle	Layer	0.50	0.14	0	0.08	**0.24**

The scores above show that the trail segment in Figure 4.14 is more relevant than the cycling infrastructure segment based on a score of 0.69 and 0.24 respectively. The QC process identified two valid location references in the comment. UrbanContext found both location references and matched the p-footprint to the most reasonable trail and bike features within the vicinity of the VGI point. Therefore, this result has 2 correct matches, 0 false matches, and 0 missed matches. The GR result table indicates the cluster score is 0 for both p-footprints; this is reasonable because neither feature intersects an VGI cluster. Thus proximity, co-location, and topicality are the most important metrics in this analysis.

The second sample VGI comment is a more complex comment located at the north edge of the Iron Horse Trail. The comment contains multiple location references that are within proximity of the VGI data point and location references are well defined in the comment.

Figure 4.14 The GR-footprint for VGI data point 371, 3 valid p-footprints were matched

UrbanContext matched the comment above to three GR-footprint features as shown in Figure 4.14 despite initially identifying 5 p-footprint candidates as shown in Table 4.11. The location references identified by the gazetteer are highlighted in yellow in the comment below and listed in Table 4.11. The rows that are highlighted red indicate records that were initially matched but later filtered out using GR metrics.

> *"This area, at the north end of Belmont at Union is a field. There are many trees. It could be re-landscaped and redesigned, so that trees hide the backs of the buildings, and it provides an entrance into Belmont Village from the trail."*

It appears that cluster scores played a larger role in this GR analysis as they show more variation. However, on further inspection, it appears that the trail segment was improperly assigned cluster score because the segment has a score of 0.76 but the segment does not intersect any clusters.

Table 4.11: P-footprint VGI data point 371, the red rows are filtered out due to negative proximity scores

gID	Table	Column	Search Term	Search Category	Proximity	Co-location	Cluster	Topicality	Geographic Relevance
371	roads	street_nam	BELMONT	Attribute	0.49	0.44	0.52	0.13	**0.35**
371	roads	street_nam	UNION	Attribute	0.52	0.44	0.52	0.13	**0.36**
371	roads	street_nam	NORTH	Attribute	-0.46	0.44	0.52	0.13	**0.03**
371	roads	street_nam	VILLAGE	Attribute	-1.11	0.44	0.52	0.13	**-0.18**
371	trails		trail	Layer	0.47	0.58	0.76	1.00	**0.69**

This comment had no ambiguous location references, but the UrbanContext system identified false p-footprints due to errors in place name disambiguation. As shown in Table 4.11, North and Village were falsely identified as streets, but the GR proximity and cluster ranking module were very effective at generating relevance metrics that indicate that the features are not part of the context of the VGI comment. The QC process identified five valid explicit and implicit location references. Three spatial entities were accurately matched. Two additional location references were matched, but the GR rating system pre-emptively identified these entities as false positives.

The third sample comment is a complex comment because it has multiple implicit and explicit location references. The comment also makes valid references to locations that are a significant distance

from the VGI data point. This comment highlights the variances in content and structure that are endemic to VGI and VGI data.

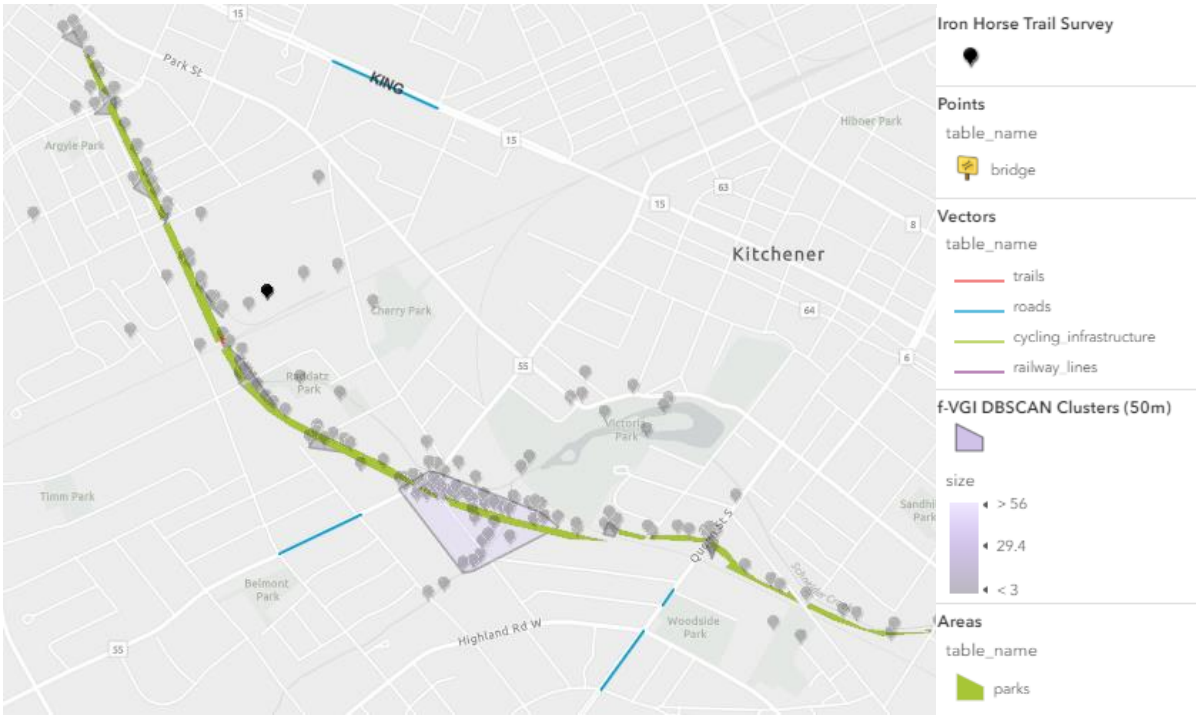

Figure 4.15 Comment 97, the scale of this VGI comment varies significantly from other comments

Figure 4.15 shows and VGI comment that is located near the middle of the study area and some distance from the trail and any comment clusters. The location references identified by the gazetteer within the VGI comment are highlighted below and listed in Table 4.12.

> *"I am glad to see that this proposed trail is here, as there are really few good ways to bike right downtown from the iron horse trail. Neither Victoria nor Queen street have bike lanes all the way to king street. There should be AT LEAST ONE primary trail"*

The UrbanContext system matched the comment in Figure 4.15 to seven features in the framework data as shown in Table 4.12. Three of the matched locations reference trails, two of the matches are trail entities while one match is a park entity. The analysis gave the trail entities higher GR scores due to higher proximity, cluster and topicality scores. The analysis also failed to recognize key location adjectives such as 'proposed trail' and 'primary trail'.

Table 4.12: GR footprint for VGI point 97, red rows are filtered features due to negative proximity scores.

gID	Table	Column	Search Term	Search Category	Proximity	Co-location	Cluster	Topicality	Geographic Relevance
97	trails	route_name	IRON HORSE TRAIL	Attribute	0.54	0.58	0	1.00	0.71
97	trails		trail	Layer	0.41	0.58	0	1.00	0.67
97	roads	street_nam	VICTORIA	Attribute	0.65	0.44	0	0.13	0.40
97	parks	park	IRON HORSE TRAIL	Feature	0.42	0.66	1	0.08	0.39
97	roads	street_nam	KING	Attribute	0.13	0.44	0	0.13	0.23
97	roads	street_nam	QUEEN	Attribute	-0.26	0.44	0	0.13	0.10
97	roads	street_nam	STREET ONE	Attribute	-2.33	0.44	0	0.13	-0.59

The other four p-footprints are street entities that earned very low GR scores. The system effectively filtered out the false location reference of 'Street One' but it also filtered out the valid Queen Street location reference. The other valid location references of Victoria and Queen street were not filtered out, but they received very low GR scores.

The following paragraphs depict clusters of GR-Footprints generated by UrbanContext. The exact GR metrics and scores associated with the following figures can be reviewed in Appendix B and the visualizations can be explored using the UrbanContext web application. The following paragraphs look at groups and clusters of VGI comments to review the consistency of results and effectiveness of GR metrics. The review looks at consistency and reasonableness of GR-footprint matches, it does not look at individual GR scores or GR rankings of features. Q-footprints, proximity, and cluster are primary metrics used to filter and match features while topicality and co-location are used to determine relevance ranking of features. The following paragraphs focus on the filtering and matching of features. Thus q-footprints, proximity scores, and cluster scores will be the focus of the review.

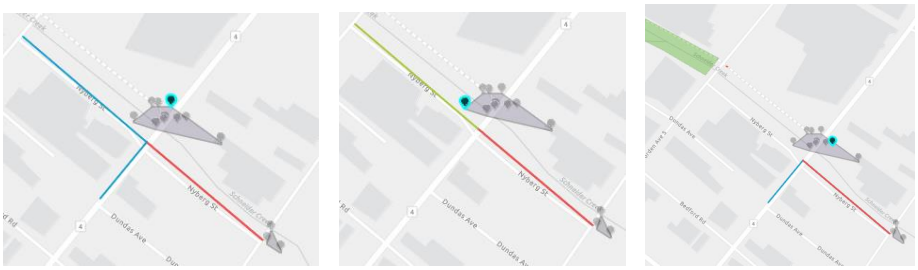

Figure 4.16: Comments 113, 193 & 189 reference trails, parks, and cycling networks

Figure 4.16 presents three VGI points located at the southern end of the Iron Horse trail; most of these comments discuss the connection between the trail and the cycling network at the intersection.

UrbanContext was effective across the cluster at consistently identifying key features that were discussed by all comments. UrbanContext was particularly effective at selecting features from different layers that intersect as described in the comments. The figures reinforce the idea that the cluster metric has significance, and VGI data points tend to cluster around features that are considered important. Figure 4.16 also reflects how complicated some comments are, three or more connected location references are made in all comments.

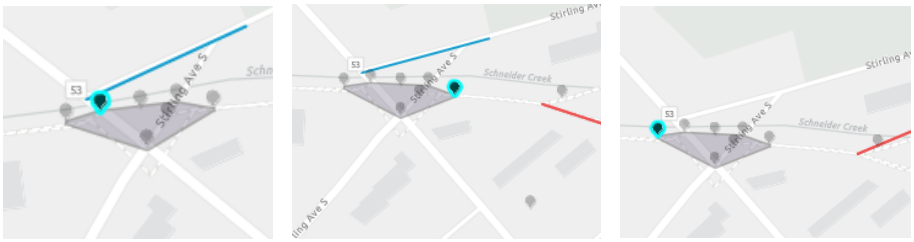

Figure 4.17 Comments 8, 242, 47 reference road and trail segments

Comments in Figure 4.17 discuss the Iron Horse trail in relation to the road. This cluster of results show inconsistencies in the selection criteria because trail segments seem to be erroneously matched. The road segments selected by UrbanContext appear to match the GR criteria from section 4.3 where p-footprints are selected based on the q-footprint and proximity, the cluster metric does not apply because the referenced road segment does not intersect a cluster. The trail segments selected by UrbanContext do not appear to match GR selection criteria because the selected trail segments are not the segments that intersect the cluster or have the highest proximity score. The cluster metric seems to have been ineffective for the results in Figure 4.17. It is possible that cluster scores were erroneously assigned or issues with the segmentation of geometry features are causing attributes from a feature in a cluster to be assigned to features outside of the cluster.

Figure 4.18 Comments 63, 19 & 348 reference trails, parks, roads, and railway segments.

Figure 4.18 looks at a set of VGI data points located in the central section of the Kitchener Iron Horse trail along downtown Kitchener and Victoria Park. The cluster of comments selected is close to the largest cluster detected in the study, the size of the neighbouring cluster should influence the selection of features. However, the UrbanContext application predominantly matches VGI data points to trail and park features that touch the comment cluster indicating that the proximity metric is prioritized over the cluster metric. All three comments in Figure 4.18 have large extents for their GR footprints where several features referenced in the comments are located a significant distance from the comment. The middle image in Figure 4.18 above depicts one such example where a railway line is matched to the comment despite the distance between the two features. The match is wrong because the comment discusses an intersection between the trail and the railway line and it is unclear why that segment was matched by UrbanContext. In general, valid references to significant features that are distant are difficult for UrbanContext to handle because the proximity threshold filters out distant features even if they are valid matches. The UrbanContext system would benefit from a model that helps determine threshold distance based upon the importance of features. Figure 4.18 indicates that there are some inconsistencies in the systems' approach to using GR metrics to match features, it is unclear whether this inconsistency is created by flaws in the conceptualization of GR metrics or if these results are caused by software implementation errors.

4.5.2 GR Summary Results

Evaluating the validity of the UrbanContext is challenging because of the challenges associated with visualizing one to many relationships and the subjective nature of linking a location reference to a place. Even humans who review the VGI data must be familiar with the study area to match VGI comments to the referenced location. In this analysis, 50 random individual VGI points and their associated GR-footprints were manually reviewed using the UrbanContext web application, the evaluation resulted in the review of 196 individual p-footprint records. Each p-footprint was compared to the text content of the VGI data point as well as the spatial context of the geotag to determine if the correct semantic term was used to match data and if the VGI point is matched to the correct geometry. The semantic and spatial quality checks are boolean measures where matches are either considered correct or incorrect based upon the reviewer's judgment and understanding of the spatial context. The overall result table from the review can be found in Appendix B. Figure 4.19 presents the error rate of semantic matches and spatial matches. Semantic matches refer to correct place name matches for the q-footprint where UrbanContext correctly identified place names or location references within the comment to generate the q-footprint. The review shows that UrbanContext had an 82% accuracy when

generating q-footprints but it's important to note that the most common reference in the data set was "trail". The UrbanContext application was effective at identifying simple location references such as trail or park but complex multi-word location references tended to result in errors.

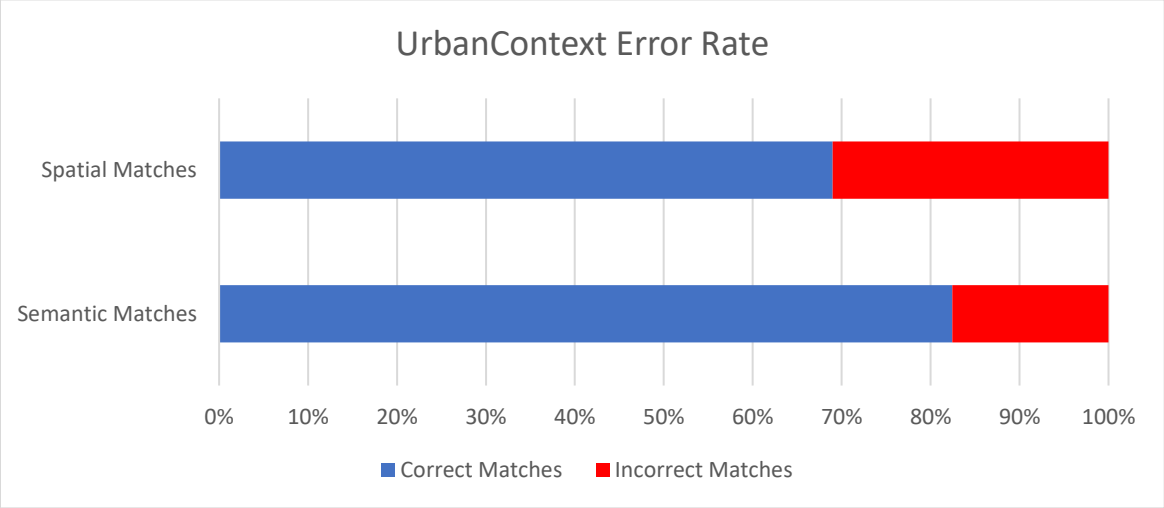

Figure 4.19 The error rate of semantic matches and spatial matches in UrbanContext.

Correct spatial matches are determined using the UrbanContext Visualization app and to compare location references in the comment to surrounding features. If the p-footprint fits the semantic description of the location reference and if the feature geometry is located in a reasonably correct place it is considered correct. As shown in Figure 4.19, the percent of correct spatial matches of p-footprints for VGI points is 69%. The spot checks of data indicated that the cluster metric did not work consistently and there are several cases where the UrbanData platform selected distant features over close features as p-footprints. Some of these errors may be caused by the way geometry is structured in the framework database where lines and polygons can be represented as multi geometries that share attributes. This can lead to incorrect assignments of relevance scores to spatial features.

The summary map below is a sample application that depicts approaches that can be used to create summary maps and visualizations using the outputs of the UrbaContext analysis. The application below shows the most important locations in the City of Kitchener according to UrbanContext. The bubbles are linked to framework features such as parks, trails, roads and more. The size of the bubbles represents the number of VGI comments that refer to the location and the color represents aggregate geographic relevance scores from all the VGI data points that reference the location. The application was made using aggregate operators such as sum() and count() on the GR-Footprint output tables. This application depicts how UrbanContext results can be used in planning studies, the summary map depicts

a set of key locations and features across the study site that should be prioritized in the planning process. The official Kitchener Iron Horse Trail Improvement Strategy staff report identifies several key locations and important sections of the Iron Horse trail to direct investment and planning resources. These locations are manually identified by city staff by reviewing feedback provided by citizens. The report divides the Iron Horse Trail into the Northern section between Victoria and Union, the central section between Victoria and Queen and the Southern section between Queen and Ottawa. The central section is identified as the primary priority of the trail improvement plan. Key locations identified in the report by Josh Joseph (2015) include;

- Victoria Park
- Downtown Kitchener
- Victoria Street
- Cherry Street
- Borden Street
- Gildner Green
- Radatz Park
- Small Parks
- Intersections

The report generally identified road and trail intersections as a safety concern that need to be addressed. Parks and green spaces near the trail are identified as opportunities for improvement (Josh Joseph, 2015). The report also identified the central portion of the trail between Victoria and Queen as the most important section of the trail for future infrastructure projects (Josh Joseph, 2015).

Figure 4.20: Most important locations discussed in the survey (urban-context-summary.glitch.me)

The summary map in Figure 4.20 highlights the most important location identified by UrbanContext. The application was able to identify and prioritize many of the key locations identified in the report such as Victoria Street, Borden Street, and Victoria Park. The application also highlights numerous locations that are not identified in the report, some of these locations are erroneous matches and others are valid locations that are not prioritized in the report. Figure 4.20 shows key locations

identified by UrbanContext near Victoria Street and downtown Kitchener. The most important features identified in Figure 4.20 are Victoria Street, two Iron Horse Trail segments near Victoria Street and Victoria Park. The Victoria Street segment is referenced 23 times and has an aggregate GR score of 7.5 while Victoria Park has three references and an aggregate GR score of 1.2. The full summary map in Figure 4.21 illustrates how UrbanContext was able to identify key roads, parks, and features across the whole study site. The summary highlights numerous valid locations, but some of the most important features in the summary map are false matches. The most common observed errors are matching adjectives such as green, north or west to street names such as Green Street or North Street. In general, the application was effective at identifying important locations in the report, but the results contain a lot of noise and some locations are not prioritized as much as they should be.

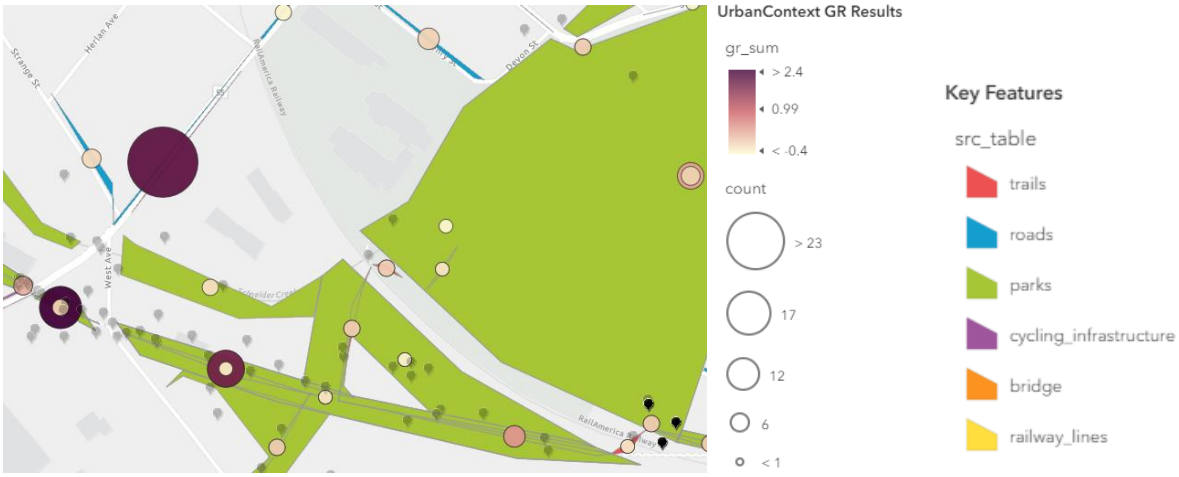

Figure 4.21 The three images highlight major locations from the analysis.

The staff report also identified several safety concerns along the trail but specific locations were not identified (Josh Joseph, 2015). In general, survey responses indicated that trail and road intersections are safety concerns (Josh Joseph, 2015). This map is generated by filtering VGI data using a keyword search on *"danger"*. The UrbanContext dangerous locations summary map shown in Figure 4.22 effectively identifies road segments that are of concern such as Borden Street or Victoria Street, the application does not correctly identify specific segments and intersections consistently. UrbanContext is reasonably effective at matching discrete point and polygon geometries such as parks or bridges but matching specific trail or road segments have been inconsistent. The UrbanContext results are enough for a general summary map but a human review is required to identify specific locations and features where resources need to be directed.

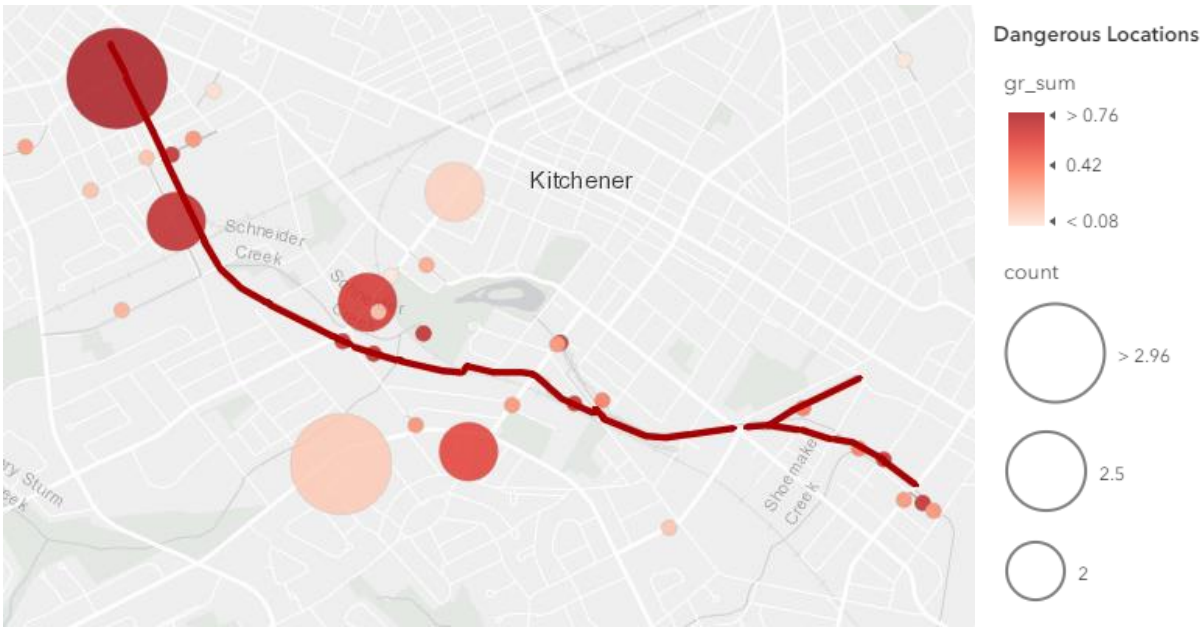

Figure 4.22 A map of GR-Footprints that contain the danger keyword (urbancontext-safety.glitch.me)

The UrbanContext summary map shows promising results that compare well against official studies of the same data set. However, the results contain a lot of noise and there is still a need for human interpretation of results. There are opportunities to reduce noise and improve the reliability of results using better data, and there is room to improve the implementation of individual GR metrics.

4.6 Discussion and Conclusion

This Chapter has developed a GR model to organize and filter heterogeneous and unstructured VGI data. The UrbanData model discussed in this Chapter built on ideas presented in Chapter 3 to address a data analysis problem that is an order of magnitude more complicated than the problem explored in Chapter 3. The challenge addressed in this Chapter was using GR metrics to match heterogeneous and unstructured geotagged comments to features and locations within a framework database. The GR model developed in this Chapter had to address challenges such as managing many to on relationships, filtering noise, accounting for spatially inaccurate data and handling vague location references. The UrbanContext model was not expected to handle all of these challenges completely, it is a test model to evaluate the feasibility of using GR metrics to handle some of these challenges. The UrbanContext case study demonstrated an approach to couple query footprints with GR metrics in order to handle multiple location references within a free form comment and to evaluate geographic relevance between one to many objects. The case study demonstrated that there is merit to this approach as several analysis results effectively replicated human judgement. However, the system was

prone to noise and it was not effective at filtering out false matches in the relevance analysis. The GR metrics did not work as effectively as desired to improve relevance analysis and there is a need to revise the GR analysis model used in UrbanContext. It appears that q-footprints are an effective approach to handling multiple location reference and supporting relevance analysis for one to many features. The gazetteer used to generate q-footprints can undoubtedly be improved to handle complex location references, lexicon and taxonomies. However, handling such complexity requires large volumes of data and resources to build ontologies that can account for differences in location references, user perspective and terminology. It is very difficult to address all the complexities of this problem and it is beyond the scope of this thesis. The q-footprints worked as desired in the UrbanContext model and future work looking at similar problems would likely benefit from adopting this approach.

The effectiveness of the GR metric is not clear as results varied across the data set. It may be possible that GR metrics such as cluster, co-location and topicality are not useful indicators of relevance in this study context, but there is reason to believe that these metrics can be good indicators of relevance if the implementation of the GR model and the GR metrics are improved. Improving the GR model requires an improved understanding of GR metrics through parameter testing and iterative evaluation model using step-by-step processes. Parameter testing could involve testing multiple values to determine optimal values for parameters such as the proximity threshold distance, the co-location search distance or the cluster size and distance thresholds. Parameter values can also potentially be dynamically generated by evaluating characteristics of the input data set. Evaluating criteria such as the average distance between features, the density of features, the average size of features and the extent of the layer are all potential characteristics that can be used to determine parameters such as the proximity metric threshold distance and more. GR metrics should also be evaluated at different scales. The proximity and cluster metric where implemented between features while the co-location and topicality where implemented between layers. A possible improvement for UrbanContext would be to evaluate relevance between feature categories when it is not possible to evaluate relevance between individual features. An example of comparing feature categories could be assessing co-location patterns between VGI comments that reference safety and road classes such as highways or avenues. Similarly, the topicality metric can be improved by categorizing VGI comments using topics such as ‘biking’ and comparing the semantic similarity between comment categories and feature categories such as ‘biking’ and road classes such as ‘highway’. Changing the GR analysis model in this manner would increase the complexity and computational cost of analysis but it should result in better relevance scores that can be used to differentiate the relevance of features. Future improvements on the UrbanContext model

should consider comparing GR metrics to ensure that metrics are not redundant or being double counted. The UrbanContext study did not indicate that any of the metrics where redundant, but future tests may indicate that there is a need to adjust the equal weighted score aggregation model.

From an application perspective, the UrbanContext case study demonstrated that there is potential to use relevance analysis to support public engagement and urban planning. The UrbanContext data processing application and the UrbanContext Viz application linked data in a manner that made it far easier to review and verify contents of individual comments and it provided guidance on identifying key locations from the survey comments. UrbanContext can not be relied upon to independently produce summary figures without oversight due to the prevalence of noise in the output. However, the UrbanContext system is not designed to operate independently because planners should be a part of the comment review process. At small scales, the UrbanContext system will likely be unnecessary. But, if the volume of online engagement increases there will be a need for a system like UrbanContext to handle increased data loads. The Kitchener case study showed that a relatively small city like Kitchener can generate large volumes of content through public engagement. Larger metropolitan cities can likely generate double or triple the volume of data assessed in this study. Overall there is potential to significantly improve the GR model presented in this chapter and the case study indicates that advances in GR could make it easier to analyze and summarize the growing volumes of unstructured VGI data produced on the web.

5 Conclusion

The goal of this thesis was to identify a general set of geographic relevance metrics that could be used to evaluate spatial relevance between structured and unstructured spatial data in different context and at different scales. The basic assumption is that there are a general set of geographic relevance metrics that can provide a comprehensive model for measuring relevance to support spatial data management tasks such as data matching, data filtering and data sorting. The case study analyses demonstrate that GR metrics have considerable promise to aid real world data relevance challenges, however further refinement of the methods is needed. Conceptually, it is likely possible to define GR using a general set of criteria. The models built in Chapter 3 and Chapter 4 were developed using a common set of basic metrics which include topicality, proximity, co-location and cluster. However, the effectiveness of the GR applications varied across the two case studies which indicates that it may not be feasible to develop a general GR model that can be used as a general GIS tool.

5.1 Research Objectives Review and Discussion

The literature review showed that literature on spatial relevance analysis is spread across a number of research fields that include VGI data quality (Goodchild & Li, 2012), VGI enrichment (Spinsanti & Ostermann, 2013), fitness for use (Jonietz et al., 2016), open data retrieval (Corti et al., 2018), geographic information retrieval (Purves et al., 2018) and geographic relevance (Reichenbacher et al., 2016). The literature review guided the identification and implementation of the four GR metrics used in this thesis.

The literature also provides guidance on expected results of the two case studies and an indication of how to interpret results such as the seemingly ineffective topicality metric in Chapter 3 and 4 or the seemingly redundant co-location and cluster co-location metrics. In Chapter 3 and 4 it was noted that the topicality metric appeared to be an ineffective indicator of relevance and there may be cause to remove the topicality metric from GR criteria. However, there is an overwhelming volume of literature that indicates that topicality is an essential measure of relevance between spatial data (Bordogna et al., 2012; McKenzie et al., 2014; Reichenbacher et al., 2016; Spinsanti & Ostermann, 2013). The topicality metric is generally difficult to implement due to the complexity of natural language processing. In this case study, the WordNet library did not fit the use case. Future work should consider the use of another ontology or explore the development of a spatial ontology to better address this dimension of the GR problem. The literature also provides models to aggregate GR scores using weighted models that

account for redundancy in metrics such as cluster and co-location (Reichenbacher et al., 2016). These models where not adapted in this study, but future research should consider adapting these models.

The literature review helped direct the second and third objectives of identifying a common set of GR criteria that can be used to evaluate relevance in two significantly different contexts. The second objective was explored in Chapter 3 looked at relevance analysis between large data sets. The third objective explored in Chapter 4 looked at relevance analysis between structured and unstructured spatial features. The overall objective was to evaluate if a common set of GR criteria can be used to solve increasingly complex GIS analysis problems. However, the goal of using a common set of conceptual GR metrics was not fully achieved as the cluster metric was significantly adjusted between Chapter 3 and Chapter 4 and the proximity metric was adjusted slightly to fit. In Chapter 3, the cluster co-location metric essentially measures similar distribution of features between two layers. In Chapter 4, cluster metric evaluates the importance of individual features based upon their location within a cluster. In Chapter 3, proximity is evaluated using layer centroids and in Chapter 4 the nearest points between features are used. The formula used to generate the proximity score was the same for Chapter 3 and Chapter 4 but the formulas for cluster co-location and cluster were different. The topicality, spatial proximity and co-location metrics worked well as general GR criteria but the cluster metric may not be applicable to all use cases.

The fourth major objective of this thesis was to evaluate the effectiveness of the conceptual GR criteria using sample applications and case studies. This objective was addressed in Chapter 3 and 4 with the UrbanContext and UrbanData application. The applications demonstrated how to implement GR metrics in a software application, and it showed how the GR model can be used to organize and filter spatial data. The UrbanData and UrbanContext case studies provide insight on the effectiveness of GR metrics in different scenarios and it brings to light some of the challenges with implementing and using GR metrics for relevance analysis. The development of UrbanData and UrbanContext demonstrated that data management is a significant issue in GR systems as relevance analysis is computationally expensive, implementing GR at scale will require innovation around improving the efficiency of GR functions. The results of the case study indicate that parameter generation and score aggregation are key problems that are not addressed well in this thesis or the GR literature. The UrbanData case study indicated that most GR metrics were good indicators of relevance between layers while the UrbanContext analysis indicated the GR metrics where not reliable indicators of relevance between framework features and VGI data points. However, the complexity of the GR analysis model in UrbanContext makes it difficult to

determine the impacts of the various GR metrics. Future work should consider the development of models that make it possible to isolate variables and use regression models to evaluate the performance of individual GR metrics in relation to other GR metrics.

The development of the UrbanData and UrbanContext application also raises questions about societal impact of GR applications. Due to the complexity of GR models and its dependence on input data and input parameter values, it is very possible to manipulate GR models to prioritize corporate interests over the public good with little oversight. Similar concerns have been expressed concerning the use of machine learning and artificial intelligence in the public realm (Batty, 2018; Kwan, 2016). A possible approach to ensure democratic and equitable use of this type of technology is through open source technologies and open standards. The GIS community has a strong history of maintaining open source projects and standards as exemplified by projects such as QGIS, OpenLayers and PostGIS. Creating low level Python or PL/SQL packages to enable GR analysis between spatial data sets or spatial features would ensure that GR applications such as UrbanData and UrbanContext are broadly available and it makes it possible for an open community to oversee the continued development of GR tools. The UrbanData application is primarily a tool to evaluate relevance between spatial data and it is envisioned that it would primarily be coupled with search systems in open data portals. UrbanData may be prone to generating filter bubbles in search portals if the GR relevance modules are coupled with profile data about the user which may or may not be available. UrbanContext and UrbanContext Viz are purpose-built tool that are focused on VGI feedback analysis. It is possible to skew UrbanContext analysis results by selectively adding framework data to the analysis system. The only way to ensure integrity of results is to publish the UrbanContext framework data sets with any output reports to ensure that report readers understand how summary figures where generated. All the applications are still unknown, and it is difficult to make recommendations on all the potential applications of this technology. It is primarily envisioned as a research tool and it is likely that the open source community is the best steward of this technology.

5.2 Contributions

This thesis has demonstrated new applications and implementations of GR provide direction for future research and highlight some of the current gaps in GR research. The results of this thesis also provide guidance on the characteristics of GR metrics and their behaviour in different contexts. The evaluation of the topicality, proximity, co-location and cluster GR metrics over two case studies demonstrate that proximity is an essential measure of spatial relevance while the other metrics of

relevance vary in importance across different study contexts. This thesis demonstrates that topicality, proximity and co-location can be considered to be general metrics of GR that can be used to evaluate relevance across multiple contexts; this finding aligns with the ideas proposed by Reichenbacher et al (2016) regarding criteria of GR. However, the case studies also indicated that the cluster metric is not a GR metric that is effective at multiple scales and future work in GR should evaluate if the cluster metric should be used as a GR criterion. Study results also support grouping GR metrics during score aggregation as suggested by Reichenbacher et al. (2016) because some GR metrics may be double counted or overweighed.

This thesis has demonstrated an approach to analyze unstructured data and handle many to one relationships. Analyzing heterogeneous unstructured VGI data contributed by a large number of users with unique perspectives, lexicons and taxonomies is a very complex task. There is large variance in the accuracy of location data and location references within unstructured comments can be very complex. The use of query footprints with GR metrics appears to be a promising solution for analyzing unstructured VGI data that is collected using social media or map-based surveys. Q-footprints can not solve the inherent ambiguity and uncertainty associated with natural language analysis, but it provides a model to filter and organize data in order to enable GR analysis between structured and unstructured features. As noted in Chapter 4, the GR metrics and aggregation model developed for UrbanContext needs to be improved but the basic model is a blueprint that can be used for future analysis.

From an application perspective, this thesis demonstrates that GR has applications outside of mobile search. The UrbanData case study demonstrated that GR metrics can be used to develop spatial data retrieval systems that are not as reliant on metadata as current solutions (Florance et al., 2015; Ivanova et al., 2013) and evaluate spatial data characteristics when determining relevance. UrbanContext improves on existing models that prioritize semantic relevance and spatial proximity (McKenzie et al., 2014; Spinsanti & Ostermann, 2013), to incorporate relevance metrics that evaluate the geographic environment of features.

The concept of geographic relevance has been generally defined in the literature (Raper, 2007; Reichenbacher et al., 2016), but criteria of geographic relevance has been defined using context specific criteria (Reichenbacher et al., 2016). This thesis emphasizes the need to better define the concept of GR, the criteria of GR and the application of GR within the broader context of GIS. It is unclear whether GR is intended to be a generally applicable to GIS, much like Tobler's law, or if GR is a concept that is intended to solve a specific problem in geographic information retrieval.

5.3 Limitations

This thesis has presented two studies on GR with limited scope due to the finite amount of time to conduct the study and limited computing resources to execute the analysis. These limitations where necessary due to the scope of the study but they are issues that should be addressed in future research.

The first limitation is the use of external modules to evaluate semantic relevance. The topicality metric did not produce expected results in both the UrbanData and the UrbanContext analysis. The use of the Python NLTK WordNet module made it difficult to identify the causes of inconsistencies of scores due to the limited understanding of the WordNet database and the Python NLTK module. Both the NLTK module and the WordNet database are open source tools and future researchers may have an opportunity to manipulate these systems to attain better results.

The second study limitation is the lack of parameter testing, spatial GR metrics such as proximity, co-location, and cluster rely on assumptions around threshold distance or cluster size. All the parameter values used in this thesis rely on the author's limited understanding of the study context, this approach to parameter generation is unreliable and prone to errors. As noted in Chapter 3 and 4, a systematic approach is needed to determine key parameters such as threshold distance or cluster size based upon characteristics of input data sets and the study site.

The third limitation is the manual review of results by a single reviewer, both studies conducted in this thesis relied on manual review of the data using limited spot checks of results. This is a reasonable approach for reviewing large data sets, but the samples selected for review may not be representative of all analysis results. Furthermore, manual review of results was conducted by the author using a limited understanding of the study context. Therefore, conclusions made from the review of data are biased by the authors' perception and understanding of the study context. A more comprehensive review of results would also provide more insight into the usefulness of the UrbanData and UrbanContext applications. This research was conducted on the premise that organizing and retrieving spatial data is currently a challenge for cities and citizens but there was minimal formal feedback collected on the need for the UrbanContext and UrbanData applications.

Lastly, the scope of the studies in Chapter 3 and 4 were very limited in size due to the limited computing resources available for this thesis. The UrbanData and UrbanContext studies processed several thousand individual features using computationally expensive spatial calculations. The

computational cost of the UrbanData and UrbanContext analysis resulted in studies with limited size and scope.

5.4 Directions for Future Research

Though this thesis was limited in scope, this research has demonstrated novel ways to use GR metrics to address challenges associated with growing volumes of web-based spatial data. Future research in GR should look at parameter testing and sensitivity analysis of GR metrics and aggregation models. Future work can improve the UrbanData and UrbanContext models by implementing feature level co-location and topicality metrics, improving the topicality implementation and gathering feedback on results. Parameter testing and sensitivity analysis for key input values such as threshold distance for proximity, search distance for co-location and cluster size for the cluster metrics are needed to optimize GR models.

Measuring geographic relevance is a complex multi-criteria problem that is not easily delimited because relevance is an ambiguous concept. As a result, parameter values and GR metric weights are subjectively defined. These values are the greatest source of uncertainty in GR models due to their impact on the final GR score. As discussed in multi-criteria analysis literature, sensitivity analysis of GR metrics can help reduce uncertainty in GR models and improve the reliability of outputs by demonstrating the impact of small changes in parameter values and GR weights on outcomes (Robert Feick & Hall, 2004). Sensitivity analysis of GR models is challenging because parameter values should need to account for the study scale and spatial data characteristics. For example, the distance threshold value for generating proximity scores should be higher when comparing spatial data sets and lower when comparing spatial features. Parameter values can be evaluated by manually testing multiple parameter values using applications such as UrbanData and UrbanContext. However, the preferred approach to parameter testing would be a systematic approach to parameter generation. A systematic model for parameter generation would evaluate characteristics of spatial data sets and the study context. Parameters can be generated using criteria such as the average distance between features, the study extent and feature density. Future research should acknowledge that changes in parameter values will have varying impacts on different types of analysis. For example, the UrbanData proximity metric is likely not very sensitive to variances in the proximity threshold distance because layers tended to occupy city extents which means layers tended to be very close or very far. Thus, minor changes in the threshold distance should not alter results. In contrast, the UrbanContext GR model is very sensitive to changes in the threshold distance for spatial proximity because the distance between VGI features and

relevant features can legitimately vary between 1m, 5m, 50m, 500m and 2km. Parameter tests should use a benchmark data set to measure how well changes in parameter values improve GR models such as UrbanData and UrbanContext.

Future evaluation of GR metrics and parameter testing of GR metrics should also consider the use of regression models to evaluate correlation between GR metrics to identify redundancy in GR metrics in order to inform the development of GR aggregation models. Regression models should compare scores of one each GR metric against scores of other GR metrics to determine whether there is overlap and redundancy between metrics. The regression analysis should inform the GR aggregation model and weighting scheme. The GR model developed by Reichenbacher et al. (2016) provides a good conceptual example for grouping GR metric types and generating aggregate GR scores. In this model, GR metrics are categorized, and metrics are aggregated using equal weights into a categorical GR score which is then used to generate an equal weighted GR score. A comparable model should be explored in future GR research. The performance of GR aggregation models can also be evaluated using backwards step wise regression models to evaluate the performance of individual GR metrics. It would also be interesting to explore GR models that dynamically add and remove GR metrics based upon the effectiveness of the metric.

A major challenge of this thesis was developing the GR metrics and evaluation model for Chapter 4 due to the complexity of implementing GR metrics between individual features. The topicality and co-location metric were not good measures of relevance between features because they were implemented between layers. The co-location and topicality metrics can likely be implemented between feature categories. This adjustment in the metric may improve complexity of analysis but it should contribute to more relevant metrics.

The Python module and WordNet database used to implement the topicality metric did not measure semantic similarity accurately. The implementation of this metric can potentially be improved by leveraging other natural language processing modules such as Amazon Comprehend or customizing the Python NLTK module to generate relevance scores that are representative of semantic relevance between spatial concepts.

Lastly, it is important that GR researchers continue to expand the scope of studies to handle the growing number of complex GIS problems emerging on the web and to gather feedback from potential users of these solutions. Prototype applications should be developed to allow users to review the

validity of results from applications such as UrbanContext and to provide feedback on the usefulness of the developed solution. GR research should be guided by user feedback in order to validate results and to ensure that efforts are directed at solving societal needs while acknowledging the ethical challenges associated with any new technology.

References

Abdollahi, A., & Riyahi Bakhtiari, H. R. (2017). Roads Data Conflation using Update High Resolution Satellite Images. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (Vol. 4, pp. 3–7). https://doi.org/10.5194/isprs-annals-IV-4-W4-3-2017

Abelson, J., Forest, P.-G., Eyles, J., Smith, P., Martin, E., & Gauvin, F.-P. (2003). *Deliberations about deliberative methods: issues in the design and evaluation of public participation processes*. *Social Science & Medicine* (Vol. 57).

Acheson, E., Wartmann, F. M., & Purves, R. S. (2018). Generating spatial footprints from hiking blogs. In *Lecture Notes in Geoinformation and Cartography* (pp. 5–7). Springer, Cham.

Adams, B., Li, L., Raubal, M., & Goodchild, M. (2007). A General Framework for Conflation. In *Geographic Information Science* (pp. 1–5). Santa Barbara.

Ali, A., & Schmid, F. (2014). Data Quality Assurance for Volunteered Geographic Information. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8728*, 126–141.

Ballatore, A., & Bertolotto, M. (2011). Semantically enriching VGI in support of implicit feedback analysis. *Web and Wireless Geographical Information Systems*, 78–93.

Ballatore, A., & Bertolotto, M. (2018). Semantically Enriching VGI in Support of Implicit Feedback Analysis. *International Symposium on Web and Wireless Geographical Information Systems*, 78–93.

Ballatore, A., Bertolotto, M., & Wilson, D. (2013). Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems*, *37*(1), 61–81.

Barua, S., & Sander, J. (2014). Mining Statistically Significant Co-location and Segregation Patterns. *IEEE Transactions on Knowledge and Data Engineeering*, *26*(5), 1185.

Basiouka, S., & Potsiou, C. (2012). VGI in Cadastre: a Greek experiment to investigate the potential of crowd sourcing techniques in Cadastral Mapping. *Survey Review*, *44*(325), 153–161.

Batty, M. (2018). Artificial intelligence and smart cities. *Environment and Planning B: Urban Analytics and City Science*, *45*(1), 3–6. https://doi.org/10.1177/2399808317751169

Batty, M., Hudson-Smith, A., Milton, R., & Crooks, A. (2010). Map mashups, Web 2.0 and the GIS revolution. *Annals of GIS*, *16*(1), 1–13. https://doi.org/10.1080/19475681003700831

Beaudreau, P., Johnson, P. A., & Sieber, R. E. (2012). Strategic Choices in Developing a Geospatial Web 2.0 Application for Rural Economic Development. *Journal of Rural and Community Development*, *7*(3), 95–105.

Bordogna, G., Ghisalberti, G., & Psaila, G. (2012). Geographic information retrieval: Modeling uncertainty of user's context. *Fuzzy Sets and Systems*, *196*, 105–124.

Bordogna, G., & Psaila, G. (2008). Modeling Soft Conditions with Unequal Importance in Fuzzy Databases based on the Vector p-norm.

Brown, G. (2012). Public participation GIS (PPGIS) for regional and environmental planning: Reflections on a decade of empirical research. *URISA Journal*, *25*(2), 7–18.

Brown, G., & Kyttä, M. (2014). Key issues and research priorities for public participation GIS (PPGIS): A synthesis based on empirical research. *Applied Geography*.

Brown, G., & Raymond, C. M. (2014). Methods for identifying land use conflict potential using participatory mapping. *Landscape and Urban Planning*, *122*, 196–208.

Burns, R., & Thatcher, J. (2015). Guest Editorial: What's so big about Big Data? Finding the spaces and perils of Big Data. *GeoJournal*, *80*(4), 445–448.

Calderon, A., Carfi, N., & De Luca, A. (2015). *International Open Data Charter*.

Camponovo, M. E., & Freundschuh, S. M. (2014). Assessing uncertainty in VGI for emergency response. *Cartography and Geographic Information Science*, *41*(5), 440–455.

Centre for International Governance Innovation. (2018). *A National Data Strategy for Canada Key Elements and Policy Considerations*.

Charalabidis, Y., Gionis, G., Ferro, E., & Loukis, E. (2010). Towards a systematic exploitation of web 2.0 and simulation modeling tools in public policy process. In *Lecture Notes in Computer Science* (Vol. 6229 LNCS, pp. 1–12). Springer, Berlin, Heidelberg.

Chen, H., Zhang, W. C., Deng, C., Nie, N., & Yi, L. (2017). Volunteered Geographic Information for Disaster Management with Application to Earthquake Disaster Databank. *IOP Conference Series: Earth and Environmental Science*, *57*(1), 012015.

Cinderby, S. (2010). How to reach the "hard-to-reach": The development of Participatory Geographic Information Systems (P-GIS) for inclusive urban design in UK cities. *Area*, *42*(2), 239–251.

Cinnamon, J., & Schuurman, N. (2013). Confronting the data-divide in a time of spatial turns and volunteered geographic information. *GeoJournal*, *78*(4), 657–674.

Cisco. (2018). Cisco Global Cloud Index : Forecast and Methodology , 2016-2021. *White Paper*, 1–46.

City of Kitchener. (2015). *City of Kitchener Iron Horse Trail Improvement Strategy*. Kitchener.

City of Toronto. (2018). *Toronto Open Data Master Plan 2018-2022*. Toronto. Retrieved from https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-master-plan/

Clough, P. D., Joho, H., & Purves, R. (2006). Judging the spatial relevance of documents for GIR. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3936 LNCS, pp. 548–552).

Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T., & Rau, R. (2011). OSMonto - An Ontology of OpenStreetMap Tags. *State of the Map Europe (SOTM-EU)*, (November).

Comber, A., Fonte, C., Foody, G., Fritz, S., Harris, P., Olteanu-Raimond, A. M., & See, L. (2016). Geographically weighted evidence combination approaches for combining discordant and inconsistent volunteered geographical information. *GeoInformatica*, *20*(3), 503–527.

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

Corti, P., Lewis, B., & Kralidis, A. T. (2018). Hypermap registry: an open source, standards-based geospatial registry and search platform. *Open Geospatial Data, Software and Standards*, *3*(1), 8.

https://doi.org/10.1186/s40965-018-0051-x

De Sabbata, S. (2013). *Assessing geographic relevance for mobile information services*. University of Zurich.

De Sabbata, S., & Reichenbacher, T. (2012). Criteria of geographic relevance: An experimental study. *International Journal of Geographical Information Science*, *26*(8), 1495–1520.

Deng, M., He, Z., Liu, Q., Cai, J., & Tang, J. (2017). Multi-scale approach to mining significant spatial co-location patterns. *Transactions in GIS*, *21*, 1023–1039.

Derungs, C., Wartmann, F., Purves, R., & Mark, D. (2013). The meanings of the generic parts of toponyms: Use and limitations of gazetteers in studies of landscape terms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8116 LNCS*, 261–278.

Devillers, R., Bédard, Y., & Jeansoulin, R. (2013). Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use Within GIS. *Photogrammetric Engineering & Remote Sensing*, *71*(2), 205–215.

Devillers, R., & Jeansoulin, R. (2006). *Fundamentals of Spatial Data Quality*. (R. Devillers & R. Jeansoulin, Eds.). ISTE.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., & Shi, W. (2010). Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Transactions in GIS*, *14*(4), 387–400.

DigitalGlobe. (2018). OpenStreetMap | DigitalGlobe Blog. Retrieved May 18, 2019, from http://blog.digitalglobe.com/tag/openstreetmap/

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, *54*(4), 86. https://doi.org/10.1145/1924421.1924442

Dominich, S. (2008). *The Modern Algebra of Information Retrieval*. *The Modern Algebra of Information Retrieval* (Vol. 24). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77659-8

Doytsher, Y., Filin, S., & Ezra, E. (2001). Transformation of Datasets in a Linear-based Map Conflation Framework. *Surveying and Land Information Systems*, *61*(3), 159–169.

Dujmović, J. J. (2007). Continuous preference logic for system evaluation. *IEEE Transactions on Fuzzy Systems*, *15*(6), 1082–1099. https://doi.org/10.1109/TFUZZ.2007.902041

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

Esmaeili, R., & Karimipour, F. (2015). Cartographic representation of spatial data quality in VGI for users with different semantics. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* (Vol. 40, pp. 81–84).

ESRI Inc. (2019). ArcGIS Hub - Open Data. Retrieved February 14, 2019, from http://hub.arcgis.com/pages/open-data

Evans-Cowley, J. S., & Griffin, G. (2012). Microparticipation with Social Media for Community Engagement in Transportation Planning. *Transportation Research Record: Journal of the Transportation Research Board*, *2307*(1), 90–98.

Feick, Rob, & Roche, S. (2013). Understanding the value of VGI. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (Vol. 9789400745, pp. 15–29). Dordrecht: Springer Netherlands.

Feick, Robert, & Hall, B. G. (2004). A method for examining the spatial dimension of multi-criteria weight sensitivity. *International Journal of Geographical Information Science*, *18*(8), 815–840.

Florance, P., McGee, M., Barnett, C., & McDonald, S. (2015). The Open Geoportal Federation. *Journal of Map and Geography Libraries*, *11*(3), 376–394.

Fogliaroni, P., D'Antonio, F., & Clementini, E. (2018). Data trustworthiness and user reputation as indicators of VGI quality. *Geo-Spatial Information Science*, *21*(3), 213–233.

Fu, G., Jones, C. B., & Abdelmoty, A. I. (2005a). Building a Geographical Ontology for Intelligent Spatial Search on the Web. *Names*, *pages*(June 2014), 167–172.

Fu, G., Jones, C. B., & Abdelmoty, A. I. (2005b). Ontology-based spatial query expansion in information retrieval. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3761 LNCS*, 1466–1482.

Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *CEUS*, *61*, 172–186.

Gao, Y., Jiang, D., Zhong, X., & Yu, J. (2016). A Point-Set-Based Footprint Model and Spatial Ranking Method for Geographic Information Retrieval. *ISPRS International Journal of Geo-Information*, *5*(7), 122.

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

Goodchild, M. (2007a). Citizens as sensors: The world of volunteered geography. *GeoJournal*.

Goodchild, M. (2007b). Editorial: Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, *2*, 24–32.

Goodchild, M. (2009). NeoGeography and the nature of geographic expertise. *Journal of Location Based Services*, *3*(2), 82–96. https://doi.org/10.1080/17489720902950374

Goodchild, M., & Hill, L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, *22*(10), 1039–1044.

Goodchild, M., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120.

Government of Canada. (2014). *Canada's Action Plan on Open Government 2014–16*. Ottawa.

Government of Ontario. Planning Act, R.S.O. 1990 (2019). Toronto: Government of Ontario.

Graham, M., & De Sabbata, S. (2015). Mapping information wealth and poverty: the geography of gazetteers. *Environment and Planning A*, *47*(6), 1254–1264.

Grira, J., Bédard, Y., & Roche, S. (2010). Spatial data uncertainty in the VGI world: Going from consumer to producer. *Geomatica*, *64*(1), 61–72.

Gruen, N., Houghton, J., & Tooth, R. (2014). *Open for Business : How Open Data Can Help Achieve the G20 Growth Target*. Victoria.

Guidoin, S., Marczak, P., Pane, J., & McKinney, J. (2014). *Identifying recommended standards and best practices for open data*. Montreal.

Hacar, M., & Gökgöz, T. (2019). A New, Score-Based Multi-Stage Matching Approach for Road Network Conflation in Different Road Patterns. *ISPRS International Journal of Geo-Information*, *8*(2), 81.

Hahmann, S., Purves, R., & Burghardt, D. (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, *9*(9), 1–36.

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. (2013). Neogeography and the delusion of democratisation. *Environment and Planning A*.

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

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

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology*, *61*(2), 217–237. https://doi.org/10.1002/asi.21261

Hodge, G., & Gordon, D. L. A. (2008). *Planning Canadian communities : an introduction to the principles, practice, and participants*. Thomson Nelson.

Huang, Y., Shekhar, S., & Xiong, H. (2004). Discovering colocation patterns from spatial data sets: A general approach. *IEEE Transactions on Knowledge and Data Engineering*, *16*(12), 1472–1485.

IBM Marketing Cloud. (2017). *10 Key Marketing Trends for 2017*.

Innes, J. E., & Booher, D. E. (2004, December). Reframing public participation: Strategies for the 21st century. *Planning Theory and Practice*. https://doi.org/10.1080/1464935042000293170

International Association of Public Participation. (2019). International Association for Public Participation. Retrieved May 19, 2019, from https://www.iap2.org/

Ivanova, I., Morales, J., de By, R., Beshe, T., & Gebresilassie, M. (2013). Searching for spatial data resources by fitness for use. *Journal of Spatial Science*, *58*(1), 15–28.

Jacobs, J. (1992). *The Death and Life of Great American Cities*. Retrieved from http://www.amazon.com/dp/0679644334

Jankowski, P., Czepkiewicz, M., Młodkowski, M., Zwoliński, Z., & Wójcicki, M. (2019). Evaluating the scalability of public participation in urban land use planning: A comparison of Geoweb methods with face-to-face meetings. *Environment and Planning B: Urban Analytics and City Science*, *46*(3),

511–533.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268.

Johnson, P. A., & Sieber, R. E. (2012). Motivations driving government adoption of the Geoweb. *GeoJournal*, *77*(5), 667–680.

Jones, C., & Purves, R. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, *22*(3), 219–228.

Jonietz, D., Antonio, V., See, L., & Zipf, A. (2017). Highlighting Current Trends in Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, *6*(7), 202.

Jonietz, D., Zipf, A., Jonietz, D., & Zipf, A. (2016). Defining Fitness-for-Use for Crowdsourced Points of Interest (POI). *ISPRS International Journal of Geo-Information*, *5*(9), 149.

Josh Joseph. (2015). *Staff Report: Iron Horse Trail Improvement Strategy*. Kitchener.

Kalvelage, K., Dorneich, M. C., Seeger, C. J., Welk, G. J., Gilbert, S., Moon, J., … Kalvelage Michael Dorneich Christopher J Seeger Gregory J Welk Stephen Gilbert Jon Moon Imad Jafir Phyllis Brown, K. C. (2018). Assessing the validity of facilitated-volunteered geographic information: comparisons of expert and novice ratings. *GeoJournal*, *83*(3), 477–488.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881–892. https://doi.org/10.1109/TPAMI.2002.1017616

Kleinhans, R., Ham, M. Van, & Evans-Cowley, J. (2015). Using Social Media and Mobile Technologies to Foster Engagement and Self-Organization in Participatory Urban Planning and Neighbourhood Governance. *Planning, Practice & Research*, *30*(3), 237–247.

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.

Krishnan, A., Deepak, P., Ranu, S., & Mehta, S. (2018). Leveraging semantic resources in diversified query expansion. *World Wide Web*, *21*(4), 1041–1067. https://doi.org/10.1007/s11280-017-0468-7

Kunz, R. (2009). *Evaluation of spatial relevance in geographic information retrieval*. *Department of Geography*. University of Zurich.

Kwan, M.-P. (2016). Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge. *Annals of the American Association of Geographers*, *106*(2), 274–282. Retrieved from http://www.tandfonline.com/action/journalInformation?journalCode=raag21

Laurini, R. (2014). A conceptual framework for geographic knowledge engineering. *Journal of Visual Languages & Computing*, *25*(1), 2–19.

Leibovici, D. G., Evans, B., Hodges, C., Wiemann, S., Meek, S., Rosser, J., & Jackson, M. (2015). On data quality assurance and conflation entanglement in crowdsourcing for environmental studies. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (Vol. 2, pp. 195–202). Multidisciplinary Digital Publishing Institute.

Li, D., Zhang, J., & Wu, H. (2012). Spatial data quality and beyond. *International Journal of Geographical*

Information Science, *26*(12), 2277–2290.

Li, M., Sun, Y., & Fan, H. (2015). Contextualized Relevance Evaluation of Geographic Information for Mobile Users in Location-Based Social Networks. *ISPRS International Journal of Geo-Information*, *4*(2), 799–814.

Li, N., Raskin, R., Goodchild, M., & Janowicz, K. (2012). An Ontology-Driven Framework and Web Portal for Spatial Decision Support. *Transactions in GIS*, *16*(3), 313–329.

Li, W., Goodchild, M., & Raskin, R. (2014). Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth*, *7*(1), 17–37.

Li, W., Yanga, C., & Yang, C. (2010). An active crawler for discovering geospatial Web services and their distribution pattern - A case study of OGC Web Map Service. *International Journal of Geographical Information Science*, *24*(8), 1127–1147. https://doi.org/10.1080/13658810903514172

Lin, W. (2018). Volunteered Geographic Information constructions in a contested terrain: A case of OpenStreetMap in China. *Geoforum*, *89*, 73–82.

Liu, H., Bao, H., & Xu, D. (2012). Concept vector for semantic similarity and relatedness based on WordNet structure. In *Journal of Systems and Software* (Vol. 85, pp. 370–381).

Machado, I. M. R., de Alencar, R. O., Campos, R. de O., & Davis, C. A. (2011). An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, *17*(4), 267–279.

MapBox. (2019). Mapbox Streets | Mapbox. Retrieved May 18, 2019, from https://www.mapbox.com/about/maps/

Marzouki, A., Mellouli, S., & Daniel, S. (2018). Spatial, temporal and semantic contextualization of citizen participation. *Proceedings of the 19th Annual International Conference on Digital Government Research Governance in the Data Age - Dgo '18*, 1–8.

Mata-Rivera, F., Torres-Ruiz, M., Guzmán, G., Moreno-Ibarra, M., & Quintero, R. (2015). A collaborative learning approach for geographic information retrieval based on social networks. *Computers in Human Behavior*, *51*, 829–842.

McKenzie, G., Janowicz, K., & Adams, B. (2014). A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science*, *41*(2), 125–137.

Merrouni, Z. A., Frikh, B., & Ouhbi, B. (2019). Toward Contextual Information Retrieval: A Review and Trends. In *Procedia Computer Science* (Vol. 148, pp. 191–200).

Minano, A., Johnson, P. A., & Wandel, J. (2018). Visualizing flood risk, enabling participation and supporting climate change adaptation using the Geoweb: the case of coastal communities in Nova Scotia, Canada. *GeoJournal*, *83*(3), 413–425. https://doi.org/10.1007/s10708-017-9777-8

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

Neuhaus, F. (2018). What is an Ontology?

Newfoundland Labrador Office of Public Engagement. (2013). *Public Engagement Guide*. Retrieved from https://ope.gov.nl.ca/publications/pdf/OPE_PEGuide.pdf

Nguyen, H., Richards, R., Chan, C. C., & Liszka, K. J. (2016). RedTweet: recommendation engine for reddit. *Journal of Intelligent Information Systems*, *47*(2), 247–265.

Noskov, A., & Zipf, A. (2019). Open-data-driven embeddable quality management services for map-based web applications. *Big Earth Data*, *2*(4), 395–422. https://doi.org/10.1080/20964471.2019.1592077

OpenStreetMap. (2018). CanVec - OpenStreetMap Wiki. Retrieved May 18, 2019, from https://wiki.openstreetmap.org/wiki/CanVec

Overell, S. E. (2009). *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. *Departamento de Computación*. University of London.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web, 1–17. https://doi.org/10.1.1.31.1768

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet: Similarity - Measuring the Relatedness of Concepts. *Intelligent Systems Demonstration*, 1024–1025. https://doi.org/10.1109/DICTA.2008.79

PlaceVision Inc. (2019). Community Engagement Examples - Community Remarks. Retrieved September 8, 2019, from https://communityremarks.com/projects/

Planning Institute Australia. Public Participation (2011). Planning Institute Australia.

Poser, K., & Dransch, D. (2015). Volunteered Geographic Information for Disaster Management with Application to Rapid Flood Damage Estimation Originally published as : with application to rapid flood damage estimation. *Geomatica*, *1*(January 2010), 89–98.

Purves, R., Clough, P., Jones, C., Arampatzis, A., Bucher, B., Finch, D., … Yang, B. (2007). The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, *21*(7), 717–745.

Purves, R., Clough, P., Jones, C., Hall, M., & Murdock, V. (2018). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends® in Information Retrieval*, *12*(2–3), 164–318.

Rabari, C., & Storper, M. (2015). The digital skin of cities: Urban theory and research in the age of the sensored and metered city, ubiquitous computing and big data. *Cambridge Journal of Regions, Economy and Society*, *8*(1), 27–42. https://doi.org/10.1093/cjres/rsu021

Ramos, J., Vandecasteele, A., & Devillers, R. (2014). Introduction Semantic Integration of Authoritative and Volunteered Geographic Information (VGI) using Ontologies. In *Association of Geographic Information Laboratories for Europe (AGILE) Conference.* (p. 6).

Raper, J. (2007). Geographic relevance. *Journal of Documentation*, *63*(6), 836–852.

Reichenbacher, T., De Sabbata, S., Purves, R. S., & Fabrikant, S. I. (2016). Assessing geographic relevance for mobile search: A computational model and its validation via crowdsourcing. *Journal of the Association for Information Science and Technology*, *67*(11), 2620–2634.

Roche, S., Propeck-Zimmermann, E., & Mericskay, B. (2013). GeoWeb and crisis management: Issues and perspectives of volunteered geographic information. *GeoJournal*, *78*(1), 21–40.

Ross, H., Baldwin, C., & Carter, B. (2016). Subtle implications: public participation versus community

engagement in environmental decision-making. *Australasian Journal of Environmental Management*, *23*(2), 123–129.

Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, *2*(2), 169–194.

Schlossberg, M., & Shuford, E. (2005). Delineating " Public " and " Participation " in PPGIS. *URISA Journal*, *16*(2), 15–26.

Schweitzer, L. (2014). Planning and social media: A case study of public transit and stigma on twitter. *Journal of the American Planning Association*, *80*(3), 218–238. https://doi.org/10.1080/01944363.2014.980439

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., … Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, *5*(5), 55.

Seeger, C. J. (2008). The role of facilitated volunteered geographic information in the landscape planning and site design process. *GeoJournal*, *72*(3–4), 199–213.

Seltzer, E., & Mahmoudi, D. (2012). Citizen Participation, Open Innovation, and Crowdsourcing: Challenges and Opportunities for Planning. *Journal of Planning Literature*, *28*(1), 3–18.

Shipley, R., & Utz, S. (2012). Making it Count: A Review of the Value and Techniques for Public Consultation. *Journal of Planning Literature*.

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., & Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, *30*(4), 378–399.

Specht, D. (2015). Book Review: The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences. *Media, Culture & Society*, *37*(7), 1110–1111.

Spinsanti, L., & Ostermann, F. (2013). Automated geographic context analysis for volunteered information. *Applied Geography*, *43*, 36–44.

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, *78*(2), 319–338.

Swedish Standards Institute. (2019). ISO/TC 211 - Geographic information/Geomatics. Retrieved March 14, 2019, from https://www.iso.org/committee/54904.html

The Open Government Partnership. (2016). *What's in the new OGP National Actions Plans*.

The World Wide Web Foundation. (2017). *Open Data Barometer*.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in Detroit Region. *Economic Geography*, *46*, 234–240. https://doi.org/doi:10.2307/143141

Touya, G., Coupé, A., Jollec, J., Dorie, O., & Fuchs, F. (2013). Conflation Optimized by Least Squares to Maintain Geographic Shapes. *ISPRS International Journal of Geo-Information*, *2*(3), 621–644.

Tulloch, D. L. (2008). Is VGI participation? From vernal pools to video games. *GeoJournal*.

Turner, A. (2006). *Introduction to neogeography*. O'Reilly.

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*, *XL-2/W1*, 143–148.

Vandecasteele, A., & Devillers, R. (2015). Improving volunteered geographic information quality using a tag recommender system: The case of OpenStreetMap. *Lecture Notes in Geoinformation and Cartography*, (9783319142791), 59–80.

Verplanke, J., McCall, M. K., Uberhuaga, C., Rambaldi, G., & Haklay, M. (2016). A Shared Perspective for PGIS and VGI. *Cartographic Journal*, *53*(4), 308–317.

Webler, Seth Tuler, T. (1999). Voices from the Forest: What Participants Expect of a Public Participation Process. *Society & Natural Resources*, *12*(5), 437–453. https://doi.org/10.1080/089419299279524

Wentz, E. A., & Shimizu, M. (2018). Measuring spatial data fitness-for-use through multiple criteria decision making. *Annals of the American Association of Geographers*, *108*(4), 1150–1167.

Wilkinson, D. M., & Huberman, B. A. (2007). Assessing the value of cooperation in Wikipedia. *First Monday*, *12*(4). https://doi.org/10.5210/fm.v12i4.1763

Yoo, J. S., & Shekhar, S. (2006). A joinless approach for mining spatial colocation patterns. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1323–1337. https://doi.org/10.1109/TKDE.2006.150

Yu, F., West, G., Arnold, L., McMeekin, D., & Moncrieff, S. (2016). Automatic geospatial data conflation using semantic web technologies, (February), 1–10.

Zhang, Z., Gentile, A. L., & Ciravegna, F. (2013). Recent advances in methods of lexical semantic relatedness-a survey. *Natural Language Engineering*, *19*(04), 411–479. https://doi.org/10.1017/S1351324912000125

Zhu, G., & Iglesias, C. A. (2017). Sematch: Semantic similarity framework for Knowledge Graphs. *Knowledge-Based Systems*, *130*, 30–32. https://doi.org/10.1016/j.knosys.2017.05.021

Zook, M., Graham, M., & Boulton, A. (2015). Crowd-Sourced Augmented Realities: Social Media and the Power of Digital Representation. In *Mediated Geographies and Geographies of Media* (pp. 223–240). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-9969-0_14

Appendix A: Code Samples from UrbanData and UrbanContext

Gazetteer Query

```
CREATE VIEW gazetteer AS SELECT row_number() OVER () AS id,
    a.table_schema,
    a.table_name,
    a.column_name,
    a.search_term,
    a.category
   FROM ( SELECT a_1.table_schema,
            a_1.table_name,
            a_1.column_name,
            a_1.search_term,
            a_1.category
           FROM ( SELECT pg_stat_user_tables.schemaname AS table_schema,
                    pg_stat_user_tables.relname AS table_name,
                    NULL::text AS column_name,
                    pg_stat_user_tables.relname AS search_term,
                    'Layer'::text AS category
                   FROM pg_stat_user_tables
                  WHERE (pg_stat_user_tables.schemaname = 'kitchener'::name)) a_1
        UNION
         SELECT 'kitchener'::name AS table_schema,
            'roads'::name AS table_name,
            'street_nam'::text AS column_name,
            roads.street_nam AS search_term,
            'Attribute'::text AS category
           FROM kitchener.roads
          GROUP BY roads.street_nam
        UNION
         SELECT 'kitchener'::name AS table_schema,
            'trails'::name AS table_name,
            'route_name'::text AS column_name,
            trails.route_name AS search_term,
            'Attribute'::text AS category
           FROM kitchener.trails
          WHERE (trails.route_name IS NOT NULL)
          GROUP BY trails.route_name
        UNION
         SELECT 'kitchener'::name AS table_schema,
            'parks'::name AS table_name,
            'park'::text AS column_name,
            parks.park AS search_term,
            'Feature'::text AS category
           FROM kitchener.parks
          WHERE (parks.park IS NOT NULL)) a
  ORDER BY a.category;
```

Q-Footprint Query

```
CREATE MATERIALIZED VIEW iht_survey_q AS SELECT a.gid,
    a.dsc,
    a.geom,
    b.id,
    b.table_schema,
    b.table_name,
    b.column_name,
    b.search_term,
    b.category
   FROM ( SELECT iht_survey.gid,
            iht_survey.dsc,
            iht_survey.geom
           FROM kitchener.iht_survey) a,
    ( SELECT static_gazetteer.gid AS id,
            static_gazetteer.table_schema,
            static_gazetteer.table_name,
            static_gazetteer.column_name,
            static_gazetteer.search_term,
            static_gazetteer.category
           FROM kitchener.static_gazetteer) b
  WHERE (to_tsvector((a.dsc)::text) @@ plainto_tsquery(replace((b.search_term)::text,
' '::text, '|'::text)));
```

Topicality Score Script

```
from nltk.corpus import wordnet as wn
import csv



def compare(tag1,tag2):
    try:
        word1 = wn.synsets(tag1)[0]
        word2 = wn.synsets(tag2)[0]

        return(word1.path_similarity(word2))
    except:
        return 0

outputCSVString = ""
outputFile = "urban_data_topicality.csv"

for dlayer in domain_layers:

    dtags = layers[dlayer]

    for layerName in layers:
```

```
        lyrTags = layers[layerName]
        top_score = compare(dlayer,layerName)

        for domTag in dtags:
            for lyrTag in lyrTags:
                new_score = compare(domTag,lyrTag)
                if new_score > top_score:
                    top_score=new_score


        print(dlayer,layerName,top_score)

        row = [dlayer,layerName,top_score]

        with open(outputFile, 'a') as csvFile:
            writer = csv.writer(csvFile)
            writer.writerow(row)

        csvFile.close()
```

Cluster Score Query

```
CREATE VIEW cluster50 AS SELECT row_number() OVER () AS gid,
    layerA.cluster,
    count(layerA.gid) AS size,
    st_convexhull(st_collect(st_buffer(layerA.geom, (1)::double precision))) AS geom
   FROM ( SELECT trail_survey.gid,
            st_clusterdbscan(trail_survey.geom, (50)::double precision, 3) OVER () AS
cluster,
            trail_survey.dsc,
            trail_survey.geom
           FROM kitchener.trail_survey) layerA
WHERE (layerA.cluster IS NOT NULL)
GROUP BY layerA.cluster;
```

```
SELECT max(b.density_metric) AS cluster_density_sc
FROM kitchener.cluster50 b
WHERE st_intersects(st_transform(a.geom, 26917), st_transform(b.geom, 26917))
LIMIT 1
```

Co-location Score Function

```
DECLARE
    colocation_score FLOAT;
  BEGIN

    EXECUTE format('SELECT
                  count(distinct(a.gid))::float/(SELECT count(*) FROM %I.%I)::float
          FROM
            %I.%I a
            JOIN %I.%I b
            ON

st_dwithin(st_transform(st_makevalid(a.geom),26917),st_transform(st_makevalid(b.geom),
26917),%L)',

input_schema,input_table,input_schema,input_table,target_schema,target_table,buffer_di
stance) INTO colocation_score;

    RETURN colocation_score;
END;
```

Cluster Co-location Score Function

```
DECLARE
    colocation_score FLOAT;
  BEGIN

    EXECUTE format('
                   SELECT count(*)::float/
                    (SELECT count(*)
                     FROM
                      (SELECT

distinct(st_clusterdbscan(st_transform(st_makevalid(geom),26917), (50)::double
precision, 3) OVER ()) AS cluster
                       FROM %I.%I) g
                     )::float as clusters_total
                  FROM
                    (SELECT gid,cluster,
                      (SELECT count(*)>0 FROM %I.%I k WHERE
st_dwithin(st_transform(st_makevalid(k.geom),26917),cluster.geom,%L)) colocated
                     FROM
                       (SELECT row_number() OVER () AS gid,
                           ham.cluster,
                           count(ham.gid) AS size,
                           st_convexhull(st_collect(st_buffer(ham.geom, (1)::double
precision))) AS geom
                        FROM ( SELECT gid,

st_clusterdbscan(st_transform(st_makevalid(geom),26917), (50)::double precision, 3)
OVER () AS cluster,
                 st_transform(st_makevalid(geom),26917) as geom
                     FROM %I.%I) ham
                      WHERE (ham.cluster IS NOT NULL)
                      GROUP BY ham.cluster) cluster) h
                  WHERE
                    colocated = true
                   ',
                    input_schema,input_table,
                    target_schema,target_table,
                    buffer_distance,
                    input_schema,input_table
                   ) INTO colocation_score;
    RETURN colocation_score;
END;
```

Appendix B: Output Tables for UrbanContext

Table B-0.1 UrbanContext Result table for figures in Chapter 4

GID	Match ID	Table Name	Column Name	Search Term	Proximity	Cluster	Co-location	Topicality	GR Score
8	1951	roads		roads	0.50	0.52	0.44	0.13	0.35
19	1941	railway_lines		railway_lines	0.50	0.14	0.09	0.06	0.22
19	1302	roads	street_nam	NORTH	0.14	0.52	0.44	0.13	0.24
19	1665	parks	park	VICTORIA PARK	0.49	0.76	0.66	0.08	0.41
19	900	roads	street_nam	PARK	0.29	0.52	0.44	0.13	0.29
19	1555	parks	park	IRON HORSE TRAIL	0.49	0.76	0.66	0.08	0.41
19	1950	trails		trails	0.49	0.76	0.58	1.00	0.69
19	986	roads	street_nam	VICTORIA	0.23	0.52	0.44	0.13	0.26
19	1293	trails	route_name	IRON HORSE TRAIL	0.49	0.76	0.58	1.00	0.69
19	1957	railway_lines		railway	0.50	0.14	0.09	0.06	0.22
19	1954	trails		trail	0.49	0.76	0.58	1.00	0.69
19	1949	parks		parks	0.49	0.76	0.66	0.08	0.41
47	1950	trails		trails	0.50	0.76	0.58	1.00	0.69
47	1954	trails		trail	0.50	0.76	0.58	1.00	0.69
63	1950	trails		trails	0.50	0.76	0.58	1.00	0.69
63	1954	trails		trail	0.50	0.76	0.58	1.00	0.69
63	1949	parks		parks	0.49	0.76	0.66	0.08	0.41
63	900	roads	street_nam	PARK	0.30	0.52	0.44	0.13	0.29
97	577	roads	street_nam	QUEENS	-0.26	0.52	0.44	0.13	0.10
97	986	roads	street_nam	VICTORIA	0.15	0.52	0.44	0.13	0.24
97	1293	trails	route_name	IRON HORSE TRAIL	0.41	0.76	0.58	1.00	0.67
97	1954	trails		trail	0.41	0.76	0.58	1.00	0.67
97	1555	parks	park	IRON HORSE TRAIL	0.42	0.76	0.66	0.08	0.39
97	92	roads	street_nam	QUEEN	-0.26	0.52	0.44	0.13	0.10
97	1499	roads	street_nam	KING	0.13	0.52	0.44	0.13	0.23
97	199	roads	street_nam	STREET ONE	-2.33	0.52	0.44	0.13	-0.59
97	1950	trails		trails	0.41	0.76	0.58	1.00	0.67
113	164	roads	street_nam	OTTAWA	0.50	0.52	0.44	0.13	0.35
113	1954	trails		trail	0.50	0.76	0.58	1.00	0.69
113	1950	trails		trails	0.50	0.76	0.58	1.00	0.69
113	380	roads	street_nam	NYBERG	0.48	0.52	0.44	0.13	0.35
189	1950	trails		trails	0.49	0.76	0.58	1.00	0.69
189	1951	roads		roads	0.49	0.52	0.44	0.13	0.35
189	164	roads	street_nam	OTTAWA	0.49	0.52	0.44	0.13	0.35
189	1954	trails		trail	0.49	0.76	0.58	1.00	0.69

189	1293	trails	route_name	IRON HORSE TRAIL	0.49	0.76	0.58	1.00	0.69
189	1555	parks	park	IRON HORSE TRAIL	0.50	0.76	0.66	0.08	0.41
193	1950	trails		trails	0.49	0.76	0.58	1.00	0.69
193	812	roads	street_nam	WOOD	-1.29	0.52	0.44	0.13	-0.24
193	1955	cycling		bicycle	0.50	0.14	0.14	0.08	0.24
193	1954	trails		trail	0.49	0.76	0.58	1.00	0.69
193	630	roads	street_nam	WINDING	-2.18	0.52	0.44	0.13	-0.54
193	1273	roads	street_nam	WINDING WOOD	-3.12	0.52	0.44	0.13	-0.85
242	73	roads	street_nam	STIRLING	0.49	0.52	0.44	0.13	0.35
242	1950	trails		trails	0.50	0.76	0.58	1.00	0.69
242	1954	trails		trail	0.50	0.76	0.58	1.00	0.69
348	1949	parks		parks	0.50	0.76	0.66	0.08	0.41
348	1665	parks	park	VICTORIA PARK	0.50	0.76	0.66	0.08	0.41
348	986	roads	street_nam	VICTORIA	0.25	0.52	0.44	0.13	0.27
348	900	roads	street_nam	PARK	0.31	0.52	0.44	0.13	0.29
348	1950	trails		trails	0.50	0.76	0.58	1.00	0.69
348	233	roads	street_nam	WATER	0.30	0.52	0.44	0.13	0.29
348	1954	trails		trail	0.50	0.76	0.58	1.00	0.69
371	248	roads	street_nam	BELMONT	0.49	0.52	0.44	0.13	0.35
371	1052	roads	street_nam	UNION	0.49	0.52	0.44	0.13	0.35
371	1950	trails		trails	0.47	0.76	0.58	1.00	0.69
371	1379	roads	street_nam	VILLAGE	-1.11	0.52	0.44	0.13	-0.18
371	1302	roads	street_nam	NORTH	-0.46	0.52	0.44	0.13	0.03
371	1954	trails		trail	0.47	0.76	0.58	1.00	0.69

Table B-0.2: UrbanContext maual spot check results where results are semantically or spatially correct

GID	Match ID	Table Name	Search Term	Proximity	Cluster	Co-location	Topicality	GR Score	Semantic	Spatial
7	1954	trails	trail	0.47	0.76	0.58	1.00	0.69	TRUE	TRUE
7	1950	trails	trails	0.47	0.76	0.58	1.00	0.69	TRUE	TRUE
8	1951	roads	roads	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
17	1555	parks	IRON HORSE TRAIL	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
17	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
17	164	roads	OTTAWA	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
17	1293	trails	IRON HORSE TRAIL	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
17	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
17	1093	roads	FAIRVIEW	0.15	0.52	0.44	0.13	0.24	TRUE	FALSE
19	900	roads	PARK	0.29	0.52	0.44	0.13	0.29	FALSE	FALSE
19	1949	parks	parks	0.49	0.76	0.66	0.08	0.41	TRUE	TRUE
19	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
19	1957	railway_lines	railway	0.50	0.14	0.09	0.06	0.22	TRUE	FALSE
19	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
19	1941	railway_lines	railway_lines	0.50	0.14	0.09	0.06	0.22	TRUE	FALSE
19	1302	roads	NORTH	0.14	0.52	0.44	0.13	0.24	FALSE	FALSE
19	986	roads	VICTORIA	0.23	0.52	0.44	0.13	0.26	FALSE	FALSE
19	1555	parks	IRON HORSE TRAIL	0.49	0.76	0.66	0.08	0.41	TRUE	TRUE
19	1665	parks	VICTORIA PARK	0.49	0.76	0.66	0.08	0.41	TRUE	TRUE
19	1293	trails	IRON HORSE TRAIL	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
26	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
26	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
31	577	roads	QUEENS	-0.06	0.52	0.44	0.13	0.17	TRUE	TRUE
31	92	roads	QUEEN	0.05	0.52	0.44	0.13	0.20	TRUE	TRUE
38	1951	roads	roads	0.47	0.52	0.44	0.13	0.34	TRUE	TRUE
38	1956	cycling_infrastructure	cyclist	0.41	0.14	0.14	0.08	0.21	TRUE	TRUE
47	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
47	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
63	1949	parks	parks	0.49	0.76	0.66	0.08	0.41	TRUE	TRUE
63	900	roads	PARK	0.30	0.52	0.44	0.13	0.29	TRUE	TRUE
63	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
63	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
64	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
64	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
64	237	roads	WENTWORTH	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
95	1956	cycling_infrastructure	cyclist	0.38	0.14	0.14	0.08	0.20	FALSE	FALSE
95	1943	bridge	bridge	0.49	0.10	0.07	0.11	0.22	TRUE	TRUE
95	603	roads	BRIDGE	-1.76	0.52	0.44	0.13	-0.40	TRUE	TRUE
95	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE

95	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
98	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
98	659	roads	BROCK	0.40	0.52	0.44	0.13	0.32	TRUE	TRUE
98	569	roads	HIGHLAND	0.37	0.52	0.44	0.13	0.31	TRUE	TRUE
98	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
105	346	roads	ROCKWAY	0.25	0.52	0.44	0.13	0.27	FALSE	FALSE
105	1472	roads	SCHNEIDER	-0.45	0.52	0.44	0.13	0.04	TRUE	TRUE
105	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
105	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
109	1052	roads	UNION	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
109	1951	roads	roads	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
113	380	roads	NYBERG	0.48	0.52	0.44	0.13	0.35	TRUE	TRUE
113	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
113	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
113	164	roads	OTTAWA	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
123	900	roads	PARK	0.31	0.52	0.44	0.13	0.29	FALSE	FALSE
123	1665	parks	VICTORIA PARK	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
123	1949	parks	parks	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
123	986	roads	VICTORIA	0.32	0.52	0.44	0.13	0.29	FALSE	FALSE
141	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
141	986	roads	VICTORIA	0.49	0.52	0.44	0.13	0.35	TRUE	FALSE
141	997	roads	WEST	0.47	0.52	0.44	0.13	0.34	TRUE	FALSE
141	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
150	249	roads	HENRY	0.32	0.52	0.44	0.13	0.29	FALSE	FALSE
150	710	roads	GREEN	-0.14	0.52	0.44	0.13	0.14	FALSE	FALSE
150	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
150	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
158	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
158	1955	cycling_infrastructure	bicycle	0.26	0.14	0.14	0.08	0.16	TRUE	FALSE
158	1555	parks	IRON HORSE TRAIL	0.49	0.76	0.66	0.08	0.41	TRUE	TRUE
158	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
158	986	roads	VICTORIA	0.21	0.52	0.44	0.13	0.26	TRUE	TRUE
158	1293	trails	IRON HORSE TRAIL	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
158	1956	cycling_infrastructure	cyclist	0.26	0.14	0.14	0.08	0.16	TRUE	FALSE
158	1476	roads	WALKER	-2.72	0.52	0.44	0.13	-0.72	FALSE	FALSE
189	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
189	1293	trails	IRON HORSE TRAIL	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
189	1951	roads	roads	0.49	0.52	0.44	0.13	0.35	FALSE	FALSE
189	164	roads	OTTAWA	0.49	0.52	0.44	0.13	0.35	TRUE	TRUE
189	1555	parks	IRON HORSE TRAIL	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
189	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE

190	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
190	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
190	986	roads	VICTORIA	0.38	0.52	0.44	0.13	0.31	TRUE	TRUE
193	630	roads	WINDING	-2.18	0.52	0.44	0.13	-0.54	FALSE	FALSE
193	1273	roads	WINDING WOOD	-3.12	0.52	0.44	0.13	-0.85	FALSE	FALSE
193	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
193	812	roads	WOOD	-1.29	0.52	0.44	0.13	-0.24	TRUE	FALSE
193	1955	cycling_infrastructure	bicycle	0.50	0.14	0.14	0.08	0.24	TRUE	TRUE
193	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
199	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
199	1424	roads	GLASGOW	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
199	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
199	1949	parks	parks	0.50	0.76	0.66	0.08	0.41	FALSE	FALSE
199	900	roads	PARK	0.28	0.52	0.44	0.13	0.28	FALSE	FALSE
214	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
214	1046	roads	SYDNEY	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
214	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
217	1954	trails	trail	0.47	0.76	0.58	1.00	0.69	TRUE	FALSE
217	1950	trails	trails	0.47	0.76	0.58	1.00	0.69	TRUE	FALSE
231	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
231	1555	parks	IRON HORSE TRAIL	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
231	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
231	377	roads	CHERRY	0.48	0.52	0.44	0.13	0.35	TRUE	TRUE
231	1293	trails	IRON HORSE TRAIL	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
232	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
232	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
232	1293	trails	IRON HORSE TRAIL	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
232	1555	parks	IRON HORSE TRAIL	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
242	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
242	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
242	73	roads	STIRLING	0.49	0.52	0.44	0.13	0.35	TRUE	TRUE
270	346	roads	ROCKWAY	0.23	0.52	0.44	0.13	0.26	FALSE	FALSE
270	164	roads	OTTAWA	0.47	0.52	0.44	0.13	0.34	TRUE	TRUE
270	682	roads	GARDEN	-0.56	0.52	0.44	0.13	0.00	FALSE	FALSE
270	1954	trails	trail	0.47	0.76	0.58	1.00	0.69	TRUE	TRUE
270	1950	trails	trails	0.47	0.76	0.58	1.00	0.69	TRUE	TRUE
270	1202	roads	SOUTH	-0.59	0.52	0.44	0.13	-0.01	FALSE	FALSE
270	1848	parks	ROCKWAY GARDENS	0.49	0.76	0.66	0.08	0.41	TRUE	TRUE
279	1202	roads	SOUTH	-0.32	0.52	0.44	0.13	0.08	FALSE	FALSE
279	1950	trails	trails	0.45	0.76	0.58	1.00	0.68	TRUE	TRUE
279	1954	trails	trail	0.45	0.76	0.58	1.00	0.68	TRUE	TRUE

282	1951	roads	roads	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
282	1950	trails	trails	0.45	0.76	0.58	1.00	0.68	TRUE	FALSE
282	1954	trails	trail	0.45	0.76	0.58	1.00	0.68	TRUE	FALSE
292	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
292	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
314	92	roads	QUEEN	0.48	0.52	0.44	0.13	0.35	TRUE	FALSE
314	577	roads	QUEENS	0.35	0.52	0.44	0.13	0.30	TRUE	FALSE
335	986	roads	VICTORIA	0.44	0.52	0.44	0.13	0.33	TRUE	TRUE
335	900	roads	PARK	0.30	0.52	0.44	0.13	0.29	FALSE	FALSE
335	997	roads	WEST	0.45	0.52	0.44	0.13	0.34	TRUE	TRUE
335	1665	parks	VICTORIA PARK	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
335	1949	parks	parks	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
348	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
348	900	roads	PARK	0.31	0.52	0.44	0.13	0.29	FALSE	FALSE
348	1949	parks	parks	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
348	233	roads	WATER	0.30	0.52	0.44	0.13	0.29	FALSE	FALSE
348	1665	parks	VICTORIA PARK	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
348	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
348	986	roads	VICTORIA	0.25	0.52	0.44	0.13	0.27	FALSE	FALSE
356	900	roads	PARK	0.20	0.52	0.44	0.13	0.25	FALSE	FALSE
356	377	roads	CHERRY	0.49	0.52	0.44	0.13	0.35	FALSE	FALSE
356	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
356	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
356	1533	parks	CHERRY PARK	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
356	1949	parks	parks	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
381	1956	cycling_infrastructure	cyclist	0.42	0.14	0.14	0.08	0.21	FALSE	FALSE
395	986	roads	VICTORIA	-0.33	0.52	0.44	0.13	0.08	TRUE	FALSE
402	1642	parks	HENRY STURM GREENWAY	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
402	249	roads	HENRY	0.25	0.52	0.44	0.13	0.27	FALSE	FALSE
412	1957	railway_lines	railway	0.50	0.14	0.09	0.06	0.22	TRUE	TRUE
419	1379	roads	VILLAGE	-0.98	0.52	0.44	0.13	-0.14	FALSE	FALSE
419	248	roads	BELMONT	0.49	0.52	0.44	0.13	0.35	TRUE	TRUE
419	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
419	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
425	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
425	1259	roads	DANIEL	-1.95	0.52	0.44	0.13	-0.46	FALSE	FALSE
425	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
446	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
446	900	roads	PARK	0.31	0.52	0.44	0.13	0.29	FALSE	FALSE
446	1949	parks	parks	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
446	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE

446	1665	parks	VICTORIA PARK	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
446	986	roads	VICTORIA	0.37	0.52	0.44	0.13	0.31	FALSE	FALSE
453	1848	parks	ROCKWAY GARDENS	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
453	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
453	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
453	268	roads	MONTGOMERY	-0.16	0.52	0.44	0.13	0.13	TRUE	TRUE
453	164	roads	OTTAWA	0.49	0.52	0.44	0.13	0.35	TRUE	TRUE
453	682	roads	GARDEN	-0.52	0.52	0.44	0.13	0.01	FALSE	FALSE
453	346	roads	ROCKWAY	0.19	0.52	0.44	0.13	0.25	FALSE	FALSE
459	1555	parks	IRON HORSE TRAIL	0.50	0.76	0.66	0.08	0.41	TRUE	TRUE
459	1293	trails	IRON HORSE TRAIL	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
459	1954	trails	trail	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
459	1950	trails	trails	0.49	0.76	0.58	1.00	0.69	TRUE	TRUE
462	52	roads	MAY	-0.73	0.52	0.44	0.13	-0.05	FALSE	FALSE
462	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
462	1956	cycling_infrastructure	cyclist	0.24	0.14	0.14	0.08	0.15	TRUE	FALSE
462	1364	roads	GAGE	0.50	0.52	0.44	0.13	0.35	TRUE	
462	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
479	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
479	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	FALSE
482	380	roads	NYBERG	0.49	0.52	0.44	0.13	0.35	TRUE	TRUE
482	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
482	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
482	164	roads	OTTAWA	0.50	0.52	0.44	0.13	0.35	TRUE	TRUE
483	900	roads	PARK	0.22	0.52	0.44	0.13	0.26	FALSE	FALSE
483	248	roads	BELMONT	0.46	0.52	0.44	0.13	0.34	TRUE	TRUE
483	1552	parks	BELMONT PARK	0.50	0.76	0.66	0.08	0.41	FALSE	FALSE
483	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
483	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
483	1949	parks	parks	0.50	0.76	0.66	0.08	0.41	FALSE	FALSE
489	1949	parks	parks	0.50	0.76	0.66	0.08	0.41	FALSE	FALSE
489	52	roads	MAY	-0.47	0.52	0.44	0.13	0.03	FALSE	FALSE
489	1950	trails	trails	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
489	1954	trails	trail	0.50	0.76	0.58	1.00	0.69	TRUE	TRUE
489	900	roads	PARK	0.25	0.52	0.44	0.13	0.27	FALSE	FALSE