Theses and Dissertations (Comprehensive)

2014

Integrated spatial analysis of volunteered geographic information

Haydn Lawrence
Wilfrid Laurier University, haydnlawrence@gmail.com

INTEGRATED SPATIAL ANALYSIS OF VOLUNTEERED GEOGRAPHIC INFORMATION

by

Haydn Roger Lawrence

BCompSc, University of New Brunswick, 2002
MA, University of New England, 2010

THESIS

Submitted to the Department of Geography and Environmental Studies

in partial fulfillment of the requirements for

MASTER OF SCIENCE

Wilfrid Laurier University

© Haydn Lawrence 2014

Wilfrid Laurier University

INTEGRATED SPATIAL ANALYSIS OF VOLUNTEERED GEOGRAPHIC INFORMATION

by

Haydn Roger Lawrence

BCompSc, University of New Brunswick, 2002
MA, University of New England, 2010

Supervisory Committee

_____

Dr. Colin Robertson, Supervisor
(Department of Geography and Environmental Studies, Wilfrid Laurier University)

_____

Dr. Rob Feick, Member
(School of Planning, University of Waterloo)

_____

Dr. Trisalyn Nelson, Outside Member
(Department of Geography, University of Victoria)

Abstract

Volunteered Geographic Information (VGI) is becoming a pervasive form of data within geographic academic research. VGI offers a relatively new form of data, one with both potential as a sensitive way to collect information about the world, and challenges associated with unknown and heterogeneous data quality. The lack of sampling control, variable expertise in data collection and handling, and limited control over data sources are significant research challenges. In this thesis, data quality of VGI is tackled as a general composite measure based on coverage of the dataset, the evenness in the density of data, and the relative evenness in contributors to a given dataset. A metric is formulated which measures these properties for VGI point pattern data. The utility of the metric for discriminating qualitatively different types of VGI is evaluated for different forms of VGI, based on a relative comparison framework. The metric is used to optimize both the spatial grains and spatial extents of several VGI study areas. General methods are created to support the assessment of data quality of VGI datasets at several spatial scales.

_____

Dr. Colin Robertson, Supervisor
(Department of Geography and Environmental Studies, Wilfrid Laurier University)

_____

Dr. Rob Feick, Member
(School of Planning, University of Waterloo)

_____

Dr. Trisalyn Nelson, Outside Member
(Department of Geography, University of Victoria)

_____

Dr. Robert McLeman, External Examiner
(Department of Geography and Environmental Studies, Wilfrid Laurier University)

Acknowledgements

I would like to thank my supervisor, Dr. Colin Robertson, for the boundless and immeasurable help he provided in the creation of this thesis. His near instantaneous replies to questions or help with ideas are one of the key factors to keeping myself on track while consistently pushing myself just a little bit harder to accomplish something I can be proud of. Coupled with his devotion to seeing his students succeed, whether in academia or elsewhere, I could not have accomplished near as much without his time and support. I would also like to thank both Dr. Rob Feick and Dr. Trisalyn Nelson, both of whom offered substantially more time and support than I have seen of most other supervisory committee members at the Masters level. Their ideas allowed for a much more comprehensive thesis. I would also like to thank all the members of the Spatial Lab at Wilfrid Laurier University for their support and conversations, especially throughout the final stages of the thesis

Table of Contents

List of Tables

List of Figures

Figure 3.1: Voronoi polygons of test data highlighting the density metric component. The image on the right denotes a Voronoi tessellation created from generic test data. The four polygons found are all adjacent in geographic (real) space (a) and close together in the Moran's Scatterplot (b). The

Chapter One

1.0 Introduction

1.1 Background

Participation of the general public in the creation, collection, analysis, and/or communication of geospatial data, whether volunteered or unknowingly, has become ubiquitous within society as many aspects of our social, cultural, and economic lives interface with digital technologies that record and store digital traces of these processes. User-generated data has been facilitated by rapid advancements in internet technology, faster data transfer speeds, online social networking websites, easy-to-use interactive mapping and collaboration tools, and the proliferation of mobile computing devices (Goodchild, 2007; Gura, 2013). Geographers have shown great interest in the research value of data obtained from these non-authoritative sources, described as user-generated content (UGC), or specifically, the subset of these data that record geospatial information commonly referred to as volunteered geographical information (VGI) (Goodchild, 2007). The forms of VGI vary from data created without any geographic purpose in mind but nevertheless associated with a geographic footprint (e.g. a geo-tweet) to spatial data purposefully collected and contributed to a specific project. Numerous examples of these forms of data driving research and applications in environmental and social sciences exist, such as tracking wildlife poaching (Stevens et al., 2013), urban noise pollution (Gura, 2013), responses to climate change (Beaubien & Hamann, 2011; Worthington et al., 2012), and biogeography and citizen science (Dickinson et al., 2012). Yet these developments are occurring without an understanding of how these data differ from traditional data sources, nor the theory or tools to assess these differences. There is growing consensus that issues of data quality must first be addressed before VGI as a data source, and a method of obtaining data about the world, can be harnessed. In an era of constrained financial resources for

research, can VGI be a valuable and cost-effective tool for researchers seeking empirical data across broad spatial areas and/or extended time periods?

In order to answer this question, a stronger theoretical foundation for VGI is necessary, including methodological tools that will support the analysis and evaluation of these emergent forms of geographical information. The most direct approach to assess how VGI differs from the authoritative forms of data sources that have been used previously, such as census data or government mapping. The first studies (Haklay 2010; Mooney et al., 2010; Zielstra, 2010) into data quality for VGI leveraged this comparative approach. However this framework fails when faced with data that do not exist in authoritative form. More general methods are required to evaluate VGI when reference data is lacking. VGI and Big Data offers researchers unprecedented opportunities for analysis across large spatial and temporal scales while maintaining fine-grained samples and measurements, and constraining these opportunities to those for which reference data already exists. The research reported in this thesis aims to develop general methods for data quality evaluation for VGI in absence of reference data.

One of the principle challenges, and only levers of control over analysis on behalf of the researcher, is deciding on the spatial scale required for analysis of VGI. Wiens (1989) summarizes the trade-offs inherent in arbitrary choices of spatial scale, and these issues are increasingly evident in empirical analyses of VGI. VGI is often obtained with modifiable extents – ranging from a single neighbourhood to the entire Earth. This lack of structure is especially relevant for the more ambient forms of VGI as there is no central authority in which to control/suggest where and what kind of data should be created. While predetermined limitations may arise, such as when comparing VGI to satellite imagery available at a specific resolution for example, spatial grain and

extent are generally at the discretion of the researcher, and too often chosen arbitrarily. Kelly

(2011), Galpern (2012), and Calderón-Patrón (2013) all analyse study areas at several different grain

sizes and discuss the benefits of these multi-scale analyses when analyzing different types of data,

including VGI. The choice of grain can pose similar problems when examining patterns in VGI as

trends may be found at certain scales while lost at others (Feick & Robertson, 2013). Neis (2012)

and Haklay (2010) both found that there are areas that show less reliability within their study areas

of OSM, particularly in rural areas of Germany (Neis) and lower socio-economic regions of London,

England (Haklay). These 'lower quality' areas would cause problems if these extents were used for

analysis without prior review analyses. In addition, with some forms of VGI, the relevant spatial

scales may be unknown to a researcher, such as in newly gentrified areas or culturally delineated,

though unofficial, neighbourhoods (Sester, 2014).   If researchers are to use these new forms of

data, general methods are required to support the assessment of data quality of VGI datasets at

several spatial scales.

1.2 Objectives

This thesis will approach the problem of measuring data quality of VGI, and focus specifically on

identifying optimal spatial scale. There are three main components to the research: the

development of a metric to evaluate data quality in point-based VGI, an implementation of the

metric in a multi-scale analysis setting, and the development of optimization based on the defined

data quality metric. The research objectives of this thesis are as follows:

1)   Develop a metric to assess data quality VGI point datasets, both ambient and active.

2)   Develop a methodology for optimizing spatial grain for the analysis of VGI based on the

data quality metric defined in objective 1.

3)   Develop a methodology for optimizing spatial extent for the analysis of VGI based on the data quality metric defined in objective 1.   The methodology should be robust to different forms of aggregation unit (regular and irregular polygons).

4)   Implement the methods in objectives 1-3 and demonstrate their use in the analysis of publicly available VGI point data as well a VGI citizen-science case study.

References

Beaubien, E. G., & Hamann, A. (2011). Plant phenology networks of citizen scientists: recommendations from two decades of experience in Canada. *International journal of biometeorology*, *55*(6), 833–841.

Calderón-Patrón, J. et al., 2013. Vertebrate Dissimilarity Due to Turnover and Richness Differences in a Highly Beta-Diverse Region: The Role of Spatial Grain Size, Dispersal Ability and Distance. *Plos One*. 8 (12).

Conrad, C. C., & Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental monitoring and assessment*, *176*(1-4), 273–291.

Feick, R. & Robertson, C., 2014. A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems*.

Galpern, P., Manseau, M., Wilson, P., 2012. Grains of connectivity: Analysis at multiple spatial scales in landscape genetics. *Molecular Ecology*. 20 (16), p. 3996-4009.

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

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

Gura, T. (2013). Citizen science: Amateur experts. *Nature*, *496*(7444), 259–261.

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

Kelly, M., Tuxen, K., Stralberg, D., 2011. Mapping changes to vegetation pattern in a restoring wetland: Finding pattern metrics that are consistent across spatial scale and time. *Ecological Indicators*. 11 (2), p. 263-273.

Li, L., Goodchild, M., & Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40, p. 61-77.

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 514–517). New York, NY, USA: ACM. doi:10.1145/1869790.1869875

Neis, P., Zielstra, D., & Zipf, A., 2012. The Street Network Evolution of Crownsourced Maps: OpenStreetMap in Germany 2007-2011. *Future Internet*. 4 (1), p. 1-21.

Sester, M., Jokar, J., Arsanjani, R., Burghardt, D., & Haunert, J., 2014. "Integrating and Generalising Volunteered Geographic Information." In *Abstracting Geographic Information in a Data Rich World – Lecture Notes in Geoinformation and Cartography*. Springer International Publishing. p. 119-155. Print.

Silvertown, J., Cook, L., Cameron, R., Dodd, M., McConway, K., Worthington, J., Juan, X. (2011). Citizen Science Reveals Unexpected Continental-Scale Evolutionary Change in a Model Organism. *PLoS ONE*, *6*(4), e18927. doi:10.1371/journal.pone.0018927

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

Stevens, M., Vitos, M., Lewis, J., & Haklay, M. (2013). Participatory monitoring of poaching in the Congo basin. Retrieved from http://www.geos.ed.ac.uk/~gisteac/proceedingsonline/GISRUK2013/gisruk2013_submission_12.pdf

Wiens, J., 1989. Spatial Scaling in Ecology. *Functional Ecology*. 3 (4), p.385-397.

Worthington, J. P., Silvertown, J., Cook, L., Cameron, R., Dodd, M., Greenwood, R. M., Skelton, P. (2012). Evolution MegaLab: a case study in citizen science methods. *Methods in Ecology and Evolution*, *3*(2), 303–309.

Zielstra, D. & Zipf, A., 2010. "A comparative study of proprietary geodata and volunteered geographic information for Germany." *13th AGILE international conference on geographic information science*.

Chapter Two

2.0 Identifying optimal study areas and spatial aggregation units for point-based VGI from multiple sources.

Haydn Lawrence, Colin Robertson, Rob Feick, and Trisalyn Nelson

ABSTRACT:

Researchers in the social sciences are increasingly using new sources of geo-referenced data such as social media posts, web traffic, and other forms of user-generated content to gain new insights about how people perceive and experience their surroundings. Many studies have shown us the applicability of VGI as a data source that supports a variety of interesting applications. However, few methods exist to offer guidance on whether sufficient VGI is available for a specific research task or, more fundamentally, what scale VGI can be aggregated and analysed at within a specific locale. In lieu of this guidance and given the patchy and heterogeneous nature of VGI, it is difficult for researchers to address questions such as: "Are there enough data for a successful study?" and "How clustered are they?"

In this paper, we introduce a new metric for evaluating feasible VGI study areas and the appropriateness of different aggregation unit sizes through three different components of data quality: coverage, density, and user-heterogeneity. Two popular sources of passive VGI are used for initial testing of the metric: Twitter and Flickr. We compare the component and aggregate measures for different simulated point processes and demonstrate the properties of this metric. The three components are assessed iteratively for the point user generated data (tweets and photos) on a local basis by altering grain sizes. This refinement of cell size continues until the coverage, density, and user-heterogeneity component are optimized relative to pre-determined ranges for both high quality and low quality. We demonstrate the application of this metric with

Flickr and Twitter data obtained for three Canadian cities as initial study areas, including

Vancouver, Toronto, and Moncton. The utility of the metric for discriminating qualitatively different

types of VGI is evaluated for each of these areas based on a relative comparison framework. Finally,

we present a use-case for this metric: identifying the optimal spatial grain and extent for a given

data set. The results of this analysis will provide a methodology for preliminary evaluation of VGI

quality within a given study area, and identify sub-areas with desirable characteristics.

2.1 Introduction

In 2007, Goodchild (2007) first coined the subset of web-based user-generated content, including blogs, Facebook posts, or tweets with geographic properties, as volunteered geographic information (VGI). Since then, VGI has figured prominently in the GIScience research agenda, highlighting the geospatial properties of data from applications such as Flickr, Twitter, and OpenStreetMap (OSM). VGI encompasses many types of data, and can be further subdivided into ambient VGI, represented by Twitter and Flickr (Stefanidis, 2013), and active VGI, represented by OSM. Ambient VGI is less 'volunteered' (and may even be considered coerced – see Mackenzie and Janowicz 2014) as users are likely not creating the data for a specific research purpose. These same sources are now being used to open up new spatial research questions in the social sciences (Goodchild and Janelle, 2010). Many have speculated that VGI gives researchers the capability to gauge the sentiment of a geographically defined study population without use of traditional methods of qualitative research, such as questionnaires or direct observation. For example, VGI was recently employed to examine sentiments among community members in the path of a hurricane (Lachlan et al, 2014), allowing people to freely express their emotions, and importantly, enabling timely, and spatially defined assessment of population sentiment. Similarly, VGI may enable access to otherwise difficult to access populations (Stefanidis, 2011). While methods for evaluating data quality for new sources of VGI is a burgeoning research area of GIScience (Neis and Zielstra, 2014; Jeffery, 2014; McKenzie et al, 2014; Hollenstein, 2010), there are few tools available to assess the characteristics of VGI for a given research problem or application area.

The majority of studies into data quality assessment for VGI have been narrowly defined in terms of geographic scale and in the data sources considered (e.g. testing specific sources in

isolation of others). Two frequently cited studies of OpenStreetMap that investigated data quality in VGI found OSM data to be very accurate within urban areas and sporadic in remote areas (Zielstra and Zipf, 2010; Haklay, 2010). However these studies were limited to narrow geographic and temporal scales by choosing areas with expected high OSM data. While initial investigations of data quality of VGI have indicated there is a real potential for VGI as a valuable source of data, these studies have largely focused on areas where large amounts of VGI were expected to be found, such as Tokyo (Stefanidis, 2012) and major population centres of Europe (Zielstra and Zipf, 2010; Haklay, 2010). Others have focused on discrete space-time events such as wildfires (Goodchild and Glennon, 2010) and earthquakes (Zook et al, 2012). Other studies such as Hollenstein and Purves' study of Flickr (2010) or Crooks' study of Twitter (2013) are also limited by their source (Flickr or Twitter only). The specific foci found in the studies mentioned above can be seen as particularly problematic when the ephemeral qualities of technology are considered, illustrated by the downfall of the once popular MySpace or the restricted access and terms of service policies for privately held data from social-networking companies such as Facebook. Examples from the literature analyzing multiple sources of VGI simultaneously have begun to emerge, for example Croitoru et al. (2013) using Flickr, Twitter, and YouTube while Li et al. (2013) use both Flickr and Twitter densities for their study of user-demographics in California, though these studies do not look at the data quality characteristics of the data as the central focus. In addition, Mearns (2014) highlights the potential of moving from just Twitter based analysis, which the study focuses on, to data from multiple social media platforms. They also describe a system which works in real time, which would be a very useful tool. Unfortunately, not all sources of VGI currently allow for real-time/streaming access.

Li et al. (2013) examined socioeconomic variables in relation to Twitter and Flickr submissions, finding a narrow subset of the underlying population (mostly rich and educated) were over-represented. However, the Li et al. study (2013) analysis of point densities was temporally limited to a period that was considered optimal and considered only the United States on a geographic scale with the demographic comparison analysis limited to a region of California. While the studies mentioned previously (Hollenstein and Purves, 2013; Stefanidis, 2013; Li et al., 2013) are instrumental in our understanding of these relatively new forms of data, the isolation of the studies' data to specific scales (geographic, temporal) and individual sources of VGI limits the generality of their results. There is a considerable research need for general data quality assessment tools specific for VGI.

Geographic information has typically been assessed through traditional methods such as the authority or reputation of the data collectors, industry and/or international standards (e.g., ISO 19157:2013), or metrics based on a comparison to reference data (e.g., root mean square error). However, these methods may not be available to researchers attempting to assess VGI quality. A lack of standards, multiple and anonymous data collectors, a lack of comparable reference data, and even multiple and sometimes conflicting contributor motivations all contribute to knowledge gaps in data quality assessment methodology for VGI (Coleman, 2009; Foody, 2013; Mooney, 2013b). This constitutes the necessity of a VGI assessment tool stemming from the problem that while traditional data collection methods exert control over sampling plans and study areas, VGI, by definition, allows for virtually no control over the data collection process. The sampling plan when using most types of VGI is necessarily post-hoc. For large-scale web-based social media applications, researchers are forced to collect data within the scope of the given APIs, for available time periods and geographic locations. However, four parameters of

control available for establishing some research design for VGI include spatial and temporal

scales of the study, the thematic focus, and the sources of VGI used in the study (i.e. changing to

a core area and using Twitter instead of Flickr). In this paper, we consider whether tools for

assessment of VGI data quality can begin to be developed based on explicit consideration of

these four parameters.

Few methods exist to offer guidance on whether sufficient VGI is available for a specific research

task or, more fundamentally, to what scale VGI can be aggregated and analysed within a specific

locale. In lieu of this guidance and given the patchy and heterogeneous nature of VGI, it is

difficult for researchers to address questions such as: "What areas have enough VGI of a given

type for my analysis?", "If a study area is defined externally, what resolution or spatial units of

analysis can be used?", "How representative is the VGI as measured by user-heterogeneity?", or

"What correspondence is there between the VGI and pre-defined zones such as census tracts or

ecozones?" This paper will explore these issues through the use of a metric designed to evaluate

ambient point-based VGI, information with geographic footprints though not actively created as

geographic information (Stefanidis, 2013). We will examine the metric, computed over various

aggregation unit sizes through three components of VGI quality: coverage, density, and user-

heterogeneity. The methods described here are generic in that they are designed as a

standalone VGI assessment tool, without reference to any authoritative or expert comparative

datasets. A key practical outcome of this research will be a set of open source tools which will

be directed at VGI evaluation / assessment from a user-oriented perspective, irrespective of

scale or source (e.g. Twitter, Foursquare, or Yelp). The last aspect is important as private

entities may change their data dissemination policies at any time, potentially restricting access.

We start with a definition of the aforementioned metric followed by an examination of its

properties for two types of spatial point processes. We follow with a case study of point data

obtained from Twitter and Flickr for three different Canadian cities: Toronto, Vancouver, and

Moncton, measuring the metric at different grains within two different extents for each city. We

conclude with a discussion of current limitations and possibilities for future research for

evaluating data quality in VGI point patterns, especially in the context of optimizing initial

research study areas.

2.2   Methods

2.2.1 Developing a metric for evaluating VGI point patterns

VGI datasets are highly variable. Data can vary with the particular VGI data source's sharing

policies such as Twitter's tiered access model which ranges from 1% of data being available for

free to 100% of the data (i.e., 'the fire hose') available at a significant cost. Additionally, due to

the nature of VGI creation, these data are prone to uncertainty. For example, studies of

geotagged photographs often contend with variability introduced by two mechanisms of

tagging: in-situ tagging while photos are taken, and post-hoc bulk-tagging of photos during

management and upload to online sharing sites (Hollenstein and Purves 2010). These problems

necessitate a method of determining if a chosen study area meets a researcher's needs,

preferably before in-depth analysis begins. As the social sciences commonly use auxiliary

aggregated datasets, such as census data (Li et al., 2013; Granell, 2014), we develop a metric for

assessing VGI in aggregated grids of different spatial grains to allow for associations between

VGI and other datasets. One of the important aspects of this approach is to mitigate the fact

that aggregations, especially with authoritative boundaries, tend to be chosen arbitrarily

(Jeffery, 2014) and these boundaries can impact analytic results (Openshaw and Taylor 1979).

Thus we want a framework for evaluating VGI across aggregated geographies that can be easily

quantified in order to test different areal units. Like the K-function, which assesses clustering or

dispersion of point patterns at multiple distances (Ripley 1977, the approach here is first explored for multiple grains within the study area for all three metric components.

There are three components to the metric: coverage, density, and user-heterogeneity. Coverage is defined as the ratio of the number of cells that contain data to the total number of cells within the study area grid. While a lower numeric limit on the number of aggregated points within a grid cell could be implemented, this paper considered a cell to contain data on a binary scale – it either contains or doesn't contain data. This component is used as a global indicator of overall VGI coverage, on which Mooney et al. (2013a) describe a very discernable contrast between urban and remote areas and even within different urban areas (such as parts of Dublin compared to parts of Paris). This could be used to delineate remote areas or as a socio-economic indicator as Neis et al. (2013) show that lower coverage (completeness) can be attributed to lower socio-economic standing. Values approaching one indicate a greater coverage over the study area (i.e. most cells contain data) while values closer to zero are indicative of the opposite.

Density is assessed by calculating the areas of the four quadrants of a Moran's scatterplot and finding the difference between the largest area and the smallest area, normalized by study area size. This method differs from the local Moran's I value as it accounts for extreme x values instead of being averaged out within the algorithm, helping to find outlier core areas. Experiments with standard measures of spatial autocorrelation revealed that outliers were an important component the metric would need to be sensitive to, as often user-contributed data is patchy. For example, a downtown core or entertainment district may be a severe outlier in terms of number of submissions (high density), but depending on the aggregation scale, this

area may or may not be flagged as unusual in a spatial statistic based on adjacent neighbouring areas.

User-heterogeneity is a measure of the number of VGI submissions to the number of unique users averaged over the study area. Values close to one indicate a similar number of contributions among unique users, which could be considered higher quality as per Linus's Law. However, values nearer to zero could be useful in a study of high value users and their impact on social media (Stefanidis, 2013). These three components were chosen for their ability to assess a point dataset irrespective of any thematic content inherent in the data, such as a tweet's text or Flickr photo's tags or image. This helps to maintain the overall goal of this paper to assess VGI point patterns in a general way without creating specific criteria based on the technology or application used. User-heterogeneity was chosen as a data quality component, density was chosen for sensitivity to spatial clustering and extreme outlier detection, and coverage for considering the spatial extent of the data relative to the aggregation unit. The density and user-heterogeneity metrics were designed specifically for assessment of VGI. Whereas most methods for local spatial analysis are tuned for identifying clustering of high or low values, we aim to detect clustering in similar values (not high, not low). Similarly for user-heterogeneity, no existing methods captured the dynamics of the relationship between variance in submissions per user and hypothesized data quality.

2.2.2 Implementing the metric

The study area is rasterized into a grid specified by the bounding box size of the point pattern and the current grain size. The raster is intersected with the polygonal study area and all cells outside of the polygonal study area are set to NA. All points within each cell are aggregated by total counts and by counts of unique users. A queen's case neighbourhood matrix is created

and the three components built as per the descriptions above. The three components are weighted to sum to a value of 1 and the final metric computed from these weightings.

2.3   Simulation Study
2.3.1 Simulation study data

Two spatial point processes were chosen to simulate point patterns that could constitute possible VGI distributions under conditions of spatial randomness and spatial clustering. Our objective was to generate point pattern realizations and evaluate the metric under each scenario to examine its sensitivity to different configurations of locations and users. These patterns were created on the unit square, though transformed to a square of 100 by 100 units. The two processes were run 999 times with the metric components computed for each run and averaged for final results. The simulations were done in the R programming language using the spatstat package (Baddeley, 2005). The delineation between a poor quality distribution versus a good quality distribution would be dependent upon the goals of the research in question. For the purposes of this paper, highly clustered areas or areas with extreme outliers are considered poor. Low user-heterogeneity is also considered poor as we consider the idea of community consensus to be the ideal context for many types of analyses of VGI. As such, areas where data is dominated by a one or a small number of users would reflect be less useful than areas with broad participation from many users (Figure 1).

A uniform marked multitype Poisson process was used to realize random point patterns where points were randomly labelled (marks constitute the user IDs in VGI point data). A randomly distributed point pattern would indicate that each location has an equal probability of a point event, which we consider to be high quality data. Similarly, as points are randomly labelled to

reflect random spatial allocation of users, each area has equal chance of being visited by each

user.

$$X \sim Poisson(\lambda)$$
where λ=10, types=50

A stationary Matern cluster process was used to simulate a clustered area of VGI, creating a

point pattern mirroring a few users based in a handful of areas creating the majority of data.

$$X \sim Matern(\kappa, r, \mu)$$
where κ=5, r(radius)=0.15, μ=100

2.3.2 Simulation study results

The simulations match the predicted results for each component of the metric (Figure 2).

Coverage for the completely spatial random (CSR) point patterns is very high except at very

small grain sizes. This is caused by having too many small cells which contain no points. The

clustered pattern yields a low metric value except at high grain sizes, caused by having too few

grid cells with only one point needed to be captured by the coverage component. Density

findings were consistent with expected spatial autocorrelation values of CSR and clustered point

patterns. The CSR point pattern, except at extremely small grains, revealed little spatial

autocorrelation or few extreme values of data while the clustered simulations produced values

that indicate clustering. User-heterogeneity was found to be very low in a clustered area with

few users and high amounts of data. At the smaller grain size (2 units), user-heterogeneity is

high in a clustered area as the cell size is so small that very few points are captured by each cell.

The opposite can be seen with the CSR point pattern with large grains creating grids of only a

few cells capturing a large number of points decreasing user-heterogeneity. However, user-

heterogeneity follows the predicted outcomes at most grain sizes for the CSR point pattern by

having a high value showing a closer ratio of unique users to data points. These both follow

predicted patterns of user-heterogeneity.

2.4   Empirical Case Study
2.4.1 Study areas and Data Collection

Three Canadian cities were chosen as study areas: Toronto, Vancouver, and Moncton. Toronto

and Vancouver were chosen as large cities with dense population centres while Moncton was

chosen as a contrast to the two larger cities while maintaining a large enough population density

to have some VGI (Table 2.1).

City	Population	Area (km^2)	Population Density / km^2
Toronto	5,583,064	5,905	945.4
Vancouver	2,313,328	2,882	802.5
Moncton	138,644	2,406	57.6

Table 2.1: Statistics Canada 2011 data

The study areas shown in Figure 3 were chosen using rectangular bounding boxes required for

accessing data from APIs. The second grouping of study areas is population centres from the

2011 Canadian census, based on a total population of at least 100,000 of which 50,000 or more

live in the core (Figure 4). The population centres can be found in Figure 4 for the three cities.

These polygons are very different from the previous rectangular bounding boxes, including only

the areas with a high density of people allowing for a much higher distribution of VGI. The final

study area is a rectangular bounding box of downtown Toronto. It includes the University of

Toronto in the top left and the two major sporting event locations and the CN Tower in the

bottom.

Data was analysed using the open APIs for Flickr and Twitter, both through Python. In total, 1,541,170 tweets and 63176 photos were from Toronto, 398,811 tweets and 44,061 photos for Vancouver, and 25,533 tweets and 1571 photos for Moncton. The tweets were analysed over a four month period from September 2013 while the Flickr photos are all from 2010. This difference in time is an aspect taken into account as one of the purposes of this research is to view the point patterns of different data sets from different sources which might have different collection times or resolutions.

2.4.2 Case Study Results

Overall, results indicated higher metric values for CMAs compared to study areas based on bounding boxes (Figure 5). Coverage and user-heterogeneity of the bounding box study areas show a difference between the two larger cities compared to Moncton.   However, density shows similarity between all three cities.

The population centre study areas had very different results. Full coverage was found in the population centres at larger grain sizes while showing similar trends from the lowest grains. As anticipated, density varied between the larger cities compared to the smaller city, Moncton. Finally, user-heterogeneity maintained differentiation between the larger cities compared to Moncton. Exact metric component values can be found in Figures 6, 7, and 8. User-heterogeneity was never found to be higher than 0.34 (Toronto downtown core). While cursory analyses of only the Rogers Centre and of the University of Toronto did find higher values (0.56 and 0.43 respectively), user-heterogeneity was not found with high values within the Canadian urban city study areas.

The overall trends in all three types of study areas showed decreases in coverage and density as grain size decreased while user-heterogeneity increased as grain size decreased. The downtown Toronto core showed the highest metric values among the three study area types. Coverage is complete at all grain sizes and density and user-heterogeneity is much higher in the Toronto core. Figure 9 shows the metric values with all three components combined at different weightings for the random study areas and the population centres. When all three components are set at equal weighting, there is little differentiation between the three cities.

2.5   Discussion
2.5.1 Large vs. small cities

A distinct difference in metric values was found between Toronto/Vancouver and Moncton. This could be explained by socio-economic factors, as detailed in Li et al.'s conclusions (2013) where tweet density was correlated with well-educated people with advanced degrees, high income, while Flickr photo density was correlated with White and Asian people with advanced degrees. Moncton had an Asian population of less than 1% and the percent of people with advanced degrees (20.1%) was less than the national average (22.9%), Toronto (33.6%), and Vancouver (30.7%) as per the 2006 Canadian census. The physical size difference between Vancouver and Moncton is negligible, though population density is very different (Table 1) which may also be a factor in the variation. When using the population centres as opposed to the random bounding boxes, similar values can be seen.

The opposite can be seen for density as the population centres show a difference between big/small cities with Moncton having higher density values than the other two. The clustering in Figure 3 of Moncton follows the three major streets and the downtown and shopping areas.

The random bounding box created a similar density for all three as it is averaging the water/ remote areas.

User-heterogeneity shows similar results for all three cities. User-heterogeneity is not reliant on population density nor area with the number of people who tweet or post photos being the more important factor. This is similar to the results found by Mooney and Corcoran (2013a) in their study of OSM in London, Paris, and Berlin. Even though Berlin has a much lower population density than Paris, Paris showed less contributed data and only a third of the number of unique users. Further study can be done using different sized cities to see if there is a connection between user-heterogeneity in urban areas in contrast to remote areas or within different urban areas. One important aspect is that the study area did not affect user-heterogeneity, as both the metric values for the bounding box and the population centres were similar. This is due to the fact that while many low user/high data areas would be found in residential/ remote areas, the heavily traveled areas such as downtown cores or shopping centres would likely have high user/high data ratios.

2.5.2 Study area and grain size

The study area had a large effect on coverage and density. It also had a lesser effect on user-heterogeneity, especially with the use of the downtown Toronto core. This is not particularly surprising as correctly choosing a study area is a fairly important aspect of any study. The one point counter to this stems from the use of the population centre. An inset rectangular bounding box for the Toronto core still produced higher results in all three metric component values. While researchers may be looking for high or low values dependent upon the research at hand, the Toronto core produced results much closer to complete spatial randomness than the population centres derived from the authoritative data sets (i.e. Canadian census). This

leads to the idea that a well-chosen study area is superior to using authoritative-driven boundaries such as a census tract, and perhaps suggests data-driven methods for defining study areas is required for the types of VGI analyzed here.

Changes in grain size led to marked differences in the metric values calculated for each study area. When grain size is too large, it creates a grid of very few cells which compares to just using the study area as a whole. For all three cities, a 10,000 metre (m) grid cell size (10 km by 10 km) did not prove to be informative. The 2000 m grid cell size gave a stronger idea of what is happening within each study area and while the 500 m grid cell size showed much lower metric values when compared to a complete spatially random point pattern, you can start to see trends within certain sub-areas better. The downtown Toronto core (Figure 8), the 250 m grid cell size gives a good indication that the data are following either a certain street or possibly a subway line. The 100 m grid cell size doesn't quite give the same overall information, but gives a much more detailed view of the University of Toronto, Rogers Centre, and Air Canada Centre, all considered areas to be relatively high in VGI.

A key point found through both the simulations and the case studies is that there are grains where the metric starts to show less of an increase or decrease (inflection point), where the slope of the values to grain size becomes more level. These inflection points offer the ability to assess diminishing returns in the metric as grain size becomes larger or smaller, allowing for the selection of an optimal grain size. As with traditional data analysis such as identifying priority areas for biodiversity (Jenkins, 2013) or in the clustering of disease (Jeffery, 2014), finding the optimal grain is critical for analysis of point data. Figure 5 demonstrates that both coverage and user heterogeneity don't follow constant rates of increase/decrease, possibly indicating optimal grain sizes at 2000 m or 5000 m for coverage and 5000 m for user-heterogeneity. These grains

might indicate an optimal spatial grain which could be used by a researcher. While there are strong similarities between Toronto and Vancouver, Moncton tends to differ in overall metric values and sometimes even in trend. Moncton has approximately 140,000 inhabitants compared to roughly 2 million and 5 million found in Vancouver and Toronto respectively. One reason for this could be a limitation on the detectable patterns imposed by data volume of the smaller city. In addition, though Flickr is used for many purposes, there would be much more tourism found within Toronto and Vancouver, which could impact the overall spatial heterogeneity of the dataset, where tourist centres and landmarks in major cities capture a disproportionate share of the data. This is an interesting finding and may allow for the use of metrics such as these to help delineate qualitatively different types of VGI distributions linked to city size and function, though further study with more cities would be required.

2.6 Conclusions and Future Work

It has always been important to find a proper study area and grain size when using aggregated point data however VGI leaves researchers at the mercy of the data. Without control of the sampling procedures, researchers need to find ways to assess their study areas for data quality based upon the conditions of their studies. Current study areas tend to include authoritative delineations of areas (spatial grain) such as census tracts or census metropolitan areas (CMAs) which are built from authoritative data or historic boundaries that do not capture the transient nature of VGI. Strong increases in the three component values of the current metric are shown to happen as the study area is refined and grain size is modified from finer to coarser resolutions. This provides a basis for the creation of an algorithm using a metric similar to the metric described in this paper to find an optimal study area and spatial grain inherent to a pre-defined criteria. For example, such criteria could be realized by modifying the weights of the three components and optimizing an aggregated measure.

Other properties could potentially be incorporated into the methodology described here. The temporal characteristics of VGI, especially in the sense of user-heterogeneity of a study area, for example, could provide information about the transience of users based on the times they were observed within one area to gauge if they are resident to the area, traveling to the area, or commuting to the area. Similarly, semantic measures of text data could be examined using similar analytical methods. This paper focused solely on the spatial point patterns of VGI though further analysis using the text could prove useful for study area optimizations. Finally, the grids for defining units of analysis were based on a regular lattice, and alternate tessellations could be explored.

By definition, researchers have far less control over VGI compared to traditional data collection methods. However, given the increasing ubiquity of geographically referenced data, it is unlikely for VGI to become less used, and may become normalized into the toolkit of all researchers. Given that traditional research methods were defined on a paradigm of researcher control over the research design, there is cause for greater understanding of how VGI differs from traditional data, and what its value is in a given context. One of VGI's greatest benefits for the social sciences is the potential for a more nuanced, organic and evolving way to sense people (e.g. assessing emotion) and places (e.g., areas people take photos and how they characterize them). To realize this, baseline measures of data quality are required, and the analysis presented here is a first attempt at the creation of a relativistic assessment tool for these new types of data.

2.7 References

Baddeley, A. and Turner, R., 2005. Spatstat: an R package for analysing spatial point pattern. *Journal of Statistical Software*. 12, pp. 1-42.

Coleman, D., Georgiadou, Y., and Labonte, J., 2009. Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*. 4, pp. 332-358.

Croitoru, A., Crooks, A., Radzikowski, J., and Stefanidis, A., 2013. Geosocial gauge: A system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science* 27(12), pp. 2483-2508.

Crooks, A., Croitoru, A., Stefanidis, A., and Radzikowski, J., 2013. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS* 17(1), pp. 124-147.

Foody, G. et al., 2013. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Transactions in GIS*. 17(6), pp. 847-860.

Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, pp. 211–221.

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

Goodchild, M. and Janelle, D., 2010. Toward critical spatial thinking in the social sciences and humanities. *GeoJournal* 75, pp. 3–13.

Granell, C., Belmonte, O., and Diaz, L., 2014. Geospatial information infrastructures to address spatial needs in health: collaboration, challenges, and opportunities. *Future Generation Computer Systems* 31, pp. 213-222.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design* 37, pp. 682-703.

Hollenstein, L. and Purves, R., 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, pp. 21–48.

Jeffery, C., Ozonoff, A., and Pagano, M., 2014. The effect of spatial aggregation on performance when mapping a risk of disease. *International Journal of Health Geographics* 13(9).

Jenkins, C., Pimm, S., and Joppa, L., 2013. Global patterns of terrestrial vertebrate diversity and conservation. *PNAS* 110(28), pp. E2602-E2610.

Lacklan, K., Spence, P., and Lin, X., 2014. Expressions of risk awareness and concern through Twitter: On the utility of using the medium as an indication of audience needs. *Computers in Human Behavior* 35, pp. 554-559.

Li, L., Goodchild, M., and Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40, pp. 61-77.

McKenzie, G., Janowicz, K., and Adams, B., 2014. A weighted multi-attribute method for matching user-generated points of interest. *Cartography and Geographic Information Science* 41(2), pp. 125-137.
McKenzie, G. & Janowicz, K. Coerced Geographic Information: The Not-so-voluntary Side of User-generated Geo-content.

Mooney, P. and Corcoran, P., 2013a. Understanding the roles of communities in volunteered geographic information projects in *Progress in Location-Based Services*. Springer, Berlin Heidelberg, pp. 357-371.

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

Neis, P., Zielstra, D., and Zipf, A., 2013. Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future Internet* 5, pp. 282-300.

Openshaw, S., and Taylor, P., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Sciences.* Pion, London. 127-144.

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

Zielstra, D. & Zipf, A., 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. *13th AGILE International Conference on Geographic Information Science* 2010.

Zook, M., Graham, M., Shelton, T., and Gorman, S. 2012. Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy* 2, pp. 7–33.

2.8 Figures
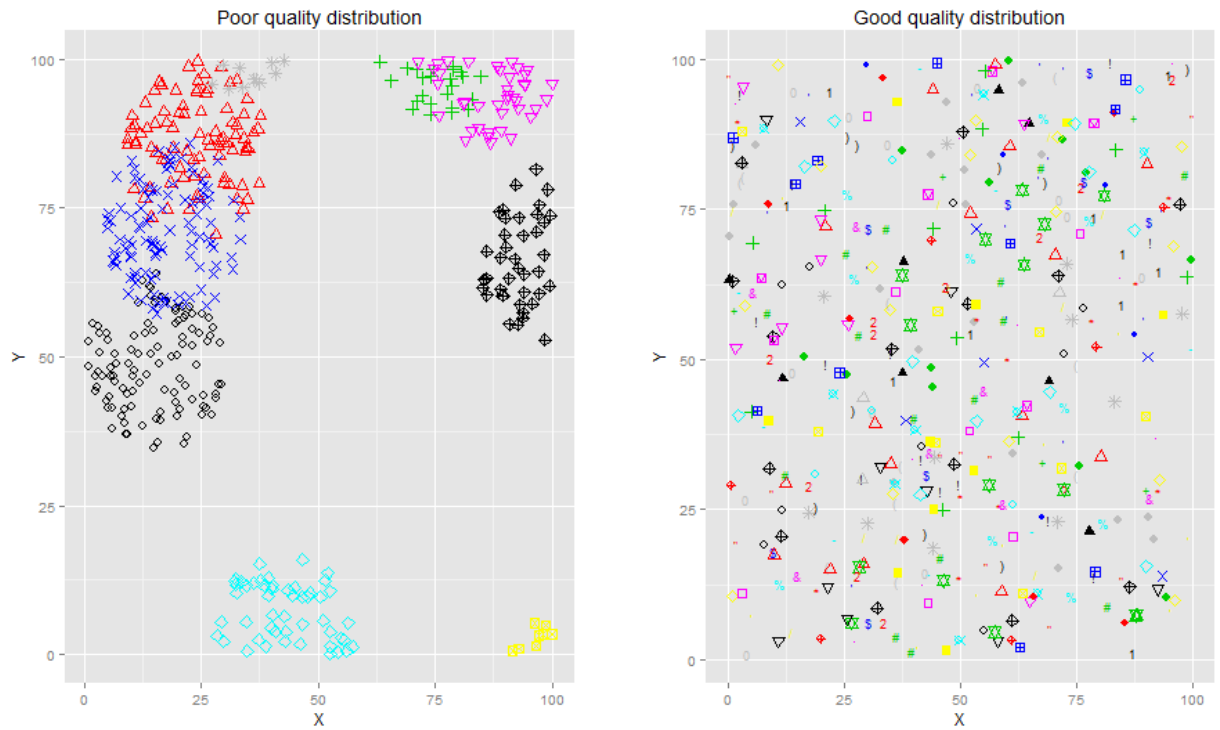
Figure 2.1: Matern clustered realization (left) and a uniform marked multitype Poisson realization (right). Distinct users are shown in different colours and symbols. These simulations were done using generalized parameters in the R programming language using the spatstat package (Baddeley, 2005). The purpose of these figures is to outline what is considered a poor quality distribution and a good quality distribution within the scope of this paper.
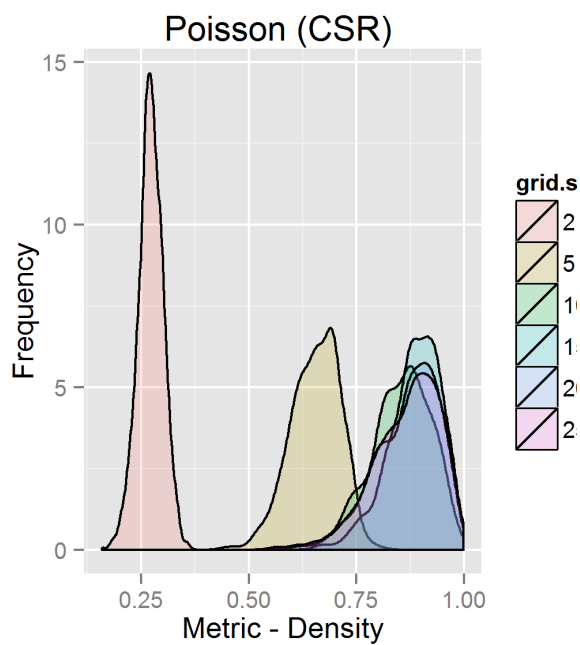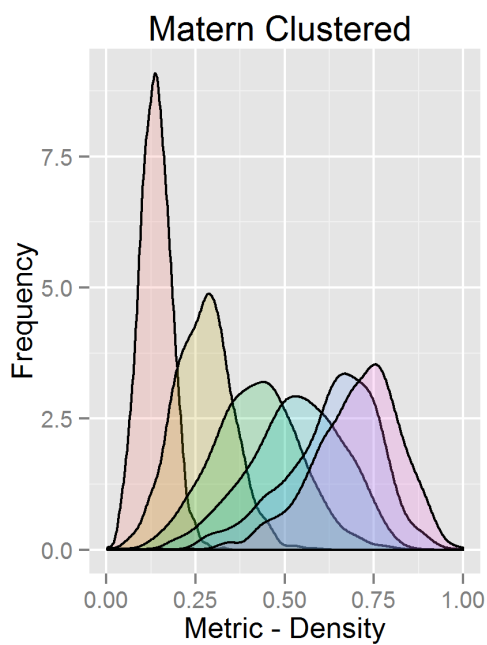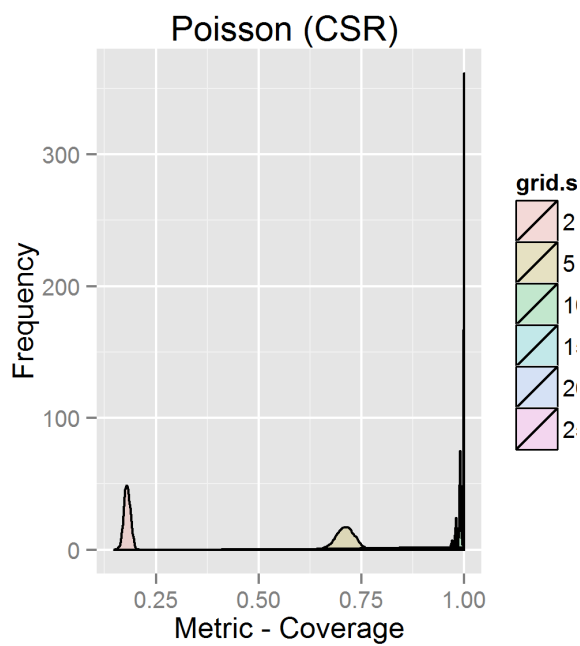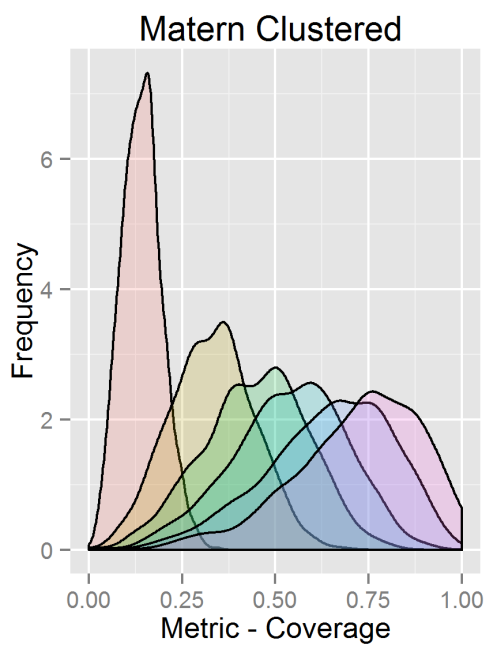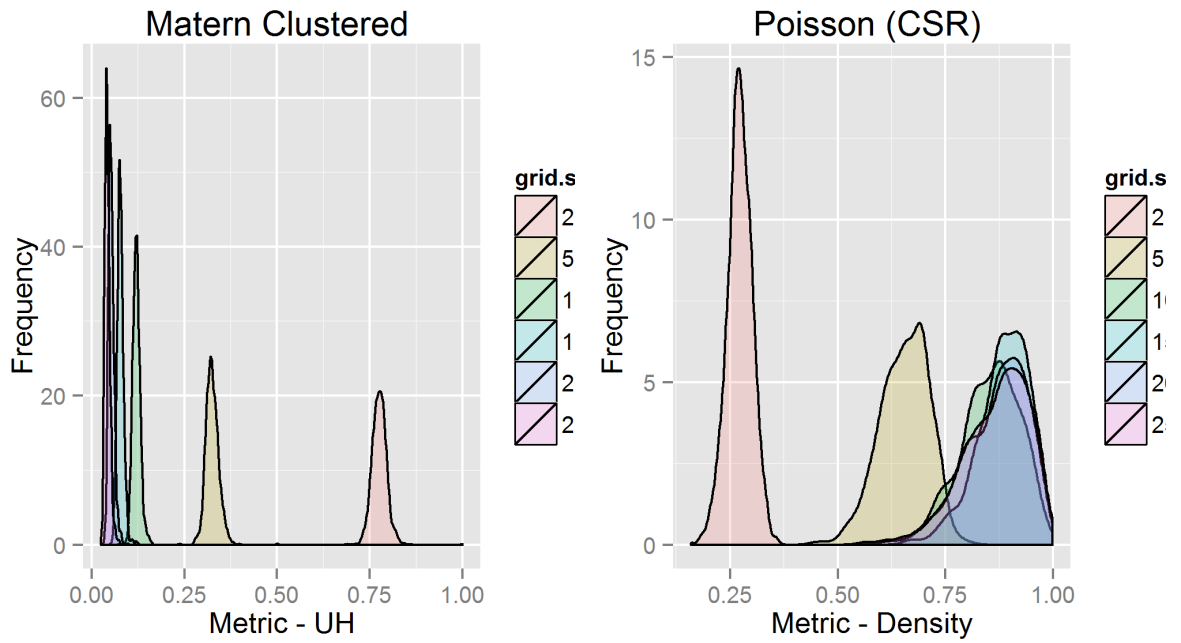
Figure 2.2: Coverage (top), density (centre), and user-heterogeneity (bottom) for clustered point pattern (left) and complete spatially random point pattern (right). These show the mean metric values over 999 iterations of the respective process at the six different grain sizes.
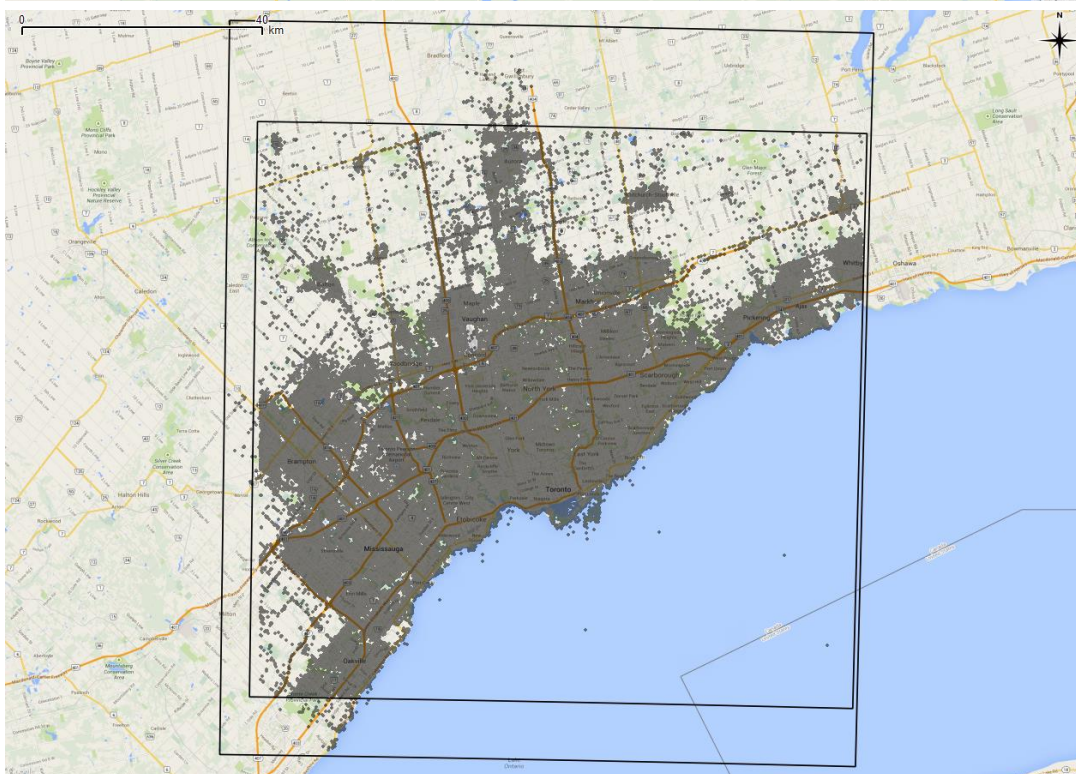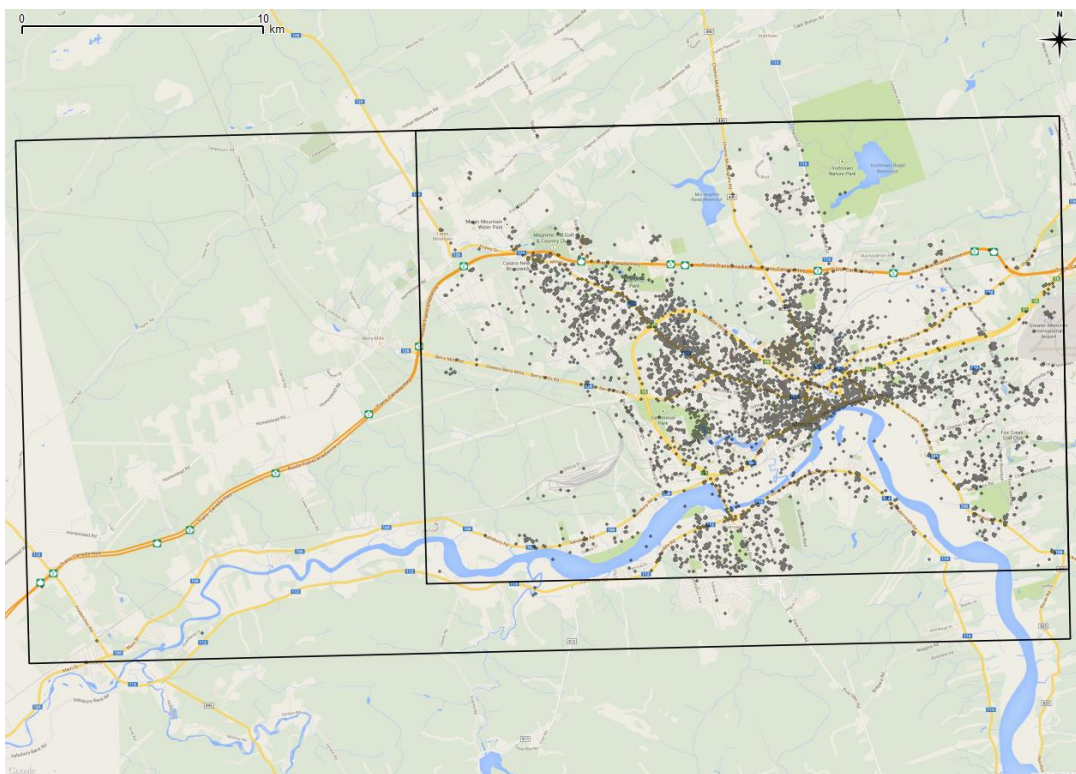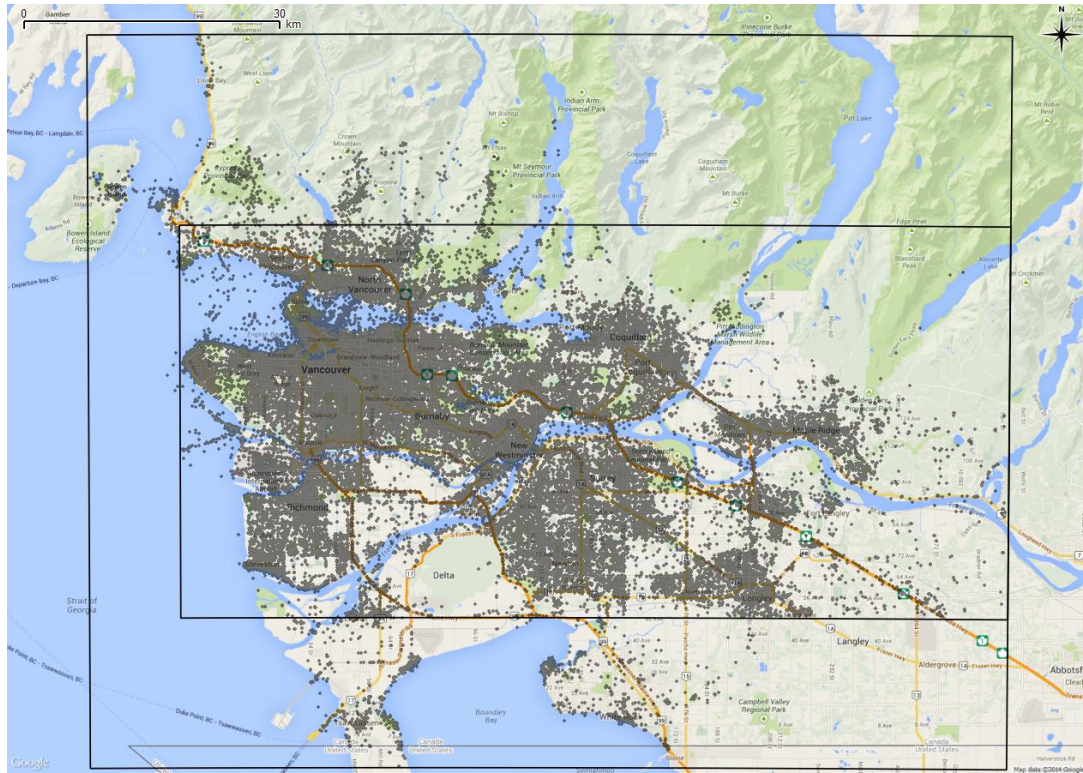
Figure 2.3: Moncton, NB (top), Toronto, ON (centre), Vancouver, BC (bottom). These are the data for both Flickr and Twitter. The inside bounding box contains both types of data while the outside box contains only Flickr data. This was done to test two different sources with two different spatial extents.
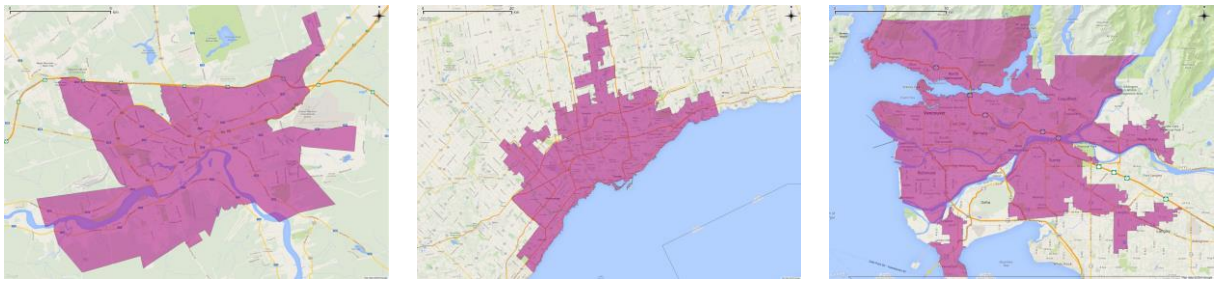
Figure 2.4: Moncton (left), Toronto (centre), and Vancouver (right) Canadian census population centres (CMAs).
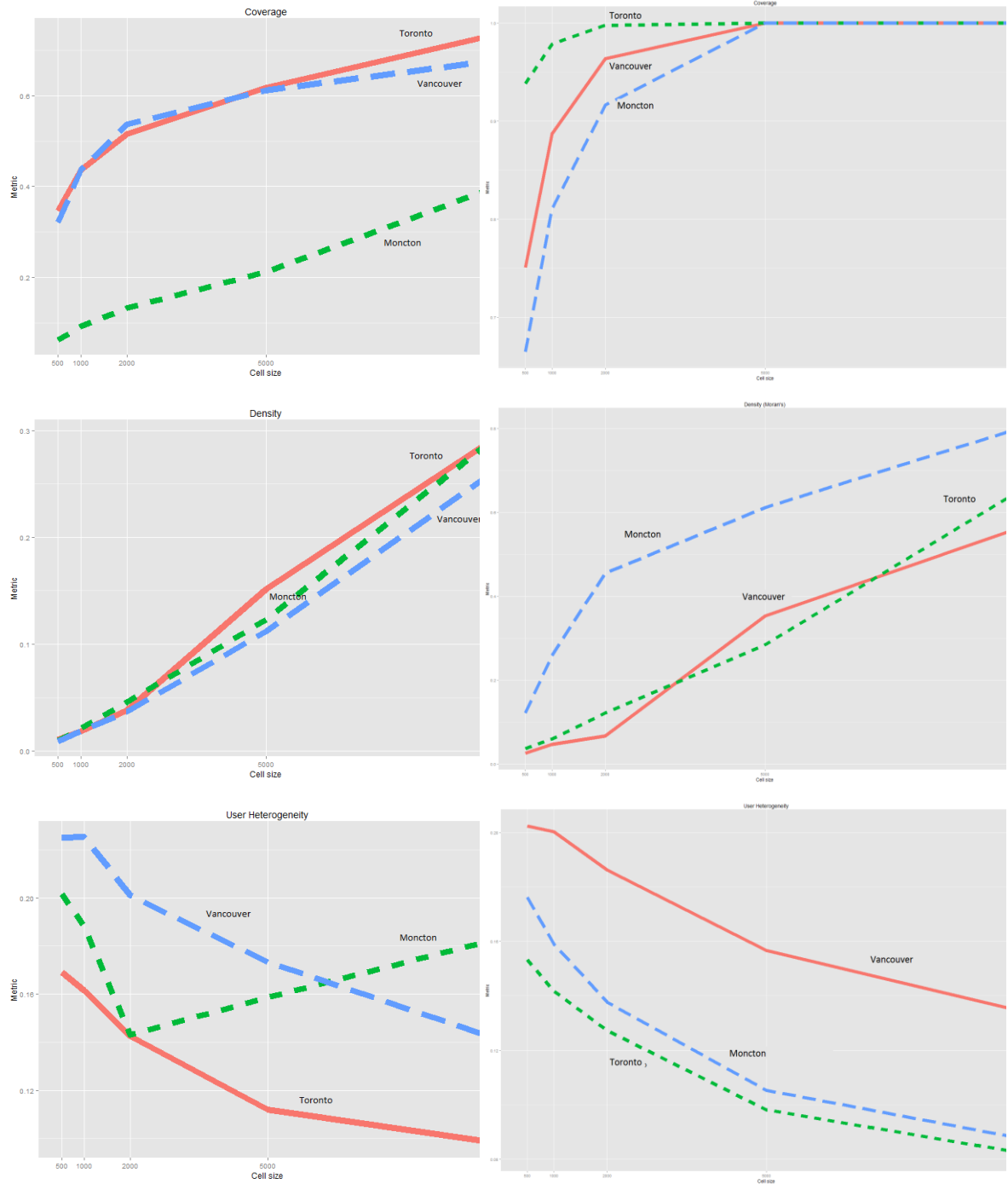
Figure 2.5: Metric component values for the randomly chosen bounding boxes (left) and census population centres (right) – coverage (top), density (centre), and user-heterogeneity (bottom):

Coverage	0.28	
Density	0.05	
User Heterogeneity	0.20	

Coverage	0.52	
Density	0.22	
User Heterogeneity	0.14	

Coverage	0.78	
Density	0.58	
User Heterogeneity	0.13	

Coverage	0.35	
Density	0.01	
User Heterogeneity	0.17	

Coverage	0.52	
Density	0.04	
User Heterogeneity	0.14	

Coverage	0.74	
Density	0.29	
User Heterogeneity	0.10	

Coverage	0.32	
Density	0.01	
User Heterogeneity	0.23	

Coverage	0.54	
Density	0.04	
User Heterogeneity	0.20	

Coverage	0.67	
Density	0.26	
User Heterogeneity	0.14	

Figure 2.6: Metric component values for the randomly chosen bounding boxes at 500 (left), 2000 (centre), and 10000 (right) grain sizes.

Coverage	0.66	Coverage	0.92	Coverage	1.00
Density	0.12	Density	0.46	Density	0.81
User Heterogeneity	0.18	User Heterogeneity	0.14	User Heterogeneity	0.09
Coverage	0.94	Coverage	0.99	Coverage	1.00
Density	0.04	Density	0.12	Density	0.67
User Heterogeneity	0.16	User Heterogeneity	0.13	User Heterogeneity	0.09
Coverage	0.75	Coverage	0.96	Coverage	1.00
Density	0.03	Density	0.07	Density	0.57
User Heterogeneity	0.20	User Heterogeneity	0.19	User Heterogeneity	0.13

Figure 2.7: Metric component values for the census population centres bounding boxes at 500 (left), 2000 (centre), and 10000 (right) grain sizes.
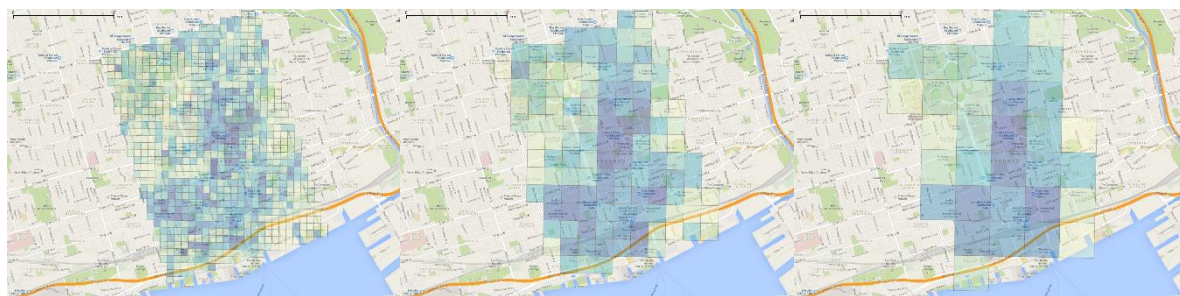
Coverage	0.998	Coverage	1.00	Coverage	1.00
Density	0.16	Density	0.51	Density	0.79
User Heterogeneity	0.34	User Heterogeneity	0.31	User Heterogeneity	0.27

Figure 2.8: Metric component values for the Toronto city core bounding boxes at 100 (left), 250 (centre), and 500 (right) grain sizes.
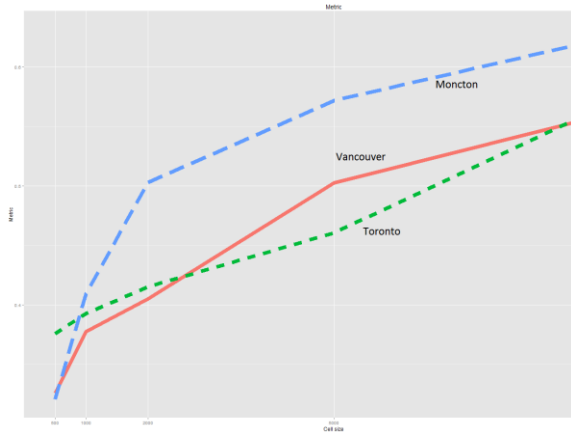
C:0.33,D:0.33,UH:0.33
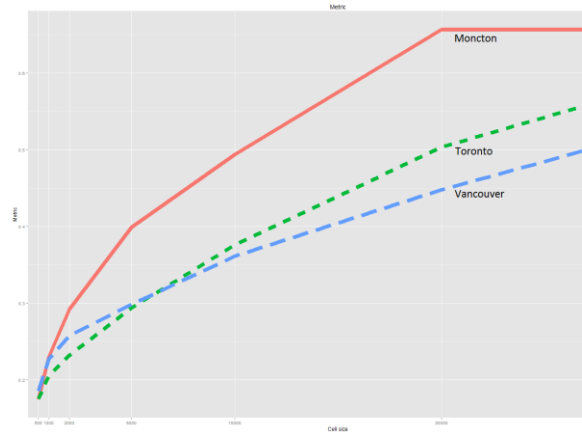
C:0.33,D:0.33,UH:0.33
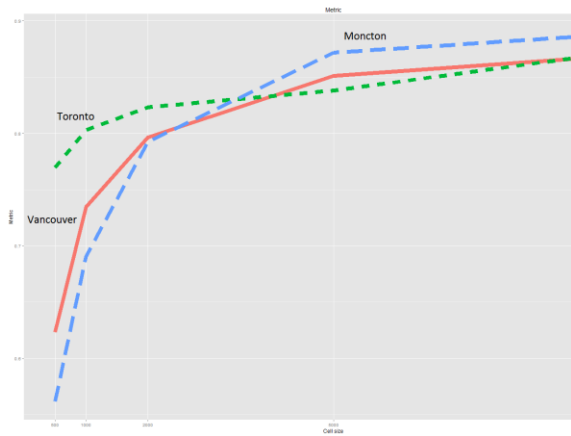
C:0.8,D:0.10,UH:0.10
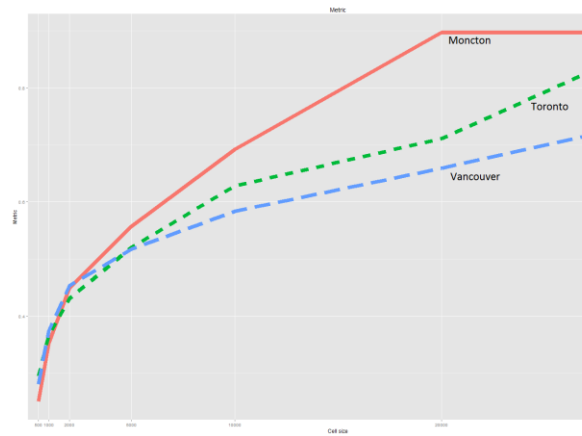
C:0.8,D:0.10,UH:0.10
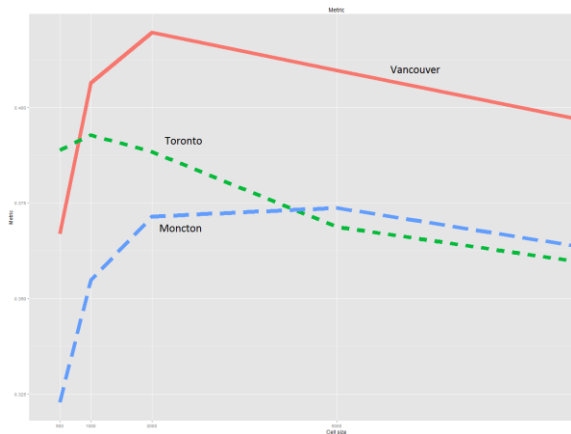
C:0.30,D:0.00,UH:0.70
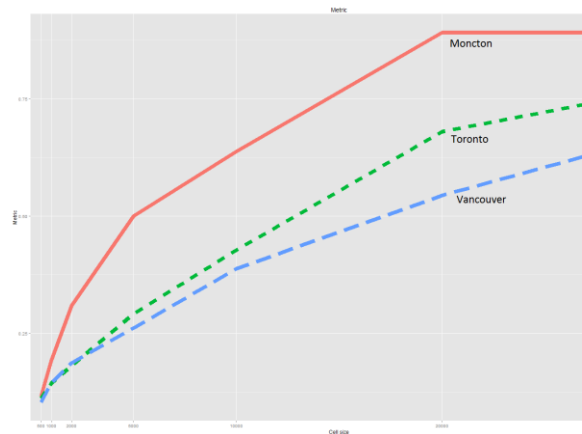
C:0.30,D:0.00,UH:0.70

Figure 2.9: Metric component values for the randomly chosen bounding boxes (left) and census population centres (right) by different metric weightings (coverage, density, and user-heterogeneity).

Chapter Three

3.0 Identifying optimal spatial extent of VGI data for analysis based on predefined quality metrics

Haydn Lawrence, Colin Robertson, Rob Feick

ABSTRACT:

Technology has become ubiquitous in our modern lives. As with most technology, social media use has become an ordinary facet of many peoples' everyday lives. Many of these data include geographically characterized data freely created by users (Volunteered Geographic Information, or VGI), whether through tweeting, emailing, or uploading pictures and videos. Flickr, Twitter, and Facebook are all sources of VGI in which the user may not be knowingly adding the data for public consumption (ambient VGI). Some data is created through more active participation in projects such as OpenStreetMap or citizen science initiatives (active VGI). Many of these data are accessible to the public and indeed researchers have started to harness these new types of data. However, these data are fairly chaotic when compared to the data researchers have been using previously, usually created by an authoritative source by experts in a field. In addition to these problems, VGI is often obtained with modifiable extents – ranging from a single neighbourhood to the entire Earth. With some forms of VGI, the relevant spatial scales may be unknown to a researcher, such as in newly gentrified areas. If researchers are to use these new forms of data, general methods are required to support the assessment of data quality of VGI datasets at several spatial scales. In this paper we focus on the impact of spatial scale in the analysis of VGI, specifically focusing on the optimal extent or study area based on a previously defined data quality metric for VGI. We present a scale-sensitive approach to the analysis of active VGI using a case study (RinkWatch project).

3.1 Introduction

The rise of "Web 2.0" technologies continue to transform society into a data-dependent one, where academic research, commerce, and governmental processes increasingly rely on data produced in full or in part by contributions of anonymous citizens. Web technologies that depend on user-contributions and feedback (YouTube, Facebook, Twitter, Flickr, FourSquare, Wikipedia, blogs, etc,) are being repurposed for a variety of applications including some that include geographical information. The user-generated content (UGC) created from these applications may have hidden scientific, social or cultural value, and tools and methods are being developed to unlock these values through machine learning, data mining and computational methods (Li et al, 2012; Elwood, 2011;Goetz & Zipf, 2012). However, there is little consensus on how to deal with the greater uncertainties inherent in data produced by non-experts. In the geographical literature, research into data quality evaluation for VGI has been at the forefront of VGI research (Barron et al, 2013; Girres & Touya, 2010; Mooney et al, 2010). As analysis of VGI has become increasingly common (Neis and Zielstra, 2014; Jeffery, 2014; McKenzie et al, 2014; Hollenstein, 2010), research has started to coalesce around the importance of spatial scale in the analysis of certain types of VGI (e.g., Feick and Robertson, 2014). Should data be aggregated, and if so, where should the focus of analysis be? Can we continue to use traditional units of aggregation, such as census tracts, or should new approaches be considered? What repurposing of VGI is suitable for what types of research questions, and how can we assess the impact of data quality parameters on research outputs? In this paper we focus on the impact of spatial scale in the analysis of VGI, specifically focusing on the optimal extent or study area based on a previously defined data quality metric for VGI (Lawrence et al, 2014). We present a scale-sensitive approach to the analysis of active VGI using a case study (RinkWatch project).

3.2 VGI Data Quality

Goodchild first coined the subset of UGC, which include geographic information, as Volunteered

Geographic Information (VGI) (Goodchild, 2007). The idea behind Goodchild's optimistic view of

VGI's potential was that there were over 7 billion potential citizen sensors, and that similar to

the success of open source software development and projects such as Wikipedia, any errors or

biases in data produced by these contributors would, over time, be averaged out (i.e., Linus's

Law). However, the reality of VGI has been more complicated. This can be easily seen in a

cursory view of one of the largest VGI initiatives, OpenStreetMap (OSM). While OSM has been

the subject of numerous data quality assessments and results have demonstrated that data can

be of high quality relative to reference data (Haklay 2010), there have also been demonstrations

of biases relative to demographic and socioeconomic class. In geosocial data such as geotagged

tweets or photos, biases have been found for urban areas relative to rural (Li et al, 2013; Hecht

and Stephens, 2014). There may be patterns in the type of data quality issues and the form of

VGI under consideration. VGI can be thought of as either ambient in nature (Stefanidis, 2013; Li,

2013), consisting of data created without any geographic purpose in mind but nevertheless

associated with a geographic footprint (e.g. a geo-tweet), or actively created with an intentional

geographic component, such as in OpenStreeMap (OSM) or various citizen science initiatives. In

this paper, we focus on active VGI through a citizen science case study.

Data quality has always been an important issue concerning spatial data and GIS. With VGI, the

possibilities for data quality variability is exacerbated by the researcher's lack of control over the

sampling design and a lack of standards and training on the part of contributors. One of the

most widely studied examples of active VGI is still that of OSM. The first studies were on the

veracity/quality of its data in comparison to authoritative datasets including Haklay's study on

OSM to the Ordnance Survey datasets of London (2010), OSM in Germany from 2007 to 2011

(Neis, 2012), and OSM in France (Girres, 2010). These studies showed that OSM had high quality in comparison to ground-truthed datasets, though only for urban areas and even within urban areas, poorer neighbourhoods were lacking in data. Dickenson (2010) pointed out several aspects of VGI data, specifically citizen science in this case, that can be problematic: trained vs. amateur data collectors, age of users, and length of the project, where the first year of a project may acquire many amateurs but over subsequent years, these users will become better at the task required. The '90-9-1' rule is another problem found in crowdsourced data, including VGI (Neis, 2002). The 90-9-1 rule states that 90% of the users contributing information do so a few times at most, while 9% contribute intermittently and in different ways. Only 1% of the users are considered active users who contribute data on a regular basis and through many different avenues in regards to active VGI. This leads directly to the notion articulated by Haklay, that analysis of VGI is really analysis of outliers rather than typical contributors (Hacklay 2013).

Irrespective of the these issues of VGI's quality, it has become a popular method of collecting data (Goodchild, 2010; Wiersma, 2010; Freifeld, 2010) in part due to allowing the general public to take on the burdens of normal sampling and data collection (e.g. time, financial, or manpower costs). Elwood et al. (2012) used the example of a project which attempted to spatially differentiate linguistic trends via the use of "pop" vs. "soda" for carbonated soft drinks among contributors through an active VGI website specifically created to ask respondents to select the term they use for carbonated soft drinks. This project asked people to enter their zip code/postal code and which term they use which is then mapped online. The issues arose with the sampling required by rigorous studies, as this did not provide a random sampling nor did it have any information about the characteristics of the sample population. The authors acknowledged this but do not find it particularly debilitating. The reasoning was that surveys have many stages and this kind of project would at least be very useful in the selection of study

sites or the formulation of hypotheses. As an example, they stated that significant research time and funding could be saved with just one Flickr.com image showing some form of problem in an area that would have been sampled.

A further issue with VGI was studied by Coleman (2009) who looked at the types of users creating these data and their motivations while Li et al. (2013) showed that georeferenced Tweets were over-represented by mostly rich and educated users. Haklay (2010) found similar results in the fact that poorer neighbourhoods were underrepresented by OSM data in London. This inherently leads to questions about the general quality of the data, especially as it pertains to sampling bias. Unfortunately, VGI does not allow much control over the sampling process for a study, even with planned active VGI initiatives, such as NatureWatch (www.naturewatch.ca). Dickenson (2010) again lists numerous citizen science initiatives where sampling bias formed, both temporally (Royle, 2005; 2007) with the standardization of sampling effort, and spatially (Betts, 2007; Niemuth, 2007) looking at roadside surveys and their representativeness of the surrounding areas. Several other studies of the sampling of VGI include Goodchild and Li's study of different approaches to quality assurance (2012), Foody (2013) demonstration of issues of representativeness for land cover interpretations from volunteers, and Dickenson's (2010) notes showing the same issue in ecological citizen science.

The spatial distribution of VGI can be highly heterogeneous. Haklay (2010), Zielstra (2010), and Neis (2012) all describe areas of low coverage in their study areas of OSM. The distribution of data in these studies tended to reflect boundaries delineated by socio-economic factors and the urban-rural divide.   When researchers would like to study an area using a VGI dataset (such as Twitter, OSM, or a specifically created initiative), they may not have the time or ability to do a full analysis of the data to gauge its potential usefulness to their research. One of the only

levers of control accessible to a researcher employing VGI for a study is the decision of where to define the study area boundaries, which given the above discussion, may capture varying degrees of spatial heterogeneity in submissions.

In addition to deciding on a study area extent, when analyzing active VGI point datasets, researchers have to decide if they will aggregate the points to a certain scale (unit of aggregation). There are many reasons for aggregation, including being able to analyse the data with authoritative datasets (e.g. census data) (Clark, 2013) or to protect privacy, especially in instances of health or medical data (Jeffery, 2009).

The importance of spatial scale in geographical analysis has a long history. Wiens (1989) describes the relationship between grain size and internal versus external variance, and the importance of choosing an aggregation size that effectively captures the process being investigated. The choice of grain size can be affected by many factors beyond process factors, such as previous studies which researchers wish to compare findings (Calderón-Patrón, 2013; Kelly, 2011), where researchers are bound by standard spatial grains within a discipline (Galpern, 2012), or by the technology available such as spatial resolution of satellite imagery. In addition, spatial grain can depend on the scale of auxiliary dataset required for a study such as census tracts for socio-economic comparisons. In some cases, spatial grain can even be arbitrarily chosen (Chase, 2013).

Researchers are now starting to view multi-scale approaches to analysis as beneficial (Kelly, 2011; Galpern, 2012) to absolutely essential (Calderón-Patrón, 2013). Li et al. (2013) examined socio-economic information associated with ambient VGI users in California by using counties instead of census tracts as people tend to work and live in the same county though not necessarily the same census tracts. However, any results deriving from the analysis of that data

would reflect only a subset of the population. Lawrence et al. (2014) iterated over square tessellations at different grain sizes to study how the grain size would affect a predefined metric over several defined extents (Toronto, Vancouver, and Montreal). This paper showed that grain size can be quantifiably delineated at points where the metric becomes optimal and therefore useful for further analysis. Feick and Robertson (2014) used a hexagonal tessellation at various sizes to help elucidate cross-scale patterns of spatial agreement and disagreement in Flickr tags in Vancouver, Canada.

How does the inherent heterogeneity within a VGI dataset affect they study area's outcomes (90-9-1 rule)? How can we address the differences in users and user submissions in relation to active VGI? If a study area is chosen with non-uniform coverage, how dispersed or clustered can the data be for an analysis, and should we rather focus on a smaller sub-area? We have seen from Lawrence et al. (2014), that the choice of grain for analysis can be optimized with respect to the general data quality parameters for VGI; density, coverage and user heterogeneity. Here, we explore how extent can be optimized based on a predefined metric consistent with the study's overlying goals. As VGI becomes more prominent in research, tools developed to aide researchers in quickly gauging optimal grains and extents will be needed. This paper details different approaches to finding the optimal extent of an active VGI study area using the case study of RinkWatch (www.rinkwatch.org) – see Appendix for complete description of the RinkWatch project. It uses a predefined metric (Lawrence et al. 2014) to create a benchmark for an optimal extent for the study area, which is also dynamically changed using gridded tessellations, authoritative polygonal boundaries, and geometric Voronoi polygonal methods.

3.3 Methods

The overall optimization of study areas was completed using two different algorithms, accounting for both regular and irregular types of tessellations. The regular, or gridded, tessellations used a quad-tree approach while irregular tessellations (authoritative datasets, such as census tracts and dissemination areas, and Voronoi tessellations) used a recursive adjacency algorithm. Both algorithms use a predefined metric to assess optimal areas which returns a value between 0 and 1.

The metric constitutes three different components: coverage, density, and user-heterogeneity (Lawrence, 2014). To determine the metric value, individual data points are aggregated within the specified type of area (grid or polygon). Coverage assesses if the area under question, either a quad-tree node or an irregular polygon, contains a minimum threshold of data points, and is given as the percentage of cells in a candidate study area that are 'covered' by the minimum number of points. General approaches to a minimum threshold could be statistical in nature (e.g. examining the distribution through a histogram or excluding a certain percentage from each tail of the distribution). The minimum threshold could also be data-driven, as census data requires a certain count to maintain anonymity or based on natural factors such as a study of viable rinks located only within appropriate climates. There was no maximum threshold applied as this would be a function of the optimization. The purpose of coverage is to identify areas with appropriate amounts of data for analysis and to exclude potentially ineligible sub-areas such as lakes, protected areas, or institutions (e.g., military base) unsuitable for analysis.

User-heterogeneity measures the ratio of data submissions to the number of unique contributors in an area. In terms of OSM, this would be a ratio of the number of edits to the number of unique users in a given geographical area. The intuition behind this formulation is

that where edits are distributed over a large number of contributors, there may be broader consensus on the representations than if contributed by a single user. In the context of thematic analysis of geosocial data, high user-heterogeneity may indicate more certainty in an identified pattern or trend (Linus's Law).

The final component is a density measure which uses the data space, or similarity among the aggregated data counts, within a Moran's Scatterplot to estimate how evenly distributed values in a study area are by evaluating the variance in the quadrants of the scatterplot defined by the measured values on the X-axis and the spatially lagged values on the Y-axis.   Moran's I is not used as it would average the outlying data counts (x values) into the statistic, which in the context of analysis of VGI (particularly geosocial data), extreme outliers may be common (e.g., a landmark or downtown core captured by one cell on the map). Density, when paired with the adjacency aspects of the simulations mentioned later, finds areas of similarity both within spatial space, or actual geographic space, and within data space (Figure 1 for a general illustration). Density is thus formulated such that values near 1 indicate evenness in the Moran's scatterplot quadrants, and values near zero indicate high variance in the Moran's scatterplot quadrants.

There were three types of aggregation methods applied to the RinkWatch data: grids (quadtrees), authoritative datasets (census tracts and dissemination areas), and geometric (Voronoi polygons). The quadtree gridded approach was chosen for its ability to create variable resolutions recursively, while the authoritative datasets were used for later association with ancillary census data for analysis. The final aggregation method, Voronoi polygons, were used for their geometric properties as the RinkWatch data, as stated earlier, has individual locations

that tend not to overlap (e.g. there would probably not be three backyard rinks on the same street).

Optimizations for regular tessellations were performed using quad trees. A quad tree is a data structure where each level has four spatially identical nodes branching from a parent node (Bereuter, 2013). Bereuter explains that this type of data structure can speed up spatial queries by continuously subdividing a homogenous study area until heterogeneous data values are found. Examples of the use of quadtrees can be found in image retrieval (Ramanathan, 2011) and environmental modelling (Popinet, 2011). The quad-tree implementation was recursive, moving down the tree until all four child nodes' metric values were less than the parent node's metric value plus a small recursion tolerance (0.05 in the case of this implementation, where the metric value is between 0 and 1). For all the child nodes, if a child node's value was greater than the parent's node metric value plus the buffer, the recursion would move to a level deeper. When the algorithm reached the bottom, the current grid area and metric value would be passed up the tree and merged with each of the other nodes that matched the abovementioned criteria, determining if the new grid area creates a greater metric value whereupon the algorithm would send up the newly merged grid area. Grid areas must be adjacent to be merged and would be discontinued if found to be an island.   At the end of the algorithm's run, the area with the largest area and highest metric would be identified as the optimal area (Figure 2).

Optimization for irregular tessellations used an algorithm which iteratively merged adjacent polygons from a seed polygon. The algorithm iterates over every polygon present in the study area, treating each as a seed. The seed polygon's neighbours are evaluated to find the neighbouring polygon that would create the highest overall metric value of the two combined.

If this metric was higher than the metric of the current seed, the seed was replaced with the new merged polygon. This process would be repeated until no adjacent polygons were found to create a higher metric value. Adjacency was determined non-recursively using a global adjacency matrix similar to the AMEOBA algorithm (Valles, 2014). First and second order adjacency were used for analyzing the dataset (Figure 3).

All data points within a polygon were aggregated for metric value determination with the metric normalized by polygon area. The polygon delineations used were census tracts, dissemination areas, and Voronoi polygons. This paper set a lower limit of 1000 m^2 for the threshold of gridded areas. Metric component weightings for analysis were varied: 50% density and coverage, and 33% density, coverage, and user-heterogeneity.   Hereafter, these will be referred to as 505 weighting and 333 weighting respectively.   The 505 weighting was used for the analysis of the active VGI dataset case study as user-heterogeneity was considered a less useful component. This is due to the fact that user heterogeneity measures the ratio of data points to unique users and the case study will use a minimum of 20 data points per user for inclusion, causing a small user-heterogeneity metric component value.

3.4 Case Study

Several citizen science initiatives are being created to harness VGI, including RinkWatch (www.RinkWatch.org), NatureWatch (www.naturewatch.ca), and a site to generate data on the distribution of plants and animals (Lukyanenko, 2011). The RinkWatch (www.RinkWatch.org) project was designed to exploit the popularity of outdoor skating by recruiting citizens to contribute to a web-based log about the quality of ice on their homemade rinks. It was anticipated that the change over time in the length of the outdoor skating season and intra-seasonal changes in skating conditions could be used as proxy indicators of winter weather

trends, complementing weather station observations. More importantly, observed changes in the indicator activity would likely trigger greater public response than would observing changes in weather station readings, especially as many weather stations are situated at airports which tend to be outside of city limits. RinkWatch was officially launched on January 8th, 2013 and ran in the winters of 2013 and 2014.Users were asked to register and add daily readings of whether their rinks were "skateable" or "not skateable". Current research (Damyanov et al., 2013) shows that rinks tend to be skateable after 3 days of consistent temperatures of -5 degrees Celsius. For the analysis, skateability is defined as the percentage of days that the users recorded that their rinks were skateable over the season (with skateable being a binary yes/no variable).

The case study of evaluating RinkWatch data to identify optimal areas was focused on the Kitchener/Waterloo (KW) area of southern Ontario, Canada. The RinkWatch data is slightly unique in comparison to other implementations of active VGI like OSM. OSM may still have several users contributing data on the same point (i.e. the same location) while RinkWatch would more likely be one user per geographic location (e.g. the person in charge of the rink or homeowner). The KW area has an estimated population of 317,933 and an average winter temperature of -7.7 degrees Celsius over the last three years (statscan). This area was chosen due to the high participation rates with the RinkWatch initiative in addition to the familiarity of the area to researchers. 985 rinks (unique users) were created over the two year study period, with 22661 individual readings (data points) specified. There are 21 rinks in the KW area that will be used for the skateability analysis, each with 20 or more readings (Figure 4). The goal is to find a study area within the KW area that would allow an optimized analysis of local temperatures for the winter season, more specifically ignoring outliers (both high and low density areas of readings) while defining an area of adjacent (i.e., connected) cells within the study area. The minimum threshold for aggregation of points by a unique rink for this study was

set at 20, or approximately three weeks of continuous participation. Final analysis of the optimal study area found will be done using a spline interpolation of skateability on the entire study area and the optimal area. The smoothness of the optimal area will be compared to the entire area to gauge the suitability and usability of the optimal area.

3.5 Results

Measures of spatial pattern inherent to the metric were able to determine an optimal geographic area for subsequent analysis of a VGI point dataset. Through use of the RinkWatch active VGI dataset, it was shown that the optimization algorithms, using the metric, discovered suitable sub-areas of study when the focus of skateability analysis was the goal. In addition, there were several differences between the regular tessellations when compared to the irregular tessellations, with the irregular tessellations showing higher metric values and more homogenous data values compared to the grid areas which incorporate several extreme outliers. The regular tessellation algorithm returned optimal areas with very low metric scores (0.148/0.216 for the 333/505 metric inputs) compared to both the irregular tessellation (census tracts) and geometric tessellation (Voronio polygons), both of which showed very similar values (0.656/0.68 and 0.968/0.9997) (Figures 5 and 6).

The gridded area (Figures 5a and 6a) found an area of like values [22, 30, 25, 27, 67, 76] though this same area contains much larger values [158, 394, 179] which do not meet the criteria needed for the skateability analysis described earlier. The metric value is also very low for both the 505 and 333 weightings. Coverage would play a large role in the lower values as the grid area is rectangular by design leaving a lot of empty space. In addition to this, the addition of the large outliers to the list of values would decrease user heterogeneity.

The irregular polygon analyses found higher metric values overall (Figures 5b, 5c, 6b, 6c). The

census tract analysis identified four polygons with similar numbers of readings [20, 22, 65, 67].

The individual polygonal metrics were normalized by the area of the polygons. The Voronoi

polygon analysis identified the same region [20, 22, 27]. User heterogeneity decreased the

metric values of both of these analyses as each aggregated point has just one unique user (rink).

The final analysis was using a lag order of 2 (Figure 7), or allowing a polygon that is connected by

one empty polygon. This showed no real difference with a weighting of 333 but the 505

weighting found a different final optimal area [20, 22, 27, 35, 76]. Dissemination areas were

tested but proved to be a too fine for an analysis of a point dataset of this number.

The skateability of each rink can be seen in Figure 8. These percentages represent the amount

of time the rinks were skateable over a variable length period dependent upon user

contributions.   A cursory analysis of the skateability percentages does not yield any definite

areas of local skateability variation nor do comparisons of the skateability (Figure 8) show a

relationship to the overall counts of data (Figure 4) (e.g. high skateability does not necessarily

mean high number of readings). High counts of readings and of skateability can be found both

in the central areas of the area and in the more suburban fringes.

Overall, the spline interpolation analysis shows a much smoother spline surface created than

the surface created from the entire dataset (number of readings >= 20) (Figures 9 and 10). This

allows us to better gauge the variability in the local area for skateability. While local variation

would show differences from the official weather data stations (located at the airport in the KW

region) and among different areas of the KW region (e.g. rural vs. urban), the variation shown

using the entire dataset for the study areas is unusable (left images of Figures 9b and 10b).

Though there are many levels of users, from people who have been building rinks for decades to

people learning for the first time, there should not be the amount of variability in the skateability of rinks as shown in the left images. The optimal area using both weightings (505 and 333) demonstrate much smoother surfaces than the overall area with similar values of skateability. The major outliers were not included in the optimal area. Figure 9 shows a 505 weighting with the [77.3, 100] values not included. These are too high and probably denote low numbers of readings and therefore unreliable data. The 333 weighting (Figure 10) shows similar results, with large areas of [0] denoted by the -596% - 0% and too many areas of large areas of [100] denoted by the 101% - 231%. It should be noted that a spline creates the best surface based on fitting a smooth surface to the data, causing the values under 0% and over 100%.

3.6 Discussion

Lawrence et al. (2014) showed that analysis of ambient VGI (Flickr and Twitter) at different grains can result in very different outcomes. His paper focused on utilizing a metric created to assess the quality of VGI at various spatial grain sizes. For this paper, through the use of several types of optimization methods, we aimed to use the same data quality metric to define an optimal study area, or extent, from the complete input study area based on the spatial characteristics inherent in general active VGI point patterns. The results showed that the optimization methods were able to find optimal study areas that matched the predefined criteria for the active VGI case study of skateability of RinkWatch: similarity in data counts by each unique contributor, spatial adjacency, and a minimum amount of data counts. The analysis of the results of the Voronoi polygon optimal area to the Voronoi polygon complete data area showed a much more realistic pattern of skateability than the same analysis conducted on the entire data set. Filtering out outliers and preserving spatial connectedness in spatial analysis of VGI point patterns may be an important pre-processing step for future studies of user-contributed data.

The findings of metric values showed that the differences between different spatial extents can be quite diverse. The first optimization method used was the quadtree gridded algorithm. While the optimal area result encapsulated a reasonable subset of the overall data matching the criteria needed for a skateability analysis, it could not overcome the inherent deficiency of standard grids being used. The optimal area found for the RinkWatch data contained several aggregated counts of readings that were much higher than the surrounding counts. This would bias the skateability analysis as a rink with 394 contributions would be much more accurate than the average rink with 73 contributions. While there could be uses for finding the most prolific contributors, such as an analysis to assess socio-economic factors that may explain strong participation rates, this was not the focus of this paper. Often, multi-scale approaches to an analysis may be constrained to a gridded approach, especially in cases of comparison to raster data, though there are efforts to combat this such as Galpern's study of landscape gene flow at multiple spatial scales using landscape connectivity (2012).

The irregular polygon study areas (census tracts, dissemination areas, and Voronoi polygons) were then tested and found to mostly provide higher metric value optimal areas based on the predetermined criteria mentioned above. The dissemination area, however, was too fine grained to be useful for the necessary analysis. While multiple lag orders were attempted (i.e. allowing for adjacency of n orders, or allowing polygons to be included even if they were separated by n empty polygons), the dissemination areas did not provide any useful results which are sparsely populated over an area. There would be very little use in creating several rinks in the same location in one neighbourhood, especially at the distance required to be adjacent using the dissemination areas. The census tract and Voronoi polygon optimal areas were much more conducive to a skateability analysis.

The census tract and Voronoi optimal areas proved to capture the best areas of all the different types of study areas tested, showing good compatibility with the criteria needed for the final skateability analysis. There was very little difference between using a metric component weighting of 333 compared to the 505 weighting. The 333 weighting was used to see if user-heterogeneity would change the outcomes of the optimizations, which it did not. This is not surprising as user-heterogeneity would always be very low due to the fixed location of the active VGI in question (fixed physical rink location). A lag order of two did change the optimal area for the 505 weighting, shown in Figure 7. The overall characteristics of this area are similar to the ones resultant from a lag order of one in average counts and number of polygons when normalized by area (Figure 6).

The skateability analysis using spline interpolation helps to visualize the difference between using the complete study area versus the optimal Voronoi area. Figures 9 and 10 show that skateability within the extent shown is not as high as the complete set of points would have us believe. The value of 100% (bottom right of Figure 9) is a highly suspect value which should be further analysed and most likely omitted from the analysis. These spline surfaces help to address the question about local weather variation within the KW region. We are looking for areas which are similar to each other yet give enough information, in this case having a minimum of 20 readings. We can then compare these (optimal) areas to each other and the official weather station data for further analysis in the future.

The optimization proved to be highly beneficial to a cursory skateability analysis of the RinkWatch data in the case study. Optimization of grain could also pose problems with further study though. This paper used grid size iterations of 500 m^2 to check for the optimal grain size to reduce computation complexity with 1000 m^2 proving to be the optimal grain, though smaller

blocks could provide higher metric values (iterations of 10 m^2 for example). While possibly

providing a more optimal extent, this could create problems in future comparisons to other

studies or within different geographic areas of the same study. Also, in general, it is easier to

understand a gridded tessellation of 1000 m^2 as opposed to less specific numbers such as 670

m^2. However, studies such as Chaudhry and Mackaness (2012), in which the grid sizes of 500,

1000, 2000, and 4000 m^2 were arbitrarily chosen for their study of Flickr tag similarity in the city

of Edinburgh, would benefit from optimization.

While studies such as this may benefit from an optimization of study area extent and grain size,

researchers should always base their study area on the theoretical criteria they are looking for

and the natural characteristics of the dataset. Along these lines, how data quality is defined is

determined by the researchers conducting the study. This paper was not looking at user-

heterogeneity (neither high nor low) as there was only a necessity to have a lower minimum to

the number of readings each unique contributor had created. A further study of user

participation may require the inclusion of user-heterogeneity looking at user participation. This

further addresses the need for tools to assess a study area for VGI, especially in cases where the

researchers have limited access to data (e.g. Twitter's 1% stream), based on data quality

measures at different grains and extents.

The need for methods to optimize study area extents and grains can be seen in other VGI

studies. Feick and Robertson's study (2014) of Flickr tags in Vancouver, Canada used multiple

geographic aggregation units (from 0.25 ha to 1024 ha) showing that different patterns in the

tagging similarity occur across scales, and fine-scale patterns can be obfuscated when

aggregated at larger scales. Even well-established fields utilizing expert level sampling

procedures can have similar difficulties. Galpern (2012) states that the optimal spatial grain is

usually unknown before a landscape genetic analysis and gives the example of looking at various

grains to filter out unimportant variation in the landscape of wide-ranging organism gene flow.

As understanding of various sources of VGI are still in their infancy, an approach to the

utilization of a VGI dataset without a multiscale focus may leave important information hidden

or ignored by the grain or extent chosen. Sester et al. (2014) states that these data have no

explicit scale and generalizes VGI as mostly large scale data (POIs as an example) where scale

must be determined otherwise integration and visualization become problematic. They also

discuss how many delineations that may be found in VGI are not official and may have semi-

permanent boundaries, as in the case of the "uptown" of a city, a "bad" neighbourhood, or

cultural delineations.

Overall, the optimizations were successful in identifying sub-study areas that met criteria

defined by the data quality metric. Finding the right study area without predefined borders can

be challenging. The data is created by users based on certain criteria they have assumed from

the project, irrespective of the researchers needs and through this, researchers need to find

new methods of studying this somewhat chaotic form of data. The need to identify sub-areas

with like characteristics becomes apparent when attempting to analyse VGI, exemplified by the

optimal sub-areas found in this study allowing for a less biased analysis of the chosen VGI

dataset (i.e. skateability). This partitioning method could be used for a new focused study area

or a better idea of the overall study area's characteristics. This paper has shown that spatial

grain and extent can be optimized through simple algorithms using a predefined metric,

allowing for easier and faster study of data or comparison of different grains and extents to

better understand how the data quality of a study using VGI can be used for analysis. This can

be of particular use for public consumption of real-time data collection and immediate

dissemination (Sester et al., 2014). While this paper focused on the optimal extent using

different grains and optimization approaches, there is no limit to the number of areas that could

be returned for further study in any study. For example, the algorithm could have a minimum

threshold for the metric values it returns, allowing for several study areas of interest (extents) to

be further analysed by the researchers. The use of measures of spatial pattern (the metric in

this case) helped to determine the optimal geographic grain and extent using the RinkWatch

active VGI dataset, locating areas of similar characteristics while excluding outlier data points

which could create bias, and thereby facilitating a cursory analysis of skateability in the case

study region. We anticipate that these findings will facilitate a wider adoption of multi-scale

analysis of VGI and promote the data quality assessment of user-generated content in

geography and other disciplines.

3.7 References

Barron, C., Neis, P., & Zipf, A., 2013. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*. Doi: 10.1111/tgis.12073

Bereuter, P. & Weibel, R., 2013. Real-time generalization of point data in mobile and web mapping using quadtrees. *Cartography and Geographic Information Science*, 40 (4), p. 271-281.

Calderón-Patrón, J. et al., 2013. Vertebrate Dissimilarity Due to Turnover and Richness Differences in a Highly Beta-Diverse Region: The Role of Spatial Grain Size, Dispersal Ability and Distance. *Plos One*. 8 (12).

Chase, J., Knight, T., 2013. Scale-dependent effect sizes of ecologogical drivers on biodiversity: why standardised sampling is not enough. *Ecology Letters.* 16, p. 17-26.

Clark, A. & Scott, D., 2013. Understanding the Impact of the Modifiable Areal Unit Problem on the Relationship between Active Travel and the Built Environment. *Urban Studies*. 51 (2), p. 284-299.

Coleman, D., Georgiadou, Y., and Labonte, J., 2009. Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*. 4, p. 332-358.

Dickinson, J., Zuckerberg, B., & Bonter, D., 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics*. 41, p. 149-172.

Damyanov, N. N., Matthews, H. D., & Mysak, L. A. (2012). Observed decreases in the Canadian outdoor skating season due to recent winter warming. *Environmental Research Letters*, *7*(1).

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

Elwood, S., 2011. Geographic Information Science: Visualization, visual methods, and the geoweb. *Progress in Human Geography*. 35 (3), p. 401-408.

Feick, R. & Robertson, C., 2014. A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems*.

Foody, G.M. et al., 2013. Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project. *Transactions in GIS*. 17 (6), p. 847-860.

Freifeld, C. et al., 2010. Participatory Epidemiology: User of Mobile Phones for Community-Based Health Reporting. *PLoS Med*. 7 (12), p. 1-5.

Galpern, P., Manseau, M., Wilson, P., 2012. Grains of connectivity: Analysis at multiple spatial scales in landscape genetics. *Molecular Ecology*. 20 (16), p. 3996-4009.

Girres, J. & Touya, G., 2010. Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS.* 14, p. 435–59.

Goetz, M., & Zipf, A., 2012. Towards Defining a Framework for the Automatic Derivation of 3D CityGML Models from Volunteered Geographic Information. *International Journal of 3-D Information Modeling*. 1 (2).

Goodchild, M.F., 2013. The quality of big (geo) data. *Dialogues in Human Geography*. 3 (3), p. 280-284.

Goodchild, M.F. & Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: A research frontier*. International Journal of Digital Earth.* 3, p. 231–41.

Goodchild, M.F., 2007. Citizens and Sensors: the world of volunteered geography. *Geojournal*. 69, p. 211-221.

Haklay, M., 2013. "Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation." In *Crowdsourcing Geographic Knowledge*. Springer Netherlands, p. 105-122. Print.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design.* 37, p. 682-703.

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

Hannak, A., et al., 2012. "Tweetin'in the Rain: Exploring Societal-Scale Effects of Weather on Mood." In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12),* Dublin.

Jeffery, C., Ozonoff, A., & Pagano, M., 2014. The effect of spatial aggregation on performance when mapping a risk of disease. *International Journal of Health Geographics,* 13 (9), p. 847-854.

Lawrence, H., Robertson, and Feick, R., 2014. Accepted. Identifying optimal study areas and spatial aggregation units for point-based VGI from multiple sources. *Proceedings of the*

ISPRS/IGU Joint International Conference on Geospatial Theory, Processing, Modeling and Applications. Toronto, Canada, October 6-8th.

Kelly, M., Tuxen, K., Stralberg, D., 2011. Mapping changes to vegetation pattern in a restoring wetland: Finding pattern metrics that are consistent across spatial scale and time. *Ecological Indicators*. 11 (2), p. 263-273.

Li, L., Goodchild, M., & Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40, p. 61-77.

Li, W., Goodchild, M., Church, R., Zhou, B., 2012. Geospatial Data Mining on the Web: Discovering Locations of Emergency Service Facilities. *Advanced Data Mining and Applications: Lecture Notes in Computer Science*. 7713, p. 552-563.

Lukyanenko, R. et al., 2011. "Citizen Science 2.0: Data Management Principles to Harness the Power of the Crowd." In *Service-Orientated Perspectives in Design Science Research*. Springer Berlin Heidelberg. P. 465-473. Print.

Mackaness, W. & Chaudhry, O., 2013. Assessing the Veracity of Methods for Extracting Place Semantics from Flickr Tags. *Transactions in GIS*. 17 (4), p. 544-562.

Mooney, P., Corcoran, P., Winstanley, A., 2010. Towards quality metrics for OpenStreeMap. *GIS '10 Proceedings fo the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. P. 514-517

National Audubon Society, 2014. History of the Christmas Bird Count. Accessed July 1, 2014. http://birds.audubon.org/history-christmas-bird-count

Neis, P., Zielstra, D., & Zipf, A., 2012. The Street Network Evolution of Crownsourced Maps: OpenStreetMap in Germany 2007-2011. *Future Internet*. 4 (1), p. 1-21.

Popinet, S., 2011. Quadtree-adaptive tsunami modelling. *Ocean Dynamics.* 61 (9), p. 1261-1285.

Ramanathan, V., Mishra, S., & Mitra, P., "Quadtree decomposition based extended vector space model for image retrieval," *Applications of Computer Vision (WACV), 2011 IEEE Workshop*, p.139-144, 5-7 Jan. 2011.

Raymond, E.S., 2001 *The Cathedral and the Bazaar-Musings on Linux and Open Source by an Accidental Revolutionary (Revised edition)*. Sebastopol, CA. O Reilly. Print.

Sester, M., Jokar, J., Arsanjani, R., Burghardt, D., & Haunert, J., 2014. "Integrating and Generalising Volunteered Geographic Information." In *Abstracting Geographic Information in a*

Data Rich World – Lecture Notes in Geoinformation and Cartography. Springer International Publishing. p. 119-155. Print.

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

Sui, D., Goodchild, M., & Elwood, S., 2013. "Volunteered Geographic Information, the Exaflood, and the Growing Digital Divide." In *Crowdsourcing Geographic Knowledge*. Springer, p. 1-12. Print.

Valles, G., 2014. AMOEBA: A Multidirectional Optimum Ecotope-Based Algorithm. R package version 1.1. http://CRAN.R-project.org/package=AMOEBA

Wiens, J., 1989. Spatial Scaling in Ecology. *Functional Ecology*. 3 (4), p.385-397.

Zielstra, D., & Zipf, A., 2010. "A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany." In *13th AGILE International Conference on Geographic Information Science*, 2010.

3.8 Figures
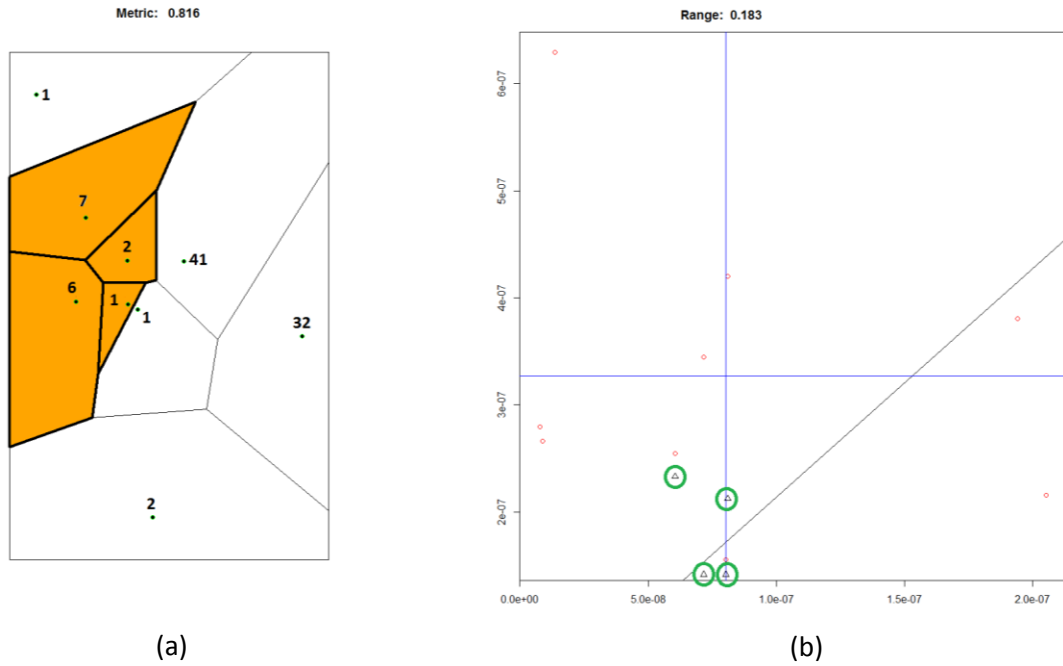
(a)   (b)

Figure 3.1: Voronoi polygons of test data highlighting the density metric component. The image on the right denotes a Voronoi tessellation created from generic test data. The four polygons found are all adjacent in geographic (real) space (a) and close together in the Moran's Scatterplot (b). The spatial lags for the Moran's Scatterplot (y-axis) and the count values normalized by area (x-axis) show that all four of the polygons found are similar.

a)

5	2
1	7

3

1	3	
4	2	#
#	8	8
	5	8

#	#	
#	#	#
	1 3 2 1	
#	7 6 6 2	
	#	6 5
		2 3

b)

#	#	
#	#	#
#	**8**	
	#	

5 | #

8

# #
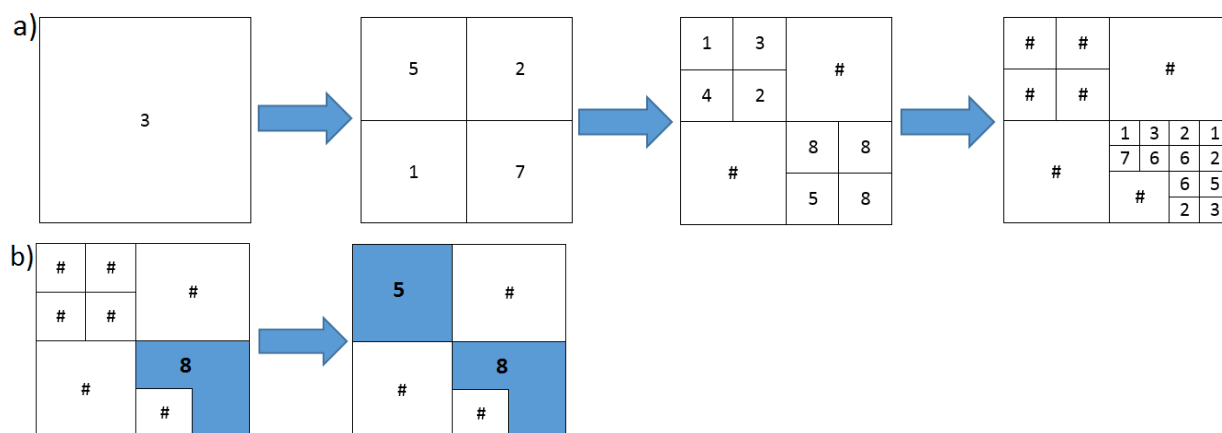
Figure 3.2a: Quadtree branch creation - The first square (metric value of 3) is divided into four nodes. The top left and bottom right nodes metric values (5 and 7) are greater than the parent node's metric value (3) and are further subdivided. In the third diagram, the top left quadrant has no nodes with metric values higher than the parent node, so it stops dividing though the bottom right continues until it also reaches the condition that the four child nodes have lower metrics than the parent.

Figure 3.2b: Quadtree return values - The algorithm then takes the nodes from the lowest branch and merges them. It finds the metric value for them (8 in this general case) and checks that value with the parent node's value (7). If the value is higher, the merged polygon is sent up the tree instead of the complete parent polygon. This example returned two polygons with metrics 9 and 5.
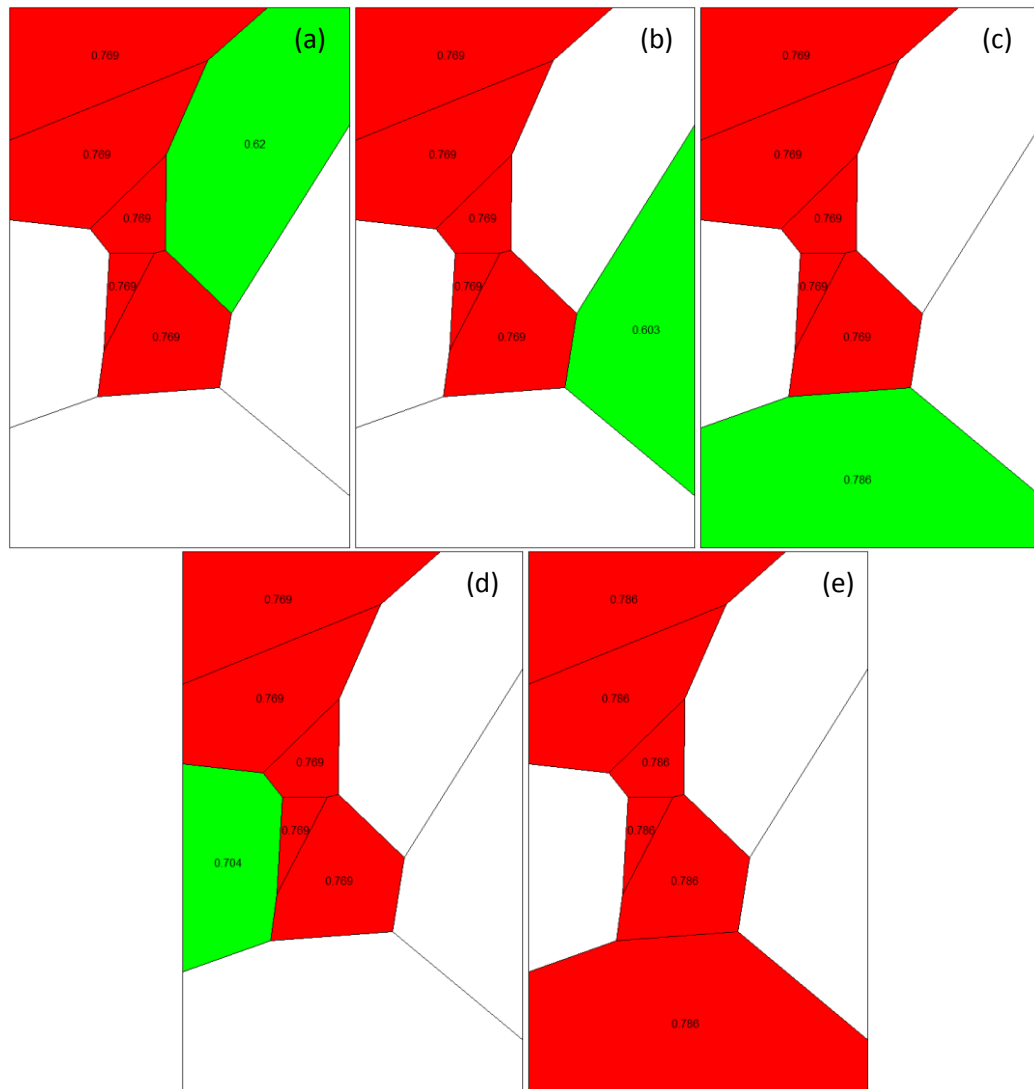
Figure 3.3: The polygon adjacency algorithm – The algorithm checks each neighbour (green) around the merged polygon (red). Images (a) through (d) show one iteration of polygon choice based on metric values and merging. It checks for the overall higher metric (green and red polygons together) and is merged into the red polygon (e) if found to be the higher than the starting value. Adjacent polygons are then checked based on the new polygon until no adjacent polygons create a higher overall metric value.
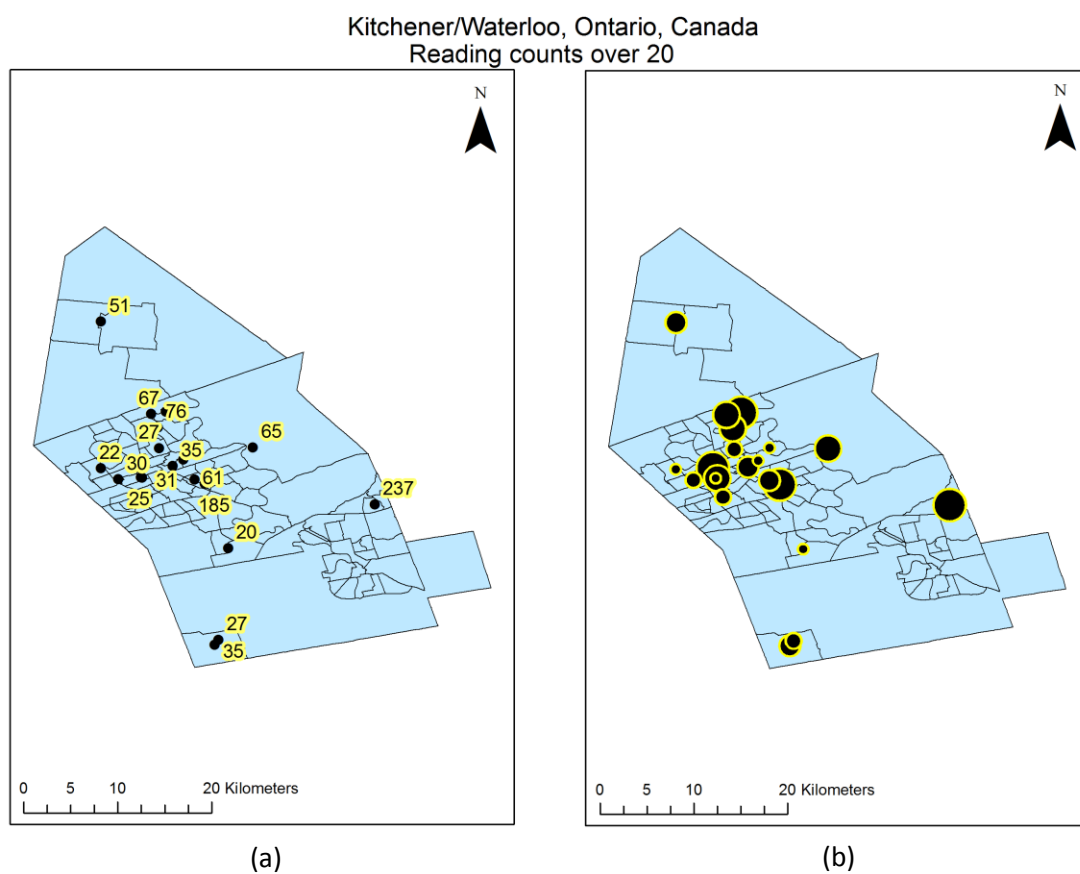
Figure 3.4: Aggregated reading counts of unique rinks with over 20 readings. Figure (a) shows the exact reading counts while (b) shows the counts proportionally (same data).

Metric: 0.148 - Grid size: 1000 sq m   **Metric: 0.656 - Lag order: 1**   **Metric: 0.68 - Lag order: 1**

(a)

coverage: 0.333 , user-heterogeneity: 0.333 , density: 0.333

(b)

coverage: 0.333 , user-heterogeneity: 0.333 , density: 0.333

(c)

coverage: 0.333 , user-heterogeneity: 0.333 , density: 0.333
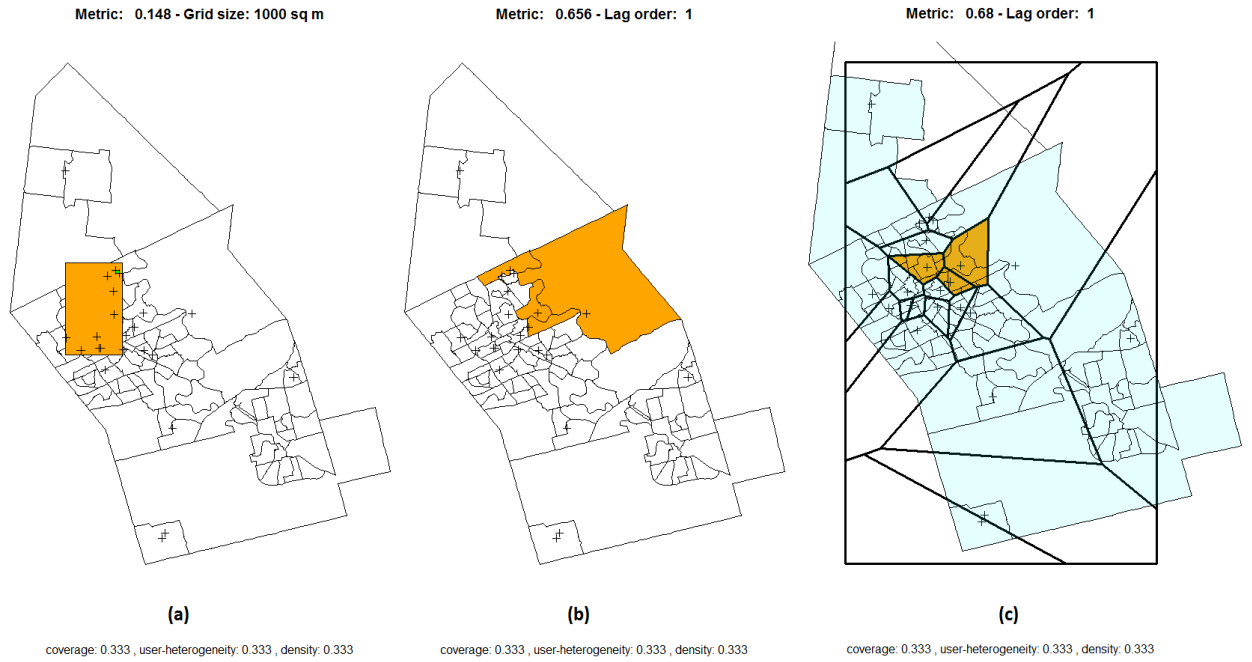
Figure 3.5: Three types of spatial grains used to compute the optimal area at metric input - coverage 33%, user-heterogeneity 33%, and density 33%. (a) Quad-tree gridded polygons are used left, (b) census tracts centre, and (c) Voronoi polygons right.
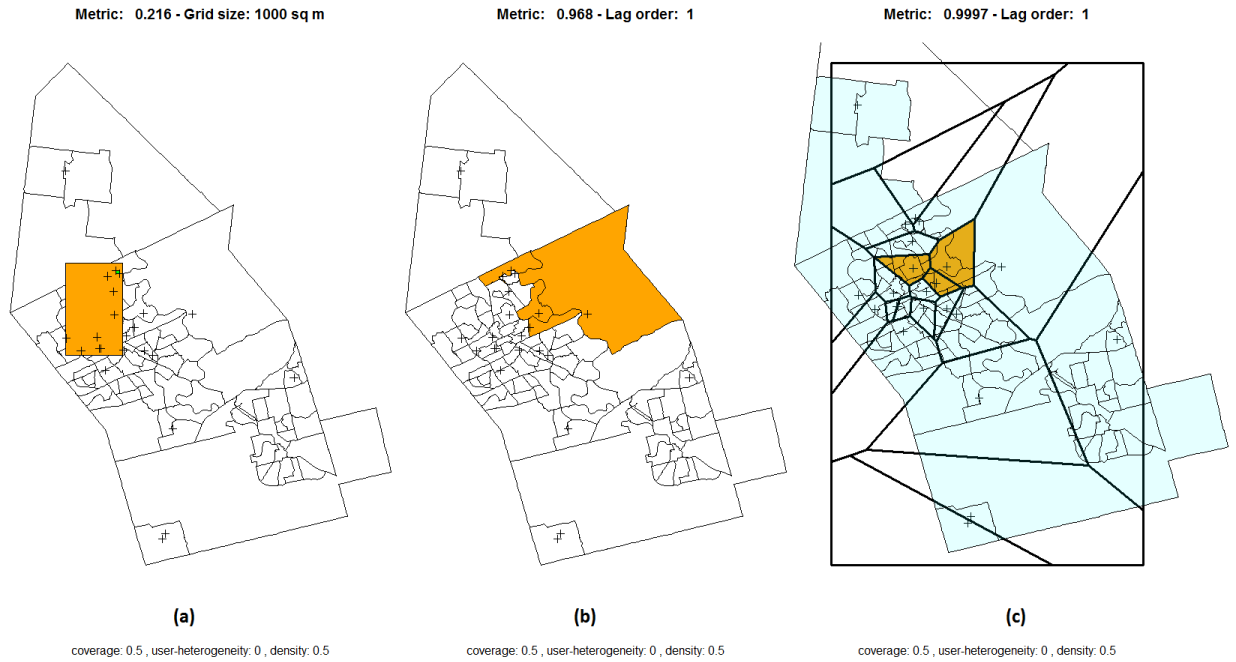
Figure 3.6: Three types of spatial grains used to compute the optimal area at metric input - coverage 50%, user-heterogeneity 0%, and density 50%. (a) Quad-tree gridded polygons are used left, (b) census tracts centre, and (c) Voronoi polygons right.

Metric: 0.656 - Lag order: 2   **Metric: 0.966 - Lag order: 2**

(a)   (b)

coverage: 0.333 , user-heterogeneity: 0.333 , density: 0.333   coverage: 0.5 , user-heterogeneity: 0 , density: 0.5
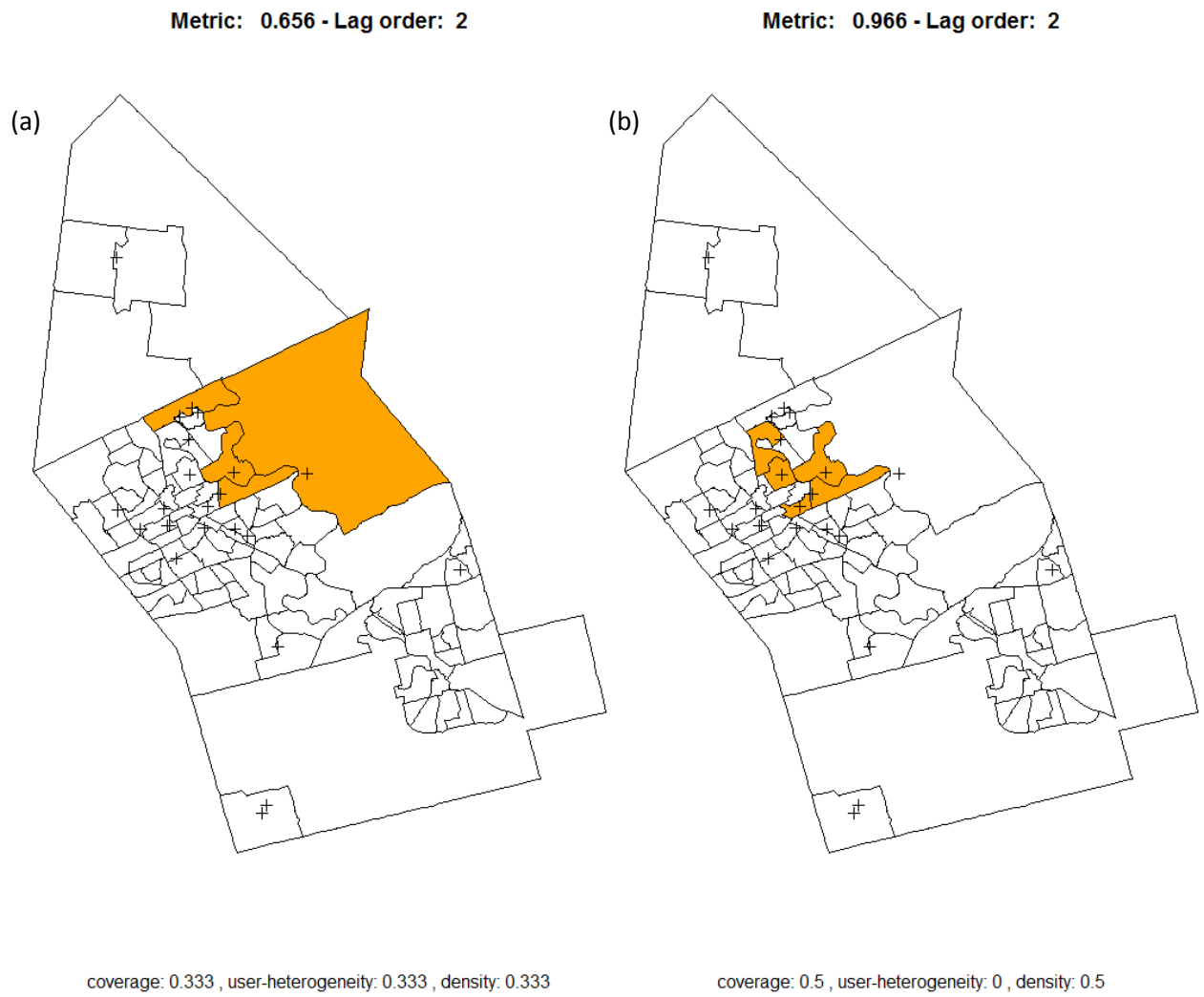
Figure 3.7: Metric values for the optimal area using metric inputs of coverage 33%, user-heterogeneity 33%, and density 33% and a lag order (adjacency) of 2.

Kitchener/Waterloo, Ontario, Canada
Reading counts over 20

(a)

(b)
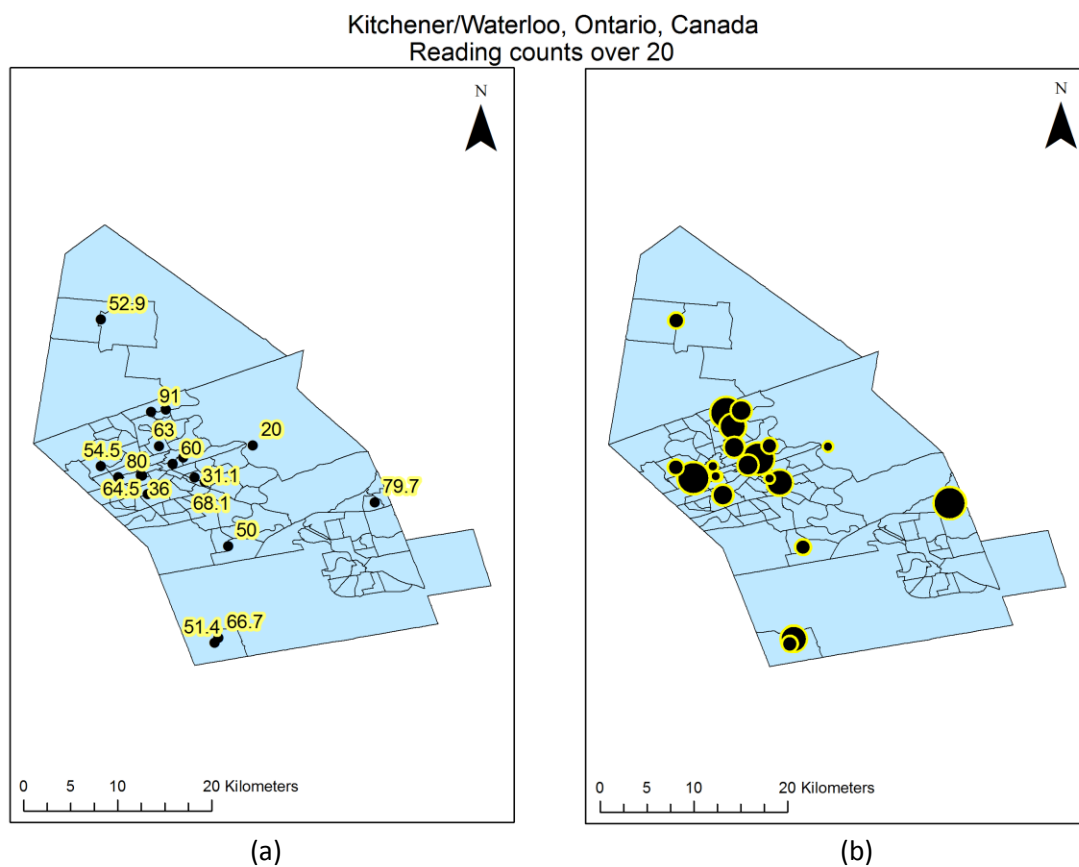
Figure 3.8: Skateability percentage based on readings of Kitchener/Waterloo, Ontario, Canada (minimum 20 readings). Figure (a) shows the skateability percentage over the RinkWatch season while (b) shows the percentages proportionally (same data).
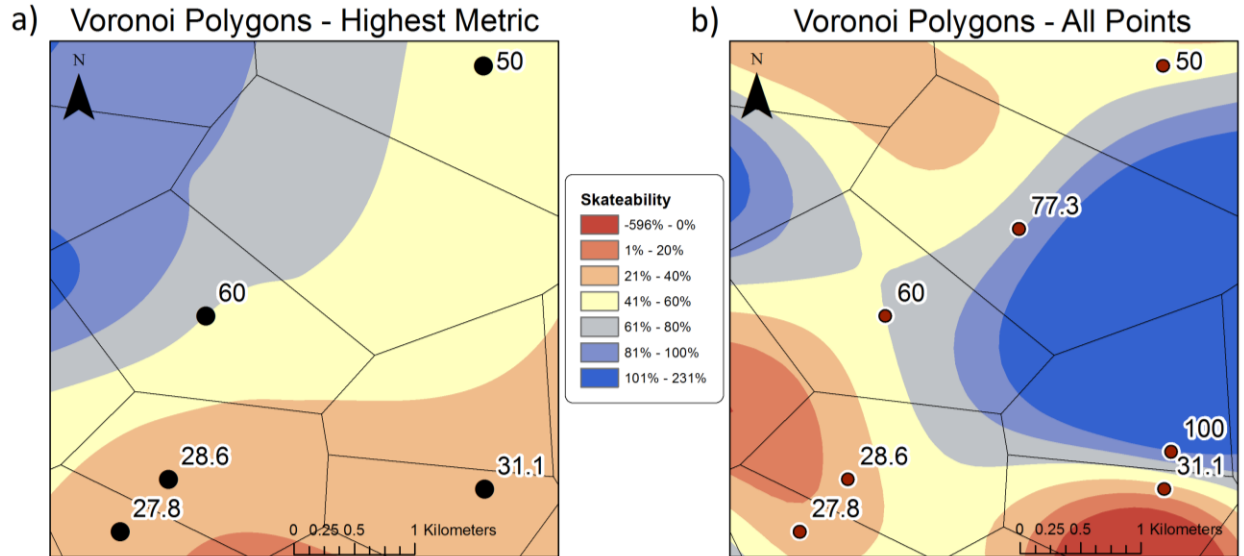
Figure 3.9: Using Voronoi Polygons, a spline interpolation surface fit based on rink skateability. The left surfaces (a) are the optimal areas with the right surfaces (b) showing the complete study area. Metric component weightings are coverage 33%, user-heterogeneity 33%, and density 33%. A spline creates the best surface based on the curvature created by the heights of the data points, causing the values under 0% and over 100%. The optimal area shows a smoother interpolation surface, allowing for higher trust in an analysis of the data.
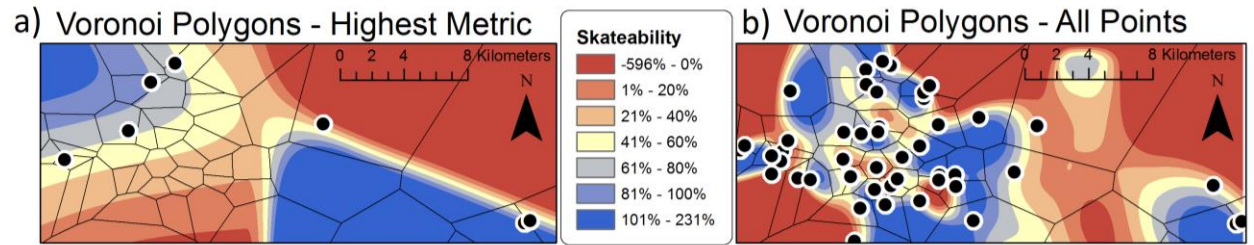
Figure 3.10: Using Voronoi Polygons, a spline interpolation surface fit based on rink skateability. The left surfaces (a) are the optimal areas with the right surfaces (b) showing the complete study area. Metric component weightings are coverage 50%, user-heterogeneity 0%, and density 50%. A spline creates the best surface based on the curvature created by the heights of the data points, causing the values under 0% and over 100%. The optimal area shows a smoother interpolation surface, suggesting a more realistic skateability surface in an analysis of the data.

Chapter Four

4.0 Conclusion

4.1 Discussion and Conclusions

VGI data quality assessment is a timely issue for academic research, as VGI offers an inexpensive and highly sensitive data source with vast potential to transform how researchers collect and analyze data. There are still many issues before VGI can be adopted as a 'normal' data source, though studies from Haklay, Girres, Goodchild, and others have shown VGI can be of 'high quality'. This thesis focused on one the aspects of VGI in which a researcher can control – spatial scale. The research had four objectives to help assess the measurement of data quality at multiple spatial scales:

1) A metric assessing data quality of ambient and active VGI.

2) A methodology to optimize spatial grain using the metric from 1)

3) A methodology to optimize spatial extent using the metric from 1)

4) A case study using the optimal spatial scale demonstrating the overall quality of the optimal area

Through these objectives, the research showed that a multiscale approach to finding the optimal grain and extent can be beneficial to researchers before they start a large scale analysis. The spatial grain analysis of the three Canadian cities elicited several differentiations between the cities that, with further study, could be used by researchers to better understand their study areas. The reasons behind these differences remain speculative (due to tourism for example), however the methods introduced here allow a pre-filtering of VGI to identify sub-areas that support more comprehensive analysis. In addition, the optimization algorithms used in the case study of RinkWatch showed that areas of similarity and high quality can be found easily and

efficiently, with the case study analysis showing a vast improvement from the overall study area when the optimal study area was considered instead.

Chapter Two found that there were quantifiable differences in the metric used when assessing the two larger cities of Vancouver and Toronto, Canada compared to the smaller city of Moncton (~140,000 inhabitants). This finding is important as it may denote the ability to assess different types of cities and possibly different areas within large urban centres in a quantifiable way, allowing for optimization algorithms to assess VGI data for certain characteristics needed for an analysis. The graphs in Figure 2.5 clearly show inflection points where the metric values start to decrease, signifying optimal grain sizes, which could then be utilized in an optimization algorithm. While further examples with more diverse types of cities should be analysed, these preliminary findings allow for ability to hypothesize methods for grain size optimization. The use of the maximal value would not be optimal due to diminishing returns. Both coverage and user-heterogeneity for the CMAs (Figure 2.5 – right) show a strong decrease in the trend of the metric component values at 5000 m^2, possibly signifying an optimal grain allowing for the creation of an algorithm to quantify these inflection points and comparing them to pre-set thresholds of change.

Additionally, an interesting finding was shown through the analysis of the downtown core 100 m^2 grid (Figure 2.8) when underlaid with the street map of Toronto, showing high metric values along major transportation lines (subway line and University Ave.). An algorithm similar to the ones found in Chapter Three might be able to delineate urban facilities, such as transportation lines. This would require further research to verify, however all of the methods reported here could be fully automated into a VGI analysis and evaluation system.

Chapter Three used several optimization methods to create optimal extents of an active VGI

citizen science dataset. The key findings showed that the areas returned from the optimizations

provided a better study area in which to gauge localized skateability (Figures 3.9 and 3.10). The

optimal areas were defined to find spatially adjacent polygons/grids with similar values. In the

case of a RinkWatch skateability analysis, the researchers would want to omit outliers from the

study. A participant who created over 300 data would skew the results if the overlying region's

other participants had a much lower average (72 readings in the case study). While some

variation would be acceptable in a region the size of Kitchener/Waterloo, the variation seen in

the spline interpolations (Figures 3.9b and 3.10b) highlights an example of the difficulty of

analysis of 'raw' VGI. One caveat to this would be the levels of expertise of the participants (i.e.

rink builders/maintainers) which could also cause some variation as new rink builders may have

created substandard rinks. We believe though that the variation in Figures 3.9b and 3.10b still

don't account for these issues.

There was not a large difference found among the different irregular tessellation optimizations.

The Voronoi polygons and the authoritative census tracts both returned similar metric values.

However, these two tessellations showed much higher metric values than the regular, quadtree

optimization. This is due to the fact that the square grids capture all points within them. If the

grain size is set too small, the optimization methods cannot find enough adjacency between

optimal areas though if the grain is too large, it captures too many points due to its shape. A

fluid, moving window would be a significant improvement to a regular tessellation optimization

method. Further study could evaluate how these differences would change for point patterns of

different densities.

Overall, the objectives were met and the goal of assessing data quality through the use of a

multiscale approach proved to be successful. The ability to filter outliers and find similarity

between VGI data points (active and ambient) was shown to be plausible. In addition, while this

research returned only one optimal area in Chapter three, it is possible to return a number of

optimal areas by setting a lower (and possibly upper) threshold allowing researchers to view and

assess the different sub-areas given, which could be useful in other applications of the metric.

The thesis was limited in several aspects. The first was the use of only three Canadian cities in

Chapter Two. The use of more diversity would be much more telling of possible differences

found through use of the metric. This could include different sized cities and cities from

different areas of the world. Another factor that could play a large role in the differentiation of

different areas of a study area would be through the inherent characteristics of the VGI – tags

for Flickr, tweets for Twitter, and various pieces of information for OSM (e.g. POIs). This thesis

did not use these attributes to maintain a generality among the findings, but the use of these

attributes could be used in the future. The final limitation was the quadtree implementation.

While it has the advantage of being an efficient algorithm, O(log n), it does not meet the needs

of finding optimal areas in a study area well based on the patterns explored here As stated

previously, a moving window of varied size would prove to be a better method.

4.2 Research Contributions

This thesis adds to the growing research of VGI, focusing on data quality issues through the use

of multiscale analyses. The creation of methods to quickly guide an analysis using VGI to

optimal study areas and grains would be beneficial to all. In addition to this, the ability to view

VGI at different scales could glean significantly more information about a study area and the

characteristics inherent in the data that might not be readily apparent. Along these lines, this

research doesn't necessarily pertain only to future studies, but could be used in previous studies of VGI possibly resulting in new findings.

While the evidence in support of multiscale approaches is a key contribution, the thesis also produced several other findings that contribute to the overall research on VGI. The differentiation found among the graphs in Chapter Two (Figures 2.5) show observable points where the spatial grain being used creates a lower/higher metric, allowing for optimal spatial grains to be computed. While more examples within the case study would have to be included, this is a very interesting finding. In addition, while outside the scope of this research, Chapter Two results showed interesting pictures of the Toronto City Core (Figure 2.8) with high metric grid cells at the 100 m^2 grain size possibly delineating transportation lines. This also requires further study, but as VGI is often a real-time data source, it would be interesting if certain characteristics of an urban area can be characterized by specific aspects of VGI.

Finally, all of the methods reported in this thesis were developed using open source software and will be made available to the research community as R scripts. This will greatly facilitate the uptake and further extension of these methods and contribute to the academic and industrial uses and study and user-generated data. We envision that as the methods developed in this thesis mature, we may develop them into an R package on the Comprehensive R Archive Network (CRAN) and/or develop a map interface for implementing data quality assessment tools for publicly available geosocial data streams.

Overall, this thesis has presented a very optimistic view of data quality in terms of VGI, both active and ambient. Through this, several mutliscale processes were created which can be easily implemented and used in the study of VGI, even if in a very cursory, exploratory fashion. While it is hoped that these ideas are appropriated by the research community, current research

suggests that multiscale, multi-applicational (e.g. Flickr and Twitter by Li et al. 2013) practices

are increasingly the focus of VGI research (Feick & Robertson, 2013; Calderón-Patrón, 2013;

Kelly, 2011; Galpern, 2012; Chase, 2013).

4.3 References

Girres, J. & Touya, G., 2010. Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS.* 14, p. 435–59.

Goodchild, M.F. & Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: A research frontier*. International Journal of Digital Earth.* 3, p. 231–41.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design.* 37, p. 682-703.

Calderón-Patrón, J. et al., 2013. Vertebrate Dissimilarity Due to Turnover and Richness Differences in a Highly Beta-Diverse Region: The Role of Spatial Grain Size, Dispersal Ability and Distance. *Plos One*. 8 (12).

Chase, J., Knight, T., 2013. Scale-dependent effect sizes of ecologogical drivers on biodiversity: why standardised sampling is not enough. *Ecology Letters.* 16, p. 17-26.

Feick, R. & Robertson, C., 2014. A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems*.

Galpern, P., Manseau, M., Wilson, P., 2012. Grains of connectivity: Analysis at multiple spatial scales in landscape genetics. *Molecular Ecology*. 20 (16), p. 3996-4009.

Kelly, M., Tuxen, K., Stralberg, D., 2011. Mapping changes to vegetation pattern in a restoring wetland: Finding pattern metrics that are consistent across spatial scale and time. *Ecological Indicators*. 11 (2), p. 263-273.

Li, L., Goodchild, M., & Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40, p. 61-77.

Appendix

Introduction - RinkWatch

Active participation of the general public in the collection, analysis, and/or communication of geospatial data for environmental science has a long tradition, with the example of the Audubon Society's Christmas bird count dating back to 1900 (Dickinson et al., 2012; Silvertown et al., 2011). In recent decades, resource managers and public environment agencies have created a range of citizen-based monitoring programs that seek to gather evidence for decision-making (Conrad & Hilchey, 2011). Geographers are showing growing interest in the research value of data obtained from interested citizens (often called volunteered geographical information, or VGI) which include applications in tracking wildlife poaching (Stevens, Vitos, Lewis, & Haklay, 2013), urban noise pollution (Gura, 2013), biological responses to climate change (Beaubien & Hamann, 2011; Worthington et al., 2012), and the well-known Cornell-based e-Bird initiative (Dickinson et al., 2012). Studies have shown that data collected from well-constructed citizen science projects such as these can be reliable and robust (Michael F. Goodchild & Glennon, 2010; Haklay, 2010; Mooney, Corcoran, & Winstanley, 2010). While still a relatively new phenomena, best practices for online, VGI-based citizen-science are starting to emerge. For example, Newman et al. (Newman et al., 2010) describe a number of key attributes of successful projects that include allowing users to contribute data, making outcomes derived from user data visible and accessible to users, and reducing data entry errors by simplifying user interface designs - as important factors in successful web-based citizen science.   When sound project design and data collection principles are followed, citizen science data has proven to be useful for scientific use, with example studies including models of plant phenology (Jeong, Medvigy, Shevliakova, & Malyshev, 2013), climate change-driven adaptations in indicator species (Silvertown et al., 2011), protein structure modelling through game-driven crowdsourcing

(FoldIt) (Eiben et al., 2012), and numerous papers using data from The Cornell Lab of Ornithology projects (Bonney et al., 2009).

There are many important design considerations when creating a VGI initiative, with resources like Bonney et al. (Bonney et al., 2009) and Dickinson and Bonney (Dickinson et al., 2012) being very useful in this regard. However, even after a project has been designed and is ready for launch, there is no guarantee that people will participate in it. The recruitment of new participants and retention of existing participants are important project considerations that are ideally considered from the outset of the project design phase (Chu, Leonard, & Stevenson, 2012). This can have important consequences on the sampling design in the resultant project – as the data obtained for analysis will be directly related to the success of participant recruitment and retention. Despite the critical role of users (i.e., citizens) in citizen science, much more attention has been paid to technology, while the recruitment and retention aspects of citizen science project design have received less attention in the literature.

In this paper, we report on design guidelines for recruitment of citizen science participants in VGI through the experience of www.RinkWatch.org, a VGI-based project that collects data about skating conditions on backyard and neighborhood outdoor ice rinks across North America. Launched in January 2013, by the end of its first season RinkWatch had approximately 1,300 registered users providing approximately 14,000 discrete entries from nearly 1,000 outdoor rinks. Such levels of participation vastly exceeded initial expectations of this unfunded initiative. The aim of this paper is to identify deliberate project design choices and attributes that worked in the project's favour and simple instances of good luck that helped make the project so popular. We also compare the impact of traditional mainstream media with social media in

attracting users to the project website, while describing some missed opportunities as well. In

doing so, we suggest our experience with RinkWatch can provide useful lessons to researchers

contemplating VGI-initiatives.

Background

Individuals' understanding of climate change, and the willingness of the general public to

engage proactively in responses to it, are complex phenomena (O'Connor, Bord, & Fisher, 1999;

Semenza et al., 2008; Weber, 2010; Wolf & Moser, 2011). Despite increasing availability of

information and growing public awareness that climate change is an important challenge, it

nonetheless remains difficult to get the general public to take actions that enhance their ability

to adapt to its potential impacts (Burch, 2010). Part of this may be due to the uncertainty of the

impacts of climate change, especially at local levels, which makes action unlikely unless the issue

is framed such that individuals contextualize potential implications within their own

communities (see Morton, Rabinovich, Marshall, & Bretschneider, 2011). If the success of e-Bird

and similar broad-based initiatives are indicative, citizen science, with its emphasis on engaging

individuals in the creation of environmental knowledge, may provide an opportunity to enhance

awareness of the implications of climate change at local and household levels, thereby creating

the potential for a more actively engaged public. The same way that charismatic mega-fauna

species are used in the environmental conservation movement to catalyze conservation

resources (Clucas, McHugh, & Caro, 2008; Wilson, 1985), a key element of climate citizen

science recruitment lay in identifying activities that enjoy broad-based familiarity and are also

linked to changes in climate (or any target issue), which we'll refer to as charismatic indicator

activities (CIA). The concept of characteristic mega-fauna, or a flagship species, is described as

follows: "*A species that has become a symbol and leading element of an entire conservation

campaign*"(Simberloff, 1998). We propose that the same concept be applied to identifying

problems with great potential for citizen science, and in particular in the recruitment of participants.

In many parts of the Northern Hemisphere, outdoor skating is an exemplary CIA, enjoying longstanding and widespread popularity across much of Canada, the northern US, and northern Europe. Although statistics are unavailable, in many communities and neighborhoods where the climate allows, temporary rinks are made in backyards and neighborhood parks by individuals who flood the ground repeatedly during freezing conditions. These rinks range in size and investment from simple, ad hoc ice sheets to more elaborate structures. Popular media and anecdotal accounts from people living in rinkmaking areas such as southern Ontario, southern Quebec, and New England suggest that milder winters have made maintaining a rink more difficult in recent decades, consistent with studies using climate data that rising average temperatures will make it increasingly difficult to do so (Damyanov, Matthews, & Mysak, 2012).

The geographically widespread popularity of outdoor rinkmaking, its obvious link to winter weather conditions, and the fact that it is an activity engaged in by families of various socio-economic backgrounds make it an excellent opportunity for engaging the general public in VGI-based citizen science (i.e., a case study CIA). Further, we believe that through the empirical collection of reports from citizens about the quality of ice on their homemade rinks, the change over time in the length of the outdoor skating season and intra-seasonal changes in skating conditions can be tracked and used as proxy indicators of winter weather trends, complementing weather station observations. More importantly, observed changes in the CIA would likely trigger greater public response and outcry than changes in weather station observations. In creating RinkWatch we set four goals: to obtain data about outdoor skating

conditions to assess its potential value as a proxy indicator of winter weather conditions and trends; to assess the potential of VGI as a means of acquiring large amounts of data across wide spatial scales; to encourage participation in environmental citizen science more generally; and, to use RinkWatch as a means of generating a wider public discussion about the impacts of climatic variability and change on citizens' day-to-day quality of life. Other examples of winter-based climate citizen science include IceWatch (www.naturewatch.ca), which seeks ice-on/ice-off data for lakes and rivers, and #Snowtweets (http://snowcore.uwaterloo.ca/snowtweets/), which asks users to take snow depth measurements and send these to the research team via Twitter. We hypothesized that RinkWatch, because of its explicit link to a popular outdoor activity, would possibly catalyze the public more so than these efforts. In this paper we explore this hypothesis through the narrative of the recruitment and retention of participants during RinkWatch's launch.

RinkWatch Launch

RinkWatch was officially launched on January 8th, 2013. It was created and maintained by a small team of three researchers using open source tools (Django, Apache, and PostgreSQL) on a server in *The Spatial Lab* housed at Wilfrid Laurier University. Users were asked to register and add daily readings of whether their rinks were "skateable" or "not skateable". As usership increased, functionality was added and the team grew to maintain the social media aspects of the project. One of the first additions to the site post launch was a forum for users, used immediately as a make-shift gallery for pictures of the users' rinks. A map interface was also added showing all the rinks that were skateable and not skateable the previous day, while redefining the word "skateable" to:

"Was the ice solid enough for you to skate, even if you chose not to for other reasons (e.g. too cold outside, didn't feel like shovelling, had a hockey tournament, etc.)?"

By the end of the outdoor skating season, there were 1,334 registered users, 979 rinks, and

13,797 readings in the RinkWatch database. There were 845 rinks located in Canada, 123 in the

United States, 1 in Norway, 1 in Turkey (an indoor ice rink), and several that were either

erroneously placed or deliberately misplaced, such as a rink in South Africa named "skating on

sand". The user forum had a total of 94 topics with 403 posts. The percentage of return users,

defined by having at least 7 readings (one week), was quite high (52.4%). The average user

added 17.4 (median: 7, s.d. 27.5) readings over the season. In addition, while there were more

rinks found in urban areas, rural rink participation was comparable to urban rinks. Figures

Appendix-1a and Appendix-1b show the coverage of rinks in North America and the users who

inputted the most number of readings as of June, 2013. Overall, the season elicited many more

readings than had been expected. Exploratory analysis of temperature readings revealed that

patterns of skateability in areas near weather stations are consistent with official meteorological

stations recordings.

Media Events and Website Visits

The first successful event for RinkWatch occurred with the Montreal Gazette newspaper article

that appeared online on January 8th and in print the following day. The Gazette story started a

chain reaction of interview requests from large Canadian newspaper chains including a live

interview on a popular morning radio show in Toronto CBC Radio's *Metro Morning,* which has an

average listenership of over 1 million ("Consider this: CBC Radio's Metro Morning in Toronto a

model of success," 2012). By the end of February, thirty-five radio interviews in both English

and French had been given by project team members, most in Canada, and stories about

RinkWatch had run in most large daily Canadian newspapers (see Table 1). On January 23rd, a

feature story about RinkWatch was aired on both CBC's national evening television and radio

news. The audio broadcast on CBC radio began at 1800H eastern time and the full video version

aired on national television at 2100H Eastern. The impact of this story is shown in Figure

Appendix-2 which tracks deviation of the number of unique visitors from the hourly means

found between January 16 and April 1 on a weekly basis. The number of visitors during the

broadcasts was much larger than the weekly mean of unique visitors for those hours.

Canadian media coverage of RinkWatch was extensive. Figure Appendix-3 presents the number

of daily visits to RinkWatch.org from mid-January to late February, with the media sources of

important spikes identified. Media coverage of RinkWatch in the US was both smaller in volume

and different in nature, but was nonetheless highly effective in attracting visitors to the website.

RinkWatch was featured on the home page of Scientific American for several weeks, from which

over 300 unique visitors proceeded to the project website. The most one-day visitors to the

project website took place on February 7th after there appeared almost simultaneously a short

story about RinkWatch in the *Huffington Post* (via Grist.org) and a posting about RinkWatch was

made by an influential US blogger, Andrew Sullivan (*The Dish*).

Social media generally proved to be an important source of visitors to the project website. The

project team established a Facebook page for RinkWatch, which contained photos and stories

but was primarily designed to channel people to the main project website. A RinkWatch Twitter

account was created, which was actively used to disseminate online articles about the project

and connect RinkWatch team members with reporters and citizens through retweeting and

following. Table 2 lists the RinkWatch website visitors by originating or referring source; as can

be seen, online media was very effective at generating visitors, with Facebook being the most

effective in generating visits. However, visitors referred from social media showed much less

interaction both in average visit duration and in the number of pages visited, as compared to

more traditional media news sites. Twitter, Facebook, and *The Dish* referrals showed the lowest

interaction through duration and pages visited, while CBC.ca, Yahoo! News, and *Scientific*

American show higher rates of interaction. Google search and direct visitors had the highest

interaction, but these would also capture return visitors and regular users. One of the most

noticeable features of visitors was their interaction with the site based on the referral site they

used.

Overall, forty-four percent of visitors stayed on the website for less than 10 seconds and viewed

only one page, with another 9% staying less than 30 seconds. This leaves approximately 46% of

visitors staying at least 30 seconds or more, which can be seen as a more than acceptable rate of

initial user retention. Durations of 61-600 seconds suggest users who might have more than a

spontaneous interest in the site. A visitor on the site for more than 600 seconds would be

interacting with the site in a meaningful way, most likely browsing or contributing to the forum

though may also signify being away from the computer or browsing other sites while leaving the

window open. Canadians appeared to have higher pages per visit and average visit duration

compared to Americans and most other countries other than Norway (Canada/US: 6.17/3.61

page views and 3:37/2:08 minutes). Canadians also had the lowest bounce rate (leaving the site

from the initial landing page instead of viewing other pages in the site). Popular social media

site referrals tended to show less interaction with the site than those from more "authoritative"

sources, such as news outlets. Google referrals probably contain referrals from radio and

television reports with direct visitation (i.e. from using a URL) consisting of return users.

Rink readings data input pages showed the most page views, with the initial splash page (output

information) and forum being the next most viewed pages. This does show that the site was

being used for its intended purpose; the input of readings with output pages being almost as

important. The "About" page, which describes the project's reasoning and methods, also received many page views in addition to the longest average time on page. Many of the web based media articles on RinkWatch linked to the "About" page.

Discussion – Lessons Learned

A number of important lessons were learned in launching and promoting RinkWatch. Firstly, media attention and user participation rates exceeded our original beliefs about the extent of public interest in outdoor skating in North America. We deliberately counted on this interest when we conceived of RinkWatch. We expected it would resonate well with Canadians, for whom outdoor skating is as much a cultural icon as it is a pastime (e.g. the Canadian five dollar note bears an image of children skating), but had not expected such an enthusiastic response from Americans. Had we undertaken greater preparatory research of our potential audience we might have recognized this sooner and adjusted our initial press release distribution accordingly. These points support the idea that ice rinks and skating culture are good examples of CIAs, a concept which could be further developed for future projects.

Secondly, there appear to be differences between user-interaction on the website and whether they arrived at the site via traditional mainstream media versus online / social media. While both platforms delivered website visitors in sudden bursts, blogs, online only newspapers, and Facebook and Twitter generated predominately short-duration visits. Visitors who learned of the site via traditional media were more likely to remain on the site. For our project, radio seemed to be a very effective means of publicizing the site and generating usership. We should note that many different types of radio programs featured stories on RinkWatch, including local and national news programs, morning talk-shows, daytime AM talk radio, and extended call-in programs; all were effective in generating visits to the website commensurate with the size of

the listening audience. An important lesson is therefore to not overlook 'old' media in the pursuit of 'new' media. However, launching a web-based VGI project without a complementary Facebook page and Twitter account would seem to be missing an inexpensive and easy vehicle for funnelling visitors to a project website and for cultivating a relationship with reporters, most of who use Twitter regularly.

Finally, when eliciting data from the public, the details of the data required must be explicit and well defined otherwise the data may not meet the desired standards. The current definition will be further refined for the 2013/2014 winter season. This is important, as shown in several of the afore mentioned studies (Bonney et al., 2009; Dickinson et al., 2012; Gura, 2013), citizen science data use is gaining acceptance in many fields as a valid way to collect certain types of data. It is also important that users have the ability to provide feedback outside of the data they may be offering. Email contacts and social media help but the largest repository of feedback was found in the forums directed at the RinkWatch project team and among themselves. There were few problems when it came to data input and management. One issue that arose was users wanting to be able to input more data than what the system allowed, including historic data from many years of logging their rinks' details. The site was changed to allow for historic data for the 2012/2013 winter season, but functionality for past seasons has not been added yet. Users also wanted to add comments to their data inputs to further explain their data, especially before the more concrete version of "skateable" was added to the site. This was done in the form of a text input field. In retrospect, it should not have been surprising that people had been tracking information on their rinks outside of official needs and historic data input should have been allowed for from the beginning. Data output is also an important factor that

has yet to be fully implemented in RinkWatch. Currently, efforts are underway to improve data analytics and visualization for users on the site to promote user retention.

With the new 2013/2014 season, it is hoped that usership will not only be maintained at current levels, but can be increased. To help retain current usership in the new winter season of 2013/2014, more outputs for individual users and the general public will be made available in addition to the creation of a gallery to showcase their rinks. A badge or reputation game-like reward system will also be implemented, giving the users a personal attachment to their participation. It is also hoped that new usership can be generated outside of North America, allowing for a better spread of data for analysis. Northern Europe and Asia would be key areas that would greatly help in analysis of the data for environmental/climate change purposes. The data collection will be updated, using a more concise definition of "skateable" and allowing for more specific data about the individual rinks to be inputted. This will allow for a more detailed analysis of the data and allowing for further separation into useful subsections (i.e. publically vs. individually maintained rinks).

Overall, the greatest question is that of what elements of RinkWatch captured public interest in a way to warrant such media attention and how can this be replicated for other citizen science projects. In the case of RinkWatch, the project aims allowed citizens to become champions of the project, taking ownership, providing recommendations and feedback, and developing a community structure. Furthermore, RinkWatch taps into a very personal aspect of community life and culture that allows people to take part in an aspect of climate change research in the same way that the allure of lions or pandas help garner support for conservation efforts. These ideas lend well to media attention, but researchers must be vigilant and take the initiative

during the launch of a citizen science project. Social media and traditional media both played a

large role in RinkWatch's success. However, a project must be able to quickly adapt, both

technologically and conceptually, to unexpected media attention and user suggestions, with

RinkWatch's post-launch addition of a forum serving as a good example.

References

Beaubien, E. G., & Hamann, A. (2011). Plant phenology networks of citizen scientists: recommendations from two decades of experience in Canada. *International journal of biometeorology*, *55*(6), 833–841.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, *59*(11), 977–984.

Burch, S. (2010). Transforming barriers into enablers of action on climate change: Insights from three municipal case studies in British Columbia, Canada. *Global Environmental Change*, *20*(2), 287–297.

Chu, M., Leonard, P., & Stevenson, F. (2012). Growing the Base for Citizen Science. *Citizen Science: Public Participation in Environmental Research*, 69.

Clucas, B., McHugh, K., & Caro, T. (2008). Flagship species on covers of US conservation and nature magazines. *Biodiversity and Conservation*, *17*(6), 1517–1528.

Conrad, C. C., & Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental monitoring and assessment*, *176*(1-4), 273–291.

Consider this: CBC Radio's Metro Morning in Toronto a model of success. (2012, September 6). *The Globe and Mail*. Retrieved from http://m.theglobeandmail.com/commentary/consider-this-cbc-radios-metro-morning-in-toronto-a-model-of-success/article599921/?service=mobile

Damyanov, N. N., Matthews, H. D., & Mysak, L. A. (2012). Observed decreases in the Canadian outdoor skating season due to recent winter warming. *Environmental Research Letters*, *7*(1), 014028. doi:10.1088/1748-9326/7/1/014028

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, *10*(6), 291–297.

Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., Baker, D. (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature Biotechnology*, *30*(2), 190–192. doi:10.1038/nbt.2109

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

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

Gura, T. (2013). Citizen science: Amateur experts. *Nature*, *496*(7444), 259–261. doi:10.1038/nj7444-259a

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

Jeong, S.-J., Medvigy, D., Shevliakova, E., & Malyshev, S. (2013). Predicting changes in temperate forest budburst using continental-scale observations and models. *Geophysical Research Letters*, *40*(2), 359–364. doi:10.1029/2012Gl054431

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 514–517). New York, NY, USA: ACM. doi:10.1145/1869790.1869875

Morton, T. A., Rabinovich, A., Marshall, D., & Bretschneider, P. (2011). The future that may (or may not) come: How framing changes responses to uncertainty in climate change communications. *Global Environmental Change*, *21*(1), 103–109.

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

O'Connor, R. E., Bord, R. J., & Fisher, A. (1999). Risk perceptions, general environmental beliefs, and willingness to address climate change. *Risk analysis*, *19*(3), 461–471.

Semenza, J. C., Hall, D. E., Wilson, D. J., Bontempo, B. D., Sailor, D. J., & George, L. A. (2008). Public perception of climate change: voluntary mitigation and barriers to behavior change. *American Journal of Preventive Medicine*, *35*(5), 479–487.

Silvertown, J., Cook, L., Cameron, R., Dodd, M., McConway, K., Worthington, J., Juan, X. (2011). Citizen Science Reveals Unexpected Continental-Scale Evolutionary Change in a Model Organism. *PLoS ONE*, *6*(4), e18927. doi:10.1371/journal.pone.0018927

Simberloff, D. (1998). Flagships, umbrellas, and keystones: is single-species management passé in the landscape era? *Biological conservation*, *83*(3), 247–257.

Stevens, M., Vitos, M., Lewis, J., & Haklay, M. (2013). Participatory monitoring of poaching in the Congo basin. Retrieved from http://www.geos.ed.ac.uk/~gisteac/proceedingsonline/GISRUK2013/gisruk2013_submission_12.pdf

Weber, E. U. (2010). What shapes perceptions of climate change? *Wiley Interdisciplinary Reviews: Climate Change*, *1*(3), 332–342.

Wilson, E. O. (1985). The biological diversity crisis. *BioScience*, *35*(11), 700–706.

Wolf, J., & Moser, S. C. (2011). Individual understandings, perceptions, and engagement with climate change: insights from in-depth studies across the world. *Wiley Interdisciplinary Reviews: Climate Change*, *2*(4), 547–569.

Worthington, J. P., Silvertown, J., Cook, L., Cameron, R., Dodd, M., Greenwood, R. M., Skelton, P. (2012). Evolution MegaLab: a case study in citizen science methods. *Methods in Ecology and Evolution*, *3*(2), 303–309.
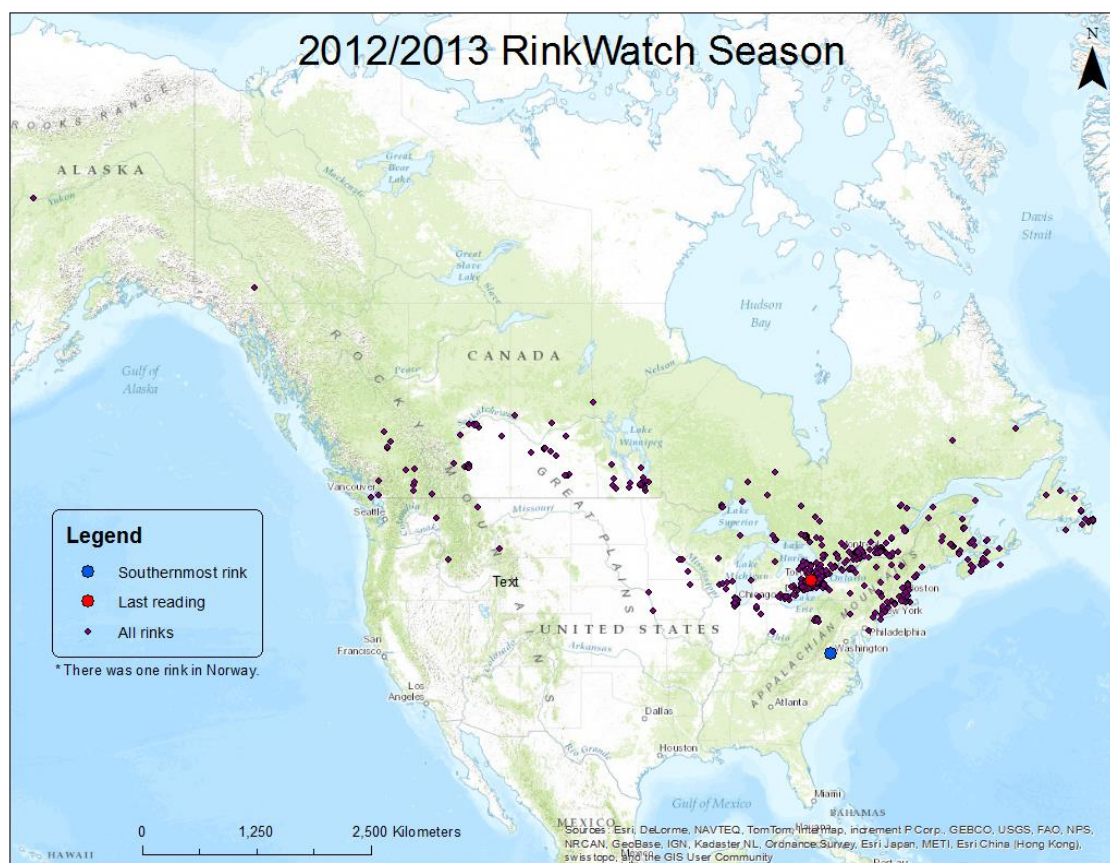
Figures

Figure Appendix-1a: All Rinks for 2013 season (the rink from Norway is not shown for clarity)

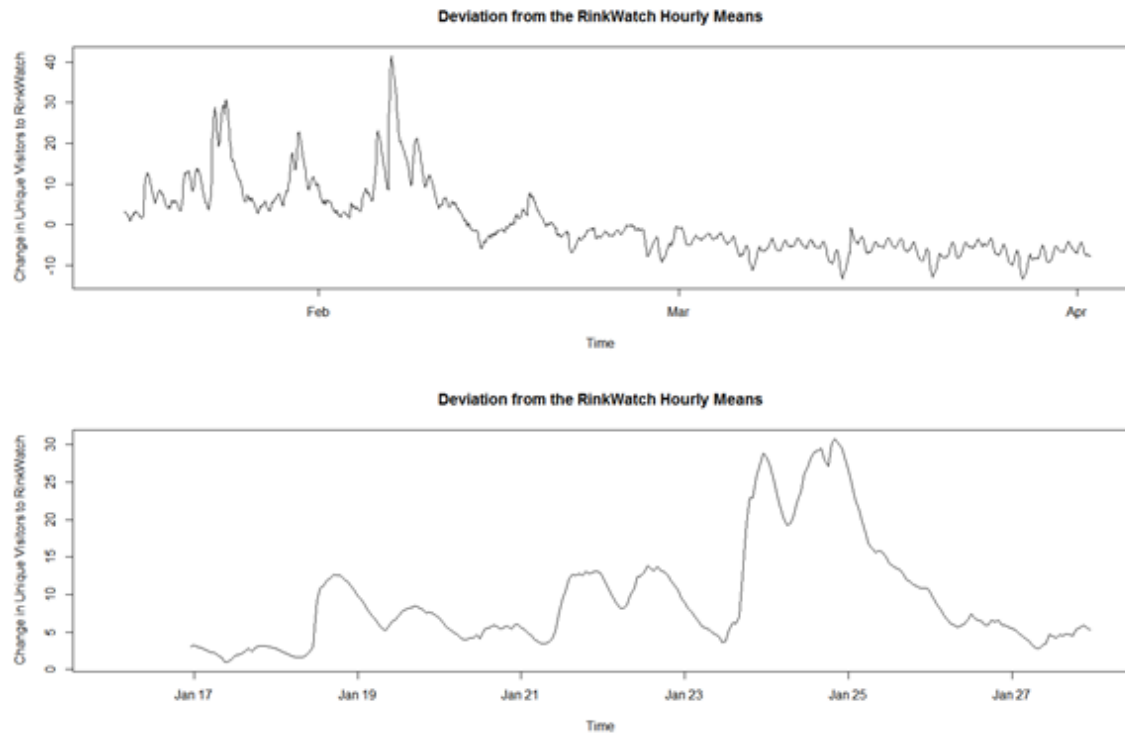Figure Appendix-1b: Top 5% active users

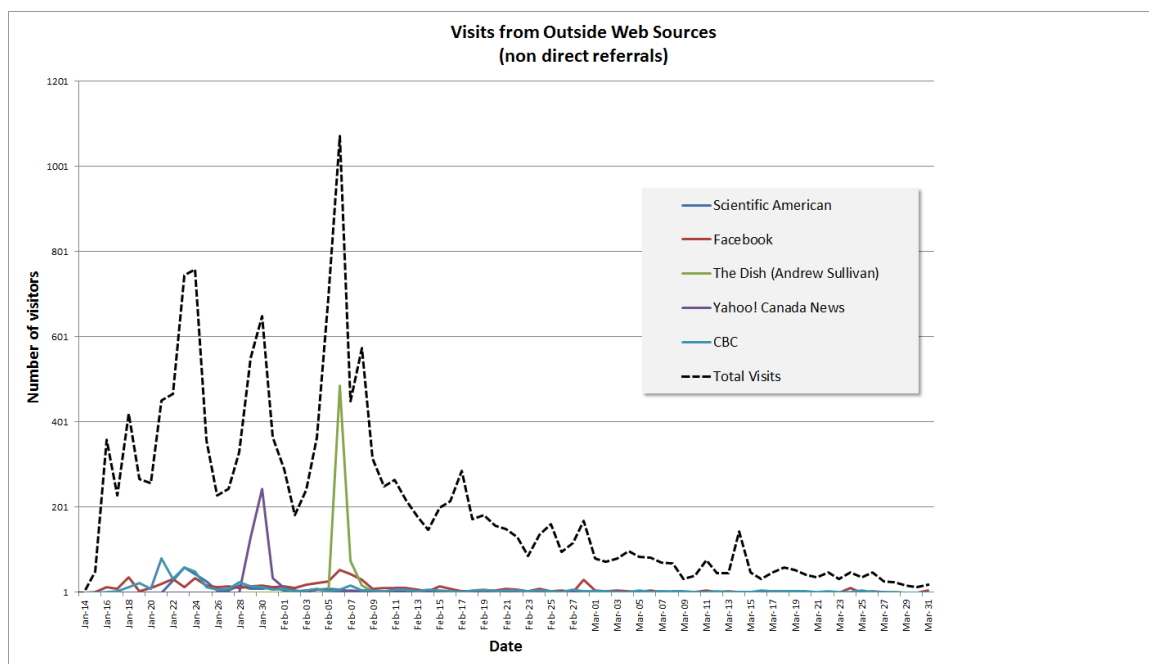Figure Appendix-2: Deviation from the RinkWatch hourly means for January 23/24

Figure Appendix-3: Visits from outside web sources (non-direct referrals)