# Abstraction and Cartographic Generalization of Geographic User-Generated Content

## Use-Case Motivated Investigations for Mobile Users

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Meysam Aliakbarian

aus dem Iran

Promotionskommission

Prof. Dr. Robert Weibel (Vorsitz)

Prof. Dr. Ross Purves

Prof. Dr. Dirk Burghardt

Prof. Dr. Haosheng Huang

Zürich, 2021

# Summary

On a daily basis, a conventional internet user queries different internet services (available on different platforms) to gather information and make decisions. In most cases, knowingly or not, this user consumes data that has been generated by other internet users about his/her topic of interest (e.g. an ideal holiday destination with a family traveling by a van for 10 days). Commercial service providers, such as search engines, travel booking websites, video-on-demand providers, food takeaway mobile apps and the like, have found it useful to rely on the data provided by other users who have commonalities with the querying user. Examples of commonalities are demography, location, interests, internet address, etc. This process has been in practice for more than a decade and helps the service providers to tailor their results based on the collective experience of the contributors. There has been also interest in the different research communities (including GIScience) to analyze and understand the data generated by internet users.

The research focus of this thesis is on finding answers for real-world problems in which a user interacts with geographic information. The interactions can be in the form of exploration, querying, zooming and panning, to name but a few. We have aimed our research at investigating the potential of using geographic user-generated content to provide new ways of preparing and visualizing these data. Based on different scenarios that fulfill user needs, we have investigated the potential of finding new visual methods relevant to each scenario. The methods proposed are mainly based on pre-processing and analyzing data that has been offered by data providers (both commercial and non-profit organizations). But in all cases, the contribution of the data was done by ordinary internet users in an active way (compared to passive data collections done by sensors).

The main contributions of this thesis are the proposals for new ways of abstracting geographic information based on user-generated content contributions. Addressing different use-case scenarios and based on different input parameters, data granularities and evidently geographic scales, we have provided proposals for contemporary users (with a focus on the users of location-based services, or LBS). The findings are based on different methods such as semantic analysis, density analysis and data enrichment. In the case of realization of the findings of this dissertation, LBS users will benefit from the findings by being able to explore large amounts of geographic information in more abstract and aggregated ways and get their results based on the contributions of other users. The research outcomes can be classified in the intersection between cartography, LBS and GIScience. Based on our first use case we have proposed the inclusion of an extended semantic measure directly in the classic map generalization process. In our second use case we have focused on simplifying geographic data depiction by reducing the amount of information using a density-triggered method. And finally, the third use case was focused on summarizing and visually representing relatively large amounts of information by depicting geographic objects matched to the salient topics emerged from the data.

# Acknowledgments

The research that led to the preparation of this dissertation has been done in the Department of Geography at University of Zurich. The main funding has been generously provided by the Swiss National Science Foundation (SNF) under grant number 200021_149823 (Place-based Map Generalization, PlaceGen) which made it possible for me and my PhD project mate Azam (Raha) Bahrehdar to start and finish this journey.

Despite the fact that this document carries the name of one person as the author, doing a PhD from the start to the end is always related to and affected (positively, negative or even both) by a lot of other people. First and foremost, I would like to thank my main supervisor and boss (who is very far from a cliché boss) Robert Weibel. I have enjoyed working in his team and also being in his neighboring office, which included getting a lot of information about various topics. Besides him, in the completion of this thesis I also enjoyed being supervised by Haosheng Huang, who is truly famous for being efficient in providing feedback. I also would like to thank my committee members Ross Purves and Dirk Burghardt for their constructive feedback and ideas during meetings and through email correspondence. Towards the end of writing up this dissertation, I had the chance to work with Tumasch Reichenbacher and Sara Fabrikant for the project MapOnTap. I really enjoyed this time as it helped me to brush up some of my skills and also to practice interdisciplinary work.

During the years I was working and researching at the Department of Geography (a.k.a. GIUZ), my path crossed with a lot of interesting people and obviously that included doing a lot of interesting activities together. Hereby I would like to thank my GIUZ colleagues Azim, Peter R., Michelle, Alex, Beni, Christian, George, Peter J., Katya, Ali, Sascha, Oliver, Raha, Irene, Sanne, Hossein, Parviz, Devis, Michele, Kenan, Gilian, Ismini, Nico, Hoda, Annica, Olga, Manu, Armand, Reik, Daniel, Zhiyoung, Arzu, Max, Bingjie and EK. Interestingly enough, during most of my time in the office, I did not have any officemate but still I had the chance to share the office with Eugenie, Diego, Shivangi, Ramya and Pia.

Looking at my life from a broader view, I always see my family beside me, supporting and motivating me to discover further. I cordially thank my parents Afsaneh and Hossein and my siblings Misagh, Maryam and Ehsan. Last but not least, I thank Nati Bibeli from bottom of my heart for her continuous and genuine support. I also should not forget TinTin and Captain Haddock, who could be more supportive by meowing less around me but (un)fortunately they do not understand what it means to write a dissertation!

# Table of Contents

# List of Abbreviations

| | |
|---|---|
| API | Application Programming Interface |
| CSR | Complete Spatial Randomness |
| DBSCAN | Density-Based Spatial Clustering of Applications with Noise |
| GeoJSON | Geo- JavaScript Object Notation |
| GR | Geographic Relevance |
| GraphQL | Graph Query Language |
| IR | Information Retrieval |
| JSON | JavaScript Object Notation |
| KDE | Kernel Density Estimation |
| LBS | Location-Based Services |
| LDA | Latent Dirichlet Allocation |
| NMA | National Mapping Agency |
| NNI | Nearest Neighbor Index |
| NO | Number of Objects |
| NV | Number of Vertices |
| OLL | Object Line Length |
| OPTICS | Ordering Points To Identify the Clustering Structure |
| OS | Ordnance Survey |
| OSM | OpenStreetMap |
| POI | Point Of Interest |
| PI | Proximity Indicator |
| PV | Proximity Value |
| REStful | REpresentational State Transfer-ful |
| ROI | Region Of Interest |
| SDK | Software Development Kit |
| TF-IDF | Term Frequency–Inverse Document Frequency |
| UGC | User-generated Content |
| VGI | Volunteered Geographic Information |

# 1 Introduction

## 1.1 Motivation

Nowadays a lot of ordinary daily decisions are made based on accessing information from digital sources over the internet (e.g. search engines). When planning an activity (e.g. holidays), a conventional user consults internet sources like search engines, location based social networks, rating websites, bargain offering websites and the like. These services rely on different data sources to offer the best results based on different factors such as user's current query, earlier queries, location, language, weekday, etc. Behind the scenes, a lot of data is collected, evaluated, matched, aggregated, indexed, analyzed and finally presented. A portion of such data is accessed from data that has been generated by other users. Both service providers and researchers have found interest in making use of the data generated by internet users (user-generated content, UGC) but this task is by no means trivial. A lot of challenges (e.g. data quality issues, population bias, users' privacy concerns, vandalism) and interesting discoveries (e.g. people's perception of neighborhoods based on their uploaded photos) are concurrently hidden in analyzing UGC. Besides that, the analysis of internet services reports continuous growth in using the internet via mobile devices. With the general objective of working towards better abstraction of UGC aimed at common (mobile) users, this thesis presents further investigations in different use cases. In the use cases, the focus is on a typical mobile user who queries and explores geographic information based on UGC (with different query parameters). In order to address this objective, new methods are proposed. The methods are based on further investigations of three different aspects of UGC (namely *semantic content*, *spatial aspects,* and *user preferences*).

Web 2.0 was born in the early 2000s and has considerably affected how the internet and the World Wide Web (WWW) has been used afterward. This change has been seen as a revolution both in techniques and usages of the internet. The main change in the usage of the Web has been in enabling previously passive internet users to produce, edit and share content besides merely consuming them. That is why a large number of authors call it *user-centric* or *participative* web. This phenomenon is an attractive research topic, which is by no means limited to a single research field. Researchers have looked at UGC and behavior of users (producers and consumers) to investigate matters that might have not been investigated before or became easier to investigate. Such investigations have brought interesting findings into light. UGC research has, for instance, helped to investigate the correlation between social phenomena and place representation in users' contributions in Ballatore & De Sabbata (2020), characterize urban landscapes in Frias-Martinez et al. (2012) and forecast political elections in Heredia, Prusa & Khoshgoftaar (2018).

An early critique of such findings is the problem of representability of population samples. Common examples are the absence of silent users in data and the limited demographics of typical contributors on different platforms, which limit the generalizability of the findings. An example of this critique is presented by H. Chen, Zheng & Ceran (2016). Despite some shortcomings, findings based on UGC remain attractive and central to a large number of researchers' work all around the world and in diverse fields of research. Elwood, Goodchild & Sui (2012) take a look at this phenomenon in the geographic research context. From a geographic perspective, investigating UGC has attracted researchers to look at different aspects of the collaboratively generated data. Most of the data (or metadata) implicitly (e.g. a placename in the textual content) or explicitly (e.g. coordinates) include a geographic component (Hecht & Gergle, 2010; Stefanidis, Crooks & Radzikowski, 2013).

### 1.1.1 Analysis of UGC

UGC data is generated and consumed differently compared to earlier conventional internet content. In most cases, the analysis of UGC demands new methods for analysis. There are also new facets to consider when performing such an investigation. For instance, research has been invested in understanding not only the content but also the lifecycle of contributors in Bégin, Devillers & Roche (2018) and their social ties in Shan, Ren & C. Li (2017), which was not very common previously. In another research work, basing on user contributions, the authors have calculated multi-dimensional similarity scores between street segments in Bahrehdar, Adams & Purves (2020). The nature of data and research objectives have demanded methodologies with roots in graph theory, text analysis, image analysis and statistical modeling, to name but a few.

Generically, UGC includes a spatial component. This component can be in the main data (e.g. a tweet about East Side Gallery in Berlin), in the metadata (e.g. the geolocation of a Flickr[1] photo) or attached to the contributor (e.g. profile location of the user writing a review on Foursquare[2]) and potentially can be based on different granularities Croitoru et al. (2013). In the GIScience research community, different approaches have been taken in order to understand and make use of the spatial component of UGC. A common approach is to adapt conventional methods to include UGC data e.g. Crooks et al. (2014). Another solution is to spatialize conventional analysis methods in order to include the spatial component of data e.g. Eisenstein et al. (2010).

### 1.1.2 (Geo-) Visualization of UGC

A straightforward way of analyzing UGC is by visualizing the content (or derivatives of the content) Egger & Lang (2013). Visualization helps the researchers (and public audience) to explore and explain data. Visualization can be seen as a step in the whole analysis process or as the goal of the process. When considering (geo-)visualizing data, UGC is different than classic data (mainly provided by official organizations or business companies). Peculiarities of UGC rely on different characteristics, such as being more dynamic (causing more frequent updates and visualizations), more contested (causing the need for interpretation and investigation), generally more subjective, and having lower quality (causing the need for uncertainty visualization and interpolations), to name but a few. Clearly, there is a need to include the characteristics of the data when working with it.

## 1.2 Use cases and research objectives

Despite the fact that there has been substantial research on abstraction and visualization of UGC information, there is still room for new investigations. Considering a mobile user who queries and explores activity possibilities in an urban study area, the user might (explicitly or implicitly) include some parameters in their query. Different situations can be considered in which the user might search for a specific word or category of activities, for example when the user is interested in exploring what is offered/available around him/her without

[1] https://www.flickr.com – accessed Nov. 2019

[2] https://foursquare.com/cities – accessed Nov. 2019

providing any query. Besides including the current location of the user, the system can also include other factors such as time and user's history or favorites. Considering different user input parameters and contextual information helps us to provide specifications for the output of the search query.

The **overall research objective of this dissertation** is to provide data enrichment and geographic visualization proposals to depict UGC contributions based on different query parameters of a conventional LBS user. Focusing on this general objective, we have detected different specific use cases as specific research objectives. The methods introduced in this dissertation are motivated by and answer for these use cases. The first use case shows a situation where a user has provided his/her personal preferences (i.e. favorite object(s)) and is interested to get a map of offered services nearby. There have not been many findings on including similarity measures (extracted from UGC sources) into the process of map generalization targeted for this user. The second use case describes a user searching for a keyword. Earlier research and conventional products have not often addressed the need to generate regions (polygons) as results rather than data points (or symbol clusters). In the third use case, the user is considered to explore the study area (e.g. as a tourist) to find important objects. In order to visualize these objects on the map, there is a need to find accordance between important topics and geographic objects. It has been found that earlier research findings have not addressed this matching process (between physical objects and topic words). Moving one step further, different visualization methods are needed to visualize the topics based on their object matches.

In order to investigate and solve the above-mentioned use cases, the following research objectives are set to be the building blocks of the scientific contributions of this dissertation.

- Research Objective 1: Based on the first use case of the user who is searching and exploring nearby services, the need to include their personal interests is investigated. This is approached by measuring the semantic similarity between the user's favorite objects and the map objects. Methods exist to measure the similarity between objects, but no earlier research has addressed measuring similarity based on keys and values (emerging from our UGC source). Moving one step further, the investigation proposes methods to include the measured semantic similarity in filtering and aggregating objects on the map. To address the latter part there is a need to adapt classic methods of cartographic generalization to include semantics extracted from the data. The findings of this objective will be helpful in providing mobile maps which include objects that are semantically closer to user preferences.
- Research Objective 2: Based on the use case of searching the data source for a certain query keyword (the second use case), the objective is to provide a better cartographic way to provide an overview of the region being queried. Earlier research works have relied on the visualization of query results in the form of points or aggregation of points which are not very helpful to visualize overviews of regions (e.g. services in urban areas). The main focus of this research objective is to develop a scale-aware method to generate and generalize a coarse level overview of points of interest (POIs) in the form of regions of interest (ROIs). This objective should be addressed by providing methods to provide scale-aware and cartographically improved visualization of the results. The findings of this objective will be helpful in providing overview maps of regions that are scale-aware and more readable.
- Research Objective 3: Based on the third use case, i.e. finding significant objects in a study area (e.g. a city), first there is a need to find salient objects and then investigate finding relationships between the topics and the objects. Having marked the matchings, the next step is to provide appropriate visualization of the

topics. Earlier research findings in data conflation and matching have not addressed the need to apply matching between topics and geographic objects. To address this objective, the matching should be followed by proposals for the visualization of the topics. The findings of this objective will be helpful in pre-processing and generation of aggregated maps based on textual contributions of users.

All three objectives are in the direction of either adapting conventional methods or developing new methods. In all three objectives, the focus is on proposing visual abstraction methods rather than evaluating them with user studies. The latter step could be part of follow-up research but is out of the focus of this dissertation.

## 1.3 Context of research

As sketched in the previous section, the scientific contribution of this dissertation is in providing new methods or adapting well-established earlier methods. In achieving that, there is a need to focus more on certain aspects of the data on which the research is based. This helps to put the contributions of this dissertation more into context. Three aspects of UGC are being considered and investigated. Firstly, there is a need to consider and inspect the semantic data included. A lot of research around UGC is related to understanding the semantic content of the data. Secondly, the spatial component of data should be considered for analysis and visualization. Thirdly, the preferences of the user(s) must be addressed and considered. When visualizing UGC data for an end-user, it is important to be able to include the different users' needs (search query, preferences, etc.).

### Semantic content

There is a lot of potential in investigating the semantics of UGC, which makes it possible to ask questions that cannot be answered by analyses merely based on official data. In a semantic analysis process, more importance is given to the implicit or explicit meanings in the contributions. This is approached through analyzing the textual or visual content of the data to find out what the entities, entity classes, user characteristics and user emotions related to contributions are. For example, in Cantador, Konstas & Jose (2011), based on the textual content of tags attached to Flickr photos, the authors classified the tags by checking them against an ontology with the aim of improving recommendations Cantador, Konstas & Jose (2011). In order to apply such analysis, there is a need to process the textual content to find out the entities mentioned in the tags and try to map the entities onto a well-defined ontology. In another exemplary earlier work, Steiger, Resch & Zipf (2016) used a combination of spatial, temporal and semantic clustering methods (and a subsequent visualization approach) to investigate and visualize the contents of Twitter[3] contributions Steiger, Resch & Zipf (2016). This was feasible only after applying a semantic analysis of the tweets.

[3] https://twitter.com – accessed Nov. 2019

### Spatial aspect

In order to understand and use UGC better, the analysis of spatial content plays an important role. Spatial analysis typically considers the first- or second-order effects of phenomena. First-order investigations are related to measuring and modeling a phenomenon based on sample points (in our case users' individual or collective contributions), where second-order investigations are about interaction(s) between phenomena. Different models of distance measures, topological relations and spatial correlations provide a means to investigate spatial phenomena. For example, Y. T. Zheng et al. (2011) have worked on how to reconstruct trajectories of tourists based on their photos shared on a UGC platform Y. T. Zheng et al. (2011). This investigation needs measurement and analysis of distances (photo–photo and photo–attraction). Measuring distances helps researchers both in quantitative and qualitative methods. A direct use of measuring distance between phenomena is in the detection of clusters. With a higher magnitude of data in a UGC analysis activity, detecting the clusters is typically part of UGC analysis.

### User preferences

With new content production data flow, the delivery and representation of data has become more user-customized and user-centric. UGC proves to have potential in providing information for better data depiction. Factors such as user location, demographic class, earlier searches and temporal patterns are used to filter, modify and adapt what is being shown to other similar users (or to the same user). Mobile phones as newer platforms for data production and data consumption simplify this matter as they provide more information in this regard. For example, based on Flickr photos, G. Chen et al. (2012) present a method for recommending touristic destinations to the user based on common interest between the user and other users. For this purpose, there is a need to consider user preferences in the form of photos taken and shared by this user. Applying this methodology had been impossible in pre-UGC times, where potentially the only solution to recommending a touristic destination was to explicitly ask the user for his/her preferences and then to query state-provided information Goossen et al. (2009). Other ways of reflecting user preferences are to include user's search history and/or favorite items (e.g. places).

Figure 1.1 helps to put the research objectives of this dissertation into place in regard to the three aspects introduced earlier. In this figure, each axis is related to one aspect of UGC analysis. For clarification, three research objectives of the current thesis are shown, and they could be compared with three exemplars of earlier research. On the semantic content axis, the exemplary paper is by Hirsch, Hosking & Grundy (2009), in which the researchers proposed visualization of Wikipedia and Freebase concepts in form of graph networks. The research is based on semantic analysis of UGC (with less regard to the other two aspects). Considering the spatial axis, the exemplar research is from reported by J. Li et al. (2015). In this paper, the authors have proposed a method to automatically detect geographic features based on UGC contributions. Regarding the user preferences axis, the relevant research paper is from Waldner & Vassileva, (2014). This paper proposes visualizing a Twitter timeline with a method that filters the tweets based on the user's preference. These three exemplars of earlier research are meant to show an earlier research activity per axis in which the significant part of the methodology is related to that axis. As can be seen, the research objectives of this dissertation collectively cover a large portion of this space, but in general, they have more focus on the spatial content of the UGC in general. This is due to the interest of the author and

also the importance of this aspect in the analysis. Furthermore, this places and pinpoints this dissertation in the realm of GIScience research.

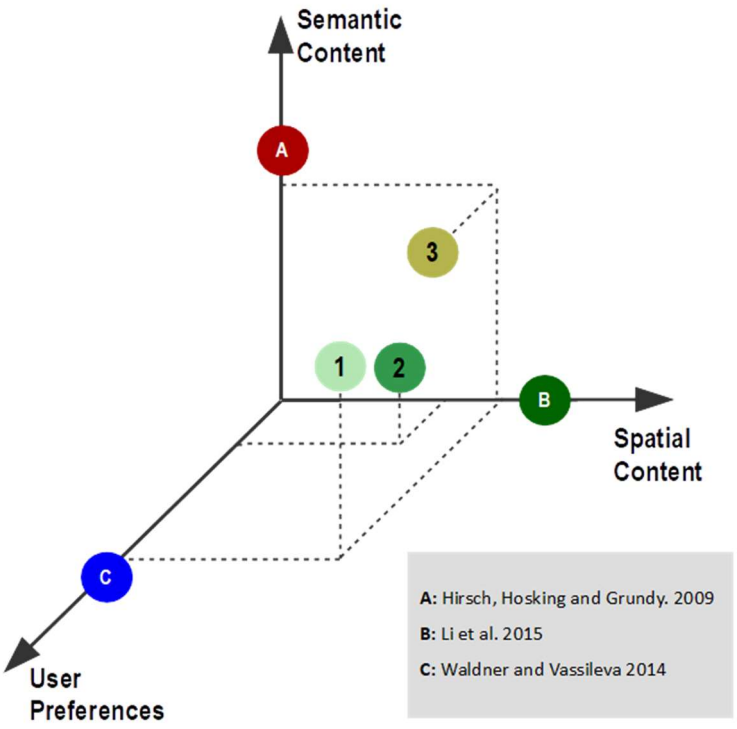


Figure 1.1 – Research objectives addressed in this dissertation. Numbers are research objectives and letters show exemplary earlier research.

## 1.4 Structure of the thesis

Following this chapter is a background chapter covering an overview of concepts and earlier research (Chapter 2). The content of this chapter covers the background which is common between the three use cases at a general level, with deeper details about the earlier research provided in subsections. In Chapters 3–5, use cases and fulfillment of research objectives are presented and the investigations are provided in more detail. As mentioned before, each use-case is coupled with a research objective. Chapter 6 provides a discussion on the findings of this dissertation, and finally, Chapter 7 provides the conclusion and outlook.

# 2 Background

The previous chapter provided an overview of the contents and pillars of this dissertation. In this chapter, we start by shortly reviewing the earlier research in user-generated content (UGC) and the challenges exposed by using such a source of data in a research context. It is important to be aware of the nature of UGC and the implicit challenges of using UGC data, as it is well related to the research objectives of this dissertation (more specifically research objectives 1 and 3). After that, a review of research in GIScience, cartography, and visualization of geographic information with a focus on the developments of the last two decades is provided. This shapes the core foundation for all of the research objectives. Subsequently, a review of past and current research findings in cartographic map generalization (related to all research objectives of this dissertation) is provided. Finally, the research gaps addressed in this dissertation are described.

## 2.1 User-generated content (UGC)

The concept of UGC is based on the concept of Web 2.0, which came to attention in the early 2000s. Web 2.0 is seen as the second generation or phase of the Web. The primary difference between the first and generations is seen to be technical; more importantly, the main difference relies on the new applications and changes in how the internet could be used.

When it comes to the applications, the greatest change that Web 2.0 brought is in content production and consumption. Prior to this era, digital content on the internet was mostly produced by governments, companies, universities and media houses. But due to Web 2.0 inventions, a lot of digital content is produced, shared and accessed by typical internet users. The application changes have helped to make a more interactive, collaborative, participatory and user-centered web (Maguire, 2007; Murugesan, 2007). The most popular Web 2.0 examples and services have been blogs, Wikis, social bookmarking platforms, and media sharing platforms Anderson (2007). Such services also provide reactions between users (such as comments, ratings and following), which are very useful for further investigating the content Casoto et al. (2010).

An important aspect of Web 2.0 platforms is the data generated by users, or UGC. Paying attention to what users express was the main goal when researchers started considering the analysis of UGC. The novelty of investigations based on UGC is what is sometimes referred to as harnessing collective intelligence or relying on the wisdom of the crowd Anderson (2007). Such investigations provide insight into understanding the collective contribution of internet users about different phenomena. Another phenomenon that has commonality with UGC is crowdsourcing. Crowdsourcing is based on the idea of sourcing a task to the crowd and to collectively perform it. Based on a geographic context, See et al. (2016) provide a comprehensive review of terminologies and relevant platforms. Enabling users to produce content was a step in facilitating connections between users over internet services. This has been the base for the development of social network platforms in which users can connect to each other and can exchange information Kaplan & Haenlein (2010). Examples are Twitter, Facebook[4] and Instagram[5]. Analysis of network properties of social networks and users'

[4] https://www.facebook.com – accessed Nov. 2019

[5] https://www.instagram.com – accessed Nov. 2019

interactions in different platforms are provided in Benevenuto et al. (2009) and Mislove et al. (2007).

### 2.1.1 UGC vs. volunteered geographic information (VGI)

UGC in different formats might contain a geospatial component (whether implicit or explicit). The explicit geospatial component can be in the form of a coordinate attached to the content, and the implicit geospatial component can be in different forms (e.g. a toponym in a blog post). When considering UGC with a geospatial component, there exists a similar concept coined by Goodchild (2007) as volunteered geographic information (VGI). Geospatial UGC and VGI have similarities and differences. Similarities lie in the fact that in both, the geospatial data is recorded by non-expert and non-paid individuals. The main difference however, is that UGC sources do not explicitly have a voluntary nature, but this is presumed in VGI Purves, Edwardes & Wood (2011). It should also be mentioned here that some researchers have named the phenomena of creating own maps or providing geographic information by non-experts through internet means (e.g. using map mashups) as "neogeography" or even "web mapping 2.0" (Batty et al., 2010; Crampton, 2009; Haklay, Singleton, & Parker, 2008; Maguire, 2007). This concept is out of the main focus of this dissertation. Here, the focus is on considering UGC data that contains a geospatial component, not on the intention of contributors (whether voluntarily or not). Therefore, the term UGC will be used.

### 2.1.2 UGC research pipeline

Based on the nature of UGC data, research efforts which base their investigation on UGC typically include common steps. The pipeline starts with data collection/access (e.g. via application programming interfaces, or APIs), which in most cases consists of sending a call through the API over the Web and then storing the results. The next step is to cleanse the data and therefore omit different types of noises (missing, incorrect and inconsistent data). Cleansed data is taken as the input for further analysis Batrinca & Treleaven (2014). This is based on the domain and the expected results. The following analysis steps might include descriptive or inferential statistics, cluster detection or classification, to name but a few. Based on the nature of the contributions, specific analysis methods are applied. For example, textual data might typically go through sentiment analysis or entity recognition (Casoto et al., 2010; Schmunk et al., 2014; L. Zhang, S. Wang & Liu 2018) where data of imagery nature will be analyzed using image analysis methods M. Zhang & Luo (2019). This is then followed by the visualization of the results in the form of diagrams, word clouds, maps and interfaces for subsetting the data (e.g. a timeline selection tool), to name but a few. Considering the volume of individual user contributions, different methods have been implemented to produce reports and visualize summaries of a large number of user contributions. Wu et al., (2016) provide a review of methods from a visual analytics point of view.

A review of methods and sub-steps for social media data (a subset of UGC) is provided by S. Chen, Lin & Yuan (2017) and a generic blueprint of subprocesses is illustrated in Figure 2.1.

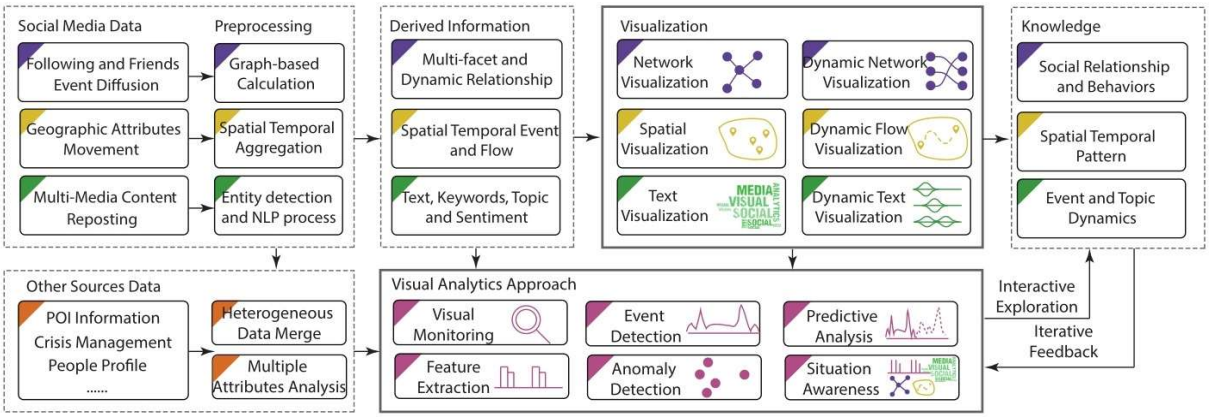


Figure 2.1- General pipeline process for typical social media processing investigations (S. Chen, Lin & Yuan 2017, p. 564). Different colors are related to different categories of data and methods. Figure is used with permission.

In the realm of GIScience, earlier research on content analysis of UGC has worked on the semantic, spatial and temporal analysis of content generated by users. Focusing more on the spatial aspects of UGC, research activities have reported findings such as recognizing users' behavioral and movement patterns in (Andrienko et al., 2013; Noulas et al., 2011), detection of urban placenames and characteristics in (Adams & McKenzie, 2013; Gao, Janowicz, & Couclelis, 2017; Mackaness & Chaudhry, 2013; Zhou et al., 2014), users' collective textual contributions in Adams, Mckenzie & Gahegan (2015) or analysis of the contributors' behavior in (Juhász et al., 2020; Thebault-Spieker, Hecht, & Terveen, 2018).

### 2.1.3 Entities and granularity

Analyzing UGC typically contains investigating different entities that offer different findings. The most straightforward approach is to analyze the content of contributions. This can be inspecting the content of blogs, micro-blogs, photos, videos and likewise. When considering the subjective side of contributions, there is a need to analyze users both individually and collectively. Interactions between users lead the researchers to find out users' common interest and also to identify influential users (or provide different classes of users) (Cha et al., 2010; Welser et al., 2011; Weng et al., 2010). Content creators provide the main contribution, and other users interact with the contribution differently. Different interactions such as commenting, following and sharing are typical.

Another aspect to consider in UGC analysis is text summarization of contents Nasar, Jaffry & Malik (2019). Users employ keywords generically to mark and classify the content. Such keywords, called tags, are helpful to obtain a summary of a web content objects. Collective social bookmarking has been considered as a form of UGC and has resulted in the introduction of folksonomies (Heymann, Koutrika & Garcia-Molina 2008; Hotho et al., 2006). A similar concept is using hashtags. These tags are meant to be used for facilitating the retrieval of resources relevant to a certain tag. Analyzing tags (and hashtags) helps to understand the content of the contribution Nazir et al. (2019).

When analyzing text data – whether considering tags generated and used by users, or keywords extracted (via an algorithm) from contributions – the next step is to understand the relationship between the contributions. This leads to a collective analysis of contributions and

textual topics. The commonality of contributions can be based on time, user, semantic content, location or another factor.

As it has been sketched in this section, several different granularities of UGC can be the subject of research investigation. Individual contributions, group of contributions, tags, individual users (and interactions between them), groups of users or topics could be subject of the research investigations.

### 2.1.4 Matching and conflating UGC

Analyzing UGC data might include a process of accessing data from different sources. In most cases, this is followed by the need to match and/or conflate the data from different sources. From a geospatial perspective, a review of different matching methods is provided in Xavier, Ariza-López, & Ureña-Cámara (2016). The matching process might be based on different entities such as users, locations, contributions (e.g. tweets, photos, and posts), activities and interactions. Generally, the level of relatedness or commonality is measured (an example is provided in McKenzie & Adams, 2018). The classical way to do so is to compare objects with other objects of a similar schema Milo & Zohar (1998). Finding matches might result in combining the data (based on the detected matches). In the geospatial research community, this activity is often called (vertical or horizontal) conflation (Lei, 2020; L. Li & Valdovinos, 2017; Samal, Seth & Cueto 2004; Yuan & Tao, 1999). Considering VGI as a potential conflation source, VGI integration is understood to be part of the research agenda in the community Yan et al. (2020). Based on the UGC data, there has been research in matching objects of similar or different entities. Table 2.1 provides a classification of matching methods related to UGC research.

Most user-generated content sources carry a textual part either as the main contribution (e.g. a blog post) or as extra information attached to a contribution (e.g. a description attached to a YouTube video). This textual content has been the subject of earlier research. A common starting point to the analysis is using text analytic methods on the textual content in order to find important topics within the contributions (Aggarwal & Wang, 2011; Batrinca & Treleaven, 2014). Earlier researchers have worked on detecting salient topics such as places Rattenbury, Good & Naaman (2007), events Becker, Naaman & Gravano (2011) and activities Hasan, Zhan & Ukkusuri (2013). Whether explicitly or implicitly, UGC contributions also include spatial information Kitchin (2013). Mining the location from the contributions has been of interest to researchers (e.g. Stock (2018) provides a review). There has also been interest in finding/investigating the location of user contributions or in understanding if a contribution is about a specific geographic object (e.g. X. Li et al., 2015). For example, Steiger, Ellersiek & Zipf (2014) have identified objects based on the footprint of a collection of UGC contributions (Tweets). In their method, they found salient topics and then generated object geometries by filtering the individual contributions using the topics, and lastly matched the geometries against a database (OpenStreetMap, or OSM). Based on the same data source (Twitter), Hahmann, Purves & Burghardt (2014) investigated the relationship between tweet content and the OSM objects close to the location of the tweet using classification methods. Mackaness & Chaudhry (2013) have used information retrieval methods to detect topics and have analyzed relevant topics based on a manual selection of geographic objects.

Table 2.1 – Classification of matching methods related to UGC research. "Contribution" is a user-generated item (e.g. a tweet), "Object" is a geographic object with a geometry and additional properties (e.g. Tower Bridge[6]), "Topic" is a word that is a collective theme of a group of user contributions (e.g. Oktoberfest or sky).

| Matching sides | | Analysis method | Example |
|---|---|---|---|
| Contribution | Contribution | Clustering | Clustering of Instagram photos for urban pattern recognition (Rodríguez Domínguez et al., 2017) |
| Contribution | Object | Ranking | Photo to POI matching X. Li et al. (2015) |
| Contribution | Topic | Topic extraction | Topic modeling (LDA) based on Twitter data (Ferrari et al., 2011), based on Twitter and Foursquare Kling & Pozdnoukhov (2012) and based on Foursquare (Bauer et al., 2012). Density-based cluster based on Instagram in Rodríguez Domínguez et al. (2017) |
| Object | Object | Object Matching | POI matching between Flickr and Yelp in McKenzie, Janowicz & Adams (2014)<br>Evaluation of UGC source by matching with an official source in Haklay (2010)<br>Image matching in Chamoso et al. (2017) |
| Topic | Topic | Topics Analysis | Classification of topics from Twitter in Lansley & Longley (2016), detecting Geotopics using LDA in Tenney, Hall & Sieber (2019) |

### 2.1.5 Difficulties and peculiarities

The process of using UGC needs to contain a critical view of the quality of the contributions. On a UGC platform, data is provided in large amounts for a cheap price, but less gatekeeping is used to evaluate data quality and credibility in Flanagin & Metzger (2008). In order to improve the quality of user contributions, UGC or social media platforms have considered different approaches internally. These processes are implemented in order to improve the quality, authenticity and credibility of users' contributions. For example, Wikipedia uses a distributed quality assurance system in order to improve the quality in Kittur & Kraut (2008). Social media platforms include more subjective content and therefore invest more in assuring the users' authenticity and also rely on violation reports.

Besides internal concerns and efforts, researchers have also looked at UGC data quality externally. In a geographic context and based on OSM, the most popular UGC source with an explicit geographic component, Zielstra & Zipf (2010) have performed an analysis of spatial completeness of the data based on the dataset covering Germany. A more complete approach to data quality based on the UK data is presented in Haklay (2010). A review of quality assessments in different VGI-related research (namely map-, image- and text-based VGI) is provided in Senaratne et al. (2017). Mooney, Corcoran & Winstanley (2010) propose a more

[6] https://www.openstreetmap.org/way/378541210 accessed Dec. 2019

general approach for evaluating the spatial quality of OSM data. All studies report heterogeneity in data quality where quality is mainly higher in more populated urban areas.

Looking closer at how OSM works, users contribute objects their coordinates and added information. Added information is generally in form of tags which are textual key-value pairs that extend meanings of geographic objects (e.g. the key-value pair of leisure=park describes a park) Mooney & Corcoran (2012)b. OSM data comes with its own peculiarities. Besides the aforementioned typical data quality concerns, when using the added information, some other concerns are present. Firstly, unlike other UGCs that typically use tags (only keys) as the description of the objects, tags in OSM are in the form of keys and values. Secondly, tags do not have a strict structure in the OSM project and generally reflect the particular view of users over mapped geographic objects. Different users might use different key-value pairs to tag the same geographic object. This differs between individual users and also users' regional communities (Ballatore, Bertolotto & Wilson, 2013; Mooney & Corcoran, 2012a). Tagging strategies are negotiated and agreed upon in the OSM wiki[7]. However, there is no single widely accepted regime of tagging. This results in a dataset in which entities have different sets of attributes even though these entities belong to the same geographic object type (e.g., park). Considering mapping based on OSM contributions, earlier research has focused mainly on challenges of data quality and data harmonization (Olteanu-Raimond et al., 2017; Touya et al., 2017) but mainly on geometric aspects of the data. Deeper research with the subjective of mapping based on UGC (and the special case of OSM) considering the data peculiarities with the focus on attributes is missing.

Addressing the issue of information heterogeneity in the level of tags in OSM exposes some inconsistencies. These inconsistencies regarding the number of attributes and their key-value pairs used pose many challenges when using OSM data. A use-case is using OSM data for generation and generalization of geographic maps. The existing map generalization methods are mostly developed for official data sources with well-defined data structures that are not present when considering using OSM. Figure 2.2 schematically shows these differences. In a typical case in the map generalization process (left), a set of geographic features and their attributes (which have consistent and fixed structures) are taken from an official data source. In contrast, as shown on the right, data fetched from OSM often contains different sets of attributes (i.e. key-value pairs) even for the same feature type. An example is about parks in the city of Zurich, Switzerland where official Open Data Zurich[8] contains 117 entries each having specific fields (e.g. name, category, postal code, website and infrastructure information), OSM yields to 312 entries with a different number of attributes per entry (data ranging from park name in different languages, address and website to opening hours and accessibility with horses). Tackling such an issue is the main objective of research objective 1 and will be addressed in more detail in Chapter 3.

[7] https://wiki.openstreetmap.org – accessed Nov. 2019

[8] https://data.stadt-zuerich.ch/dataset/geo_park – accessed Jan. 2020

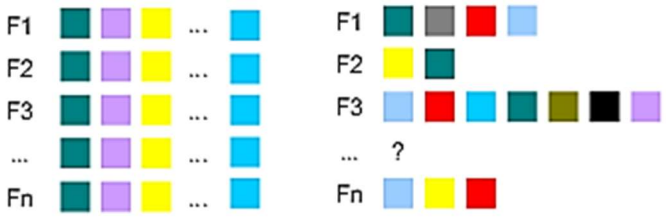


Figure 2.2 – A schematic comparison of data from an official source (left) containing well-defined, structured and homogeneous properties in form of a relational database with the situation of data from OSM (right) in which data is semi-structured and more heterogeneous.

Moving a step forward and when analyzing UGC with enough data quality considerations, there are also other steps needed to gain useful data. There are examples of pre-processing UGC in order to cleanse the data for further process. An example based on Twitter is presented in Singh & Kumari (2016). A primary issue is missing data or having inconsistent data Batrinca & Treleaven (2014). Another common problem to overcome is heterogeneity Ma, Sandberg & Jiang (2015). Besides the classification of contributors and contribution methods, a review of challenges and solutions for understanding and improving VGI data quality is provided in Bordogna et al. (2016).

## 2.2 Modern cartography

Providing visualization of geographic information in the form of maps or other formats has changed in the last decade. The process of printing and distributing maps is being replaced Jenny, Jenny & Räber (2008). The new pipeline is almost fully digital (different steps are performed using digital means). When considering the internet facilitation in the process of providing geographic maps, some authors have used the terms “GeoWeb” or “GeoSpatial Web” Maguire (2007); Venkateswaran (2015). When looking at this phenomenon in 2000s and later, the important changes are enabling users with different web platforms. Examples are designing dynamic map pages and making map mashups Batty et al. (2010); Haklay & Weber (2008). Besides the dynamic visualization, the data itself could also be dynamic. Similar to changes rooted in Web 2.0, the changes have been based on the facilitation of user collaboration.

Geographic data by nature offers notions of scale and level of detail and visualization of geographic information (in form of maps) always offers overviews. Generating overviews as awareness of the content and structure in order to offer navigation and exploration based on reduced data is the objective of overview maps Hornbæk & Hertzum (2011). For example, by having a high number of objects, a straightforward way of visualizing query results on a conventional digital map is using markers for each result record. In case of having visual clutter (as the result of markers being very close to each other), different solutions are typically considered (e.g. combining markers as marker clusters). This is a common issue when visualizing the results of user queries. Generating an overview of regions that represent or relate to individual points has been investigated before for different objectives. This has been named before as region of interest (ROI) and such regions are used to generate aggregated objects of interest. The motivation is to delineate a region that represents the points (e.g. measurement points, geo-referenced documents, etc.). The region can serve as an aggregated overview of the covered points.

Earlier research has tried to generate regions to investigate a hypothesis (for example, about vernacular placenames in Hollenstein & Purves, 2010) but the focus was not found to be on visualizing the results as regions. Other examples are provided by Cao et al. (2014) and Lamprianidis et al. (2014), where ROIs have been generated in order to represent a high density of POIs (related to a query). Yu et al. (2017) have investigated the relevance of locations to query words. Zhu et al. (2019) address the challenge by focusing on finding the semantic structure of the points.

Moving a step further, the research of this dissertation investigates visualization of results in a more cartographic way. The earlier research findings (e.g. those by Cao et al. 2014 and Lamprianidis et al. 2014) did not mainly address the visualization of ROIs in different scales and map extents. The focus of earlier research (e.g. Cao et al., 2014, Lamprianidis et al., 2014 and Yu et al., 2017) has been mostly on finding a region relevant to a query, but no investigation has gone further in querying for different scales and map extents (e.g. by taking the current map extent into account). For generating such a region, point density can be used to take high densities of points as higher levels of confidence in representing the points. Different approaches in modeling and analyzing density have been proposed, among which density-based clustering and generation of density surfaces are very common.

Kernel density estimation (KDE) is a typical method of modeling density for (geo)spatial problems. For example Jones et al., (2008) have used this method to delineate vague regions where Grothe & Schaab, (2009) used KDE (and also support vector machines) to generate region footprints from geo-tagged photos. Henrich, Lüdecke & Blank (2008) have used a modified KDE in order to relate terms to regions.

Based on geo-tagged photos and with the objective of delineating city cores, Hollenstein and Purves have also used KDE generated surfaces Hollenstein & Purves (2010). Also in Yingjie Hu et al. (2015), the authors have worked on defining and understanding urban areas based on geo-tagged photos. In Gao et al. (2017), the authors have worked on delineating cognitive regions based on UGC. Combining density maps in order to result in aggregated density maps has been investigated in (Lu et al., 2015; Scheepens et al., 2011, 2012). The authors have addressed combining different map topics, but map extent and scale have not been addressed.

Most of the earlier work mentioned here has been investigations in information retrieval. However, by focusing on the visualization step, conventional ways of visualizing (geo-) information are mostly interactive systems that provide the results based on the user's current view and interactions with the system. Providing methods that cover different scales and map extents enables us to adapt the result based on the viewer's current view over the data (i.e. zoom level and current map view). Substantial work has been done in order to find and delineate ROIs, but when considering visualizing ROIs in a valid manner there is a need to define and evaluate ways of visualizing ROIs with the objective of generating a cartographic map. In Adams, McKenzie & Gahegan (2015), the authors performed a method partially similar to the methodology of research objective 2, but their objective and the results are different. Their approach is similar in using KDE to estimate the density of terms and also to relate KDE bandwidth to zoom level, but the differences are generating regions (vs. density maps) and the parameterization of the method. An exemplar use-case is also provided in Bereuter (2015) in which density estimation is being used to show regions with higher densities of points (activity keywords). In that research work however, KDE parameters are pre-calculated and are not based on the data. Emphasizing the visualization aspect of generating overviews over the study area is the focus of research objective 2 and will be presented in more detail in Chapter 4.

## 2.3 LBS

A generation of new services that have partial foundations in geographic information has been classified under the term location-based services (LBS). These are services that have geographic information as a backbone and are focused on accessing services over mobile networks Jiang & Yao (2006); Raper et al. (2007). LBS are based on different context factors such as location, profile, time and history Grifoni, D'Ulizia & Ferri (2018); Steiniger, Neun & Edwardes (2006). Some implementations of LBS offer a high level of personalization based on their individual and collective user profiles H. Huang et al. (2018); Mokbel et al. (2011).

An integral part of LBS lies in submitting different queries and then receiving the results and applying appropriate visualization of the results subsequently. Another property of LBS is their usage of digital maps. Usage of this new medium needs specific considerations such as dealing with limited screen size. Besides handling dynamic data, designing for limited screens made it more important to develop data abstraction methods Edwardes, Burghardt & Weibel (2005); Gartner, Bennett & Morita (2008); Reichenbacher (2004). Based on the inherited nature of LBS, which typically (but not always) involves working on smaller screens, providing overviews over the whole datasets is of high importance. More classical literature such as Information Seeking Mantra suggests "overview first, zoom and filter, then details on demand" (Shneiderman, 1996, p. 336), and this view is well adapted in classic geographic information visualization Keim (2002); Keim, Panse & Sips (2005); Y. Zheng et al. (2016) and geographic maps Kraak (2011).

Visualizing high volumes of UGC content on an LBS in the form of overview maps requires aggregating source data in various ways. This step relies on aggregating data in which a crucial component is to aggregate information based on the geospatial location of user contributions S. Chen, Lin & Yuan (2017). Spatially aggregating user contributions is approached by analyzing the density or proximity of contribution points S. Chen, Lin & Yuan (2017). Current research in LBS is moving towards further considerations of context-awareness H. Huang et al. (2018); Yürür et al. (2016).

The content of this dissertation has some commonality with research in the field of LBS. Several concepts such as tackling visualization of geographic information on limited screens, including user location in queries, providing abstractions and including user preferences in the visualization of information are the similarities between the contents of this dissertation and the past and ongoing research in LBS. However, concepts and challenges such as indoor positioning, user privacy and wayfinding are not the focus of this dissertation.

## 2.4 Cartographic generalization

### 2.4.1 Classic research in cartographic generalization

The process of map generalization has been an integral part of cartography and map-making. Different conditions and situations are taken as triggers for this process, of which scale and map purpose are the most common Brassel & Weibel (1988); Shea & Mcmaster (1989). While the scale is rather a geometric trigger, map purpose is related to information content and user needs. Similar to many other phenomena, generalization has also gone through a change from being a manual process (human knowledge) to an automated counterpart (digitally modeled knowledge). Besides that, the research community has worked in the direction of modeling this process both holistically and in detail Harrie & Weibel (2007). In breaking down the process into smaller sub-processes and sub-models, different entities such as operators, rules and constraints have been introduced and investigated.

Generalization operators are atomic sub-processes each performing a specific task on the geographic map objects Cecconi (2003); Foerster, Stoter & Köbben (2007). Examples are selection, aggregation and exaggeration operators. The operators generally work separately from each other; therefore, they need a super-process to control and orchestrate them. Having different combinations of running order and input parameters for each operator makes it necessary to have qualitative and quantitative measures to be able to select relatively better results. This has been approached by introducing rules Nickerson (1986) and later constraints Beard (1991) in the process of generalization Harrie & Weibel (2007). The introduction of constraints has helped to gain quantitative measures of penalties and thus the quality of the generalization process. The goal of introducing these constraints was to measure to what extent the process conformed to the constraints which were modeled based on the experience of cartographers and the common practices in national mapping agencies (NMAs). Defining these constraints (e.g. topology constraints) has been done both by researchers (e.g. Harrie, 2003) and map production experts (e.g. Spiess et al., 2002), facilitating the introduction of optimization procedures Højholt (2000); Sester (2005).

Generalization constraints are triggered based on different situations (visual clutter to be a common one). This happens when a high number of geographic objects are visualized very close to each other, and thus it is hard for the map user to see the objects individually. When a visual clutter constraint is triggered, besides (re-) applying the classic solution of using generalization operators (such as selection or typification), other solutions can also be considered. When generating dynamic digital maps, another conventional way to solve visual clutter is to use marker clusters (showing one marker representing several POIs H. Huang & Gartner (2012)). There has been also another method in which space is deformed in order to provide a better representation of data points Bereuter & Weibel (2017); Edwardes, Burghardt & Weibel (2005). Having visual clutter is taken as the trigger for the methodology of research objective 2. Further details are provided in Chapter 4.

### 2.4.2 Conventional research trends

Research in the realm of cartographic generalization and geographic information abstraction is coping with new challenges motivated by either research or application demands or a combination of both.

Moving from predefined fixed-scale generalization design, newer research efforts explore possibilities of providing continuous or vario-scale generalization Meijers et al., (2020); Sester & Brenner (2004); van Oosterom & Meijers (2011). Examples are provided in Peng, Wolff & Haunert (2016) for administrative boundaries, in Peng & Touya (2017) for built-up areas and in Šuba, Meijers & van Oosterom (2016) for road network objects. Besides addressing the challenge of vario-scale generalization and regarding the need to provide real-time cartographic products, there have been efforts in providing algorithms that can provide real-time outputs for different generalization tasks. This is often called on-the-fly generalization and Bereuter & Weibel, (2013) have provided an example set of algorithms.

Another important research thread for further automating the generalization process is working towards the development of ontologies and geographic knowledge engineering. This thread of research is after making implicit knowledge more explicit, machine-understandable and transferable. Examples are Gould & Mackaness (2016), W. Huang & Harrie (2019) and W. Huang et al., (2020). When combining this goal with the usage of UGC, this becomes even more challenging when considering the semi-structured nature of UGC data (Ballatore, 2016; Lemmens et al., 2016). Tackling this challenge requires investigation of semantic structures of the data and the processes involved. This investigation helps us to generate maps that are closer to user needs and also easier to share between different systems.

The process of map-making is traditionally based on standardized data collection, processing and visualization of geographic data. While most research in map generalization to date has focused on topographic maps where semantics is typically well-structured, maps relying on UGC will display a wider variety of semantics, as UGC-based data sources are more heterogeneous than the ontologies defined for topographic data of official sources Sester et al. (2014); Touya et al. (2017). Consequently, there is a need to invest more in integrating semantics in the generalization procedure. For example, Burghardt, Dunkel and Gröbe have provided a review of the potentials of including UGC-based findings in the process of map generalization (on generalization operator level) in Burghardt, Dunkel and Gröbe (2017).

Using semantically driven measures in process of generalization (in the level of operators) is a step forward to include user personalization in the generalization of mobile maps. The motivation for integrating UGC knowledge into map generalization is to move toward maps that convey not only the data generated by users but also the knowledge that is engrained in user contributions, e.g. in the form of special semantics and tags that users attach to features captured in UGC, and which can be extracted from the UGC Sester et al. (2014). As it has been sketched before, earlier research has shown interest in providing the means to include users' preferences (e.g. user interest) in query parsing and visualization. The task is based on matching the map content to the user's interest. For example, Pippig, Burghardt & Prechtel (2013) have investigated including semantic similarity (between map objects and a selected theme) in route planning. Their approach is based on data from Wikipedia[9] and generates routes based on similarity to a selected theme. Also, earlier research in the framework of Geographic Relevance (GR) has proposed some context in Crease & Reichenbacher (2011); de Sabbata & Reichenbacher (2012). However, no investigation has moved towards further steps to not only provide the results based on semantic similarity measures but also include semantic-driven measures in the process of cartographic generalization. Research objective 1 provides further steps on this track and provides a means of integration of UGC-driven semantics in the process of map generalization. The focus is mainly on two generalization operators (selection and aggregation) which are found to be relevant when considering the inclusion of UGC-based data Burghardt, Dunkel & Gröbe (2017). This research objective is presented in chapter 3. The aim is to detect and utilize semantic measures derived from geographic UGC (OSM contributions) and investigate their potential in the generalization process.

A UGC-related research trend in the research society is the need to conflate and fuse data from different sources. The sources might be from different providers, in different formats, with different granularities (temporal and spatial) and with different semantic modeling. A discussion and use cases on this thread are provided in Sester et al. (2014).

### 2.4.3 Quantitative evaluation

Research on generalization has always been aware of the need for evaluation and quality control of the process. In order to evaluate and optimize the generation of cartographic maps, there is a need to quantify and formalize appropriate visualizations. The evaluation covers aspects of objects representation as well as map readability Stoter et al. (2014). Cartographic evaluation of generalization results takes place before, during and after generalization Stoter et al. (2014). When considering to evaluate visualization and generalization of map features

[9] https://www.wikipedia.org – accessed Feb. 2020

in a quantitative manner, constraint-based approaches are taken into account where certain situations are evaluated based on the violation of constraints Harrie & Weibel (2007).

There has been research reported about measuring map readability measures through user tests Harrie & Stigmar (2010). The motivation was to extract map readability measures in order to include them as constraints. When considering map readability and quantifying information content on maps, different measures (e.g. number of objects on the map) have been introduced and evaluated Harrie, Stigmar & Djordjevic (2015); Stigmar & Harrie (2011). Harrie, Stigmar and Djordjevic have categorized map readability measures into four categories: the amount of information, spatial distribution, object complexity and graphical resolution. Based on a practical user study, they reported higher importance of the two former measure categories Harrie, Stigmar & Djordjevic (2015). In Bereuter, (2015), the author based her quantitative generalization evaluation on categories of data reduction, conflict reduction, data enhancement, displacement, maintenance of spatial patterns, homogenization and cluster maintenance. These criteria have been adapted to the objectives of point generalization.

## 2.5 Research gaps

Based on the literature presented earlier in this chapter, the research gaps of this dissertation are the following.

### 2.5.1 Integration of UGC-backed semantics in map generalization

In order to advance towards semantically enriching location-based service (LBS) and on-demand mapping, there is a need to focus on UGC-backed semantics in the process of generating a map based on user's intention and spatial context. In fulfilling this need, an important step is to integrate the semantic knowledge into the process of map generalization. When considering integrating semantic measures in the generalization process, as discussed before in section 2.4.2, the utilization of measures based on OSM contributions (considering the peculiarities) has not been formerly addressed. As mentioned in the previous chapter, this research gap is addressed by a use case of a typical mobile user searching for urban services around him/her. The following investigations are provided and discussed further in Chapter 3:

- Integration of semantic similarity measures into map generalization (with focus on selection and aggregation)
- Extension of a semantic similarity method to adapt to OSM data structure

### 2.5.2 Visualization of geographic data points in the form of regions

Both research and product development in LBS and web mapping have dealt with the challenge of point clutter on digital maps. This has been approached either by applying selection (based on a ranking algorithm) or marker clustering (a form of aggregation). As reviewed earlier in sections 2.2 and 2.3, regarding this challenge (e.g. how to manage visual clutter while visualizing results of a query as data points on the map), no earlier research work has reported a methodology to generate regions to provide an overview of the underlying points for visualization purposes. This gap is addressed by providing the methodology to

provide cartographically improved results. As mentioned in the previous chapter, this research gap is based on the use case of a user querying a dataset to get an overview of the relevant results. The following investigations are provided and discussed further in Chapter 4:

- Providing the methodology of generating ROIs to represent POIs
- Evaluating POI density as the suitability measure for the generation of ROIs
- Examining the ROI generation process based on a combination of quantitative cartographic measures

### 2.5.3 Visualization of textual topics using matching between the topics and geographic objects

As presented in Table 2.1 and section 2.1.4, despite the fact that there has been substantial research on matching UGC to different sources and based on different granularities, no earlier research work has addressed matching between important topics and geographic objects. Applying the matching between objects and topics is advantageous in at least two ways. The first advantage is regarding visualizing topics in the form of abstracts. By performing matching, the topics become spatially more concrete. The second advantage is that applying the matching between objects and topics provides a ground for investigating further data sources by using the matched objects as a query. In addressing this research gap here, we focus more on the visual abstraction advantages.

With the assumption of having a list of matched topics and geographic objects, the next step involves working towards generating appropriate visualizations of those salient topic words. In the research reported here, we take the text content of UGC as a proxy to find which geographic objects are salient. In order to move to the visualization step, there is also a need to provide a relevant visualization methodology. Conventional ways of visualizing UGC contributions are generally in the form of charts and patterns placed as foreground data on the map. There exists a research gap to relate collective contributions (extracted topics) directly to the map content rather than visualizing them as a foreground layer. This gap is addressed by providing a methodology to match between textual topics and geographic objects. As mentioned in the previous chapter, this research gap is based on the use case of a user exploring a study area to find important geographic objects and regions around him/her. The following investigations are provided and discussed further in Chapter 5:

- Providing the methodology to match between geographic objects and textual topics
- Analyzing matching results quantitatively for the recognition of patterns
- Visualizing the topics based on the recognized patterns

# 3 Integration of semantic measures in map generalization

## 3.1 Introduction

Responding to the first research objective introduced in Chapter 1, the research reported in this chapter investigates the integration of knowledge derived from UGC contributions into the process of map generalization. The knowledge may then be used in different ways, e.g. when making decisions during the map generalization or adapting data collections to users' search terms and thematic interests. Here we work towards the derivation of semantic measures that will help us develop new generalization operators. We use OSM as an example of UGC, focusing primarily on the semantics that is captured in OSM features in form of tags, and how this information can be exploited. The main contributions of this chapter are threefold:

- We show how the user-contributed semantics contained in OSM feature tags can be exploited as semantic measures (to be used in the generalization process later)
- For the selection and aggregation generalization operators, we introduce changes in a theoretical level in order to include semantic measures
- We show generalization operator changes based on an applied example (LBS use case)

## 3.2 Methodology

The present study aims to fill the research gaps mentioned in the previous chapter. First, an extension of dice similarity (Dice, 1945; Markines et al., 2009) is introduced and then the usage of this similarity measure in map generalization is introduced afterward.

### 3.2.1 Semantic similarity based on OSM contributions

As mentioned, in OSM dataset tags are given in the form of keys and values. We introduce and use a semantic similarity measure (extension of dice similarity) which measures feature-to-feature similarity.

#### Preprocessing

In order to apply the feature-to-feature similarity analysis, there is a need to preprocess the data. The main step iterates over the features and removes tags that express spatial information and retains only tags that exclusively express semantic information. The logic behind this step is that the geometry of features contains sufficient spatial information and thus tags expressing spatial information are not needed and should be removed if the focus is on semantic analysis. Examples of such spatial tags are postal-code, addr:street and addr:housenumber. Besides removing tags with spatial content, other tags that carry information not relevant to the focus of this analysis have also been removed (e.g. contact information of the businesses). The full list of blacklisted tags is included in Appendix A.

#### Processing

With appropriate tags for each feature, we are ready to calculate the similarity between features. In order to calculate semantic similarity, several measures are commonly used by Markines et al. (2009): Jaccard, dice and cosine similarity. For Jaccard and dice measures,

having tags of the two features to be compared is enough. Therefore, one can calculate the length of both sides, their intersection, and their union. As cosine similarity is defined in a vector space, there is a need to know the dimensions of the space. These dimensions are determined by the union of tags of all features in the study area. Cosine similarity is also a frequency-based measure. In the case of OSM, each feature cannot have a key more than once (e.g. it is impossible to have memorial=statue and memorial=war_memorial), therefore it is impossible to measure the frequency of terms for each feature.

The above similarity measures are based on set-theory intersections or term frequencies, but if we have key-value pairs it is important to include values in the similarity measurement. Measuring similarity solely based on keys results in a misconception of feature matches and miscalculation of their level of similarity. In order to overcome such a problem a new measure is proposed here, KeyValue similarity, which can be seen as an extension of the dice similarity:

$$sim_{\text{KeyValue}}(X,Y) = \frac{\left(\frac{2|K_X \cap K_Y|}{|K_X| + |K_Y|} + \frac{|V_X \cap V_Y|}{|K_X \cap K_Y|}\right)}{2}$$

where KX represents keys of feature X, VX represents values of feature X. This equation normalizes the features' commonalities. While shared keys are normalized with the length of feature tags, shared values of those keys are normalized with the number of shared keys.

The proposed measure generally yields lower values than the other aforementioned measures. It reports a value of 1.0 only if all of the keys and values are equal, while the other measures report 1.0 only when the keys are the same (and not necessarily the values).

### 3.2.2 Semantic measures in map generalization

In this section, we study modifying generalization operators based on semantic similarity measures. Selection and aggregation operators are considered here. Our initial hypothesis is to apply two thresholds to similarity measurements: a lower threshold α and an upper threshold β, where a value of less than α is taken as dissimilarity, a value between α and β as similarity, and a value greater than β as a candidate for equality (i.e. two equal features). In other words:

$$s = Sim\,(X,Y): \begin{cases} s < \alpha \rightarrow dissimilar \\ \alpha \leq s \leq \beta \rightarrow \text{similar} \\ s > \beta \rightarrow test\ if\ X = Y \end{cases}$$

Equation 3.1 – Determining decisions based on object-object semantic KeyValue similarity

Based on semantic similarity score results, we have set our α and β to be the lower and higher sensitive values. More information is provided later in Section 3.3.2.

## Selection

The motivation for applying a selection is to reduce the number of features on a map, thus aiming for less data or less visual clutter. The filter criteria may be spatial (e.g. overlap or congestion) or semantic (e.g. an importance ranking or classification function deciding to retain or remove features).

Using the

Equation 3.1 similarity-based selection as defined here will take a feature as a representative (or search feature or exemplar) and decide about the similar and dissimilar features. Two approaches are possible: retain the similar features and eliminate the dissimilar ones; or take the feature as the representative and remove the similar ones as they are already represented. Both options are shown schematically in Figure 3.1. It should be mentioned here that as we are considering the properties (tags) of the search feature, this search feature should not necessarily be in the same study area, but rather its properties should be accessible.

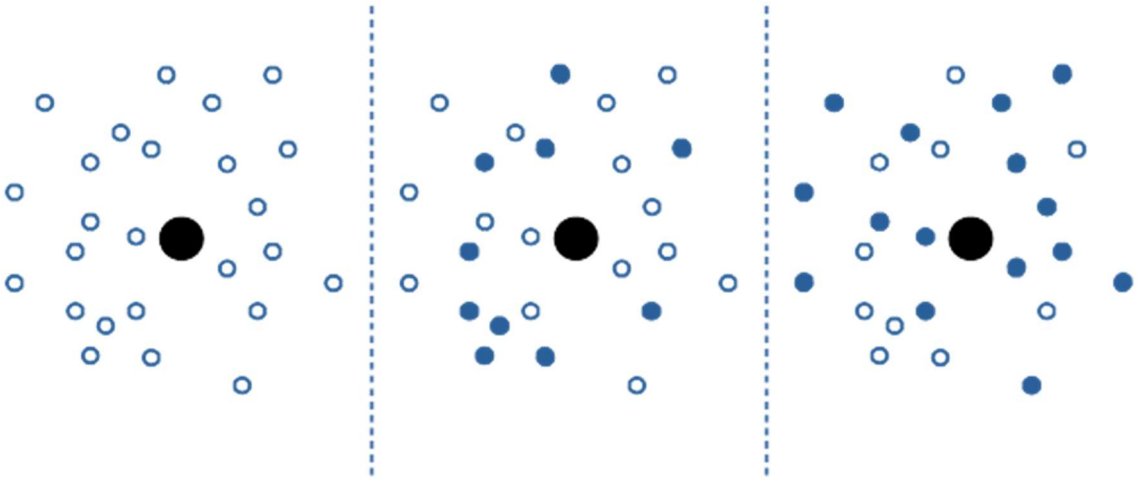

Figure 3.1 – Schematic illustration of similarity-based selection (search feature shown in black, selected features shown with filled-in circles, filtered features shown with hollow circles). Left: initial situation. Middle: the selection process has selected the semantically similar features and eliminated the dissimilar ones. Right: the selection process has selected the semantically dissimilar features and eliminated the similar ones.

## Aggregation

Aggregation is the operator that merges features in order to decrease the number of features or the detail of rendered features. A group of features would be aggregated if they are close enough to each other and have enough similarity to be taken as one feature. Generally, the aggregated feature represents the group of features that have been taken into aggregation. This aggregation process is done by considering both semantic similarity and spatial restriction. In the first step, a set of potential features to be aggregated is selected by applying the semantic similarity (

Equation 3.1) over all the features in a map area. In the second step, these selected features are aggregated based on how close they are, which might result in several clusters. This step can

be achieved by applying spatial clustering methods like density-based spatial clustering of applications with noise (DBSCAN, Ester et al. (1996)) and ordering points to identify the clustering structure (OPTICS, Ankerst et al. (1999)).

Take DBSCAN as an example, which has two input parameters:  (epsilon) specifies how close points should be to each other to be considered a part of a cluster, and minPts sets the minimum number of points required to form a cluster. Epsilon exactly addresses the spatial restriction. By setting an appropriate epsilon value, the selected features from the first step can be grouped into clusters. For each cluster, its members are aggregated to generate an aggregated feature. Features that do not belong to any cluster can be still visualized and shown on the map as individual features. Examples are provided in the results section.

A schematic example is given in Figure 3.2. The resulting feature of this aggregation can be placed on the anchor point (search feature if applicable), the centroid of the collection of features, or as a minimum convex polygon. An important constraint is to limit the process of finding candidates within a meaningful radius, as aggregation of very far features is not meaningful. In the case of visualizing the points and the new generated aggregated feature(s) on the same map, the symbology should be communicated to the user to highlight that one symbology represents single features and the other symbology represents an aggregated feature (multiple features).

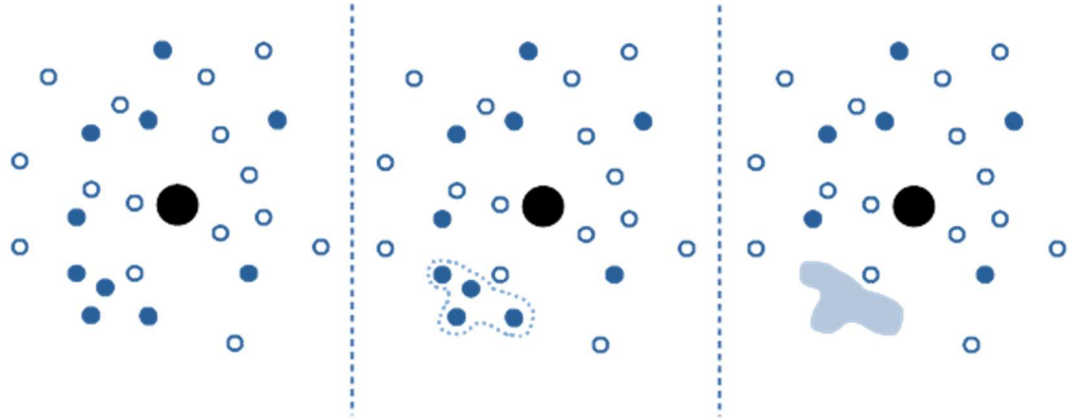

Figure 3.2 – Schematic illustration of similarity-based aggregation. Left: initial situation in which points are selected based on the similarity-based selection. Middle: detection of a spatial cluster of selected features. Right: visualization of the aggregated feature (along with other selected features).

## 3.3 Results

Here we would like to present a case study based on the methodology introduced earlier. For the sake of better understanding, we are presenting the case study based on a real-life example of how the map generalization process is modified in order to serve the needs of potential map users.

### 3.3.1 Test case

In a situation, our sample user, Paul, is interested in dining out. He is currently located at a certain coordinate (X,Y) and searches for the places around him which offer food. In this search, two different classes of constraints should be considered:

Spatial constraints: Where is he located now and how far away should the results be?

Semantic constraints: Which characteristics of the features are interesting for him?

Paul's query has been translated to a compound query, and a list of restaurants/fast food locations is returned after executing the query on OSM data. The query is based on the bounding box covering the metropolitan area of the city where he is at the time of querying (in this example in Barcelona, Spain). The following combinations of keys and values have been considered:

- combination of "amenity = restaurant"
- combination of "amenity = fast_food"
- combination of "amenity = food_court"
- "cuisine" as a key
- combination of "food=yes". This would fetch features that are not restaurants or fast food locations but offer food (e.g. pubs)

This combination has been manually selected based on analyzing OSM Wiki[10] and TagInfo[11].

After getting the initial results that contain a lot of possibilities (Figure 3.3), Paul is interested in filtering the results because successful filtering would help him to have less visual clutter and also make the decision-making process easier. A basic way to filter the results would be to consider Paul's location in the feature selection and map visualization. Applying a spatial filter would result in a situation similar to Figure 3.4. The spatial filter applied here is based on the travel time between his position and the position of the features.

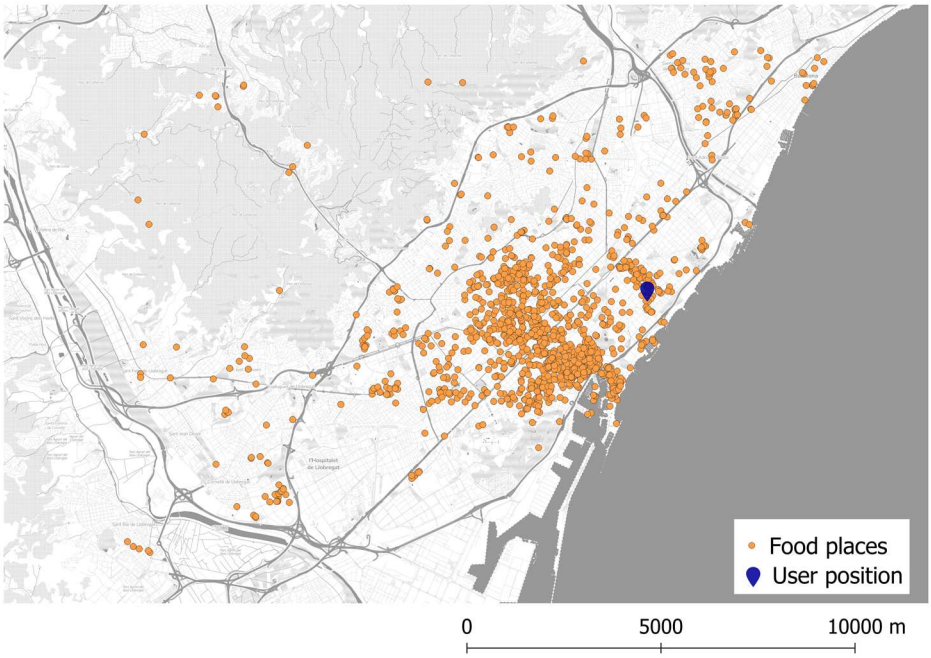


Figure 3.3 – Example situation of Paul searching for places around him that offer food in Barcelona. The dark blue marker is his location and the blue points are the places fetched from OSM.

[10] https://wiki.openstreetmap.org – accessed Nov. 2019

[11] https://taginfo.openstreetmap.org – accessed Nov. 2019

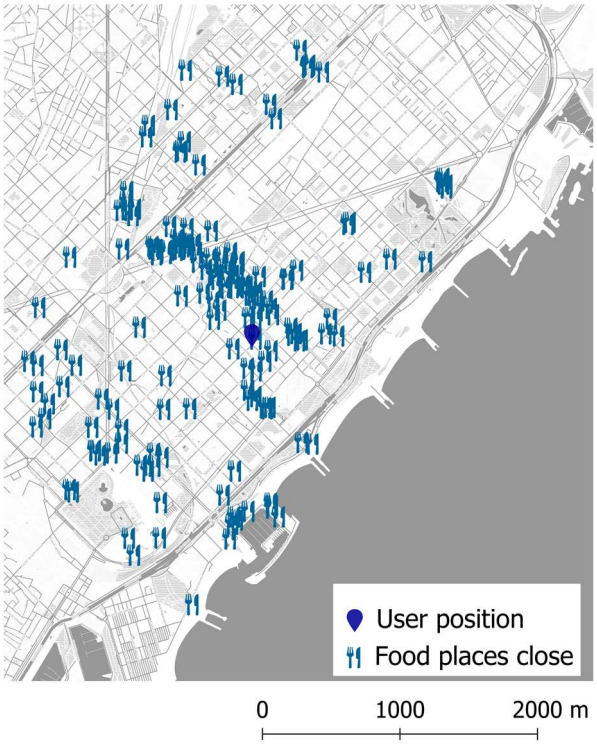


Figure 3.4 – Example situation of Paul searching for places around him that offer food. The dark blue marker is his location and the restaurant / fast foods around him have been filtered based on walking time (threshold set to 25 minutes).

In order to make the results closer to the user's needs, a meaningful step could be made in direction of the thematic constraints mentioned earlier. A way to retrieve more thematically relevant results for him would be to consider his previous successful experiences of the same activity (eating food in a restaurant / fast food). This means that the LBS will consider one of Paul's favorite places in order to find places similar to them. By applying this filter, he will see a more filtered version of the map. This would be similar to (Figure 3.5 left). In this map, the fetched features are first filtered spatially and then thematically based on their semantic similarity to an exemplar feature (Paul's favorite restaurant). The semantic similarity is measured based on the key-value pairs describing features based on the methodology introduced earlier in the methodology section. From this situation, Paul might take different actions. One option would be to apply a higher thematic filter (e.g. concentrating more on the similarity to the exemplar search feature) and keep the spatial filter, as is visualized in Figure 3.5, middle. Another option would be to consider places that are further from him but are still similar enough (e.g. keeping the semantic filter and apply a spatial filter with a higher threshold), as is visualized Figure 3.5, right. This situation is when the user is willing to travel longer in order to get more favorable results.

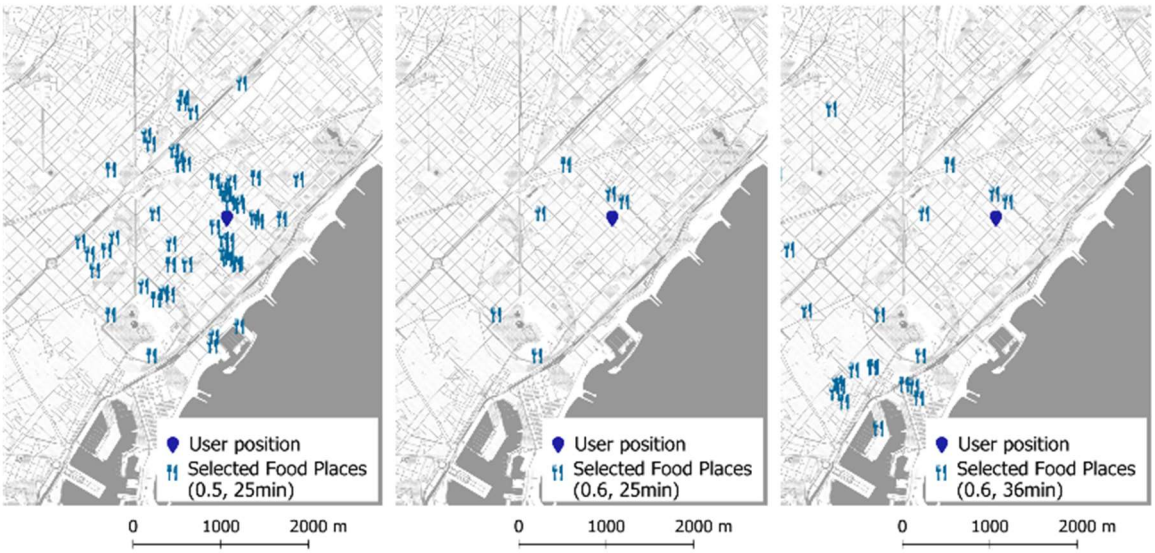


Figure 3.5 – Three situations of Paul searching for places around him that offer food. The dark blue marker is his location. Threshold values are as the following (semantic, spatial): left: (0.5, 25) middle: (0.6, 25) right: (0.6, 36)

Another process that would be considered in these situations is aggregation in which several map features are aggregated together. This becomes necessary when zooming out or by having many features on the map (which will necessitate having features being combined together). Two examples are provided in Figure 3.6. Both results are generated having the results of the selection (spatial and semantic) process as the input to the semantic aggregation. The aggregation needs to detect clusters of features in order to generate new aggregated features based on them. Therefore, DBSCAN (Ester et al., 1996) (MinPts= 3, Eps=100m) is applied. The output of the clustering would be considered as the collections of features to be put together as a new feature replacing them. In these examples, the geometries of the new features are generated after applying a convex hull around the detected cluster and then applying a buffer. The motivation of the aggregation in the example shown in Figure 3.6, left, has a large number of results, therefore replacing the aggregated features in place of the individual features would decrease the visual clutter. The example shown in Figure 3.6, right, is based on the situation when the user is willing to consider further places for more favorable options. This could be useful where an aggregated feature could be interesting for the user to pan and zoom in (reversing aggregation).

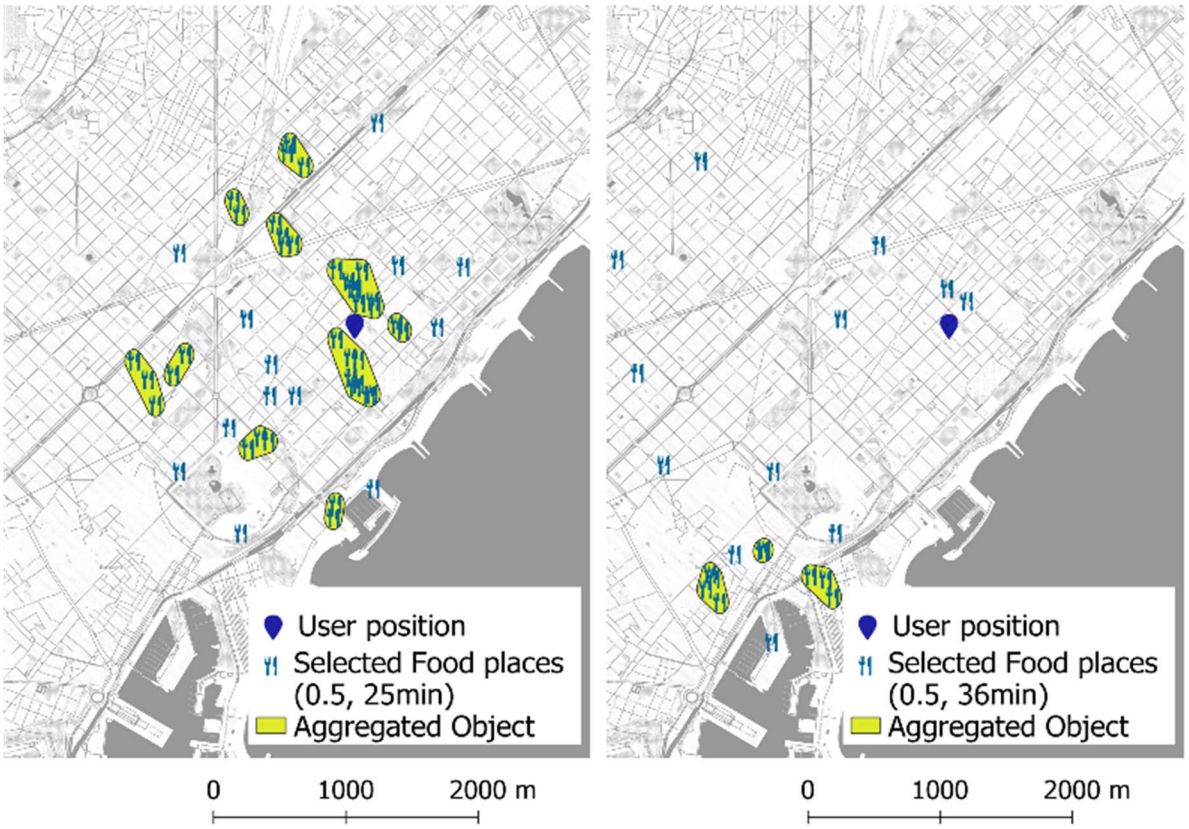


Figure 3.6 – Two situations of Paul searching for places around him that offer food. The dark blue marker is his location. On the left, the semantic threshold is set to 0.5 and the spatial filter is set to 25 minutes. On the right, the semantic threshold is set to 0.6 and the spatial filter is set to 36 minutes. The places were aggregated using the DBSCAN algorithm.

### 3.3.2 Threshold sensitivity

We have explored the semantic similarity measure introduced in this chapter regarding its sensitivity to variations in the threshold value. Based on data located in four different European cities, we have tested the similarity threshold (similar to the scenario introduced earlier in 3.3.1) and we have observed that the threshold shows sensitivity close to two values. First, a minimum value (0.2) which differentiates between the dissimilar objects and the objects that are slightly similar. We have used this value as our α (cf. Section 3.2.2). This threshold helps us to make a sharp differentiation between dissimilar objects and objects with a low similarity score. Second, the value (0.6) is taken as the maximum similarity, which helps us to differentiate between objects that have higher similarity values and the objects that are candidates to be checked for being copies (rather than different objects). We have taken this value as our β. When filtering out the duplicates, a similar approach to our latter case is reported in Novack, Peters and Zipf (2018) where the authors use a string matching score (and not a semantic similarity score) to decide whether certain matching candidates are duplicates or not. Figure 3.7 provides details of the similarity scores versus the number of objects having that similarity score. We have used the lower and higher sensitive values as our thresholds to produce the results presented in this chapter (Figure 3.3 to Figure 3.6).

The sensitivity of the threshold can be used to filter a large number of features, but at the same time, it limits the flexibility of selecting the number of features on the map.

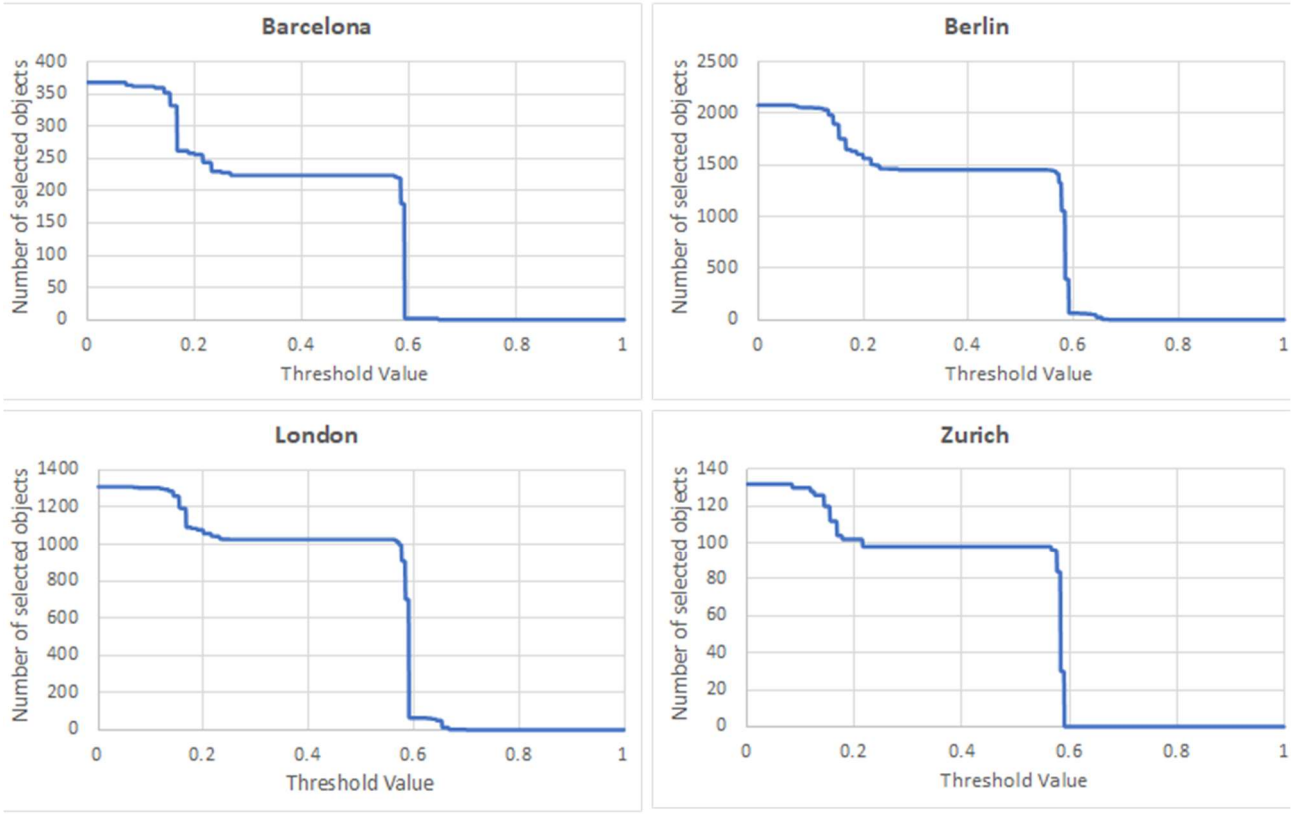


Figure 3.7 – Number of selected objects vs. threshold value in four different cities. The semantic similarity threshold shows sensitivity around 0.2 and 0.6. We have used these values as our thresholds $\alpha$ and β.

## 3.4 Discussion

The research reported here has focused on the adaptation of map generalization process in order to consider knowledge hidden in users' contributions (in UGC). Representation of users' knowledge has been considered in the form of tags (a combination of keys and values). As mentioned before, the usage of certain keys and values are considered as users' agreement on a feature's attributes and definition. The definitions have been taken into account when processing the information to be visualized on the map. Based on the semantic analysis applied on the features attributes, the process has been informed (in the form of parameterization).

The analysis was based on a measure introduced in this chapter and two generalization operators. The idea behind this proposal is to include knowledge derived from user-generated data in the process of map generalization. When analyzing the features space, we considered a search feature (or exemplar feature). Firstly, this could be seen as a feature of interest where the user's query takes into account what a user considers as favorable. Map features go through a process of being semantically examined to find to which extent they are similar to the exemplar feature. The exemplar feature can be a feature the user likes, a feature being recommended to him/her, or even a feature found elsewhere with a positive reputation. A second way of considering the exemplar search is the fact that the search feature contains tags

that all together form the query. Therefore, one could modify this query by addition or removal of tags in order to make it more specific or generic, respectively.

Applying the methods introduced here, modification of the behavior of generalization operators (selection and aggregation) has been formalized and tested (in an example case visualized in Figure 3.5 and Figure 3.6). By the inclusion of semantic similarity measures, the similarity-based selection process fetched different sets of features (which are closer to the exemplar-based on their tag definitions). The results are promising in different ways. Firstly, the selection manages to resolve the visual clutter and results in a smaller collection of more relevant results. Secondly, the operator fetches features from different categories. Taking the test case as an example, instead of merely selecting restaurants or fast foods (a single class in the dataset), the selection also fetches features of other classes which include tags related to food (the user's intention). In some cases, a feature of other class (e.g. a pub) is favored over a restaurant, as it has key-value combination(s) that lead us to more semantic proximity (e.g. cuisine=regional). Depicting results from other classes is a useful behavior where the user is interested in map features which can be categorized in different classes but all afford similar activity to take place. The similarity-based aggregation operator has resulted in combinations of features to be merged into one feature (but it remains to be discussed which properties the new generated feature should inherit). The results of this operator are interesting, as, firstly, they could contribute to resolving visual clutter. In addition, the aggregated features are helpful in visualizing smaller scales where the user is interested in panning and zooming in an explorative way.

## 3.5 Conclusion

In this chapter, the inclusion of a semantically driven measure in map generalization has been suggested, formalized and tested. The similarity measure is based on the semantics of geographic features hidden in key-value form and is measured comparing study area features to a reference feature. The application of this measure in two map generalization operators has been introduced and tested.

We believe that this approach has the potential to be extended in different ways and also could be related and/or combined with other relevant research efforts. Firstly, there is room for investigating the relationship between the semantic similarity measures and the number of map features selected based on semantic similarity selection. This would lead to a less sensitive and more efficient method of parameterizing the operator. Another way to enrich the methodology would also be to consider a weighing regime based on assigning different weights to different tags (based on frequency, hierarchy or prominence, etc.).

Secondly, as mentioned in the discussion and the test case, the mapping between OSM tags, on the one hand, and user intentions and activities, on the other hand, was done manually. We believe there is room to investigate activities related to map features and use them in forming the pre-processing phase. This has been partially done and will be presented in Chapter 5, where we have provided a mapping between keywords (extracted from textual UGC content) and combinations of OSM key-values.

Thirdly, as mentioned earlier, the research reported here has some relationship to what has been done earlier under the umbrella of GR. One way to consider the situation is to assume our research as a plug-able component to a GR system. From another perspective, a GR system could be seen as an input to the methodology introduced here. An output of the GR system is a ranked list of geographic features based on their relevance to the query. The ranked list generated based on the GR output could be seen as the input to our generalization operators.

With such a setup the operator takes the GR-based ranking as the similarity measure input to the operator parameterization. We believe there is room for such investigations in both ways.

The work presented here contributes mainly to research about enabling a map generalization process to include semantics hidden in UGC. Map content filtering was based on different factors, including similarity to a query. The inclusion of semantics paves the way of bridging the gap between what is available in the dataset and what the user is after. This is in line with the direction of on-demand mapping, where we are interested in modifying the map content according to users' requests and intentions. It is also one step further in moving the generalization process to be able to reflect people's (i.e. users') definition and perception of geographic phenomena.

As mentioned here, and in regard to the user's interaction with the cartographic results, a probable user action is zooming out. In such situations, the user is interested in getting an overview of the data of the study area. Chapter 4 provides the methodology of generating overviews of the study area by using keywords as search queries. This is then followed by Chapter 5, where the visual representation of the data is based on topics extracted from the data itself (without the usage of keywords or users' favorite objects, in contrast to Chapter 4 and this chapter respectively).

# 4 Generation and generalization of regions of interest (ROIs) based on user-generated points of interest (POIs)

## 4.1 Introduction

In the previous chapter, we proposed the methodology of adapting generalization operators to include semantics based on the user's list of favorite objects. In this chapter, we follow by providing the methodology to visualize areas rather than points of interest (POIs, the results of a textual query). In the visualization of POIs, having a large number of points in a dense region can cause problems such as visual clutter and confusion in making decisions. This problem is more obvious when considering mobile phones (with small displays) as the main media of delivering conventional digital maps. An example is shown in Figure 4.1 based on the mobile app Foursquare, where the user has searched for a specific type of food (burger). The search results are shown to the user. Because of visual clutter, this map at this scale does not help to make a decision, but the user can observe some regions that offer more options. Searching for another word (dance) as a probable follow-up activity, the second figure is shown to the user. It is visible to the user that there are regions which include places offering both query terms. These observations are based on densities of places relevant to user search terms. It would be useful if the user could see *regions* (a group of POIs close to each other to be taken as one aggregated object) offering what he/she is searching for – or, optionally, regions offering one search term which is close enough to regions of the other search term.

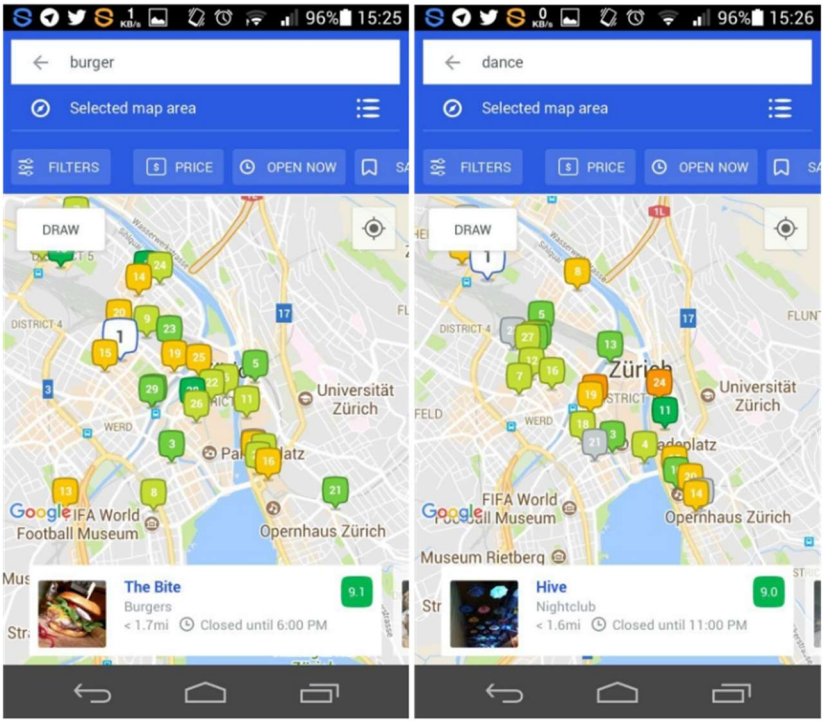


Figure 4.1 – Searching for two different terms on the app Foursquare. The results are shown in the form of points, each representing an individual object (venue entry in Foursquare database).

However, these methods do not help to get more information about the underlying POIs and the geographic structure enclosing them. In order to provide a solution, it is feasible to investigate ways of generating *regions*, which represent an aggregated view of a group of POIs, thus reducing clutter and complexity. This has been named before as regions of interest (ROIs), which are used in order to generate aggregated objects of interest. Different methods of generating ROIs have been reviewed in Chapter 2.

Responding to the second research objective introduced in Chapter 1, this chapter focuses on generalizing POIs in the form of ROIs to support the overview of a study area: not only to represent a higher density of POIs (based on user contributions), but also to handle different scales and different map themes (with each search term taken as a map theme). As it will be presented later, points with a clustering behavior are needed to apply the method introduced here. It should be also mentioned here that the switch from point representation to region representation should be reversed when either the user zooms in more than a threshold (i.e. not an overview map anymore) or if the density of the points is low (i.e. the points are not clustered).

Based on the second research gap introduced in Chapter 2 and by fulfilling the research objective introduced in Chapter 1, the contributions of the research reported in this chapter are as following:

- Suggesting and parameterizing the method to generate ROIs (with respect to the user's current map view)
- Evaluating ROIs generated based on measures of map readability and spatial distribution

## 4.2 Methodology

### 4.2.1 Overview

In order to check whether generating ROIs is relevant, we check the clustering behavior of the points using the L-function. In the case of having a clustered behavior, the ROI generation method is triggered.

In the next step, POIs are taken as input to a kernel density estimation (KDE) function. Important parameters in the KDE step are the kernel function and bandwidth. Different density functions can be considered (e.g. Gaussian, cosine or Epanechnikov). The kernel estimation results (in the form of rasters) are normalized to the range of [0..1].

In case of having more than one search term, there is a need to overlay the resulting rasters. This optional step is done after the generation of KDE surfaces. Aside from the method to combine the rasters, different weight scores can be assigned (e.g. in order to reflect importance). By having either a single KDE surface or the result of combined overlaid KDE surfaces, it is feasible to derive regions by generating contour lines. Contour values are generated using a threshold in the range of [0..max].

Contour lines are then converted to polygons. Polygons are then checked for validity. If a polygon violates a validity criterion, there is a need to fix that violation by changing the threshold value (while keeping the density bandwidth). After repeating the process of generating polygons based on different thresholds (and in the case of achieving a situation without violating criteria), the set of polygons is taken as the valid representation of the POIs.

In the case that the clustered nature of the points does not hold (e.g. by zooming in and having a lesser number of points in the map extent), point visualization would replace the region visualization method. As mentioned before, this is evaluated using L-function. Figure 4.2 shows an overview of the methodology in a visual way.

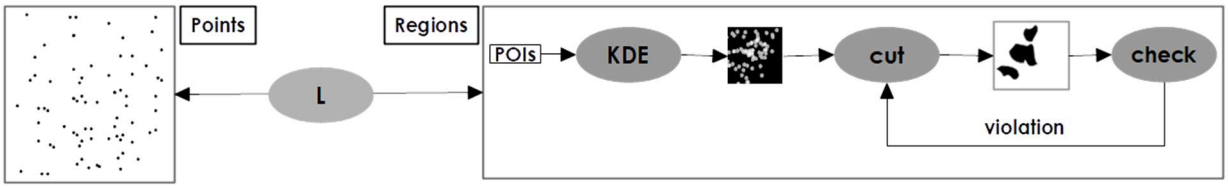


Figure 4.2 – Overview of the methodology. The switch between point and region visualization is controlled by applying L-function.

The following pseudo-code provides more detail on how the ROI generation works (the right block in Figure 4.2). The method needs input values (e.g. a query word); example values are provided as inputs to the pseudo-code. The implementation of the methodology has been based on Python programming language and additional libraries from QGIS[12] and GRASS GIS[13]. Further details of each step are provided in the next sections.

12 qgis.org - accessed Feb. 2021

13 grass.osgeo.org - accessed Feb. 2021

**Inputs:** query_words = ["food", "dance"], bandwidth = 416m, kernel_func = triweight, threshold = 0.3, min_area = 0.2, max_area = 0.8

**Output:** rois

```
for query_word in query_words
        points[query_word] = query_db(query_word)
        l = riply_l(points[query_word])
        if l>riply_l.clustered: /*checking the L function*/
                raster[query_word] = KDE(points[query_word], kernel_func, bandwidth)
        else continue
        raster = normalize(raster)
        violation = true
        while violation:
                polylines = threshold_cut(raster, threshold)
                polygons = polyline_to_polygon(polylines)
                alpha = ((min_area|max_area) / polygons.area.sum)
                new_threshold = calculate_curve_fit(threshold, alpha)
                threshold = new_threshold
                violation = (polygons.area.sum < min_area) OR (polygons.area.sum > max_area)
        rois.add(polygons)
return rois
```

### 4.2.2 Changing the visualization method

The research reported here proposes a methodology to replace the visualization of POIs in the form of points, with the generation of ROIs based on the density of data points. This change in the form of visualization is meaningful when the density of the points in a region (i.e. map extent) is high and the points are very close to each other. In order to make this decision based on a quantitative measure, we use L-function (based on Ripley's K-function (Ripley 1977)). This function is a distance-based point pattern analysis method, which normalizes the value of the K-function and helps us to find out distances that comply with a clustered behavior (O'Sullivan & Unwin 2003). Selecting appropriate distances is based on the calculated value of the L-function in comparison to the theoretical CSR (Complete Spatial Randomness) and the simulated values for L. For each test distance (d), if the value of the L-function is more than a maximum value of the simulated envelope (based on the assumption of CSR), the points show a significant clustered behavior. Figure 4.3 shows exemplary results based on our input data. In our test cases, the whole study area (e.g. top row in Figure 4.3) was significantly clustered and therefore valid for applying the methodology reported here. Based on the same data, different map extents show different behaviors: clustered behavior, partial clustered behavior, and not clustered behavior. In the case of a partial clustered behavior (e.g. bottom right in Figure 4.3), the generation of KDE surfaces with a bandwidth greater than distances that comply with clustered behavior is avoided (e.g. 750m for the example in Figure 4.3, bottom right). As it can be observed, "dance" POIs in map extent C (bottom right) show a significant clustered behavior up to ~750m, where the POIs for "burger" in the same extent

(bottom left) show a clustered behavior (though not significant) up to the same range. Figure 4.13 visualizes the POIs of the lower row.

Obviously, this process can be reversed (i.e. switching back to a point visualization setup) when the points do not show a clustered behavior. This happens naturally when the user starts from the overview of the data and zooms in.

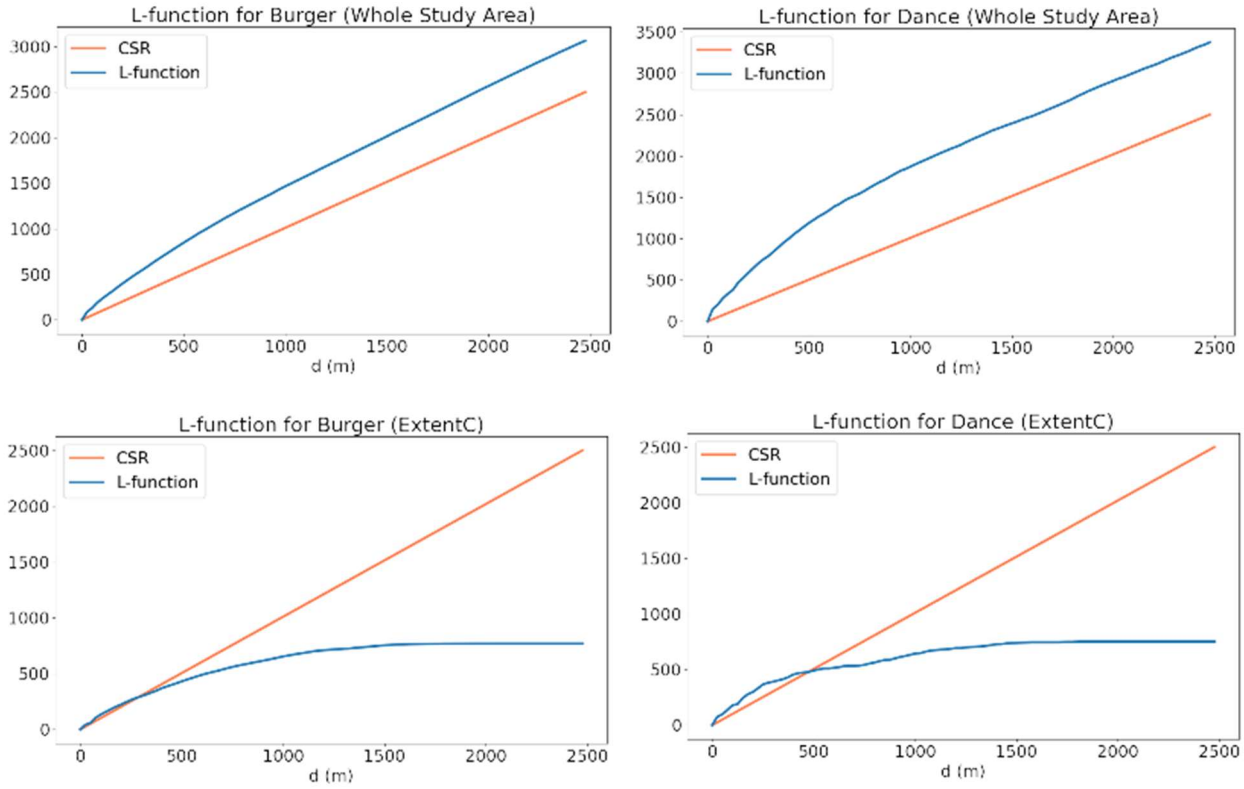


Figure 4.3 – Investigating the cluster behavior of POIs using an L-function for two query words based on the whole study area and the map extent. The L-function is shown in blue and the theoretical CSR is shown in orange. The dashed lines show the values for the simulated envelopes. The graphs at top left and right row show the L-function for the whole study area (top left for the query "burger" and top right for "dance"). The bottom row shows the L-function for map extent London C (more information provided in Section "Source data and study areas").

### 4.2.3 Methodology considerations

Intermittent steps in switching between point visualization to region visualization include the generation of density estimation surface(s) using KDE. Alternatively, one could base the method on generating polygons by firstly detecting POI clusters and then extracting a shape (e.g. convex hull) for each cluster. Here we provide some reasons to support our decision for using a density surface. The first aspect to consider is the membership flexibility. Cluster membership is typically a binary state where density surfaces provide large number of state possibilities. Parker & Downs (2013) have reported this flexibility to be an advantage in the visualization of geographic data. Another important aspect is that methods based on KDE typically result in smoother shapes which are generally more desirable (De Berg & Speckmann 2011). Besides flexibility and smoothness based on the continuous nature of (KDE) surfaces, using them enables us to combine different datasets (e.g. results from different queries).

Considering analysis based on clustered behavior of points, alternative methods based on the detection of clusters can be applied. For example, other studies have used density-based clustering algorithms (e.g. DBSCAN Ester et al. 1996) or derivative/improved alternatives) and have reported their results in form of polygons (using a concave hull, convex hull, alpha shape or similar methods) based on clusters as the results (Galton & Duckham 2006). As expressed before, the flexibility provided by continuous intermittent steps is the main reason for selecting KDE surfaces. Density-based estimations tend to be criticized as they do not distinguish between data points and outliers. Although this could be potentially a disadvantage in our methodology, because we are mostly looking at a subset of the data (already fitted to the current map view), this effect is not large.

There are also practical reasons for basing the method on generating density surfaces. Considering a multi-scale visualization scenario, it is important to relate between POIs and ROIs. When using an estimated surface (and contour lines), it is more feasible to quantitatively measure the relationship between POIs and ROIs (in the form of generated polygons). This is done by measuring the (closest) distance between the POI and ROI polygons. The surface resulting from a kernel density function reports the intensity over the study area (thus in certain coordinates, e.g. POIs). This is helpful when considering relating POIs to ROIs (e.g. switching from ROI view to POI view). When using clusters, typically we can only test whether the POI belongs to a cluster or not (or to which cluster does the POI belong). Another point is to consider the importance. In order to include the importance of terms as a parameter (e.g. when expressing the importance of the generated region being close to public transportation), we can increase/decrease the threshold of contour lines, something difficult to implement in a continuous manner when considering a clustering approach.

In our suggested methodology, the KDE surface is used to generate the polygons as regions representing the points. This step is helpful to enable the user to make decisions (e.g. to visit a venue or not). Earlier studies found that a combination of KDE surfaces and KDE contour lines is helpful for communicating geospatial data in an exploratory manner (Gibin, Longley & Atkinson, 2007; Yujie Hu et al., 2018). The conventional usage of KDE surface visualization is in risk or exposure analysis (Chainey, Tompson & Uhlig, 2008; Hart & Zandbergen, 2014; Yujie Hu et al., 2018). But in our scenario, the focus is on making decisions that are mostly discrete decisions (selecting between objects, in other words, to decide to visit a venue or not). Therefore, the discretization step is needed to help the user in the decision-making process. In summary, for our objective and based on the mentioned points, using kernel density surfaces and the discretization is selected based on theoretical and practical reasons.

Another important point to mention before presenting the details of the methodology is the comparison of the method proposed here with the conventional marker clustering method (commonly used in web and mobile mapping). As visually depicted in Figure 4.4, marker clusters represent a group of markers close to each other by showing the number of elements per cluster and placing the symbol on the cluster center. As it can be seen, the ROI generation result(s) tend to provide regions that not only can have an arbitrary shape (to be used in finding the region at other scales), they are also able to cover most of the points that have contributed to generating them. In marker cluster representations, however, the cluster markers are placed at the midpoint of each cluster. This tends to result in cases where the marker representing a particular cluster is far from most of (or all) the points. This will end up with the users zooming into clusters but simultaneously losing the connection with the surrounding points, as cluster centers may 'jump' between scales, which is challenging to the user. We take this specific shortcoming of this particular, commonly used visualization method as both a motivation and baseline for developing the methodology proposed in this chapter.

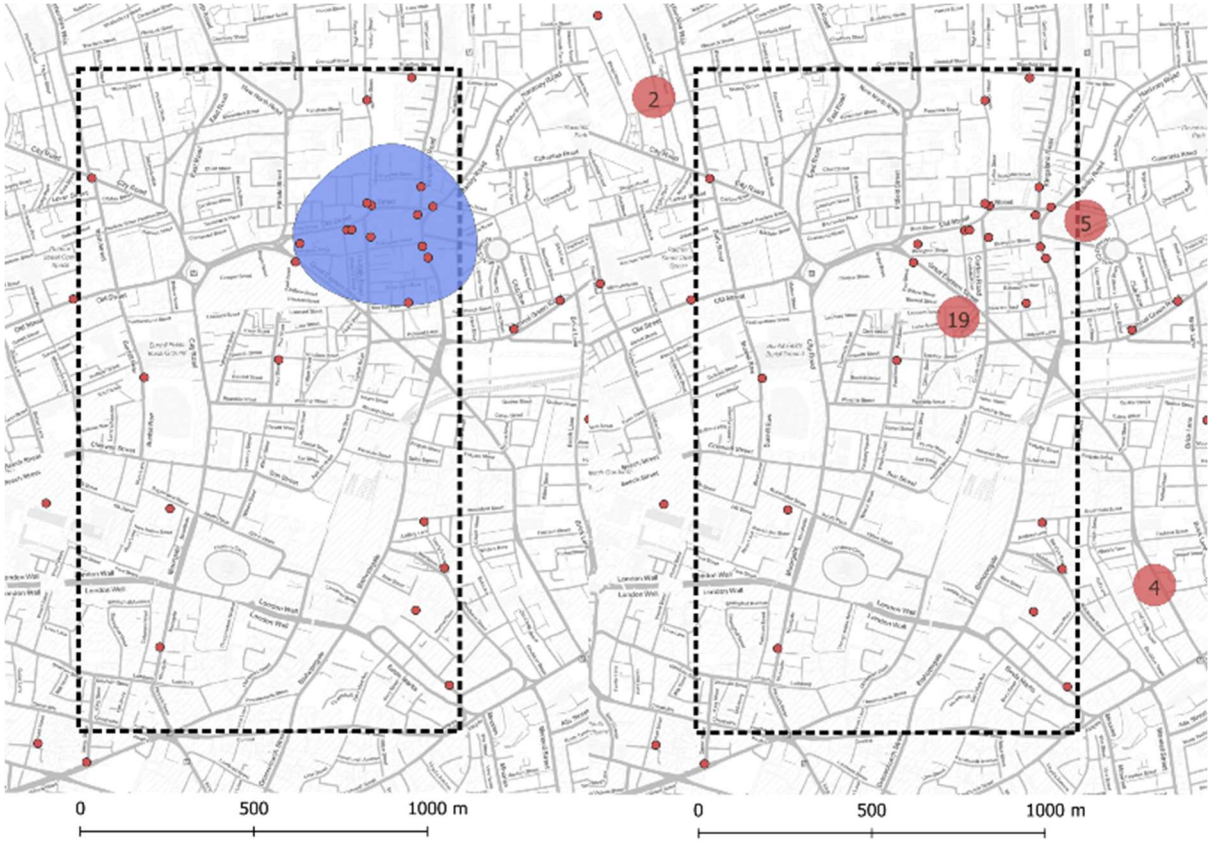


Figure 4.4 – Comparing the ROI results (left) with the marker clustering results (right) for query word "dance".

### 4.2.4 Kernel functions

When applying density estimation, different functions are typically considered. An important factor in selecting the estimator is the phenomena being estimated. But it has been also reported that the selection of kernel function does not have an effect on the reported results (Calenge 2011; Wand & Jones 1995). In the investigation reported in this chapter, we have focused more on "Triweight" function. This decision is based on the reason that the smoothing effect stops at *1*bandwidth* from the kernel center; therefore, it offers better control over results (when parameterizing the bandwidth).

### 4.2.5 Overlay method

In case of having more than one search term (e.g. when the user searches for more than one search term and therefore more than one estimated raster surface is generated), there is a need to combine the resulting rasters. In order to do that, different raster operations that take input values from the raster layers can be used to combine estimated raster values. Other operations such as the conversion of the values into binary values are also feasible.

We have used a weighted sum of input rasters using the following formula:

$$V_{i,j} = \frac{\sum_{l=1}^{L} F_{i,j,l} w_l}{\sum_{l=1}^{L} w_l}$$

where $V_{i,j}$ is the raster value at coordinate $i, j$ and $w_l$is the weight for the set for raster layer $l$. $F_{i,j,l}$ is the function value at coordinate $i, j$ for layer $l$. Normally $F_{i,j,l}$ is set to be the raster value. Later in Section 4.3.2, we provide two examples of using addition (+) and subtraction (–) operators. By using addition, the method provides the presence of both query words. By using subtraction, presence of one query word and lack of the other word is desirable.

### 4.2.6 Parameterization

The results are affected by setting parameters. Among these parameters, the most important numerical parameters are *density bandwidth* and *contour threshold*. In order to get valid and meaningful results, it is important to parameterize these two parameters based on the objective. Here we consider two ways of deriving parameters: ***query*** and ***cartographic*** approaches. When considering query constraints, we focus more on having appropriate results based on user queries; when considering cartographic constraints, the focus is more on valid results from the cartographic point of view.

#### Density bandwidth

In order to generate ROIs, we aim to generate regions representing a high density of the desired points related to the search term (e.g. "parking"). Selecting the bandwidth is important as it can be related to the maximum distance between two points to be seen as members of the same region. The effect of selecting different bandwidths is shown in Figure 4.5. The selection of the bandwidth affects the number, shape and size of ROIs.


Figure 4.5 – The effect of different kernel bandwidths(bw). Six different bandwidths are shown. The contour threshold value is set to 0.4 and the POIs are based on the query word "beautiful".

We consider two ways of deriving the bandwidth parameter. When considering a **cartographic** method to derive the bandwidth parameter, a mapping between bandwidth and map extent is needed. When the map user is looking at the map, the generated ROIs should not be too big to cover the majority of the screen, but they should also not be too small. Considering a map extent that covers an area (Figure 4.6), we can derive a bandwidth that generates results that are appropriate for the current map extent view. Here we assume that having regions (similar to the blue ellipse in the figure) that cover a certain part of the screen (e.g. 25%) is a compromise that fulfills number and area limitations (compromising for too many regions with a small area and vice versa). With such an assumption, the bandwidth value is set to be the average of the semi-minor and semi-major axes of this ellipse:

$$h = \frac{a + b}{2 * n}$$

where $h$ is the density bandwidth, $n$ is the number of ellipses to fit, $a$ is the current map view width and $b$ is the current map view height.

When considering a **query** approach, the bandwidth is set when submitting the query. Again, we aim for ROIs with POIs close enough to each other to be considered in the same region.

Having a user in mind, this distance can be seen as the distance a user is willing to move between points (i.e. close enough to be considered in the same region). This value can be expressed either as a distance or a temporal constraint. The temporal constraint can be transformed into a distance constraint after making the mode of transportation clear, thus indicating an average speed (e.g. 10 minutes of walking when considering 5 km/h as the average speed results in 833 meters). If this is taken as the maximum distance the user is considering, half of this constraint should be set as the bandwidth value (i.e. 416m in our example). Half of the distance is taken as the minimum distance so that the kernels of the two points will touch each other (therefore to be taken as the same region).

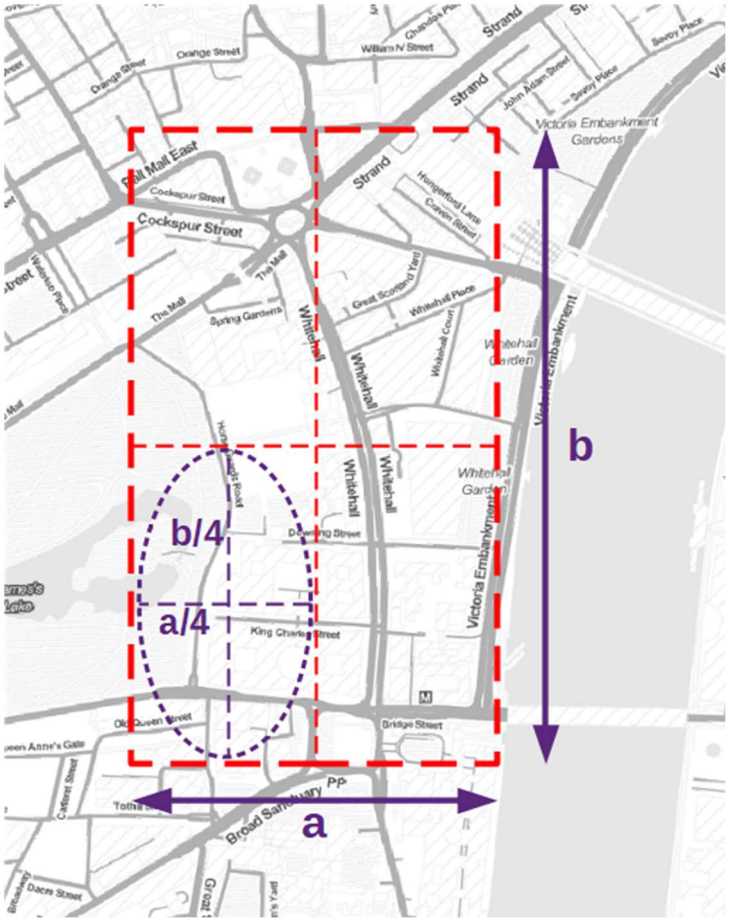


Figure 4.6 – Limiting an ROI inside a map extent (dashed rectangle). This figure shows the case of n=4.

## Contour threshold

The cutoff value for the generation of contour lines is set to cut in the normalized range of [0..1]. This proportion is taken as density importance of the region(s) being generated. If a high certainty (therefore a higher density) is desired, then the threshold value is set higher. Figure 4.7 shows the concept in a three-dimensional view. The threshold plane extracts the regions that have a value more than the threshold. Figure 4.8 provides an example of different contour values based on the same data and the same bandwidth but with different contour thresholds. The parameter also affects the number, size and shape of ROIs.

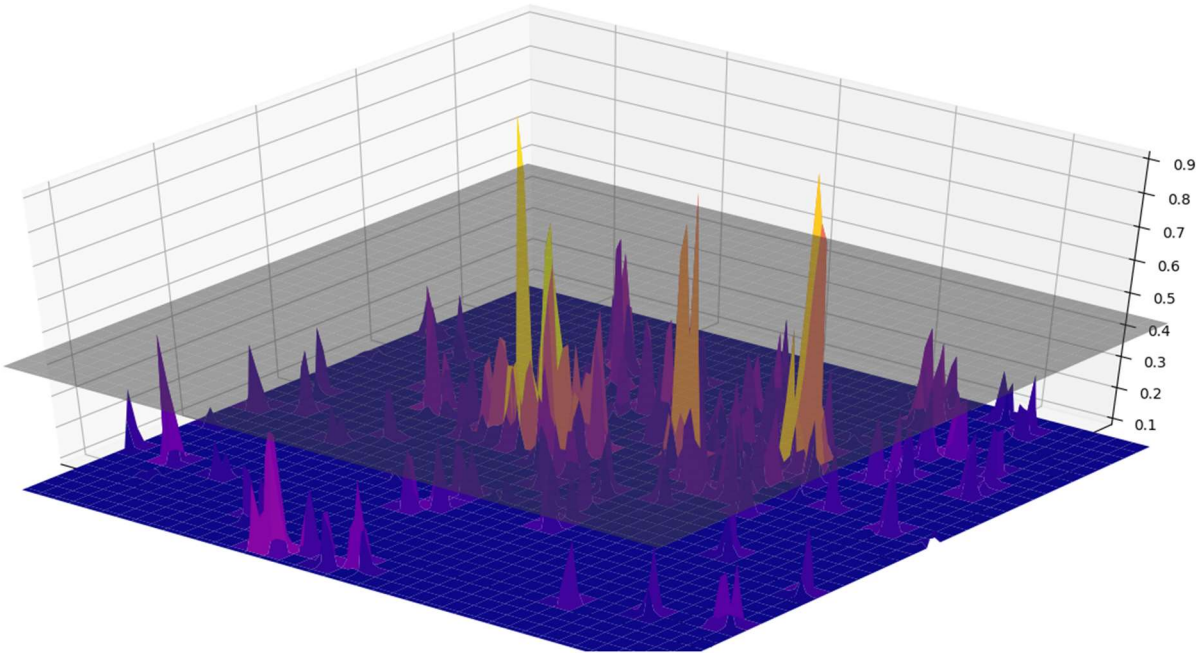


Figure 4.7 – Cutting the KDE surface with a threshold (0.4 in this illustration)


Figure 4.8 – The effect of different contour threshold cuts. Six different thresholds are shown. The bandwidth is set to 350m and the POIs are based on the query word "beautiful".

Similar to the previous section when considering contour thresholds, there are two ways of deriving the parameter. When considering a **cartographic** method to solve the problem, again

the interest is towards getting cartographically appropriate results. The contour threshold has also an effect on the size of the ROIs generated. In order to get appropriate ROIs, the area of the generated regions is checked to ensure that it is not smaller or bigger than area thresholds (e.g. between 5% and 20% of the map extent). The following condition is checked for each region:

$$A_{min} \leq A_{ROI} \leq A_{max}$$

In the case of violation, the contour threshold is changed to generate new regions. If the area is smaller (or bigger) than the minimum valid area, in order to get a bigger (or smaller) region, the contour threshold would be decreased (or increased). Based on the non-linear nature of kernel density functions and also the effect of close POIs interacting on each other, threshold changes are not linear. In order to solve this problem, we use curve fitting to determine threshold changes. Figure 4.9 shows an example. Observations of the threshold are generated by testing different threshold values (e.g. [0.0..1.0] with 0.1 steps). For each observation, the sum of the area generated for the whole study area is divided into the maximum area (whole map area). Using a non-linear least squares curve fit based on an exponential function on the observation points, we have now a function estimating the relation between area growth /shrink and the different threshold cuts. Two points should be mentioned here. The first point is that this solution is an estimation and leads us to an estimated value that can be optimized by repetitions. Another point is that the fit curve is valid for the whole study area and not for individual map extents; therefore, it leads to estimations. In our tests, we have reached an answer with a maximum of five repetitions (exemplary results are presented later in Section 4.3).

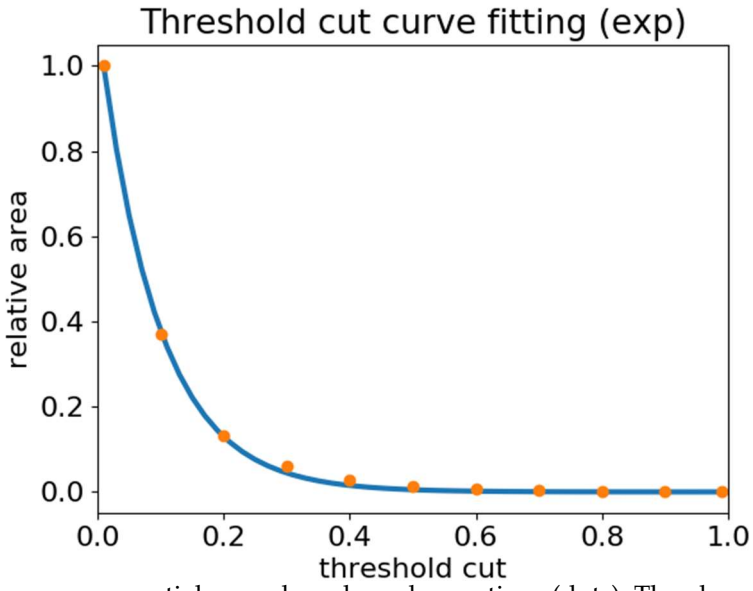


Figure 4.9 – Fitting an exponential curve based on observations (dots). The observations are generated based on the whole study area. Observations are made by generating ROIs for thresholds of [0.0..1.0] with 0.1 steps. The relative area is calculated by dividing the sum of ROIs area to the maximum area.

When approaching the problem from a **query** point of view, the solution to the challenge is to adjust the threshold based on the user's desired importance/confidence for that query. It should be mentioned here that based on experiments, low (less than 0.10) or high (more than

0.85) thresholds result in regions that are not very meaningful (because of being too big or too small respectively).

### 4.2.7 POI-ROI relationship

It is important to be able to link the POIs to the ROIs they belong (are assigned) to. Besides evaluating the methodology, this is also useful to support switching between point and region visualization modes. This could be seen as a transition step between regions to points (e.g. when the user zooms in). In order to measure this relationship, two different methods are considered:

- Distance: measuring the closest distance between the POI and the closest ROI
- Density estimated value: checking the value of the overlaid raster at the POI coordinate helps to be able to compare to which extent the POI is close/far from the ROI

By setting measures as thresholds, POIs that are close enough to the ROI would be stored as related to that ROI. Supporting the objective of switching from an aggregated view (ROI) to an individual view (POIs), the ROI-POI relationship is used to replace the ROI with the relevant POIs. An example is shown in Figure 4.10 where POIs close to a region (a distance less than 300m) are taken as related POIs.


Figure 4.10 – Finding related POIs. POIs visualized in blue are close (closer than 300m) and are therefore taken as related to generated ROIs. POIs visualized in light red are far (farther than 300m) from the ROI.

## 4.3 Experiments and results

This section includes experiments highlighting the methodology introduced in this chapter. Firstly, the source data and the study areas are introduced and the effect of different

parameterization of the method is shown consequently. Secondly, an example use case is provided in which the method is presented in an applied manner. The results of the evaluation of the methodology are provided in Section 4.4.

### 4.3.1 Source data and study areas

In order to query for textual terms, we evaluate our method on two datasets (Foursquare[14] and Yelp[15]) and two different cities in the UK (London and Manchester).

#### Foursquare

Foursquare provides information on venues in urban areas. Besides searching for venues based on a query and/or location, the users can also share their experience in the venues (by providing photos, ranking, checking-in, etc.). The users can interact with the service through web and conventional mobile platforms. The service also provides a means for venue owners to provide information about their venues (contact information, opening hours, etc.). Interested users can access the data through different ways such as RESTful APIs and also SDKs.

The venues dataset provides a means to search for venues based on different query parameters (such as name, category or location). By using a rectangular cell grid over the whole study area, we queried the dataset through Foursquare API for each grid cell. Cell data were aggregated and stored in GeoJSON format and later were used to test the methodology. In our Foursquare dataset for the city of London, we used a collection of 56,822 venues based on Foursquare data from February 2018. An important data aspect for each venue is the tips written by users. We used tips in our queries to retrieve tips that include the query word. An example tip about a venue in London reads, “Be ready to dance and enjoy great music” which included the search query “dance”.

#### Yelp

Yelp is an online directory that provides data on businesses in urban areas. Similar to Foursquare, Yelp also provides a means to interact with data as an end-user in the form of searching for businesses (e.g. by category or location). Information on business also includes information that has been shared by users (e.g. reviews, ranks and photos). The service is available both from the Web and from apps on conventional mobile platforms. Business owners can also use the service to input/update the information about their businesses. Interested users can access the data through different ways such as RESTful APIs and GraphQL endpoints.

The businesses dataset provides a means to search for businesses based on different query parameters (such as name, cuisine or location). By using a cell grid over the whole study area, we queried the dataset through Yelp API for each grid cell. Cells data were aggregated and stored in GeoJSON format and later were used in the methodology. In our Yelp dataset for the city of Manchester, we used a collection of 1508 businesses based on Yelp data from July 2019. An important data aspect for each business is the review written by users. We used reviews in our queries to retrieve tips that include the query word. An example review about

[14] https://foursquare.com – accessed Aug. 2019

[15] https://www.yelp.com – accessed Aug. 2019

a business in Manchester reads, “Burgers are a bit small but they’re priced accordingly. Bun was toasted a bit too crisp but not a ruined one.” which included the search query “burger”.

## London

Foursquare data for the city of London was accessed as one of the datasets. We used a bounding box of 11750 × 11000m (from [525242, 174992] to [536992, 185992] in the Ordnance Survey National Grid reference system EPSG 27700). An important consideration in selecting cities in England is the availability of data in the English language. London also offers the advantage of having a rich set of data. We also used several mobile map extents (with different sizes and generally where a concentration of the POIs is present) in order to test our method on different scales. Figure 4.11 shows the study area and the map extents. Table 4.1 provides data on the size of the mentioned extents. Map extents are set with the same ratio of 16:9, similar to conventional mobile phone displays.

Table 4.1 – London map extents and their dimensions.

| Reference | Extent Width (m) [East-West] | Extent Height (m) [North-South] |
|---|---|---|
| A | 197 | 350 |
| B | 560 | 995 |
| C | 1100 | 1955 |

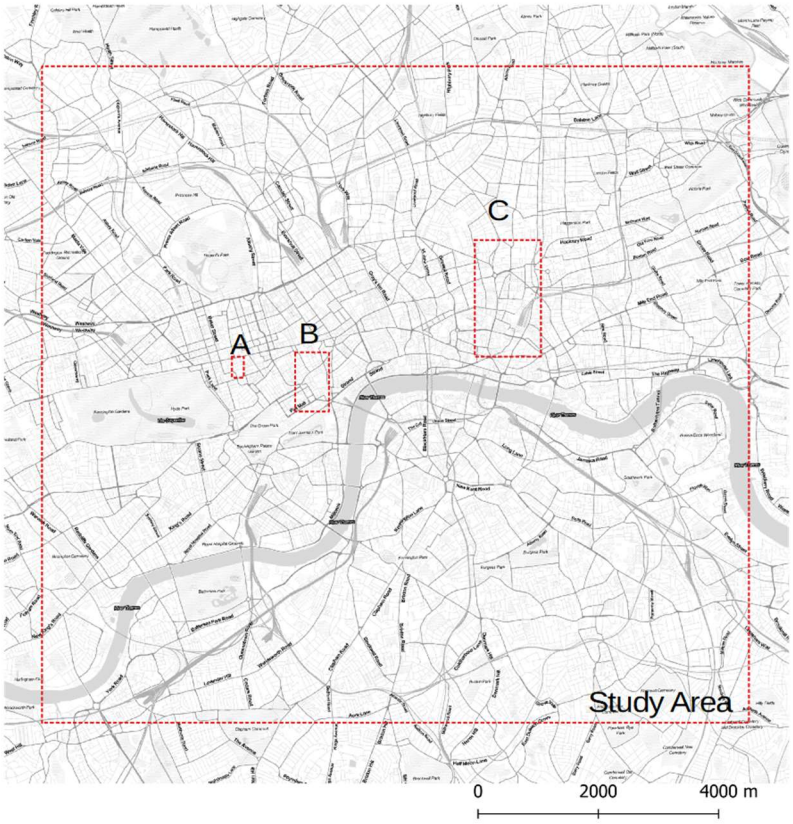


Figure 4.11 – Study area and map extents for the city of London

## Manchester

A second study area was considered based on the city of Manchester. Yelp data for this city has been accessed as the second dataset. We used a bounding box of 9544 × 8378m (from [388951, 401665] to [379407, 393287] in the Ordnance Survey National Grid reference system EPSG 27700). As mentioned before, our selection of cities in England was based on the availability of data in the English language. We also used several mobile map extents in order to test our method on different scales. Figure 4.12 shows the study area and the map extents. Table 4.2 provides data on the size of the mentioned extents. Map extents are set with the same ratio of 16:9, similar to conventional mobile phone displays.

Table 4.2 – Manchester map extents and their dimensions.

| Reference | Width (m) [East-West] | Height (m) [North-South] |
|---|---|---|
| X | 2292 | 4074 |
| Y | 1279 | 2271 |
| Z | 713 | 1268 |

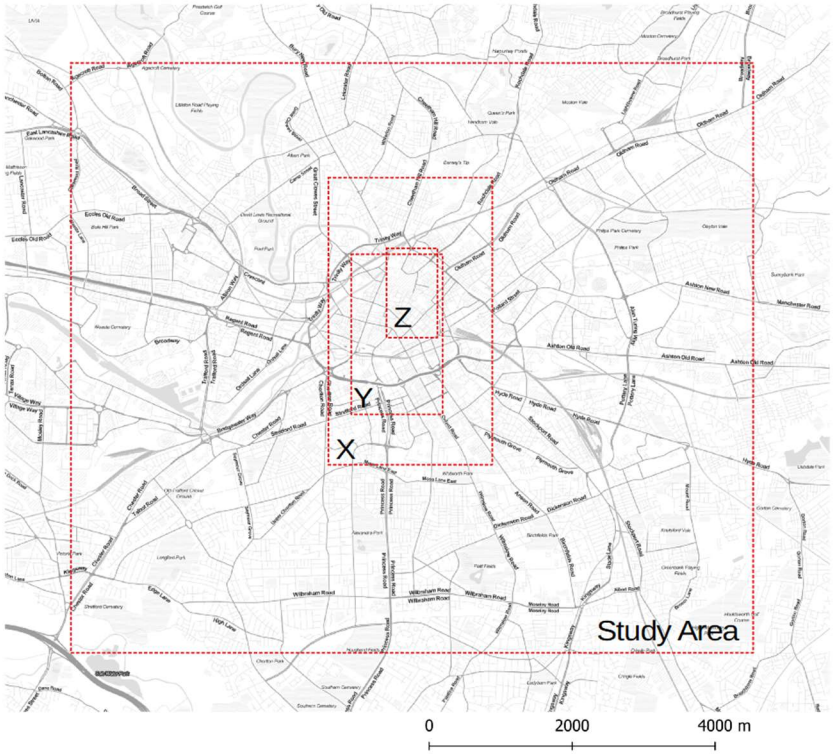


Figure 4.12 – Study area and map extents for the city of Manchester

### 4.3.2 Effects of method parameterization

In this section, the effect of different parameterization of the methodology is presented through examples. Firstly, the effect of bandwidth is investigated based on two selected map extents. This is followed by the investigation on the effect of contour threshold selection and finally by overlay method effects.

#### Bandwidth selection

The selection of appropriate bandwidth for the KDE function is important as it can affect the results of the estimated surface. As mentioned before, we have considered two different approaches. When considering the problem from a cartographic point of view, the bandwidth is derived from the map extent of the user (based on the formula introduced earlier in Section 4.2.6) and when considering the problem from a query point of view, the bandwidth is set by the user. The user sets the amount of time he/she is willing to spend to move between POIs. This would be translated into the bandwidth (e.g. 10 minutes of walking leads to 832m, therefore taking the half of this distance).

Figure 4.13 illustrates two sets of POIs based on London map extent C. We used two different words to fetch the POIs ("dance" on the left and "burger" on the right), and the ROIs are generated based on the POIs. Three different bandwidths are generated: 382m (cartographic approach), 416m and 582m (query approach based on 10min and 14min thresholds). ROIs generated based on "dance" (left) reflect the concentration of the POIs on the top-right corner of the map very well.

The difference between bandwidths is in covering two different POIs by the generate regions. All three ROIs cross the map's right border; therefore, in this case, applying threshold change or modifications such as displacement or morphing can potentially be considered. For the other case (for query word "burger"), there are more POIs in the map extent and an obvious cluster is well covered by all three different bandwidths. The second cluster of POIs are also detected (bottom right), and all three bandwidths result in ROIs in that region (with different sizes), but the ROIs are not very representative (covering only one POI and close to about six other POIs). Therefore, this ROI can be a candidate to be removed from the overview visualization. The difference between bandwidths is in covering four POIs on the top example. In both cases, the range of bandwidth seems not to generate very big ROIs (which cover the majority of the map extent) or very small (which are not visible).

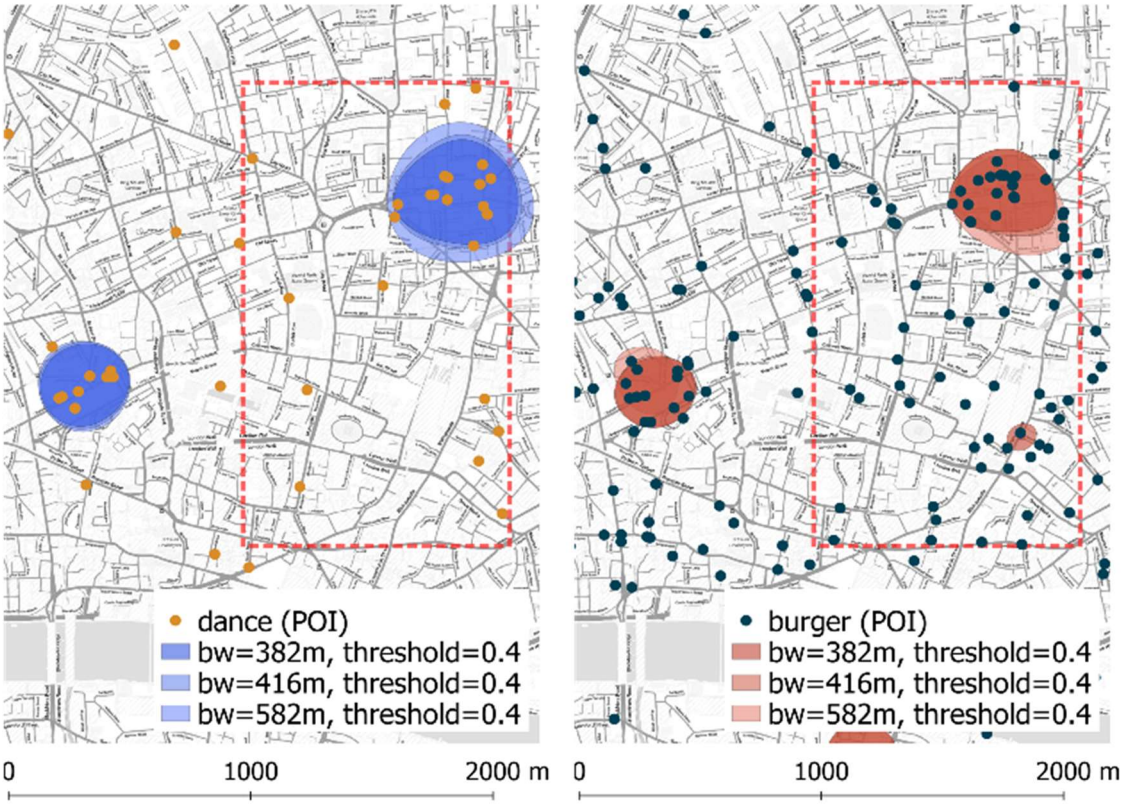


Figure 4.13 – Effect of bandwidth selection based on London Foursquare dataset and map extent C. On the left, ROIs are derived from the word "dance". On the right, they are generated by using POIs of query word "burger". Threshold cuts are set to 0.4.

In the other example and based on the Manchester dataset, we have used the same query words and three different bandwidths: 444m (cartographic approach), 416m and 582m (query approach based on 10min and 14min thresholds). Figure 4.14 illustrates two sets of POIs based on Manchester map extent Y. ROIs on the left ("dance"), seem to well represent the POIs' concentration on the middle of the map. The longer bandwidths (444 and 582m) result in one ROI, and the shorter bandwidth (416m) results in two ROIs (caused by a break in the zone with lower density). This could be understood as if the user is willing to move for a maximum of 10 minutes, he/she would get two regions to decide between but by willing to move 14

minutes, he/she would get one region to consider. ROIs on the right (for "burger"), represent a concentrated region on the top right corner. There are two concentrated POI regions in the center and the left side of the map which represent a smaller set of points (therefore candidates for elimination). The ROI in the bottom right corner of the map represents seven POIs and is going to be retained.

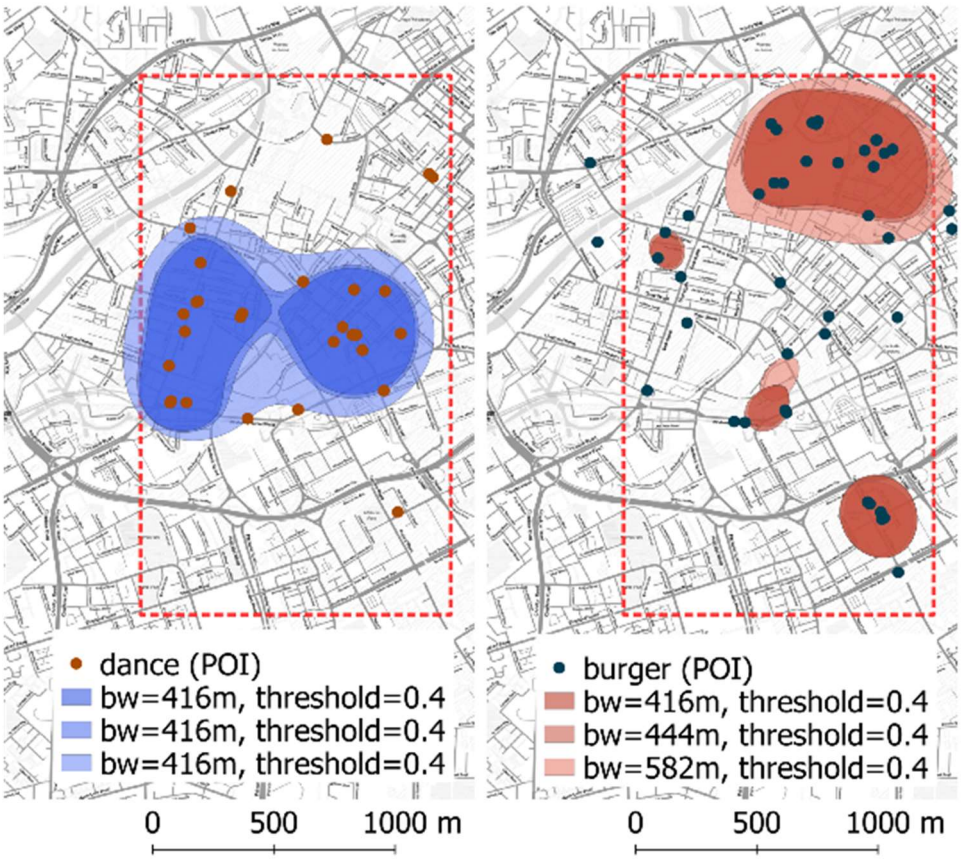


Figure 4.14 – Effect of bandwidth selection based on Manchester Yelp dataset and map extent Y. On the left, the ROIs are derived from the word "dance". On the right, they are generated by using POIs of query word "burger". Threshold cuts are set to 0.4.

### Contour threshold

Likewise, we approach the problem of defining the contour threshold from two different perspectives. One perspective solves the problem with regards to the cartographic view of the generated regions and the other one is based on values assigned by the user.

Figure 4.15 shows an example based on the query approach. The examples are based on Yelp data in the city of Manchester on map extent X. Based on the bandwidth equation, the bandwidth is set to 795m. The user has initially assigned 0.6 for the value of the threshold cut. The resulting ROI represents the cluster of POIs, but in case of interest to get bigger regions, two other thresholds are tested (0.5 and 0.4 consequently). The result shows the process of the ROI morphing towards other concentrations of POIs. Because of the fact that the user is in control of the process, he/she selects the preferred result (e.g. by rejecting 0.6 ROI because of size and 0.4 ROI because of tail shape, therefore selecting 0.5 ROI).

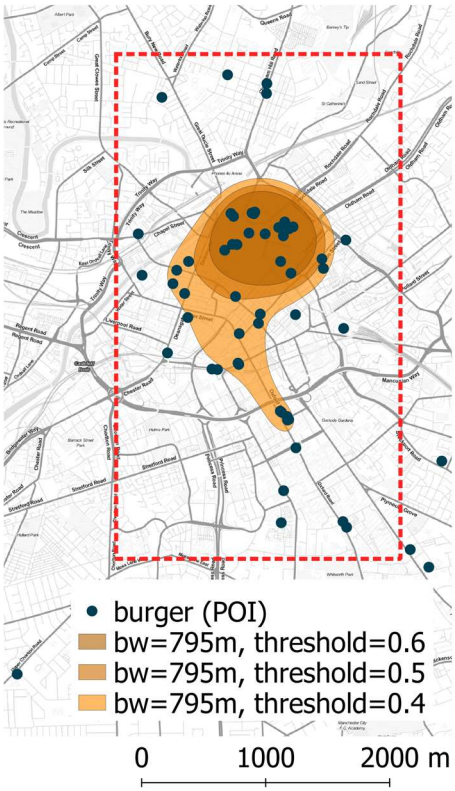


Figure 4.15 – Changing the threshold value for Manchester Yelp dataset on the map extent X. The bandwidth is 795m and the threshold cuts are 0.6, 0.5, and 0.4 from the smallest to the biggest polygon.

When considering the cartographic approach, the threshold parameter is set based on a cartographic criterion. An important factor to control is the size of the generated regions. (Other cartographic factors such as proximity between the ROIs could also be additionally added to the process.) In order to do that, we approach the problem by going through iterations. An example is provided in Table 4.3 and is illustrated in Figure 4.16. In the first iteration and based on a default threshold (e.g. 0.5), the ROIs are generated.

The results are then evaluated to see whether a case of violation exists. We test for relative area violations by dividing the area of the ROI to the area of the whole map view. In this example, the generated ROI takes 3.2572% of the map view. Our desired area range is set to be between 10% and 35% of the map view. Therefore, we should generate a new set of ROIs. To calculate the new threshold, we first calculate the change coefficient by dividing the desired relative area (in our example case 15%) to the current relative area. For the first iteration of the process, our change coefficient is 4.6052. By using the current threshold (0.5) and the change coefficient, we can calculate the new threshold based on the curve fitting method (introduced earlier in Section 4.2.6). In our example we calculate the new threshold to be 0.3479. By repeating these steps (calculating the relative area and change coefficient) for each iteration, our estimations get closer to the desired value (i.e. the change coefficient converges to 1). Here we have described an example of a growing ROI but the process is similar when we consider a shrinking ROI. Both cases are illustrated in Figure 4.16 and the data is provided in Table 4.3. Figure 4.16 provides visual representations and more detail is provided in the text.

Table 4.3 – Contour threshold changes in grow and shrink scenarios.

| Grow example | | | |
|---|---|---|---|
| Iteration | Contour value | ROI area % | Change coefficient |
| 1 | 0.5 | 3.2572 | 4.6052 |
| 2 | 0.3479 | 10.6266 | 1.4116 |
| 3 | 0.3136 | 13.7583 | 1.0903 |
| 4 | 0.3050 | 14.4865 | 1.0283 |
| 5 | 0.3022 | 14.8600 | 1.0094 |
| Shrink example | | | |
| Iteration | Contour value | ROI area % | Change coefficient |
| 1 | 0.4 | 40.2528 | 0.8695 |
| 2 | 0.4130 | 38.6603 | 0.9053 |
| 3 | 0.4223 | 37.2923 | 0.9385 |
| 4 | 0.4282 | 36.7132 | 0.9533 |

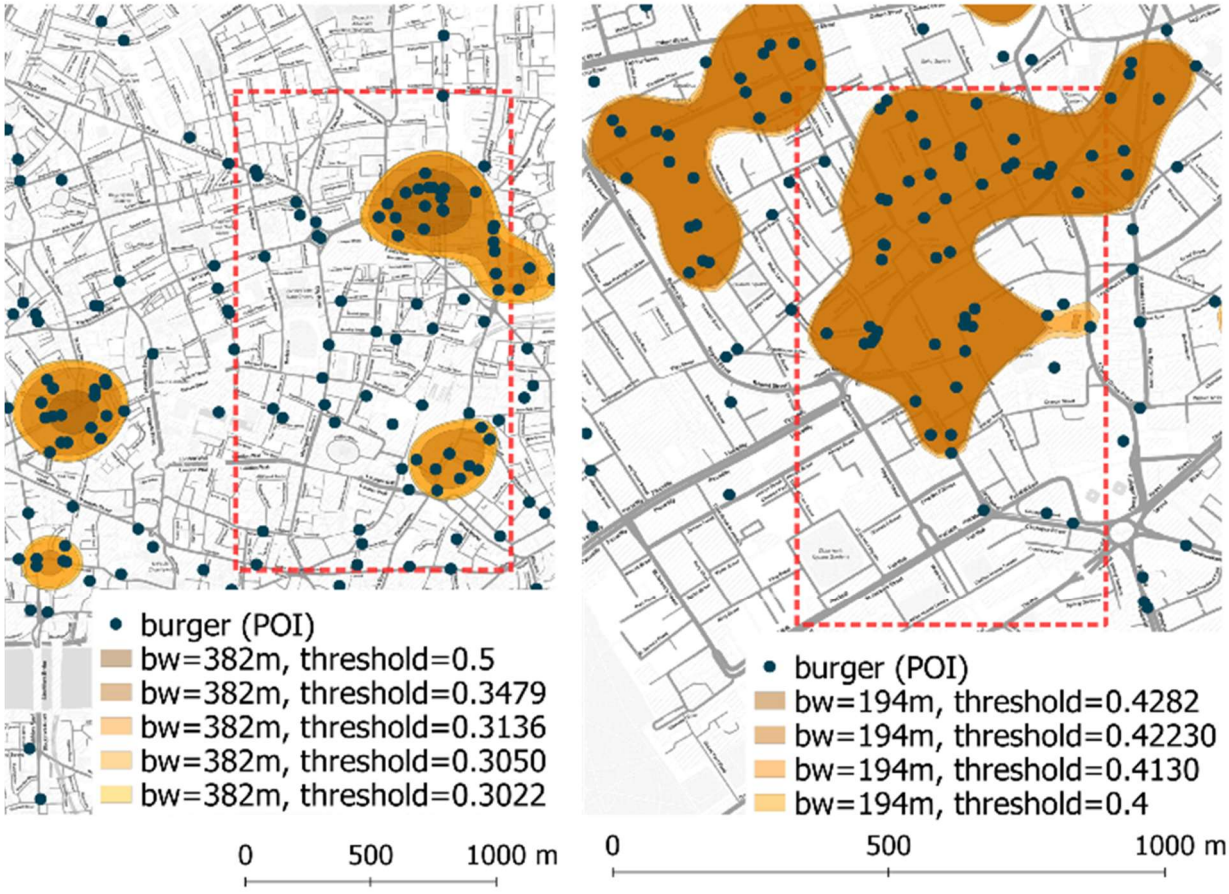


Figure 4.16 – Contour threshold changes in two different scenarios. In both cases, POIs are based on venues where data contains the word "burger". The left case is a scenario where the generated ROIs are small and should grow (to reach a minimum relative size), based on London extent C. The right

case is a case where the generated ROIs are big and should shrink (to reach a maximum relative size), based on London extent B.

### Themes overlay

Another important aspect of the parameterization of the methodology is overlaying the themes by combining the rasters of different query words together. This enables us to generate results that comply with a combination of criteria. Examples are generating ROIs that include two query words (by adding raster values) or ROIs that include one query word but exclude the other (by subtraction). Applying other combinations of raster operations is possible, but enough attention should be paid in order to derive meaningful results that can be related to meaningful queries. Figure 4.17 provides an example of generating results by adding the query rasters (in order to include both in the results), and Figure 4.18 shows the results when using a subtraction operator (to include one query word but exclude the other one). The results of addition overlay show the inclusion of both POI subsets, but morphing towards the shape of one theme (word) is obvious as different weights have been assigned. The combined ROIs cover more "burger" points than the "dance" ROI and vice versa. When considering the subtraction case, the resulting ROIs show the polygons moving away from the negative raster (related to the query word "loud"). Therefore, the results provide regions that cover higher densities of the positive term while moving away (and therefore not covering) the POIs related to the negative query term.

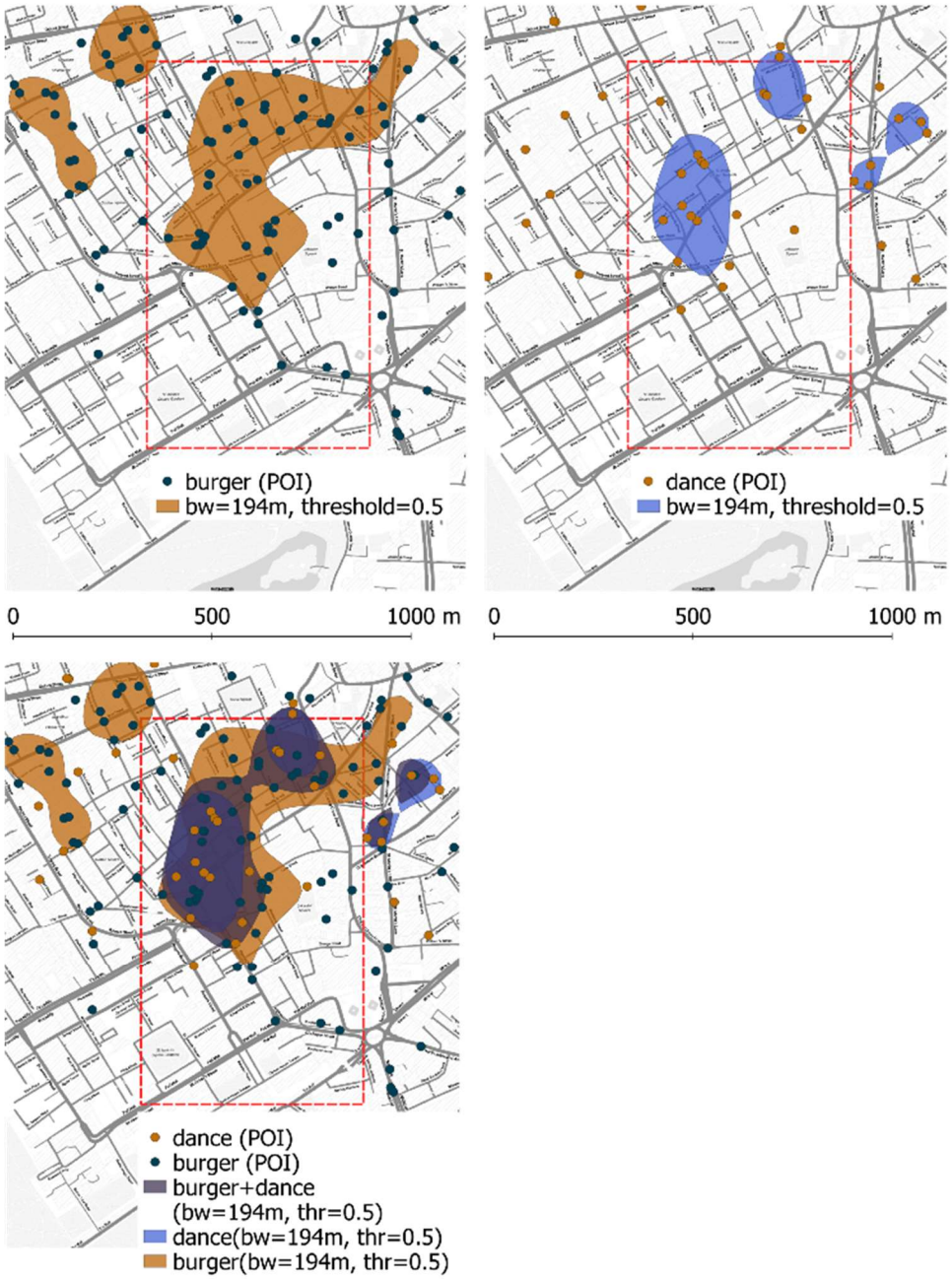


Figure 4.17 – Overlaying two queries using addition based on London extent B: POIs and ROIs of query word “burger” (top left), results for the word “dance” (top right), and results of the overlay (bottom). In order to express importance, different weights are assigned (0.7 to “dance” and 0.3 to “burger”).

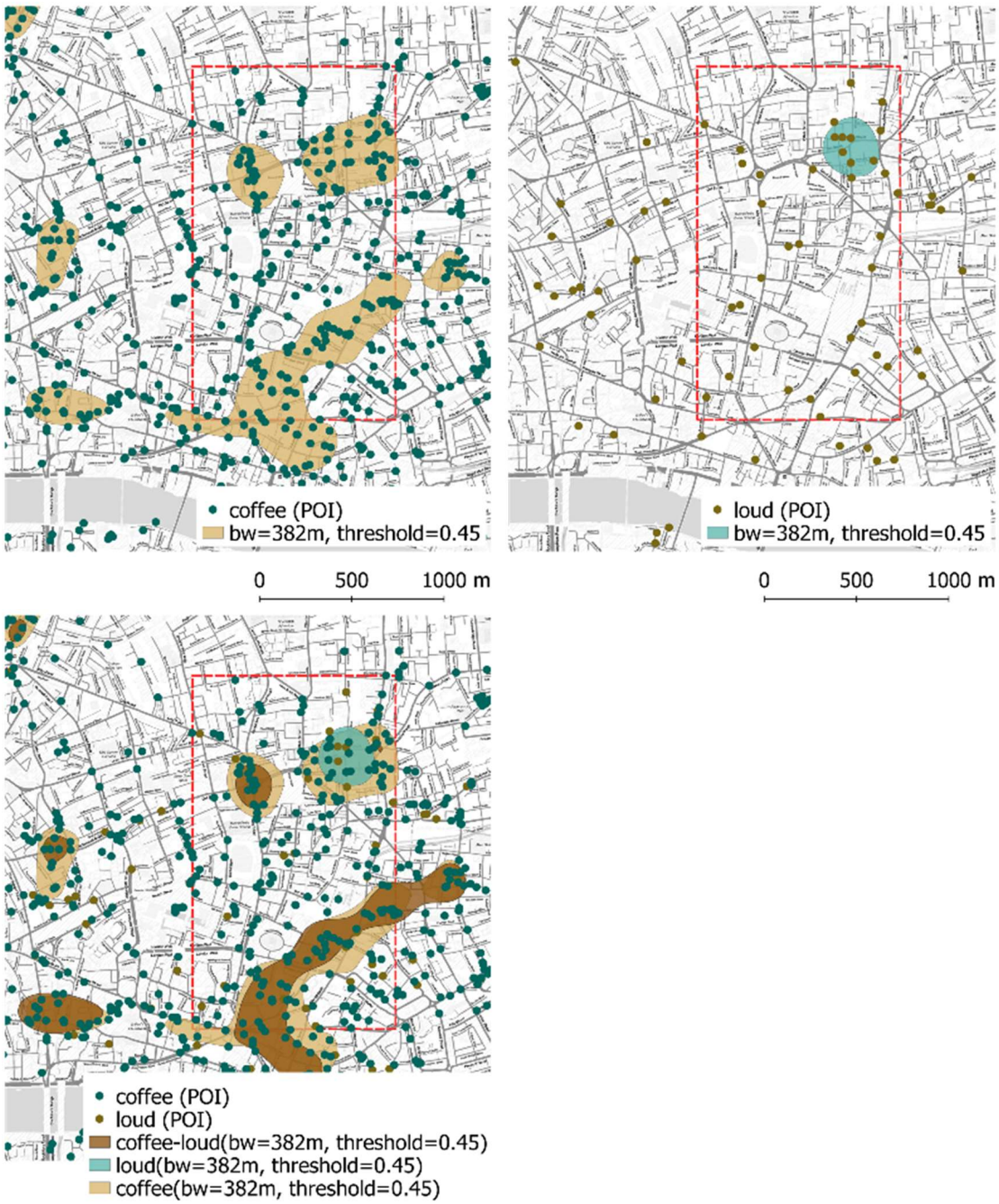


Figure 4.18 – Overlaying two queries using subtraction based on London extent C: POIs and ROIs of query word "coffee" (top left), results for the word "loud" (top right), and results of the overlay (bottom).

### 4.3.3 Sample use case

Our applied use case is based on a mobile user (named Paul) who is visiting the city of London for the first time and is planning his afternoon in the city. As London is a big city with a large number of sightseeing and venues to visit, Paul consults his mobile service to firstly get an overview of regions which provide activities he is interested in and secondly to obtain a closer view over those regions. Paul considers two words ("architecture" and "coffee") related to

two activities that he is interested in doing: visiting a place with architectural value and drinking a coffee afterward. Figure 4.19 shows the primary results. The points on the maps are POIs that have been retrieved from the source dataset (Foursquare venues for London).

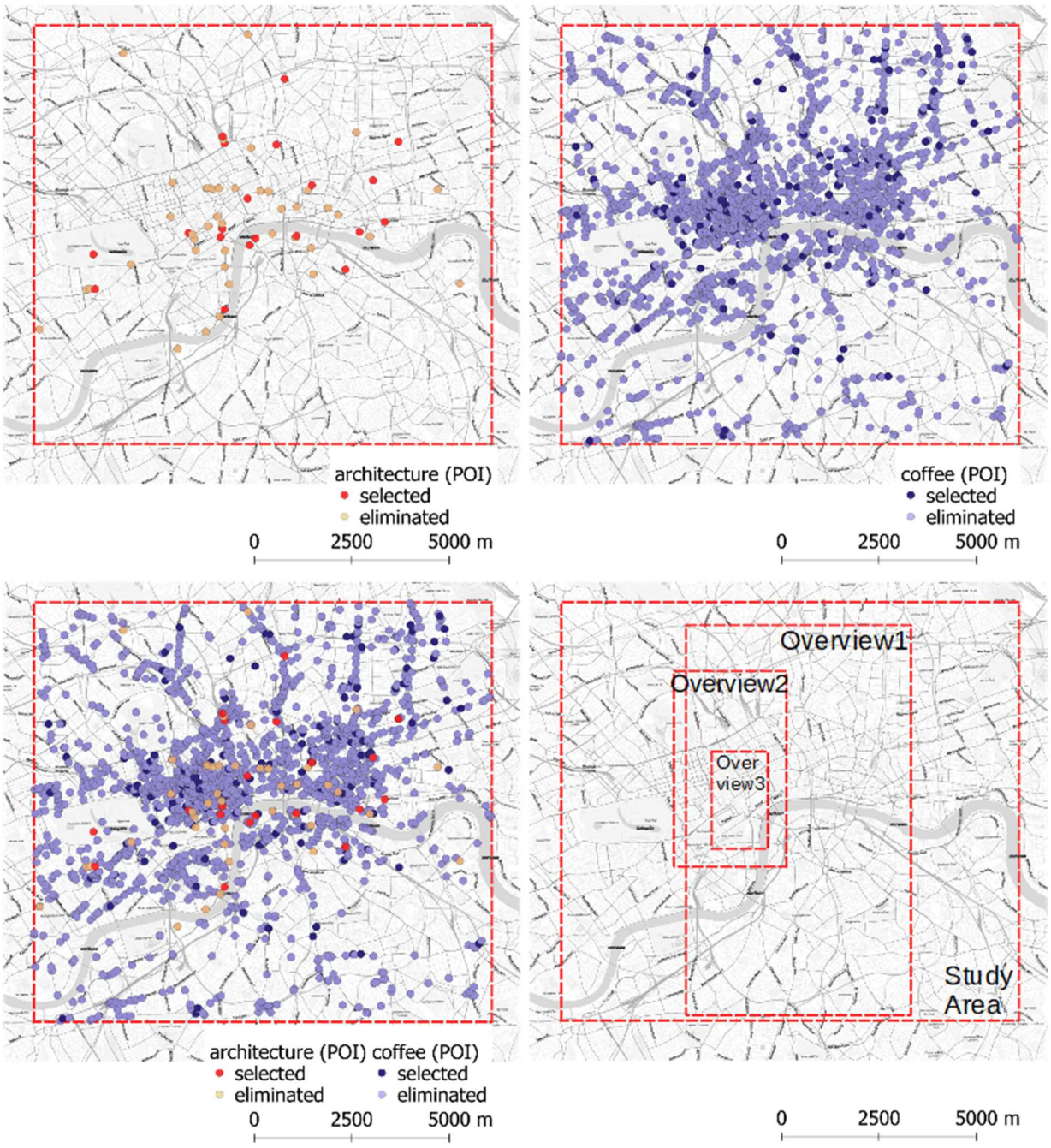


Figure 4.19 – Initial search results for Paul's search keywords "architecture" and "coffee": results for "architecture" (top left), results for "coffee" (top right), and the combination of POIs from both (bottom left). Darker points show a subset of the results by applying a selection criterion (venue ratings more than a threshold, e.g. more than 9 out of 10) that a conventional mobile app applies to the data. The map in bottom right shows the map extents that Paul is using to zoom on the dense areas.

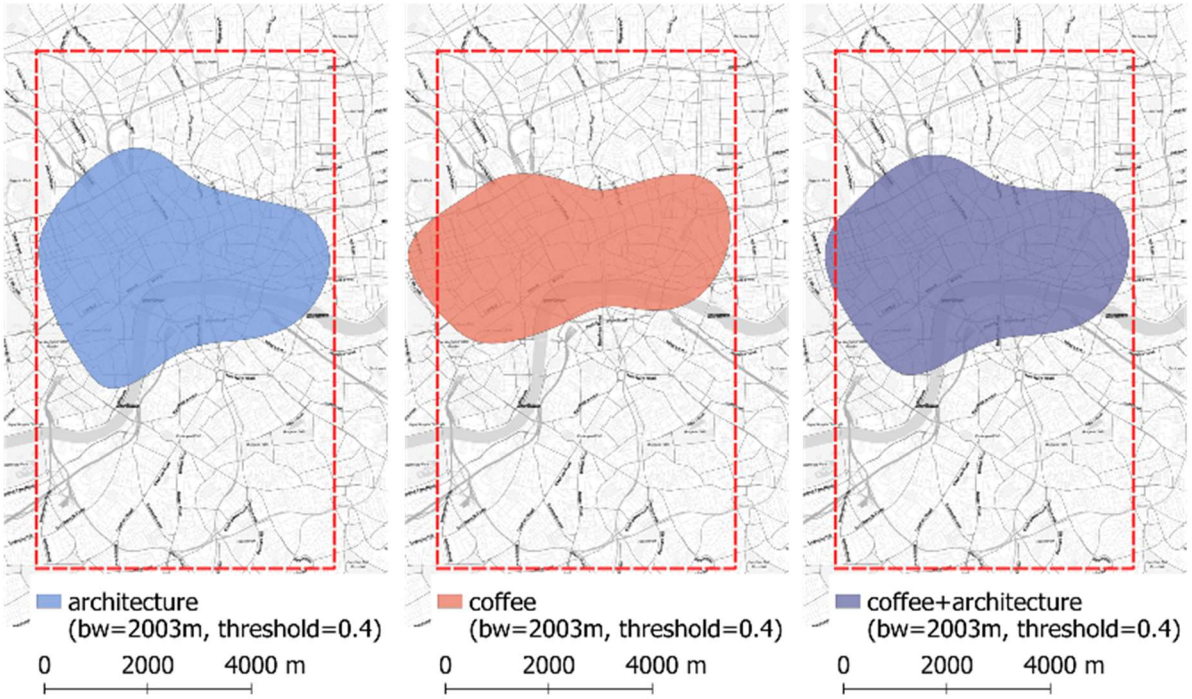


Figure 4.20 – The first set of ROI maps shown to Paul: "architecture" (left), "coffee" (middle), and the combination of both (right). ROIs are generated based on a 2003m bandwidth and a threshold cut of 0.4.

With a high density of POIs on the maps, there is a need to apply a generalization/abstraction process (such as selection/filtering). This is what a conventional map provider service would do by considering a (combined) criterion, on which a subset of the results is selected and visualized. In the maps in Figure 4.19, the darker points are selected by applying a user rating score criterion. It is obvious that some regions of the map provide more results that attract Paul to consider those regions to visit. Paul's interaction with the map is based on observing regions of higher densities (through ROIs) and zooming on them. The bottom right map on Figure 4.19 shows Paul's overview maps based on his map interactions (subsequent zooming in).

By entering the query words ("architecture" and "coffee"), and by setting weights for the combination map (0.7 for "architecture" and 0.3 for "coffee", as Paul finds architecture more important), the ROI generation process gives the results shown in Figure 4.20 to Paul. The map extent is an overview (5769 × 10256m) and the bandwidth is set to 2003m. The generated ROIs provide regions of dense POIs. Because of the weighting, the shape of the combined map is more similar to that of architecture. As this is still an overview covering a large area, Paul zooms in on the left side of the map and gets a new map.

The second overview extent is shown in Figure 4.21. The map extent complies with a 2885 × 5128m rectangle and yields to a 1002m bandwidth. The primary ROIs for this zoom level (which covers 25% of the last map extent) are bigger than the map extent and could go through the relative area constraint process introduced earlier. Paul gets an understanding of the concentration and zooms in one more level.

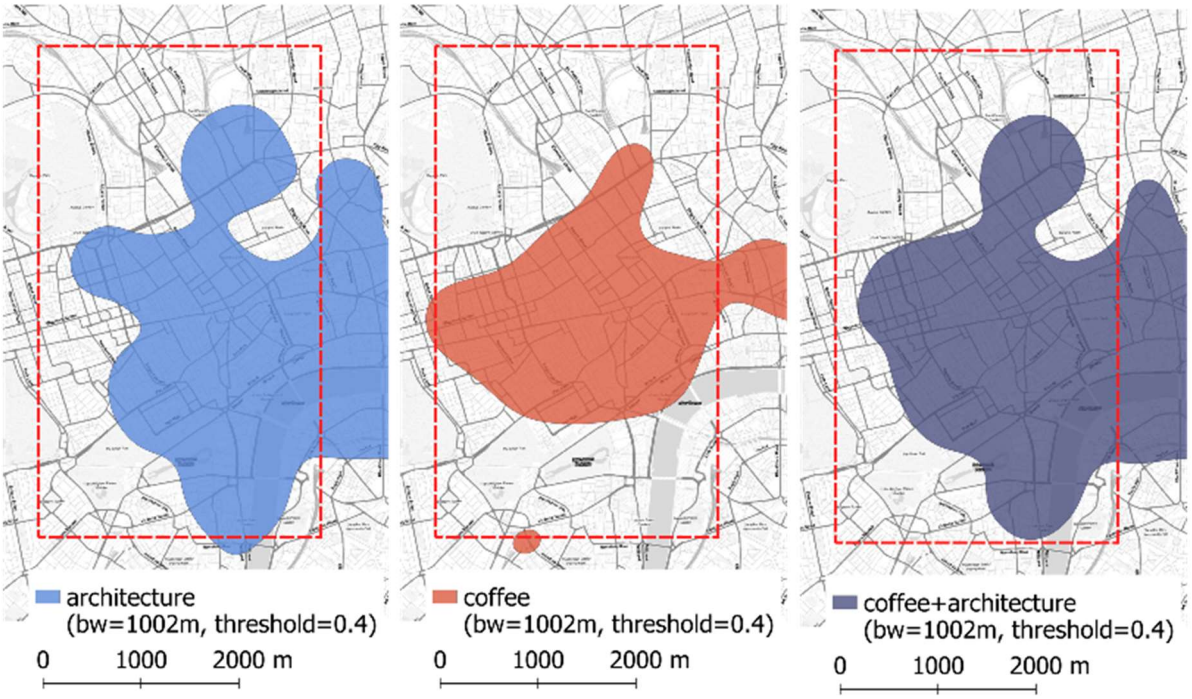


Figure 4.21 – Second set of ROI maps shown to Paul: "architecture" (left), "coffee" (middle), and the combination of both (right). ROIs are generated based on a 1002m bandwidth and a threshold cut of 0.4.

Figure 4.22 shows the results of the third overview map. The results are based on a 1442 × 2564m map that yields to a 501m bandwidth. As architecture POIs show a lower concentration, the ROI breaks into smaller ROIs but the high concentration of coffee POIs results in a region that still crosses the map view. The combination map, however, shows a size that is well considerable for this map view. The bottom right map in Figure 4.22 shows the POIs. The generated region covers 12 architecture POIs and 302 coffee POIs. With the previous steps, Paul could focus on the regions of high density (and high interest for him). The ROI representation process can continue further (as far as the POIs are clustered). This visual representation helps Paul to avoid interacting with a confusing number of objects, but rather to focus on regions that potentially include venues that are relevant to his planned afternoon – and in a more readable manner.

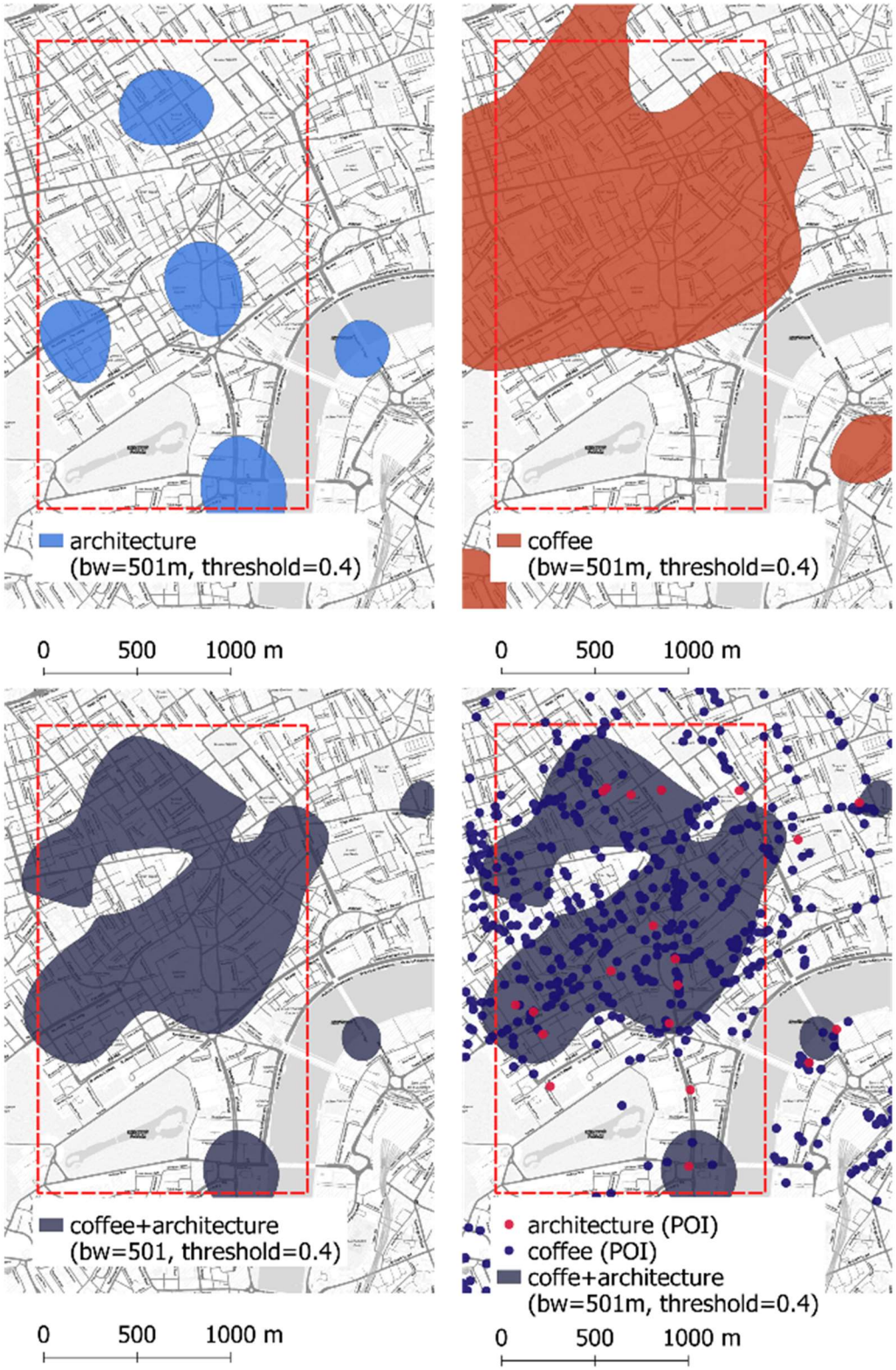


Figure 4.22 – Third set of ROI maps shown to Paul. Top row maps are the maps based on "architecture" and "coffee" POIs. The bottom row shows the result of the combination: bottom left the combined ROI map and bottom right the ROI map, along with the POIs of query POIs.

## 4.4 Cartographic evaluation

In order to evaluate the methodology introduced in this chapter, we could not base our comparison on comparing with earlier methods of generating ROIs (such as Cao et al., 2014, Lamprianidis et al., 2014 and Yu et al., 2017), as they had different motivation and for instance did not include any map view and scale related parameters. Therefore, we focused on comparing the situation before and after switching between POIs and the ROI-based visualization method based on cartographic evaluation measures. Hence, the aim was to explore the effect of the proposed method on key cartographic criteria that can be quantitatively assessed.

As mentioned in Section 2.4.3, in order to evaluate the generalization process in a quantitative manner, we use measures that help us evaluate map readability. These measures reflect either the amount of information or spatial distribution of the map objects (e.g. proximity). Based on the earlier work in cartographic generalization evaluation (Harrie, Stigmar & Djordjevic, 2015; Stigmar & Harrie, 2011), we selected a group of measures that fit our problem and the suggested solution:

- Number of objects (NO): number of objects on the map.
- Number of vertices (NV): number of vertices on the map. In the case of point objects, this number is equal to the number of objects. In the case of lines and polygons, shape points are extracted and counted.
- Object line length (OLL): the sum of the length of all objects on the map. In the case of point objects, the perimeter of the bounding rectangle is used.
- Nearest neighbor index (NNI): the ratio of mean nearest neighbor distance to the expected nearest neighbor distance.
- Proximity value (PV): a value that shows whether disjointed objects are too close to each other. This is calculated by dividing the sum of the area of intersections between neighboring object buffers (buffer size based on keeping a minimum distance of 0.3 mm between objects based on Harrie, Stigmar & Djordjevic, 2015) to the area of the whole study area.
- Proximity indicator (PI): number of object pairs that are too close to each other. This is calculated by comparing object distance to a threshold value (0.2 mm based on Harrie, Stigmar & Djordjevic, 2015).

We selected NO and NV, as they represent the number of objects. A map containing a large number of objects has a higher potential to have visual clutter and also high complexity (i.e. higher readability issues). OLL provides insight into the length of objects on the map (i.e. the space occupied by the borders of objects). NNI, PV and PI are rather measures of objects close to each other (and therefore hard to distinguish between). The latter three measures were selected to help us understand how condensely the objects are visualized. We included these measures because it is possible to have a small number of objects but all of them condensely visualized very close to each other. NO is based on the results reported by Stigmar & Harrie (2011), NV, OLL and PI are based on results reported by (Harrie, Stigmar & Djordjevic, 2015; Stigmar & Harrie, 2011) and finally, PV is based on Harrie, Stigmar & Djordjevic, (2015). In order to investigate the retaining spatial distribution of the data, we used NNI (in the measures above) and analysis of quadrat counts.

Based on one of the cases presented before, Figure 4.23 and Table 4.4 provide the measures and diagrams of ROI and POI situations. These measures are based on two words ("burger" and "beautiful") and two different bounding boxes (the whole study area and London extent B). For the case of both example words, when looking at the whole study area, all measures

except PV report decrease. For the case of both example words, when looking at the selected map extent (Extent B), the behavior of NNI, NO, PI and PV do not change but NV and OLL report an increase (instead of a decrease). Looking at the case of NV, by applying a shape simplification algorithm, e.g. that from Douglas & Peucker (1973), the increased number of vertices could be avoided (using a tolerance of 5 meters). To summarize, in general, the results of these measures report a general decrease (with the overall exception of PV and situational exception of OLL) and thus an improvement in map readability measures.

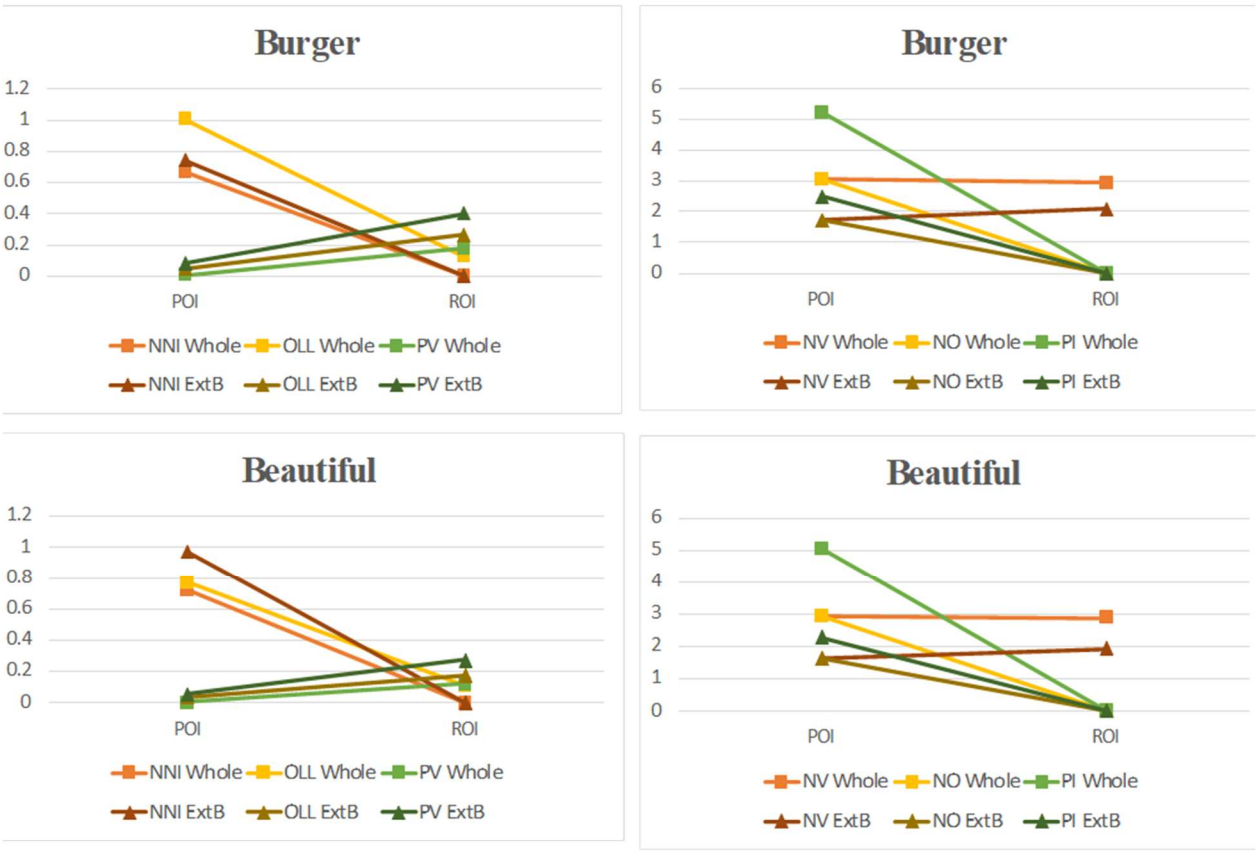


Figure 4.23 – Results of cartographic evaluation based on the word "burger" and "beautiful". The left and right results are separated due to different scales of the measures. More detail on the numbers and abbreviations is provided in Table 4.4.

We are aware that improving the measures merely by decreasing the amount of information at the cost of losing the spatial distribution of data is not desirable. Therefore, we applied a quadrat count analysis on the data both before and after the generation of ROIs. The analysis is based on a square grid cell of 200m, and we counted the number of POIs and ROIs in each cell. When looking at the distribution of the number of objects per cell, for each case we divided the cells into two classes: one having more objects than the threshold (1) and one having less than the threshold (0). The threshold value was calculated based on the distribution of the number of objects per cell and by cutting 90% of the cells. For an example case of the word "burger", the POIs led to a threshold value of 2 and the ROIs led to a threshold value of 1.

Table 4.4 – Cartographic evaluation measures based on the words "burger" and "beautiful". For each word, two different bounding boxes are investigated (the whole study area and also London extent B).

| **Query word** | **Burger** | | | |
|---|---|---|---|---|
| | **Whole study area** | | **Extent B** | |
| **Measure** | **POI** | **ROI** | **POI** | **ROI** |
| Nearest Neighbor Index (NNI) | 0.66319 | 0 | 0.73923 | 0 |
| Number of Vertices (NV) | 1115 | 875 | 52 | 123 |
| Number of Objects (NO) | 1115 | 1 | 52 | 1 |
| Object Line Length (OLL) | 1.0035 | 0.12908 | 0.0468 | 0.26263 |
| Proximity Indicator (PI) | 168485 | 0 | 315 | 0 |
| Proximity Value (PV) | 0.00663 | 0.17589 | 0.08324 | 0.39716 |
| | | | | |
| **Query word** | **Beautiful** | | | |
| | **Whole study area** | | **Extent B** | |
| **Measure** | **POI** | **ROI** | **POI** | **ROI** |
| Nearest Neighbor Index (NNI) | 0.72097 | 0.0 | 0.96832 | 0 |
| Number of Vertices (NV) | 856 | 765 | 43 | 85 |
| Number of Objects (NO) | 856 | 1 | 43 | 1 |
| Object Line Length (OLL) | 0.7704 | 0.11094 | 0.0387 | 0.17461 |
| Proximity Indicator (PI) | 114643 | 0 | 190 | 0 |
| Proximity Value (PV) | 0.00494 | 0.12215 | 0.05704 | 0.27360 |

In order to take ROI generation method as a good representation, the method should represent the spatial distribution of the POIs. This translates to having cells overlap with POIs (more than the threshold) and with ROIs (more than the threshold). The abovementioned case leads us to Table 4.5 and Figure 4.24. True positive cells are the cells that included POIs and ROIs. True negative cells are the cells that had neither POI nor ROI. False positive cells are the cells that were reported to have an ROI but did not have (enough) POIs. False negative cells are the cells that included POIs but were not reported for having ROIs. Based on the data, we have an accuracy of 94.6%. In other tests (using other words and grid cell sizes), the accuracy has been observed to be between 92% and 97%.

Table 4.5 – Evaluating the spatial distribution of POIs and ROIs, counting the overlap between POI and ROI cells.

| | | **POI** | |
|---|---|---|---|
| | | **Yes (1)** | **No (0)** |
| **ROI** | **Yes (1)** | 96 (3.01 %) | 38 (1.19 %) |
| | **No (0)** | 135 (4.23 %) | 2921 (91.57 %) |

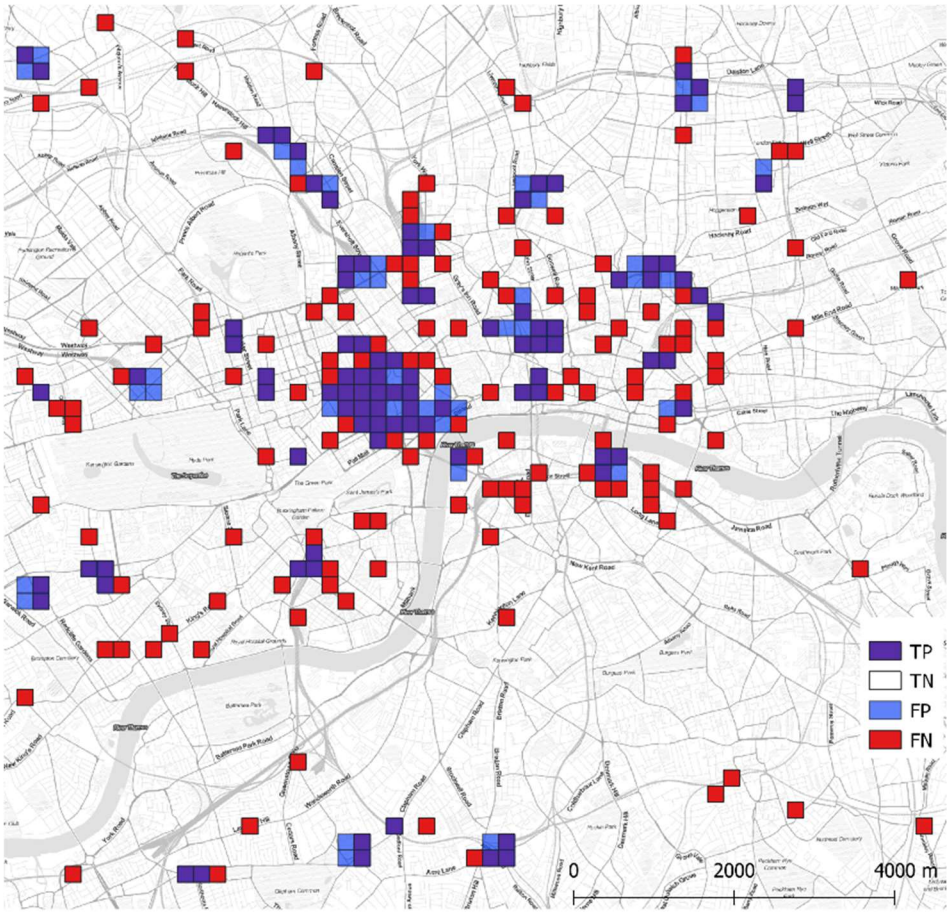


Figure 4.24 – Overlapping the cell with ROIs onto cells with POIs

## 4.5 Discussion

In this chapter, the method of generating ROIs to replace POIs has been proposed. The POIs have been results of queries sent to a database to fetch points related to a certain keyword. The motivation is based on the need to provide abstractions of geographic information and avoid clutter at the same time. Generation of ROIs have been investigated in early research (mostly in information retrieval), but earlier research works (e.g. Cao et al., 2014; Lamprianidis et al., 2014; and Yu et al., 2017) have not provided the support for different scales. We have approached this problem by relating the generation of the ROIs to the user's map view and have proposed both the methodology and the quantitative measure for triggering the method (i.e. switching to ROI visualization).

We tested our results by comparing the ROI visualization to POI visualization of the same data. The comparison was based on measures of map readability (specifically, a measure of the amount of information and spatial distribution), as reported in Section 4.4. For example, in the case presented in Table 4.4, we observed a great decrease in NV, NO and PI and a slight increase in PV when looking at the data of the whole study area. When looking at the data of a certain map extent, we observed a great decrease in NO and PI but an increase in NV, PV and OLL. Looking at the case further, we observed that the increase in NV is due to the shape complexity of the generated ROIs which can be solved by applying a simplification algorithm

(e.g. Douglas Peucker). This step, which was not part of the initial methodology, can be seen as a post-process improvement step.

With a shape simplification step, the decrease in NV holds for smaller extents. This means that by applying the method introduced in this chapter, the map readability measures of NO, NV and PI report great decrease where the other measure of OLL and PV report a slight increase. Considering the cartographic evaluation measures that are based on the amount of information, we can conclude that the switch from POI visualization to ROI visualization results in significantly more readable maps (with the consideration of line simplification as a post-process) and is therefore an appropriate abstraction method.

Another point to consider is whether the whole spatial distribution of the data is retained or not. We tested this by applying the quadrat count analysis (exemplary results in Table 4.5 and Figure 4.24) and reported accuracy between 92% and 97% in quadrat representability of ROIs. An important point is to look at false positives (reported as ROI but not POI) and false negatives (reported as not ROI but include POIs more than the threshold). The case of false negatives is more frequent, mainly due to the concentration of points in a single quadrat cell not covered by an ROI. This is more tolerable compared to the case of false positives, where the density of POIs is lower than the threshold but the cell is included in an ROI. Looking at the data, we observed the latter to be generally quadrat cells neighboring POI quadrat cells or gap cells in regions of high concentration of POI quadrat cells (Figure 4.24). This phenomenon is seen as a shortcoming of the representation of POIs with ROIs.

In summary, when applied appropriately, the method introduced here results in more readable maps by reducing the amount of information while keeping the spatial distribution of the data. The benefits of the method are achievable by triggering the process correctly and by using appropriate parameters. The parameter (L-function) is calculated based on the data and enables the automation of parameter selection. The important initial decision lies in applying the method to situations of high densities. Despite the fact that the method generates results based on the low density of points, the generated regions are not better visual representations of them. Besides this factor, selecting appropriate bandwidths and contour thresholds are of high importance.

The main difference of our method compared to other similar research works (e.g. Cao et al., 2014 and Lamprianidis et al., 2014) lies in establishing the relationship between the map view of the user and the generation of the ROIs. This relationship functions towards parameterizing the KDE surface based on the user's interactions rather than extracting parameters from the data, with the advantage that the results are adapted to the user's demands. The disadvantage, however, lies in the fact that user interactions with the data (e.g. panning and zooming) should generate new KDE surfaces (and contours). This makes the process computationally expensive. A feasible solution is to pre-compute the KDE-surfaces (based on pre-defined zoom levels). Another consideration regarding optimizing this method is to minimize the iterations in the generation of contour lines. We have proposed solving this problem by estimating the desired value, and there is potential to consider this problem differently (e.g. by finding a local threshold value rather than a global value).

Finally, it should be mentioned here that the research reported in this chapter describes proposed a methodological approach that has been quantitatively evaluated. Obviously, there exists the possibility of testing this methodology on subjects and investigating the advantages and disadvantages. This step can be seen as a follow-up investigation based on the findings of this chapter.

## 4.6 Conclusion

In this chapter, we have presented a method to generate regions (ROIs) that represent point objects (based on but not limited to a search query). Besides introducing the overall method, we presented solutions to parameterizing ROIs in order to fit in a selected map extent (e.g. a mobile phone screen). We also provided a method of quantitatively triggering the visualization change from points to regions (and vice versa). Consequently, we presented the cartographic evaluation of our method based on map readability measures.

The work presented here contributes mainly to the work of location-based services (LBS) and cartographic generalization and is a step to provide a solution for the situations of having visual clutter because of high point density (e.g. on mobile apps and web maps).

In Chapter 3, we proposed deriving a visualization parameter (i.e. semantic similarity) from a list of provided objects (e.g. user's favorite objects). Another important factor in generating the results in that chapter was the current location of the user, and we mainly focused on maps of higher scales (comparing to this chapter). In this chapter, we minimized the user's input to a search keyword and focused more on generating overview maps of the study area (thus lower-scale maps). In Chapter 5, we base our methodology on topics derived from the dataset (without any user input or query) and without including the user's location.

# 5 Abstraction of textual data by matching between geographic objects and topics

## 5.1 Introduction

In the previous chapter, we proposed the methodology of generating ROIs to represent POIs as an abstract view over a study area. The input to the methodology was POIs that have been filtered by a search term (query). In this chapter however, we propose the methodology of generating abstract visualizations without using any search term but rather by using important textual topics extracted from the data.

Making use of the quantity of the data generated by millions of internet users demands a means of abstracting it. Besides the necessity of providing abstract views over the data and by considering the diversity of the data generated by internet users (e.g. from tourism attraction reviews to records of sports activities), matching and fusing data from different sources become crucial. A conventional internet user needs a way to grasp an overview of the data that includes important aspects of the data. Abstractions of geographic information are a common way of starting visualization sequences (starting by an abstract visualization and followed by more detailed views), which is in line with Visual Information Seeking Mantra (Shneiderman 1996). Basing our methodology on UGC contributions and in response to the third research objective introduced in Chapter 1, we aim to provide visual geographic abstractions of salient keywords (topics) to provide overviews over the study area. By addressing this research objective, the contributions of the research reported in this chapter are the following:

- Matching methodology between topics and geographic objects based on a textual method
- Pattern recognition method for matching results
- Visualization of matching results based on the detected patterns

## 5.2 Methodology

Figure 5.1 provides a visual representation of the methodology of this chapter. The first step in the process is the extraction of textual topics from the UGC source. This is done by extracting tokens (words) from the textual content by calculating the term frequency–inverse document frequency (TF-IDF) score for each token. We have filtered the tokens based on their TF-IDF score to extract the most important topics. Extracted topics are then passed to a matching process. In the matching process, the topics are checked with a database of geographic objects taken from OSM. In the case of finding matches between words and the list of geographic objects, the list of objects is then passed to a pattern recognition process. The pattern recognition process classifies the input (based on quantitative measures, i.e. the number of clusters and the K-function derivative) into three different classes. Based on the assigned class, the visualization method is determined.

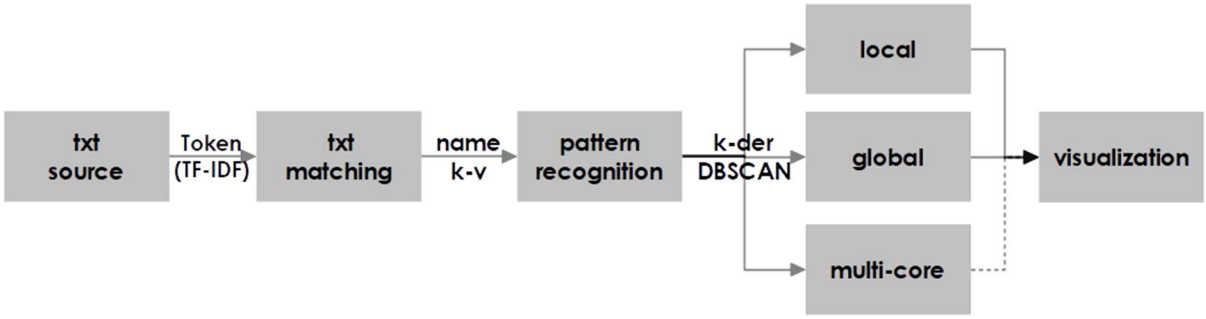


Figure 5.1 – Overview of the methodology. k-v stands for key-value and k-der stands for K-function derivative over distance.

### 5.2.1 Data

We have used data from three different data sources: Flickr, OpenStreetMap and Foursquare. OSM data is taken as the geographic database against which salient topics are tested. Flickr textual data is extracted by using tags attached to photos. In this step, a list of tags has been excluded from the data. These tags are mainly (but not limited to) tags about photographic properties of the photo (e.g. camera or lens information) or tags including geographic information that was not relevant in the granularity of our research (e.g. the tag "europe"). Foursquare data is extracted from the reviews that users have written about the venues in the dataset. This dataset includes data similar to the form of natural language, therefore more preprocessing steps were necessary. These steps included removing stop words and excluding tokens based on their part of speech (e.g. exclusion of adjectives and adverbs). All of the data has been accessed through the Web via (RESTful) APIs. For our tests, we focused on data from the city of London in England as our study area. We used a bounding box of 11750 × 12500m (from [ 525200, 174950 ] to [ 537700, 186700 ] in the Ordnance Survey National Grid reference system EPSG 27700).

### 5.2.2 Extraction of topics

In order to extract salient topics from datasets, the study area was divided into grids cells of 625 × 470 meters. Although we have not investigated an optimum cell size, a cell size of ~500 meters has been reported to be optimal for the extraction of topics on this study area (Bahrehdar & Purves 2018). All of the textual data extracted from the contributions (photo tags in the case of Flickr and venue reviews in the case of Foursquare) in each grid cell are taken as one text document.

After extracting the tokens, the documents are then analyzed with their token TF-IDF scores in order to extract the most important tokens of each document. Calculating TF-IDF scores is a common method in information retrieval in order to detect important words in documents. As mentioned in the previous section, the tokens went through a preprocessing step in which non-relevant textual information has been removed from the dataset. Initially, by limiting the TF-IDF score to a minimum of 0.001, we found 8402 and 28418 unique tokens in Flickr and Foursquare, respectively. By observing the TF-IDF scores, we extracted thresholds for passing the tokens to the next step (textual matching). The threshold was calculated based on the tokens who represent 50% of the top cumulative TF-IDF scores. This resulted in taking the top 222 and 1007 tokens from Flickr and Foursquare, respectively.

Figure 5.2 provides the diagram of the top 500 words TF-IDF score from Flickr and Foursquare. Table 5.1 provides the statistics of the TF-IDF scores. Table 5.2 provides the top 20 words for each dataset along with their scores.

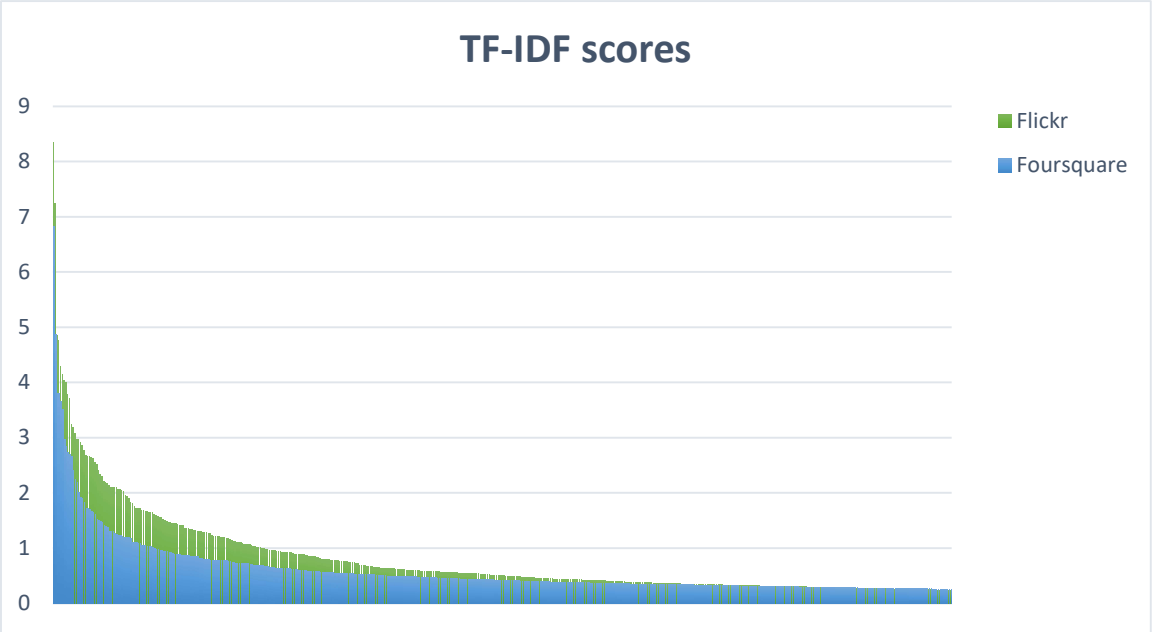


Figure 5.2 – TF-IDF scores for the top 500 words from both datasets

Table 5.1 – Statistical summary of the TF-IDF scores for Flickr and Foursquare

| | **Flickr** | **Foursquare** |
|---|---|---|
| **n** | 8402 | 28418 |
| **min.** | 0.001002 | 0.001 |
| **max.** | 8.342227 | 6.834461 |
| **mean** | 0.072712 | 0.028288 |
| **standard deviation** | 0.283739 | 0.115973 |

Table 5.2 – Top 20 words from Flickr and Foursquare along with their scores.

| **Flickr** | | **Foursquare** | |
|---|---|---|---|
| word | score | word | score |
| square | 8.342227 | food | 6.834461 |
| hackney | 7.237971 | place | 4.872726 |
| islington | 4.853711 | staff | 3.810804 |
| deptford | 4.764054 | service | 3.804566 |
| battersea | 4.290357 | coffee | 3.65025 |
| camberwell | 4.147036 | wa | 3.507205 |
| peckham | 4.040944 | get | 2.969784 |

| | | | |
|---|---|---|---|
| southwark | 4.011444 | amazing | 2.850857 |
| city | 3.790356 | pub | 2.740491 |
| architecture | 3.717355 | beer | 2.697774 |
| park | 3.243569 | try | 2.660817 |
| train | 3.195589 | one | 2.41699 |
| thames | 3.081353 | go | 2.240814 |
| building | 2.980117 | park | 2.189823 |
| people | 2.960763 | lunch | 2.005699 |
| clapham | 2.914723 | bar | 1.922227 |
| art | 2.868612 | breakfast | 1.906325 |
| dalston | 2.672131 | view | 1.832262 |
| chelsea | 2.6605363 | pizza | 1.725505 |
| bermondsey | 2.6585836 | chicken | 1.723986 |

### 5.2.3 Textual matching

Two methods were used to match between salient topics and geographic objects. The first method is a name search in which the name of the object is compared to the list of salient topics. OSM data was taken as the database of geographic objects. This dataset contains objects which might have the name in their properties (a large number of objects do not have a name, therefore they are excluded from this analysis). The name may contain several words (e.g. "King's Cross Baptist Church"), and one or more of these words can be in the list of topics. In such a case, the object is selected as a match. Based on the number of words in the object's name, a matching score is calculated. For example, in the case of "King's Cross Baptist Church", we matched by searching for the keyword "church". With a match based on one word out of the four-word name, the matching score was set to 0.25.

The second method considers objects' semantic information by accessing their properties. In OSM, properties are attached by using keys and values.[16] In order to include the semantic properties of the objects, we searched for certain keys and/or values. The query keys and values were taken from a list (provided in Appendix A) in which words are linked to the combination(s) of keys and values. The content of this list is prepared using two methods of manual search and also querying online services, TagInfo[17] and TagFinder[18]. A more systematic approach is proposed in Yousaf & Wolter (2019). TagInfo provides the statistics of the usage of keys and values on the whole OSM dataset. TagFinder provides a search service in which one can search for words and find a ranked list of suggested tags for that word. An example of mapping between topic words and key-value combinations is matching between the word "market" and key-value combinations of *amenity=marketplace*; *shop=market*; *amenity=market*; *market=**; *marketplace=**.

By applying the matching between the topics and OSM objects, we are able to use those topics to detect objects that are related to (or represent) them. An important point regarding the motivation behind this step is that, although it is possible to apply a similar matching activity

[16] For example, in case of the mentioned church, the following keys and values are the important properties: amenity=place_of_worship; building=church; religion=christian; denomination=baptist

[17] https://taginfo.openstreetmap.org – accessed Nov. 2019

[18] http://tagfinder.herokuapp.com – accessed Nov. 2019

on the source data (i.e. applying matching between topics and Foursquare objects), doing so on OSM data provides several advantages. The first advantage is that OSM data provides more categorical diversity in objects compared to a limited set of object categories (i.e. venues in Foursquare). Besides the categorical diversity of objects, the geometry of the objects in OSM is more detailed. The superior degree of detail also holds regarding the semantic information stored in OSM. Last but not least (and from a pragmatic point of view), OSM as a collaborative non-profit source of data potentially offers a more stable long-term data source.

### 5.2.4 Pattern recognition

After applying the matching method (by name match, key-value match or both), each word was mapped to a group of geographic objects. After manually removing invalid results (objects matched to tokens such as colors, animal names and the like) and by observing the objects, three different patterns have been identified:

- Clustered objects with one center. The words matched to these objects are mainly city regions or neighborhoods. An example is the word “Hackney”, a district in London. We name this group “local pattern”.
- Clustered objects with more than one center. The words matched to these objects are concepts or activities with a locational nature (therefore they are not dispersed). An example is the word “Quay”. We name this group “multi-core pattern”.
- Dispersed objects. The words matched to these objects are concepts, which do not have a certain location but are available in many places in the whole study area. An example is the word “Square”. We name this group “global pattern”.

Figure 5.3 provides examples of these three patterns. The Foursquare dataset did not result in a multi-core example, but the Flickr dataset resulted in all three patterns.

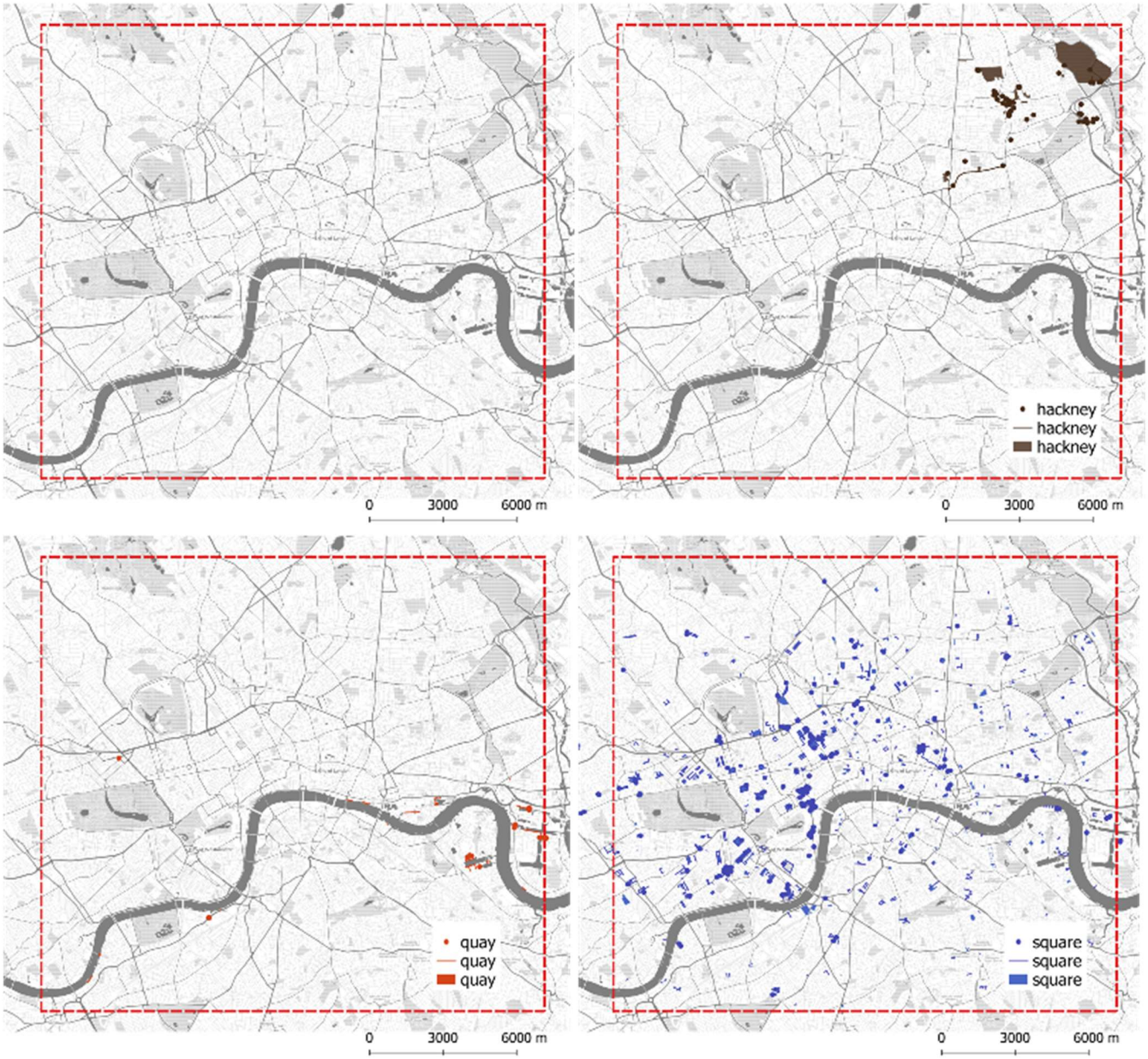


Figure 5.3 – Map of matched tokens. At the top left is the whole study area, at the top right is the map of objects matched to the word "Hackney", which is identified with the local pattern. The bottom left is the map of objects matched to the word "Quay" which is identified with the multi-core pattern. The bottom right is the map of objects matched to the word "Square", which is identified with the global pattern.

The patterns were first detected manually and by exploring and investigating the data. In order to automate the pattern recognition step, further investigation was done based on point pattern analysis methods over the data. For the objects, which were already modeled as points, the original data was used. For line and polygon objects, their centroid has been taken into consideration. This step enabled us to apply point pattern methods to the data that was not originally in the form of points.

To complete this classification task, we used different methods to examine different features of the data. Two features of the data were found helpful to perform this pattern recognition

task: the first derivative of Ripley's K-function (Ripley 1977) over distance (d) and the number of DBSCAN clusters. K-function by nature is an accumulative function, therefore the rate of change is not easily visible. Therefore, we used the first derivative of this function over the distance (d) to be able to observe the changes by increasing the distance. In the case of DBSCAN, we applied this clustering function (with different minimum points and epsilon values). By changing the epsilon parameter, we get different numbers of clusters, but the behavior of the number of clusters per distance (as parameter epsilon) is indicative.

Figure 5.4 shows exemplary results based on the three keywords of "Hackney", "Quay" and "Square". As it can be seen in this figure, the first derivative of the K-function helps us to detect the global pattern. Based on the nature of data in this pattern by adding to the search ratio (*d*) and up to a maximum value, there is always more data being added; therefore, the value of the function is growing. After reaching the absolute maximum, however, the function starts to decrease. On the other hand, the other two patterns have more than one extremum (none of them being an absolute maximum). After reaching a certain *d*, the value stays on a fixed number around 0.

We used the number of DBSCAN clusters for the next step and in order to differentiate between local and multi-core patterns. As can be seen in

Figure 5.4, the number of clusters for a multi-core pattern would stay equal to the number of cores or more. On the other hand, in the case of a local pattern, the number of clusters descends to 1 in short range (obviously in all cases that have a large epsilon, all the patterns converge to 1). In our test cases (based on the examples provided in this chapter), this convergence happened with epsilons smaller than 1000 meters.

Detecting the pattern helps us to firstly decide on the visual abstraction method and the map content. Besides that, the pattern clarifies the map extent and granularity of objects visualized.

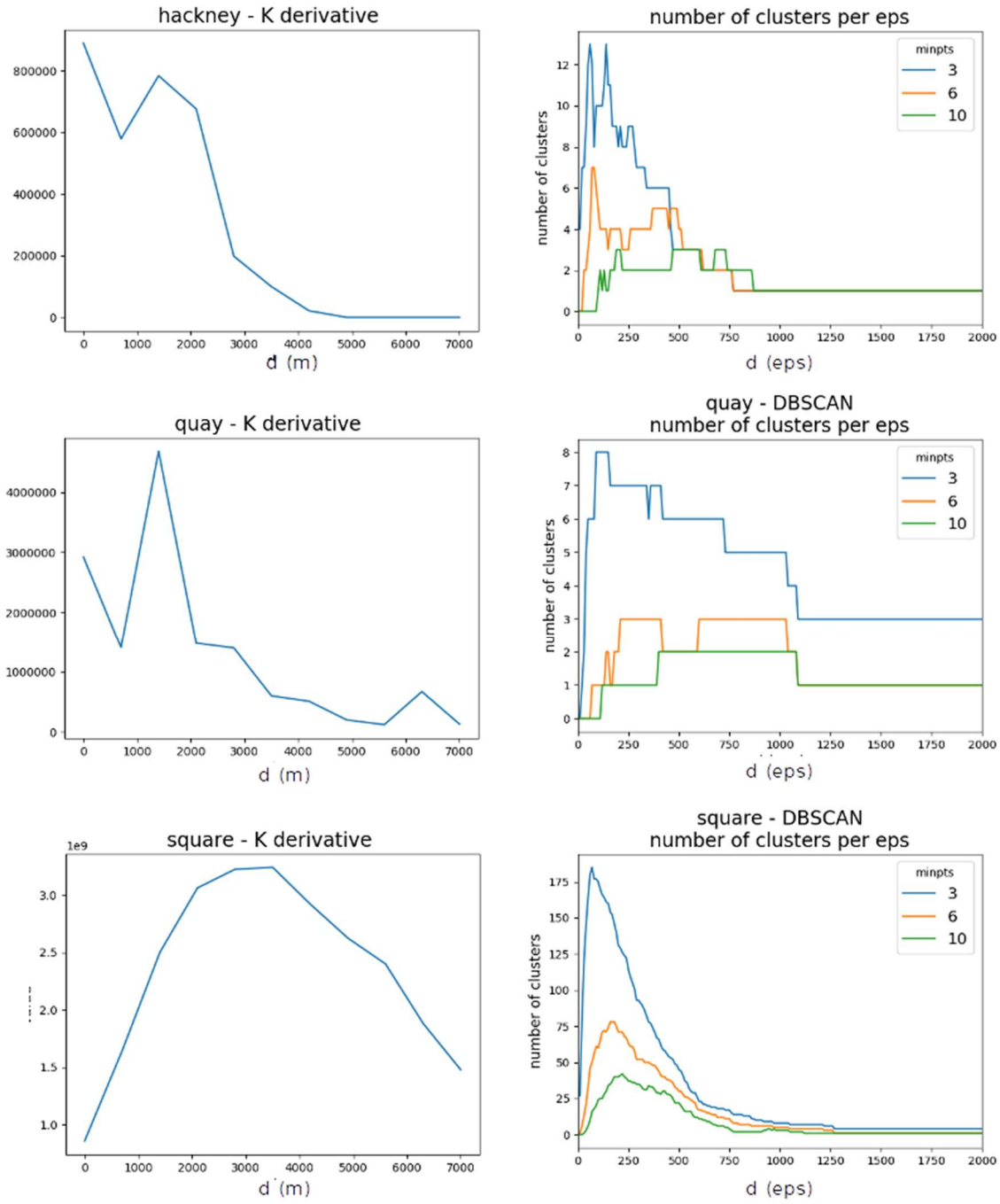


Figure 5.4 – Using two features of the data to detect the local, multi-core or global patterns. Diagrams on the left are the first derivative of K function. Diagrams on the right are the number of DBSCAN clusters. From top to bottom, the diagrams are based on “Hackney”, “Quay” and “Square”.

### 5.2.5 Visualization

Following the pattern detection step, the next step is triggered by the detected pattern. Here we propose the visualization methodology of two patterns (local and global patterns). The topics that are classified under the multi-core pattern should be investigated deeper by

understanding their cores. The decision on the visual representation of the multi-core pattern is based on the nature of the cores.

As it has been mentioned before and in the case of a "local" pattern, the topics are city neighborhoods. Therefore, in the case of having this pattern, we work towards visualizing a neighborhood by visualizing the representative objects of it. And by having the "global" pattern, we have observed that different classes of spatially dispersed features (based on a specific word in their name or based on a common key-value) are in the dataset.

### Local pattern

The words, which were detected to be a local pattern, are names of administrative divisions or neighborhoods (in different administrative levels). Therefore, in order to visualize the topic, we aim to find objects that represent a neighborhood. This task is similar to classical problems in map generalization, due to the fact that different urban object types are present in dense areas, and there is a need to make decisions about the classes to include in the map of certain intent and scale.

In order to base our method of visualizing neighborhoods/districts we have studied the cartographic conventions of the local National Mapping Agency (NMA). Ordnance Survey (OS), as the local NMA, offers different datasets and cartographic products.[19] The OS District Map[20] product contains settlements, named places, roads, woodlands and administrative boundaries. Based on OS recommendations, it is recommended to use this dataset for scales between 1:15,000 and 1:30,000. Another relevant dataset is the OS OpenMap Local,[21] which includes buildings, roads, sites, railways, hydrology, coastline, woodland, and cartographic text. OS recommends using this dataset for scales between 1:3,000 and 1:20,000. These two datasets have many commonalities. Their main differences are in labeling rules and level of detail (e.g. generalization of residential areas). We take the set of object classes from these two datasets and then query our database of geographic objects (OSM) to retrieve the relevant objects. It should be mentioned here that in order to have clear neighborhood boundaries (in different administrative levels), we used an official data source[22] openly available from the local government.

### Global pattern

In the case of a global pattern, we are firstly interested in understanding the reason for the matching of each object. If the matching is mostly based on a name match, there is a chance that the majority of the matched objects include a (combination of) specific tag(s). For example, in our data for the word "music", 90% percent of the objects are selected because they had the query word in their name and 10% of the objects are selected because of having the tag *music*. When considering the properties (keys and values) of the matched objects, some significant properties are present that are helpful for finding similar objects (those which have common properties but do not include the query word in their name).

[19] https://www.ordnancesurvey.co.uk/opendatadownload/products.html – accessed Nov. 2019

[20] https://www.ordnancesurvey.co.uk/business-government/products/vectormap-district – accessed Nov. 2019

[21] https://www.ordnancesurvey.co.uk/business-government/products/open-map-local – accessed Nov. 2019

[22] Statistical GIS Boundary Files for London – https://data.gov.uk/dataset/6cdebf5d-c69b-4480-8c9c-53ab8a816b9d/statistical-gis-boundary-files-for-london – accessed Nov. 2019

Based on the nature of OSM data and the lack of a clear and agreed-upon definition for using attached properties (keys and values) (Mooney & Corcoran, 2012a), some conventional analysis methods could not be used (e.g. a property like "*maxspeed*" can have different values: "none", "20" and "30_mph"). Therefore, in the first step, we considered key-value combinations with the lowest percentage of missing values and ranked the tags based on this property. This is a common data reduction technique to filter out properties that probably carry less information. It should be mentioned here that, due to the mixed nature of the values (e.g. numbers, categories, text), other data reduction methods have been ruled out. The second step was done to find correlated other key-values by looking at the correlation between the key-values in the dataset (matched objects). An example is the word "walk". The top three key-values with the least missing values are "*highway=footway*", "*maxspeed=20 mph*" and "*lit=yes*". Looking at the top three correlated key-values, we got (*foot=yes*, *surface=paved*, *bicycle=no*), (*highway=residential*, *sidewalk=both*, *surface=asphalt*) and (*sidewalk=both*, *surface=asphalt*, *maxspeed=30 mph*), respectively. Now it is possible to use the combination of key-values along with their correlated other key-values to query the dataset to give us objects which are not among the primary set of results. The interesting outcome is that, not only can we further enrich our results (by refining the search method mentioned in Section 5.2.3), but also this is more in the direction of reasoning between our search word (e.g. "walk") to combination(s) of key-values which describe the relevant objects based on their mutual properties (in other words, connecting between the search query "walk" and a *footway* with *lights* and a *speed limit*). Based on the example of the word "walk", Table 5.3 provides the key-values and the statistics of objects having the (combination of) key-values.

Table 5.3 – The top three key-values with the lowest missing values along with their most correlated keys-values. On the left are the key-values with the lowest missing values. On the right are the top key-values correlated to them. The number of objects shows the number of objects in the dataset that had the key-value or the combination of the main key-value and its correlated key-values.

| Missing values | | | Correlated key-values | | |
|---|---|---|---|---|---|
| key-value | % | number of objects | key-value | correlation coefficient | number of objects |
| highway=footway | 68.860 | 19299 | foot=yes | 0.1940 | 2059 |
| | | | surface=paved | 0.1767 | 130 |
| | | | bicycle=no | 0.1563 | 9 |
| maxspeed=20 mph | 78.143 | 15884 | highway=residential | 0.5340 | 8102 |
| | | | sidewalk=both | 0.2977 | 1100 |
| | | | surface=asphalt | 0.2596 | 504 |
| | | | lit=yes | | 451 |
| lit=yes | 78.966 | 18795 | sidewalk=both | 0.4528 | 5788 |
| | | | surface=asphalt | 0.4269 | 3323 |
| | | | maxspeed=30 mph | 0.3613 | 968 |
| | | | highway=tertiary | | 101 |

### 5.2.6 Automating the methodology

The methodology introduced here has the potential to be automated. Above, we introduced the different steps, their inputs/outputs and the important parameters in each step. Based on the current status of the methodology, it works in a semi-automatic manner, and in some steps, there is a need to check the results manually.

Figure 5.5 depicts the workflow of the proposed methodology, highlighting the steps that are not yet fully automated and which thus require human intervention. Starting from the beginning of the workflow, the data extraction from text, the textual matching and the pattern recognition steps are all automatic. Up to this step, the only non-automatic input is the mapping between words and the list of keys and values (introduced earlier in Section 5.2.3). In further investigations, we found it to be possible to also automate the extraction of this mapping (an example thereof is provided in Section 6.3.2). There is also a need to check the results of the textual matching and pattern recognition (mostly regarding the multi-core pattern) before moving further. In the visualization step and for the global pattern, we also included manually checking the results. In the latter case (as introduced earlier in Section 5.2.5), the reason to do the manual check is to limit the number of objects on the map. The methods to select the objects (the percentage of missing values and the correlation coefficient between keys and values) are well automatable but the human intervention is more about letting the user control avoiding to have too few or too many objects on the map. This could also potentially be replaced by a quantitative measure (e.g. number of objects), but the results should be approved by doing user tests.

In summary, while several steps of the proposed methodology are currently only semi-automated and require manual intervention, these steps are mostly about selecting thresholds and could be automated following further experimentation or the use of cartographic rules of thumb. They have not yet been automated in this thesis because the focus was on developing an overall methodology and demonstrating the effects of applying it in a particular case study.

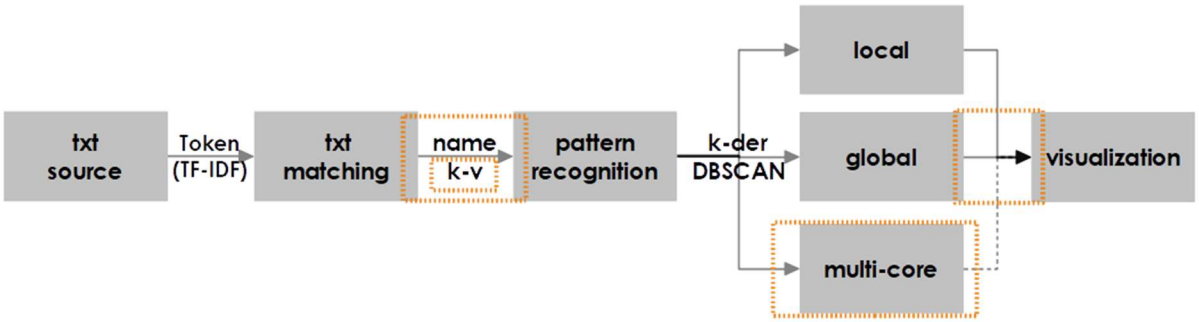


Figure 5.5 – In orange: the methodological steps which currently include manual intervention. k-v stands for key-value and k-der stands for K-function derivative over distance.

## 5.3 Results

In this section, we report the results of the methodology introduced in this chapter. We reported the results of the TF-IDF for the keywords based on our two different datasets (Flickr and Foursquare) already in Section 5.2.2. Here, we firstly report the results of the textual matching. Then, the results of the pattern recognition for a selected number of keywords along with their visualization results are reported.

### 5.3.1 Textual matching

The results of the textual matching (using a threshold value of 0.33) between the top twenty topics of each dataset are shown in Table 5.4 and Table 5.5 for Flickr and Foursquare datasets, respectively.

As it can be seen, the Flickr results include more topics that could be seen as placenames. These tokens tend to match to objects mostly based on their name, not the objects' attached properties, e.g. train stations with that name. The Flickr data's topics tend to match with the collection of objects that include point, line and polygon geometries. Generally, the number of lines (mostly streets or water bodies) is small, but in cases such as "Thames", this number is larger. In the case of Foursquare-matched objects, the number of line objects is mostly zero. It is straightforward to observe that the set of words extracted from Foursquare are more classes of objects (e.g. pub), as they were extracted from reviews attached to venues. In the case of Flickr dataset and based on the nature of data (tags attached to photographs), however, the extracted topics tend to include more placenames.

Table 5.4 – Statistics of the matched objects based on the topics extracted from the Flickr dataset. The number of objects that are matched based on their name and properties is reported for each topic.

| Flickr | | | | | |
|---|---|---|---|---|---|
| token | points | lines | polygons | name matches | property matches |
| square | 290 | 947 | 306 | 1536 | 7 |
| hackney | 78 | 18 | 42 | 138 | 0 |
| islington | 54 | 25 | 27 | 106 | 0 |
| deptford | 39 | 55 | 27 | 121 | 0 |
| battersea | 71 | 81 | 49 | 201 | 0 |
| camberwell | 38 | 45 | 24 | 107 | 0 |
| peckham | 57 | 67 | 34 | 158 | 0 |
| southwark | 48 | 55 | 33 | 136 | 0 |
| city | 204 | 265 | 155 | 624 | 0 |
| architecture | 5 | 0 | 54 | 6 | 53 |
| park | 834 | 825 | 1100 | 2001 | 758 |
| train | 1478 | 14 | 238 | 2 | 1728 |
| thames | 15 | 615 | 22 | 652 | 0 |
| building | 358 | 13 | 103157 | 252 | 103276 |
| people | 28 | 0 | 4 | 32 | 0 |
| clapham | 111 | 85 | 35 | 231 | 0 |
| art | 567 | 9 | 203 | 128 | 651 |
| dalston | 29 | 29 | 16 | 74 | 0 |
| chelsea | 73 | 54 | 57 | 184 | 0 |
| bermondsey | 31 | 27 | 17 | 75 | 0 |

Table 5.5 – Statistics of the matched objects based on the topics extracted from the Foursquare dataset. The number of objects that are matched based on their name and properties is reported for each topic.

| Fousquare | | | | | |
|---|---|---|---|---|---|
| token | points | lines | polygons | name matches | property matches |
| food | 1142 | 0 | 940 | 341 | 1741 |
| place | 439 | 1673 | 238 | 2034 | 316 |
| staff | 4 | 1 | 4 | 9 | 0 |
| service | 21 | 3 | 28 | 52 | 0 |
| coffee | 269 | 0 | 185 | 256 | 198 |
| pub | 745 | 0 | 815 | 16 | 1544 |
| beer | 24 | 0 | 25 | 25 | 24 |
| park | 834 | 825 | 1100 | 2001 | 758 |
| lunch | 1 | 0 | 0 | 1 | 0 |
| bar | 763 | 0 | 358 | 462 | 659 |
| breakfast | 9 | 0 | 3 | 12 | 0 |
| view | 20 | 12 | 14 | 46 | 0 |
| pizza | 102 | 0 | 69 | 171 | 0 |
| chicken | 49 | 0 | 59 | 108 | 0 |
| selection | 0 | 0 | 1 | 1 | 0 |
| room | 33 | 4 | 96 | 133 | 0 |
| time | 10 | 1 | 7 | 18 | 0 |
| burger | 46 | 0 | 22 | 68 | 0 |
| restaurant | 2365 | 0 | 1506 | 34 | 3837 |
| garden | 82 | 28 | 1544 | 198 | 1456 |

### 5.3.2 Pattern recognition and visualization

#### Global pattern

We selected two global patterns from Flickr (namely *park* and *art*) and two global patterns from Foursquare (namely *food* and *coffee*). Figure 5.6 shows the first derivative of the K-Function for the selected topic words.

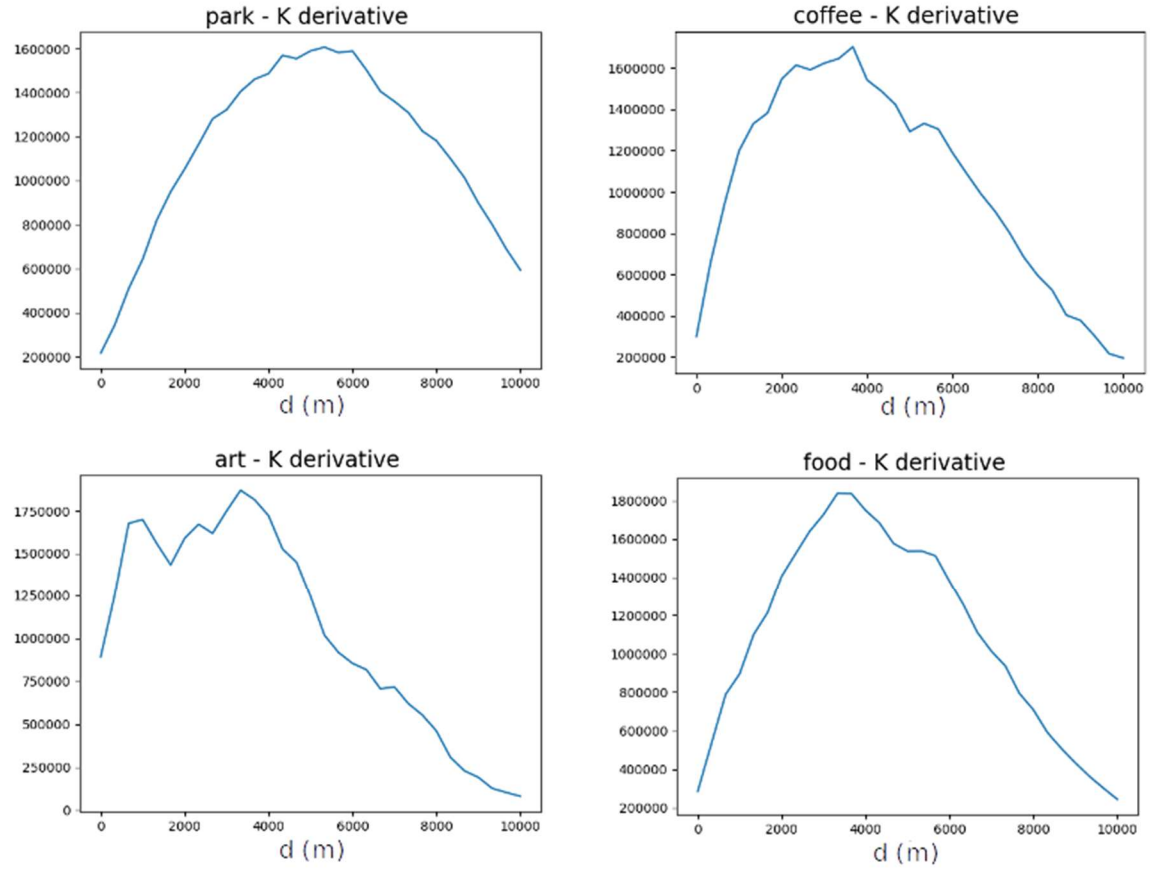


Figure 5.6 – First derivate of the K-function for the selected global patterns (park, art, coffee, food)

Following the methodology introduced before, the next step is to check the key-values of each dataset to find the key-values with the lowest missing value ratio. This is then followed by finding other key-values that have the most correlation to that key-value. We have applied thresholds for the missing values and the correlation coefficient. Based on observing our results, we have limited the ratio of missing values to be less than 80% and for the correlation coefficient, we have limited our results to coefficient with the value 0.2 as the minimum.

Table 5.6 – Analysis of key-values for "park". The gray cells are excluded from the visualization step.

| Missing values | | | Correlated key-values | |
|---|---|---|---|---|
| key-value | % | number of objects | key-value | correlation coefficient |
| leisure=park | 56.4550 | 872 | barrier=fence | 0.1617 |
| | | | access=private | 0.0781 |
| | | | landuse=park | 0.07304 |
| | | | landuse=grass | 0.05687 |
| lit=yes | 82.6332 | | surface_asphalt | 0.6814 |
| | | | maxspeed=20 mph | 0.6121 |
| | | | Sidewalk=both | 0.6053 |

| | | | | |
|---|---|---|---|---|
| maxspeed=20 mph | 87.8586 | | lit=yes | 0.6121 |
| | | | highway=tertiary | 0.4705 |
| | | | highway=residential | 0.4159 |
| | | | highway=primary | 0.3818 |

For the first example ("park"), Table 5.6 provides the data on the key-value analysis and Figure 5.7 shows the results. Based on the data and the values, we have only included the combination of "leisure=park" in our visualization step. This has helped us to find objects that are tagged as parks but do not have the word park in their name (e.g. "peckham rye common" and "mabley green"). Another advantage is to filter out less relevant objects which have the word "park" in their names (e.g. "Deptford Park School" bus stop) but are not parks (based on our query of leisure=park).

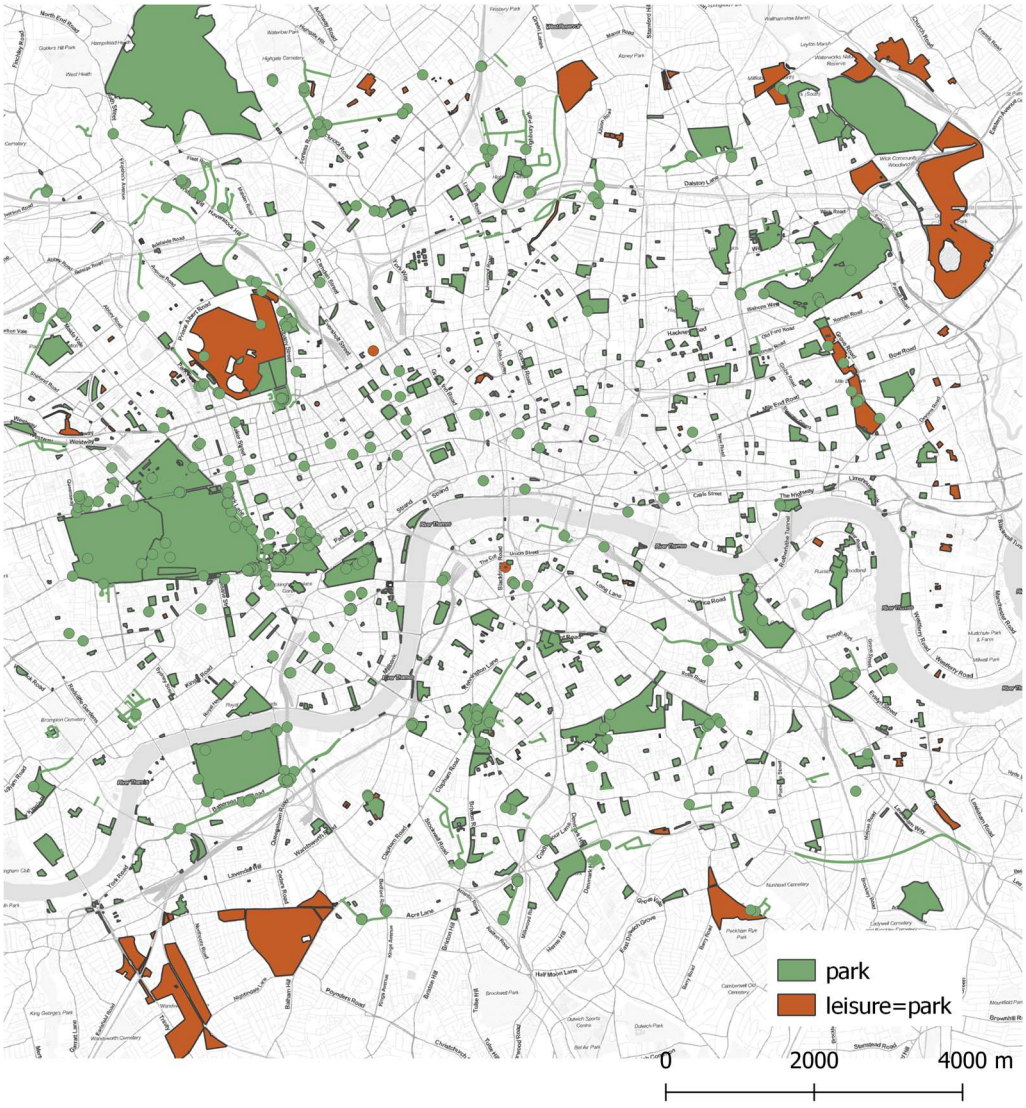


Figure 5.7 – Original data for the park (green) and the added data based on the key-value leisure=park (brown).

Table 5.7 – Analysis of key-values for art. The gray cells are excluded from the visualization step.

| Missing values | | | Correlated key-values | |
|---|---|---|---|---|
| key-value | % | number of objects | key-value | correlation coefficient |
| shop=art | 62.4516 | | level=0 | 0.1824 |
| | | | building:levels=4 | 0.1496 |
| | | | building:colour=white | 0.1238 |
| tourism =artwork | 64.5161 | | artwork_type=sculpture | 0.6139 |
| | | | artwork_type=statue | 0.2822 |
| | | | artwork_type=mural | 0.2019 |
| artwork_type= sculpture | 80.9032 | | tourism_artwork | 0.6139 |
| | | | artwork_group_year_of_the_bus_sculpture_trails | 0.1818 |
| | | | historic:tourism_artwork | 0.1659 |

For the second example (Art), Table 5.7 provides the data on the key-value analysis and Figure 5.8 shows the results. Based on the data and the values, we investigated the usage of two key-values. The first key-value is *shop=art*, which filters our result to show objects tagged as art shops. Looking at the data we observe a loss in the concentration of the objects and the general coverage of data over the whole study area. There are also not enough correlated key-values to continue the analysis based on this key-value. Based on the key-value with the second lowest ratio of missing values (*tourism=artwork*), the resulting objects are interesting, as not only the whole coverage is kept (and is therefore more appropriate for an overview visualization), but also some object are present in the results that carry a key-value (*tourism=artwork*) which was not among our list of tags in the textual matching step (list in Appendix B). These objects also do not have the word "art" in their name; therefore, these objects were not present in the initial dataset (the result of the matching step) but could be taken as relevant objects for the visualization step.

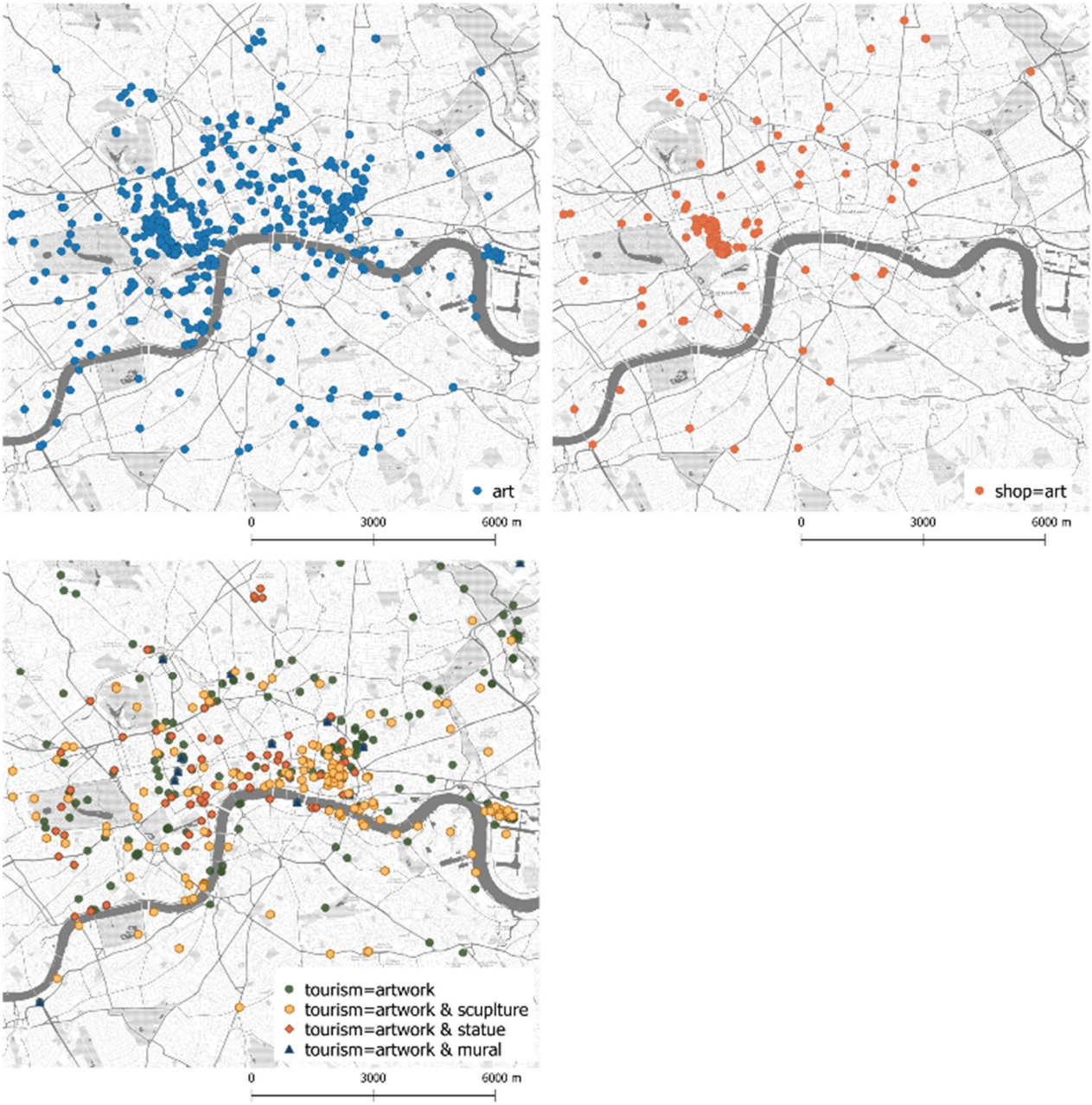


Figure 5.8 – Original data for "art" (top left) and the query based on the selected key-values (top right and bottom). The top right map shows the results for the query having *shop=art* and the bottom map shows the results for the query having *tourism=artwork* and three combinations ("sculpture", "statue", "mural").

For the third example ("coffee"), Table 5.8 provides the data on the key-value analysis and Figure 5.9 shows the results. Based on the data and the values, we investigated the usage of two key-values. The third key-value (*building=yes*) was excluded from our analysis due to a large number of retrieved objects. The first key-value is *amenity=cafe*, which filters our results to show objects tagged as cafes. This set of results is the objects that are explicitly tagged as cafes (directly relevant to our search word coffee). The second set of results retrieves objects by filtering for the key-value pair *cuisine=coffee_shop*. This set of results has a good coverage with the last set of results (for objects being tagged as both *amenity=cafe* and *cuisine=coffee_shop*) but the interesting part of the results are the objects that are tagged as classes other than cafes (e.g. restaurants) but offer coffee as part of their service.

Table 5.8 – Analysis of key-values for "coffee". The gray cells are excluded from the visualization step.

| Missing values | | | Correlated key-values | |
|---|---|---|---|---|
| key-value | % | number of objects | key-value | correlation coefficient |
| amenity=cafe | 8 | | cuisine=coffee_shop | 0.1828 |
| | | | outdoor=seating_yes | 0.0745 |
| | | | internet=access_wlan | 0.0684 |
| cuisine=coffee_shop | 49.5556 | | amenity=cafe | 0.1828 |
| | | | internet=access_wlan | 0.1695 |
| | | | internet=access_no | 0.1641 |
| building=yes | 75.3333 | | building:material_brick | 0.2387 |
| | | | building:levels_4 | 0.2242 |
| | | | building:colour_brown | 0.1852 |
| | | | building:levels_3 | 0.1765 |

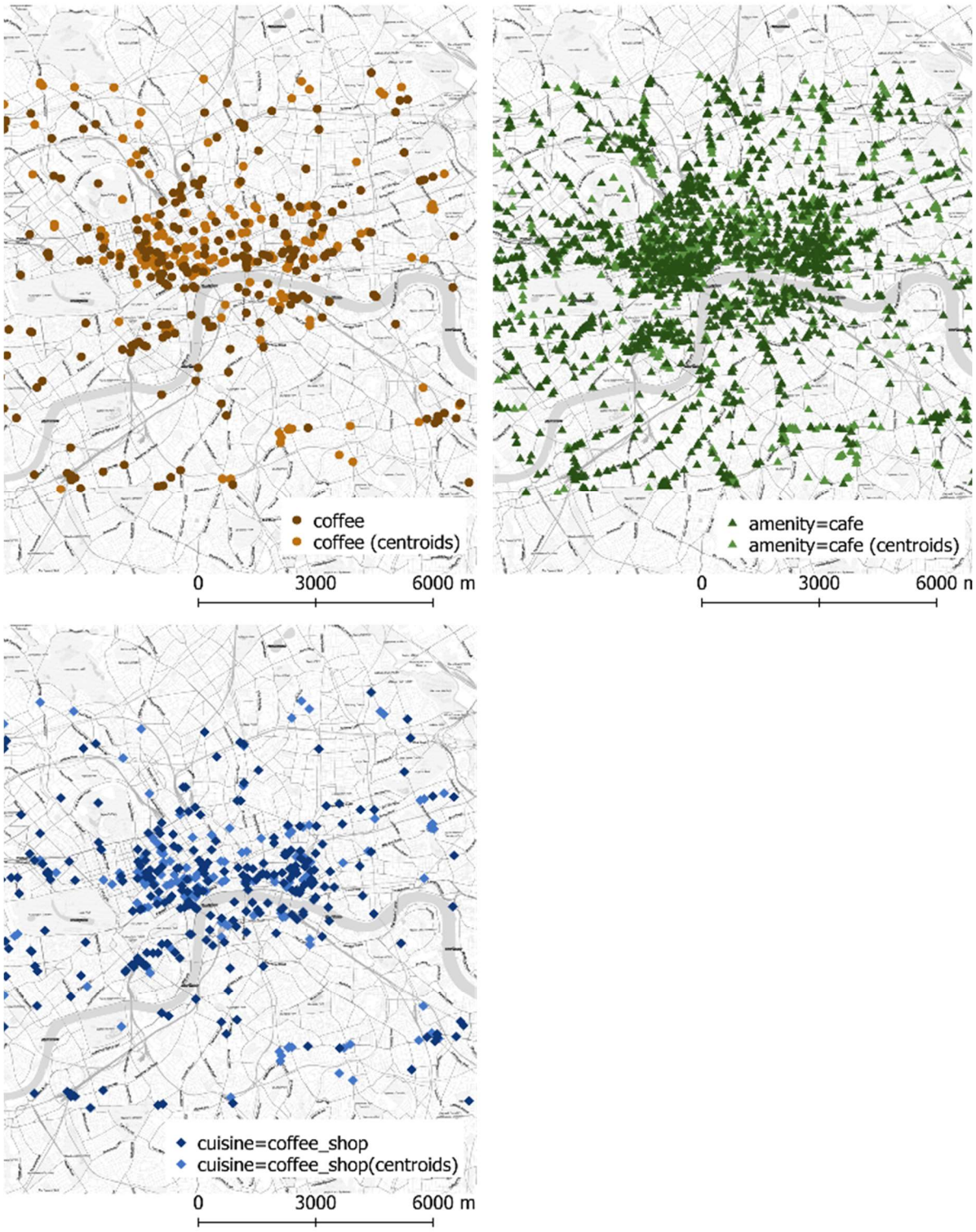


Figure 5.9 – Original data for coffee (top left) and the query based on the selected key-values (top right and bottom). All three maps depict both point and polygon (as centroids due to low scale) objects. The top right map shows the results for the query of *amenity=cafe* and the bottom map shows the results for the query of *cuisine=coffee_shop*.

Table 5.9 – Analysis of key-values for "food". The gray cells are excluded from the visualization step.

| Missing values | | | Correlated key-values | |
|---|---|---|---|---|
| key-value | % | number of objects | key-value | correlation coefficient |
| amenity=fast_food | 28.7317 | | takeaway=yes | 0.17166 |
| | | | cuisine=sandwich | 0.1537 |
| | | | cuisine=chicken | 0.1445 |
| | | | cuisine=burger | 0.1340 |
| building=yes | 74.2927 | | building:levels=4 | 0.2709 |
| | | | amenity=pub | 0.2548 |
| | | | food=yes | 0.2401 |
| food=yes | 88.1463 | | amenity=pub | 0.8941 |
| | | | real_ale=yes | 0.6226 |
| | | | toilets=yes | 0.4546 |

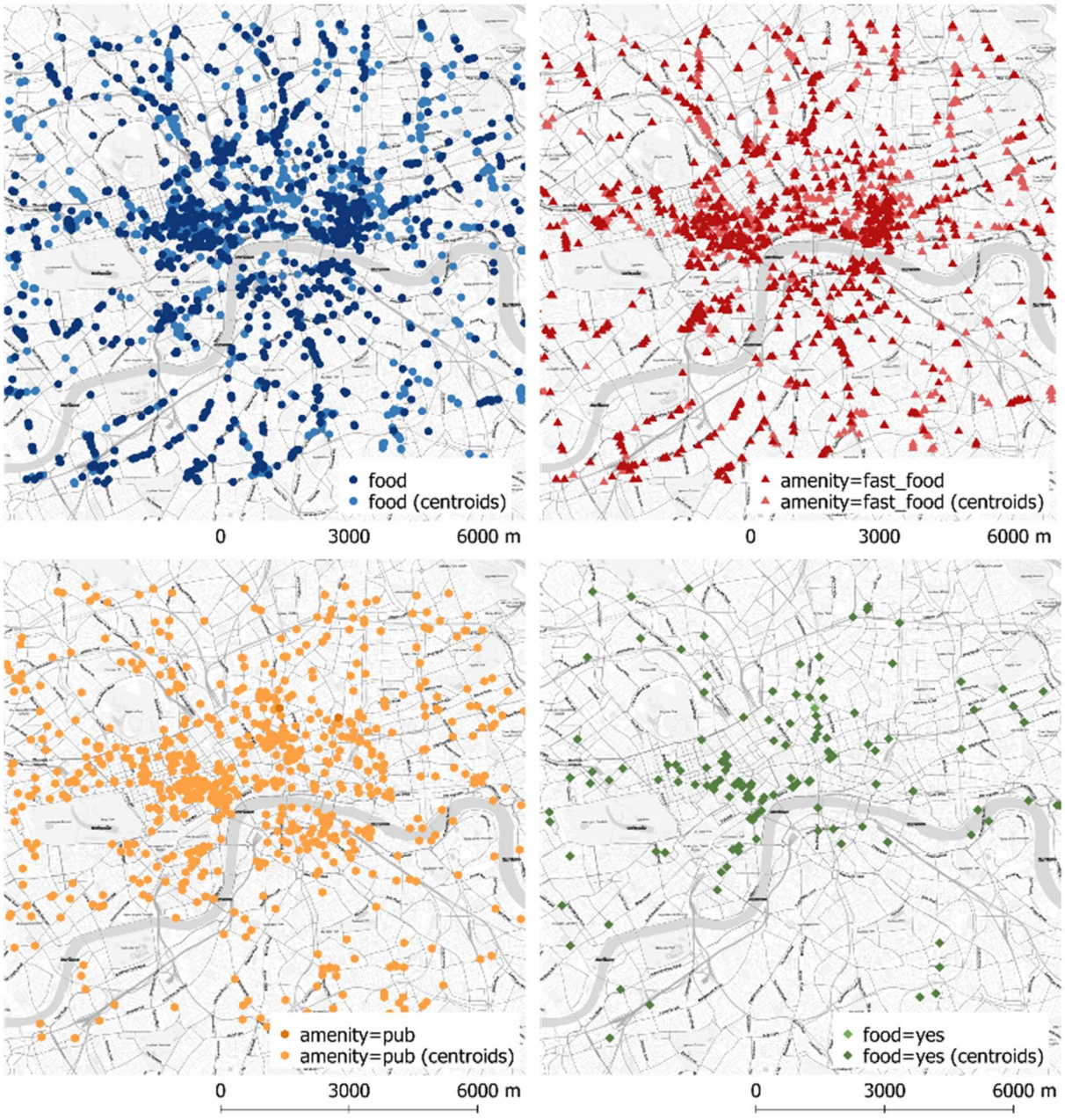


Figure 5.10 – Original data for "food" (top left) and the query based on the selected key-values. In all maps, in case of having the data in the form of polygon objects, their centroids are shown (due to low scale). The top right map shows the results for the query having *amenity=fast_food*, the lower left map shows the results for the query of *building=yes* and *amenity=pub* and the lower right map shows the results for the query of *building=yes* and *food=yes*.

For the fourth example ("food"), Table 5.9 provides the data on the key-value analysis and Figure 5.10 shows the results. Based on the data and the values, we investigated the usage of three key-value combinations. The first key-value is *amenity=fast_food,* which filters our result to show objects are explicitly tagged as fast food providers. Looking at the data, we observe some additional attributes in the form of keys and values that describe the properties of these objects (e.g. cuisine). The second key-value property to consider is *building=yes*. Similar to the case of "coffee", we avoided visualizing the results based on this key-value

combination as it is very general and also the number of objects is very large. The key-value with the highest correlation with *building=yes* is also not very helpful (*building:levels=4*), but we used the second and third top-correlated keys and values (namely *amenity=pub* and *food=yes*). By considering the combination of *building=yes* and *amenity=pub*, we get the pubs that provide food. The combination of *building=yes* and *food=yes*, however, provides a more varied set of results (which are typically amenities such as bars, pubs and restaurants). It should be mentioned here that in the mapping keywords and key-values, we considered two different entries for food and restaurant therefore the results of food did not contain many objects marked as restaurant.

The method of querying and visualization that has been proposed here is not always effective but is helpful to enrich the query process and also has the means of controlling the order of the results (e.g. in case of need to filter the results for avoiding visual clutter). This process is in the direction of making the implicit (the disperse presence of properties in the dataset) more explicit (the data-driven generation of queries).

The output is a typical generalization problem and, as a lot of semantic analysis has been already done, it is meaningful to apply geometrically driven rules (e.g. selection based on a minimum size threshold).

## Local pattern

We have selected two local patterns from Flickr (namely "*Hackney*" and "*Clapham*"). Figure 5.11 shows the first derivative of the K-function and the DBSCAN diagram for the selected topic words.

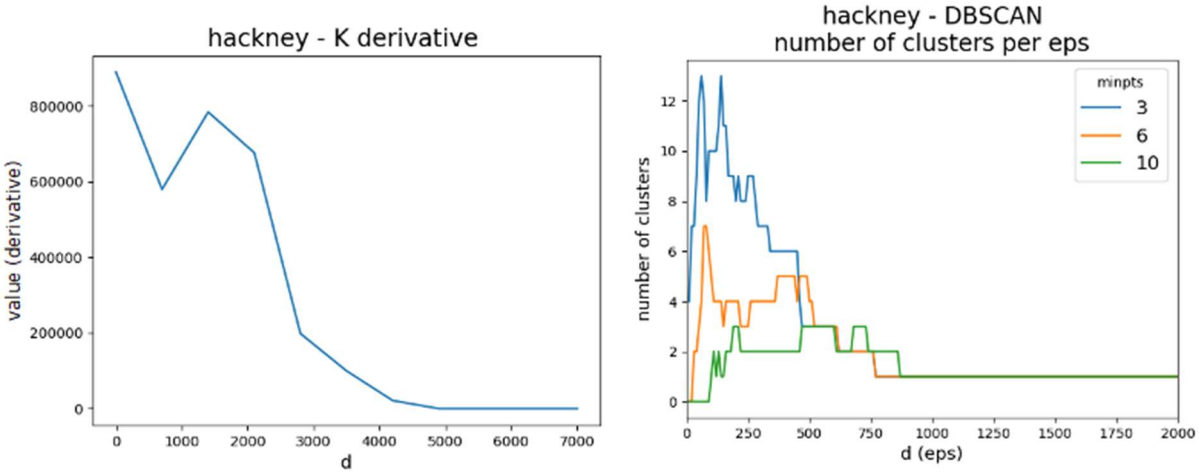


Figure 5.11 – K-function derivative and DBSCAN for "Hackney".

Based on the set of representative object types of residential areas, transportation network, urban green area and water bodies (based on the OS datasets and mentioned earlier in Section 5.2.5), we query OSM for the keys and values mentioned in Table 5.10. Figure 5.12 provides the result of the abstraction for the word Hackney.

Table 5.10 – Selected key-values for querying OSM for the local pattern

| Object class | Key-values |
|---|---|
| Residential areas | landuse=residential |

| | |
|---|---|
| Transportation network | highway=primary, highway=secondary, highway=motorway railway=rail, railway=subway |
| Urban green area | leisure=park, leisure=garden, leisure=pitch, landuse=forest |
| Water bodies | water=*, natural=water, leisure=nature_reserve, landuse=reservoir |

Figure 5.12 shows the results of querying OSM for the mentioned object classes for the example case of Hackney, while Figure 5.13 shows the results for the topic word "Clapham".

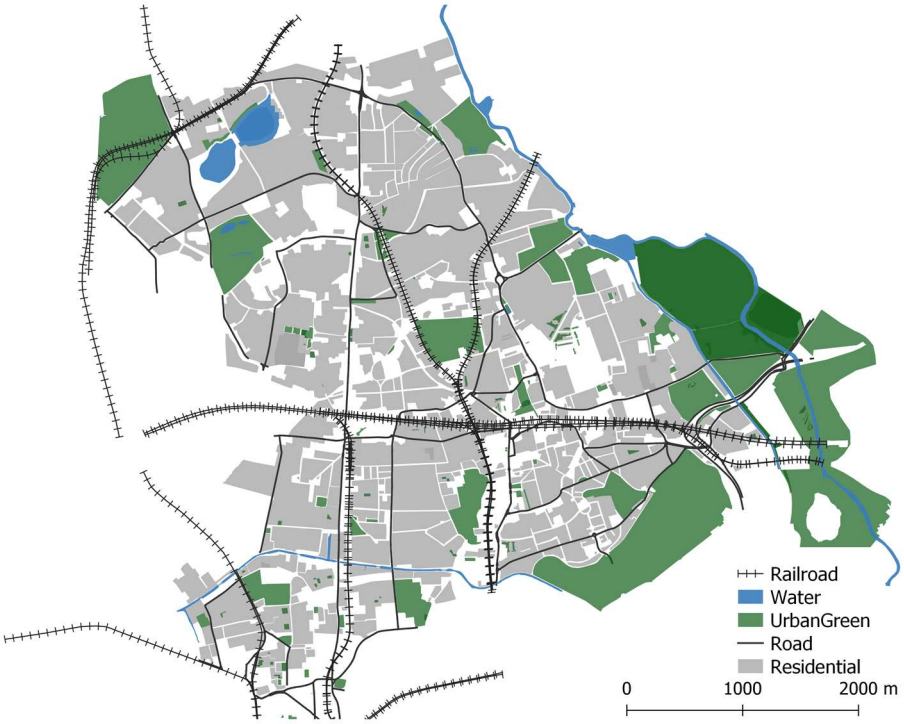


Figure 5.12 – Visualizing the results of querying OSM for the local topic "Hackney"

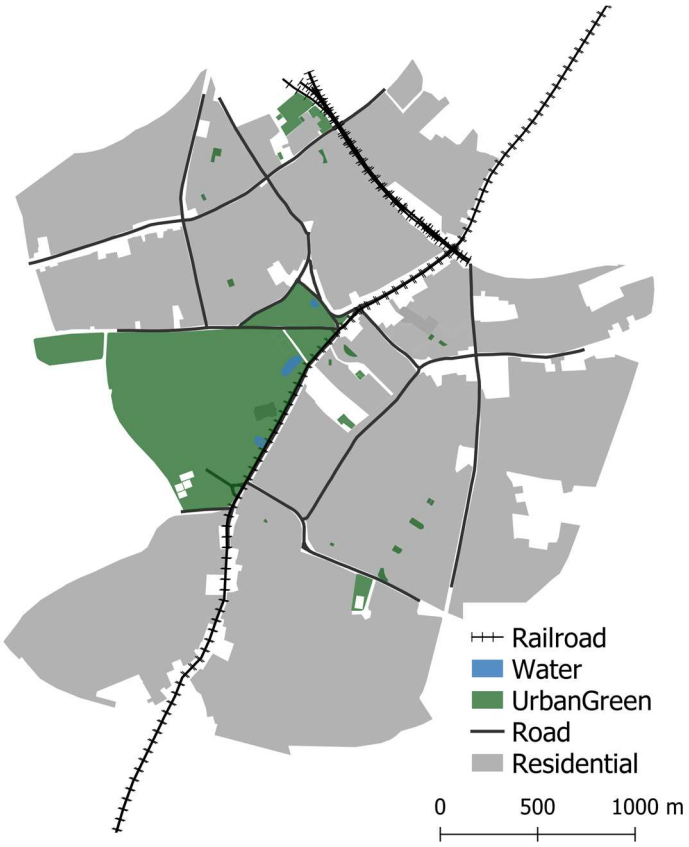


Figure 5.13 – Visualizing the results of querying OSM for the local topic "Clapham"

Similar to the global pattern, after querying the geographic database (OSM), the results are typical map generalization problems. There exists the potential to apply generalization operators such as selection (e.g. based on minimum size), aggregation (e.g. based on proximity and belonging to a similar class), displacement and the like.

## 5.4 Discussion

In this chapter, the method of providing abstractions in the form of maps based on textual UGC contributions has been provided. We observed different sets of topics extracted from different sources (as presented before in Table 5.1 and Table 5.2). Already in this step, we noticed the difference between the results of Flickr and Foursquare. This fact is based on the different nature of the data of these sources. But after polishing our method by testing the results of Foursquare (which is more in the form of natural language), we believe that our method is well generalizable to be applied to other textual sources.

Based on the results of the matching step (presented in Table 5.4), we could retrieve relevant objects from OSM dataset by using a combination of a name matching and a property matching method. As we have observed earlier in Table 5.4, different topics have different ratios of results triggered by each matching method. For example, the topic "coffee" resulted in a balanced set of results, where the topic "peckham" resulted in more name matches than property matches and the topic "art" resulted in more property matches than name matches.

When interpreting this case, it is straightforward to observe that the nature of the query words plays an important role in the balance of matched objects.

We deepened our analysis by providing the pattern recognition method, which helped us to differentiate between the local and global patterns and to filter results that need further considerations (multi-core pattern). When analyzing the results of the global pattern, we found benefits (e.g. results that do not carry the query word in their name but that carry properties that have high commonality with the retrieved objects) in enriching our queries using the statistical method introduced in Section 5.3.2.

The methodology introduced here has some limitations. Firstly, the result of the topic extraction step (mostly in the case of Flickr) reports words that are frequent in photos (e.g. colors and names of animals) but that could not be matched to geographic objects. Another limitation that could lead to future work is to base the analysis on a semantic model that is driven from OSM. This helps to detect words that are different but are about the same concept (e.g. cemetery and graveyard). The motivation behind using the mapping between words and key-value pairs (see Appendix B) was the lack of such a model.

## 5.5 Conclusion

In the research reported in this chapter, we started with a large number of UGC contributions (from two different sources) without a query word. In the interim steps, we extracted tokens and provided the matching methodology to match the tokens to geographic objects to be shown as abstract, data-driven maps. Finally, we provided query and visualization methods to depict the results to users.

The work presented here contributes mainly in areas of cartographic generalization and geographic information retrieval. The contribution of this chapter lies in proving steps in data fusion and abstraction of textual information in a geographic context.

# 6 Discussion

So far, we have presented the results for three research objectives and have provided the discussion on the findings of each objective separately in their respective chapters. Each objective is generally based on a need to provide methods of visualizing geographic information based on UGC. Comparing the research objectives, the main differences were in the criteria of the problem being addressed in each objective (input parameters, UGC sources and the overall goal). Revisiting the context of research of this dissertation introduced earlier in Section 1.3, each research context aspect is addressed differently in the solutions provided for each research objective. Research objectives 1 and 3 have focused deeper on the semantic aspect of data but this aspect was also included in the methodology of research objective 2. When considering the spatial aspect of the data, all three research objectives include spatial measures and concepts (e.g. distance and point/object clusters) in their methodology. On the user preferences aspect, the depth of including user preferences is declining from research objectives 1 to 3. In this chapter, we discuss the findings of this dissertation in an overarching and inter-objective manner.

## 6.1 Research objective 1: Integration of UGC-backed semantic measures in map generalization

The first research objective was aimed at providing a means to make better use of UGC-backed semantics in the process of map generalization. We approached this problem by basing our findings on extracting semantic measures from UGC. The problem definition was based on initial parameters, such as the location of the user as well as a list of his/her favorite object(s). We used relatively more information as input parameters than in the other research objectives. Also, the fact that more parameters are included in the queries makes the results more user specific and tailored for the user inputs.

### 6.1.1 Contributions

The contributions of this research objective are twofold. Firstly, we have extended a semantic similarity measure to be able to make better use of the data of a certain UGC source (OSM). Secondly, we have adapted two generalization operators to include the semantic similarity measure in their behavior.

#### Extension of semantic similarity measure

In the methodology proposed for this research objective, we have provided an extension to the dice semantic similarity measure Dice, (1945). This extension assisted us in including UGC knowledge from OSM by including the objects' semantic properties in the form of keys and values in the calculation (rather than including only keys). The motivation was to tackle the peculiarities related to this specific UGC source. Extending the semantic similarity calculation has provided the base to adapt two generalization operators (selection and aggregation operators).

Regarding the special case of OSM, in comparison to other UGC sources (e.g. Flickr), the fact that each object's additional properties are in the form of keys and values required the modification in semantic similarity calculations. In this case and based on the absence of formal semantics, there have been earlier efforts in modeling the contributions semantically.

Earlier proposals are not able to fulfill what our semantic similarity is able to capture. For example, the semantic ontology proposed by Codescu et al. (2011) is not able to differentiate values of with the same key (i.e. *amenity=university* is considered as close to *amenity=restaurant* as *amenity=waste_basket*). The tag-tag similarity proposed by Ballatore, Bertolotto & Wilson (2013) also can not capture the object-object similarity between map objects (as it is not clear how to project the tag-tag similarity onto the object-object similarity based on the objects' shared and non-shared set of tags). Therefore, the extension of the dice similarity measure has been considered.

The object-object semantic similarity helped us to find similar objects to be fed to the generalization operators. As provided in Section 3.4 and based on the test cases provided in Chapter 3, we have shown the benefit of the extended semantic similarity in finding objects from different classes that were selected because of having common key-value combinations. The semantic similarity was used by applying a threshold filter that has shown to be sensitive (cf. Section 3.3.2 and Figure 3.7). This could be improved by smoothing the sensitivity of the measure (e.g. by applying weights to the key-value pairs).

#### Adaptation of generalization operators

The previous contribution led us to include the extended dice similarity into the process of map generalization on the level of operators. This helped us to provide maps that rely on UGC-backed semantics. Earlier research has provided the means to include the semantics of data in the process of map generalization (e.g. aggregation of buildings based on their object class in (Lee et al. 2017). There has also been research in improving the adaption of OSM data into the process of map generalization (e.g. the harmonization of OSM data level of detail in (Touya & Baley, 2017). However, these adaptations have not addressed the inclusion of the semantics of OSM into the process of generalization. By inclusion of the UGC-backed semantics into the process of map generalization, we could improve tailoring our results to the users' queries. In this step, we mainly focused on selection and aggregation operators. These have been marked as very relevant when considering to include data from LBS sources (based on use cases from a user-generated LBS) (Burghardt, Dunkel & Gröbe, 2017).

For example, based on the selection operator and as illustrated before in Figure 3.4 and Figure 3.5, using the semantic similarity and by applying spatial and semantic weights, we could provide different sets of results. Using a rather high semantic similarity weight resulted in selecting objects that are more similar to the user's search but are farther from his/her location. This could be seen as a compromise that the user is willing to take in order to find better matches to his/her search by moving larger distances. Figure 3.6 illustrates this phenomenon for the aggregation operator. In general, we could combine the spatial (distance) and semantic (semantic similarity) thresholds to find the best ratio between the two weights.

Considering a more general view over the findings of this objective, we can base our geographic representations (in a mobile context) on the semantic content of the data (provided by other users) for users by knowing their preferences (through the proxy of favorite objects). This is a step into enabling future LBS to be able to provide maps for user needs.

### 6.1.2 Limitations

As introduced before and discussed in this chapter, the first contribution of this research objective is source dependent. The focus is on a UGC source (OSM) with the lack of a defined semantic structure as well as a clear regime of tagging but with data in a specific format (keys and values). In the case of using another source, there might be a need to change the semantic

similarity calculation method (e.g. by replacing it with a graph distance measure based on the ontology of the source data). The proposed method of calculating semantic similarity based on keys and values also includes a limitation when comparing values with each other. The current method only tests the values for being *equal* and does not provide a deeper way of comparing different values of the same key. As an example, the most common values for the key "wheelchair" are "yes", "no", "limited", "designated" and "bad". When comparing these values with each other, as they are not equal, the value part of the similarity would not get any point and only the key similarity gets a point but it is obvious that the (dis)similarity between these values is not the same. Another limitation is the generalization operators that have been proposed for adaptation. Besides selection and aggregation, the possibility of adapting other operators to include UGC-backed semantics should be investigated. Based on the rather large amount of information on different UGC datasets, the operators that facilitate abstraction and reduction of information (in contrast to operators such as enlargement or exaggeration) are desired. In further investigations in the semantic adaptation of cartographic generalization operators, these operators should be prioritized.

We have proposed our methodology for two operators. Further investigations might address other candidate operators, for example simplification, displacement and typification. Using UGC-backed semantic measures (e.g. semantic similarity or semantic importance) in making decisions and parameterization of simplification (to which extent each object class should be simplified), displacement (to which maximum extent each object classes is allowed to be displaced) and typification (which object class should be typified) might be considerations of follow-up research investigations.

## 6.2 Research objective 2: Generation and generalization of Regions Of Interest (ROIs)

In the second research objective, we focused on the generation and generalization of ROIs based on a set of POIs. In contrast to research objective 1, we did not take the coordinates of the user into account, and the parameters to the analysis were the query word of the user. The input objects are a set of objects that have been fetched from a UGC source and based on the query keyword. The focus of the methodology of this objective is on providing better visual representations; therefore, less focus is given to the semantic analysis of the content. In comparison to research objective 1, this objective is based on a more specific task (switching the visual representation of points) and is an investigation in geographic visualization rather than utilization of UGC semantics.

### 6.2.1 Contributions

The main contributions of research objective 2 are twofold. Firstly, we provided the methodology of switching between POIs and ROIs along with a quantitative measure to trigger this process. Secondly, we provided the criteria to evaluate the results. Considering a more general view over the findings of this objective, we have proposed a geographic visualization method (in a mobile context) to provide an overview of regions while reducing the amount of information and keeping the spatial distribution.

### Generation of ROIs

In addressing this objective, firstly we provided a quantitative trigger measure to switch the visualization from points (POIs) to regions (ROIs). Basing our methodology on the L-function (Besag 1977), we could assure that switching to the ROI representation (and back to POI in larger scales) is a reliable measure to be taken as the trigger to switch between POI and ROI representation. We have provided more detail in Section 4.3.2.

Besides providing the quantitative trigger measure, our contribution was the process of generating the ROIs. More specifically and in comparison to earlier research, e.g. Lamprianidis et al. (2014) and Adams, McKenzie & Gahegan (2015), the important aspect was to relate an ROI parameter (i.e. bandwidth) to the current view of the user. This has not been addressed before, mainly because of its different research focus. In our case, we have focused on the visualization rather than information retrieval aspects. By achieving this contribution, we provide a methodology that can generate ROIs based on the user's current map view without being constrained to certain scales.

Using KDE, we calculated the surface corresponding to the input parameters (bandwidth and an initial threshold). In case of violating constraints (mainly relative area occupied by the ROI), the methodology calculates threshold cut changes until meeting the criteria (more details on this process are provided in Section 4.3) have been suggested. An important aspect that potentially can be used in LBS is the possibility of combining the KDE surface in order to combine queries. This possibility enables the users to query the database for visualizations of objects that fulfill multiple criteria. To enhance this possibility, the KDE surfaces could be combined by applying weights to emphasize the importance of different layers (surfaces). This aspect of our contribution is also not present in the earlier research, e.g. Adams, McKenzie & Gahegan (2015).

### Evaluation of ROI generation

In order to validate the generated ROIs, we have proposed a set of quantitative map readability measures (comprehensive details provided in Section 4.5). The proposed combination of measures has been based on earlier research (Harrie, Stigmar & Djordjevic, 2015; Stigmar & Harrie, 2011) and also contributions from this dissertation (tailoring the combination to our problem, using NNI and verifying the results using the quadrat count analysis). Looking at the measures based on the generated results, we have concluded, in general, that by reducing information (and replacing POIs with ROIs), the readability measures improve. The exception was the number of vertices (NV), which in some cases worsened (increased). To tackle this issue, we have proposed to apply a shape simplification algorithm. Having the set of map readability measures besides the L-function trigger measure helps us to apply POI to ROI representation change in a controlled and optimal manner. This is important to be considered in addition to the ROI generation method.

## 6.2.2 Limitations

The methodology of this objective focuses on the POIs that are present in the current map view of the user. Triggering the POI to ROI change via an L-function also takes the POIs in the current map view into account. Therefore, if the user pans or zooms in or out, the calculations need to be redone. This characteristic introduces a rather high load of calculations. The solution to this problem is to pre-calculate certain ROIs (based on frequent keywords and zoom levels/scales). This solution needs user tests to find out candidate query

words and zoom levels. Another important aspect that has not been addressed is limiting the results based on physical geographic constraints (e.g. to avoid generating ROIs on water bodies). This could already be partially addressed by assigning negative weights wherever the physical constraints are present (e.g. on lakes). A more general solution would be to modify the kernel function to include spatial constraints. Candidate solutions to this problem are directed KDE (Krisp & Peters, 2011), shape-constrained KDEs (e.g. Du et al., 2013) or network KDE functions (e.g. Tang et al., 2016).

## 6.3 Research objective 3: Abstraction of textual data by matching between geographic objects and themes

In the third research objective, we focused on extracting salient topics from two UGC sources (namely Flickr and Foursquare). Using the detected topics, we investigated matching between the set of geographic objects (OSM objects in the study area) and the extracted topics. By analyzing the matching results, we could understand the underlying patterns of the matched objects better. Finally, and based on the detected patterns, we could provide suggestions on the visualization of the results in the form of geographic maps.

In contrast to research objective 1, we did not include any user preference parameters, and therefore the results are more objective. Unlike in research objective 2, we did not include any current map view or query word from the user. Therefore, in contrast to the other research objectives, the results of research objective 3 are not based on queries or user inputs, but rather based on the data itself. In this research objective, we mainly focused on understanding salient topics and ways to indirectly geo-reference (by matching them to geographic objects) and abstract them.

### 6.3.1 Contributions

The contributions of research objective 3 are threefold. The first contribution was the proposal of a textual method to match between geographic objects (from a UGC source) and textual themes (derived from other UGC sources). Based on the results of the textual matching process, in our second contribution, we developed a quantitative pattern recognition method to classify the matching results into three different classes (local, global and multi-core). The third contribution was providing steps in the visualization of the patterns (based on two of the patterns).

#### Matching method between geographic object and topics

Analyzing UGC data typically requires matching different sources or clustering the contributions together based on their similarity or proximity. Integration (and matching) of UGC is also among the community research agenda (Yan et al., 2020). When applying the matching, the sides of the matching can be of different granularities. We have provided our matching method to match between geographic objects on one side (basing our methodology on OSM) to topics generated from textual analysis of UGC contributions (provided examples based on the data from Flickr and Foursquare). In our matching criteria, we used a combination of matching names and properties (similar to McKenzie, Janowicz & Adams, 2014). The name matching component was calculated based on finding the topic keyword in the name of the object, and the properties matching component was based on a mapping between topic keywords and combinations of OSM keys and values. A more systematic

approach to this mapping is provided by Yousaf & Wolter (2019). To our knowledge (as highlighted in Section 2.1.4), a matching investigation with similar matching sides has not been presented in earlier research. Our contribution is the textual matching proposal to combine name and properties on the granularity level of geographic objects on one side to the topic keywords on the other side. The results have helped us to geo-reference the topics (by using geographic objects as the proxy). As provided in Section 5.3.1, data from Flickr and Foursquare have resulted in different combinations of matching results. Flickr results included neighborhood names and object type names (e.g. building). Foursquare results included words more related to place types (e.g. pub and bar).

#### Pattern recognition

In order to move further in the investigation of the matched geographic objects, we analyzed the objects matched to different keywords. We have observed three patterns (local, global and multi-core) present in the data. The observation was initially done manually and, by analyzing the findings with quantitative spatial measures, we found the first derivation of K-function and the number of DBSCAN clusters to be appropriate measures to detect the patterns. Having these measures helps us achieve a reproducible means for future investigations. Considering the calculated measures also helps us to understand the topics better (e.g. their dispersion and number of cores).

#### Proposal for visual abstraction

We developed the methodology of visualization of each topic based on their quantitative measures from the previous contribution. For the global topics, we provided the method to enrich the input data by inferring the determinant keys and values for each topic. This was then followed by analyzing the correlated keys and values to be used as the query parameters for the enrichment step. For the local topics, we based our method on the practices of a local NMA and provided the proposal to query and fetch the relevant object classes to represent the detected neighborhoods. In both cases, we ended with typical cartographic generalization problems. When considering earlier research on providing visualization of topics or themes, earlier research has mainly focused on placing words (e.g. Martin & Schuurman, 2017) or word-clouds (Cidell, 2010; Tessem et al., 2015). Instead of representing topics by merely showing them as words, we found matches between the words and geographic objects. Therefore, we could show those objects as representations of the topic. For instance, instead of labeling certain regions with the word “food”, we found objects that are places that offered food or have included food-related words (e.g. “cuisine”). We then represented this word with those objects.

### 6.3.2 Limitations

Some building blocks of the research objective require manual procedures. An example is the mapping between keywords and OSM key-value sets. This step has been done by consulting OSM tag analysis tools such as TagInfo and TagFinder. Ideally, this mapping should be replaced by an automatic process or a similar concept (e.g. a graph). A candidate replacement for the manual approach is provided in Yousaf & Wolter (2019). In their proposal (more specifically their word-to-value similarity), the authors provide similarity measure calculations based on NLP analysis of OSM keys and values.

Another limitation of the research is about the findings of research objective 3 to be source specific. All the contributions of this research objective are OSM specific and therefore, by changing the geographic data source, the methodology should also be reconsidered. Finally, a limitation of the contributions of this objective is in finding no matching objects for some topic words. The extraction of topics from textual data often includes intervention and interpretation of the results, which is not a trivial task (Lansley & Longley 2016). In the case of finding no match, the process stopped at the first step and we were not able to visualize those topic words.

## 6.4 Practical LBS implementation considerations

When considering implementing the methodologies proposed in this dissertation in an LBS (or several LBSs), there exist several points to consider. The first point is the importance of performing user tests to verify the findings from a user perspective (e.g. what is the optimized weighing between semantic and spatial similarity for the methodology of the research objective 1?). Besides performing user tests for all three research objectives, each of the objectives might need further parameterization and fine-tuning. For example, and as it has been mentioned earlier in Section 4.5, practical implementation of the methodology of research objective 2 should include pre-computing KDE surfaces for certain map scales (LBS zoom levels). This could be done by selecting the list of most frequently searched keywords to generate their KDE surfaces. In another step, there is potential in improving the process of finding the appropriate threshold cut by using 3-dimensional triangulation. On the example of the proposed methodology for research objective 3, from a user point of view, it would be important to include a temporal aspect for the detected topics. The user would find the topics and their maps to be more useful if they reflect the temporality of the contributions (e.g. including carnival as a topic in a city with carnival tradition and ceremonies). In order to do so, the data that has been used for extracting topics should be filtered for different time snapshots.

The above-mentioned suggestions are provided as initial considerations and obviously there will be other practical challenges along the way when implementing the methods as conventional LBS applications aimed at end-users.

# 7 Conclusion

In this dissertation, we have made proposals about how using UGC can change the way geographic information is visualized. The motivation behind the investigations was the potential of UGC contributions (with a focus on textual data) and the fact that we could benefit by using them in the process of geographic information abstraction and visualization guided by cartographic generalization. Based on this general motivation, we have defined three use cases and consequently identified three research objectives. The use cases were always about a mobile user interacting with geographic information to get visual abstractions. In the first use case, the user's preferences were included in the visualization of the results to get maps that were semantically closer to his preferences. In the second use case, the user searched for a keyword to get an overview visualization of the results around him/her. In the third use case, the user provided no query keyword but was interested to explore the salient objects in his/her city of interest. In Chapters 3 to 5, we provided methodological proposals as responses to research objectives raised by investigating the use cases. In the following sections, we review the achievements, mention open issues and give an outlook.

## 7.1 Achievements

The main achievements of this thesis are the following:

- Adaptation of cartographic generalization operators to include UGC-backed semantics
- Methodology of density-triggered ROI visualization
- Textual matching method between topics and OSM objects
- Visualization method for local and global patterns

In the following sections, we focus on concluding these achievements.

### Adaptation of cartographic generalization operators to include UGC-backed semantics

This achievement was based on the need to adapt the process of map generalization to include UGC contributions. In order to address this challenge, firstly we proposed to extend the dice semantic similarity measure to deal with OSM data, owing to the fact that earlier methods were unable to handle the OSM key-value format. Secondly, we provided a solution to include the extended semantic similarity measure in novel selection and aggregation operators, thus allowing modification of the behavior of these generalization operators. With the combination of these two steps, we could provide visual responses (maps) to the mobile user by providing objects that have been semantically and spatially close to their queries (using their current position and favorite objects). By applying different sets of weights to the semantic and spatial similarity, it is possible to tweak the results to fulfill the user's needs (i.e. by providing the balance between the distance the user is willing to compensate to find the objects similar to his/her favorite objects).

### Methodology of density-triggered ROI visualization

In addressing the challenge of representing high densities of POIs, we have proposed to switch to ROI visualization. This is a need when representing large amounts of search results on conventional (mobile) maps. In this process and in order to be able to provide results at different scales, we were the first to provide a mapping between the current map view and kernel density bandwidth. We also provided the methodology to combine density surfaces in order to provide maps that fulfill multiple criteria (i.e. several query keywords). The transition from points to a region-based representation is meant to reduce the information load on the user and to replace a high number of points with a rather smaller number of regions. The developed methodology helps to reduce the amount of information on overview maps to a great extent (while retaining the spatial distribution).

### Textual matching method between topics and OSM objects

In order to find the representation of textual data in the form of geographic objects, we focused on finding matching between topics and OSM objects. Addressing this need led us to propose a textual matching method to find matches between the topics extracted from a UGC source (e.g. Foursquare) and OSM objects (with their key-value structure). We were the first to propose the methodology of matching between the topics on one side to the geographic objects on the other side. This helped us to geo-reference the topics (through OSM objects as proxies). Moreover, we proposed a quantitative method to analyze the results of the matching process.

### Visualization method for local and global patterns

As a follow up to the textual matching method, we proposed a visualization method for the local and global patterns. The local pattern visualization proposal was a mapping between OSM keys and values on the one side and the practiced methods of urban area visualizations on the other side. The global pattern visualization proposal was based on an iterative query method to address the question of why a group of objects are present in the matched objects. This has been approached by considering the statistics of the results. By applying correlation analysis (between different combinations of key and value pairs), we could successfully enrich our results to include the objects that have not been in the primary result set. The combination of this achievement and the textual matching method has helped us to go through a chain of steps to start with a large number of textual contributions and to end with a geographic visualization of the most important topics of the contributions. The overall process is beneficial to provide the users with abstract overview maps of the study area.

## 7.2 Open issues

Although this dissertation has led to some insightful achievements, some issues are still left open. First and foremost, there exists a high potential for evaluating the performance of the methods proposed in this dissertation by user testing. The focus of this dissertation was on developing methods for abstraction and visualization of geographic information based on UGC contributions. Different aspects of each research objective can potentially be evaluated through user tests based on the definition of different user tasks.

Another point to mention (more relevant to research objectives 1 and 3) is the lack of a semantic model based on OSM data. Although there have been earlier efforts on providing proposals for such models, it is still possible to develop models that provide deeper structures and will better fit problems similar to ours. We overcame this problem by using the key-value information from OSM in a different manner (in research objectives 1 and 3). With the

presence of an appropriate model based on the recorded semantic properties of OSM objects, the methods of this dissertation might have been developed differently. For example, by having a detailed ontological model of the OSM key-value semantics (e.g. in the form of a knowledge graph similar to the approach suggested by Patel, Paraskevopoulos, & Renz, 2018), the semantic similarity calculation of the OSM objects would have been based on this model rather than our key-value similarity measurement (research objective 1). Another example relates to the case of research objective 3, whereby having this ontological model we could have used the knowledge graph in our pattern recognition and visualization steps.

Finally, the research presented in this dissertation has not focused on the temporal aspects of UGC contributions (e.g. analysis of historical changeset of objects in OSM). We always accessed the data from our sources (OSM, Flickr and Foursquare) by taking a snapshot of the data for our study area. We believe that the methodological proposals of this dissertation are not greatly affected by the temporal dimension of the data. However, there exists the possibility of reconsidering and revisiting the objectives of this dissertation by taking the temporal aspects of the contributions into account.

## 7.3 Outlook

UGC-backed semantics in cartographic generalization

We have proposed including the extended dice similarity into the rather classic cartographic generalization model. We approached this by modifying the behavior of two operators. However, there exists a great potential to approach this problem from a more general perspective. This might lead to reconsidering and remodeling some components of cartographic generalization. One main concern would be to enhance the inclusion of conventional data (our focus is on UGC but other examples are point clouds, images and videos) in comparison to predefined, standard and (mainly) vector data from the NMAs. As part of this reconsideration, there is a great capacity for the inclusion of the UGC-backed semantics. The classic cartographic generalization model considers the semantics in a rather limited depth and assigns more room to considerations of the spatial/geometric component of the data. However, it is possible to focus on finding semantic measures to be fed to a generalization process that focuses on personalization, representation (of the contributors) and being vario-scale rather than providing maps with predefined purposes, content, audience and scales. For example, future investigations on personalization might include the means to provide the map content (e.g. foreground objects, object aggregation, level of detail, complexity of the symbology, labeling language, etc.) based on the user's behavior in working with the mobile apps. The investigations might focus on the extraction of different semantic measures (e.g. objects of interest) and on classifying the user based on UGC contributions.

OSM Semantic model

As mentioned in Section 7.2, a (detailed) semantic data model based on the contributions of OSM is missing. Earlier approaches, such as Codescu et al. (2011), cannot fulfill various needs (e.g. inclusion of keys and values) and are also relatively outdated. Taking OSM as one of the commonly used geospatial UGC sources, this should be further investigated. The future model should be able to provide answers to semantic queries (concepts relation and

similarity) and also be flexible in connecting to other geospatial sources (e.g. gazetteers) and semantic sources (e.g. knowledge databases). An example of the latter integration is the inclusion of Wikidata[23] in OSM. Some OSM objects include links to Wikidata nodes, but no means of using the Wikidata knowledge in map generalization has been implemented yet.

### POI visualization methods for LBS

We have focused on the proposal of replacing high densities of points with regions and believe this proposal to be useful for the presented cases. However, there exists the potential to investigate finding point visualization alternatives in other situations that we have not focused on. Examples are focusing on other use cases (e.g. indoor navigation), other data types (e.g. 3D points) and also other mediums (e.g. circular displays in Serrano, Roudaut & Irani, 2017).

### Other UGC data

Another important topic for future work is to extend our methods to focus on other types of user-generated contributions. Besides focusing on the geospatial component of the data, we have based our methods mainly on the textual component of UGC, where future research might focus on other types of contributions (e.g. user-generated georeferenced drone videos). To achieve this, there is a need to understand the nature of the data and also to include information extraction methods (e.g. video processing) in the abstraction and generalization method.

[23] https://www.wikidata.org – accessed May 2020

# Bibliography


Adams, B., & McKenzie, G. (2013). Inferring Thematic Places from Spatially Referenced Natural Language Descriptions. In *Crowdsourcing Geographic Knowledge* (pp. 201–221). https://doi.org/10.1007/978-94-007-4587-2

Adams, B., Mckenzie, G., & Gahegan, M. (2015). Frankenplace : Interactive Thematic Mapping for Ad Hoc Exploratory Search Categories and Subject Descriptors. *Proceedings of the 24th International World Wide Web Conference (WWW 2015)*, 12–22. https://doi.org/10.1145/2736277.2741137

Aggarwal, C. C., & Wang, H. (2011). Text Mining in Social Networks. In *Social Network Data Analytics* (pp. 353–378). https://doi.org/10.1007/978-1-4419-8462-3

Anderson, P. (2007). *What is Web 2.0?: ideas, technologies and implications for education*.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic Patterns in Georeferenced Tweets Through Space-Time Visual Analytics. *Computing in Science & Engineering*, *15*(3), 72–82.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *ACM Sigmod Record*, *28*(2), 49–60. https://doi.org/10.1145/304182.304187

Bahrehdar, Azam R., & Purves, R. S. (2018). Description and characterization of place properties using topic modeling on georeferenced tags. *Geo-Spatial Information Science*, *21*(3), 173–184. https://doi.org/10.1080/10095020.2018.1493238

Bahrehdar, Azam Raha, Adams, B., & Purves, R. S. (2020). Streets of London: Using Flickr and OpenStreetMap to build an interactive image of the city. *Computers, Environment and Urban Systems*, *84*(December 2019), 101524. https://doi.org/10.1016/j.compenvurbsys.2020.101524

Ballatore, A. (2016). Semantic Challenges for Volunteered Geographic Information. In *European Handbook of Crowdsourced Geographic Information* (pp. 87–95). Ubiquity Press Ltd. https://doi.org/10.5334/bax.g

Ballatore, A., Bertolotto, M., & Wilson, D. C. (2013). Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems*, *37*, 61–81. https://doi.org/10.1007/s10115-012-0571-0

Ballatore, A., & De Sabbata, S. (2020). Los Angeles as a digital place: The geographies of user-generated content. *Transactions in GIS*, *24*(4), 880–902. https://doi.org/10.1111/tgis.12600

Batrinca, B., & Treleaven, P. C. (2014). Social media analytics: a survey of techniques, tools and platforms. *AI and Society*, *30*(1), 89–116. https://doi.org/10.1007/s00146-014-0549-4

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

Bauer, S., Noulas, A., Seaghdha, D. O., Clark, S., & Mascolo, C. (2012). Talking places: Modelling and analysing linguistic content in foursquare. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 348–357. https://doi.org/10.1109/SocialCom-PASSAT.2012.107

Beard, K. (1991). Constraints on Rule Formation. In B. P. Buttenfield & R. B. McMaster (Eds.), *Map Generalization: Making Rules for Knowledge Representation* (pp. 121–135). Longman Scientific & Technical.

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*, 438–441. https://doi.org/10.1.1.221.2822

Bégin, D., Devillers, R., & Roche, S. (2018). The life cycle of contributors in collaborative online communities -the case of OpenStreetMap. *International Journal of Geographical Information Science*, *32*(8), 1611–1630. https://doi.org/10.1080/13658816.2018.1458312

Benevenuto, F., Rodrigues, T., Cha, M., & Almeida, V. (2009). Characterizing User Behavior in Online Social Networks. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, 49–62.

Bereuter, P. (2015). *Quadtree-based Real-time Point Generalisation for Web and Mobile Mapping*. University of Zurich.

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), 271–281. https://doi.org/10.1080/15230406.2013.779779

Bereuter, P., & Weibel, R. (2017). Variable-scale maps in real-time generalisation using a quadtree data structure and space deforming algorithms. *International Journal of Cartography*, *3*(1), 134–147. https://doi.org/10.1080/23729333.2017.1304189

Besag, J. (1977). Contribution to the discussion of Dr. Ripley's paper. *JR Statist. Soc. B*, *2*, 193–195.

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., & Rampini, A. (2016). On predicting and improving the quality of Volunteer Geographic Information projects. *International Journal of Digital Earth*, *9*(2), 134–155. https://doi.org/10.1080/17538947.2014.976774

Brassel, K. E., & Weibel, R. (1988). A review and conceptual framework of automated map generalization. *International Journal of Geographical Information Systems*, *2*(3), 229–244. https://doi.org/10.1080/02693798808927898

Burghardt, D., Dunkel, A., & Gröbe, M. (2017). Generalisation and Multiple Representation of Location-Based Social Media Data. *20th ICA Workshop on Generalisation and Multiple Representation*.

Calenge, C. (2011). Home range estimation in R: the adehabitatHR package. *Office National de La Classe et de La Faune Sauvage: Saint Benoist, Auffargis, France*.

Cantador, I., Konstas, I., & Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. *Journal of Web Semantics*, *9*(1), 1–15. https://doi.org/10.1016/j.websem.2010.10.001

Cao, X., Cong, G., Jensen, C. S., & Yiu, M. L. (2014). Retrieving Regions of Interest for User Exploration. *Proceedings of the VLDB Endowment*, *7*(9), 733–744. https://doi.org/10.14778/2732939.2732946

Casoto, P., Dattolo, A., Omero, P., Pudota, N., & Tasso, C. (2010). Accessing, Analyzing, and Extracting Information from User Generated Contents. In *Handbook of Research on Web 2.0, 3.0, and X.0* (pp. 312–328). IGI Global. https://doi.org/10.4018/978-1-60566-384-5.ch018

Cecconi, A. (2003). Integration of cartographic generalization and multi-scale databases for enhanced web mapping [University of Zurich]. In *University of Zurich*. https://doi.org/OFEV_integrationSIG

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *Fourth International AAAI Conference on Weblogs and Social Media*.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, *21*, 4–28. https://doi.org/doi: 10.1057/palgrave.sj.8350066

Chamoso, P., Rivas, A., Martin-Limorti, J. J., & Rodriguez, S. (2017). A Hash Based Image Matching Algorithm for Social Networks. *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017. PAAMS 2017. Advances in Intelligent Systems and Computing*, *619*, 183–190. https://doi.org/10.1007/978-3-319-61578-3

Chen, G., Woodward, J., Chen, L., Hussain, I., Mirza, H. T., & Majid, A. (2012). A context-aware personalized travel recommendation system based on geotagged social media data mining. *International Journal of Geographical Information Science*, *27*(4), 662–684. https://doi.org/10.1080/13658816.2012.696649

Chen, H., Zheng, Z., & Ceran, Y. (2016). De-biasing the reporting bias in social media analytics. *Production and Operations Management*, *25*(5), 849–865. https://doi.org/10.1111/poms.12509

Chen, S., Lin, L., & Yuan, X. (2017). Social Media Visual Analytics. *Computer Graphics Forum*, *36*(3), 563–587. https://doi.org/10.1111/cgf.13211

Cidell, J. (2010). Content clouds as exploratory qualitative data analysis. *Area*, *42*(4), 514–523. https://doi.org/10.1111/j.1475-4762.2010.00952.x

Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T., & Rau, R. (2011). Osmonto an ontology of openstreetmap tags. *State of the Map Europe (SOTM-EU)*. http://www.informatik.uni-bremen.de/~okutz/osmonto.pdf

Crampton, J. W. (2009). Cartography: Maps 2.0. *Progress in Human Geography*, *33*(1), 91–100. https://doi.org/10.1177/0309132508094074

Crease, P., & Reichenbacher, T. (2011). Adapting Cartographic Representations to Improve the Information Seeking of LBS Users. *Proceedings of The International Cartographic Conference*.

Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: A system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, *27*(12), 2483–2508. https://doi.org/10.1080/13658816.2013.825724

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., Karagiorgou, S., Efentakis, A., & Lamprianidis, G. (2014). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, *29*(5), 720–741. https://doi.org/10.1080/13658816.2014.977905

De Berg, M., & Speckmann, B. (2011). Delineating Imprecise Regions via Shortest-Path Graphs. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 271–280.

de Sabbata, S., & Reichenbacher, T. (2012). Criteria of geographic relevance: An experimental study. *International Journal of Geographical Information Science*, *26*(8), 1495–1520. https://doi.org/10.1080/13658816.2011.639303

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. https://doi.org/https://doi.org/10.2307/1932409

Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. *The Canadian Cartographer*, *10*(2), 112–122. https://doi.org/10.1002/9780470669488.ch3

Du, P., Parmeter, C. F., & Racine, J. S. (2013). Nonparametric kernel regression with multiple predictors and multiple shape constraints. *Statistica Sinica*, *23*(3), 1347–1371.

Edwardes, A., Burghardt, D., & Weibel, R. (2005). Portrayal and generalisation of point maps for mobile information services. *Map-Based Mobile Services: Theories, Methods and Implementations*, 11–30. https://doi.org/10.1007/3-540-26982-7_2

Egger, M., & Lang, A. (2013). A Brief Tutorial on How to Extract Information from User-Generated Content (UGC). *KI - Künstliche Intelligenz*, *27*(1), 53–60. https://doi.org/10.1007/s13218-012-0224-1

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A Latent Variable Model for Geographic Lexical Variation. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, MIT, Massachusetts, USA, 9-11 October 2010*, *October*, 1277–1287. https://doi.org/10.1038/nrm2900

Elwood, S., Goodchild, M. F., & Sui, D. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590. https://doi.org/10.1080/00045608.2011.595657

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 226–231. https://doi.org/10.1.1.71.1980

Ferrari, L., Rosi, A., Mamei, M., & Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN '11*, 1. https://doi.org/10.1145/2063212.2063226

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3), 137–148. https://doi.org/10.1007/s10708-008-9188-y

Foerster, T., Stoter, J., & Köbben, B. (2007). Towards a formal classification of generalization operators. *Proceedings of the 23rd International Cartographic Conference, Moscow, Russia*.

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 239–248. https://doi.org/10.1109/SocialCom-PASSAT.2012.19

Galton, A., & Duckham, M. (2006). What Is the Region Occupied by a Set of Points? *International Conference on Geographic Information Science*, 81–98.

https://doi.org/10.1007/11863939_6

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, *21*(April), 446–467. https://doi.org/10.1111/tgis.12289

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J. A., McKenzie, G., Ju, Y., Gong, L., Adams, B., & Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, *31*(6), 1245–1271. https://doi.org/10.1080/13658816.2016.1273357

Gartner, G., Bennett, D. A., & Morita, T. (2008). Towards Ubiquitous Cartography. *Cartography and Geographic Information Science*, *34*(4), 247–257. https://doi.org/10.1559/152304007782382963

Gibin, M., Longley, P., & Atkinson, P. (2007). Kernel Density Estimation and Percent Volume Contours in General Practice Catchment Area Analysis in Urban Areas. *Proceedings of GISRUK*.

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

Goossen, M., Meeuwsen, H., Franke, J., & Kuyper, M. (2009). My Ideal Tourism Destination: Personalized Destination Recommendation System Combining Individual Preferences and GIS Data. *Information Technology & Tourism*, *11*(1), 17–30. https://doi.org/10.3727/109830509788714587

Gould, N., & Mackaness, W. (2016). From taxonomies to ontologies: Formalizing generalization knowledge for on-demand mapping. *Cartography and Geographic Information Science*, *43*(3), 208–222. https://doi.org/10.1080/15230406.2015.1072737

Grifoni, P., D'Ulizia, A., & Ferri, F. (2018). Context-Awareness in Location Based Services in the Big Data Era. In G. Skourletopoulos, G. Mastorakis, C. Mavromoustakis, C. Dobre, & E. Pallis (Eds.), *Mobile Big Data* (pp. 85–127). Springer, Cham. https://doi.org/doi:10.1007/978-3-319-67925-9_5

Grothe, C., & Schaab, J. (2009). Automated Footprint Generation from Geotags with Kernel Density Estimation and Support Vector Machines. *Spatial Cognition and Computation*, *9*(3), 195–211. https://doi.org/10.1080/13875860903118307

Hahmann, S., Purves, R., & Burghardt, D. (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, *9*(9), 1–36. https://doi.org/10.5311/JOSIS.2014.9.185

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*(1), 682–703. https://doi.org/10.1068/b35097

Haklay, M., Singleton, A., & Parker, C. (2008). Web Mapping 2.0 : The Neogeography of the GeoWeb. *Geography Compass*, *2*(6), 2011–2039. https://doi.org/10.1111/j.1749-8198.2008.00167.x

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

Harrie, L. (2003). Weight-Setting and Quality Assessment in Simultaneous Graphic

Generalization. *The Cartographic Journal*, *40*(3), 221–233. https://doi.org/10.1179/000870403225012925

Harrie, L., & Stigmar, H. (2010). An evaluation of measures for quantifying map information. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(3), 266–274. https://doi.org/10.1016/j.isprsjprs.2009.05.004

Harrie, L., Stigmar, H., & Djordjevic, M. (2015). Analytical Estimation of Map Readability. *ISPRS International Journal of Geo-Information*, *4*(2), 418–446. https://doi.org/10.3390/ijgi4020418

Harrie, L., & Weibel, R. (2007). Modelling the Overall Process of Generalisation. In A. Ruas, W. A. Mackaness, & T. Kilpeläinen (Eds.), *Generalisation of Geographic Information: Cartographic Modelling and Applications* (pp. 67–87).

Hart, T., & Zandbergen, P. (2014). Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting. *Policing: An International Journal of Police Strategies & Management*, *37*(2), 305–323. https://doi.org/10.1108/PIJPSM-04-2013-0039

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13*, 1. https://doi.org/10.1145/2505821.2505823

Hecht, B., & Gergle, D. (2010). On the "Localness" of User-Generated Content. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 229–232.

Henrich, A., Lüdecke, V., & Blank, D. (2008). Approaches for Determining the Geographic Footprint of Arbitrary Terms for Retrieval and Visualization. *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

Heredia, B., Prusa, J. D., & Khoshgoftaar, T. M. (2018). Social media for polling and predicting United States election outcome. *Social Network Analysis and Mining*, *8*(1), 1–16. https://doi.org/10.1007/s13278-018-0525-y

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve web search? *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 195–206. https://doi.org/10.1145/1341531.1341558

Hirsch, C., Hosking, J., & Grundy, J. (2009). Interactive visualization tools for exploring the semantic graph of large knowledge spaces. *CEUR Workshop Proceedings*, *443*(May).

Højholt, P. (2000). Solving Space Conflicts in Map Generalization: Using a Finite Element Method. *Cartography and Geographic Information Science*, *27*(1), 65–74. https://doi.org/10.1559/152304000783548028

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, *1*(1), 21–48. https://doi.org/10.5311/JOSIS.2010.1.3

Hornbæk, K., & Hertzum, M. (2011). The notion of overview in information visualization. *Journal of Human Computer Studies*, *69*(7–8), 509–525. https://doi.org/10.1016/j.ijhcs.2011.02.007

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Emergent Semantics in BibSonomy.

*GI Jahrestagung (2)*, 305–312.

Hu, Yingjie, Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, *54*, 240–254. https://doi.org/10.1016/j.compenvurbsys.2015.09.001

Hu, Yujie, Wang, F., Guin, C., & Zhu, H. (2018). A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied Geography*, *99*, 89–97. https://doi.org/10.1016/j.apgeog.2018.08.001

Huang, H., & Gartner, G. (2012). A technical survey on decluttering of icons in online map-based mashups. In M. P. Peterson (Ed.), *Online Maps with APIs and WebServices* (pp. 157–175). Springer. https://doi.org/10.1007/978-3-642-27485-5_11

Huang, H., Gartner, G., Krisp, J. M., Raubal, M., & Van de Weghe, N. (2018). Location based services: ongoing evolution and research agenda. *Journal of Location Based Services*, *12*(2), 63–93. https://doi.org/10.1080/17489725.2018.1508763

Huang, W., & Harrie, L. (2019). Towards knowledge-based geovisualisation using Semantic Web technologies: a knowledge representation approach coupling ontologies and rules. *International Journal of Digital Earth*, *0*(0), 1–22. https://doi.org/10.1080/17538947.2019.1604835

Huang, W., Kazemzadeh, K., Mansourian, A., & Harrie, L. (2020). Towards Knowledge-Based Geospatial Data Integration and Visualization: A Case of Visualizing Urban Bicycling Suitability. *IEEE Access*, *8*, 85473–85489. https://doi.org/10.1109/ACCESS.2020.2992023

Jenny, B., Jenny, H., & Räber, S. (2008). Map design for the Internet. *Lecture Notes in Geoinformation and Cartography*, *9783540720287*, 31–48. https://doi.org/10.1007/978-3-540-72029-4_3

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

Jones, C. B., Purves, R., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, *22*(10), 1045–1065. https://doi.org/10.1080/13658810701850547

Juhász, L., Novack, T., Hochmair, H. H., & Qiao, S. (2020). Cartographic vandalism in the era of location-based games-the case of open street map and Pokémon GO. *ISPRS International Journal of Geo-Information*, *9*(4), 1–20. https://doi.org/10.3390/ijgi9040197

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Keim, D. A. (2002). Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, *7*(1), 100–107.

Keim, D. A., Panse, C., & Sips, M. (2005). Information Visualization: Scope, Techniques and Opportunities for Geovisualization. *Exploring Geovisualization*, 21–52. https://doi.org/10.1016/B978-008044531-1/50420-6

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks.

*Dialogues in Human Geography*, *3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Kittur, A., & Kraut, R. E. (2008). Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, 37–46. http://kittur.org/files/KitturKraut_2008_CSCW_QualityCoordination.pdf

Kling, F., & Pozdnoukhov, A. (2012). When a City Tells a Story: Urban Topic Analysis. *Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12*, 482. https://doi.org/10.1145/2424321.2424395

Kraak, M.-J. (2011). Is There a Need for Neo-Cartography? *Cartography and Geographic Information Science*, *38*(2), 73–78. https://doi.org/10.1559/1523040638273

Krisp, J. M., & Peters, S. (2011). Directed kernel density estimation (DKDE) for time series visualization. *Annals of GIS*, *17*(3), 155–162. https://doi.org/10.1080/19475683.2011.602218

Lamprianidis, G., Skoutas, D., Papatheodorou, G., & Pfoser, D. (2014). Extraction , Integration and Analysis of Crowdsourced Points of Interest from Multiple Web Sources. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, 16–23.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Lee, J., Jang, H., Yang, J., & Yu, K. (2017). Machine Learning classification of buildings for map generalization. *ISPRS International Journal of Geo-Information*, *6*(10). https://doi.org/10.3390/ijgi6100309

Lei, T. L. (2020). Geospatial data conflation: a formal approach based on optimization and relational databases. *International Journal of Geographical Information Science*, 1–39. https://doi.org/10.1080/13658816.2020.1778001

Lemmens, R., Falquet, G., De Sabbata, S., Jiang, B., & Bucher, B. (2016). Querying VGI by semantic enrichment. In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, & R. Purves (Eds.), *European Handbook of Crowdsourced Geographic Information* (pp. 185–194). Ubiquity Press Ltd. https://doi.org/http://dx.doi.org/10.5334/bax.j

Li, J., Dai, H., Yuan, Z., Qin, Q., & Jiang, H. (2015). Extracting map information from trajectory and social media data. *ICSDM 2015 - Proceedings 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, 18–21. https://doi.org/10.1109/ICSDM.2015.7298018

Li, L., & Valdovinos, J. (2017). Information Fusion and Intelligent Geographic Information Systems. *Information Fusion and Intelligent Geographic Information Systems*, 227–241. https://doi.org/10.1007/978-3-319-59539-9

Li, X., Pham, T.-A. N., Cong, G., Yuan, Q., Li, X.-L., & Krishnaswamy, S. (2015). Where you Instagram? Associating Your Instagram Photos with Points of Interest. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, 1231–1240. https://doi.org/10.1145/2806416.2806463

Lu, Y., Hu, X., Wang, F., Kumar, S., Liu, H., & Maciejewski, R. (2015). Visualizing social media

sentiment in disaster scenarios. *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*, 1211–1215. https://doi.org/10.1145/2740908.2741720

Ma, D., Sandberg, M., & Jiang, B. (2015). Characterizing the heterogeneity of the openstreetmap data and community. *ISPRS International Journal of Geo-Information*, *4*(2), 535–550. https://doi.org/10.3390/ijgi4020535

Mackaness, W. A., & Chaudhry, O. (2013). Assessing the veracity of methods for extracting place semantics from flickr tags. *Transactions in GIS*, *17*(4), 544–562. https://doi.org/10.1111/tgis.12043

Maguire, D. J. (2007). GeoWeb 2.0 and Volunteered GI. *Workshop on Volunteered Geographic Information*, 104–106.

Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. *Proceedings of the 18th International Conference on World Wide Web, WWW' 09, ACM*, 641–650. https://doi.org/10.1145/1526709.1526796

Martin, M. E., & Schuurman, N. (2017). Area-Based Topic Modeling and Visualization of Social Media for Qualitative GIS. *Annals of the American Association of Geographers*, *107*(5), 1028–1039. https://doi.org/10.1080/24694452.2017.1293499

McKenzie, G., & Adams, B. (2018). A data-driven approach to exploring similarities of tourist attractions through online reviews. *Journal of Location Based Services*, *12*(2), 94–118. https://doi.org/10.1080/17489725.2018.1493548

McKenzie, G., Janowicz, K., & Adams, B. (2014). A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science*, *41*(2), 125–137. https://doi.org/10.1080/15230406.2014.880327

Meijers, M., van Oosterom, P., Driel, M., & Šuba, R. (2020). Web-based dissemination of continuously generalized space-scale cube data for smooth user interaction. *International Journal of Cartography*, *6*(1), 152–176. https://doi.org/10.1080/23729333.2019.1705144

Milo, T., & Zohar, S. (1998). Using Schema Matching to Simplify Heterogeneous Data Translation. *Proceedings of the 24rd International Conference on Very Large Data Bases (VLDB*, 122–133.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and Analysis of Online Social Networks. *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement,* 29–42.

Mokbel, M. F., Bao, J., Eldawy, A., Levandoski, J. J., & Sarwat, M. (2011). Personalization,Socialization, and Recommendations in Location-based Services 2.0. *5th International VLDB Workshop on Personalized Access, Profile Management and Context Awareness in Databases (PersDB)*.

Mooney, P., & Corcoran, P. (2012a). Characteristics of heavily edited objects in openstreetmap. *Future Internet*, *4*(1), 85–305. https://doi.org/10.3390/fi4010285

Mooney, P., & Corcoran, P. (2012b). The Annotation Process in OpenStreetMap. *Transactions in GIS*, *16*(4), 561–579. https://doi.org/10.1111/j.1467-9671.2012.01306.x

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). Towards quality metrics for

OpenStreetMap. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems GIS 10*, 514–517. https://doi.org/10.1145/1869790.1869875

Murugesan, S. (2007). Understanding Web 2.0. *IT Professional*, *9*(4), 34–41. https://doi.org/10.1109/MITP.2007.78

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2019). Textual keyword extraction and summarization: State-of-the-art. *Information Processing and Management*, *56*(6). https://doi.org/10.1016/j.ipm.2019.102088

Nazir, F., Ghazanfar, M. A., Maqsood, M., Aadil, F., Rho, S., & Mehmood, I. (2019). Social media signal detection using tweets volume, hashtag, and sentiment analysis. *Multimedia Tools and Applications*, *78*(3), 3553–3586. https://doi.org/10.1007/s11042-018-6437-z

Nickerson, B. G. (1986). evelopment of a rule-based system for automatic map generalization. *Proceedings of the 2nd International Symposium on Spatial Data Handling*.

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An Empirical Study of Geographic User Activity Patterns in Foursquare. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, *January*, 570–573. https://doi.org/papers3://publication/uuid/557455DB-AC4A-4C73-968A-31E7A663BC4E

Novack, T., Peters, R., & Zipf, A. (2018). Graph-based matching of points-of-interest from collaborative geo-datasets. *ISPRS International Journal of Geo-Information*, *7*(3), 1–17. https://doi.org/10.3390/ijgi7030117

O'sullivan, D., & Unwin, D. (2003). *Geographic information analysis*. John Wiley & Sons.

Olteanu-Raimond, A. M., Hart, G., Foody, G., Touya, G., Kellenberger, T., & Demetriou, D. (2017). The Scale of VGI in Map Production: A Perspective on European National Mapping Agencies. *Transactions in GIS*, *21*(1), 74–90. https://doi.org/10.1111/tgis.12189

Parker, J. K., & Downs, J. A. (2013). Footprint generation using fuzzy-neighborhood clustering. *GeoInformatica*, *17*(2), 285–299. https://doi.org/10.1007/s10707-012-0152-0

Patel, H., Paraskevopoulos, P., & Renz, M. (2018). GeoTeGra: A system for the creation of knowledge graph based on social network data with geographical and temporal information. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, 617–620. https://doi.org/10.1109/ASONAM.2018.8508674

Peng, D., & Touya, G. (2017). Continuously Generalizing Buildings to Built-up Areas by Aggregating and Growing. *3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, 1–8. https://doi.org/10.1145/3152178.3152188

Peng, D., Wolff, A., & Haunert, J. (2016). Continuous generalization of administrative boundaries based on compatible triangulations. *Geospatial Data in a Changing World: AGILE 2016, Lecture Notes in Geoinformation and Cartography*, 399–415. https://doi.org/10.1007/978-3-319-33783-8

Pippig, K., Burghardt, D., & Prechtel, N. (2013). Semantic similarity analysis of user-generated content for theme-based route planning. In *Journal of Location Based Services* (Vol. 7, Issue 4, pp. 223–245). Taylor & Francis. https://doi.org/10.1080/17489725.2013.804214

Purves, R., Edwardes, A. J., & Wood, J. (2011). Describing place through user generated content. *First Monday*, *16*(9).

Raper, J., Gartner, G., Karimi, H., & Rizos, C. (2007). A critical evaluation of location based services and their potential. *Journal of Location Based Services*, *1*(1), 5–45. https://doi.org/10.1080/17489720701584069

Rattenbury, T., Good, N., & Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07*, 103. https://doi.org/10.1145/1277741.1277762

Reichenbacher, T. (2004). *Mobile Cartography – Adaptive Visualisation of Geographic Information on Mobile Devices*. Technical University Munich.

Ripley, B. D. (1977). Modelling Spatial Patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(2), 172–212.

Rodríguez Domínguez, D., Díaz Redondo, R. P., Fernández Vilas, A., & Ben Khalifa, M. (2017). Sensing the city with Instagram: Clustering geolocated data for outlier detection. *Expert Systems with Applications*, *78*, 319–333. https://doi.org/10.1016/j.eswa.2017.02.018

Samal, A., Seth, S., & Cueto, K. (2004). A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, *18*(5), 459–489. https://doi.org/10.1080/13658810410001658076

Scheepens, R., Willems, N., Van De Wetering, H., Andrienko, G., Andrienko, N., & Van Wijk, J. J. (2011). Composite density maps for multivariate trajectories. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2518–2527. https://doi.org/10.1109/TVCG.2011.181

Scheepens, R., Willems, N., Wetering, H. Van De, & Wijk, J. J. Van. (2012). Interactive Density Maps for Moving Objects. *IEEE Computer Graphics and Applications*, *June 2014*. https://doi.org/10.1109/MCG.2011.88

Schmunk, S., Höpken, W., Fuchs, M., & Lexhagen, M. (2014). Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC. *Information and Communication Technologies in Tourism 2014*, 253–265. https://doi.org/10.1007/978-3-319-03973-2_19

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Mil inski, G., Nikši , M., Painho, M., P dör, A., Olteanu-Raimond, A.-M., & Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, *5*(5), 55. https://doi.org/10.3390/ijgi5050055

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

Serrano, M., Roudaut, A., & Irani, P. (2017). Visual composition of graphical elements on non-rectangular displays. *Conference on Human Factors in Computing Systems - Proceedings*, *2017-May*, 4405–4416. https://doi.org/10.1145/3025453.3025677

Sester, M. (2005). Optimization approaches for generalization and data abstraction. *International Journal of Geographical Information Science*, *19*(8–9), 871–897.

https://doi.org/10.1080/13658810500161179

Sester, M., Arsanjani, J. J., Klammer, R., Burghardt, D., & Haunert, J.-H. (2014). Integrating and generalising volunteered geographic information. In D. Burghardt, C. Duchêne, & W. Mackaness (Eds.), *Abstracting geographic information in a data rich world* (pp. 119–155). Springer International Publishing. https://doi.org/10.1007/978-3-319-00203-3

Sester, M., & Brenner, C. (2004). Continuous generalization for fast and smooth visualization on small displays. *International Archives of Photogrammetry and Remote Sensing*, *34*(4), 1293–1298. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.9129&rep=rep1&type=pdf

Shan, S., Ren, J., & Li, C. (2017). The dynamic evolution of social ties and user-generated content: a case study on a Douban group. *Enterprise Information Systems*, *11*(10), 1462–1480. https://doi.org/10.1080/17517575.2016.1177204

Shea, K. S., & Mcmaster, R. B. (1989). Cartographic Genralization in a Digital Environment: When and How to Generalize. *Proceedings of AutoCarto*, *9*, 56–67.

Shneiderman, B. (1996). Eyes have it: a task by data type taxonomy for information visualizations. *IEEE Symposium on Visual Languages, Proceedings*, 336–343. https://doi.org/10.1109/vl.1996.545307

Singh, T., & Kumari, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. *Procedia Computer Science*, *89*, 549–554. https://doi.org/10.1016/j.procs.2016.06.095

Spiess, E., Baumgartner, U., Arn, S., & Vez, C. (2002). Topographic Maps - Map Graphics and Generalisation. In *Cartographic Publication Series*. Swiss Society of Cartography.

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, *78*(2), 319–338. https://doi.org/10.1007/s10708-011-9438-2

Steiger, E., Ellersiek, T., & Zipf, A. (2014). Explorative public transport flow analysis from uncertain social media data. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. https://doi.org/10.1145/2676440.2676444

Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, *30*(9), 1694–1716. https://doi.org/10.1080/13658816.2015.1099658

Steiniger, S., Neun, M., & Edwardes, A. (2006). *Foundations of Location Based Services Lesson 1 CartouCHe 1- Lecture Notes on LBS: Vol. 1.0*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.1844

Stigmar, H., & Harrie, L. (2011). Evaluation of Analytical Measures of Map Legibility. *The Cartographic Journal*, *48*(1), 41–53. https://doi.org/10.1179/1743277410Y.0000000002

Stock, K. (2018). Mining location from social media: A systematic review. *Computers, Environment and Urban Systems*, *71*(March), 209–240. https://doi.org/10.1016/j.compenvurbsys.2018.05.007

Stoter, J., Zhang, X., Stigmar, H., & Harrie, L. (2014). Evaluation in Generalisation. In D. Burghardt, C. Duchêne, & W. Mackaness (Eds.), *Abstracting Geographic Information in*

*a Data Rich World* (pp. 259–297). Springer. https://doi.org/10.1007/978-3-319-00203-3

Šuba, R., Meijers, M., & van Oosterom, P. (2016). Continuous Road Network Generalization throughout All Scales. *ISPRS International Journal of Geo-Information*, *5*(8), 145. https://doi.org/10.3390/ijgi5080145

Tang, L., Kan, Z., Zhang, X., Sun, F., Yang, X., & Li, Q. (2016). A network Kernel Density Estimation for linear features in space–time analysis of big trace data. *International Journal of Geographical Information Science*, *30*(9), 1717–1737. https://doi.org/10.1080/13658816.2015.1119279

Tenney, M., Hall, G. B., & Sieber, R. E. (2019). A crowd sensing system identifying geotopics and community interests from user-generated content. *International Journal of Geographical Information Science*, *33*(8), 1497–1519. https://doi.org/10.1080/13658816.2019.1591413

Tessem, B., Bjørnestad, S., Chen, W., & Nyre, L. (2015). Word cloud visualisation of locative information. *Journal of Location Based Services*, *9*(4), 254–272. https://doi.org/10.1080/17489725.2015.1118566

Thebault-Spieker, J., Hecht, B., & Terveen, L. (2018). Geographic Biases are "Born, not Made": Exploring Contributors' Spatiotemporal Behavior in OpenStreetMap. *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, 71–82. https://doi.org/10.1145/3148330.3148350

Touya, G., Antoniou, V., Christophe, S., & Skopeliti, A. (2017). Production of Topographic Maps with VGI: Quality Management and Automation. In G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond, C. costa Fonte, & V. Antoniou (Eds.), *Mapping and the Citizen Sensor* (pp. 61–91). Ubiquity Press Ltd. https://doi.org/https://doi.org/10.5334/bbf.d

Touya, G., & Baley, M. (2017). Level of Details Harmonization Operations in OpenStreetMap Based Large Scale Maps. In M. Leitner & J. J. Arsanjani (Eds.), *Citizen Empowered Mapping* (pp. 3–25). Springer, Cham. https://doi.org/10.1007/978-3-319-51629-5_1

van Oosterom, P., & Meijers, M. (2011). Towards a true vario-scale structure supporting smooth-zoom. *14th ICA/ISPRS Workshop on Generalisation and Multiple Representation*, 1–19. http://www.gdmc.nl/publications/2011/True_vario-scale_structure.pdf

Venkateswaran, R. (2015). *Extracting and Linking Locations and Activities from the Geospatial Web in Support of Mobile GIS Applications*.

Waldner, W., & Vassileva, J. (2014). A visualization interface for twitter timeline activity. *CEUR Workshop Proceedings*, *1253*, 45–52.

Wand, M. P., & Jones, M. C. (1995). *Kernel Smoothing*. CRC Press.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. *Proceedings of the 2011 IConference*, 122–129. https://doi.org/10.1145/1940761.1940778

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). TwitterRank: finding topic-sensitive influential twitterers. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 261–270. http://portal.acm.org/citation.cfm?doid=1718487.1718520%5Cnpapers3://publication/doi/10.1145/1718487.1718520

Wu, Y., Cao, N., Gotz, D., Tan, Y. P., & Keim, D. A. (2016). A Survey on Visual Analytics of Social Media Data. *IEEE Transactions on Multimedia*, *18*(11), 2135–2148. https://doi.org/10.1109/TMM.2016.2614220

Xavier, E. M. A., Ariza-López, F. J., & Ureña-Cámara, M. A. (2016). A Survey of Measures and Methods for Matching Geospatial Vector Datasets. *ACM Computing Surveys (CSUR)*, *49*(2), 1–34. https://doi.org/10.1145/2963147

Yan, Y., Feng, C. C., Huang, W., Fan, H., Wang, Y. C., & Zipf, A. (2020). Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience. *International Journal of Geographical Information Science*, 1–27. https://doi.org/10.1080/13658816.2020.1730848

Yousaf, M., & Wolter, D. (2019). How to identify appropriate key-value pairs for querying OSM. *ACM International Conference Proceeding Series*, 1–6. https://doi.org/10.1145/3371140.3371147

Yu, Z., Wang, C., Bu, J. jun, Hu, X., Wang, Z., & Jin, J. he. (2017). Finding map regions with high density of query keywords. *Frontiers of Information Technology and Electronic Engineering*, *18*(10), 1543–1555. https://doi.org/10.1631/FITEE.1600043

Yuan, S., & Tao, C. (1999). Development of conflation components. *Proceedings of Geoinformatics*, *2*(2), 1–13. https://doi.org/10.2202/1948-4682.1069

Yürür, Ö., Liu, C. H., Sheng, Z., Leung, V. C. M., Moreno, W., & Leung, K. K. (2016). Context-awareness for mobile sensing: A survey and future directions. *IEEE Communications Surveys and Tutorials*, *18*(1), 68–93. https://doi.org/10.1109/COMST.2014.2381246

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), 1–25. https://doi.org/10.1002/widm.1253

Zhang, M., & Luo, L. (2019). Can User-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp. *SSRN*. https://doi.org/http://dx.doi.org/10.2139/ssrn.3108288

Zheng, Y. T., Li, Y., Zha, Z. J., & Chua, T. S. (2011). Mining travel patterns from GPS-tagged photos. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6523 LNCS*(PART 1), 262–272. https://doi.org/10.1007/978-3-642-17832-0_25

Zheng, Y., Wu, W., Chen, Y., Qu, H., & Ni, L. M. (2016). Visual Analytics in Urban Computing: An Overview. *IEEE Transactions on Big Data*, *2*(3), 276–296. https://doi.org/10.1109/tbdata.2016.2586447

Zhou, B., Liu, L., Oliva, A., & Torralba, A. (2014). Recognizing city identity via attribute analysis of geo-tagged images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8691 LNCS*(PART 3), 519–534. https://doi.org/10.1007/978-3-319-10578-9_34

Zhu, X., Wu, Y., Chen, L., & Jing, N. (2019). Spatial keyword query of region-of-interest based on the distributed representation of point-of-interest. *ISPRS International Journal of Geo-Information*, *8*(6). https://doi.org/10.3390/ijgi8060287

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 Guimarães, Portugal*, *1*, 1–15.

https://doi.org/10.1119/1.1736005

# Appendixes

## Appendix A: OSM tags black list

The following tags have been filtered from the properties of OSM features prior to application of key-value pairs in the semantic similarity analysis (research objective 1). The list was completed manually and is based on observing the data and test results.

'@id', 'name', 'alt_name', 'old_name', 'name:en', 'name:zh', 'note:name:en', 'note:name:zh', 'contact:phone', 'phone', 'email', 'website', 'contact:website','contact:email', 'contact:fax', 'url', 'wikimedia_commons', 'wikipedia', 'created_by', 'opening_hours', '@relations', 'postal_code', 'uk_postcode_centroid', 'postcode', 'source', 'ref', 'local_ref', 'uic_ref', 'uic_name', 'mobility:station_id', 'collection_times', 'ele', 'addr:district', 'addr:full', 'addr:housenumber', 'addr:suburb', 'addr:postcode', 'addr:street', 'addr:city', 'addr:country', 'addr:housename', 'addr:flat', 'addr:flats', 'addr:interpolation', 'addr:name', 'addr:number', 'addr:place', 'address', 'fixme', 'naptan', 'naptan:AtcoCode', 'naptan:CommonName', 'naptan:Bearing', 'naptan:Indicator', 'naptan:Street', 'naptan:verified', 'naptan:NaptanCode', 'naptan:Landmark', 'fhrs:id', 'fhrs:rating', 'fhrs:authority', 'fhrs:inspectiondate', 'fhrs:rating_date', 'fhrs:hygiene', 'fhrs:confidence_management', 'fhrs:local_authority_id', 'fhrs:structural'

# Appendix B: Mapping between tokens and OSM keys and values

| Token | Key-value combination(s) |
|---|---|
| academy | school:type=academy |
| allotment | landuse=allotments; place=allotments; allotments=* |
| animal | attraction=animal; landuse=animal_keeping |
| appliance | shop=electronics; shop=appliances |
| appliances | shop=electronics; shop=appliances |
| aquarium | tourism=aquarium |
| architecture | artwork=architecture; building:architecture=* |
| army | military=*; tank=* |
| art | shop=art; tourism=gallery; artwork=*; artwork_type=* |
| badminton | sport=badminton |
| baker | bakery=*; shop=bakeryl; craft=bakery |
| baking | shop=bakery; industrial=bakery; bakery=* |
| banksy | artist_name=Banksy |
| bar | amenity=bar; bar=* |
| basketball | sport=basketball |
| beach | natural=beach; leisure=beach_resort; beach=* |
| beer | beer_garden=*; drink:beer=*; beer=* |
| bicycle | bicycle=* |
| book | books=*; amenity=library; shop=books |
| bouldering | climbing=bouldering;sport=bouldering |
| brick | building:material=brick; building:Awalls=brick; building:cladding=brick; wall:material=brick; material=brick; wall=brick; building:structure=brick; building:facade:material=brick; surface=brick; |
| bridge | bridge=*; man_made=bridge; seamark_type=bridge; building=bridge; historic=bridge |
| building | building=*; historic=building |
| cafe | amenity=cafe; cafe=*; shop=cafe |
| canal | waterway=canal; water=canal; canal=* |
| castle | historic=castle; building=castle; castle=* |
| cathedral | building=cathedral |
| cemetery | landuse=cemetery; cemetery=* |
| chapel | amenity=place_of_worship; building=chapel; place_of_worship=chapel; historic=chapel; chapel=*; amenity=chapel |
| cheese | shop=cheese; produce=cheese; cuisine=cheese |
| church | amenity=place_of_worship; building=church; historic=church |
| circus | theatre:genre=circus |
| clock | amenity=clock; clock=*; tower:type=clock |
| club | club=*; leisure=club; amenity=club |
| coffee | cuisine=coffee_shop; shop=coffee; coffee=* |
| college | amenity=college; building=college |
| concert | amenity=concert_hall; theatre=concert_hall |
| concrete | surface=concrete; material=concrete |

| | |
|---|---|
| cottage | tourism=chalet; building=bungalow |
| court | amenity=courthouse; |
| cricket | sport=cricket |
| cycling | cycleway=*; sport=cycling; cycling=*; bicycle=* |
| dock | waterway=dock; dock=* |
| dockland | waterway=dock; dock=* |
| dome | building:roof:shape=dome; tower:construction=dome; building=dome; landmark=dome |
| eat | amenity=cafeteria; amenity=fast_food; food=*; amenity=food_court; shop=food; fast_food=* |
| eating | amenity=cafeteria; amenity=fast_food; food=*; amenity=food_court; shop=food; fast_food=* |
| event | advertising=event; amenity=event_venue |
| exhibition | amenity=exhibition_centre |
| farm | building=farm; landuse=farm; place=farm; farmland=*; farm=* |
| fashion | shop=fashion; clothes=fashion |
| ferry | route=ferry; amenity=ferry_terminal; ferry=* |
| ferris | route=ferry; amenity=ferry_terminal; ferry=* |
| festival | festival=*; amenity=festival_grounds |
| field | farmland=field; crop=field_cropland; landuse=field; field=* |
| flats | building:flats=*; building=apartments; |
| food | amenity=fast_food; food=*; amenity=food_court; shop=food; fast_food=* |
| football | sport=football; sport=soccer |
| fountain | amenity=fountain; water=fountain; fountain=* |
| fruit | fruit=* |
| gallery | tourism=gallery; shop=gallery; amenity=gallery |
| garage | building=garage; building=garages; amenity=garages; |
| garden | leisure=garden |
| gate | barrier=gate; historic=city_gate; gate:type=*; gate=*; entrance=gate |
| glass | building:material=glass; roof:material=glass; windows=glass; material=glass; |
| graffiti | artwork_type=graffiti |
| grave | amenity=grave_yard; cemetery=grave; grave=*; landuse=grave_yard; historic=grave; memorial=grave |
| graveyard | amenity=grave_yard; cemetery=grave; grave=*; landuse=grave_yard; historic=grave; memorial=grave |
| grocery | shop=grocery; shop=convenience; shop=supermarket |
| hall | building=hall; amenity=public_hall; amenity=community_hall; amenity=village_hall; amenity=concert_hall; amenity=hall |
| hamlets | place=hamlet |
| hill | natural=hill |
| history | historic=*; history=*; historical=* |
| hotel | tourism=hotel; building=hotel; hotel=* |
| house | building=house; building:type=house; historic=house |
| industry | building=industrial; landuse=industrial; usage=industrial; industrial=*; building:use=industrial; industry=*; amenity=industry |
| industrial | building=industrial; landuse=industrial; usage=industrial; industrial=*; building:use=industrial; industry=*; amenity=industry |

| | |
|---|---|
| inn | destination=Inn |
| installation | artwork_type=installation |
| junction | junction=*; highway=motorway_junction; railway=junction |
| korfball | sport=korfball |
| lake | water=lake; water:type=lake; waterway=lake; natural=lake |
| lawn | landuse=grass; |
| library | amenity=library; building=library; library=* |
| mall | shop=mall; building=mall |
| market | amenity=marketplace; shop=market; amenity=market; market=*; marketplace=* |
| meadow | landuse=meadow; meadow=*; natural=meadow |
| meat | shop=butcher; butcher=* |
| medical | shop=medical_supply; medical_supply=*; shop=medical; office=medical |
| memorial | historic=memorial; memorial=* |
| military | military=*; landuse=military; usage=military; access=military; building=military; historic=military |
| monument | historic=monument; monument=* |
| mound | natural=termite_mound; man_made=mound; site_type=tumulus |
| museum | tourism=museum; building=museum; museum=* |
| music | music=*; club=music |
| nature | nature=*; natural=* |
| natural | nature=*; natural=* |
| nonprofit | ownership=private_nonprofit; ownership=public_nonprofit |
| obstacle | obstacle=*; barrier=obstacle |
| office | office =*; building=office; building:use=office; building_type=office; amenity=office |
| opera | theatre:genre=opera |
| palace | castle_type=palace; historic=palace; building=palace |
| park | leisure=park |
| parliament | amenity=parliament; building=parliament |
| party | shop=party |
| petrol | amenity=fuel; |
| photographer | craft=photographer; shop=photographer |
| pitch | leisure=pitch |
| place | place=*; |
| plant | natural=grassland; leisure=garden; building=greenhouse |
| police | amenity=police; building=police |
| polling | polling_station=*; amenity=polling_station |
| pond | water=pond; landuse=pond; waterway=pond; pond=* |
| pray | amenity=place_of_worship; building=mosque; building=chapel |
| pub | amenity=pub; pub=* |
| quay | man_made=quay |
| railway | railway=*; landuse=railway; route=railway; historic=railway |
| ramp | ramp=* |
| recycling | amenity=recycling; recycling_type=* |
| restaurant | amenity=restaurant; restaurant=*; building=restaurant |

| | |
|---|---|
| river | waterway=river; waterway=riverbank; water=river; riverbank=*; river=* |
| road | highway=road; route=road; road=* |
| roads | highway=road; route=road; road=* |
| route | route=*; route=road; type=route |
| school | amenity=school; building=school; landuse=school; school=*; building:use=school |
| sculpture | artwork_type=sculpture; artwork=sculpture; amenity=sculpture; memorial=sculpture |
| ship | ship=* |
| shop | shop=*; amenity=shop; building:use=shop; building=shop; room=shop |
| show | show=* |
| slide | playground=slide |
| square | landuse=square; place=square |
| station | railway=station; amenity=bus_station; public_transport=station; building=train_station; station=* |
| statue | memorial=statue; artwork_type=statue; landmark=statue; monument=statue; artwork=statue |
| studio | amenity=studio; studio=* |
| suburbs | place=suburb |
| suburb | place=suburb |
| surgery | amenity=hospital |
| tennis | sport=tennis |
| temple | building=temple |
| theatre | amenity=theatre; building=theatre; theatre=* |
| therapist | office=therapist |
| tobacco | shop=tobacco; tobacco=* |
| tour | tourism=* |
| tower | man_made=tower; building=tower; tower=*; historic=tower; building=bell_tower |
| toy | shop=toys |
| train | train=*; building=train_station; station=train; attraction=train |
| transport | public_transport=*; type=public_transport; transport=*; transportation=* |
| tree | natural=tree; natural=tree_row; trees=*; landcover=trees; landuse=trees |
| tunnel | tunnel=*; type=tunnel |
| tyre | shop=tyres; service=tyres |
| valley | natural=valley; valley=*; |
| wall | barrier=city_wall; wall=castle_wall; historic:barrier=wall; historic=wall; historic=city_wall; historic=stone_wall |
| war | memorial=war_memorial; tomb=war_grave; memorial:type=war_memorial; cemetery=war_cemetery; monument=war_memorial |
| water | natural=water; water=*; amenity=drinking_water; man_made=water_tower; man_made=water_well; man_made=water_works; type=water; leisure=water_park; amenity=water_point; water_supply=*; water_source=*; substance=water; man_made=water_tank; attraction=water_slide; amenity=water |
| wedding | clothes=wedding; shop=wedding |
| wildlife | zoo=wildlife_park; leisure=nature_reserve; |

| | |
|---|---|
| woodland | natural=wood; landuse=forest |
| zoo | tourism=zoo; zoo=* |